Analysis of Crime & Poverty in Washington During 2017

Report

Student:

Pooja Nagrecha, Ryan Thomas, Ana Gill & Chris Krokus

Crime Dataset: Exploration

During the initial exploration of the Crime Dataset, the following steps were taken:

- Overview of dataset
  - Reviewing the initial condition of the dataset
  - Isolate the columns that possible contain meaningful data
  - Identify common data in both the Crime & Poverty Dataset
- Cleaning the data
  - Rename columns for easier data manipulation
  - Identifying any missing fields
  - Assigning fields to their appropriate data types
- Creating Visualizations
  - Investigate trends in crime over time
  - Investigate distribution of crime over a geographic space
  - Review the various types of offenses and weapons
- Preliminary Statistical Analysis
  - Confirm whether any of the data presents a normal distribution
  - Explore possible correlations within the dataset
  - Complete Chi-square Analysis

Overview of Dataset

After renaming the columns with more manageable names, the following summary of the dataset was available:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33082 entries, 0 to 33081
Data columns (total 23 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   CCN              33082 non-null  int64  
 1   Report           33082 non-null  object 
 2   Shift            33082 non-null  object 
 3   Method           33082 non-null  object 
 4   Offense          33082 non-null  object 
 5   Block            33082 non-null  object 
 6   XBlock           33082 non-null  int64  
 7   YBlock           33082 non-null  int64  
 8   Ward             33082 non-null  int64  
 9   ANC              33082 non-null  object 
 10  District         33079 non-null  float64
 11  PSA              33079 non-null  float64
 12  Neighborhood     32712 non-null  object 
 13  Block_Grou       32998 non-null  object 
 14  Census_Trac      32998 non-null  float64
 15  Voting_Precinct  33082 non-null  object 
 16  Latitude         33082 non-null  float64
 17  Longitude        33082 non-null  float64
 18  Bid              5830 non-null   object 
 19  Start_Date       33082 non-null  object 
 20  End_Date         31585 non-null  object 
 21  Object_ID        33082 non-null  int64  
 22  Octo_Recor       33082 non-null  object 
dtypes: float64(5), int64(5), object(13)
memory usage: 5.8+ MB
```

The following observations were made:

- Many fields represented geographical information (e.g. XBlock, YBlock, Ward, District, Neighborhood, Block Groc, Census Tract, Voting Precinct, Longitude & Latitude)
- With respect to time, Shift reported on whether crimes occurred in the morning or evening. More importantly, Start_Date and End_Date captured a timestamp of

when incidents took place. This two timestamps were originally typed as strings and would need to be modified to create time series plots.

- Method reported on the weapon used in committing the crime
- Offense detailed the various types of crime

Type of Crime	Frequency in 2017
THEFT/OTHER	8170
THEFT F/AUTO	5538
MOTOR VEHICLE THEFT	1692
ROBBERY	1494
ASSAULT W/DANGEROUS WEAPON	1405
BURGLARY	1025
SEX ABUSE	194
HOMICIDE	94
ARSON	3

This column features various types of theft and there was thought given to merging categories. Although there are THEFT F/AUTO and MOTOR VEHICLE THEFT, one may reference a theft involving a vehicle whereas the other is likely theft of a vehicle. Similarly, ROBBERY is theft of personal property and BURGLARY involves entering a building to commit theft. There was sufficient distinction to leave this categories independent.

- Census_Tract information was present in both datasets. This is a geographical area defined for the purpose of taking a census.

Cleaning the Data

Identify Missing Data

Referencing the Pandas output above, the dataframe contained **33082** rows. However, the highlighted fields contained some missing fields.

#	Column	Non-Null	Count	Dtype
0	CCN	33082	non-null	int64
1	Report	33082	non-null	object
2	Shift	33082	non-null	object
3	Method	33082	non-null	object
4	Offense	33082	non-null	object
5	Block	33082	non-null	object
6	XBlock	33082	non-null	int64
7	YBlock	33082	non-null	int64
8	Ward	33082	non-null	int64
9	ANC	33082	non-null	object
10	District	33079	non-null	float64
11	PSA	33079	non-null	float64
12	Neighborhood	32712	non-null	object
13	Block_Grou	32998	non-null	object
14	Census_Trac	32998	non-null	float64
15	Voting_Precinct	33082	non-null	object
16	Latitude	33082	non-null	float64
17	Longitude	33082	non-null	float64
18	Bid	5830	non-null	object
19	Start_Date	33082	non-null	object
20	End_Date	31585	non-null	object
21	Object_ID	33082	non-null	int64
22	Octo_Recor	33082	non-null	object

Columns Bid and End_Date were not used in the analysis. With respect remaining columns with missing data, an appropriate value was found to fill the missing fields. In most cases, this was the median() value.

Assign Appropriate Data Types

- Start_Date was converted to a timestamp using the following command

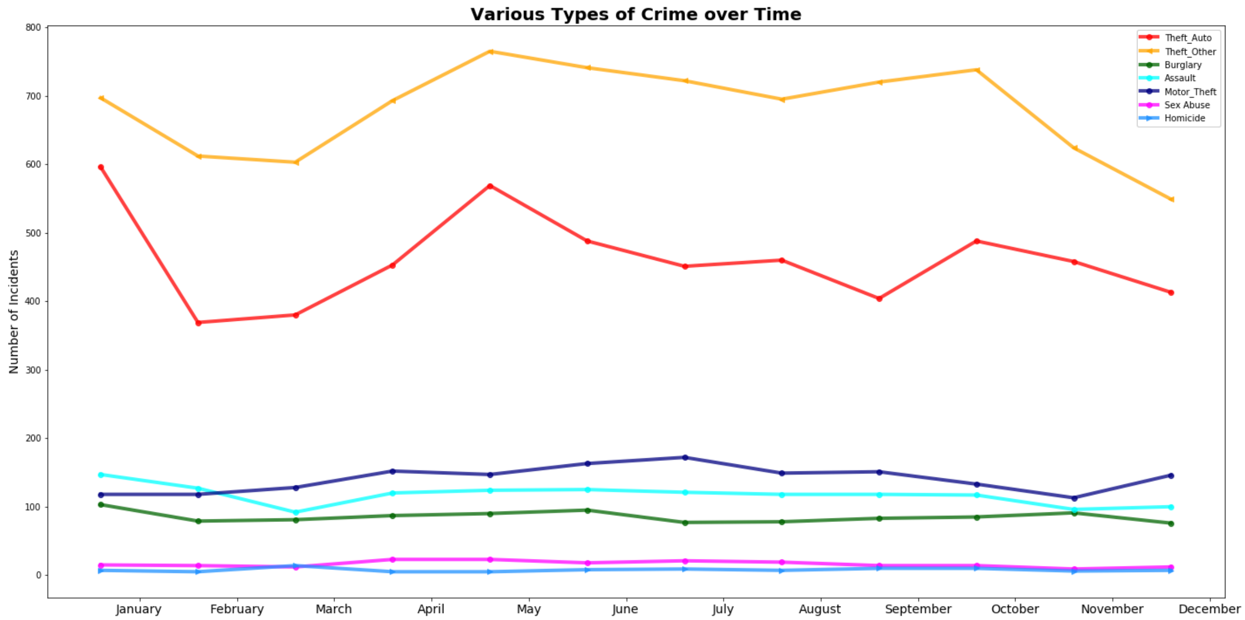
```
df['Start_Date'] = pd.to_datetime(df['Start_Date'], format='%Y-%m-%dT%H:%M:%S.%f')
```

Creating Visualizations

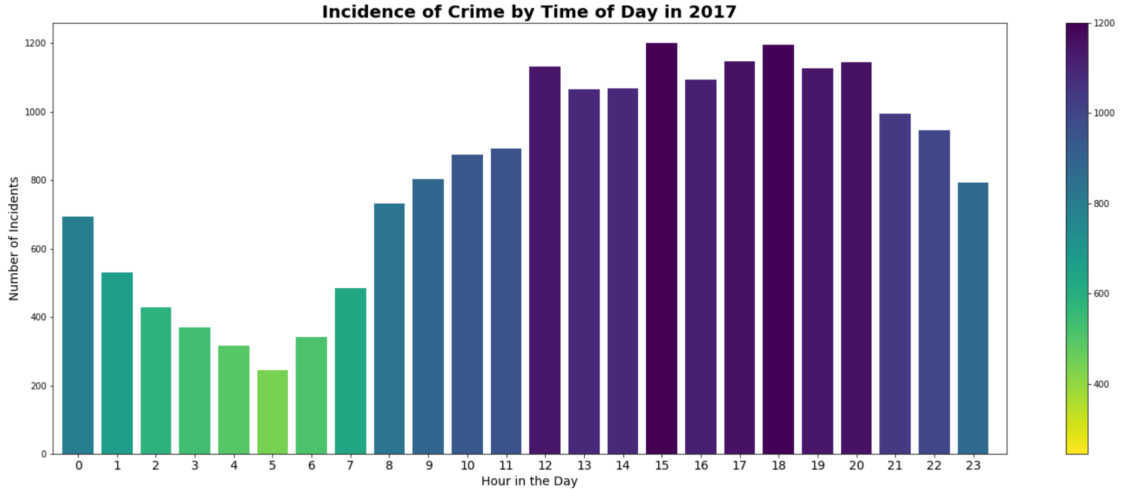
Investigate Trends in Crime over Time

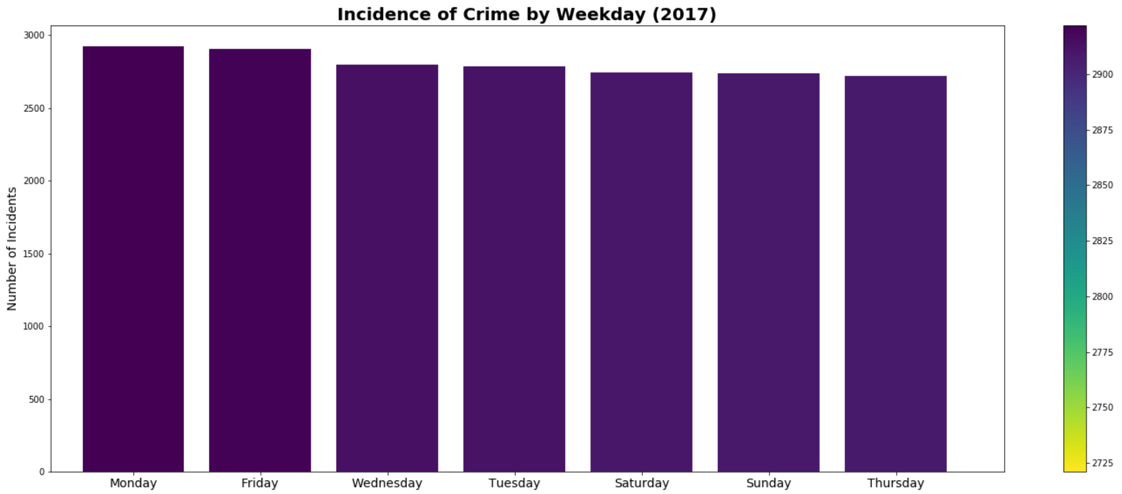
Using the timestamps in Start_Date, it was now possible to:

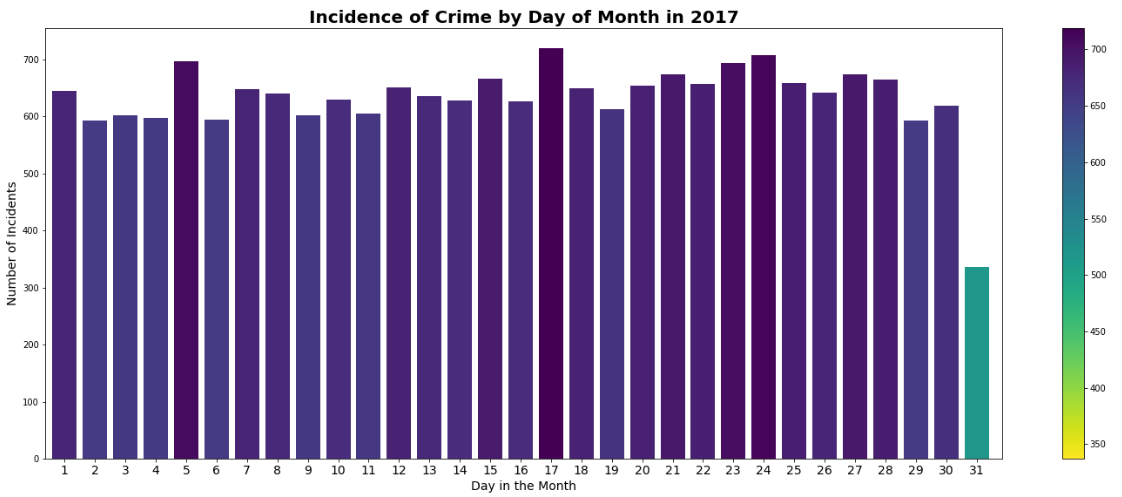
- create a time series of different types of crime over time

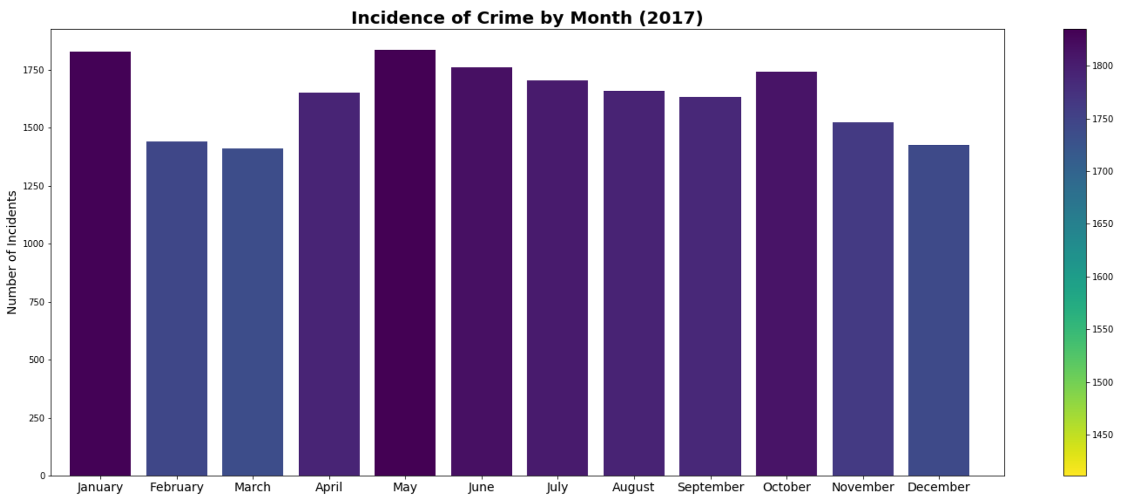


- aggregate incidence of crime over different periods (e.g. time of day, weekdays, month and year




Incidence of Crime by Weekday (2017)
Number of Incidents
Monday
Friday
Wednesday
Tuesday
Saturday
Sunday
Thursday


Incidence of Crime by Day of Month in 2017
Number of Incidents
Day in the Month

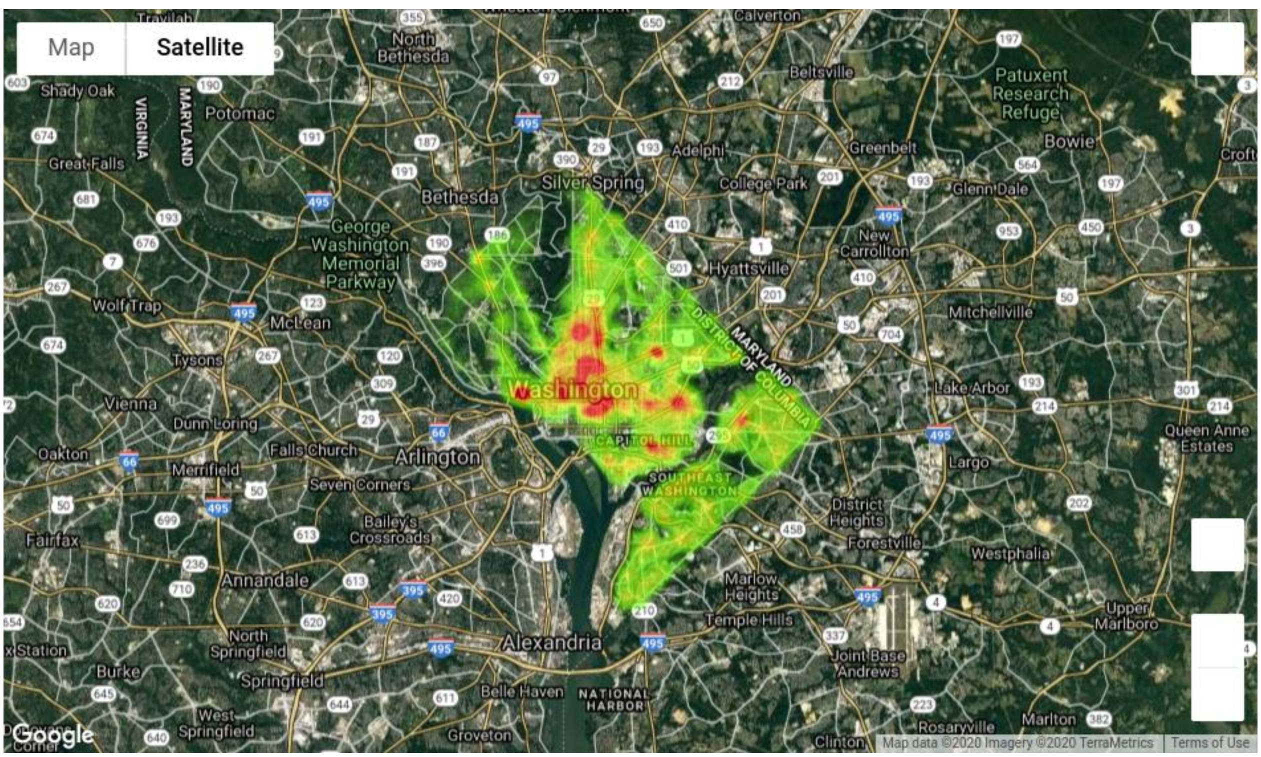

Incidence of Crime by Month (2017)
Number of Incidents
January
February
March
April
May
June
July
August
September
October
November
December

It was found that Homicide, Sex Abuse, Assault, Burglary and Motor Theft remainder fairly consistent throughout the year. However, Theft_Auto and Theft_Other demonstrated more range with peaks in January and April 2017.

With respect to looking at the total volume of crime during the year, the incidence of crime remains relatively steady during the week, Over the course of the month, only the 31st day illustrated a significant drop and this like due to only 7 months having 31 days. Reviewing variations in crime during the day, it was observed that most incidents took place between 3pm and 6pm.

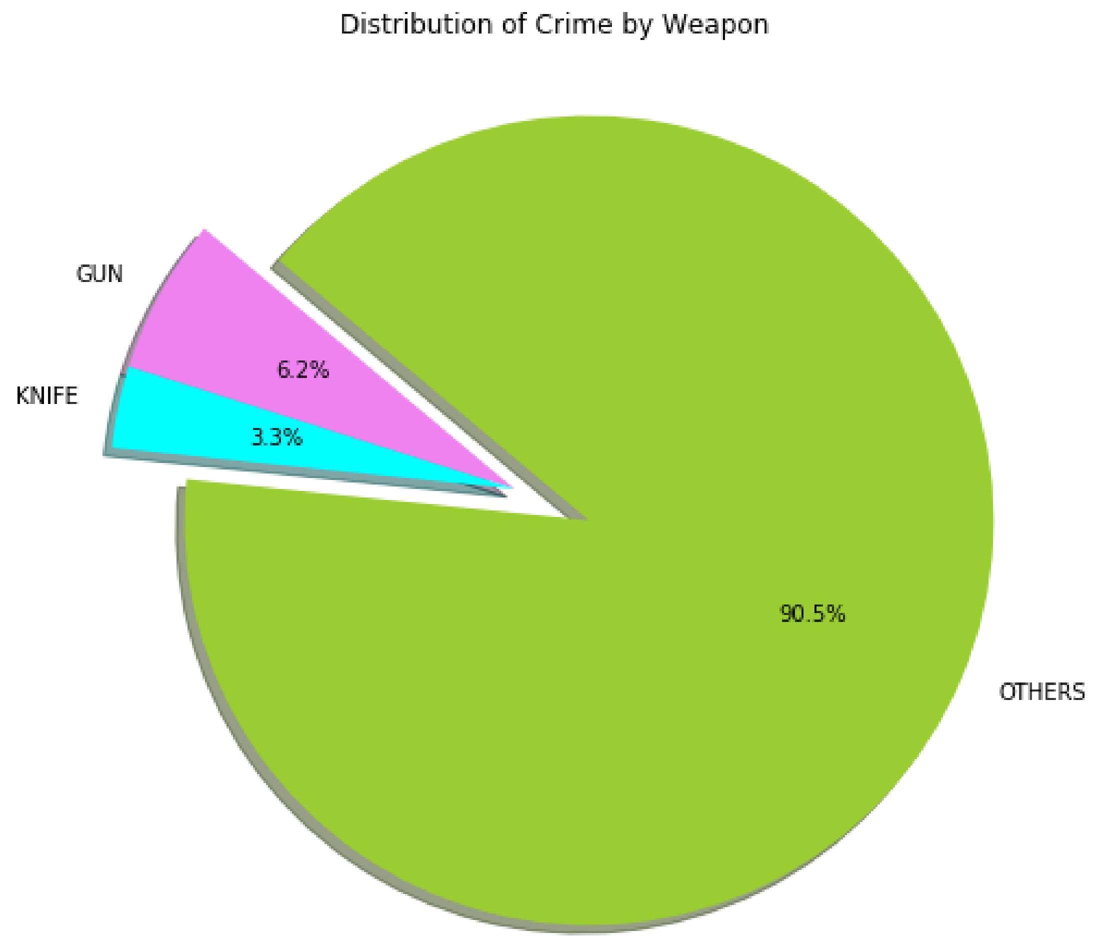
Investigate Distribution of Crime over Washington

Leveraging Latitude and Longitude fields in the dataset, a Google Maps heatmap was generated to illustrate the distribution of crime over the Washington area. It can be seen that high volumes are crime are concentrated on the city center.



Types of Weapons

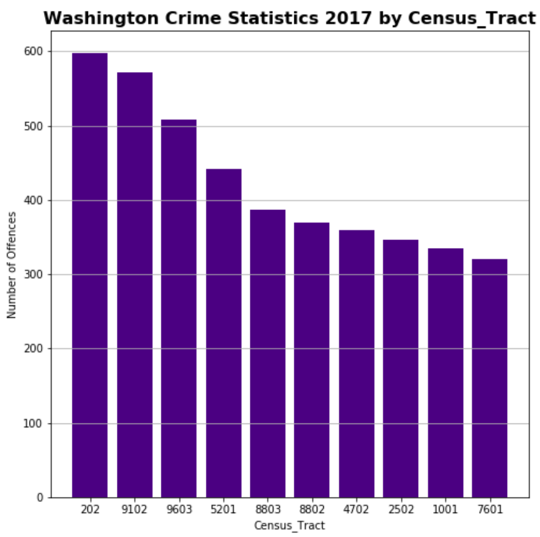
Method was not an insightful field. This datapoint lacked granularity with more than 90% of the data being uncategorized.

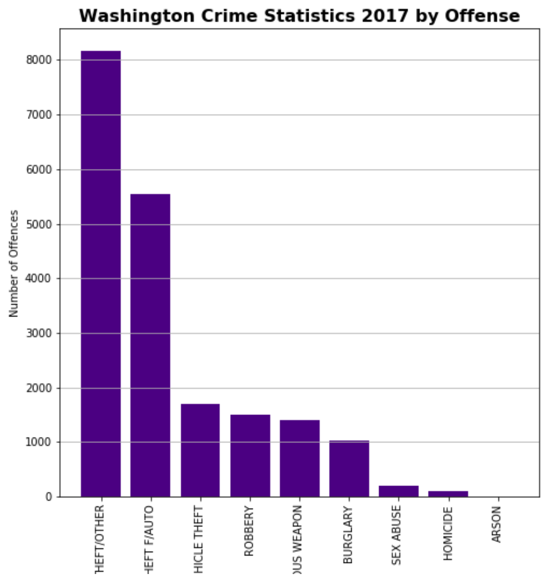


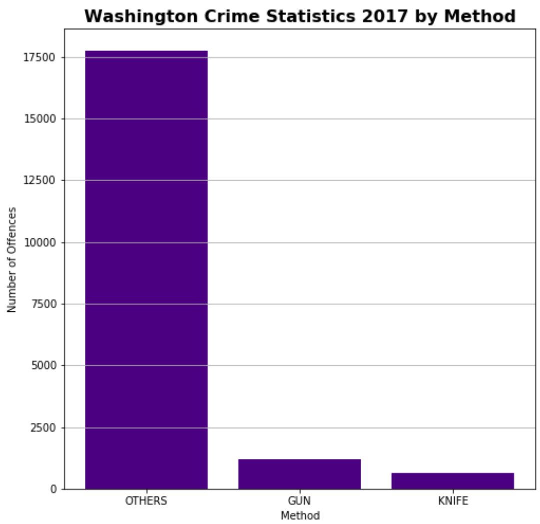
Bar Charts

Below are chart illustrating:

- top 10 census tracts
- total number of offenses by type
- total number of offenses by method



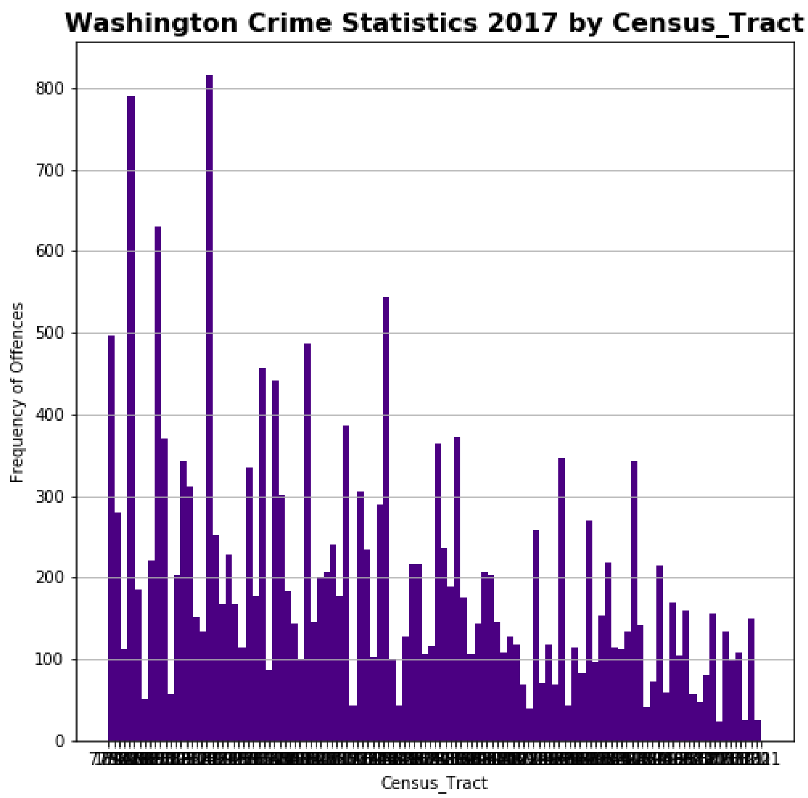




Preliminary Statistical Analysis

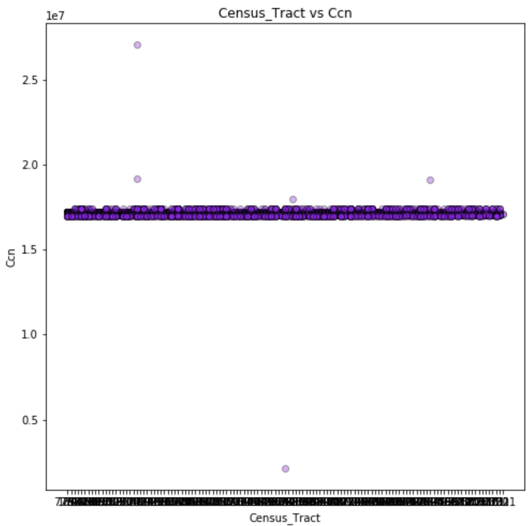
Histograms

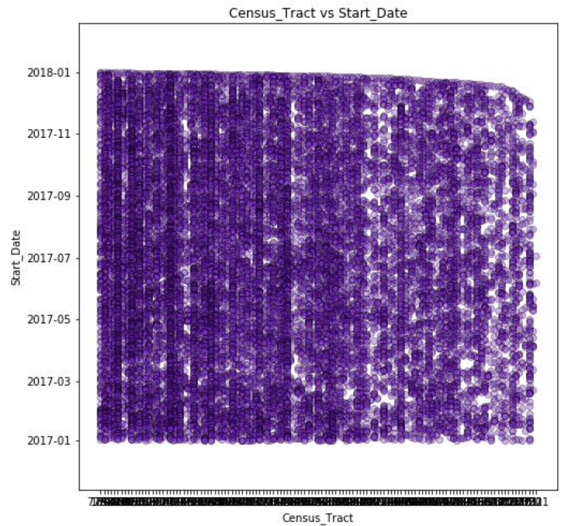
Histograms were generated for each field to confirm the nature of the distribution. None of the columns were found to demonstrate a normal distribution

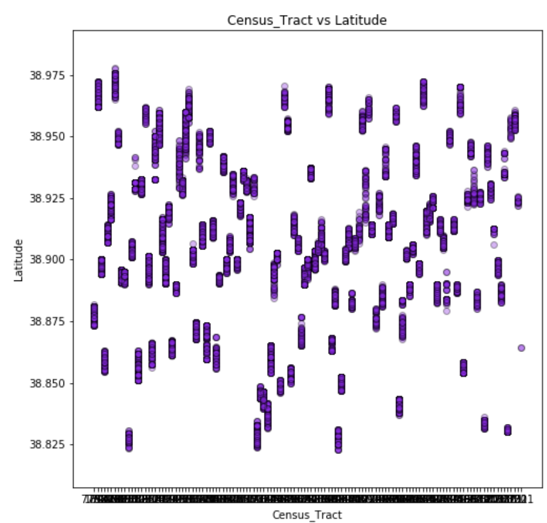


Scatter Plots

Various scatter plots were generated and did not highlight any correlation between columns.



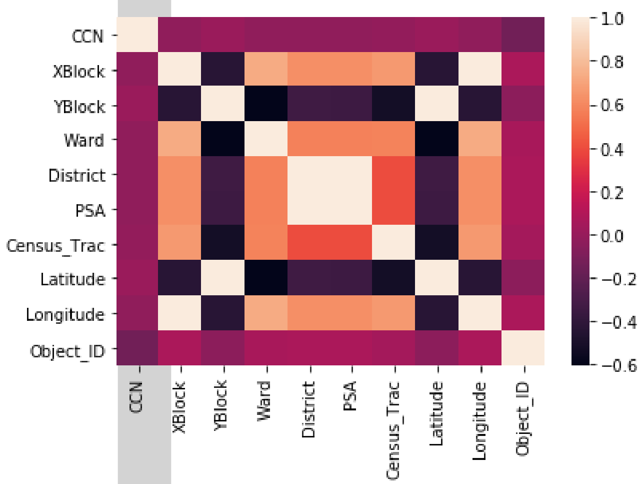




Correlation

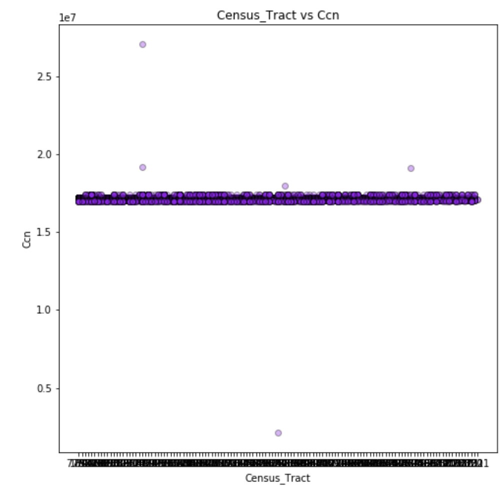
As highlighted previously, a large portion of the dataset represented geographic information. No meaning correlation was observed.

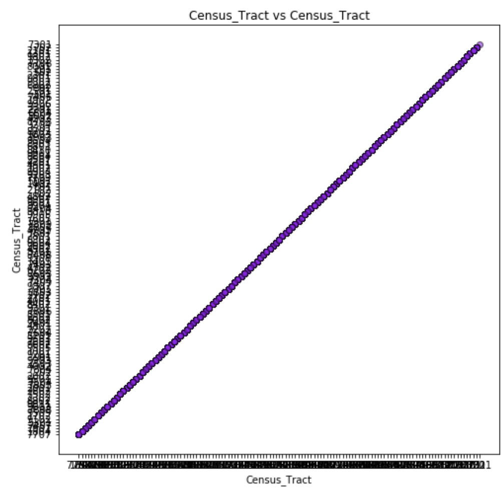
CCN	XBlock	YBlock	Ward	District	PSA	Census_Trac	Latitude	Longitude	Object_ID	Column1
CCN	1	-0.023778	0.013871	-0.027237	-0.023326	-0.023277	-0.015033	0.013869	-0.023775	-0.137417
XBlock	-0.023778	1	-0.431536	0.731881	0.627243	0.627632	0.662248	-0.431537	1	0.077595
YBlock	0.013871	-0.431536	1	-0.604422	-0.339318	-0.342631	-0.521944	1	-0.431603	-0.037003
Ward	-0.027237	0.731881	-0.604422	1	0.576764	0.57667	0.582619	-0.604424	0.731845	0.062215
District	-0.023326	0.627243	-0.339318	0.576764	1	0.999924	0.398355	-0.339336	0.627175	0.075858
PSA	-0.023277	0.627632	-0.342631	0.57667	0.999924	1	0.399889	-0.342648	0.627565	0.075729
Census_Trac	-0.015033	0.662248	-0.521944	0.582619	0.398355	0.399889	1	-0.521892	0.662361	0.051251
Latitude	0.013869	-0.431537	1	-0.604424	-0.339336	-0.342648	-0.521892	1	-0.431604	-0.037001
Longitude	-0.023775	1	-0.431603	0.731845	0.627175	0.627565	0.662361	-0.431604	1	0.077593
Object_ID	-0.137417	0.077595	-0.037003	0.062215	0.075858	0.075729	0.051251	-0.037001	0.077593	1

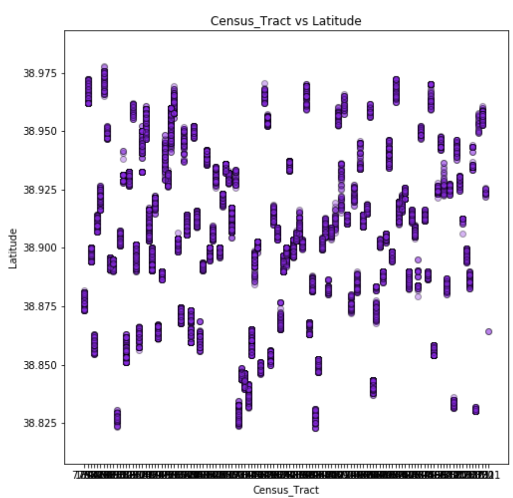


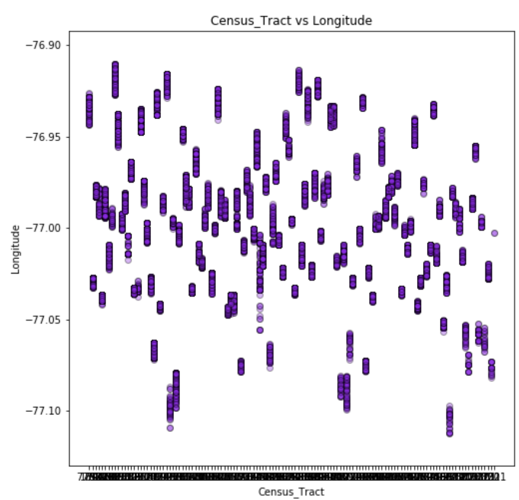
Appendix

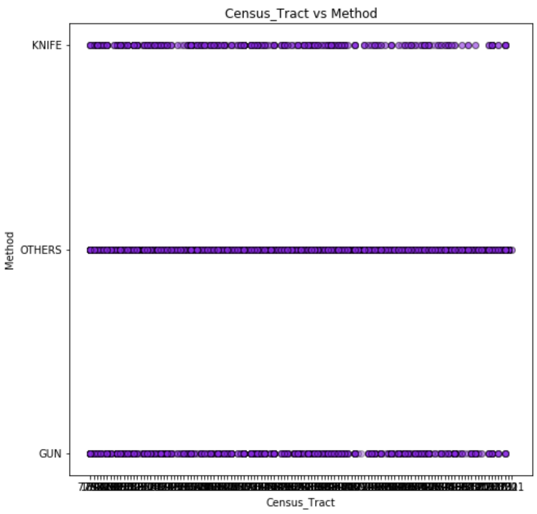
Scatter Plots

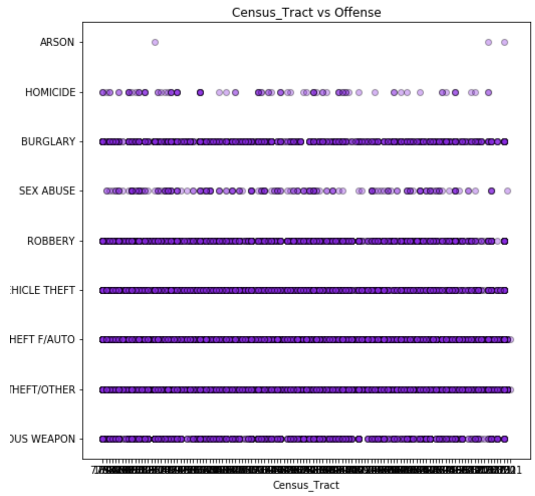


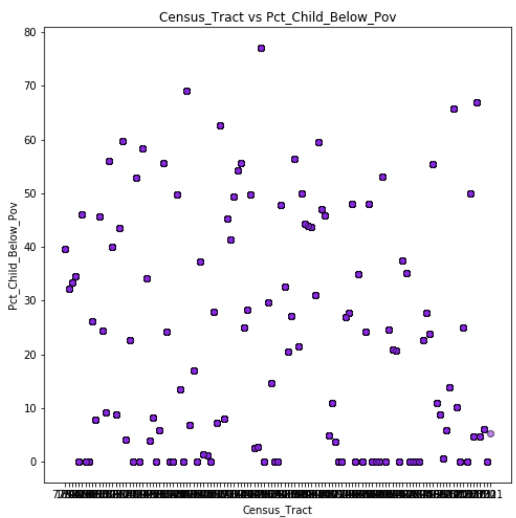


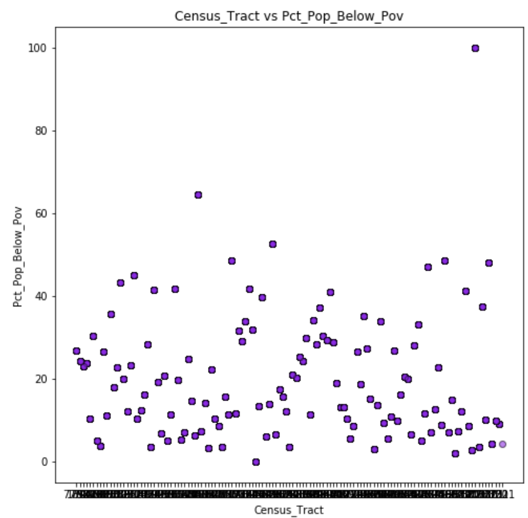


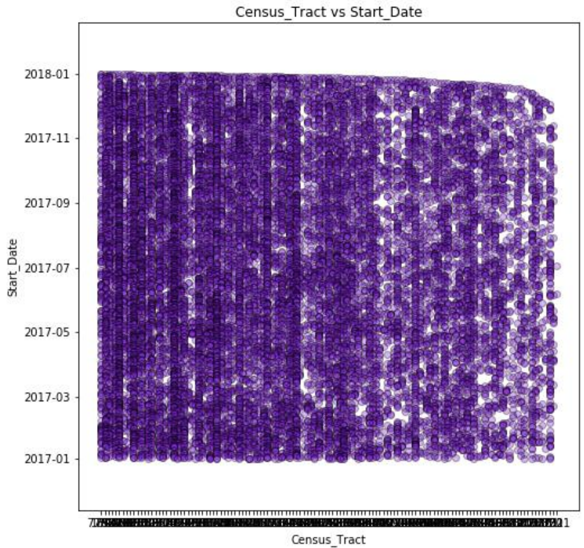



Census_Tract vs Method
KNIFE
OTHERS
GUN
Method
Census_Tract

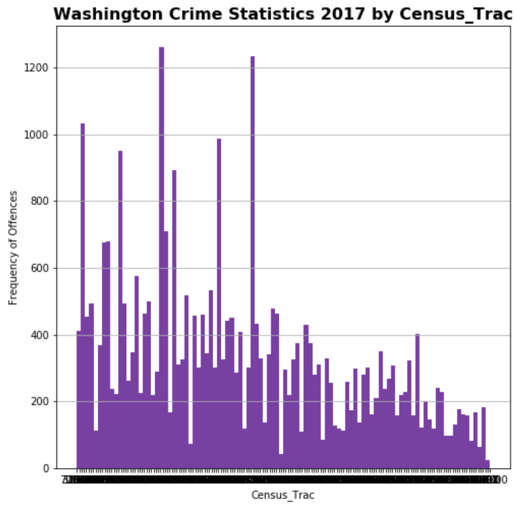

Census_Tract vs Offense
ARSON
HOMICIDE
BURGLARY
SEX ABUSE
ROBBERY
HICLE THEFT
HEFT F/AUTO
HEFT/OTHER
US WEAPON
Census_Tract


Census_Tract vs Pct_Child_Below_Pov
0
10
20
30
40
50
60
70
80
Pct_Child_Below_Pov
Census_Tract

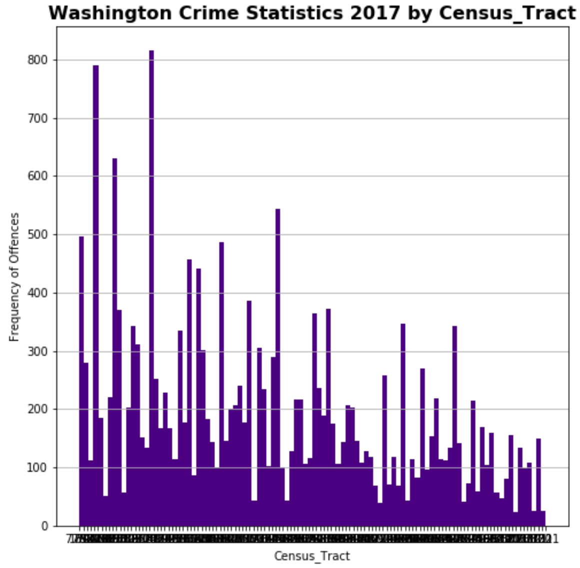

Census_Tract vs Pct_Pop_Below_Pov
0
20
40
60
80
100
Pct_Pop_Below_Pov
Census_Tract


Census_Tract vs Start_Date
Start_Date
2018-01
2017-11
2017-09
2017-07
2017-05
2017-03
2017-01
Census_Tract

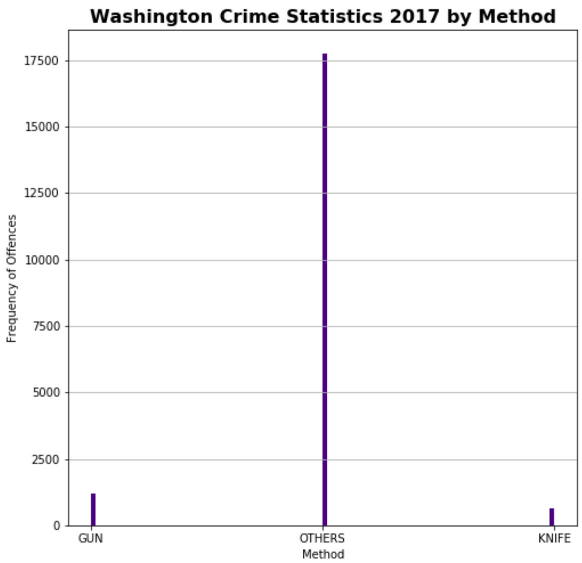
Histograms



Washington Crime Statistics 2017 by Census_Trac
Frequency of Offences
Census_Trac

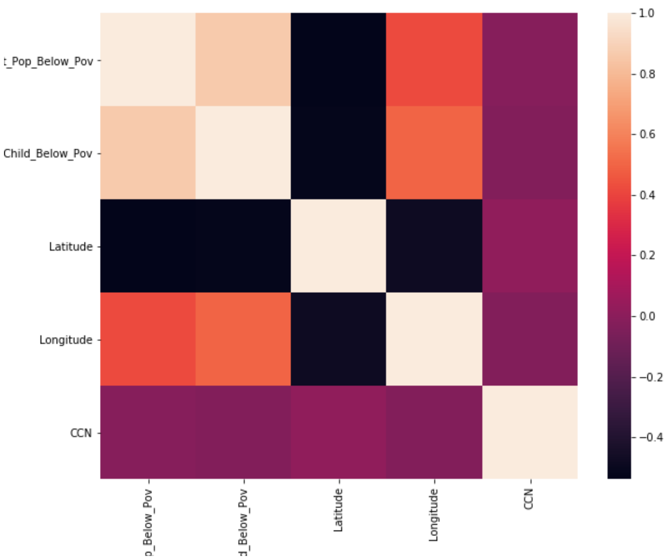


Washington Crime Statistics 2017 by Census_Tract
Frequency of Offences
Census_Tract



Washington Crime Statistics 2017 by Method
Frequency of Offences
GUN
OTHERS
KNIFE
Method

Heatmap



t_Pop_Below_Pov
Child_Below_Pov
Latitude
Longitude
CCN
p_Below_Pov
d_Below_Pov
Latitude
Longitude
CCN
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4