Assignment-based Subjective Questions

1. Here are some of the inferences on the analysis of the categorical variables and their effect on the dependent variable.

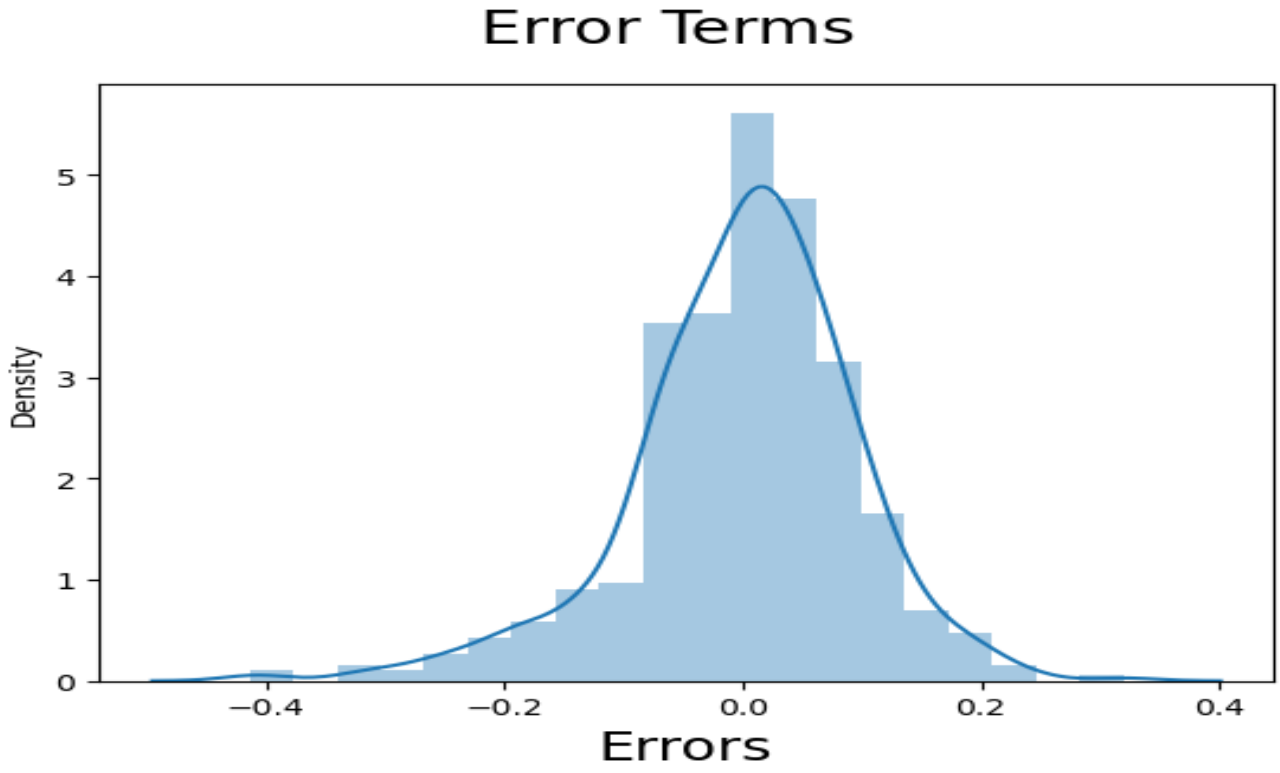
   a. The season of Fall has the highest median followed by summer as they have the best weather conditions.
   b. The median bike rentals have increased in the year 2019 compared to the year 2018. This may be due to the people getting conscious about the environment.
   c. The bike rentals are more on non-holiday days compared to holiday. This indicates that people prefer to spend time at home during the holidays.
   d. The months of Fall-June to October have a higher median value.
   e. The overall median for the weekdays and working-days are the same. 6. The Clear weather situation has the highest median while the weather situation of Light snow has the least. The count of bike sharing is Zero for the weather situation - 4 'Heavy Rain + Ice Fillets + Thunderstorm + Mist, Snow + Fog'.

2. It is important to use drop first=True as it helps in reducing the extra column created during dummy variable creation. It helps to reduce the correlations created among dummy variables. Example: Let's say we have 3 types of values in a categorical column and we want to create dummy variable for that column. If one variable is not furnished and not semi-furnished, then it is obvious that it is unfurnished. So, we do not need a 3rd variable to identify the category of unfurnished.
   Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. The numerical variable 'atemp' has the highest correlation with the target variable 'cnt' with a value of '0.65' followed by 'temp' with a value of 0.64".

4. We validate the assumptions of the Linear Regression by plotting a distplot of the residuals and analysing it to see if it is a normal distribution or not and if it has a mean = 0. The diagram below

shows that it is normally distributed with mean = 0.



5. The Following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:
   a. temp(Temperature) - A coefficient value of '0.5480' indicates that a unit increase in temp variable, increases the bike hire numbers by 0.5480 units.
   b. Light Snow(weathersit) - A coefficient value of -0.2838' indicates that, a unit increase of this variable, decreases the bike hire numbers by -0.2838 units.
   c. Yr(Year)-A coefficient value of '0.2328' indicates that, a unit increase of this variable, increase the bike hire numbers by 0.2328 units.

General Subjective Questions

1. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

   Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.

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.

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

3. The Pearson correlation coefficient (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. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association.

4. Scaling is a common dental procedure for patients with gum disease.

   This is a type of dental cleaning that reaches below the gumline to remove plaque buildup. The process of scaling and root planning the teeth is often referred to as a deep cleaning.

   Standardization is divided by the standard deviation after the mean has been subtracted. Data is transformed into a range between 0

and 1 by normalization, which involves dividing a vector by its length.

5. If there is perfect correlation, 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.

6. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

   The power of Q-Q plots lies in their ability to summarize any distribution visually. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.