Potholes in Boston

DATA SCIENCE TRACK – CAPSTONE PROJECT 1

Thibault Dody | Exploratory Data Analysis| 09/12/2017

Springboard

https://github.com/tdody/Springboard-Capstone-1

TABLE OF CONTENTS

1 INTRODUCTION

The purpose of this report is to present the preliminary conclusions of the analysis of the pothole issue is Boston.

2 DATA SOURCES

The datasets used for this project are the following:

1. The weather data of the city obtained from the National Centers for Environmental Information (NOAA). The dataset was available upon request on the agency website.
2. The neighborhoods population and size dataset was obtained from ZipAtlas[1].
3. The neighborhoods map file was downloaded from the Boston Open Data website[2].
4. The pothole dataset was obtained from the City of Boston website[3]

The folder containing the project files is located in the GitHub repository Springboard-Capstone-1[4] (It contains the following:

- 00_Data_Wrangling-Weather.ipynb
- 01_Data_Wrangling_Boston.ipynb
- 02a_Data_Wrangling_Potholes.ipynb
- 02b_Google_Geo_API_Fetcher.ipynb
- 03_Data_Story_Telling.ipynb
- Folder Original Data
- Folder Intermediate Data
- Folder Cleaned Data

The following sections summarize the cleaning process used for each of the Jupyter Notebooks.

The methodology used to perform the data cleaning is described in the "Data Wrangling Report" stored as a PDF file in the GitHub repository.

3 SOFTWARE VERSION

For this analysis, the following software are used:

- Jupyter NoteBook v5.0.0
- Python v3.6.1
- Anaconda v4.4.0

[1] http://zipatlas.com/us/ma/zip-code-comparison/population-density.htm
[2] http://bostonopendata-boston.opendata.arcgis.com/datasets/53ea466a189b4f43b3dfb7b38fa7f3b6_1
[3] https://data.cityofboston.gov/City-Services/Requests-for-Pothole-Repair/n65p-xaz7/data
[4] github.com/tdody/Springboard-Capstone-1

4 QUESTIONS TO BE ANSWERED

As part of this preliminary analysis, the following questions were identified as key points of the study:

1. How does the weather impact the number of claims?
2. How does the number of claims differ by neighborhood?

In order to answer these questions, the main contributing features need to be identified. We will use Exploratory Data Analysis (EDA) methods in order to answer the questions.

5 CORRELATION WITH WEATHER DATA

5.1 Claims per season

In order to optimize the allocation of the resources needed to repair the pothole, it is primordial to understand how demand fluctuates over the year. Based on basic knowledge and experience, potholes tend to appear at certain points during the year. In order to refine our analysis, the total number of claims received every two weeks are summed and each month is assigned to a season.

Table 5-1: Number of claims per season

Property	Spring	Summer	Fall	Winter
Mean	**458**	216	108	202
Standard Deviation	**272**	77	35	192
Min	61	66	58	**37**
25th Percentile	**242**	146	84	89
Median	**425**	223	99	139
75th Percentile	**559**	260	125	213
Max	**1220**	409	201	762

From the data presented in Table 5-1, we identify Spring as being the season with the maximum of claims. The other three seasons (Summer, Fall, and Winter) have similar properties. Figure 5-1 depicts the variation of claims measured bi-weekly over the July 2011 to June 2017 period. From this plot, we can make the following observations:

- The number of claim seems to be correlated to the strength of the winter. Indeed, the winters of 2014, 2015, and 2017 were particularly cold and large snowfalls occurred frequently.
- There seems to be a time lag between the cold period and the claim peak.

Based on our primary conclusions, we will now investigate two factors that are specific to winter times. We will try to see if the number of claims is positively correlated with the number of snow days and the number of freezing days.

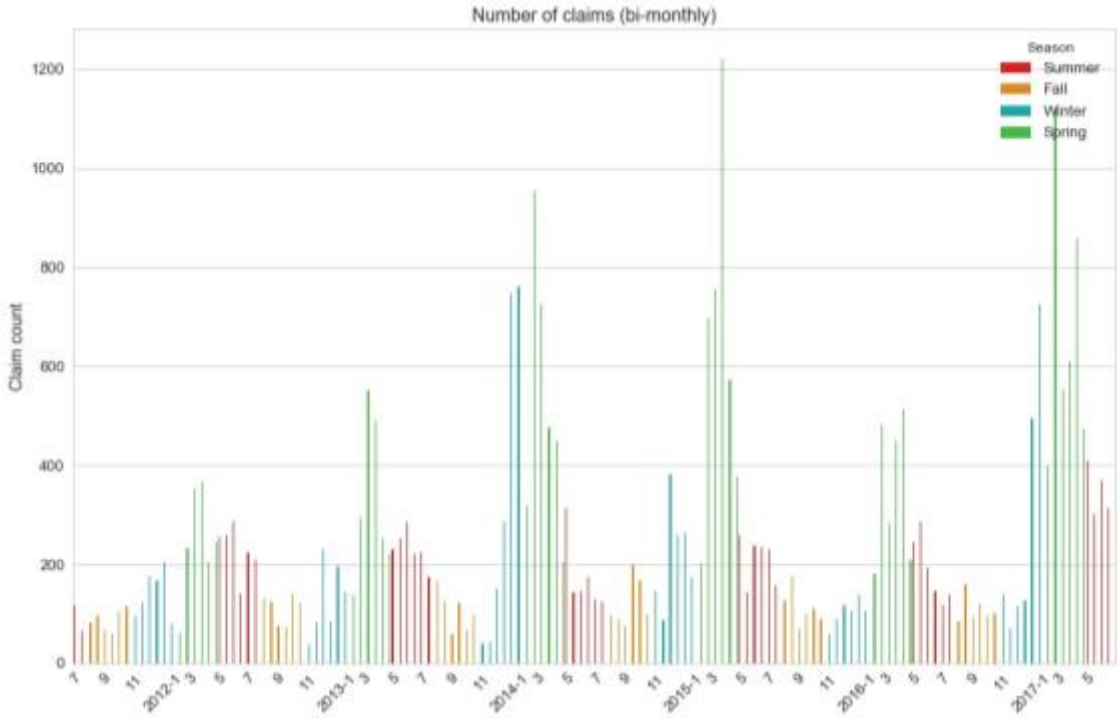

Figure 5-1: Bi-monthly claim count

The average number of snow days and freezing days are depicted in Figure 5-3. As expected, these two features are only non-null during winter time. However, the peak for these two parameters seems to happen in January and February. From this observation, we can now estimate the time-lag to be between one and two months.

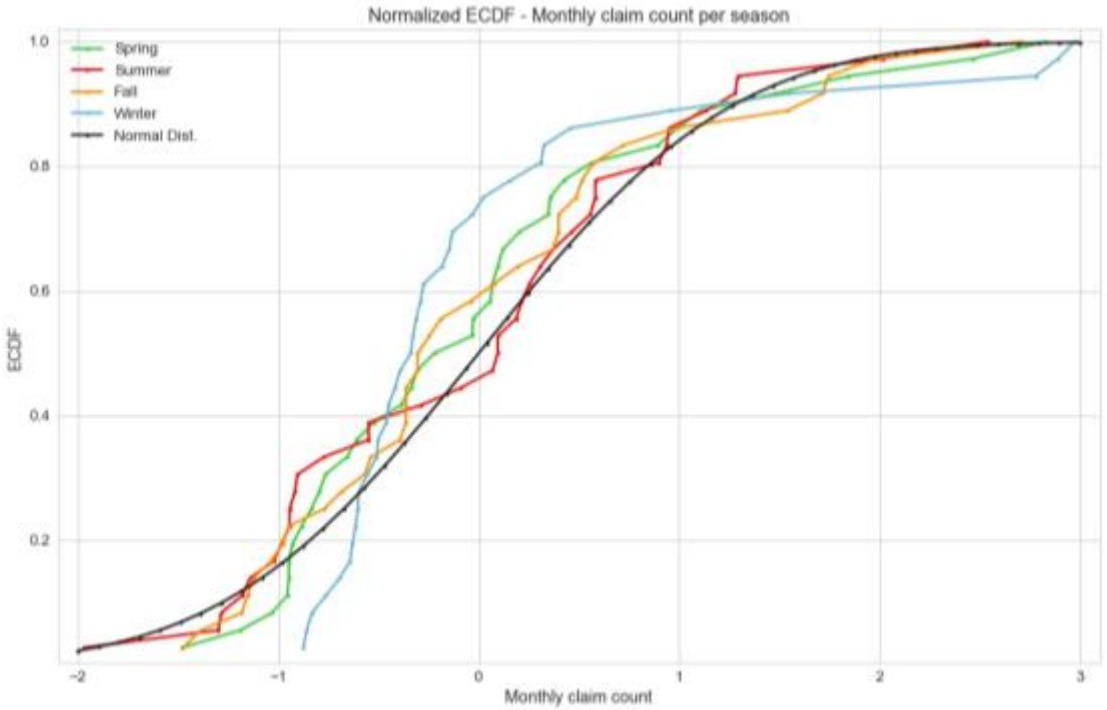

Figure 5-2: Bi-monthly number of claims – Empirical cumulative distribution functions

Figure 5-2 depicts the normalized empirical cumulative distribution functions. While the distribution for the Summer data is relatively close to the normal distribution, the other three seasons are not. We use the scipy library to perform hypothesis testing:

Null hypothesis: Ho -> The claim count is normally distributed.
Alternative hypothesis Ha -> The claim count is not normally distributed.

Table 5-2: Number of claims - Hypothesis testing

Item	Spring	Summer	Fall	Winter
p-value	0.01214	0.63233	0.09237	3.459E-06
Conclusion	Ho true	Reject Ho	Reject Ho	Ho true

Based on the p-values listed above, the spring and winter distributions can be assumed to be normally distributed.

Now that we have a better understanding of the correlation between the winter weather and the number of claim, we will try to compute the time-lag that maximizes the correlation between the two data sets.

Figure 5-4 depicts the correlation matrix between the shifted pothole data and the weather data. In order to produce this plot, the pothole data was shifted incrementally from one month two six months. The Pearson's correlation factors were then computed between the number of monthly potholes and

the weather data. After inspection of the correlation map, the following results were identified as significant:

- Shift 0 with Cooling Degree Days (season-to-date): -0.64
- Shift -1 with DSNW: 0.79
- Shift -1 with Number of days with minimum temperature <= 32 degrees Fahrenheit: 0.74
- Shift -1 with Number of days with maximum temperature <= 32 degrees Fahrenheit: 0.76
- Shift -1 with Extreme minimum temperature for month: -0.70
- Shift -1 with Highest daily snowfall in the month/year: 0.75
- Shift -1 with Extreme maximum temperature for month/year: -0.67
- Shift -1 with Heating Degree Days: 0.74
- Shift -1 with Total Monthly Snowfall: 0.75
- Shift -1 with Average Monthly/Annual Temperature: -0.71
- Shift -1 with Monthly/Annual Maximum Temperature: -0.71
- Shift -1 with Monthly/Annual Minimum Temperature: -0.72

In conclusion, the monthly number of claims is strongly positively correlated with the amount of snow fall and the number of freezing days. It is also strongly negatively correlated with the temperature measurements.

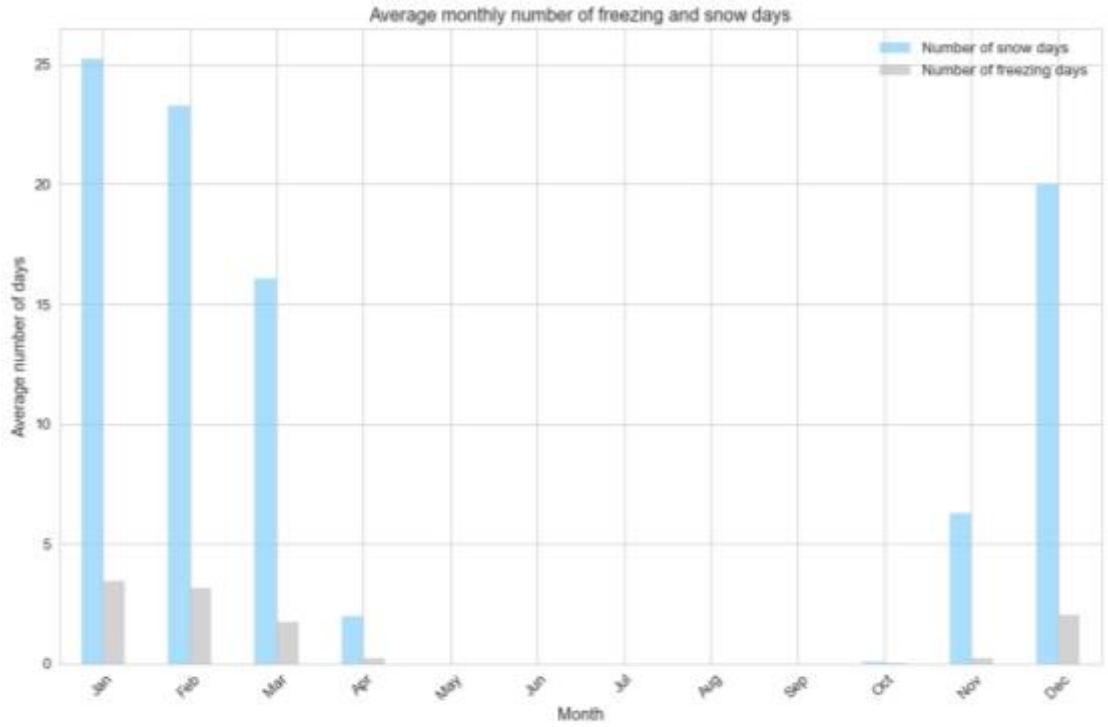

Figure 5-3: Monthly average of freezing and snow days

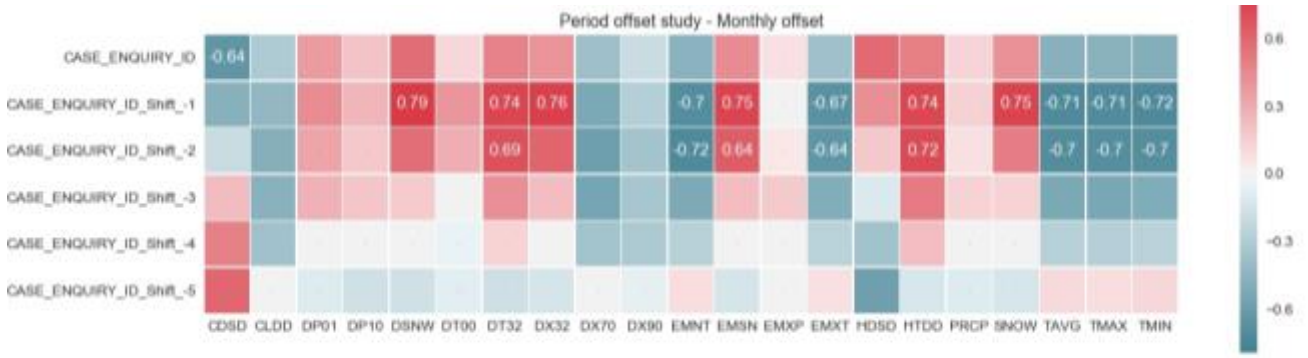

Figure 5-4: Correlation heat map - Number of claims and weather data

5.2 Linear Regressions

Section 8, Figure 8-1 to Figure 8-6 depict the linear regression (and residuals) of the various regressions that were performed in order to assess the relationships between the weather data and the monthly number of claims.

Table 5-3: Regression results

	Snow days	Freezing days	Average temperature	Precipitation
Slope	174.5	29.5	-17.7	45.2
Intercept	334.3	273.3	1424.0	230.0

Table 5-3 lists the results of the linear regression between the monthly number of claims and the weather data. Interestingly, the effect of a freezing day with snow has a contribution almost six times greater than the contribution of a freezing days without snow on the number of claim.

Figure 8-5 depicts the residuals of the regression between the monthly average temperature and the monthly number of claims. It is interesting to notice that the residual seems to decrease (in absolute value) as the monthly average temperature decreases. Our assumption is that the higher variance at lower temperature (near freezing) is probably explained by the amount of snow.

6 CORRELATION BETWEEN THE NUMBER OF CLAIM AND NEIGHBORHOOD DATA

In this section, we will use the demographic data from the different zip-codes of the city. The purpose of this analysis is to identify the main contributors leading to a higher than average number of claims.

The heat map presented below provides valuable insights regarding the correlation between the neighborhood properties and the number of claims. The results are interesting as the population density barely influences the number of claims while the neighborhood area and population are strongly positively correlated with the number of claims.

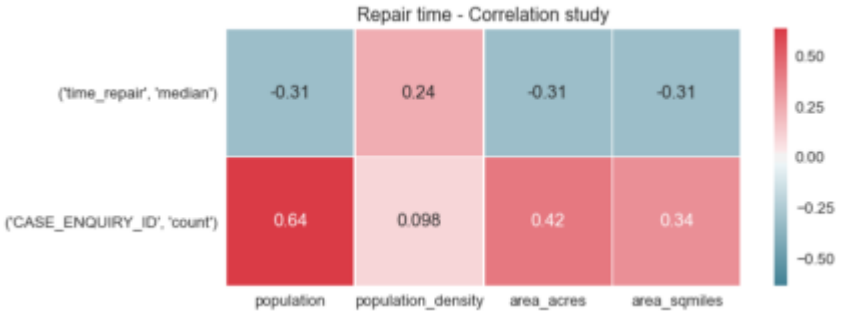

Figure 6-1: Heat map population versus number of claims and repair time

The linear regression plots are presented in Figure 8-7 and Figure 8-8. The results of the linear regression are listed in Table 6-1.

Table 6-1: Regression results

	Population	Area [acres]
Slope	0.0146	0.134
Intercept	269.6	415.1

Now that we have targeted the main contributors of the number of claims. We will focus our study on the discrepancy in the number of claims when considering the neighborhood individually.

Figure 9-1 depicts the discrepancy in the pothole density per neighborhoods. A few neighborhoods have an extremely high pothole density.

Table 6-2: Outliers, pothole density

LOCATION_ZIPCODE	CASE_ENQUIRY_ID	population	population_density	area_acres	Latitude	Longitude	area_sqmiles	pothole_density
02210	318	592.0	757.88	499.92	42.35	-71.04	0.78	53.72
02110	405	1428.0	8630.93	105.89	42.36	-71.05	0.17	28.36
02108	769	3446.0	12377.16	178.19	42.38	-71.06	0.28	22.32
02109	442	3428.0	20752.98	105.72	42.36	-71.05	0.17	12.89
02136	2951	28392.0	6048.39	3004.25	42.26	-71.13	4.69	10.39

Table 6-2 lists the five outliers of the pothole density distribution. Now that we have targeted the outliers, the goal is to understand why they have such a high ration.

First, we will investigate their population density. Indeed, if a neighborhood does not have many people living in but many working in, the roads will be subjected to high traffic while the population count would remain low. The first feature they have in common is their locations, all three are located in the center of the financial district. These neighborhoods are known for their old, narrow streets.

Based on the population density ranking, zip code 02210 and 02110 appear to be located in the bottom half of the table in term of population density. In conclusion, while having fewer people living in these areas, these two neighborhoods are located at the intersection of the South of Boston and the cities of Cambridge and Somerville (both located North of the Charles river). We saw that discrepancies appear when comparing the number of claims per neighborhoods. We will now refine the study of the number of claims distribution by looking at a finer mesh. The size of the mesh is based on the range of latitude and longitude of the pothole claims.

Figure 6-2 shows the distribution of the number of claims for a finer mesh. From this representation, we can identify smaller area with high number of claims. These correspond to West Boston and the historical city center.

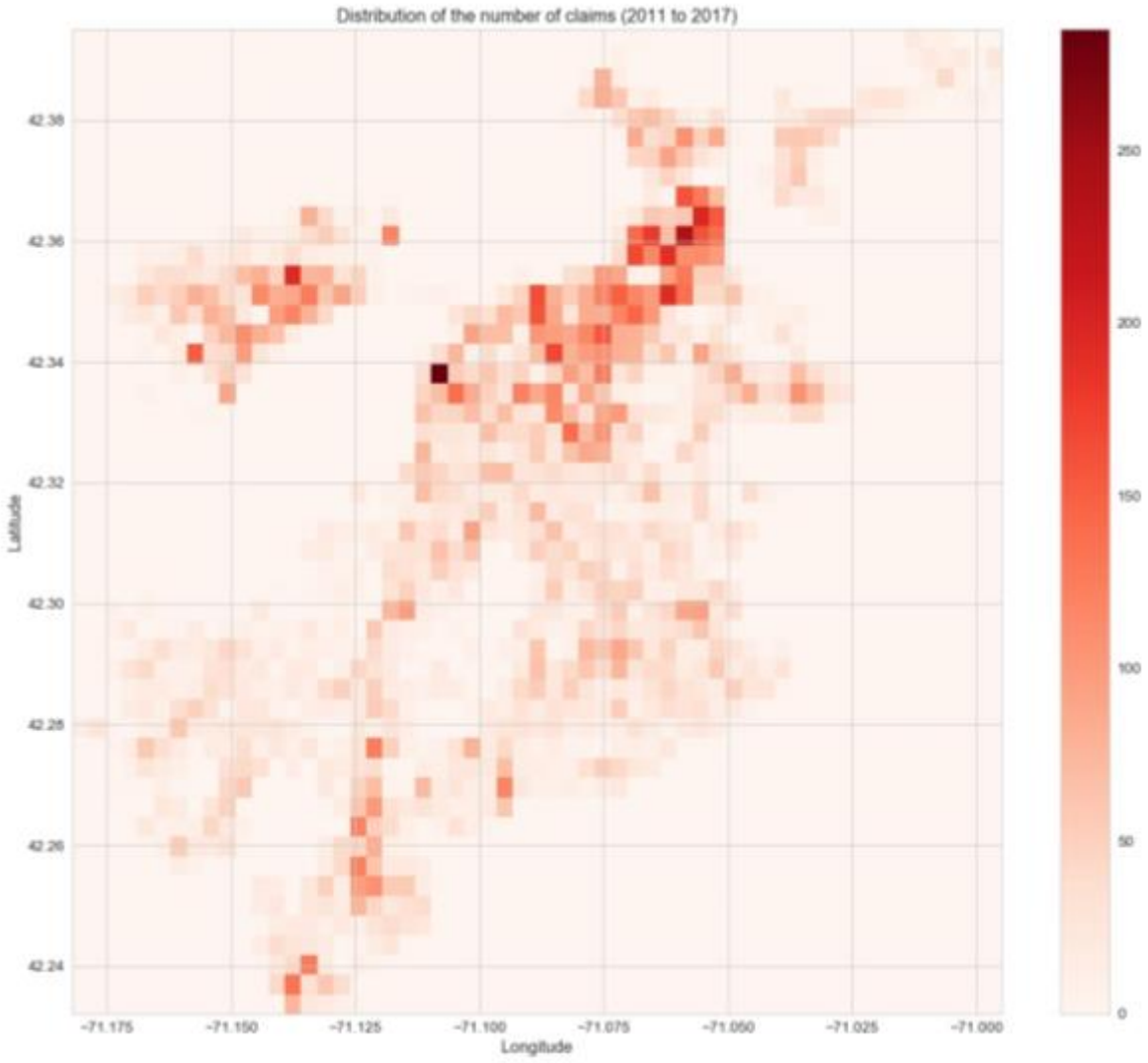

Figure 6-2: Number of claims (2011 to 2017)

7 CONCLUSION

The results of the analysis presented in this report lead to the following conclusions:

1. As expected, the weather has a major impact on the frequency of appearance of potholes. However, we were able to rule out the rain and the "just cold" weather as the number of claims is directly correlated to the number of freezing days and the amount of snow fall. Moreover, our study shows that there is a lag effect of one month between a period of bad weather and a peak in pothole claims.
2. When considering the discrepancy in the number of claims per neighborhood, our analysis shows that most of the neighborhood have similar claim density (number of claims per 100

inhabitants). However, areas with denser traffic and older infrastructures account for a much higher density.

Note: This report only contains the analysis of the number of pothole feature, a similar extensive study of the repair time is performed and documented in the notebook: *03_Data_Story_Telling*.

8 ATTACHMENT A – REGRESSION PLOTS

8.1 Weather Data

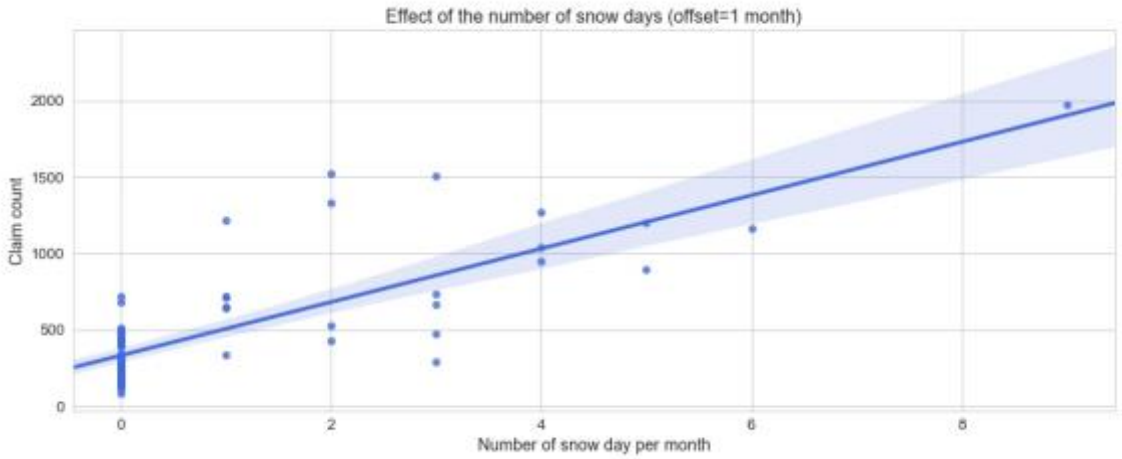

Figure 8-1: Effect of the number of snow days (offset = 1 month)

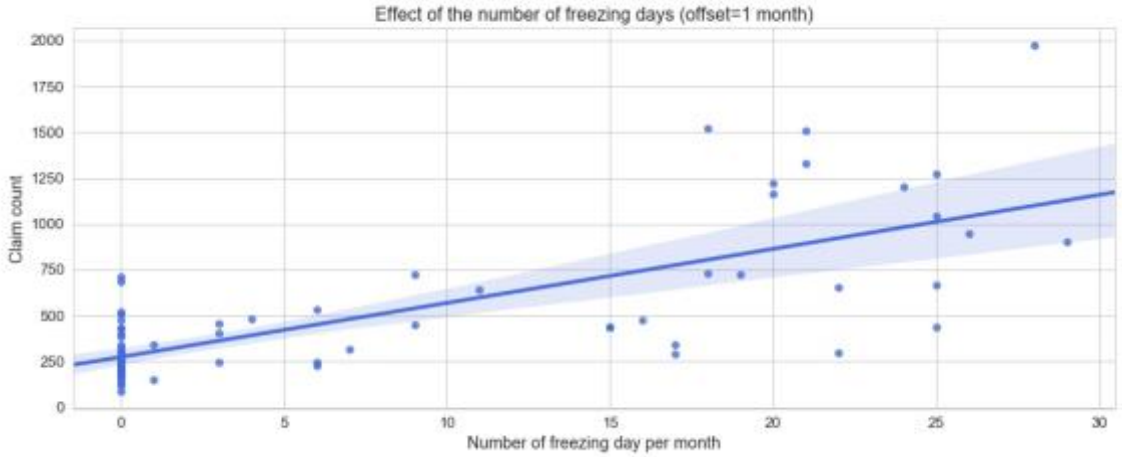

Figure 8-2: Effect of the number of snow days (offset = 1 month)

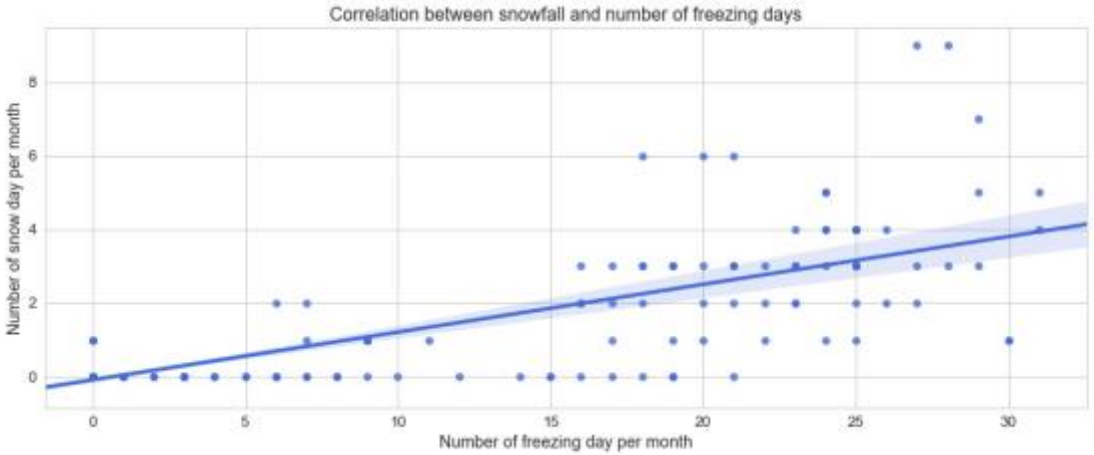

Figure 8-3: Correlation between the snowfall and the number of freezing days

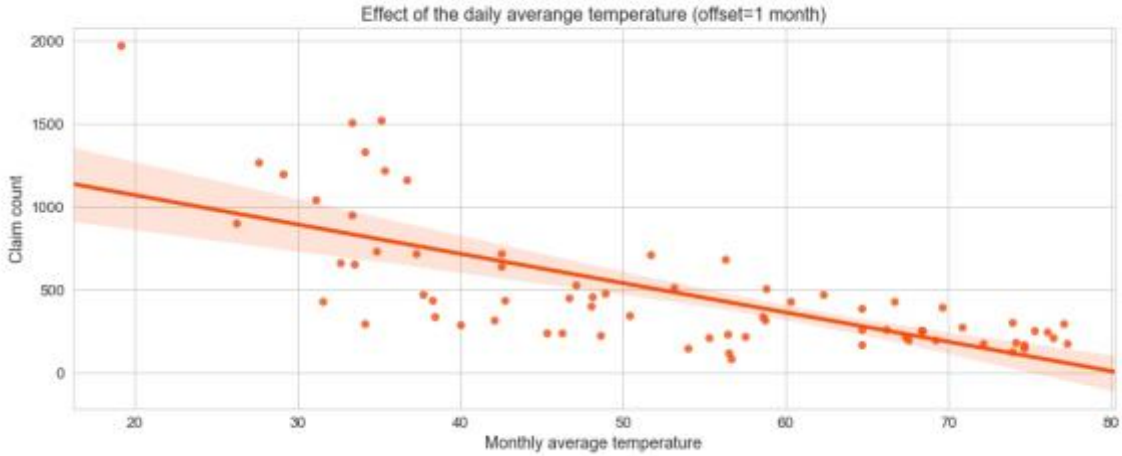

Figure 8-4: Effect of the monthly temperature on the number of claims

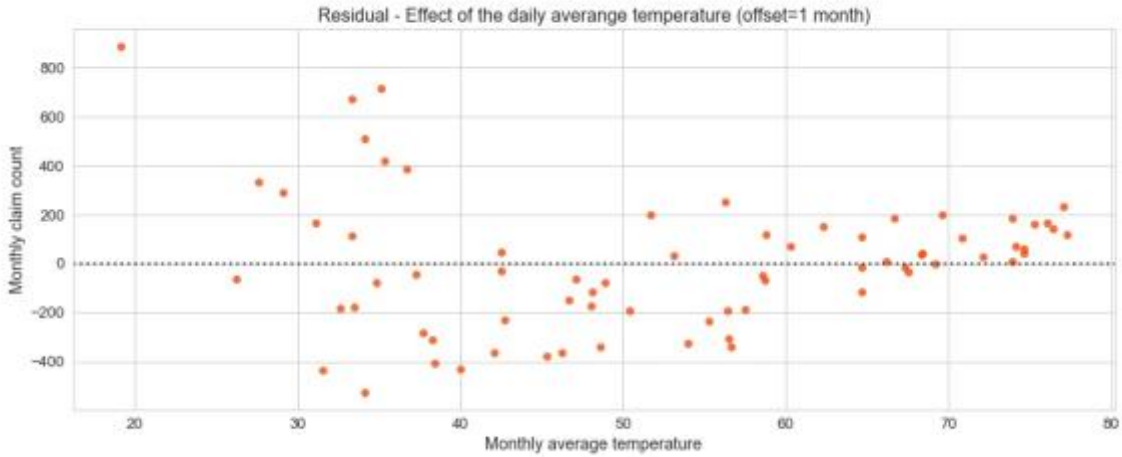

Figure 8-5: Residual of the effects of the monthly temperature on the number of claims

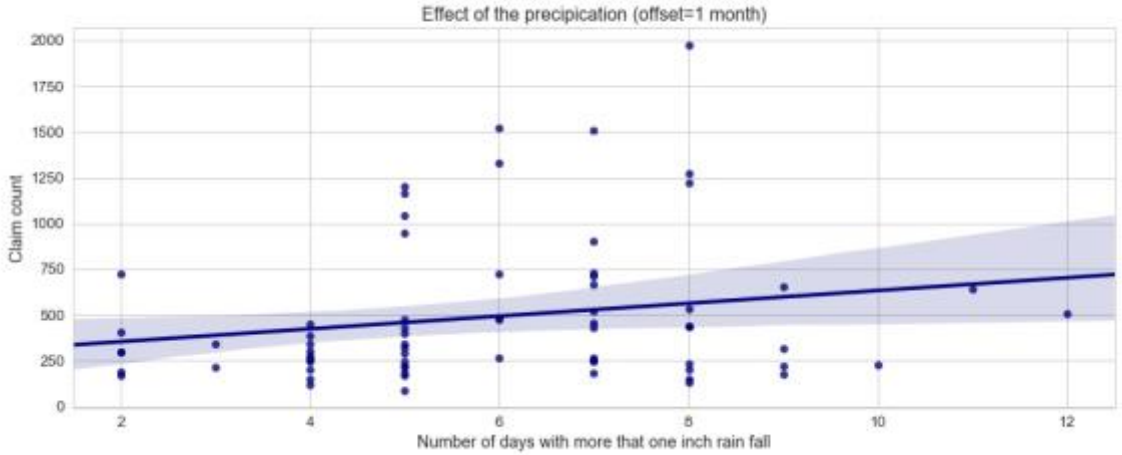

Figure 8-6: Effect of the precipitation on the number of claims

8.2 Population Data

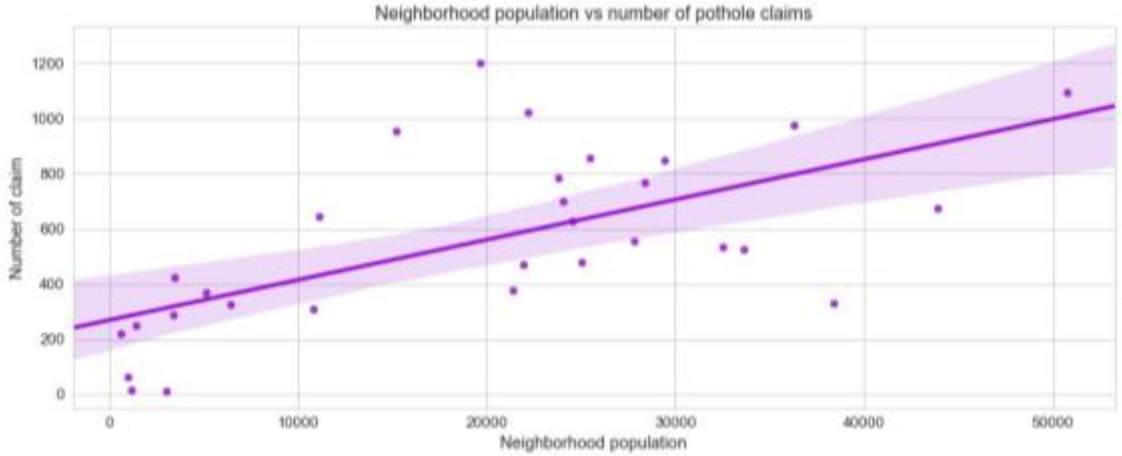

Figure 8-7: Effect of the population on the number of claims

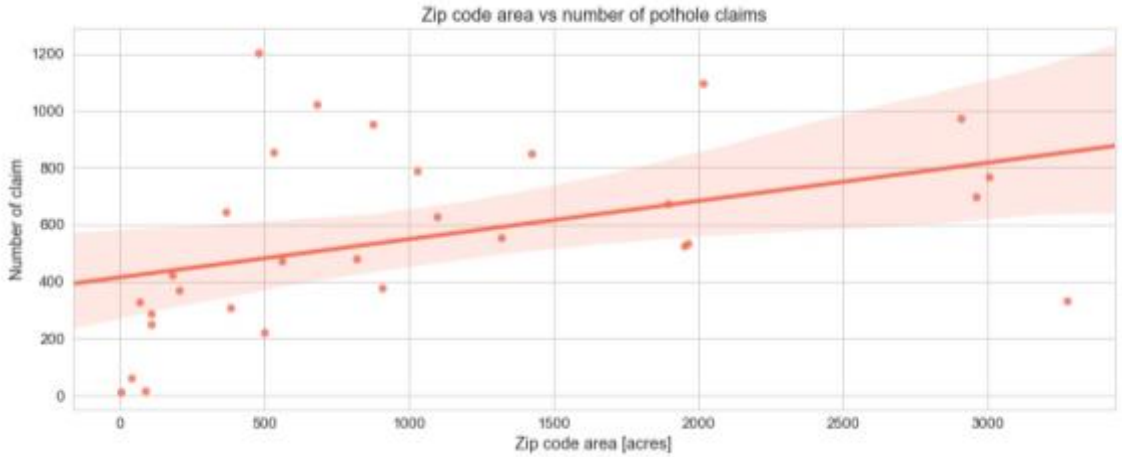

Figure 8-8: Effect of the area on the number of claims

9 ATTACHMENT B – MAPS

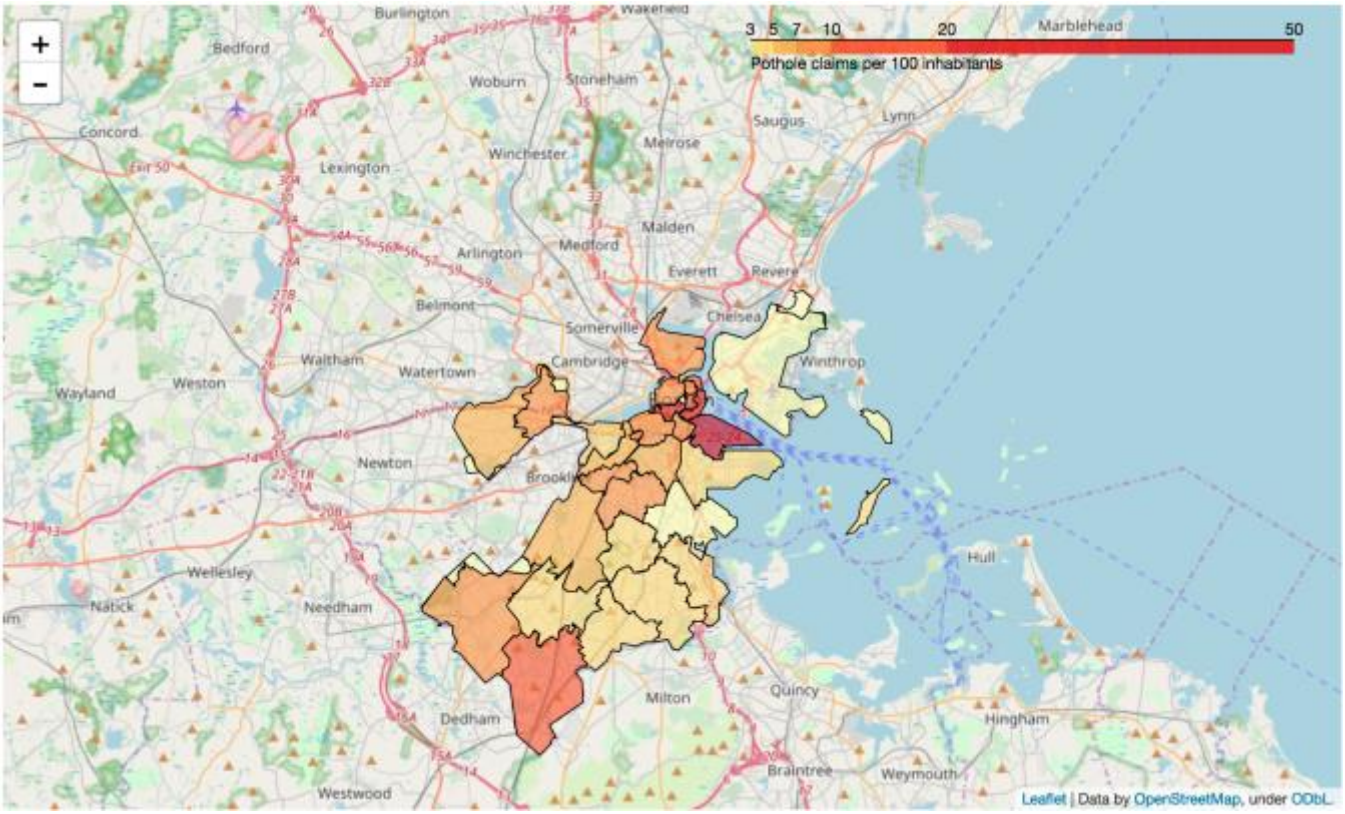

Figure 9-1: Pothole density per 100 inhabitants