Project

Gaveen Alwis

2023-10-22

Final Project

Introduction

NOAA's National Weather Service Storm Prediction Center has provided a collection of data, in CSV format, on past tornado's that have swept across America between 1950-2022. The data has a plentiful number of rows,totaling 67,945 rows, and 27 columns. The columns after tidying consist of the data types: int,factor,date(strings were provided, however they were recoded into to factors for the purpose of analysis).Prior to the analysis, a series of cleaning steps were required to be completed.To begin, it was observed that the data naming scheme followed the conventional format, so nothing was changed. Following on from this, the observation was made that some of the data in character format and dbl format could be changed to factor and integer format respectively for ease of analysis (as shown in section 1.1). This completed the majority of the cleaning that could be done by eye. To clean the data that could not be seen by eye, a skim was conducted (shown in figure 1.2). This revealed that the column "sg" had a constant value of 1 for every row in the data set. The lack in variance, would not provide any useful information in the analysis, thus it was removed.Following on from this, it was observed that the columns "stn" and "closs" were not provided any contextual information from the github source the data was retrieved from. It was later discovered that these columns were "discontinued" as per the github,thus they were removed (shown in figure 1.4).Returning to the skim, it was observed that one row did not fit the conventional positive integers between 1-5 structure that was followed for the rest of the magnitude column. With context that the mag column indicates the magnitude of a tornados, and magnitudes do not go into the negatives, it was safe to assume that this row of data was a outlier, therefore it was removed.The skim data was then run one last time, and no other anomalies were found in the response, hence it was assumed that the data was fully clean (as shown in section 1.7.

Graphs

To begin analysis on tornado data set, a comparison between a variety of different variables were conducted through the means of graphing.

Firstly, a comparison was made between each state in America and the average magnitude of tornado's it had indured. Both magnitude and state are categorical variables, thus a bar-chart served as the most appropriate means of displaying the data (shown in section 2.2 and figure 1.0). From observation it can be seen that on average the state of "AR" has the highest magnitude tornado's. On the other hand, the states of "VI" and "DC" has the lowest magnitude tornado's. This could indicate that these respective states are more likely to be hit with tornado's of low magnitude, whilst a state like "AR" are more likely to be hit with tornado's of a greater magnitude due to the topology and geography. The graph itself has no shape,location or spread, as the x-axis contains categorical qualitative data, which has not unified order, so the shape of the graph will change from instance to instance. In addition, no definite outlier can be found, as all values hover around the mean value, which will be discussed furthr in the standard deviation section

below. However, it can be inferred that the magnitude of 0 states may be outliers, as they have no values, which could indicate a hole in the gathering proccess, however this cannot be proved. As shown in section 2.3, the graph's calculated mean is 0.691, with a standard deviation of 0.31 and median of 0.804. This means that on average across all states, the magnitude of a tornado is most likely going to be around 0.691. The median of 0.804 demonstrates that the body of the magnitudes lie towards the 0 mark rather than the 5 mark,for context, an even distribution would be more likely to see a median around 3. Next, the standard deviation calculated was 0.31. Standard Deviation refers to the the spread of variation of the data from the mean value. In this scenario, the standard deviation value is close to the mean value, indicating a low level of variance between the data points and the mean. Therefore, it can be inferred that the magnitude's are bound to be consistent from state to state, a factor that can also be observed via the graph. Lastly, the 95% confidence interval was calculated to be between 0.6051224-0.7758915. This means that with 95% confidence it can be claimed that the population mean Magnitude given a set of tornados is between the points 0.605 and 0.776. So if this were to be repeated on a new set of data with a similar mean, we can assume that the mean Magnitude would lie between the magnitudes of 0.605 and 0.776.

Secondly, a comparison was made between the magnitude of a tornado and its influence on the number of injuries to grasp an understanding on the severity of each class of magnitude. As magnitude is a categorical variable, whilst injuries is a quantitative variable, a box plot was selected to demonstrate the relationship between the two(shown in section 2.4, figure 2.0). When observing the mean values of the box plot, it can be noticed that an exponential growth in the number of injuries is present, as the magnitude of the tornado's increases. From observation, the magnitude 5 tornado has the greatest average number of injuries, hovering around the 400 mark. Whilst the magnitude 0 and 1 tornado's had the minimum number of injuries, with both hovering around the 0-5 on average. From observation it is evident that the data set has a a range of outliers, the main body of which lies around the magnitudes of 0-2, whilst severely decreasing between the magnitudes of 3-5 as shown by the dots above and below the box plots. This could be due to a variety of reason, one of which is the number of values in each magnitude category. To expand on this, as seen in the bar graph previously, a large body of the data lies in the magnitudes between 0-2, such that these categories are more prone to errors, the more data, the more opportunities of errors to arise. On the other hand, the magnitudes of 3-5 have less data points, thus less opportunity for errors to arise. Moreover, a standard deviation of 18.214 was calculated alongside the box plot, which in context is very high. The standard deviation values represents the distance of the data points from the mean of the data. A high standard deviation indicates a larger spread of data and a lower level of confidence. The high standard deviation may be a result of the large number of outliers found in the box plots. A large number of outliers are likely to cause large fluctuations in numbers, thus increasing the distance between data points and the mean, inherently increasing the standard deviation of the data. However, this does not rule out the influence of the box plot graphed. It is possible that these data values will make sense in relation to other data. Suppose an example where a low magnitude tornado's in a highly populated area compared to a high magnitude tornado in a low density area. This would cause the low magnitude tornado to have a greater injury rate than the high magnitude tornado.Lastly, the 95% confidence interval for the number of injuries given a random set of tornados lies between the values of 1.297354-1.571260. Therfore, it can be claimed with 95% confidence that the true population mean of the number of injuries lies between the points 1.297354-1.571260. So if the test were to be done with a subsection of the data, their is a 95% chance that the population mean lies in between 1.297354-1.571260 injured people.

Following on from the high standard deviation of the injuries and the hypothesis that was established, a comparison was made between the amount of property loss (in dollars) against the number of injuries. Cost of property loss was chosen to validity the hypothesis that the number of injuries is influenced by the population density. Property loss will give an indication to whether the area impacted by the tornado was a high or low density area, as a high property loss would indicate that tornado hit a high populated area, whilst a low property loss would indicate that a tornado hit a low populated area. For the hypothesis to be true, we would expect the number of injuries to linearly increase as the property loss increased. A scatter plot was chosen for this comparison as the two variables that are being compared are both numerical values, and a scatterplot would be able to appropriately graph the two (shown in figure 3.0 and section 2.7). From the output of the scatterplot, it can be seen that on average the number of injuries is evenly distributed among the amount of damage done, with a high concentration of the data lying on the left side of the graph and

the bottom. By viewing the peaks and generating a trendline, it can be inferred that the scatter plot was bi-modal with a mean of 117976.9. The standard deviation of loss was 9034341, considering the mean was 117976.9 it can be inferred that this is a very high standard deviation value.This indicates a large distribution of data, which is reflected in the scatterplot. The data points of the left of the graph demonstrate that in low density areas it is possible to have a large number of injuries, and the more dense areas, have similar injuries rates as the low density areas. Therefore it can be assumed that the hypothesis established prior which suggested a connection between population density and the number of injuries to be false. Lastly, a 95% confidence interval for the amount of property damage given a tornado was calculated to lie between 50045.03-185908.77. Therefore it can be claimed that with 95% confidence, the true population mean of a the amount of property damage is given between $50045.03 and 185908.77.If the test were to be repeated with a subsection of the data, the population mean of the value would lie in between these two points, 95% of times.

Next, a comparison was made between the number of tornado's and the month they occurred. A histogram was chosen for this task as histograms are able to effectively communicate the relationships between a categorical and a quantitative variable. The graph is slightly right skewed with the maximum on the 5th Month (May). This suggests that tornado's are most likely to occur in the month of May. On the the hand, the minimum of the graph lies in January, thus it can be inferred that tornados are least likely to occur in January. The shape of the graph is a uni modal with a mean at 5.969166 and a standard deviation of 2.449539.Thus, when compared with the mean, it can be inferred that the months column is tightly distributed, meaning the data can be used with a high level of confidence. From the graph alone, no outliers are evident.The unimodal nature of the graph indicates that the number of tornado's in a given year only increases once, thus suggesting that this is a seasonal occurrence. This hypothesis aligns with what it known on how tornado's form and the influence of weather and temperature. To conclude, a 95% confidence interval was calculated for the month of a given set of tornado to be in between 5.950747-5.987585. It can be inferred with 95% confidence that the true population mean lies in between 5.950747-5.987585 (May-June). If the tests were to be redone with new set of data or a subsection of the current data, the population mean of the number of months would lie between 5.950747-5.98758, 95% of the time.

Lastly, the relationship between the cost of damage against the magnitude of the tornado was calculated. Due to the nature of cost being very large in addition to the categorical and quantitative nature of the variable, a line graph was chosen to provide the greatest amount of clarity. The graph is left skewed with the maximum being at the magnitude of 3. Whilst the magnitude of 1 and 2 have the least. The graph is unimodal with mean at 3. The standard deviation of magnitude is 0.895758, indicating that the data is tightly bound. In relation to the mean and the standard deviation, it can be inferred that the graph has a high level of certainty. Following on from this, the cost related to a magnitude of 5 is unexpected, it was expected that the cost of a magnitude 5 tornado would be greater than a magnitude 4 tornado, however this was not the case in the diagram. This could be due to a variety of factors. The current hypothesis is that the data has not provided enough sample of magnitude 5 tornado for it to be represented in the dataset.Due to this the graph is unimodal, whilst it was expected to be exponential. However, it could also be assumed to be a collection of outlier in the data. Finally, the 95% confidence interval was calculated for the magnitude of a tornado set to be in between 1.771953-1.785424. Therefore it can be claimed be 95% confidence that the true population mean, in regards to magnitude, lies in between 1.771953-1.785424. If the tests were to be repeated, their would be a 95% chance that the population mean would lie somewhere in between these points.

Modelling

Following on from this a series of predictive models were constructed to predict a variety of useful information. The predictive model that were used include: Straight line prediction models (linear and multiple regression) followed by curve line prediction models(Logistic regression, Poisson Regression and Regression Trees).

To begin, a linear regression predictive model was constructed to test the relationship between loss and injury. Linear regression models allow for predicting one variable by using another variable through a linear equation. The variable loss and injury were chosen as they would allow scientist to determine the amount of destruction

caused by a tornado directly after it occurred using the number of injuries reported to hospitals without endangering themselves. Section 3.0 represents the output of this linear regression, which has calculated the equation y=21236x+87518, in which y represents the price of damage and x represents the number of injuries. Suppose the example where a journalist is wanting to write an article directly after a tornado on it's effects. The journalist identifies 15 injuries, thus the amount of loss would be y=21236*15+87518, totaling $406,058 of damage. However, the validity of the model must be tested via the 4 assumption before practical use. The four assumptions that determine the validity of a linear model is Linearity, Homoscedasticity, Normality and Independence. To begin, Linearity is determines whether a linear function was the best option for the combination of variables. The linearity test conducted in section 3.1 shows a roughly straight horizontal trendline, confirming the linearity of the model. The trendline does curve downwards towards the end, however these values are a large distance away from the main body of data, and thus can be assumed to be outliers. Homoscedasticity tests refers to the variance of noise terms throughout the the dataset. When tested in section 3.2, the graph generated a positive trendline with a slight bump on the far left. However, a correct homoscedasticity tests should have a straight horizontal line throughout the entire graph, thus this does not meet the requirement of homoscedasticity. It can be inferred that this was caused by the concentration of errors on the right hand side of the graph, indicating a larger proportion of errors are concentrated towards the larger values of injuries column. Thirdly, the normality of linear models determines whether the noise terms are normally distributed. Normality is tested in section 3.3 which shows a collection of data that for the majority lies on the dotted line, thus meeting the requirements for normality. A few data points on the left and right of the graph deviate from the dotted line, however these do not influence the normality of the data as the bulk of the data still lies on the dotted lines. It can be inferred that these values that deviate from the dotted lines are examples of outliers. Lastly, independence refers to if the error terms are independent. Independence cannot be demonstrated by a graph, thus the question "Could the observations from one subject somehow give us more information about the other observations?" must be used to determine the independence. In this instance, the number of injuries do not give us information about the cost of damage, thus the answer to the question is no. Therefore it can be inferred that the linear regression model is independent.From analysis of these 4 assumption, it can be inferred that the linear regression model is mostly valid model, however as it did not pass to homoscedastictity test, it cannot be used in a practical sense.

Similar to linear regression, multiple regression attempts to build a relationship between variables. Unlike linear regression, multiple regression takes a series of independent variables. The multiple regression built in this assignment attempts to predict the number of injuries from the independent variables, magnitude,state,length of tornado and width of tornado (shown in section 3.4). Following this, the Anova() function was run to determine the P-Values of the independent variables as shown in section 3.5. From observation, it is evident that none of the P-Values for the independent variables fall under the minimum of 0.05. Therefore it is inferred that all the independent variables chosen were valid and useful. This model could be used in practice to prepare hospitals with the right amount of equipment for an incoming tornado. Similar to the linear regression model, the 4 assumptions were tested on the multiple regression to determines it's validity. To begin, the linearity of the model was tested, to determine whether the relationship between the dependent and independent variables are linear, as multiple regression is best suited for linear relationships. As shown in section 3.6, the data points of the model follow a horizontal straight line, and thus meet the requirements of linearity. Secondly the homoscedasticity was tested, to determine whether the noise terms all have the same variance.Homoscedasticity was tested in section 3.7, in which the trendline contained a strongly positive relationship, almost linear. However for the data to be claimed to be homoscedasticity, it must be a straight horizontal line, which this model does not contain. Thus this model can not be claimed to meet the requirement for homoscedasticity. Following on from this, the normality of the model was calculated to determine whether the noise terms are normally distributed. This was done in section 3.8, which shows the majority of the data lying on the grey dotted line.Thus, the multiple regression models meet the requirements of normality. Lastly, independence was tested through the question "Could the observations from one subject somehow give us more information about the other observations?" to determine whether the error terms were independent. It can be inferred that the number of injuries do not give us more information on the independent variable, and thus meeting the requirements for independence. As the model did not meet all 4 requirements of validity, it cannot be used in a practical sense. However, it

is thought that the influence of errors has not hindered the model greatly. Thus, it can be inferred that the model can still be used to get approximations of the number of injured people after a tornado.

Next, a 3 types of curve line prediction models were built, these 3 were Logistic regression, Poisson Regression and Regression Trees.

To begin, the poisson regression model, is a tool that often provides model comparing count variables to a series of independent variables.For this scenario, it was chosen for the price of damage (loss) to be the response variable, and the predictors to be length,width,magnitude and state.Using this model, a prediction can be made of the average cost of damage, depending on the length,width,magnitude of the tornado and state of where it has occurred. Loss was chosen as the response variable for this model as it would be able to provide greater accuracy than the linear model build prior. However, this does not devalue the linear model, the linear model will provide results quicker, and does not require as many predictor variables. As the response variable for poisson regression models are required to be positive, whole number, the loss values were rounded to the nearest whole number.The output of this model is shown in section 3.10. It is common practice for any "PR" values greater than 0.05 to be removed. However,in this instance, the only variables greater than 0.05 is a small number of states that do not greatly influence the final model, it was decided that they were not going to be removed. This model is now able to efficently calculte a approximate cost of damage given length,width,magnitude and state.

Following on from this, a regression tree was built which incorporates all the variables in the dataset and attempts to calculate the magnitude based on the values of the other variables. This model is displayed through a tree diagram, as shown in 3.11. A regression tree attempts to model a relationship between a response variable and a series of predictor variables from the dataset. The appropriate predictor variables are chosen by the program, and will only select the most influential variables that impact the end result. From the tree generated, it can be observed that the influential predictor variables are the number of injured, price of damage(loss),width,length and date. Using these variables, this graph is able to provide a rough estimate of the magnitude of the tornado. Further inspection of the graph, however revealed that the magnitude values only reach 3.9, which can be assumed to be 4. Therefore, magnitude of 5 tornado's will never be calculated in this model. This error may have arisen due to the number of outlier values, or a miscalculation that happened in the generation process. Furthermore, in section 3.13 the most valuable predictor variable was calculated, from the graph, it is evident that the injury variable is the most important. Therfore, it can be inferred that the injury column has the greatest relationship to the magnitude of a tornado.Moreover in section 3.14, the regression tree is tested by the program.From the results which were outputted, it can be inferred that the regression tree is a viable model for practical use.

Logistic regression model attempt to predict a binary outcome, given a number of predictor variable. For this report, a logistic regression model was built to predict whether the number of injuries would be greater than 5, given the magnitude of the tornado, the state in which it occurred, cost of damage, width of tornado and length of tornado (shown in section 3.15). When the output of the model was observed, it was found that the state, for a majority, and price of damage possessed P-Values greater than 0.05, indicating that they were less significant, therefore the decision was made to remove States and loss altogether and rerun the model,as shown in section 3.16.From inspection of the output, the values seemed to be correct, and all p-values were less than 0.05, thus it was assumed that the logistic regression was correct and complete. To test the validity of the model, a combination of the predict() function and roc curves were used. Using this, a graph was generated, as shown in section 3.17.From observation, the graph met expectation. Just guessing whether their would be more than 5 injuries, would be represented by the 45degree dotted line, and the trendline which represents the logistic regression, is much more accurate, represented by the prominent asymptotic shape. Therefore, it can be inferred that the logistic regression was successful and a valid model. This model could be used in a real world context by changing the number of injuries amount and calculating the probability of injuries, better preparing hospitals.

Reveiw

In Recollection, a variety of different factors and information were learnt throughout the undertaking of this assignment in regards to Tornado's and how to predict a variety of different factors from it. To begin,

via the bar chart it was observed that different states in America endure different scales of tornado's, with some enduring on average magnitude 1 tornado's, whilst other enduring none. From what is known about tornado's, it can be inferred that the topology, geography and weather determine whether a tornado is formed. Following on from this, through the box plot and the scatter plot, it was observed that the magnitude of a tornado influences the injuries, whilst the price in damage does not. In addition, to observing that the month of may had the most number of tornado's, which reinforces the ideas established prior that the weather has an influence on the creation of tornado.Furthermore, it was discovered that magnitude 3 tornado's had the greatest amount of damage on property. From the graph's onwards, the amount of new information on tornado's was limited as the graphs often do most of the heavy lifting. However, it was identified that the linear relationship between price of damage and number of injuries is y=21236x+87518. In addition to learning that injury has the greatest influence on the magnitude of a tornado.

In consideration of the all the data presented and the model constructed, a large number of observation were made.To begin, it was evident that throughout all of the models and graphs, that a large number of errors and outliers were present within the data set. These errors and outliers have negatively influenced the graphs and models created. In construction of the models and graphs,the severity of the errors and outliers were not evident,however the impact of the errors and outlier were clear when the analysis of each graph and model came into fruition.If this report were to be done again, the outlier and the errors would be removed prior to the construction of the graphs and models. Moreover, moving forward, graphs and models will be analysed directly after they have been built, stopping any errors from bleeding into other models.Next,it was evident whilst analyzing the data that although the graph had a large number of different data and data types, it was found that only a few of the data was actually usable in analysis and the construction of models. The result of this is multiple models and graphs using similar, sometimes identical, predictor and response variables. Moving forwards, datasets containing a larger variety of data will be selected, in addition to better selecting predictor and response variable when model and graphs are constructed.

Conclusion

In conclusion, this report attempts to analyse and interpret the data from the Tornado's dataset, provided by the NOAA's National Weather Service Storm Prediction Center. The dataset consists of a series of variables which inform about the tornado itself and it's after affects on the residents it hit.Before any analysis could be conducted, the dataset had to be cleaned, by reallocating data types to meet the requirements of the graphs and models being built. In addition to removing unused or non-complete columns and removing evident outliers. Following the cleaning, 5 graphs were built to analyse and demonstrate the relationships between variables.The bar graph compared the states to the average magnitude of the tornado's in that state. From this, it could be understood that certain states are more likely to have higher magnitude tornado than other, whilst others are likely to have a lower magnitude. Thus it can be inferred that the typography and geography of a state influences the magnitude of the tornado's it endures.Next, a bar chart demonstrated the relationship between each magnitude and the number of injuries it causes. This graph demonstrated a exponential relationship between the two variable. Thus, it can be inferred that the the greater the magnitude of a tornado, the greater chance of civilians being injured.Following on from this, a scatter plot was drafted up comparing the price of damage to the number of injuries, to test the hypothesis that more dense areas have a greater number of casualties. This was done by representing price of damage as highly dense areas, as it is common sense that largely dense areas hit by a tornado's are more likely to have a high price of damage. However from inspection of the graph outputted, no such conclusion can be made due to the large spread of the data throughout the entire graph. This infers that the density of the area does not influence the number of injured civilians.Furthermore, a histogram was built to determine which months are more likely to have tornado's. The output was a rightly skewed unimodal dataset which contained a maximum at month=5.Therfore it can be inferred that tornados are most likely to occur in the month of May. Lastly, a line graph was generated to determine wheather the magnitude of the tornados influences the amount of damage it cause to properties. The graph was a unimodal left skewed set of data, which had a maximum at magnitude 4. Surprisingly, tornados of magnitude 5 do not have greater amount of damage on property as a tornado with magnitude 4. It could be inferred that magnitude 5 tornado are more likely to occur is deserted area, thus impacting less properties. However, this could also be due to the little

