Potholes in Boston

DATA SCIENCE TRACK – CAPSTONE PROJECT 1

Thibault Dody | Milestone Report| 09/15/2017

Springboard

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

TABLE OF CONTENTS

1 INTRODUCTION

Potholes and bad road conditions are probably Bostonian's favorite discussion topics during winter times. The purpose of this project is to assess to situation using real data and address whether discrepancies, in the number of claims or time for repair, exist between the different part of the city. Once the pothole is formed, it represents a real risk for road users, especially bikers and riders. The damage suffered by cars due to the pothole problem is estimated to cost $3 billion a year (Source: American Automobile Association). To this amount, the cost of the repair and care need to be added to this issue.

In conclusion, potholes and road deterioration are a real problem that has a real impact on our community. Improving our response and efficiency to treat the problem can reduce the cost to cities (road repair) and individual (car repair and health costs) but more importantly prevent traffic injuries and casualties.

1.1 Pothole mechanism

The mechanism leading to the formation of a pothole starts with a deterioration of the sub-bas layer (located under the pavement). When the soil compresses, shifts, and weakens, an air or water pocket forms. As car apply load on the now unsupported pavement, it bends and cracks leading to the formation of the hole.

1.2 Client

For this project, we will be working for the city of Boston and more specifically for its Department of Public Works. The city has created a service to report potholes through a 3-1-1 call. People are now able to report pothole and provide a specific location. This first step to the "smart response" to the problem will be our starting point. By using the data collected by the city, we will perform an analysis of the data in order to identify major contributing factors to this issue.

1.3 Objectives

The objectives of this investigation are defined as follow:

1. Obtain a better understanding of the problem and its contributors
2. Inspect the relationship between weather and pothole formation
3. Investigate discrepancies between neighborhoods
4. Use a predictive model to estimate future trends and repair time

1.4 Data Sets

The datasets used for this project are the following:

1. The pothole dataset was obtained from the City of Boston website[1]
2. 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.
3. The neighborhoods population and size dataset was obtained from ZipAtlas[2].
4. The neighborhoods map file was downloaded from the Boston Open Data website[3].

The file organization of this project is presented in APPENDIX C.

The pothole problem is a complex issue as it is tied to many parameters. For instance, the weather of the observed region is a major contributor, the quality of the road pavement, and the traffic density. In this report, we will only focus on the impact of the weather properties and the characteristics of the population living in the observed area. As part of a follow-up project, including traffic data through the monitoring program set up by the city could bring essential complementary information to our analysis.

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

1.4.1 Pothole data set

The pothole data set contains the 3-1-1 claims dated from 07/01/2011 to 06/27/20147. It was obtained by querying the 3-1-1 Boston databased. The set contains 41370 claims. Using this dataset constitute a challenge as most of the fields are user input. This means that they are prone to typos, error, and missing information. A particular attention will be required in order to clean the database and extract only meaning full records.

1.4.2 Weather data set

Since harsh weather is considered by experts as a major contributor leading to pothole formation, the National Centers for Environmental Information (NOAA) is added into this analysis. The data set contains monthly weather data from January 2000 to July 2017. We chose to query features that were deemed important for our investigation. This data set is relatively cleaned and will not require a lot of transformation to be incorporated into our analysis package.

1.4.3 Boston neighborhoods data set

In order to compare the pothole distribution over the city of Boston, it was important to group the data by area. Using the grouping by zip code allows a more accurate description of the problem as the data can be described as a function of the population density or the neighborhood area. The data set is extracted for the ZipAtlas website. It features 31 zip codes along with the zip code area and population from 2010. This data set is relatively clean; however, we will have to perform a compatibility analysis before we can join it with our pothole data set.

1.4.4 Boston neighborhood map

In order to create effective visual tools to present the results of our analysis, it was necessary to be able to plot data on a map. To do so, we will use the information provided by the website of the City of Boston. The file is a reshape file that we converted into a JSON file.

2 DATA PREPARATION

2.1 Weather data set

2.1.1 Data cleaning

The data set contains 221 records and 26 features. Since the choice was made to use only the data from one weather station, the features *STATION* and *NAME* are removed from the data frame since they only contain the same value for each record. Moreover, the features *DSND* and *EMSD* do not contain data for the entire time period. For some unknown reasons, these parameters were no longer monitored by the station. Therefore, they were removed from the data set. The data is presented as a daily record, since the set contains only one record per day, the *DATE* feature was chosen as the index of the data frame. This will facilitate the manipulation of the set.

2.1.2 Data validation

After cleaning the data set, it was crucial to explore the data and asses its quality. Using different plotting technics, the following plots were produced:

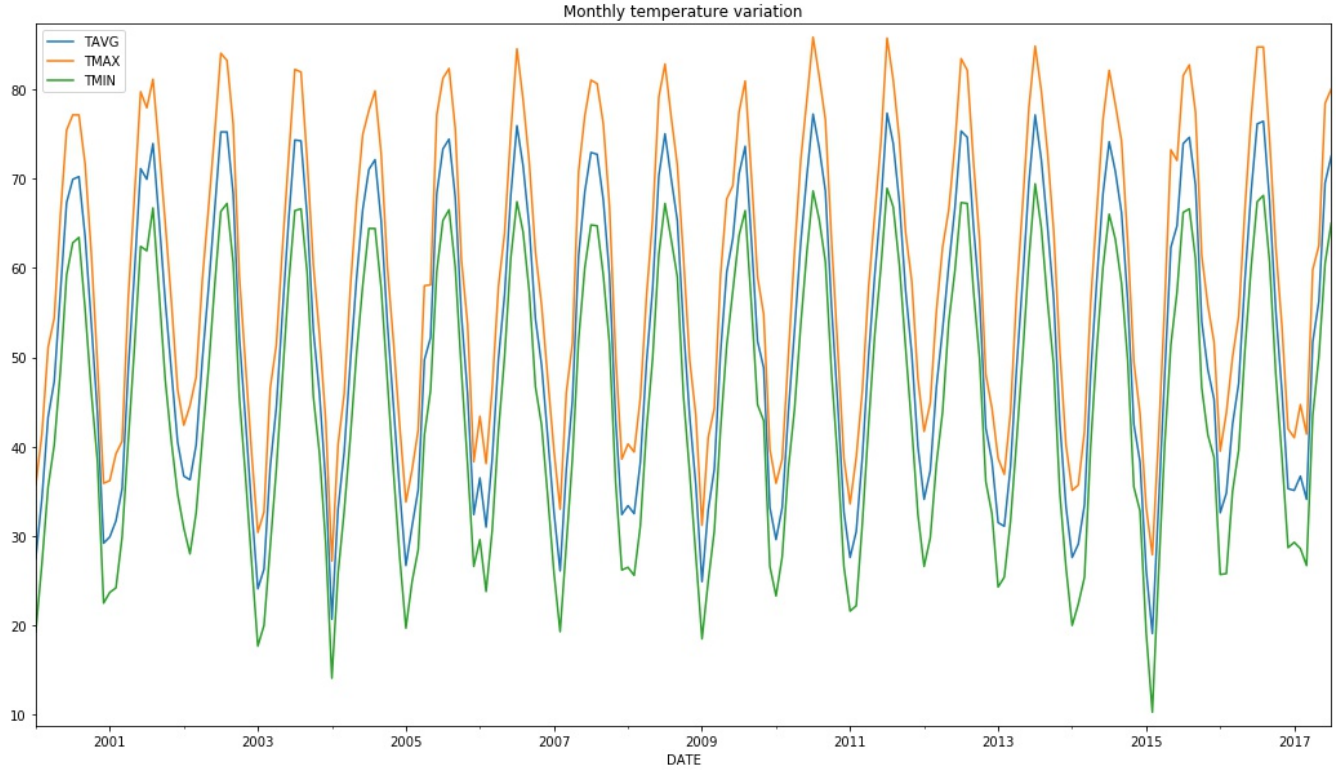


Figure 2-1: Monthly temperature variation

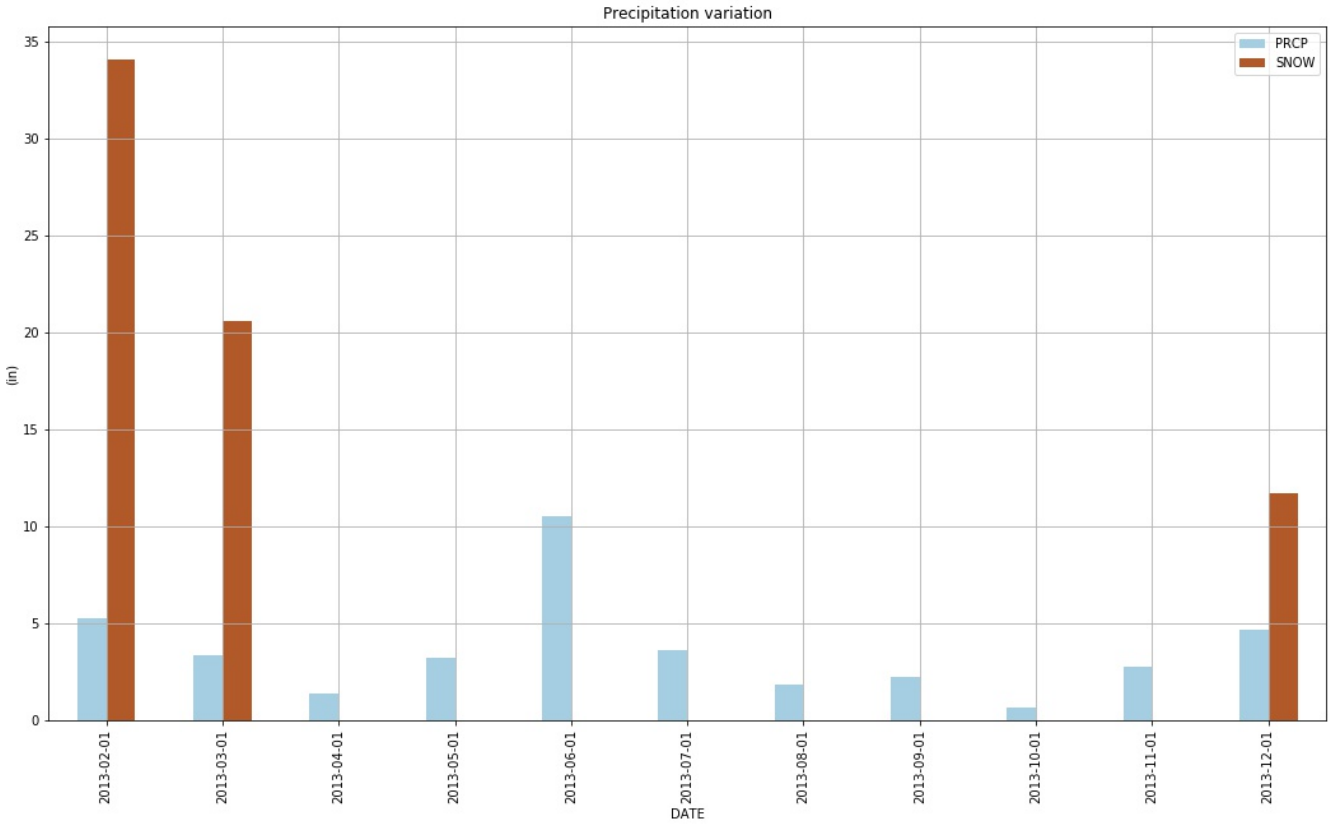


Figure 2-2: Precipitation variation in 2013

After inspection, the data set is ready to be used as is.

2.2 Boston neighborhood data set

2.2.1 Data cleaning

The data set contains 31 records and 24 features. The last four features (*Unnamed 7* to *Unnamed 10*) do not contain anything, they are removed from the set. In order to be able to merge our various data frames, we rename the features of this data frame to merge the feature names of the pothole dataset.

The *Latitude* and *Longitude* features are created by extracting their values from the *Location* feature using pandas extract method. The *Location* feature is then removed from the set. Moreover, for clarity of the plots, the dataset is sorted by zip codes and re-indexed.
Since the dataset includes the population density and the population of each neighborhoods, the neighborhood area is computed by dividing the population by the density of the area. For convenience, the results are converted in acres.

2.2.2 Data validation

The main aspect of the data exploration and validation is first to verify that we have all the Boston neighborhoods contain in the set. Using the folium[4] package, we plot the centers of the neighborhoods (Latitude and Longitude) on top of the neighborhoods map. The map is obtained as a reshape file (converted into a JSON file using mapshaper[5] website). The result is provided below. When inspecting the map, we notice two important ideas:

- First the neighborhood dataset does not contain the small zip codes used to "square" the city boundary. These were created along the year as the city's boundaries were adjusted. This is not an issue for us as these "ghost" zip codes do not contain any (or only a few) house.
- Moreover, Boston is not a common city when it comes to the shape of its neighborhoods. Indeed, several blocks located downtown (for instance the city hall) possess their own zip codes. These are not included in the dataset but do not alter the quality of the data and our analysis.
- Finally, the neighborhood Chestnut Hill (West-most on the map below) does not entirely belong to the city. Therefore, the population and neighborhood area that we posse applies to the entire neighborhood while the pothole data might only cover the area under Boston's jurisdiction.

