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?

Answer:

- People prefer to rent bike when temperatures are high. In our dataset temperature was generally in range of 11-30 degrees.
- Snowing, cloudy and high wind-speeds have negative impact on the rentals.
- From above points we can say that people like to rent bike, in clear weather conditions. Bad weather is bad for the business
- Fewer people rent bike in Spring.
- 2019 has shown a good growth in terms of more rentals in comparison to the previous year. This is good for business.
- From EDA we noticed that:
    - Mid-year booking during May, June, Jul, Aug, Sept, and Oct are high. It increase from the start of the year, and start decreasing post October, till the end of the year.
    - Nobody rents a bike on a heavy rain day.
    - More people rent bike on a working day.

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

Answer:

- During dummy variables creation, we use **drop_first=True** to avoid multicollinearity and prevent the occurrence of the "dummy variable trap."
- The dummy variable trap can lead to creation of unstable or misleading model coefficients.
- For a categorical variable with n categories, we typically need n-1 dummy variables.

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

Answer:

- **temp** has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?**Answer:**

- Dependent variables and independent variables should have a linear relationship.
- All independent variables should have insignificant correlation with each other.
- Error terms should be normally distributed.
- Error terms should have constant variance

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

Answer:

- temp (0.43)
- Light Snow (-0.29). Negative indicates it negatively impact the demand.
- yr (0.24)

  We should keep these features into consideration before making a plan.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer:

Linear regression is a commonly used statistical technique that predicts the connection between a dependent variable (y) and one or more independent variables (x_1, x_2, ..., x_n).

It assumes a linear relationship between the independent variables and the dependent variable, where the dependent variable can be expressed as a linear combination of the independent variables with some added noise.

The equation for simple linear regression (with one independent variable) can be written as:

$y = \beta_0 + \beta_1 * x_1 + \varepsilon$

where y represents the dependent variable, x_1 is the independent variable, β_0 is the y-intercept (the value of y when x_1 is zero), β_1 is the slope of the line (representing the change in y for a unit change in x_1), and ε represents the error term.

For multiple linear regression (with multiple independent variables), the equation expands as:

$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_n * x_n + \varepsilon$

In this case, there are n independent variables (x_1, x_2, ..., x_n), and β_1, β_2, ..., β_n represent the slopes or coefficients associated with each independent variable.

There are assumptions of linear regression:

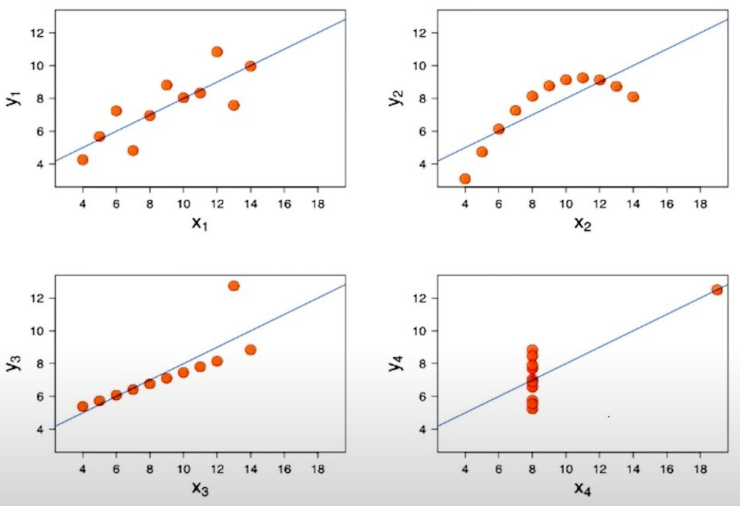
- Dependent variables and independent variables should have a linear relationship.
- All independent variables should have insignificant correlation with each other.
- Error terms should be normally distributed.
- Error terms should have constant variance

2. Explain the Anscombe's quartet in detail.
Answer:

Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties but exhibit distinct patterns when plotted. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and the limitations of relying solely on summary statistics. The quartet is often used to illustrate the concept of "numerical summary equivalence" or "statistical indistinguishability."

Consider the following example with same mean of X and Y, same standard deviation for X and Y, and yet when you visualize them, they are totally different.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



3. What is Pearson's R?
Answer:

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. The formula for Pearson's correlation coefficient is as follows:

$r = (\Sigma[(x_i - \mu_x)(y_i - \mu_y)]) / (n * \sigma_x * \sigma_y)$

where:

r represents Pearson's correlation coefficient.
x_i and y_i are the individual data points of the two variables.
μ_x and μ_y are the means of x and y, respectively.
σ_x and σ_y are the standard deviations of x and y, respectively.
n is the number of data points.

It can take values between -1 and 1. The sign (+/-) of the coefficient indicates the direction of the relationship: positive if the variables have a positive linear association, negative if they have a negative linear association. The magnitude of the coefficient represents the strength of the linear relationship, with values closer to -1 or 1 indicating a stronger linear correlation, while values closer to 0 indicate a weaker correlation or no linear relationship.

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

Scaling refers to the process of transforming variables to a specific range or distribution. It is often performed as a preprocessing step in data analysis and machine learning to bring different variables onto a similar scale, making them more comparable and facilitating meaningful analysis.

Scaling is performed for several reasons:

- Comparable Magnitudes: Variables often have different units or scales, such as age in years and income in dollars. Scaling ensures that the variables are on a similar scale, preventing the dominance of one variable over others due to its larger magnitude.

- Algorithm Requirements: Scaling helps in achieving convergence faster, avoiding numerical instability, and ensuring fairness in the influence of different variables on the algorithm.

- Interpretability: Scaling makes it easier to interpret and compare the coefficients or weights assigned to different variables in regression or linear models. Without scaling, variables with larger values may dominate the model and give a false impression of their importance.

Key differences between normalized scaling and standardized scaling
- Minimum and maximum values are used for normalized scaling, while standardized scaling uses mean and standard deviation.
- We use normalized scaling when features are of different scales, while standardized scaling is used to ensure zero mean, and a unit standard deviation.
- Standardized scaling is less affected by outliers, whereas normalized is really affected by them.
- In Normalized scaling, scaled values are between [-1,1], whereas in standardized scaling there is no range.

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

In case of perfect correlation we have VIF = Infinity.

The VIF for a predictor variable is calculated as the ratio of the variance of the estimated coefficient when that variable is included in the model to the variance of the estimated coefficient when that variable is excluded from the model. Mathematically, the VIF for a predictor variable Xj is computed as:

VIF(Xj) = 1 / (1 - R^2 j)

where R^2 j is the coefficient of determination of the regression model with predictor variable Xj as the response and all other predictor variables as predictors.

In some cases, the VIF value can be infinite. This occurs when the coefficient of determination (R^2j) for the predictor variable Xj is equal to 1. This means that the predictor variable can be perfectly predicted by the other variables in the model, indicating perfect multicollinearity.

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to check if a dataset follows a specific theoretical distribution, like the normal distribution. It compares the quantiles of the dataset against the quantiles of the expected distribution, such as the normal distribution. This plot helps to visually evaluate the similarity between the observed data and the expected distribution.

A Q-Q plot in linear regression is used to assess the assumption of normality for the residuals. It helps determine if the residuals follow a normal distribution. This is important because it ensures the validity of statistical inference and accurate predictions. Deviations from a straight line in the Q-Q plot indicate departures from normality, which can impact the reliability of the regression model and its conclusions.