Final Project Data 606

Jaya Veluri

12/7/2021

INTRODUCTION

Car insurance premiums are varied among different states in America. Why are some states having higher premiums than others? In this project, I wanted to study the association between driver records and car insurance premiums. It's good to know the percentage of bad drivers and your potential car insurance premium before you are planning to move to another state.

Data

The data is collected in 2017 from National Highway Traffic Safety Administration and National Association of Insurance Commissioners by FiveThirtyEight.

There are 51 cases/observations in the given data set(including District of Columbia). Each case represents a state in the United States. I will be studying the Insurance Loss and Insurance Premium variables.

Insurance Loss is the independent variable that is quantitative and Insurance Premium is the response variable that is also quantitative.

This is an observational study. I will draw my conclusions based on analyzing the existing data.

The population of interest is anyone who drives in the United States, hence this study can be generalized to the general population.

The data may or may not be used to establish causal links between Insurance Loss and Insurance Premiums(even if there is a relationship between the two), since there could be more factors taken into consideration when the premiums were being priced.

```
library(ggplot2)
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
```

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

head(bad_drivers)
```

```
##        State Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
## 1    Alabama                                                              18.8
## 2     Alaska                                                              18.1
## 3    Arizona                                                              18.6
## 4   Arkansas                                                              22.4
## 5 California                                                              12.0
## 6   Colorado                                                              13.6
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
## 1                                                                  39
## 2                                                                  41
## 3                                                                  35
## 4                                                                  18
## 5                                                                  35
## 6                                                                  37
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
## 1                                                                          30
## 2                                                                          25
## 3                                                                          28
## 4                                                                          26
## 5                                                                          28
## 6                                                                          28
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
## 1                                                                         96
## 2                                                                         90
## 3                                                                         84
## 4                                                                         94
## 5                                                                         91
## 6                                                                         79
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accid
## 1
## 2
## 3
## 4
## 5
## 6
##   Car.Insurance.Premiums....
## 1                     784.55
## 2                    1053.48
## 3                     899.47
## 4                     827.34
## 5                     878.41
## 6                     835.50
##   Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
## 1                                                                       145.08
## 2                                                                       133.93
## 3                                                                       110.35
## 4                                                                       142.39
```

```
## 5                                                                    165.63
## 6                                                                    139.91
```

```
names(bad_drivers) <- c("State", "Drivers.Collisions",
"Drivers.Perc.Speeding", "Drivers.Perc.Alcohol",
"Drivers.Perc.Not.Distracted", "Drivers.Perc.No.Pre.Accident",
"Drivers.Insurance.Premium", "Drivers.Insurance.Loss")
```

Data Analysis

```
datatable(bad_drivers)
```

```
summary(bad_drivers)
```

```
##     State           Drivers.Collisions Drivers.Perc.Speeding
##  Length:51          Min.   : 5.90      Min.   :13.00
##  Class :character   1st Qu.:12.75      1st Qu.:23.00
##  Mode  :character   Median :15.60      Median :34.00
##                     Mean   :15.79      Mean   :31.73
##                     3rd Qu.:18.50      3rd Qu.:38.00
##                     Max.   :23.90      Max.   :54.00
##  Drivers.Perc.Alcohol Drivers.Perc.Not.Distracted Drivers.Perc.No.Pre.Accident
##  Min.   :16.00        Min.   : 10.00              Min.   : 76.00
##  1st Qu.:28.00        1st Qu.: 83.00              1st Qu.: 83.50
##  Median :30.00        Median : 88.00              Median : 88.00
##  Mean   :30.69        Mean   : 85.92              Mean   : 88.73
##  3rd Qu.:33.00        3rd Qu.: 95.00              3rd Qu.: 95.00
##  Max.   :44.00        Max.   :100.00              Max.   :100.00
##  Drivers.Insurance.Premium Drivers.Insurance.Loss
##  Min.   : 642.0            Min.   : 82.75
##  1st Qu.: 768.4            1st Qu.:114.64
##  Median : 859.0            Median :136.05
##  Mean   : 887.0            Mean   :134.49
##  3rd Qu.:1007.9            3rd Qu.:151.87
##  Max.   :1301.5            Max.   :194.78
```

Summary statistics for characteristics of bad drivers

```
summary(bad_drivers)
```

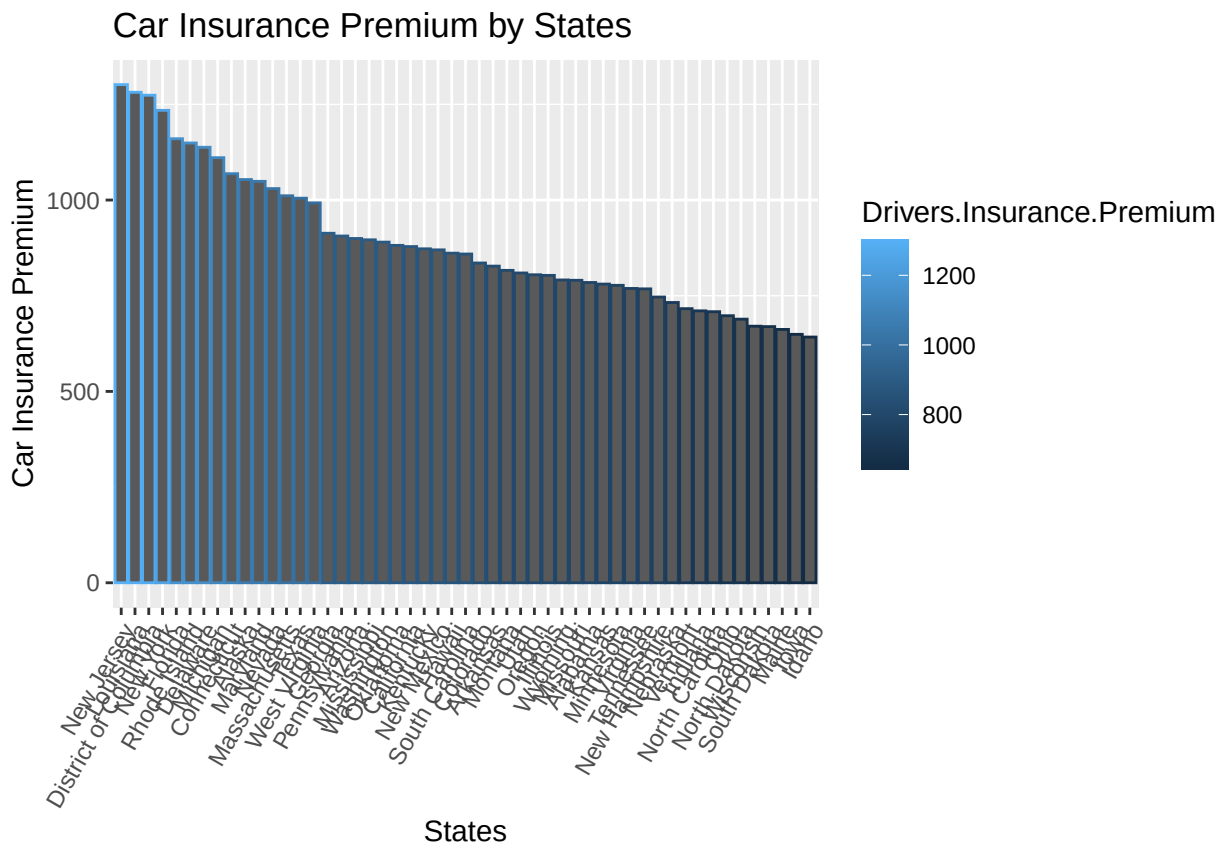
```
##     State           Drivers.Collisions Drivers.Perc.Speeding
##  Length:51          Min.   : 5.90      Min.   :13.00
##  Class :character   1st Qu.:12.75      1st Qu.:23.00
##  Mode  :character   Median :15.60      Median :34.00
##                     Mean   :15.79      Mean   :31.73
##                     3rd Qu.:18.50      3rd Qu.:38.00
##                     Max.   :23.90      Max.   :54.00
##  Drivers.Perc.Alcohol Drivers.Perc.Not.Distracted Drivers.Perc.No.Pre.Accident
```

```
##  Min.   :16.00        Min.   : 10.00             Min.   : 76.00
##  1st Qu.:28.00        1st Qu.: 83.00             1st Qu.: 83.50
##  Median :30.00        Median : 88.00             Median : 88.00
##  Mean   :30.69        Mean   : 85.92             Mean   : 88.73
##  3rd Qu.:33.00        3rd Qu.: 95.00             3rd Qu.: 95.00
##  Max.   :44.00        Max.   :100.00             Max.   :100.00
##  Drivers.Insurance.Premium Drivers.Insurance.Loss
##  Min.   : 642.0            Min.   : 82.75
##  1st Qu.: 768.4            1st Qu.:114.64
##  Median : 859.0            Median :136.05
##  Mean   : 887.0            Mean   :134.49
##  3rd Qu.:1007.9            3rd Qu.:151.87
##  Max.   :1301.5            Max.   :194.78
```

Barplot to see car insurance premiums in all the states ranking from highest to lowest.

```
bad_drivers %>% ggplot(aes(x=reorder(State, -Drivers.Insurance.Premium),
  y=Drivers.Insurance.Premium, color=Drivers.Insurance.Premium)) +
  geom_bar(stat = "identity") +
  guides(fill = FALSE) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  xlab("States") + ylab("Car Insurance Premium") +
  ggtitle("Car Insurance Premium by States")
```

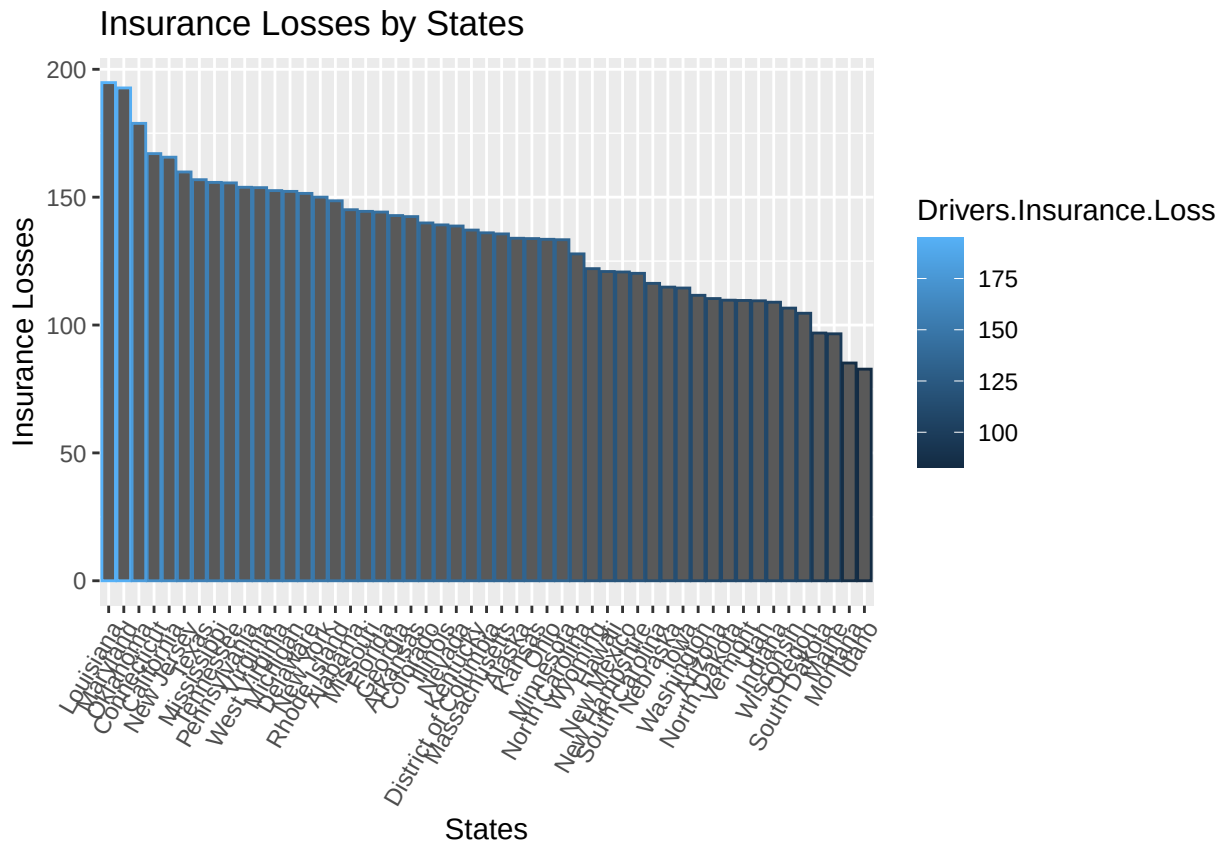
```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



Barplot to see insurance losses in all the states ranking from highest to lowest.

```
bad_drivers %>% ggplot(aes(x=reorder(State, -Drivers.Insurance.Loss),
 y=Drivers.Insurance.Loss, color=Drivers.Insurance.Loss)) +
  geom_bar(stat = "identity") +
  guides(fill = FALSE) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  xlab("States") + ylab("Insurance Losses") +
  ggtitle("Insurance Losses by States")
```

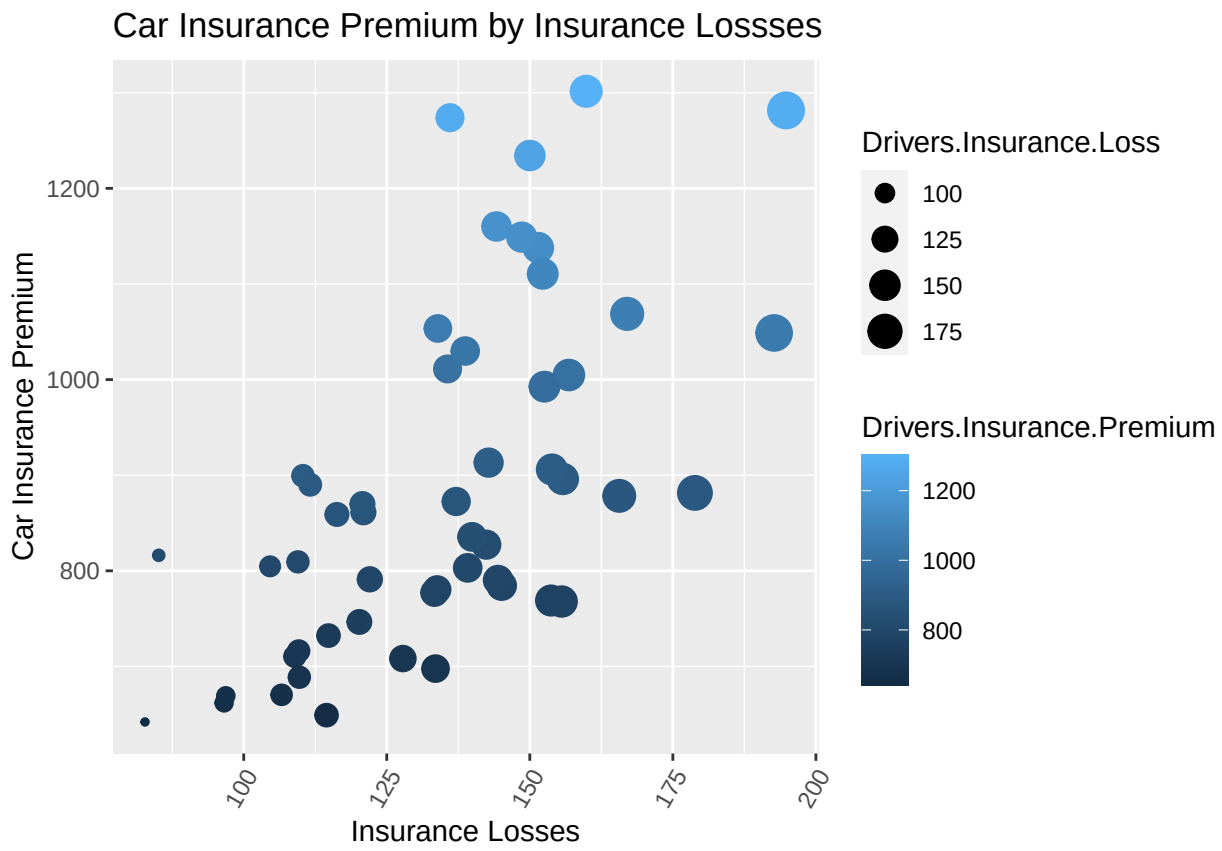
```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



Relationship between insurance losses and premiums from a scatterplot.

```
bad_drivers %>% ggplot(aes(x=Drivers.Insurance.Loss,
    y=Drivers.Insurance.Premium, color=Drivers.Insurance.Premium,
    size=Drivers.Insurance.Loss)) +
  geom_point() +
  guides(fill = FALSE) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  xlab("Insurance Losses") + ylab("Car Insurance Premium") +
  ggtitle("Car Insurance Premium by Insurance Lossses")
```

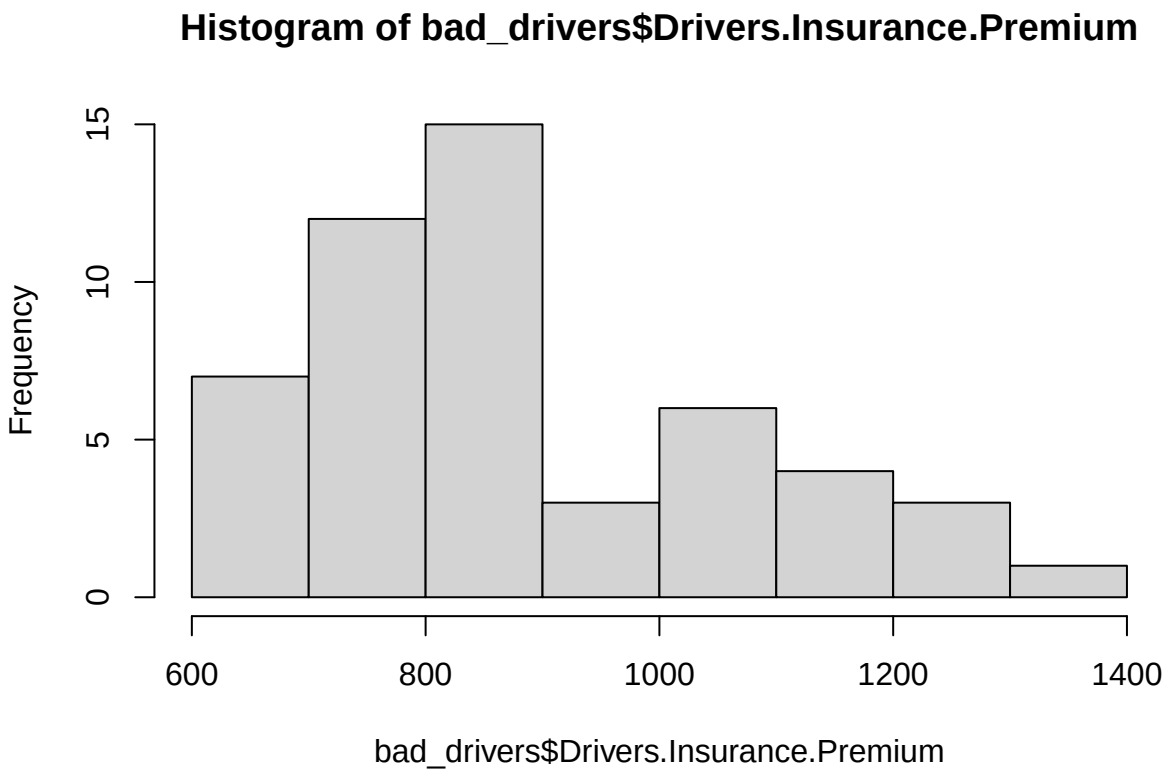
```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



There is an upward movement towards the upper right corner which indicates higher losses lead to higher premium

Inference

```
hist(bad_drivers$Drivers.Insurance.Premium)
```



Running a linear regression model:

```
m_loss <- lm(Drivers.Insurance.Premium ~ Drivers.Insurance.Loss, data = bad_drivers)
summary(m_loss)
```

```
##
## Call:
## lm(formula = Drivers.Insurance.Premium ~ Drivers.Insurance.Loss,
##     data = bad_drivers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213.33  -96.75  -40.11  112.24  379.97
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             285.3251   109.6689   2.602   0.0122 *
## Drivers.Insurance.Loss   4.4733     0.8021   5.577 1.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.9 on 49 degrees of freedom
## Multiple R-squared:  0.3883, Adjusted R-squared:  0.3758
## F-statistic:  31.1 on 1 and 49 DF,  p-value: 1.043e-06
```

The linear regression model suggests that the formula used to predict the insurance premium by loss is: premium = 285.33 + 4.47 * loss

Conclusion

From this study, I would conclude that there appears to be association between car insurance losses and insurance premiums. At this point we can only be certain that Idaho is the safest state which also has the lowest car insurance premium and New Jersey has the highest car Insurance Premium, (even though it ranks 6th in insurance losses). So in some states, additional to insurance losses there are other factors which effects car Insurance Premium.

Extending The Study

1) Is there a linear relationship between insurance premiums and percent of drivers caught drinking.
2) Is there a linear relationship between insurance premiums and percent of drivers caught speeding.
3) Which Region/state has the worst drivers and best drivers