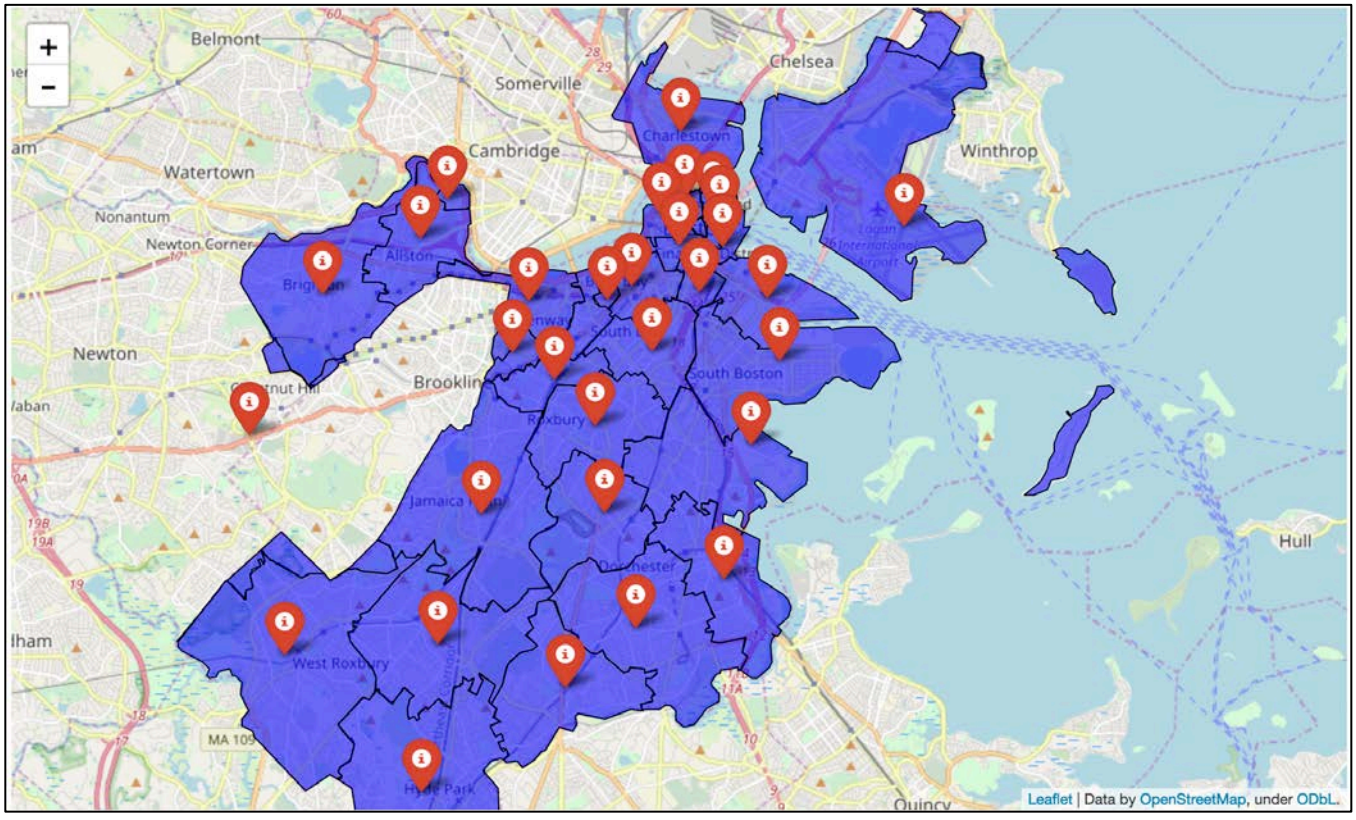


Figure 2-3: Boston zip codes and neighborhood centers

The following plots are created to assess the quality of the data by comparing the results per zip codes.

[4] https://folium.readthedocs.io/en/latest/
[5] http://mapshaper.org/

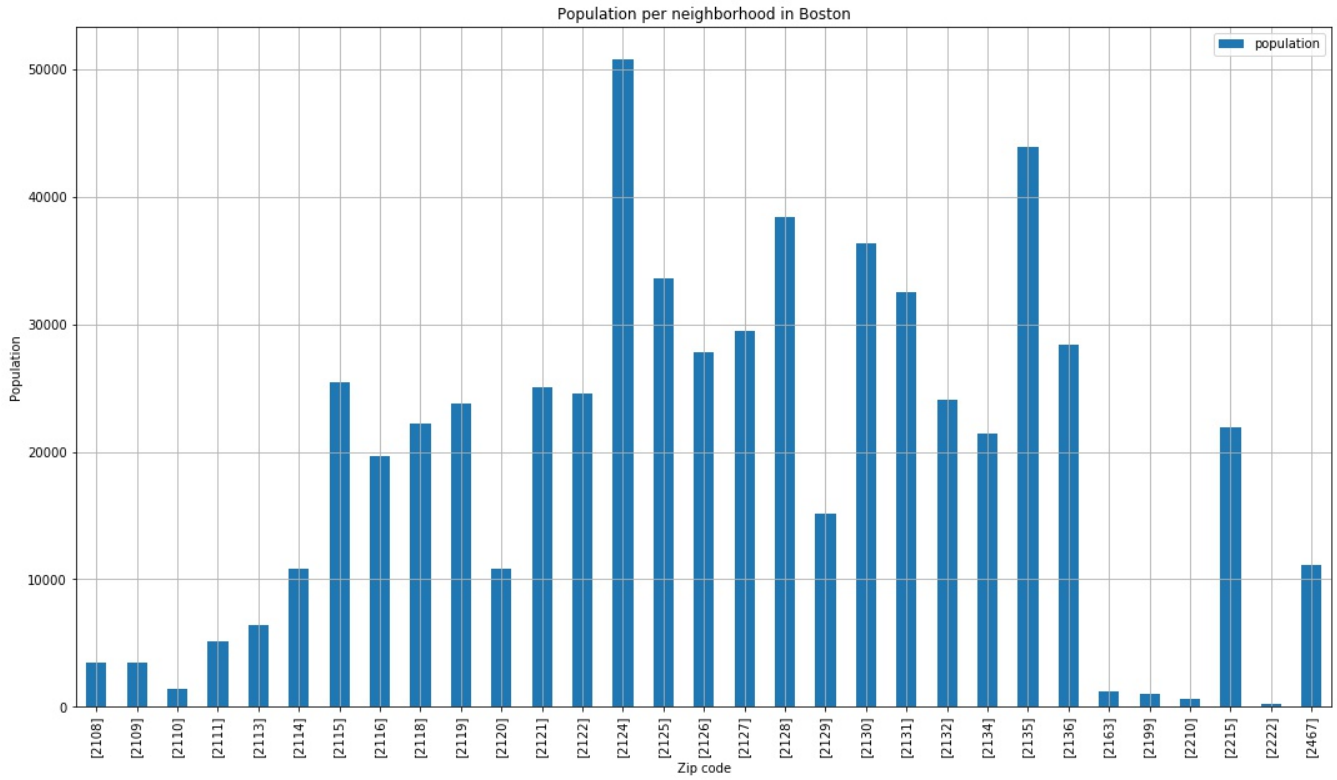


Figure 2-4: Population per neighborhood in Boston

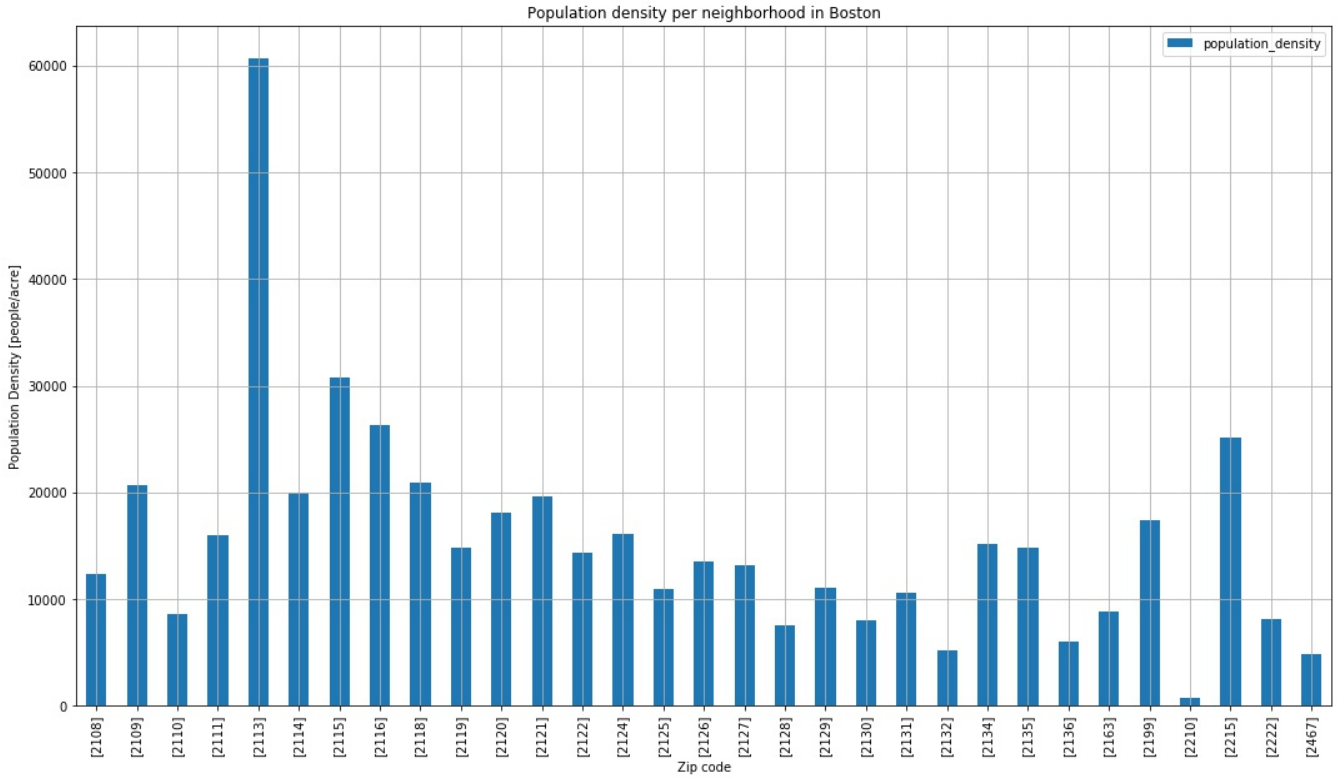


Figure 2-5: Population density per neighborhood in Boston

2.3 Boston neighborhood data set

2.3.1 Data cleaning

Empty features:

After a dive into the data frame structure, several columns were deleted. The features *land_features, Property_Type, Property_ID*, and *Geocoded_location* do not contain any data and are therefore removed from the data frame.

We ran an analysis function on the data set that returns the number of unique values per feature. This tool lead to the deletion of the *SUBJECT, REASON, TYPE,* and *Department* features since they only contain a single unique value for the entire column.

Claim text boxes:

One of the main challenges when dealing with user input text is the lack of consistency. In our case, the claim form contains a text box than can be filled by the employee in charge of the repair. Unfortunately, the conclusion of the claim is contained within this feature. For instance, if the pothole was repaired, the box will contain “Case Resolved”. However, if for any reason the pothole could not be repaired, the case is closed by filling the same text box. Because of the large variety of reason and the different writing styles, this field contains everything and anything from “wrong address” to the most unexpected: “Case Closed Case Noted no pot hole shows picture of coffee mug.” Using Excel for convenience, the *CLOSURE_*REASON column was inspected to identify patterns amongst all the invalid case. The following list was used as a filter:

- *CLOSURE_REASON* contains the string "Case Resolved" => Case is acceptable
- *CLOSURE_REASON* contains the word "duplicate" => Case to be removed. (typos include duplcate, duplicte)
- *CLOSURE_REASON* contains the word "invalid" => Case to be removed.
- *CLOSURE_REASON* contains the words "better location" => Case to be removed.
- *CLOSURE_REASON* contains the words "please contact" => Case to be removed.
- *CLOSURE_REASON* contains the words "please call" => Case to be removed.
- *CLOSURE_REASON* contains the word "test" => Case to be removed.
- *CLOSURE_REASON* contains the words "could not find" => Case to be removed.
- *CLOSURE_REASON* contains the word "cannot" => Case to be removed.
- *CLOSURE_REASON* contains the word "private" => Case to be removed. (versions include prvt)
- *CLOSURE_REASON* contains the word "wrong" => Case to be removed.
- *CLOSURE_REASON* contains the word "nothing" => Case to be removed.
- *CLOSURE_REASON* contains the word "re-subnmit" => Case to be removed. (versions include resubmit)
- *CLOSURE_REASON* contains the words "no pot hole" => Case to be removed. (versions include no potholes, no sink hole)

Ward names:

Another inconsistency in the input format appears in the *ward* feature. Indeed, some of them are provided using a simple number, others are given using the string "Ward X". In order to homogenize the data, this feature was converted into a integer by trimming the unnecessary letters.
Once this transformation performed, we looked at the missing entries for the ward feature. In order to retrieve the ward number from a record, we needed at least the location of the pothole. For this reason, all records without location and ward were simply removed from the dataset.

