# Model Building Part 2

Charlie Bertelsen

2023-12-09

## Project Questions and Methods:

### *1. Is there a correlation between the number of Covid cases/deaths and the number of traffic collisions in Los Angeles California?*

### *2. What is this relationship and what can we learn from it?*

To answer these questions the three main methods that I will be using are polynomial regression, negative binomial regression, and principal component regression. I chose to use both linear regression and polynomial regression because I want to compare the two regression methods and see which one is a better fit for my data. Based on the scatter plots in my data exploration the data does not appear to follow a very linear pattern so I am expecting polynomial models to fit much better. In the last model building assignment I used a Poisson model but this model did not do a great job at fitting the data. I chose to switch from a poisson model to the negative binomial model because I found that there is over dispersion in my data. This is evident from the extremely low p-value (2.2e-16) and the high dispersion value (3.03) from the dispersion test below. My dependent variable (traffic collisions) is a count variable with overdispersion so I am expecting negative binomial regression to perform and fit the data better but this may not be the case. I will be using PCA as well because I found some potential issues with multicollinearity in my data using vif analysis. This brief analysis is below and I found that two of my independent variables (cases and deaths) have vif scores above 5 which means that the coefficient estimates may be unstable and the results of standard regression may be harder to interpret.

## Method 1: Linear and Polynomial Regression for Individual Variables

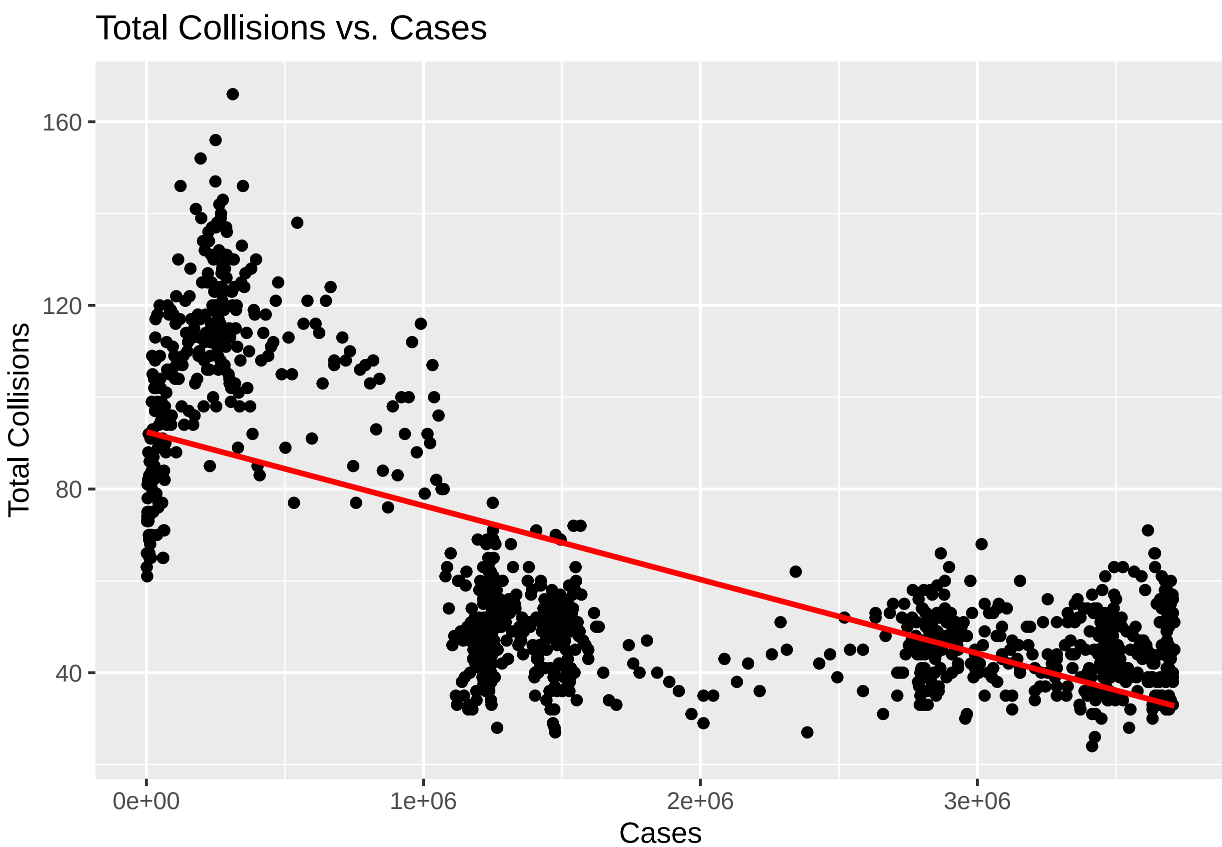
### Linear Regression:

```
##
## Call:
## lm(formula = Total_Collisions ~ cases, data = CT2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.089 -17.270   0.754  12.896  78.562
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.245e+01  1.087e+00   85.07   <2e-16 ***
## cases       -1.608e-05  4.842e-07  -33.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.77 on 1110 degrees of freedom
## Multiple R-squared:  0.4985, Adjusted R-squared:  0.498
```

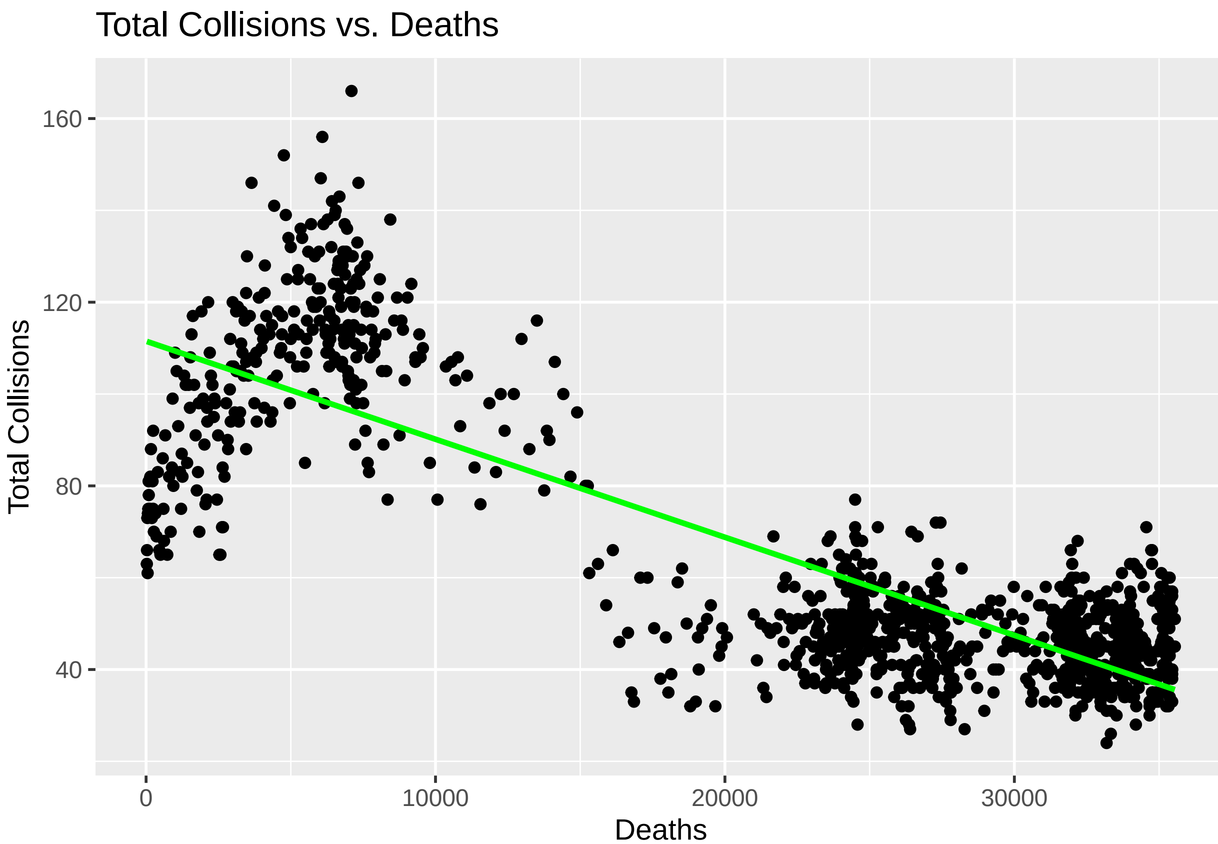
```
## F-statistic:  1103 on 1 and 1110 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Total_Collisions ~ deaths, data = CT2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.393  -9.596   0.295   9.534  69.644
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.115e+02  1.064e+00  104.77   <2e-16 ***
## deaths      -2.135e-03  4.171e-05  -51.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16 on 1110 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.7022
## F-statistic:  2621 on 1 and 1110 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



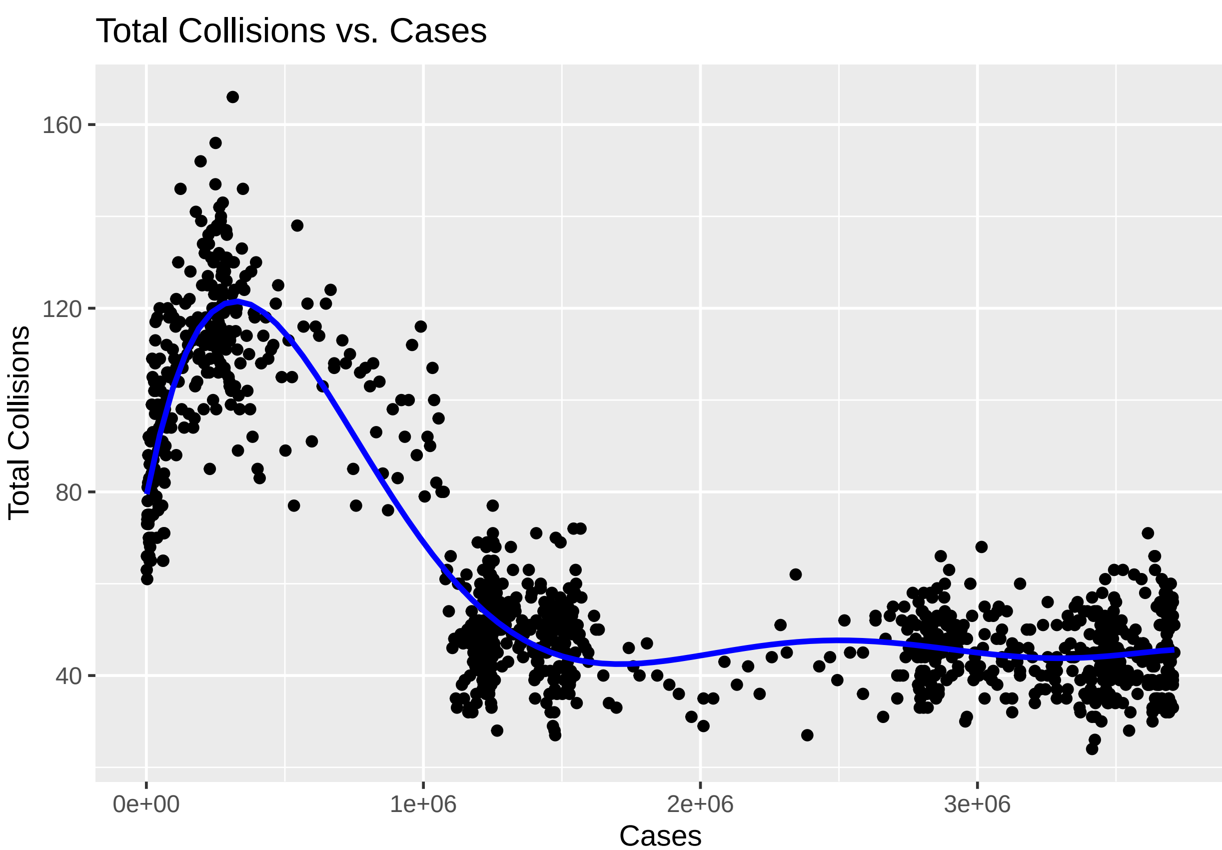

After fitting linear regression models to the data, the low p-values and considerable multiple R-squared values suggest a relationship between traffic collisions and the number of covid cases/deaths. This helps me to answer my question 1 by showing me that there is a correlation between these variables. While there may be a relationship between these variables a linear regression model does not fit the data very well based on the graphs which means other regression methods might be more appropriate.

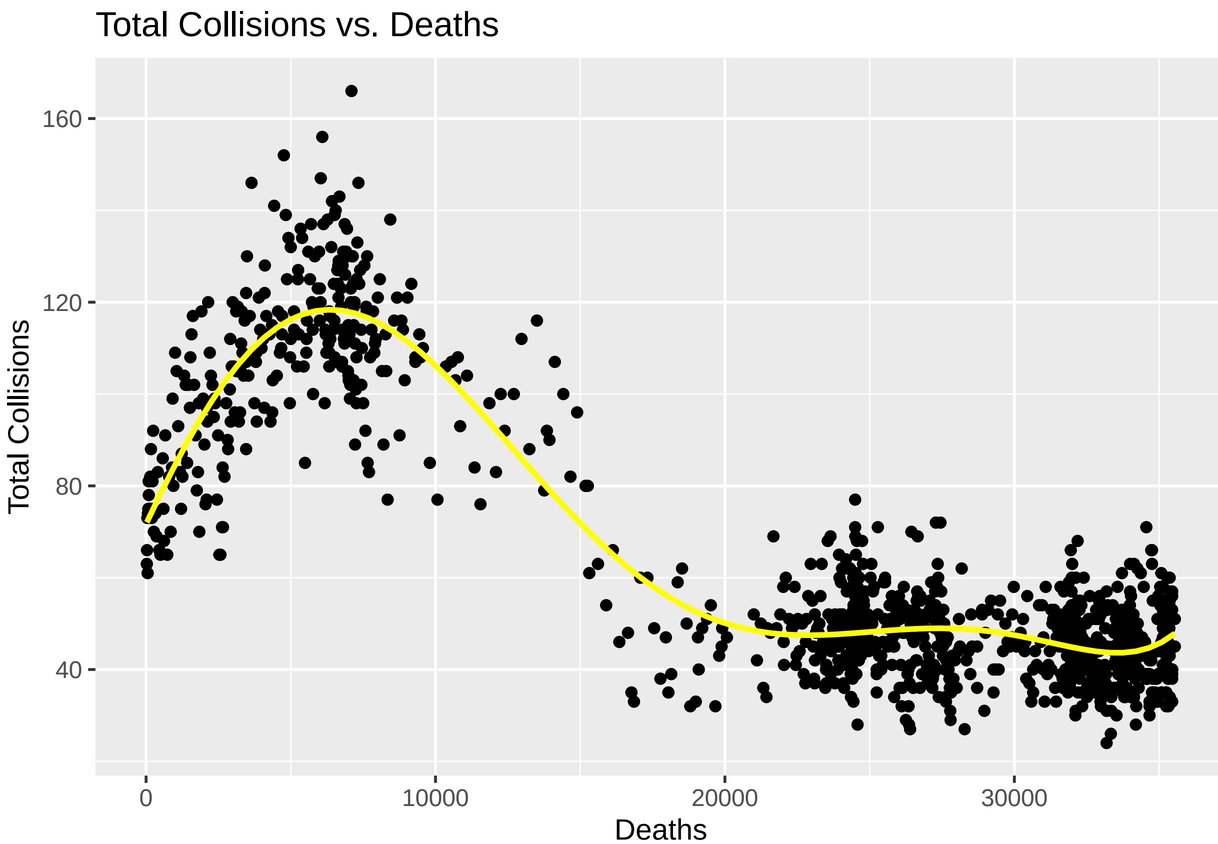
**Polynomial Regression:**

```
##
## Call:
## lm(formula = Total_Collisions ~ poly(cases, degree = 6), data = CT2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.709  -6.772  -0.437   6.366  46.223
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                62.8705     0.3242 193.917  < 2e-16 ***
## poly(cases, degree = 6)1 -689.8474    10.8115 -63.807  < 2e-16 ***
## poly(cases, degree = 6)2  439.6390    10.8115  40.664  < 2e-16 ***
## poly(cases, degree = 6)3  -27.6478    10.8115  -2.557   0.0107 *
## poly(cases, degree = 6)4 -307.9324    10.8115 -28.482  < 2e-16 ***
## poly(cases, degree = 6)5  233.2162    10.8115  21.571  < 2e-16 ***
## poly(cases, degree = 6)6  -80.0870    10.8115  -7.408 2.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 10.81 on 1105 degrees of freedom
## Multiple R-squared:  0.8647, Adjusted R-squared:  0.864
## F-statistic:  1177 on 6 and 1105 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Total_Collisions ~ poly(deaths, degree = 6), data = CT2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.390  -6.513  -0.068   6.004  48.220
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 62.8705     0.3086 203.697  < 2e-16 ***
## poly(deaths, degree = 6)1 -818.9360    10.2923 -79.568  < 2e-16 ***
## poly(deaths, degree = 6)2  102.8119    10.2923   9.989  < 2e-16 ***
## poly(deaths, degree = 6)3  326.3721    10.2923  31.710  < 2e-16 ***
## poly(deaths, degree = 6)4 -205.4198    10.2923 -19.959  < 2e-16 ***
## poly(deaths, degree = 6)5   56.5018    10.2923   5.490 4.99e-08 ***
## poly(deaths, degree = 6)6   67.3299    10.2923   6.542 9.28e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 1105 degrees of freedom
## Multiple R-squared:  0.8774, Adjusted R-squared:  0.8767
## F-statistic:  1318 on 6 and 1105 DF,  p-value: < 2.2e-16
```





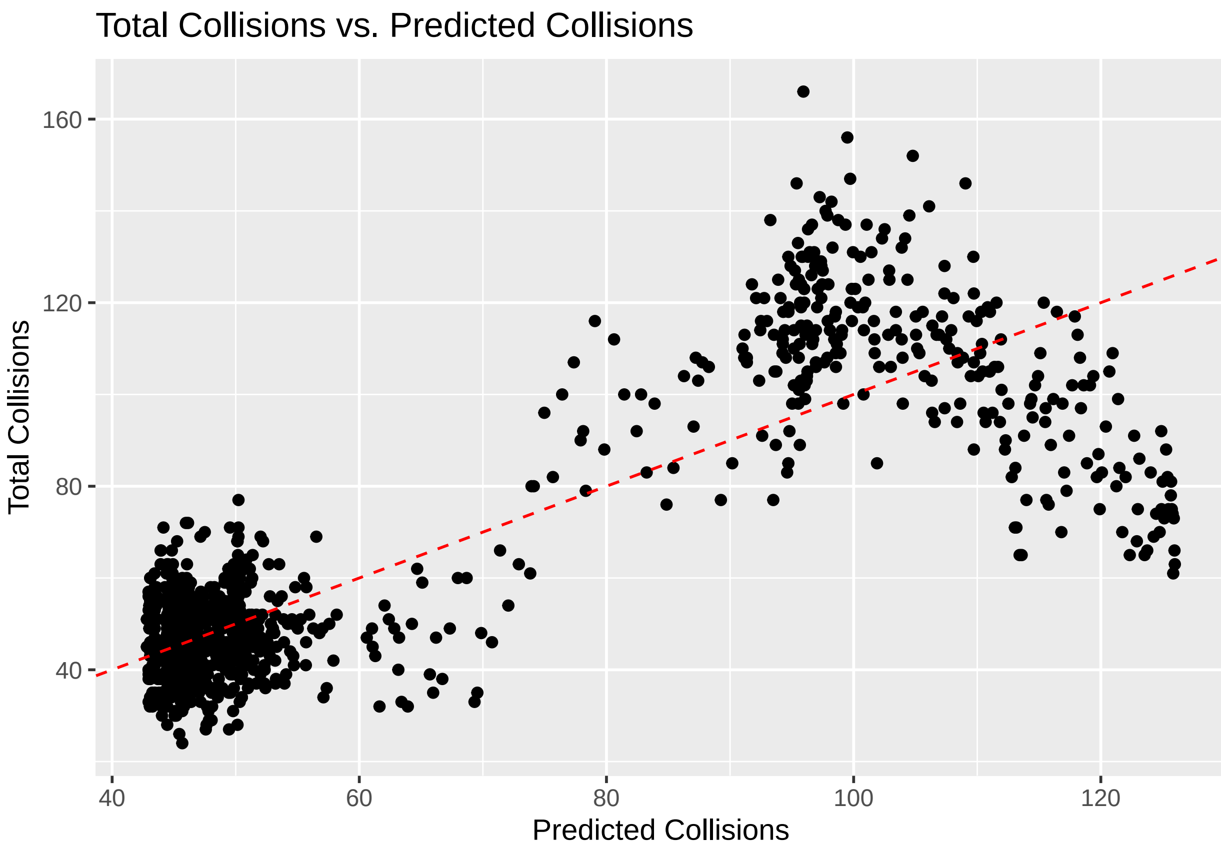

Fitting polynomial models produced significantly better results than the linear models. Not only are the p-values very small but the Multiple R-squared values for the polynomial models (cases: 0.8647, deaths: 0.8774) are much higher than the values for the linear models (cases: 0.499, deaths: 0.703). This tells me that the predictor variables explain much more of the variance in the response variable when fitting the polynomial models to the data. Overall these models help me to see that there is a very clear correlation between traffic collisions and cases/deaths. When traffic collisions were high cases/deaths were low vice versa.

### Method 2: Negative Binomial Regression

```
##
##  Overdispersion test
##
## data:  poisson_model
## z = 13.598, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   3.030811
```

```
##
## Call:
## glm.nb(formula = Total_Collisions ~ cases + deaths, data = CT2023,
##     init.theta = 32.68165911, link = log)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.837e+00  1.485e-02 325.668   <2e-16 ***
## cases        1.347e-07  1.381e-08   9.749   <2e-16 ***
```

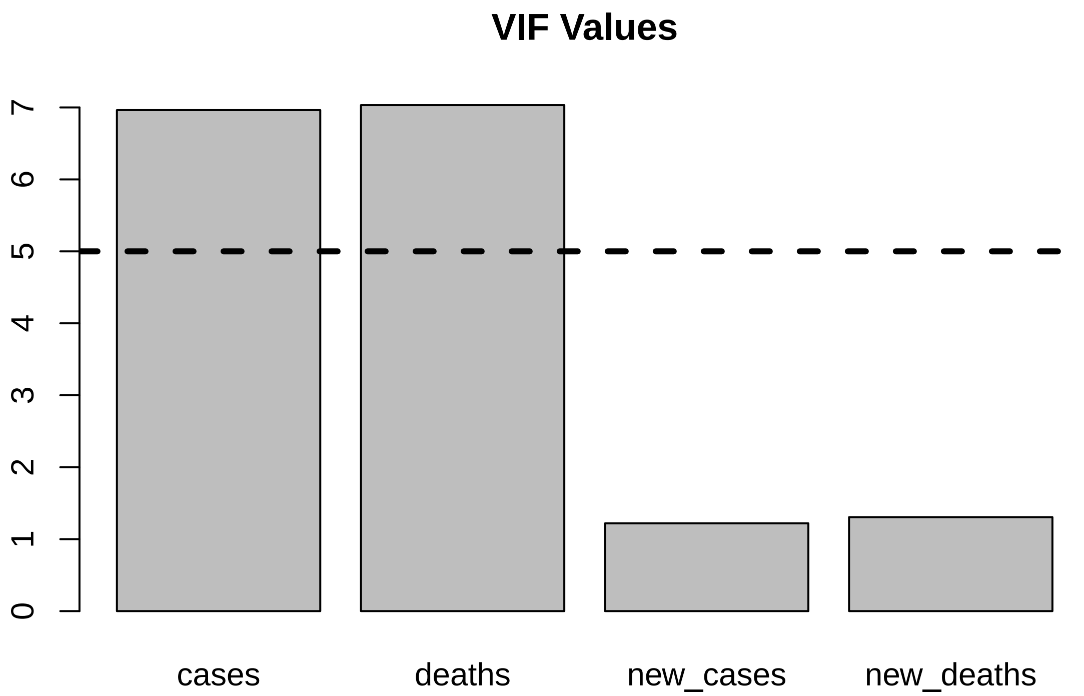
```
## deaths      -4.445e-05  1.499e-06 -29.660   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(32.6817) family taken to be 1)
##
##     Null deviance: 4405.9  on 1111  degrees of freedom
## Residual deviance: 1057.9  on 1109  degrees of freedom
## AIC: 8768
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  32.68
##           Std. Err.:  2.05
##
##  2 x log-likelihood:  -8760.011
```




The negative binomial regression model summary supports my conclusions from the previous linear and binomial models that there is a strong correlation between the number of traffic collisions and the number of cases/deaths. The coefficients of the model suggest a small positive relationship between the number of cases and collisions and a stronger negative relationship between the number of deaths and collisions. While this model identified this relationships as significant, the graph shows that the regression line does not fit the model very well much like the linear regression models. This model was much more succesful than the poisson model from my last analysis.

### Method 3: Principal Component Analysis and Regression

```
##      cases     deaths  new_cases new_deaths
##   6.963344   7.032442   1.219450   1.305612
```



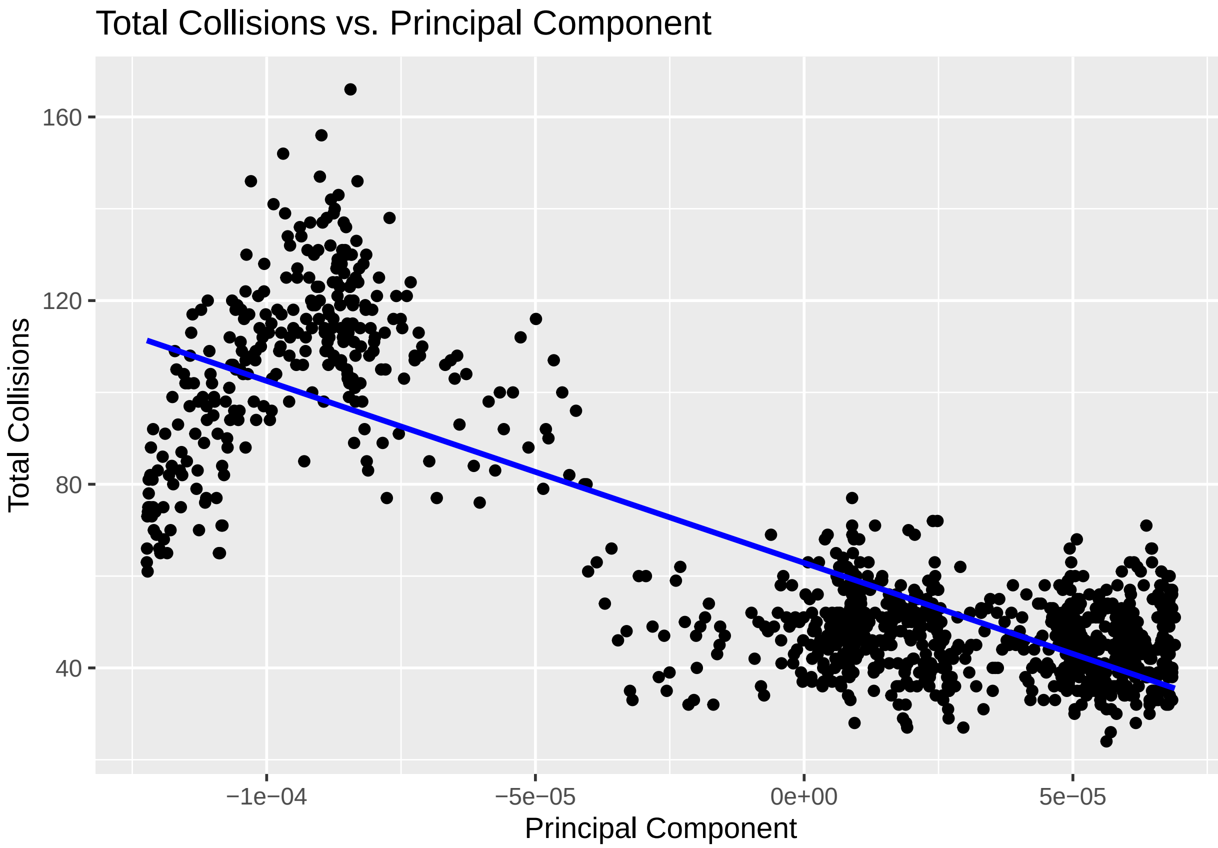

A quick VIF analysis of the four predictor variables shows a potential issue with high multicollinearity between cases and deaths. For this reason, PCA will be a good choice to eliminate multicollinearity and reduce the dimensions of the data.

```
## Importance of components:
##                           PC1     PC2
## Standard deviation     1.3875 0.27334
## Proportion of Variance 0.9626 0.03736
## Cumulative Proportion  0.9626 1.00000
```

```
##
## Call:
## lm(formula = Total_Collisions ~ principal_components, data = CT2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.270  -9.586   0.353   9.577  69.685
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.287e+01  4.806e-01  130.82   <2e-16 ***
## principal_components -3.963e+05  7.760e+03  -51.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.03 on 1110 degrees of freedom
## Multiple R-squared:  0.7014, Adjusted R-squared:  0.7012
```

```
## F-statistic:  2608 on 1 and 1110 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




After completing principal component analysis the high standard deviation, proportion of variance and cumulative proportion of principal component 1 means it captures more than enough of the variability to be the only component used in the regression model. The regression summary using principal component 1 as the predictor variable shows a high statistic significance between traffic collisions and covid cases/deaths. This is supported by the low p-value (<0.05) and the high Multiple R-squared. This relationship is strong and negative as the number of traffic collisions was higher when there were lower cases/deaths and vice versa. Overall this model provides further evidence to support a strong correlation between the number of traffic collisions and the number of cases/deaths.

## Conclusion:

The main changes that I made from the model building part 1 assignment were adding polynomial regression, changing poission regression to negative binomial regression, and completing regression analysis on the principal components. These changes led me to new conclusions about my data that I otherwise may not have discovered.

After implementing 4 different regression models the results were overwhelming that there is a strong correlation between the number of traffic collisions and the number of covid cases/deaths. All of the models proved that this relationship was statistically significant and negative, meaning that as covid cases/deaths increased the number of collisions decreased.

At first glance, the scatter plots of collisions vs cases/deaths are not linear. This is further supported by the results of fitting the polynomial regression models which explained more of the variance and fit better. The polynomial model was by far the best fitting model as can be seen on the collisions/cases-deaths plot that includes the regression line.

While there is a strong correlation between collisions and covid cases/deaths this does not mean that the increasing number of covid total cases and deaths directly caused the decrease in traffic collisions. During the period 2020-2023 the number of daily traffic collisions decreased significantly in Los Angeles county. It is important to note that the numbers of cases/deaths are the accumulated totals for LA recorded with their corresponding dates. Overall, my analysis shows that at some point during the pandemic there was a combination of different factors that ultimately led to a large decrease in the daily number of traffic collisions in LA that has stayed consistent even after life has mostly returned to normal. In further analysis, if a complete record of daily new cases was found it would interesting to compare how daily collisions coincided with daily reported new cases to see if there was any relationship.