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 1:

   I can infer below effects on the dependent variable (count of bike bookings) based on the analysis of Categorical variables:

   1. The bike booking has **increased** substantially in **year 2019** than **2018**.
   2. There are **more** bookings in **fall season** than other seasons.
   3. Bookings are **increased** in **September month** by **0.0767** unit.
   4. Bookings **reduce** are done in **July month** by **0.0524** unit.
   5. On **holiday** there is **decrease** of **0.0980** units of bike bookings.
   6. Every time the **weather** is **Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** the bike booking is **reduced** by **0.2852** units.
   7. When **weather** is **Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist** there is **decrease** of **0.0816** units in the bike bookings.
   8. **Winter season** there is **increase** of **0.0831** units of bike bookings.
   9. **Summer season** there is **increase** of **0.0453** units of bike bookings.
   10. **Spring season** there is **decrease** of **0.0669** units of bike bookings.

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

   Answer 2:

   1. In Linear Regression categorical variables are converted to numerical variables by dummy variable creation.
   2. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. When all new dummy columns have value 0 it indicates nth level exist implicitly.
   3. This is done by get_dummies() function of pandas library which uses parameter **drop_first**, when set True **will not create new dummy column for level 'n'** and is by default False**.**
   4. **This will reduce correlation of one dummy column with target variable.**

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

   Answer 3: '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 4:

   I validate the assumptions of Linear Regression after building the model of the training set based on below factors:

   1. **Normality of Error terms:** Error terms should be normally distributed.
   2. **Multicollinearity:** Detecting associations between predictor variables and removing it.
   3. **Linearity:** There should be linear relationship between feature variable and target variable.
   4. **Homoscedasticity:** Error terms should have constant variance or there should be no pattern in residual values.
   5. **No auto-correlation:** Error terms should be independent of each other

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 5:

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes, based on final model:

   1. temp
   2. year
   3. weather_light (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

General Subjective Questions

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

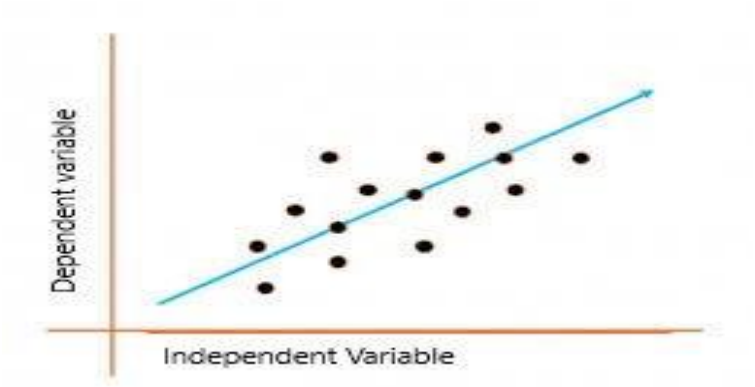
Answer 1:

1. Linear regression predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables.
2. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values.
3. Applied in various domains like economics and finance, this method analyses and forecasts data trends.
4. It can extend to multiple linear regression involving several independent variables and logistic regression, suitable for binary classification problems

Linear Regression is classified into Simple and Multiple Linear Regression.

Simple Linear Regression

- In a simple linear regression, there is one independent variable and one dependent variable.
- The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables.
- The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.



To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$\mathbf{Y_i = \beta_0 + \beta_1 X_i}$

where Y_i = Dependent variable, β_0 = constant/Intercept, $\boldsymbol{\beta_1}$ = Slope/Intercept, $\mathbf{X_i}$ = Independent variable.

<u>Residuals:</u>

In regression, the difference between the observed value of the dependent variable($\mathbf{y_i}$) and the predicted value(**predicted**) is called the residuals.

$\boldsymbol{\varepsilon_i = y_{predicted} - y_i}$ **where** $\mathbf{y_{predicted} = B_0 + B_1 X_i}$

Best-Fit Line:

The best fit line is a line that fits the given scatter plot in the best way. Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares (RSS)

Cost Function for Linear Regression:

In Linear Regression, generally **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the $y_{predicted}$ and y_i. We calculate MSE using simple linear equation y=mx+b:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (B1x_i + B0))^2$$

Using the MSE function, we'll update the values of B_0 and B_1 such that the MSE value settles at the minima.

Gradient Descent is one of the optimization algorithms that optimize the cost function (objective function) to reach the optimal minimal solution which can be used.

Evaluation Metrics for Linear Regression:

- The strength of any linear regression model can be assessed using various evaluation metrics.
- These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

Coefficient of Determination or R-Squared (R2)

Mathematically it can be represented as,

$R^2 = 1 – (RSS/TSS)$

- **Residual sum of Squares (RSS)** is defined as the sum of squares of the residual for each data point in the plot/data. It is the measure of the difference between the expected and the actual observed output.

$$RSS = \sum_{i=1}^{n}(y_i - b_0 - b_1x_i)^2$$

- **Total Sum of Squares (TSS)** is defined as the sum of errors of the data points from the mean of the response variable. Mathematically TSS is,

$$TSS = \sum (y_i - \bar{y}_i)^2$$

Hypothesis in Linear Regression:

The Null and Alternate hypotheses in this case are:

H_0: $B_1 = 0$

H_A: $B_1 \neq 0$

To test this hypothesis, we use **t statistic**, **F statistic**

Multiple linear regression:

- Multiple linear regression is a technique to understand the relationship between a *single* dependent variable and *multiple* independent variables.
- The formulation for multiple linear regression is also similar to simple linear regression with the small change that instead of having one beta variable, you will now have betas for all the variables used.
- The formula is given as:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_p X_p + \boldsymbol{\varepsilon}$$

Assumption of Linear Regression:

1. **Normality of Error terms:** Error terms should be normally distributed.
2. **Multicollinearity:** Detecting associations between predictor variables and removing it
3. **Linearity:** There should be linear relationship between feature variable and target variable.
4. **Homoscedasticity:** Error terms should have constant variance or there should be no pattern in residual values.
5. **No auto-correlation:** Error terms should be independent of each other

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

Answer 2:

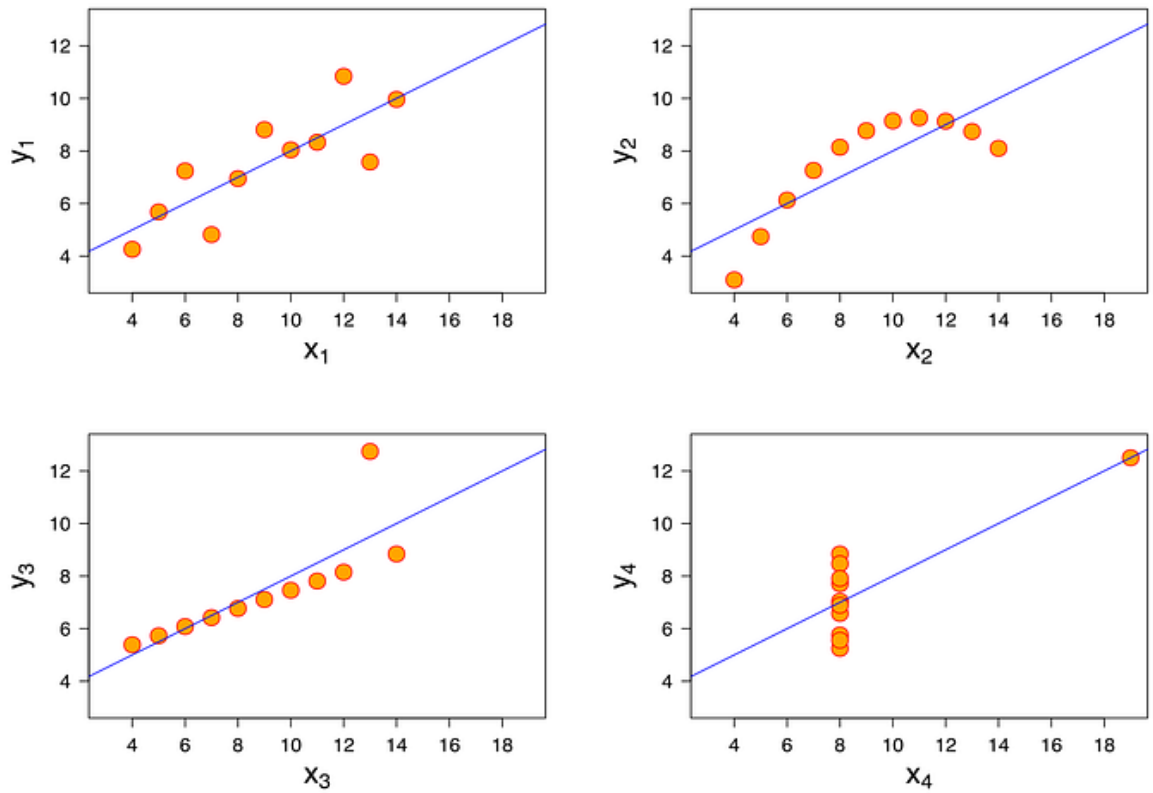
- **Anscombe's quartet** comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.
- The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
- The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths.
- Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
- **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.
- It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

	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 plotted differently:



- 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.

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 3:

- The **Pearson correlation coefficient (*r*)** or **Pearson's R** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
- It summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Below is the formula to find Pearson's R:

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

The Pearson's R is a good choice to measure a correlation when:

- Both variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear

Interpretation of Pearson's R:

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.

- The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.
    - When *r* is 1 or –1, all the points fall exactly on the line of best fit.
    - When *r* is greater than .5 or less than –.5, the points are close to the line of best fit.
    - When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit.
    - When *r* is 0, a line of best fit is not helpful in describing the relationship between the variables.

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

Answer 4:

- 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: If feature scaling method is not used in an algorithm, then it can consider the value 5000 grams to be greater than 6 kg which is not true in this case. The algorithm will give wrong predictions. So, Feature Scaling is used to bring all values to same magnitudes.

Below are the differences between Normalized and Standardized Scaling

Feature	Normalized Scaling	Standardized Scaling
Definition	This technique re-scales a feature or observation value with distribution value between 0 and 1.	It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.
Formula	$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$	$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$

Outliers	It is affected by outliers.	It is less affected by outliers.
Usage	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Range	Scales value ranges between range [0, 1] or [-1, 1]	It is not bounded to a certain range.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer 5:

- If there is perfect correlation between feature and target variable, 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 (R^2) =1, which lead to VIF = 1/ (1-R^2) to 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 6:

- 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.