Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)**

Answer:

As per the analysis of the categorical variables using boxplot and bar plot, below are the points we can infer from the visualization:

- Fall season has the highest number of bookings with respect to all seasons. And in each season the booking count has increased from 2018 to 2019.

- Most of the bookings were done during May, June, July, August, September and October Month. The number of bookings increases till the mid of year and gradually decreases as we move towards the end of the year.

- 2019 attracted more bookings with respect to previous year i.e. 2018 which shows good business progress and more profits.

- Thursday, Friday, Saturday and Sunday have more booking w.r.t to pother days of the weeks.

- Number of bookings for working and non-working days are almost same.

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

Answer:

It is important to use drop_first=True as it helps in reducing the extra column which is created during the dummy variable creation. This way the correlation created between the dummy variables is reduced.

Syntax for using the drop_first is: **drop_first= bool**. The default value is false. Used to get n-1 variables out of n categorical levels by removing the first level.

Example: We have 3 categorical columns as X, Y,Z and we want to create the categorical column for them. Then there is no need to create a variable to identify Z because if variable is NOT X and NOT Y then variable is obviously Z.

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

Answer:

Looking at the pair plot, the ‘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:

The assumptions of Linear regression Model can be validated based on the 5 assumptions:

- Normality of errors – Error terms are normally distributed.
- Multicollinearity – There should be insignificant multicollinearity among variables.
- Linear Relationship – There should be linearity among the variables.
- Homoscedasticity – There should be no visible pattern in residual values.
- Independence of residuals – There should be no 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:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- temp
- winter
- year

General Subjective Questions

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

Answer:

Linear regression can be defined as the statistical model that analyzes the linear relationship between the dependent and independent variables with given set of independent variables.
Linear relationship between variables means that the value of one or more independent variables will change (increase/decrease), the value of dependent variables will change accordingly (increase/decrease).
The output variables are called as the dependent variables and the input variables are called the independent variables.

Mathematically the relationship can be represented with the help of following equation −
$Y = mX+ c$

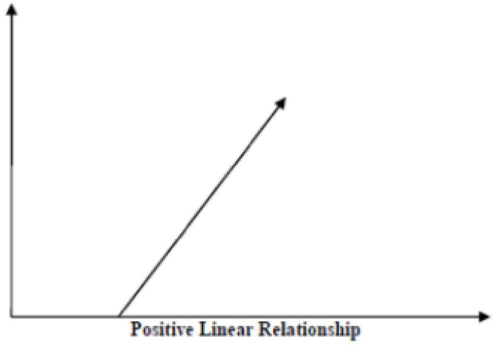
Here,

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

If X = 0, Y would be equal to c.

Based on the nature of relation between the dependent and independent variables the linear regression can be classified as:

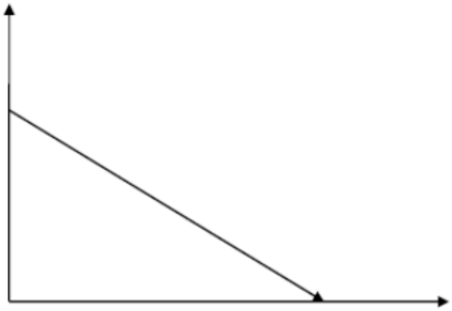
1. Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.
   Graph for Positive Linear Regression is:



Positive Linear Relationship

2. Negative Linear Relationship: A linear relationship will be called positive if independent increases and dependent variable decreases.

   Graph of Negative Linear Regression is:



Negative Linear Relationship

Based on the number of variables, Linear regression can be classified into 2 types:

1. Simple Linear Regression: Linear regression only has one independent variable impacting the slope of the relationship between the dependent and independent variable.

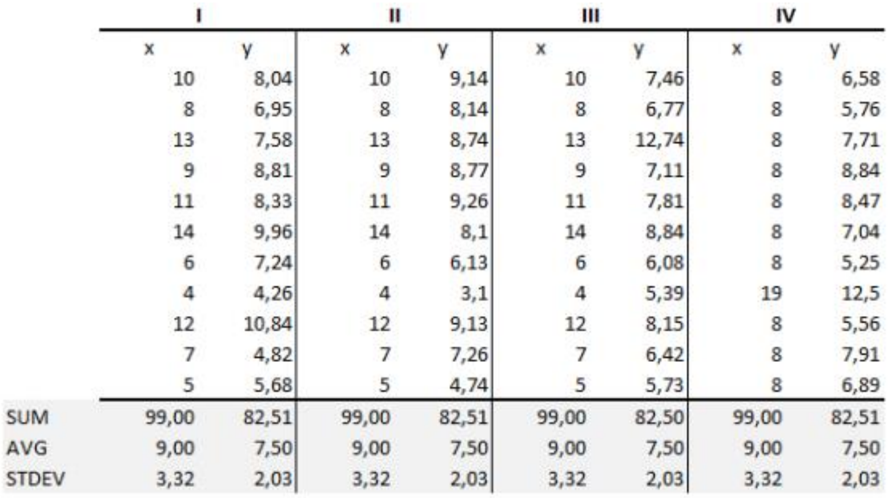
2. Multiple Linear Regression: Multiple regression incorporates multiple independent variables.

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

Answer:

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc.) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

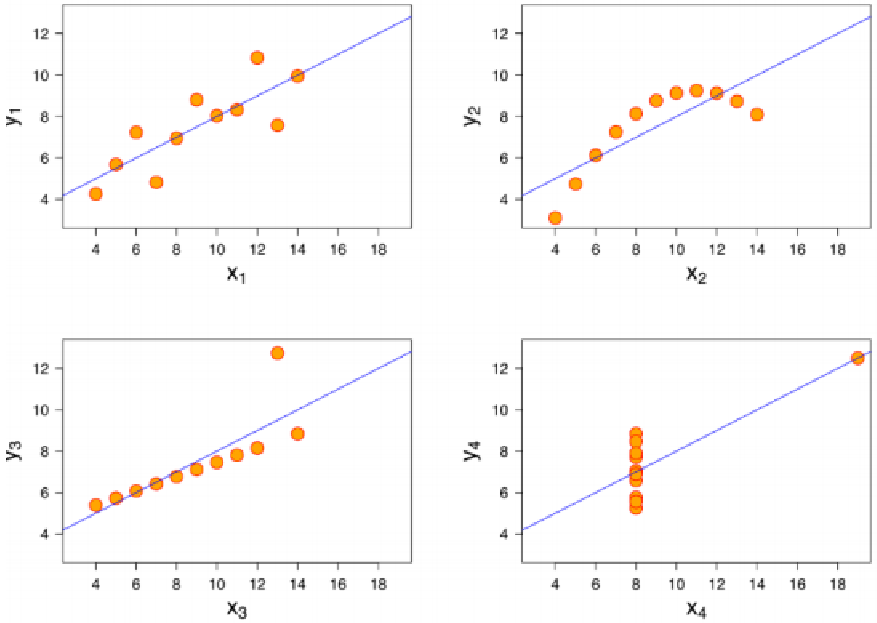


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15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

From the above table we can see the following details:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- The variance of x is 11 and variance of y is 4.13 for each dataset

But, 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:



From the above plot we can produce the following details:

- Dataset 1 appears to have clean linear models.
- Dataset 2 is not distributed normally.
- In Dataset 3 the distribution is linear, but the calculated regression is disturbed by an outlier.
- Dataset IV shows high correlation coefficient

So from analysis using Anscombe's quartet we can prove the importance of visualization in Data Analysis.

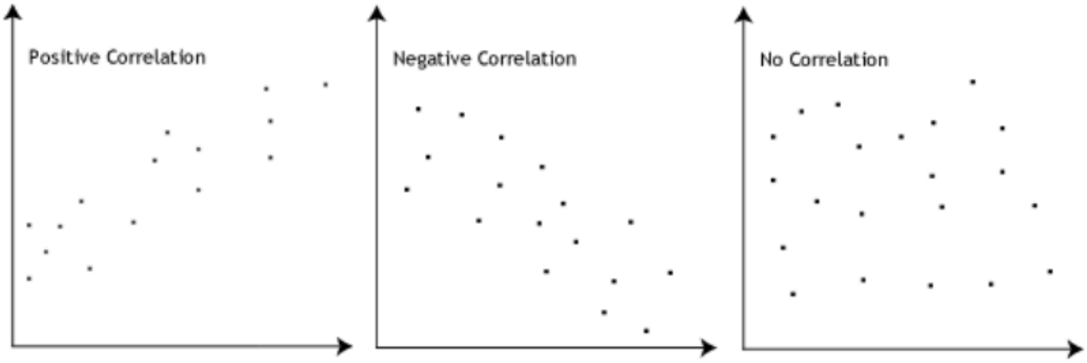
3. What is Pearson's R? (3 marks)

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the different 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.

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

Answer:

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example:

Suppose in an algorithm we are calculating the height of humans. If some heights are in centi-meter and some in meters. In such cases it can lead to wrong predictions.

To avoid such wrong predictions, the range of all features are scaled so that each feature contributes proportionately and model performance improves drastically.

Difference between Normalized and Standardized scaling:

Sr. No.	Normalized Scaling	Standardized Scaling
1.	In this case the maximized and minimized values are used for scaling.	Here mean and Standard deviation is used for scaling.
2.	It is affected by outliers.	It is not affected by outliers.
1.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
2.	It is called as Scaling Normalization	It is called as Z-Score Normalization.
3.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.

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. 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 independents 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)

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.