



Data Science Project Management

Case Study 1
Group-2



Team Members	Tasks
Ahmad Iqbal	Data Visualization + Presentation
Arshpreet kaur	Data Visualization + Documentation
Abhiram Pazhuvelil Sathyan	Documentation + Github
Basant Singh	CNN + Image PreProcessing
Keerthi Reddy Dokuri	Dashboard + Data Preprocessing
Meenakshi Remadevi	Data Visualization (Tableau) + Presentation
Muhammad Anas	Data Visualization + Data Preprocessing
Muhammad Hasan	Dashboard + Image PreProcessing
Nigel George	Github + Data Visualization (Tableau)
Tony Regi Jacob	CNN + Model Tuning

Table of Content

- Tasks To Do
- Update of Dashboard
- Data Visualization and Insights
- Data PreProcessing
- CNN Model
- Results
- Links

Tasks To Do

- Updates to the project process in both the code and the project board
- Python notebook, sequentially executed code and an HTML file
- Updated Project Board and Git repository

Update on Dashboard

Added **Milestone** On Dashboard to create Data Insights using Visualization and to build Initial CNN Based Model

Milestones

[Follow](#) [Edit Milestone](#) [...](#)

JUNE
15
Thursday



Data Insights & Intial Model Building

You are responsible

To pre-process data and visualise it using Tableau & Python techniques, such as univariate and bivariate analysis, in order to get some practical modeling or insight.

To create an initial CNN-based model and test various tuning parameters in order to achieve high accuracy

Published

[Edit](#) [Delete](#) [Add Tag](#) [Attach task list](#)

Today

Update on Dashboard

Created and **assigned** Tasks among all team members to get a real time update from each member about every single task.

Used **Tags** and **Priority** tasks to know about **severity** and **current status** of tasks.

BrandDatasetVisual > My List

6 Pre - Process Data

Pre-processing data entails modifying raw data to make it acceptable for machine learning or analysis tasks. Data cleaning (handling missing values, outliers, and inconsistencies), data integration (combining data from different sources), data transformation (converting variables into appropriate formats), feature selection/extraction (choosing relevant features or creating new ones), data discretization (dividing continuous variables into categories), handling imbalanced data (addressing class imbalances), and data splitting (creating training, validation, and test sets) are the steps. These procedures guarantee data quality and get the data ready for more modeling or analysis.

[See less](#)

Assigned to



Dates

Jun 4th - [Today](#)

Priority



Medium

Tags



[Published](#) [×](#)

Board column



Completed

Update on Dashboard

Created **subtasks** for every team member and **details added** to subtasks along with declaring **prerequisite** as well as deadlines.

Added **Code Screenshots** to dashboard to keep the the members updated about other's task.

Created by

AI Sun 4th June 2023, 15:06

Progress

Set progress: 90%

Task ID

#27025475

See less

Subtasks

6

Hide ^

> 1 Review

Jun 4th - Today

AI

✓ Normalizing and scaling features

Jun 4th - Today

M

✓ Handling missing or null values

Jun 4th - Today

B

✓ Visualize and dataset exploration

Jun 4th - Today

M

✓ Draw assumptions based on visualization

Jun 4th - Today

A

✓ Tableau Visualization Of Data

Jun 4th - Jun 19th

A

+ Add subtask

Tasks Breakdown

Tasks :

Q Assignee or task name



Incomplete Tasks

Anytime



+ Add Task List



Task Name	Assign...	Due Date	Priority	Estimated time	Logged time	Tags	+
▼ My List ... + *							
▼ 4 Pre - Process Data 4	M	Today	Medium	—	—	Published x	...
> 1 Review 1	AI	Today	High	—	—		...
✓ Normalizing and scaling features	M	Today	High	—	—		...
✓ Visualize and dataset exploration	M	Today	High	—	—	Published x	...
✓ Draw assumptions based on visualization	A	Today	High	—	—		...
+ Add a subtask							
⊖ Initial CNN-Bases Model 3	T	—	High	—	—	Published x	...
> ⊖ Split Data into Training, Testing and Validation 3	M	Jun 25th	Medium	—	—	Draft x	...
> ⊖ Image Pre Processing 6	B	Jul 27th	High	—	—	Draft x	...
> ⊖ Models Building 5	A	Aug 4th	High	—	—	Draft x	...
+ Add a task 6-completed							

Getting started 3

Data Visualization and Insights

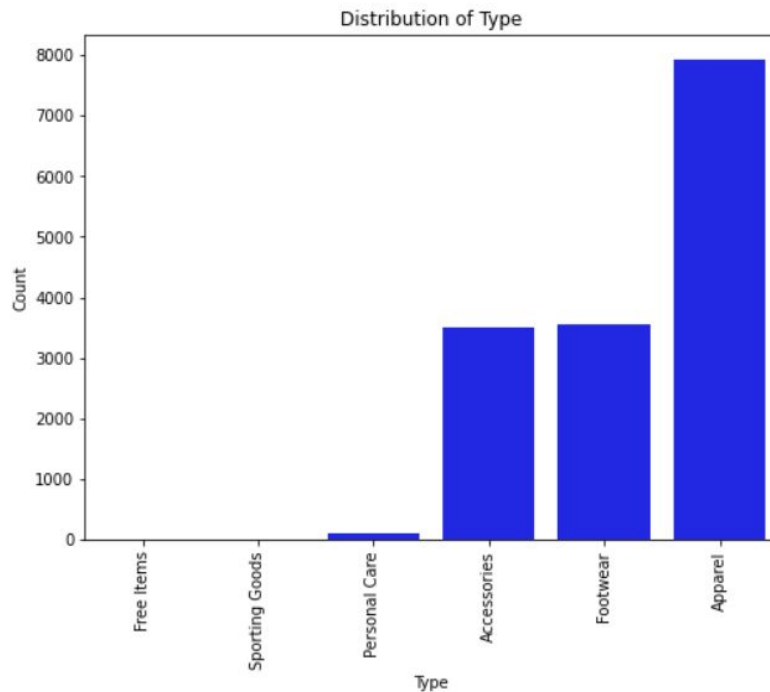
We have 9 unique categorical variables in our dataset among which further types are there

ID	GenderType	Type	SubType	Article	PrimaryColor	Seasonal	Year	Use	Brand
39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0	Casual	Peter England
59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0	Casual	Titan
53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0	Casual	Puma
29114	Men	Accessories	Socks	Socks	Navy Blue	Summer	2012.0	Casual	Puma
9204	Men	Footwear	Shoes	Casual Shoes	Black	Summer	2011.0	Casual	Puma

Data Visualization and Insights

The graph shows the types of Items **Wearable Items**

Apparel, Footwear and **accessories** are the highest used items by customers.

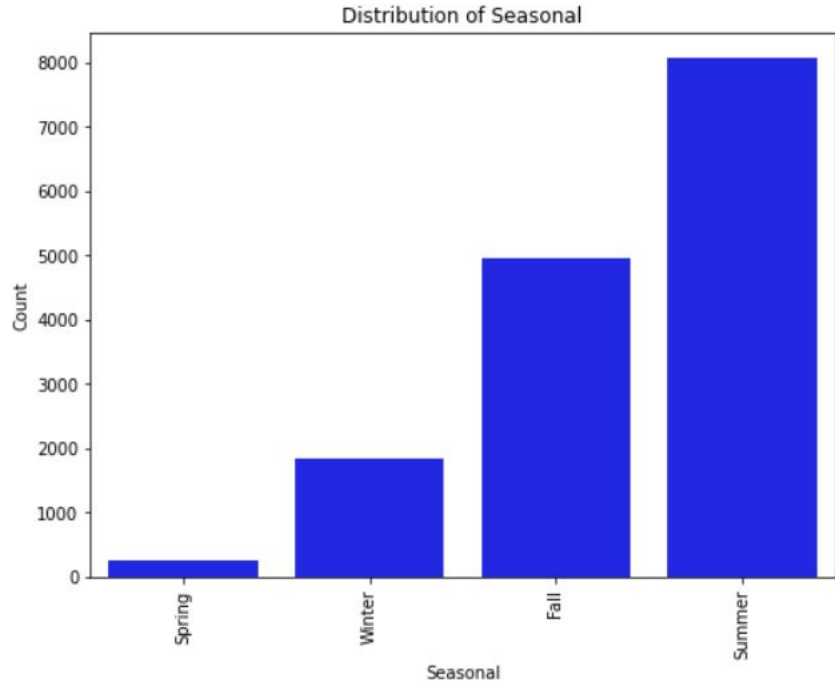


Data Visualization and Insights

The graph shows the **high** tendency of **summer's** product usage by customers

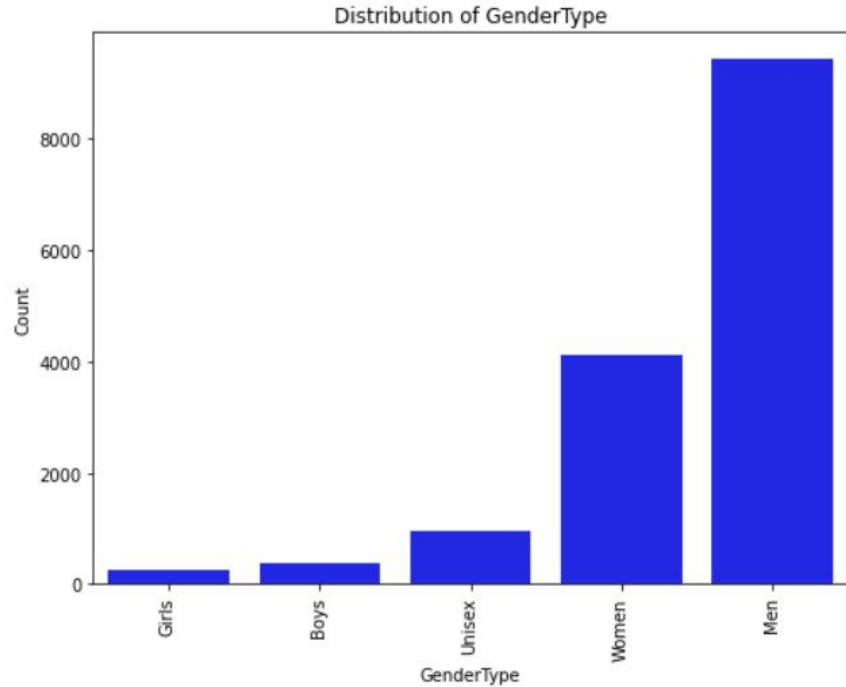
Fall is the 2nd highest with **winter** 3rd.

While **spring** is the least.



Data Visualization and Insights

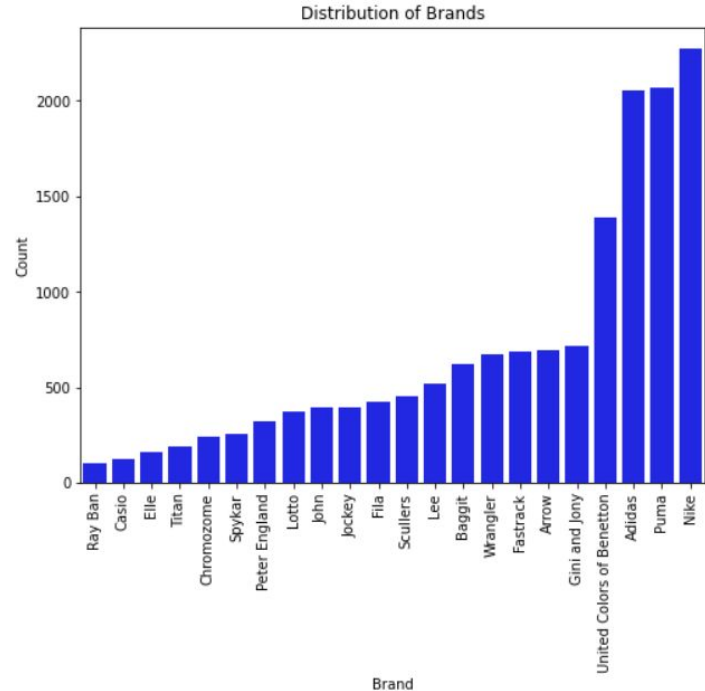
Men's are the most user of products with women's second highest



Data Visualization and Insights

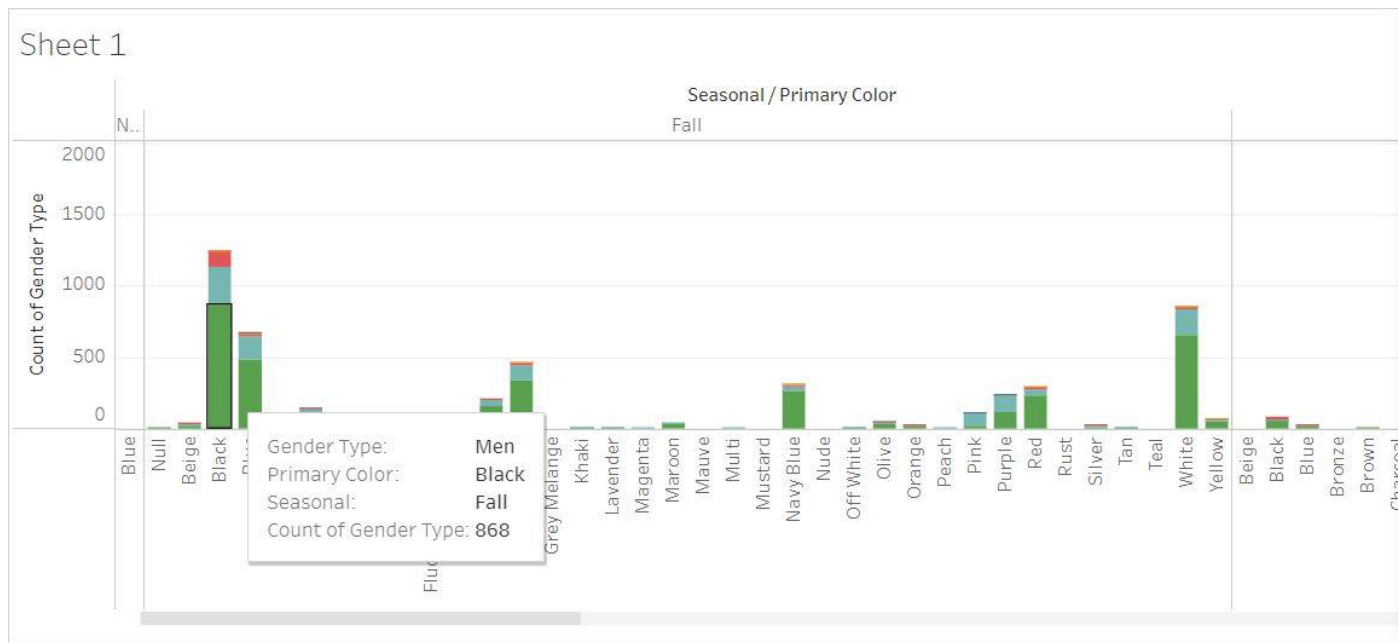
The graph indicates that customers loves to buy products from Nike, Puma and adidas

Whereas very few number of customers buys from Ray ban.



Data Visualization (Tableau)

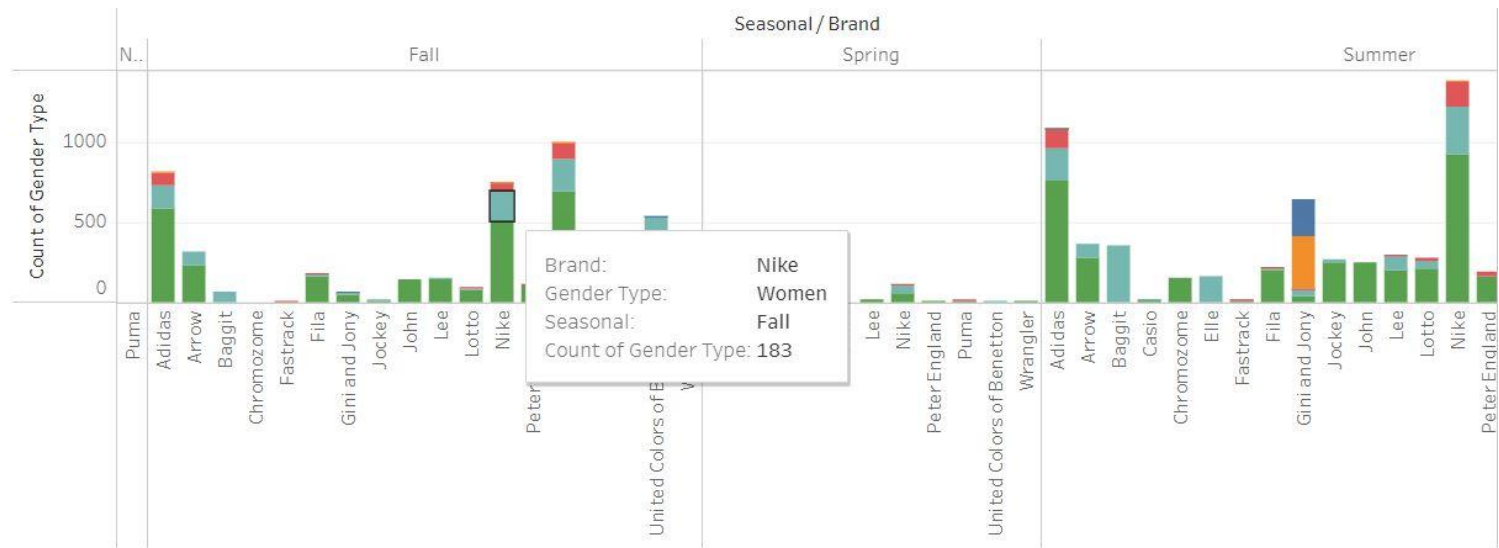
The **colour** most worn by particular genders in a specific season



Data Visualization (Tableau)

The **Brands** worn by different gender types in different seasons

Sheet 1



Data Visualization (Tableau)

The **items** and **number** of item users among all seasons

Sheet 1

Gender	Sub Type	Seasonal				
		Null	Fall	Spring	Summer	Winter
	Eyewear				2	200
	Flip Flops		103	8	170	7
	Fragrance			55		
	Free Gifts				1	3
	Gloves		5			
	Headwear		29	1	24	2
	Innerwear		7		366	91
	Loungewear and Nightwe..		4		13	11
	Mufflers		6		5	
	Sandal		132	14	138	9
	Scarves		2		8	
	Shoes		767	32	1,300	17
	Socks		60	1	145	1
	Ties		3		33	
	Topwear	1	2,040	23	1,909	36
	Wallets		16		94	13
	Watches				31	448

Data Pre-Processing

- Data Preprocessing is done to distinguish between **Real** and **fake Images** available in dataset
- Separated **IDs** for Real and Fake images used for Model
- Train Test **split** is done in **70:30** ratio
- Processed the images using various techniques that includes **gray scale, rotation range, zoom range , width and height range, horizontal flip, and fill mode.**

CNN

- Built and used **CNN** Model to Classify target variable
- **Relu** and **Sotmax** functions were used as activation function
- **Adam** Optimizer helped to increase the accuracy to a great extent
- Used **12** epoch for the model to run and learn every time

Results

We only scored **69.79%** for our initial Model, which is not very good.

We obtained an **AUC** of **0.771%**, which indicates that the model can distinguish between positive and negative outputs with reasonable accuracy, but still we were unable to do so.

Out of 14000 data points, we got **3900** True Negative Values , **661** False Positive Values, **7429** False Negative Values and **1307** True Positive Values which shows that our model has potential to learn but still we need to work more on model parameters.

After visualization , we came to know we have lots of **categorical** and highly **unbalanced** data that make it difficult to train **CNN** on it.

Due to highly unbalanced data, AUC and other kernel would be very less, and ultimately it would be tough for our model to learn from data.

Links

Github

<https://github.com/Tonyri/Data-Science>

Project Dashboard

<http://surl.li/ibtnd>