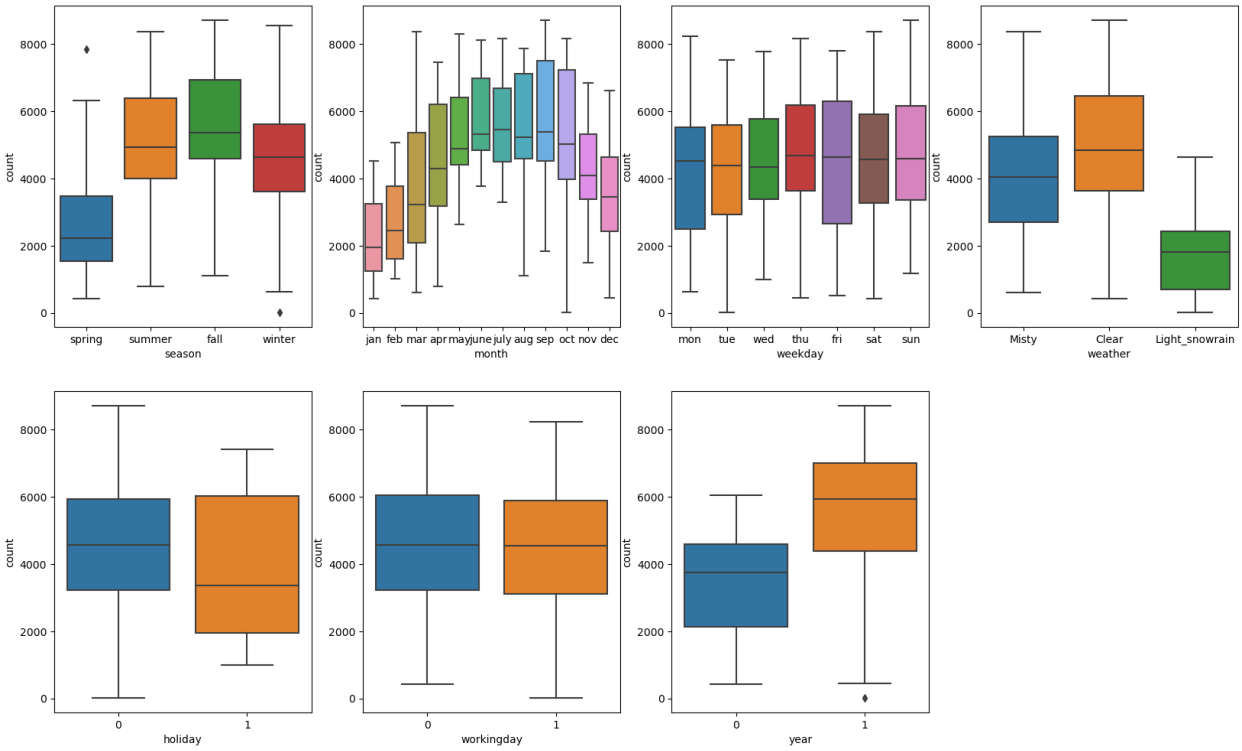
Assignment-based Subjective Answers

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

Ans: From my analysis of the categorical variables from the dataset, I can say:

- ★ Bike demand in the fall is the highest.
- ★ Bike demand takes a dip in spring.
- ★ Bike demand in 2019 is higher as compared to 2018.
- ★ Bike demand is high in the months from May to October.
- ★ Bike demand is high if the weather is clear or with Mist cloudy while it is low when there is light rain or light snow.
- ★ The demand for bikes is almost similar throughout the weekdays.
- ★ Bike demand doesn't change whether the day is a working day or not.



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

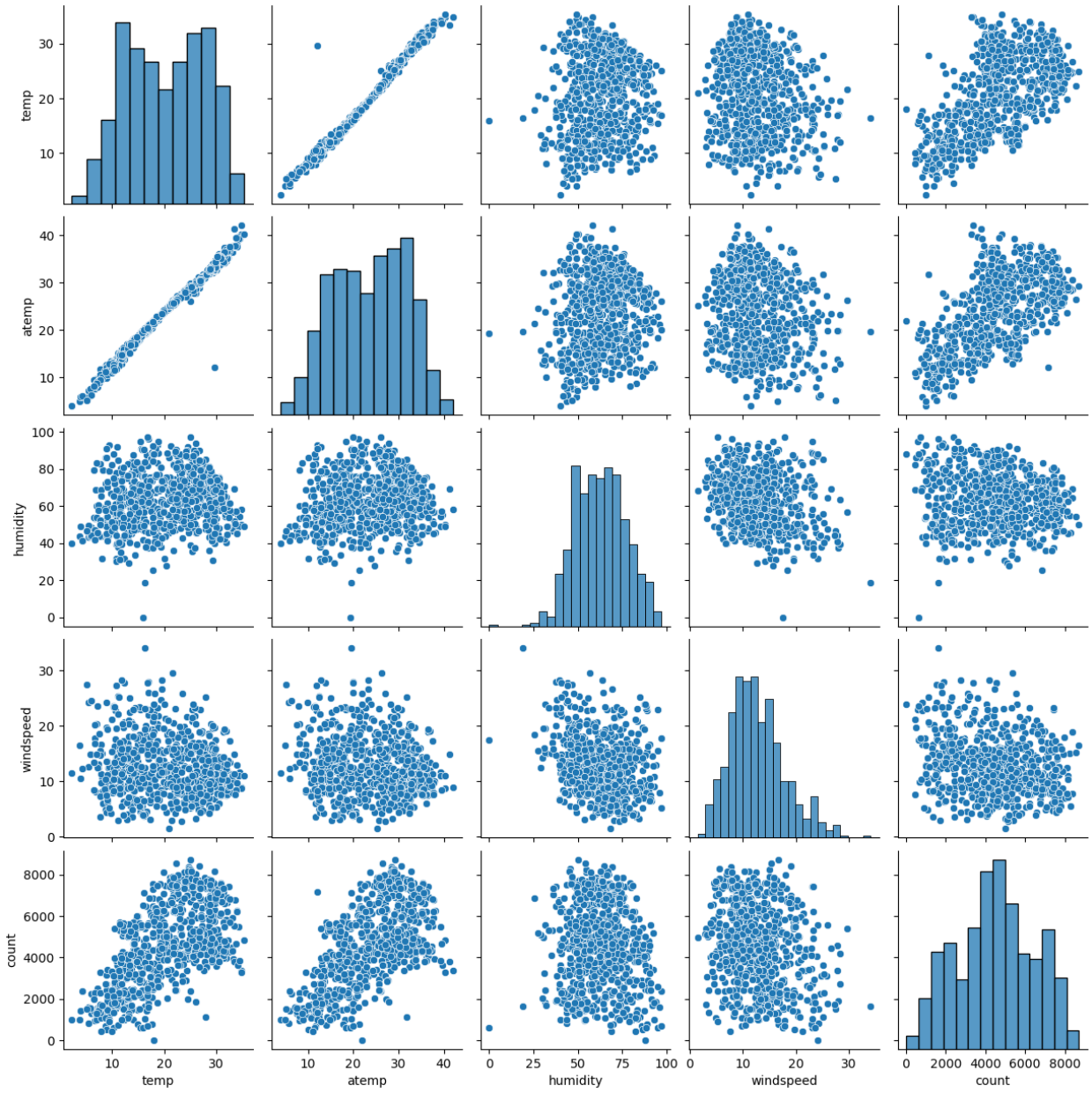
Ans. Use of *drop_first = True* is important in order to achieve k-1 dummy variables as it can be used to delete extra columns while creating dummy variables.

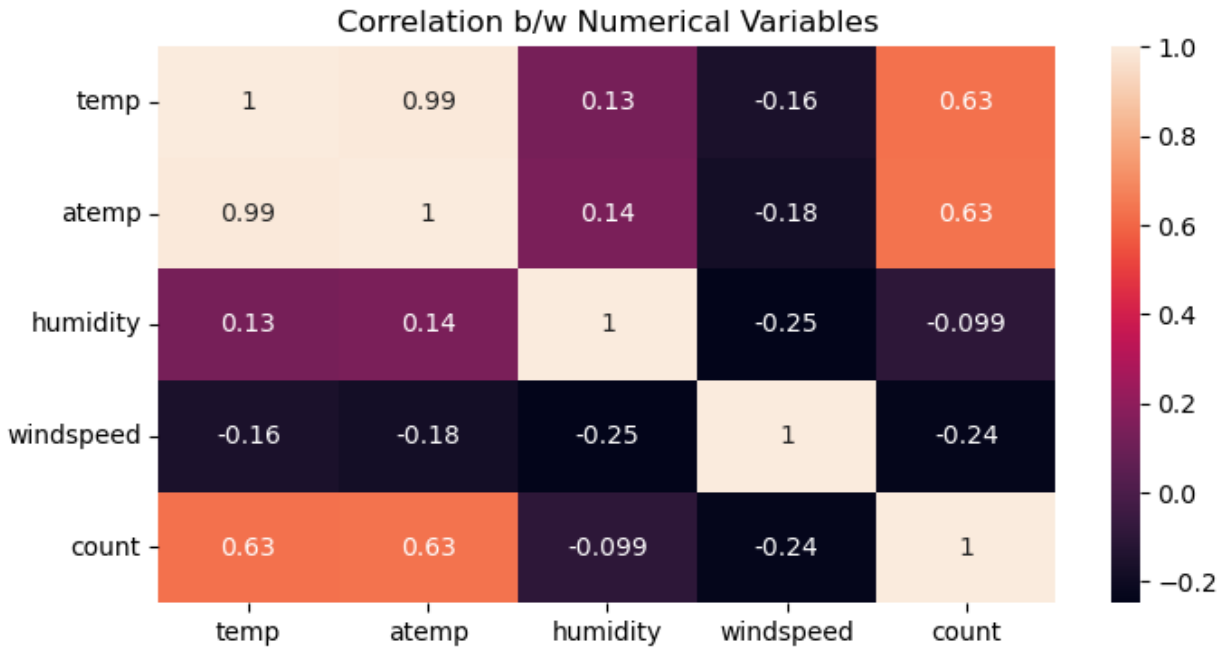
For Example: We have three variables: Furnished, Semi-furnished and unfurnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished, So we can remove it.

It is also used to reduce the collinearity between dummy variables.

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

Ans. By looking at the pair plot & heat map, temp & atemp variables have the highest (0.63) correlation with target variable 'cnt' (which was renamed as 'count').


temp
atemp
humidity
windspeed
count
temp
atemp
humidity
windspeed
count



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

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

- weathersit_Light_Snow(Negative correlation).
- year(Positive correlation).
- temp(Positive correlation).

General Subjective Answers

1. Explain the linear regression algorithm in detail.

Linear regression:

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

Regression analysis is used for three types of applications:

- Finding out the effect of Input variables on Target variables.
- Finding out the change in Target variable with respect to one or more input variables.
- To find out upcoming trends.

Here are the types of regressions:

1. Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Polynomial Regression

“Linear regression is one of the very basic forms of machine learning in the field of data science where we train a model to predict the behaviour of your data based on some variables. linear regression as the name suggests variables which are linearly correlated to each other on x-axis and y-axis.”

In some cases, the value will be linearly upward, that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example of that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

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

Mathematically, we can write a linear regression equation as:

$$y=a+bx$$

Or

$$y=mx+c$$

Or

$$y=\beta 0+\beta 1x$$

what-is-linear-regression-2
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

Use Cases of Linear Regression:

Prediction of trends and Sales targets - To predict how industry is performing or how many sales targets industry may achieve in the future.
Price Prediction - Using regression to predict the change in price of stock or product.
Risk Management - Using regression to the analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe's quartet in detail.

<u>Anscombe's Quartet</u>:

"'Anscombe's Quartet' can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots."

