Data Extraction

Following columns were extracted for WomenJean, MenCasualShirt and MenPoloTShirt verticals.

For Sales Data	For Returns Data
product_title	product_title
product_id	product_id
brand	brand
color	color
weight	weight
length	length
breadth	breadth
height	height
type	type
variant	variant
design	design
pattern	pattern
occasion	occasion
outer_material	outer_material
material	material
sleeve	sleeve
euro_size	euro_size
uk_india_size	uk_india_size
us_size	us_size
designed_for	designed_for
size_in_number	size_in_number
size	size
new_display_size	new_display_size
size_chart	size_chart
order_external_id	order_item_id
order_id	order_id
account_id	return_reason
order_item_date	return_sub_reason
	return_comments
	return_item_date

Directory structure

Puth both script/ and data/ directories with following directory structure in same directory.
Note:- Replace <vertical_name> with suitable vertical name in file.

```
scripts/
├── cleaning.py
├── method_1
│   ├── lib/
├── method_2
│   ├── clustering.py
│   ├── helper.py
├── method_3
│   ├── clustering.py
│   ├── helper.py
│   └── lib/
├── method_4
│   ├── clustering.py
│   └── lib/
└── persona_analysis.py


data/
├── <vertical_name>
│   ├── results/
│   │   ├── method_1/
│   │   ├── method_2/
│   │   │   ├── fastgreedy/
│   │   │   │   ├── returns/
│   │   │   │   ├── sales/
│   │   │   ├── infomap/
│   │   │   │   ├── returns/
│   │   │   │   ├── sales/
│   │   │   ├── walktrap/
│   │   │   │   ├── returns/
│   │   │   │   ├── sales/
│   │   ├── method_3/
│   │   │   ├── fastgreedy/
│   │   │   │   ├── returns/
│   │   │   │   ├── sales/
│   │   │   ├── infomap/
│   │   │   │   ├── returns/
│   │   │   │   ├── sales/
│   │   │   ├── walktrap/
│   │   │   │   ├── returns/
│   │   │   │   ├── sales/
│   │   ├── method_4/
│   │   │   ├── sales/
│   │   │   ├── returns/
├── returns/
│   ├── data/
```

```
│   │   ├── <vertical_name>_returns_data.csv
│   ├── pickles/
│   └── plots/
└── sales/
    ├── data/
    │   ├── <vertical_name>_sales_data.csv
    ├── pickles/
    └── plots/
```

Language and Dependencies

Python language was used for coding purposes. Following python packages needs to be installed for the programs to work.

- numpy
- pandas
- matplotlib
- seaborn
- networkx
- sklearn
- pickle
- igraph

File Usage

- **cleaning.py**
    - For cleaning the dataset and dividing it into training,validation and testing set. First run this file to obtain training, validation and test sets. Change the vertical names passed from main function accordingly.
- **persona_analysis.py**
    - For persona analysis of user to see if single user buys product from same vertical for other persons. Change the vertical names passed from main function accordingly.

Methods

- **Method-1**
    - Clustering (Brand,Size)s with Random Walk with Restart(RWR). *Note:- Change the vertical names passed from main function accordingly.*
- **Method-2**
    - Clustering (Brand,Size)s with standard community detection algorithms(Fastgreedy,Infomap and Walktrap). *Note:- Change the vertical names passed from main function accordingly.*

- **Method-3**
    - Clustering (Brand,Size)s with standard community detection algorithms(Fastgreedy,Infomap and Walktrap). Than compress those clusters with Random Walk with Restart(RWR). *Note:- Change the vertical names passed from main function accordingly.*
- **Method-4**
    - Converting (Brand,Size) nodes into vectors using node2vec and subsequently applying k-means algorithm to get clusters of (Brand,Size) nodes. *Note:- Change the vertical names passed from main function accordingly.*