**Assignment 4**

**Kishan Sarpangala**

**M04307966**

Question 1

Consider the wine datasets given at this [LINK.](https://archive.ics.uci.edu/ml/datasets/wine+quality) One dataset at this site is for red wines and the other is for white wines. Data has eleven features and the twelfth column is the target attribute: “wine-quality”. The values of wine-quality range between 0 and 10. The goal of this task is to learn a regression model that predicts “wine-quality” for a given set of features. Consider the **white wine dataset** and perform the following tasks with this dataset. Use Matlab or Python SKLearn libraries to perform these tasks.

**1.A**

Do linear regression to learn the single-feature regression models, one model for each of the 11 features. Find the R2 and AIC values for each of these models. Report these values for the models.

**[8 Marks]**

**What is AIC?**

* Akaike Information Criterion Metric is helpful in describing the efficiency/quality of our model for a the data provided. Akaike Information Criterion Metric is also referred as AIC in short.
* Goodness of fit is rewarded by AIC. Hence, say we have a set of candidate attributes or models for a given data set the preferred way to model is the one which will give us the minimum Akaike Information Criterion Metric value. [1]
* We need a loss function in order to optimize our model. Akaike Information Criterion Metric (AIC) provides us a way to calculate the loss function. If this loss function is optimally minimized in the right direction this will gives us an "optimal" model which fits our given data.
* In other words while fitting a model if you increase the number of parameters you will improve the log likelihood but will run into the danger of over fitting. The AIC penalizes for increasing the number of parameters thus minimizing the AIC selects the model where the improvement in log likelihood is not worth the penalty for increasing the number of parameters. [1]

**What is R2**

* **It’s a statistic metric which** will give some information about the goodness of fit of a given model. In regression, the **R2** coefficient of determination is a **statistical** measure of how well the regression predictions approximate the real data points. [2]
* Ideally an **R2** of 1 indicates that the regression predictions perfectly fit a given data.
* In this assignment we used AIC and R2 to compare different models & more dynamic hypothesis testing to explore the relationship between different attributes which would define the quality of the white wine.
* We end up doing univariate analysis for this question.
* Here the model is built using OLS with all the parameters against quality Check summary data, R-squared value, it’s not that great, we need to improve the model by removing unwanted parameters. [2]
* 

4898 data rows are available. 11 features are used to classify.

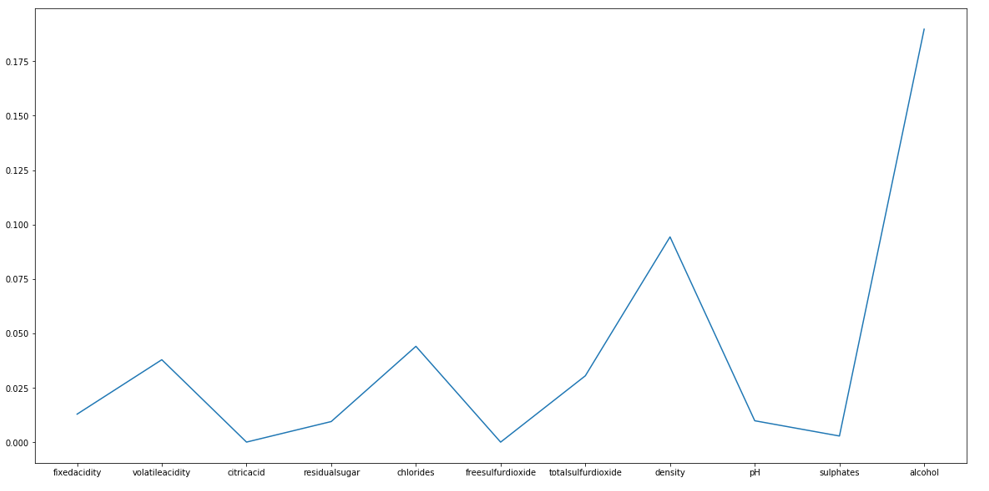
Below table will help us determine the features which are best in predicting the quality of the wine while performing univariate analysis.

|  |  |  |
| --- | --- | --- |
| **Models/Features** | **rsquared** | **AIC** |
| fixedacidity | 0.012919239085696144 | 12649.542674671113 |
| volatileacidity | 0.03791703462546092 | 12523.903238229133 |
| residualsugar | 8.480735490934421e-05 | 12712.81801327696 |
| citricacid | 0.009521237537143956 | 12666.37496494396 |
| chlorides | 0.044072456899563295 | 12492.465076996094 |
| Freesulfurdioxide | 6.655405919031576e-05 | 12712.90742468611, |
| totalsulfurdioxide | 0.030533095213563288 | 12561.351623049086 |
| density | 0.09432472922515445 | 12227.96672220093 |
| pH | 0.00988577719477901 | 12664.571954152203 |
| sulphates | 0.002881314493483056 | 12699.10036825579, |
| alcohol | 0.1897253327492563 | 11682.782413524452 |

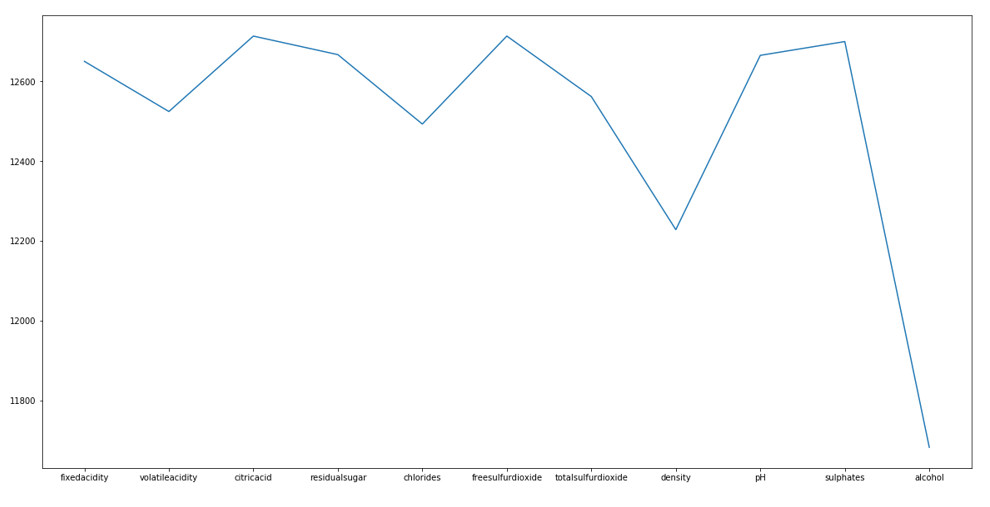
Alcohol was found to have highest **R2** with lowest AIC value (I have marked it in the able above)

**R2**

**Graph for Univariate (below)**



**AIC Graph for Univariate (below)**



* The main features in the data set was found to be alcohol, density, chlorides and volatile acidity while performing univariate analysis.

**1.B**

Select the model with the highest R2 value, combine with its feature other features, one at a time, and thus generate all bivariate regression models (models containing two features). One of these two features is from the selected single-feature model and the other is from one of the remaining 10 features.

**[8 Marks]**

* For this question we end up doing Bivariate Analysis
* Two more significant independent variables correlating to quality are alcohol and volatile acidity. These combined features best determine or help us predict the quality of wine.
* In terms of relationships between independent variables, some strong correlations are observed

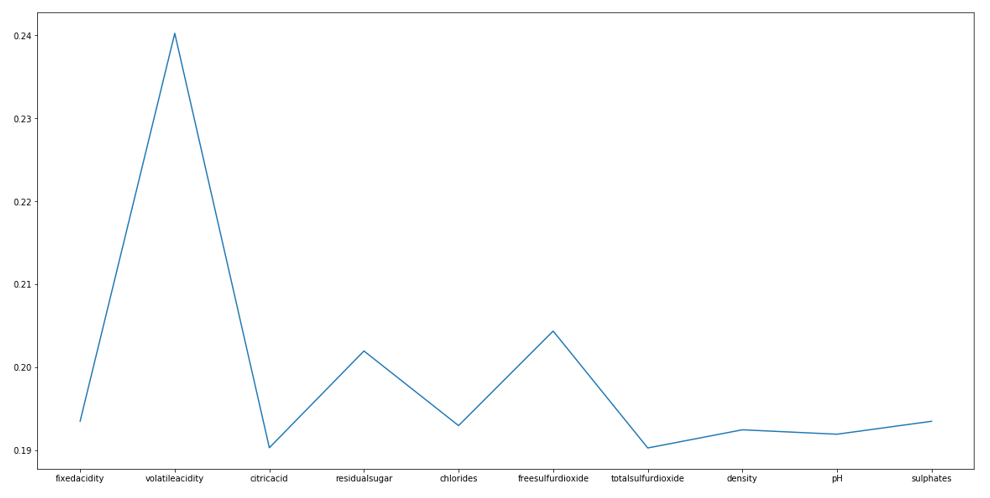
Below table will help us determine the features which are best in predicting the quality of the wine while performing bivariate analysis.

|  |  |  |
| --- | --- | --- |
| **Models/Features** | **rsquared** | **AIC** |
| Aclhocol + fixedacidity | 0.19350275804257744 | 11661.89500108597 |
| Aclhocol + volatileacidity | 0.2402311847533618 | 11369.551595515226 |
| Aclhocol + residualsugar | 0.19029391204459856 | 11681.344222676827 |
| Aclhocol + citricacid | 0.2019510008159675 | 11610.316711315549 |
| Aclhocol + chlorides | 0.19295861042568707 | 11665.198591190954 |
| Aclhocol + Freesulfurdioxide | 0.20435197005222794 | 11595.558628692195 |
| Aclhocol + totalsulfurdioxide | 0.190266569777101 | 11681.509616227759 |
| Aclhocol + density | 0.19245487662184757 | 11668.254839006066 |
| Aclhocol + pH | 0.191923201096876 | 11671.478547200913 |
| Aclhocol + sulphates | 0.1934806280289244 | 11662.029398718263 |

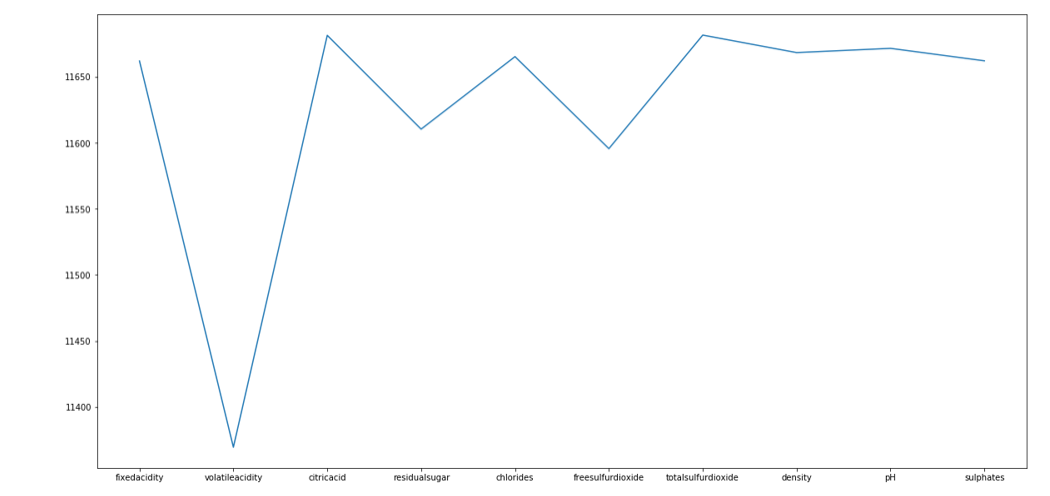
Aclhocol + volatileacidity was found to have highest **R2** with lowest AIC value (I have marked it in the able above)

**R2**

**Graph for Bivariate (below)**



**AIC Graph for Bivariate (below)**



**1.C**

Select the bivariate model with the highest R2 value as the Best model at this stage. Combine a third feature from the remaining nine features with this selected bivariate model to build (and then select the best) 3-feature regression models. Report the R2 and AIC values of all these models.

**[8 Marks]**

Below table will help us determine the features which are best in predicting the quality of the wine while performing multi-variate analysis.

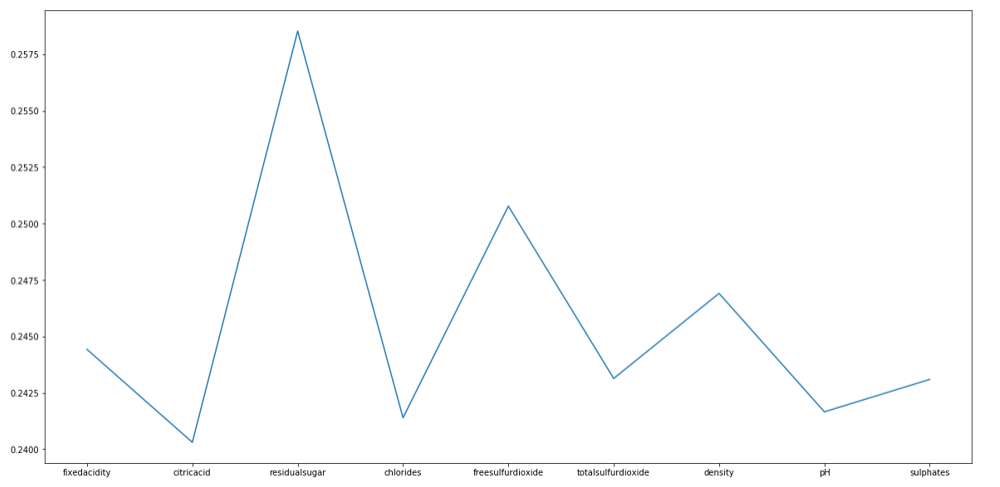
|  |  |  |
| --- | --- | --- |
| **Models/Features** | **rsquared** | **AIC** |
| Aclhocol + volatileacidity + fixedacidity | 0.24442512013197892 | 11344.439663086534 |
| Aclhocol + volatileacidity + citricacid | 0.24030961624094338 | 11371.04594531183 |
| Aclhocol + volatileacidity + residualsugar | 0.25852615806597834 | 11252.166211350774 |
| Aclhocol + volatileacidity + chlorides | 0.24139567574890275 | 11364.038714964241 |
| Aclhocol + volatileacidity + Freesulfurdioxide | 0.2507712371048688 | 11303.127342561762 |
| Aclhocol + volatileacidity + totalsulfurdioxide | 0.24313806051114784 | 11352.775904923396 |
| Aclhocol + volatileacidity + density | 0.24690635619896484 | 11328.328628723433 |
| Aclhocol + volatileacidity + pH | 0.24166031225998164 | 11362.329766401646 |
| Aclhocol + volatileacidity + sulphates | 0.24309658826833835 | 11353.044283434197 |

Aclhocol + volatileacidity + residualsugar was found to have highest **R2**  with lowest AIC value ( I have marked it in the able above)

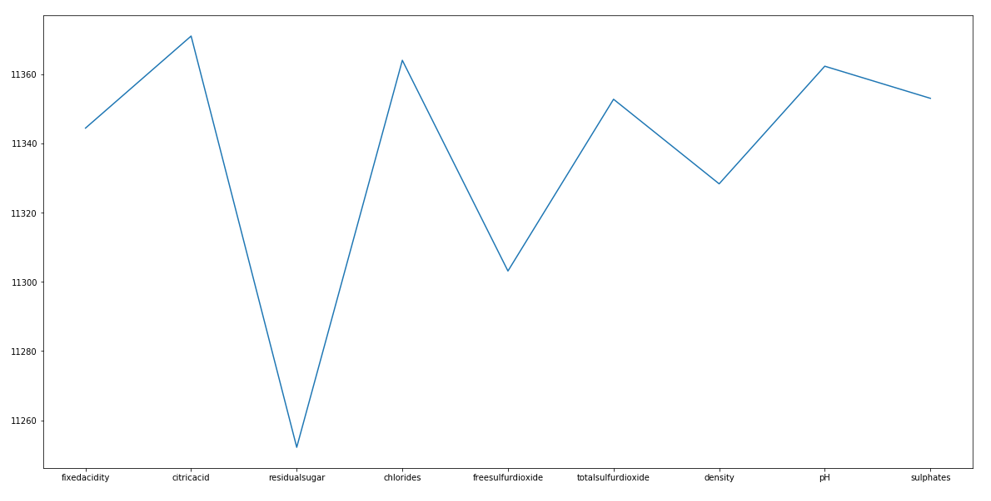
* This is obtained after multivariate analysis

**R2**

**Graph for multivariate (below)**



**AIC Graph for Multivariate (below)**



**1.D**

Repeat the steps above to generate (k+1)-feature models from the k-feature models until the following situation arises: all the (k+1)-feature models have an AIC value higher than the AIC value of the k-feature model from which they are being generated. Stop the process and report the k-feature model found as being the best regression model for this data. Report the features included, their coefficients, and p-values for the coefficients. Comment on the magnitudes of the p-values.

**[Marks 14]**

**Features included:**

['alcohol', 'volatileacidity', 'residualsugar', 'freesulfurdioxide', 'density', 'pH', 'sulphates', 'fixedacidity']

1. **Alcohol**
2. **Volatile Acidity**
3. **Residual Sugar**
4. **Free Sulfur dioxide**
5. **Density**
6. **pH**
7. **Sulphates**
8. **Fixed Acidity**

**Features which did not get included are**

critic acid

2) chlorides

3) Freesulfurdioxide

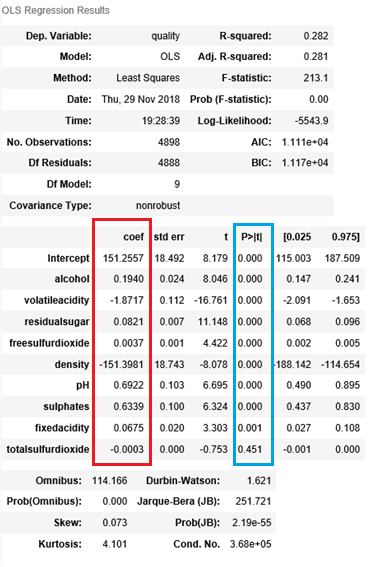
4) totalsulfurdioxide.

**What is P-Value?**

* The **p**-**value** is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event.

The **p**-**value** is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. [3]

* P-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). It’s noted that a low p-value which is below < 0.05 means that you can reject the null hypothesis. What it means is that, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. [8]
* Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response. [8]



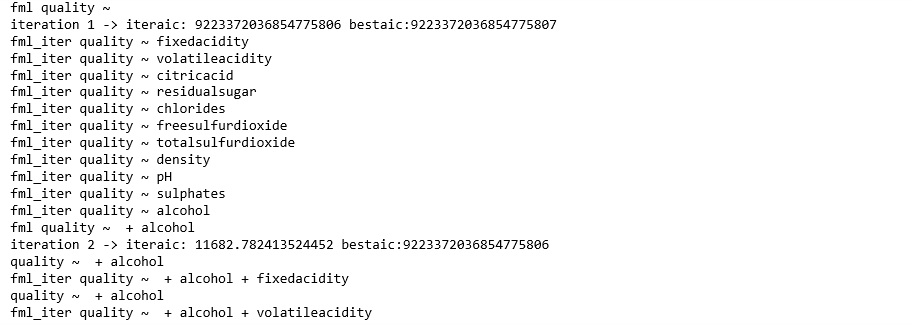
* Coefficients and also the P values are presented above.
* Multiple R-squared will always increase if you add more independent variables.
* But Adjusted R-squared will decrease if you add an independent variable that doesn’t help the model.  
  This is a good way to determine if an additional variable should even be included in the model.
* If the p value is less 0.05 than the feature is significant
* All the P values are tending to zero expect for the features fixed acidity and total sulfur dioxide
* For the following features the p value is zero which are alcohol, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates, fixed acidity.
* alcohol, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates, fixed acidity, fixed acidity and total sulfur dioxide have p value less than 0.05

**Main gist of the algorithm**:

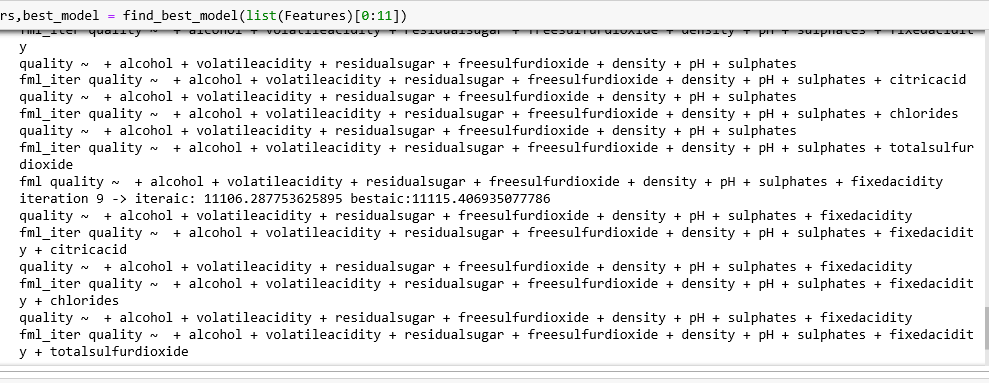
Inner loop takes each feature with highest R2 and keep appending to the model. The outer loop will look at the final model which has newly set of features if the AIC of the overall model is lower than keep on continuing with the algorithm. Stop running the algorithm when the AIC goes higher.

**Already I have answered this question above for more results/information refer to .pynb**

**Final result:**



**More information can be obtained from the .pynb**



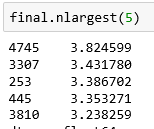
**1.E**

Find the five wines that have the largest magnitudes of difference between the predicted and the actual wine-quality values. Look at the regression model, the rest of the data, and comment on why you think these wines are outliers.

**[Marks 7]**

**Outliers:**

An *outlier* is an observation that lies an abnormal distance from other values in a random sample in our data.



|  |  |  |  |
| --- | --- | --- | --- |
| **Serial Number** | **Quality value from Excel** | **X\_pred value** | **Magnitude** |
| 253 | 3 | 6.3867023003270695 | 3.386702 |
| 445 | 3 | 6.353270875189918 | 3.353271 |
| 3307 | 3 | 6.43177978734515 | 3.431780 |
| 3810 | 3 | 6.2382586252068055 | 3.238259 |
| 4745 | 3 | 6.824598915816989 | 3.824599 |

The above tale shows how the magnitude is calculated for the “five wines that have the largest magnitudes”

* We can observe from the table there is existence of mild outlier and extreme outlier. A data point beyond an inner fence on either side is considered a **mild outlier** while a data point that is beyond an outer fence is considered an **extreme outlier**.
* One or more feature values have high variance for the above-mentioned serial numbers – 253, 445, 3307, 3810, 4745
* The main reason is high variance in independent features.
* Below I have presented how I found X\_pred values
* For 253 serial number we should be looking at the row 255 in the excel file. Similarly for 445 we need to be looking at the 447 row of the excel file.
* Free Sulphur dioxide is the feature which creates this high variance which is depicted in the table below

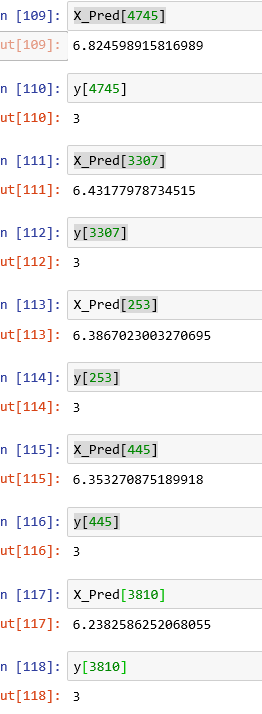
|  |  |
| --- | --- |
| **Serial Number** | **Free Sulphur dioxide of white wine** |
| 253 | 5 |
| 445 | 16 |
| 3307 | 124 |
| 3810 | 42 |
| 4745 | 289 |

The mean for the free Sulphur dioxide of the white wine was found to be

**Mean (μ): 95.2**

**Variance (σ2): 11125.36**

**This is highest amongst all the features selected for the model.**



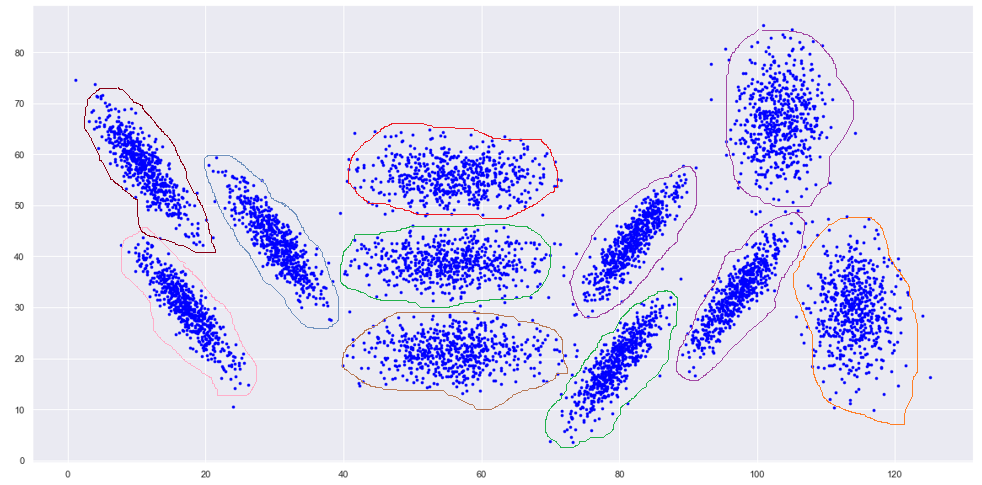
**Question 2**

Consider the data-file attached with this homework containing 6600 data points on a 2-D plane. You will need to use the BIC metric to determine the quality of a clustering. This is computed here as: BIC = n\*log(SSE/n) + log(n)\*c\*(d+1) where n is the number of data points, c is the number of clusters, and d is the number of features (dimensionality of the data). Remember to use the sum of SSEs for all the clusters in any clustering.

**2.A**

Plot the data on a 2-D scatter plot and mark by hand the boundaries of the ideal clusters that you would like discovered in this dataset.

**[3 Marks]**



Ideally, I would like to see the above 11 clusters to be found by the algorithm.

**Note:**

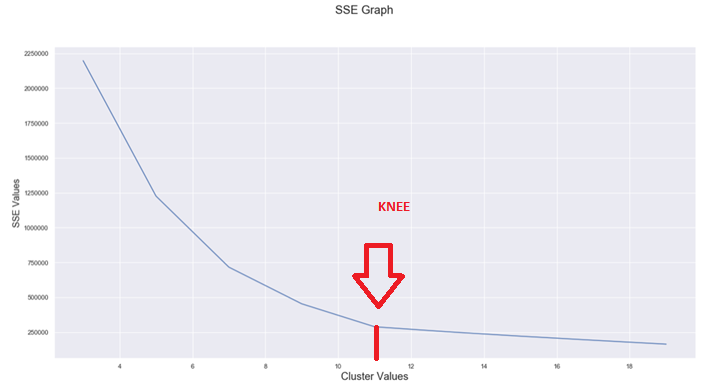
(1) In above graph all the data points need to belong to either of the 11 clusters based on clustering algorithm. I have used paint hence its difficult to put each point to a cluster. Above diagram gives us a rough intuitive idea where each point should belong.

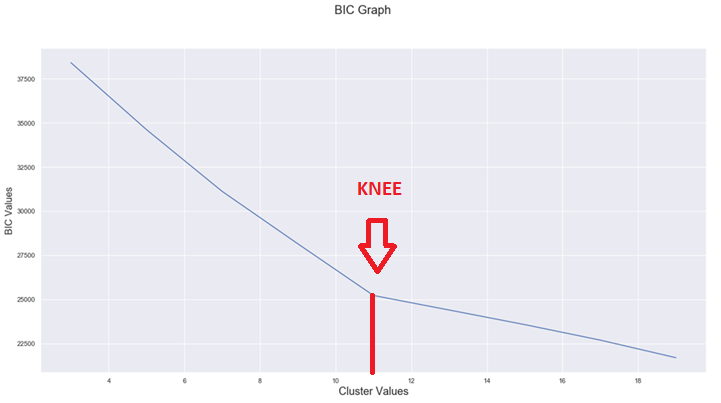
(2) Ideally it should be spherical

**2.B**

Run the k-means algorithm for k = 3, 5, 7, 9, 11, 13, 15, 17 and 19. Plot the total SSE and BIC values for the above values of k. What is the best number of clusters for this dataset? How did you find the best number of clusters, briefly explain.

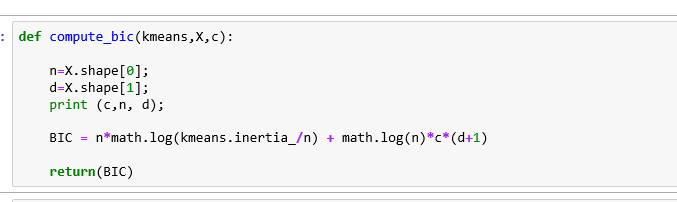
**[12 Marks]**



****

**Note**:

* The best cluster as provided by SSE and also BIC is 11
* Knee helps us find the best cluster value from SSE and BIC curves
* One method to validate the number of clusters is the Elbow or Knee method. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 0 to 10 as show an in the figure above), and for each value of k calculate the sum of squared errors (SSE) and BIC. BIC algorithm/function is shown below



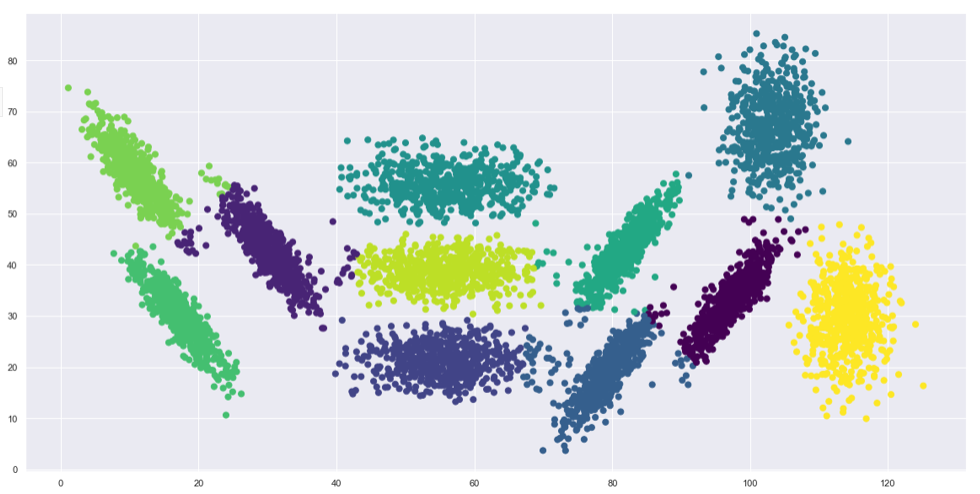
* The main objective of the k-means algorithm is to minimize the SSE(squared error) is a reasonable choice if you replace every object by its nearest centroid.

**2.C**

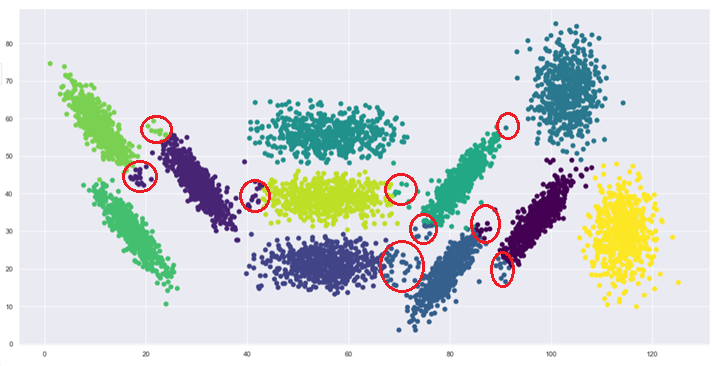
For the best number of clusters selected above, plot the scatter plot of the data showing the points of each cluster with a different color/symbol. Mark the points on the scatter plot that belong to clusters other than what your intuition says. Why did k-means algorithm place them in these different clusters – explain very briefly.

**[8 Marks]**

* 11 was the best cluster selected from above question
* In image below, I have used 11 different colors to depict the different clusters.



Below are the misclassified points; which I have marked using red color. (I marked the points on the scatter plot that belong to clusters other than my intuition )





The reason for the misclassification of certain points from my intuition is because of the following disadvantage which are inherent to k-means

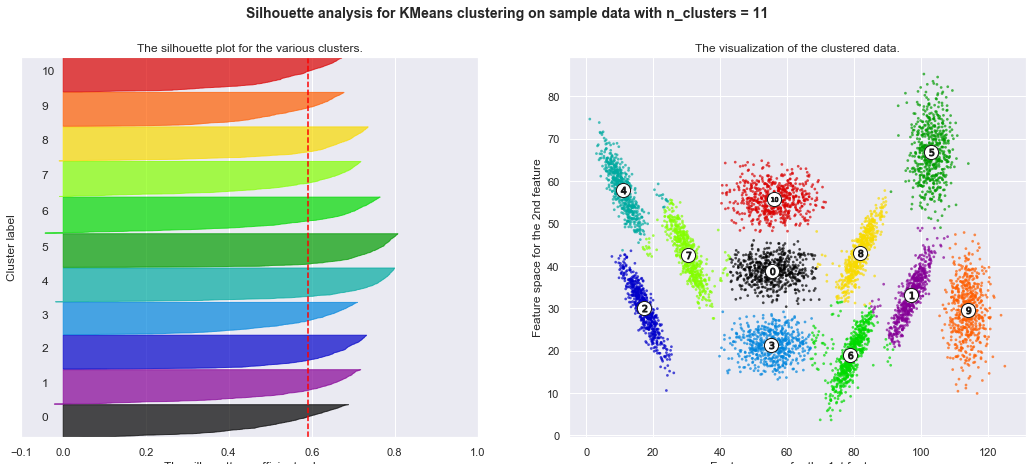
* The selection of the initial seeds has an effect on the final results.
* The ﬁnal result is also affected by the order of the data.
* There is high susceptibility for a poor initial cluster centroid selection which will eventually lead to bad convergence at a local minimum level; hence we are observing this misclassification depicted above. Also, the objective of k-means is to minimize the squared error; it’s a good way to calculate if you replace every object by its nearest centroid but this might lead to misclassification each time you run k-means because of the randomness of centroid selection.
* Also, k-means is very sensitive to outliers in a given data set. K-means algorithm always tries to optimize SSE or sum of squares because of which there is high probability that a large deviation; such as an existence of an outliers will get lot of importance/weightage while trying to reduce the overall SSE.

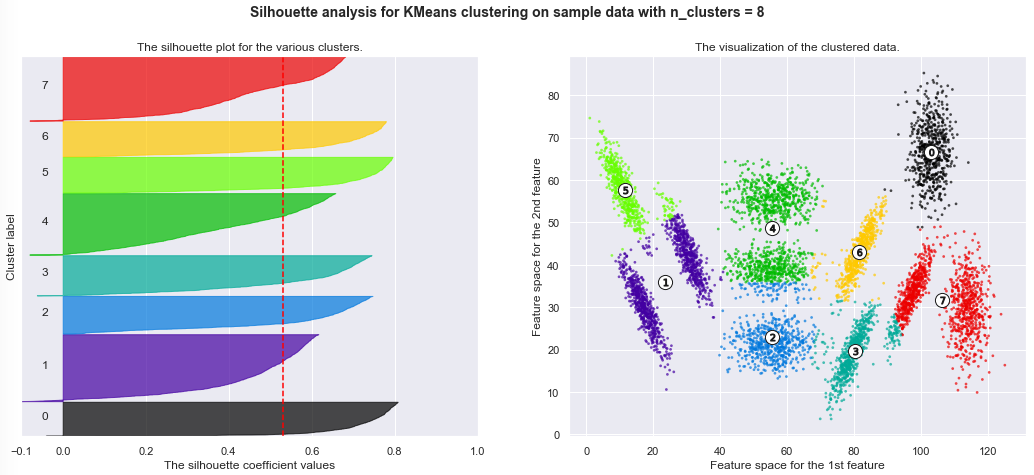
**2.D**

Plot the silhouette diagram for the best clustering you have selected. Comment on the characteristics of the silhouette diagram that you think are informative about this clustering. Comment using the cluster numbers and their plots on the silhouette diagram.

**[5 Marks]**

* If you observe the images below for clusters =11 and clusters = 8 you would notice that how each close each point in 1 cluster is to the data points in the neighboring clusters.
* Silhouette provides a way to assess the features/parameters with the use of number of clusters visually present
* Also from the thickness of the silhouette plot depicts the cluster size which can be visualized below.
* So, the Silhouette coefficients are above average value of 0.6 which is depicted by the red color.
* I plotted cluster 11 and cluster 8 and we can see that Silhouette diagram for 11 is better than 8 because the thickness is equal among all the features and also the silhouette coefficients are above average value of 0.6





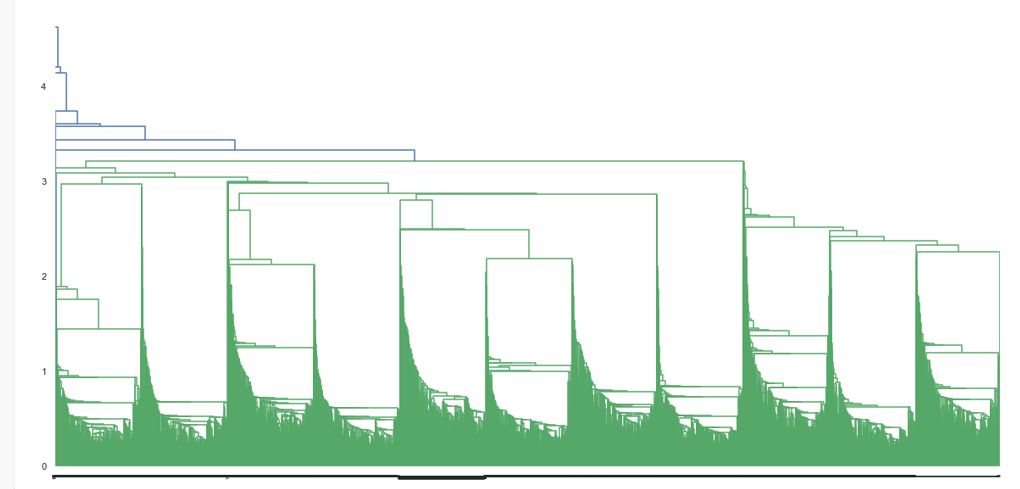
**2.E**

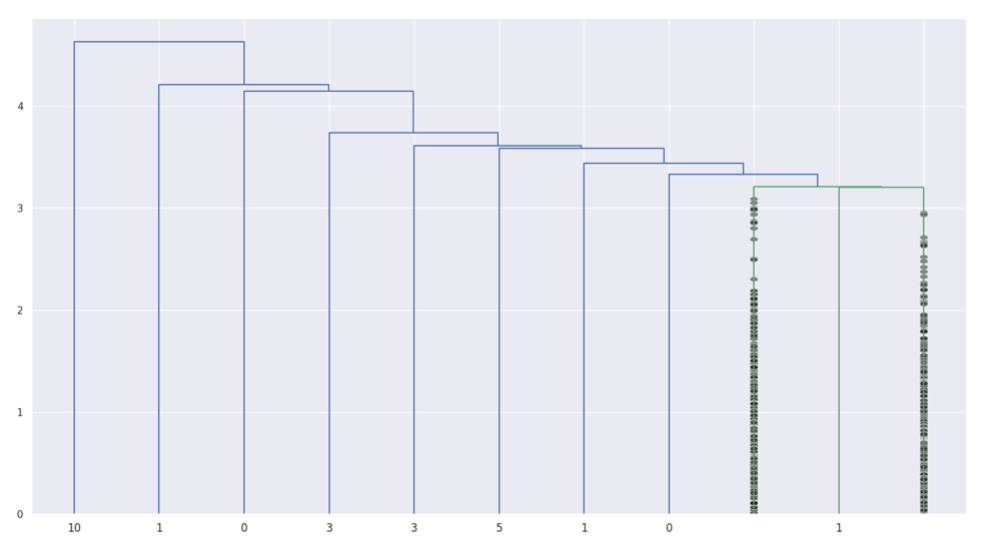
Perform single-linkage hierarchical clustering for this data and cut the dendrogram to obtain 11 clusters. There are options/parameters in most toolboxes to generate a given number of clusters. Plot the 2-D scatter plot of the dataset showing data points of each of the 11 clusters with different color/symbol.

**[12 Marks]**

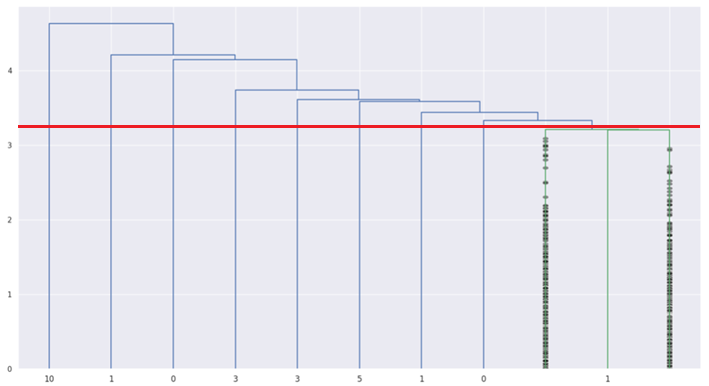
* **Single-linkage clustering** is one of several methods of agglomerative clustering or hierarchical clustering.
* in agglomerative clustering grouping clusters in bottom-up fashion is the standard approach.
* In agglomerative clustering with each step there is combining of 2 clusters occurs. The combining is done in such a way that the closest pair of data elements not yet belonging to the same cluster as each other are grouped together.

single-linkage hierarchical clustering and dendrogram to obtain 11 clusters is depicted below:

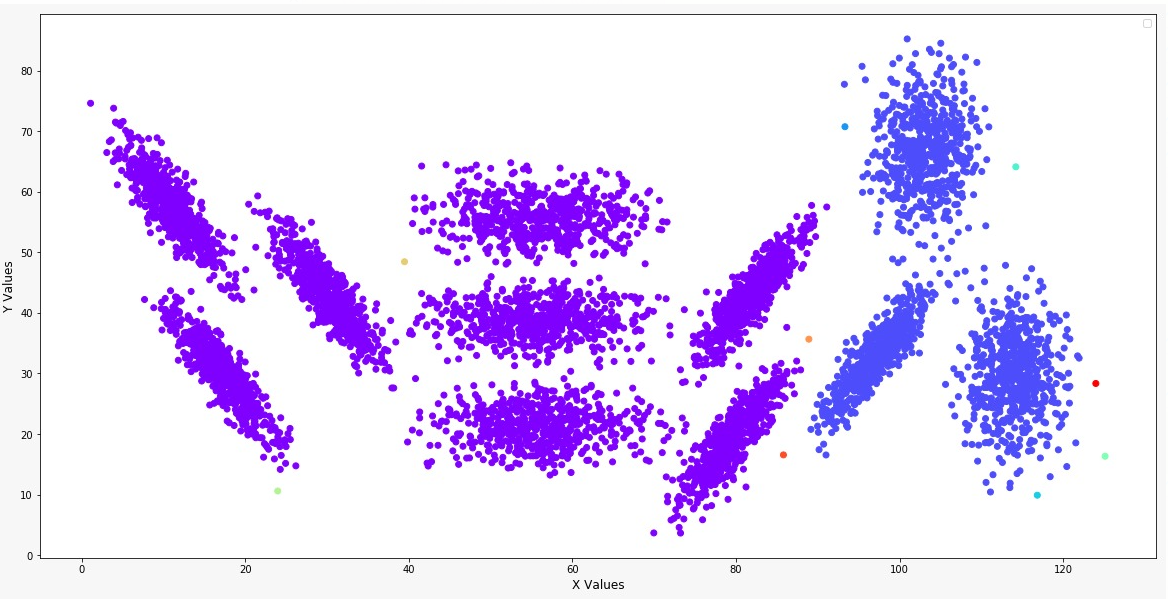




single-linkage hierarchical clustering for our data is depicted below & the dendrogram is cut to obtain 11 clusters:

****

2-D scatter plot of the dataset showing data points of each of the 11 clusters



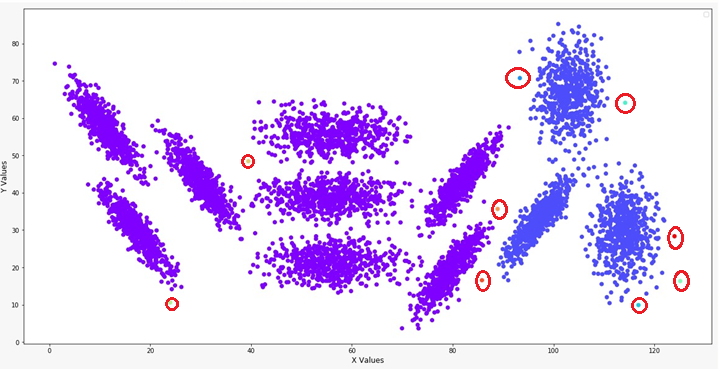
There are 11 clusters formed in the above figure but it’s not well balanced. A depiction of agglomerative clustering

**2.F**

Mark any data points on this scatter plot that are clustered differently from your intuitive view of the correct clusters. Explain why Single-linkage clustering may have placed them in counter-intuitive clusters

**[5 Marks]**

Marked the data points which are clustered differently than my intuitive view:

****

* Incrementally the single linkage agglomerative/[hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering) develops/builds; hence producing a dendrogram this is the reason why in the image below the clusters aren’t well balanced.
* Highly densed data points form a cluster and we end up cutting at 11 because of which we get different set of clusters which can be represented as outliers
* In agglomerative [hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering); it is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around hence we cannot do global optima hence single-linkage clustering has placed data points differently from the intuitive view.
* Agglomerative hierarchical clustering results in a potentially a suboptimal solution the reason is it aims at finding the best step at each cluster fusion (greedy algorithm) achieves a local optimum while k-means tries to optimize a global goal (variance of the clusters)
* Our data set is not of hierarchical structure if it was than hierarchical clustering would have worked; than hierarchical clustering can recover the hierarchy. Hence single-linkage clustering has placed data points differently from the intuitive view.
* K-Means may produce higher clusters than hierarchical clustering hence in Question 2 D we see 11 clusters which are well balanced while in above pic we can notice that though we have 11 clusters it’s not well balanced. There are more high dense regions.
* The outliers data points effect single linkage agglomerative/[hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering) as it is sensitive to it. Outliers effects its performance.

**References:**

1. [**https://en.wikipedia.org/wiki**](https://en.wikipedia.org/wiki)**/**
2. [**https://stackoverflow.com/questions/13989419/removing-outliers-from-a-k-mean-cluster**](https://stackoverflow.com/questions/13989419/removing-outliers-from-a-k-mean-cluster)
3. [**https://scikitlearn.org/stable/auto\_examples/cluster/plot\_kmeans\_silhouette\_analysis.html**](https://scikitlearn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)
4. [**http://www.alcula.com/calculators/statistics/variance/**](http://www.alcula.com/calculators/statistics/variance/)
5. [**https://www.packtpub.com/mapt/book/big\_data\_and\_business\_intelligence/9781784397180/6/ch06lvl1sec112/the-drawbacks-of-k-means**](https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781784397180/6/ch06lvl1sec112/the-drawbacks-of-k-means)
6. [**http://stp.lingfil.uu.se/~santinim/ml/2016/Lect\_10/10c\_UnsupervisedMethods.pdf**](http://stp.lingfil.uu.se/~santinim/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf)
7. [**https://stats.stackexchange.com/questions/246061/what-are-the-advantages-of-the-pre-defined-initial-centroids-in-clustering**](https://stats.stackexchange.com/questions/246061/what-are-the-advantages-of-the-pre-defined-initial-centroids-in-clustering)
8. [**http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients**](http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients)