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?
    a. The categorical variables in the dataset are
        i. season - We see the winter and summer positively impact bike demand. During the spring season, the demand is low
        ii. month , - Sept month we see the highest demand for bikes, compared to another month Jan and Jul also has some demands
        iii. weekday , During the week day we see the demand are similar , only during holiday , demand is reduced
        iv. weathersit – Demand is high when the weather situation is clear.
        v. Yr – year 2019 the demand has increased compared to 2018

2. Why is it important to use drop first=True during dummy variable creation?

    drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.
    using drop_first=True during dummy variable creation helps address multicollinearity, improves interpretability, reduces model complexity, and ensures consistency between the training and prediction phases.

3. Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable?

    Looking at the pairs-plot 'temp' and 'atemp' has the highest correlation

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. I validated the assumptions using
        i. Error Distribution: In the model created the error terms were normally distributed
        ii. Multicollinearity check - Validating Absence of Multicolinearity
        iii. Validating homoscedasticity - almost no relation between Residual & Predicted Value.
        iv. Linear Relation Prediction : There was a linear relationship between predicted and actual values

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

    The top 3 Features contributing significantly towards explaining the demand were

    a. Temperature
    b. mnth_Sep
    c. year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s)

The formula for a simple linear regression is
y = b0 + b1*x1 + b2*x2 + ... + bn*xn
where:

- y is the predicted value of the dependent variable y for any given value of the independent variable x).
- b0 is the intercept, the predicted value of y when the x is 0.
- b1 is the regression coefficient – how much we expect y to change as x increases.
- x is the independent variable (the predictor).

The algorithm works by finding the values of B0 and B1 that minimize the mean squared error (MSE) between the predicted and actual values of y This is done by using various methods such as gradient descent, normal equation, or least squares
Linear regression has some assumptions that need to be met for it to work properly. These assumptions are:

- Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
- Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
- Normality: The data follows a normal distribution.
- Linearity: The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight

2. Explain Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that have identical statistical properties but differ significantly when visualized. The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and not solely relying on summary statistics.
The four datasets in Anscombe's quartet have the following properties:
Dataset I:

- x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset II:

1. x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
2. y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

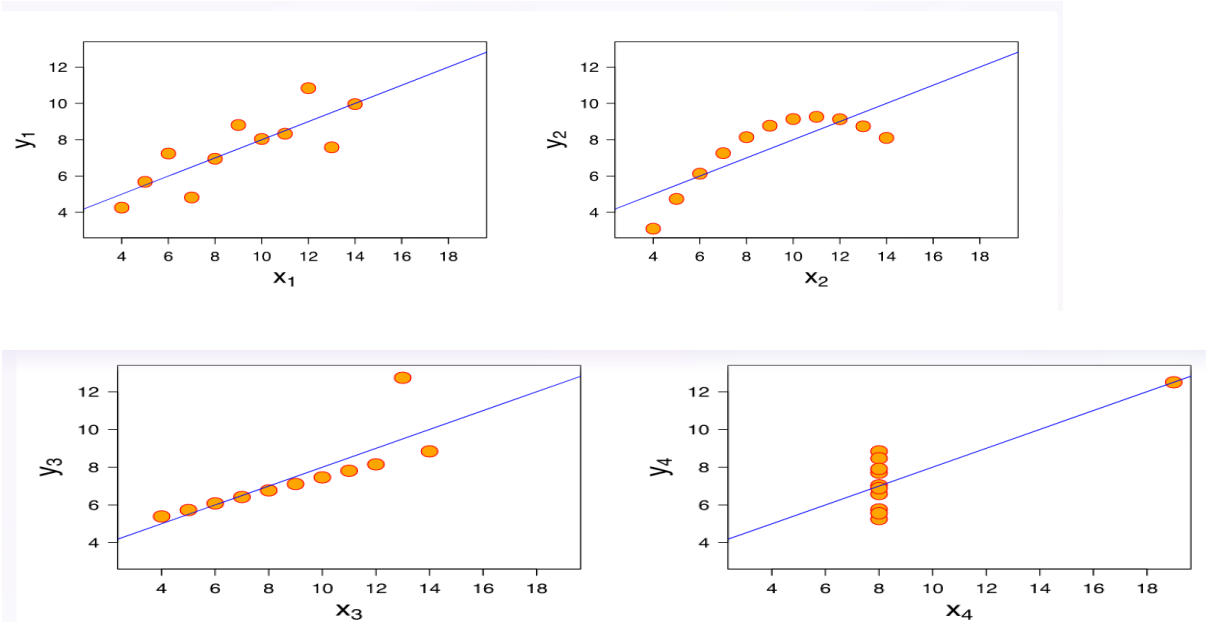
Dataset III:

- x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset IV:

- x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
- y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Although the summary statistics (mean, variance, correlation, etc.) for all four datasets are nearly identical, the datasets exhibit different patterns when visualized. This demonstrates the importance of data visualization in understanding and interpreting data.
When plotting the datasets, the following observations can be made:



- Dataset I: Shows a relatively linear relationship between x and y, with some scatter around the line.
- Dataset II: Also shows a linear relationship, but with an outlier that affects the regression line.
- Dataset III: Appears to have a quadratic relationship, with a slight curve.
- Dataset IV: Appears to have an outlier that drastically affects the correlation coefficient and regression line. It demonstrates the importance of examining the entire dataset and not relying solely on summary statistics.

Anscombe's quartet serves as a powerful reminder that summary statistics alone can be misleading, and visualizing data is crucial for gaining deeper insights and understanding the underlying relationships within the data.

3. What is Pearson's R

The Pearson correlation coefficient (***r***) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- A Pearson's R of 0 means there is no linear correlation between the variables.
- A Pearson's R of 1 means there is a perfect positive linear correlation between the variables, meaning that they increase or decrease together in the same proportion.

- A Pearson's R of –1 means there is a perfect negative linear correlation between the variables, meaning that one variable increases as the other decreases in the same proportion.

Pearson's R can also be visualized as a measure of how close the observations are to a line of best fit. The closer the points are to the line, the higher the absolute value of Pearson's R.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a process of transforming the data values of a variable to a different range or scale.

Scaling is performed for various reasons, such as:
To make the data comparable across different variables that have different units or ranges.
To reduce the effect of outliers or extreme values on the data analysis.
To improve the performance of some machine learning algorithms that are sensitive to the scale of the data.
There are two common types of scaling techniques:

1. Normalized Scaling (Min-Max Scaling):
   - Normalized scaling rescales the data to a specific range, typically between 0 and 1.
   - It preserves the relative relationships and distribution of the original data.
   - The formula for normalized scaling is:
   - *X_scaled = (X - X_min) / (X_max - X_min)*
   - The minimum value of the variable becomes 0, and the maximum value becomes 1.

2. Standardized Scaling (Z-Score Scaling):
   - Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.
   - It centers the data around 0 and scales it based on the standard deviation.
   - Standardized scaling assumes that the data follows a normal distribution.
   - The formula for standardized scaling is:
   - X_scaled = (X - X_mean) / X_std
   - The mean of the variable becomes 0, and the standard deviation becomes 1.

The main difference between normalized scaling and standardized scaling lies in the range and distribution of the scaled values. Normalized scaling preserves the range and distribution of the original data, whereas standardized scaling centers the data around 0 and adjusts it based on the standard deviation.

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

   VIF stands for variance inflation factor, which is a measure of how much the variance of a regression coefficient is increased due to multicollinearity.
   VIF can be calculated as 1 / (1 - R^2), where R^2 is the coefficient of determination of regression of one independent variable on the others.

   Sometimes, the value of VIF is infinite. This happens when there is a perfect correlation between two or more independent variables. In other words, one variable can be expressed exactly by a linear combination of other variables. This causes the denominator of the VIF formula to be zero, resulting in an infinite value.

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

   A Q-Q (quantile-quantile) plot is a graphical tool used to assess the distributional similarity between a given sample of data and a theoretical distribution. It helps to determine whether the observed data follow a specific distribution or if it deviates from it.

   The use and importance of a Q-Q plot in linear regression are as follows:

   - It can be used to check the **normality assumption** of the error terms in a linear regression model.
   - Compare the distribution of residuals across different models or subsets of data. If the Q-Q plots of different models or subsets show similar patterns, it suggests that they have similar error distributions. This is important for checking the homoscedasticity assumption of linear regression, which requires that the error variance is constant across different levels of the predictor variables.
   - It can be used to **detect outliers** or influential observations in the data
   - Model Selection: Q-Q plots can aid in comparing the distributional fit of different models. By comparing the Q-Q plots of the residuals from different models, you can determine which model provides a better fit to the assumed distribution.