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)
    - Count of total rental bikes is seen to be high on the fall season.
    - Business is doing better in 2019 wrt 2018. Rentals increased in 2019.
    - September has highest number of bookings while median for bike rental count is highest for July. Trend of booking bikes increases from July and by October the booking trend decreases as winter approaches.
    - Thursday, Friday, Saturdays have more rentals than other days of the week.
    - There are less rentals on holidays than non-holidays.
    - In Clear, Few clouds, Partly cloudy, Partly cloudy weather conditions, there are more bookings.
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   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 n-1 dummies out of n categorical levels by removing the first level.
   Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not Alpha and Beta, then It is obvious Gamma. So we do not need 3rd variable to identify the Gamma.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Temp and atemp have 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)
   I have 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)
   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
    - atemp
    - light_snowrain

- year

General Subjective Questions

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

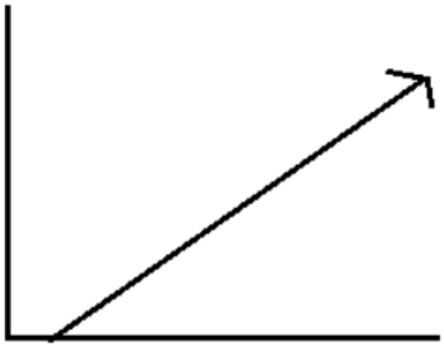
   Linear regression can be defined as a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to find the best-fitting linear equation that predicts the dependent variable based on the independent variable(s). The algorithm calculates the slope and intercept of the line that minimizes the difference between predicted and actual values, allowing for predictions of the dependent variable for new input data.

   The equation for a simple linear regression with one independent variable can be represented as:
   $Y = mX + c$
   where m is the slope of the line.
   c is the y-intercept,
   X is the independent variable,
   Y is the dependent variable.
   if X = 0, Y would be equal to c

   The linear relationship can be positive or negative in nature:

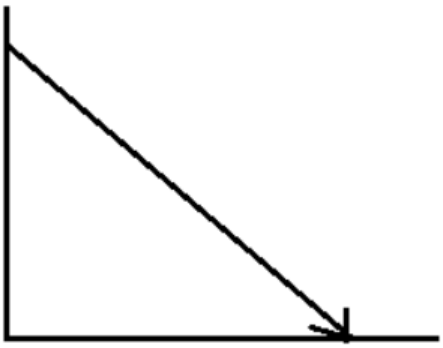
   Positive Linear Relationship:

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



   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
2. Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

1. Multi-collinearity –

Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

2. Auto-correlation –

Another assumption Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

3. Relationship between variables –

Linear regression model assumes that the relationship between response and feature variables must be linear.

4. Normality of error terms –

Error terms should be normally distributed.

5. Homoscedasticity –

There should be no visible pattern in residual values

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven paired observations (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. Despite having similar means, variances, correlations, and regression lines, they exhibit very different patterns when plotted.

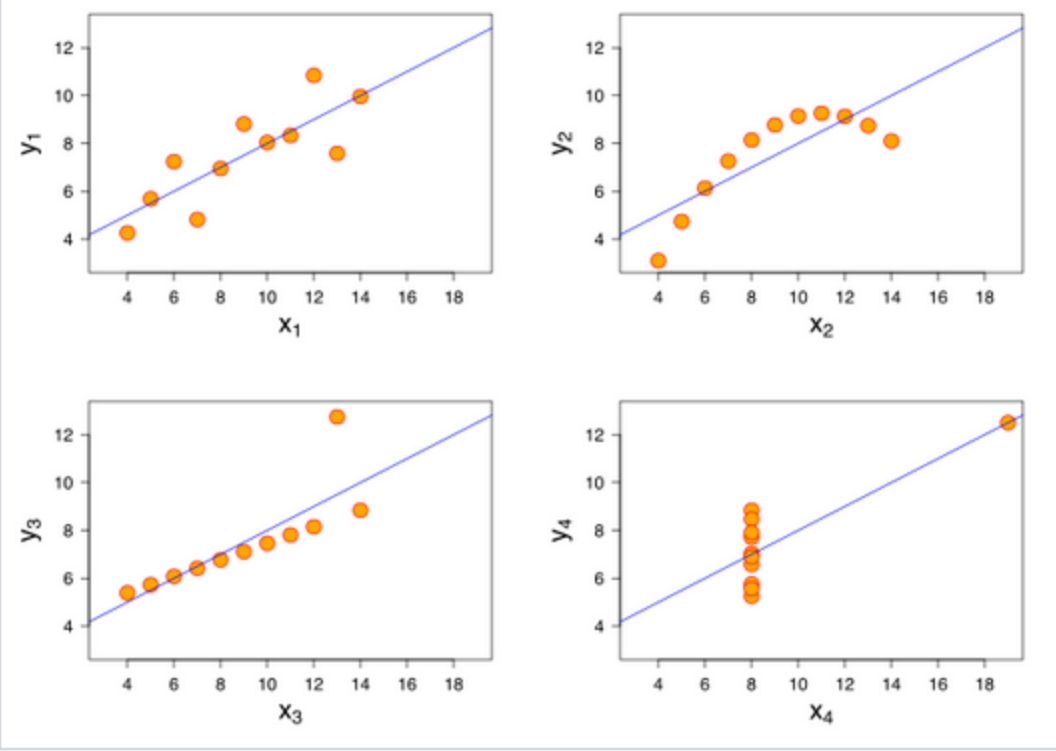
This quartet emphasizes the importance of graphical exploration in understanding the nature of data and highlights that datasets with similar statistical properties can lead to different interpretations and conclusions.

It's often used to demonstrate the limitations of summary statistics and the significance of data visualization in 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

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 are 0.816 for each dataset



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Dataset I appear 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)

Pearson's r or correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

r=1 indicates a perfect positive linear relationship (as one variable increases, the other also increases proportionally).

r=−1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally).

r=0 indicates no linear relationship between the variables.

Formula

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

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)

Scaling is the process of transforming the range of variables or features to a consistent scale. It is often performed to ensure that all variables contribute equally to the analysis and to avoid issues caused by differences in the scales of different features.

Scaling is performed for 2 main reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent method

Differences:

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation
Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
It is really affected by outliers.	It is much less affected by outliers
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)

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)

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.