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)

Answer:- The categorical columns of the dataset consist of season, year, months, holiday, weekday, working day, and weather sit(conditions). Their effect on dependent variables is as follows:

1. Seasons like winter and summer are the most demanding for bike-sharing, whereas cloudy, humid, light, and heavy rain/snow are the worst time for bike-share.
2. As the rentals are increasing, 2019 has more bike rentals than 2018 as it might be getting more popular or people becoming aware of the environment.
3. As seasons make a huge difference in bike-sharing the months will be equally related. September, winter months as December, and summer months are good for rentals, however, January, July, and rainy months have less demand.
4. People prefer spending quality time with family and mostly at home, maybe because the demand is less for holidays.
5. Weekdays cannot be inferred much but Saturdays and Sundays are holidays, so those weekdays will not have many rentals.
6. Working day can be a demanding time for bike-sharing as it will be helpful for office goings, college students, and others.
7. Weather conditions can play an important role in the demand for bike share. Light snow, heavy snow/rain will have fewer rentals whereas clear or partly cloudy days have more rentals.

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

Answer:- When creating dummy variables, **drop_first=True** is an important parameter to consider. It is used to drop the first category of a categorical variable, which is used as a reference.

The reason for dropping the first category is to avoid the issue of multicollinearity, which can occur when using all the categories of a categorical variable. Multicollinearity is a situation where one independent variable is highly correlated with another independent variable in the same model. This can lead to difficulties in interpreting the effect of each independent variable on the outcome variable.

By using **drop_first=True**, we can ensure that the model is not influenced by the baseline category, which can improve the overall performance of the model. It can also reduce the number of dummy variables required to represent a categorical variable, leading to a clearer model and reducing the computational burden.

Let's say we have a categorical variable called "color" with three categories: "red", "green", and "blue". If we make a dummy variable without drop_first=True then the output will be:

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

Answer:- Among the numerical variables 'temp' and 'atemp' are the only variables having the highest correlation with the target variable. The pair-plot shows the temperature and count of total rentals as increasing, which means rentals increase when temperature increases.

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

Answer:- Validating the assumptions of linear regression is an important step after building a model on the training set to ensure that the model is appropriate for making predictions on new data. Here are the steps to validate the assumptions of linear regression:

1. **Check for linearity**: The first assumption of linear regression is that the relationship between the independent and dependent variables is linear. This can be checked by plotting the residuals against the predicted values. The plot should show a random pattern without any clear trends or patterns.
2. **Check for homoscedasticity**: The second assumption of linear regression is that the variance of the residuals is constant across all levels of the independent variable(s). This can be checked by plotting the residuals against the predicted values or against each independent variable. The plot should show a random scatter of points without any clear patterns.
3. **Check for normality**: The third assumption of linear regression is that the residuals are normally distributed. This can be checked by plotting a histogram or a Q-Q plot of the residuals. The plot should show a normal distribution of the residuals.
4. **Check for multicollinearity**: The fourth assumption of linear regression is that the independent variables are not highly correlated with each other. This can be checked by calculating the correlation matrix of the independent variables. If there are high correlations between two or more variables, then one of them should be removed from the model.
5. **Check for outliers**: Outliers and influential observations can have a significant impact on the regression coefficients and the overall model performance. Outliers can be identified by plotting the residuals against the predicted values, and influential observations can be identified by calculating the leverage and Cook's distance of each observation.
6. **Check for model fit**: Finally, the overall fit of the model should be evaluated by checking the goodness of fit measures such as R-squared, adjusted R-squared, and root mean squared error (RMSE).

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:- Based on the final model, the top 3 features contributing to the demand for shared bikes are Temperature, Year, and winter season. Temperature is the highest feature to contribute followed by year and winter season.
Temperature is having the highest coefficient of 1122.1742 units. With the greater temperature greater will be the demand for bike rentals. So, temperature plays an important role in the demand for bike rentals.
The year coefficient is in increasing order with a coefficient of 1122.1742 units. It can be said that the year 2019 has a greater number of bike rentals than the year 2018 maybe because of an increase in its popularity or awareness of the environment in people.
The third-highest coefficient is for the winter season which is 473.6542 units. The season also plays an important role in the demand for bike rentals. The rentals increase in the winter season can be due to easygoing in snow or traffic.

General Subjective Questions

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

Answer:- Linear regression is a statistical technique that is used to model the linear relationship between a dependent variable and one or more independent variables. It is a widely used algorithm in machine learning and statistical analysis, as it is simple and easy to interpret.
The basic idea behind linear regression is to find a line that best fits the data points, such that the difference between the actual and predicted values is minimized. This line is called the regression line or the line of best fit. The regression line is defined by the equation:
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, and b1, b2, ..., bn are the regression coefficients that represent the slope of the line for each independent variable.
The most common method to estimate the regression coefficients is the ordinary least squares (OLS) method and the VIF method. Once the regression coefficients are calculated on train data, the model is used to make predictions on test data by plugging in the values of the independent variables into the regression equation.
Linear regression can be extended to multiple linear regression when there are more than one independent variables. Linear regression has several assumptions, including linearity, homoscedasticity, normality, independence, and absence of multicollinearity, which should be checked before interpreting the results and making predictions.

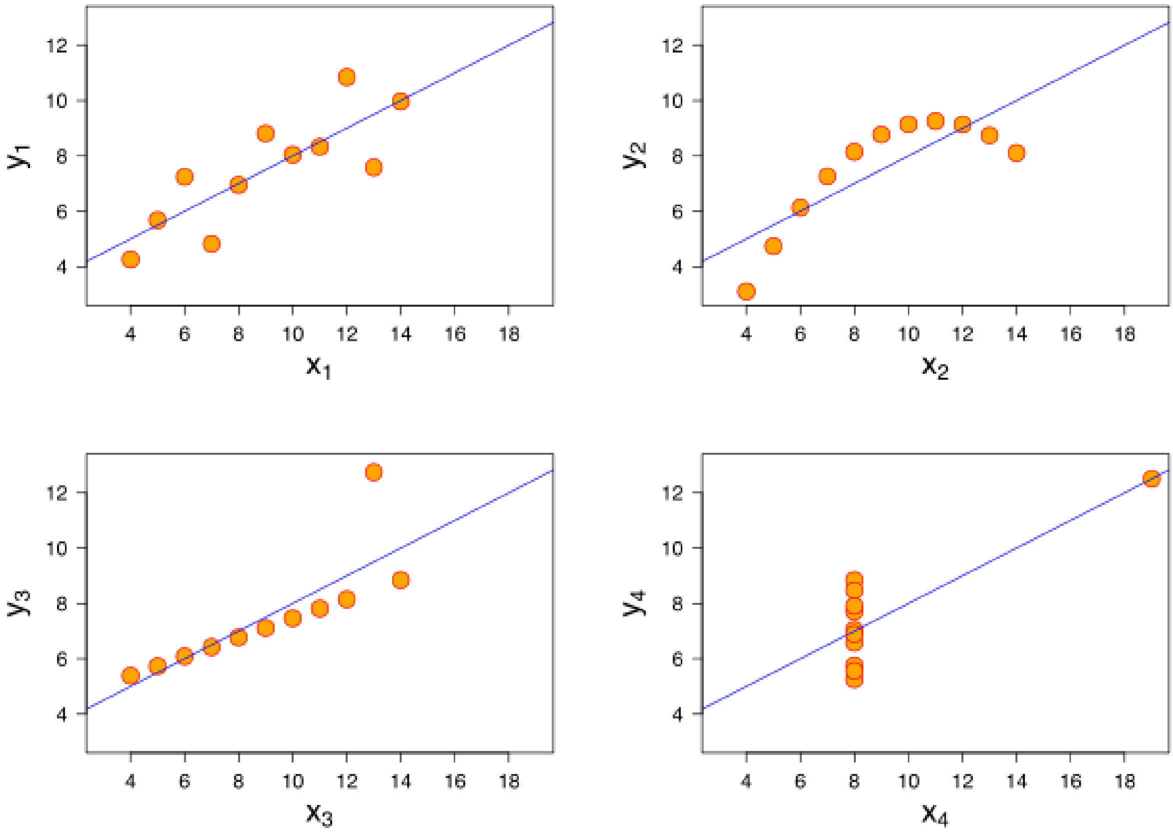
If the assumptions are not met, appropriate measures such as data transformation, removing outliers, or removing highly correlated variables can be taken to improve the model performance.

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

Answer:- Anscombe’s quartet is a set of four datasets that have identical summary statistics but exhibit vastly different patterns when plotted graphically. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphical analysis in data visualization and to caution against overreliance on summary statistics.

Each dataset in Anscombe's quartet consists of 11 (x, y) pairs, which represent the independent and dependent variables, respectively.

When plotted graphically, each dataset in Anscombe's quartet reveals a unique pattern. Here are the plots of the four datasets in Anscombe's quartet:



Four datasets of Anscombe's quartet

The first dataset shows a linear relationship between x and y, while the second dataset has a curved relationship. The third dataset shows the impact of an outlier on the slope and intercept, while the fourth dataset has no apparent relationship between x and y, except for an outlier that has a large effect on the summary statistics.

3. What is Pearson’s R? (3 marks)

Answer:- Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is denoted by the symbol 'r' and takes values between -1 and 1, where -1 represents a perfect negative correlation, 0 represents no correlation, and 1 represents a perfect positive correlation.
Pearson's R is calculated by dividing the covariance between two variables by the product of their standard deviations. The formula for Pearson's R is as follows:

$$\mathbf{r = (\Sigma(x_i - x)(y_i - y)) / (\sqrt{\Sigma(x_i - x)^2} * \sqrt{\Sigma(y_i - y)^2})}$$

where x_i and y_i are the values of the two variables, x, and y are their means, and Σ denotes the sum of all the values.
Pearson's R is a valuable tool in data science because it provides a numerical value that summarizes the relationship between two variables, which can help in making data-driven decisions. For example, in marketing, Pearson's R can be used to determine the relationship between advertising spending and sales, which can help to optimize marketing budgets.
By analyzing the correlation between variables with Pearson’s R, data scientists can identify highly correlated variables and remove redundant features, reducing the dimensionality of the data and improving the performance of machine learning models.
Pearson's R is an essential tool in data science for analyzing the association between variables, identifying patterns and trends, and improving the performance of machine learning models.

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

Answer:- In data preprocessing, scaling refers to the process of transforming the numerical values of features in a dataset to a standardized range. The purpose of scaling is to improve the accuracy and efficiency of machine learning algorithms, particularly those that are sensitive to the scale of the input features.
Scaling is performed because many machine learning algorithms use distance-based metrics to measure the similarity between observations. If the features are not on the same scale, then the algorithm may overweight the features with larger values, leading to biased results. Scaling the features to a similar range can help to mitigate this problem and improve the accuracy of the model.
There are two commonly used scaling techniques: normalized scaling and standardized scaling.

1. **Normalized scaling**, also known as min-max scaling, transforms the values of features to a range between 0 and 1. The formula for normalized scaling is:

   $$\mathbf{X_{norm} = (X - X_{min}) / (X_{max} - X_{min})}$$

   where X is the original value of the feature, X_min is the minimum value of the feature, and X_max is the maximum value of the feature. This scaling method preserves the relative relationships between the values of the feature.

2. **Standardized scaling**, also known as z-score scaling, transforms the values of features to have zero mean and unit variance. The formula for standardized scaling is:

   X_std = (X - X_mean) / X_std

   where X is the original value of the feature, X_mean is the mean value of the feature, and X_std is the standard deviation of the feature. This scaling method centers the feature around zero and scales it to a range that is relative to its variability.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the relative relationships between the values of the feature, while standardized scaling centers the feature around zero and scales it to a range relative to its variability. Both scaling methods are useful for improving the performance of machine learning algorithms, depending on the specific application and the nature of the data.

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

(3 marks)

Answer:- VIF (Variance Inflation Factor) is a measure of multicollinearity between independent variables in a regression model. It quantifies how much the variance of the estimated regression coefficient for an independent variable is increased due to correlation with other independent variables in the model.

A VIF of 1 means that there is no correlation among independent variables, whereas a VIF value greater than 1 indicates that the independent variable is correlated with other independent variables. A VIF value of 5 or above is generally considered to indicate high multicollinearity between the independent variables.

Sometimes, the value of VIF can be infinite. This happens when there is perfect multicollinearity between independent variables. Perfect multicollinearity occurs when one or more of the independent variables can be expressed as a linear combination of other independent variables in the model.

The formula of VIF is:

VIF = $\frac{1}{1-\mathrm{R}\hat{}2}$

where R^2 is the coefficient of determination for the regression model that uses the predictor variable as the dependent variable and all other predictor variables as the independent variables. When the value of R gets equal to 1 then the value of VIF gets infinite.

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

(3 marks)

Answer:- In linear regression, Q-Q plots are commonly used to assess the normality of the residuals, which are the differences between the observed values and the predicted values of the dependent variable. If the residuals are normally distributed, then the Q-Q plot will show a straight line, indicating that the distribution of the residuals is consistent with a normal distribution. If the residuals are not normally distributed, then the Q-Q plot will show a curved line or deviations from the straight line, indicating that the distribution of the residuals deviates from a normal distribution.

The use and importance of a Q-Q plot in linear regression can be summarized as follows:

1. **Assumption of normality:** One of the key assumptions of linear regression is that the residuals are normally distributed. Q-Q plots provide a visual method for assessing the normality assumption and can help to identify departures from normality, such as skewness, or outliers.
2. **Model validation**: Q-Q plots are an important tool for validating the linear regression model. If the Q-Q plot shows a straight line, it suggests that the model is a good fit for the data, and the residuals are normally distributed. If the Q-Q plot shows deviations from the straight line, it indicates that the model may not be a good fit for the data, and further investigation may be necessary.
3. **Outlier detection**: Q-Q plots can also be used to identify outliers in the data. Outliers are observations that deviate significantly from the normal distribution and can have a strong influence on the regression estimates. Q-Q plots can help to identify outliers by showing points that deviate from the straight line.