Project Report

on

Analysis of New York Crime, Pedestrian and Ports Entry data

Author:

Rohit Keshav Bewoor

Matriculation Number 11011831

Big Data and Business Analytics

SRH Hochschule Heidelberg

September 12, 2019

ABSTRACT

This report presents the method and results to evaluate a link, if any, between

- the crimes in New York City,
- number of pedestrians in the city, and
- the number of persons entering the USA through two airports in the vicinity of the five boroughs of New York City.

After cleaning the data from three sources, each set of data was put into a dedicated MongoDb collection using python scripts. Then another python script was used to extract the data from the three collections, perform various manipulations like aggregations and consolidation, and output a final CSV file.

Using Tableau to access the above final CSV file, basic visualisation techniques were used to create dashboards. These can be used to further drill down and explore relationships.

In addition, the statistical tool R, was used to explore a link between the variables mentioned earlier using simple and multiple linear regression.

Table of Contents

1 Introduction

The administration department of New York City (referred to as NYC from hereon) and various government organisations of the USA make data available periodically for free. Three data sources were decided on for this project:

- *First* is a Pedestrian Volume Index – a Bi-Annual Pedestrian Counts [1]
  An index of pedestrian volumes tracking the long-term trends of neighbourhood commercial corridors. Data is collected at 114 locations, including 100 on-street locations. Counts are taken biannually during May and September and Pedestrian volumes at 50 sample locations around the City are combined. The exact dates on which the counts are taken varies from area to area.
  Data set has around 120 rows with 100 columns. Data is provided by the five boroughs of NYC - Brooklyn, Bronx, Manhattan, Queens and Staten Island.

- *Second* is the NYPD Complaint Data Historic[2].
  This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2017. Each record has information about type of crime, reporting time, when crime occurred, location including borough and lat-long, precinct, and suspect and victim demographics.
  Data set is a CSV file with 35 columns and 6,500,871 including header row.

- *Third* is tourist arrival data via major ports of entry into the USA[3].
  These are excel files with number of entries from Overseas based on I-94 entries. For all years from 2007 to 2018, except for 2011, month wise data is available based on the port of entry. While there are between 15 to 30 ports of entry, we will only use the New York and Newark data as these ports are closest to the five boroughs of NYC.

2 Motivation

The idea was to pick a project involving multiple data sources and find a way to marry the information and then attempt to establish a link, if any, between the data. Predicting crime could be a useful public safety tool. In general the motive was also to use publicly available data and to seek insights that would encourage governments and institutions to continue to release such data periodically.

3 Problem Statement

To test the following hypothesis using the data extracted from the three sources:

1) The number of inbound arrivals at the ports of entry near NYC, i.e. New York and/or Newark, affects the number of pedestrians found in New York City.

2) The total number of crimes reported to the NYPD is affected *by a combination of, or individually*, the number of arrivals via the ports of entry of New York and Newark *and* the number of pedestrians in NYC.
Thus, these two data points can be used to predict how many crimes will be reported to the NYPD.
If possible, to also explore such a link by using more granular data by the type of crime, e.g. robbery, harassment, drugs, felony, assault, etc.

3) From an exploratory perspective:

- visualize the change in the number of pedestrians, the number of port arrivals and the number of the crimes over the time period for which the data is available i.e. May and September of 2007 to 2017.

- analyse the data in other suitable ways.

4 Literature Review

No useful information was found regarding the analysis of the data being used in this project.

However, while there is a Kaggle link[4] for the crime data, there are only kernels present dealing with basic scatter plots and histograms, etc.

Thus the data sources considered and the analysis done is novel to best of the authors' knowledge.

5 Proposed Solution

The following tools and languages were used to implement the solution:

- Database: MongoDb
- Programming language: Python
- Statistical Analysis: R package
- File formats for data loading and data extraction: CSV and Excel files
- Visualisation: Tableau

Brief outline of the steps used in the implemented solution:

a) Perform high level analysis of the data sources.

b) Clean and prepare the data by converting it to CSV files.

c) Use Python scripts to use the CSV files and load into MongoDB collections.

d) Use a Python script to "marry" the data from the three collections and process them into rows to be written to a final output CSV file.

e) Use the final output CSV file to load into Tableau.

f) Find and use a shape file for the five boroughs of NYC and connect to the data loaded in step 5.

g) Create dashboards that address the problems statements and allow exploratory analysis of the data to find trends and insights.

h) Use some statistical software package (R or scikit of Python) to perform regression analysis to test the hypothesis mentioned in the problem statement.

6 Contributions

Not applicable as this is an individual submission.

7 Architecture

The raw data from the three sources was converted into CSV files. Using python scripts, these file were used to load data into individual MongoDb collections. One python script was written to access MongoDb data and process it.

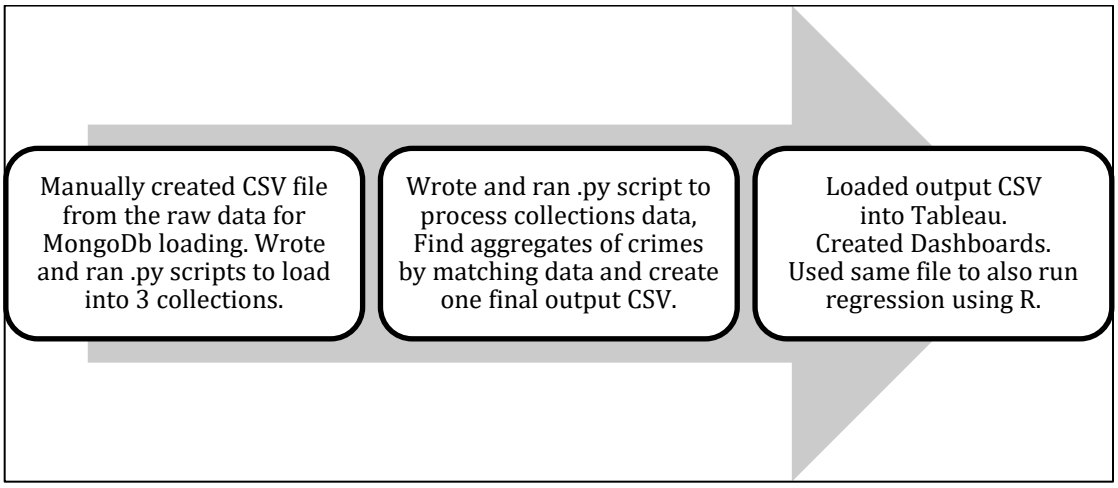
As data for Pedestrian information was only available for two months viz. May and September each year, it was decided to create one consolidated output CSV file. Each row would for a particular – Year, Month, Borough – combination. Thus the output file had 110 rows (11 years x 2 months x 5 boroughs = 110). The idea was to find the aggregate of total number of crimes and total of crimes by various sub-criteria and capture it in the rows being written to the output CSV file. E.g. a row for year=2014, month=September and Borough=Queens, would now have around 450 columns consisting of aggregate of the number of crimes in that Year-Month-Borough combination from different perspectives.

This was achieved by using the granular details present in the source data: these are broadly based on the Key Code with 74 unique values, Offense Description with 70

unique values, Premise Type with 72 unique values and Law Category Code with 3 unique values.

A python script that ran continuously for almost 24 hours created the final output CSV file. Due to memory and run-time issues, many of the granular data fields had to be ignored. Hence, the final output file had to be reduced in size from the original plan of including around 2000 variables to only processing around 200 variables.

Figure 1: Architecture Pipeline



7.1 Data Cleaning and Preparation

About the raw data:
Data was available as one CSV and multiple excel files from the sources mentioned. The crime data was a CSV file (approx. 2GB size) with 6.5 million rows and 35 columns per row. The Pedestrian data was available in an excel file on one worksheet. For the Arrivals at Ports information, data was in one excel file for years 2012-2018; with one worksheet for each year containing the monthly totals directly. But for the years 2007-2010, each years data was in a different excel file; with cumulative counts for each month on a different sheets. Data for 2011 was missing completely.

Preparation:
For the Crime data, the source CSV file was suitable without any pre-processing required.

For the ports data, the missing values for 2011 were assigned taking an average of the values for 2010 and 2012. Two issues had to dealt with:

- Inconsistent date format (mm/dd/yyy vs dd/mm/yyyy)
- Year to month cumulative data to be converted to month-wise data

For the Pedestrian data:

- All the data was on one excel file, but each borough had multiple data rows.
- The counts for each borough were found manually and the CSV file for MongoDB load could then be created.

Thus, after the above steps, there were finally three CSV files available.

Loading to MongoDb:
Three scripts were written in python to take the three input CSV files from earlier step. The loading into MongoDb for the 6.5 million rows had to be done in parts due to large data size, with each loading instance processing 1 million data rows.

Figure 2: Running scripts to load Crime data to MongoDb

```
Command to load the first 1 million rows of Crime data into MongoDb
python dataInsertMongoNYCrime5.py "D:/EverythingD/01SRH-BDBA
Acads/Blk7-DataStoryTelling/Data4Analysis/finalDataSourcesUsed4MongoAGAIN/NYPD_Complaint_Data_Historic.csv"
1000000 0 db2story coll21nyccrimedata LOG_dataInsertMongoNYCrime5-nr1000k-nsk0k.log

Command to load the second 1 million rows of Crime data into MongoDb
python dataInsertMongoNYCrime5.py "D:/EverythingD/01SRH-BDBA
Acads/Blk7-DataStoryTelling/Data4Analysis/finalDataSourcesUsed4MongoAGAIN/NYPD_Complaint_Data_Historic.csv"
1000000 1000000 db2story coll21nyccrimedata LOG_dataInsertMongoNYCrime5-nr1000k-nsk1000k.log

Command to load the final set of 1 million rows of Crime data into MongoDb
python dataInsertMongoNYCrime5.py "D:/EverythingD/01SRH-BDBA
Acads/Blk7-DataStoryTelling/Data4Analysis/finalDataSourcesUsed4MongoAGAIN/NYPD_Complaint_Data_Historic.csv"
1000000 5000000 db2story coll21nyccrimedata LOG_dataInsertMongoNYCrime5-nr1000k-nsk5000k.log
```

Figure 3: Running scripts to load Pedestrian and Ports data to MongoDb

```
Command to load the entire data for Pedestrian data into MongoDB
python dataInsertMongoNYPedestrian5.py "D:/EverythingD/01SRH-BDBA
Acads/Blk7-DataStoryTelling/Data4Analysis/NYCPedestrianDataPrepared-20190905-clean5.csv" -1 0 db2story coll23pedesdata
LOG_dataInsertMongoNYPedestrian5-loadAll.log



Command to load the entire data for Ports Data data into MongoDB
python dataInsertMongoNYPortsdata5.py "D:/EverythingD/01SRH-BDBA
Acads/Blk7-DataStoryTelling/Data4Analysis/PortsDataPrepared-20190906-clean3.csv" -1 0 db2story coll23portsdata
LOG_dataInsertMongoNYPortsdata5-loadAll.log
```

Some details of the data layout and the important fields is shown below:

1. For Crime data layout:

Figure 4: MongoDb data layout - Crime data document

```
NYPD crime data
CMPLNT_NUM              659261697
CMPLNT_FR_DT            05/12/2007
CMPLNT_FR_TM            22:15:00
CMPLNT_TO_DT            05/12/2007
CMPLNT_TO_TM            22:22:00
ADDR_PCT_CD             109
RPT_DT                  05/12/2007
KY_CD                   106
OFNS_DESC               FELONY ASSAULT
PD_CD                   109
PD_DESC                 "ASSAULT 2,1,UNCLASSIFIED"
CRM_ATPT_CPTD_CD        COMPLETED
LAW_CAT_CD              FELONY
BORO_NM                 QUEENS
LOC_OF_OCCUR_DESC       INSIDE
PREM_TYP_DESC           AR/NIGHT CLUB
JURIS_DESC              N.Y. POLICE DEPT
JURISDICTION_CODE       0
PARKS_NM                NA
HADEVELOPT
HOUSING_PSA             NA
X_COORD_CD              1030368
Y_COORD_CD              214723
SUSP_AGE_GROUP
SUSP_RACE               ASIAN / PACIFIC ISLANDER
SUSP_SEX                M
TRANSIT_DISTRICT
Latitude                40.7559295
Longitude               -73.83353931
Lat_Lon                 "(40.755929496,-73.833539312)"
PATROL_BORO             PATROL BORO QUEENS NORTH
STATION_NAME
VIC_AGE_GROUP           18-24
VIC_RACE                ASIAN / PACIFIC ISLANDER
VIC_SEX                 M
```

a. While a link to the data dictionary is here, a brief introduction to the data is given below:

   i. Unique complaint number: CMPLNT_NUM

   ii. Complaint Date/Time: CMPLNT_FR_DT, CMPLNT_FR_TM

   iii. Location Info: Latitude, Longitude, BORO_NM

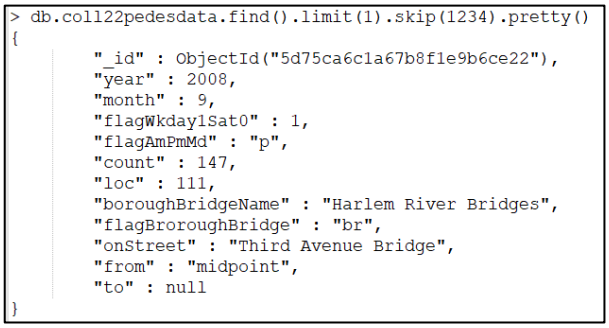
   iv. High level classification (Law Category Code with values as "Felony", "Misdemeanor" and "Violation") : LAW_CAT_CD

   v. Other granular data classifying by other types:
   KY_CD (74 unique values), OFNS_DESC (71 unique values),
   PD_CD (427 unique values), PD_DESC (415 unique values),
   PREM_TYP_DESC (73 unique values)

   vi. Suspect and Victim demographic info: SUSP_AGE_GROUP,
   SUSP_RACE , SUSP_SEX , VIC_AGE_GROUP , VIC_RACE , VIC_SEX

2. For Pedestrian data:

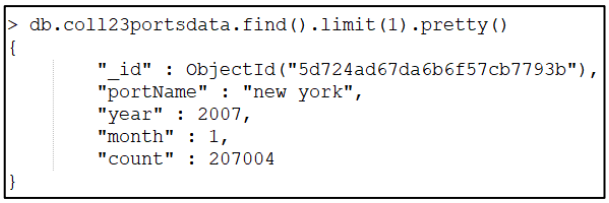
Figure 5: MongoDb data layout - Pedestrian data document



```
> db.coll22pedesdata.find().limit(1).skip(1234).pretty()
{
        "_id" : ObjectId("5d75ca6c1a67b8f1e9b6ce22"),
        "year" : 2008,
        "month" : 9,
        "flagWkday1Sat0" : 1,
        "flagAmPmMd" : "p",
        "count" : 147,
        "loc" : 111,
        "boroughBridgeName" : "Harlem River Bridges",
        "flagBroroughBridge" : "br",
        "onStreet" : "Third Avenue Bridge",
        "from" : "midpoint",
        "to" : null
}
```

3. For Ports data:

Figure 6: MongoDb data layout - Ports data document



```
> db.coll23portsdata.find().limit(1).pretty()
{
        "_id" : ObjectId("5d724ad67da6b6f57cb7793b"),
        "portName" : "new york",
        "year" : 2007,
        "month" : 1,
        "count" : 207004
}
```

Retrieval of MongoDb data and creating output CSV for Tableau processing:
The stumbling block was that Pedestrian data was only available for May and September in each year. Therefore all the data from the three sources was to be matched up

carefully to ensure that only data for these months was extracted, processed and written to the output CSV file.
Initially a python script was written to compute the number of crimes for each of the types of granular level information – almost 1060 types in all. But this approach caused memory issues, run time length was too long; and it was impossible to extract the necessary information. So the PD_CD and PD_DESC data was completely disregarded for processing and the number of crimes was only calculated for granular perspective by including the following fields: KY_CD, OFNS_DESC, PREM_TYP_DESC and LAW_CAT_CD. This reduced the number for 1060 to 218, which allowed processing to occur successfully.
Please refer the data dictionary [5] for field level details.

A few code snippets of the python script used to extract data from MongoDb, process and write the output CSV file:

Figure 7: Script for data processing and creating final output CSV file - 1

Figure 8: Script for data processing and creating final output CSV file - 2

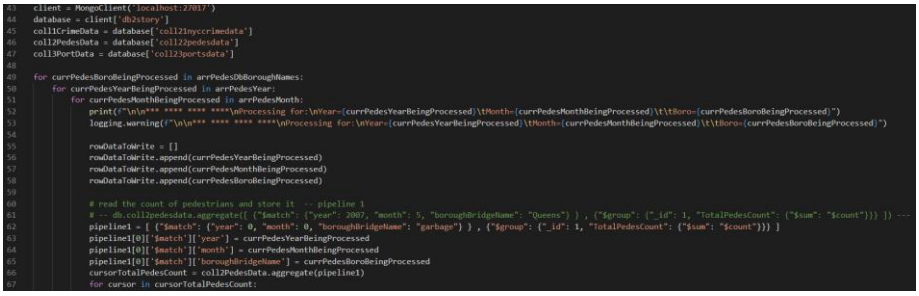
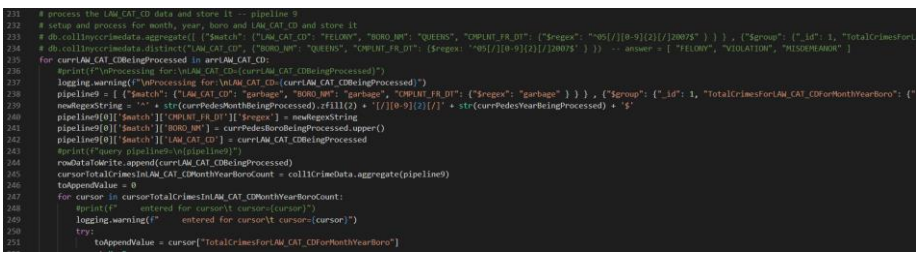


Figure 9: Script for data processing and creating final output CSV file - 3



Below is a snapshot of the start of the final output CSV showing the header and first data row (the data row is highlighted in blue):

Figure 10: Layout of the final output CSV file

Using Tableau:

The final output CSV file had 110 rows with around 450 columns. A spatial file for the five boroughs was found on the internet [6] and used for visualisation.

Regression using R:

Simple and Multiple Linear regression was done – please refer Results section for details.

7.2 Visualisation techniques suitable for the chosen dataset

The following visualisation techniques could fit this use case:

a) Spatial (geographic based): Since there was longitude and latitude data available, such a technique is well suited.

b) Heat maps: could be used thereafter.

c) Pie-charts: As there are only 5 boroughs one could use pie charts to show various comparisons borough-wise.

d) Bar and Line Charts: As there is time series data, this approach is suitable for trend analysis.

7.3 Justification of the chosen visualisation technique

A spatial map for some dashboards was chosen as the starting point for two reasons:

- the original dataset had latitude and longitude data which might have been used later;
- Secondly, seeing the five boroughs on a map allows the exploratory research to be done with a better bird's eye view of what one is attempting to look for.

Since there are several years of data and it is time series based, bar charts and line charts are preferred over pie charts.

To allow easier analysis and comparison of trends, the choice was made to include following elements in the dashboard design:

- juxtaposing of multiple graphs
- Interaction plots

7.4 Marks and Channels

Marks used in the visualisation are Lines, Points, Bar.

Channels used in the dash boards are length, position, color, dual axis.

7.5 Two visual design proposals

One option was to provide data in tabular form to allow exploratory searching.

The overall design is made with the assumption that the end-user is not interested at this stage to get into the numbers.

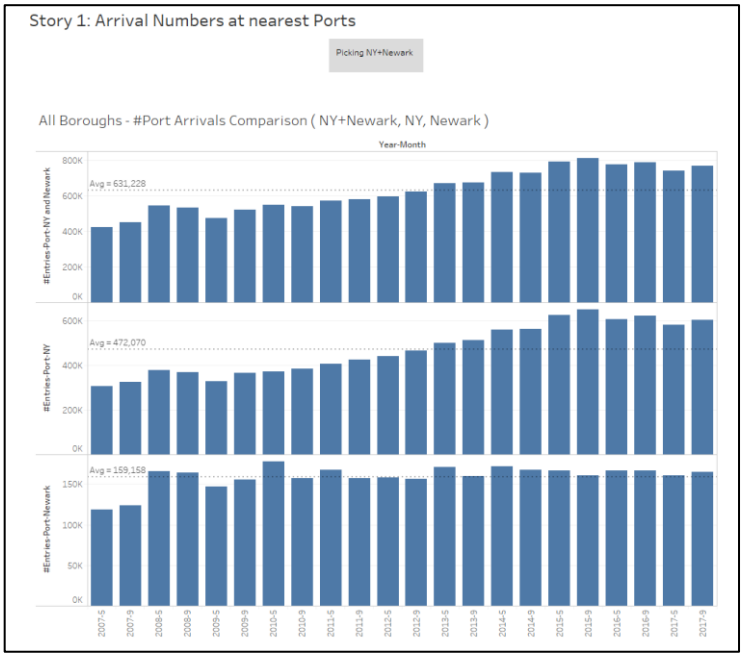
- Hence, a choice was made for an interaction plot with spatial map to make easier navigation and make the exploration interesting.
- There are so many fields that it is not feasible to inspect all the information individually. So the aim is provide an interesting graphical interface for the user to delve deeper into the data to find insights by simply visual inspection of the dashboards.
- The user can later on turn to other persons to draw numerical insights using the output CSV file.

7.6 Dashboard Design

7.6.1 Story 1 Dashboard

Why the choice was to include port entry numbers of NY and Newark together for analysis, rarther than consider the numbers for each port individually.

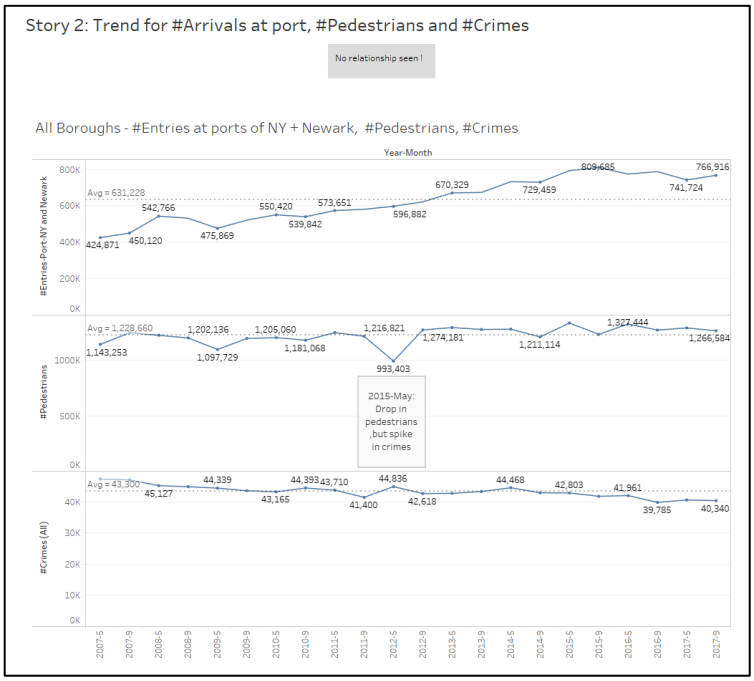
Figure 11: Dashboard#1



7.6.2 Story 2 Dashboard

Visually inspecting relationship between the Pedestrian, Port Arrivals and Total Crimes.

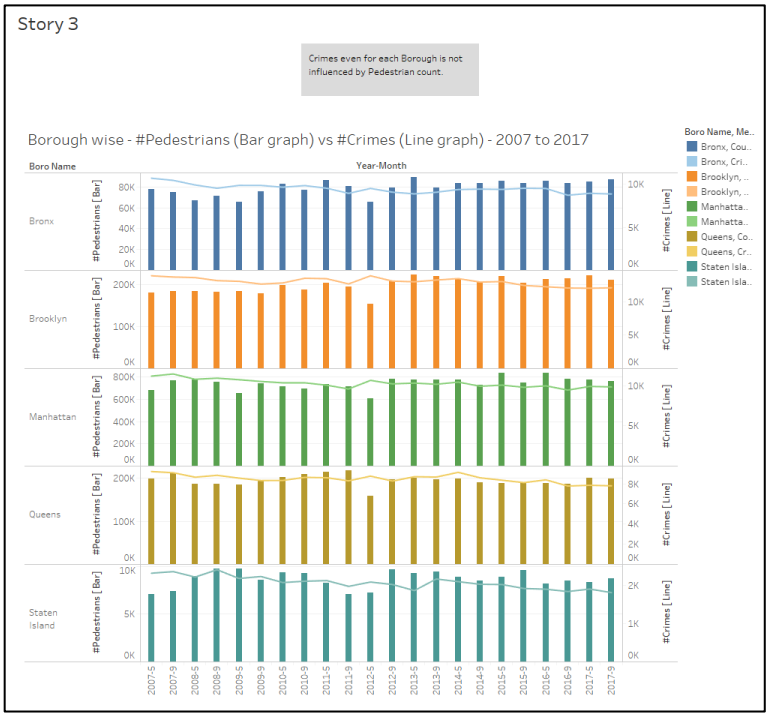
Figure 12: Dashboard#2



7.6.3 Story 3 Dashboard

A more in-depth visual inspection of Pedestrian counts v/s Total Crimes data is shown via use of a dual axis for easier comparison. Pedestrian data is shown using Bar graphs and the Total Crimes using a Line graph.

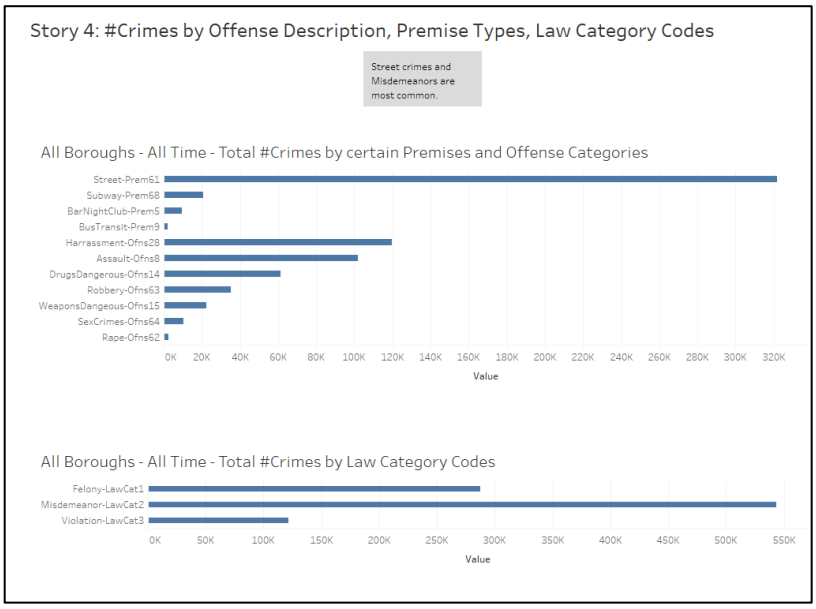
Figure 13: Dashboard#3



7.6.4 Story 4 Dashboard

Introduction to the types of crimes that occur frequently and the number of such crimes at an aggregate level i.e. for all the 11 years and the five boroughs together.

Figure 14: Dashboard#4



7.6.5 Story 5 Dashboard

Interactive dashboard allows the ability to click on borough in the map, which affects the middle visual. Then one can independently choose the years for which to view the last visual of crime breakdown by years using the Slider filter.
Focus of this dashboard is on number of Crimes by Law Category Codes.

Figure 15: Dashboard#5 – initial view

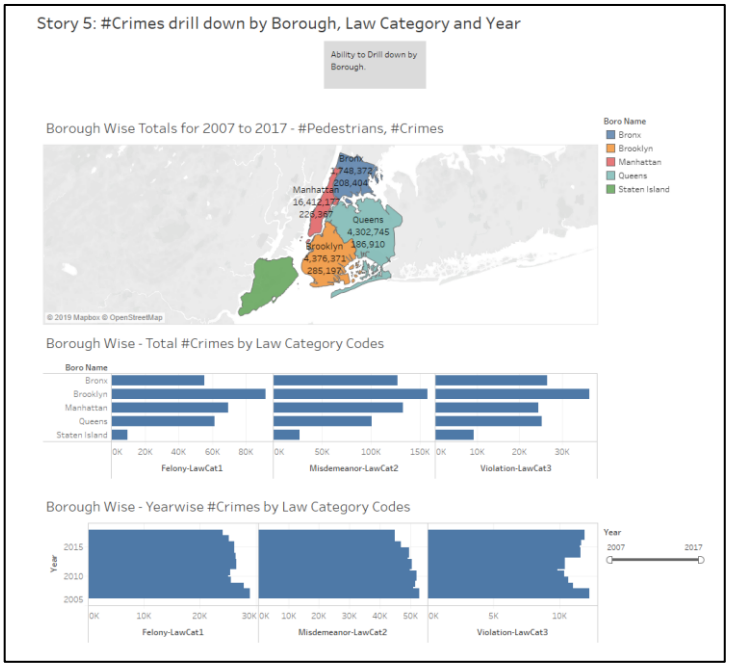


Figure 16: Dashboard#5 - on clicking some borough

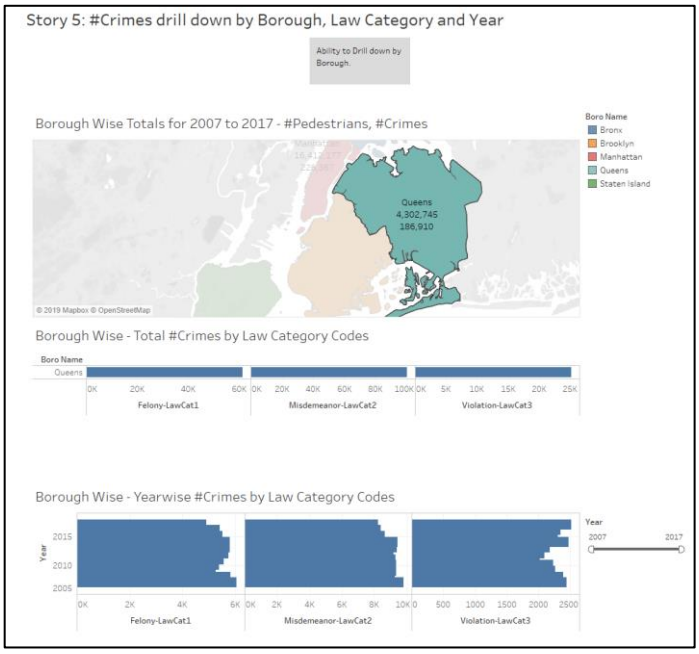
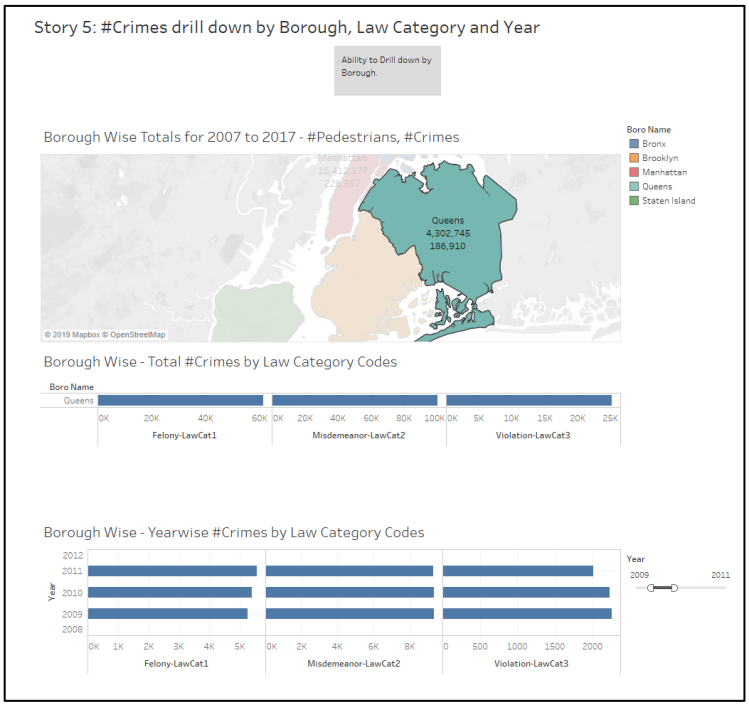


