King Saud University
College of Computer and Information Sciences
Information Technology Department

TamimiMarket data analytic

IT465 Data analytics & Visualization

Phase #3

Prepared by

Group#3 : TamimiMarket	
Hana Saad Almalki	437200270
Sarah Mohammed Alsadan	437200565
Shrooq Abdulrahman Albahooth	437202434
Haifa Yousef Alkhudair	437200221

Supervised by

Dr. Ahlam abdulghni

1) Introduction

Tamimi Markets is one of the fastest-growing supermarket chains in Saudi Arabia, and Saudi shoppers named Tamimi Markets one of the Top 100 Saudi Brands. The vision of Tamimi is to be the premier one-stop online grocery destination in the kingdom. Also, its promise to deliver a friendly online shopping experience with the widest selection.

2) Data

Data source: Tamimi Market website (https://shop.tamimimarkets.com/)

The website of the Tamimi market has a lot of features and services such as ordering from the website with multiple payment methods including (cash on delivery, Mada, visa). Products are categorized based on the type. The main category in the navigation bar is Hot Deals, FRESH, FOOD & BEVERAGES, BABY, HEALTHY LIVING, HEALTH & BEAUTY, HOUSEHOLD, PETS each one of them has a subcategory. Each product has some information such as Name, Brand, Size, Price. Also, when select and navigate to the product page it appears at the bottom of the page recommendation products. The website shows weekly offers and the most purchased products. so, we decided to collect the data by web scraping from their website.

Choose Your Dataset	
Dataset Name	Tamimi Market Products
Dataset Size	21100
Link	Attached in the file
The goal from using this dataset	• Extract the relationship between category and price. • Determine which product category has the most offers. • Determine what is the brand that gives more offers. • Determine what are the average prices of baby products. • Determine what are the least expensive brand for pet products

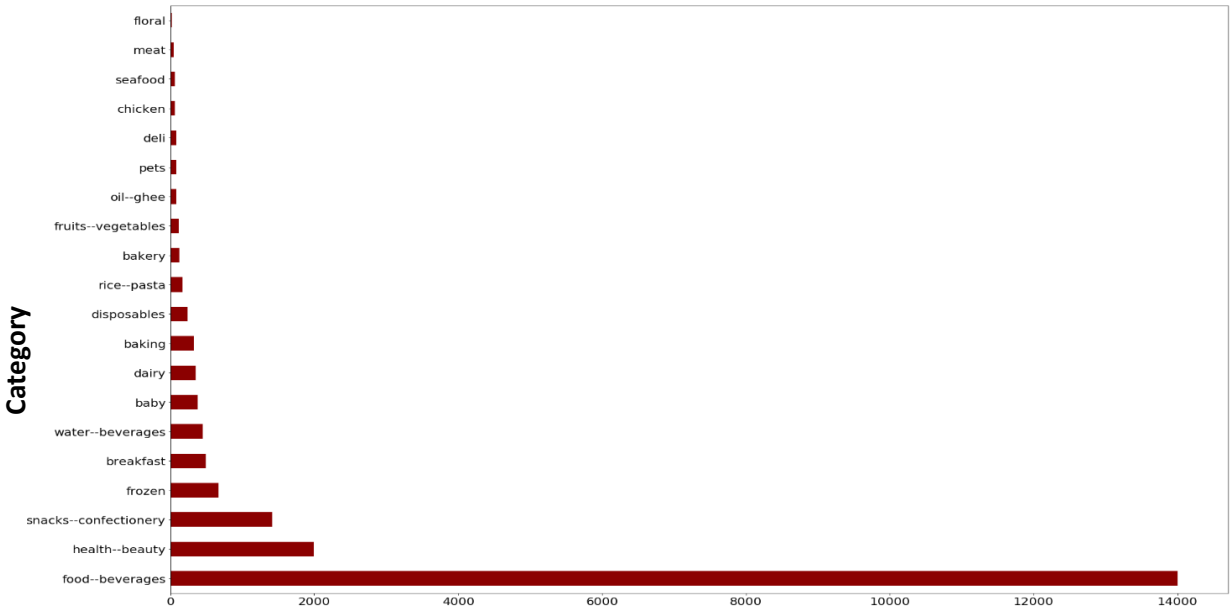
❖ Meta Data:

Attribute	Description	Quantitative vs qualitative	Data Type
ID	Every product has unique value	Qualitative	Nominal
Prodect_Name	Define the product by (not unique) name	Qualitative	Nominal
Price	The price of the product is represented in the Saudi riyal currency	Quantitative	Continuous
Category	Department of product in the supermarket, there are 21 Departments	Qualitative	Nominal
Brand	Shows the company of the product	Qualitative	Nominal
Offers	Is the product among the discounts?	Quantitative	Nominal

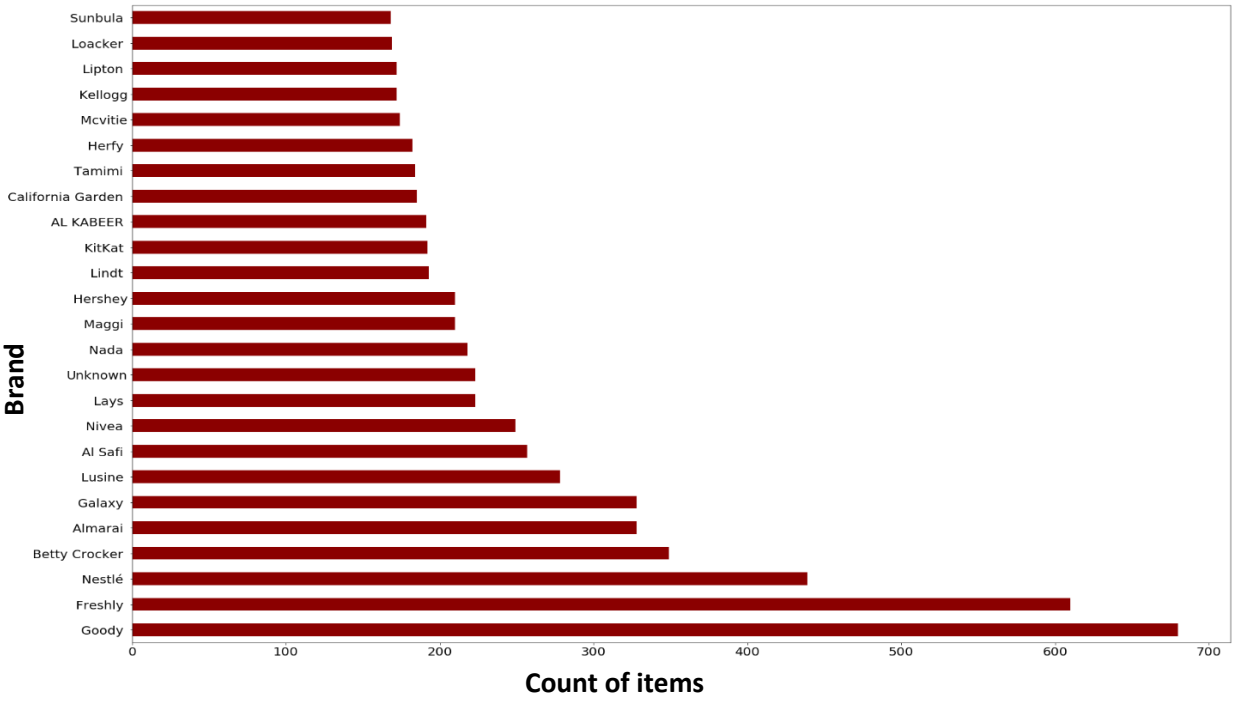
3) Cleaning and Preprocessing

1. Clean products category from extra words and unify them.
2. Clean products name from extra words and unify them.
3. Clean products price from extra characters and convert them to float type.
4. Clean products brand from extra words and unify them.
5. Clean products offer from extra characters and convert them to float type and unify products that do not have offers by setting the value of 0.0 for them.
6. Handle missing price value by deleting one row as it does not affect the data set.
7. Handle missing brand value by setting it as "unknown" brand because the brand is unknown and not specified or classified.
8. Add new feature "Different" have the original price "offer" then the price after offering "price" and how much of money will customer save.

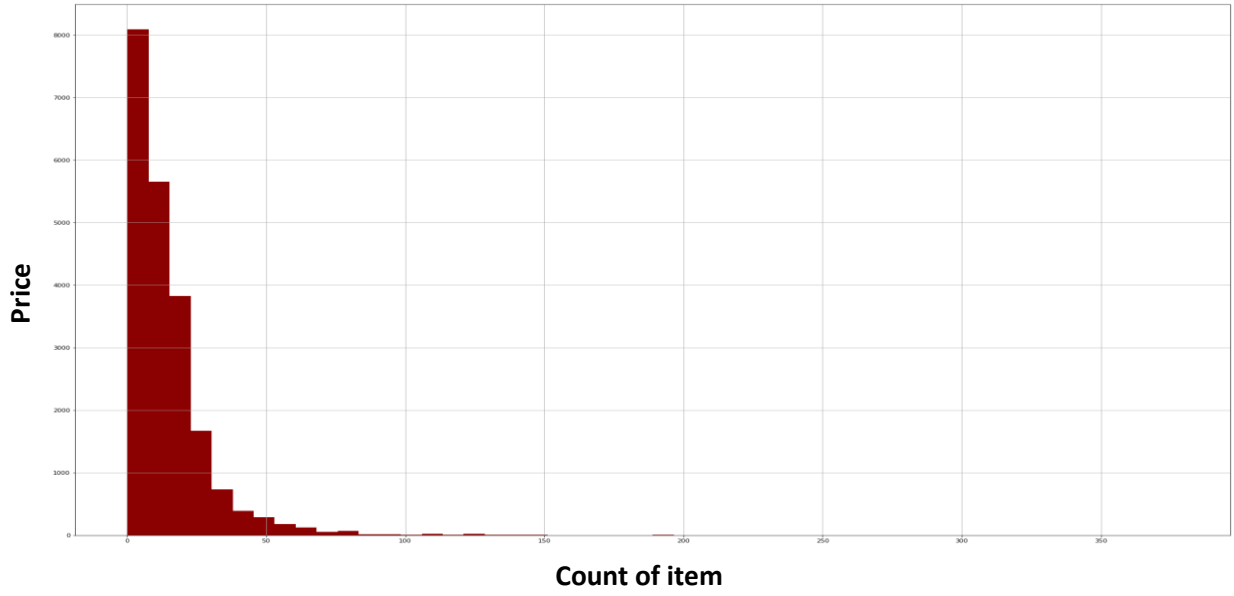
4) Visualize the Data



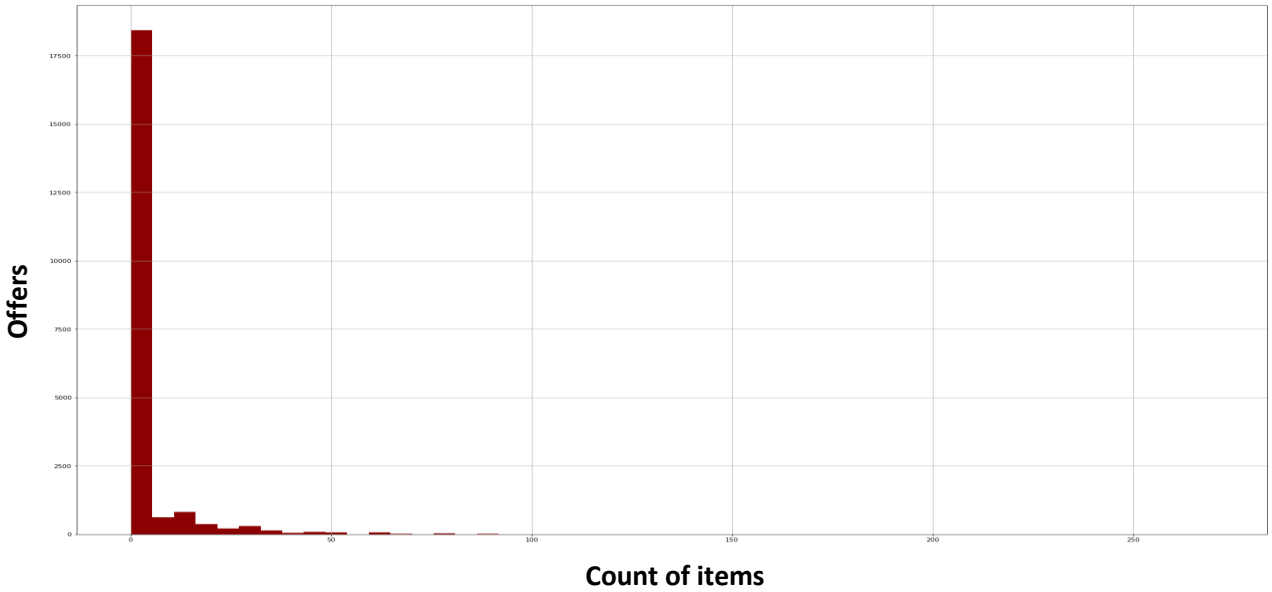
About 14000 items are food beverages, and other categories have less than 2000 items. So, the food beverages are the most frequent category with about 14000, then we have health beauty as the second frequent category with about 2000.



Goody brand is the most frequent brand with about 680 items, then we have Freshly as the second frequent brand with about 610 items, then Nestlé as the third frequent brand with about 610 items. All the most frequent brands are food brands.



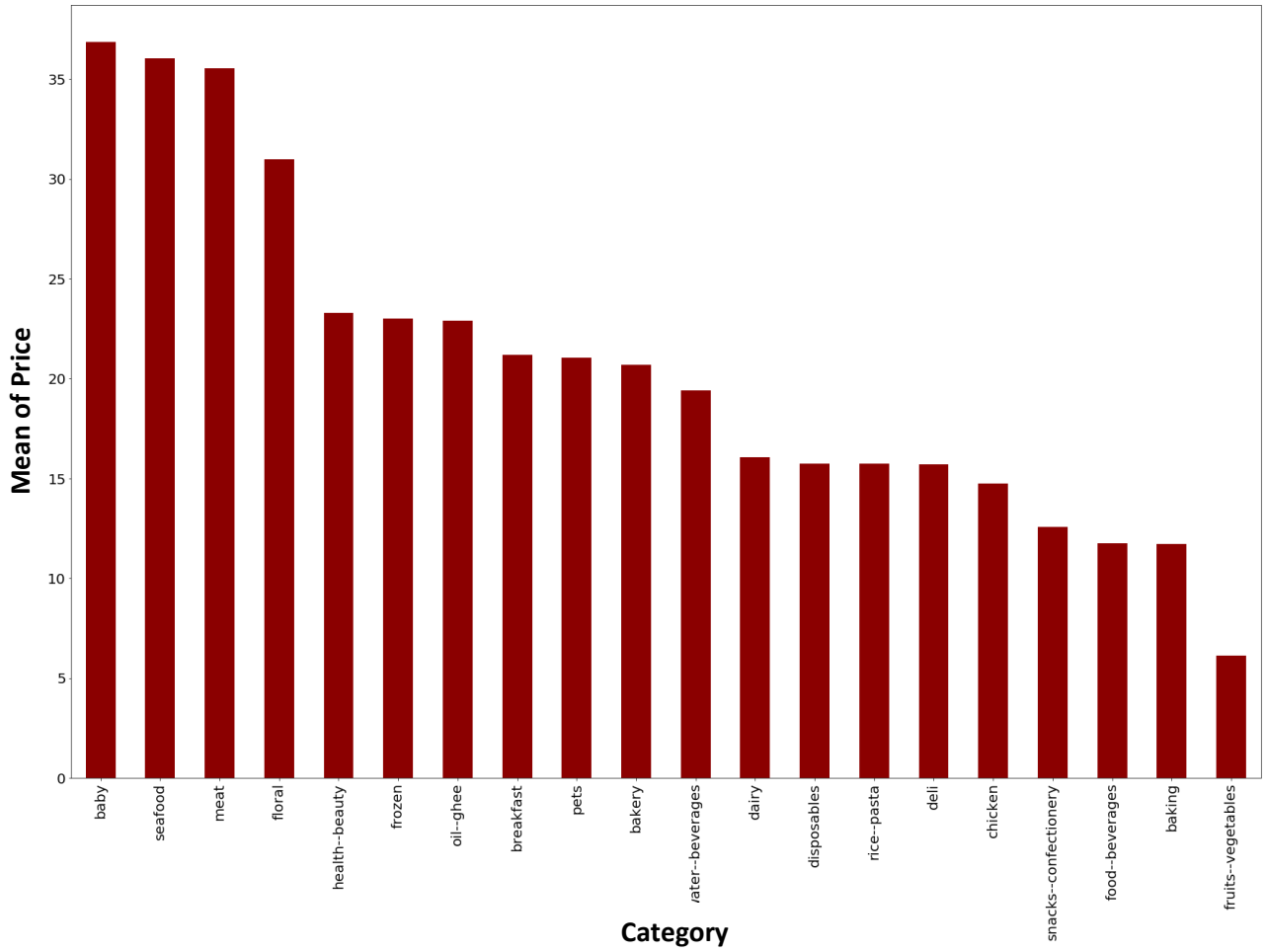
As we see in the above graph the prices between 0.25 and 377. More than 75% of items are less than 50.00 SAR, and half of items are under 25.00 SAR.



Only 25% of items have offer, more than half of the items with an offer has under 50.00 SAR offer.

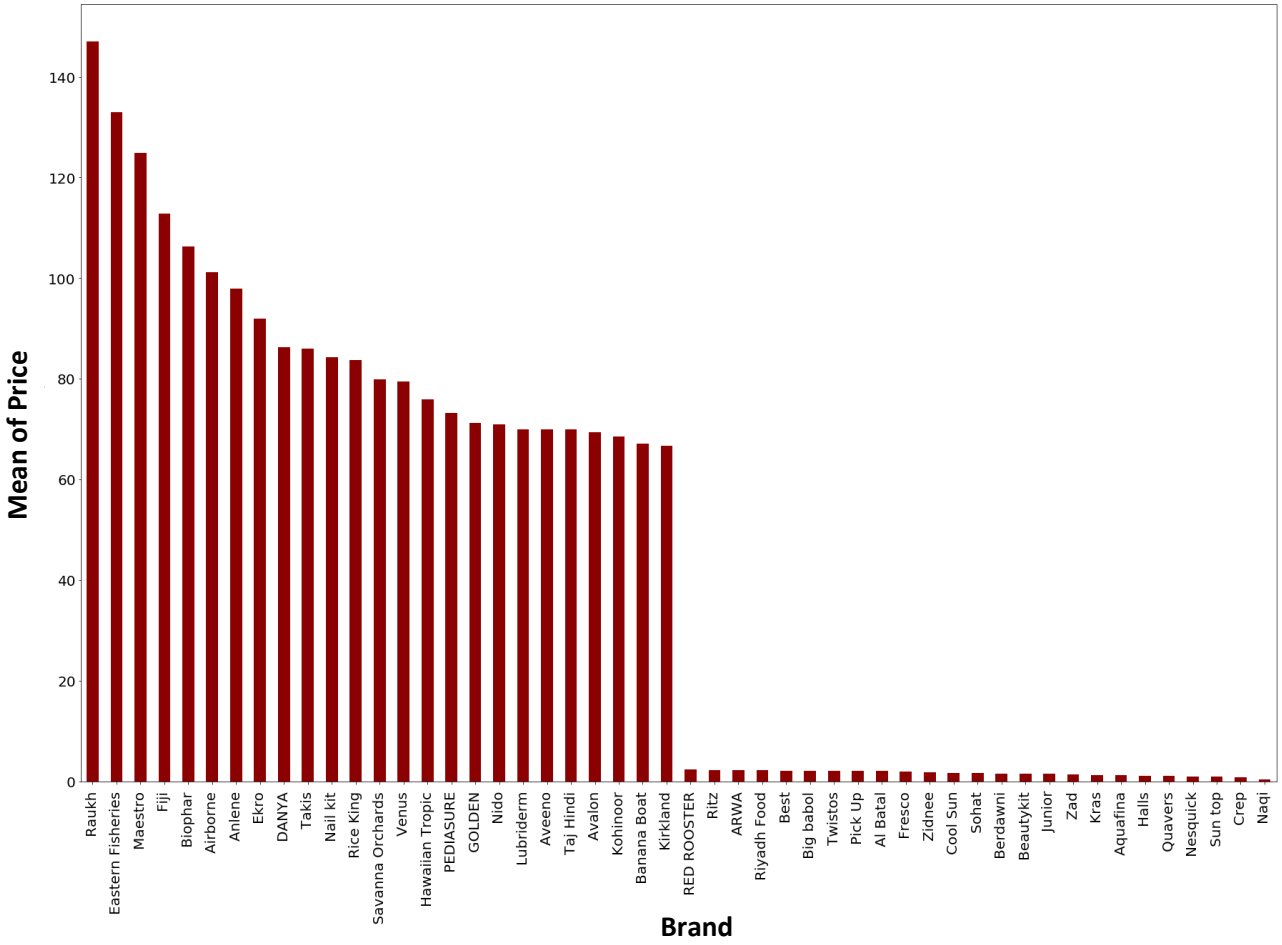
5) Statistical Analysis

- The mean price for each category



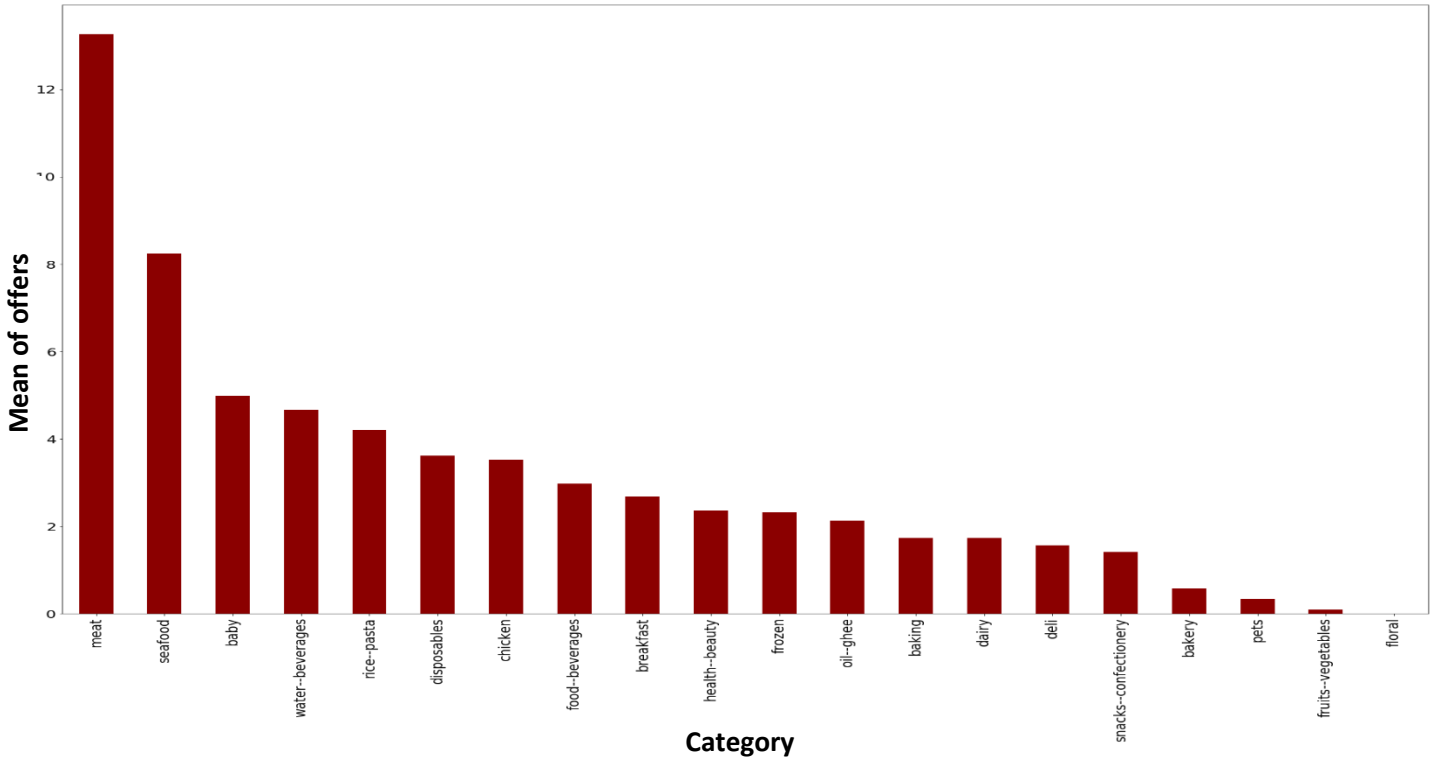
After calculating the mean price for each category. Baby category has the most expensive items with about 37 SAR, seafood category is the second most expensive items with about 36 SAR, meat category is the third most expensive items with about 35 SAR. The cheapest category is fruits and Vegetables, then baking and food beverages become after fruits and Vegetables in the cheapest category.

- The mean price for each Brand with the higher 25 and Lower 25



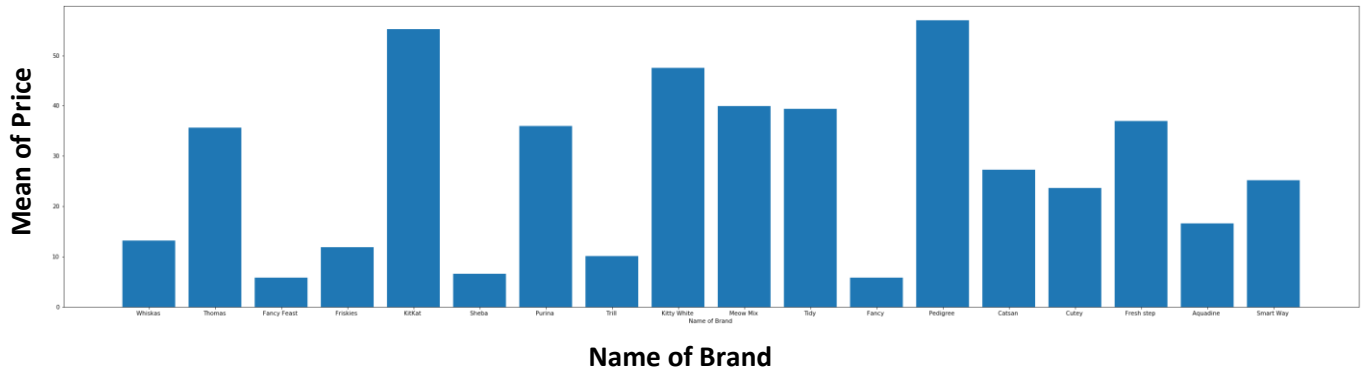
After calculating the mean price for each brand. Raukh brand has the most expensive items with about 147 SAR. Raukh is a juice brand. Eastern Fisheries brand is the second most expensive items with about 132 SAR. Eastern Fisheries is a fish brand. Maestro brand is the third most expensive items with about 124 SAR. Maestro is olive and oil brand. The cheapest brands are Naqi & Crep & Sun top with less than 1.00 SAR mean price. Sun Top is a juice brand. Crep is chocolate brand. Naqi is water brand.

- The mean of offer with each category



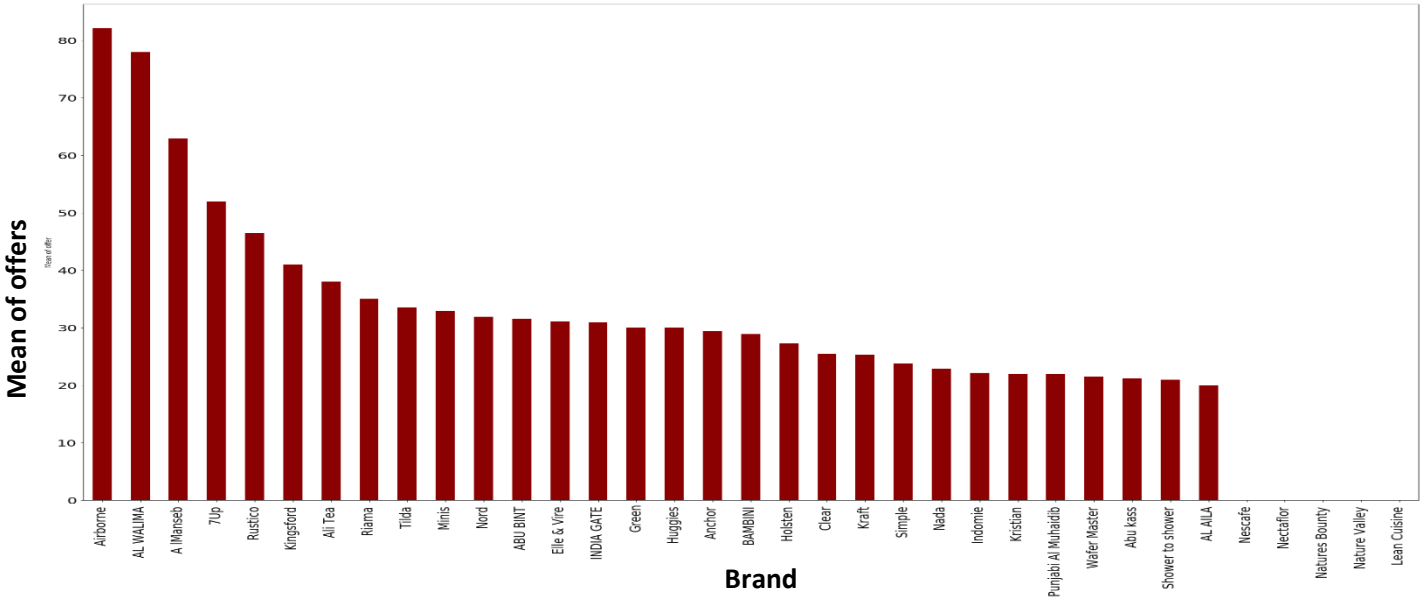
After calculating the mean offer for each category. Meat category items have the highest mean offer, floral is the only category with no offers.

- The mean of price with each brand



After calculating the mean price for each brand. Pedigree and Kitkat brands are having the highest mean of price. Fancy, Fancy Feast and sheba brands are having the lowest mean of price.

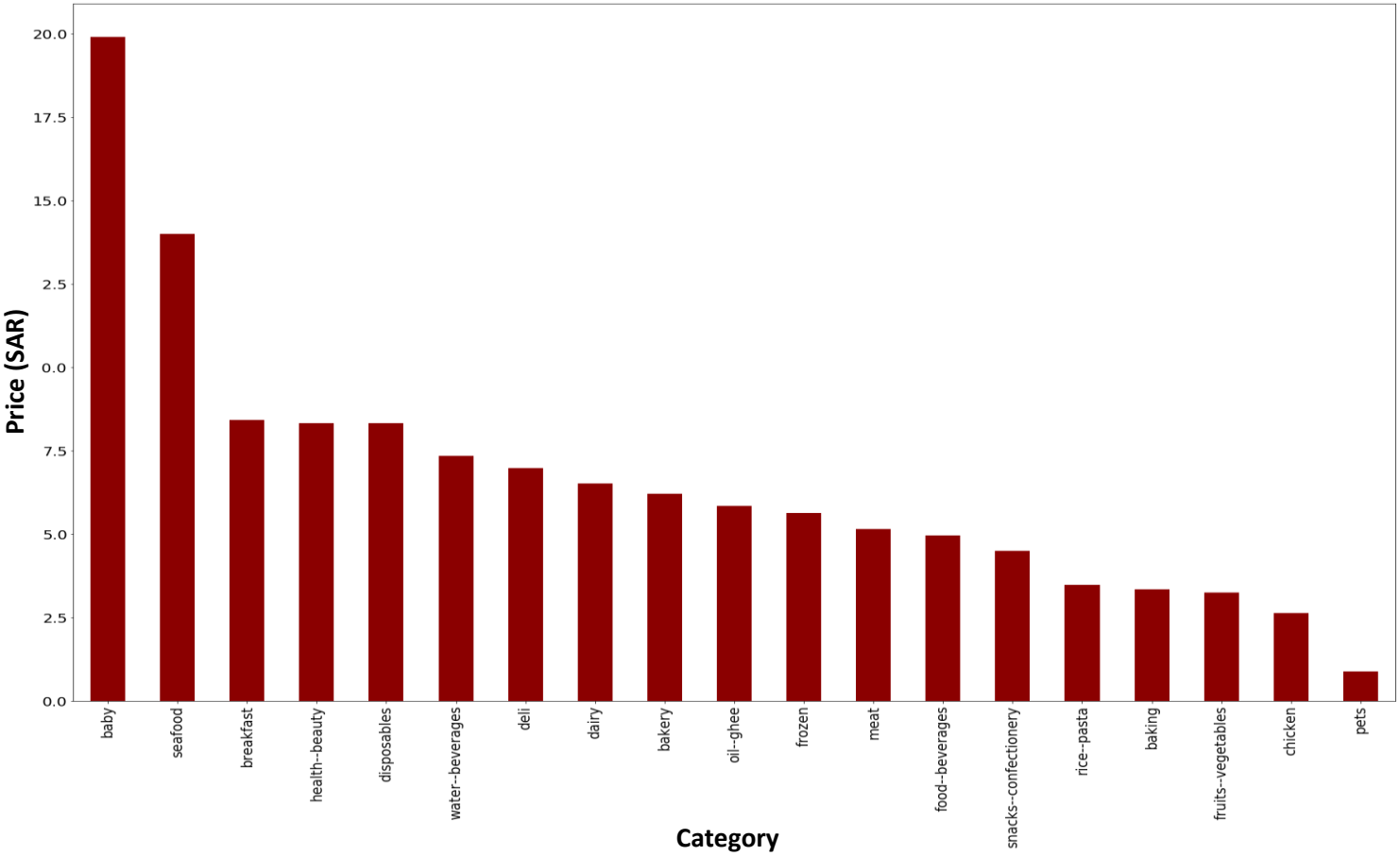
- The mean of offer with each brand



After calculating the mean offer for each brand. Airborne brand is having the highest mean offer, Nescafe, Nectaflor, Natyres Bounty, Nature Valley, Lean Cuisineare are only brand with no offers.

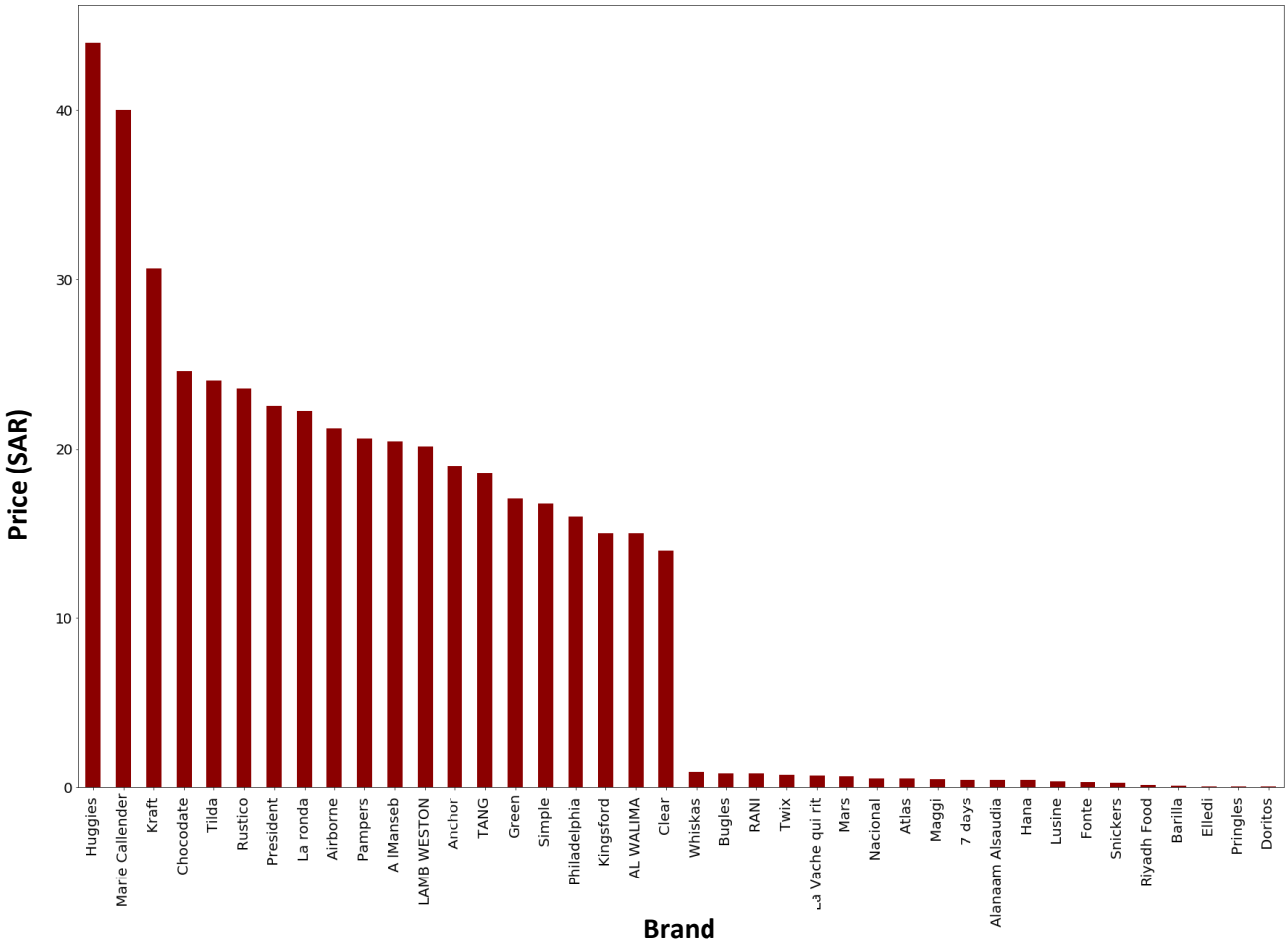
We create new dataframe for offers. this new df have the ordinal price "offer" then the price after offer "price" and how much of money will customer save.

- The discount from each category



Baby category has the best offers with mean 20.00 SAR discounts, then seafood has the second best offers with mean 14.00 SAR discounts. both baby & seafood are the most expensive categories, so we can say the most expensive categories have the best offers.

- The discount from each brand



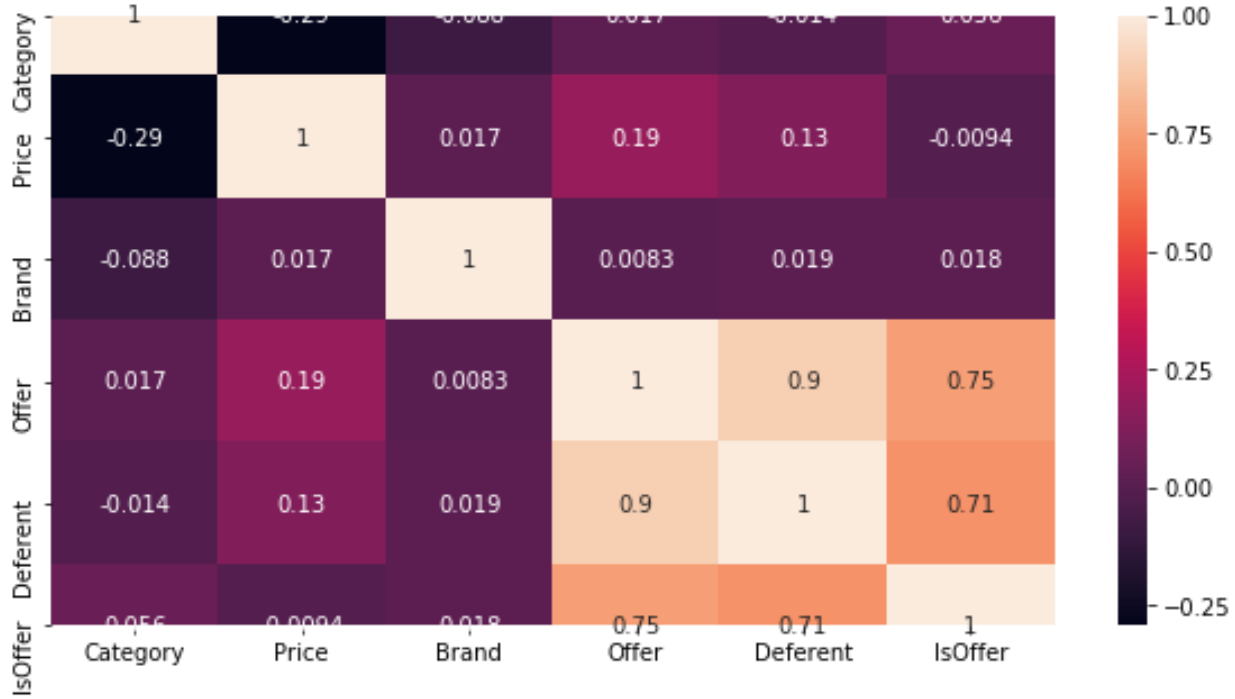
Huggies brand has the best offers with mean 44.00 SAR discounts, then Marie Callender brand has the second best offers with mean 40.00 SAR discounts. 50% of brands that give offers are given offers under 8.00 SAR discounts. Only 25% of brands that give offers are given offers above 20 SAR discounts.

6) Modeling

We build two models; the first model predicts if the items should have discount or not. This model can help web and app developers to add recommendations for each item should have offer or not. The second model and the third model are for predict the category. This model can help web and app developers to have well understand of the data.

We add new column "isOffer", if the item has offer or not. Then we transform the data to find the correlation of columns with the "isOffer".

- The Correlation Between Variables:



- First Model:

We trying to predict is the item has offer. Category, Price and Brand were used as features, IsOffer was used as label. Then we split 80/20 training and test set, random state controls the shuffling applied to the data before applying the split. The machine learning type is supervised classification. The models are SVM, Logistic Regression and Naïve Bayes.

- Second Model:

We trying to predict the category, Offer, Price and Brand were used as features, Category was used as label. Then we split 80/20 training and test set, random state controls the shuffling

applied to the data before applying the split. The machine learning type is supervised classification. The models are SVM, Logistic Regression and Naïve Bayes.

- Third Model:

We trying to predict the category by the name of the item using Tf-idf:

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

So, Category was used as label. Name was used as feature and we converted the letters to lowercase letters and removed the numbers and punctuation marks if any, then we put it in TfidfVectorizer. We split 80/20 training and test set, random state controls the shuffling applied to the data before applying the split. The machine learning type is supervised classification. The models are SVM, Logistic Regression and Naïve Bayes.

7) Evaluation

	Model 1	Model 2	Model 3
Logistic Regression	The Accuracy is: 86% The error rate is: 14%	The Accuracy is: 64% The error rate is: 36%	The Accuracy is: 83% The error rate is: 17%
Support Vector Machine	The Accuracy is: 32% The error rate is: 68%	The Accuracy is: 66% The error rate is: 34%	The Accuracy is: 87% The error rate is: 13%
Naïve Bayes	The Accuracy is: 86% The error rate is: 14%	The Accuracy is: 61% The error rate is: 39%	The Accuracy is: 79% The error rate is: 21%
The Result	Logistic regression and naïve bayes has same accuracy. SVM has very bad accuracy.	All models have similar accuracy. SVM considered the best model, NB considered the worst model.	Logistic regression and SVM have similar accuracy. SVM considered the best model, NB considered the worst model.