Data Exploration Report

Charlie Bertelsen

2023-10-20

PROBLEM STATEMENT:

During the pandemic, stay-at-home orders and lock down procedures greatly decreased the total number of drivers on the roads. In the second largest city in the U.S. with some of the worst traffic, Los Angeles (LA), the impact of the pandemic on the roads was very apparent. Understanding the relationship between COVID-19 cases/deaths and the occurrence of traffic collisions would be extremely beneficial for public safety planning and policy-making. Identifying potential correlations or trends can provide valuable insights into the interplay between pandemic-related restrictions, human mobility, and road safety. By understanding these dynamics, local authorities can develop targeted strategies to enhance public health measures and optimize traffic management protocols to create a safer environment for all residents.

LA COVID data:

https://catalog.data.gov/dataset/la-county-covid-cases

LA Traffic Collision Data:

https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present

SUMMARY OF DATA ENGINEERING

Data Cleaning:

The first step in my data cleaning process was to remove columns that will not be used in my analysis. These columns were county, state, fips code, people tested (empty), state cases, state deaths, new state cases, and new state deaths. For my traffic data set, I removed DR number, date reported, area ID, crime code and description, lat/long, address and cross street location. Most of this information was the same for the entire column such as county or crime code. Any other removed columns were removed because they would not be needed to answer my main question. The second step in my data cleaning process was to check for NULL values in both of the data sets. What I found was that the only column in both of the data sets with any missing values was the "Victim.Age" column in the Traffic data set. This meant I did not have to remove any rows because no vital information was missing such as the date. The next step of my cleaning process was sorting each data set by date. I did this by first formatting the date columns as dates with as.date() instead of characters, and then ordering them in ascending order with order(). The final thing that I had to do to clean the data before joining it was eliminating duplicate rows. I did this for both dataframes using the distinct() function to create two new dataframes with only the first occurrence of a row.

Data Joining:

I used a full join to join my covid and traffic datasets because I wanted to keep the traffic data from approximately 2010-2020 so I could have pre-pandemic data to compare traffic collisions before and during the pandemic. Before I was able to make the join, I first had to change the column name "Date.Occured" in the Traffic dataset to just "date" to match the covid dataset. Once this was done, I created a new dataframe with just the date and number of traffic collisions which I was able to combine with the COVID data set using the full_join() function for a seamless operation.

DATA VISUALIZATIONS:

For better data visualizations, I performed more data cleaning steps. I also used the ggplot2 package for many of the plots instead of the standard r plots for better visualizations and made sure to include labels and titles on each of the graphs.

Univariate Charts:

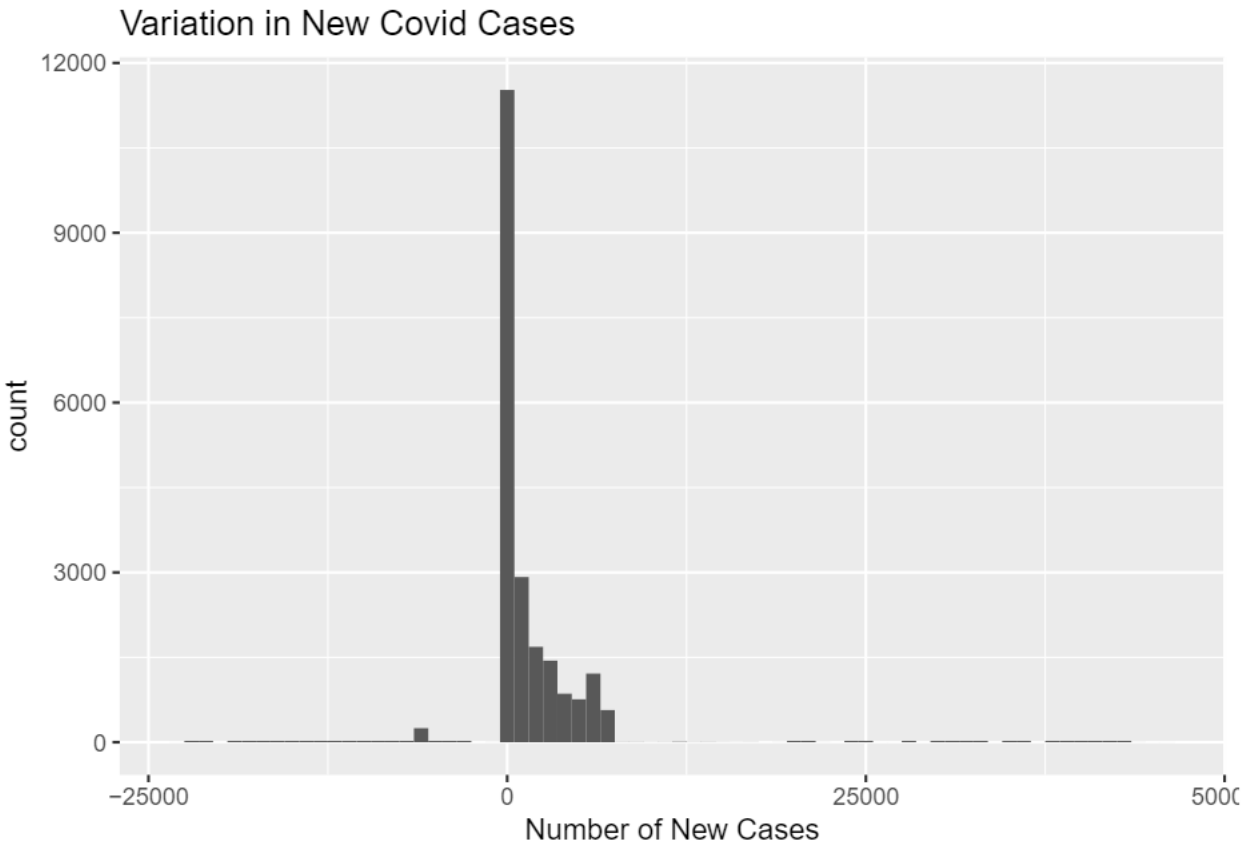

Figure 1: Variation in the number of new covid cases each day in Los Angeles California.

In this improved version of the plot I used ggplot and added labels and a title. This Variation in Cases chart represents the variation in the number of new cases recorded in LA. This graph is important because it tells us that many of the dates in the dataset recorded very small changes in new cases. The large axis of the graph tells us that there are some outliers or some weird patterns that will have to be examined within the data.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

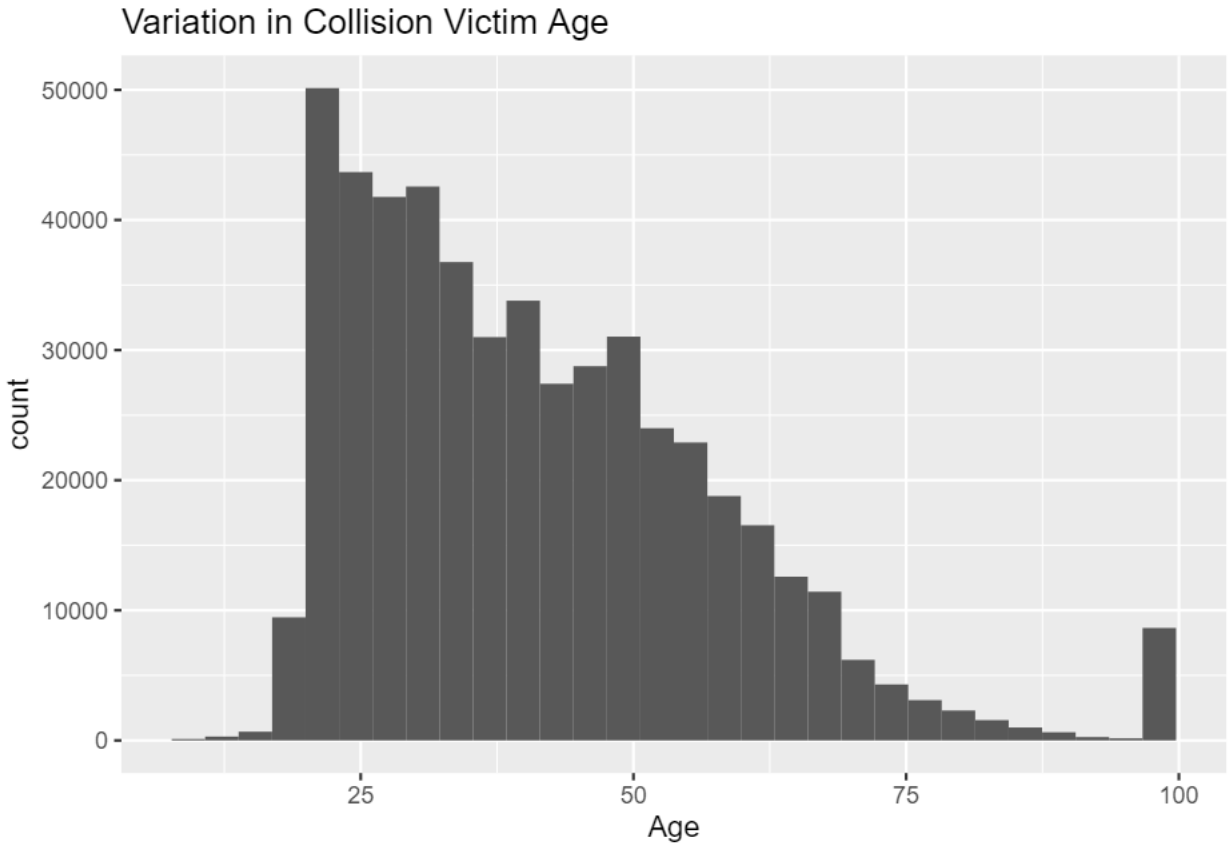

Figure 2: Histogram representing the variation in victim age for traffic collisions from 2010-2023.

This graph represents the variation in ages that were involved in traffic collisions. From the graph we can see that there is a relatively normal distribution with a peak around the early to late twenties meaning new drivers are getting in the most traffic collisions. In this improved version of the plot I used ggplot and added labels and a title.

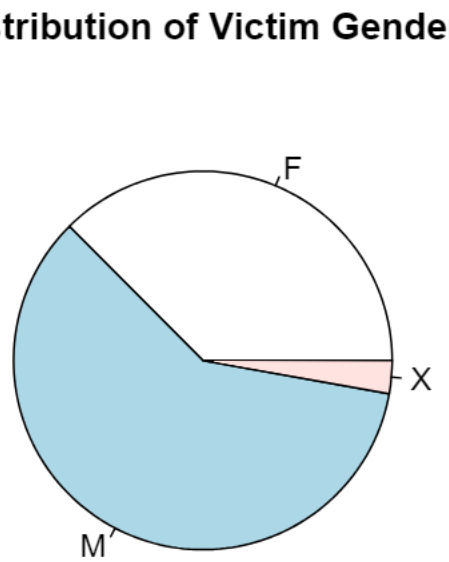

Figure 3: Distribution of genders involved in traffic collisions where M = Male, F = Female, X = Unknown.

This Pie chart represents the distribution of traffic collisions by gender (M: Male, F: Female, X: Other). From the graph we can see that males account for a significantly larger portion of the collisions than females. This information is important because it gives me some more background on the demographic most frequently involved in traffic collisions.

Bivariate Charts:

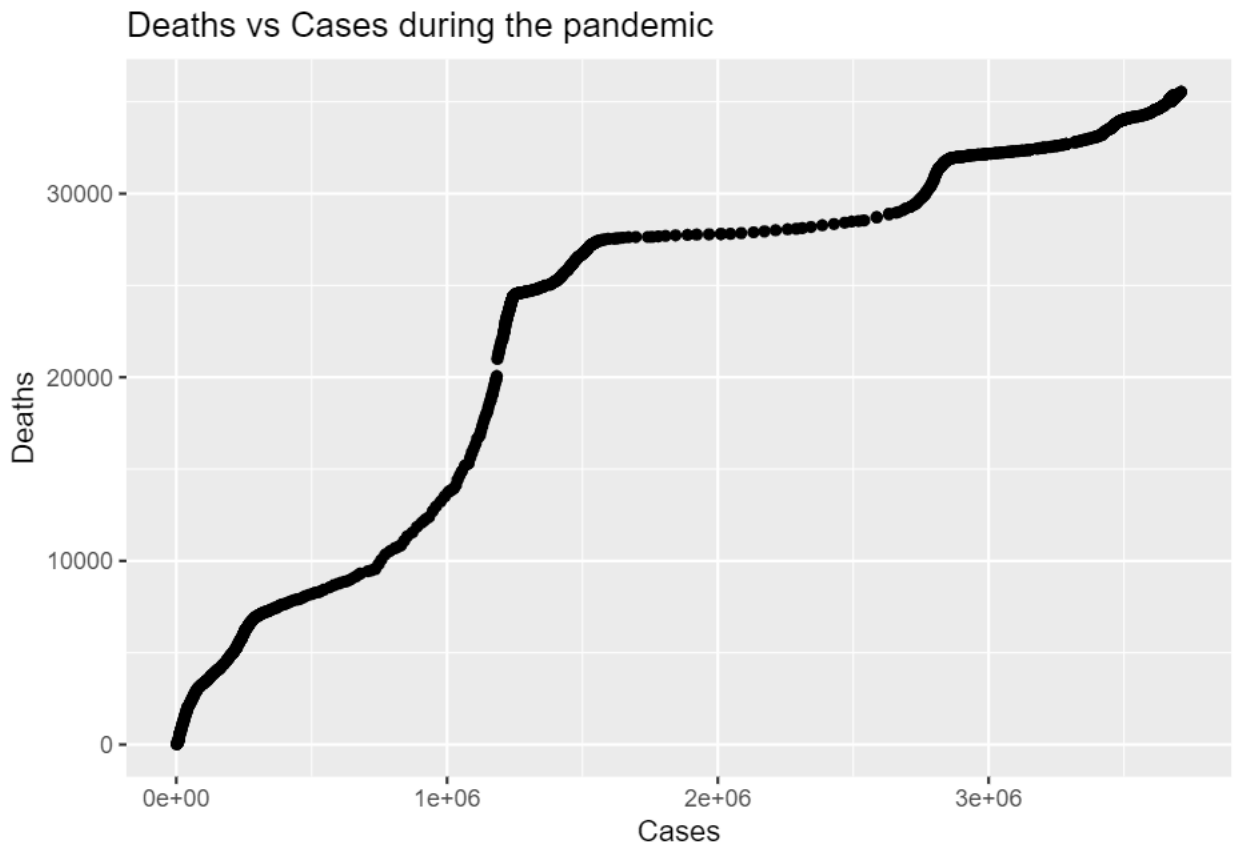

Figure 4: Scatterplot showing the relationship between the total number of cases and the total number of deaths in LA.

This chart represents the relationship between total cases and deaths in Los Angeles. There is a clear linear trend and strong relationship between these two variables. This is important because it helps me to see how strongly these variables are correlated and helps me figure out some more background on Covid during this time.

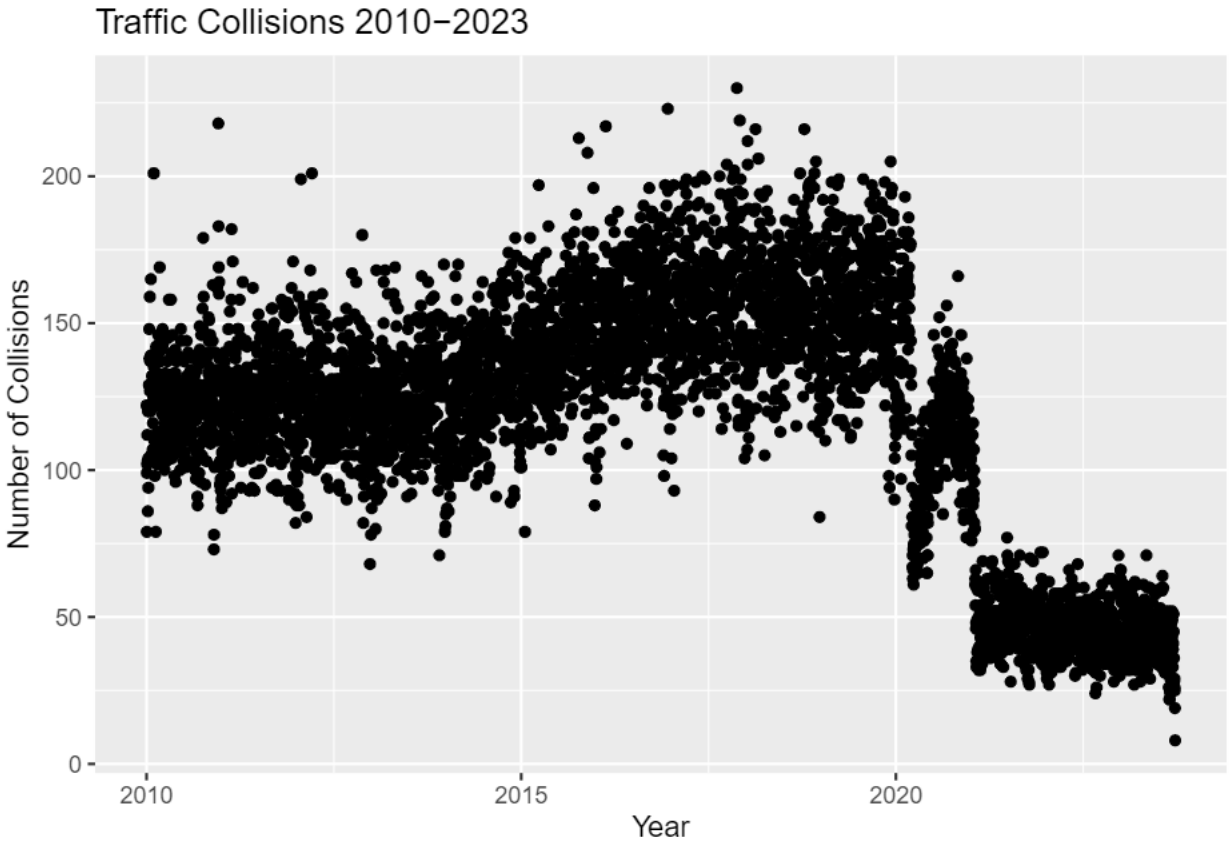

Figure 5: Scatterplot showing the number of traffic collisions on each day from 1/1/2020 to 9/23/2023.

