

Analysis and Visualization of Super market Data using Data Science Techniques

M V Jerabandi, Keerti Rayaraddi, Achala Shirol, Sneha Mugalkod, Sushmita Tirlapur

mark.jerabandi@gmail.com

Dept of CSE, Rural Engineering College

Hulkoti, Karnataka, India

ABSTRACT

One of the most motivating areas of research is data science that becomes continuously popular in supermarket analysis. For discovering new trends, mining plays an important role in supermarket analysis which is helpful for all parties associated with this field. The process of data science is to extract data by automatic or semi-automatic means. Data science consists of artificial intelligence, machine learning and database management to extract new patterns for huge data sets and the knowledge associated with these patterns. So, we can use data science in supermarket application, through which management of supermarket get converted into knowledge management. The growth of supermarkets in most populated cities is increasing market competitions at higher rate. Data science can also help the supermarkets to predict shopping trends and buyer habits, this data helps the store not just to know but also how to place them in the store. Hence, in this project work, it is proposed to build a model to analyse supermarket data by using various data science techniques and tools and visualization of data.

1.0 Introduction

The goal of every supermarket is to make profit. This is achieved when more goods are sold and the turnover is high. A major challenge to increasing sales of supermarket lies in the ability of manager of forecast sales pattern and know readily beforehand when order and replenish inventories as well as plan for manpower and staffs. The amount of sales data has steadily been on the increase in recent years and the ability to leverage this gold of data separates high performing supermarket from the others. One of the most valuable assets a supermarket can have data generated by customers as they interact with various supermarket. Within these data, lies important patterns and variables that can be modelled using machine learning algorithm; this can to very high degree of accuracy correctly forecast sales. There exist several techniques to forecasting supermarket sales and historically, many supermarkets have relied on these traditional statistical models. However, machine learning has grown to be an important area of data science that has gained ground due to its high predictive and forecasting powers and as such as become the go to for highly accurate sales forecasting as well as other important areas. To correctly forecast a future event, a machine learning model trained on data from which it learns patterns that are used to predict future instances. An accurate forecasting model can greatly increase supermarket revenue and is generally of great importance to the organization as it improves profit as well as provides insights into the way customers can be better served. Supermarkets are big business and they use data on a big scale. Originating in the US in the 1930s, supermarkets have since gradually taken over a bigger and bigger share of the retail and grocery market. Giants like Wal-Mart, Aldi and Carrefour are among the largest retailers in the world with revenues approaching the hundreds of billions. As such many have invested heavily in big data, with analytics and data science forming a core part of their decision making. The

traditional supermarket occupies a large amount of floor space, usually on a single level. It is usually situated near a residential area in order to be convenient to consumers.

The basic appeal is the availability of a broad selection of goods under a single roof, at relatively low prices. Other advantages include ease of parking and frequently the convenience of shopping hours that extend into the evening or even 24 hours of the day. Supermarkets usually allocate large budgets to advertising, typically through newspapers. They also present elaborate in-shop displays of products. Supermarkets typically are chain stores, supplied by the distribution centers of their parent companies thus increasing opportunities for economies of scale. Supermarkets usually offer products at relatively low prices by using their buying power to buy goods from manufacturers at lower prices than smaller stores can. They also minimize financing costs by paying for goods at least 30 days after receipt and some extract credit terms of 90 days or more from vendors. Certain products (typically staple foods such as bread, milk and sugar) are very occasionally sold as loss leaders so as to attract shoppers to their store. Supermarkets make up for their low margins by a high volume of sales, and with of higher-margin items bought by the attracted shoppers. Self-service with shopping carts (trolleys) or baskets reduces labor cost, and many supermarket chains are attempting further reduction by shifting to self-service checkout.

Data science is a discipline that uses modern tools and techniques to collect and interpret vast amounts of data. It can also make informed business decisions. Data science is a concept to unify statistics, data analysis, informatics, and their related methods. in order to understand and analyze actual phenomena with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing award winner Jim Gray imagined data science as a fourth paradigm of science and asserted that everything about science is changing because of the impact of information technology and the data deluge.

Data science has found its applications in almost every industry.

Healthcare: Health care companies are using data science to build sophisticated medical instruments to detect and cure diseases.

Gaming: Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

Image recognition: Identifying patterns in images and detecting objects in an image is one of the most popular.

Recommendation system: Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms. Logistics: Data science is used by logistics companies to optimize routers to ensure faster delivery of products and increase operational efficiency.

Fraud detection: Banking and financial institutions use data science and related algorithms to detect fraudulent transactions Data science plays an important role in analysing the data of supermarket.

The data analysts examine data sets to identify trends and draw conclusions, data analysts collect large volumes of data, organize it and analyse it to identify relevant patterns. After the analysis, they strive to present their findings through data visualizations methods like charts, graphs etc. The analysis is usually conducted via a rule, mining algorithm.

It collects the useful from the data, a special function accepts the data, splits it according to some different factors and drops the useless or not required data. Data has become of great

importance for those willing to take profitable decisions during the business. Data needs to be very beneficial for every company's decision maker, this analysis of a vast amount of data allows influencing or rather manipulating the customers decisions.

1.1 Related Work

The following are some of the papers which have been referred by us for better understanding of the concept supermarket data analysis and the data visualization.

An analysis of supermarket pricing, published on April 1, 2016: According to this paper, The purpose of study is to explore price fluctuations (tracking of pricing trends) in essential consumer items among identified Supermarkets in Gaborone. The prices were read from shop displays at the beginning of the month, mid-month and at the end of the month.

Objectives of the study

- To find the pricing strategies used by the selected supermarkets.
- To identify price fluctuations in selected commodities among identified supermarkets.
- To find out whether there are any clear patterns of pricing within the branches of a supermarket.
- To assess consumer awareness on regular prices and specials.

Grewal et. al, presented a paper in which retailers today are experimenting different pricing models to test the one that will lead to higher purchases, and enrich their retail mix. hinted those prices are also being changed based on the prices of competitors, time of the day or even conversion rate. Competition between supermarkets appears to be much more intense than ever, as supermarkets devote 80% of their hours to managing promotions and only 20 % of the retail sales come from those promotions.

Fassnacht et. al, (2013) According to the study performed on this paper, the pricing strategy is seen as one of the five most important priorities in retail management. While supermarkets compete along many dimensions, the pricing strategy is clearly one of the most important factors that stores take on board for their successful operation. In many retail businesses, pricing strategy can be categorized as a choice between offering relatively stable prices across a wide range of items, popularly known as Every Day Low Pricing (EDLP) or adopt a big and frequent discount on a smaller set of items known as High-Low Pricing (Hi/Lo).

International Journal of Engineering research and technology (2018) The study of this paper describes that there are many techniques proposed in order to efficiently process large volume of medical record which has to enhance the processing of Supermarket, they have a proposed a series of Big Data in Super Market by using Hadoop. This paper describes the concept of reader-based trolley for supermarket automation. The automation RFID will be included in each and every product, with a unique ID number so that the product can be recognized.

Fan W, Bifet A (2013). Mining big data, current status, and forecast to the future. ACM SIGKDD Explorations Newsletter This paper proposed a frame work which focus on improving the performance of MapReduce workloads and maintain the system. DHSA will focuses on the maximum utilization of slots by allocating map (or reduce) slots to map and reduce tasks dynamically.

Journal of Student Research Fourth Middle East College Student Research Conference, Muscat, Sultanate of Oman "Data Mining Applications for Sales Information System Using Market Basket Analysis on Stationery Company" Tasks that are functional to data mining such as

Association the procedure of finding the relations between items.

Sequence same association but in more than one period.

Clustering the process of collecting similar data to groups, each group has similar data.

Classification makes a class of an object based on its concept.

Regression the method of approximating the value expected by the patterns based on the dataset.

The solution: a procedure of finding a problem and explaining it to give some information to decision making.

Market basket analysis is an association in data mining to find elements that appear in one time. This method can know the customer buying patterns by finding an association between different items in different invoices. The results from these methods can be used by retail shopping such as sales company/supermarket to improve a marketing plan through items that may be purchased concurrently by consumers.

The supermarket economic bases can be explained by 4 concepts of utility:

Place utility, time utility, ownership utility, shape utility

It can be understood from above part from a consumer perspective the supermarket offers him the products and services he needs in the required quantity, at required location of time.

Paper proposed Aug (2017) wageninga: According to the study proposed on this paper the dataset for measuring customer mind set the marketing performance customer chain the aspect customer mindset contains of measures brand equity, perceived quality, satisfaction and attitudinal loyalty. Dataset for measuring product market performance: Performance outcomes achieved in the market place in which product is offered it consists of measure: unit, sales, revenue, premium, market place and new product success.

A frame work proposed for computing fast and reliable data analysis and mining feedbacks (Rodriguen & Chiplunleas 2018): They gave real time twitter input in to generate for getting results of the analysis to generate fast feedback through sentiment analysis.

Sones yildim jan,27 2021 The main purpose of this article is to demonstrate various pandas functions that help to analyze tabular data the parameters of pandas functions are highly important as they make the functions more powerful and versatile.

David's Stepanick Aug 6, 2021 Retail data analytics how data science improves retail market. Investing in retail data analytics is essential for any business that wants to understand their customer needs better which products are in true demand exclusively over a specific period of time. In 2016 has a huge focus on shopper loyalty with consumers. Two distinct types of data Scan data that is collected in store when items are sold or scanned at the checkout. Panel data adds more depth to the data such as age of customer.

1.2 Motivation

Until few years ago, retail industry focused only on marketing and customer service. Now, the focus is on collecting data, analysing it, and then using the conclusions attained from the analysis to improve the marketing and customer service strategies. Data Science in retail helps to protect the company's reputation. It helps in personalizing the customer experience. Businesses collect customer data from many different channels, including physical retail, e-commerce, and social media so that it can observe the customers interest and can set target communications. Tableau is a data visualization tool which can handle millions of rows of data with ease. Different types of visualization can be created with a large amount of data without impacting the performance of the dashboards. Also, there is an option in Tableau where the user can make "live" to connections to different data sources like SQL etc. Supermarket data analysis and visualization These points of the domain data science and the data visualization

tool tableau motivated us to take up the project in order to achieve our project's objectives that are stated in the upcoming chapters.

1.3 Problem Statement

To design and develop a model to analyze Big Mart/Supermarket data and visualization using suitable data science tools and techniques.

1.4 Objectives of Proposed Work

- To analyze, predict, generate and visualize the various reports based on different parameters using python libraries.
- To analyze, predict, generate and visualize the various reports based on different parameters using data visualization tool like tableau.
- To study the performance of data visualization tool like tableau for various datasets.

2.0 Proposed Methodology

The proposed methodology for the results and the visualization to the parameters considered is given below. Initially the datasets required for supermarket data analysis and visualization were collected, then the datasets were uploaded to the google collab after the uploading was completed the results were obtained with respect to our queries. Then the dataset was uploaded in the data visualization tool that is Tableau and the results were obtained in the form of charts, graphs and reports. Datasets The datasets that are used are collected from Kaggle i.e., <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data> The dataset that we have used for the testing the performance of the data visualization tool Tableau is collected from Kaggle. i.e., <https://in.docworkspace.com/d/sIMyc48xn7cGmlQY>. The main dataset used in the project includes the attributes consists of the attributes such as ItemIdentifier, ItemFatContent, Item_Visibility, Item_Type, Item_MRP, Outlet_Identifier, Outlet_Establishment_Year, Outlet_Size, Outlet_Location_Type, Outlet_Type, Item_Outlet_Sales.

2.1 Data Exploration

A quantitative study was conducted to build a model which visualizes the results of the queries. Here some of the datasets that are included for the data analysis. The model includes the phases such as collection of the datasets, data pre-processes, outputs. Block diagram data cleaning & charts & Datasets data analyzing graphs.

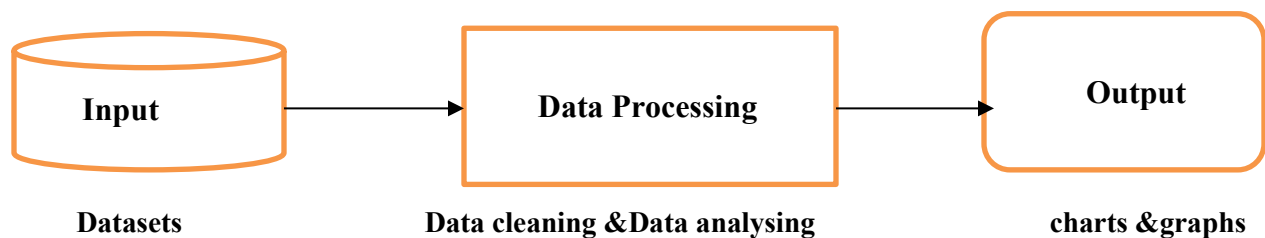


Fig. 1 Block Diagram of Proposed Model

2.2 Data collection

Data collection was an essential and protracted process. The accuracy of the data collection is essential to maintain cohesion. This is the dataset in the retail and FMCG sector using Pandas, the various datasets that are collected from Kaggle website <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data>

select=Train.csv. Our dataset has 8523 rows and 12 columns.

2.3 Data Pre-processing

Data pre-processing is an important process in development of data science model. We need to import some useful libraries that will help us to import the dataset into python environment, manipulate and analyze the same. The data collected is often loosely controlled with out-of-range value, missing value etc. imputation of missing values in our data missing values have been handled by using simple imputer from python package.

2.4 Data Cleaning

Data Cleaning is the process that removes data that does not belong in the dataset, while the methods or techniques used for data cleaning may vary according to type of data the company stores. Data cleaning includes removal of duplicates or irrelevant data, fixing structural errors filtering unwanted outliers, handling missing data and validation.

2.5 Data Analysing

Data analysing is a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, conclusions and supporting decision making from the datasets.

2.6 Output

After the data is being passed through the phases of pre-processing, cleaning, and analysing various outputs are produced in accordance to our queries. The outputs are obtained in the form of graphs, reports, and charts.

2.7 Tools and Technologies

2.7.1 Python Libraries

The proposed methodology incorporates various tools such as: Python Libraries, Tableau, and methods to investigate the business dataset of the supermarket store.

Python is the most widely used programming language today. When it comes to solving data science tasks and challenges, Python never ceases to surprise its users. Most data scientists are already leveraging the power of Python programming every day. python is an easy-to-learn, easy to-debug, widely used, object-oriented, open-source, high-performance language, and there are many more benefits to python programming. Python has been built with extraordinary Python Supermarket data analysis and visualization libraries for data science that are used by programmers every day in solving problems. Here's the top 10 Python libraries for data science: The procedure incorporates the different python libraries for accomplishing the objectives or the destinations.

The libraries include:

NumPy: (Numerical Python) is the fundamental package for numerical computation in Python; it contains a powerful N-dimensional array object. It has around 18,000 comments on GitHub and an active community of 700 contributors. It's a general-purpose array-processing package that provides high-performance multidimensional objects called arrays and tools for working with them. NumPy also addresses the slowness problem partly by providing these multidimensional arrays as well as providing functions and operators that operate efficiently on these arrays. NumPy is an essential bundle which is utilized for logical processing in python. It is the library in python that gives multidimensional cluster object, different inferred objects and an arrangement of schedules for quick procedure on exhibits, including numerical, legitimate, arranging, selecting, basic straight variable-based math, essential factual tasks, arbitrary reproduction and considerably more. Vectorization portrays the shortfall of any express circling, ordering etc., vectorized code is a lot more straightforward to read, it has not many lines of code

and thus few bugs. NumPy completely upholds an item situated approach, many of its techniques are reflected by capacities in the peripheral NumPy name space, which permits the developer to code in whichever worldview they like.

Pandas: Pandas is quick and it has superior execution and efficiency for clients. Pandas are by and large utilized for information science yet have you asked why? This is on the grounds that pandas are utilized related to different libraries that are utilized for information science. It is based on the highest point of the NumPy library which implies that a great deal of constructions of NumPy are utilized or imitated in Pandas. The information created by Pandas are regularly utilized as contribution for plotting elements of matplotlib, factual investigation in SciPy, AI calculations in scikit learn.

Matplotlib: Matplotlib has powerful yet beautiful visualizations. It's a plotting library for Python with around 26,000 comments on GitHub and a very vibrant community of about 700 contributors. Because of the graphs and plots that it produces, it's extensively used for data visualization. It also provides an object-oriented API, which can be used to embed those plots into applications. It is an astounding perception library in Python for 2D plots of exhibits. Matplotlib is a multi-stage information perception library based on NumPy exhibits and intended to work with the more extensive SciPy stack.

Perhaps the best advantage of representation is that it permits us visual admittance to enormous measures of information in effectively absorbable visuals. Matplotlib comprises of a few plots like line, bar, dissipate, histogram and so forth Matplotlib accompanies a wide assortment of plots. Plots assists with getting patterns, designs, and to make connections. Supermarket data analysis and visualization

Count plot: A count plot is useful when managing straight out qualities. It is utilized to plot the recurrence of the various classes. The section sex contains downright information in the titanic information, i.e., male and female

KDE Plot: A Kernel Density Estimate (KDE) Plot is utilized to plot the circulation of consistent information. Circulation plot A Distribution plot is like a KDE plot. It is utilized to plot the circulation of consistent information.

Disperse plot: Disperse plots assist with understanding co-connection between information.

Jointplot: A Joint Plot is likewise used to plot the relationship between information.

Pair plots: Seaborn allows us to plot different disperse plots. It's a decent choice when you need to get a fast outline of your information.

By removing significant experiences from crude information, general store information investigation assists with planning better methodologies for expanding incomes and lessening costs. Both are significant for amplifying deals limit in the exceptionally serious commercial center. Store information examination, a major information and information science-based Supermarket data analysis and visualization methodology is tied in with gathering and concentrating on client information to find client inclinations and distinguish patterns. These are crucial to settling on more exact business choices. Taking an essential, information driven gander at retail exchanges yields priceless experiences into purchaser needs. It additionally informs a great deal regarding the exhibition of stores, items, and sellers. General store information investigation can furnish retailers with data about client information, similar to item look through requests. An inside and out investigate what items (or sets of items) clients look for online assists them with finding precise client needs, even in the tightest portions. With the key experiences, retailers can seriously extend the scope of their items to incorporate those that will really sell. You will know what items (or item goes) are sought after and consequently have

the option to refine your contribution as needs be. Revelations to be made after examination Relation of clients with Super Market Products connection with amounts. Types of items and their deals. Products and their evaluations.

Tableau: Tableau is a data visualization software that is packed with powerful graphics to make interactive visualization. The most important aspect of tableau is its ability to interface with databases, spreadsheets, OLAP (online analytical processing) cubes, etc. it also has the ability to visualize geographical data and for plotting longitudes and latitudes in maps.

1. Tableau can extract data from a database like pdf, excel, text documents, R, Hadoop, Python, or SAS to cloud databases like Flipkart, Google sheet, Netflix, Amazon.
2. The data is dragged off to the data engine of Tableau, also called the Tableau desktop. Here, the business analyst works on data, generates a dashboard, and shares it with the user, where the user reads this on the screen called Tableau Reader.
3. Data is published with various supported features like collaboration, models of security, automation, distribution, etc.
4. In the end, the user will be able to download a visualized data file on emails, desktop, or mobile. (Understand the basics and types of data visualization in Business Analytics).

Google Collaboratory: Google Collab is an online notebook-like coding environment that is well-suited for machine learning and data analysis. It comes equipped with many Machine Learning libraries and offers GPU usage. It is mainly used by data scientists and ML engineers.

2.8 Implementation of Test Cases

The following are the snippets of the code used in the implementation.

Snippet for outlet sales versus overall item outlet sales

```
outlet_type_sales = df.groupby("Outlet_Type") ["Item_Outlet_Sales"]. sum ()  
pd.set_option ('display. float format', lambda x: '%.3f % x') #converting from scientific notation to numerical format  
outlet_type_sales.sort_values(by=['Item_Outlet_Sales'], ascending=[False]). reset_index ()
```

snippet for item outlet sales versus outlet location type

```
pd.set_option ('display. float format', lambda x: '%.3f % x')  
location_type_sales = df.groupby("Outlet_Location_Type") ["Item_Outlet_Sales"]. Sum ().  
reset_index ()  
location_type_sales.sort_values(by=['Item_Outlet_Sales'], ascending=[False])
```

snippet for item outlet sales versus outlet size

```
pd.set_option ('display. float format', lambda x: '%.3f % x')  
outlet_size_sales = df.groupby("Outlet_Size") ["Item_Outlet_Sales"]. sum (). reset_index ()  
outlet_size_sales.sort_values(by=['Item_Outlet_Sales'], ascending=[False])
```

snippet for item category versus overall sales

```
df.groupby("Item_Type") ["Item_Outlet_Sales"].max (). sort_values(by=['Item_Outlet_Sales'],  
ascending=[False]). reset_index ()
```

snippet for item visibility versus its overall sales

```
pd.set_option ('display. float format', lambda x: '%.3f % x')  
item_visibility_sales = df.groupby("Item_Visibility") ["Item_Outlet_Sales"]. sum ().  
sort_values(by=['Item_Outlet_Sales'], ascending=[False]). reset_index ()  
item_visibility_sales.sort_values(by=['Item_Outlet_Sales'], ascending=[False])
```

snippet for weight of the item versus overall sales

```
pd.set_option ('display. float format', lambda x: '%.3f % x')
```



```
item_visibility_sales = df.groupby("Item_weight") [["Item_Outlet_Sales"]].sum ()  
sort_values(by=['Item_Outlet_Sales'], ascending=[False]).reset_index ()  
item_visibility_sales.sort_values(by=['Item_Outlet_Sales'], ascending=[False])  
snippet for overall sales of each product in tier 1 cities when compared to tier2 and tier 3 cities  
df_sales=pd.concat ([df_tier1_list, df_cities_list], axis=1)  
df_sales
```

3.0 Results and Discussion

The proposed work has been implemented using python libraries, tableau the various parameters are considered to analyze, predict, generate and visualize the various reports based on the analysis of two datasets downloaded from Kaggle.

The various parameters considered are:

- To analyze if outlet type has impact on overall sales.
- To predict the locations which make most sales.
- To know whether the outlet size has an impact on overall sales.
- Analysing which category of products sells the most and least.
- Inspect if visibility and weight have impact on sales of the products.
- To know the average MRP of products that sells the most and the least, and to which category it belongs.
- Checking better selling of products in tier 1 cities as compared to tier 2 and tier 3 cities.

3.1 Results using Python Libraries

While using platform like google colab, we need to connect our google drive and store data or directly provide path of the file location to pandas.

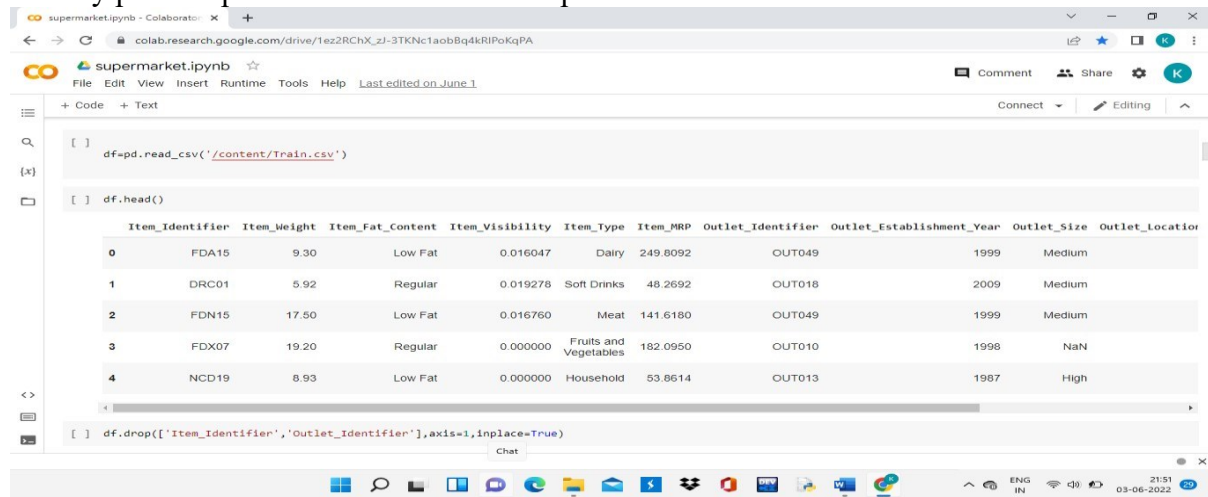


Fig. 2 The dataset which has been used for the data analysis

The following output shows that only columns have missing values in them, namely, Item_weight and the outlet_size.

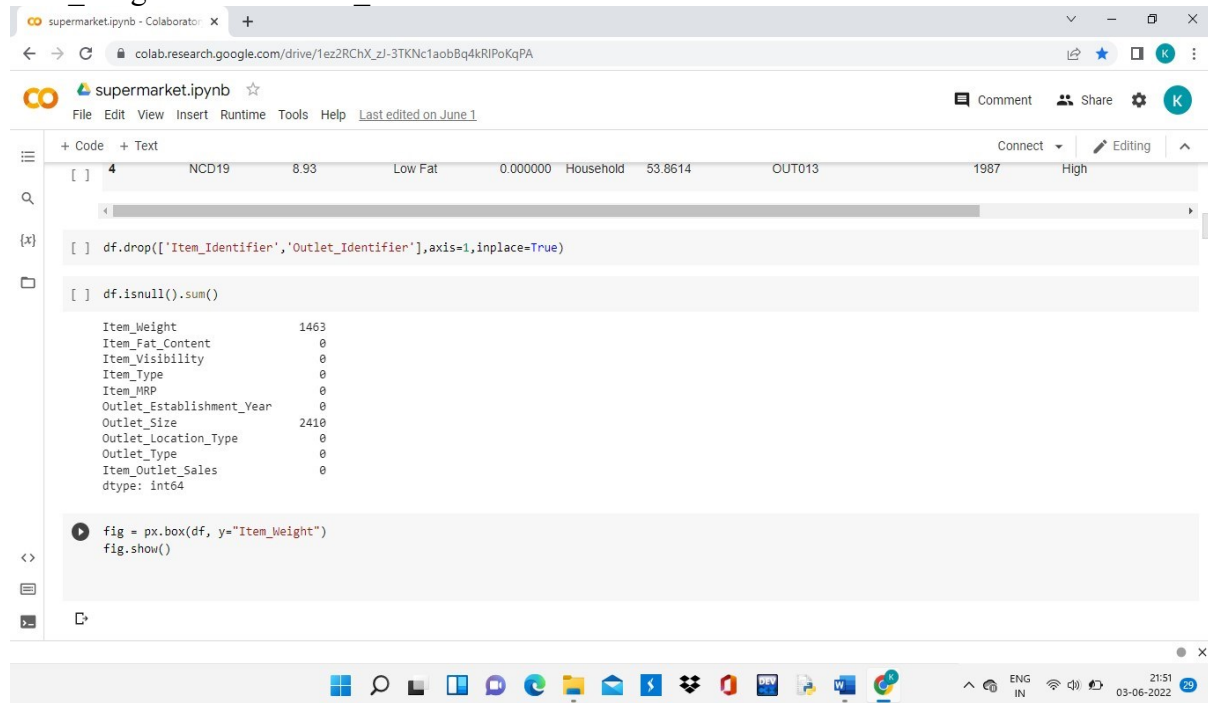


Fig. 3 list of columns having missing values

The figure below is the boxplot of Item_weight to check for the presence of outliers.

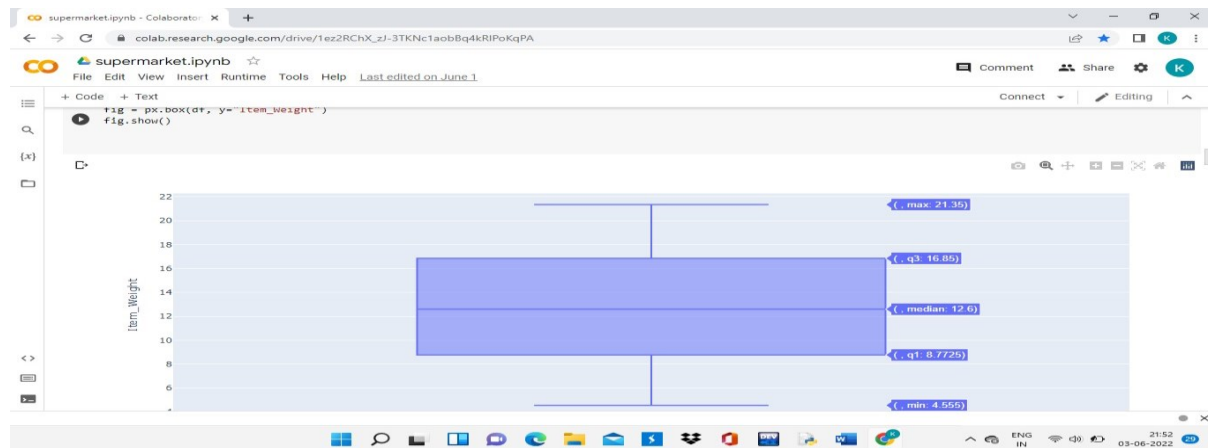


Fig. 4 Boxplot to check the outliers

The screenshot is the plot self-explanatory; we can observe the outlet type verses the overall sales.

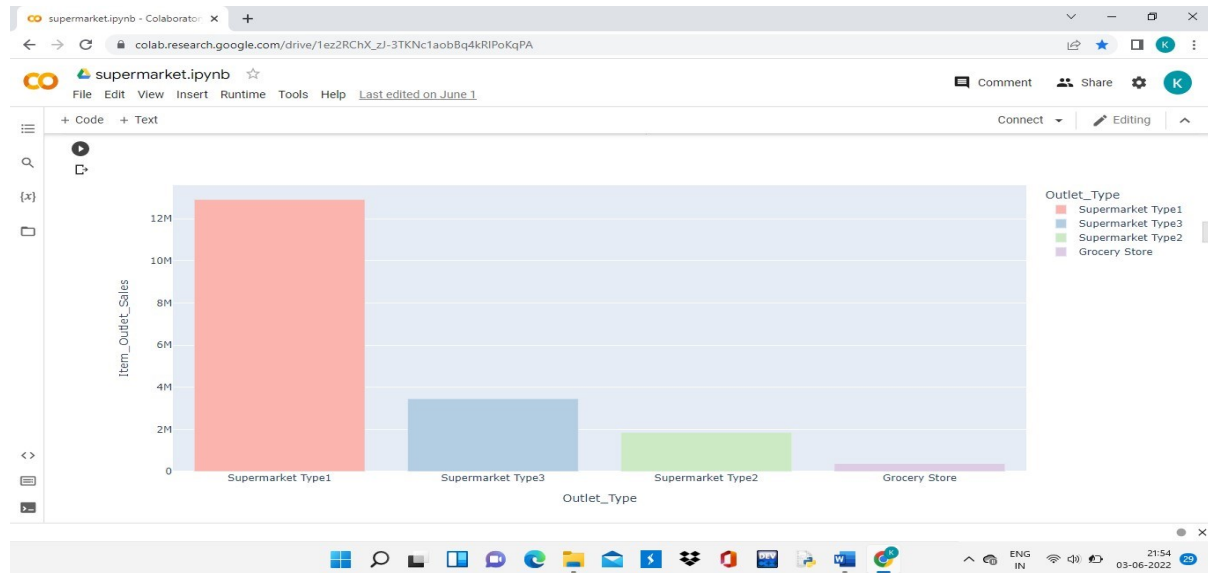


Fig. 5 Outlet_type versus overall Item-outlet_sales

The figure below gives the sales per outlet location by grouping the data with respect to the outlet_location and aggregating the corresponding sales, and tier3 cities perform the best.

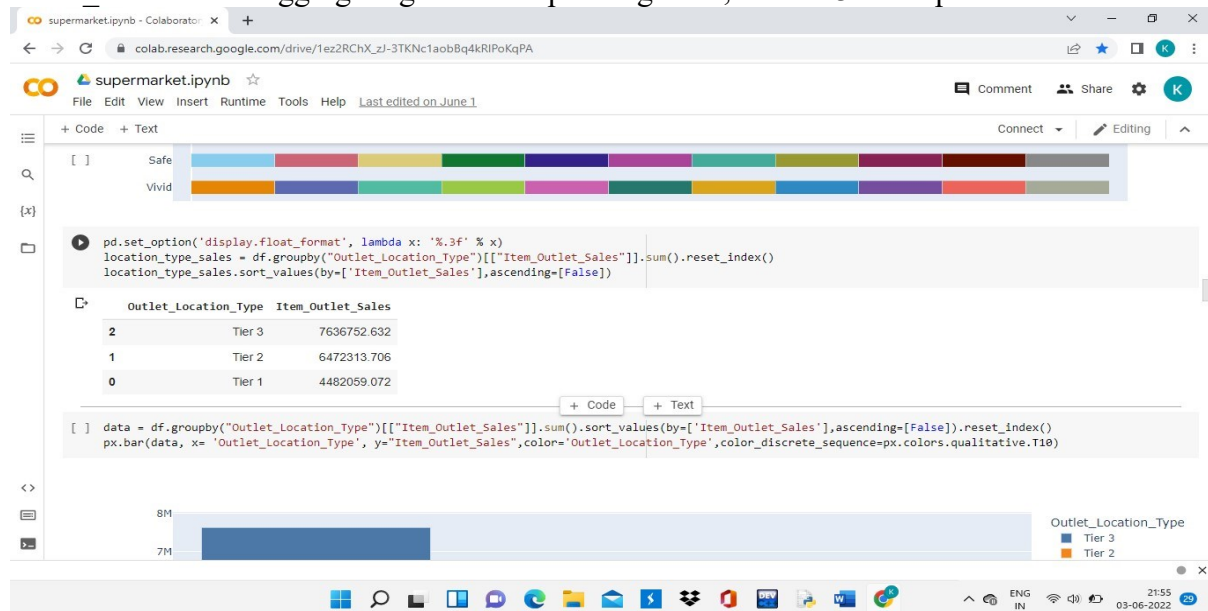


Fig. 6 outlet_location versus overall_sales

The chart below is the representation of the Item_outlet_sales verses the outlet_location_type.

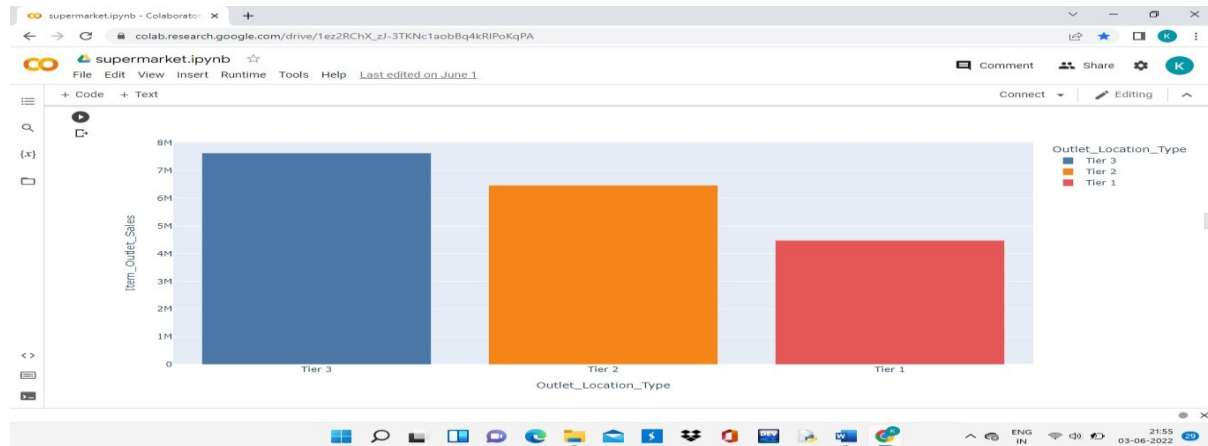


Fig. 7 Item_outlet_sales versus the outlet_location_type

The figure below is the output obtained on comparing Item_outlet_sales to the outlet_size, and medium-sized outlets have outperformed small and high sized outlets.

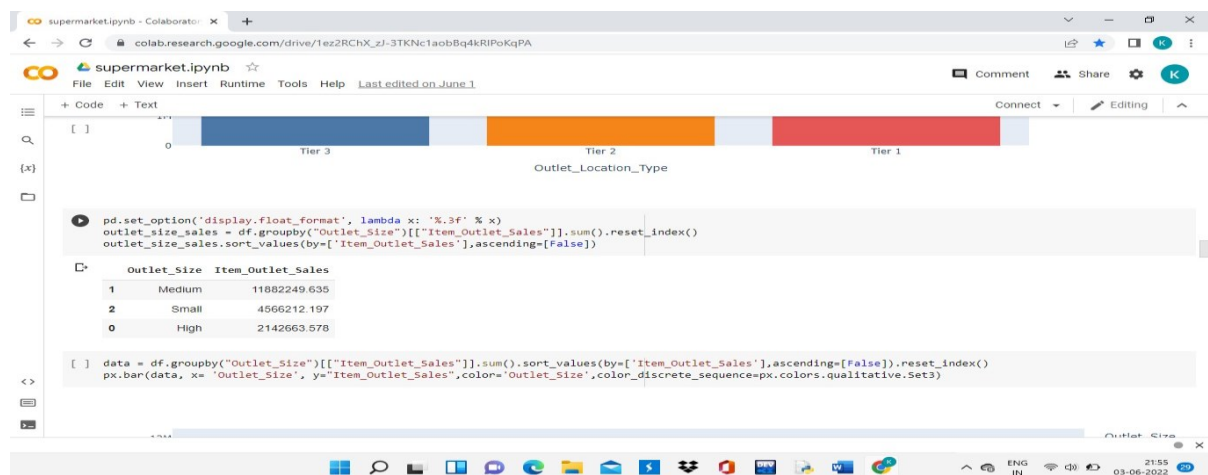


Fig. 8 Item_outlet_sales to the outlet_size

The below screenshot of the plot result of the outlet_size verses the overall sales.

