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?

Answer:

- Summer and Fall are the most popular season for renting bikes, followed by spring and winter, which is expected given that the weather is ideal for motorcycling.

- Most of the bookings has been done during the month of may, june, july, aug, sep and oct . Number of booking for each month seems to have increased from 2018 to 2019.

- In comparison to previous year, i.e 2018, booking increased for each weather situation in 2019. Clear weather seems to have attracted more booking.

- After visualising weekday column we conclude that wednesday, thurseday, Friday and Saturday have more number of bookings as compared to the start of the week.

- After visualising workingday column we conclude that Booking seemed to be almost equal either on working day or non-working day. But, the count increased from 2018 to 2019.

- Finally, the data indicates that clear skies are optimal for renting bikes as they provide ideal temperate conditions with little humidity and cooler temperatures.

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

- **Answer**:
- It is important to use drop_first=True during dummy variable creation to avoid multicollinearity in the dataset. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other. In the case of dummy variable creation, if we include all the dummy variables in the model, then one of the dummy variables can be predicted with 100% accuracy based on the values of the other dummy variables. This can lead to issues such as overfitting and unstable estimates of regression coefficients.
- By using drop_first=True, we drop the first category of each categorical variable and create k-1 dummy variables for a variable with k categories. This removes the issue of perfect multicollinearity and also helps to simplify the model interpretation by setting a baseline category for each variable.

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

Answer: The term “temp” exhibited the highest correlation with a same coefficient of 0.63.

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

Answer: There are several assumptions that need to be validated when building a linear regression model on the training set. Here are some common techniques to validate these assumptions:

- **Linearity:** The relationship between dependent variable and the independent variables should be linear. One way to check this is to plot a scatter plot of the dependent variable against each independent variable. If the points are scattered randomly, then the relationship is linear.
- **Normality:** The dependent variable should be normally distributed. One way to check this is to plot a histogram of the dependent variable and look for a bell-shaped curve. Another way is to use a normal probability plot to check whether the data points fall on a straight line.
- **Homoscedasticity:** The variance of the errors should be constant across all values of the independent variable. One way to check this is to plot a scatter plot of the residuals (the differences between the predicted values and the actual values) against the predicted values. If the points are scattered randomly, then the variance is constant.
- **Independence:** The residuals should be independent of each other. One way to check this is to plot a scatter plot of the residuals against the independent variables. If there is no pattern in the plot, then the residuals are independent.

These assumptions can be validated using various statistical tests as well, such as the Shapiro-Wilk test for normality, the Breusch-Pagan test for homoscedasticity, and the Durbin-Watson test for independence. It is important to validate these assumptions to ensure that the linear regression model is reliable and accurate.

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

Answer: The following are the top three qualities that significantly contribute to the need for shared bikes:

- September
- Yr
- July

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are multiple independent variables.

The goal of linear regression is to find the best-fit line or hyperplane that can predict the dependent variable with the least amount of error. This is done by minimizing the sum of the squared errors between the predicted values and the actual values.

The linear regression model is represented as:

y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn

Where:

- y is the dependent variable
- x1, x2, ..., xn are the independent variables
- b0 is the intercept (the value of y when all independent variables are 0)
- b1, b2, ..., bn are the coefficients (the change in y for a unit change in the corresponding independent variable)

The coefficients (b1, b2, ..., bn) are estimated using the method of least squares, which involves finding the values of the coefficients that minimize the sum of the squared errors between the predicted values and the actual values.

Once the coefficients are estimated, the linear regression model can be used to predict the dependent variable for new values of the independent variables. The accuracy of the predictions can be evaluated using various metrics such as mean squared error, mean absolute error, and R-squared.

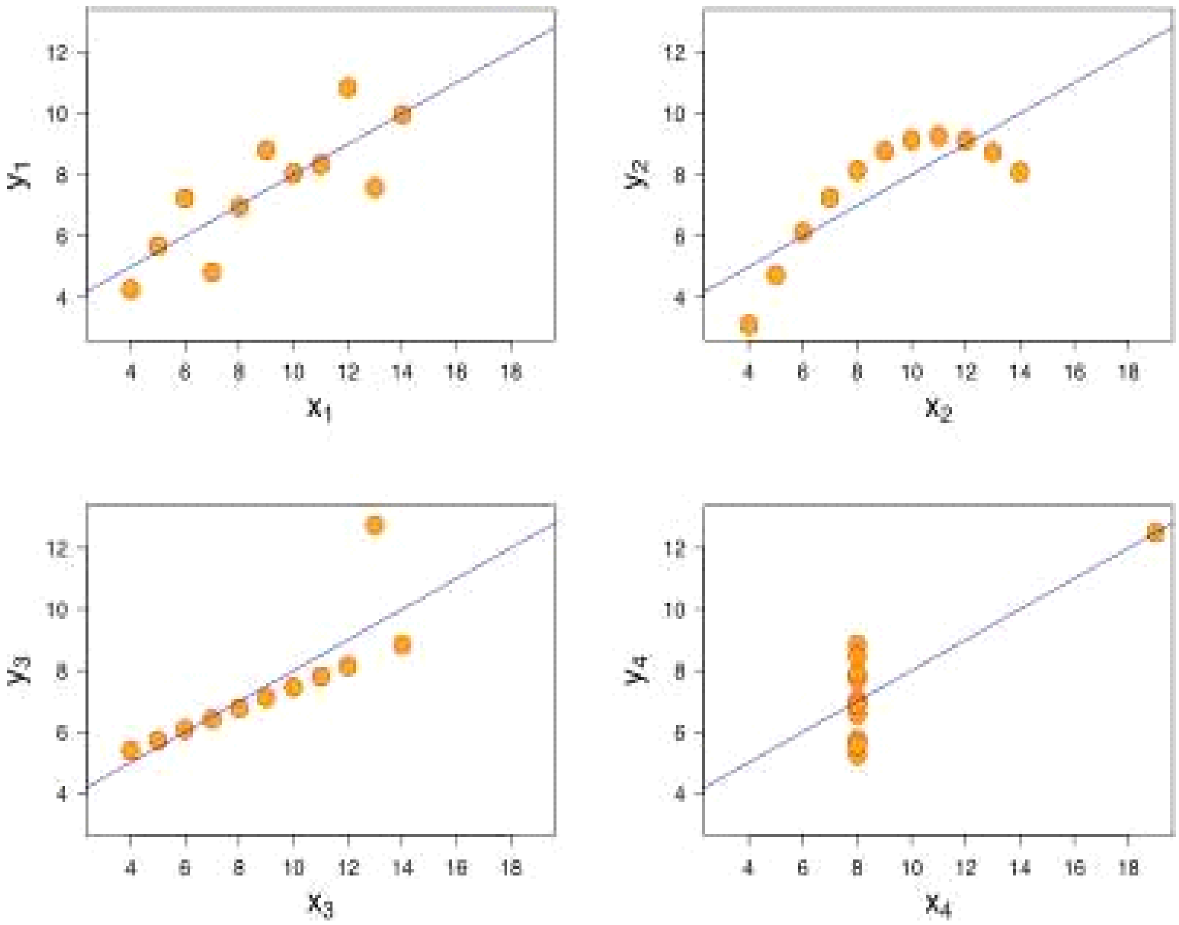
Linear regression can be used for both simple and complex datasets and is widely used in various fields such as finance, economics, and engineering for modeling and prediction.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet consists of four data sets with virtually similarsimple descriptive statistics, but when represented graphically, the distributions are very different.

The mean, sample variation of x and y, correlation coefficient, linearregression line, and R-Square value make up the simple statistics.

Anscombe's Quartet demonstrates how graphing can nevertheless reveal significant differences between numerous data sets with many comparablestatistical features. The charts are displayed below:



- The first plot (top left) seems to represent a straightforward linearrelationship.

- The correlation coefficient is useless because the second figure (top right)depicts a nonlinear relationship and is not normally distributed.
- The third plot is linear but uses a different regression line (bottom left). This istaking place as a result of the outliers in the data.

- The fourth plot (bottom right) does not demonstrate a linear relationship, butthe data were changed because of outliers.

To put it simply, it is better to visualise data and eliminate outliers beforeexamining it.

3. What is Pearson's R?

Answer: The strength of a relationship between two variables is measured byPearson's R.
It is calculated by dividing the covariance of two variables by the sum of theirstandard deviations. Its range of values is +1 to -1.

- A value of 1 denotes a complete linear positive correlation. It implies that ifone variable rises, the others will follow suit.

- Zero indicates there is no association.

- A score of -1 indicates a completely negative association. It implies that ifone variable rises, another will fall.

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

Answer: Scaling is a preprocessing step in machine learning where the features of a dataset are transformed to have a specific range or distribution. Scaling is performed to ensure that all the features in a dataset have equal importance during modeling.

The difference between normalized scaling and standardized scaling is as follows:

Normalized scaling involves scaling the features so that they have a minimum value of 0 and a maximum value of 1. This is also known as min-max scaling. Normalized scaling is useful when we need the features to be on a common scale and the distribution of the data is not important.

Standardized scaling involves scaling the features so that they have a mean of 0 and a standard deviation of 1. This is also known as Z-score normalization. Standardized scaling is useful when the features have different units of measurement or when the distribution of the data is important. Standardized scaling transforms the data into a standard normal distribution with a mean of 0 and a standard deviation of 1.

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

Answer: In some cases, the value of VIF can be infinite. This happens when the predictor variable is a linear combination of other predictor variables in the model. In other words, the predictor variable can be perfectly predicted from the other predictor variables in the model. As a result, the estimated regression coefficient for this predictor variable is not identifiable and the VIF value becomes infinite.

For example, if we have three predictor variables (X1, X2, X3) in a linear regression model, and we define X3 = X1 + X2, then X3 is a linear combination of X1 and X2. When we calculate the VIF for X3, it will be infinite because X3 can be perfectly predicted from X1 and X2, and the estimated regression coefficient for X3 is not identifiable.

In such cases, it is necessary to remove the redundant variable(s) from the model to resolve the issue of infinite VIF. This can be done using techniques such as principal component analysis (PCA) or by simply removing one of the correlated predictor variables.

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

Answer: A Q-Q plot, or quantile-quantile plot, is a graphical technique used to assess whether a set of data is normally distributed. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of a theoretical normal distribution. If the data is normally distributed, the Q-Q plot will result in a straight line.

The use importance of a Q-Q plot in linear regression are as follows:

- Checking normality: Linear regression assumes that the dependent variable follows a normal distribution. By plotting the sample data against a theoretical normal distribution, a Q-Q plot can help to check whether the data is normally distributed. If the Q-Q plot shows a straight line, it suggests that the data is normally distributed.
- Identifying outliers: A Q-Q plot can also be used to identify outliers. Outliers are observations that are significantly different from the rest

of the data and can have a disproportionate impact on the regression model. If there are outliers in the data, the Q-Q plot may show points that deviate significantly from the straight line.

In summary, a Q-Q plot is a useful tool in linear regression analysis to check the normality of data, assess the assumption of homoscedasticity, and identify outliers.