****

**DEPARTMENT OF COMPUTER SCIENCE**

**MSc Big Data Technologies**

**MODULE CODE : 7BUIS008W**

**MODULE TITLE :** **Data Mining & Machine Learning**

**MODULE LEADER : DR. Panagiotis Chountas**

**COURSEWORK - 1**

**Handed Out: 20th May 2021 Submission Due Date: 24th June 2021**

**Student ID: w1813148**

**STUDENT NAME: PATEL BHAUMIKKUMAR SHAILESH**

|  |  |  |
| --- | --- | --- |
|  | Content | Page Number |
| 01 | **Data Description-Global Happiness** | **03** |
| 02 | **Data Set Information-Retail Data** | **17** |
| 03 | **References** | **24** |

|  |
| --- |
| 1. Data Description-Global Happiness |

At the beginning, we need to importing all the necessary libraries that is the important part of the project.



Also import the Happiness-Data using .csv format. I used my internal drive link in the path of the dataset. Used describe method that statistical values for all numeric variables as count, max, mean and quantiles min to max.



Using this code, we can easily major the none value of the dataset. In this dataset data has no Null value.

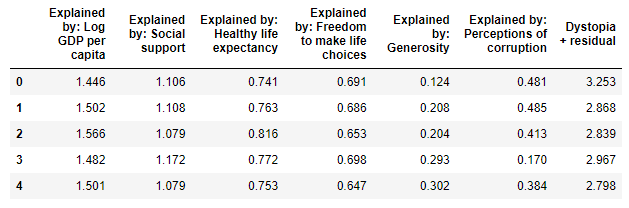




Using the df.head(5) code, we can see the top five happiest countries and the top Happiness Score is 7.842 for Finland.



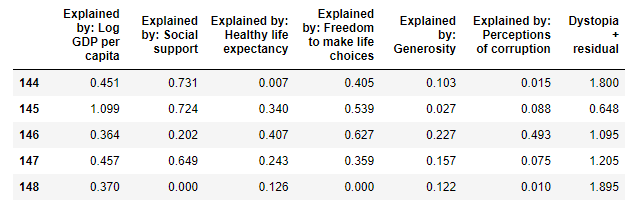
The output of the code above looks like this.



Same as df.tail(5) code used for bottom five and Afghanistan is the least Happiness Score is 2.523 from 148 countries.



The output of the code above looks like this.



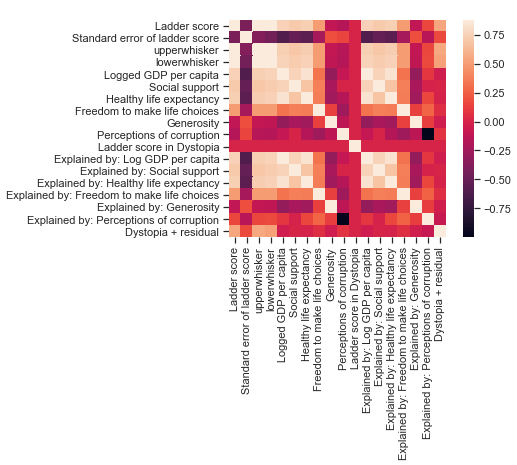
This method we use for drop some column that we need to erase form the dataset.



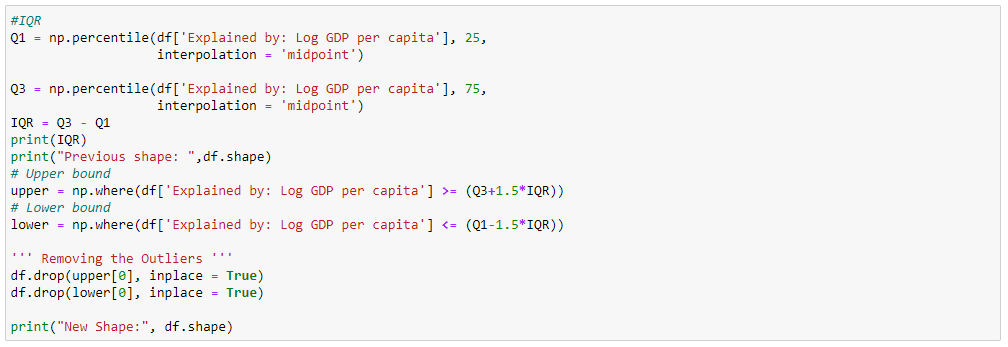
Most of the values are in float type, only Country name and regional indicator is object. There are none of the integer.



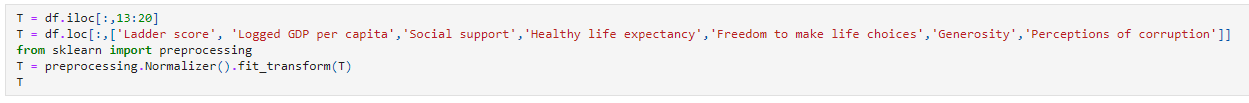


We can visualize the heatmap using the below code and the heatmap looks stunning visually. The graphs demonstrate the correlation between each variable. The standard error of ladder score has negative correlation while Ladder score (Happiness Score) is increasing. We will analyze the Happiness score with the factors of logged GDP per capita, Freedom to make life choices, Perceptions of corruption as these values are extremely high correlated to each other. The main factor that’s effect Social support, Healthy life expectancy and Freedom to make life choices is Logged GDP per capita.

The below interquartile range code is measure of the statistical dispersion, being equal to the diverse between 75th and 25th percentile or between upper and lower quartiles. This method we used various time for the diverse variables like “Log GDP per Capita”,”Social Support”,”Healthy Life expectancy”,Freedom to make life choices”,”Generosity”,”Perceptions of corruption”. When we have IQR code using IQR=Q3-Q1 after that #upper bound code and #lower bound code mention below will give an output with some true and false values. Now the True values the presence of an outlier while False means values are valid.



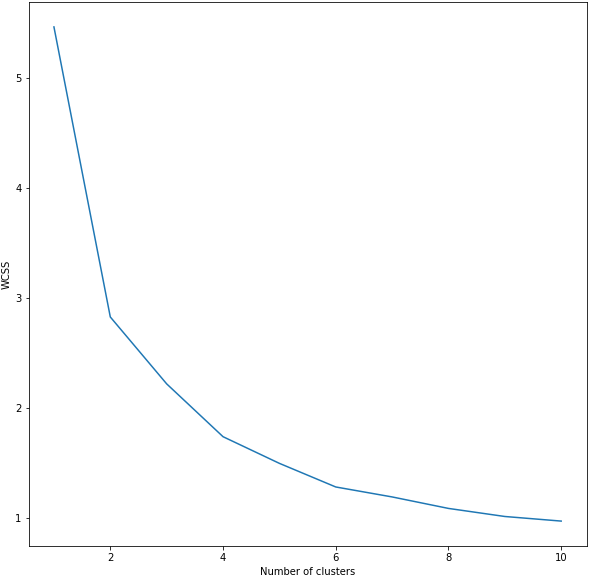
Now, we use normalizer for more radical transformation. The main point of the normalization is changing our observation so that they can described the dataset as a normal distribution. If we need to use machine learning or statistics techniques so that we really want to normalize data.



After the data normalizer we see that the perfect useful data set using below code. Principal component analysis (PCA) is a one of the techniques that is useful for the classification as well as compressed the data. The main purpose of the PCA is redundance of the dimensionality of a dataset via finding a new set of variables. Dimensionality reduction after and before k-means.



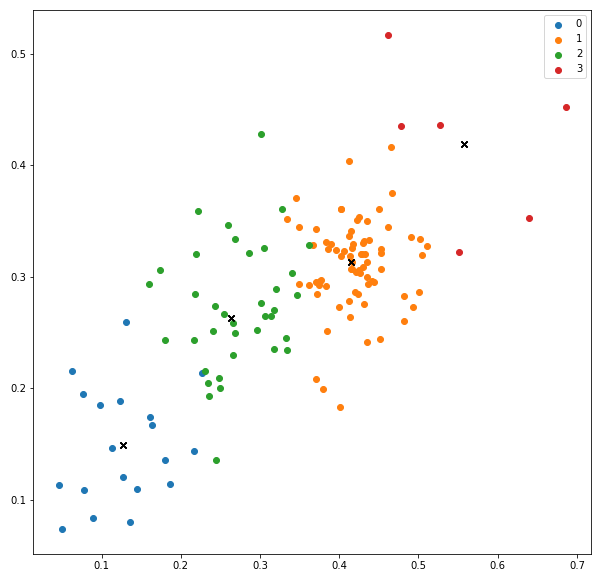
The output of the code above looks like this.



Above coding illustrate the first model of the original data as well as I know the dataset that it contains groups but still, we will implement the “elbow method” to determine the number of clusters. This can be done by using WCSS and the below k-means using as four clusters as well as initialization purpose.



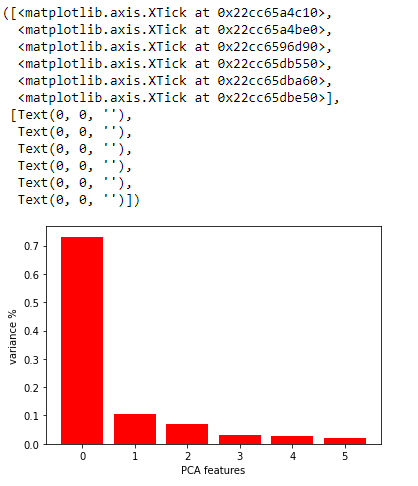
The output of the code above looks like this.

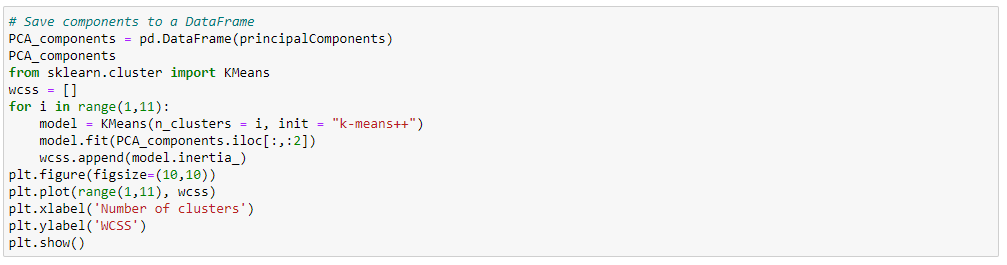


We can perform k-means after the performing dimensionality reduction. I specify any number of dimensions to reduce to while initializing PCA functions but for simplicity’s sake I have used two. If I used two components with the highest value and visualize the dataset. Here we used fit\_transform () function to both fit the dataset.

