Final Assignment

Sumayya & Sumayah

6/24/2019

R·I·T Dubai
جامعة روتشستر للتكنولوجيا – نيويورك، أمريكا
Rochester Institute of Technology, NY, USA

Introduction

NYC311 Service Requests & Resolution Analysis is to Explore and analyze NYC311 Service requests (historical data sets) to understand diverse patterns, regular themes and trends, as well as community satisfaction levels and sentiment pulse (social network feeds) derived from resolution categories and timing.

When the New York City Department of Transportation (DOT) maintains roadways in a timely manner, few people thank them. But when conditions deteriorate, serious collisions can occur, resulting in loss of life.

Managing a city operation is a challenge. With the increase in data collection and reporting, NYC employees and residents have powerful tools to answer questions and identify problem spots. The result of this data analysis could impact funding, hiring, citywide initiatives, and project management at all levels. Comparing the roadway condition complaints to Census data has great potential for statistical discovery.

This Report will explore NYC 311 data and shows if there is a correlation between race and per capita roadway condition complaints across zip codes, and if there is a correlation between income and roadway condition complaints.

The Data Set

The Nyc311 dataset contains 9,124,937 objects and 52 variables for the period 2003-2015 that will be explored, filtered and analysed in this report. As well as it will give a preliminary description of the data by focusing on the complaints which sent to Department of Transport (DOT), specially the complaints related to road conditions such as pothole, Rough, pitched or cracked road, hummock, cave in and failed street repairs, as well as complaints related to line marking either after reaping or faded one.

Sources

New York City 311 Data Provided by the Khalil Darwish as 311 Service Requests from 2010 to Present

US Income Data US Census Factfinder Portal 2010-2014 American Community Survey 5-Year Estimates

US Population and Demographic Data US Census Factfinder Portal 2010-2014 American Community Survey 5-Year Estimates ACS DEMOGRAPHIC AND HOUSING ESTIMATES

Data Collection

The roadway data was collected by New York City 311 through phone calls and web forms from individuals contacting the city to complain or inquire about a roadway situation. Phone calls are entered directly into the 311 database by phone operators.

The US Census demographic and population data was collected by the US government. From the Census website: The American Community Survey (ACS) is a mandatory, ongoing statistical survey that samples a small percentage of the population every year.

Street Conditions

I this report, seven of the 17 complaints were selected. The five asphalt-related complaints were combined into the asphalt complaint category. The remaining two were the two options for marking complaints: faded and missing.

Brief of statistics:

The following is a summary of the statistics. For this report, each case is a zip code. Each case is a group of all 311 complaints for that zip code for the selected time period.

Statistic	Description
8,429,127	New York City population, estimated.
2010 to 2014	Selected date range of 311 complaints.
526,668	Street condition complaints in date range.
406,875	Street condition complaints with a reported zip code.
335,140	Asphalt complaints.
1,339	Missing markings complaints.
5,948	Faded markings complaints.
198	Number of zip codes in NYC, considered as cases.

The raw data set

The raw dataset consists of 9,124,937 rows and 52 columns and this is very large dataset.

For this project, the data was filtered as follows:

1. Only Department of Transportation data.
2. Only complaints which had a zip code as part of the complaint, to make data analysis possible.
3. Only complaints which were for a street condition.
4. Only the five primarily asphalt-related street conditions and two markings-related street conditions.

Asphalt conditions

- Pothole
- Rough, Pitted or Cracked Roads
- Hummock
- Cave-in
- Failed Street Repair

Marking conditions

- Line/Marking - After Repaving
- Line/Marking - Faded

To further simplify, all asphalt conditions were combined into one roadway surface complaint group, named "asphalt."

Missing values were deleted.

As a result, 17.6% of complaints send to Department of Transport (DOT), which consists of 1,601,604 complaints.

Tidying the dataset

It is important to check if subsetted data is tidy and put in the concentration the three rules which make a dataset tidy:

1) Each variable must have its own column.
2) Each observation must have its own row.
3) Each value must have its own cell.

If we validate the first two rules, we will automatically be validating the third rule.

We started by checking if there are any duplicate values in the 'Unique.Key' column in order to confirm that the data is at the level of a service request and the 'Unique.Key' is an identifier of the service request. Thus, there are no duplicates. Also, As there are missing values especially in location type and incident zip, it must be deleted form the dataset. In addition, as the period is important for analyzing, we separated it.

As a result, the data become more unique and has 406,875 complaints for street condition in all borough during the study period.

Joining Data

In order to gain more insights, we added some demographic data about the population number and income per zip code to the main 311 data. On the other hand, we add some population characteristic like: race.

dplyr provides a nice and convenient way to combine datasets. A join with dplyr adds variables to the right of the original dataset. The beauty is dplyr is that it handles four types of joins:

1. Left_join()
2. right_join()
3. inner_join()
4. full_join()

Both datasets were merged and transformed to display the required data by using left_join R function. Therefore, we get 198 zip codes and 12 variables

---

Data Dictionary

The following is attribute names and descriptions for the new combined dataset.

Table 2: Data Dictionary

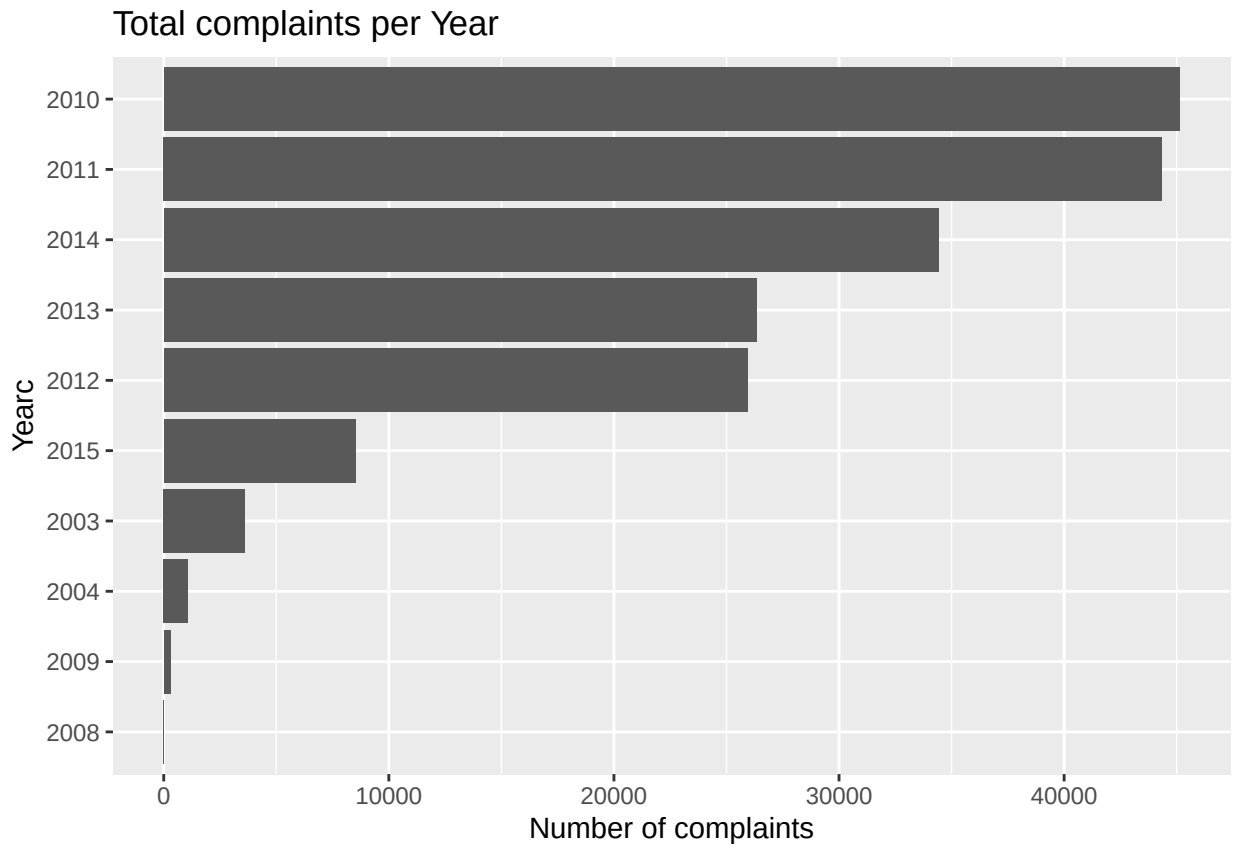
Variable	Description
zip1	zip code with ZIP_ prefix
borough	borough of NYC
asphalt	number of asphalt condition complaints
missing	number of missing markings complaints
faded	number of faded markings complaints
pcomp	complaints per capita for zip code
pcompa	asphalt complaints per capita for zip code
pcompm	missing markings complaints per capita for zip code
pcompf	faded markings complaints per capita for zip code
complaints	total number of complaints
income	income for zip code
pop	population for zip code
pwhite	percent white population for zip code
pblack	percent black population for zip code

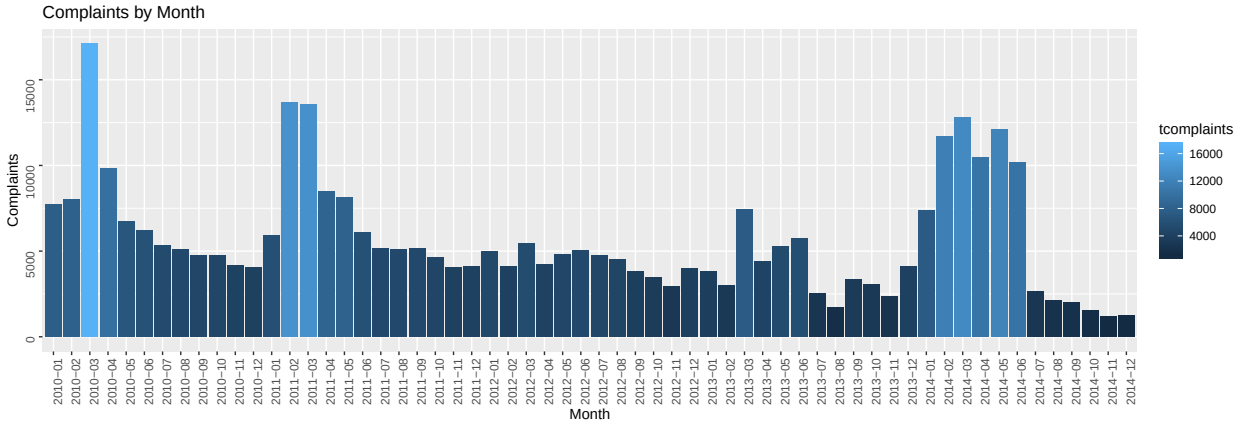
Analysis and Findings

In this section, we are going to show our findings by analyzing the narrowed data.

