Name- Ayush Mehta (ayush10mehta@gmail.com)

Linear Regression Assignment

Assignment-based Subjective Questions

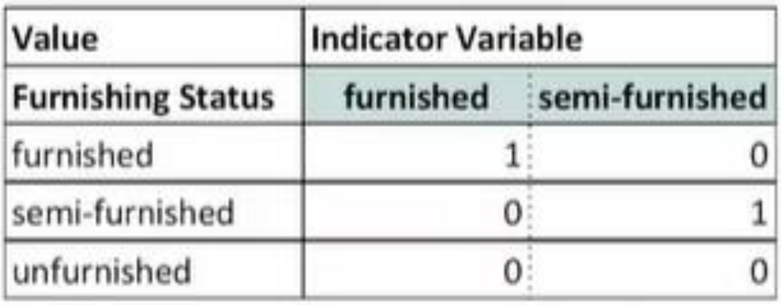
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- People are most likely to rent bike on holidays.
- People are most likely to rent bike when the weather is weathersit 1 that is when the weather is mostly clear or with few clouds or partly cloudy.
- Around 60% bike rent capacity has been increased from year 2018 to 2019.
- The bike rent capacity increases every fall that is the third season (season 3).
- Derived a variable quarter from the date column and got to know that the bike rent capacity reaches to its peak in every 3^{rd} quarter.

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

drop_first=True is important to use, because it reduces the extra column produced when dummy variables are formed. As a result, it reduces the correlations created among dummy variables.

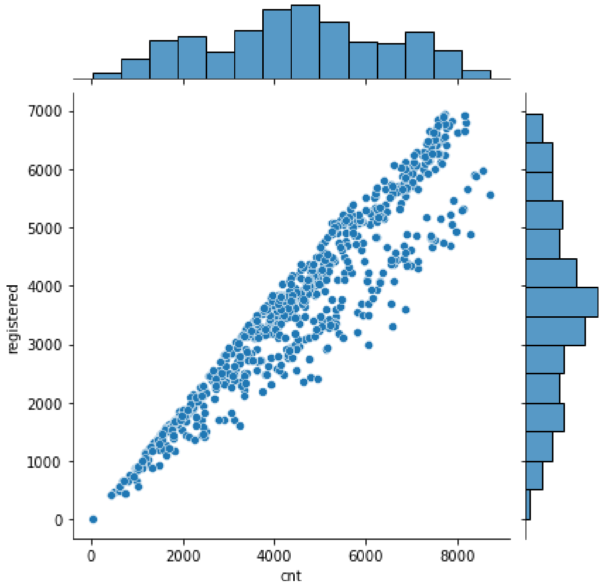
Let's imagine we want to build a dummy variable for a categorical column that has three different types of data (Furnished, unfurnished, semi- furnished). One variable is obviously unfurnished if it is neither furnished nor semi-furnished. Therefore, there is no need for a third variable to identify unfurnished.

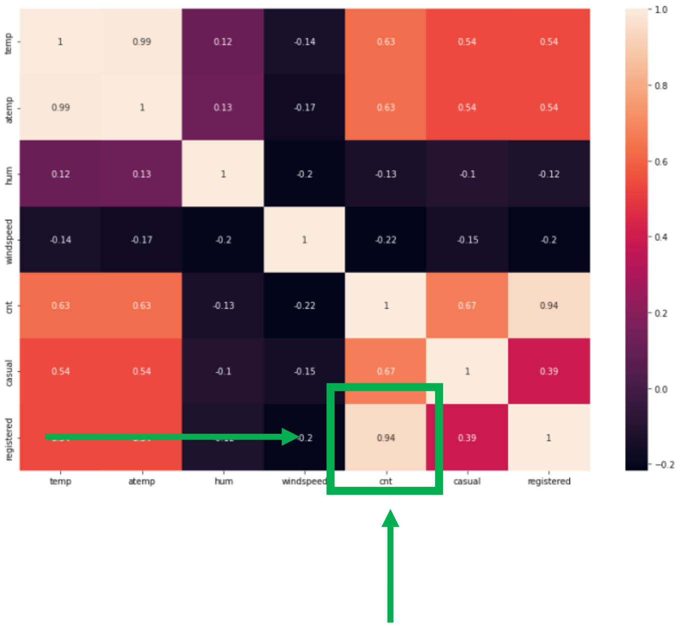


Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

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

‘Registered’ numerical variable has the highest correlation with the target variable ‘cnt’ having a correlation value of +0.94.

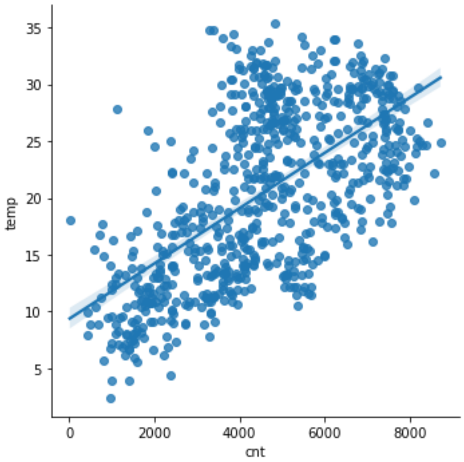


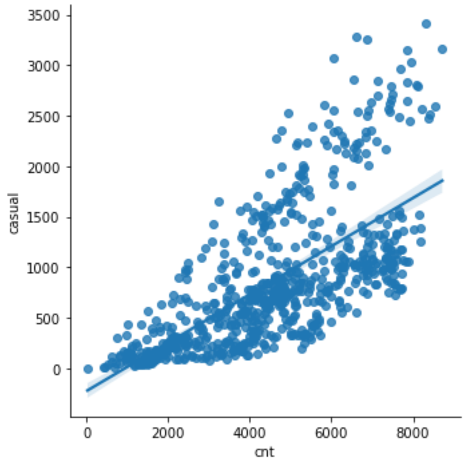


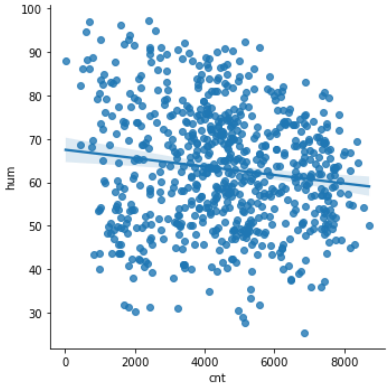
4.How did you validate the assumptions of Linear Regression after building the model on the training set?

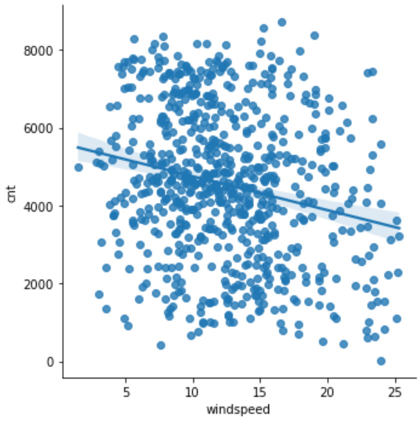
There are 5 basic assumptions of Linear Regression Algorithm:

1.Linear Relationship between the features and target:
According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

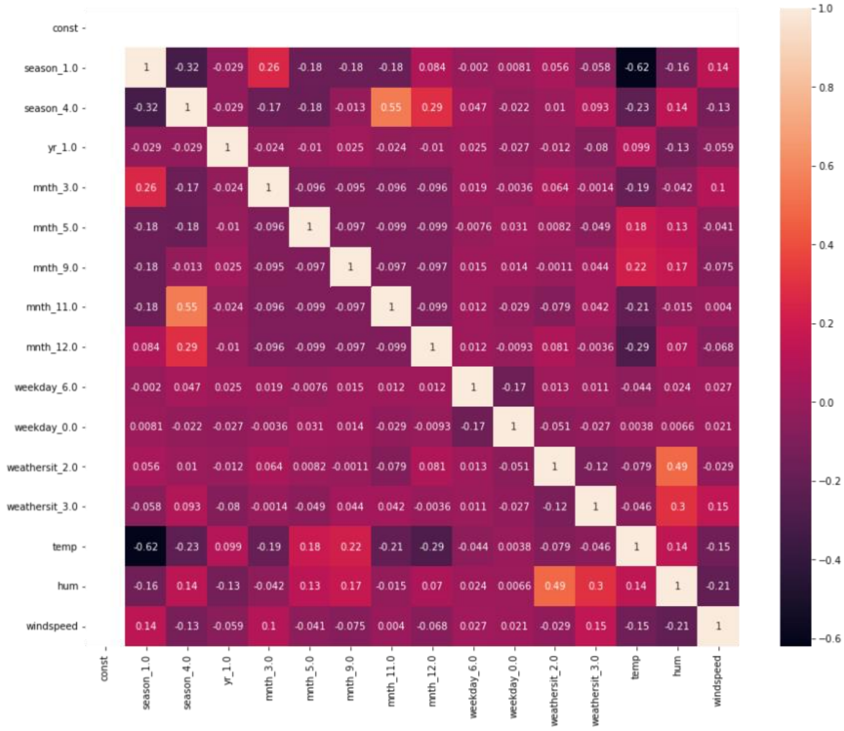




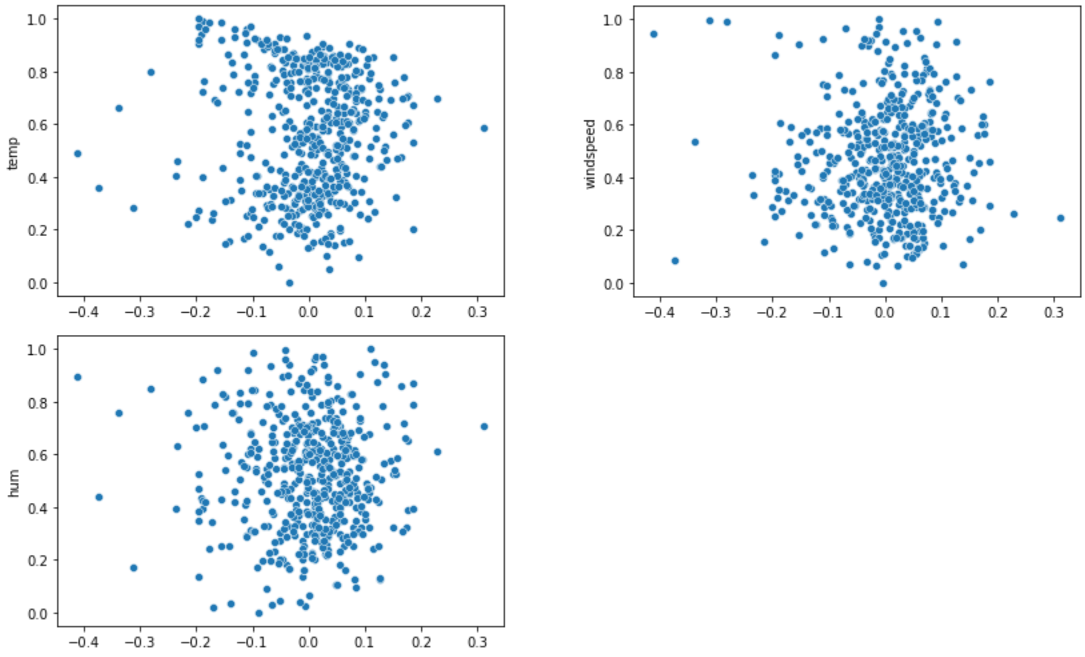




2.Little or no Multicollinearity between the features:
Multicollinearity is a condition in which there are extremely strong correlations or relationships between the independent variables. As a result, it is a form of disturbance in the data that, when present, reduces the regression model's statistical power. For locating highly connected characteristics, pair plots and heatmaps (correlation matrices) can be employed.

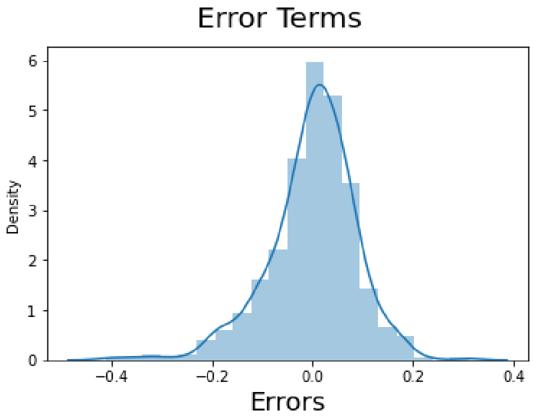


3.Homoscedasticity Assumption:
When the error term—the "noise" or random disturbance in the relationship between the features and the target—is the same for all values of the independent variables, the situation is referred to as homoscedastic. A good technique to check for homoscedasticity is to plot the residual values against the anticipated values in a scatter plot. The distribution shouldn't show any obvious patterns, and if there is one, the data is heteroscedastic.



4.Normal distribution of error terms:

The fourth assumption is that the error(residuals) follow a normal distribution.

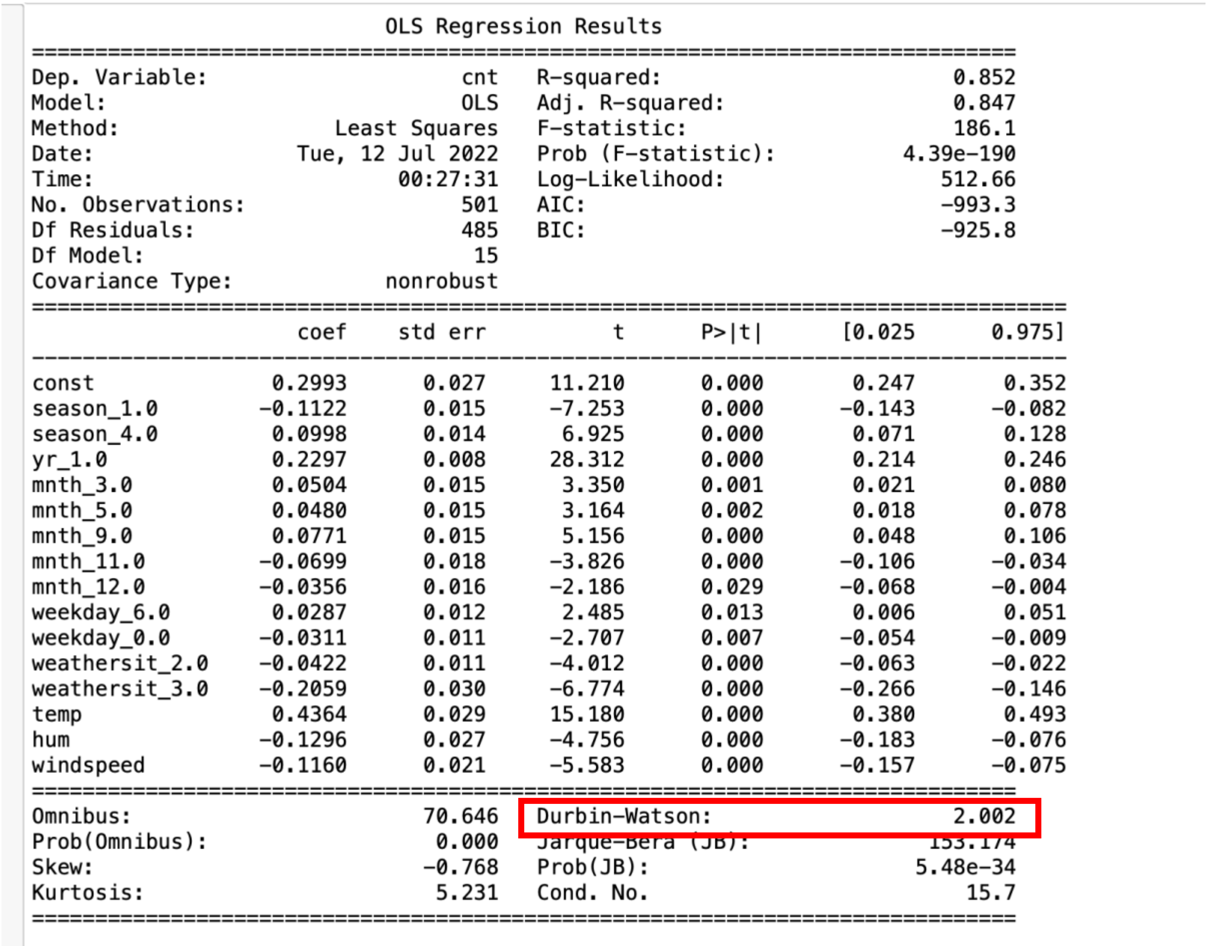


5.Little or No autocorrelation in the residuals:

Autocorrelation occurs when the residual errors are dependent on each other. The presence of correlation in error terms drastically reduces model’s accuracy. This usually occurs in time series models where the next instant is dependent on previous instant.

The Durbin-Watson test can be used to examine autocorrelation.

This number will always range from 0 to 4. There is stronger support for positive serial correlation when the statistic is closer to 0. There is increasing support for negative serial correlation the closer we get to 4.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.852
Model:                            OLS   Adj. R-squared:                  0.847
Method:                 Least Squares   F-statistic:                     186.1
Date:                Tue, 12 Jul 2022   Prob (F-statistic):          4.39e-190
Time:                        00:27:31   Log-Likelihood:                 512.66
No. Observations:                 501   AIC:                            -993.3
Df Residuals:                     485   BIC:                            -925.8
Df Model:                          15
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.2993      0.027     11.210      0.000       0.247       0.352
season_1.0        -0.1122      0.015     -7.253      0.000      -0.143      -0.082
season_4.0         0.0998      0.014      6.925      0.000       0.071       0.128
yr_1.0             0.2297      0.008     28.312      0.000       0.214       0.246
mnth_3.0           0.0504      0.015      3.350      0.001       0.021       0.080
mnth_5.0           0.0480      0.015      3.164      0.002       0.018       0.078
mnth_9.0           0.0771      0.015      5.156      0.000       0.048       0.106
mnth_11.0         -0.0699      0.018     -3.826      0.000      -0.106      -0.034
mnth_12.0         -0.0356      0.016     -2.186      0.029      -0.068      -0.004
weekday_6.0        0.0287      0.012      2.485      0.013       0.006       0.051
weekday_0.0       -0.0311      0.011     -2.707      0.007      -0.054      -0.009
weathersit_2.0    -0.0422      0.011     -4.012      0.000      -0.063      -0.022
weathersit_3.0    -0.2059      0.030     -6.774      0.000      -0.266      -0.146
temp               0.4364      0.029     15.180      0.000       0.380       0.493
hum               -0.1296      0.027     -4.756      0.000      -0.183      -0.076
windspeed         -0.1160      0.021     -5.583      0.000      -0.157      -0.075
==============================================================================
Omnibus:                       70.646   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              153.174
Skew:                          -0.768   Prob(JB):                     5.48e-34
Kurtosis:                       5.231   Cond. No.                         15.7
==============================================================================
```

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

Temperature, weekday_0.0, yr_1.0 This are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning.

Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

One of the most fundamental types of machine learning in the field of data science is linear regression, where we train a model to forecast the behaviour of your data based on a few variables. As you can see from the name, linear regression requires that the two variables that are on the x- and y-axes have a linear correlation.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

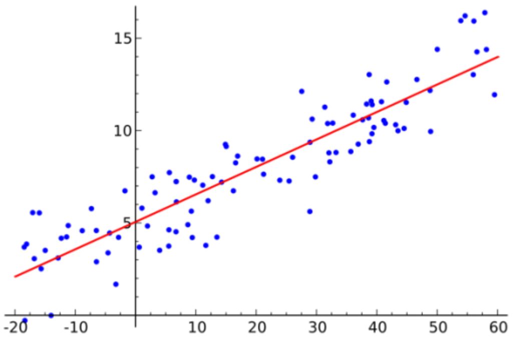
$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

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

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation



shown above.

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
Modelling the linear relationship between the explanatory (independent) factors and response (dependent) variables is the aim of multiple linear regression. Because multiple regression takes into account several explanatory variables, it can be thought of as an extension of ordinary least-squares (OLS) regression.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Cost Function

The cost function helps us to figure out the best possible values for a and b which would provide the best fit line for the data points. Since we want the best values for a and b, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

Residuals (ei) = y_actual – y_predicted

RSS (Residual sum of squares) = e1^2 +e2^2 + ------------- en^2

To find the minimum equate the equations derivative to 0 and find the value of a & b.

Evaluation - There are 3 main metrics for model evaluation in regression:

R Square/Adjusted R Square - R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Mean Square Error(MSE)/Root Mean Square Error(RMSE) - While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Mean Absolute Error(MAE) - Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

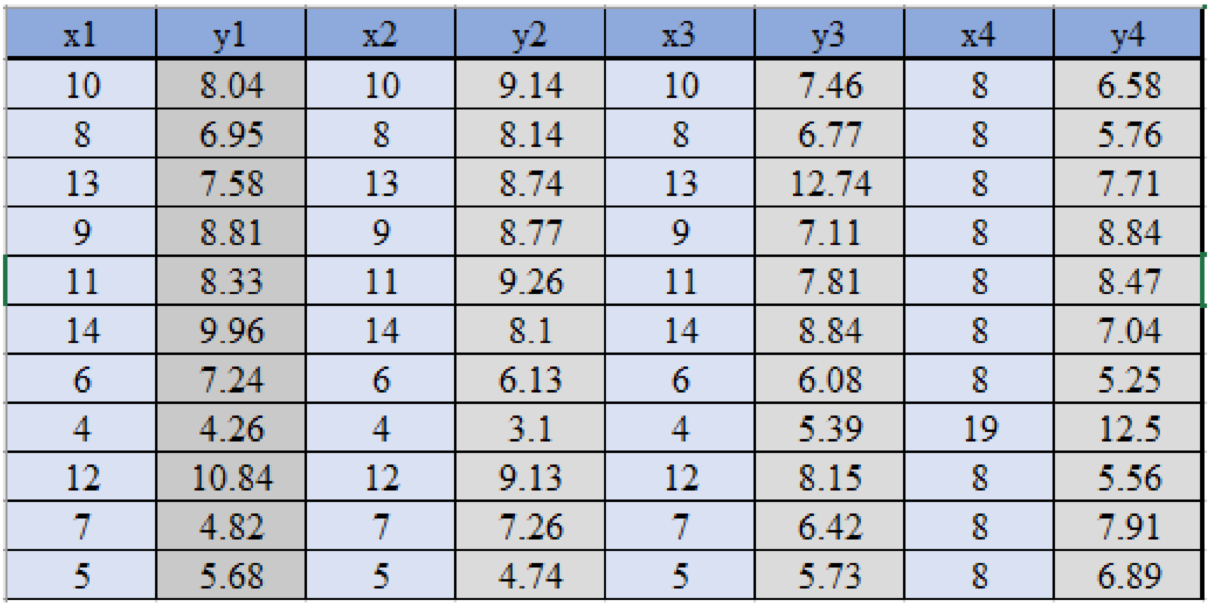
Assumptions of Linear Regression - Please refer to the question no. 4 of Assignment-based Subjective Questions

2.Explain the Anscombe’s quartet in detail.

Even though it's absolutely false, most people tend to think that "the numerical computations are correct, but the graphs are rough.

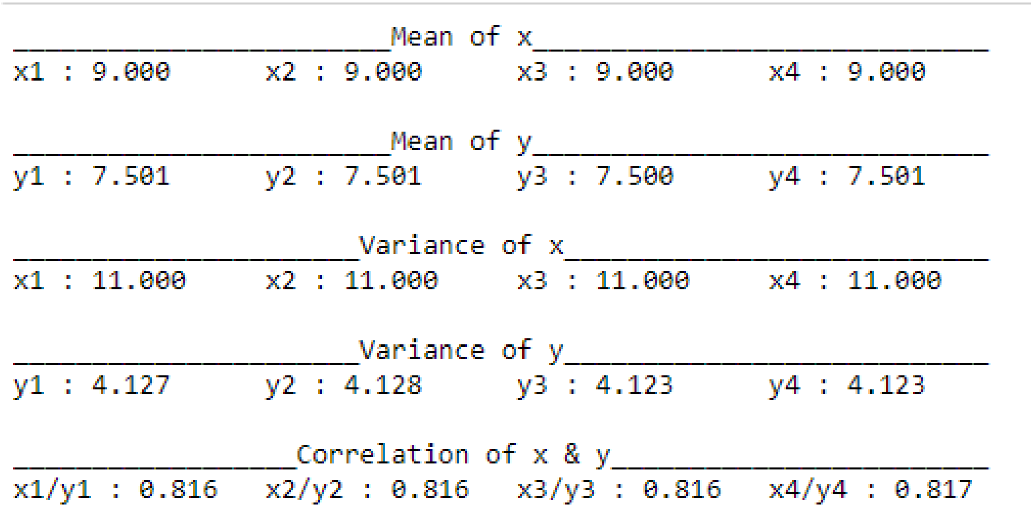
Anscombe's Quartet, created in 1973 by statistician Francis Anscombe to emphasise the value of visualizing data before evaluating it using statistical features, serves as the modal example to illustrate the significance of data visualisation. It is made up of four data sets, each of which has eleven (x,y) points. The fundamental characteristic of these data sets that needs to be analysed is that they all have distinct graphical representations but the same descriptive statistics (mean, variance, standard deviation, etc.). Regardless of the statistical analysis, each graph plot demonstrates a different behaviour.

Data set



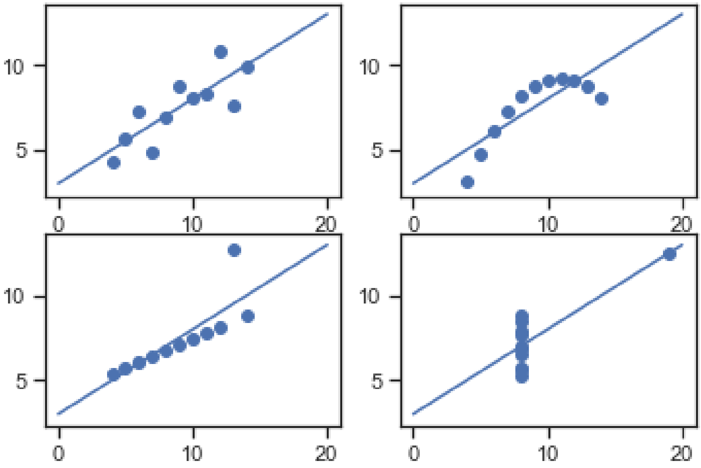
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7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

Statistical analysis



```
________________________Mean of x_____________________________
x1 : 9.000       x2 : 9.000       x3 : 9.000       x4 : 9.000

________________________Mean of y_____________________________
y1 : 7.501       y2 : 7.501       y3 : 7.500       y4 : 7.501

______________________Variance of x___________________________
x1 : 11.000      x2 : 11.000      x3 : 11.000      x4 : 11.000

______________________Variance of y___________________________
y1 : 4.127       y2 : 4.128       y3 : 4.123       y4 : 4.123

__________________Correlation of x & y________________________
x1/y1 : 0.816    x2/y2 : 0.816    x3/y3 : 0.816    x4/y4 : 0.817
```

plotting
e results



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?

Pearson correlation coefficient, often known as Pearson's correlation coefficient or Pearson's r, as the measurement of the strength of the relationship and association between two variables.

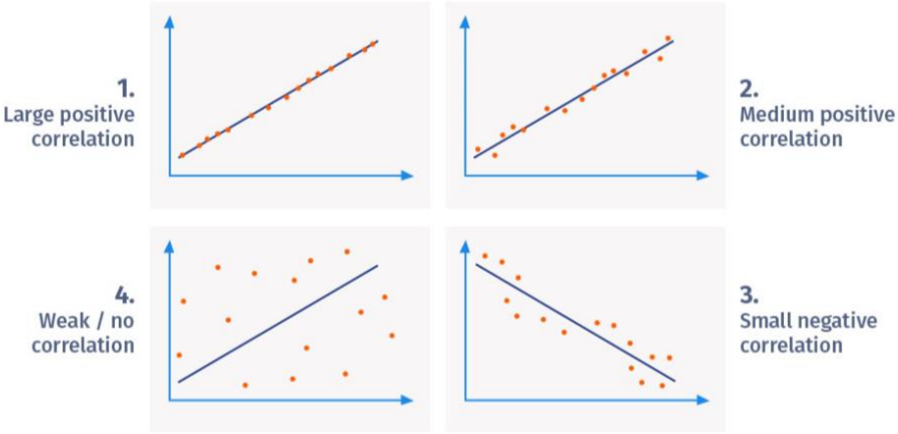
Simply said, Pearson's correlation coefficient determines the impact of a change in one variable on a change in the other.

The statistical significance of the Pearson coefficient connection is very high. It examines the connection between two variables. It aims to depict the link between two variables by drawing a line through their data. With the help of the Pearson correlation coefficient calculator, the relationship between the variables is measured. There are two possible outcomes for this linear relationship.

- Positive linear relationship: Generally speaking, as a person gets older, their income rises.
- Negative linear relationship: As the vehicle's speed rises, so does the amount of time it takes to travel, and vice versa.

The correlation coefficient formula determines how the variables are related. It gives back values in the range of -1 and 1.

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



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

To standardise the data within a specific range, independent variables are subjected to the data pre-processing technique known as scaling. Additionally, it help in accelerating algorithmic calculations.

The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range. If scaling is not done, the algorithm will only consider magnitude and not units, which will result in inaccurate modelling. We must scale all the variables to the same degree of magnitude in order to resolve this problem. The t-statistic, F-statistic, p-values, R-squared, etc. are unaffected by scaling, which is significant because they are all dependent on the coefficients.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

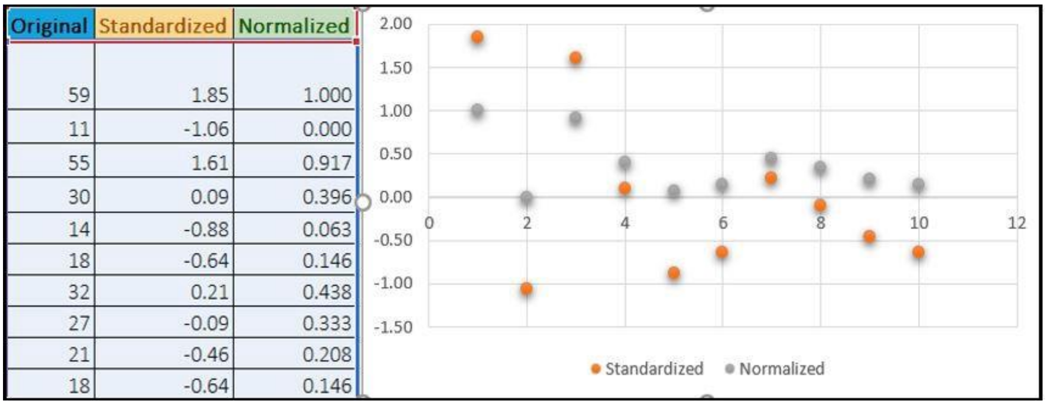
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Diagram helps us in visualizing clearly the difference between normalization and standardization.

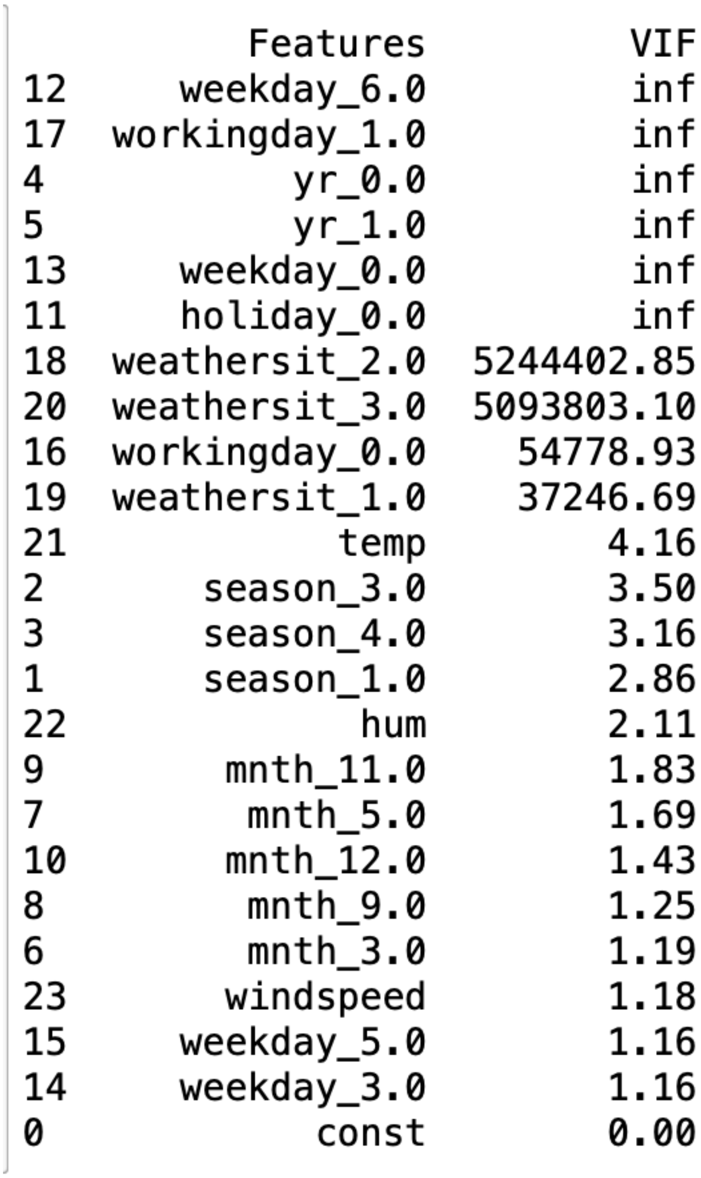


Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

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

VIF = infinity if there is perfect correlation. This demonstrates that the two independent variables have an exact association. If the correlation is perfect, we have R2 = 1, which results in 1/(1-R2) infinite. The variable that is producing this perfect multicollinearity must be removed from the dataset in order to solve the issue.

An infinite VIF value means that a linear combination of other variables can precisely express the associated variable (which show an infinite VIF as well).



```
          Features          VIF
12       weekday_6.0          inf
17    workingday_1.0          inf
4             yr_0.0          inf
5             yr_1.0          inf
13       weekday_0.0          inf
11       holiday_0.0          inf
18    weathersit_2.0   5244402.85
20    weathersit_3.0   5093803.10
16    workingday_0.0     54778.93
19    weathersit_1.0     37246.69
21              temp         4.16
2         season_3.0         3.50
3         season_4.0         3.16
1         season_1.0         2.86
22               hum         2.11
9          mnth_11.0         1.83
7           mnth_5.0         1.69
10         mnth_12.0         1.43
8           mnth_9.0         1.25
6           mnth_3.0         1.19
23         windspeed         1.18
15       weekday_5.0         1.16
14       weekday_3.0         1.16
0              const         0.00
```

We dropped few of the columns one by one and rechecked the VIF value to solve this issue in our model.

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

Using the Quantile-Quantile (Q-Q) plot, we may visually determine if a collection of data is likely to have originated from a theoretical distribution like the Normal, Exponential, or Uniform distribution. Identifying if two data sets originate from populations with a similar distribution is also helpful.

This is useful when performing a linear regression since it allows us to verify via a Q-Q plot that the training and test data sets are from populations with similar distributions.

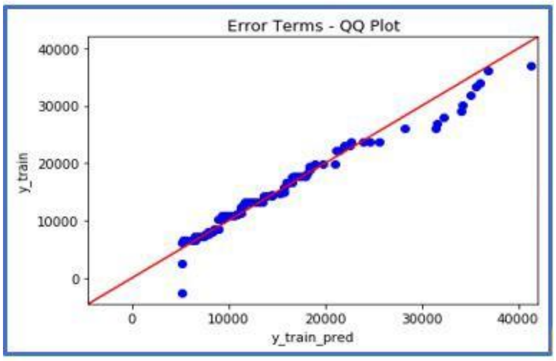
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

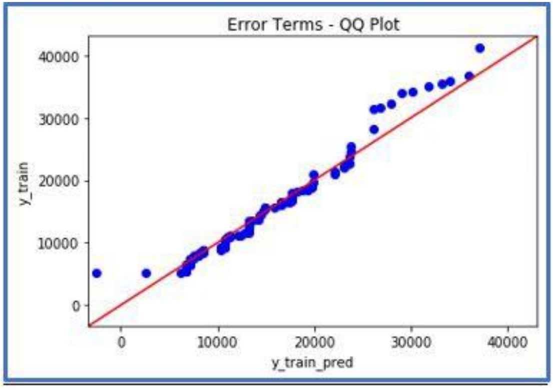
Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

- Y-values < X-values: If y-quantiles are lower than the x-quantiles.



- X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis