St. Paul Bike Accident Analysis

Kenny Waite

2/27/2020

# Introduction

As a society, Americans are trying to develop a more active lifestyle for both the health benefits as well for the enjoyment that comes with getting outside. One of the most popular recreational activities in America is cycling. Cities are becoming more aware of the desire for bike trails and greenways, however, many major cities still lack an extensive trail system for bicyclists. If you would like to cover long distance on your cycling outing, you most likely will have to at some point share the road with vehicles. For a majority of reasons, this can be problematic, both for the bicyclists and the driver of the vehicle. Until sufficient bike lanes are created and greenway trail systems are matured, the road will have to be shared.

# Problem Statement

For my project, I will be attempting to identify the high volume locations of bike accidents in St. Paul Minnesota and identify what factors lead to an area being high risk for bike accidents in St. Paul. This analysis is important for multiple reasons. For starters, bikers can use this data to avoid specific areas when planning their bike routes in order to minimize their chances of getting in an accident with a vehicle. Also, when planning future streets/bike lanes/greenways, cities will be able to use this analysis to assist them in designing roads and locations that reduce bike accidents. ## The Process I used three models in an attempt to find the most accurate way to predict bike accidents in each of the three districts. Regression, Multiple Regression and Logistical Regression. The below outlines the findings from running the data through these three models. ## The Data I will be using data collected by the [St. Paul Police Department](https://www.opendatanetwork.com/dataset/information.stpaul.gov/9qtp-ayhq) to determined what characteristics are related to areas of high bike and pedestrian crashes. As you can see from the below code, there was a lot of data cleanup and manipulation I had to perform. As well as creating new fields, I had to convert data types, and remove any incorrect data. I added comments to the code below so you can see specifically what I did to clean up the data.

bikeDataRaw <- fromJSON("https://information.stpaul.gov/resource/bw92-5h94.json?crash\_type=Bike")  
bikeData <- bikeDataRaw[, c(1, 3, 5, 9, 10, 17, 22, 23, 24, 25)]  
# Convert the date\_time to a datetime format from a character  
bikeData$date\_time <- as\_datetime(bikeData$date\_time)  
# Split the date and time values up from the date\_time value  
bikeData$date <- as.Date(bikeData$date\_time)  
time <- strftime(bikeData$date\_time, format = "%H:%M:%S")  
bikeData$time <- chron(times = time)  
bikeData$month <- month(bikeData$date)  
# Create district accident and citation to counts  
bikeData$central\_Accidents <- ifelse(bikeData$district == "Central District", 1,   
 0)  
bikeData$eastern\_Accidents <- ifelse(bikeData$district == "Eastern District", 1,   
 0)  
bikeData$western\_Accidents <- ifelse(bikeData$district == "Western District", 1,   
 0)  
bikeData$citation\_Biker <- ifelse(bikeData$citation\_to == "Biker", 1, 0)  
bikeData$citation\_Driver <- ifelse(bikeData$citation\_to == "Driver", 1, 0)  
# Convert values from character to numerics  
bikeData$driver\_age <- as.numeric(bikeData$driver\_age)  
bikeData$speed\_limit <- as.numeric(bikeData$speed\_limit)  
bikeData$lanes\_of\_traffic <- as.numeric(bikeData$lanes\_of\_traffic)  
bikeData$biker\_age <- as.numeric(bikeData$biker\_age)  
# Create new fields  
bikeData$season <- ifelse(bikeData$month %in% 10:12, "Fall", ifelse(bikeData$month %in%   
 1:3, "Winter", ifelse(bikeData$month %in% 4:6, "Spring", ifelse(bikeData$month %in%   
 4:6, "Summer", "Unknown"))))  
bikeData$driver\_skill\_level <- ifelse(bikeData$driver\_age <= 24, "New Driver", ifelse(bikeData$driver\_age %in%   
 25:64, "Experienced Driver", ifelse(bikeData$driver\_age >= 65, "Senior Driver",   
 "Unknown")))  
# Remove incorrect data  
bikeData <- subset(bikeData, bikeData$district != "Transit" & bikeData$district !=   
 "State Patrol" & bikeData$district != "Metro Transit PD" & bikeData$injury\_to\_biker !=   
 "Unknown" & bikeData$driver\_skill\_level != "Unknown" & bikeData$season != "Unknown")  
summary(bikeData)  
 crash\_type date\_time district   
 Length:136 Min. :2016-01-25 09:43:00 Length:136   
 Class :character 1st Qu.:2016-10-21 02:09:00 Class :character   
 Mode :character Median :2017-06-12 10:28:00 Mode :character   
 Mean :2017-09-14 20:22:55   
 3rd Qu.:2018-06-14 10:12:15   
 Max. :2019-06-25 07:03:00   
   
 citation\_to biker\_age injury\_to\_biker lanes\_of\_traffic  
 Length:136 Min. : 6.00 Length:136 Min. :0.000   
 Class :character 1st Qu.:15.00 Class :character 1st Qu.:2.000   
 Mode :character Median :27.00 Mode :character Median :2.000   
 Mean :30.95 Mean :2.838   
 3rd Qu.:44.00 3rd Qu.:4.000   
 Max. :79.00 Max. :7.000   
 NA's :4 NA's :31   
 signal\_present speed\_limit driver\_age date   
 Length:136 Min. :10.00 Min. :17.00 Min. :2016-01-25   
 Class :character 1st Qu.:30.00 1st Qu.:28.00 1st Qu.:2016-10-20   
 Mode :character Median :30.00 Median :41.50 Median :2017-06-12   
 Mean :29.63 Mean :42.54 Mean :2017-09-14   
 3rd Qu.:30.00 3rd Qu.:54.25 3rd Qu.:2018-06-14   
 Max. :40.00 Max. :92.00 Max. :2019-06-25   
 NA's :28   
 time month central\_Accidents eastern\_Accidents  
 Min. :02:53:00 Min. : 1.000 Min. :0.0000 Min. :0.0000   
 1st Qu.:09:20:00 1st Qu.: 5.000 1st Qu.:0.0000 1st Qu.:0.0000   
 Median :12:01:00 Median : 6.000 Median :0.0000 Median :0.0000   
 Mean :11:35:17 Mean : 6.228 Mean :0.3676 Mean :0.2426   
 3rd Qu.:13:56:00 3rd Qu.: 6.000 3rd Qu.:1.0000 3rd Qu.:0.0000   
 Max. :23:00:00 Max. :12.000 Max. :1.0000 Max. :1.0000   
   
 western\_Accidents citation\_Biker citation\_Driver season   
 Min. :0.0000 Min. :0.00000 Min. :0.0000 Length:136   
 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000 Class :character   
 Median :0.0000 Median :0.00000 Median :0.0000 Mode :character   
 Mean :0.3897 Mean :0.07407 Mean :0.2889   
 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:1.0000   
 Max. :1.0000 Max. :1.00000 Max. :1.0000   
 NA's :1 NA's :1   
 driver\_skill\_level  
 Length:136   
 Class :character   
 Mode :character

# The Analysis

Now, what you have all been waiting for, the ***ANALYSIS***. This is the part of the report where we dig into the data and see what kind of insights we can identify.

## Linear Regression

The first approach I took to analyzing the data was using linear regression models to identify if there were any strong relationships between the data points and the amount of bike accidents in each district. Linear regression cannot be performed with categorical data points, so I had to create three new data sets for each district excluding the categorical values.

bikeDataCentral <- bikeData[, c(5, 7, 9, 10, 14)]  
bikeDataEastern <- bikeData[, c(5, 7, 9, 10, 15)]  
bikeDataWestern <- bikeData[, c(5, 7, 9, 10, 16)]

Once the new data sets were created, I then used the contrivance matrix below to see what type of relationship each variable had to the amount of accidents in each district. ***Central District***

cov(bikeDataCentral, use = "complete.obs")  
 biker\_age lanes\_of\_traffic speed\_limit driver\_age  
biker\_age 335.9900019 0.6812269 3.0013590 20.7082120  
lanes\_of\_traffic 0.6812269 1.4661231 0.7202485 -0.8258591  
speed\_limit 3.0013590 0.7202485 8.3770142 2.8276063  
driver\_age 20.7082120 -0.8258591 2.8276063 280.5486313  
central\_Accidents -1.2581052 0.1349253 0.1679286 -0.7256843  
 central\_Accidents  
biker\_age -1.2581052  
lanes\_of\_traffic 0.1349253  
speed\_limit 0.1679286  
driver\_age -0.7256843  
central\_Accidents 0.2427684

***Eastern District***

cov(bikeDataEastern, use = "complete.obs")  
 biker\_age lanes\_of\_traffic speed\_limit driver\_age  
biker\_age 335.9900019 0.68122695 3.0013589594 20.7082120  
lanes\_of\_traffic 0.6812269 1.46612308 0.7202484954 -0.8258591  
speed\_limit 3.0013590 0.72024850 8.3770141720 2.8276063  
driver\_age 20.7082120 -0.82585906 2.8276062900 280.5486313  
eastern\_Accidents -0.8408076 -0.07357795 -0.0009706853 -0.5171811  
 eastern\_Accidents  
biker\_age -0.8408076102  
lanes\_of\_traffic -0.0735779460  
speed\_limit -0.0009706853  
driver\_age -0.5171811299  
eastern\_Accidents 0.1918074160

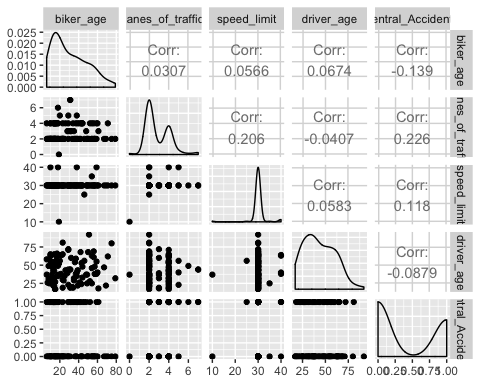
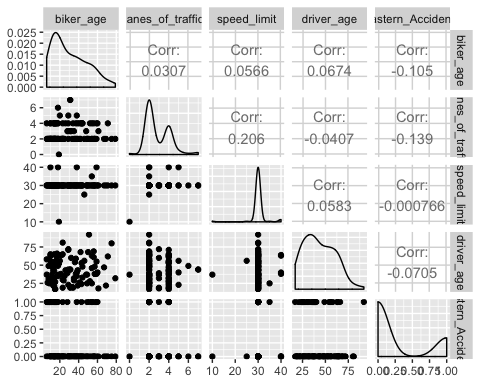
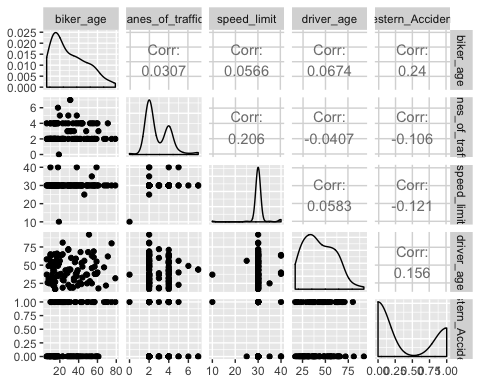
***Western District***

cov(bikeDataWestern, use = "complete.obs")  
 biker\_age lanes\_of\_traffic speed\_limit driver\_age  
biker\_age 335.9900019 0.68122695 3.0013590 20.7082120  
lanes\_of\_traffic 0.6812269 1.46612308 0.7202485 -0.8258591  
speed\_limit 3.0013590 0.72024850 8.3770142 2.8276063  
driver\_age 20.7082120 -0.82585906 2.8276063 280.5486313  
western\_Accidents 2.0989128 -0.06134731 -0.1669579 1.2428655  
 western\_Accidents  
biker\_age 2.09891283  
lanes\_of\_traffic -0.06134731  
speed\_limit -0.16695787  
driver\_age 1.24286546  
western\_Accidents 0.22762570

***Positive or Negative Relationship Between The Variables and The Accidents by District***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| District | Biker Age | Lanes of Traffic | Speed Limit | Driver Age | Month |
| Central | Negative | Positive | Positive | Negative | Positive |
| Eastern | Negative | Negative | Negative | Negative | Positive |
| Western | Positive | Negative | Negative | Positive | Negative |

The table above highlights the positive or negative relationship between the different variables and the amount of accidents per district.

Next I created three correlation matrices for each district to get a quick snapshot of which variables had the greatest impact on the relationship to bike accidents occurring. ***Central District***  ***Eastern District***  ***Western District*** 

The next table shows the correlation coefficients for between each variable and the amount of accidents per district. ***Positive or Negative Relationship Between The Variables and The Accidents by District***

Variable Central Eastern Western  
1 Biker Age -0.1847 -0.09236 0.2628  
2 Lanes of Traffic 0.2389 -0.1524 -0.1035  
3 Speed Limit 0.004035 0.07481 -0.07324  
4 Driver Age -0.08265 -0.05687 0.1317

Every variable has a weak positive or negative correlation between the variable and the number of accidents in the district.

The next table shows the coefficient of determination for between each variable and the amount of accidents per district. ***Positive or Negative Relationship Between The Variables and The Accidents by District***

Variable Central Eastern Western  
1 Biker Age 0.03411 0.00853 0.06906  
2 Lanes of Traffic 0.05707 0.02323 0.01071  
3 Speed Limit 1.628e-05 0.005597 0.005364  
4 Driver Age 0.006831 0.003234 0.01734

Looking at this table, a lot of the variables do not account for more than 1% of the variation in the amount of accidents in each district. Some of the larger coefficient of determinations are shown in the table below. However, most of the percentages are low leaving more than 90% of the variability in the amount of bike accidents to be accounted for by other variables.

District Variable Percentage\_of\_Variation  
1 Central Biker Age 3%  
2 Central Lanes of Traffic 6%  
3 Eastern Lanes of Traffic 2%  
4 Western Biker Age 7%

### Linear Regression Conclusion

Based on my analysis, you cannot say that the Biker Age, Lanes of Traffic, Speed Limit or Driver Age are a direct cause of more bike accidents. However, what you can say is that there was a weak, sometimes positive, sometimes negative relationship depending on the variable(see the able above) between Biker Age, Lanes of Traffic, Speed Limit or Driver Age and the amount of bike accidents per district.

## Regression Model

I then decided to run a few of the variables through a regression model to see what the model would say about the variables and their impact of the variation of bike accidents in St. Paul. Below is the raw results of the regression model between the selected variables and their corresponding districts. ***Regression Analysis Between Central District Accidents and Biker Age***

Call:  
lm(formula = central\_Accidents ~ biker\_age, data = bikeData,   
 na.action = "na.omit")  
  
Coefficients:  
(Intercept) biker\_age   
 0.520912 -0.004837

Call:  
lm(formula = central\_Accidents ~ biker\_age, data = bikeData,   
 na.action = "na.omit")  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.4919 -0.4097 -0.2597 0.5516 0.7742   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.520912 0.081316 6.406 2.51e-09 \*\*\*  
biker\_age -0.004837 0.002257 -2.143 0.0339 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4785 on 130 degrees of freedom  
 (4 observations deleted due to missingness)  
Multiple R-squared: 0.03413, Adjusted R-squared: 0.0267   
F-statistic: 4.594 on 1 and 130 DF, p-value: 0.03395

***Regression Analysis Between Central District Accidents and Lanes of Traffic***

Call:  
lm(formula = central\_Accidents ~ lanes\_of\_traffic, data = bikeData,   
 na.action = "na.omit")  
  
Coefficients:  
 (Intercept) lanes\_of\_traffic   
 0.11397 0.09743

Call:  
lm(formula = central\_Accidents ~ lanes\_of\_traffic, data = bikeData,   
 na.action = "na.omit")  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.6011 -0.3088 -0.3088 0.4963 0.6912   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.11397 0.12018 0.948 0.3452   
lanes\_of\_traffic 0.09743 0.03902 2.497 0.0141 \*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4783 on 103 degrees of freedom  
 (31 observations deleted due to missingness)  
Multiple R-squared: 0.05707, Adjusted R-squared: 0.04791   
F-statistic: 6.234 on 1 and 103 DF, p-value: 0.01412

***Regression Analysis Between Eastern District Accidents and Lanes of Traffic***

Call:  
lm(formula = eastern\_Accidents ~ lanes\_of\_traffic, data = bikeData,   
 na.action = "na.omit")  
  
Coefficients:  
 (Intercept) lanes\_of\_traffic   
 0.42660 -0.05635

Call:  
lm(formula = eastern\_Accidents ~ lanes\_of\_traffic, data = bikeData,   
 na.action = "na.omit")  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.4266 -0.3139 -0.2012 0.6861 0.7988   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.42660 0.11088 3.848 0.000207 \*\*\*  
lanes\_of\_traffic -0.05635 0.03600 -1.565 0.120576   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4413 on 103 degrees of freedom  
 (31 observations deleted due to missingness)  
Multiple R-squared: 0.02324, Adjusted R-squared: 0.01375   
F-statistic: 2.45 on 1 and 103 DF, p-value: 0.1206

***Regression Analysis Between Western District Accidents and Biker Age***

Call:  
lm(formula = western\_Accidents ~ biker\_age, data = bikeData,   
 na.action = "na.omit")  
  
Coefficients:  
(Intercept) biker\_age   
 0.178578 0.006959

Call:  
lm(formula = western\_Accidents ~ biker\_age, data = bikeData,   
 na.action = "na.omit")  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.6031 -0.3804 -0.2656 0.5239 0.7518   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.178578 0.080740 2.212 0.02873 \*   
biker\_age 0.006959 0.002241 3.105 0.00233 \*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4751 on 130 degrees of freedom  
 (4 observations deleted due to missingness)  
Multiple R-squared: 0.06906, Adjusted R-squared: 0.0619   
F-statistic: 9.644 on 1 and 130 DF, p-value: 0.002332

The below table highlights, in the first column of what the model for each variable predicted for the number of bike accidents. For bike lanes, I chose 4 and for biker age I chose 35. This is selfishly because I usually bike on 4 lane roads and am getting close to 35. The second column identifies whether if you had chose the the mean value of the number of accidents, would it have resulted in a significantly better or worse prediction.

Regression\_Model Accidents\_Predicted\_by\_Model Mean\_Better\_or\_Worse  
1 Central~Biker Age 0.4 Better  
2 Central~Lanes of Traffic 0.5 Better  
3 Eastern~Lanes of Traffic 0.2 Significantly Better  
4 Western~Biker Age 0.4 Better  
 F\_Ratio  
1 0.034  
2 0.014  
3 0.121  
4 0.0023

### Regression Model Conclusion

Based on the f Ratios and and R2 values, we would need to collect more data and additional data types to use these models when predicting the number of bike accidents by district in St. Paul, MN.

## Multiple Regression

I am now going to try and see if a Multiple Regression Model is something I can use with this data in an attempt to identify high risk causes for accidents by district. ***Multiple Regression Analysis Between Central District Accidents and Biker Age, Lanes of Traffic, Speed Limit and Driver Age***

Call:  
lm(formula = central\_Accidents ~ biker\_age + lanes\_of\_traffic +   
 speed\_limit + driver\_age, data = na.exclude(bikeData))  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.6143 -0.3902 -0.2236 0.5429 0.8376   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.076576 0.512814 -0.149 0.8816   
biker\_age -0.003915 0.002622 -1.493 0.1386   
lanes\_of\_traffic 0.085313 0.040464 2.108 0.0376 \*  
speed\_limit 0.014855 0.016957 0.876 0.3832   
driver\_age -0.002196 0.002873 -0.764 0.4464   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.481 on 97 degrees of freedom  
Multiple R-squared: 0.08454, Adjusted R-squared: 0.04679   
F-statistic: 2.24 on 4 and 97 DF, p-value: 0.07031

***Multiple Regression Analysis Between Eastern District Accidents and Biker Age, Lanes of Traffic, Speed Limit and Driver Age***

Call:  
lm(formula = eastern\_Accidents ~ biker\_age + lanes\_of\_traffic +   
 speed\_limit + driver\_age, data = na.exclude(bikeData))  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.3853 -0.2910 -0.2076 0.4519 0.8952   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.380870 0.467840 0.814 0.418  
biker\_age -0.002331 0.002392 -0.975 0.332  
lanes\_of\_traffic -0.053073 0.036915 -1.438 0.154  
speed\_limit 0.005920 0.015470 0.383 0.703  
driver\_age -0.001887 0.002621 -0.720 0.473  
  
Residual standard error: 0.4389 on 97 degrees of freedom  
Multiple R-squared: 0.03564, Adjusted R-squared: -0.00413   
F-statistic: 0.8962 on 4 and 97 DF, p-value: 0.4694

***Multiple Regression Analysis Between Western District Accidents and Biker Age, Lanes of Traffic, Speed Limit and Driver Age***

Call:  
lm(formula = western\_Accidents ~ biker\_age + lanes\_of\_traffic +   
 speed\_limit + driver\_age, data = na.exclude(bikeData))  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.7022 -0.3248 -0.1987 0.4562 0.8800   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.695706 0.491307 1.416 0.1600   
biker\_age 0.006246 0.002512 2.487 0.0146 \*  
lanes\_of\_traffic -0.032239 0.038767 -0.832 0.4077   
speed\_limit -0.020775 0.016246 -1.279 0.2040   
driver\_age 0.004084 0.002752 1.484 0.1412   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4609 on 97 degrees of freedom  
Multiple R-squared: 0.1038, Adjusted R-squared: 0.06686   
F-statistic: 2.809 on 4 and 97 DF, p-value: 0.02964

The next tables highlights how much better the multiple regression model predicts the variation in bike accidents per district compared to the regression model.

District Multiple\_Regression\_Percentage\_of\_Variation Additional\_Variation  
1 Central 8% 2%  
2 Eastern 4% 2%  
3 Western 10% 3%

Next I will calculate he standardized betas for the predictors in the multiple regression models: ***Central District Regression Model Betas***

biker\_age lanes\_of\_traffic speed\_limit driver\_age   
 -0.14563794 0.20965375 0.08726232 -0.07466148

The Beta for Biker Age: **-0.15** This value of **-0.15** indicates that as the bikers age decreases by one standard deviation(**19**), Central District bike accidents decrease by **-0.15** standard deviations. The standard deviation for Central District bike accidents is **0.48** and so this constitutes a change of **-0.07** Central District bike accidents. Therefore, for every **19 years** decrease in biker age, an extra **-0.07** is added to the amount of Central District bike accidents. This interpretation is true only if the effects of Lanes of Traffic, Speed Limit and Driver Age are held constant. The Beta for Lanes of Traffic: **0.21** This value of **0.21** indicates that as the number of lanes increases by one standard deviation(**1.2**), Central District bike accidents increase by **0.21** standard deviations. The standard deviation for Central District bike accidents is **0.48** and so this constitutes a change of **0.1** Central District bike accidents. Therefore, for every **1.2** increase in lanes of traffic, an extra **0.1** is added to the amount of Central District bike accidents. This interpretation is true only if the effects of Biker Age, Speed Limit and Driver Age are held constant. The Beta for Speed Limit: **0.087** This value of **0.087** indicates that as the Speed Limit increases by one standard deviation(**4.4**), Central District bike accidents increase by **0.087** standard deviations. The standard deviation for Central District bike accidents is **0.48** and so this constitutes a change of **0.042** Central District bike accidents. Therefore, for every **4.4** increase in speed limit, an extra **0.042** is added to the amount of Central District bike accidents. This interpretation is true only if the effects of Biker Age, Lanes of Traffic and Driver Age are held constant. The Beta for Driver Age: **-0.075** This value of **-0.075** indicates that as the Driver Age decreases by one standard deviation(**16**), Central District bike accidents decrease by **-0.075** standard deviations. The standard deviation for Central District bike accidents is **0.48** and so this constitutes a change of **-0.036** Central District bike accidents. Therefore, for every **4.4 years** decrease in Driver age, an extra **-0.036** is added to the amount of Central District bike accidents. This interpretation is true only if the effects of Biker Age, Lanes of Traffic and Speed Limit are held constant. ***Eastern District Regression Model Betas***

biker\_age lanes\_of\_traffic speed\_limit driver\_age   
 -0.09757828 -0.14673323 0.03912097 -0.07217808

The Beta for Biker Age: **-0.098** This value of **-0.098** indicates that as the bikers age decreases by one standard deviation(**19**), Eastern District bike accidents decrease by **-0.098** standard deviations. The standard deviation for Eastern District bike accidents is **0.43** and so this constitutes a change of **-0.042** Eastern District bike accidents. Therefore, for every **19 years** decrease in biker age, an extra **-0.042** is added to the amount of Eastern District bike accidents. This interpretation is true only if the effects of Lanes of Traffic, Speed Limit and Driver Age are held constant. The Beta for Lanes of Traffic: **-0.15** This value of **-0.15** indicates that as the number of lanes decreases by one standard deviation(**1.2**), Eastern District bike accidents increase by **-0.15** standard deviations. The standard deviation for Central District bike accidents is **0.43** and so this constitutes a change of **-0.063** Eastern District bike accidents. Therefore, for every **1.2** decrease in lanes of traffic, an extra **-0.063** is added to the amount of Eastern District bike accidents. This interpretation is true only if the effects of Biker Age, Speed Limit and Driver Age are held constant. The Beta for Speed Limit: **0.039** This value of **0.039** indicates that as the Speed Limit increases by one standard deviation(**4.4**), Eastern District bike accidents increase by **0.039** standard deviations. The standard deviation for Eastern District bike accidents is **0.43** and so this constitutes a change of **0.017** Central District bike accidents. Therefore, for every **4.4** increase in speed limit, an extra **0.017** is added to the amount of Eastern District bike accidents. This interpretation is true only if the effects of Biker Age, Lanes of Traffic and Driver Age are held constant. The Beta for Driver Age: **-0.072** This value of **-0.072** indicates that as the Driver Age decreases by one standard deviation(**16**), Eastern District bike accidents decrease by **-0.072** standard deviations. The standard deviation for Eastern District bike accidents is **0.43** and so this constitutes a change of **-0.031** Eastern District bike accidents. Therefore, for every **4.4 years** decrease in Driver age, an extra **-0.031** is added to the amount of Eastern District bike accidents. This interpretation is true only if the effects of Biker Age, Lanes of Traffic and Speed Limit are held constant. ***Western District Regression Model Betas***

biker\_age lanes\_of\_traffic speed\_limit driver\_age   
 0.23997681 -0.08182033 -0.12602948 0.14336124

The Beta for Biker Age: **0.24** This value of **0.24** indicates that as the bikers age increases by one standard deviation(**19**), Western District bike accidents decrease by **0.24** standard deviations. The standard deviation for Western District bike accidents is **0.49** and so this constitutes a change of **0.12** Western District bike accidents. Therefore, for every **19 years** increase in biker age, an extra **0.12** is added to the amount of Western District bike accidents. This interpretation is true only if the effects of Lanes of Traffic, Speed Limit and Driver Age are held constant. The Beta for Lanes of Traffic: **-0.082** This value of **-0.082** indicates that as the number of lanes decreases by one standard deviation(**1.2**), Western District bike accidents increase by **-0.082** standard deviations. The standard deviation for Western District bike accidents is **0.49** and so this constitutes a change of **-0.04** Western District bike accidents. Therefore, for every **1.2** decrease in lanes of traffic, an extra **-0.04** is added to the amount of Western District bike accidents. This interpretation is true only if the effects of Biker Age, Speed Limit and Driver Age are held constant. The Beta for Speed Limit: **-0.13** This value of **-0.13** indicates that as the Speed Limit decreases by one standard deviation(**4.4**), Western District bike accidents increase by **-0.13** standard deviations. The standard deviation for Western District bike accidents is **0.49** and so this constitutes a change of **-0.062** Western District bike accidents. Therefore, for every **4.4** increase in speed limit, an extra **-0.062** is added to the amount of Western District bike accidents. This interpretation is true only if the effects of Biker Age, Lanes of Traffic and Driver Age are held constant. The Beta for Driver Age: **0.14** This value of **0.14** indicates that as the Driver Age decreases by one standard deviation(**16**), Western District bike accidents decrease by **0.14** standard deviations. The standard deviation for Western District bike accidents is **0.49** and so this constitutes a change of **0.07** Western District bike accidents. Therefore, for every **4.4 years** decrease in Driver age, an extra **0.07** is added to the amount of Western District bike accidents. This interpretation is true only if the effects of Biker Age, Lanes of Traffic and Speed Limit are held constant. ***Central District Confidence Intervals***

2.5 % 97.5 %  
(Intercept) -1.094368938 0.941217788  
biker\_age -0.009118081 0.001288520  
lanes\_of\_traffic 0.005003780 0.165621542  
speed\_limit -0.018799585 0.048509963  
driver\_age -0.007898350 0.003505781

In this multiple regression model, only one predictors (lanes of traffic) has a small confidence interval and does not cross zero, indicating that the estimates for this parameters is likely to be representative of the true population values and is significant. ***Eastern District Confidence Intervals***

2.5 % 97.5 %  
(Intercept) -0.547663139 1.309403138  
biker\_age -0.007078404 0.002415541  
lanes\_of\_traffic -0.126339096 0.020192528  
speed\_limit -0.024783585 0.036622932  
driver\_age -0.007089265 0.003314726

In this multiple regression model, none of the predictors do not cross zero, indicating that the estimates for these parameters are not likely to be representative of the true population values and are not significant. ***Western District Confidence Intervals***

2.5 % 97.5 %  
(Intercept) -0.279403400 1.670814551  
biker\_age 0.001261129 0.011231296  
lanes\_of\_traffic -0.109180251 0.044701496  
speed\_limit -0.053018217 0.011468492  
driver\_age -0.001379377 0.009546485

In this multiple regression model, only one predictors (biker age) has a small confidence interval and does not cross zero, indicating that the estimates for this parameters is likely to be representative of the true population values and is significant. I will now compare the regression ### Multiple Regression Model Conclusion Based on the findings above, I cannot say that this model, although better than the regression model, accurately predicts the number of bike accidents per district based on the low importance on the variables predicting the number of accidents per district. Not all the assumptions were met and we can not assume this model would not generalize to district’s bike accident volume.

## Logistical Regression Model

The next model I am attempting to use to help predict the number of accidents per district is the Logistical Regression Model. I think this model will be more useful because I will be able to pull in additional categorical data points that I was not able to use in the Regression and Multiple Regression models above.

***Logistical Regression Analysis Between Central District Accidents and Biker Age, Injury to Biker, Lanes of Traffic, Signal Present, Speed Limit, Driver Age, Season and Driver Skill Level***

[1] TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE FALSE TRUE  
[13] TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE  
  
Call:  
glm(formula = central\_Accidents ~ biker\_age + injury\_to\_biker +   
 lanes\_of\_traffic + signal\_present + speed\_limit + driver\_age +   
 season + driver\_skill\_level, family = binomial(), data = train)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-1.8378 -0.8981 -0.5778 1.0215 2.1559   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -9.099e-01 3.177e+00 -0.286 0.7746   
biker\_age -2.653e-02 1.508e-02 -1.759 0.0785 .  
injury\_to\_bikerYes 9.539e-01 7.031e-01 1.357 0.1749   
lanes\_of\_traffic 3.282e-01 2.442e-01 1.344 0.1790   
signal\_presentRRFB -1.440e+01 1.455e+03 -0.010 0.9921   
signal\_presentStop Sign -8.717e-03 6.793e-01 -0.013 0.9898   
signal\_presentTraffic Signal 2.043e+00 1.516e+00 1.347 0.1778   
signal\_presentYes 7.178e-01 6.316e-01 1.136 0.2558   
speed\_limit 2.624e-02 1.003e-01 0.262 0.7936   
driver\_age -3.956e-02 2.533e-02 -1.561 0.1184   
seasonSpring -1.421e-01 5.884e-01 -0.242 0.8091   
seasonWinter 1.426e-01 8.893e-01 0.160 0.8726   
driver\_skill\_levelNew Driver -3.071e-01 8.819e-01 -0.348 0.7277   
driver\_skill\_levelSenior Driver 1.215e+00 1.292e+00 0.940 0.3472   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 116.44 on 87 degrees of freedom  
Residual deviance: 100.01 on 74 degrees of freedom  
 (21 observations deleted due to missingness)  
AIC: 128.01  
  
Number of Fisher Scoring iterations: 14  
 10 19 26 36 57 62 77 87   
 NA NA NA NA 0.3495657 NA 0.4938779 NA   
 105 111 123 144 155 160 169 198   
0.1887856 0.2531563 NA 0.1604926 0.1753293 NA NA 0.4699789   
 211 222 235 243 255 262 269 287   
0.3890956 NA 0.3746673 0.3224919 0.6576479 0.3795355 0.5577541 NA   
 297 306 314   
0.2758949 NA NA   
 4 8 9 13 15 17   
1.328550e-01 6.427085e-01 4.283554e-01 6.192531e-01 NA 1.414539e-01   
 22 23 27 28 29 31   
7.320406e-01 4.255080e-01 NA 2.822946e-01 5.937347e-01 1.520071e-01   
 32 33 41 42 44 45   
2.416870e-01 5.804671e-01 NA 1.917945e-01 3.007932e-01 4.891010e-01   
 47 59 60 61 69 70   
1.331231e-01 NA 2.305555e-01 4.145487e-01 6.351365e-01 4.019921e-01   
 79 82 83 84 85 86   
 NA 1.472628e-01 4.337122e-01 2.740169e-01 8.836636e-02 5.646773e-01   
 88 89 90 91 98 107   
1.736770e-07 5.754734e-01 1.891953e-01 NA 4.492139e-01 NA   
 108 110 112 119 125 127   
 NA 1.191286e-01 4.636094e-01 NA 7.855420e-01 2.406372e-01   
 128 130 134 143 145 146   
3.700213e-01 4.746509e-01 5.585519e-01 7.764901e-01 3.645152e-01 1.716267e-01   
 149 152 154 156 157 158   
 NA NA 3.391509e-01 5.449397e-01 2.761821e-01 NA   
 163 168 173 175 176 189   
8.194923e-01 3.616140e-01 3.705385e-01 NA 3.974298e-01 8.152686e-01   
 190 191 199 200 201 202   
 NA NA 1.555134e-01 1.629952e-01 1.965731e-01 4.075404e-01   
 203 215 216 217 227 231   
4.456132e-01 NA 2.353169e-01 6.576362e-01 3.881726e-01 3.266818e-01   
 236 237 238 239 240 242   
3.921972e-01 NA 5.224498e-01 2.464082e-01 5.328386e-01 3.960739e-01   
 245 247 252 253 254 257   
1.762931e-01 2.754375e-01 6.659975e-01 2.750435e-01 2.253110e-01 3.624984e-01   
 260 261 265 267 271 274   
7.399270e-02 2.222937e-01 NA 1.584925e-01 3.071447e-01 4.559969e-01   
 275 277 278 286 288 291   
1.486378e-01 5.844642e-01 6.812163e-01 NA 5.926831e-01 NA   
 294 295 296 300 301 302   
6.783019e-01 2.137253e-01 2.754050e-01 3.194149e-01 3.294509e-01 2.014412e-01   
 309 310 316 317 318 320   
2.183011e-01 6.141905e-01 8.194923e-01 1.281735e-01 9.788662e-02 9.198681e-02   
 324   
 NA   
 Predicted\_Value  
Actual\_Value FALSE TRUE  
 0 46 9  
 1 18 15  
[1] 0.6931818

***Logistical Regression Analysis Between Eastern District Accidents and Biker Age, Injury to Biker, Lanes of Traffic, Signal Present, Speed Limit, Driver Age, Season and Driver Skill Level***

[1] TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE TRUE  
[13] TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE  
  
Call:  
glm(formula = eastern\_Accidents ~ biker\_age + injury\_to\_biker +   
 lanes\_of\_traffic + signal\_present + speed\_limit + driver\_age +   
 season + driver\_skill\_level, family = binomial(), data = train)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-1.8998 -0.8081 -0.5621 1.0110 1.8940   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.18518 2.83978 0.417 0.6764   
biker\_age -0.01639 0.01563 -1.048 0.2945   
injury\_to\_bikerYes -1.38453 0.69726 -1.986 0.0471 \*  
lanes\_of\_traffic -0.44668 0.29092 -1.535 0.1247   
signal\_presentRRFB -14.48428 1455.39763 -0.010 0.9921   
signal\_presentStop Sign -0.14586 0.70008 -0.208 0.8350   
signal\_presentTraffic Signal 0.74500 1.61770 0.461 0.6451   
signal\_presentYes -0.22877 0.69432 -0.329 0.7418   
speed\_limit 0.03484 0.08702 0.400 0.6889   
driver\_age 0.01001 0.02669 0.375 0.7076   
seasonSpring -0.99936 0.59642 -1.676 0.0938 .  
seasonWinter -1.47674 1.18055 -1.251 0.2110   
driver\_skill\_levelNew Driver 0.36349 1.02910 0.353 0.7239   
driver\_skill\_levelSenior Driver -0.67142 1.35798 -0.494 0.6210   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 101.836 on 85 degrees of freedom  
Residual deviance: 89.647 on 72 degrees of freedom  
 (23 observations deleted due to missingness)  
AIC: 117.65  
  
Number of Fisher Scoring iterations: 14  
 10 13 26 36 57 59 77 87   
 NA 0.2193999 NA NA 0.1908197 NA 0.5084624 NA   
 105 107 123 144 155 156 169 198   
0.6447504 NA NA 0.1250455 0.1950542 0.4911440 NA 0.2645417   
 211 215 235 243 255 257 269 287   
0.2453478 NA 0.4810417 0.1803635 0.3368457 0.2446330 0.1029600 NA   
 297 300 314   
0.3178042 0.3563716 NA   
 4 8 9 15 17 19   
8.505579e-02 5.414765e-02 6.211504e-01 NA 3.118712e-01 NA   
 22 23 27 28 29 31   
4.893349e-02 2.493796e-01 NA 2.805406e-01 1.586550e-01 3.703418e-01   
 32 33 41 42 44 45   
3.108487e-01 1.290470e-01 NA 5.927730e-01 8.354571e-01 1.514929e-01   
 47 60 61 62 69 70   
2.646410e-01 3.251835e-01 1.053925e-01 NA 2.826343e-01 1.108725e-01   
 79 82 83 84 85 86   
 NA 7.060776e-01 3.558157e-01 5.716981e-01 7.703482e-02 5.137927e-01   
 88 89 90 91 98 108   
1.736770e-07 1.663483e-01 5.120436e-01 NA 1.961774e-01 NA   
 110 111 112 119 125 127   
5.162848e-01 5.039179e-01 2.889449e-01 NA 3.772336e-01 4.422995e-01   
 128 130 134 143 145 146   
6.954947e-02 2.662761e-01 8.354800e-02 3.408103e-02 2.725780e-01 4.198334e-01   
 149 152 154 157 158 160   
 NA NA 9.770609e-02 3.281145e-01 NA NA   
 163 168 173 175 176 189   
2.282090e-01 2.002544e-01 3.810501e-01 NA 3.401455e-01 2.927927e-01   
 190 191 199 200 201 202   
 NA NA 3.315011e-01 5.216394e-01 2.030265e-01 5.019431e-01   
 203 216 217 222 227 231   
1.051179e-01 7.753821e-02 3.111997e-01 NA 4.449220e-01 2.392868e-01   
 236 237 238 239 240 242   
2.404931e-01 NA 2.222103e-01 2.853380e-01 2.887646e-01 1.607448e-01   
 245 247 252 253 254 260   
7.414731e-01 1.456779e-01 2.264453e-01 1.465495e-01 1.076934e-01 3.731375e-01   
 261 262 265 267 271 274   
1.333197e-01 3.669787e-01 NA 1.678788e-01 3.029185e-01 3.370951e-01   
 275 277 278 286 288 291   
2.504105e-01 2.887583e-01 1.022385e-01 NA 1.491035e-01 NA   
 294 295 296 301 302 306   
1.705990e-01 2.901844e-01 1.673248e-01 2.223236e-01 1.119052e-01 NA   
 309 310 316 317 318 320   
3.821885e-01 2.191393e-01 2.282090e-01 3.623626e-01 1.955262e-01 3.466083e-01   
 324   
 NA   
 Predicted\_Value  
Actual\_Value FALSE TRUE  
 0 57 5  
 1 17 7  
[1] 0.744186

***Logistical Regression Analysis Between Western District Accidents and Biker Age, Injury to Biker, Lanes of Traffic, Signal Present, Speed Limit, Driver Age, Season and Driver Skill Level***

[1] TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
[13] TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE  
  
Call:  
glm(formula = western\_Accidents ~ biker\_age + injury\_to\_biker +   
 lanes\_of\_traffic + signal\_present + speed\_limit + driver\_age +   
 season + driver\_skill\_level, family = binomial(), data = train)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-1.4166 -0.8294 -0.6150 0.9770 2.5207   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -7.23539 4.80246 -1.507 0.1319   
biker\_age 0.03022 0.01629 1.855 0.0636 .  
injury\_to\_bikerYes 0.86308 0.76925 1.122 0.2619   
lanes\_of\_traffic -0.05307 0.28371 -0.187 0.8516   
signal\_presentRRFB 17.76922 3956.18036 0.004 0.9964   
signal\_presentStop Sign -0.04289 0.73643 -0.058 0.9536   
signal\_presentTraffic Signal -17.89619 1756.11347 -0.010 0.9919   
signal\_presentYes 0.26835 0.65079 0.412 0.6801   
speed\_limit 0.10138 0.14569 0.696 0.4865   
driver\_age 0.04452 0.02613 1.704 0.0884 .  
seasonSpring 0.39889 0.60827 0.656 0.5120   
seasonWinter -0.95743 1.07580 -0.890 0.3735   
driver\_skill\_levelNew Driver -0.04426 1.06835 -0.041 0.9670   
driver\_skill\_levelSenior Driver -0.49754 1.21603 -0.409 0.6824   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 106.548 on 81 degrees of freedom  
Residual deviance: 88.081 on 68 degrees of freedom  
 (28 observations deleted due to missingness)  
AIC: 116.08  
  
Number of Fisher Scoring iterations: 16  
 17 31 33 36 61 84 86   
0.06074693 0.78033132 0.15568693 NA 0.56823703 0.32811375 0.11382607   
 87 110 130 143 144 158 189   
 NA 0.23238131 0.18856483 0.16614190 0.77349233 NA 0.10821435   
 191 198 217 239 242 243 261   
 NA 0.20052635 0.13457982 0.42238897 0.09139326 0.84792599 0.67209267   
 277 286 287 302 320   
0.39364275 NA NA 0.20191876 0.84580181   
 4 8 9 10 13 15   
9.282643e-01 2.335996e-01 6.083599e-02 NA 1.492673e-01 NA   
 19 22 23 26 27 28   
 NA 4.151643e-01 2.255638e-01 NA NA 3.725575e-01   
 29 32 41 42 44 45   
2.114938e-01 4.289833e-01 NA 1.312509e-01 4.171054e-02 3.021015e-01   
 47 57 59 60 62 69   
5.925938e-01 5.132084e-01 NA 6.660782e-01 NA 1.890431e-01   
 70 77 79 82 83 85   
3.805167e-01 3.418522e-01 NA 2.594975e-01 1.905656e-01 8.533607e-01   
 88 89 90 91 98 105   
1.000000e+00 5.010076e-08 2.794834e-01 NA 5.477337e-01 1.865195e-01   
 107 108 111 112 119 123   
 NA NA 3.564088e-01 2.048856e-01 NA NA   
 125 127 128 134 145 146   
5.970920e-08 7.469252e-01 6.907354e-01 4.834318e-01 2.907594e-01 1.954477e-01   
 149 152 154 155 156 157   
 NA NA 1.366018e-01 6.333707e-01 2.110052e-01 3.928195e-01   
 160 163 168 169 173 175   
 NA 4.739541e-09 3.019250e-01 NA 4.810030e-01 NA   
 176 190 199 200 201 202   
3.416716e-01 NA 6.005066e-01 5.084680e-01 4.637376e-01 2.083167e-01   
 203 211 215 216 222 227   
2.547147e-01 2.674290e-01 NA 8.222625e-01 NA 3.546887e-01   
 231 235 236 237 238 240   
3.197549e-01 2.819662e-01 2.646372e-01 NA 2.836154e-01 1.585207e-01   
 245 247 252 253 254 255   
1.243835e-01 4.057364e-01 2.280038e-01 3.563601e-01 6.258433e-01 1.848982e-01   
 257 260 262 265 267 269   
2.819946e-01 6.186880e-01 8.918555e-02 NA 5.928770e-01 1.506127e-01   
 271 274 275 278 288 291   
3.411762e-01 1.531265e-01 6.114822e-01 2.597953e-01 3.876811e-01 NA   
 294 295 296 297 300 301   
2.934862e-01 4.738520e-01 5.508543e-01 4.745170e-01 2.307921e-01 3.170042e-01   
 306 309 310 314 316 317   
 NA 4.232898e-01 2.786105e-01 NA 4.739541e-09 2.052184e-01   
 318 324   
5.836764e-01 NA   
 Predicted\_Value  
Actual\_Value FALSE TRUE  
 0 49 4  
 1 14 15  
[1] 0.7804878

### Logistical Regression Conclusion

District Model\_Accuracy Variables\_Greatest\_Influence  
1 Central 67% None  
2 Eastern 74% Season - Spring  
3 Western 66% Biker Age

The table above identified that the model is 66% accurate and above for all three districts, with the Eastern District coming in at 74% accurate. This tells me that the Logistic Regression Model is a much better use case for this data set and predicting the number of bike accidents by district.

# Conclusion

Based on the analysis of the data with three models - Regression, Multiple Regression and Logistical - I can say that the correct model for predicting bike accidents in St. Paul, MN is the Logistical model. The regression and multiple regression models could only account for 10% or less of the variability in bike accident volumes per district. This is not a good predictor and we would need to collect more quantitative data in order to strengthen those models.

The Logistical Regression model however did a much better job at predicting the volume of bike accidents for each district. The accuracy of over 66% was much higher than the regression and multiple regression models and I believe this is because of the additional categorical variables that were included in this model helped the model make a more accurate prediction versus only using numerical values.

66%, 67% and 74% accuracy rate for the three models are great numbers, but this model could always do better and get stronger if it were to be used in a real life setting. I think to increase this accuracy, we would need to collect more accurate data (data quality checking) and see if there are additional metrics we can add to the model that would help strengthen the models ability to predict the volume of bike accidents based on the variables input into the model.