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?

Categorical variables are playing an important role in predicting the demand for the Bike sharing service. The major variable that has become part of the model are
spring
Light Snow + Rain
Mist + Cloudy

For this variable to become part of the model dummy variable had to be used, so that individual column for each variable is created and then these variables can be used for predicting the demand for bike sharing in future.

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

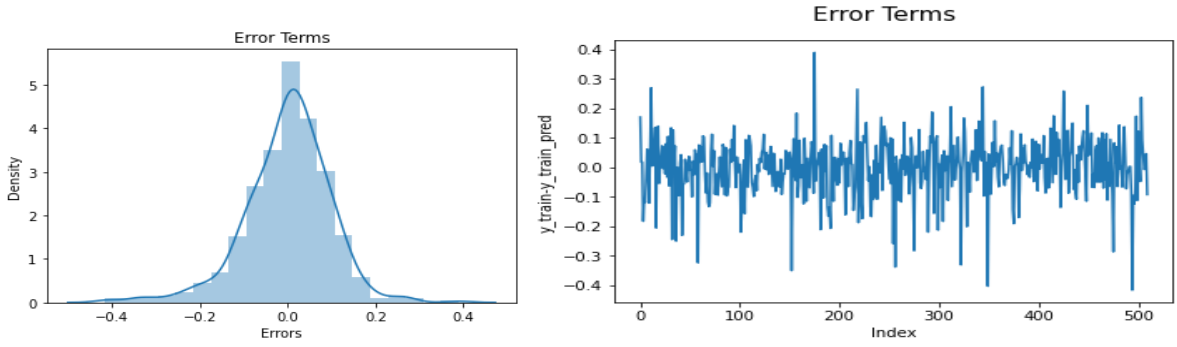
drop_first**bool, default False**

It is part of pd.get_dummies. This is used when we want to have n-1 dummies out of n categorical levels by removing the first level.

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

The Variable which has the highest correlation with the target variable is Temp.

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



The plots are created using y_train – y_train_pred

By using these two plots, assumptions related to Linear Regression were confirmed.

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

Followinr are the THREE variables that are playing a key role

1. Temp
2. Yr
3. Light Snow + Rain

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear Regression is one of the forms of machine learning. It can be developed using single independent variable which is called as Simple Linear Regression.

When multiple variables are to be introduced as used ,then that regression is called as Multiple Linear Regression.

The Linear Regression follows the Line equation, which is

y = mX +c

Where

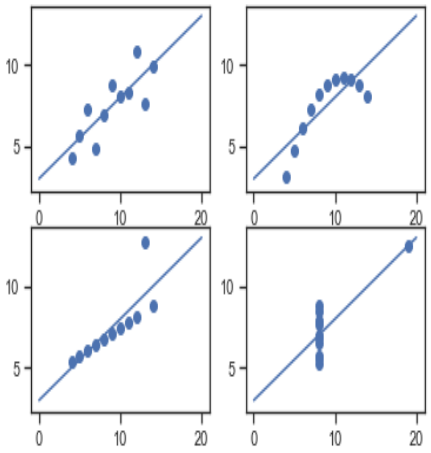
C – intercept
M – slope
X – independent variable
Y – dependent variable.

Based on the data available , a Best Line Fit is created.

2.Explain the Anscombe's quartet in detail.

Anscombe's Quartet is an example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).

Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.

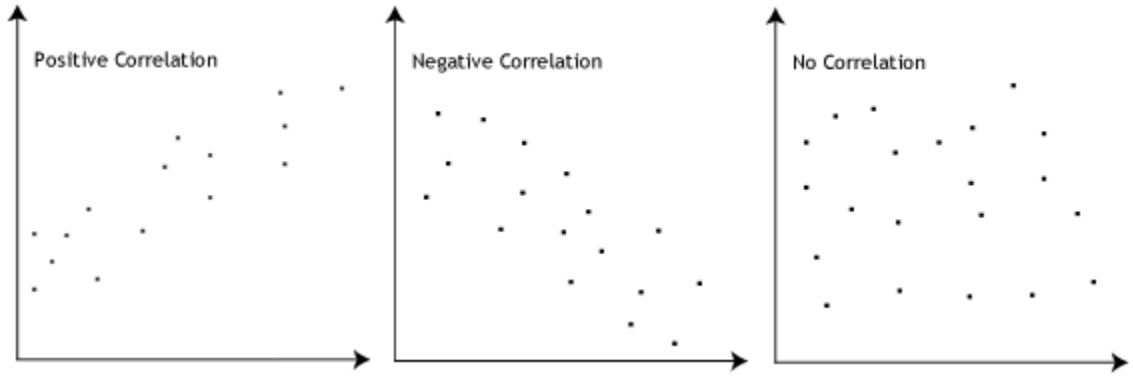
Data-set IV — looks like the value of x remains constant, except for one outlier as well.

3.What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- 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



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is one of the important step in of data preparation which is applied to continuous independent variable to normalize the data within a particular range. Scaling of data helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization:

Also know as Min/Max Scaling, it brings all of the data in the range of 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

VIF is used to check if there are any correlation between two independent variables. If the value of VIF = Infinity, it means that two independent variables are having perfect correlation.

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity

The infinity problem can be solved by dropping one of the variables from the dataset which is causing this perfect multicollinearity.

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

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. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Importance: -

1. It can be used with sample sizes also.
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
3. It is used to check following scenarios:

   If two data sets —

   i. come from populations with a common distribution

   ii. have common location and scale

   iii. have similar distributional shapes

   iv. have similar tail behaviour