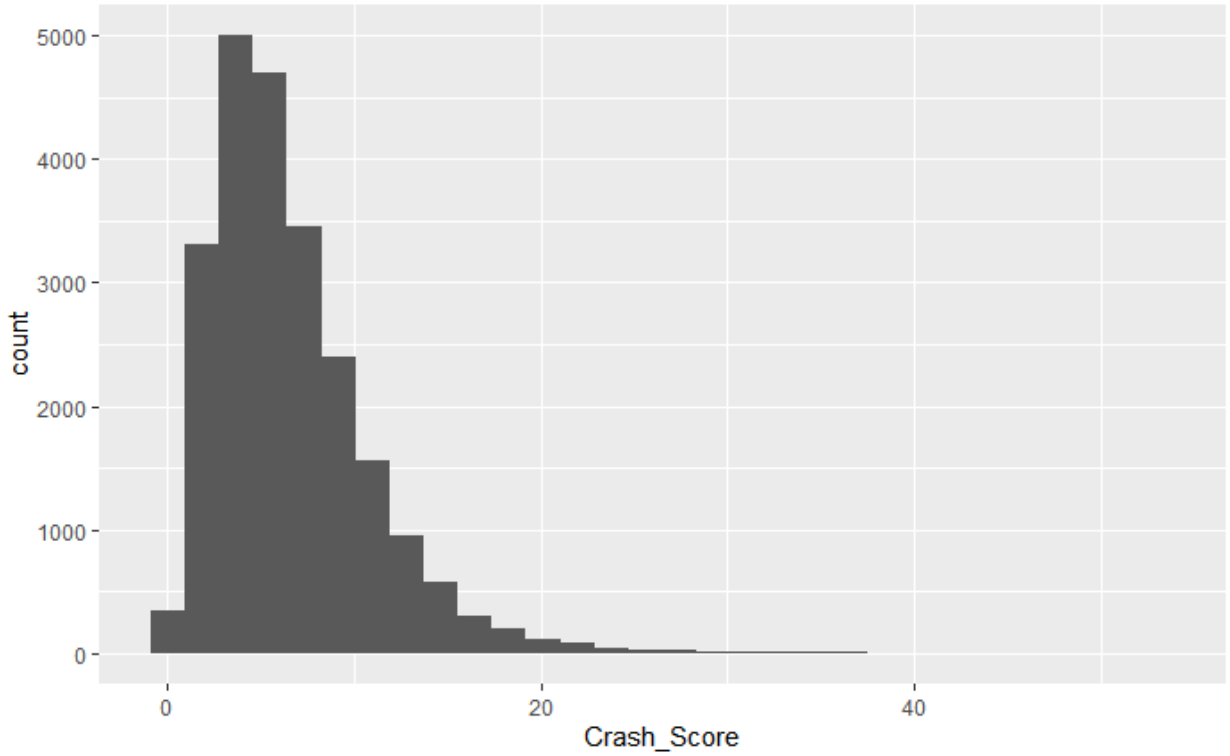
Practice Exam – Traffic Safety Solution

Task 1 – Explore the relationship of each variable to the *Crash* (5 points)

Note that the crash scores are positive and right skewed with a mean of 6.6 and a median of 5.7. The values range from 0 to 40 and measure the severity or extent of the crash.

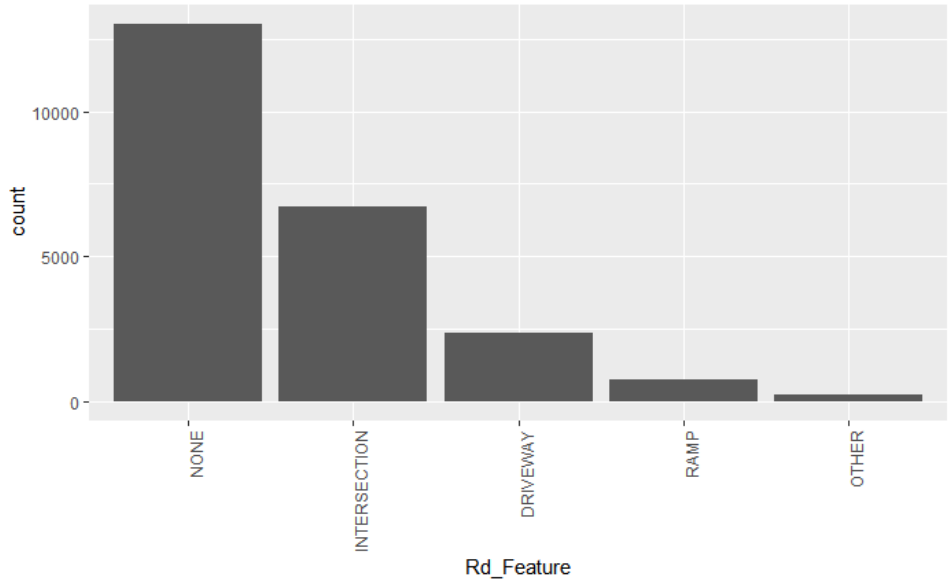


The summary stats support the following conclusions:

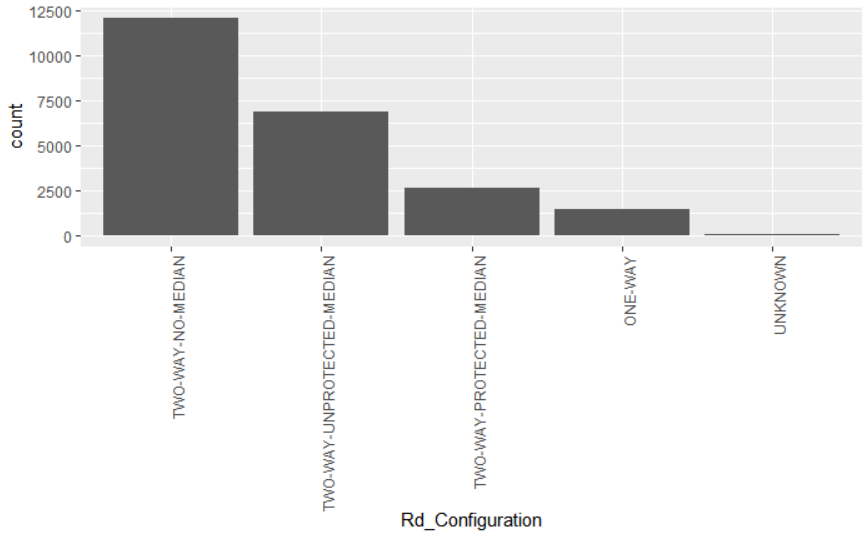
- Rd_feature values are highest in intersections;
- The highest median crash scores in Rd_Character values occur on straight roads. The mean has not been used as the distribution is skewed;
- Rd_Class values show that the most severe crashes are on highways;
- Rd_Configuration values are highest for intersections and two-way roads;
- Rd_Surface values are highest for grooved concrete;
- Most severe crashes occur when Rd_Conditions are icy, snowy, or slushy
- Analysis of light data shows that most severe crashes occurs during times of poor visibility: e.g. dawn, dusk, and at night;
- Most severe crashes occur in bad weather (e.g. snow, rain);
- The highest crash scores occur at stop signs; and,
- Most severe crashes are located in roadwork/construction zones.

The most significant bar charts included:

- Rd_Feature – the majority have no crashes, followed by the categories INTERSECTIONS and DRIVEWAYS.



Rd_Configuration data show that most crashes occur on two-way roads. There are fewer on one-way roads. This makes sense as dealing with oncoming traffic may make driving more dangerous.



Task 2 – Reduce the number of factor levels where appropriate (5 points)

I reduced the factor levels for the following variables:

Rd_feature

- The category RAMP was grouped into OTHER. .

Rd_Character

- Rd_configuration was simplified into 2 categories - OTHER and Two-Way.

Rd_Surface

- GROOVED CONCRETE was included in the OTHER category.

Traffic_Control

- Was limited to 2 categories - OTHER or STOP.

Task 3 – Use observations from principal components analysis (PCA) to generate a new feature (9 points)

Principal component analysis is a dimensionality reduction method which collapses data into fewer principal components, thereby reducing the dimensionality. Each principal component (PC) is a linear combination of the original variables in which the loadings serve as weights.

First, all factors are converted into numeric 0/1 indicators. These variables include Rd_Conditions, Light, and Weather. The weight information is then used to construct features capturing information about poor weather conditions. This includes:

- -0.51 Rd_ConditionsDRY
- 0.5 Rd_ConditionsWET
- -0.46 WeatherCLEAR

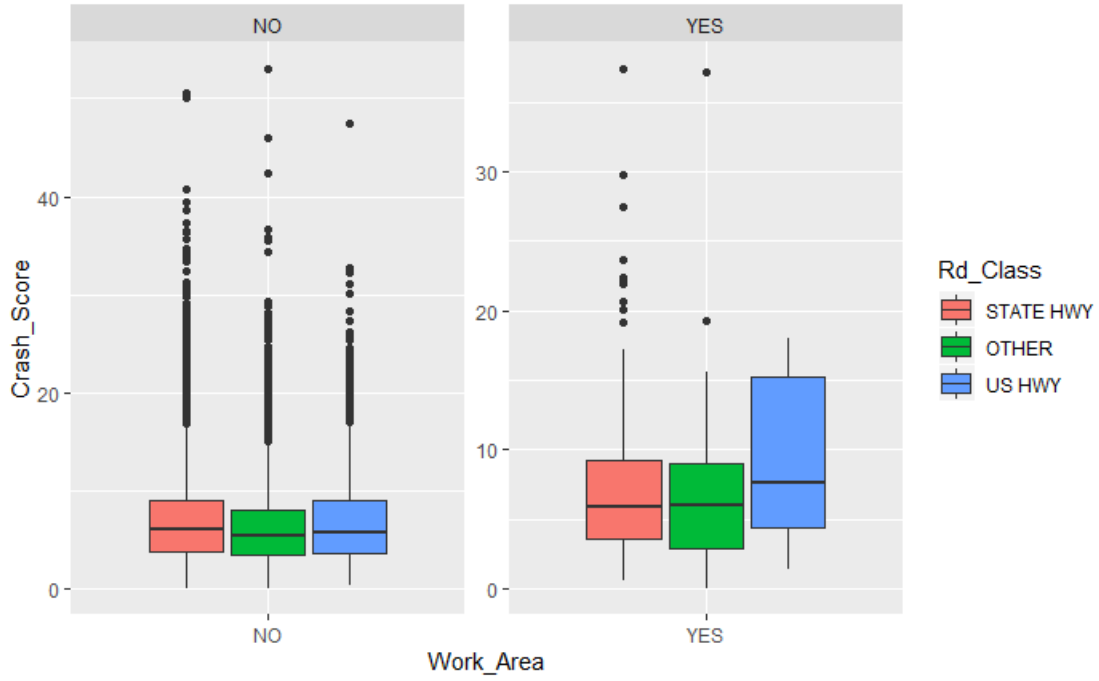
This feature is then added to the data.

Task 4 –Select an interaction (7 points)

An interaction is when the impact that a variable has on the target depends on the value of another variable. This case considers two interactions: work area/road class and time of day/road configuration.

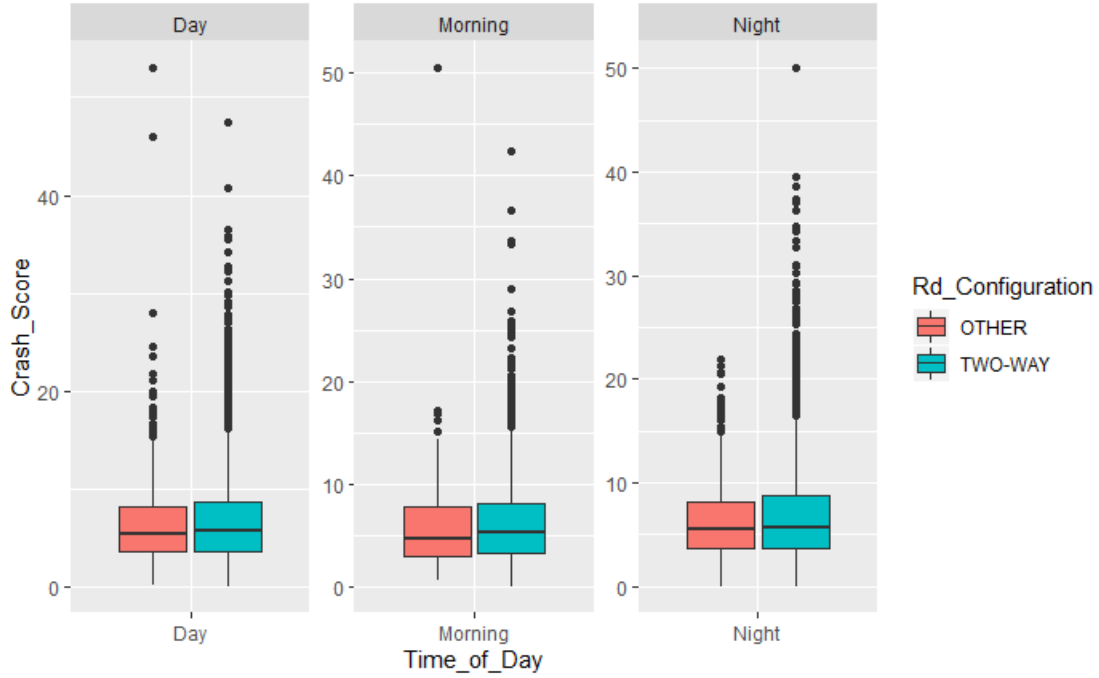
Work Area and Road Class

Construction zones on highways have a different risk of accidents than those on main roads. The graph below shows the distribution of crash scores for non-work areas (left) and work areas (right). Driving on US highways is riskier in work areas than in non-work areas. Data shows that there is an interaction.



Time of Day and Road Configuration

The graphs below show accidents during the day, morning, and night times. It is possible that highways are riskier at night than during the day. However, all six plots show about the same distribution. The median crash score is about 7 and is flat, which means that no interaction is present.



Task 5 – Select a distribution and link function (10 points)

A GLM is a model that relates the mean of a response distribution to a linear predictor through a link function. The crash score distribution is right skewed and strictly positive. This means that probability distributions such as an Inverse Gaussian or Gamma would be good choices for analysis.

The mean of Gamma and Inverse Gaussian distributions are always positive, therefore they require a link function which results in a positive mean. This suggests that a log, square root, or inverse squared are appropriate. A log enables simpler interpretation of the model. This results in the model being multiplicative and the coefficients being interpreted as percentage changes in the crash score.

The two options selected are:

- A Gamma response family with a log link; and
- An Inverse Gaussian response family with a log link.

Before the model was fitted, a training and test set was created.

An OLS model is fit with all variables, resulting in:

```
[1] "AIC"
[1] 99449.51
[1] "RMSE"
[1] 4.284356
```

A Gamma with a log link shows:

```
[1] "AIC"
[1] 93272.25
[1] "RMSE"
[1] 4.284246
```

The AIC is based on the log of the likelihood of generating the data with an added penalty term for the number of parameters. A lower AIC value indicates that the model is a better fit. The Gamma has a lower (worse) AIC than the OLS. The OLS model has a worse (higher) RMSE as well.

The inverse Gaussian model does not converge; the Gamma with a log link is the chosen model.

Task 6 – Select features using AIC or BIC (12 points)

For L (log likelihood) and p (number of parameters) the AIC is -2L + 2p.

The BIC considers sample size *n* so the penalty on the likelihood is greater for larger datasets than for smaller ones. The BIC is -2L + log(n)p.

In our data, the log(n) is about 10, which means that the BIC favors models which are simpler than AIC. In order to create the desired interpretable model, fewer variables are needed, favoring BIC over AIC. BIC is better because it creates a simpler model; however, it became too simple as it resulted in only two variables being included. Therefore AIC is ultimately chosen, resulting in the remaining categories of Rd_Class, RD_feature, Traffic_Control, Rd_Surface, and Rd_Character.

Many of the variables are determined to be not significant.

```
Call:
glm(formula = Crash_Score ~ Rd_Class + Rd_Feature + Traffic_Control +
    Rd_Surface + Rd_Character + year, family = Gamma(link = "log"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3101  -0.5539  -0.1390   0.2794   3.4335

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               15.791616   6.765289   2.334 0.019596 *
Rd_ClassOTHER             -0.088038   0.011979  -7.349 2.08e-13 ***
Rd_ClassUS HWY             0.031072   0.019377   1.604 0.108832
Rd_FeatureINTERSECTION     0.043222   0.014679   2.944 0.003239 **
Rd_FeatureDRIVEWAY         0.017420   0.017511   0.995 0.319844
Rd_FeatureRAMP            -0.086850   0.033118  -2.622 0.008739 **
Rd_FeatureOTHER           -0.020332   0.047514  -0.428 0.668722
Traffic_ControlSIGNAL      0.045761   0.015285   2.994 0.002759 **
Traffic_ControlSTOP-SIGN   0.063688   0.018835   3.381 0.000723 ***
Traffic_ControlYIELD       0.042579   0.049472   0.861 0.389441
Traffic_ControlOTHER       0.061765   0.050920   1.213 0.225151
Rd_SurfaceCOARSE ASPHALT   0.035421   0.017531   2.020 0.043352 *
Rd_SurfaceCONCRETE        -0.081702   0.030131  -2.712 0.006704 **
Rd_SurfaceGROOVED CONCRETE 0.009109   0.040665   0.224 0.822765
Rd_SurfaceOTHER            0.044427   0.091253   0.487 0.626369
Rd_CharacterSTRAIGHT-GRADE 0.008948   0.015651   0.572 0.567509
Rd_CharacterCURVE-LEVEL   -0.058870   0.029138  -2.020 0.043364 *
Rd_CharacterSTRAIGHT-OTHER -0.069276   0.029007  -2.388 0.016941 *
Rd_CharacterCURVE-GRADE   -0.045811   0.031398  -1.459 0.144566
Rd_CharacterCURVE-OTHER    0.078848   0.048841   1.614 0.106461
Rd_CharacterOTHER         -0.089933   0.198028  -0.454 0.649732
year                      -0.006898   0.003355  -2.056 0.039821 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.4173638)

    Null deviance: 7311.8  on 17353  degrees of freedom
Residual deviance: 7220.7  on 17332  degrees of freedom
AIC: 93254

Number of Fisher Scoring iterations: 5
```

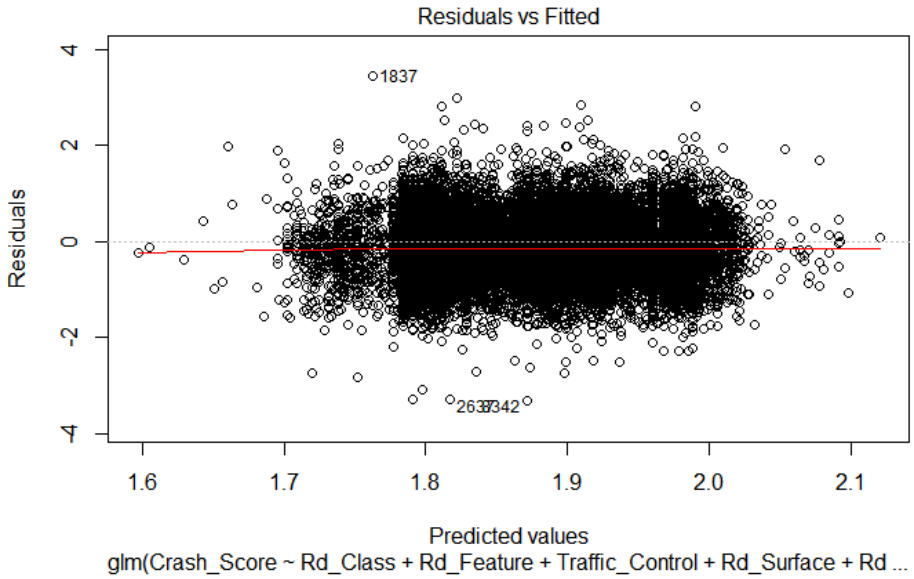
A forward direction starts with no variables in the model and then only adds those which improve the log likelihood (AIC or BIC). A backward direction starts with all variables and removes all that are not significant. In order to create a simpler model, the forward direction was used.

Task 7 – Validate the model (6 points)

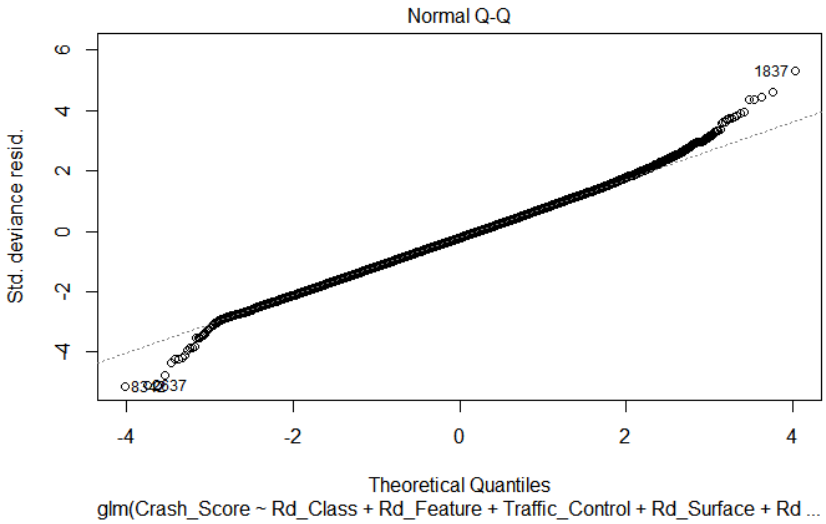
The AIC and RMSE are both better (lower) than the OLS model. The AIC is better (lower) after running a stepwise selection, but the RMSE is worse.

```
[1] "AIC"
[1] 93253.88
[1] "RMSE"
[1] 4.286682
```

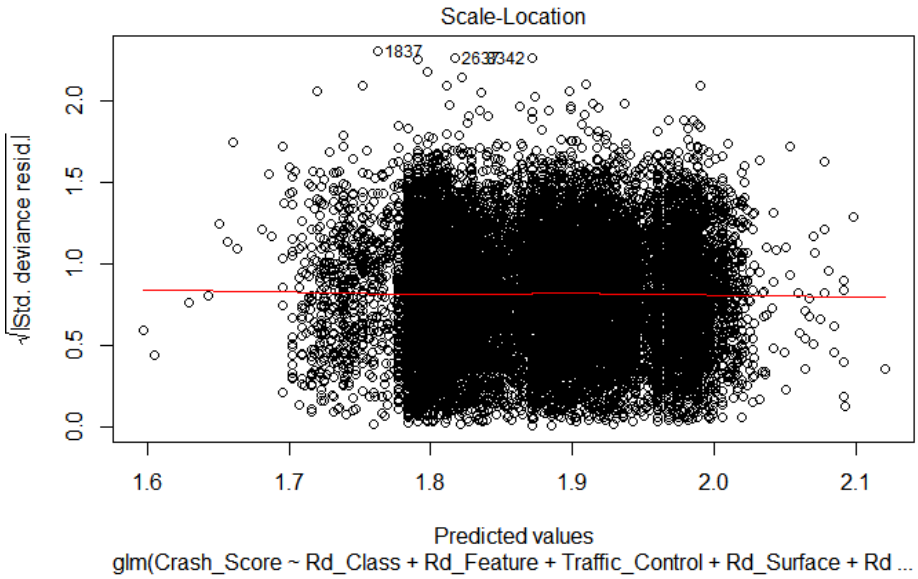
The raw residuals are centered at zero and are random, a good result.



A QQ plot shows the theoretical quantiles of the deviance residuals against the actual quantiles. These are approximately normal for most values. There is some deviation along the tails, but it is generally a good fit.



The deviance residuals against the predicted values are also centered at zero and are random.



Task 8 – Interpret the model (9 points)

A full data set was used to rerun the model. This helps to make the coefficients more consistent, as the larger training size results in smaller error rates. The target variable is the crash score, which is higher for accidents which are more severe and have more vehicles involved.

```
Call:
glm(formula = Crash_Score ~ Rd_Class + Rd_Feature + Traffic_Control +
    Rd_Surface + Rd_Character + year, family = Gamma(link = "log"),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3108  -0.5545  -0.1409   0.2785   3.4167

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.4185738)

    Null deviance: 9740.7  on 23136  degrees of freedom
Residual deviance: 9612.2  on 23115  degrees of freedom
AIC: 124324

Number of Fisher Scoring iterations: 5
```

As is shown below, roads with different characteristics demonstrate varied coefficients, the percentage difference in crash score. The crash score is:

- 9% lower for non-highways and non-state highways;
- 4% higher on US highways compared to non-highways;
- 5% higher for intersections compared to non-intersections;
- 3% higher for driveways;
- 9% lower for ramps;
- 2% lower for non-driveways and ramps;
- 5% higher for traffic signals; and,
- 8% higher at stop signs.

- 5% higher for non-stop signs.

A similar interpretation can be made based on road surface types and whether a road is straight or curved.

Term	Percentage Difference in Crash Score
Rd_ClassOTHER	-8.7%
Rd_ClassUS_HWY	4.4%
Rd_FeatureINTERSECTION	4.6%
Rd_FeatureDRIVEWAY	2.7%
Rd_FeatureRAMP	-8.8%
Rd_FeatureOTHER	-2.1%
Traffic_ControlSIGNAL	5.4%
Traffic_ControlSTOP-SIGN	7.5%
Traffic_ControlYIELD	5.5%
Traffic_ControlOTHER	4.3%
Rd_SurfaceCOARSE_ASPHALT	2.4%
Rd_SurfaceCONCRETE	-6.5%
Rd_SurfaceGROOVED_CONCRETE	0.0%
Rd_SurfaceOTHER	-3.7%
Rd_CharacterSTRAIGHT-GRADE	-0.1%
Rd_CharacterCURVE-LEVEL	-5.6%
Rd_CharacterSTRAIGHT-OTHER	-5.8%
Rd_CharacterCURVE-GRADE	-7.1%
Rd_CharacterCURVE-OTHER	3.5%
Rd_CharacterOTHER	-7.1%
year	-0.4%

Task 9 – Investigate ridge and LASSO regressions (12 points)

Ridge and Lasso are types of penalized regressions. Using the parameters *alpha* and *lambda*, we can impose a penalty on the model score based on the coefficient sizes. This helps to remove unneeded variables and make the coefficients smaller.

When alpha = 1, there is a LASSO, which means that the penalty is the L1 norm, or the sum of the absolute values of the coefficients. The parameter lambda controls how much this impacts the model. As lambda increases, the model is made simpler as variables are dropped. Cross-validation was used to choose the best lambda.

The RMSE of the lasso was better than that of the Gamma GLM with a forward stepwise selection.

RMSE LASSO

```
[1] 4.287291
```

RMSE Gamma GLM

```
4.286682
```

While the LASSO results in a better RMSE, there are 13 variables instead of 10 in the GLM. The Department of Transportation is looking for a model that is easy to interpret, therefore the GLM is better suited to this purpose. The Bias Variance tradeoff says that as the model's flexibility increases, or as more variables are added, the performance will decrease. In this case, although the Lasso has more parameters and is more flexible, it is also harder to interpret. The GLM has fewer parameters and is easier to interpret. If a backward selection had been used, there would have been more variables. If Ridge had been used instead of the Lasso, it would have included all variables and the error rate would probably be lower but there would have been problems with interpretation.

Task 10 – Consider a decision tree (5 points)

A decision tree separates crashes with a high score from those with a low score based on a series of yes/no questions. Advantages of using a tree include:

- Automatically detects interactions;
- Handles missing values (already managed by the assistant);
- Simple to interpret; and,
- Detects non-linearities (not relevant in this case because all variables are categorical).

The disadvantages include:

- As each prediction is an average of training samples, predicted crash scores would be stepwise (not relevant here as all predictors are categorical);
- The predictive power of a single tree tends to be lower than GLM, random forest, or GBM
- High variance leads to changing results after training on new data.

Task 11 – Executive summary (20 points)

The North Carolina Department of Transit wants help understanding the causes of vehicle crashes. Using predictive analytics, we have identified the leading causes of crashes, including road conditions, weather, and terrain. These findings can be used in order to determine risk factors for road accidents . The Department of Transit can use this information to improve the most dangerous thoroughfares.

Using historical data about vehicle crashes, we have identified which factors increase the frequency and severity of accidents. We used *crash score*, a numeric measure of both the number of vehicles involved and the amount of injury sustained. Our results show that the crash score is:

- 9% lower for non-highways and non-state highways;
- 4% higher on US highways compared to non-highways;

- 5% higher for intersections compared to non-intersections;
- 3% higher for driveways;
- 9% lower for ramps;
- 2% lower for non-driveways and ramps;
- 5% higher for traffic signals; and,
- 8% higher at stop signs.
- 5% higher for other types of traffic signs.

I also have similar relevant findings about the significance of road surface (asphalt or concrete) as well as the type of road (curved or straight).

I used data collected from 23,137 observations of vehicle crashes and compared information on road conditions, weather, lighting, time of day, season, and other road and traffic characteristics to the crash score. I made some observations regarding the number of accidents and crash scores at different areas and on different types of roads. I looked at summary stats and graphs. I then simplified the number of factors used based on relevance. As some information within the original data was too nuanced to be directly used in a model, the factors were simplified.

As three variables were related to weather and time of day, we used a dimensionality reduction technique known as PCA in order to combine these factors into a single variable. This took into account whether the road conditions were wet or dry, and if the weather was clear. This simplified information was the input to our predictive model.

I also looked at interactions, the relationships between variables. For example, construction on a highway may elevate the risk of accident when compared to those not undergoing roadwork. I looked at construction areas, road class, and road configuration (one- or two-way) and found that interactions between variables such as construction and road type did affect crash scores.