Assignment-based Subjective Questions

Question1:- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:-

For analysis of categorical variables , I have used barplots. Few points that we can infer from barplots are mention below

1. highest demand of bikes in fall season. Followed by summer and least demand in the spring season
2. Demand of bikes in 2019 is approximately twice the year 2018
3. maximum and approximately same demand in august, June, July, September and followed by October, May, November, April and least demand in December, February and January.
4. as expected, maximum demand on good weather day, least demand on bad weather day
5. on holiday, there is high decrease in demand

Question 2 :- Why is it important to use drop_first=True during dummy variable creation?

Answer:-

drop first= True is useful since it reduces the unnecessary column produced during dummy variable creation. As a result, it reduces the correlations formed between dummy variables.

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

Answer:-

‘temp’ and ‘atemp’ variable has the the highest correlation with target variable ‘count’

Question 4:- How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:-

- Multicollinearity check:- vif value of all variable less than 5 and by ploting heatmap of correlation of variables
- Normality of error terms :- The residuals should be normally distributed. This can be checked by creating a histogram
- Homoscedasticity: The variance of the residuals should be constant across the levels of the independent variables. This can be checked by plotting the residuals against the predicted values. If the residuals form a funnel shape, this suggests that the variance is not constant and the homoscedasticity assumption is violated.
- Linearity: The relationship between the independent and dependent variables should be linear. We can check this by plotting the residuals against the predicted values. If the residuals are randomly scattered around the horizontal axis, this suggests that the linearity assumption is met.

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

Answer:-

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:-

- Temp
- Year_2019
- Season_winter

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer:-

Linear regression is a method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting straight line through the data points. This line can be used to make predictions about the dependent variable based on the independent variables.

There are two types of linear regression:

- Simple linear regression
- Multiple linear regression

Simple linear regression:-

The equation for a simple linear regression with one independent variable is: $y = b_0 + b_1 * x$

where y is the dependent variable, x is the independent variable, b0 is the y-intercept, and b1 is the slope of the line. The y-intercept is the value of y when x is zero, and the slope is the change in y for each unit change in x.

To find the best-fitting line, we need to determine the values of b0 and b1 that minimize the difference between the predicted values of y and the actual values of y. This is done by minimizing the sum of the squared differences between the predicted values of y and the actual values of y. The sum of squared differences is also known as the residual sum of squares (RSS). The goal is to find the values of b0 and b1 that minimize the RSS.

The process of finding the values of b0 and b1 that minimize the RSS is called ordinary least squares (OLS) estimation. There are a variety of numerical optimization algorithms that can be used to find the OLS estimates. Gradient descent is a popular algorithm for solving this problem.

Multiple Linear Regression :

When we have multiple independent variable, the equation takes the form: Y = b0 + b1*x1 + b2*x2 + ... + bn*xn

where n is the number of independent variables, and x1, x2, ..., xn are the independent variables. The coefficients b1, b2, ..., bn represent the change in y for each unit change in the corresponding independent variable, while holding all other independent variables constant. Once the coefficients are determined, the model can be used to make predictions about y given new values of x1, x2, ..., xn. However, it's important to keep in mind that Linear Regression assumes

1. Linear relation between dependent and independent variables
2. Error terms are normally distributed.
3. Error terms are independent of each other
4. Homoscedasticity:- error terms have same or constant variance

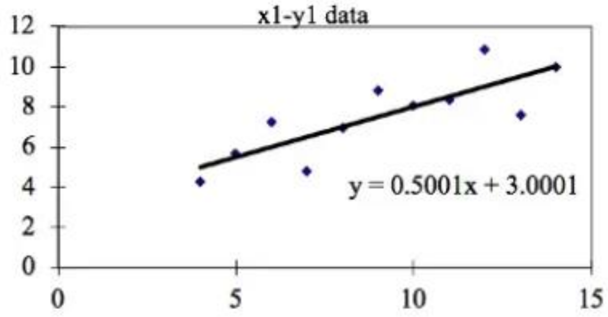
Question 2 :- Explain the Anscombe's quartet in detail.

Answer:-

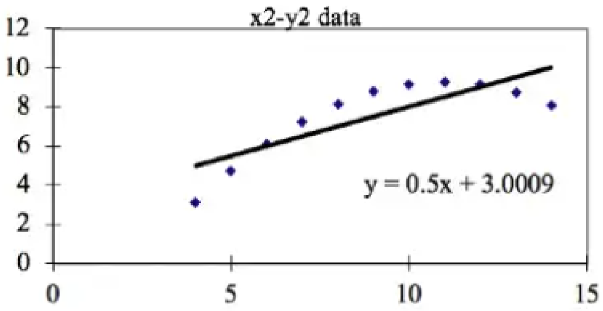
Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe to find out the importance of visualizing data before analyzing it. Each of the datasets contains 11 (x, y) points, and they all have the same summary statistics (mean, variance, correlation, and linear regression line), but they look very different when visualized.

Here are the four datasets:

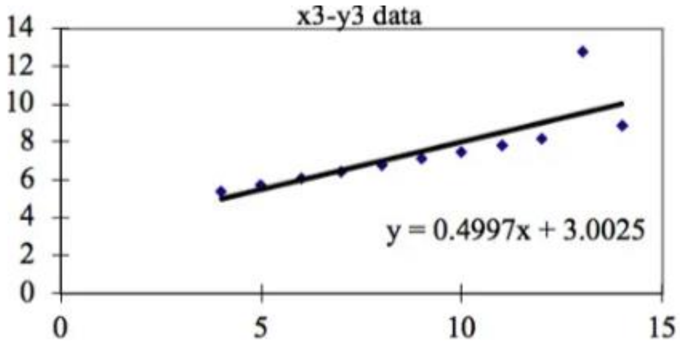
Dataset 1: when we plot the points of this data set a straight line is formed, and the linear regression line is a good fit for the data.



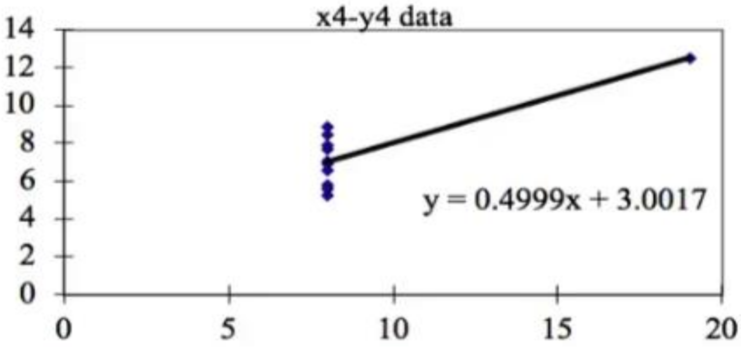
Dataset 2: when we plot the points of this data set line bends, and the linear regression line is not a good fit for the data.



Dataset 3: This dataset contains a set of discrete points and these points do not form a line when plotted. The linear regression line is not a good fit for these data points and correlation is low .



Dataset 4: This dataset contains a set of discrete points and when we plot these points, an almost perfect line is formed. The linear regression line is a good fit for these data points but has an outlier point which creates high correlation.



Anscombe's quartet emphasises the significance of visually representing data before evaluating it. The dataset's summary statistics may be extremely similar, while the particular data points may be very different. We can detect patterns and relationships in the data by looking at a graph rather than just the summary statistics. But all four datasets have the same correlation coefficient, displaying the data shows that datasets 1, 2, and 4 have a linear relationship while dataset 3 does not.

Question 3. What is Pearson's R?

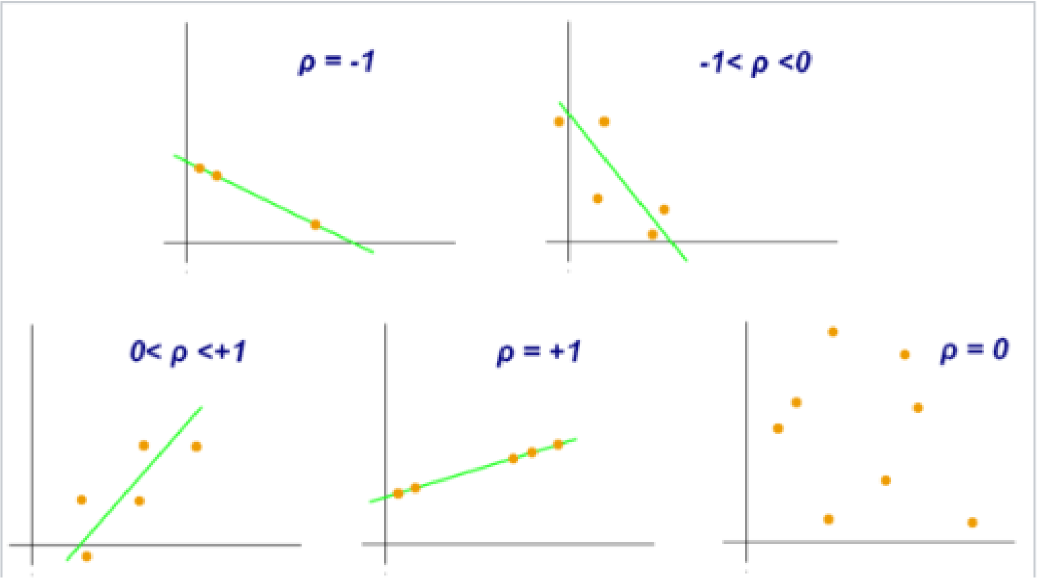
Answer:-

Pearson's R, also known as Pearson's correlation coefficient. it is a measure of the linear correlation between two variables. It ranges from -1 to 1, where -1 indicates that there is a perfect negative linear relationship between variables, 0 indicates that three is no linear relationship between variables, and 1 indicates that there is a perfect positive linear relationship between variables. The correlation coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations.

The formula for Pearson's R is as follows:

ρ = cov(x,y) / (std(x) * std(y))

where x and y are the two variables being measured, cov(x,y) is the covariance of x and y, std(x) is the standard deviation of x, and std(y) is the standard deviation of y.



Question 4:- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:-

It is the step of data preparation in linear regression which is applied on independent numerical variables to normalize the data points with in a particular range . it also helps in speeding up the mathematical calculation in linear regression algorithm. Handle variations in units of measurement: When dealing with data that comes from different sources or is measured in different units, scaling can be used to ensure that all variables are measured on the same scale.

Difference between Normalization and standardization:-

Normalization:

1. Normalization is a technique that scales a variable to have a values between 0 and 1. This is done by subtracting the minimum value of variable from its value and then dividing the result by the difference between the variable's maximum and minimum values.
2. Formula for normalizing variable x = (x-min(x)) / (max(x)-min(x))
3. This method is affected by the outliers
4. Scikit learn library provides the transformer called MinMaxScaler()

Standardization:

1. Standardization is a technique that scales a variable to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from its value and then dividing the result by the standard deviation.
2. Formula for standardizing variable x = (x-mean(x)) / std(x)
3. This method is less affected by outliers
4. Scikit learn library provides the transformer called standardScaler()

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

Answer:-

The value of VIF can become infinite. When a variable is a perfect linear combination of the other predictor variables, it means that this variable can be perfectly predicted by the other predictor variables and can be written as a linear equation of the other variables. It means this is the case of perfect correlation and we know for perfect correlation, r2 =1 hence VIF=infinite. To avoid such situations we have to drop such column

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

Answer:-

A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used to determine if a sample of data fits into a specific probability distribution. A Q-Q plot compares the quantiles of sample data to the quantiles of a theoretical distribution (such as a normal distribution) that is being tested for. The points in the Q-Q plot will fall nearly on a straight line if the sample data matches the theoretical distribution. If the sample data does not match to the theoretical distribution, the points will vary from a straight line in a given way, indicating which distribution the sample data is more likely to originate from.

A Q-Q plot in linear regression can be used to test the assumption of normality of the model's errors (also known as residuals). The normality assumption is critical in linear regression since it is used to sample distributions and estimate intervals. If the errors are not normally distributed, the estimators' properties may not hold, leading to inaccurate conclusions. A Q-Q plot, which plots the model's residuals against the normal distribution, can help us illustrate this assumption. The points in the Q-Q plot will fall nearly on a straight line if the residuals are fairly normally distributed. When the residuals stray from normalcy, the dots diverge from a straight line, indicating the deviation from normality.