When inspecting the range of values taken by the ward column, we noticed that a missing/inaccurate location was sent to the ward 0. Indeed, all the potholes that could not be located were given the following properties:

- Ward: 0
- Latitude: 42.3594
- Longitude: -71.0587

Later on, the Google Geo-API will be used. But because of the utilization limit, it was chosen to manually locate the addresses corresponding to the Ward 0 using Google Maps and replace their coordinates and zip in the data frame. This subset consisted of eight unique addresses. The retrieved data was merged to the main data frame.

Missing closing date:

Since the purpose of this study is to assess trends in the pothole reparations, we need to work with cases that were closed. Therefore, any record without a *CLOSED_DT* is removed from the set.

Default location:

After inspection, a large number of cases were assigned to the default location presented above. A mapping by zip code was first suggested, however because of the relatively messy correlation between neighborhood designation and zip codes, a one-to-one match could only be found for two zip codes which would apply to three cases only. To save time, these cases were not fixed with this method.

Since this subset is too large to be dropped, we decided to come up with a location fetcher algorithm. The first step consisted on a feasibility study, to do so, we looked at the record in the subset and defined if whether or not the available data was sufficient to geo-localize the addresses. After a first glance, the data was judged good enough to move on to the next step of the retrieval process. The subset of the potholes without zip code and assigned to the default location (See previous) was extracted and saved as a csv file. The reason behind this choice is that the limitation of the API we decided to use would not work well if integrated in the main note book. Therefore, we created a specific dedicated to the address search. The Google Geo-API was used, it is based on the algorithm behind Google Maps. Our algorithm will go through the csv file and restart to where it previously stopped, it will extract a missing location, send the request to the API and try to retrieve the zip code, latitude, and longitude corresponding to the location description. When creating our API request function, we had to consider two limitations. First, the API has a search limit of 2500 requests per day (we had roughly 4500 locations to search) and the API limits the frequency of the searches. This last constraint forced us to use a time delay of 0.2s (value

based on trial and error) between each iteration in the main loop. Once a location has been retrieved, it is stored in an output csv file.

Moreover, they were two groups of addresses that could potentially lead to issues. Addresses without specific address number were assign a 0, others were assigned a address number range. For the first group, we drop the zero and search for the street name alone (by default, the API locates the address corresponding to the middle of the street). For the second group, we decided for simplicity to keep the second street number (12-15 First Street becomes 15 First Street).

Once this process was completed, several locations were still missing, out of solution to retrieve their location, we decided to drop them from the set. It was determined after inspection that these locations were not retrieved because the zip code did not match the street name. Finally, the retrieved locations were merged with the main data frame.

Precinct:

Since the precinct map is regularly updated, it was decided not to use this feature in the final analysis. The *precinct* column is removed from the set.

2.3.2 Additional feature

Photos:

The original data set contained the *SubmittedPhoto* and *ClosedPhoto* features. Since each feature contained the URL of the pictures taken to support the claim or the picture of the repair, we judge important to keep these features for the analysis. After all, a claim with picture might for a quicker repair… However, the provided URL addresses were not really useful as is. In order to be integrated in the analysis, both of them were converted into the Boolean features *SubmittedPhoto_Bool* and *ClosedPhoto_Bool.* If a record contains an URL, its corresponding Boolean feature is set to True.

Time to repair:

By subtracting the claim closed date and the claim creation date, we create a new feature *time_repair*. When examining the results, we found that some of the results were negative or extremely large. We decided to remove record with negative repair time or repair time greater that six months. Indeed, with constant road constructions going on in the city and the harsh winters, it is possible for a pothole to be inaccessible for six months. For larger repair time, these values are either extremely unlikely or simply incorrect.

Intersections:

During the location retrieval process, we noticed that many pothole cases were listed on road intersection. In order to assess whether or not the potholes develop more often on intersection, a new feature was added to the set: *is_intersection*. If a pothole location contains the word “intersection”, then a True value is assigned to the record.

3 DATA ANALYSIS

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 weather impact the repair time?
3. How does the number of claims differ by neighborhood?
4. Does the repair time appear to be similar amongst all neighborhoods?
5. Are intersections more prone to develop potholes?
6. Does reporting the claim with a photo improve the repair time?

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.

3.1 Number of monthly claims

3.1.1 Correlation with weather data

3.1.1.1 Number of 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 3-1: 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 3-1, we identify Spring as being the season with the maximum of claims. The other three seasons (Summer, Fall, and Winter) have similar properties. Figure 3-1 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.

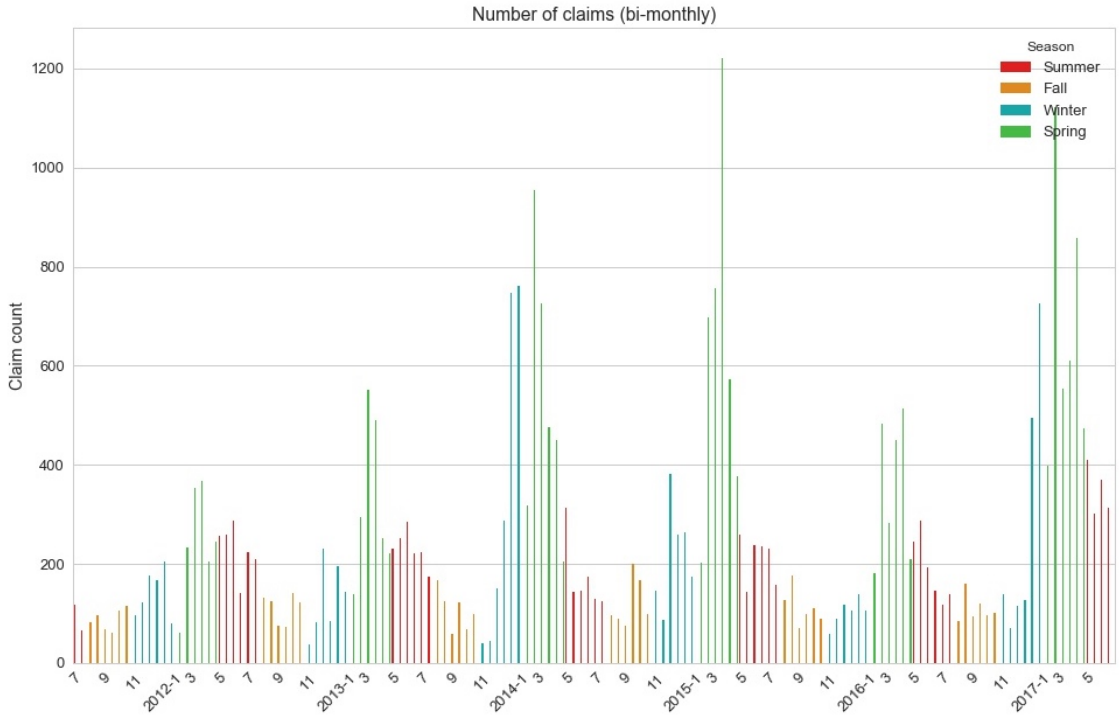


Figure 3-1: Bi-monthly claim count

The average number of snow days and freezing days are depicted in Figure 3-3. 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.

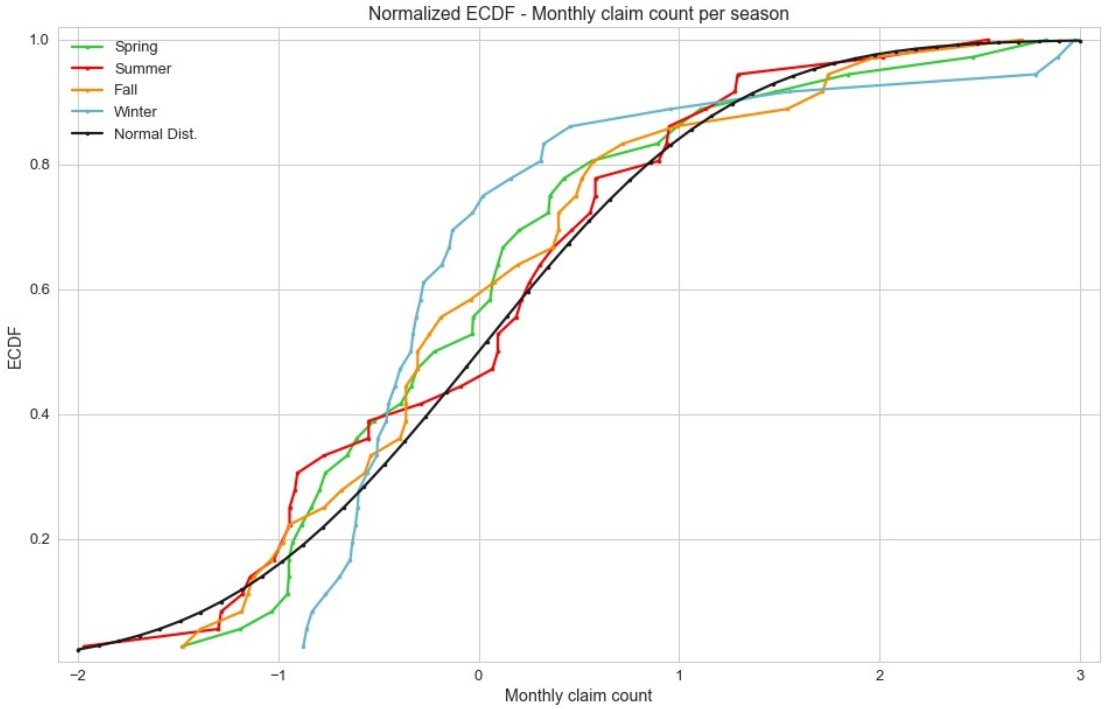


Figure 3-2: Bi-monthly number of claims – Empirical cumulative distribution functions

Figure 3-2 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 3-2: 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 3-4 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.

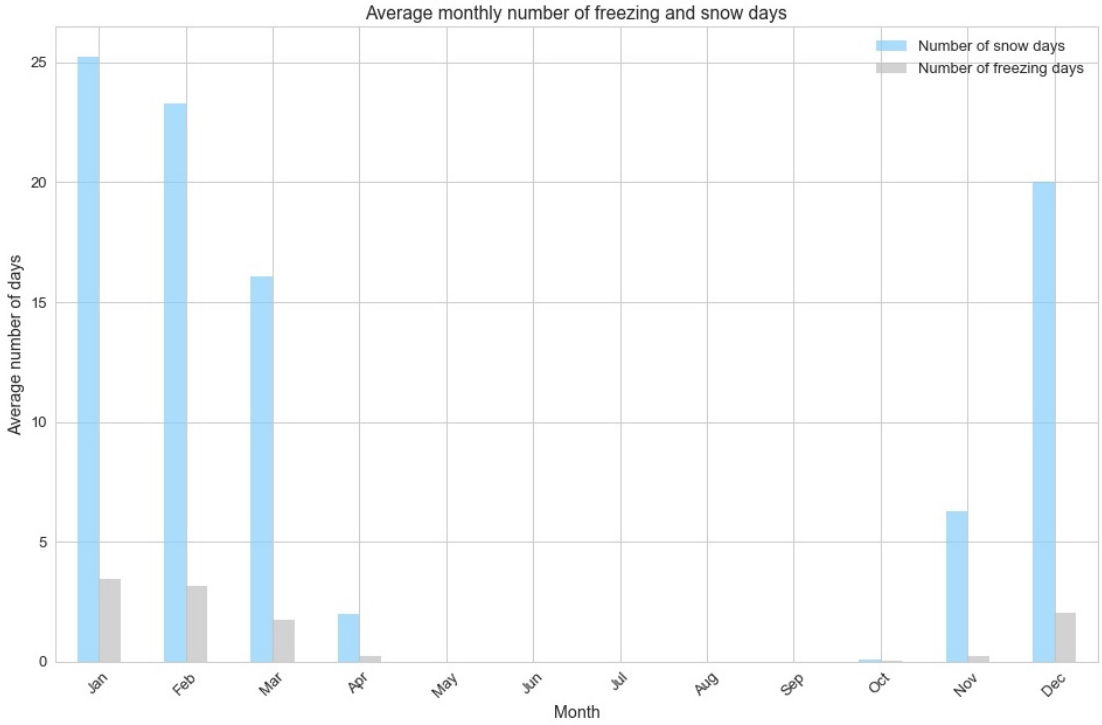


Figure 3-3: Monthly average of freezing and snow days

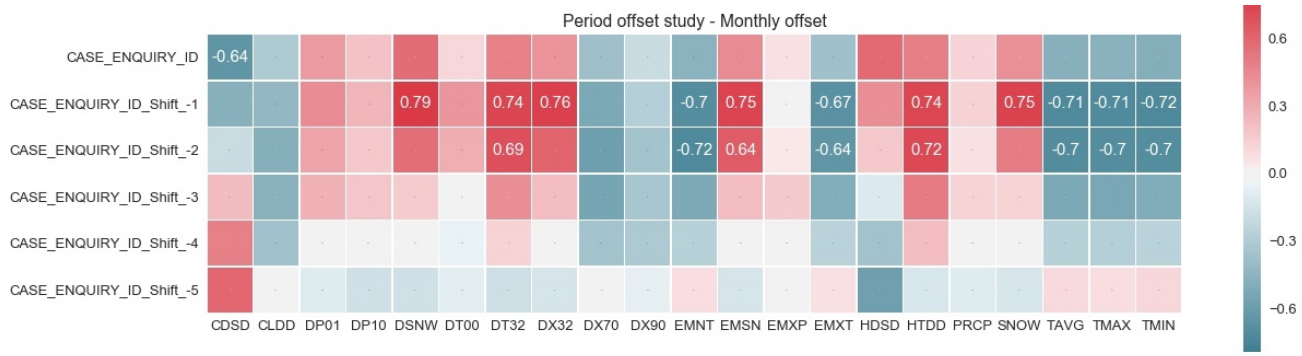


Figure 3-4: Correlation heat map - Number of claims and weather data

3.1.1.2 Linear regressions

Appendix D contains 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 3-3: 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 3-3 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.

When considering the residuals of the regression between the monthly average temperature and the monthly number of claims, t 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.

3.1.2 Correlation with neighborhood data

3.1.2.1 Number of claims per neighborhood

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 3-5: Heat map population versus number of claims and repair time

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.

Appendix E presents the discrepancy in the pothole density per neighborhoods. A few neighborhoods have an extremely high pothole density.

Table 3-4: 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 3-4 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 3-6 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.

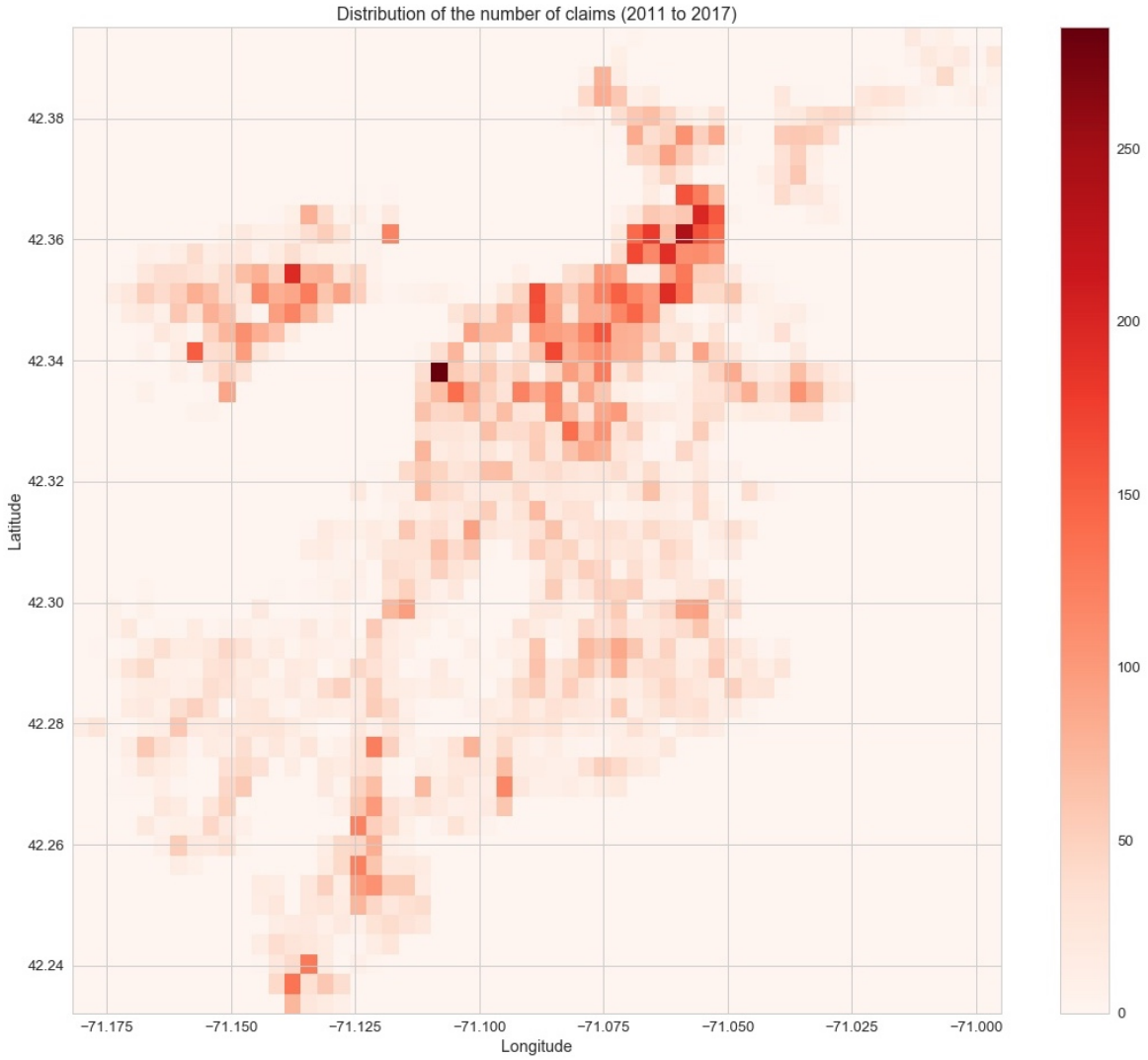


Figure 3-6: Number of claims (2011 to 2017)

3.1.2.2 Linear regressions

The linear regression plots are presented in Appendix F. The results of the linear regression are listed in Table 3-5.

Table 3-5: Regression results

	Population	Area [acres]
Slope	0.0146	0.134
Intercept	269.6	415.1

3.2 Number of claims and intersections

When looking at the raw data, it seems that abnormally large portion of the claims are made concerning pothole case on intersection. In this section, we will investigate if the distribution of claims over the intersection feature. Table 3-6 presents the claim distribution per year.

Table 3-6: Intersection claims

OPEN_DT	is_intersection	CASE_ENQUIRY_ID	Percent
2011	FALSE	875	68.57%
	TRUE	401	31.43%
2012	FALSE	2885	68.03%
	TRUE	1356	31.97%
2013	FALSE	3122	64.54%
	TRUE	1715	35.46%
2014	FALSE	4715	64.84%
	TRUE	2557	35.16%
2015	FALSE	4412	67.11%
	TRUE	2162	32.89%
2016	FALSE	3068	66.61%
	TRUE	1538	33.39%
2017	FALSE	4267	64.38%
	TRUE	2361	35.62%

As shown above, the number of potholes located within intersection is extremely high compared to the proportion of road that constitutes intersections. This can be explained by the failure mechanism of the top layer of the road. Indeed, the asphalt works well in compression but is not as good to support shear loads which happen when a car changes direction.
If we conservatively assume that intersection accounts for 10% of the roads in the city, while they account for ~33% of the pothole claims.

3.3 Repair time

3.3.1 Correlation with weather data

3.3.1.1 Repair time per season

Now that we know the cold weather has a significant impact on the road damage frequency, we will be looking at the impact of the weather on the repair time. The repair time is defined as the difference between the claim closure date and the claim creation date. The pothole database can be modified by making a request to 311 or by city workers. A significant number of potholes are discovered and fixed at

the same time by city workers patrolling in the street. In order to only focus on claims made by "normal" users, we will restrain the data set to potholes that took more than half a day to be fixed.

Because of the effect of potential outliers on the mean repair time, we will be using the median repair time. The results provided in Table 3-7 indicates a similar response time between season. When looking at the actual distribution of the media repair time by bi-monthly interval, we concluded that with only a few outliers and a median time of 2 days to fix a pothole, the city has a good response time. Moreover, the response time barely varies from one season to the other, which indicates that the city responds well to pothole claims even during the winter.

Table 3-7: Median repair time per season

Season	Spring	Summer	Fall	Winter
Repair Time [Days]	1.20	1.10	1.18	1.11

Figure 3-7 depicts the median repair time for claims grouped over a two-week period. The first month of the spring of 2015 appears to be an outlier. Let's investigate why the time to repair was so long. Before looking at the data for this specific month, we can make a hypothesis, this increase in time repair is probably due to the snow falls. Indeed, New England experienced the worst winter in decades. The accumulation of snow, the long lasting freezing temperatures prevented the city to fix the potholes.

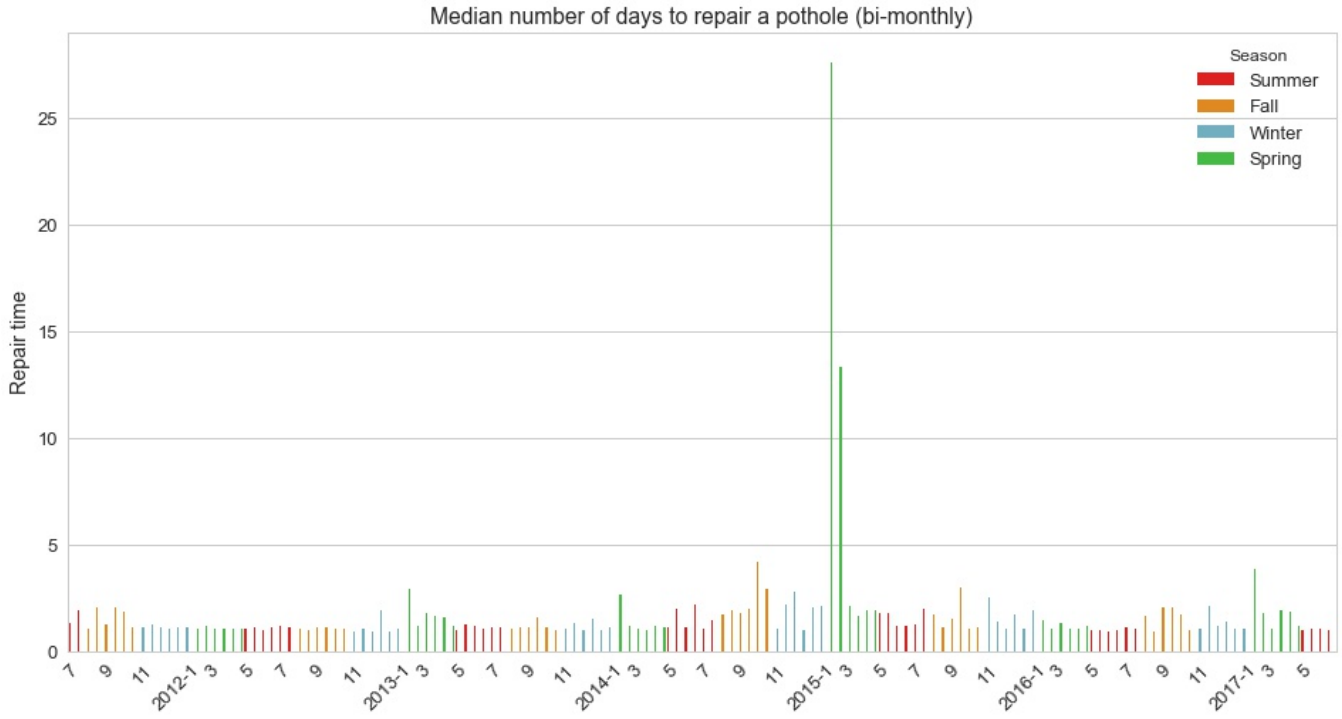


Figure 3-7: Median repair time over from 2011 to 2017

As shown above, the distribution of the repair time over the month of February 2015 is spread from 0 to 80 days with a linear decrease. The data for the month of February 2014 is mostly contained in the 0 to 5 day-bin. There are two factors that can explain the difference:

1. The city workers in 2015 were also allocated to snow removal, therefore, the number of the team available to repair pothole was less than that of 2014.
2. Because of the expected heavy snowfall, the city waited for the multiple storms to end and started fixing pothole once the bad weather was over.

Figure 3-8: Heat map repair time versus weather data

As shown above, there is no clear correlation between the weather data and the repair time. This confirms that except a few outliers, the city has a fast and consistent response when it comes to fixing potholes reported through 3-1-1 calls.

3.3.1.2 Linear regressions

Appendix G contains the linear regression of the various regressions that were performed in order to assess the relationships between the weather data and the median repair time.

Table 3-8: Regression results

	Snow days	Freezing days	Average temperature	Precipitation
Slope	0.036	-0.001	0.001	-0.005
Intercept	1.281	1.318	1.264	1.336

The results of the linear regressions combined with the correlation coefficients lead to the conclusion that the weather only has a minor effect on the repair time. For instance, the linear regressions involving the number of freezing days and the average temperature have small slope values.

3.3.2 Correlation with neighborhood data

3.3.2.1 Repair time per neighborhood

The purpose of this section is to investigate if pothole repair takes longer in certain neighborhood. Table 3-9 lists the median repair time per zip code. From the results, we can conclude that even with a two times ration between the longer repair time compared to the quicker one, two days is an acceptable

delay for public work action. Overall, the city is efficient and responsive not matter where the claim is made.

Table 3-9: Median repair time per neighborhood

Zip code	Repair time (days)	Zip code	Repair time (days)
2136	1.00	2467	1.47
2163	1.01	2134	1.50
2120	1.01	2135	1.56
2215	1.05	2115	1.59
2119	1.07	2210	1.60
2124	1.07	2127	1.67
2126	1.07	2129	1.67
2122	1.07	2118	1.68
2131	1.07	2113	1.72
2132	1.07	2116	1.74
2130	1.14	2108	1.76
2121	1.18	2111	1.80
2125	1.28	2128	1.95
2110	1.41	2199	1.96
2109	1.42	2114	2.03

3.3.2.2 Linear regressions

Appendix I contains the linear regression plot of the median repair time as a function of the population.

Table 3-10: Regression results

	Population
Slope	-7.43e-6
Intercept	1.57

As shown in Table 3-10, the median repair time is negatively correlated with the population number. This counter-intuitive result leads to the conclusion that larger neighborhoods are more reactive to claims.

3.4 Number of claims and repair time

At this stage of the analysis, we have studied the number of claims and the repair time separately. In this section, we will investigate how the number of claims impact the repair time. The linear regression between the two properties is presented in Appendix J.

Table 3-11: Regression results

	Number of claims
Slope	6.14e-4
Intercept	1.17

This time, the results validate our expectations; it takes longer for a neighborhood to fix a pothole if the neighborhood is prone to receiving more claims.

3.5 Repair time and photos

Recently, the 3-1-1 claim platform was updated to include the option to attach a picture to the claim. This feature is meant to help the repair team locate the pothole. In this section, we will investigate whether including a picture with your claim reduces the repair time. We will do so by performing a hypothesis testing.

Null hypothesis: Ho -> The mean repair time is the same for cases with picture or without picture.
Alternative hypothesis: Ha -> The mean repair time for cases with pictures is different from the ones without pictures.

For this evaluation, we do not make any assumption regarding the distribution of the repair time. We will therefore perform the test using a bootstrap method.

Mean repair time with picture: 4.02 days
Mean repair time without picture: 3.80 days
Empirical mean difference: 0.22 days

Mean repair time for entire set of claims: 3.89 days

We generate shifted sets based on the assumption of equal mean:

$$X_{shifted} = X + \bar{Y} - \bar{X}$$

$Where$
$X\ is\ the\ set\ of\ claims\ with\ or\ without\ picture$
$\bar{Y}\ is\ the\ mean\ repair\ time\ for\ the\ entiere\ set\ including\ claims\ with\ and\ without\ pictures$
$\bar{X}\ is\ the\ mean\ repair\ time\ of\ the\ set\ with\ or\ without\ picture$

We obtain a p-value of 0.49, which means that we cannot reject the null hypothesis. In conclusion, attaching a picture to the claim does not have a statistical impact on the repair time.

4 Preparation for Predictive Models

We now have a better understanding of which parameters appear to be the major contributors to the pothole problem. The next step of this case-study is to create two different models using machine learning techniques to predict the evolution of the claims in a neighborhood at a certain time of the year and to predict the repair time associated to a new claim.

5 Preliminary Conclusions

After a detailed review and evaluation of the three sets of data, the following answers can be provided:

1. On average, the city is efficient when it comes to fixing potholes reported by 311 calls. However, at least three zip codes are not as efficient as the rest as fixing a pothole takes more than 20 days on average. Our hypothesis to explain these results is the fact that these three neighborhoods are small with few people with roads that are used by many to commute to work.
2. 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. Finally, our study shows that there is a lag effect of one month between a period of bad weather and a peak in pothole claims.
3. Surprisingly, the city is doing a good job at maintaining an efficient service during and right after a tough winter. Except for the winter of 2015 (historic snowfall record), the city is responsive and potholes are typically fixed within couple days on average.
4. As expected, the number of claims is positively correlated with the size (area and population of a neighborhood).
5. A few neighborhoods report a larger pothole density (count of claims per 100 inhabitants). These areas correspond to the older neighborhood that are small but densely populated.
6. Intersections are prone to pothole formation.
7. Attaching a picture to a claim does not improve the repair time.

APPENDIX A: VARIABLE DEFINITION

Pothole data set

Source: https://data.cityofboston.gov/City-Services/Requests-for-Pothole-Repair/n65p-xaz7/data
Format: CSV
Features:

- *CASE_ENQUIRY_ID*: Case number assigned to the request to repair the pothole.
- *OPEN_DT*: Date and time of the repair request.
- *TARGET_DT*: Scheduled time for repair.
- *CLOSED_DT*: Date and time the case was closed.
- *OnTime_Status*: ONTIME if *CLOSED_DT*>TARGET_DT
- *CASE_STATUS*: Case status.
- *CLOSURE_REASON*: Reason for the case closure.
- *CASE_TITLE*: Request type. In this case, the type is "Request for Pothole Repair".
- *SUBJECT*: The city department in charge of the request.
- *REASON*: Reason for the case opening.
- *TYPE*: Specific reason for the case opening. In this case, the type is "Request for Pothole Repair".
- *QUEUE*: Code corresponding to the department per neighborhood in charge of the repair.
- *Department*: Code corresponding to the department in charge of the repair.
- *SubmittedPhoto*: URL of the photo taken to support the claim.
- *ClosedPhoto*: URL of the photo taken to support the repair.
- *Location*: Address of the pothole.
- *fire_district*: Fire district corresponding to the pothole location.
- *pwd_district*: Public Work district corresponding to the pothole location.
- *city_council_district*: City Council district corresponding to the pothole location.
- *police_district*: Police district corresponding to the pothole location.
- *neighborhood*: Neighborhood corresponding to the pothole location.
- *neighborhood_services_district*: Neighborhood Services district corresponding to the pothole location.
- *ward*: Ward corresponding to the pothole location.
- *precinct*: Precinct corresponding to the pothole location.
- *land_usage*: Blank column.
- *LOCATION_STREET_NAME*: Street number and street name corresponding to the pothole location.
- *LOCATION_ZIPCODE*: Zip code corresponding to the pothole location.
- *Property_Type*: Blank column.
- *Property_ID*: Blank column.
- *LATITUDE*: Latitude of the pothole location.
- *LONGITUDE*: Longitude of the pothole location.
- *Source*: Source of the request.
- *Geocoded_Location*: Blank column

Weather data set

Source: Upon request on NOAA site
Format: CSV
Features:

- *STATION*: (17 characters) is the station identification code.
- *NAME*: (max 50 characters) is the name of the station (usually city/airport name). This is an optional output field.
- *DATE*: is the year of the record (4 digits) followed by month (2 digits) and day (2 digits).
- *CDSD*: Cooling Degree Days (season-to-date). Running total of monthly cooling degree days through the end of the most recent month. Each month is summed to produce a season-to-date total. Season starts in January in Northern Hemisphere and July in Southern Hemisphere. Given in Celsius or Fahrenheit degrees depending on user specification.
- *CLDD*: Cooling Degree Days. Computed when daily average temperature is more than 65 degrees Fahrenheit/18.3 degrees Celsius. CDD = mean daily temperature - 65 degrees Fahrenheit/18.3 degrees Celsius. Each day is summed to produce a monthly/annual total. Annual totals are computed based on a January – December year in Northern Hemisphere and July – June year in Southern Hemisphere. Given in Celsius or Fahrenheit degrees depending on user specification.
- *DP01*: Number of days with >= 0.01 inch/0.254 millimeter in the month/year.
- *DP10*: Number of days with >= 1.00 inch/25.4 millimeters in the month/year
- *DSND*: Number of days with snow depth >= 1 inch/25 millimeters.
- *DSNW*: Number of days with snowfall >= 1 inch/25 millimeters.
- *DT00*: Number of days with maximum temperature <= 0 degrees Fahrenheit/-17.8 degrees Celsius.
- *DT32*: Number of days with minimum temperature <= 32 degrees Fahrenheit/0 degrees Celsius.
- *DX32*: Number of days with maximum temperature <= 32 degrees Fahrenheit/0 degrees Celsius.
- *DX70*: Number of days with maximum temperature >= 70 degrees Fahrenheit/21.1 degrees Celsius.
- *DX90*: Number of days with maximum temperature >= 90 degrees Fahrenheit/32.2 degrees Celsius.
- *EMNT*: Extreme minimum temperature for month/year. Lowest daily minimum temperature for the month/year. Given in Celsius or Fahrenheit depending on user specification.
- *EMSD*: Highest daily snow depth in the month/year. Given in inches or millimeters depending on user specification.
- *EMSN*: Highest daily snowfall in the month/year. Given in inches or millimeters depending on user specification
- *EMXP*: Highest daily total of precipitation in the month/year. Given in inches or millimeters depending on user specification.
- *EMXT*: Extreme maximum temperature for month/year. Highest daily maximum temperature for the month/year. Given in Celsius or Fahrenheit depending on user specification.
- *HDSD*: Heating Degree Days (season-to-date). Running total of monthly heating degree days through the end of the most recent month. Each month is summed to produce a season-to-date total. Season starts in July in Northern Hemisphere and January in Southern Hemisphere. Given in Celsius or Fahrenheit degrees depending on user specification.
- *HTDD*: Heating Degree Days. Computed when daily average temperature is less than 65 degrees Fahrenheit/18.3 degrees Celsius. HDD = 65(F)/18.3(C) – mean daily temperature. Each day is summed to produce a monthly/annual total. A
- *PRCP*: Total Monthly/Annual Precipitation. Given in inches or millimeters depending on user specification.
- *SNOW*: Total Monthly/Annual Snowfall. Given in inches or millimeters depending on user specification

- *TAVG*: Average Monthly/Annual Temperature. Computed by adding the unrounded monthly/annual maximum and minimum temperatures and dividing by 2.
- *TMAX*: Monthly/Annual Maximum Temperature. Average of daily maximum temperature given in Celsius or Fahrenheit depending on user specification
- *TMIN*: Monthly/Annual Minimum Temperature. Average of daily minimum temperature given in Celsius or Fahrenheit depending on user specification

Boston neighborhoods data set

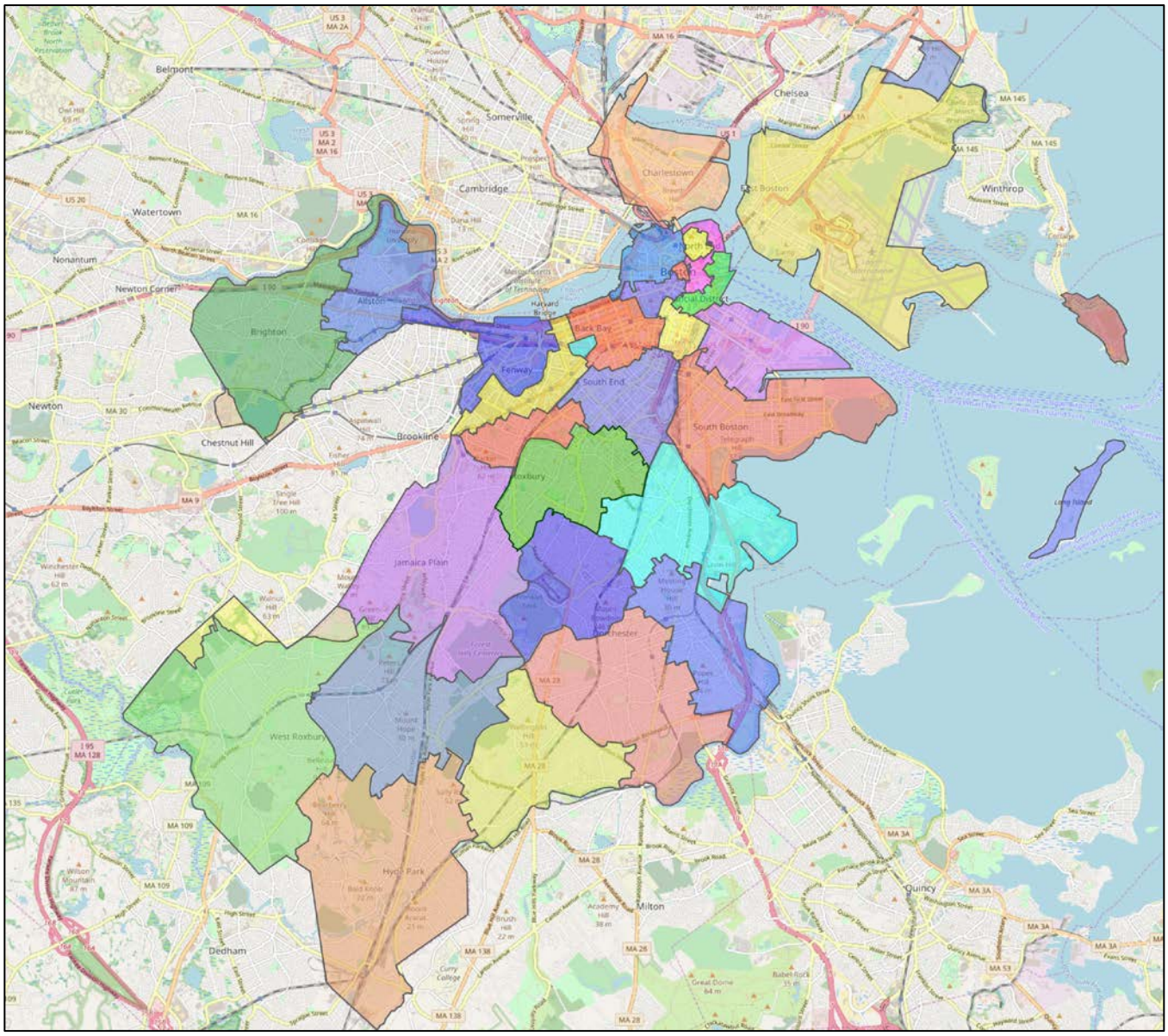
Source: http://zipatlas.com/us/ma/zip-code-comparison/population-density.htm
Format: CSV
Features:

- *#*: Index
- *Zip Code*: Neighborhood zip code
- *Location*: Neighborhood longitude and latitude
- *City*: Neighborhood city
- *Population*: Neighborhood population
- *People / Sq. Mile*: Neighborhood density
- *National Rank*: Density national rank
- *Unnamed: 7*: Blank column
- *Unnamed: 8*: Blank column
- *Unnamed: 9*: Blank column
- *Unnamed: 10*: Blank column

Boston neighborhoods map

Source: http://bostonopendata-boston.opendata.arcgis.com/datasets/53ea466a189b4f43b3dfb7b38fa7f3b6_1
Format: RESHAPE then JSON
Conversion: http://mapshaper.org/

APPENDIX B: BOSTON NEIGHBORHOOD MAP



APPENDIX C: PROJECT FILES AND SOFTWARE

File Organization

The set of files are located in the GitHub repository: https://github.com/tdody/Springboard-Capstone-1

Name	Format	Description
00_Data_Wrangling-Weather	iPython	Data cleaning of the weather data set
01_Data_Wrangling_Boston	iPython	Data cleaning of the neighborhood data set
02a_Data_Wrangling_Potholes	iPython	Data cleaning of the pothole data set
02b_Google_Geo_API_Fetcher	iPython	Location retrieval tools using Google API
03_Data_Story_Telling	iPython	Data analysis and descriptive statistics
Cleaned Data	Folder	Contains the cleaned CSV output from the notebooks 00 to 03
Figures	Folder	Contains the JPG figures of the plots from the notebooks 00 to 03
Original Data	Folder	Contains the existing CSV and JSON files
Intermediate Data	Folder	Intermediate data set used by the 02b notebook
REAME	Text	Repository Description
Data Wrangling Report	PDF	Presents the results of the 00 to 02b notebooks
Capstone Project 1 – Proposal	PDF	Initial project idea
EDA Report	PDF	Presents the results of the 03 notebook

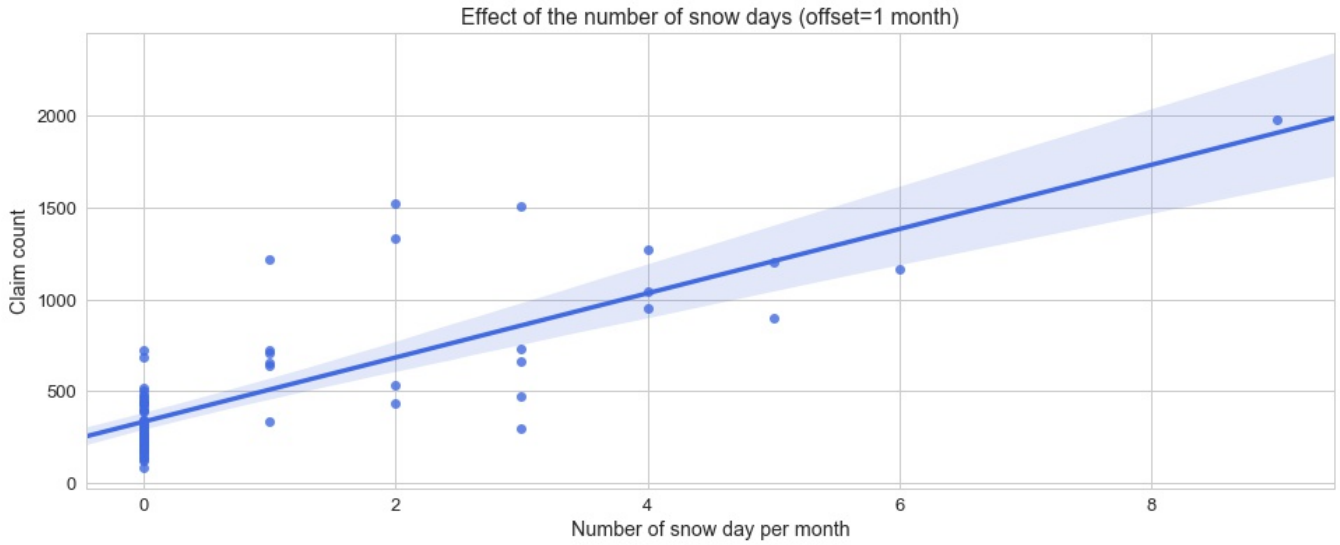
Software

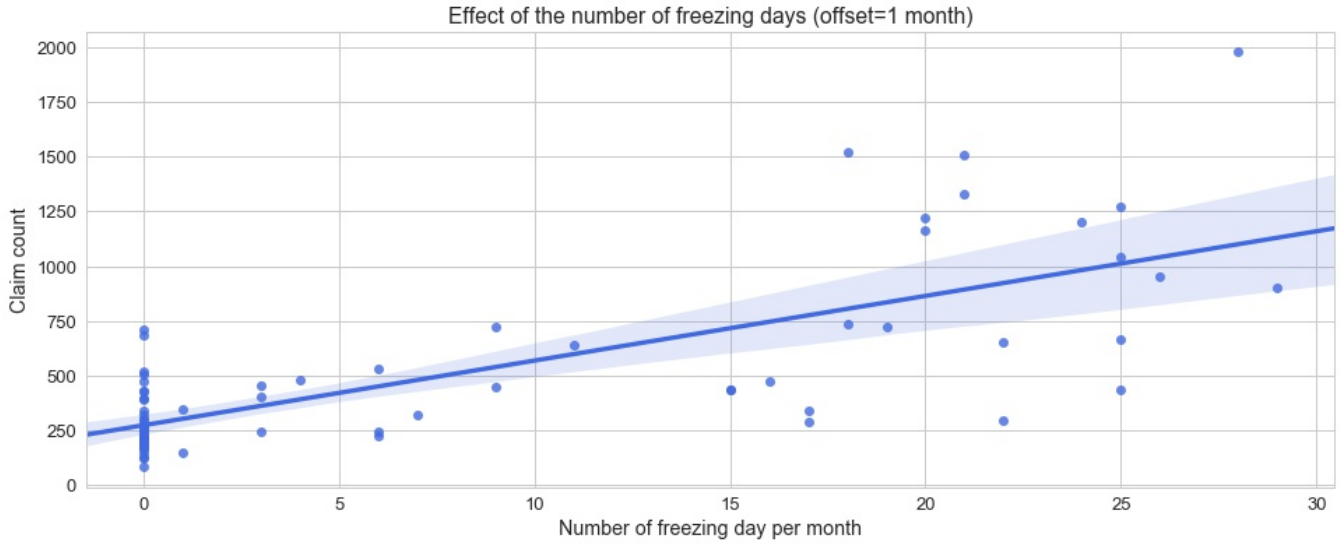
Name	Version
Python	3.6.1
Jupyter NoteBook	5.0.0
Anaconda	4.4.0

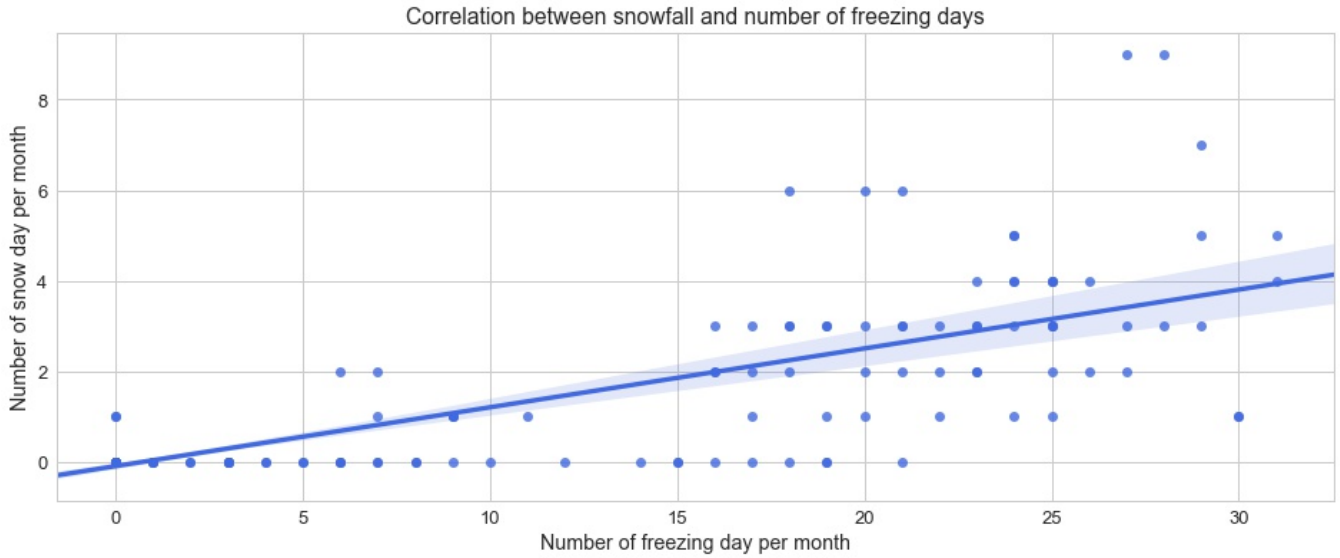
Modules

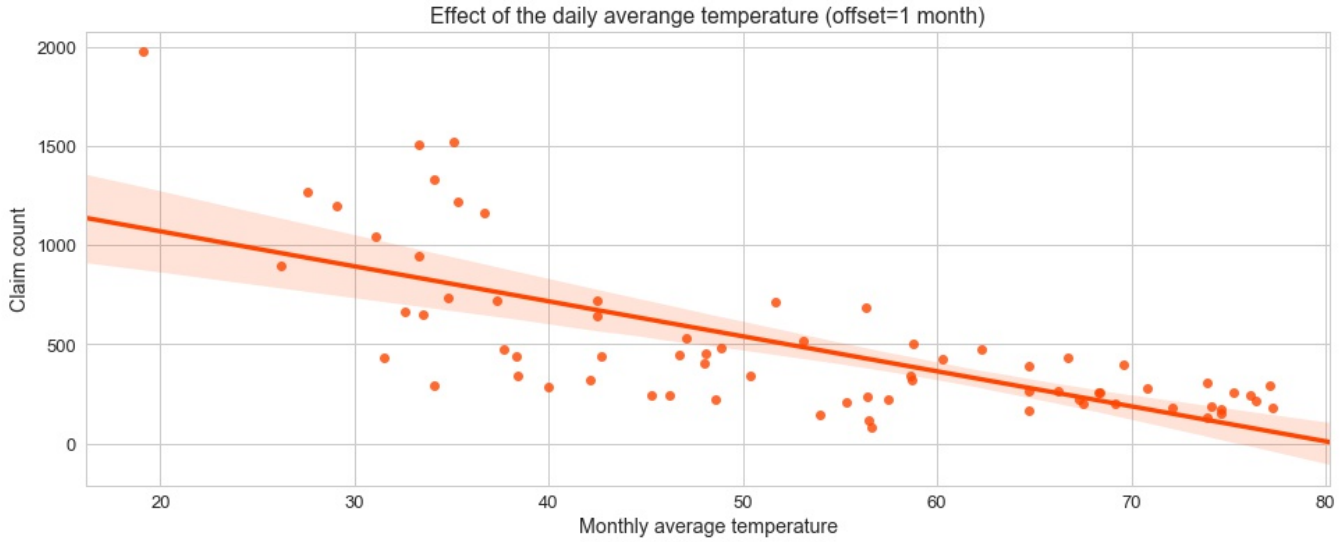
Name	Version
Pandas	0.20.1
Numpy	1.12.1
Matplotlib	2.0.2
Seaborn	0.7.1
Scipy	0.19.0
Folium	0.3.0

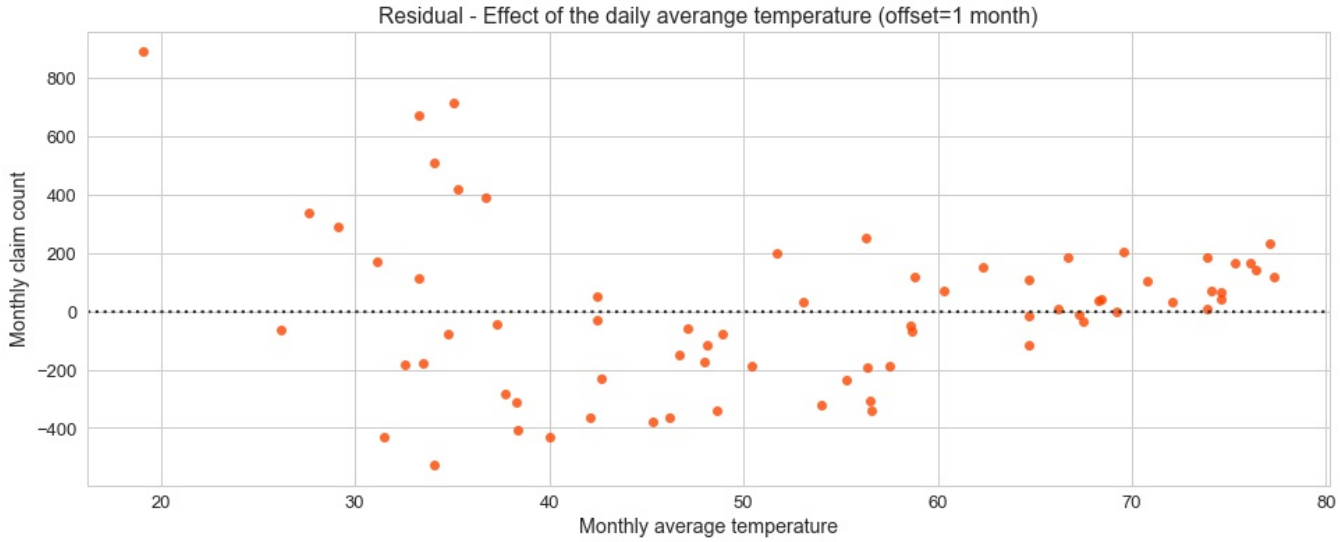
APPENDIX D: REGRESSION PLOTS – NUMBER OF CLAIMS VS. WEATHER DATA

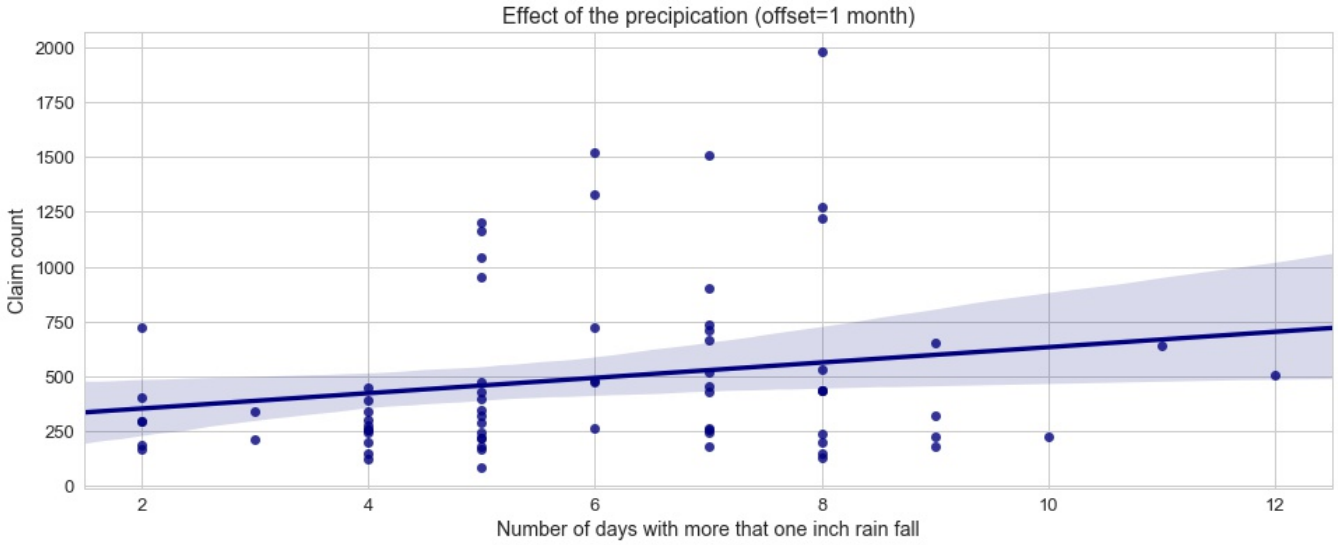




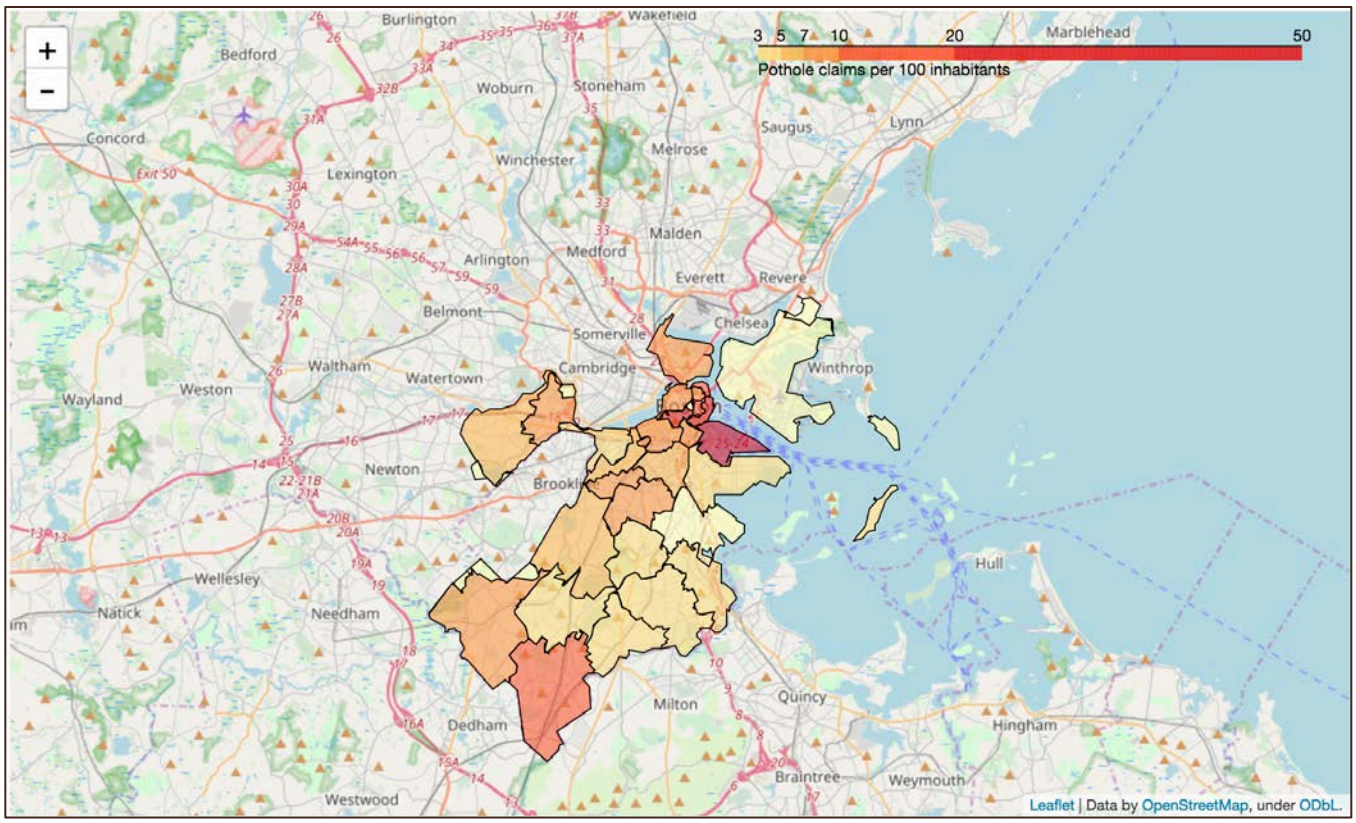

Correlation between snowfall and number of freezing days
Number of snow day per month
Number of freezing day per month
0
2
4
6
8
0
5
10
15
20
25
30


Effect of the daily averange temperature (offset=1 month)
Claim count
Monthly average temperature
0
500
1000
1500
2000
20
30
40
50
60
70
80

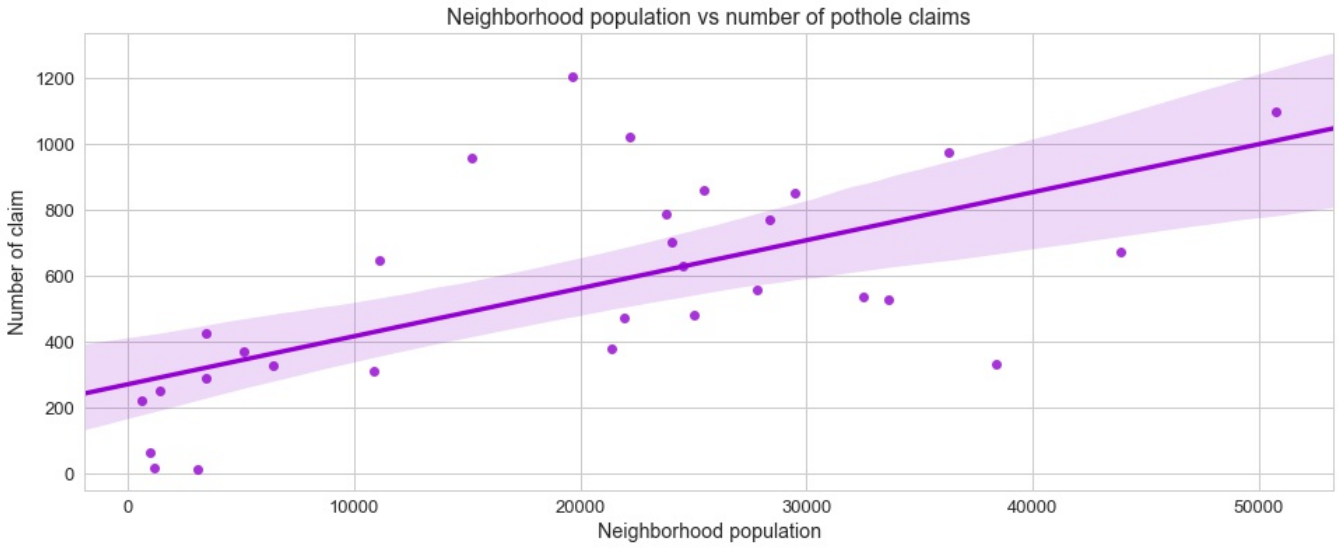

Residual - Effect of the daily averange temperature (offset=1 month)
Monthly claim count
Monthly average temperature
−400
−200
0
200
400
600
800
20
30
40
50
60
70
80


Effect of the precipication (offset=1 month)
Claim count
2000
1750
1500
1250
1000
750
500
250
0
2
4
6
8
10
12
Number of days with more that one inch rain fall

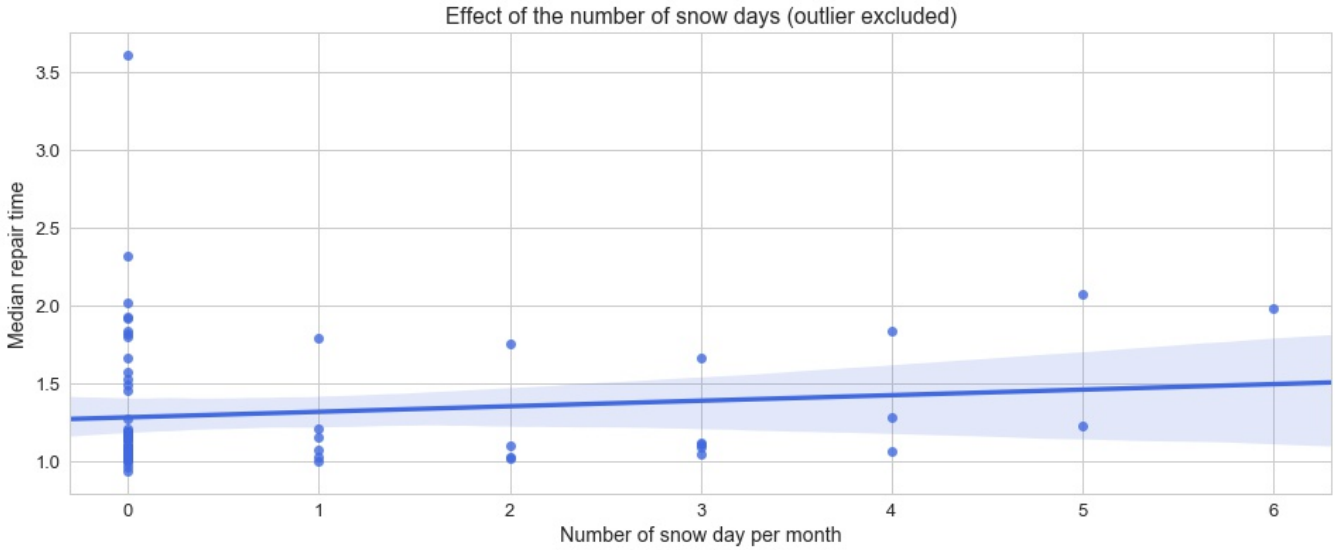
APPENDIX E: MAP – POTHOLE DENSITY

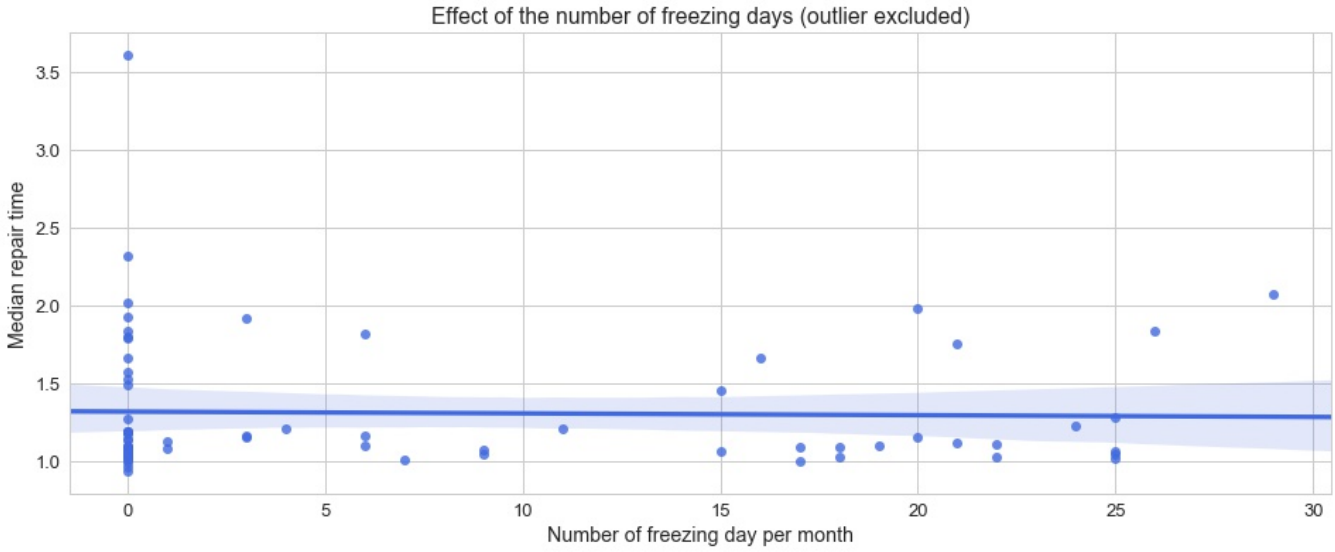


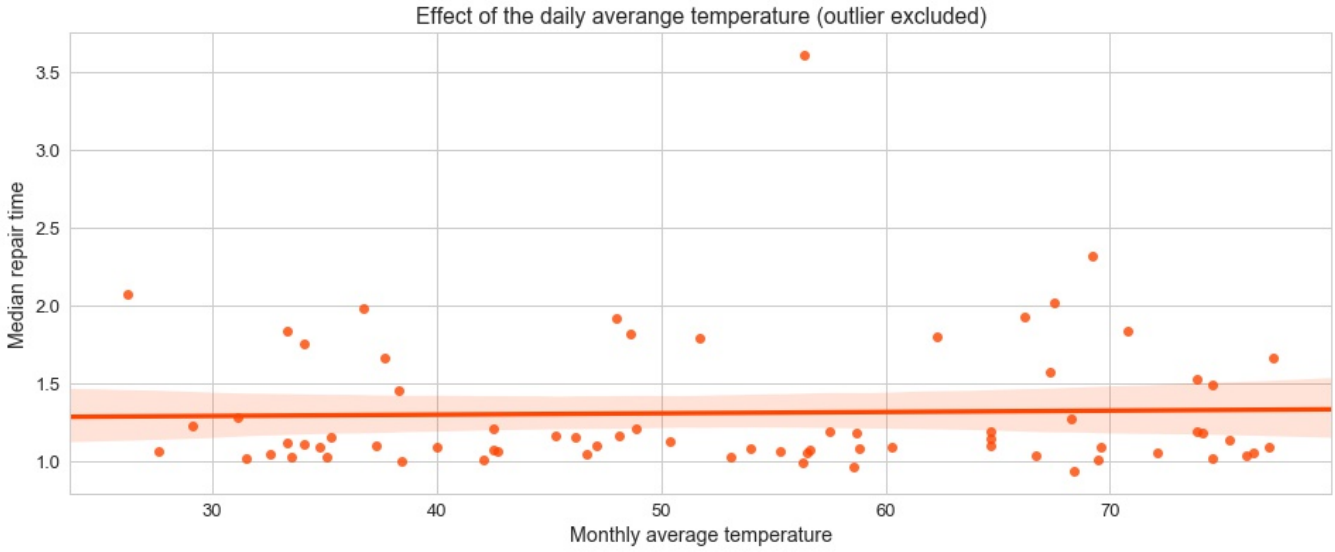
Appendix F: Regression Plots – Number of Claims vs. Neighborhood Data

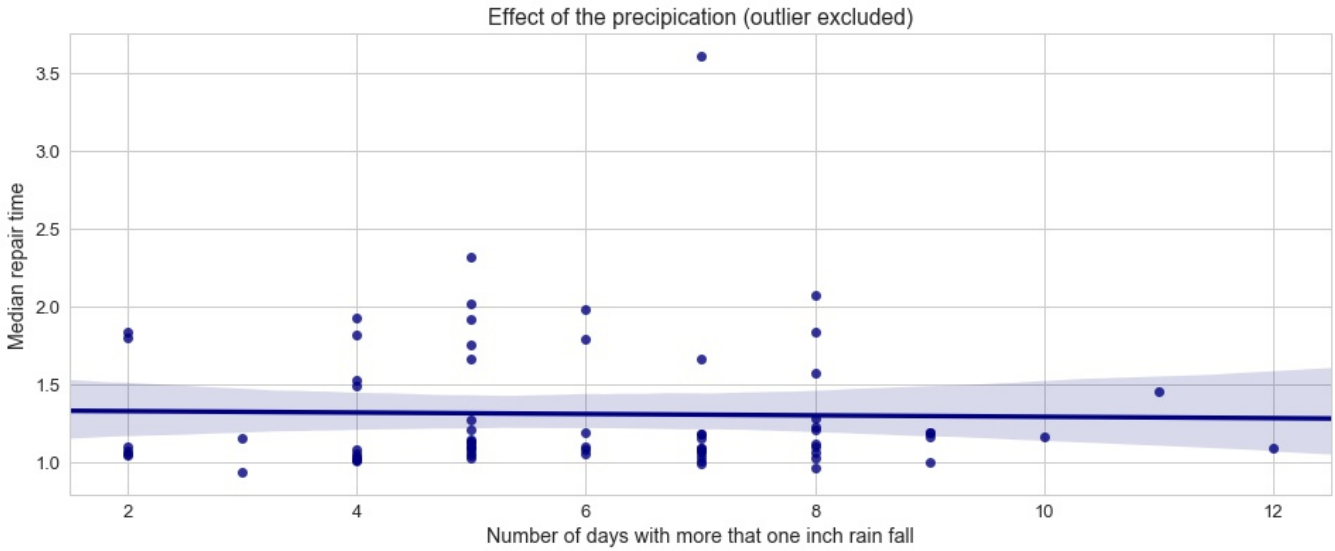


Appendix G: Regression Plots – Repair Time vs. Weather Data

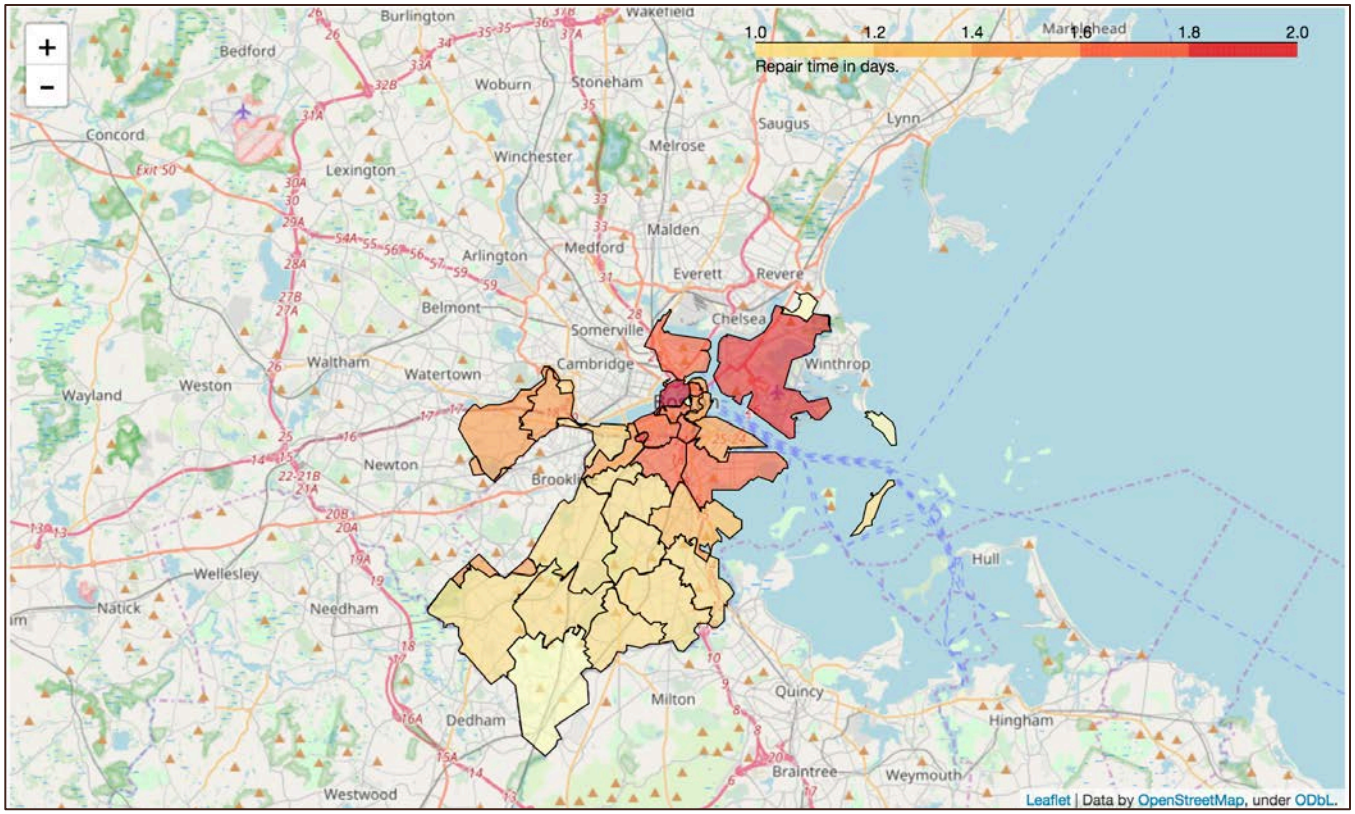




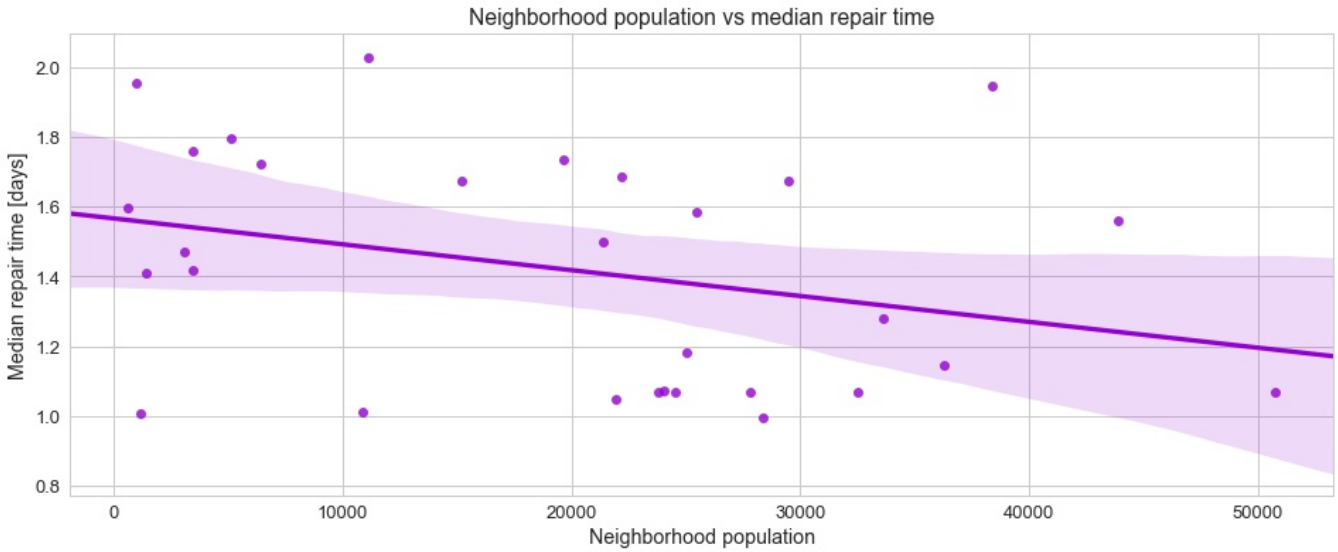

Effect of the daily averange temperature (outlier excluded)
Median repair time
3.5
3.0
2.5
2.0
1.5
1.0
30
40
50
60
70
Monthly average temperature


Effect of the precipication (outlier excluded)
Median repair time
3.5
3.0
2.5
2.0
1.5
1.0
2
4
6
8
10
12
Number of days with more that one inch rain fall

APPENDIX H: MAP – REPAIR TIME



APPENDIX I: REGRESSION PLOTS – REPAIR TIME VS NEIGHBORHOOD DATA



APPENDIX J: REGRESSION PLOTS – NUMBER OF CLAIMS VS REPAIR TIME