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: In the sprint season demand for bike is less when compared with other seasons.

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

Answer: it helps in reducing the extra column created during dummy variable creation, because drop_first=True is important to use. Hence it reduces the correlations created from the dummy variables.

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

Answer: From all the features, the numerical variable 'registered' has the highest correlation with the target variable 'cnt'.

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

Answer: If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables

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

Answer: To achieve maximum demand for bike it’s recommended to give the below mentioned variables

1. Temperature (0.552)
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
3. year (0.256)

* Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

* Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in * Weathersit3 variable decreases the bike hire numbers by 0.3070 units.

* Year (yr) - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

General Subjective Questions

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

Answer: It’s a machine learning algorithm based on supervised learning. We train a model to predict the behaviour of data based on some variables. The two variables which are on the x-axis (feature/independent/input) and y-axis (target/dependent/output) should be linearly correlated.

it is mostly used for finding out the relationship between variables and forecasting.

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

Answer: Plotting data to confirm the validity of the model fit, like the 4 datasets that have nearly identical simple statistical properties also we can say like equal in terms of the mean and variance of x and y values, but appears different when graphed.

The Pearson correlation between the x and y values is the same, to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

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

Answer: Pearson's R is the measurement of the strength of the relationship between two variables and their association with each other.

Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

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

Answer: **What is scaling:** Data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed: we have to do scaling to bring all the variables to the same level of magnitude like most of the times collected data set contains features highly varying in magnitudes, units and range. If scaling is not done the algorithm takes magnitude in account and not units hence incorrect modelling.

Note: Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between normalized scaling and standardized scaling:

Normalization Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.

The one disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Answer: The perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1 which it led to 1/(1-R2) infinity. To overcome we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer: Quantile-Quantile plots is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

It helps us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution, it's 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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis