DATA PRE-PROCESSING

RATIONALE BEHIND SELECTING THE FOLLOWING ADDITIONAL FEATURES:

To make our solution more relevant, we have added the following features:

1. **Population density**- We found population density to be more correlated with sales revenue of a city/town than population or area of that city/town.

$$\text{Population density} = \frac{\text{Population of city/town}}{\text{Area of city/town}}$$

2. **Percentage of population using plastic for roof and wall**- Since consumption of plastic commodities was a SKU in the given dataset, we thought this particular variable will have a considerable impact on the sales revenue for the city/town.

3. **Percentage of population exposed to banking facilities**- Usually cities tend to have better banking facilities than towns. We considered this feature as a distinguisher between cities and towns.

4. **Percentage of population using mobiles/landlines/TV/Radio/Computer-** Since consumption of electronic commodities was a SKU we needed to keep in mind, we thought this feature would have decent impact on the sales revenue.

5. **Breaking down total population into 4 categories namely kids, teen, working, oldage-** Segregating population data on the basis of age was extremely necessary because different age groups were found to influence different SKUs. For example, sales revenue collected from toys will be solely influenced by kid's population of that city/town and not by the total population.

6. **Number of females in that city/town-** Number of women in a city/town has a huge impact on the sales revenue of that city/town. This is because women usually have an affinity towards fashion, shoes and home essentials.

7. **Number of literates in that city/town-** If the number of literates are high for a particular region, then the sales of stationeries are going to be high. Hence we considered number of literates in a particular city/town as an important feature. We thought number of literates is a better parameter than literacy rate because literacy rate depends on the total population of a particular city/town while number of literates is independent of it.

8. **Per capita spending-** Perhaps the most important feature to be considered while we are dealing with sales revenue, per capita spending helps us distinguish between cities and rural areas. Also, a region having higher per capita spending is the one we would like to target for our market strategy.

9. **Percentage of population below poverty line (BPL) -** To gauge the poverty of a particular city/town, percentage of population below poverty line is an important index. Usually larger the fraction of population below poverty line, lower would be the sales revenue for luxury SKUs like home fashion, electronics and so on.

10. **NSDP per capita-** One of the most important economic indicators of a state, Net State Domestic Product helps not only gives an idea about the consumption in the state but also about various other important economic parameters like government spending, private investments and exports.

TECHNIQUES USED FOR DATA SCRAPING:

Besides using data from standard government reports and other trusted sources, we have used regression tools to capture data that was not available in the internet. For example, we failed to find standard per capita spending for different cities/towns for the year 2014. We collected per capita spending for the years 1973, 1977, 1983, 1993 and 1999 and used regression tools to fit the data using a quadratic polynomial. The R^2 value for the fit came out to be approximately 0.995. We fitted in the year 2014 in this quadratic polynomial and thus obtained per capita consumer spending data which we have used in our model. Also, due to absence of accurate BPL data for the year 2014, we have used a more reliable data of 2011-12 census, assuming that there would not be major changes in % BPL population in between these 2 years.

FEATURES NOT SELECTED:

We did not intentionally include a few features in our dataset in order to make a strong model. Some of these features are:

