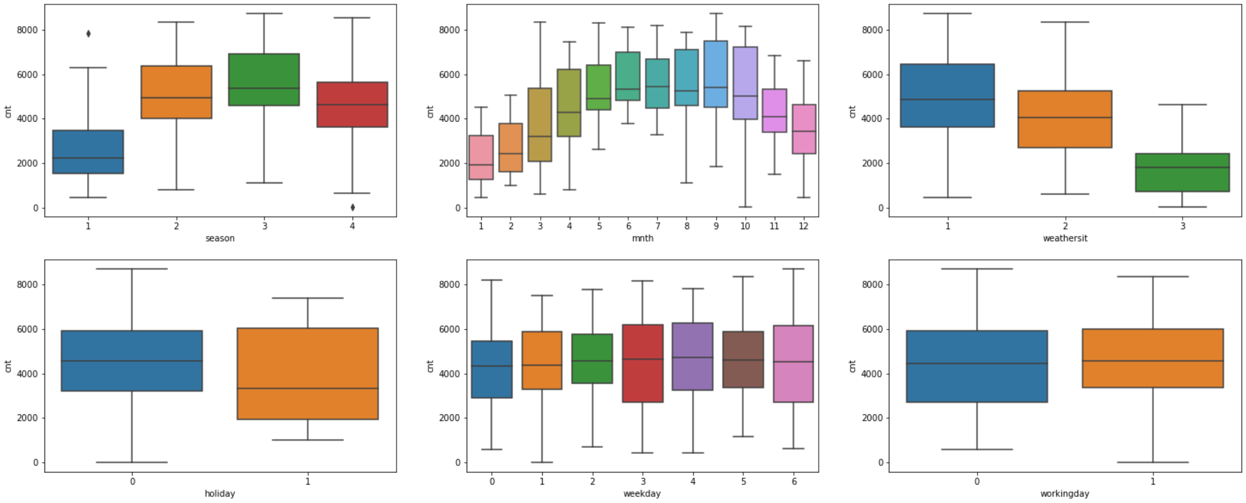
Assignment based - Subjective Questions:

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

Ans:

- In the dataset, columns ‘season’, ‘mnth’, ‘weathersit’, ‘holiday’, ‘weekday’, ‘workingday’ are the categorical variables.
- Variables descriptions are:
  - mnth: month of the bike booked by user
  - weathersit: Weather situation when user has booked the bike
  - holiday: wether the booking day is holiday or not
  - weekday: Day of the week when user booked the bike
  - workingday: States wether the bike booked day is working day or not.
- Lets have a look at the Box plots for the above variables so, that we can provide some information how its effect the dependent variable(cnt).



- We can able to make the following insights from the above box plots wrt to target variable 'cnt'

1. Working day:

- Almost 69% of users books bike in working day which is closes to 5000
- This indicates, workingday can be a good predictor for the dependent variable

2. Weekday:

- Almost all weekdays, the no.of bike users count was similar and it is around in between 3000-6000
- Medians of all the weekdays are around in between the 4000-6000 that means, more than 50% of people using bikes in all days of a week irrespective of the day of the week.

- Difference/distance between the 25% and 75% of box is more for weekdays 3(Wednesday) & 6(Saturday) but not a significant difference when compared with others [Considering start of week as Sunday]

3. Holiday:

- Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

4. Weathersit:

- Most of the bike users are in weathersit 1(Clear, Few clouds, Partly cloudy, Partly cloudy), followed by 2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) which is almost 67% of users.
- This was followed by weathersit2 with 30% of total booking
- Very less no.of bike users are available in weathersit 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).
- This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

5. Month:

- Most of the bookings are happening around the months 6, 7, 8, 9 (more than 5000 bookings are happening. Almost 10%)
- Where as months 1 & 2 are having less bookings (Less than 3000)
- In almost all months the differenct between the 25% to 75% is similar but for months 3,4,9,10 is having significantly more difference
- This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

6. Season:

- For season 3 having more no.of bike users (More than or equal to 5000 users. Almost 32%) followed by season2 & season4 with 27% & 25% (greater than 4000 and near to 5000)
- Season 1 is having less users which is less than 3000
- Almost all seasons are having the difference between 25% and 75% is significantly having no difference among them.
- This indicates, season can be a good predictor for the dependent variable.

---

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

Ans:

- While we are making predictions using regression model. Model predictors should be of numeric type. So, we convert categorical variables to dummy variables

For example: Lets have a simple example of categorical variable **Gender.** Which will have two possible values Male, Female.

Lets assume ‘1’ as Male and Female as ‘0’. Then its obviously converts data from categorical to numeric. It looks as follows

	Male	**Female**
Male	1	0
Female	0	1

Now, if we observe above data we can see that, knowing one variable can interpret the other variable easily. This is nothing but reducing the level of the categorical variable values.

Lets, see how it looks now

	Male
Male	1
Female	0

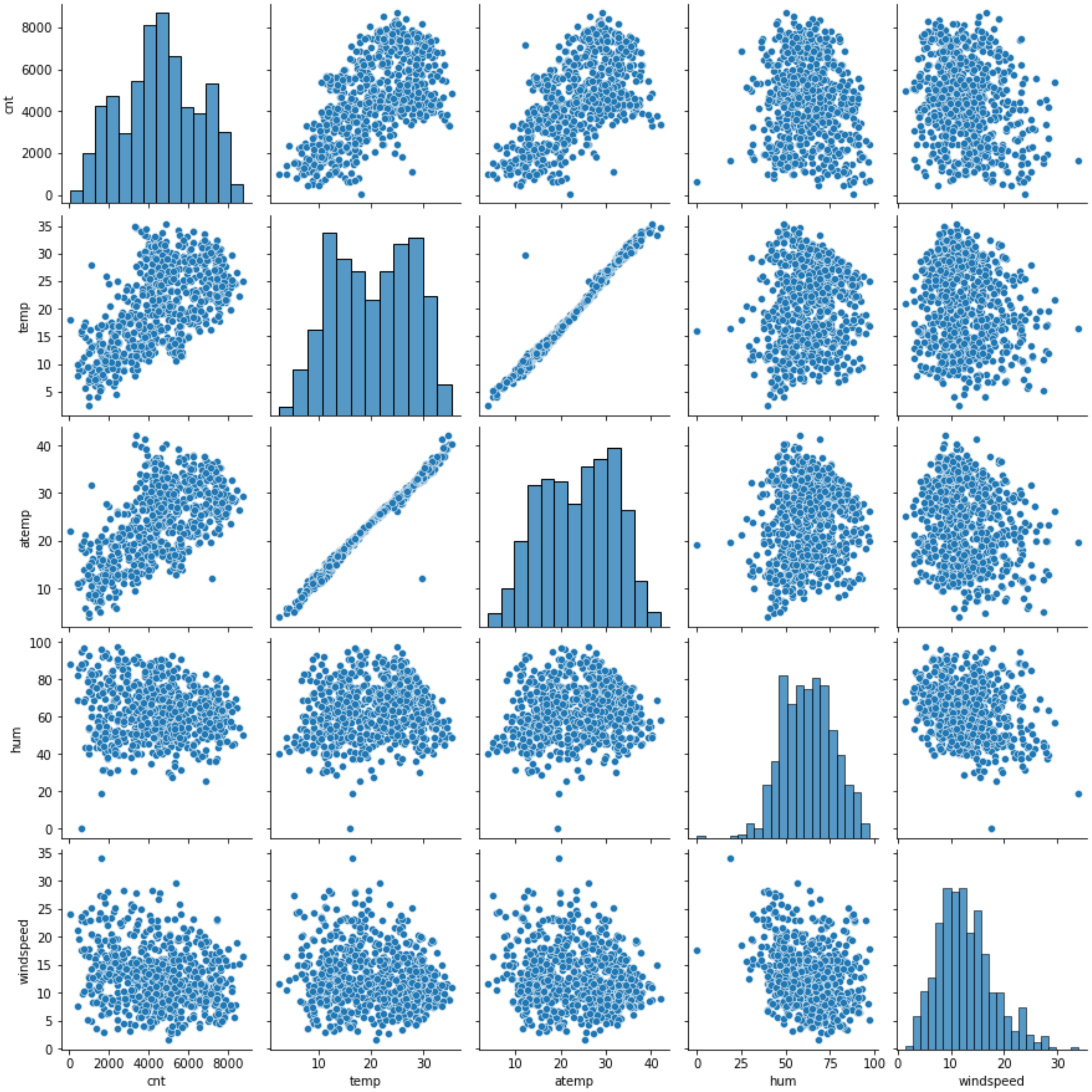
From the above table, by using only ‘Male’ as a variable we can predict ‘Female’ as well. Because if user is not Male i.e if value is not 1 then it is obviously 0 (Female). In this fashion we can reduce the level of the categorical variable and make dummies.

- Generally, if the categorical variable having the **‘n’** levels then dummy variables will be having **‘n-1’** levels in it.
- Pandas have the inbuilt method to convert the column into dummy variables. Its syntax is as follows
  pd.get_dummies(data, prefix=None, prefix_sep=’_’, dummy_na=False, columns=None, sparse=False, drop_first=False, dtype=None)
  Lets see what each input element tells us:
  - Data: Data frame object which Is having the categorical variables
  - Prefix: List of names/ name that needs to be prefixed with the dummy column names after conversion
  - Prefix_sep: If we append prefix to the columns we needs to use separator for column name and prefix.
  - Dummy_na: If there are any column with null data then if we need to create any separate to mention it as Null then we set this property as True. Default it is True.
  - Columns: Columns that needs to be dummied in data frame. If nothing is mentioned then it bydefault convert category and object data type variables into dummy variables
  - Sparse: wether to make the dummy-encoded columns as numpy array or sparse array.
  - Dtype: Data type to be mentioned for the newly created columns
  - **Drop_first:** It is default False.
    - If it is False, then it creates dummy variables for all k-level of categorical variable.
    - If it is True, then pandas create categorical variable with k-1 level.

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

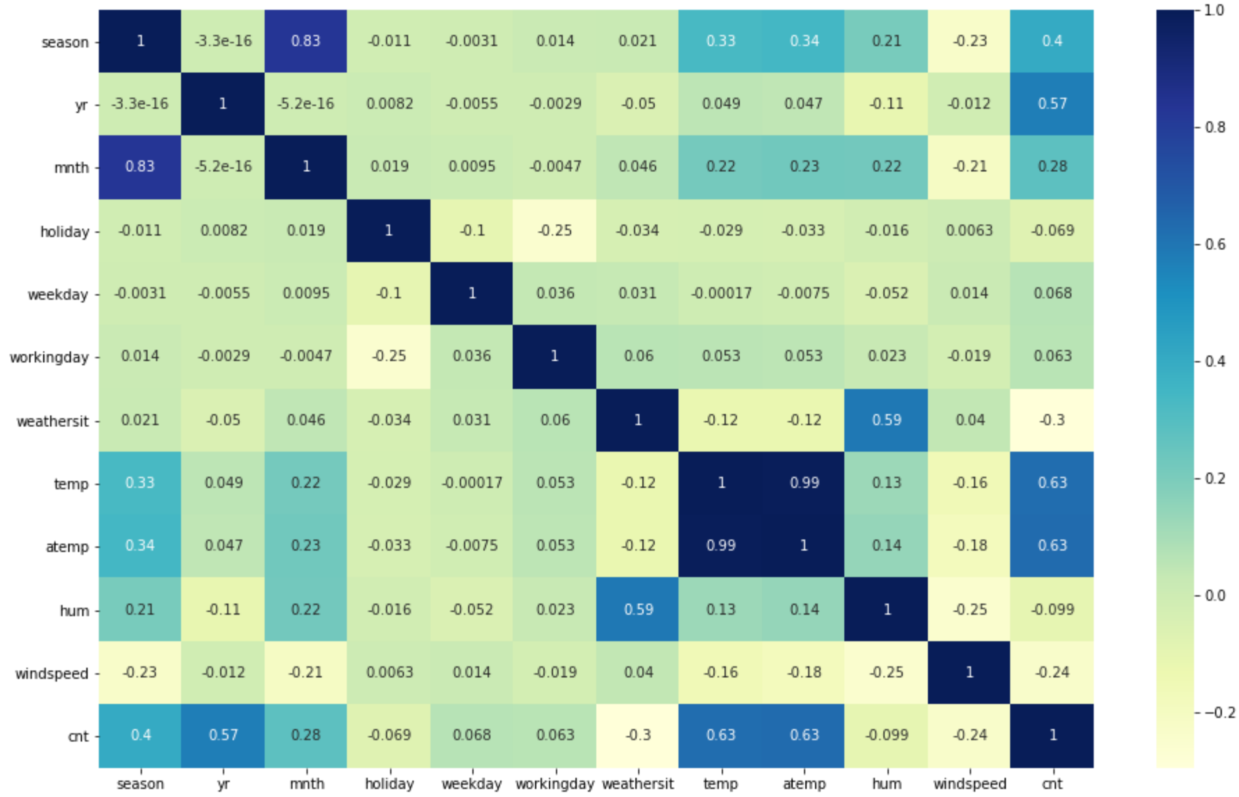
Ans:

Lets have a look at the pair plot among the numeric variables get the details.



From the above pair plot it is clearly visible that target variable(cnt) is having highest correlation with columns **'temp', 'atemp'.**

Lets confirm the same by looking at the heat map.



	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt
season	1	-3.3e-16	0.83	-0.011	-0.0031	0.014	0.021	0.33	0.34	0.21	-0.23	0.4
yr	-3.3e-16	1	-5.2e-16	0.0082	-0.0055	-0.0029	-0.05	0.049	0.047	-0.11	-0.012	0.57
mnth	0.83	-5.2e-16	1	0.019	0.0095	-0.0047	0.046	0.22	0.23	0.22	-0.21	0.28
holiday	-0.011	0.0082	0.019	1	-0.1	-0.25	-0.034	-0.029	-0.033	-0.016	0.0063	-0.069
weekday	-0.0031	-0.0055	0.0095	-0.1	1	0.036	0.031	-0.00017	-0.0075	-0.052	0.014	0.068
workingday	0.014	-0.0029	-0.0047	-0.25	0.036	1	0.06	0.053	0.053	0.023	-0.019	0.063
weathersit	0.021	-0.05	0.046	-0.034	0.031	0.06	1	-0.12	-0.12	0.59	0.04	-0.3
temp	0.33	0.049	0.22	-0.029	-0.00017	0.053	-0.12	1	0.99	0.13	-0.16	0.63
atemp	0.34	0.047	0.23	-0.033	-0.0075	0.053	-0.12	0.99	1	0.14	-0.18	0.63
hum	0.21	-0.11	0.22	-0.016	-0.052	0.023	0.59	0.13	0.14	1	-0.25	-0.099
windspeed	-0.23	-0.012	-0.21	0.0063	0.014	-0.019	0.04	-0.16	-0.18	-0.25	1	-0.24
cnt	0.4	0.57	0.28	-0.069	0.068	0.063	-0.3	0.63	0.63	-0.099	-0.24	1

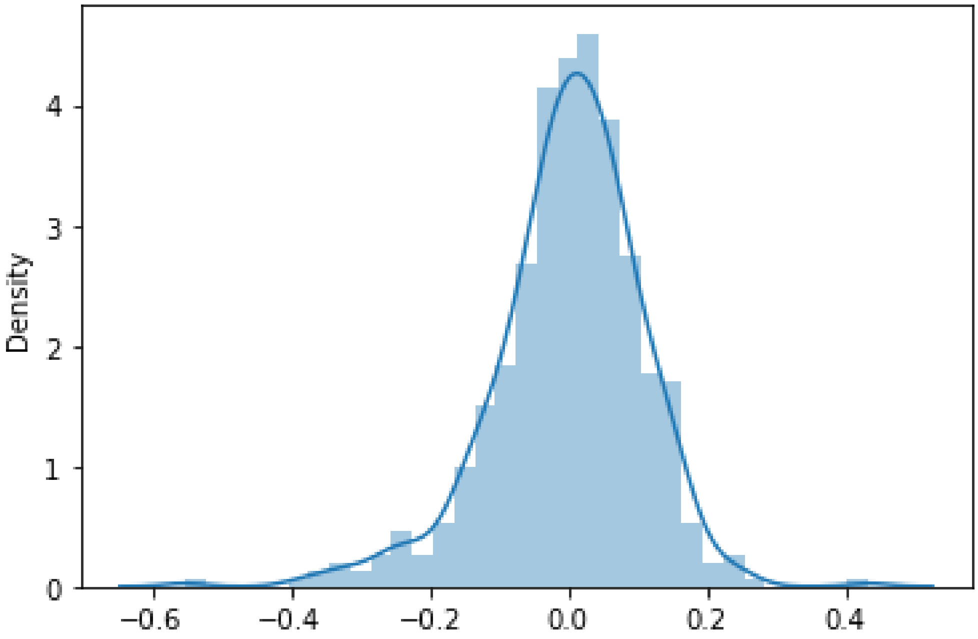
- From the above heatmap it is proved that target variable(cnt) is having highest correlation with variables **‘temp’, ‘atemp’**.

---

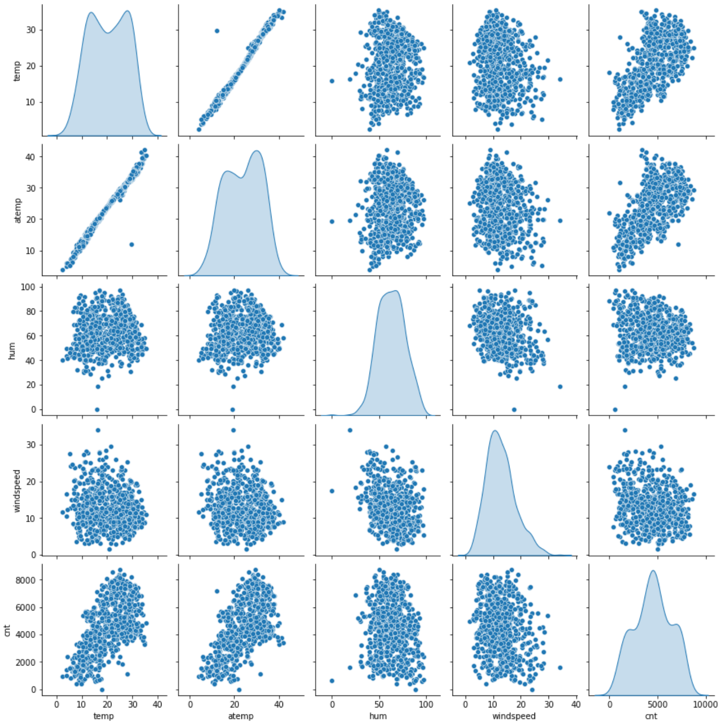
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

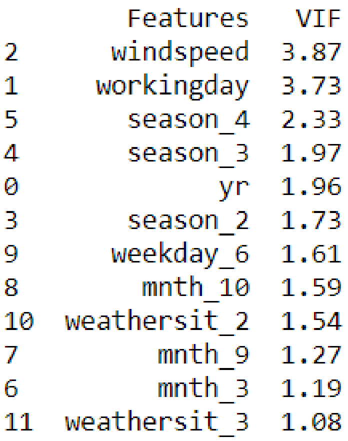
- After the training the model, we will validate the assumptions as follows:
  - **Error terms are normally distributed with mean 0 (not x,y)**
    - To prove it we can make a residual analysis using the train data set.
    - Initially lets create the y_predicted data using the final regression model using the X_train data.
    - Find the difference between the actual y_train and the y_predicted data. And that difference is called the residual.
    - On plotting the residuals we can observe that residuals are normally distributed and having the mean around 0.



- **There is a linear relationship between X & Y**
    - Without having atleast one variable that is linearly correlation with Y then we cannot create a Regression model
    - So, by plotting the pairplot in between the numeric variables and the target variable we can observe wether we are having the linear relationship between X & Y



        - From the above graph we can see that cnt is having linear relationship with temp and atemp.
    - **There is no multicollinearity between the predictor variables**
        - In the model if the predictor variables are having the multicollinearity between them, then the model will not be efficient.
        - To prove that we use VIF to calculate the multicollinearity among the predictor variables
        - If any variable having the VIF greater than 5, we can say that predictors are having the multicollinearity among them
        - Following are the VIF values of the predictors



```
        Features   VIF
2      windspeed  3.87
1     workingday  3.73
5       season_4  2.33
4       season_3  1.97
0             yr  1.96
3       season_2  1.73
9      weekday_6  1.61
8        mnth_10  1.59
10  weathersit_2  1.54
7         mnth_9  1.27
6         mnth_3  1.19
11  weathersit_3  1.08
```

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

Ans:

According to the final model occurred to me. Top 3 features contributing significantly towards explaining the demand of the shared bikes are

`season_3` , `season_2`, `yr` with its coefficients 0.316649, 0.268363, 0.243876 respectively.

General Subjective Questions

1. Explain the linear regression algorithm in detail.?

Ans:

- Regression analysis is a technique of predictive modelling that helps us to find the relationship between the input and the target variable.
- Linear regression is a part of the Regression analysis technique.

Linear Regression:

- Linear regression is a machine learning algorithm based on the supervised learning.
- In this linear regression we train a model to predict the behavior of our data on some variables.
- In In this linear regression the two variables which are on x-axis and y-axis should be linearly correlated.
- In linear regression model we will try to create a straight line which can predict the Y(Target) value with the help of the predictive/input variables.

Mathematically we write this equation as follows:

$Y = mx + c$

where

y -> Target Variable

x -> Input/Predictor variable

m -> Slope of the line.

c -> Interceptor.

Slope: It tells us the relationship between the x & y variables. Like, if x increases by 1 unit then y will be increased by m units.

Mathematically it can be written as

$m = \tan(\theta)$ => (change in Y/change in X)

Intercept: It tells us the point where the line intersects the y-axis. If a line is passing through the Origin then the intercept will be 0.

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector

Best Fit Line:

- For any given data, if we plot data points in the graph it will be scattered in the graph.
- Through the scattered points, we can draw n number of lines among them and each line will be having its own slope and intercept. Among those lines we need to find the best fit line such that slope and coefficients are optimal.
- That best fitted line is our regression model, where we can predict the target(y) variable for any future input(x)
- To find the best fit line we use a technique called **"Residuals".**
    - **Residual:** Residual is actually an error in our prediction model. That error will be represented as **'e'.** And the error is actually the difference between the actual 'Y' value and the predicted 'Y' value.
- For each and every line this residuals will be varied and in each line each y-value will be having its own residual error value.
- We will get the best line, such that whose **Residual Sum of Squares**(RSS) are minimum and this technique is called the **Ordinary Least Squares**(OLS) method

Assumptions of Linear Regression: We make some assumptions to work on the linear regression model, since we are making the model based on the sample data from the population but not on the whole population data.

- There is a linear relationship between the X & Y.
- Error terms(Residuals) are normally distributed (not X, Y).
- Error terms are independent of each other.
- Error terms have the constant variance.

R Square:

- R Square is a measurement used to determine the percentage of variance of the data in the predicted model using the predictors used in the model.
- It helps us to tell the strength of the linear regression model.
- If R-Square is less then data points are scatter in the plot vigorously among the line and the predicted model will not be more accurate.
- If R-Square is high then the data points are not much scattered among the line and the model is more accurate in the predictions
- If R-Square value is 1, then all the data points are lies on the predicted straight line.

Steps to follow to create Linear Regression model using Python:

- In python we use statsmodels api and the sciket learn modules to create the linear regression model.
- Let us see the steps for creating model:
    - Reading the data set and understanding it
        - We check for null data if any
        - We check for any cleaning required or not
    - Visualizing the data
        - We use scatter plots, box plots, heatmap to visualize the data.
        - Scatter plot and heat maps tells us the correlation among the X & Y values
        - Box plots helps in understanding the categorical variables
    - Preparing data for model
        - Here we create any new variables if required for the model
        - Here we change the categorical data into numeric data by creating them as dummy variables.
    - Splitting the data into training and test sets
        - Lets us assume the target variable and the rest as the predictor variables from data set
        - Split the data into training and testing data sets as 70-30 or 80-20 percentages.
    - Rescale the features
        - We can use either Normalization or standardization for rescaling.
        - We need to rescale the data so that, all the columns data will be significantly in same scale and prediction will be more accurate.
    - Train the model
        - Here we train the model using training data set
        - In every model building we analyze the data using R-Squares and the VIF Values and eliminate the features which are not significant(Using p=values)
        - We again rebuild model with the remaining columns and rebuilds the model again
        - We follow this procedure until we finds our model is having significant predictors and predictors with low VIF values.
    - Evaluating the model using test set
        - Once, the final model is obtained, use the same model on test data set and predict the y-values
        - Obtains the r-square for the test data set and compare it with training data set r-square value. There should not be any much significant difference among them.
- The final model provides us the details of m, c. using that we can predict the y-values using the future input x variables.

---

2. Explain the Anscombe's quartet in detail.

Ans:

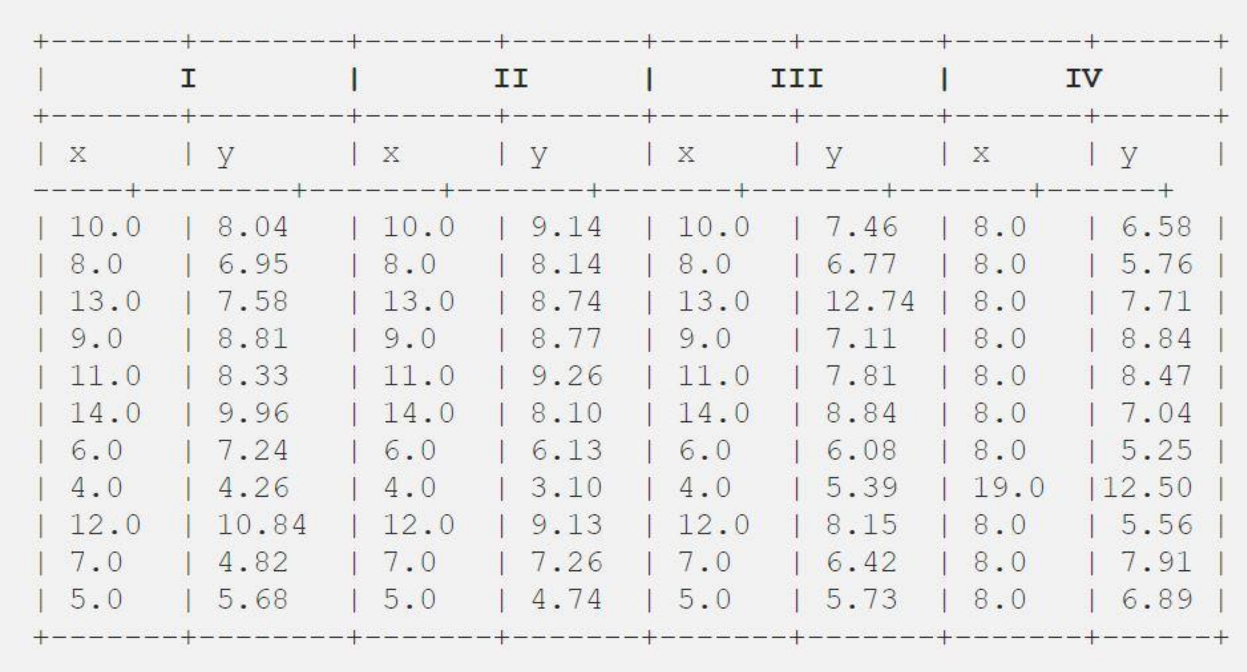
- Anscombe's quartet comprises of four data sets that have nearly identical simple statistical properties, yet appear very different when graphed.

- Each data set consists of eleven (x,y) points, which 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.

History of Anscombe's quartet:

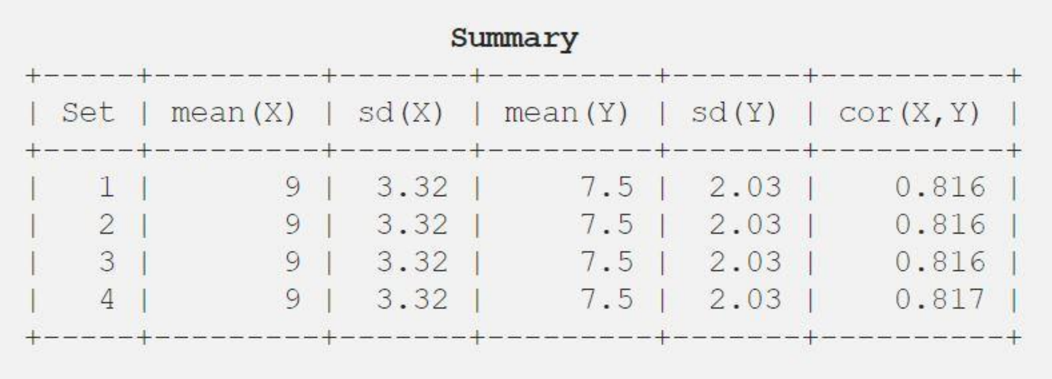
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.

Those data is as 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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

