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:
   I have done analysis on categorical columns using box plot following are inferences:
   - The counts are lowest in the spring season.
   - Summer, fall, and winter have higher median counts compared to spring, with fall having the highest median count.
   - The year 2019 has higher counts compared to the year 2018. This indicates an increase in counts over time.
   - Counts are lowest in January and December, and highest from May to October, indicating a peak usage during these months.
   - Clear weather has the highest counts.
   - Mist has lower counts than clear weather.
   - Light rain has the lowest counts.
   - There is not a significant difference between holidays and non-holidays, but non-holidays have a slightly higher median count.

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

   Answer:
   Using drop_first=True is important to use as it helps to discard one extra column created by dummy variables hence avoids redundant data that can cause multicollinearity.

   Example:
   Suppose you have a categorical variable with three categories: A, B, and C. Creating dummy variables without drop_first=True would result in:

A	B	C
1	0	0
0	1	0
0	0	1

   Using drop_first=True, you drop the first category resulting in:

B	C
0	0
1	0
0	1

   Here, the reference category is A, and the presence of B or C is indicated by the dummy variables, avoiding redundancy and multicollinearity.

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

   Answer:
   'temp' 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:
   - **Normality of Residuals:** Checked by plotting a histogram of residuals to see if they follow a normal distribution.
   - **Homoscedasticity:** Ensured there is no visible pattern in residual values.
   - **Multicollinearity:** Assessed using Variance Inflation Factor (VIF) to ensure predictors are not highly correlated.

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:
   Below are top 3 features contributing significantly towards explaining the demand of the shared bikes
   - Year (yr)
   - Temperature (temp)
   - Winter (winter)

General Subjective Questions

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

   Answer:
   Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting linear relationship between the dependent variable y and the independent variables X.
   This equation takes the form:
   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n$
   Where y is dependent variable
   x_1 x_2 ….. x_n are independent variable
   β_1 β_2 ….. β_n are the coefficients for the independent variables x_1 x_2 ….. x_n
   β_0 is the intercept

   The linear relationship can be either positive or negative in nature:

   - Positive Relationship: If the dependent variable increases as the independent variable increases, the relationship is positive.
   - Negative Relationship: If the dependent variable decreases as the independent variable increases, the relationship is negative.

Assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.
- Normality: The residuals of the model are normally distributed.
- Homoscedasticity: There is no visible pattern in residual values.

Goal of Linear Regression

Linear regression aims to find the best-fit line that minimizes the difference between the observed values and the predicted values. This best-fit line allows us to accurately predict the target variable based on the given features.

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

Answer:

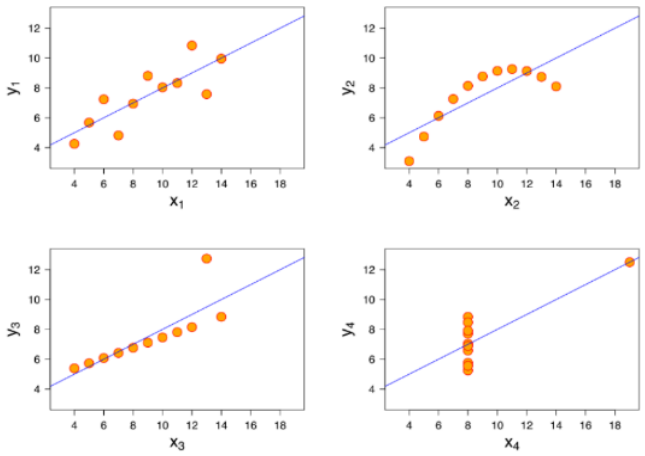
Anscombe's quartet comprises four datasets with nearly identical statistical properties (mean, variance, correlation, etc.) but very different distributions and relationships between variables.

	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 stats shows that mean and variances were identical for x and y

- Mean of x is 9 and of y is 7.50 for each dataset
- Variance of x is 11 and of y is 4.13 for each dataset
- Correlation coefficient between x and y is 0.816 for each dataset

When we plot these four dataset we can observe that they show the same regression line but each dataset represents a different behaviour.



- Dataset I – shows a linear relationship between x and y with some variance around the line.

- Dataset II – displays a curved relationship, indicating that x and y are not linearly related.
- Dataset III –strong linear relationship between x and y with one significant outlier which drastically affects the statistical properties of the data.
- Dataset IV – x remains constant, except for one outlier that makes it challenging to determine any meaningful relationship between x and y.

It demonstrates the importance of visualising data even if datasets have similar statistical summaries, their graphical representations can reveal crucial differences that affect the interpretation and conclusions of the analysis.

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

Answer:
Pearson's R or Pearson correlation coefficient shows the strength and direction of linear relationships between two variables. Just imagine that you have two sets of data such as height and weight; Pearson's R can tell whether these sets move together or not, if so, in what manner. It varies from -1 to 1 where 1 denotes perfect positive linear relationship, -1 is perfect negative linear relationship and 0 represents no linear relationship. For example if Pearson's R for height against weight is 0.8 it means that taller people generally weigh more with a strong relationship . On the other hand, when it is -0.8 one variable increase will lead to decrease in the other also known as a strong relationship between them. If it's around 0, it means height and weight don't have any linear relationship. Pearson's R gives a quick and easy way to understand relationships in your data and is widely used in statistics to explore the connection between two variables.

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

Answer:
Scaling is the process of transforming the features of a dataset so that they are on a similar scale. It is a crucial step in data pre-processing, especially when dealing with features having varying values. Without feature scaling, machine learning algorithms might incorrectly weigh greater values more heavily and consider smaller values as lesser, regardless of their actual significance.

Sr 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	Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
3	It is affected by outliers.	It is less affected by outliers.
4	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:
   A Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of other predictor variables.This means that there is a perfect correlation between two or more independent variables. In such cases, the R-squared (R^2) value becomes 1, leading to a situation where VIF, calculated as $1/(1 - R^2)$ approaches infinity.When VIF is infinite, it indicates that the model cannot estimate the regression coefficients properly due to this perfect correlation. A high VIF value generally suggests multicollinearity which can distort the regression results. To address this issue, one needs to drop one of the correlated variables from the dataset to resolve the perfect multicollinearity. This step helps in reducing the VIF to a manageable level, ensuring more reliable coefficient estimates in the regression model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

   Answer:
   A Q-Q (Quantile-Quantile) plot is a graphical technique for determining if two dataset come from populations with common distribution or not. Q-Q plot is generally used to determine if the dataset is normally distributed or if it follows some other known distribution.

   Use of Q-Q plot:
   In a Q-Q plot, the quantiles of one dataset are plotted against the quantiles of another dataset. A quantile is a value below which a certain percentage of data points fall. For example, the 0.3 (or 30%) quantile is the value below which 30% of the data points lie. If the two datasets come from the same distribution, the points on the Q-Q plot will fall along the 45-degree reference line. Deviations from this line indicate differences between the distributions of the two datasets.

   Importance of Q-Q plot in Linear Regression:
   In linear regression, it is important to check if the residuals (errors) follow a normal distribution. A Q-Q plot of the residuals can help assess this. If the residuals are normally distributed, the points on the Q-Q plot will lie along the reference line. This validation is crucial because many inferential statistics in regression analysis, such as confidence intervals and hypothesis tests, assume normality of residuals.