Baltimore Crime Trends

CSE 140 Homework 9 – Final Project

Sasha Babayan

Summary of Research Questions and Results:

1. Do crimes happen more frequently farther away from district police stations?
   For this question I am trying to determine if you see an increase in crimes as you move farther away from the police station of that crime's corresponding district. This will help the law enforcers in Baltimore to be better placed within the city and could, over time, decrease crime rates.
   Results: There is no correlation between distance from police stations and frequency of crimes.

2. Is there a correlation between time of day and the number crimes that occur?
   I am interested in testing to see if there are certain times of the day that see a lot of crime and when those times are. This will help policemen, as well as the citizens, to be more alert during times of the day with high crime rates.
   Results: There is a strong correlation between the time of day and the number of crimes that occur during that time.

Motivation and Background:

Baltimore is well-known for its high crime rates and is considered to be one of the more dangerous cities in the nation. In analyzing this dataset I hope to uncover trends that could help law enforcement better detect opportunities for crime and be better prepared for when crimes do occur. By having insight into the likely locations and times of crime occurrences, policemen can prevent crimes before they occur. Hopefully through analyses like these Baltimore can eventually decrease their crime rate by a significant amount and ultimately become a safer place.

Dataset:

I used two datasets for this assignment, the first I took from Baltimore city's government website, and the second I created myself. The dataset I exported from Baltimore's government website was a large spreadsheet of crimes that have occurred in the city dating back to 2008. This dataset includes information on the date and time of the crime, the type of crime, the location (in address form and as latitude/longitude coordinates) and the district and neighborhood in which the crime occurred. In order to obtain my two datasets follow the instructions below.

Dataset Instructions:

1. Crime Rates:
    a. Go to https://data.baltimorecity.gov/Crime/BPD-Part-1-Victim-Based-Crime-Data/wsfq-mvij
    b. There are 8 colored tabs towards the right hand side of the page, select the light blue colored tab labeled "Export"
    c. Select the first option "CSV", download the dataset and save it as "Baltimore_Crime_Data"

2. Police Stations:
    a. Create a new .csv file titled "Police_Stations" with three columns labeled "District", "Address", and "Location"
    b. Go to http://www.baltimorepolice.org. Under the tabs Your Community>Your District> select each district one by one and include their district name (in uppercase letters) under the "District" column, and their address in the "Address Column"
    c. For each station's address, find the corresponding latitude, longitude coordinates using http://stevemorse.org/jcal/latlon.php and taking the "from google" decimal results. Store these coordinates in the "Location" column.

Methodology:

a. Create a master list of the data in each row of the .csv file "Baltimore_Crime_Data". Extract the district name, the coordinates, the neighborhood, the time, and the type of crime. If a row of data is missing a district name, missing location coordinates, or the location coordinates are (0, 0) or fall outside of the range (35<lat<45, -85<long<-74) do not store information for that row.
b. Create a master list of the data in each row of the .csv file "Police_Stations". Extract the district name and the coordinates.
c. Find the distance from the nearest police station (the police station in the same district) for each crime.
d. Create a scatter plot of 10,000 randomly sampled crimes using latitude and longitude as your x and y axes. On that sample plot, using a different color than what was used to plot the crimes, plot each police station's location. Use the domain (39.2, 39.4) for the x-axis and the range (-76.75, -76.5) for the y-axis. Save the scatter plot to a .png file titled "loc_scatter_plot.png".
e. Create a histogram bar graph plotting the number of distances that were less than 1 mile, 1-2 miles, 3-4 miles, 4-5 miles, and greater than 5 miles away from the district police station. Save the histogram to a .png file titled "distance_frequency_hist.png".

f. Create a pie chart of the number of the ratios of the number of reported crimes for each individual district. Save the chart to a .png file titled "district_pie_chart.png".
g. Find and print the Pearson correlation coefficient and p-value for all distances (distance from the location of the crime to the district's police station) against the frequency of each distance.
h. Print the neighborhood name, district name, and number of crimes in that neighborhood for the top 10 neighborhoods with the highest number of crimes. If two neighborhoods have the same number of crimes, print the one that appears first in lexicographical order of its district.
i. Create a histogram bar graph of the number of crimes that occur during each hour of the day. Use military time, where the 0^{th} hour is 12:00 AM. Save the plot to a .png file titled "time_frequency_hist.png".
j. Find and print the Pearson correlation coefficient and p-value for the hour of day against the number of crimes that have been reported during that hour.
k. Print each unique type of crime and its frequency. Print in sorted descending order of frequencies, if two frequencies are the same print in lexicographical order.

Results:

1. For my first research question - do crimes happen more frequently farther away from district police stations – I chose my null hypothesis to be: as you move farther away from district police stations, the number of crimes will increase. I assumed that areas closer to police stations would be safer because the officers are within closer range. However, I found the Pearson correlation coefficient to be 0.0337799340836 with a p-value of 3.74587245373e-24 (or essentially 0). This disproves my null hypothesis and is in favor of the alternative hypothesis with great certainty because the p-value is so small. Since my correlation coefficient is also small (.034) we say that there is no correlation between distance to police station and frequency of crimes. According to my dataset, the neighborhood that saw the most crimes was Downtown Baltimore, belonging to the central district. However in terms of districts, the Northeastern district had the highest number of crimes (15.4%), almost double that of the Eastern (8.2%) and Western (8.3%) districts.

2. For my second research question – is there a correlation between time of day and the number crimes that occur – I set my null hypothesis to be: there is a correlation between time of day and frequency of crimes. The Pearson correlation coefficient was 0.733212704933 with a p-value of 4.58454303789e-05. This leads me to believe that there is a strong correlation between time of day and frequency of crimes. This is because the correlation coefficient is within the range of 0.5 to 1.0 and the p-value is very close to zero, thus suggesting great certainty. From the produced histogram, it is evident that the

times of day that see the least amount of crime are from 2:00 – 8:00 AM and the times that see that most amount of crime are 3:00 – 8:00 PM. This is somewhat unexpected, I had assumed that the most amount of crimes occur in the night. However, because the amount of robberies and larcenies highly exceeds the amount of shooting and homicide cases, it makes sense that these crimes occur during the day because this is the time when most people are away from their home and at work.

Reproducing Your Results:

To reproduce my results, simply store the python program and both .csv files in the same directory, then run the python code. The output will be printed and the plots will be saved as .png files under their respective names in that same directory. In order for my code to run, however, the Haversine package must be installed (https://pypi.python.org/pypi/haversine).

Collaboration:

Though no one helped me on the assignment I did seek help for plotting histograms and pie charts on the internet. I used some of the code from http://matplotlib.org/examples/pylab_examples/pie_demo.html to help with my pie chart, and some code from http://stackoverflow.com/questions/5926061/plot-histogram-in-python to help with one of my histograms.

Reflection:

For this project I did not give enough thought to my dataset and what the assignment was asking for. Though my initial dataset seemed interesting (Baltimore Parking dataset) it wasn't flexible nor standardized enough for me to be able to draw any substantial conclusions from it. In addition it was so large that I decided to work with only one particular month of the data which limited me greatly in the analyses I could do. Halfway through writing my code I decided to switch to this dataset (Baltimore Crime data) because it seemed more manageable to work with and a lot of the data was already in the format I needed it in. I do think I could have saved myself a lot of time if I really thought more about what the project was asking us to do and took into account my limitations, rather than jumping into a dataset right away.