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: Identified following as categorical columns to see how predictor variables stands against the target variable.

- Prepared Box plots for columns like season, mnth, weekday, weathersit, holiday, workingday, yr
- Prepared Bar plots for the same against count

Following is the inference from the plots:

- Season Fall has more bookings compared to others seasons, there is a increment in the demand from year 2018 (red bar) to 2019 (blue bar)
- Relatively June month have more bookings compared to other months during year 2018, whereas in 2019 November month has highest demand.
- Clear weather has highest bookings, Misty weather has occupied next position.
- Bookings are almost same entire week, there is no much difference between weekday or weekend. This is applicable for both the years.
- As discussed in Point 1, 2019 have about 60% extra bookings compared to the year 2018

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

Answer:

- drop_first=True drops the first column during dummy variable creation. Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".
- We do NOT need another column for "Uknown".
    - It can be necessary for some situations, while not applicable for others. The goal is to reduce the number of columns by dropping the column that is not necessary. However, it is not always true. For some situations, we need to keep the first column.

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 assumption of Linear Regression Model based on below 5 assumptions -

- Normality of error terms : Error terms should be normally distributed

- Multicollinearity check: There should be insignificant multicollinearity among variables.
- Linear relationship validation: Linearity should be visible among variables
- Homoscedasticity: There should be no visible pattern in residual values.
- Independence of residuals: 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: I see mainly 2 features like, Weather Situation (Clear weather) and Season (Fall) have contributed significantly and third feature could be children holiday season (month).

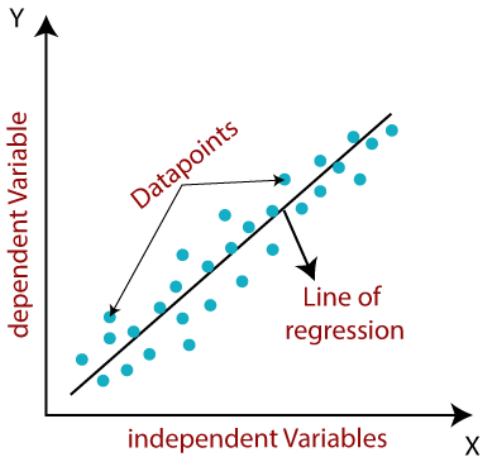
General Subjective Questions

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

Answer: Linear regression is one of the earliest and most popular Machine Learning alrgorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as Sales, salary, age, product price etc.,

inear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

y= a0+a1x+ ε

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

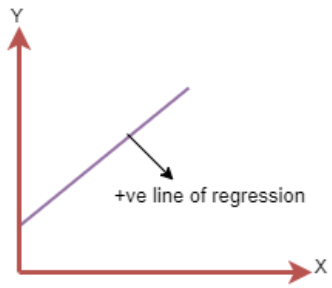
Linear regression can be further divided into two types of the algorithm:

- *Simple Linear Regression:*
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- *Multiple Linear regression:*
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

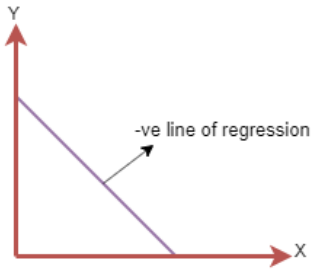
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- *Positive Linear Relationship:*
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- *Negative Linear Relationship*:
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

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

Answer:

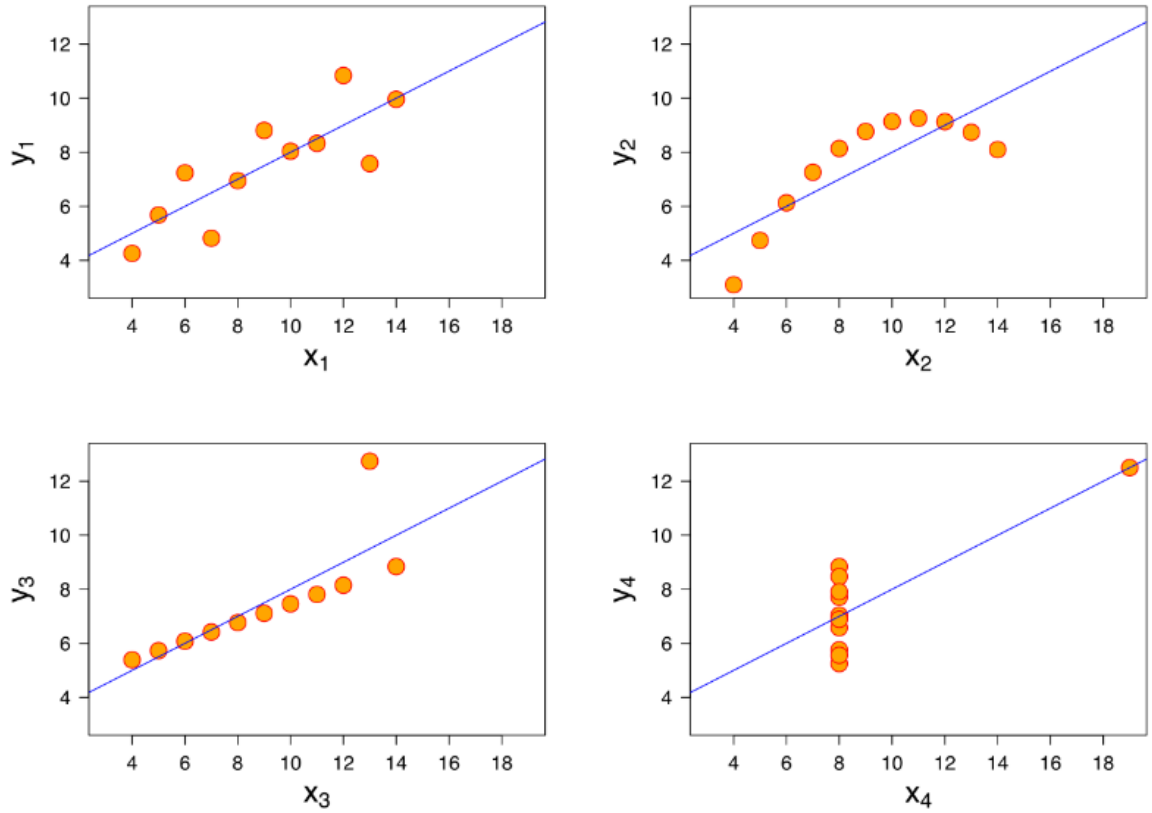
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when 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

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.
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? (3 marks)

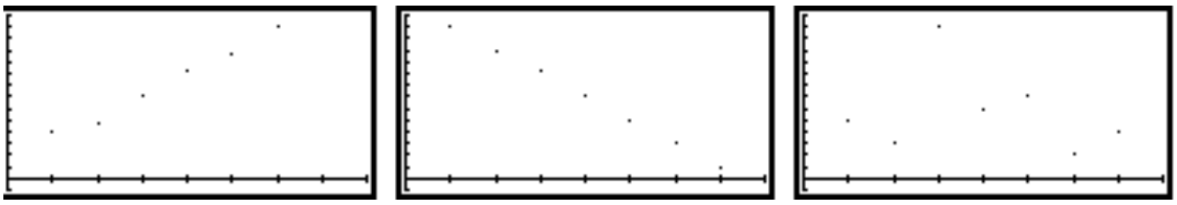
Answer:

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'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

The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233



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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

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.

Difference between Normalized scaling and standardized scaling:

Sl.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, 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 independent 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.