Big Data Analytics – Project Report
Anubhav Reddy N
axn155530

Objective: To analyze the police incidents report from https://www.dallasopendata.com/ using big data tools such as hive, pig and spark and identify patterns in the incidents reports.

## Preprocessing of the data

- Download and clean the raw data file from https://www.dallasopendata.com/
- Shortlist the columns required for analysis
- Load the data into Hadoop

The steps and code to achieve the above objectives are provided in the file – "Police_Incidents_Cleaning_Steps.pdf"

- Create a folder called Project and move the file Police.tsv into this folder. Then load the folder into hive.
- Navigate to folder named "project" and execute hive file "police_text.hive" to load the data into hive table

## Processing Data – Hive query

- Number of incidents per year – run file "incidents_per_year.hive" – save to incidents_per_year.csv
- Number of incidents by UCR offence description by year – run file "UCR_offense_incidents.hive" – save to UCR_offense_incidents.csv
- Number of incidents by Beat, year – run file "beat_incidents_per_year.hive" – save to beat_incidents_per_year.csv
- Number of incidents by Division, year – run file "div_incidents_per_year.hive" – save to div_incidents_per_year.csv
- Number of incidents by Zip Code – run file "zip_incidents.hive" – save to zip_incidents.csv
- Number of incidents by Complainant race != TEST, Complainant gender = F, year != 2014 – run file "race_genderF.hive" – save to race_genderF.csv
- Number of incidents by Complainant race != TEST, Complainant gender = M, year != 2014 – run file "race_genderM.hive" – save to race_genderM.csv
- Modus operandi of incidents – run file "modus_operandi.hive" – save to "modus_operandi.txt"
- Type of incident – run file "type_incident.hive" – save to "type_incident.txt"
- Type of location – run file "type_location.hive" – save to "type_location.txt"

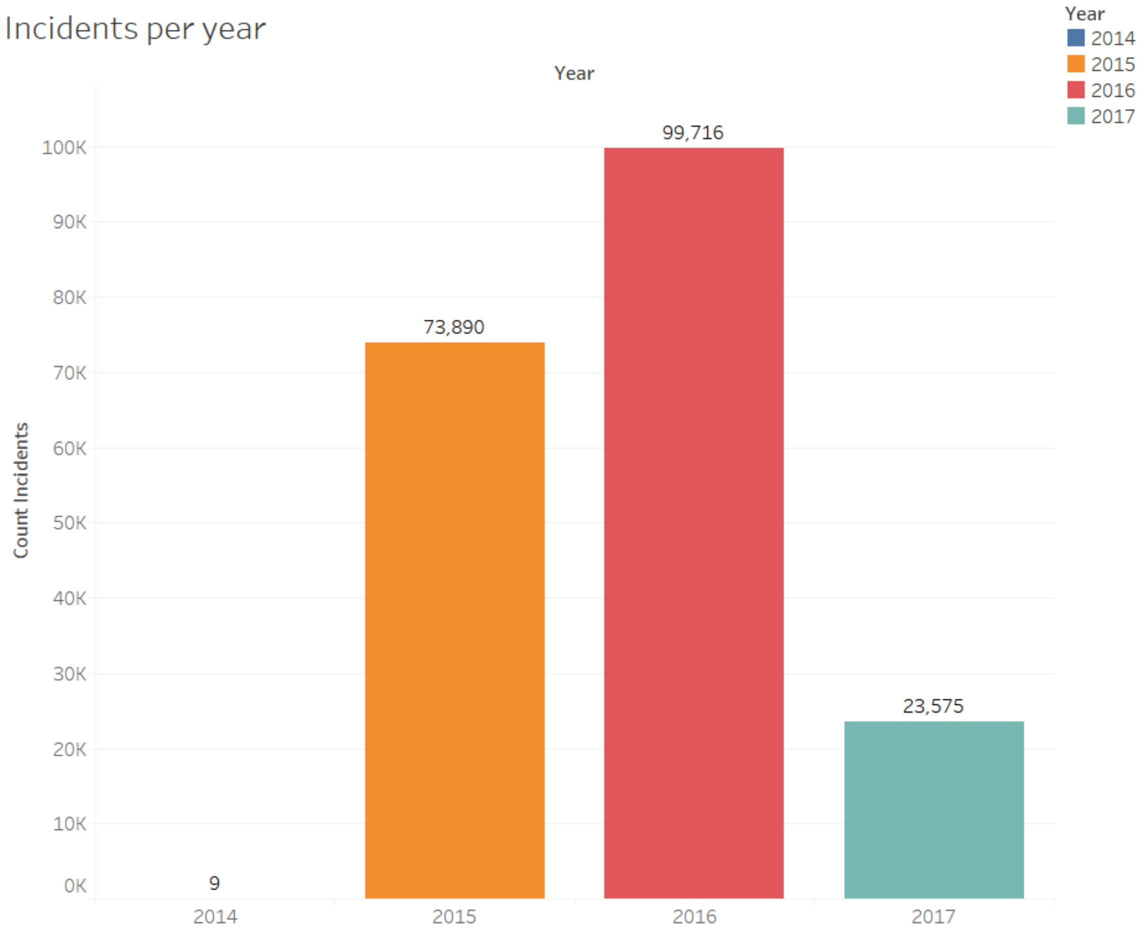
***All code file, data files and Tableau workbook will be attached with the project report document to help reproduce this work.

