Urban Big Data Centre

Data and expertise for improving cities

www.ubdc.ac.uk

An exploration of fuel poverty in the private rental housing market

Transcript from webinar video recording

1

00:00:00,720 --> 00:00:03,120

[Qunshan Zhao] Today I will just briefly

2

00:00:04,000 --> 00:00:09,920

introduce to you the research I have done. So, can you see my slides now?

3

00:00:15,120 --> 00:00:22,480

Muir can you see my slides? I'll just make sure. Yes, you can see, ok.

4

00:00:23,200 --> 00:00:33,120

So today the webinar topic is about

exploration of fuel poverty in the private rental

5

00:00:33,120 --> 00:00:40,560

housing market in the city of Glasgow. As Muir introduced, I'm a lecturer in urban analytics

6

00:00:41,200 --> 00:00:47,840

in the University of Glasgow and also based in Urban Big Data Centre, which is one of the

7

00:00:49,360 --> 00:00:55,920

research centres and also a data infrastructure funded by the UKRI

Economic and Social Research

8

00:00:55,920 --> 00:01:05,840

Council called ESRC. So just a bit of the personal introduction of myself. So, I mean the

9

00:01:05,840 --> 00:01:12,560

general research interest I have is using an urban analytical approach to soft social economy

10

00:01:12,560 --> 00:01:18,720

and environmental problems in our cities. Particularly with the new forms of urban big

11

00:01:18,720 --> 00:01:26,320

data right now. So, comparing to probably 10, 20 years ago we have traditional urban data. But

12

00:01:26,320 --> 00:01:33,760

under the big scene, the research scene, a few approaches I use including the GIScience

13

00:01:33,760 --> 00:01:39,840

all the way back to my original background in remote sensing and geography information systems.

14

00:01:40,480 --> 00:01:47,520

Also, photogrammetry and also using some of the spatial analytical methods. But besides that, in

15

00:01:47,520 --> 00:01:54,240

the more geography urban planning background, I have also used many different other things such as

16

00:01:54,240 --> 00:02:00,800

machine learning, general statistics, operations research, and you've probably heard about it and

17

00:02:00,800 --> 00:02:07,120

people call it optimisation. Since the network, that's like remote

18

00:02:07,120 --> 00:02:13,280

sensing by connecting all the different sensors - sensor network, stationary sensors,

19

00:02:13,280 --> 00:02:19,680

portable sensors, etc. And also, I've done a piece of work on urban climate modeling and instrumentation.

20

00:02:20,640 --> 00:02:27,840

So that's the approaches I use in my research and many things can be adding up later on. In terms of

21

00:02:27,840 --> 00:02:35,200

teaching, so I handle three masters courses in the University of Glasgow and I'm also a

22

00:02:35,200 --> 00:02:43,840

Masters in Urban Analytics course convener here in Glasgow. And actually, Réka is one of our very

23

00:02:43,840 --> 00:02:53,200

best students from the first year of the Masters in Urban Analytics. So, she's helping today. So the

24

00:02:53,200 --> 00:02:58,560

course I teach including the big data urban analytics, more like a conceptual understanding of the big

25

00:02:58,560 --> 00:03:06,480

data of analytics. The programming tools talk about a few different programming tools including Python,

26

00:03:07,440 --> 00:03:13,520

SQL and R. And today, in the second half of the tutorial, I will talk about some of how I actually

27

00:03:13,520 --> 00:03:21,360

achieve my research results. So, I took the notebook. And also, the final course I teach is called Urban

28

00:03:21,360 --> 00:03:29,760

Analysis Group Project. So that's the kind of particular course for all our students.

29

00:03:29,760 --> 00:03:34,960

Basically, three or four students

work together as a group and then we

30

00:03:34,960 --> 00:03:40,880

have some of the real-world group projects from industry, from third sector.

31

00:03:41,920 --> 00:03:46,720

And they work together for three or four months. So, if you're interested in more of my background

32

00:03:47,440 --> 00:03:52,240

feel free to go to my website, to look at it, and also you can look at some of my publications there.

33

00:03:54,800 --> 00:03:57,840

So yeah, Muir already briefly introduced the

34

00:03:59,040 --> 00:04:06,080

Urban Big Data Centre and I just

want to give you slightly more introduction here.

35

00:04:06,880 --> 00:04:14,240

So Urban Big Data Centre is a research centre right now and that is jointly funded by ESRC and

36

00:04:14,240 --> 00:04:21,760

the University of Glasgow starting from 2019 to 2024. And we are already here for 10 years,

37

00:04:21,760 --> 00:04:28,560

we get started in 2014 and the first five years we serve as a data service centre. In the second

38

00:04:28,560 --> 00:04:35,200

five years we switch to a research centre but also maintain our data service continuously.

39

00:04:36,000 --> 00:04:42,720

So, the general objectives of UBDC is we try to promote the use of big data

40

00:04:42,720 --> 00:04:48,080

and innovative research methods to improve social, economic, and environmental well-being in cities.

41

00:04:48,080 --> 00:04:53,840

I'm quite proud and I will talk more

examples today to let you understand how it works.

42

00:04:54,640 --> 00:05:01,520

So, a few pillars of our centre,

including world-leading urban research - that's

43

00:05:01,520 --> 00:05:08,560

what we want to do. And we also host a very good amount of data collection and data services,

44

00:05:09,200 --> 00:05:17,360

which means you can go to the UBDC website and I see Muir has shared that in the chat channel as well. So, you

45

00:05:17,360 --> 00:05:22,240

can go to that website and you can find many other datas there and you can apply too. And

46

00:05:22,240 --> 00:05:27,680

particularly for a few of the commercial data set, typically you need to pay to obtain it, including

47

00:05:27,680 --> 00:05:33,680

the Zoopla data I use today. But through UBDC, if you want to do non-commercial research you

48

00:05:33,680 --> 00:05:41,280

can actually go ahead apply and we will share the data with you under the formalised data license.

49

00:05:42,320 --> 00:05:48,560

We've also done teaching and capacity building, like today. I mean we have a series of

50

00:05:48,560 --> 00:05:55,200

data dive events. So that's actually helping us to help more people to

51

00:05:55,200 --> 00:06:04,560

understand how to use the new forms of data in urban research. So in UBDC you have

52

00:06:05,280 --> 00:06:12,800

four existing research packages including education,

urban governance, housing, and transportation. And we

53

00:06:12,800 --> 00:06:21,760

will soon open a new research package called Urban

Sensing & Analytics in due course. Alright,

54

00:06:22,560 --> 00:06:29,680

so, after a brief introduction of the background, I will go to the research today.

55

00:06:29,680 --> 00:06:35,840

So today we'll talk about fuel poverty. So, the first thing is why that's

56

00:06:37,200 --> 00:06:44,720

emerging questions we want to look at. And I just found out some interesting statistics here.

57

00:06:45,280 --> 00:06:53,040

So, the on the left-hand side, that's the rise of fuel poverty in the UK. And you can find out by,

58

00:06:54,400 --> 00:07:00,160

back to 2002 there's 1.3 million

people under fuel poverty and that's

59

00:07:00,960 --> 00:07:11,440

about 20 percent of UK households. But after 10 years, in 2012, we actually get 6.5 million people,

60

00:07:11,440 --> 00:07:19,920

25 percent of UK households under fuel poverty status. And after four years we've got even more in

61

00:07:19,920 --> 00:07:24,960

nine million people and basically thirty two percent of UK households under fuel poverty.

62

00:07:25,920 --> 00:07:32,720

And I'm sure if we have a finger for 2020, under the COVID-19 situation we'll probably get more

63

00:07:32,720 --> 00:07:40,480

people because the economy is actually heavily hit in the UK. So, we keep seeing the increase of

64

00:07:41,360 --> 00:07:50,000

fuel poverty status, maybe not only in the UK but in many other places. So how to actually help

65

00:07:50,000 --> 00:07:57,440

then, what actually caused the fuel poverty. And the government is actually thinking of the problems right

66

00:07:57,440 --> 00:08:03,200

so, its thinking what's the solutions. So, people need to spend a lot of fuel, pay a lot of money

67

00:08:03,840 --> 00:08:10,560

to actually make, during the winter in the UK, they can make their home warm enough.

68

00:08:11,600 --> 00:08:19,760

So, on the right-hand side the Scottish Government published a new roadmap about how to improve

69

00:08:19,760 --> 00:08:27,440

fuel poverty over entire Scotland. And that's from the report and you can see down in the

70

00:08:28,400 --> 00:08:33,840

second reference, that's a government publication called The Fuel Poverty (Target, Definition and

71

00:08:33,840 --> 00:08:40,960

Strategy) (Scotland) Bill back to 2018. So, it says the homes with fuel poor households to reach

72

00:08:40,960 --> 00:08:49,840

EPC C, basically c-band. So just to give you an idea. So, A is the best in terms of the fuel, in terms

73

00:08:49,840 --> 00:08:56,720

of energy performance in the UK, and G is the worst. And it says if the household is fuel poor

74

00:08:57,440 --> 00:09:05,040

by 2030 their house needs to reach

EPC C and by 2040 it needs to reach EPC B.

75

00:09:07,280 --> 00:09:11,680

But you will need a lot of, I would say, you will need a lot of investment because the household,

76

00:09:11,680 --> 00:09:17,840

always a house needs to be updated and there's many. And fuel poverty is not only the house,

77

00:09:17,840 --> 00:09:24,880

right. It's not only the building but also the income level of the population. So let's

78

00:09:24,880 --> 00:09:30,080

for these two figures, I just want to give you a brief idea of how it is emerging, this question.

79

00:09:32,320 --> 00:09:36,480

So, to understand fuel poverty, so

we have a few different methods.

80

00:09:37,200 --> 00:09:42,640

So, the traditional methods, including the household and the building survey, is basically

81

00:09:43,280 --> 00:09:48,240

traditional ways to do...you think of census, like people that go into the

82

00:09:48,240 --> 00:09:56,080

household and ask how much money you spend on electricity and gas. And also, the building

83

00:09:56,080 --> 00:10:01,040

companies, they actually go to look at each of the buildings and see what are the energy

84

00:10:01,040 --> 00:10:08,960

efficiencies in terms of the building itself and also appliances. And a few other methods, a few

85

00:10:08,960 --> 00:10:14,080

other literatures that use the official

statistics. They say they use the census data,

86

00:10:15,040 --> 00:10:22,400

whatever, in different countries and also they obtain energy cost data from electricity companies

87

00:10:22,960 --> 00:10:29,920

and they use those data to look at, to understand, right. So, from sensors they can easily

88

00:10:29,920 --> 00:10:37,040

obtain income data for the household and they also have the energy cost. They can

89

00:10:37,040 --> 00:10:47,200

understand how much money you spend on energy compared to your income and is that a

90

00:10:47,200 --> 00:10:54,320

particularly high portion of the percentage there. So that's one way to look at it. And also, there's a

91

00:10:54,320 --> 00:11:01,760

lot of, I mean fuel poverty is more like energy policy questions. So, a lot of policy debates,

92

00:11:01,760 --> 00:11:07,360

discussions around this topic. And if you're going into literature the Baker etc

93

00:11:07,360 --> 00:11:15,520

2018 is a common paper published in Nature Energy, I believe. It's one very good journal. And talk about

94

00:11:15,520 --> 00:11:21,200

the fuel poverty, how it looks like

and the qualitative and conceptual discussions

95

00:11:22,080 --> 00:11:29,520

and how to actually achieve it, what's the factors to influence it. So, this is the

96

00:11:30,400 --> 00:11:36,640

brief literature review, like how it looks like in an academic literature review. But what's

97

00:11:36,640 --> 00:11:42,240

next? So that's the question. Is there any other way to actually tackle this problem?

98

00:11:43,760 --> 00:11:50,320

So basically, here in this analysis

today, I want to share with you, is try to use

99

00:11:50,960 --> 00:11:57,120

some of the new forms of urban big data to understand fuel poverty. So, the data we have,

100

00:11:57,120 --> 00:12:04,720

including the Zoopla data. So many of you, I know, today are audiences from all the world so you may not

101

00:12:04,720 --> 00:12:13,360

be quite familiar with Zoopla. But it's similar to other online platforms. In the US

102

00:12:13,360 --> 00:12:19,680

you'll probably use Zillow and Redfin and in the UK you probably use Rightmove as well. And Zoopla

103

00:12:19,680 --> 00:12:27,920

is also another one of the largest online house rental and selling platforms. And if you're

104

00:12:27,920 --> 00:12:35,360

in Asia, and I know this, there are many similar platforms there as well. So, Zoopla data includes

105

00:12:36,400 --> 00:12:44,800

say the rental price, it includes information including when they actually post it and

106

00:12:45,440 --> 00:12:51,040

so a series of data will post it online and people can actually look at it and to choose the

107

00:12:51,040 --> 00:12:59,760

rental property or like buy or sell their properties. So luckily from UBDC we got a budget to purchase

108

00:13:01,440 --> 00:13:08,080

a portion of the Zoopla data. I will

talk more later in the slides.

109

00:13:08,080 --> 00:13:15,440

They provide the data to us in a machine-readable format in CSV. And in the second half,

110

00:13:15,440 --> 00:13:20,720

in the tutorial, I will actually show you a summary of the Zoopla data. So, this data I cannot really

111

00:13:20,720 --> 00:13:26,400

put in on GitHub right now because it needs to go through data license. Because that's

112

00:13:26,400 --> 00:13:30,080

a commercial data set, we purchased it and we have agreement with Zoopla and

113

00:13:30,880 --> 00:13:36,640

they allow us to share with people if they want to use it for non-commercial research. But everything

114

00:13:36,640 --> 00:13:42,640

needs to go through the UBDC entire paperwork and you will eventually get a delivery of the data.

115

00:13:43,520 --> 00:13:48,560

EPC data is come from the

government. It's more openly available.

116

00:13:50,240 --> 00:13:59,680

And you can imagine so right now there's a wave of people, I mean researchers,

117

00:13:59,680 --> 00:14:06,480

I mean public stakeholders, like just shout to the government and say "oh can you just make more

118

00:14:06,480 --> 00:14:12,560

of the government data openly available? Because that will give more and more benefits to the

119

00:14:12,560 --> 00:14:18,800

entire society". And I'm not quite sure, I mean I don't know about your background but one very

120

00:14:18,800 --> 00:14:25,360

good example is the satellite image, one called Landsat from the United States. And back to I

121

00:14:25,360 --> 00:14:33,680

think about 12 or 13 years ago they started to make it entirely free. And actually, making it free,

122

00:14:35,360 --> 00:14:39,840

adding much more values to

the entire society comparing to

123

00:14:40,400 --> 00:14:46,320

putting the data under the paywall and receive the money just by selling the data.

124

00:14:47,040 --> 00:14:53,280

So that's one for a good example. So, the EPC data right now is getting more widely available. So, for

125

00:14:53,280 --> 00:14:59,840

England and Wales, you can actually download four EPC data online studio portal and

126

00:14:59,840 --> 00:15:06,480

they provide a pretty detailed information. I think they provide a full postcode. For Scotland you can

127

00:15:06,480 --> 00:15:12,800

download individual data on the Energy Saving Trust website and I will show you later.

128

00:15:13,680 --> 00:15:17,840

But for the machine-readable data the Scottish Government, in terms of privacy

129

00:15:17,840 --> 00:15:23,840

consideration, they only give us the postcode sector, which is not a full postcode. But

130

00:15:24,880 --> 00:15:31,600

we are trying to push that, for the government to release the full version of the

131

00:15:31,600 --> 00:15:34,320

data set. And that will help us to better understand.

132

00:15:37,200 --> 00:15:43,120

So, a bit more on the Zoopla data. So, as I mentioned you can obtain Zoopla from the UBDC data service,

133

00:15:43,120 --> 00:15:48,800

there's a link, and also if you go to

my GitHub repository for today's seminar

134

00:15:49,520 --> 00:15:57,360

you can find all the links on the top of the notebook. And just a brief

135

00:15:57,360 --> 00:16:07,600

idea of the data range. The data I have is from January 2010 to the end of March 2019. And we'll

136

00:16:07,600 --> 00:16:12,720

continue to receive data from then because we are kind of negotiating contracts with them to get

137

00:16:12,720 --> 00:16:22,080

the continuous coverage of the Zoopla data. So, all data is 188 CSV files and totalling 44.3 gigabyte,

138

00:16:22,080 --> 00:16:26,640

which is certainly big data. It's not

something that you can handle by using

139

00:16:26,640 --> 00:16:34,480

Microsoft Excel or ArcGIS. So, it's not something that can easily be handled by traditional software. So

140

00:16:34,480 --> 00:16:41,360

the coverage we have including England, Wales, and Scotland. The geography unit we have is a

141

00:16:41,360 --> 00:16:48,720

full postcode for Zoopla data. And actually, for each of the Zoopla data we also get some

142

00:16:48,720 --> 00:16:57,520

address information. Also, a link to the real Zoopla website so if we want to really go

143

00:16:59,280 --> 00:17:05,840

further we can look at those information to see if we can identify the address. But that would be

144

00:17:05,840 --> 00:17:15,440

a bit more complicated because it involves some of the text analysis etc. So, for the EPC data

145

00:17:15,440 --> 00:17:23,040

you can obtain through the Scottish Government website and for the machine-readable CSVs.

146

00:17:23,040 --> 00:17:27,680

And as I said if you want to find out individual houses, like say you rent a flat or you

147

00:17:27,680 --> 00:17:33,200

purchase a house in Scotland and you want to see what's the EPC. And typically, you'll get it because

148

00:17:33,200 --> 00:17:38,480

if you have those kind of activities. But if you want to look at some other people's

149

00:17:38,480 --> 00:17:44,800

you can go to the energy saving charts, there's a link, and you can type in the postcode and find out

150

00:17:44,800 --> 00:17:53,440

if there's an archive EPC certificate. So, EPC data machine readable version, the time ranges

151

00:17:53,440 --> 00:18:02,560

from October 2012 to March 2020 and the EPC data is actually 2.5 gigabytes with 30 CSVs.

152

00:18:03,120 --> 00:18:09,760

The coverage is for Scotland and the geographical units are postcode sector. But as I said England, Wales

153

00:18:09,760 --> 00:18:18,080

they have a different portal, and it covers the English coverage pretty good and Wales it's

154

00:18:18,080 --> 00:18:26,640

about half of the coverage and it has a geographical unit of postcode. But from

155

00:18:26,640 --> 00:18:32,800

my discussion here for these two data sets, you can find out these two data are not a comprehensive

156

00:18:33,600 --> 00:18:40,320

representation of the entire private rental market or EPCs. So, the Zoopla, only the

157

00:18:41,520 --> 00:18:48,640

houses or flats under advertisement you might get information. If some flat and

158

00:18:48,640 --> 00:18:53,840

the homeowners live there for 30 years, it's never going on the market, you will not get information

159

00:18:53,840 --> 00:19:00,800

for the data for the house. So, for the Scottish EPC there also are a

160

00:19:00,800 --> 00:19:06,160

few regulations and we'll talk about that in more detail in the tutorial. But generally speaking

161

00:19:06,960 --> 00:19:15,280

only those houses required to take the EPC mean you will get information. So, a few other data.

162

00:19:15,840 --> 00:19:22,320

So, to understand fuel poverty, one of the very important things is the

163

00:19:22,320 --> 00:19:27,040

income level. And for Scottish income information I think they stopped in

164

00:19:27,040 --> 00:19:31,600

the census to obtain that information. But they have a weekly household average income

165

00:19:32,320 --> 00:19:40,320

from another housing survey. So, here's the data I use. And because we have different geographical

166

00:19:40,320 --> 00:19:46,240

units, the weekly household average incomes under the data zone level and the postcode

167

00:19:46,240 --> 00:19:54,160

sector for EPC data and for postcode for the Scottish Zoopla data. So, we need to have some

168

00:19:54,160 --> 00:20:01,280

data alignments through different spatial units. So, I use the Scottish postcode directory files

169

00:20:02,960 --> 00:20:13,440

from the National Registers of Scotland website. So, go to analysis. So, we have

170

00:20:13,440 --> 00:20:22,000

the Zoopla data pretty big - 44 gigabytes - but how it looks like internally. And I just

171

00:20:22,000 --> 00:20:26,640

get excited every time to play around with the data and find out that kind of

172

00:20:27,840 --> 00:20:33,840

explore the unknown situation. And I first read the data in and it takes about

173

00:20:34,800 --> 00:20:41,360

five, seven minutes to read the entire data set into my Python Jupyter Notebook.

174

00:20:41,360 --> 00:20:47,840

It's original data has 1.3 million

rows of records all the way for about

175

00:20:48,640 --> 00:20:56,880

more than eight years of data. And I further narrowed down to Glasgow, a big Glasgow region so

176

00:20:57,680 --> 00:21:04,240

it's not Glasgow city but a big Glasgow region. Then it narrowed down from 1.3 million to 124k

177

00:21:05,280 --> 00:21:12,640

records. And then goes to the Glasgow city and I use the particular postcodes

178

00:21:13,440 --> 00:21:21,280

we will understand like a central area for Glasgow city and that's 71,000 rows of data.

