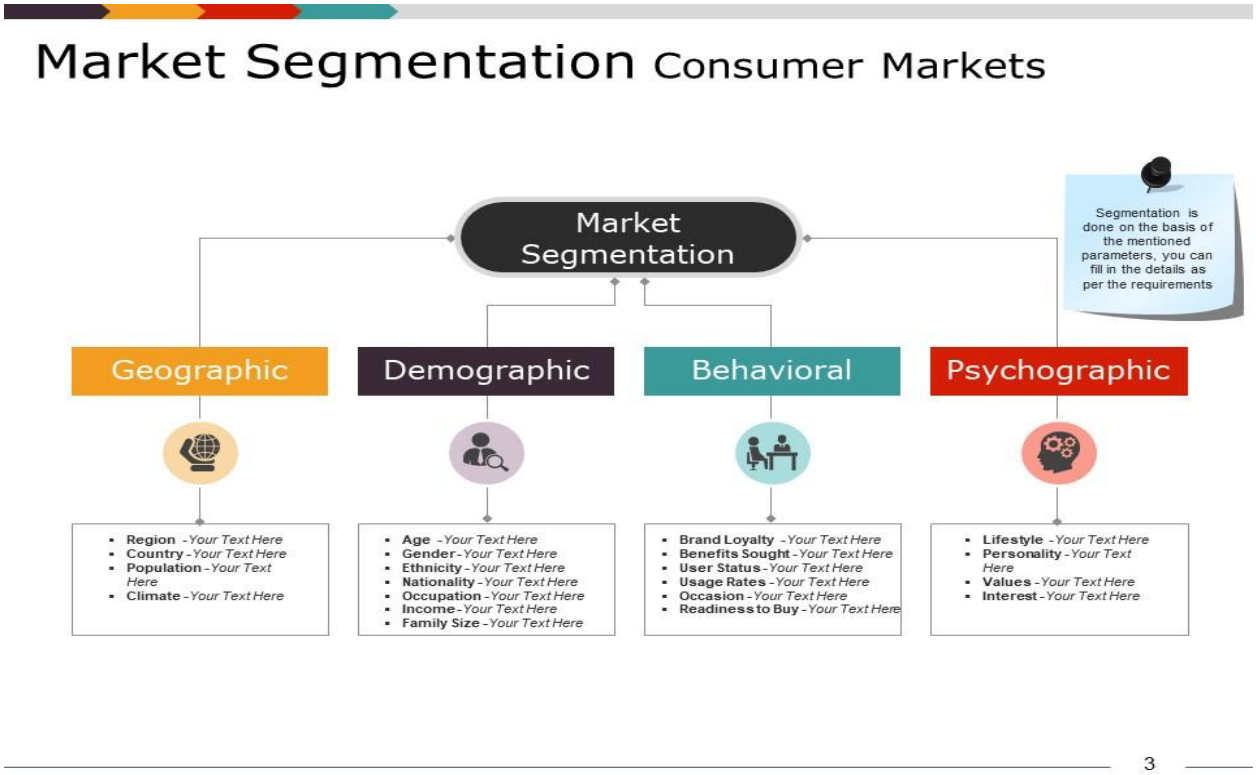
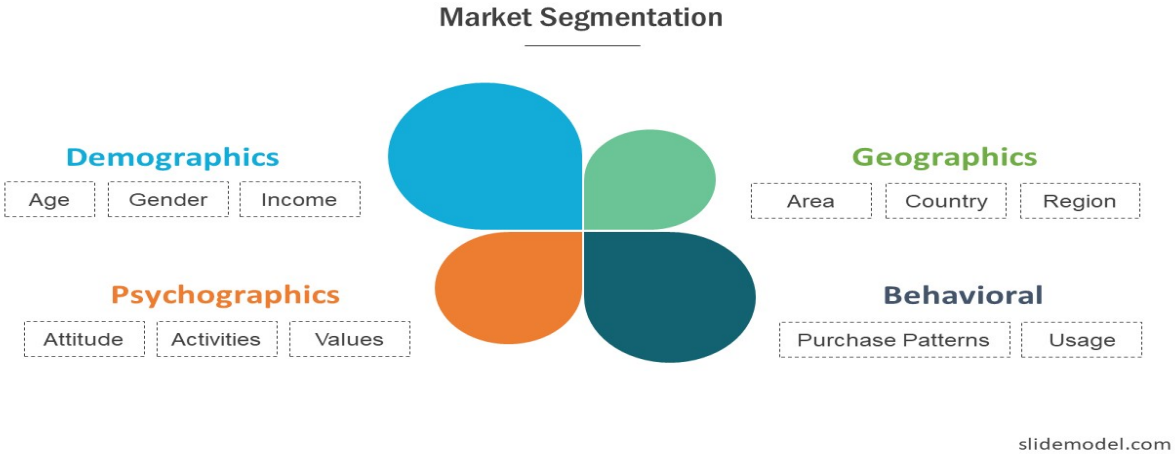
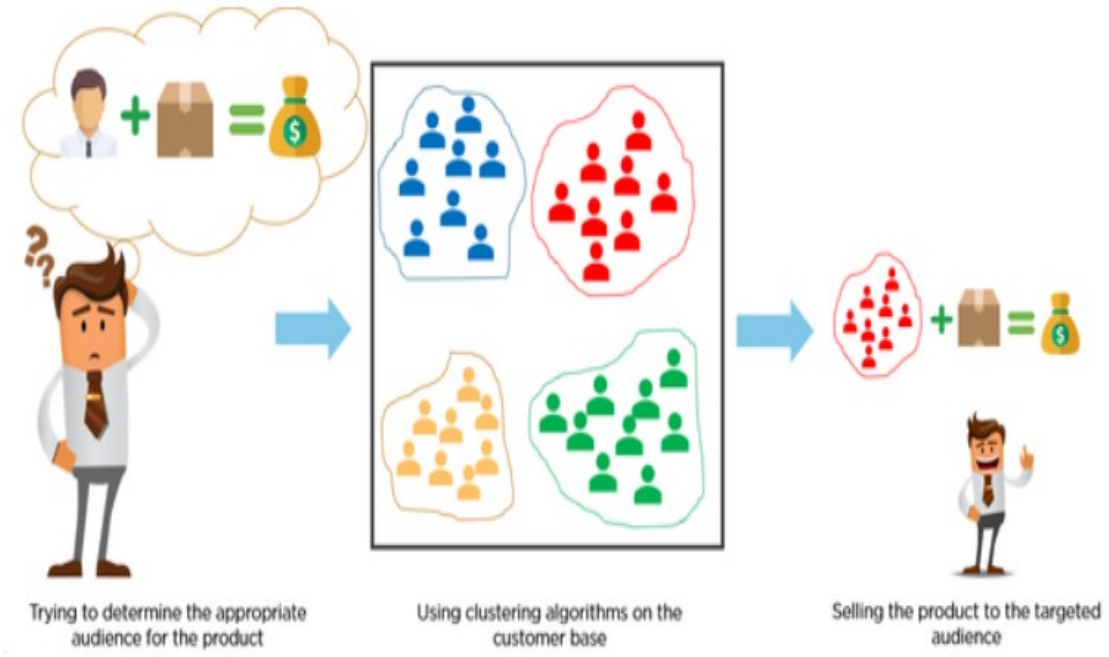
Market Segmentation Analysis
-By Saikat Adhikary





