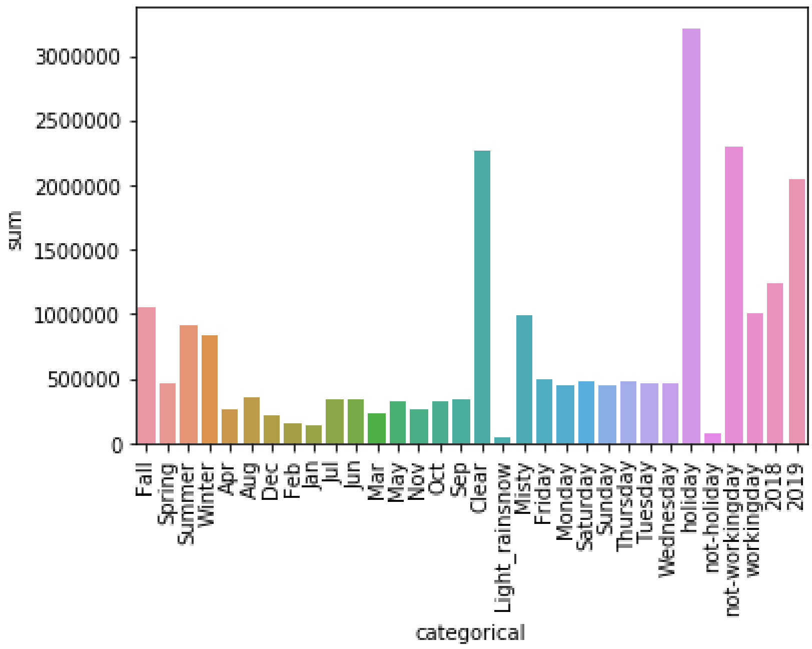
Assignment-based Subjective Questions

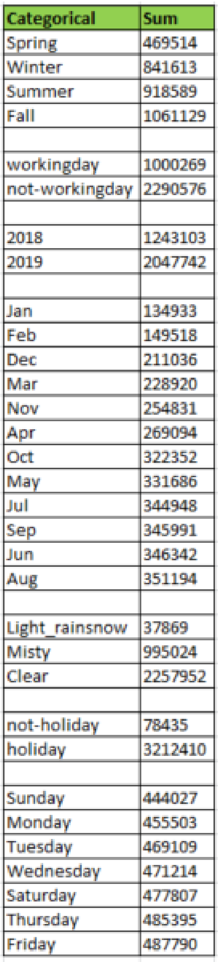
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer :



1. The demand of bike is less in Spring season, as compared to other seasons. The demand of bike is more in Fall season as compared to other seasons
2. The demand of bike is more in year 2019 as compared to year 2018
3. The demand of bike is lowest in month of January and it is highest in month of August.
4. Bike demand is more in Clear weather (Clear, Few clouds, Partly cloudy, Partly cloudy) as compared to other weather condition. For Light Snow (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), the demand of bike is lowest (there isn’t any data available for heavy snow in the input dataset provided)
5. Bike demand is more in holidays as compared to not a holiday
6. The demand of bike is more on a weekday Friday followed by Thursday
7. Bike demand is more in not-working day as compared to working days

The total bike rentals for each categorical variable listed under column ‘Categorical’ can be referred in below screenshot under the column ‘Sum’.



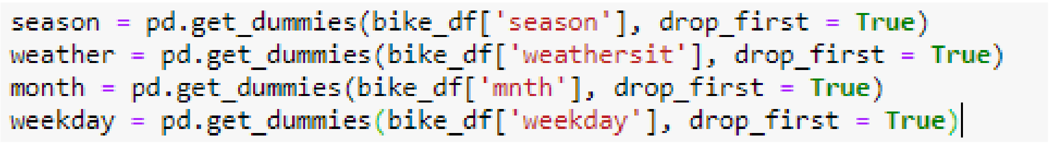
Categorical	Sum
Spring	469514
Winter	841613
Summer	918589
Fall	1061129
workingday	1000269
not-workingday	2290576
2018	1243103
2019	2047742
Jan	134933
Feb	149518
Dec	211036
Mar	228920
Nov	254831
Apr	269094
Oct	322352
May	331686
Jul	344948
Sep	345991
Jun	346342
Aug	351194
Light_rainsnow	37869
Misty	995024
Clear	2257952
not-holiday	78435
holiday	3212410
Sunday	444027
Monday	455503
Tuesday	469109
Wednesday	471214
Saturday	477807
Thursday	485395
Friday	487790

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

Answer :
We set "drop_first=True" during dummy variable creation. It drops the first column or first level to get 'n-1' dummies out of 'n' categorical levels. It helps in reducing the extra column created during dummy variable creation. Hence, reducing the correlations created among dummy variables.

Example : -
If we have 3 types of values in Categorical column [furnished, semi-furnished and unfurnished] and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then, it is obvious to be unfurnished. So, that means one doesn't need the third variable to identify the unfurnished category.

In the assignment, for below categorical variables 'get_dummies' and 'drop_first = True' has been used.

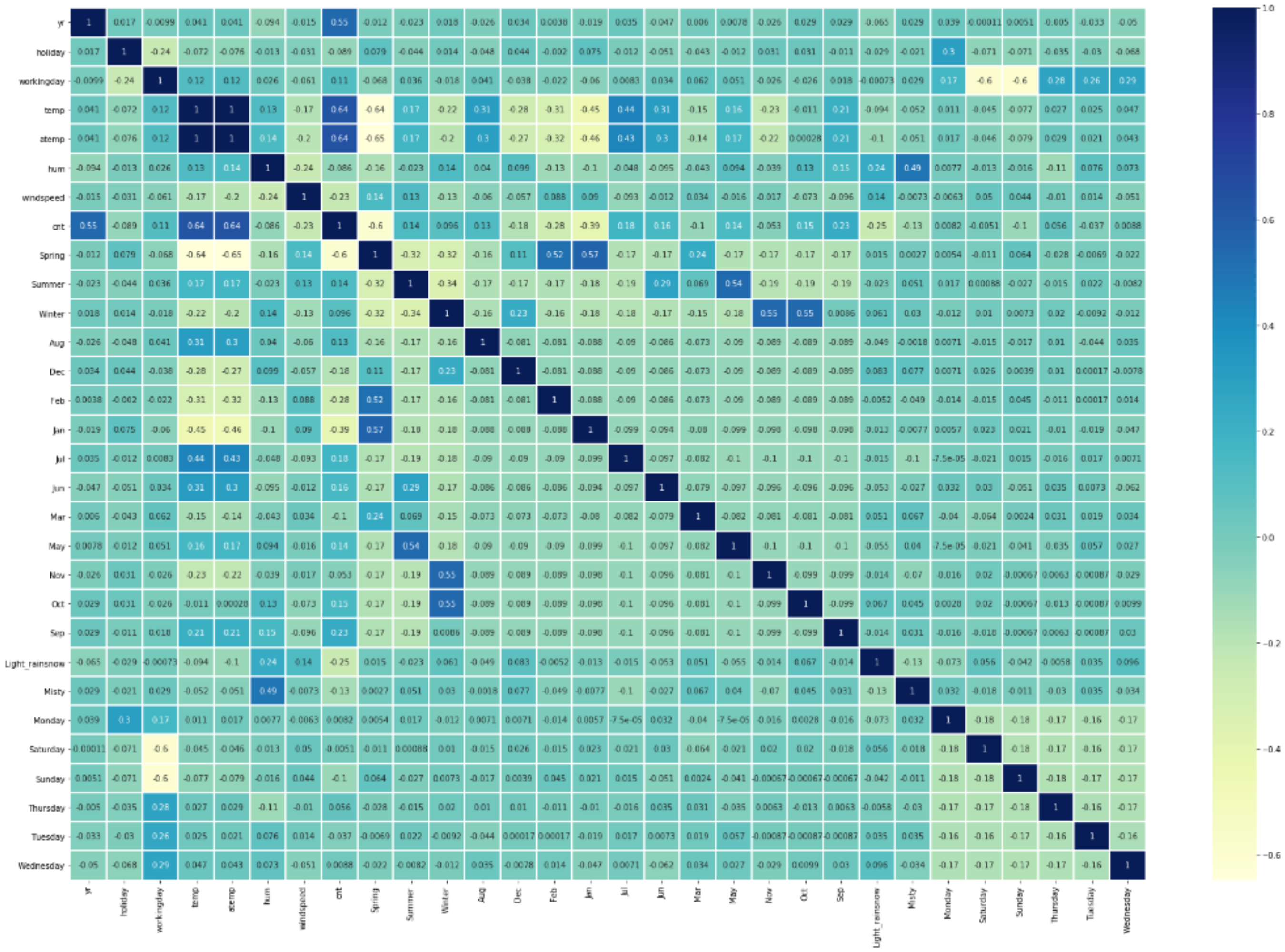


```
season = pd.get_dummies(bike_df['season'], drop_first = True)
weather = pd.get_dummies(bike_df['weathersit'], drop_first = True)
month = pd.get_dummies(bike_df['mnth'], drop_first = True)
weekday = pd.get_dummies(bike_df['weekday'], drop_first = True)
```

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

Answer :

Features ‘temp’ and ‘atemp’ has very high correlation with value ‘1’. We can use only one of these two variables for analysis and building a model to avoid multicollinearity.

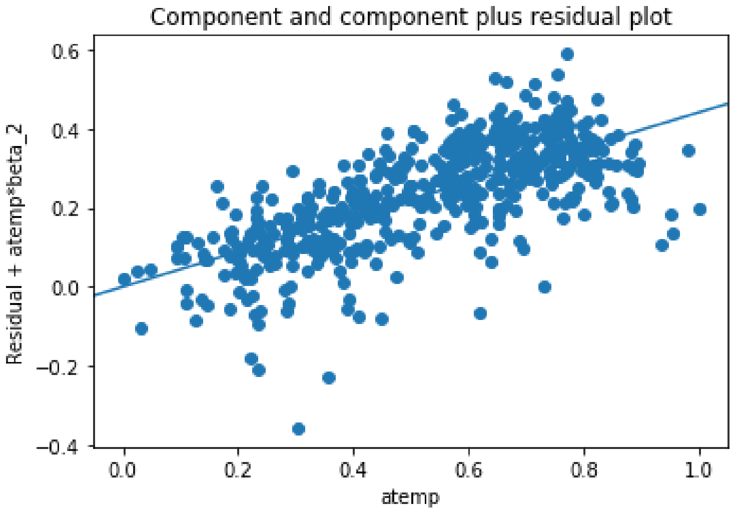


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

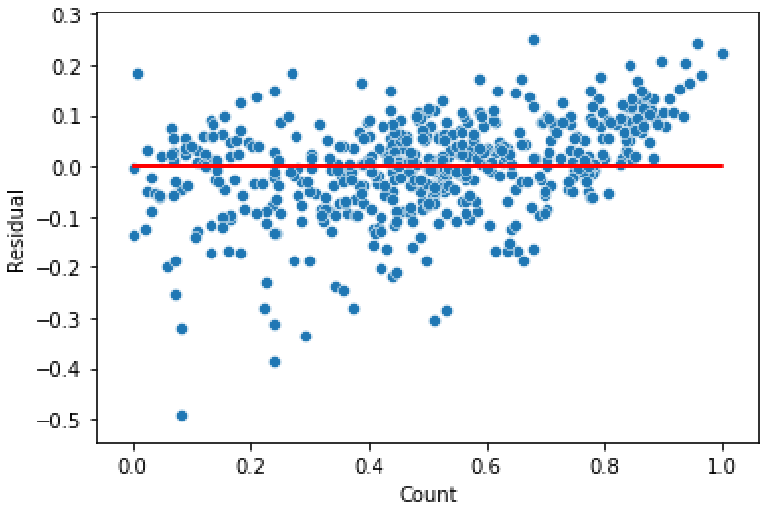
Answer :

Below are the data points on which validation was done while building a Linear Regression Model :

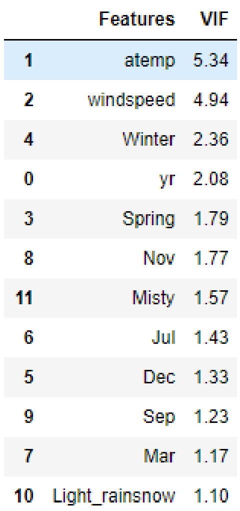
- Linear Relationship
  - The Dependent variable and Independent variable must have a linear relationship
    - As per the data, the plots (like pairplot, plot_ccpr) represents the relationship between the model and the predictor variables. With the plots we can see that the linearity is well preserved



- Homoscedasticity
  - As per the data, there is no visible pattern in residual values, thus homoscedacity is well preserved



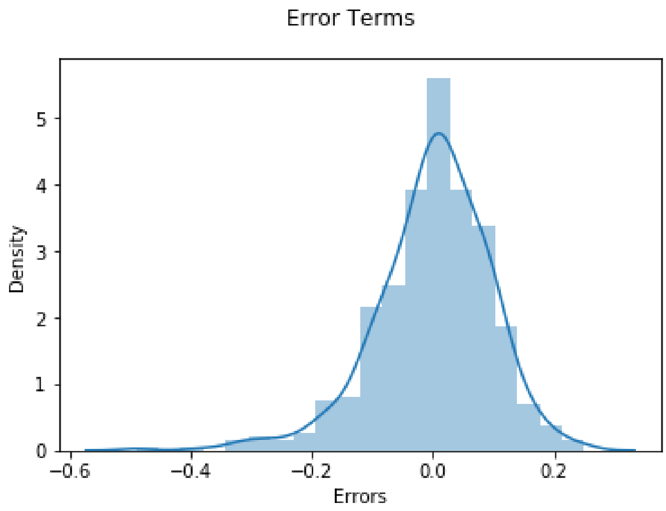
- Absence of Multicollinearity
  - As per the data, all the predictor variables have VIF value around 5. So, we can consider that there isn't any multicollinearity among the predictor variables



	Features	VIF
1	atemp	5.34
2	windspeed	4.94
4	Winter	2.36
0	yr	2.08
3	Spring	1.79
8	Nov	1.77
11	Misty	1.57
6	Jul	1.43
5	Dec	1.33
9	Sep	1.23
7	Mar	1.17
10	Light_rainsnow	1.10

- Independence of residuals
  - As per the data, the Durbin-Watson value for Final Model (lr 4) is 2.066. This shows the value closer to 2, the less auto-correlation there is between the various variables

- Normality of Errors
  - As per the data, from the plot, we can conclude that error terms are following a normal distribution with mean zero



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 :

As per the final model, the top 3 predictor variables that influence bike booking are:

- Temperature (atemp)
  - A coefficient value of ‘0.44183’ indicates that a temperature has significant impact on bike rentals
- Light_rainsnow (weathersit =3)

    - A coefficient value of '-0.265918' indicates that the light snowrain deters people from renting out bikes
- Year (yr)
    - A coefficient value of '0.241595' indicates that year-wise the rental numbers are increasing (year 2019 has more rentals as compared to 2018)

Final Equation :

cnt = 0.236468 + 0.241595yr + 0.44183atemp + (-0.0815windspeed) + (-0.162485Spring) + 0.083225Winter +(-0.068584Dec) + (-0.053022Jul) + 0.054613Mar + (-0.0853Nov) + 0.049708Sep + (-0.265918Light_rainsnow) + (-0.084282Misty)

General Subjective Questions

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

Answer :

In Cross Industry Standard Process for Data Mining (CRISP–DM) framework, for the Model building (analytical core of the data mining process which uses machine learning algorithms, in which the machine learns from the data) and Evaluation (ensures an exact comparison of the created data models & selects the most best fit model) phase, we use predictive analysis on cleaned and prepared data. Linear regression is most commonly used method of predictive analysis.

Machine learning models can be classified into 3 types based on the task performed, the nature of the output & whether they are part of supervised or unsupervised learning methods, that is,

Supervised Learning Method	Unsupervised Learning Method
Regression: The output variable to be predicted is a continuous variable Example of Regression : Linear Regression, Logistic Regression etc.	Clustering: No predefined notion of label allocated to groups or clusters
Classification: The output variable to be predicted is a categorical variable	

Regression analysis is used to find the equation that describes well or fits the data. With regression equation, one can use the model to make predictions. Linear regression is a type of regression analysis which uses linear relationships between a dependent variable known as target & one or more independent variables known as predictors to predict the future of the target. It is of two types: Simple Linear Regression & Multiple Linear Regression.

Simple Linear Regression is where one independent/predictor variable is present & the model has to find the linear relationship of it with the target variable. Whereas, Multiple Linear Regression is where, there are more than one independent/predictor variables for the model to find the relationship with the target variable.

Equation of Simple Linear Regression, where β_0 is the intercept, β_1 is coefficient / slope, X is the predictor variable and Y is the dependent variable with n is the sample size.

$Y = \beta_1 * X + \beta_0$

Intercept β_0 is calculated as : $[\sum Y * \sum X^2 - (\sum X) * (\sum XY)] / [n * \sum X^2 – (\sum X)^2]$
Slope β_1 : $[(n * \sum XY) – ((\sum X) * (\sum Y))] / [n * \sum X^2 – (\sum X)^2]$

Equation of Multiple Linear Regression, where β_0 is the intercept, $\beta_{1,} \beta_{2,} \beta_{3,....,} \beta_n$ are coefficients / slopes of the predictors variables $X_{1,} X_{2,} X_{3,....,} X_n$ and Y is the dependent variable.

$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + ... + \beta_n * X_n$

Steps involved in model building for linear regression (simple linear regression [SLR] and multiple linear regression [MLR]):

1) Reading, understanding and visualising the data
   a. To visualize the linear relationship between numerical variables : use Pairplot
   b. To visualize the linear relationship between categorical & continuous variable : use Boxplot
2) Preparing the data for Modelling
   a. Encoding - Creation of Dummy/indicator variables
   b. Splitting the data into Train set and Test test
   c. Feature scaling using Normalization or Standardisation
3) Training the model
   a. 1st option : Adding highly correlated variable (from heatmap) one by one in building a model
   b. 2nd option : Build model with all the features, then keep on removing features one-by-one from the model based on p-value and VIF based on below :
      i. High p value, High VIF – Remove the features
      ii. High - Low :
         i. High p, Low VIF : Remove these Features First
         ii. Low p, High VIF : Remove these Features after the [High p, Low VIF] condition

      Low p value, Low VIF – We would keep them in model as they are significant variables to the model.
   c. 3rd option : Build model with top n feature using RFE (Recursive feature elimination) and using rfe ranking and rfe support (True) its decided which features are to be considered in model building.
4) Residual analysis
   a. To check if the error terms are normally distributed, plot the histogram of the error terms and verify if the mean is Zero, plot a scatter plot to verify if error terms are independent of each other
5) Predicting and Evaluation on the test data
   a. On Test data, derive the predicted values and use scatter plot to visualize the value of Y test and Y predicted
   b. Verifying the RMSE - high RMSE is bad and a low RMSE is good

c. Verifying the Adjusted R-squared on Train and Test data and compare - Higher R^2, better the model fits the data. In general, the difference between Train and Test should be within 5% means a good stable model.

d. Prob(F-Statistic) low (< 0.05) represents overall model fit is significant and the fit is not by chance. Prob(F-Statistic) > 0.05, might represent the Model fit might be by chance.

Below are the formulas for RSS, TSS, R-squared, RMSE, VIF and Adjusted R-squared :

For a model, best-fit line is found by minimising the expression of RSS (Residual Sum of Squares), which is equal to the sum of squares of the residual for each data point in the plot. Residuals is calculated for a data point as subtracting predicted value of dependent variable from actual value of dependent variable. A small RSS would indicate a tight fit of the model to the data. Y_i values of the y-variable in a sample and X_i values of the x-variable in sample.

- $RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

Total sum of squares (TSS) is calculated as sum of squares of measurement minus their mean (Y_{mean}). It shows how much the data.

- $TSS = \sum_{i=1}^{n} (Y_i - Y_{mean})^2$

The coefficient of determination R^2 shows the magnitude of the association. It shows how close the data are to the fitted regression line. It takes value between 0 and 1. Higher R^2, better the model fits the data.

- $R^2 = (1 – RSS/TSS)$

Root Mean Squared error (RMSE) is a metric that gives the deviation of the predicted values (Y_{pred}) from the actual observed value (Y_i). Since, it is a sort of error term, it is better to have a low RMSE. So, a high RMSE is bad and a low RMSE is good.

- $RMSE = \sqrt{} [\sum_{i=1}^{n} (Y_i – Y_{pred})^2 / n]$

<u>For Multiple Linear Regression (MLR) :</u>

For detecting multicollinearity, that is detecting association with other variables VIF is used. Variables for which VIF is less than 5 should be considered in the model. Higher R^2 would mean high VIF, which would interpret the variable is having associations with other variable. The association can be explained by,

- Variation Inflation Factor (VIF_i) = $1 / (1 – R_i^2)$

In MLR, R^2 will always either increase or remain same when more features are added. But Adjusted R^2 is used to penalize model using higher number of predictors. So, if a variable is added in the model and the Adjusted R^2 drops, it represents the added variable as insignificant and shouldn’t be used in model. The adjusted R^2 value increases only if the new added feature improves the model more than would be expected by chance.

- Adjusted $R^2 = 1 – [(1 – R^2) (N – 1) /(N – p – 1)]$, p =predictor variables, N = Sample Size

P-value states the probability of observing a similar or more extreme observation given the null hypothesis is true. It is calculated from the cumulative probability for a given t-score using the T-table.

- Low p-value (< 0.05) would mean the variable is significant. High p-value (> 0.05) would mean the variable is not significant and hence won't help much in prediction.

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

Answer :

Anscombe's Quartet is defined as a group of 4 data sets, which are nearly identical with same statistical observations with same statistical information (variance, standard deviation, correlation and mean) for all data points but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions & appear differently when plotted on scatter plots.

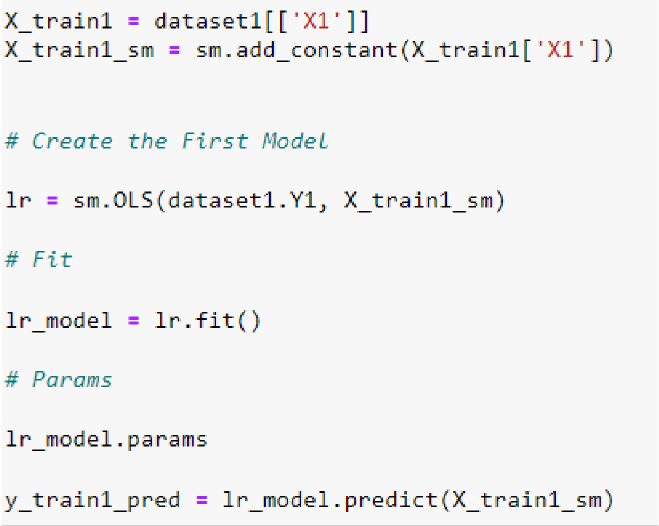
Anscombe's Quartet was developed in 1973 by statistician Francis Anscombe, to demonstrate the importance of graphing the data before analysing it with statistical properties & the effect of outliers.

The 4 dataset are is as below :

Observation	X1	Y1	X2	Y2	X3	Y3	X4	Y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
Summary Statistics								
N (Sample Size)	11	11	11	11	11	11	11	11
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

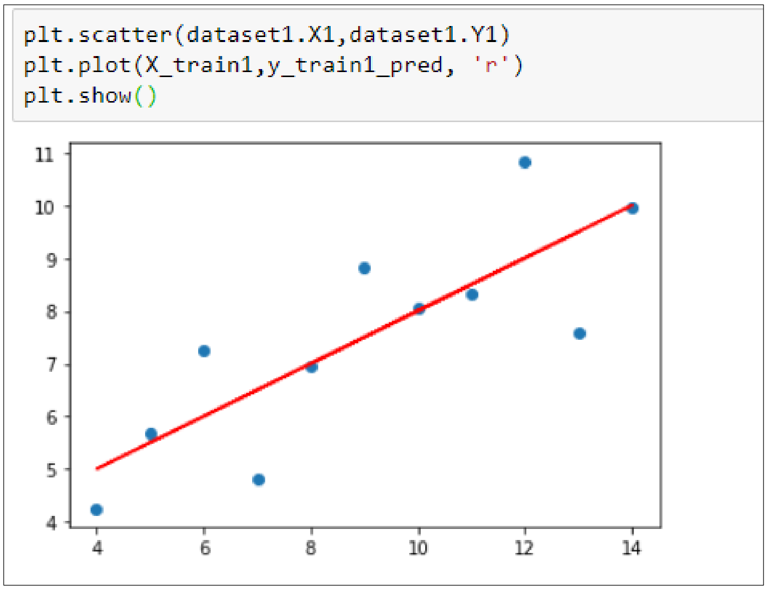
These 4 datasets can be described as follows:

For plotting scatter plots to view how well the regression line fits, have calculated the Y_predicted values using OLS from the initial datasets. Sample code for one of the dataset is as below :



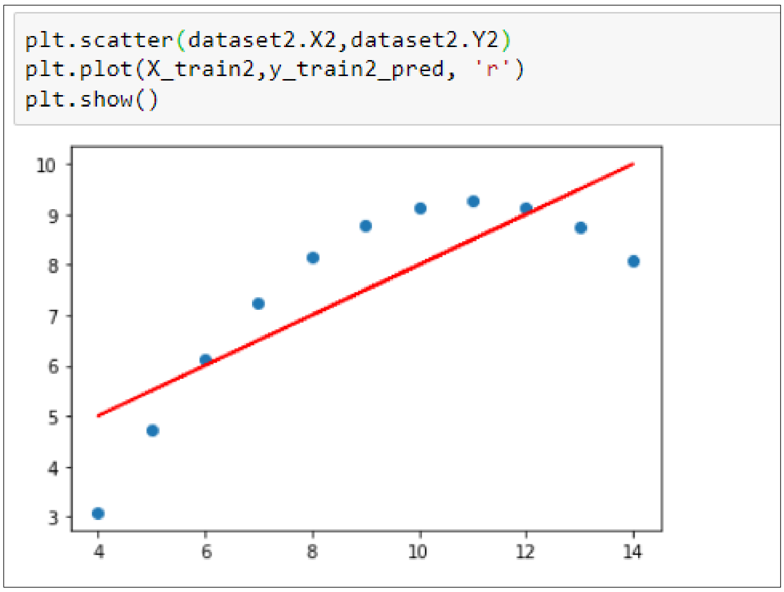
```
X_train1 = dataset1[['X1']]
X_train1_sm = sm.add_constant(X_train1['X1'])


# Create the First Model

lr = sm.OLS(dataset1.Y1, X_train1_sm)

# Fit

lr_model = lr.fit()

# Params

lr_model.params

y_train1_pred = lr_model.predict(X_train1_sm)
```

1) Dataset 1 [X1 and Y1]: This dataset fits the linear regression model well.



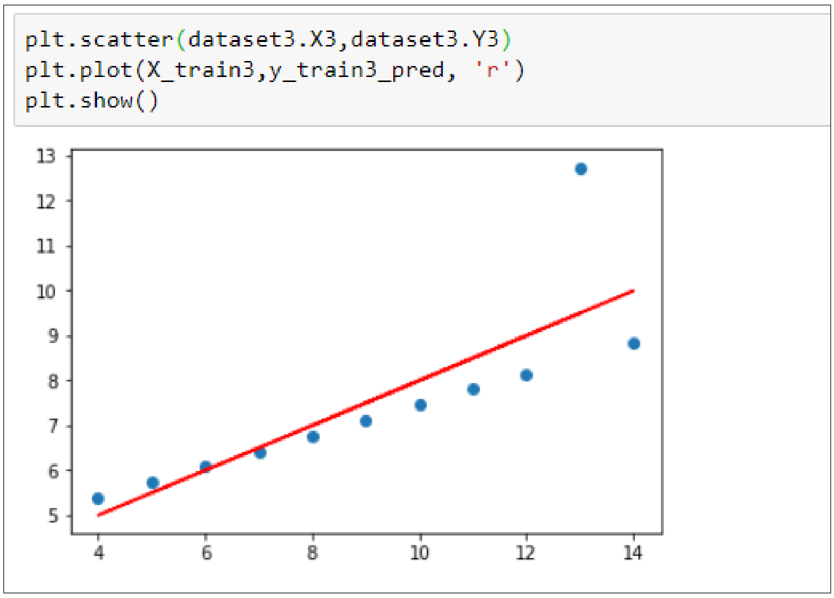
```
plt.scatter(dataset1.X1,dataset1.Y1)
plt.plot(X_train1,y_train1_pred, 'r')
plt.show()
```

2) Dataset 2 [X2 and Y2]: this dataset couldn't fit linear regression model quite well as the data is non-linear.



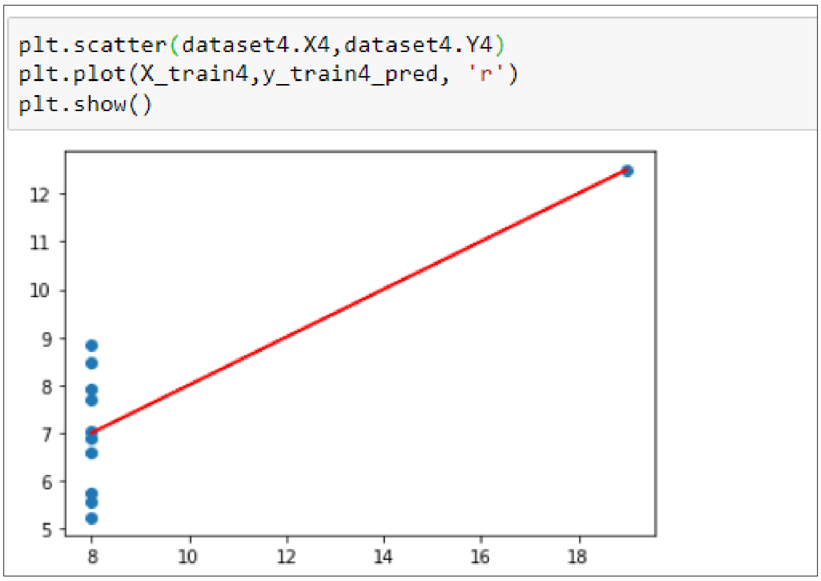
```
plt.scatter(dataset2.X2,dataset2.Y2)
plt.plot(X_train2,y_train2_pred, 'r')
plt.show()
```

3) Dataset 3 [X3 and Y3]: This has outliers involved in the dataset which cannot be handled by linear regression model



```
plt.scatter(dataset3.X3,dataset3.Y3)
plt.plot(X_train3,y_train3_pred, 'r')
plt.show()
```

4) Dataset 4 [X4 and Y4]: This has outliers involved in the dataset which cannot be handled by linear regression model



```
plt.scatter(dataset4.X4,dataset4.Y4)
plt.plot(X_train4,y_train4_pred, 'r')
plt.show()
```

3. What is Pearson’s R? (3 marks)

Answer :

Correlation coefficients are to measure how strong a relationship is between two variables. Pearson’s correlation or Pearson’s R is a correlation coefficient commonly used in linear regression. Pearson's correlation ‘r’ is a numerical summary of the strength of the linear association between the variables. For calculating Pearson ‘r’ correlation, the variables should be normally distributed, continuous with no significant outliers and linear relationship and even the error terms follow homoscedasticity.

If variables tend to go up & down together, the correlation coefficient would be considered as positive. If variables tend to go up and down in opposition, that is, low values of one variable is associated with high values of other variable, the correlation coefficient would be considered negative.

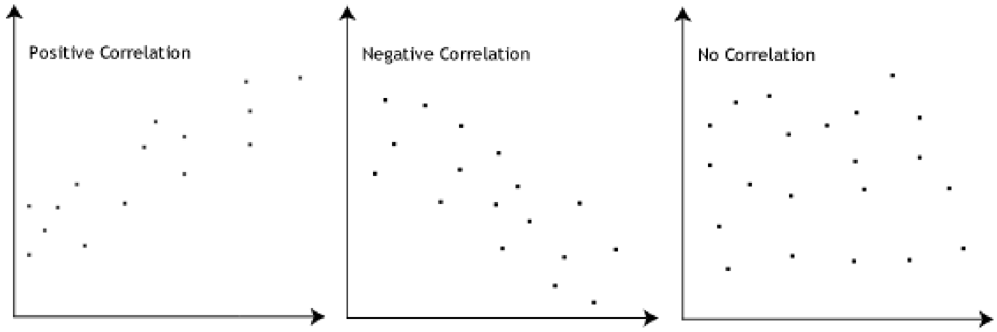
Y_i values of the Y-variable in a sample and X_i values of the X-variable in sample. x_{mean} and y_{mean} are the average values for X and Y variables.

Pearson's correlation $r = \sum (x_i - x_{mean}) * (y_i - y_{mean}) / \sqrt{} [\sum (x_i - x_{mean})^2 * (y_i - y_{mean})^2]$

The Pearson's correlation coefficient value varies between -1 and +1 and the values can be described as follows :

- r = 1, refers that the data is perfectly linear with a positive slope and for every positive increase in one, there is a positive increase of a fixed proportion in the other variable
- r = -1, refers that the data is perfectly linear with a negative slope and for every positive increase in one, there is a negative decrease of a fixed proportion in the other variable
- r = 0, refers that there is no linear association and the two variables aren’t related
- r > 0 < 5 refers that there is a weak association between variables
- r > 5 < 8 refers that there is a moderate association between variables

- r > 8 refers that there is a strong association between variables



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer :
Feature scaling is to bring all the variables at the same level. It is an important data pre-processing step applied to independent variables to normalize the data within a particular range in machine learning. Scaling helps in speeding up the calculations in an algorithm.

In general a collected data set can contain features with highly varying magnitudes, units and range. If scaling isn't done on the dataset, then algorithm will only takes magnitude into consideration & not units, hence algorithm would result in incorrect modelling. To solve this issue, scaling needs to be done so as to bring all the variables at the same level for correct interpretation.
Scaling doesn't change the shape of original distribution, it just scales them. It affects only the coefficients & no other parameters like t-statistic, F-statistic, p-values, R-squared etc.

Most common methods of scaling are Normalization (also known as min-max scaling) and Standardization.

- Normalization refers to rescale the values into a range of [0,1]. It is useful when we don't know about the distribution or when dataset doesn't follow a Normal distribution. It is also known as scaling normalization. Normalization takes care of the outliers by mapping all the maximum values to 1. If dataset has outliers, normalizing data will scale most of the data to a small interval, which means all features will have the same scale and hence it will not handle outliers well.

  Normalization is calculated as = (x – min(x)) / [max(x) – min(x)] , max(x) and min(x) are the maximum and the minimum values of the feature (x) respectively

- Standardization refers to rescales data to have a mean of 0 and a standard deviation of 1. It is useful when the feature distribution is a Normal or Gaussian distribution. It is also known as z-score normalization. Standardization has an added advantage when compared with Normalization, that is, it doesn't compress the data between a particular range. Standardization is more robust to outliers because there is no predefined range of transformed features.

Standardization is calculated as = (x – x̄) / σ , Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature (x) vector.
σ = √ [∑ (x – x̄ $)^2$ / n] , n = sample size

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

(3 marks)

Answer :
The variance inflation factor (VIF) quantifies the extent of correlation between predictors in a model.

For a perfect correlation, VIF would be infinite. An infinite VIF value would indicate that the corresponding variable might get expressed exactly by a linear combination of other variables. Value infinite means a perfect correlation between the independent variables.

For a perfect correlation, R^2 =1, VIF = 1/(1-R^2) would be infinite. To solve this problem, one variable needs to be dropped from the dataset which is causing the perfect multicollinearity.

A large value in VIF would indicate that there is a correlation between the variables. VIF greater than 10 would mean that there is multicollinearity and the feature needs to be eliminated from model.

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

(3 marks)

Answer :
Quantile-Quantile (Q-Q) plot, is a graphical method for determining whether two samples of the datasets came from same population or not. A Q-Q plot is a plot of 2 quantiles, that is, the quantile of first dataset against the quantiles of the second dataset.

This plot can be used to test distribution amongst 2 different datasets. Using this plot it is possible to compare the distributions of the datasets to see if they are indeed the same. This would be helpful in machine learning, where the dataset is split into train-validation-test to verify if the distribution is the same for both. It is also used in the post-deployment scenarios to identify covariate shift or dataset shift or concept shift visually [it refers to change in distribution of input variables present in training & test data].

If the 2 distributions being compared are similar, the points in Q–Q plot would approximately lie on the line represented as Y = X. If the distributions are linearly related, then the points in the Q–Q plot would approximately lie on a line, but not certainly on the line represented by Y = X.

For Q–Q plot, the sample sizes needn’t be equal and various distributional aspects can simultaneously be tested using this plot like shifts in location and scale, changes in symmetry, and the presence of outliers. So, if 2 data sets come from populations whose distributions differ only by a shift in location, then, the points would lie along a straight line that is displaced either up or down from the reference line.

The Quantile-Quantile plot is used for the following purpose:

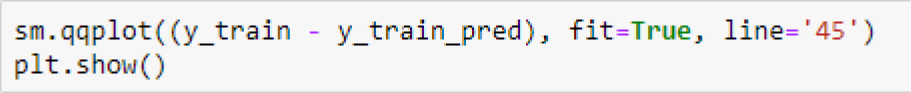
- To determine whether 2 samples comes from the same population.
- To determine whether 2 samples have the same tail
- To determine whether 2 samples have the same distribution shape
- To determine whether 2 samples have common location behaviour

Plotting a Q-Q plot :

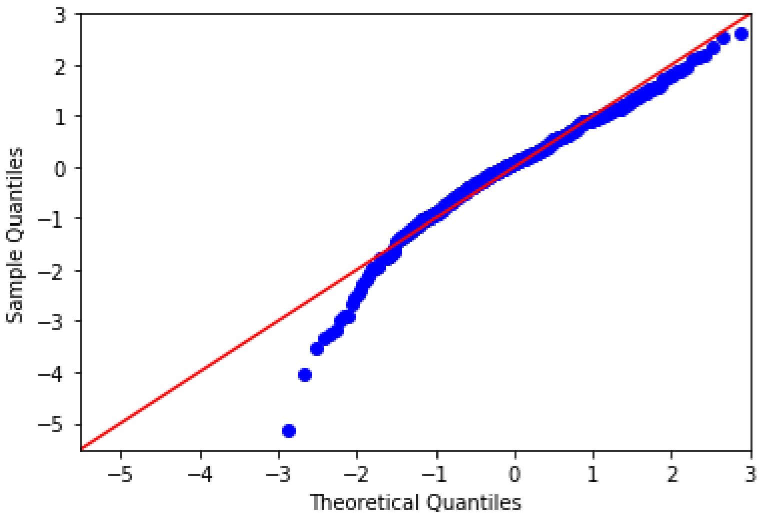
- First collect the data for plotting the quantile-quantile plot
- Next, sort the data in an order (ascending or descending)
- Plot a normal distribution curve
- Calculate the z-value (that is, cut-off point) for each segment
- Last, plot the dataset values against the normalizing cut-off points

If the data points lie approximately in a straight line, it can be concluded that the data point is normally distributed.

Sample Q-Q plot while doing the Bike sharing assignment on day.csv dataset :



```
sm.qqplot((y_train - y_train_pred), fit=True, line='45')
plt.show()
```


Sample Quantiles
Theoretical Quantiles