Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?** (3 marks)

   - Bike demand for the year 2019 was more than the year 2018
   - Bike demand is similar throughout the weekdays
   - Demand doesn't change much whether its working day or not
   - Demand of the bike is more if the weather is clear or with mist cloud

2. **Why is it important to use drop_first=True during dummy variable creation?** (2 mark)

   It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.
   For Example: We have 3 variables: male female and others. We can only take 2 variables as male will be 10, female will be 01, so we don't need others as we know 00 will indicate others. So we can remove it.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?** (1 mark)

   Both temp and atemp have same correlation as 0.63 which is highest among all.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

- By checking the linearity between response and predictor variables.
- By checking the error distribution
- By checking the P value .General standard is its should be less than 0.05
- By checking the VIF value low VIF value generally less than 5 is good

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)

- mist_fewcloud
- sep
- Sat

General Subjective Questions

1. **<u>Explain the linear regression algorithm in detail</u>**. (4 marks)

   Linear regression is a supervised learning algorithm that compares input (X) and output (Y) variables based on labeled data. It is also used to find the relationship between the two variables and predicting future results based on past relationships information. The measure of relationship between two variables is denoted by the correlation coefficient. A linear regression model is represented by the below equation.

Simple Linear Regression:

A general linear equation looks like

$Y = a + bX$

X = the independent variable
Y = the dependent variable
b = the slope of the line
a = the intercept (intercept value is y value when x = 0)

Generalizing the equation with respect to Simple Linear Regression

$Y = B_0+B_1X$

Where

B_0 is a constant

B_1 is the regression coefficient

Multiple Linear Regression:

A multiple linear regression model can extent to several explanatory variables. Generalizing equation used to explain Multiple Linear Regression

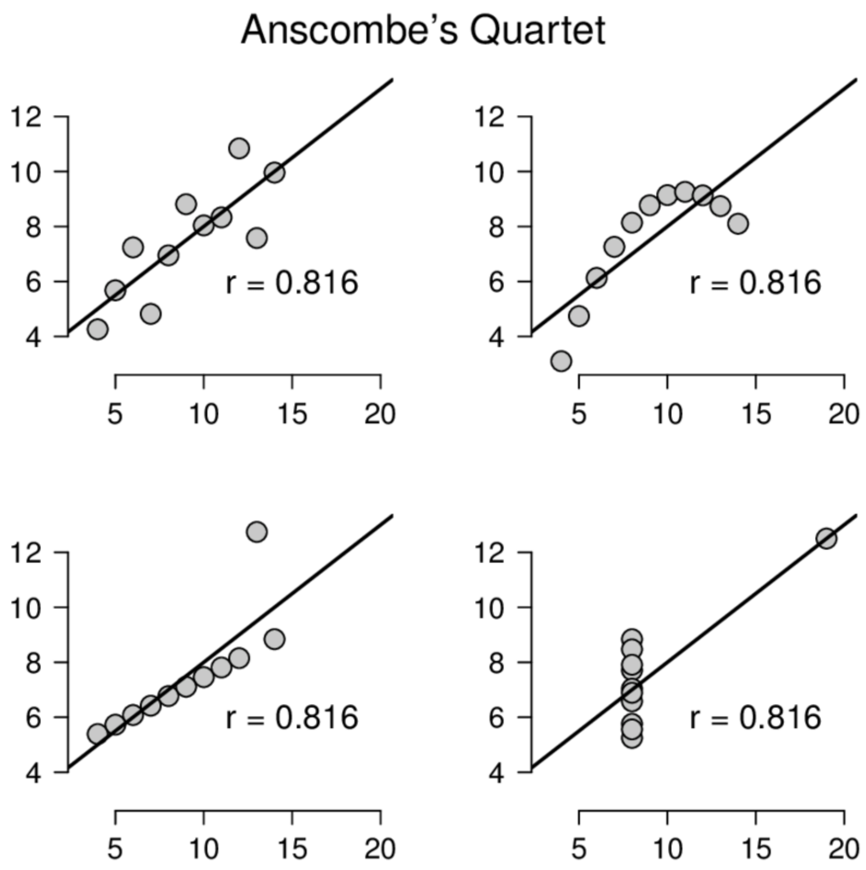
$Y_i = B_0 + B_1X_{i1} + B_2X_{i2} + B_3X_{i3}$ ----------- $B_pX_{ip} + E$

Y_i = dependent variable
X_i = explanatory variables
B_0 = y-intercept (constant term)
B_p = slope coefficients for each explanatory variable
E = the model's error term (also known as the residuals)

2. **Explain the Anscombe's quartet in detail.** **(3 marks)**

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
It comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

Below is a sample figure representing the same


Anscombe's Quartet

If you see the above figure your will see that each has the Pearson's correlation R between the x and y values as 0.816 that is the four different data sets are also equal in terms of the mean and variance of the x and y values but the graphical displays reveal that the patterns are very different from one another, and that the Pearson's correlation (a linear measure of association) is only valid for the data set from the top left panel.

3. **What is Pearson's R?** (3 marks)

Pearson's Correlation Coefficient, often denoted as *r*, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1,

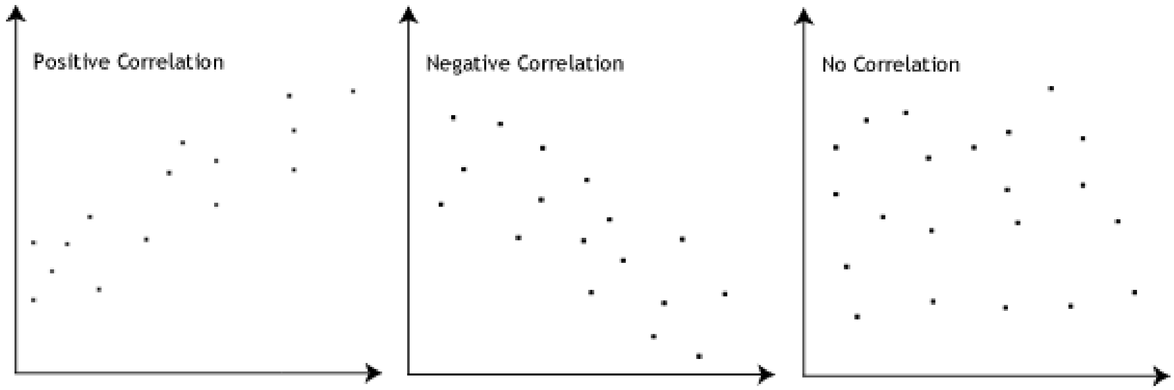
Where:
r = 1 denotes perfect positive linear relationship
r = -1 denotes perfect negative linear relationship
r = 0 denotes no linear relationship

The formula to calculate r value is given below

The formula is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where
$\mathbf{Y}_\mathbf{i}$ and $\mathbf{X}_\mathbf{i}$ individual data points while $\bar{\mathbf{Y}}$ and **X** are means of the variables
Sample graph representing the r values cases



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** (3 marks)

Scaling in the context of data analysis and machine learning refers to the process of transforming the range of variables or features of a dataset. The goal of scaling is to bring all features to a similar scale or range, which can be important for certain algorithms and analyses. There are different methods of scaling, and two common ones are normalized scaling and standardized scaling.

Why Scaling is Performed:

Algorithm Sensitivity: Some machine learning algorithms are sensitive to the scale of the input features. For example, distance-based algorithms (e.g., k-nearest neighbors, support vector machines) can be influenced by the magnitude of features.

Convergence Speed: Gradient-based optimization algorithms, such as those used in neural networks, may converge faster when features are on a similar scale.

Interpretability: Scaling can make it easier to interpret coefficients in linear models, as they represent the change in the output per unit change in the input.

Normalized Scaling:

Range: Normalized scaling, or Min-Max scaling, involves transforming the data to a specific range, usually between 0 and 1.

Formula: The formula for normalized scaling is given by:

$$y = \frac{x \ - min(x)}{max(x) - \min(x)}$$

Standardized Scaling (Z-score normalization):

Centering and Scaling: Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

Formula: The formula for standardized scaling is given by:

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

Differences between Normalized Scaling and Standardized Scaling:

Scale Range: Normalized scaling confines data to a specific range (e.g., 0 to 1), while standardized scaling centers data on the mean with a standard deviation of 1.

Sensitivity to Outliers: Standardized scaling is less sensitive to outliers compared to normalize scaling.

Use Cases: Normalized scaling is often used when the distribution of data is known and the algorithm requires features to be within a specific range. Standardized scaling is useful when the distribution of the data is not known or when dealing with algorithms that assume the features are normally distributed.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to assess the individual contributions of each variable. VIF is calculated for each predictor variable, and a high VIF indicates a high degree of multicollinearity.

A VIF value of infinity typically occurs when there is a perfect linear relationship between a set of independent variables. This means that one or more variables can be expressed as a perfect linear combination of others, leading to an infinite VIF.

Dealing with infinite VIF values is crucial to maintain the integrity of the regression analysis. Here are some steps you can take:

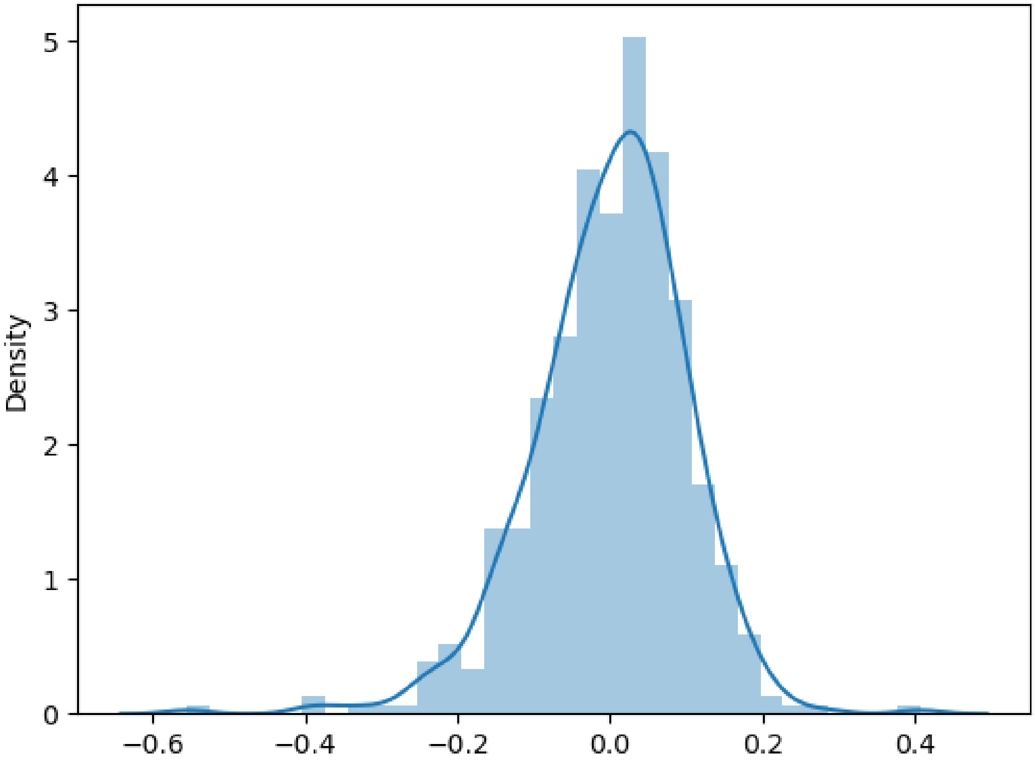
- Consider removing one or more of the highly correlated variables
- If the model includes interactions or polynomial terms, reconsider whether these are necessary. Simplifying the model may help reduce multicollinearity.
- Increasing the sample size can sometimes alleviate multicollinearity issues.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**.

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Sample image from my test model


Density