Importing Liabraries

In [1]:

```
# for data manipulation

import numpy as np
import pandas as pd


# for data visualization

import matplotlib.pyplot as plt
import seaborn as sns

# for Fearure Extraction and Recommendation System

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# miscellaneous

import warnings
warnings.filterwarnings("ignore")
```

Data Gathering

In [2]:

```
df = pd.read_csv(r"C:\Users\Dell\Downloads\store_zara.csv")
df.head()
```

Out[2]:

	brand	url	sku	name	description	price	currency	images
0	Zara	https://www.zara.com/us/en/basic-puffer-jacket...	272145190-250-2	BASIC PUFFER JACKET	Puffer jacket made of tear-resistant ripstop f...	19.99	USD	['https://static.zara.net/photos///2023/I/0/2/...
1	Zara	https://www.zara.com/us/en/tuxedo-jacket-p0889...	324052738-800-46	TUXEDO JACKET	Straight fit blazer. Pointed lapel collar and ...	169.00	USD	['https://static.zara.net/photos///2024/V/0/1/...
2	Zara	https://www.zara.com/us/en/slim-fit-suit-jacke...	335342680-800-44	SLIM FIT SUIT JACKET	Slim fit jacket. Notched lapel collar. Long sl...	129.00	USD	['https://static.zara.net/photos///2023/I/0/2/...
3	Zara	https://www.zara.com/us/en/stretch-suit-jacket...	328303236-420-44	STRETCH SUIT JACKET	Slim fit jacket made of viscose blend fabric. ...	129.00	USD	['https://static.zara.net/photos///2024/V/0/1/...
4	Zara	https://www.zara.com/us/en/double-faced-jacket...	312368260-800-2	DOUBLE FACED JACKET	Jacket made of faux leather faux shearling wit...	139.00	USD	['https://static.zara.net/photos///2024/V/0/2/...

In [3]:

```
df.columns
```

Out[3]:

```
Index(['brand', 'url', 'sku', 'name', 'description', 'price', 'currency',
       'images', 'scraped_at', 'terms', 'section', 'error', 'image_downloads'],
      dtype='object')
```

In [4]:

```
# drop unnecessary columns
```

In [5]:

```
df.drop(["brand","url","currency","images","scraped_at","error","image_downloads"],axis=1,inplace=True)
```

In [6]:

```
df
```

Out[6]:

	sku	name	description	price	terms	section
0	272145190-250-2	BASIC PUFFER JACKET	Puffer jacket made of tear-resistant ripstop f...	19.99	jackets	MAN
1	324052738-800-46	TUXEDO JACKET	Straight fit blazer. Pointed lapel collar and ...	169.00	jackets	MAN
2	335342680-800-44	SLIM FIT SUIT JACKET	Slim fit jacket. Notched lapel collar. Long sl...	129.00	jackets	MAN
3	328303236-420-44	STRETCH SUIT JACKET	Slim fit jacket made of viscose blend fabric. ...	129.00	jackets	MAN
4	312368260-800-2	DOUBLE FACED JACKET	Jacket made of faux leather faux shearling wit...	139.00	jackets	MAN
...	...	...	...	...	...	...
3124	311307129-999-99	TUBEROSE 100 ML	ZARA TUBEROSE GLITTER EDP 100 ML (3.4 FL. OZ)....	22.90	bags	WOMAN
3125	311287165-712-3	WOOL ALPACA FRINGED SCARF	Scarf made of 14% alpaca and 14% wool. Fringed...	49.90	bags	WOMAN
3126	311302993-712-3	ALPACA AND WOOL BLEND SEQUIN SCARF	Alpaca blend scarf and sequin appliqués.\n\nDI...	49.90	bags	WOMAN
3127	326448329-999-99	NaN	NaN	27.90	bags	WOMAN
3128	323826890-709-3	CROSSBODY BAG	Crossbody bag with flap. Removable and adjusta...	29.99	bags	WOMAN

3129 rows × 6 columns

Exploratory Data Analysis

In [7]:

```
# basic information
```

In [8]:

```
df.shape
```

Out[8]:

```
(3129, 6)
```

In [9]:

```
df.columns
```

Out[9]:

```
Index(['sku', 'name', 'description', 'price', 'terms', 'section'], dtype='object')
```

In [10]:

```
df["terms"].unique()
```

Out[10]:

```
array(['jackets', 'puffers', 'pants', 'jeans', 'sweaters', 'cardigans',
       'hoodies', 'sweatshirts', 't-shirts', 'overshirts', 'linen',
       'shorts', 'suits', 'blazers', 'tracksuits', 'coats', 'shoes',
       'bags', 'dresses', 'skirts', 'tops', 'bodysuits', 'knitwear'],
      dtype=object)
```

In [11]:

```
# overall data information
```

In [12]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3129 entries, 0 to 3128
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   sku          3129 non-null   object 
 1   name         3065 non-null   object 
 2   description  3059 non-null   object 
 3   price        3129 non-null   float64
 4   terms        3129 non-null   object 
 5   section      3129 non-null   object 
dtypes: float64(1), object(5)
memory usage: 146.8+ KB
```

In [13]:

```
# statistical information
```

In [14]:

```
df.describe().T                                                    # because only two numerical co
```

Out[14]:

	count	mean	std	min	25%	50%	75%	max
price	3129.0	64.10078	49.492635	1.99	39.9	49.9	69.9	869.0

In [15]:

```
# check for missing values
```

In [16]:

```
df.isna().sum()
```

Out[16]:

```
sku             0
name           64
description    70
price           0
terms           0
section         0
dtype: int64
```

In [17]:

```
df = df.dropna()
```

In [18]:

```
df.isna().sum()
```

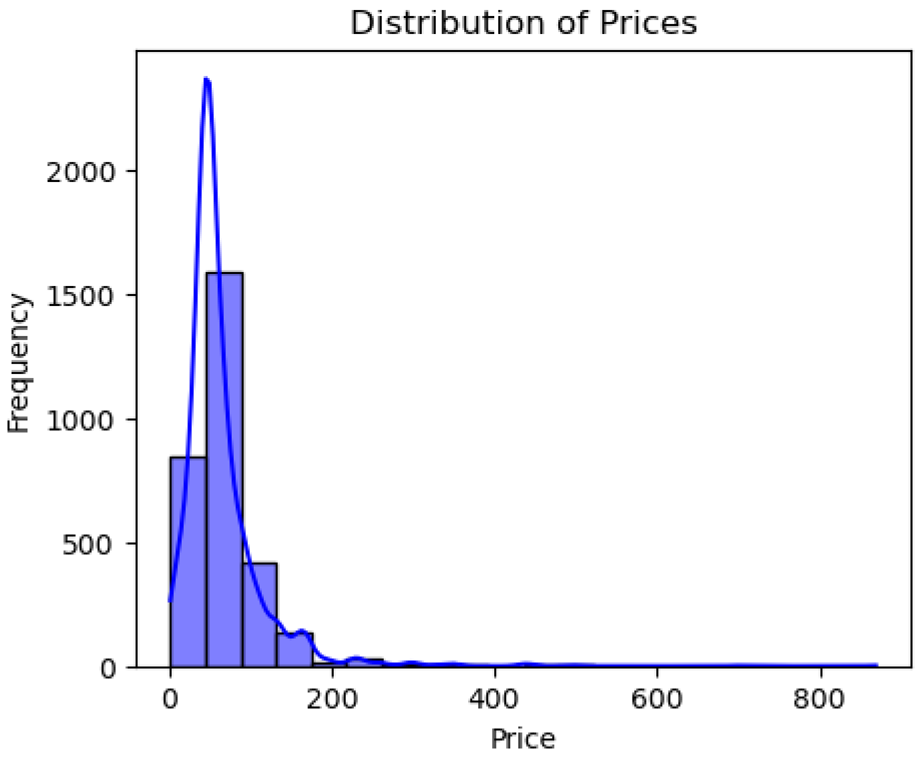
Out[18]:

```
sku            0
name           0
description    0
price          0
terms          0
section        0
dtype: int64
```

Data Visualization

In [19]:

```
# Histogram of 'price'
plt.figure(figsize=(5, 4))
sns.histplot(df["price"], bins=20, kde=True,color='blue')
plt.title("Distribution of Prices")
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.show()
```



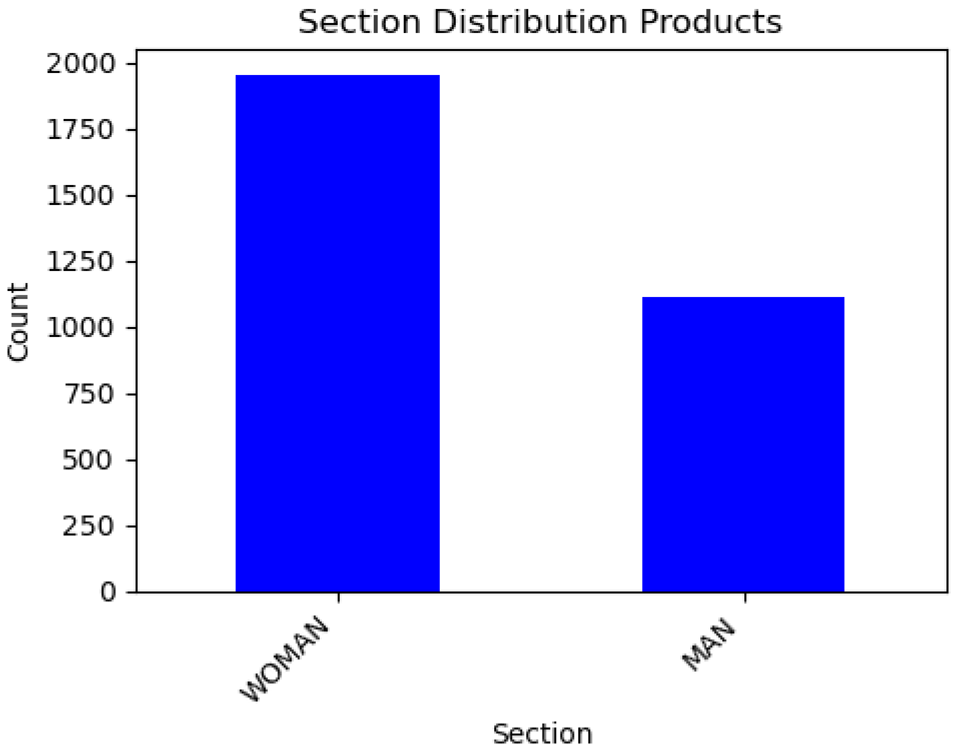
In [20]:

```
# The above histplot illustrates that over 1500 products are priced within the range of 0-200 USD.
```

In [21]:

```
# Assuming 'section' column contains the section information
section_distribution = df['section'].value_counts()

# Plotting the distribution
plt.figure(figsize=(5, 4))
section_distribution.plot(kind='bar', color='blue')
plt.title('Section Distribution Products')
plt.xlabel('Section')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



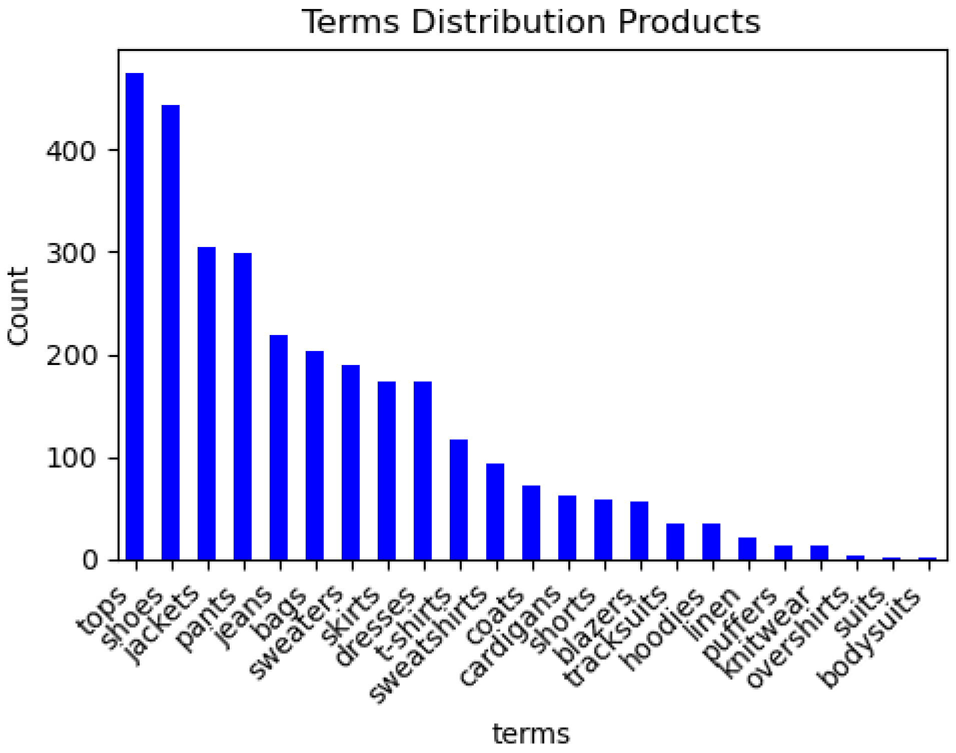
In [22]:

```
# This plot indicates that there are more products available in the women's section compared to the men's sect
```

In [23]:

```
# Assuming 'section' column contains the section information
term_distribution = df['terms'].value_counts()

# Plotting the distribution
plt.figure(figsize=(5, 4))
term_distribution.plot(kind='bar', color='blue')
plt.title('Terms Distribution Products')
plt.xlabel('terms')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



In [24]:

```
# The plot illustrates a notable level of demand for terms associated with both tops and shoes.
```

Feature Engineering

In [25]:

```
# convert numerical values to string for count vectorization
```

In [26]:

```
df['price'] = df['price'].astype(str)
```

In [27]:

```
# creating new column considered for recomendation
# we are recomending "SKU" on the basis of ("name","description","price","terms","section")
```

In [28]:

```
df["features"] = df["name"]+df["description"]+df["price"]+df["terms"]+df["section"]
```

In [29]:
```
df.head()
```

Out[29]:

	sku	name	description	price	terms	section	features
0	272145190-250-2	BASIC PUFFER JACKET	Puffer jacket made of tear-resistant ripstop f...	19.99	jackets	MAN	BASIC PUFFER JACKETPuffer jacket made of tear-...
1	324052738-800-46	TUXEDO JACKET	Straight fit blazer. Pointed lapel collar and ...	169.0	jackets	MAN	TUXEDO JACKETStraight fit blazer. Pointed lape...
2	335342680-800-44	SLIM FIT SUIT JACKET	Slim fit jacket. Notched lapel collar. Long sl...	129.0	jackets	MAN	SLIM FIT SUIT JACKETSlim fit jacket. Notched l...
3	328303236-420-44	STRETCH SUIT JACKET	Slim fit jacket made of viscose blend fabric. ...	129.0	jackets	MAN	STRETCH SUIT JACKETSlim fit jacket made of vis...
4	312368260-800-2	DOUBLE FACED JACKET	Jacket made of faux leather faux shearling wit...	139.0	jackets	MAN	DOUBLE FACED JACKETJacket made of faux leather...

Feature Extraction

In [30]:
```
cv = CountVectorizer(max_features=5000,stop_words="english")
```

In [31]:
```
cv.fit_transform(df['features']).toarray().shape
```

Out[31]:
```
(3059, 2664)
```

In [32]:
```
vectors = cv.fit_transform(df['features']).toarray()
```

```
In [33]: vectors[0]
```

```
Out[33]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

```
In [34]: len(cv.get_feature_names_out())
```

```
Out[34]: 2664
```

Apply cosine similarity

```
In [35]: cosine_similarity(vectors).shape
```

```
Out[35]: (3059, 3059)
```

```
In [36]: similarity = cosine_similarity(vectors)                                    # so we dont have to write above ful
```

```
In [37]: similarity[0]
```

```
Out[37]: array([1.        , 0.32102894, 0.28109135, ..., 0.        , 0.        ,
                0.14547859])
```

```
In [38]: similarity[0].shape
```

```
Out[38]: (3059,)
```

In [39]:

```
sorted(list(enumerate(similarity[0])), reverse=True, key=lambda x:x[1])[1:6]
```

Out[39]:

```
[(55, 0.618115099963687),
 (9, 0.6180700462007376),
 (101, 0.5819143739626462),
 (6, 0.578351744823806),
 (197, 0.5619514869490164)]
```

Function For Recommendation System

In [40]:

```
def recommend(input1):
    # Ensure input1 is converted to string
    input_sku = str(input1)
    try:
        # Get index of input SKU
        input_index = df[df['sku'] == input_sku].index[0]
        # Retrieve similarity scores for the input SKU
        distances = similarity[input_index]
        # Sort indices based on similarity scores in descending order, excluding the input index
        similar_indices = sorted(range(len(distances)), key=lambda i: distances[i], reverse=True)[1:4]
        # Print recommended SKUs
        print("Recommended Products:")
        for i in similar_indices:
            print(df.iloc[i]['sku'],df.iloc[i])
    except IndexError:
        print("SKU not found!")
```

In [41]:

```
pd.set_option('display.max_column',None)                    # to see all values
pd.set_option('display.max_rows',None)
```

In [42]:

```
df
```

Out[42]:

	sku	name	description	price	terms	section	features
0	272145190-250-2	BASIC PUFFER JACKET	Puffer jacket made of tear-resistant ripstop f...	19.99	jackets	MAN	BASIC PUFFER JACKETPuffer jacket made of tear-...
1	324052738-800-46	TUXEDO JACKET	Straight fit blazer. Pointed lapel collar and ...	169.0	jackets	MAN	TUXEDO JACKETStraight fit blazer. Pointed lape...
2	335342680-800-44	SLIM FIT SUIT JACKET	Slim fit jacket. Notched lapel collar. Long sl...	129.0	jackets	MAN	SLIM FIT SUIT JACKETSlim fit jacket. Notched l...
3	328303236-420-44	STRETCH SUIT JACKET	Slim fit jacket made of viscose blend fabric. ...	129.0	jackets	MAN	STRETCH SUIT JACKETSlim fit jacket made of vis...
4	312368260-800-2	DOUBLE FACED JACKET	Jacket made of faux leather faux shearling wit...	139.0	jackets	MAN	DOUBLE FACED JACKETJacket made of faux leather...
5	320298385-807-2	CONTRASTING COLLAR JACKET	Relaxed fit jacket. Contrasting lapel collar a...	79.9	jackets	MAN	CONTRASTING COLLAR JACKETRelaxed fit jacket. C...

In [43]:

```
recommend("272145190-250-2")
```

```
Recommended Products:
267133943-711-2 sku                                              267133943-711-2
name                                   LIGHTWEIGHT PUFFER JACKET
description    Padded jacket made of technical fabric. High c...
price                                                      19.99
terms                                                    jackets
section                                                      MAN
features       LIGHTWEIGHT PUFFER JACKETPadded jacket made of...
Name: 55, dtype: object
312372602-800-2 sku                                              312372602-800-2
name                             100% FEATHER FILL PUFFER JACKET
description    Puffer jacket made of shiny finish technical f...
price                                                      169.0
terms                                                    jackets
section                                                      MAN
features       100% FEATHER FILL PUFFER JACKETPuffer jacket m...
Name: 9, dtype: object
267186163-643-2 sku                                              267186163-643-2
name                                     HOODED TECHNICAL JACKET
description    Jacket made of technical fabric with brushed i...
price                                                      19.99
terms                                                    jackets
section                                                      MAN
features       HOODED TECHNICAL JACKETJacket made of technica...
Name: 103, dtype: object
```

In [44]:

```
recommend("311292672-800-2")
```

```
Recommended Products:
311282759-806-2 sku                                              311282759-806-2
name                                    FAUX SUEDE BOMBER JACKET
description    Jacket made of faux suede fabric. Rib elastic ...
price                                                       69.9
terms                                                    jackets
section                                                      MAN
features       FAUX SUEDE BOMBER JACKETJacket made of faux su...
Name: 37, dtype: object
311309526-800-2 sku                                              311309526-800-2
name                                  FAUX LEATHER BOMBER JACKET
description    Jacket made of faux leather fabric. High colla...
price                                                       69.9
terms                                                    jackets
section                                                      MAN
features       FAUX LEATHER BOMBER JACKETJacket made of faux ...
Name: 25, dtype: object
311292541-802-2 sku                                              311292541-802-2
name                                        FLEECE BOMBER JACKET
description    Jacket made of faux shearling fabric. Rib elas...
price                                                      109.0
terms                                                    jackets
section                                                      MAN
features       FLEECE BOMBER JACKETJacket made of faux shearl...
Name: 66, dtype: object
```

In [45]:

```
recommend("324052738-800-46")
```

```
Recommended Products:
328594167-800-46 sku                                             328594167-800-46
name                                        STRAIGHT SUIT JACKET
description    Straight fit blazer. Notched lapel collar and ...
price                                                      129.0
terms                                                    jackets
section                                                      MAN
features       STRAIGHT SUIT JACKETStraight fit blazer. Notch...
Name: 53, dtype: object
339688935-711-46 sku                                             339688935-711-46
name                                 VISCOSE - LINEN SUIT JACKET
description    Straight fit jacket made of viscose and linen....
price                                                      169.0
terms                                                    jackets
section                                                      MAN
features       VISCOSE - LINEN SUIT JACKETStraight fit jacket...
Name: 184, dtype: object
329706743-401-46 sku                                             329706743-401-46
name                                     HOUNDSTOOTH SUIT JACKET
description    Straight fit blazer. Notched lapel collar and ...
price                                                      139.0
terms                                                    jackets
section                                                      MAN
features       HOUNDSTOOTH SUIT JACKETStraight fit blazer. No...
Name: 104, dtype: object
```

In [46]:

```
def recommend_quick(input1):
    df_index = df[df['sku']==input1].index[0]
    distances = similarity[df_index]
    output_list = sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:6]

    for i in output_list:
        print(df.iloc[i[0]].sku)
```

In [47]:

```
recommend_quick("272145190-250-2")
```

```
267133943-711-2
312372602-800-2
267186163-643-2
278112470-800-2
312372092-800-2
```