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)

Ans: In this dataset year, holiday, season (spring, winter etc.), month (January, July, September etc.) Situation (Light_Snow_Rain, Mist_Cloudy etc.), Weekday are the categorical variables.

- Year- The bike rental count has increased in 2019 compared to 2018.
- Holiday: When it's a holiday (Holiday = 1), the demand is found lower compared to non-holidays.
- Season: Highest bike rental in winter and lower in spring season.
- Month: In May to October bike rental demand is high. In September having maximum customer.
- Weather Situation: Demand is higher when the weather is clear.
- Weekday: No big change in bike hire demand on specific weekdays.

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

Ans: drop_first=True is important because this helps to reduce extra column in dummy variables. This helps in reducing correlation among dummy variables. This approach reduces the multicollinearity among the variables.

eg. Assume that the variable "Gender" is categorical with the following levels: Male, Female, and Other. We omitted the 'Other' and generated two dummy variables: Male and Female.

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

Ans: In pairplot 'temp' to 'atemp' is 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? (3 marks)

Ans: Assumptions of Linear Regression after building the model on the training set are as follows-

- Dataset having a linear relationship and strong best-fit line, which we can see in the regression plot in the notebook.
- No Multicollinearity: In the independent variable dataset VIF (Variance Inflation Factor) is less than 5.
- Error terms are normally distributed with mean 0, which we can see Residual Analysis part.

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

Ans: Based on the final model Following are the top 3 features are contributing significantly-

- Temperature – Temperature increases the bike hire numbers also increase.
- Light Snow Rain- In Light Snow and Rain situations bike hire numbers decrease.
- Year- After a year the bike hire numbers also increase.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression algorithm is a powerful tool used to predict and understand the relationship among data. This prediction is based on one or more predictors (independent variable). The algorithm is used as follows-

The basic concept in linear regression is to find the best-fitting line which is straight and this line is through the data points. The equation of this straight line is bellowed:

$$y=\beta_0+\beta_1 x$$

Here, y = dependable variable, x= Independent variable, β0 = intercept to y axis, β1= Slope of the line (change in y if we increase by 1 unit in x)

We can use (*y= mx + c) this* formula also, where m=slope and c= intercept.

For multiple linear regression. Independent variables are multiple so the equation for that is below:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Here, y = dependable variable, x= Independent variable, β0 = intercept to y axis, β1= Slope of the line (change in y if we increase by 1 unit in x), x_1, x_2.. are the independent variables.

Assumptions of Linear Regression:

- Linearity: The relationship is linear between dependent and independent variables.
- Normality: The residuals of the model are normally distributed.
- No Multicollinearity: Not too high correlation between the independent variables.

Evaluation of the model:

- R- Squared- Ranges from 0 to 1 (higher values indicate best fit)
- P- values: <0.05 (Decides which predictor to keep in model)
- VIF: <5 (No Multicollinearity)

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

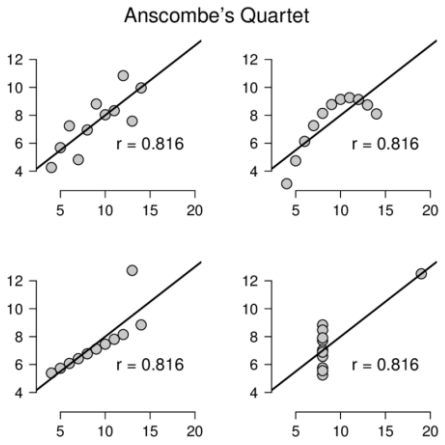
Ans: Anscombe's quartet is a powerful illustration of the limitation of summary statistics. In this technique, we can understand the true nature of data by graphing the data. This method

demonstrates the importance of visualizing the data and shows that only summary statistics can be misleading.

Anscombe's quartet comprises four datasets, which have identical descriptive properties in terms of mean, variance, R- Squared, correlations, and linear regression lines.

- Importance of Anscombe's quartet: Graphical Analysis for pattern- relationship- anomalies, Understanding context, highlights outlier influence, selection of model.

Following is a representative image as example, how this technique can help us to present the limitations present in summary statistics.



Here is how 4 datasets work is shown below:

Dataset No.1: Shows typical relationships, that match the regression line through the data.

Dataset No.2: Shows as a curve. Which is a non-linear relationship, that Linear regression does not capture well.

Dataset No.3: Contains Outlier that significantly affects regression line.

Dataset No.4: most points having the same x value except one outlier, create vertical line, except one point.

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

Ans: In statistics Pearson's R called as Pearsons Correlation Coefficient. This is Correlation Coefficient that measures the linear correlation between two variables. In simple words Pearsons R helps us to understand how two variables are related. It tells us if they tend to increase together or decrease together or no such pattern in it.

Pearson's R number tells us how two variables are related to each other.

Pearson's R ranges from -1 to 1

Example: if Pearson's R is 0.9, it shows that strong positive relationship between two variables. This means if one variable goes up other also goes up. if Pearson's R is -0.8, it shows that strong negative relationship between two variables. This means if one variable down other also goes down. If Pearson's R is 0.2 it means that there is weak correlation between two variable, they don't move together.

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

Ans: Scaling is the process of adjusting the values of independent variables so they will be within a similar range or distribution. This process we do to ensure that each feature contributes equally

Why is scaling performed:

- Equal Contribution: Different features having different units or scales eg. Kilometer, temperature, height etc. without scaling large range can dominate the analysis.
- Improves Performance: If similar scale present in data, then model can work faster.
- Reduce Biases- Due to adjusting data within specific range, model reduces the biases.

Difference between normalized scaling and standardized scaling:

Normalized scaling	Standardized scaling
Adjust the values to a fixed range, usually 0 to 1	Transform the data to having mean of 0 and Standard Deviation of 1
It subtract the minimum value of the feature and divide by the range	It Subtract the mean value of the feature and divide by the Standard Deviation
Formula $$\text{Normalized Value} = \frac{\text{Value} - \text{Min}}{\text{Max} - \text{Min}}$$	Formula $$\text{Standardized Value} = \frac{\text{Value} - \text{Mean}}{\text{Standard Deviation}}$$
Use: Data with outlier	Use: Data is normally distributed

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

(3 marks)

Ans: VIF (Variance Inflation Factor) measures how much variance of regression coefficient is inflated due to multicollinearity among predictor variables.

VIF Formula:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

VIF becomes infinite due to the perfect multicollinearity, where one predictor variable is exactly linearly depend on one or more predictor variables(R squared Equal to 1), then denominator in the VIF calculation become 0 and overall value become infinite.

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

(3 marks)

Ans: Q-Q (Quantile- Quantile) Plot is graphical tool used to compare distribution of data with the theoretical distribution, typically normal distribution. it becomes important to check weather the both data come from same background or not.

This provides the goodness of fit graphically, used to compare two theoretical distributions to each other.