Data Visualization Project

Analysis of New York Crime, Pedestrian and Ports Entry data

Presented on 12.09.2019 by:
Rohit Keshav Bewoor
Matriculation Number 11011831
Big Data and Business Analytics
SRH Hochschule Heidelberg

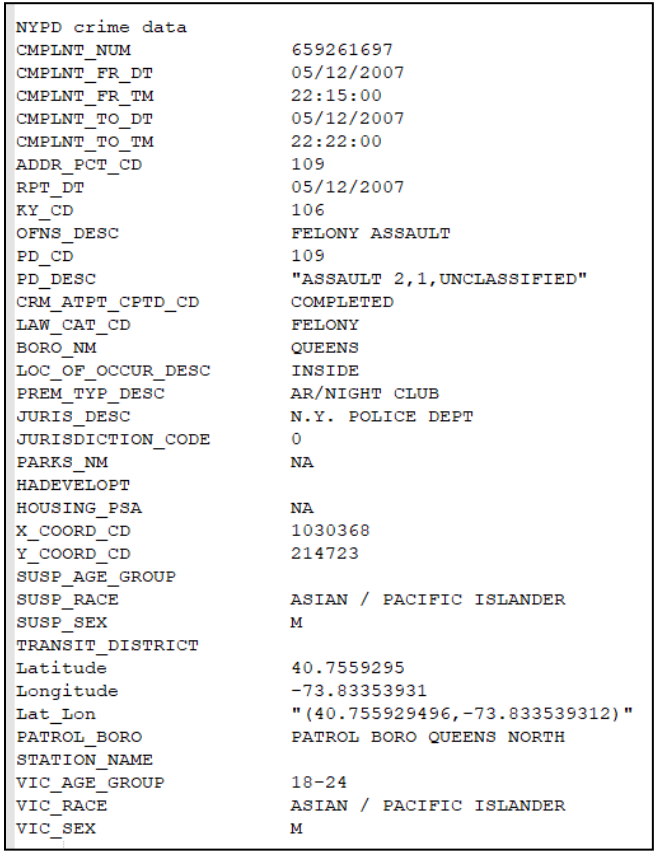
SRH HOCHSCHULE HEIDELBERG
Intelligence in Learning

INTRODUCTION

- 3 datasets were used to create consolidated data
- Pedestrian Volume Index
  - May and September, from 2007 to 2017 data available.
  - Excel file – approx. 120 rows with 100 columns
- NYPD Crime Historic Data
  - Each reported crime from 2006 to 2017
  - Information like date, borough, classification of crime in various ways, suspect and victim demographics
  - CSV file – approx. 6.5 million rows with 35 columns
- Port of Entry data regarding Arrival numbers into USA
  - Multiple Excel files – year wise 2007 to 2010, 2011 missing, consolidated file for 2012 to 2018.
  - Multiple Ports including New York and Newark.
  - Inconsistent date formats, layouts in some years.

How does the Data look?

Example of Crime Data Record after loading into Mongodb



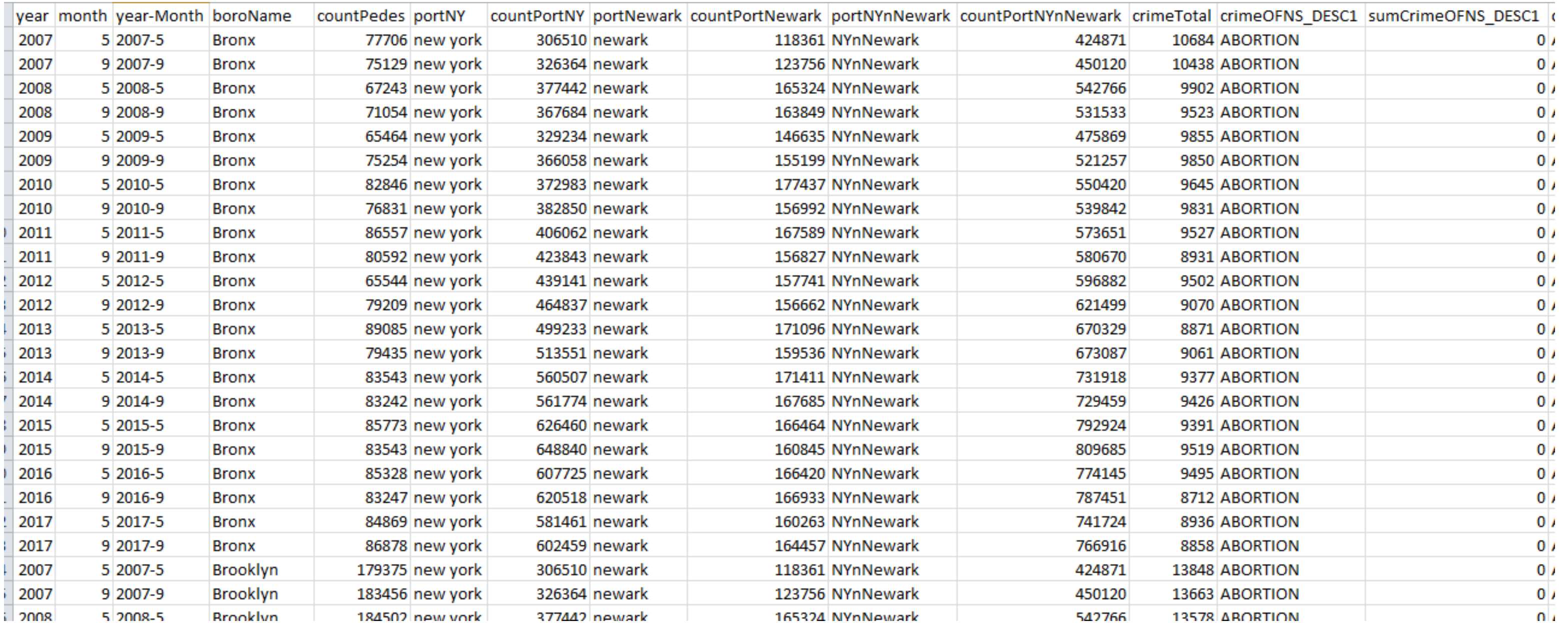
```
NYPD crime data
CMPLNT_NUM               659261697
CMPLNT_FR_DT             05/12/2007
CMPLNT_FR_TM             22:15:00
CMPLNT_TO_DT             05/12/2007
CMPLNT_TO_TM             22:22:00
ADDR_PCT_CD              109
RPT_DT                   05/12/2007
KY_CD                    106
OFNS_DESC                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
```

- Unique complaint number: CMPLNT_NUM
- Complaint Date/Time: CMPLNT_FR_DT, CMPLNT_FR_TM
- Location Info: Latitude, Longitude, BORO_NM
- High level classification (Law Category Code with values as “Felony”, “Misdemeanor” and “Violation”) : LAW_CAT_CD
- Other granular data classifying by other types: KY_CD (74 unique values), OFNS_DESC (71 values), PD_CD (427 values), PD_DESC (415 values), PREM_TYP_DESC (73 values)
- Suspect and Victim demographic info: SUSP_AGE_GROUP, SUSP_RACE , SUSP_SEX , VIC_AGE_GROUP , VIC_RACE , VIC_SEX

Data – How does it look?

STAATLICH
ANERKANNTE
HOCHSCHULE

Final Output CSV File used for Tableau: 110 rows with around 450 columns in all

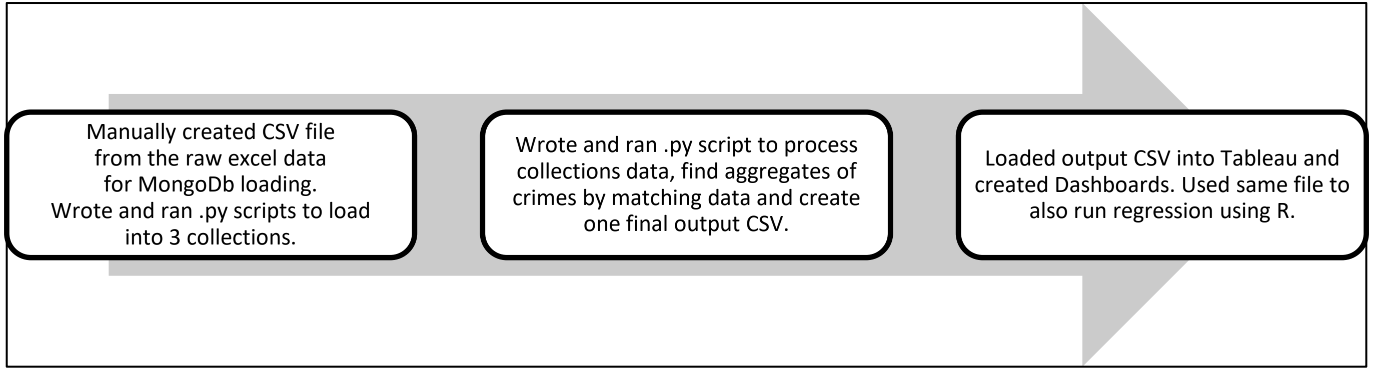


year	month	year-Month	boroName	countPedes	portNY	countPortNY	portNewark	countPortNewark	portNYnNewark	countPortNYnNewark	crimeTotal	crimeOFNS_DESC1	sumCrimeOFNS_DESC1
2007	5	2007-5	Bronx	77706	new york	306510	newark	118361	NYnNewark	424871	10684	ABORTION	0
2007	9	2007-9	Bronx	75129	new york	326364	newark	123756	NYnNewark	450120	10438	ABORTION	0
2008	5	2008-5	Bronx	67243	new york	377442	newark	165324	NYnNewark	542766	9902	ABORTION	0
2008	9	2008-9	Bronx	71054	new york	367684	newark	163849	NYnNewark	531533	9523	ABORTION	0
2009	5	2009-5	Bronx	65464	new york	329234	newark	146635	NYnNewark	475869	9855	ABORTION	0
2009	9	2009-9	Bronx	75254	new york	366058	newark	155199	NYnNewark	521257	9850	ABORTION	0
2010	5	2010-5	Bronx	82846	new york	372983	newark	177437	NYnNewark	550420	9645	ABORTION	0
2010	9	2010-9	Bronx	76831	new york	382850	newark	156992	NYnNewark	539842	9831	ABORTION	0
2011	5	2011-5	Bronx	86557	new york	406062	newark	167589	NYnNewark	573651	9527	ABORTION	0
2011	9	2011-9	Bronx	80592	new york	423843	newark	156827	NYnNewark	580670	8931	ABORTION	0
2012	5	2012-5	Bronx	65544	new york	439141	newark	157741	NYnNewark	596882	9502	ABORTION	0
2012	9	2012-9	Bronx	79209	new york	464837	newark	156662	NYnNewark	621499	9070	ABORTION	0
2013	5	2013-5	Bronx	89085	new york	499233	newark	171096	NYnNewark	670329	8871	ABORTION	0
2013	9	2013-9	Bronx	79435	new york	513551	newark	159536	NYnNewark	673087	9061	ABORTION	0
2014	5	2014-5	Bronx	83543	new york	560507	newark	171411	NYnNewark	731918	9377	ABORTION	0
2014	9	2014-9	Bronx	83242	new york	561774	newark	167685	NYnNewark	729459	9426	ABORTION	0
2015	5	2015-5	Bronx	85773	new york	626460	newark	166464	NYnNewark	792924	9391	ABORTION	0
2015	9	2015-9	Bronx	83543	new york	648840	newark	160845	NYnNewark	809685	9519	ABORTION	0
2016	5	2016-5	Bronx	85328	new york	607725	newark	166420	NYnNewark	774145	9495	ABORTION	0
2016	9	2016-9	Bronx	83247	new york	620518	newark	166933	NYnNewark	787451	8712	ABORTION	0
2017	5	2017-5	Bronx	84869	new york	581461	newark	160263	NYnNewark	741724	8936	ABORTION	0
2017	9	2017-9	Bronx	86878	new york	602459	newark	164457	NYnNewark	766916	8858	ABORTION	0
2007	5	2007-5	Brooklyn	179375	new york	306510	newark	118361	NYnNewark	424871	13848	ABORTION	0
2007	9	2007-9	Brooklyn	183456	new york	326364	newark	123756	NYnNewark	450120	13663	ABORTION	0
2008	5	2008-5	Brooklyn	184502	new york	377442	newark	165324	NYnNewark	542766	13578	ABORTION	0

Motivation

- Try to merge disparate data from multiple sources for analysis of a comprehensive dataset.
- Learn to handle large files in projects and effective querying with MongoDb
- If a prediction model for crime analysis could be found, that could be useful for public safety and law enforcement

ARCHITECTURE


Manually created CSV file from the raw excel data for MongoDb loading. Wrote and ran .py scripts to load into 3 collections.

Wrote and ran .py script to process collections data, find aggregates of crimes by matching data and create one final output CSV.

Loaded output CSV into Tableau and created Dashboards. Used same file to also run regression using R.

- Analyse raw data, prepare and convert into CSV after cleanup. Load with Python scripts.
- Retrieve the data from three Mongo collections and aggregate into a suitably structured CSV file to use later on.
  - Output CSV file has 110 rows, 450 columns.
  - Took almost 24 hours to run the Python script to create the data.
    Many more columns could have been extracted but kept crashing due to memory and run time issues.
- Used Tableau for visualisation by processing the CSV output at previous stage.

Task

- Hypothesis testing:
  - 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.
  - 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.
  - 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.

- Exploratory analysis:
  - Visualize the change in the number of pedestrians, the number of arrivals and the number of the crimes for months of May and September from 2007 to 2017.
  - Analyse the data in other suitable ways.

VISUALIZATION TECHNIQUES

Techniques used in the project:

- Geographic (spatial)
- Bar graphs
- Line graphs
- Interactive dashboard
- Miscellaneous: R used for numerical analysis and more rigorous hypothesis testing.

Marks and Channels used:

- Marks: Line graph, Bar graph, Points, Annotations.
- Channels: Position, Length, Color, Dual Axis

Justification

- Spatial technique usage:
  - Doing borough wise analysis is slightly easier with birds eye view; and also makes analysis interesting for user.

- Bar graphs and Line graphs as time series data (11 years with 2 months in each year)

- Dashboard with multiple visuals juxtaposed for easier comparison.

RESULTS FROM R ANALYSIS

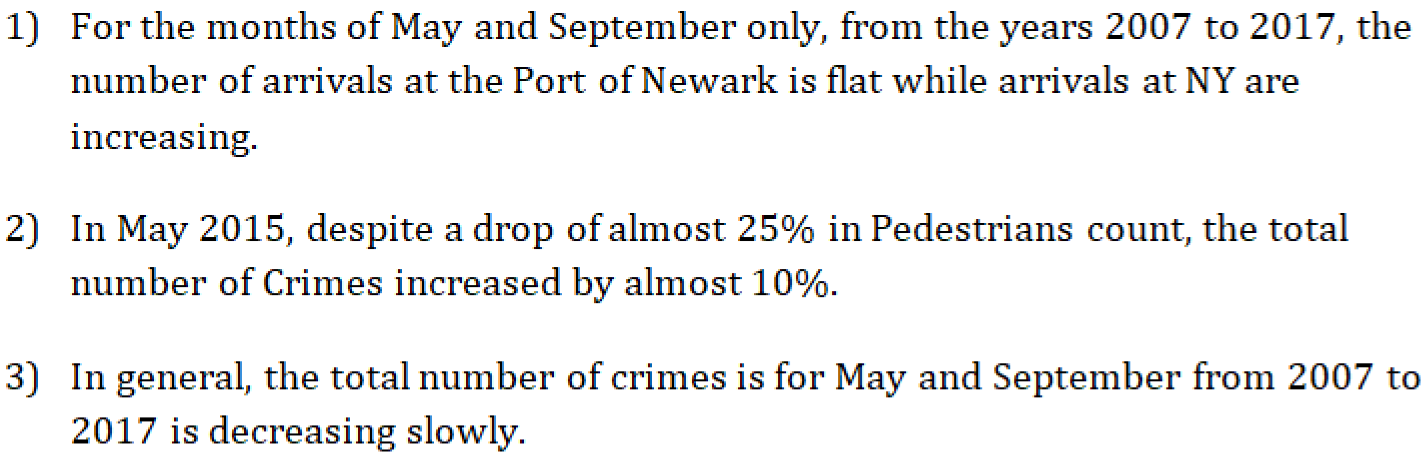
- Number of persons arriving at the Ports of New York and Newark together, **has no impact** on the Number of Pedestrians found in NY City.

- Regarding Total number of Crimes (all Types combined):
  - Pedestrian data effect: an increase of 5.94 in crimes for every 1000 increase in Pedestrians. However the data only explains 17% variation in the Total Crimes.
  - number of persons arriving at the Ports of New York and Newark together has no impact.

RESULTS FROM R ANALYSIS

- Regarding only Crimes with Premise Type as “Street”:
  - Multiple linear regression with both Ports arrival data and the Pedestrian counts together is better than simple linear.
  - Crime numbers increase by 1.26 for every 1000 increase in Pedestrians.
  - Crime decreases by 2.3 for every 1000 increase in the number of arrivals at the Ports of New York and Newark combined.

INSIGHTS FROM DASHBOARDS

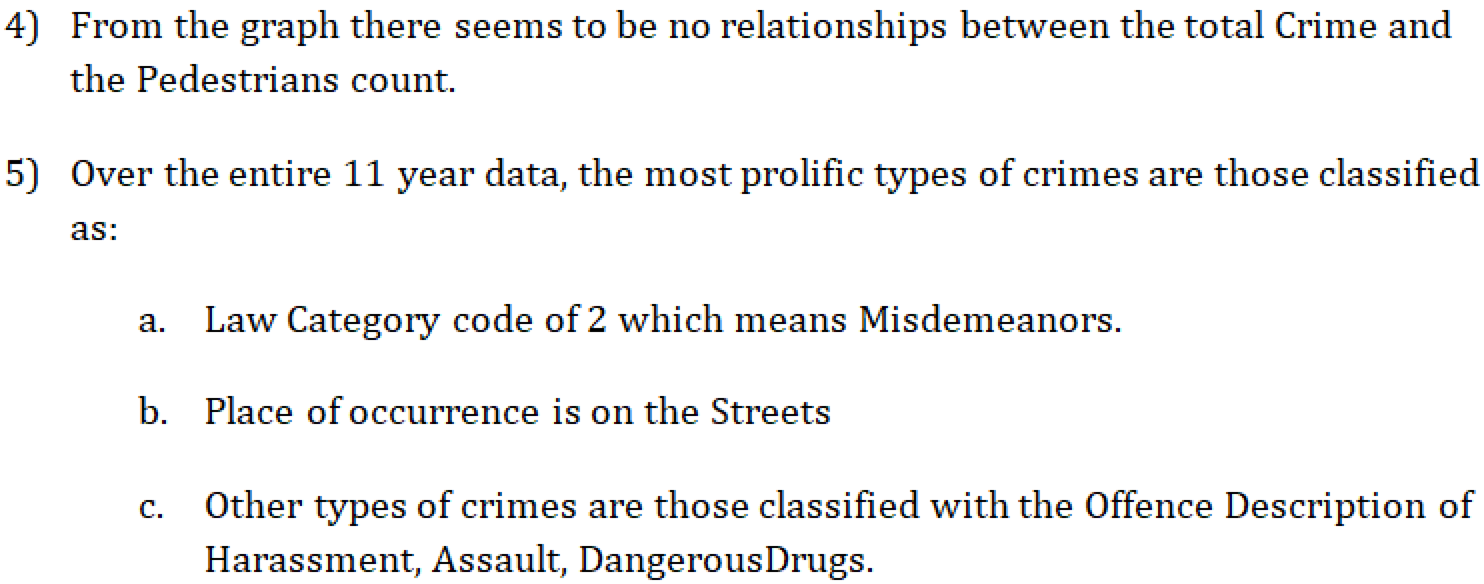


1) For the months of May and September only, from the years 2007 to 2017, the number of arrivals at the Port of Newark is flat while arrivals at NY are increasing.

2) In May 2015, despite a drop of almost 25% in Pedestrians count, the total number of Crimes increased by almost 10%.

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

INSIGHTS FROM DASHBOARDS



4) From the graph there seems to be no relationships between the total Crime and the Pedestrians count.

5) Over the entire 11 year data, the most prolific types of crimes are those classified as:

   a. Law Category code of 2 which means Misdemeanors.

   b. Place of occurrence is on the Streets

   c. Other types of crimes are those classified with the Offence Description of Harassment, Assault, DangerousDrugs.

DASHBOARD DEMO

Thank you for your attention. Questions?