Staying Safe in NYC: A Crime Analysis of New York City

Van Steenbergen Nicolaas

Fish Alex(acfish)

Visiting the Big Apple can be a thrilling adventure, but also a potentially dangerous one. Knowing where crime takes place and what to look out for can keep you safe in unfamiliar settings. By doing an exploratory analysis of crime in New York City we learn how to safely navigate all the boroughs and neighborhoods. Using data provided by the NYPD and US government, we examine correlations between a variety of factors such as: types of crime, time, location, and weather. We created an customized map that shades areas of New York City different colors based on its safety using the factors listed above. This product can be utilized by tourists to get a data-driven visual representation of which areas are the safest.

1. Intro

Having the ability to travel is becoming cheaper every year which means more lost and unsure tourists in new cities. It is known that tourists are a common target for criminals because they are unfamiliar with the new environment[1]. If tourists have in depth knowledge of criminal history in the city they are traveling to, it will help them avoid dangerous areas they might wander into otherwise. This analysis seeks to enlighten travelers where crime occurs along with the type and severity. By using crime complaint reports provided by NYPD and historical weather provided by the National Centers for Environmental Information we are able to identify the most dangerous seasons and locations in New York City.

1.1 Document Reproducibility

This report is arranged using the **R** [R-base][2] package *knitr* [R-knitr][3]. This project may be imported into the RStudio environment and compiled by researchers wishing to reproduce this work for future data sets.

2. The Data

All data comes from open source government agencies. The NYPD Complaints data set[4] is provided by the NYPD and the historical weather data set[5] is provided by the National Centers for Environmental Information. The neighborhood data set[6] is provided by the New York department of city planning. The scope of the data sets include every report of crime, daily summaries of weather, and population of specific neighborhoods. In the raw format there exists a lot of unnecessary variables for exploring crime and the location of crime that we needed to clean.

2.1 Obtaining the Data

The weather and neighborhood data sets come from government agencies while the New York City crime data is provided by the New York Police Department; all data sources are publicly available. The data was downloadable by csv format so reading the data into R did not provide major issues. One precaution we had to take was reading the data before filtering the year we wanted because the files were so large. Once we had our filtered data we were able to do a quick analysis of each set to learn how to best approach the cleaning process.

2.1.1 Weather Data

Our approach with the weather data was to first understand the structure of the data set and only use the necessary columns. With the help of plyr[8] and dplyr[9], cleaning the data frame did not provide any challenges, but one modification we had to make was adding an average temperature column. The data set provided the maximum and minimum temperatures, so by finding the average of them we were able to make a new column with averages. We also had to only pull data from 2015 and the data set did not include a date-time column that could be easily understood by R so we used stringr[10] commands and regular expressions to grab all dates from the 2015 year.

2.1.2 NYPD Complaint Data

The largest data set was the NYPD Complaint data; because of its size we used a regular expressions to filter the date by the year we wanted. After narrowing it down to one year, the data set was more manageable so we could start a structural analysis. Having to decode the coded column names used by the NYPD took some time, despite the documentation provided. The data set had a lot of columns that were not useful for our analysis. After clearing them out, we were left with a complete data set of only the variables we needed to analyze the time, type, and location of crime. It was fortunate that the data set included the borough the complaint came from, so we did not have to track its location manually and were able to use the column extensively in our analysis.

2.1.3 Map Data

Obtaining and cleaning the data for the map was the most tedious step due to the lack of neighborhood names in the original data set. While we were given the borough, latitude, and longitude; we needed to use a variety of mapping packages to plot each latitude and longitude and assign it an accurate neighborhood. Packages such as: leaflet[11], sp[12], ggmap[13], rgeos[14], maptools[15], broom[16], httr[17], rgdal[18], spData[19], and tigris[20] all contributed to the pinpointing of crime locations in neighborhoods. The complexity of using latitude and longitude to be assigned within an area proved to be our biggest hurdle. Once we were able to get a neighborhood associated with coordinates we were able to merge the original data set with the created one containing the neighborhood names. To display the static maps we were able to use ggplot2[21] which resulted in an understandable and accurate map.

2.2 Short Comings

With the NYC Complaint crime, the date the complaint was made did not always exist so we had to make the assumption that the complaint date was within 24 hours of the NYPD reporting date. This allows us to take the complaint time and pair it with report date so we have the near-exact time and location of every complaint reported.

Along with missing data, the variety of crime names was overwhelming. We took notice that many of the crimes fell under similar categories so we had to manually group them into larger categories to get a more manageable set of crimes. While it was not technically difficult, the time it took and precision required was a temporary, but necessary short coming.

As mentioned in section 2.1.3, the tedious task of getting neighborhood names associated with latitude and longitude was our biggest issue. Despite accomplishing the task of connecting coordinate with neighborhoods merging the two data frames gave us some struggles. While knowing to merge based on latitude and longitude of both sets, the files were so large that our computers could not handle the merge. It wasn't until we used the data.table[22] package to turn our data frames into data tables, which took the merge time from indefinite to instant.

3. Methods

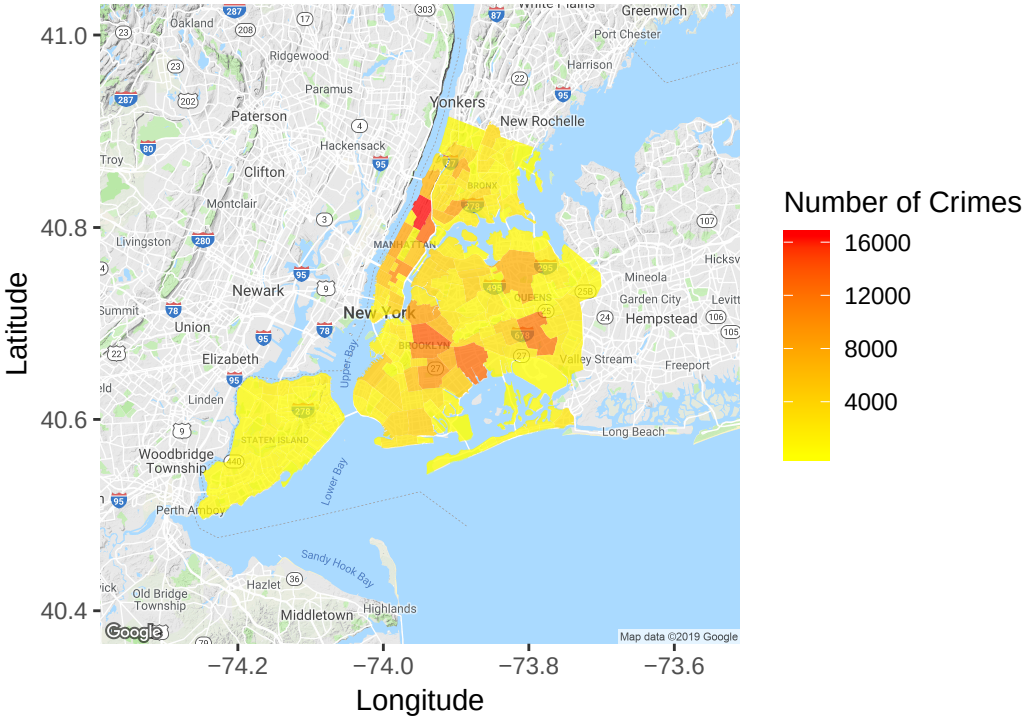
Having access complete access NYPD Complaint Historical data we are able to provide the exact latitude and longitude location of each crime as well as the type of crime. This information is the basis of our data product and is the main focus of our analysis.

3.1 Data Product

Our data product is a historical trends of the areas where crime happens, and what severity of crime happens in those areas. An interactive shiny[23] version of *figure 1* has been created to show the crime dense regions, and can be faceted by severity of the crime. This reflects the main goal of the analysis, which is help provide a safer experience when visiting NYC. Having an accurate and understandable map is the most important piece of our data product, so making sure the map is of the highest quality was a priority. If the traveler can check historical trends of the area they want to visit, they can have better judgement of where and when to travel.

Figure 1

Total Crimes per Neighborhood in NYC in 2015



4. Initial Findings

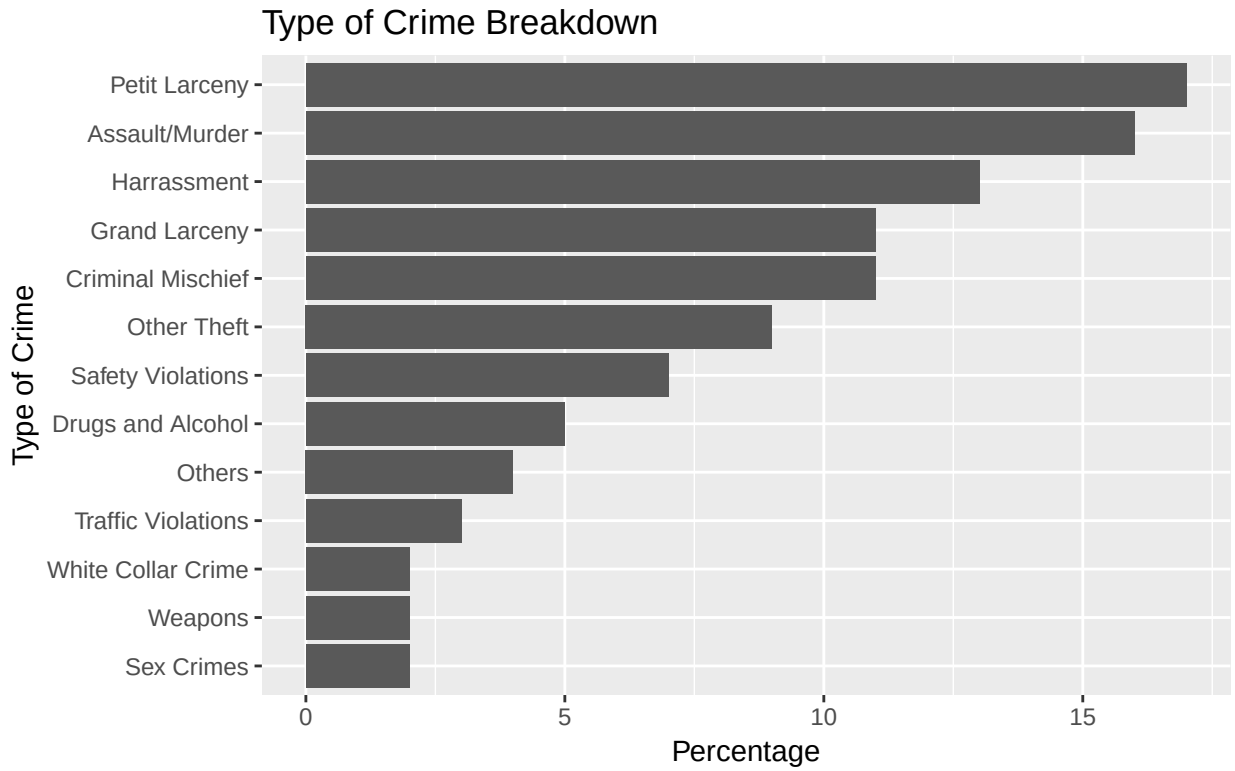
An initial analysis gave insight to what type of crime occurs. As seen in *table 1* , there are only 15 major types of crime that make up 96% of all the crime in NYC. We are able to immediately understand that Petit Larceny, which is low-level theft, is the most common crime reported. We can infer that tourists make up a portion of the victims, because they are commonly targeted by pick pocketers.

Table 1: Crime Breakdown

Crime	Percentage
Petit Larceny	17
Assault/Murder	16
Harrassment	13
Criminal Mischief	11
Grand Larceny	11
Other Theft	9
Safety Violations	7
Drugs and Alcohol	5
Others	4
Traffic Violations	3
Weapons	2
White Collar Crime	2
Sex Crimes	2

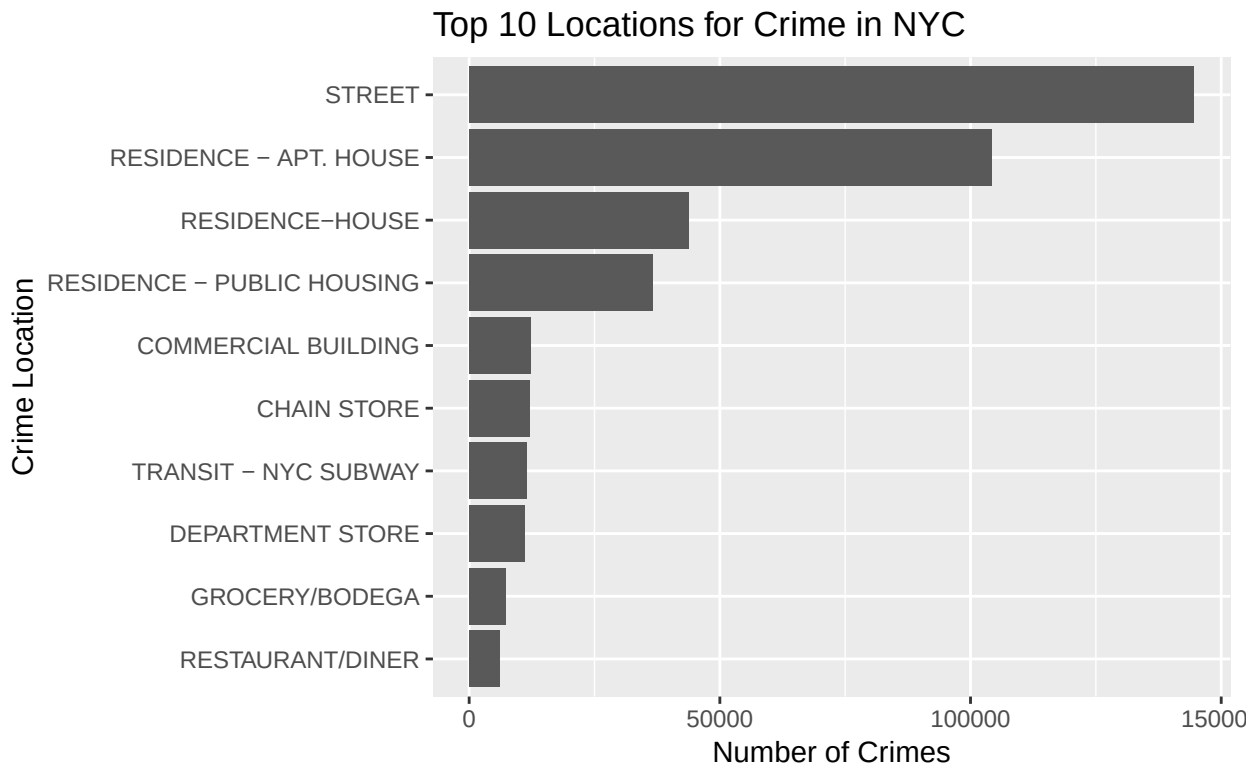
These categories can be understood better with visual representations of the crime breakdown as seen in *figure 2*.

Figure 2:



Knowing where within the neighborhoods it takes place can add another layer of protection. As seen in *figure 3*, a lot of the crime in NYC happens on the street. So if a tourist is wanting to maximize their safety, taking a cab whenever possible is the best option.

Figure 3

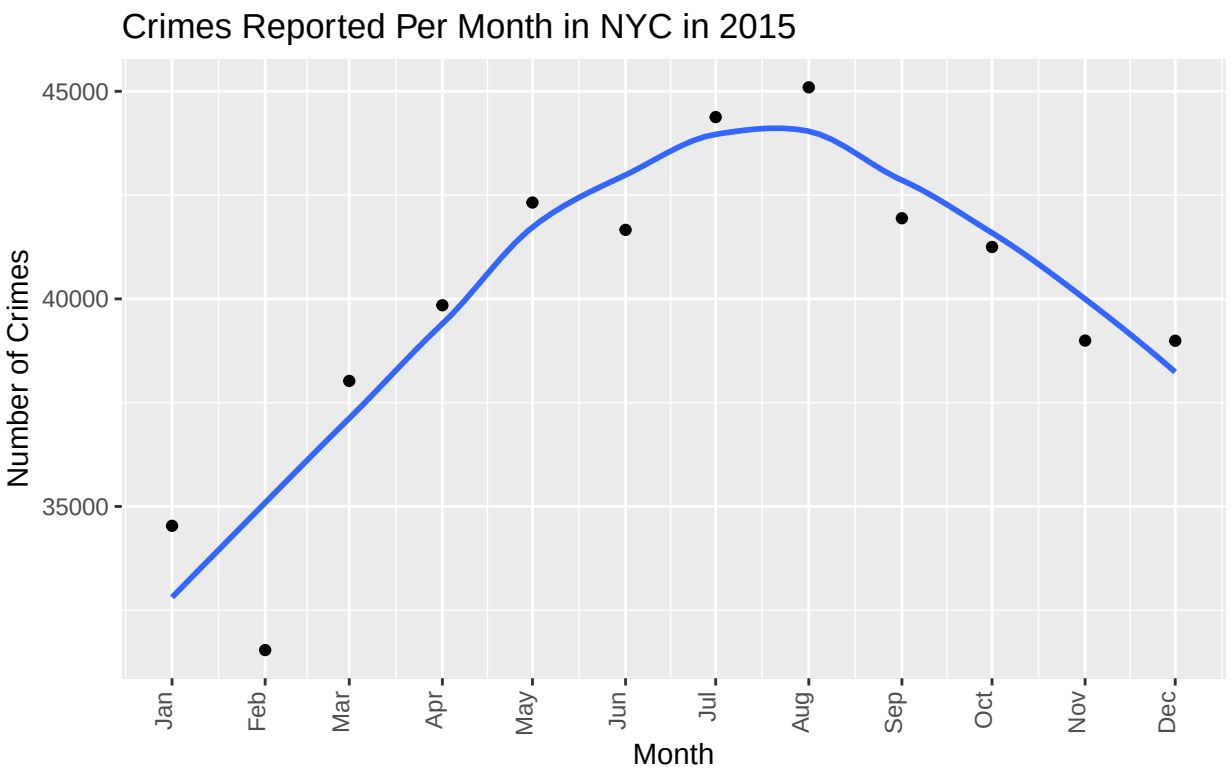


4.1 Frequency of Crime

4.1.1 Total Crimes by Month

In *figure 4* below, it shows that the summer months NYC have more crime, so if tourists would like to visit in the summer, they will be at a higher risk that time of the year. If traveling during the summer is their only option, visiting in May and June would be safer than visiting in July and August.

Figure 4

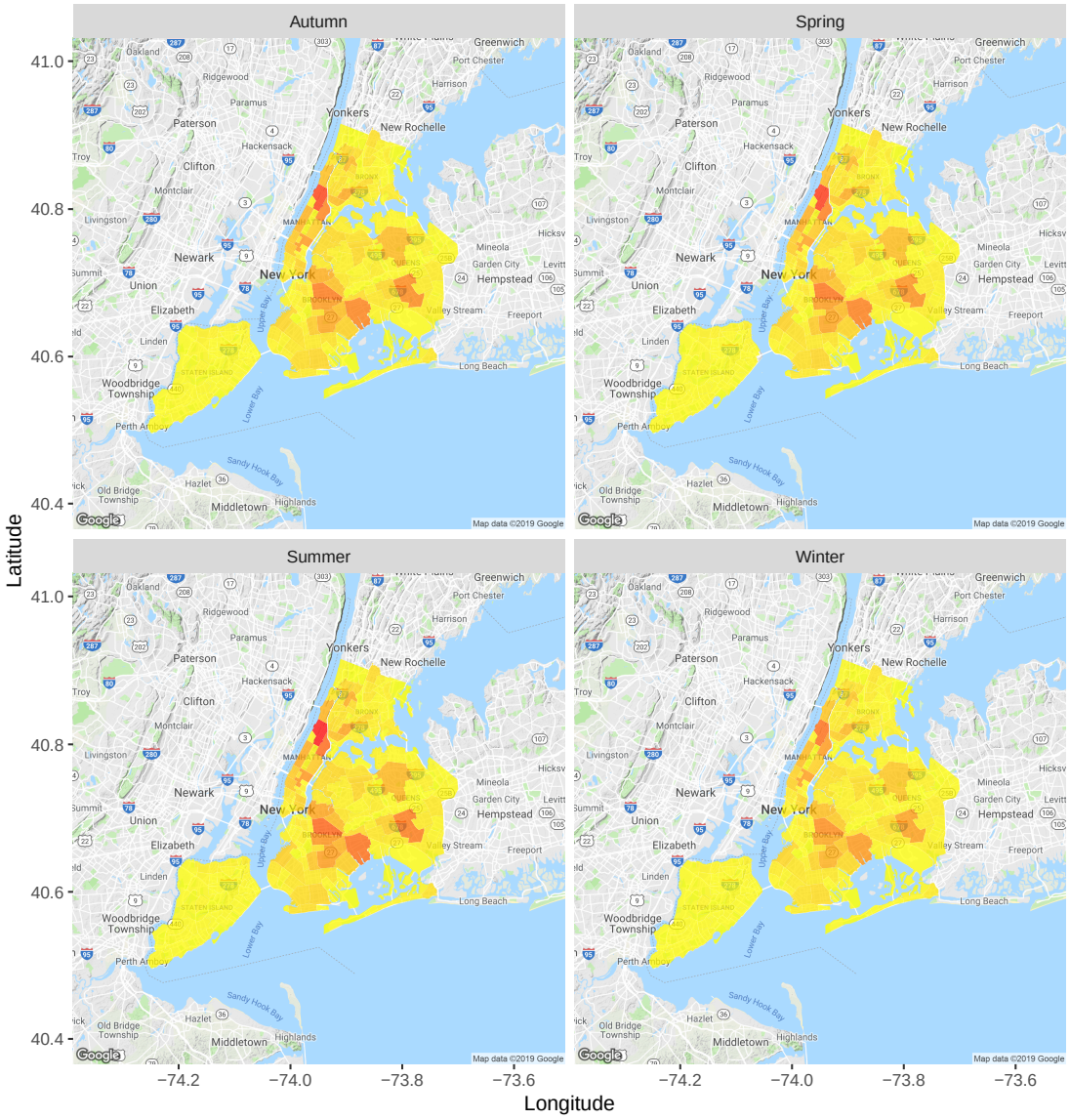


4.1.2 Total Crime by Season

Taking a wider look at the seasons we can see in *figure 5* that it stays consistent throughout the year, that is, there are not crime 'hot spots' that follow a seasonal pattern. The neighborhoods with the highest crime rates remain consistently dangerous with exception of winter, where there is a noticeable decrease in certain Manhattan neighborhoods. With *figure 4* and *figure 5* we can start to notice that there is a consistent fluctuation of crime throughout the year.

Figure 5

Total crimes per Neighborhood

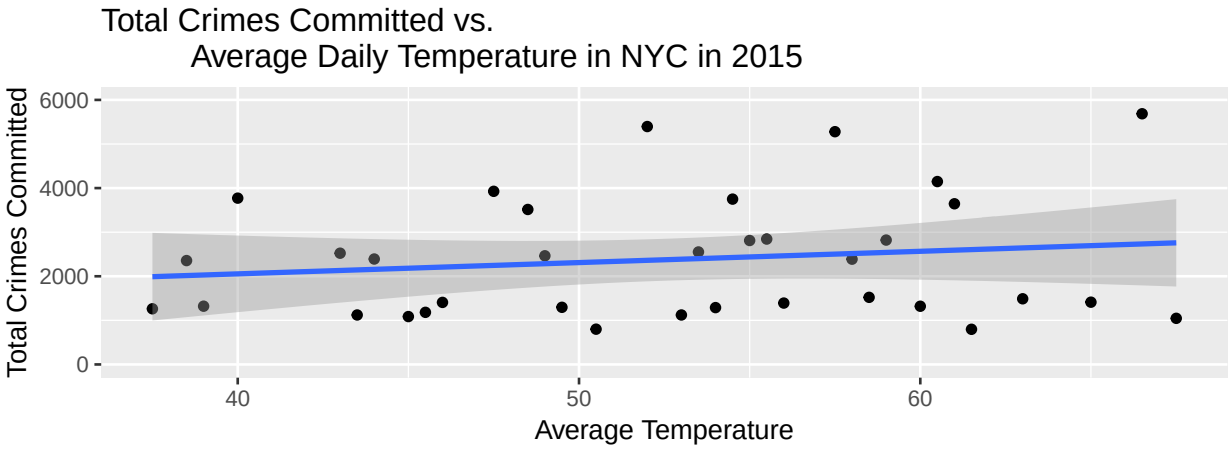


4.2 Crime Rate and Temperature

We also looked into how average temperature had an affect in the rate of crime with the assumption less people are out when the weather is colder. Literature from Chicago Tribune[7] backs up the assumption that weather does have an affect on crime.

Along with *figure 4* and *figure 5*, we take a look at *figure 6* to understand that the average temperature has an affect on overall crime in New York City. While subtle, it can be seen that the warmer average temperature leads to a higher rate of crime. We agree with the Chicago Tribune and assume this is because more people are out during the summer so the more crowded places are, the more crime there will be.

Figure 6



4.3 Severity of Crime

We investigated the severity of crime because violations and misdemeanors are not as dangerous as felonies, and might artificially inflate the overall crime but not increase overall danger. While all three are still reported complaints, it will make a difference for travelers visiting to know if the area is more common for felonies, violations, or misdemeanors.

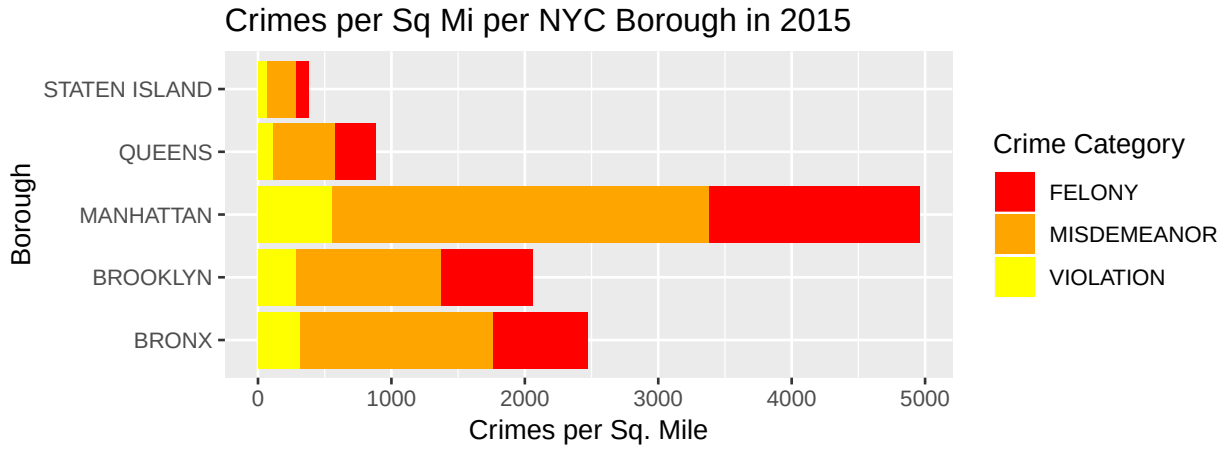
While all crime is dangerous it is important to note the different severity reported. Within the complaints data set there were three different types of reports according to the NYPD: Felony, Misdemeanor, and Violation. Notice that in all of the boroughs, misdemeanors made up the majority of the complaints recorded. We can see that Brooklyn had the largest percentage of felonies while Staten island's percentage of felonies are smaller comparatively.

Our data product gives the option to facet all of NYC by severity of crime so it is possible to compare the neighborhood safety based on severity.

4.3.1 Severity of Crime Per Square Mile

It can be seen in *figure 7*, that when based off of crimes per square mile, Manhattan is the most crime dense area of all the boroughs. This is due to Manhattan having the smallest area of all five boroughs. It is important to note that this does not imply that the most crime happens in Manhattan, just that it is the most crime dense borough.

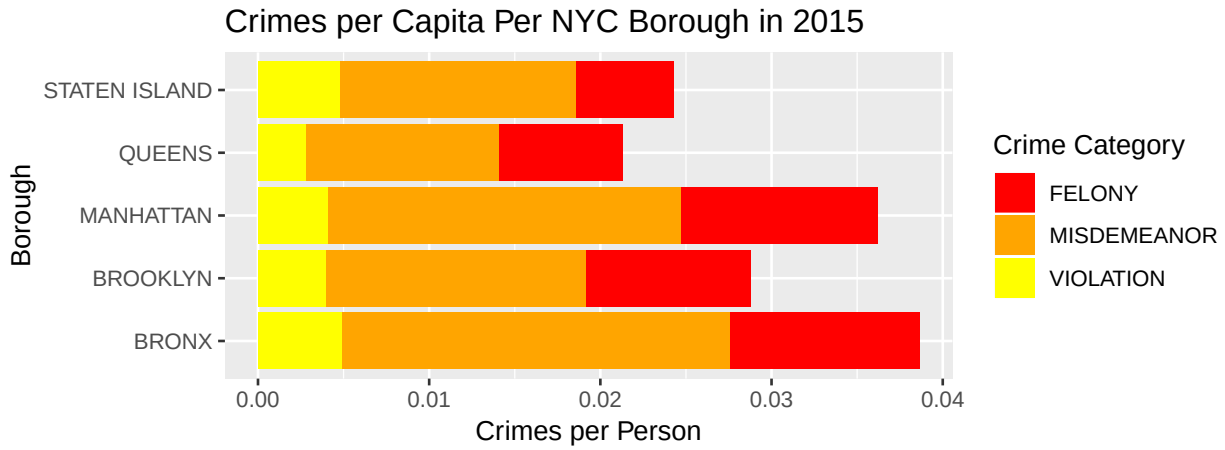
Figure 7



4.3.2 Severity of Crime Per Capita

In contrast with *figure 7*, *figure 8* looks at crime per capita. This will give a better estimate on how much crime happens in each borough, rather than the density of the crime. The biggest differences are seen with Staten Island and Manhattan; where Staten Island has a huge increase and Manhattan is now on par with the other boroughs. Overall, the crime per capita is fairly even across all five boroughs.

Figure 8



4.4 Neighborhoods of Intrest

While investigating the NYC area we discovered some neighborhoods of interest, particularly the most dangerous and least dangerous. What we found interesting is the top five most dangerous neighborhoods are fairly spread out. Indicating there is not a 'bad side' of New York, rather hot spots of crime are all around the city, with exception of Staten Island.

Table 2: Most Dangerous

Neighborhood	Number of Crimes
Harlem	16460
Bedford-Stuyvesant	13461
East Harlem	12743
East New York	12550
Jamaica	11604

Harlem and East Harlem are in Manhattan. Bedford-Stuyvesant and East New York are in Brooklyn. While Jamaica is in Queens. Notice in *table 2* that none of the top 5 most dangerous are in Bronx or Staten Island.

Table 3: Least Dangerous

Neighborhood	Number of Crimes
Fort Wadsworth	12
Ferry Point Park	9
Great Kills Park	7
LaGuardia Airport	7
Port Ivory	5

Fort Wadsworth, Port Ivory, and Great Kills Park are in Staten Island. Ferry Point Park and LaGuardia Airport are in Queens. Queens and Staten Island make up the lowest crime areas. Note that in *table 3*, all of the neighborhoods listed appear to be in more rural and less populated areas; which explains the lack of crime. How little crime is actually reported is most shocking, specifically with Port Ivory having only 5 reports the entire year. After investigating, it is known that Port Ivory has very little residents and is mostly industrial. The majority of residents live in mobile homes and there are few single family homes in the area.

4.4.1 Harlem and East Harlem

Harlem and East Harlem were of specific interest because they both were among the top three most dangerous neighborhoods. We took a look into what types of crimes happens in these neighborhoods to see if it was reflective of NYC as a whole.

According to *table 4* it closely relates to the overall crime scene in NYC, but the proportion of Assault/Murders and Drugs/Alcohol was a lot higher than the average. This might indicate a higher level of violent crime in the area and should be avoided by tourists wandering around the city.

Table 4: Harlem & East Harlem

Crime	Occurances of Crime
Assault/Murder	4815
Criminal Mischief	4100
Petit Larceny	4002
Harrassment	3787
Drugs and Alcohol	2946

The crime rate of these neighborhoods was so high we investigated further to find out that the two neighborhoods account for 26.38% of the crime in the Manhattan area, which has over 50 neighborhoods.

Furthermore, we wanted to understand how much of the Assault/Murder in Manhattan these neighborhoods accounted for, which came out to 34.58%.

4.4.2 Low Crime Areas

Low crime areas were also of interest, but did not lead to any major insights. With such low crime rate, there were no trends to investigate; rather, a lack of crime consistency suggests most randomly occur and are rare as seen in *table 5*.

Table 5: Least Dangerous

Crime	Occurances of Crime
Petit Larceny	7
Criminal Mischief	6
Grand Larceny	6
Harrassment	6
Weapons	4

4.4.3 Times Square (Midtown)

We also wanted to look into Times Square, which is apart of the Midtown neighborhood in Manhattan. We suspected that Petit Larceny would be the top crime by far due to tourists being pick-pocketed during their visit. To no one's surprise Petit and Grand Larceny were by far the most reported crimes and harassment in a distant third as seen in *table 6*.

Table 6: Midtown

Crime	Occurances of Crime
Petit Larceny	3723
Grand Larceny	2573
Harrassment	858

5. Conclusion

After gaining an understanding of crime in NYC we are able to several data-driven claims. Crime does not appear to have any patterns regarding neighborhoods or boroughs. While there are several hot spots in the boroughs, the connection between location and crime appears to be random. Perhaps socioeconomic factors play a more major role in these hot spots rather than location.

We can also conclude that seasons and temperature have an affect on the crime rate. While temperature alone has a subtle effect, seasons show a more dramatic change. This is most likely because warmer weather promotes more activity outside of homes and the more people out, means more opportunity for crime.

When comparing the boroughs it is important to distinguish how the crime rate is being calculated because it can paint a very different picture. Our conclusion of with the boroughs is that Manhattan is the most crime dense area, while Brooklyn contains the most crime per capita. This is most likely due to Manhattans small area compared to the other boroughs.

Diving deeper into individual neighborhoods of interest solidified the conclusion that boroughs as a whole are not good or bad, but usually a few neighborhoods within them are. By looking into Harlem and East Harlem we were able to understand that the majority of Manhattan is generally safe, it is just the two neighborhoods raising the overall crime rate. We were also able to better understand the types of crime in each area, like

with Times Square in Midtown. Knowing it is a very tourist heavy part of the city thieves will go there more to target them knowing tourists are unfamiliar with the city; thus, adding to the crime rate.

6. Future Goals

Going forward we would like to learn how Google maps creates routes from location to location and use that with our data product. It would prioritize safer areas for tourists and could even display safe zones, such as police stations and hospitals.

Having the data product become a real-time updated system would make it more accurate for tourists as well as have crime alerts so they know what areas to avoid immediately. Making this into a mobile app and expanding the cities covered would be the ultimate goal for this project.

References

[1] Allen, *Crime Against International Tourists*, https://www.bocsar.nsw.gov.au/Documents/CJB/cjb43.pdf

[2] R Core Team,R: A Language and Environment for Statistical Computing, R Foundation for Statistical-Computing, Vienna, Austria,http://www.R-project.org/, 2014

[3] Yihui Xieknitr: A general-purpose package for dynamic report generation in R,http://yihui.name/knitr/, 2014

[4] NYC OpenData, *NYPD Complaint Data Historic*, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[5] National Centers for Environmental Information(NOAA), *Daily Weather Summaries Central Park, NY*, https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

[6] NYC OpenData, *Neighborhood*, https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhc
swpk-hqdp

[7] Chicago Tribune, *Does a hot summer mean more crime? Here's what the data show*, http://www.chicagotribune.com/news/data/ct-crime-heat-analysis-htmlstory.html

[8] Hadley Wickham, plyr, https://cran.r-project.org/web/packages/plyr/plyr.pdf

[9] Hadley Wickham, dplyr, https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

[10] Hadley Wickham, stringr, https://cran.r-project.org/web/packages/stringr/stringr.pdf

[11] Joe Cheng, leaflet, https://cran.r-project.org/web/packages/leaflet/leaflet.pdf

[12] Edzer Pebesma, sp, https://cran.r-project.org/web/packages/sp/sp.pdf

[13] David Kahle & Hadley Wickham, ggmap, https://cran.r-project.org/web/packages/ggmap/ggmap.pdf

[14] Roger Bivand, rgeos, https://cran.r-project.org/web/packages/rgeos/index.html

[15] Roger Bivand, maptools, https://cran.r-project.org/web/packages/maptools/maptools.pdf

[16] David Robinson, broom, https://cran.r-project.org/web/packages/broom/index.html

[17] Hadley Wickham, httr, https://cran.r-project.org/web/packages/httr/httr.pdf

[18] Roger Bivand, rgdal, https://cran.r-project.org/web/packages/rgdal/rgdal.pdf

[19] Roger, Bivand, spData, https://cran.r-project.org/web/packages/spData/spData.pdf

[20] Kyle Walker, tigris, https://cran.r-project.org/web/packages/tigris/tigris.pdf

[21] Hadley Wickham, ggplot2, https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

[22] Matt Dowle, data.table, https://cran.r-project.org/web/packages/data.table/index.html

[23] Winston Chang, shiny, https://cran.r-project.org/web/packages/shiny/index.html