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)

   The regression equation from the model:

   cnt=4546.36 * const + 1102.94 * temp + 978.27 * yr + 642.27 * season_4 + 344.12 * season_2 + 236.06 * mnth_9 + 201.17 * workingday + 189.93 * weekday_6- 107.22 * mnth_12 - 133.33 * mnth_11 - 201.41 * partly clouded - 213.62 * hum - 223.81 * windspeed - 358.4 * weathersit_3 - 201.4066* weathersit_2

   Based on the data available, the most favourable seasons for biking are summer and winter.
   Most favourable weather condition is the Clear, Few clouds, Partly cloudy
   Year - 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
   Weekday - bike usage is preferred on Saturdays
   Weathersit - weather condition is the clean/few clouds days is good for biking. even on the light rainy days, ebikes are being used for daily commute.
   Month - favourable months are November, December.

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

   We use pd.get_dummies() to convert categorical variables into dummy variables for regression models. The drop_first=True parameter is important for avoiding a common problem called multicollinearity. Creating dummy variables prevents issues where the model can't tell the unique effect of each variable because they are too closely related. It keeps the model accurate and simple.

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

   “temp” is the variable which has the highest correlation with target variable

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

   - Normal Distribution of error terms: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0.
   - QQ Plot : In linear regression, one key assumption is that the residuals (the differences between the observed and predicted values) are normally distributed. For the results of a linear regression to be reliable, the residuals should be normally distributed. The Q-Q plot helps in visually checking this assumption. If the residuals deviate significantly from normality, it can imply problems with the model, such as omitted variables or the need for transformation.

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

   Temperature, year, weathersit (time of the year)

General Subjective Questions

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

   linear regression is used to understand the relationship between two variables. One variable is the predictor (independent variable), and the other is the response (dependent variable). The goal of linear regression is to find the best-fitting straight line through the data points that predicts the response variable based on the predictor variable.
   In linear regression, the relationship between variables is described with a linear equation: y=mx+c.

   - y: the response variable we want to predict.
   - x: the predictor variable.
   - m: the slope of the line (shows how much y changes for a unit change in x).
   - c: the intercept (the value of y when x is 0).

   Use least method square to find the line that best fits the data. This line minimises the sum of the squared differences between the observed values and the predicted values.

   Assumptions in linear regression:

   - Linearity: The relationship between x and y is linear.
   - Independence: The observations are independent of each other.
   - Homoscedasticity: The residuals (differences between observed and predicted y values) have constant variance.
   - Normality: The residuals are normally distributed.

   By finding the best-fitting line through data points, it helps us understand relationships between variables and make predictions

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

   Anscombe's quartet is a set of four datasets that are used to demonstrate the importance of visualising data before analysing it. Each dataset has nearly identical simple statistical properties, yet they look very different when graphed. It demonstrates that datasets with similar statistical properties can have very different distributions and relationships between variables.

   Only statistical measures are not good enough to depict the data sets. Summary statistics like mean, variance, correlation, and regression lines can give a false sense of understanding about the data.They can be misleading if the data has outliers or a non-linear relationship. Anscombe's underscores the importance of visualising data to get a complete and accurate understanding. The datasets show that a linear model is not always suitable, even if summary statistics suggest so. Visual plots can indicate when a different model might be necessary.

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

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the relationship between two variables. Pearson's R is a number that tells how strongly two variables are related and whether the relationship is positive or negative. The value of Pearson's R ranges from -1 to +1.
+1 - perfect positive relationship. As one variable increases, the other also increases in a perfectly linear way.
-1: perfect negative relationship. As one variable increases, the other decreases in a perfectly linear way.
0: No linear relationship. The variables do not show any trend of increasing or decreasing together.
All other values lies within these limits.

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

Scaling is the process of adjusting the range of features or variables in your data. It is an essential step in data preprocessing, especially when the features have different units or different ranges. Any feature should not get any undue advantage or disadvantage due to its skewed (too large or too small) scale as compared to other features. It ensures that all features contribute equally to the model, rather than letting features with larger ranges dominate the training process.Some machine learning algorithms, like gradient descent-based methods, perform better when features are on a similar scale. scaling can help algorithms converge faster by preventing any one feature from dominating the learning process.

Two types of Scaling:

Normalised scaling : Normalisation (also known as Min-Max scaling) rescales the features to a fixed range, usually 0 to 1 or -1 to 1. The rescaled values fall within the specified range, typically 0 to 1.Outliers can skew the min and max values, affecting the scaling. It is used when distribution of data does not have outliers

Standardised scaling : Standardisation (also known as Z-score normalisation) rescales the features so that they have a mean of 0 and a standard deviation of 1. The transformed data will have a mean of 0 and a standard deviation of 1. Standardisation is less affected by outliers compared to normalisation. It is used when data contains outliers and distribution of the data is unknown or not bounded. Generally using linear and logistic regression.

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

VIF (Variance Inflation Factor) measures how much a variable is correlated with the other variables in a regression model.

$$VIF = \frac{1}{1 - R^2}$$

Sometimes, the VIF value can become infinite, indicating a problem with multicollinearity.
So an infinite value can indicate perfect multicollinearity where one predictor variable in the regression model is an exact linear combination of other predictor variables.
Once cause can be that predictor cannot be uniquely determined; there's redundancy in the data.
Similarly it can happen that even if one variable is almost, but not exactly, a linear combination of others, VIF can be extremely high, approaching infinity.

It can be fixed by Remove Redundant Variables, or by combining variables that combines the correlated variables can solve the issue. Also regularisation techniques such as Ridge Regression and Lasso can handle multicollinearity better than OLS.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool to compare two distributions. It helps us see if a dataset matches a particular theoretical distribution, usually the normal distribution.

In linear regression, one key assumption is that the residuals (the differences between the observed and predicted values) are normally distributed. Checking this assumption is crucial because it affects the validity of the regression model.

On the plot, the x-axis shows the expected quantiles if the data were perfectly normal. The y-axis shows the actual quantiles from your data. If the points closely follow the straight line, data is approximately normally distributed.
If there are deviations to the plot, the model needs to be adjusted accordingly by transforming variables or changing the regression model .

For the results of a linear regression to be reliable, the residuals should be normally distributed. The Q-Q plot helps in visually checking this assumption. If the residuals deviate significantly from normality, it can imply problems with the model, such as omitted variables or the need for transformation.