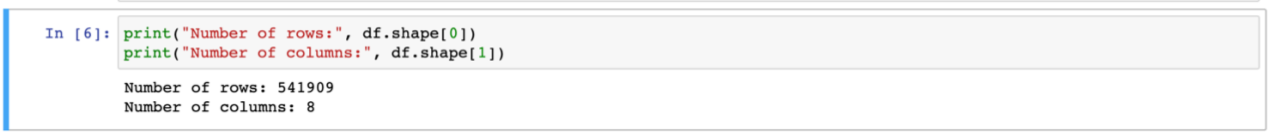
STA6714-DATA PREPARATION

ASSIGNMENT 7 Final: Online Retail

Dataset-

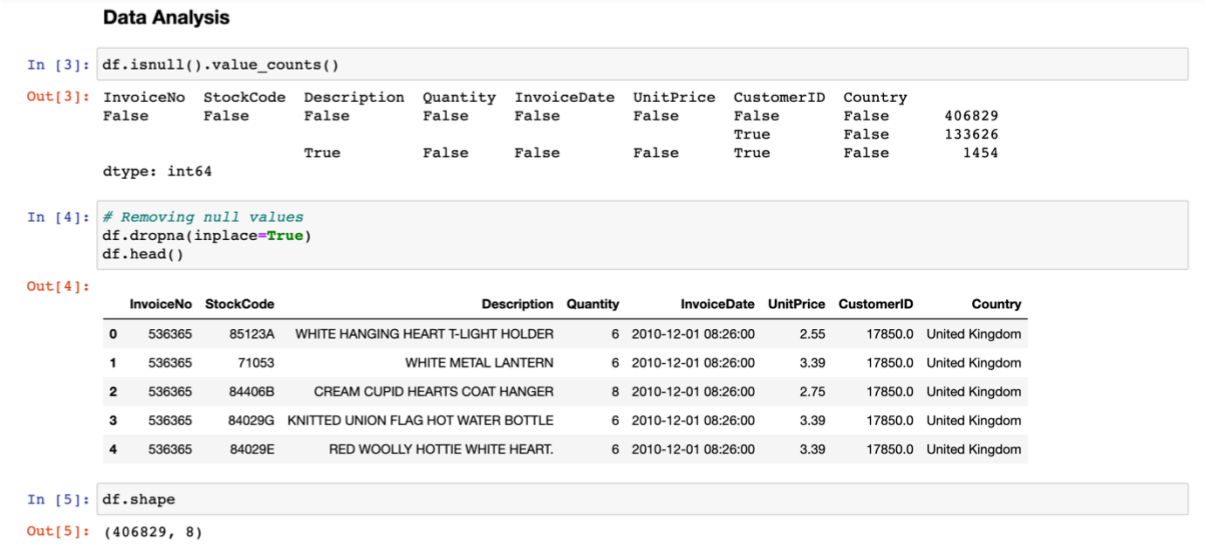
Online Retail.xlsx from UCML repository



```
In [6]: print("Number of rows:", df.shape[0])
        print("Number of columns:", df.shape[1])

        Number of rows: 541909
        Number of columns: 8
```

Exploratory Data Analysis-

Removing Null Values-



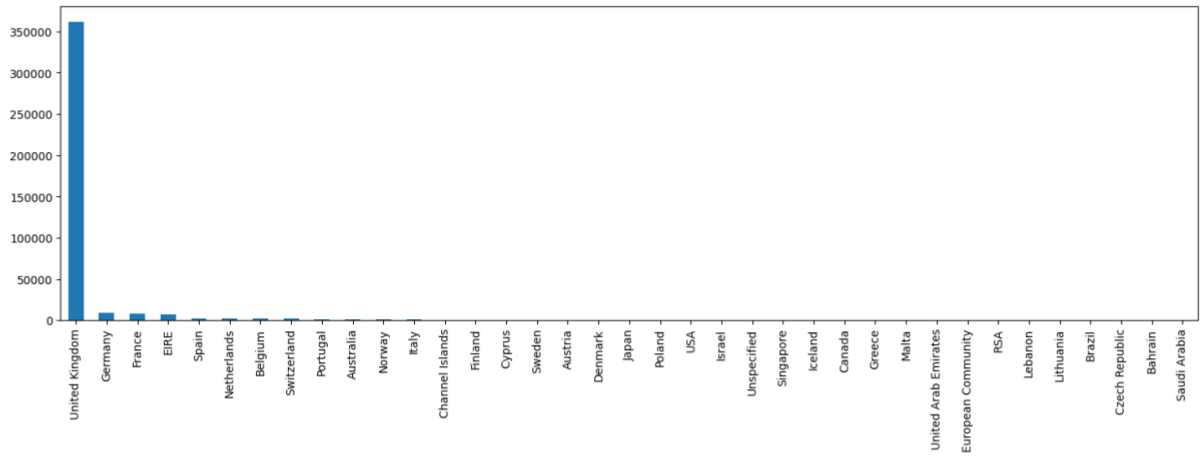
Data Analysis

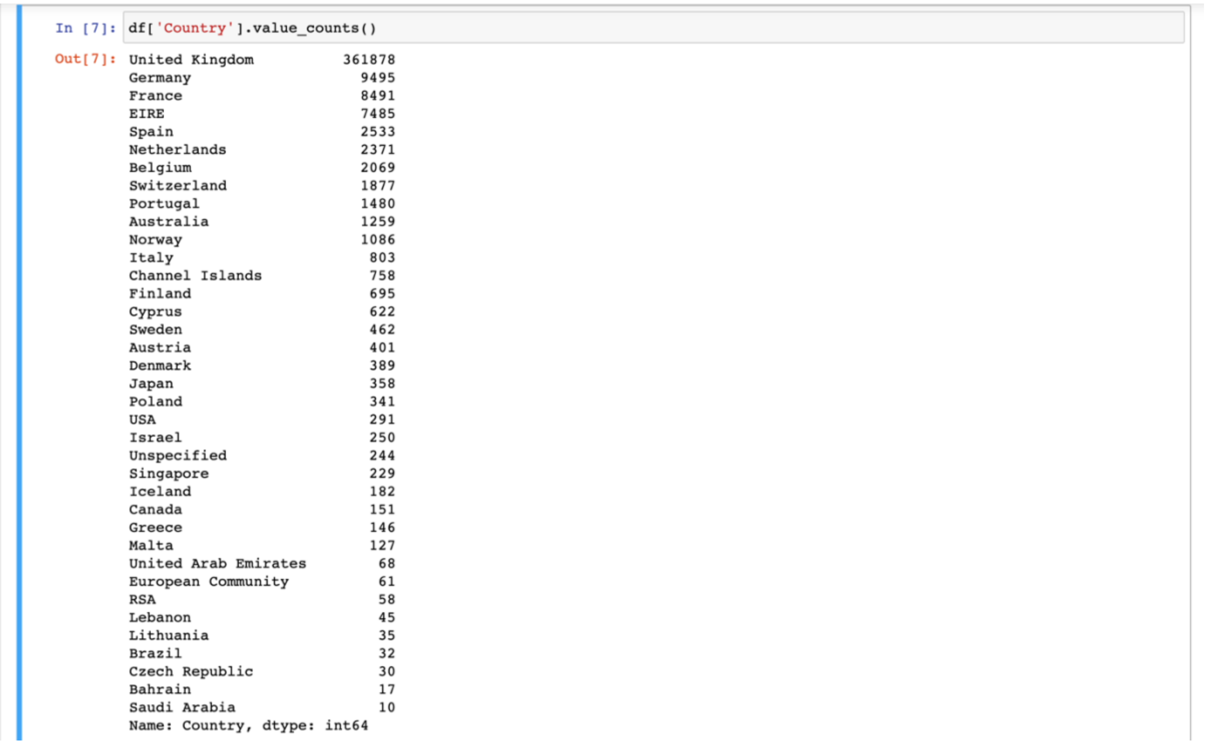
```
In [3]: df.isnull().value_counts()

Out[3]: InvoiceNo  StockCode  Description  Quantity  InvoiceDate  UnitPrice  CustomerID  Country
        False      False      False        False     False        False      False       False      406829
                                                                             True        False      133626
                              True         False     False        False      True        False        1454
        dtype: int64

In [4]: # Removing null values
        df.dropna(inplace=True)
        df.head()

Out[4]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

```
In [5]: df.shape

Out[5]: (406829, 8)
```

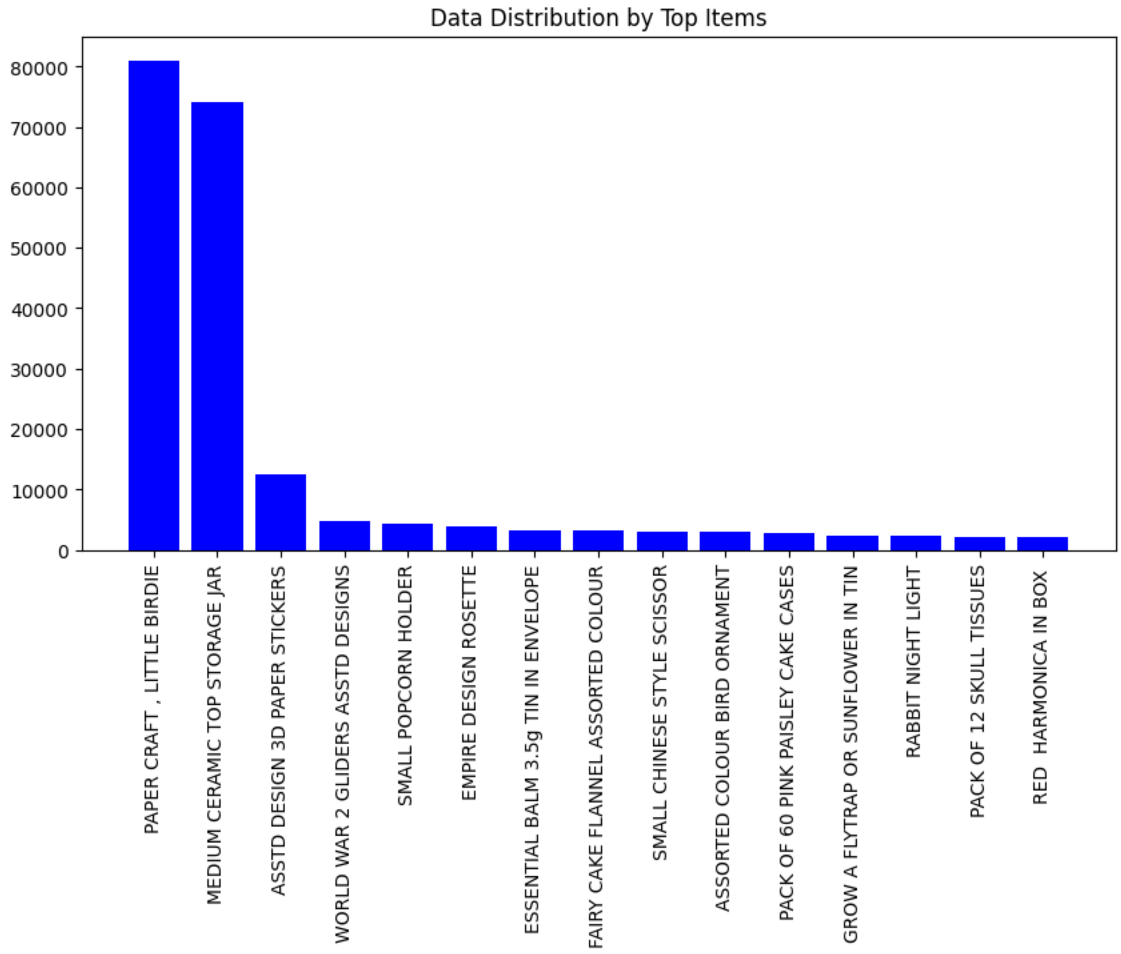
Data Distribution by Country-





```
In [7]: df['Country'].value_counts()

Out[7]: United Kingdom          361878
        Germany                   9495
        France                    8491
        EIRE                      7485
        Spain                     2533
        Netherlands               2371
        Belgium                   2069
        Switzerland               1877
        Portugal                  1480
        Australia                 1259
        Norway                    1086
        Italy                      803
        Channel Islands            758
        Finland                    695
        Cyprus                     622
        Sweden                     462
        Austria                    401
        Denmark                    389
        Japan                      358
        Poland                     341
        USA                        291
        Israel                     250
        Unspecified                244
        Singapore                  229
        Iceland                    182
        Canada                     151
        Greece                     146
        Malta                      127
        United Arab Emirates        68
        European Community          61
        RSA                         58
        Lebanon                     45
        Lithuania                   35
        Brazil                      32
        Czech Republic              30
        Bahrain                     17
        Saudi Arabia                10
        Name: Country, dtype: int64
```

Data Distribution by Top Items-

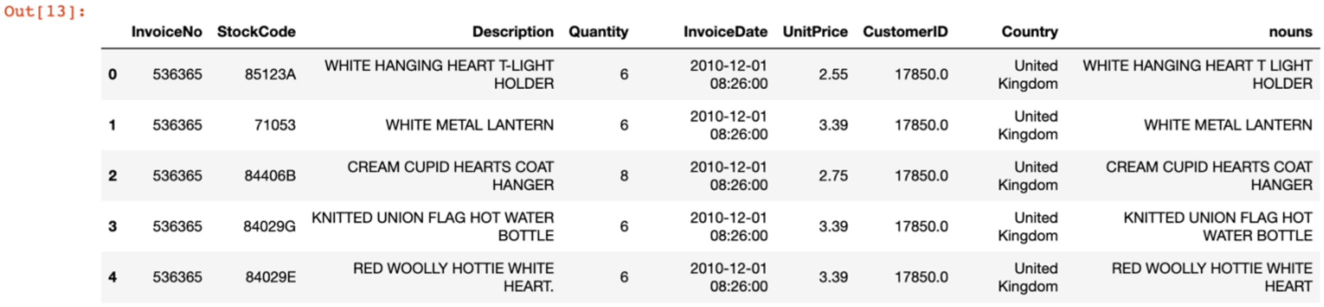


Generating Word Cloud for Descriptions-

UNION JACK
PACK
LUNCH BAG
FAIRY CAKE
BAG PINK
HAND WARMER
LUNCH BOX
ALARM CLOCK
PANTRY DESIGN
HOME SWEET
DRAWER KNOB
CLOCK BAKELIKE
RED RETROSPOT
METAL SIGN
CHAIN KIT
ASSORTED COLOUR
CAKE CASES
HOT WATER
PINK POLKADOT
BAG VINTAGE
WHITE HEART
VINTAGE DOILY
HEART DECORATION
JUMBO BAG
BAG RED
SET RED
DESIGN LUNCH
CAKE TIN
HEART T
KEEP CALM
LOVE LONDON
DOLLY GIRL
PAPER CHAIN
VINTAGE LEAF
T LIGHT
LIGHT HOLDER
COOKIE CUTTER
WHITE FINISH
SWEET HOME
CHARLOTTE BAG
HANGING HEART
WATER BOTTLE
VINTAGE CHRISTMAS

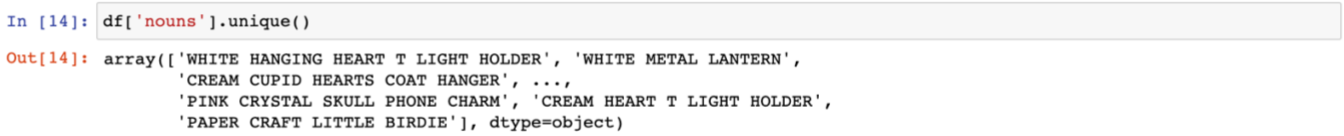
Data Pre-Processing-

- Used natural language processing (NLP) to extract nouns and proper nouns from each item description and then grouped similar items into categories based on the most common words in the item descriptions.
- Then found the most similar category to a customer's previous purchase and recommended other items from that category.
- Two methods are defined, namely:
    - *extract_nouns* which extracts the nouns and proper nouns from a given string of text using spaCy.
    - *create_category_name* which generates a category name based on the most common words in a list of item descriptions.
- Then extracted the unique values of the 'Description' column.
- Created a dictionary to map each unique value to its extracted nouns using extract_nouns, and then mapped the extracted nouns back to each row in the 'Description' column in the original DataFrame.



Out[13]:

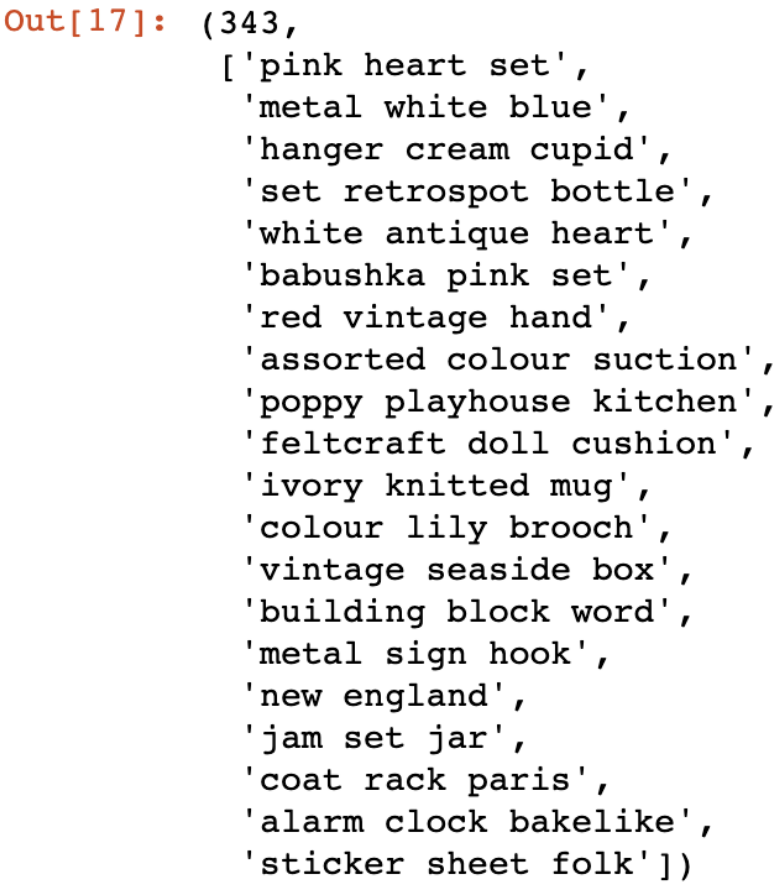
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	nouns
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	WHITE HANGING HEART T LIGHT HOLDER
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	WHITE METAL LANTERN
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	CREAM CUPID HEARTS COAT HANGER
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	KNITTED UNION FLAG HOT WATER BOTTLE
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	RED WOOLLY HOTTIE WHITE HEART



```
In [14]: df['nouns'].unique()

Out[14]: array(['WHITE HANGING HEART T LIGHT HOLDER', 'WHITE METAL LANTERN',
                'CREAM CUPID HEARTS COAT HANGER', ...,
                'PINK CRYSTAL SKULL PHONE CHARM', 'CREAM HEART T LIGHT HOLDER',
                'PAPER CRAFT LITTLE BIRDIE'], dtype=object)
```

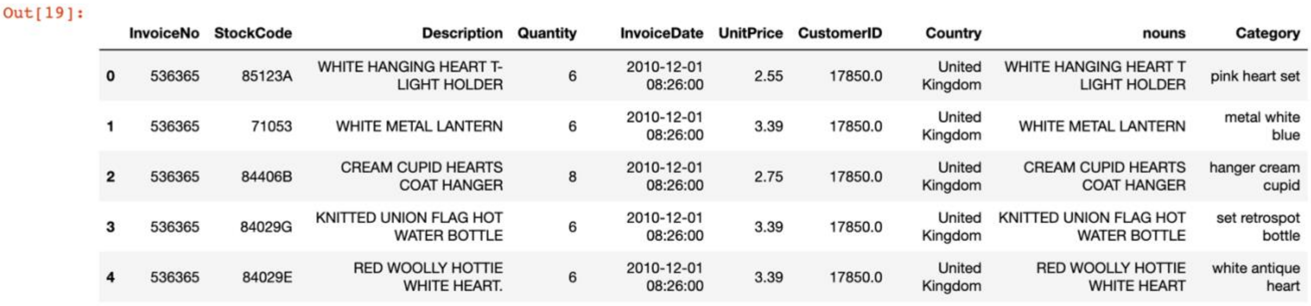
- Next, defined levenshtein_similarity which calculates the similarity between two strings using the Levenshtein distance.

- Then extracted the unique items from the dataset and assigns each item to a category based on its similarity to existing items using levenshtein_similarity. Created a dictionary to store the item categories and a dictionary to store the category names based on the most common words in the item descriptions. Then mapped the categories to the original DataFrame.