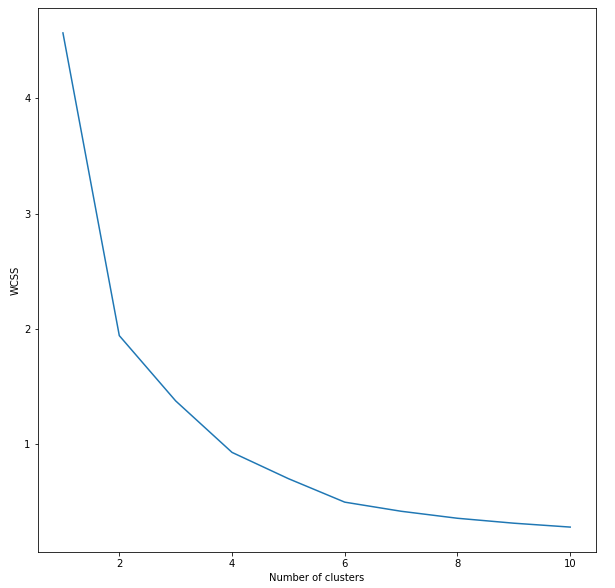


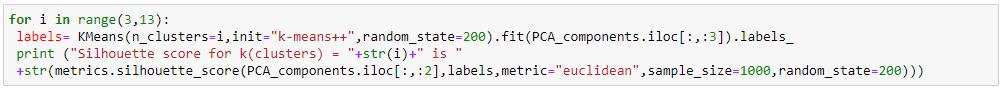
The output of the code above looks like this.



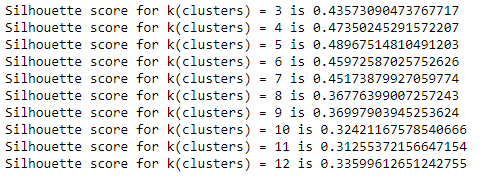


The output of the code above looks like this.





We got the silhouette Score for k clusters from 3 to 12.



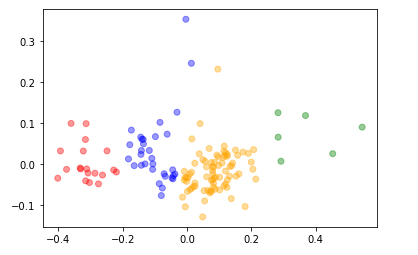
Now we will train my model based on the new features generated by PCA (). Since we have only two principal components, we will get 2D figure with 4 clusters. After three clusters at the elbow the change in the value of inertia is no longer significant and most likely neither is the variance of the rest of the data after the elbow point. There we can discard everything after k=4 and proceed to the last step in the process.



Plot the results

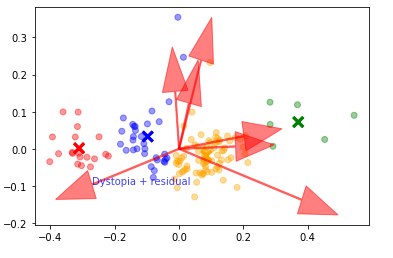




The output of the code above looks like this.



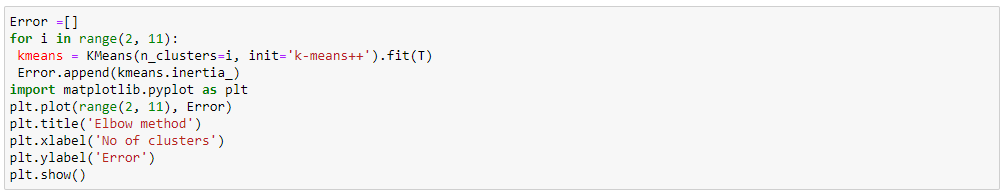
The output of the code above looks like this.



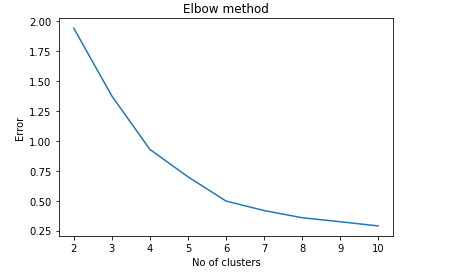
The arrows are the projection of each feature on the principal axis component and represent the level of importance of each feature in the multidimensional scaling. We also conclude that some variables should be places place separately where as others should place close to each other. If we have a large number of features then it is better to perform dimensionality reduction first and then use the clustering Algorithm.

We will first fit multiple k-means models and in each successive model, we will increase the number of clusters. We will store the inertia value of each model and then plot it to visualize the result.

Fitting multiple k-means algorithms and storing the values in an empty list using



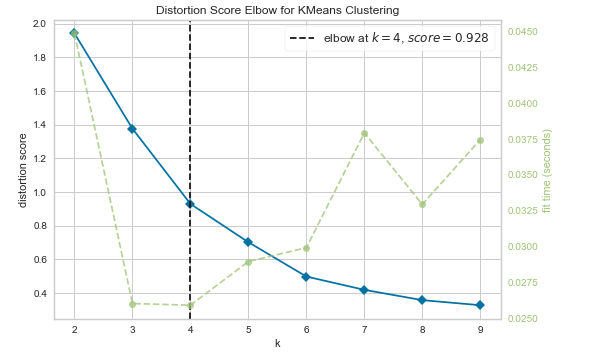
The output of the code above looks like this.



Elbow method with yellowbrick visualizer



The output of the code above looks like this.



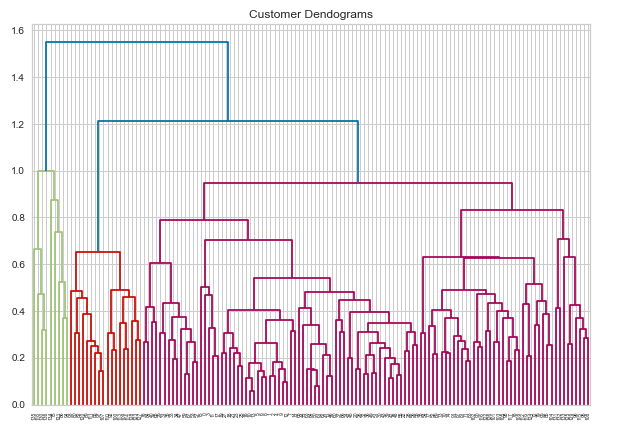
In this code, we will perform hierarchical clustering on real-world data and see how it can be used to solve an actual problem. To cluster this data into groups we will follow the same steps that we already performed in previous steps

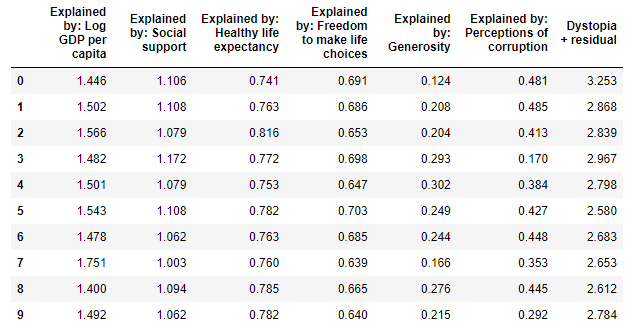
Next, we need to know the clusters that we want our data to be split to. We will again use the spicy library to create the dendrograms for our dataset. Execute the following script to do so



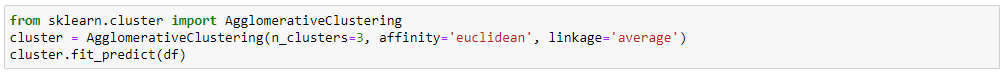
In the script above we import the hierarchy class of the spicy.cluster library as shc. The hierarchy class has a dendrogram method which takes the values returned by the linkage method of the same class. The linkage method takes the dataset and the method to minimize distance as parameters. We use ‘average’ since it minimizes then variants of distance between the clusters.

The output of the script above looks like this

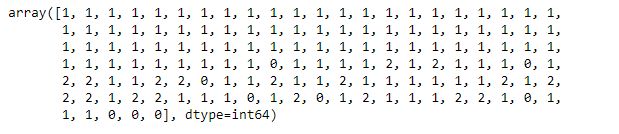




Now we know the number of clusters for our dataset, the next step is to group the data points into three clusters. To do so we will use the AgglomerativeClustering class of the sklearn.cluster library.



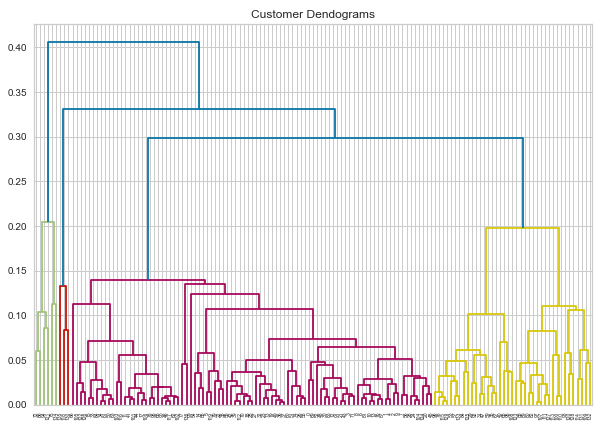
The output of the script above looks like this



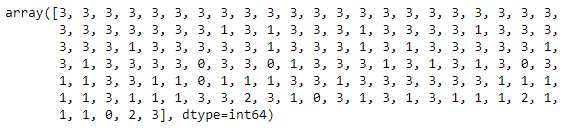
Now, we need to go back work for the above script again as per the assessment question. So, we got these results



The output of the code above looks like this.





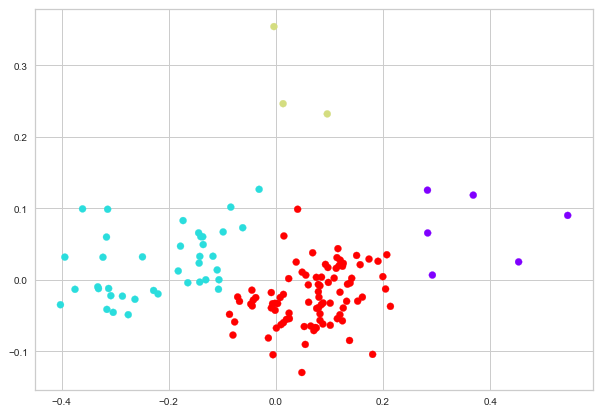


We can see the clusters from all of the data points. Since we have four clusters we have four lables in the output.

As a final step, let’s plot the clusters to see how actually my data has been clusted.



The output of the code above looks like this.



Above analysis demonstrates that there are none of the factor which can elaborate the happiness of people. Some factors like GDP, Social support, Perceptions of corruption and Healthy life expectancy have major role on happiness. This suggests us when we analyze the happiness, we should consider all factors together.

I already know that we unable to buy happiness by ourself but we can provide other factor that to be happier like Healthy life expectancy, Perceptions of corruption, Freedom to make life choices, income inequality and Generosity. Most of the factor’s catalyzer by the GDP.

As a result, creating strong foundation one of the result evidence that belongs to happiness via demonstrated by many countries. This is the time to build social trust & healthy lives so let’s hold our leaders to this fact.

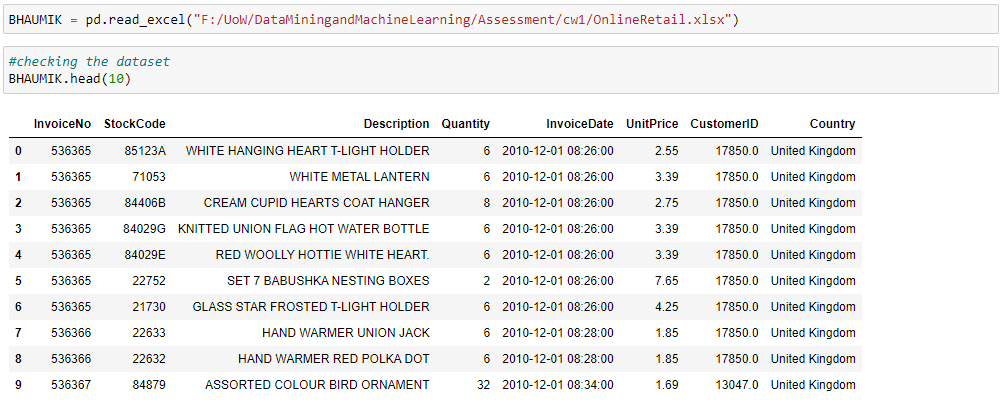
Developed countries will keep their status and they will be happier than the rest of the world. Unhappy countries score will not increase with these circumstances, even their score will decrease every year. Happiness has several factors; GDP is the powerful factor but not the only one. I wish we could have a better world in future but i am afraid this will not happen.

|  |
| --- |
| 1. Data Set Information: Retail Data |

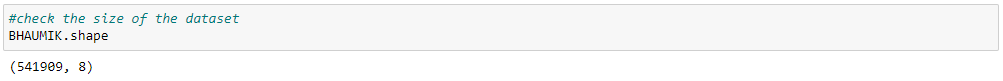
First of all, import libraries that required to do this analysis.

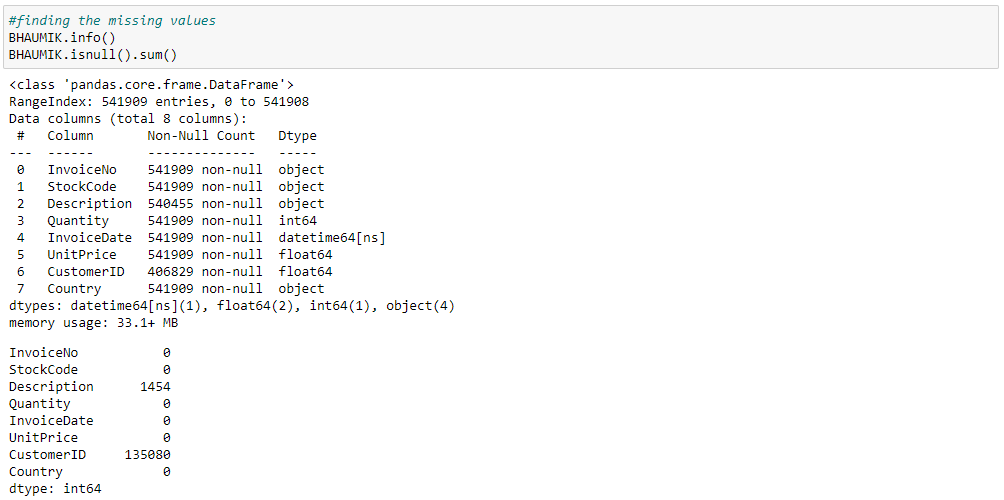


Add dataset from the internal drive and after that checking dataset that successfully imported



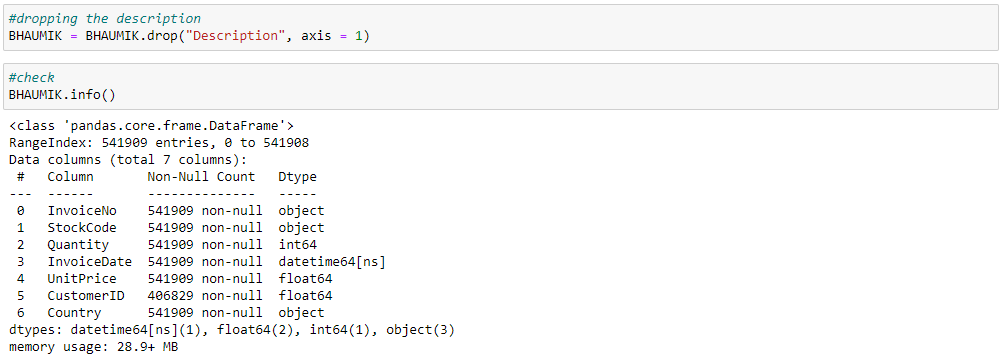
Using the shape command script, we can check the size of the dataset



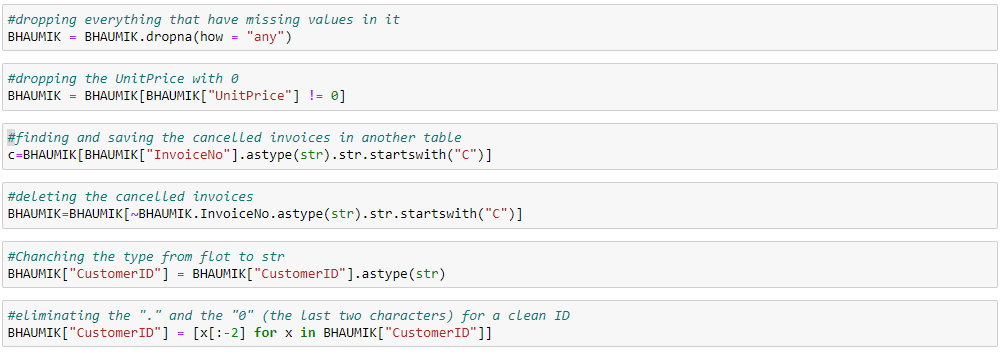


Before starting the information analysis, we want to scrub the dataset from redundant data and missing values. By running the code to seek out any missing values within the dataset, we found 1454 missing values in Description, and 135,080 missing values in CustomerID. We will Drop the outline column because the Stockcode identifies the identical item, and also the description can be different while the code is that the same. This could also prevent the following line to drop more significant data from other values as we don’t know where they’re specifically

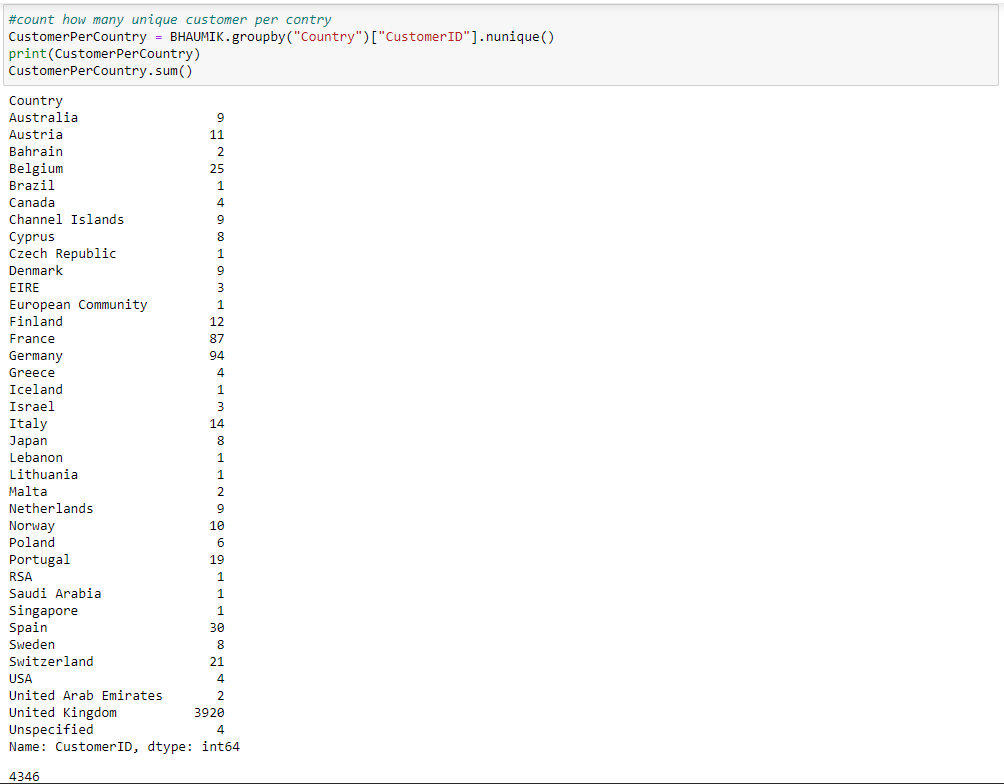
With all of this, we are now left with 541909 entries, which is the complete dataset. Although we discontinue a decent chunk of information (26.58%), it's still within the acceptable margin of omission. All the omitted data could not be averaged with mathematical operations; therefore, the omission is that the best solution.



Then, we are able to proceed with dropping all the missing values from the table, as they were only within the Description and CustomerID, which cannot be averaged. We will also drop any UnitPrice which is adequate zero. The cancelled invoices are going to be deleted from the database too. When checking the database, we will see that the CustomerID is taken into account a float when it should be nominal therefore it should be converted into a string. However, leave the CustomerID as same as before, which is 17850.0 rather than 17850. We added another step to the method which is cancelling the last two characters of the string

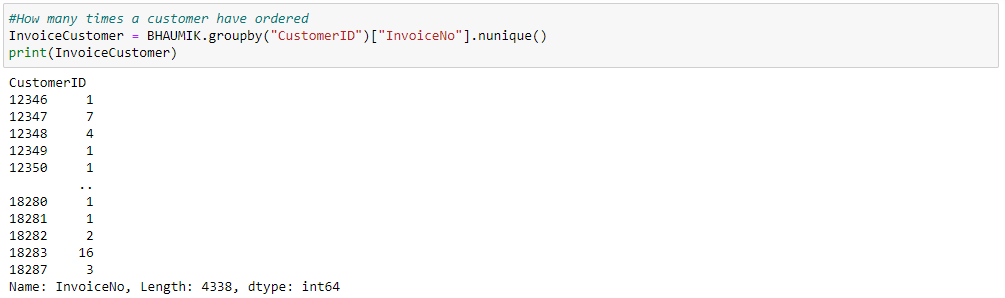


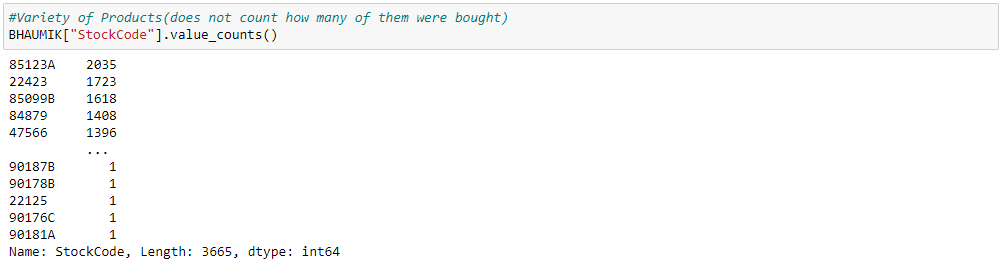
We can now count what number unique customers there are per country, it indicates that there are in total 4346 unique customers, with 3920 unique clients from the uk, which is 90.2% of the full customer base, making it the most country that the business operates with. The second highest in Germany (94), and third from France (87). The business has customers everywhere the globe, with the bulk from the European countries.

