**# 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? (6 marks)

**Solution**: 1. Season: The demand of bike is less in the spring season when compared with other seasons.

2. year: The Bike demand in year 2019 is higher as compared to 2018.

3. Month: The month bar plots indicates that more bikes are on rent during may to october.

4.Holiday and workingday: The working day and holiday bar plots indicate that more bikes are on rent during normal working days than on weekends or holidays.

5. Weathersit: The weathersit bar plot shows that more bikes are on rent during clear while it is low when there are Few clouds and Partly cloudy.

6.Weekdays: There is no significant change in the weekdays on the demand of bike.

2: Why is it important to use *drop\_first=True* during dummy variable creation? (6 mark)

**Solution:**

* It is used to reduce the collinearity between dummy variables.
* It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

1. How did you validate the assumptions of Linear Regression after building the model on

the training set? (6 marks)

**Solution**:

* Linearity of relationship between response and predictor variables.
* Constant variance of the error.
* Normality of the error distribution (Normal distribution of error terms).
* Less Multi-collinearity between feature (low VIF).
* The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y which I have performed in the EDA.
* If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

1. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (6 marks)

**Solution:**

1. Weathersit
2. Temp
3. Season

**# General Subjective Questions**

1. Explain the linear regression algorithm in detail. (10 marks)

**Solution:**  
It is a form of regression, where the target variable is continuous. It estimates the

relationship between a target variable and one or more predictor variables.

The Equation of linear Regression is y= m1x1 + m2x2 + m3x3 + ……….. + m(n) x(n) + c .

Where y is target variable and x1, x2 , x3 …… xn are predictor variables . And we have two

unknowns, m, and c, and we need to choose those values of m and c, which provides us with

the minimum error. We need to get the best fit line which is the line that has the minimum

error. In linear regression, when the error is calculated using the sum of squared error, this

type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression.

Error function is explained by ‘e = - y’, and error depends on the values of ‘m’ and ‘c’. Our aim

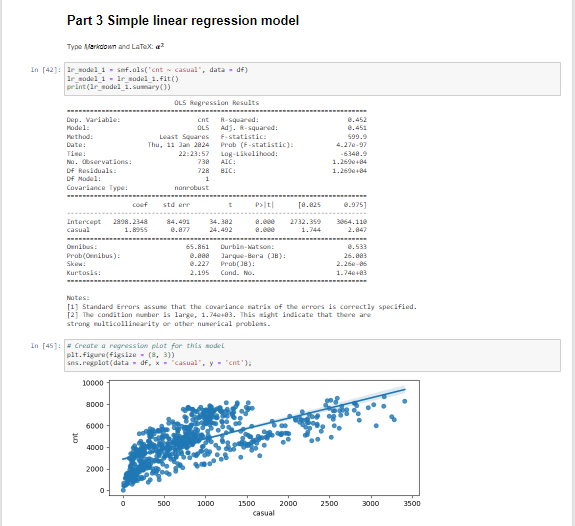
is to build an algorithm which can minimize the error.

And in order to do so we use cost function of Linear Regression, Which is:

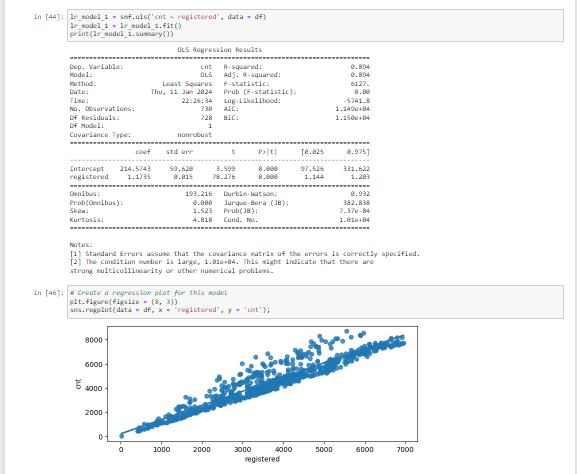
J ( mi, c )= (1/2n) Σ (yi – yp)^2

Where yi and yp are expected values and predicted values.

Our main aim is to minimize J by changing m and c and it can be done using Gradient Descent Algorithm.Cost function measures the performance of a Machine Learning model for given data.



With the bike sharing assignment example we can see that count on y axis and casual on axis, the value of R-squared values is 0.45 which is 45%. we have coefficient values we can use in the formula.



With respect to count on y axis and registered on x axis here we can see very strong linear regression line. R-squared value we have is 0.89 (89%).

1. What is R-squared? And What is adjusted R-squared? (4 marks)

**Solution**:

* R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable in a regression model.
* R-squared only works as intended in a simple linear regression model with one explanatory variable. With a multiple regression made up of several independent variables, the R-squared must be adjusted.
* The adjusted R-squared compares the descriptive power of regression models that include diverse numbers of predictors. Every predictor added to a model increases R-squared and never decreases it.
* Thus, a model with more terms may seem to have a better fit just for the fact that it has more terms, while the adjusted R-squared compensates for the addition of variables; it only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance.
* In an overfitting condition, an incorrectly high value of R-squared is obtained, even when the model actually has a decreased ability to predict. This is not the case with the adjusted R-squared. Essentially, R-squared is a statistical analysis technique for the practical use and trustworthiness of betas of securities.

1. What is Scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (6 marks)

**Solution** :

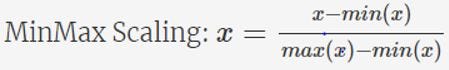
It is a step of data Pre-Processing which is applied to independent variable to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling 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/Min-Max Scaling:

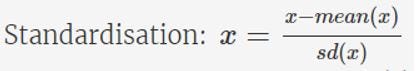
* It brings all of the data in the range of 0 and

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



# Standardization Scaling:

* 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).



* sklearn.preprocessing.scale helps to implement standardization in python.
* One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

happen? (6 marks)

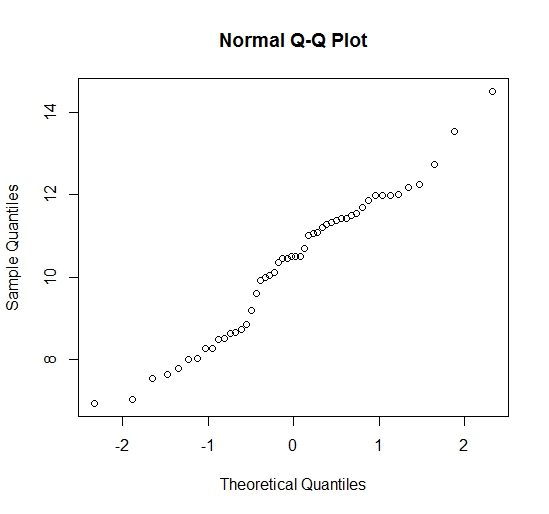
**Solution** :

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:  
  
X\_1=C+ α\_2 X\_2+α\_3 X\_3+⋯  
  
〖VIF〗\_1=1/(1-R\_1^2 )  
  
Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:  
  
X\_2=C+ α\_1 X\_1+α\_3 X\_3+⋯  
  
〖VIF〗\_2=1/(1-R\_2^2 )  
  
If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. 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. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

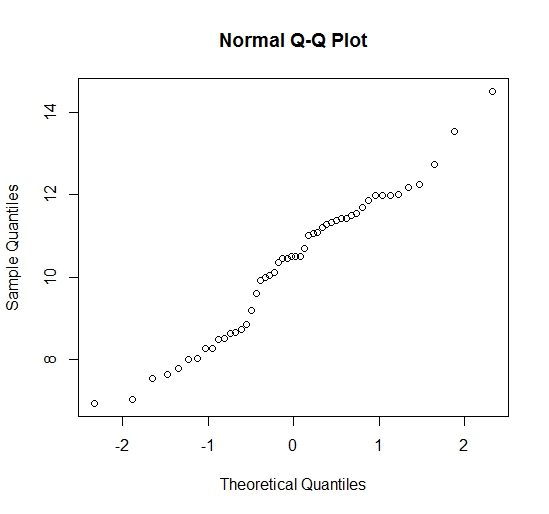
1. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (10 marks)

**Solution** :

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.



A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.



For example

