Problem 3, DS730 – Chris Austin

Overview of dataset: <https://www.kaggle.com/currie32/crimes-in-chicago>

Link to data: <https://www.kaggle.com/currie32/crimes-in-chicago/data>

I combined all four data files (2001-2016 full, 2017 partial) from Kaggle and cleaned out header information. It came to ~1.8 GB in total, and over 7million records.

A couple of small files to join with are in the tar file as well. One is seasons.csv, and categorizes months into seasons. The second is TimesOfDay.csv, and buckets times of day into 4-hour segments for categorization of crime data into intervals so that the intervals can be considered for frequency.

The Hive code is in the problem3 folder of the tar file, and is also commented with much of the information below:

prob3\_0.hive loads the main Chicago data table and also loads the two lookup tables. It also creates a couple of views used in the later problems. IT SHOULD BE RAN FIRST and has the DROP/CREATE commands not specific to particular problems.

====================================================================================

prob3\_1.hive is used to ask: Are violent crimes (Assaults, Homicides, Battery, Criminal Sexual Assault) more prevalent as a percentage of crimes reported in the summer months than in the rest of the year? In other words, are they more likely to occur as a fraction of that season's reported offenses?

Homicides per 100,000 crime reports. Summer is the highest season.

120.1 Fall HOMICIDE

104.78 Spring HOMICIDE

129.06 Summer HOMICIDE

99.21 Winter HOMICIDE

Batteries (as a % of all reports). Summer is the second-highest season, but barely.

17.41 Fall BATTERY

19.04 Spring BATTERY

18.91 Summer BATTERY

17.11 Winter BATTERY

Criminal Sexual Assault per 100,000. Summer is the highest season.

369.16 Fall CRIM SEXUAL ASSAULT

360.74 Spring CRIM SEXUAL ASSAULT

389.25 Summer CRIM SEXUAL ASSAULT

385.29 Winter CRIM SEXUAL ASSAULT

Assaults per 1,000. Summer is second-highest.

59.12 Fall ASSAULT

64.03 Spring ASSAULT

62.54 Summer ASSAULT

56.21 Winter ASSAULT

Other questions that the data could shed light on: Do we have similar Summer reporting ratios for non-violent types of crime? Are there others that are counter-cyclical to violent crimes?

====================================================================================

prob3\_2.hive is used to ask: Are robberies and other crimes of opportunity (Theft, Burglary, Prostitution, Criminal Trespass, etc.) more prevalently reported at a specific time of day? In other words, are they at a higher percentage of that time's calls?

Answer:

Thefts are more commonly reported as a percentage of reports in that timeframe in the early afternoon and late morning.

24.42 Early Afternoon THEFT

18.82 Early Morning THEFT

15.7 Evening THEFT

22.36 Late Afternoon THEFT

23.6 Late Morning THEFT

17.53 Overnight THEFT

Robberies are more commonly reported as a percentage of reports in that timeframe in the early morning and overnight.

3.06 Early Afternoon ROBBERY

4.89 Early Morning ROBBERY

4.29 Evening ROBBERY

3.87 Late Afternoon ROBBERY

2.47 Late Morning ROBBERY

4.96 Overnight ROBBERY

Prostitution more commonly reported as a percentage of reports in that timeframe in the early morning and evening.

0.36 Early Afternoon PROSTITUTION

2.52 Early Morning PROSTITUTION

2.09 Evening PROSTITUTION

0.55 Late Afternoon PROSTITUTION

0.91 Late Morning PROSTITUTION

0.95 Overnight PROSTITUTION

Burglaries are more commonly reported as a percentage of reports in that timeframe in the early morning and late morning,

5.84 Early Afternoon BURGLARY

11.13 Early Morning BURGLARY

3.96 Evening BURGLARY

5.23 Late Afternoon BURGLARY

8.5 Late Morning BURGLARY

4.64 Overnight BURGLARY

====================================================================================

prob3\_3.hive is used to ask: Does the Chicago police department have a better arrest percentage for violent crimes (Assaults, Homicides, Battery, Criminal Sexual Assault) in the first four years of the dataset

(2001-2004) than in the last four years of the dataset (2012-2015) ? 2016 and partial 2017 have been thrown out, as investigations may still be pending.

Answer: (Column headers are report period, report type, arrest count, report count, ratio)

2012-15 ASSAULT 16657 71805 23.2 %

2012-15 BATTERY 47027 211483 22.24 %

2012-15 HOMICIDE 717 1846 38.84 %

2012-15 CRIM SEXUAL ASSAULT 705 5314 13.27 %

2001-04 ASSAULT 26160 121417 21.55 %

2001-04 BATTERY 77859 363538 21.42 %

2001-04 HOMICIDE 1996 3172 62.93 %

2001-04 CRIM SEXUAL ASSAULT 1491 7079 21.06 %

Assault arrest rates have increased marginally since the early years of the dataset. Same goes for battery but criminal sexual assault and homicide arrest rates have decreased sharply since the early years of the dataset.

====================================================================================

prob3\_4.hive is used to ask: Which 10 city blocks have the highest count of narcotic offenses over the life of the data, and what were their arrest-to-report ratios?

Answer: (Column headers are report count, arrest count, ratio, and city block)

2670 2619 98.09 % 033XX W FILLMORE ST

2485 2463 99.11 % 005XX E BROWNING AVE

1803 1785 99.0 % 038XX W ROOSEVELT RD

1800 1775 98.61 % 023XX S STATE ST

1565 1525 97.44 % 027XX S DEARBORN ST

1369 1343 98.1 % 029XX S STATE ST

1329 1322 99.47 % 039XX W GRENSHAW ST

1225 1221 99.67 % 012XX N BURLING ST

1215 1214 99.92 % 051XX W MADISON ST

1136 1134 99.82 % 048XX W MONROE ST