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 categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.

- **“season” –**
  - Based on the data available, the most favourable seasons for biking are summer and fall.
    - Higher targets can be planned in summer and fall with strategic advertising.
  - Spring has significant low consumption ratio.

- **“workingday” –**
  - Working day represents weekday and weekend/holiday information.
  - The registered users are renting bikes on working days whereas casual users prefer the bikes on non-working days. This effect is nullified when we look at the total count because of the contradictory behaviour of registered and casual users.
  - Registered and casual users’ identity and relevant strategy for working and not working days shall help to increase the numbers.
- **“weathersit” –**
  - Most favourable weather condition is the clean/few clouds days.
  - Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace.
  - There is no data available for heavy rain/snow days.
- **“weekday” –**
  - If we consider “cnt” column we do not find any significant pattern with the weekday.
  - However if the relation is plotted with “registered” users, we observe that bike usage is higher on working days. And with “casual” users it opposite.

- **“yr” –**
  - 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
- **“holiday” –**
  - Holiday consumption of bikes if compared within “registered” and “casual” users then the observation is “casual” users are using bikes more on holiday.
- **“mnth” –**
  - The bike rental ratio is higher for June, July, August, September and October months.
  - 75 quantile grows in the months mentioned in point 1.

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

Using one-not encoding the dummy variables are created to cover the range of values of categorical variable
Each category is transformed into a set of binary variables (0 or 1) indicating the presence or absence of that category.

using drop_first=True when creating dummy variables is a crucial step in handling categorical data in regression models. It eliminates multicollinearity, prevents the dummy variable trap, simplifies interpretation, and makes the model more efficient and stable. It's a best practice when working with categorical variables in regression analysis.

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 i.e. 0.63.

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

Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

Top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and winter season.

General Subjective Questions

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

Linear Regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

One variable, denoted x, is regarded as the **predictor**, **explanatory**, or **independent** variable. The other variable, denoted y, is regarded as the **response**, **outcome**, or **dependent** variable. It assumes a linear relationship between the predictors and the target.

There are 3 different types of linear regression,
Simple Linear Regression: When there is only one independent variable.
Multiple Linear Regression: When there are multiple independent variables.
Polynomial Regression: When the relationship between variables is nonlinear, and polynomial terms are introduced.

In simple linear regression with one independent variable (X) and one dependent variable (Y), the model can be represented as:

$$\mathbf{Y = \beta 0 + \beta 1 * X + \epsilon}$$

Y is the dependent variable (target) you want to predict.
X is the independent variable (feature) that you use to make predictions.
β0 is the intercept, representing the value of *Y* when *X* is 0.
β1 is the coefficient of *X*, representing the change in *Y* for a one-unit change in *X*.
ϵ is the error term, representing the variability in *Y* that is not explained by *X*. It is assumed to be normally distributed and have a mean of 0.

The primary goal of linear regression is to find the best-fitting linear relationship (a straight line) between the independent and dependent variables. This involves estimating the coefficients (β0 & β1) that minimize the sum of the squared differences (residuals) between the observed data points and the predicted values.

Once the coefficients are estimated, we can use the linear regression equation to make predictions.

Common metrics used or evaluating the performance of a linear regression model are:
Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.
R-squared (R^2): Represents the proportion of the variance in the dependent variable that is explained by the independent variables. A higher R^2 indicates a better fit.
Adjusted R-squared: Adjusts R^2 for the number of predictors, preventing overfitting.
Residual Analysis: Examining the distribution of residuals to check for violations of linear regression assumptions (e.g., normality, homoscedasticity).

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

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
$Σx^2$ = the sum of squared x scores
$Σy^2$ = the sum of squared y scores

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

- What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in multiple linear regression models. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, which can lead to unstable coefficient estimates and reduced interpretability of the model. VIF is used to quantify the degree of multicollinearity, and it is calculated for each predictor variable in the model.

The formula to calculate VIF for a predictor variable is:

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

Where:

R^2 is the coefficient of determination of the predictor variable regressed on all the other predictor variables in the model.
A VIF value of 1 indicates no multicollinearity (i.e., the predictor variable is not correlated with other predictors). As VIF values increase beyond 1, it suggests increasing multicollinearity.

However, VIF can become infinite (or extremely large) in some situations. This occurs when there is perfect multicollinearity, which means that one or more of the predictor variables can be perfectly predicted by a linear combination of the other predictor variables in the model. Perfect multicollinearity leads to an R^2 value of 1 in the VIF formula, resulting in an infinite VIF.

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

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of

type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- Interpretations
    - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
    - Y values < X values: If y-values quantiles are lower than x-values quantiles.
    - X values < Y values: If x-values quantiles are lower than y-values quantiles.
    - Different distributions – If all the data points are lying away from the straight line.
- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.