179

00:21:22,320 --> 00:21:28,880

And particularly we're interested in the private rental market right here. So, in Zoopla data

180

00:21:30,880 --> 00:21:38,480

there's a column that differentiates the rental property and, I mean, the homeowners. And after

181

00:21:39,120 --> 00:21:49,840

pulling that out the data further reduced to 18,000 rows. And after the first four steps of the

182

00:21:51,040 --> 00:21:57,840

filtering we kind of got rental

properties information in Glasgow city

183

00:21:58,560 --> 00:22:02,960

in Zoopla data. And that's us doing more of the data cleaning including

184

00:22:02,960 --> 00:22:08,880

removing the missing values, remove the outliers, some of the rental prices -

185

00:22:10,800 --> 00:22:17,680

- it'd be something unbelievable and I would say probably that one's a

186

00:22:17,680 --> 00:22:24,160

kind of very expensive house for people to try to rent it out. So, we remove those outliers. Because we look at

187

00:22:24,160 --> 00:22:29,680

a few properties, right. So, we look at the low incomes and the low rent and also in terms of a

188

00:22:29,680 --> 00:22:38,080

time range we need to correspond with the EPC data. So, if you still remember we have Zoopla from 2010

189

00:22:38,080 --> 00:22:47,840

to 2019 and EPC data from the end of 2012 and all the way through to 2020. And in this analysis

190

00:22:47,840 --> 00:22:57,040

I use the data between 2013 to 2018, I believe, to just make sure that both data is covered the same

191

00:22:58,640 --> 00:23:08,800

period of time. But after doing all those data cleanings, we first removed 3,000 data rows

192

00:23:09,520 --> 00:23:16,400

and then the final step is to calculate a yearly rental cost at a postcode sector level. So

193

00:23:16,400 --> 00:23:22,160

it involves some of the data manipulations and I will present that later in the tutorial.

194

00:23:23,840 --> 00:23:30,800

So, after the first step for the Zoopla data, which is the probably most exciting and most

195

00:23:31,600 --> 00:23:37,200

time consuming and I go to the EPC

data, it's much better because the government

196

00:23:37,200 --> 00:23:41,840

certainly they have a data scientist. They already clean the data in a very good format.

197

00:23:42,560 --> 00:23:50,400

And for EPC data, first I filter the time range to match the Zoopla data. And then I calculate a unit

198

00:23:50,400 --> 00:23:56,720

energy cost per year per square meter from EPC data. So, in this data they have a column called

199

00:23:56,720 --> 00:24:05,280

current energy costs in the next three years and then they also have the indoor

200

00:24:05,280 --> 00:24:14,400

areas for the house or for the flat. And so, I can easily calculate these measures from the EPC data.

201

00:24:15,200 --> 00:24:20,240

And the next step is to actually join the EPC data with the Zoopla data based on a postcode sector.

202

00:24:21,760 --> 00:24:29,120

And after the spatial drawing, based on the postcode, and I calculate the energy cost per year

203

00:24:29,760 --> 00:24:35,840

based on the bedroom numbers through Zoopla data and the unit energy cost through EPC data. So

204

00:24:36,400 --> 00:24:43,840

I found a reference here, so it says in the UK, generally speaking, one bedroom

205

00:24:43,840 --> 00:24:51,920

flat is around 35 square meters and two-bedroom flats, I just make it a bit consistent,

206

00:24:51,920 --> 00:25:00,400

so, two bedroom would be 70 and three bedroom 105. But that's just a rough estimate not

207

00:25:01,840 --> 00:25:09,280

just from the source I see it described like that. But for sure the housing sizes range

208

00:25:09,280 --> 00:25:16,000

from different ways and we can certainly find other sources and find more reliable

209

00:25:16,000 --> 00:25:24,160

resources. But that's the way I calculate - I use the unit energy cost times the

210

00:25:24,160 --> 00:25:28,960

52 weeks and also times the average bedroom in each of the postcode sectors

211

00:25:29,680 --> 00:25:36,720

and get the information for energy costs per year. And then the next step is to actually import the

212

00:25:36,720 --> 00:25:42,960

weekly income data and connect with the postcode shapefiles and just to merge all the data together.

213

00:25:42,960 --> 00:25:49,840

So, it's quite easy to say here but in the actual data science process it's a fun

214

00:25:49,840 --> 00:25:55,840

process to find out all the joins. And the final part is actually to look at the fuel poverty measures.

215

00:25:56,720 --> 00:26:03,600

And here I use three ratios between

energy and rental cost and income also

216

00:26:03,600 --> 00:26:10,400

energy plus rental and income. So, it goes to the next slides about fuel property measures.

217

00:26:11,520 --> 00:26:16,000

So originally, I have this idea to use

Zoopla to look at fuel poverty because

218

00:26:17,920 --> 00:26:21,040

I'm thinking, for a lot of rental properties,

219

00:26:22,320 --> 00:26:30,160

the tenants typically will not have the right to further improve the building. So typically

220

00:26:30,160 --> 00:26:38,240

you are not allowed to do it. And if they

221

00:26:38,240 --> 00:26:45,280

pay a lot of money for rent,

expensive rent, but they also have to live in a

222

00:26:46,080 --> 00:26:49,600

very low energy efficiency house, that would actually

223

00:26:50,240 --> 00:26:57,200

exacerbate their poverty levels. So that's the very beginning.

224

00:26:58,560 --> 00:27:05,680

So, idea says oh why I should look at this and see if that's true from my understanding.

225

00:27:05,680 --> 00:27:11,600

So here I have an estimation from data analysis. I have a yearly energy cost, I have a

226

00:27:11,600 --> 00:27:18,880

yearly rental cost based on the postcode sector and also the yearly household income based on

227

00:27:18,880 --> 00:27:26,160

also at the same spatial unit. And I can actually calculate ratios, say the energy cost divided by

228

00:27:26,160 --> 00:27:33,120

the income, I call it fuel property. But it's actually just a ratio, like how much money

229

00:27:33,120 --> 00:27:39,520

you spend on fuel out of your entire household income. And the second one, I call

230

00:27:39,520 --> 00:27:45,760

it rental poverty but it's like how much money you spend, what's the ratio of money you spend

231

00:27:47,600 --> 00:27:52,080

on rent compared to income. And if we add it together that's the entire ratios.

232

00:27:55,040 --> 00:27:59,040

So, let's go to look at some of the maps here, which is exciting.

233

00:27:59,040 --> 00:28:05,280

So, this is the first one. So, the ratios

between the energy cost and the household

234

00:28:05,280 --> 00:28:12,160

income. So, I don't know how many of you are from Glasgow or will be familiar with Glasgow.

235

00:28:12,720 --> 00:28:18,720

We can certainly find out in the east end we've got a bit more fuel poverty because

236

00:28:18,720 --> 00:28:24,960

that's the kind of areas with low incomes and also old houses in and around the city centre.

237

00:28:25,680 --> 00:28:32,800

But we also see some of the interesting patterns in the north west and north east. So

238

00:28:33,360 --> 00:28:40,720

they have to spend a lot of money on their fuel. So that's the first map from this

239

00:28:40,720 --> 00:28:48,560

analysis. But if you look at the rental properties or say what's the ratio of rent they pay out of

240

00:28:48,560 --> 00:28:57,840

their incomes. And, very straightforward, we find out the city centre is obviously the most expensive.

241

00:28:59,840 --> 00:29:07,120

And you can see the west end, it's also a pretty expensive area where you need the

242

00:29:07,120 --> 00:29:13,680

University of Glasgow. And also, you can find when the distance is growing, moving

243

00:29:13,680 --> 00:29:22,240

away from city centre, you'll see you pay less rent out of your income in terms of a ratio. And

244

00:29:22,240 --> 00:29:30,800

you see a lot smaller is the 17 percent to 22 percent and city centre is around 36 percent to 47 percent.

245

00:29:31,440 --> 00:29:36,000

So, it's pretty high. But if we go to the fuel poverty, go to the previous slide. So the

246

00:29:36,000 --> 00:29:42,800

highest ratio for the fuel poverty is you actually pay about 30 percent of your money

247

00:29:42,800 --> 00:29:48,960

up to the electricity and gas. So that's actually a big chunk of money.

248

00:29:50,160 --> 00:29:57,520

So how it looks when you add it up. So, this is the last map. And you will find out

249

00:29:58,240 --> 00:30:06,080

since the city centre has a pretty high rent and we can kind of estimate that's the areas. But the

250

00:30:06,080 --> 00:30:12,800

finding is quite interesting. We find out the areas near the city centre have the highest ratio to pay

251

00:30:13,760 --> 00:30:23,600

rent and energy together. The highest from around 60 to 75 percent is around city centre and

252

00:30:23,600 --> 00:30:30,320

also the north of the city centre, which is not actually showing up in a previous map. And this kind of

253

00:30:30,320 --> 00:30:37,520

surprised me and you will certainly

need further investigation in future research.

254

00:30:41,680 --> 00:30:48,480

Ok, so after the brief results and a few summary points here.

255

00:30:48,480 --> 00:30:54,880

So, I think the first thing to mention is that fuel poverty has been observed in the east end of Glasgow. And as well as

256

00:30:54,880 --> 00:31:00,640

the northwest and southwest of the city. And high rental cost has been confirmed in the city centre

257

00:31:00,640 --> 00:31:04,880

and it gradually decreased from the

centre to the outskirts of Glasgow.

258

00:31:05,840 --> 00:31:11,040

And we find out some of the most deprived areas are in the north of the city centre in Glasgow,

259

00:31:11,040 --> 00:31:16,960

which is quite interesting and needs further investigation. So here is the

260

00:31:16,960 --> 00:31:22,640

preliminary results of my research, but the result is still quite,

261

00:31:24,160 --> 00:31:28,480

I mean it's just in the first stage, right. So, we don't really go into any of our regressions.

262

00:31:28,480 --> 00:31:34,000

But the results illustrate the value of using new forms of urban big data to understand

263

00:31:34,000 --> 00:31:39,680

fuel poverty. And more generally that's a new way to tackle traditional urban questions.

264

00:31:42,560 --> 00:31:46,880

Alright, so a few future works. So, this one certainly has some capacity

265

00:31:46,880 --> 00:31:52,480

to make it more comprehensive. So, the first thing is that we can extend the study area.

266

00:31:53,120 --> 00:32:00,560

but we can extend it to Glasgow city region. It contains I think 8 or 9 city councils.

267

00:32:00,560 --> 00:32:06,400

We can extend it to major cities in Scotland including Edinburgh including Glasgow including

268

00:32:06,400 --> 00:32:15,280

Aberdeen and many other cities. And these data, the Zoopla data, is available across the UK. So, we can even

269

00:32:15,280 --> 00:32:21,760

extend it to the entire Great Britain including England, Wales, and Scotland. But I'm sure that

270

00:32:21,760 --> 00:32:27,840

will require a lot of data processing power because the data sets are much bigger.

271

00:32:29,360 --> 00:32:34,480

And also, here we look at private rental markets, but how about homeowners? So, it would

272

00:32:34,480 --> 00:32:38,720

be quite interesting to look at homeowners as well. And also, for homeowners we also

273

00:32:38,720 --> 00:32:46,640

get Registers of Scotland data hosted in UBDC. And that can be another resource to look at

274

00:32:46,640 --> 00:32:55,600

and also to combine with the Zoopla data. So, there are many more analyses that can be done. So right now

275

00:32:55,600 --> 00:33:01,120

we don't really explore what parameters actually, I mean, right now we only look at the rental cost

276

00:33:01,840 --> 00:33:09,040

and the energy cost, right. But we can look at what's the reason to cause this

277

00:33:09,040 --> 00:33:15,200

pretty high, resulting in this high cost. And the census data or we can use the social demographic

278

00:33:15,200 --> 00:33:21,120

information to predict also to estimate either through many different ways

279

00:33:21,120 --> 00:33:26,560

in terms of data. So, we use machine learnings, we

use spatial regressions, we use geographically

280

00:33:26,560 --> 00:33:35,120

weighted regressions. So, a few other things we can do in a future analysis. So, I have one related

281

00:33:35,120 --> 00:33:42,160

publication, Réka's here so it's great to highlight her, and so we collaborated on one of

282

00:33:42,160 --> 00:33:48,480

the fuel poverty and income deprivation research. We used the England data set in Bristol and

283

00:33:48,480 --> 00:33:56,800

It was presented at GISRUK this year and also this research will be presented later online.

284

00:33:59,120 --> 00:34:05,120

So that's pretty much what I have today. Thank you very much for your time to

285

00:34:05,120 --> 00:34:12,320

listen. And I hope this research should give you ideas of how to use the new forms

286

00:34:12,320 --> 00:34:20,000

of data to look at housing questions. And also helped you to have a better idea

287

00:34:20,000 --> 00:34:26,560

of how the data looks and how you

can get it from UBDC. And in the second half

288

00:34:27,280 --> 00:34:32,960

I will go to the notebook, the Jupyter Notebook, by using Python. And I will go into more detail

289

00:34:32,960 --> 00:34:40,240

of the analysis and give you some ideas like how I actually achieve all these results from my

290

00:34:41,440 --> 00:34:53,760

programming. Thank you very much. [Réka Vonnák] So we have

some questions, or so far only one question.

291

00:34:53,760 --> 00:34:59,760

But please post all of your questions in the webinar chat. So, Will asked in the beginning of

292

00:34:59,760 --> 00:35:07,200

the presentation, how is fuel poverty defined? There are various ways to

293

00:35:07,200 --> 00:35:15,840

define it, so I don't know what is the exact one that you used. [Qunshan Zhao] So let me see,

294

00:35:19,120 --> 00:35:25,440

right, so the fuel poverty, so it actually has many different ways to define and one of the

295

00:35:25,440 --> 00:35:36,400

literature you can refer to is the Baker etc 2018. In the Nature Energy paper, they

296

00:35:36,400 --> 00:35:42,080

actually describe in different ways, in different areas, they actually have different thresholds.

297

00:35:42,080 --> 00:35:47,120

Some areas say if more than twenty percent of your income is spent on energy there will be fuel

298

00:35:47,120 --> 00:35:51,920

poverty. In some areas say if more than twenty-five percent it will be fuel poverty. And it also

299

00:35:52,560 --> 00:35:59,280

depends on the locations like if you're talking about Scotland and some

300

00:35:59,280 --> 00:36:05,840

like Spain you'll be different

in terms of the areas. Like UK needs a lot of

301

00:36:07,120 --> 00:36:12,000

central heating but previously I lived in Phoenix, so there's also a fuel

302

00:36:12,000 --> 00:36:18,240

poverty problems but that's basically people using AC to cool down a house. So, it's actually,

303

00:36:20,080 --> 00:36:29,200

it really depends on different countries and different areas. [Réka Vonnák] Thank you. Any questions?

304

00:36:29,200 --> 00:36:35,120

We don't have other questions in the chat so we can still wait a couple of minutes. Maybe someone

305

00:36:35,120 --> 00:36:53,840

will come up with another one. [Qunshan Zhao] Yeah feel free to

put questions in the Q&A, I'm happy to answer.

306

00:36:58,880 --> 00:37:04,720

[Réka Vonnák] I don't see any other ones, so maybe we can have the break now and let's come back in 15

307

00:37:04,720 --> 00:37:12,480

minutes for the tutorial. [Qunshan Zhao] Sure, yeah, that's good. Let's come back after 15 minutes.

308

00:37:13,440 --> 00:37:18,720

And so, before we go, I will show,

309

00:37:20,960 --> 00:37:26,080

I will show the website. So, if you go to my, I don't know how many of

310

00:37:26,080 --> 00:37:33,920

you have used GitHub before, but if you go to my GitHub account and my name is qszhao

311

00:37:35,200 --> 00:37:44,000

and there's a repository called UBDC Data Dives 2020. And I put the

312

00:37:44,800 --> 00:37:52,640

slides and also a notebook layer. So, if you open a notebook you can see

313

00:37:52,640 --> 00:38:00,400

the code layer. So, I will briefly introduce the analysis to you in about 15-20 minutes.

314

00:38:02,880 --> 00:38:09,040

[Réka Vonnák] Yes, I posted the link in the chat so everyone can have a look at the GitHub record.

315

00:38:09,040 --> 00:38:17,840

[Qunshan Zhao] Very good, alright, so I will see you all back at around 11.

316

00:38:21,040 --> 00:38:25,440

You'll find a few repositories

317

00:38:26,400 --> 00:38:36,240

I use here but the UBDC Data Dives 2020 is the one we use today. And here we have two files, one is

318

00:38:36,240 --> 00:38:42,080

the slides that we explained in the first session and the second one is called Zoopla fuel poverty.

319

00:38:42,080 --> 00:38:49,280

I click on notebook, which is the code we will go through.

320

00:38:50,320 --> 00:38:58,880

So here we go, so if you want to have the data you can go to here, the main page of this

321

00:38:58,880 --> 00:39:05,200

repository, and you go to code - it's green buttons - and if you don't know how to use git

322

00:39:05,200 --> 00:39:12,000

that's fine. And also, I don't know how many of you have a background of Python and doing the

323

00:39:12,000 --> 00:39:18,560

programming and using the git but the easy part is you can just download a zip file directly

324

00:39:19,200 --> 00:39:24,240

and you will get an entire folder. You'll get one folder for all the data here,

325

00:39:24,800 --> 00:39:31,600

all the files here. So, I don't put the

data in a repository because they are pretty huge.

326

00:39:32,240 --> 00:39:36,960

Even EPC data is 2.5 gigabytes, so I don't upload it here.

327

00:39:37,760 --> 00:39:45,920

But you can find out in the first page, your first chunk, you will find a few of the data source

328

00:39:45,920 --> 00:39:51,600

and you can click the link actually through your website if you opened it. And the Zoopla data

329

00:39:52,240 --> 00:40:00,160

that's the UBDC website about the Zoopla data and it introduces a variety of different details.

330

00:40:01,120 --> 00:40:08,560

So, we also have the API to get Zoopla data and also a few of the

331

00:40:09,520 --> 00:40:17,200

processing historical data information, the data tables. So, in UBDC we have a team of data scientists

332

00:40:17,760 --> 00:40:24,480

to focus on processing data and analysing data and to provide the best service to our data

333

00:40:25,200 --> 00:40:30,720

users. So, you can find out a lot of information here on the web page. And if you want to apply

334

00:40:30,720 --> 00:40:37,840

you can just go to 'apply to use data' and you will receive correspondence from our staff.

335

00:40:50,000 --> 00:40:54,160

Ok, so the EPC data, if you opened it,

336

00:40:58,880 --> 00:41:04,640

Yeah, it's a bit slow here. But this is

the 2020 data set from Scottish Government.

337

00:41:04,640 --> 00:41:10,320

You can download it directly from

their website and that's the

338

00:41:11,200 --> 00:41:22,720

source. And Energy Saving Trust

Scotland, if you search 'EPC data Scotland'

339

00:41:25,200 --> 00:41:32,800

and you can find something called Scottish EPC Register. And here you can download the individual EPC,

340

00:41:32,800 --> 00:41:40,240

search by postcode. So, this is the way you actually get a PDF, but if you want

341

00:41:40,240 --> 00:41:46,880

a machine readable you will go here to the website. Ok, so weekly household incomes, you can

342

00:41:46,880 --> 00:41:50,880

open a link, I will not do it here, and also the postcode shapefile you can open a link here.

343

00:41:53,040 --> 00:42:03,520

so, it actually contains, including postcode unit, postcode sector, postcode district. And also, in postcard unit it

344

00:42:03,520 --> 00:42:09,040

has corresponding data zones in that information. So, you'll be

345

00:42:09,040 --> 00:42:14,480

pretty happy that's a pretty helpful data set. So, I don't know how many of you know Python

346

00:42:14,480 --> 00:42:19,840

but generally speaking, you can just simply install Anaconda Python distribution if you want.

347

00:42:20,400 --> 00:42:25,840

And then you can run, I mean I only use pandas and geopandas in my code so that's

348

00:42:25,840 --> 00:42:34,960

not a lot of package. Let me see the GitHub repository. Alright, so you can

349

00:42:34,960 --> 00:42:46,320

see my code through GitHub but I will open a notebook on my

350

00:42:49,760 --> 00:42:52,800

computer and you can see it.

351

00:42:53,600 --> 00:43:00,000

So, I already explained some of the information here, so I will not go through that, but here

352

00:43:03,200 --> 00:43:07,040

is very simple, I mean if you

know Python. I don't know how many have

353

00:43:07,040 --> 00:43:11,840

Python, but I will just give a brief

introduction of the entire code process.

354

00:43:12,880 --> 00:43:18,000

And if you have any more questions, we can discuss them. So here I just

355

00:43:18,000 --> 00:43:23,680

import the package, the pandas. And the glob package, it helps to read the Zoopla data in.

356

00:43:24,400 --> 00:43:31,600

and here's the ID for the path of the data and then just adding all the Zoopla data

357

00:43:32,400 --> 00:43:40,960

to the dataframe. And I already ran it here, but it takes about on my very good desktop

358

00:43:40,960 --> 00:43:47,520

computer in the University and that takes about probably seven to eight minutes. So, it's reading

359

00:43:47,520 --> 00:43:54,800

a huge amount, 44 gigabytes of data into the dataframe. And here's a brief idea of

360

00:43:54,800 --> 00:44:02,080

how it looks like for the Zoopla data set. So, it has the listing IDs, property IDs and I

361

00:44:02,080 --> 00:44:08,720

can actually show you the Zoopla data here. So, here's the Zoopla data you can see in a spreadsheet.

362

00:44:09,840 --> 00:44:14,640

so you have listing IDs, property IDs and a lot of different information. Like this is

363

00:44:14,640 --> 00:44:21,840

the price. We have the postcode outcode and incode. And also, they have the price change.

364

00:44:24,080 --> 00:44:30,640

So many others: number of

bedrooms, number of floors, number of bathrooms,

365

00:44:31,440 --> 00:44:42,960

the last marketed date. Many information here, postcodes, yeah many information. Ok, so you

366

00:44:42,960 --> 00:44:51,840

can see like the entire dataframe is 1.3 million, no it's 13 million actually. It's a

367

00:44:51,840 --> 00:44:59,840

huge data set. And I fill the data to Glasgow by using post town and it quickly filters down to

368

00:45:00,880 --> 00:45:10,320

124,000 rows. And then I further filter the data based on the city postcode.

369

00:45:10,320 --> 00:45:17,840

So, you can see I have a list of like from G1 to G5. G12 is the University of Glasgow.

370

00:45:18,800 --> 00:45:25,440

And it's all the way to G53. And the code is basically filtering to the Glasgow

371

00:45:25,440 --> 00:45:32,880

city and now we have 71,000 rows of data. And since we're interested in the private rental market

372

00:45:32,880 --> 00:45:40,320

in this analysis. So, we use the listing status and set it to rent and the data is further reduced to

373

00:45:40,320 --> 00:45:48,320

17,000. And now we start to do some data cleaning. And first thing I print out all the column names.

374

00:45:48,320 --> 00:45:53,520

It's just helpful for me to remember what's there in the data set. So, the first thing I want to use

375

00:45:53,520 --> 00:46:00,800

is the last marketed date, which is this thing has been removed from the Zoopla website.

376

00:46:00,800 --> 00:46:07,680

So, it kind of means that it's reached the end of the advertisement and the rental activity

377

00:46:07,680 --> 00:46:15,200

is finished. But the first thing is I find out a lot of zeros in the data set. So, I

378

00:46:15,840 --> 00:46:24,400

basically say ok if you have zeros, I'll just remove it. And it shows that we reduce from

379

00:46:24,400 --> 00:46:32,480

17,000 to 16,000. But there's

other ways, like we have the first marketed day and

380

00:46:32,480 --> 00:46:37,760

we look at that information if we want to use it. So, you know data analysis there's always different

381

00:46:37,760 --> 00:46:44,640

rationales to make decisions. So, a lot of things can be discussed across the entire process.

382

00:46:46,240 --> 00:46:54,320

And the next thing I want to do is

to actually convert the panda series to datetime

383

00:46:54,320 --> 00:47:04,080

format. And the things like this will make it possible to look at different years and to

384

00:47:04,080 --> 00:47:12,320

match up with the EPC data. So, the next one I try to analyse the bedroom numbers and we find out

385

00:47:12,320 --> 00:47:17,520

we have a lot of zero bedrooms. I don't really know what happens there, so maybe studios,

386

00:47:17,520 --> 00:47:25,920

and how they count it. And we also have 22 bedrooms in your data set. So, I think those are

387

00:47:25,920 --> 00:47:31,200

quite, I mean a lot of them are basically I would say outliers from a statistic perspective.

388

00:47:31,200 --> 00:47:36,560

So, I only keep the bedroom numbers ranging from one to five and remove outliers. But basically

389

00:47:36,560 --> 00:47:42,000

we only remove about, I mean it's only more than one percent of the data here. So, it's not a

390

00:47:42,000 --> 00:47:51,840

quite a big deal. So, we reduce to almost 16,000 rows of data and that's the new dataset we have.

391

00:47:53,120 --> 00:48:00,480

And in the Zoopla data list there's also a status column that shows the status of the rental. And you can

392

00:48:00,480 --> 00:48:08,240

see that we have the rent 'under offer', 'rented' or 'to rent'. So generally speaking, we can say

393

00:48:08,240 --> 00:48:16,080

ok probably 'to rent' is not yet rented out but just for analysis I want to make it

394

00:48:16,080 --> 00:48:22,880

simple and I just include all the data here. So, this can be discussed if we have, we look at UK-wide

395

00:48:22,880 --> 00:48:28,960

or Scotland-wide - do we still want to include those two rental parameters in the data.

396

00:48:31,040 --> 00:48:38,400

So, the next cleaning is based on time and since as I said EPC data and Zoopla data have a different

397

00:48:38,400 --> 00:48:48,080

time range and we want to align them together. So here I just basically clean the data to from

398

00:48:48,080 --> 00:48:58,640

January 2013 to end of December at the end of 2018

399

00:48:59,200 --> 00:49:06,000

to ensure they have the same time coverage. And one thing it is worth mentioning, I don't really separate

400

00:49:06,000 --> 00:49:13,440

different years but that is something that can be done accordingly because we can only look at like in

401

00:49:14,240 --> 00:49:21,600

2013 how it looks like, 2014 how it looks like, or say we can look at 2013 plus 2014 and we can look

402

00:49:21,600 --> 00:49:29,280

at 2017 and 18 and see if we have actually any rent increase. Sometimes that happens. So, by

403

00:49:29,280 --> 00:49:33,920

that reason I don't really separate

them out, but that's a potential direction

404

00:49:33,920 --> 00:49:40,880

to pursue in the future. So, you can look at how many datas we have in 2013 to 18. We

405

00:49:40,880 --> 00:49:50,800

kind of cover the majority of information for sure. So, we further reduced to around 15,000 rows of data.

406

00:49:53,520 --> 00:50:02,000

And since the Zoopla data provide a weekly rental price, so I just sort the rental price from

407

00:50:02,000 --> 00:50:09,760

the highest to lowest and just find out some interesting outliers that one week you pay 87,000

408

00:50:09,760 --> 00:50:19,920

UK pounds and to me it's quite strange data and it can be either errors or it can be outliers.

409

00:50:20,640 --> 00:50:29,040

So, I set the weekly rental price less than 500, but 500 is still pretty high actually so you have 500

410

00:50:29,040 --> 00:50:37,200

times four, it's like two thousand pounds a month. It's in a fluid area in the UK

411

00:50:37,200 --> 00:50:44,640

in Glasgow, that's probably happened, but that's just more like

412

00:50:44,640 --> 00:50:51,680

experience, like roughly estimate. But actually, when I do it, it removes the top one percent.

413

00:50:53,360 --> 00:51:00,640

So, here's the code, how I do it in pandas, and it only removes the top one percent of the outliers.

414

00:51:00,640 --> 00:51:06,000

So, I don't think that really influenced that much. But that really helps to us to

415

00:51:06,000 --> 00:51:11,360

focus on fuel poverty. So that's the data cleaning process. You can find

416

00:51:11,360 --> 00:51:18,080

more processes to looking at the data. So, in any of the data science projects

417

00:51:18,080 --> 00:51:23,200

you always need to look at the data and look at different fields you have and you need

418

00:51:23,200 --> 00:51:29,040

to play around with different fields and see what kind of criteria can help you to better

419

00:51:30,080 --> 00:51:35,840

extract the data set from a huge data set. So, when we talk about urban big data, right, so

420

00:51:36,560 --> 00:51:44,800

44 gigabytes, but we're eventually down to like 14,000 rows in this analysis. If I export the

421

00:51:44,800 --> 00:51:53,280

CSV it's probably only several megabytes. So big data can be narrowed down to small data,

422

00:51:53,280 --> 00:51:58,960

relatively, I mean it's not real small data, it's still big. But comparing to the

423

00:51:58,960 --> 00:52:04,960

original data set the useful information is actually quite limited. And that's how

424

00:52:04,960 --> 00:52:12,320

it looks like for those kind of new forms of data. So, a few other manipulations. So, I calculate

425

00:52:12,320 --> 00:52:20,240

a yearly rent by the weekly price times 52 weeks and generate the full postcode columns easily.

426

00:52:21,360 --> 00:52:26,960

And also, oh there's a typo here, but generate a postcode sector column to match with the EPC data.

427

00:52:27,520 --> 00:52:34,240

Just to remove the last two digits. And here I show a quick statistics of Zoopla listing

428

00:52:34,240 --> 00:52:44,160

in each of the postcode sectors. And you can find out the smallest Zoopla listing number is in G15

429

00:52:44,800 --> 00:52:51,600

7, there's only 7 listings across from, I mean, thirteen to eighteen

430

00:52:53,200 --> 00:53:05,440

across six years. But the highest is all the way to 639 in G41 3. So, the next thing we'll

431

00:53:05,440 --> 00:53:12,400

do, we'll do an average to calculate average rental cost in each of the postcode sectors and

432

00:53:12,400 --> 00:53:19,600

these numbers actually will be pretty important because if we only have seven that's probably,

433

00:53:19,600 --> 00:53:25,280

it depends on the area of the postcode area, I mean the size of a postcode area, but it really

434

00:53:26,560 --> 00:53:36,720

shows the representative of the price. Averaging seven listing price comparing to averaging 600

435

00:53:36,720 --> 00:53:42,240

listing price, certainly you'll get a different, I mean the data is not in the same level

436

00:53:42,240 --> 00:53:50,000

of reliability, right. Ok, so one thing we can do we can remove some of these

437

00:53:50,000 --> 00:53:55,840

small numbers but to ensure we have a full spatial coverage I don't do it here.

438

00:53:56,960 --> 00:54:01,200

So yeah, another thing is that in the Zoopla data we have property type information

439

00:54:01,840 --> 00:54:08,960

and in our final Glasgow city data set the majority is actually flats.

440

00:54:08,960 --> 00:54:17,280

It's about 13,000 and we have 400 terrace houses. We have some detached

441

00:54:17,280 --> 00:54:24,160

houses, we have some cottage houses and we also have some semi-detached houses, and we have some

442

00:54:24,160 --> 00:54:29,520

bungalows, you can find out. So different types here. So yeah, just give you, I mean majority is

443

00:54:29,520 --> 00:54:33,840

still flats and you can imagine that's kind of how it happens, like in the city

444

00:54:33,840 --> 00:54:40,240

centre and the west end most of the flats are rented out frequently and change tenants so that's how

445

00:54:40,240 --> 00:54:46,480

it looks like. So, the next one is to calculate the mean value for yearly rental cost in each

446

00:54:46,480 --> 00:54:55,760

postcode sector. And I use the group by function in pandas. And you can find out the average

447

00:54:56,800 --> 00:55:04,080

yearly rent is ranging from around 5,000 pounds a year all the way in G34 and all the

448

00:55:04,080 --> 00:55:12,560

way goes to the highest and goes to 12,000 pounds a year in G2. So that's the Zoopla data cleaning.

449

00:55:14,240 --> 00:55:20,960

So, the next thing is I go to the EPC data. Zoopla data is finished here.

450

00:55:20,960 --> 00:55:28,160

And the EPC data, as I mentioned the introductions in the first part, so that's

451

00:55:28,160 --> 00:55:34,960

a few of the information. Not all EPCs are available on the register and buildings newly constructed

452

00:55:35,680 --> 00:55:43,440

after January 2013 and also if the dwellings sold or rented out after December 2008

453

00:55:43,440 --> 00:55:48,240

they are required to have a new EPC. And also, the non-domestic building sold or

454

00:55:48,240 --> 00:55:54,480

rented to a new tenant and the public buildings from January 2013 they need to have an EPC.

455

00:55:54,480 --> 00:56:01,440

So, it's not a full coverage of the entire building and the entire housing market in Scotland but we

456

00:56:01,440 --> 00:56:04,960

keep accumulating right, the data keeps accumulating. We have more data.

457

00:56:07,280 --> 00:56:16,960

So, EPC has 2.5 gigabytes, 30 CSVs. We have coverage in Scotland and so I can show you the

458

00:56:16,960 --> 00:56:23,040

EPC data here a little bit. So, here's

the EPC data. You can download by yourself as well.

459

00:56:23,040 --> 00:56:30,000

So, I use the date of assessment, as I mentioned, that's when they generate EPC certificates.

460

00:56:30,000 --> 00:56:35,040

And they have a total floor areas and that's what I use to calculate a unit

461

00:56:35,920 --> 00:56:44,000

energy cost. But one thing that is quite useful is the current energy rating efficiency. And

462

00:56:44,560 --> 00:56:48,960

one thing I just don't have time to do is to actually generate histograms to

463

00:56:48,960 --> 00:56:58,720

show what's the current Scotland EPC bands and that will be useful, right. So

464

00:56:59,360 --> 00:57:02,320

according to the government documents they want to

465

00:57:02,960 --> 00:57:09,120

move all the private, I think they want to move all private housing EPC to at least E

466

00:57:10,000 --> 00:57:14,640

by 2020, I think this year. I don't remember exactly in the government documents, but they have a few

467

00:57:14,640 --> 00:57:20,480

goals. But you'll be good to look at the data and see how difficult to achieve it. So, there are many

468

00:57:20,480 --> 00:57:26,720

information here and you can explore like on water heating how much you spend and

469

00:57:27,760 --> 00:57:34,880

also the environmental information. Ok let's back to the

470

00:57:36,320 --> 00:57:44,960

code. So, we read the CSV file the same as previously the Zoopla data into the EPC frame.

471

00:57:44,960 --> 00:57:53,840

So, we have about one 1.2 million. And I use the date of assessment as the EPC obtained date.

472

00:57:54,720 --> 00:58:06,800

The code actually finds out a few errors in EPC data. So, it says the sum of years like in 30

473

00:58:06,800 --> 00:58:14,560

12. So actually they have typos in data sets so that can be actually corrected

474

00:58:14,560 --> 00:58:22,720

by the government. And you can find out some 1970s exist and it's also 2103 something like. So those

475

00:58:22,720 --> 00:58:28,000

kind of things you can find out in the data sets actually error data. And I further just

476

00:58:30,960 --> 00:58:40,000

eliminate and filter down to 2013

2018. There's a code here and also,

477

00:58:40,640 --> 00:58:47,600

as we know, the EPC data is only

at postcode sector and this part is just to

478

00:58:48,560 --> 00:58:54,880

extract the EPC data to Glasgow city based on the postcode, same as the Zoopla. And then

479

00:58:56,400 --> 00:59:04,320

the next step is to actually calculate the energy cost per year per square meter.

480

00:59:04,320 --> 00:59:11,200

And based on the total current energy cost over three years, divided by three, right, and then also

481

00:59:11,200 --> 00:59:17,840

divided by the total floor areas. So, this is how I calculate it. You'll need cost.

482

00:59:20,000 --> 00:59:24,640

So, here's the same as Zoopla data,

we want to know how many EPC records

483

00:59:24,640 --> 00:59:34,560

in each of the postcode sectors. And we find out we have a few ones like G20, G21, G32 5.

484

00:59:35,440 --> 00:59:47,120

And but we have really large areas that goes to 2,660. In G53 7 we have a 2,600 EPC certificates.

485

00:59:47,680 --> 00:59:53,200

So, this is actually quite an interesting phenomenon because

486

00:59:53,200 --> 00:59:59,200

you can find out what's the, so there's three conditions to obtain new EPCs and

487

00:59:59,760 --> 01:00:08,720

lodged in the Energy Saving Trust register. And I would probably say that G25 is a very

488

01:00:09,440 --> 01:00:15,920

stable neighbourhood, not like many people actually sell their house or rent their house

489

01:00:15,920 --> 01:00:22,400

out. It's quite a stable neighbourhood. But some other postcodes, that's actually

490

01:00:23,040 --> 01:00:29,360

frequently change tenants, frequently buying and selling houses. So that can also kind of

491

01:00:29,360 --> 01:00:36,880

represent the neighbourhood dynamics in some way. So, the next thing is I calculate the mean

492

01:00:36,880 --> 01:00:43,200

of unit energy costs in each postcode sector by using a groupby, same as the first part in Zoopla.

493

01:00:44,640 --> 01:00:49,920

And you can find the unit energy cost from around seven pounds

494

01:00:53,040 --> 01:00:59,680

per square meter per year to all the

way to 23 pounds. Really large but mostly

495

01:01:00,320 --> 01:01:08,320

up to 12 pounds. So, I finished all the cleaning and manipulations. And then next is join

496

01:01:08,320 --> 01:01:16,400

two data together and I use the merge function and based on the postcode sector. And the next one

497

01:01:16,400 --> 01:01:23,520

is calculate, as I mentioned in my first part, calculate energy cost per year based on

498

01:01:23,520 --> 01:01:30,240

bedroom numbers from Zoopla data and a unit energy cost from EPC data. I mean that's

499

01:01:30,240 --> 01:01:36,080

that's just the average number of bedroom statistics. You can find out the average you're

500

01:01:36,080 --> 01:01:42,720

bound to around two bedrooms. But still remember I remove zeros, remove really high bedroom numbers.

501

01:01:42,720 --> 01:01:49,600

So, we have only range from one to five here and so the minimum is around 1.3

502

01:01:50,240 --> 01:01:53,760

and the mean is around two and

maximum is around three bedrooms.

503

01:01:57,520 --> 01:02:02,720

And as I mentioned I use the space

standard for homes I think from RIBA

504

01:02:02,720 --> 01:02:11,360

Royal Institute of British Architects and it says it has a regulation for a one bedroom

505

01:02:11,360 --> 01:02:17,600

house, two-bedroom house, three-bedroom house. And then I use the correspond measures to

506

01:02:17,600 --> 01:02:23,840

calculate the energy cost per year. So basically, average number of bedroom times 35

507

01:02:24,720 --> 01:02:30,720

times Zoopla EPC unit energy

cost and I generate this new view.

508

01:02:32,000 --> 01:02:38,400

And the next thing is to basically

add up information, adding up energy cost

509

01:02:38,400 --> 01:02:47,040

and rental cost and you can see that's from the smallest at 5,600 a year all the way

510

01:02:47,040 --> 01:02:54,160

to 13,000 a year spend totally on rent and energy, which is quite a lot of money.

511

01:02:57,760 --> 01:03:04,880

So here I finish all the analysis for Zoopla and the EPC. And the next thing is to add up the

512

01:03:05,760 --> 01:03:11,680

spatial components of this analysis by using the postcode shapefile and also

513

01:03:12,240 --> 01:03:18,560

add the weekly income information. But that's spatially so it's

514

01:03:18,560 --> 01:03:25,120

using pandas. And I used geopandas to read the shapefiles and the spatial

515

01:03:25,120 --> 01:03:30,960

data set. And that's the National Records of Scotland for the postcode sector shapefiles.

516

01:03:32,320 --> 01:03:42,480

Yeah, geopandas is not as widely used as pandas but it's quite a useful Python package to deal

517

01:03:42,480 --> 01:03:47,920

with the spatial data set, particularly in data manipulations, cleanings around the different

518

01:03:47,920 --> 01:03:57,440

shapefiles. Here's where I just did the postcode district

519

01:03:57,440 --> 01:04:05,920

and filter the shapefile to the Glasgow city area. So same as previously. And I merged

520

01:04:05,920 --> 01:04:12,880

Zoopla data and EPC data with the shapefile based on the postcode sector. And here's just a brief

521

01:04:16,320 --> 01:04:23,920

data, how it looks like. So, you can find a lot of data already connected based on the same sector

522

01:04:23,920 --> 01:04:34,160

number. Then I read the income data from Scottish Government website - panda read CSV. And in the income

523

01:04:34,160 --> 01:04:41,040

data it has the mean value and the median values for each of the data zones. So, it's under the data

524

01:04:41,040 --> 01:04:49,120

zone geography unit. And in this analysis since you always use average values in the Zoopla data

525

01:04:49,120 --> 01:04:57,120

and EPC data, so I also use the average income in this analysis, so I remove all the median data

526

01:04:57,920 --> 01:05:07,840

in the CSV. And since the income

is at data zone level and I will need to

527

01:05:08,640 --> 01:05:14,640

convert it from data zone to

postcode sector and then I can join that with

528

01:05:14,640 --> 01:05:21,360

the previous shapefile with the Zoopla data and EPC data. So, this kind of spatial unit

529

01:05:21,360 --> 01:05:29,760

aggregation is relatively time consuming if you use ArcGIS. I have done that before

530

01:05:29,760 --> 01:05:34,880

but I find R through Python sometimes actually straightforward by using geopandas

531

01:05:34,880 --> 01:05:42,640

and pandas. So, this can be a reference for you to use. And here I filter the new

532

01:05:43,920 --> 01:05:50,320

shapefile to Glasgow city and I find this shapefile is quite useful because you can

533

01:05:50,320 --> 01:05:56,720

find that it has the full postcode, postcard district, postcode sector. It also has

534

01:05:56,720 --> 01:06:04,080

the council numbers. You also have

data zones at 2011 and you also have a few other

535

01:06:04,080 --> 01:06:09,680

things. So it's a very good reference file to actually connect different spatial units together.

536

01:06:11,840 --> 01:06:16,800

And then I merge weekly income data with the shapefile based on the data zone number.

537

01:06:17,600 --> 01:06:24,800

Featurecode here and find out this is the data set. And I calculated mean weekly income for

538

01:06:24,800 --> 01:06:33,840

each postcode sector by using groupby again. And you can find the weekly income ranging from 425 in

539

01:06:33,840 --> 01:06:45,040

G31 all the way to the highest 915 one week. So that's an average value. So that's

540

01:06:45,040 --> 01:06:54,160

how it looks like in terms of income in Glasgow city. So, we go to the final stage to merge

541

01:06:54,160 --> 01:07:01,760

all the data together based on the postcode sector. And this is the final Glasgow shapefile I have.

542

01:07:03,200 --> 01:07:10,320

And after that, you still remember the three maps on my slides? So that's the ratios between

543

01:07:10,320 --> 01:07:18,640

energy cost and income. And here I calculate those ratios and since the weekly

544

01:07:18,640 --> 01:07:26,640

incomes are times 52 weeks. And this is the final data frame. You can find all the information like

545

01:07:26,640 --> 01:07:35,440

all the ratios have been added in the data frame. And another thing is to write the

546

01:07:36,560 --> 01:07:42,160

data frame to a shapefile and then

you can do whatever you want to. So, if you

547

01:07:42,160 --> 01:07:48,080

want to do some of the visualisation you can use ArcGIS. If you want to do

548

01:07:48,080 --> 01:07:53,600

some more analysis, like say if you want to do further analysis like spatial regressions you

549

01:07:53,600 --> 01:08:01,360

can use GeoDa. You want to do GWR you can use the GWR soft files. But also, those kind of things can be

550

01:08:01,360 --> 01:08:08,640

done in Python as well. So, when you have a good shapefile, you have a good data frame, that's easy

551

01:08:08,640 --> 01:08:16,080

plugging to different Python packages. You can use Python to do spatial regression GWR. And if

552

01:08:16,080 --> 01:08:22,240

you want to do visualisations you can use Folium. There are a few other things you can do here.

553

01:08:22,800 --> 01:08:29,760

So, here's the end of the data

analysis. So, I don't do any of virtualisation

554

01:08:29,760 --> 01:08:37,200

here, but I hope this is helpful for you to kind of understand how I reach my data science

555

01:08:37,200 --> 01:08:45,920

data analysis results in my slides. Ok, so it's about 30 minutes, so I will stop here. Any questions?

556

01:08:48,720 --> 01:08:55,280

[Réka Vonnák] We don't have any questions in the chat, but there

seems to be an issue with the EPC data download

557

01:08:55,280 --> 01:08:58,800

but I think that goes down to

personal security settings on

558

01:08:59,600 --> 01:09:07,600

the laptops or computers. [Qunshan Zhao] Yeah, I think if you go

here you should be able to download the EPC data.

559

01:09:08,160 --> 01:09:14,320

So, let's open data set. I don't think

they have any limits, or you need to do any

560

01:09:14,320 --> 01:09:20,400

logins, no, so just freely download. Let's see for today, all the data is fully

561

01:09:20,400 --> 01:09:25,840

available except the Zoopla data. So, if you apply for the Zoopla data from UBDC

562

01:09:27,200 --> 01:09:33,920

and once you have it you can actually run through my notebook easily with all the data. And all

563

01:09:33,920 --> 01:09:38,000

you need to do is just to change

the data path, right. So, you need to change your

564

01:09:38,000 --> 01:09:43,680

data path to your own path. And that's the only thing you need to do and then you can

565

01:09:43,680 --> 01:10:03,440

play around with all this information. Ok? Any more questions? [Réka Vonnák] No

566

01:10:11,040 --> 01:10:15,840

I don't think we have any more questions.

567

01:10:23,680 --> 01:10:30,320

Karen is asking how long this

page will be available to us? [Qunshan Zhao] Yeah so

568

01:10:30,320 --> 01:10:36,400

the GitHub account I will just leave it there I think.

569

01:10:37,680 --> 01:10:45,360

So, you will be able to access it and I don't think I will remove the code. It's quite a

570

01:10:45,360 --> 01:10:51,840

there's no magic there so I only use two package pandas and geopandas.

571

01:10:52,400 --> 01:11:00,320

So, in mostly just the data analysis but I want to just give you, this can serve as a reference to

572

01:11:00,320 --> 01:11:06,000

you. Basically, if you want to do similar things. Probably not the fuel poverty but similar data

573

01:11:06,000 --> 01:11:13,680

analysis by using Python, that can be one of the resources you can look at. [Réka Vonnák] Sarah is saying that she doesn't

574

01:11:13,680 --> 01:11:19,920

have any experience with Python so it's a lot to take in. Yes, I think it is a lot to take in but it's

575

01:11:19,920 --> 01:11:26,400

a good reference to use in the future if you want to improve. [Qunshan Zhao] Yeah if you're

576

01:11:26,400 --> 01:11:33,520

interested in this analysis you can consider joining us in our MSc in Urban Analytics. And we have

577

01:11:34,240 --> 01:11:41,360

world-leading researchers here to teach you how to use it and Réka is one of the

578

01:11:41,360 --> 01:11:48,240

previous students. [Réka Vonnák] I felt the same way. [Qunshan Zhao] Yeah, I believe she learnt a lot

579

01:11:48,240 --> 01:11:54,080

across the one-year period of time and she can do all these things right now I believe. So

580

01:11:56,400 --> 01:12:02,880

[Réka Vonnák] Yeah apply for the masters. [Qunshan Zhao] Yeah if you're from the UK, from Scotland

581

01:12:02,880 --> 01:12:07,840

we also have a Data Lab

scholarship that you can apply to,

582

01:12:12,480 --> 01:12:22,880

Yes, thanks Karen for your statement. Yeah, I think that's helpful but

583

01:12:22,880 --> 01:12:28,560

a lot of these can be presented to the government, right. So sometimes they

584

01:12:28,560 --> 01:12:36,160

don't have the data analysis capacities and this will help them to understand how,

585

01:12:38,640 --> 01:12:46,320

so they have regulations but to accurately identify the areas, that's sometimes not very easy.

586

01:12:46,880 --> 01:12:51,600

And these can actually support their decisions. [Réka Vonnák] Yeah Sarah is asking how do you use your

587

01:12:51,600 --> 01:12:59,520

research to try to influence government? [Qunshan Zhao] Yeah

that's a very good question. And so

588

01:12:59,520 --> 01:13:06,400

the first thing is like for us, as academics and also we're working in social science.

589

01:13:06,400 --> 01:13:12,320

We always try to talk with the government, right. So, we want to establish collaborations and

590

01:13:12,320 --> 01:13:19,840

particularly in UBDC we have a really good connection with Glasgow City Council and Scottish Government.

591

01:13:19,840 --> 01:13:26,400

We have a few projects ongoing between us, I mean on the transportation side but also on the housing side.

592

01:13:27,120 --> 01:13:33,360

And so that's one thing is to actually make contacts with them

593

01:13:34,000 --> 01:13:40,640

and to talk with them frequently, I mean not frequently, but basically we want

594

01:13:40,640 --> 01:13:48,320

to find the right person to talk to. And also, we do research, we publish papers, but after

595

01:13:48,320 --> 01:13:56,160

that it's actually very important for academics to push to the next step, to move your publication

596

01:13:56,160 --> 01:14:02,560

to generate impacts. But how to do it? So, it's like publications, sometimes because you always need to

597

01:14:02,560 --> 01:14:09,440

have some new things, right, you need to have new findings. And sometimes you will be slightly

598

01:14:09,440 --> 01:14:16,240

too advanced for the public sector

readers. So, you want to present them clearly, like

599

01:14:16,800 --> 01:14:24,080

the findings, and also fit to their interest. So sometimes we have similar goals but

600

01:14:24,080 --> 01:14:30,320

slightly different routes. So, one of the important things that we want to merge those two routes and

601

01:14:30,320 --> 01:14:37,040

work together. I think that's the way I believe will be helpful.

602

01:14:37,040 --> 01:14:55,840

[Réka Vonnák] Any more questions? We still have some time so if you have questions just post them.

603

01:15:07,040 --> 01:15:11,280

Oh, we have a new question from

Tatiana. I'm wondering about the

604

01:15:11,280 --> 01:15:18,080

the data on income and its integration with the Zoopla data. You didn't not consider

605

01:15:19,120 --> 01:15:23,360

different years, right? I'm just wondering here is there an assumption that income has

606

01:15:23,360 --> 01:15:29,200

not increased much since 2012 and is this why you only have one data set for income?

607

01:15:30,160 --> 01:15:36,560

[Qunshan Zhao] Yeah so, the income data set is not widely available in Scotland and so they stop to

608

01:15:36,560 --> 01:15:44,560

collect the income data on the 2011 sensors but we have another one coming in 2021.

609

01:15:44,560 --> 01:15:50,400

I don't know if they will have the income data. So, the 2014 income data is the newest data I can find

610

01:15:50,400 --> 01:15:57,360

but there's another option to use the SIMD, the Scottish Index of Multiple Deprivation, so

611

01:15:57,360 --> 01:16:04,400

it updates at data zone level and I think quite frequently, every 18 months. And Réka has used

612

01:16:04,400 --> 01:16:12,000

it before for England, the IMD data. So that's another option but SIMD, there's not exact

613

01:16:12,000 --> 01:16:19,920

numbers of income, it's just a quantile. Like you can range from one to zero and one probably means

614

01:16:19,920 --> 01:16:27,120

the poor ratio, 10 means the richest. But because we want to calculate the cost

615

01:16:27,760 --> 01:16:36,000

exactly, so here I use this Scottish income data. So, I will not say like after six years

616

01:16:36,000 --> 01:16:42,640

the income will not change but also our data for the EPC and Zoopla that's from 2013 to 2018.

617

01:16:43,280 --> 01:16:50,320

So that's kind of within the range of the whole Zoopla and the EPC data sets.

618

01:16:50,880 --> 01:16:58,400

So always the data alignment is difficult spatially, as I show you in the notebook. I have done a few

619

01:16:58,400 --> 01:17:06,000

things to align on different postcode sectors, postcodes, but also I mean temporally we need

620

01:17:06,000 --> 01:17:13,200

to consider when the data actually happened. So, we have Zoopla data, EPC data and we try to

621

01:17:13,200 --> 01:17:21,200

align them together. And income data is relatively ok within the range. And so, we always have some

622

01:17:21,200 --> 01:17:25,920

limitations in the data analysis, but

we just need to find out a way to do it.

623

01:17:27,120 --> 01:17:32,560

I hope that answers your question. [Réka Vonnák] We also have another question from Oleg about using

624

01:17:32,560 --> 01:17:37,280

Python instead of like classic GIS and how you understand that the shape is good enough

625

01:17:37,280 --> 01:17:47,520

to use in Python and not the classic ArcGIS. [Qunshan Zhao] So I think those two are complementary. So

626

01:17:48,800 --> 01:17:53,680

if you really favour open source you can use Python

627

01:17:55,840 --> 01:18:01,760

plus QGIS - that's the open-source package, I mean open-source software of QGIS.

628

01:18:02,960 --> 01:18:09,520

But generally speaking, Python has many packages. And right now, we have more and more GIS related

629

01:18:10,320 --> 01:18:17,760

packages. It's amazing. So geopanda is one of the data manipulations. Typically, previously what I

630

01:18:17,760 --> 01:18:24,560

have done, I have done all a lot of these analyses in ArcGIS and they have like small database

631

01:18:25,360 --> 01:18:30,560

operation systems, so the attribute tables, right, and in QGIS you can do it as well.

632

01:18:30,560 --> 01:18:37,680

But to handle this 45 gigabyte data it's almost impossible to read the data into a

633

01:18:37,680 --> 01:18:44,880

software because typically it can handle small data sets, probably of one gigabyte. I don't

634

01:18:44,880 --> 01:18:48,720

know that will work and it will take

probably take a long time to just read the data.

635

01:18:49,520 --> 01:18:54,320

And because you use a lot of memory. But using Python is much easier. So

636

01:18:54,320 --> 01:19:01,200

I don't want to say like you fully stick to Python and don't learn any of GIS.

637

01:19:02,160 --> 01:19:08,240

I think you need to have some GIS knowledge and particularly not everything available in

638

01:19:08,240 --> 01:19:15,120

say ArcGIS or QGIS and that's available in Python. So, the open-source community will

639

01:19:15,120 --> 01:19:21,840

continue to develop new code. Like for one of my research interests in spatial optimisation

640

01:19:22,640 --> 01:19:28,080

location allocation, location analysis. In ArcGIS you can do it directly, but in Python

641

01:19:28,080 --> 01:19:34,480

right now, it still has a lot of caveats to like how to actually run it in a good way. And also, the

642

01:19:34,480 --> 01:19:42,320

packages how to actually align the GIS data set with the optimisation.

643

01:19:43,040 --> 01:19:48,560

And there's still some barriers. In spatial statistics it has a pretty good progress

644

01:19:48,560 --> 01:19:55,840

in terms of making open source, but in a lot of other things like hydrologies, geologies I mean

645

01:19:55,840 --> 01:20:02,000

ArcGIS can be very versatile and so that's something you need to balance. Also, in

646

01:20:02,000 --> 01:20:10,160

our mastering of analytics we still teach Python, R and GIS together. So, I would say at this moment you

647

01:20:10,160 --> 01:20:20,720

still need to learn all of them but probably after 10, 15 years Python is good enough. [Réka Vonnák] Thanks.

648

01:20:20,720 --> 01:20:25,680

There's another question from Will, which I think is a very good point. If we looked at the

649

01:20:25,680 --> 01:20:32,080

variability within high level postcode areas, so if there are smaller private houses and high rise, but

650

01:20:32,080 --> 01:20:37,120

it might compromise the calculation of average values? I think that's a good question as well.

651

01:20:39,680 --> 01:20:45,840

[Qunshan Zhao] Yeah let me see.

652

01:20:50,640 --> 01:20:56,880

Right, so I would say like a lot of high-rise buildings certainly, say in the city

653

01:20:56,880 --> 01:21:06,480

centres, you see more high-rise buildings. So, one thing is that, what we can do,

654

01:21:06,480 --> 01:21:14,720

So, we can separate out house and flat in the Zoopla data. So many things we can do, right. So

655

01:21:14,720 --> 01:21:22,400

we can further filter the data so we

only look at flat and we compare to house. But

656

01:21:22,400 --> 01:21:27,360

particularly in Zoopla data, we don't really have information to say that's a high-rise building.

657

01:21:27,360 --> 01:21:32,160

I mean we have house

information, but we don't really differentiate

658

01:21:32,160 --> 01:21:38,000

different types of flats. I mean a lot right now near the university there's a lot of like

659

01:21:38,000 --> 01:21:44,080

brand new buildings of student flats and those kind of data are probably not even going into

660

01:21:44,080 --> 01:21:50,000

Zoopla because they have their own systems. So those kind of high-rise flats will not

661

01:21:50,560 --> 01:21:53,840

actually even show up in the Zoopla data set. So

662

01:21:53,840 --> 01:22:01,440

that's what I mentioned. The new forms of urban big data are always biased. So that is not a

663

01:22:01,440 --> 01:22:08,160

like a comprehensive survey, like census. We have stratified samples and then we can

664

01:22:08,880 --> 01:22:15,280

say ok, that's a good representative

for the entire population of the analysis. But here

665

01:22:15,280 --> 01:22:23,200

the Zoopla data, that's a private rental market. And also, another thing is, we can look at Airbnb. So

666

01:22:23,200 --> 01:22:29,280

that's a short-term letting and Zoopla is more like long-term letting or say

667

01:22:29,280 --> 01:22:36,320

long-term renting. And we can also look at Airbnb in Glasgow, that's basically the short-term renting,

668

01:22:36,320 --> 01:22:43,040

and see what's the differences across these two behaviours and do they influence each other.

669

01:22:43,040 --> 01:22:50,080

And time will be quite sensitive information here because I mean in COVID-19 we don't have

670

01:22:50,080 --> 01:22:54,320

that much tourism, particularly in Edinburgh. So, Glasgow is less like a tourism city but

671

01:22:54,960 --> 01:23:02,720

you can see probably a lot of Airbnbs they become long-term, they show up in Zoopla

672

01:23:03,920 --> 01:23:13,440

right now, because there's not many tourists. So many things can be combined and many new

673

01:23:13,440 --> 01:23:22,320

research questions can be raised and can be analysed. So, I think there is huge potential in this

674

01:23:22,320 --> 01:23:29,680

domain to push forward like with different types of data, how to connect them together and how to draw

675

01:23:29,680 --> 01:23:35,680

a useful policy recommendation and information. Yeah, but that's a great question.

676

01:23:37,680 --> 01:23:44,000

[Réka Vonnák] I don't see any more questions. If you have them just please post them.

677

01:23:45,360 --> 01:23:51,920

[Qunshan Zhao] Yeah, we can be here for a few more minutes. If you have a question just post it there.

678

01:23:53,520 --> 01:24:00,640

Or if you have a general question like for our Centre, for our masters programmes or like for

679

01:24:01,680 --> 01:24:08,080

other research I'm happy to answer. [Réka Vonnák] You can also email any questions.

680

01:24:08,080 --> 01:24:16,080

[Qunshan Zhao] Yeah feel free to email myself or Réka for the questions.

681

01:24:16,080 --> 01:24:25,760

We're happy to help. [Réka Vonnák] I don't see any more questions

coming up, so I think we answered everything.

682

01:24:28,960 --> 01:24:29,460

[Qunshan Zhao] I hope so.

683

01:24:32,080 --> 01:24:38,480

Alright, so I will stop here. Let me go back to

684

01:24:42,080 --> 01:24:44,880

Ok I got a question in the Q&A.

685

01:24:47,760 --> 01:24:53,520

So, from Tatiana, ok so wondering

about data of all kinds integration. [Réka Vonnák] Yeah, we answered that.

686

01:24:55,840 --> 01:25:01,360

[Qunshan Zhao] Ok. The assumption that income has not increased much.

687

01:25:03,120 --> 01:25:04,880

Ok, yeah, I answered that question.

688

01:25:06,320 --> 01:25:11,840

[Réka Vonnák] Thanks all for coming. [Qunshan Zhao] Good to see Muir come back.

689

01:25:14,240 --> 01:25:20,400

Oh, you're on mute. [Muir Houston] Sorry, I just did a PhD supervision I was doing there so

690

01:25:22,160 --> 01:25:28,800

hopefully the session went well. [Qunshan Zhao] Yeah, we've kind of drawn to the end of the session and we had

691

01:25:28,800 --> 01:25:36,960

a few good questions but if we don't have any more questions we will stop here.

692

01:25:36,960 --> 01:25:43,040

I would say thank you very much for joining today's webinar and thanks Muir for a nice

693

01:25:43,040 --> 01:25:52,640

housekeeping introduction and thanks Réka for help maintaining the Q&As and helping the entire

694

01:25:52,640 --> 01:25:59,600

process. And I hope this session will give you ideas how to use the new forms of urban big data,

695

01:26:00,160 --> 01:26:09,440

to do some housing research. But also, as I mentioned, we as a data service centre, we host

696

01:26:09,440 --> 01:26:18,320

a variety of different data sets and you have a lot of opportunities to do it. And so

697

01:26:18,320 --> 01:26:26,320

recently, I just want to do an advertisement, we have five posts in UBDC. They are

698

01:26:26,320 --> 01:26:32,320

short-term posts of about four to five months til the end of March. And we have a few options to do

699

01:26:32,960 --> 01:26:41,040

public transfer index, to do mobility research, to do video analytics through the CCTV cameras

700

01:26:41,040 --> 01:26:46,080

and also through the dashboard development. So, if you have an interest you can look at those posts.

701

01:26:46,720 --> 01:26:55,120

And also, the last one, we will have our last UBDC Data Dive by our very own Senior

702

01:26:55,120 --> 01:27:02,640

Lecturer Jing Yao on this Friday about using

703

01:27:02,640 --> 01:27:10,000

GIS to do some of the location allocation analysis. So, look at the website

704

01:27:10,000 --> 01:27:15,600

and if you're interested please register. [Muir Houston] Just to say that we will be putting

705

01:27:16,560 --> 01:27:21,840

these sessions online on YouTube but we need to make sure we've got transcripts and all that for

706

01:27:21,840 --> 01:27:28,000

them due to the new accessibility regulations. So, they'll maybe take a couple

707

01:27:28,000 --> 01:27:34,800

of weeks for us but keep an eye on the UBDC site and you'll see the videos.

708

01:27:35,520 --> 01:27:40,160

[Qunshan Zhao] So I will leave my GitHub repository there, so you have access to it and

709

01:27:40,160 --> 01:27:44,320

the data, you can download open data by yourself and you can apply for the Zoopla data.

710

01:27:44,960 --> 01:27:51,120

And the final video with the transcription will go to the YouTube channel of UBDC.

711

01:27:51,840 --> 01:27:58,560

And keep an eye on our Centre. If you want to subscribe to our newsletter, and we always

712

01:27:58,560 --> 01:28:04,000

have a lot of things going on and also this is the main research

713

01:28:04,000 --> 01:28:11,280

direction in the Economic and Social Research Council, to invest in the data infrastructure

714

01:28:11,280 --> 01:28:19,281

so we are part of it. Alright, thank you so much, you have a good day. [Réka Vonnák] Thank you. [Qunshan Zhao] Bye.