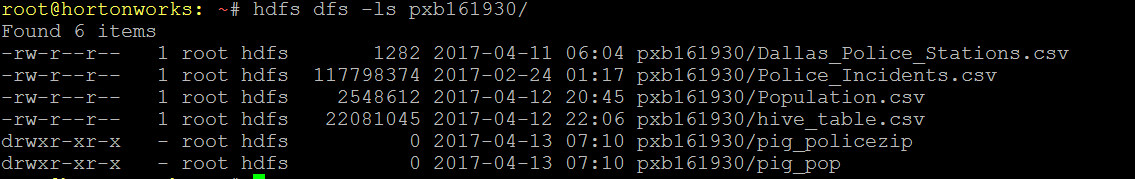
Final project Report

Data Cleaning:

* Loaded the datasets as spark data frame using sqlContext and removed the unnecessary columns.
* Cleaned all the columns names of the dataset and casted data time columns.
* Created Hive tables using sqlContext and loaded them into hive and then to pig to perform data analysis.



**Objective 2:** I joined Dallas Police Incidents data to another dataset that provides police station data. Checked the response rate is getting effected by distance of the police station from the incident location. I used Hive to store and join my data. I will use the Hive Query Language (HQL) to analyze the data.

Code Execution:

Load the Police\_Incidents.csv and Dallas\_police-Stations.csv dataset to pxb161930/ folder in hdfs using the below command

Hdfs dfs -put /root/Police\_Incidents.csv pxb161930/

Hdfs dfs -put /root/ Dallas\_police-Stations.csv pxb161930/

Once the datasets are loaded in to hdfs run the below command to execute the “Q2\_code.py” python file.

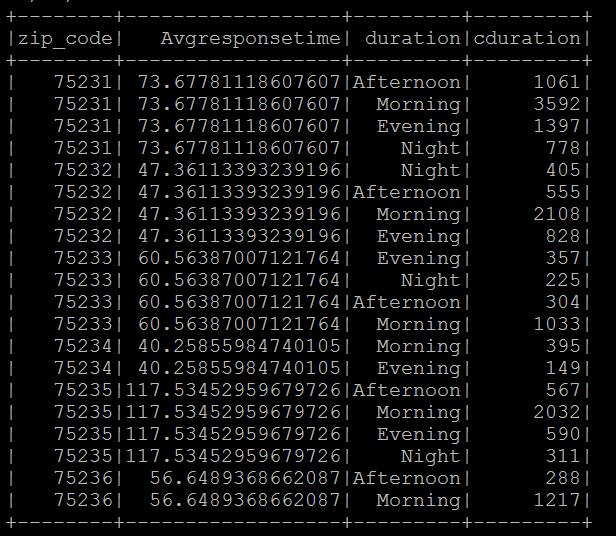
pyspark --packages com.databricks:spark-csv\_2.10:1.2.0 Q2\_code.py

Hive query description:

* Joined Police Incidents data with police station data.
* Wrote logics to extract Average response time, duration of the day and count of crime for each duration.

This command performs the data cleaning, creates hive tables and executes the hive queries through sqlContext and the output is saved as /root/q2data/q2\_Avg\_Response\_time.csv

Output:

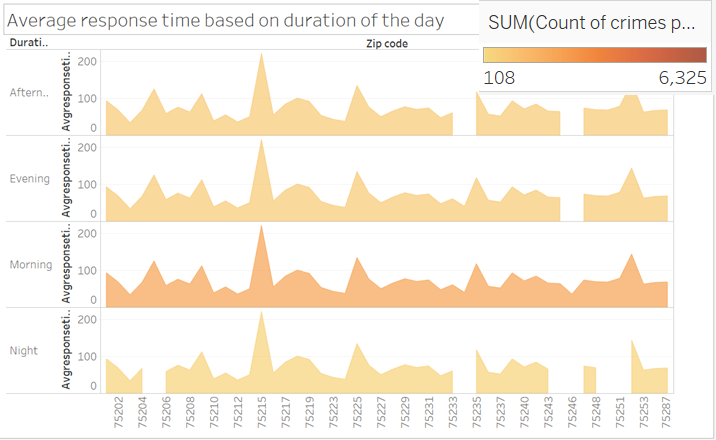


Insight:

* The response time doesn’t follow any pattern with police station location.
* But surely the response time is getting effected by the time of the occurrence of the crime.
* Crimes which are happening during mornings have AVG response time more than the crimes which occurred during evenings, nights and afternoons.

Used /root/q2data/q2\_Avg\_Response\_time.csv to visualize the data and to understand the trends and patterns better.

Below are the trends based on each duration of the day and each zip code.



**Objective 3:**

I used hive to do data analysis and check if the incidents which are reported by the witness are addressed and resolved in short duration compared to the incidents which are directly reported by the victim or which are identified by the police directly. If incidents reported by witness are addressed better, then check if there is any relation between the person (race) reporting and the address from which the incident is reported to response time.

Code Execution:

Load the Police\_Incidents.csv to pxb161930/ folder in hdfs using the below command

Hdfs dfs -put /root/Police\_Incidents.csv pxb161930/

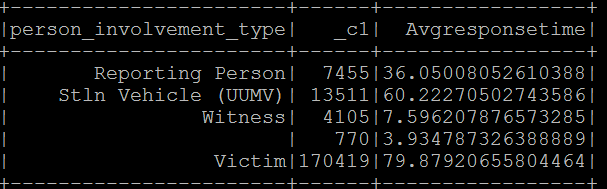
execute the “Q3\_code.py” python file using the below command.

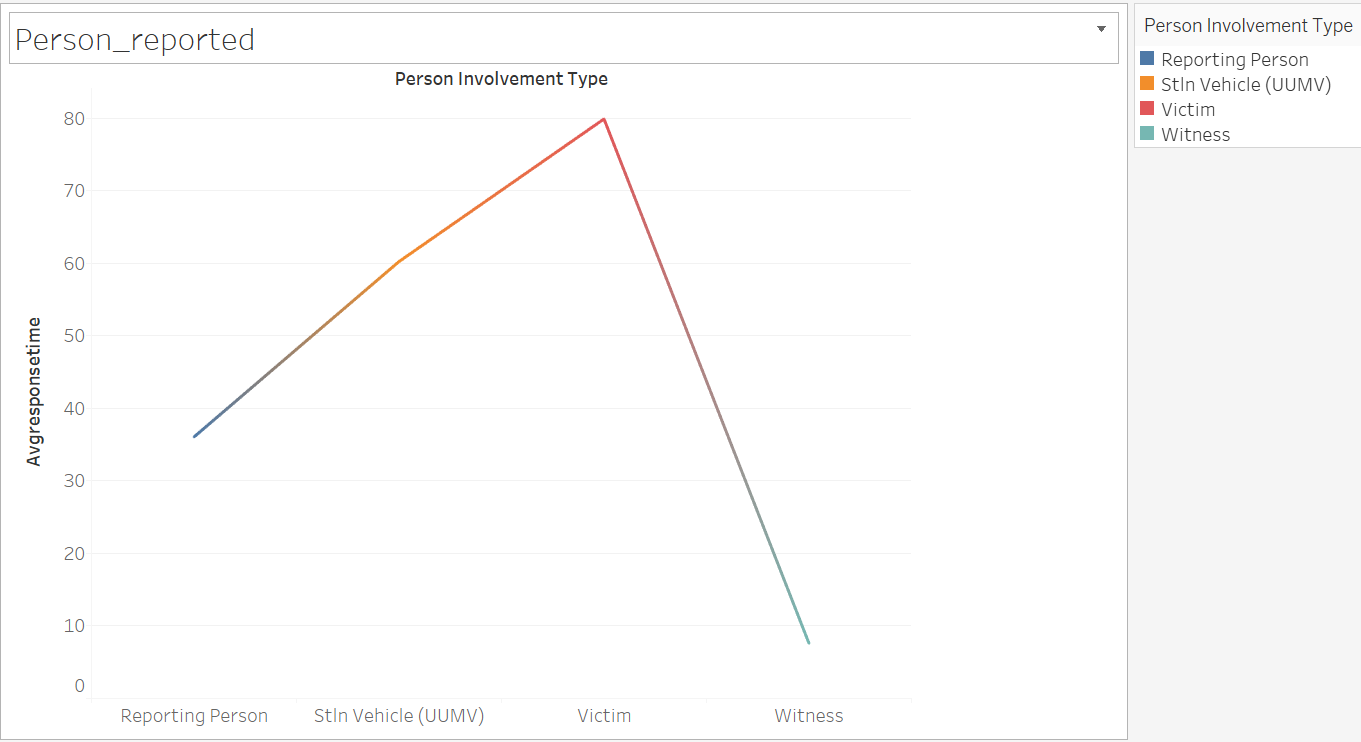
pyspark --packages com.databricks:spark-csv\_2.10:1.2.0 Q3\_code.py

This command performs the data cleaning, creates “default.q3table” hive table and executes the hive queries through sqlContext.

I have then performed the analysis on the q3table data and generated outputs as below: ‘**/root/q3data\_1/q3\_Respone\_time.csv’** and ‘**/root/q3data\_2/q3\_Complainant\_table.csv’**

Output1:

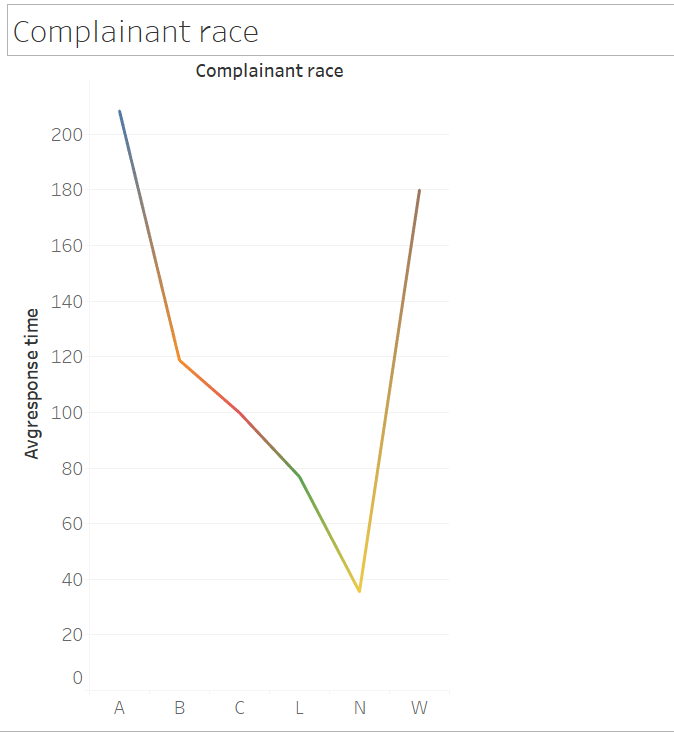
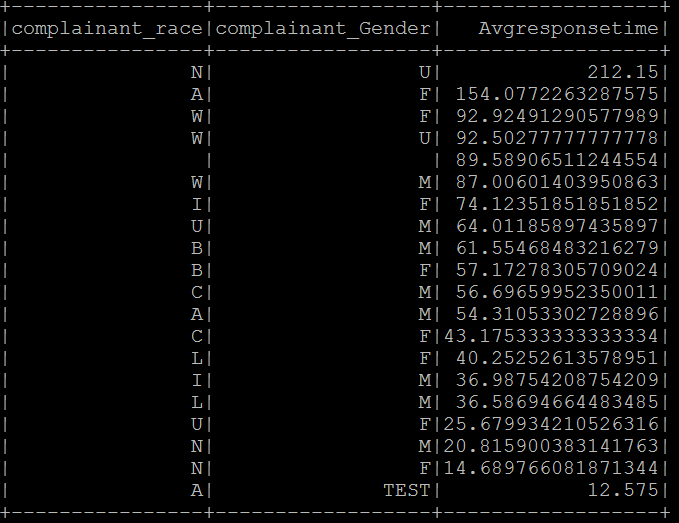




Insights:

The crimes which are reported by the witness are resolved faster with an average of 7.59 minutes.

Output2:



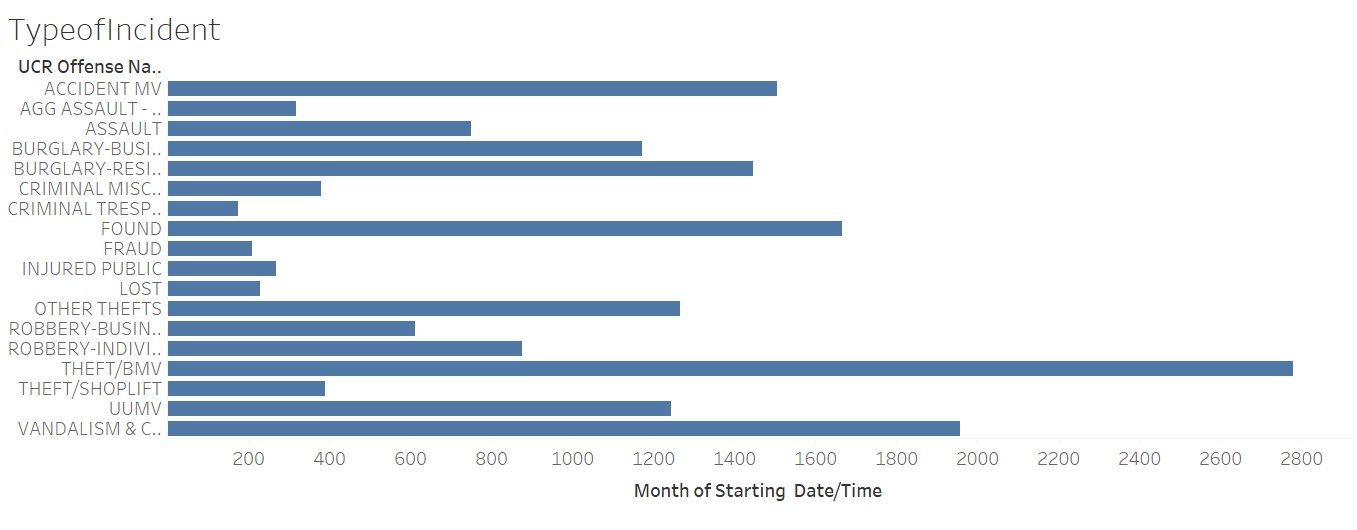
Insights:

Complainants race with ‘N’ has reported the crimes faster than the others.

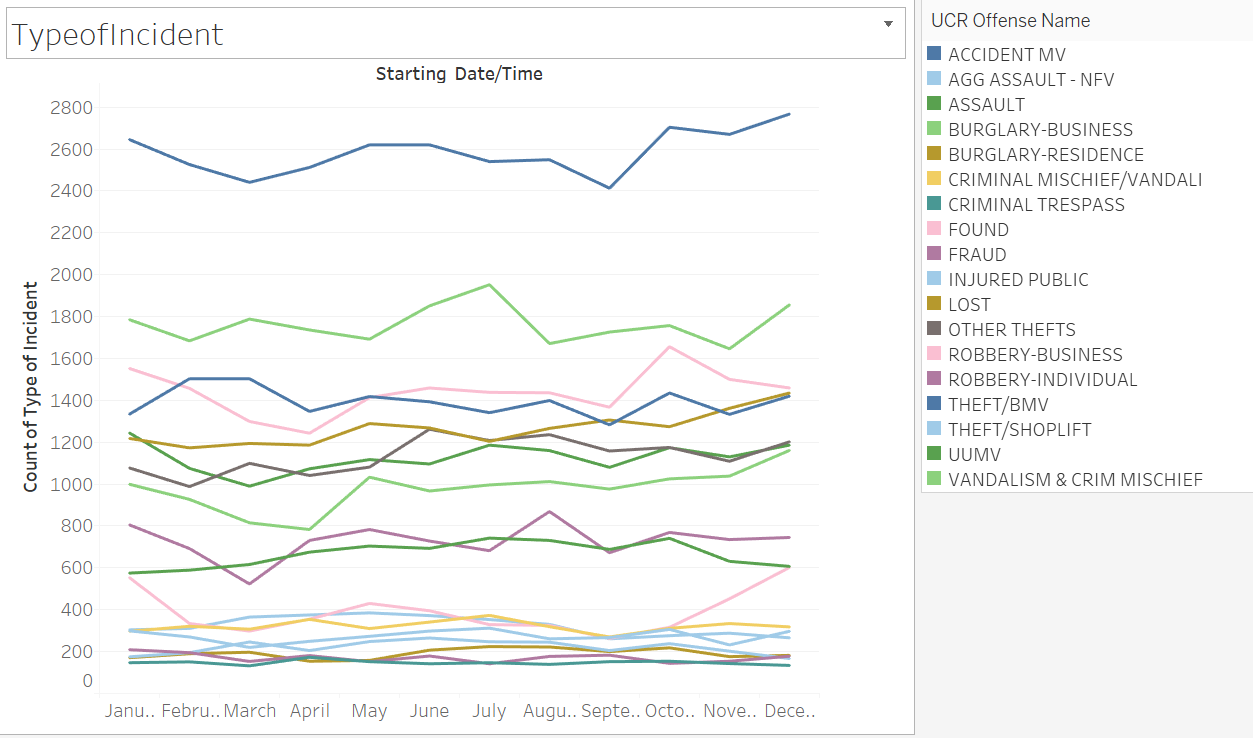
**Objective 4:**

I will perform Time series analysis to check which type of crimes are prevailing during different months in a year? I will use tableau to visualize these outcomes.

I used hive to clean the data and loaded it in to tableau.



Here we can clearly see that the THEFT/BMV crimes, Accident MV, BURGLARY, found – ABANDONED Property, UUMV and VANDALISM & CRIME MISCHIEF are more prevalent.



Some of the crimes are more prevalent throughout the year. But the fact is that all these most prevalent crimes tend to increase during the months of November and December.

**Objective 5:**

Join the population density table per zip code with our dataset using pig and find out if high populous areas are having more crime rate when compared to less population density area and to find out correlation with population density and number of crimes.

Code Execution:

Load the Population.csv dataset to pxb161930/ folder in hdfs using the below command:

# hdfs dfs -put /root/ Population.csv pxb161930/

Run the below query:

# pyspark --packages com.databricks:spark-csv\_2.10:1.2.0 Q5\_code.py

run the below commands on command prompt:

# cat /root/q5data/000000\_0 /root/q5data/000001\_0 > pxb161930/q5data.csv

# hdfs dfs -put /root/pxb161930/q5data.csv pxb161930/

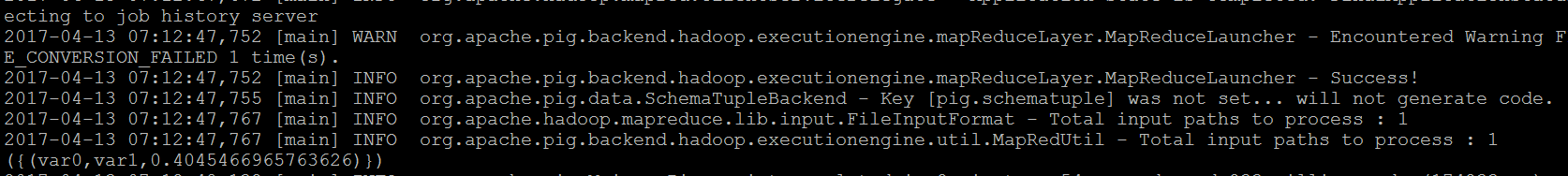
Once we have the data in place run the below command:

# Pig Q5\_code.pig

Pig Query description:

* Loaded the q5data.csv and Population.csv datasets as pig relations.
* Joined both the dataset based on the zip code and calculated correlation between crime rate and population density.

Output:



Insight:

Correlation between Crime rate and population density has turned out to be 40%. We can inference that the crime rate and population has positive significant correlation between both.