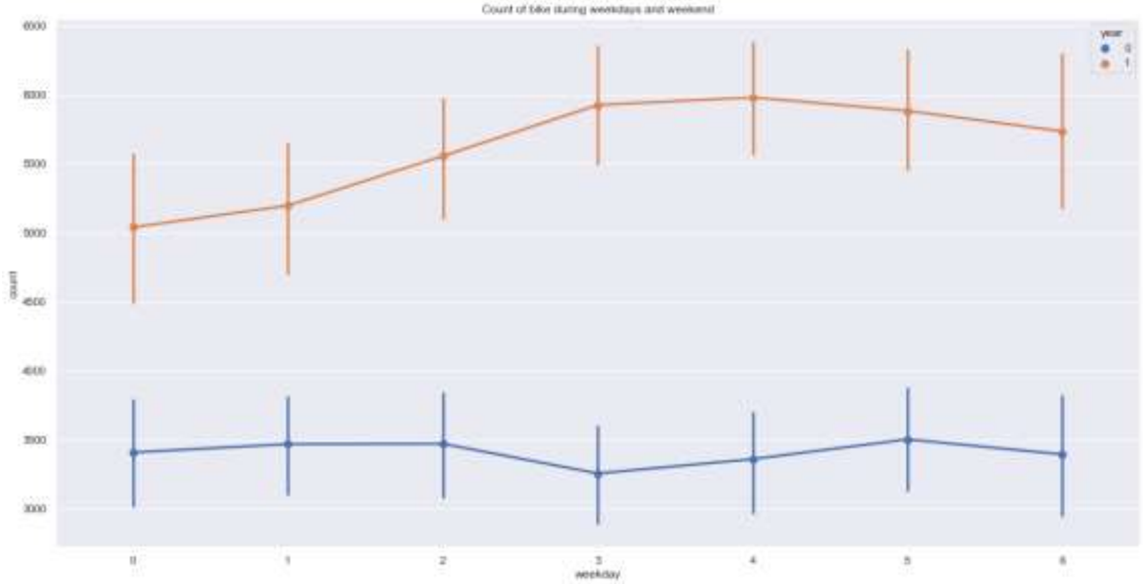
Assignment-based Subjective Questions

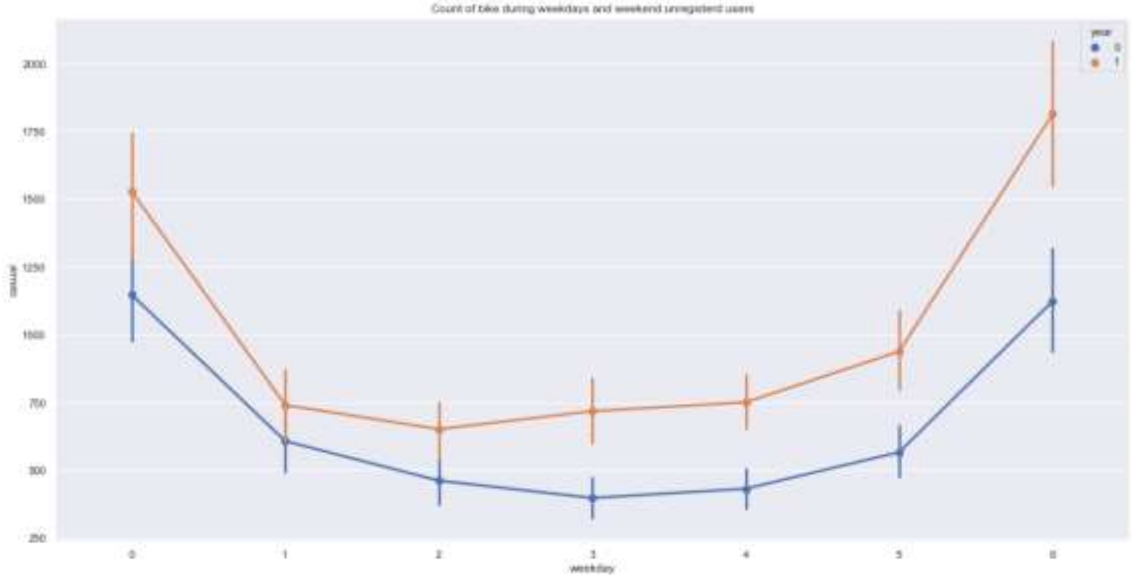
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

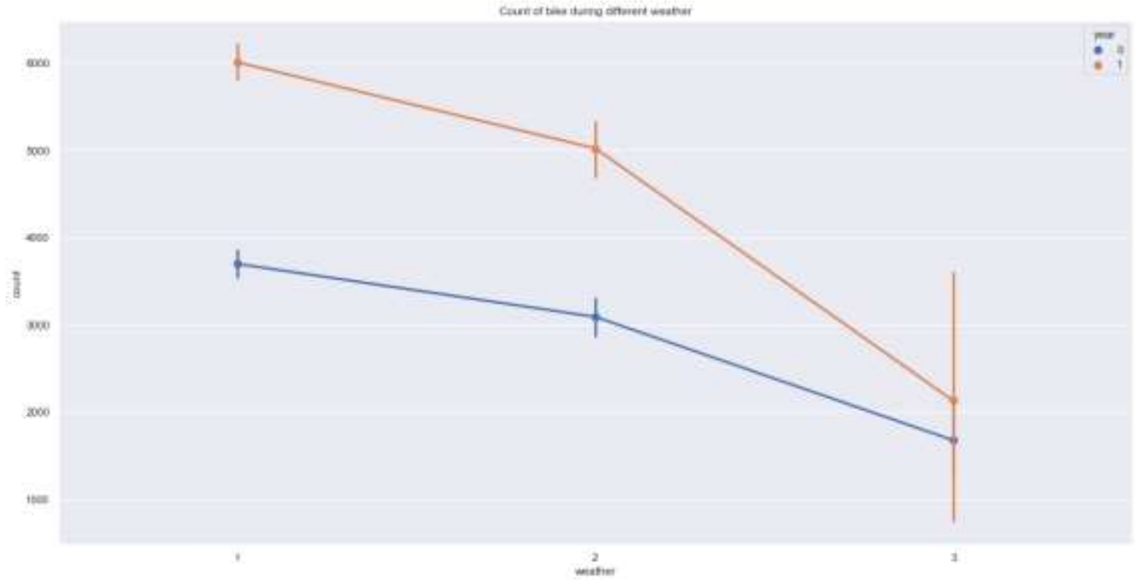
Here are the following Inferences drawn from the data:



From the above plot we observe that in the year 2019 count of bikes booked are more compared to year 2018



From the previous plot we already know the magnitude of people booking is more in year 2019 compared to 2018, but here in the above plot we observed the number of casual users' book more as the week progresses towards the end. i.e., Saturday and Sunday. The traffic is very busy in Boom Bikes



It is clear from the above plot bookings are more when weather is clear

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

Answer:

This is called as dummy variable trap.

This occurs when we create k dummy variables instead of k-1 dummy variables.

When this happens, at least two of the dummy variables will suffer from perfect multicollinearity. That is, they'll be perfectly correlated. This causes incorrect calculations of regression coefficients and their corresponding p-values.

Let's consider categorical variable weathersit (Renamed as weather in code during analysis). The data for this particular column having three values 1,2,3 respectively. If we perform get_dummies on this so we will get below table:

Weathersit_1	Weathersit_2	Weathersit_3
0	0	1
0	1	0
1	0	0

If I delete Weathersit_1 column in that case also I will be able to express the existence of Weathersit_1 when column Weathersit_2 and Weathersit_3 value is 0.

Hence if we have categorical variable with k-levels, then we need to use k-1 columns to represent the dummy variables.

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

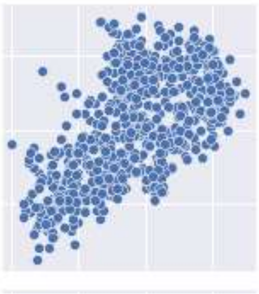
Answer:

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable "cnt" (Renamed as count during analysis)

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

Answer:

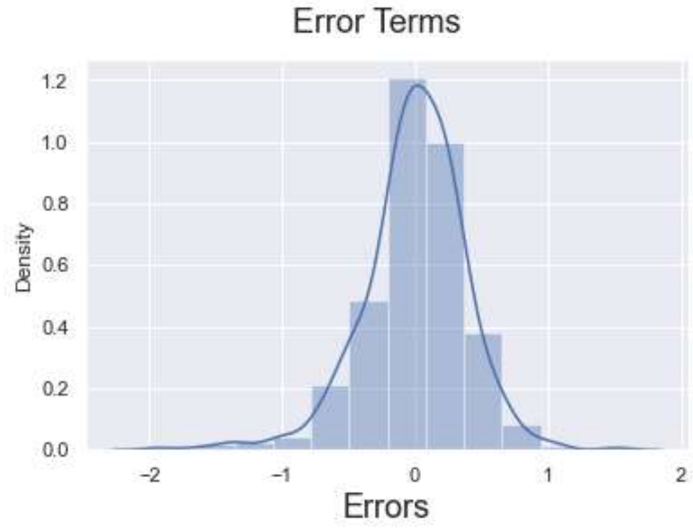
One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables. If you try to fit a linear relationship in a non-linear data set, the proposed algorithm won't capture the trend as a linear graph, resulting in an inefficient model. Thus, it would result in inaccurate predictions.



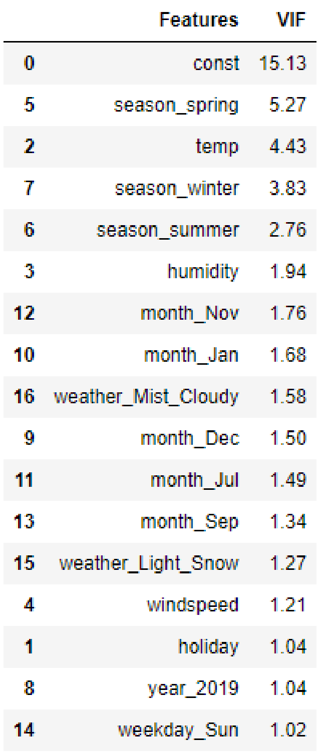
(Temp vs cnt scatter plot)

From the above plot we observe that there is a linear relationship between the dependent variable (temp) and independent variable (cnt)

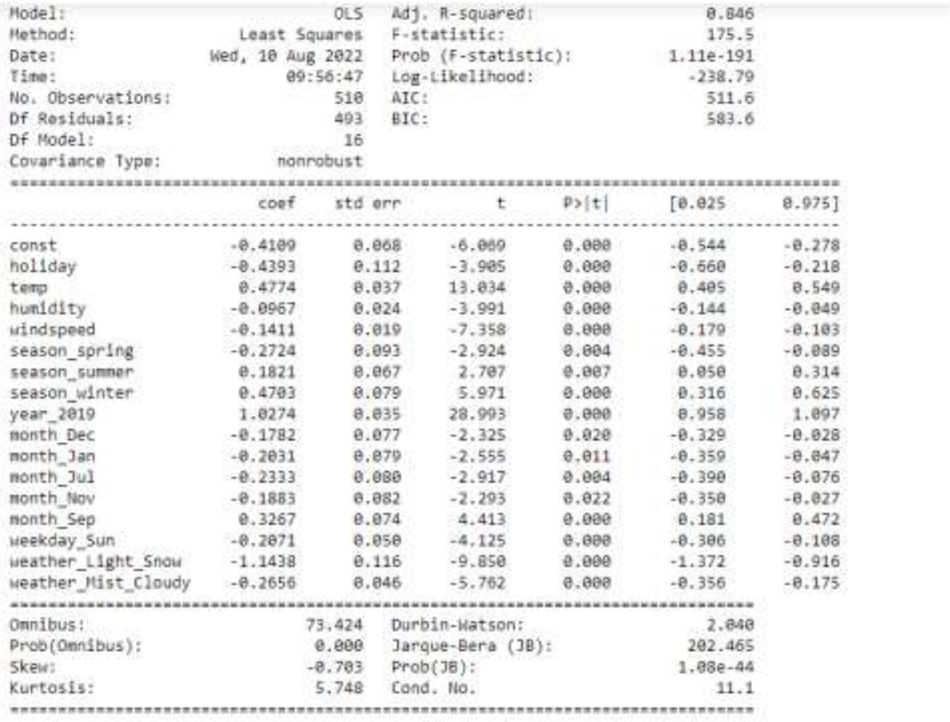
The residuals (error terms) are independent of each other. In other words, there is no correlation between the consecutive error terms of the time series data. The presence of correlation in the error terms drastically reduces the accuracy of the model. If the error terms are correlated, the estimated standard error tries to deflate the true standard error.



The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model. In essence, it is difficult to explain the relationship between the dependent and the independent variables. In other words, it is unclear which independent variables explain the dependent variable.



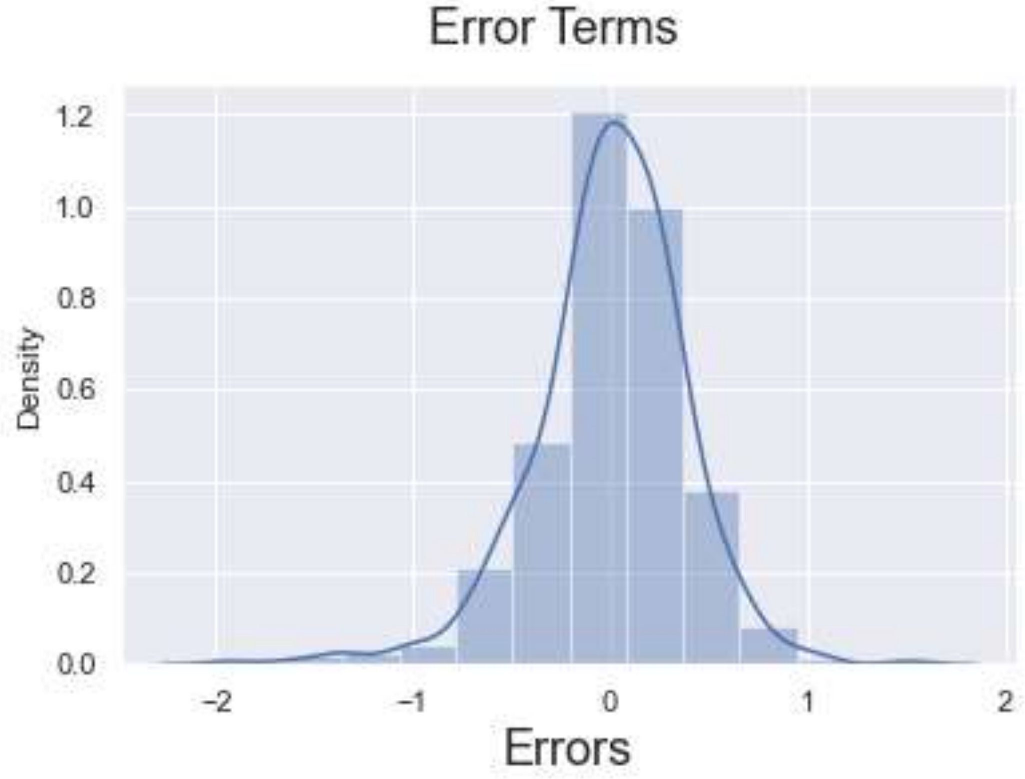
	Features	VIF
0	const	15.13
5	season_spring	5.27
2	temp	4.43
7	season_winter	3.83
6	season_summer	2.76
3	humidity	1.94
12	month_Nov	1.76
10	month_Jan	1.68
16	weather_Mist_Cloudy	1.58
9	month_Dec	1.50
11	month_Jul	1.49
13	month_Sep	1.34
15	weather_Light_Snow	1.27
4	windspeed	1.21
1	holiday	1.04
8	year_2019	1.04
14	weekday_Sun	1.02



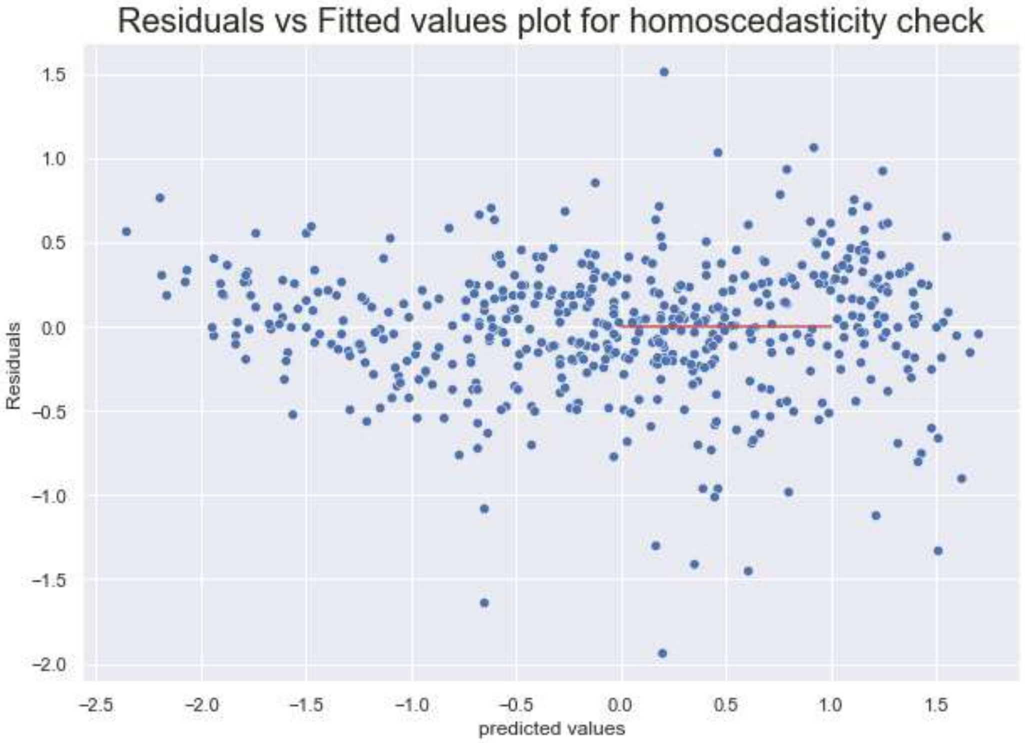
```
Model:                            OLS   Adj. R-squared:                  0.846
Method:                 Least Squares   F-statistic:                     175.5
Date:                Wed, 10 Aug 2022   Prob (F-statistic):          1.11e-191
Time:                        09:56:47   Log-Likelihood:                -238.79
No. Observations:                 510   AIC:                             511.6
Df Residuals:                     493   BIC:                             583.6
Df Model:                          16
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  -0.4109      0.068     -6.069      0.000      -0.544      -0.278
holiday                -0.4393      0.112     -3.905      0.000      -0.660      -0.218
temp                    0.4774      0.037     13.034      0.000       0.405       0.549
humidity               -0.0967      0.024     -3.991      0.000      -0.144      -0.049
windspeed              -0.1411      0.019     -7.358      0.000      -0.179      -0.103
season_spring          -0.2724      0.093     -2.924      0.004      -0.455      -0.089
season_summer           0.1821      0.067      2.707      0.007       0.050       0.314
season_winter           0.4703      0.079      5.971      0.000       0.316       0.625
year_2019               1.0274      0.035     28.993      0.000       0.958       1.097
month_Dec              -0.1782      0.077     -2.325      0.020      -0.329      -0.028
month_Jan              -0.2031      0.079     -2.555      0.011      -0.359      -0.047
month_Jul              -0.2333      0.080     -2.917      0.004      -0.390      -0.076
month_Nov              -0.1883      0.082     -2.293      0.022      -0.350      -0.027
month_Sep               0.3267      0.074      4.413      0.000       0.181       0.472
weekday_Sun            -0.2071      0.050     -4.125      0.000      -0.306      -0.108
weather_Light_Snow     -1.1438      0.116     -9.850      0.000      -1.372      -0.916
weather_Mist_Cloudy    -0.2656      0.046     -5.762      0.000      -0.356      -0.175
==============================================================================
Omnibus:                       73.424   Durbin-Watson:                   2.040
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              202.465
Skew:                          -0.703   Prob(JB):                     1.08e-44
Kurtosis:                       5.748   Cond. No.                         11.1
==============================================================================
```

In the code presented Model 13 is the final model because all the variable has p value less than 0.05 and VIF<5 except for season_spring. VIF for season_spring is just above 5 but at the same time coefficient is negative (-0.2724), so we can take that into consideration

The error terms should follow a normal distribution**.** The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one). This is what is observed on Model 13 in the code



scatter plot shows the residual vs predicted value. The data points are spread across equally without a prominent pattern, it means the residuals have constant variance (homoscedasticity)



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

Answer:

The Top 3 features contributing significantly towards the demands of share bikes are:

- weathersit_Light_Snow (negative correlation).
- yr_2019 (Positive correlation).
- temp (Positive correlation).

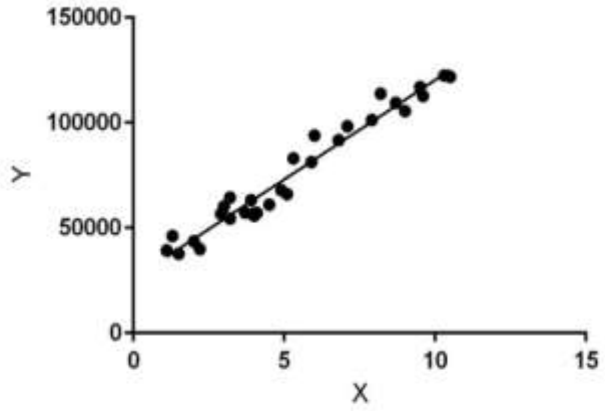
General Subjective Questions

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

Answers: -

- It is a machine learning algorithm based on supervised learning. It performs a regression task.
- Regression models a target prediction value based on independent variables.
- It is mostly used for finding out the relationship between variables and forecasting.
- Different regression models differ based on – the kind of relationship between dependent and independent variables taken into consideration and the number of independent variables getting used.
- Equation for simple linear regression is y=mx+B where m is regression coefficient and B is intercept.
- In the above equation Y is Target variable and x is predictor variable.
- For multi-linear regression the equation will change to y=B1x1+B2x2+B3x3………Bnxn+C.
- Regression coefficient calculation formula is below: -

$$\beta = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2},$$

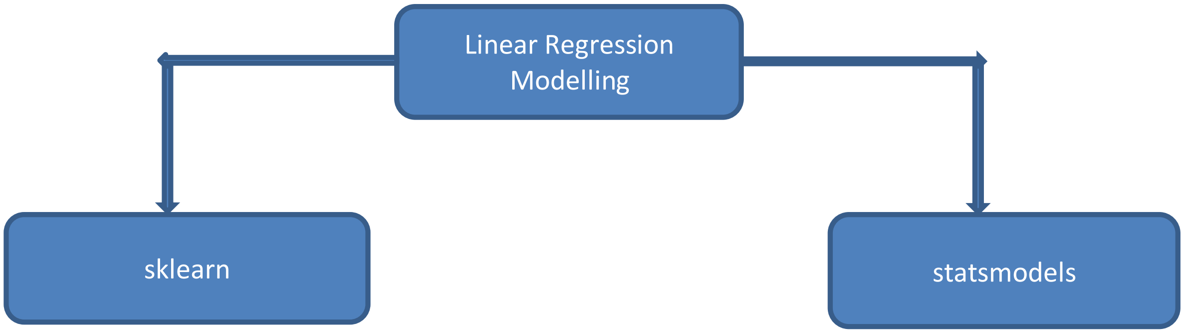


- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

- So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

- In the above figure, X (input) is the income and Y (output) is the tax paid by a person. The regression line is the best fit line for our model.

The two major library which are use for linear regression model is: -



Below are the major library's list which helps to perform linear regression: -

- import pandas as pd

- import numpy as np

- import seaborn as sns

- import matplotlib. pyplot as plt

- import warnings
- warnings.filterwarnings (action = 'ignore')
- from sklearn.model_selection import train_test_split
- from sklearn.preprocessing import MinMaxScaler
- import statsmodels.api as sm
- from sklearn.feature_selection import RFE
- from sklearn.linear_model import LinearRegression
- from sklearn.preprocessing import OneHotEncoder
- from statsmodels.graphics.gofplots import qqplot
- from sklearn.metrics import r2_score

There are three major steps we follow to perform linear regression: -

i. Analyzing the correlation and directionality of the data,

ii. Estimating the model, i.e., fitting the line

iii. evaluating the validity and usefulness of the model.

Below are the few important point for linear regression algorithm: -

i. In ideal situation simple linear regression must follow the equation y=Bx+c equation and for multi linear regression it must follow equation y=B1x1+B2x2+B3x3…Bnxn

ii. Error terms are independent, normally distributed with mean zero with constant variance

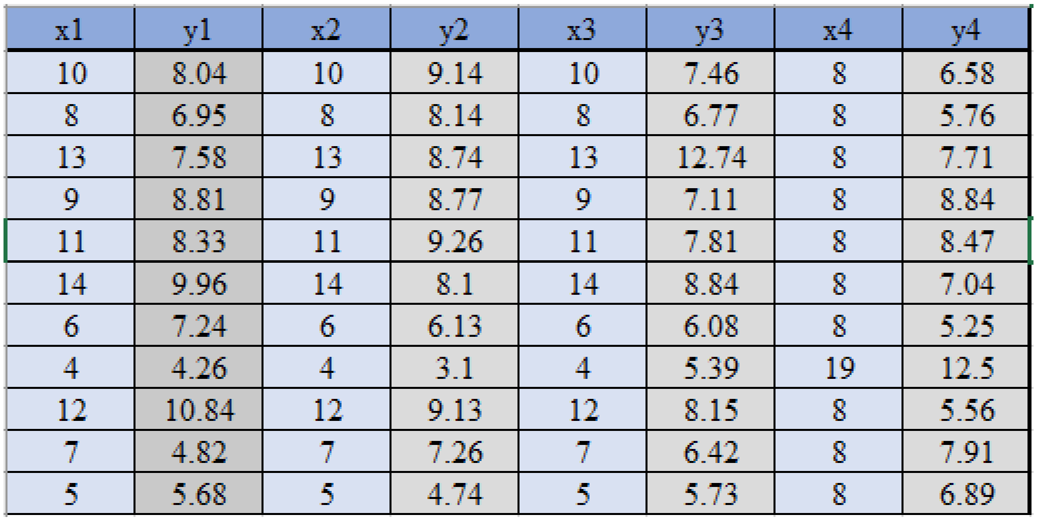
iii. Coefficients is obtained by minimising the sum of squared errors.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer: -

Anscombe’s Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc.) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

For Example: -



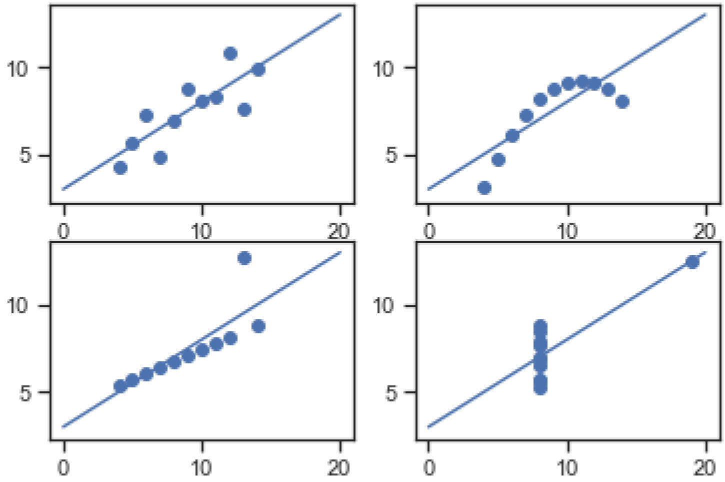
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

When we apply the statistical formula on the above data-set, we get below results:-

Average Value of x = 9, Average Value of y = 7.50, Variance of x = 11, Variance of y =4.12, Correlation Coefficient = 0.816

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.

Linear Regression Equation: y = 0.5 x + 3



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

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

Answers: -

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

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)

Answer: -

It is a step of data Pre-Processing which is applied to independent variables 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.

Difference between normalized scaling and standardized scaling are below: -

S.No	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.
6	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

Answers: -

When there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

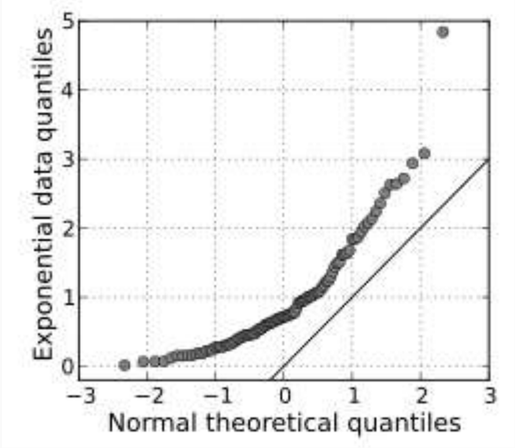
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answers: -

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. 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.

A Q-Q plot showing the 45-degree 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. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.