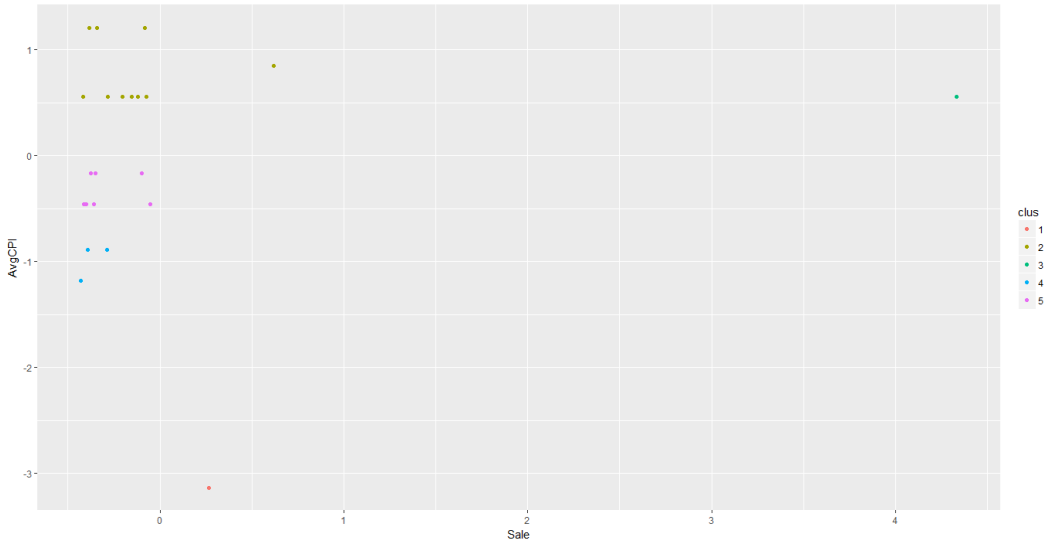
1. **Number of sports academies in each city/town**: We thought the sales revenue for the SKU sports would be highly affected by the number of sports academies in that city/town. However, we could not find accurate data in this respect and hence, had to drop our plan.

2. **Number of hospitals in each city/town**: We thought health care revenue would be affected by number of hospitals/health care centers in each city/town. However, we did not find accurate data in this respect and so did not include this feature in our dataset.

- Tried to related sales, city and SKU by making a heat map plot.

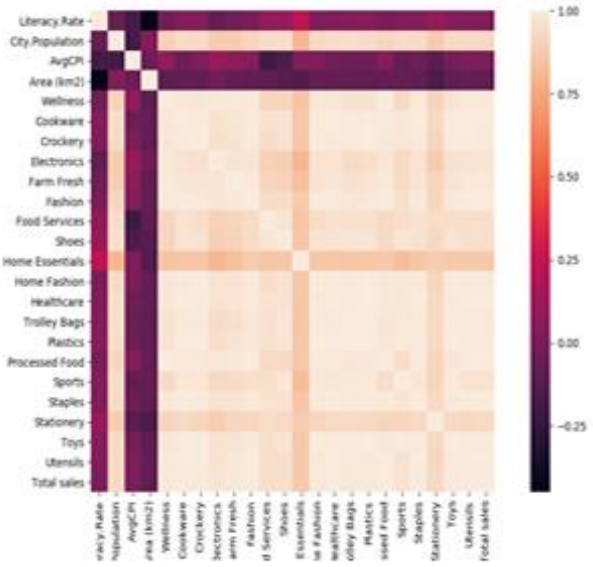
- Relating sales/population , city and SKU.

- Clustering using the sales and cpi for a single type of store(wellness).

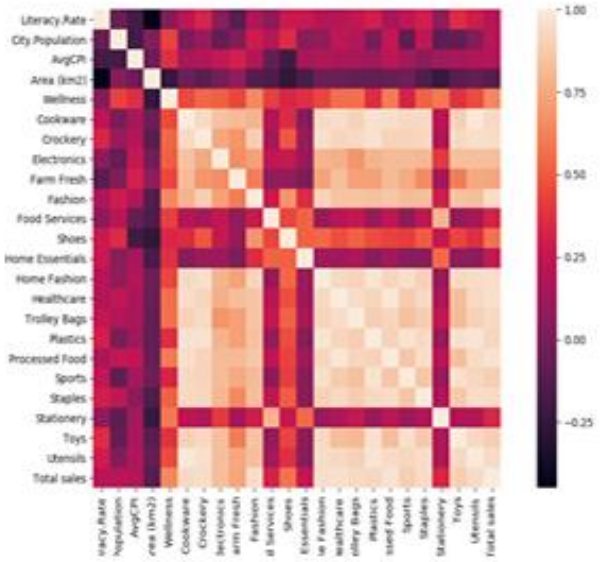


EXPLORATORY DATA ANALYSIS

The two sheets present in the given dataset were combined in two types: one had each of the SKUs as a column, the values of which represented the sales of that particular SKU for the given city. The other was a One-Hot Label Encoded data set for each of the categories of the SKUs.

