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)

- There is high correlation between Cnt and Year
- There is high correlation between Cnt and Temp/atemp , indication of multicollinearity
- Cnt is highest negatively correlated to Spring Season
- There is also negative correlation in Jan and Feb month
- Sat and Sun are highly -ve correlated
- Humidity and Windspeed is negatively correlated
- Casusal + Registered = cnt , so these two variables can as well be out of modelling
- There is Year on Year growth in rental services
- Growth is significant from May to Sep
- Bikes are rented more in Fall Season & Clear Weather
- Irrespective of working or non-working day, growth is same.
- Year 2018 Everyday rental services are almost similar wherein Year 2019, there is more service towards the weekend

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

Drop_first is important in reducing number of dummy variables. It also helps in reducing the correlation among dummy variables. Reduces redundancy, for example in case of Gender, we don't need both Male and Female variables in dataset. We can have only Male=1, if Male =0, it can be interpreted as Female.

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

Temperature has the highest correlation

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

There are following assumptions of Linear Regression

- Linear Relationship
- Homoscedasticity
- Absence of Multicollinearity
- Independence of residuals
- Normality of Errors

After building model, I predicted the values on training set and drew few graphs to check if there is linear relationship as per model.

VIF < 5 for all independent variables in model

There is normal distribution of Error

Scatter plot between Residual and Target Variable

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

1. Temp

2. Year

3. Season

As per Model 6, coefficient of Temp, Yr and season details like Summer/Winder are having high significance

```
const          0.157371
yr             0.229702
holiday       -0.097564
workingday    -0.029735
temp           0.525834
windspeed     -0.136403
2_summer       0.091796
4_winter       0.136862
Cloudy        -0.085511
Light Snow    -0.266263
Sep            0.106777
```

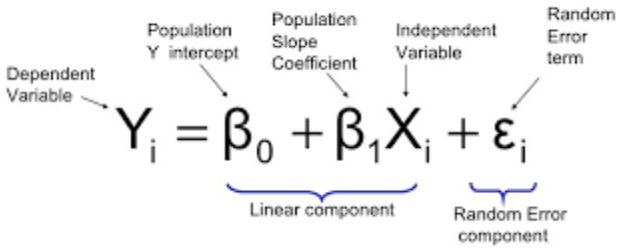
General Subjective Questions

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

Linear Regression is a machine learning algorithm, which helps in identifying the linear relationship between the variables, which best fits based on the available data points. This is done based on Residual Sum of Squares (RSS) method.
There are two types of linear regression – Simple & Multiple Linear Regression.
The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

Straight line between two variables when plotted on scatter plot is represented as **Y=m X + c**
Y= dependent variable
X= Independent variable
c= Intercept of a line
m is the slope of line

We can also write the linear equation as follows



As part of linear regression mainly we follow below steps to come up with a model

1. Analyzing the correlation and directionality of data, it can be positive or negative relationship. Based on correlation between the dependent & independent variables, we either add all the independent variable and start dropping one at a time to keep reviewing the model's performance. We can as well perform forward process wherein one variable is added at a time and review the performance.

2. Estimating the model or fitting the line based on R^2 and p-value. Independent variable's p-value should be lesser than 0.05 to be included in model. Higher the value of R^2 better the model performance would be.

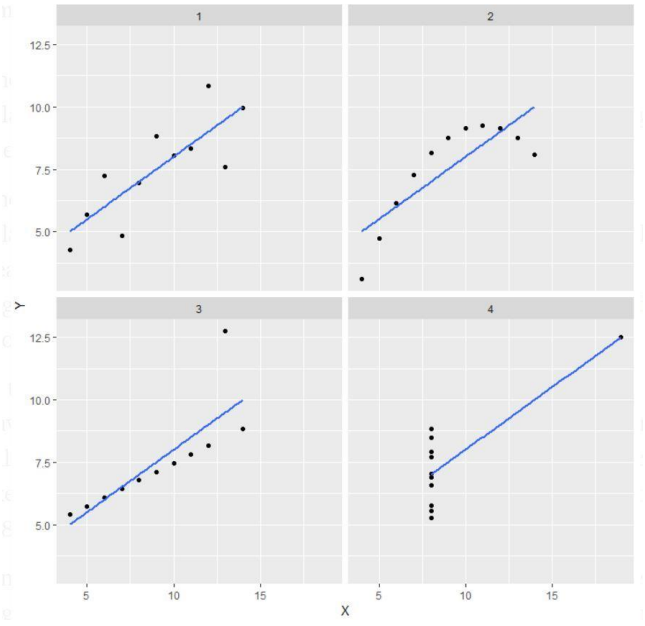
3. Evaluating the validity and performance on test data. Once the model is ready, we use test data to check the performance of model.

4. Need to perform residual analysis and make sure the distribution is normal.

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

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.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.



As shown above all the four data sets have similar stats like Number of observations, Mean, Standard Deviation but when plotted as scatter plot, they are telling different story. You can even see the straight line is fit in all dataset but looking at data points only in chart 1 has linear relationship, rest others have either the outliers or non-linear relationship.

It is very much important to perform data visualization and build model which best fits the data point otherwise it is very easy to fool linear regression.

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

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's.

Pearson's correlation (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression. Below is the formula for Pearson's R

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
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

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

Scaling is a method to get all features at same level, so that model is computed properly. For example, there are few variables in KG and few in Grams, so we need to convert them to same scale like in Grams for better calculation.
Scaling can be performed one data is split into train & test, we can perform scaling on all variables which may be part of model.

Mainly there are four method of scaling as mentioned below

1. Normalizing
2. Standardizing
3. Min-Max Method
4. Unit Vector

Normalizing is the method where data values are scaled between 0 and 1

For continuous variables like height, we can use below formula, where **df** is data frame and **height** is column name

```
df["height_normal"] = (df["height"] - df["height"].mean()) /
                                    (df["height"].max() - df['height'].min())
```

Feature like sex, which is categorical, we can convert them as 0=female and 1=Male

Standardizing is done by taking each value of your column, subtracting the mean of the column, and then dividing by the standard deviation of the column.

$$x' = \frac{x - \bar{x}}{\sigma}$$

In similar height example , we can calculate the standard value as mentioned below

```
df["height_standard"] = (df["height"] - df["height"].mean()) / df["height"].std()
```

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

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

VIF = 1 / (1-R^2)

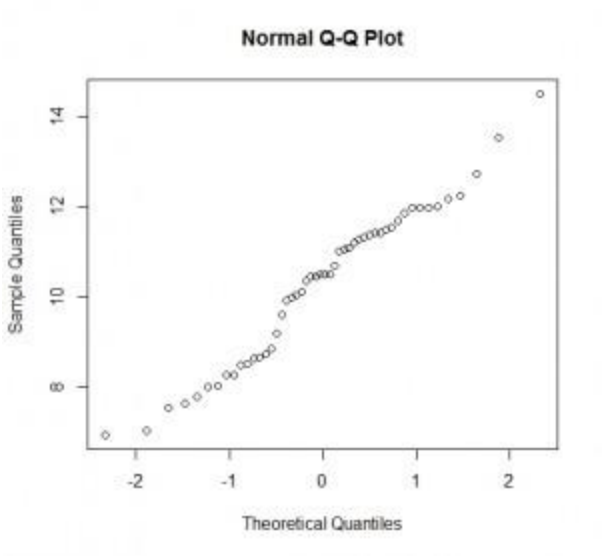
If there is perfect correlation, then VIF is infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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. 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 also known as quantile-quantile plot, this is a scatter plot of two set of quantiles against each other. A quantile in a data is percentage of point below the given value in data set, this is also referred as percentile.

If the 2 sets came from a population with the same distribution, the points fall approximately along a straight line.



Q-Q plot helps in answering the following

1. If data sets come from population with a common distribution
2. If distribution has similar shape
3. If they have common location and scale
4. If they have similar tail behavior

There are few advantages of Q-Q plot

1. Sample size can be unequal
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Making a q-q plot

1. Sort all the data values in ascending order
2. Draw a normal distribution curve
3. Find the z-value for each segment, segment will be n-1, where n is number of observations
4. Plot data sets values against the z-values