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)**

   Ans:

   I have analyse categorical variable using the boxplot and pair plot. Below are the few points we can infer from the visualization

   1) Season 3(fall) has highest demand among all the seasons

   2) Year 1(2019) has more demand then the year 0(2018)

   3) Demand increase as month increases till September after demand decreases.

   4) Demand increase when day is holiday or weekend.

   5) Clear weather has more demand.

   6) weekday we don't have clear idea about this.

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

   Ans:

   The use of drop_first=True is important in creating dummy variables to avoid multicollinearity and the creation of extra columns. By setting drop_first=True, we can reduce the number of columns created when creating dummy variables by removing the first level of the categorical variable. This helps in reducing the correlations among the dummy variables.

   For example, if we have a categorical variable with three levels (A, B, and C) and we create dummy variables without dropping the first level, we will end up with three dummy variables. However, we can infer the third level C from the absence of A and B, so we don't need a separate dummy variable for C.

   Therefore, using drop_first=True will result in only two dummy variables being created (representing B and C), which is sufficient to represent all three levels of the categorical variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

   with the target variable? (1 mark)

   Ans:

   'temp' and 'atemp' 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)

   Ans:

   I have checked the assumptions of the Linear Regression Model and ensured that they have been met. These assumptions include:

   - Normality of error terms:
     - The error terms should be normally distributed in order to have a well-behaved model.
   - Multicollinearity check:
     - The variables should not exhibit significant multicollinearity, or high correlation with each other.
   - Linear relationship validation:
     - There should be a visible linear relationship among the variables in the model.
   - Homoscedasticity:
     - There should be no visible pattern in the residual values, meaning that the variance of the residuals should be constant.
   - Independence of residuals:
     - There should be no auto correlation among the residuals, meaning that the residuals 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?

   Ans:

   The regression analysis of the bike hire numbers has yielded the following coefficient values:

   1) Feeling Temperature (atemp): A coefficient value of 0.5779 suggests that there is a positive relationship between atemp and bike hire numbers, indicating that a unit increase in the temperature variable results in a corresponding increase in bike hire numbers by 0.5779 units.
   2) Weather Situation 3 (weathersit_3): A coefficient value of -0.2821 suggests that there is a negative relationship between weathersit_3 and bike hire numbers, indicating that a unit increase in the weathersit_3 variable leads to a decrease in bike hire numbers by 0.2821 units, with respect to weathersit_1.
   3) Year (yr): A coefficient value of 0.2341 suggests that there is a positive relationship between yr and bike hire numbers, indicating that a unit increase in the year variable leads to an increase in bike hire numbers by 0.2341 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans:

   Linear regression is a statistical model used to analyze the linear association between a dependent variable and a set of independent variables. A linear relationship between the variables indicates that as the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes proportionally (increases or decreases).

   The relationship can be expressed mathematically using the following equation:

   Y = mX + c,

   where Y is the dependent variable being predicted,

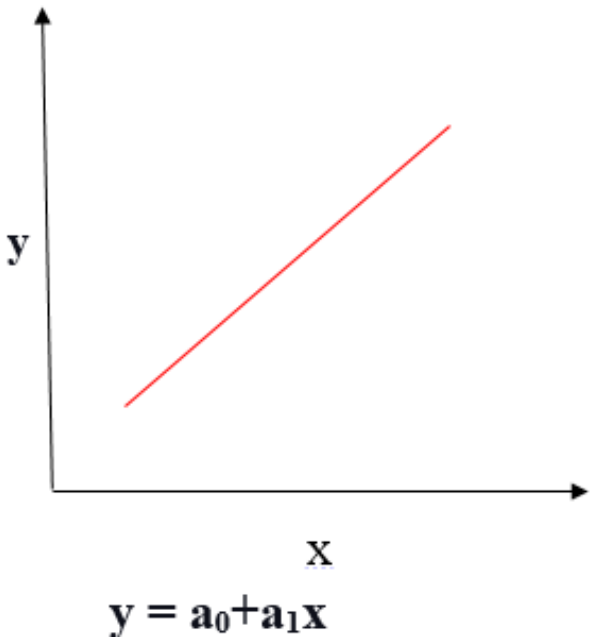
   X is the independent variable being used to make predictions,

   m is the slope of the regression line that represents the effect of X on Y,

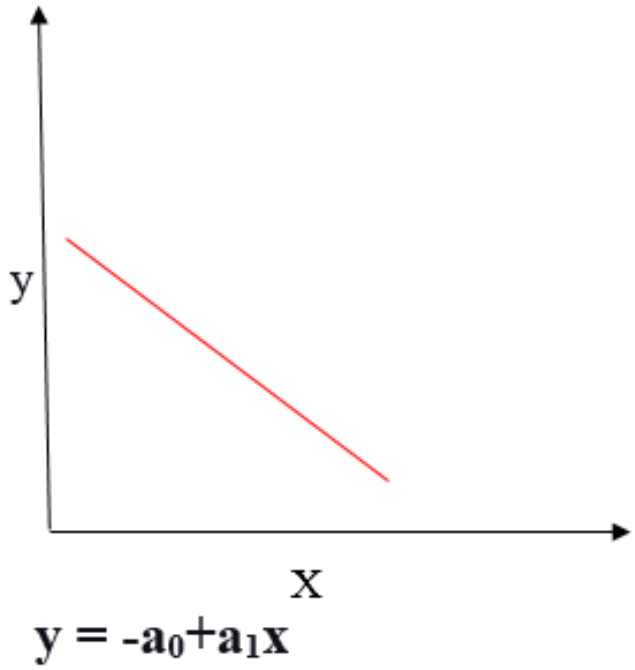
   and c is a constant known as the Y-intercept. When X is equal to 0, Y would be equal to c.

   Additionally, the linear relationship between variables can be either positive or negative, as described below −

   1. Positive Linear Relationship: A linear relationship is considered positive if both the independent and dependent variables increase together. This can be visualized using the following graph:



2. Negative linear relationship: A linear relationship is called negative if the independent variable increases and the dependent variable decreases. This can be visualized using the following graph:



Linear regression models rely on certain assumptions about the dataset, including:

1. Multi-collinearity: The model assumes that there is little or no multi-collinearity present in the data. Multi-collinearity occurs when the independent variables or features are interdependent.
2. Auto-correlation: The model also assumes that there is little or no auto-correlation present in the data. Auto-correlation occurs when there is a relationship between the residual errors.
3. Linear relationship: Linear regression models assume that the relationship between the response and feature variables is linear.
4. Normality of error terms: The error terms should be normally distributed.
5. Homoscedasticity: The residual values should not exhibit any visible pattern.

2. Explain the Anscombe's quartet in detail.

Ans:

Francis Anscombe developed a dataset known as Anscombe's Quartet. It consists of four datasets, each containing eleven (x, y) pairs. What is striking about these datasets is that they share the same descriptive statistics. However, when graphed, each dataset tells a completely different story, despite their similar summary statistics. The mean of x is 9 and the mean of y is 7.50 for each dataset, and similarly, the variance of x is 11 and the variance of y is 4.13 for each dataset. Additionally, the correlation coefficient between x and y is 0.816 for each dataset. Although these four datasets show the same regression lines, plotting them on an x/y coordinate plane reveals that each dataset is conveying a distinct narrative.

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Across the groups, the mean of x is 9 and the mean of y is 7.50 for each dataset, with the variance of x being 11 and the variance of y being 4.13 for each dataset. Additionally, the correlation coefficient (which indicates the strength of the relationship between two variables) between x and y is 0.816 for each dataset. However, when we plot these four datasets on an x/y coordinate plane, we can observe that they share the same regression lines, yet each dataset conveys a unique story.

3. What is Pearson’s R?

   Ans:

   Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to 1.
   A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable increases proportionally. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases proportionally. A value of 0 indicates no correlation, meaning that there is no linear relationship between the two variables.
   Pearson's R is commonly used in various fields such as finance, social sciences, and engineering, to examine the degree of association between two variables. It is particularly useful in regression analysis, where it helps to identify the strength of the relationship between the independent and dependent variables.

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

   Ans:

   Feature Scaling is a crucial data pre-processing technique used to normalize independent features in a fixed range. It helps to manage the issue of highly varying magnitudes or values

or units present in the data. If feature scaling is not applied, machine learning algorithms tend to prioritize higher values and neglect smaller values, irrespective of the unit of the values. For example, suppose an algorithm doesn't use feature scaling. In that case, it may consider the value 3000 meters to be greater than 5 km, which is incorrect. In such cases, the algorithm may provide wrong predictions. Therefore, Feature Scaling is used to bring all values to the same magnitudes, allowing us to overcome this issue.

Normalisation	Standardisation
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)

   Ans:

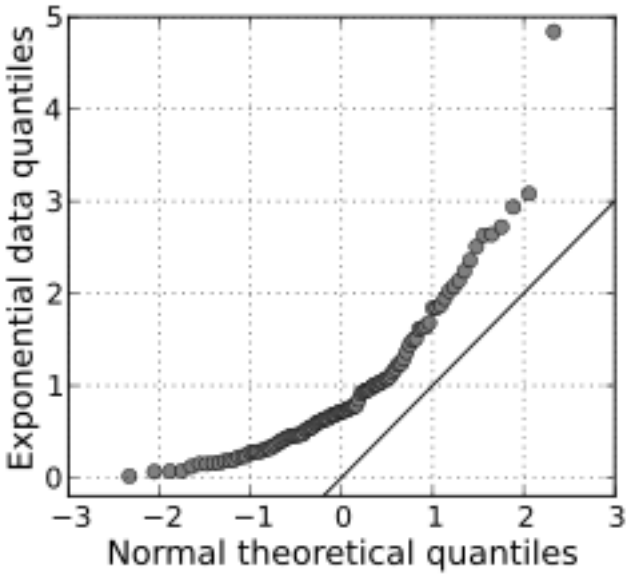
   The Variance Inflation Factor (VIF) is a measure of the correlation between the predictor variables in a regression analysis. A VIF value of infinity indicates that there is perfect correlation between two independent variables. On the other hand, a large VIF value suggests that there is a significant correlation between the variables, which can lead to an inflated variance of the model coefficient.
   For instance, a VIF value of 4 indicates that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
   In case of perfect correlation between two independent variables, the value of R-squared (R2) becomes 1, which leads to a division by zero error of 1/(1-R2) resulting in infinity. To address this, we need to remove one of the variables from the dataset that is causing the perfect multicollinearity.

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

   (3 marks)

   Quantile-Quantile plots (Q-Q plots) are diagrams that display two quantiles against each other. A quantile is a fraction where a specific proportion of values fall below that fraction. For instance, the median is a quantile where 50% of the data falls below that point and 50% lies above it. The main objective of Q-Q plots is to identify whether two data sets have a similar

distribution. A 45-degree line is drawn on the Q-Q plot, and if the two data sets have a common distribution, the points will fall on that reference line.



When comparing two distributions, if they are similar, the points on the Q-Q plot will generally fall on or close to the line y = x. However, if the distributions have a linear relationship, the points on the Q-Q plot may fall on a line other than y = x. Additionally, Q-Q plots can be used to estimate parameters in a location-scale family of distributions using a graphical approach. Overall, Q-Q plots provide a visual representation of the similarities and differences in properties such as location, scale, and skewness between two distributions.