Introduction: *The emergence of many competitors and entrepreneurs has caused a lot of tension among competing businesses to find new buyers and keep the old ones. As a result of the predecessor, the need for exceptional customer service becomes appropriate regardless of the size of the business.Furthermore, the ability of any business to understand the needs of each of its customers will provide greater customer support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service. Each segment has customers who share the same market features. Big data ideas and machine learning have promoted greater acceptance of automated customer segmentation approaches in favor of traditional market analytics that often do not work when the customer base is very large. In this paper, the k-means clustering algorithm is used for this purpose. The Sklearn liabrary was developed for the k-Means algorithm and the program is trained using a 100-pattern two-factor dataset derived from the retail trade. Characteristics of average number of customer purchases and average number of monthly customers.*



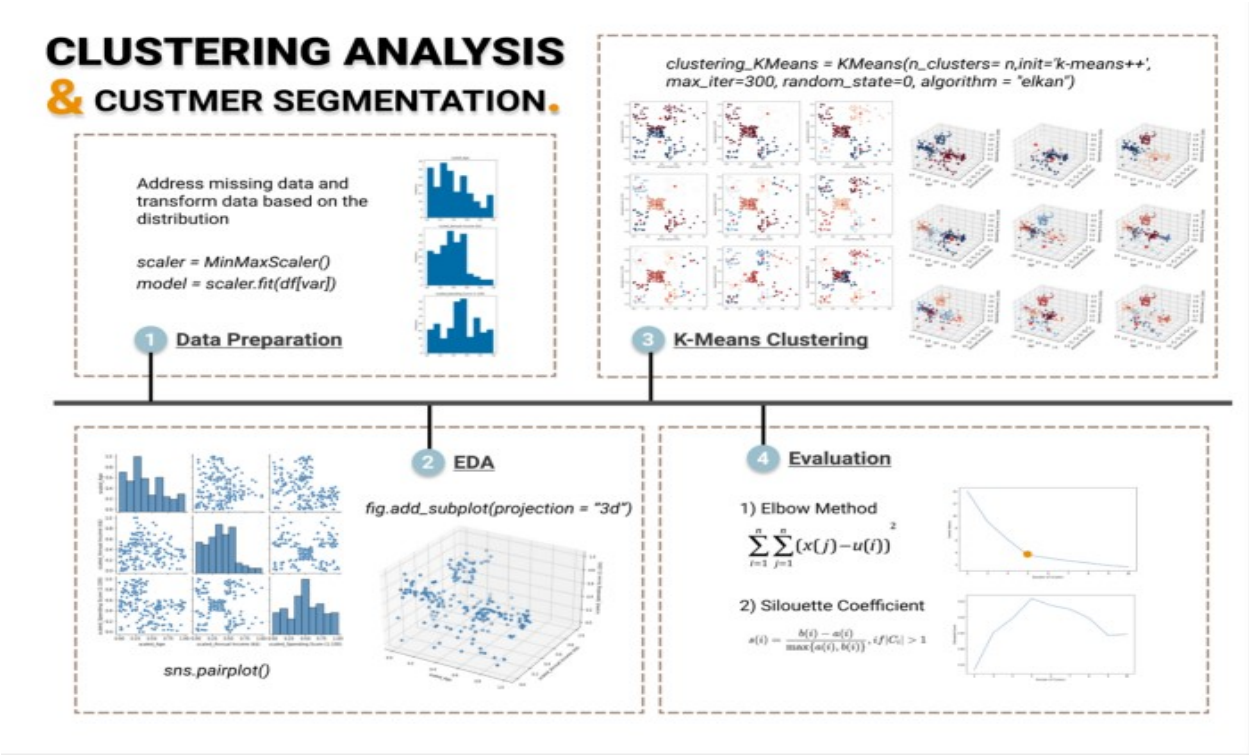
Data collection:kaggle.com

Github: https://github.com/saikatpythondev/ml-internship/blob/master/Untitled3.ipynb

Methods of customer classification : There are many ways to partition, which vary in severity, data requirements, and purpose. The following are some of the most commonly used methods, but this is not an incomplete list. There are papers that discuss artificial neural networks, particle determination and complex types of ensemble, but are not included due to limited exposure. In future articles, I may go into some of these options, but for now, these general methods should suffice. Each subsequent section of this article will include a basic description of the method, as well as a code example for the method used. If you do not have the expertise, well, just skip the code and you have to get a good handle on each of the 4 sub-sections included in this article.

Group analysis :Group analysis is an integration or unification, approach to consumers based on their similarity. There are 2 main types of categorical group analysis in market policy: hierarchical group analysis, and classification. In the meantime, we will discuss how to classify groups, called k-methods.

K. Means encounter :The K-means clustering algorithm is an algorithm often used to draw insights into formats and differences within a database.In marketing, it is often used to build customer segments and understand the behavior of these unique segments. Let's try to build an assembly model in Python's environment.



Proposed Model

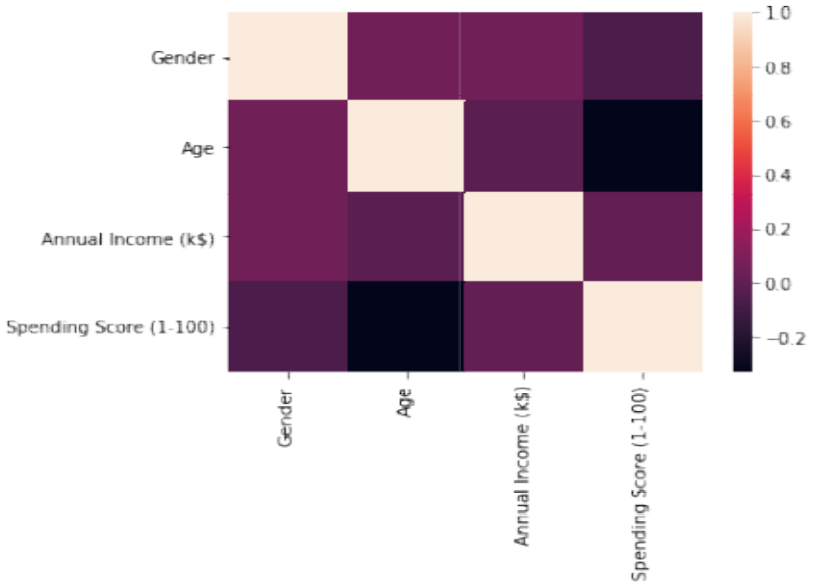
A) Import packages and data: To begin, we import the necessary packages to do our analysis and then the csv data file.If you want to follow up with the same data, For this example, I place the csv file in the folder (directory) where I present Jupyter notebook.

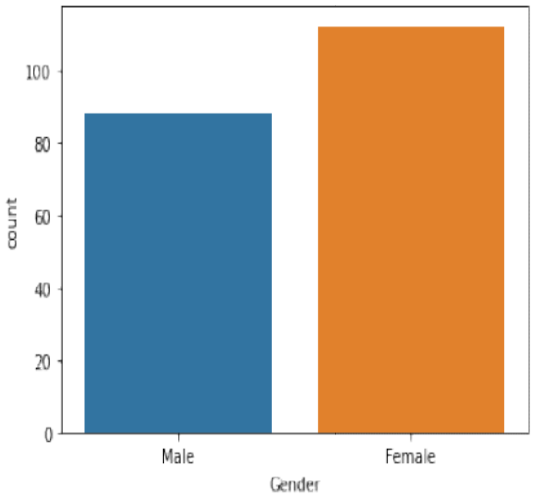
B) Data cleaning: After importing the package and data, we will see that the data is not as helpful as that, so we need to clean and organize this data in a way that we can create more actionable insights.

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

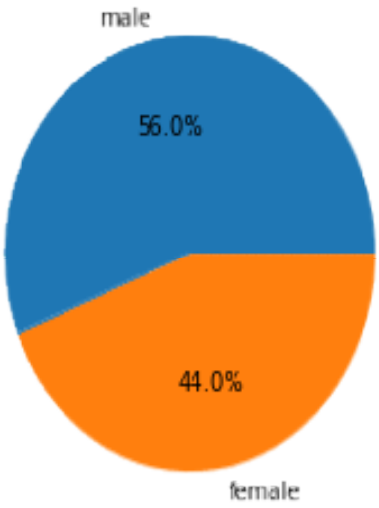
$c) EDA:$ after importing data we have to do the statistical analysis of the data

the correlation of each feature in the data is shown in the figure:

