Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Bike demand in the fall is the highest

Bike demand takes a dip in spring

Bike demand in year 2019 is higher as compared to 2018

Bike demand is high in the months from May to October

Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or snow

Bike demand doesn't change whether day is working or not.

Q2. Why is it important to use drop _ first=True during dummy variable creation?

It is important in order to achieve k-1 dummy variables as it can be used to delete extra columns while creating dummy variables.

It also reduce multicollinearity among the categorical variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

'Atemp' and temp both have same correlation with target variable of 0.63 which is the Linearity of relationship between response and predictor variables.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set

Less multi-collinearity among the features.

Normality of the error term distribution.

Constant variance of the errors.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes

Temp

Winter

Spring

Q6. Explain the linear regression algorithm in detail?

It is a form of regression, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables. The Equation of linear Regression is y= m1x1 + m2x2 + m3x3 + ……. + m(n) x(n) + c . Where y is target variable and x1, x2 , x3 …… xn are predictor variables . And we have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error. We need to get the best fit line

WHICH IS THE LINE THAT HAS THE MINIMUM ERROR. IN LINEAR REGRESSION, WHEN THE ERROR IS CALCULATED USING THE SUM OF SQUARED ERROR, THIS TYPE OF REGRESSION IS KNOWN AS OLS, I.E., ORDINARY LEAST SQUARED ERROR REGRESSION. ERROR FUNCTION IS EXPLAINED BY 'E = - Y', AND ERROR DEPENDS ON THE VALUES OF 'M' AND 'C'. OUR AIM IS TO BUILD AN ALGORITHM WHICH CAN MINIMIZE THE ERROR. AND IN ORDER TO DO SO WE USE COST FUNCTION OF LINEAR REGRESSION, WHICH IS: J (MI, C)= (1/2N)Σ(YI – YP)^2 WHERE YI AND YP ARE EXPECTED VALUES AND PREDICTED VALUES. OUR MAIN AIM IS TO MINIMIZE J BY CHANGING M AND C AND IT CAN BE DONE USING GRADIENT DESCENT ALGORITHM. COST FUNCTION MEASURES THE PERFORMANCE OF A MACHINE LEARNING MODEL FOR GIVEN DATA.

Q7. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, /each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Q8. What is Pearson's R?

Pearson's Correlation Coefficient, often denoted as *r*, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

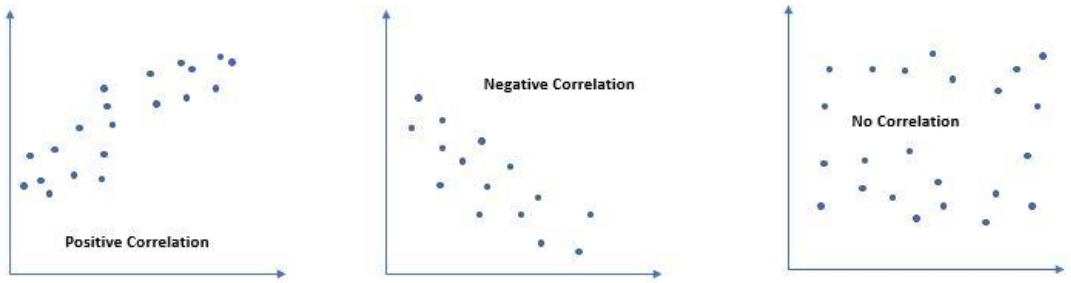
- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

The formula is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

Value of ‘r’ ranges from ‘-1’ to ‘+1’. Value ‘0’ specifies that there is no relation between the two variables. A value greater than ‘0’ indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable. Value less than ‘0’ indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.



Pearson correlation draws a line of best fit through two variables, indicating the distance of data points from this line. A ‘r’ value near +1 or -1 implies all data points are close to the line. An ‘r’ value close to ‘0’ suggests data points are scattered around the line.

Q9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

What?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

Sklearn . preprocessing. scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Q10.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is perfect correlation then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 = 1, Which lead to 1/(1-r2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear relationship of other variables (which show an infinite VIF as well).

Q11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

The power of Q-Q plots lies in their ability to summarize any distribution visually. QQ plots is very useful to determine

If two populations are of the same distribution

If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.

Skewness of distribution