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?

Answer:

From the analysis on categorical columns we can infer following from the visualization :

1: Fall season has attracted more booking and in each season the booking count has increased drastically from 2018 to 2019.

2: Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.

3: Good weather (Clear, Few clouds, Partly cloudy, Partly cloudy) attracted more booking.

4: Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.

5: When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.

6: Booking seemed to be almost equal either on working day or non-working day but booking count has increased from 2018 to 2019 for both working and non-working day.

7 : 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

2: Why is it important to use **drop_first=True** during dummy variable creation?

Answer :

It is important to use **drop_first=True** during dummy variable creation as it helps in reducing extra column created during dummy variable creation. Hence it reduces collinearity among dummy variables.

Keeping k number of dummy variable for k categories will create a redundant column which is not needed as the combination of other dummy variables will uniquely represent this redundant column.

3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

Pair-plot among variables indicate that '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?

Answer:

1: Errors should be normally distributed and centred around 0 , this is done by producing a distribution plot and the error terms must follow normal distribution pattern.

2: Multicollinearity is checked by using VIF and all the features should be below 5.

3: Error terms should be independent of each other , this is done by plotting graph of index with error terms and observe that there should be no visible pattern

4: Linearity validation is done by plotting graph of component with component+residual.

5: Plotting graph of actual and predicted data and both should follow the same patterb

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:

Based on the final model the top 3 contributing feature are:

1: temp

2: yr

3: season_winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

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 when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

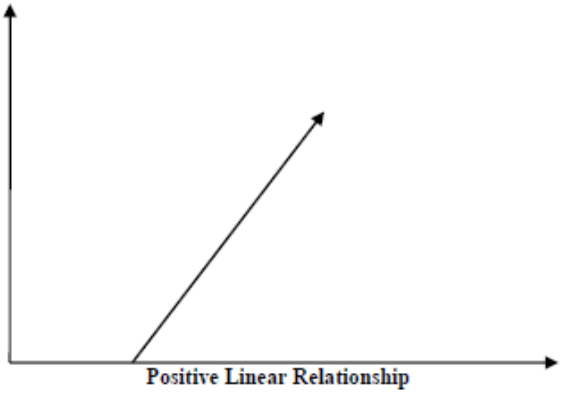
m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

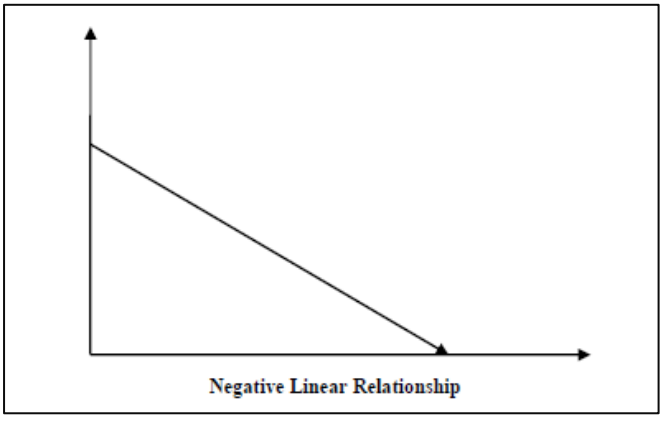
1 : Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



2: Negative Linear relationship:

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Linear regression is of the following two types −

1:Simple Linear Regression : SLR is used when dependent variable is predicted using a single independent variable.

2:Multiple Linear Regression : MLR is used when dependent variable is predicted using a multiple independent variable.

2. Explain the Anscombe's quartet in detail.

Answer:

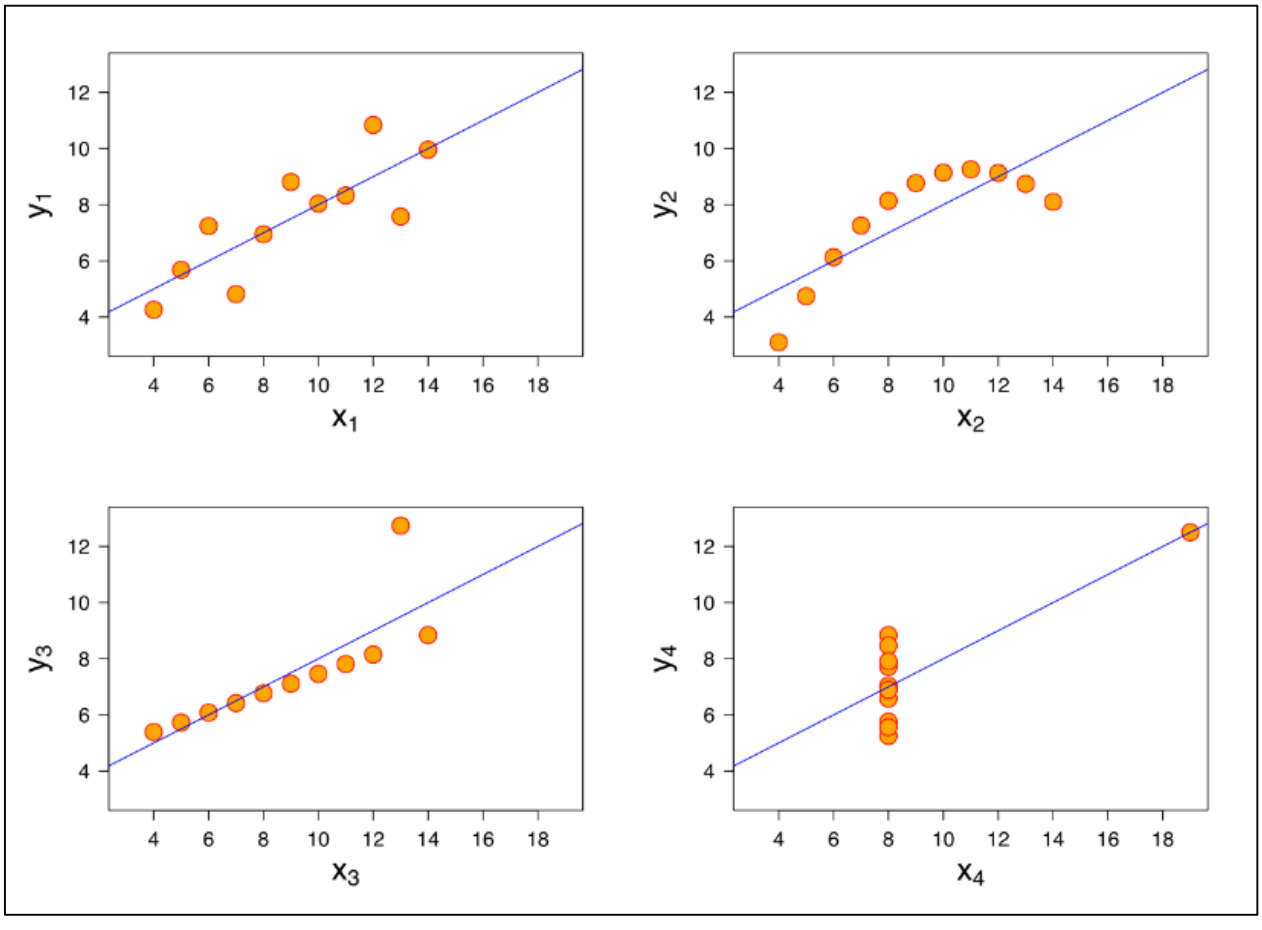
Anscombe's Quartet comprises four datasets, each containing eleven (x, y) pairs which share the same descriptive statistics but things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

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

	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

- We can see that in each dataset mean of x is 9 and mean of y is 7.50 .
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient 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 produces different observation :



- 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.

3. What is Pearson's R?

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. 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?

Answer:

Scaling is a technique used to normalize independent variable within a specified range.

It is performed during data pre-processing to handle highly varying values or units .

If scaling not done then machine learning algorithm weighs greater value high and smaller value lower irrespective of the unit of values or we can say that the algorithm take magnitude only in account not units hence resulting in incorrect modelling .

Difference between normalized scaling and standardized scaling is as follows:

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].	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?
Answer:

The value of VIF becomes infinite when there is a perfect correlation between two variables. If VIF is 3 it means that the variance of the model is inflated by a factor of 3 due to the presence of multicollinearity.

VIF = 1/(1-R^2)

VIF becomes infinity when R^2 becomes 1 which indicates perfect correlation . To solve this we need to drop one of the variable from the dataset which is causing multicollinearity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.