This scatterplot shows how the number of traffic collisions changed from 2010-2023. Based on this plot it is clear that there was a significant decrease in traffic collisions around the pandemic. This means that I was correct in assuming a relationship between Covid cases/deaths and collisions, and I will be able to explore this relationship further.

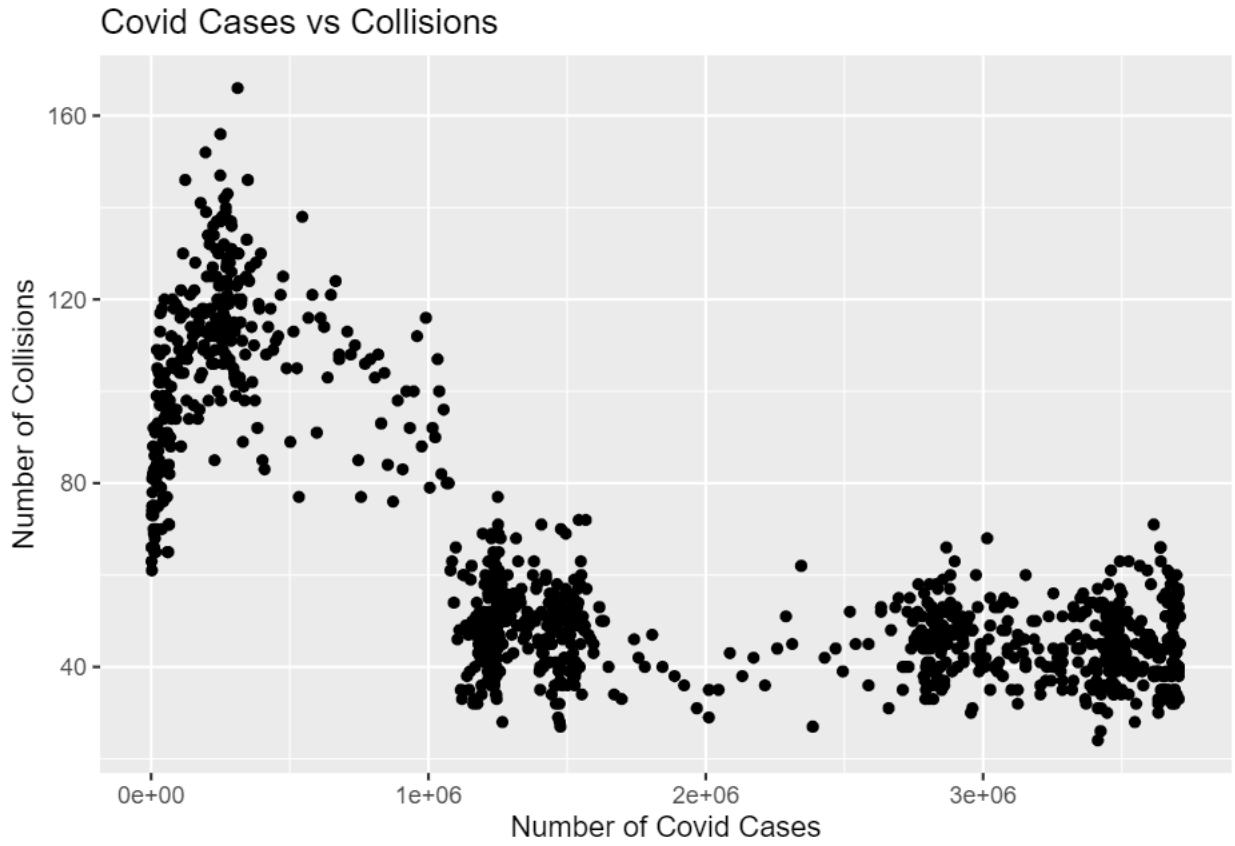

Figure 6: Scatterplot showing the number of covid cases vs the number of traffic collisions.

This chart is a scatterplot showing the relationship between the number of Covid cases and the number of traffic collisions. Based on the graph there is a pretty apparent inverse relationship between collisions and cases, as cases increase the number of collisions decreases. There are some large clumps of data in the scatterplot which means that cluster analysis may be a useful technique to discover patterns in the data.

High-Dimensional Charts:

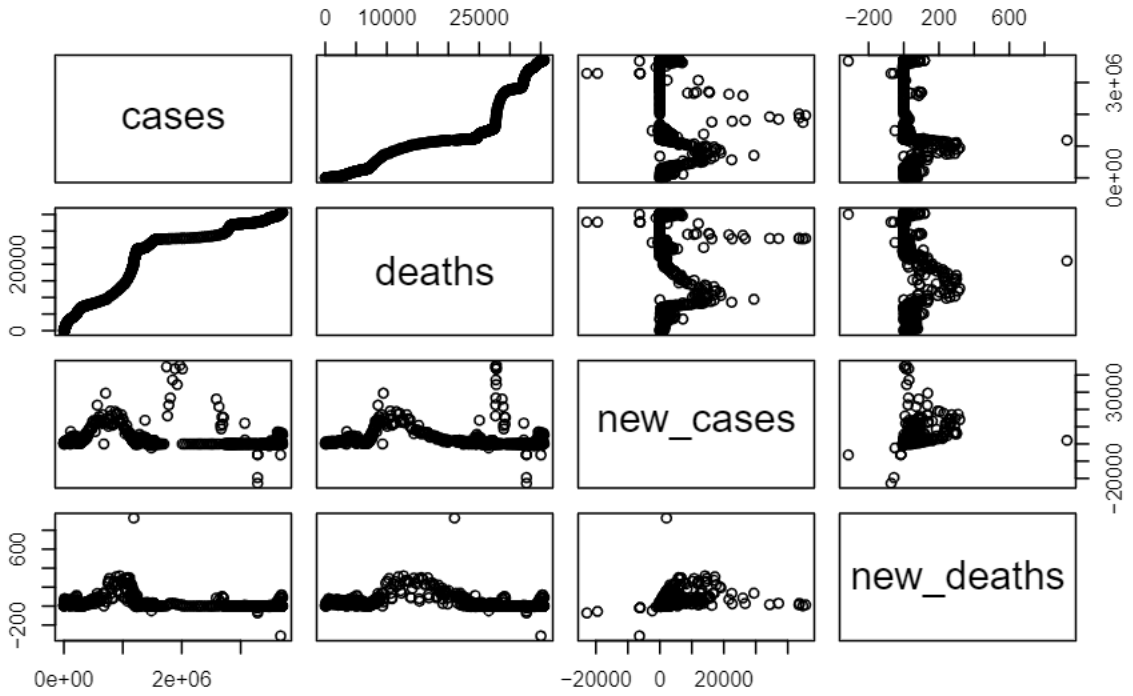

Figure 7: Scatterplot matrix chart showing the relationships between each of the variables in the Covid dataset.

This scatterplot matrix shows the relationship between the numeric values in the COVID dataset. This dataset shows that some variables have much stronger correlations than others. In this improved version of the scatterplot matrix I removed the columns that were not needed in my analysis such as state statistics. This allowed me to see that all variables have pretty significant relationships except for the relationship between new cases and new deaths.

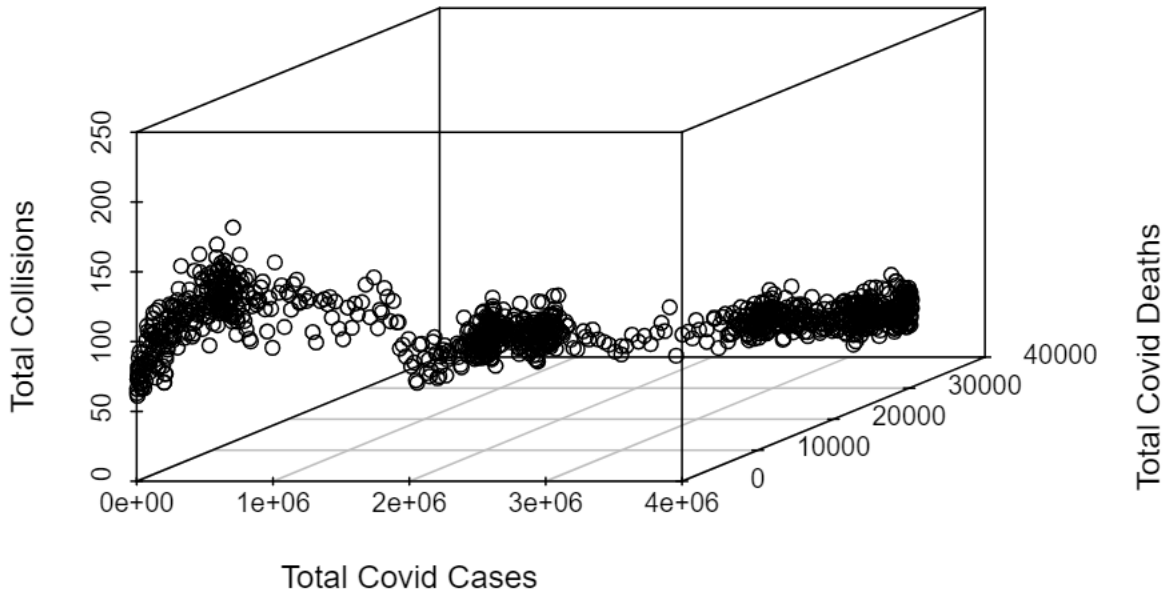

Figure 8: Scatterplot matrix chart showing the relationships between each of the variables in the Covid dataset.

This second high dimensional chart is a 3d scatterplot showing the cases, deaths and traffic collisions in LA. In this new chart I used traffic collisions on the Z-axis rather than new cases. Based on the plot cases and deaths have a strong correlation and there is a significant enough relationship between these two and traffic collisions to make it worth my time to evaluate the relationships between these three variables.

Other Charts:

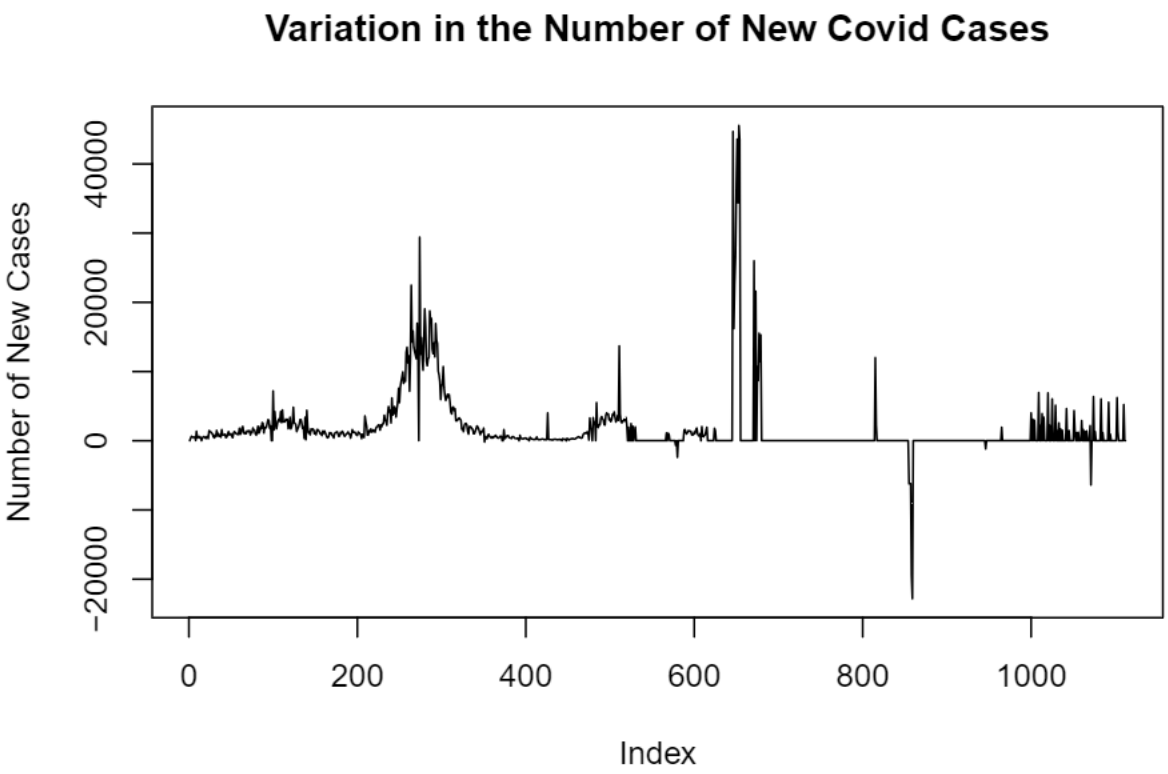

Figure 9: Line graph showing the changes in the new cases column.

This last chart is a line graph of the number of new cases in Los Angeles. In this improved version of my previous line chart, you can clearly see that after the additional data cleaning steps the issue with weird random negative new case numbers is gone. This chart is important because it tells me that new cases were not consistently reported throughout the data and this column may not be a reliable column to use in my analysis.

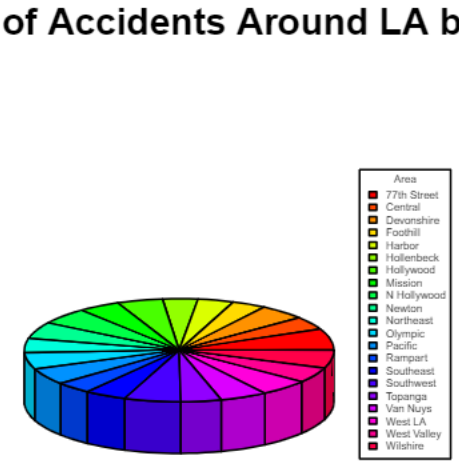

Figure 10: Pie chart showing the distribution of accidents in the different areas.

This 3d pie chart shows the distribution of traffic collisions by area. I created this chart to see if some areas are more susceptible to traffic collisions but that is clearly not the case. This is important because this tells me that it may not be worth my time to explore the differences between accidents in different parts of LA.

In this improved version of the chart, there is a legend that shows which area corresponds to each of the different slices.

THREE DATA EXPLORATION TECHNIQUES:

Method 1: Outlier Detection

I selected outlier detection as my first data exploration technique because it is a data mining method that can tell you if you need to do more data cleaning. Outlier detection is very relevant to my final project outcome because it is important to have the data as clean and ready as possible before fitting models and performing further exploration to make results more accurate. Finding outliers allows me to determine if there are any columns in my data that need further processing. The three main columns in which outliers will be important are covid cases, covid deaths, and traffic collisions. Outlier analysis on this column is below.

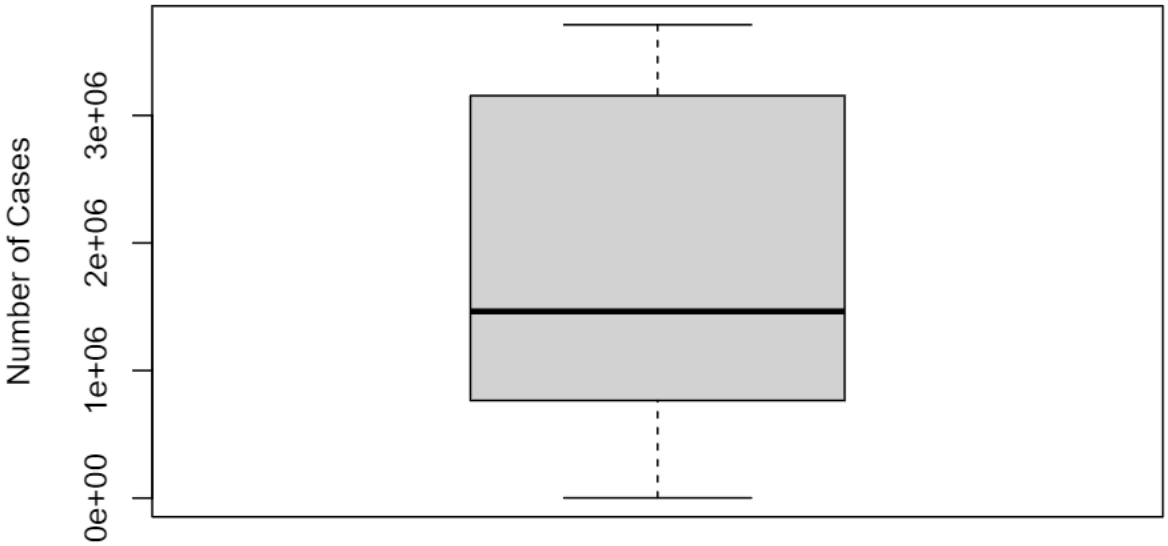

Figure 11: Boxplot showing the min, max, mean and quartiles of covid cases.

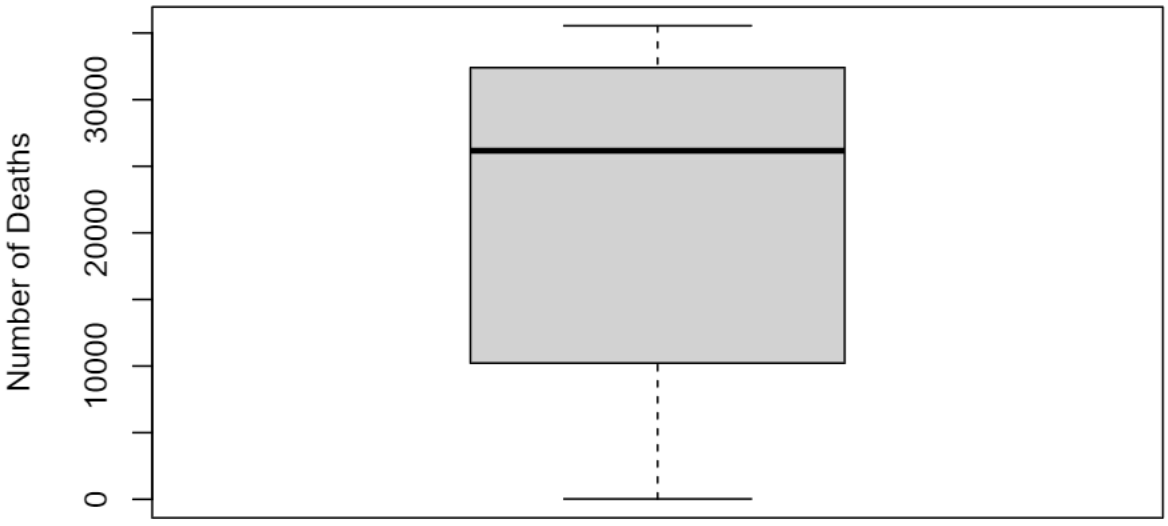

Figure 12: Boxplot showing the min, max, mean and quartiles of covid deaths.

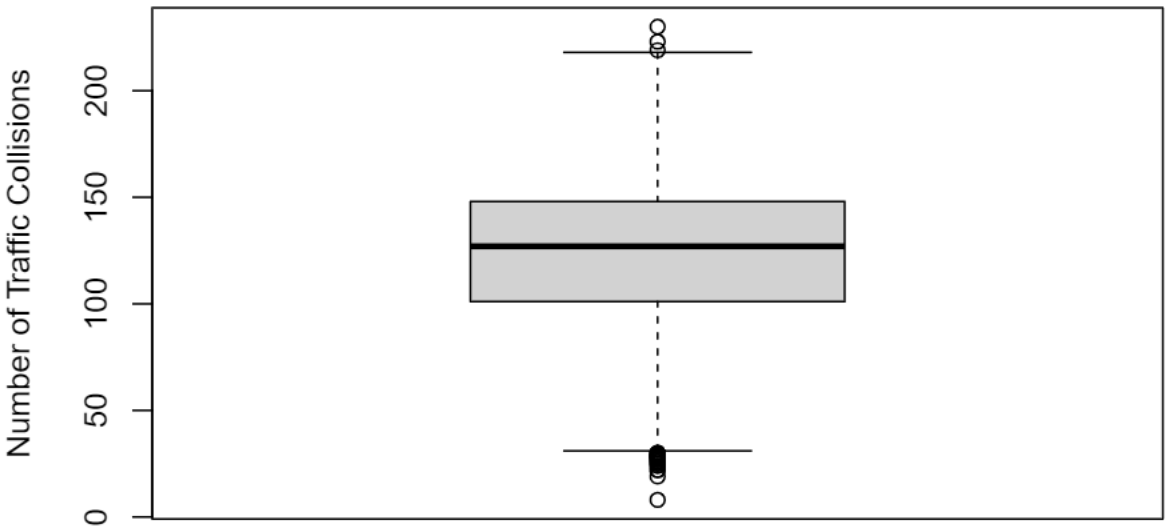

Figure 13: Boxplot showing the min, max, mean and quartiles of traffic collisions.

Based on the boxplots above, the only column that may have outliers is number of traffic collisions:

```
## 2.5%
##   36
```

```
## 97.5%
##   183
```

```
##   [1] 201 218 199 201 197 187 213 208 196 217 185 185 188 186 190 196 188 186
##  [19] 188 186 184 197 190 223 195 184 197 186 188 185 188 197 194 199 190 185
##  [37] 198 190 191 200 199 189 185 200 194 204 192 190 194 186 191 199 184 200
##  [55] 202 188 230 195 186 219 199 194 188 194 212 204 189 188 216 189 206 194
##  [73] 184 193 195 188 185 185 184 192 201 188 185 190 186 216 196 193 197 198
##  [91] 196 201 205 192 184 188 198 192 184 187 184 197 199 186 185 185 185 199
## [109] 191 197 189 194 185 184 191 189 198 188 205 194 196 185 187 193 186  35
## [127]  33  35  32  33  32  34  34  34  33  28  35  34  32  29  32  28  27  34
## [145]  34  33  31  29  35  35  27  31  35  33  35  33  33  35  30  31  35  35
## [163]  32  35  34  35  35  33  32  35  24  31  26  31  34  30  35  34  34  34
## [181]  35  34  28  32  30  32  33  35  35  35  35  33  33  33  35  35  32  35
## [199]  35  32  34  34  33  27  35  32  32  32  34  35  28  33  31  33  34  32
## [217]  34  29  35  35  34  33  31  35  32  26  31  22  22  25  30  24  29  35
## [235]  33  28  29  26  25  19   8
```

Based on this percentile evaluation method anything below 36 (2.5%) or above 183 (97.5%) could be a potential outlier. After searching for all values that do not fall within this range, the minimum collisions was 8 while the maximum was 230. Both of these are completely reasonable numbers for an entire city which means I will not have to remove any rows, and there are no incorrectly entered values which may have a significant impact on my results. Using this clean data and knowing that there are no significant outliers in the columns that I will be using, I can move on to further data exploration and move closer to my project goals knowing my results will be more accurate.

Method 2: Regression

I selected multiple regression as my second advanced exploration technique because it is a very good tool for finding details about variable relationships. I chose regression instead of other techniques because it is very straightforward and it can help me to understand the direction and magnitude of the relationship between Covid cases/deaths and traffic collisions, which is a major part of my project goal.

```
##
## Call:
## lm(formula = Total_Collisions ~ cases + deaths, data = Covid_Dates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.064  -7.609   0.154   7.914  68.596
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.172e+02  1.112e+00  105.42   <2e-16 ***
## cases        1.100e-05  9.266e-07   11.88   <2e-16 ***
## deaths      -3.274e-03  1.036e-04  -31.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.07 on 1109 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7356
## F-statistic:  1546 on 2 and 1109 DF,  p-value: < 2.2e-16
```

Using the summary table above I can see that the p-values for Covid cases and deaths are both extremely small ($p<0.05$) which means that it is very likely that there is a statistically significant relationship between these two variables and traffic collisions. In addition to this, the multiple R-squared value is 0.736 which means that the predictor variables account for about 74% of the variation in traffic collisions. This high R-squared value further proves to me that this relationship is worth exploring. This information tells me that I am on the right track to understanding the relationship between Covid cases/deaths and traffic collisions which I will be able to use to accomplish my project goal of improving road safety with statistically backed traffic claims.

Method 3: K-means Clustering

The third data exploration technique that I chose is K-means clustering. I chose to use a clustering method because I wanted to explore the 3-4 clusters that can be seen in the above cases vs collisions graph as well as the 3d scatterplot. In both cases there appears to be 3-4 close clusters of data which means there is likely relationship between traffic collisions during around different dates.

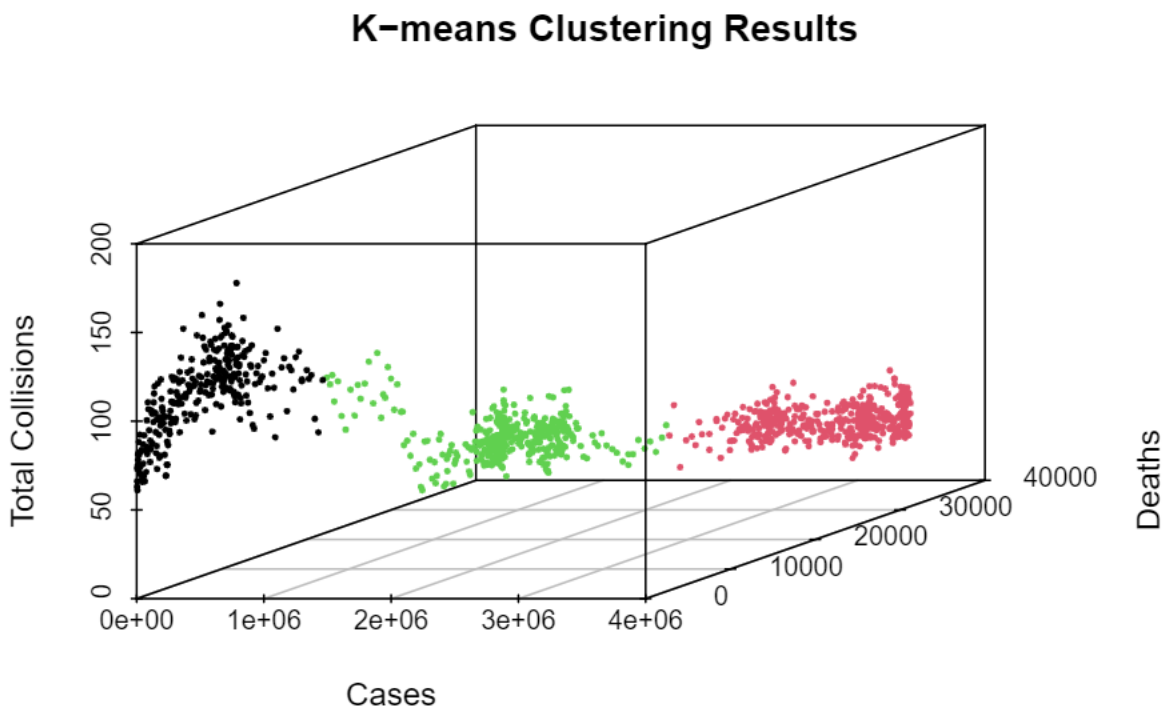

Figure 14: 3D scatterplot showing the relationship between Covid cases/deaths and traffic collisions divided into clusters.

It is clear after looking at the clustered 3d scatterplot that there are three significant groups when looking at cases, deaths and collisions as a whole. These three groups were likely formed during different stages of the pandemic which would explain the large difference between them. You can see from this chart that as the pandemic went on, the number of traffic collisions decreases and the number of cases and deaths increase. This change is shown by the slow fall from the top left of the 3d box to the bottom right where collisions are low, and cases/deaths are high. I can use this information to look for different dates during the pandemic where the inverse relationship between traffic collisions and cases/deaths was strong to better support my arguments.

CONCLUSION:

My outlier analysis showed me that the datasets that I chose did not contain any significant points that would have an affect on further data exploration. This is important because it allows me to continue my project with higher confidence of my results. My regression analysis proved that I was right in assuming a relationship between Covid cases/deaths and traffic collisions. The regression summary showed a strong inverse relationship that I will be able to explore further. My cluster analysis helped me to see the inverse relationship in a 3-dimensional space and showed me that there are certain time periods where values were similar which may be useful when proceeding with my analysis. Overall I will use the insights that I gained from these data exploration methods to continue my analysis and find evidence to support the relationship between Covid cases/deaths and traffic collisions. Using this evidence I will be able to better advocate for road safety and policies that decrease per capita automobile use bringing me closer to my final project goals.