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: From the boxplots, we can see that count of total rental bikes(dependent variable) is :

   - higher during the fall season and relatively low in spring season
   - increases significantly in the year 2019 from 2018
   - increases somewhat linearly in first half of the year and then it starts decreasing. During September, bike sharing is more.
   - higher on days (other than holidays), probably because people commute to office
   - count of total rental bikes is higher during clear weather and dips during rainy days
   - Additionally, correlation between temp and atemp is quite high
   - Weekday is not giving clear picture about demand.

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

   Ans: Because for any categorical variable with p levels, we need (p-1) dummy variables to explain the values. Hence, we use **drop_first=True t**o limit the dummy predictor variable columns.

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

   Ans: temp and atemp have the highest correlation with Target variable(cnt)

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

   Ans: By Predicting and Evaluating on test set. Applied same pre-processing transformation steps to test data set as well as to the training data set. Computed R^2 for test set and then compared R^2 values for train and test set.

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: Top 3 variables contributing significantly towards the demand of shared bikes:

   a) Temp
   b) yr
   c) winter season

General Subjective Questions

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

   Ans: Linear Regression algorithm attempts to explain the relationship between Dependent variable and independent variables using a straight line. The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables. The aim is to find the best fit line on basis of training data to evaluate output for test data.

   - Model Representation:

   The linear regression equation for a simple case with one independent variable is:

   $y = b_0 + b_1X_1 + b_2X_2 + b_3xX_3 + \ldots\ldots\ldots.b_nX_n + e$

   y represents the dependent variable.

   x_i represents the independent variable.

   b_0 is the intercept (where the regression line crosses the y-axis).

   b_i is the coefficient (the change in y for a unit change in x ensuring other dependent variables are constant).

   e represents the error term (the difference between the observed and predicted values).

   - Assumptions:

   Linearity: The relationship between variables is linear.

   Independence: The residuals (errors) are independent of each other.

   Homoscedasticity: The variance of residuals is constant across all levels of the independent variable.

   - Steps to perform Multiple Linear Regression:

   1) Reading and understanding the data
   2) Cleaning the data
   3) Visualizing the data using EDA
   4) Preparing the data for modelling(train-test split, rescaling)
   5) Training the model
   6) Residual Analysis
   7) Predictions and Evaluations on test set

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

   Ans:

   Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties but display vastly different patterns when graphed. It highlights the limitations of relying solely on summary statistics and emphasizes the importance of visualizing data to truly understand its nature and relationships. Each dataset in Anscombe's quartet has its unique set of x and y values, and their respective equations will vary depending on the relationships present in those datasets. These equations are derived through statistical methods to best fit the data points within each dataset.

   It is used for:

- Data Visualization Advocacy: It underscores the significance of visualizing data to truly understand its nature. It encourages analysts to graphically explore datasets to uncover hidden patterns, relationships, and potential outliers that might not be apparent through summary statistics alone.
- Statistical Analysis Awareness: It highlights the limitations of relying solely on summary statistics, stressing the need for caution when drawing conclusions based solely on these numerical measures.
- Model Validation and Assumptions Testing: It serves as a practical tool for assessing the assumptions of statistical models. Analysts use it to check if models assume linear relationships, homoscedasticity, and normality of residuals hold true, thereby ensuring the reliability of statistical analyses.

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

   Ans: Pearson's correlation coefficient (often denoted as "r") is a measure that tells how strongly two variables are related in a linear manner. It ranges from -1 to 1, where:

   - 1 indicates a perfect positive linear relationship: As one variable increases, the other variable also increases proportionally.

   -1 indicates a perfect negative linear relationship: As one variable increases, the other decreases proportionally.

   -0 indicates no linear relationship between the variables.

   Pearson's "r" helps in understanding the direction (positive or negative) and strength of the relationship between two continuous variables. However, it specifically measures linear relationships and might not capture non-linear associations.

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

   Ans: Scaling is a process implemented on variables to standardize the values so that we get optimal and standard coefficients and an effective Linear Model.

   Scaling is important for:

   - Correct interpretation of coefficients

   - Gradient descent method(which minimizes cost function) becomes more effective

   There are two major methods to scale the variables, i.e., standardisation and MinMax scaling. Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

   - Min-max scaling (normalisation): Values Between 0 and 1
     - Formula: (x-xmin)/(xmax-xmin)
   - Standardisation (mean - 0, sigma-1)
     - Formula: (x - mu)/sigma

   It is advisable to use min max scaling because it takes care of Outliers.

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

Ans: Value of VIF for any ith variable is infinite because that ith variable is highly correlated to other predictor variables. Because pf which R squared error value for that variable is 1. As a result, when we feed the same to VIF formula(VIF(i) = 1/(1-r^2)), VIF values comes out to be 1/0 which is infinite

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans: A Q-Q plot, short for quantile-quantile plot, is a graphical method used to compare two probability distributions. In the context of linear regression, it's often utilized to assess if the residuals (the differences between observed and predicted values) follow a normal distribution, which is an assumption of many linear regression models.

Here's how it works and why it's important in linear regression:

- Understanding the Q-Q Plot:

The Q-Q plot compares the quantiles of the residuals to the quantiles of a theoretical normal distribution.
If the residuals are normally distributed, the points in the Q-Q plot will fall along a straight line (the line of equality) at a 45-degree angle.
Deviations from this straight line indicate departures from normality in the residuals.

- Use in Linear Regression:

Normality of residuals is an essential assumption in linear regression. If the residuals are not normally distributed, it might affect the reliability of the regression results.
A Q-Q plot helps to visually assess the assumption of normality. If the points on the plot deviate significantly from the straight line, it suggests that the residuals might not follow a normal distribution.

- Importance in Model Assessment:

Identifying non-normality in residuals through a Q-Q plot prompts further investigation. It could indicate issues like outliers, heteroscedasticity (unequal variance), or other problems that might affect the validity of the regression model.
Addressing issues found in the Q-Q plot might involve data transformation, using different models, or employing robust regression techniques to make the model more reliable.