DSP Kumar

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: Based on the analysis of the categorical variables, the following inferences can be drawn:
   - **Seasonal Demand:** The fall season (season 3) sees the highest demand for bike rentals.
   - **Yearly Increase:** There is a noticeable growth in bike rental demand in the following year.
   - **Monthly Trend:** Demand consistently increases each month until June reaches its peak in September, and then declines after September.
   - **Holiday Effect:** Bike rental demand decreases on holidays.
   - **Weekday Impact:** The influence of weekdays on demand is unclear.
   - **Weather Conditions:** Clear weather conditions have the highest bike rental demand.

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

   Answer: Using drop_first=True is important during dummy variable creation because it helps eliminate the extra column that is being created during the creation of the dummy variable. This reduction prevents increased correlations among dummy variables. If we do not drop one of the dummy variables derived from a categorical variable, it becomes redundant in the dataset. This redundancy, coupled with the constant variable can lead to multicollinearity issues. Additionally, it simplifies the model by reducing the number of predictors.

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 and has a strong linear relationship with the target variable "cnt".

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   Answer: I validated based on the following assumptions:
   - The error terms follow a normal distribution with a mean of 0.
   - There is no visible pattern in the error terms.
   - The linearity was checked.
   - Multicollinearity was checked using Variance Inflation Factors (VIFs).
   - Overfitting was evaluated by comparing the R^2 value with the Adjusted R^2 value.

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: The top three features significantly contributing to explaining the demand for shared bikes are "holiday", "temp", and "hum".

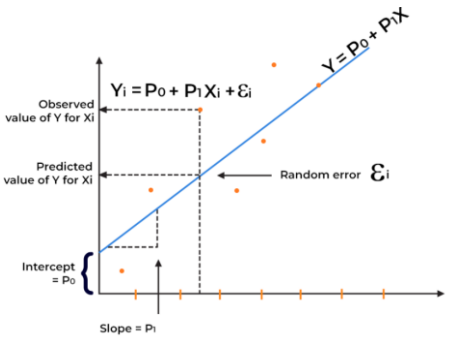
General Subjective Questions:

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

Answer: Linear Regression is a supervised learning algorithm used in machine learning. The relationship between a dependent variable (commonly represented as Y) and one or more independent variables (typically represented as X) can be modeled mathematically using linear regression. It is predicated on the idea that Y can be roughly represented as a linear function of X plus an error term. It determines the best-fitting straight line for the given data to establish the optimal linear relationship between the independent and dependent variables, typically using the Sum of Squared Residuals method.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Where β0 is the intercept, $\beta 1$, $\beta 2$, … , βp are the coefficients of the independent variables, and ϵ represents the error term.



When there is only one independent variable, it is called Simple Linear Regression, and when there are multiple independent variables, it is referred to as Multiple Linear Regression.

Simple Linear Regression:
This is the most basic form of linear regression, involving just one independent variable and one dependent variable. The equation for simple linear regression is:

$$y=\beta_0+\beta_1 X$$

Where: Y is the dependent variable, X is the independent variable, β0 is the intercept and β1 is the slope.

Multiple Linear Regression:
There are multiple independent variables and one dependent variable in this. For multiple linear regression, the equation is:

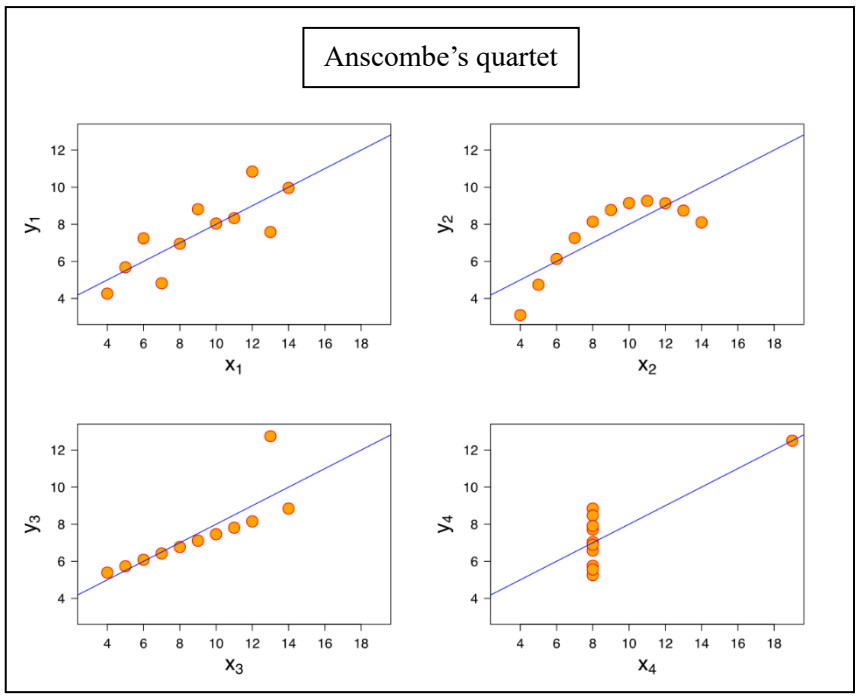
$$y=\beta 0 + \beta 1 X + \beta 2 X + \ldots\ldots \beta n X$$

Where: Y is the dependent variable X1, X2, …, Xp are the independent variables β0 is the intercept β1, β2, …, βn are the slopes.

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

Answer: Anscombe's quartet consists of four small datasets that, although being visually depicted in a different way, contain almost identical simple descriptive statistics (such as

mean, variance, correlation, and regression line). This demonstrates the significance of data visualization in interpreting information and identifying trends or patterns.
Anscombe's quartet demonstrates the importance of exploratory data analysis by showing how datasets with identical summary statistics can exhibit vastly different relationships when graphically visualized.



The four datasets of Anscombe's quartet:

- Dataset 1: The first set of data matches the linear regression model fairly well.
- Dataset 2: Due to its non-linear nature, the data cannot be fitted using a linear regression model.
- Dataset 3: Demonstrates the outliers in the dataset that the linear regression model is unable to manage.
- Dataset 4: Presents the outliers in the data set that the linear regression model is unable to handle.

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

Answer: Pearson's R which is commonly known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables which are denoted as X and Y. It ranges from -1 to +1 where +1 indicates a perfect positive linear correlation, 0 indicates no linear correlation that is the variables are independent of each other and -1 indicates perfect negative linear correlation.

To obtain a standardized measure of linear dependency, one can calculate Pearson's *R* R by dividing the covariance of two variables by the product of their standard deviations.

Formula:

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

Where,

N = Number of pairs of scores

$\sqrt{x}$= Sum of x Scores

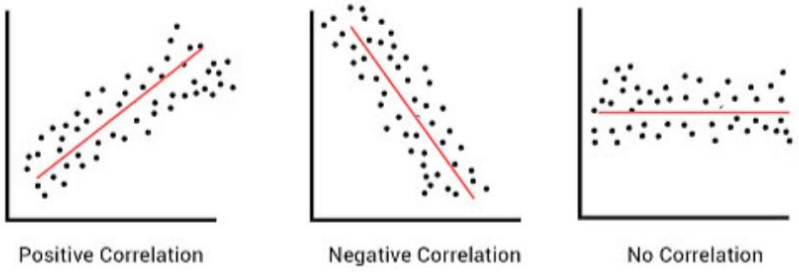
$\sqrt{y}$= Sum of y scores

$\sqrt{xy}$= sum of the products of paired scores

$\sqrt{x^2}$= sum of the squared x scores

$\sqrt{y^2}$= sum of the squared y scores

The Pearson coefficient indicates a correlation between variables and does not imply causation.



9. 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 data into a standardized range which helps to improve the performance of the machine learning algorithms. Without scaling, regression algorithms may incorrectly prioritize larger numerical values over smaller ones, affecting the accuracy of predictions by not treating all features equally. Scaling affects the coefficients and not the other parameters.

Scaling is done in order to:

- Make sure every attribute has an equal impact on model fitting and prediction.
- To increase gradient descent methods' convergence.
- To prevent domination of certain features that are in large numbers.
- Boost the efficiency of models like SVMs and K-nearest neighbors that are sensitive to the size of the input data.

Scaling can be achieved through different methods, such as normalization and standardization: normalization, which scales a variable to a range between 0 and 1, and standardization, which transforms data to have a mean of 0 and a standard deviation of 1.

The formula for **normalized scaling:**

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

The formula for **standardized scaling:**

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

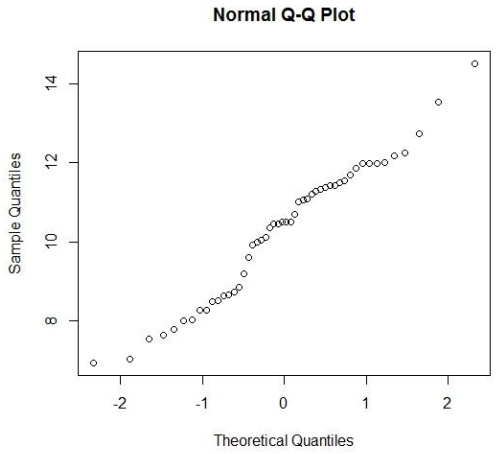
10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: When one or more independent variables in a regression model are perfectly collinear with one another, the Variance Inflation Factor (VIF) is likely to have an infinite value. That is when the R square value is equal to 1 then VIF's value becomes infinite. One variable can be precisely predicted from another due to perfect multicollinearity, which results in an infinite or undefinable VIF value. In practice, this happens when predictors have a linear connection with one another, i.e., when one variable in the dataset is a constant multiple of another.

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

Answer: A graphical technique called a Q-Q plot or Quantile-Quantile plot is used to determine if a dataset belongs to a specific theoretical distribution, such as the normal distribution. It also helps to determine whether two datasets come from populations with a common distribution by visually comparing their quantiles.

The Q-Q plot indicates similarity or consistency between the datasets or distributional assumptions if the points roughly follow a straight line. Below an example of the normal Q-Q plot is shown.



Use of Q-Q Plot:

- Q-Q plots are useful for comparing empirical data distributions against theoretical distributions, even with sample sizes.

- **Distribution Comparison:** It provides a visual comparison between a dataset's distribution and a theoretical distribution.
- **Normality Check:** In linear regression, Q-Q plots are crucial for verifying the assumption of normality of residuals (error terms).

Importance of Q-Q Plot:

- Q-Q plots do not require equal sample sizes.
- They can simultaneously test various distributional aspects such as shifts in location, scale, changes in symmetry, and the presence of outliers.
- Q-Q plots provide deeper insights into distributional differences compared to analytical methods