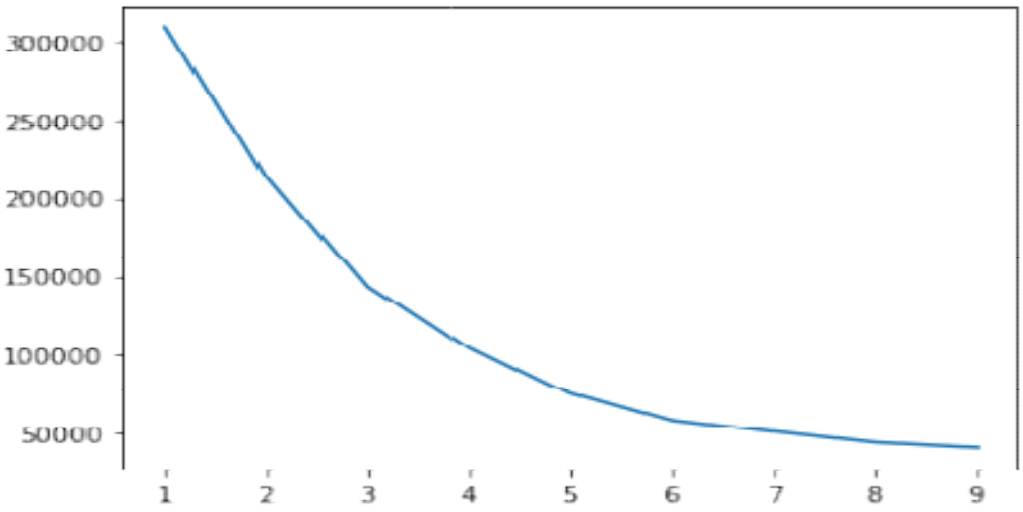




The count & pie plot of gender is also shown in the figure:

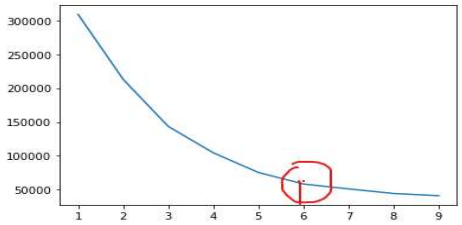


D)Elbow criterion method (with the sum of squared errors) (SSE): The idea behind the elbow method is to run a k-mean correlation in the data given for the k value (num_clusters, e.g. k = 1 to 9), and for each k value, calculate the sum of the squared errors (SSE). is. Then, adjust the SSE line for each k value.Here, we want to reduce SSE. SSE usually falls to 0 as we go up k (and SSE is 0 where k is equal to the number of data points, because where each data point has its own set, and there is no error between it and its trunk) . The objective is therefore to select a smaller value of k, which still has a lower SSE, and the cone usually represents where it begins to return negatively with increasing.



Finally ,optimum value for kmeans algorithm we set the k=6 for the result

```
Out[60]: [<matplotlib.lines.Line2D at 0x21dede06c70>]
```




```
In [64]: km=KMeans(n_clusters=6)
         km.fit(file)

Out[64]: KMeans(n_clusters=6)
```

E) Explaining customer segment:

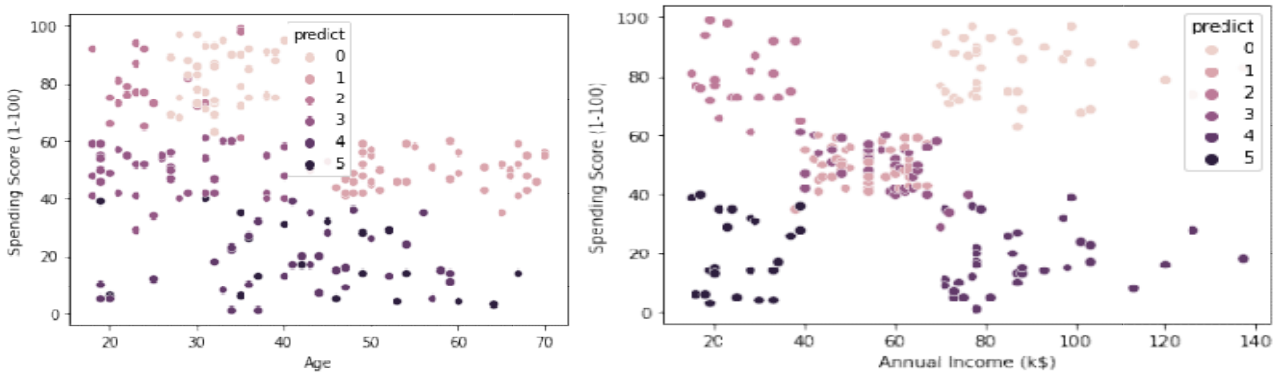
```
file["predict"]=pred
```

```
file
```

	Gender	Age	Annual Income (k$)	Spending Score (1-100)	predict
0	1	19	15	39	5
1	1	21	15	81	2
2	0	20	16	6	5
3	0	23	16	77	2
4	0	31	17	40	5
...	...	...	...	...	...
195	0	35	120	79	0
196	0	45	126	28	4
197	1	32	126	74	0
198	1	32	137	18	4
199	1	30	137	83	0

Now we have to combine the matrix of integration and see what we can gather from the standard data for each cluster.

f) Cluster visualization: after expending new column for cluster we visualize the cluster with previous feature that how they are depending ,we use seaborn package to make the scatter plots:



Result:

Predict the dependency of cluster with other features:

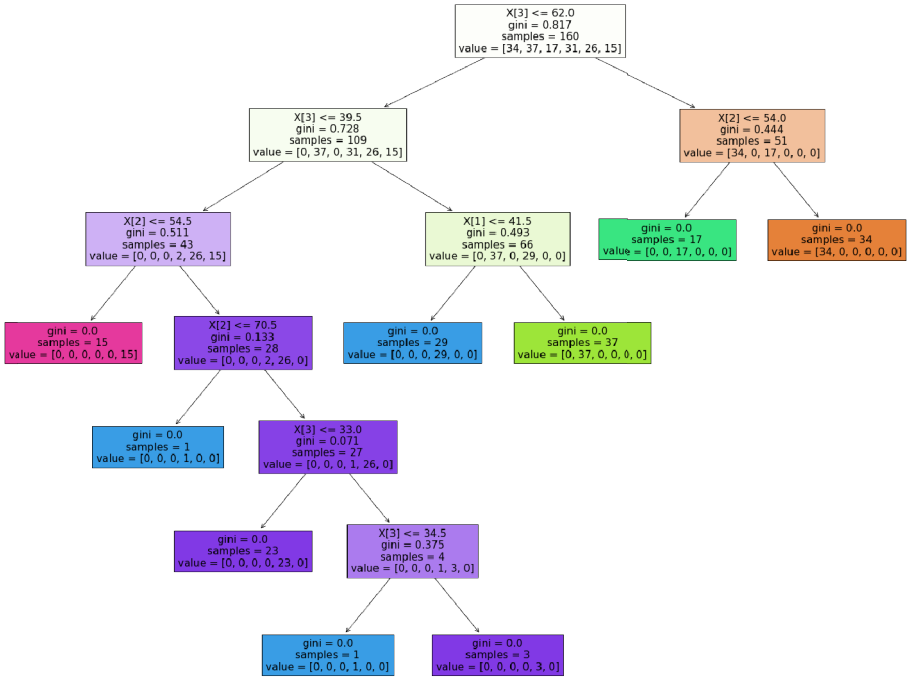
Here we use supervised learning algorithm to predict how the cluster vary with other features of the dataset we mainly use the decision tree classifier to make the model ,here the dependent feature is the column :"predict"

i.e,the cluster and independent feature is other column in the dataset

so finally we got the model with 92.5% accuracy

and the Tree looks like :

```
|--- feature_3 <= 62.00
|   |--- feature_3 <= 39.50
|   |   |--- feature_2 <= 54.50
|   |   |   |--- class: 5
|   |   |--- feature_2 >  54.50
|   |   |   |--- feature_2 <= 70.50
|   |   |   |   |--- class: 3
|   |   |   |--- feature_2 >  70.50
|   |   |   |   |--- feature_3 <= 33.00
|   |   |   |   |   |--- class: 4
|   |   |   |   |--- feature_3 >  33.00
|   |   |   |   |   |--- feature_3 <= 34.50
|   |   |   |   |   |   |--- class: 3
|   |   |   |   |   |--- feature_3 >  34.50
|   |   |   |   |   |   |--- class: 4
|   |--- feature_3 >  39.50
|   |   |--- feature_1 <= 41.50
|   |   |   |--- class: 3
|   |   |--- feature_1 >  41.50
|   |   |   |--- class: 1
|--- feature_3 >  62.00
|   |--- feature_2 <= 54.00
|   |   |--- class: 2
|   |--- feature_2 >  54.00
|   |   |--- class: 0
```



Acknowledgment:

We would like to thank the administration of Feynn lab for providing us with opportunity to carrying out such a detailed analysis and providing us with the exposure to industry business analytics.

We would also like to thank our parents ,friends and associates for their constant support ,aid and advice