UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Engineering, Built Environment and Information Technology

# MIT 805

## Big Data

Assignment 1

u18023143

**Caleb Siyasiya**

By signing this assignment I confirm that I have read and are aware of the University of Pretoria's policy on academic dishonesty and plagiarism and I declare that the work submitted in this assignment is my own as delimited by the mentioned policies. I explicitly declare that no parts of this assignment have been copied from current or previous students' work or any other sources (including the internet), whether copyrighted or not. I understand that I will be subjected to disciplinary actions should it be found that the work I submit here does not comply with the said policies.

August 23, 2024

# Contents

# List of Figures



# List of Tables

# 1 Introduction

Online shopping has revolutionized the beauty and fashion industry by changing how businesses interact with customers. More and more businesses are turning to e-commerce platforms to increase their client engagement in order to better understand their behaviour and needs. Thus e-commerce platforms offer a treasure trove of information for analytics that can be used to better understand market place trends and create tailored client experiences thereby increasing platform engagement, sales and hopefully profits. In this assignment, analysis will be performed on data for a fashion company to convey how e-commerce data can be used to derive insights that can create business value.

# 2 Data Set

The Zara data set, which contains information about various products listed on the Zara website was used for this assignment. The data set was constructed using various web scrapping tools to scrap data from the Zara Website[1]. The data was obtained from Kaggle and is available under a MIT license.

## 2.1 Volume

The full size of the data set (uncompressed file) is 12.5GB[1].This is primarily made up of the 25 100 images of the different products within the data set.They are 3129 unique items within the data set describing each of the products which can be found in the CSV file( 5MB). The search queries used to construct the data set can be found in the JSON file[1].

## 2.2 Variety

The data set comprises of both structured and unstructured data in the form of images(.jpg), JSON and CSV files. Each product in the dataset is described using the following fields:

**Table 1:** Data fields

| **Field** | **Description** | **Data Type** |
|---|---|---|
| Brand | Brand of the product (Zara) | String Object |
| Name | Name of Product | String Object |
| Description | Description of product | String Object |
| Price | Price of product | Float |
| SKU | Stock Keeping Unit of product | String Object |
| Currency | Currency used to indicate price | String Object |
| URL | URL of product on Wbeiste | List of String Object |
| Images | List of Image urls for product | List of String Object |
| Scraped_at | Timestamp at which data scrapped from webesite | String Object |
| Terms | Term used to search for product on website | String Object |
| Error | empty column | String Object |
| Image_downloads | Name of downloaded images for product | List of String Object |

## 2.3 Value

Data Analysts can use this data set to understand fashion trends from the Zara brand which is one of the largest fashion brands in the world. This data set can allow analysts to identify which products and styles are most popular their by tailoring their own product lines and seasonal offerings to align with the trend from an industry leading brand. Such analysis could also help analyst's and researches understand the competitive marketing strategy behind Zara's product offering.

The prices indicated in the data set can also be used to analyze pricing trends of various clothing items which can inform organization's pricing strategy allowing them to be more competitive. Companies that sell Zara products on their own online platforms can also use this data set to train and build Recognition Systems that recommend Zara products based on a client's behaviour. Finally, this data set can be used to build new A.I tools such as Image Recognition algorithms applications for clothing items

## 2.4 Veracity

The data set is roughly 6 months old and thus it should be noted that the data set does not necessarily reflect the most up to date product offering and prices from Zara[1]. In addition to this, because web scrapping tools were used to collect the information it may not be fully representative of all available data for Zara products making it somewhat incomplete. Thus caution must be exercised in applying this data set for decision-making.

## 2.5 Velocity

The data set is static and thus does not fulfill the velocity component for Big Data.

# 3 Pre-Processing

The pre-processing of the data was performed using Python. The code used for the pre-processing can be accessed here. The GitHub code can be accessed here.

The pre-processing involved the following steps:

- Dropping columns not used in the data analysis(e.g. url columns)
- Replacing null values
- exploratory Data Analysis of the data set
- Visualizing images within the dataset

## 3.1 Exploratory Data Analysis

Figure 1 shows the distribution of prices for the different items. As can be seen from the figure, over 75% of items cost less than $200.00 with the average item costing $64.10.

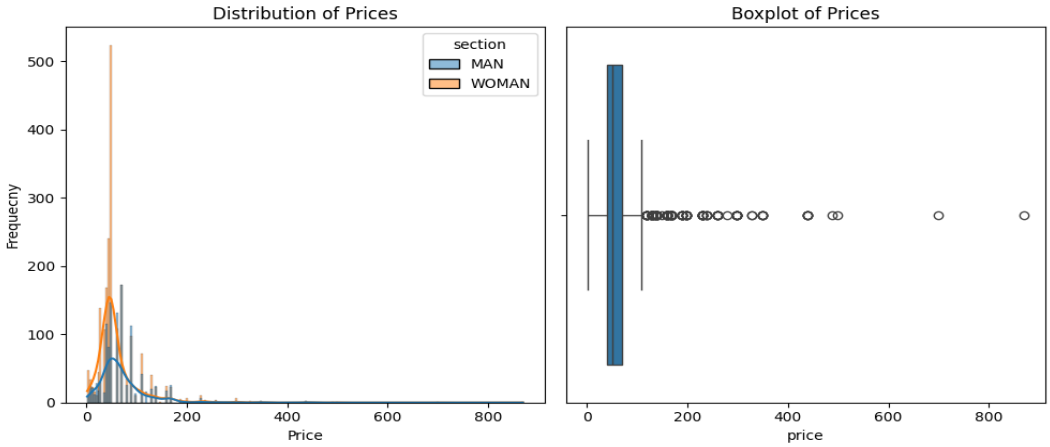


**Figure 1:** Distribution of Prices

Looking at the split of products, as shown in Figure 2, it can be seen that roughly two-thirds of the items are women's items.

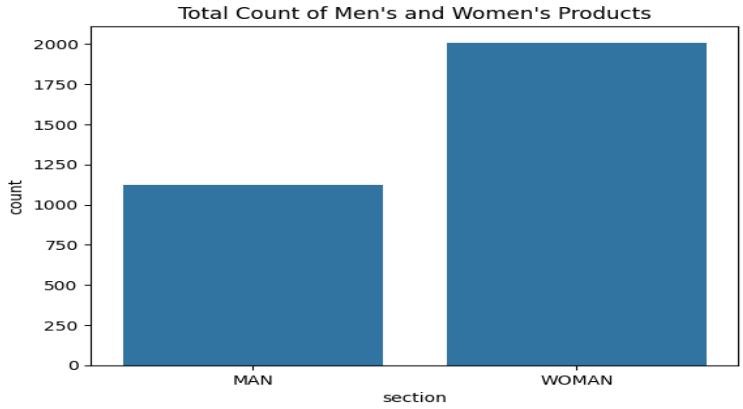


**Figure 2:** Total count of men's and women's items

Figure 3 shows the count of the various men's and women's products.

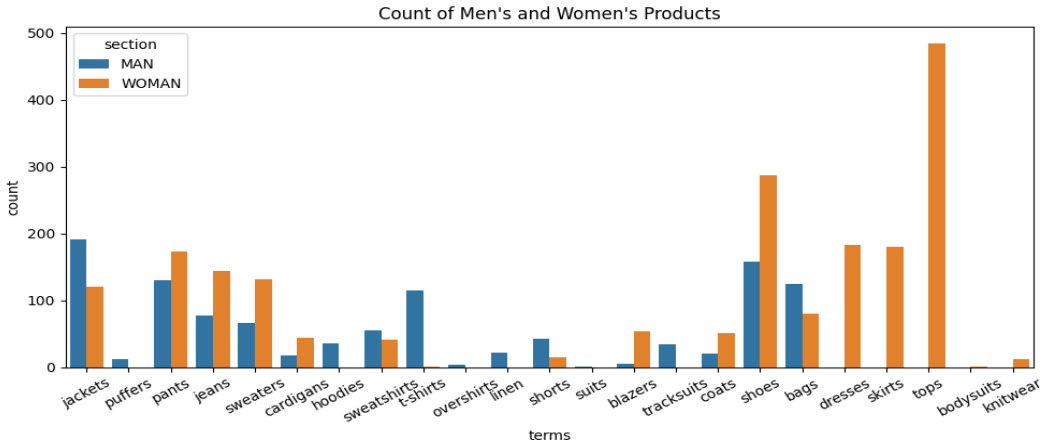


**Figure 3:** Count of Product various product types

# 4 Predicted Relationships

As can be seen from Figure 3, women's tops, shoes, dresses and skirts are the most commonly occurring product on the Zara website followed by men's jackets, pants and shoes. Thus, it is predicted that these items will account for majority of the total expected revenue which was calculated as $200 571.34.

It is also expected that a relationship exists between the number of images provided for a product on the website and the product's price, with more expensive items having more pictures associated with them to make them more desirable, increasing the chances of selling these items.

# References

[1] Mario Parreno Lara, "Zara products," 2024, accessed: August 20, 2024. [Online]. Available: https://www.kaggle.com/datasets/maparla/zara-products

This Overleaf document can be viewed here