***REPORT WORK - ANSWERS***

**Part 2.1:-**

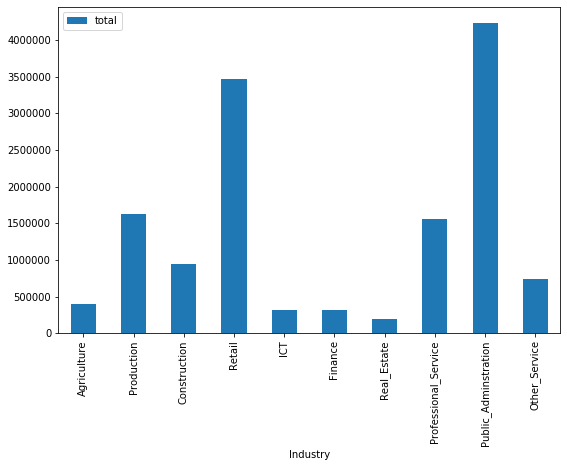
In this part of data analysis I used all the numeric columns and applied summation on it and created a new column named ‘total’ to our dataset.

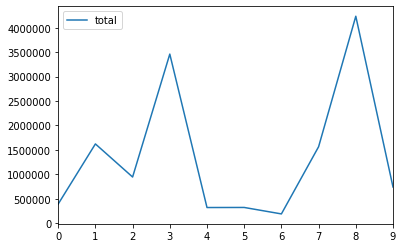
By using min() and max() function on ‘total’ column of data frame I got the required results as:

1. Name of Industry with LOWEST numbers of employed workers is : Real\_Estate
2. Name of Industry with HIGHEST numbers of employed workers is : Public\_Adminstration

After it I plotted this column (‘total’) with Industry column.

Plotted results are plotted as:





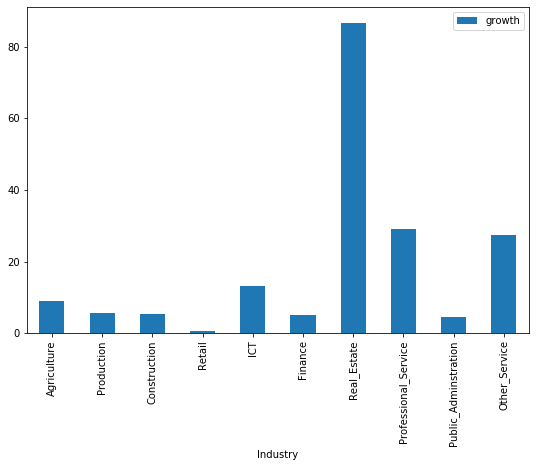
**Part 2.2:-**

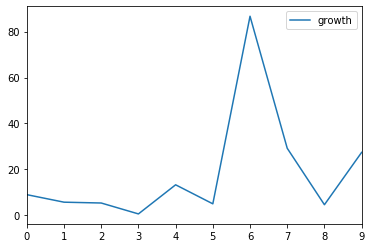
In this part I created a column ‘growth’ by finding total growth in percentage from year 2009 to 2018.

After it by applying min() and max() functions on data frame and got the industry names with highest and lowest overall growth as:

1. Real\_Estate has HIGHEST OVERALL GROWTH
2. Retail has HIGHEST OVERALL GROWTH

Then plotted these results as:



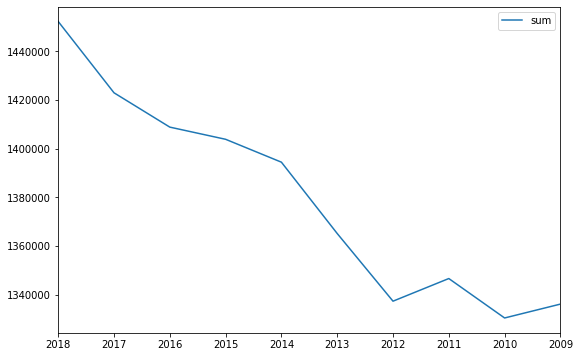


**Part 2.3:-**

In this part of data analysis I have find worst and best performed year by first finding summation of each year’s total employed workers in all the industries. Then used these results to get the required results as:

1. Year 2010 is Worst performed year with lowest employment.
2. Year 2018 is Best performed year with highest employment.

Then Plotted these results as:



**Part 4.2:-**

Does the aforementionedindustries are also correlated over the each year?

