Car Accidents and Population Growth in Baton Rouge, LA, Tempe, AZ, and Seattle, WA

Alejandro Calderon, Dylan Craker, Trent Douglas

6/07/2021

1. Introduction

For our final report we decided to study car crash data in 3 US cities, Baton Rouge, Louisiana; Seattle, Washington; and Tempe, Arizona. We chose these three cities because of their similarities and differences. Baton Rouge is a relatively stagnant city with respect to population and area, neither having grown much in the past 10 years, with population actually slightly declining. Seattle, because it has not increased in area, but its a;ready high population has increased at the third fastest rate in the country. Finally, we chose Tempe because has also been increasing a rapid rate, but it has a lower population than Seattle. The differing populations in these cities and their trends are what inclined us to chose them, as well as having the most available and most well documented data that we could find. Our research question in these three cities was: how accurate of a predictor is population density with respect to fatal car accidents?

Our hypothesis is that fatalities increase as population density increases. We believe this is the case, because the growth of a city is often correlated with increased economic activity, bringing more upper and middle class people into the city. This results in more cars in the city, and more drivers on the road. This can also create more foot traffic in areas of highly populated cities and result in more accidents, and therefore more fatal accidents. This is our hypothesis, we believe it is important to research this because it could help provide legislators and government officials with the information necessary to prevent excessive building permits, regulate road construction, and create overall better systems for traffic and prevent fatal accidents if possible. The opposite of our hypothesis would be that population density is not an accurate predictor of fatal car accidents.

2. Data Descriptions

We found our data on data.gov, a site that compiles statistical data that the government produces. There were tons of different datasets to chose from, whether national, state, city, town, county, etc. However, after looking through all the options, the Baton Rouge, Seattle, and Tempe datasets, seemed more complete, most detailed, and most up to date. Once we started looking at the data, we decided to narrow down our focus to more specific variables such as year, fatality, injury, and factors/causes of accidents. Though some of the variables had different names across datasets, they are generally represent the same things.

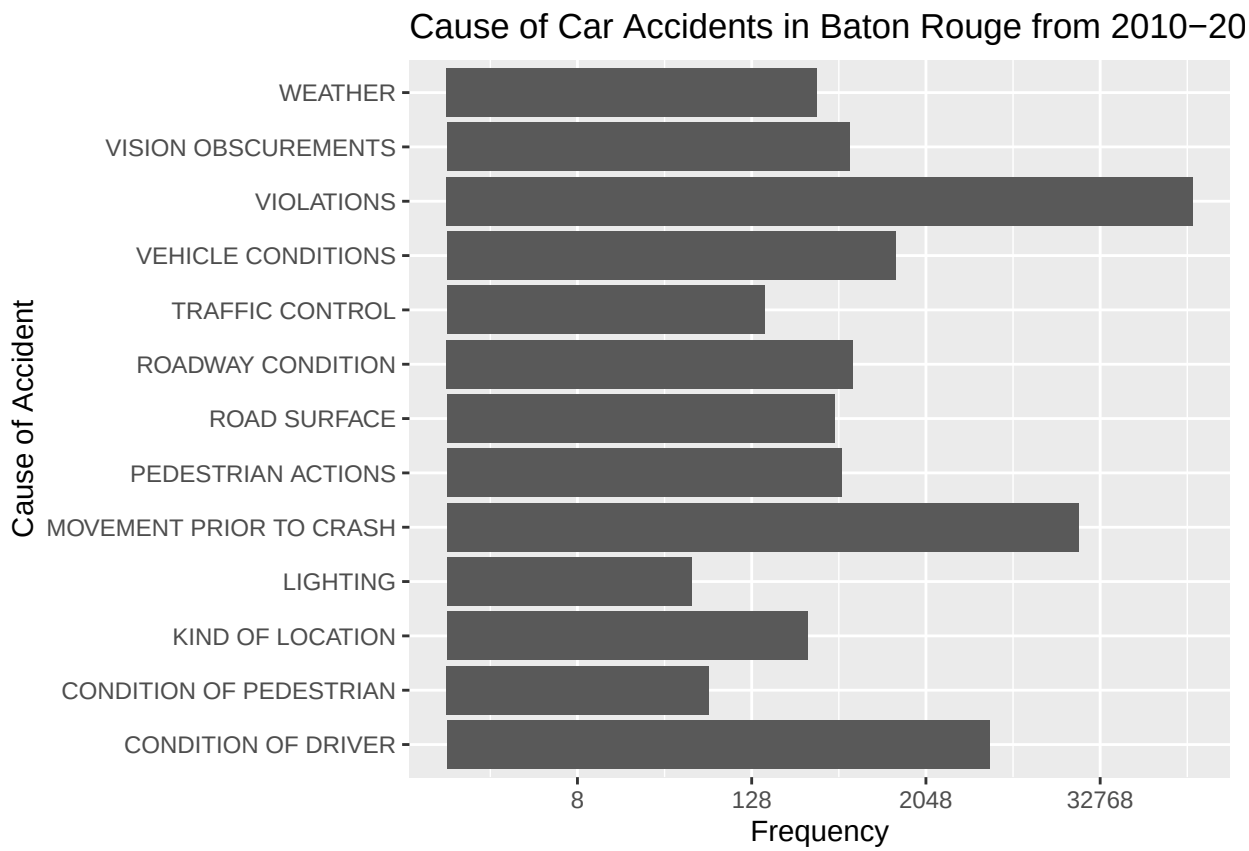
Link to database: `https://catalog.data.gov/dataset?tags=crash&_res_format_limit=0`

Depending on the city, there was slightly different information available for the accident. For example, the Baton Rouge data has a very clear variable "Primary.Factor" that labels the primary cause of the accident. In the Tempe data, the details label what might have caused the accident to occur. This includes alcohol influence, crashing at a junction, and the action the driver took that lead to the crash. The information and data from all cities is all similar but can be interpreted and read differently. In order for us to analyze the number of fatalities from accidents in these cities we narrowed down our variables by year. Per year, we kept number of fatalities, number of accidents without fatalities, population, and population density. With

these variables for each city we can rug a regression to find the relationship between the size of the city and how fatal the accidents are.

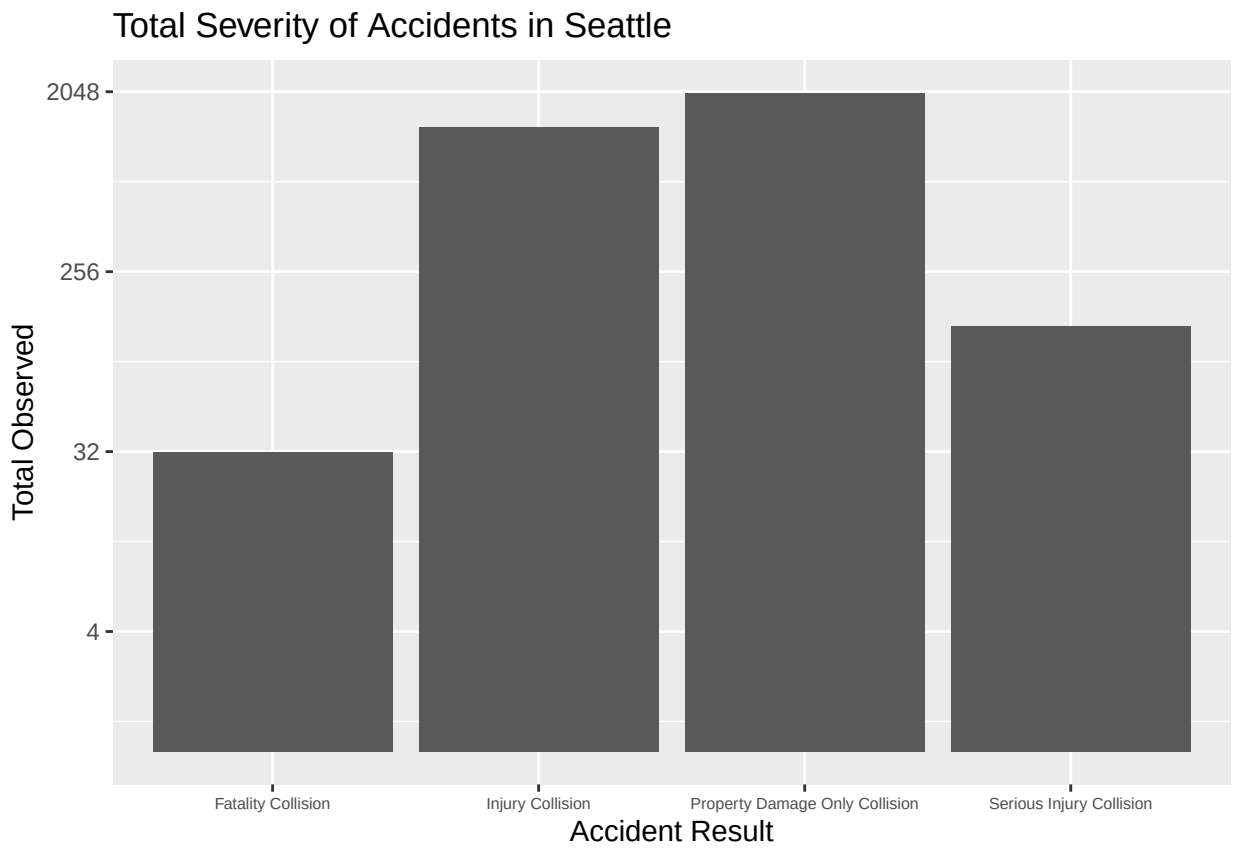
For the sake of understanding causes for most of the data, we looked into the primary factors of accidents to see where their distributions lie. The Baton Rouge barplot shows the causes of accidents in the city during the time period. This was done to ensure that the distribution of the data was fair, and not overly representing drunk driving accidents, or terrible road conditions.

```
print(br_desc_plot + ggtitle("Cause of Car Accidents in Baton Rouge from 2010-2021"))
```

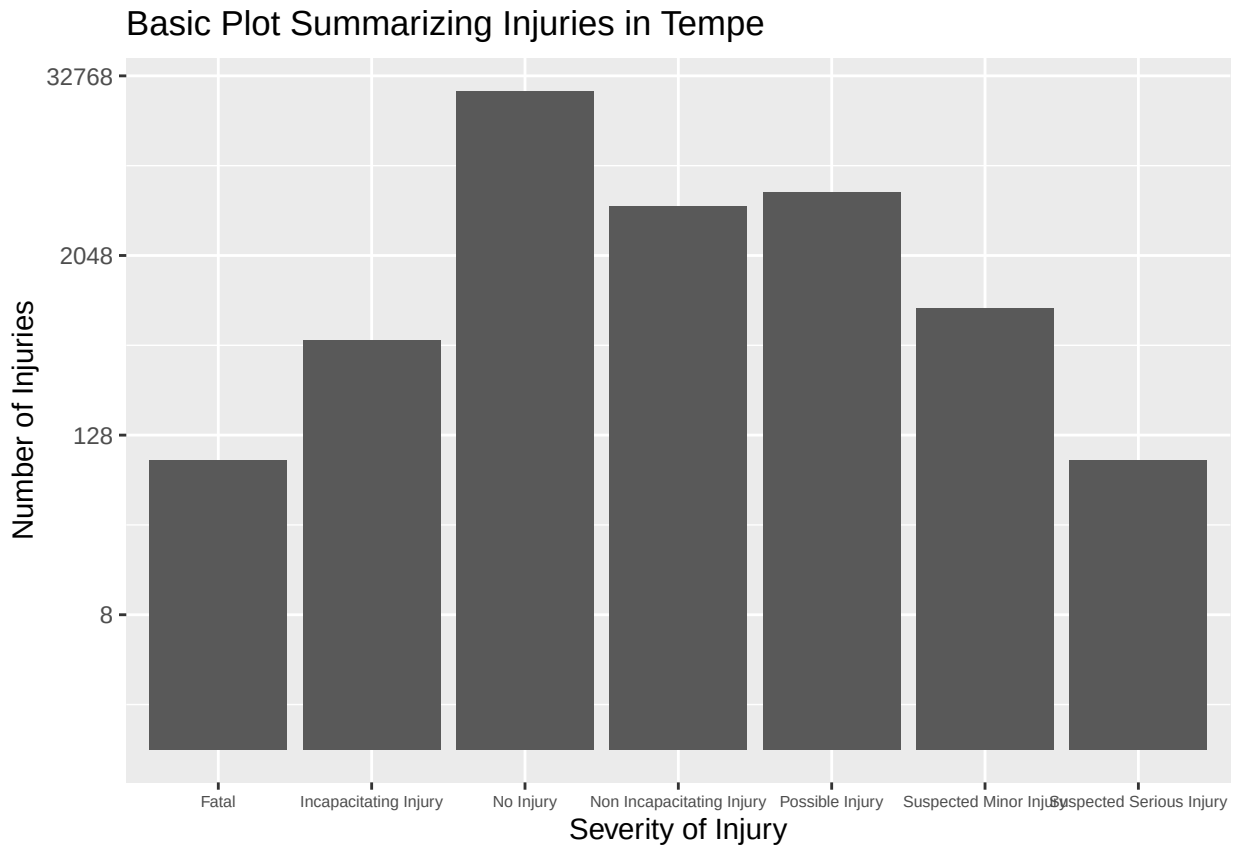


The Seattle and Tempe plots focused on the severity of accidents. The Seattle plot shows distribution of severity among the accidents, and the Tempe plot summarizes the degree of injury.

```
print(sea_severity + ggtitle("Total Severity of Accidents in Seattle"))
```




```
print(tempe_severity_plot + ggtitle("Basic Plot Summarizing Injuries in Tempe"))
```



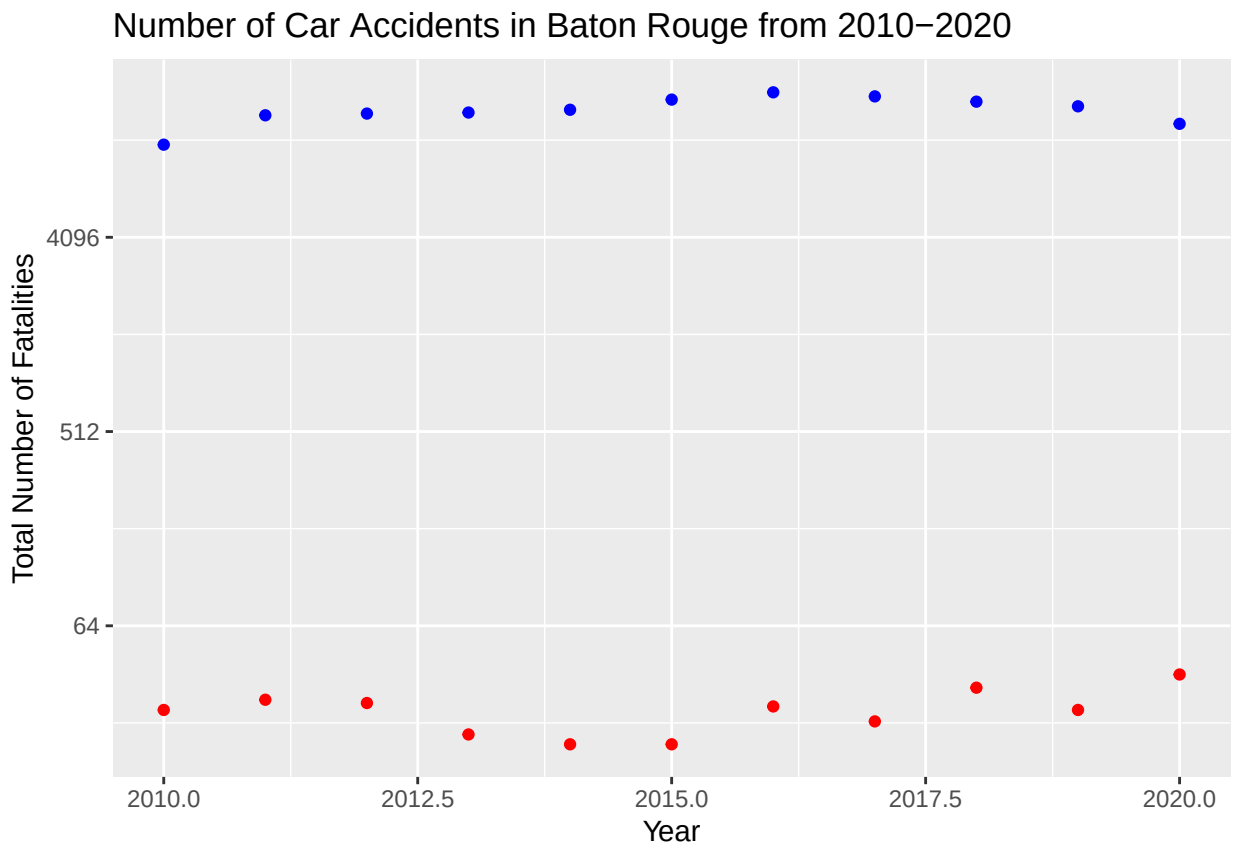
Our response variable is the amount of fatalities occurring as a function of population density. The degree of severity and causes of accidents are relevant because it could be possible that with an increase in population density, there is an increase in total accidents but not severity. It could also be the case that the causes of accidents become more and more concentrated on violations and conditions of drivers.

3. Results - Scatterplots and Regressions

Now we will show the results of our data.

In Baton Rouge, there was a stark contrast between the number of fatal car accidents each year and the total number of accidents. After accounting for population changes during each year between 2010 and 2020, we were able to see that the changes between each year showed no significant pattern. This was further confirmed after running a regression on fatality counts and population density with respect to each year. The results of the regression showed that the Baton Rouge R squared is too small to be described as accurately predicting the data. The p-value is too large to describe the correlation between car accident fatality and population changes from 2010-2019, seeing as it is greater than .05. From the size of the p-value, we fail to reject the null hypothesis with the Baton Rouge data set.

```
print(fatality_plot + ggtitle("Number of Car Accidents in Baton Rouge from 2010-2020"))
```



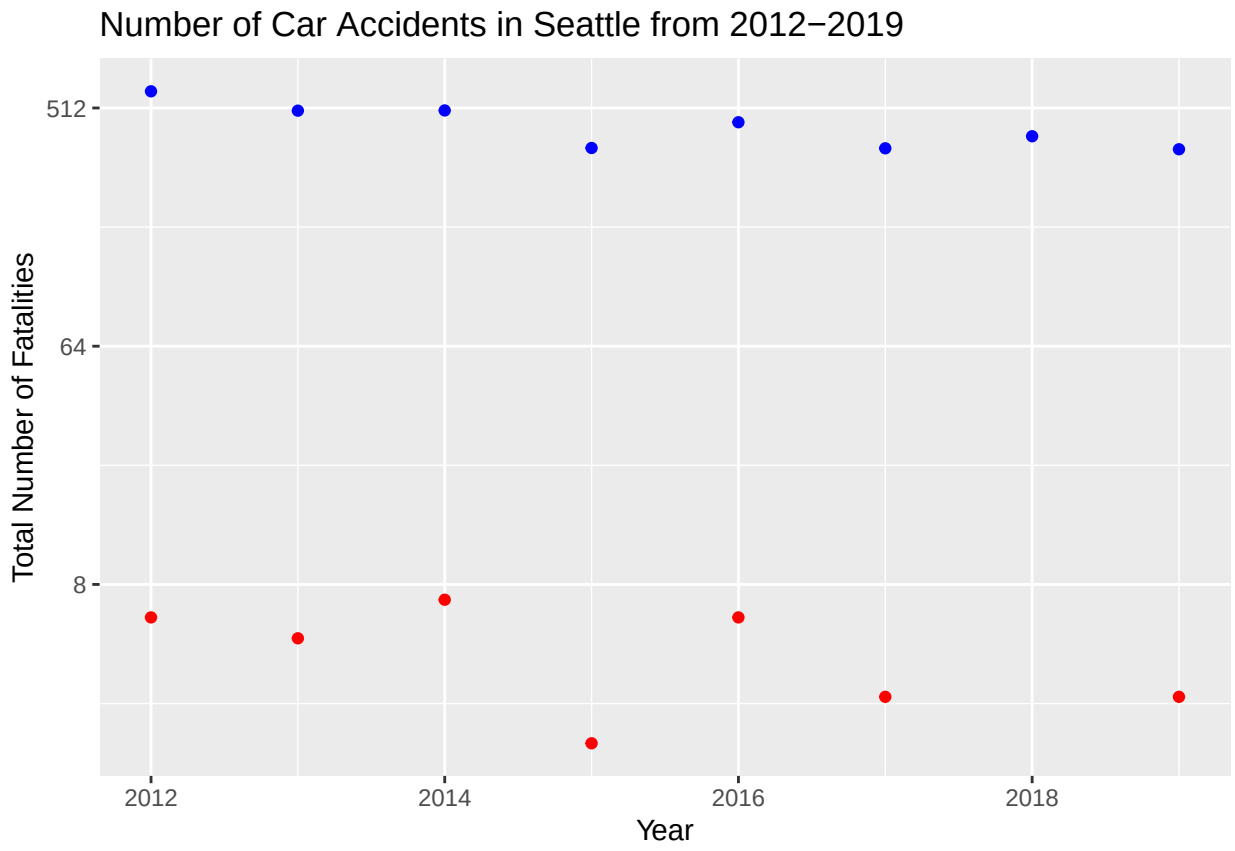

```
baton_rouge_regression
```

```
##
## Call:
## lm(formula = fatality_counts ~ pop_density + year, data = baton_rouge_totals)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.444 -3.699 -1.141  3.516  5.911
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4595.1332  2907.3534   1.581    0.158
## pop_density   -0.2030     0.1135  -1.789    0.117
## year          -2.0114     1.3097  -1.536    0.168
##
## Residual standard error: 4.643 on 7 degrees of freedom
## Multiple R-squared:  0.3191, Adjusted R-squared:  0.1246
## F-statistic:  1.64 on 2 and 7 DF,  p-value: 0.2605
```

In Seattle, there was also a contrast between the total number of fatal and non fatal accidents each year. Again, when accounting for population changes, they made no significant impact on the frequency of fatal car accidents. The only two trends that one can see in the geom point plot is that the number of fatalities and number of non fatalities in car accidents stay relatively consistent year over year. The bar plot confirms that the lowest frequency of accident was the most severe while the least severe accident is most common.

Furthermore, a regression on fatality counts and population density shows that r-squared is too small to accurately describe the data. The p-value is also too large to determine a correlation between car accident fatality and population changes in Seattle, as it is greater than 0.05. Therefore, we fail to reject the null hypothesis with the Seattle data set.

```
print(fatality_plot_sea + ggtitle("Number of Car Accidents in Seattle from 2012-2019"))
```

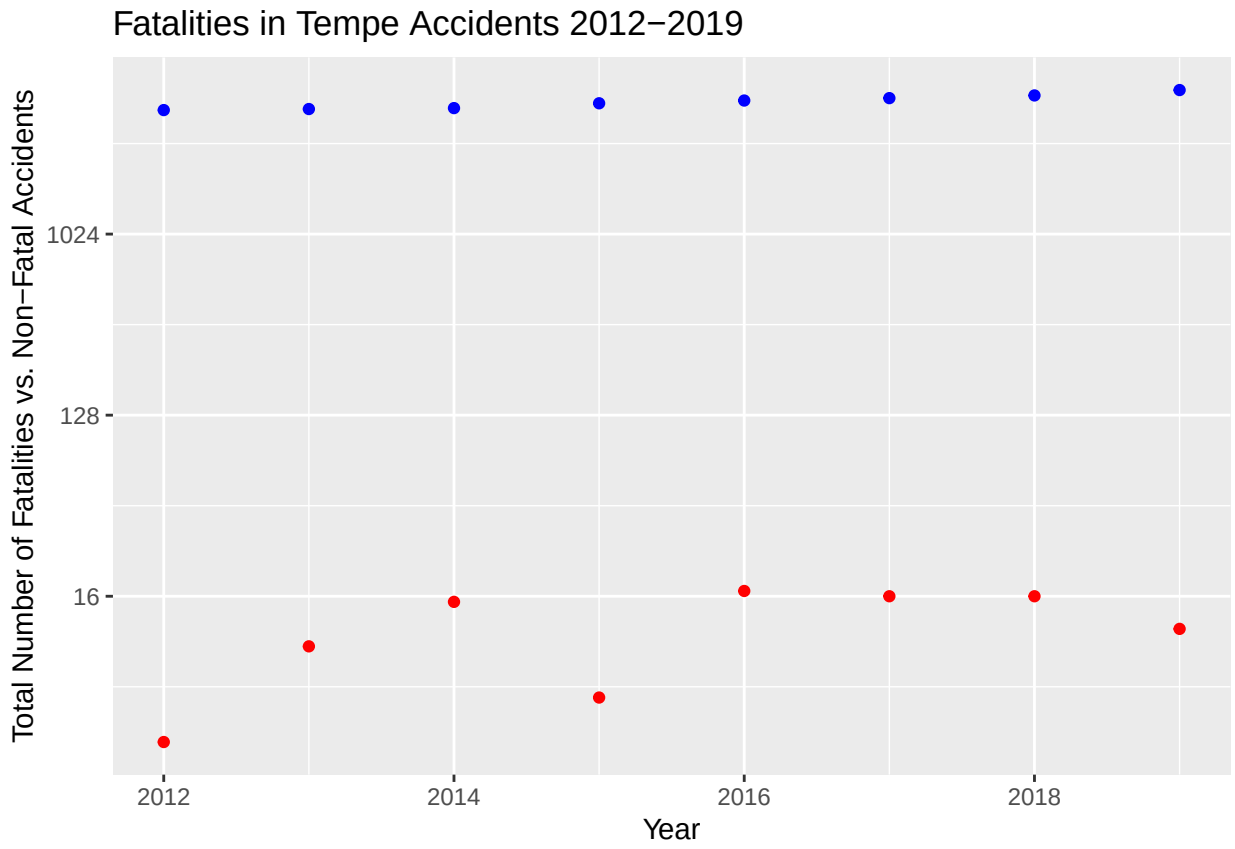



```
seattle_regression
```

```
##
## Call:
## lm(formula = fatality_counts ~ pop_density + Year, data = seattle_totals)
##
## Residuals:
##        1        2        3        4        5        6        7
## -0.02573 -0.54749  1.89974 -2.62203  1.88720 -0.60356  0.01187
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.367e+03  1.994e+04  -0.069    0.949
## pop_density -2.597e-03  2.279e-02  -0.114    0.915
## Year         7.309e-01  1.034e+01   0.071    0.947
##
## Residual standard error: 1.918 on 4 degrees of freedom
## Multiple R-squared:  0.3225, Adjusted R-squared:  -0.01618
## F-statistic: 0.9522 on 2 and 4 DF,  p-value: 0.4589
```

Similarly in Tempe, there was a much higher number of non-fatal accidents compared with he number of fatal accidents. The change in population over the years did not have a significant impact on the number of fatalities but did seem to increase the number of accidents. The population in Tempe is known to have been growing over the recent years but again this impact was not significant based on the regression. The regression shows the r-squared value represents a somewhat high amount of variability and does not accurately predict the data. The p-value is also too large to determine a correlation in fatalities and population in Tempe. Once again, we fail to reject the null hypothesis with the Tempe data.

```
print(tempe_fatality_plot + ggtitle("Fatalities in Tempe Accidents 2012-2019"))
```




```
tempe_regression
```

```
##
## Call:
## lm(formula = Fatalities ~ pop_density + Year, data = tempe_totals)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -1.4657  0.5089  4.0211 -7.8932  4.1272  1.9581  2.3757 -3.6321
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.084e+03  8.059e+03  -1.127    0.311
## pop_density -3.794e-02  4.582e-02  -0.828    0.445
## Year         4.594e+00  4.095e+00   1.122    0.313
```

```
##
## Residual standard error: 4.911 on 5 degrees of freedom
## Multiple R-squared:  0.4089, Adjusted R-squared:  0.1725
## F-statistic: 1.729 on 2 and 5 DF,  p-value: 0.2686
```

4. Conclusion

In conclusion, our hypothesis was that as population density increased there would be a rise in fatal car accidents. The null hypothesis was that as population density increased, there would be little to no change in fatal car accidents. Based on the R-squared and p-values recorded in the Baton Rouge, Seattle and Tempe regressions, we fail to reject the null hypothesis. All of our p-values were $>.05$ which indicated that we could not reject the null hypothesis. We do not find a lot of support for our hypothesis based on our data and analysis, so more research is necessary. A major limitation of our analysis is that we only looked at 10 years. By looking at a trend over 50 years we could learn more about the causes of fatal accidents, their frequency, and the trend in them with respect to growing and declining populations in cities. We were also limited in the quality of data that we used. Since a lot of this data comes from police reports, there is the possibility of human error and a misrepresentation of some aspects of data. We can also not say with certainty that all accidents were reported, and we cannot measure how many were unreported.