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 – Categorical variables (specially temp, year, season) have significant impact on dependent variable. The impact is both positive and negative for different variables.

   Following are the observations with respect to effect of categorical variable on dependent variable:

   a.) The demand bike increased in the year 2019 when compared with year 2018 – Shows growth.
   b.) During Jun to Sep, bike demand is high. In Jan, it is the lowest.
   c.) The demand of bike is almost similar throughout the weekdays.
   d.) The bike demand is high when the weather is clear or when there are few clouds. However demand is less in case of light snow or light rainfall. Data for other type of weather is not there in the dataset.
   e.) There is no significant change in bike demand for working day and non-working day.

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

   Answer – It helps in reducing the extra column created during dummy value creation. And hence it reduces the correlations created among dummy variables.

   Ex - If we have say 3 types of values in categorical column and we want to create dummy variable for that column. If one variable is not 'furnished' and also not 'semi_furnished', then It is for sure unfurnished. So we do not need 3rd variable to identify the unfurnished.

	Furnished	Semi-Furnished
Furnished	1	0
Semi-Furnished	0	1
Nor Furnished	0	0

   Here with only 2 dummy variables we can represent all 3. This is achieved using **drop_first=True.**

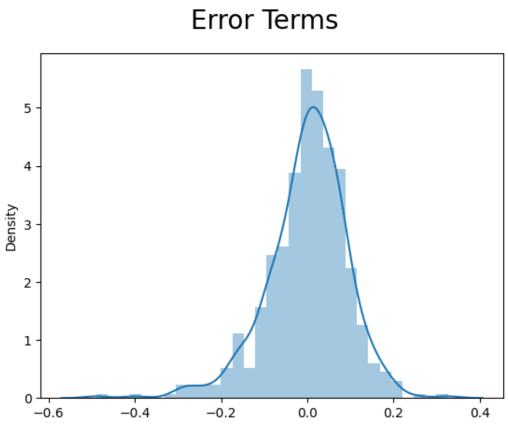
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Answer – 'temp' and 'atemp' have the highest correlation with the target variable 'cnt' (0.63).

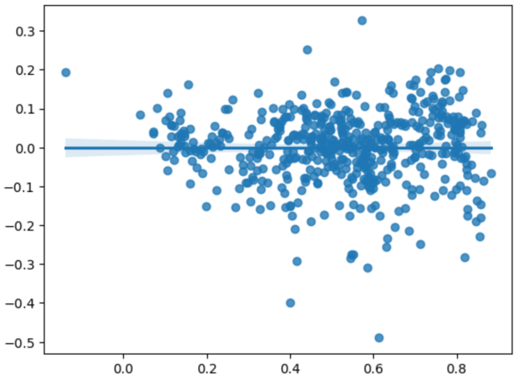
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer –

1.) Calculated ‘y_train_pred’ and compared it with the original data to derive residual. Plotted the error term and found out they are normally distributed



2.) Plotted scatter plot of residual values vs predicted values is to check for patterns homoscedasticity. There were no patterns.



3.) Calculated VIF to make sure that there is no collinearity.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer *– Top 3 features contributing significantly towards explaining the demand of shared bikes:*

1. ***temp*** *- Positive Correlation (Coeff = 0.3546)*
2. ***yr*** *- Positive Correlation (Coeff = 0.2406)*
3. ***Light_Snow_Or_Rain*** *- Negative Correlation (Coeff = - 0.2949)*

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer - Linear Regression is a machine learning algorithm based on supervised learning. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value based on a given independent variable(s). So, this regression technique finds out a linear relationship between input and output.

   There are 2 types of Linear Regression :

   1.) **Simple Linear Regression** - Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.
   2.) **Multiple Linear Regression** - Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable.

   Assumptions of Linear Regression

   - Linear relationship between the features and target
   - Small or no multicollinearity between the features
   - Homoscedasticity - there should be no clear pattern distribution of data in the scatter plot.
   - Normal distribution of error terms
   - No autocorrelations

2. Explain the Anscombe's quartet in detail.

   Answer - Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

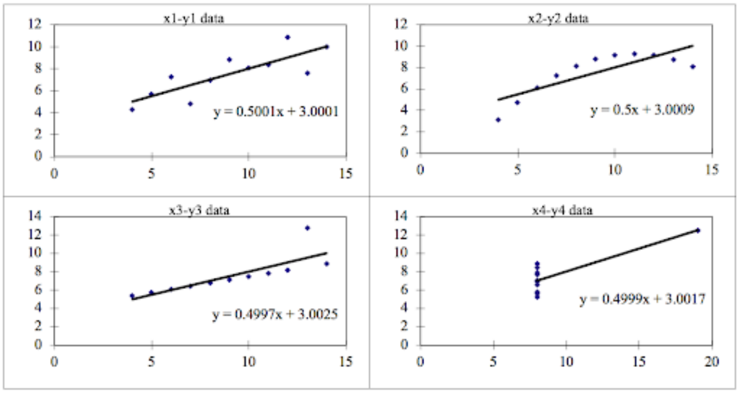
   ANSCOMBE'S QUARTET FOUR DATASETS

   Data Set 1: fits the linear regression model pretty well.

   Data Set 2: cannot fit the linear regression model because the data is non-linear.

   Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

   Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.



Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

Answer - Pearson's r is a numerical summary of the strength of the linear association between the variables. It is a statistic that measures the linear correlation between two variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

Answer - It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why? - The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range. If scaling is not done, the algorithm will only consider magnitude and not units, which will result in inaccurate modelling. We must scale all the variables to the same degree of magnitude in order to resolve this problem. The t-statistic, F-statistic, p-values, R-squared, etc. are unaffected by scaling, which is significant because they are all dependent on the coefficients.

There are 2 types of scaling:

1.) **Min-Max Scaling / Normalization** – It basically brings all the data in the range of 0 and 1.

    It helps for data which have outliers.

    x = (x – min(x)) / (max(x) – min(x))

2.) **Standardized Scaling** – It is another scaling technique where the values are centred around the mean with a standard deviation of 1.

    x = (x – mean(x)) / sd(x)

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

   Answer - VIF = infinity basically means there is perfect correlation. In the case of perfect correlation, we get R2 = 1, which lead to 1/(1-R2) infinity. In this case, we need to drop the variable from the dataset which is causing this perfect multicollinearity.

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

   Answer – Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

   This helps in a scenario of linear regression when we receive training and test data separately. In this case we can make use of Q-Q plot to find out if both the data sets are from populations with same distribution.