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?

   These are the categorical features present in the dataset - 'season', 'mnth', 'workingday', 'yr', 'weekday', 'holiday', 'weathersit'. I plotted the box plot and pie chart distributions for these categorical variables with respect to the target variable 'cnt' and had the following observations:

   1. ***season***: Almost 32% of the bike booking were happening in fall season with a median of over 5000 booking (for the period of 2 years). This was followed by summer season & winter season with 27% & 25% of total booking. This indicates, season might be a good predictor for the dependent variable.

   2. ***mnth***: Almost 10% of the bike booking were happening in the months May, June, July, August & September with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

   3. ***weathersit***: Almost 67% of the bike booking were happening when weather is either clear or has few clouds with a median of close to 5000 booking (for the period of 2 years). This was followed by mostly cloudy weather with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

   4. ***holiday***: Almost 97% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday might not be a good predictor for the dependent variable.

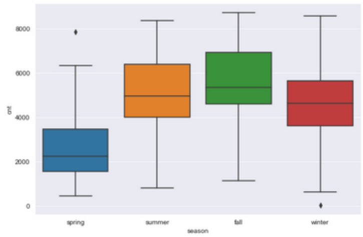
   5. ***weekday***: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.

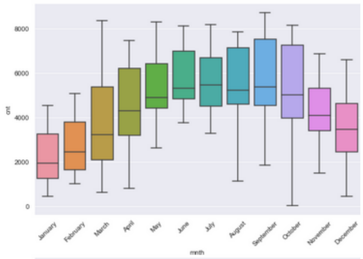
   6. ***workingday***: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday might be a good predictor for the dependent variable.

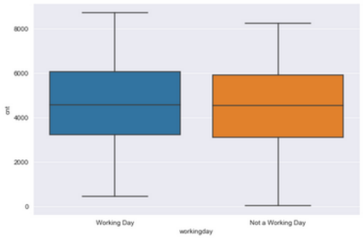
   7. ***year***: Year 2019 has greater number of bookings as compared to 2018 with alomost north of 62% of the total bookings in 2 years.

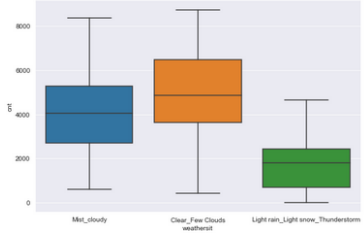
   Below are the box plots and pie charts for these categorical features:

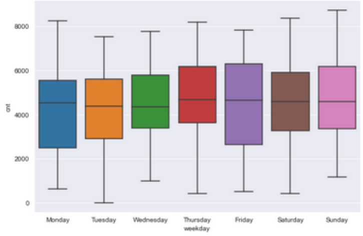
Box Plots:

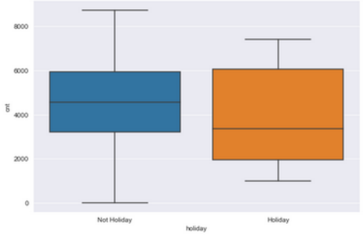


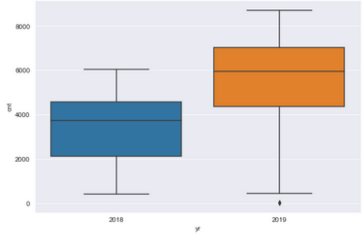




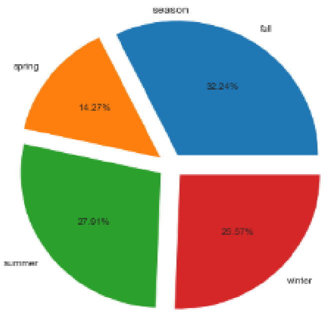


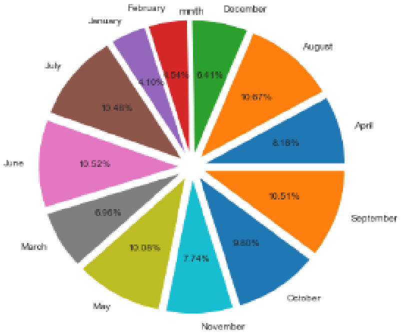


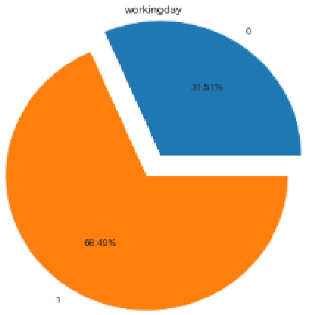


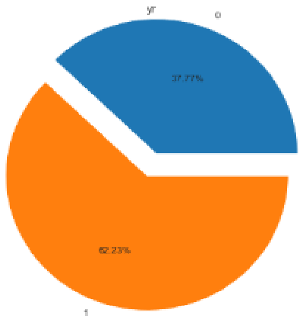


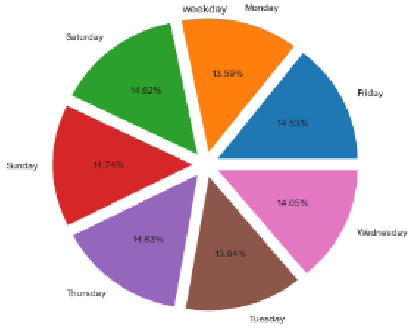
Pie Charts:

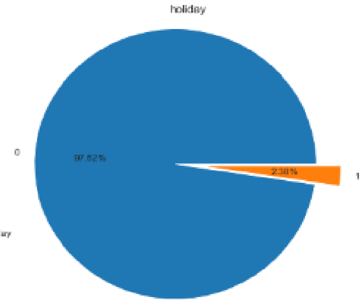


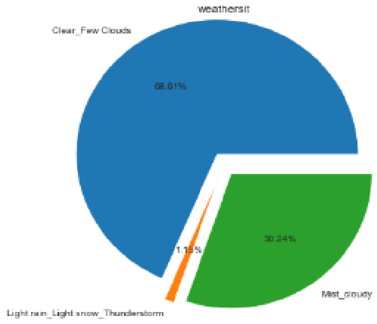












2. Why is it important to use **drop_first=True** during dummy variable creation?

   We create dummy variables to do one hot encoding for categorical features so that these can be used as predictors in Linear Regression models. If we don't drop the first column then the dummy variables will be correlated (redundant) and could eventually lead to multicollinearity in the model. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For instance, iterative models may have trouble converging and lists of variable importance's may be distorted. To keep this under control, we lose drop the first column.

   For example, if we have a column 'gender' which has two values either male or female. Now, when we create dummy variables for the column gender, we'll get something like below:

gender_male	gender_female
0	1
1	0
0	1

   It is understood from the above table that, where the column 'gender_male' is 1, that person is male and where the column 'gender_female' is 1, that person is female. However, we don't actually require two columns to represent this information and it could actually be depicted using one column as below:

gender_male
0
1
0

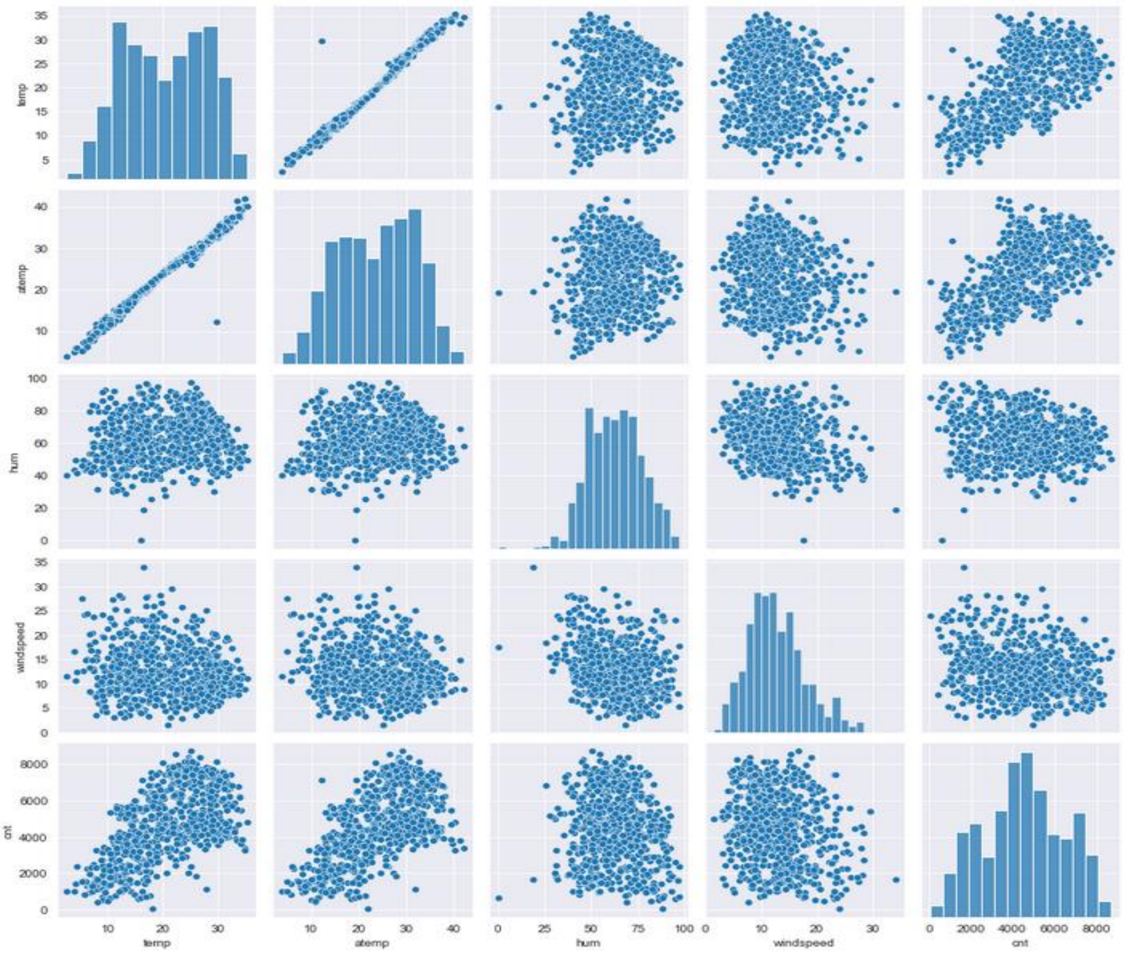
   We can clearly interpret from the above table that, where 'gender_male' column is 1, that person is male and where it is 0, that person is female. Hence, we could represent the information using only one column. Therefore, to avoid the problems mentioned above we drop the first variable while creating dummy variables.

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

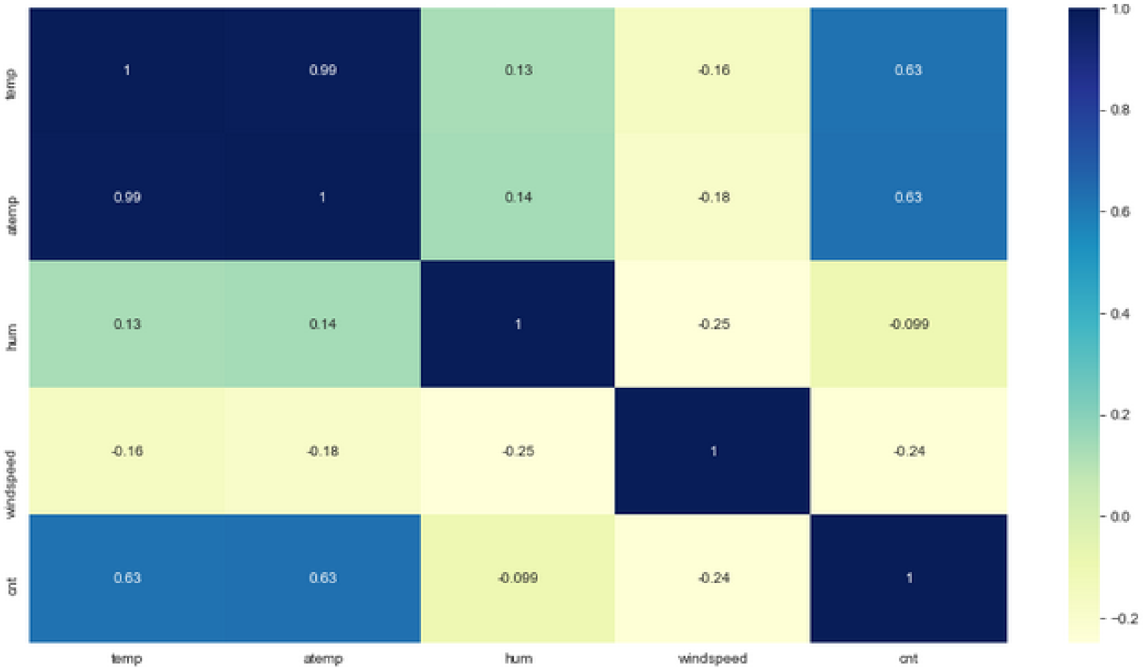
   These are the numerical variables in our dataset - 'temp', 'atemp', 'hum', 'windspeed'. I plotted the pair plot and correlation heatmap for these features and observed that 'temp' and 'atemp' have a strong positive correlation with the target variable 'cnt' with a value of around 0.63.

   Below are the screenshots for the pair plots and correlation heatmap.

Pair Plots:



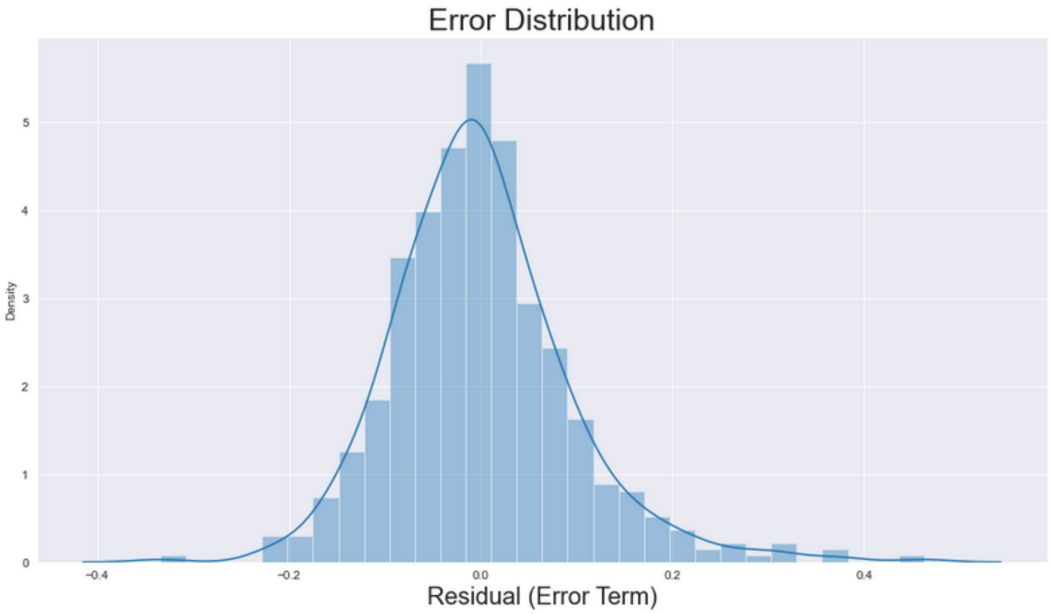
Correlation Heatmap:



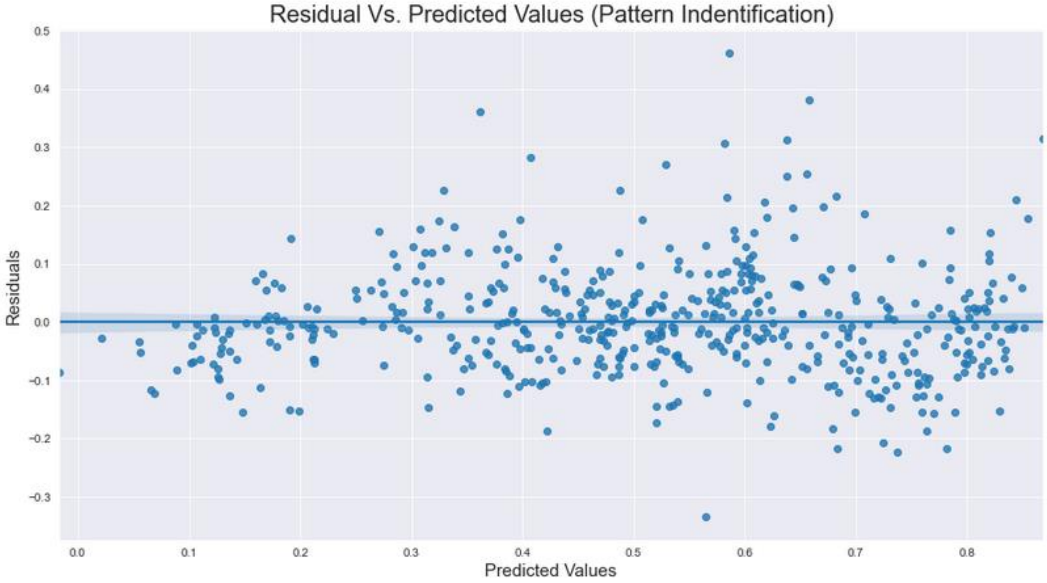
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The following assumptions were validated after building the model:

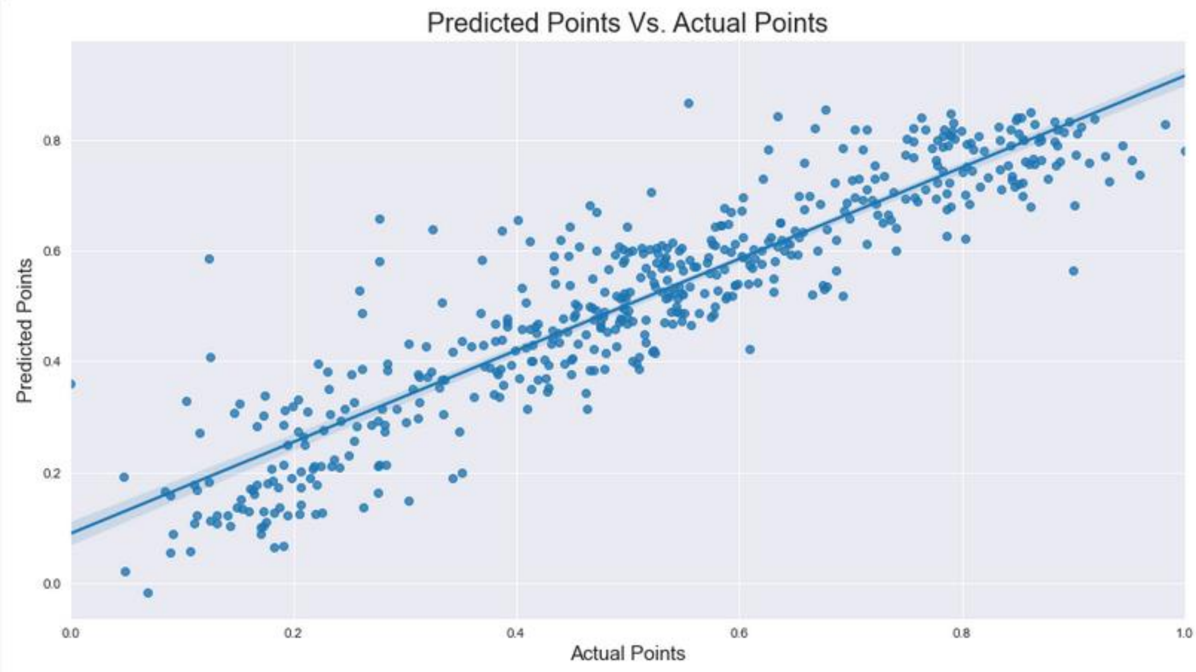
1. ***Distribution of Error Terms*** - Residuals should follow normal distribution and should be centered about 0(mean = 0). We validated this assumption about residuals by plotting a distribution plot of residuals and see if residuals are following normal distribution or not. The below diagram shows that the residuals are distributed about mean = 0



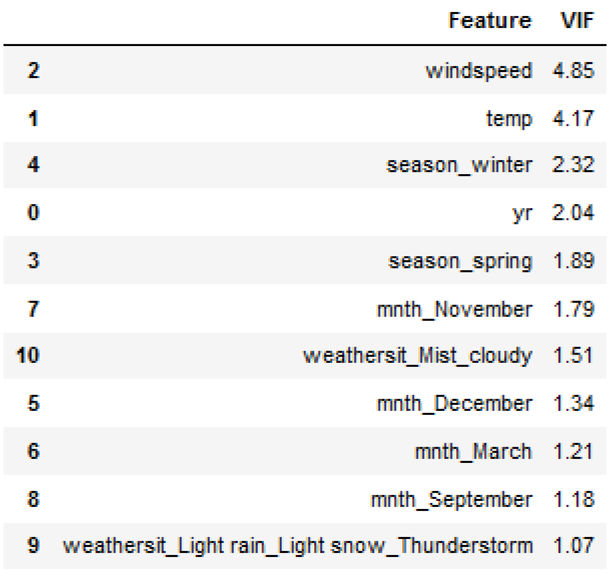
2. ***Error Terms are Independent to Each Other*** – I plotted a scatter plot between the residuals and predicted values for the training data set and we can observe that there is almost no relation (any identifiable pattern) between residuals and predicted values. This is what we had expected from our model to have no specific pattern. Additionally, the personr value 1.061650767297806e-15 i.e. ~0 indicates that there is no relation between the residuals and the predicted values.



3. ***Homoscedasticity*** – I plotted a scatter plot between predicted values and the actual values and from the graph below we can say that residuals are equally distributed across predicted value. This means we see equal variance and we do NOT observe either high or low concentration of data points in certain regions.



4. ***Multicorrelation*** – Checked the VIF for all the predictors and it is less than 5. Hence, there is very low (insignificant) multicollinearity between the predictors.



	Feature	VIF
2	windspeed	4.85
1	temp	4.17
4	season_winter	2.32
0	yr	2.04
3	season_spring	1.89
7	mnth_November	1.79
10	weathersit_Mist_cloudy	1.51
5	mnth_December	1.34
6	mnth_March	1.21
8	mnth_September	1.18
9	weathersit_Light rain_Light snow_Thunderstorm	1.07

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

   The top 3 features are:

   1. ***temp*** - A coefficient value of ‘0.394242’ indicates that a unit increase in temp variable, increases the bike hire numbers by 0.394242 units.
   2. ***yr*** - A coefficient value of ‘0.233564’ indicates that a unit increase in yr variable, increases the bike hire numbers by 0.233564 units.
   3. ***weathersit_Light rain_Light snow_Thunderstorm***: A coefficient value of ‘-0.314140’ indicates that a unit increase in weathersit_Light rain_Light snow_Thunderstorm variable, decreases the bike hire numbers by 0.314140 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is a type of supervised Machine Learning algorithm that is generally used for the prediction of continuous numeric values. It is one of the very basic forms of machine learning where we train a model to predict the behavior of target/dependent variable based on some independent/predictor variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Linear Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

   Mathematically, we can write a linear regression equation as:

   $$y = a + bx, \text{ where:}$$

   b = Slope of the line
   a = y-intercept of the line
   x = Independent variable from dataset
   y = Dependent variable from dataset

   Linear Regression is broadly divided into simple linear regression and multiple linear regression.

   1. ***Simple Linear Regression***: SLR is used when the dependent variable is predicted using only one independent variable. The equation for simple linear regression is same as above.
   2. ***Multiple Linear Regression***: MLR is used when the dependent variable is predicted using multiple independent variables. The equation for multiple linear regression is as below:

   $$Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon, \text{ where:}$$

   $\beta1$ = coefficient for X1 variable
   $\beta2$ = coefficient for X2 variable
   $\beta3$ = coefficient for X3 variable and so on...
   $\beta0$ is the intercept (constant term)
   ϵ = Error Term

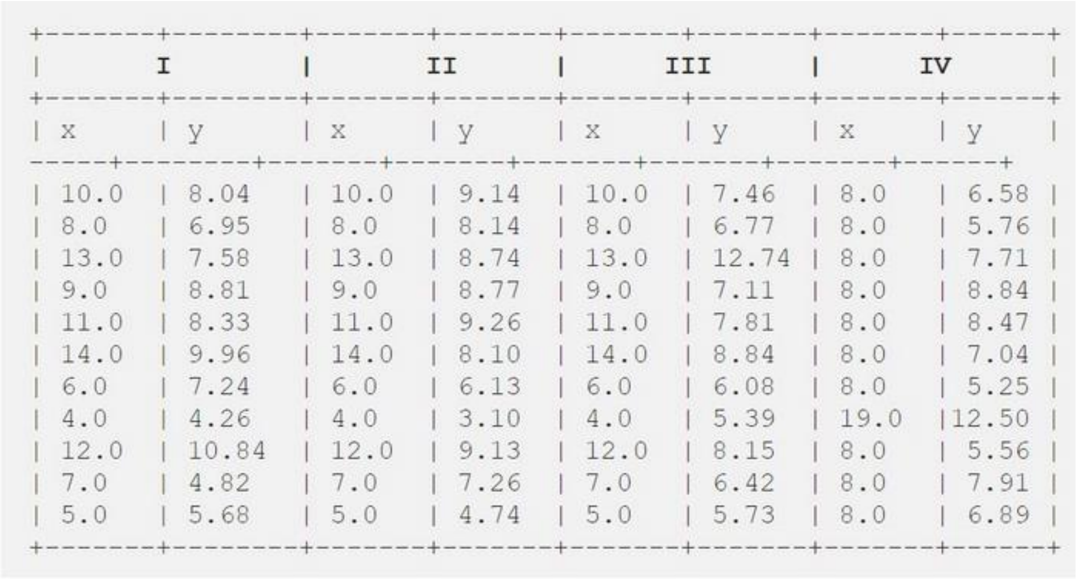
   There following assumptions are associated with a linear regression model:

   1. ***Linearity***: The relationship between X and the mean of Y is linear.
   2. ***Homoscedasticity***: The variance of residual is the same for any value of X.
   3. ***Independence***: Predictor variables are independent of each other.
   4. ***Normality***: The error terms (residuals) are normally distributed.

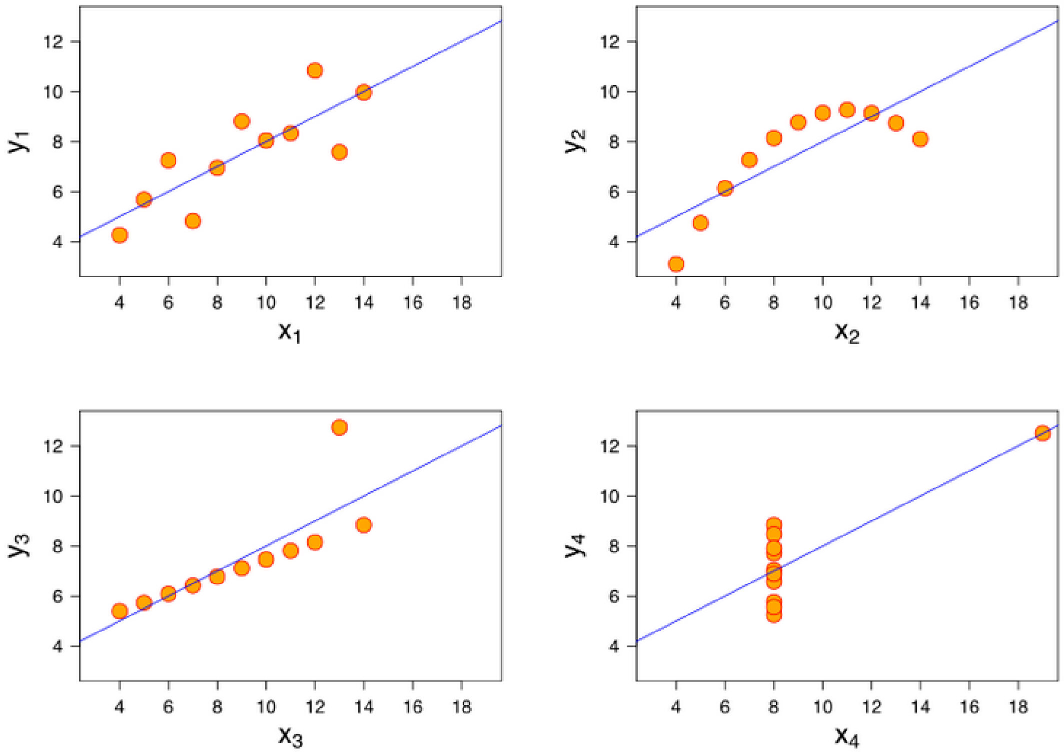
2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Below are the screenshots of the dataset and the corresponding graphs:



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



1. The first scatter plot (top left) appears to be a simple linear relationship.
2. The second graph (top right) is not distributed normally; while there is a relation between them, it’s not linear.
3. In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient, or the bivariate correlation, 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 deviations; thus 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

Pearson correlation coefficient formula:

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

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

   Scaling is a step applied to independent variables during the data pre-processing stage to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. If scaling is not done then algorithm only takes magnitude in account and not units and hence, results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

   Now, let's discuss about normalized and standardized scaling and their differences:

   Normalized Scaling – It is generally used when we know that the distribution of dataset does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors or Neural Networks.

   Standardized Scaling – This can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true.

   The below are the differences between the two:

Normalization	Standardization
Minimum and maximum value of features are used for scaling.	Mean and standard deviation value of features are used for scaling.
It is used when features are of different units.	It is used when we want to ensure zero mean and unit standard deviation.
Scaled values lie between [0, 1] or [-1, 1].	It does not have a bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.

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

   VIF (Variance Inflation Factor) – VIF basically is a measure of the amount of multicollinearity in a set of multiple regression variables. It is calculated using the below formula:

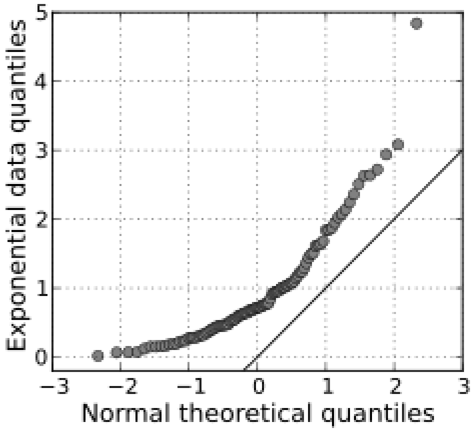
$$VIF = 1/ (1-R2)$$

   If there is perfect correlation between two independent variables, then we get R2 = 1. Hence, using the formula above, we get VIF = 1/0 = infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   A Q-Q plot is basically a scatter plot of quantiles of two datasets against each other. The purpose of Q-Q plots is to find out whether two sets of data come from the same distribution or not. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

   Below is Q-Q plot showing a 45 degrees reference line:



   If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.