Project 2

Dan Do

October 10, 2023

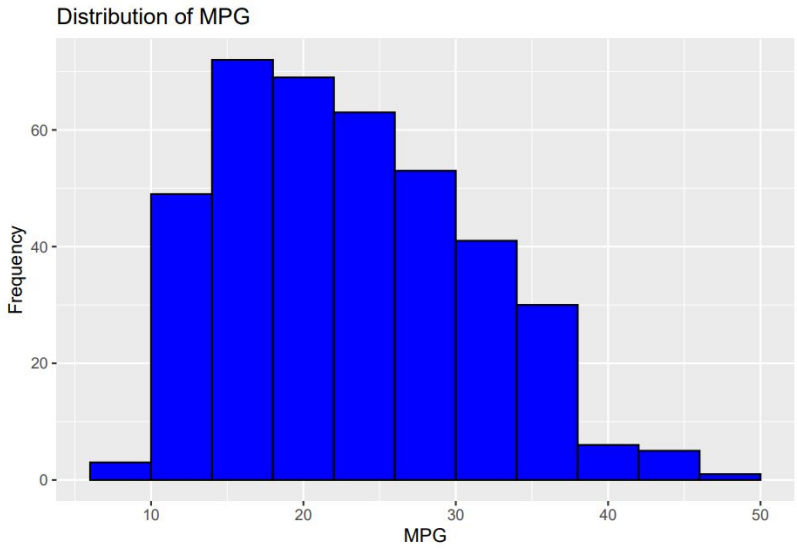
Abstract

The goal of this project is to devise 5 questions using the Auto dataset. The questions are relevant to what someone would be interest to if they want to learn more about the auto industry. The questions will be answered using the data analytics skills that we have learned in the class. All the resources that I used in the project will be entirely from the class's materials.

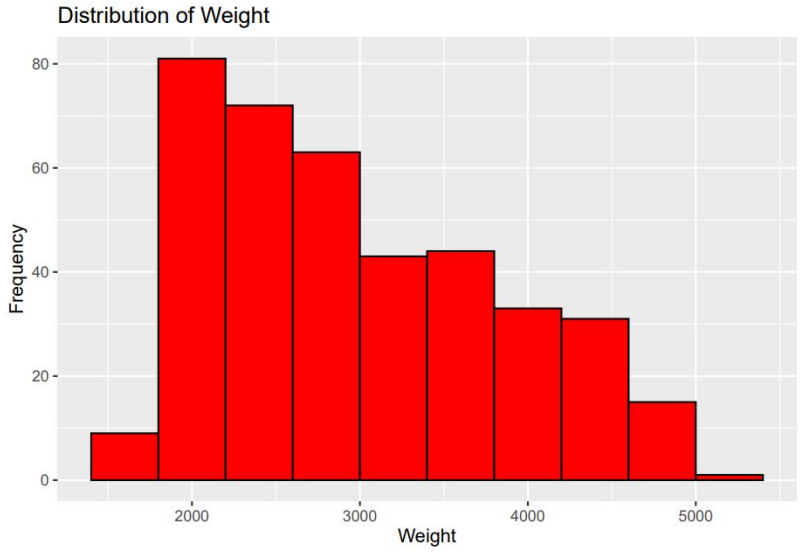
In this comprehensive report on the Auto dataset, I examined the factors impacting miles-per-gallon (mpg) in vehicles, offering valuable insights for both potential car buyers and auto industry enthusiasts. Firstly, cars with 4 or 5 cylinders engine have the best mpg and that cars originating from Japan are found to have the highest mpg. In addition, I found that the lighter the weight of a car, the less fuel needed, and cars from Japan have these features the best. Lastly, there are no significant interaction effect between the country of origin and the number of cylinders on mpg.

In my report, I used generalized linear models and two-way ANOVA model to provide data-driven transparency on these factors. Using generalized linear model so that I can build a linear relationship between the independent and dependent variables. Using two-way ANOVA method to compare and determine the effect of two independent variables on one dependent variable.

Question 1: How is the distribution of data?



Looking at the histogram that shows the mpg distribution, we see that there are a lot of cars with mpg in the range of 15 to close to 20 miles per gallon with over 70 cars. Second place with mpg range from 20 to 25, there are about over 60 cars within that range. There are under 10 cars with the least mpg and about less than 5 cars with the most mpg.



Based on the weight distribution, we can see that there are a lot of cars with weight in range of 1500 to 2300 pounds with about 82 cars. The frequency of cars that weigh 5000 pounds or more is very low, with probably about less than 5 cars. This makes sense because average weight of a car looks to be about 3000 pounds. If a car goes over the weight limit, it can lead to tire failure and potentially be dangerous.

Question 2: What factors are influencing mpg?

```
##
## Call:
## glm(formula = mpg ~ as.factor(cylinders) + displacement + horsepower +
##     weight + acceleration + year + as.factor(origin), data = data)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.208e+01  4.541e+00  -4.862 1.70e-06 ***
## as.factor(cylinders)4  6.722e+00  1.654e+00   4.064 5.85e-05 ***
## as.factor(cylinders)5  7.078e+00  2.516e+00   2.813  0.00516 **
## as.factor(cylinders)6  3.351e+00  1.824e+00   1.837  0.06701 .
## as.factor(cylinders)8  5.099e+00  2.109e+00   2.418  0.01607 *
## displacement           1.870e-02  7.222e-03   2.590  0.00997 **
## horsepower            -3.490e-02  1.323e-02  -2.639  0.00866 **
## weight                -5.780e-03  6.315e-04  -9.154  < 2e-16 ***
## acceleration           2.598e-02  9.304e-02   0.279  0.78021
## year                   7.370e-01  4.892e-02  15.064  < 2e-16 ***
## as.factor(origin)2     1.764e+00  5.513e-01   3.200  0.00149 **
## as.factor(origin)3     2.617e+00  5.272e-01   4.964 1.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.595561)
##
##     Null deviance: 23819.0  on 391  degrees of freedom
## Residual deviance:  3646.3  on 380  degrees of freedom
## AIC: 2012.7
##
## Number of Fisher Scoring iterations: 2
```

To find the answer to this question, I used the generalized linear model to understand the relationship between the mpg and all the other variables.

According to the result, it can be seen that the p-value for cars with 6 cylinders, which is 0.6701, and acceleration which has a p-value of 0.78021 are greater than 0.05, which indicates that they do not significantly influence mpg. Thus they should not be used to estimate mpg.

On the other hand, cars with 4, 5, and 8 cylinders are significant because they have p-values less than 0.05. Displacement, horsepower, weight, year, and origin variables have p-values less the than 0.05. That indicates these variables are significantly influence mpg.

Question 3: Which one is positively influencing mpg and which one is negatively influencing mpg?

```
##
## Call:
## glm(formula = mpg ~ as.factor(cylinders) + displacement + horsepower +
##     weight + acceleration + year + as.factor(origin), data = data)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.208e+01  4.541e+00  -4.862 1.70e-06 ***
## as.factor(cylinders)4  6.722e+00  1.654e+00   4.064 5.85e-05 ***
## as.factor(cylinders)5  7.078e+00  2.516e+00   2.813  0.00516 **
## as.factor(cylinders)6  3.351e+00  1.824e+00   1.837  0.06701 .
## as.factor(cylinders)8  5.099e+00  2.109e+00   2.418  0.01607 *
## displacement           1.870e-02  7.222e-03   2.590  0.00997 **
## horsepower            -3.490e-02  1.323e-02  -2.639  0.00866 **
## weight                -5.780e-03  6.315e-04  -9.154  < 2e-16 ***
## acceleration           2.598e-02  9.304e-02   0.279  0.78021
## year                   7.370e-01  4.892e-02  15.064  < 2e-16 ***
## as.factor(origin)2     1.764e+00  5.513e-01   3.200  0.00149 **
## as.factor(origin)3     2.617e+00  5.272e-01   4.964 1.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.595561)
##
##     Null deviance: 23819.0  on 391  degrees of freedom
## Residual deviance:  3646.3  on 380  degrees of freedom
## AIC: 2012.7
##
## Number of Fisher Scoring iterations: 2
```

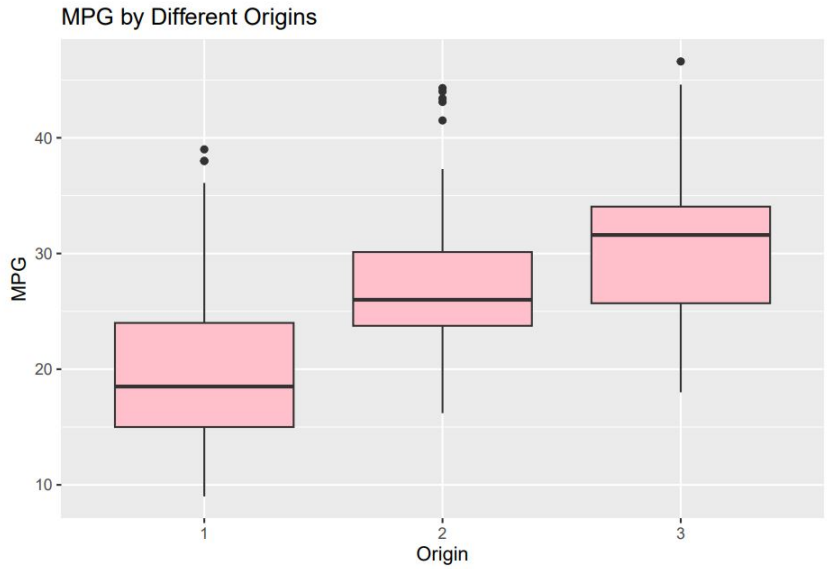
According to the table, we are looking at the t-values and estimates to determine the linear relationship between the factors and mpg.

Using weight as an example, it has a t-value of -9.154, which means that as the weight increases, the mpg for the vehicle would decreases. Thus, if we want to have vehicle with a higher mpg, then we would want a lighter vehicle. And also using estimation, we see that the estimate for weight is -0.005780, which suggests that the lower the weight, the better the mpg the car has.

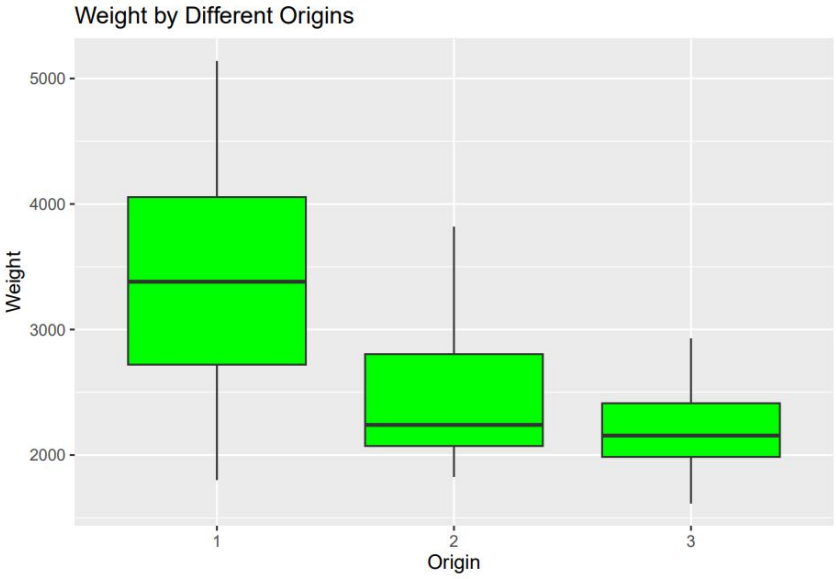
Using cylinder 3 as the base and compare them to cars with cylinder 4, 5, 6, and 8, we can see that they have a higher mpg than 3-cylinders cars. Cars with 4 or 5 cylinders engine have the best mpg because their estimates are high.

Also, as the year gets more recent, the cars have better mpg because the estimate is 0.737.

Question 4: How do the countries/origins have different cars? Compare the cars structure by the weight and mpg to their origins.



For this question, I used the box plot to check for the data distribution. From the plot, we can see that origin 3, which is Japan, has the highest median in terms of mpg.



Combining these two box plots, I use them to test for the statement I did on Question 3 about weight and mpg. From the box plot, it shows that origin 3 (Japan) produces cars with the least weight. And thus, since origin 3's cars are the lightest and provide the most mpg, from the two box plots, they support the statement.

Question 5: Is there a significant difference in mpg between vehicles from different countries, and does this difference depend on the number of cylinders in the engine?

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## origin             1   7609    7609 334.258 <2e-16 ***
## cylinders          1   7325    7325 321.769 <2e-16 ***
## origin:cylinders   1     52      52   2.293  0.131
## Residuals        388   8833      23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since there are two independent variables which are origins and cylinders on one dependent variable, mpg, I use the two-ways ANOVA method to find the effect of this combination.

Two-ways ANOVA test three null hypotheses:

1) The means of observations grouped by one factor (origin) are the same
2) The means of observations grouped by the other factor (cylinders) are the same
3) There is no interaction between the two factors

Based on the result of the R code, it can be seen that origin has a highly significant effect on mpg. P-value for origin variable is less than 0.05 indicates that there are significant differences in mpg between the three origin categories. The cylinders factor also has a highly significant effect on mpg, with a very low p-value (<2e-16) that is less than 5% significance level. This suggests that the number of cylinders in the engine significantly impacts fuel efficiency. Finally, the interaction between origin variable and cylinders variable does not appear to be statistically significant. The p-value for this interaction term is 0.131, which is higher than the typical significance level of 0.05. This suggests that there is no strong evidence of a significant interaction effect between the country of origin and the number of cylinders on mpg. Thus, we would reject the first and second null hypothesis, but do not reject the third null hypothesis.

Limitations of Report:

The first factor that can put a limitation to the report is other variables that a car can have. This report only allows variables such as mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin, and name. There are many other variables that can influence the data like models (sedan, truck, SUV, etc.), engine size, and price. Those are factors that very important in the auto industry and would allow an in-depth analytics.

The second limitation could be that the relationship between the variables are different due to technological advancement in the auto industry over time. As technology grows, the materials that are used to produce the cars can be changed and/or improved. For example, a car made in the 80s might use materials that allow lighter weight to the car and a more efficient engines, compared to a car made in the 70s. These temporal variables are also need to be taken into consideration.

The third limitation is that by looking at the Auto dataset, there are very few cars with 3 cylinders (about 4 cars) as well as 5 cylinders (about 3 cars) while there are 199 cars with 4 cylinders, 83 cars with 6 cylinders, and 103 cars with 8 cylinders. Since there are so little cars with 3 and 5 cylinders, this might cause the data to be skewed.

R code

```r
# Load the packages
library(ISLR)
library(ggplot2)
library(dplyr)

# Load the Auto dataset and show the structure of it
data("Auto")
str(Auto)

# Assign Auto dataset to variable data
data = Auto

# Question 1
ggplot(data, aes(x = mpg)) +
  geom_histogram(binwidth = 4, fill = "blue", color = "black") +
  labs(title = "Distribution of MPG",
       x = "MPG",
       y = "Frequency")

ggplot(data, aes(x = weight)) +
  geom_histogram(binwidth = 400, fill = "red", color = "black") +
  labs(title = "Distribution of Weight",
       x = "Weight",
       y = "Frequency")

# Question 2 + 3
ggplot(data, aes(x=factor(origin), y=mpg)) +
  geom_boxplot(fill = "pink") +
  labs(title = "MPG by Different Origins",
       x = "Origin",
```

```
       y = "MPG")


ggplot(data, aes(x=factor(origin), y=weight)) +
  geom_boxplot(fill = "green") +
  labs(title = "Weight by Different Origins",
       x = "Origin",
       y = "Weight")


# Question 4
model1 = glm(formula = mpg ~ as.factor(cylinders) + displacement + horsepower + weight + acceleration +
summary(model1)


# Question 5
aov_2way = aov(mpg ~ origin * cylinders, data = Auto)
summary(aov_2way)


# Total number of cars with each factor of cylinders
data %>%
  group_by(data$cylinders) %>%
  summarise(Count = n())
```