

Part 1: Loading the data and creating tables.

Step 1. Download the Dallas Police Incidents and clean the dataset and upload it into cyberduck.

Step 2. Use **awk -F "\t" '{print \$2 "\t" \$3 "\t" \$7 "\t" \$8 "\t" \$10 "\t" \$11 "\t" \$17 "\t" \$18 "\t" \$19 "\t" \$22 "\t" \$23 "\t" \$25 "\t" \$26 "\t" \$32 "\t" \$58 "\t" \$59 "\t" \$60 "\t" \$61 "\t" \$62 "\t" \$63 "\t" \$64 "\t" \$65}' Modified_Police_Incidents.tsv > Police.tsv** command to create another file with the necessary fields.

Step 3. Upload the compressed project folder to cyberduck and unzip using 'unzip project'. Run hive and pig scripts using **hive -f scripts/xxxxxxx.hive** and **pig -f scripts/xxxxxxx.pig** respectively in the project folder. Run **/hive -f scripts/xxxxxx.hive > ~/ xxxxxx.txt** to get the results of the query into a text file to complete the visualization.

Step 4. Change the directory to project using 'cd project' and run the **create_pop_data_text.hive** query to create the police incidents table from the 'Police.tsv' created in step2.

Step 5. Run the following hive query '**create_housing_data_text.hive**' to create housing data that contains the number of housing units per each zip code. The data is present in the 'Housing_Data' file of the data folder in project directory.

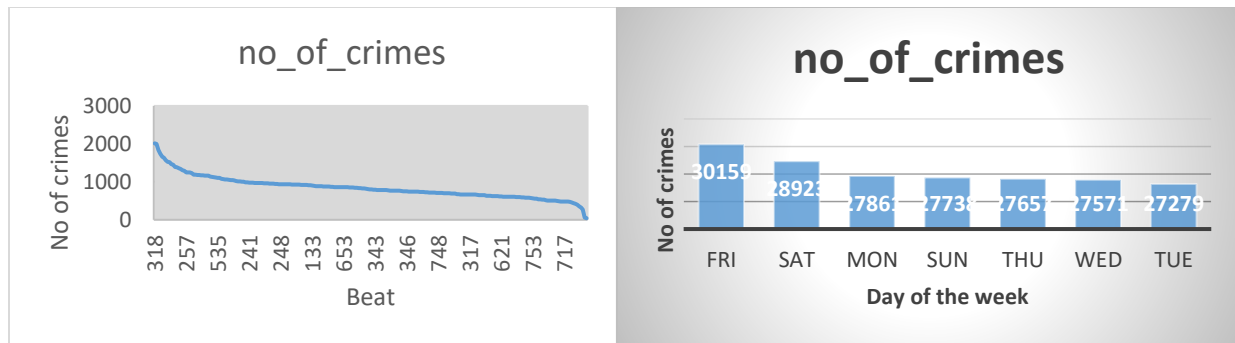
Step 6. Run the following hive query '**create_hhincome_data_text.hive**' to create income data that contains the average income of families per each zip code. The data is present in the 'HH_Income_Data' file of the data folder in project directory.

Step 7. Run the following hive query '**create_families_data_text.hive**' to create families' data that contains the number of families per each zip code. The data is present in the 'Families_Data' file of the data folder in project directory.

Part 2: Data Analysis

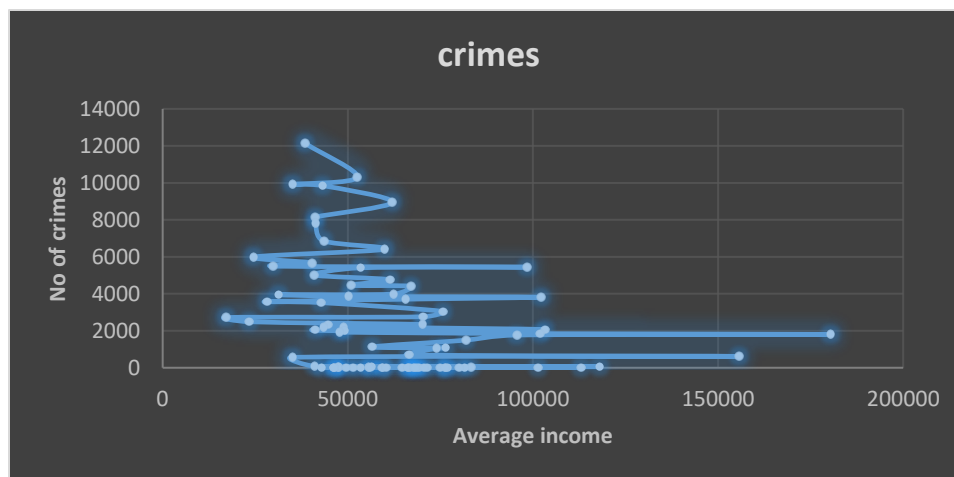
Run **crimes_by_beat.hive**, **crimes_by_reporting_area.hive**, **crimes_by_sector.hive** to get the total number of incidents that are happening in each beat, reporting area and sector respectively.

We can see from the graphs below that beat 318 has the highest number of crimes and the number of crimes is high on Friday followed by the number of crimes on Saturday.



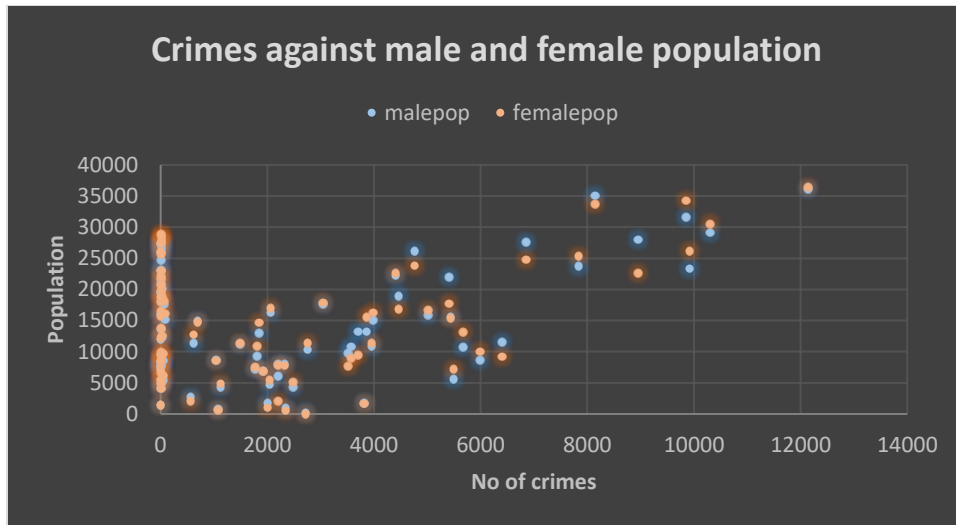
Run ***crimes_by_day.hive***, ***crimes_by_day_zipcode.hive*** to get the total number of incidents on each day of the week and the total number of incidents on each day of the week in each zip code respectively.

Run ***crimes_against_average_income.hive*** to join the police incidents table and the average income table by zip code and get the number of incidents in each zip code and the average income of that zip code. The output of this query is in the 'crimes_against_average_income' file of the output data folder.



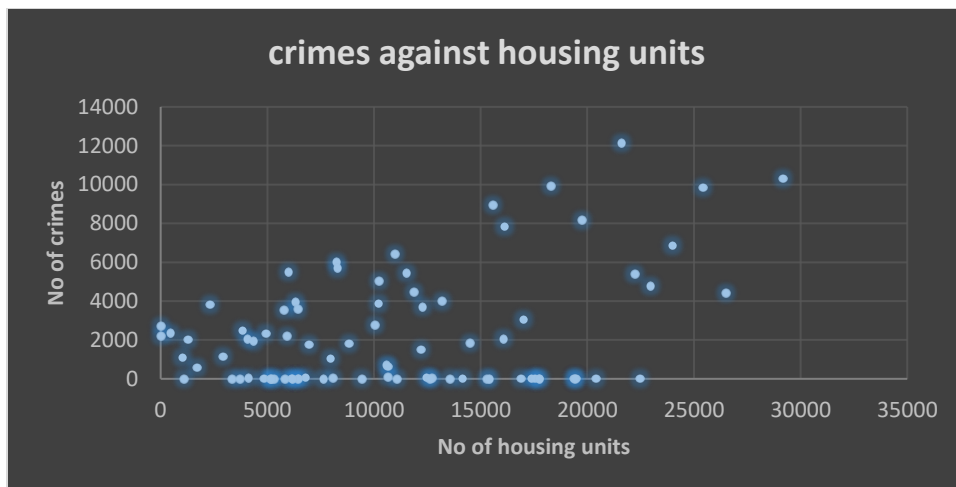
We can see from the plot above that as the average income of the zip code increases there is a decrease in the number of crimes happening in that zip code area(except for some of the outliers that are present).

Run ***crimes_against_pop_data.hive*** to join the police incidents table and the population table by zip code and get the number of incidents in each zip code and the population of that zip code. The output of this query is in the 'crimes_against_pop_data' file of the output data folder.



We can clearly see that as the population of a zip code increases the number of crimes happening in that zip code also increases (except for some of the outliers that are present).

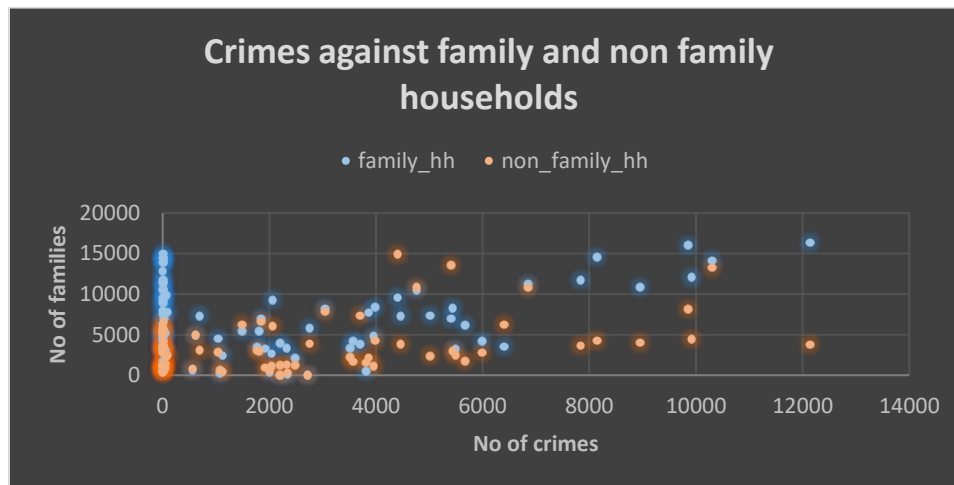
Run ***crimes_against_housing_data.hive*** to join the police incidents table and the housing table by zip code and get the number of incidents in each zip code and the number of housing units in that zip code. The output of this query is in the 'crimes_against_housing_data' file of the output data folder.



We can see that as the number of housing units increase the number of crimes happening in that zip code also increases (except for some of the outliers that are present).

Run ***crimes_against_families_data.hive*** to join the police incidents table and the families table by zip code and get the number of incidents in each zip code and the number of family and non-family

households in that zip code. The output of this query is in the 'crimes_against_families_data' file of the output data folder.

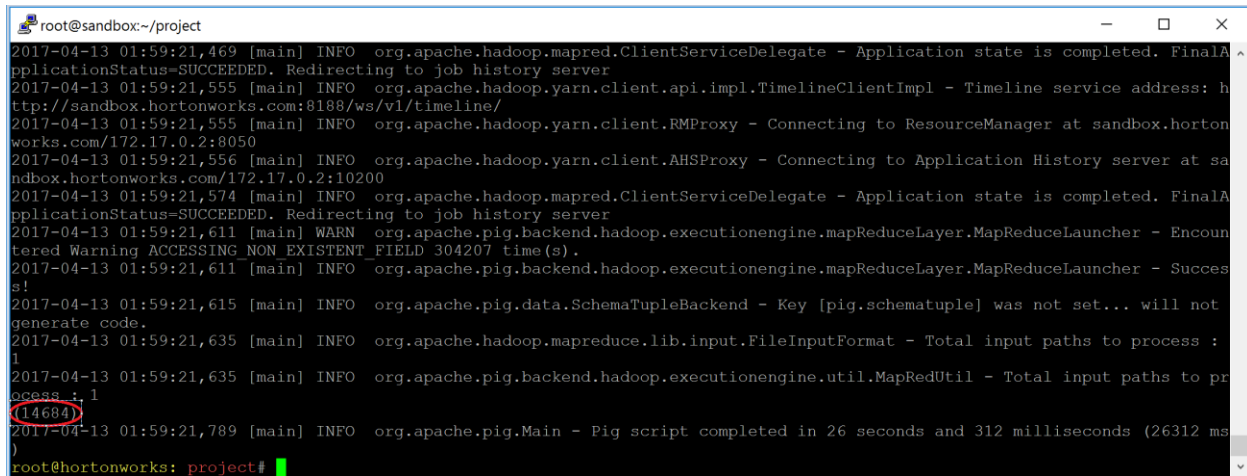


We can see that as the number of families increase the number of crimes happening in that zip code also increases (except for some of the outliers that are present). And also the number of crimes in the family households is almost equal to that of the non-family households.

Run '**count_of_assault_crimes.pig**' to create a relation which stores the necessary fields of the Police Incidents data and then count the number of crimes which are of the type 'assault' by matching the column with the expression '**.*ASSAULT.***'. We can see that '9975' assault type of incidents present in the Dallas police data.

```
root@sandbox:~/project
2017-04-13 01:55:15,502 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCEEDED. Redirecting to job history server
2017-04-13 01:55:15,586 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: h
ttp://sandbox.hortonworks.com:8188/ws/v1/timeline/
2017-04-13 01:55:15,586 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox.horton
works.com/172.17.0.2:8050
2017-04-13 01:55:15,586 [main] INFO org.apache.hadoop.yarn.client.AHSPProxy - Connecting to Application History server at sa
ndbox.hortonworks.com/172.17.0.2:10200
2017-04-13 01:55:15,594 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCEEDED. Redirecting to job history server
2017-04-13 01:55:15,619 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encoun
tered Warning ACCESSING_NON_EXISTENT_FIELD 205722 time(s).
2017-04-13 01:55:15,619 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succes
s!
2017-04-13 01:55:15,624 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not
generate code.
2017-04-13 01:55:15,638 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
1
2017-04-13 01:55:15,638 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pr
ocess: 1
(9975)
2017-04-13 01:55:15,754 [main] INFO org.apache.pig.Main - Pig script completed in 32 seconds and 268 milliseconds (32268 ms
)
root@hortonworks: project#
```

Run '*count_of_robbery_crimes.pig*' to create a relation which stores the necessary fields of the Police Incidents data and then count the number of crimes which are of the type 'robbery' by matching the column with the expression '*.ROBBERY.*'. We can see that '14684' robbery type of incidents present in the Dallas police data.



```
root@sandbox:~/project
2017-04-13 01:59:21,469 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCEEDED. Redirecting to job history server
2017-04-13 01:59:21,555 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: h
ttp://sandbox.hortonworks.com:8188/ws/v1/timeline/
2017-04-13 01:59:21,555 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox.horton
works.com/172.17.0.2:8050
2017-04-13 01:59:21,556 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sa
ndbox.hortonworks.com/172.17.0.2:10200
2017-04-13 01:59:21,574 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCEEDED. Redirecting to job history server
2017-04-13 01:59:21,611 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encoun
tered Warning ACCESSING_NON_EXISTENT_FIELD 304207 time(s).
2017-04-13 01:59:21,611 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succes
s!
2017-04-13 01:59:21,615 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not
generate code.
2017-04-13 01:59:21,635 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
1
2017-04-13 01:59:21,635 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pr
ocess : 1
14684
2017-04-13 01:59:21,789 [main] INFO org.apache.pig.Main - Pig script completed in 26 seconds and 312 milliseconds (26312 ms
)
root@hortonworks: project#
```

NOTE: The demographic datasets used in the join tables are from US Census 2000 datasets (<https://www.census.gov>). The visualizations were completed in Excel (Got the output of each query in a text file and completed the visualization).