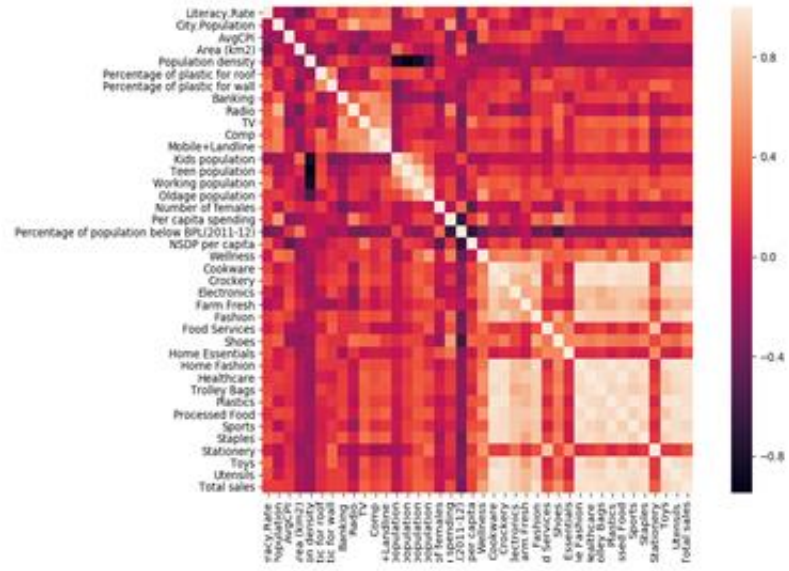


Heat Map for features without addition of data in (type-1 dataset)
Inference: Clearly the big white patch implies a lot of variables are correlated. If we look in details, the 'Population' feature is highly correlated with 'Total Sales', and correspondingly, all the sales of value of each SKU type, gets correlated. The white patch indicates, that not only these values are correlated, but additionally they are linearly correlated. It would only be justified to divide the sales of each by the population. We obtain the following heatmap.



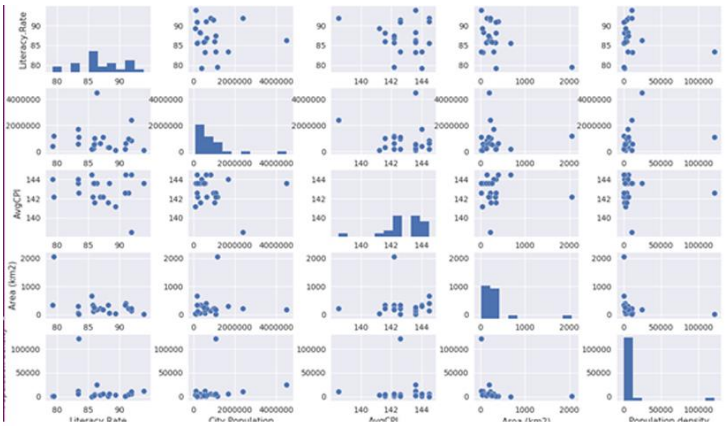
It was observed that the sales of certain SKUs like ‘Home Essential’, ‘Food services’ and ‘Shoes’ are almost not correlated with any other features, which is only justified because these are the items most households need. Even then, we double ensured its non-correlation by other methods mentioned in the forthcoming sections.

The existing data set had very few features. We brainstormed over the various factors that might affect the sales of various SKUs and collected data about the same, which were available from reliable sources. The following graph represents the heatmap for correlation matrix, for the features that we included.



The heatmap shows the correlation of the features (after certain features were scaled per the population). We can clearly see some SKU features being correlated among themselves, more than other SKU features, like the group of (‘Farm Fresh’, ‘Cookware’, ‘Crockery’), (‘Trolley Bags’, ‘Fashion’, ‘Plastics’) and so on, which aptly describes the intuitive correlation between these features.

We also did pair plot analysis of each of the non-SKU features with the different SKU features, and also various SKU types among themselves, to see, if certain two features, were in fact correlated, but non-linearly, and hence, were failed to be represented in the heatmap representation of the correlation matrix. Since, there were lots of pair plot analysis, we are attaching only one such analysis (SKU - Wellness):

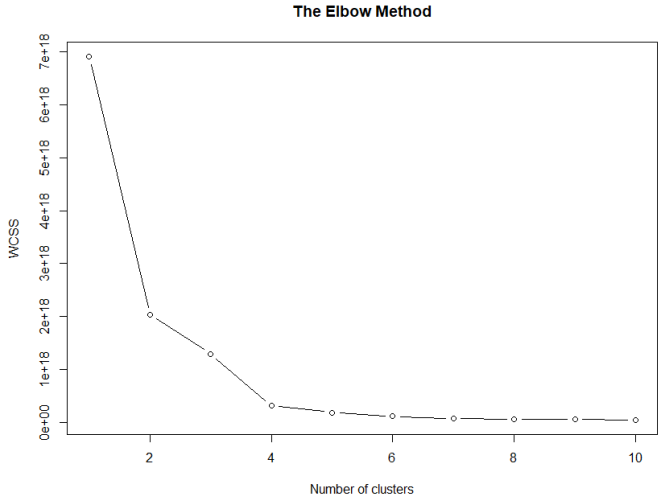


From these pair plot analysis, we were able to infer a lot many things, and got an intuitive idea about how the various features are correlated. One most of the important conclusion was that we found two cities behaving as an outlier outrageously. One of them being Kolkata (plausible reasons: highest population density, largest population in the given data, the only metropolitan city and so on) and the other being Darjeeling (plausible reasons: being a tourist place, the sales of the marts in that city doesn't aptly reflect the demographics, economics and transactional parameters of that place).

DIFFERENT TYPES OF CLUSTERING USED:

KMEANS CLUSTERING

- We concatenated the 2 datasets into one larger dataset.
- Then we converted the categorical variables into binary variables of 0 and 1.
- Then we applied K-Means clustering directly on the dataset with different number of cluster from 1 to 10.
- After then we plotted the WCSS (Within Cluster Sum of Squares) vs number of clusters to check how many clusters are the best.



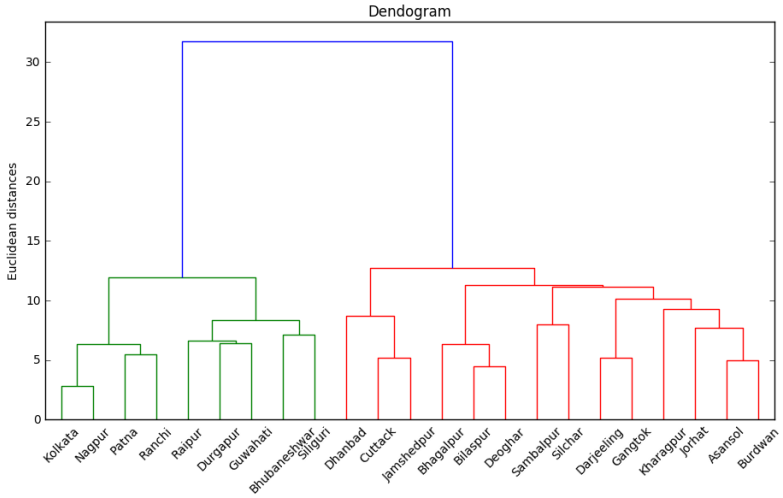
- Number of clusters = 4 is preferable

- Applied **spectral clustering** to the dataset as it was in its normal form. Didn't get any good result and value of silhouette coefficient went negative.
- Tried to apply **K-Means** on a dataset which had many different columns containing all the data that was available. There was one missing value which was replaced by 950.00. This figure was found out by applying mice() algorithm on the dataset.
- Accuracy was best calculated for clusters = 3
- Accuracy decreased as the number of clusters increased.
- Then applied agglomerative clustering to the same dataset and got a silhouette score of 0.716 , when distance was considered as Manhattan distance
- Then applied agglomerative clustering to the same dataset and got a silhouette score of 0.67 , when distance was considered as Euclidean distance

PRESUMING CLUSTERS BY INTUITION:

Converted all the variables in a categorical manner according to their 5-point summary values. i.e. 25%,50%,75% values. 0-25% values were labeled as 0, next 25% as 1, next 25% as 2 and the last 25% as 3. First column which was named as "City" was instead used as the index of the data frame. Column named "State was removed". Column named "City Type" was encoded.

- Applied **Hierarchical Clustering** to the data frame. Got this dendrogram as the result :



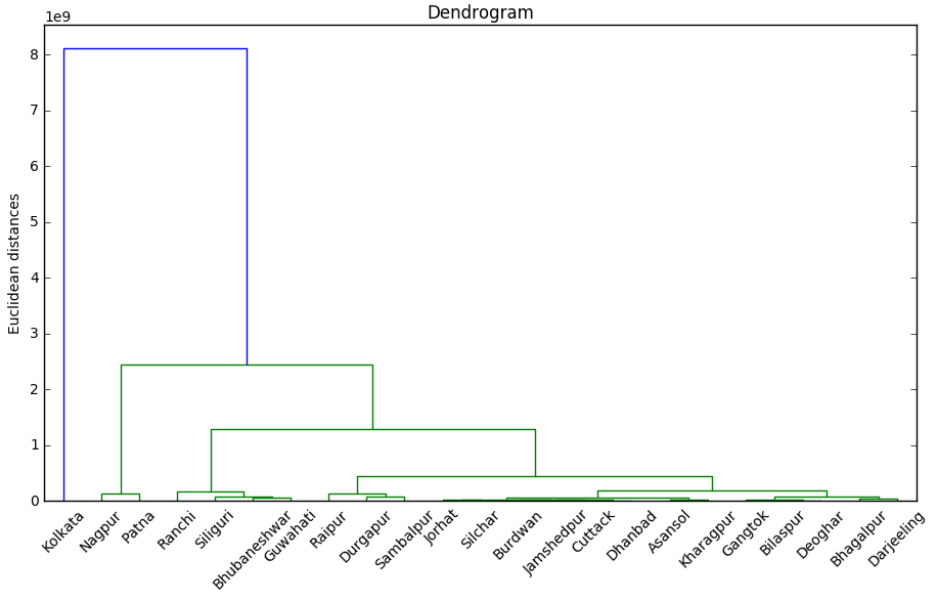
Then applied K-Means on the same dataset.

First performed both steps for clusters = 2. By cross-checking with Adjusted Rand Index we got a value of 1.0 which signified that the clusters were far apart and both clustering methods resulted in formation of same clusters. Value of Silhouette Index for both the clustering was found to be 0.33

Then we performed the clustering for clusters = 3. This time value of Adjusted Rand Index was found to be 0.71, which was also quite good as the clusters could still be very much

differentiated. Silhouette index was found to be 0.19 in case of Hierarchical Clustering and 0.20 for K Means clustering.

- Then applied hierarchical clustering on the whole dataset without changing anything. This was the result of the dendrogram plot:



This way it looks like Kolkata is a kind of outlier. As it is having way too much distance from other clusters. Silhouette Index of the Hierarchical clustering was found to be 0.76 for clusters = 3 and this value was same for K-Means.
We were getting perfect matching results for clusters = 3.

For clusters = 2, we got similar results but with silhouette index of 0.88.

VALIDATION OF THE CLUSTERS

- Dividing the sale row with 1000 to bring it in a suitable range for clustering with all other variables.
- Split the data into training and test by 80% and 20%.
- Preparing a model by applying K-Means on the training set and applying this model on the test set.
- Preparing a model by applying K-Means on test set.
- The numbering of the clusters won't match in both the distributions. So we would use parameters attribute and check the centroids of any one variable say sale.
- Now we will map one-one between the 3 centroids of train and 3 centroids of test as the centroids that are nearest will form similar cluster in both the cases.
- Then according to the mapping we will make a confusion matrix and find out the number of points clustered perfectly.
- We will run this algorithm say 100 times and take the average of all the accuracies (86%).
- Applied the algorithm again and calculated the silhouette coefficient (0.64).

- Also applying the Dunn index (0.04).

CHANGING THE ENCODING OF THE SKUs

- Changed the type of encoding of the SKU. Instead of making binary variables, tried to encode it according to the average sale of each SKU and numbering it to 1-19.
- Applied all the same validation test again.
- Got an increase in the silhouette index to 0.65
- The average of accuracies increases to 87%.

SILHOUETTE COEEFICIENT METHOD

For a clustering model we try to maximize our silhouette coefficient to 1

METHOD:

- From all the available variables, we will run a for loop which takes only two variables and cluster it with K-Means and check for its silhouette coefficient.
- After all the column combinations are taken into clustering we will take those 2 variables that give us the largest Silhouette coefficient.
- Taking these 2 variables as important we will take a third variable (in a for loop) and again check for the maximum silhouette coefficient.
- Continue this till we get a very little increase in the silhouette index
- This may give us the nest features for clustering this dataset

REASON FOR NOT USING IT:

- The variables obtained were not related to any of the sales of the SKUs, which does not help us to infer any predictions
- Silhouette coefficient is not the best cluster validation index for this dataset

FEATURE IMPORTANCE:

Method1:

- For K-Means, WCSS calculates the Euclidean distance between the points in a cluster to the centroid of that cluster.
- So we can assume that the factor that controls the WCSS value, will have the most importance in predicting the clusters.
- So we will calculate the accuracy of the model with K-Means validation model described earlier
- We will try to make new variables/ use feature scaling by dividing some of the variables with the mean of that particular variable.

- After obtaining a desired accuracy we will use WCSS value for feature importance,
- More the WCSS, more is the importance.
- Features that were important in the original dataset were population, literacy rate, area, CPI

Method2:

Initial steps:

- Using additional data obtained from net, we applied stepwise regression on the features keeping output variable as Total sales.
- Dropping the sales of all the rest of the SKUs and categorical variables we tried to apply this regression so as to find the relation of all these variables on the total sales
- In this method we tried to find all the variables having P value less than 0.05(confidence value = 5%)
- Eventually we got 12 important features:
  - ✓ **Avg. CPI of the city**
  - ✓ **Population density**
  - ✓ **Percentage of plastic for roof**
  - ✓ **Percentage of plastic for wall**
  - ✓ **Banking**
  - ✓ **Radio**
  - ✓ **computers**
  - ✓ **Mobiles + Landlines**
  - ✓ **Kids' population**
  - ✓ **Old age population**
  - ✓ **Per capita spending**
  - ✓ **NSDP per capita.**
- Assuming these features to be reliable, we proceeded further.

Method3:

PRINCIPAL FEATURE ANALYSIS (PFA)
For paper citation see : http://venom.cs.utsa.edu/dmz/techrep/2007/CS-TR-2007-011.pdf .

The method exploits the structure of the principal components of a set of features to choose the principal features, which retain most of the information, both in the sense of maximum variability of the features in the lower dimensional space and in the sense of minimizing the reconstruction error.

This was then applied to the whole dataset of ours in which we had all the features considered as per different cities. So the dataset had 23 rows and 43 columns at the start.

At start applied a bit of data preprocessing, by changing the index of the dataset to be the name of the respective city to which the data belonged to, deleting the state column and then also label encoding the city type variable.

Then we fit the dataset to the algorithm and found out the following results. We decided to find the 10 most important features of the dataset. We found that the following columns were the most important features:

1. Area (km2)
2. Fashion
3. Healthcare
4. Food Services
5. Home Fashion
6. Staples
7. Crockery
8. Processed Food
9. Trolley Bags
10. Utensils

This way we could see that most important features were generally the sale values of different type of SKUs. Except Area all of the variables were sale. So maybe this is happening due to fact that the variation in sales is very high, sales figures are higher in comparison to other figures and because there are multiple large values of sales in our dataset so the dataset is generally becoming too much dependent on sales values and characterizing the points solely on based on them.

Now implemented the same algorithm but this time we didn't use the features which described sales. The 8 most important features that we got were:

1. Average CPI for the period
2. Oldage population
3. Teen population
4. Computers
5. NSDP per capita
6. Working population
7. Total sales
8. Literacy Rate
   Where, all the population data were taken as the fraction of total population.

Then we appended the dataset containing only these particular features with our dataset which contained the sales characteristics.

ALGORITHM FOR MARKETING STRATEGY

- We found the most important features from the above techniques
- Now we will find the correlation of all these features with the sales of all the SKUs
- Now we will sort this correlation and find out which SKU is most positively and negatively correlated with these features
- So these SKUs will have that particular feature as the most important one
- Now we divide the sale of the SKU with the corresponding feature and sorted this values in an increasing order

- We will get the top 5 cities with the least value of this ratio
- So we will apply a marketing strategy for that particular SKU on these cities only

Wellness	Bhagalpur(0) , Gangtok(0) , Dhanbad , Jorhat , Nagpur , Cuttack, Raipur , Deoghar , Cuttack
Farm Fresh	Darjeeling(0) , Bhagalpur(0) , Dhanbad, Silchar, Jamshedpur, Cuttack, Jorhat, Bilaspur
Shoes	Gangtok(0), Asansol, Dhanbad, Burdwan, Jorhat, Bhagalpur, Jamshedpur, Durgapur
Food Services	Bhagalpur(0), Silchar, Jorhat, Burdwan, Deoghar, Asansol
Crockery	Bhagalpur, Jorhat, Dhanbad, Jamshedpur, Cuttack, Asansol, Bilaspur
Staples	Bhagalpur(0), Jorhat, Dhanbad, Jamshedpur, Cuttack, Burdwan, Raipur, Bhubaneshwar
Home Essentials	Silchar(0), Sambalpur, Jorhat, Bhagalpur, Burdwan, Asansol, Patna, Dhanbad
Electronics	Gangtok (0), Ranchi , Bhagalpur, Guwahati , Jorhat, Dhanbad , Raipur
Stationery	Bhagalpur, Jorhat, Dhanbad, Guwahati, Deoghar , Bilaspur, Sambalpur , Cuttack
Processed Food	Bhagalpur , Jamshedpur , Jorhat , Cuttack ,Raipur
Utensils	Bhagalpur , Dhanbad , Nagpur , Kolkata , Jamshedpur , Raipur
Fashion	Gangtok, Dhanbad , Jorhat , Asansol , Nagpur

MARKETING STRATEGY:

1. Wellness:
    - The stores that sell wellness products can be strategically placed near Farm Fresh products, to psychologically signify boost the sale of either products.
    - Placing ads at major gyms and community fitness parks and so on, would be a very good strategy.
    - The mart could also become sponsors to some well-known sports championship in the local area for effective publicity.
    - The sale of these products are influenced by average per capita disposable money post expenses of a city. The store could also tie up with online retail giants, to boost their sales.
2. Farm Fresh

- Farm Fresh are something that people would buy everyday. So, the prices of such items should be highly competitive with the market outside. This is a very effective strategy followed by Ikea, one of the leading super-markets, wherein food prices in their restaurants accompanying the stores are very cheap, giving a psychological impression that all the goods in the stores are priced competitively.
- The stores of Farm Fresh and Processed food, should be in vicinity of each other, so that people "find" things worth buying while they are purchasing similar items.

3. Shoes
    - Location of Shoes stores should be near those of Fashion.
    - Seasonality of this sector should be taken into account. Use advanced machine learning algorithms (like RNNs) to predict the sales of shoes in a particular season, so that least amount of products remain unsold.
4. Fashion
    - The same features as that of Shoes are applicable to the Fashion sector as usual.
    - This is the sector where we could boost revenue even without increasing the sales of the sector significantly. The prices should be fluctuated intentionally and incrementally, so as to find the right balance regarding the prices of the fashion goods.
    - Different sections could be allocated to unique and expensive boutique or designer collections, where prices could go beyond a competitive margin. This section could be targeted for the riches and upper-middle classes.
5. Food Services
    - Food services unit should be as far away from the Crockery and Processed Foods as possible.
    - This place should have a better ambience with fewer crowd and more spacing
    - Tie ups could be made with food delivery services, and that would lead to increased sales and also better marketing.
6. Stationary
    - Office stationary are generally not brought from the marts. So, it should be better customized to meet the stationary requirements of school and college going students.
    - Could partner with educational institutions to give away gift hampers from the inventory of the Staples unit, for publicity.
    - Differential pricing (special pricing during peak seasons, like the beginning of the semester or academic year) could be adopted.
7. Home Essential
    - Home essentials are bought by almost every household, on a regular basis. In such cases, profit should be intended to be made out of smaller margins on each product spread over large sales. Hence, prices should be competitive.
    - Provide as many as choices available for each item in Home Essential, so that people would have the tendency to try out new products resulting in increased sales.
8. Electronics
    - Surprise promotional pricing approach should be adopted, to keep people incentivized to come to the store, even though a single person may not visit this unit as frequently as he/she visits other units.

- People who buy electronics from the mart are more likely to go for a online price search before actual purchase, or even more so, go for an online purchase altogether. In such cases, we have to follow value pricing method.
- The promotion of this unit should go into Television ads and online advertisements because prospective customers are more likely to be influenced by these.

9. Staples
    - The principles of promoting bulk purchases at competitive prices to increase the revenue, is the best method to adopt. This could be done by giving discounts for larger purchases.
10. Processed Food
    - Processed Food could also be placed in or nearby the Staples unit, to increase compulsive and impulsive purchases.
    - Tie ups with brands of goods that sell such processed goods, to mention the location of store, when they market their product.
    - Offer differential pricing on special occasions or on a scheduled basis, when expected sales are low.
11. Utensils
    - Utensils unit if a major contributor to the total sales, should be marketed through television and radio ads, which is more common amongst households with larger families.
    - The prices should be competitive and the quality of items should be good, because the best promotion of these goods, occur through word of mouth.
12. Crockery
    - Prices could be fancy, the fancy prices could themselves be a Unique Selling Point.
    - Place expensive crockery sets to tap the upper-middle and the rich class, to entice them to come to the mart more frequently.

WHAT CAN WE DO WHEN CPI OR THE NSDP CHANGES ?

We found the correlation matrix between CPI , NSDP per capita and sales of all the different kinds of SKUs.

For CPI : We found that the linear correlation was highest between Avg. CPI and Wellness (0.36) and was closest to zero between Avg. CPI and Staples (0.06).
For NSDP per capita : We found that the the linear correlation was highest between NSDP and Staples (0.36) and was closest to zero between NSDP and Farm Fresh (0.00).

In case of decrease in CPI of India, we will assume that this decrease will linearly affect decrease in value of Avg CPI of all the cities. By this assumption we can also say that the sales of Wellness SKU will decrease along with the decrease in CPI and that the sales of Staples SKU will remain almost the same as it is not too much correlated with CPI. So what we can do is introduce special offers that will correlate the sales of Wellness SKU with Staples SKU, i.e. Give special discounts on buying some products related to Wellness SKU or bundle up the products of Staples SKU with buying of products of Wellness SKU.

Similarly we can do in case of decrease in NSDP per capita of a state. In this case we can say that decrease in NSDP per capita will result in significant decrease in sales of products of Staples SKU but the sales of

products of Farm Fresh SKU will remain mostly the same. So we can bundle products of Farm Fresh SKU along with buying of products of Staples SKU.

The reverse process can be implemented when the CPI or the NSDP will increase. I.e. we will bundle up products of Staples SKU along with products of Farm Fresh SKU.