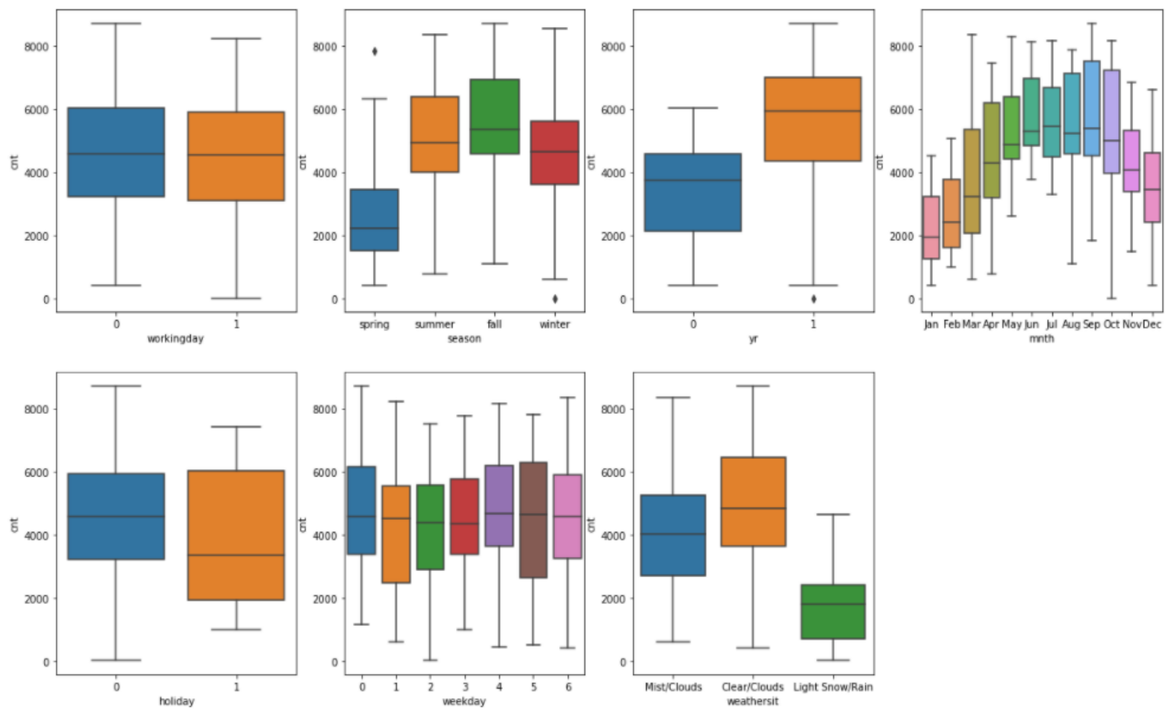
Assignment-based Subjective Questions

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

My inference from the analysis of categorical variables is as mentioned below:

- demand of share bike is same on weekday and weekend.
- Demand of share bike is most in falls in least in spring season.
- Demand has increased drastically in year 2019 as compared to year 2018.
- Demand is more in the month of Aug, Sep and Oct of the year and least in Jan, Feb.
- Demand is least when the weather is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; and is high when it is Clear, Few clouds, Partly cloudy, Partly cloudy type of weather.
- on holidays, the 25th and 50th quartile has less demands as compared to non-holidays.

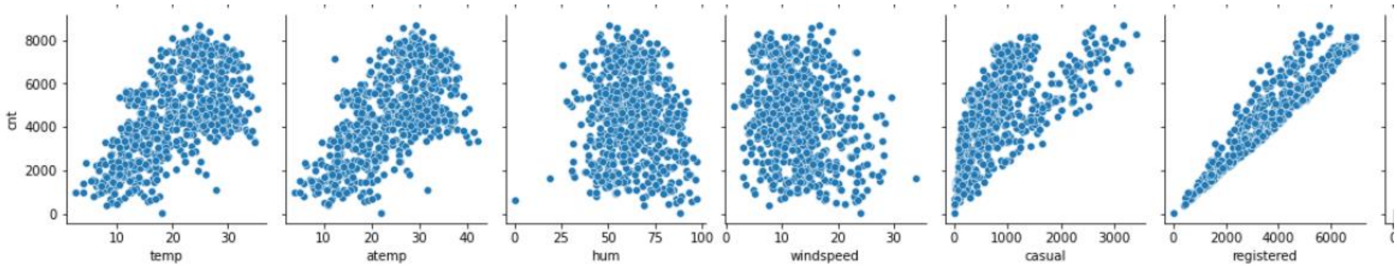


2. Why is it important to use drop_first=True during dummy variable creation?

Drop_first=True is used during dummy variable creation to avoid redundant features and hence to improve the VIF (multi-collinearity). As the created features could be correlated and this should be avoided in linear regression model building process.

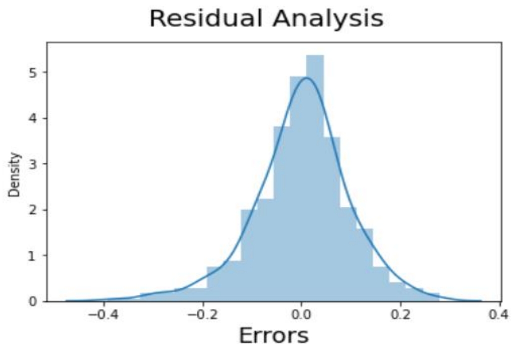
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The feature ‘Registered’ has highest correlation with the target variable ‘cnt’, but we cannot use the ‘registered’ feature in model building process as cnt is summation of casual and registered so it will always have high correlation with it. After dropping ‘Registered’ feature, ‘temp’ is having highest correlation with the target variable ‘cnt’.

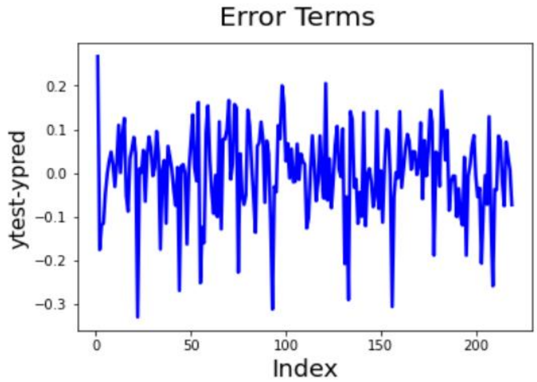


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

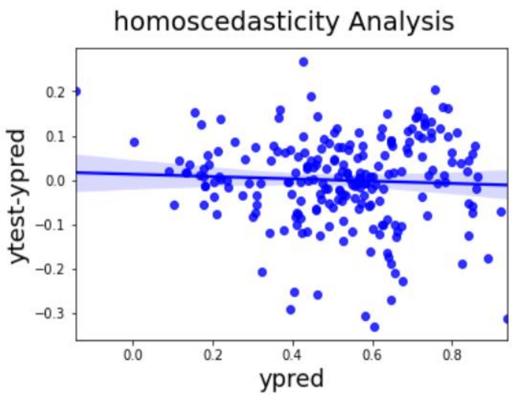
- Residual Analysis is done to check the distribution of error terms. Error terms (y_actual – y_pred) should have normal distribution as shown below:



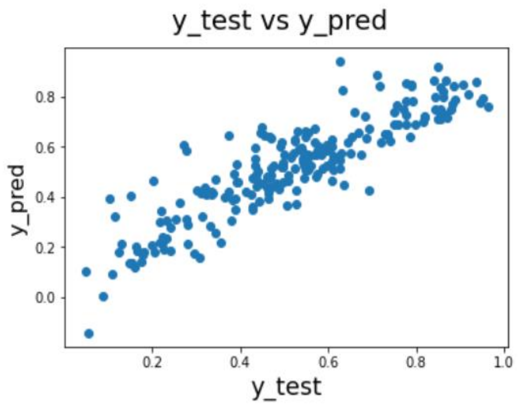
- Plotting the error terms shows no pattern and hence they are independent of each other and shows the constant variance as shown below:



- After plotting scatter plot of ypred against ytest, we can see the pattern that depicts the homoscedasticity:



- The relationship between ypred and ytest is as shown below:



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

Top 3 features contributing significantly towards explaining the demand of the shared bikes are as mentioned below:

- temp
- hum
- yr

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.815
Model:                            OLS   Adj. R-squared:                  0.812
Method:                 Least Squares   F-statistic:                     244.7
Date:                Sat, 03 Jul 2021   Prob (F-statistic):          7.01e-177
Time:                        16:19:17   Log-Likelihood:                 468.75
No. Observations:                 510   AIC:                            -917.5
Df Residuals:                     500   BIC:                            -875.1
Df Model:                           9
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                0.2687      0.027      9.822      0.000       0.215       0.322
yr                   0.2253      0.009     25.751      0.000       0.208       0.242
holiday             -0.1010      0.028     -3.668      0.000      -0.155      -0.047
temp                 0.6575      0.023     28.516      0.000       0.612       0.703
hum                 -0.2770      0.033     -8.336      0.000      -0.342      -0.212
windspeed           -0.1897      0.027     -6.948      0.000      -0.243      -0.136
winter               0.1143      0.011     10.765      0.000       0.093       0.135
Light Snow/Rain     -0.1941      0.027     -7.217      0.000      -0.247      -0.141
Jul                 -0.0911      0.018     -4.950      0.000      -0.127      -0.055
Sep                  0.0629      0.017      3.766      0.000       0.030       0.096
==============================================================================
Omnibus:                       31.085   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.472
Skew:                          -0.442   Prob(JB):                     1.81e-11
Kurtosis:                       4.243   Cond. No.                         15.1
==============================================================================
```

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is supervised machine learning algorithm used for below purpose:

- To predict continuous values
- To identify the driver features
- Measure the impact of independent variables(predictors) on target dependent variables.

Based on the number of independent features, it is classified into 2 categories:

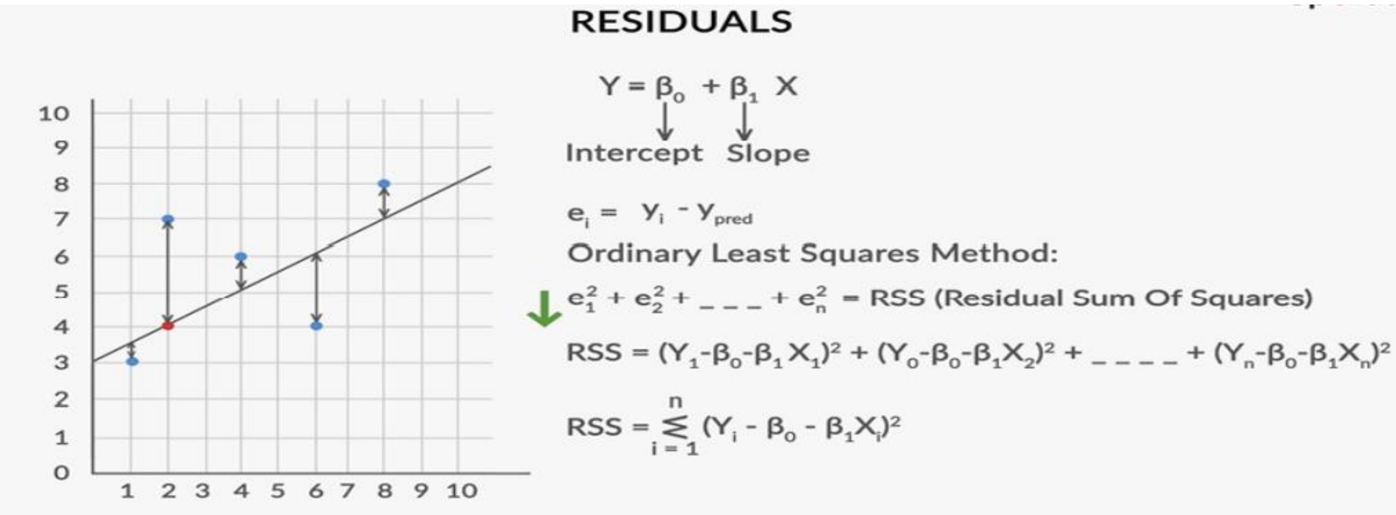
1) Simple linear regression → 1 independent variable. We get a line when the predicted value is plotted against the independent variable. The formula of simple linear regression is:

   y = mx+c (where m is the coefficient and c is the intercept)

2) Multi linear regression → more than 1 independent variables. We get a hyperplane when the predicted value is plotted against the predictors. The formula of multi-linear regression is:

Y=c+m1x1+m2x2+m3x3+….

Since there are infinite lines possible from a given data point, hence simple linear regression aims at finding the best fit line. The 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 for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



A small RSS means a tight fit of the model to the data.

Cost function and Gradient descent:

The cost function determines how far the prediction is from the original dependent variable. The formula is as shown below:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\Sigma(h_i - y_i)^2$$

Where m is the total number of data points, hi is the actual value of the dependent variable and yi is the predicted value of the dependent variable.

The idea of any machine learning algorithm is to minimize the cost function so that the hypothesis is close to the original dependent variable. We need to optimize the theta value to do that. If we take the partial derivative of the cost function based on theta0 and theta1 respectively, we will get the gradient descent. To update the theta values, we need to deduct the gradient descent from the corresponding theta values:

$$\theta_0 = \theta_0 - \alpha\frac{d}{d\theta_0}J(\theta_0)$$
$$\theta_1 = \theta_1 - \alpha\frac{d}{d\theta_1}J(\theta_1)$$

Where alpha is the learning rate which is the speed at which we want to move towards negative of the gradient. Here we keep on calculating the Theta value until the cost function does not change further. That cost function is said to be reduced and the resulting equation of the line is called **best fit line**.

The strength of linear regression model can be assessed using 2 metrics:

1) R^2 or Coefficient of determination → It explains what portion of the given data variation is explained by the developed model. It takes values between 0 and 1. Higher the R^2, better the model. Mathematical

formula of R^2 is R^2 = 1 - (RSS / TSS). Where RSS is residual sum of squares and TSS is total sum of squares which is sum of squares of difference between actual and mean of actual output values.

2) Residual standard Error (RSE) → It helps in measuring the lack of fit of a model on the given data points. Mathematically, it is calculated as square root of RSS/df. Where RSS is the residual sum of squares and df is the degree of freedom which is no.of data points – 2

Assumptions of simple linear regression:

1) What does simple linear regression do:
    a. At each x, there is a distribution of the values of y
    b. Model predicts a single value, therefore there is a distribution of error terms.
2) Linear relationship between x and y
3) Error terms (Residuals) are normally distributed.
4) Error terms are independent of each-other.
5) Error terms of constant variance(homoscedasticity).

Hypothesis testing in Linear regression → Tests whether the coefficient is significant or not

1) Start by saying that the coefficient is not significant which is our null hypothesis.
2) Hence, the alternate hypothesis is that the coefficient is significant.
3) Calculate the t-score of the coefficient. Then derive the p-value from the cumulative probability for the given t-score using t-table.
4) Make the decision based on p-value. If p-value < 0.05 then reject the null hypothesis which means the relationship is significant else non-significant.

Parameters to assess a linear model are:

1) T-test → Used to determine the p-value and hence helps in determining whether the coefficient is significant or not. P-value should not be more than 0.05 for a coefficient to be significant.
2) F-stat → It says if the overall model fit is significant or not. Generally, the higher the value of F-stat the more significant the model turns out to be. P(F-stat) is inversely proportional to the F-stat.
3) R-squared → After it has been concluded that the model fit is significant, the R^2 tells the extent of the fit. It explains how well the straight line describes the variance in the data. Its value ranges from 0 to 1 with 1 being best fit and 0 the worst fit.

<u>Multi – Linear regression</u>:

As explained earlier, It is the model used to predict one target variable based on more than one independent features. The mathematical formula looks like below:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

- Interpretation of co-efficient → Change in the mean response E(Y), per unit increase in the variable when other predictors are held constant.
- Here coefficients are still obtained by minimising sum of squared error (least square criteria).
- For inference, the assumptions from simple linear regression still hold here along with few new considerations as mentioned below:
    - Model may overfit by becoming too complex with too many features. Symptoms are below:
        - Model fits the train data set too well and hence does not generalise.
        - High train accuracy on train dataset and low accuracy on test dataset.

- Multicollinearity → association between predictor variables. Associated features should be avoided to remove redundancy. Hence feature selection becomes an important aspect here.
  - Affects the interpretation (Change in Y), Inference (Coefficient swings wildly, signs can invert, and p-value does not remain reliable)
  - Does not affect the prediction and precision of the prediction. Stats such as R^2 are not impacted.
  - Multi-collinearity detection → VIF (Variance Inflation Factor). High VIF (more than 5) indicates multi-collinear feature. The mathematical formula is 1/(1-R^2)
  - Dealing with multi-collinearity
    - Dropping variables. Drop highly correlated variables and pick the business interpretable variables.
    - Create new variables using interaction of older variables and drop original features.
- Dealing with Categorical variables → use dummy indicator values (0 and 1). If the categorical variable has n levels, then n-1 new dummy variables are created.
- Scaling a variable → In order to bring all the independent features on the same scale, all the numerical variables are scaled. It is done to make the coefficients reliable. Scaling does not impact model accuracy and stats (p-value, R^2, f-stat, t-stat). There are 2 ways to perform scaling:
  - Standardization (where mean = 0 and standard deviation = 1)
  - MinMax Scaling (between 0 and 1). This method is used frequently.

Selecting the best model:
Since there are many ways to build models, then the question arises on which is the best model. The best is chosen based on trade-off between explaining highest variance and keeping the model simple. Key idea is to penalise the models for using high number of predictors. Hence, we use a new stat called **adjusted R^2** instead of R^2. The mathematical formula of adjusted R^2 looks like below:

$$1 - \left[\frac{(1 - R^2) \times (n - 1)}{(n - k - 1)} \right]$$

where n = sample size, k = number of predictors.

Feature Selection:
In case of multi-linear regression in order to reduce the complexity and improve the model fit, we need to select the features carefully. There are 3 approaches of feature selection as described below:

- Manual feature elimination → This is used when the number of predictors is relatively low (10 to 20)
  - Build model with all the features
  - Drop features one by one that are least helpful in prediction (high p-value, high VIF, low correlation) and rebuild the model and repeat the process
- Automated approach
  - Select top 'n' features using RFE(Recursive feature elimination) process
  - Build the model using these features
- Mixed or balanced approach → First model is built using automated approach and then manual approach is applied on the first build model. This is the preferred option.

Finally, the steps involved in building a Linear/Evaluating a Linear model:

1) Load the data
2) Perform basic cleaning
3) Inspect the meta data → rows, columns, distribution, shape etc
4) Split the train and test datasets
5) Perform missing value imputation on train data set

6) Perform outlier treatment on train data set
7) Perform scaling on train data set
8) Perform feature selection using RFE
9) Model building
10) Residual analysis
11) Model evaluation on test data set
12) Target prediction

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough". It has been rendered as an actual musical quartet.

The 4 data sets are as shown below:

Anscombe's quartet

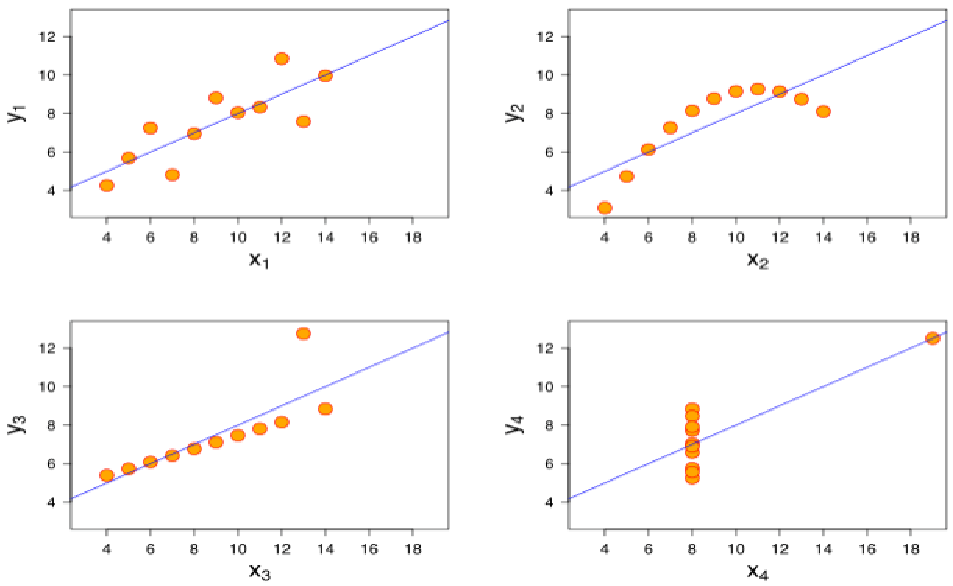
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Stats summary of the 4 data sets as shown below:

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 |
+-----+---------+-------+---------+-------+----------+
```

It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed as shown below:



Explanation of this output:

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

Application: The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

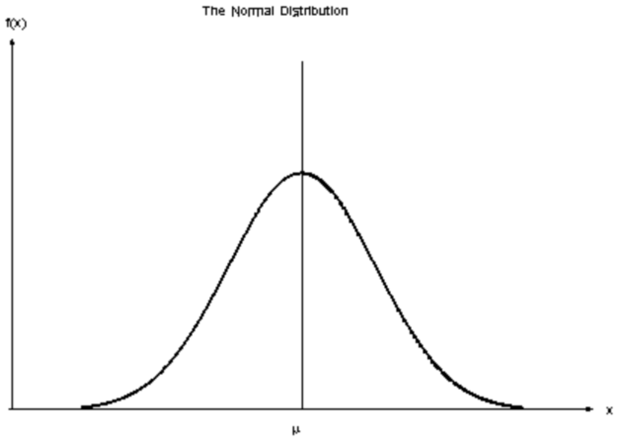
Pearson's correlation coefficient (Pearson's R) is the test statistics that measures the statistical relationship, or association, between two continuous variables.
It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Kinds of questions a Pearson correlation answers:

- Is there a statistically significant relationship between age and height?
- Is there a relationship between temperature and ice cream sales?
- Which two variables have the strongest co-relation between age, height, weight, size of family and family income?
- Is there a relationship among job satisfaction, productivity, and income?

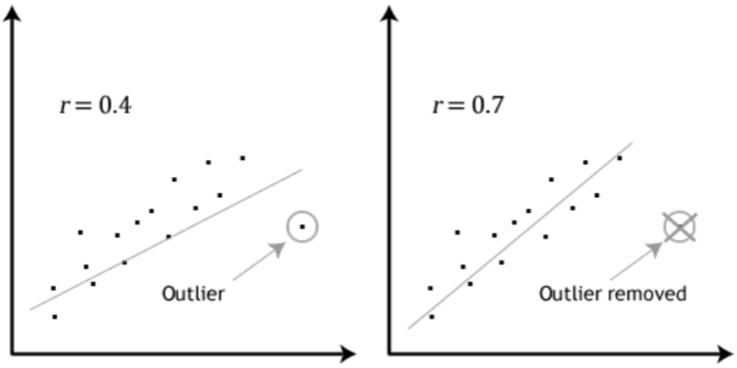
Assumptions:

1) For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve' as shown below:



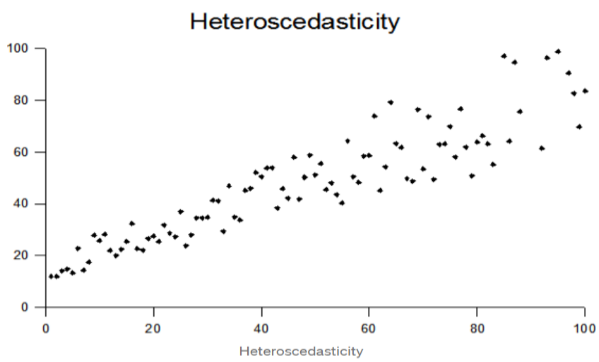
Normal Distribution

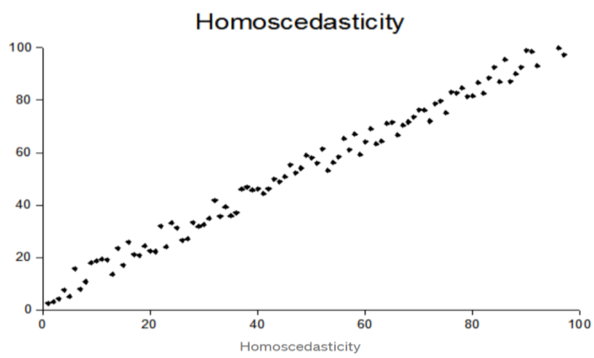
2) There should be no significant outliers. Pearson’s correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means that including outliers in the analysis can lead to misleading results.



Outliers

3) Each variable should be continuous.
4) The two variables have a linear relationship. Scatter plots helps to tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption.
5) The observations are paired observations. For every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example, if you are calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.
6) **Homoscedasticity** → It describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. A scatterplot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedasticity is heteroscedasticity (the violation of homoscedasticity) which is present when the size of the error term differs across values of an independent variable.





Properties:

1) Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
2) Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3) Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

1) Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2) High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
3) Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4) Low degree: When the value lies below + .29, then it is said to be a small correlation.
5) No correlation: When the value is zero.

Mathematical Formula:

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

What do the terms strength and direction mean?

The terms 'strength' and 'direction' have a statistical significance as mentioned below:

1) Strength: Strength signifies the relationship correlation between two variables. It means how consistently one variable will change due to the change in the other. Values that are close to +1 or -1 indicate a strong relationship. These values are attained if the data points fall on or very close to the line. The further the data points move away, the weaker the strength of the linear relationship. When there is no practical way to draw a straight line because the data points are scattered, the strength of the linear relationship is the weakest.
2) Direction: The direction of the line indicates a positive linear or negative linear relationship between variables. If the line has an upward slope, the variables have a positive relationship. This means an increase in the value of one variable will lead to an increase in the value of the other variable. A negative correlation depicts a downward slope. This means an increase in the amount of one variable leads to a decrease in the value of another variable.

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

Scaling 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 must 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.

Normalized scaling → It is also known as MinMax scaling. It brings all the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python. The mathematical formula of MinMax scaling is as below:

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

Standardized scaling → Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean 0 and standard deviation 1. sklearn.preprocessing.scale helps to implement standardization in python. The mathematical formula of standardization is as shown below:

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

Comparison of Normalisation Vs Standardisation:

S.No.	Normalisation	Standardisation
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.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

If all the independent variables are orthogonal to each other, then VIF = 1.0.
If 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). For example, as shown below the first 3 features are highly correlated and hence the VIF is 'inf' here:

	Features	VIF
9	Clear/Clouds	inf
10	Light Snow/Rain	inf
11	Mist/Clouds	inf
6	spring	5.20
3	temp	3.97
8	winter	3.77
7	summer	2.67
4	hum	1.91
12	Jan	1.57
13	Jul	1.49
14	Nov	1.48
15	Sep	1.31
5	windspeed	1.21
1	yr	1.04
2	holiday	1.03
0	const	0.00

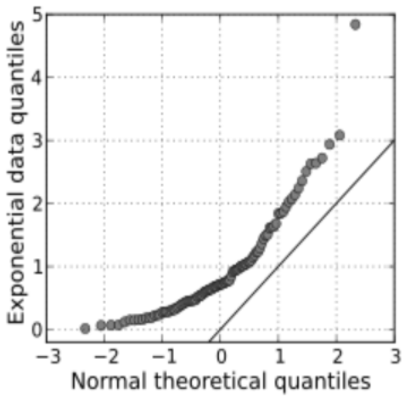
If VIF is large and multicollinearity affects the analysis results, then you need to take some corrective actions before proceeding multiple regression. Here are the various options:

- One approach is to review the independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if the inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then we may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe the independent variables. Using this approach will get rid of the multicollinearity problem but it may be hard to interpret the meaning of these "new" independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.

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

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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.



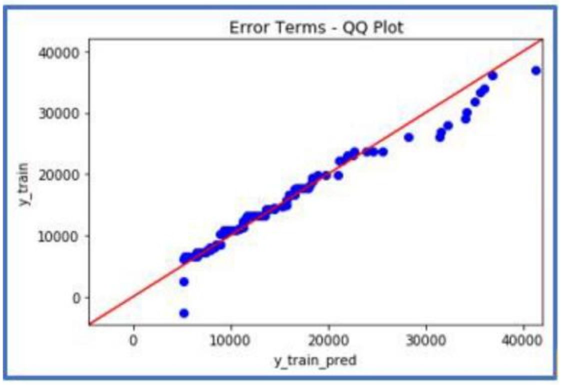
The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It is being compared to a set of data on the y-axis. This Q-Q plot is called a normal quantile-quantile (QQ) plot. The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

It is used to check if two data sets:

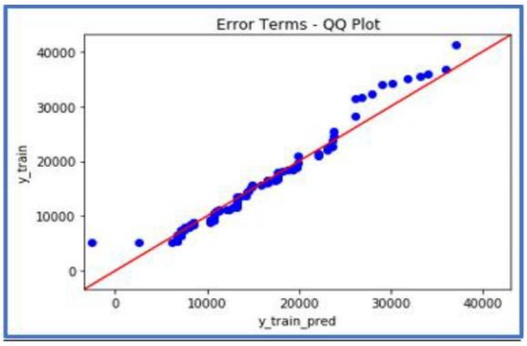
- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

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.

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

Q-Q Plots and the Assumption of Normality - The assumption of normality is an important assumption for many statistical tests. Let us assume we are sampling from a normally distributed population. The normal Q-Q plot is one way to assess normality. However, we do not have to use the normal distribution as a comparison for the data; we can use any continuous distribution as a comparison if we can calculate the quantiles. In fact, a common procedure is to test out several different distributions with the Q-Q plot to see if one fits the data well.