Final Project Step 2

Dipika Sharma

May 23, 2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Add Citations

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

How to import and clean my data?

For the Data Acquisition step, I am gathering data from link https://raw.githubusercontent.com/fivethirtyeight/data/master/bad-drivers/bad-drivers.csv using the read.csv.

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

We can learn 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....
```

As we can see we have 8 different columns for 51 states of United State. I will rename the column as it is currently very long and it is always preferable to use small and meaningfull name of the columns.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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.Bee
         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 ...
```

What does the final data set look like?

To have the better idea we can look at some of the data using head function. Also check the complete data to see if we have any nulls in the data. for now I checked already we do not have any NULLS so we are fine.

```
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
```

Questions for future steps.

I am planning to cover following questions in future steps:

1. Which states have the bad drivers?
2. Which State has best drivers?
3. Which States has maximum fata collision and which has less?
4. Which states has more car Insurance premium and which has less?
5. Which States Insurance company incurred losses?

What information is not self-evident?

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. There is also no information about the region in a data set. I think it would be interesting to see which region has worst drivers?

What are different ways you could look at this data?

I am planning to look at every aspect of the data to see and understand which states has worst driver in United States. I will try to use each column to see its relations with collision in states. With the help of different plots and function available in R, it will be easier for me to get information from data.

How could you summarize your data to answer key questions?

```
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 ...
```

Using the str function I can see the structure of the bad drivers data set, it is giving me the idea about the classes of the 8 variable and the number of obseravtion is 51

```
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
```

Head function will give me better idea about the data stored in 8 variables.

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

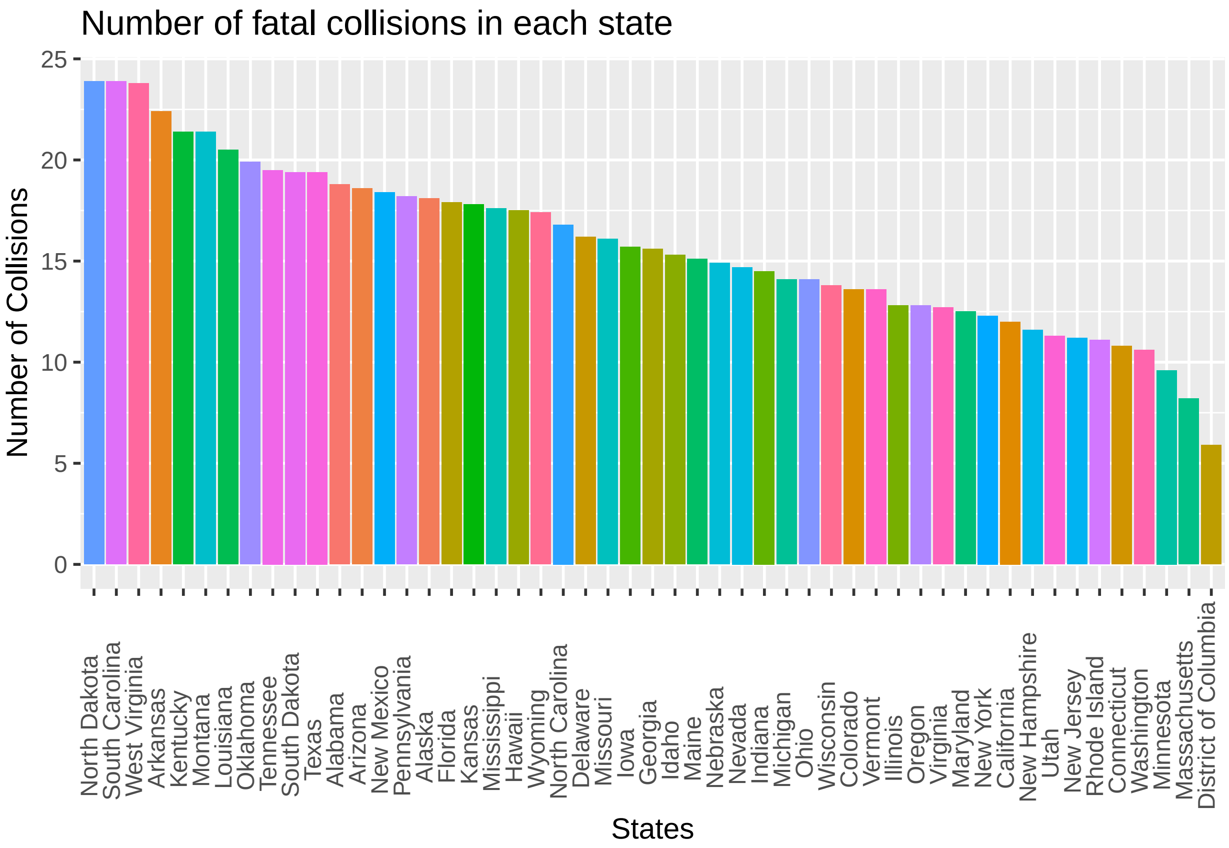
We already saw using str and head fucntion that bad drivers dataframe is 51 observations of the 8 variables. Now in order to get a better idea of the distribution of your variables in the dataset I am using the summary function. it will give me the descriptive statistic of each variable in data set. By looking at the function output we can see the mean, median or range of the numerical variables this will give us better understanding of what plots I can use, what will be the range and others.

What types of plots and tables will help you to illustrate the findings to your questions?

state vs Number of fatal collision

The first plot I want to analyze is state vs Number of fatal collision, It will tell me which state has most collision and which has minimum collision for every billion miles traveled

```
library(ggplot2)
ggplot(baddrivers_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))
```



Looking at the graph we can clearly see that North Dakota and South Carolina have most fatal collision of 23.9 where as District of Columbia has less collision of 5.9 for every billion miles traveled.

```
summary(baddrivers_df$driver_fatalities)
```

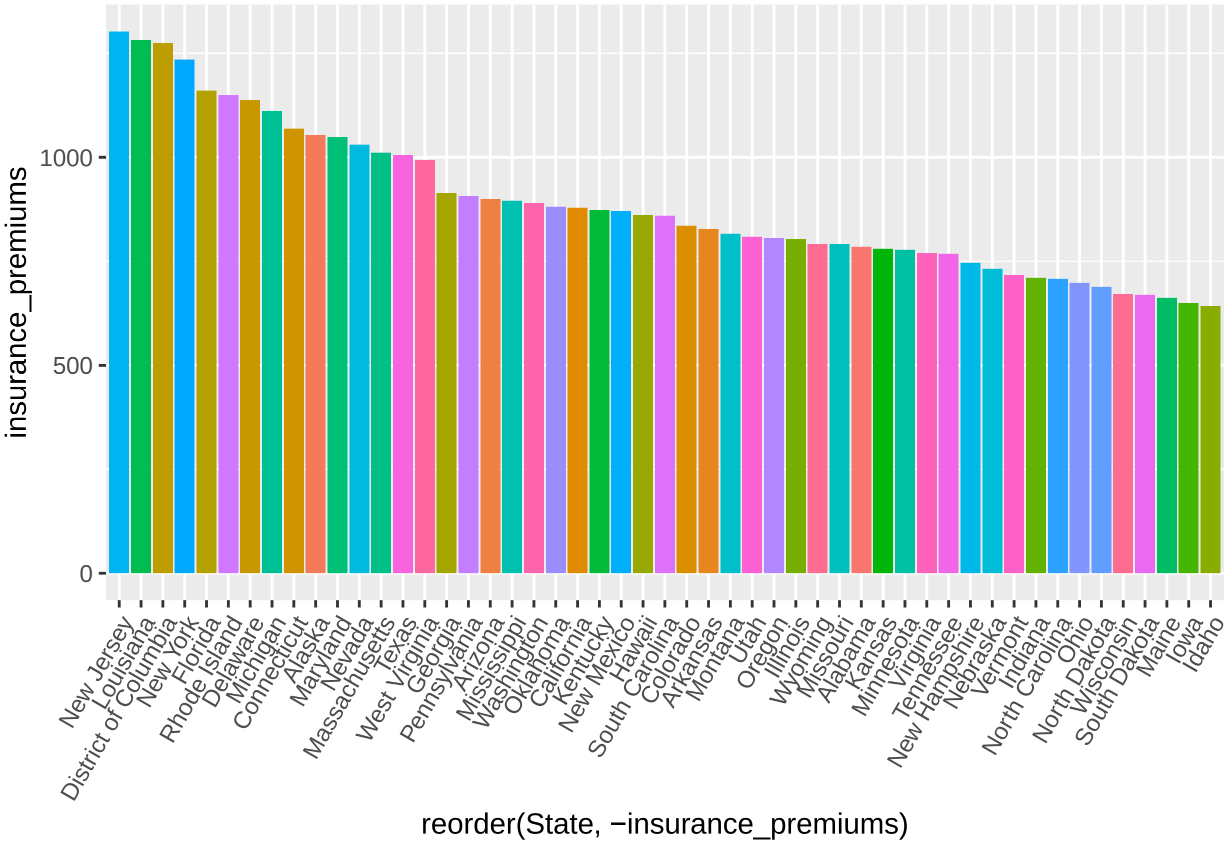
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.90   12.75   15.60   15.79   18.50   23.90
```

Looking at the average of driver fatalities we can see the fatal collision count in state North Dakota and South Carolina is higher then the average collision.

State vs car Insurance Premium

I am using ggplot to see Car Insurance premium of all 51 states of United State. It will give us the blink of data and also the understanding of car insurance premium works in different states.

```
library(ggplot2)
baddrivers_df %>% ggplot(aes(x=reorder(State, -`insurance_premiums`), y=`insurance_premiums`, fill=Stat
  geom_bar(stat = "identity") +
  guides(fill = FALSE) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



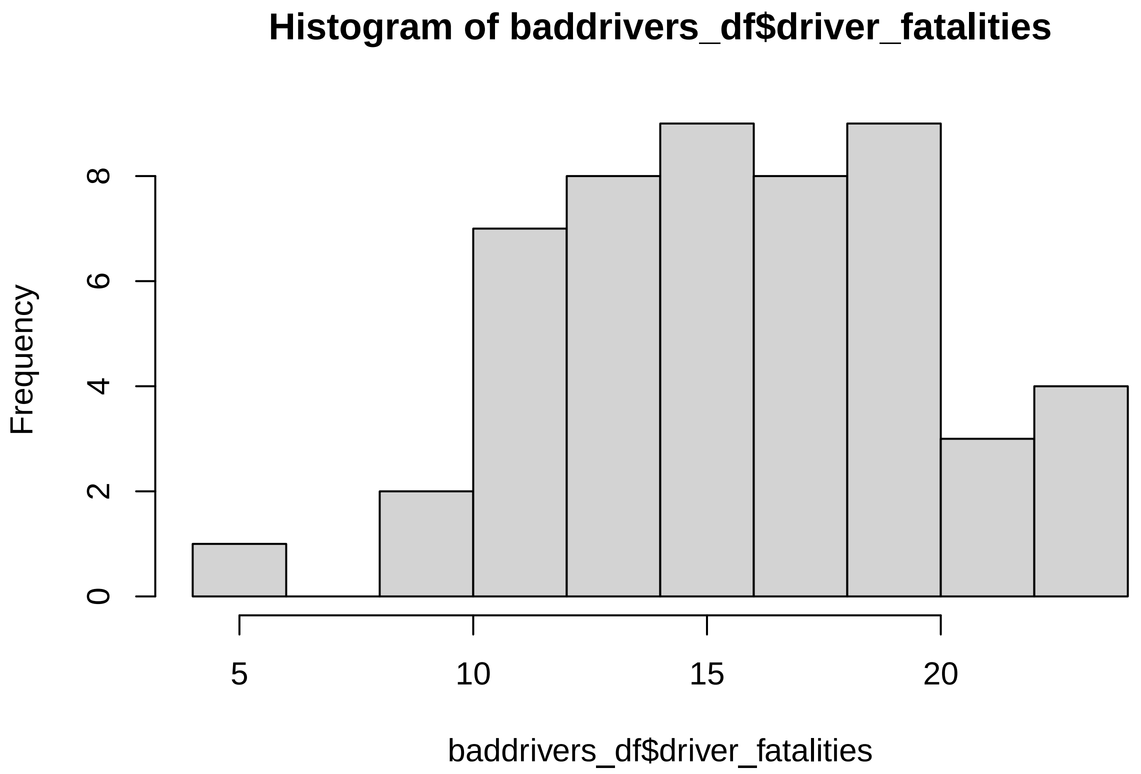
```
summary(baddrivers_df$insurance_premiums)
```

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

Looking at the plot we can stat that Idaho state has less car insurance premium of 642 where as the New Jersey state has highest car insurance premium 1301.5 which is higher from the average i.e. 859.

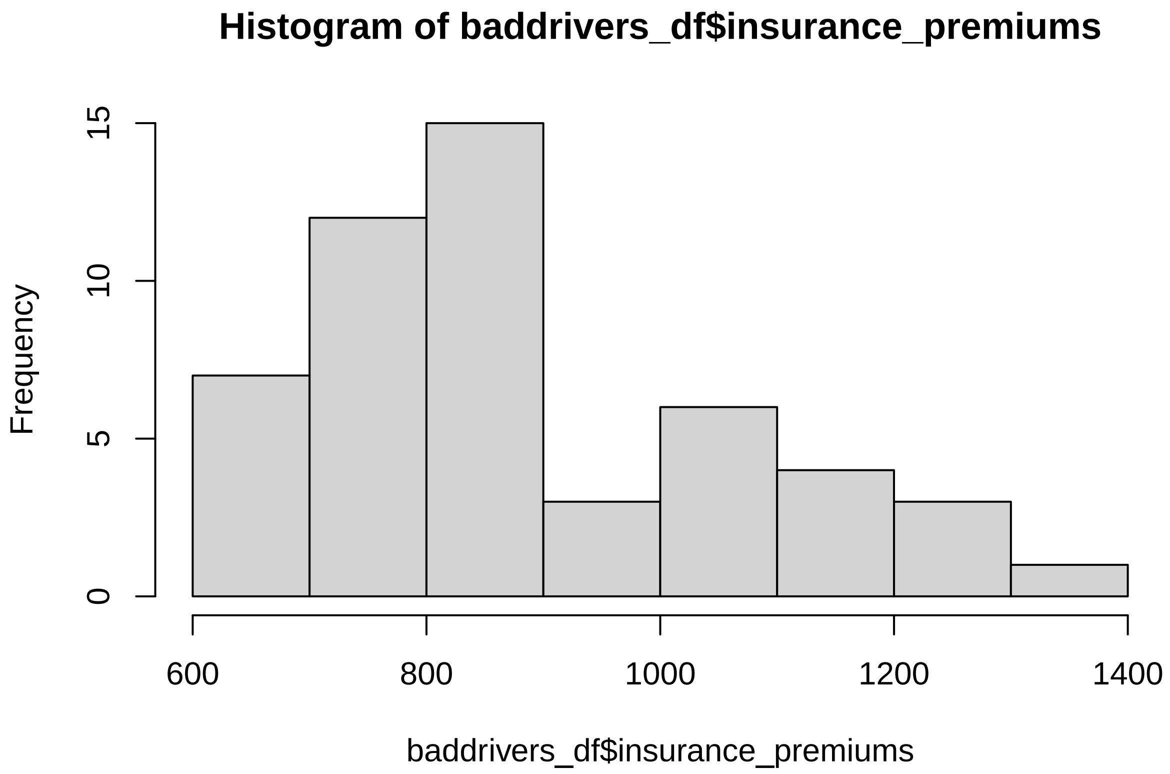
Distribution of driver_fatalities and Car Insurance Premium.

```
hist(baddrivers_df$driver_fatalities)
```



Looking at the above histogram we can see that the distribution bi-model, slightly left skewed. Also we can see the that District of Columbia is an outlier with less number of collision.

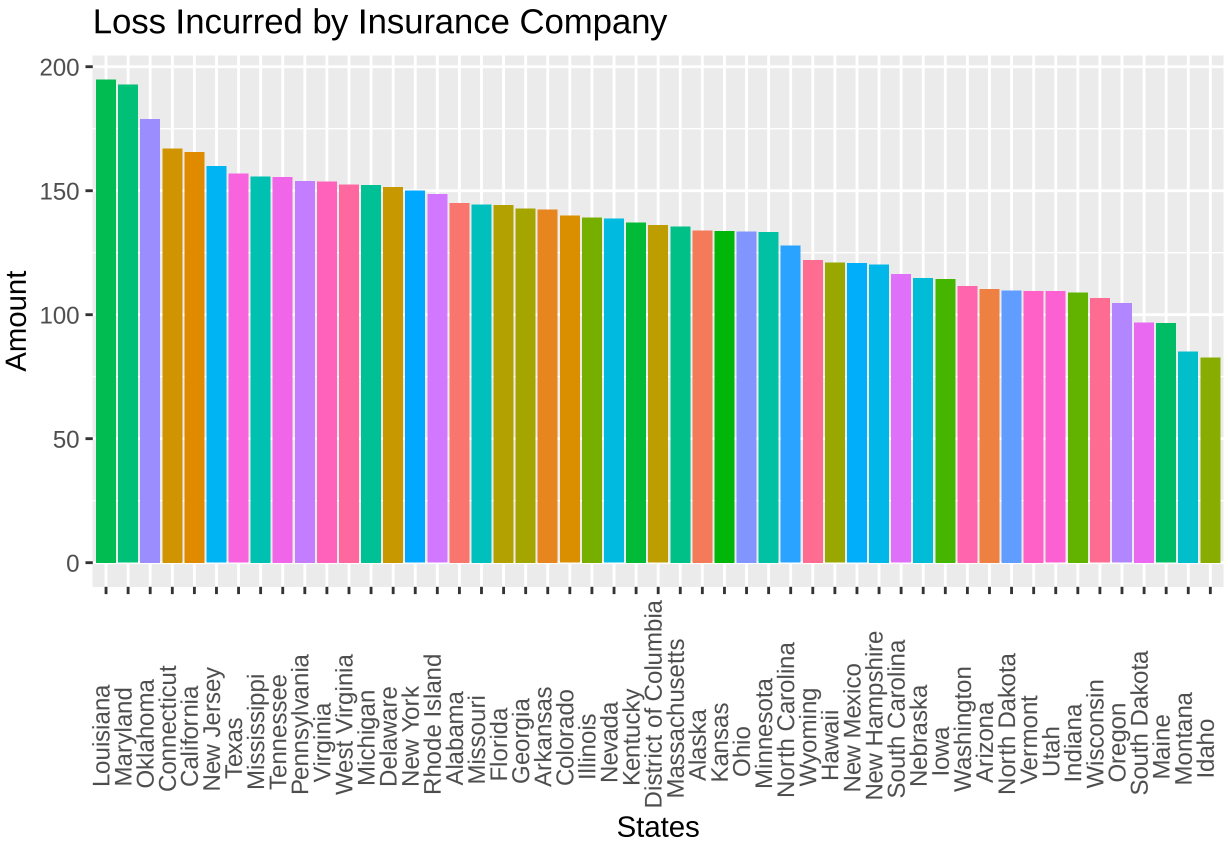
```
hist(baddrivers_df$insurance_premiums)
```



This distribution of car insurance premium is skewed right and unimodal.

State vs Insurance company losses

```
ggplot(baddrivers_df,aes(x=reorder(State,- insurance_companies_losses), y= insurance_companies_losses, 
  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(baddrivers_df$insurance_companies_losses)
```

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

The above plot stat 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. The average losses incurred is 136.05.

Lets try to understand if we have any relationship between fatal collision and insurance premium. For this I am using the linear regression model.

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

```
##
## Call:
## lm(formula = insurance_premiums ~ driver_fatalities, data = baddrivers_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
```

By looking at the summary of linear regression model we can say the driver collision is strongly associated with Car insurance premium as they shwing low p value and also the 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.

Since there is a relationship between Car Insurance and fatal collision we can say 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 state Where as North Dakota, South Carolina, New jersey and Louisiana has worst drivers in United State.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes I have used the linear regression to find out the relationship between the continuous variable, one is dependent variable and other is independent variable. In our case the dependent variable is Car Insurance premium and independent variable is Driver fatalities. This relationship help me to achieve my goal of getting answer to my question which state has worst drivers?

Questions for future steps.

I would love to see the given data by region to find out which region of United States has worst drivers and which has best drivers? Also it would be interesting to see how the weather condition and road condition can change my analysis. So for future research I want to used region dataframe along with dataframes which has weather and roads information in it.

Refrences

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.