HW3

Q1:

In R, we can calculate pairwise correlations using the `cor()` function. Here's how we can calculate and explain the pairwise correlation among the variables 'cylinders', 'displacement', 'horsepower', and 'weight' from your dataset:

The correlation between the number of cylinders and engine displacement is nearly perfect (0.951) because more cylinders often mean a larger engine, as each cylinder contributes to displacement. Cylinders also correlate positively with horsepower (0.843) due to more combustion chambers, though it's influenced by factors like engine design and efficiency. Cylinders correlate with vehicle weight (0.898) since larger engines are in heavier vehicles, but materials and design play a role. Engine displacement strongly correlates with weight (0.933) because larger engines tend to be in heavier vehicles.

In summary, these correlations result from the interconnectedness of engine attributes and vehicle characteristics, but other factors impact them.

```
> data <- read.csv("C:/Users/faiya/OneDrive - Texas Tech University/Texas tech course/fall 23/software analytics//HW1/auto.csv")
> correlation <- cor(data[, c('cylinders', 'displacement', 'horsepower', 'weight')])
> print(correlation)
              cylinders displacement horsepower    weight
cylinders     1.0000000    0.9508233  0.8429834 0.8975273
displacement  0.9508233    1.0000000  0.8972570 0.9329944
horsepower    0.8429834    0.8972570  1.0000000 0.8645377
weight        0.8975273    0.9329944  0.8645377 1.0000000
```

	cylinders	displacement	horsepower	weight
cylinders	1.0000000	0.9508233	0.8429834	0.8975273
displacement	0.9508233	1.0000000	0.8972570	0.9329944
horsepower	0.8429834	0.8972570	1.0000000	0.8645377
weight	0.8975273	0.9329944	0.8645377	1.0000000

Q2:

The correlation between acceleration and horsepower is negative (-0.689), indicating an inverse relationship. As horsepower increases, acceleration tends to decrease, and vice versa. This correlation is less perfect because other factors, such as vehicle weight, transmission efficiency, and gear ratios, influence a car's acceleration performance. While higher horsepower generally means more power, it doesn't guarantee faster acceleration due to the complex interplay of these additional factors. Therefore, the negative correlation suggests that higher horsepower vehicles may not always accelerate more quickly, highlighting the nuanced relationship between these two variables in automotive performance.

```
# Calculate the correlation
q2 <- cor(data[, c('acceleration', 'horsepower')])
print(q2)
```

```
> # Calculate the correlation
> q2 <- cor(data[, c('acceleration', 'horsepower')])
> View(q2)
> View(q2)
> print(q2)
             acceleration horsepower
acceleration    1.0000000 -0.6891955
horsepower     -0.6891955  1.0000000
```

	acceleration	horsepower
acceleration	1.0000000	-0.6891955
horsepower	-0.6891955	1.0000000

Q3:

The correlation between 'model_year' and 'mpg' is positive and relatively strong (0.580), indicating a moderate positive relationship. This suggests that, on average, newer model years tend to have better fuel efficiency (higher miles per gallon) compared to older model years. This positive correlation suggests that there may have been advancements in automotive technology, fuel efficiency standards, or changes in consumer preferences over time that have led to more fuel-efficient vehicles in newer model years.

```
> q3 <- cor(data[, c('model_year', 'mpg')])
> print(q3)
           model_year      mpg
model_year   1.000000 0.580541
mpg          0.580541 1.000000
```

Regarding the negative correlation between 'model_year' and the group of 'cylinders,' 'displacement,' 'horsepower,' and 'weight,' it suggests that as model years increase, there is a tendency for these engine-related variables (cylinders, displacement, horsepower, and weight) to decrease slightly. This could reflect a trend towards smaller, more fuel-efficient engines and lighter vehicles in newer model years, which aligns with the emphasis on fuel efficiency and emissions reduction in the automotive industry. However, the negative correlation is relatively small, indicating that other factors, such as design trends and consumer preferences, also play a role in these changes.

```
> correlation_q3 <- cor(data[, c('model_year', 'cylinders', 'displacement', 'horsepower', 'weight')])
> print(correlation_q3)
              model_year  cylinders displacement horsepower     weight
model_year     1.0000000 -0.3456474   -0.3698552 -0.4163615 -0.3091199
cylinders     -0.3456474  1.0000000    0.9508233  0.8429834  0.8975273
displacement  -0.3698552  0.9508233    1.0000000  0.8972570  0.9329944
horsepower    -0.4163615  0.8429834    0.8972570  1.0000000  0.8645377
weight        -0.3091199  0.8975273    0.9329944  0.8645377  1.0000000
>
```

Q4:

It doesn't make sense to analyze the correlation of 'origin' with variables like 'mpg' or 'horsepower' because 'origin' is a categorical variable that represents the country or region of origin of a vehicle, while 'mpg' and 'horsepower' are continuous numerical variables. Correlation is typically used to measure the strength and direction of linear relationships between two continuous variables. Categorical variables like 'origin' don't have a natural linear relationship with continuous variables.

We can use ANOVA analysis to further analyze, by using aov function in R :

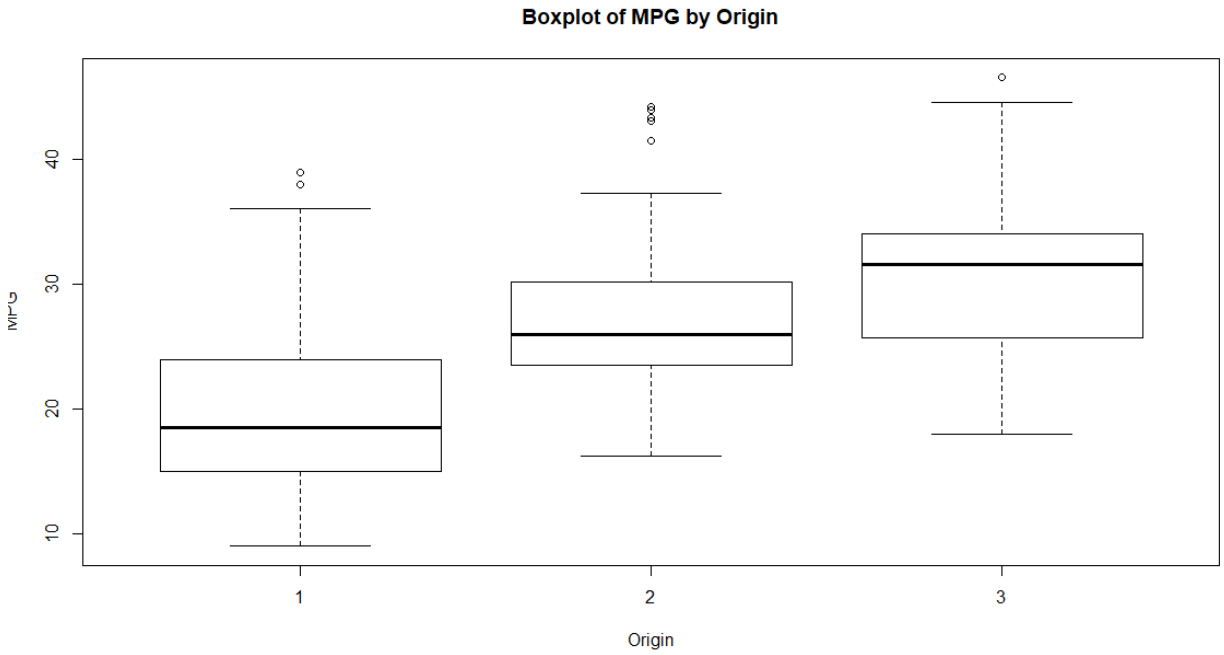
```
> # Perform ANOVA
> anova_result <- aov(mpg ~ origin, data = data)
> # Print ANOVA summary
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
origin        1   7609    7609   183.1 <2e-16 ***
Residuals   390  16210      42
---
```

The ANOVA (Analysis of Variance) summary presented indicates that there is a statistically significant relationship between the 'origin' variable and the 'mpg' variable. The p-value (Pr(>F)) is significantly less

than 0.05 (typically the significance threshold), suggesting that there are significant differences in 'mpg' among different 'origin' categories. In other words, the country of origin appears to be related to variations in fuel efficiency (mpg). However, this analysis does not provide information about the specific differences between origin categories; additional post-hoc tests or pairwise comparisons may be needed to explore those differences further.

We can further analyze- using boxplot- shows that there is a significant difference in the median mpg of cars from different countries. Japanese cars have the highest median mpg, followed by European cars and American cars. American cars have the widest IQR, indicating that they have the greatest variability in mpg.

```
# Create a boxplot
boxplot(data$mpg ~ data$origin, xlab = "Origin", ylab = "MPG", main = "Boxplot of MPG by Origin")
```



Q5:

Analyzing the correlation of zipcodes, which are essentially categorical identifiers for geographical areas, with other numerical variables like house price or household income doesn't make sense for several reasons.

Firstly, zipcodes are nominal categorical data, not continuous numerical data. Correlation measures, such as Pearson's correlation coefficient, are designed for assessing relationships between continuous variables. Zipcodes, being numerical, might suggest an ordinal relationship, but in reality, the numerical values assigned to zipcodes are arbitrary and do not represent a meaningful quantitative relationship. For example, the difference between two zipcodes doesn't convey any numerical significance.

Secondly, zipcodes are primarily administrative boundaries, and the numerical values assigned to them are not suitable for mathematical operations. Even though two zipcodes might be numerically close, it doesn't imply any meaningful proximity in terms of geographical or socioeconomic factors.

To compare house prices between Lubbock, TX, and Dallas, TX (which have different zipcodes), we should use the actual address or location data associated with each property. Instead of relying solely on zipcodes, we can collect precise geographical coordinates or street addresses for the properties in question. Then, use these coordinates or addresses to conduct a meaningful comparison of house prices. Various real estate and mapping tools are available that can help you obtain property data based on location, allowing for accurate and meaningful comparisons between different areas, even if they have distinct zipcodes.