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:

- Outliers are not causing alarming distortion hence no need to remove them.
- In the year 2019 number of passengers have increased across all categories which indicates the business will do better in years ahead.
- Fall is most popular season then Summer.
- Maintenance should be carried out and all bikes should be in perfect condition during Apr-Oct to maximise revenue.
- On Holidays usage is below mean indicating people spend time with families. However there is not much difference between weekends and weekdays.
- Clear weather is most preferred.
- Cnt is correlated to atemp as per the pair plot and heatmap.

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

Answer:

The function get_dummies creates as many new columns as there are levels in the category. Thus for k levels we get k new columns. However we can depict the category by using k -1 columns hence we drop the first new column.

If variable has 3 types of data A, B, and C. get_dummies will create three columns A, B and C . However we do not need column A because if B and C is 0 indicates data is A.

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

Answer:

Temp/atemp has highest correlation.

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

Answer:

- The difference between the y_train and y_train_pred gives a normal distribution.
- R^2 value is 0.828 and Prob(F-Statistic) is close to 0
- As per pair plot there exists linear relation between atemp and cnt

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:

As per my model top 3 features are

- Atemp

- Windspeed
- Summer season

General Subjective Questions

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

Answer:

It is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line.

The standard equation of the regression line is given by the following expression: $Y = β_o + β_1X$
Y – Dependent variable we are trying to predict
$β_o$ - Intercept
$β_1$ - slope
X – Independent variable.

Linear Regression makes below assumptions on the data set –

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other (Multicollinearity)
4. Error terms have constant variance (homoscedasticity)

The strength of the linear regression model can be assessed using 2 metrics:
1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

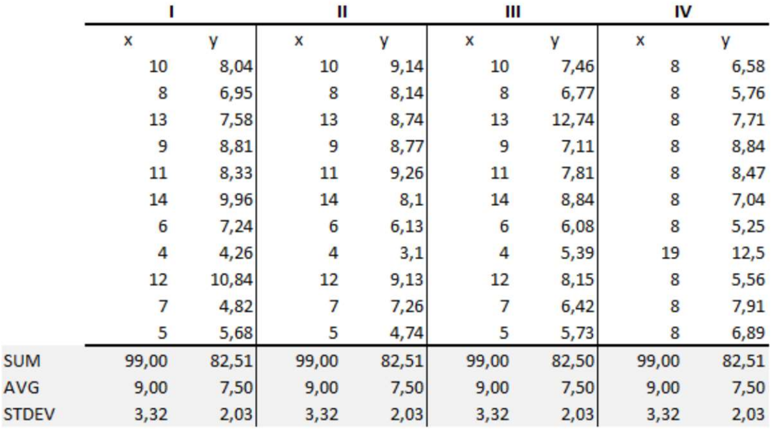
Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

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

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

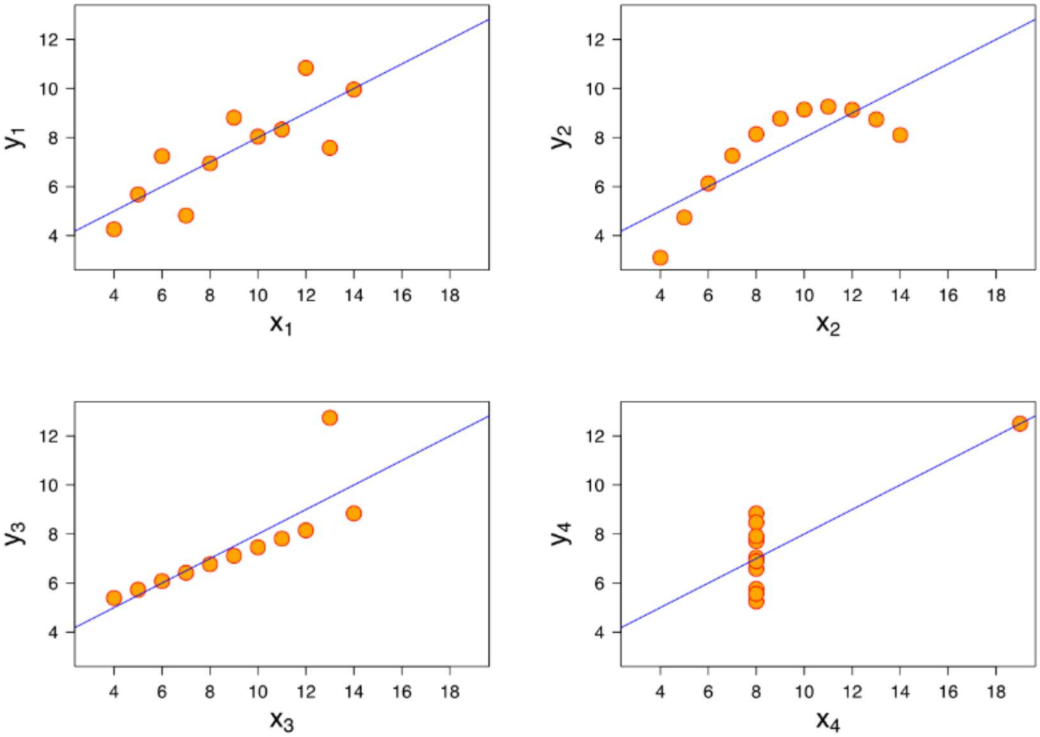


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.

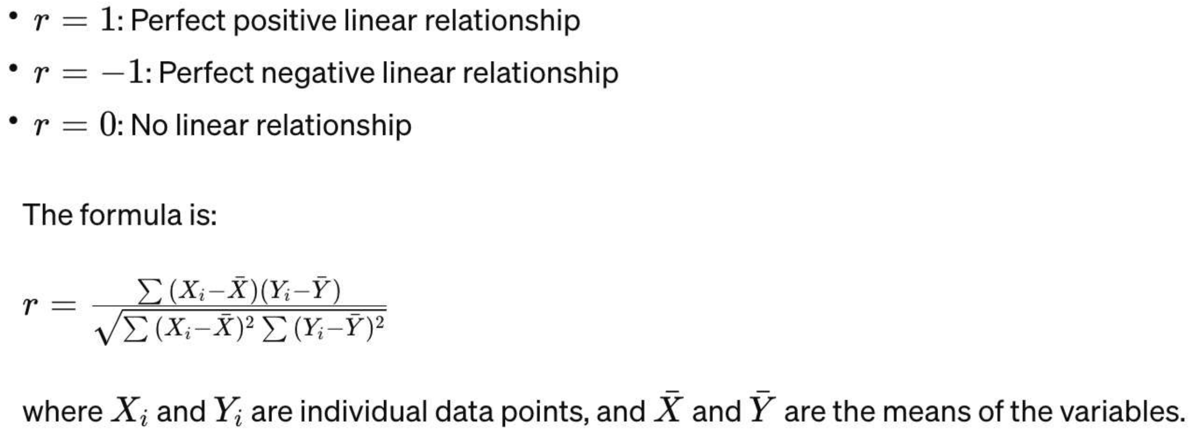
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

4. What is Pearson's R? (3 marks)
Answer:

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:



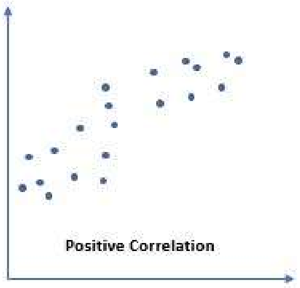
- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

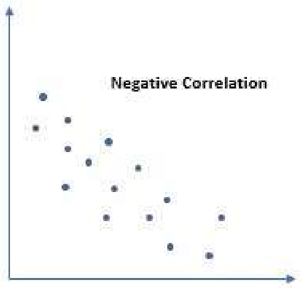
The formula is:

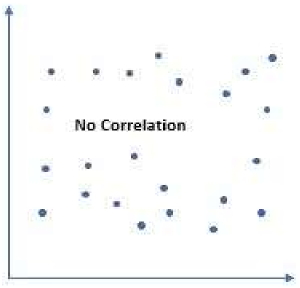
$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

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.

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

Answer:

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods You can scale the features using two very popular method:

- **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one.
- **MinMax Scaling (Normalizing):** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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-square, etc.

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

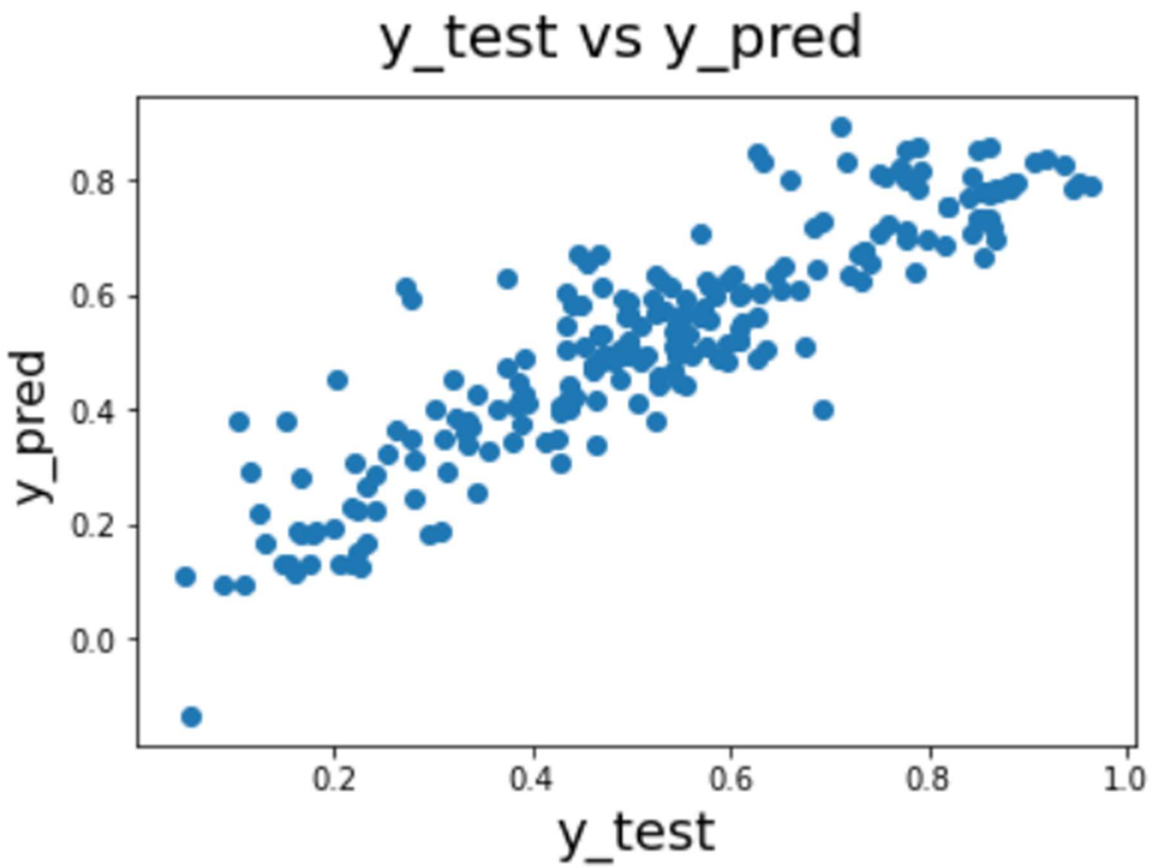
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

7. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.


y_test vs y_pred
y_pred
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
y_test