Subjective Questions – Linear Regression

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)

Ans: The below categorical variables seems to have the below inference –
1. Summer and Fall season has the highest bike booking
2. Apr to Oct has a steady rise of bike booking with a dip in July
3. Clear weather seems to have attracted the maximum customers
4. Non-holiday seems to have more bike booking than holiday probably people either relax or do other things on holidays
5. Weekday seems quite consistent with bike booking
6. Booking seems to be same between weekday vs non-weekday

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

Ans: drop_fist=True allows you to remove the reference variable used for encoding dummy variables. When you create dummy variables, n-1 levels is ok as one of the combination will be uniquely representing the redundant column.
As an example – if we have 3 types of values for a categorical variable and we want to create dummy variables, then n-1 (3-1=2) levels is good enough.

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

Ans: From the correlation matrix, in fact both temp and atemp seems have the highest correlation with 'cnt'

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

Ans: The following assumptions where validated -
1. Multicollinearity – There should be no multicollinearity between the variables. The RSE methodology in conjunction with recalculation of VIF proves this
2. Linear Relationship – It's clearly visible that the predictors and target variable have a linear relationship
3. Error Distribution – Error terms are normally distributed with mean at 'zero'
4. Homoscedasticity – No visible pattern observed in the residuals
5. Residuals – Residuals does not show auto correlation

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

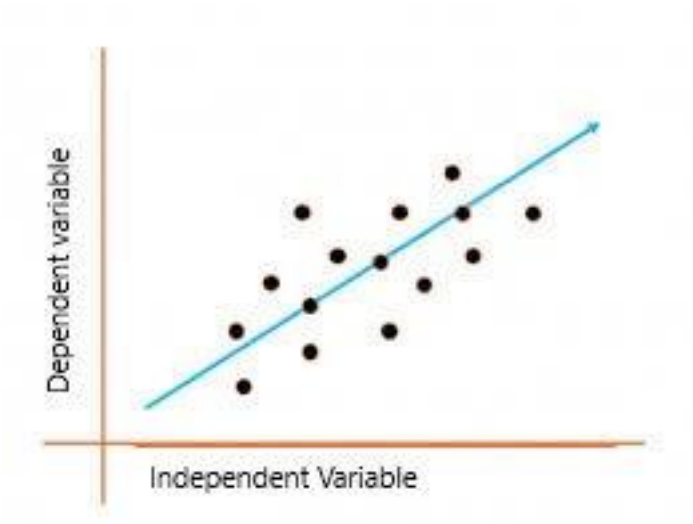
Ans: The following 3 variables have the maximum impact on bike booking:
1. Temperature
2. Season
3. Weather

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression: Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values. Linear Regression is classified as "Supervised Learning" where one has the past data and labels associated with the data.

Simple Linear Regression - In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.



The graph above presents the linear relationship between the output(y) and predictor(X) variables. The blue line is referred to as the *best-fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.
*T*o calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,
$\mathbf{Y_i = \beta_0 + \beta_1 X_i}$

where Y_i = Dependent variable, β_0 = constant/Intercept, β_1 = Slope/Intercept, X_i = Independent variable.

Multiple linear regression is a technique to understand the relationship between a *single* dependent variable and *multiple* independent variables.
The formulation for multiple linear regression is also similar to simple linear regression with
the small change that instead of having one beta variable, you will now have betas for all the variables used. The formula is given as:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p + \varepsilon$$

The most used metrics are,

1. Coefficient of Determination or R-Squared (R2)
2. Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

The assumptions of Linear Regression are -

1. Linearity relationship between dependent and independent variables
2. Independence of residuals
3. Normal distribution of residuals
4. Homoscedasticity
5. Multi-collinearity

Accessing the model –

1. p-value < 0.05
2. VIF of predictor variables <5
3. F statistic should be high
4. R2 value should be high
5. Adjusted R2
6. r2 score
7. RSME
8. MAE

2. Explain the Anscombe's quartet in detail.

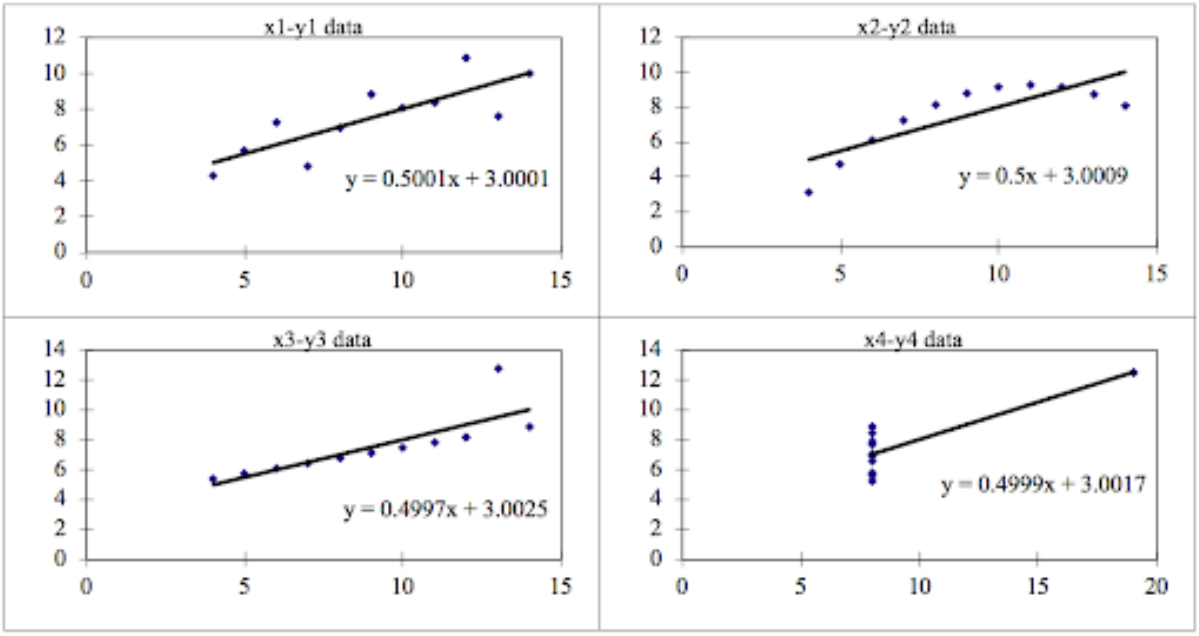
Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY,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

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

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



We can describe the four data sets as:

- **Data Set 1:** fits the linear regression model pretty well.

- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

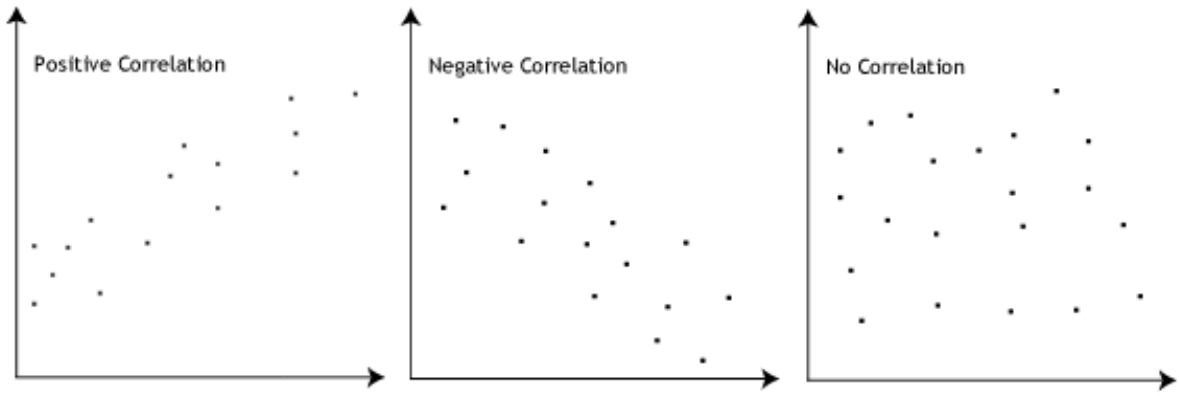
Ans: It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Ans: Scaling 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 is scaling needed?

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.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

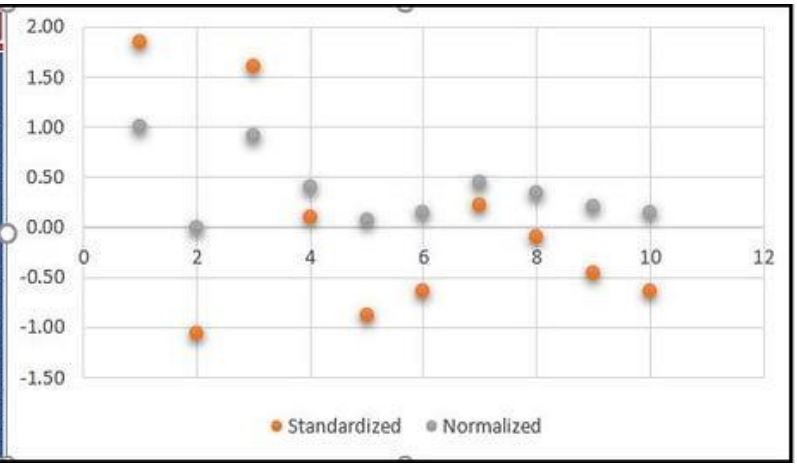
Standardization Scaling:

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**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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.

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

Ans: 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.

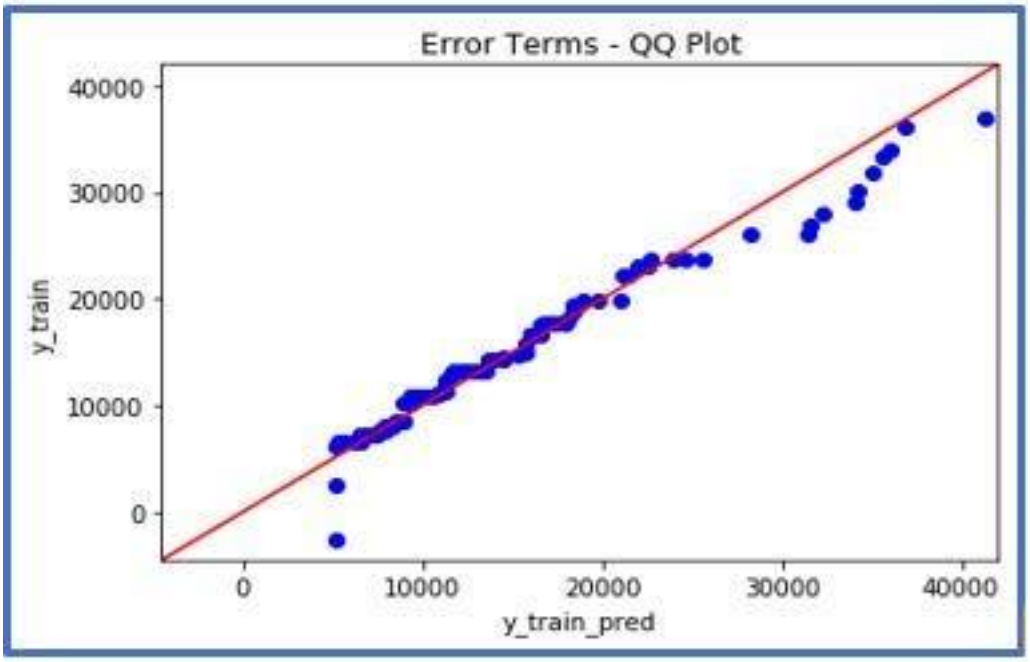
QQ plots is very useful to determine

- 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

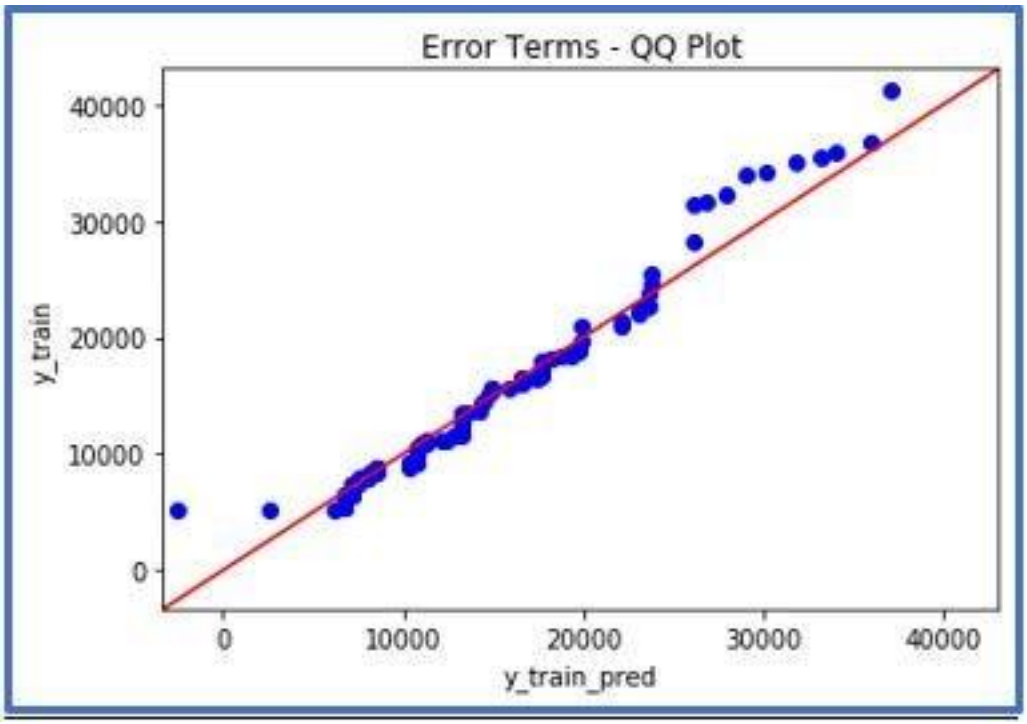
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis