DSC520 Final Project

Andrea Fox

13 November 2019

Section 1 – Week 9 – Getting Started

1. Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
    a. Watching the news every morning in Kansas City, Missouri it seems like there is always a story regarding gun violence. Citizens are becoming concerned and calling for action . The new mayor, Quentin Lucas, ran on the platform that he would work on decreasing violence. The real issue is that crime in general has increased in the city over the past three years, but the media is using gun violence stories because it is a hot topic nationwide.
2. Draft 5-10 Research questions that focus on the problem statement.
    a. What is the frequency of gun violence happening in comparison to other crimes reported?
    b. Are there specific areas that are more heavily targeted?
    c. Are the crimes reported only involve suspects or does it also include victims?
    d. Is crime really increasing or just seems like it because of heavier reporting from media?
    e. Does crime increase at a specific time of day?
    f. Is crime heavier during specific parts of the year?
3. Provide a concise explanation of how you plan to address this problem statement.
    a. I plan on addressing this topic by looking at the previous three years of crime reported by Kansas City Police Department. I will look heavily at the crime description and whether the firearm flag has an indication of yes. After gathering, cleaning, and modeling the data I plan to bring this information to the local news so that the city is receiving accurate reporting, and to KCPD so that they can target areas that may need additional patrols.
4. Discuss how your proposed approach will address (fully or partially) this problem.
    a. If my hypothesis is accurate than it can bring accurate awareness to the actual crime taking place in the city. This allows citizens to be more informed and perhaps feel somewhat safer knowing that not everything happening crime-wise is gun related. A secondary benefit of this study is to bring to KCPD so that they can see what areas may have a heavier burden of crime in comparison to other parts of the city. This will allow them to adjust patrols to areas in need as well as having better predictions on time crime is most likely to happen.
5. Do some digging on a dataset that you can use to address the issue. Original source where the data was obtained is cited and, if possible, hyperlinked. Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
    a. KCPD Crime Data 2018
        i. https://data.kcmo.org/Crime/KCPD-Crime-Data-2018/dmjw-d28i
        ii. Description: This is the most current information as of the date of upload. This provides the user the ability to view the most current crime information within Kansas City, Missouri. The displayed information is the most current information

from the data source as of the date of upload. The data source is dynamic and therefore constantly changing. Changes to the information may occur, as incident information is refined. While the Board of Police Commissioners of Kansas City, Missouri (Board) makes every effort to maintain and distribute accurate information, no warranties and/or representations of any kind are made regarding information, data or services provided. The Board is not responsible for misinterpretation of this information and makes no inference or judgment as to the relative safety to any particular area or neighborhood.

iii. Date Created: 12-18-2017
iv. Last Updated: 04-19-2019

b. KCPD Crime Data 2017
   i. https://data.kcmo.org/Crime/KCPD-Crime-Data-2017/98is-shjt
   ii. Description: This is the most current information as of the date of upload. This provides the user the ability to view the most current crime information within Kansas City, Missouri. The displayed information is the most current information from the data source as of the date of upload. The data source is dynamic and therefore constantly changing. Changes to the information may occur, as incident information is refined. While the Board of Police Commissioners of Kansas City, Missouri (Board) makes every effort to maintain and distribute accurate information, no warranties and/or representations of any kind are made regarding information, data or services provided. The Board is not responsible for misinterpretation of this information and makes no inference or judgment as to the relative safety to any particular area or neighborhood.
   iii. Date Created: 01-03-2017
   iv. Last Updated: 03-05-2018
c. KCPD Crime Data 2016
   i. https://data.kcmo.org/Crime/KCPD-Crime-Data-2016/wbz8-pdv7
   ii. Description: This is the most current information as of the date of upload. This provides the user the ability to view the most current crime information within Kansas City, Missouri. The displayed information is the most current information from the data source as of the date of upload. The data source is dynamic and therefore constantly changing. Changes to the information may occur, as incident information is refined. While the Board of Police Commissioners of Kansas City, Missouri (Board) makes every effort to maintain and distribute accurate information, no warranties and/or representations of any kind are made regarding information, data or services provided. The Board is not responsible for misinterpretation of this information and makes no inference or judgment as to the relative safety to any particular area or neighborhood
   iii. Date Created: 01-08-2016
   iv. Last Updated: 04-19-2019

6. Identify the Packages that are needed for your project.
   a. ggplot2
   b. readr
   c. class
   d. ggm
   e. caret
   f. caTools
   g. lattice

    h. devtools
    i. tidyverse
    j. tidyr
7. What types of plots and tables will help you to illustrate the findings to your research questions?
    a. Scatterplots to get an idea of where my data might be trending. I plan to look at gun violence by year and gun violence in comparison to other crimes on a year to year basis.
    b. Histogram showing gun violence by month and year
    c. Histogram showing all crime for on a year by year basis
    d. Histogram showing crime by area. (haven't decided if I want to look at each year separately or together or both)
    e. Histogram showing crime rates by time of day
    f. Histogram of all crimes where a firearm is involved (this may not necessarily be a shooting).
    g. Chart of predictions based on crime type, firearm involved, time of year, time of day, and area in KC
8. What do you not know how to do right now that you need to learn to answer your research questions?
    a. If cleaning data is involved how is that done?
    b. Will I need to split the data and set up train/test like week 8?
    c. What methods are needed?
    d. Will I need to set up a few different models and cross validate to verify the accuracy of my hypothesis?
    e. Is there bias in the data?

Section 2 – Week 10 – Cleaning Your Data and Exploratory Data Analysis

1. Data importing and cleaning steps are explained in the text and in the DataCamp exercises (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.
    a. See RMD file
2. With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
    a. See RMD file
3. What do you not know how to do right now that you need to learn to import and cleanup your dataset?
    a. I ran into a lot of problems when cleaning the data with the biggest of not knowing really where to start. I decided to go through Cleaning Data Using R on DataCamp to see if I could pick up a few things, which really helped me follow a list of ideas. I had an exceptionally hard time figuring out the correct way to create my main_kcpd18 dataset from the original. It kept telling me I had an undefined column. Turns out I had originally combined Offense and Description before creating the subset and it didn't recognize Crime. So I deleted that out and moved it underneath creating another dataset of main2_kcpd18. For now this looks to be the bulk of what I'm going to work with to answer the majority of my questions.
    b. Will I use the other columns like Age or Sex?
    c. Did I clean everything or will I experience?
    d. Issues I tried looking at outliers, but most of my data is not numerical, so not sure if that will be an issue here

Section 3 – Week 11 – Starting Your Writeups

1. Discuss how you plan to uncover new information in the data that is not self-evident.
    1. Originally when I cleaned my data I was only using one dataset, but that wasn't going to work for the complete project. I re-defined my clean-up process, which I included under Data Import and Cleaning. My first thought to uncover new information was to take the three separate datasets and combine them into one so I could look at the entire picture of crime.
    2. I also separated out my Reported_Date column as it made it hard to plot. By doing this I could focus more on the months and years.
    3. I also plan on using several different plots/graphs to help me get a better visual of the data that will help me drive some of the more technical aspects.I may also consider creating a few different variables to getting a more detailed look at the data.
2. What are different ways you could look at this data to answer the questions you want to answer?
    1. First I looked at crime in KC by Offense over the last 3 years
    2. Then I looked at Offense code by zipcode
    3. Looked at Offense code by year and month
    4. I also need to look at the crime and see the difference in crime with or without a firearm, but I'm still working on those plots as they don't come out correctly initially.
3. Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
    1. As part of my clean-up process I decided to join my 3 separate dataframes into one to make analyzing the data easier. I figure if I need to look at data by month or year I can call that out in my analysis or plots. I also created a subset of data that contains the most relevant values in answering my original questions.
4. How could you summarize your data to answer key questions?
    1. Compare KC crime using firearms and not using firearms
    2. Look at the trend of crime across years and months. This will give us an idea if 1) crime is increasing each year, and 2) is more crime happening at certain times of year.
    3. Determine if there is a certain area in the KC area that experiences more crime
5. What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
    1. I plan to use mostly bar graphs as they give the best visuals for the data I have.
6. What do you not know how to do right now that you need to learn to answer your questions?
    1. One thing I messed up on when cleaning my data was switching Day and Month around when I separated them.
    2. Need help in figuring out how I can visualize my crime with and without a firearm. I'm thinking stacked bar graphs, but I can't seem to get the size of my bars right.
    3. Need help adjusting my bar sizes to be more ledgible.
7. Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
    1. I do plan on using some machine learning however I haven't reached the point of using it yet. I want to be able to predict the trend in crime in the next year or two. I think it's

important to find out whether crime is staying at an average number each year or if we're really experiencing an increase.

Section 4 – Week 12

1. Overall, write a coherent narrative that tells a story with the data as you complete this section.
2. Summarize the problem statement you addressed.
    1. The problem that I was trying to address was is there really an increase in crime involving a firearm or could it just be that's what sells in the media right now. I also wanted to know if there was an increase in crime overall in the Kansas City area.
3. Summarize how you addressed this problem statement (the data used, and the methodology employed).
    1. Once the data was cleaned up, I created a subset of all 3 years combined that consisted Report, Reported Date, Reported Time, Offense, Description, Zip code, Firearm Flag, and Area. I then broke my years up into 3 different subsets for each year with the same columns selected as that's what I had decided was going to give me the most beneficial information.
    2. After I broke everything up I created graphs that would give me a visual of Offense, zipcode, year, month, hour, and firearm used. I did this for all 4 of my subsets to see if there were any differences trend wise.
        1. I wanted to see if there were any trends in time of year or time of day where crime was more likely to happen. I noticed in the graphs that it varied across the different years.
    3. I then created another set of graphs and layered in whether firearms were flagged with Y/N. This gave me a stacked bargraph, but gave me a great visual of the fact that there weren't nearly that many crimes committed using a firearm.
    4. I also looked at what my top 5 crimes were, so I could try and analyze the likelihood of whether a firearm would be used. This is based on the code and then assumption on my part. I couldn't figure out how to graph my top five crimes in relation to a firearm.
        1. 1401 Property Damage – Firearm not likely
        2. 802 Non-Aggravated Assault Domestic – Since it's non-aggravated I'd say firearm not likely again
        3. 640 Stealing from Auto – Firearm not likely since this is usually a crime of opportunity
        4. 702 Auto Theft – Firearm is likely here
        5. 501 Burglary Residence – Firearm is likely
    5. I then plotted my Offenses by year to get a better visual. This was interesting because looking at 2017 and 2018 the crime stayed fairly consistent across the months, but 2016 looked pretty extreme in some months especially December.
    6. I then finished off by looking at linear model, covariance, and correlation. Here I found that my variables that I thought might have some similarities was incorrect and there were pretty huge numbers that showed these weren't really related at all.
4. Summarize the interesting insights that your analysis provided.
    1. I honestly thought firearms would be more of an issue in KC crime. However, as a friend of mine pointed out when I was speaking to her this is all cirme in KC and not all crime is necessarily a crime in which it would "make sense" to use a firearm.

    2. I believe more firmly that the news only reports the “sensational” crimes like homocide and shootings when in reality property damage is the highest Offense seen both year by year and all years combined. When I looked at it for the past 3 years combined it had 34, 535 records.
5. Summarize the implications to the consumer (target audience) of your analysis.
    1. At first glance the data shows that firearms are not nearly as involved in crime in Kansas City as portrayed in the news.
    2. July-October seems to be the heaviest months of crime in the KC area. This was contrary to what I originally believed. I thought it would be heavier around the holidays, but the past three years it was on a downward trend starting in October.
    3. 10am-7pm is the heaviest time period for crime in the area.
6. Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
    1. There were quite a few limitations on my part. I spent many hours trying to slice and dice the data several different ways. One of the biggest issues I had was trying to turn my Y/N from Firearm.Used.Flag into numbers. If could have figured that out I might have opened myself up to more tests and methods to analyze.
    2. One limitation I hadn’t considered is this data is only the crime that was reported, but I’m sure there are several that happened and went unreported. This could definitely skew my data.
    3. I think my biggest issue is that when I started this course R was new to me as well as statistics, so it took me longer to figure things out. I think I could have done a better job had I had more time.
7. In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.