District vs. Crime in Denver, Colorado

Wyatt Romero

Table of Contents

Section I: Research Question

The research question that this analysis will be asking is as follows: Does district influence the amount of crime experienced in Denver, Colorado? This research question hypothesizes that there is a statistically significant correlation that can aid predictions between district and crime rate in the City of Denver. Denver, Colorado, and the surrounding area is heavily populated, the city is divided into seven districts managed by the Denver Police Department. The justification behind asking this question lies in providing the Denver Police Department with information that will allow them to delegate resources towards districts with increased crime and provide constituents of the districts with quick and reliable police response.

Section II: Data Collection

The data that was collected for this analysis comes directly from the City of Denver's Open Data Catalog. The City of Denver describes the dataset as including "criminal offenses in the City and County of Denver for the previous five calendar years plus the current year to date" (Denver, 2022). This dataset is updated Monday through Friday with data based on the National Incident Based Reporting System. The data contains 20 variables and approximately 360,000 incidents reported within the last five years. The variables and their characteristics are presented below:

- incident_id – type: num. Unique ID assigned to each reported crime
- offense_id – type: num. Unique ID assigned to each offense reported
- OFFENSE_CODE – type: chr. A unique code assigned to type of offense
- OFFENSE_CODE_EXTENSION – type: num. A unique numeric code assigned to type of offense
- OFFENSE_TYPE_ID – type: chr

- OFFENSE_CATEGORY_ID – type: chr
- FIRST_OCCURENCE_DATE – type: chr
- LAST_OCCURENCE_DATE – type: chr
- REPORTED_DATE – type: chr
- INCIDENT_ADDRESS – type: chr
- GEO_X – type: num
- GEO_Y – type: num
- GEO_LON – type: num
- GEO_LAT – type: num
- DISTRICT_ID – type: num
- PRECINCT_ID – type: num
- NEIGHBORHOOD_ID – type: chr
- IS_CRIME – type: num
- IS_TRAFFIC – type: num
- VICTIM_COUNT – type; num

An advantage of using this dataset is that it is readily available and updated consistently, creating a more accurate dataset with each update. This reason also presents a disadvantage of using the dataset; there is a lot more cleaning of the data that must be done to analyze the data properly. Therefore, the cleaning must be done on a consistent basis to create and maintain an accurate, up-to-date analysis. The data is readily available to the public though the City of Denver's website, presenting little to no challenges in collecting the data, it is just a matter of gathering the updated data consistently, to maintain an accurate model.

Section III: Data Extraction and Preparation

The dataset was readily usable as a .csv file and required no SQL management to perform analysis. After reading the .csv file into Rstudio, the dim() function is called on the dataset to see the dimensions, or shape, of the data. The function, str(), is then called to view the structure of the data, along with the summary() function to gain more information about each variable. The dataset is then reduced to the variables needed to analyze crime rate per district. On this newly formed, reduced dataset, a check is performed for null or missing values, those values are then imputed or omitted, based on number of values missing or null. The data set is then written as a new .csv file.

```
crime <- read_csv('Downloads/Crime.csv')
```

```
Rows: 361027 Columns: 20
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (8): OFFENSE_CODE, OFFENSE_TYPE_ID, OFFENSE_CATEGORY_ID, FIRST_OCCURRENCE_DATE, LAST_OCCURRENCE_DATE, REPORTED_DATE...
dbl (12): incident_id, offense_id, OFFENSE_CODE_EXTENSION, GEO_X, GEO_Y, GEO_LON, GEO_LAT, DISTRICT_ID, PRECINCT_ID, IS_...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(crime)
```

```
[1] 361027     20
```

```
str(crime)
```

```
spec_tbl_df [361,027 × 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ incident_id           : num [1:361027] 2.02e+10 2.02e+07 2.02e+07 2.02e+07 2.02e+07 ...
 $ offense_id            : num [1:361027] 2.02e+16 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...
 $ OFFENSE_CODE          : chr [1:361027] "2999" "2999" "2999" "2999" ...
 $ OFFENSE_CODE_EXTENSION: num [1:361027] 0 0 0 0 0 0 0 0 0 0 ...
 $ OFFENSE_TYPE_ID       : chr [1:361027] "criminal-mischief-other" "criminal-mischief-other" "criminal-mischief-other" "criminal-mischief-other" ...
 $ OFFENSE_CATEGORY_ID   : chr [1:361027] "public-disorder" "public-disorder" "public-disorder" "public-disorder" ...
 $ FIRST_OCCURRENCE_DATE : chr [1:361027] "1/4/2022 11:30:00 AM" "1/3/2022 6:45:00 AM" "1/3/2022 1:00:00 AM" "1/3/2022 7:47:00 PM" ...
 $ LAST_OCCURRENCE_DATE  : chr [1:361027] "1/4/2022 12:00:00 PM" NA NA NA ...
 $ REPORTED_DATE         : chr [1:361027] "1/4/2022 8:36:00 PM" "1/3/2022 11:01:00 AM" "1/3/2022 6:11:00 AM" "1/3/2022 9:12:00 PM" ...
 $ INCIDENT_ADDRESS      : chr [1:361027] "128 S CANOSA CT" "650 15TH ST" "919 E COLFAX AVE" "2345 W ALAMEDA AVE" ...
 $ GEO_X                 : num [1:361027] 3135366 3142454 3147484 3136478 3169237 ...
 $ GEO_Y                 : num [1:361027] 1685410 1696151 1694898 1684414 1705800 ...
 $ GEO_LON               : num [1:361027] -105 -105 -105 -105 -105 ...
 $ GEO_LAT               : num [1:361027] 39.7 39.7 39.7 39.7 39.8 ...
 $ DISTRICT_ID           : num [1:361027] 4 6 6 4 5 6 3 6 3 1 ...
 $ PRECINCT_ID           : num [1:361027] 411 611 621 411 512 621 312 623 311 123 ...
 $ NEIGHBORHOOD_ID       : chr [1:361027] "valverde" "cbd" "north-capitol-hill" "valverde" ...
 $ IS_CRIME              : num [1:361027] 1 1 1 1 1 1 1 1 1 1 ...
 $ IS_TRAFFIC            : num [1:361027] 0 0 0 0 0 0 0 0 0 0 ...
 $ VICTIM_COUNT          : num [1:361027] 1 1 1 1 1 1 1 1 1 1 ...
```

summary(crime)

```
  incident_id          offense_id        OFFENSE_CODE       OFFENSE_CODE_EXTENSION OFFENSE_TYPE_ID    OFFENSE_CATEGORY_ID
 Min.   :2.020e+04   Min.   :2.020e+10   Length:361027      Min.   :0.0000         Length:361027      Length:361027
 1st Qu.:2.018e+09   1st Qu.:2.018e+15   Class :character   1st Qu.:0.0000         Class :character   Class :character
 Median :2.020e+09   Median :2.020e+15   Mode  :character   Median :0.0000         Mode  :character   Mode  :character
 Mean   :5.677e+09   Mean   :5.677e+15                      Mean   :0.2633
 3rd Qu.:2.022e+09   3rd Qu.:2.022e+15                      3rd Qu.:0.0000
 Max.   :2.021e+12   Max.   :2.021e+18                      Max.   :5.0000

 FIRST_OCCURRENCE_DATE LAST_OCCURRENCE_DATE REPORTED_DATE      INCIDENT_ADDRESS       GEO_X              GEO_Y
 Length:361027         Length:361027        Length:361027      Length:361027      Min.   :       1   Min.   :       1
 Class :character      Class :character     Class :character   Class :character   1st Qu.: 3139841   1st Qu.: 1683183
 Mode  :character      Mode  :character     Mode  :character   Mode  :character   Median : 3146086   Median : 1694802
                                                               Mean   : 3156584   Mean   : 1693516
                                                               3rd Qu.: 3164305   3rd Qu.: 1701690
                                                               Max.   :40674766   Max.   :10890452
                                                               NA's   :4738       NA's   :4738
    GEO_LON           GEO_LAT        DISTRICT_ID     PRECINCT_ID     NEIGHBORHOOD_ID       IS_CRIME   IS_TRAFFIC
 Min.   :-115.5   Min.   : 0.00   Min.   :1.00   Min.   :111.0   Length:361027      Min.   :1   Min.   :0
 1st Qu.:-105.0   1st Qu.:39.71   1st Qu.:2.00   1st Qu.:222.0   Class :character   1st Qu.:1   1st Qu.:0
 Median :-105.0   Median :39.74   Median :3.00   Median :324.0   Mode  :character   Median :1   Median :0
 Mean   :-104.9   Mean   :39.73   Mean   :3.65   Mean   :382.9                     Mean   :1   Mean   :0
 3rd Qu.:-104.9   3rd Qu.:39.76   3rd Qu.:5.00   3rd Qu.:523.0                     3rd Qu.:1   3rd Qu.:0
 Max.   :   0.0   Max.   :39.90   Max.   :7.00   Max.   :759.0                     Max.   :1   Max.   :0
 NA's   :5321     NA's   :5321    NA's   :585    NA's   :585
  VICTIM_COUNT
 Min.   : 1.000
 1st Qu.: 1.000
 Median : 1.000
 Mean   : 1.019
 3rd Qu.: 1.000
 Max.   :32.000
```

```
crime <- select(crime, c('incident_id', 'DISTRICT_ID'))

head(crime)
```

A tibble: 6 × 2

incident_id <dbl>	DISTRICT_ID <dbl>
20226000193	4
20223319	6
20223093	6
20224000	4
20223956	5
20223903	6

6 rows

```
crime[crime == "?"] <- NA

colSums(is.na(crime))



crime <- na.omit(crime)
```

```
incident_id DISTRICT_ID
          0         585
```

```
colSums(is.na(crime))
```

```
incident_id DISTRICT_ID
          0           0
```

```
write.csv(crime, "crime_capstone.csv", row.names = TRUE)
```

The entire dataset was then imported into Tableau to view the geographical data and continue the analysis on a geographical scale and create an environment in which the Denver Police Department could then view and interact on the crime data with. The Tableau environment

will be provided as an external attachment, but results will be discussed further in the analysis section.

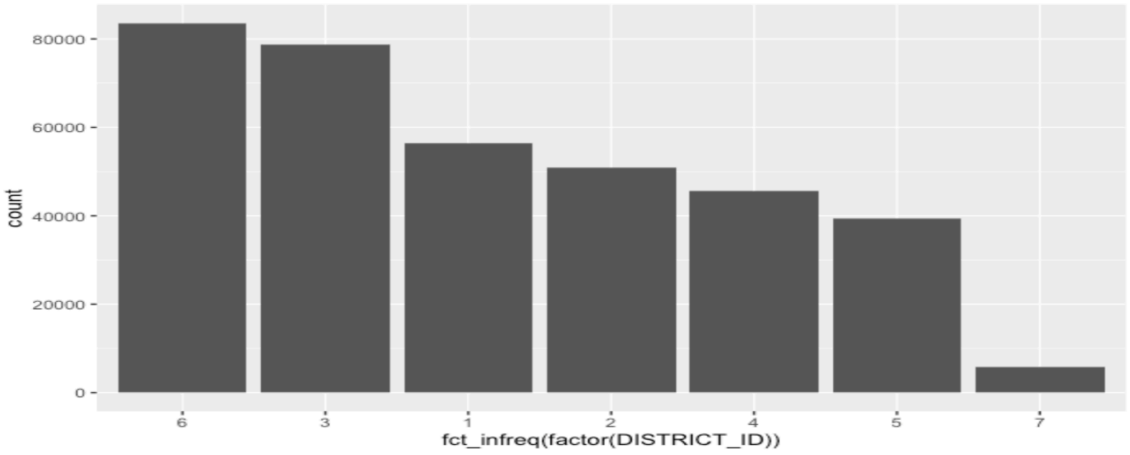
RStudio, along with R, was utilized in the data preparation process due to its statistical prowess, extensive library, and visualization capabilities, but a disadvantage of R lies in that it tends to run rather slowly due to many of its functions being spread across many packages. Tableau is used for its ability to easily view the geographical data and create an interactive environment. A disadvantage of Tableau lies in the fact that it requires knowledge of SQL syntax, if the police department themselves would like to do some analysis, they would need some knowledge of SQL syntax, otherwise, if they have a question that could be answered through Tableau, the department would need to have another government department look into it, taking resources from those departments in the meantime.

Section IV: Analysis

To further analyze the data, the number of crimes reported per district will be calculated through the table() function, as well as creating visualizations to determine the percentage of crime each district experiences compared to the other districts in Denver. The table() function is a simple analysis technique that will count and display the frequency of a given variable. In the case of this analysis, the function will count and display the frequency of each DISTRICT_ID associated with an incident_id. With the statistical confirmation gained from the analysis in R, the same data will be represented in Tableau to view the districts on a map and create the interactive environment to view crime statistics. Due to the geographical nature of the data, Tableau is an incredibly advantageous tool to use for its ease of use and user-friendly interactivity, but the amount of data that we have, can make it difficult to pinpoint a specific data

point as the districts are crowded with data, but the amount of data also helps define the district lines.

```
graph <- ggplot(filter(crime), aes(fct_infreq(factor(DISTRICT_ID)))) + geom_bar()
graph
```




```
as.data.frame(table(crime$DISTRICT_ID))
```

Description: df [7 × 2]

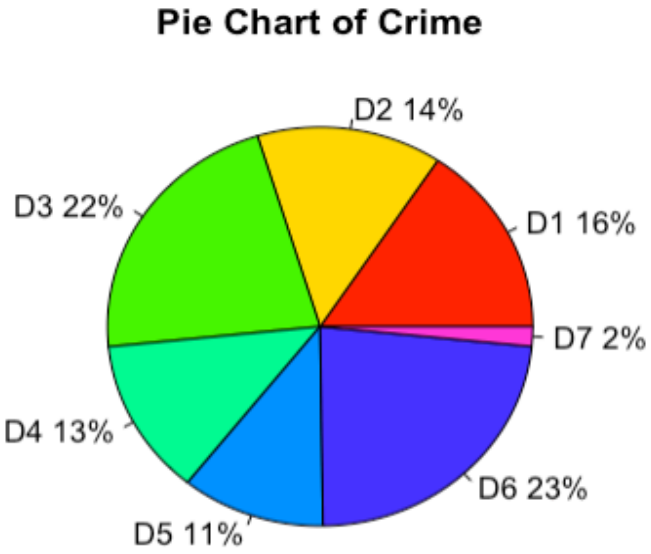
Var1 <fctr>	Freq <int>
1	56478
2	50802
3	78771
4	45707
5	39354
6	83586
7	5744

7 rows

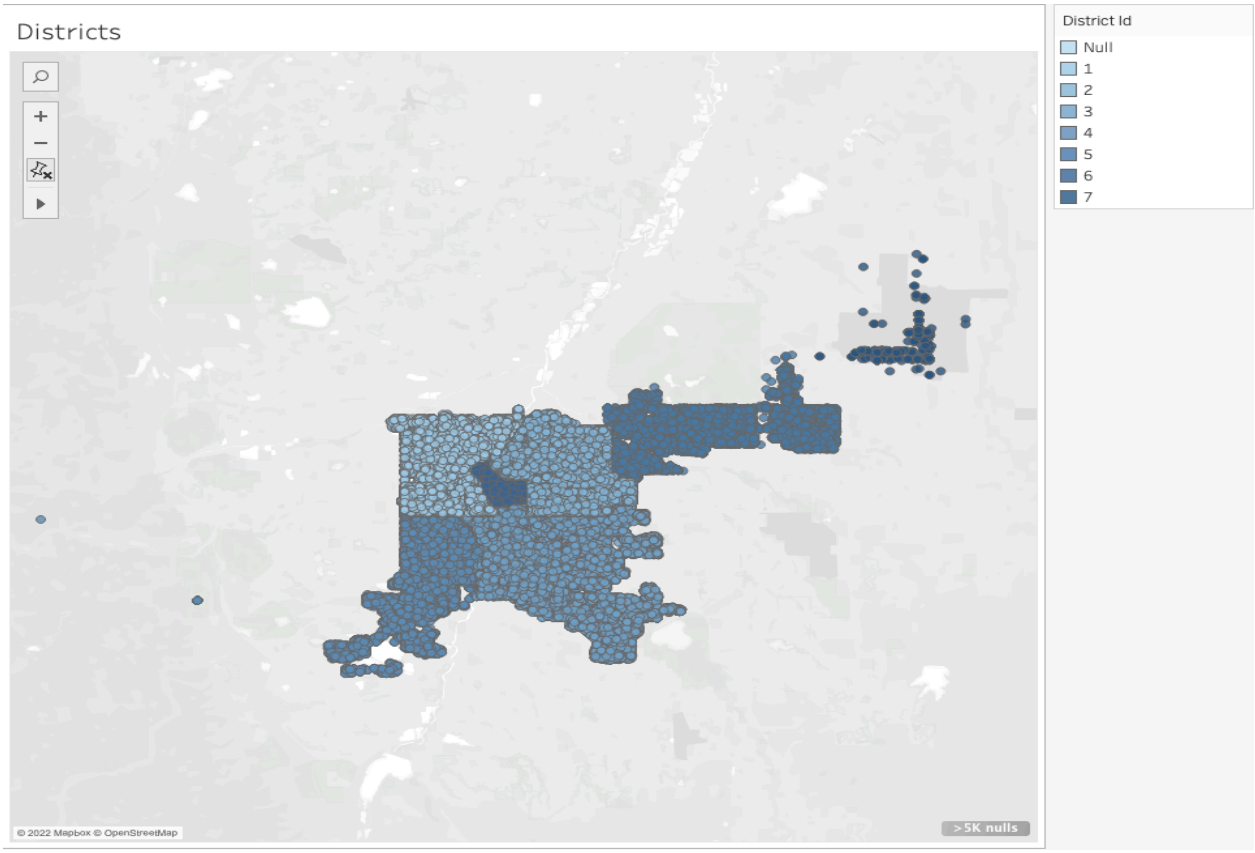
```
slices <- c(56478, 50802, 78771, 45707, 39354, 83586, 5744)
labels <- c("D1", "D2", "D3", "D4", "D5", "D6", "D7")
percent <- round(slices/sum(slices)*100)
labels <- paste(labels, percent)
```

```
labels <- paste(labels, "%", sep="")

pie(slices, labels = labels, main = "Pie Chart of Crime", col=rainbow(length(labels)))
```



The Tableau map was created by importing the .csv file and creating a new sheet. The Geo Lat and Geo Lon variables are then set into the Rows and Columns spaces respectively. This will generate a map. The Incident Id variable should then be converted from dimension to measure and then be renamed to Incident Cnt. The Incident Cnt variable should then be brought to the marks card a detail which then create a dot for each reported crime. The District Id variable should be brought to the marks card as a color, which will then color each dot according to the district it belongs to, providing visual representations of each district. Then the tooltip should be edited to include statistics relating to each crime. Bring the Sum of the Victim Count variable to the marks card as a tooltip as well as the Offense Category Id variable. Click into the tooltip extension of the marks card to edit the text and import the newly added variables to the marks card. Now, when any point on the map is highlighted, the tooltip will display information regarding the district, the incident id, the victim count, and the offense category.



Section V: Data Summary and Implications

Through the analysis in R, it is identified that District 6 and District 3 experience the most crime out of any of the districts, and combined, nearly 50% of all crime in Denver. District 6 experiences 23% of the crime, and District 3, 22% of the crime. An interesting point that is revealed through the Tableau Analysis, is that District 6 is also the smallest district out of the 7 in Denver, yet it experiences 23% of the crime, further analysis of neighborhood identifications may reveal further details as to why this may be. A limitation of this analysis is that, as the data is consistently updated, these statistics are not set in stone and could change significantly over time. From the analysis above, it can be concluded that, as of June, 2022, 2 districts dominate the overall crime rate in Denver, Colorado. Those districts are 6 and 3, both with over 20% of the crime reported over the last 5 years experienced in their districts, the next closest district is

district 1 with 16% of the crimes reported. These results should lead the Denver Police Department to delegate resources towards District 6 and District 3 to provide those districts with resources to reduce crime while also increasing police presence to provide constituents of those districts with quick and reliable police response.

For further analysis on this data, I suggest using the IS_TRAFFIC and IS_CRIME variables to determine what percentage of crimes committed are traffic related and if that is influenced by the number of cars in Denver, Colorado. One could also go in the direction of figuring the crime statistics at the neighborhood level to identify the neighborhoods that experience the most crime in Denver, Colorado.

Section VI: Sources

Denver. (2022). *Crime*. Denver Open Data Catalog: Crime. Retrieved June 30, 2022, from https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime

Google Developers. (2021, January 13). *K-means advantages and disadvantages | clustering in machine learning | google developers*. Google. Retrieved June 30, 2022, from https://developers.google.com/machine-learning/clustering/algorithm/advantages- disadvantages

TechVidvan. (2021, July 6). *Should you start learning R? weigh the pros and cons of R programming*. TechVidvan. Retrieved June 30, 2022, from https://techvidvan.com/tutorials/pros-and-cons-of-r/

Thinklytics. (2020, December 14). *What are the Pros & Cons of tableau?* Thinklytics. Retrieved June 30, 2022, from https://thinklytics.com/what-are-the-pros-cons-of-tableau/