General Subjective Questions

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

   Ans:-

   Linear Regression is a Machine Learning algorithm.

   Regression line is defined by the equation of a straight line which is y = mx+c where m is the angle of slope and c is the intersection.

   There are 2 types of regression:-

   Simple Linear Regression:-

   - This is the model with only 1 independent variable.
   - A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.
   - The independent variable is also known as the predictor variable and the dependent variable is also called as output variable.

   Multiple Linear Regression:-

   - Model with more than 1 independent variable.
   - It represents the relationship between two or more independent input variables and a response variable.
   - Multiple linear regression is needed when one variable might not be sufficient to create a good model and make accurate predictions.

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

   Ans:-

   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some uniqueness in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

   It was constructed in 1973 by statistician Francis Anscombe to illustrates
   the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

   There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

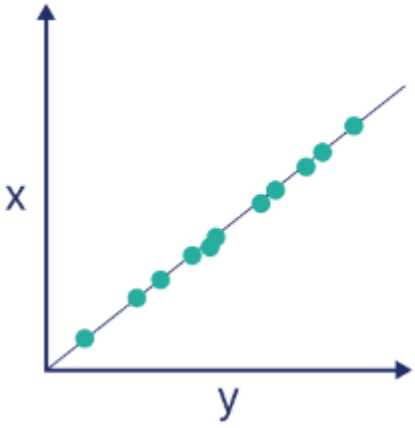
   Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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

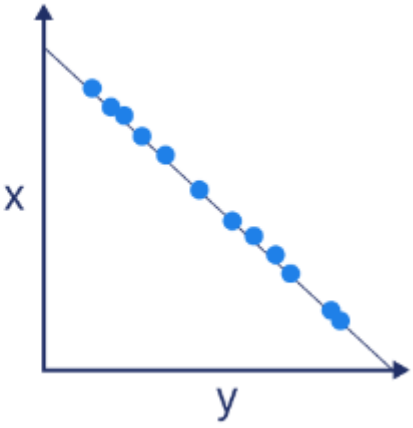
3. What is Pearson's R? (3 marks)

Ans:-

- The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
- If the Pearson's R lies between 0 and 1 then it is called as the positive correlation. If it is 0 then there is no correlation and if it lies between 0 and -1 then there is negative correlation.
- The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.
- The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.
- When *r* is 1 or –1, all the points fall exactly on the line of best fit:

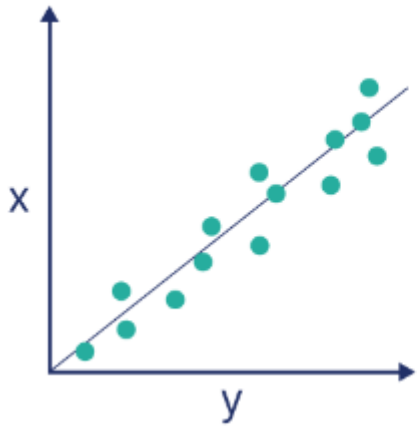


Positive Correlation



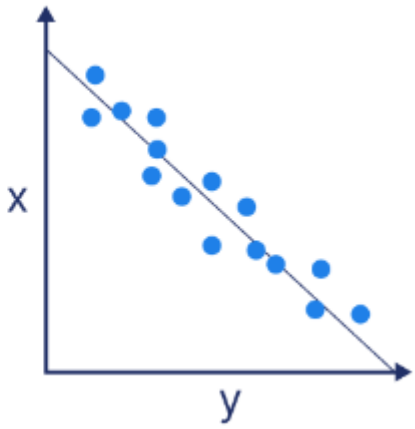
Negative Correlation

- When *r* is greater than .5 or less than –.5, the points are close to the line of best fit:



Strong Positive Correlation

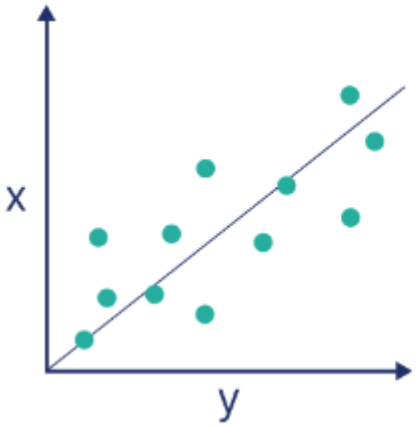
r > .5



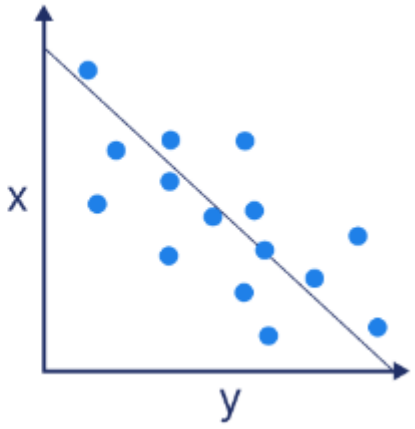
Strong Negative Correlation

r < .5

- When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:

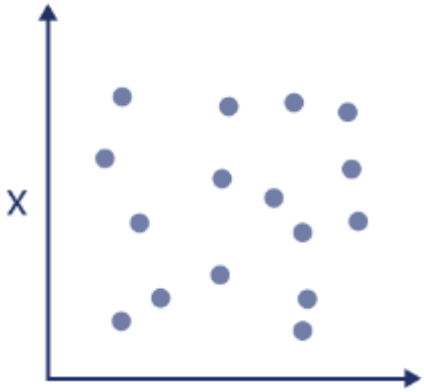


Weak Positive Correlation



Weak Negative Correlation

- When *r* is 0, a line of best fit is not helpful in describing the relationship between the variables:



No Correlation

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

Ans:-

- 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.
- 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.
- Minmax scaling is defined as **x = x – min(x) / max(x) – min(x)**

Standarized Scaling:-

- Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- Standardization is defined as **x = x – mean(x) / sd(x)**
- 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.

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.

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 plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles.
Also, it helps to determine if two data sets come from populations with a common distribution.
Uses:-
- It can be used with sample sizes also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set**.**
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Importance:-

- A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

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:-
   - Book bike rent count is more in the year 2019.
   - Bike rent is higher in the fall.
   - Bike rental count is more on the weekends and holiday than on weekdays.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans:-
drop_first=True is important to use as it helps in reducing the extra column created during dummy variable creation.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans:-
By checking the p-value and the VIF of the variables.

We could have:

- High p-value, high VIF
- High-Low: High p , low VIF: remove these first LOw p
- high VIF: remove these after ones above
- Low p, low VIF

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans:-
   Holiday
   Spring Season
   July Month