**Visualizing the data – Tableau**

**Data File:** incidents_per_year.txt

Import all the data files extracted from Hadoop into tableau to create visualizations in order to understand the data and its distribution.

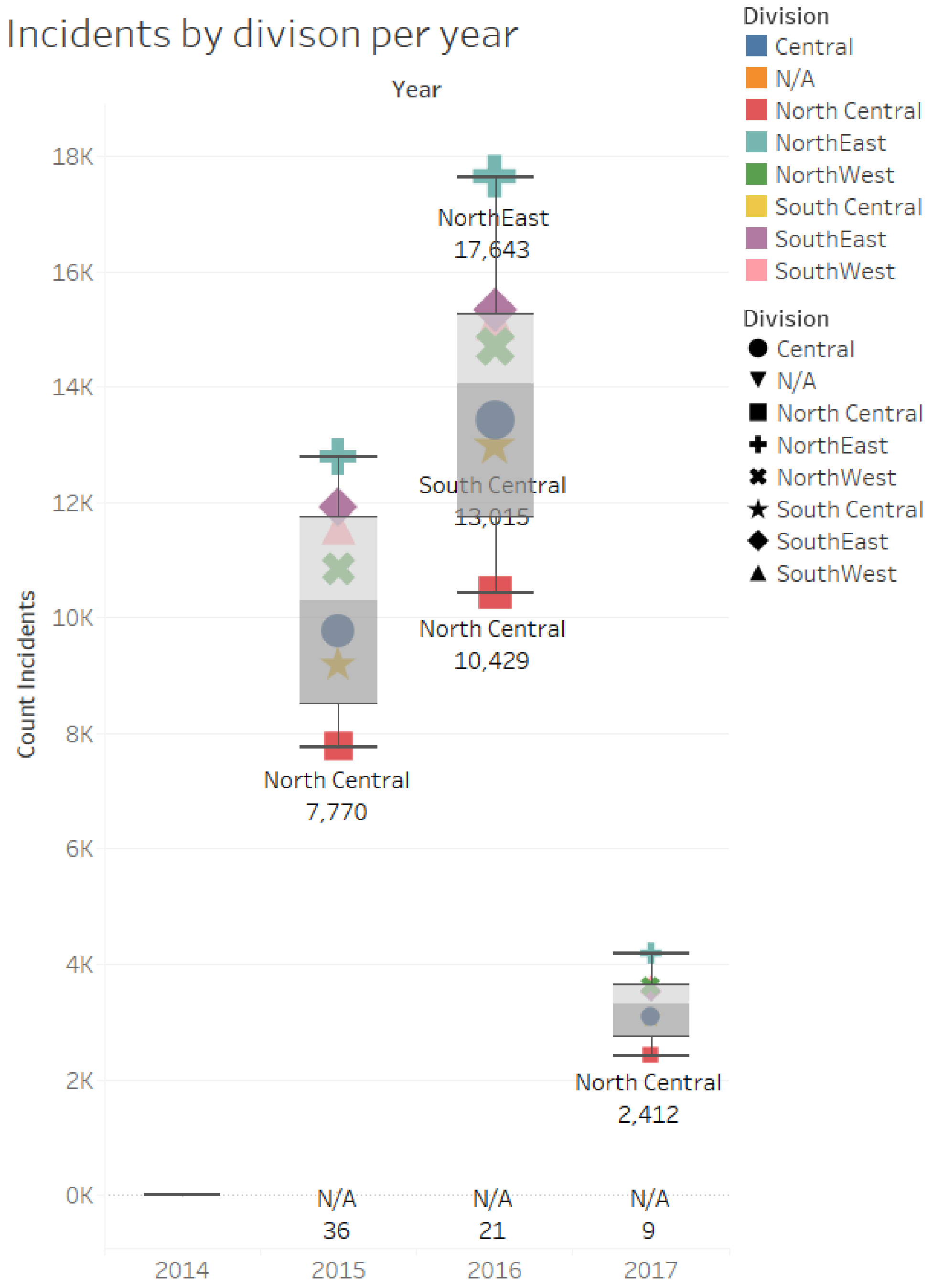
**Incidents per year:**



The incidents per year show the distribution of incidents over the years. We can see that crime in Dallas has increased from 2015 to 2016.

**Incidents by Division per Year**

**Data File:** div_incidents_per_year.csv

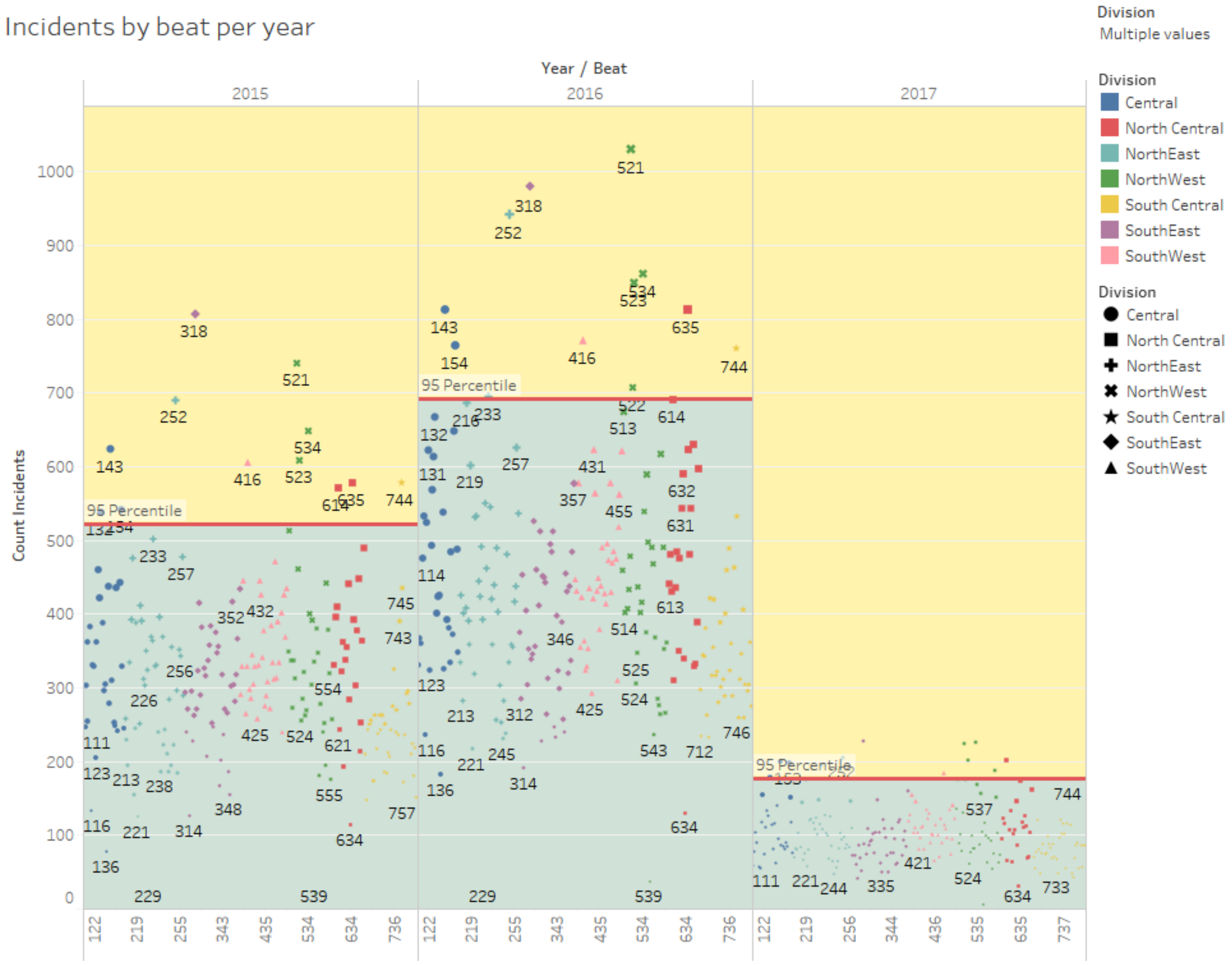


The chart below reflects a box plot created to understand the distribution of crime over the police divisions in Dallas. The North East, South East and South West have largest incidents of crime over the year 2015 and 2016. North Central and South Central continue to report less number of incidents.

## **Incidents by Beat per Year**

**Data File:** beat_incidents_per_year.csv

The chart below shows that some beats respond to more incidents than other beats. Beats 318, 521, 252, 534, 523, 143 have consistently responded to more incidents than other beats i.e. in the 95th percentile. The Dallas police department should look into inserting more Police officers for these beats.

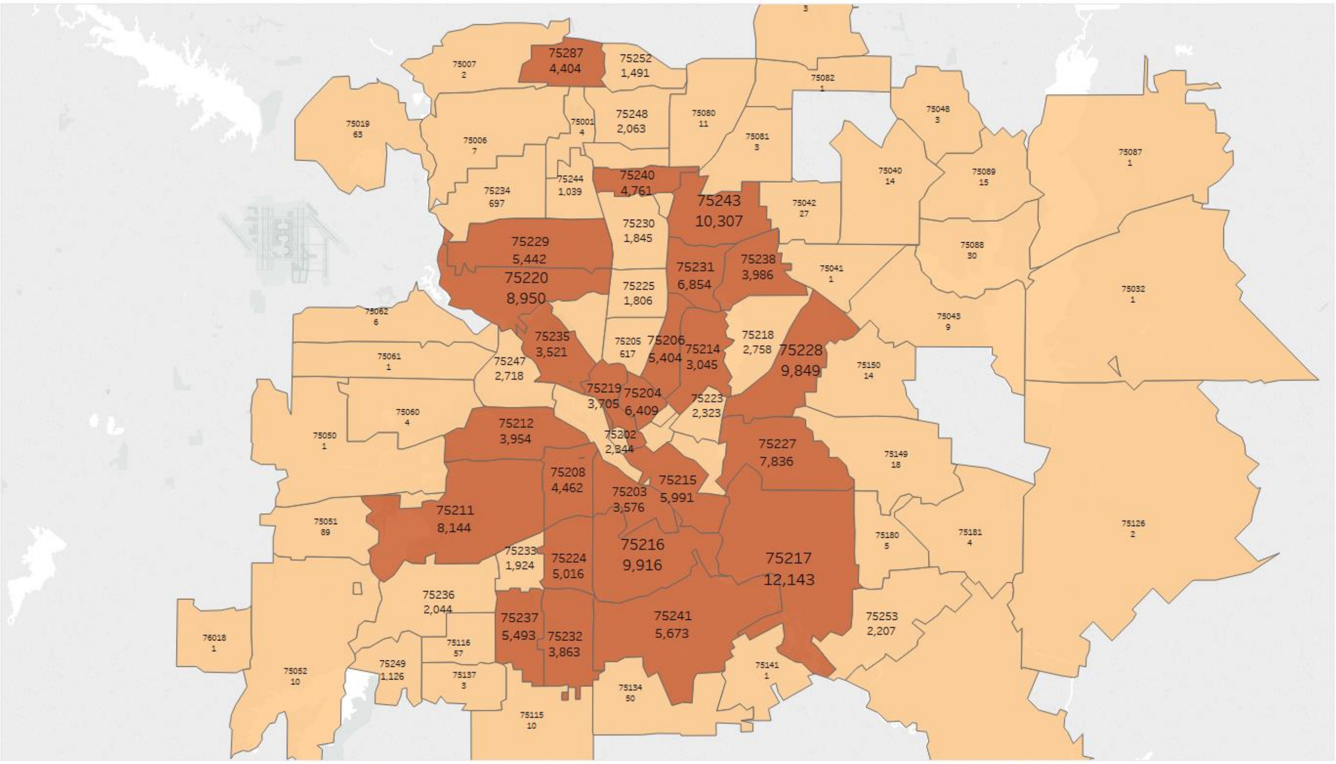


## Incidents by Zip Code

**Data File:** zip_incidents.csv

The Map below shows the distribution of incidents by Zip Codes on a map of Dallas. We can see the zip codes in Brown color report extremely high incidents compared to other zip codes. People living in these zip codes are should be more susceptible to crime and need to be more cautious.

Incidents by Zipcode



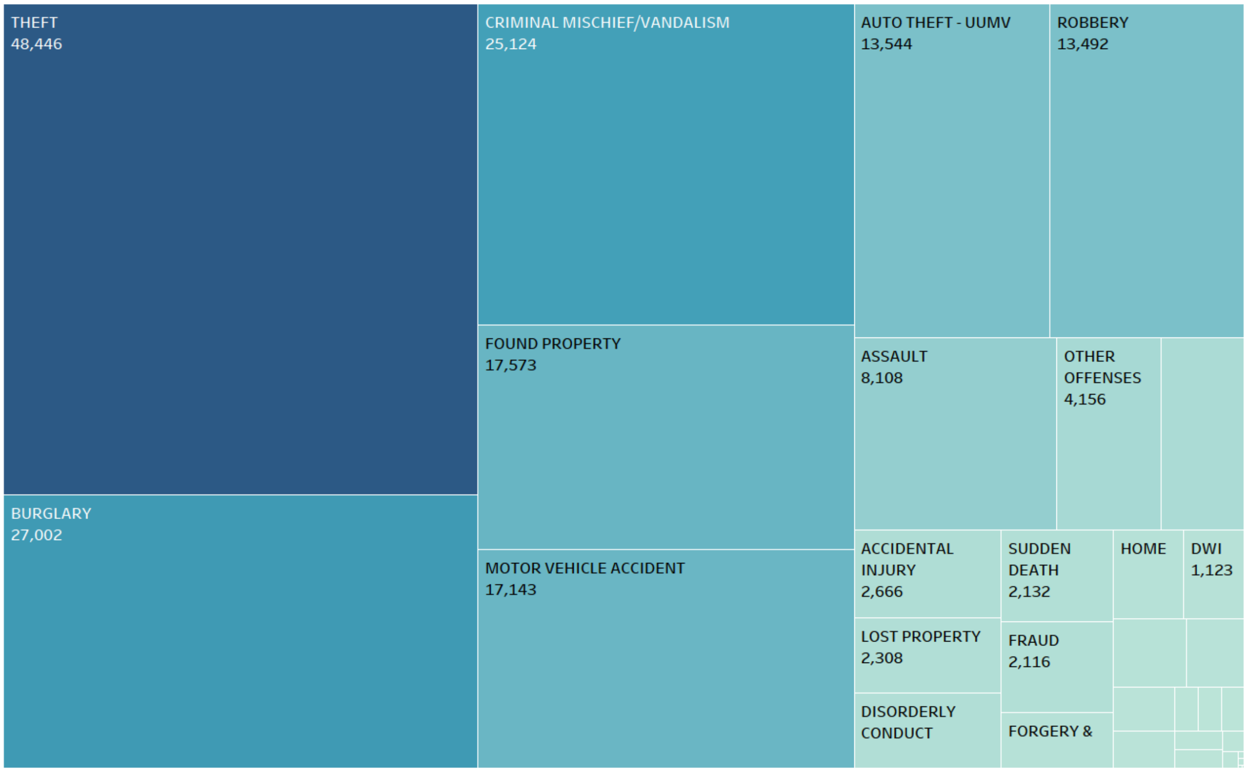
Map based on Longitude (generated) and Latitude (generated). Color shows details about High Incidents Flag. Size shows sum of Count Incidents. The marks are labeled by Zip Code and sum of Count Incidents. Details are shown for Zip Code. The data is filtered on Year, which keeps 2015, 2016 and 2017. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

## Incidents by UCR offense description

**Data File:** UCR_offense_incidents.csv

Most reports of incidents involve Theft, Burglary, Vandalism, Property Found and Motor Vehicle incident. Theft, Burglary and Robbery constitute approximately 50% of all the incidents reported in the year 2015, 2016 and 2017.

Incidents by UCR offense description

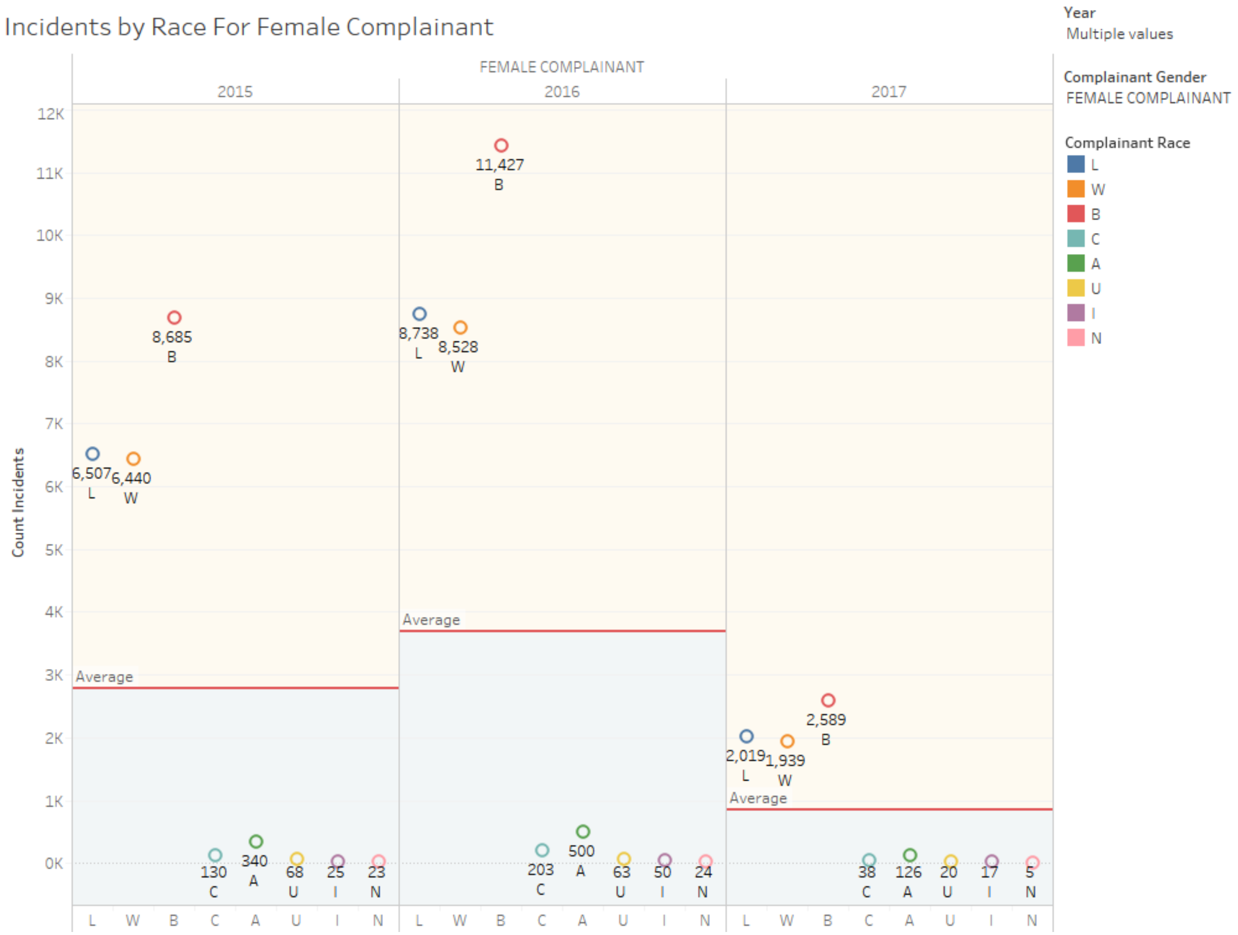


Ucr Offense Description and sum of Count Incidents. Color shows sum of Count Incidents. Size shows sum of Count Incidents. The marks are labeled by Ucr Offense Description and sum of Count Incidents.

**<u>Incidents by Race for Female Complainants:</u>**

**<u>Data File:</u>** race_genderF.csv

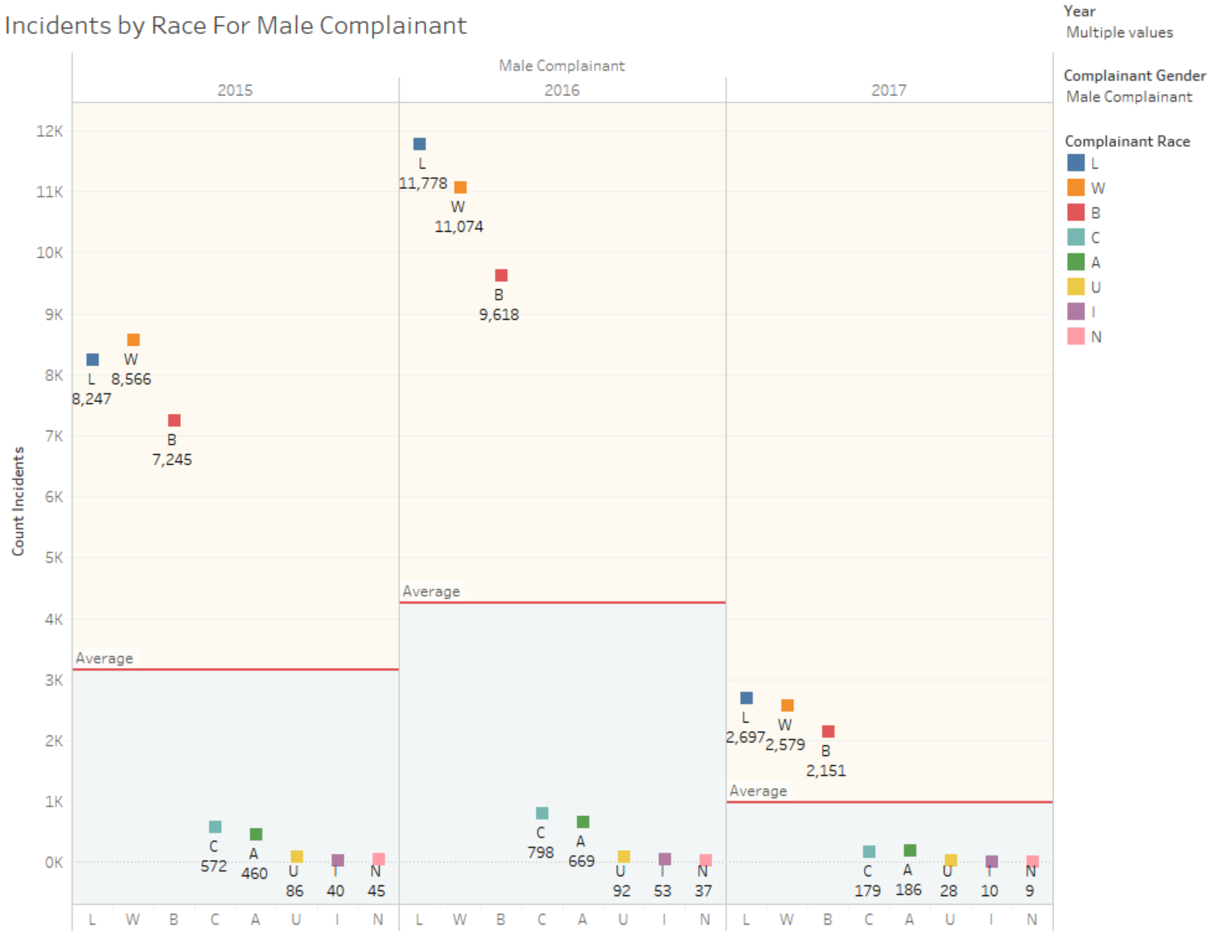
The chart below shows that the incidents reported by female complainants are mostly reported by Black / African American, Latino and White females. The population of Latino and African American is much less compared to White Americans we can say that Latino / Hispanic and African American / Black females are more vulnerable to criminal incidents than other people belonging to other races. This could be driven by economic and geographic factors which need to be investigated to arrive at more concrete understanding.



**<u>Incidents by Male Complainant</u>**

**<u>Data File:</u>** race_genderM.csv

The chart below shows that the incidents reported by male complainants are mostly reported by Latino, White and Black / African American males. The population of Latino and African American is much less compared to White Americans therefore, we can say that Latino / Hispanic and African American / Black males are more vulnerable to criminal incidents than other people belonging to other races. This could be driven by economic and geographic factors which need to be investigated to arrive at more concrete understanding.



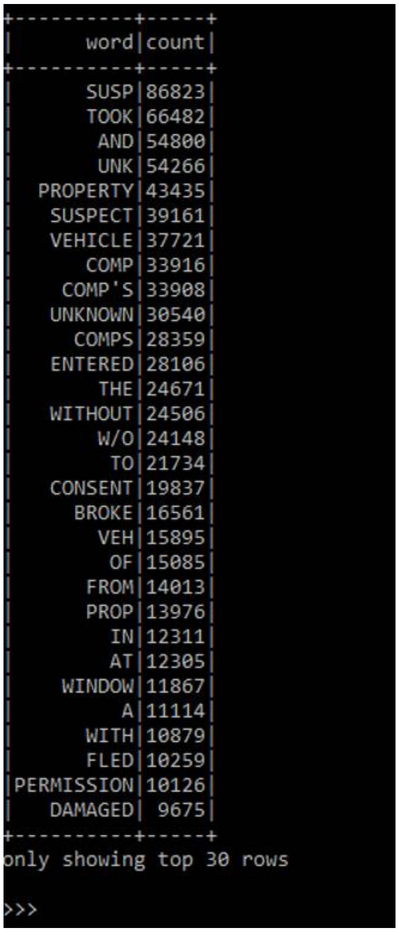
**Text Mining – PySpark**

**Data File:** modus_operandi.txt, type_incident.txt and type_location.txt. (Load file into Spark)

Text mining was executed on the data points from Modus_Operandi, type of incident and type of location in order to generate word count of the most frequent words used. This can help us understand the underlying sentiment of act of most crimes, type of crime and location of crime.

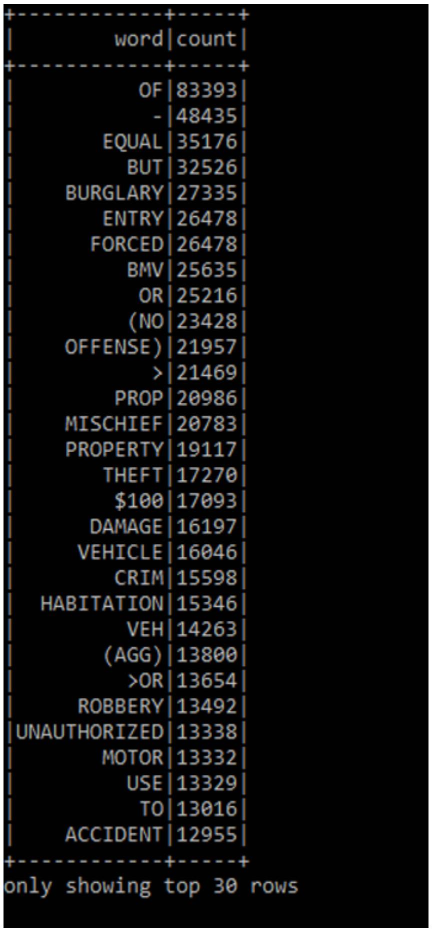
**This can be achieved using the code provided in the file “Word_Count_PySpark.txt”.

MODUS OPERANDI



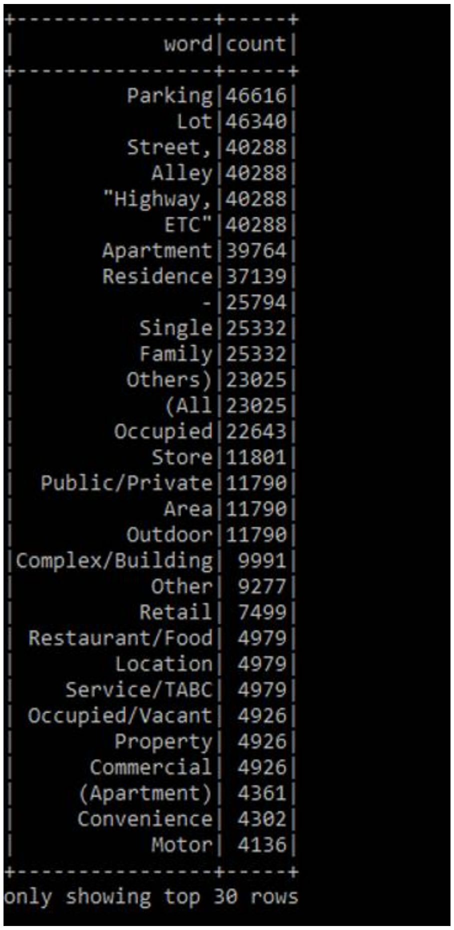
```
+----------+-----+
|      word|count|
+----------+-----+
|      SUSP|86823|
|      TOOK|66482|
|       AND|54800|
|       UNK|54266|
|  PROPERTY|43435|
|   SUSPECT|39161|
|   VEHICLE|37721|
|      COMP|33916|
|    COMP'S|33908|
|   UNKNOWN|30540|
|     COMPS|28359|
|   ENTERED|28106|
|       THE|24671|
|   WITHOUT|24506|
|       W/O|24148|
|        TO|21734|
|   CONSENT|19837|
|     BROKE|16561|
|       VEH|15895|
|        OF|15085|
|      FROM|14013|
|      PROP|13976|
|        IN|12311|
|        AT|12305|
|    WINDOW|11867|
|         A|11114|
|      WITH|10879|
|      FLED|10259|
|PERMISSION|10126|
|   DAMAGED| 9675|
+----------+-----+
only showing top 30 rows

>>>
```

TYPE OF INCIDENT



```
+------------+-----+
|        word|count|
+------------+-----+
|          OF|83393|
|           -|48435|
|       EQUAL|35176|
|         BUT|32526|
|    BURGLARY|27335|
|       ENTRY|26478|
|      FORCED|26478|
|         BMV|25635|
|          OR|25216|
|         (NO|23428|
|    OFFENSE)|21957|
|           >|21469|
|        PROP|20986|
|    MISCHIEF|20783|
|    PROPERTY|19117|
|       THEFT|17270|
|        $100|17093|
|      DAMAGE|16197|
|     VEHICLE|16046|
|        CRIM|15598|
|  HABITATION|15346|
|         VEH|14263|
|       (AGG)|13800|
|         >OR|13654|
|     ROBBERY|13492|
|UNAUTHORIZED|13338|
|       MOTOR|13332|
|         USE|13329|
|          TO|13016|
|    ACCIDENT|12955|
+------------+-----+
only showing top 30 rows
```

TYPE OF LOCATION



```
+----------------+-----+
|            word|count|
+----------------+-----+
|         Parking|46616|
|             Lot|46340|
|         Street,|40288|
|           Alley|40288|
|       "Highway,|40288|
|            ETC"|40288|
|       Apartment|39764|
|       Residence|37139|
|               -|25794|
|          Single|25332|
|          Family|25332|
|         Others)|23025|
|            (All|23025|
|        Occupied|22643|
|           Store|11801|
|  Public/Private|11790|
|            Area|11790|
|         Outdoor|11790|
|Complex/Building| 9991|
|           Other| 9277|
|          Retail| 7499|
| Restaurant/Food| 4979|
|        Location| 4979|
|     Service/TABC| 4979|
|  Occupied/Vacant| 4926|
|        Property| 4926|
|      Commercial| 4926|
|     (Apartment)| 4361|
|     Convenience| 4302|
|           Motor| 4136|
+----------------+-----+
only showing top 30 rows
```

**Modus Operandi:** Most incident reports involve an unknown suspect who takes property / vehicle / computer without consent and enters the premises / vehicle through window. The suspect also causes damage to property and flees the spot.

**Type of Incident:** Most incident reports involve Burglary, Robbery and Theft with forced entry. Stealing money or property worth more than $100. Most incidents also involve a vehicle theft or accident.

**Type of Location:** Most incidents take place outdoors in a parking lot. High number of incidents also take place in apartment occupied by either a single occupant or a family.