Food Sales Prediction Project

Christina Ha

Table of Contents

Overview

Item_Identif	Item_Weight	Item_Fat_Co	Item_Visibili	Item_Type	Item_MRP	Outlet_Ident	Outlet_Estab	Outlet_Size	Outlet_Locat	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.0160473	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket	3735.138
DRC01	5.92	Regular	0.01927822	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket	443.4228
FDN15	17.5	Low Fat	0.01676008	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket	2097.27
FDX07	19.2	Regular	0	Fruits and Ve	182.095	OUT010	1998		Tier 3	Grocery Stor	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket	994.7052
FDP36	10.395	Regular	0	Baking Good	51.4008	OUT018	2009	Medium	Tier 3	Supermarket	556.6088
FDO10	13.65	Regular	0.01274109	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket	343.5528
FDP10		Low Fat	0.12746986	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket	4022.7636
FDH17	16.2	Regular	0.01668711	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket	1076.5986
FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket	4710.535
FDY07	11.8	Low Fat	0	Fruits and Ve	45.5402	OUT049	1999	Medium	Tier 1	Supermarket	1516.0266
FDA03	18.5	Regular	0.04546377	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket	2187.153
FDX32	15.1	Regular	0.1000135	Fruits and Ve	145.4786	OUT049	1999	Medium	Tier 1	Supermarket	1589.2646
FDS46	17.6	Regular	0.04725733	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket	2145.2076
FDF32	16.35	Low Fat	0.0680243	Fruits and Ve	196.4426	OUT013	1987	High	Tier 3	Supermarket	1977.426
FDP49	9	Regular	0.06908896	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket	1547.3192
NCB42	11.8	Low Fat	0.00859605	Health and H	115.3492	OUT018	2009	Medium	Tier 3	Supermarket	1621.8888
FDP49	9	Regular	0.06919638	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket	718.3982
DRI11		Low Fat	0.03423768	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket	2303.668
FDU02	13.35	Low Fat	0.10249212	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket	2748.4224
FDN22	18.85	Regular	0.13819028	Snack Foods	250.8724	OUT013	1987	High	Tier 3	Supermarket	3775.086
FDW12		Regular	0.03539992	Baking Good	144.5444	OUT027	1985	Medium	Tier 3	Supermarket	4064.0432
NCB30	14.6	Low Fat	0.02569813	Household	196.5084	OUT035	2004	Small	Tier 2	Supermarket	1587.2672
FDC37		Low Fat	0.057557	Baking Good	107.6938	OUT019	1985	Small	Tier 1	Grocery Stor	214.3876
FDR28	13.85	Regular	0.02589649	Frozen Foods	165.021	OUT046	1997	Small	Tier 1	Supermarket	4078.025
NCD06	13	Low Fat	0.0998871	Household	45.906	OUT017	2007		Tier 2	Supermarket	838.908
FDV10	7.645	Regular	0.06669344	Snack Foods	42.3112	OUT035	2004	Small	Tier 2	Supermarket	1065.28
DRJ59	11.65	low fat	0.01935613	Hard Drinks	39.1164	OUT013	1987	High	Tier 3	Supermarket	308.9312
FDE51	5.925	Regular	0.16146653	Dairy	45.5086	OUT010	1998		Tier 3	Grocery Stor	178.4344
FDC14		Regular	0.0722218	Canned	43.6454	OUT019	1985	Small	Tier 1	Grocery Stor	125.8362
FDV38	19.25	Low Fat	0.17034855	Dairy	55.7956	OUT010	1998		Tier 3	Grocery Stor	163.7868
NCS17	18.6	Low Fat	0.08082937	Health and H	96.4436	OUT018	2009	Medium	Tier 3	Supermarket	2741.7644
FDP33	18.7	Low Fat	0	Snack Foods	256.6672	OUT018	2009	Medium	Tier 3	Supermarket	3068.0064
FDO23	17.85	Low Fat	0	Breads	93.1436	OUT045	2002		Tier 2	Supermarket	2174.5028
DRH01	17.5	Low Fat	0.09790403	Soft Drinks	174.8738	OUT046	1997	Small	Tier 1	Supermarket	2085.2856
NCX29	10	Low Fat	0.08929114	Health and H	146.7102	OUT049	1999	Medium	Tier 1	Supermarket	3791.0652
FDV20		Regular	0.05951181	Fruits and Ve	128.0678	OUT027	1985	Medium	Tier 3	Supermarket	2797.6916
DRZ11	8.85	Regular	0.11312389	Soft Drinks	122.5388	OUT018	2009	Medium	Tier 3	Supermarket	1609.9044
FDX10		Regular	0.12311145	Snack Foods	36.9874	OUT027	1985	Medium	Tier 3	Supermarket	388.1614
FDB34		Low Fat	0.02648095	Snack Foods	87.6198	OUT027	1985	Medium	Tier 3	Supermarket	2180.495
FDU02	13.35	Low Fat	0.1025115	Dairy	230.6352	OUT046	1997	Small	Tier 1	Supermarket	3435.528
FDK43	9.8	Low Fat	0.02681843	Meat	126.002	OUT013	1987	High	Tier 3	Supermarket	2150.534
FDA46	13.6	Low Fat	0.11781835	Snack Foods	192.9136	OUT049	1999	Medium	Tier 1	Supermarket	2527.3768
FDC02	21.35	Low Fat	0.06910283	Canned	259.9278	OUT018	2009	Medium	Tier 3	Supermarket	6768.5228
FDL50	12.15	Regular	0.04227787	Canned	126.5046	OUT013	1987	High	Tier 3	Supermarket	373.5138
FDM39	6.42	LF	0.08949893	Dairy	178.1002	OUT010	1998		Tier 3	Grocery Stor	358.2004
NCP05	19.6	Low Fat	0	Health and H	153.3024	OUT045	2002		Tier 2	Supermarket	2428.8384
FDV49	10	Low Fat	0.02587958	Canned	265.2226	OUT045	2002		Tier 2	Supermarket	5815.0972
FDL12	15.85	Regular	0.12163272	Baking Good	60.622	OUT046	1997	Small	Tier 1	Supermarket	2576.646
FDS02		Regular	0.2553949	Dairy	196.8794	OUT019	1985	Small	Tier 1	Grocery Stor	780.3176
NCL17	7.39	Low Fat	0.06777971	Health and H	143.8812	OUT046	1997	Small	Tier 1	Supermarket	3134.5864
FDM40	10.195	Low Fat	0.15980385	Frozen Foods	141.5154	OUT013	1987	High	Tier 3	Supermarket	850.8924
FDR13	9.895	Regular	0.02869693	Canned	117.0492	OUT013	1987	High	Tier 3	Supermarket	810.9444
FDA43	10.895	Low Fat	0.06504158	Fruits and Ve	196.3794	OUT017	2007		Tier 2	Supermarket	3121.2704
NCP18	12.15	Low Fat	0.02876001	Household	151.4708	OUT017	2007		Tier 2	Supermarket	4815.0656

- Given a large dataset, how can we help retailers understand the properties of products and outlets to understand trends and predict sales?

Part 1: Set up

- Import libraries
- Create a DataFrame using pandas
- View first 5 rows of the DataFrame
- Examine features and values

Part 1: Set up

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.ensemble import RandomForestRegressor

df = pd.read_csv('/content/drive/MyDrive/Data Science Bootcamp/Project 1: Food Sales Prediction/sales_predictions.csv')

df.head()
```

Out[2]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Loca
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3

Part 2: Cleaning

- Explore dataframe (shape, datatypes)
- Check for and address any duplicates, missing values, inconsistent categories
- Obtain summary statistics of each numerical category (min, max, mean)

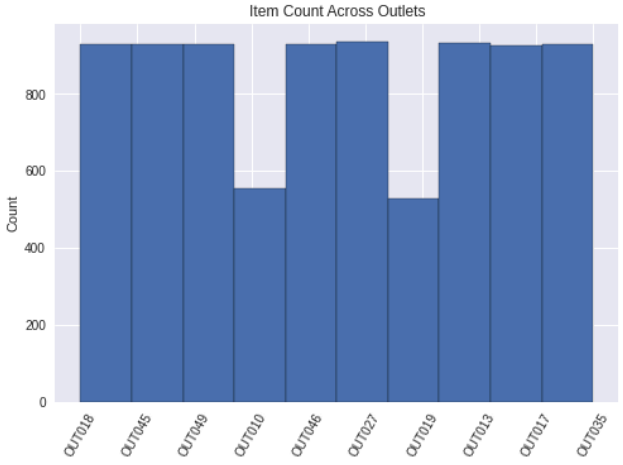
```
#5) Address the missing values
#missing values in Item_Weight and Outlet_size

#MISSING ITEM_WEIGHT
null_item_weight = pd.isnull(df['Item_Weight'])
null_item_weight_filter = df.loc[null_item_weight]
null_item_weight_filter

year_filter = df["Outlet_Establishment_Year"] == 1985
df[year_filter].describe()
#the Item_Weight values are missing when Outlet_Establishment_Year == 1985 (1463 total missing weight values)

df['Item_Identifier'].value_counts()

#sorting the dataframe by Item Identifier and Outlet Establishment Year so we can grab the missing item weights from the nearby rows
df.sort_values(by=['Item_Identifier', 'Outlet_Establishment_Year'], ascending=False, inplace=True)
df.head(50)

#filling in the missing values with the nearby row
df.loc[:, 'Item_Weight'].fillna(method='ffill', inplace=True)
```

Part 3: Statistical Analysis

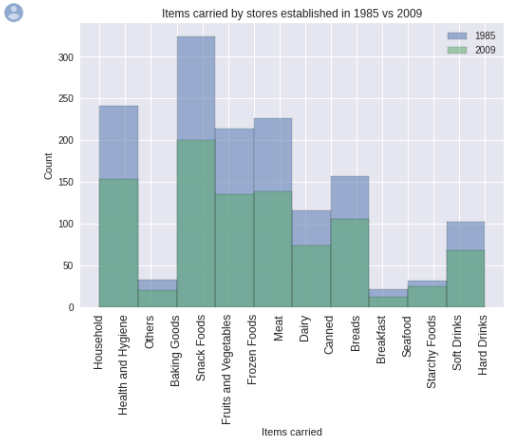
- Complete statistical analysis to help understand and explain data

Histograms

```
#Viewing distribution of quantity of items carried across individual outlets
plt.style.use('seaborn')
df['Outlet_Identifier'].hist(bins=10, edgecolor = "black")
plt.xticks(rotation = 60)
plt.title("Item Count Across Outlets")
plt.ylabel('Count');
```

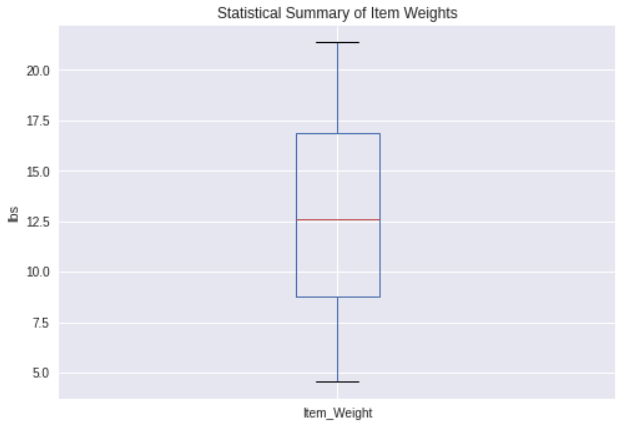



```
#Comparing distributions of item types in stores established in 1985 and 2009
year1985_filter = df["Outlet_Establishment_Year"] == 1985
year2009_filter = df["Outlet_Establishment_Year"] == 2009

plt.hist(df[year1985_filter]['Item_Type'], alpha=0.5, edgecolor="black", label="1985")
plt.hist(df[year2009_filter]['Item_Type'], alpha=0.5, edgecolor="black", label="2009")
plt.title('Items carried by stores established in 1985 vs 2009')
plt.xlabel('Items carried')
plt.ylabel('Count')
plt.legend()
plt.xticks(fontsize=12, rotation = 90);
```

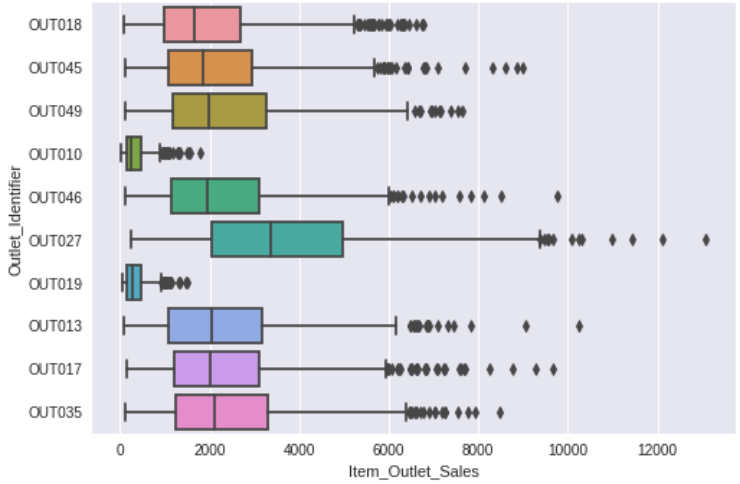


Boxplots

```
#Viewing boxplot of weight of items in the dataset
df.boxplot(['Item_Weight'])
plt.title('Statistical Summary of Item Weights')
plt.ylabel('lbs');
```

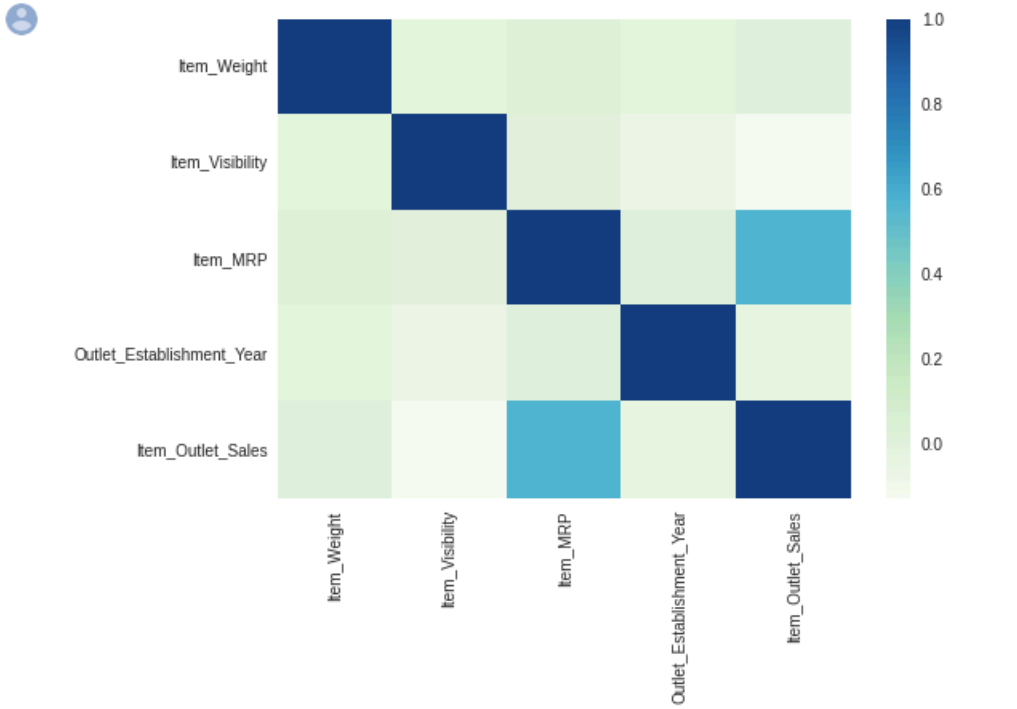



```
#Viewing statistical summaries of item outlet sales by inidividual outlets

sns.boxplot(x=df["Item_Outlet_Sales"], y=df["Outlet_Identifier"]);
```



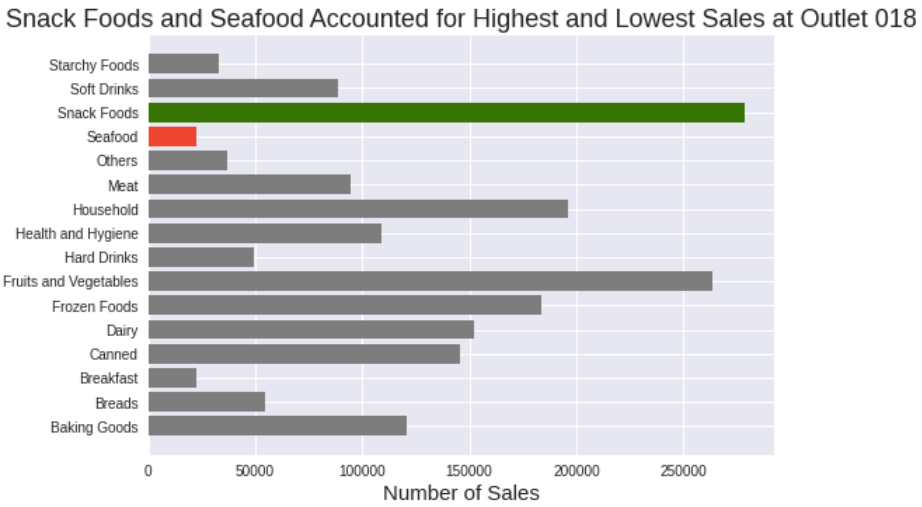
Correlations

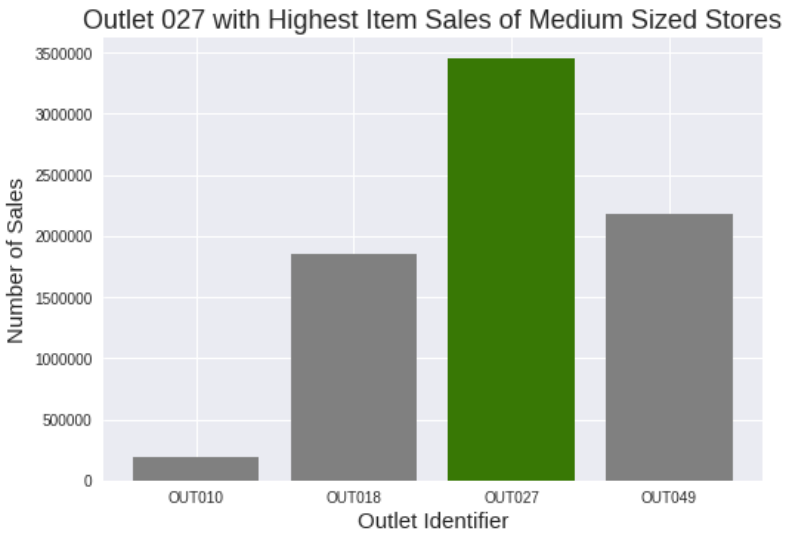
```
#3. Heatmap of the correlation between features

corr = df.corr()

sns.heatmap(corr, cmap="GnBu");
```



Part 4: Explanatory Analysis

- Build data visualizations to better understand trends in the data





Part 5: Linear Regression

- Create and evaluate a model to predict sales
- Transform categorical variables into numbers
- Use Dummy Encoding, OHE, and hashing where appropriate
- Assign the “Item_Outlet_Sales” column as the target and the remaining variables as the features matrix
- Train, test, split the data set
- Build a linear regression model
- Evaluate the test set results using R2 and RMSE

Evaluate the test set results using R2

```
[ ] print("R2 Train Score:", r2_score(y_train, reg_model.predict(X_train)))
    print("R2 Test Score:", r2_score(y_test, reg_model.predict(X_test)))
```

```
R2 Train Score: 0.5622383373007961
R2 Test Score: 0.5663209914828912
```

Evaluate the test set results using RMSE

```
[ ] print('Testing RMSE:', np.sqrt(mean_squared_error(y_test, reg_model.predict(X_test))))
```

```
Testing RMSE: 1093.8512397475317
```

Part 6: Tree Based Models

- Build and evaluate:
  - Decision Tree Model
  - Bagged Tree Model
  - Random Forest Model
- Compare R2 and RMSE scores
- Recommend using the Random Forest Model because of lowest RMSE and Highest R2 Test Scores

Conclusion

Item_Identif	Item_Weigh	Item_Fat_Co	Item_Visibili	Item_Type	Item_MRP	Outlet_Ident	Outlet_Estab	Outlet_Size	Outlet_Locat	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.0160473	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket	3735.138
DRC01	5.92	Regular	0.01927822	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket	443.4228
FDN15	17.5	Low Fat	0.01676008	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket	2097.27
FDX07	19.2	Regular	0	Fruits and Ve	182.095	OUT010	1998		Tier 3	Grocery Stor	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket	994.7052
FDP36	10.395	Regular	0	Baking Good	51.4008	OUT018	2009	Medium	Tier 3	Supermarket	556.6088
FDO10	13.65	Regular	0.01274109	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket	343.5528
FDP10		Low Fat	0.12746986	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket	4022.7636
FDH17	16.2	Regular	0.01668711	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket	1076.5986
FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket	4710.535
FDY07	11.8	Low Fat	0	Fruits and Ve	45.5402	OUT049	1999	Medium	Tier 1	Supermarket	1516.0266
FDA03	18.5	Regular	0.04546377	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket	2187.153
FDX32	15.1	Regular	0.1000135	Fruits and Ve	145.4786	OUT049	1999	Medium	Tier 1	Supermarket	1589.2646
FDS46	17.6	Regular	0.04725733	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket	2145.2076
FDF32	16.35	Low Fat	0.0680243	Fruits and Ve	196.4426	OUT013	1987	High	Tier 3	Supermarket	1977.426
FDP49	9	Regular	0.06908896	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket	1547.3192
NCB42	11.8	Low Fat	0.00859605	Health and H	115.3492	OUT018	2009	Medium	Tier 3	Supermarket	1621.8888
FDP49	9	Regular	0.06919638	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket	718.3982
DRI11		Low Fat	0.03423768	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket	2303.668
FDU02	13.35	Low Fat	0.10249212	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket	2748.4224
FDN22	18.85	Regular	0.13819028	Snack Foods	250.8724	OUT013	1987	High	Tier 3	Supermarket	3775.086
FDW12		Regular	0.03539992	Baking Good	144.5444	OUT027	1985	Medium	Tier 3	Supermarket	4064.0432
NCB30	14.6	Low Fat	0.02569813	Household	196.5084	OUT035	2004	Small	Tier 2	Supermarket	1587.2672
FDC37		Low Fat	0.057557	Baking Good	107.6938	OUT019	1985	Small	Tier 1	Grocery Stor	214.3876
FDR28	13.85	Regular	0.02589649	Frozen Foods	165.021	OUT046	1997	Small	Tier 1	Supermarket	4078.025
NCD06	13	Low Fat	0.0998871	Household	45.906	OUT017	2007		Tier 2	Supermarket	838.908
FDV10	7.645	Regular	0.06669344	Snack Foods	42.3112	OUT035	2004	Small	Tier 2	Supermarket	1065.28
DRJ59	11.65	low fat	0.01935613	Hard Drinks	39.1164	OUT013	1987	High	Tier 3	Supermarket	308.9312
FDE51	5.925	Regular	0.16146653	Dairy	45.5086	OUT010	1998		Tier 3	Grocery Stor	178.4344
FDC14		Regular	0.0722218	Canned	43.6454	OUT019	1985	Small	Tier 1	Grocery Stor	125.8362
FDV38	19.25	Low Fat	0.17034855	Dairy	55.7956	OUT010	1998		Tier 3	Grocery Stor	163.7868
NCS17	18.6	Low Fat	0.08082937	Health and H	96.4436	OUT018	2009	Medium	Tier 3	Supermarket	2741.7644
FDP33	18.7	Low Fat	0	Snack Foods	256.6672	OUT018	2009	Medium	Tier 3	Supermarket	3068.0064
FDO23	17.85	Low Fat	0	Breads	93.1436	OUT045	2002		Tier 2	Supermarket	2174.5028
DRH01	17.5	Low Fat	0.09790403	Soft Drinks	174.8738	OUT046	1997	Small	Tier 1	Supermarket	2085.2856
NCX29	10	Low Fat	0.08929114	Health and H	146.7102	OUT049	1999	Medium	Tier 1	Supermarket	3791.0652
FDV20		Regular	0.05951181	Fruits and Ve	128.0678	OUT027	1985	Medium	Tier 3	Supermarket	2797.6916
DRZ11	8.85	Regular	0.11312389	Soft Drinks	122.5388	OUT018	2009	Medium	Tier 3	Supermarket	1609.9044
FDX10		Regular	0.12311145	Snack Foods	36.9874	OUT027	1985	Medium	Tier 3	Supermarket	388.1614
FDB34		Low Fat	0.02648095	Snack Foods	87.6198	OUT027	1985	Medium	Tier 3	Supermarket	2180.495
FDU02	13.35	Low Fat	0.1025115	Dairy	230.6352	OUT046	1997	Small	Tier 1	Supermarket	3435.528
FDK43	9.8	Low Fat	0.02681843	Meat	126.002	OUT013	1987	High	Tier 3	Supermarket	2150.534
FDA46	13.6	Low Fat	0.11781835	Snack Foods	192.9136	OUT049	1999	Medium	Tier 1	Supermarket	2527.3768
FDC02	21.35	Low Fat	0.06910283	Canned	259.9278	OUT018	2009	Medium	Tier 3	Supermarket	6768.5228
FDL50	12.15	Regular	0.04227787	Canned	126.5046	OUT013	1987	High	Tier 3	Supermarket	373.5138
FDM39	6.42	LF	0.08949893	Dairy	178.1002	OUT010	1998		Tier 3	Grocery Stor	358.2004
NCP05	19.6	Low Fat	0	Health and H	153.3024	OUT045	2002		Tier 2	Supermarket	2428.8384
FDV49	10	Low Fat	0.02587958	Canned	265.2226	OUT045	2002		Tier 2	Supermarket	5815.0972
FDL12	15.85	Regular	0.12163272	Baking Good	60.622	OUT046	1997	Small	Tier 1	Supermarket	2576.646
FDS02		Regular	0.2553949	Dairy	196.8794	OUT019	1985	Small	Tier 1	Grocery Stor	780.3176
NCL17	7.39	Low Fat	0.06777971	Health and H	143.8812	OUT046	1997	Small	Tier 1	Supermarket	3134.5864
FDM40	10.195	Low Fat	0.15980385	Frozen Foods	141.5154	OUT013	1987	High	Tier 3	Supermarket	850.8924
FDR13	9.895	Regular	0.02869693	Canned	117.0492	OUT013	1987	High	Tier 3	Supermarket	810.9444
FDA43	10.895	Low Fat	0.06504158	Fruits and Ve	196.3794	OUT017	2007		Tier 2	Supermarket	3121.2704
NCP18	12.15	Low Fat	0.02876001	Household	151.4708	OUT017	2007		Tier 2	Supermarket	4815.0656

- Ask questions
- Get creative
- Stay patient