Assignment-based Subjective Answers

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   (3 marks)

   The categorical variables in the bike sharing dataset were season, yr, holiday, weekday, working day, and weathersit and mnth. The box plot is used for data visualization.

   These variables had the following effect on our dependent variable:-

   - Season – It is observed that the spring season had the least value of cnt, whereas fall had the maximum value of cnt. Summer and winter had intermediate values of cnt.
   - Yr - The number of rentals in 2019 was more than 2018
   - Holiday - Rentals are reduced during the holiday period.
   - Working day – The median count of users is constant throughout the week.
   - Weathersit - There are no users when the season has heavy rain/ snow indicating that the weather is extremely unfavourable. The highest count was seen when the weathersit was 'Clear, Partly Cloudy'.
   - Weekday - The count of rentals is almost even throughout the week.
   - Mnth - September saw the highest number of rentals while December had the lowest number. The weather situation in December is usually heavy snow due to which the rentals might have dropped.

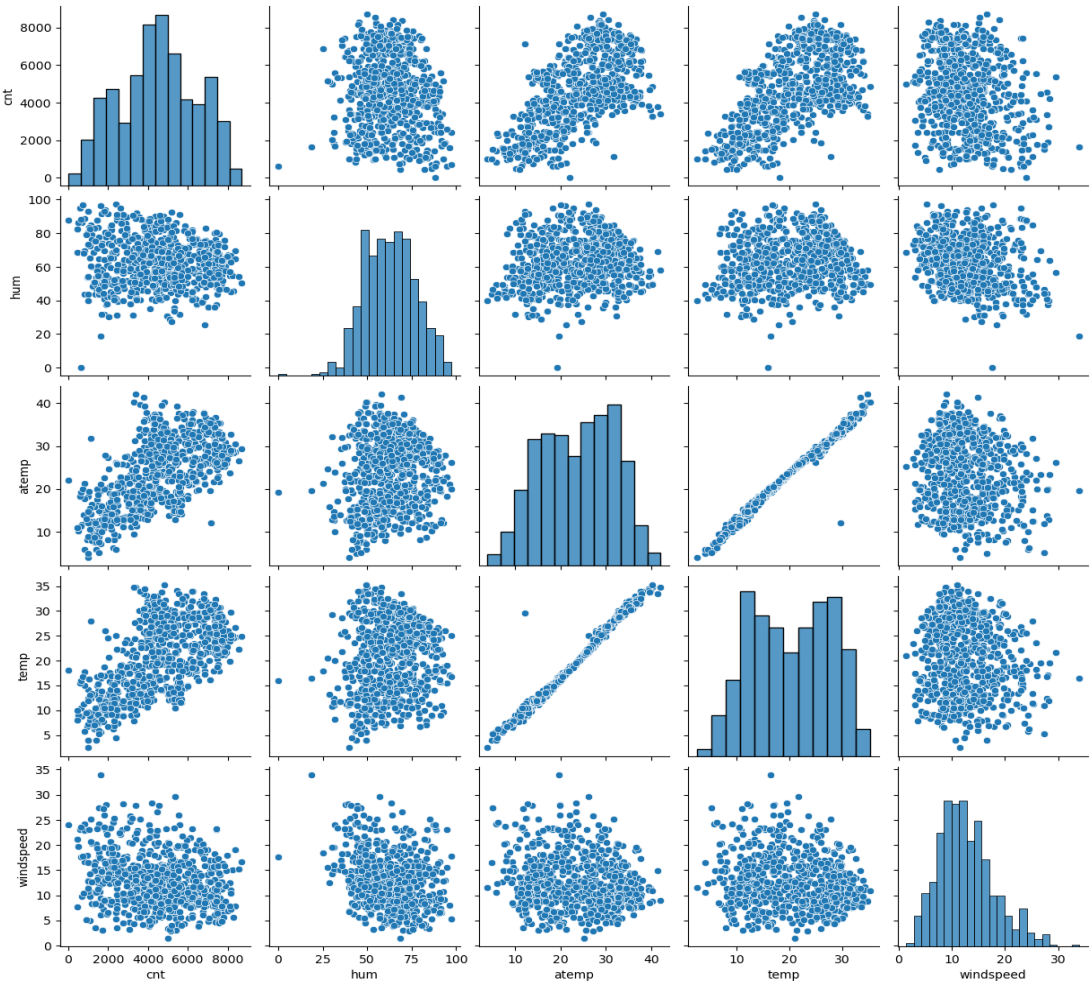
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   A dummy variable is a numeric variable that encodes categorical information. If we don't drop the first column then the dummy variables will be correlated. This may affect some models adversely. It helps in reducing the extra column created during dummy variable creation. It is also important in order to achieve k-1 dummy variables as it can be used to delete extra columns while creating variables.

   For example: We have three variables: furnished, semi-furnished, and un-furnished. We can only take 2 variables as furnished will be 1-0, semi furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. Hence, we can remove it.

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

From the pair plot below we can see that, “temp” and “atemp” are the two numerical variables which are highly correlated 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)

The following tests were done to validate the assumptions of linear regression:

a. One way to assess linearity is to plot the dependent variable against each independent variable. We visualized the numeric variables using a pair plot to see if the variables were linearly related or not.

b. Secondly, residual distribution should follow a normal distribution and be centered around 0 (mean = 0). We validated by plotting a distplot of residuals and saw if residuals were following a normal distribution.

c. Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get a quantitative idea of how much the feature variables are correlated with each other.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of shared bikes?
   (2 marks)
   The top 3 features are:

   a. temp - coefficient: 0.472823

   b. yr - coefficient: 0.234361

   c. weathersit_Light Snow & Rain - coefficient: -0.291727

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   a. Linear regression is one of the most popular Machine Learning algorithms.
   b. It is a powerful algorithm in data science. Since the Linear Regression algorithm represents a linear relationship between a dependent and one or more independent variables, it is known as Linear Regression.
   c. It is easier to implement and interpret.
   d. Linear regression is based on the popular equation "y = mx + c".
   e. Regression is broadly divided into simple linear regression and multiple linear regression.
      (i) Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
      (ii) Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.
   f. Applications include Sales Forecasting and Finance Applications to Predict Stock prices and investment evaluation.

2. Explain the Anscombe's quartet in detail. (3 marks)
   (i) Anscombe's Quartet was developed by statistician Francis Anscombe. Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines.

(ii) Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

(iii) Anscombe's quartet helps us to understand the importance of data visualization.

(iv) It was developed to emphasize both the importance of graphing data and the effect of outliers and other influential observations on statistical properties.

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

The Pearson correlation coefficient (***r***) is the most common way of measuring a linear correlation. The Pearson correlation coefficient is a descriptive statistic, which means that it summarizes the characteristics of a dataset. It measures the strength between different variables and their relationships. Whenever a statistical test is conducted, it is advised to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

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

Feature engineering is a critical step in building effective machine learning. Scaling is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. it also aids in accelerating algorithmic calculations.

(i) Normalization or Min-Max Scaling is used to transform features to be on a similar scale. Normalization typically means rescaling the values into a range of [0,1]. The new point is calculated as:

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

(ii) Standardization (or Z-score normalization) is the process of rescaling the features. The formula for standardization is:

$$x\_standardized = (x\text{-}mean) / standard\_deviation$$

Standardization is useful when the distribution of the features is not Gaussian (i.e., normally distributed).

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

(3 marks)

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is a perfect correlation, then VIF = infinity. The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite. It denotes perfect correlation in variables. The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required.

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

(3 marks)

(i) Q-Q plots are also known as Quantile-Quantile plots. It plots the quantiles of a sample distribution against the quantiles of a theoretical distribution

(ii) It helps to determine if two data sets come from populations with a common distribution.

(iii) The advantages of the q-q plot are: The sample sizes do not need to be equal and many distributional aspects can be simultaneously tested.

(iv) In linear regression, it helps us to compare the sample distribution of the variable at hand against any other possible distributions graphically.