OEIT6 - Data Analytics

Experiment 1: Exploratory Data Analysis

Name: Gaurav Panchal

UID: 2019120046

Data Set Link: https://drive.google.com/file/d/1sDZ08Sh-EJ_kkucuNziUduJQzIpiesFa/view?usp=sharing

Importing required Libraries

```
from pydrive.auth import GoogleAuth
from google.colab import drive
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
import matplotlib
matplotlib.rc('figure', figsize = (20, 8))
matplotlib.rc('font', size = 14)
matplotlib.rc('axes.spines', top = False, right = False)
matplotlib.rc('axes', grid = False)
matplotlib.rc('axes', facecolor = 'white')
```

Connecting Dataset to Google Colab

```
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
file_id = '1sDZ08Sh-EJ_kkucuNziUduJQzIpiesFa' #<-- You add in here the id from you google
download = drive.CreateFile({'id': file_id})
```

```
download.GetContentFile('modcloth_final_data.json')
df  = pd.read_json("modcloth_final_data.json", lines=True)
```

Using the pd.read_json() function the json file is brought into a pandas DataFrame, with the lines parameter as True- because every new object is separated by a new line.

```
df.head()
```

	item_id	waist	size	quality	cup size	hips	bra size	category	bust	height	use
0	123373	29.0	7	5.0	d	38.0	34.0	new	36	5ft 6in	
1	123373	31.0	13	3.0	b	30.0	36.0	new	NaN	5ft 2in	sydneybrac
2	123373	30.0	7	2.0	b	NaN	32.0	new	NaN	5ft 7in	
3	123373	NaN	21	5.0	dd/e	NaN	NaN	new	NaN	NaN	alexm
4	123373	NaN	18	5.0	b	NaN	36.0	new	NaN	5ft 2in	dbe

EDA - Exploratory Data Analysis

We can already make few observations here, by looking at the head of the data:

- There are missing values across the dataframe, which need to be handled.
- Cup-size contains multiple preferences- which will need handling, if we wish to define cup sizes as 'category' datatype.
- Height column needs to be parsed for extracting the height in a numerical quantity, it looks like a string (object) right now.
- Not so important, but some columns could do with some renaming- for removing spaces.
- Firstly, we handle the naming of columns for ease-of-access in pandas.

Number of Instances:

```
df.columns
```

```
    Index(['item_id', 'waist', 'size', 'quality', 'cup_size', 'hips', 'bra_size',
           'category', 'bust', 'height', 'user_name', 'length', 'fit', 'user_id',
           'shoe_size', 'shoe_width', 'review_summary', 'review_text'],
          dtype='object')
```

We can see that column names are inconsistent and has *spaces* in it. Let's Clean it

```
df.columns = [names.replace(' ', '_') for names in  df.columns]
```

Attribute Information:

[ ] ↳ *31 cells hidden*

Feature Engineering

Creating a new feature of first_time_user Building on our observations above, it makes sense to identify the transactions which belong to first time users. We use the following logic to identify such transactions:

If bra_size/cup_size have a value and height, hips, shoe_size, shoe_width and waist do not- it is a first time buyer of lingerie.

If shoe_size/shoe_width have a value and bra_size, cup_size, height, hips, and waist do not- it is a first time buyer of shoes.

If hips/waist have a value and bra_size, cup_size, height, shoe_size, and shoe_width do not- it is a first time buyer of a dress/tops.

Below we will verify the above logic, with samples, before we create the new feature.

1. Looking at the few rows where either bra_size or cup_size exists, but no other measurements are available.
2. Looking at the few rows where either shoe_size or shoe_width exists, but no other measurements are available.
3. Looking at the few rows where either hips or waist exists, but no other measurements are available.

[ ] ↳ *1 cell hidden*

EDA via visualizations

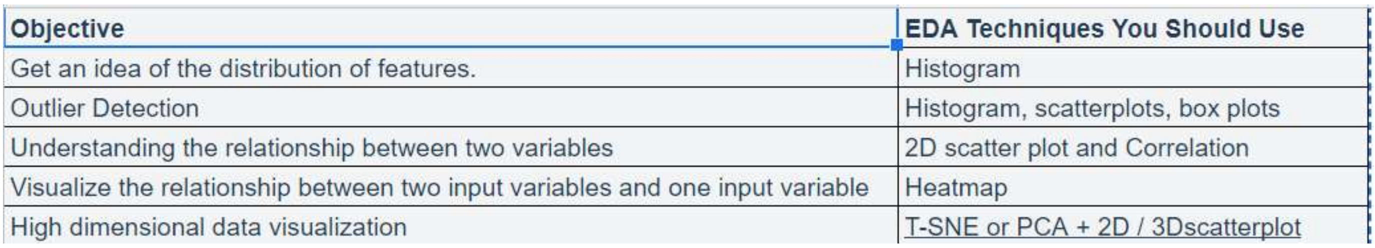
1. Distribution of different features over Modcloth dataset
2. Categories vs. Fit/Length/Quality Here, we will visualize how the items of different categories fared in terms of - fit, length, and quality. This will tell Modcloth which categories need more attention!

[ ] ↳ *20 cells hidden*

Inference:

Exploratory data analysis is the most important step in any data science task. The main objectives of the EDA are:

1. Analyze data distribution
2. Detect outliers and anomalies
3. Select the most important features
4. Remove unnecessary columns
5. Removing and filling in missing values
6. Discover the hidden motives
7. Better understanding of patterns within the data.
8. Find interesting relations among the variables.
9. Using statistics and visualizations to analyze and identify trends in data sets.



Objective	EDA Techniques You Should Use
Get an idea of the distribution of features.	Histogram
Outlier Detection	Histogram, scatterplots, box plots
Understanding the relationship between two variables	2D scatter plot and Correlation
Visualize the relationship between two input variables and one input variable	Heatmap
High dimensional data visualization	T-SNE or PCA + 2D / 3Dscatterplot