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: I have done analysis of the categorical variables using plots. Below are few points we can infer from the visualization.

a) Fall season seems to have attracted more bookings, and in each season booking count has increased drastically from 2018 to 2019.

b) Clear weather attracted more booking which seem obvious from plot..

c) Wednesday, Thursday and Saturday have more number of bookings .

d) Booking have been increased till September month and then gradually decreased.

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

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

SYNTAX: drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in Categorical column and we want to create dummyvariable for that column. If one variable is not X and Y, then it is obvious Z. So we do not need 3rd variable to identity the Z.

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

Answer: 'Temp' variable has the highest correlation with the target variable.

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

Answer: Validated the assumptions based on below:

a) The error terms are normally distributed.

b) The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.

c)The predicted values have linear relationship with the actual values.

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: Below are the top 3 features contributing towards explaining the demand of the shared bikes -

> Temp

> Spring

> Saturday

General Subjective Questions :

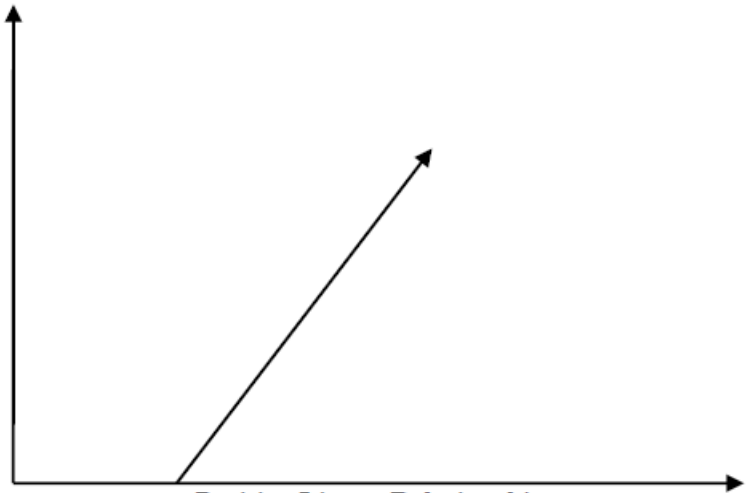
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that hen the value of one or more independent variables will change, the value of dependent variable will also change accordingly (i.e,) either increase or decrease.

> POSITIVE LINEAR RELATIONSHIP:

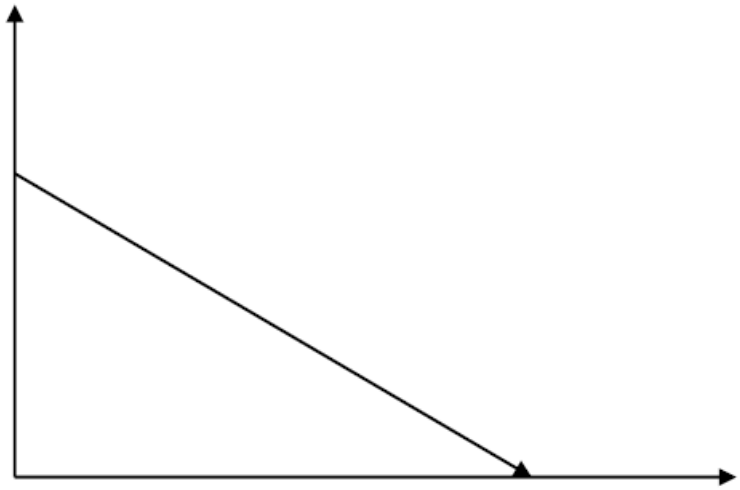
A Linear relationship will be called positive if both independent and dependent variable increases. It can be understood from the graph below



Positive Linear Relationship

> NEGATIVE LINEAR RELATIONSHIP:

A Linear relationship will be called negative if both dependent and independent variable increases and dependent variable decreases



Negative Linear Relationship

TYPES OF LINEAR REGRESSION:

1) Simple Linear Regression

2) Multi Linear Regression

ASSUMPTIONS:

Following are some of the assumptions about the dataset that is made by Linear Regression model -

> MULTICOLLINEARITY

> AUTOCORRELATION

> RELATIONSHIP BETWEEN VARIABLES

> NORMALITY OF ERROR TERMS

> HOMOSCEDASTICITY

2. 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.

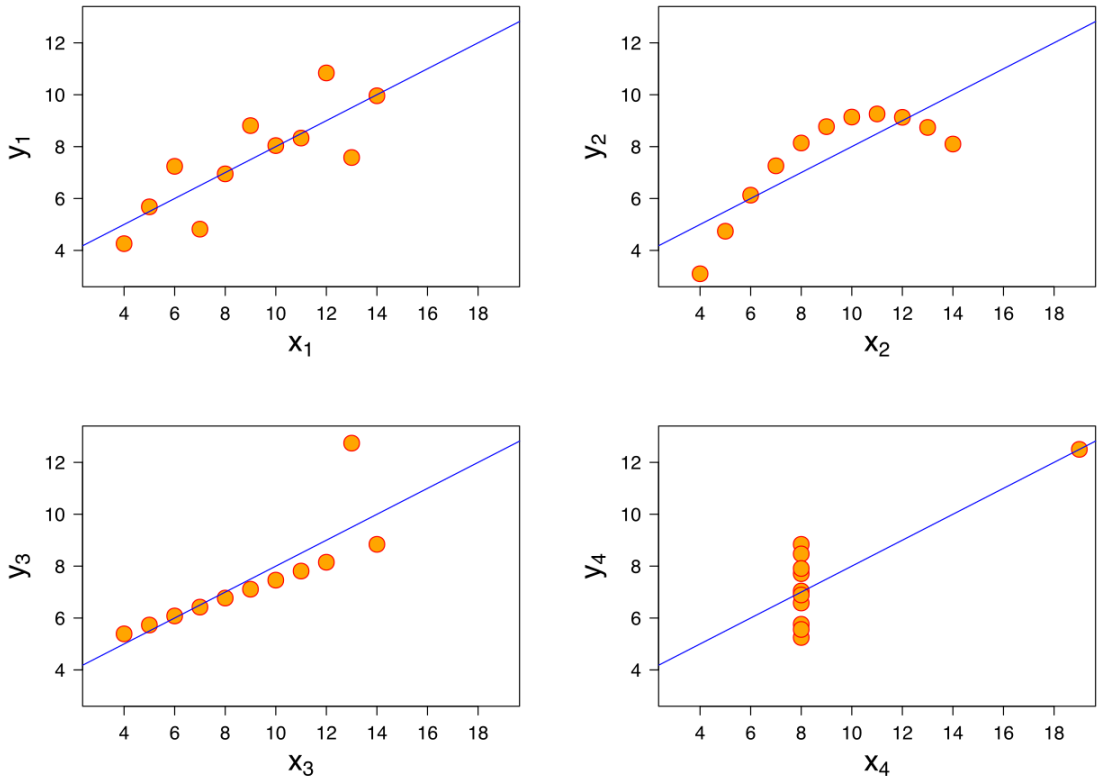
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

Quartet's Summary Stats

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.

3. What is Pearson’s R? (3 marks)

Answer:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

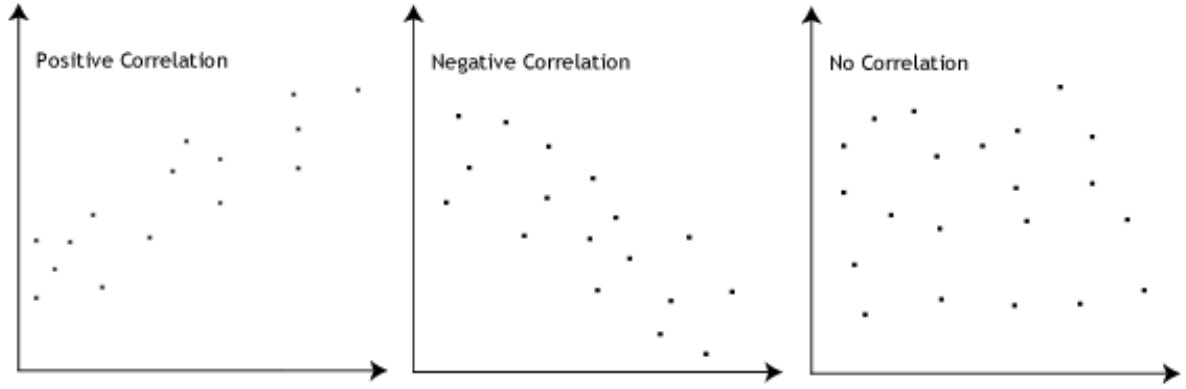
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

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



Pearson r formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

r=correlation coefficient

x_{i}=values of the x-variable in a sample

\bar{x}=mean of the values of the x-variable

y_{i}=values of the y-variable in a sample

\bar{y}=mean of the values of the y-variable

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

and standardized scaling? (3 marks)

Answer:

SCALING:

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.

Why Scaling:

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).

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

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)

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. 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 for determining if two data sets come from populations with a common distribution such as a Normal, Exponential, or Uniform distribution. This helps in a scenario of linear regression when we have the training and test data set received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with the same distributions.

It is a probability plot for comparing two probability distributions by plotting their quantiles against each other. Quantiles are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities. “Q” stands for quantile. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. If the two data sets have come from populations with different distributions then the data points will be far from the reference line.

The q-q plot is used to find out the following:

Whether the two data sets come from populations with a common distribution, Whether the two data sets have a common location and scale, Whether the two data sets have similar distributional shapes, Whether the two data sets have similar tail behavior.

THANK YOU.