Retail Analytics Trends

Big Data Processing and Application

Shakib Ibna Shameem (2307673)
Akash Sinha Bappy (2307938)
K H M Burhan Uddin (2307264)
Mitun Paul (2307325)

Introduction to Big Data in Retail

Big Data in Retail

- Huge records generated everyday
- Massive volume of structured and unstructured data from various sources.

Importance

- Essential for analyzing trends, customer behavior
- Enhancing operational efficiency.

Big Data Tools

- Parallel processing of multiple sources of data
- Process as stream

Objectives of the Project

Apply Big Data Tools

- Using pySpark for parallel processing
- MLib for predictive analysis

Descriptive Analysis

- Exploring patterns and trends in historic data
- Setting future objectives
- Performance measurements

Predictive Analysis

- Inventory stock prediction
- Sales forecasting

Dataset

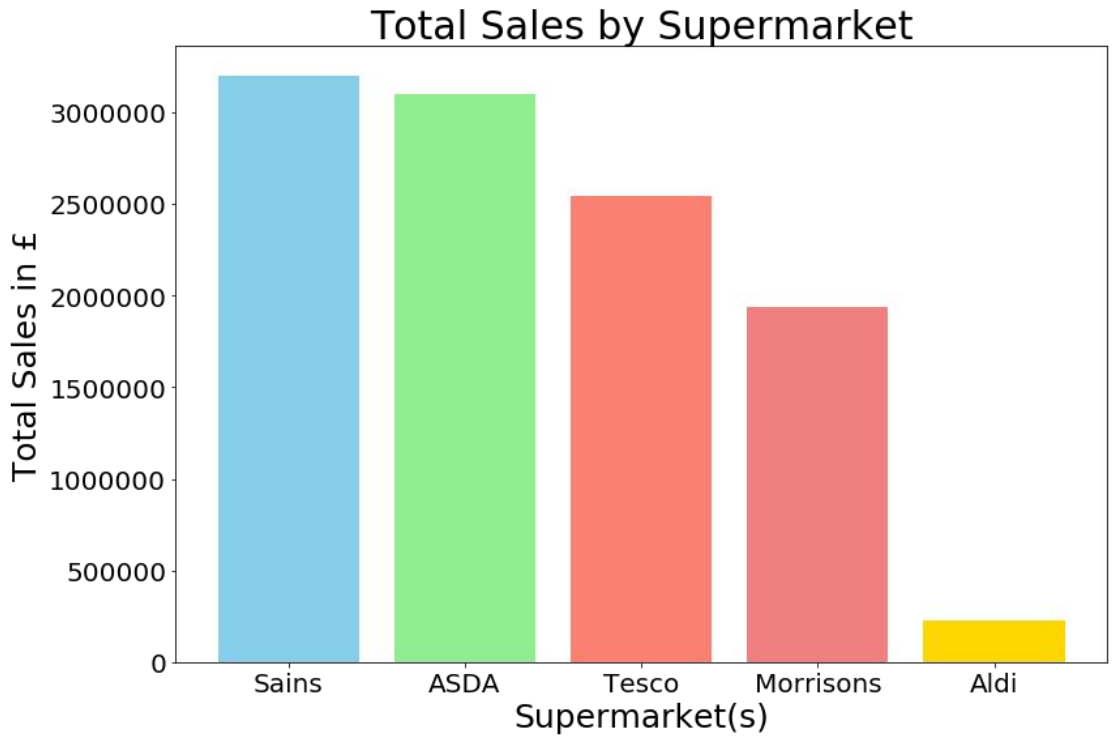
Provides a comprehensive view of supermarket prices, enabling analysis of pricing trends, unit costs, and brand presence across various categories.

```
root
 |-- supermarket: string (nullable = true)
 |-- date: date (nullable = true)
 |-- month: integer (nullable = true)
 |-- year: integer (nullable = true)
 |-- product_category: string (nullable = true)
 |-- item_count: integer (nullable = true)
```

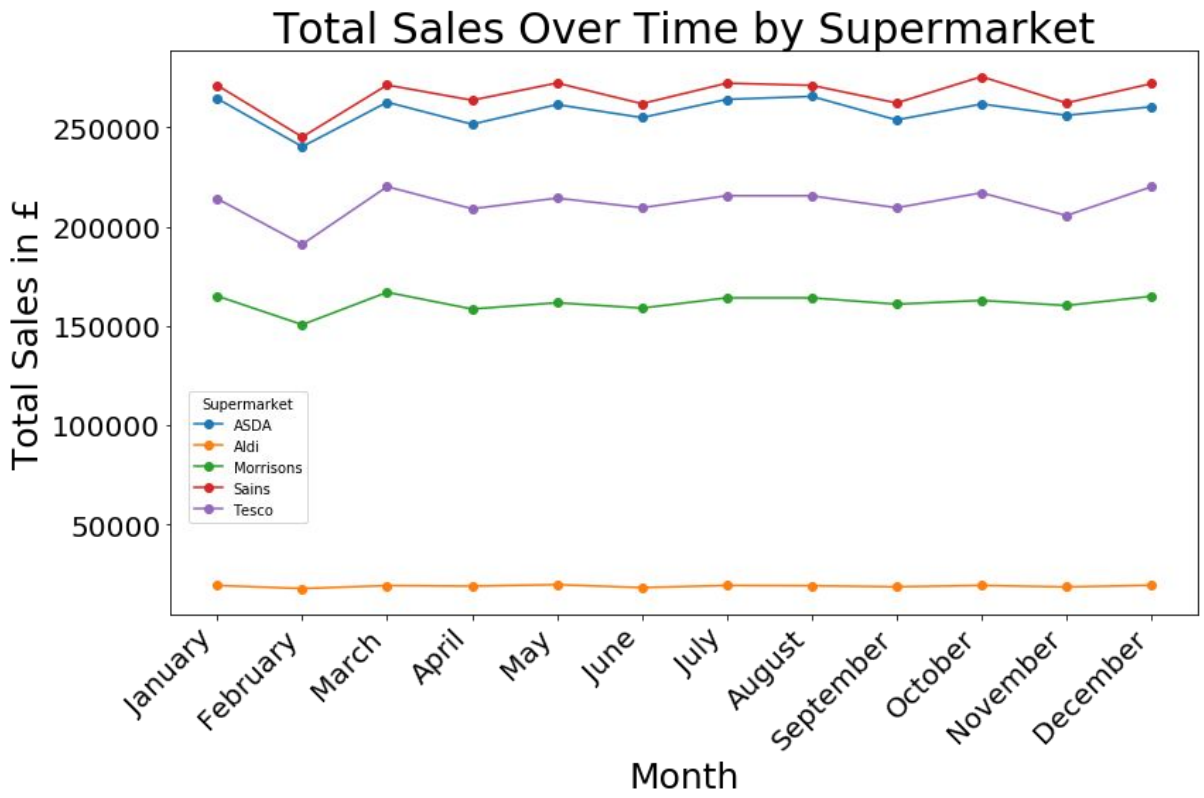
Source: https://www.kaggle.com/datasets/willianoliveiragibin/retail-analytics-trends

Exploratory Data Analysis

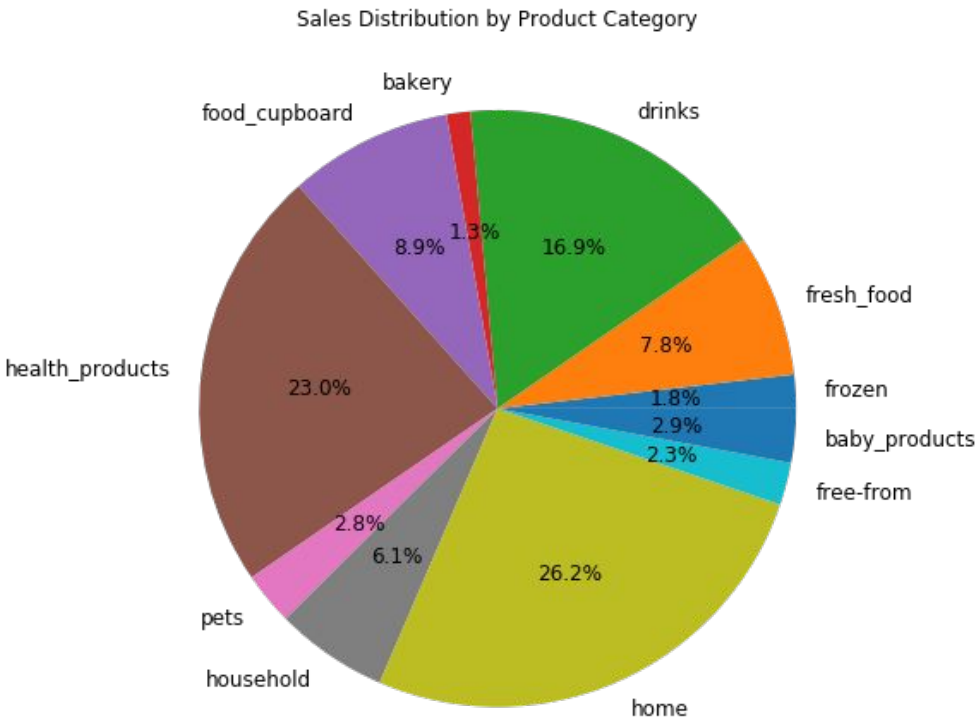
Total Sales by Supermarket



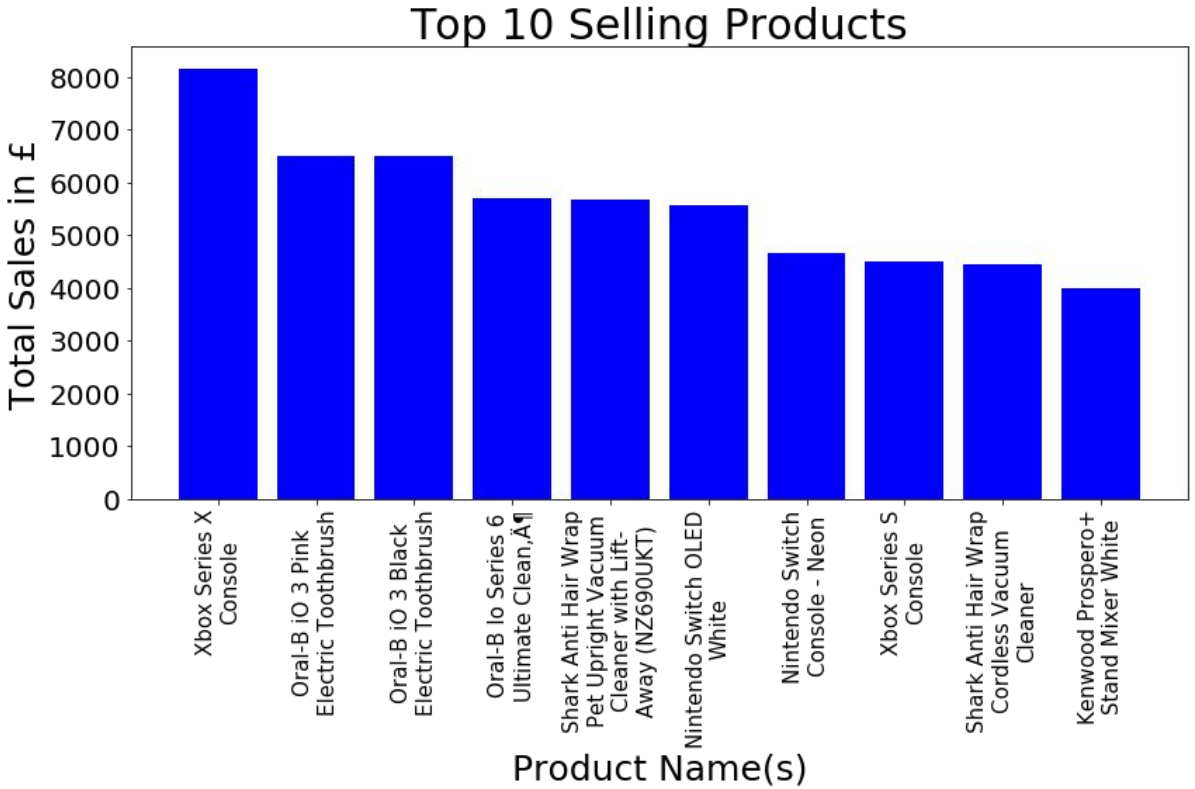
Sales by Month



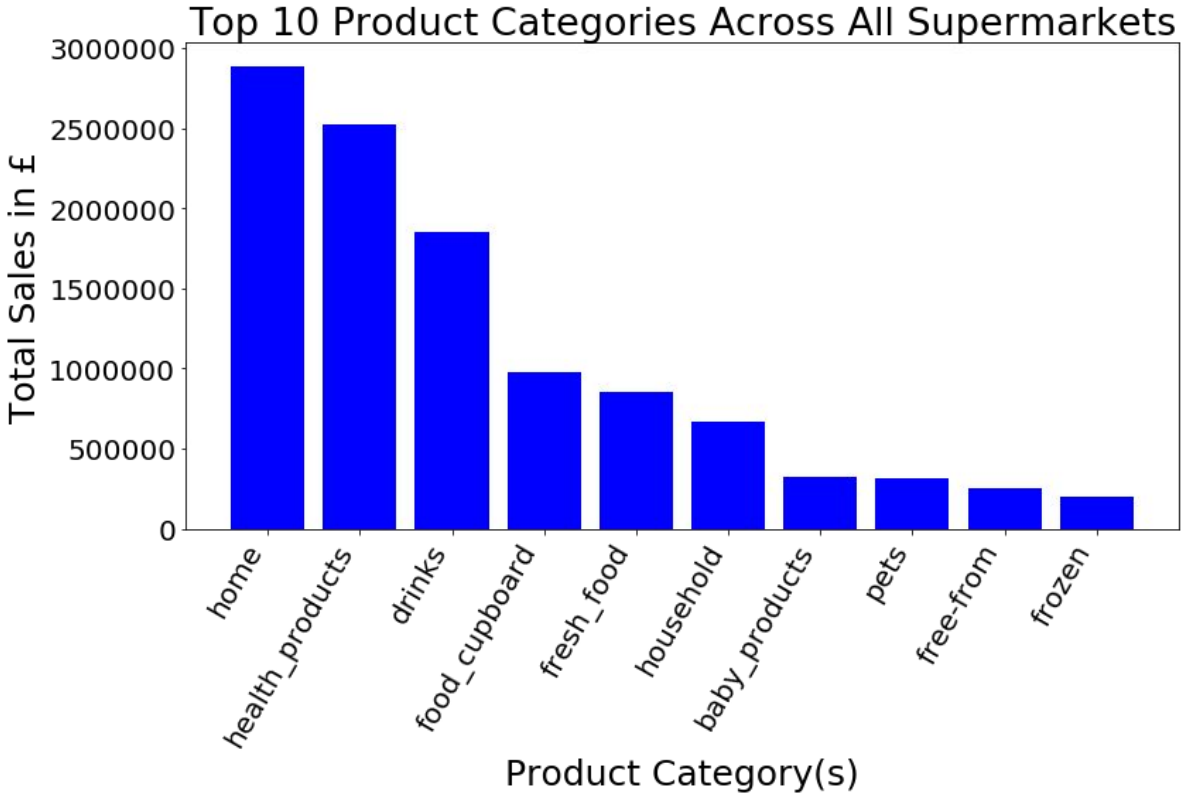
Sales Distribution by Product Category



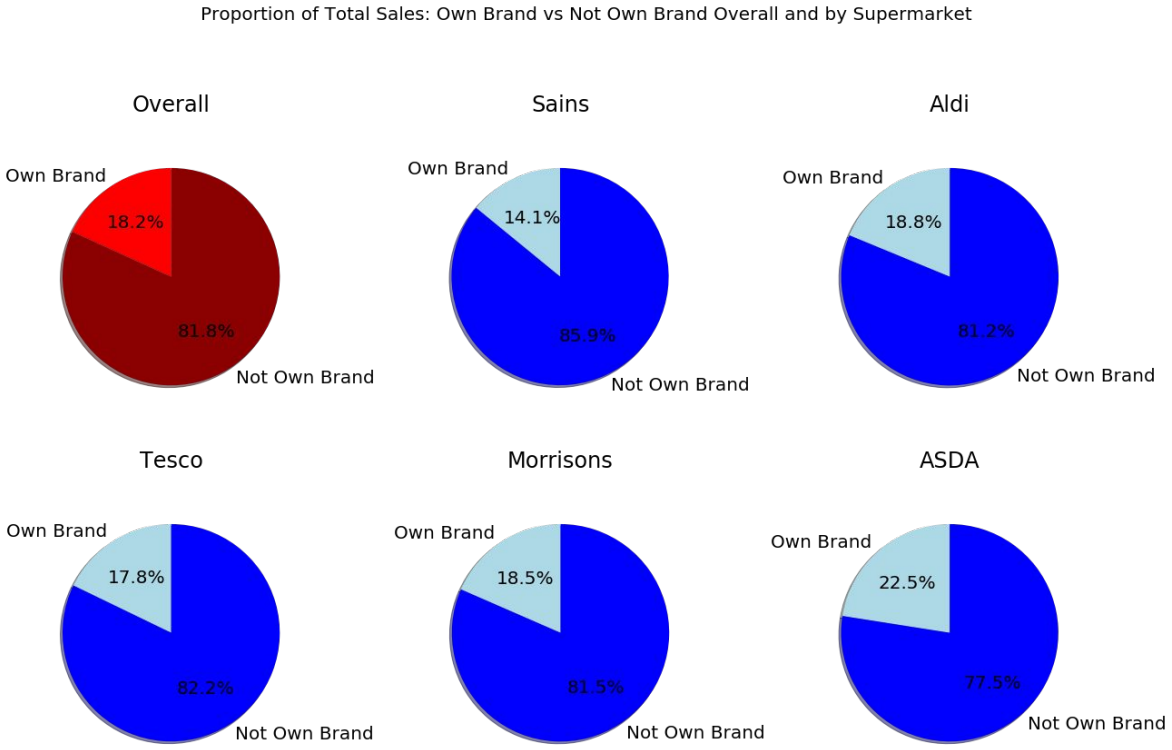
Top Sales


Top 10 Selling Products
Total Sales in £
8000
7000
6000
5000
4000
3000
2000
1000
0
Xbox Series X Console
Oral-B iO 3 Pink Electric Toothbrush
Oral-B iO 3 Black Electric Toothbrush
Oral-B Io Series 6 Ultimate Clean,Ã¶
Shark Anti Hair Wrap Pet Upright Vacuum Cleaner with Lift-Away (NZ690UKT)
Nintendo Switch OLED White
Nintendo Switch Console - Neon
Xbox Series S Console
Shark Anti Hair Wrap Cordless Vacuum Cleaner
Kenwood Prospero+ Stand Mixer White
Product Name(s)

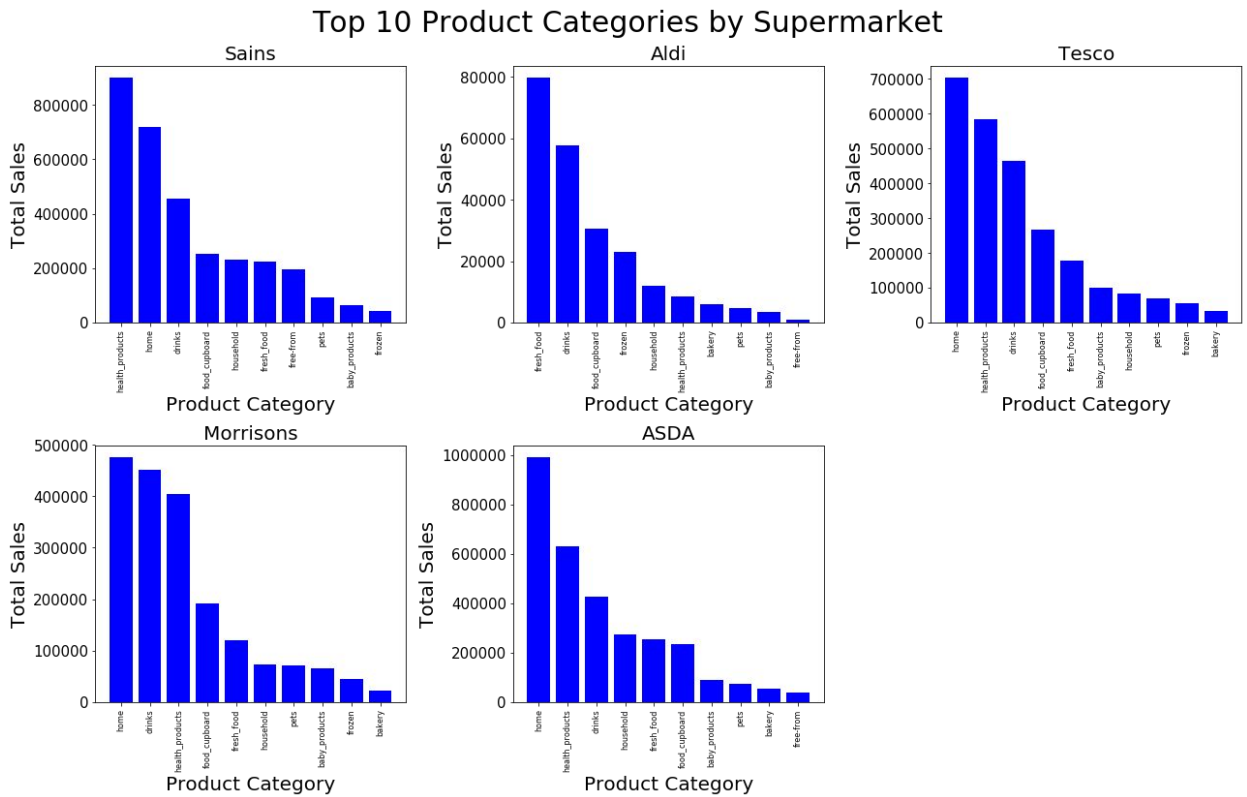
Top Sales by Category



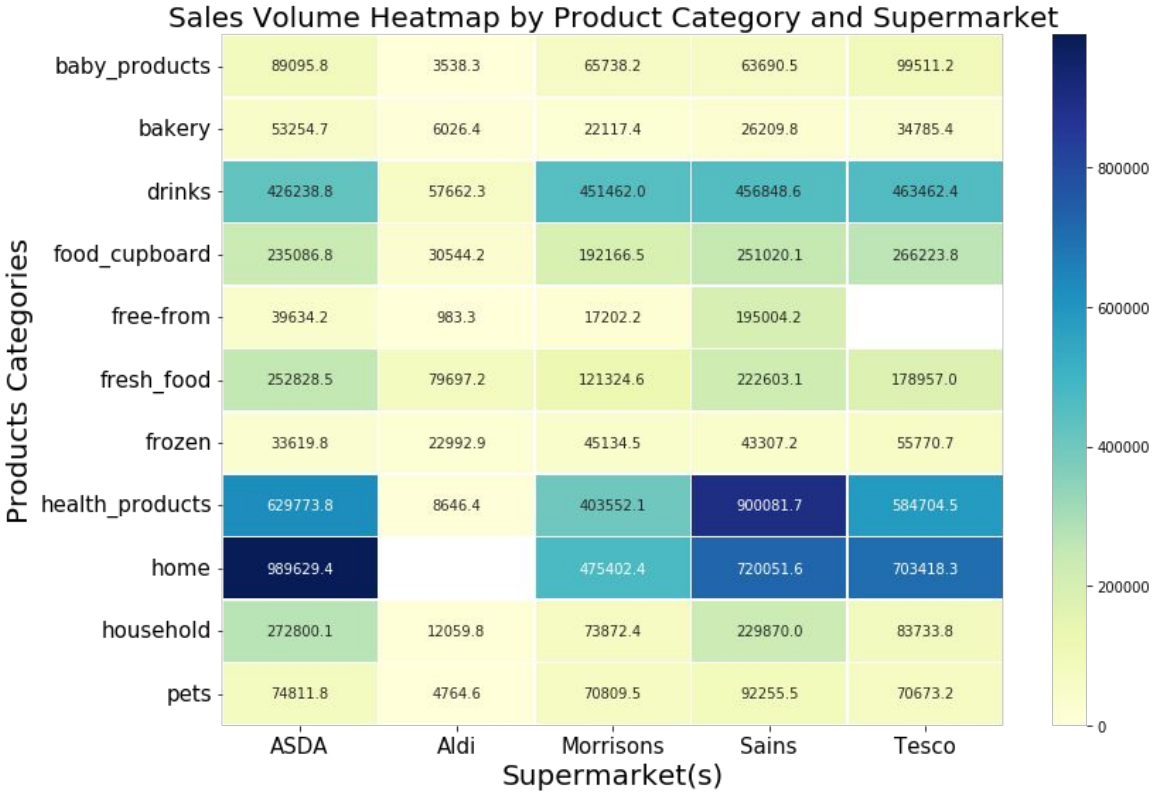
Own vs other brands sell across super markets



Product Categories by Super Market



Product Categories by Super Market



Predictive Data Analysis

Data Processing

Aggregation

```
root
 |-- supermarket: string (nullable = true)
 |-- date: date (nullable = true)
 |-- month: integer (nullable = true)
 |-- year: integer (nullable = true)
 |-- product_category: string (nullable = true)
 |-- item_count: integer (nullable = true)
```

summary	supermarket	month	year	product_category	item_count
count	19283	19283	19283	19283	19283
mean	NULL	6.5246590260851525	2023.0	NULL	109.12912928486232
stddev	NULL	3.448263466319703	0.0	NULL	94.50654972258147
min	ASDA	1	2023	baby_products	1
max	Tesco	12	2023	pets	396

Dataset Processing

Label Encoding & Features Vectors

month	item_count	supermarket_cat	product_category_cat	day
1	332	2.0	5.0	18
1	323	2.0	2.0	26
1	39	2.0	8.0	28
1	47	3.0	4.0	12
1	117	1.0	1.0	2

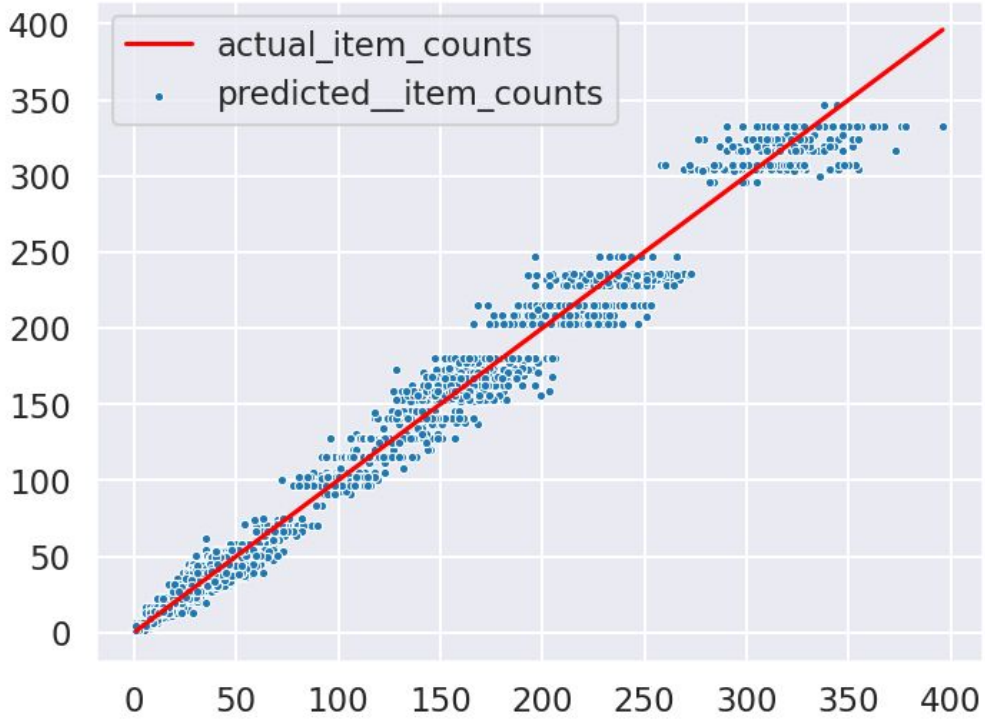
only showing top 5 rows

month	item_count	supermarket_cat	product_category_cat	day	feature_vectors
1	332	2.0	5.0	18	[1.0,18.0,2.0,5.0]
1	323	2.0	2.0	26	[1.0,26.0,2.0,2.0]
1	39	2.0	8.0	28	[1.0,28.0,2.0,8.0]
1	47	3.0	4.0	12	[1.0,12.0,3.0,4.0]
1	117	1.0	1.0	2	[1.0,2.0,1.0,1.0]
1	154	1.0	1.0	7	[1.0,7.0,1.0,1.0]
1	129	1.0	3.0	25	[1.0,25.0,1.0,3.0]
1	12	4.0	0.0	16	[1.0,16.0,4.0,0.0]
1	5	4.0	10.0	6	[1.0,6.0,4.0,10.0]
1	186	2.0	10.0	16	[1.0,16.0,2.0,10.0]

only showing top 10 rows

Result

```
R Squared Error on test data = 0.9861
Root Mean Squared Error on test data = 11.2828
Mean Squared Error on test data = 127.301
Mean Absolute Error on test data = 8.12435
```



Thank You