Mall Customer Dataset

Name: Sakshi Mhatre, Gopikha Menon, Rithwik Vedpathak

Student: Comps College name: Vidyavardhini's College of Engineering and Technology

---

1. ***ABSTRACT:***

In Mall Customer Dataset, the main aim is to optimize sales and policies for an organization. Catering to target audience thus boosts profits, and understanding which group sustains losses allows marketing to understand various aspects of improvement. Taking into factor many key-points, the appropriate solution can be derived.

2. ***INTRODUCTION:***

Customer Segmentation can be a powerful means to identify unsatisfied customer needs. This technique can be used by companies to outperform the competition by developing uniquely appealing products and services.

The most common ways in which businesses segment their customer base are:
1.Demographic information: such as gender, age, familial and marital status, income, education, and occupation.
2.Psychographics: such as social class, lifestyle, and personality traits.
3.Behavioral data: such as spending and consumption habits, product/service usage, and desired benefits.

3. ***DATA COLLECTION:***

There are various websites and datasets available consisting of a mall customer dataset. The dataset usedfor this article was taken from Kaggle.com.

DATA DESCRIPTION:

data - DataFrame

Index	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
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72
10	11	Male	67	19	14
11	12	Female	35	19	99
12	13	Female	58	20	15
13	14	Female	24	20	77

We will take into consideration 200 customers, their gender, age, annual income (k$) and spending score. These factors will help us determine and arrive at our desired conclusion. The above table only shows a few entries.

CLEANING AND PREPARING DATA:

Procuring the data was not that difficult as it was readily available on kaggle.com, the cleaning process did not take much time either.Invalid data such as negative values, misprinted values and outliers wereremoved along with duplicate values.

4. ***MACHINE LEARNING MODELS:***

Will be using various techniques such as exploratory data analysis, Linear Regression. Packages like pandas, seaborn, matplotlib, numpy, plotly were used.

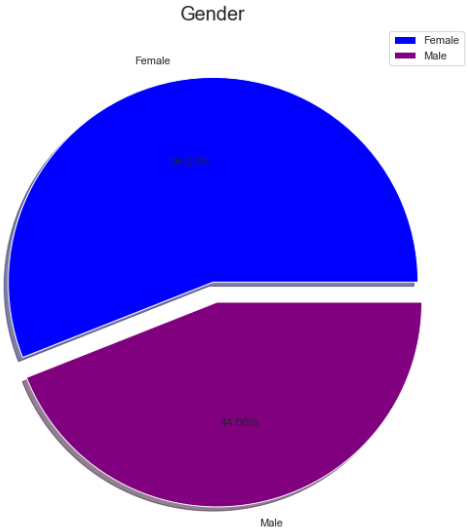
K-MEANS CLUSTERING:

The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset.

- Specify number of clusters *K*.
- Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters is not changing.

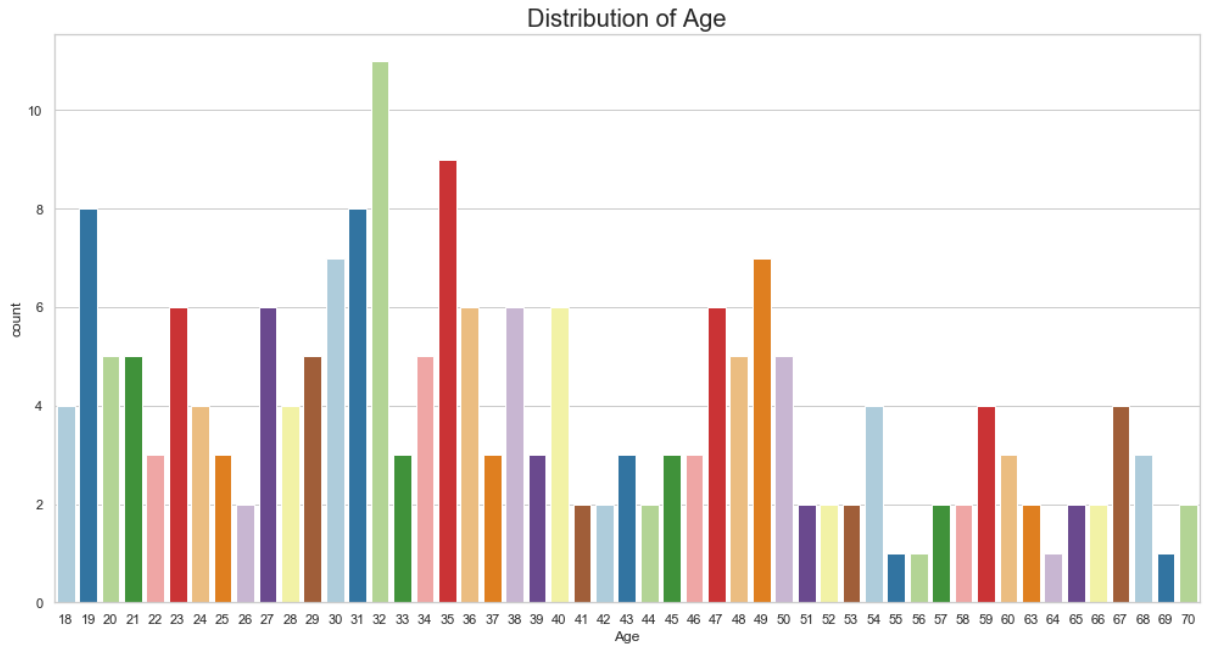
5. ***EXPERIMENTAL RESULTS:***

- **Distribution of Gender:**



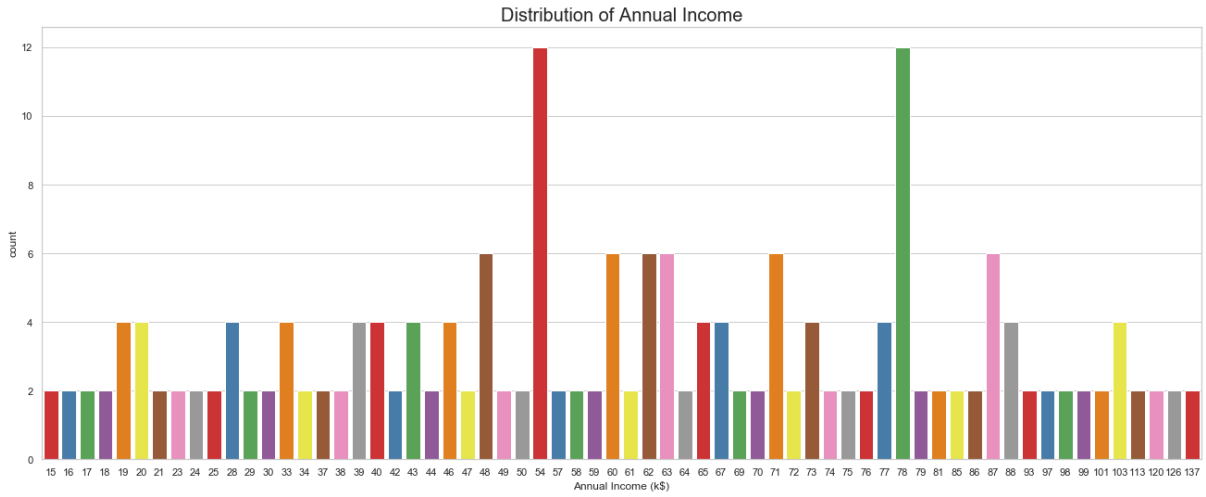
Females are the dominant with a staggering 56% whereas men trail behind with 44%, thus not shopping as much.

- **Distribution of Age:**



Although the age ranges are all over the place, we can infer that the age group of 30-40 are the most frequent to shop, with age group of 32 the highest. As we go towards the senior citizens, the graph begins to taper down as they do not visit as frequently. Younger people, less than 30 years old, are quite frequent as well. Lesser than those in their 30's, yet frequent.

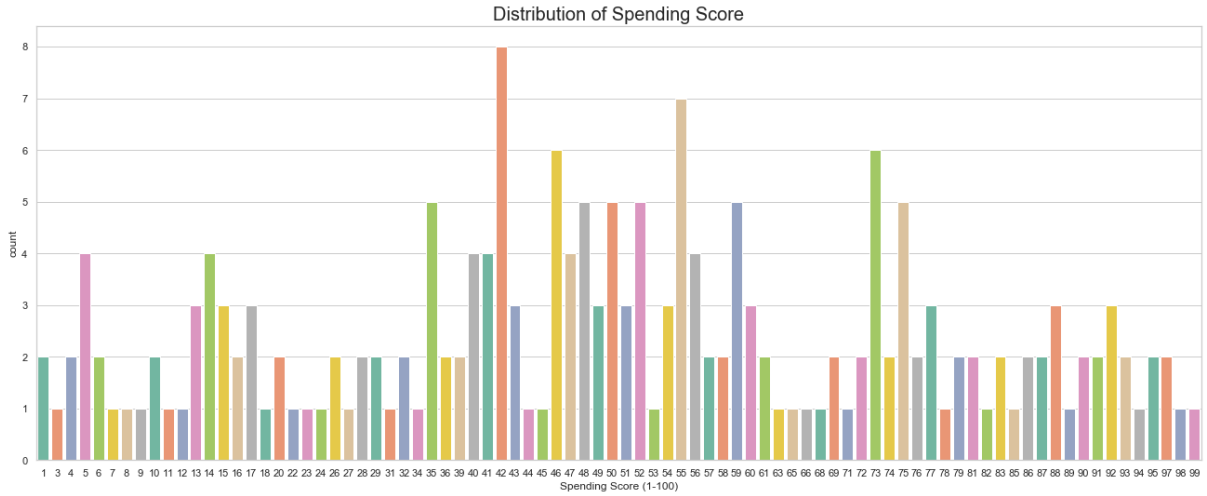
- **Distribution of Annual Income:**



Showing the distribution level of income, we can confer that the most frequently visiting customers at the mall have an annual income of 54k or 78k. The extreme

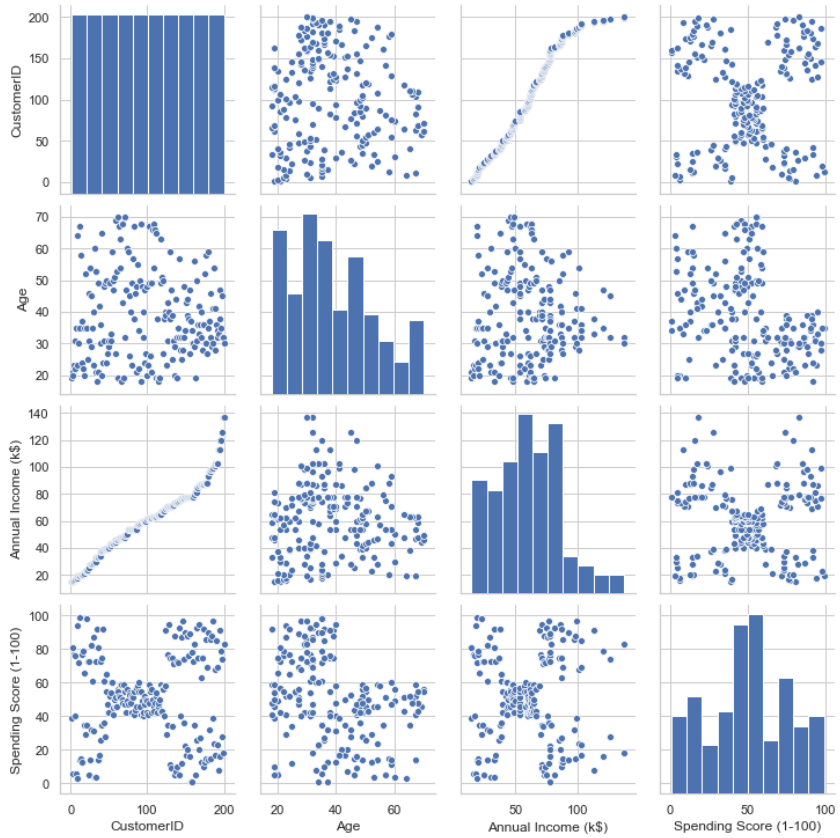
wealthy and financially struggling are not as likely to visit.

- **Distribution of Spending Score:**



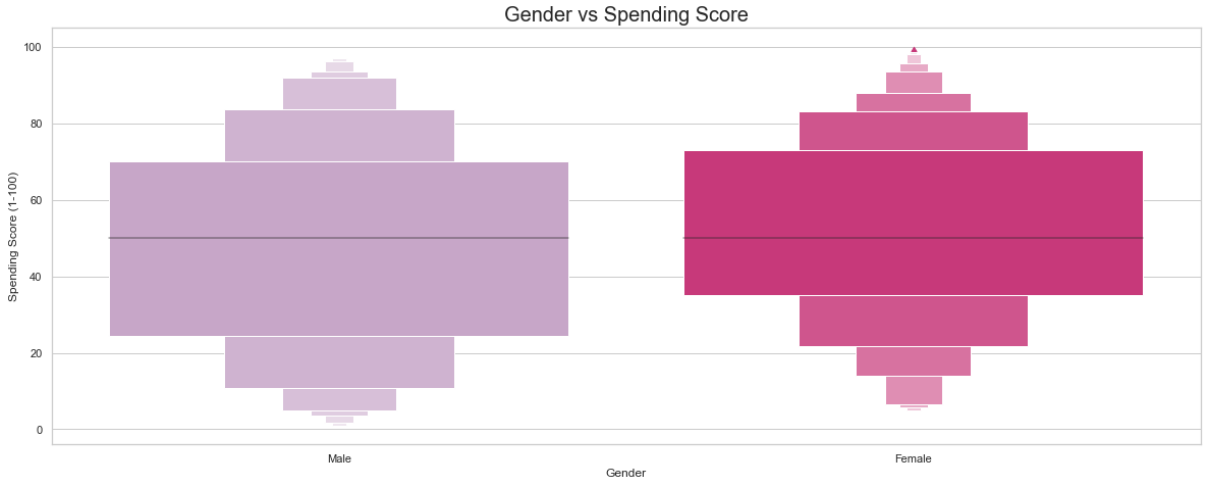
Overall, we may conclude that most customers have their spending scores in the range of 35-60. There are quite a few with a spending score of 80+ as well as less than 30. This is an equal balance considering people who visit just a few times while passing by like tourists and the other end of the spectrum for those who possibly live nearby thus the mall being convenient for everyday needs as well as other things.

- **Pairplot of Data:**



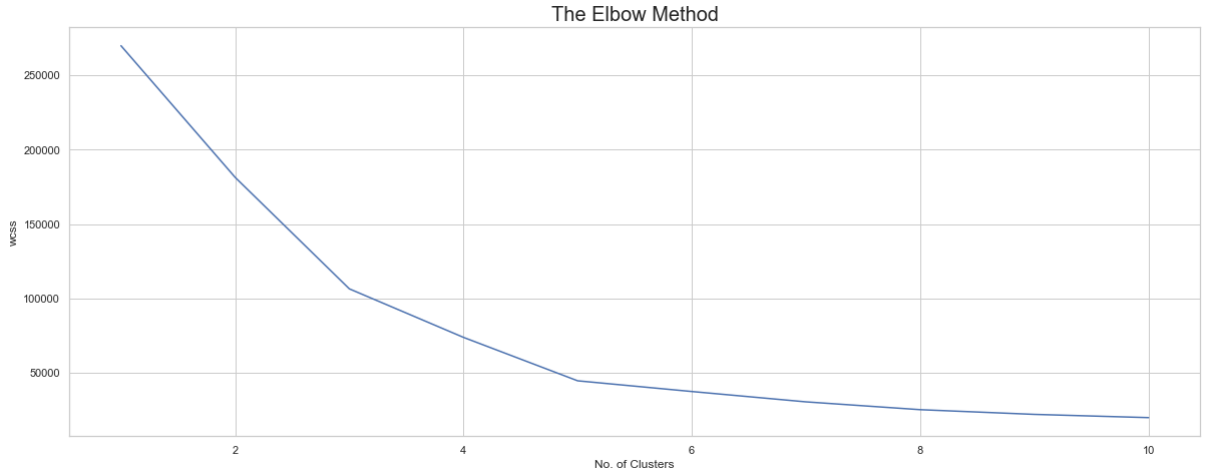
A pairplot of every attribute from the dataset. Thus, letting us explore many points.

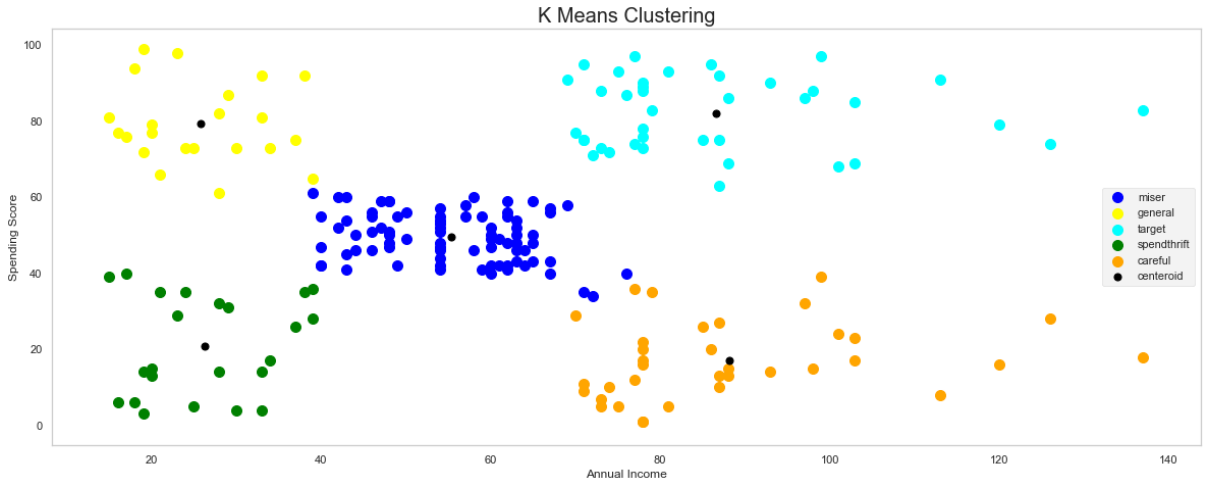
- **Gender vs Spending Score :**



Whilst men have a Spending score of around 25k -70k, women have a spending score of 38k-75k. This bi-variate analysis allows us to understand that female customers tend to spend more than male customers.

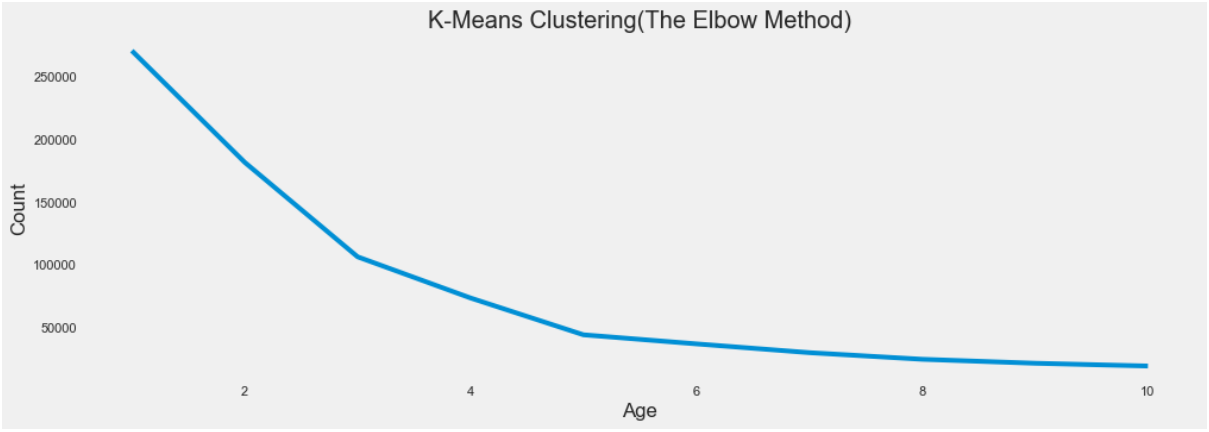
- **Annual Income vs Spendscore (K Means Alg):**

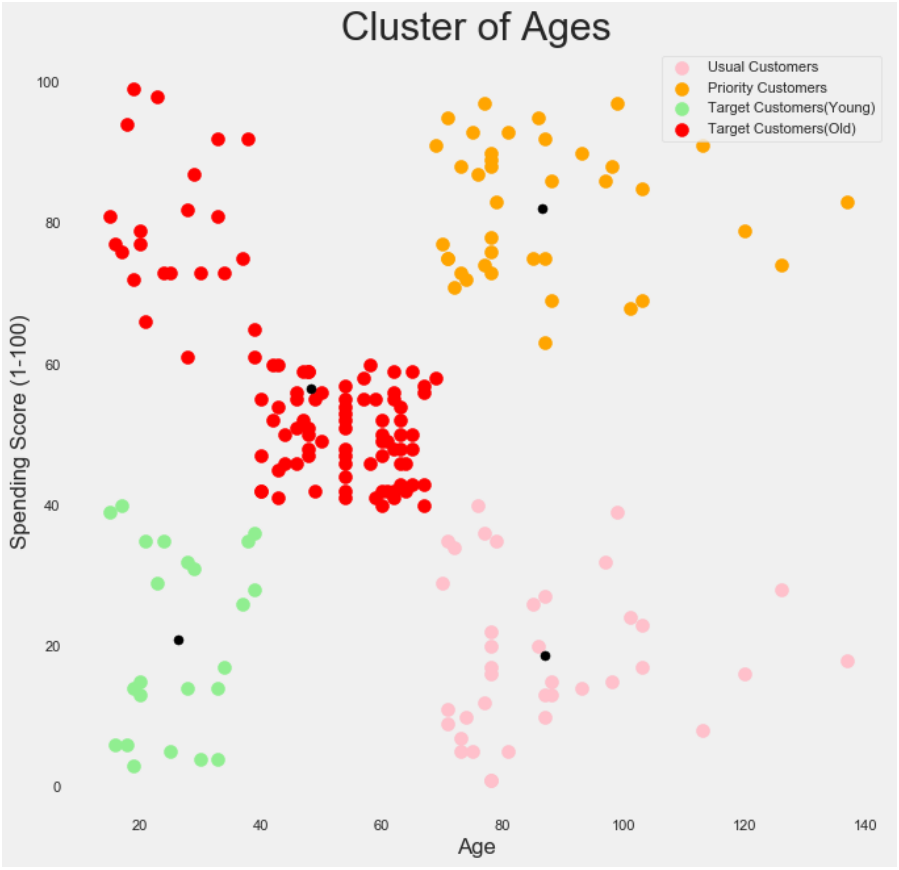




Clustering Analysis gives us a noticeably clear insight about the different segments of the customers, namely miser, general, target, spendthrift and careful. We can segregate them based on their Annual Income and Spending Score. Here, once we have identified the target audience, we can make sure that the mall caters to them the most, as they are the ones who would in turn procure a profit. The miser category of customers can be avoided, and we must not be catering to it. The rest can be promotionally catered to according to the category of products.
The target audience in this would be those with an annual income of 70k-100k and a spending score of 60-100. There are also some outliers which should be taken into consideration as they are the ones with high annual income and high spending score.

- **Age vs Spending Score (K Means Alg):**





Here, the cluster of age and spending score gives a more detailed idea of what marketing strategies and policies would optimize the spending scores of the customers in the mall. We have separated the customers into 4 categories- Usual Customers, Priority Customers, Senior Customers(target), Young Customers(target).
Here the Older target audience is between the ages of 45-70 and have a spending score of 40-60. The Younger Target audience is of age 20-40 with a spending score of 40-60. These are target audiences as they possess the potential to spend more and thus increase sales and profits.

6. Conclusion:

By analyzing the given data into appropriate graphs and liner regressions, we have

optimized market strategies and ensured that efficiency is maintained. An overall look on all the customers has been done with respect to all the attributes. Thus, understanding the gains and losses with each. The target audience has been narrowed down and the non-target audience as well. This is necessary as to avoid losses.

7. Future Work:

With the help of limited data, one can only predict the overall sales and expenditure from the customers. If more data can be provided then we can find specific profits/losses of a product, sector etc. Hence, to improve efficiency, more data is necessary.