Final Project

Dipika Sharma

June 5, 2021

Add Citations

- R for Everyone (Lander 2014)
- Discovering Statistics Using R (Field, Miles, and Field 2012)

Introduction

For my final project I have analyzed drivers data. As we know Road accidents have become so common these days. We have heard or read about accidents at an average of 1/week.

```
bad_drivers <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/bad-drivers/bad-
```

I have learned about the structure of our data using str function.

```
str(bad_drivers)
```

```
## 'data.frame':    51 obs. of  8 variables:
##  $ State
##  $ Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
##  $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
##  $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
##  $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
##  $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Acci
##  $ Car.Insurance.Premiums....
##  $ Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
```

The dataset we are using have 8 different columns for 51 states of United State. I have renamed all the columns to make it simple and readable.

```
library(dplyr)

baddrivers_df <- bad_drivers %>%
  rename(driver_fatalities = "Number.of.drivers.involved.in.fatal.collisions.per.billion.miles",
         speeding_percent = "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding",
         alcohol_percent = "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaire
         not_distracted_percent = "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Dist
         no_prior_accident_percent = "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Be
         insurance_premiums = "Car.Insurance.Premiums....",
         insurance_companies_losses = "Losses.incurred.by.insurance.companies.for.collisions.per.insure

str(baddrivers_df)
```

```
## 'data.frame':    51 obs. of  8 variables:
##  $ State                     : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ driver_fatalities         : num  18.8 18.1 18.6 22.4 12 13.6 10.8 16.2 5.9 17.9 ...
##  $ speeding_percent          : int  39 41 35 18 35 37 46 38 34 21 ...
##  $ alcohol_percent           : int  30 25 28 26 28 28 36 30 27 29 ...
##  $ not_distracted_percent    : int  96 90 84 94 91 79 87 87 100 92 ...
##  $ no_prior_accident_percent : int  80 94 96 95 89 95 82 99 100 94 ...
##  $ insurance_premiums        : num  785 1053 899 827 878 ...
##  $ insurance_companies_losses: num  145 134 110 142 166 ...
```

The final data looks like below:

```
head(baddrivers_df)
```

```
##        State driver_fatalities speeding_percent alcohol_percent
## 1    Alabama              18.8               39              30
## 2     Alaska              18.1               41              25
## 3    Arizona              18.6               35              28
## 4   Arkansas              22.4               18              26
## 5 California              12.0               35              28
## 6   Colorado              13.6               37              28
##   not_distracted_percent no_prior_accident_percent insurance_premiums
## 1                     96                        80            784.55
## 2                     90                        94           1053.48
## 3                     84                        96            899.47
## 4                     94                        95            827.34
## 5                     91                        89            878.41
## 6                     79                        95            835.50
##   insurance_companies_losses
## 1                     145.08
## 2                     133.93
## 3                     110.35
## 4                     142.39
## 5                     165.63
## 6                     139.91
```

```
tail(baddrivers_df)
```

```
##            State driver_fatalities speeding_percent alcohol_percent
## 46       Vermont              13.6               30              30
## 47      Virginia              12.7               19              27
## 48    Washington              10.6               42              33
## 49 West Virginia              23.8               34              28
## 50     Wisconsin              13.8               36              33
## 51       Wyoming              17.4               42              32
##    not_distracted_percent no_prior_accident_percent insurance_premiums
## 46                     96                        95            716.20
## 47                     87                        88            768.95
## 48                     82                        86            890.03
## 49                     97                        87            992.61
## 50                     39                        84            670.31
## 51                     81                        90            791.14
##    insurance_companies_losses
```

```
## 46                    109.61
## 47                    153.72
## 48                    111.62
## 49                    152.56
## 50                    106.62
## 51                    122.04
```

Lets get the state region data. I found the state region data set in kaggle https://www.kaggle.com/omer2040/usa-states-to-region which i am using to see which region has worst drivers in United States.

```
## Set the working directory to the root of your DSC 520 directory
setwd("/Users/dipikasharma/R_Projects/DSC520")

## Load the `data/states.csv` to
states_df <- read.csv("data/states.csv")
```

Lets join the bad drivers data with states dataframe to get the region.

```
bad_drivers_df <- merge(x=baddrivers_df,y=states_df,by="State",all.x=TRUE)
str(bad_drivers_df)
```

```
## 'data.frame':    51 obs. of  11 variables:
##  $ State                     : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ driver_fatalities         : num  18.8 18.1 18.6 22.4 12 13.6 10.8 16.2 5.9 17.9 ...
##  $ speeding_percent          : int  39 41 35 18 35 37 46 38 34 21 ...
##  $ alcohol_percent           : int  30 25 28 26 28 28 36 30 27 29 ...
##  $ not_distracted_percent    : int  96 90 84 94 91 79 87 87 100 92 ...
##  $ no_prior_accident_percent : int  80 94 96 95 89 95 82 99 100 94 ...
##  $ insurance_premiums        : num  785 1053 899 827 878 ...
##  $ insurance_companies_losses: num  145 134 110 142 166 ...
##  $ State.Code                : chr  "AL" "AK" "AZ" "AR" ...
##  $ Region                    : chr  "South" "West" "West" "South" ...
##  $ Division                  : chr  "East South Central" "Pacific" "Mountain" "West South Central" .
```

As we can see now the data set structure has 51 observation of 11 variables.

The problem statement you addressed.

My purpose of analyzing the drivers data to find out Which states has the worst drivers in United States? Is number of accidents and premium are related to each other? I have studied How the premiums getting effected by increase rate of accidents? Which state has high and low premium in United States? Apart from that I have also added the region information in our data set to see which region in United States has worst drivers.

How you addressed this problem statement

```
summary(bad_drivers_df)
```

```
##     State           driver_fatalities speeding_percent alcohol_percent
##  Length:51          Min.   : 5.90     Min.   :13.00    Min.   :16.00
##  Class :character   1st Qu.:12.75     1st Qu.:23.00    1st Qu.:28.00
##  Mode  :character   Median :15.60     Median :34.00    Median :30.00
##                     Mean   :15.79     Mean   :31.73    Mean   :30.69
##                     3rd Qu.:18.50     3rd Qu.:38.00    3rd Qu.:33.00
##                     Max.   :23.90     Max.   :54.00    Max.   :44.00
##  not_distracted_percent no_prior_accident_percent insurance_premiums
##  Min.   : 10.00         Min.   : 76.00            Min.   : 642.0
##  1st Qu.: 83.00         1st Qu.: 83.50            1st Qu.: 768.4
##  Median : 88.00         Median : 88.00            Median : 859.0
##  Mean   : 85.92         Mean   : 88.73            Mean   : 887.0
##  3rd Qu.: 95.00         3rd Qu.: 95.00            3rd Qu.:1007.9
##  Max.   :100.00         Max.   :100.00            Max.   :1301.5
##  insurance_companies_losses  State.Code           Region
##  Min.   : 82.75             Length:51          Length:51
##  1st Qu.:114.64             Class :character   Class :character
##  Median :136.05             Mode  :character   Mode  :character
##  Mean   :134.49
##  3rd Qu.:151.87
##  Max.   :194.78
##    Division
##  Length:51
##  Class :character
##  Mode  :character
##
##
##
```

We already learned about the data structure and have gained the little idea of drivers data frame using str, head and tail function. In order to get a better idea of the distribution of your variables in the dataset I have used summary function to know about the descriptive statistic of each variable in data set. As we can see summary function telling us about the mean, median or range of the numerical variables. I can use appropriate plots with correct range.

My plan is to use the different plots like state vs Number of fatal collision plot which will give me understanding about which state has most collision and which has minimum collision for every billion miles traveled.

Once i learned about the fatal collision, I want to see which states in United States has high insurance premium. It will give my knowledge How car insurance premium works in different states?

Since we added the region data in our final data set, it would be interesting to see which region has most fatal collision.

After gaining knowledge about the fatal collision and Insurance premium in different states, my priority shift to understand if there is a any relationship between the two for which i will use plots to see the distribution of driver_fatalities and Car Insurance Premium.

I am planning to see the fatal collision relationship with region, alcohol percent, and Insurance company losses.

To have better understanding about the premium i am planning to plot Insurance company losses to see which states has high losses in United states.

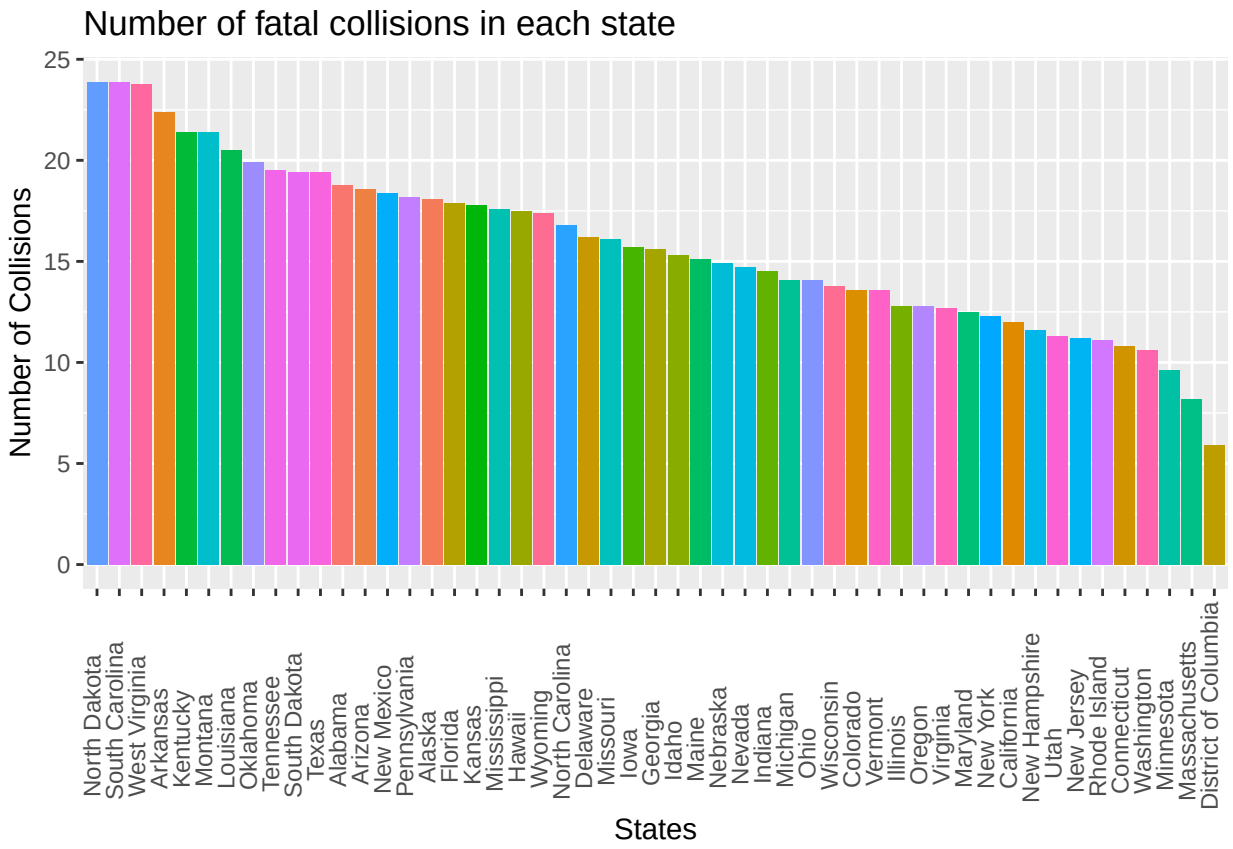
Lastly I want to see which region has most speeding ticket and drivers with alcohol percent.

Analysis

state vs Number of fatal collision

I have analyzed the first plot state vs Number of fatal collision which gave me the states with maximum collision North Dakota and South Carolina with fatal collision of 23.9 and state with less collision District of Columbia with collision of 5.9 for every billion miles traveled.

```
library(ggplot2)
ggplot(bad_drivers_df,aes(x=reorder(State,- driver_fatalities),
                          y = driver_fatalities, fill=State) )+
  geom_bar(stat = "identity")+
  xlab("States")+
  ylab("Number of Collisions")+
  ggtitle("Number of fatal collisions in each state")+
  guides(fill = FALSE) +
  theme(axis.text.x=element_text(angle=90,hjust=0.2,vjust=0.2))
```




```
summary(bad_drivers_df$driver_fatalities)
```

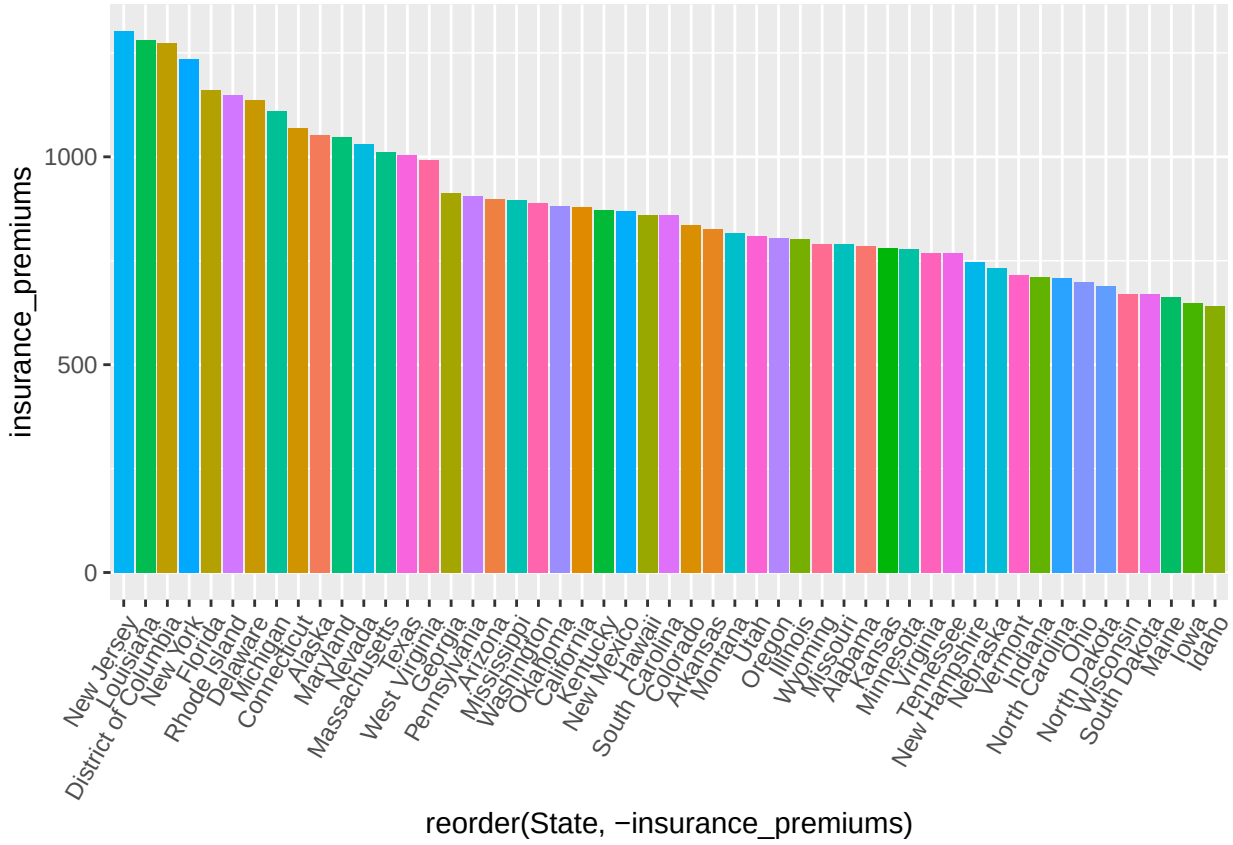
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.90   12.75   15.60   15.79   18.50   23.90
```

I have used the summary function to find out the average of driver fatalities and found that the fatal collision count in state North Dakota and South Carolina is higher then the average collision.

State vs car Insurance Premium

Used the ggplot to see Car Insurance premium of all 51 states of United State.It gave me understanding about How car insurance premium works in different states? Idaho state has less car insurance premium of 642 where as the New Jersey state has highest car insurance premium 1301.5.

```
bad_drivers_df %>% ggplot(aes(x=reorder(State, -`insurance_premiums`),
                              y=`insurance_premiums`, fill=State)) +
  geom_bar(stat = "identity") +
  guides(fill = FALSE) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```
summary(bad_drivers_df$insurance_premiums)
```

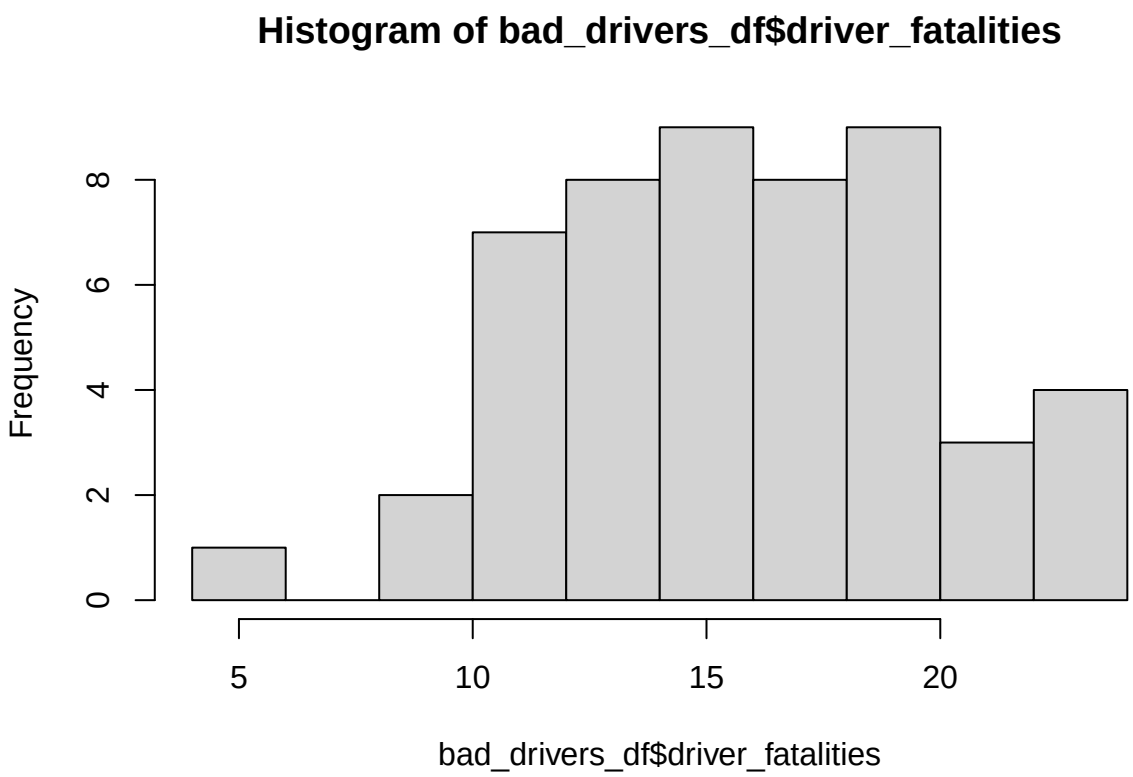
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   642.0   768.4   859.0   887.0  1007.9  1301.5
```

New Jersey state has car insurance premium is higher then the average of 859.

Distribution of driver_fatalities and Car Insurance Premium.

Histogram of Driver Fatalities

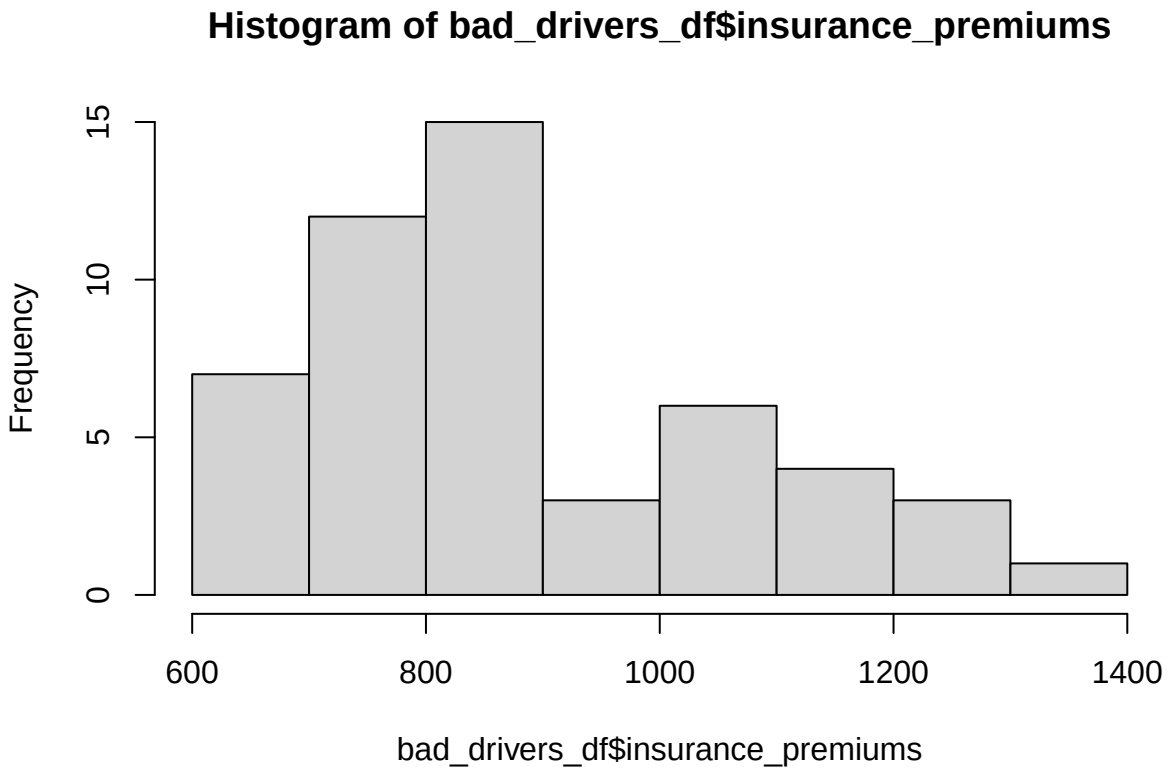
```
hist(bad_drivers_df$driver_fatalities)
```



Histogram of Insurance preimum

With the help of histogram I found that District of Columbia is an outlier with less number of collision and the model distribution is bi-model and slightly left skewed.

```
hist(bad_drivers_df$insurance_premiums)
```

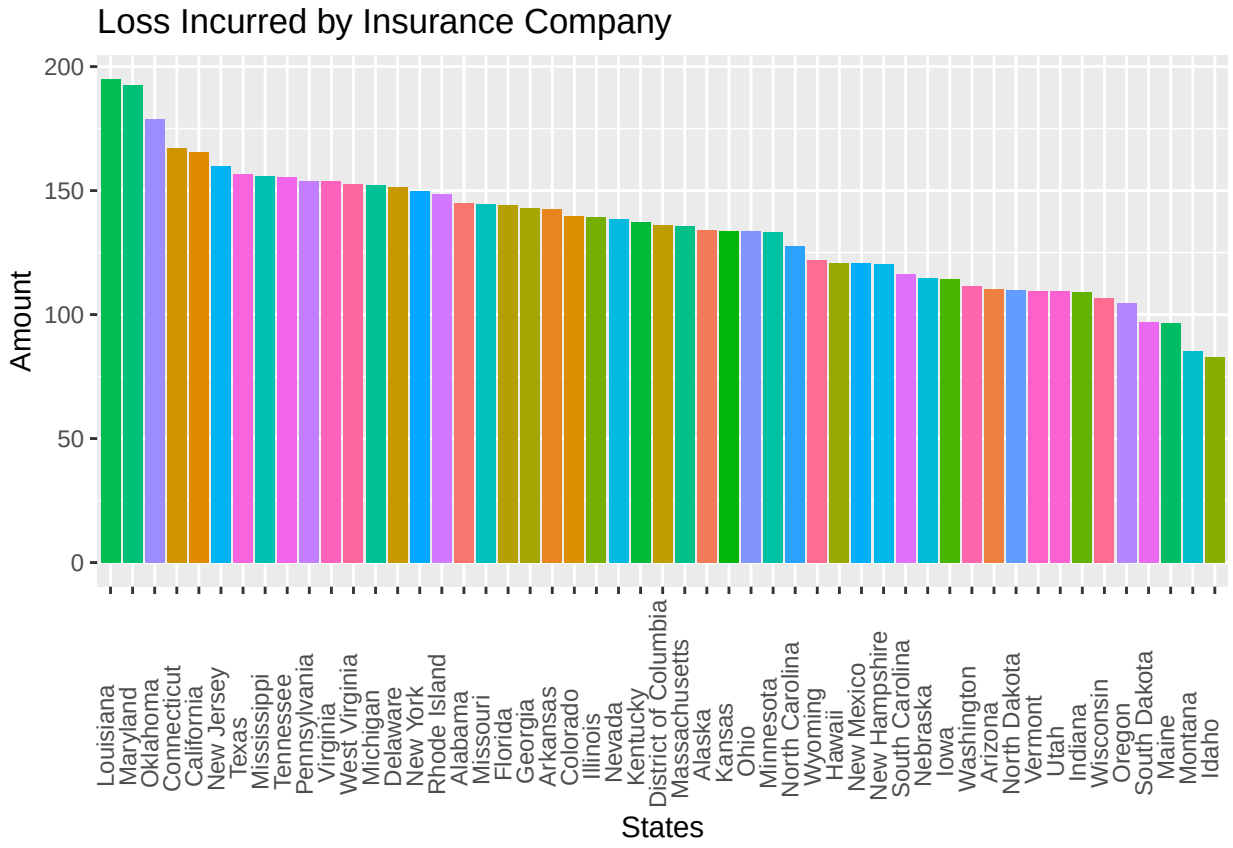


Whereas the the distribution of car insurance premium is skewed right and unimodal.

State vs Insurance company losses

Using below plot we found that Louisiana state has most expensive losses incurred by insurance company of 194.78 where as Idaho state has less losses incurred by insurance company of 82.75.

```
ggplot(bad_drivers_df,aes(x=reorder(State,- insurance_companies_losses),
                          y= insurance_companies_losses, fill=State) )+
  geom_bar(stat = "identity")+
  xlab("States")+
  ylab("Amount")+
  ggtitle("Loss Incurred by Insurance Company")+
  guides(fill = FALSE) +
  theme(axis.text.x=element_text(angle=90,hjust=0.2,vjust=0.2))
```




```
summary(bad_drivers_df$insurance_companies_losses)
```

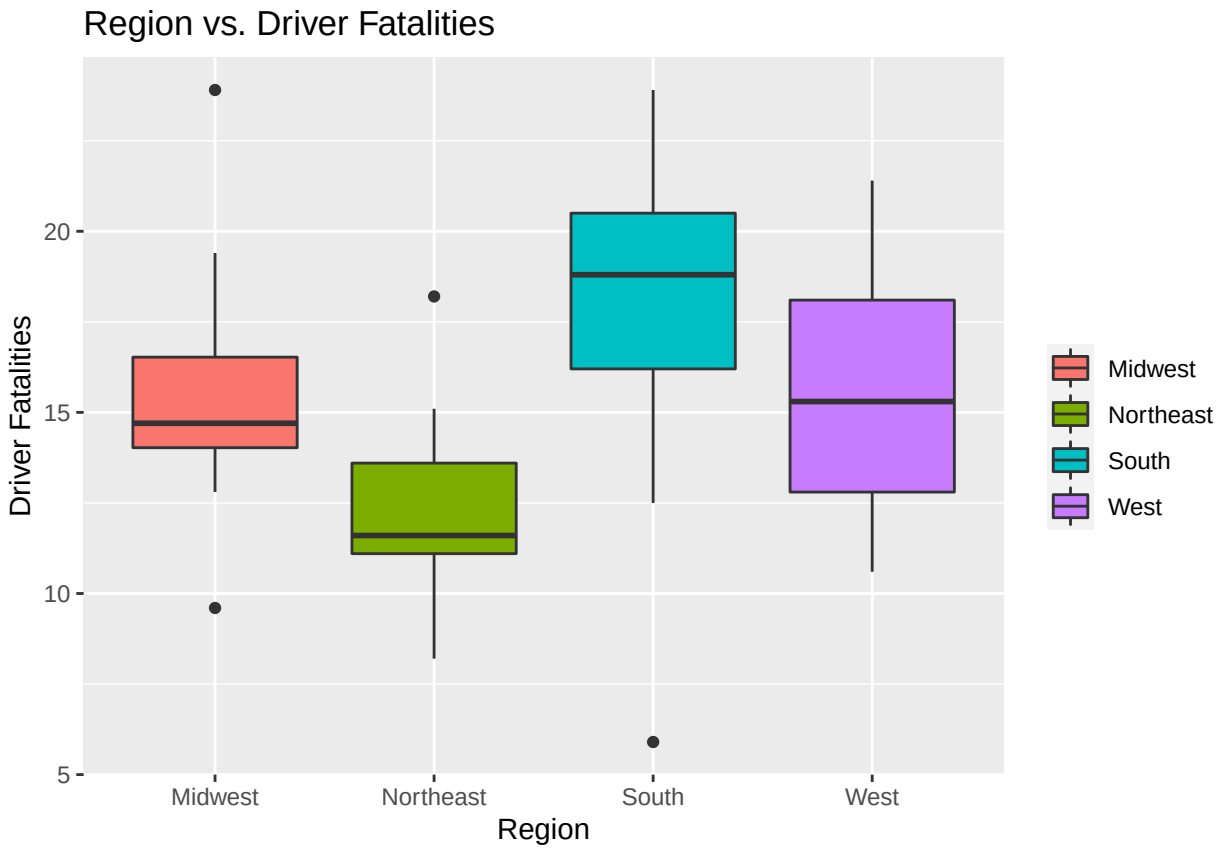
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   82.75  114.64  136.05  134.49  151.87  194.78
```

Louisiana state has losses incurred by insurance company of 194.78 which is more than the average losses incurred of 136.05.

Driver Fatalities by Region

I am using the boxplot to see how the data spread out from the center.