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

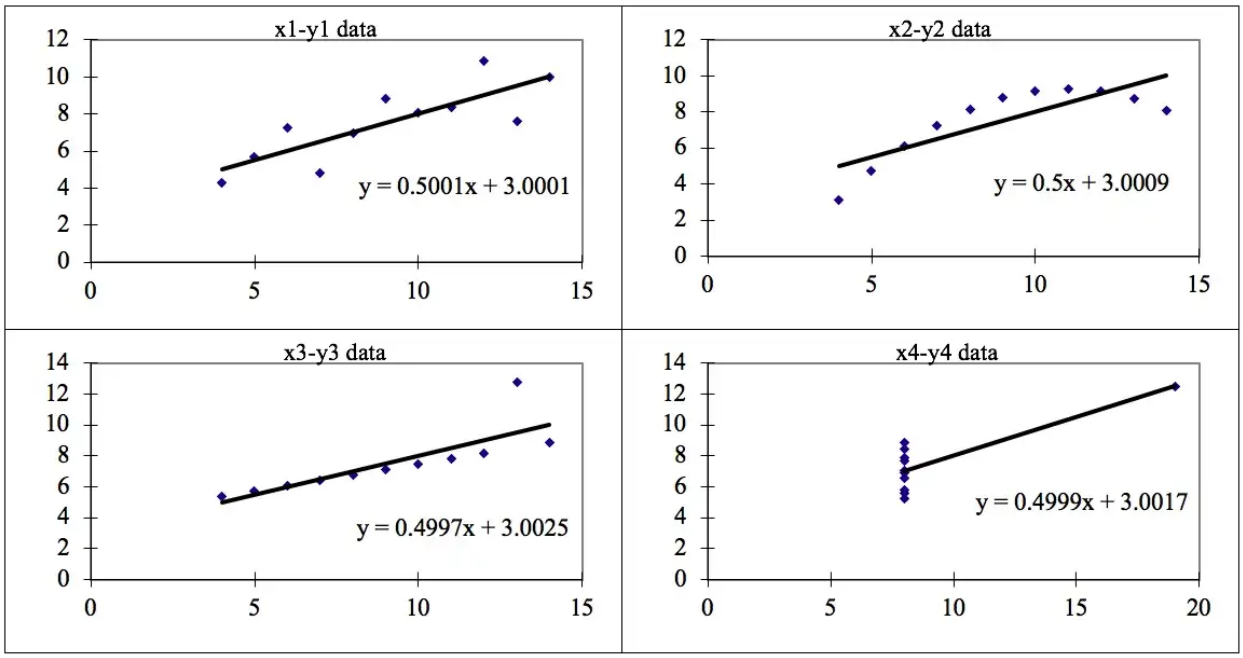
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

Anscombe's Data										
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

The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data										
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	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		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	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well (look at x1, y1).
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear (look at x2, y2).
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by a linear regression model (look at x3, y3).
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by a linear regression model (look at x4, y4).

We can **conclude** by saying, we have described the four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

<u>Pearson's R</u>:

'Pearson correlation coefficient' is a measure of the strength of a linear association between two variables - denoted by r.

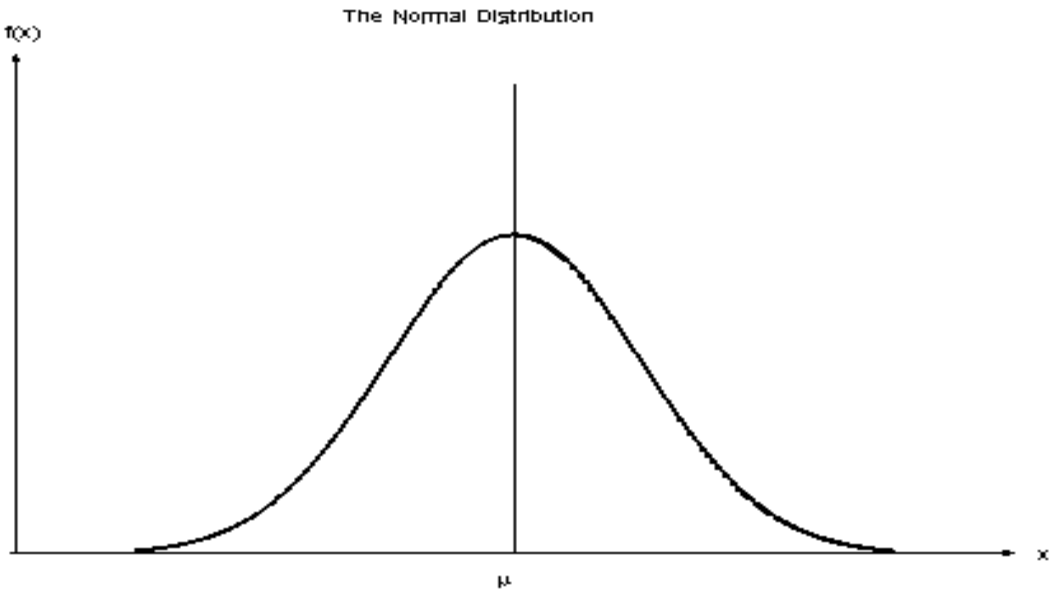
In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the 'Pearson product-moment correlation coefficient' (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

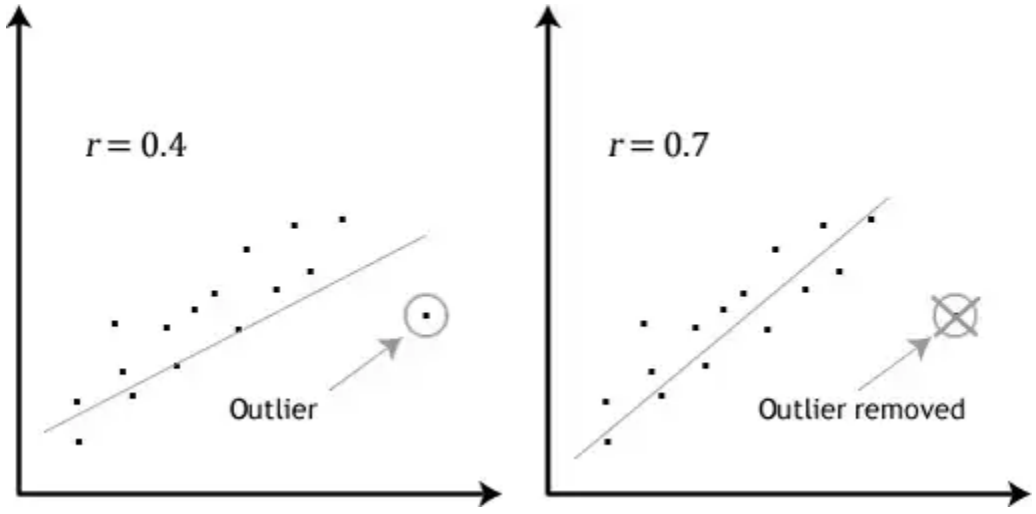
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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'**. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.

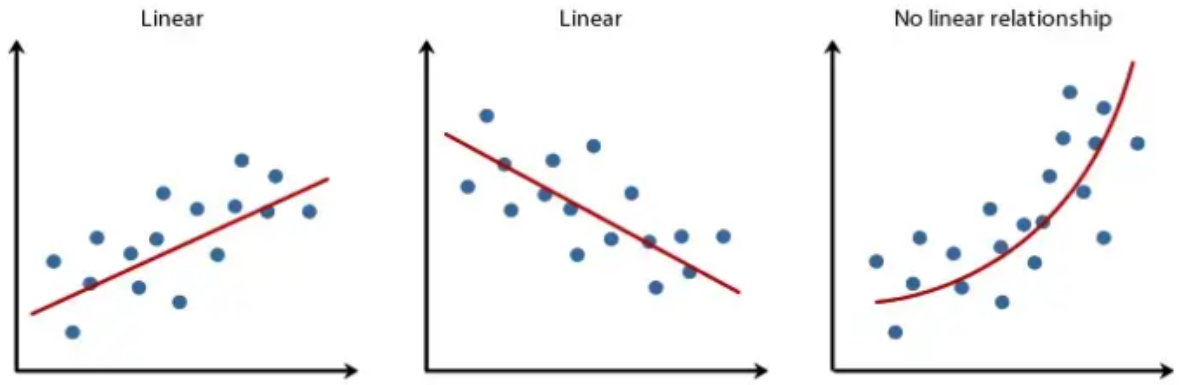


2. There should be no significant outliers. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means - including outliers in your analysis can lead to misleading results.



3. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

4. The two variables have a linear relationship. Scatter plots will help you 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. If the data you have is not linearly related you might have to run a non-parametric .

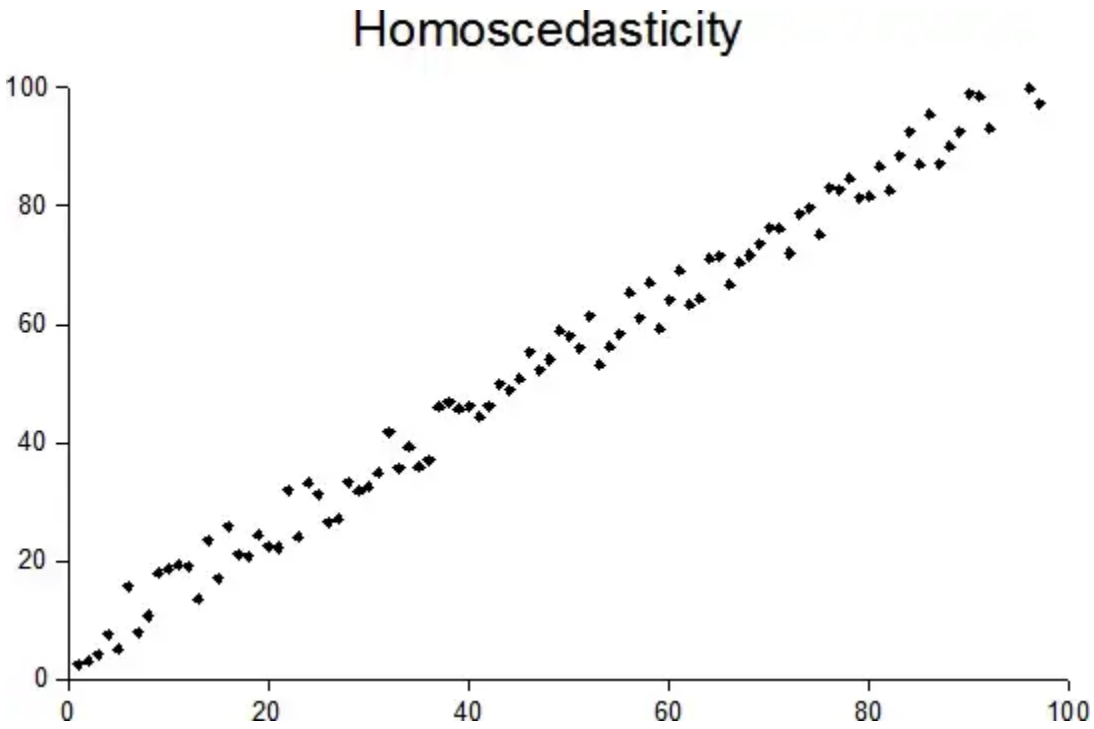


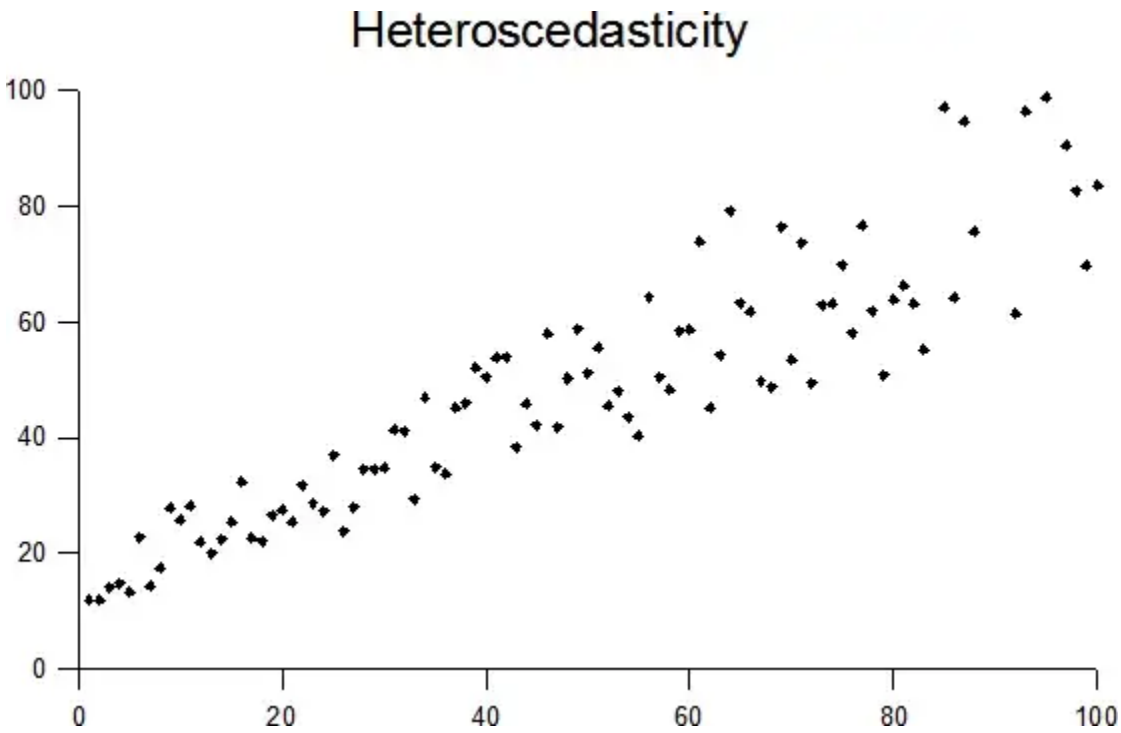
Linear and Non-Linear Relationships

5. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example: if you're 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. **Homoscedascity**: I've saved the best for last. It is hard to pronounce but the concept is simple. Homoscedasticity 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 scatter-plot 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

‘Heteroscedascity’ (the violation of homoscedasticity) which is present when the size of the error term differs across values of an independent variable.





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

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.

Why is scaling performed:

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between Normalized Scaling and Standardized Scaling:

- **'Normalization/Min-Max Scaling'** 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' 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).

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

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.

- Normalization typically rescales the values into a range of [0,1].

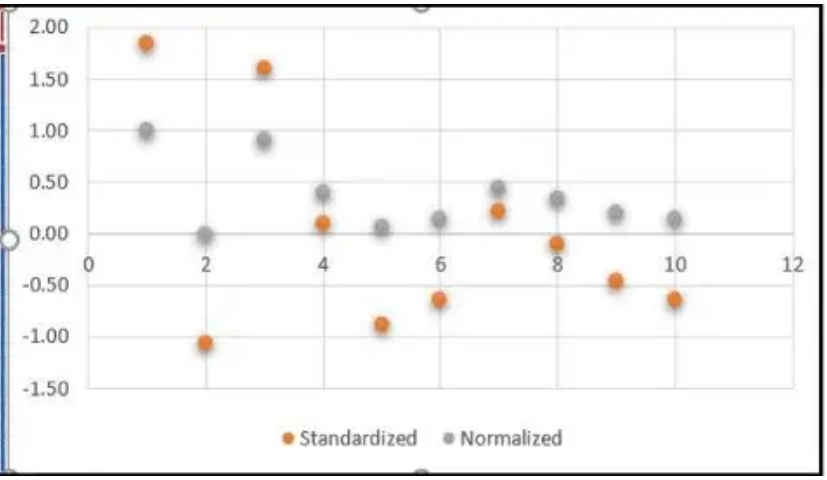
  Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.no	Normalization/Min-Max	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bound 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.

Example: Below shows an example of Standardized and Normalized scaling on original values.

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:

The '**Variance Inflation Factor' (VIF)** measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

VIF is another commonly used tool to detect whether multicollinearity exists in a regression model. It measures how much the variance (or standard error) of the estimated regression coefficient is inflated due to collinearity. In 'Ordinary Least Squares (OLS)' regression analysis, multicollinearity exists when two or more of the independent variables demonstrate a linear relationship between them. For example, to analyze the relationship of company sizes and revenues to stock prices in a regression model, market capitalizations and revenues are the independent variables. With multicollinearity, the regression coefficients are still consistent but are no longer reliable since the standard errors are inflated. It means that the model's predictive power is not reduced, but the coefficients may not be statistically significant with a Type II error.

Therefore, if the coefficients of variables are not individually significant - cannot be rejected in the t-test, respectively - but can jointly explain the variance of the dependent variable with rejection in the F-test and a high coefficient of

determination (R2), multicollinearity might exist. It is one of the methods to detect multicollinearity.

VIF is infinite:

VIF = infinity shows a perfect correlation between two independent variables. As we are aware:

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{Tolerance}$$

Where, 'i' refers to the ith variable.

If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite. It denotes perfect correlation in variables.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

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

Q-Q Plot:

"Q-Q plots are also known as **Quantile-Quantile plots**. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential."

“**Quantile-Quantile (Q-Q) plot**, is a graphical tool to help us assess if a set of data plausibly came from some theoretical 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.”

“Quantile-Quantile plot or **Q-Q plot** is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against”. For example, if you are testing, if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members’ age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding of the underlying distribution of a variable(s).It 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 the same distributions.

