Phase 4: Data exploration

Team name: **THE MINERS**
Team members:

1) Rishi Krishna Thodupunuri – rthod1@unh.newhaven.edu
2) Nikhil Teja Tangella – ntang2@unh.newhaven.edu
3) Sai Teja Gattu - sgatt4@unh.newhaven.edu

GIT link: https://github.com/rishi-krishna/THE_MINERS.git

---

2. Please introduce your selected data set and research question.

The dataset is called the "Zara US Fashion Products Dataset" and it contains information about products sold by the Zara US stores. Here are some details about the dataset:

- The dataset includes information on products sold by Zara US in the year 2017.
- Each product is represented by a unique ID(Product ID) number and includes information such as the product name, product size, Product Category, sizes, Colors, price, state age and Date of sale.
- The dataset also includes information about the category and subcategory of each product, as well as its color and composition.
- In addition to product information, the dataset also includes information about Zara US store locations state wise.
- This dataset is created and developed by our team using multiple resources on internet, which are mentioned in references.

Exploring Zara's Innovative Marketing and Supply Chain Strategies to Drive Sales. Identification of Deviations in sales and supply chain based on Customer reviews and sentiment analysis.

---

3. Please put a list of the exploration techniques, which you used in this work.

As part of Data exploration first we took raw data and pre-processed it by removing unnecessary data from dataset and we structured the data into following columns “Product ID, Product Category, Product Name, Product price, Sizes, Coors, State, AGE, Date”.

Below are data exploration techniques we used in this project:

Pre-processing: Converting the raw data into use full data, To create any data model we need to pre-process the un-useful data.

- ➔ Filtering: Selecting a subset of the data based on certain criteria, such as state or date range.
- ➔ Aggregation: Grouping data by one or more variables to get a summary of the data.
- ➔ Outlier detection: Identifying and analyzing values that are significantly different from the rest of the data.

Data cleaning: Removing unnecessary data and structuring the remaining data into columns.

- ➔ If there are any Null values and empty cells we removed those Null values using 'NOT NULL' Function.

Descriptive statistics: Analyzing the dataset to get a general sense of the data, such as mean, median, mode, standard deviation, etc.

- ➔ We used methods to find overall sales date wise:
- ➔ 'DATE TIME' function
- ➔ example- strptime(x,'%m/%d/%Y')) : converts the string value to exact date
- ➔ strftime('%b'), converts the numeric date(month) to string date(month).

Data visualization: Creating charts and graphs to visually represent the data, such as scatter plots, histograms, box plots, etc.

- ➔ We used functions like matplotlib-pyplot, scatter…etc, seaborn, to represent data visually.
- ➔ Bar graph, we used to represent state wise visualization
- ➔ Histographic, point plots …etc

Cross-tabulation: Analyzing the relationships between two or more variables by creating a contingency table.

- ➔ We used functions like crosstab ..etc

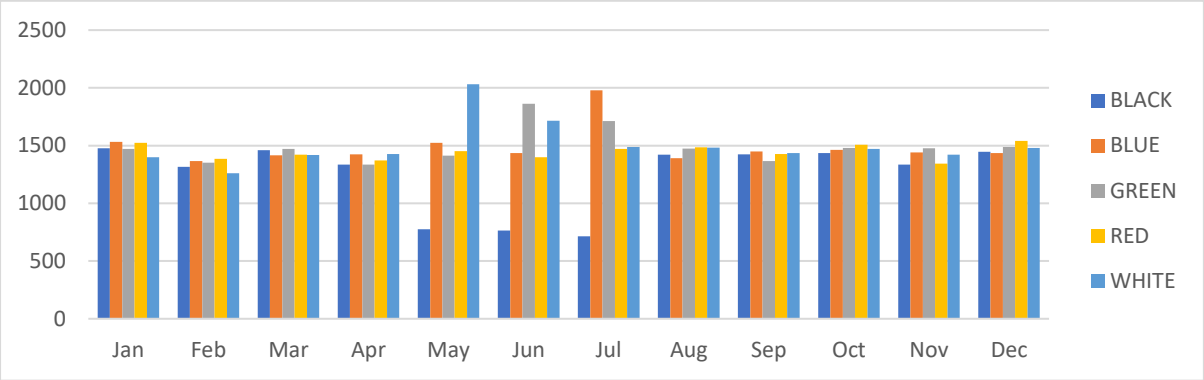
---

4. Please describe your data explorations from different perspectives using varied visualization techniques such as tables and charts. Finally, you should conclude your data exploration in a paragraph, which describes your findings based on the data exploration.

Below are few tables and charts that we generated using visualization techniques:

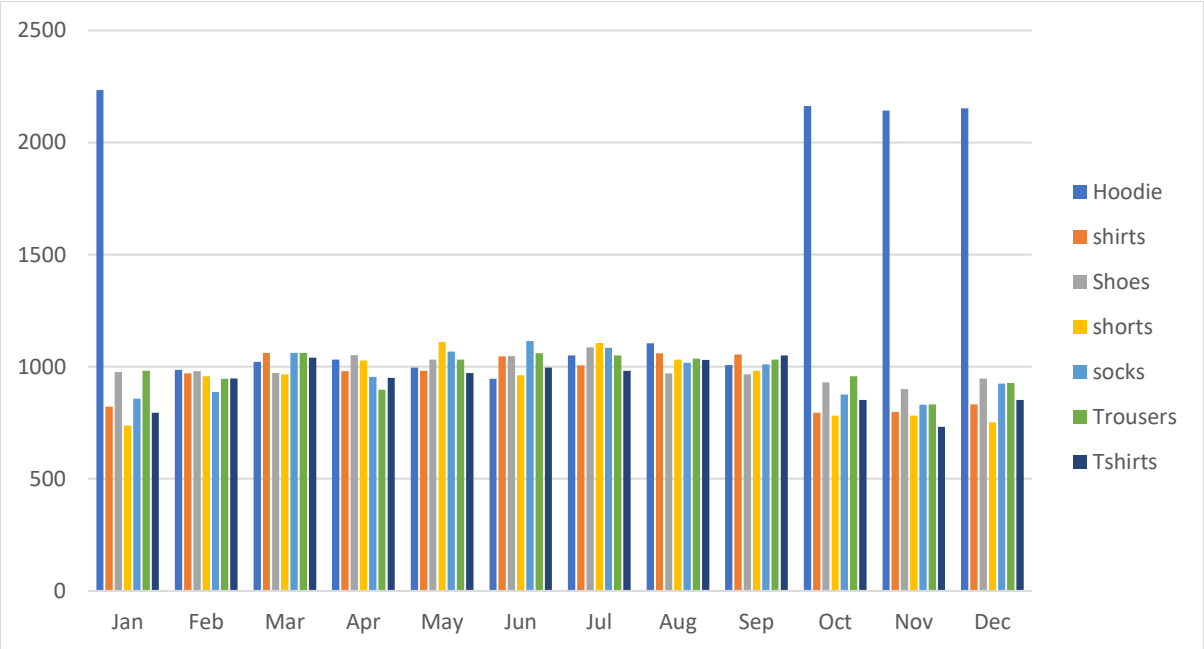
……..

➔ Below Bar graph represents the relation between color and monthly sales of the Zara dataset.
➔ This graph is generated using Matplotlib functions.
➔ X-axis represents Month and Y-axis represents number of sales.
➔ From this graph we can conclude that "BLACK" color products has less sales in months of MAY, June and July.



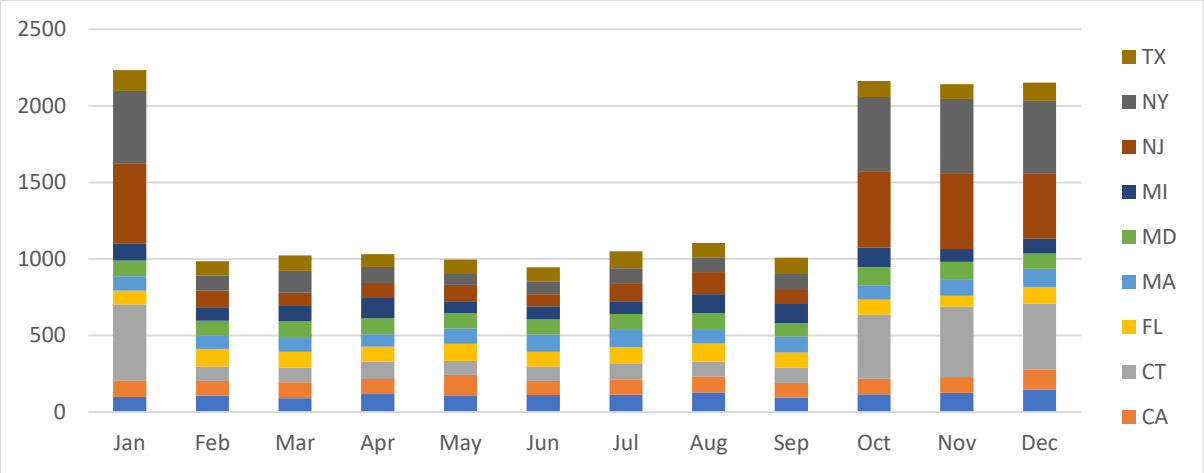
……..

➔ Below Bar graph represents the relation between product- Hoodie and monthly sales of the Zara dataset.
➔ This graph is generated using Matplotlib functions.
➔ X-axis represents Month and Y-axis represents number of sales for hoodies.
➔ From this graph we can conclude that "Hoodies" products has more sales in months of Jan, Oct, Nov and Dec.



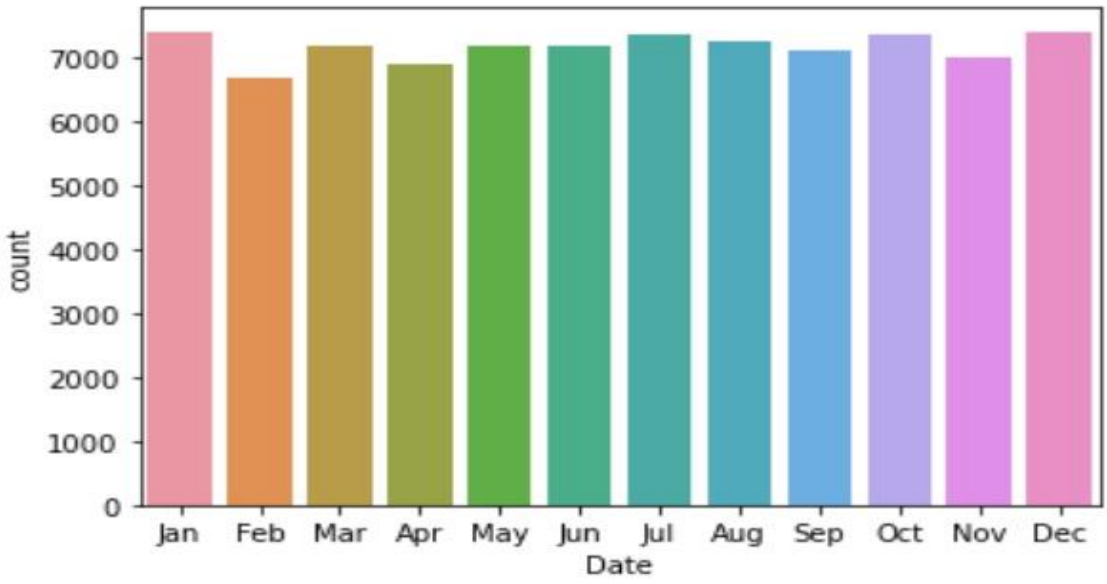
……..

➔ Below Bar graph represents the relation between product- Hoodie, monthly sales, and state wise of the Zara dataset.
➔ This graph is generated using Matplotlib functions.
➔ X-axis represents Month and Y-axis represents number of sales for hoodies.
➔ From this graph we can conclude that “Hoodies” products has more sales in states of NY, NJ and CT.



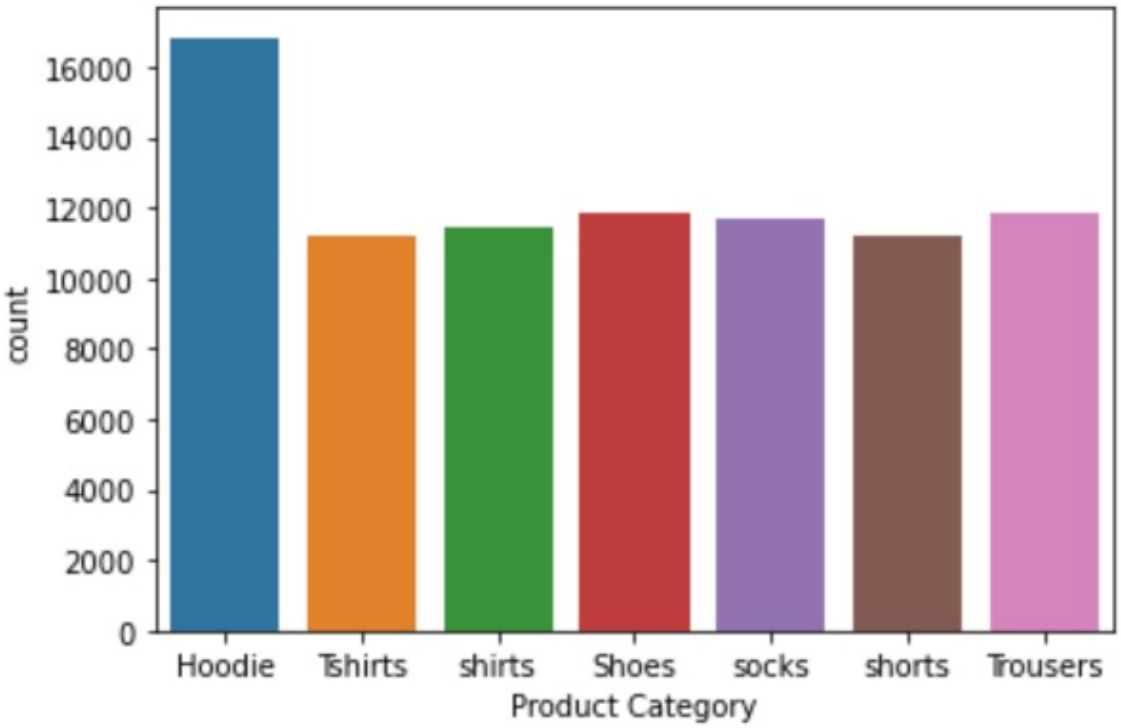
……..

➔ Below Bar graph represents the relation between total sales and monthly wise of the Zara dataset.
➔ This graph is generated using seaborn functions.
➔ X-axis represents Month and Y-axis represents number of sales of all products.
➔ From this graph we can get total sales for every month in a year



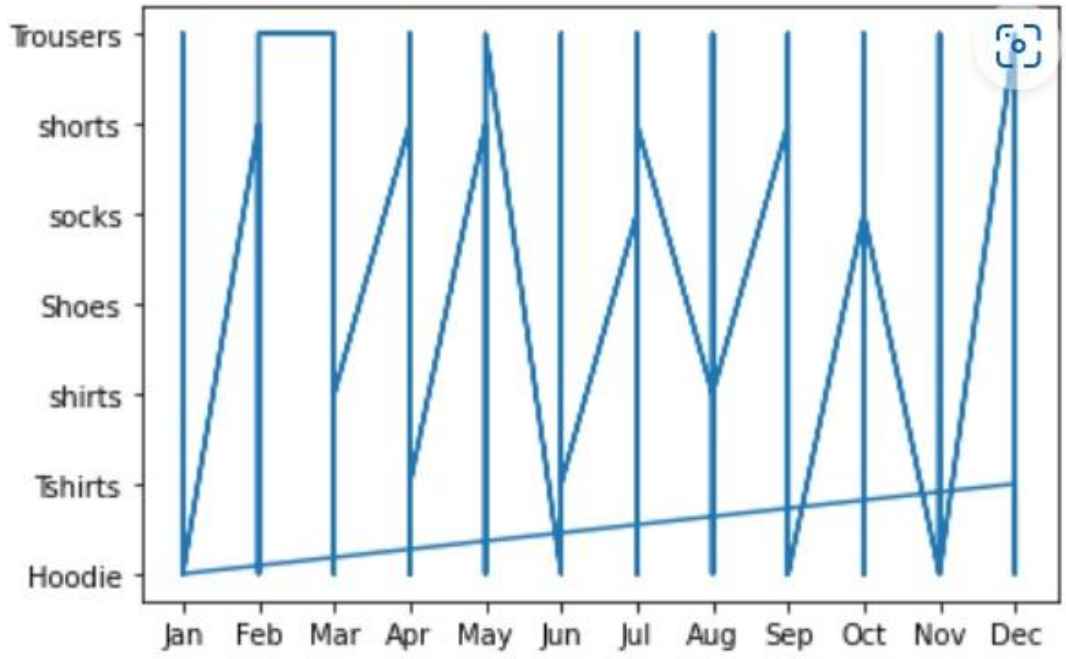
……..

- ➔ Below Bar graph represents the relation between total sales and product category wise of the Zara dataset.
- ➔ This graph is generated using seaborn functions.
- ➔ X-axis represents Product category and Y-axis represents number of sales(count) of all products.
- ➔ From this graph we can get total sales of all products based on category.



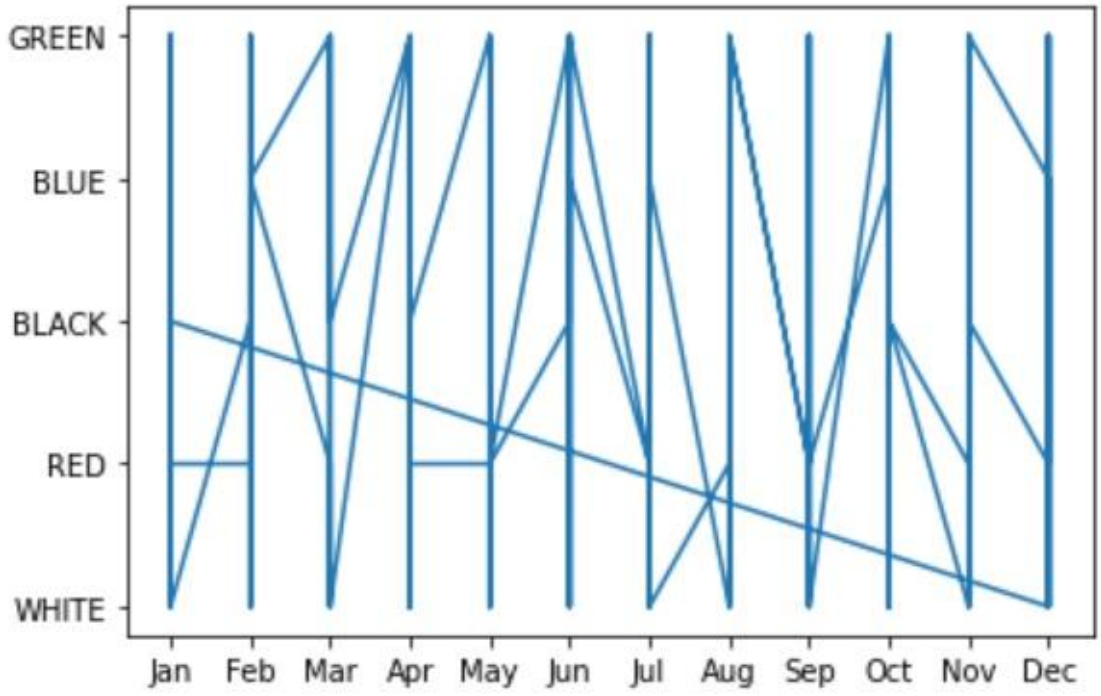
……..

- ➔ Below Line graph represents the relation between product category and month wise of the Zara dataset.
- ➔ This graph is generated using matplotlib functions.
- ➔ X-axis represents product category and Y-axis represents Product category of all products.
- ➔ From this graph can get sales variation across the year for particular category, all presented in one graph.



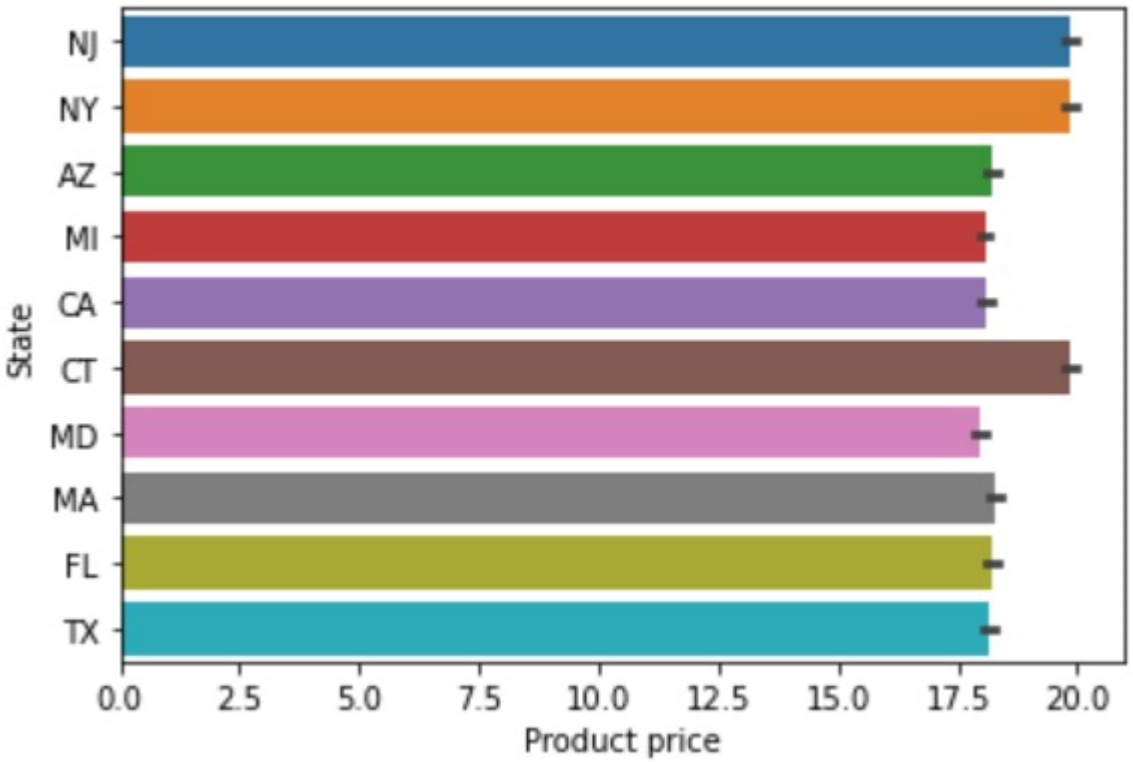
……..

- ➔ Below Line graph represents the relation between product color and month wise of the Zara dataset.
- ➔ This graph is generated using matplotlib functions.
- ➔ X-axis represents product category and Y-axis represents Product color of all products.
- ➔ From this graph can get sales variation across the year for particular product color, all presented in one graph.



……..

- ➔ Below bar graph represents the relation between product price and state wise of the Zara dataset.
- ➔ This graph is generated using seaborn functions.
- ➔ X-axis represents product prices and Y-axis represents states.
- ➔ From this graph can get sales variation across the states, all presented in one graph.


NJ
NY
AZ
MI
CA
CT
MD
MA
FL
TX
State
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Product price

References:-

https://ecommercedb.com/reports/zara-com-brand-report-poland-2021/284

https://www.dataandsons.com/data-market/product-lists/zara-uk-fashion-data-in-csv-format

https://www.kaggle.com/crawlfeeds/zara-us-fashion-products-dataset

https://search.vi-seem.eu/dataset/zara_dataset

https://www.statista.com/forecasts/1218316/zara-revenue-development-ecommercedb

https://www.kaggle.com/datasets/thedevastator/fast-fashion-eco-data

https://www.volza.com/p/zara-clothes/

https://www.globaldata.com/data-insights/consumer/the-sales-of-zara-in-clothing---footwear-industry-in-china-1861343/

https://www.aggdata.com/clothing_store_locations/zara_spain

https://agenty.com/marketplace/stores/zara-retail-store-locations-in-germany

https://github.com/fediazgon/zara-data-challenge-19