Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Our Final model is

*cnt = 1664.98 + 2038.16 * yr + 4163.18 * temp - 1301.92 * windspeed - 504.21 * season_Spring + 540.12 * season_Summer + 796.78 * season_Winter + 758.38 * mnth_September - 670.12 * weathersit_Cloudy/Misty - 2441.59 * weathersit_Rainy/Snowy*

which includes the following Categorical variables **yr, season (Spring, Summer, Winter),** ***mnth_September*** *and* ***weathersit (Cloudy/Misty, Rainy/Snowy)***

- **yr** has very high coefficient because as we pass each year, the demand is increasing.
- **season_Spring** has negative impact because of the weather of that particular season. We might have negative temperatures which is a major roadblock for demand.
- **season_Summer** will have clear, hot and sunny days where as **season_Winter** will have cool and ideal temperatures due to which the demand might be on the higher side which is evident for the coefficient.
- **mnth_September** is the first month of Fall season which is Summer to Winter transition where we find suitable weather due to which citizens may opt to come out of their home for their daily needs.
- **weathersit_Cloudy/Misty** and **weathersit_Rainy/Snowy** have negative coefficients due to adverse/unsuitable weather because of which citizens might not opt to come out of their homes.

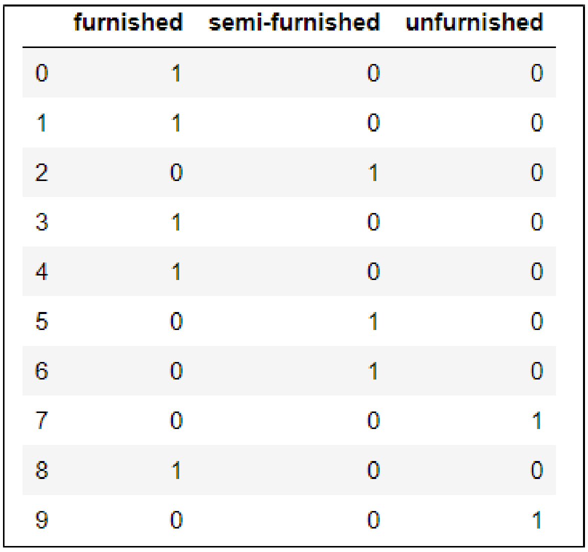
The below variable does not feature in our final model, due to below reasons:

- From Model 3, we have a high p-value for **holiday**, which means the relationship is insignificant wrt target variable and the variable is also a highly imbalanced variable. Hence, we dropped this variable.
- From EDA we have seen that **workingday** is a highly imbalanced variable with high 1's (499 records). Hence, we have dropped this variable to prevent the model from giving biased results.
- From EDA we have seen that the demand is almost same across all **weekdays**. So, the 15 variables which were obtained from RFE feature only **weekday_Saturday.** But from Model 7, the p-value of weekday_Saturday is ~0.05 and the demand on weekdays is almost equal on all days. Hence, this variable is insignificant wrt target variable and was dropped from our analysis.

Q2. Why is it important to use drop_first=True during dummy variable creation?

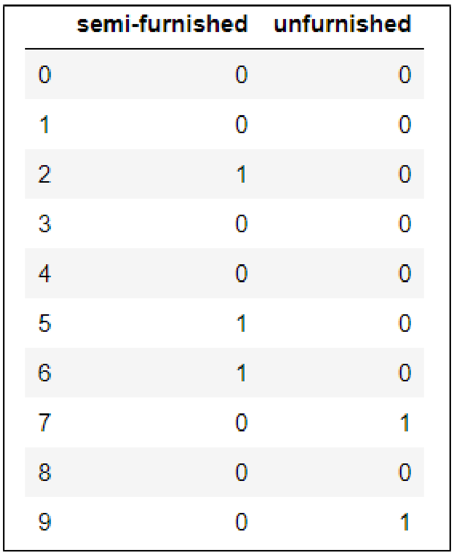
drop_first=True will drop the first dummy variable of a categorical column and returns n-1 dummy variables for a column with n categorical levels. The dropped column will become our Base State and the other variables may have more or less impact wrt Base State.

For example, we have categorical column with values “furnished”, “semi-furnished” and “unfurnished”. In fig 1 we have not opted to use drop_first. Hence get_dummies() has returned 3 dummy variables for 3 categorical levels.



	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0
5	0	1	0
6	0	1	0
7	0	0	1
8	1	0	0
9	0	0	1

fig 1



	semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0
5	1	0
6	1	0
7	0	1
8	0	0
9	0	1

fig 2

In fact, we can represent all 3 categorical levels with just 2 dummy variables as shown in fig 2 where

- 00 will correspond to furnished
- 01 will correspond to unfurnished
- 10 will correspond to semi-furnished

We can also reduce the redundancy in the information provided by dummy variables.

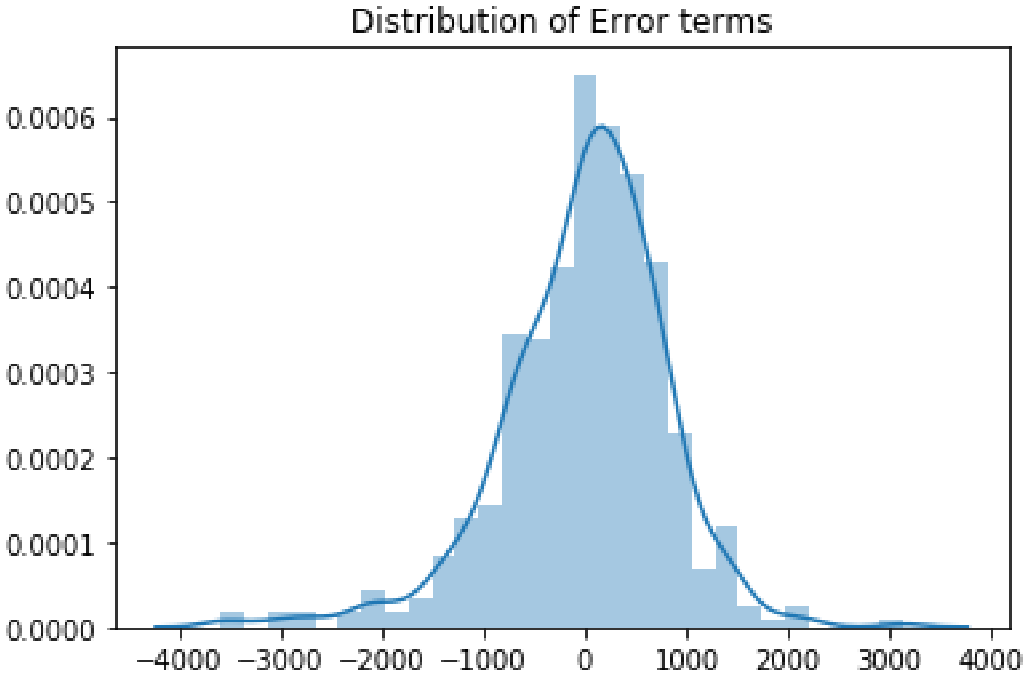
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

From Pairplot, **temp** and **atemp** are highly correlated with target variable and highly correlated with each other. Hence, we have to use only one variable for our analysis to avoid multi-collinearity.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1. **Normality of Error terms**

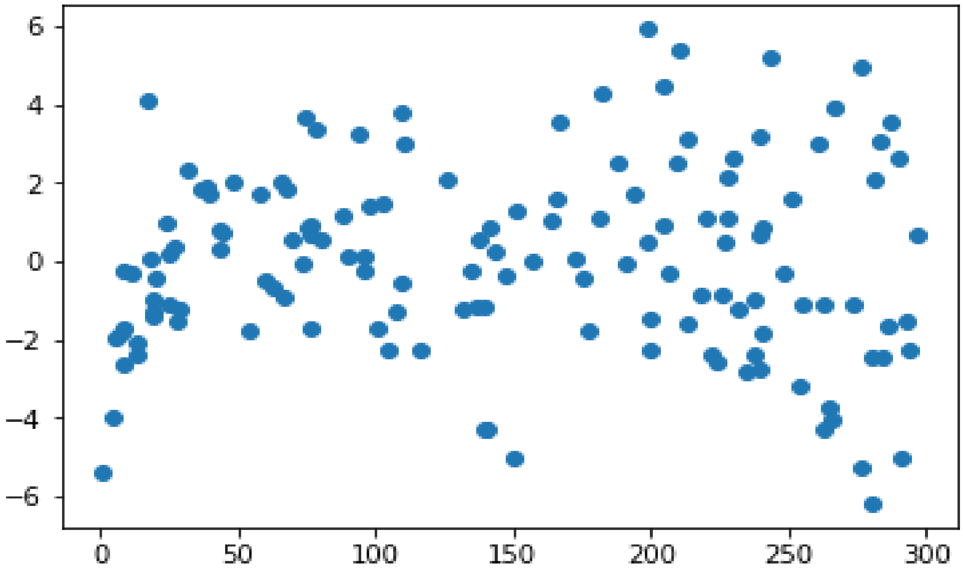
In the assignment, we have verified “Normality of Error terms” by plotting a histogram of difference between y_train and y_train_pred which is shown in below figure.



From above figure, the error terms are normally distributed and the mean is also centred at 0. Hence, “Normality of Error Terms” holds True in our analysis.

2. **Error terms are independent of each other**

We can also verify the “Error terms are independent of each other” using a scatter plot.



The above plot is from a Simple Linear Regression model, where x corresponds to X_train and y corresponds to residuals.

The plot shows that error terms are scattered across the plot and there is no pattern associated with the error terms.

We can conclude that “Normality of Error Terms” holds True.

3. Error terms have constant variance

From fig 3, we can say that the error terms have same standard distribution for most part of the plot unlike fig 4 where there is constant increase in the standard deviation with increase in x.

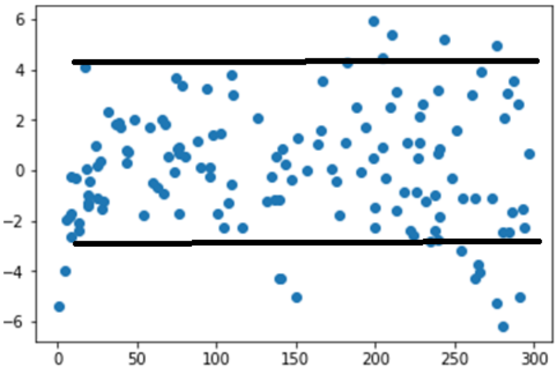


fig 3

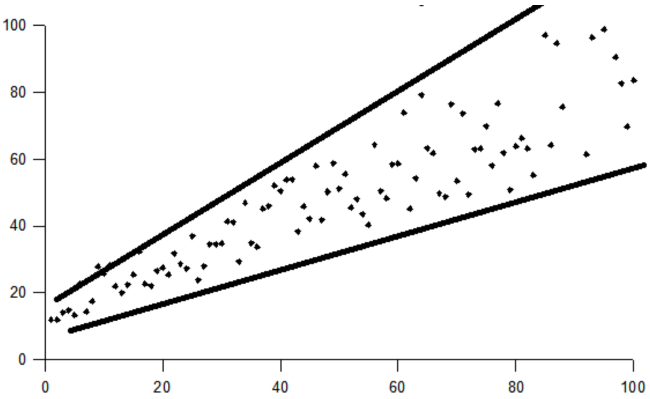


fig 4

If the error terms have same variance across the plot then “Error terms have constant variance” holds True.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Based on the final model, **yr, temp** and **season_Winter** are the 3 significant features with 2038.16, 4163.18 and 796.78 as their coefficients respectively.

- **yr** has very high coefficient because as we pass each year, the demand is increasing.
- **temp** has the highest coefficient as it is highly correlated with the target variable because the demand varies based of the weather of a given day.
- **season_Winter** will have cool and ideal temperatures due to which the demand might be on the higher side which is evident for the coefficient.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is used to establish a relationship between target variable and the predictors to know the variables which are driving the dependant variable. A Linear Regression algorithm comprises of following steps:

➢ **Import Libraries:**

- Import necessary libraries that are required for our analysis.

➢ **Reading and understanding the data:**

- Read the data and load it into a data frame.
- Perform routine checks on data like verifying the shape, Info and descriptive summary of columns.

➢ **Data Pre-processing:**

- Drop unnecessary columns and the columns/rows which have high null values.
- Perform the data quality checks.
- Identify and treat the outliers.

➢ **Exploratory Data Analysis:**

- Perform Univariate, Bivariate and Multivariate analysis to extract insights from the given data which are useful in building the model.

➢ **Data Preparation:**

- Identify the Categorical and Continuous variables in the data frame.
- Create dummy variables for Categorical variables.
- Split the data into df_train and df_test data sets.
- Scale the Continuous variables using any scaling approach.
- Create X-train and y_train using training dataset.

➢ **Building a model:**

- Perform Feature selection by using Automation approach like RFE to ensure that we don't end up in creating a complex model. If we fail to do so, the complex model will learn too much from train data which may lead to overfitting.
- Build a model using the variables given by RFE.
- Calculate VIFs for variables used in building the model.
- Drop a variable manually by inspecting the VIFs and p-values.
- Repeat this process until all the variables used in building the model have VIFs < 5 and p-values < 0.05

- Verify the R-squared value after building each model. If there is any significant drop in R-squared make sure to add it back to the model and ensure that VIFs and p-values of all variables are under check.

➢ **Residual Analysis:**

- Predict the values of X_train using the final model.
- Calculate the difference between predicted values and y_train.
- Verify Linear Regression assumptions.

➢ **Making predictions using the final model:**

- Drop the variables from test data that were not used in final model.
- Scale Continuous variables using transform() method.
- Create X_test and y_test.
- Predict the values using final model and X_test.
- Plot a scatter plot using y_test and y_pred to understand the spread.
- Calculate the R-squared value using y_pred and y_test.
- Verify the R-squared and Adj. R-squared for both train data and test data.
- If the R-squared value of test data is within the range of +5 or – 5 of R-squared value of train data then our model is generalized model.

➢ **Making predictions using the final model:**

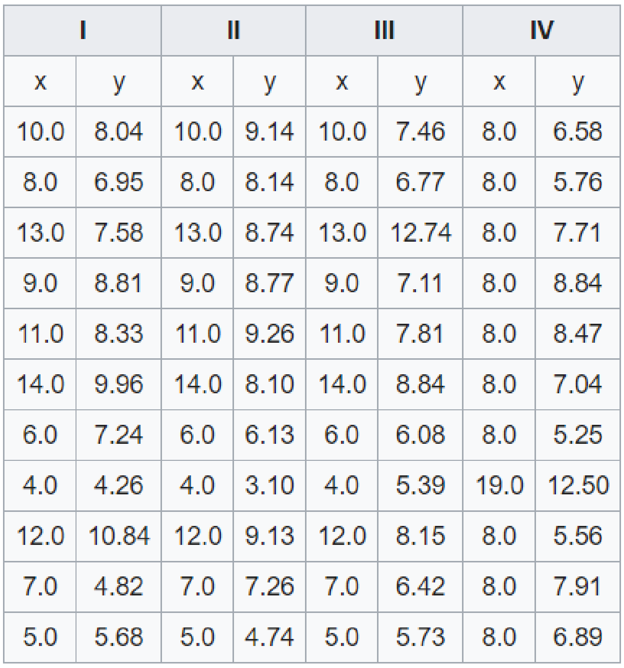
- Report the final model which is simple and interpretable.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet demonstrates how important it is to plot the data rather than simply relying on summary statistics alone.

In fig 5, we have 4 data sets with below properties

Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	y = 3.00 + 0.500x
Coefficient of determination of the linear regression	0.67



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

fig 5

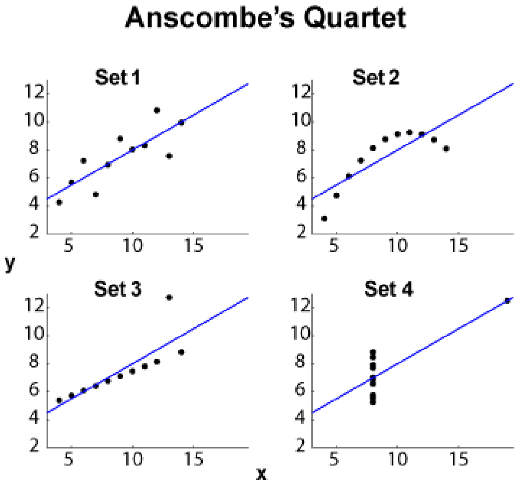


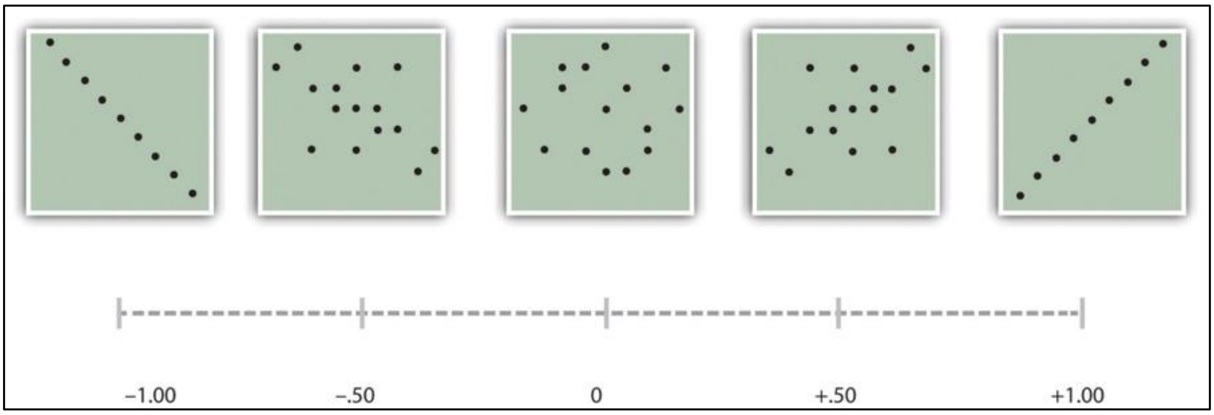
fig 6

- When we plot the 1st, data set we get the line as shown in Set 1 of fig 6 which is a simple Linear Relationship between x and y.
- When we plot the 2nd data set which has same properties of 1st data set, we expect regression line to be similar to the graph of set 1. But it appears to be as set 2 of fig 6. While there us an obvious relationship between x and y, it is not linear.
- When we plot the 3rd data set which has same properties of 1st and 2nd data sets, the data points are very close to regression line. So, we expect that the correlation would be very close to 1. But there is an outlier which exerts enough influence to bring down the correlation coefficient to 0.816.
- When we plot the 4th data set which has same properties of 1st, 2nd and 3rd data sets, all data points lie on x=8 line. There are couple of points which are very close to the line that are sufficient to give a correlation value of 0.816, even though the other data points do not indicate any relationship between the variables.

So, we have seen how different data sets with same properties could have different distributions. So, it is imperative to plot the data and not to simply rely on Summary statistics of columns.

Q3. What is Pearson's R?

- Pearson's R is also referred as Pearson's Correlation Coefficient which is a measure of linear relationship between X and y variables.
- It is also known as Bivariate Correlation.
- Value always lies in between -1 and 1.



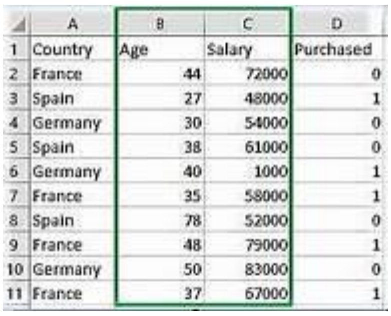
- In above figure, we have 5 plots with -1, -0.5, 0, 0.5 and 1 as their Correlation Coefficient.
- From above figure, If the correlation coefficient is 1 then X and y are strongly coupled in positive direction i.e. if there is an increase in X then there is an increment in y as well.
- If the correlation coefficient is 1, then it is also called as Perfect Positive Correlation.
- If the correlation coefficient is -1 then X and y are strongly coupled in negative direction i.e. if there is an increase in X then there is a decrement in y.
- If the correlation coefficient is -1, then it is also called as Perfect Negative Correlation.
- If the correlation value is 0 then there is not linear relationship between X and y.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a method used to normalize the range of independent variables or features of data. In simple words, scaling means converting a value of independent variables from 1 scale to another (common) scale.

Scaling is performed for below reasons:

- **Ease of interpretation:**
  From the below figure, there are 2 variables "Age" and "Salary" which are used in building a model to classify weather a person is likely to purchase or not. From the final model, we came to know that Salary is most significant variable which have high coefficient and Age is the least significant variable which have low coefficient.

	A	B	C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	0
3	Spain	27	48000	1
4	Germany	30	54000	0
5	Spain	38	61000	0
6	Germany	40	1000	1
7	France	35	58000	1
8	Spain	78	52000	0
9	France	48	79000	1
10	Germany	50	83000	0
11	France	37	67000	1

  If the variables which are used in the building the model are on different scales then it would be difficult to interpret the coefficients. The most significant variable might have very less coefficient and the least significant variable might have very high coefficients.

  If the variables which are used in the building the model are in same range then the interpretation of coefficients would be easy and helpful in finding the actual significance of the variables.

- **Faster Convergence for Gradient Descent Methods:**
  At the backend typically we have a Gradient Descent method going on. If we have lot of features in different scales, then it takes lot of time for convergence. If they are in same range then the convergence will become much faster.

Difference between normalized scaling and standardized scaling:

- **MinMaxScaler:**
  MinMaxScaler is also called as Normalization which uses below formula.

  $$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

  It will subtract the minimum value of the variable from the data point and divides the result with the difference of Minimum and Maximum values of the variable.

  “It brings all the variables in the range from 0 to 1” and we see its impact only on Continuous variables but not on Binary and dummy variables whose values are 0 and 1.

- **StandardScaler:**
  StandardScaler is also referred as Standardization which uses below formula.

  $$(x_i - mean(x))/stdev(x)$$

  If will subtract the Mean of variable from the data point and divides it with Standard deviation. Thus, converting the data point to z-score.

  On performing StandardScaler on a variable, it will return a Mean of 0 and Standard deviation of 1. Hence, the distribution is centred at 0.

  As this deals with Mean and Standard deviation, it will scale the Binary and dummy variables along with Continuous variables.

Scaling changes the coefficients and will not impact the p-values or accuracy or overall fit of the model.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIFs are used to detect the multicollinearity between the Independent variables.

Sometimes the pair wise plots and correlations may not be enough to detect multicollinearity as there might be a chance that 1 independent variable is explained by 2 or more independent variables. In these scenarios, pair wise plots and correlations are not a good idea to detect multicollinearity.

So, we build a model that predicts X1 using other independent variables. This process is repeated to all independent variables. This is the main idea behind Variance Inflation Factor approach to detect multicollinearity.

To quantify the relationships, we use Variance Inflation Factor which uses the below formula.

$$VIF_i = \frac{1}{1 - R_i^2}$$

If the VIF is > 10 then it is very high, if the VIF > 5 then it is considered as high else the VIF is under check.

Some times, the value of VIF for a particular variable will be infinate. This is because the denominator values is 0 which means the R-Squared value is 1.

If the R-Squared value is 1 then the variable is in perfectly correlated in positive direction with other variable and hence, is a redundent column of our data.

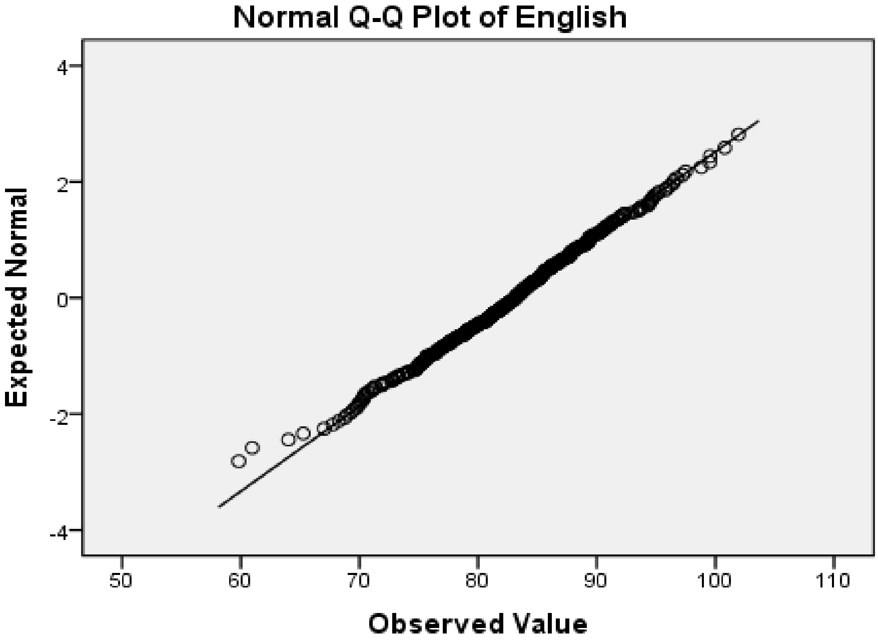
In other words, if the variable is in perfect positive correlation with other variable(s) then the VIF of that particular variable will become Infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile is the percentage of points below a given value in our data set. For example, Median is a quantile where 50% of the data fall below that point and the other 50% above it.

A Q-Q plot also referred as Quantile_Quantile plot, which is a scatter plot of two sets of quantiles against each other.

If the 2 sets come from a population with the same distribution, the points fall approximately along a 45-degree reference line.



How to make a Q-Q plot:

- Sort the numbers from smallest to largest.
- Draw a normal curve and divide the curve into n+1 segment's where n is the number of data points in our sample.
- Find the z-value for each segment.
- Plot the data set values against the z-values.

Q-Q plot helps us in answering the following questions:

- If 2 data sets are from populations with same distribution.
- If the distributions have similar shape.
- If they have common location and scale.
- If they have similar tail behaviour.

Advantages of Q-Q plot:

- It is not mandatory to have samples of equal size.
- It gives insights on multiple distribution features like shifts in location, Scale and Symmetry and also the presence of outliers.