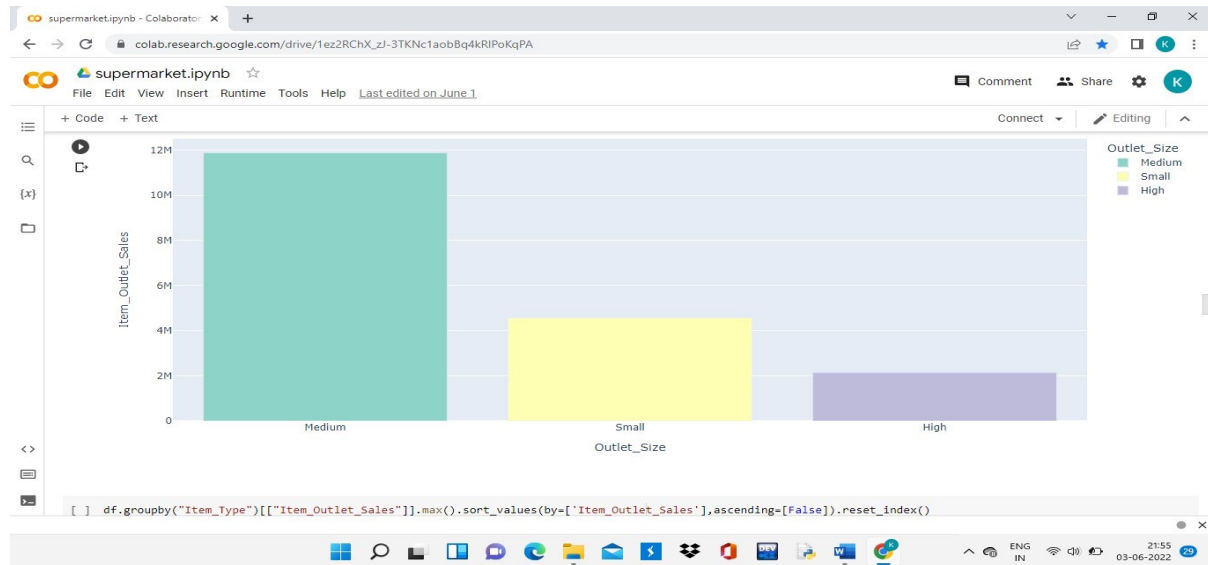


Fig. 9 plot representation of Item_outlet_sales to the outlet_size

The below figure is the result obtained when Item_category is compared with the overall sales, and we observe that Household item category has sold the most.

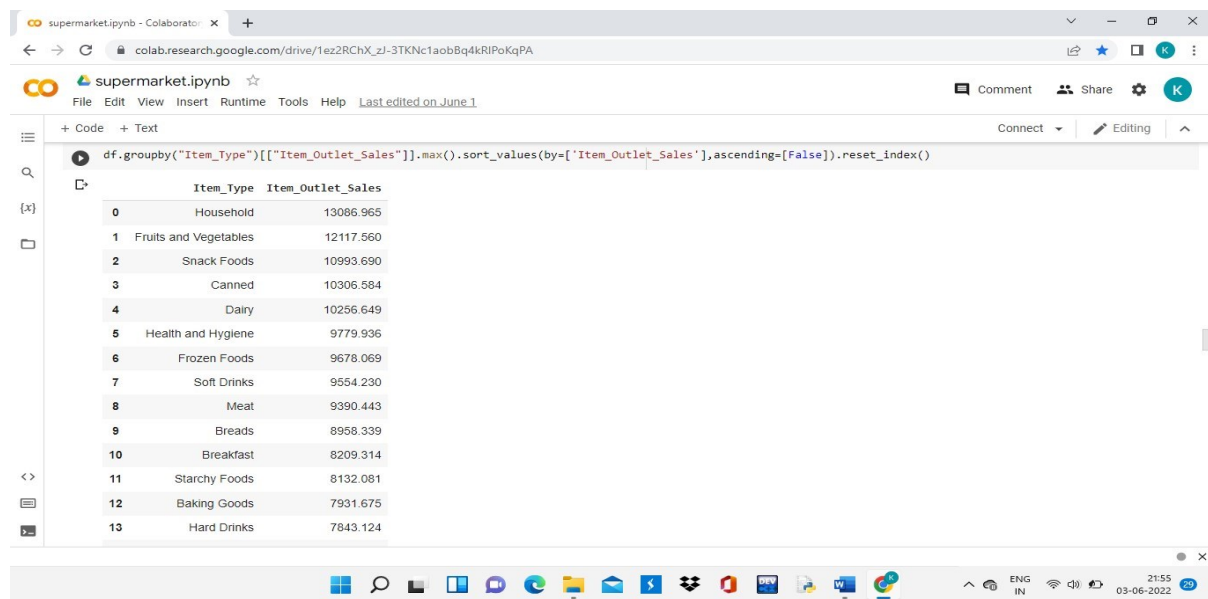


Fig. 10 Item_category versus the overall sales

The below is the plotly result obtained on comparing Item_category with the overall sales.



Fig. 11 plotly result of Item_category versus the overall sales.

The below plot is scatter plot as Item_visibility has a total of 7880 rows, in such cases bar graphs are not suitable plotting option. And we observe that visibility between 0.005-0.18 make most sales and visibility with more than 0.2 make least sales.



Fig. 12 Item visibility versus its overall sales.

The screenshot below shows the result of the Item_weight verses the overall sales, and the product with the weight of 12.85 units has made most sales and item with 9.1 has least.



Fig. 13 weight of the items versus their overall sales.

The figure below gives the results of the item type for the Item_MRP making highest sales and the lowest sales. The items with MRP 244.996 make highest i.e. the "Fruits and Vegetables" and "Household", and "soft drinks" with MRP 31.290 make the lowest sales.

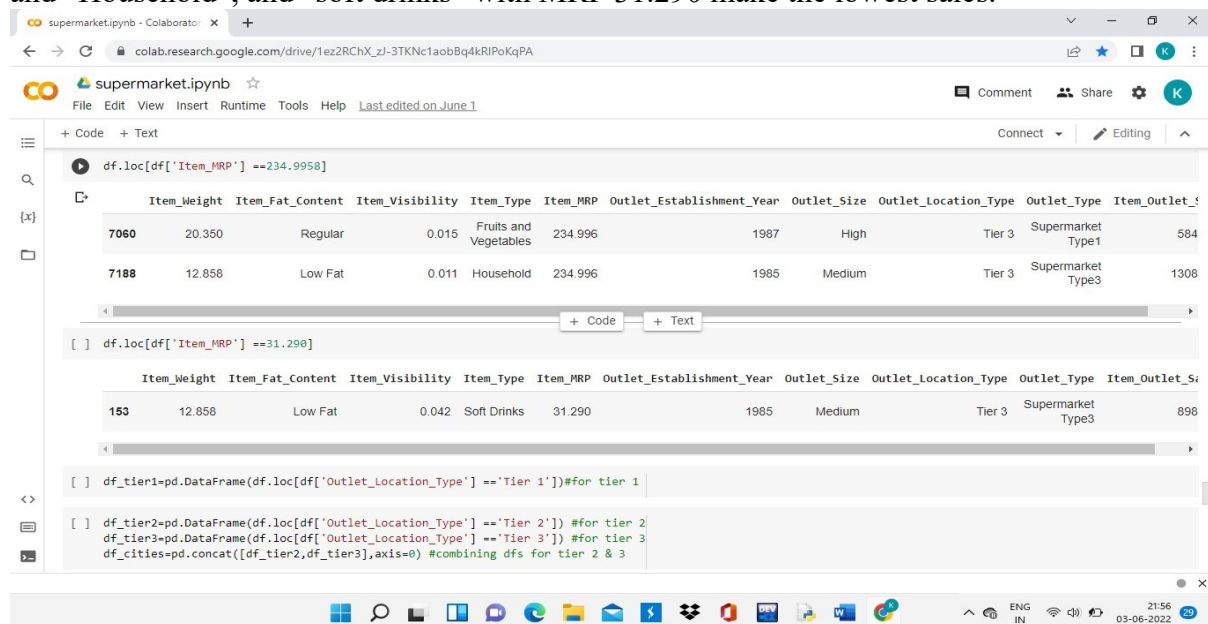
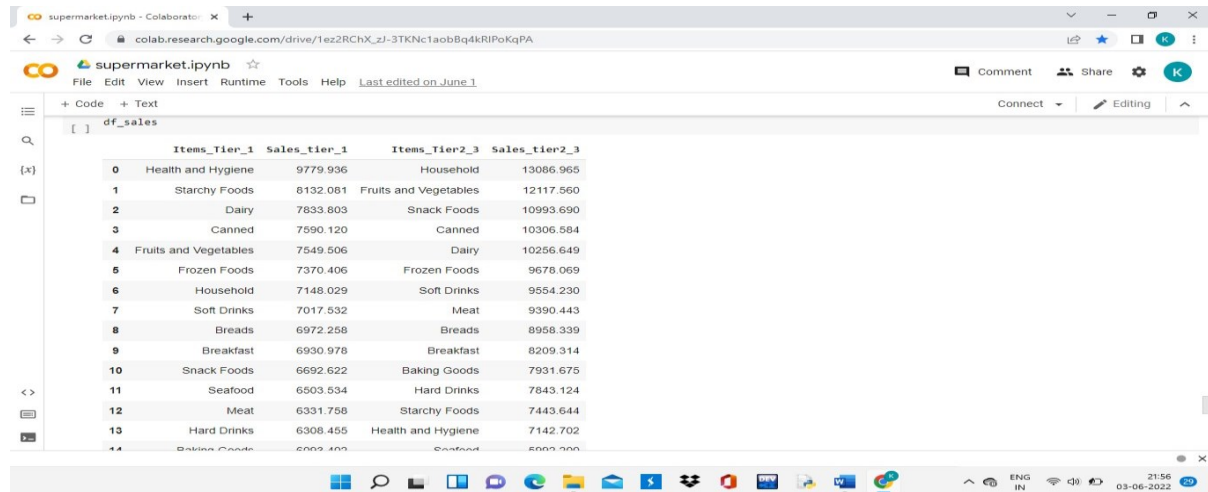