Advantages: Below are few advantages of **Q-Q Plot:**

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios, if two data sets-

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

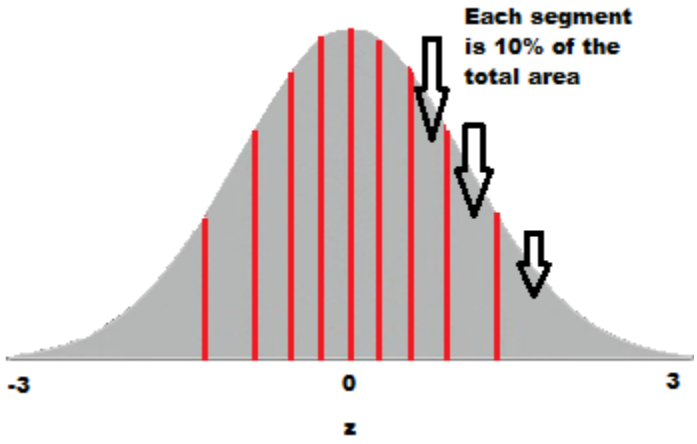
How to Make a Q Q Plot:

Example: Do the following values come from a normal distribution?
7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

Step 1: Order the items from smallest to largest.

3.77
4.25
4.50
5.19
5.89
5.79
6.31
6.79
7.19

Step 2: Draw a normal distribution curve. Divide the curve into n+1 segments. We have 9 values, so divide the curve into 10 equally-sized areas. For this example, each segment is 10% of the area (because 100% / 10 = 10%).



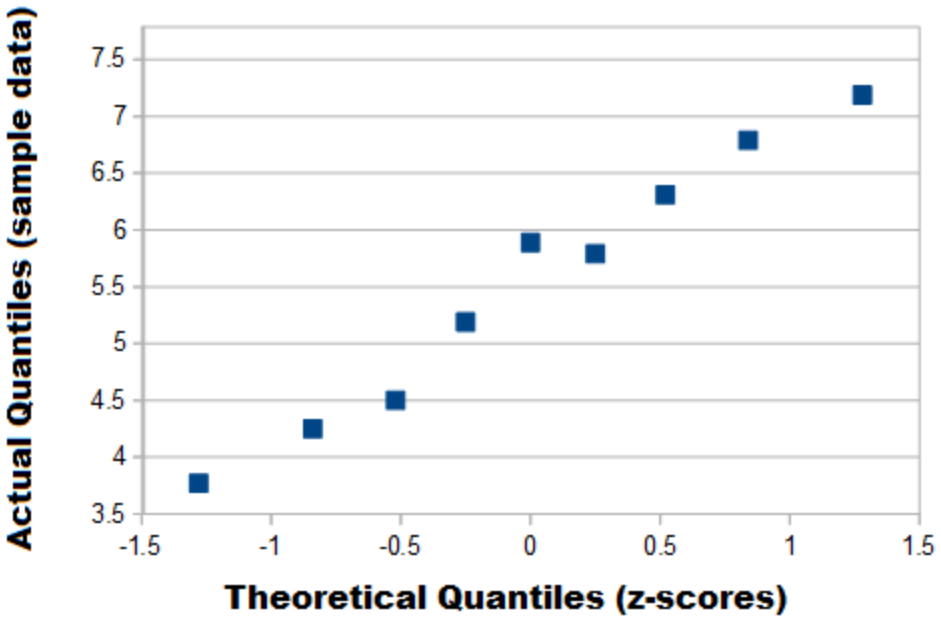
Step 3: Find the z-value (cut-off point) for each segment in Step 3. These segments are areas, so refer to a z-table (or use software) to get a z-value for each segment. The z-values are:

10% = -1.28
20% = -0.84
30% = -0.52

40% = -0.25
50% = 0
60% = 0.25
70% = 0.52
80% = 0.84
90% = 1.28
100% = 3.0

A few of the z-values plotted on the graph.

Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3).



The (almost) straight line on this q q plot indicates the data is approximately normal.

Q Q Plots and the Assumption of Normality:

The assumption of normality is an important assumption for many statistical tests; you assume, you are sampling from a normally distributed population. The normal Q Q plot is one way to assess normality. However, you don't have to use the normal distribution as a comparison for your data; you can use any continuous distribution as a comparison (for example a Weibull distribution or a uniform distribution), as long as you 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 your data well.