Assignment-based Subjective Questions

Q.1 : **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

A.1 : I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization:

- Fall season seems to have attracted more bookings. And, in each season thebooking count has increased drastically from 2018 to 2019.
- Most of the bookings have been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more bookings as compared to the start of the week.
- When it's not a holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day

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

A.2 : drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax: drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

A.3 : **'temp'** variable has the highest correlation with the target variable.

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

A.4 : I have validated the assumption of Linear Regression Model based on below 5 assumptions:

1. Normality of error terms
2. Multicollinearity check
3. Linear relationship validation
4. Homoscedasticity
5. Independence of residuals

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

A.5 : Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. temp
2. winter
3. sep

General Subjective Questions

Q.1 : **Explain the linear regression algorithm in detail.**

A.1 : Linear regression is a popular and fundamental supervised machine learning algorithm used for predictive modeling and understanding the relationship between a dependent variable (target) and independent variables (features).

The goal of linear regression is to find the best-fitting linear relationship between the independent variables (features) and the dependent variable (target) in a dataset. This relationship is expressed as a linear equation of the form:

$Y=mX+b$

Where:

Y is the dependent variable (the target you want to predict).
X is the independent variable (one or more features).
m is the slope (coefficient) of the line, representing the effect of X on Y.
b is the y-intercept, indicating the value of Y when X is zero.

Machine learning models can be classified into the following three types based on the task performed and the nature of the output:

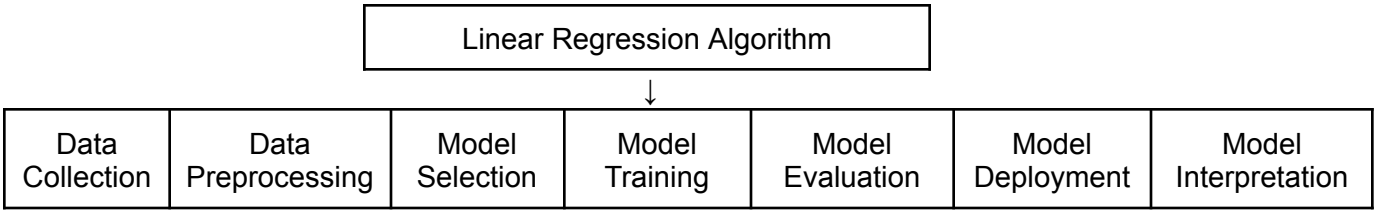
1. **Regression**: The output variable to be predicted is a continuous variable.
2. **Classification**: The output variable to be predicted is a categorical variable.
3. **Clustering**: No predefined notion of label allocated to groups/clusters formed.

Here, There is two main type of linear regression:

1. **Simple linear regression**: The simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line.

2. **Multiple linear regression**: Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables.

Linear Regression Algorithm:



Linear Regression is a simple yet powerful algorithm for modeling the relationship between input variables and a continuous target variable.

Q.2 : **Explain the Anscombe's quartet in detail.**

A.2 : Anscombe's quartet is a famous statistical example consisting of four distinct datasets, each containing 11 (x, y) data points.The uniqueness of Anscombe's quartet lies in the fact that all four datasets have nearly identical simple descriptive statistics, such as means, variances, correlations, and linear regression coefficients.

Here's a detailed explanation of each dataset in Anscombe's quartet:

Dataset 1:

This dataset appears to have a straightforward linear relationship between the x and y variables.
The linear regression line is a good fit.
The correlation coefficient is close to 1, indicating a strong positive linear relationship.

Dataset 2:

Dataset 2 also seems to have a linear relationship, but with a different slope and intercept compared to Dataset 1.
The data points are more spread out than in Dataset 1.
The correlation coefficient is close to 1, suggesting a strong linear relationship.

Dataset 3:

This dataset does not have a linear relationship.
It consists of two distinct groups of points, with one group forming a linear pattern and the other an inverted U-shaped pattern.
The linear regression line does not accurately capture the relationship.
The correlation coefficient is close to 0, indicating a very weak linear relationship.

Dataset 4:

Dataset 4 also does not have a linear relationship.
It consists of mostly constant x values with one outlier.
The outlier strongly influences the linear regression line, leading to a high slope and a correlation coefficient close to 1.
However, the relationship is not truly linear, as removing the outlier significantly changes the pattern.

Q.3 : **What is Pearson's R?**

A.3 : The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

If r = 1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases in a linear fashion.
If r = -1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases in a linear fashion.
If r = 0, it indicates no linear relationship between the two variables. However, note that there might still be other types of relationships or associations present.

Key points about Pearson's R:

- Pearson's r quantifies the strength and direction of linear relationships between variables. It may not capture nonlinear relationships.
- Outliers in the data can strongly influence the value of Pearson's r. A single outlier can artificially inflate or deflate the correlation coefficient.
- The coefficient's range makes it easy to interpret. Values close to -1 or 1 indicate strong relationships, while values close to 0 suggest a weak or no linear relationship.
- A high correlation does not imply that one variable causes the other. Correlation measures association, but causation requires additional evidence and analysis.
- Pearson's correlation coefficient assumes that both variables follow a normal distribution.

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

A.4 : Scaling is a preprocessing technique used in data analysis and machine learning to transform the features (variables) of a dataset so that they have a consistent scale or range.

The primary purpose of scaling is to ensure that all the features contribute equally to the analysis and modeling process. Scaling is important because many machine learning algorithms are sensitive to the scale of the input features. If the features have different scales, it can lead to issues such as slow convergence, unstable model performance, or the domination of one feature over others in the learning process.

Different between normalized scaling and standardized scaling:

Range:

- Normalization scales the data to a specific range, typically between 0 and 1.
- Standardization transforms the data to have a mean (average) of 0 and a standard deviation of 1.

Purpose:

- Normalization is primarily used to scale features when the algorithm being used or the application domain requires that all features be within a consistent, bounded range (e.g., between 0 and 1). It helps maintain the relative relationships and proportions between data points but may not handle outliers well.
- Standardization is primarily used to make data suitable for algorithms that rely on the distribution of data or assume that data follows a normal (Gaussian) distribution. It centers the data around the mean and scales it by the standard deviation, making it robust to outliers.

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

A.5 : Infinite VIF values occur when there is perfect multicollinearity in a regression model. It's crucial to identify and address the root causes of multicollinearity to ensure the reliability of your regression analysis and model interpretation.

This situation arises for the following reasons:

- Perfect Collinearity: Infinite VIF occurs when one or more independent variables in your regression model can be perfectly predicted from a linear combination of other independent variables. In other words, there is a perfect linear relationship among some of your predictor variables. This perfect multicollinearity leads to an inability to estimate the coefficients accurately.

- Linear Dependency: In a regression model, if two or more variables are linearly dependent, it leads to an unstable and singular matrix when calculating the correlation matrix or inverses. This, in turn, results in the VIF being calculated as infinity.

- Model Overparameterization: Sometimes, including too many independent variables in a regression model without careful consideration can lead to perfect multicollinearity.

- Data Issues: Infinite VIF can also be an indication of data problems or errors, such as duplicate entries or incorrect data transformations.

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

A.6 : Q-Q plots are valuable tools in linear regression for assessing the normality assumption of the residuals and predictor variables. They help you visually inspect the distribution of the data and identify departures from normality, which can inform data preprocessing and model improvement decisions.

Creating a Q-Q Plot:

- Sorting Data: First, you sort the residuals (or predictor variable values) in ascending order.

- Computing Expected Quantiles: You compute the quantiles (percentiles) that correspond to the sorted residuals if they were to follow a theoretical normal distribution. These expected quantiles are based on the mean and standard deviation of the residuals.

- Plotting: You create a scatterplot with the expected quantiles on the x-axis and the actual sorted residuals on the y-axis.

Use and Importance of Q-Q Plots in Linear Regression:

- Assumption Checking: One of the key assumptions in linear regression is that the residuals are normally distributed. Deviations from this assumption can impact the validity of regression results, including confidence intervals and hypothesis tests.

- Visual Assessment: A Q-Q plot provides a visual comparison between the observed residuals and the quantiles expected from a normal distribution. If the points on the plot closely follow a straight line, it suggests that the residuals are approximately normally distributed.

- Identification of Departures: Departures from normality, such as skewness or heavy tails, are readily apparent in a Q-Q plot. If the points deviate from the straight line in a systematic manner, it indicates a departure from the normal distribution.

- Decision Making: Based on the Q-Q plot, you can make informed decisions about whether any transformations or modifications to the data or model are necessary.

- Residual Diagnostics: Q-Q plots are also used to assess the normality of residuals, which is crucial in examining the assumptions of constant variance and independence of errors. Non-normal residuals can be indicative of model misspecification.