The School of Mathematics

THE UNIVERSITY *of* EDINBURGH

Identifying houses with internal lead piping across Scotland

by

Ruitong Liu, s1700307

July 2020

Supervised by
Bruce Worton, Gemma Aitchison and Mine Cetinkaya

Own Work Declaration

I confirm that the work contained in this report is my own except where otherwise indicated.

Acknowledgement

I am grateful to the supervisors of the project, Bruce Worton, Gemma Aitchison and Mine Cetinkaya and Maria Tovar Gallardo for their advice and guidance. I would also like to thank my parents for supporting me.

Word Count

5000 words(including the executive summary, main text and references; excluding appendices).

Executive summary

Background

Lead is widely present in nature and is considered one of the most serious environmental pollutants. Even small amounts of lead consumed over time can cause lead poisoning. Scottish Water must meet strict regulatory standards on the level of lead in drinking water. The maximum limit (PCV – Prescribed Concentration or Value) for lead in drinking water is 10 microgrammes per litre. Not all house in Scotland have lead pipes but older houses built before 1970 may still have lead piping in plumbing systems or storage tanks.

Research Question

The DWQR would like to identify houses throughout Scotland at increased risk of having internal lead piping and to identify areas in Scotland with relatively large numbers of houses with internal lead piping.

Data

In order to achieve DWQR's goal, I prepare a single data set by collating information from 9 data sets made available by Scottish Water and the Scottish Assessors Association. I merge all useful variables to postcode in standard format to ensure that all information is available at the street postcode level.

Work

In data cleaning section, I first create a cleaning function to delete irrelevant variables or variables contain missing values above 50 percent and keep useful variables. Besides, I convert dates to consistent format using the global "ISO 8601" standard and use consistent variable names such as latitude and longitude. Finally, I merge data through rig and collate variables to a single data set.

In the exploratory data analysis, I perform descriptive statistical analysis of important variables and also give some correlation analysis with plots. Besides, I drew a heat map based on latitude and longitude to better show the distribution of water pollution in Scotland.

Findings

I find that lead level in drinking water are regional. Tay area has the largest number of lead pipes, but houses in this area have a relatively short construction time and house prices are low, so there is a risk of increasing lead level. The number of houses containing lead pipes in the Nith area is second, and the age of the houses in this area is longer. Forth and Ayr have less lead pipes and the proportion of houses built after 1970 is relatively high. So the price of house is also relatively high and the risk of exceeding lead is low.

Through analysis of points that lead exceeds the standard in the samples . I think it is urgent to replace the lead pipes and give water treatment for Nith, Ness and Tay for DWQR.

Contents

1 Introduction

As is all known, lead is considered one of the most important environmental pollutants and is a major threat to human health. Many studies have proved that lead exposure can cause disorders of various body systems such as respiratory, neuralgic, digestive, cardiovascular and urinary diseases. More seriously, children absorb more lead than adults due to their growing bones and other organs which lead can become deposited in. The signs and symptoms in young children include irritability and fatigue, loss of appetite and weight loss, abdominal pain, hearing loss, developmental delay and learning difficulties (Scottish Water).

Therefore, the problem arises when drinking water comes into contact with lead supply pipes, lead tanks, lead solder joints on copper pipes, or inferior quality brass fittings and taps, particularly for longer periods (e.g. overnight). This can result in high lead levels in the drinking water supply. Although the use of lead service pipes was phased out during the 1960's and officially became illegal in 1969 (Fact sheet7, Scottish Water). There may still be a small chance that lead pipes exist in houses built before 1970 or private water supply systems. So the Drinking Water Quality Regulator (DWQR) want to identify houses throughout Scotland at increased risk of having internal lead piping and to identify areas in Scotland with relatively large numbers of houses with internal lead piping.

This report aims to contribute to DWQR's goals by cleaning data, merging data sets and showing exploratory data analysis. I draw heat map to show the distribution of lead pollution in Scotland in an intuitive, clear and precise way. I hope that a better understanding of the information in data sets can help DWQR remove lead pipes and create an entirely lead-free water system in Scotland.

2 Literature Review

Lead is widely present in nature, lead dust and lead-containing oxides in the air can pollute the soil and water, the pollution of lead to soil is cumulative and irreversible (Liu GuoFeng,2006). Toxicity correlates with lead concentration in blood and progresses from biochemical and subclinical abnormalities, at levels around 10 g/dL, to coma and to death at levels more than 100 g/dL(Swaran J.S. Flora,2006). As a result, lead exposure may cause respiratory, neuralgic, digestive, cardiovascular and urinary diseases. While several adverse health effects of heavy metals have been known for a long time, exposure to heavy metals continues, and is even increasing in some parts of the world, in particular in less developed countries, though emissions have declined in most developed countries over the last 100 yearsJärup, L2003). In fact the incidence of occupational and adult lead poisoning has declined in the recent past, the problem still exists. When water lies in contact with lead service pipes, lead-lined tanks, or lead solder, particularly for longer periods (e.g. overnight) it can absorb lead. In Scotland, drinking water must meet strict regulatory standards on the level of lead, that is, the maximum limit (PCV – Prescribed Concentration or Value) is 10 microgrammes per litre(Scotish Water Factsheet 7, 2013).

3 Data

3.1 Data source

This paper uses 9 data sets including 6 from Scotland Water Company and other sources.

1. From Scotland Water Company:
   SW-All Lead WQ Samples (2010-18), SW-Comm pipe data,SW-Lead Comm Pipe, Replacements (2004-2018), SW-Phosphate Dosing WTWs Y or N, SW-Postcodes linked to SW Zonal Structure, SW-Scottish Water Zonal Phosphate Levels.

2. From Other sources:
   Other-UK-HPI-full-file-2019-03, Other-SAA_PropertyAgeData, Other-Postcode_ household count_ urban class.

3.2 Data Cleaning Steps

My process of collating the information in the 9 data sets can be summarized into the following five steps:

1. Preliminary sorting variables
   Firstly, it is found that there are a large number of missing values in multiple data sets, which are filled in with NA. Then I established a cleaning function to delete variables which missing rate is over 50 percent in each data set. This does not mean that they are invalid or useless variables, such as variables related to the price index of the house, ID variables related to regional structure information. In the process of exploring data, I also conducted some exploratory analysis of them.

2. Connect and merge information in different data sets
   I match and link data sets with the same variable, and call merge to collate information. I merge'SW-Scottish Water Zonal Phosphate Levels.xls' with 'SW-Phosphate Dosing WTWs Y or N.xlsx' through 'Rig'variable. Then I also combined the information in 'Common piped data.csv' through 'WTW1' connection to get 'cleaned_pipe.csv'. Besides, I merge data information in 'Other-Postcode_household count_urban class.csv' and'Other-UK-HPI-full-file-2019-03.csv' through 'CouncilArea2018Code','Date' to form 'cleaned_ house data.csv'.

3. Adjust Unit of variables
   Here, I create another function to deal unit of variables. For example, I use function to split unit of meter in 'Estimated length of pipe in meters'. Besides, I also check the unit for Lead, 'Phosphorus' ,'Temperature' and 'Hydrogen ion' and put their unit in separate columns.

4. Standardized the format of date data, split date and time
   There are various date and time format in raw data sets, I converted the date format and reset it to a standardized form YY-MM-DD so that it can be used when fitting models.

5. Adjust variables according to meanings
   In the data set named'cleaned_pipe.csv ', there are 45 columns refer to various variables and 185952 observations in total. Firstly, I am not interested in some variables since they are almost unrelated to DWQR's goal, including 'AR10_PROPERTYID', 'Inspected by', 'Survey this toby?''Toby cover photo reference''How certain is the identification?', 'Comments''Rig', 'WTW1', 'OSAPR'. I delete them directly in final data sets. Then for pipe age, I keep 'AGE', 'Pipe age in years from 2019' to represent the pipe or house age. I delete 'Is the property age pre 1970? , 'Date_Commissioned'.
   For location, I consider that 'Description of Location' can be replace by whether the observation is domestic or not. And 'FULL_ADDRESS', 'Nearside or Farside?' can be

replace by latitude and longitude. ‘District postcode’ has already contains in street postcode. So I delete ‘FULL_ADDRESS’, ‘Nearside or Farside?’,’ District postcode’. I first compute Longitude and Latitude from ’Easting’ and ’Northing’ however I find that there are 149509 missing values in 185952 sample values. So I use ’NORTHING_NUM’ and ’EASTING_NUM’ to get Latitude and Longitude and drop ‘Easting’, ‘Northing’, ‘NORTHING_NUM’ and ‘EASTING_NUM’.
Finally, I delete observations that neither the communication pipe nor water supply pipe is made of lead to form the final data set ‘DF1.csv’.

3.3 Processed Data

The processed data set ‘DF1.csv’ includes 156766 sets of observation points of 26 variables in 8 regions of Scotland. I made the variable dictionary in Table 1.

Variable Names	Description of Variable
AR10_MATERIALL	Type of material of communication pipe
Pipe Material	Type of material of pipe
MAIN_DISTANCE	Distance from the mains pipe
POST_TOWN	Town name of corresponding postcode
Street postcode	Street postcode include district code
RZ_REF	Reference ID of water treatment Zone
RZ_NAME	Name of water treatment Zone
REGION	8 Regions of Scotland
CONFIDENCE_GRADE	Alphanumeric score which qualifies data reliability
AGE	Year communication pipe was installed range 1883-1970
Pipe age in years from 2019	Age of communication pipe until 2019 range:49-136
Date	Date of pipe inspection
Domestic or Non-domestic	if property is domestic or non-domestic
Property Type	Type of property
How many properties are supplied by this comms pipe	Number of properties supplied by this pipe
toby box acessed?	Yes/No
Is there a meter box?	Yes/No
Is there any evidence of leakage?	Yes/No evidence of leakage of lead
Is the property age pre 1970?	Yes/No the property age is before 1970
Estimated length of pipe in meters	Length of pipe(m)
Hydrogen ion	Hydrogen ion level (ug/l)
Lead	Lead level(ugPb/l)
Phosphorus	Phosphorus level(ug/l)
Temperature	Water temperature($^{\circ}$C)
Longitude	replace Easting or EASTING_NUM
Latitude	replace Northing or NORTHING_NUM

Table 1: Variable Dictionary

Apart from ’DF1.csv’ , I also use several data sets in the process of exploratory data analysis including the cleaned house price data and data sets used to find relationships. I analyze the lead communication pipe and home water supply lead pipe separately because DWQR are responsible for lead level in communication pipe and homeowners are usually responsible for the water supply pipe up to the property boundary.

4 Exploratory Data Analysis

4.1 summary data and variables

1. Lead
In the sample data set, the average lead level in drinking water in Scotland is about 4.87 ugPb/l, the minimum value is 0.2 ugPb/l and the maximum value is 474.7 ugPb/l which is much greater than the average value and also far from strict regulatory standards, that is, the maximum limit (PCV – Prescribed Concentration or Value) is 10 microgrammes per litre(Scotish Water Factsheet 7, 2013).

Region	Average	Max	Min	Count
Ayr	3.89	21.50	0.60	1160
Clyde	1.50	59.50	0.20	7572
Don	5.85	124.10	0.70	13134
Forth	10.41	122.80	0.20	5313
Ness	6.37	339.80	0.20	15712
Nith	5.23	376.50	0.20	46454
Tay	4.28	112.3	0.20	63680
Tweed	3.89	474.70	0.20	32927

Table 2: Lead Level Statistics

In Table 2, I find that there are cases where lead pollution is seriously exceeded in Tweed, Nith and Ness. The highest lead level points in these regions are 474.70 ugPb/l, 376.50 ugPb/l and 339 ugPb/l separately.

Besides, Forth have the largest lead average level(10.41 ugPb/l) which above the maximum limit(10 ugPb/l). Although the sample size in this area is relatively small (5313), DWQR still needs to conduct further investigation and data collection in this area in order to reduce the level of lead pollution in Forth.

In contrast, Clyde and Ayr have good drinking water qualities. For Clyde has lowest average lead level, 1.50 ugPb/l and smaller maximum lead level points 59.50. Ayr has lowest maximum lead level and its average lead level is just 3.89 ugPb/l. The situation of lead pollution in Tay, Don are roughly similar, further exploration and analysis still need to be conduct.

In all, most water samples tested contain below the legal limit of lead. But the data shows that lead pollution has a certain regional. From lead level statistics, DWQR should collect more information in Forth, Ness and Nith.

2. Pipe
a. Pipe materials
In Figure , we can see that the materials used of communication pipe changes with time.

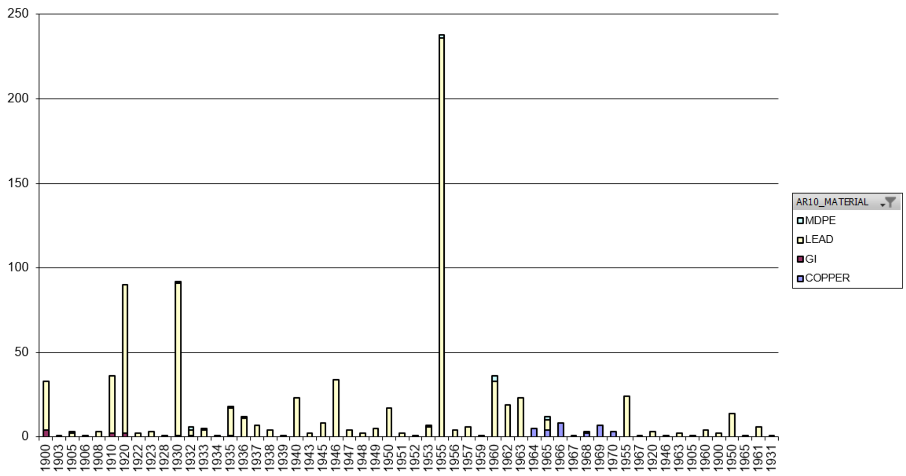

Figure 1: Type of communication pipe

From Figure 1, we can see that from 1900 to 1970, the materials used in the pipeline changed continuously as the year increased. GI materials have been used in water supply systems in Scotland from 1900 to 1920, but from about 1937, GI materials almost disappeared in water supply pipelines in Scotland. Besides, lead was once widely used in water supply systems, especially in 1910, 1920, and 1930. Then, in 1955, the use of lead in the Scottish water supply system reached its peak and then it fell sharply. In 1970, lead materials were no longer used which was replaced by copper. While MDPE and COPPER are both new materials and have been used since about 1955 and 1964 respectively.

b. Pipe leakage
Statistical analysis shows that in the pipes installed before 1970, evidence of lead leakage can be found in 1692 pipes and there is no evidence of lead leakage in 161184 pipes. While only 293 of pipes installed after 1970 can be found evidence of lead leakage and 22783 observation points does not reveal evidence.
This shows that the risk of lead leakage in water pipes installed before 1970 is greater.
Apart from evidence, I select all observation points from 'DF1.csv' that lead level is above 10 ugPb/l.

Region	Average Lead	Average Phosphorus	Average Hydrogen ion	count	percentage
Ayr	21.50	360.00	8.40	57	0.49%
Clyde	44.30	442.00	8.00	9	0.08%
Don	22.02	1611.62	7.73	1150	9.82%
Forth	60.04	664.73	7.90	605	5.17%
Ness	29.95	1773.53	7.64	1886	16.10%
Nith	46.31	401.32	8.04	2171	18.53%
Tay	17.70	1057.68	7.80	5271	45.00%
Tweed	84.78	748.79	7.62	564	4.82%

Table 3: Lead leakage Location

From Table 3, all 11713 observation points meet lead level that above standard 10 ugpb/l. While the largest average lead leakage area in Scotland is Tweed with 84.78 ugpb/l accounts for 4.82 percent. Forth also has larger average lead level at leakage points which is 60.04ugpb/l. Apart from that, we should notice that the largest percentage for having lead exceed limit situations is Tay. Nith and Ness are also accounted for a large proportion.

c. Pipe or Property age

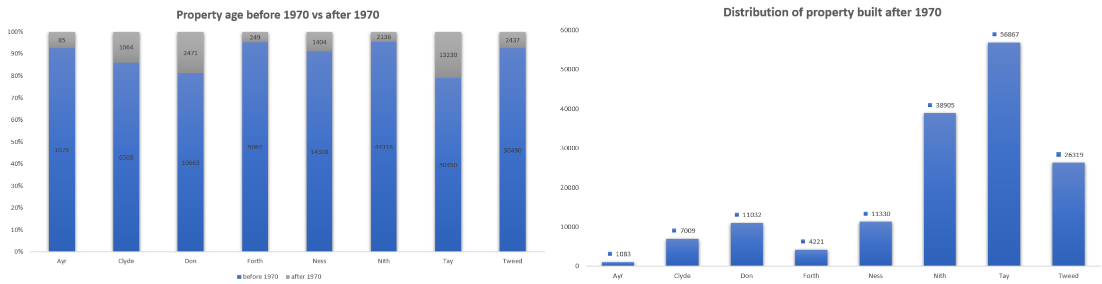

Figure 2: Property Age in Regions

Figure 2 reflects the situation of house ages in 8 regions of Scotland. Obviously, old house built before 1970 account for large proportion in Tay. Besides, Tweed, Don and Nith have large numbers of old house too. In contrast, there are only 85 old houses built before 1970 in Ayr in sample data set and house in Forth are new.
In the right part of Figure 2, we can find that Tay has the largest number of new houses since 1970 which is 56867. Nith ranks second with 38905. Ayr only has the smallest amount of new houses built after Ayr in sample data set.

d. Lead Pipe Distributions

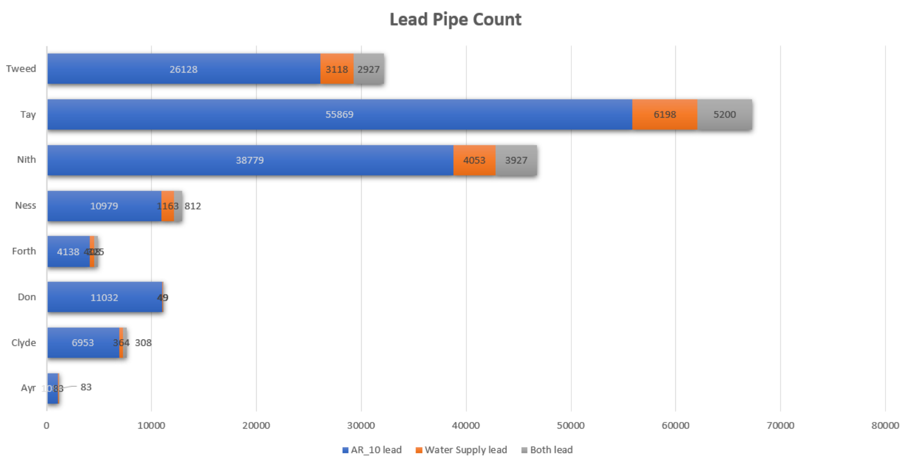

Figure 3: Type of communication pipe

In Figure 3, I find that Tay has the largest number of lead pipes, but the house construction time in this area is relatively short which requires further analysis to determine whether there is risk of increased lead level.
The number of lead pipes in Nith ranks second. Meanwhile, the age of the houses in Nith is long. So lead in the water quality samples exceeds the limit level. There is a situation of lead pollution which means that there is an urgent need to replace lead pipes and give water treatment in Nith. Besides, There are also many lead pipes in Tweed and further analysis of samples is required to determine the lead contamination.
Forth and Ayr have less lead pipes and a relatively higher proportion of houses built after 1970 so I consider the risk of lead pollution is lower.

3. House Average Price
   In the cleaned sample data set of House average price, the average price of houses in Scotland is about 118715.6246 pound per square meter. According to the street postcode positioning, in the sample, the average house price in Nith and Ayrare both lower than the average level. The lowest average house price is in the Nith which is about 98159.86 pound per square meter. At the same time, the highest average house price is located in Tay, with an average house price of approximately 17,1800 pound per square meter, about twice that of Nith area.

   House price in the Nith area are much lower than the overall average house price in Scotland, and the lead content in water is also higher than ones in other areas, indicating that the houses in the area are older. Therefore, there is reason to doubt that lead pipes still remains in Nith's water supply system. In addition, the statistical results can also indicate that there may be an inverse correlation between lead pollution levels and housing prices in different regions.

Postcode	Average Price	District
PH1,PH2, PH7	17862.02	Tay
FK7	164029.63	Tay
PH3, PH5	156093.36	Ness
PH5	140324.71	Ness
TD1, TD5	147682.77	Tweed
KY5,KY7,KY8	17862.02	Tay
KA1	108698.47	Ayr
KA2	107613.25	Ayr
KA3	103869.85	Ayr
KA4	103245.955	Ayr
ML5, ML6	98159.86	NitH

Table 4: House Average Price

4.2 Exploratory Data Analysis

1. Correlation between all variables

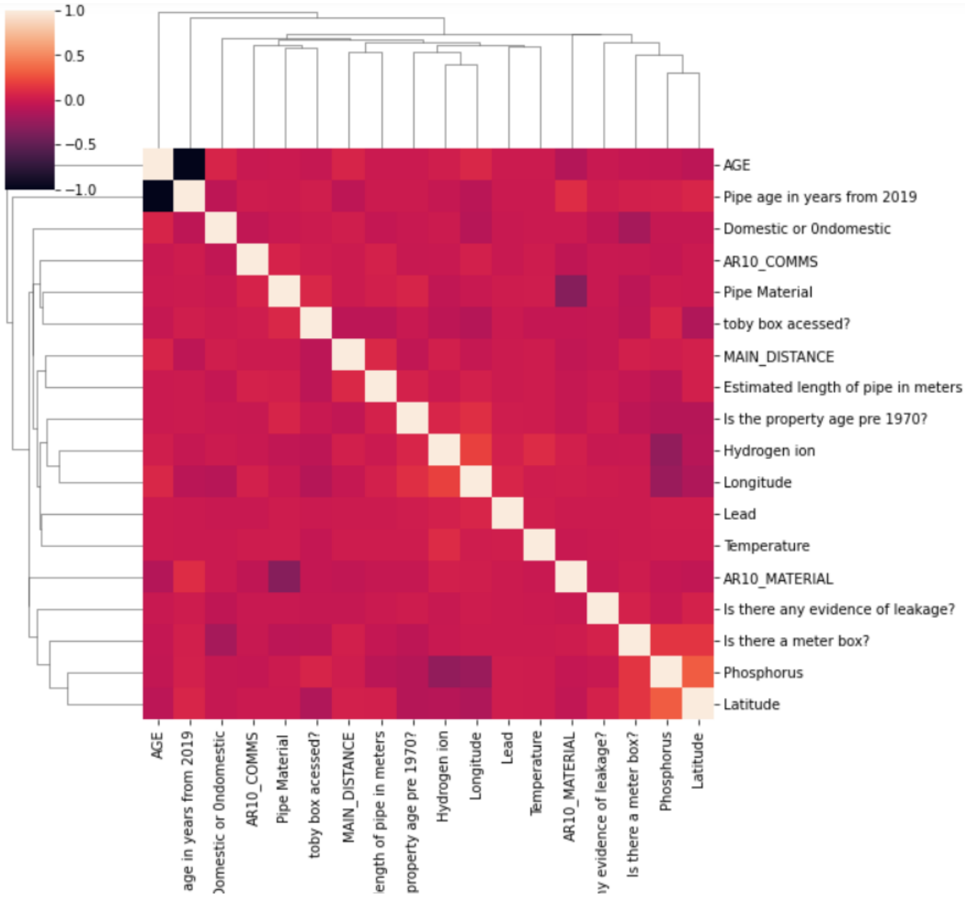

Figure 4: House Price Distribution

I draw a map including all variables in 'DF3_PCA.csv'. In figure 4, the lighter the color of the small square is, the stronger the correlation between variables. I find that there might be positive relationship between Longitude and Hydrogen ion, Latitude and phosphorus, age and leakage and so on.

Specifically, if longitude becomes larger, the Hydrogen ion also increases.So the eastern regions of Scotland contains higher level of Hydrogen ion than western regions. When latitude becomes larger, phosphorus gets larger, which suggests the northern area of Scotland contains more phosphorus.

2. Correlation of Lead, Phosphorus, Hydrogen ion

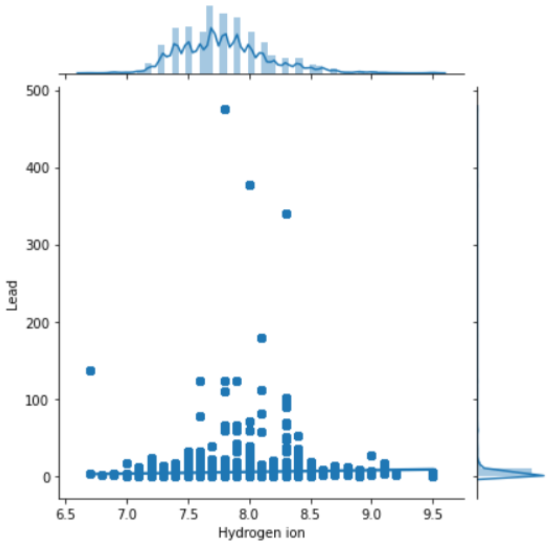

Figure 5: Correlations of Hydrogen ion and Phosphorus

In Figure 5, when the Hydrogen ion concentration is between 7.58.5, lead level reaches its peaks and is relatively dense. When the Phosphorus concentration is between 500 and 800,lead content is prone to peak. Besides, when the PH concentration is around 1400, the lead level in water is also high.

3. Correlation of Length and Lead

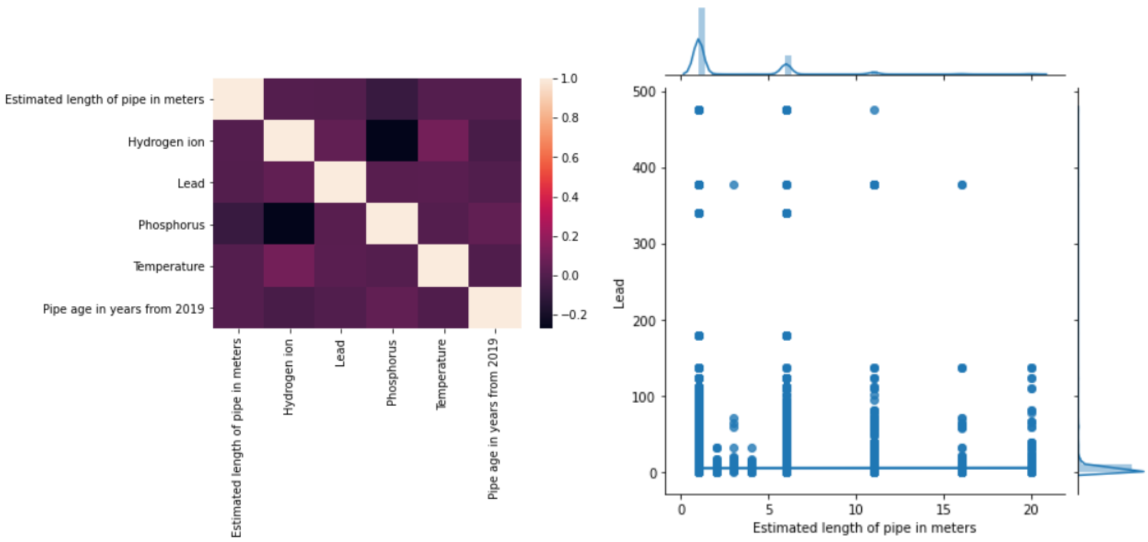

Figure 6: Correlations

From Figure 6, there is no obvious correlation between the length of the lead pipe and the lead content in water. But length of pipe has a very weak negative correlation with the Phosphorus content in the water, the correlation coefficient is -0.2. Besides, from figure, we can also find that as the age of lead pipes increases, Phosphorus could decrease slightly.

4.3 Heatmap Visualization

1. Lead Level Heat Map

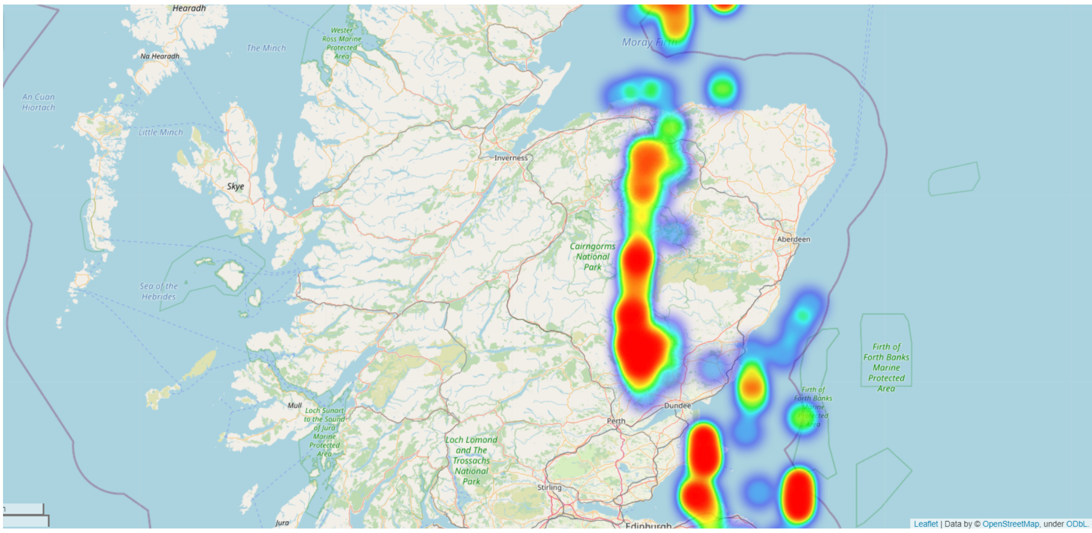

Figure 7: House Price Distribution

To better understanding the lead pollution distribution across Scotland, I highlight in visualisations areas of higher risk. In Figure 7, I use Latitude and Longitude to draw a heat map. The map shows the level of lead in the observation points in water supply pipeline systems in various regions. Besides, it also presents the distribution of sample observation points.

Figure 8: Correlations

Specifically, through heat map, we can roughly find lead pipe's location at street level. Here We can see that the red part suggests that the main pipe along the street may be lead. According to the map, the lead pollution in Ness, Tay, Nith are serious including Tain, Kinlochleven, lanlark, Biggar and so on.

2. House Price Heat Map

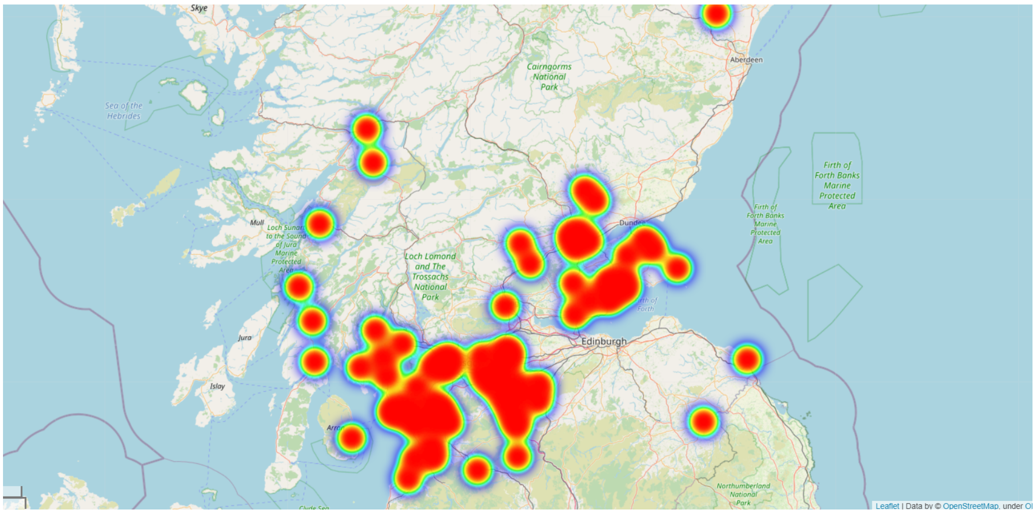

Figure 9: House Price Distribution

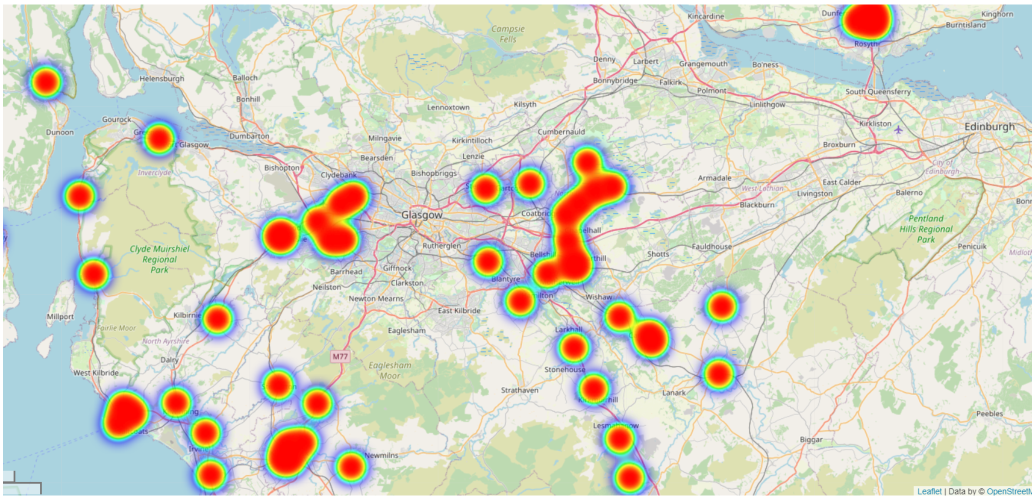

Figure 10: House Price Distribution

I also draw a heat map of average house price in different parts of Scotland. I find that houses in southern area are more expensive than northern area. Besides, from the analysis about house age, I know that these areas also include more houses that are built after 1970. As the older the house are, the greater possibility of lead pipes they have. So I think there is a negative correlation between house price and lead level in water.

The lead pollution in the Nith area is serious and the houses in Nith are older. Therefore, we can find that the house price in Nith region is low. This is also consistent with the conclusion drawn in the previous section on the regional analysis of housing price in Scotland. Apart from that, comparing Figure 7 and 10, we can see that there is almost no lead pollution in the house in the Forth area and the house prices are higher.

5 Conclusions and Recommendations

I prepare the single clean data set named 'DF1.csv' and list the descriptions of all variables in data dictionary.

I find that lead level in drinking water are regional. Tay area has the largest number of lead pipes, but houses in this area have a relatively short construction time and house prices are low, so there is a risk of increasing lead level. The number of houses containing lead pipes in Nith area is second, and the age of the houses in this area is longer. Forth and Ayr have less lead pipes and the proportion of houses built after 1970 is relatively high. So the price of house is relatively high and the risk of exceeding lead is low.

Through analysis of points that lead in the samples exceeds the standard. It is urgent to replace the lead pipes and water treatment especially for Nith and Ness regions. They should also be the key area for DWQR. Besides, there are also many lead pipes in Tweed, further analysis of the information in their water quality samples is required to determine their lead pollution.

In summary, according to summary statistics, correlation analysis and data visualisation, I find that there might be a certain relationship between phosphorus, Hydrogen ion, the house age, structure type, price of the house,the material and length of the communication pipe and lead levels.

It is recommended that we need large sample size to analyze Nith, Ness and Tweed. Because houses in these areas are old. Meanwhile, a regression analysis model such as logistic model could be established on these variables to predict and supervise the lead level in drinking water through Scotland.

References

Boskabady M, Marefati N, Farkhondeh T, Shakeri F,Farshbaf A and Boskabady MH (2018).The effect of environmental lead exposure on human health and the contribution of inflammatory mechanisms, a review. Environment International 120 404_420.

Chen,S. Miller,T., Golemboski,K.,Dietert, R.,1997.Suppression of macrophage metabolite production by lead glutamate in vitro is reversed by meso2, 3dimercaptosuccinic. Acid(DMSA). Toxicol. In Vitro 10,351-358.

Eckerson W.W.Data quality and the bottom lineachieving business success through a commitment to high quality data[R].The Data Ware housing Institute,2002.

J.Archbold and K.Bassil. 2014. Health Impacts of Lead in Drinking Water. Technical Report.

Järup, L(2003). Hazards of heavy metal contamination. British Medical Bulletin Volume 68, 2003, Pages 167-182.

Juhola M, Laurikkala J. Missing values: how many can they be to preserve classification reliability [J/ O L]. Artificial Intelligence Review, 2011.(20110801).

Liu GuoFeng(2006). The screening and studying of Pb contaminated characteristic and Pb hyperaccumulators in heavy metal mines. Fujian Agriculture and Forestry University. FuZhou, Fujian, P.R.C.350002.

Monni S M, Salemaa. Copper resistance of calluna vulgaris originating from the pollution gradient of a Cu-Ni smejter in southwest Finland[j]. Environmental pollution, 2000, 109: 211-219.

Pyle D. Data preparation for data mining [M].San Francisco: Morgan Kaufmann,1999.

Rocha A and Trujillo KA (2019) Neurotoxicity of low-level lead exposure: History, mechanisms of action, and behavioral effects in humans and preclinical models. NeuroToxicology 73 5880.

Swaran J.S. Flora, Govinder Flora and Geetu Saxena (2006). Environmental occurrence, health effects and management of lead poisoning. Lead. pp. 158-228.

Appendices

A Methods in Exploratory Data Analysis

I mainly use Correlation analysis to find the relationship between variables. However, I think it is better to conduct Principal Component Analysis, so I make the data set named ' DF3_PCA.csv'.

I change many variables in this data set into category variables however losing their meanings. And many researchers think it is improper to conduct PCA on category variables. Besides, I think all these information I get through correlation analysis and data visualisation method is enough to identify the houses at increased risk of having lead piping and find areas in Scotland with relatively large numbers of houses with internal lead piping.

Data visualisation method is very attractive to me. They show information in a more clear precise and in intuitive way which helps me a lot.

B Code Clean

```
import pandas as pd
import os
import seaborn as sns
import matplotlib.pyplot as plt
#the path to all raw datasets in one file, 'data' such as 'C:/Users/Rainie Liu/
                                                Desktop/DATA'
path = 'data'
#define basic dataframe cleaning function, delete variables contains missing
                                                values above 50%
def clean_one_df(df):
    thre = 0.5
# print(df.info())
    for col in df.columns:
        if df[col].count() < thre * len(df):
            df.drop(col, axis=1, inplace=True)
    df.dropna(inplace=True)
# print(df.info())
    return df
# clean lead data and draw boxplots of lead and Water Zones
# clean variables in 'SW - All Lead WQ Samples (2010-18).xls'
df1 = pd.read_excel(os.path.join(path, 'SW - All Lead WQ Samples (2010-18).xls')
                                                , encoding='utf-8')[
    ['Result Numeric Entry', 'Sample Date', 'Eastings', 'Northings', 'Street
                                                    Postcode', 'DMA Id', 'RSZ Id', 'WOA
                                                    Id',
     'WSZ Id', 'RSZ Water System Id']]
df1 = clean_one_df(df1)
df1.columns = ['Result Numeric Entry', 'Date', 'Easting', 'Northing', 'Street
                                                Postcode', 'DMA Id', 'RSZ Id', 'WOA Id',
     'WSZ Id', 'RSZ Water System Id']

for col in ['DMA Id', 'RSZ Id', 'WOA Id', 'WSZ Id', 'RSZ Water System Id']:
    x = df1[col]
    sns.boxplot(y=df1['Result Numeric Entry'], x=x)
    plt.show()

for col in ['Easting', 'Northing']:
    sns.lineplot(x=col, y='Result Numeric Entry',data=df1)
    plt.show()
print(df1.info())
# transform Easting and Northing variables to longitude and latitude to draw
                                                heatmap
```

```
df1['Longitude'] = df1['Easting'] / 100000 - 6.54
df1['Latitude'] = df1['Northing'] / 10000 - 14
df1.to_csv('cleaned_lead_wq.csv', index=0, encoding='utf-8-sig')

# clean pipe data
df2 = pd.read_excel(os.path.join(path, 'SW - Comm pipe data.xls'), encoding='utf
                                        -8')
df2_2 = pd.read_excel(os.path.join(path, 'SW - Scottish Water Zonal Phosphate
                                        Levels.xls'), encoding='utf-8',
                      sheet_name='Sheet1')[['Hydrogen ion', 'Lead', 'Phosphorus'
                                                            , 'Temperature', '
                                                            Rig']]
df2_2 = clean_one_df(df2_2)
df2_3 = pd.read_excel(os.path.join(path, 'SW - Phosphate Dosing WTWs Y or N.xlsx
                                        '), encoding='utf-8')
df2_3 = clean_one_df(df2_3)
df2_3.columns = ['Rig', 'WTW1', 'temp']
df2_3.drop('temp', axis=1, inplace=True)
# connect datasets through Rig variable, split 'Zone'
df2_2['Rig'] = df2_2['Rig'].apply(lambda x: x.split('Zone')[0].upper() + 'WTW ')
print(df2_2.head())
print(df2_3.head())
# print(list(df2_2['Rig']))
# print(list(df2_3['Rig']))
# connect and merge two datasets through 'Rig' into df2_4 dataset
df2_4 = pd.merge(df2_2, df2_3, on=['Rig'], how='left')
print(df2_4.head())
print(df2.columns)
print(df2_2.columns)
l_index = []
#keep lead material
for ind in df2.index:
    if pd.isna(df2['Specify If Other '].loc[ind]):
        l_index.append(ind)
df2 = df2.loc[l_index]

df2 = clean_one_df(df2)
print(df2.info())
#create a function to deal 'm', unite the representation of length
def func(x):
    if 'm' in str(x):
        return int(x[:-1])
    elif '-' in str(x):
        return int(x.split('-')[0])
    elif '>' in str(x):
        return int(x.split('>')[1])
    else:
        return int(x)
# use function deal unit
df2['Estimated length of pipe in meters'] = df2['Estimated length of pipe in
                                        meters'].apply(lambda x: func(x))
print(df2.info())
print(df2_4.info())
# merge datasets through 'WTW1'
df2 = pd.merge(df2, df2_4, on=['WTW1'], how='left')
df2 = clean_one_df(df2)
print(df2.info())

# find out Easting refers to Longitude; Nothing refers to Latitude
# Longitude should contain negative values so I check several points to compute
                                        the relationship
df2['Longitude'] = df1['Easting'] / 100000 - 6.54
df2['Latitude'] = df1['Northing'] / 10000 - 14
df2.to_csv('cleaned_pipe.csv', index=0, encoding='utf-8-sig')
for col in ['AR10_MATERIAL', 'AR10_COMMS', 'POST_TOWN', 'Street postcode', '
                                        DISTRICT POSTCODE', 'RZ_REF', 'RZ_NAME',
```

```
            'REGION', 'CONFIDENCE_GRADE', 'Pipe Material ', 'Nearside or Farside
                                                        ?']:
    sns.countplot(x=col, data=df2)
    plt.show()
# deal house data in 'Other-Postcode_household count_urban class.csv'
# the sample contain too many missing values
df3 = pd.read_csv(os.path.join(path, 'Other - Postcode_ household count_ urban
                                         class.csv'), encoding='utf-8')[
    ['Street postcode', 'Latitudeitude', 'Longitudegitude', 'DateOfIntroduction'
                                            , 'CouncilArea2018Code']]
df3 = clean_one_df(df3)
# clean Date data
df3['Date'] = df3['DateOfIntroduction'].apply(lambda x: pd.to_datetime(x.split()
                                         [0]))
print(df3.head())
# delete Date of Introduction
df3.drop('DateOfIntroduction', axis=1, inplace=True)
# clean data in 'Other-UK-HPI-full-file-2019-03.csv'
df4 = pd.read_csv(os.path.join(path, 'Other - UK-HPI-full-file-2019-03.csv'),
                                         encoding='utf-8')[['CouncilArea2018Code'
                                         ,
 'AveragePrice', 'Date']]
df4 = clean_one_df(df4)
# transform Date data
df4['Date'] = pd.to_datetime(df4['Date'])

print(df4.head())
# merge house data by time and region code
df5 = pd.merge(df3, df4, on=['CouncilArea2018Code', 'Date'], how='inner')
print(df5.head())
# save cleaned house dataset
df5.to_csv('cleaned_house_data.csv', index=0, encoding='utf-8-sig')
```

C Code Explore

```
# -*- coding: utf-8 -*-
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import mode
get_ipython().magic('matplotlib inline')
import datetime
from pandas import Series
from datetime import datetime
from numpy import mean, median

df1 = pd.read_csv('C:/Users/Rainie Liu/Desktop/DF2.csv')
df2 = pd.read_csv('C:/Users/Rainie Liu/Desktop/DF3_PCA.csv')
df3 = pd.read_csv('C:/Users/Rainie Liu/Desktop/DF4_level.csv')
df4 = pd.read_csv('C:/Users/Rainie Liu/Desktop/DF5_AR10_lead.csv')
df5 = pd.read_csv('C:/Users/Rainie Liu/Desktop/DF6_supplylead.csv')

# 156766 total numbers of observations in 'Lead' dataset
num = 156766
# Maintain data type consistency
Length = np.array(df['Estimated length of pipe in meters'][0:num],dtype = float)
Lead = np.array(df['Lead'][0:num],dtype = float)
Pho = np.array(df['Phosphorus'][0:num],dtype = float)
Temp = np.array(df['Temperature'][0:num],dtype = float)
Hyd = np.array(df['Hydrogen ion'][0:num],dtype = float)
#Lead
#Kernel density estimate which shows the statistical distribution of lead
                                         content levels
```

```
sns.kdeplot(df1['Lead'], shade=True)
# draw lead dist plot
sns.distplot(df1['Lead'])

#Lead pipe length and Lead level
# if communication AR10_pipe material is lead, view correlation
sns.jointplot(x = 'Estimated length of pipe in meters', y = 'Lead', data = df2 ,
                                        kind = 'reg')
# if water supply pipe material is lead, view correlation
sns.jointplot(x = 'Estimated length of pipe in meters', y = 'Lead', data = df2 ,
                                        kind = 'reg')

# Compute correlation coefficients between all variables
# df2 (DF3) have all Categorical variables
print(df1.corr())
print(df2.corr())
print(df4.corr())
print(df5.corr())

# plots
#Dependency graph is spectacular!
sns.pairplot(df1)
#I want to see all variables relationship
sns.clustermap(df2.corr())
# View the relationship between Lead,Phosphorus,Hydrogen ion,Temperature
sns.pairplot(df3)
# I want to see the relevance heat map between variables
sns.heatmap(df3.corr())
# View relationship
sns.jointplot(x = 'Phosphorus', y = 'Lead', data = df3 ,kind = 'reg')
sns.jointplot(x = 'Hydrogen ion', y = 'Lead', data = df3 ,kind = 'reg')

#Draw Heatmap of Lead
import numpy as np
import pandas as pd
import seaborn as sns
import folium
import webbrowser
from folium.plugins import HeatMap
#'path' on my computer to data' such as 'C:/Users/Rainie Liu/Desktop/DF1.csv'
data = pd.read_csv('path',low_memory=False)

num = 156766
#get latitude, longitude value and Lead concentration and Convert to numpy
                                        floating point
lat = np.array(data["Latitude"][0:num],dtype=float)
lon = np.array(data["Longitude"][0:num],dtype=float)
lead = np.array(data["Lead"][0:num],dtype=float)
# Make the data into the form of [lats,lons,weights]
data1 = [[lat[i],lon[i],lead[i]] for i in range(num)]
#Draw a Map first
# set latitude and longitude of central point of Scotland as the center of the
                                        original picture
# set the zoom level to 10 times
map_osm = folium.Map(location=[70.5500,5.5510],
               control_scale = True, zoom_start=10)
# Add the heat map to the map created earlier
HeatMap(data1).add_to(map_osm)
# save as html file
file_path = r"C:/Users/Rainie Liu/Desktop/DATA.html"
map_osm.save(file_path)
#Use default browser open heatmap
webbrowser.open(file_path)

#Draw Heatmap of House average price
```

```
#'path' on my computer to data' such as 'C:/Users/Rainie Liu/Desktop/
                                        cleaned_house_data.csv'
data = pd.read_csv('path',low_memory=False)
#count in data set
num = 133
#get latitude, longitude value and House average price and Convert to numpy
                                        floating point
lat = np.array(data["Latitude"][0:num],dtype=float)
lon = np.array(data["Longitude"][0:num],dtype=float)
price = np.array(data["AveragePrice"][0:num],dtype=float)
# Make the data into the form of [lats,lons,weights]
data1 = [[lat[i],lon[i],price[i]] for i in range(num)]

#Draw a Map first
# set latitude and longitude of central point of Scotland as the center of the
                                        original picture
# set the zoom level to 10 times
map_osm = folium.Map(location=[70.5500,5.5510],
               control_scale = True, zoom_start=10)
# Add the heat map to the map created earlier
HeatMap(data1).add_to(map_osm)
# save as html file
file_path = r"C:/Users/Rainie Liu/Desktop/DATA.html"
map_osm.save(file_path)
#Use default browser open heatmap
webbrowser.open(file_path)
```