Case Study on Business Governance using Data Analytics

Author: Anoushka M., Apurva R., Aamuktha R., Bindu C.

Bachelor of Technology - Computer Science and Engineering, GITAM University, Hyderabad

Abstract: *In this case study, our main focus is to understand the governance of data and its applications in the business domain. For retail businesses to thrive in a data-centric world, it is imperative that they retain control over their data - from how it's created to who has access to it, to where it's being stored. Over the past few years, retail industries have undergone a massive transformation as customers became digital natives and their buying habits changed. The transformation hasn't been easy and today, retailers deal with immense amounts of data to analyse consumer trends, manage inventories and predict future demands. In light of this, our main objective through this paper is to show how data mining can help discover relevant knowledge contained in databases obtained from a well-based Superstore. These findings can be used to help business owners of a certain retail firm keep track of how their company is performing and find out the weak areas over which the company can develop in order to make profit. On the whole, as the retail industry will undergo more changes in the future, it is necessary to be prepared to establish a culture of data governance to derive improved business outcomes.*

Keywords: *Data Mining, Knowledge Discovery Database, Data Analysis, Data Visualization, Cost-Benefit Analysis, Clustering, Data Classification, Business Analytics*

1. Introduction

In businesses today, we have multitudes of companies that go above and beyond to work efficiently in order to gain a competitive advantage over others. A rapid-growing and incredibly popular technology that can help gain this advantage is data mining. Data mining technology allows a company to use the mass quantities of data that it has compiled, and develop correlations and relationships among this data to help businesses improve efficiency, understand their customers, make better organizational decisions, and build on effective planning strategies.

Data Mining is segregated into three major constituents - **Clustering or Classification, Association Rules and Sequence Analysis**. Data Mining acts as a tool that extracts predictive information from immense amounts of data. By using mathematical and statistical calculations to uncover various trends and corrections among the data stored in a database. It is a blend of Artificial Intelligence technology, Statistics, Data Warehousing, and Machine Learning.

Data mining set base with statistics. Statistical functions such as standard deviation, regression analysis, and variance are all valuable tools that allow people to study them to solve problems. Furthermore, data mining tools are conceptually designed to allow you to predict future trends. Business intelligence experts use dedicated software tools and advanced mathematical algorithms to extract patterns from large quantities of data and evaluate trends and the probability of future events along the same.

2. Data Analysis and Visualization

Data analysis and data visualization might have different views, but they also exist to work together efficiently. This way, one is presented with the opportunity to create visual analytics that showcase complex data sets in an easy and understandable way.

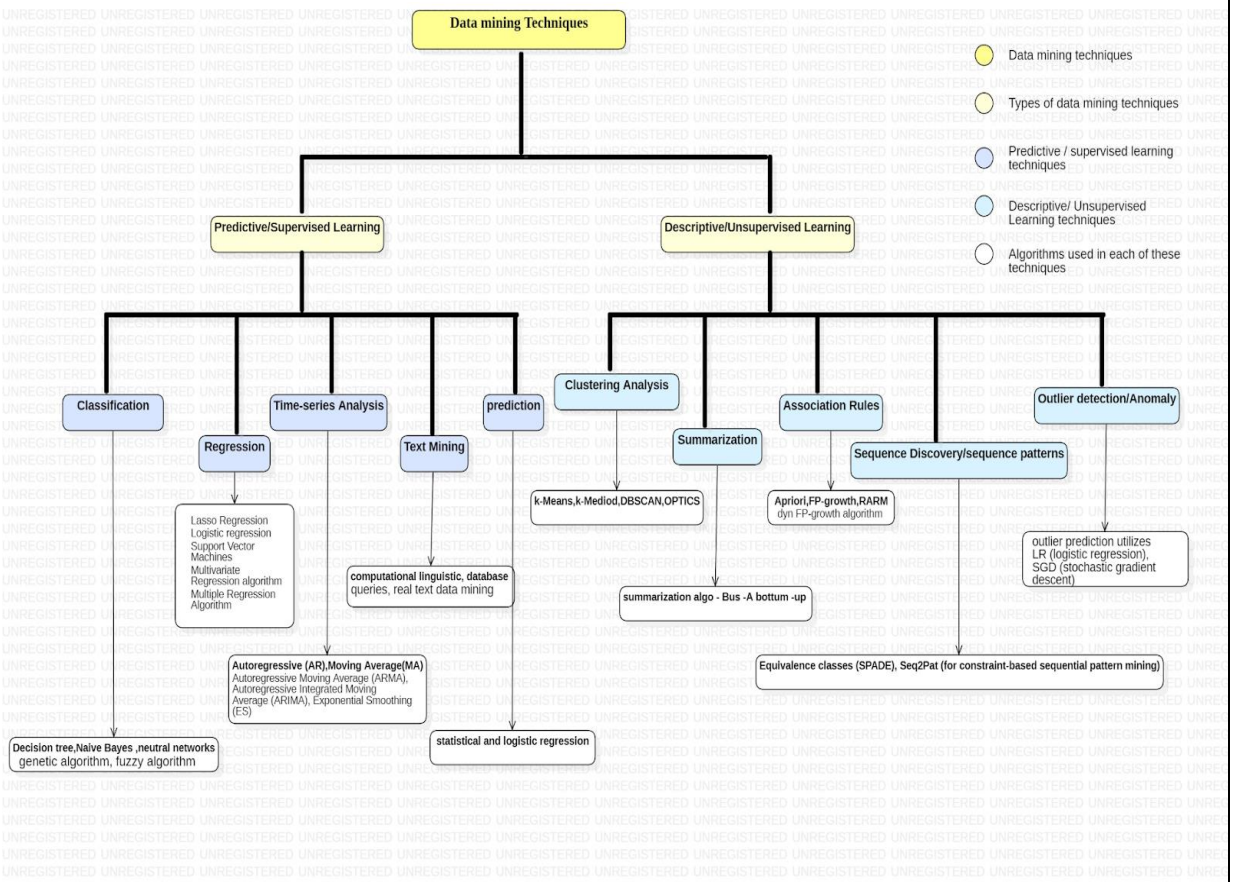
Presentation is key to gaining deeper insights about the data at hand and it allows one to make better decisions in order to connect with the audience.

When we look at data analysis, we're branching into a field that is an exploratory, fact-finding process that expects one to ask the right questions in order to make sense of answers that are usually hard to come across. Alternatively, we have data visualization where data is visually interpreted. It can consist of anything ranging from a single chart to a detailed dashboard, an infographic or a data story. The ultimate goal of data visualization is to pull down the time it takes for a certain audience to understand the data that is being put forward.

Acting as a method of structuring and ordering the collected data, data analysis ensures that data is being turned into information that a certain team or an organisation can use. By using systematic methods, analysis is performed to figure out trends, groupings, or other relationships between different types of data. Acting as a method used at putting data into a chart, graph, or other visual format that helps inform analysis and interpretation, we use data visualization to do the trick. By using this way of representation, one can present the analysed data in ways that not only engage different stakeholders but are majorly accessible as well. As evolution demands, to facilitate larger change processes and use data in an informed way, multiple visuals may be required at times.

2.1. Data Mining Techniques

The flowchart given below provides a detailed analysis of the various techniques and algorithms of data mining.



2.2. Data Mining Algorithms

In this paper, we'll be thoroughly examining one algorithm, i.e., K-Means Clustering.

K-means clustering

According to "Datanovia", "K-means clustering is regularly used unsupervised machine learning algorithms. It is used for distributing a given data set into a collection of k groups (i.e., k clusters), where k signifies the amount of groups that were specified by the analyst in advance. It analyses objects in various groups called clusters, if particularised things within the classified clusters are highly alike, then it is called high intra-class similarity, whereas if objects from different sets are mostly different it is known as low inter-class similarity. In k-means clustering, the centre signifies each cluster, which correlates to the mean of points assigned to the cluster.

The Elbow method could be a extremely talked-about technique, and also an approach is to run k-means clustering for a variety of clusters k, and for every value, we are calculating the aggregate of squared distances from each point to its assigned centre (distortions). In the K-means algorithm, the initial step when using k-means clustering is to point to the number of clusters (k) created within the phylogenetic extermination. The algorithm begins by randomly picking a few k objects from the data set to perform the initial centres for the clusters. The extracted objects are called cluster means or centroids. Next, to its closest centroid, the left objects are assigned, here the most immediate is determined using the Euclidean distance separating the item and the cluster mean. This action is termed the "cluster assignment step".

