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)**

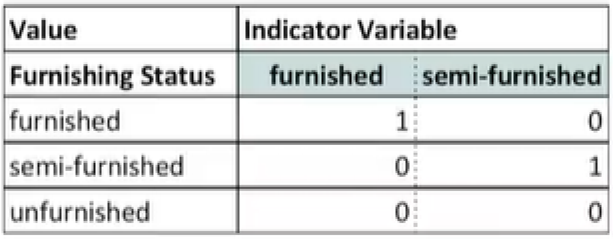
A. From the analysis of categorical variables from the dataset, below are the points we can infer:

- Among all the seasons, the fall season had more bookings in both the years 2018 and increased in 2019.
- The booking trend started going up in June and September with the highest bookings the trend started coming down from then.
- Bookings are widely affected by the weather situation which thus shows clear weather is having more bookings.
- The number of bookings on working days is slightly more compared to holidays which seems quite reasonable.
- The given dataset has data for 2 years and we can conclude that the business has gone up compared to the previous year.

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

A. It is important because:

- 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.
- Let’s say we have 3 types of values in the Categorical column and we want to create a dummy variable for that column. If one variable is not furnished and semi_furnished, then It is unfurnished. So we do not need 3rd variable to identify the unfurnished.



Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

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

A. 'temp' variable has the highest correlation with the target variable.

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

A. Validation is done in 4 kinds:

- Normality of error terms: error terms are normally distributed
- Linear relationship: linearity is present among variables
- Homoscedasticity: no visible pattern in residual values
- Independence of residuals: no auto-correlation

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

A. the top 3 features that contribute significantly towards explaining the demand for shared bikes are:

- temp
- mnth_sep
- season_winter

General Subjective Questions

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

A. Linear regression is a very basic form of machine learning in which a model is trained to predict the behavior of data based on some variables. In simple terms when you consider two variables X and Y and put them on the graph and you see they are linearly correlated then you can say the variables are in a linear relationship.

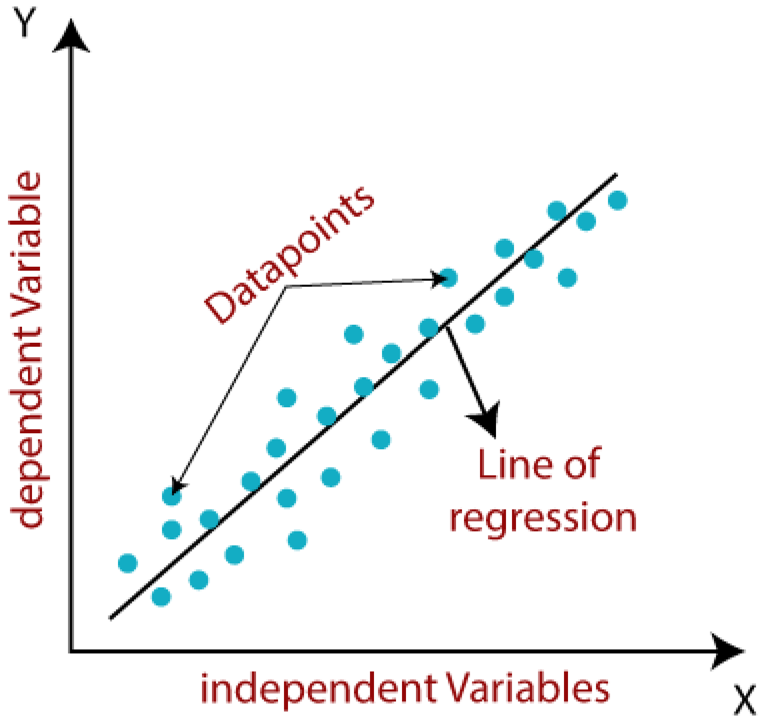


Fig - 1

In the above figure, we can see how X and Y are in a linear relationship. Mathematically we can represent the linear regression in the form of the equation which is **y = mx + c** where:

y = target variable

x = independent variable

m = slope of the line

c = constant also known as the y-intercept

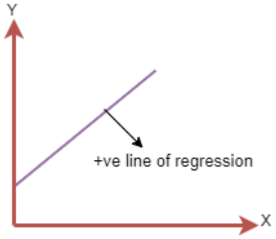
Linear regression can be divided into two types:

- **Simple linear regression:** In which only one independent variable is used for prediction.

- **Multiple linear regression:** In which more than one independent variable is used for prediction.

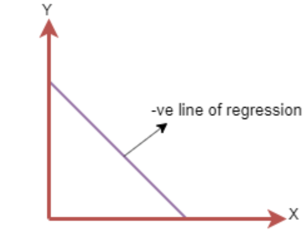
In the above Fig - 1 we see there is a line of regression that can show two types of relationships:

- **Positive linear relationship:** If the dependent variable increases on Y-axis and the independent variable increases on X-axis then it is said to be a positive linear relationship.



The line equation will be: $\mathbf{Y= a_0+a_1x}$

- **Negative linear relationship:** If the dependent variable decreases on Y-axis and the independent variable increases on X-axis then it is said to be a positive linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

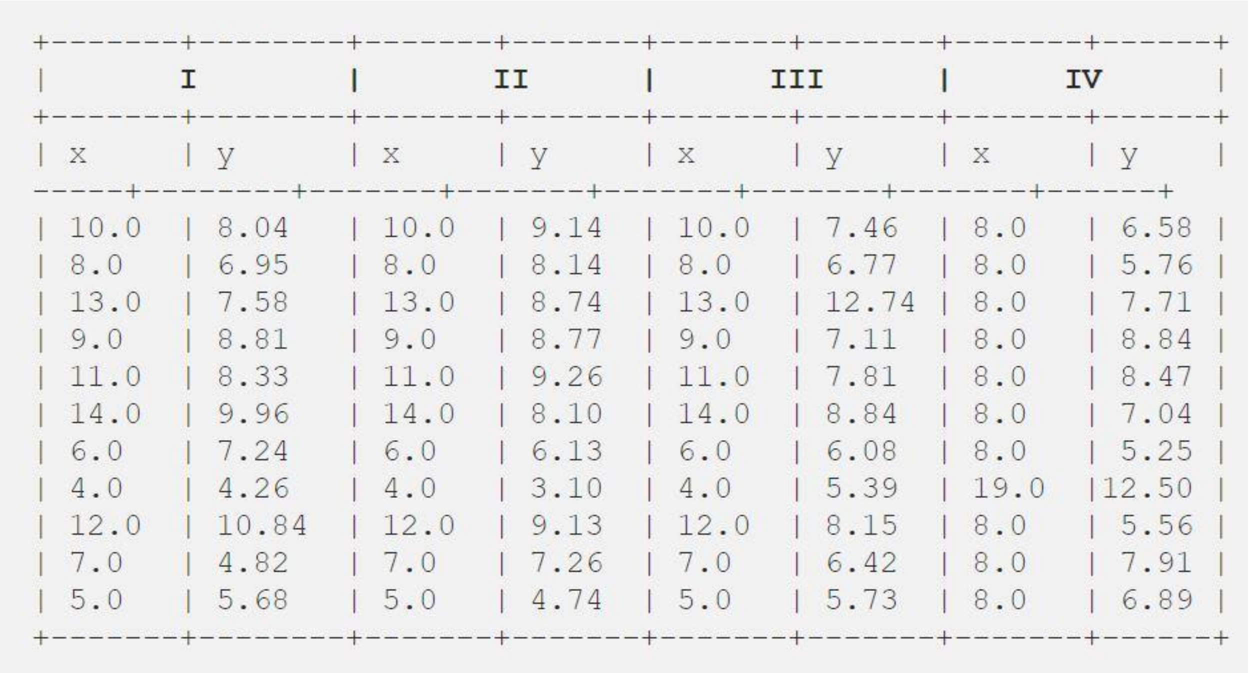
The four assumptions of linear regression are:

1. Linear Relationship.
2. Independence.
3. Homoscedasicity.
4. Normality.

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

A. Anscombe's quartet has 4 datasets that have similar statistical properties but appear very different when plotted on the graph. Each dataset consists of eleven points. In 1973 statistician Anscombe constructed it to demonstrate the importance of data before analyzing it and the effect of outliers on statistical properties.

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data points are given below.

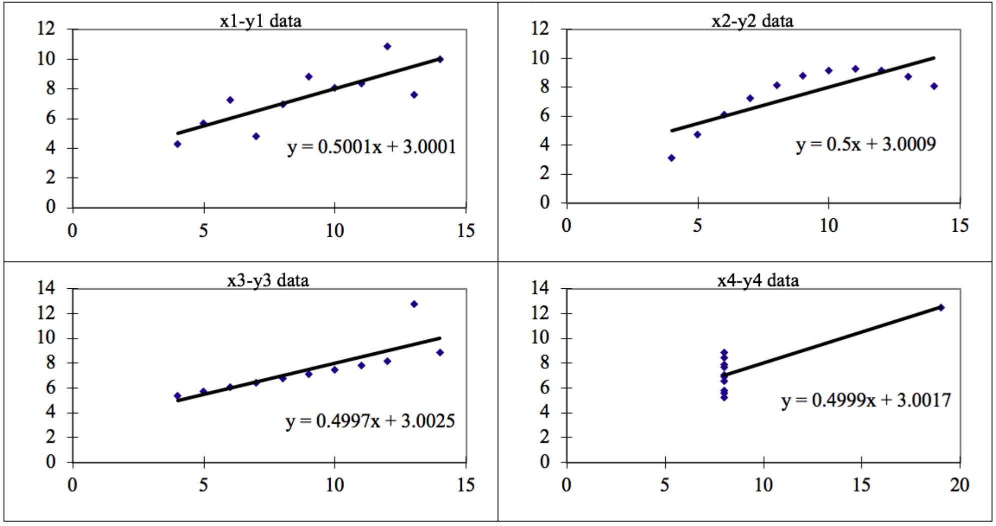


```
+-------+--------+-------+--------+-------+--------+-------+------+
|       I        |       II       |       III      |       IV     |
+-------+--------+-------+--------+-------+--------+-------+------+
| x     | y      | x     | y      | x     | y      | x     | y    |
-----+--------+-------+-------+-------+--------+-------+------+
| 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 |
+-------+--------+-------+--------+-------+--------+-------+------+
```

- The mean of x is 9 and the mean of y is 7.50 for each dataset.
- The variance of x is 11 and the variance of y is 4.13 for each dataset.
- The correlation coefficient between x and y is 0.816 for each dataset.

When we plot graphs for all these 4 datasets we see the same line of regression but each dataset tells us a different story which we can see below:

Anscombe's Quartet

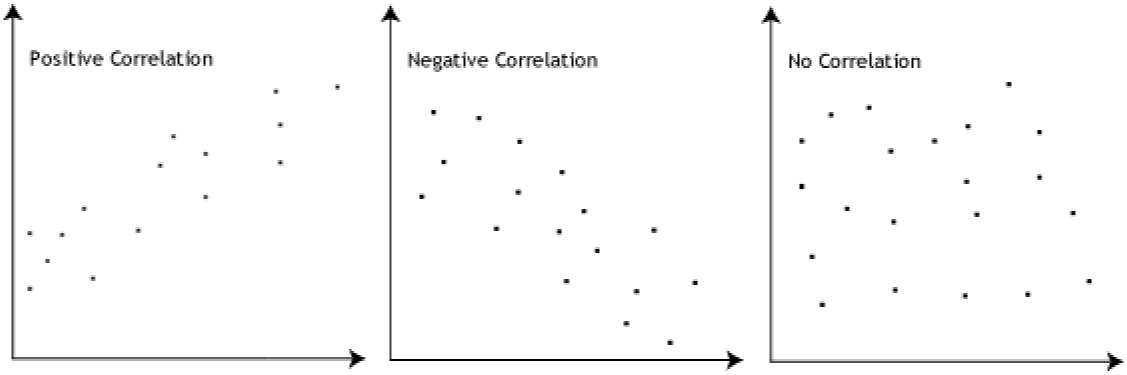


Explanation of graphs:

- In the first one if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one, you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier and is indicated to be far away from that line.
- Finally, the fourth one shows an example of when one high-leverage point is enough to produce a high correlation coefficient.

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

A. Pearson's R 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 deviation. It is essentially a normalized 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

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

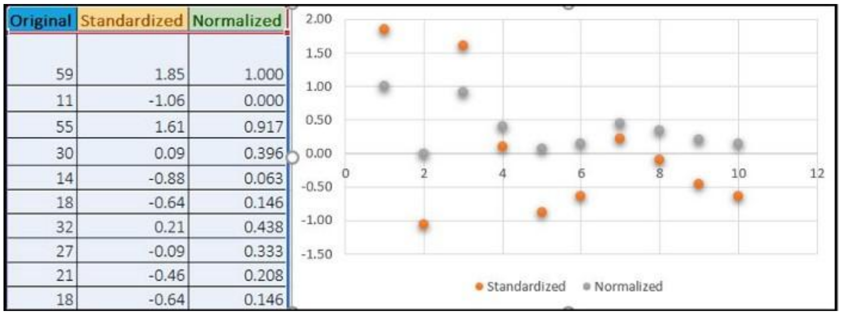
A. Scaling is a step of data pre-processing applied on independent variables to normalize the data within a range. This is mainly performed to make calculations easy in the algorithm.
When data is collected it will be of huge range with several magnitudes and if scaling is not done the model will only cosider certain magnitude and thus making the model incorrect,scaling is performed to bring all the magnitudes into consideration.
Scaling will alter the coefficient value but not R2 value or F-statsitic.
Normalized Scaling:
It is also known as MinMax Scaling and this will range the the data between 0 and 1.
sklearn.preprocessing.MinMaxScaler is used for implementation.

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

Standardized Scaling:
Standardized scaling is done on the basis of Z-Score, the values are replaced with Z-Score and brings the entire data into standard normal distribution which will have mean as 0 and standard deviation as 1.
sklearn.preprocessing.scale is used for implementation.

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

The important difference between normalized scaling and standardized scaling is that the standardized scaling looses some data especially outliers which makes it a disadvantage for standardized scaling and making MinMax scaling the preferred one.

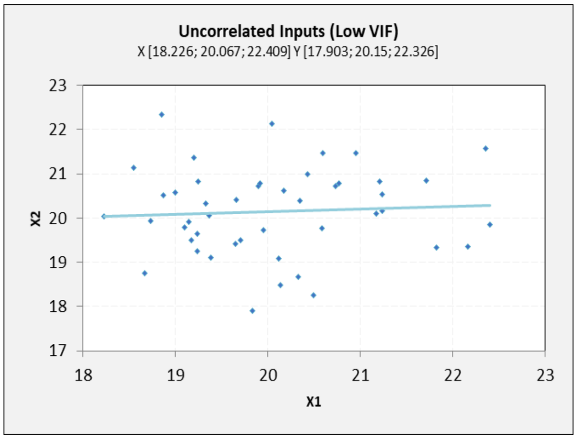


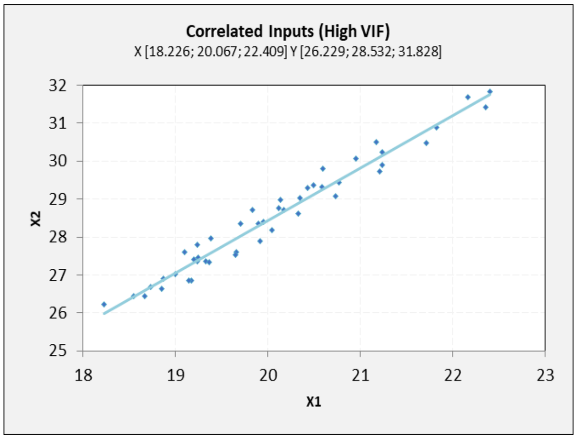
Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

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

A. If the value of VIF = inf then it shows that there is perfect correlation of that particular variable with the target variable, if we consider this situation between two independent variables and calculate the R2 it will be 1 which shows perfect correlation and it leads to 1/(1-R2) infinity. The variables with such VIF value will not contribute to the model rather alter the results so inorder to avoid it they are often dropped from the dataset and thus avoiding multicollinearity.

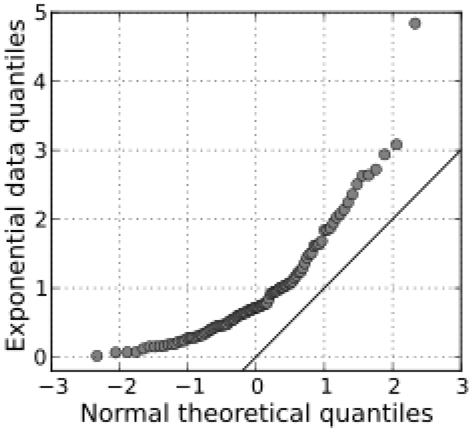
In simple terms, inf as the value of VIF will indicate that the corresponding variable may be expressed exactly by a linear combination of other variables.





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

A. Q-Q plots also known as Quantile-Quantile plots are simply plots of two quantiles against each other. A qunatile is nothing but a fraction where certain values fall below the quantile. If we consider median it is often 50-50 which means 50% of data lies below the point and 50% lies above the point. The main purpose of using a Q-Q plot is to understand whether or not both the data sets are coming from same distribution and in this case the plot will be with 45 degree angle if they come from same distribution and the points will be on the reference line. The more distance of points from reference line will infer that the two datasets come from different distribution.



It is often important to know weather the two datasets are from same distribution or not as we often get data which is not created by us and inorder to get more clarity on the model and the population of dataset population and their distribution we use Q-Q plot.

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

statsmodels.api will provide the feature of plotting qqplot and one more interesting thing is we also do have qqplot_2samples to plot graph for two different datasets.