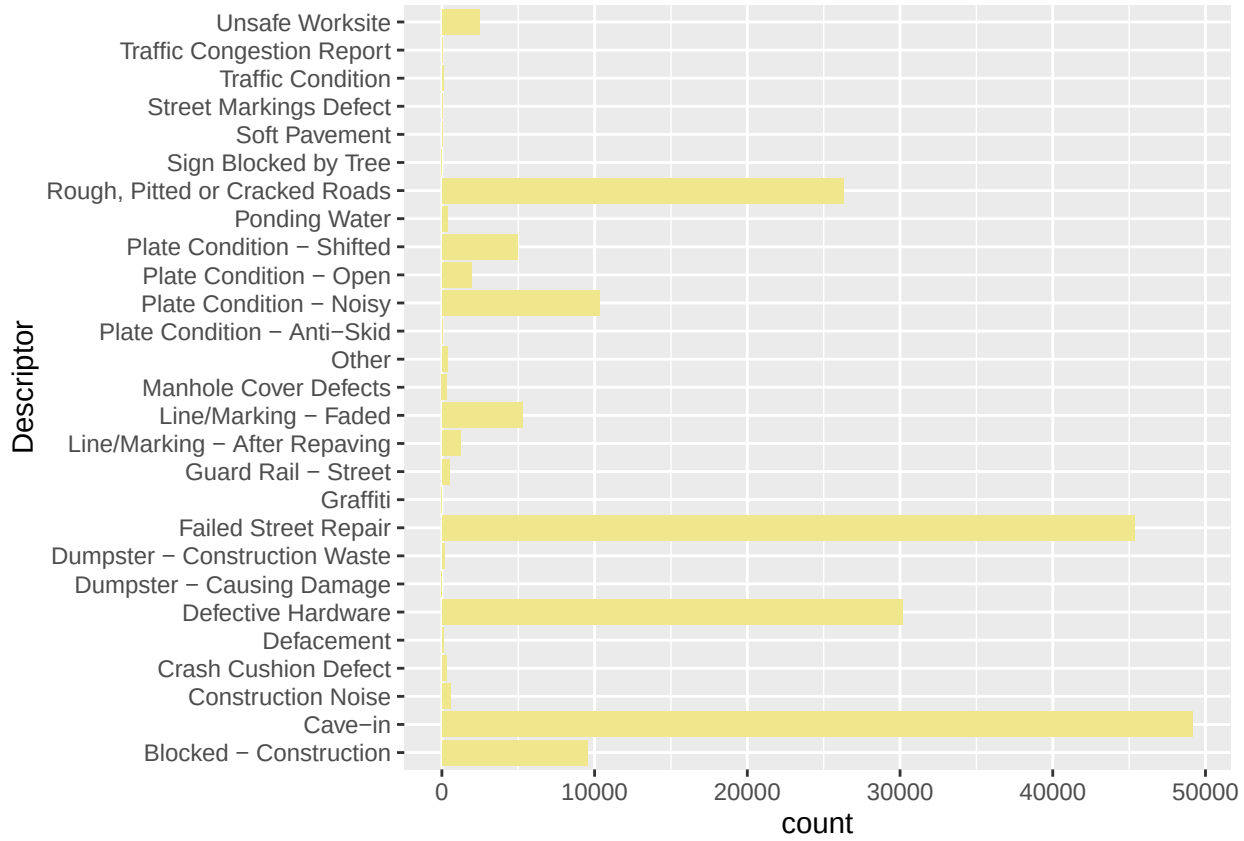
Complaints by time

The below charts show that 2010 has the highest number of complaints. While in monthly base, the months of February until May have the highest number of complaints.





Most frequent complaint type

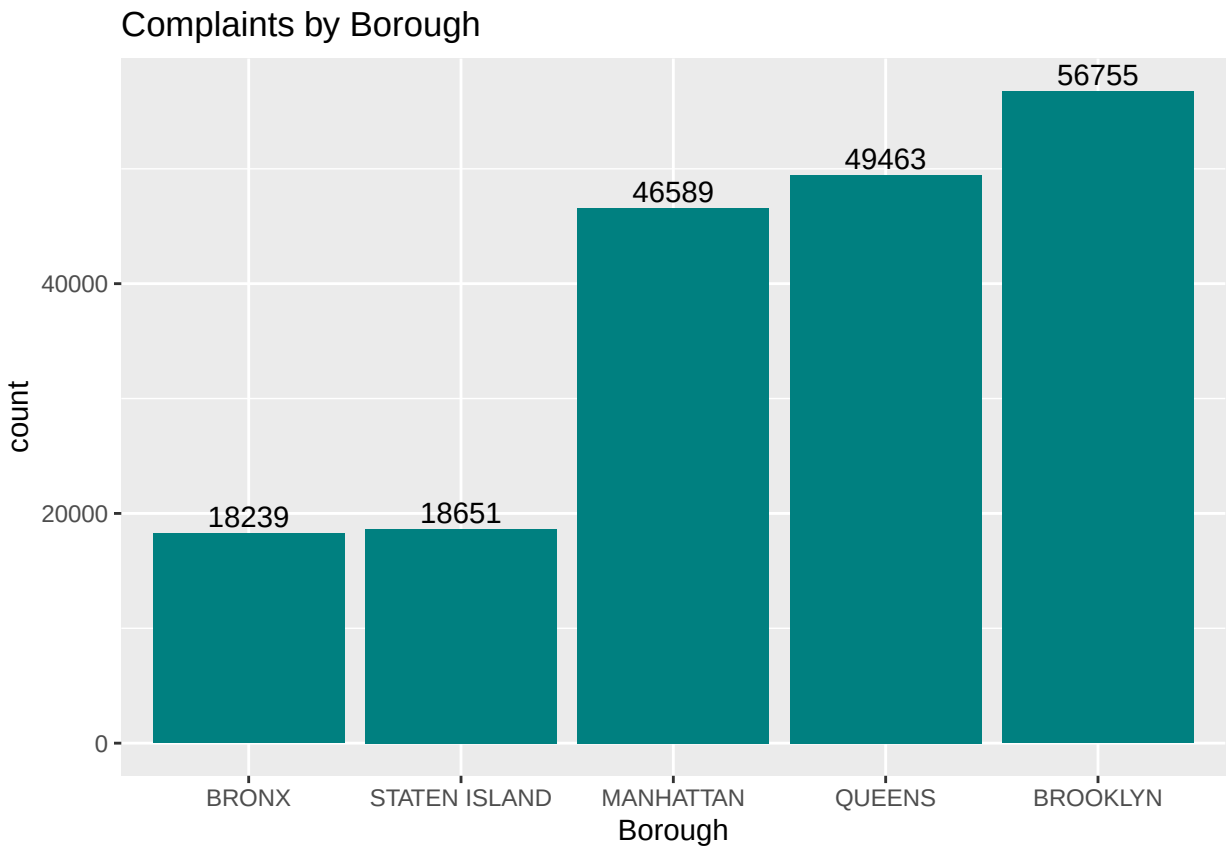


The plot above shows the most frequent complaint type related to street condition. Therefore, the most type that people are complained about is related to Asphalt conditions which is: Cave-in.

Complaints by Borough

```
## <ggproto object: Class CoordFlip, CoordCartesian, Coord, gg>
##     aspect: function
##     backtransform_range: function
##     clip: on
##     default: FALSE
##     distance: function
##     expand: TRUE
##     is_free: function
##     is_linear: function
##     labels: function
##     limits: list
##     modify_scales: function
##     range: function
##     render_axis_h: function
##     render_axis_v: function
##     render_bg: function
##     render_fg: function
##     setup_data: function
##     setup_layout: function
```

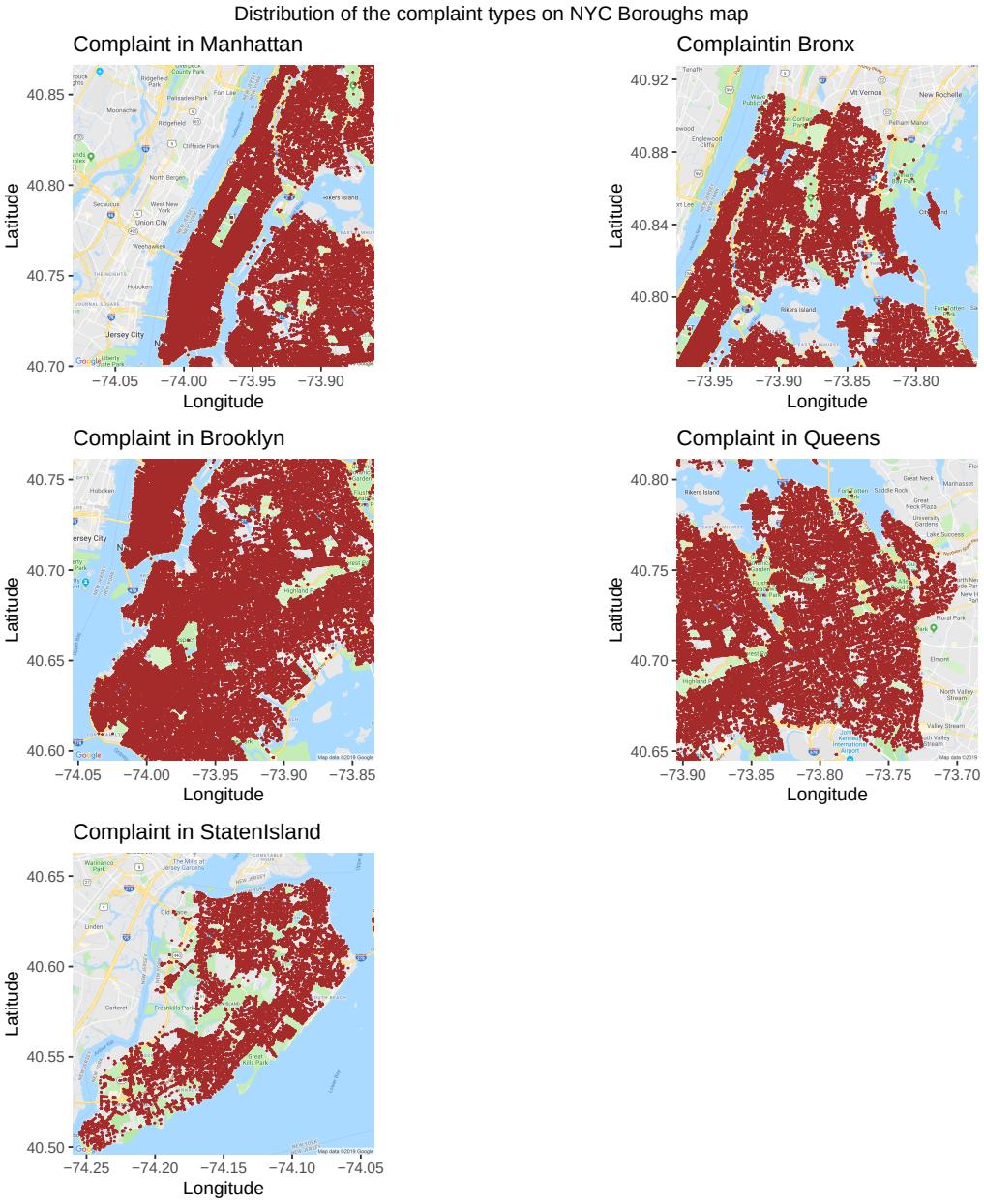
```
##     setup_panel_params: function
##     setup_params: function
##     transform: function
##     super:  <ggproto object: Class CoordFlip, CoordCartesian, Coord, gg>
```



From the above plot, it is clear that the highest number of complaints is from "BROOKLYN", then "QUEENS" and "MANHATTAN".

Distribution of the complaint types on NYC Boroughs map

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.78306,-73.97125&zoom=12&size=640x6

## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.84478,-73.86483&zoom=12&size=640x6

## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.67818,-73.94416&zoom=12&size=640x6

## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.72822,-73.79485&zoom=12&size=640x6

## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.57953,-74.1502&zoom=12&size=640x64

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

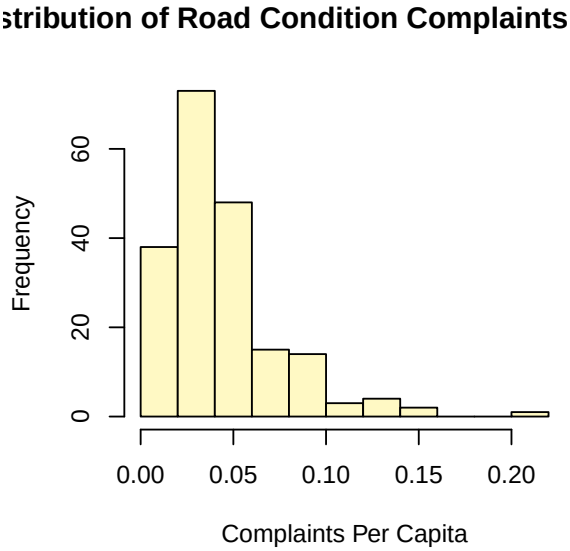


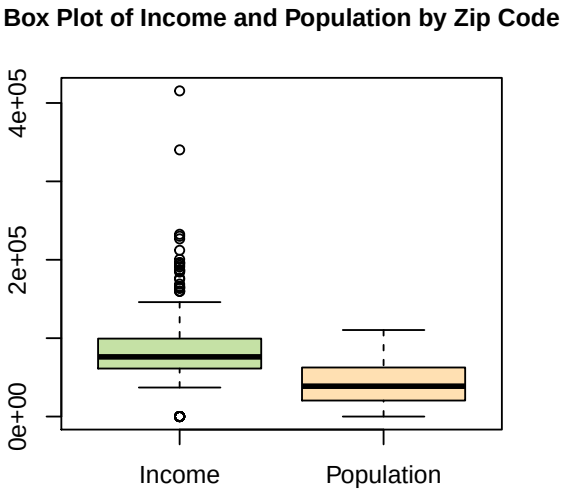
In general, complaints occur most in BROOKLYN then QUEENS and MANHATTAN.

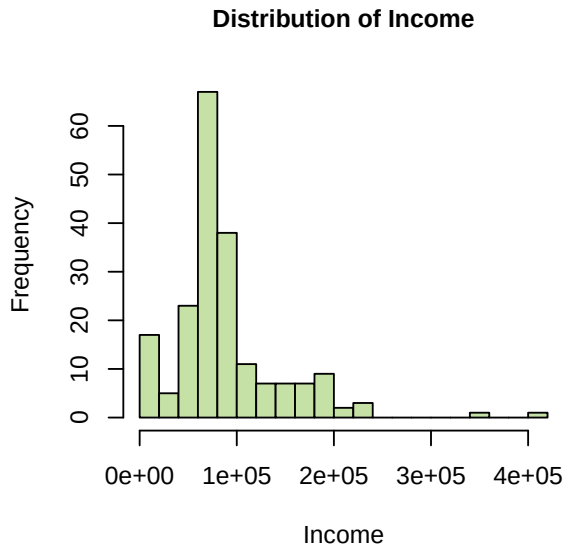
Analysing the joined data.

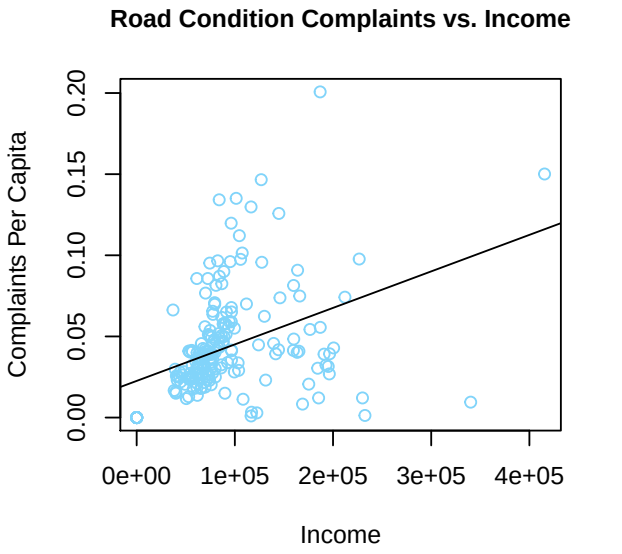
In this section, we will show some plots and charts that will be discused in the following section.

Distribution of Road Condition Complaints Per Capita and income



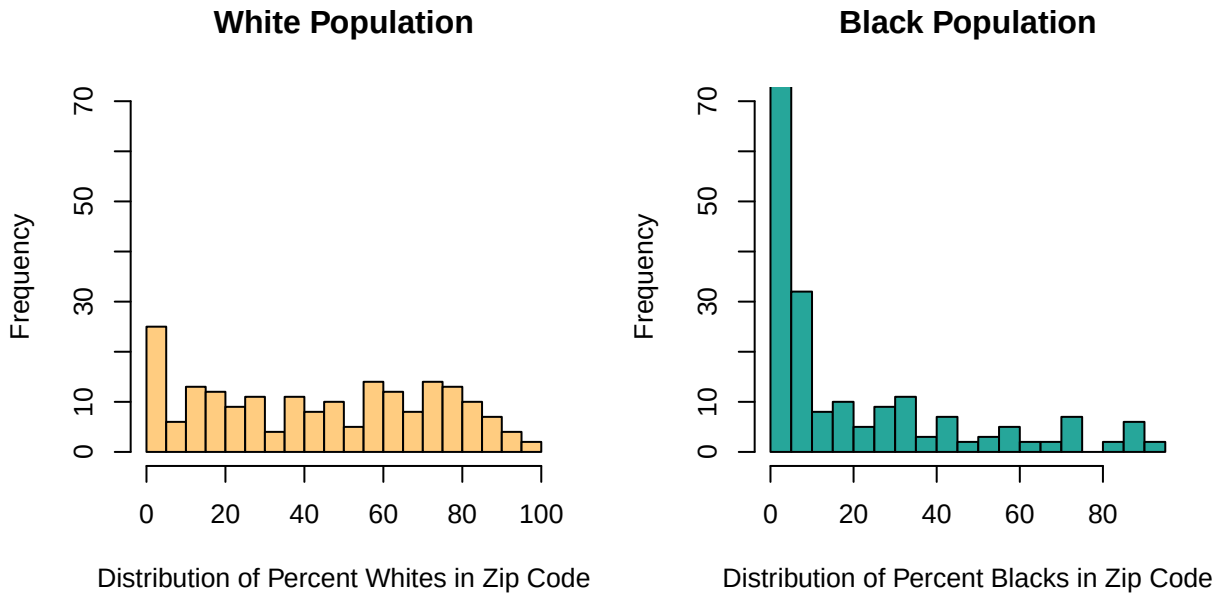


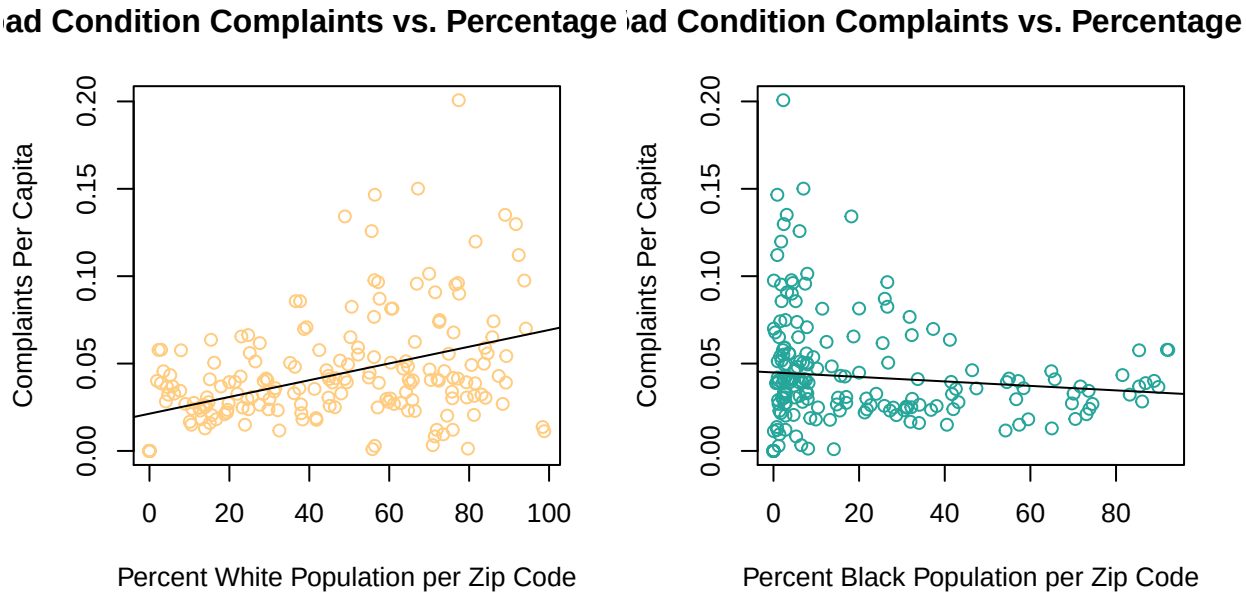




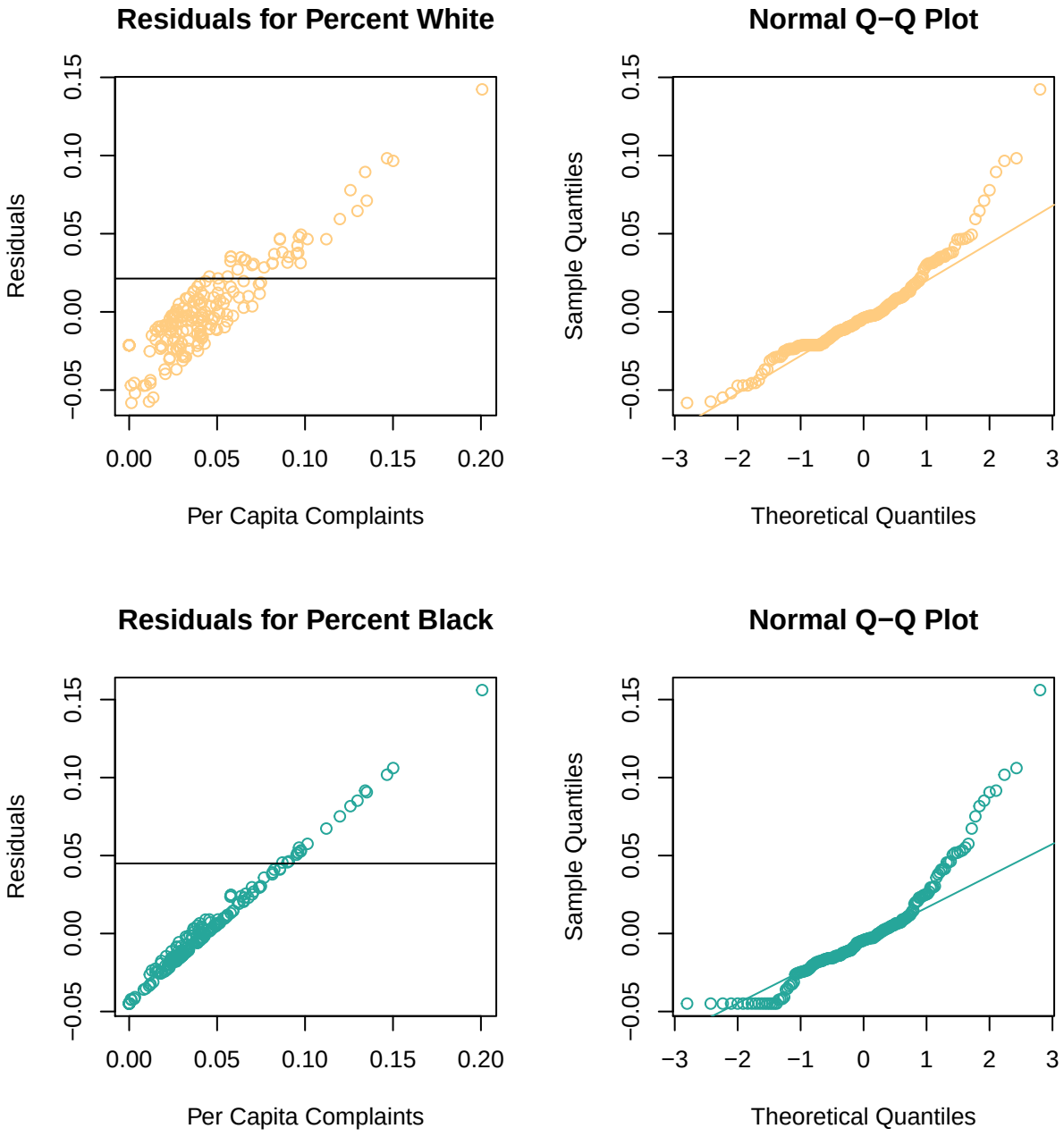
It is clear that there is different distribution between the income and population, as well, there is also outlier in the income plot per zip code.

Correlation for Percent White people and Black people

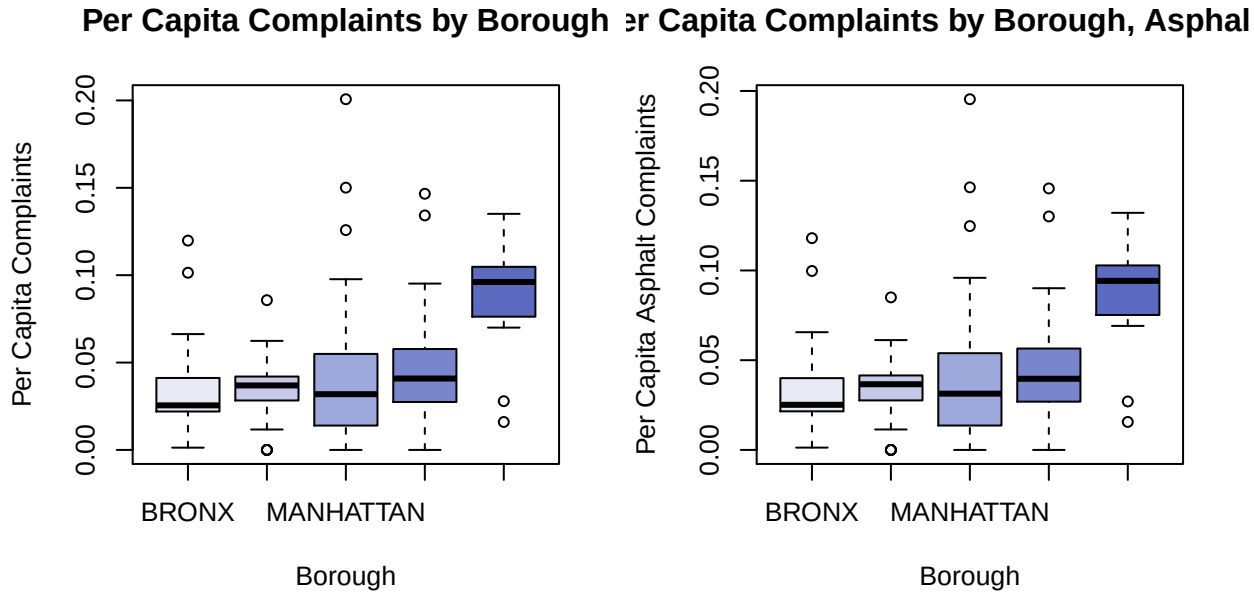


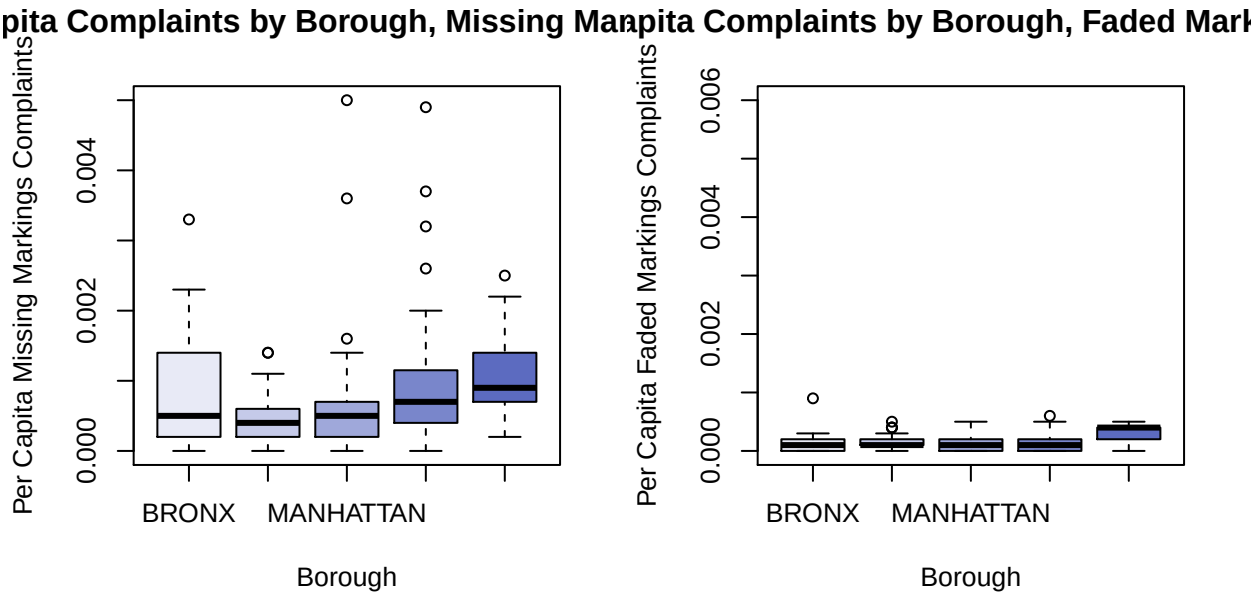


Residuals for Percent White and Black



Per Capita Complaints by Borough





Moreover, means by borough as the following:

```
## combined_data1$borough: BRONX
## [1] 0.03488846
## ------------------------------------------------------
## combined_data1$borough: BROOKLYN
## [1] 0.03510488
## ------------------------------------------------------
```

```
## combined_data1$borough: MANHATTAN
## [1] 0.0408125
## ------------------------------------------------------------
## combined_data1$borough: QUEENS
## [1] 0.04375625
## ------------------------------------------------------------
## combined_data1$borough: STATEN ISLAND
## [1] 0.08669091
```

Results and Outliers

There were several interesting outliers. The graph of per capita income had many outliers in the upper ranges, expected for New York City.

The box plots for borough response time had more outliers in Manhattan at the upper range of per capita complaints. This could indicate that Manhattan has higher density in general, or a dedicated group of concerned citizens.

The per capita complaints for Staten Island were significantly greater than other boroughs. Staten Island has historically had issues with road conditions. They are geographically separated from the other four boroughs and have many sprawled suburban areas which are difficult to address in a concise and timely manner.

There were high income outliers for zip codes. In one case, a Westchester County zip code was being included in the NYC zip code list, possibly due to a shared border street with the Bronx. (ZIP_10803, mean income 232513, Bronx.)

A high income outlier for Queens was a zip code in Long Island city across the river from Manhattan, where dozens of new skyscraper condominiums have risen in recent years on the site of previously industrial neighborhoods. (ZIP_11109, mean income 168940, Queens.)

A high outlier for per capita complaints was ZIP_10004 in Lower Manhattan. Interesting, this zip code includes the NYC DOT headquarters, which could potentially be skewing the data due to employees using 311 web forms as part of their job or interest.

Some zip codes in the US Census dataset showed as 0 income or NA income. Researching these zip codes identified these as parks, airports, large office buildings, and in one case the former World Trade Center zip code, which has been discontinued.

Testing the Hypothesis

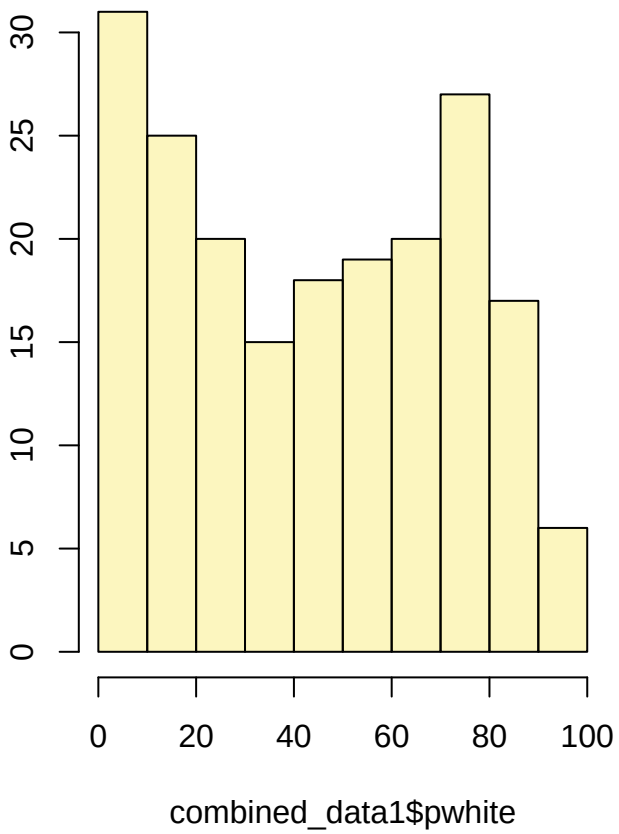
Satisfying conditions for inference

The conditions for inference do appear to be satisfied. The sample size is greater than 30; the datasets follow a uni-modal normal distribution; the samples are random.

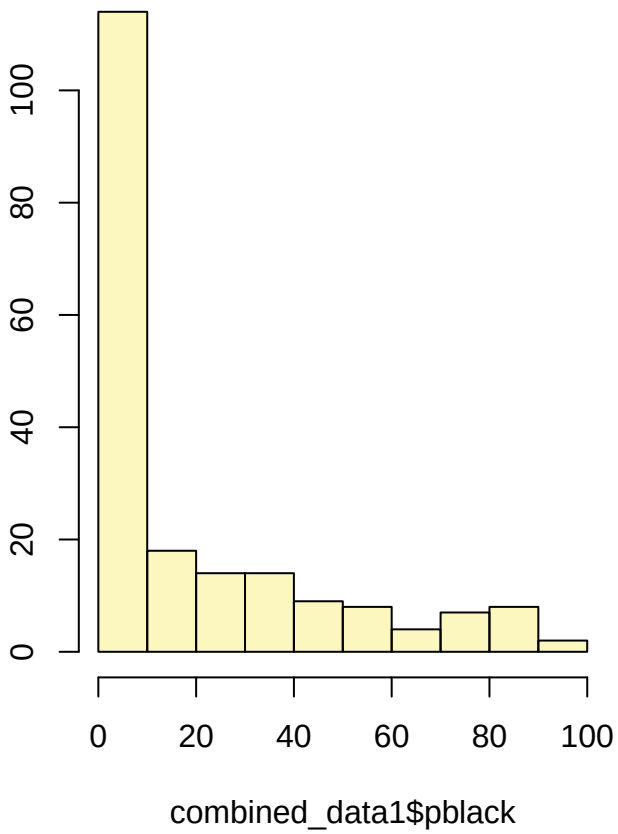
Inference

Summary statistics:

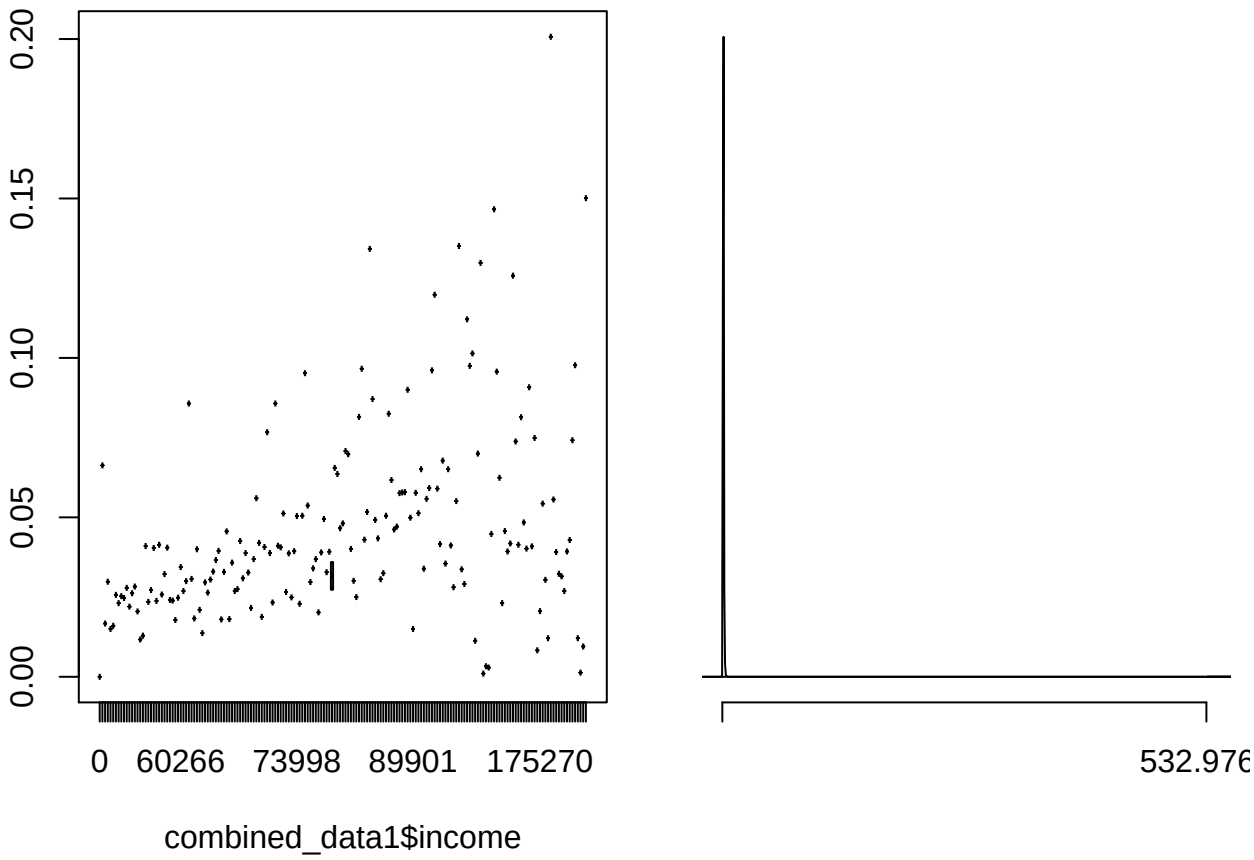
```
## Single mean
## Summary statistics:
```




```
## mean = 43.9035 ;  sd = 28.8732 ;  n = 198
## Standard error = 2.0519
## 95 % Confidence interval = ( 39.8818 , 47.9252 )

## Single mean
## Summary statistics:
```




```
## mean = 19.9793 ;  sd = 25.1686 ;  n = 198
## Standard error = 1.7887
## 95 % Confidence interval = ( 16.4736 , 23.485 )
```



Conclusion

Summary

New York City 311 is a free public service that allows individuals to register complaints or inquiries on city conditions. Comparing the roadway condition complaints to Census data has great potential for statistical discovery.

Through this report, we can see the characteristics of Street conditions complaints in New York City and it allows to understand it in New York City with multiple perspectives.

The initial question was to determine if a correlation exists between road condition complaints and income or race. Using plots, linear modeling, and statistical analysis, both race and income did appear to be correlated with per capita complaints.

The validity of the data was indicated by summary statistics for the chosen variables, which showed low p-values less than 0.05, as well as normality and qqplots for residuals.

Further data exploration beyond the original scope of the hypothesis resulted in the discovery of significant correlations in complaints by borough, as well as clear seasonal trends in complaint volume. Staten Island has a greater percentage of complaints per capita. Early spring months have the greatest volume of complaints, which is logical as they are after snow plow contact, salt treatment, and snow melting, and when individuals increase outdoor activities.

Insights

Complaints with blank zip codes were excluded to make data analysis possible. However, excluding complaints where the complainant provided voluntary information omit disenfranchised neighborhoods or individuals. Median income might have been a better indicator than mean since high incomes in large cities skew the average higher.

Missing and faded markings might have identical meaning to 311 callers, even though these are significantly different for DOT processes. For operational purposes, we was looking for accurate data since these are handled through different processes. Grouping race percentage by zip code may not be a reliable indicator of the behavior of a certain race. For example, blacks in majority-white zip codes may be more likely to complain to 311 than blacks in majority-black zip codes, but some of this level of analysis is lost in grouping.