number of magnitude 5 tornado's in the dataset.Proceeding with the information gathered from the graphs, a series of models were built to predict various different factors in a real world setting, begging with a linear regression. The linear regression model builds a relationship between the response variable,cost of damage, and the predictor variables, number of injuries. Using the model the linear equation y=21236x+87518 was derived, which could be used in a real world context to make quick calculation of the total damage costs based off the number of injuries reported to the nearby hospitals. Next a multiple linear regression was built to compare the response variable,number of injuries, to the predictor variables,magnitude,state,length and width. Using this information, and the predict() function, the number of injuries can be calculated, using the predictor variable, to better prepare hospitals.Following on from this,a poisson regression model was built to determine price of damage depending on the length,width,magnitude and the state in which the tornado occurred. From observation of the p-Values and the estimates, it can be assumed that this model is valid and provides a clear and accurate way to calculate an estimate of the amount of damage a town or state sustains from a tornado. Following on from this, a regression tree was built to calculate the magnitude of a tornado. The graph that was generated provided a easy way for the average consumer to calculate the magnitude of a tornado. In addition to this, it was shown through tests, that magnitude has the greatest relationship with injury, thus it can be inferred that injury influences the calculation of magnitude. Lastly, a logistic regression was built which allows for prediction on whether their would be greater than 5 injuries. This model could be used to calculate to better prepare hospitals for incoming tornado's. The confidence of this model was calculated via a ROC curve, which followed a asymptotic trendline as expected. Throughout the duration of this report, a series of errors negatively influenced the final result of the models and the graphs built. The biggest of these errors were the number of outliers in the data. Retrospectively, it should have been assumed that in a data of greater than 65,000 rows, their would be an abundance of outliers. These outliers caused graphs and models to be miscalculated, however it can be inferred that the deviation from the true values were minor, as the number of outliers were small. Moving forward, outliers will be removed in the cleaning step of analysis, so that it does not influence the final result.

```
## Section 1.0-Loading in the required R Libraries and data
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(rsample)
library(vip)
```

```
##
## Attaching package: 'vip'
##
## The following object is masked from 'package:utils':
##
##     vi
```

```
library(skimr)
library(parsnip)
library(rpart)
library(broom)
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.1.1 --
## v dials        1.2.0     v tune         1.1.2
## v infer        1.0.5     v workflows    1.1.3
## v modeldata    1.2.0     v workflowsets 1.0.1
## v recipes      1.0.8     v yardstick    1.2.0
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x dials::prune()    masks rpart::prune()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(knitr)
tornados<-read_csv("https://www.spc.noaa.gov/wcm/data/1950-2022_actual_tornadoes.csv")
```

```
## Rows: 68701 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): mo, dy, st, stf
## dbl  (23): om, yr, tz, stn, mag, inj, fat, loss, closs, slat, slon, elat, el...
## date  (1): date
## time  (1): time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(tornados)
```

```
## # A tibble: 6 x 29
##      om    yr mo    dy    date       time      tz st    stf     stn   mag   inj
##   <dbl> <dbl> <chr> <chr> <date>     <time> <dbl> <chr> <chr> <dbl> <dbl> <dbl>
## 1   192  1950 10    01    1950-10-01 21:00      3 OK    40       23     1     0
## 2   193  1950 10    09    1950-10-09 02:15      3 NC    37        9     3     3
## 3   195  1950 11    20    1950-11-20 02:20      3 KY    21        1     2     0
## 4   196  1950 11    20    1950-11-20 04:00      3 KY    21        2     1     0
## 5   197  1950 11    20    1950-11-20 07:30      3 MS    28       14     1     3
## 6   194  1950 11    04    1950-11-04 17:00      3 PA    42        5     3     1
## # i 17 more variables: fat <dbl>, loss <dbl>, closs <dbl>, slat <dbl>,
## #   slon <dbl>, elat <dbl>, elon <dbl>, len <dbl>, wid <dbl>, ns <dbl>,
## #   sn <dbl>, sg <dbl>, f1 <dbl>, f2 <dbl>, f3 <dbl>, f4 <dbl>, fc <dbl>
```

```
## Section 1.1-Establishing Column Data Types
tornados$om<-as.integer(tornados$om)
tornados$yr<-as.factor(tornados$yr)
tornados$mo<-as.integer(tornados$mo)
tornados$dy<-as.integer(tornados$dy)
tornados$tz<-as.integer(tornados$tz)
tornados$st<-as.factor(tornados$st)
tornados$stf<-as.integer(tornados$stf)
tornados$stn<-as.integer(tornados$stn)
tornados$mag<-as.integer(tornados$mag)
tornados$inj<-as.integer(tornados$inj)
tornados$fat<-as.integer(tornados$fat)
tornados$ns<-as.integer(tornados$ns)
tornados$sn<-as.integer(tornados$sn)
tornados$sg<-as.integer(tornados$sg)
tornados$f1<-as.integer(tornados$f1)
tornados$f2<-as.integer(tornados$f2)
tornados$f3<-as.integer(tornados$f3)
tornados$f4<-as.integer(tornados$f4)
tornados$fc<-as.integer(tornados$fc)
head(tornados)
```

```
## # A tibble: 6 x 29
##      om yr       mo    dy date       time      tz st      stf   stn   mag   inj
##   <int> <fct> <int> <int> <date>     <time> <int> <fct> <int> <int> <int> <int>
## 1   192 1950     10     1 1950-10-01 21:00      3 OK       40    23     1     0
## 2   193 1950     10     9 1950-10-09 02:15      3 NC       37     9     3     3
## 3   195 1950     11    20 1950-11-20 02:20      3 KY       21     1     2     0
## 4   196 1950     11    20 1950-11-20 04:00      3 KY       21     2     1     0
## 5   197 1950     11    20 1950-11-20 07:30      3 MS       28    14     1     3
## 6   194 1950     11     4 1950-11-04 17:00      3 PA       42     5     3     1
## # i 17 more variables: fat <int>, loss <dbl>, closs <dbl>, slat <dbl>,
## #   slon <dbl>, elat <dbl>, elon <dbl>, len <dbl>, wid <dbl>, ns <int>,
## #   sn <int>, sg <int>, f1 <int>, f2 <int>, f3 <int>, f4 <int>, fc <int>
```

```
## Section 1.2-Skimming Data to observe outliers or errors in the data
## NOTE: SKIMS HAD TO BE COMMENTED OUT DUE TO ERROS WHEN RENDERING USING KNIT
##skim(tornados)
```

```
## Section 1.3- Removing "sg" as the column is not viable for analysis
tornados<-tornados%>%select(-sg)
head(tornados)
```

```
## # A tibble: 6 x 28
##      om yr       mo    dy date       time      tz st      stf   stn   mag   inj
##   <int> <fct> <int> <int> <date>     <time> <int> <fct> <int> <int> <int> <int>
## 1   192 1950     10     1 1950-10-01 21:00      3 OK       40    23     1     0
## 2   193 1950     10     9 1950-10-09 02:15      3 NC       37     9     3     3
## 3   195 1950     11    20 1950-11-20 02:20      3 KY       21     1     2     0
## 4   196 1950     11    20 1950-11-20 04:00      3 KY       21     2     1     0
## 5   197 1950     11    20 1950-11-20 07:30      3 MS       28    14     1     3
## 6   194 1950     11     4 1950-11-04 17:00      3 PA       42     5     3     1
## # i 16 more variables: fat <int>, loss <dbl>, closs <dbl>, slat <dbl>,
```

```
## #   slon <dbl>, elat <dbl>, elon <dbl>, len <dbl>, wid <dbl>, ns <int>,
## #   sn <int>, f1 <int>, f2 <int>, f3 <int>, f4 <int>, fc <int>
```

```
## Section 1.4-"STN" and "Closs" were removed, due to being referenced as "discontinued"
## by the source (github)
tornados<-tornados%>%select(-stn)
tornados<-tornados%>%select(-closs)
head(tornados)
```

```
## # A tibble: 6 x 26
##      om yr       mo    dy date       time      tz st      stf   mag   inj   fat
##   <int> <fct> <int> <int> <date>     <time> <int> <fct> <int> <int> <int> <int>
## 1   192 1950     10     1 1950-10-01 21:00      3 OK       40     1     0     0
## 2   193 1950     10     9 1950-10-09 02:15      3 NC       37     3     3     0
## 3   195 1950     11    20 1950-11-20 02:20      3 KY       21     2     0     0
## 4   196 1950     11    20 1950-11-20 04:00      3 KY       21     1     0     0
## 5   197 1950     11    20 1950-11-20 07:30      3 MS       28     1     3     0
## 6   194 1950     11     4 1950-11-04 17:00      3 PA       42     3     1     0
## # i 14 more variables: loss <dbl>, slat <dbl>, slon <dbl>, elat <dbl>,
## #   elon <dbl>, len <dbl>, wid <dbl>, ns <int>, sn <int>, f1 <int>, f2 <int>,
## #   f3 <int>, f4 <int>, fc <int>
```

```
## Section 1.5-Reskimming data to find any missed errors, and to check if data is fully cleaned
## NOTE: SKIMS HAD TO BE COMMENTED OUT DUE TO ERROS WHEN RENDERING USING KNIT
##skim(tornados)
```

```
## Section 1.6- From the skim it can be seen that the mag has a negative value and the rest
##seem to lie around the 1 mark. By looking through the data it seems to indicate that -9
##is an outlier. This could made it to two different values, either NA or 0. But NA was
##chosen as we don't know for sure what the value could have been.
tornados<-tornados[tornados$mag!=-9,]
## NOTE: SKIMS HAD TO BE COMMENTED OUT DUE TO ERROS WHEN RENDERING USING KNIT
##skim(tornados)
```

```
## Section 1.7-Final Cleaned Dataset.
head(tornados)
```

```
## # A tibble: 6 x 26
##      om yr       mo    dy date       time      tz st      stf   mag   inj   fat
##   <int> <fct> <int> <int> <date>     <time> <int> <fct> <int> <int> <int> <int>
## 1   192 1950     10     1 1950-10-01 21:00      3 OK       40     1     0     0
## 2   193 1950     10     9 1950-10-09 02:15      3 NC       37     3     3     0
## 3   195 1950     11    20 1950-11-20 02:20      3 KY       21     2     0     0
## 4   196 1950     11    20 1950-11-20 04:00      3 KY       21     1     0     0
## 5   197 1950     11    20 1950-11-20 07:30      3 MS       28     1     3     0
## 6   194 1950     11     4 1950-11-04 17:00      3 PA       42     3     1     0
## # i 14 more variables: loss <dbl>, slat <dbl>, slon <dbl>, elat <dbl>,
## #   elon <dbl>, len <dbl>, wid <dbl>, ns <int>, sn <int>, f1 <int>, f2 <int>,
## #   f3 <int>, f4 <int>, fc <int>
```

```
## Section 2.1- Building another table with the mean magnitude of tornado's per state in America.
tornados_meanMag<-tornados%>%group_by(st)%>%summarise(meanMag=mean(mag,na.rm = TRUE))
tornados_meanMag
```

```
## # A tibble: 53 x 2
##    st    meanMag
##    <fct>   <dbl>
##  1 AK      0
##  2 AL      1.02
##  3 AR      1.13
##  4 AZ      0.413
##  5 CA      0.383
##  6 CO      0.398
##  7 CT      0.992
##  8 DC      0
##  9 DE      0.884
## 10 FL      0.511
## # i 43 more rows
```

```
## Section 2.2-Bar graph comparing each state in America against it's respective mean magnitude.
tornados$mag<-as.factor(tornados$mag)
ggplot(tornados_meanMag,aes(x=st,y=meanMag))+geom_bar(stat="identity")+
  labs(title="States In America against their respective mean magnitude of Tornado's")+
  theme(axis.text.x = element_text(angle=90) )+
  annotate("text",x=Inf,y=-Inf,
           label="Figure 1.0-Bar graph comparing each state in America against it's respective mean magn
```

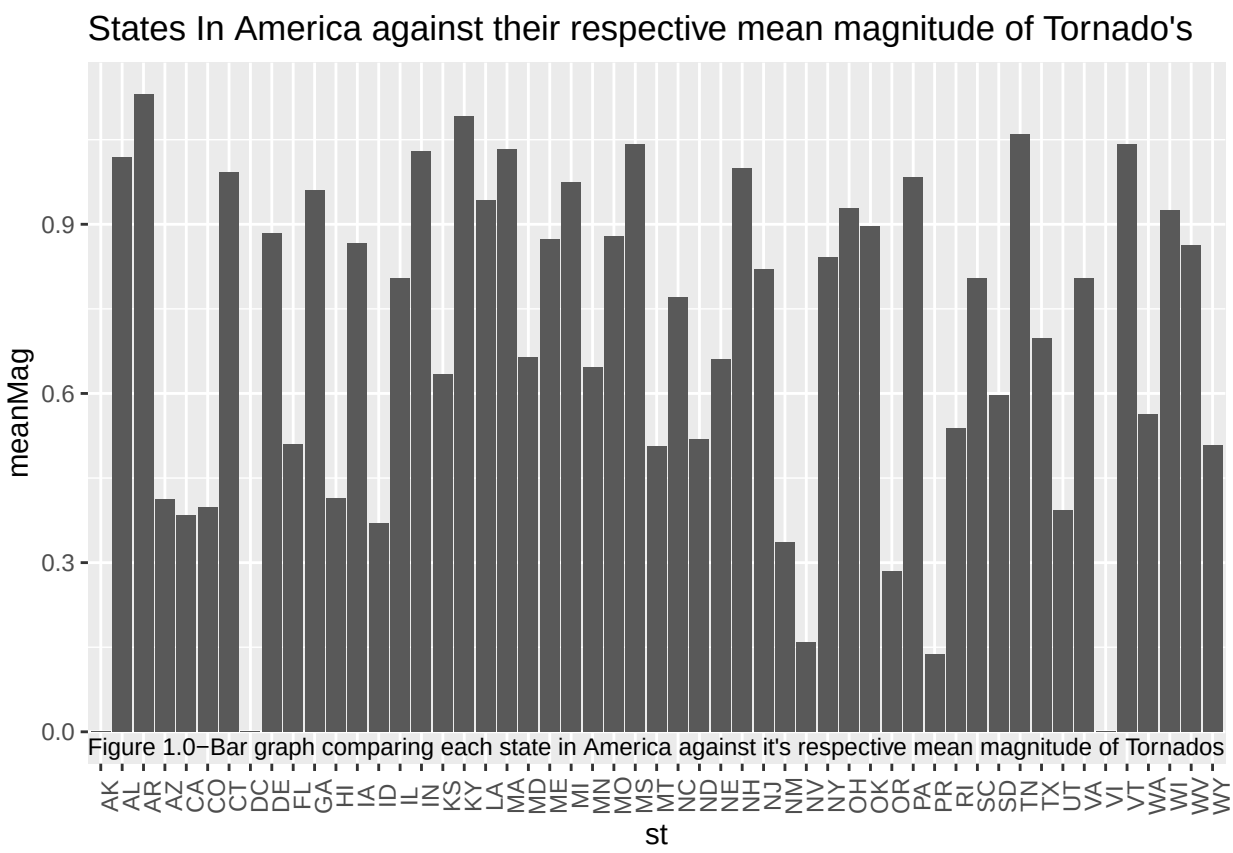


Figure 1.0−Bar graph comparing each state in America against it's respective mean magnitude of Tornados

```
## Section 2.3- Mean,Median and SD of the mean magnitude values.
mag_state_mean<-mean(tornados_meanMag$meanMag)
```

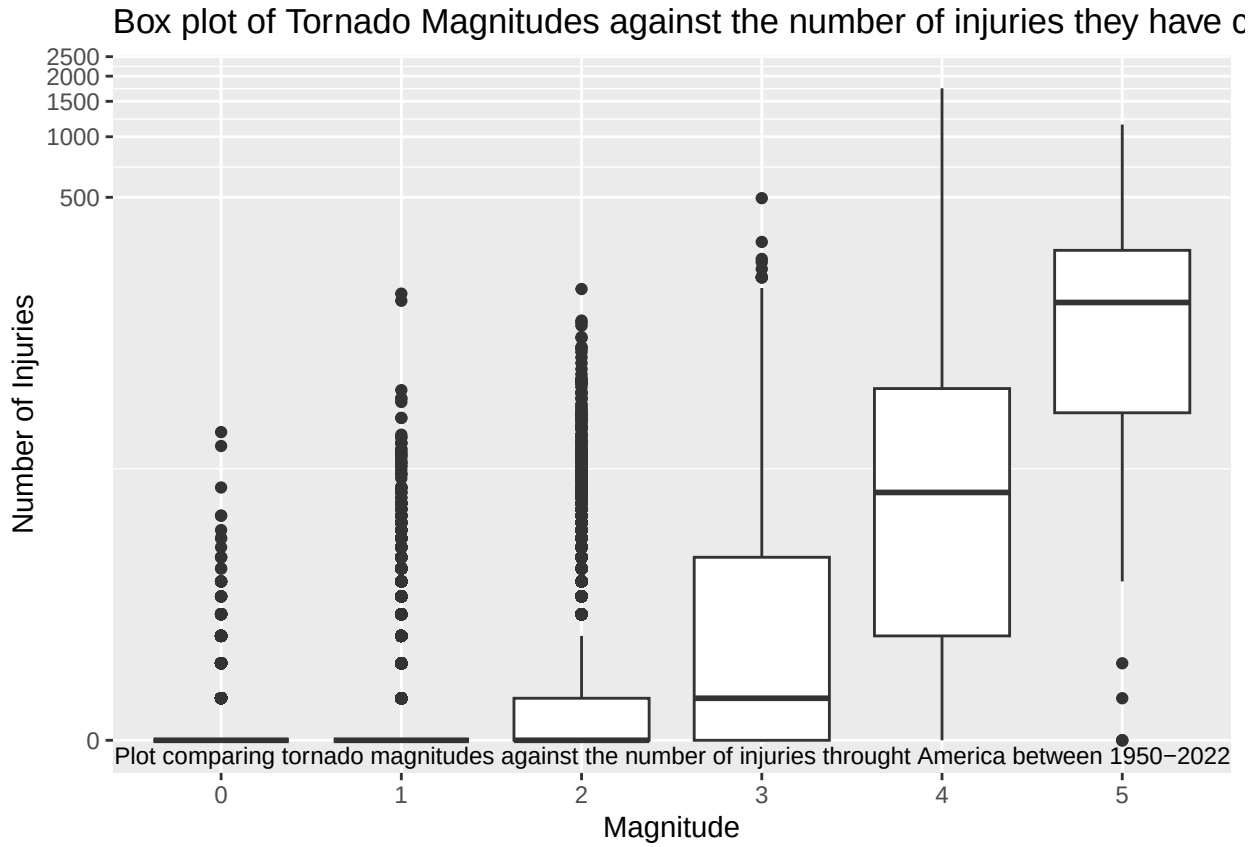
```r
mag_state_sd<-sd(tornados_meanMag$meanMag)
mag_state_median<-median(tornados_meanMag$meanMag)
mag_state_length<-length(tornados_meanMag$meanMag)
mag_state_summary<-
  data.frame(Observations=c("Mean","SD","Median","Length"),
             Values=c(mag_state_mean,mag_state_sd,mag_state_median,mag_state_length))
mag_state_summary
```

```
##   Observations     Values
## 1         Mean  0.6905070
## 2           SD  0.3097749
## 3       Median  0.8042142
## 4       Length 53.0000000
```

```r
## From Week 5 Modules:
t <- qt(p = 0.025, df = mag_state_length-1, lower.tail = FALSE)
lwr <- mag_state_mean - t * mag_state_sd / sqrt(mag_state_length)
upr <- mag_state_mean + t * mag_state_sd / sqrt(mag_state_length)
ci <- c(lwr = lwr, upr = upr)
ci
```

```
##       lwr       upr
## 0.6051224 0.7758915
```

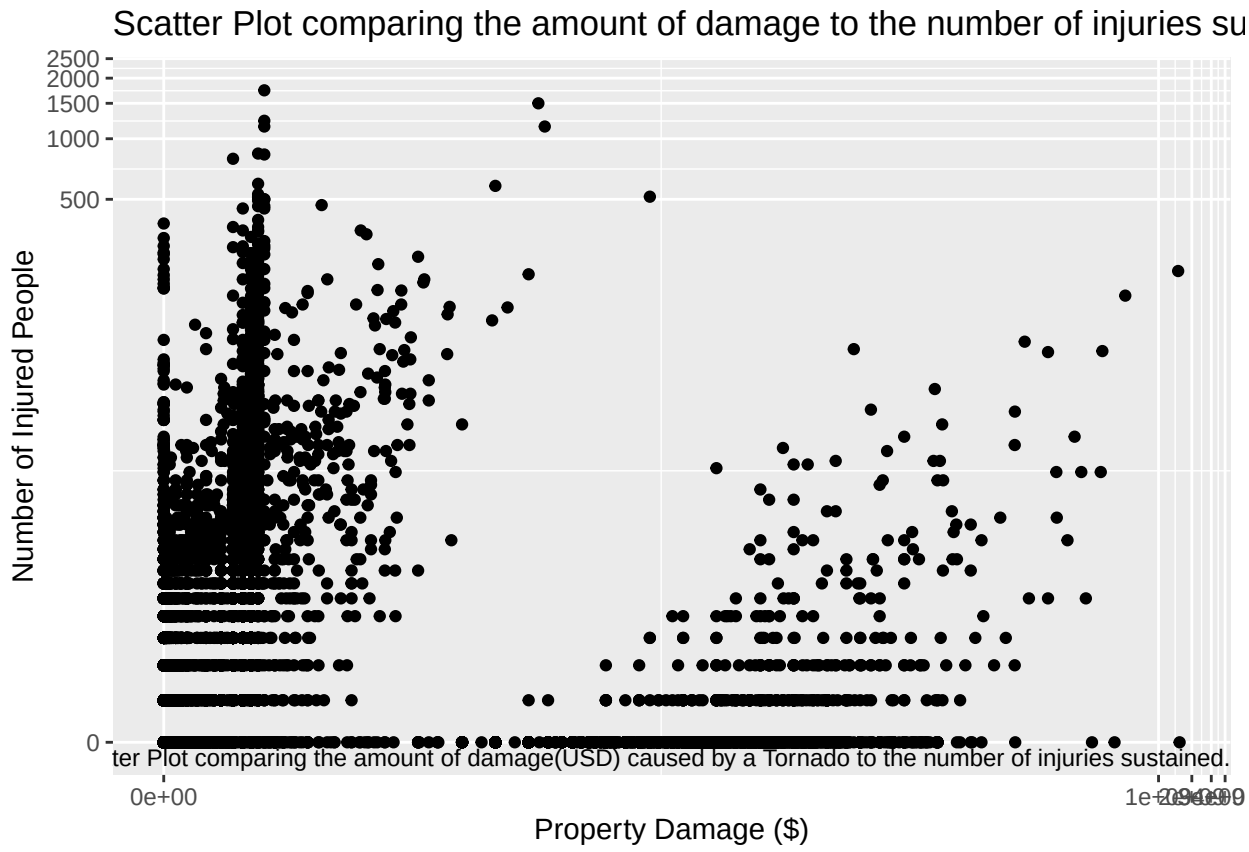
```r
## Section 2.4-Box Plot comparing each stage of magnitude values to the number of injuries they have ca
ggplot(tornados,aes(mag,inj))+geom_boxplot()+
  labs(title="Box plot of Tornado Magnitudes against the number of injuries they have caused.")+scale_y
  annotate("text",x=Inf,y=-Inf,
           label="Figure 2.0- Box Plot comparing tornado magnitudes against the number of injuries throu
  labs(x="Magnitude",y="Number of Injuries")
```




```
## Section 2.5-Calculating the Standard Deviation of the number of injuries
inj_sd<-sd(tornados$inj)
## Section 2.6-Calculating the mean number of injuries sustained in a tornado
inj_mean<-mean(tornados$inj)
inj_length<-length(tornados$inj)
t_sd <- qt(p = 0.025, df = inj_length-1, lower.tail = FALSE)
lwr <- inj_mean - t_sd * inj_sd / sqrt(inj_length)
upr <- inj_mean + t_sd * inj_sd / sqrt(inj_length)
ci_length <- c(lwr = lwr, upr = upr)
ci_length
```

```
##      lwr      upr
## 1.297354 1.571260
```

```
## Section 2.7-Scatter Plot comparing the amount of damage(USD) to the number of injuries sustained
ggplot(tornados,aes(x=loss,y=inj))+geom_point()+
  labs(title="Scatter Plot comparing the amount of damage to the number of injuries sustained")+
  scale_x_continuous(trans=scales::pseudo_log_trans(base=2))+
  scale_y_continuous(trans=scales::pseudo_log_trans(base=10))+
  annotate("text",x=Inf,y=-Inf,
           label="Figure 3.0- Scatter Plot comparing the amount of damage(USD) caused by a Tornado to th
  labs(x="Property Damage ($)",y="Number of Injured People")
```

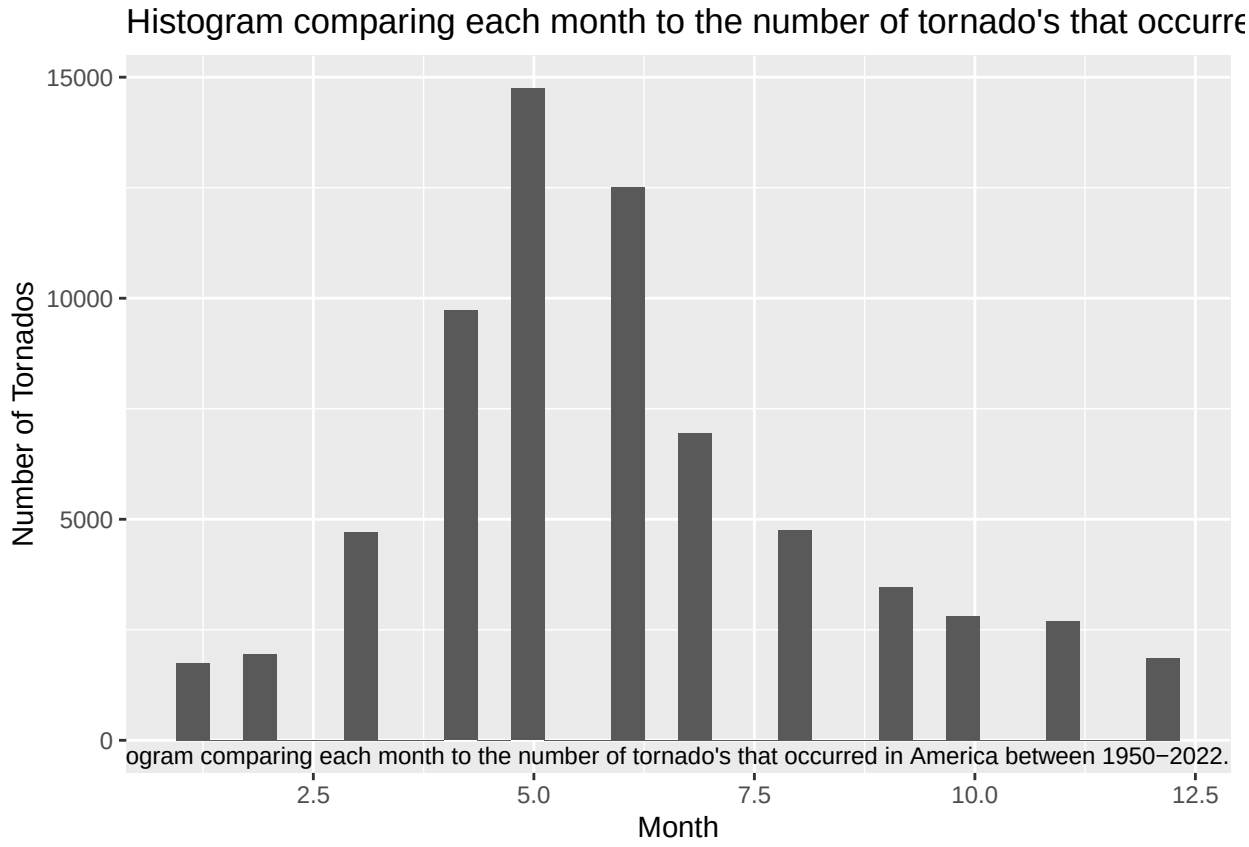



```
mean_loss<-mean(tornados$loss)
sd_loss<-sd(tornados$loss)
length_loss<-length(tornados$loss)
t_loss <- qt(p = 0.025, df = length_loss-1, lower.tail = FALSE)
lwr <- mean_loss - t_loss * sd_loss / sqrt(length_loss)
upr <- mean_loss + t_loss * sd_loss / sqrt(length_loss)
ci_loss <- c(lwr = lwr, upr = upr)
ci_loss
```

```
##       lwr       upr
##  50045.04 185908.78
```

```
## Section 2.8-Histogram comparing each month to the number of tornado's that occurred.
ggplot(tornados,aes(mo))+geom_histogram()+
  labs(title="Histogram comparing each month to the number of tornado's that occurred.")+
  annotate("text",x=Inf,y=-Inf,
           label="Figure 4.0- Histogram comparing each month to the number of tornado's that occurred i
  labs(x="Month",y="Number of Tornados")
```

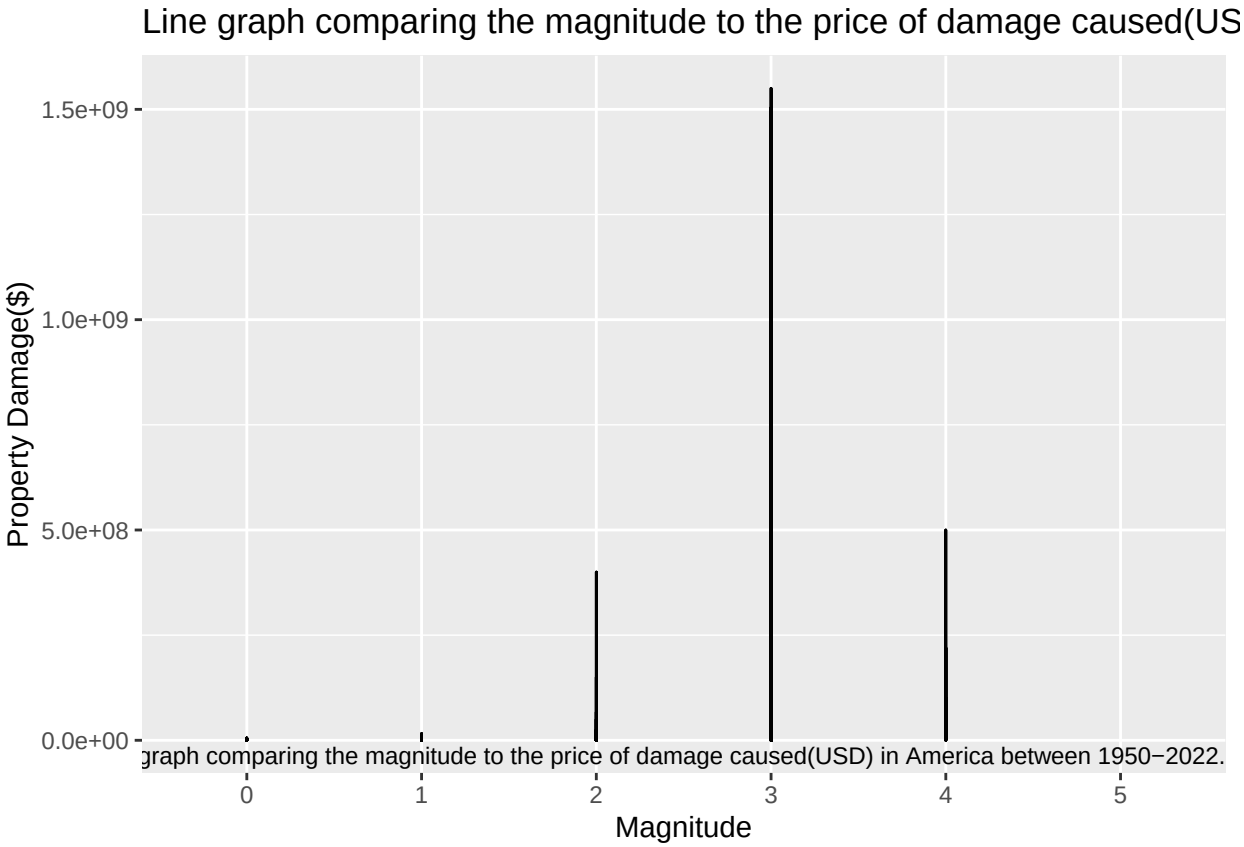
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mean_mo<-mean(tornados$mo)
sd_mo<-sd(tornados$mo)
length_mo<-length(tornados$mo)
t_mo <- qt(p = 0.025, df = length_mo-1, lower.tail = FALSE)
lwr <- mean_mo - t_mo * sd_mo / sqrt(length_mo)
upr <- mean_mo + t_mo * sd_mo / sqrt(length_mo)
ci_mo <- c(lwr = lwr, upr = upr)
ci_mo
```

```
##      lwr      upr
## 5.950747 5.987585
```

```
## Section 2.9-Line graph comparing the magnitude to the price of damage caused(USD)
ggplot(tornados,aes(mag,loss))+geom_line()+
  labs(title="Line graph comparing the magnitude to the price of damage caused(USD)")+
  annotate("text",x=Inf,y=-Inf,label="Figure 5.0- Line graph comparing the magnitude to the price of dam
  labs(x="Magnitude",y="Property Damage($)")
```




```
tornados$mag<-as.integer(tornados$mag)
mag_sd<-sd(tornados$mag)
mag_mean<-mean(tornados$mag)
mag_length<-length(tornados$mag)
t_mag <- qt(p = 0.025, df = mag_length-1, lower.tail = FALSE)
lwr <- mag_mean - t_mag * mag_sd / sqrt(mag_length)
upr <- mag_mean + t_mag * mag_sd / sqrt(mag_length)
ci_mag <- c(lwr = lwr, upr = upr)
ci_mag
```

```
##      lwr      upr
## 1.771953 1.785424
```

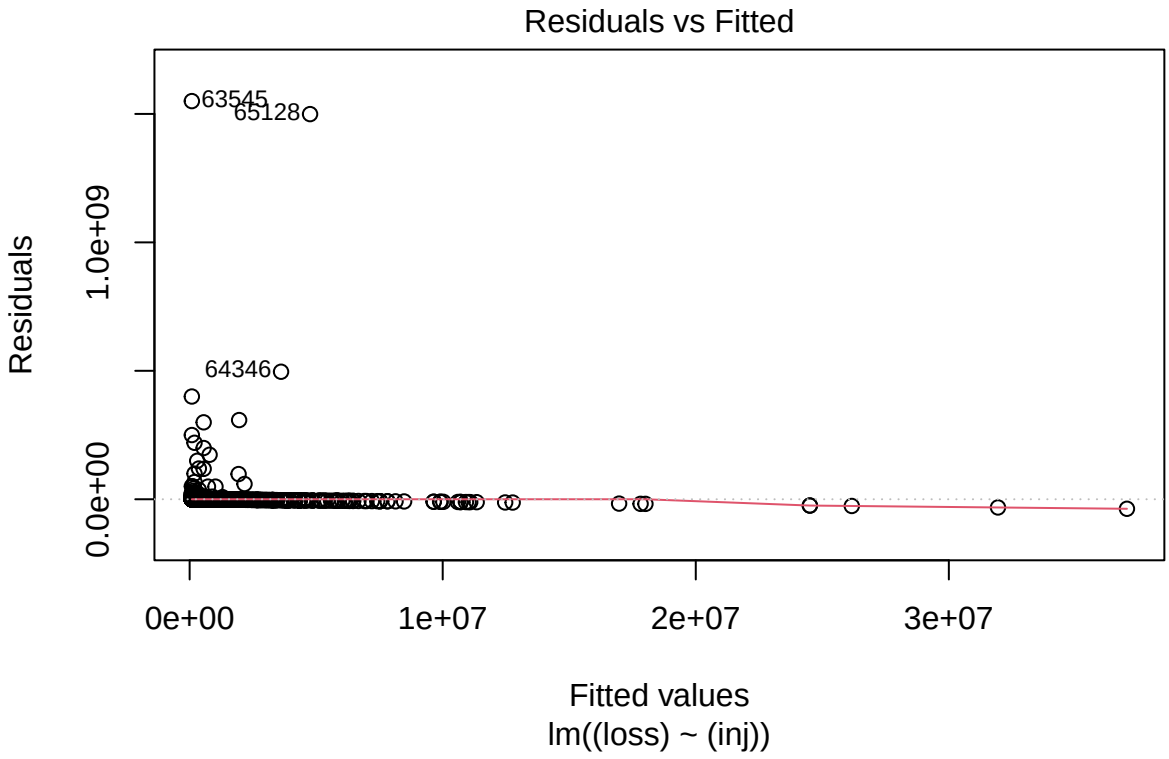
```
tornados$mag<-as.factor(tornados$mag)
```

```
## Section 3.0-Linear Regression model predicting Price of damage(USD) from the number of injuries.
inj_lm<-lm((loss)~(inj),data=tornados)
summary(inj_lm)
```

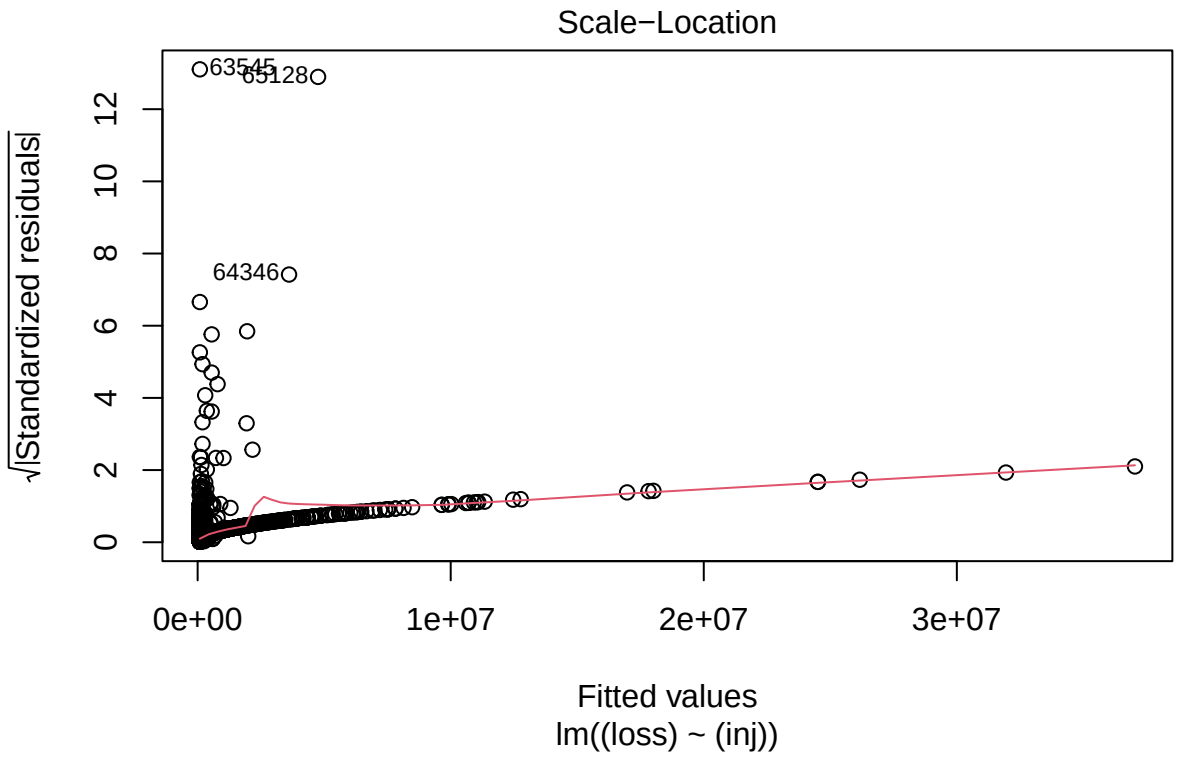
```
##
## Call:
## lm(formula = (loss) ~ (inj), data = tornados)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
##  -37038363     -87518     -87518     -87515 1549912482
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    87518      34735    2.52   0.0118 *
## inj            21236       1901   11.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9026000 on 67943 degrees of freedom
## Multiple R-squared:  0.001833,   Adjusted R-squared:  0.001818
## F-statistic: 124.8 on 1 and 67943 DF,  p-value: < 2.2e-16
```

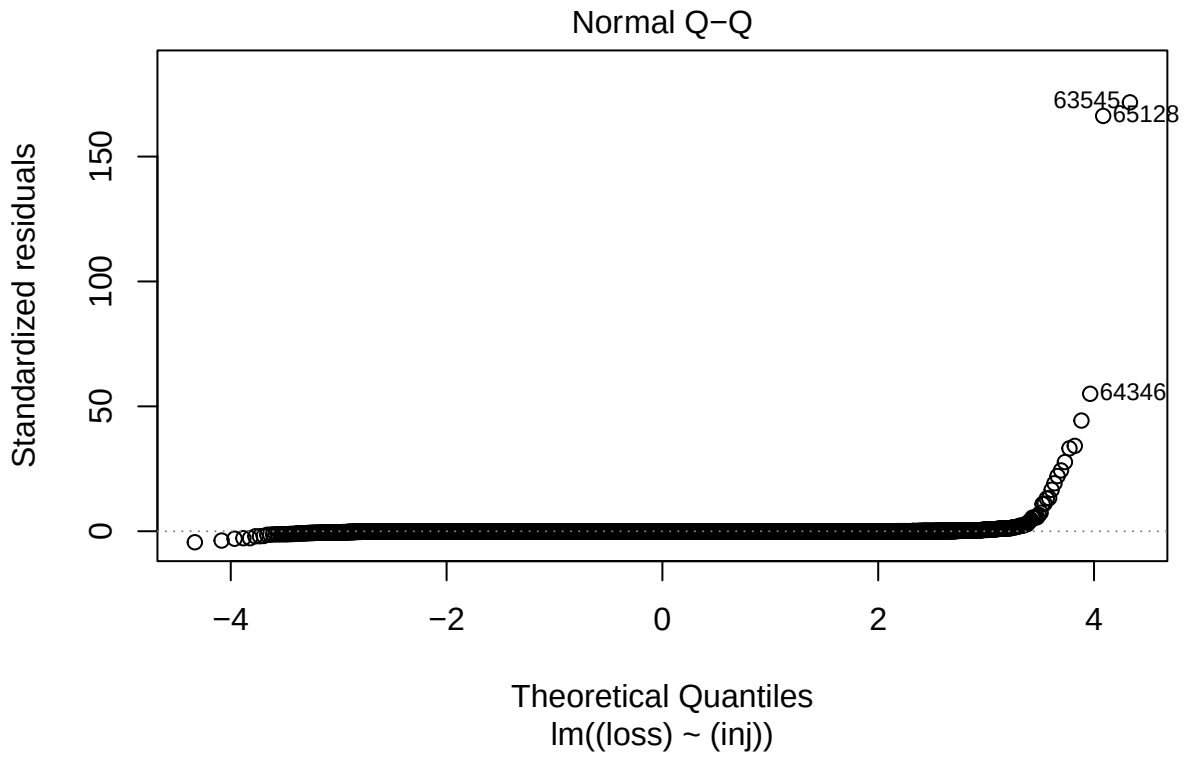
```
## Section 3.1-Linearity Test of linear Regression Model in section 3.0.
plot(inj_lm, which = 1)
```




```
## Section 3.2-Homoscedasticity Test of linear Regression Model in section 3.0.
plot(inj_lm, which = 3)
```




```
## Section 3.3-Normality Test of linear Regression Model in section 3.0.
plot(inj_lm, which = 2)
```




```
## Section 3.4-Multiple Regression model predicting the number of injured citizen from a tornado given
inj_multiple_lm<-lm((inj)~(mag)+st+len+wid,data=tornados)
summary(inj_multiple_lm)
```

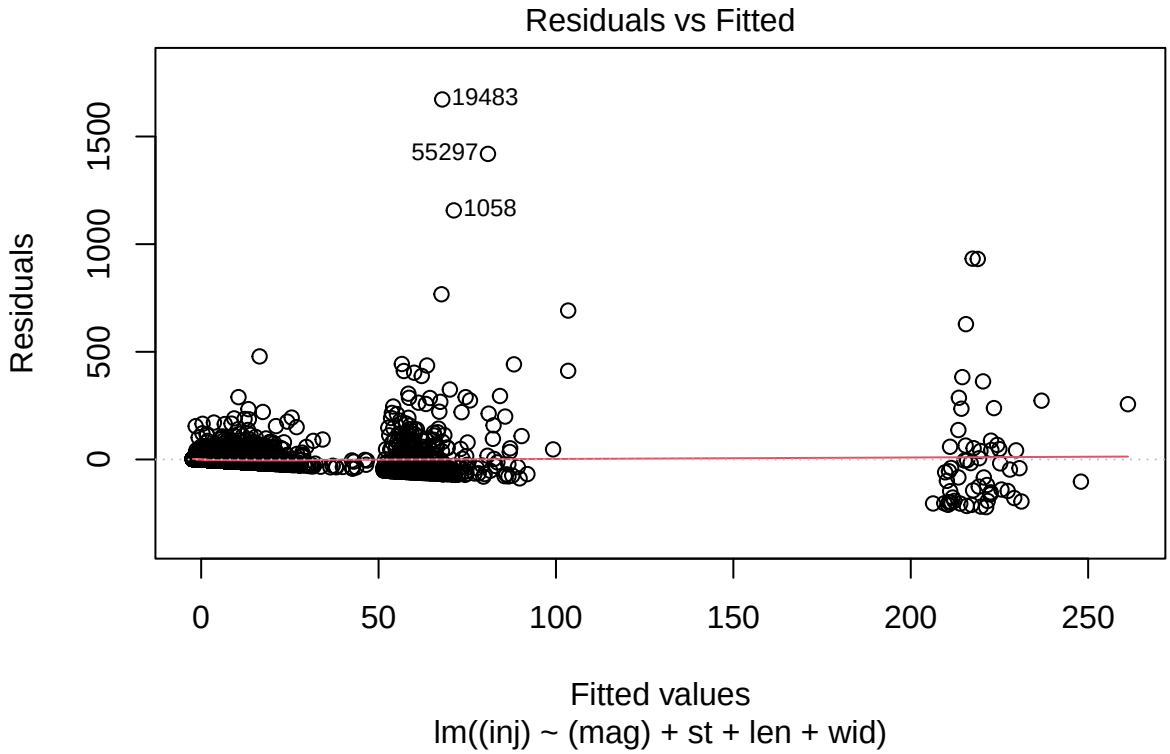
```
##
## Call:
## lm(formula = (inj) ~ (mag) + st + len + wid, data = tornados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -221.24   -0.73    0.06    0.78 1672.05
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.001e-02  7.930e+00  -0.003   0.9980
## mag2        -6.160e-01  1.421e-01  -4.336 1.45e-05 ***
## mag3        -4.028e-01  1.974e-01  -2.040   0.0414 *
## mag4         4.892e+00  3.588e-01  13.636  < 2e-16 ***
## mag5         5.244e+01  7.150e-01  73.346  < 2e-16 ***
## mag6         2.078e+02  2.101e+00  98.894  < 2e-16 ***
## stAL         3.433e-01  7.936e+00   0.043   0.9655
## stAR        -4.275e-02  7.938e+00  -0.005   0.9957
## stAZ         1.489e-03  7.987e+00   0.000   0.9999
## stCA        -3.197e-02  7.964e+00  -0.004   0.9968
## stCO        -1.554e-01  7.937e+00  -0.020   0.9844
## stCT         4.259e+00  8.061e+00   0.528   0.5973
```

```
## stDC        -4.934e-02  1.211e+01  -0.004   0.9967
## stDE         2.386e-01  8.157e+00   0.029   0.9767
## stFL         4.854e-01  7.934e+00   0.061   0.9512
## stGA         4.697e-01  7.939e+00   0.059   0.9528
## stHI         5.189e-02  8.307e+00   0.006   0.9950
## stIA        -1.769e+00  7.936e+00  -0.223   0.8236
## stID        -2.322e-01  8.002e+00  -0.029   0.9769
## stIL        -2.438e-01  7.936e+00  -0.031   0.9755
## stIN         5.779e-01  7.940e+00   0.073   0.9420
## stKS        -1.398e+00  7.933e+00  -0.176   0.8602
## stKY         5.800e-01  7.944e+00   0.073   0.9418
## stLA         9.651e-03  7.937e+00   0.001   0.9990
## stMA         7.305e+00  8.019e+00   0.911   0.3623
## stMD         4.556e-02  7.970e+00   0.006   0.9954
## stME        -3.483e-01  8.048e+00  -0.043   0.9655
## stMI         5.726e-01  7.945e+00   0.072   0.9425
## stMN        -7.919e-01  7.937e+00  -0.100   0.9205
## stMO        -4.839e-01  7.936e+00  -0.061   0.9514
## stMS        -4.181e-01  7.936e+00  -0.053   0.9580
## stMT        -6.224e-01  7.966e+00  -0.078   0.9377
## stNC         4.789e-01  7.941e+00   0.060   0.9519
## stND        -9.793e-01  7.939e+00  -0.123   0.9018
## stNE        -1.339e+00  7.935e+00  -0.169   0.8660
## stNH        -5.426e-01  8.094e+00  -0.067   0.9466
## stNJ        -4.688e-01  8.022e+00  -0.058   0.9534
## stNM        -1.116e-01  7.955e+00  -0.014   0.9888
## stNV        -2.068e-01  8.097e+00  -0.026   0.9796
## stNY        -5.039e-01  7.963e+00  -0.063   0.9495
## stOH         1.647e+00  7.943e+00   0.207   0.8357
## stOK        -9.399e-01  7.934e+00  -0.118   0.9057
## stOR         2.153e+00  8.057e+00   0.267   0.7893
## stPA        -1.197e-01  7.947e+00  -0.015   0.9880
## stPR        -1.626e-01  8.459e+00  -0.019   0.9847
## stRI         8.344e-01  9.068e+00   0.092   0.9267
## stSC        -9.943e-02  7.944e+00  -0.013   0.9900
## stSD        -7.471e-01  7.938e+00  -0.094   0.9250
## stTN         7.125e-01  7.942e+00   0.090   0.9285
## stTX        -2.208e-01  7.931e+00  -0.028   0.9778
## stUT         4.462e-01  8.051e+00   0.055   0.9558
## stVA         1.064e-01  7.950e+00   0.013   0.9893
## stVI        -1.212e-01  1.773e+01  -0.007   0.9945
## stVT        -2.042e-01  8.254e+00  -0.025   0.9803
## stWA        -3.650e-01  8.055e+00  -0.045   0.9639
## stWI        -1.288e+00  7.941e+00  -0.162   0.8711
## stWV         1.370e-01  8.038e+00   0.017   0.9864
## stWY        -4.440e-01  7.952e+00  -0.056   0.9555
## len          2.530e-01  8.605e-03  29.406  < 2e-16 ***
## wid          2.942e-03  3.369e-04   8.732  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.86 on 67885 degrees of freedom
## Multiple R-squared:  0.2425, Adjusted R-squared:  0.2418
## F-statistic: 368.3 on 59 and 67885 DF,  p-value: < 2.2e-16
```

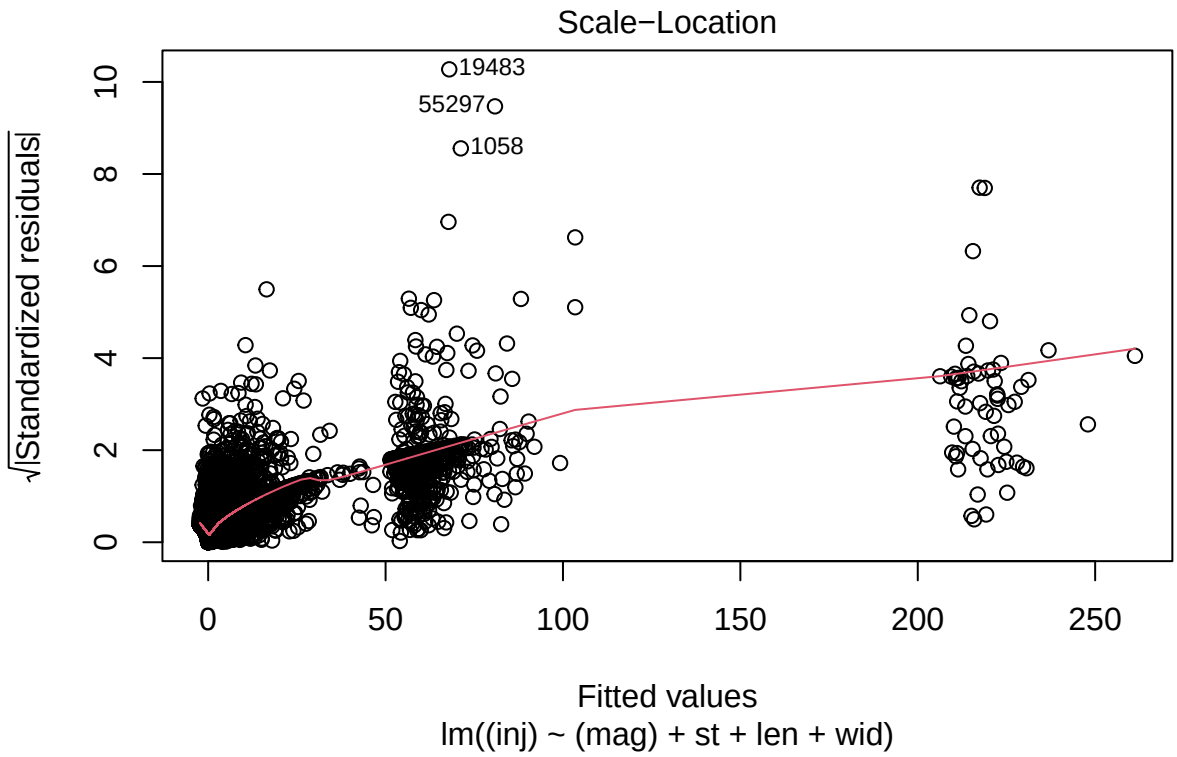
```r
## Section 3.5-ANOVA Test of multiple Regression Model in section 3.4.
anova(inj_multiple_lm)
```

```
## Analysis of Variance Table
##
## Response: (inj)
##              Df   Sum Sq Mean Sq   F value    Pr(>F)
## mag           5  5146086 1029217 4092.1729 < 2.2e-16 ***
## st           52    45127     868    3.4505  6.84e-16 ***
## len           1   255300  255300 1015.0745 < 2.2e-16 ***
## wid           1    19177   19177   76.2483 < 2.2e-16 ***
## Residuals 67885 17073670     252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

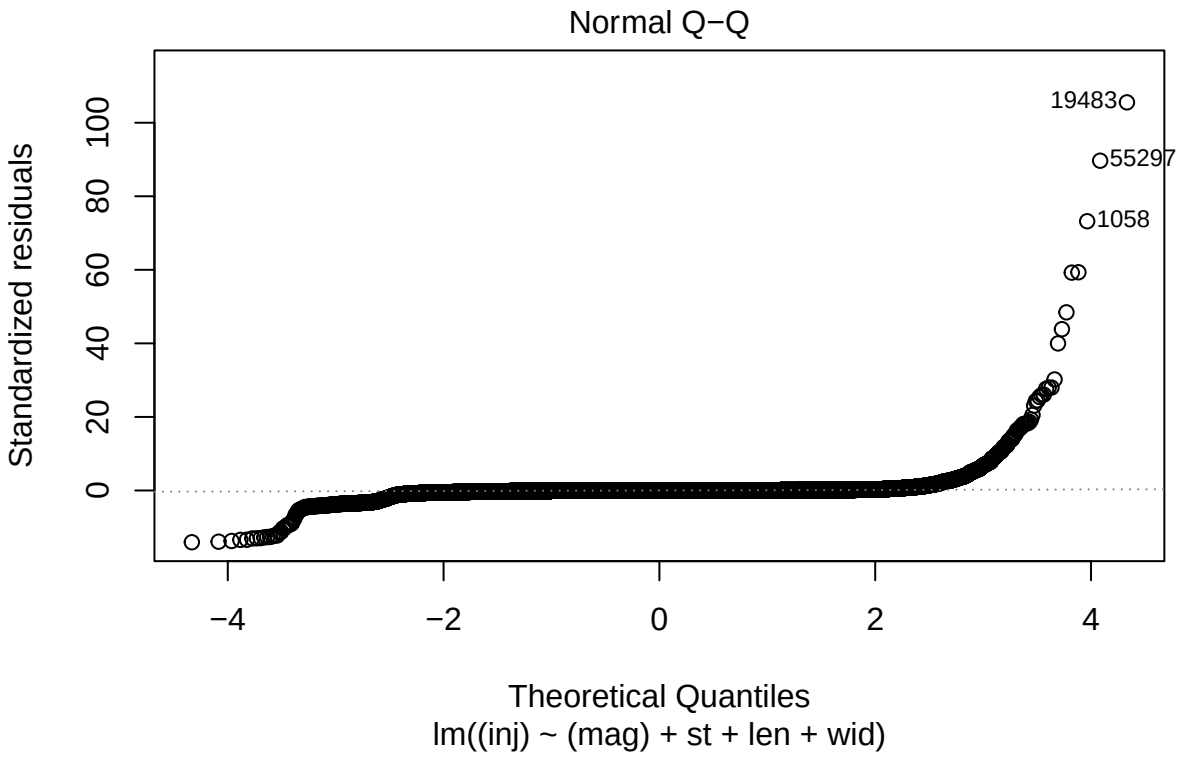
```r
## Section 3.6-Linearity Test of multiple Regression Model in section 3.4.
plot(inj_multiple_lm, which = 1)
```




```r
## Section 3.7-Homoscedasticity Test of multiple Regression Model in section 3.4.
plot(inj_multiple_lm, which = 3)
```




```
## Section 3.8-Normality Test of multiple Regression Model in section 3.4.
plot(inj_multiple_lm, which = 2)
```




```
## Section 3.9-Prediction Test of multiple Regression Model in section 3.4.
predict(inj_multiple_lm,newdata = tibble(mag="3",st="IL",len=2.8,wid=50))
```

```
##         1
## 0.1890425
```

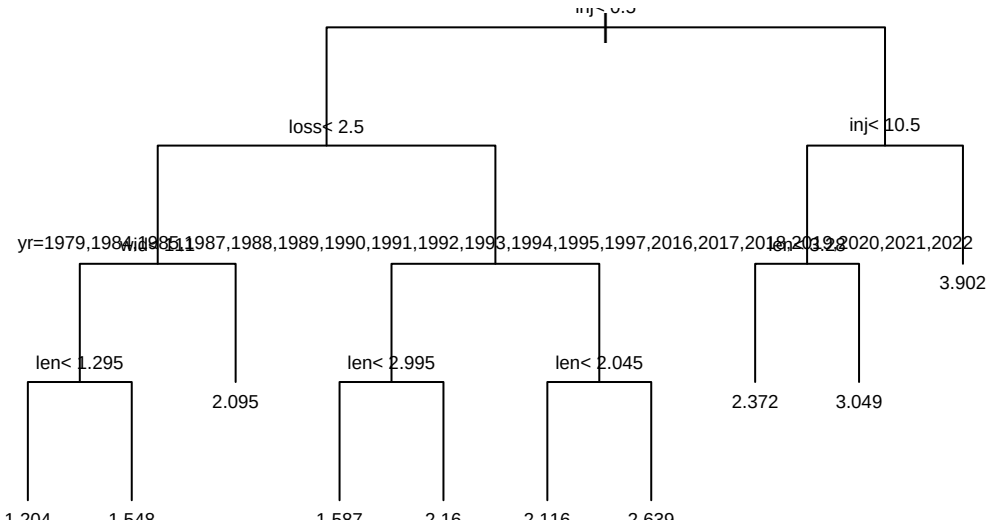
```
## Section-3.10-Poisson Regression Model
tornados$loss <- round(tornados$loss)
tornadosPoiss<-glm(loss~len+wid+mag+st,family = poisson(link="log"),
                   data=tornados)
summary(tornadosPoiss)
```

```
##
## Call:
## glm(formula = loss ~ len + wid + mag + st, family = poisson(link = "log"),
##     data = tornados)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -20117    -190     -85     -28  107016
##
## Coefficients:
##               Estimate Std. Error    z value Pr(>|z|)
## (Intercept) -6.976e-01  7.071e-01     -0.987 0.323839
## len          5.257e-03  4.002e-07  13134.131  < 2e-16 ***
```

```
## wid          1.676e-03  1.519e-08 110360.219  < 2e-16 ***
## mag2         1.617e+00  1.100e-04  14699.195  < 2e-16 ***
## mag3         3.089e+00  1.044e-04  29597.905  < 2e-16 ***
## mag4         5.321e+00  1.020e-04  52152.350  < 2e-16 ***
## mag5         4.158e+00  1.099e-04  37817.120  < 2e-16 ***
## mag6        -5.098e+00  1.117e-02   -456.278  < 2e-16 ***
## stAL         5.066e+00  7.071e-01      7.165 7.78e-13 ***
## stAR         8.631e+00  7.071e-01     12.207  < 2e-16 ***
## stAZ         6.358e+00  7.071e-01      8.992  < 2e-16 ***
## stCA         6.221e+00  7.071e-01      8.798  < 2e-16 ***
## stCO         4.536e+00  7.071e-01      6.416 1.40e-10 ***
## stCT         6.911e+00  7.071e-01      9.773  < 2e-16 ***
## stDC         1.102e+01  7.071e-01     15.588  < 2e-16 ***
## stDE        -9.172e-01  7.112e-01     -1.290 0.197204
## stFL         8.849e+00  7.071e-01     12.514  < 2e-16 ***
## stGA         8.633e+00  7.071e-01     12.209  < 2e-16 ***
## stHI        -2.814e-01  7.181e-01     -0.392 0.695184
## stIA         9.247e+00  7.071e-01     13.077  < 2e-16 ***
## stID         2.090e+00  7.073e-01      2.955 0.003125 **
## stIL         7.870e+00  7.071e-01     11.129  < 2e-16 ***
## stIN         6.751e+00  7.071e-01      9.547  < 2e-16 ***
## stKS         6.245e+00  7.071e-01      8.831  < 2e-16 ***
## stKY         7.215e+00  7.071e-01     10.204  < 2e-16 ***
## stLA         9.531e+00  7.071e-01     13.479  < 2e-16 ***
## stMA         8.504e+00  7.071e-01     12.027  < 2e-16 ***
## stMD         8.091e+00  7.071e-01     11.442  < 2e-16 ***
## stME         2.126e+00  7.073e-01      3.007 0.002642 **
## stMI         8.595e+00  7.071e-01     12.156  < 2e-16 ***
## stMN         6.660e+00  7.071e-01      9.418  < 2e-16 ***
## stMO         8.724e+00  7.071e-01     12.337  < 2e-16 ***
## stMS         7.463e+00  7.071e-01     10.554  < 2e-16 ***
## stMT         6.110e+00  7.071e-01      8.641  < 2e-16 ***
## stNC         9.022e+00  7.071e-01     12.759  < 2e-16 ***
## stND         6.718e+00  7.071e-01      9.500  < 2e-16 ***
## stNE         4.618e+00  7.071e-01      6.531 6.54e-11 ***
## stNH         2.666e+00  7.072e-01      3.770 0.000164 ***
## stNJ         8.164e+00  7.071e-01     11.546  < 2e-16 ***
## stNM         7.979e+00  7.071e-01     11.284  < 2e-16 ***
## stNV        -3.334e-01  7.186e-01     -0.464 0.642711
## stNY         6.656e+00  7.071e-01      9.413  < 2e-16 ***
## stOH         1.059e+01  7.071e-01     14.970  < 2e-16 ***
## stOK         5.929e+00  7.071e-01      8.385  < 2e-16 ***
## stOR         8.165e+00  7.071e-01     11.547  < 2e-16 ***
## stPA         7.332e+00  7.071e-01     10.370  < 2e-16 ***
## stPR         8.325e+00  7.071e-01     11.773  < 2e-16 ***
## stRI         9.371e+00  7.071e-01     13.252  < 2e-16 ***
## stSC         8.672e+00  7.071e-01     12.263  < 2e-16 ***
## stSD         6.371e+00  7.071e-01      9.010  < 2e-16 ***
## stTN         1.042e+01  7.071e-01     14.740  < 2e-16 ***
## stTX         9.591e+00  7.071e-01     13.564  < 2e-16 ***
## stUT         7.985e+00  7.071e-01     11.292  < 2e-16 ***
## stVA         8.850e+00  7.071e-01     12.516  < 2e-16 ***
## stVI        -1.262e+01  4.693e+02     -0.027 0.978555
## stVT         5.715e+00  7.071e-01      8.082 6.39e-16 ***
```

```
## stWA         7.870e+00  7.071e-01     11.129  < 2e-16 ***
## stWI         6.588e+00  7.071e-01      9.317  < 2e-16 ***
## stWV         7.628e+00  7.071e-01     10.787  < 2e-16 ***
## stWY         4.568e+00  7.071e-01      6.460 1.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.1819e+11  on 67944  degrees of freedom
## Residual deviance: 6.3085e+10  on 67885  degrees of freedom
## AIC: 6.3085e+10
##
## Number of Fisher Scoring iterations: 11
```

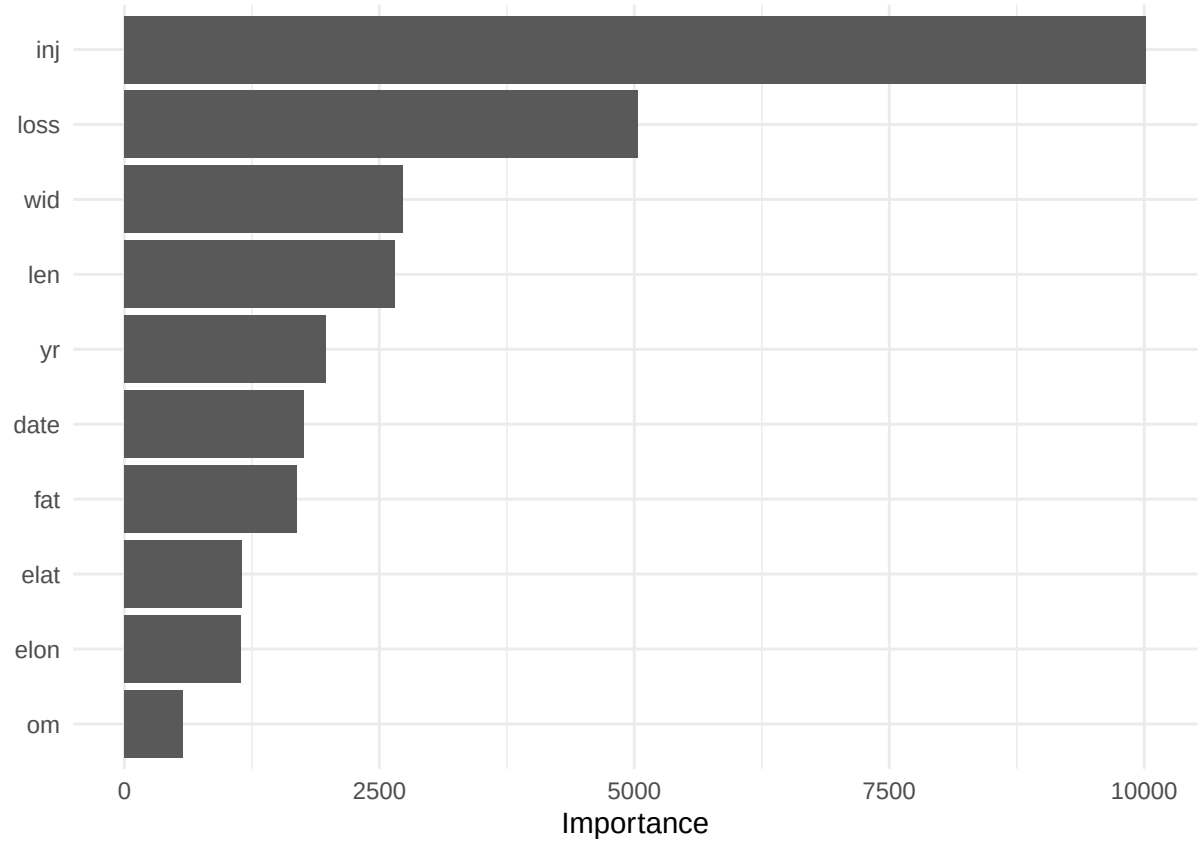
```
## Section 3.11- Regression Tree
set.seed(1223)
tornados_split<-initial_split(tornados)
tornados_train<-training(tornados_split)
tornados_test<-testing(tornados_split)
## Code from Week 9: Regression Trees - An Example
tornados_train$mag<-as.integer(tornados_train$mag)
reg_tree_spec<-decision_tree( mode = "regression" ) %>% set_engine( "rpart" )
tornados_tree<-reg_tree_spec%>%fit(mag~.,data=tornados_train)
plot(tornados_tree$fit,uniform = TRUE,cex=0.8)
text(tornados_tree$fit,pretty=0,cex=0.6)
```



```
## Section 3.12-Checking if Tree is in a tree structure
tornados_tree
```

```
## parsnip model object
##
## n= 50958
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 50958 40966.6800 1.779779
##    2) inj< 0.5 45114 26194.9900 1.630093
##      4) loss< 2.5 29404 11842.8300 1.399299
##        8) wid< 111 25289  6939.5470 1.286132
##         16) len< 1.295 19253  3950.6090 1.204020 *
##         17) len>=1.295 6036  2445.0670 1.548045 *
##        9) wid>=111 4115  2589.0380 2.094775 *
##      5) loss>=2.5 15710  9854.4890 2.062062
##       10) yr=1979,1984,1985,1987,1988,1989,1990,1991,1992,1993,1994,1995,1997,2016,2017,2018,2019,20
##         20) len< 2.995 4577  1699.3800 1.587284 *
##         21) len>=2.995 2287  1496.1070 2.160472 *
##       11) yr=1950,1951,1952,1953,1954,1955,1956,1957,1958,1959,1960,1961,1962,1963,1964,1965,1966,19
##         22) len< 2.045 6038  2943.1500 2.116429 *
##         23) len>=2.045 2808  1709.8330 2.638889 *
##    3) inj>=0.5 5844  5957.5500 2.935318
##      6) inj< 10.5 4788  3753.0000 2.722013
##       12) len< 3.28 2313  1416.4980 2.372244 *
##       13) len>=3.28 2475  1789.0840 3.048889 *
##      7) inj>=10.5 1056   998.9536 3.902462 *
```

```
## Section 3.13- Most Influential Variable in the tree
tornados_tree%>%vip(num_features=)+theme_minimal()
```




```
## Section 3.14- Testing Prediction capability of regression tree using predict()
## NOTE: THIS HAD TO BE COMMENTED OUT DUE TO ERRORS DURING KNITTING
##tornados_test$mag<-as.factor(tornados_test$mag)
##tornados_testing_prediction<-as.factor(tornados_testing_prediction$mag)
##tornados_testing_prediction<-tornados_tree%>%
  ##predict(new_data = tornados_test)%>%bind_cols(tornados_test)
##tornados_testing_prediction%>%metrics(truth=mag,estimate=.pred)
```

```
## Section 3.15-Logistic Regression Model
tornados$inj_over5<-ifelse(tornados$inj>5,1,0)
tornados_logistic<-glm(inj_over5~mag+st+loss+len+wid,data=tornados,family=binomial())
summary(tornados_logistic)
```

```
##
## Call:
## glm(formula = inj_over5 ~ mag + st + loss + len + wid, family = binomial(),
##     data = tornados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7562  -0.1708  -0.0582  -0.0318   3.9080
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.757e+01  1.978e+03  -0.009    0.993
```

```
## mag2         2.624e+00  2.276e-01  11.531  < 2e-16 ***
## mag3         4.554e+00  2.231e-01  20.417  < 2e-16 ***
## mag4         6.042e+00  2.263e-01  26.692  < 2e-16 ***
## mag5         7.489e+00  2.445e-01  30.635  < 2e-16 ***
## mag6         9.133e+00  5.354e-01  17.058  < 2e-16 ***
## stAL         1.073e+01  1.978e+03   0.005    0.996
## stAR         1.052e+01  1.978e+03   0.005    0.996
## stAZ         1.109e+01  1.978e+03   0.006    0.996
## stCA         1.008e+01  1.978e+03   0.005    0.996
## stCO         9.469e+00  1.978e+03   0.005    0.996
## stCT         1.093e+01  1.978e+03   0.006    0.996
## stDC        -4.404e-02  3.022e+03   0.000    1.000
## stDE         1.097e+01  1.978e+03   0.006    0.996
## stFL         1.113e+01  1.978e+03   0.006    0.996
## stGA         1.102e+01  1.978e+03   0.006    0.996
## stHI        -2.339e+00  2.052e+03  -0.001    0.999
## stIA         9.086e+00  1.978e+03   0.005    0.996
## stID        -2.113e+00  1.992e+03  -0.001    0.999
## stIL         1.030e+01  1.978e+03   0.005    0.996
## stIN         1.019e+01  1.978e+03   0.005    0.996
## stKS         8.852e+00  1.978e+03   0.004    0.996
## stKY         1.084e+01  1.978e+03   0.005    0.996
## stLA         1.049e+01  1.978e+03   0.005    0.996
## stMA         1.010e+01  1.978e+03   0.005    0.996
## stMD         1.035e+01  1.978e+03   0.005    0.996
## stME        -3.049e+00  2.003e+03  -0.002    0.999
## stMI         1.045e+01  1.978e+03   0.005    0.996
## stMN         1.003e+01  1.978e+03   0.005    0.996
## stMO         1.032e+01  1.978e+03   0.005    0.996
## stMS         1.027e+01  1.978e+03   0.005    0.996
## stMT         7.832e+00  1.978e+03   0.004    0.997
## stNC         1.093e+01  1.978e+03   0.006    0.996
## stND         8.783e+00  1.978e+03   0.004    0.996
## stNE         8.811e+00  1.978e+03   0.004    0.996
## stNH         9.050e+00  1.978e+03   0.005    0.996
## stNJ         1.066e+01  1.978e+03   0.005    0.996
## stNM         1.067e+01  1.978e+03   0.005    0.996
## stNV        -1.046e+00  2.014e+03  -0.001    1.000
## stNY         1.041e+01  1.978e+03   0.005    0.996
## stOH         1.105e+01  1.978e+03   0.006    0.996
## stOK         9.922e+00  1.978e+03   0.005    0.996
## stOR         1.037e+01  1.978e+03   0.005    0.996
## stPA         1.033e+01  1.978e+03   0.005    0.996
## stPR        -9.089e-01  2.092e+03   0.000    1.000
## stRI         1.245e+01  1.978e+03   0.006    0.995
## stSC         1.078e+01  1.978e+03   0.005    0.996
## stSD         8.559e+00  1.978e+03   0.004    0.997
## stTN         1.096e+01  1.978e+03   0.006    0.996
## stTX         9.954e+00  1.978e+03   0.005    0.996
## stUT         1.068e+01  1.978e+03   0.005    0.996
## stVA         1.055e+01  1.978e+03   0.005    0.996
## stVI        -1.274e-02  4.423e+03   0.000    1.000
## stVT         1.007e+01  1.978e+03   0.005    0.996
## stWA        -2.874e+00  2.002e+03  -0.001    0.999
```

```
## stWI         9.997e+00  1.978e+03   0.005    0.996
## stWV         1.095e+01  1.978e+03   0.006    0.996
## stWY         8.671e+00  1.978e+03   0.004    0.997
## loss         2.533e-10  1.035e-09   0.245    0.807
## len          2.606e-02  1.737e-03  14.999  < 2e-16 ***
## wid          3.994e-04  6.893e-05   5.794 6.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20190  on 67944  degrees of freedom
## Residual deviance: 11684  on 67884  degrees of freedom
## AIC: 11806
##
## Number of Fisher Scoring iterations: 16
```

```
## Section 3.16-Logistic Regression Model, but state and loss removed due to the high p-values.
tornados_logistic<-glm(inj_over5~mag+len+wid,data=tornados,family=binomial())
summary(tornados_logistic)
```

```
##
## Call:
## glm(formula = inj_over5 ~ mag + len + wid, family = binomial(),
##     data = tornados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7163  -0.1517  -0.0458  -0.0358   3.8374
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.368e+00  2.181e-01 -33.781  < 2e-16 ***
## mag2         2.762e+00  2.268e-01  12.178  < 2e-16 ***
## mag3         4.654e+00  2.220e-01  20.965  < 2e-16 ***
## mag4         6.044e+00  2.246e-01  26.914  < 2e-16 ***
## mag5         7.323e+00  2.401e-01  30.500  < 2e-16 ***
## mag6         8.733e+00  5.244e-01  16.653  < 2e-16 ***
## len          2.401e-02  1.636e-03  14.677  < 2e-16 ***
## wid          3.675e-04  6.535e-05   5.624 1.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20190  on 67944  degrees of freedom
## Residual deviance: 12314  on 67937  degrees of freedom
## AIC: 12330
##
## Number of Fisher Scoring iterations: 9
```

```
## Section 3.17- Using ROC Curves to test validity of Logistic Regression Model
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
predict_logistic<-predict(tornados_logistic,type="response")
roc_logistic<-roc(tornados$inj_over5,predict_logistic)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
plot(roc_logistic)
```