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)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

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

(3 marks)

Answers to Assignment Based Subjective Questions:

1. From the analysis we can infer the below points:

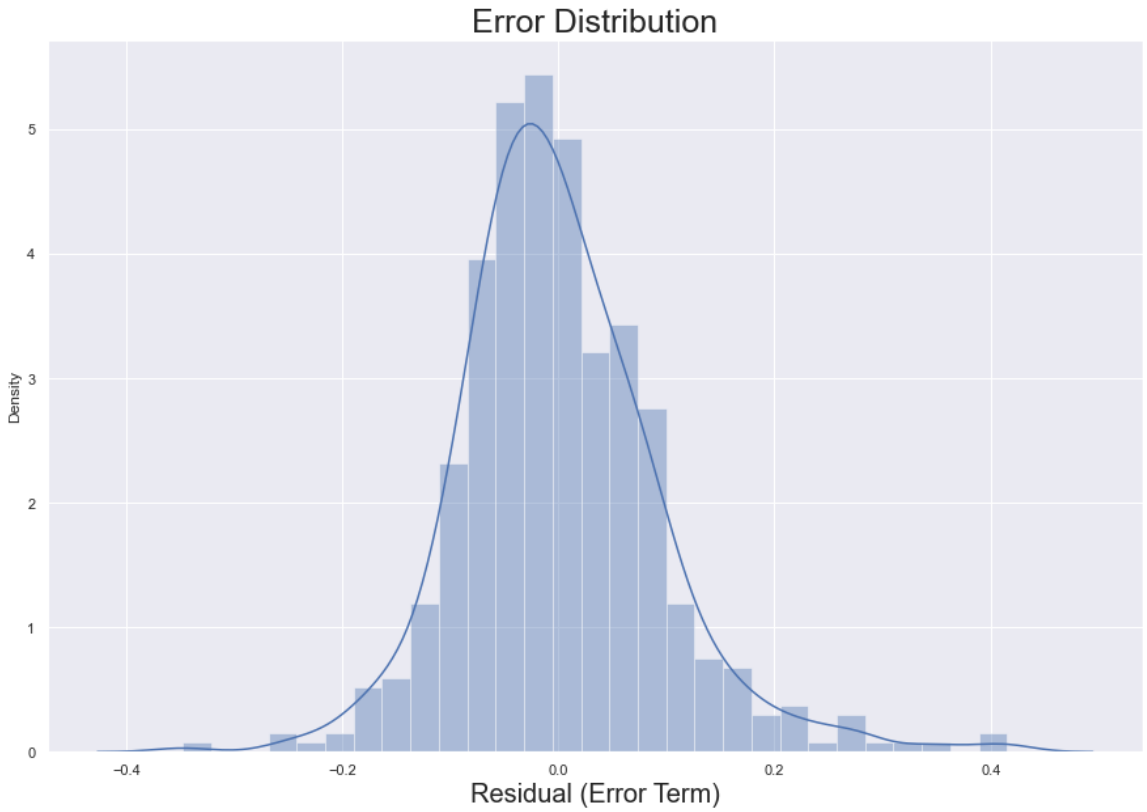
   - Seasons: We could see business was operating similar days in all four seasons.
   - Yr: Number of days operation in both the year are almost same.
   - Month: We could see business was operating similar days in all 12 months.
   - Holiday: Business was operating in 3% days of holiday
   - weekdays: We could see business was operating similar percentage in all weekdays.
   - Working day: Business was operating in 68% in working days and 32% in nonworking days.
   - Weathersit: From the above analysis it is being observed that there is no data for 4th category of Weathersit i.e., Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog. May be the company is not operating on those days or there was no demand of bike.

2. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. From the pair plot we could observe that, `temp` has highest positive correlation with target variable `cnt`

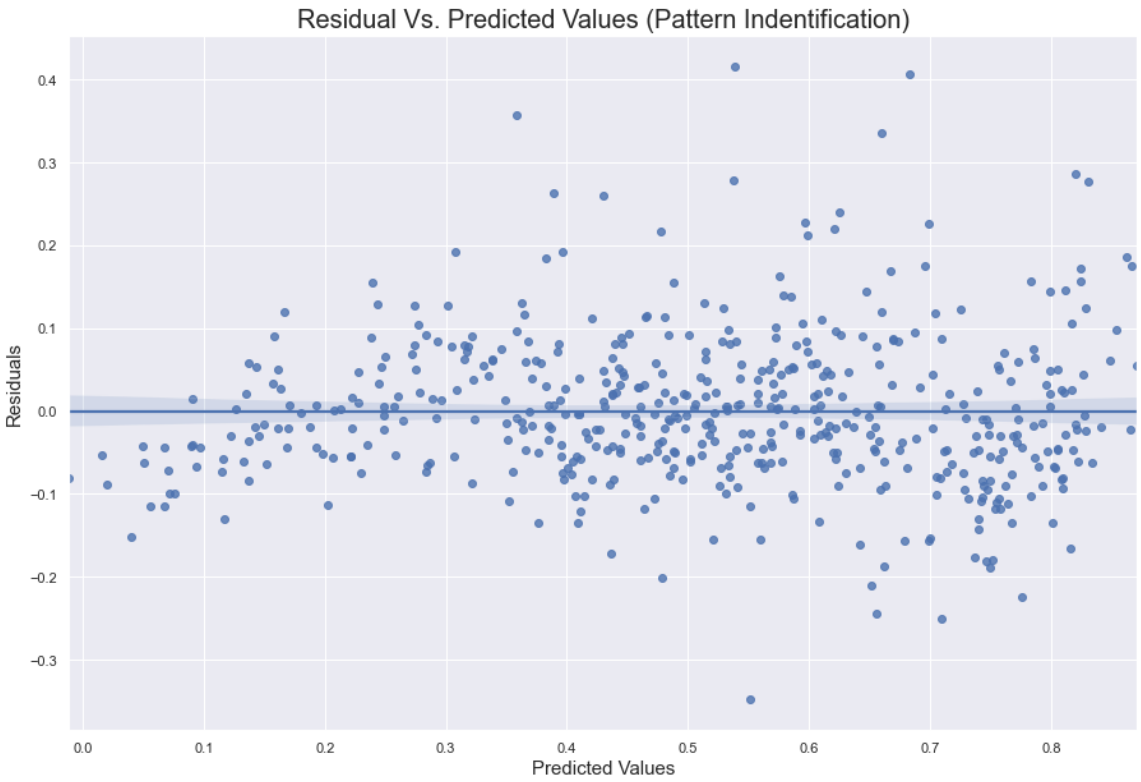
4. Validation of the assumptions of Linear Regression was done using the below proofs:

    - Assumption of Normally Distributed Error Terms



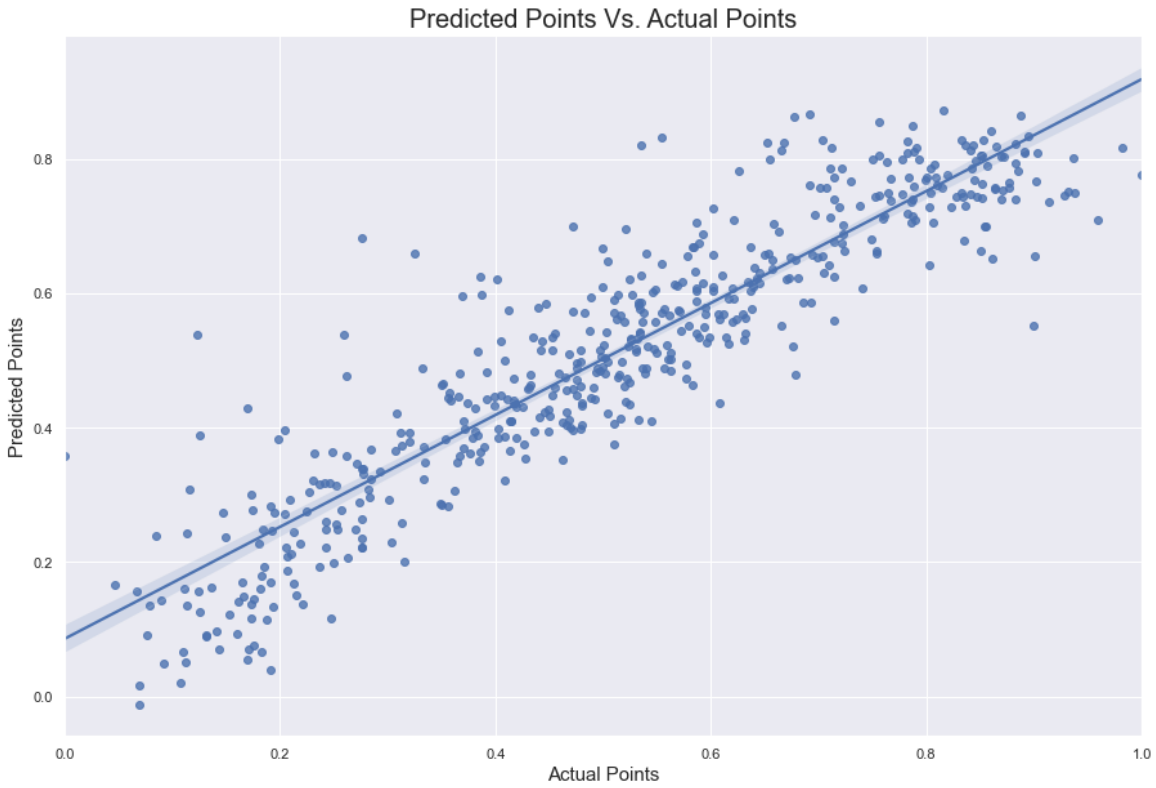
From the above graph, it is evident that Error Distribution is normally distributed across 0, which indicates that our model has handled the assumption of Error Normal Distribution properly.

- Assumption of Error Terms being independent



From the above graph, we see that there is almost no relation between Residual & Predicted Value. This is what we had expected from our model to have no specific pattern.

- Homoscedasticity



From the above graph, we can say that residuals are equal distributed across predicted value.

This means we see equal variance and we do NOT observe high concentration of data points in certain region & low concentration in certain regions.

This proves Homoscedasticity of Error Terms

- Multicorrelation

This assumption is already taken care of while building model by calculating VIF of every predictor. Following is the final VIF value of all the predictors used in the model.

	Features	VIF
0	temp	5.13
1	windspeed	4.61
2	season_spring	2.74
3	season_summer	2.24
4	yr	2.07
5	season_winter	1.77
6	mnth_January	1.61
7	mnth_July	1.59
8	weathersit_mist	1.56
9	mnth_September	1.33
10	weathersit_light	1.08

5. Based on final model top three features contributing significantly towards explaining the demand are:

    1. Temperature
    2. weathersit: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    3. year

Answers to General Subjective Questions:

1. **Linear Regression**

   Linear regression is a machine learning algorithm and is a part of the branch supervised learning.

   Linear regression does the target value prediction based on the independent variables.

   On the basis of independent variables there are two types of linear regression:

   - Simple Linear Regression: Model with only one independent variable
   - Multiple Linear Regression: Model with more than 1 independent variables.

Mathematically a linear regression equation can be written as =

$$y = mx + c$$

where m is the slope of the line

c is the y intercept of the line

x is the independent variable from the dataset

y is the dependent variable from the dataset

Linear Regression can be widely used in the below business areas:

1. Sales target predictions and trends in sales

2. Price Prediction

3.Risk Management

2. **Anscombe's quartet**

Anscombe's quartet means four datasets that are nearly identical in simple statistical properties and still are appearing differently when they are plotted on a graph.

This is used to prove the importance of plotting the data graphically before analyzing it as there could be a possibility that the data looks accurate statically but could be completely different when represented graphically.

Example:

Consider the below 4 datasets for example and now we will be plotting them to view the difference between each of the dataset graphically.

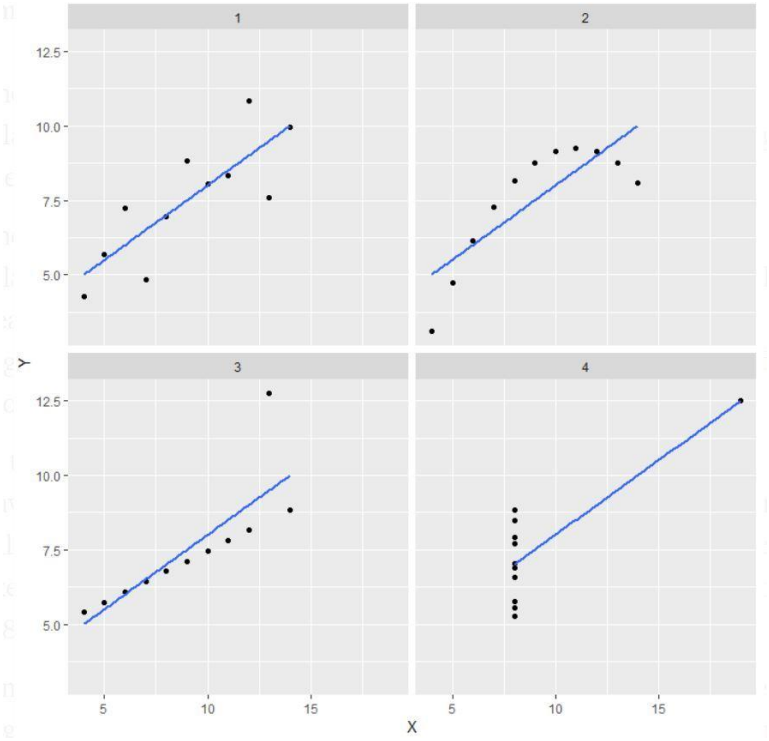
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

If we analyze these data sets statistically, we would get the below results:

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Now we would represent these 4 datasets graphically to infer more details about them.



By looking at the above diagram we can see that all four datasets have different graphs and below are the other observations:

- For graph 1, there is a linear relationship between x and y
- For graph 2, there is a non-linear relationship between x and y
- For graph 3, there is a perfect linear relationship for all the data points except for one outlier which is indicated far away from the actual line.
- For graph 4, this shows that a high leverage point is enough to produce a high correlation coefficient.

3. Pearson's R

- Pearson's R is also known as Pearson's correlation coefficient (PCC).
- It is the covariance of two variables, divided by the product of their standard deviations.
- It is a normalized measurement of the covariance, such that the result always has a value between -1 to 1.

4. Scaling and why it is performed

- Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- If scaling is not done, then a ML algorithm would be weigh towards greater values and ignore the units altogether. Hence, the algorithm would give wrong predictions.
- Hence, to avoid wrong predictions, we use scaling to bring all values to the same magnitudes.

There are two techniques to perform feature scaling:

- Normalized scaling: This technique rescales a feature or observation value with distribution value between 0 and 1.

- Standardized scaling: This is an effective technique which rescales a feature so that it has distribution with 0 mean value and variance equals to 1.

5. VIF value is equal to Infinity

- If there is a perfect correlation, the VIF value is equal to infinity. This shows perfect correlation between two independent variables.
- In case of perfect correlation, we get R2 value is equal to 1 which lead to 1/(1-R2) infinity.
- To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. Q-Q Plot and importance of it in linear regression

- Q-Q plot is referred to as Quantile-Quantile plot.
- This is a graphical technique for determining if two data sets come from populations with a common distribution.
- A Q-Q plot is a plot of the quantile of the first data set against the quantiles of the second data set.
- By Quantile we mean the fraction of points below the given value.
- A 45-degree reference line is also plotted. If two sets come from a population with the same distribution, the points should fall approximately along this reference line.