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?

Ans: Based on the analysis of the categorical variables, it can be inferred that :

- **Season -** There is a high number of bike rentals in the fall season, followed by summer season and then by winter season.
- **Weekday** - There is a high number of bike rentals on Thursday, followed by Wednesday and Saturday. The bike rentals are least on Tuesday and Sunday
- **Weathersit** - There is a high number of bike rentals on clear sky days, and the lowest number of bike rentals on Light snow days.
- **Holiday** - There is a high number of bike rentals on non-holiday days.
- **Workingday** - The maximum number of bike rentals are made on working days, bu there is not significant difference in rentals on working and non-working days
- **Month** - The bike rentals are very much high during the months of September, October and August, and the lowest during the months of January and February.
- Bike rentals is high in 2019, as compared to in 2018.

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

Ans: It is very important to use drop_first = True, as it helps to avoid multicollinearity and improves the performance and stability of our model.

- **Multicollinearity** - When we create dummy variables without dropping the original column, then, each category of the original categorical variable is represented as a separate dummy variable. This creates a multicollinearity issue.
- **Improved Model Performance -** Removing one dummy variable for each categorical variable reduces the dimensionality of the dataset. Fewer variables can lead to faster model training and better model performance.

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

Ans : By looking at the pair-plot among the numerical variables, the numerical variable that has the highest correlation with the target variable, is the “temp” variable. We can see that their is a Linear relationship.

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

Ans: The assumptions of Linear Regression were validated by the following:

- **Residual Analysis -** The residuals were calculated by subtracting the predicted values from the actual target values for the training set.
- **Linearity Assumption** - We plotted the residuals against the predicted values. A scatter plot was created to show a linear pattern.

- **Outlier Detection** - The data was treated for outlier values.
- **Multicollinearity** - The model was checked for multicollinearity using VIF, where, a column having a VIF >10 was dropped.

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

Ans : The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- **temp**(Temperature) - With a coefficient of 0.4914, the 'temp' variable is contributing significantly towards the target variable.
- **yr**(Year) - With a coefficient of 0.2334 , the 'yr' variable is contributing significantly towards the target variable.
- **Light Snow -** With a coefficient of -0.3041 , the 'Light snow' variable is contributing significantly towards the target variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans - Linear Regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features or predictors). It aims to find a linear relationship that best describes how changes in the independent variables are associated with changes in the dependent variable.

When the number of the independent feature is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

Assumption for Linear Regression Model -

1. **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
2. **Independence**: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3. **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
4. **Normality**: The errors in the model are normally distributed.

5. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

- **Hypothesis function for Linear Regression** : y = mx + c,
  Where, m = coefficient of x
  C = intercept
- **Objective** - The goal of Linear Regression is to find the values of m1,m2,m3.., such that the sum of the squared differences between the observed values of the dependent variable and the predicted values generated by the linear equation, is minimum.
- **Model Training -** To train a Linear Regression model, we need a dataset containing both the independent variables (features) and the dependent variable (target). The dataset is then splitted into train and test data. Train data is the data on which the model is trained or learns, and test data is the one on which the model is tested.
- **Making Predictions** - After training, the model is used to make predictions. Given new values of the independent variables (X1, X2, ... Xn), the model predicts the corresponding value of the dependent variable (Y).
- **Evaluation**: The performance of the Linear Regression model can be evaluated using various metrics, such as:
  R-squared (R^2): A measure of how well the model explains the variance in the dependent variable.
  Mean Squared Error (MSE): A measure of the average squared difference between predicted and actual values.
  Root Mean Squared Error (RMSE): The square root of MSE, providing the error in the original units of the target variable.
- **Assumptions** - Linear Regression assumes that the relationship between the variables is linear, the residuals (error terms) are normally distributed, and the residuals have constant variance (homoscedasticity). It also assumes that the independent variables are not highly correlated (no multicollinearity).

2. Explain the Anscombe's quartet in detail.

Ans - **Anscombe's quartet** is a set of four datasets that have nearly identical simple descriptive statistics (e.g., mean, variance, correlation) but exhibit very different patterns and relationships when plotted visually. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. It serves as a powerful reminder that data visualization is crucial for understanding and interpreting data. Here's a detailed explanation of Anscombe's quartet:

Dataset 1:

X1 (independent variable): 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
Y1 (dependent variable): 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82
Dataset 2:

X2 (independent variable): 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
Y2 (dependent variable): 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26

Dataset 3:

X3 (independent variable): 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
Y3 (dependent variable): 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42
Dataset 4:

X4 (independent variable): 8, 8, 8, 8, 8, 8, 8, 19, 8, 8
Y4 (dependent variable): 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91
Key Observations:

Similar Descriptive Statistics: When we calculate basic summary statistics (mean, variance, correlation) for each of the four datasets, we can see that they are nearly identical. This can mislead someone who relies solely on these statistics into thinking that the datasets are similar.

Different Data Patterns: When we plot the data from these datasets, we can see starkly different patterns:

Dataset 1: A relatively linear relationship.
Dataset 2: Linear with an outlier.
Dataset 3: Non-linear relationship.
Dataset 4: Dominated by a single outlier, which significantly affects the correlation.

In summary, Anscombe's quartet is a collection of four datasets that demonstrate the importance of data visualization in understanding data and the limitations of relying solely on summary statistics. It illustrates how different datasets with the same basic statistics can have very different patterns and relationships when visualized.

3. What is Pearson’s R?

Ans - Pearson's correlation coefficient, often denoted as "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is used to assess the degree to which two variables are linearly related to each other.

Key characteristics of Pearson's correlation coefficient (r):

Range: The value of r lies between -1 and 1.

r = 1: Perfect positive linear correlation (as one variable increases, the other increases proportionally).

r = -1: Perfect negative linear correlation (as one variable increases, the other decreases proportionally).
r = 0: No linear correlation (variables are not linearly related).

Direction: The sign of r (+ or -) indicates the direction of the linear relationship:

Positive r: Indicates a positive linear relationship (both variables tend to increase or decrease together).
Negative r: Indicates a negative linear relationship (one variable tends to increase as the other decreases).

Strength: The magnitude (absolute value) of r quantifies the strength of the linear relationship. The closer r is to 1 or -1, the stronger the linear correlation. Values closer to 0 represent weaker or no linear correlation.

Assumptions:

1. Assessing relationships between variables in scientific research.
2. Analyzing the strength and direction of associations in social sciences.
3. Identifying correlations in finance, such as the relationship between two stock prices.
4. Evaluating the correlation between features in machine learning to select relevant predictors.
5. Checking for multicollinearity in regression analysis.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans - Scaling is a preprocessing technique used in data analysis and machine learning to transform the numerical values of different features (variables) into a similar scale or range. The primary goals of scaling are to make the data more suitable for certain algorithms and to improve the interpretability of the results.

Normalized Scaling (Min-Max Scaling) -

- Normalized scaling scales the data to a specific range, typically between 0 and 1.
- The formula to perform Min-Max scaling for a feature :

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- The minimum value in the original data is mapped to 0, and the maximum value is mapped to 1. All other values are linearly scaled between these two extremes.
- Normalized scaling is suitable when you want to preserve the relationships between the original values but ensure they fall within a specific range.

Standardized Scaling (Z-score Scaling):

Standardized scaling (also known as Z-score scaling) transforms the data so that it has a mean of 0 and a standard deviation of 1.

The formula to perform standardized scaling for a feature

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

The data distribution is centered around 0, and the spread of the data is adjusted so that it has a standard deviation of 1.
Standardized scaling is appropriate when we want to eliminate the scale entirely, making features comparable without regard to their original units. It is commonly used in techniques like Principal Component Analysis (PCA) and clustering.

In summary, scaling is a crucial preprocessing step in data analysis and machine learning to ensure that numerical features are on a similar scale, making models more effective and interpretable.

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

Ans - The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple linear regression model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it challenging to isolate the individual effects of these variables on the dependent variable. High VIF values indicate a strong correlation between a predictor variable and the other predictor variables in the model.

- To solve this problem, we need to drop one of the variables from the dataset, that is causing perfect multicollinearity.
- If removing variables is not an option, we can use regularization techniques like Ridge regression or Lasso regression, which can help mitigate multicollinearity by adding a penalty term to the coefficients.

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

Ans - A Q-Q plot (Quantile-Quantile plot) is a graphical tool used in statistics and data analysis to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. It is a visual comparison between the quantiles of the observed data and the quantiles of the theoretical distribution. Here's an explanation of the use and importance of a Q-Q plot in linear regression:

How to Interpret a Q-Q Plot:

In a Q-Q plot:

The x-axis represents the quantiles of the theoretical distribution (e.g., the expected quantiles if the data were normally distributed).
The y-axis represents the quantiles of the observed data. The points on the Q-Q plot are typically scatter-plotted, with each point corresponding to a quantile from the observed data.

Q-Q plots are a valuable tool in linear regression and data analysis to assess the normality assumption of residuals. They provide a visual way to check whether the observed data follows a theoretical distribution, helping analysts make informed decisions about model validity, data transformations, and the need for alternative statistical techniques.