Yes, the industries are also correlated over the each year. Because the Correlation is between numerical columns of a dataset. As first we found correlation between average and Labeled­\_industries and after we found the correlation between Labeled\_industries and all the years. The average and the values of the years are related with each others.

**Part 5.1:-**

In this part I have to apply K-Means Clustering for on the results obtained from part 2.3 (Data frame with column 2018 and 2010 only) with K = 3 & K = 2.

Before moving further let’s have some idea about K-Means Clustering.

K-Means Clustering is one of the simplest and most popular Un-Supervised Machine-Learning-Algorithm and its main objective is “grouping similar data points and discover underlying patterns”.

To find the best results from this technique first we have to find value of K, by some methods like “Elbow” method.

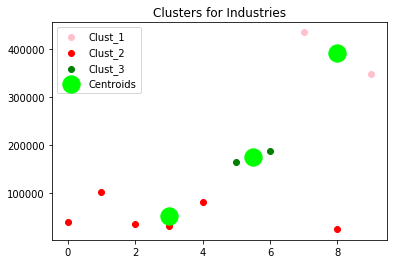
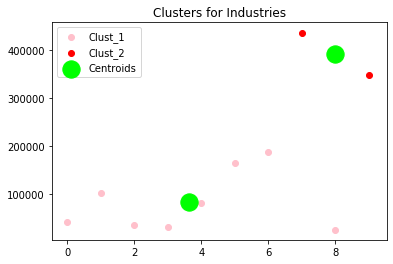
But Here, it is already given to use K = 2 & 3.

So using K = 2 and 3 on our data frame we got our data in form clusters (2 clusters and 3 clusters resp.).

K-Means Clustering can be called by Scikit Learn Library.

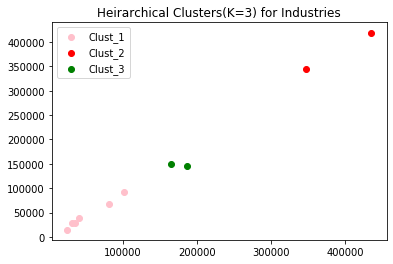
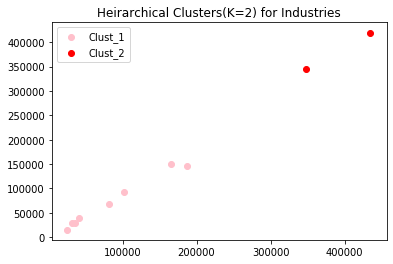
Below graph will show the findings for this part:

**K= 3 K = 2**

**Part 6:- Discussion**

1. The employment workers data of Wales from the StatsWales data source were taken (downloaded) as needed.
2. As our requirements we have a dataset with columns (“Industry”, 2018, 2017, …., 2010, 2009) and 10 rows (Industries for each year – ‘Professional\_Service', 'Public\_Adminstration', 'Other\_Service', 'Agriculture', 'Production', 'Construction', 'Retail', 'ICT', 'Finance', 'Real\_Estate').
3. We got the required dataset by performing data preprocessing task using **Pandas** library.
4. After it checked null values in dataset. There were no null values.
5. Then applied Data Analysis (Part 2) on dataset. After completing Part 2 we will move to Part 3.
6. In next part (Part 3) we will visualize the dataset by plotly library. So we applied Bubbled\_Scatter plot for change in years (consecutive years) and between year 2009 and 2018. After completing Part 3 we will move to part 4.
7. In next part (part 4 with 2 subsection 4.1 and 4.2) 1st subsection we have to find correlation between ‘Industries’ and ‘average’ column and next we have to find correlation between industries and all the years.
8. In 4.1 we got "Public\_Administration" industry as highest correlated industry (0.713420) and "Professional\_Service" industry as lowest correlated industry (0.046310).
9. Next in 4.2 we find correlation results for each years with industries.
10. Next I used first K = 3 and then K = 2 and applied K-Means Clustering on data frame (results from part 2.3) and got the results in form of clusters. (Fig. in Part 5.1)
11. Next used same dataset with same values of clusters for the Hierarchical Clustering and got the results in form of clusters as below:

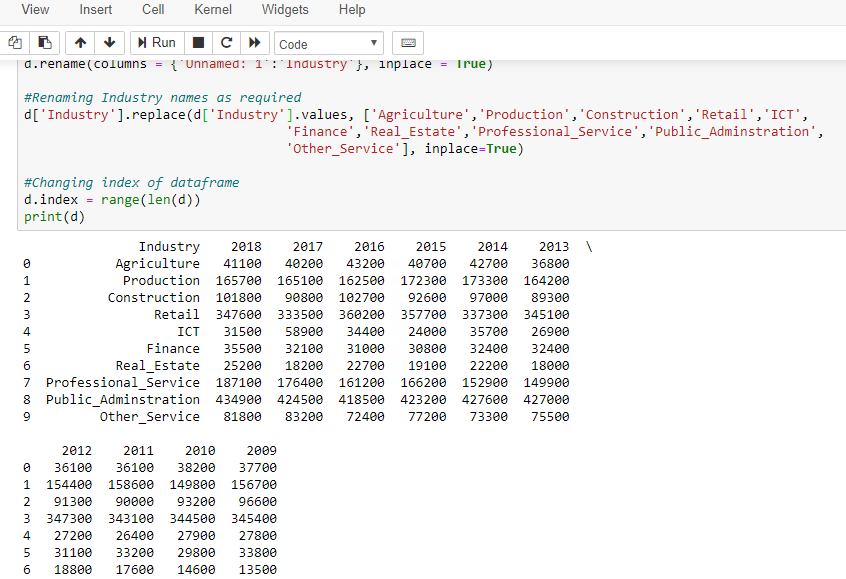
 

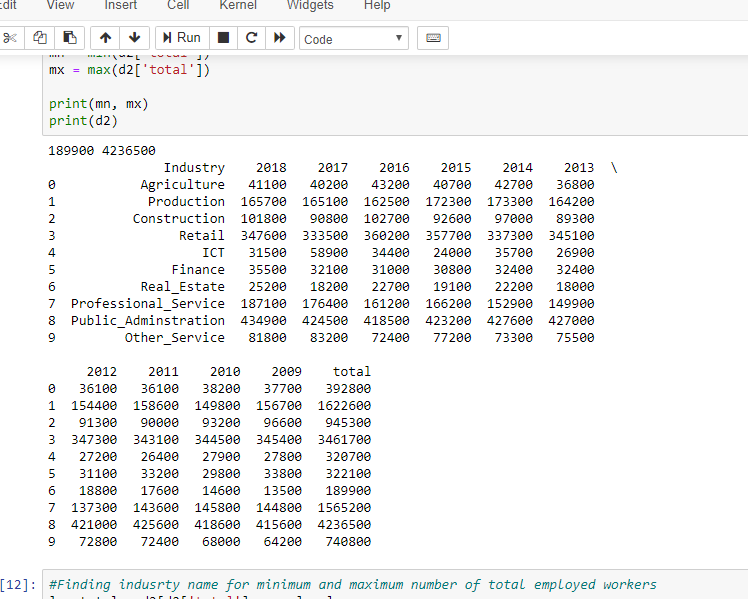
1. Comparing Both the Clusterings :

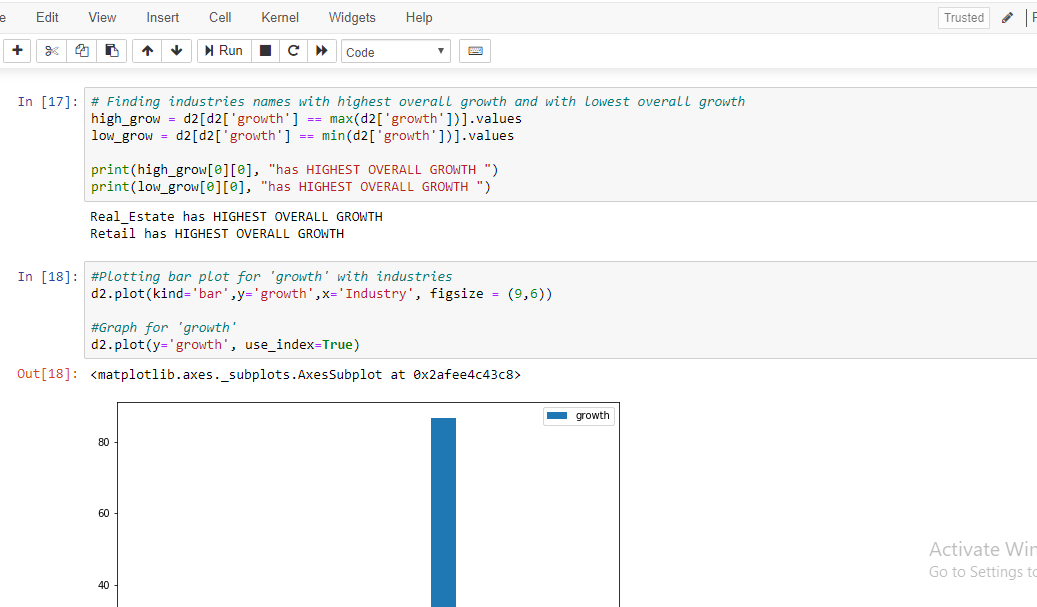
The results we got about both the clustering (K\_Means and Heirarchical) shows that they are almost similar (Same dataset for both).

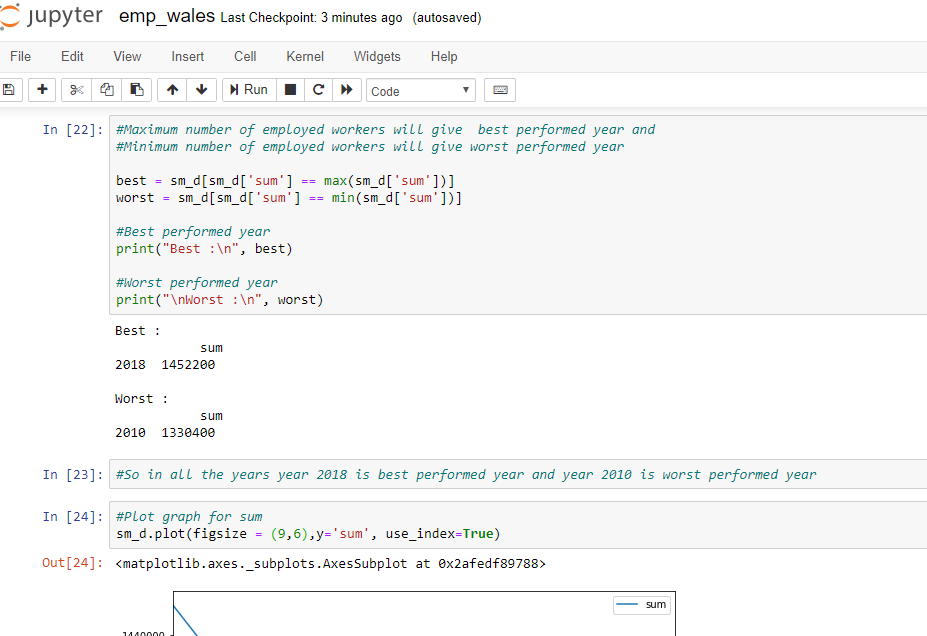
It may happen because we have a small dataset, the shape of clusters differ a little. However, along with many similarities, these two techniques have some differences also.

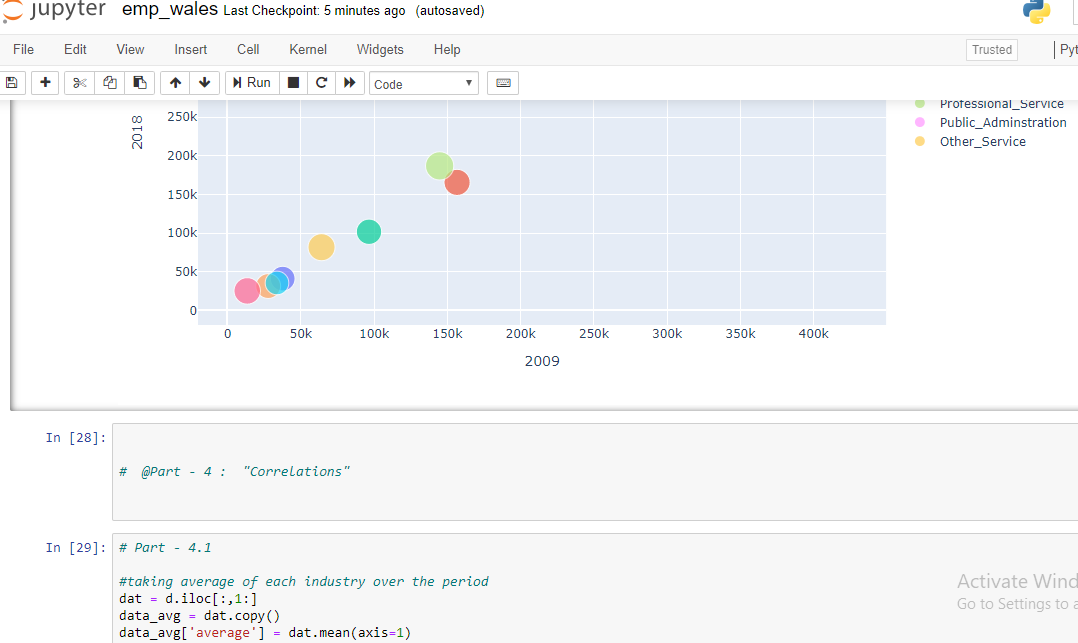
**SCREENSHOTS:-**

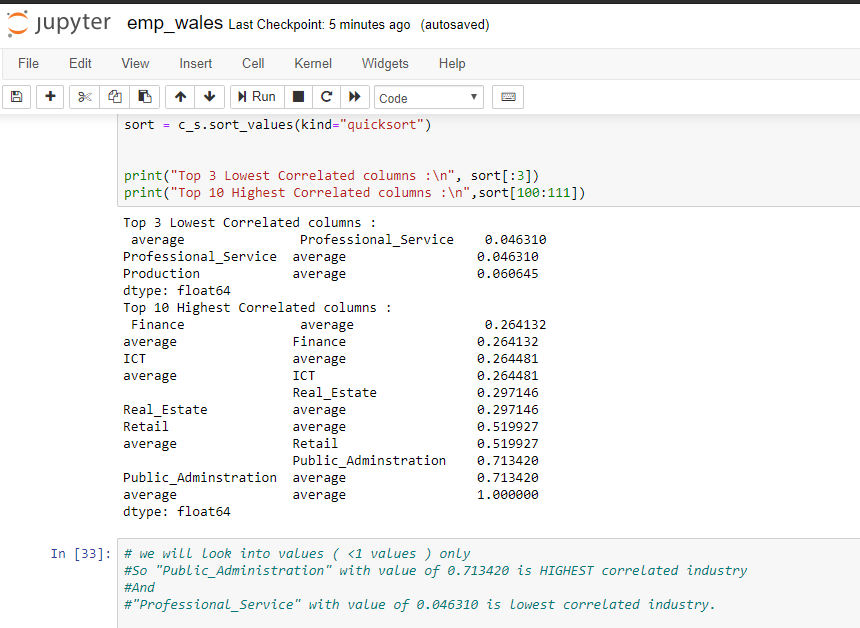


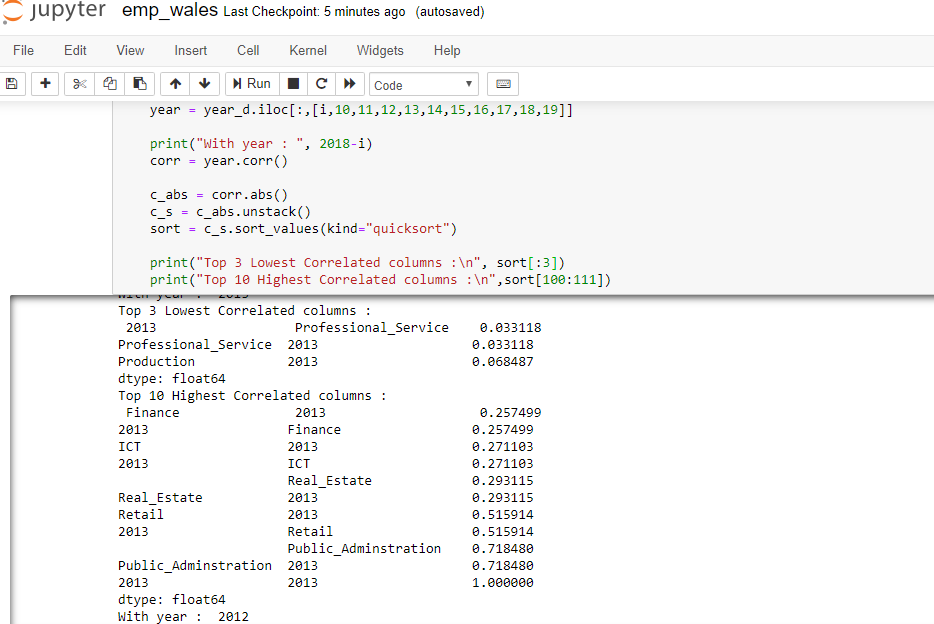


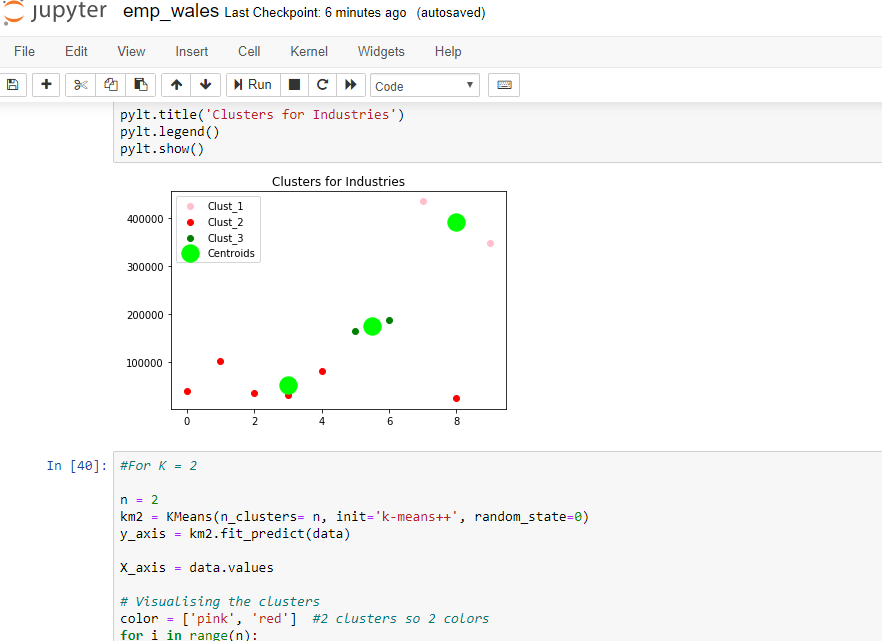


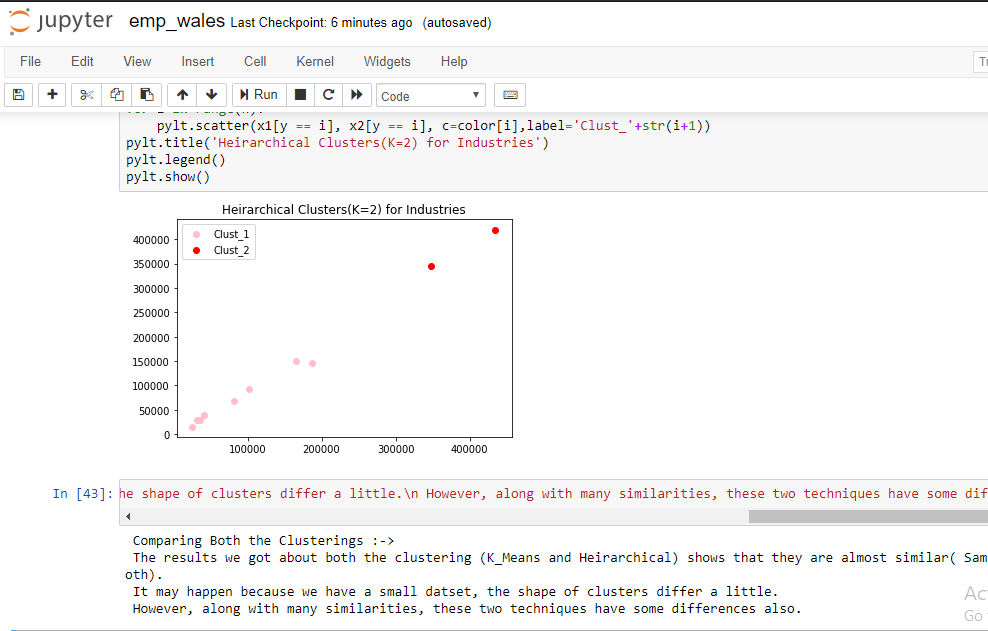












**# END #**