Fig. 14 Highest and lowest selling items with their MRP.

The screenshot below gives the information on how each product is performing for tier1 cities.



	Items_Tier_1	Sales_tier_1	Items_Tier_2_3	Sales_tier_2_3
0	Health and Hygiene	9779.936	Household	13086.965
1	Starchy Foods	8132.081	Fruits and Vegetables	12117.560
2	Dairy	7833.803	Snack Foods	10993.690
3	Canned	7590.120	Canned	10306.584
4	Fruits and Vegetables	7549.506	Dairy	10256.649
5	Frozen Foods	7370.406	Frozen Foods	9678.069
6	Household	7148.029	Soft Drinks	9554.230
7	Soft Drinks	7017.532	Meat	9390.443
8	Breads	6972.258	Breads	8958.339
9	Breakfast	6930.978	Breakfast	8209.314
10	Snack Foods	6692.622	Baking Goods	7931.675
11	Seafood	6503.534	Hard Drinks	7843.124
12	Meat	6331.758	Starchy Foods	7443.644
13	Hard Drinks	6308.455	Health and Hygiene	7142.702
14	Baking Goods	6003.400	Seafood	6003.400

Fig. 15 sales of each product in tier1 cities.

The below graph is the result obtained when each product's performance is compared with tier1 cities with tier2 and tier3 cities. And the result says that no item has performed better in tier1 cities than that of tier2 cities and tier3 cities.



Fig. 16 overall sales of each product in tier 1 cities when compared to tier 2 and tier 3 cities.

3.2 Results obtained by data visualization tool Tableau

The screenshots below are the results obtained when the main dataset was uploaded in the tableau tool.

There are few steps which need to be followed to work with tableau tool, the steps are as follows:

- Firstly, install the tableau desktop.
- Connect it to your data, i.e., upload your dataset to the tableau.
- Now drag and drop the attributes to the row and the column section to see the results.

The screenshot below corresponds to our first objective that is outlet types when compared to overall sales.

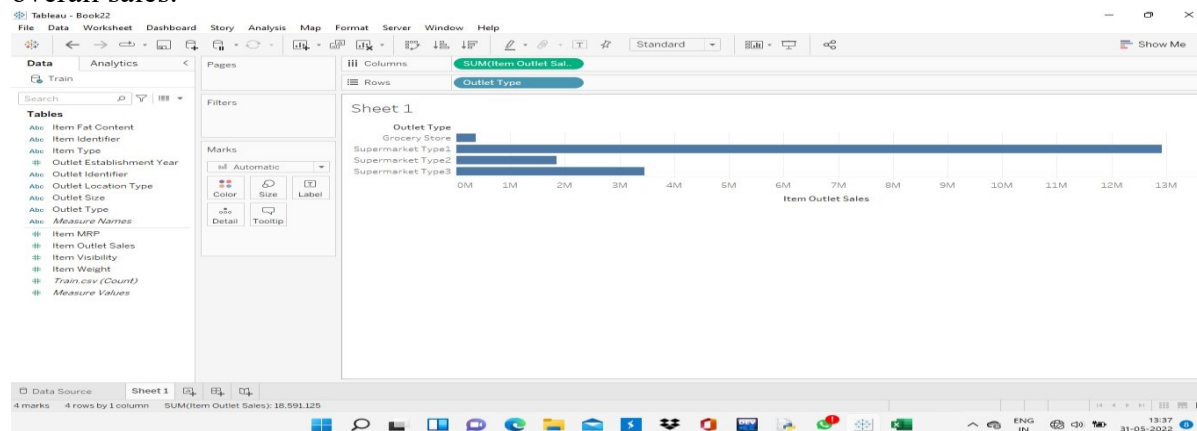


Fig. 17 outlet types versus their overall sales

The result obtained below corresponds to the second objective that is which location make

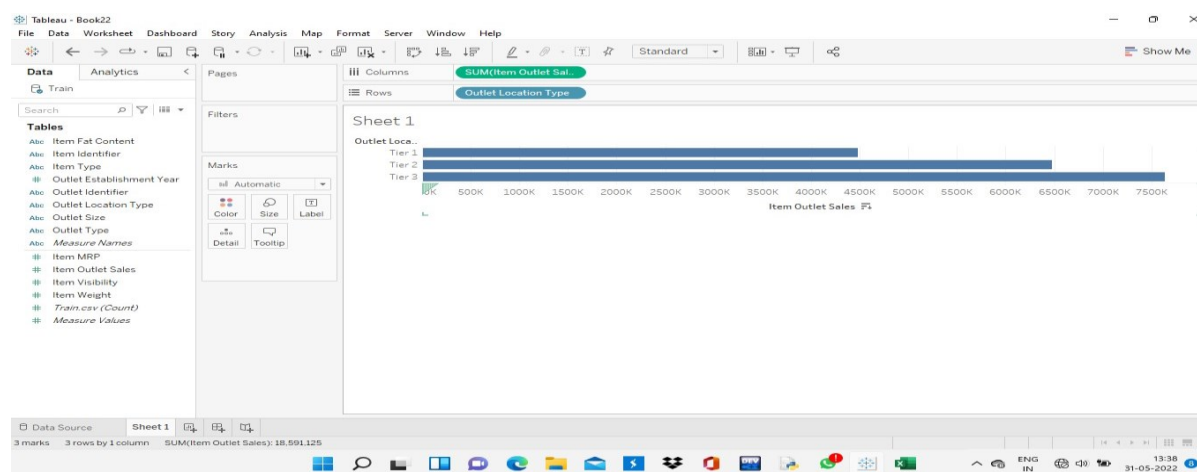


Fig. 18 cities that make the most sales

The result obtained below is the result for the objective that is whether the outlet_size has impact on sales.

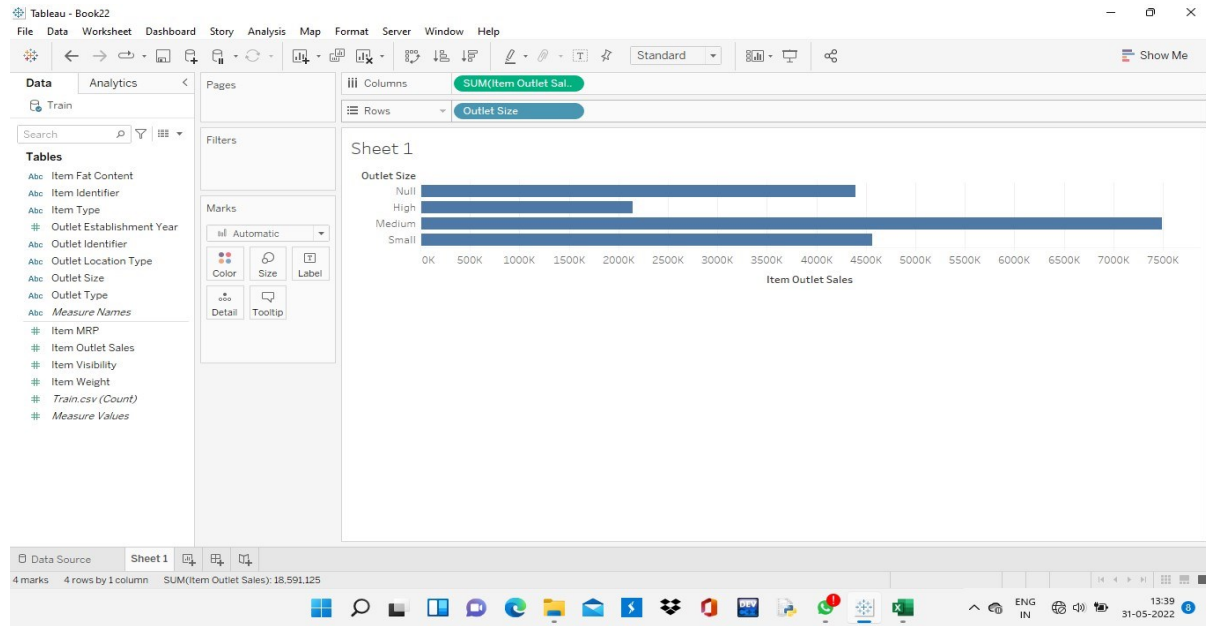


Fig. 19 Outlet size versus the overall sales

The below is the result obtained for the objective, the category of product which sells the most and the least.

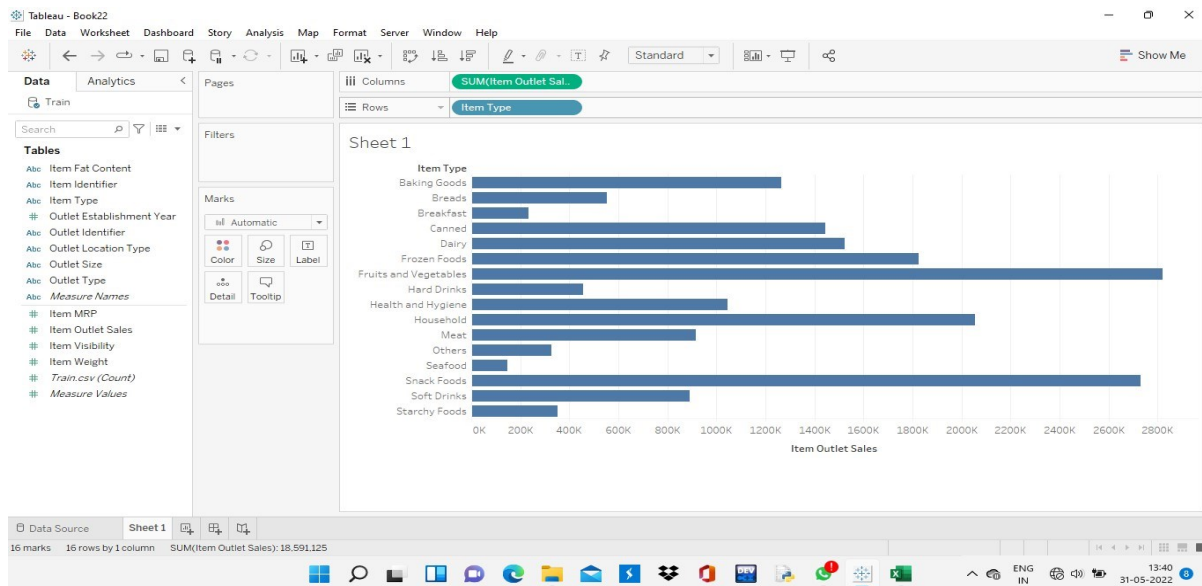


Fig. 20 Category of products which sells the most and the least

The below obtained table of result corresponds to the objective, visibility and weight has any impact on the overall sales.

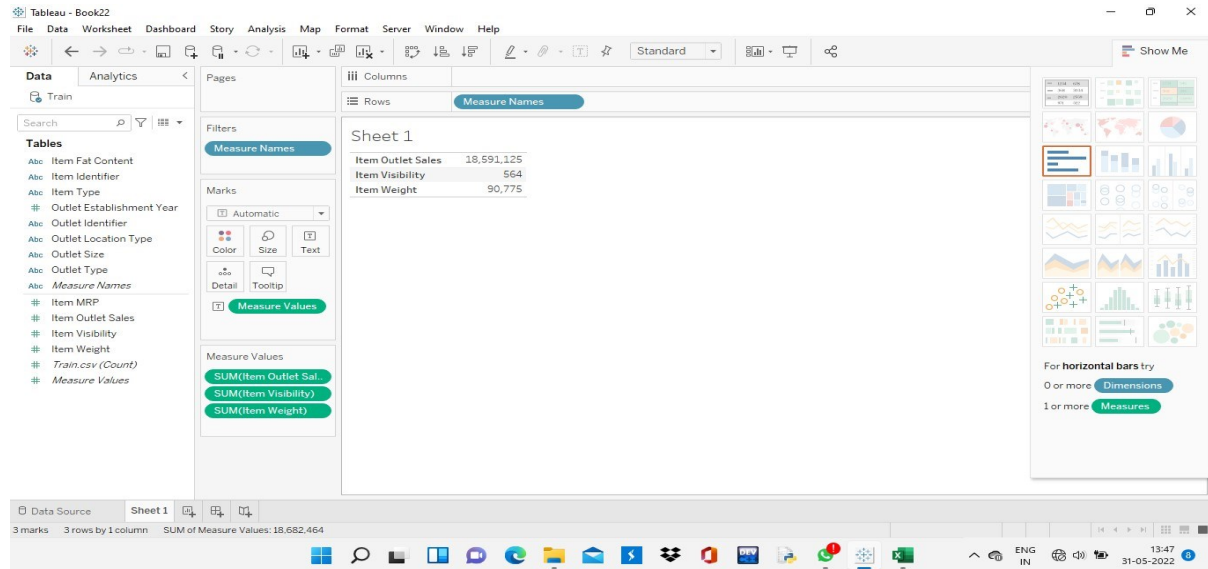


Fig. 21 Impact of Visibility and weight versus the overall sales of product

The screenshot below is the result to our objective, the average MRP of the products that sells the least.

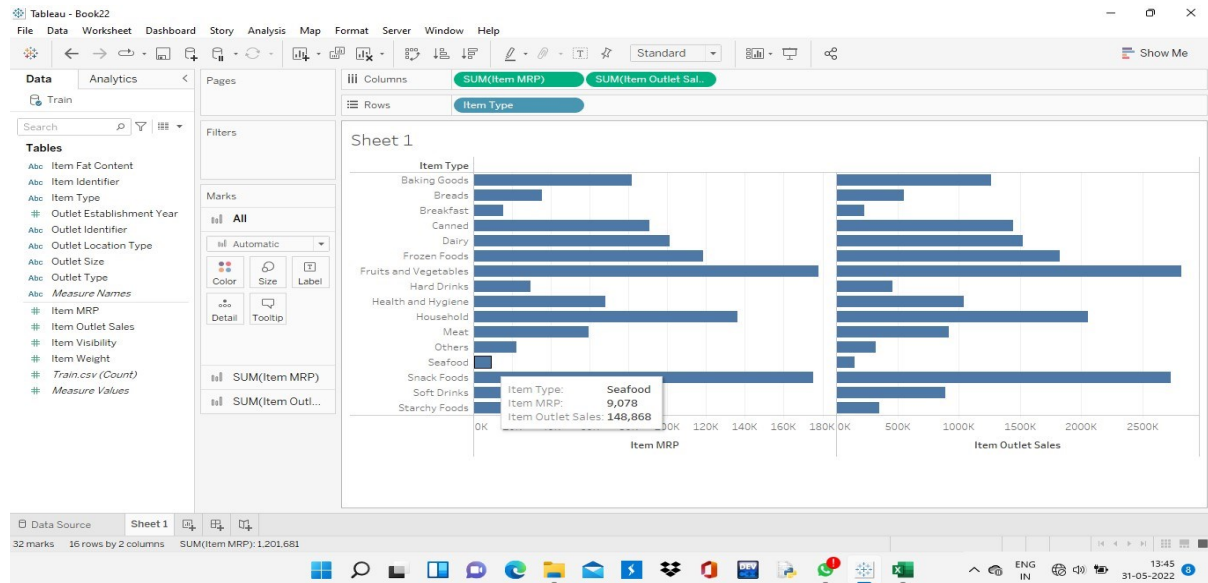


Fig. 22 Average MRP of the product that sells the least

The screenshot below is the result to our objective, the average MRP of the products that sells the most.

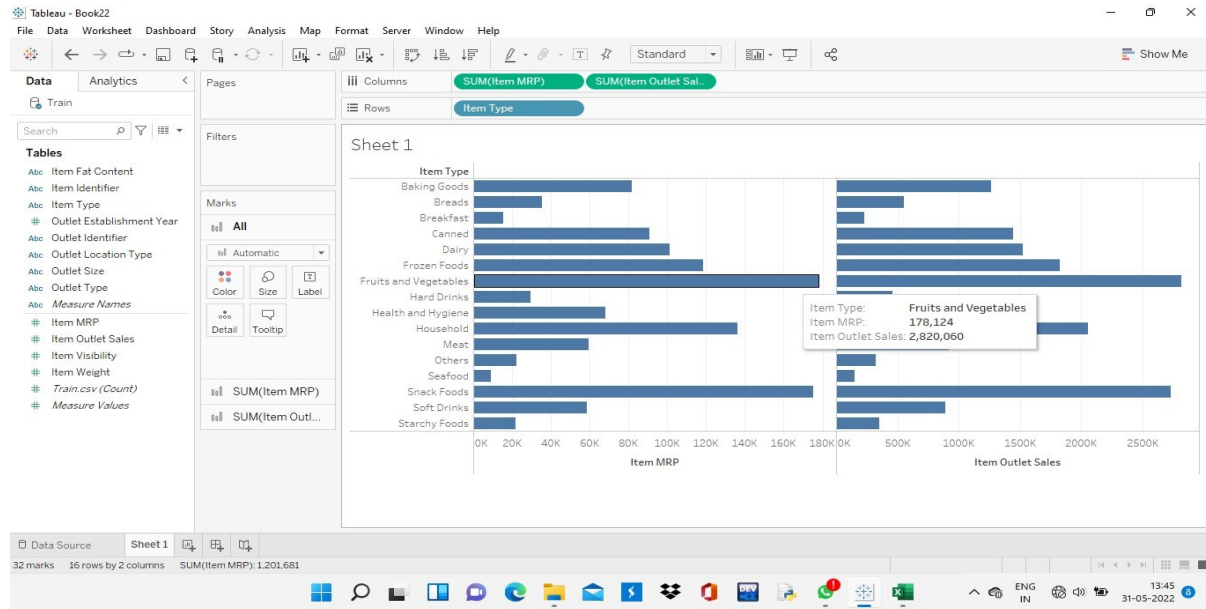


Fig. 23 Average MRP of the product that sells the most

The below given figure shows the result for the objective, the products which sell better in tier 1 cities when compared to tier2 and tier3 cities.

