quantium

August 3, 2024

```
[6]: import pandas as pd
     import numpy as np
```

```
[10]: df=pd.read_excel(r"C:\Users\Divya\Downloads\QVI_transaction_data.xlsx")
```

```
[9]: pip install openpyxl
```

```
Requirement already satisfied: openpyxl in
c:\users\divya\appdata\local\programs\python\python312\lib\site-packages (3.1.5)
Requirement already satisfied: et-xmlfile in
c:\users\divya\appdata\local\programs\python\python312\lib\site-packages (from
openpyxl) (1.1.0)
Note: you may need to restart the kernel to use updated packages.

[notice] A new release of pip is available: 24.0 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
[11]: df
```

```
[11]:
         DATE  STORE_NBR  LYLTY_CARD_NBR  TXN_ID  PROD_NBR  \
0       43390          1            1000       1         5
1       43599          1            1307     348        66
2       43605          1            1343     383        61
3       43329          2            2373     974        69
4       43330          2            2426    1038       108
…       …        …             …     …       …
264831  43533        272          272319  270088        89
264832  43325        272          272358  270154        74
264833  43410        272          272379  270187        51
264834  43461        272          272379  270188        42
264835  43365        272          272380  270189        74

                                       PROD_NAME  PROD_QTY  TOT_SALES
0         Natural Chip        Compny SeaSalt175g         2        6.0
1                       CCs Nacho Cheese    175g         3        6.3
2         Smiths Crinkle Cut  Chips Chicken 170g         2        2.9
3         Smiths Chip Thinly  S/Cream&Onion 175g         5       15.0
```

```
4       Kettle Tortilla ChpsHny&Jlpno Chili 150g         3       13.8
…                                          …        …        …
264831   Kettle Sweet Chilli And Sour Cream 175g         2       10.8
264832             Tostitos Splash Of  Lime 175g         1        4.4
264833                  Doritos Mexicana    170g         2        8.8
264834   Doritos Corn Chip Mexican Jalapeno 150g         2        7.8
264835             Tostitos Splash Of  Lime 175g         2        8.8

[264836 rows x 8 columns]
```

[12]:
```
df.shape
```

[12]:
```
(264836, 8)
```

[13]:
```
df.describe
```

[13]:
```
<bound method NDFrame.describe of          DATE  STORE_NBR  LYLTY_CARD_NBR
TXN_ID  PROD_NBR  \
0       43390          1            1000       1         5
1       43599          1            1307     348        66
2       43605          1            1343     383        61
3       43329          2            2373     974        69
4       43330          2            2426    1038       108
…       …        …             …      …      …
264831  43533        272          272319  270088        89
264832  43325        272          272358  270154        74
264833  43410        272          272379  270187        51
264834  43461        272          272379  270188        42
264835  43365        272          272380  270189        74

                                       PROD_NAME  PROD_QTY  TOT_SALES
0         Natural Chip        Compny SeaSalt175g         2        6.0
1                         CCs Nacho Cheese    175g         3        6.3
2         Smiths Crinkle Cut  Chips Chicken 170g         2        2.9
3         Smiths Chip Thinly  S/Cream&Onion 175g         5       15.0
4       Kettle Tortilla ChpsHny&Jlpno Chili 150g         3       13.8
…                                          …        …        …
264831   Kettle Sweet Chilli And Sour Cream 175g         2       10.8
264832             Tostitos Splash Of  Lime 175g         1        4.4
264833                  Doritos Mexicana    170g         2        8.8
264834   Doritos Corn Chip Mexican Jalapeno 150g         2        7.8
264835             Tostitos Splash Of  Lime 175g         2        8.8

[264836 rows x 8 columns]>
```

[17]:
```
df.isnull().sum()
```

[17]:

```
DATE              0
STORE_NBR         0
LYLTY_CARD_NBR    0
TXN_ID            0
PROD_NBR          0
PROD_NAME         0
PROD_QTY          0
TOT_SALES         0
dtype: int64
```

[18]:

```
df.dropna()
```

[18]:

```
         DATE  STORE_NBR  LYLTY_CARD_NBR  TXN_ID  PROD_NBR  \
0       43390          1            1000       1         5
1       43599          1            1307     348        66
2       43605          1            1343     383        61
3       43329          2            2373     974        69
4       43330          2            2426    1038       108
…       …        …             …     …       …
264831  43533        272          272319  270088        89
264832  43325        272          272358  270154        74
264833  43410        272          272379  270187        51
264834  43461        272          272379  270188        42
264835  43365        272          272380  270189        74

                                       PROD_NAME  PROD_QTY  TOT_SALES
0         Natural Chip        Compny SeaSalt175g         2        6.0
1                       CCs Nacho Cheese    175g         3        6.3
2         Smiths Crinkle Cut  Chips Chicken 170g         2        2.9
3         Smiths Chip Thinly  S/Cream&Onion 175g         5       15.0
4       Kettle Tortilla ChpsHny&Jlpno Chili 150g         3       13.8
…                                          …       …        …
264831   Kettle Sweet Chilli And Sour Cream 175g         2       10.8
264832             Tostitos Splash Of  Lime 175g         1        4.4
264833                  Doritos Mexicana    170g         2        8.8
264834   Doritos Corn Chip Mexican Jalapeno 150g         2        7.8
264835             Tostitos Splash Of  Lime 175g         2        8.8

[264836 rows x 8 columns]
```

[19]:

```
df.head()
```

[19]:

```
    DATE  STORE_NBR  LYLTY_CARD_NBR  TXN_ID  PROD_NBR  \
0  43390          1            1000       1         5
1  43599          1            1307     348        66
2  43605          1            1343     383        61
3  43329          2            2373     974        69
```

```
4  43330          2            2426    1038       108

                                  PROD_NAME  PROD_QTY  TOT_SALES
0    Natural Chip        Compny SeaSalt175g         2        6.0
1                  CCs Nacho Cheese    175g         3        6.3
2    Smiths Crinkle Cut  Chips Chicken 170g         2        2.9
3    Smiths Chip Thinly  S/Cream&Onion 175g         5       15.0
4  Kettle Tortilla ChpsHny&Jlpno Chili 150g         3       13.8
```

[22]:
```
df["PROD_NAME"].value_counts()
```

[22]:
```
PROD_NAME
Kettle Mozzarella   Basil & Pesto 175g      3304
Kettle Tortilla ChpsHny&Jlpno Chili 150g    3296
Cobs Popd Swt/Chlli &Sr/Cream Chips 110g    3269
Tyrrells Crisps     Ched & Chives 165g      3268
Cobs Popd Sea Salt  Chips 110g              3265
                                            …
RRD Pc Sea Salt     165g                    1431
Woolworths Medium   Salsa 300g              1430
NCC Sour Cream &    Garden Chives 175g      1419
French Fries Potato Chips 175g              1418
WW Crinkle Cut      Original 175g           1410
Name: count, Length: 114, dtype: int64
```

[23]:
```
df["PROD_NAME"].unique
```

[23]:
```
<bound method Series.unique of 0           Natural Chip        Compny
SeaSalt175g
1                         CCs Nacho Cheese    175g
2           Smiths Crinkle Cut  Chips Chicken 170g
3           Smiths Chip Thinly  S/Cream&Onion 175g
4         Kettle Tortilla ChpsHny&Jlpno Chili 150g
                            …
264831     Kettle Sweet Chilli And Sour Cream 175g
264832               Tostitos Splash Of  Lime 175g
264833                    Doritos Mexicana    170g
264834     Doritos Corn Chip Mexican Jalapeno 150g
264835               Tostitos Splash Of  Lime 175g
Name: PROD_NAME, Length: 264836, dtype: object>
```

[24]:
```
df["TOT_SALES"].mean()
```

[24]:
```
7.3041995801175075
```

[25]:
```
df["TOT_SALES"].max()
```

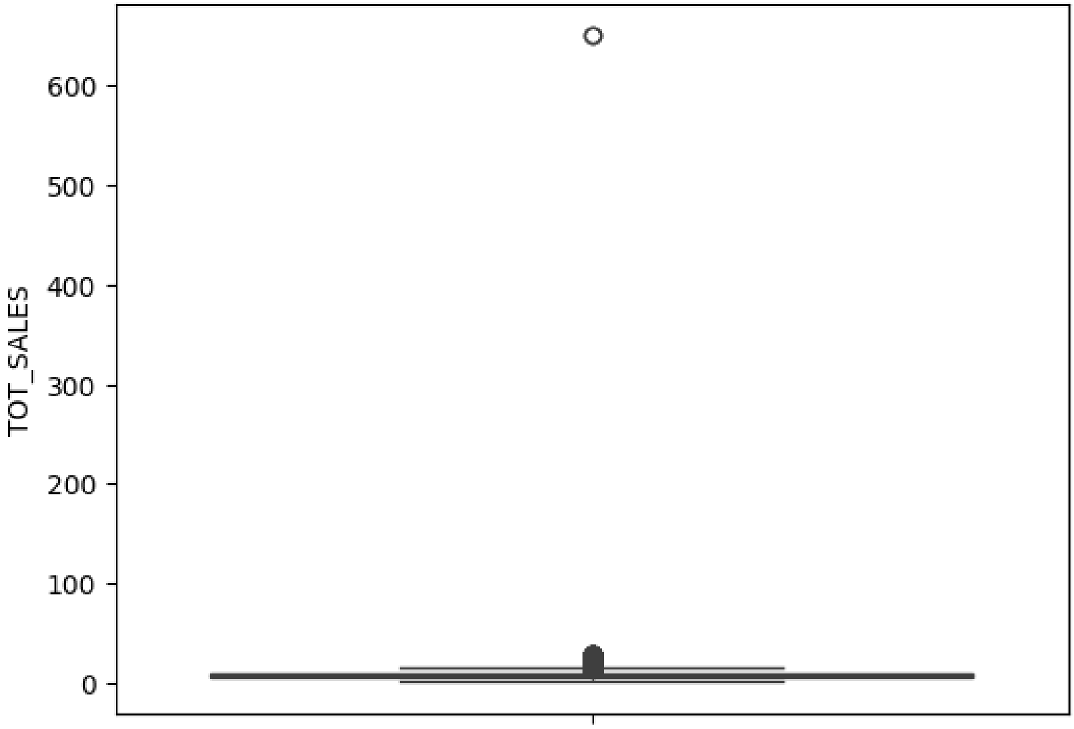
[25]: 650.0

```
[26]: df["TOT_SALES"].min()
```

[26]: 1.5

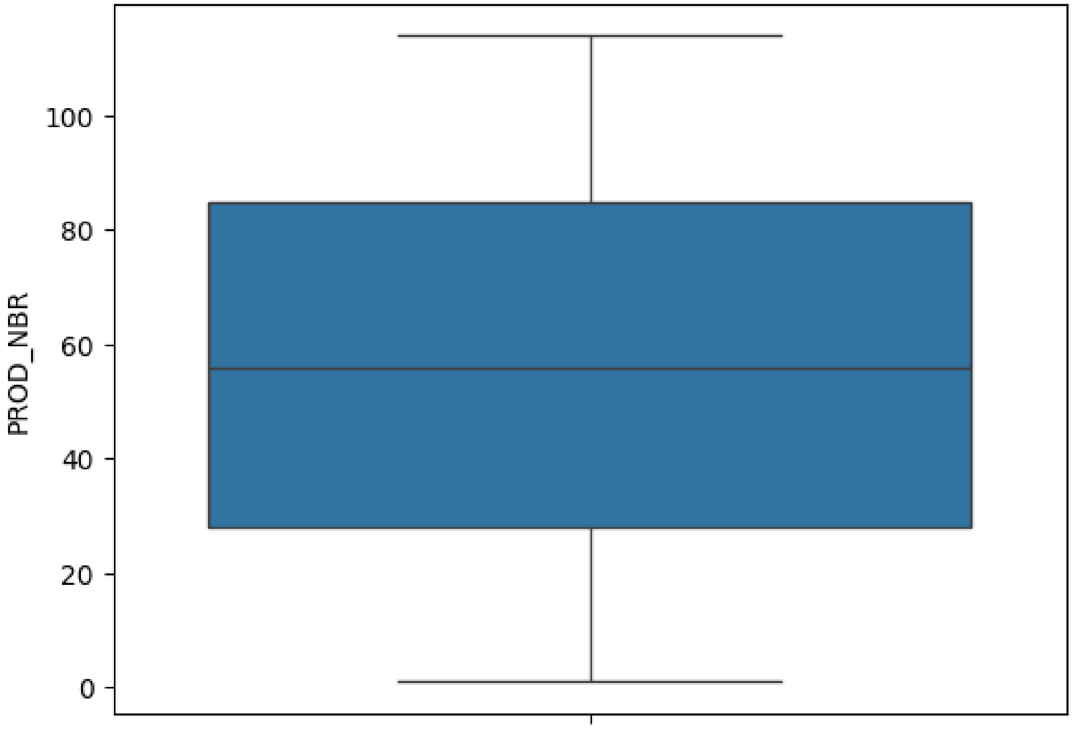
```
[27]: # Box Plot
      import seaborn as sns
      sns.boxplot(df['TOT_SALES'])
```

[27]: <Axes: ylabel='TOT_SALES'>



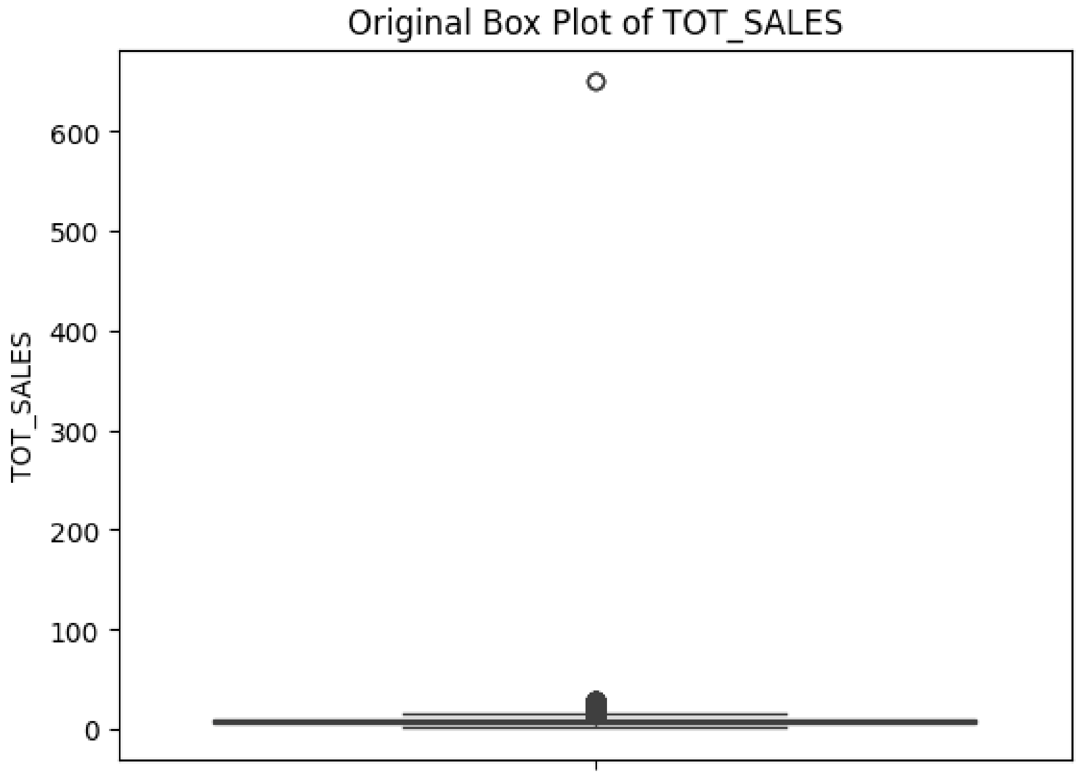
```
[29]: import seaborn as sns
      sns.boxplot(df['PROD_NBR'])
```

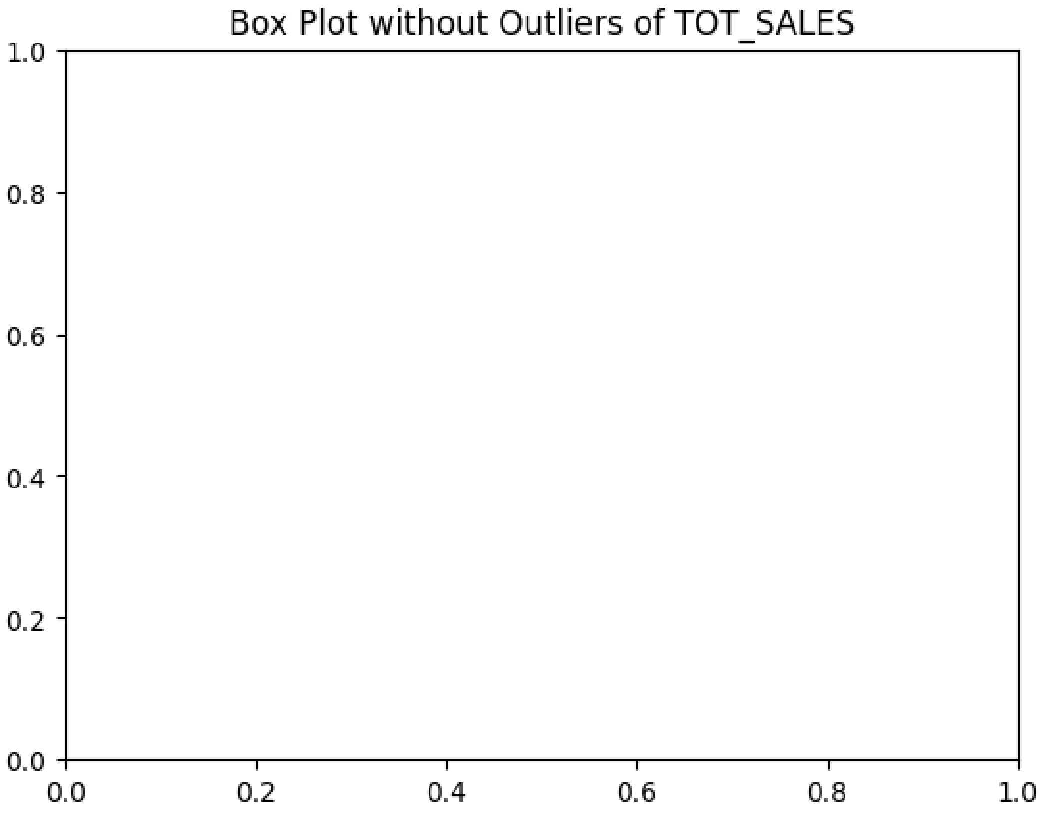
[29]: <Axes: ylabel='PROD_NBR'>



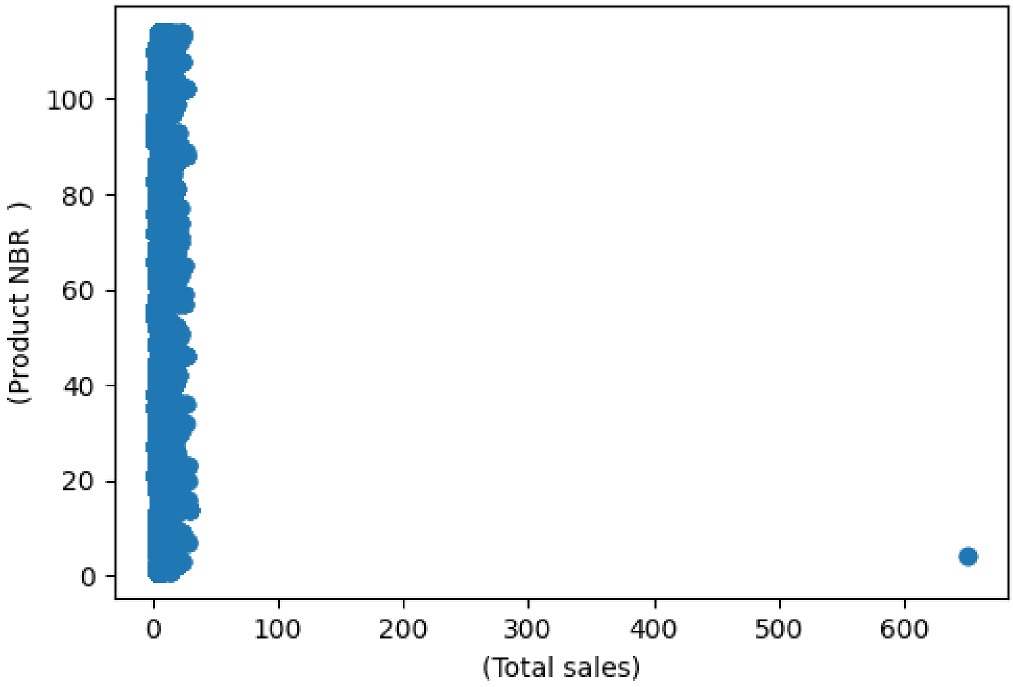
[30]:

```
import seaborn as sns
import matplotlib.pyplot as plt


def removal_box_plot(df, column, threshold):
        sns.boxplot(df[column])
        plt.title(f'Original Box Plot of {column}')
        plt.show()

        removed_outliers = df[df[column] <= threshold]

        sns.boxplot(removed_outliers[column])
        plt.title(f'Box Plot without Outliers of {column}')
        plt.show()
        return removed_outliers


threshold_value = 0.100

no_outliers = removal_box_plot(df, 'TOT_SALES', threshold_value)
```


Original Box Plot of TOT_SALES
TOT_SALES
0
100
200
300
400
500
600



[34]:
```
fig, ax = plt.subplots(figsize=(6, 4))
ax.scatter(df['TOT_SALES'], df['PROD_NBR'])
ax.set_xlabel('(Total sales)')
ax.set_ylabel('(Product NBR  )')
plt.show()
```



[32]:
```
df.dtypes
```

[32]:
```
DATE                 int64
STORE_NBR            int64
LYLTY_CARD_NBR       int64
TXN_ID               int64
PROD_NBR             int64
PROD_NAME           object
PROD_QTY             int64
TOT_SALES          float64
dtype: object
```

[33]:
```
df['PROD_NAME'] = df['PROD_NAME'].astype('string')
```

[35]:
```
df.dtypes
```

[35]:
```
DATE                             int64
STORE_NBR                        int64
LYLTY_CARD_NBR                   int64
TXN_ID                           int64
PROD_NBR                         int64
PROD_NAME           string[python]
PROD_QTY                         int64
TOT_SALES                      float64
```

```
dtype: object
```

[37]:

```
df["PROD_NAME_CLEAN"]=df["PROD_NAME"].str.replace("\d+g", "") #   Removing the␣
 ↪package sizes from the product names, and storing them in a separate column.
df["PROD_SIZE"]=df["PROD_NAME"].str.extract("(\d+)")     #   Extracting the␣
 ↪package sizes from the product names, and storing them in a separate column.
df["PROD_NAME"]=df["PROD_NAME_CLEAN"] #   Assigning the PROD_NAME_CLEAN column␣
 ↪to the PROD_NAME column.
df=df.drop("PROD_NAME_CLEAN", axis=1) #   Dropping the PROD_NAME_CLEAN column␣
 ↪from the pandas.DataFrame.
df["BRAND_NAME"]=df["PROD_NAME"].str.split().str[0]     #   Extracting the brand␣
 ↪names from the product names, and storing them in a separate column.
df=df.loc[:, ["DATE", "STORE_NBR", "LYLTY_CARD_NBR", "TXN_ID", "PROD_NBR",␣
 ↪"PROD_NAME", "PROD_SIZE", "BRAND_NAME", "PROD_QTY", "TOT_SALES"]]    #  ␣
 ↪Rearranging the columms of the pandas.DataFrame.
df
```

```
<>:1: SyntaxWarning: invalid escape sequence '\d'
<>:2: SyntaxWarning: invalid escape sequence '\d'
<>:1: SyntaxWarning: invalid escape sequence '\d'
<>:2: SyntaxWarning: invalid escape sequence '\d'
C:\Users\Divya\AppData\Local\Temp\ipykernel_21176\69326714.py:1: SyntaxWarning:
invalid escape sequence '\d'
  df["PROD_NAME_CLEAN"]=df["PROD_NAME"].str.replace("\d+g", "") #   Removing the
package sizes from the product names, and storing them in a separate column.
C:\Users\Divya\AppData\Local\Temp\ipykernel_21176\69326714.py:2: SyntaxWarning:
invalid escape sequence '\d'
  df["PROD_SIZE"]=df["PROD_NAME"].str.extract("(\d+)")     #   Extracting the
package sizes from the product names, and storing them in a separate column.
```

[37]:

```
          DATE  STORE_NBR  LYLTY_CARD_NBR  TXN_ID  PROD_NBR  \
0        43390          1            1000       1         5
1        43599          1            1307     348        66
2        43605          1            1343     383        61
3        43329          2            2373     974        69
4        43330          2            2426    1038       108
…          …        …              …       …       …
264831   43533        272          272319  270088        89
264832   43325        272          272358  270154        74
264833   43410        272          272379  270187        51
264834   43461        272          272379  270188        42
264835   43365        272          272380  270189        74

                                        PROD_NAME PROD_SIZE BRAND_NAME  \
0          Natural Chip        Compny SeaSalt175g       175    Natural
1                        CCs Nacho Cheese    175g       175        CCs
2          Smiths Crinkle Cut  Chips Chicken 170g       170     Smiths
```

```
3         Smiths Chip Thinly  S/Cream&Onion 175g       175     Smiths
4       Kettle Tortilla ChpsHny&Jlpno Chili 150g       150     Kettle
…                                          …       …        …
264831   Kettle Sweet Chilli And Sour Cream 175g       175     Kettle
264832             Tostitos Splash Of  Lime 175g       175   Tostitos
264833                  Doritos Mexicana    170g       170    Doritos
264834   Doritos Corn Chip Mexican Jalapeno 150g       150    Doritos
264835             Tostitos Splash Of  Lime 175g       175   Tostitos

        PROD_QTY  TOT_SALES
0              2        6.0
1              3        6.3
2              2        2.9
3              5       15.0
4              3       13.8
…          …        …
264831         2       10.8
264832         1        4.4
264833         2        8.8
264834         2        7.8
264835         2        8.8

[264836 rows x 10 columns]
```

[ ]: