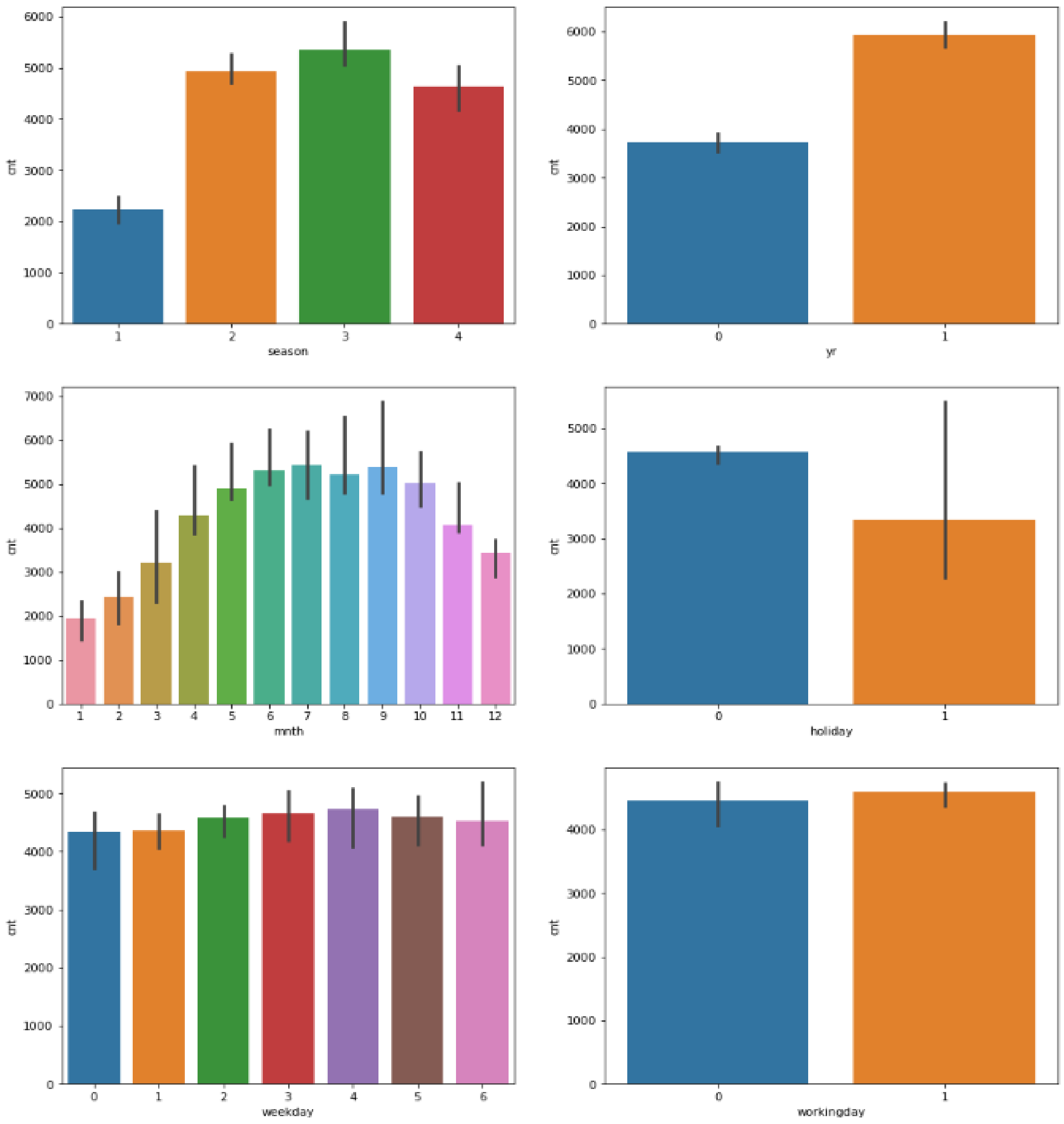
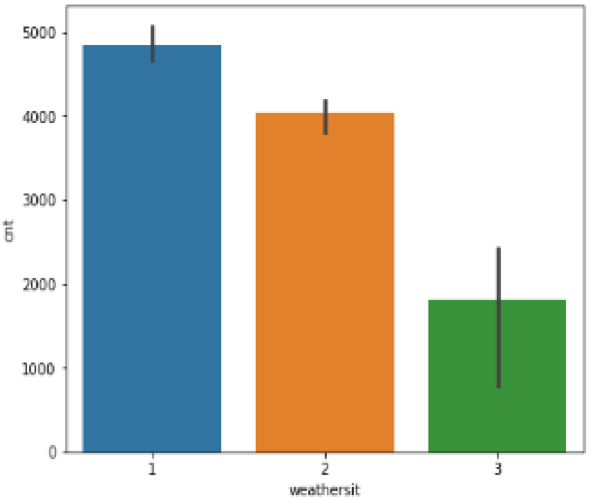
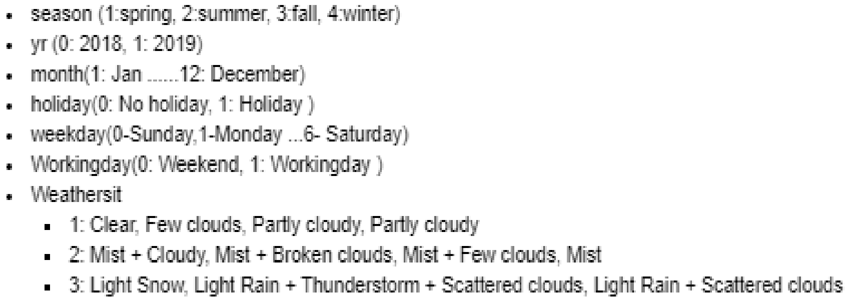
Assignment-based Subjective Questions

Question No.1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Categorical variables in data set are ***season***, ***yr***, ***mnth***, ***holiday***, ***weekday***, ***workingday***, ***weathersit***. Our dependent variable is **cnt.** Effect of categorical variables on dependent variable is as under:







- season (1:spring, 2:summer, 3:fall, 4:winter)
- yr (0: 2018, 1: 2019)
- month(1: Jan ......12: December)
- holiday(0: No holiday, 1: Holiday)
- weekday(0-Sunday,1-Monday ...6- Saturday)
- Workingday(0: Weekend, 1: Workingday)
- Weathersit
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

a. **Season:** There is a trend that max users are in fall and summer seasons and user count falls significantly in spring and winter season.

b. **Year:** Bike-sharing systems are slowly gaining popularity as awareness is increasing, the demand for these bikes is increasing every year.

c. **Month:** There are more users in the midyear (with maximum users in July, Aug & September) compared to starting and ending of year probably due to winters.

d. **Holiday:** On holidays there are relatively less demand compared to no holiday, however it is not consistent, indicating presence of outliers.

e. **Weekday/Working day:** There are comparably slightly more demand of bikes on working day compared to weekend.

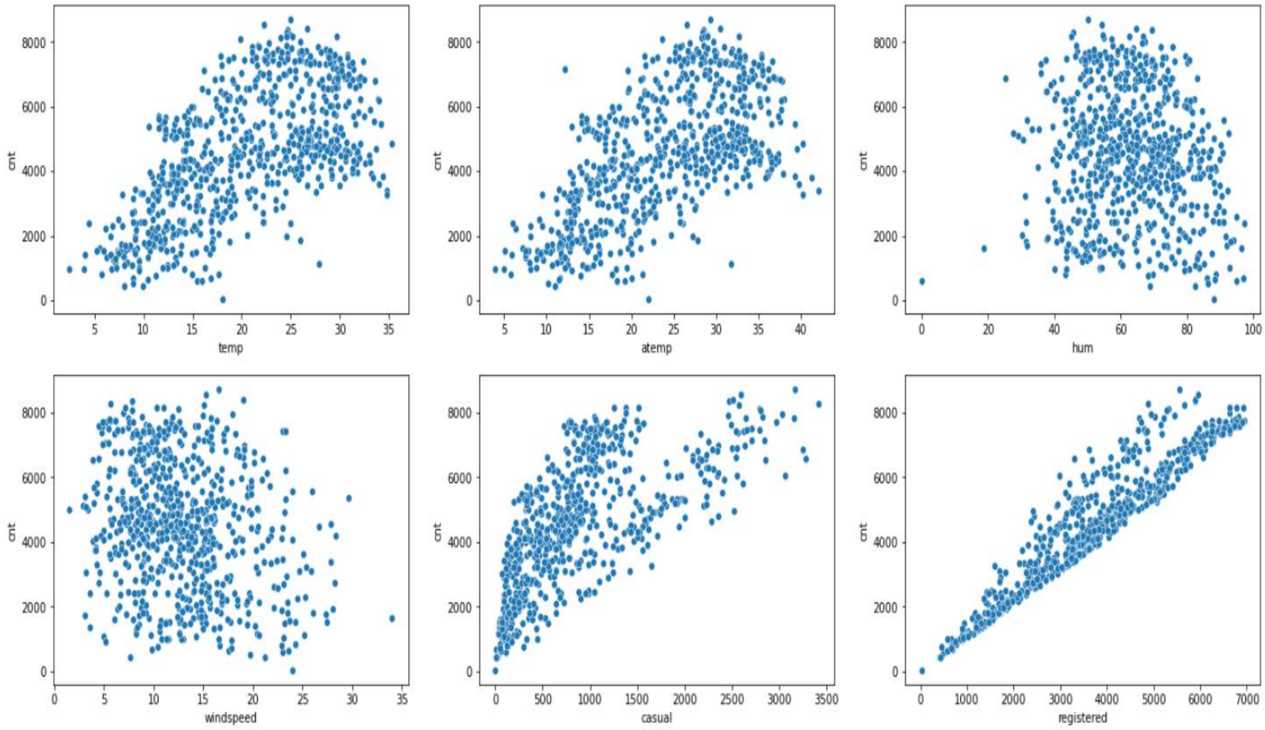
f. **Weathersit***:* As expected on rainy/snowy days demand is less compared to clear weather**.**

Question No.2 Why is it important to use drop_first=True during dummy variable creation?

Answer: As we know for categorical variable with 'n' no. of levels, 'n-1' no. of dummy variables is required for explaining the categorical variable, hence while creating dummy variables one dummy variable need to be deleted using drop_first=True argument which deletes first variable.

Question No.3 Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: Variation of numerical variables with target variable 'cnt' is as under:

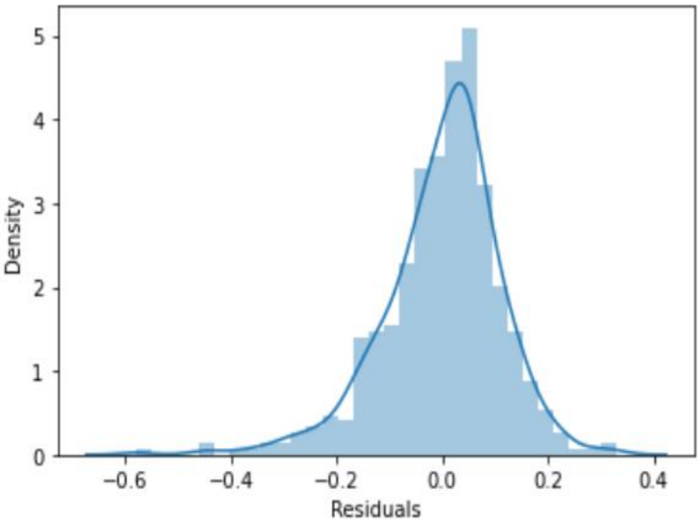


As evident from plots above, **target variables** is having highest correlation with ***registered users*** variable with a Pearson correlation coefficient of 0.95.

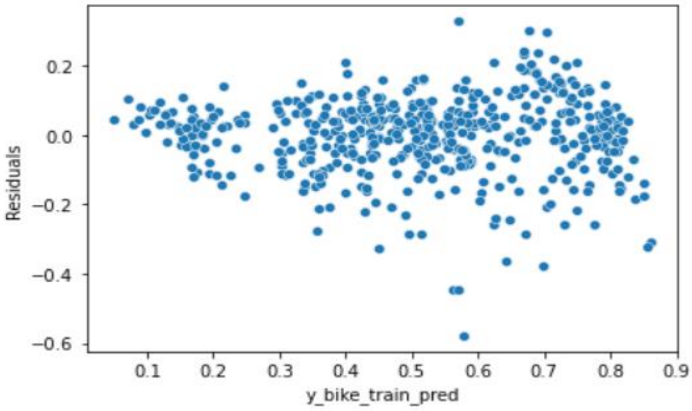
Question No.4 How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: After building the model on training set, Residual Analysis was carried out on training set as under:

1. Error terms /Residuals was calculated as
   Res= y(train)- y(train_pred)
2. Thereafter histogram was plotted for residual terms and it was found that residual/error terms are normally distributed with mean around zero. Thus validating the assumption that error terms are normally distributed with bell type curve.

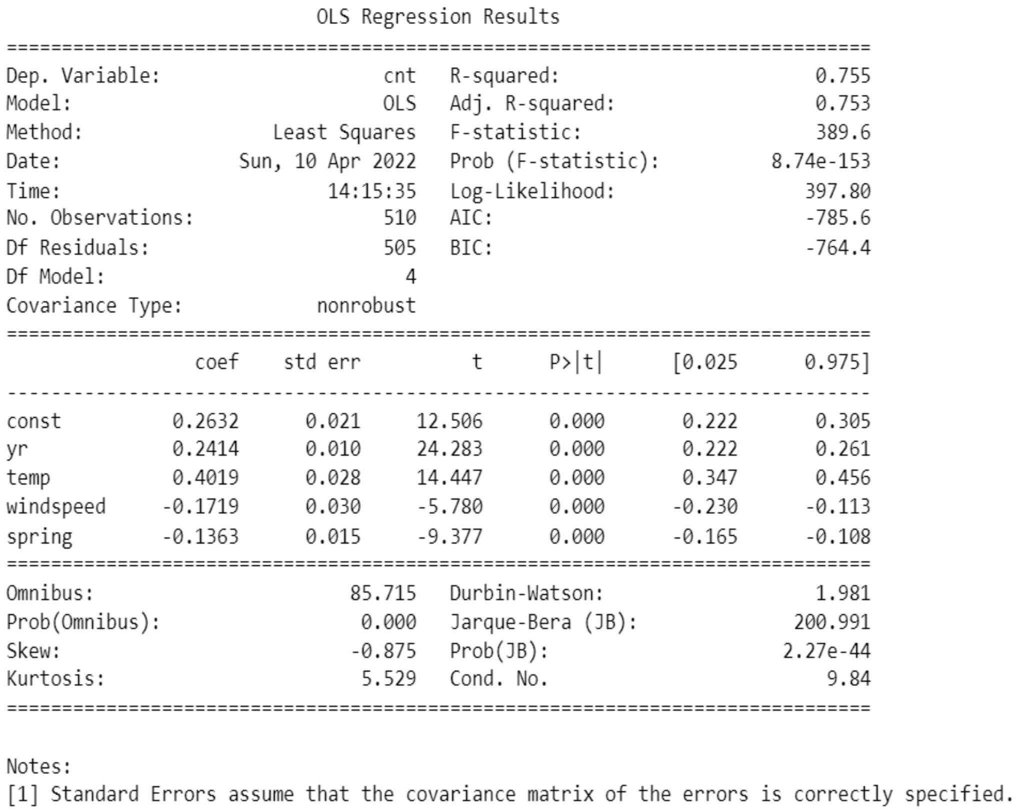


3. Subsequently, in order to validate assumption that error terms are independent , residuals/errors were plotted against y(train_pred) and it was found the residuals were randomly distributed and there is no clear pattern among the residuals /error terms i.e. homoscedasticity as under:



Question No.5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Based on final model, top three features contributing significantly towards explaining the demand of the shared bikes based on states model are as under:



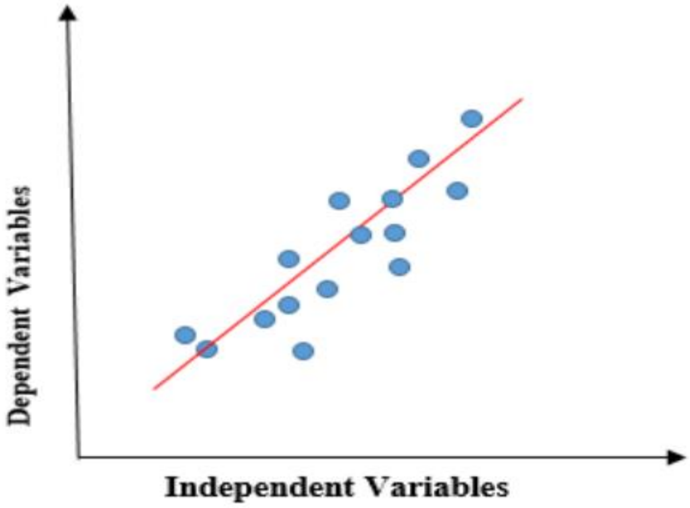
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.755
Model:                            OLS   Adj. R-squared:                  0.753
Method:                 Least Squares   F-statistic:                     389.6
Date:                Sun, 10 Apr 2022   Prob (F-statistic):          8.74e-153
Time:                        14:15:35   Log-Likelihood:                 397.80
No. Observations:                 510   AIC:                            -785.6
Df Residuals:                     505   BIC:                            -764.4
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2632      0.021     12.506      0.000       0.222       0.305
yr             0.2414      0.010     24.283      0.000       0.222       0.261
temp           0.4019      0.028     14.447      0.000       0.347       0.456
windspeed     -0.1719      0.030     -5.780      0.000      -0.230      -0.113
spring        -0.1363      0.015     -9.377      0.000      -0.165      -0.108
==============================================================================
Omnibus:                       85.715   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              200.991
Skew:                          -0.875   Prob(JB):                     2.27e-44
Kurtosis:                       5.529   Cond. No.                         9.84
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

a. **Temperature:** Temperature is having positive correlation with total no. of bike rentals. As the temperature rises after winters demand for bike increases and with decrease in temperature demand of bikes decreases.
b. **Year:** As the awareness about bikes for public transportation is increasing, over the years demand for rental bikes is also increasing.
c. **Wind speed:** Demand for rental bike is in negative relation to wind speed as expected.

General Subjective Questions:

Question No. 1: Explain the linear regression algorithm in detail?

Answer: Linear regression in Machine Learning is a supervised algorithm and the one of the most used regression algorithm. Linear regression means fitting the best fit line between independent and target variables with the least mean square error concept.



The above graph represents the linear relationship between the dependent variable and independent variable. The red line above is the best fit straight line. When the value of x (independent variable) increases, the value of y (dependent variable) is also increasing.

Before implementing linear regression, we should check whether the data is following these assumptions:

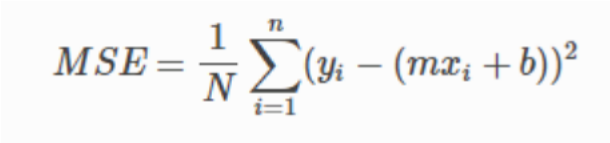
- Data should be linear
- No Multicollinearity
- No auto-correlation
- Error terms should be normally distributed
- Homoscedasticity should be there in residuals/errors

As data is linear, simple linear equation is as under:

$$Y = mx+b$$

We can calculate MSE (mean square error) as under:

If y = actual values, y_i = predicted values



$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

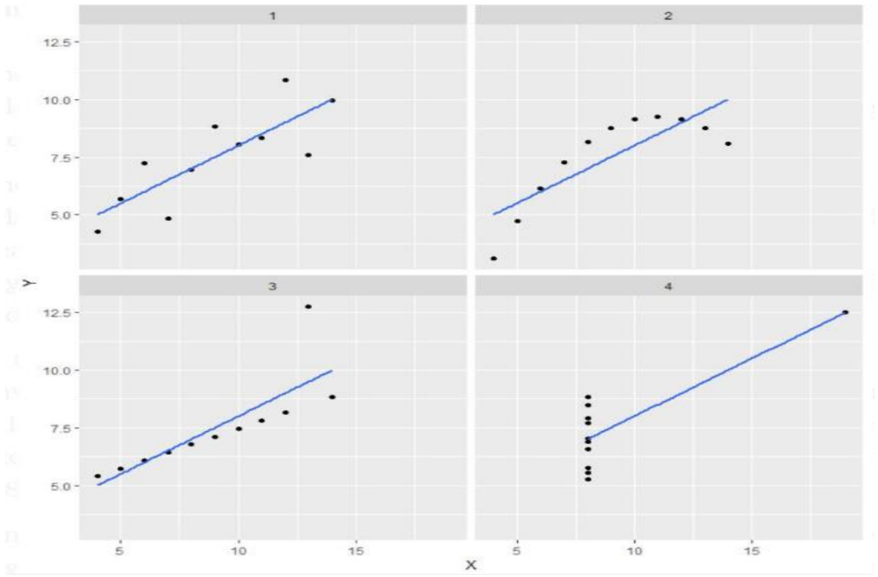
Now above value has to be minimized, which can be achieved by Gradient Descent approach by using suitable Learning Rate(alpha).

Linear Regression can be implemented in Python using Sci-Kit Learn and Statmodels libraries and the general approach is as under:

1. Reading and Understating Data.
2. Visualizing Data
3. Data Preparation
4. Splitting Data into train and test set.
5. Scaling data
6. Building a Linear model using Sci-kit Learn or Statmodels library
7. Residual Analysis
8. Making Predictions Using the Final Model on test data
9. Model Evaluation.

Question No. 2: Explain the Anscombe’s quartet in detail?

Answer: Anscombe’s quartet comprises four datasets that have almost same statistical properties, yet they have very different appearance when plotted as brought out below for different values of X and Y.



<u>Explanation of this output:</u>

- In the first one(top left) if we look at the scatter plot we will see than there is nealry linear relationship between x and y.
- In the second one(top right) if we look at this plot we can see that there is a non-linear relationship between x and y.
- In the third one(bottom left) we can say when there is a perfect linear relationship for all the data points except one which looks like to be an outlier which is indicated far away from the line.
- Finally, the fourth one(bottom right) shows an example when one distant point is enough to produce a high correlation coefficient.

However, if we check mean, Standard deviation and correlation for all above plots, they are same for each plot.

<u>Application:</u>

The quartet is often used to show the importance of looking at a data set graphically before starting to analyze according to a particular type of relationship, and the in effectiveness of basic statistic properties for describing real datasets.

Question No. 3: What is Pearson’s R?

Answer: Pearson’s R is a statistic that measures the linear correlation between two variables. It has a numerical value that lies between -1.0 and +1.0. In Statistics, the Pearson's r is also referred to as Pearson's Correlation Coefficient, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. Pearson's Correlation Coefficient is named after Statistician Karl Pearson. He formulated the correlation coefficient in the 1880s. Pearson's correlation coefficient cannot identify nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. The formula for calculating Pearson's correlation coefficient is as under:

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

Where,

N = the number of pairs of dataset

Σxy = the sum of the products of paired dataset

Σx = the sum of x in dataset

Σy = the sum of y in dataset

Σx2 = the sum of squared x in dataset

Σy2 = the sum of squared y in dataset

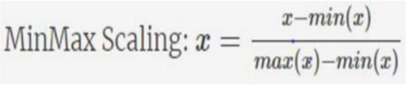
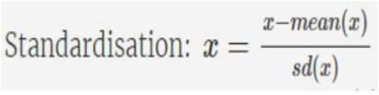
Question No. 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: **Scaling:** Scaling is a step of data Pre-Processing which is applied to variables to convert the data within a particular range. Scaling can be done through Normalization or Standardization using SciKit Library in python.

Reasons for Scaling:

1. Sometimes data set contains features highly varying in magnitudes, units and ranges. If scaling is not done then model only takes magnitude in account and not units hence leading to incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Also scaling helps in interpreting effect of each feature in final model.
2. It also helps in speeding up the calculations in an model as gradient descent is achieved early.

Difference between normalized scaling and standardized scaling:

Sl. No.	Normalized scaling	Standardized scaling
1.	Formula for Normalization is as under:  MinMax Scaling: $x = \frac{x-min(x)}{max(x)-min(x)}$	Formula for standardization is as under:  Standardisation: $x = \frac{x-mean(x)}{sd(x)}$
2.	The Range of converted values is between 0 to 1 or -1 to 1.	The Range of converted values is between not necessarily 0 to 1, range of values are distributed between positive and negative values.
3.	Mean and standard deviation is non zero.	Mean of values is zero and standard deviation is one.
4.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization
5.	It is really affected by outliers.	It is much less affected by outliers.
6.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

Question No. 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: VIF = infinity shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) as 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).

Question No. 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Answer: Q-Q (Quantile-Quantile) plot, is a graphical tool to help us assess if a set of data possibly came from some distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Q-Q plot also helps in Linear regression, when train and test set are recived separately and then we can confirm using Q-Q plot that both the train and test data came from population with same distribution.

Few advantages:

a) It can be used with sample sizes also

b) Q-Q plot can be used to detect many distributional aspects:

- Shifts in location,
- Shifts in scale,
- Changes in symmetry
- Presence of outliers can all be detected from this plot.