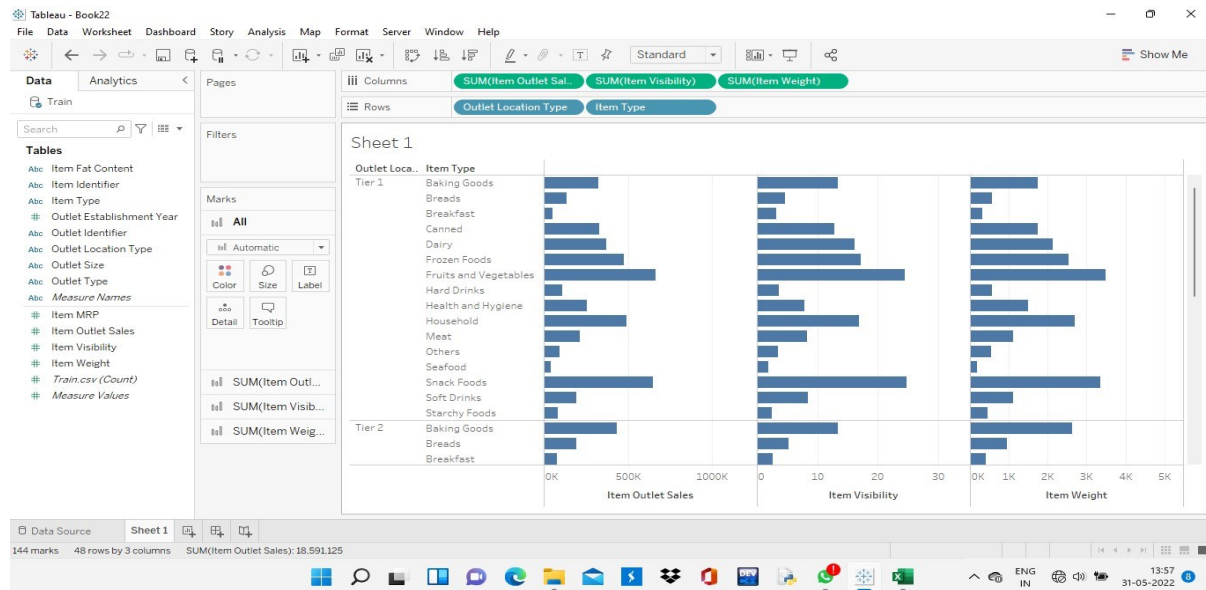


Fig. 24 Category of products sold better in tier 1 cities

The below is the result obtained for the objective, the products which sell better in tier2 cities when compared to tier1 cities.

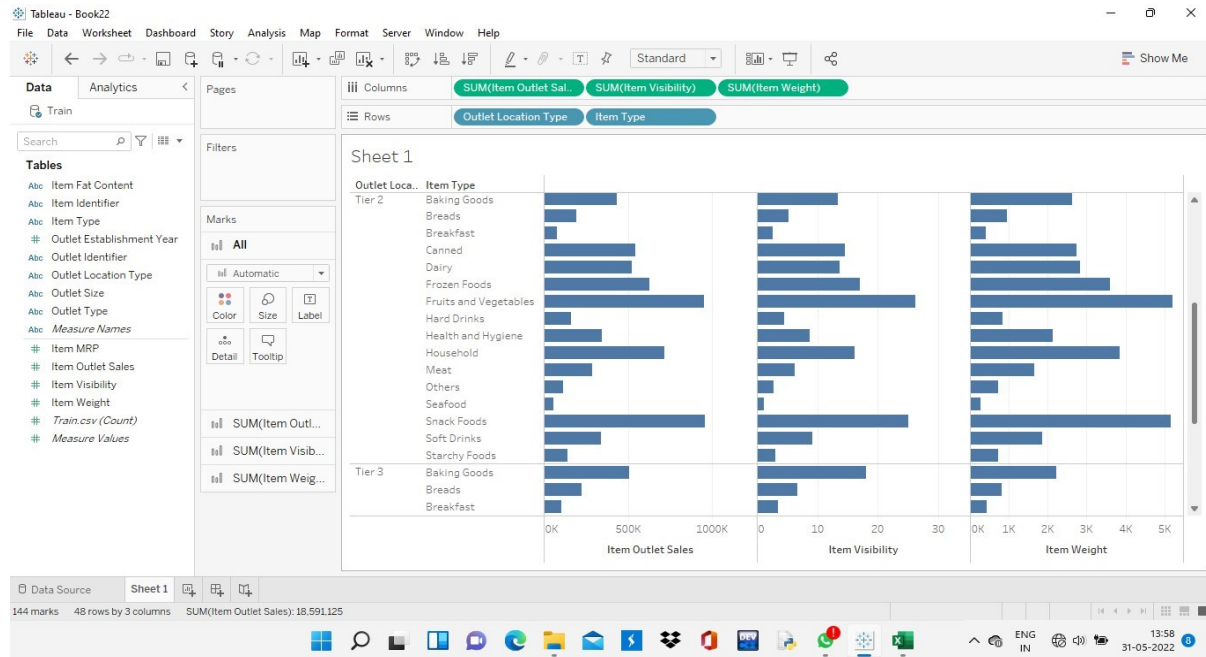


Fig. 25 Each Category of products sold in tier 2 cities

The below is the result obtained for our last objective, the products which sell better in tier3 cities when compared to tier1 cities.

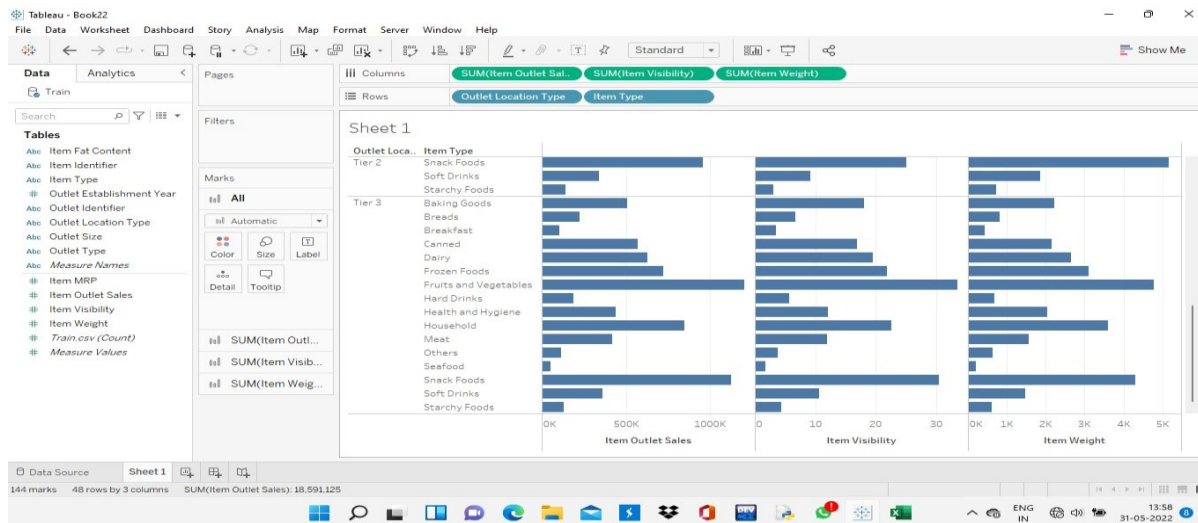


Fig. 26 Each category of products sold in tier 3 cities

4.0 Conclusion and Future Work

The proposed work helps in producing different kind of reports by Analysing the data set of super market using different data science tools such as: Python Libraries, Tableau. With the help of python library i.e., pandas, plotting library called plotly to ensure that the objectives of this project are achieved. The project helps to predict whether the terminologies such as type of outlet, locations, outlet size, weight, visibility has any impact on the overall sales and also analyze which category products are sold most and the least and checking where the products are most sold i.e., in tier1, tier2 or tier3 cities.

References:

1. Chirayath R Sathyamoorthi, Paul Mburu, An analysis of supermarket pricing: The case of selected supermarkets in Gaborone, Botswana March 2016 journal of management research 8(2):66 doi:10.5296/jmr.v8i2.9089
2. Grewal, P.D.S. (2014) A critical conceptual analysis computer engineering. IOSR journal of computer engineering, 16, 9-13. [doi.org 10.9790/0661-16210913](https://doi.org/10.9790/0661-16210913).
3. Martin Fassnacht, henning mohr, pricing luxury brands: specificities, conceptualization, and performance impact march 2013 doi:10.15358/0344-1369_2013_2_104
4. International Journal of Engineering research and technology (2018)
5. W, Fan, Albert Bifet Mining big data: current status, and forecast to the future January 2014. ACM SIGKDD Explorations Newsletter 16(1):1-5.
6. Journal of Student Research Fourth Middle East College Student Research Conference, Muscat, Sultanate of Oman 2019.
7. Daljeet Kaur, Jagroop Kaur Data mining in supermarket, number 8(2017), pp. 1945-1951.
8. Soner Yildirim supermarket data analysis with sql, January 27, 2021.