Customer Shopping Trends Insights

Table of Contents

Appendix

Exploratory Data Analysis

Figure 1 - Total Sales by product

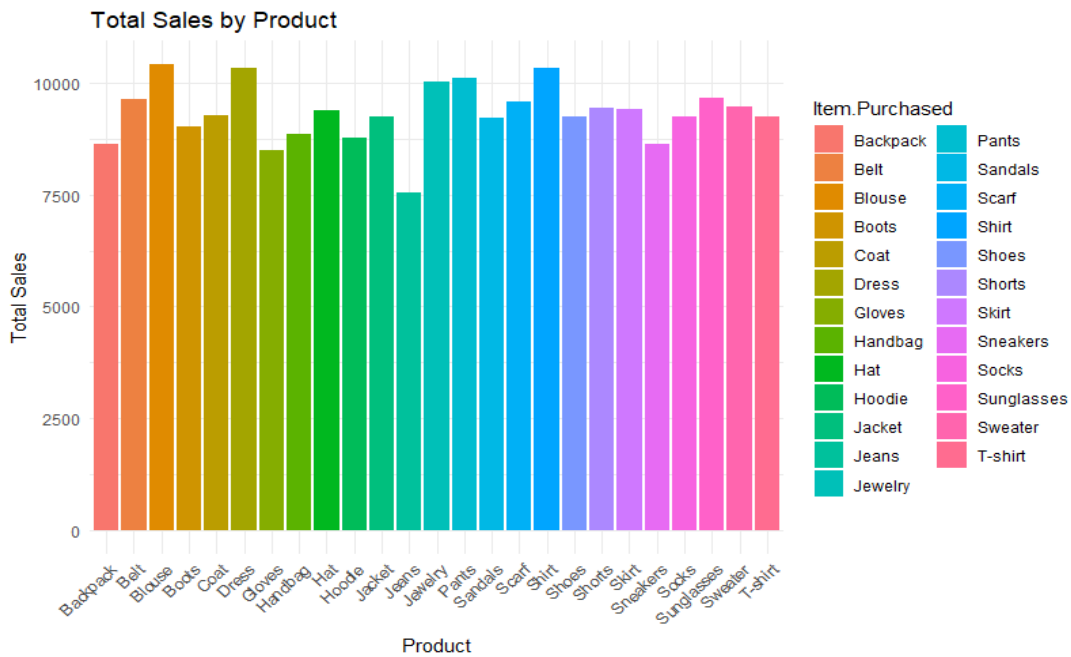


Figure 2 - Total sales by Category

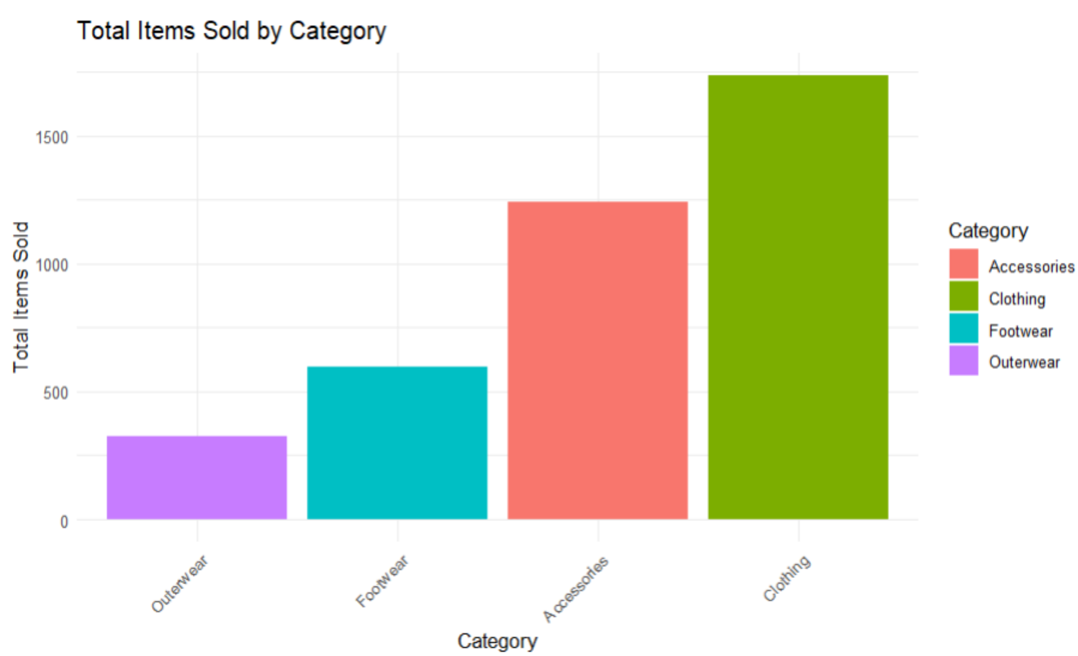


Figure 3 - Consumer Segments by age analysis

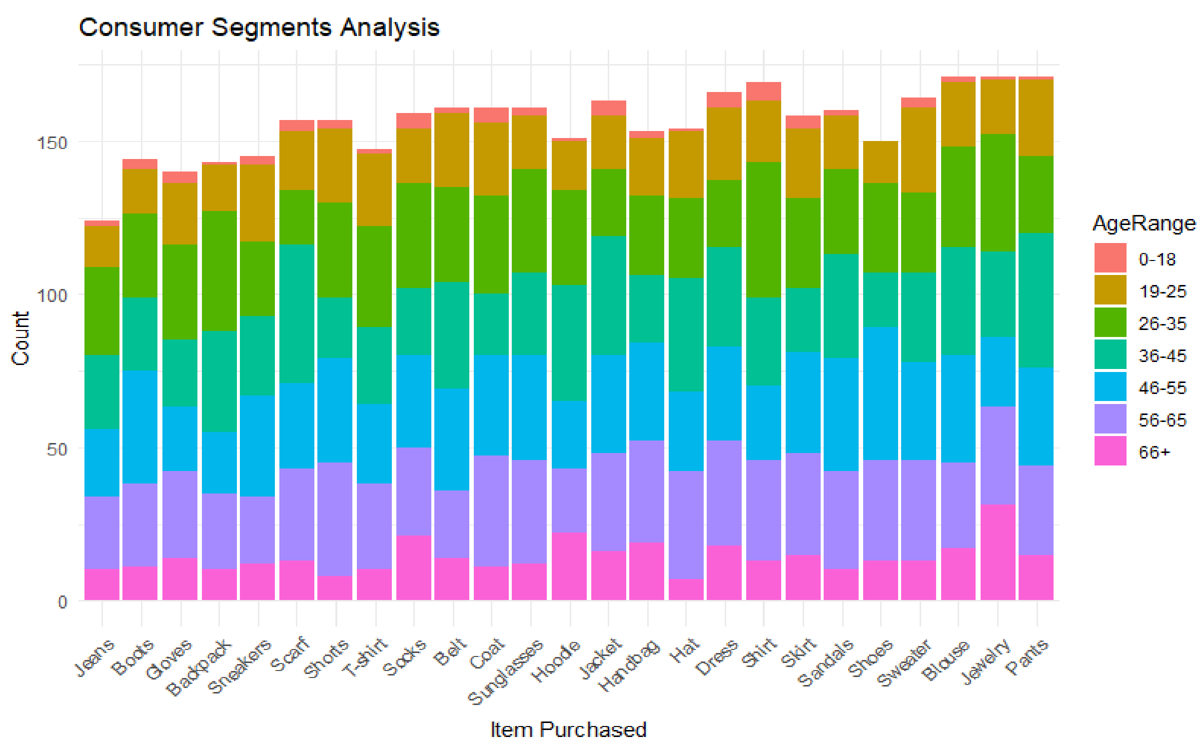


Figure 4 - Seasonal product analysis

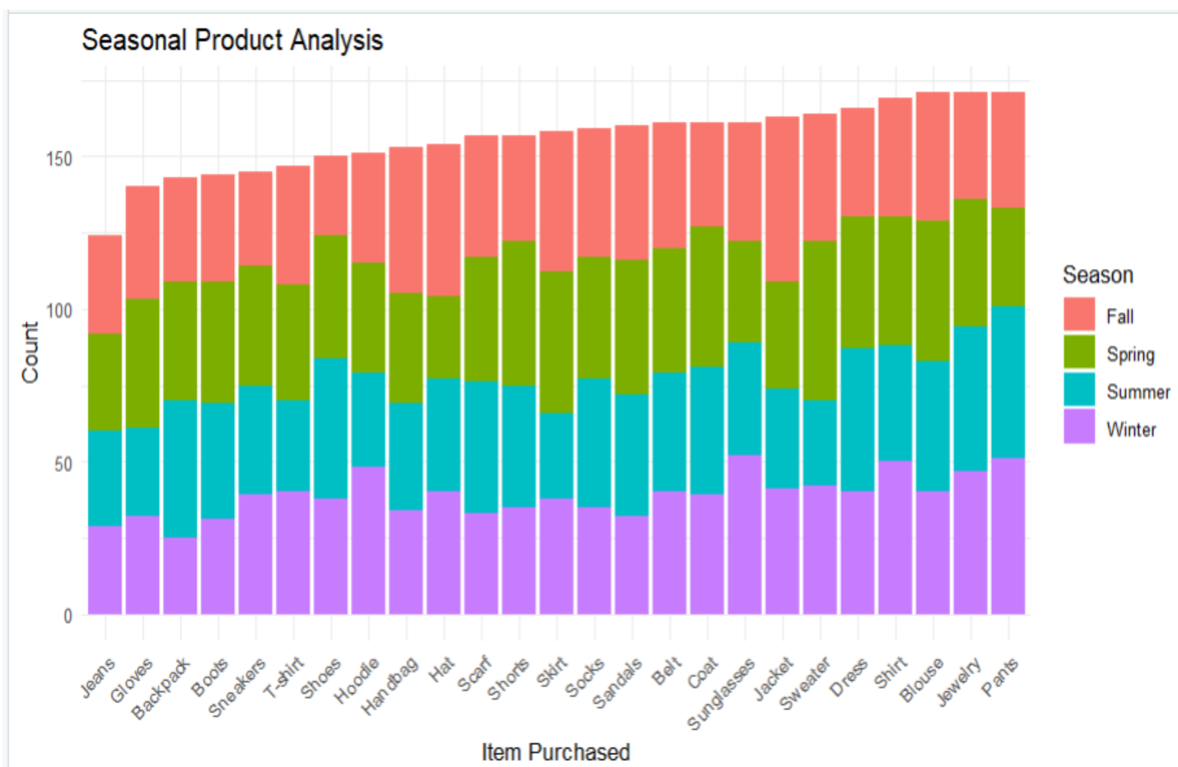


Figure 5 - Count of categories sold by size

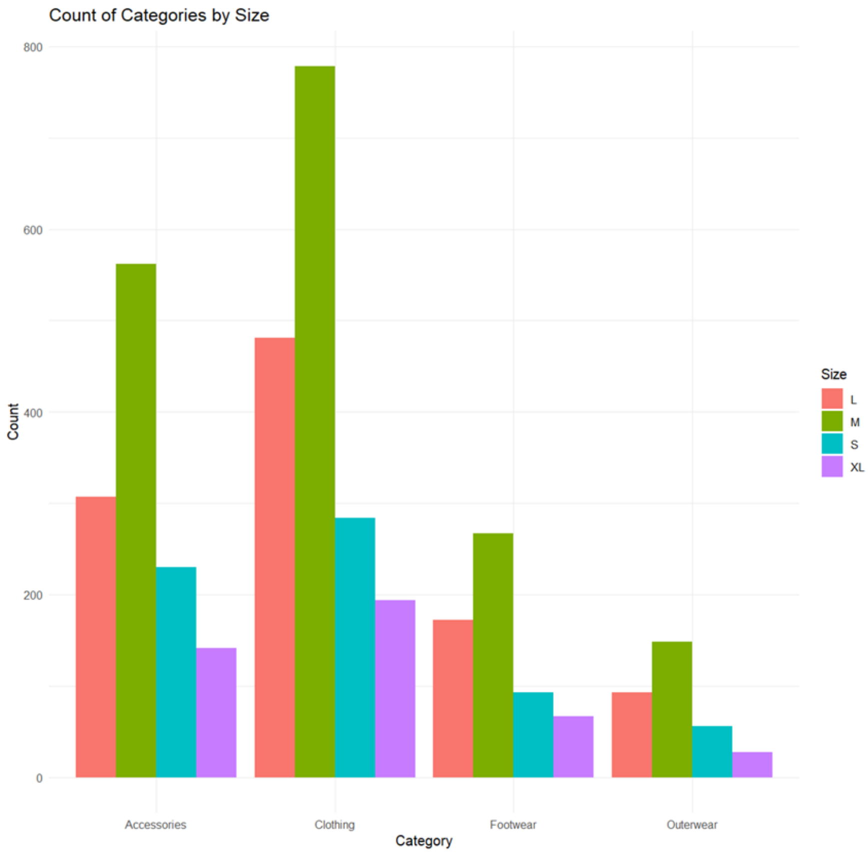
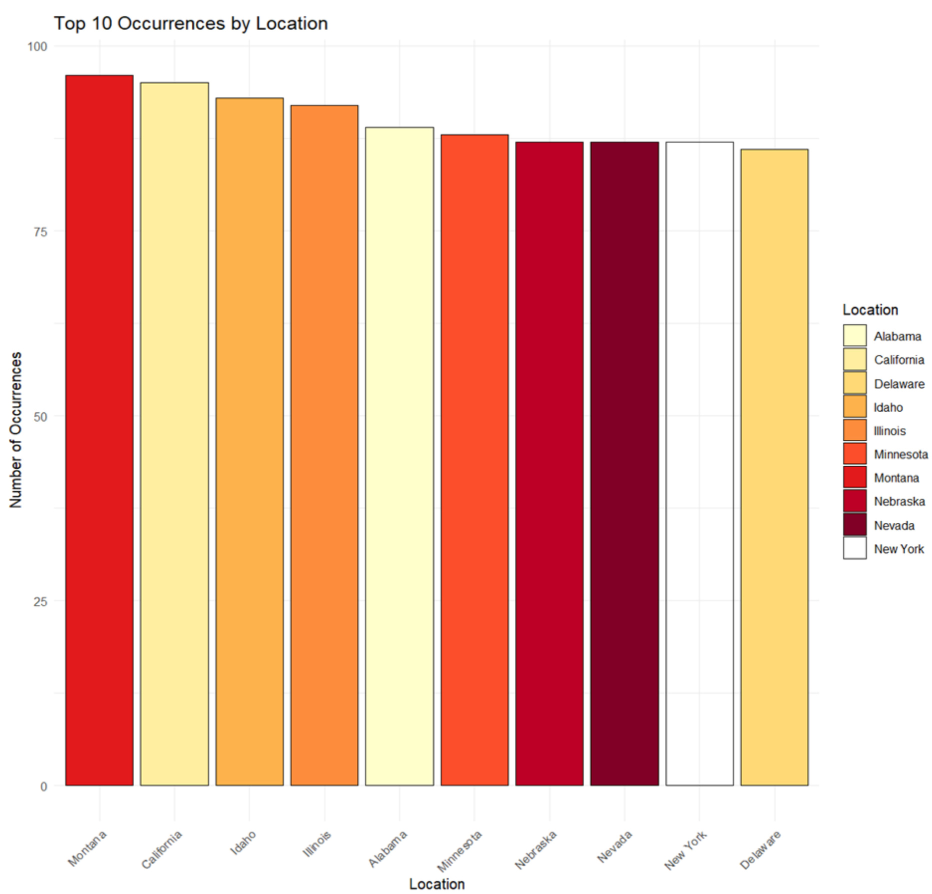


Figure 6 - Transactions by location



Code

R File: https://drive.google.com/file/d/1tOQE162t6Qw9lVS30tTdbg4oWv5tXLW4/view?usp=drive_link

Dataset:

https://drive.google.com/file/d/1Tw3h8hPIU2j2BeXM6dyNi5SED7ItO8gh/view?usp=drive_link

Google Drive: https://drive.google.com/drive/u/1/folders/0AAMmevo0wg5-Uk9PVA

Machine Learning

Figure 7 - K-Means Clustering by Age group to determine average Review Ratings

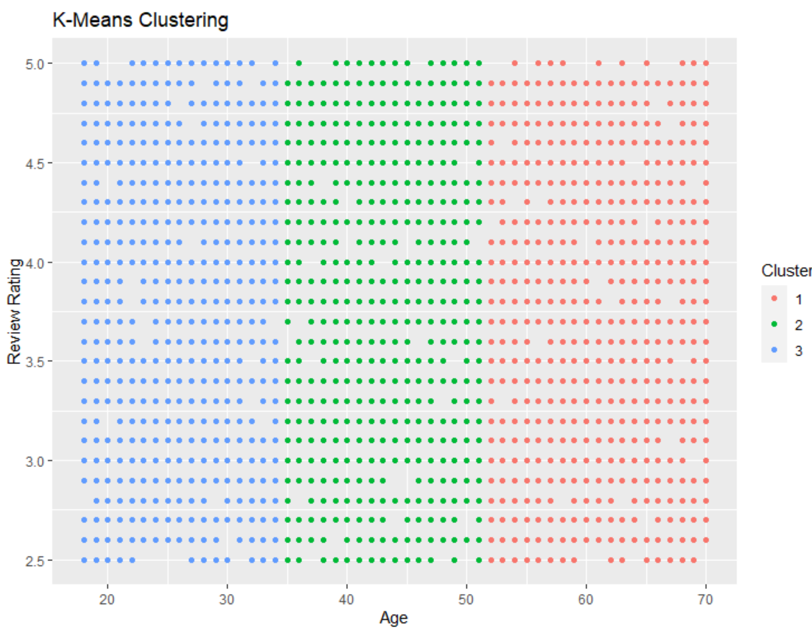


Figure 8 - K-Means Clustering by Age to determine average purchase amounts. (2 clusters)

Figure 9 - K-Means Clustering by Age to determine average purchase amounts. (3 clusters)

Business Problem

In a rapidly changing environment of retail shopping, the main business problem for managers is providing shoppers with a relevant selection of products. This is an industry where consumer preferences change constantly with new emerging trends. Providing shoppers with an appealing inventory poses a significant challenge to these businesses. To solve this problem, managers must gain insights into consumer behavior by **identifying best selling items within each category offered, understanding different customer segments & preferences, and seasonal fashion trends.** The analysis of such characteristics is necessary in meeting the needs of consumers and prepares managers for the constant changing environment of fashion retail.

By employing modeling techniques, **our objective is to offer managers with valuable insights on consumer behavior.** We hope to provide them with the necessary tools to increase sales performance by optimizing their inventory and improve marketing strategies according to the market. **We will offer recommendations based on our analysis to aid the process of making informed business decisions.**

Data Description

This dataset was obtained from Kaggle and it contains various variables that describe **characteristics of products such as their color, category, consumer's behavior and purchasing patterns** around these various items. The dataset has a wide range of variables to describe **consumer characteristics such as age, gender, product category, purchase amount, frequency of purchases, customer reviews, applied discounts, and season** when the purchase was made.

Relevance of the data

This dataset has a range of 3900 consumer records that can be used to understand the customer's behavior and purchasing patterns. This will allow us to analyze consumer preferences and market trends which will assist the businesses in their decision-making processes. **This dataset can be used to build machine learning models with the goal of increasing sales by helping businesses to accommodate their production according to the trending demands of consumers.**

Exploratory Data Analysis

Our analysis begins with a focus on descriptive analytics where we identify the trends among the shoppers. This will provide managers with an insightful overview of how current business is performing and what beginning steps they should begin to take to improve business performance.

- We have first calculated the Amount of Sales by Item purchased. To achieve that, we grouped the 'item purchased' column from the shopper's dataset and summarized the purchased amount($USD) according to each purchased item. Finally, we sorted the values in descending order.
- In figure 1, we have graphed the total sales distribution for each product. It reveals that blouses, dresses, and shirts are the most purchased items. This observation tells us there is a high demand for these clothing items. It would be advisable to managers to focus on inventory levels to maintain a steady supply to keep up with demand and increase sales. Also, tailoring marketing efforts to advertise to customers the availability of these items.
- Next we wanted to identify which category has the most sales which can be seen in figure 2. We used the category column and counted the total sales per category. Our findings reveal clothing is the dominant category in total sales with accessories closely following. Managers should leverage this data to put more resources towards these main categories which drive revenue. Keeping

sufficient inventory levels and marketing new arrivals of items in these categories to attract more shoppers.

- In figure 3 our analysis focused on consumer segments, using age as the key parameter. The categories were determined as followed:

  age_ranges <- c(0, 18, 25, 35, 45, 55, 65, Inf)

  age_labels <- c("0-18", "19-25", "26-35", "36-45", "46-55", "56-65", "66+")

- A new column, 'age range,' was created to categorize shoppers. Notably, shoppers under 18 exhibited a greater demand for socks, coats, dresses, and shirts. Shoppers aged 66 and above showed a significant interest in jewelry products. This was an interesting observation as it could be linked to their financial capacity.. In contrast, younger age groups have lower occurrences of buying jewelry. Managers should strategically tailor marketing strategies to appeal to respective age groups. For example, advertising the new availability of socks, coats, dresses, and shirts on social media to attract younger shoppers.
- To identify seasonal trends with products, we counted the number times a product sold within each season for each product. This allows us to see how the time of year influences sales which is shown in figure 4. We are able to identify specific trends such as jeans having the lowest performing sales throughout the year. In fall, shoppers prefer to buy jackets which shows they are preparing for the coming winter season. In the spring, shoppers tend to buy sweaters for the mild weather. During the summer, shoes, backpacks, and dresses are the most popular products. From this analysis, managers can consider when to increase inventory levels for each item that is popular based on the time of year. For example, just before the fall season, it would be wise to start stocking up popular fall items like jackets.
- To identify the most frequently purchased sizes by category type, we counted the items sold per each category and are shown in figure 5. Again, clothing being the most popular category, and in this category M' followed by 'L' sizes are the most bought items across all categories. This is a

noteworthy trend which gives managers an overview of what sizes shoppers require. Managers can incorporate this information into their logistical planning in terms of production and procurement and making sure there is an optimal amount of inventory for these specific sizes.

- Figure 6 shows the distribution of sales across the locations of stores. It reveals that Montana, California, and Idaho are the top three locations with the highest amount of transactions. From this, managers should prioritize the allocation of resources to these locations as they drive the most business. Also, tailoring marketing strategies to less popular locations such as Delaware to attract more shoppers to this location.

Machine Learning Techniques

KNN:

<u>Subscription-Status</u>

We tried using KNN to predict if a customer would sign up to a 'subscription' given their different shopping behaviors. For this we used the "caret" library to utilize the trainControl() method of this package and applied k-fold validation for 5 segments. After testing with different values for the number of neighbors 'k', we finally chose tunegrid from 1 to 30 neighbors but were able to achieve an accuracy of 72.92% for k equal to 28.

```
  17  0.7241027   5.557093e-03
  18  0.7241021   4.680000e-03
  19  0.7251287  -6.349777e-05
  20  0.7256395   9.403945e-04
  21  0.7271777   4.862550e-04
  22  0.7241011  -6.428402e-03
  23  0.7261510  -1.526436e-03
  24  0.7269213   8.181756e-04
  25  0.7266662  -5.750624e-03
  26  0.7276918  -1.134545e-03
  27  0.7276918  -3.716197e-03
  28  0.7292306   1.081971e-03
  29  0.7284617  -2.199093e-03
  30  0.7284614  -1.334006e-03

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 28.
> 
```

Category

We also tried to predict the ‘Category’ of a product a person would predict based on their shopping habits.

```
   90  0.4366686   0.0000782326
   91  0.4394891   0.0040451955
   92  0.4397468   0.0042753920
   93  0.4405131   0.0048192781
   94  0.4397432   0.0033418174
   95  0.4384612   0.0009380022
   96  0.4384615   0.0005848246
   97  0.4369230  -0.0025087130
   98  0.4374372  -0.0013521131
   99  0.4356410  -0.0052475235
  100  0.4376923  -0.0019529105

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 93.
>
```

But we were able to achieve only 44.05% for a huge value of k =93, so we tried using random forest instead for both of these observations.

Random Forest:

For random forest we kept the value of mtry=5 because the optimal value for random forest classifier model should be near to ‘p/3’ and we have 17 columns in our dataset. We created a ‘fold’ vector to apply k-fold cross validation of 5 segments on the model to properly train the model for both these observations.

Subscription-Status

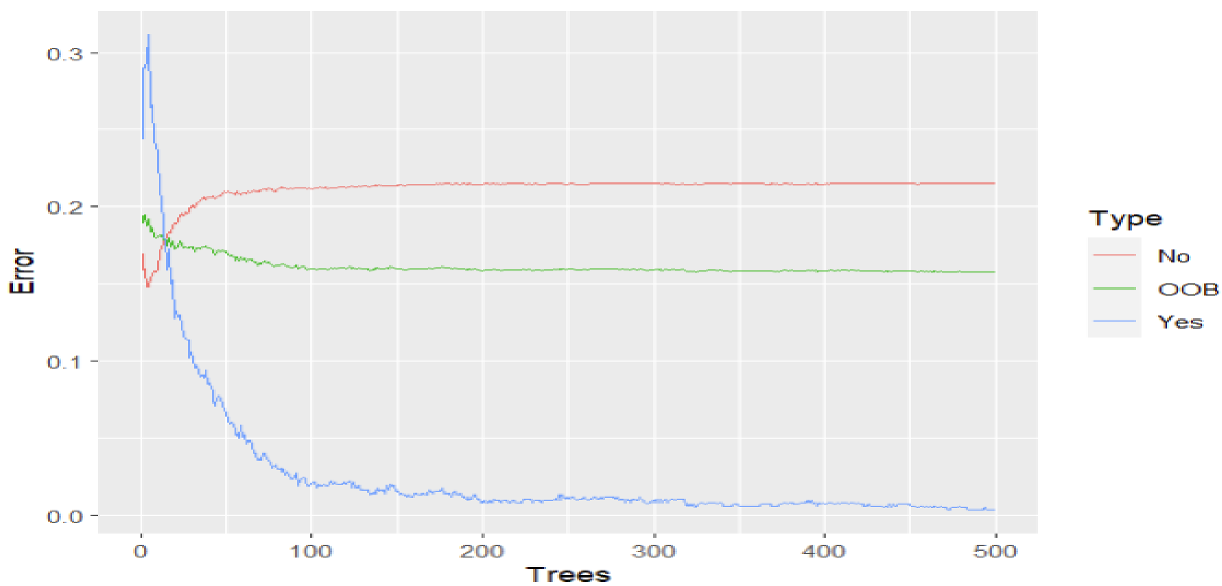
We were able to achieve accuracy of around 84% with random forest for predicting ‘Subscription Status’.

By default, the model used 500 trees.

```
Call:
 randomForest(formula = Subscription.Status ~ ., data = train_set,      mtry = 5)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 15.76%
Confusion matrix:
      No Yes class.error
No  1785 489 0.215039578
Yes    3 845 0.003537736
> 
```

We wanted to check if the model would still work fine if we used trees less or more than 500, so we tried building a graph to check how error varied for different values of 'ntree' and we found that the test error didn't significantly decrease after ntree= 200, so 200 trees were enough.



We built a model with ntree= 200 and found that the error rate was approximately the same, but we decided to keep the final model with mtry= 5 and ntree= 500 because the accuracy was still better with few decimal points.

```
Call:
 randomForest(formula = Subscription.Status ~ ., data = train_set,      mtry = 5, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 5

        OOB estimate of  error rate: 15.82%
Confusion matrix:
      No Yes class.error
No  1788 486 0.213720317
Yes    8 840 0.009433962
>
```

```
> importance(rf_subs2)
                         MeanDecreaseGini
Age                             28.731824
Gender                          55.799474
Item.Purchased                 101.081647
Category                         6.736897
Purchase.Amount..USD.           30.369095
Location                       169.530136
Size                            12.785550
Color                          105.339497
Season                          13.412954
Review.Rating                   24.460674
Payment.Method                  25.757230
Shipping.Type                   25.388804
Discount.Applied               255.773340
Promo.Code.Used                274.385182
Previous.Purchases              30.048249
Preferred.Payment.Method        25.563849
Frequency.of.Purchases          32.586839
>
```

Category

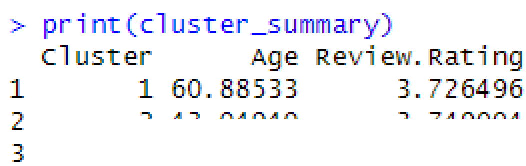
For category too we tried applying random forest and the error rate decreased to 19.02%, hence accuracy of 80.98%.

```
                     Number of trees: 10
No. of variables tried at each split: 5

        OOB estimate of  error rate: 19.02%
Confusion matrix:
            Accessories Clothing Footwear Outerwear class.error
Accessories         840       74       38        23   0.1384615
Clothing            117     1221       26        26   0.1215827
Footwear             69       72      326        14   0.3222453
Outerwear            57       53       20       121   0.5179283
> importance(rf_Cat)
```

K-Means:

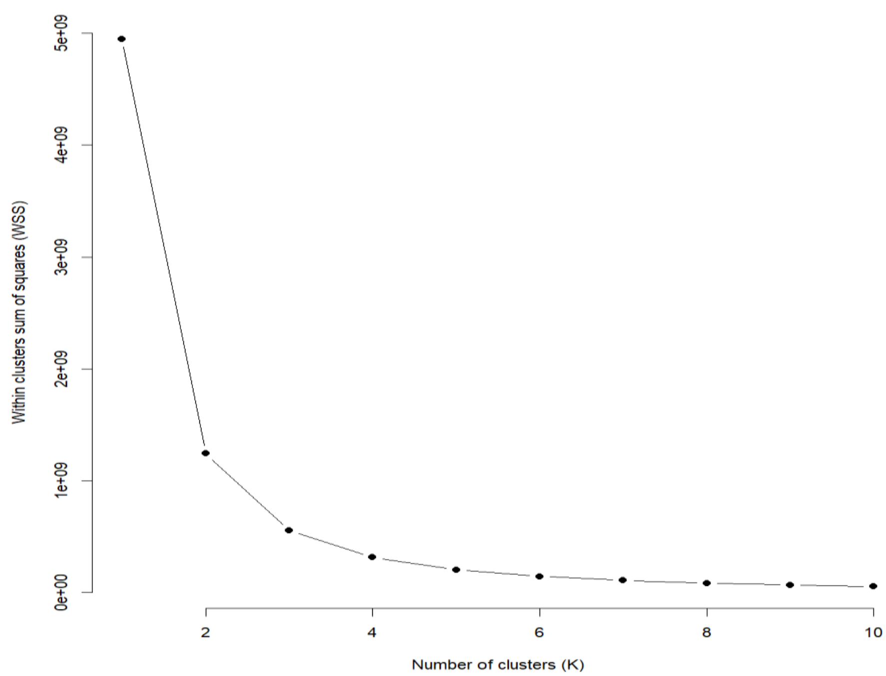
We explored the application of K-means clustering. The goal of this method is to create segments or clusters among the dataset based on similarities of the data points. Each data point is assigned to a K cluster where K is defined by a specific category in the data *(Education Ecosystem, 2018)*. With our dataset, we were able to cluster each shopper into 3 segments by 'Age' & approximate "Review Rating" they would give to a product which can be seen in figure 7. This gives us an average age for each category with the average customer review given to the business. Here we are able to see across the 3 age segments, shoppers give an average of 3.7 out of 5. This tells managers that across all age groups, rating levels are mediocre and should be improved by creating a better customer shopping experience by increased inventory levels and improved marketing tactics that are tailored to different customer segments.



```
> print(cluster_summary)
  Cluster      Age Review.Rating
1       1 60.88533      3.726496
2
3
```

K Means Clusters for Shoppers

As per the K Means clustering on 0-10 cluster sizes, cluster size 2 and 3 seemed to be the best cluster size with the lowest sum of squares (WSS) - after which point, increasing the cluster size would not lead to any significant reduction in WSS.



K Means Clusters (2 clusters) figure 8

	Customer.ID	Age	Gender	Item.Purchased	Category	Purchase.Amount..USD.	Location
1	2924.5	43.96974	0.64	11.91487	1.275897	60.20359	24.60308
2	974.5	44.16718	0.00	11.90256	1.271282	59.32513	24.28769

	Size	Color	Season	Review.Rating	Subscription.Status	Shipping.Type
1	1.428205	12.08103	1.517949	3.752462	1.00	2.485641
2	1.371795	11.88410	1.473846	3.747436	0.46	2.484103

	Discount.Applied	Promo.Code.Used	Previous.Purchases	Payment.Method
1	1.00	1.00	24.92051	2.488205
2	0.14	0.14	25.78256	2.471282

	Frequency.of.Purchases	AgeRange
1	3.061026	3.344615
2	3.005128	3.355385

The cluster insights (for 2 & 3 clusters) reflects how the average 'Age' is related to the 'Purchase Amount Spent' on Shopping ~ Average Age for the shopper is around 43-44 years, and they spent around $59-60 ~ so they can target this age group accordingly with their marketing strategies and have an idea of how much these shoppers will spend each time they visit a store.

K Means Clusters (3 clusters) figure 9

	Customer.ID	Age	Gender	Item.Purchased	Category	Purchase.Amount..USD.	Location	Size	Color	Season	Review.Rating
1	1948.5	43.89769	0.0000000	12.00077	1.282308	59.17846	24.51308	1.376154	12.09308	1.487692	3.773077
2	649.0	44.32333	0.0000000	11.90223	1.276366	59.86528	24.07390	1.403387	11.85681	1.480370	3.738029
3	3249.0	43.98463	0.9592621	11.82321	1.262106	60.24904	24.74865	1.420446	11.99769	1.519600	3.738739

	Subscription.Status	Shipping.Type	Discount.Applied	Promo.Code.Used	Previous.Purchases	Payment.Method	Frequency.of.Purchases	AgeRange
1	1.0000000	2.520000	0.7092308	0.7092308	25.30077	2.506923	3.038462	3.337692
2	0.1893764	2.470362	0.0000000	0.0000000	26.13395	2.494996	2.989992	3.371055
3	1.0000000	2.464258	1.0000000	1.0000000	24.62106	2.437356	3.070715	3.341276

Conclusion

In conclusion, our comprehensive analysis through descriptive and machine learning methods of customer shopping trends is able to provide a manager of a retail store with valuable insights in order to help improve shopper experience. We have put together an **"Shopper Persona"** which depicts what we have learned through our exploratory analysis.

Shopper Persona

"A typical shopper is a person from Montana & California (in the US) of around 30-45 years of age, who prefers shopping the most in Fall & Spring, with Clothing in size 'M' & Accessory being the top categories they shop for, & PayPal & Cash being the preferred medium they use for paying".

Through machine learning, we are able to determine -

- Predicting subscription status is dependent on discount applied and promo codes available at the time of purchase.
- The category of an item that a customer is going to buy is depending on the color of an item and the location of where it is being sold.
- The average rating based on different customer age groups being 3.7/5 across all segments
- The average purchase amounts by age groups in clusters of 1-2 being $60 & 1-3 $59.

References

Customer Shopping Trends Dataset. (n.d.). Kaggle. Retrieved 2023, from https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data

Education Ecosystem. (2018, September 12). *Understanding K-means Clustering in Machine Learning | by Education Ecosystem (LEDU)*. Towards Data Science. Retrieved December 9, 2023, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1