**Introduction:**

This work focuses on an insight into the effects that led to the outcome of traffic fatalities. The dataset used in these findings is the 2015 Traffic Fatalities courtesy of www.Kaggle.com.

This dataset looks into accident as well as other data that factored into motor vehicle collisions that lead to fatalities. Two CSV files from this dataset were used for this analysis:

Accident.CSV

which looked into many attributes about the crash data such as when and where the accidents occurred, as well as the type of conditions of the crashes such as weather and the involvements of impaired drivers.

Distraction.CSV

looked into if the crashes included distractions that the drivers reported during the time of the crash.

**Results:**

Step 1:

The first portion only looked into the accident.csv dataset. Information about the dataset was collected using the str() function as well as taking table() and hist() of some of the attributes.

The dplyr package was used for this dataset for data manipulation

Step 2:

The number of accidents with x amount of people were found by using the group\_by() function to group by the number of persons in the accidents, and summarized using the summarize() function.

Results are shown as:

PERSONS sum\_persons

*<int>* *<int>*

1 0 0

2 1 13303

3 2 17650

4 3 14187

5 4 9416

6 5 6495

7 6 4296

8 7 2590

9 8 1776

10 9 810

# ... with 26 more rows

Step 3:

Next, the number of deaths per accident was found by again using the group\_by() and summarize() functions, and shows as:

FATALS sum\_persons

*<int>* *<int>*

1 1 29816

2 2 3876

3 3 900

4 4 308

5 5 120

6 6 54

7 8 8

8 10 10

Step 4:

For the following section, the data was slightly manipulated by filtering one of the values in the HOUR category in the accident.csv dataset, to account for technicalities in the recording of the hours of the day. Any hour greater than 24 was removed using the filter() function:

accident\_hour <- filter(accident, HOUR <24)

Step 5:

The states that the accidents occurred in were then shown using the group\_by() and summarize() functions to result in:

STATE sum\_persons

*<int>* *<int>*

1 1 849

2 2 65

3 4 893

4 5 531

5 6 3176

6 8 546

7 9 266

8 10 126

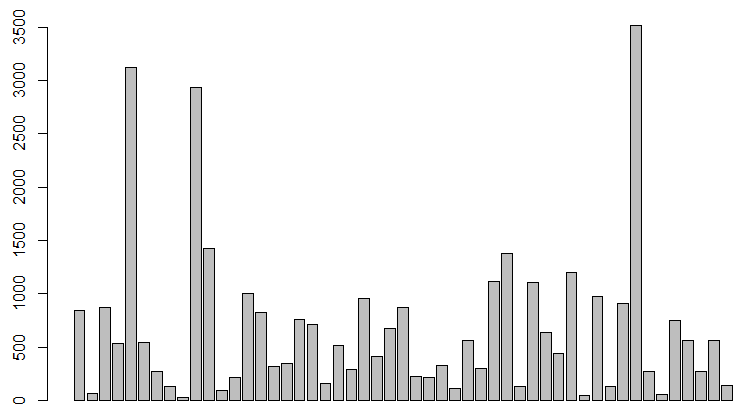
9 11 23

10 12 2939

# ... with 41 more rows

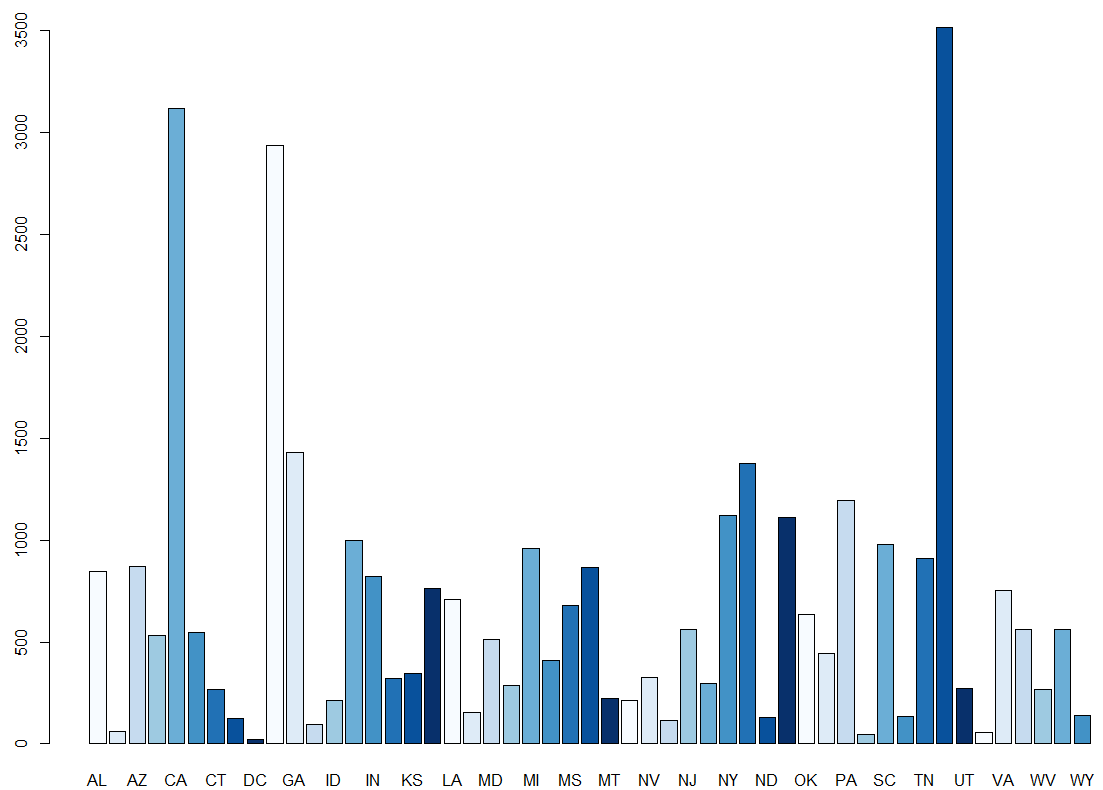
Step 6:

The deaths per state were then shown in a barplot which included the names of all the states alongside the bars:



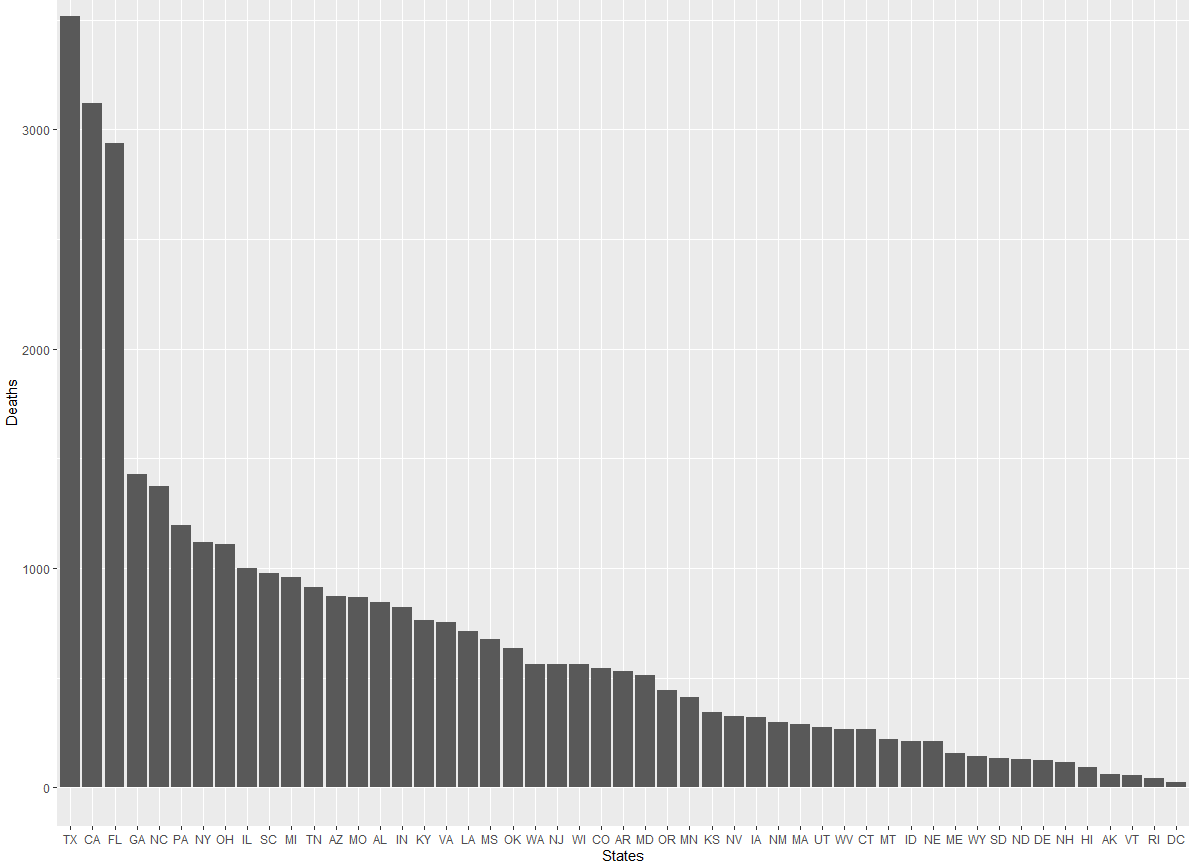
Step 7:

The packages ggplot2() and RColorBrewer() were then used to display the data in a colour format as well as the inclusion of the data labels:



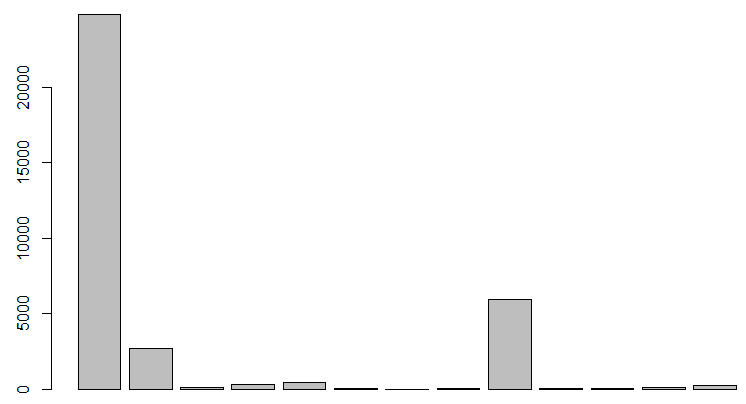
Step 8:

The deaths per state were then rearranged from highest to least using ggplot()



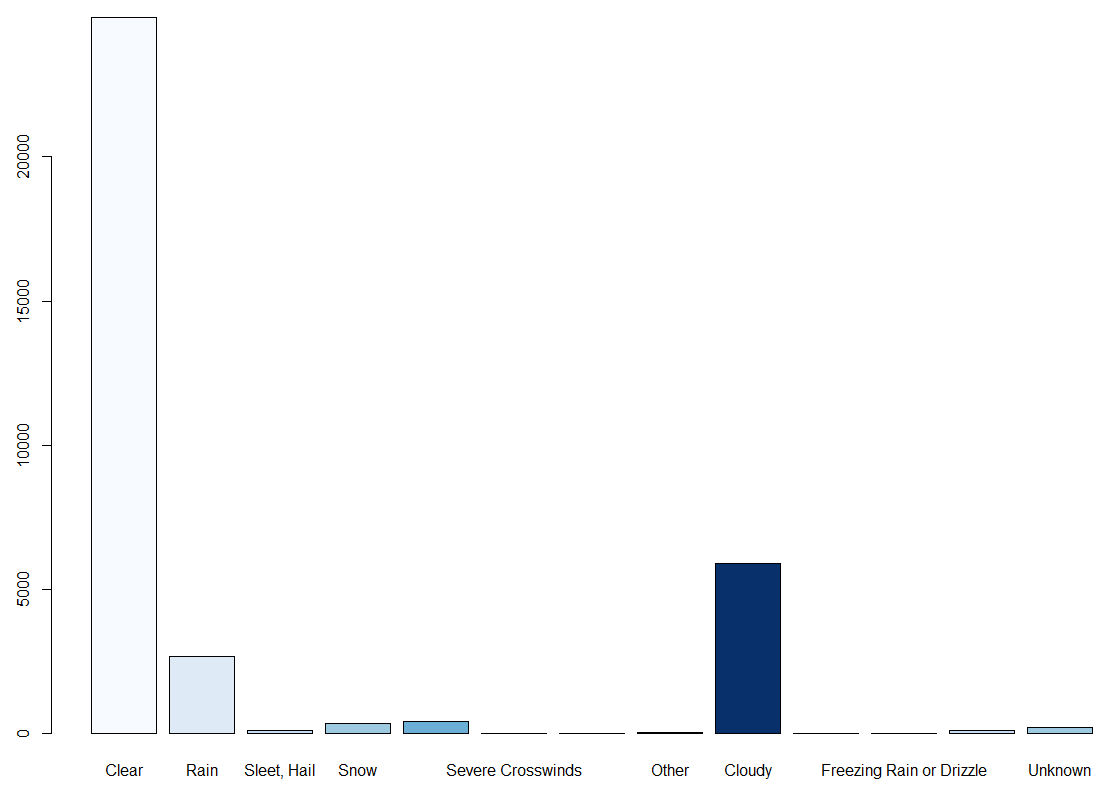
Step 9:

Data labels for the different weather codes were added using the metadata provided by Kaggle, to account for the different weather types that were recorded during the accidents:



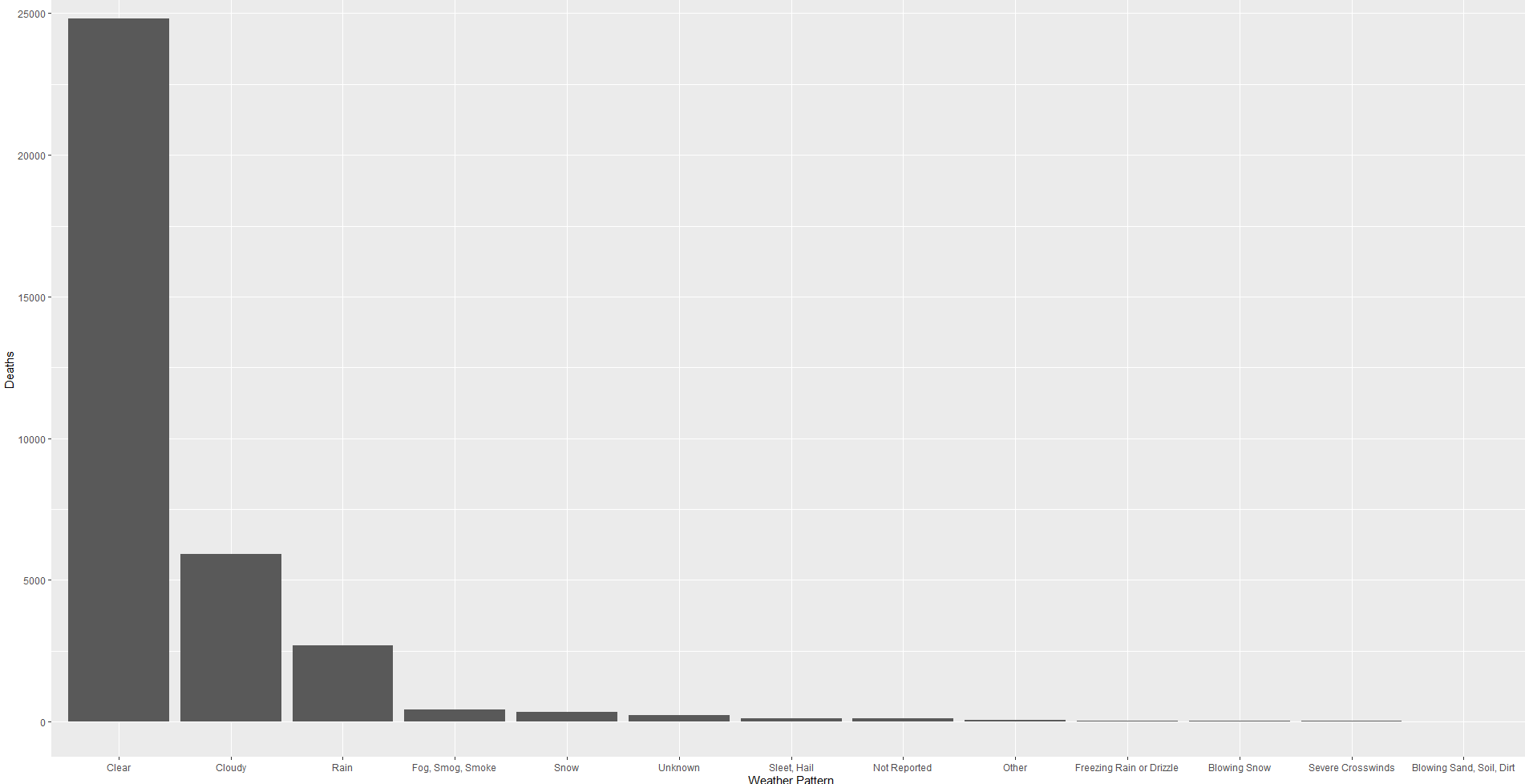
Step 10:

Data labels and colouring were then added:



Step 11:

The weather data was then rearranged in order of highest to lowest deaths per weather pattern, again using the ggplot() function:



Step 12:

A classification model was then attempted to determine if the factors of

* STATE
* CITY
* DAY OF THE WEEK
* NHS
* ROAD SYSTEM
* DRUNK DRIVER

were factors that could lead to the prediction of the number of fatalities in a given crash.

The decision tree model had the following table for the training data for cases where 1-10 fatalities were observed:

1 2 3 4 5 6 10

1 20667 1345 219 52 16 6 1

2 0 0 0 0 0 0 0

3 0 0 0 0 0 0 0

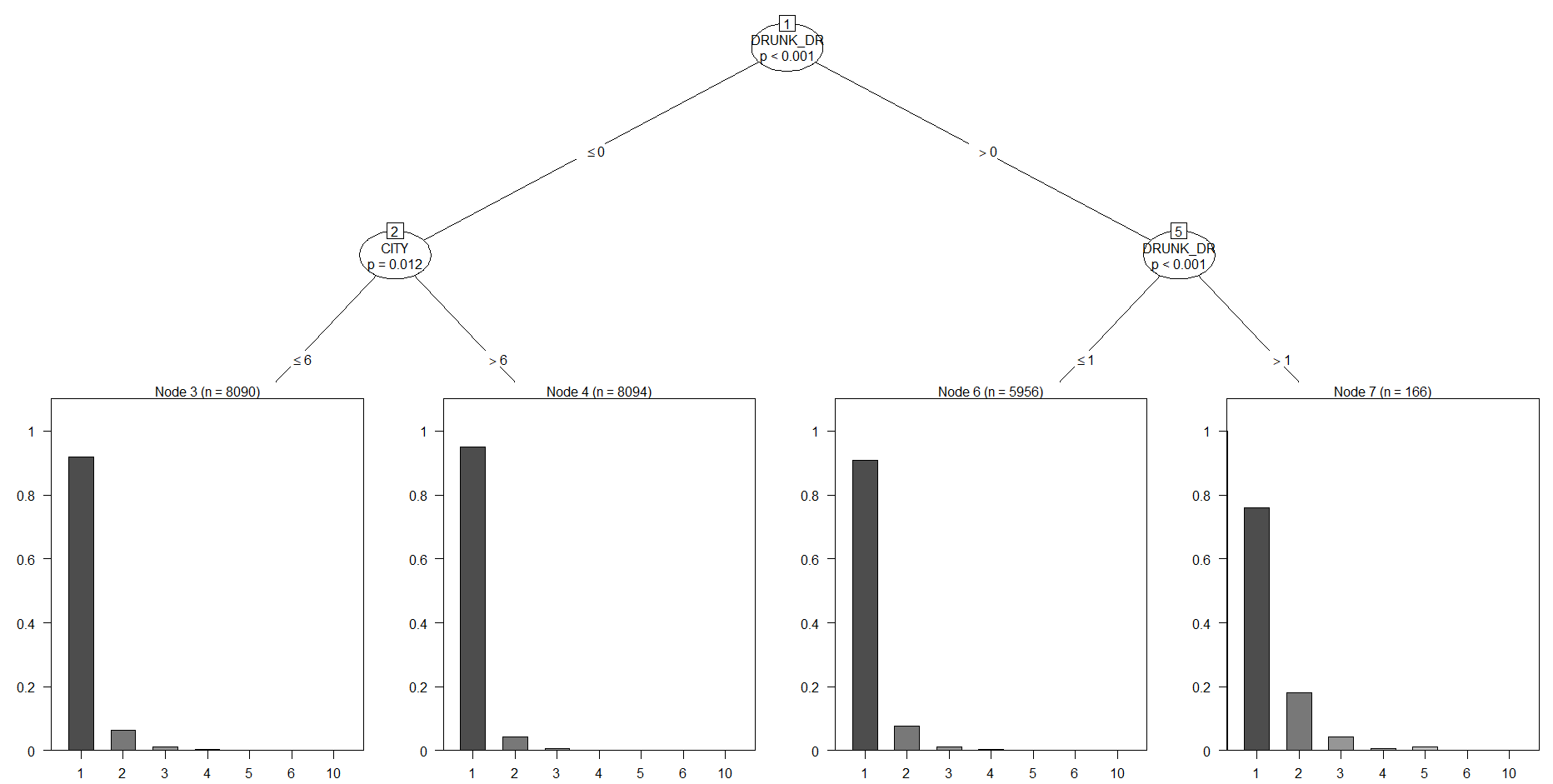
4 0 0 0 0 0 0 0

5 0 0 0 0 0 0 0

6 0 0 0 0 0 0 0

10 0 0 0 0 0 0 0

With the following decision tree:



Step 13:

The testing data was then used in the model to determine if the number of deaths were in fact predictable from the attributes used:

testpred 1 2 3 4 5 6 8

1 8913 581 80 25 8 3 1

2 0 0 0 0 0 0 0

3 0 0 0 0 0 0 0

4 0 0 0 0 0 0 0

5 0 0 0 0 0 0 0

6 0 0 0 0 0 0 0

10 0 0 0 0 0 0 0

**Conclusions from data:**

From the data in the earlier steps, it is observed that more than 13000 accidents involved one person whereas more than 17000 accidents involved two people. This is interesting since it shows that almost as many accidents are single vehicle crashes as those that are two-vehicle crashes. On the other hand, almost 30000 accidents involved only one death whereas only less than 4000 crashes lead to two deaths. This seems counter to the fact that there are more two-vehicle crashes, suggesting that it is much less likely for both parties involved in a collision to die than it is for only one person to.

When looking at the number of deaths per state, the three states that jump out the most are Texas, California and Florida. Texas and California are the two most populous states, and Florida is a very close 4th to New York’s 3rd. However, New York is 7th in terms of fatalities in 2015. This gap can be attributed largely to the fact that in the very highly populated of the New York City area, many people tend to go about their lives by other forms of transportation than motor vehicles.

When looking at the weather conditions in the fatal crashes, the overwhelming amount of accidents happened under Clear weather, and the second highest was in Cloudy weather. This suggests that weather is not a major factor in crashes that lead to fatalities, counterintuitively.

Finally, when looking at the classification by decision tree, the model was unable to successfully predict the number of fatalities given the attributes used. This may be due to the fact that some of the attributes were categorical and not numeric.