```
Out[17]: (343,
          ['pink heart set',
           'metal white blue',
           'hanger cream cupid',
           'set retrospot bottle',
           'white antique heart',
           'babushka pink set',
           'red vintage hand',
           'assorted colour suction',
           'poppy playhouse kitchen',
           'feltcraft doll cushion',
           'ivory knitted mug',
           'colour lily brooch',
           'vintage seaside box',
           'building block word',
           'metal sign hook',
           'new england',
           'jam set jar',
           'coat rack paris',
           'alarm clock bakelike',
           'sticker sheet folk'])
```

- Lastly, defined several helper functions to get customer invoices, invoice items, and items from the DataFrame based on a given column name and ID. Used these functions to find the most similar category to a customer's previous purchase, and then recommends five items from that category.

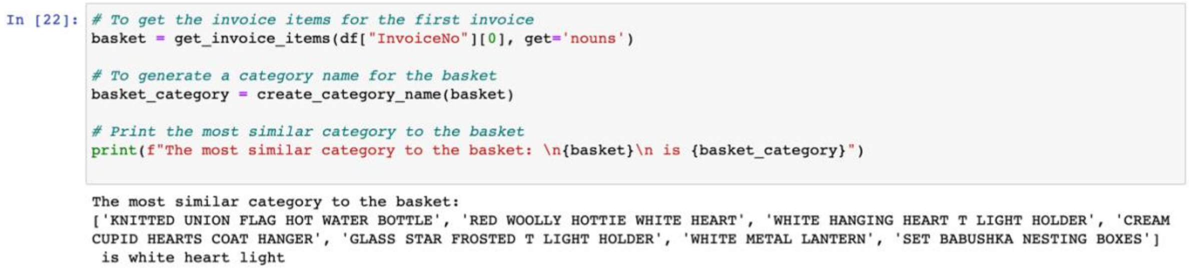


Out[19]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	nouns	Category
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	WHITE HANGING HEART T LIGHT HOLDER	pink heart set
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	WHITE METAL LANTERN	metal white blue
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	CREAM CUPID HEARTS COAT HANGER	hanger cream cupid
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	KNITTED UNION FLAG HOT WATER BOTTLE	set retrospot bottle
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	RED WOOLLY HOTTIE WHITE HEART	white antique heart

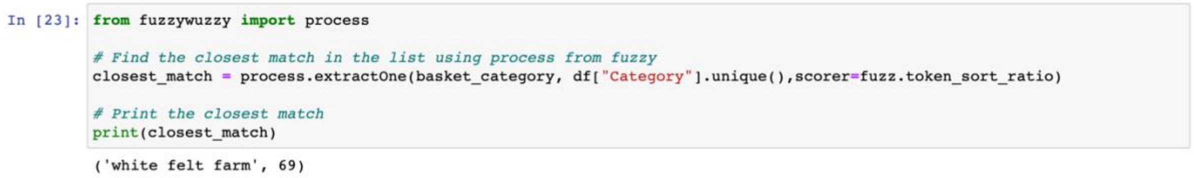
- Retrieves the items in a particular invoice, then creates a category name for that basket, and finds the most similar category from the available categories in the dataset using fuzzy string matching.
- Once the closest matching category is found, sorts the items in that category by quantity, removes the items already present in the basket, and recommends a random item from the top 3 items.

Most similar category to the basket-



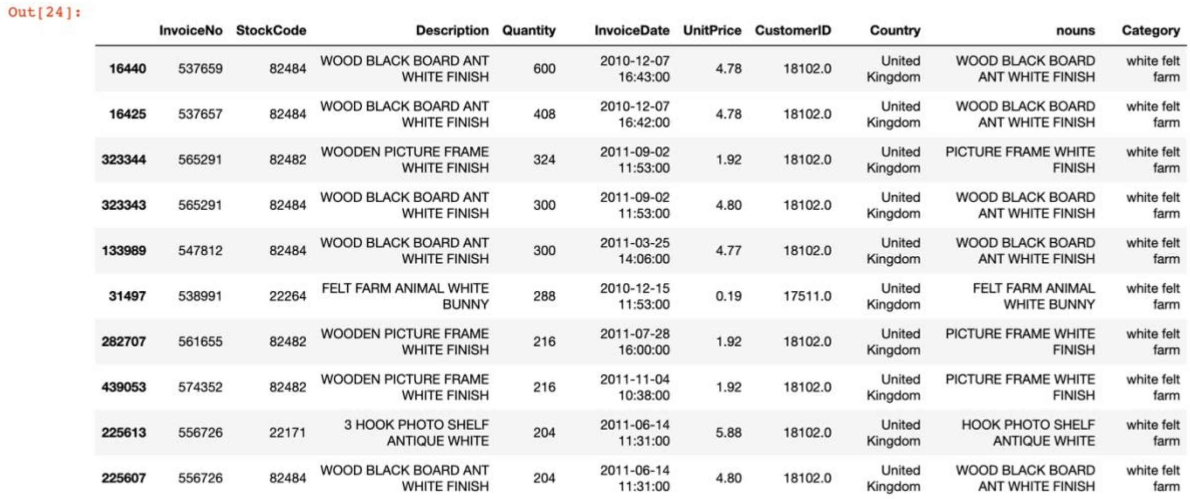
```
In [22]: # To get the invoice items for the first invoice
         basket = get_invoice_items(df["InvoiceNo"][0], get='nouns')

         # To generate a category name for the basket
         basket_category = create_category_name(basket)

         # Print the most similar category to the basket
         print(f"The most similar category to the basket: \n{basket}\n is {basket_category}")
```

```
The most similar category to the basket:
['KNITTED UNION FLAG HOT WATER BOTTLE', 'RED WOOLLY HOTTIE WHITE HEART', 'WHITE HANGING HEART T LIGHT HOLDER', 'CREAM
CUPID HEARTS COAT HANGER', 'GLASS STAR FROSTED T LIGHT HOLDER', 'WHITE METAL LANTERN', 'SET BABUSHKA NESTING BOXES']
 is white heart light
```



```
In [23]: from fuzzywuzzy import process

         # Find the closest match in the list using process from fuzzy
         closest_match = process.extractOne(basket_category, df["Category"].unique(),scorer=fuzz.token_sort_ratio)

         # Print the closest match
         print(closest_match)
```

```
('white felt farm', 69)
```



Out[24]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	nouns	Category
16440	537659	82484	WOOD BLACK BOARD ANT WHITE FINISH	600	2010-12-07 16:43:00	4.78	18102.0	United Kingdom	WOOD BLACK BOARD ANT WHITE FINISH	white felt farm
16425	537657	82484	WOOD BLACK BOARD ANT WHITE FINISH	408	2010-12-07 16:42:00	4.78	18102.0	United Kingdom	WOOD BLACK BOARD ANT WHITE FINISH	white felt farm
323344	565291	82482	WOODEN PICTURE FRAME WHITE FINISH	324	2011-09-02 11:53:00	1.92	18102.0	United Kingdom	PICTURE FRAME WHITE FINISH	white felt farm
323343	565291	82484	WOOD BLACK BOARD ANT WHITE FINISH	300	2011-09-02 11:53:00	4.80	18102.0	United Kingdom	WOOD BLACK BOARD ANT WHITE FINISH	white felt farm
133989	547812	82484	WOOD BLACK BOARD ANT WHITE FINISH	300	2011-03-25 14:06:00	4.77	18102.0	United Kingdom	WOOD BLACK BOARD ANT WHITE FINISH	white felt farm
31497	538991	22264	FELT FARM ANIMAL WHITE BUNNY	288	2010-12-15 11:53:00	0.19	17511.0	United Kingdom	FELT FARM ANIMAL WHITE BUNNY	white felt farm
282707	561655	82482	WOODEN PICTURE FRAME WHITE FINISH	216	2011-07-28 16:00:00	1.92	18102.0	United Kingdom	PICTURE FRAME WHITE FINISH	white felt farm
439053	574352	82482	WOODEN PICTURE FRAME WHITE FINISH	216	2011-11-04 10:38:00	1.92	18102.0	United Kingdom	PICTURE FRAME WHITE FINISH	white felt farm
225613	556726	22171	3 HOOK PHOTO SHELF ANTIQUE WHITE	204	2011-06-14 11:31:00	5.88	18102.0	United Kingdom	HOOK PHOTO SHELF ANTIQUE WHITE	white felt farm
225607	556726	82484	WOOD BLACK BOARD ANT WHITE FINISH	204	2011-06-14 11:31:00	4.80	18102.0	United Kingdom	WOOD BLACK BOARD ANT WHITE FINISH	white felt farm

```
array(['WOOD DRAWER CABINET WHITE FINISH',
       'WOOD BLACK BOARD ANT WHITE FINISH', 'FELT FARM ANIMAL SHEEP',
       'FELT FARM ANIMAL RABBIT', 'FELT FARM ANIMAL HEN'], dtype=object
)


Recommended item: GUMBALL MONOCHROME COAT RACK
```

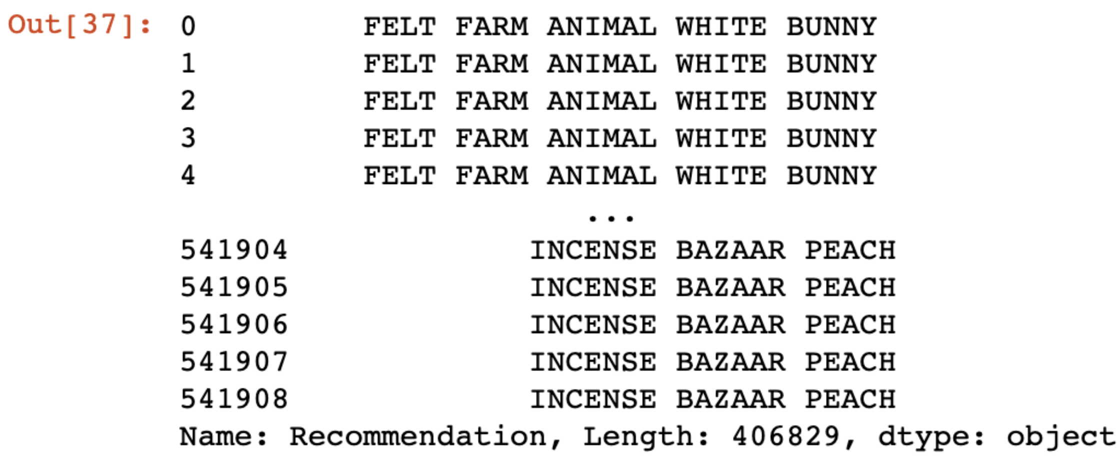
Modelling-

get_recommendation() method takes an invoice number as an input and returns a recommended item based on the items in that invoice.

get_basket_rec() that takes a basket as an input and returns a recommended item based on the items in that basket.

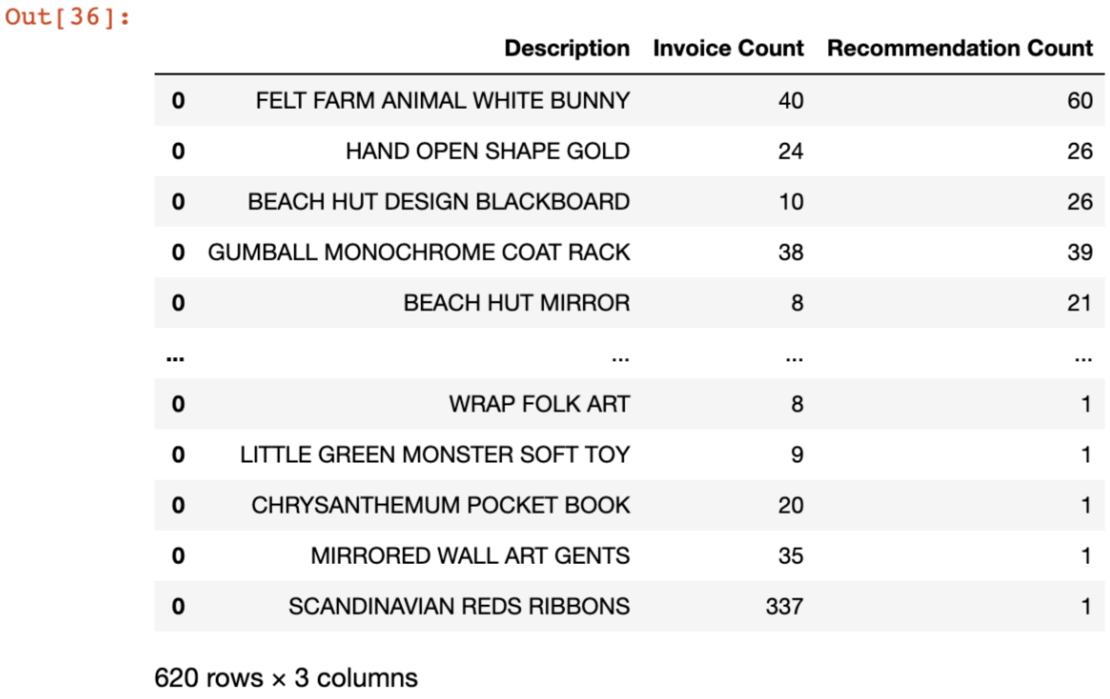
Both functions use the same method to find the recommended item: they bin the input into a category, find the most similar category, get the items of that category, remove the items already in the input, and randomly select one of the remaining items.

Recommendation Dataset-



```
Out[37]: 0          FELT FARM ANIMAL WHITE BUNNY
         1          FELT FARM ANIMAL WHITE BUNNY
         2          FELT FARM ANIMAL WHITE BUNNY
         3          FELT FARM ANIMAL WHITE BUNNY
         4          FELT FARM ANIMAL WHITE BUNNY
                                ...
         541904              INCENSE BAZAAR PEACH
         541905              INCENSE BAZAAR PEACH
         541906              INCENSE BAZAAR PEACH
         541907              INCENSE BAZAAR PEACH
         541908              INCENSE BAZAAR PEACH
         Name: Recommendation, Length: 406829, dtype: object
```

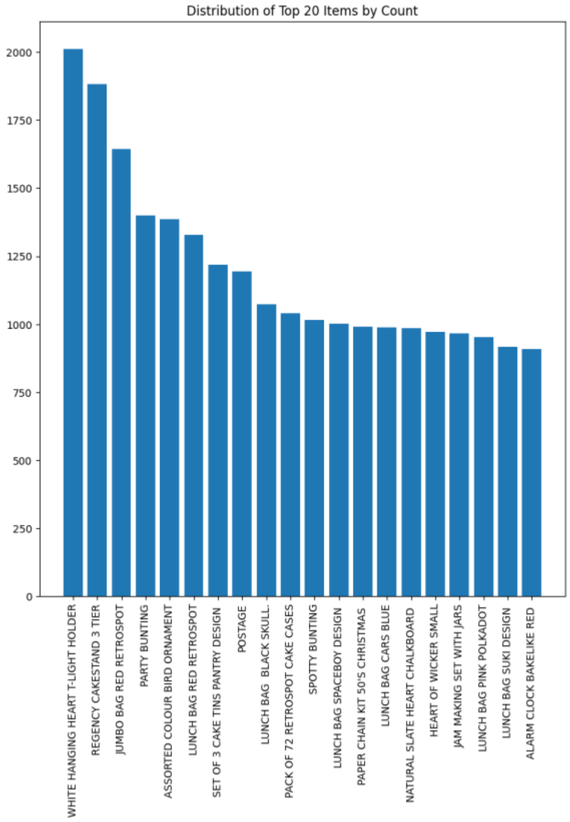
Summary Table

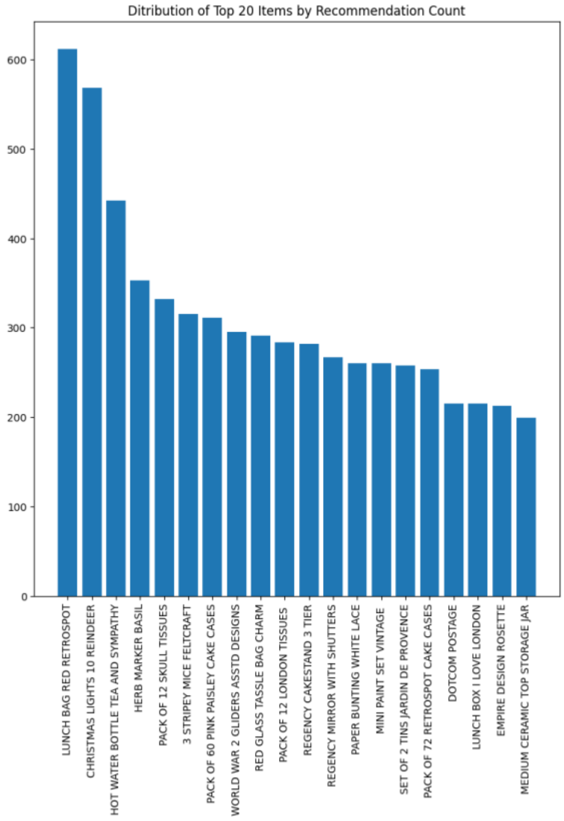


Out[36]:

	Description	Invoice Count	Recommendation Count
0	FELT FARM ANIMAL WHITE BUNNY	40	60
0	HAND OPEN SHAPE GOLD	24	26
0	BEACH HUT DESIGN BLACKBOARD	10	26
0	GUMBALL MONOCHROME COAT RACK	38	39
0	BEACH HUT MIRROR	8	21
...	...	...	...
0	WRAP FOLK ART	8	1
0	LITTLE GREEN MONSTER SOFT TOY	9	1
0	CHRYSANTHEMUM POCKET BOOK	20	1
0	MIRRORED WALL ART GENTS	35	1
0	SCANDINAVIAN REDS RIBBONS	337	1

620 rows × 3 columns


Distribution of Top 20 Items by Count
0
250
500
750
1000
1250
1500
1750
2000
WHITE HANGING HEART T-LIGHT HOLDER
REGENCY CAKESTAND 3 TIER
JUMBO BAG RED RETROSPOT
PARTY BUNTING
ASSORTED COLOUR BIRD ORNAMENT
LUNCH BAG RED RETROSPOT
SET OF 3 CAKE TINS PANTRY DESIGN
POSTAGE
LUNCH BAG BLACK SKULL.
PACK OF 72 RETROSPOT CAKE CASES
SPOTTY BUNTING
LUNCH BAG SPACEBOY DESIGN
PAPER CHAIN KIT 50'S CHRISTMAS
LUNCH BAG CARS BLUE
NATURAL SLATE HEART CHALKBOARD
HEART OF WICKER SMALL
JAM MAKING SET WITH JARS
LUNCH BAG PINK POLKADOT
LUNCH BAG SUKI DESIGN
ALARM CLOCK BAKELIKE RED


Ditribution of Top 20 Items by Recommendation Count
0
100
200
300
400
500
600
LUNCH BAG RED RETROSPOT
CHRISTMAS LIGHTS 10 REINDEER
HOT WATER BOTTLE TEA AND SYMPATHY
HERB MARKER BASIL
PACK OF 12 SKULL TISSUES
3 STRIPEY MICE FELTCRAFT
PACK OF 60 PINK PAISLEY CAKE CASES
WORLD WAR 2 GLIDERS ASSTD DESIGNS
RED GLASS TASSLE BAG CHARM
PACK OF 12 LONDON TISSUES
REGENCY CAKESTAND 3 TIER
REGENCY MIRROR WITH SHUTTERS
PAPER BUNTING WHITE LACE
MINI PAINT SET VINTAGE
SET OF 2 TINS JARDIN DE PROVENCE
PACK OF 72 RETROSPOT CAKE CASES
DOTCOM POSTAGE
LUNCH BOX I LOVE LONDON
EMPIRE DESIGN ROSETTE
MEDIUM CERAMIC TOP STORAGE JAR