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?

Ans. While performing EDA, visualized the relationship between the categorical variables and the target variable. In this assignment, consider the **effect of the categorical variable 'weathersit' on the target variable 'cnt'**. It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being 50,000 approximately. Similarly, certain inferences could be made 'season' and 'yr' as well.

When we build model, categorical features such as yr,season etc we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

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

Ans. **drop_first**=**True** is **important to use**, because it helps in reducing the extra column created **during dummy variable creation**. Hence it reduces the correlations created among **dummy variables**.

During dummy value creation (dummy encoding) it is advisable to use drop_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding.

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

Ans. **The numerical variable 'registered' has the highest correlation with the target variable 'cnt'**, if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable 'atemp' has the highest correlation with the target variable 'cnt'.

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

Ans. After building the model on the training set, we carried out the following analysis: -

1. A test for normal distribution of error terms(residuals) by visualizing a distribution plot of the error terms.

2. Eliminations and inclusion of independent variables into each model based on VIF and p-values to avoid multicollinearity.

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

The top 3 features that significantly explain the demand of the shared bikes are: -

1. 'atemp'- temperature in Celsius

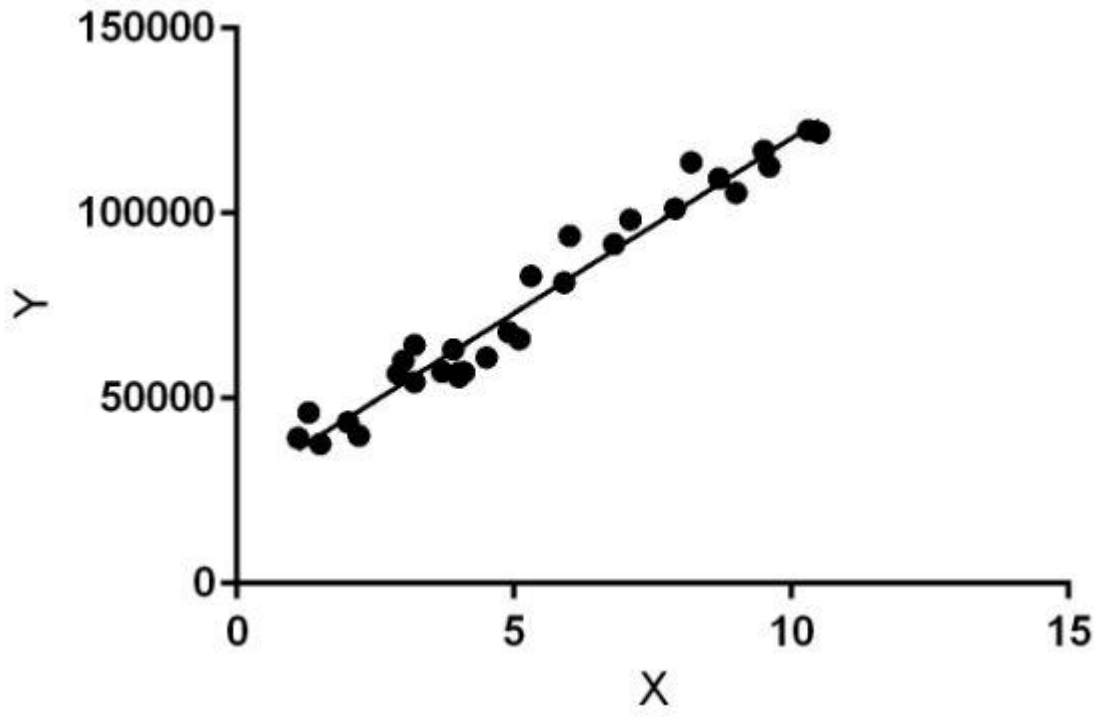
2. 'yr'- year (0: 2018, 1: 2019)

3. 'winter'- A subcategory of 'season' (4: winter)

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Ans. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

The linear regression model can be represented by the following equation:

$$y = \theta_1 + \theta_2.x$$

where,

Y is the predicted value

x is input training data (univariate – one input variable(parameter))

θ_1 is intercept
θ_2 is coefficient of x

The goal of regression analysis is to create a trend line based on the data you have gathered. This then allows you to determine whether other factors apart from the amount of calories consumed affect your weight, such as the number of hours you sleep, work pressure, level of stress, type of exercises you do etc. Before taking into account, we need to look at these factors and attributes and determine whether there is a correlation between them. Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. If the test is done over a long time duration, extensive data can be collected and the result can be evaluated more accurately.

2. Explain the Anscombe's quartet in detail.

Ans. It is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset contains of eleven

(x, y) pairs as follows: -

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics for each dataset are identical

1. The average value of x is 9.

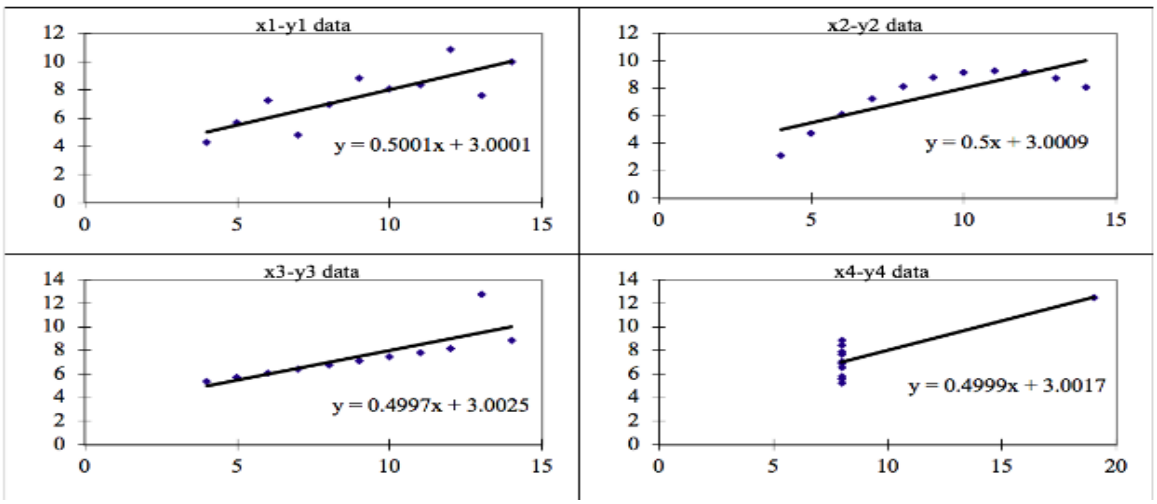
2. The average value of y is 7.5.

3. The variance for x is 11 and y is 4.12

4. The correlation between x and y is 0.816

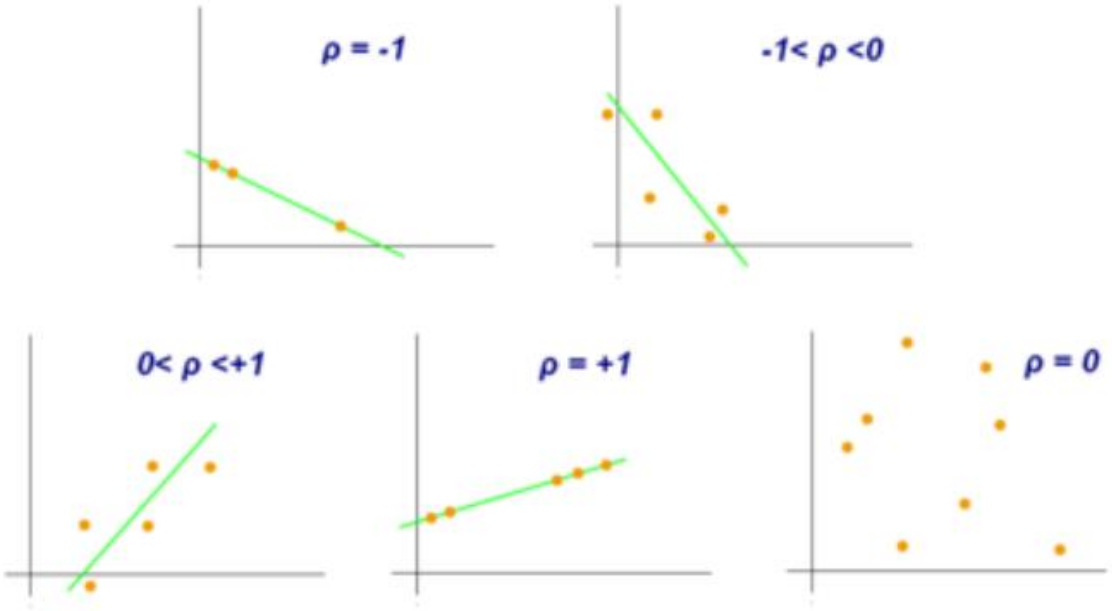
5. The line of best for is y = 0.5x + 3.

Plot gives different graphs



3. What is Pearson's R?

Ans. Pearson's R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)



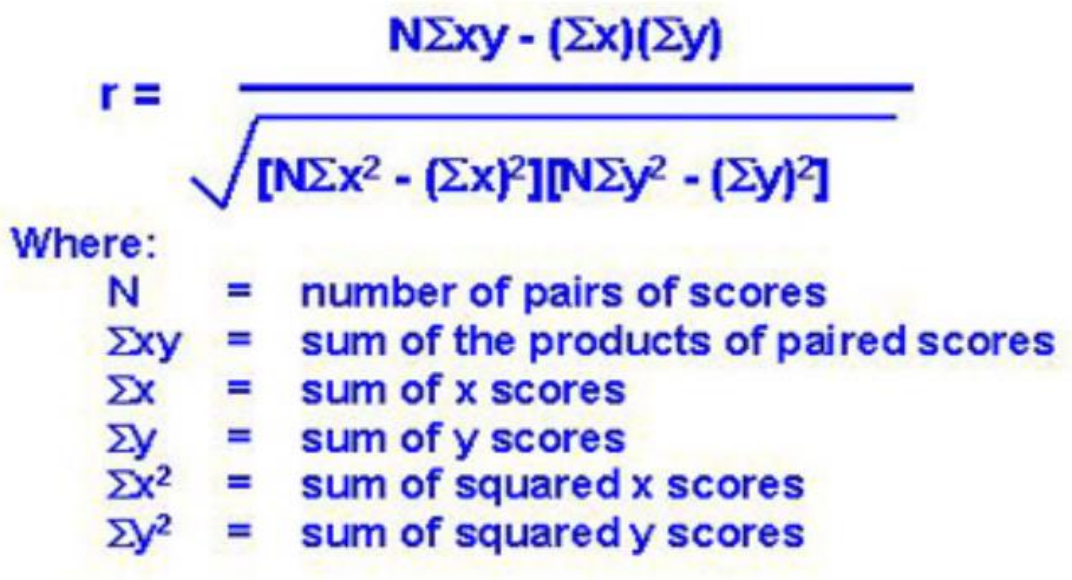
r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

Below, is the formula to calculate Pearson's R for a given dataset.



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

Ans. Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.It is extremely important to rescale the variables so that they have a comparable scale.

If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.

scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1.

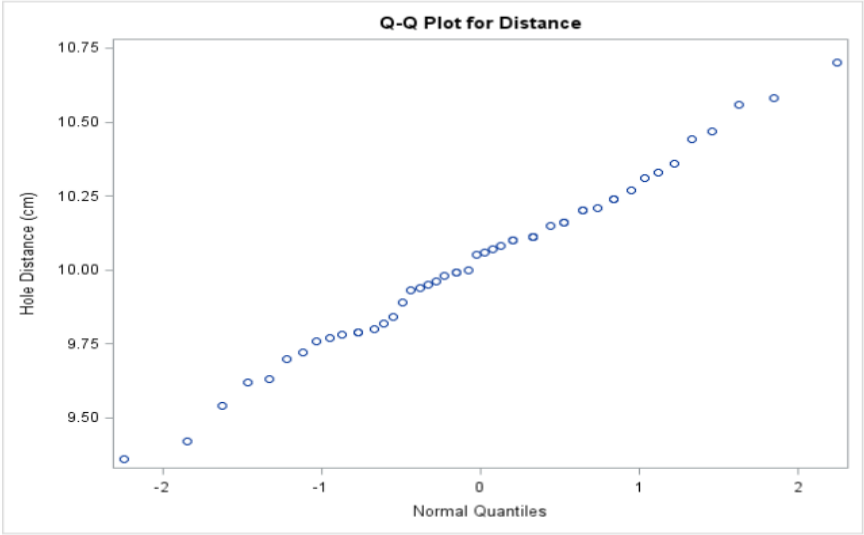
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. It directly indicates that the particulate variable has severe collinearity. Also, the corresponding variable can be expressed as a linear combination of other variables. In other words, squared multiple correlation of any predictor variable with the other predictors approaches unity.

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

Ans. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight i.e.


Q-Q Plot for Distance
Hole Distance (cm)
10.75
10.50
10.25
10.00
9.75
9.50
-2
-1
0
1
2
Normal Quantiles