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)
   - Strong co-relation between season and cnt. seasons 2 and 3 are having higher values
   - May to Sep has demand for bike rentals compared to other months

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

It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

The highest correlation was between Cnt (target) and atemp (numerical variable).

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

When I got R square value of 80% and p values are significant. Then I started checking further on the residual distribution and also have checked VIFs before building residual plot. When I saw my VIFs are under 5% and the residual distribution was normal with mean at 0.

Then I check the train model on the test sample and the R square value of target test and target predicted value came out to be 77%, which is near to the train model value.

Below are the findings to validate the Linear regression model.

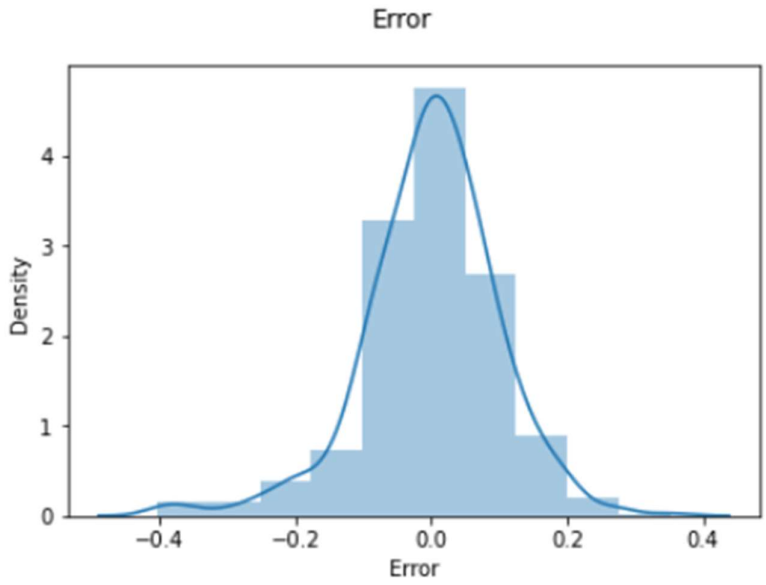
VIF below 5%

	Features	VIF
2	atemp	4.74
3	windspeed	3.39
0	weekday	2.83
1	workingday	2.81
5	weathersit_1	2.68
7	year	2.02
4	season_4	1.27
6	weathersit_3	1.11

- Adjusted r2 score is good relative to the R square and adj r2 score stands at 80.0%
- P(f-stat) is also very less
- VIF values are less than 5%

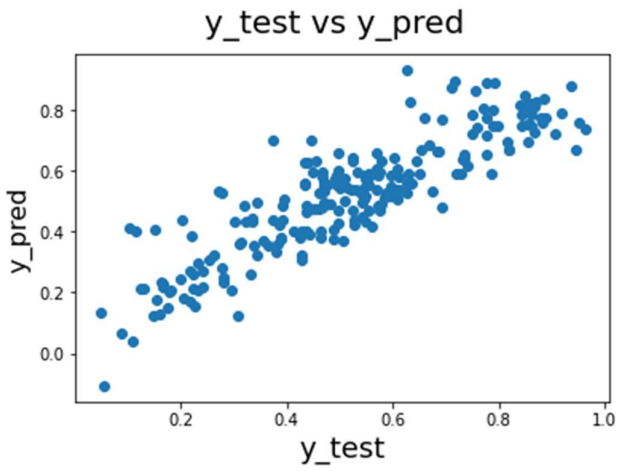
Residual Analysis of the Train data (normal distribution of residual data)

```
Text(0.5, 0, 'Error')
```



Plot of y_test vs y_pred

```
Text(0, 0.5, 'y_pred')
```




```
r2_score(y_test, y_pred_lm2)
```

```
0.7716233805475031
```

•Adjusted r2 score is good relative to the R square and adj r2 score stands at 77.7%

Best fitted equation:

We can see that the equation of our best fitted line is:

cnt = 0.009 x weekday + 0.026 x workingday + 0.621 x atemp - 0.127 x windspeed + 0.100 x season_4 + 0.074 x weathersit_1-0.199 x weathersit_3 + 0.234 x year

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

The top 3 contributing to demand of shared bikes are:

(a) Atemp – contributing the most- feel temperature
(b) Weathersit_1- is the second factor
(c) Working day- on a working day people use shared bikes more
(d) Year- in 2019 the demand has already increased (year on year growth is seen so far)

We can see that the equation of our best fitted line is:

cnt = 0.009 x weekday + 0.026 x workingday + 0.621 x atemp - 0.127 x windspeed + 0.100 x season_4 + 0.074 x weathersit_1-0.199 x weathersit_3 + 0.234 x year

General Subjective Questions

1) Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a machine learning algorithm. It is supervised learning. Linear regression models a target variable using a predictor. The predictor is an independent variable and target is dependent variable.

It is used to find relationship between dependent and independent variable.

Linear regression predicts a dependent variable value (y) based on the independent variable value(x). It shows linear relationship between x (input) and y (output).

The equation for linear regression is as below:

$$y = a_0 + a_1 * x$$

Here, y = target variable

a_0=intercept

a_1=slope/coefficient of x

x=predictor

Linear regression algorithm is to find the best values for a_0 and a_1, so that we have the best fit line.

The best-fit regression line, will give minimum difference between predicted y value to actual y value . So, it is very important to update the a_0 and a_1 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

Cost function, also known as Mean Squared Error (MSE) , is a function that measures the performance of a Machine Learning model for given data. Cost Function quantifies the error between predicted values and expected values and presents it in the form of a single real number. The equation for MSE is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Mean Squared Error

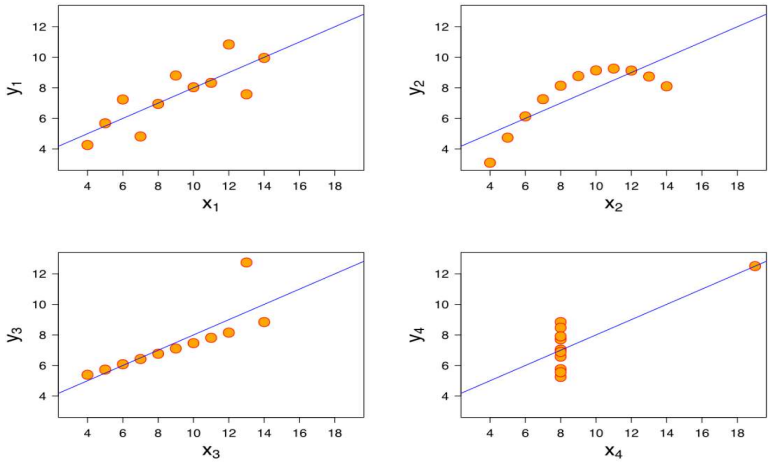
2) Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four data sets that have nearly identical simple statistics but when they are plotted, very different distributions appear and very different when graphed. Each dataset consists of eleven (*x*,*y*) points.

Anscombe's quartet shows why data visualization is important even before analysing the data.

- Here, the first plot shows simple linear relationship between two variables.
- The second one is not normally distributed but shows some relationship between variables but it is not a linear relationship.
- The third graph has a linear relationship but has different regression line. The outlier will influence the correlation in this case.
- Finally, the fourth graph shows an example when one high leveraged point is enough to produce a high correlation coefficient while other points are not showing any relationship.

Below plots shows different distribution for similar data set.



Anscombe's quartet

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

Pearson's R is a linear correlation between two variables X & Y. It has a value between +1 and −1.
Here 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
The formula is as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

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

Scaling is a method used to standardize the range of features of data. Scaling is necessary because te data value may vary widely. Scaling is important in data pre-processing while using machine learning algorithms.

During data pre-processing standardisation and normalisation are used.

Standardization

Standardization is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with mean as zero and standard deviation as 1.
μ=0 and σ=1
where μ is the mean and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

Whereas **Normalisation** is Min-Max scaling shrinks the data to a range of -1 to 1 or 0 to 1.
It works well if the distribution is not Gaussian or standard deviation is very small.
If the data is having outliers then normalisation is not good to use and in such cases standardisation can be used.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

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

The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It helps to diagnose multicollinearity in a model.

$$VIF = \frac{1}{1 - R^2}$$

VIF is computed for all the predictor in a model. If the value of VIF is 1 then predictor is not correlated to other predictors. An **infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Q-Q plot is a probability plot, a graphical method to compare two probability distributions by plotting their quantities against each other.