Linear Regression - Subjective Questions

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)

As per the boxplots shown in the Python notebook, the categorical variables and their impact can be observed.

- Season : Spring has least demand, whereas fall has highest demand.
- Weather Situation : Heavy snow and rain results in no demand. Clear weather has high demand.
- Year : Demand in 2019 was considerably more than 2018.
- Month : The months of January, February and December see least demand indicating the weather situation is related to the demand.
- Holiday : Rentals are less during holidays.
- Weekday, working day : Weekday or working day - these two do not have much impact on the demand.

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

When the dummy variable is created for ‘n’ categories, ‘n’ new columns are created. However, if we know the value of the n-1 variables, then the value of the nth variable is already known. For instance, if a bag has red, blue and green marbles only, then if we know which marbles are red or green, it is enough to tell us that the rest of the marbles are blue. Hence this helps in reducing false multicollinearity.

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

The ‘temp’ variable 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? (3 marks)

The assumptions of Linear Regression after building the model on the training set were validated as below.

- Pairplots using the data showed that the variables are linearly related. Linear regression needs the relation between dependent and independent variables to be linear.
- We plotted the error terms, which turned out to be normally distributed with mean as 0.
- Variance Inflation Factor (VIF) was calculated to quantitatively understand the extent of multicollinearity. Linear Regression should have little or no multicollinearity.

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

The top 3 features are as below.

Feature	Co-efficient
Temp (temperature)	0.509836
weathersit_Light Snow & Rain (weather situation)	-0.248777
Yr (year)	0.230502

General Subjective Questions

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

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to the observed data. It can be classified as:

- Simple Linear Regression : When there is only one independent feature
  The equation for simple linear regression is
  $y = \beta 1\, x + \beta 0 + \epsilon$
  Where y is the dependent variable
  x is the independent variable
  $\beta 1$ is the slope
  $\beta 0$ is the y-intercept
  ϵ is the error term
- Multiple Linear Regression : When there are more than one features
  The equation for multiple linear regression is
  $y = \beta 0 + \beta 1\, x1 + \beta 2 x2 + \ldots + \beta n x n + \epsilon$
  Where y is the dependent variable
  X1, x2…xn are the independent variables
  $\beta 1, \beta 2 \ldots \beta n$ are the coefficients for each independent variable
  $\beta 0$ is the y-intercept
  ϵ is the error term

Assumptions of Linear Regression are:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: The observations are independent of each other.
- Homoscedasticity: The residuals (errors) have constant variance at every level of x.
- Normality: The residuals of the model are normally distributed.

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables. The best Fit Line 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 best-fitting line is found using the Ordinary Least Squares

(OLS) method, which minimises the sum of the squared differences between the observed values and the values predicted by the linear model.

Positive linear relationship is when both the dependent and independent variable increase. Negative linear relationship is when the independent variable increases, but the dependent variable decreases.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. This helps to illustrate the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Key Features of Anscombe's Quartet

- Identical Descriptive Statistics:
    - Each dataset has the same mean, variance, correlation coefficient, and linear regression line.
    - This includes having the same values for the mean of x, the mean of y, the variance of x, the variance of y, the correlation between x and y, and the linear regression equation.
- Different Graphical Representations:
    - Despite having similar statistical properties, each dataset looks very different when plotted, illustrating how different patterns can produce the same summary statistics.
    - The visual inspection of the datasets reveals different relationships, anomalies, and structures in the data that are not evident from the statistical summaries alone.

Anscombe's quartet highlights the importance of visualising data before making conclusions based solely on statistical measures. It shows that datasets with identical statistical properties can have very different distributions and patterns, underscoring the need for data visualisation in exploratory data analysis.

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

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.

- r=1 indicates a perfect positive linear relationship
- r=−1 indicates a perfect negative linear relationship
- r=0 indicates no linear relationship

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where, r = correlation coefficient

x_i = values of x variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of Y variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Feature scaling is a method used to normalise or standardise the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as lower, irrespective of the units of the values.

• Normalisation is generally used when it is known that the distribution of data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

• Standardisation, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalisation, standardisation does not have a bounding range. So, even outliers are present in the data, they remain unaffected by standardisation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If a perfect correlation exists, then VIF is infinite. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/(1-R2) infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A quantile (q-q) plot is a graphical technique used to determine whether two data sets come from a normally distributed population.

Using a Q-Q chart: A q-q chart is a plot of the number of the first set of data compared to the number of the second set of data. By majority we mean the fraction (or percentage) of scores below a certain value. So the 0.3 (or 30%) threshold is the point where 30% of the data falls below the value and 70% falls above it. A 45-degree reference line is also

plotted. If two groups in a population have the same distribution, the points should be close to the line. The greater the deviation from this line, the greater the evidence that the two data come from populations with different distributions. Importance of the Q-Q plot: When there are two samples of data, it is often useful to determine whether the joint distribution assumption holds. If so, area and size estimators can combine the two data to obtain an area and size estimate. If two examples are different, it is important to understand some of the differences. The q-q plot may provide more insight into the nature of the differences than analytical methods such as chi-square and 2-sample Kolmogorov-Smirnov tests.