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 analysed categorical variables from dataset through a> bar plot b>box plot.

There are following points we can infer from graphs;

A>During year 2019, no of booking has increased as compared to year 2018

B>Booking is almost equal on working vs non working day

C>Booking is more when it is holiday as compared to non-holiday since people find more time to visit different places on bike

D>Fall season has attracted more bookings and has gone up from 2018 to 2019 during each season

E>Mostly booking has taken place during months of May , June, July, Aug, Sep and Oct. It has started decreasing towards end of the year

F>When we move towards end of the week , booking increases i.e from Thursday onwards

G>More booking is observed during clear weather since people prefer bike during clear weather only.

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

Answer:

Drop_first=True, reduces extra column created during dummy variable creation

It is helpful in reducing correlation created among dummy variables

Syntax: drop_first : Type- bool, default value-False, Description-It implies whether to get k-1 dummies out of k categorical levels by removing first level

Suppose we have 4 types of variables in categorical column and we create dummy variable out of it. If one variable is not P,Q,R then it will be S .we do not require 4th variable to identify S.

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 highest correlation with Target variable

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

Answer:

I have validated the assumptions of Linear Regression on training set based on following 5 assumptions;

1.>Normality of error terms-

Error terms should be normally distributed

2.>Multicollinearity check-

There should be insignificant multicollinearity among variables

3.>Linear Relationship Validation-

There should be linear relationship between variables

4.>Homoscedasticity-

There should not be visible pattern in residual values

5.>Independence of residuals-

There should not be auto-correlation

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:

Top 3 features contributing significantly towards explaining the demand of the shared bikes are;

1.temp

2.winter

3.sep

General Subjective Questions

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

   Answer:

   It is statistical model which analyses the linear relationship between dependent variable and with given set of independent variables.

Linear relationship between variables means that if there is change in value of independent variable then the value of dependent variable will change accordingly.

The equation of line representing the relationship is ;

y= mx+c

Here y=Dependent variable which is to be predicted

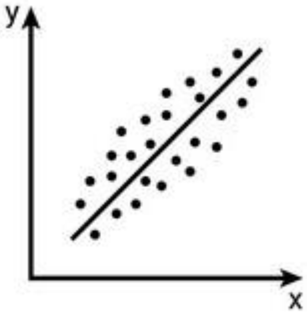
x=Independent variable

m=Slope of the regression line

c=constant, also known as y intercept

There are two types of linear relationship;

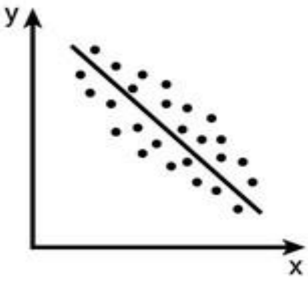
1.Positive linear relationship-As independent variable increases, dependent variable also increases



As per graph above with positive slope, with increase in X, Y also increases with positive slope.

2.Negative Linear Relationship-

The relationship between variables is called negative if with increase in one variable, there is decrease in other variable and vice versa.



There are two types of Linear regression

A. Simple linear regression

B. Multiple linear regression

There are assumptions related to linear regression

A. Multicollinearity-It assumes that there is no dependency or very insignificant relation between independent variables

B. Auto-Correlation-It assumes that there is very little or no auto correlation between residual errors hence they are not dependent on each other

C. Linear Relationship-It assumes that there should be linear relationship between dependent and independent variables.

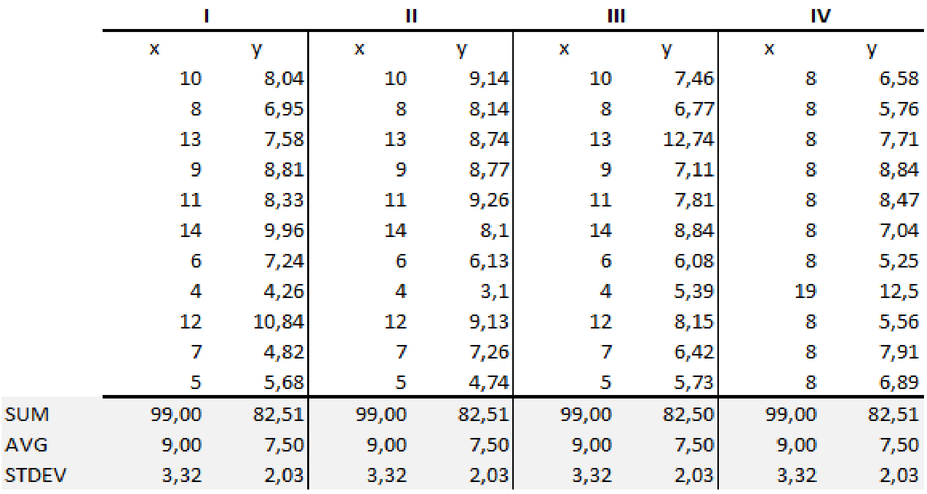
D. Normality of error terms-It assumes that error terms should be normally distributed

E. Homoscedasticity- It assumes that residuals have constant variance at every level of independent variable. we use scatter plot to check it.

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

Answer

Anscombe's quartet was formulated by Francis Anscombe. It is framed with four datasets each containing eleven (x, y) pairs. When it comes to descriptive statistics, these statistics share the same outcome but when these dataset is plotted in graphs ,the pattern shows different trends for all graphs.



	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 above summary statistics shows that there were identical variances for x and y for all groups.

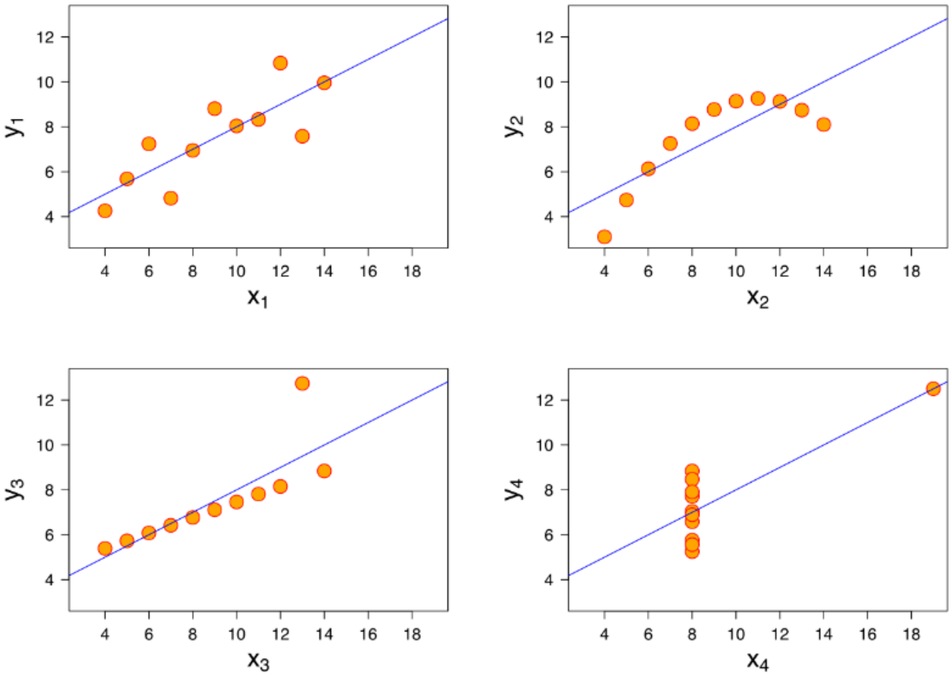
Summary statistics similarity-

Mean of x is 9 and y is 7.50 for each dataset

The correlation coefficient for all dataset is 0.816

Variance of x is 11 and y is 4.13 for all dataset

When graph is plotted for all dataset ,we can infer that they show same regression line but trends are different for all graphs.



We can infer following outcomes from above graph;

Dataset I has clean and well fitted regression line.

Dataset II is not normally distributed

Dataset III is linear, regression is thrown off because of outlier

Dataset IV shows that there may be high correlation even if there is one outlier

This explains clearly that data visualisation is important which helps to reach us to right conclusion.

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

Answer:

It is numerical outcome which shows the strength of linear relation between variables.

If variable increases and decreases together with same trend, the coefficient will be positive otherwise it becomes negative in reverse case.

The Pearson's coefficient value ranges from -1 to +1.The 0 value indicates that there is no correlation.

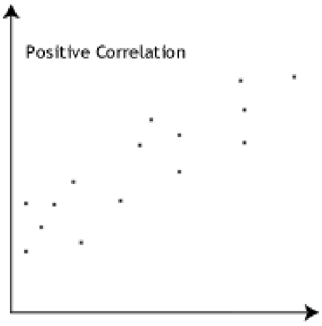
Positive value shows that with increase in one variable, there will be increase in other variable also and with decrease in one variable, there is decrease in other variable .

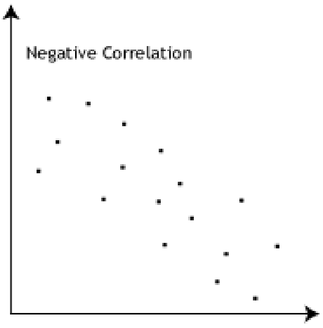
There is reverse trend when coefficient is negative.

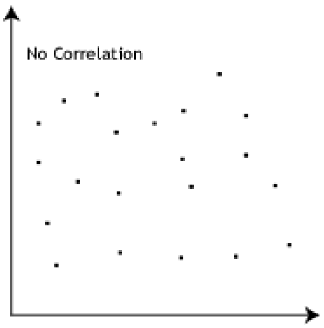
Formulae for Pearson's coefficient(r)

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

This is shown in below graphs.







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 process in data pre-processing which is applied to independent variables to normalize data within a particular range.

It is performed because algorithm takes into account the magnitude of data only and not the unit and collected data varies widely in units, range and magnitude. Hence to resolve this issue scaling is performed to bring all variables to same range of magnitude to make correct prediction.

Example -Algorithm will process the value of 1000 meter as higher than 900 KM which is wrong and hence prediction will be wrong.

Difference between normalized scaling and standardized scaling;

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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 between independent variable then VIF value comes to be infinite.

A high value of VIF indicates that there is correlation between variables.

In case of perfect correlation, we get R squared(R2) =1 and when we calculate 1/(1-R2) as infinity since denominator becomes 0

To solve this problem, we need to drop variable which causes perfect multicollinearity.

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

(3 marks)

Answer:

The quantile-quantile plot (Q-Q plot) is graphical way to conclude whether two datasets have originated from populations with common distribution.

Use of Q-Q Plot

It is a plot of quantiles of the first dataset against quantiles of second dataset. Quantile refers to fraction of points below given value. If quantile is 40%,it means that 40% of data falls below and 60% falls above it.A-45 degree reference line is plotted. If two datasets come from same population with the same distribution, these values should align approximately along this line. The more the departure, the greater the evidence that datasets have from population with different distribution.

Importance of Q-Q Plot

If there are two samples of data, we check whether the assumption of common distribution is justified or not. The location and scale estimator can pool both datasets to get common location and scale. If samples differs, we can get insights of differences.

It provides more clarity regarding nature of difference as compared to chi-square and Kolmogorov-Smirnov 2 sample tests.