```
ggplot(bad_drivers_df, aes(x = Region, y = driver_fatalities, fill = Region)) +
  geom_boxplot() +
  labs(x = "Region", y = "Driver Fatalities",
       title = "Region vs. Driver Fatalities",
       fill = "")
```



We can clearly stat that the southern region has the highes median compare to Midwest, west and Northeast.

Relationship between Fatal Collision and Insurance Premium

I used Linear Regression model to understand the relationship between the fatal collision and insurance premium. I concluded from this model that driver collision is strongly associated with Car insurance premium as they showing low p value. Driver fatalities estimate value is -8.638 which show how it Car Insurance related and how much car Insurance will get effected with fatal collision.

```
lm_df <- lm(formula = insurance_premiums ~ driver_fatalities ,
            data = bad_drivers_df)
summary(lm_df)
```

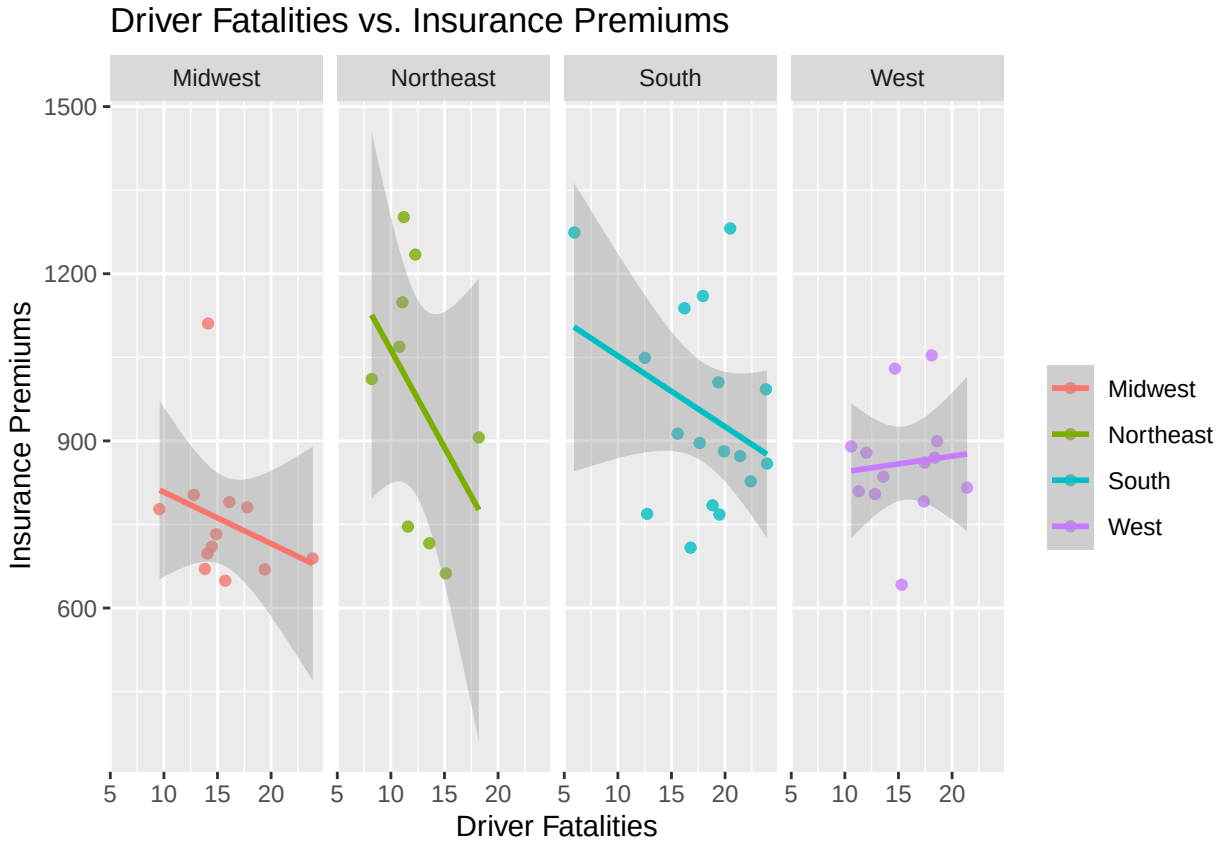
```
##
## Call:
## lm(formula = insurance_premiums ~ driver_fatalities, data = bad_drivers_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.23 -136.43  -22.29  133.45  435.28
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1023.354     98.748  10.363 6.08e-14 ***
## driver_fatalities   -8.638      6.055  -1.427     0.16
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 176.5 on 49 degrees of freedom
## Multiple R-squared:  0.03988,    Adjusted R-squared:  0.02029
## F-statistic: 2.035 on 1 and 49 DF,  p-value: 0.16
```

Fata Collision and Insurance premium by Region

We can also use scatterplot to understand the relationship between the fatal collision and Insurance premium. Here I am using the region to see the relationship based on region.

```
ggplot(bad_drivers_df, aes(x = driver_fatalities,
                           y = insurance_premiums, col = Region)) +
  geom_jitter(alpha = 0.8) + geom_smooth(method = "lm") +
  facet_grid(. ~ Region) +
  labs(x = "Driver Fatalities", y = "Insurance Premiums",
       title = "Driver Fatalities vs. Insurance Premiums",
       col = "")
```



We can see from above plot that South and Northeast has the highest Car Insurance premium and fatal collision are negatively related to Insurance premium in Midwest, Northeast and South. Both variables has the positive relation only in West region.

Fatal collision with region, alcohol percent, and Insurance company losses Model

Let build a model to see relationship between fatal collision with region, alcohol percent, and Insurance company losses.

```
model <- lm(driver_fatalities ~ Region + alcohol_percent +
              insurance_companies_losses, data = bad_drivers_df)

summary(model)
```

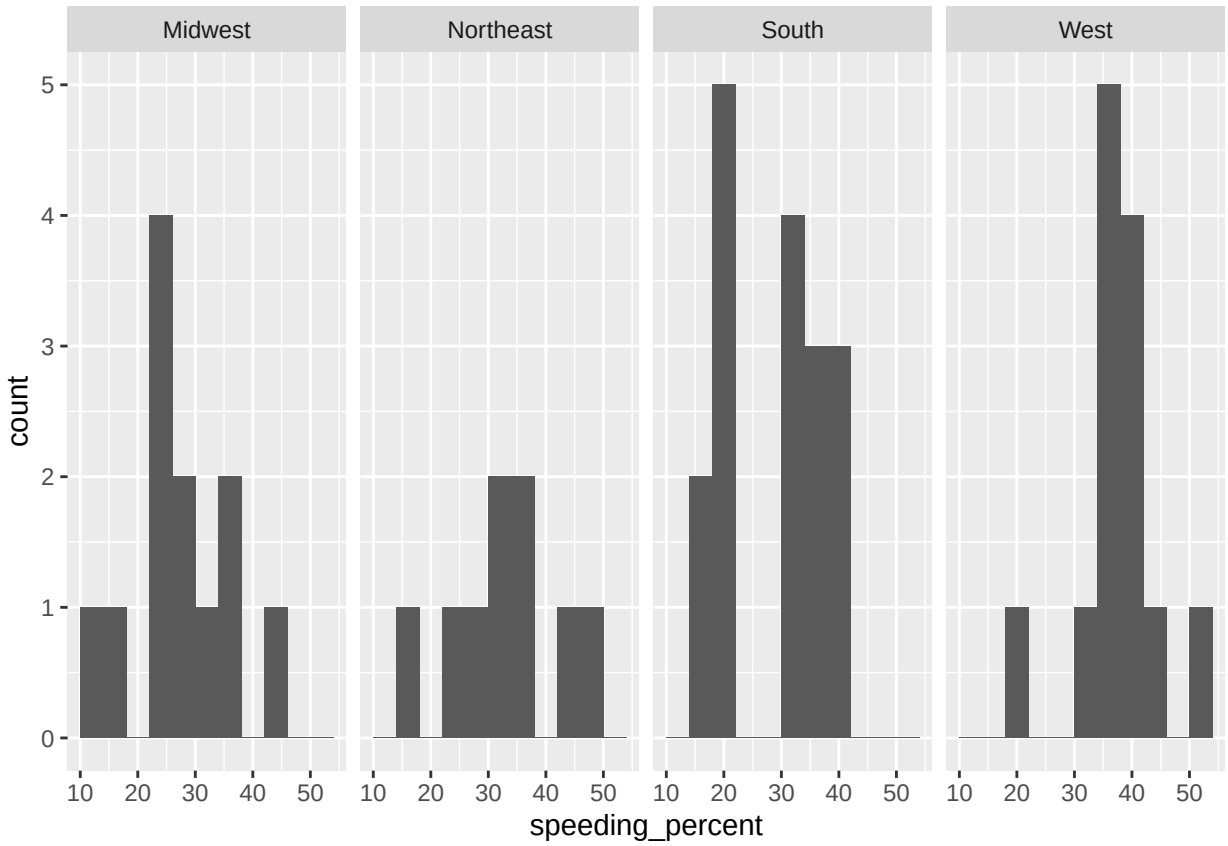
```
##
## Call:
## lm(formula = driver_fatalities ~ Region + alcohol_percent + insurance_companies_losses,
##     data = bad_drivers_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9769  -1.6737  -0.1061   2.1819   6.4751
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.34846    4.73461   2.819  0.00713 **
## RegionNortheast             -2.66848    1.62967  -1.637  0.10852
## RegionSouth                  3.68866    1.53391   2.405  0.02036 *
## RegionWest                   0.13876    1.46238   0.095  0.92483
## alcohol_percent              0.20464    0.10135   2.019  0.04944 *
## insurance_companies_losses -0.03444    0.02495  -1.380  0.17428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.61 on 45 degrees of freedom
## Multiple R-squared:  0.3099, Adjusted R-squared:  0.2332
## F-statistic: 4.041 on 5 and 45 DF,  p-value: 0.004098
```

we can clearly stat that region south with 4.33 estimate has major impact on driver fatalities where as region northeast has least impact on driver fatalities indicating that this region has good drivers compare to south region of united states.

Speeding percent by Region

I have also analyzed the speeding percent in each region and found that South and West region has more rate of getting speeding ticket compare to other region.

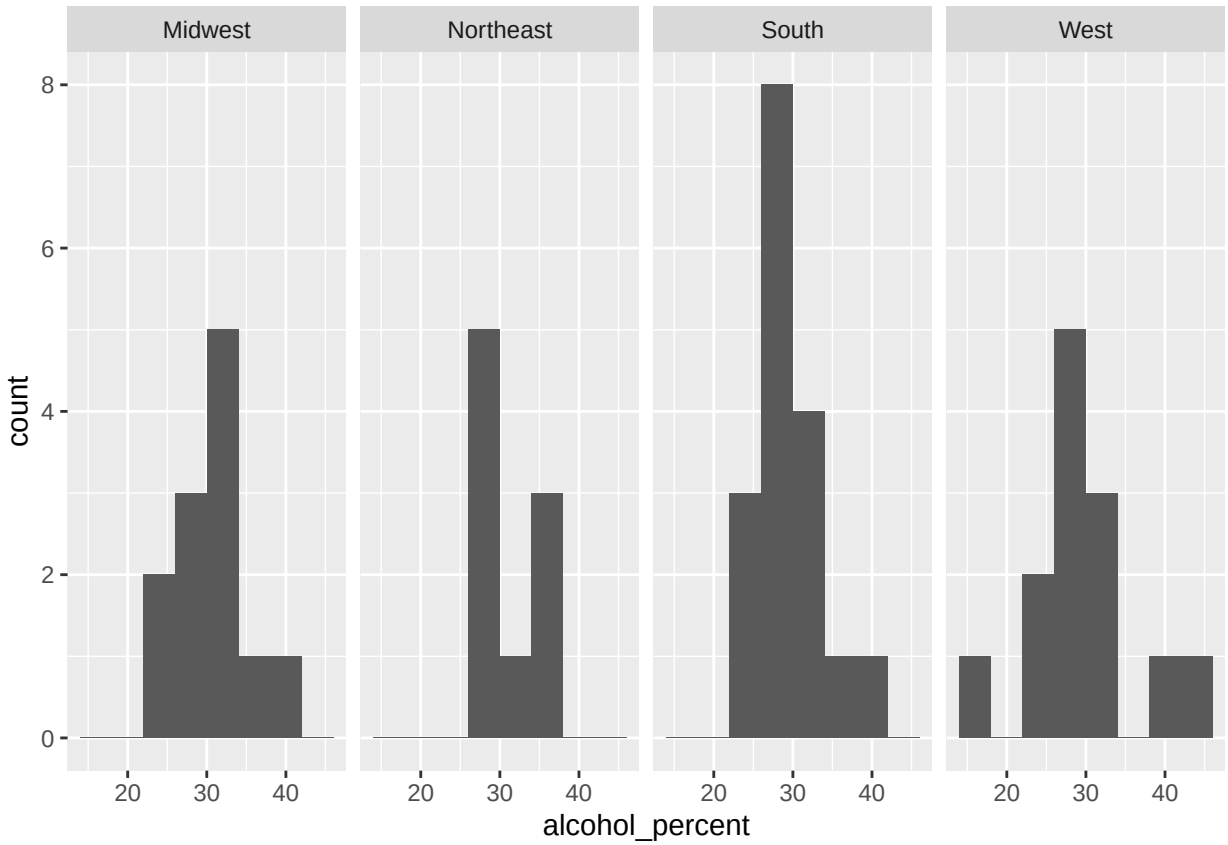
```
ggplot(bad_drivers_df, aes(x = speeding_percent)) +
  geom_histogram(binwidth = 4) + facet_grid(. ~ Region)
```



Alcohol Percent by Region

Lets learn about which region has the drivers with alcohol percent.

```
ggplot(bad_drivers_df, aes(x = alcohol_percent)) +
  geom_histogram(binwidth = 4) + facet_grid(. ~ Region)
```



As we can see the south region has most drivers with alcohol percent.

Implications

After analyzing the different cases above we can state that there is a relationship between Car Insurance and fatal collision. Idaho state has the good drivers with low Car Insurance premium, also the losses incurred by Insurance company is also low in Idaho compare to all other state. Where as North Dakota, South Carolina, New jersey and Louisiana has worst drivers in United State. Overall if we look the data via region we have found that the South region has worst drivers in United States where as NorthEast region has good drivers.

Limitations

Using the current data set we cannot say anything about the road condition of each state. Also there is no evident about the weather condition when collision happened. These two factors plays important role when comes to road accident as they can also be the reason of collision and if so in that case we cannot say drivers are bad.

Concluding Remarks

I have analyzed driver_fatalities, insurance_premiums and insurance_companies_losses, region variables of data set.Along with that I have visualized the data to see the number of car crashes in each state, the insurance premiums in each state, fatal collision in each region, Relationship between Car Insurance Premium and fatal collision, how much insurance company pay out?, speeding percent in each region and lastly the drivers with alcohol percent in all region. After learning about all we can conclude that we have good drivers in Idaho as its Car Insurance premium is low , also the losses incurred by Insurance company is also low

compare to all other state. On the other hand North Dakota, South Carolina, New jersey and Louisiana states has worst drivers in United State. Overall we can say that South region has the worst drivers in United States.

References

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using r*. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.

Lander, J. P. 2014. *R for Everyone: Advanced Analytics and Graphics*. Addison-Wesley Data and Analytics Series. Addison-Wesley. https://books.google.com/books?id=3eBVAgAAQBAJ.