After the assignment step, the algorithm calculates the new mean of every cluster. Now that the centres are recalculated, every observation made is rechecked to operate if it would be more intimate to a particular cluster. All the objects are reassigned afresh using the updated cluster means. The cluster assignment and centroid update steps are iterated until the cluster assignments stop alternating. The clusters formed inside the present iteration are indistinguishable compared to those that were obtained within the prior iteration."

3. Business Governance using Data Analytics

3.1. Paradigm

To get a better understanding of the case study that will be presented, let us first take a look at an example, where data analytics is the key tool that is used to build a company sourced in the retail industry.

One of USA's biggest bargain retailers, Target Corporation is an organization that keeps the client data in Guest ID and keeps track of a broader scope of information like buy history, card use, survey reactions, support issues, email responses, site clicks, etc. Action information as such is then further enhanced by buying segment information like age, religion, schooling, conjugal status, youngsters' number, assessed pay, work history and critical life occasions, for example, the day you last moved or the chances of you having been separated or at any point backed out of all financial obligations.

Based on the client information that Target followed, it saw a buying pattern of those who were in various periods of pregnancy. For example, during the initial 20 weeks, pregnant ladies started buying calcium, magnesium and zinc supplements. In the subsequent trimester, they started purchasing bigger pants and larger quantities of sanitizers. From this collection of data, Target had the option to recognize pregnant clients despite the fact that their pregnancy status hasn't been disclosed to Target yet. Keeping in view, the client data as forementioned, Target started targeted item advancements to the explicit purchaser section. The monetary outcome was extremely surprising for them. It boosted its income rate from $44 billion to $67 billion after they began using this data analytics.

3.2. Exploratory Data Analysis – Retail Case Study Example

Exploratory Data Analysis (EDA) is the process of learning the structure of a dataset in order to discover patterns, to spot anomalies, to test hypotheses and to check assumptions through a series of numerical and graphical techniques. It is usually described by some experts as a method for "taking a peek" at data to understand it's value and its various applications. EDA is usually a precursor to other kinds of work that deals with statistics and data.

Objective:
To examine a company's performance within the retail industry and determine the weak areas over which profit can be made and derive potential business problems based on a sample dataset.

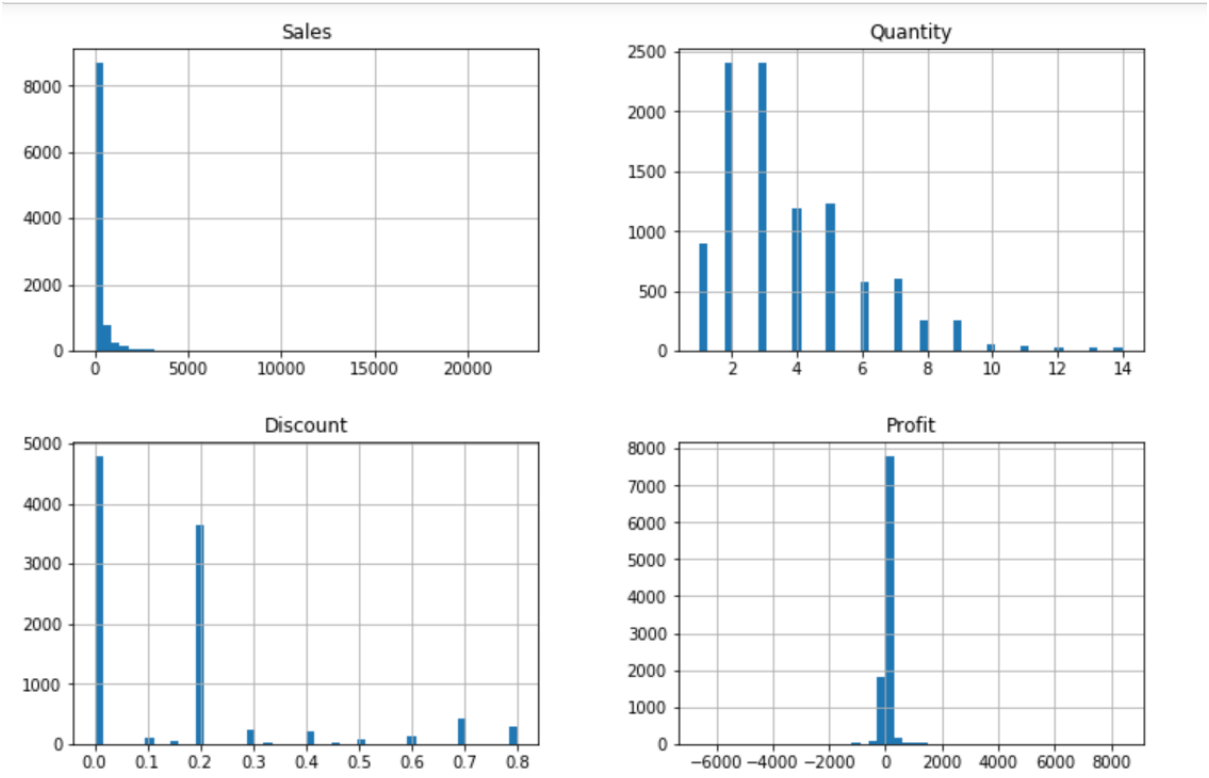
Pre-Requisites:
In this case study, we will explore a sample dataset click to open.
We will also be using Jupyter notebook. Follow the link for the Jupyter installation guide.

This dataset consists of data with customers in different countries, states who make purchases from a retail company based in the United States (US) that sells products within the categories: Furniture, Office supplies and technology.

Scenario:
The owner of a Superstore called GetMore wants to examine his company's performance. We are helping out the CMO of the company to enhance the company's campaign results. For the last few days, we have been working our way around data as a part of exploratory data analysis.

Once we explore the data given, we will find out the shape of our data. In this case, our dataset has 9994 rows × 13 columns which is considered huge.

Next, we will plot the data to get a clear visual representation.

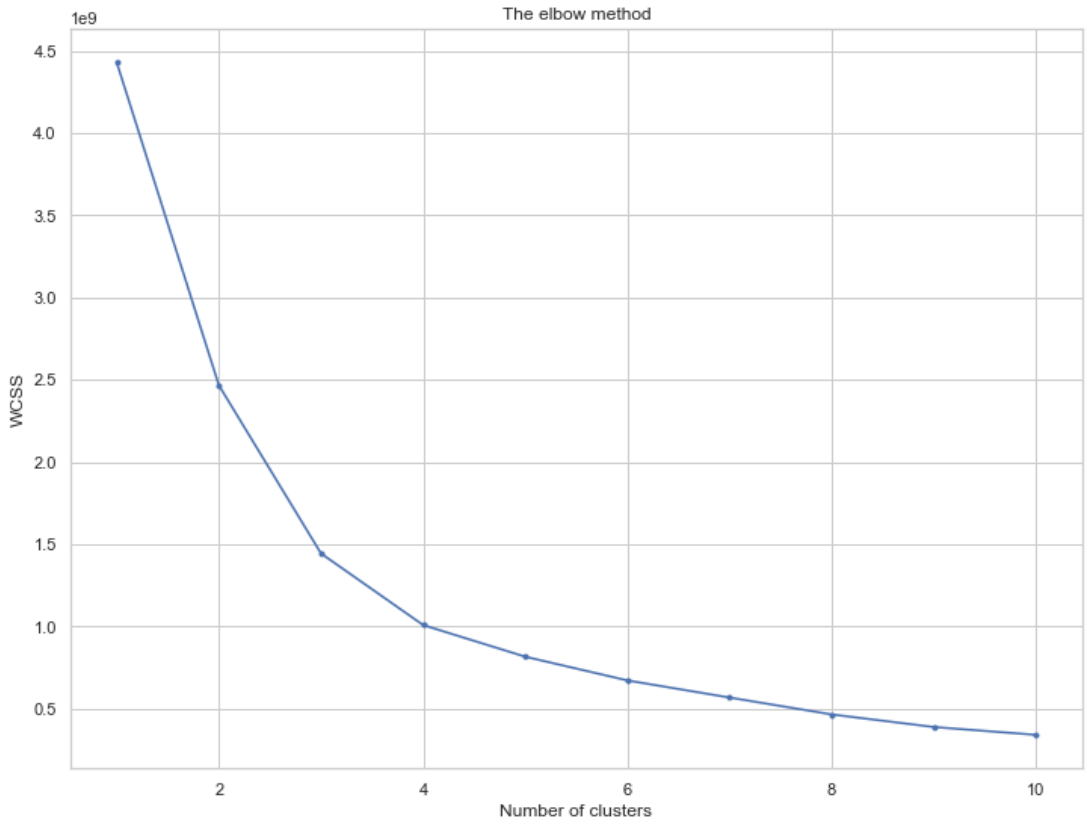


The following is "Elbow method" to find the ideal or optimal number of clusters in the data.

Source Code:

```
x = df.iloc[:, [ -1, -2, -3, -4]].values
wcss = [] #within sum of squares
for i in range (1, 11):
    kmeans = KMeans (n_clusters = i, init = 'k-means++',
                     max_iter = 300, n_init = 10, random_state = 0)
    kmeans.fit(x)
    wcss.append (kmeans.inertia_)


plt.plot(range(1, 11), wcss,marker='.')
plt.title('The elbow method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



Therefore, the optimal number of clusters in this case is 4.

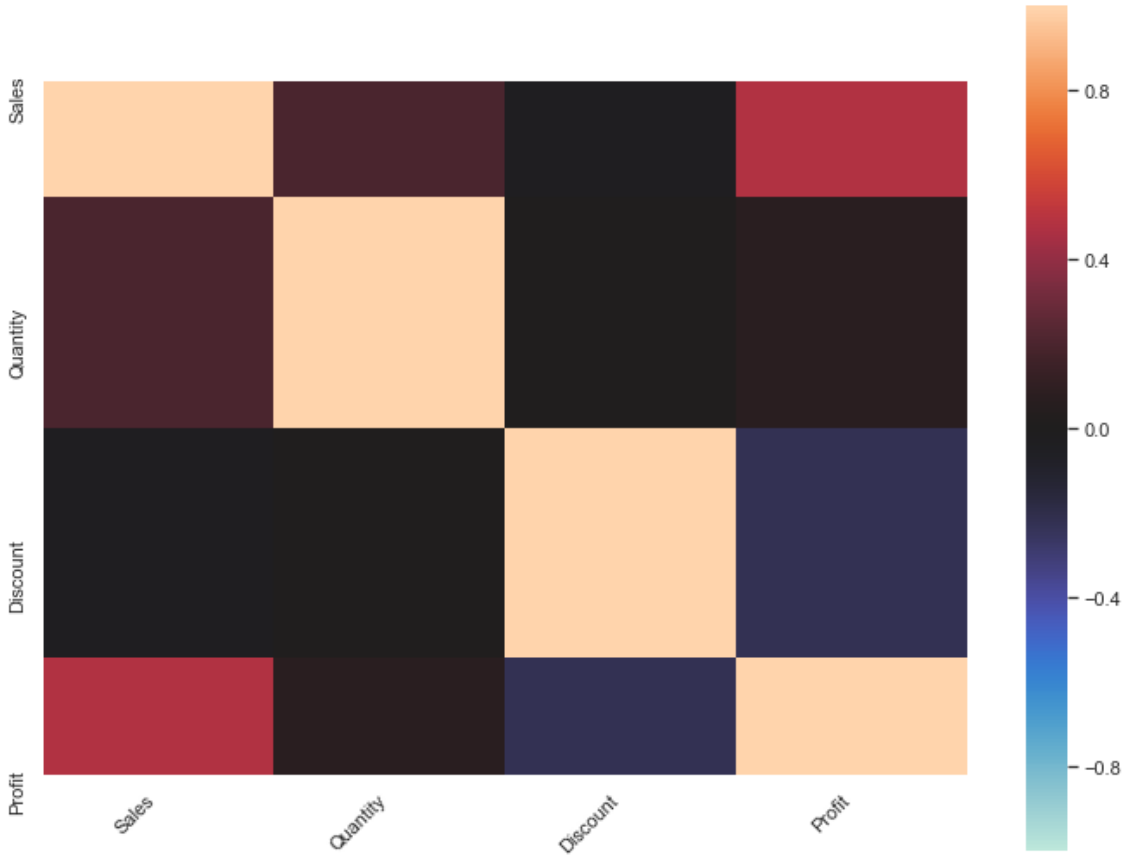
Heat Map:

Given below is a heatmap that makes visuals easier to analyse that any standard analytics report.

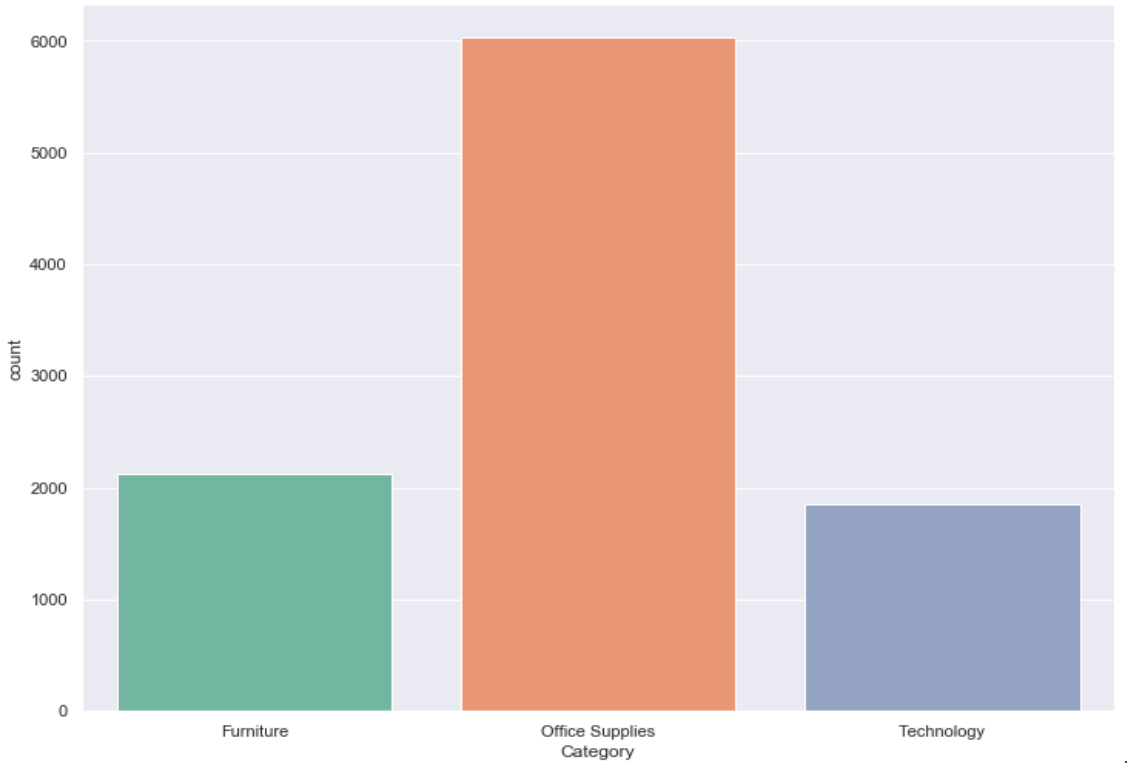
The darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour has been used.

The following is the two-dimensional plot of values which are mapped to the indices and columns of the chart.

[
Text(0.5, 0, 'Sales'),
Text(1.5, 0, 'Quantity'),
Text(2.5, 0, 'Discount'),
Text(3.5, 0, 'Profit')
]



The following distribution gives information about the difference in the count value of the categories – Furniture, Office Supplies and Technology:

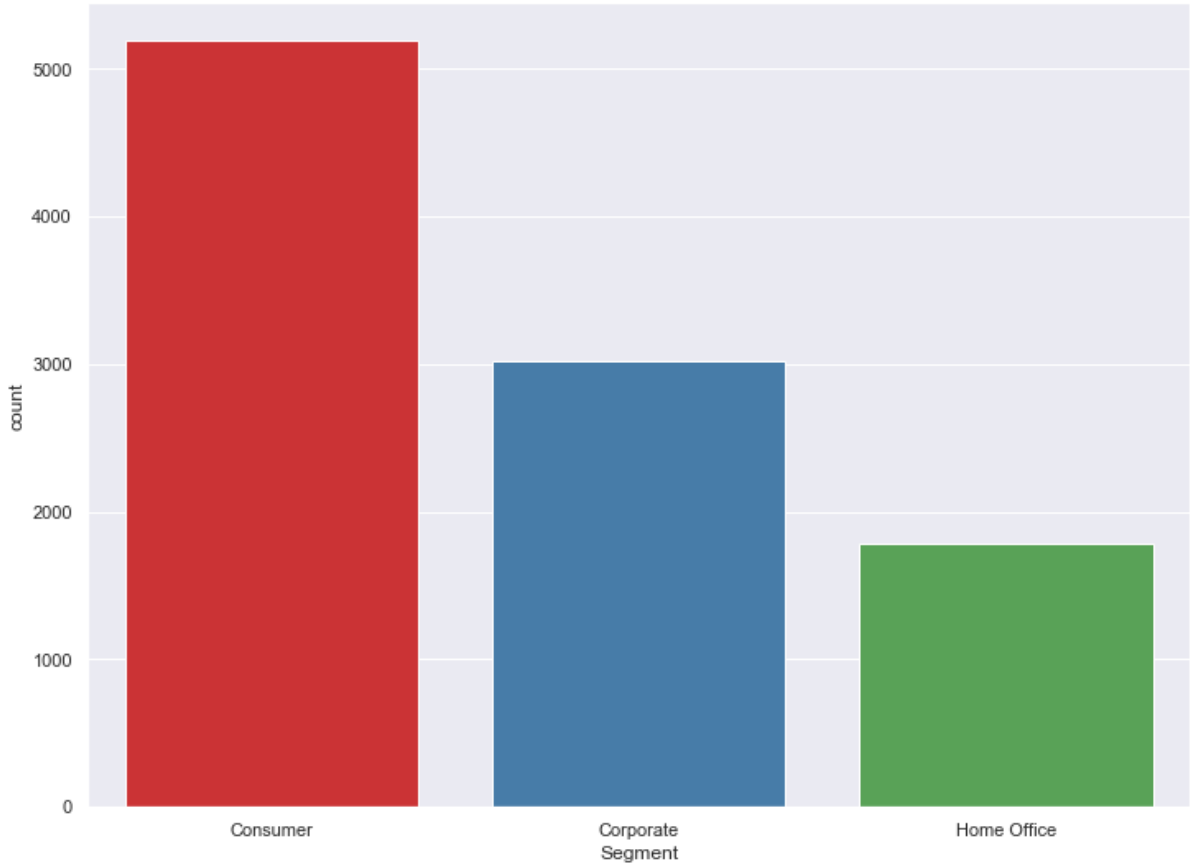


The above distribution gives the information about the difference in the count value of the categories (Furniture, Office Supplies and Technology) and the highest and lowest counts as well which is office supplies and technology respectively.

If the count value of the category is high it is expected for the sales to be high as well as the low sales can lead to loss.

This also says that more efforts should be put on the sales of office supplies as there is a scope to increase the sales value in this category.

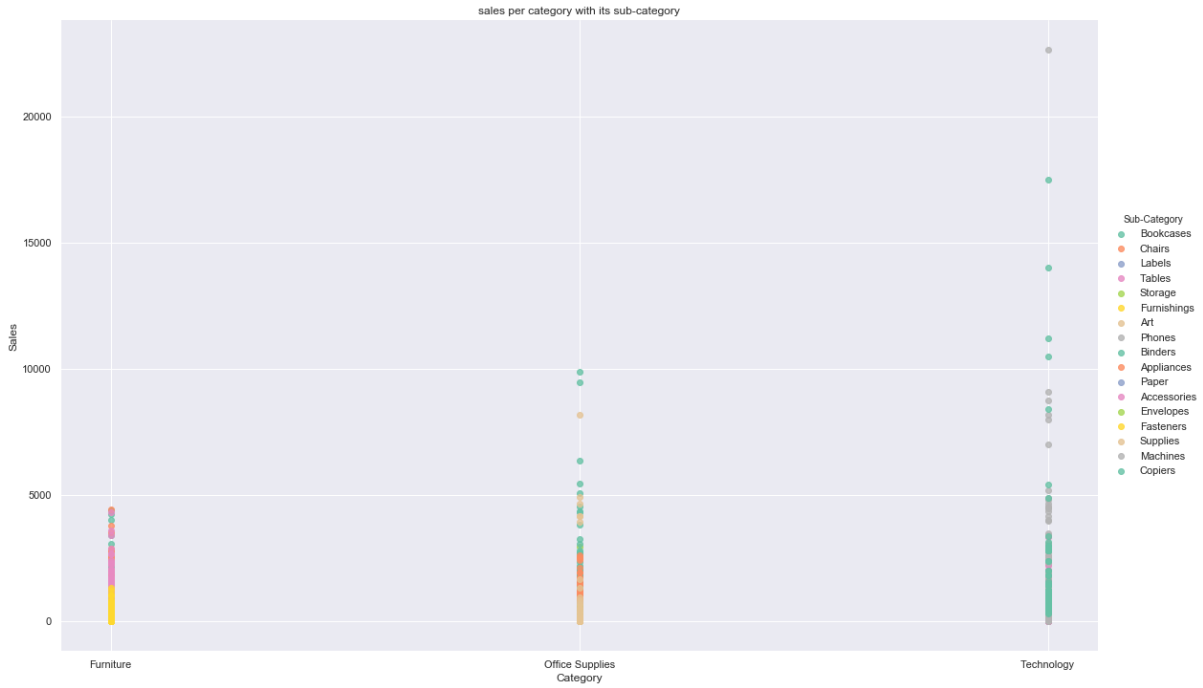
Similarly, now let's look at the "segment" which gives us the information about the kinds of people buying the products (consumer--regular customers, corporate segment, home office)



From the above distribution it is clear that consumers have the maximum count. Although, this doesn't necessarily mean that the company gains the maximum sales or profits mainly because of the consumers.

This tells us that the highest priority is to be given to the regular consumers as they are ones purchasing the highest number of goods.
But, this certainly does not mean that they are the source of profit. In the further analysis we will find out the exact sources of profits.

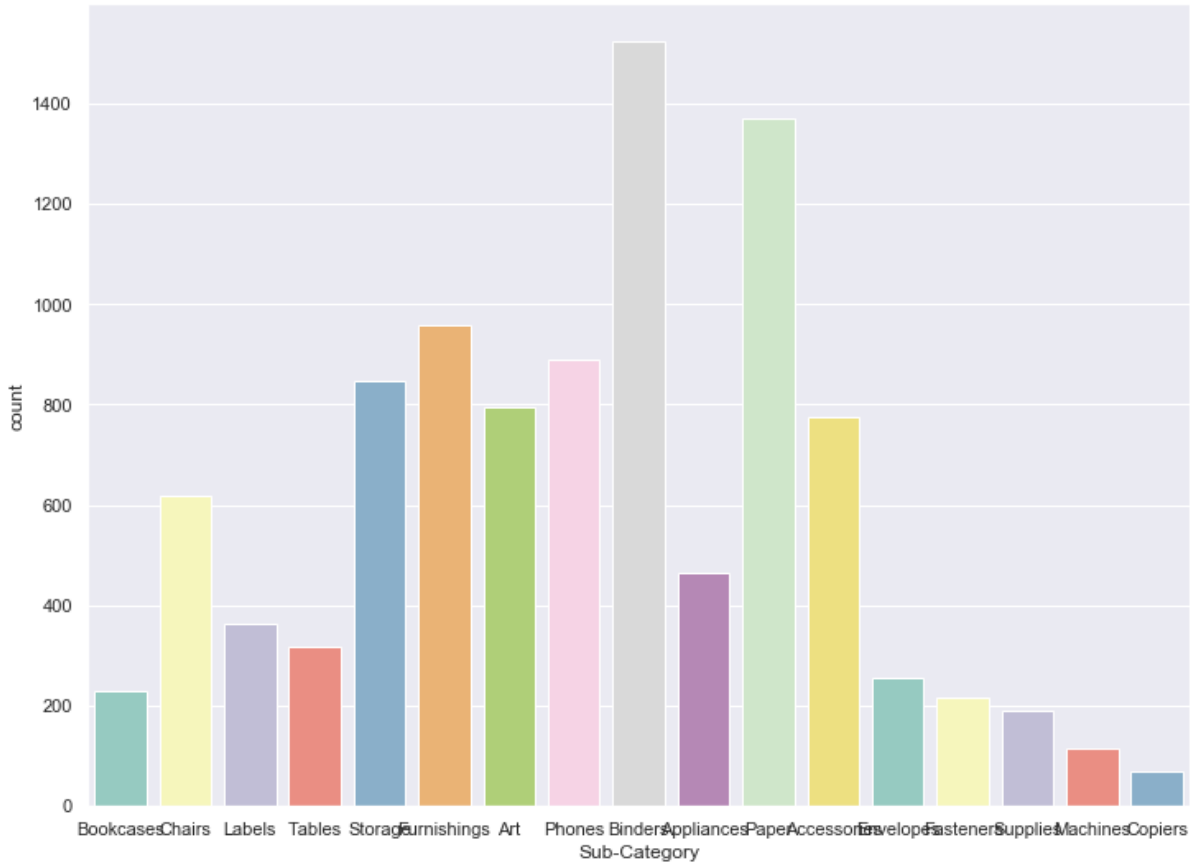
The following is one of the several interesting results and patterns that were noticed in the data. Upon analysing the distribution of number or count of sales across a number of product categories and sub-categories purchased by each customer, the following pattern was detected.



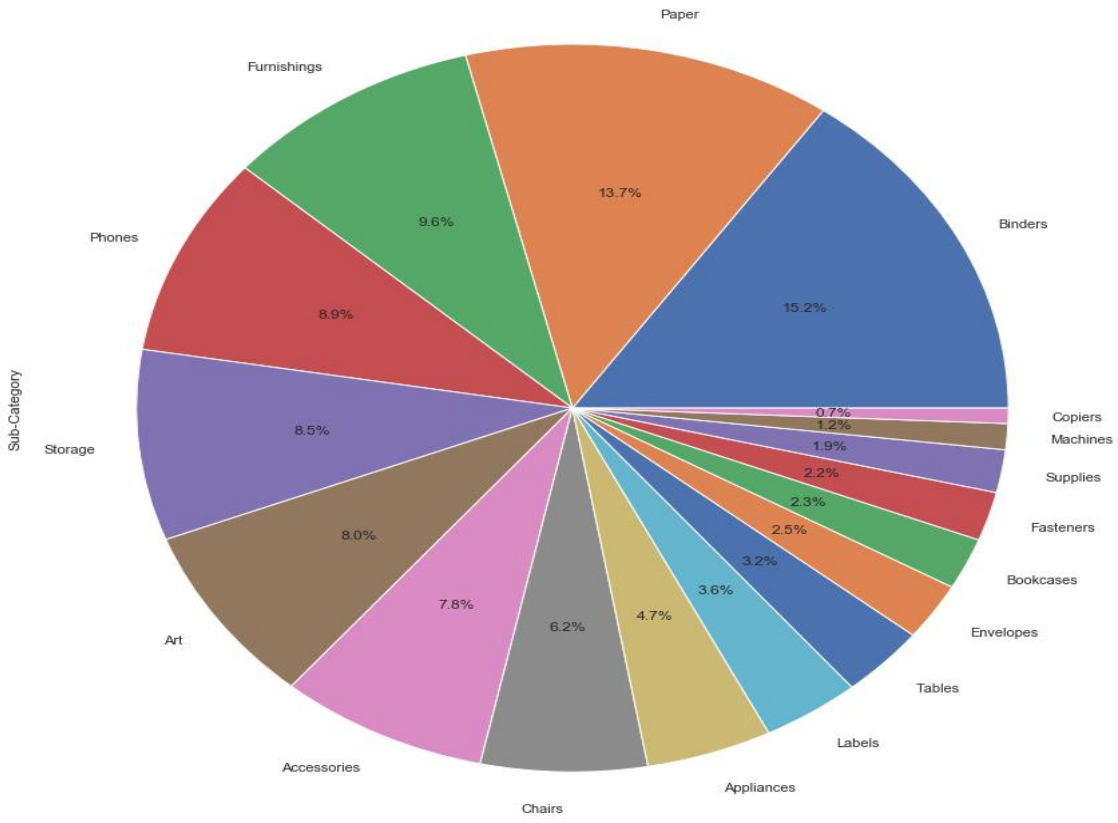
The above graph gives us the general idea about the product (subcategory) with the highest demand in each category.

In our case, furnishings, blinders and phones have the highest demand (Sales) in each of the category's furniture, office supplies and technology respectively.

Now, let's look at the details and get into the "Sub-Category" of our data, the subcategory of the data would give us the true insights and detailed information about the kind of products that are in demand, are currently a source of profit and see if there's any scope of increment in sales. This would also tell us about the products that are causing loss and help us predict why.



Below is the pie chart for visual representation of count of each subcategory.



From the graph, the following derivations have been made:

i. The category with largest sales values are Binders, Paper and Furnishing.

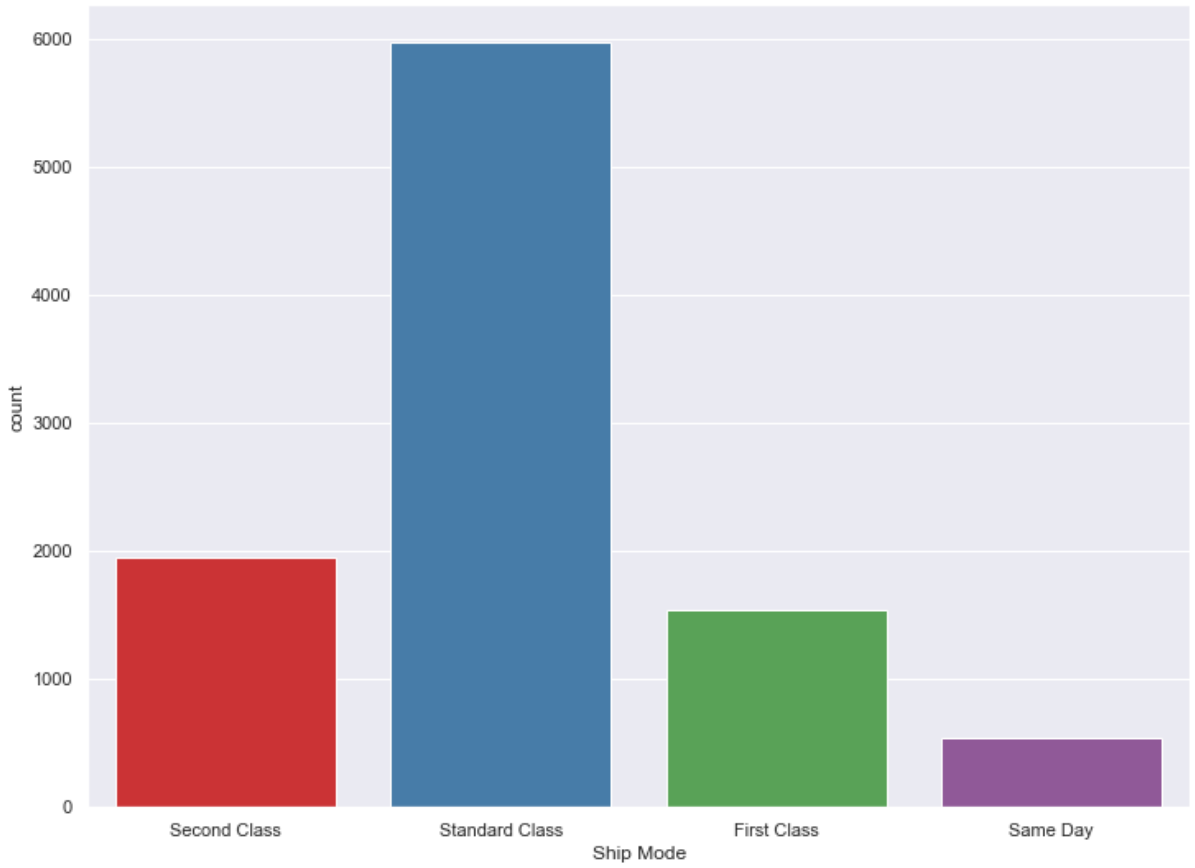
ii. The category with the least sales values are Copiers, Machines and Suppliers

iii. Furnishings have more sales value than the Tables and Chairs.

iv. Binders are sold more than Envelopes.

v. Art has less value than the furnishings

vi. We can say that the number of consumers is much higher than the producers as the sales value of machines and supplies is far less than the others.



From the graph, the following derivations have been made:

i. Most people opt for Standard Class shipping mode.

ii. The second highest method people opt for shipping is the Second-Class method.

iii. The least opted method is shipping on the Same Day.

iv. There is a huge difference between the 1st and 2nd opted shipping method where the difference in count is almost 4000.

v. First Class shipping method is opted a little less than Second Class with no huge difference.

vi. There is a possibility for the reason of the count of Standard Class shipping being far more than the other methods is that the cost is lesser compared to them.

To predict the reason, we should be having a clear knowledge about each of these shipment modes.

Standard shipping is a cost-effective option for delivery of a small to large standard package that doesn't include overnight or any other special requirements to deliver products quickly. Usually, standard shipping is done via couriers with orders possibly needing a day to process. Standard shipping time in the US generally takes anywhere between 3-5 business days.

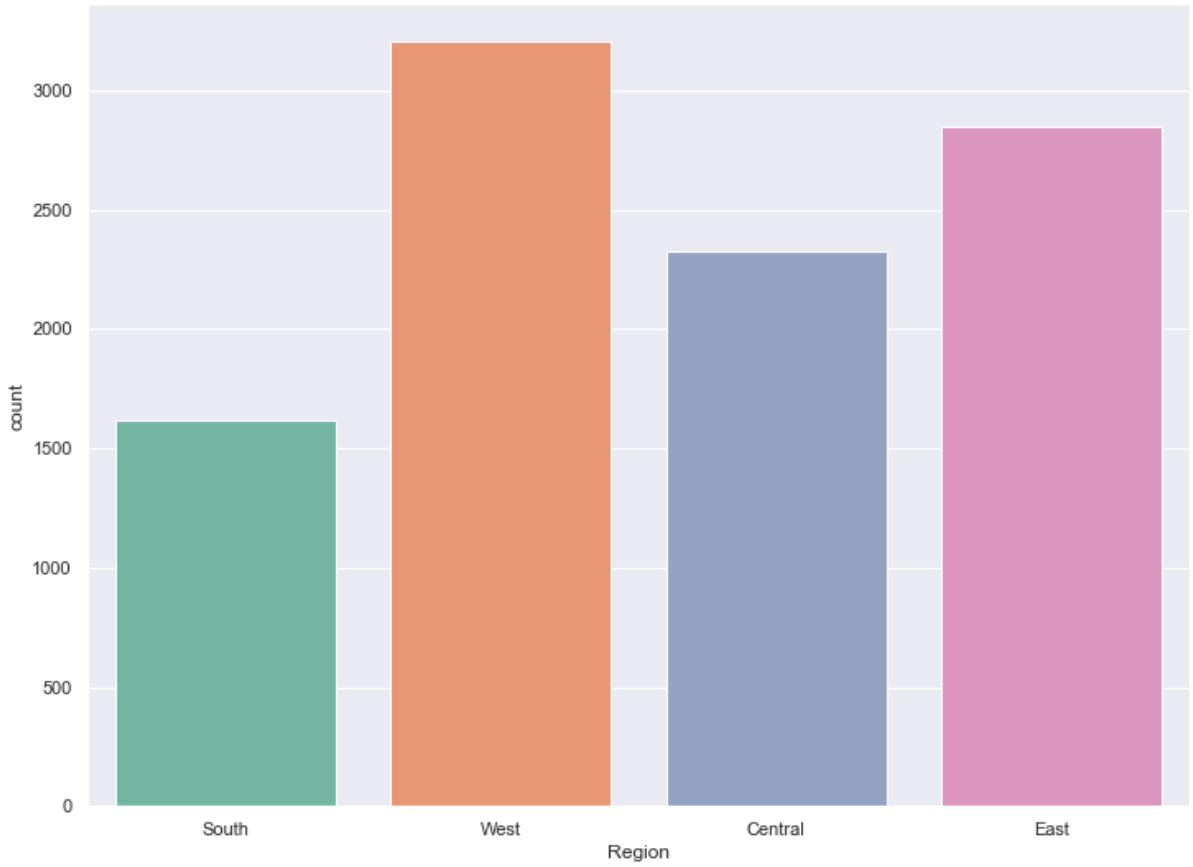
First-class should achieve delivery on the next working day. First Class is delivered faster by the USPS than Bulk or Standard. If the item is not so urgent, the consumers prefer to pay for the slower, second-class shipment. That should achieve delivery on the second or third working day after the purchase.

Shipment on the same day delivers the product on the same day.

As the first class delivers the product in one day this would be more expensive compared to second class which would take about 3-4 working days. The shipment on the same day would be the most expensive one as it demands for the arrangements and sorting's to be done in just hours.

The reason for most of the customers to prefer the standard class could be because of one of the findings we have made earlier as most of the customers are regular consumers and might not be in a hurry. So, they would prefer the cost-effective mode which is the standard-class shipment mode.

People opt for Same Day shipping only when there is a good reason for faster delivery so it is less.



From the graph, the following derivations have been made:

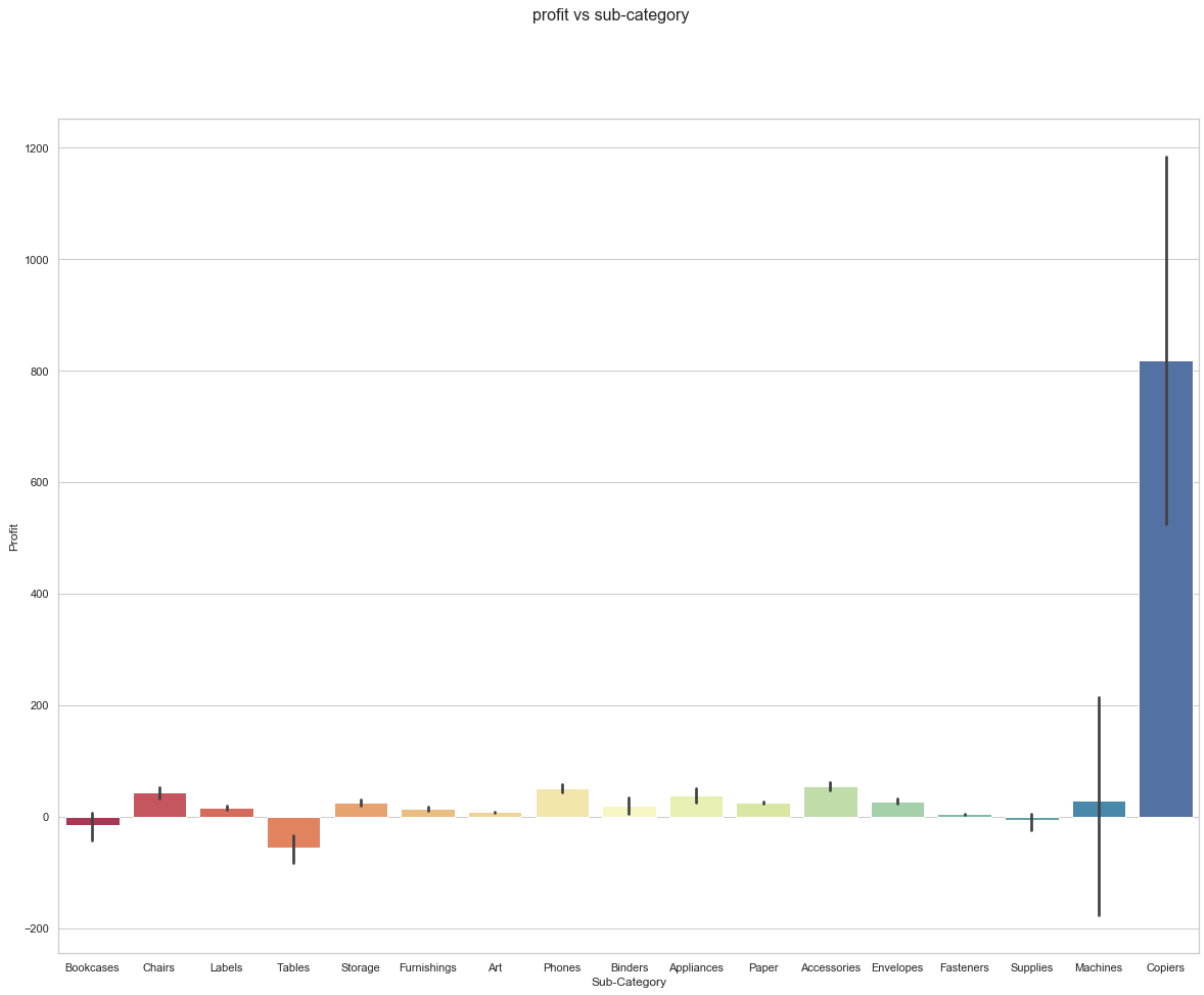
i. The region with the highest count value is West.

ii. The region with lowest count value is South.

iii. The region with the second highest count value is East with a difference in count of almost 400 lesser than West.

iv. The region with second lowest count value is Central with a difference in count of almost 600 more than South.

v. No region has count lesser than 1500.



From the graph, the following derivations have been made:

The losses are in sub category bookcases, tables, suppliers and highest amount of profit is gained through copiers with no loss, highest amount of loss through tables.

i. The highest profit is gained by the category of Copiers.

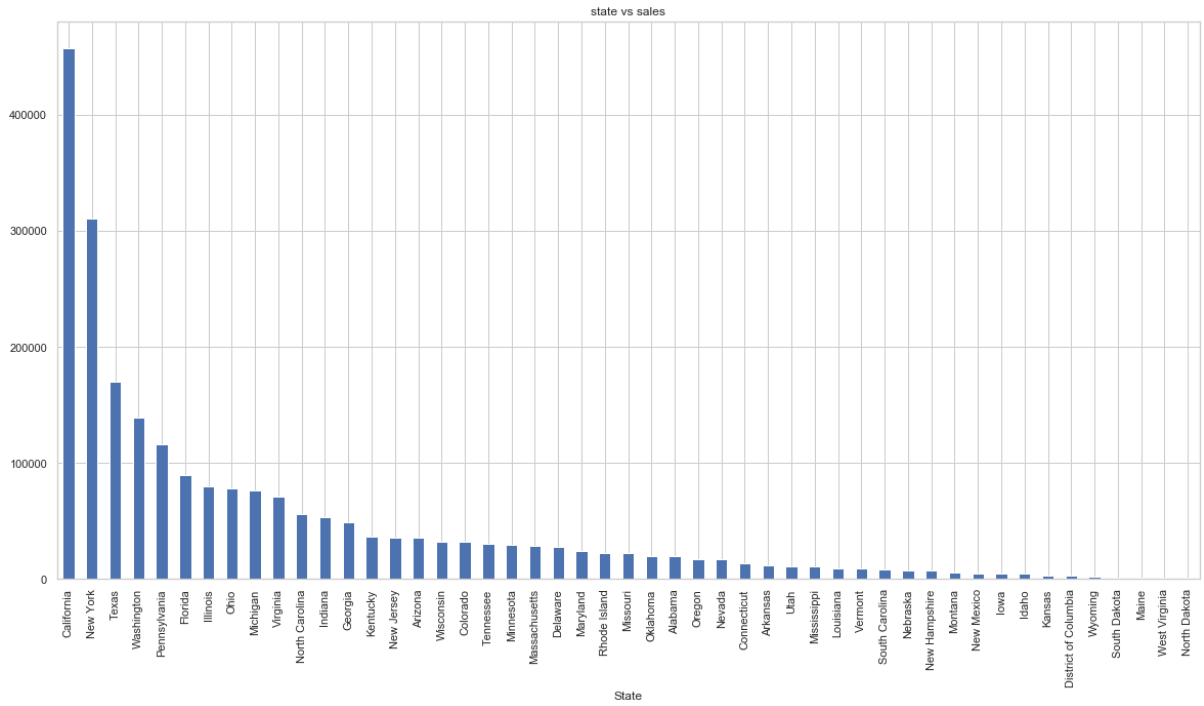
ii. The least profit is gained by the category of Fasteners.

iii. The highest loss is made from the category of Tables.

iv. The lowest loss is made from the category of Supplies.

v. There is a huge difference in profit gained by the highest profit-making category when compared to the second highest profit-making category which is Accessories.

vi. No category crosses the profit margin of 100 except Copiers.



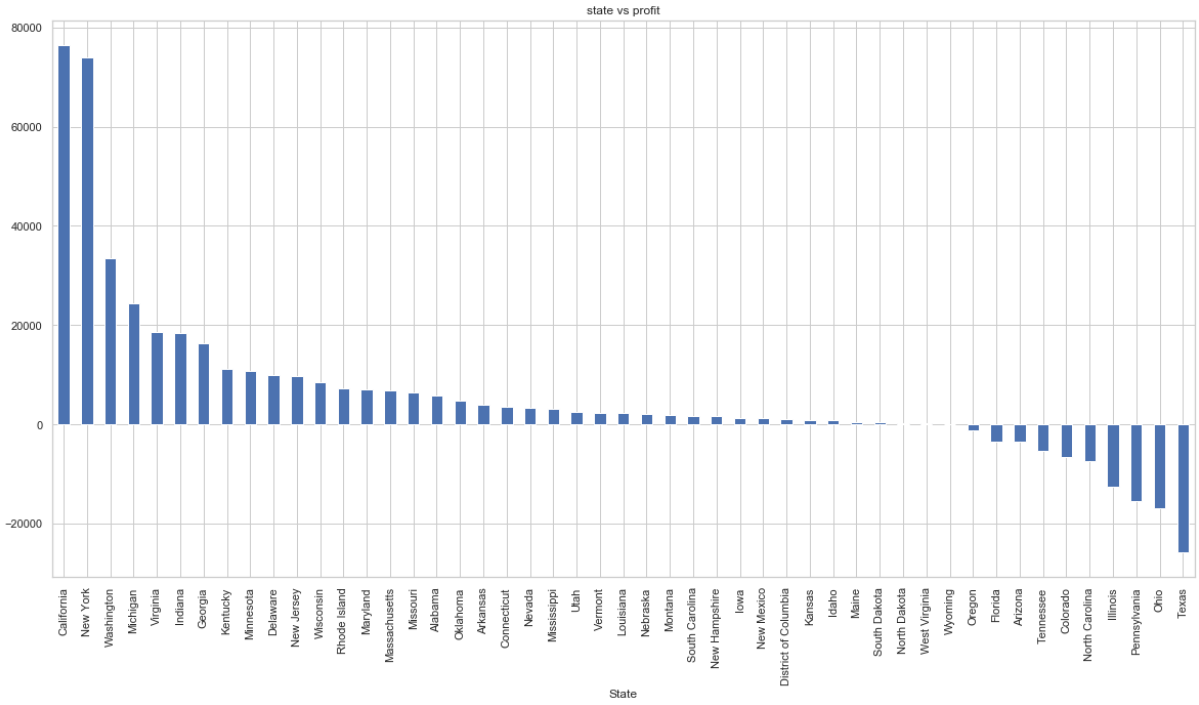
The above graph shows the relationship between states and sales in their respective states:

i. California state has the highest sales where binders have more sales.

ii. After California New York has the highest sales.

iii. States such as Kansas, District of Columbia and Wyoming can be observed to have very low sales.

iv. Further, we can also analyse that states of South Dakota, Maine, West Virginia and North Dakota have no sales.



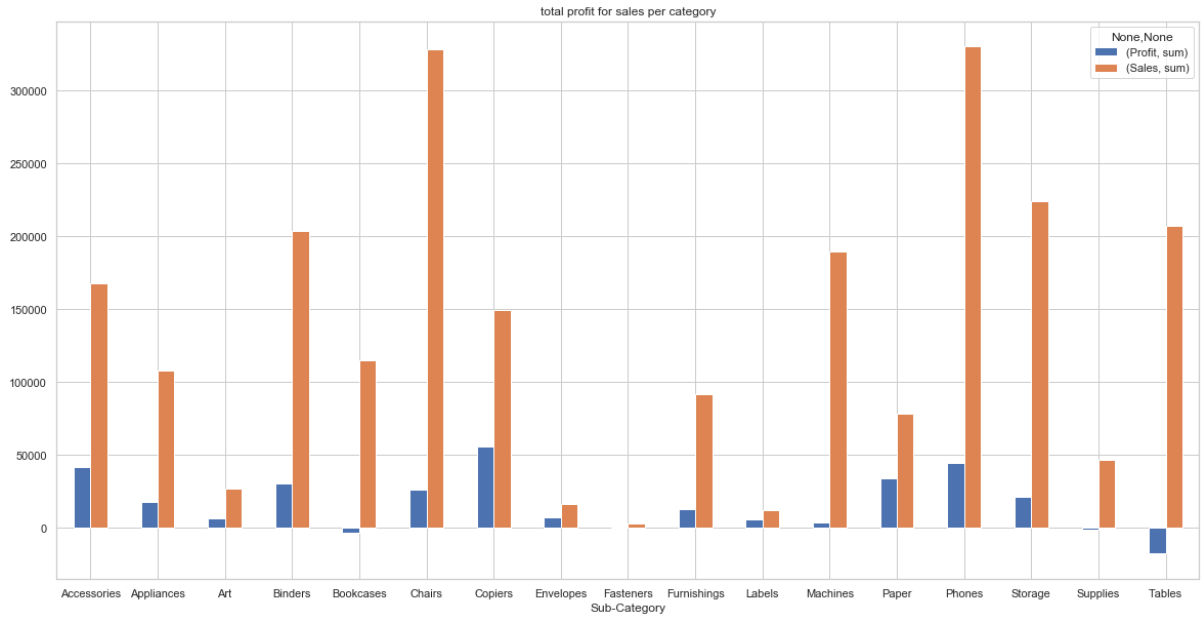
The above graph shows the relationship between the states and the profit of each states:

i. Highest profit acquired in California state which is in the west region.

ii. From the graph we can see that Texas which is in Central America is at the highest loss compared to others.

iii. Certain states like North Dakota and West Virginia does not have either profit or loss which shows that there have not been any sales in these states.

The graph below shows the relationship between total profit per sales per category. This gives us detailed information about the products(subcategories) that are making comparatively high profit but have low sales count and vice versa.



The above graph shows the relationship between total profit per sales per category:

i. Chairs and phones made the highest sales in comparison to other sub categories.

ii. Profits from copier sales are observed to be the highest even though sales where not on par as compared to chairs and phones.

iii. Tables, bookcases and supplies are in loss.

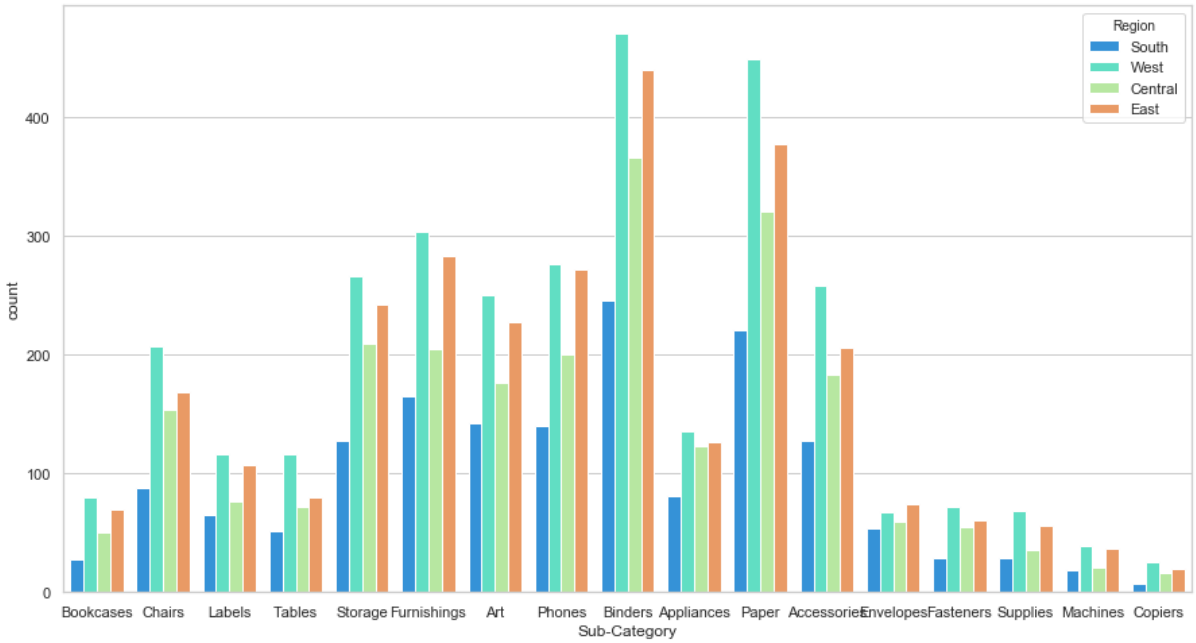
iv. We can further come to the conclusion that copiers made the highest profits and tables made a loss.

Tables, bookcases and supplies are in loss. This could be something to look into as tables clearly have a high sales count compared to more than half of the subcategories present but are still in loss.

The other categories which were the sources of loss are bookcases and supplies. We can try to improve the sales count of the supplies to stop the loss.

The sales count of phones and chairs is quite large which means that these products are in demand, but the profit made by them is very little, in order to make more profits the cost of the products could be increased.

From the previous analysis the count value of the blinders is highest and hence the sales could be increased.



The above graph shows the relationship between count value and the Sub - category products purchased:

i. In all the regions, the same pattern is observed where binders and papers have the highest count and least count is in copiers and machines.

ii. Additionally, we can further analyse that the highest peak can be seen in the west region for the count value of binders.

iii. Whereas copiers made least count value which less than 50 in all regions.

From the knowledge acquired earlier that states of South Dakota(central), Maine(east), West Virginia(east) and North Dakota(central) have zero sales looking into this, South Dakota and North Dakota belong to the central region and the products that have high demand in this region are papers, blinders followed by phones and accessories therefore to increase sales in these two states we can spread the word and let more people in corporate segment and regular consumers know about the company.

Then we could increase the supply of papers, blinders, phones and accessories to these states. Maine, West Virginia are states in the eastern region of the US. According to the patterns observed the products with high demand are blinders, papers followed by furnishings and phones which means that the purchases made are mostly by the corporate segment and home office.

4. Conclusion

Exploratory data analysis is an incredible asset. An EDA is to put your high-level business examination the correct way. EDA gives an incredible chance to test your basic business speculations and hunches prior to hopping into a thorough model structure.

We have seen that the sales in states like California are more even though profits from New York are on par with California although sales are low. Which leads to further analysis that if sales in New York can be increased we can get more profits.

As the products that are supplied to the regular customers are making more sales but if we further increase the price to get more profit, the demand of the product may decrease which might lead to decrease in sales as well.

On the whole, as most of the profits are from the corporate segment, we can offer volume pricing and bulk ordering to target these sectors to get more profits.

5. References

I. https://pipol.com/data-analysis-and-data-visualization/

II. https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/#:~:text=K%2Dmeans%20clustering%20(MacQueen%201967,pre%2Dspecified%20by%20the%20analyst.

III. https://www.talend.com/resources/data-mining-techniques/

IV. http://ucanalytics.com/blogs/exploratory-data-analysis-retail-case-study-example-part-3/

V. https://blog.meghdut.io/interesting-case-studies-of-data-analytics-in-retail-industry/