AIT - 664
Individual Class Project
Wholesale Market Descriptive Analysis

Introduction

The market for the sale of goods to a retailer is known as a wholesale market. A wholesaler receives huge quantities of goods and products from the manufacturing unit and it further, sells them off to the retail stores. The consumers buy these products from the retail stores. The price at which the wholesaler buys the goods from the manufacturer is quite less as huge stocks are bought in one go. The price is much lesser than the case in which individual retail stores buy them from the manufacturer.

Initially, wholesalers used to be closer to the markets to which they supplied, rather than the source through which they got the products. In today's scenario with the emergence of the internet and e-procurement, most of the retailers have started locating near the manufacturers. There are several types of channels which are being used to send these products.

Goal

The main goal of this project is to analyze the wholesale market in certain states of the US. It is very important to analyze the current trend, in the demand of the customers, in order to avoid heavy losses. The supply of products should be as per the demand of the market. Both of these should go hand in hand in order to obtain profits.

Columns and their data types:

Column Name	Description	Type
Location	US State	Text
Year	Year of the record.	Text
Channel	Customer's channel: Horeca - 1, Retail - 2	Text
Region	Customer's region: Lisnon - 1, Oporto - 2, Other - 3	Text
Fresh	annual spending (m.u.) on fresh products	Floating-point number
Milk	annual spending (m.u.) on milk products	Integer
Grocery	annual spending (m.u.) on grocery products	Floating-point number
Frozen	annual spending (m.u.) on frozen products	Integer
Detergents_Paper	annual spending (m.u.) on detergents and paper products	Floating-point number
Delicassen	annual spending (m.u.) on and delicatessen products.	Integer

Milestones

1. Data Acquisition

The data was acquired from the website https://archive.ics.uci.edu/ml/datasets/Wholesale+customers in the form of a csv file. The file then was introduced to python programming language.

I used this language to obtain the data from a csv file and turned it into a useable form by performing a particular cleaning technique. There are certain libraries that I used, to perform different kinds of functionality, for which I have used the python script.

Code:

```
##Final Project - AIT 664


##Importing the libraries and Dependencies
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import csv
from pandas import read_csv
import os

##Work directory setup
os.getcwd()
os.chdir('E:\GMU SEM 2\AIT 664\AIT - 664 Final Project')
os.getcwd()

##File reading using pandas
data = pd.read_csv("Wholesale customers data.csv")
```

Dataset read:

	A	B	C	D	E	F	G	H	I	J
1	LOCATION	YEAR	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
2	AL	1971	2	3	12669	9656	7561	214	2674	1338
3	AL	1972	2	3	7057	9810	9568	1762	3293	1776
4	AL	1973	2	3	6353	8808	7684	2405	3516	7844
5	AL	1974	1	3	13265	1196	4221	6404	507	1788
6	AL	1975	2	3	22615	5410	7198	3915	1777	5185
7	AL	1976	2	3	9413	8259	5126	666	1795	1451
8	AL	1977	2	3	12126	3199	6975	480	3140	545
9	AL	1978	2	3	7579	4956	9426	1669	3321	2566
10	AL	1979	1	3	5963	3648	6192	425	1716	750
11	AL	1980	2	3	6006	11093	18881	1159	7425	2098
12	AL	1981	2	3	3366	5403	12974	4400	5977	1744
13	AL	1982	2	3	13146	1124	4523	1420	549	497
14	AL	1983	2	3	31714	12319	11757	287	3881	2931
15	AL	1984	2	3	21217	6208	14982	3095	6707	602
16	AL	1985	2	3	24653	9465	12091	294	5058	2168
17	AL	1986	1	3	10253	1114	3821	397	964	412
18	AL	1987	2	3	1020	8816	12121	134	4508	1080
19	AL	1988	1	3	5876	6157	2933	839	370	4478
20	AL	1989	2	3	18601	6327	10099	2205	2767	3181
21	AL	1990	1	3	7780	2495	9464	669	2518	501
22	AL	1991	2	3	17546	4519	4602	1066	2259	2124

2. Data Preprocessing

The data contained certain missing values in few of the columns. I checked their count with the help of a predefined function. Later, I used the mean of that column and filled in the missing values with this mean value of that particular column. I converted the data type of the numeric columns to integer data type, in order to accommodate the mean value, which was a float type. I changed the data type of some fixed columns to string data type.

Code:

```
##Calculation of the number of missing values in the dataset
print("Number of null values in each column")
print(data.isnull().sum())

##Inserting NA in the missing value space.
data.fillna("NA", inplace=True)

##Conversion of the data type of the columns from string to numeric and filling
##in all the NA values with the mean value of that the column. Conversion
##of certain columns to different data types as well.
print(data)
print(data.dtypes)

data["Channel"] = data["Channel"].apply(str)
data["Region"] = data["Region"].apply(str)
data["YEAR"] = data["YEAR"].apply(str)

data["Fresh"] = pd.to_numeric(data["Fresh"], errors='coerce')
data["Fresh"] = data["Fresh"].fillna(int(data["Fresh"].mean()), downcast='infer').astype(int)

data["Milk"] = pd.to_numeric(data["Milk"], errors='coerce')
data["Milk"] = data["Milk"].fillna(int(data["Milk"].mean()), downcast='infer').astype(int)

data["Frozen"] = pd.to_numeric(data["Frozen"], errors='coerce')
data["Frozen"] = data["Frozen"].fillna(int(data["Frozen"].mean()), downcast='infer').astype(int)

data["Delicassen"] = pd.to_numeric(data["Delicassen"], errors='coerce')
data["Delicassen"] = data["Delicassen"].fillna(int(data["Delicassen"].mean()), downcast='infer').astype(int)

data["Grocery"] = pd.to_numeric(data["Grocery"], errors='coerce')
data["Grocery"] = data["Grocery"].fillna(int(data["Grocery"].mean()), downcast='infer').astype(int)

data["Detergents_Paper"] = pd.to_numeric(data["Detergents_Paper"], errors='coerce')
data["Detergents_Paper"] = data["Detergents_Paper"].fillna(int(data["Detergents_Paper"].mean()), downcast='infer').astype(int)

####Printing the data types of the column to verify if the data type has been converted
##successfully.
print(data.dtypes)

####Check to verify if all the missing values have been filled up.
print("After replacing null values with the mean value of each column")
print(data.isnull().sum())
```

Output:

```
Number of null values in each column
LOCATION             0
YEAR                 0
Channel              0
Region               0
Fresh               10
Milk                 9
Grocery              8
Frozen               8
Detergents_Paper     9
Delicassen           8
dtype: int64
```

Data types before conversion:

```
LOCATION            object
YEAR                 int64
Channel              int64
Region               int64
Fresh               object
Milk                 int64
Grocery             object
Frozen               int64
Detergents_Paper    object
Delicassen           int64
dtype: object
```

Data types after conversion:

```
LOCATION            object
YEAR                object
Channel             object
Region              object
Fresh                int64
Milk                 int64
Grocery              int64
Frozen               int64
Detergents_Paper     int64
Delicassen           int64
dtype: object
After replacing null values with the mean value of each column
LOCATION            0
YEAR                0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicassen          0
dtype: int64
```

3. Mining Tool Preparation

A summary of the dataset is displayed by using the summary function.

Code:

```
####Printing of the statistics of the data set
summary = data.describe()
print("Summary Statistics")
print(summary)
```

Output:

```
Summary Statistics
              Fresh          Milk       Grocery        Frozen  Detergents_Paper     Delicassen
count    440.000000    440.000000    440.000000    440.000000        440.000000    440.000000
mean   12027.626424   5796.265909   7968.881549   3071.931818       2887.845103   1524.870455
std    12634.330614   7380.377175   9495.985632   4854.673333       4765.992384   2820.105937
min        3.000000     55.000000      3.000000     25.000000          3.000000      3.000000
25%     3151.750000   1533.000000   2156.500000    742.250000        261.500000    408.250000
50%     8549.000000   3627.000000   4785.500000   1526.000000        820.500000    965.500000
75%    16933.750000   7190.250000  10655.750000   3554.250000       3922.000000   1820.250000
max   112151.000000  73498.000000  92780.000000  60869.000000      40827.000000  47943.000000
```

The file containing all the changes that have been made on the dataset, i.e. the clean data set, is exported to a comma separated file.

Code:

```
####Redirecting the clean data set to a csv file.
data.to_csv('Wholesale Customer Data.csv', index=False, encoding='utf-8')
```

This data was then introduced to R where the analysis was done, and the visualizations were created further.

4. Clustering Analysis

After obtaining the summary of the dataset, there is a noticeable difference between the minimum and maximum values in all the columns of the dataset. The maximum value of milk varies from 55 to 73498. The top customers would make this value to be really high, in comparison to the lower ones. However, from a business point of view, it is not really required to analyze the expenditure of the top customers.

They can basically be considered as outliers. They cannot be removed by doing any kind of normalization, but a log transformation may prove to be useful. It is not required that the algorithm for clustering requires the expenditure made by the top customers. The middle ones require clustering and segmentation.

```
> setwd("E:/GMU SEM 2/AIT 664/AIT - 664 Final Project")
> data <-read.csv("Wholesale customers data.csv",header=T)
> summary(data)
    Channel          Region          Fresh             Milk          Grocery          Frozen
 Min.   :1.000   Min.   :1.000   Min.   :     3   Min.   :   55   Min.   :    3   Min.   :   25.0
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153   1st Qu.:  742.2
 Median :1.000   Median :3.000   Median :  8504   Median : 3627   Median : 4756   Median : 1526.0
 Mean   :1.323   Mean   :2.543   Mean   : 12000   Mean   : 5796   Mean   : 7951   Mean   : 3071.9
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656   3rd Qu.: 3554.2
 Max.   :2.000   Max.   :3.000   Max.   :112151   Max.   :73498   Max.   :92780   Max.   :60869.0
 Detergents_Paper    Delicassen
 Min.   :    3.0   Min.   :    3.0
 1st Qu.:  256.8   1st Qu.:  408.2
 Median :  816.5   Median :  965.5
 Mean   : 2881.5   Mean   : 1524.9
 3rd Qu.: 3922.0   3rd Qu.: 1820.2
 Max.   :40827.0   Max.   :47943.0
```

I removed the top 5 entries in the code from all the columns and created a new data set.

```
> top.n.custs <- function (data,cols,n=5) { #Requires some data frame and the top N to remove
+   idx.to.remove <-integer(0) #Initialize a vector to hold customers being removed
+   for (c in cols){ # For every column in the data we passed to this function
+     col.order <-order(data[,c],decreasing=T) #Sort column "c" in descending order (bigger on top)
+     #Order returns the sorted index (e.g. row 15, 3, 7, 1, ...) rather than the actual values sorted.
+     idx <-head(col.order, n) #Take the first n of the sorted column C to
+     idx.to.remove <-union(idx.to.remove,idx) #Combine and de-duplicate the row ids that need to be remo
ved
+   }
+   return(idx.to.remove) #Return the indexes of customers to be removed
+ }
```

Further, I checked for the number of customers that have been removed.

```
> top.custs <-top.n.custs(data,cols=3:8,n=5)
> length(top.custs) #How Many Customers to be Removed?
[1] 19
```

Using the dataset that has been obtained from the above code, cluster analysis can be performed on it. The first four columns of the dataset, Location, Year, Channel and Region can be excluded as they cannot contribute in the clustering process.

```
data[top.custs,] #Examine the customers
data.rm.top <-data[-c(top.custs),] #Remove the Customers
set.seed(76964057) #Set the seed for reproducibility
k <-kmeans(data.rm.top[,-c(1,2,3,4)], centers=5) #Create 5 clusters, Remove columns 1,2,3 and 4
k$centers #Display cluster centers
```

Output:

```
> data.rm.top <-data[-c(top.custs),] #Remove the Customers
> set.seed(76964057) #Set the seed for reproducibility
> k <-kmeans(data.rm.top[,-c(1,2)], centers=5) #Create 5 clusters, Remove columns 1,2,3 and 4
> k$centers #Display cluster centers
      Fresh      Milk   Grocery   Frozen Detergents_Paper Delicassen
1  4189.747  7645.639 11015.277 1335.145        4750.4819  1387.1205
2 16470.870  3026.491  4264.741 3217.306         996.5556  1319.7593
3 33120.163  4896.977  5579.860 3823.372         945.4651  1620.1860
4  5830.214 15295.048 23449.167 1936.452       10361.6429  1912.7381
5  5043.434  2329.683  2786.138 2689.814         652.8276   849.8414
```

Interpretation:

Cluster 3 – Contains most of the fresh category products.
Cluster 1 – It appears to have grocery at quite a high level, fresh food at a lower level and detergents paper at a level little more than the medium one.
Cluster 5 – It mostly depicts the small customers and can be considered to be a small store of all different kinds of products.

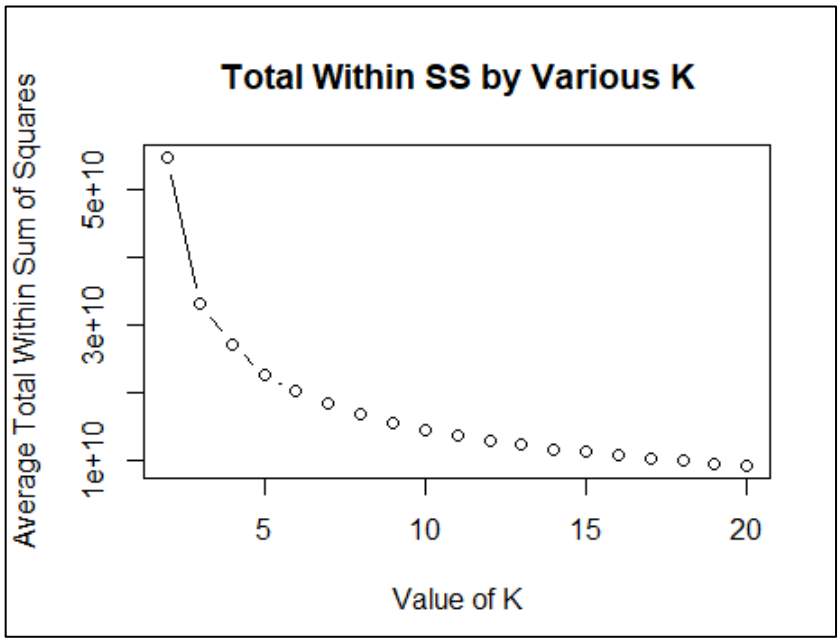
Withness and Betweenss:

Betweenss – It is the sum obtained of the squared distance between the centers of the clusters. A perfect scenario is where the centers are quite far away from each other.

Withness – It can be termed as the sum of the square of the distance between data points and the center of the clusters. A lesser value is considered to be good. There is either a requirement of creating more clusters and also the presence of several outliers.

Finding the appropriate number of clusters:

Betweenss and withness can be plotted into a graph in order to compare them. Their comparison gives the elbow point. An elbow point signifies the optimum number of clusters that should be created for a particular dataset. The point where the graph bends and does not make any more progress in withness, is the K point. I have set the range between 2 and 20 and the graph, in order to trace the elbow point.



The plot that has been obtained does not contain a distinct elbow point. As per the graph, 5 seems to be the K point for this dataset.

Now we have 5 different clusters. Count of the data points in each cluster is also obtained.

```
> table(k$cluster) #Give a count of data points in each cluster

  1   2   3   4   5
 83 108  43  42 145
```

Further analysis demands, setting up vectors for the analysis and plotting.

```
> rng<-2:20 #K from 2 to 20
> tries<-100 #Run the K Means algorithm 100 times
> avg.totw.ss<-integer(length(rng)) #Set up an empty vector to hold all of points
> for(v in rng){ # For each value of the range variable
+   v.totw.ss<-integer(tries) #Set up an empty vector to hold the 100 tries
+   for(i in 1:tries){
+     k.temp<-kmeans(data.rm.top,centers=v) #Run kmeans
+     v.totw.ss[i]<-k.temp$tot.withinss#Store the total withinss
+   }
+   avg.totw.ss[v-1]<-mean(v.totw.ss) #Average the 100 total withinss
+ }
> plot(rng,avg.totw.ss,type="b", main="Total Within SS by Various K",
+      ylab="Average Total Within Sum of Squares",
+      xlab="Value of K")
```

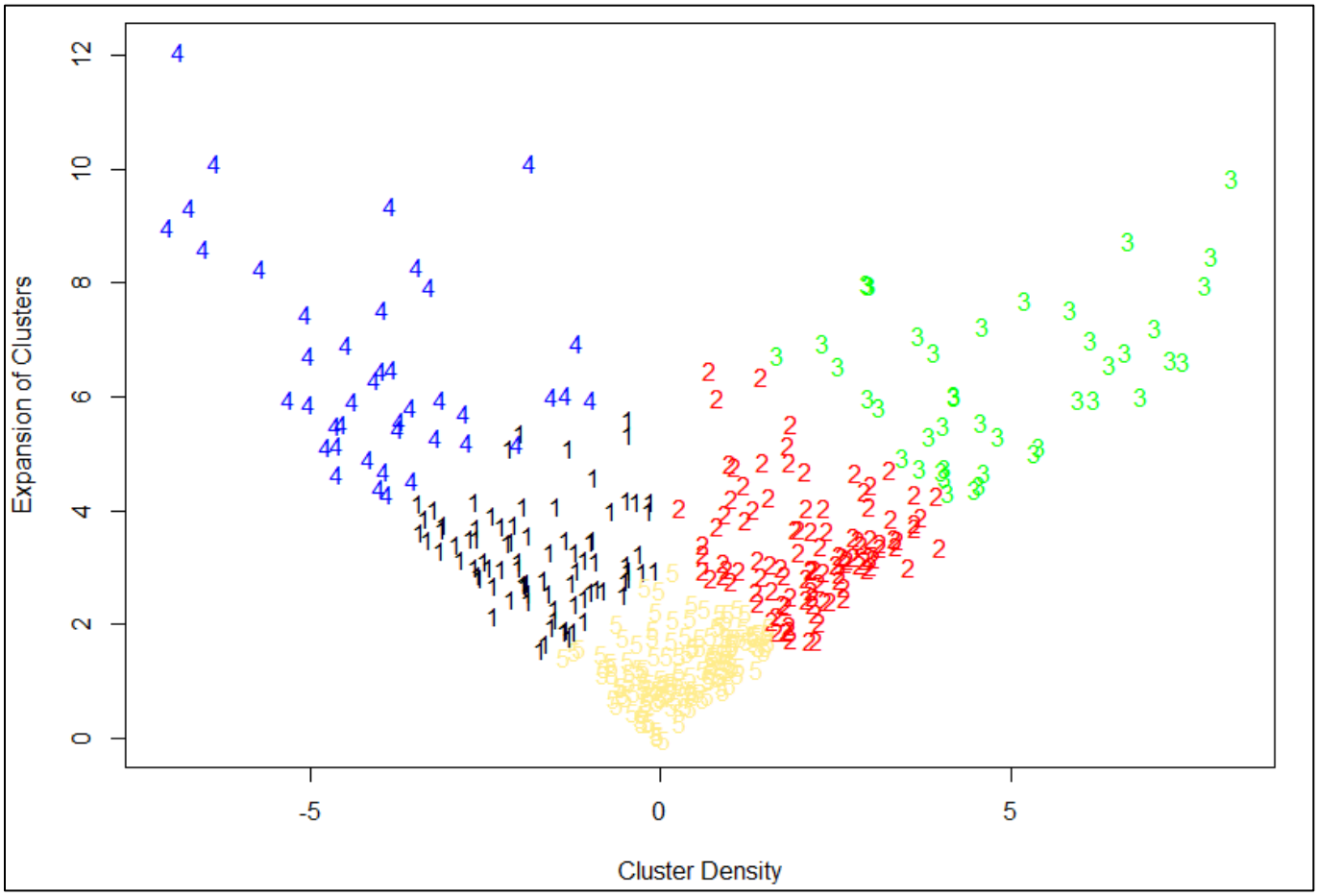
5. Visualizations

I. Plot Interpretations

Plot 1

```
> plotcluster(data.rm.top, k$cluster, xlab = "Cluster Density", ylab = "Expansion of Clusters")
```

The plot here shows the cluster density VS the expansion of the cluster. It can be clearly seen from the plot that the cluster 5 has a noticeable amount of garbage values which shows that the customers did not follow any buying trends and were just buying stuff on a random basis, but as time passed, customers changed the pattern of buying which is represented by the other clusters as they expand throughout the plot and become lesser dense, reflecting varied patterns.

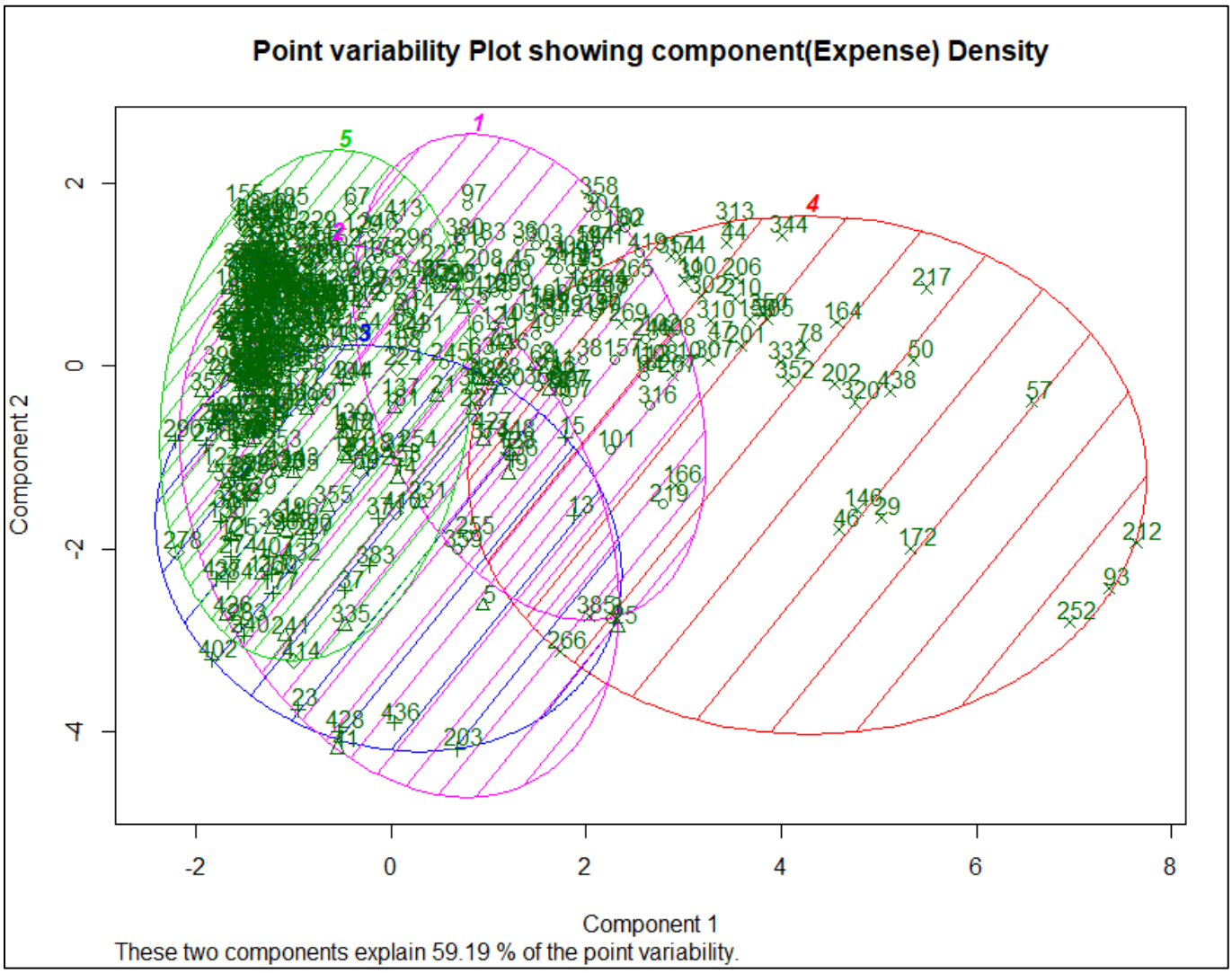


Plot 2

```
> clusplot(data.rm.top, k$cluster, color=TRUE, shade=TRUE,labels=2, lines=0, main = "Point varia
bility Plot showing component(Expense) Density")
```

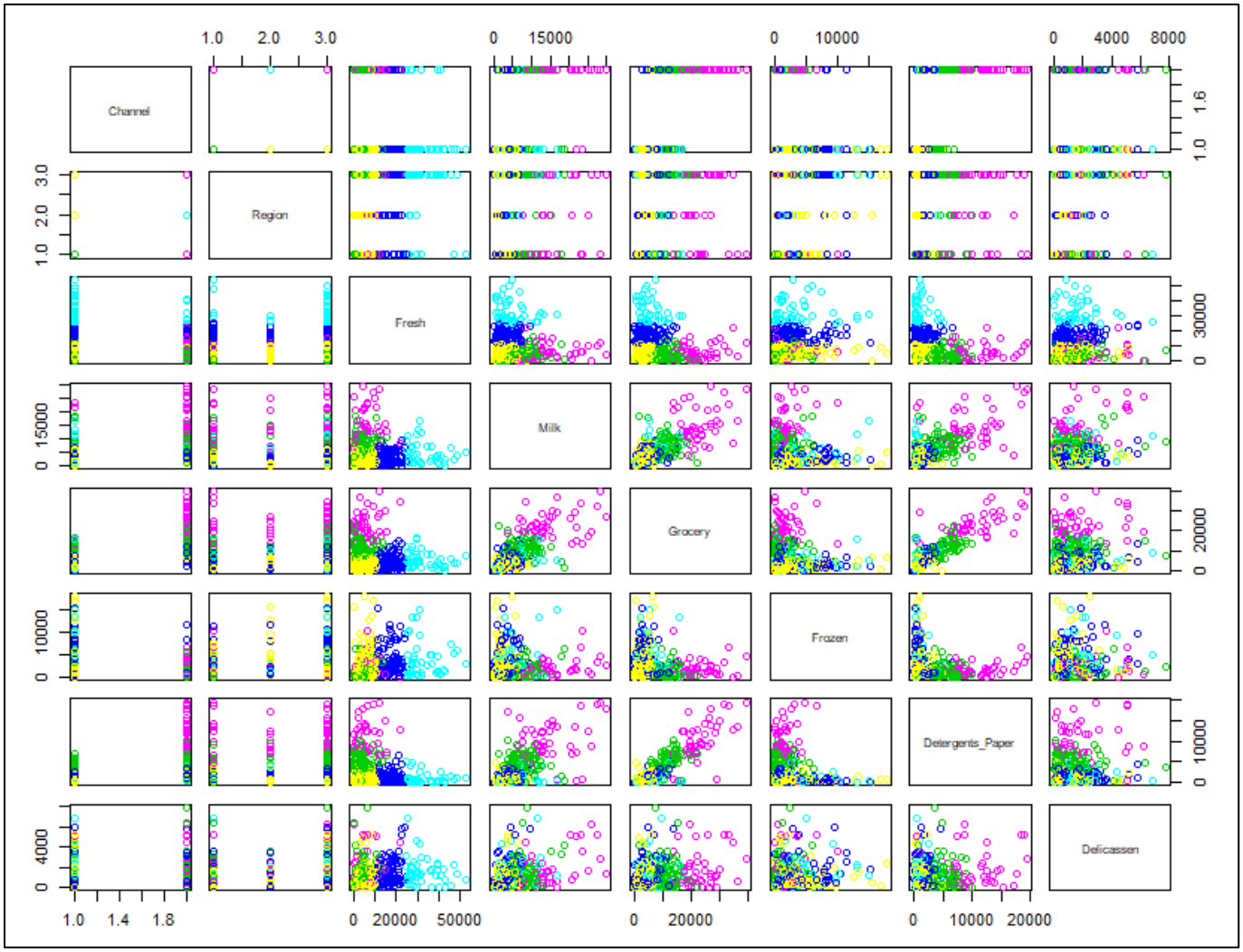
The plot shows the point variability of the component density, that is the Expense density of the wholesale goods in the market. It represents the market trends of how variably the people buy goods and how frequent is the supply to the demand.

We can clearly see that cluster 5 is the densest and cluster 4 is the most vastly expanded. This shows that there was a better trend in the goods purchase with the change in different types of purchase that happened during a long span of time.



Plot 3

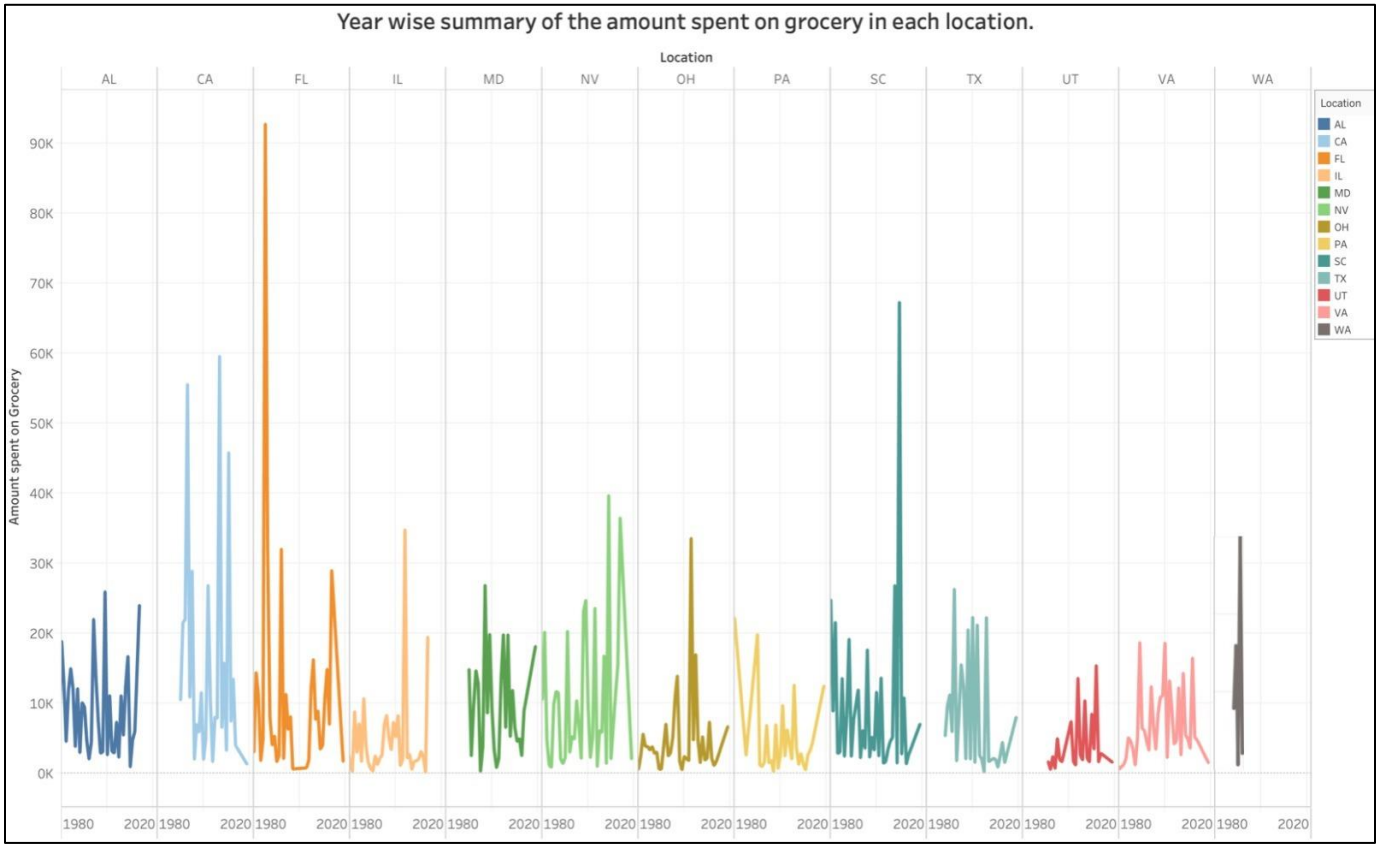
```
> with(data, pairs(data.rm.top, col=c(3:8)[k$cluster]))
```



The above plot shows the variables and their veracity among other variables. Based on these six variables and their distribution plots one can analyze which one is to be held back and which one to be sent for production. The variable goods expenses are either dense or are scarce in a corner, while some are vastly distributed. One can easily see how to manage the market demand by looking at the graph where each variable column shows the cluster density with relation to other variable goods.

Trend and Pattern Plots

1. Line Plot

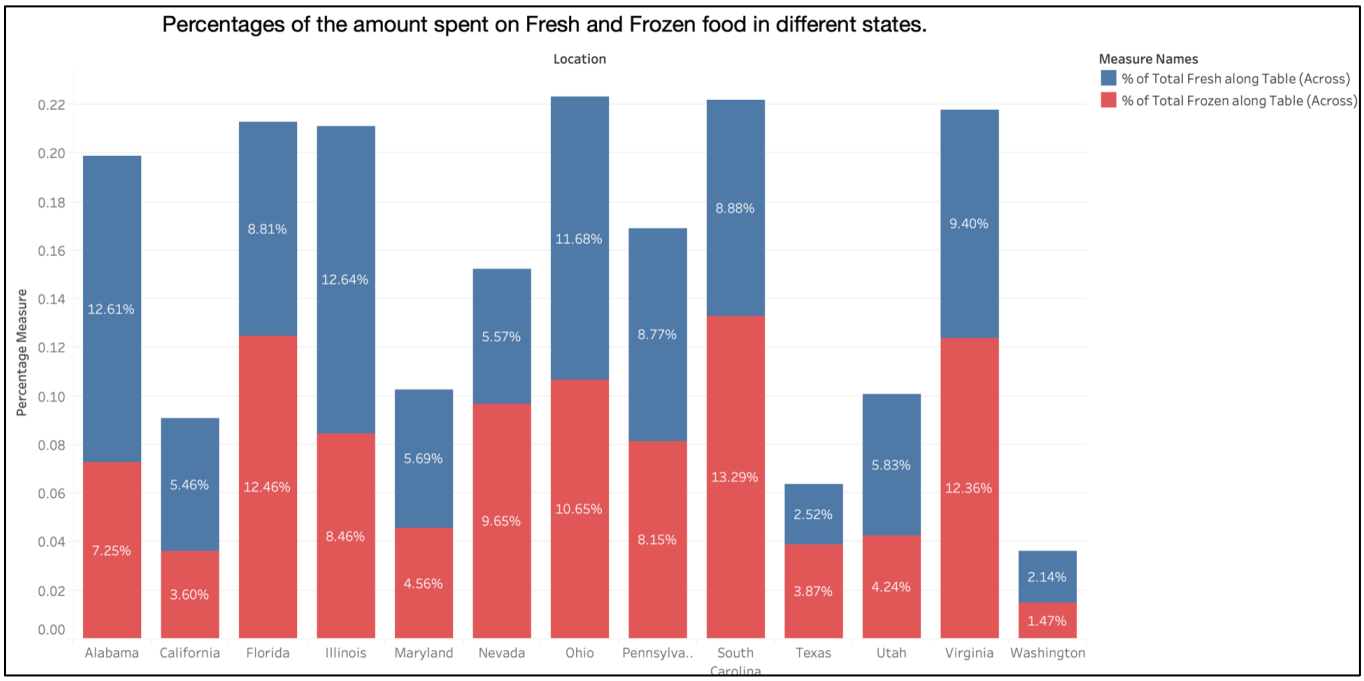


The above line plot is a representation of the buying pattern of grocery in the wholesale market of various states of the US. This is a year wise distributed summary.

As per the graph the amount that has been spent in Florida in the year range of 1980 - 2000 has been the highest until now. There could be two possible reasons for this plot being the highest. The population of Florida might be much more than that of the other states in that particular year range or else there might have been big customers who would have invested a lot in the wholesale market. However, as the line is not constantly high, this in itself, signifies that it is not the population which played a major role in this case but mostly an investment from a high-end customer of the wholesale market, who invested a huge amount in the market. The lowest variation in the pattern can be noticed in Texas. The reasons can be both, lower population or else low-end customers.

Data: Yearly expense of the amount spend on groceries in each location.
Information: The amount of expenses varies from each state to state.
Knowledge: The amount of expenses varies due to the buying trends of the people in various states.
Wisdom: I am able to understand and infer that the amount of funds spent on grocery shopping from 1980 to 2020 had a huge changeover which instilled a lot of losses to the wholesale market and thus the management needs to think about the manufacturing and selling patterns to the customer outlets.

2. Stacked bar chart

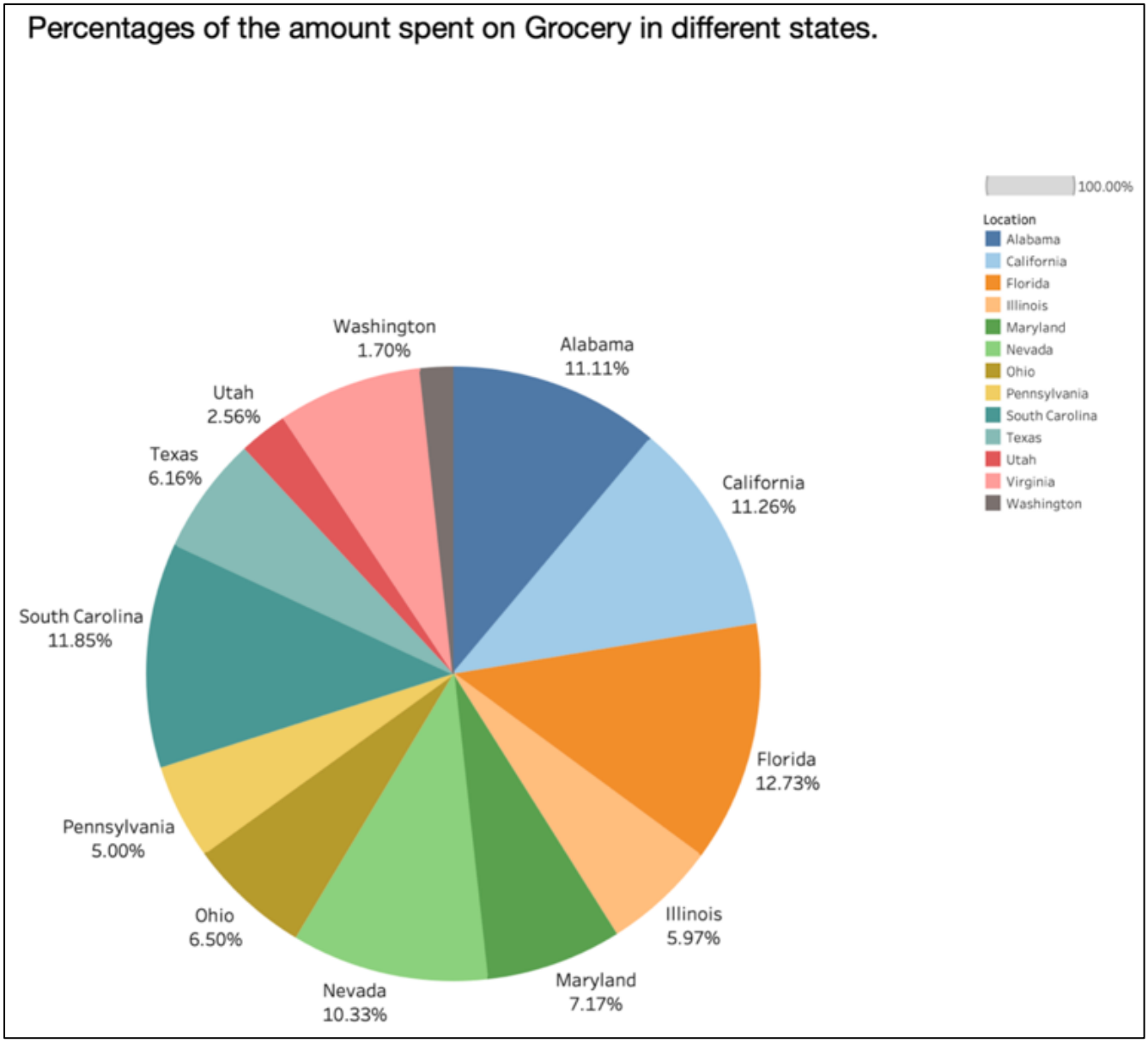


This chart signifies the percentage distribution of the amounts spent on the fresh and the frozen food available in the wholesale market. As per the graph the maximum difference between the amount spent on the two kinds of food is in the state of Alabama, i.e. 5.36. The majority of the people in this state prefer eating fresh food in comparison to the frozen food. It can be inferred from the graph, that the residents are conscious about their fitness and take good care of their health.

There is a major difference between the percentages in the state of North Carolina as well, i.e. 4.41. On the contrary, the people in this state prefer eating frozen food in comparison to the fresh food, in order to save time. They are not much concerned about their health.

Data: Percentage of the amount spend on Fresh and Frozen foods in different states.
Information: The percentage of amount that is spend on both the categories of food is quite near to a total of 22% out of 100%.
Knowledge: The amount of expenses is higher on the fresh foods side rather than the frozen ones.
Wisdom: I am able to understand and infer that the amount of funds spent on fresh foods are more because the people prefer to eat more of the fresh food everyday rather than depending on the frozen food. Also, Washington is the only state where percentage of both the categories are really low.

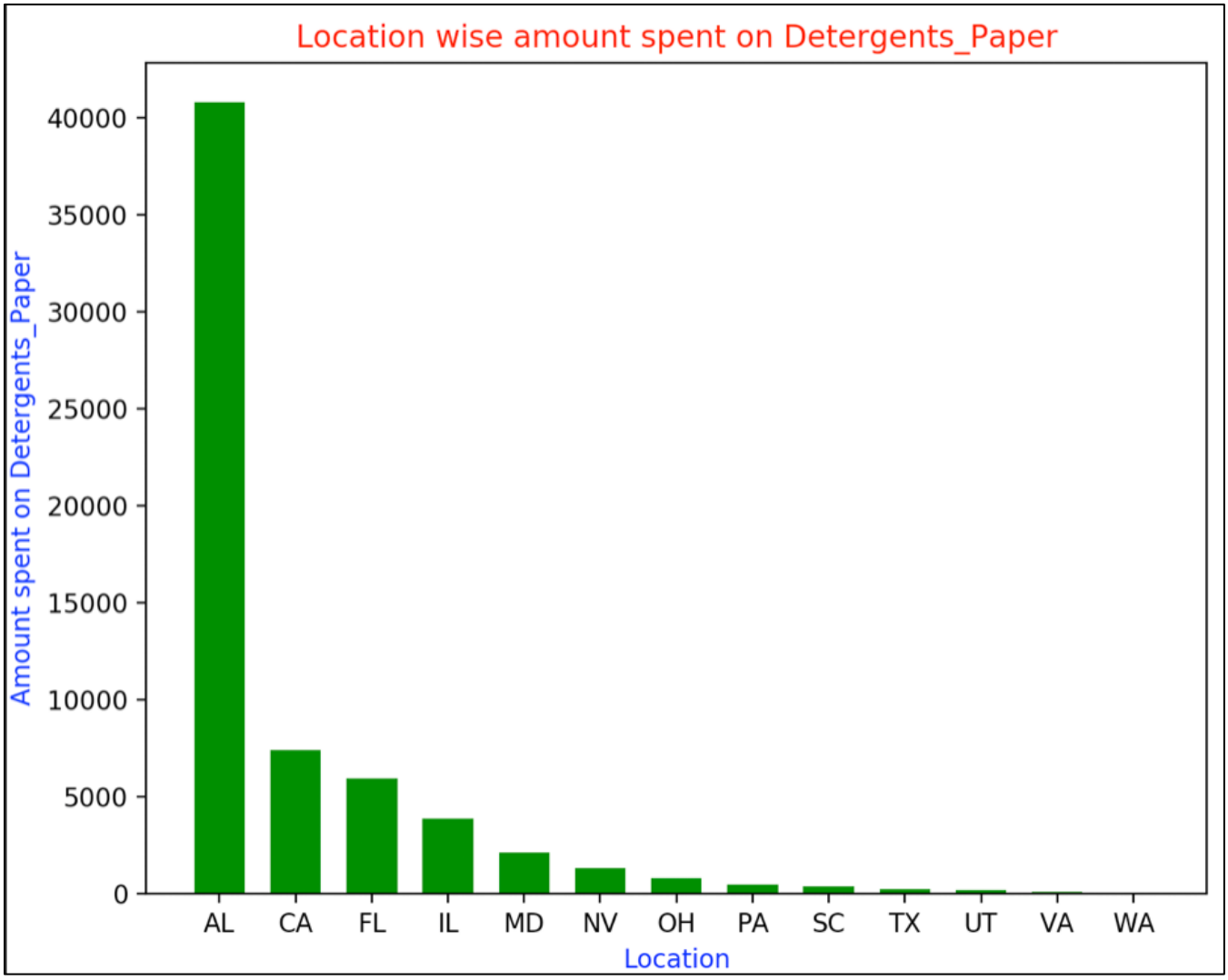
3. Pie Chart



The pie chart is a percentage comparison chart of the amount spent on the groceries, in the various states. The highest percentage is in Florida, i.e. 12.73%., and the lowest is in Washington, i.e. 1.70%. Accordingly, should their market be supplied with groceries. The channels should carry the number of goods, according to the demand in that particular state. If the supply increases the demand, heavy losses will be suffered in that case.

Data: Percentage of the amount spent on grocery in different states on the whole.
Information: The amount of expenses varies from each state to state.
Knowledge: The amount of expenses has a huge variation due to the buying trends of the people in various states.
Wisdom: I am able to understand and infer that the amount of funds spent on grocery shopping on the whole told us a lot about how the people spend and what kind of pattern is followed when calculated on a whole or say complete level.

4. Bar Chart - Detergents Paper

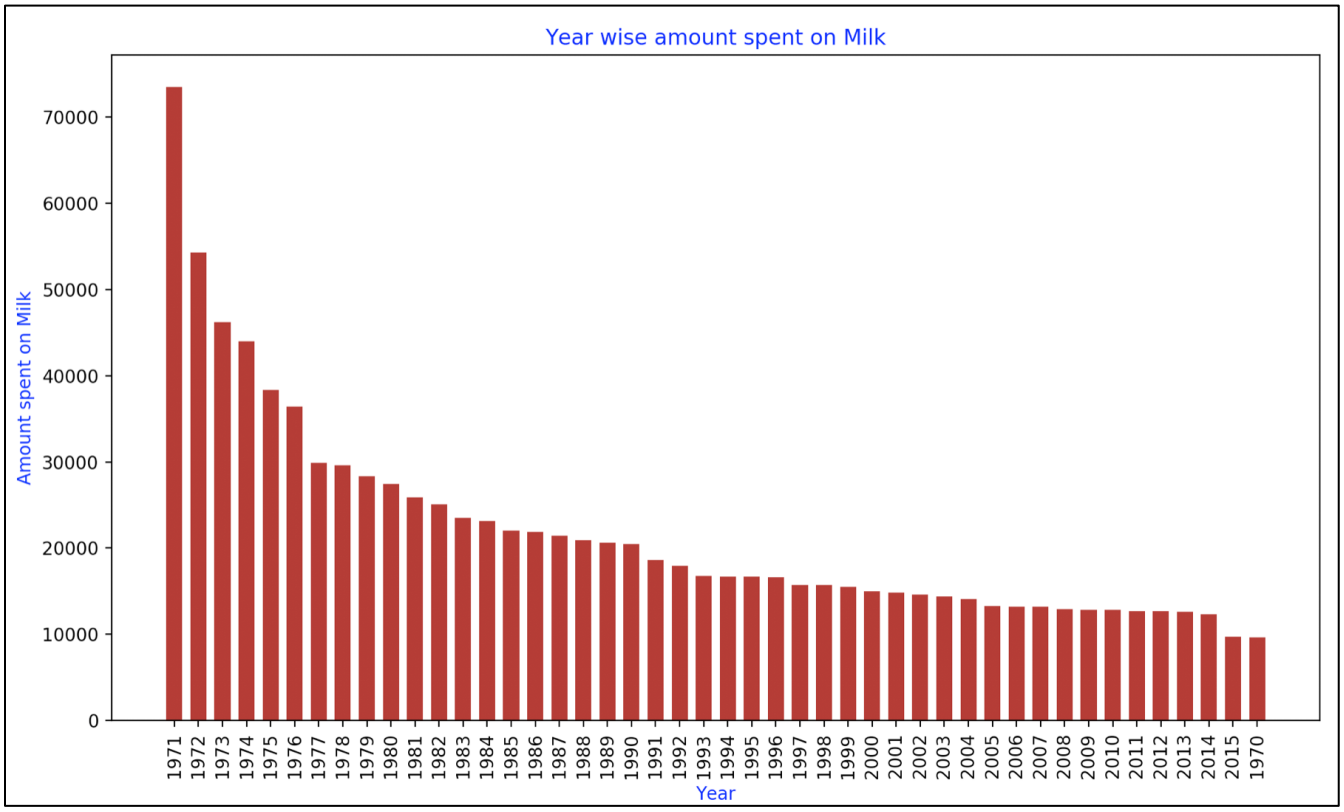


This chart shows the difference between the amount spent on Detergents Paper in the different states of US. It has been plotted by using the python script.

As per the graph the highest amount was spent in Alabama on this particular product. While, in Washington DC the amount spent on it was almost equivalent to null. The reason behind this can be that the residents of this Alabama find Detergent paper to be a more convenient option than the other detergent options. The graph has been depicted in a sorted order of expenditure.

Data: Location wise amount spend on Detergents Paper.
Information: The amount of expenses varies on a big scale from ranging from Alabama to Washington.
Knowledge: The amount of expenses varies on a very large scale and is greatest in Alabama while lowest or none in Washington.
Wisdom: I am able to understand and infer that the amount of funds spent on Detergents paper is quite large in Alabama due to the extreme usage in that state while in Washington there is none.

5. Bar Chart - Milk

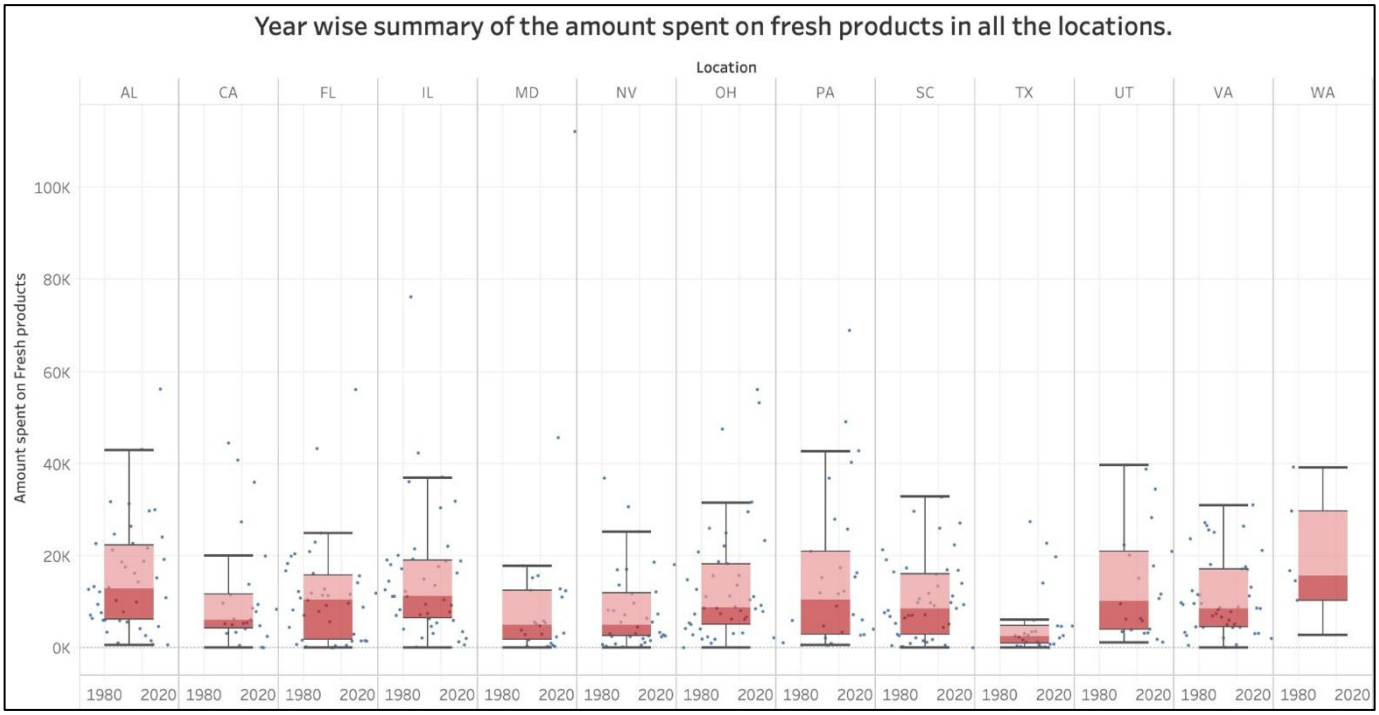


This graph contains a year wise distribution of the expenditure made on milk. On arranging the bars, it was found that with the advancement of time, the consumption of milk has reduced every year.

The only exception to this pattern is the year 1970, wherein there was a sudden rise in the following year. The reason behind this can be inferred as the introduction of dairy farms in 1971, which brought a sudden rise in the supply and consumption of milk. Whereas, the price of milk increased in the following years, which could not be afforded by the lower section of the society as they had to start using other alternatives of milk.

Data: Yearly expense of the amount spend on milk in each location.
Information: The amount of expenses varies from each year to year in a linear manner.
Knowledge: The amount of expenses varies due to dairy farms that opened in the year 1971.
Wisdom: I am able to understand and infer that the amount of funds spent on milk from 1980 to 2015 have a linear flow as can be seen in the plot, because the time after the introduction of milk farm in the US, getting milk became easier and as this is a daily usage product it was used in abundance. But as the years passed the usage trends changed and it depleted with a steep downfall and became kind of constant throughout the years with usage being normalized.

6. Box and Whiskers plot

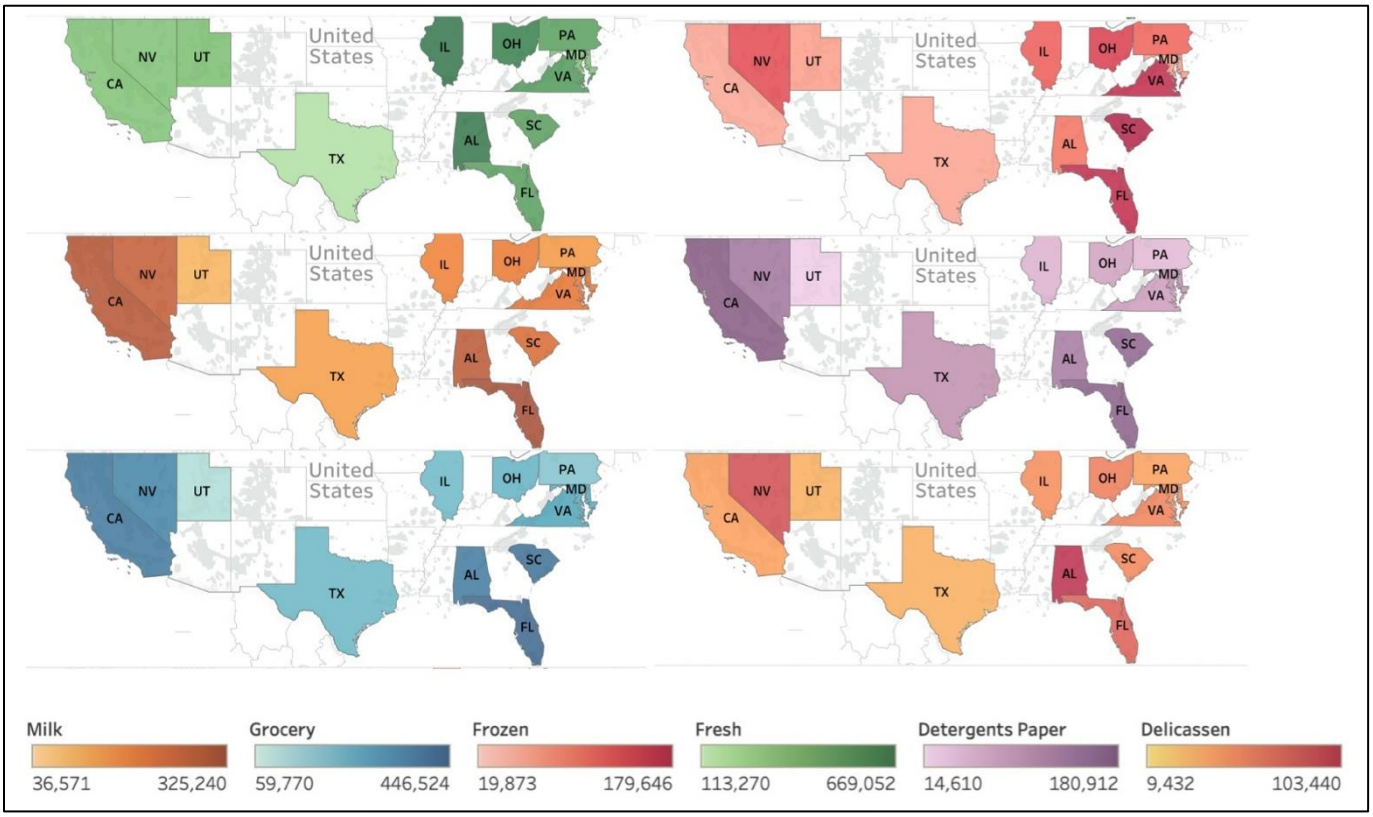


This graph helps in studying the distributional characteristics of group values and it also gives an understanding of the level of the values. The values are sorted in the very beginning and then four equal distributions are made with help of lines that are known as quartiles and the groups are termed as quartile groups.

The plots can be split into the below parts:
Whiskers: It signifies the values outside the middle 50%.
Median: It signifies the midpoint of the values.
Inter-quartile range: It is made up of the lower and upper quartiles, divided by the median. The middle 50% values come under this range.

The above box plot has been plotted based on the amount spent on the fresh food products in various states of the US. As per the plot, the box plot for Texas is the smallest. This shows that there is a very small difference between the amount that has been invested by the high and low-end customers, respectively. The box plot of Washington is not equally distributed across the median. This signifies that there is a huge difference between the amounts spent by the high-end customers, as compared to the difference between the amounts of the low-end customers.

7. US – Map Plot



The above graph shows the ranging demand and varying expenses of every good that is sold in the wholesale market of the various states in the US. It clearly shows that there are states wherein the sales have been lesser and therefore, they do not require as much of the wholesale goods as the others do. In case large amount of production and sales continues without realizing the losses that the vendors are going through, which in turn is destroying the economical balance of the industry.

According to recent facts The United States are under a debt of about $22.22 trillion, where the above given losses constitute a percentage of 19.67%, which is a huge amount and can be saved if the wholesale market functions in a well-managed manner.

Conclusion

It is a very important task to monitor the demand in all of the states in US. The supply in the various regions should be set accordingly, and the channels should be efficient enough to handle the supply of a wholesale market in a particular state. This is the only way that this business can stay in profit and avoid heavy losses. It totally depends on the choice and the level of the customers in a particular region.

The bar chart for milk has shown a clear picture of the amount spent on it getting reduced every year. This is a clear indication of its price being unaffordable. If this trend continues the milk department of the wholesale market will suffer huge losses. Therefore, efforts should be made to make the price affordable. Its price can also be altered by changing the channel that it is currently being used. There are certain taxes on different channels, which tend to make the product expensive.

The observation that is common in all the different kinds of analysis that have been done in this project is that the demand in the wholesale market for all of the products that are present in the data set, is low. Therefore, somewhere or the other this reflects that the prices for all of these products is not that affordable in this state. One way of bringing the prices down would be by opting for a cheaper channel, in order to cut short, the transport charges. However, there shouldn't be an increase in the supply this state or else it will incur heavy losses.

Citations – References

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

2. Forgey, E. (1965). “Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classification”. In: Biometrics.

3. Lloyd, S. (1982). “Least Squares Quantization in PCM”. In: IEEE Trans. Information Theory.

4. Hartigan, J. A. and M. A. Wong (1979). “Algorithm AS 136: A k-means clustering algorithm”. In: Applied Statistics 28.1, pp. 100–108.

5. MacQueen, J. B. (1967). “Some Methods for classification and Analysis of Multivariate Observations”. In: Berkeley Symposium on Mathematical Statistics and Probability