Figure 17: Dashboard#5 - on selecting year/s using the slider filter



7.6.6 Story 6 Dashboard

Interactive dashboard allows the ability to click on borough in the map, which affects the visual for total crimes by borough. Then one can inspect the last visual for the number of crimes for each time slice to see trends over the 11 year period.

Focus of this dashboard is on certain frequently occurring crimes as per the classification of the “Offense Type Description” and the “Premise Type Description”.

Figure 18: Dashboard#6 - initial view

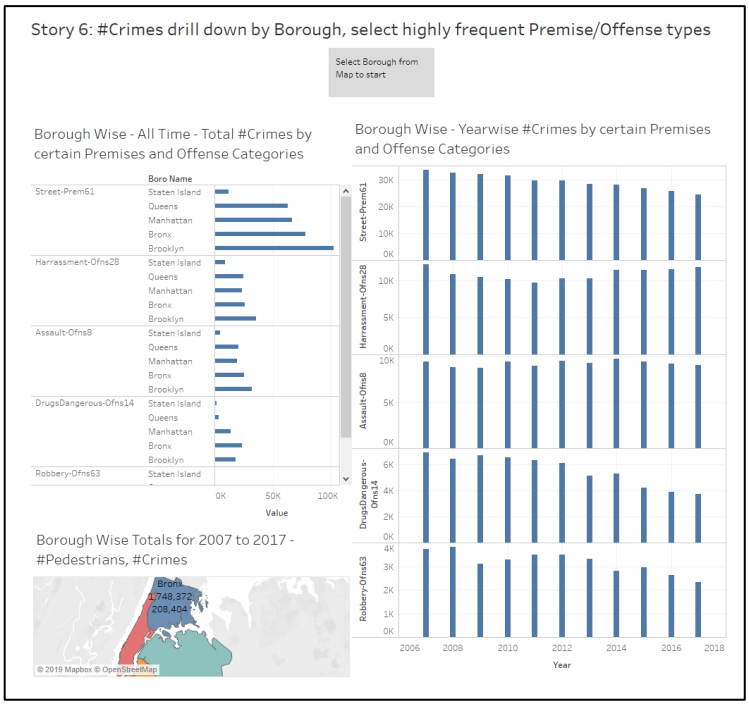
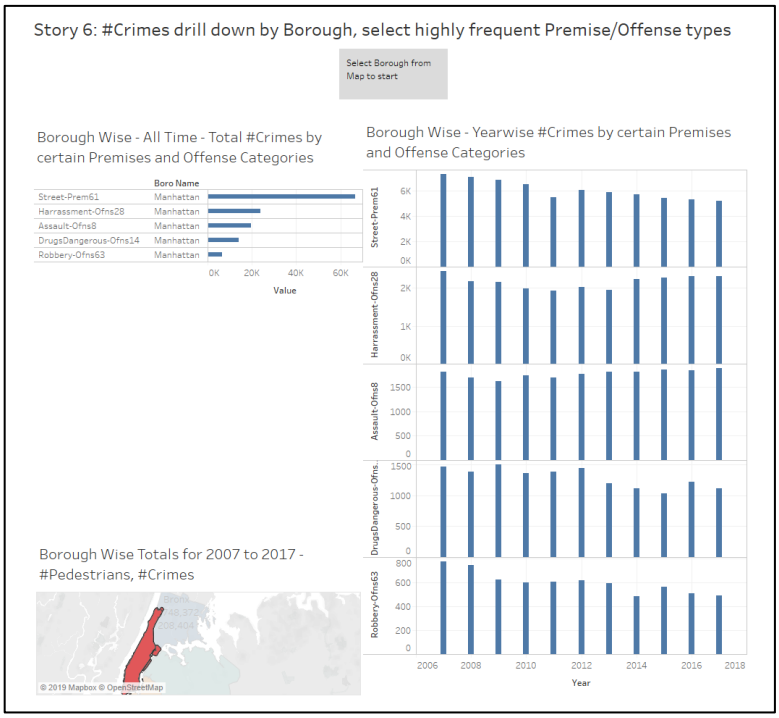


Figure 19: Dashboard#6 - on clicking some borough



8 Results

8.1 Results from Dashboard analysis

a) From Story 1 Dashboard: Regarding number of arrivals at Ports of Entry for New York, Newark and the combined numbers:

   - We can see that the number of arrivals for port Newark picks up from 2007 to 2008, but thereafter it is flattish over the entire period.
   - But for New York it increases steadily.
   - Taking the numbers for both ports together, continues to show the increasing trend.

b) From Story 2 Dashboard: Regarding the comparison of ports arrival numbers, pedestrian count and the total number of crimes:

   - There is a steady increase increase in the number of arrivals
   - The number of pedestrians seems to be fairly flat
   - The total number of crimes is decreasing at a slow but steady pace

c) From Story 3 Dashboard: Link between the number of pedestrians and the total number of crimes – for each borough:

   - No clear tread can be ascertained by visual inspection.

d) From Story 4 Dashboard:

   - Under the classification based on the type of Law Category Code, “Misdemeanor” type crimes are the most frequent and account for nearly 60% of the total.
   - Under the classification based on the type of Premise, “Street” type crimes with a count of 320k+ are the most frequent. The next most frequent crime type is “Subway” type with only around 20k count.
   - Under the classification based on the Offence Description, crimes of type “Harassment” are most frequent with a count of nearly 120k. The next highest is for “Assault” type with a count just over 100k.

e) From Story 5 Dashboard:

   - The use of an interaction plot with filters, allows the user to inspect the link between borough, Law Category Code classification of crimes, and the years in which these crimes occurred in an interesting way.

f) From Story 6 Dashboard:

- The use of an interaction plot, allows the user to inspect the link between borough, the top 5 frequently occurring types of crimes, and the years in which these crimes occurred in an interesting way.
- An example of an insight found is:
  - When comparing the reduction in the number of "DrugsDangerous" type crimes over the 11 years, for all boroughs except Manhattan, there is a fall of 50%.
    But only for Manhattan this fall is only around 30% - 7263 to 5152.

8.2 Results from Linear Regression using the statistical package R

Note: All analysis is done with confidence level of 95%.

8.2.1 Model#1: CrimesTotal = dependsOn(countPedestrians)

Figure 20: R output for Model#1

```
> ## model 1 :   crimeTotal = countPedes
> slm1 <- lm(d$crimeTotal ~ d$countPedes , data = trainingSet1)
> summary(slm1)

Call:
lm(formula = d$crimeTotal ~ d$countPedes, data = trainingSet1)

Residuals:
   Min     1Q Median     3Q    Max
 -5439  -1883    161   2118   5733

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.201e+03  4.342e+02  16.584  < 2e-16 ***
d$countPedes 5.940e-03  1.210e-03   4.909 3.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3319 on 108 degrees of freedom
Multiple R-squared:  0.1824,    Adjusted R-squared:  0.1749
F-statistic:  24.1 on 1 and 108 DF,  p-value: 3.262e-06
```

Overall model is statistically significant as p-value is 3.26e-06.
Effect of Pedestrian Count on total Crimes is thus statistically.
Beta(PedestrianCount) = 5.94e-03, which means that for every 1000 increase in Pedestrians, the Total crimes increase by 5.94.
Adjusted Rsquared is only at 17.49%.

8.2.2 Model#2: CrimesTotal = dependsOn(countPortNyNewark)

Figure 21: R output for Model#2

```
> ## model 2 :   crimeTotal = countPortNYnNewark
> slm2 <- lm(d$crimeTotal ~ d$countPortNYnNewark , data = trainingSet1)
> summary(slm2)

Call:
lm(formula = d$crimeTotal ~ d$countPortNYnNewark, data = trainingSet1)

Residuals:
    Min      1Q  Median      3Q     Max
-6858.2  -494.2   786.3  1868.0  5099.0

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.022e+04  1.894e+03   5.395 4.09e-07 ***
d$countPortNYnNewark -2.464e-03  2.949e-03  -0.836    0.405
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3659 on 108 degrees of freedom
Multiple R-squared:  0.006425,  Adjusted R-squared:  -0.002775
F-statistic: 0.6984 on 1 and 108 DF,  p-value: 0.4052
```

Overall model discarded as statistically not significant with p-value of 0.4.

8.2.3 Model#3: CrimesTotal = dependsOn(countPedestrians , countPortNyNewark)

No use exploring as including Port counts is pointless.

8.2.4 Model#4: CrimesTotalPremTypeStreet = dependsOn(countPedestrians)

Figure 22: R output for Model#4

```
> ## model 4 :   sumCrimeStreet.Prem61 = countPedes
> slm4 <- lm(d$sumCrimeStreet.Prem61 ~ d$countPedes , data = trainingSet2)
> summary(slm4)

Call:
lm(formula = d$sumCrimeStreet.Prem61 ~ d$countPedes, data = trainingSet2)

Residuals:
     Min       1Q   Median       3Q      Max
-2198.09  -839.68    97.97  1045.89  2379.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.624e+03  1.769e+02  14.839   <2e-16 ***
d$countPedes 1.225e-03  4.928e-04   2.486   0.0144 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1352 on 108 degrees of freedom
Multiple R-squared:  0.05414,   Adjusted R-squared:  0.04538
F-statistic: 6.182 on 1 and 108 DF,  p-value: 0.01444
```

Overall model is statistically significant with p-value 0.014.
Effect of Pedestrian Count on total Street Crimes is statistically significant.
Beta(PedestrianCount) = 1.225e-03, which means that for every 1000 increase in Pedestrians, the Total crimes increase by 1.22
Adjusted Rsquared is very low at 4.5%, which is even lesser than for Model#1.

8.2.5 Model#5: CrimesTotalPremStreet = dependsOn(countPortNyNewark)

Figure 23: R output for Model#5

```
> ## model 5 :   sumCrimeStreet.Prem61 = countPortNYnNewark
> slm5 <- lm(d$sumCrimeStreet.Prem61 ~ d$countPortNYnNewark , data = trainingSet2)
> summary(slm5)

Call:
lm(formula = d$sumCrimeStreet.Prem61 ~ d$countPortNYnNewark,
    data = trainingSet2)

Residuals:
    Min      1Q  Median      3Q     Max
-2699.6  -199.0   131.2   785.6  2068.6

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.324e+03  7.063e+02   6.122 1.52e-08 ***
d$countPortNYnNewark -2.215e-03  1.100e-03  -2.014   0.0465 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1365 on 108 degrees of freedom
Multiple R-squared:  0.0362,    Adjusted R-squared:  0.02728
F-statistic: 4.057 on 1 and 108 DF,  p-value: 0.04648
```

Overall model is statistically significant with p-value of 0.027.

Effect of Pedestrian Count on total Street Crimes is just about statistically significant, with a 5% confidence level, with p-value of 0.0475.
Beta(PortCountNynNewark) = -2.215e03, which means that for every 1000 increase in arrivals at Ports, the Total crimes decrease by 2.22.
Adjusted Rsquared is extremely low at only 2.7%.

8.2.6 Model 6: CrimesTotalPremStreet = dependsOn(countPedestrians , countPortNyNewark)

Figure 24: R output for Model#6

```
> ## model 6 :   sumCrimeStreet.Prem61 = countPedes + countPortNYnNewark
> mlm6 <- lm(d$sumCrimeStreet.Prem61 ~ d$countPedes + d$countPortNYnNewark , data = trainingSet2)
> summary(mlm6)

Call:
lm(formula = d$sumCrimeStreet.Prem61 ~ d$countPedes + d$countPortNYnNewark,
    data = trainingSet2)

Residuals:
     Min       1Q   Median       3Q      Max
-2418.04  -648.48   -15.41  1001.55  2137.48

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.073e+03  6.950e+02   5.861 5.17e-08 ***
d$countPedes         1.260e-03  4.850e-04   2.598   0.0107 *
d$countPortNYnNewark -2.308e-03  1.072e-03  -2.153   0.0336 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1330 on 107 degrees of freedom
Multiple R-squared:  0.09341,   Adjusted R-squared:  0.07647
F-statistic: 5.512 on 2 and 107 DF,  p-value: 0.005265
```

Overall model is statistically significant with p-value of 0.005.
Effect of Pedestrian Count and Ports Count together on total Street Crimes is statistically significant with p-value of 0.0107 and 0.0336 respectively.
Beta(PedestrianCount) = 1.26e-03, which means that for every 1000 increase in Pedestrians, the Total crimes increase by 1.26.
Beta(PortCount) = -2.308e-03, which means that for every 1000 increase in arrivals at Ports, the Total crimes decreases by 2.3.
Adjusted Rsquared still weak at 7.6%, but better than for for Model#4 and 5.

8.2.7 Model#7: countPedestrians = dependsOn(countPortNyNewark)

Figure 25: R output for Model#7

```
> ## model 7 :   countPedes = countPortNYnNewark
> slm7 <- lm(d$countPedes ~ d$countPortNYnNewark , data = trainingSet3)
> summary(slm7)

Call:
lm(formula = d$countPedes ~ d$countPortNYnNewark, data = trainingSet3)

Residuals:
    Min      1Q  Median      3Q     Max
-249368 -170675  -55585  -29903  577079

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.990e+05  1.365e+05   1.458    0.148
d$countPortNYnNewark 7.400e-02  2.126e-01   0.348    0.728

Residual standard error: 263800 on 108 degrees of freedom
Multiple R-squared:  0.00112,   Adjusted R-squared:  -0.008129
F-statistic: 0.1211 on 1 and 108 DF,  p-value: 0.7285
```

Overall model discarded as statistically not significant with p-value of 0.72.

9 Evaluation

Interactive dashboard allows the ability to click on Borough in the map, which affects the middle visual. Independently choose the years for which to view the last visual of crime breakdown by years using the Slider filter. Focus is on number of Crimes by Law Category Codes. The feedback for including such an interaction was positive with the person saying it helped them maintain the "bird eye" view of what location they were currently analysis. This was the feedback that was expected.

However, when shown the graphs on Dashboard-6 with around 10 types of crimes, even after almost 90 seconds, the user was not really being able to discern a patter and seems to concentrating too hard. Based on the feedback, it was decided to reduce the number of types of crimes to be presented on the dashboard to only 5 types. The following five were chosen as these were the most frequent types: "Street", "Harassment", "Assault", "DrugsDangerous" and "Robbery".

10 Discussions and Conclusions

The underlying assumption is that persons landing at both ports of entry – New York and Newark - will tend to visit NYC, due to proximity of both places to NYC. Therefore, as we can see variation with time for the combined total, these numbers were used for the overall analysis and *not just New York or just Newark*.

By inspecting the graphs plotted in the Data dashboards using Tableau and a more rigorous analysis using linear regression with R, the conclusions are:

1) The Number of persons arriving at the Ports of New York and Newark together, has no impact on the Number of Pedestrians found in NY city.

2) Regarding Total number of Crimes (all Types combined):

    a. There is an increase of 5.94 for every 1000 increase in Pedestrians. However the data only explains 17% variation in the Total Crimes.

    b. The number of persons arriving at the Ports of New York and Newark together has no impact.

3) Regarding Street Crimes only (with PREM_TYP_CODE = 61):

    a. Taking into account both the Ports data and Pedestrian data explains a 7% variation which is higher than on using a simple linear model with only one of these variables.

    b. There is a increase of 1.26 in crimes for every 1000 increase in Pedestrians. But there is a decrease of 2.3 in crimes for every 1000 increase in the number of arrivals at the Ports of New York and Newark combined.

From the Tableau Dashboards we notice the following:

1) For the months of May and September only, from the years 2007 to 2017, the number of arrivals at the Port of Newark is flat while arrivals at NY are increasing.
   Even after taking the combined total, there is a steady, but slower, increase in the port arrival numbers.

2) In May 2015, despite a drop of almost 25% in Pedestrians count, the total number of Crimes increased by almost 10%. This was counter-intuitive to the expectation.

3) In general, the total number of crimes for May and September from 2007 to 2017 is decreasing slowly but steadily.

4) From the graph there seems to be no relationships between the total Crime and the Pedestrians count. But when a more rigorous statistical analysis is done, there is a link that can be expressed numerically and is covered already.

5) The most prolific types of crimes were those classified as:

   a. Law Category code of 2 which means “Misdemeanors”.

   b. Premise Type as “Streets”

   c. Other types of crimes are those classified with the Offence Descriptions of “Harassment”, “Assault” and “DrugsDangerous”.

6) For crimes classified with Offense Description of “DrugsDangerous”, there is a substantial reduction of 50% for all boroughs except Manhattan. For Manhattan, this reduction is only around 30%.

Further scope:

- Due to processing power and memory limitations, the latitude and longitude data was never used. But that could show interesting trends if explored.

- There is too much granular data with over a thousand types of classifications based on the PD_CD, PD_DESC, OFNS_DESC, KY_CD, LAW_CAT_CODE and PREMISE_TYP_DESC. A suitable way to organise the data with the NYPD releasing more information on how to club the disparate classifications will aid in further analysis.

References

[1] Link for Pedestrian Count data:
https://www1.nyc.gov/html/dot/downloads/misc/nycdot-bi-annual-pedestrian-index.xlsx

[2] Link for NYPD Historic Complaints crime data:
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[3] Link for Tourist Arrival at Ports in USA:
https://travel.trade.gov/view/m-2017-I-001/documents/Final%20COR%20Port%20of%20Entry.xlsx

[4] Kaggle dataset link for NYPD Historic Crime data:
www.kaggle.com/agustinsellanes/nypd-complaint-data-historic

[5] NYPD Crime dataset Data Dictionary:
https://data.cityofnewyork.us/api/views/qgea-i56i/files/ee823139-888e-4ad0-badfe18e2674a9cb?download=true&filename=NYPD_Complaint_Historic_DataDictionary.xlsx

[6] Spatial File location for 5 NY City boroughs: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm