Midterm Project 1

Colin Wang

Introduction

This R Markdown homework is for Math189 Midterm Project 1. Unlike homework 2 that focused more on Exploratory Data Analysis (EDA), in this project, I will provide an in-depth analysis for the relationship between weights of the car and its corresponding miles per gallon value, as well as this relationship's dependency on the number of cylinders. In this report, I will also justify my analysis, including its advantages as well as potential drawbacks. I will also include discussion relevent tools covered in the first sixth lectures, and provide reasons why some tools are useful and other tools are not. The dataset used in this project is extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).[1] The dataset is downloaded from course GitHub.[2]

Data Description [3]

1. **model**: the car model
2. **mpg**: Miles/(US) gallon
3. **cyl**: Number of cylinders
4. **disp**: Displacement (cu.in.)
5. **hp**: Gross horsepower
6. **drat**: Rear axle ratio
7. **wt**: Weight (lb/1000)
8. **qsec**: 1/4 mile time
9. **vs**: V/S
10. **am**: Transmission (0 = automatic, 1 = manual)
11. **gear**: Number of forward gears
12. **carb**: Number of carburetors

Body

Load Data

```r
df = read.csv('mtcars.csv', header = TRUE)
head(df)
```

```
##               model  mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

[1] Henderson, H., & Velleman, P. (1981). Building Multiple Regression Models Interactively. Biometrics, 37(2), 391-411. doi:10.2307/2530428 [Accessed: 22-Jan-2021]

[2] T. McElroy, "Ma189Homework2" [Online]. Available: https://github.com/tuckermcelroy/ma189/tree/main/Data [Accessed: 22-Jan-2021]

[3] Christian, Motor Trend Car Road Analysis, Dec-2014. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/51431_3323677bd16347fd983ba69d2aac5d64.html. [Accessed: 13-Jan-2021].

Analysis

Relationship between weight (wt) and miles per gallon (mpg)

Sample Mean and Variance

First, we want to see the basic information of our variables. Specifically, we want to find out a typical value for weight and miles per gallon, and how various they are. Therefore, we can use sample mean and variance to check.

```
mean(mtcars$wt)
```

```
## [1] 3.21725
```

```
var(mtcars$wt)
```

```
## [1] 0.957379
```

```
mean(mtcars$mpg)
```

```
## [1] 20.09062
```

```
var(mtcars$mpg)
```

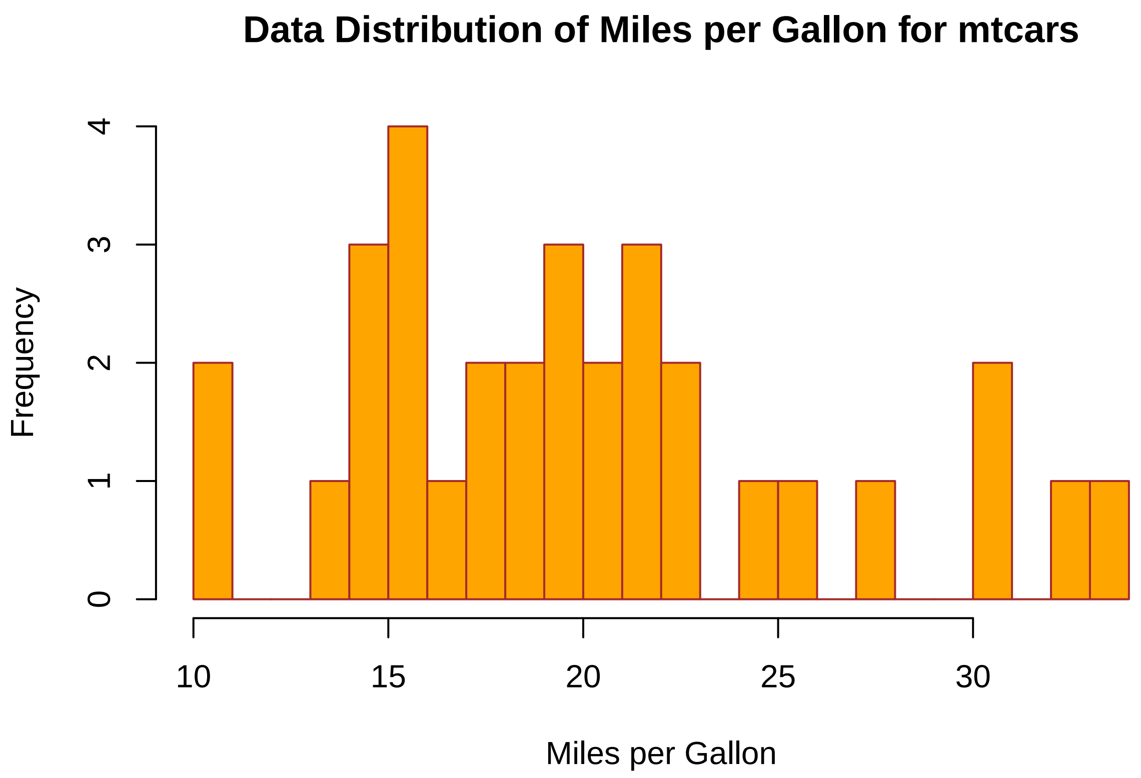
```
## [1] 36.3241
```

From the above calculation, we observe that weight has a mean of 3.22 and variance of 0.96 whereas miles per gallon has a mean of 20.1 and variance of 36.32. Means can be understood as a typical value when there are fewer amount of outliers, and variance shows how the variability of these values. Because we don't know if the data contain outliers, we need to see the distribution of these samples for each variable.

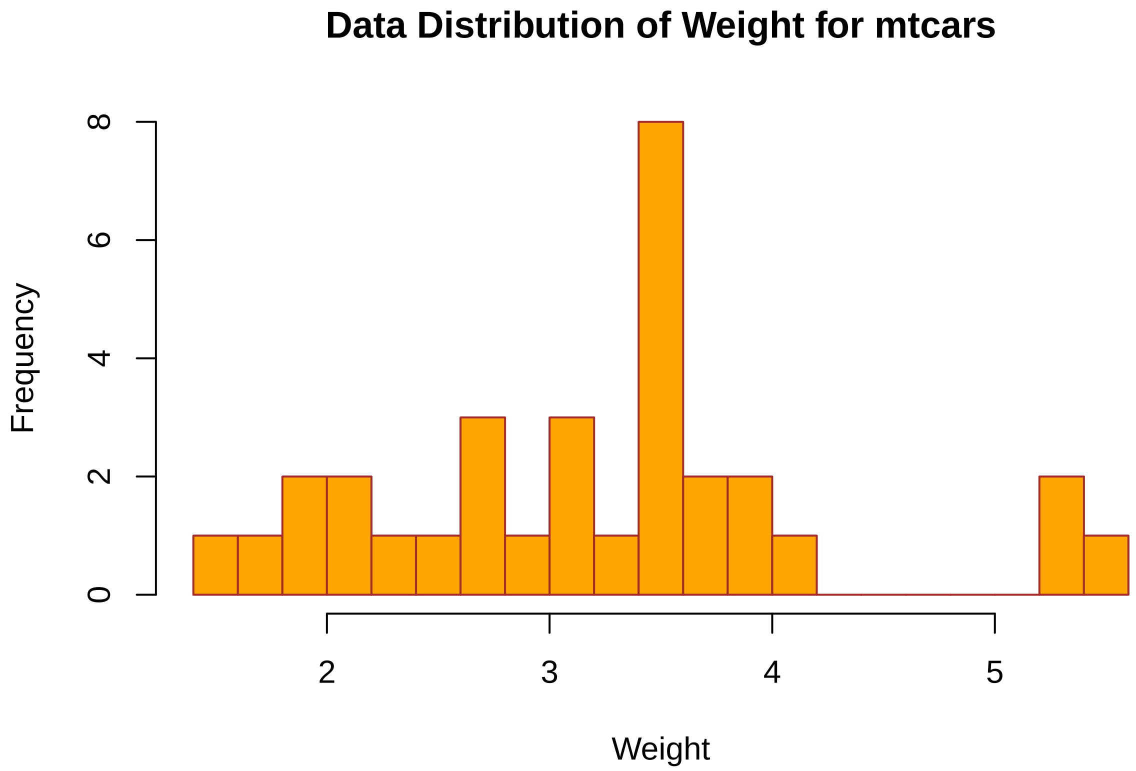
Distribution of Each Sample

As we can see from the head of the dataset, both weight and miles per gallon variable are continuous quantative variables, so we can plot two histograms for them.

```
hist(mtcars$mpg,
breaks = 20,
main = "Data Distribution of Miles per Gallon for mtcars",
xlab = "Miles per Gallon",
col = "orange",
border = "brown",
)
```



```
hist(mtcars$wt,
breaks = 15,
main = "Data Distribution of Weight for mtcars",
xlab = "Weight",
col = "orange",
border = "brown",
)
```

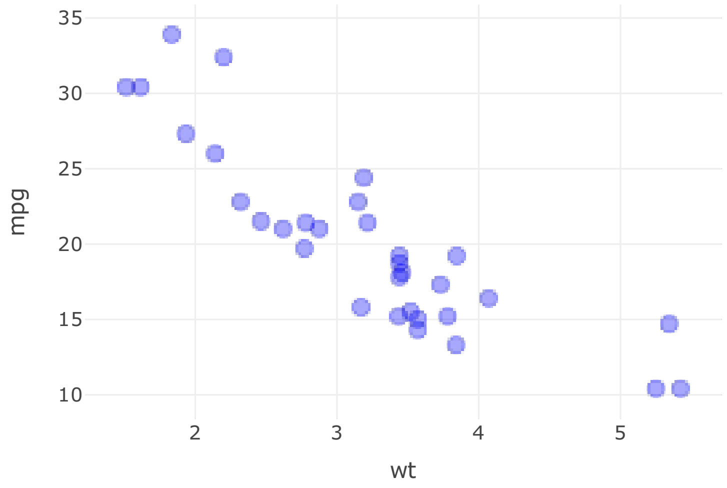


From the histogram for miles per gallon, we can see that most samples are within the range betweeen 15 and 25. We have a few extreme values that can be as low as 10 and as high as over 30-35. From the histogram for weight, we can see that there exists a majority of cars that weigh around 3.5. The distribution has a long tail toward smaller values, but also contain 3 outliers toward larger values, quantatitively weights above 5. Because of that, the ratio of the car with largest mpg to the car with smallest mpg is around 3-4, which is smaller than the ratio of the car with largest weight to the car with the smallest weight, which is around 5-7.

Correlations between Samples

Even though the distribution of these two variables are slightly different, we want to see if there is an association between weight and miles per gallon. Therefore, we plot a scatter plot between these two variables can provide an overview for their relationship.

```
fig <- plot_ly(df, x = ~wt, y = ~mpg, color = I("blue"), alpha = 0.5, size = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg')))
fig
```



From the above figure, we can see that there exists a negative relationship between weight and miles per hour in general. We also observe that the 3 outliers for large weights mentioned in the above histogram all have relatively smaller value of miles per gallon, although they do not exhibit the “linear” characteristics relative to cars with weights between 1.5 and 4. Since this is a qualitative description and we do not know how strong they are correlated, we will use the correlation coefficient to find out.

```
cov(df['wt'], df['mpg'])
```

```
##           mpg
## wt -5.116685
```

```
cor(df['wt'], df['mpg'])
```

```
##            mpg
## wt -0.8676594
```

Here, we can see that weight and miles per gallon has a covariance of -5.117 and a correlation of -0.867. Because correlation is a standardized version of covariance, we will mainly analyze this correlation coefficient.From the rule of thumb for interpreting the strength of the relationship between two variables based on the value of correlation coefficient r, a table [4] is proposed below:

r < 0.25: No relationship
0.25 < r < 0.5: Weak relationship
0.5 < r < 0.75: Moderate relationship
r > 0.75: Strong relationship

Because we have a correlation coefficient of -0.867 and both variables are without any transformation, we

[4] Zach, “What is Considered to Be a ‘Strong’ Correlation?,” Statology, 22-Jan-2020. [Online]. Available: https://www.statology.org/what-is-a-strong-correlation/. [Accessed: 23-Jan-2021].

can conclude with this quantative measurement that there exists a **strong negative linear relationship between weight and miles per gallon**.

Causation between wt and mpg

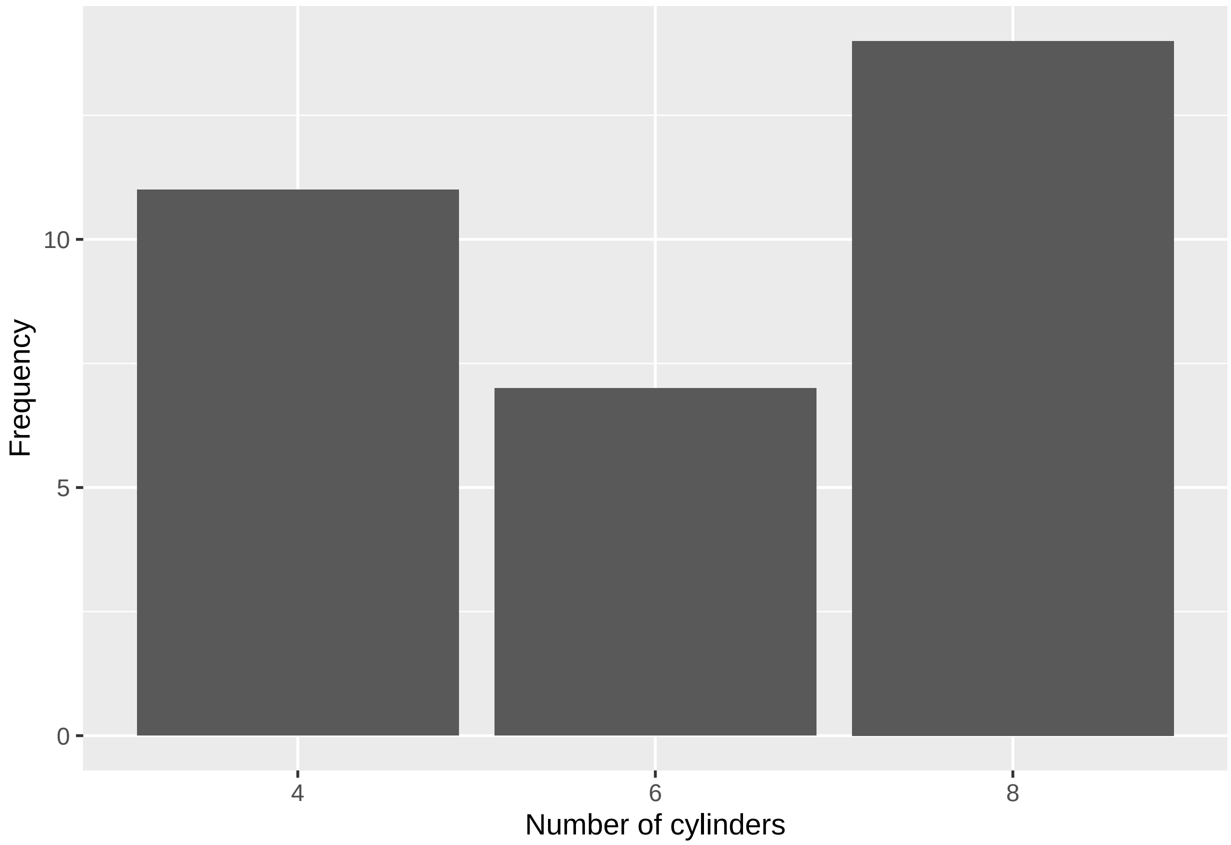
Now we know that there is a negative correlation between weight and miles per hour, we can further explore if this correlation is a causation, or there exists other confounding variables that caused this correlation. Intuitively, a car with more weight uses more fuels for each mile. Therefore, we might perform a granger causality test. Granger causality provides a much more stringent criterion for causation (or information flow) than simply observing high correlation. Although it is mostly used in time-series tasks, we can also consider changes in weight as a signal to changes in miles per gallon. But since this tool has not been covered in lecture for its applicability and suitability yet, we leave this exploration as a possible future study.

Relationship's dependency on the number of cylinders (cyl)?

Distribution of the Number of Cylenders

Now, we have introduced a new variable, number of cylinders, and we want to explore if the relationship between weight and miles per gallon is dependent on the number of cylinders. But before we start exploring this dependency, we want to look at the sample distribution of variable `cyl`. Here, we will use a barplot as we observe that number of cylinders are discrete values which only contain 4, 6, and 8.

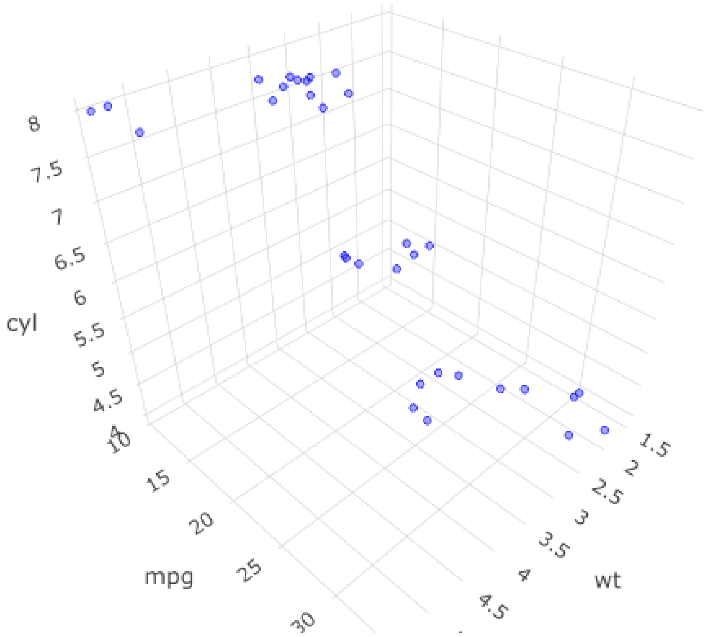
```r
agg_count <- as.data.frame(table(mtcars$cyl))
ggplot(agg_count, aes(x=Var1, y=Freq)) +
  geom_bar(stat = "identity") +
  xlab('Number of cylinders') +
  ylab('Frequency')
```



Based on the barplot, we found that most cars have a cylinder amount of 8, following by an amount of 4, and there are fewest cars having a cylinder amount of 6. To explore if the number of cylinders affects the relationship between weight and miles per gallon, we can plot a 3D scatter plot. Also, since we only have three catrgories for the number of cylinders, we can also plot three 2D scatter plot to show the correlation between these two variables under each cylinder amount category.

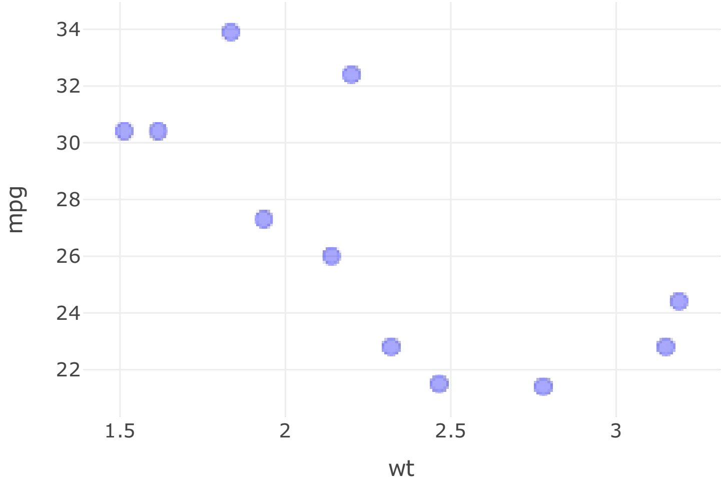
Correlations between Samples

```
library(plotly)
fig <- plot_ly(df, x = ~wt, y = ~mpg, z = ~cyl, color = I("blue"), alpha = 0.5, size = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg'),
                                   zaxis = list(title = 'cyl')))
fig
```

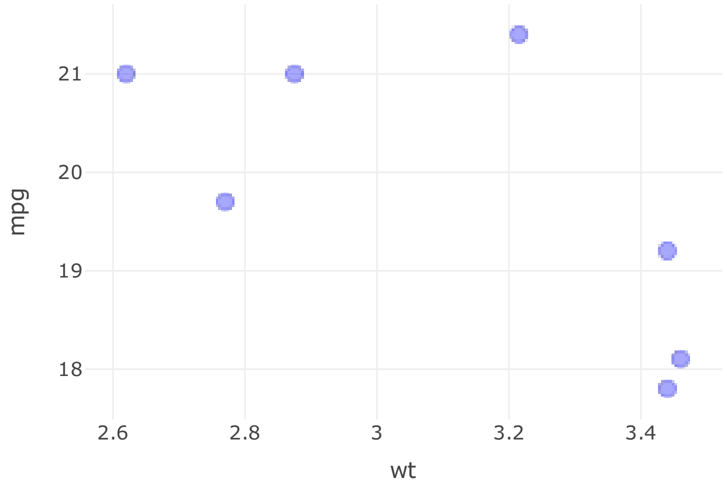


```
mtcar_4 = df[df$cyl == 4,]
mtcar_6 = df[df$cyl == 6,]
mtcar_8 = df[df$cyl == 8,]
```

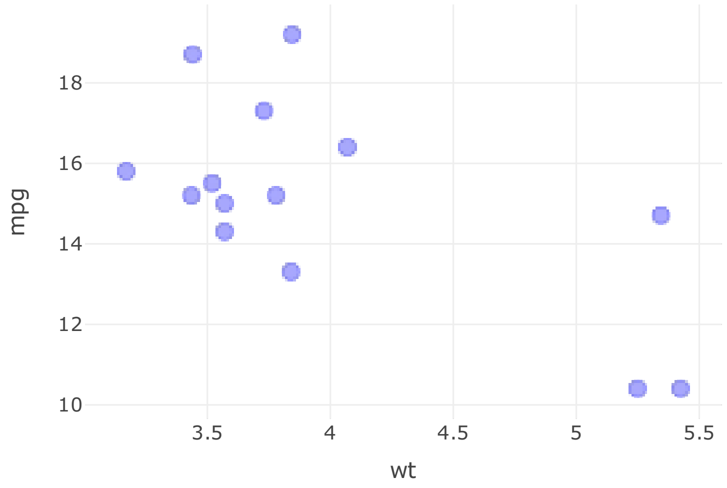
```
fig <- plot_ly(mtcar_4, x = ~wt, y = ~mpg, color = I("blue"), alpha = 0.5, size = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg')))
fig
```



```
fig <- plot_ly(mtcar_6, x = ~wt, y = ~mpg, color = I("blue"), alpha = 0.5, size = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg')))
fig
```



```
fig <- plot_ly(mtcar_8, x = ~wt, y = ~mpg, color = I("blue"), alpha = 0.5, size = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg')))
fig
```



From the above plots, we qualitatively observe the following trends: - When the number of cylinder is 4, the correlation between weight and miles per gallon is negative when weight is smaller than 2.6, and this correlation becomes positive when weight is larger than 2.6. - When the number of cylinder is 6, the correlation between weight and miles per gallon is generally negative, but there is an outlier where its weight is around 3.2. - When the number of cylinder is 8, the correlation between weight and miles per gallon is negative, including those three outliers with weight larger than 5 mentioned in the previous section.

Quantatitive Measurements

Quantatively, we can examine the correlation coefficient of all three categories.

```
cor(mtcar_4['wt'], mtcar_4['mpg'])
```

```
##           mpg
## wt -0.7131848
```

```
cor(mtcar_6['wt'], mtcar_6['mpg'])
```

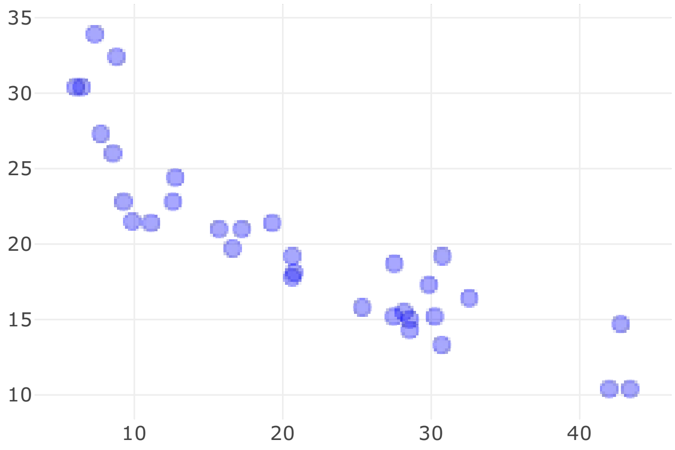
```
##           mpg
## wt -0.6815498
```

```
cor(mtcar_8['wt'], mtcar_8['mpg'])
```

```
##          mpg
## wt -0.650358
```

These correlation coefficients are consistent with our previous discussion, where all of them a smaller than the correlation when we don't split the data by their number of cylenders. With an initial impression, we can also conclude that the relationship between weight and miles per gallon does depend on the number of cylinders, as more cylenders seem to lead weaker correlation between weight and miles per gallon. In conclusion, while the number of cylinders does not inverse the relationship between weight and miles per gallon, there exists a trend where more cylinders lead to weaker negative correlations between weight and miles per gallon. An explanation for this could be that a car with more number of cylinders will have a low value of miles per gallon regardless of its weight - `cyl` becomes a more dominanent factor when its value gets bigger. To verify this, we can see if we have a higher correlation when we compare the product of weight and number of cynlinders against miles per gallon.

```
fig <- plot_ly(df, x = (df$wt)*(df$cyl), y = df$mpg, color = I("blue"), alpha = 0.5, size = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg')))
fig
```



```
cor(df['wt']*df['cyl'], df['mpg'])
```

```
##           mpg
## wt -0.8843793
```

```
cor(df['wt'], df['mpg'])
```

```
##           mpg
## wt -0.8676594
```

Both scatterplot and correlation suggests that the correlation becomes stronger when we take both weight and number of cylinders into account.

Discussion on Tools & Analysis Justification

In this data analysis project, several tools to analyze data and make inferences are used: sample mean, sample variance, covariance between two variables, correlation between two variables, 2D Scatterplots, 3D Scatterplots, Histogram, and Barcharts.

Estimators of mean are biased because samples are collected from a trend megazine where the trend indicates that samples are not randomly picked from the population but following certain biases (e.g. only cars with high performance). A solution for this is to sample data randomly from the population. By the same token, estimators of variance are also biased because of the selective sampling from the population (e.g. cars with high performance are more "characteristic" than normal cars, making sample variance higher than population variance).

Within a sample, compared to the median, sample mean is biased because samples contain outliers in only one tail (e.g. we have three cars that weigh significantly more than the rest of the cars, making the sample mean biased to a larger value). A solution for this is to calculate the mean without the influence of outliers, or use the median for more robust estimates.

Within a sample covariance between two variables are biased when compare the covariance between other pairs of variables. This is because covariance is not a standardized representation of correlation and is influenced by the scale of variance of variables. A solution for this is to use the correlation coefficient, or preprocess variables to the same scale before calculating the covariance.

Within a sample, correlations, scatterplots, bar charts, and histogram are less biased because they are either standardized representation of variables or figures that use the same scale (within a figure) to represent variables.

Conclusion

In this work, we analyzed the relationship between weight and miles per gallon on automobiles, as well as whether the number of cylinders is affecting this relationship. By performing qualitative and quantatitive examinations on these variables, we discovered that there exists a strong linear negative relationship between weight and miles per gallon, the number of cylinders is affecting this relationship where more cylinders lead to weaker relationship between weight and miles per gallon, and the number of cylinders itself can also considered as a factor influencing the miles per gallon where the product of the weight and the number of cylinders and miles per gallon produced a stronger negative correlation than weight alone. Future studies on these variables include determining if the relationship between weights and the number of cylinders is causal or not, and why the number of cylinders is affecting the relationship between weight and miles per gallon (i.e. are there confounding variables?).