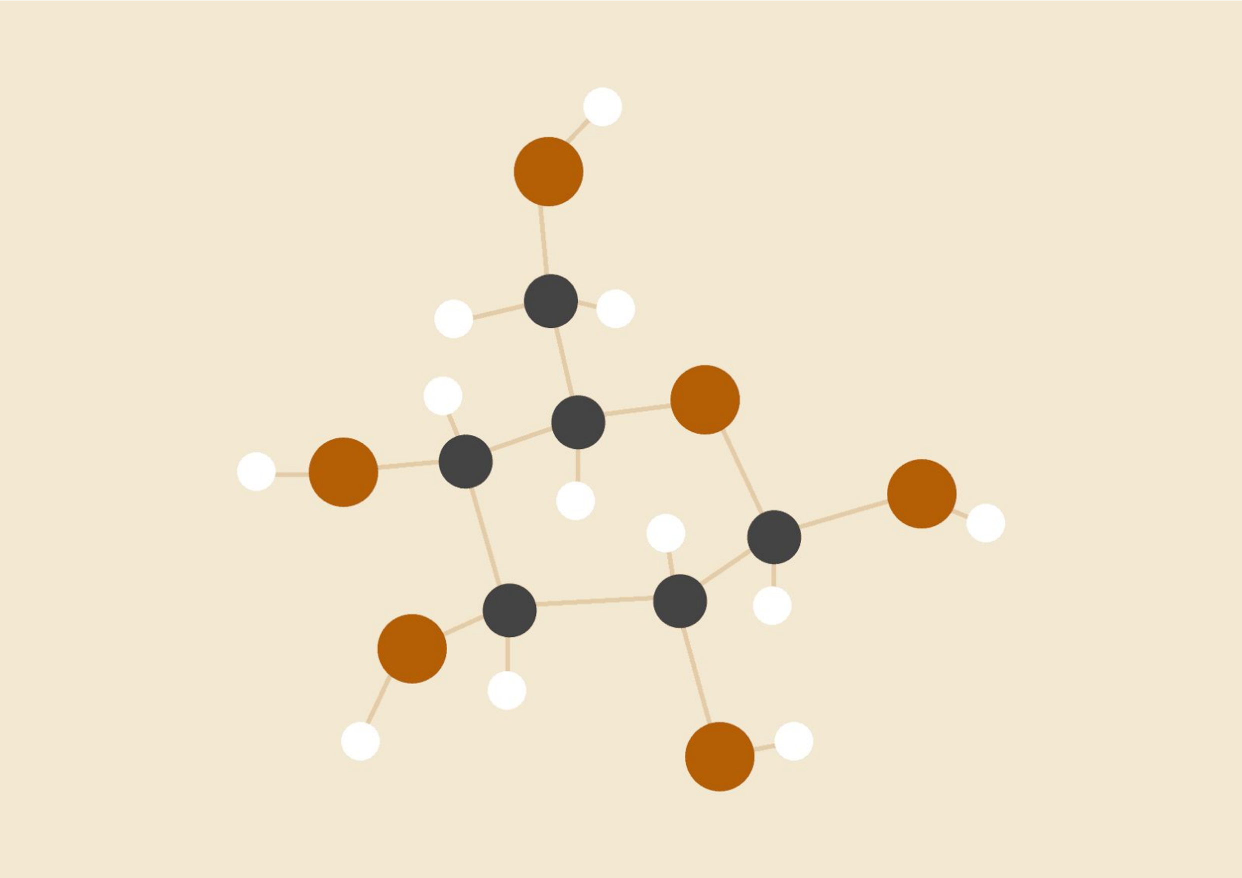
RETAIL CHAIN ANALYSIS REPORT

A detailed report about a retail chain and its sale



Dhiraj Kumar

Sophomore, IIT Kharagpur

Umang Agarwal

Sophomore, IIT Kharagpur

DATASET

The given data was given as a problem statement for Open IIT Data Analytics Competition. The original data set can be download from https://github.com/djbarnwal/Retail-Chain-DA/blob/master/Original_data.xlsx and the rearranged data set can be seen at https://github.com/djbarnwal/Retail-Chain-DA/blob/master/Original_data_rearranged.csv

The dataset contains sale records for 19 SKUs (Stock Keeping Unit), total sale and demographics parameters for 23 different cities and towns.

INTRODUCTION

The aim of this project is to analyse the given data set and describe the best strategies which could boost up the sale for stores. We would also want to know how different parameters affect our sale. Moreover if we open up a new store how much sale we might expect upon its establishment with given demographic parameters.

METHODOLOGY

We merged the two datasets given to give out final cross-sectional dataset on which analysis would be performed.

We added new variables to our dataset for each city from various Government websites like census. These extra variables are for each city individually for further analysis.

https://github.com/djbarnwal/Retail-Chain-DA/blob/master/FinalData.csv

Added demographic parameters-

1. **State** - Added just to see the effect of geographical location on total sale
2. **City Type** - Added to see the effect of urbanisation on total sale
3. **Household Size(State Wise)**
4. **Per capita income** - Added as almost all SKU's eventually depend on income.
5. **Household Income**
6. **Population under 40** - Added as sale for toys and stationery is more in areas with greater percentage of younger population.

7. **Population above 40**
8. **Population (0-20)** -This range of population spends heavily in electronics,packaged food etc.
9. **Population(40-60)** - The more the percentage of old population in a city,the more spent on healthcare and wellness.
10. **Female Population** - Added as women form a major population investing in fashion home essentials etc.
11. **GSDP** - Gross State Domestic Product.

Given demographic parameters -

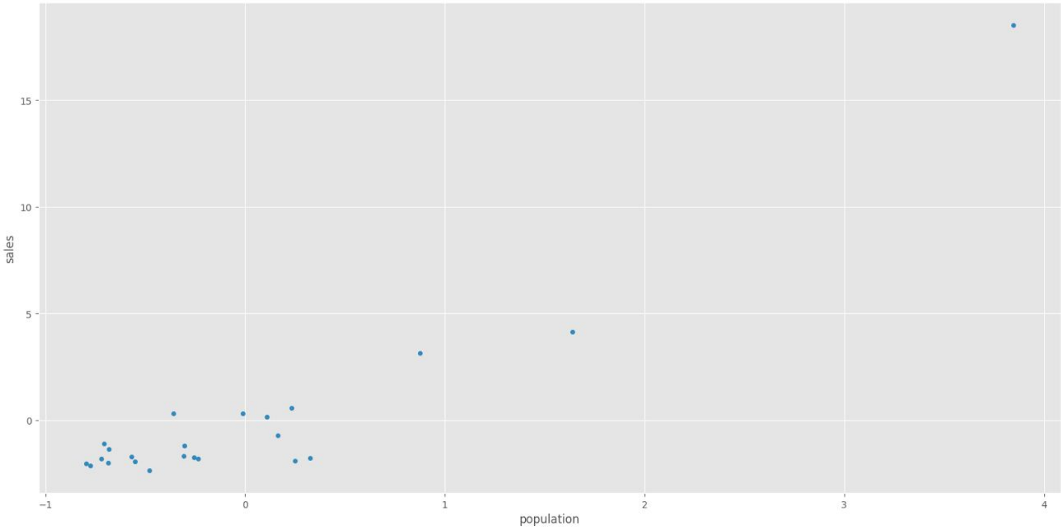
12 .**Total Population -** Population of each city/town given in the dataset.
13. **Area(km2)** - Area of the city/town in km2.
14. **Avg. CPI for the Period** - A consumer price index (CPI) measures changes in the price level of market basket of consumer goods and services purchased by households.

CLUSTERING

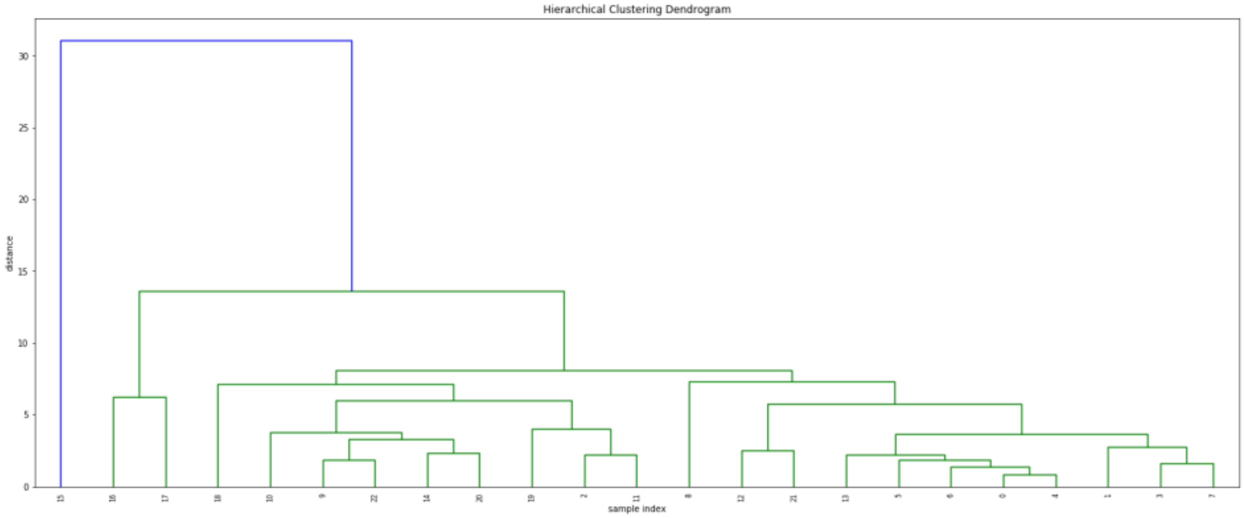
Clustering algorithm used is hierarchical clustering, a method of cluster analysis which seeks to build a hierarchy of clusters. The type of hierarchal cluster used is Agglomerative Clustering: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

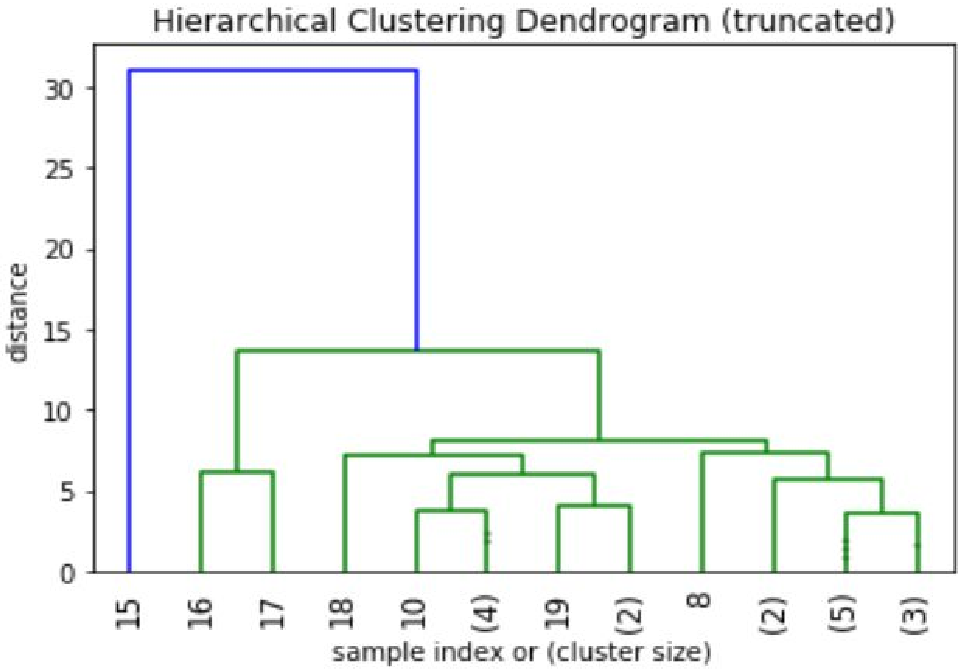
The reason for using this clustering algorithm is that the dataset with such less data points, this clustering algorithm works best for such a dataset.

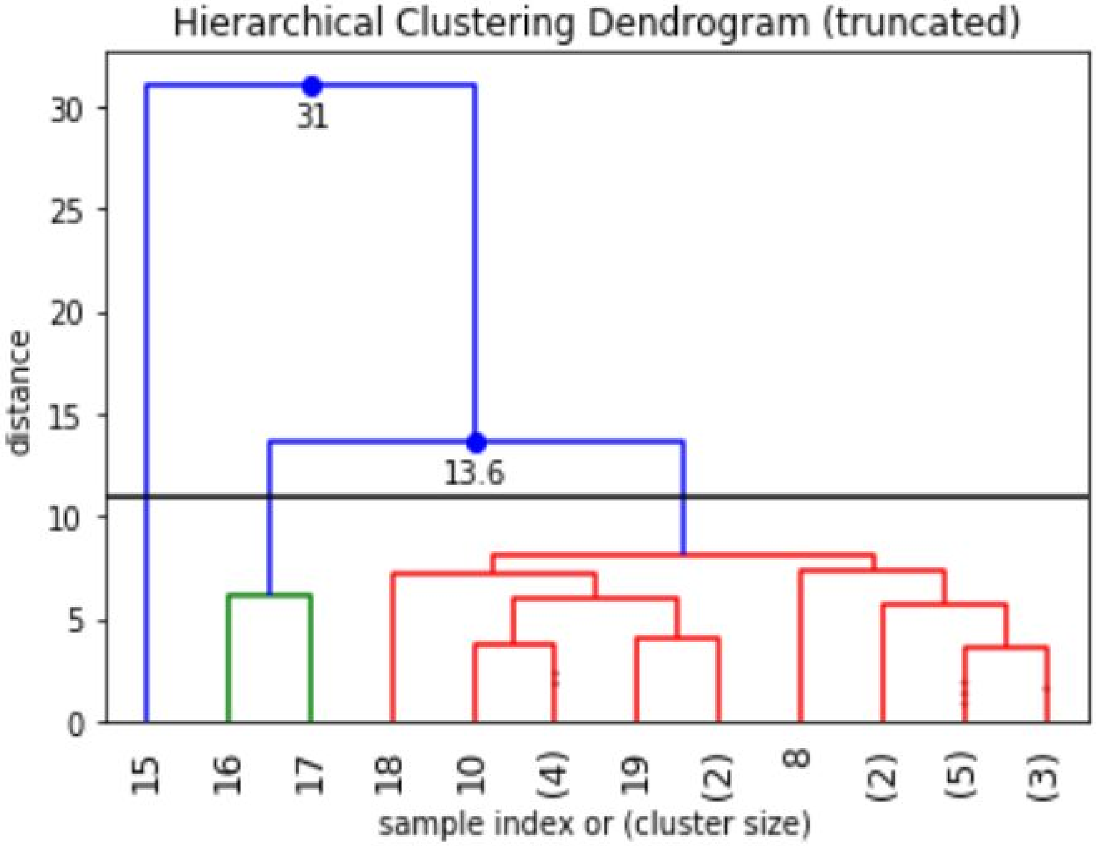
The number of clusters were best for 4 clusters and the judging criteria for our cluster analysis was Silhouette (refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster) More the silhouette scores the more clusters are distinct then others.



Our clusters were divided as follows :-



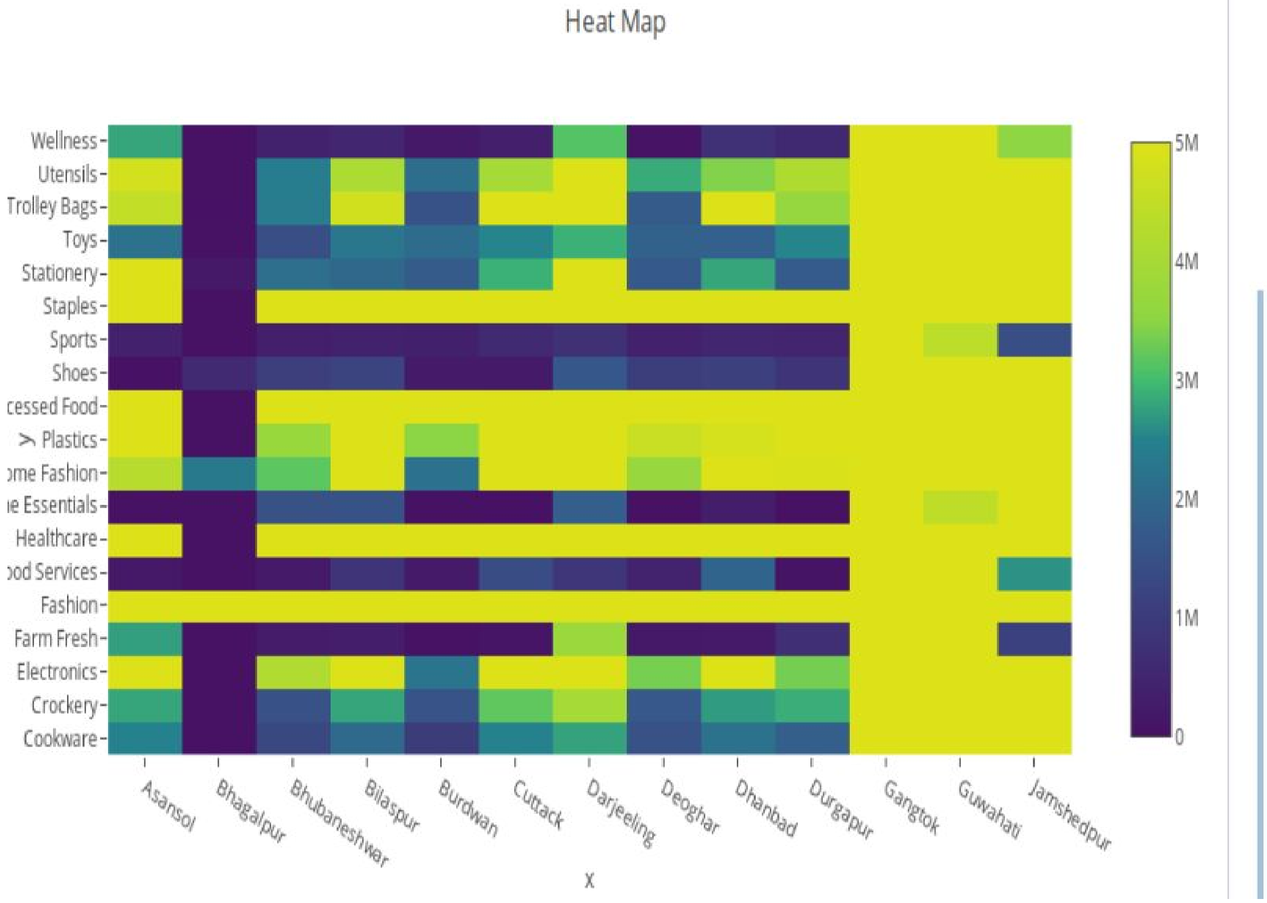

Hierarchical Clustering Dendrogram (truncated)
distance
30
25
20
15
10
5
0
15
16
17
18
10
(4)
19
(2)
8
(2)
(5)
(3)
sample index or (cluster size)


Hierarchical Clustering Dendrogram (truncated)
31
13.6
distance
30
25
20
15
10
5
0
15
16
17
18
10
(4)
19
(2)
8
(2)
(5)
(3)
sample index or (cluster size)

Kolkata - Being the only metropolitan city in our dataset Kolkata was found to be an outlier by the algorithm. This was a cluster in itself. Therefore, the marketing strategy for this is to be done differentially.
Patna and Raipur-These cities were clustered in one group different from other cities.
All cities except Patna and Raipur were found to be clustered together.
And all towns as the 4th cluster.
The analysis of cluster 1 and 2 namely Kolkata as 1 and Patna and Nagpur as another are similar since the demographics of both of these were similar.

The following given data set has very high correlation. Cookware, Crockery, Electronics etc. having minimum correlation of 0.88. therefore we decided to group all the values and then used Principal Component Analysis (PCA) to get a proper idea of the variation of data. In Principle Component Analysis we use an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. We also normalized the data to make the results more reasonable. is the process of organizing the attributes and tables a relational database to reduce data redundancy and improve data integrity. Principle Component Analysis was necessary for feature reduction and efficiency of the results because the given data set is very small.

REGRESSION



By our analysis we were able to found that the correlation between the variables population density and total sales varied a lot for cities and towns.

For cities the correlation was as high as 0.93 which indicated that total sales were highly dependent on population as well as area. More the population in a given area more the sales. This means that the sales given for the retail chain was for different stores across the city.

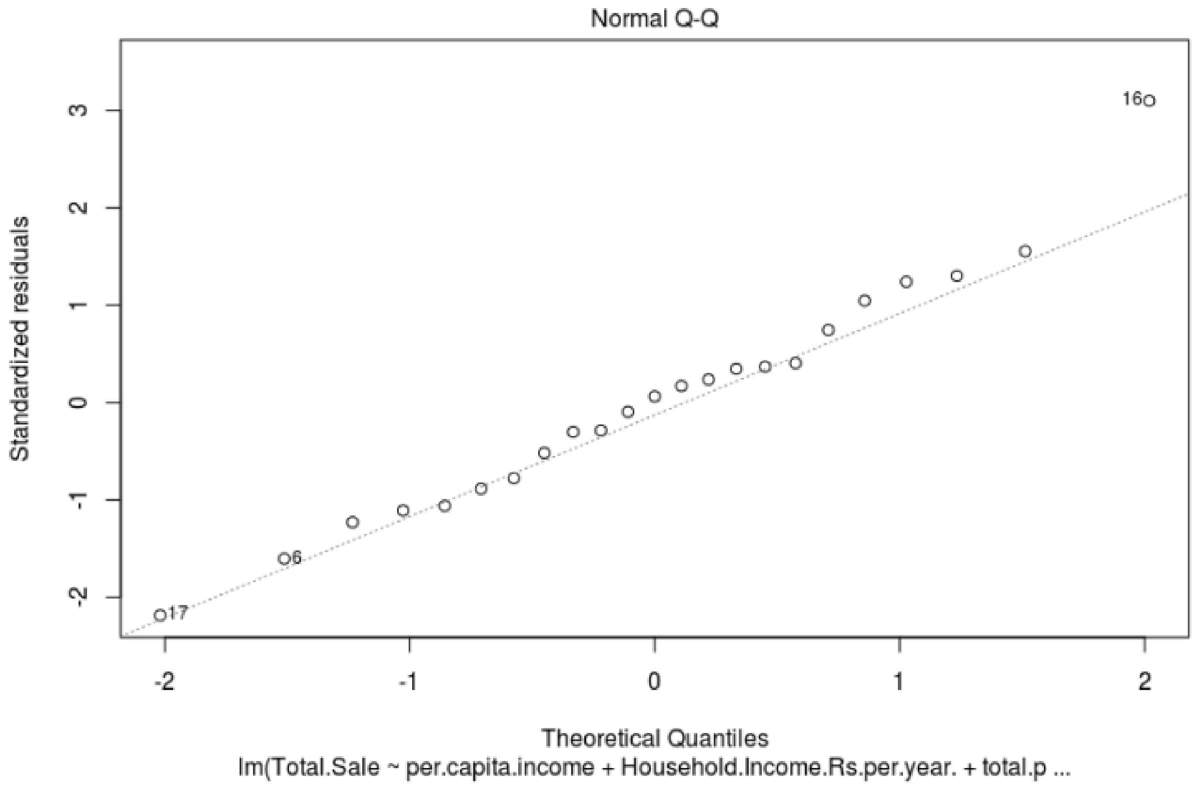
On the other hand, the correlation was nearly equal to zero(-0.05) for towns which indicated a monopoly of the retail chain in these towns as however large the area the whole population went to that particular retail chain for their needs.

1. First ,we tried to fit the regression model taking only Total sales vs population into consideration we observed the following results -
   Multiple R-squared: 0.8666,   **Adjusted R-squared**: 0.8603
   P-value: 1.195e-10
2. Then we tried to take all the demographic parameters into consideration and observed the following results that both R squared and adjusted R squared are

high but the model is overfitting.
Therefore,we used backward elimination to obtain the optimal parameters.

COEFFICIENT	ESTIMATE	STD. ERROR	T VALUE	Pr (>|t|)
(Intercept)	-3.378e+10	9.171e+09	-3.683	0.00201 **
Per capita income	-5.359e+07	2.507e+07	-2.138	0.04833 *
Household income	3.562e+07	1.706e+07	2.088	0.05316
Total population	1.060e+03	8.139e+01	13.022	6.24e-10 ***
Population 40-60	3.425e+09	1.442e+09	2.374	0.03045 *
Avg. CPI	2.304e+08	6.452e+07	3.571	0.00255 **
Area (sq. km)	-2.808e+05	1.554e+05	-1.807	0.08959

We also created a qq plot for checking the validity of this regression model and to see if the assumptions of linear regression hold true. Since the qq plot is at 45 degrees, our multiple regression model works perfectly.



MARKETING STRATEGY

For marketing strategy we have created total 4 cluster based on their demographical structure.

(1) Patna, Nagpur

(2) Kolkata

(3) All Cities except Kolkata, Patna and Nagpur

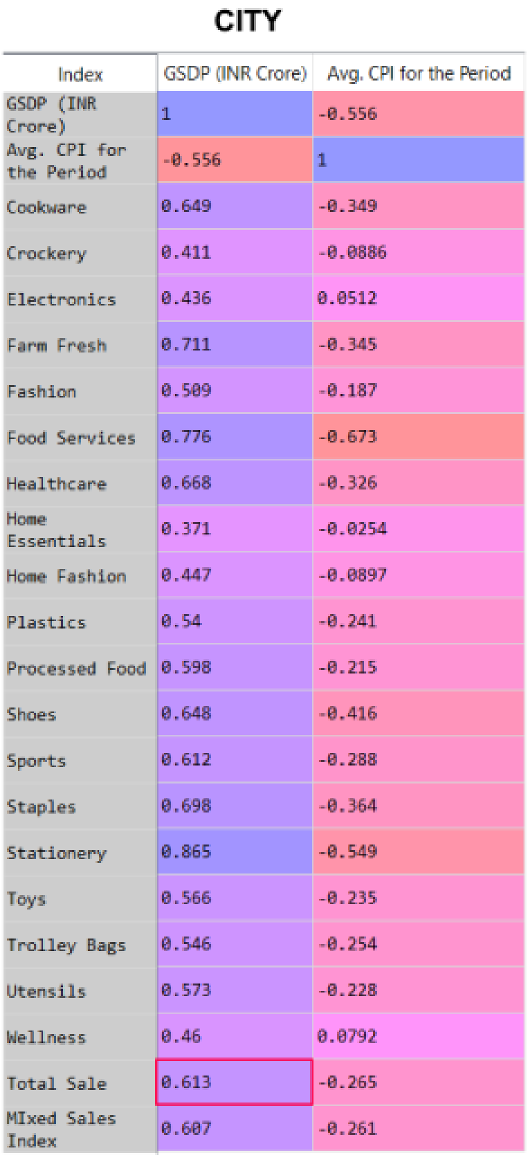
(4) All Towns

The marketing strategy should be decided for these 4 groups by managers of the retail chain. This would help them to make better decisions which would affect the common groups in a beneficial way.

Cities like Kolkata, Patna and Nagpur are relatively more economically advanced and thus are cash cows for the retail chain. Moreover in these cities home delivery can be tried on trial basis and if the results are good, they can be carried down to the other cities as well.

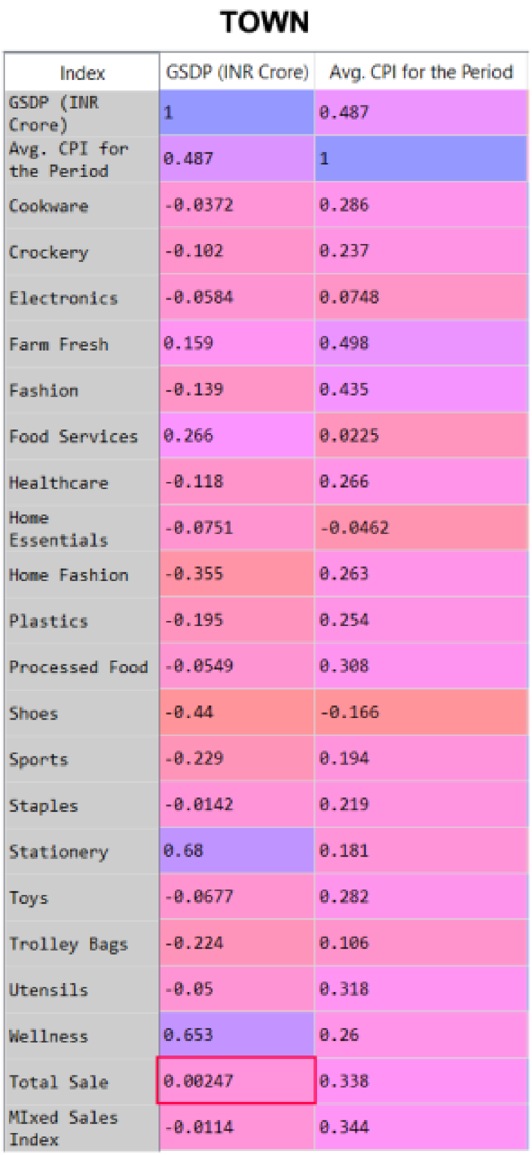
EFFECT OF GSDP

GSDP of cities is highly correlated with the total sales in the cities. On the other hand GSDP of town are close to zero correlation with the total sales as highlighted in the below heat map.



CITY

Index	GSDP (INR Crore)	Avg. CPI for the Period
GSDP (INR Crore)	1	-0.556
Avg. CPI for the Period	-0.556	1
Cookware	0.649	-0.349
Crockery	0.411	-0.0886
Electronics	0.436	0.0512
Farm Fresh	0.711	-0.345
Fashion	0.509	-0.187
Food Services	0.776	-0.673
Healthcare	0.668	-0.326
Home Essentials	0.371	-0.0254
Home Fashion	0.447	-0.0897
Plastics	0.54	-0.241
Processed Food	0.598	-0.215
Shoes	0.648	-0.416
Sports	0.612	-0.288
Staples	0.698	-0.364
Stationery	0.865	-0.549
Toys	0.566	-0.235
Trolley Bags	0.546	-0.254
Utensils	0.573	-0.228
Wellness	0.46	0.0792
Total Sale	0.613	-0.265
MIxed Sales Index	0.607	-0.261



TOWN

Index	GSDP (INR Crore)	Avg. CPI for the Period
GSDP (INR Crore)	1	0.487
Avg. CPI for the Period	0.487	1
Cookware	-0.0372	0.286
Crockery	-0.102	0.237
Electronics	-0.0584	0.0748
Farm Fresh	0.159	0.498
Fashion	-0.139	0.435
Food Services	0.266	0.0225
Healthcare	-0.118	0.266
Home Essentials	-0.0751	-0.0462
Home Fashion	-0.355	0.263
Plastics	-0.195	0.254
Processed Food	-0.0549	0.308
Shoes	-0.44	-0.166
Sports	-0.229	0.194
Staples	-0.0142	0.219
Stationery	0.68	0.181
Toys	-0.0677	0.282
Trolley Bags	-0.224	0.106
Utensils	-0.05	0.318
Wellness	0.653	0.26
Total Sale	0.00247	0.338
MIxed Sales Index	-0.0114	0.344

It can be concluded that there should be no change in the marketing strategy for towns as there is no correlation between GDP and Total sales. For cities our strategy would change slightly as total sale is affected by the change in GSDP. Increasing GSDP in an indication of a prosperous state and which results in less savings and more consumption by the consumer. In case the GSDP increase, the previous strategies as decided for the cities should be implemented more rigorously. An investment should be made to attract customers as the economy is prospering and the marginal propensity to consume increases for the consumers.

REFERENCES

1. http://censusindia.gov.in/
2. http://www.censusindia.gov.in/2011census/population_enumeration.html
3. https://en.wikipedia.org/wiki/2011_Census_of_India
4. http://mhrd.gov.in/statist
5. http://mhrd.gov.in/sites/upload_files/mhrd/files/statistics/ESG2016_0.pdf
6. https://en.wikipedia.org/wiki/List_of_Indian_states_and_union_territories_by_GDP_per_capita
7. www.arcgis.com/home/item.html?id=6cf22970ea8c4b338a196879397a76e4
8. http://ensusindia.gov.in/2011census/hlo/hlo_highlights.html
9. https://en.wikipedia.org/wiki/Indian_states_and_territories_ranking_by_sex_ratio
10. https://en.wikipedia.org/wiki/Kolkata
11. https://en.wikipedia.org/wiki/List_of_cities_in_India_by_area
12. https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population
13. https://data.gov.in/keywords/consumer-price-index