Welcome

PREDATORY PRICING DATASET ANALYSIS AND HI-VALUE CUSTOMERS IDENTIFICATION

With pandas (python) – Pre-Requirements

Note: this file is available at: https://drive.google.com/drive/folders/1KJ7EvWslt-DbA7jfHoSHojUoPKbLtmUl?usp=sharing

Disclosure: Nothing in this lecture should be considered as investment or business advices. Past performance is not necessarily indicative of future returns. **Predatory Pricing Dataset and Customers Dataset during these lectures is only taken as the general example to show, how one can do data analysis using pandas (in python).** I AM NOT REPOSIBLE FOR YOUR ANY KIND OF LOSS/PROFIT IN/ON YOUR BUSINESS/STOCKS RETURNS. Consider a financial adviser before investing or invest at your own risk.

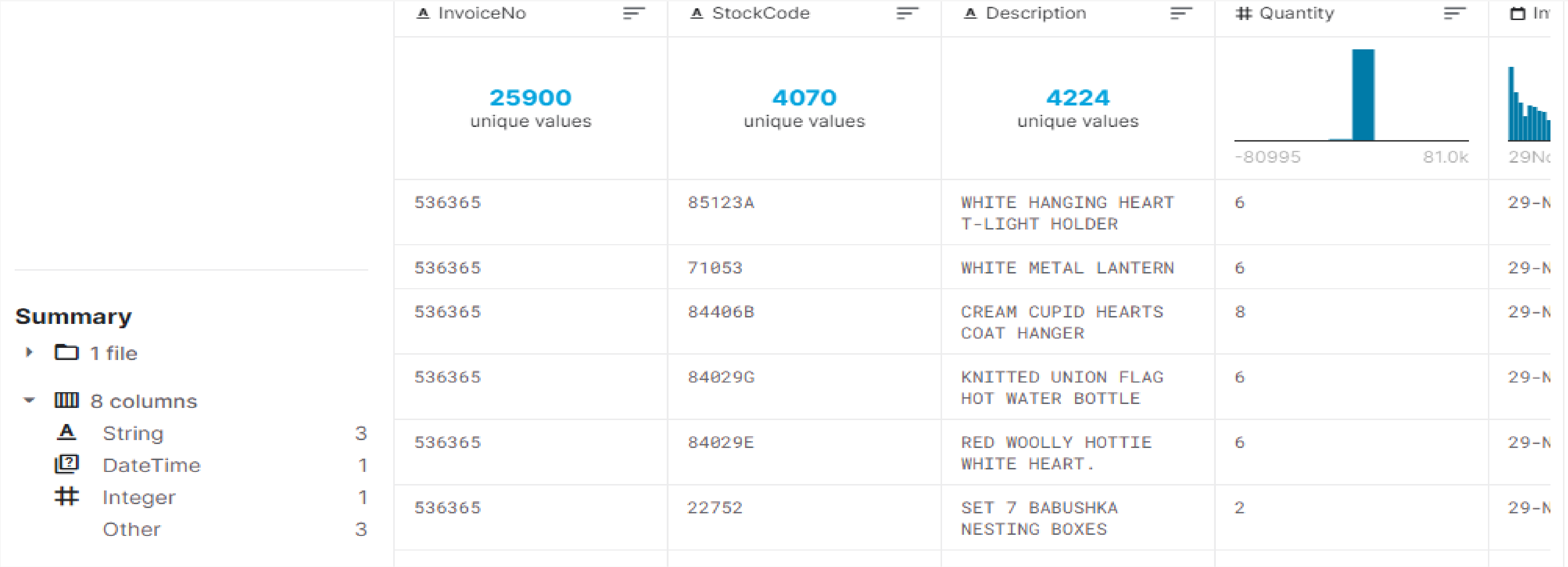
мξƏДмαĉĦʅŋΞ

About Dataset – UK-High value Customers Identification

Description: A UK-based online retail store has captured the sales data for different products for the period of one year (Nov 2016 to Dec 2017). The organization sells gifts primarily on the online platform. The customers who make a purchase consume directly for themselves. There are small businesses that buy in bulk and sell to other customers through the retail outlet channel.

Objective: Find significant customers for the business who make high purchases of their favorite products. The organization wants to roll out a loyalty program to the high-value customers after identification of segments. Use the clustering methodology to segment customers into groups.

About Dataset – UK-High value Customers Identification



Summary

- 1 file
- 8 columns

String	3
DateTime	1
Integer	1
Other	3

InvoiceNo	StockCode	Description	Quantity	In
25900 unique values	4070 unique values	4224 unique values	-80995 81.0k	29N
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	29-N
536365	71053	WHITE METAL LANTERN	6	29-N
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	29-N
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	29-N
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	29-N
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	29-N

Dataset source (kaggle): https://www.kaggle.com/vik2012kvs/high-value-customers-identification

Dataset download link, direct link: Ecommerce.csv

NOTE: Data is available under education license only. Don't use dataset other than educational purposes.

About Dataset – UK-High value Customers Identification (Conti.)

Number of rows: 541909 entries, 0 to 541908. With different number of null values in each column.

Data columns (total 8 columns): 'InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', and 'Country‘.

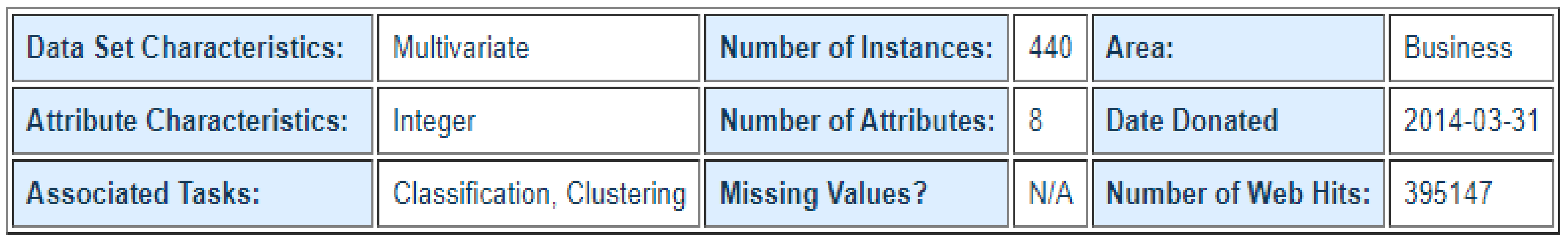
- Column 0: InvoiceNo (541909 non-null values) (datatype – object) – InvoiceNo Invoice number (A 6-digit integral number uniquely assigned to each transaction).
- Column 1: StockCode (541909 non-null values) (datatype – object) – Stock (Product/item) Code.
- Column 2: Description (540455 non-null values) (datatype – object) – Product (item) description name.
- Column 3: Quantity (541909 non-null values) (datatype – integer) – Quantity of each product (item) per transaction.
- Column 4: InvoiceDate (541909 non-null values) (datatypes – object) – The day when each transaction was generated.

About Dataset – UK-High value Customers Identification (Conti.)

- Column 6: UnitPrice (541909 non-null values) (datatype – float) – Unit price (Product price per unit).
- Column 7: CustomerID (406829 non-null values) (datatypes – float) – Country name (The name of the country where each customer resides)
- Column 8: Country (541909 non-null values) (datatype – object) – Customer number (Unique ID assigned to each customer).

More Dataset – Wholesale customers Dataset
(Not Mandatory)

Abstract: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) (i.e. the monetary unit principle, the assumption that money itself is treated as a unit of measurement, and that all transactions or economic events recorded in the accounts of a business can be expressed and measured in monetary terms by a currency) on diverse product categories.



Data Set Characteristics:	Multivariate	Number of Instances:	440	Area:	Business
Attribute Characteristics:	Integer	Number of Attributes:	8	Date Donated	2014-03-31
Associated Tasks:	Classification, Clustering	Missing Values?	N/A	Number of Web Hits:	395147

Dataset source (ics.uci): https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

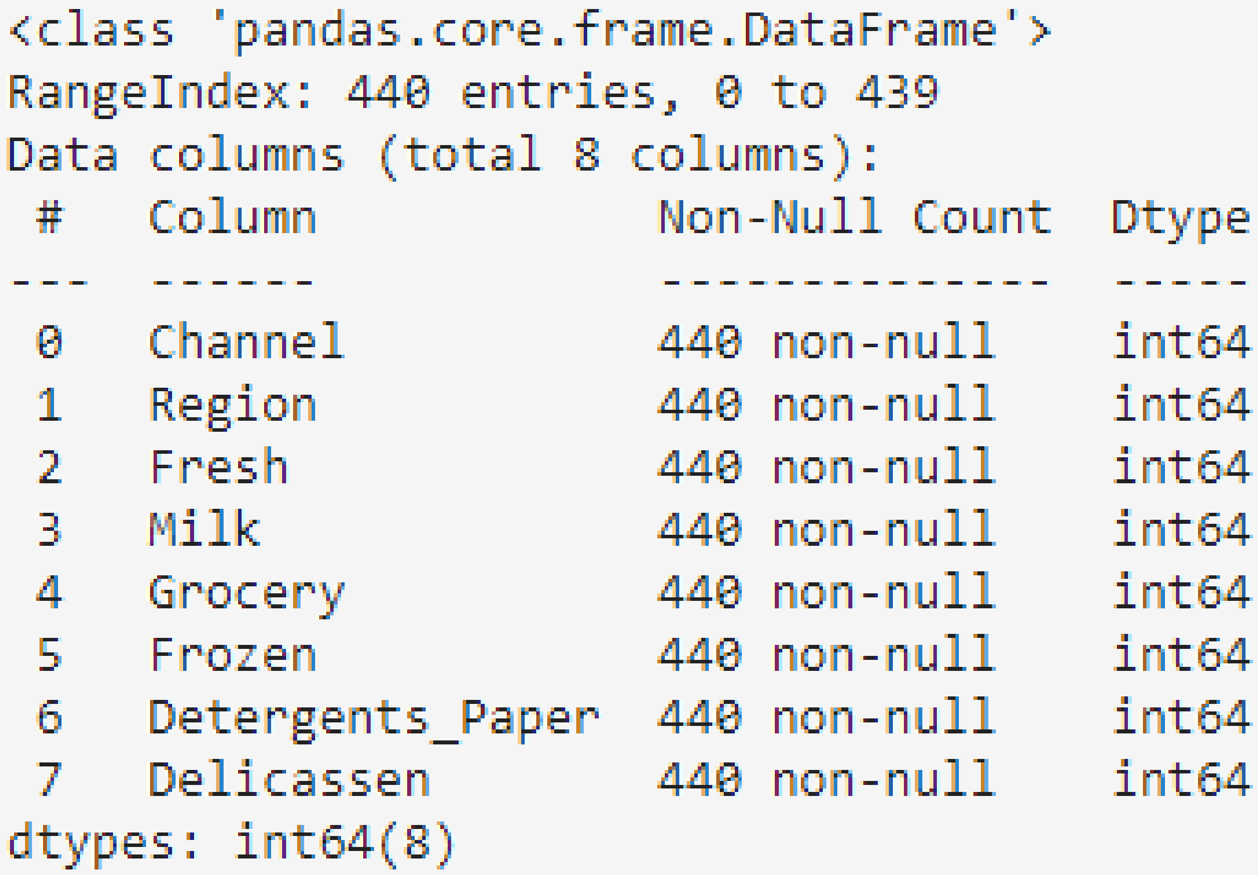
Dataset download link, direct link: Wholesal customers data.csv

Margarida G. M. S. Cardoso, margarida.cardoso@iscte.pt, ISCTE-IUL, Lisbon, Portugal.

More Dataset – Wholesale customers Dataset (Conti.)

(Not Mandatory)

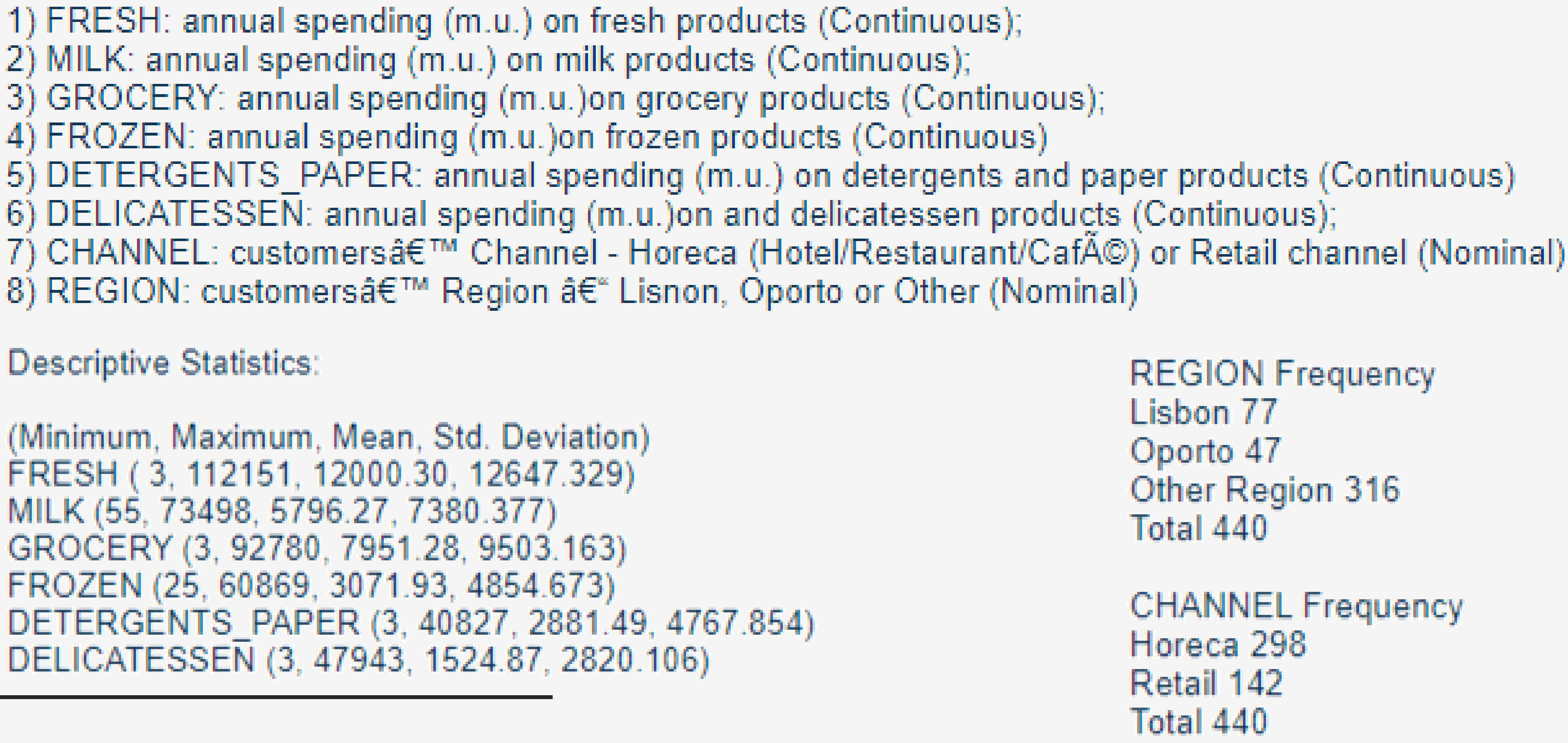
Total 440 rows, from 0 to 439. Total 8 columns (chammel, region, fresh, milk, grocery, frozen, degergents_paper, and delicassen) with no null values in each columns. All the values are of integer type.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    int64
 1   Region            440 non-null    int64
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicassen        440 non-null    int64
dtypes: int64(8)
```

More Dataset – Wholesale customers Dataset (Conti.)

(Not Mandatory; Attribute Information)



1) FRESH: annual spending (m.u.) on fresh products (Continuous);
2) MILK: annual spending (m.u.) on milk products (Continuous);
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)
5) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
7) CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)
8) REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

Descriptive Statistics:

(Minimum, Maximum, Mean, Std. Deviation)
FRESH (3, 112151, 12000.30, 12647.329)
MILK (55, 73498, 5796.27, 7380.377)
GROCERY (3, 92780, 7951.28, 9503.163)
FROZEN (25, 60869, 3071.93, 4854.673)
DETERGENTS_PAPER (3, 40827, 2881.49, 4767.854)
DELICATESSEN (3, 47943, 1524.87, 2820.106)

REGION	Frequency
Lisbon	77
Oporto	47
Other Region	316
Total	440

CHANNEL	Frequency
Horeca	298
Retail	142
Total	440

More Dataset – Amazon Customer Reviews Dataset

(Not Mandatory)

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon’s iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others. Accordingly, we are releasing this data to further research in multiple disciplines related to understanding customer product experiences. Specifically, this dataset was constructed to represent a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews.

Source: https://s3.amazonaws.com/amazon-reviews-pds/readme.html

More Dataset – Amazon Customer Reviews Dataset (Conti.)

(Not Mandatory)

Data format: Tab ('\t') separated text file (.tsv file), without quote or escape characters. First line in each file is header; 1 line corresponds to 1 record.

Rows – Different number of rows in different files.

Columns –

- marketplace - 2 letter country code of the marketplace where the review was written.
- customer_id - Random identifier that can be used to aggregate reviews written by a single author.
- review_id - The unique ID of the review.
- product_id - The unique Product ID the review pertains to. In the multilingual dataset the reviews for the same product in different countries can be grouped by the same product_id.

Source: https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

More Dataset – Amazon Customer Reviews Dataset (Conti.)

(Not Mandatory)

- product_parent - Random identifier that can be used to aggregate reviews for the same product.
- product_title - Title of the product.
- product_category - Broad product category that can be used to group reviews (also used to group the dataset into coherent parts).
- star_rating - The 1-5 star rating of the review.
- helpful_votes - Number of helpful votes.
- total_votes - Number of total votes the review received.
- vine - Review was written as part of the Vine program.
- verified_purchase - The review is on a verified purchase.
- review_headline - The title of the review.
- review_body - The review text.
- review_date - The date the review was written.

More Dataset – Amazon Customer Reviews Dataset (Conti.)

(Not Mandatory)

Sample Content:

- https://s3.amazonaws.com/amazon-reviews-pds/tsv/sample_us.tsv
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/sample_fr.tsv

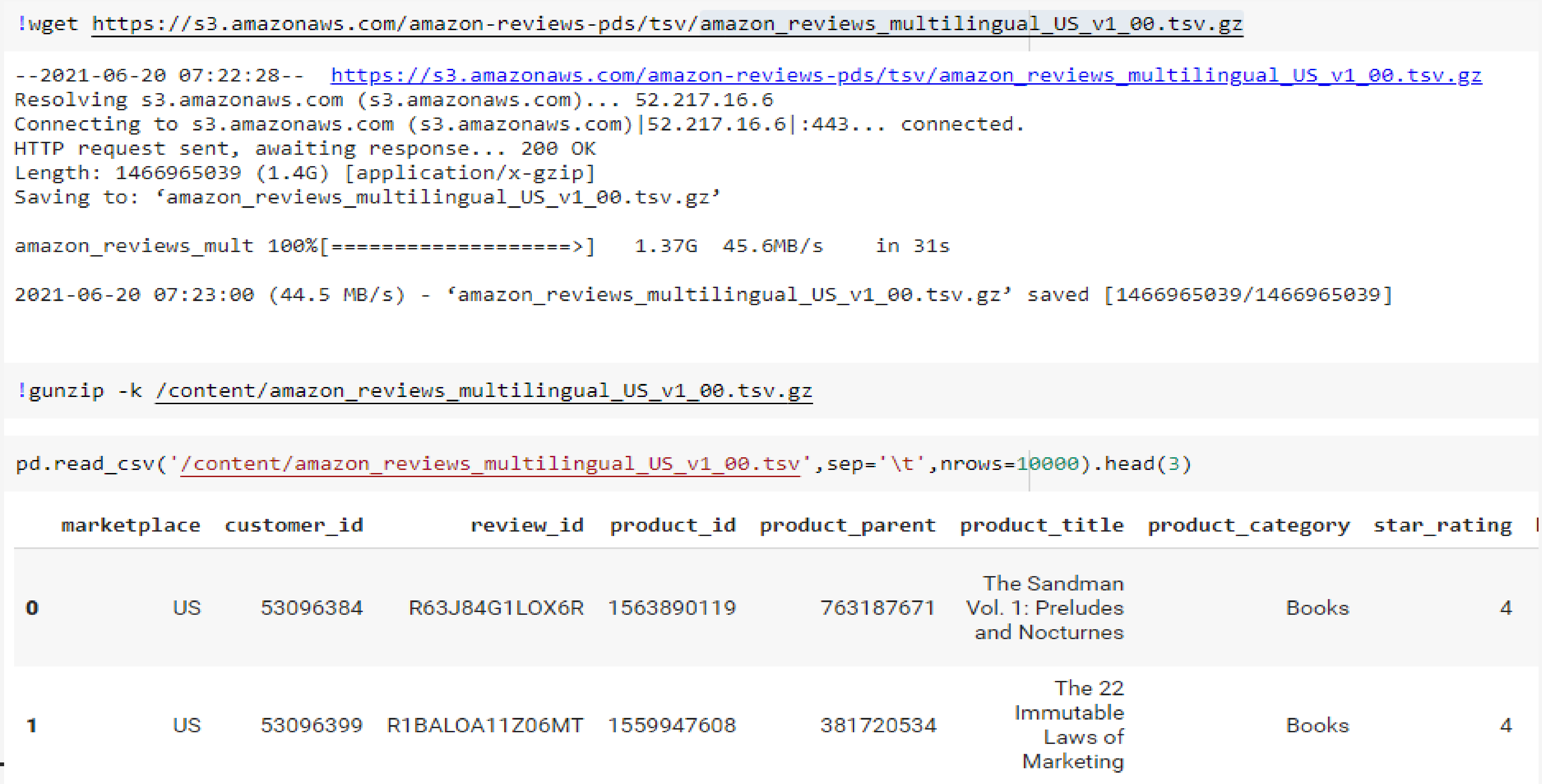
All data source –

- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_US_v1_00.tsv.gz
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_UK_v1_00.tsv.gz
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_JP_v1_00.tsv.gz
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_FR_v1_00.tsv.gz
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_DE_v1_00.tsv.gz

More links here: https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

More Dataset – Amazon Customer Reviews Dataset (Conti.)

(Not Mandatory, Get data in colab)



```
!wget https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_US_v1_00.tsv.gz
```

```
--2021-06-20 07:22:28--  https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_US_v1_00.tsv.gz
Resolving s3.amazonaws.com (s3.amazonaws.com)... 52.217.16.6
Connecting to s3.amazonaws.com (s3.amazonaws.com)|52.217.16.6|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1466965039 (1.4G) [application/x-gzip]
Saving to: ‘amazon_reviews_multilingual_US_v1_00.tsv.gz’

amazon_reviews_mult 100%[===================>]   1.37G  45.6MB/s    in 31s

2021-06-20 07:23:00 (44.5 MB/s) - ‘amazon_reviews_multilingual_US_v1_00.tsv.gz’ saved [1466965039/1466965039]
```

```
!gunzip -k /content/amazon_reviews_multilingual_US_v1_00.tsv.gz
```

```
pd.read_csv('/content/amazon_reviews_multilingual_US_v1_00.tsv',sep='\t',nrows=10000).head(3)
```

	marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating
0	US	53096384	R63J84G1LOX6R	1563890119	763187671	The Sandman Vol. 1: Preludes and Nocturnes	Books	4
1	US	53096399	R1BALOA11Z06MT	1559947608	381720534	The 22 Immutable Laws of Marketing	Books	4

More links here: https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

References

- Lecture drive link:
- Dataset source (data.world): https://data.world/data-hut/predatory-pricing-data-from-amazon
- https://www.kaggle.com/
- https://www.kaggle.com/vik2012kvs/high-value-customers-identification
- https://www.kaggle.com/vik2012kvs/high-value-customers-identification/download
- https://drive.google.com/file/d/1IxkyEQJBvnTf6SVl_C_eU9UEMBBMrQ22/view?usp=sharing
- https://archive.ics.uci.edu/ml/index.php
- https://archive.ics.uci.edu/ml/datasets/Wholesale+customers
- https://archive.ics.uci.edu/ml/machine-learning-databases/00292/
- https://archive.ics.uci.edu/ml/machine-learning-databases/00292/Wholesale%20customers%20data.csv
- https://s3.amazonaws.com/amazon-reviews-pds/readme.html
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

THANKS FOR UR PRECIOUS TIME! ☺

- Questions? 🗣 🗪

βy мξӘДмαĉĦĮŋΞ

Thank you