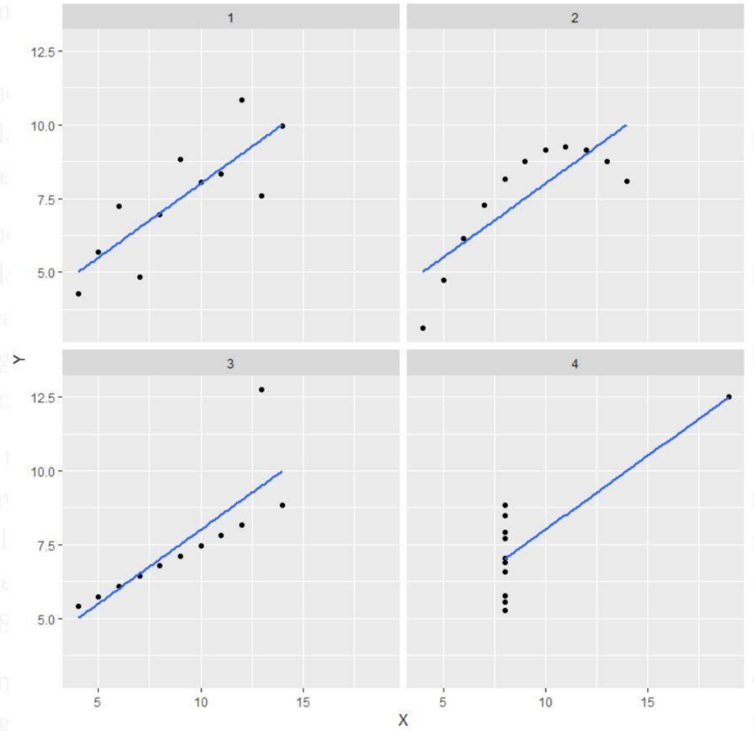
Council analyzed the data and provided the following information of mean, standard deviation, correlation between x & y as follows



```
                         Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

- On plotting the scatter plots among the data sets they have observed that, they have identical statistical properties but appeared in different way when graphed.

- **Graph1:** Looks like scatter plot but having the linear relationship between x & y
- **Graph2:** Non-linear relationship between x & y
- **Graph3:** Perfect linear relationship between all the data except for some who are outliers.
- **Graph4:** Tells us one data point is enough to produce a high correlation coefficient.



In simple, this quartlets tells us that, data having similar means, standard deviation, correlation varies when they are plotted in graph and gives us the un realistic information.

3. What is Pearson's R?

Ans:

- Correlation between the sets of data is a measure of how well they are related. One of the common measure of the correlation is **Pearson Correlation**
- Full name is **Pearson Product Moment Correlation(PPMC)**
- It shows the linear relationship between two sets of data. In simple it tells us weather we can draw a straight line in the graph to represent the data.
- For population data we use rho (ρ) to represent it, where as 'r' for a sample data.
- PPMC always varies in between -1 to +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

Disadvantages:

- PPMC is not able to tell the difference between the dependent and independent variables.
- It doesn't provide any information about the slope of the line, it just suggests weather we can draw a straight line or not.

Formulae:

$$\rho_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

where:

$\rho_{xy} =$ Pearson product-moment correlation coefficient

$\text{Cov}(x, y) =$ covariance of variables x and y

$\sigma_x =$ standard deviation of x

$\sigma_y =$ standard deviation of y

---

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

Ans:

Scaling & Its Importance:

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values

Example: If an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

Types:

- Min-Max Normalization
- Standardization.

Min-Max Normalization:

- This technique re-scales a feature or observation value with distribution value between 0 and 1.
- In this technique dummy values that are created will not gets effected since, dummy variables are also will be of either 0 or 1.
- In Min-Max Scaling, data at point **'i'** will be scaled to Xnew
- **Formulae:**

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

Where, min(X) is the minimum value of X in that column

max(X) is the maximum value of X in that column

X*i* is the value of X at ith index.

Standardization:

- It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.
- This cannot be used when data set is having dummy variables in it.
- **Formulae:**

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

Where,
X*i* is the value of X at ith Index
Xmean is the mean of the column
Standard Deviation is the standardDeviation calculated for that column.

Differences:

Normalization	Standardization
It is used when features are of different scales	It is used when we want to ensure zero mean & unit Standard Deviation
Scales values between [0,1]/[-1,1]	No certain range
Effected by outliers	Less effected by outliers
It is useful when we don’t know about the data	It is useful when feature distribution is Normal/Gaussian
Called as Scaling Normalization	Called as Z-Score normalization
MinMaxScaler class of Scikit-learner is used for scaling	**StandardScaler** class of Scikit-learner is used for scaling

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

Ans:

VIF: It is a measurement used to determine the multicollinearity between the predictors in the model. If the VIF is greater than 5, it is eligible to drop from the model.

Formulae: VIF = 1/(1-RSquare)

- From the above we can say that, VIF can be infinite if R-Square of a model is 1.0.
- If the R-Square of a model is 1.0 it means that the model generated with the predictors is having the perfect correlation among them.
- An Infinite VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.
- To get rid of from the infinity VIF value, we need to drop one of the predictor which is causing the perfect multicollinearity

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

Ans:

- Q-Q Plot is short form notation of Quantile-Quantile plot.
- It is a graphical tool to assess, if a set of data possibly came from the some theoretical distribution such as Normal, Exponential or Uniform distribution.
- Simply, it is a tool to tell weather two different data sets came from populations with same distribution or not.

Helps in linear regression:

- Since, in liner regression we divide data sets into two different data sets like training and test data set. But, if we get training and testing data set already then to determine weather they came from the populations with same distribution or not we use this Q-Q plot.

It uses to check the following scenarios:

If two data sets -

- Came from populations with same distribution or not
- Have common location and scale
- Have similar distribution shapes
- Have similar tail behavior

- In python, statsmodels.api provide **qqplot and qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.