Assessment One: Data preparation and Visualisation

DA3307 - Data Visualisation

Muhammad Hidayat Bin Mohd Yusof (20FTT2910)

Dataset Description

Variable Name	Description	Type
ProductID	The unique ID number of the product.	Categorical
Weight	The weight of the product.	Numerical
FatContent	The amount of fat content contained within the product.	Numerical
ProductVisibility	The percentage of display area allocated for this product out of all products on display.	Numerical
ProductType	The category of product.	Categorical
MRP	The maximum retail price of the product.	Numerical
OutletID	The unique ID number of the store the product is sold in.	Categorical
EstablishmentYear	The year of establishment of the outlet.	Categorical
OutletSize	The size of the store.	Categorical
LocationType	The type of city the store is located in.	Categorical
OutletType	What type of outlet it is either a grocery store or supermarket.	Categorical
OutletSales	The sales of the product in that particular store.	Numerical

Data Preparation

Correcting inconsistent categorical values

```
#Correcting inconsistent naming of FatContent nominal attribute
df.replace(to_replace={'reg':'Regular','LF':'Low Fat','low fat':'Low Fat'},inplace=True)
df['FatContent'].value_counts()
```

```
Low Fat    5517
Regular    3006
Name: FatContent, dtype: int64
```

Data Preparation

Missing Values

```
[10]
df.isnull().sum()
```

```
ProductID              0
Weight              1463
FatContent             0
ProductVisibility      0
ProductType            0
MRP                    0
OutletID               0
EstablishmentYear      0
OutletSize          2410
LocationType           0
OutletType             0
OutletSales            0
dtype: int64
```

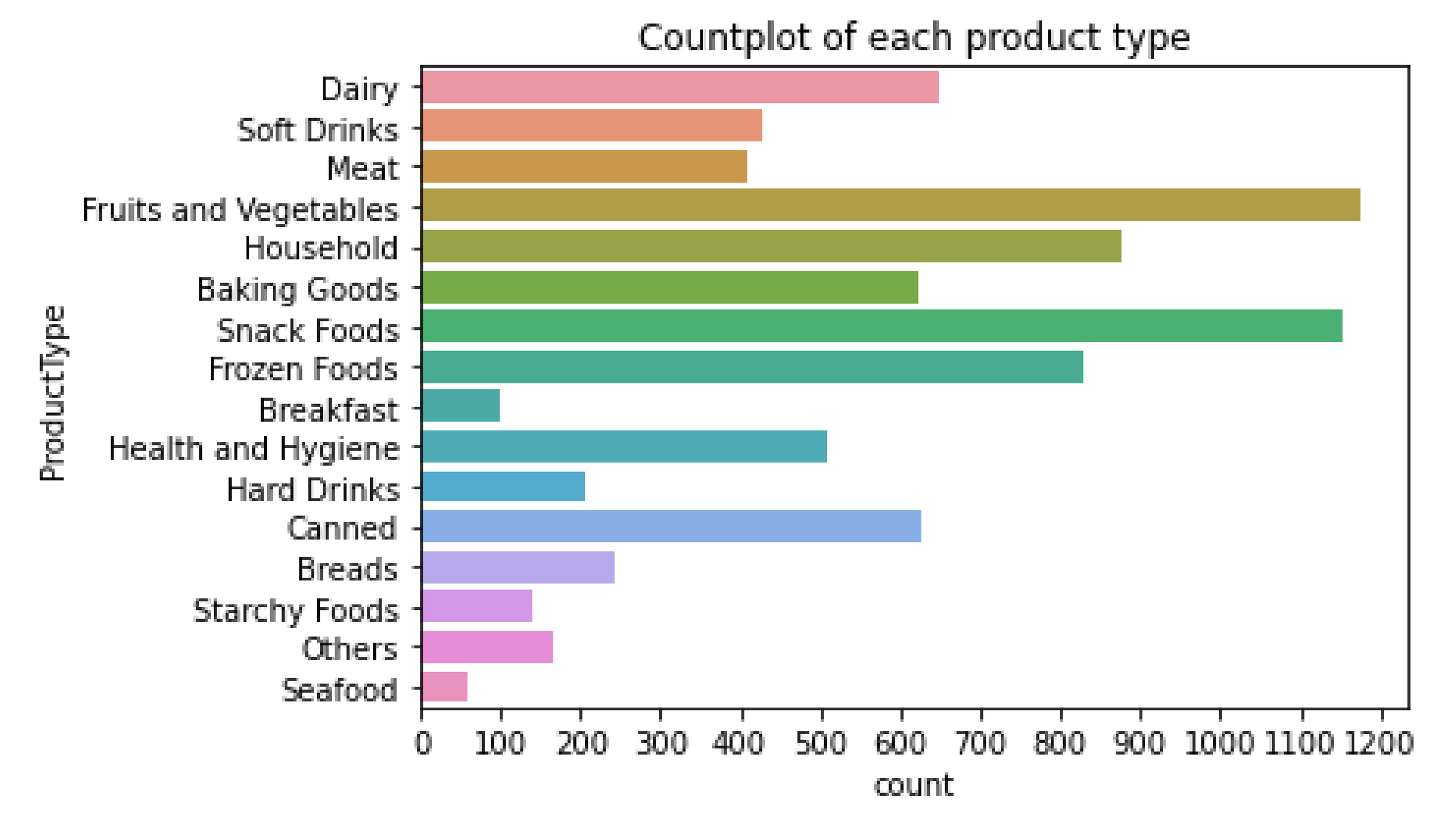
```
#Missing values in 'Weight' attribute is repalced with its mean and missing values in
# 'OutletSize' is replaced with 'Medium' which is the columns mode
df.fillna({'Weight':df['Weight'].mean(),'OutletSize':'Medium'},inplace=True)
df.isnull().sum()
```

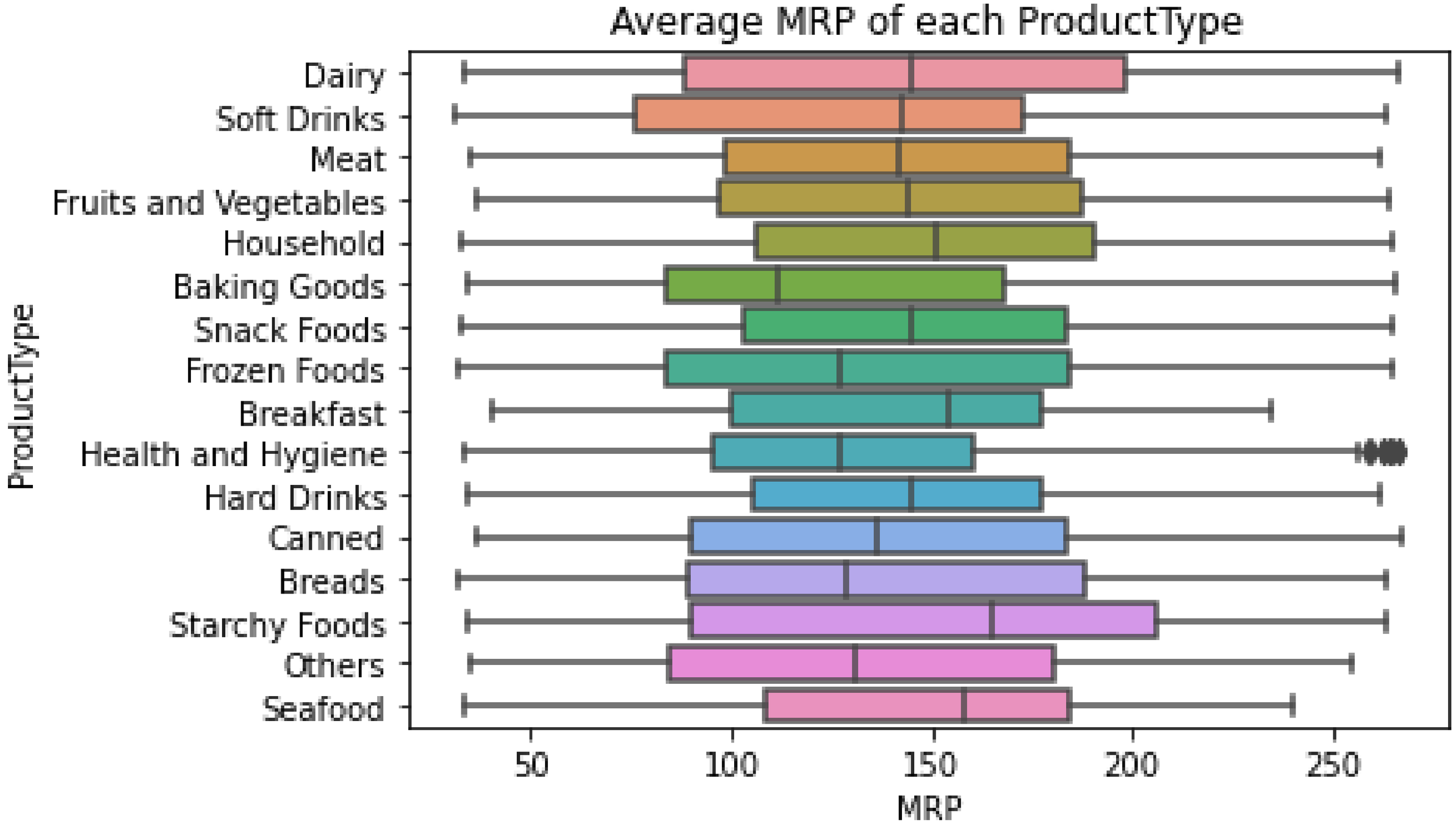
```
ProductID           0
Weight              0
FatContent          0
ProductVisibility   0
ProductType         0
MRP                 0
OutletID            0
EstablishmentYear   0
OutletSize          0
LocationType        0
OutletType          0
OutletSales         0
dtype: int64
```

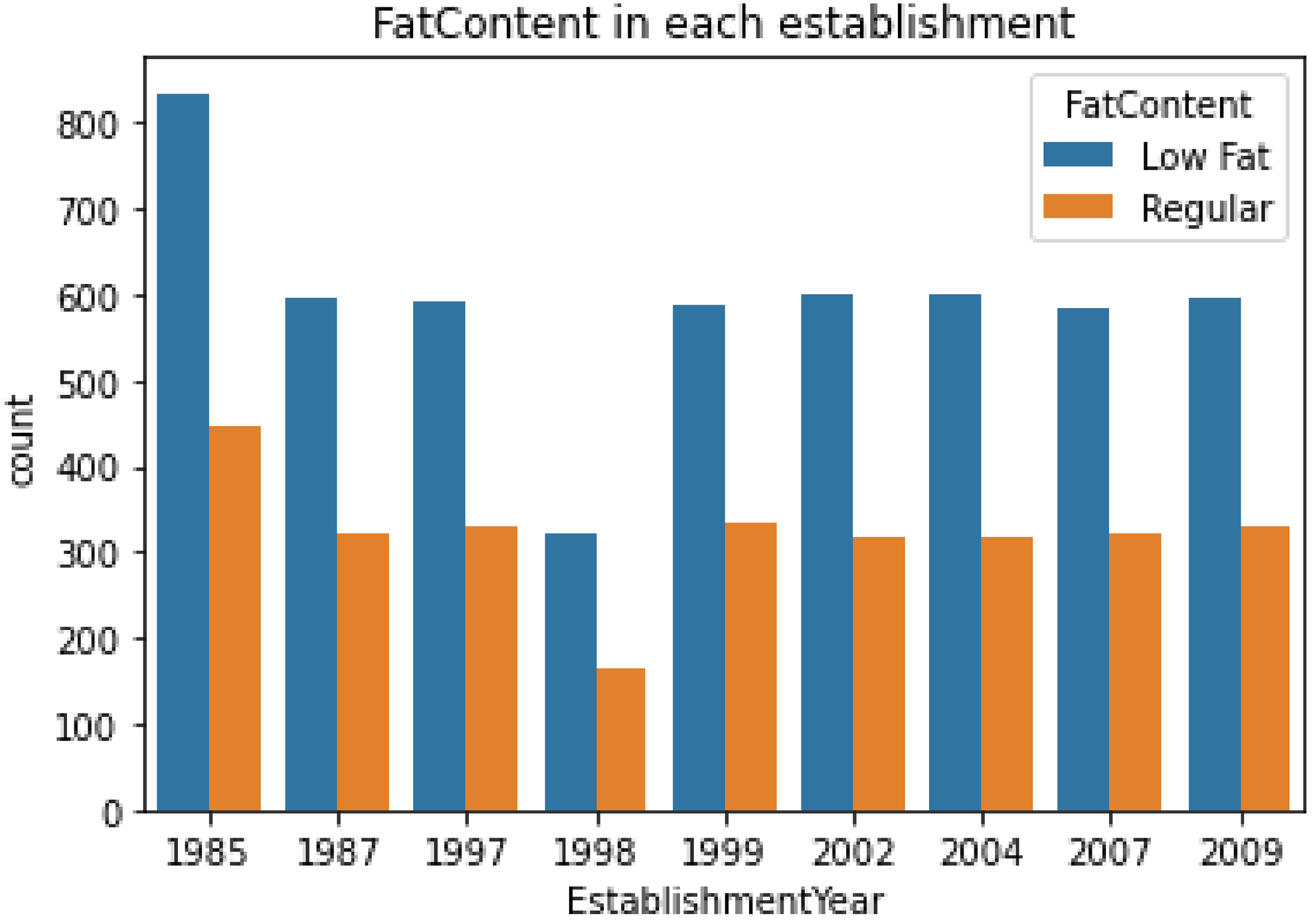
Data Preparation

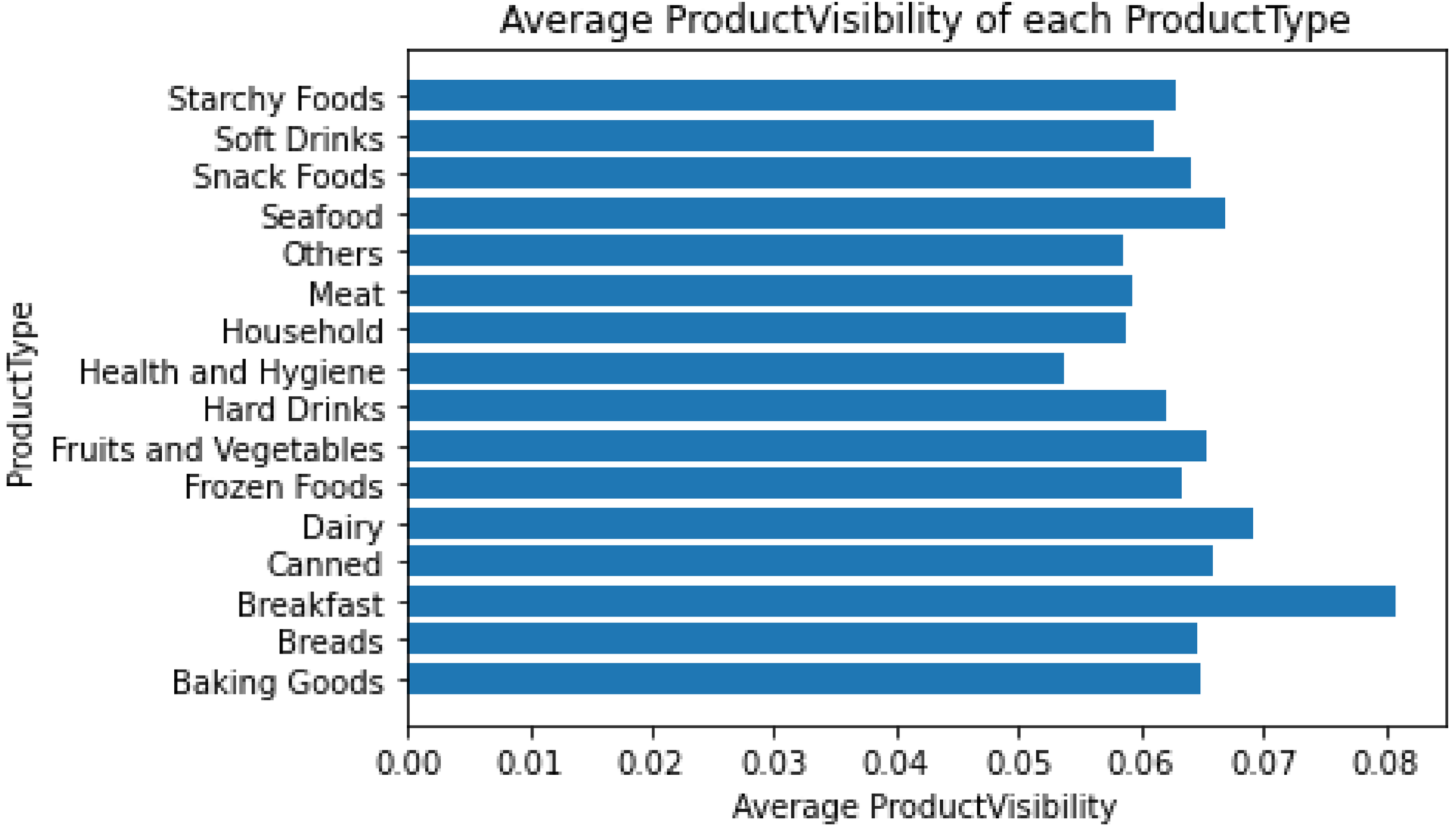
Outliers

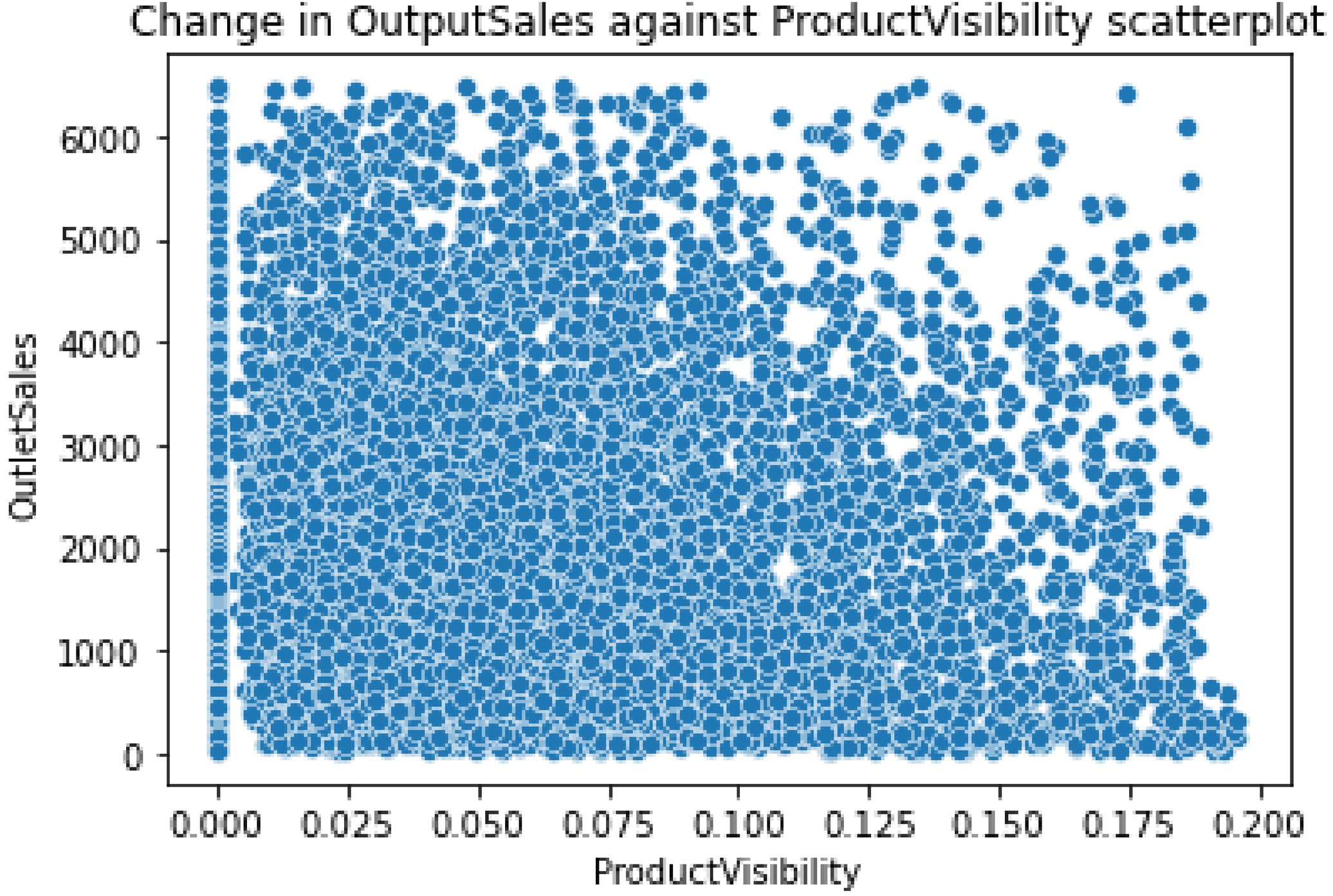
```
q1=numeric.quantile(q=0.25)
q3=numeric.quantile(q=0.75)
IQR=q3-q1
lowerbound=q1-IQR*1.5
upperbound=q3+IQR*1.5

not_outlier=numeric[~((numeric<lowerbound)|(numeric>upperbound)).any(axis=1)]

no_outler_df=df.loc[not_outlier.index]
no_outler_df
```


Countplot of each product type
Dairy
Soft Drinks
Meat
Fruits and Vegetables
Household
Baking Goods
Snack Foods
Frozen Foods
Breakfast
Health and Hygiene
Hard Drinks
Canned
Breads
Starchy Foods
Others
Seafood
ProductType
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
count


Average MRP of each ProductType
Dairy
Soft Drinks
Meat
Fruits and Vegetables
Household
Baking Goods
Snack Foods
Frozen Foods
Breakfast
Health and Hygiene
Hard Drinks
Canned
Breads
Starchy Foods
Others
Seafood
ProductType
50
100
150
200
250
MRP


FatContent in each establishment
FatContent
Low Fat
Regular
800
700
600
500
400
300
200
100
0
ount
1985
1987
1997
1998
1999
2002
2004
2007
2009
EstablishmentYear


Average ProductVisibility of each ProductType
Starchy Foods
Soft Drinks
Snack Foods
Seafood
Others
Meat
Household
Health and Hygiene
Hard Drinks
Fruits and Vegetables
Frozen Foods
Dairy
Canned
Breakfast
Breads
Baking Goods
ProductType
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Average ProductVisibility


Change in OutputSales against ProductVisibility scatterplot
OutletSales
6000
5000
4000
3000
2000
1000
0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
ProductVisibility

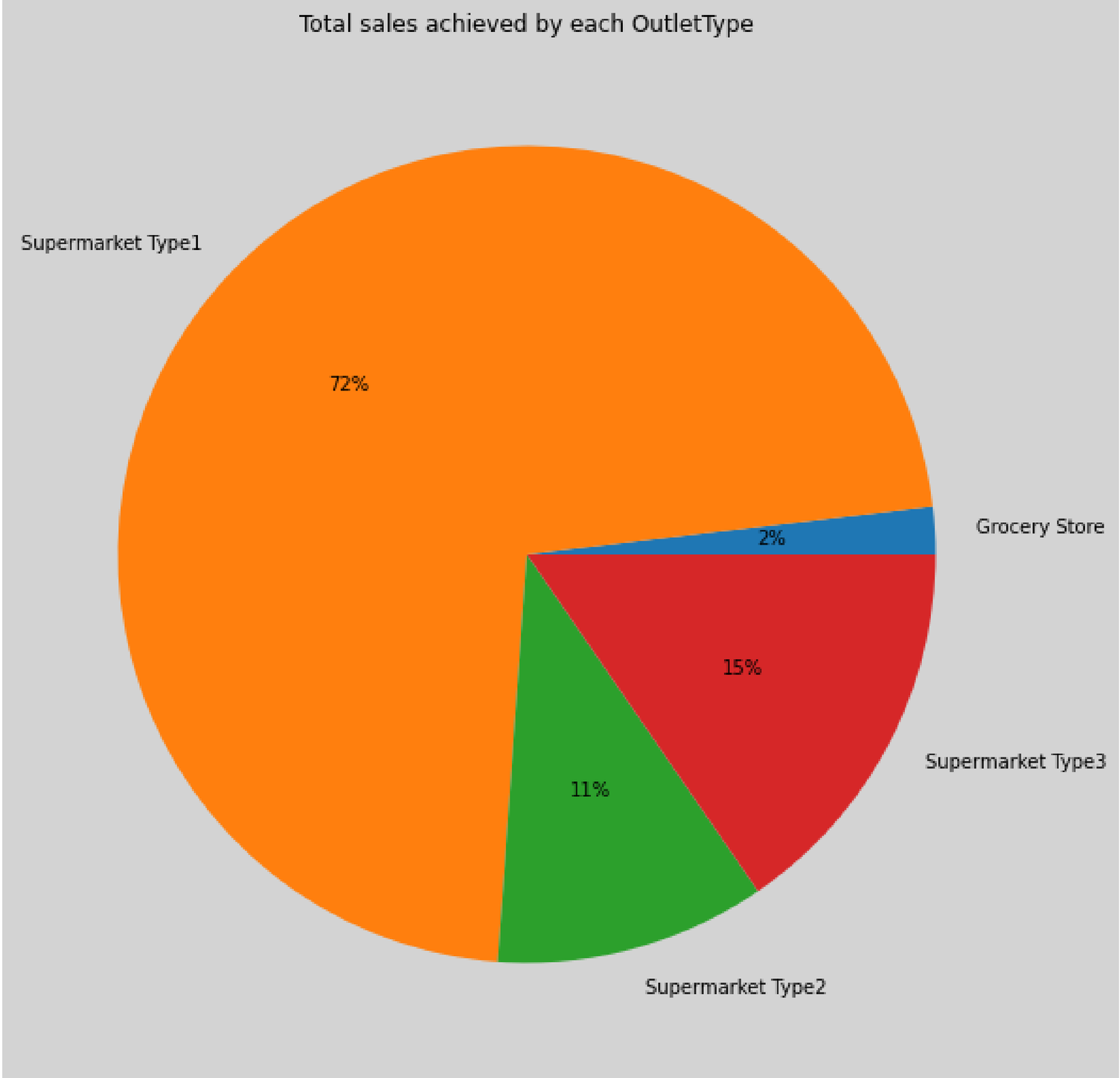

Total sales achieved by each OutletType
Supermarket Type1
72%
Grocery Store
2%
Supermarket Type3
15%
Supermarket Type2
11%

Chart Benefits and Limitations

- The charts allow thousands of rows to be condensed into useful visualizations.
- Many of which reveal interesting details that would not have been noticed from the original dataset.
- Regardless of how they are presented if the data is flawed so will the visualizations.
- the stories made for each chart still have a chance of being completely wrong because they are more or less just assumptions

Dataset limitations and recommendations

- A column listing each item bought should next time be made for market basket analysis to be conducted.
- The characteristics of the customers should also be recorded to understand their purchasing habits.

Thank you for your time!