Disha Saha, DSC 530- Final Project Paper

For my final project I was able to find a dataset on Vancouver, CA on Kaggle (Lu, 2019). Vancouver is known to be the most populous city in the province of British Columbia, and it has the highest population density in all of Canada (Wikipedia, 2020). Therefore, for my final project I wanted to find out if Vancouver city was safe? The data set that I used to investigate this issue was collected by Kangbo Lu from the city of Vancouver open data catalogue (Lu, 2019). The dataset has crime data from 2003 through 2019, which has 624039 rows, and 10 columns.

I have investigated these following questions, and this is what I can conclude:

- Is there a particular crime in Vancouver, CA that has become more prevalent?
  - According to the histograms, Theft from Vehicle shows to be most prevalent
- Is crime increasing throughout the years in Vancouver, CA?
  - According to the CDF, and histogram overall crime has gone down between 2003 and 2019, but crime has started to pick up again slowly starting 2014 and now its leveling off.
- Do colder months have less crime prevalence?
  - The hypothesis test showed that we can reject the null hypothesis, and therefore conclude that colder months have less crime prevalence in Vancouver, CA.
- Which neighborhoods have the most crime?
  - According to the histogram, Central Business District had the most crime incident occurrence in Vancouver, CA.
- Can we predict when crimes are most likely to occur?
  - The regression analysis gave me an R^2 value of 0.03 which meant that the model was not a very good fit to the data. Based on this outcome I would say it would be difficult to predict when a crime was going to occur.

However, coming to these conclusions wasn't as straightforward for me sometimes. I struggled a bit on how to plot a proper scatter plot for my dataset. I believed it wasn't as straightforward as the examples given in the book such as plotting the scatter plots for weight and height. Even though I had 5 numeric variables in my dataset, I had to figure out how to aggregate and perform calculations on my variables in order to have quantitative data for me to do the scatter plots. Looking through my scatter plots, and the results of the correlation and covariance steers me to think that I might have made some assumptions that might not have been for the best. I would have liked to have figured out relationships that had a strong positive correlation. At the same token, I did have variables that helped me with my analysis such as the crime type, year, month, day, and neighborhood. Due to having such variables I feel confident about my histogram plots, CDF, hypothesis testing, and my regression analysis results.

References:

Downey, A. B. (2015). Think Stats. Sebastopol, CA: O'Reilly Media, Inc.

Lu, k. (2019, November 18). Vancouver Crime Report, Version 2. Retrieved from https://www.kaggle.com/agilesifaka/vancouver-crime-report

Lynch, D. (2019, June 3). Weather in Vancouver, B.C.: Climate, Seasons, and Average Monthly Temperature. Retrieved from https://www.tripsavvy.com/vancouver-average-monthly-temperatures-3371376

Wikipedia. (2020, January 14). Vancouver. Retrieved from https://en.wikipedia.org/wiki/Vancouver