Assignment-based Subjective Questions -

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans : Categorical variables can have a significant role on the dependent variable in statistical analysis. It depends on the nature and type of the categorical variables and the statistical methods used for analysis.
Following are the key points which shows the effect of categorical variables on the dependent variables:

- Before including them in the analysis, we first need to encode them into the numerical format. We can call them as dummy variables. This is the important step of data preparation.
- We can then concatenate these dummy variables in the data format to check their influence based on the binary variables. Also, we can drop the variables which are having negligible effect on the dataset.
- Then we are able to analyze the effect of the categorical variables on the data set.

Q2. Why is it important to use drop_first=True during dummy variable creation?
Ans: This is important to prevent multicollinearity, which can lead to unstable and unreliable regression results. This can further result in poor model generalization and difficulties in interpreting the effects of individual variables.

So, after using **drop_first=True** avoids multicollinearity, simplifies the interpretation of the model and ensures that the intercepts have meaningful interpretation as the expected value when all the categorical variables are at their same levels.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: “atemp” and “temp” variables have the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: It is a very crucial step in building a reliable model. After building a linear regression model on the training set, we can diagnose and test the train data to validate the model assumptions. Here are the few assumptions:-

- **Residual Analysis**:- We can calculate the residuals by taking differences between the observed and expected values for the training set.
- **Linearity Assumptions**:- We can plot the residuals against the predicted values. The residuals should be randomly scattered around zero without any noticeable patterns. A pattern in the residual plot could indicate a violation of the linearity assumption.
- **Homoscedasticity**:- Homoscedasticity refers to the condition where variance of the residuals in a regression model is consistent. The scatter plot of the residuals against the predicted value or any independent variable shows homoscedasticity.
- **Normality of Residuals**:- After plotting a histogram of the residuals will show normality.
- **Multicollinearity Assumption**:- We need to calculate the VIF (Variance Inflation Factor) for each predictor variable to check the multicollinearity. High VIF (above 5 or 10) can indicate multicollinearity issues.
- **Model fit:-** We can access the overall fit of the model using metrics like R-squared, adjusted R-squared or Root Mean Square Error (RMSE)

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: The top features contributed significantly towards explaining the demand of the shared bikes are **"September", "Summer" and "Working day".**
Hence when the situation comes back to normal, the company should come up with new offers and schemes in September when the weather is pleasant and also advertise a little for Saturday as this is when business would be at its best.

General Subjective Questions -

Q1. 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 and one or more independent variables by fitting a linear equation to the observed data. It is one of the simplest and most widely used regression techniques for predictive modeling and statistical analysis. Here's a detailed explanation of the linear regression algorithm:

- **Problem Statement:**- Linear regression is typically used when we want to predict a continuous numeric output variable (dependent variable) based on one or more numeric input variables (independent variables or predictors).

- **Basic Equation**:- The linear regression model is based on a linear equation of the form:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{\square} X_{\square} + \varepsilon$$

  Where, Y represents the dependent variable (the predicted variable),
  $X_1, X_2, ..., X_{\square}$ are the independent variables (predictors).
  $\beta_0, \beta_1, \beta_2, ..., \beta_{\square}$ are the coefficients of the model, representing the intercept and slopes.
  ε represents the error term, which accounts for the random variation in the data that the model can't explain.

- **Objective**:- The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple regression) that minimizes the sum of squared differences between the observed values of the dependent variable and the values predicted by the model.

- **Model Training**:- To train the linear regression model, we need a dataset with known values of both the dependent and independent variables. The algorithm estimates the coefficients (β_0, β_1, β_2, etc.) that best fit the data.

- **Coefficient Estimation**:- The coefficients are estimated using a technique called ordinary least squares (OLS). OLS finds the values of β_0, β_1, β_2, etc., that minimize the sum of the squared residuals (the differences between the observed and predicted values). It does this by solving a system of linear equations.

- **Model Assumptions**:- Linear regression relies on several assumptions:
  - Linearity: The relationship between the independent and dependent variables is linear.
  - Independence: The residuals (errors) are independent of each other.
  - Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.
  - No multicollinearity: The independent variables are not highly correlated with each other.
- **Model Evaluation**:- After training the model, we need to evaluate its performance. Common evaluation metrics for linear regression include:
  - R-squared (R^2): Measures the proportion of the variance in the dependent variable that is explained by the model.
  - Mean Squared Error (MSE): Measures the average squared difference between observed and predicted values.

    - Root Mean Squared Error (RMSE): The square root of MSE, providing the same units as the dependent variable.

- **Interpretation**:- Linear regression coefficients (β values) provide insights into the strength and direction of the relationships between independent variables and the dependent variable. A positive coefficient indicates a positive correlation, while a negative coefficient indicates a negative correlation.

In summary, linear regression is a fundamental algorithm for modeling and predicting continuous outcomes based on linear relationships between variables. It's relatively simple and interpretable, making it a valuable tool in various fields, including economics, finance, biology, and social sciences. However, it's essential to understand its assumptions and limitations when applying it to real-world data.

Q2. Explain the Anscombe’s quartet in detail.
Ans: Anscombe's quartet is a famous dataset in statistics and machine learning that consists of four small datasets, each containing 11 data points. These datasets were created to demonstrate the importance of visualizing data and the potential pitfalls of relying solely on summary statistics like means and variances. Anscombe's quartet is often used to illustrate how different datasets with very different characteristics can have nearly identical summary statistics. Let's delve into each dataset within Anscombe's quartet:

Dataset I:
x: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Dataset II:
x: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y: [9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]

Dataset III:
x: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Dataset IV:
x: [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]

Now, let's explore some important observations and lessons from Anscombe's quartet:

Summary Statistics Can Be Deceptive: When we calculate basic summary statistics (mean, variance, correlation) for each dataset, we will find that those are nearly identical. This can lead to the false assumption that the datasets are very similar, but a graphical analysis reveals otherwise.

Visualization Is Crucial: To truly understand data, it's essential to visualize it. When we create scatterplots of these datasets, we will see that they have very different patterns. Dataset I exhibits a roughly linear relationship, Dataset II is curved, Dataset III has an outlier, and Dataset IV has a clear outlier that significantly impacts the regression line.

The Importance of Outliers: Outliers, even a single one, can have a substantial impact on the regression line and the overall interpretation of data. In Dataset IV, the presence of a single outlier at (19, 12.5) drastically changes the regression line's slope.

Context Matters: When working with real-world data, understanding the context is crucial. We can't solely rely on mathematical metrics; we need to consider the domain knowledge and the underlying processes that generated the data.

In machine learning and data analysis, Anscombe's quartet serves as a reminder of the limitations of summary statistics and the power of data visualization. It underscores the need to explore data visually before making assumptions or building models, as well as the importance of robustness against outliers and understanding the uniqueness of each dataset.

Q3. What is Pearson’s R?
Ans: Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is widely used to assess the degree to which two variables are linearly related to each other. Pearson's r is a value between -1 and 1, where:

- r = 1 indicates a perfect positive linear relationship.
- r = -1 indicates a perfect negative (inverse) linear relationship.
- r = 0 indicates no linear relationship (variables are uncorrelated).

Key characteristics of Pearson's correlation coefficient:

Linear Relationship: Pearson's r measures the linear association between two variables. It assumes that the relationship between the variables can be adequately described by a straight-line (linear) relationship.

Symmetry: The correlation coefficient is symmetric, meaning that swapping the order of the variables does not change the value of r. In other words, the correlation between variable X and variable Y is the same as the correlation between Y and X.

Normalization: Pearson's r is a normalized measure, which means it is not affected by changes in the scale or units of measurement of the variables. It standardizes the covariance between the two variables by dividing it by the product of their standard deviations.

Interpreting the value of Pearson's r:
If r is close to 1, it suggests a strong positive linear relationship, meaning that as one variable increases, the other tends to increase proportionally.
If r is close to -1, it suggests a strong negative (inverse) linear relationship, indicating that as one variable increases, the other tends to decrease proportionally.
If r is close to 0, it suggests little to no linear relationship between the variables.

It's important to note that Pearson's correlation coefficient assesses only linear relationships. It may not capture other types of associations, such as nonlinear relationships or dependencies that are not well-described by a straight line. Additionally, correlation does not imply causation; a high correlation between two variables does not necessarily mean that one variable causes the other.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is a preprocessing technique in machine learning that involves transforming the features (independent variables) of a dataset to bring them into a similar numerical range or distribution. Scaling is performed to ensure that all features contribute equally to the learning process and to improve the performance of machine learning algorithms. The primary goal of scaling is to remove any potential biases introduced by the differences in the scales of features.

Here are the key reasons why scaling is performed in machine learning:

Equal Contribution: Machine learning algorithms often use mathematical calculations to make predictions or decisions. If the features have different scales, features with larger scales can dominate the learning process and have a disproportionate influence on the model's output. Scaling helps ensure that all features contribute evenly to the model.

Faster Convergence: Many machine learning algorithms, such as gradient descent, converge faster when the features are on a similar scale. This can lead to quicker training and shorter computation times.

There are two common methods for scaling features: normalized scaling (min-max scaling) and standardized scaling (z-score scaling).

- **Normalized Scaling (Min-Max Scaling):-**
  Normalized scaling scales the features to a specific range, typically between 0 and 1. It preserves the original distribution of the data but transforms it to a common scale.Formula:-

  $$X\_scaled = (X - X\_min) / (X\_max - X\_min)$$

  where X is an individual feature value, X_min is the minimum value of that feature in the dataset, and X_max is the maximum value.

- **Standardized Scaling (Z-Score Scaling):-**
  Standardized scaling transforms the features to have a mean (average) of 0 and a standard deviation of 1. This method centers the data around zero and scales it to have a consistent spread. Formula,

  $$X\_scaled = (X - \mu) / \sigma$$

  where X is an individual feature value, μ (mu) is the mean of that feature in the dataset, and σ (sigma) is the standard deviation.

In summary, scaling is a crucial preprocessing step in machine learning to ensure that features are on a similar scale and contribute equally to model training. Normalized scaling and standardized scaling are two common methods for achieving this goal, each with its own advantages and use cases. The choice between them depends on the characteristics of our data and the requirements of our machine learning algorithm.

Q5. 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 regression model, specifically when multiple independent variables are highly correlated with each other. High VIF values indicate that one or more independent variables can be predicted from the others, which can lead to instability in the regression model's coefficient estimates. VIF is calculated for each predictor variable, and it quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity.

The formula to calculate the VIF for a predictor variable:-

$$VIF(Xi) = 1/1-(R_{Xi})^2$$

Here are some reasons why a predictor variable might have an infinite VIF:

Perfect Multicollinearity: Perfect multicollinearity occurs when one predictor variable can be expressed as a perfect linear combination of one or more other predictor variables. For example, if we have two predictor variables that are identical, we will encounter perfect multicollinearity, resulting in infinite VIF values.

Linear Dependence: When there is a linear dependence among the predictor variables.It implies that one variable can be perfectly predicted from another, leading to infinite VIF.

Dummy Variable Trap: In some cases, when encoding categorical variables using dummy variables, we may inadvertently create perfect multicollinearity if one category can be predicted from the others. This can result in infinite VIF for the dummy variables.

To address this issue, we should carefully review our dataset and the way predictor variables are defined. Detect and resolve multicollinearity by considering variable selection techniques, combining correlated variables, or removing redundant variables. In the context of dummy variables for categorical data, it's essential to use proper encoding schemes (e.g., dropping one category as the reference) to avoid the dummy variable trap.

Ultimately, detecting and handling multicollinearity is crucial to ensure the stability and interpretability of our regression model.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution or to compare two datasets for similarity in distribution. Q-Q plots are especially useful for evaluating whether a dataset is approximately normally distributed, which is a common assumption in linear regression and many other statistical techniques.

How a Q-Q Plot Works:

- Ordered Data: To create a Q-Q plot, we can start by sorting the values in our dataset in ascending order.

- Theoretical Quantiles: We can then calculate the quantiles of the theoretical distribution we want to compare our data to. For example, if we are assessing normality, we can calculate the quantiles of the standard normal (Z) distribution.

- Plotting: We can plot the ordered data against the corresponding theoretical quantiles. Each data point in the ordered dataset is matched with a quantile from the theoretical distribution. The x-axis represents the quantiles of the theoretical distribution, while the y-axis represents the quantiles of our dataset.

Interpretation and Use of a Q-Q Plot in Linear Regression:
The Q-Q plot provides a visual comparison between the empirical quantiles of our dataset and the expected quantiles of the theoretical distribution.

Here's how it is used in linear regression:
Normality Assessment: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots are often used to assess the normality of these residuals.

Ideal Case: In an ideal Q-Q plot for normally distributed residuals, the points should closely follow a straight line with a slope of 45 degrees (the line y = x). This means that our residuals have a normal distribution.

Departure from Normality: Deviations from the straight line suggest departures from normality. If the points curve upward or downward away from the line, it indicates non-normality. For example, heavy tails or skewness might be observed in the data.

Residual Diagnostics: By examining the Q-Q plot of residuals, we can detect deviations from normality. Departures from normality may suggest issues with the model assumptions, which can affect the validity of regression results.

Outlier Detection: In a Q-Q plot, extreme outliers often appear as points that deviate significantly from the expected quantiles. These outliers may represent unusual data points or errors in the model.

Model Improvements: If non-normality is detected in the residuals, we may need to consider transformations of the dependent variable or additional model adjustments to account for this departure from normality.

In summary, a Q-Q plot is an essential tool in linear regression for assessing the normality assumption of the residuals and detecting potential outliers. It provides a visual way to evaluate how well our data aligns with a theoretical distribution, helping us to make informed decisions about the validity of our regression model and the need for any adjustments.