Retail Strategy and Analytics

Data preparation and customer analytics

In [1]:
```
# in terminal (if needed)
!pip install pandas numpy matplotlib seaborn scikit-learn openpyxl notebook
```

```
Requirement already satisfied: pandas in c:\users\shubh\appdata\local\programs\pytho
n\python312\lib\site-packages (2.3.3)
Requirement already satisfied: numpy in c:\users\shubh\appdata\local\programs\python
\python312\lib\site-packages (2.3.4)
Requirement already satisfied: matplotlib in c:\users\shubh\appdata\local\programs\p
ython\python312\lib\site-packages (3.10.7)
Requirement already satisfied: seaborn in c:\users\shubh\appdata\local\programs\pyth
on\python312\lib\site-packages (0.13.2)
Requirement already satisfied: scikit-learn in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (1.7.2)
Requirement already satisfied: openpyxl in c:\users\shubh\appdata\local\programs\pyt
hon\python312\lib\site-packages (3.1.5)
Requirement already satisfied: notebook in c:\users\shubh\appdata\local\programs\pyt
hon\python312\lib\site-packages (7.4.7)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\shubh\appdata\loca
l\programs\python\python312\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\shubh\appdata\local\progra
ms\python\python312\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from matplotlib) (1.3.3)
Requirement already satisfied: cycler>=0.10 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from matplotlib) (4.60.1)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from matplotlib) (1.4.9)
Requirement already satisfied: packaging>=20.0 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in c:\users\shubh\appdata\local\programs\py
thon\python312\lib\site-packages (from matplotlib) (12.0.0)
Requirement already satisfied: pyparsing>=3 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from matplotlib) (3.2.5)
Requirement already satisfied: scipy>=1.8.0 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from scikit-learn) (1.16.3)
Requirement already satisfied: joblib>=1.2.0 in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from scikit-learn) (1.5.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from scikit-learn) (3.6.0)
Requirement already satisfied: et-xmlfile in c:\users\shubh\appdata\local\programs\p
ython\python312\lib\site-packages (from openpyxl) (2.0.0)
Requirement already satisfied: jupyter-server<3,>=2.4.0 in c:\users\shubh\appdata\lo
cal\programs\python\python312\lib\site-packages (from notebook) (2.17.0)
Requirement already satisfied: jupyterlab-server<3,>=2.27.1 in c:\users\shubh\appdat
a\local\programs\python\python312\lib\site-packages (from notebook) (2.28.0)
Requirement already satisfied: jupyterlab<4.5,>=4.4.9 in c:\users\shubh\appdata\loca
l\programs\python\python312\lib\site-packages (from notebook) (4.4.10)
Requirement already satisfied: notebook-shim<0.3,>=0.2 in c:\users\shubh\appdata\loc
al\programs\python\python312\lib\site-packages (from notebook) (0.2.4)
Requirement already satisfied: tornado>=6.2.0 in c:\users\shubh\appdata\local\progra
ms\python\python312\lib\site-packages (from notebook) (6.5.2)
Requirement already satisfied: anyio>=3.1.0 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook) (4.11.
0)
Requirement already satisfied: argon2-cffi>=21.1 in c:\users\shubh\appdata\local\pro
```

```
grams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook)
(25.1.0)
Requirement already satisfied: jinja2>=3.0.3 in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook) (3.1.
6)
Requirement already satisfied: jupyter-client>=7.4.4 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->noteboo
k) (8.6.3)
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in c:\users\shubh\appdata
\local\programs\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->n
otebook) (5.9.1)
Requirement already satisfied: jupyter-events>=0.11.0 in c:\users\shubh\appdata\loca
l\programs\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebo
ok) (0.12.0)
Requirement already satisfied: jupyter-server-terminals>=0.4.4 in c:\users\shubh\app
data\local\programs\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.
0->notebook) (0.5.3)
Requirement already satisfied: nbconvert>=6.4.4 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook)
(7.16.6)
Requirement already satisfied: nbformat>=5.3.0 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook) (5.
10.4)
Requirement already satisfied: prometheus-client>=0.9 in c:\users\shubh\appdata\loca
l\programs\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebo
ok) (0.23.1)
Requirement already satisfied: pywinpty>=2.0.1 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook) (3.
0.2)
Requirement already satisfied: pyzmq>=24 in c:\users\shubh\appdata\local\programs\py
thon\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook) (27.1.0)
Requirement already satisfied: send2trash>=1.8.2 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook)
(1.8.3)
Requirement already satisfied: terminado>=0.8.3 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook)
(0.18.1)
Requirement already satisfied: traitlets>=5.6.0 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->notebook)
(5.14.3)
Requirement already satisfied: websocket-client>=1.7 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from jupyter-server<3,>=2.4.0->noteboo
k) (1.9.0)
Requirement already satisfied: async-lru>=1.0.0 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from jupyterlab<4.5,>=4.4.9->notebook) (2.
0.5)
Requirement already satisfied: httpx<1,>=0.25.0 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from jupyterlab<4.5,>=4.4.9->notebook) (0.2
8.1)
Requirement already satisfied: ipykernel!=6.30.0,>=6.5.0 in c:\users\shubh\appdata\l
ocal\programs\python\python312\lib\site-packages (from jupyterlab<4.5,>=4.4.9->noteb
ook) (7.1.0)
Requirement already satisfied: jupyter-lsp>=2.0.0 in c:\users\shubh\appdata\local\pr
ograms\python\python312\lib\site-packages (from jupyterlab<4.5,>=4.4.9->notebook)
(2.3.0)
Requirement already satisfied: setuptools>=41.1.0 in c:\users\shubh\appdata\local\pr
```

```
ograms\python\python312\lib\site-packages (from jupyterlab<4.5,>=4.4.9->notebook) (8
0.9.0)
Requirement already satisfied: certifi in c:\users\shubh\appdata\local\programs\pyth
on\python312\lib\site-packages (from httpx<1,>=0.25.0->jupyterlab<4.5,>=4.4.9->noteb
ook) (2025.11.12)
Requirement already satisfied: httpcore==1.* in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from httpx<1,>=0.25.0->jupyterlab<4.5,>=4.4.9-
>notebook) (1.0.9)
Requirement already satisfied: idna in c:\users\shubh\appdata\local\programs\python
\python312\lib\site-packages (from httpx<1,>=0.25.0->jupyterlab<4.5,>=4.4.9->noteboo
k) (3.11)
Requirement already satisfied: h11>=0.16 in c:\users\shubh\appdata\local\programs\py
thon\python312\lib\site-packages (from httpcore==1.*->httpx<1,>=0.25.0->jupyterlab<
4.5,>=4.4.9->notebook) (0.16.0)
Requirement already satisfied: babel>=2.10 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from jupyterlab-server<3,>=2.27.1->notebook)
(2.17.0)
Requirement already satisfied: json5>=0.9.0 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from jupyterlab-server<3,>=2.27.1->notebook)
(0.12.1)
Requirement already satisfied: jsonschema>=4.18.0 in c:\users\shubh\appdata\local\pr
ograms\python\python312\lib\site-packages (from jupyterlab-server<3,>=2.27.1->notebo
ok) (4.25.1)
Requirement already satisfied: requests>=2.31 in c:\users\shubh\appdata\local\progra
ms\python\python312\lib\site-packages (from jupyterlab-server<3,>=2.27.1->notebook)
(2.32.5)
Requirement already satisfied: sniffio>=1.1 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from anyio>=3.1.0->jupyter-server<3,>=2.4.0->no
tebook) (1.3.1)
Requirement already satisfied: typing_extensions>=4.5 in c:\users\shubh\appdata\loca
l\programs\python\python312\lib\site-packages (from anyio>=3.1.0->jupyter-server<3,>
=2.4.0->notebook) (4.15.0)
Requirement already satisfied: argon2-cffi-bindings in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from argon2-cffi>=21.1->jupyter-server
<3,>=2.4.0->notebook) (25.1.0)
Requirement already satisfied: comm>=0.1.1 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from ipykernel!=6.30.0,>=6.5.0->jupyterlab<4.5,
>=4.4.9->notebook) (0.2.3)
Requirement already satisfied: debugpy>=1.6.5 in c:\users\shubh\appdata\local\progra
ms\python\python312\lib\site-packages (from ipykernel!=6.30.0,>=6.5.0->jupyterlab<4.
5,>=4.4.9->notebook) (1.8.17)
Requirement already satisfied: ipython>=7.23.1 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from ipykernel!=6.30.0,>=6.5.0->jupyterlab<
4.5,>=4.4.9->notebook) (9.7.0)
Requirement already satisfied: matplotlib-inline>=0.1 in c:\users\shubh\appdata\loca
l\programs\python\python312\lib\site-packages (from ipykernel!=6.30.0,>=6.5.0->jupyt
erlab<4.5,>=4.4.9->notebook) (0.2.1)
Requirement already satisfied: nest-asyncio>=1.4 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from ipykernel!=6.30.0,>=6.5.0->jupyterlab
<4.5,>=4.4.9->notebook) (1.6.0)
Requirement already satisfied: psutil>=5.7 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from ipykernel!=6.30.0,>=6.5.0->jupyterlab<4.5,
>=4.4.9->notebook) (7.1.3)
Requirement already satisfied: colorama>=0.4.4 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from ipython>=7.23.1->ipykernel!=6.30.0,>=6.
5.0->jupyterlab<4.5,>=4.4.9->notebook) (0.4.6)
```

```
Requirement already satisfied: decorator>=4.3.2 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from ipython>=7.23.1->ipykernel!=6.30.0,>=
6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (5.2.1)
Requirement already satisfied: ipython-pygments-lexers>=1.0.0 in c:\users\shubh\appd
ata\local\programs\python\python312\lib\site-packages (from ipython>=7.23.1->ipykern
el!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (1.1.1)
Requirement already satisfied: jedi>=0.18.1 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from ipython>=7.23.1->ipykernel!=6.30.0,>=6.5.0
->jupyterlab<4.5,>=4.4.9->notebook) (0.19.2)
Requirement already satisfied: prompt_toolkit<3.1.0,>=3.0.41 in c:\users\shubh\appda
ta\local\programs\python\python312\lib\site-packages (from ipython>=7.23.1->ipykerne
l!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (3.0.52)
Requirement already satisfied: pygments>=2.11.0 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from ipython>=7.23.1->ipykernel!=6.30.0,>=
6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (2.19.2)
Requirement already satisfied: stack_data>=0.6.0 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from ipython>=7.23.1->ipykernel!=6.30.0,>=
6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (0.6.3)
Requirement already satisfied: wcwidth in c:\users\shubh\appdata\local\programs\pyth
on\python312\lib\site-packages (from prompt_toolkit<3.1.0,>=3.0.41->ipython>=7.23.1-
>ipykernel!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (0.2.14)
Requirement already satisfied: parso<0.9.0,>=0.8.4 in c:\users\shubh\appdata\local\p
rograms\python\python312\lib\site-packages (from jedi>=0.18.1->ipython>=7.23.1->ipyk
ernel!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (0.8.5)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from jinja2>=3.0.3->jupyter-server<3,>=2.4.0
->notebook) (3.0.3)
Requirement already satisfied: attrs>=22.2.0 in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from jsonschema>=4.18.0->jupyterlab-server<3,>
=2.27.1->notebook) (25.4.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in c:\users\shub
h\appdata\local\programs\python\python312\lib\site-packages (from jsonschema>=4.18.0
->jupyterlab-server<3,>=2.27.1->notebook) (2025.9.1)
Requirement already satisfied: referencing>=0.28.4 in c:\users\shubh\appdata\local\p
rograms\python\python312\lib\site-packages (from jsonschema>=4.18.0->jupyterlab-serv
er<3,>=2.27.1->notebook) (0.37.0)
Requirement already satisfied: rpds-py>=0.7.1 in c:\users\shubh\appdata\local\progra
ms\python\python312\lib\site-packages (from jsonschema>=4.18.0->jupyterlab-server<3,
>=2.27.1->notebook) (0.29.0)
Requirement already satisfied: platformdirs>=2.5 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from jupyter-core!=5.0.*,>=4.12->jupyter-s
erver<3,>=2.4.0->notebook) (4.5.0)
Requirement already satisfied: python-json-logger>=2.0.4 in c:\users\shubh\appdata\l
ocal\programs\python\python312\lib\site-packages (from jupyter-events>=0.11.0->jupyt
er-server<3,>=2.4.0->notebook) (4.0.0)
Requirement already satisfied: pyyaml>=5.3 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from jupyter-events>=0.11.0->jupyter-server<3,>
=2.4.0->notebook) (6.0.3)
Requirement already satisfied: rfc3339-validator in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from jupyter-events>=0.11.0->jupyter-serve
r<3,>=2.4.0->notebook) (0.1.4)
Requirement already satisfied: rfc3986-validator>=0.1.1 in c:\users\shubh\appdata\lo
cal\programs\python\python312\lib\site-packages (from jupyter-events>=0.11.0->jupyte
r-server<3,>=2.4.0->notebook) (0.1.1)
Requirement already satisfied: fqdn in c:\users\shubh\appdata\local\programs\python
\python312\lib\site-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-events
```

```
>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (1.5.1)
Requirement already satisfied: isoduration in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter
-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (20.11.0)
Requirement already satisfied: jsonpointer>1.13 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from jsonschema[format-nongpl]>=4.18.0->jup
yter-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (3.0.0)
Requirement already satisfied: rfc3987-syntax>=1.1.0 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from jsonschema[format-nongpl]>=4.18.0
->jupyter-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (1.1.0)
Requirement already satisfied: uri-template in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter
-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (1.3.0)
Requirement already satisfied: webcolors>=24.6.0 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from jsonschema[format-nongpl]>=4.18.0->ju
pyter-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (25.10.0)
Requirement already satisfied: beautifulsoup4 in c:\users\shubh\appdata\local\progra
ms\python\python312\lib\site-packages (from nbconvert>=6.4.4->jupyter-server<3,>=2.
4.0->notebook) (4.14.2)
Requirement already satisfied: bleach!=5.0.0 in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from bleach[css]!=5.0.0->nbconvert>=6.4.4->jup
yter-server<3,>=2.4.0->notebook) (6.3.0)
Requirement already satisfied: defusedxml in c:\users\shubh\appdata\local\programs\p
ython\python312\lib\site-packages (from nbconvert>=6.4.4->jupyter-server<3,>=2.4.0->
notebook) (0.7.1)
Requirement already satisfied: jupyterlab-pygments in c:\users\shubh\appdata\local\p
rograms\python\python312\lib\site-packages (from nbconvert>=6.4.4->jupyter-server<3,
>=2.4.0->notebook) (0.3.0)
Requirement already satisfied: mistune<4,>=2.0.3 in c:\users\shubh\appdata\local\pro
grams\python\python312\lib\site-packages (from nbconvert>=6.4.4->jupyter-server<3,>=
2.4.0->notebook) (3.1.4)
Requirement already satisfied: nbclient>=0.5.0 in c:\users\shubh\appdata\local\progr
ams\python\python312\lib\site-packages (from nbconvert>=6.4.4->jupyter-server<3,>=2.
4.0->notebook) (0.10.2)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from nbconvert>=6.4.4->jupyter-server<
3,>=2.4.0->notebook) (1.5.1)
Requirement already satisfied: webencodings in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from bleach!=5.0.0->bleach[css]!=5.0.0->nbconve
rt>=6.4.4->jupyter-server<3,>=2.4.0->notebook) (0.5.1)
Requirement already satisfied: tinycss2<1.5,>=1.1.0 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from bleach[css]!=5.0.0->nbconvert>=6.
4.4->jupyter-server<3,>=2.4.0->notebook) (1.4.0)
Requirement already satisfied: fastjsonschema>=2.15 in c:\users\shubh\appdata\local
\programs\python\python312\lib\site-packages (from nbformat>=5.3.0->jupyter-server<
3,>=2.4.0->notebook) (2.21.2)
Requirement already satisfied: six>=1.5 in c:\users\shubh\appdata\local\programs\pyt
hon\python312\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Requirement already satisfied: charset_normalizer<4,>=2 in c:\users\shubh\appdata\lo
cal\programs\python\python312\lib\site-packages (from requests>=2.31->jupyterlab-ser
ver<3,>=2.27.1->notebook) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\shubh\appdata\local\pr
ograms\python\python312\lib\site-packages (from requests>=2.31->jupyterlab-server<3,
>=2.27.1->notebook) (2.5.0)
Requirement already satisfied: lark>=1.2.2 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from rfc3987-syntax>=1.1.0->jsonschema[format-n
```

```
ongpl]>=4.18.0->jupyter-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (1.3.1)
Requirement already satisfied: executing>=1.2.0 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from stack_data>=0.6.0->ipython>=7.23.1->ip
ykernel!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (2.2.1)
Requirement already satisfied: asttokens>=2.1.0 in c:\users\shubh\appdata\local\prog
rams\python\python312\lib\site-packages (from stack_data>=0.6.0->ipython>=7.23.1->ip
ykernel!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (3.0.1)
Requirement already satisfied: pure-eval in c:\users\shubh\appdata\local\programs\py
thon\python312\lib\site-packages (from stack_data>=0.6.0->ipython>=7.23.1->ipykerne
l!=6.30.0,>=6.5.0->jupyterlab<4.5,>=4.4.9->notebook) (0.2.3)
Requirement already satisfied: cffi>=1.0.1 in c:\users\shubh\appdata\local\programs
\python\python312\lib\site-packages (from argon2-cffi-bindings->argon2-cffi>=21.1->j
upyter-server<3,>=2.4.0->notebook) (2.0.0)
Requirement already satisfied: pycparser in c:\users\shubh\appdata\local\programs\py
thon\python312\lib\site-packages (from cffi>=1.0.1->argon2-cffi-bindings->argon2-cff
i>=21.1->jupyter-server<3,>=2.4.0->notebook) (2.23)
Requirement already satisfied: soupsieve>1.2 in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from beautifulsoup4->nbconvert>=6.4.4->jupyter
-server<3,>=2.4.0->notebook) (2.8)
Requirement already satisfied: arrow>=0.15.0 in c:\users\shubh\appdata\local\program
s\python\python312\lib\site-packages (from isoduration->jsonschema[format-nongpl]>=
4.18.0->jupyter-events>=0.11.0->jupyter-server<3,>=2.4.0->notebook) (1.4.0)
```

In [65]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
%matplotlib inline

plt.rcParams['figure.figsize'] = (10,5)
sns.set(style="whitegrid")
```

In [66]:
```
# Load Files and Quick Inspection

pb = pd.read_csv(r"C:\Users\shubh\Downloads\QVI_purchase_behaviour.csv")
tx = pd.read_excel(r"C:\Users\shubh\Downloads\QVI_transaction_data.xlsx")
```

In [67]:
```
# Info
print("Transactions shape:", tx.shape)
print(tx.dtypes)
print(tx.head())

print("Purchase behaviour shape:", pb.shape)
print(pb.dtypes)
print(pb.head())
```

```
Transactions shape: (264836, 8)
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
    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
Purchase behaviour shape: (72637, 3)
LYLTY_CARD_NBR       int64
LIFESTAGE           object
PREMIUM_CUSTOMER    object
dtype: object
   LYLTY_CARD_NBR               LIFESTAGE PREMIUM_CUSTOMER
0            1000   YOUNG SINGLES/COUPLES          Premium
1            1002   YOUNG SINGLES/COUPLES       Mainstream
2            1003          YOUNG FAMILIES           Budget
3            1004   OLDER SINGLES/COUPLES       Mainstream
4            1005  MIDAGE SINGLES/COUPLES       Mainstream
```

In [68]:
```
print("trans.shape:", tx.shape)
print("pb.shape:", pb.shape)
print("\ntrans.columns:", list(tx.columns))
print("pb.columns:", list(pb.columns))
```

```
trans.shape: (264836, 8)
pb.shape: (72637, 3)

trans.columns: ['DATE', 'STORE_NBR', 'LYLTY_CARD_NBR', 'TXN_ID', 'PROD_NBR', 'PROD_N
AME', 'PROD_QTY', 'TOT_SALES']
pb.columns: ['LYLTY_CARD_NBR', 'LIFESTAGE', 'PREMIUM_CUSTOMER']
```

In [69]:
```
# Convert to datetime
tx['DATE'] = pd.to_datetime(tx['DATE'],unit='D', origin='1899-12-30')
```

In [70]:
```
tx['PROD_NAME']
```

Out[70]:
```
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
```

In [71]:
```
# examine products that are not chips
unique_names = tx['PROD_NAME'].unique()

product_words = pd.DataFrame({
   "word" :  [w for name in unique_names for w in name.split() ]


})
```

In [72]:
```
product_words
```

Out[72]:

	word
0	Natural
1	Chip
2	Compny
3	SeaSalt175g
4	CCs
...	...
584	150g
585	Doritos
586	Salsa
587	Mild
588	300g

589 rows × 1 columns

In [73]:
```
product_words = product_words[~product_words['word'].str.contains(r'\d')]
```

In [ ]:
```

```

In [74]:
```
# Keep only words that contain A-Z
product_words = product_words[product_words['word'].str.contains(r'[A-Za-z]')]
```

In [75]:
```
product_words
```

Out[75]:

	word
0	Natural
1	Chip
2	Compny
4	CCs
5	Nacho
...	...
582	Cut
583	Bolognese
585	Doritos
586	Salsa
587	Mild

458 rows × 1 columns

In [76]:
```
#Count Frequency
product_words['word'].value_counts().head(20)
```

Out[76]:
```
word
Chips       21
Smiths      16
Cut         14
Crinkle     14
Kettle      13
Salt        12
Cheese      12
Original    10
Chip         9
Doritos      9
Salsa        9
Corn         8
Pringles     8
RRD          8
Chicken      7
WW           7
Sour         6
Sea          6
Vinegar      5
Thins        5
Name: count, dtype: int64
```

In [77]:
```
# Remove ttxhe ONLY non-chip item
tx['IS_SALSA'] = tx['PROD_NAME'].str.lower().str.contains("salsa")

chips_data = tx[~tx['IS_SALSA']].drop(columns=['IS_SALSA'])
```

In [78]:
```
# check for NULLS and possible outliers

tx.describe()
```

Out[78]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	
count	264836	264836.00000	2.648360e+05	2.648360e+05	264836.000000	2(
mean	2018-12-30 00:52:12.879215616	135.08011	1.355495e+05	1.351583e+05	56.583157	
min	2018-07-01 00:00:00	1.00000	1.000000e+03	1.000000e+00	1.000000	
25%	2018-09-30 00:00:00	70.00000	7.002100e+04	6.760150e+04	28.000000	
50%	2018-12-30 00:00:00	130.00000	1.303575e+05	1.351375e+05	56.000000	
75%	2019-03-31 00:00:00	203.00000	2.030942e+05	2.027012e+05	85.000000	
max	2019-06-30 00:00:00	272.00000	2.373711e+06	2.415841e+06	114.000000	
std	NaN	76.78418	8.057998e+04	7.813303e+04	32.826638	

Mean ~1.9

75th percentile = 2

Max = 200

In [79]:
```
### Filter the dataset to find the outlier

tx[tx['PROD_QTY'] == 200]
```

Out[79]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QT
69762	2018-08-19	226	226000	226201	4	Dorito Corn Chp Supreme 380g	2(
69763	2019-05-20	226	226000	226210	4	Dorito Corn Chp Supreme 380g	2(

There are two transactions where 200 packets of chips are bought in one transaction and both of these transactions where by the same customer It looks like this customer has only had the two transactions over the year and is not an ordinary retail customer. The customer might be buying chips for commercial purposes instead. We'll remove this loyalty card number from further analysis.

In [80]:
```
tx = tx[tx["LYLTY_CARD_NBR"] != 226000]
```

In [81]:
```
transaction_per_date = tx.groupby('DATE').size().reset_index(name="transactions")
```

In [82]:
```
transaction_per_date
```

Out[82]:

	DATE	transactions
0	2018-07-01	724
1	2018-07-02	711
2	2018-07-03	722
3	2018-07-04	714
4	2018-07-05	712
...	...	...
359	2019-06-26	723
360	2019-06-27	709
361	2019-06-28	730
362	2019-06-29	745
363	2019-06-30	744

364 rows × 2 columns

We have only 364 rows so 1 day is missing from the record between 1 July 2018 to 30 June 2019

In [83]:
```
all_dates = pd.DataFrame({
    'DATE': pd.date_range(start = tx['DATE'].min() , end= tx['DATE'].max(), freq='D
})
```

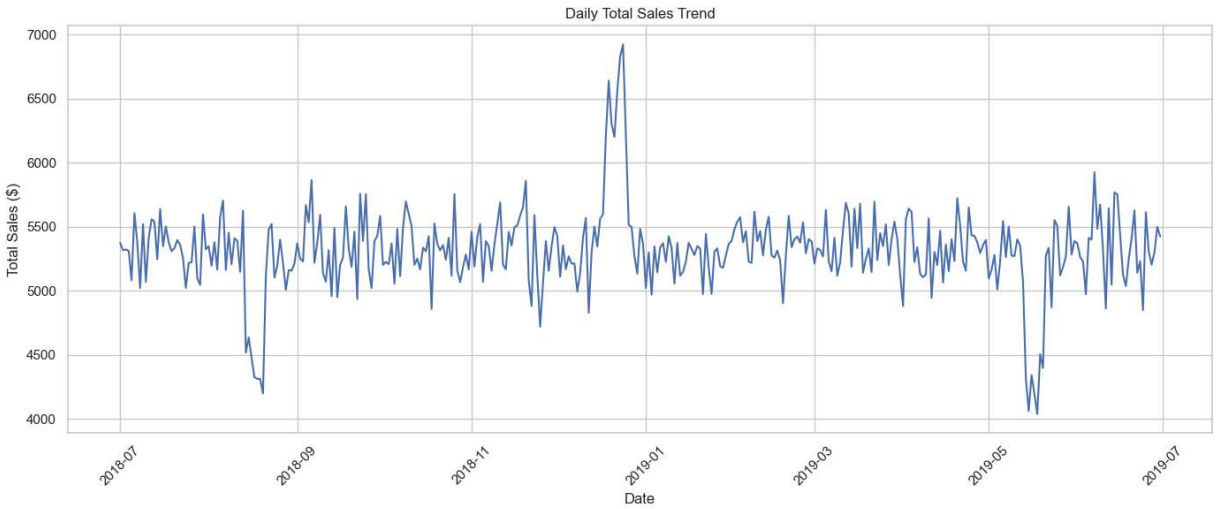
In [84]:
```
all_dates.head()
```

Out[84]:

	DATE
0	2018-07-01
1	2018-07-02
2	2018-07-03
3	2018-07-04
4	2018-07-05

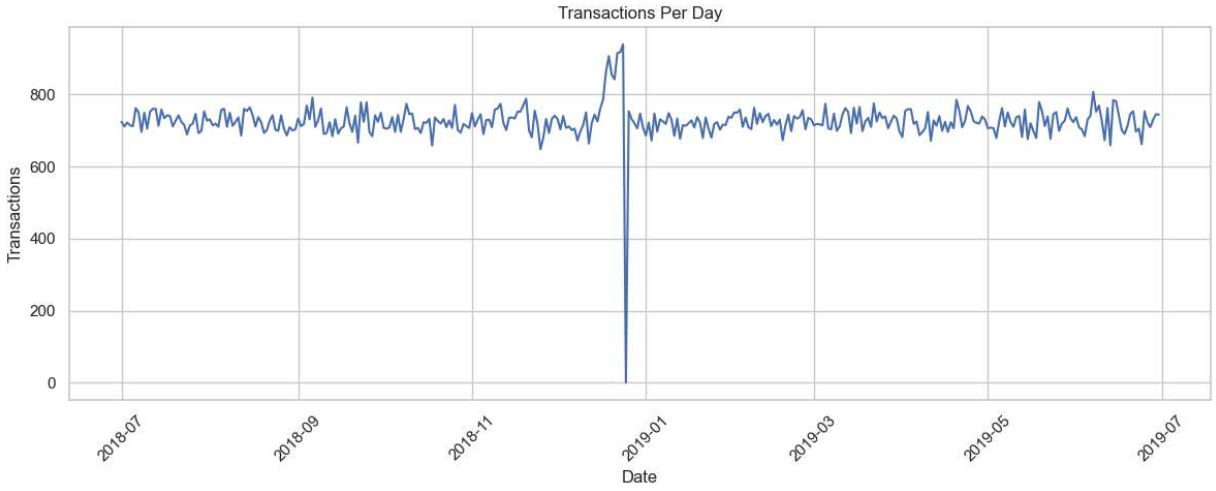
In [85]:
```
merged = all_dates.merge(transaction_per_date, on='DATE', how='left')
merged['transactions'] = merged['transactions'].fillna(0)
```

In [173…
```
import pandas as pd
import matplotlib.pyplot as plt

# Ensure DATE column is datetime
tx['DATE'] = pd.to_datetime(tx['DATE'])

# Group by DATE to get total sales per day
daily_sales = tx.groupby('DATE')['TOT_SALES'].sum().reset_index()

plt.figure(figsize=(14,6))
plt.plot(daily_sales['DATE'], daily_sales['TOT_SALES'])

plt.xlabel("Date")
plt.ylabel("Total Sales ($)")
plt.title("Daily Total Sales Trend")

plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()

plt.show()
```



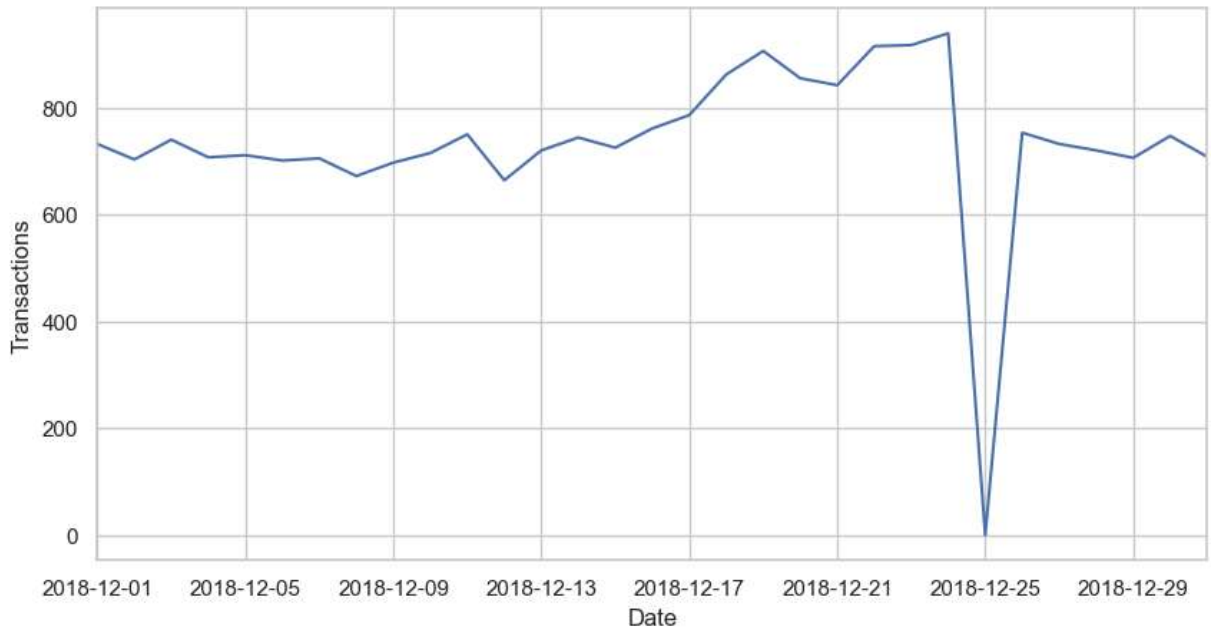
In [86]:
```
import matplotlib.pyplot as plt
```

```
plt.figure(figsize=(12,5))
plt.plot(merged['DATE'], merged['transactions'])
plt.xlabel("Date")
plt.ylabel("Transactions")

plt.title("Transactions Per Day")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



In [87]:
```
plt.plot(merged['DATE'], merged['transactions'])
plt.xlim(pd.to_datetime("2018-12-01"), pd.to_datetime("2018-12-31"))
plt.xlabel("Date")
plt.ylabel("Transactions")
```

Out[87]: `Text(0, 0.5, 'Transactions')`



We see that the sales have increased in December but sale went to zero on 25th .It can be due to christmas holiday

Extract the Pack Size and Prod_name

In [88]:
```
tx['PACK_SIZE'] = tx['PROD_NAME'].str.extract(r'(\d+)', expand=False).astype(int)
```

In [89]:
```
tx
```

Out[89]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD
0	2018-10-17	1	1000	1	5	Natural Chip Compny SeaSalt175g	
1	2019-05-14	1	1307	348	66	CCs Nacho Cheese 175g	
2	2019-05-20	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	
3	2018-08-17	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	
4	2018-08-18	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	
...	...	...	...	...	...	...	
264831	2019-03-09	272	272319	270088	89	Kettle Sweet Chilli And Sour Cream 175g	
264832	2018-08-13	272	272358	270154	74	Tostitos Splash Of Lime 175g	
264833	2018-11-06	272	272379	270187	51	Doritos Mexicana 170g	
264834	2018-12-27	272	272379	270188	42	Doritos Corn Chip Mexican Jalapeno 150g	
264835	2018-09-22	272	272380	270189	74	Tostitos Splash Of Lime 175g	

264834 rows × 10 columns

In [98]:
```
# Check if the pack_size looks sensible

tx.groupby("PACK_SIZE").size().reset_index(name="count").sort_values("PACK_SIZE")
```

Out[98]:

	PACK_SIZE	count
0	70	1507
1	90	3008
2	110	22387
3	125	1454
4	134	25102
5	135	3257
6	150	43131
7	160	2970
8	165	15297
9	170	19983
10	175	66390
11	180	1468
12	190	2995
13	200	4473
14	210	6272
15	220	1564
16	250	3169
17	270	6285
18	300	15166
19	330	12540
20	380	6416

min is 70g and maximum is 380g , which is reasonable

In [99]:
```
# EXTRACT THE BRAND NAME
```

In [103…
```
tx['BRAND'] = tx['PROD_NAME'].str.split(' ').str[0]
```

In [105…
```
tx['BRAND'].unique()
```

Out[105…
```
array(['Natural', 'CCs', 'Smiths', 'Kettle', 'Old', 'Grain', 'Doritos',
       'Twisties', 'WW', 'Thins', 'Burger', 'NCC', 'Cheezels', 'Infzns',
       'Red', 'Pringles', 'Dorito', 'Infuzions', 'Smith', 'GrnWves',
       'Tyrrells', 'Cobs', 'Woolworths', 'French', 'RRD', 'Tostitos',
       'Cheetos', 'Snbts', 'Sunbites'], dtype=object)
```

Some of the brand names look like they are of the same brands - such as RED and RRD, which are both Red Rock Deli chips. Let's combine these together.

In [107…
```
#Creating a mapping dictionary
brand_mapping = {
    "Red": "RRD",
    "Smith": "Smiths",
    "Dorito": "Doritos",
    "NCC": "Natural",
    "Grain": "GrnWves" ,
    "Snbts": "Sunbites",
    "Infzns": "Infuzions",
    "WW": "Woolworths"

}
```

In [108…
```
tx['BRAND'] = tx['BRAND'].replace(brand_mapping)
```

Analayze customer data

In [109…
```
pb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72637 entries, 0 to 72636
Data columns (total 3 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   LYLTY_CARD_NBR    72637 non-null  int64 
 1   LIFESTAGE         72637 non-null  object
 2   PREMIUM_CUSTOMER  72637 non-null  object
dtypes: int64(1), object(2)
memory usage: 1.7+ MB
```

In [111…
```
pb.describe()
```

Out[111…

	LYLTY_CARD_NBR
count	7.263700e+04
mean	1.361859e+05
std	8.989293e+04
min	1.000000e+03
25%	6.620200e+04
50%	1.340400e+05
75%	2.033750e+05
max	2.373711e+06

In [117…
```
pb['LIFESTAGE'].unique()
```

Out[117…
```
array(['YOUNG SINGLES/COUPLES', 'YOUNG FAMILIES', 'OLDER SINGLES/COUPLES',
       'MIDAGE SINGLES/COUPLES', 'NEW FAMILIES', 'OLDER FAMILIES',
       'RETIREES'], dtype=object)
```

In [118…
```
pb['LIFESTAGE'].value_counts()
```

Out[118…
```
LIFESTAGE
RETIREES                  14805
OLDER SINGLES/COUPLES     14609
YOUNG SINGLES/COUPLES     14441
OLDER FAMILIES             9780
YOUNG FAMILIES             9178
MIDAGE SINGLES/COUPLES     7275
NEW FAMILIES               2549
Name: count, dtype: int64
```

In [120…
```
pb['PREMIUM_CUSTOMER'].value_counts()
```

Out[120…
```
PREMIUM_CUSTOMER
Mainstream    29245
Budget        24470
Premium       18922
Name: count, dtype: int64
```

In [122…
```
pb.isnull().sum()
```

Out[122…
```
LYLTY_CARD_NBR      0
LIFESTAGE           0
PREMIUM_CUSTOMER    0
dtype: int64
```

In [125…
```
# Merge the transaction data and the purchase behavior data
df = tx.merge(pb, on="LYLTY_CARD_NBR" , how="left")
```

In [126…
```
df.isnull().sum()
```

Out[126…
```
DATE                0
STORE_NBR           0
LYLTY_CARD_NBR      0
TXN_ID              0
PROD_NBR            0
PROD_NAME           0
PROD_QTY            0
TOT_SALES           0
IS_SALSA            0
PACK_SIZE           0
BRAND               0
LIFESTAGE           0
PREMIUM_CUSTOMER    0
dtype: int64
```

In [127…
```
df.to_csv("output.csv" , index=False)
```

Data Analysis on customer segment

Now that the data is ready for analysis, we can define some metrics of interest to the client *Who spends the most on chips (total sales), describing customers by lifestage and how premium their general purchasing behaviour is

• How many customers are in each segment

• How many chips are bought per customer by segment

• What's the average chip price by customer segment

• The customer's total spend over the period and total spend for each transaction to understand what proportion of their grocery spend is on chips

• Proportion of customers in each customer segment overall to compare against the mix of customers who purchase chips

In [132…
```
#Total sales by LIFESTAGE and PREMIUM_CUSTOMER

sales = (df.groupby(['LIFESTAGE','PREMIUM_CUSTOMER'])['TOT_SALES']
.sum()
.reset_index(name = "SALES")
       )
```

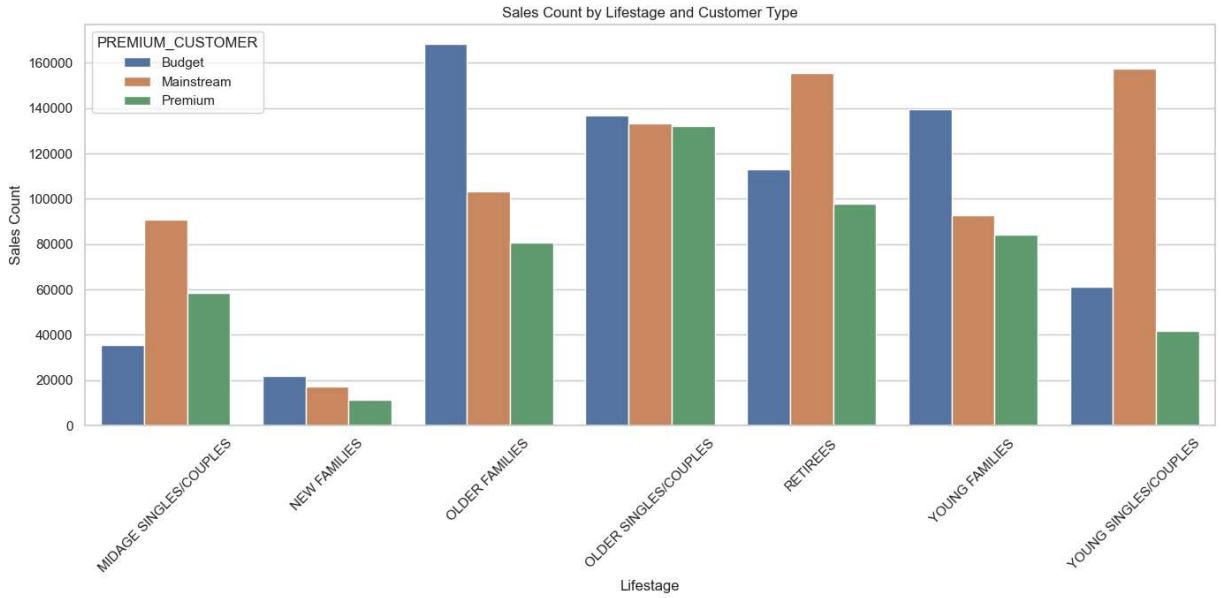
In [133…
```
df.head()
```

Out[133…

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY
0	2018-10-17	1	1000	1	5	Natural Chip Compny SeaSalt175g	2
1	2019-05-14	1	1307	348	66	CCs Nacho Cheese 175g	3
2	2019-05-20	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2
3	2018-08-17	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5
4	2018-08-18	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3

In [139…
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(14, 7))
sns.barplot(
    data=sales,
    x="LIFESTAGE",
```

```
    y="SALES",
    hue="PREMIUM_CUSTOMER"
)

plt.xticks(rotation=45)
plt.title("Sales Count by Lifestage and Customer Type")
plt.ylabel("Sales Count")
plt.xlabel("Lifestage")
plt.tight_layout()
plt.show()
```



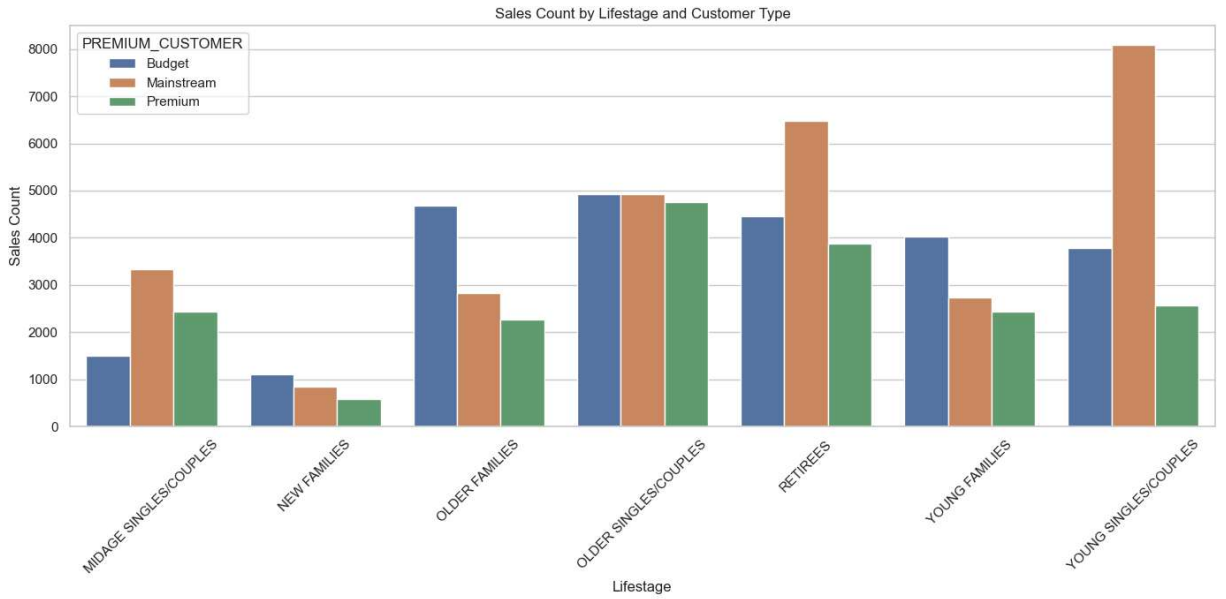
Sales are coming mainly from Budget - older families, Mainstream - young singles/couples, and Mainstream retirees

Let's see if the higher sales are due to there being more customers who buy chips.

In [143…
```
# The number of customer by LifeStage and PREMIUM_CUSTOMER

customers = pb.groupby(['LIFESTAGE' , 'PREMIUM_CUSTOMER'])['LYLTY_CARD_NBR'].nuniqu
```

In [144…
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(14, 7))
sns.barplot(
    data=customers,
    x="LIFESTAGE",
    y="Customers",
    hue="PREMIUM_CUSTOMER"
)

plt.xticks(rotation=45)
plt.title("Sales Count by Lifestage and Customer Type")
plt.ylabel("Sales Count")
plt.xlabel("Lifestage")
```

```
plt.tight_layout()
plt.show()
```

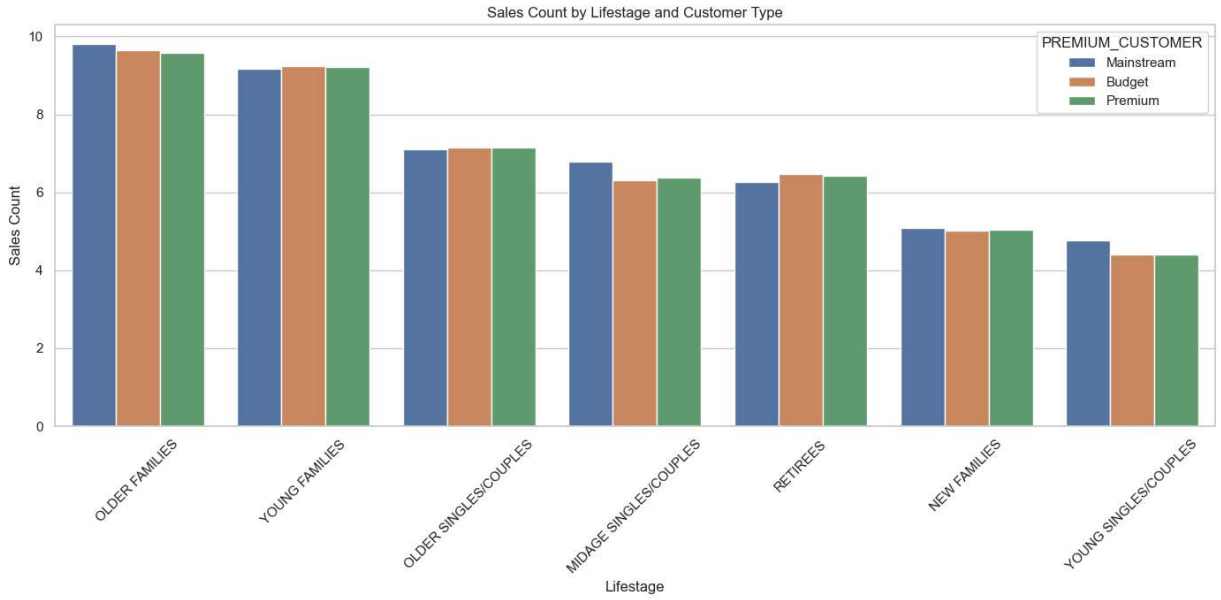


In [151…

```
#Average number of units per Customer
# Total prods / total customer
avg_units = (df.groupby(['LIFESTAGE' , 'PREMIUM_CUSTOMER'])
             .agg(
                 total_quantity = ('PROD_QTY' , "sum"),
                 unique_customers = ('LYLTY_CARD_NBR' , "nunique")
             )
             .assign(Avg_unit_per_cust= lambda d : d.total_quantity / d.unique_cust
             .reset_index()
             .sort_values("Avg_unit_per_cust", ascending=False))
```

In [153…

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

plt.figure(figsize=(14, 7))
sns.barplot(
    data=avg_units,
    x="LIFESTAGE",
    y="Avg_unit_per_cust",
    hue="PREMIUM_CUSTOMER"
)

plt.xticks(rotation=45)
plt.title("Sales Count by Lifestage and Customer Type")
plt.ylabel("Sales Count")
plt.xlabel("Lifestage")
plt.tight_layout()
plt.show()
```

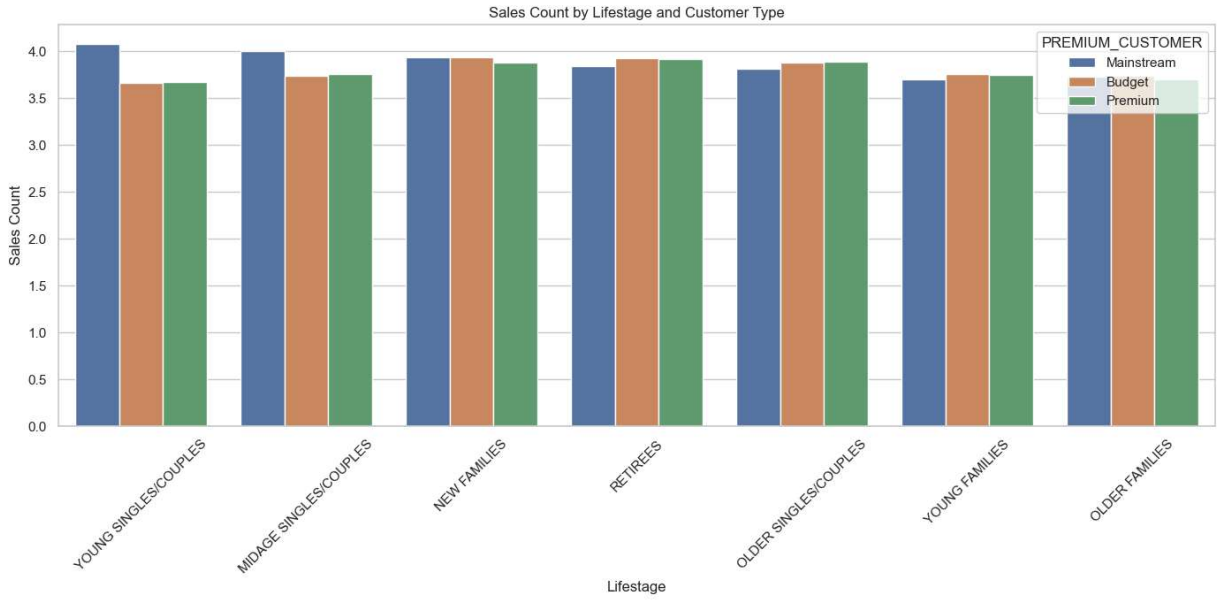


Older families and young families in general buy more chips per customer

we will also look at the av price per unit chip bought for each customer segment

In [157…
```
# Total sales / total prod quantity
avg_price_per_unit = (df.groupby(['LIFESTAGE' , 'PREMIUM_CUSTOMER'])
             .agg(
                 total_sales = ('TOT_SALES' , "sum"),
                 total_quantity = ('PROD_QTY' , "sum")
             )
             .assign(avg_price_per_unit= lambda d : d.total_sales / d.total_quantit
             .reset_index()
             .sort_values("avg_price_per_unit", ascending=False))
```

In [158…
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(14, 7))
sns.barplot(
    data=avg_price_per_unit,
    x="LIFESTAGE",
    y="avg_price_per_unit",
    hue="PREMIUM_CUSTOMER"
)

plt.xticks(rotation=45)
plt.title("Sales Count by Lifestage and Customer Type")
plt.ylabel("Sales Count")
plt.xlabel("Lifestage")
plt.tight_layout()
plt.show()
```



Compared to budget and premium counterparts, mainstream mid-age and young singles and couples are more willing to pay more for each packet of chips. This might be because high-end consumers are more likely to purchase nutritious snacks, and when they do, it's usually for amusement rather than personal consumption of chips. The decline in young and premium midage singles and couples further supports this. purchasing chips in contrast to those in the mainstream

Ques - "Is this difference just random noise? Or is it truly significant?"

We are comparing means

(price per chip unit)

✔ Between two independent groups

Group 1 → Mainstream young & mid-age singles/couples

Group 2 → Budget + Premium young/mid-age singles/couples

✔ The sample size is large

→ t-test works very well due to Central Limit Theorem.

✔ Price per unit is numeric and continuous

→ perfect case for independent Welch's t-test.

So the t-test is the standard, correct method here.

H_0 (null):

Mainstream and non-mainstream customers pay the same average price per unit.

H_1 (alternative):

Mainstream customers pay a higher price per unit.

In [163…
```
# t test

from scipy.stats import ttest_ind
```

In [164…
```
df['price'] = df['TOT_SALES'] / df['PROD_QTY']
mask_lifestage = df['LIFESTAGE'].isin(["YOUNG SINGLES/COUPLES", "MIDAGE SINGLES/COU
# Mainstream customers
mainstream = df[mask_lifestage & (df['PREMIUM_CUSTOMER'] == "Mainstream")]['price']
# Non-mainstream customers (Premium or Budget)
non_mainstream = df[mask_lifestage & (df['PREMIUM_CUSTOMER'] != "Mainstream")]['pri

# One-tailed t-test: alternative = "greater"
t_stat, p_value_two_tailed = ttest_ind(mainstream, non_mainstream, equal_var=False)

# Convert to one-tailed p-value
p_value_one_tailed = p_value_two_tailed / 2 if t_stat > 0 else 1 - (p_value_two_tai

print("Welch Two Sample t-test")
print("t =", t_stat)
print("p-value (one-tailed, greater) =", p_value_one_tailed)
print("mean of mainstream =", mainstream.mean())
print("mean of non-mainstream =", non_mainstream.mean())
```

```
Welch Two Sample t-test
t = 40.60989476220132
p-value (one-tailed, greater) = 0.0
mean of mainstream = 4.045586042532388
mean of non-mainstream = 3.688165443861052
```

the unit price for mainstream, young and mid-age singles and couples are significantly higher than that of budget or premium, young and midage singles and couples.

In [166…
```
import pandas as pd

# ----- Create the two segments -----

segment1 = df[
    (df["LIFESTAGE"] == "YOUNG SINGLES/COUPLES") &
    (df["PREMIUM_CUSTOMER"] == "Mainstream")
]

other = df[
    ~(
        (df["LIFESTAGE"] == "YOUNG SINGLES/COUPLES") &
        (df["PREMIUM_CUSTOMER"] == "Mainstream")
     )
]

# Total quantity purchased
quantity_segment1 = segment1["PROD_QTY"].sum()
quantity_other = other["PROD_QTY"].sum()
```

```
# ----- Brand proportions -----

quantity_segment1_by_brand = (
    segment1.groupby("BRAND")["PROD_QTY"]
    .sum()
    .reset_index(name="targetSegment")
)
quantity_segment1_by_brand["targetSegment"] /= quantity_segment1

quantity_other_by_brand = (
    other.groupby("BRAND")["PROD_QTY"]
    .sum()
    .reset_index(name="other")
)
quantity_other_by_brand["other"] /= quantity_other

brand_proportions = quantity_segment1_by_brand.merge(
    quantity_other_by_brand,
    on="BRAND",
    how="inner"
)

brand_proportions["affinityToBrand"] = (
    brand_proportions["targetSegment"] / brand_proportions["other"]
)

brand_proportions = brand_proportions.sort_values(
    "affinityToBrand", ascending=False
)

print("Brand Affinity:\n", brand_proportions)

# ----- Pack-size proportions -----

quantity_segment1_by_pack = (
    segment1.groupby("PACK_SIZE")["PROD_QTY"]
    .sum()
    .reset_index(name="targetSegment")
)
quantity_segment1_by_pack["targetSegment"] /= quantity_segment1

quantity_other_by_pack = (
    other.groupby("PACK_SIZE")["PROD_QTY"]
    .sum()
    .reset_index(name="other")
)
quantity_other_by_pack["other"] /= quantity_other

pack_proportions = quantity_segment1_by_pack.merge(
    quantity_other_by_pack,
    on="PACK_SIZE",
    how="inner"
)

pack_proportions["affinityToPack"] = (
```

```
    pack_proportions["targetSegment"] / pack_proportions["other"]
)

pack_proportions = pack_proportions.sort_values(
    "affinityToPack", ascending=False
)

print("Pack-size Affinity:\n", pack_proportions)

# To check what brands sell 270g packs:
brands_270 = df.loc[df["PACK_SIZE"] == 270, "PROD_NAME"].unique()
print("Products with 270g pack:\n", brands_270)
```

```
Brand Affinity:
          BRAND  targetSegment     other  affinityToBrand
19     Tyrrells       0.029587  0.023933         1.236235
18     Twisties       0.043306  0.035283         1.227401
9        Kettle       0.185649  0.154216         1.203823
17     Tostitos       0.042581  0.035377         1.203638
11          Old       0.041598  0.034753         1.196958
12     Pringles       0.111980  0.093743         1.194536
5       Doritos       0.122877  0.105277         1.167176
4          Cobs       0.041856  0.036375         1.150700
8     Infuzions       0.060649  0.053157         1.140947
16        Thins       0.056611  0.053084         1.066445
7       GrnWves       0.030674  0.029052         1.055825
3      Cheezels       0.016851  0.017370         0.970141
14       Smiths       0.093420  0.121714         0.767536
6        French       0.003702  0.005364         0.690113
2       Cheetos       0.007533  0.011240         0.670145
13          RRD       0.045377  0.068426         0.663149
10      Natural       0.018379  0.028741         0.639452
1           CCs       0.010484  0.017602         0.595599
15     Sunbites       0.005954  0.011719         0.508043
20  Woolworths       0.028189  0.057429         0.490854
0       Burger       0.002744  0.006145         0.446537
Pack-size Affinity:
     PACK_SIZE  targetSegment     other  affinityToPack
17         270       0.029846  0.023377        1.276694
20         380       0.030156  0.023832        1.265361
19         330       0.057465  0.046727        1.229814
4          134       0.111980  0.093743        1.194536
2          110       0.099658  0.083642        1.191482
14         210       0.027309  0.023401        1.167002
5          135       0.013849  0.012180        1.136997
16         250       0.013460  0.011905        1.130611
9          170       0.075740  0.075440        1.003980
18         300       0.054954  0.057263        0.959679
10         175       0.239102  0.251517        0.950641
6          150       0.155130  0.163446        0.949122
8          165       0.052185  0.058004        0.899681
12         190       0.007015  0.011590        0.605256
11         180       0.003365  0.005651        0.595459
7          160       0.006005  0.011526        0.521046
1           90       0.005954  0.011719        0.508043
3          125       0.002821  0.005623        0.501746
13         200       0.008413  0.017379        0.484086
0           70       0.002847  0.005889        0.483476
15         220       0.002744  0.006145        0.446537
Products with 270g pack:
 ['Twisties Cheese     270g' 'Twisties Chicken270g']
```

We will look into Mainstream - young singles/couples . we will find if they tend to buy a particular brand of chips

In [167…

```
# # ----- Create the two segments -----
```

In [168…

```
segment1 = df[
(df["LIFESTAGE"] == "YOUNG SINGLES/COUPLES") &
(df['PREMIUM_CUSTOMER'] == "Mainstream")
]

other = df[
    ~(
        (df["LIFESTAGE"] == "YOUNG SINGLES/COUPLES") &
        (df["PREMIUM_CUSTOMER"] == "Mainstream")
     )
]


quantity_segment1_by_brand = (
    segment1.groupby("BRAND")["PROD_QTY"]
    .sum()
    .reset_index(name="targetSegment")
)
quantity_segment1_by_brand["targetSegment"] /= quantity_segment1

quantity_other_by_brand = (
    other.groupby("BRAND")["PROD_QTY"]
    .sum()
    .reset_index(name="other")
)
quantity_other_by_brand["other"] /= quantity_other

brand_proportions = quantity_segment1_by_brand.merge(
    quantity_other_by_brand,
    on="BRAND",
    how="inner"
)

brand_proportions["affinityToBrand"] = (
    brand_proportions["targetSegment"] / brand_proportions["other"]
)

brand_proportions = brand_proportions.sort_values(
    "affinityToBrand", ascending=False
)

print("Brand Affinity:\n", brand_proportions)
```

```
Brand Affinity:
          BRAND  targetSegment     other  affinityToBrand
19     Tyrrells       0.029587  0.023933         1.236235
18     Twisties       0.043306  0.035283         1.227401
9        Kettle       0.185649  0.154216         1.203823
17     Tostitos       0.042581  0.035377         1.203638
11          Old       0.041598  0.034753         1.196958
12     Pringles       0.111980  0.093743         1.194536
5       Doritos       0.122877  0.105277         1.167176
4          Cobs       0.041856  0.036375         1.150700
8     Infuzions       0.060649  0.053157         1.140947
16        Thins       0.056611  0.053084         1.066445
7       GrnWves       0.030674  0.029052         1.055825
3      Cheezels       0.016851  0.017370         0.970141
14       Smiths       0.093420  0.121714         0.767536
6        French       0.003702  0.005364         0.690113
2       Cheetos       0.007533  0.011240         0.670145
13          RRD       0.045377  0.068426         0.663149
10      Natural       0.018379  0.028741         0.639452
1           CCs       0.010484  0.017602         0.595599
15     Sunbites       0.005954  0.011719         0.508043
20   Woolworths       0.028189  0.057429         0.490854
0        Burger       0.002744  0.006145         0.446537
```

Percentage likelihood uplift percentage uplift

percentage uplift=(affinityToBrand−1)×100

Mainstream young singles/couples are 23% more likely to purchase Tyrrells chips compared to the rest of the population

In [169…

```
quantity_segment1_by_pack = (
    segment1.groupby("PACK_SIZE")["PROD_QTY"]
    .sum()
    .reset_index(name="targetSegment")
)
quantity_segment1_by_pack["targetSegment"] /= quantity_segment1

quantity_other_by_pack = (
    other.groupby("PACK_SIZE")["PROD_QTY"]
    .sum()
    .reset_index(name="other")
)
quantity_other_by_pack["other"] /= quantity_other

pack_proportions = quantity_segment1_by_pack.merge(
    quantity_other_by_pack,
    on="PACK_SIZE",
    how="inner"
)
```

```
pack_proportions["affinityToPack"] = (
    pack_proportions["targetSegment"] / pack_proportions["other"]
)

pack_proportions = pack_proportions.sort_values(
    "affinityToPack", ascending=False
)

print("Pack-size Affinity:\n", pack_proportions)

# To check what brands sell 270g packs:
brands_270 = df.loc[df["PACK_SIZE"] == 270, "PROD_NAME"].unique()
print("Products with 270g pack:\n", brands_270)
```

```
Pack-size Affinity:
     PACK_SIZE  targetSegment     other  affinityToPack
17         270       0.029846  0.023377        1.276694
20         380       0.030156  0.023832        1.265361
19         330       0.057465  0.046727        1.229814
4          134       0.111980  0.093743        1.194536
2          110       0.099658  0.083642        1.191482
14         210       0.027309  0.023401        1.167002
5          135       0.013849  0.012180        1.136997
16         250       0.013460  0.011905        1.130611
9          170       0.075740  0.075440        1.003980
18         300       0.054954  0.057263        0.959679
10         175       0.239102  0.251517        0.950641
6          150       0.155130  0.163446        0.949122
8          165       0.052185  0.058004        0.899681
12         190       0.007015  0.011590        0.605256
11         180       0.003365  0.005651        0.595459
7          160       0.006005  0.011526        0.521046
1           90       0.005954  0.011719        0.508043
3          125       0.002821  0.005623        0.501746
13         200       0.008413  0.017379        0.484086
0           70       0.002847  0.005889        0.483476
15         220       0.002744  0.006145        0.446537
Products with 270g pack:
 ['Twisties Cheese     270g' 'Twisties Chicken270g']
```

In [ ]: