Course Code: MGT7215

Marketing Analytics – Assignment 1

Muhammad Muneeb Ullah Ansari - 40426685

Word Count : 2012

Introduction

Research in customer heterogeneity has explored various dimensions and implications of this phenomenon across different industries and contexts. One of the key areas of focus has been on segmentation strategies to identify distinct customer groups based on shared characteristics or behaviors. Bertrand and Mullainathan (2001) highlighted the importance of segmentation in targeting and personalization efforts, emphasizing the need to consider both observable and unobservable dimensions of heterogeneity.

Moreover, advances in data analytics and machine learning have enabled researchers to uncover nuanced patterns of heterogeneity within customer populations. Breiman (2001) introduced the concept of ensemble learning techniques, such as random forests and gradient boosting, which have proven effective in capturing complex interactions and nonlinear relationships among customer attributes.

KMeans clustering, proposed by MacQueen (1967), is a popular unsupervised learning algorithm used for partitioning data into distinct groups based on similarity. In their study, Jain and Dubes (1988) provided an in-depth overview of clustering algorithms, highlighting the effectiveness of KMeans in handling large datasets and its computational efficiency. However, researchers have noted the sensitivity of KMeans to the initial selection of cluster centroids (Arthur & Vassilvitskii, 2007), leading to the development of improved initialization techniques such as KMeans++ (Arthur & Vassilvitskii, 2007).

Gupta (2013) demonstrated the applicability of RFM analysis in predicting customer lifetime value and guiding targeted marketing strategies. Moreover, recent advancements in data mining techniques have enabled the integration of RFM with machine learning algorithms, enhancing its predictive capabilities (Han et al., 2019)

LDA has been widely adopted in various domains, including image recognition (Duda et al., 2000) and bioinformatics (Chang & Lin, 2011). Despite its simplicity, LDA offers several advantages, such as dimensionality reduction and robustness to multicollinearity (Sharma & Paliwal, 2015). However, researchers have also highlighted its limitations, particularly in cases where class separability is low or when dealing with imbalanced datasets (Zhang et al., 2019).

Methodology

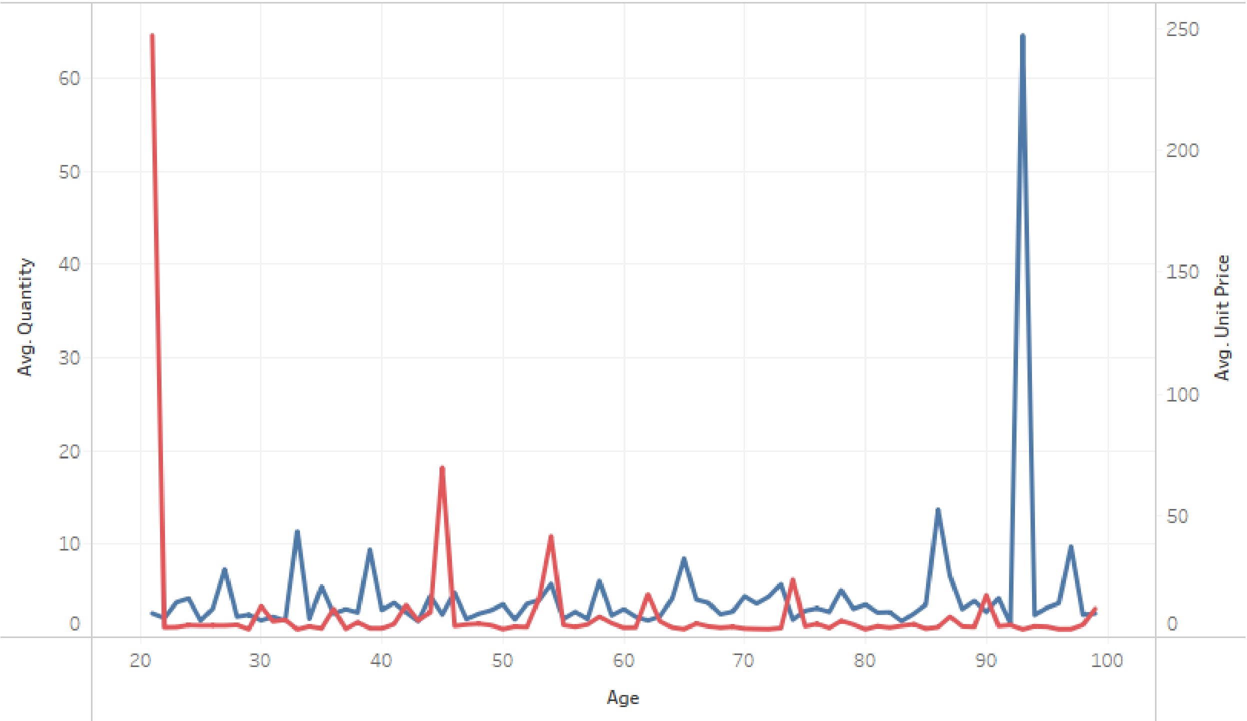
We start by loading the dataset into Tableau to perform an initial exploratory analysis of the data. We observe that the data had missing values and redundancy of observations in CustomerID. After bringing the dataset to the R environment, we perform hierarchal clustering by each of the 4 centroid linkage methods: Complete, Single, Average and Centroid. This became the basis to form our elbow plot to identify optimal value of K for K-means clustering. After forming relevant segments, we performed a second segmentation with dataset which contained no missing values for CustomerID. We then imported the results of the first segmentation to Tableau to display relevant traits of each segment formed. Following that an LDA Analysis was performed to associate relevant basis variables to each segment and create a model to predict customers into each segment. Finally an RFM analysis was performed where the dataset with no missing values was used to assign RFM scores to each customer by both independent and sequential methods.

Results and Discussion

Exploratory Analysis

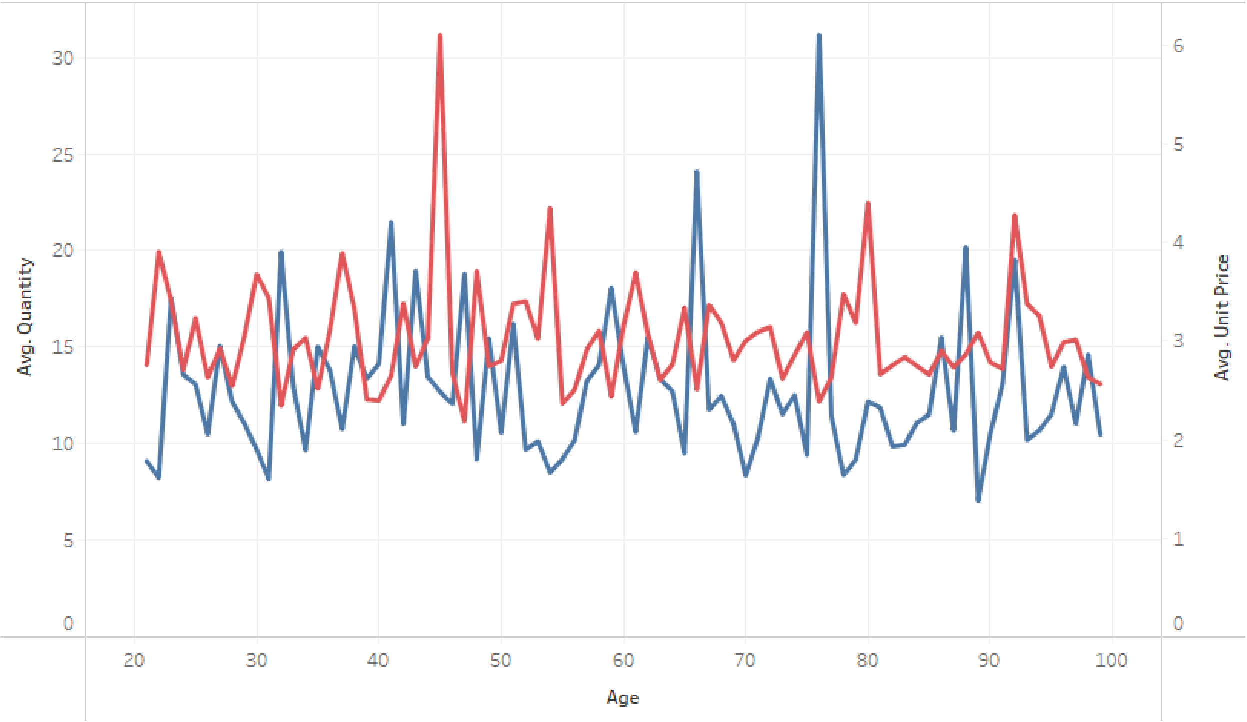
Missing values of CustomerID has very strong business implication. It means that those customers did not form an account of the e-commerce platform before making a purchase. Hence we initially look at our customers as those who did make an account for purchase and those who did not. For both these customers, analysis revealed that the only difference in their purchase patterns was in quantity and unit price. Customers who did not make an account (observations with missing values for CustomerID) revealed a lower purchase quantity but higher unit price. There was more outlier behavior for non-registered customers.

Age with Unit Price and Quantity for Non-registered Customers



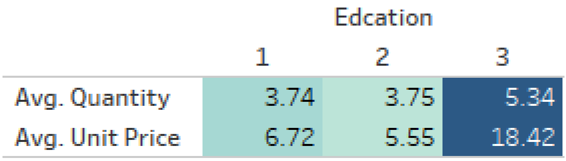
The trends of Avg. Quantity and Avg. Unit Price for Age. Color shows details about Avg. Quantity and Avg. Unit Price. The data is filtered on Customer ID, which keeps NA.

Age with Unit Price and Quantity for Registered Customers



The trends of Avg. Quantity and Avg. Unit Price for Age. Color shows details about Avg. Quantity and Avg. Unit Price. The data is filtered on Customer ID, which excludes NA.

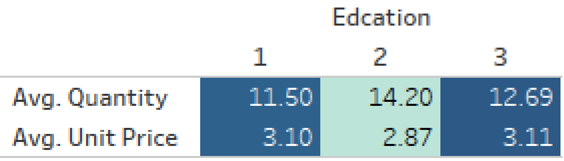
Education with Unit Price and Quantity for Non-registered Customers



	Edcation		
	1	2	3
Avg. Quantity	3.74	3.75	5.34
Avg. Unit Price	6.72	5.55	18.42

Avg. Quantity and Avg. Unit Price broken down by Edcation. Color shows Avg. Unit Price. The marks are labeled by Avg. Quantity and Avg. Unit Price. The data is filtered on Customer ID, which keeps NA.

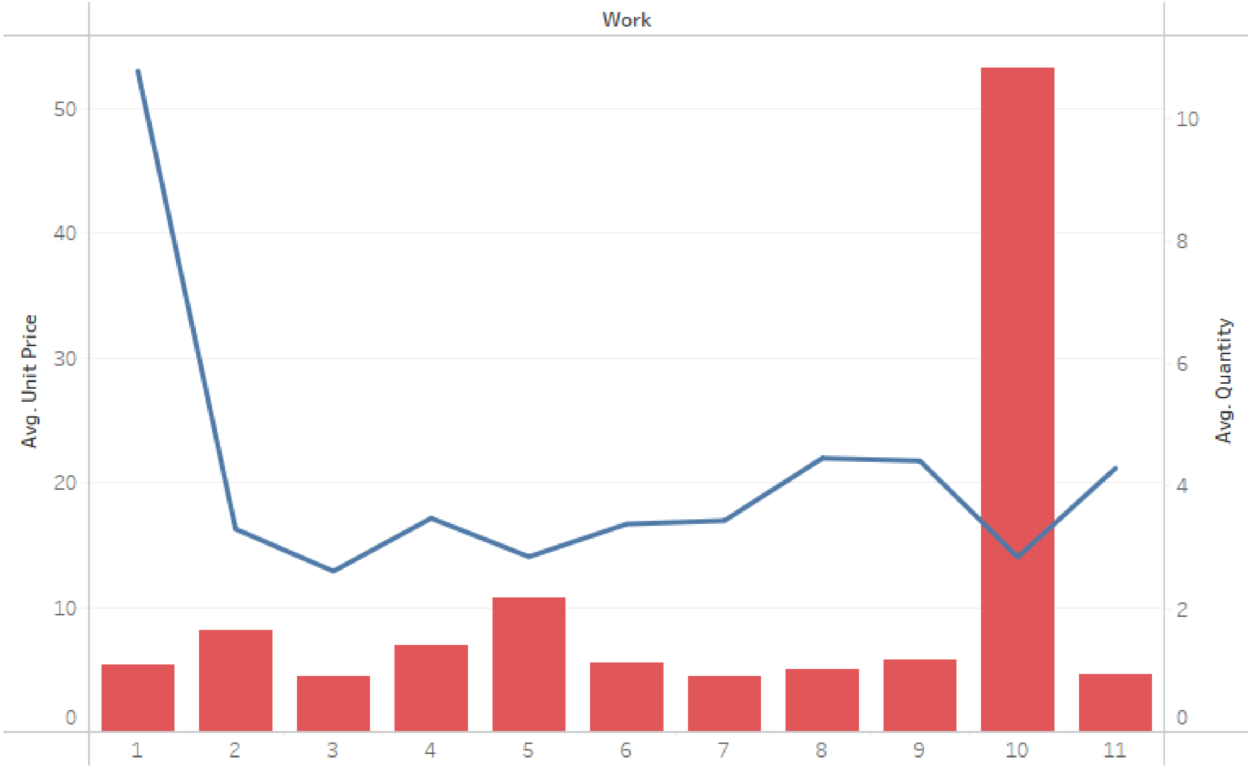
Education with Unit Price and Quantity for Registered Customers



	Edcation		
	1	2	3
Avg. Quantity	11.50	14.20	12.69
Avg. Unit Price	3.10	2.87	3.11

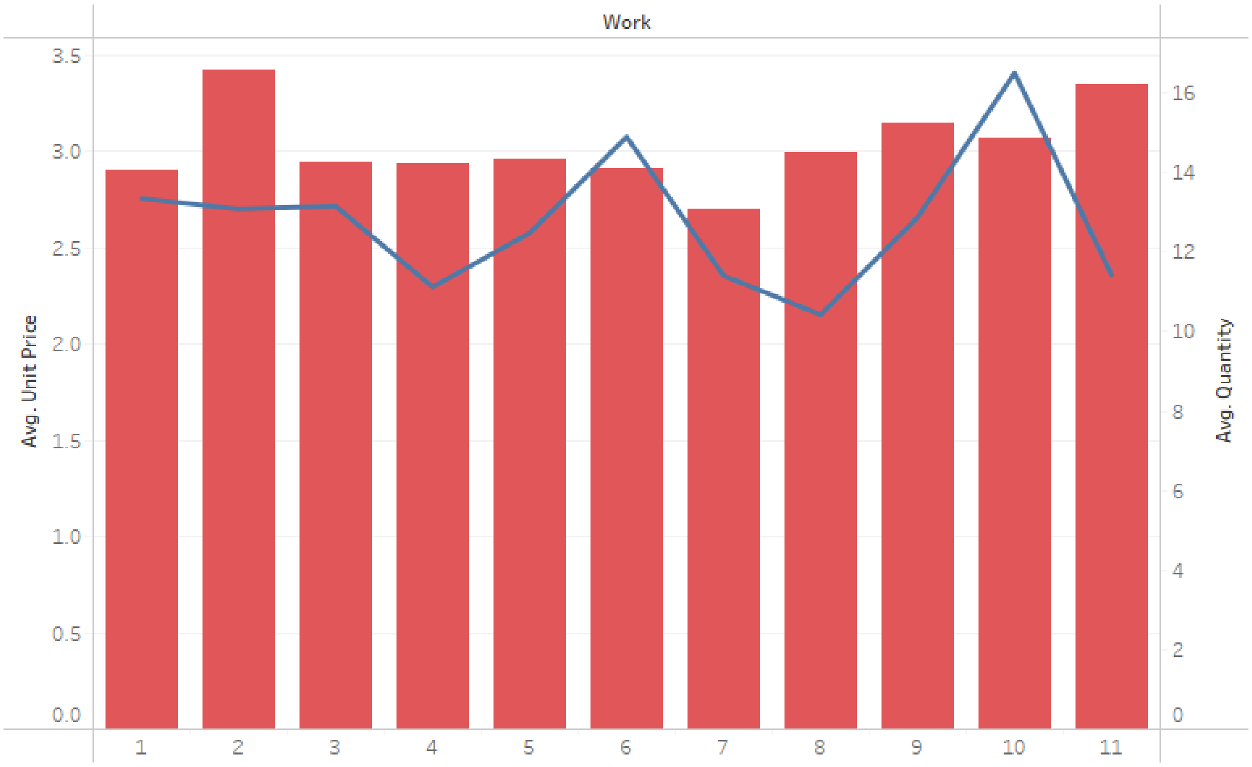
Avg. Quantity and Avg. Unit Price broken down by Edcation. Color shows Avg. Unit Price. The marks are labeled by Avg. Quantity and Avg. Unit Price. The data is filtered on Customer ID, which excludes NA.

Work by Unit Price and Quantity for Non-registered Customers

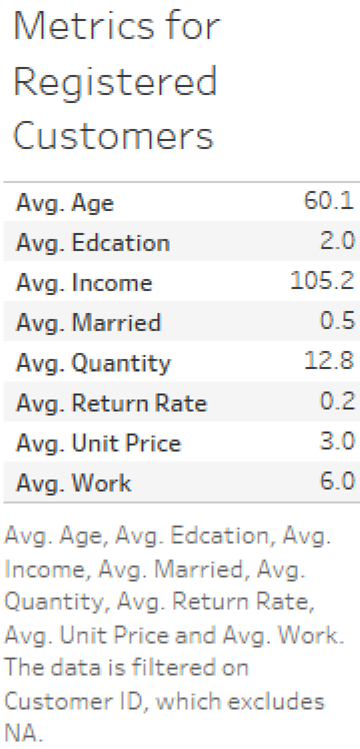


The trends of Avg. Unit Price and Avg. Quantity for Work. Color shows details about Avg. Unit Price and Avg. Quantity. The data is filtered on Customer ID, which keeps NA.

Work by Unit Price and Quantity for Registered Customers



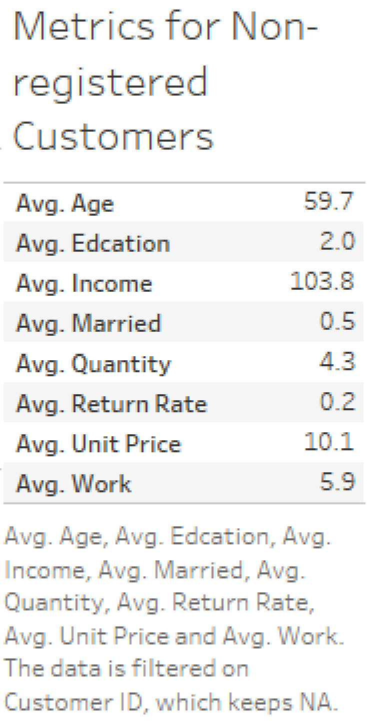
The trends of Avg. Unit Price and Avg. Quantity for Work. Color shows details about Avg. Unit Price and Avg. Quantity. The data is filtered on Customer ID, which excludes NA.



Metrics for Registered Customers

Avg. Age	60.1
Avg. Edcation	2.0
Avg. Income	105.2
Avg. Married	0.5
Avg. Quantity	12.8
Avg. Return Rate	0.2
Avg. Unit Price	3.0
Avg. Work	6.0

Avg. Age, Avg. Edcation, Avg. Income, Avg. Married, Avg. Quantity, Avg. Return Rate, Avg. Unit Price and Avg. Work. The data is filtered on Customer ID, which excludes NA.

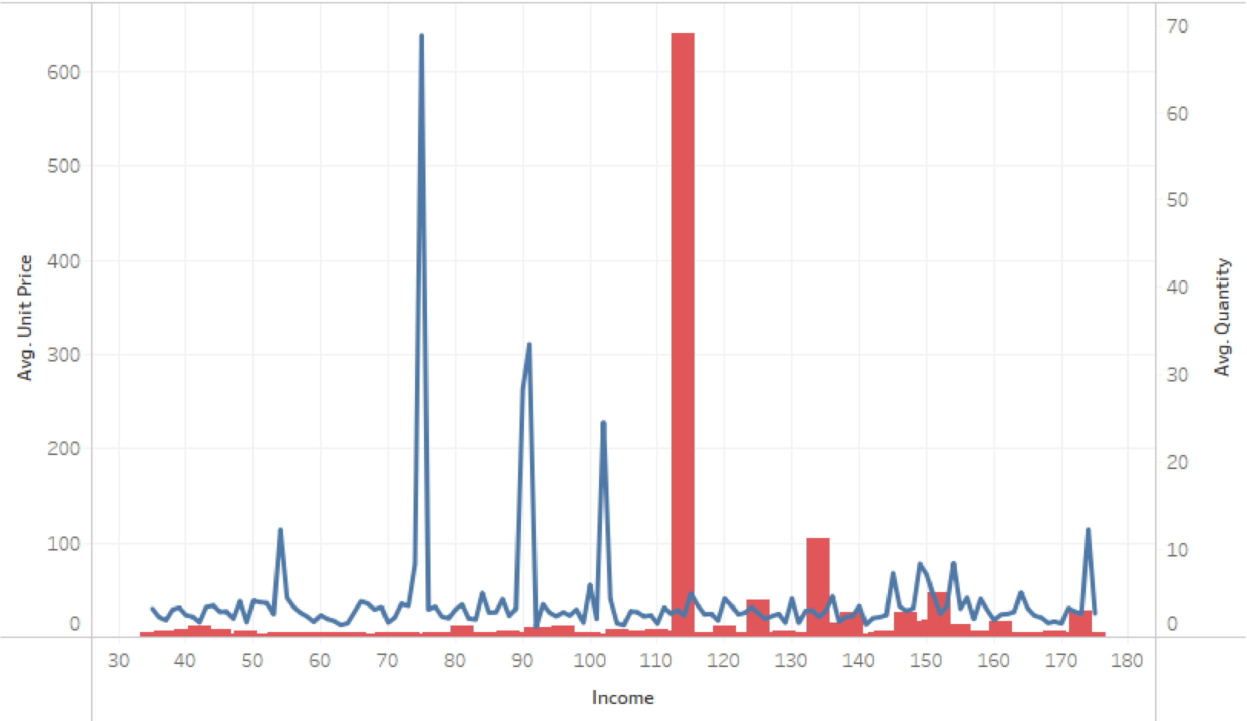


Metrics for Non-registered Customers

Avg. Age	59.7
Avg. Edcation	2.0
Avg. Income	103.8
Avg. Married	0.5
Avg. Quantity	4.3
Avg. Return Rate	0.2
Avg. Unit Price	10.1
Avg. Work	5.9

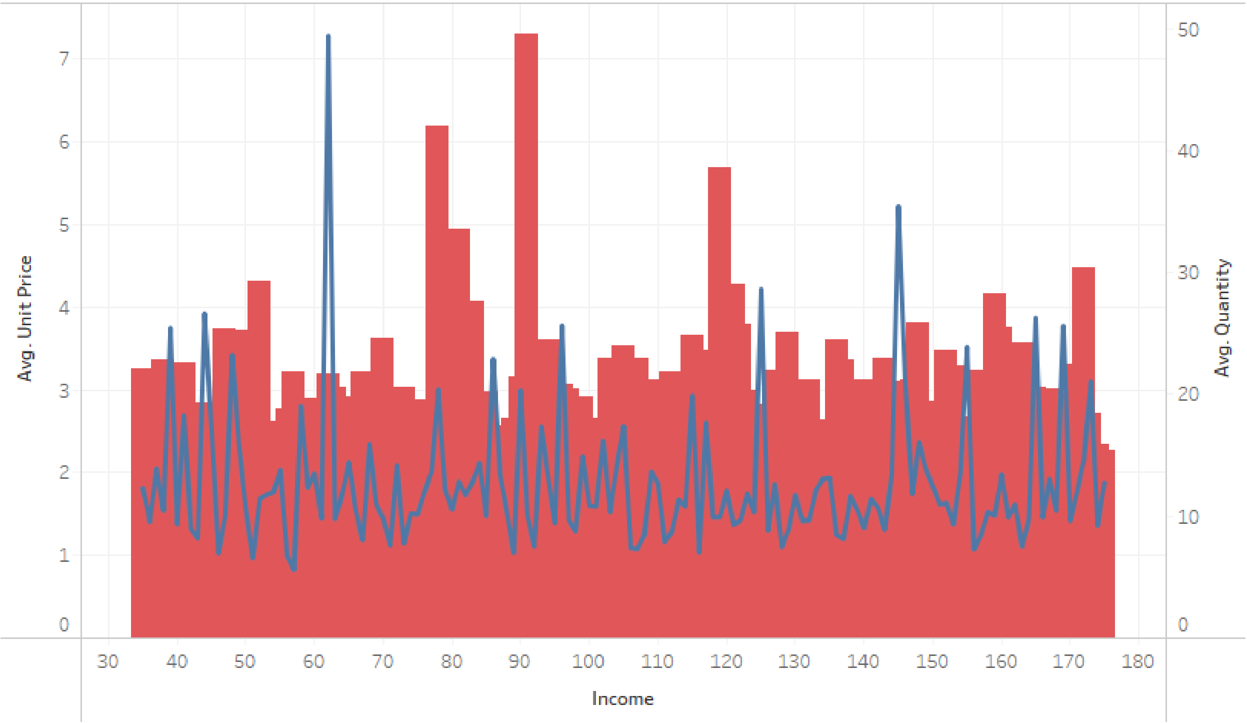
Avg. Age, Avg. Edcation, Avg. Income, Avg. Married, Avg. Quantity, Avg. Return Rate, Avg. Unit Price and Avg. Work. The data is filtered on Customer ID, which keeps NA.

Income with Unit Price and Quantity for Non-registered Customers

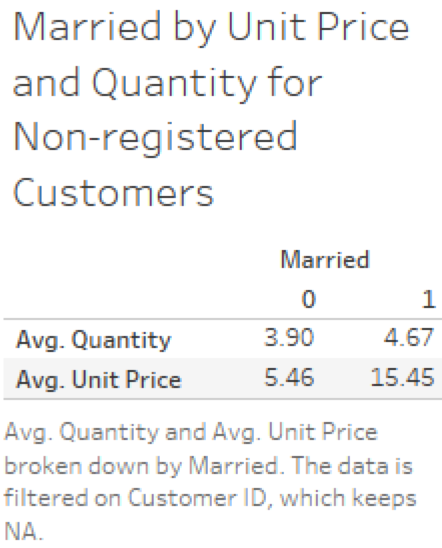


The trends of Avg. Unit Price and Avg. Quantity for Income. Color shows details about Avg. Unit Price and Avg. Quantity. The data is filtered on Customer ID, which keeps NA.

Income with Unit Price and Quantity for Registered Customers



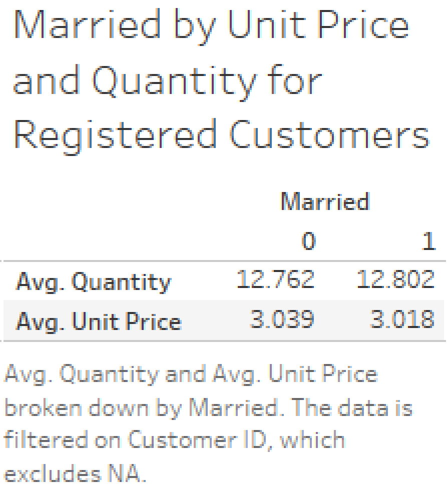
The trends of Avg. Unit Price and Avg. Quantity for Income. Color shows details about Avg. Unit Price and Avg. Quantity. The data is filtered on Customer ID, which excludes NA.



Married by Unit Price and Quantity for Non-registered Customers

	Married	
	0	1
Avg. Quantity	3.90	4.67
Avg. Unit Price	5.46	15.45

Avg. Quantity and Avg. Unit Price broken down by Married. The data is filtered on Customer ID, which keeps NA.



Married by Unit Price and Quantity for Registered Customers

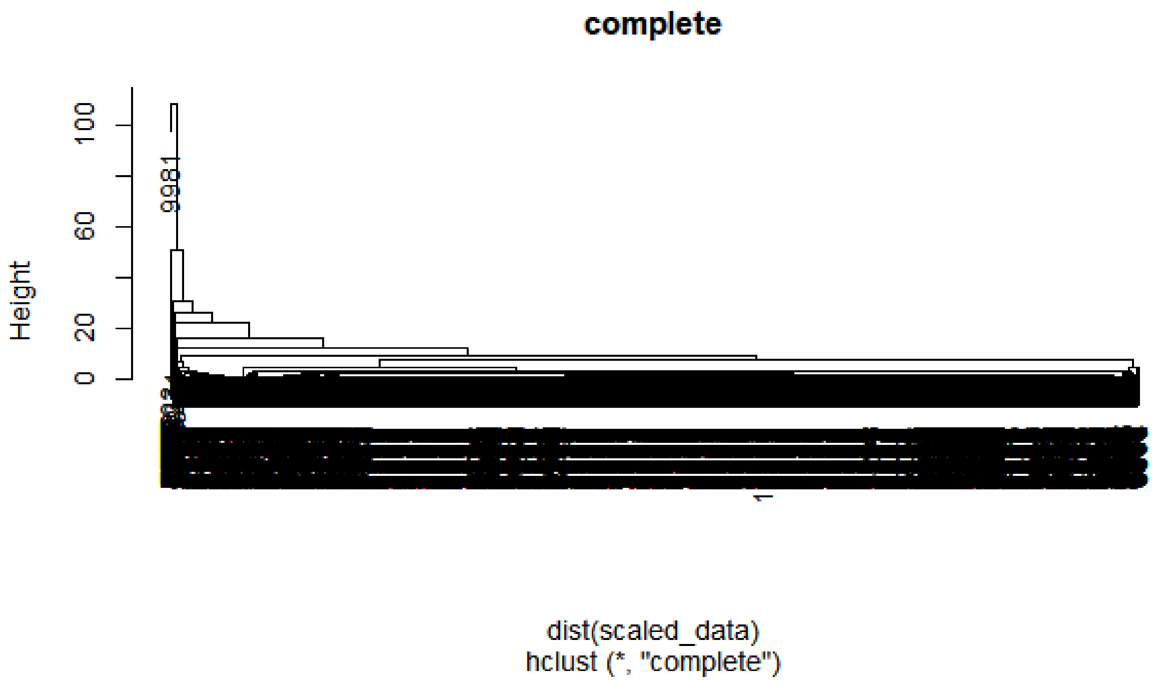
	Married	
	0	1
Avg. Quantity	12.762	12.802
Avg. Unit Price	3.039	3.018

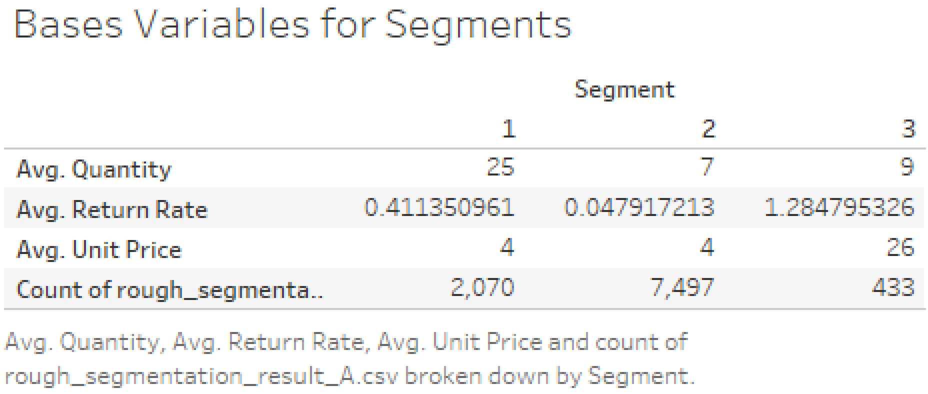
Avg. Quantity and Avg. Unit Price broken down by Married. The data is filtered on Customer ID, which excludes NA.

This mean that observations with values for CustomerID were regular registered customers who made purchases in higher quantity on average but at a lesser price. Hence this gives root to the possibility of different segments for these two different types of customers. The analysis also revealed redundancy in CustomerID where same CustomerIDs had repeated observations with different descriptor variables. This means that for customers that were registered, more than one person made a purchase through that registered account.

Segmentation

We decide to perform segmentation twice because of the possibility of different behaviors for each kind of customers. First for all the customers (including registered and non-registered) then for only registered customers. We decide not to remove redundancy in CustomerID because even if more than one person made a purchase from a specific CustomerID, it still gives us an indication of what kind of person would find the e-commerce website attractive. If we decided to reduce a CustomerID to only one customer, that would not give us a complete understanding of the people interested in the platform. It would also be difficult to determine which person should be associated to the CustomerID as this would require very strong business understanding and context which we lack in this assignment. We performed hierarchal clustering by each of the 4 centroid linkage method and opted to choose complete linkage to proceed further as clusters were more likely to have varying densities and irregular shape due to the presence of outliers in the complete dataset.





Bases Variables for Segments

	Segment		
	1	2	3
Avg. Quantity	25	7	9
Avg. Return Rate	0.411350961	0.047917213	1.284795326
Avg. Unit Price	4	4	26
Count of rough_segmenta..	2,070	7,497	433

Avg. Quantity, Avg. Return Rate, Avg. Unit Price and count of rough_segmentation_result_A.csv broken down by Segment.

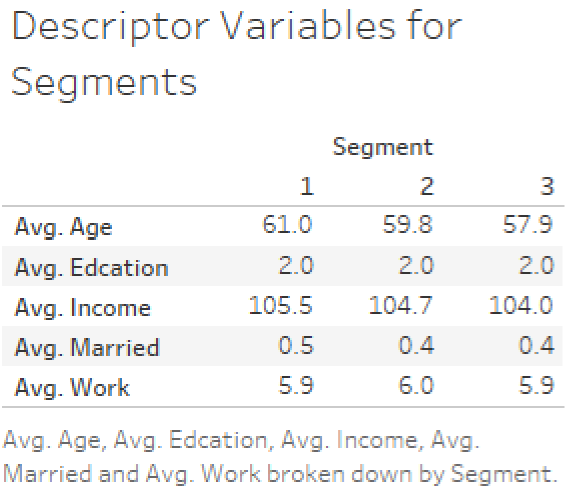
Performing segmentation on every customer revealed 3 segments. We can broadly classify each segment as follows:

Segment 1, Value Shoppers: High Quantity, Average Returns, Low Price (Minority)

Segment 2, Unattractive Shoppers: Low Quantity, Low Returns, Low Price (Majority)

Segment 3, Premium Shoppers: Low Quantity, High Returns, High Price (Barely existing)

We have a majority of customers who are unattractive shoppers. Subject to business context, it appears that the e-commerce website might not be performing well. If the business platform is positioned a known certain way, we can tell if Segment 1 or Segment 3 would be viable to target. As for the results, it would seem better to target segment 1, Value Shoppers, as they still consist of a recognizable chunk of the platform's customers. Hence offers, deals and promotions on the website should be focusing on value, pushing out more quantities for lesser price for growing this segment. This is also subject to business context and the positioning that the platform already has.



Descriptor Variables for Segments

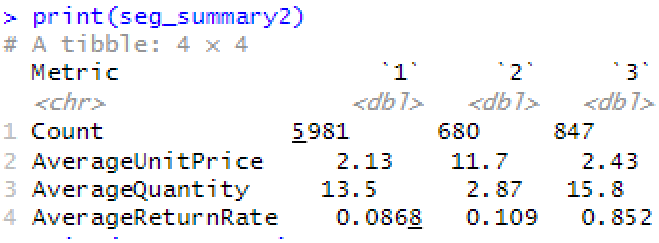
	Segment		
	1	2	3
Avg. Age	61.0	59.8	57.9
Avg. Edcation	2.0	2.0	2.0
Avg. Income	105.5	104.7	104.0
Avg. Married	0.5	0.4	0.4
Avg. Work	5.9	6.0	5.9

Avg. Age, Avg. Edcation, Avg. Income, Avg. Married and Avg. Work broken down by Segment.

As for the descriptor variables of each segment, there seems to be no prominent distinction for each segment. This issue will be addressed while performing LDA.

When the same clustering procedure was applied for dataset with CustomerID as non-NAs, the results of the segments were a bit similar. Similarities were Segment 2 barely existing as premium shoppers, Segment 1 majority of value shoppers. The only difference was of segment 3 which we can call as value shoppers but with higher return rate. So essentially Segment 1 and 3 are very similar in this case.

This means that both attempts of segmentation only vary by unattractive shoppers present in the whole dataset.

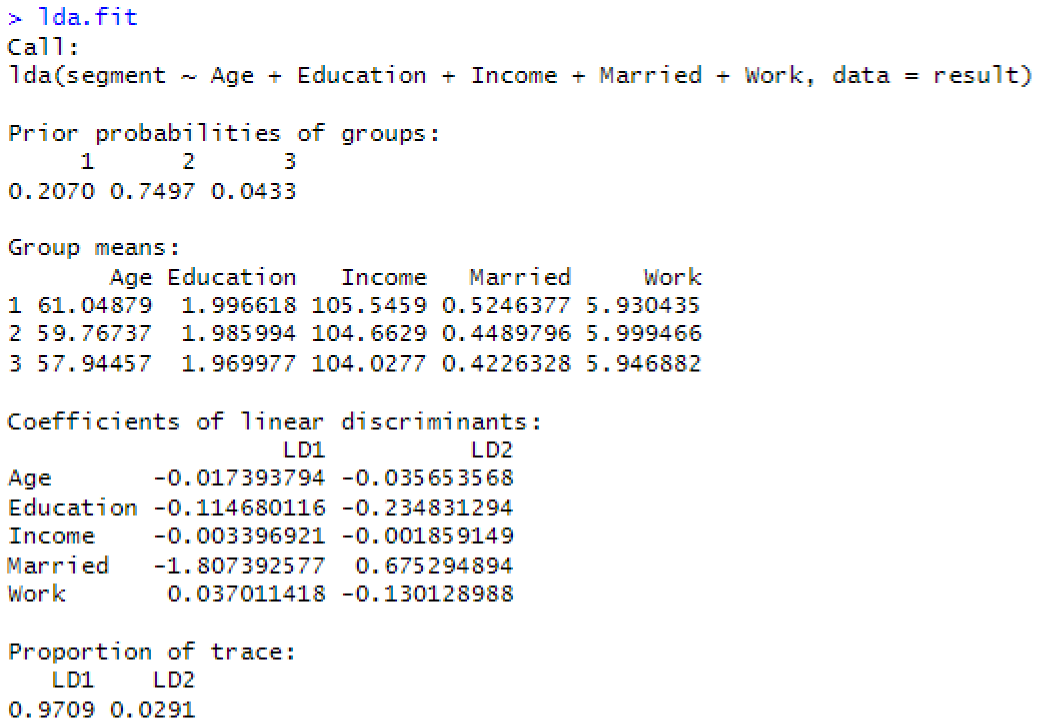


```
> print(seg_summary2)
# A tibble: 4 × 4
  Metric                 `1`     `2`     `3`
  <chr>                <dbl>   <dbl>   <dbl>
1 Count             5981     680     847
2 AverageUnitPrice     2.13   11.7     2.43
3 AverageQuantity     13.5     2.87   15.8
4 AverageReturnRate    0.0868  0.109   0.852
```

Because the complete dataset gives us a more holistic understanding of the totality of customers on the platform, we will consider the first segmentation result as a basis for our segmentation.

Linear Discriminant Analysis and ANNOVA

For LDA, we will consider the complete dataset because we want it to be consistent with the segmentation that we performed.



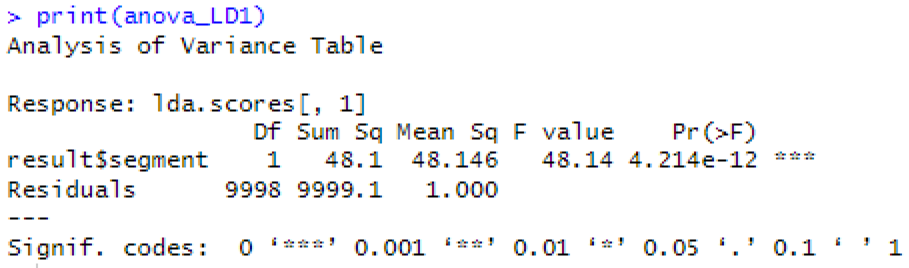
```
> lda.fit
Call:
lda(segment ~ Age + Education + Income + Married + Work, data = result)

Prior probabilities of groups:
     1      2      3 
0.2070 0.7497 0.0433 

Group means:
       Age Education   Income   Married     Work
1 61.04879  1.996618 105.5459 0.5246377 5.930435
2 59.76737  1.985994 104.6629 0.4489796 5.999466
3 57.94457  1.969977 104.0277 0.4226328 5.946882

Coefficients of linear discriminants:
                   LD1          LD2
Age       -0.017393794 -0.035653568
Education -0.114680116 -0.234831294
Income    -0.003396921 -0.001859149
Married   -1.807392577  0.675294894
Work       0.037011418 -0.130128988

Proportion of trace:
   LD1    LD2 
0.9709 0.0291 
```

3 bases variables resulted in 2 LDA equations to form where only one was significant with a p-value of less than 0.05. The prior probabilities indicate the proportion of each segment within the dataset:

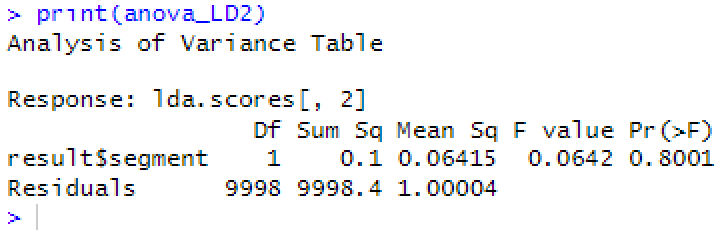
- Segment 1: 20.70%
- Segment 2: 74.97%
- Segment 3: 4.33%

The group means provide the average values for each predictor variable within each segment.

- Age: Segment 3 is slightly younger than the others.
- Education: Education level is fairly similar across segments, with slight variations.
- Income: Similar across segments, with Segment 1 having the highest average income.
- Married: Segment 1 has the highest proportion of married individuals.
- Work: Work hours are relatively consistent across segments.

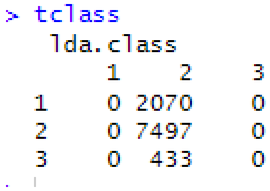


```
> print(anova_LD1)
Analysis of Variance Table

Response: lda.scores[, 1]
                 Df Sum Sq Mean Sq F value    Pr(>F)
result$segment    1   48.1  48.146   48.14 4.214e-12 ***
Residuals      9998 9999.1   1.000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



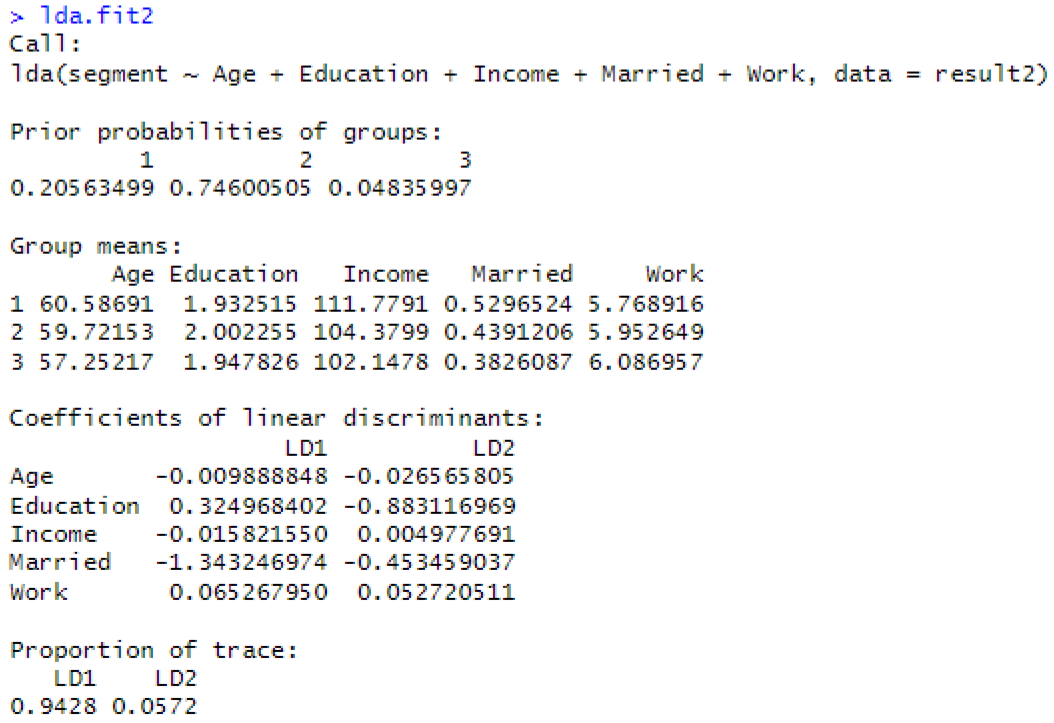
```
> print(anova_LD2)
Analysis of Variance Table

Response: lda.scores[, 2]
                 Df Sum Sq Mean Sq F value Pr(>F)
result$segment    1    0.1 0.06415  0.0642 0.8001
Residuals      9998 9998.4 1.00004
>
```

Coefficients of LD1 indicate that it is primarily influenced by “Married” and “Education” indicating that these segments most effectively differentiate between the segments. This is the equation that is statistically significant. Based on the analysis, LD1 provides significant discrimination between the segments, driven primarily by marital status and education level. Segment 2, being the largest, does not appear to be significantly different from the others based on LD2.



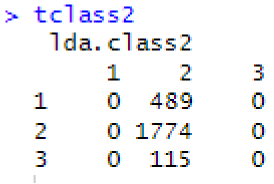
```
> tclass
   lda.class
       1    2    3
  1    0 2070    0
  2    0 7497    0
  3    0  433    0
```

After training the model and making predictions, it was apparent that the model was only assigning observations to the segment with the highest observations, which is segment 2. This is due to the nature of the data and the high weightage associated with segment 2. In conclusions, we will discuss how this problem could have been solved by advanced analytical techniques however that is out of the scope of this assignment.



```
> lda.fit2
Call:
lda(segment ~ Age + Education + Income + Married + Work, data = result2)

Prior probabilities of groups:
         1          2          3 
0.20563499 0.74600505 0.04835997 

Group means:
       Age Education   Income   Married     Work
1 60.58691  1.932515 111.7791 0.5296524 5.768916
2 59.72153  2.002255 104.3799 0.4391206 5.952649
3 57.25217  1.947826 102.1478 0.3826087 6.086957

Coefficients of linear discriminants:
                   LD1          LD2
Age       -0.009888848 -0.026565805
Education  0.324968402 -0.883116969
Income    -0.015821550  0.004977691
Married   -1.343246974 -0.453459037
Work       0.065267950  0.052720511

Proportion of trace:
   LD1    LD2 
0.9428 0.0572 
```

We also performed LDA with the dataset after removing redundancy in order to observe whether this biased prediction was being induced due to the presence of duplicated values. However resulting LDA model gave a similar biased result.

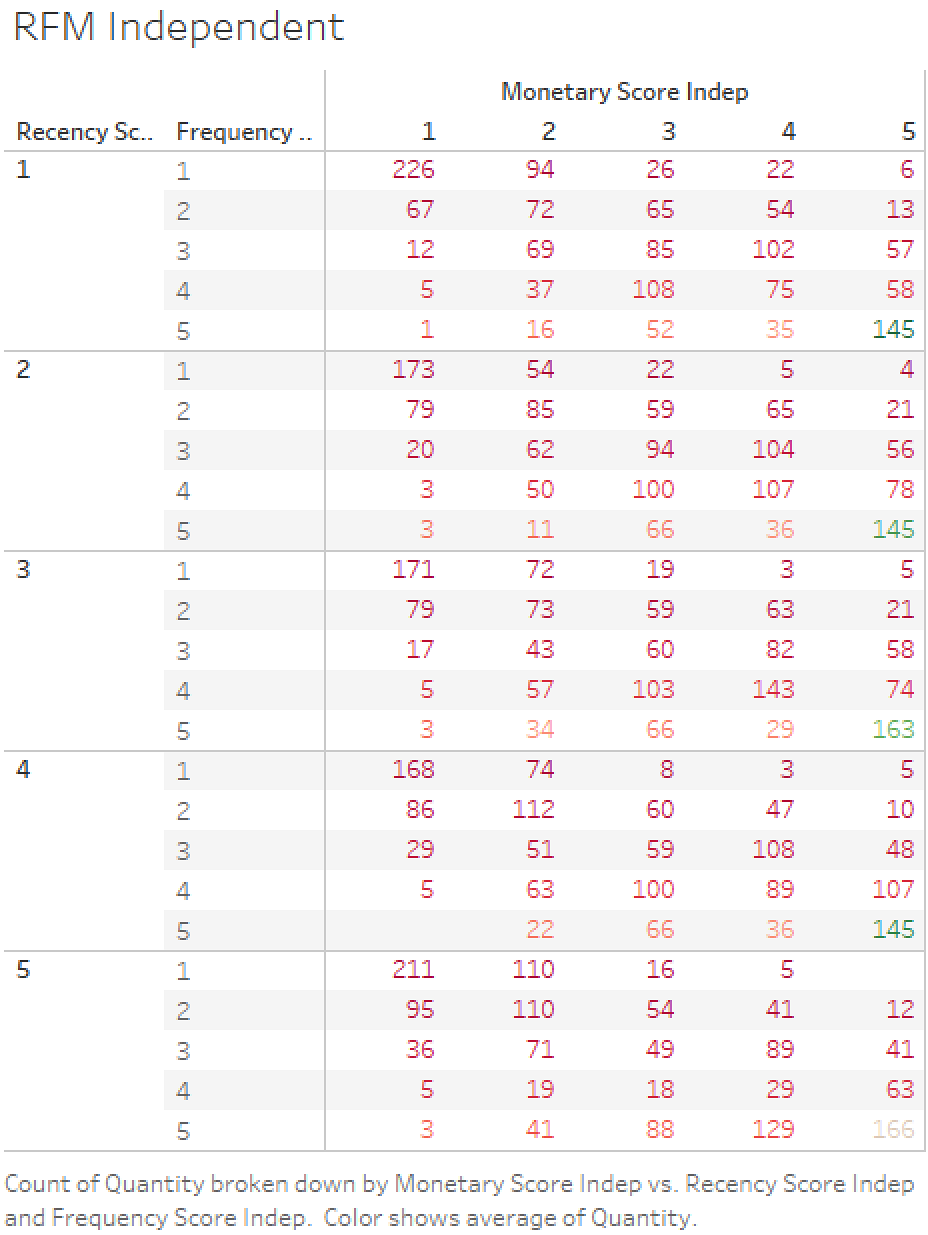


```
> tclass2
   lda.class2
       1    2    3
  1    0  489    0
  2    0 1774    0
  3    0  115    0
```

RFM Analysis

For an RFM Analysis, it is essential that we have a dataset of non-NA CustomerID unlike segmentation. That is because results of RFM are used to target individual specific accounts of customers (with attractive scores) which therefor need to be identified by a unique identification example CustomerID. That is not the case with segmentation as targeting a segment is not on the basis of a unique identifier like CustomerID, it is on the basis of widely generalizable descriptor variables of characteristics like age, income, work, marital status etc.

We opted to go for Independent RFM rather than Sequential as there is no clear order of importance among the dimensions as well as we want to capture every different aspect of customer behavior.



RFM Independent

		Monetary Score Indep				
Recency Sc..	Frequency ..	1	2	3	4	5
1	1	226	94	26	22	6
	2	67	72	65	54	13
	3	12	69	85	102	57
	4	5	37	108	75	58
	5	1	16	52	35	145
2	1	173	54	22	5	4
	2	79	85	59	65	21
	3	20	62	94	104	56
	4	3	50	100	107	78
	5	3	11	66	36	145
3	1	171	72	19	3	5
	2	79	73	59	63	21
	3	17	43	60	82	58
	4	5	57	103	143	74
	5	3	34	66	29	163
4	1	168	74	8	3	5
	2	86	112	60	47	10
	3	29	51	59	108	48
	4	5	63	100	89	107
	5		22	66	36	145
5	1	211	110	16	5	
	2	95	110	54	41	12
	3	36	71	49	89	41
	4	5	19	18	29	63
	5	3	41	88	129	166

Count of Quantity broken down by Monetary Score Indep vs. Recency Score Indep and Frequency Score Indep. Color shows average of Quantity.

For our RFM, the following criteria can be established:

- Best Customers: Those who score high on all three criteria (R=5, F=5, M=5), indicating they are recent, frequent, and high-spending. In the provided table, there are 762 such customers.

- Loyal Customers: Customers with high frequency (F=4 or 5), indicating regular purchases. They might vary in recency and monetary scores.

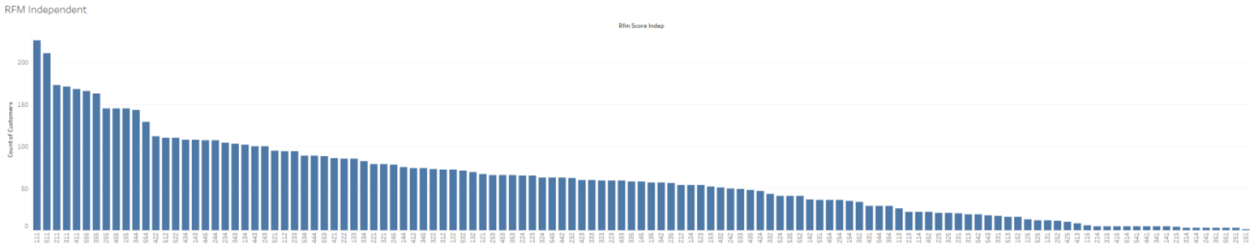
- Big Spenders: Customers with high monetary scores (M=4 or 5), indicating they spend a lot even if they do not purchase often.

- At-Risk Customers: Customers with high scores historically but low recency (R=1 or 2), indicating they used to shop frequently and spend but haven't done so recently. It's crucial to re-engage these customers.

- Lost Cheap Customers: Those with low on all three criteria (R=1, F=1, M=1), indicating they are not recent, not frequent, and low spenders. They are not usually a priority for marketing efforts.

By segmenting customers using RFM analysis, the company can tailor its marketing strategies to different groups, potentially increasing customer retention and overall profitability. For instance:

- Best Customers can be rewarded to reinforce their behavior.
- Loyal Customers might be encouraged with loyalty programs.
- Big Spenders could be targeted with upsells or premium offers.
- At-Risk Customers may need re-engagement campaigns.
- Lost Cheap Customers might be targeted selectively based on cost-benefit analysis.



Count of rfm_result.csv for each Rfm Score Indep.

Focusing on the count of customers in each RFM group in our dataset, We can see the top 5 groups are for the poorest Frequency and Monetary score. The highest count is for 111 which is alarming, followed by 511, 211, 311 and 411. Neither of these are attractive segments so we won't spend any marketing efforts on them. The next 5 counts are 555, 355, 255, 455 and 155 respectively. It would make more sense to direct marketing communications to them to improve their recency scores, it seems that profitable customers might have forgotten the e-commerce website and hence the website needs to stay relevant for them. The table view reveals that Customers in almost every group are associated with below average purchase quantities. This is in accordance with outliers of quantities in data which we cannot ignore.

Conclusions

Majority of this interpretation is subject to proper business context and understanding which we lack for this assignment example how the website is already positioned. Only drawback of this flow of analysis is the poor predictive accuracy of the LDA model which is due to the nature of the data. However methods beyond the scope of this assignment can leverage equal sampling techniques, bootstrapping, and cross-validation to improve the predictive accuracy of the model. Quadratic Discriminant Analysis (QDA) could be used as well if you left the assumption that variables are linearly related.

References

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. American Economic Review, 91(2), 67-72.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.

Gupta, S. (2013). Customer lifetime value: A comprehensive literature review. Journal of Marketing Theory and Practice, 21(4), 403-438.

Han, S., Jang, S. S., & Kim, J. H. (2019). Enhancing RFM analysis using deep learning: A case study of hotel business. International Journal of Hospitality Management, 82, 154-165.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern classification (2nd ed.). Wiley.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

Sharma, A., & Paliwal, K. K. (2015). Linear discriminant analysis: A detailed tutorial. AI Communications, 28(3), 245-258.

Zhang, S., Li, X., & Hu, Q. (2019). Linear discriminant analysis in the presence of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 33(2), 1958026.

Appendix

```
#Load Library
library(dplyr)
library(tidyr)
library(MASS)

#Load Data
data <- read.csv(file.choose())

#Exploring
str(data)
summary(data)
sum(is.na(data))
summary(as.factor(data$CustomerID))
#there are many NAs and redundancy in CustomerID

length(unique(data$CustomerID))
summary(data[data$CustomerID == 17841, ])
#tells us that there is redundancy in CustomerID

data_1 <- data[is.na(data$CustomerID), ]
data_2 <- data[complete.cases(data$CustomerID), ]
#Creating 2 datasets for all customers and registered customers

summary(data_1)
summary(data_2)

#################
#CLUSTER ANALYSIS#
```

```
#################

#Cluster Analysis with all oversations

## Load Packages and Set Seed
set.seed(1)

# Import Data
seg <- data

# Data Preprocessing: Scale the variables
scaled_data <- scale(cbind(seg$Quantity, seg$ReturnRate, seg$UnitPrice))

# Run hierarchical clustering
seg_hclust_complete <- hclust(dist(scaled_data), method = "complete")
seg_hclust_single <- hclust(dist(scaled_data), method = "single")
seg_hclust_average <- hclust(dist(scaled_data), method = "average")
seg_hclust_centroid <- hclust(dist(scaled_data), method = "centroid")

# Plot dendrogram for hierarchical clustering
plot(seg_hclust_complete, main = "complete")
plot(seg_hclust_single, main = "single")
plot(seg_hclust_average, main = "average")
plot(seg_hclust_centroid, main = "centroid")

#Making Elbow plot with each Linkage Type revealed Complete to be the best linkage
# Plot Elbow plot
x <- c(1:10)
sort_height <- sort(seg_hclust_complete$height, decreasing = TRUE)
```

```
y <- sort_height[1:10]
plot(x, y); lines(x, y, col = "blue")

# Run k-means clustering with an optimal number of clusters
# Here, we'll use 3 clusters as an example
optimal_k <- 3
seg_kmeans <- kmeans(scaled_data, centers = optimal_k)

# Add segment number back to original data
seg$segment <- seg_kmeans$cluster

#Segmentation and Clustering with No NAs

# Import Data
seg2 <- data_2

# Data Preprocessing: Scale the variables
scaled_data2 <- scale(cbind(seg2$Quantity, seg2$ReturnRate, seg2$UnitPrice))

# Run hierarchical clustering
seg_hclust_complete2 <- hclust(dist(scaled_data2), method = "complete")
seg_hclust_single2 <- hclust(dist(scaled_data2), method = "single")
seg_hclust_average2 <- hclust(dist(scaled_data2), method = "average")
seg_hclust_centroid2 <- hclust(dist(scaled_data2), method = "centroid")



# Plot dendrogram for hierarchical clustering
plot(seg_hclust_complete2, main = "complete")
plot(seg_hclust_single2, main = "single")
plot(seg_hclust_average2, main = "average")
plot(seg_hclust_centroid2, main = "centroid")

# Plot Elbow plot
x2 <- c(1:10)
sort_height2 <- sort(seg_hclust_complete2$height, decreasing = TRUE)
y2 <- sort_height[1:10]
plot(x2, y2); lines(x2, y2, col = "blue")
#here single giving us the best result, decided to go for complete for the sake of consistency

# Determine the optimal number of clusters using the elbow method
# (This step may not be directly applicable to hierarchical clustering)
# You may need to decide the number of clusters based on domain knowledge or other criteria

# Run k-means clustering with an optimal number of clusters
# Here, we'll use 5 clusters as an example
optimal_k2 <- 3
seg_kmeans2 <- kmeans(scaled_data2, centers = optimal_k2)

# Add segment number back to original data
seg2$segment <- seg_kmeans2$cluster

#View of Segments
# Load the necessary library

# Group the data by Segment and summarise the average values
seg_summary2 <- seg2 %>%
  group_by(segment) %>%
  summarise(
    Count = n(),  # Count the number of records for each Segment
    AverageUnitPrice = mean(UnitPrice, na.rm = TRUE),
    AverageQuantity = mean(Quantity, na.rm = TRUE),
    AverageReturnRate = mean(ReturnRate, na.rm = TRUE)
  ) %>%
  pivot_longer(cols = -segment, names_to = "Metric", values_to = "Value") %>%
  pivot_wider(names_from = segment, values_from = Value)

# View the resulting data frame
print(seg_summary2)


# Export data to a CSV file, complete observations
write.csv(seg, file = "rough_segmentation_result_A.csv", row.names = FALSE)


##################################
#LDA and ANNOVA with Redundancy#
#################################
result <- read.csv(file.choose()) ## Choose segmentation result file


lda.fit <- lda (segment ~ Age + Education + Income + Married + Work, data = result)


lda.fit
#2 LDA equations as 3 segments.


lda.result <- predict(lda.fit, result)
lda.result


names(lda.result)
```

```
lda.class <- lda.result$class
lda.class

tclass <- table(result$segment, lda.class)
tclass



sum(result$segment == 1)
sum(result$segment == 2)
sum(result$segment == 3)



# Extract linear discriminant scores
lda.scores <- predict(lda.fit)$x

# Perform ANOVA on LD1
lm_LD1 <- lm(lda.scores[,1] ~ result$segment)
anova_LD1 <- anova(lm_LD1)
print(anova_LD1)
#Significant

# Perform ANOVA on LD2
lm_LD2 <- lm(lda.scores[,2] ~ result$segment)
anova_LD2 <- anova(lm_LD2)
print(anova_LD2)
#Results not significant

#LDA with removed redundancy, just in case
result2 <- result[!duplicated(result$CustomerID), ]
```

```
#This removes duplicate values


lda.fit2 <- lda (segment ~ Age + Education + Income + Married + Work, data = result2)

lda.fit2

#Both equations not significant


lda.result2 <- predict(lda.fit2, result2)

lda.result2


names(lda.result2)


lda.class2 <- lda.result2$class

lda.class2


tclass2 <- table(result2$segment, lda.class2)

tclass2

#Still all predictions made of customer segment 2


#No point in considering data of removed redundancy


###################

# RFM Analysis  #

###################


set.seed(1)


## Read in RFM data

rfm <- data_2
```

```
# Convert 'InvoiceDate' to Date object

rfm$InvoiceDate <- as.Date(rfm$InvoiceDate)


# Calculate recency based on the maximum date in the dataset

last_date <- max(rfm$InvoiceDate)

rfm$recency_days <- as.numeric(difftime(last_date, rfm$InvoiceDate, units = "days"))



# How many levels for each

groups <- 5  # This will use quintiles to sort and give 125 total groups


# Independent Sort

rfm$recency_score_indep <- ntile(-rfm$recency_days, groups)

rfm$frequency_score_indep <- ntile(rfm$Quantity, groups)  # Assuming 'Quantity' represents frequency

rfm$monetary_score_indep <- ntile(rfm$UnitPrice * rfm$Quantity, groups)  # Assuming 'UnitPrice' * 'Quantity' represents monetary

rfm$rfm_score_indep <- paste(rfm$recency_score_indep * 100 + rfm$frequency_score_indep * 10 + rfm$monetary_score_indep)


# Sequential Sort

rfm$recency_score_seq <- ntile(-rfm$recency_days, groups)

rfm$frequency_score_seq <- ave(rfm$Quantity, rfm$recency_score_seq, FUN = function(x) ntile(x, groups))

rfm$monetary_score_seq <- ave(rfm$UnitPrice * rfm$Quantity, interaction(rfm$recency_score_seq, rfm$frequency_score_seq), FUN = function(x) ntile(x, groups))

rfm$rfm_score_seq <- paste(rfm$recency_score_seq * 100 + rfm$frequency_score_seq * 10 + rfm$monetary_score_seq)


# Final Output

head(rfm)
```

```
## Export RFM Results with Independent and Sequential Sort

write.csv(rfm, file = "rfm_result.csv", row.names = FALSE)
```