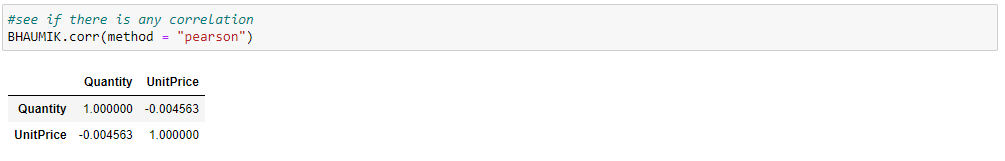


The business has customers everywhere the globe, with the bulk from the European countries. We know that there have been 18,532 invoices in total, with customers that have done business over 500 times, and many one-time customers furthermore. We can assume also the foremost popular products with their code.



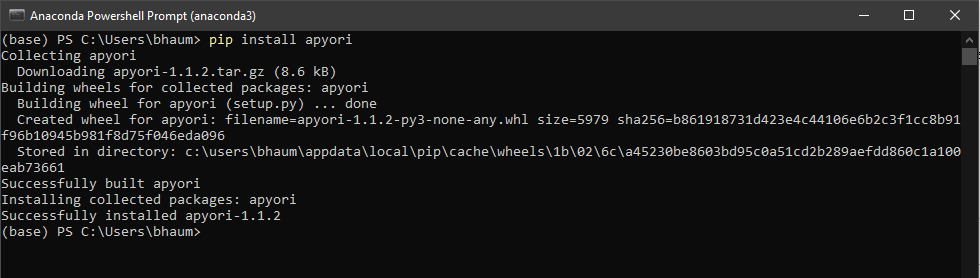




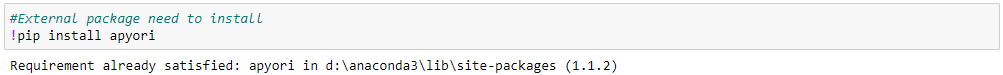


The only conclusion from the Pearson Correlation of this dataset is that the value doesn't affect in the slightest degree the quantity the client will purchase

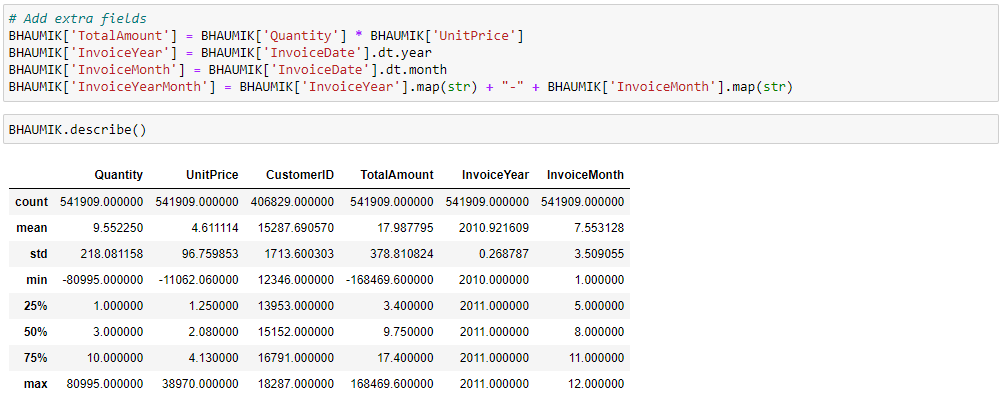
For performing Apriori, need to install apyori



After install apyori, we check in Jupiter

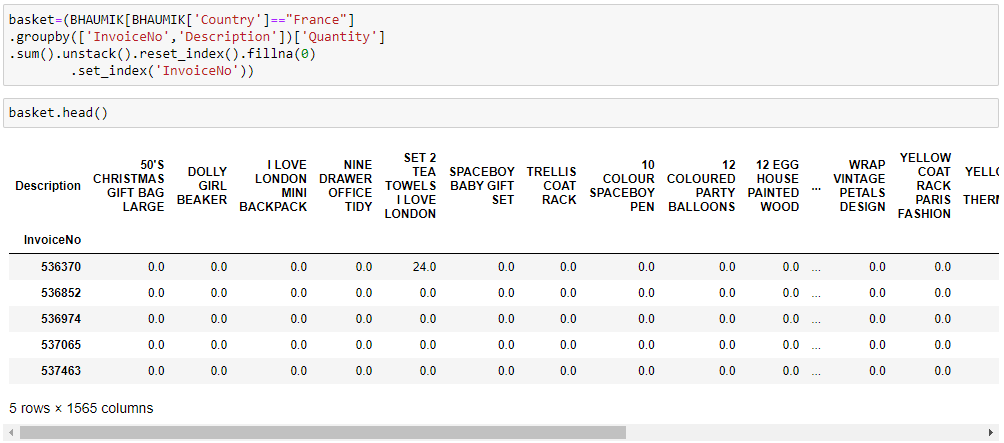


Add extra fields on the dataset and describe the dataset. I use my own name as BHAUMIK for Data set name



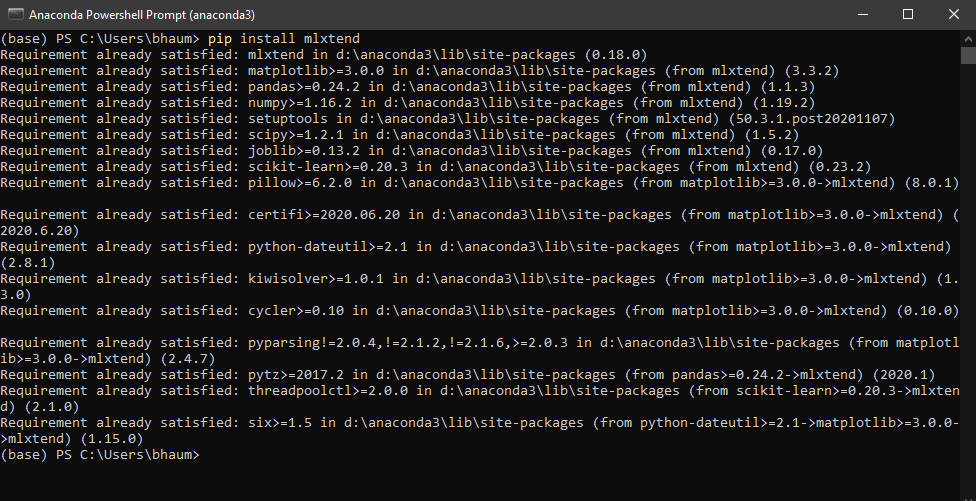
I interested in rules according to a different metric of interest so I can adjust the metric and min\_threshold arguments



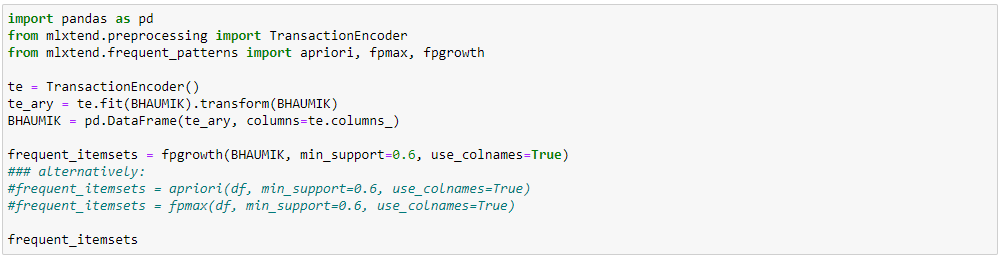


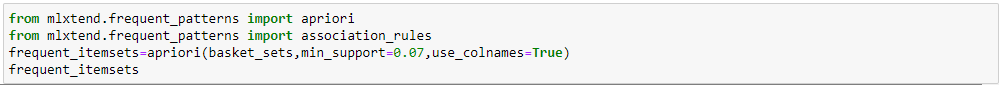


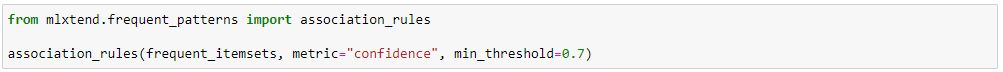
Install mlxtend through Powershall and after that confirm the installation via Jupyter



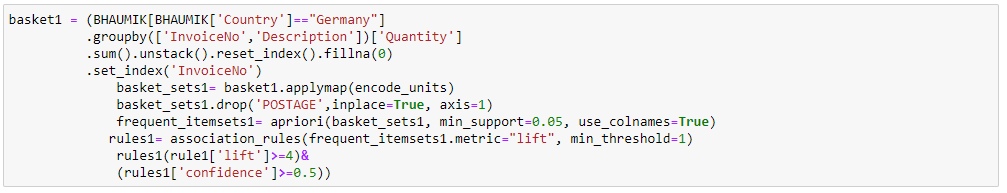
Now we can analyze the combinations of data vary by country of vast purchase. For this analysis we need to perform our patterns one by one countries



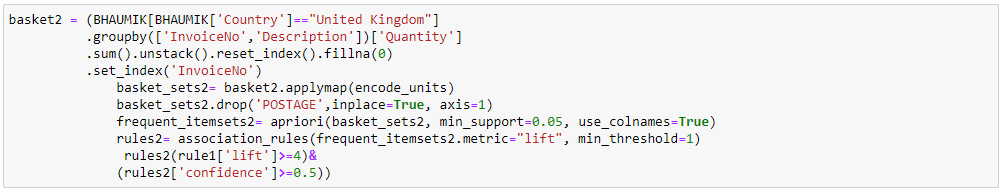




In the below script, I perform analysis on Germany and found they love Plasters in TIN Woodland Animals.



Through this analysis, United Kingdom love Green Regency Teaup and Saucer.



RITE people love Pack of 72 Retrospot Cake Cases



Now, with the reference to the analysis domain and by my own observation of the dataset. Business would be growth rapidly with the miner changes through red and green alarm clock, they put side by side to up the combined sales. Best things is to observe that both together.

On the similarities for some product put together so that people can buy that product with the product they instance to buy with a well as well as we can make some valuable offer to the customer.

Likewise, in the United Kingdom, Roses & Green Regency saucer can be together for the 85% growth climb same as Erie, Pink and Green regency Saucer consume more than Uk people least 92% confidence level.

For Example, if people come to the store to buy a Bread so we can put together Butter or Milk. So that with the one product other product also climb in the sale. This is the common and recommendation sales techniques follow by the business people.

|  |
| --- |
| 1. Reference |

* [**https://jupiter.math.nctu.edu.tw/~yuhjye/assets/file/teaching/2017\_machine\_learning/PCAsubset.pdf**](https://jupiter.math.nctu.edu.tw/~yuhjye/assets/file/teaching/2017_machine_learning/PCAsubset.pdf)
* [**http://www.dummies.com/education/math/statistics/how-to-interpret-a-scatterplot/**](http://www.dummies.com/education/math/statistics/how-to-interpret-a-scatterplot/)
* [**http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/graphs/scatterplot/interpret-the-results/key-results/**](http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/graphs/scatterplot/interpret-the-results/key-results/)
* [**http://www.independent.co.uk/life-style/health-and-families/health-news/the-secret-of-happiness-family-friends-and-your-environment-2053053.html**](http://www.independent.co.uk/life-style/health-and-families/health-news/the-secret-of-happiness-family-friends-and-your-environment-2053053.html)
* **Jupiter, Using for Coding and implementation**
* **app.aqldbm.com, Use this site for making diagram**
* **Module study Material**
* **cmaps color,https://matplotlib.org/stable/tutorials/colors/colormaps.html**
* **https://www.youtube.com/watch?v=guVvtZ7ZClw**