Empirical Analysis of Black Friday Shopping Trends

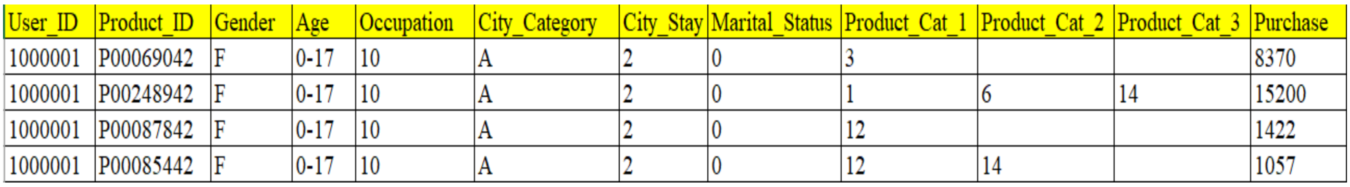
Jeswin Maria Cinthia Palthasar, Praveen Natarajan,
Nivedhaa Murali, Shabana Fathima Samsudeen

Abstract

Black Friday is one of the biggest and busiest shopping events of the year. With almost all the goods on sale, the retailers aim at making profit from the huge sales and buyers look forward to purchasing branded products at a bargain. The black Friday sales tally hit $7 billion in 2019 and it is a win-win situation for all involved parties. The flashy sale on black Friday not just brings the frequent shoppers but also the niche of shoppers who shop rarely and look for good deals. The data of this major shopping event can be made useful in more ways than one. Instead of focusing on revenue estimation and prediction, this analysis and report aims at deeply investigating the purchase patterns of the shoppers and identifying subsets of shoppers with similar behaviour. By categorizing the shoppers into groups based on their demographic and geographic information, the companies and retailers can work towards creating collective yet personalized campaigns and offers to capture the customers.

Data Summary

Black Friday raw dataset consist of 550068 observations and 12 variables. The snippet below shows structure of dataset used for visualization purpose.



User_ID	Product_ID	Gender	Age	Occupation	City_Category	City_Stay	Marital_Status	Product_Cat_1	Product_Cat_2	Product_Cat_3	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3			8370
1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200
1000001	P00087842	F	0-17	10	A	2	0	12			1422
1000001	P00085442	F	0-17	10	A	2	0	12	14		1057

Data Visualization

Data Pre-processing

The dataset has been imported into Power BI from a CSV file. The data is tidy and missing values in cells denotes that the customer did not purchase those items. Hence the missing values are not removed and are included in the analysis. This visualization is based on only one dataset and so the detection of mapping is not required.

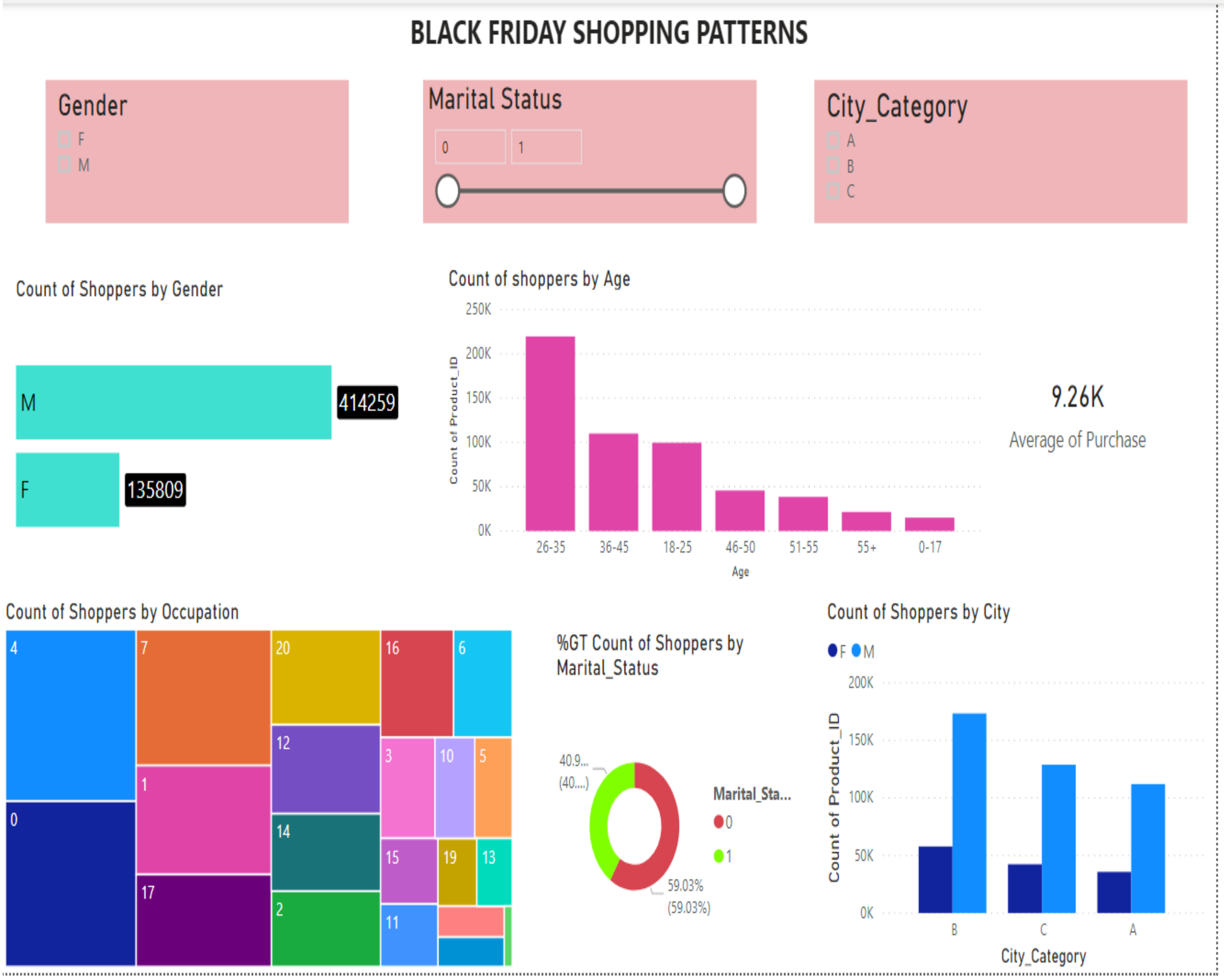
Design

- **Overview**

The objective of this dashboard is to visualize the shopping trends based on different categorical variables. Three main categorical variables such as Gender, Marital Status and City Category are chosen as filters (Slicers at the top) to observe the customers buying patterns. Bar charts, Funnel chart, Donut chart and Heat map is used to observe the trends among different gender, age group, occupation, marital status, and cities.

- **Interaction**

The interaction of the model is based on the slicers placed on the top of the dashboard. There are 3 slicers on Gender, Marital Status and City. These slicers allow us to select the value and observe the customer behaviour based on inputted values.



- **Shopping trends by Gender and Age group.**

From the horizontal bar chart, it is evident that there are a greater number of male shoppers than Female shoppers. The vertical bar chart indicates that people of 26 to 35 years of age are more interested in buying products during black Friday sale than the other age groups. If the filters are applied, say city A is selected this general trend changes. It is observed that people of age 55 years or more shop the least which is not the general trend. Hence the filters are useful to discover the hidden patterns.

- **Shopping trends by Occupation, Marital status, and City.**

From the heat map, it is evident that customers with Occupation_4 buy more products than others. The donut chart shows that single people are more interested in black Friday sale than married people. The clustered bar chart at the right end indicates that city B has the highest sale and city A has the least sale.

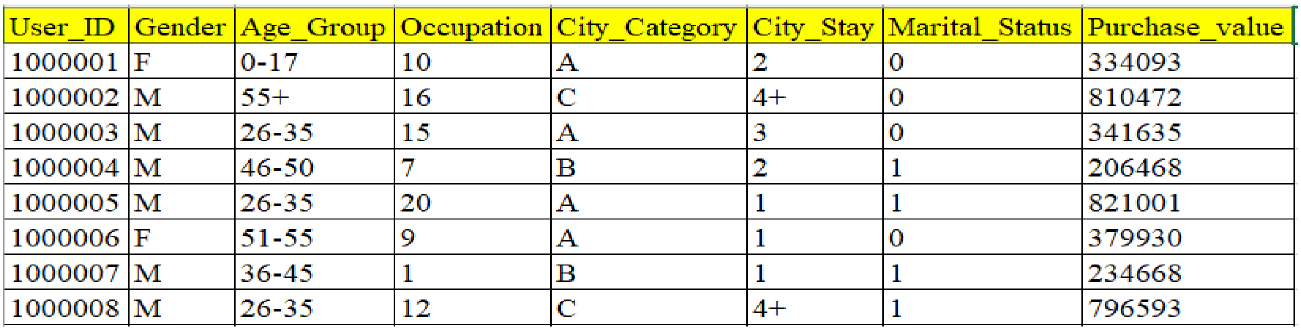
Upon conducting initial analysis using Power BI dashboards, we notice multiple patterns emerging from the dataset. Using this analysis as starting point, we proceed further to identify how these factors and categories influence one another and in turn impact the purchase trends.

Multiple Correspondence Analysis (MCA)

MCA is a statistical method for analysing and visualizing a dataset containing more than two categorical variables. MCA is typically used to analyse a dataset from a survey to identify groups of individuals with similar profiles and the association between variable categories. Black Friday dataset is analogous to the survey data in that we analyse user characteristics from the categorical variables like age, occupation, gender and demographic related information followed by the degree of association between the variables.

Data pre-processing

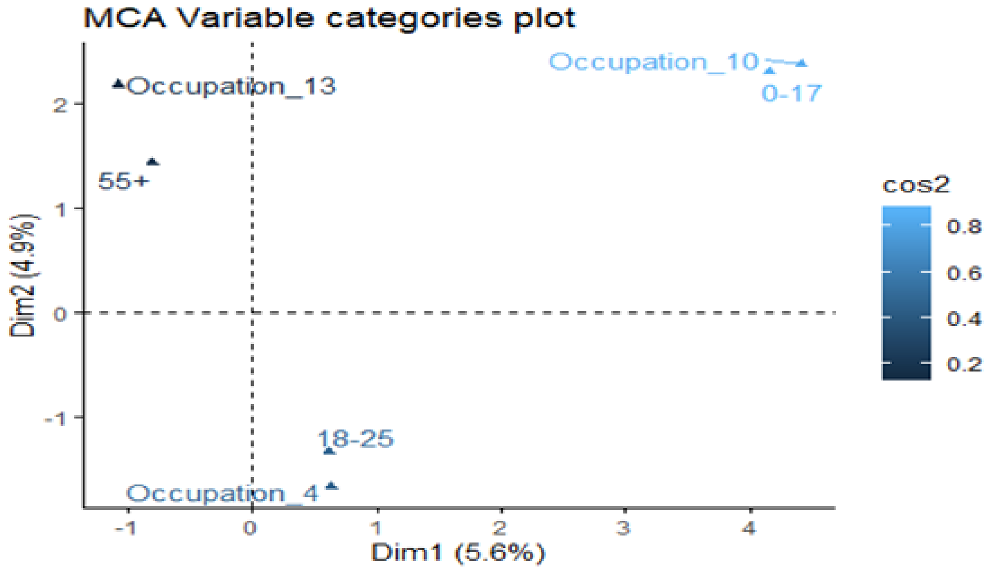
The raw dataset has multiple observations (rows) for each User ID. Our understanding of the raw dataset is that every User ID's age group, occupation, gender, marital status, residence and duration of stay in the current city are the same. However, the distinguishing variables are Product categories and the Purchase Values. The revised dataset consists of the total purchase value for each User ID. The transformed dataset consists of 5891 observations of 8 variables encompassing each User ID's features and their total purchase values as shown below.



User_ID	Gender	Age_Group	Occupation	City_Category	City_Stay	Marital_Status	Purchase_value
1000001	F	0-17	10	A	2	0	334093
1000002	M	55+	16	C	4+	0	810472
1000003	M	26-35	15	A	3	0	341635
1000004	M	46-50	7	B	2	1	206468
1000005	M	26-35	20	A	1	1	821001
1000006	F	51-55	9	A	1	0	379930
1000007	M	36-45	1	B	1	1	234668
1000008	M	26-35	12	C	4+	1	796593

Variable Categories plot to show association between the category levels

The plot shows the association between occupation and age groups variable categories. It is worth mentioning that the graph shows only the variable categories that are well-represented based on the cos2 values (quality of representation). It is recommended to segment users based on these identified groups in order to account for a higher proportion of variance in comparison to those variable categories that are clustered to the origin.



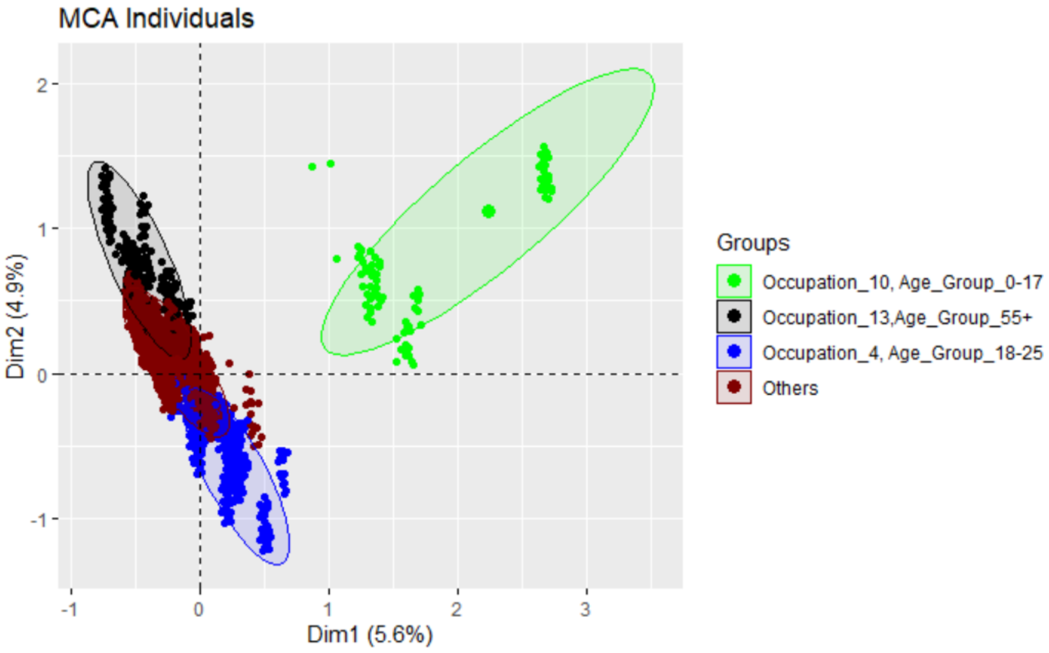
MCA Individuals plot for customers segmentation

The MCA individuals plot shows the segmentation of groups based on the user's characteristics (Occupation, Age_group etc). A concentration ellipse is highlighted around each group for better clarity. Customers are segmented into different groups based on their characteristics as follows:

Group 1: Occupation_10, Age_Group_0-17

Group 2: Occupation_13, Age_Group_55+

Group 3: Occupation_4, Age_Group_18-25
Group 4: Others

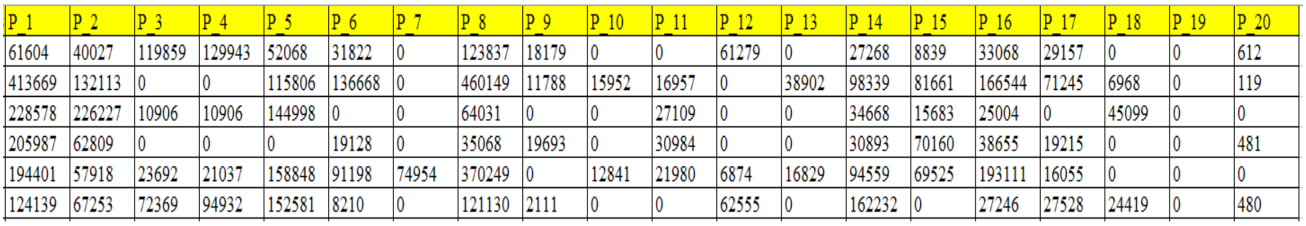


Principal Component Analysis (PCA)

In the previous section, customers were segmented based on the characteristics using the qualitative variables in the dataset. However, it is important to segment the customers based on their shopping behaviours (inclination toward product categories) using the quantitative variables leveraging the concept of Principal Component Analysis (PCA).

Data pre-processing

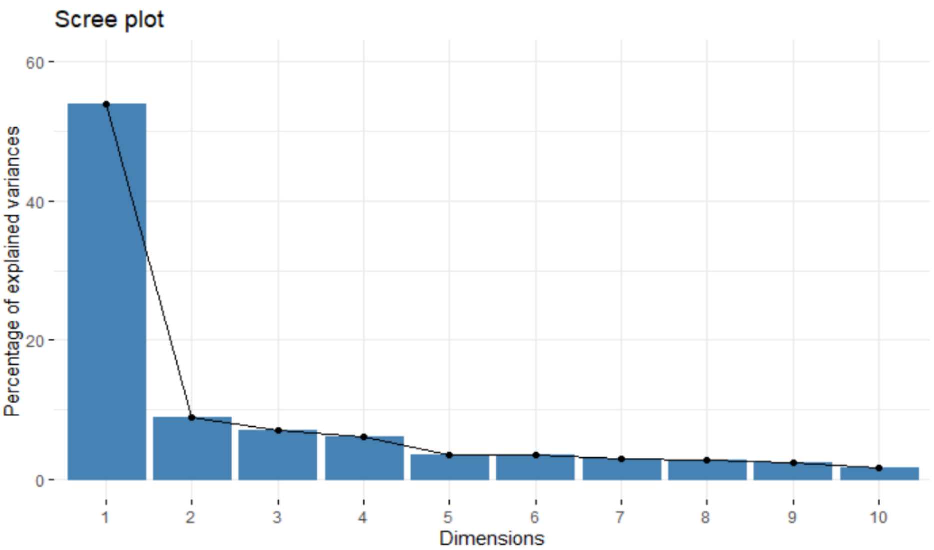
The raw data has been transformed extensively in order to implement the PCA method. From the purchase values and product categories IDs' (categorical levels from 1 to 20) in the raw dataset, the user's purchase value for each category has been determined. The transformed dataset has additional columns for each product category identifier (1 to 20) with the corresponding purchase values as shown. The table shows an excerpt of the complete dataset. The zeros in some observations imply that the user has not purchased any product in that category and therefore purchase value is zero.



P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10	P_11	P_12	P_13	P_14	P_15	P_16	P_17	P_18	P_19	P_20
61604	40027	119859	129943	52068	31822	0	123837	18179	0	0	61279	0	27268	8839	33068	29157	0	0	612
413669	132113	0	0	115806	136668	0	460149	11788	15952	16957	0	38902	98339	81661	166544	71245	6968	0	119
228578	226227	10906	10906	144998	0	0	64031	0	0	27109	0	0	34668	15683	25004	0	45099	0	0
205987	62809	0	0	0	19128	0	35068	19693	0	30984	0	0	30893	70160	38655	19215	0	0	481
194401	57918	23692	21037	158848	91198	74954	370249	0	12841	21980	6874	16829	94559	69525	193111	16055	0	0	0
124139	67253	72369	94932	152581	8210	0	121130	2111	0	0	62555	0	162232	0	27246	27528	24419	0	480

Scree plot to visualize the percentage of Explained Variance

From the Scree plot output shown below, the first dimension, also referred as the Principal Component 1 (PC1) accounts for more than 50% of the variation in the quantitative variables, while the second dimension (Principal Component 2 or PC2) accounts for just a little less than 10%. As we go further, the percentage of explained variance decreases gradually. Therefore, the first two dimensions are chosen as the principal components.



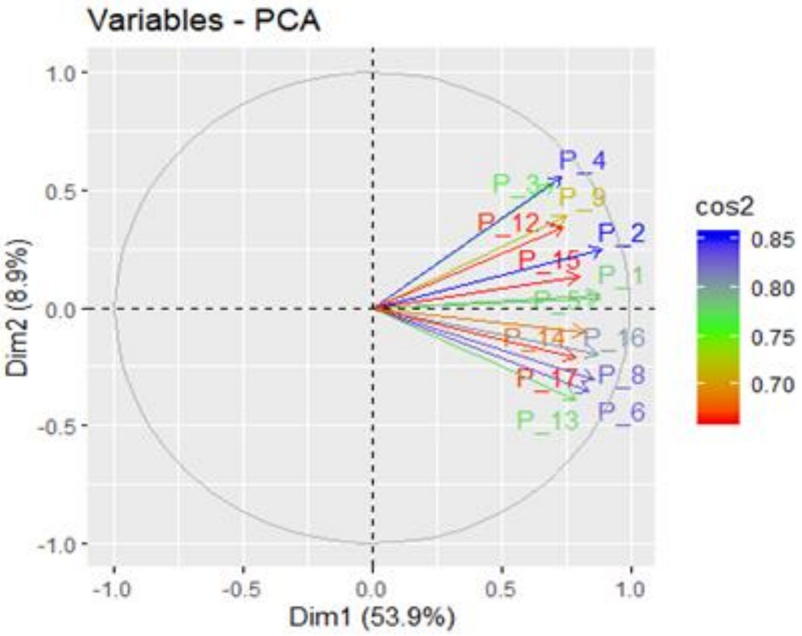
Variables plot to identify the correlation of Product Categories

It can be inferred from the Variables plot that all product category levels lie on the positive side of the first principal component (Dim1). However, when viewed from the angle of second principal component, (Dim2 as represented by the dashed vertical axis) product category levels can be segmented to two groups as follows-

Group 1: P_4, P_3, P_12, P_2, P_15, P_5, P_9, P_1

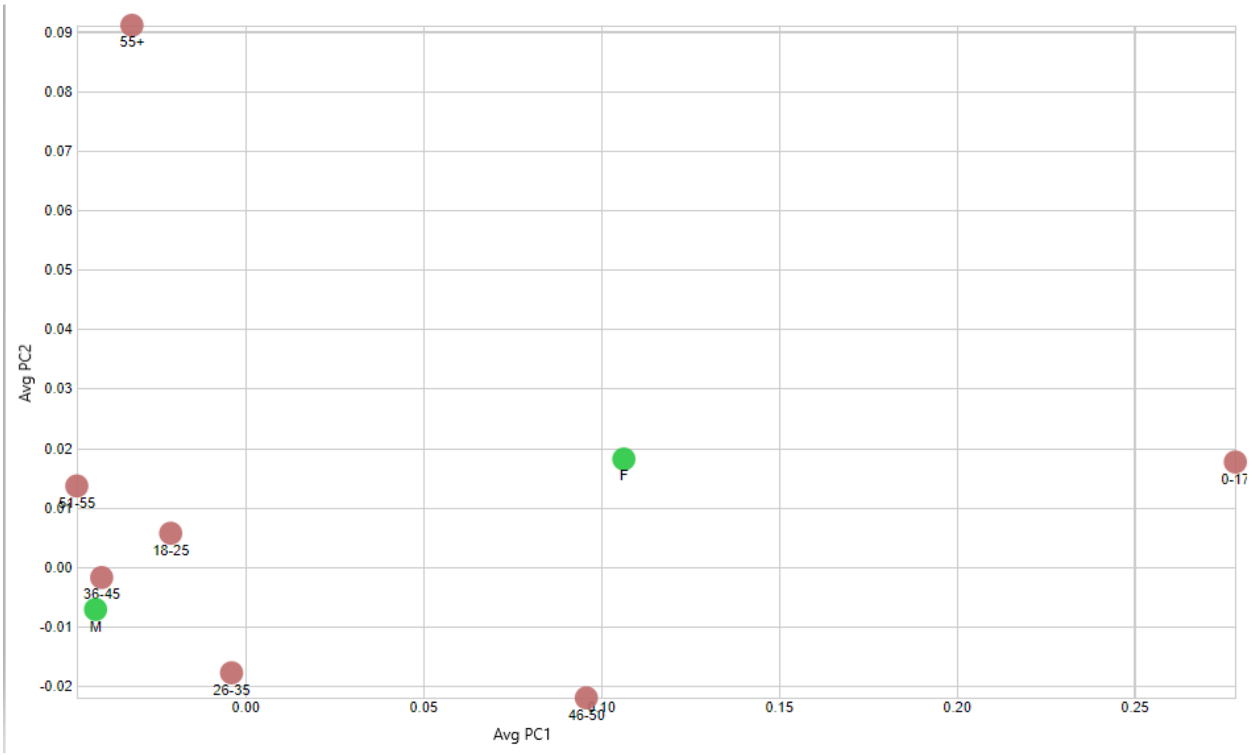
Group 2: P_14, P_16, P_8, P_6, P_17, P_13

The levels that have high quality of representation (cos2>0.6) are shown in the correlation circle and are chosen for product category group segmentation.



Scatter Plot for PCA-Variables and Categorical Variables

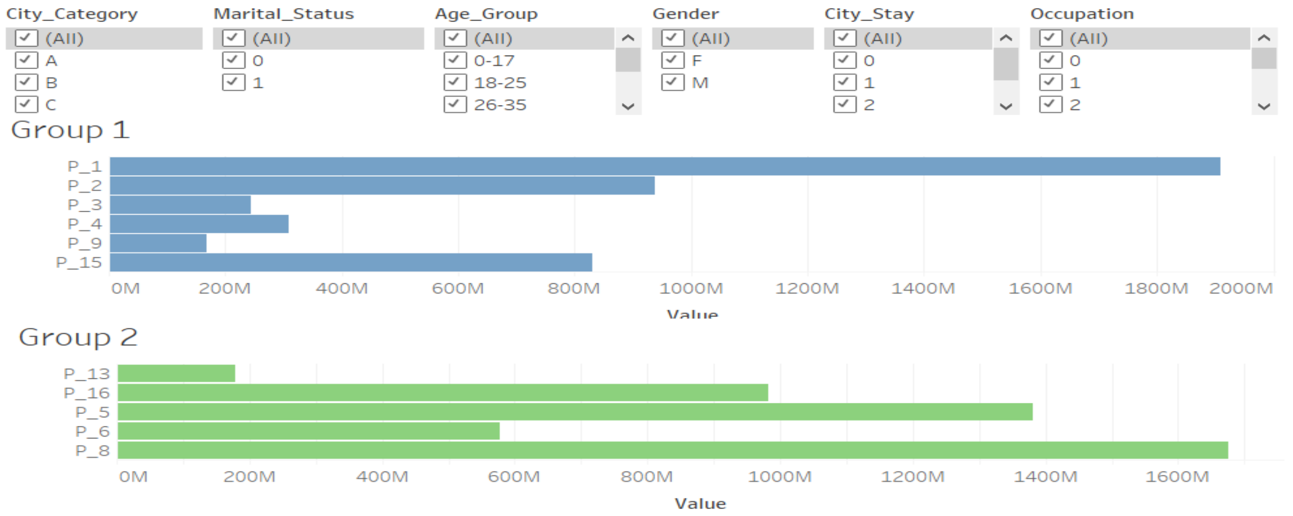
The below graph has been plotted based on average of PC1 and PC2(the chosen principal components), Age and Gender categorical variables from our data. The graph conveys that Female shoppers from age 0-17 show similar shopping trend, likewise Male shoppers in age groups of 36-45, 18-25, 26-35 and 51-55 have same shopping trends. This again can be used as an evidence that Male customers shop the most during the sale.



Recommendations

Black Friday sale is one of the most expected events of the year and almost all people splurge to buy everything they need, want and desire. The insights gained from the data if applied appropriately can help improve the shopping experience for customers and the sales for retailers not just on Black Friday but all through the year. With the details we have gathered from the data, we make a few recommendations as follows:

- The products categories segmented in the same group can be given shelf spaces next to or nearby each other. Insights from following dashboards can be used to identify customer groups with higher purchase rates and make specific product groupings within the same group.
- Multi-buy promotions or clubbed offers (BOGO) can be given on the products found under the same group to encourage buying more commodities.
- Occupation and Age are found to be the most impactful qualitative variables; hence these factors should be given serious consideration while segmenting customers.
- Products can be bundled and sold to the groups of customers identified from MCA. Packaged deals benefit the sellers as they sell more goods and buyers get discounted deals.



City_Category	Marital_Status	Age_Group	Gender	City_Stay	Occupation
☑ (All)	☑ (All)	☑ (All)	☑ (All)	☑ (All)	☑ (All)
☑ A	☑ 0	☑ 0-17	☑ F	☑ 0	☑ 0
☑ B	☑ 1	☑ 18-25	☑ M	☑ 1	☑ 1
☑ C		☑ 26-35		☑ 2	☑ 2

References

Data Source

https://www.kaggle.com/sdolezel/black-friday

Textbooks

Alboukadel Kassambara. (2017). *Practical Guide to Principal Component Methods in R.*

Andy Field., Jeremey Miles., Zoe Field. (2012). *Discovering Statistics using R.*

Discussion boards

https://piazza.com/class/k540vtexyhd28x?cid=143

https://bioinformatics.stackexchange.com/questions/3917/hierarchial-pca-clustering-with-duplicated-row-names

https://stackoverflow.com/questions/37861578/factominer-mca-error-regarding-numeric-values

https://stackoverflow.com/questions/59362413/errors-making-predictions-on-mca-object-factominer

https://stackoverflow.com/questions/54621526/error-in-dimnamesres-of-mca-factominer

https://stackoverflow.com/questions/27246276/factominers-mca-runs-out-of-memory

Blogs

https://medium.com/@u3554364/an-end-to-end-project-of-product-and-customer-segmentation-analysis-in-retail-store-with-r-and-809a4cef0ba

https://rstudio-pubs-static.s3.amazonaws.com/262425_b3168ae9be4a48028815c91747b78f8b.html

https://www.vendhq.com/blog/sales-promotion/

https://www.shopify.ca/encyclopedia/customer-segmentation

https://rpkgs.datanovia.com/factoextra/

Visualization

https://www.youtube.com/watch?v=Nr31rv9tsJ8&t=38s

https://www.youtube.com/watch?v=ql5-VDbMo0w

https://www.tableau.com/learn/training/20201

https://www.analyticsvidhya.com/blog/2017/07/data-visualisation-made-easy/

https://docs.microsoft.com/en-us/power-bi/guided-learning/