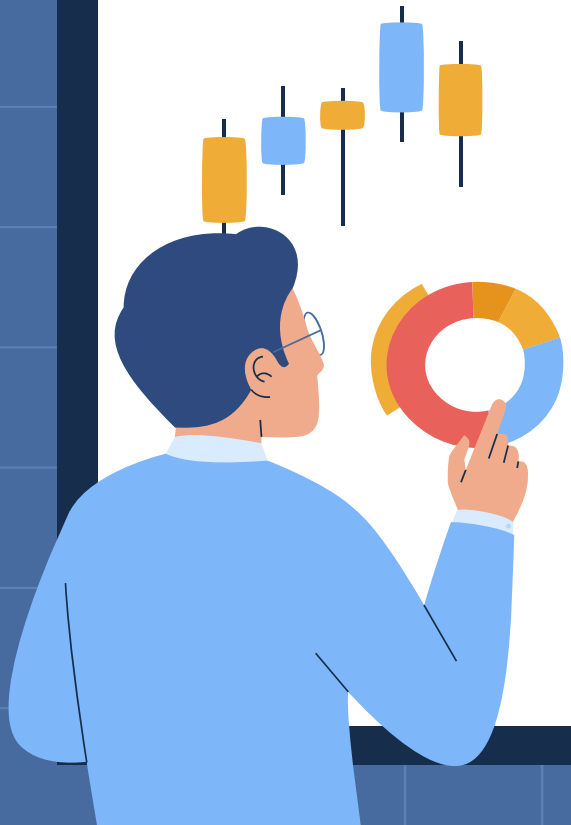
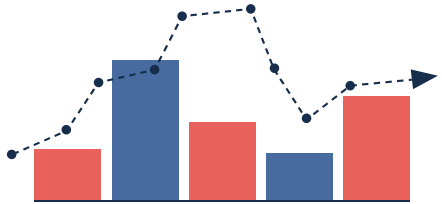


Analysis of Customer Purchasing Behavior on Black Friday

Final Report

Group 10



01

EDA Report

Project Overview

Business Objectives

– Black Friday is one of the largest shopping festivals in North America and even globally, attracting a vast number of consumers every year.
– The objective of analyzing consumer behavior on Black Friday is to enable businesses to refine their sales strategies and market positioning, ultimately leading to enhanced sales and profits.

Data Source

– A dataset containing the purchase history for users in a shopping system on Black Friday, with 12 columns and 537,577 records.

Data Dictionary

Variable	Description
User_ID	Unique identifier for the user.
Product_ID	Unique identifier for the product.
Gender	Gender of the user. (F: Female, M: Male)
Age	Age group of the user.
Occupation	Occupation code of the user.
City_Category	Category of the city where the user resides.
Stay_In _Current_City_Years	Number of years the user has lived in the current city.
Marital_Status	Marital status of the user. (0: Unmarried, 1: Married)
Product_Category_1	Product category 1 codes.
Product_Category_2	Product category 2 codes.
Product_Category_3	Product category 3 codes.

Initial Data Review

Character String Variables

Attribute	Missing Values	Unique Values
Product_ID	1145	3624
Gender	1150	3
Age	1181	8
City_Category	1136	4

Initial Data Review

Numeric Variables

Attribute	Missing Values	Unique Values	Mean	Min	Max	Standard Deviation
User_ID	1155	5892	-	-	-	-
Occupation	1153	22	-	-	-	-
Stay_In_Current_City_Years	1177	6	1.859385	0	4	1.2898630
Marital_Status	1119	3	0.4088484	0	1	0.4916216
Product_Category_1	1139	19	-	-	-	-
Product_Category_2	1119	19	-	-	-	-
Product_Category_3	1104	17	-	-	-	-
Purchase	1167	17956	9333.917	185	23961	4980.9520136

Numeric variables to convert to categories (factors):
User_ID, Occupation,
Stay_In_Current_City_Years, Marital_Status
Product_Category_1, Product_Category_2,
Product_Category_3

Initial Data Review

Convert to Factor Variables

Attribute	Missing Values	Unique Values	Mode
User_ID	1155	5892	Character
Product_ID	1145	3624	Character
Gender	1150	3	Character
Age	1181	8	Character
Occupation	1153	22	Character
City_Category	1136	4	Character
Stay_In_Current_City_Years	1177	6	Character
Marital_Status	1119	3	Character
Product_Category_1	1139	19	Character
Product_Category_2	1119	19	Character
Product_Category_3	1104	17	Character

Initial Data Review

Stay as Numeric Variable

Attribute	Missing Values	Unique Values	Mean	Min	Max	Standard Deviation
Purchase	1167	17956	9,333.917	185	23961	4980.9520136

Initial Data Review

Initial Observations

Data Quality Overview

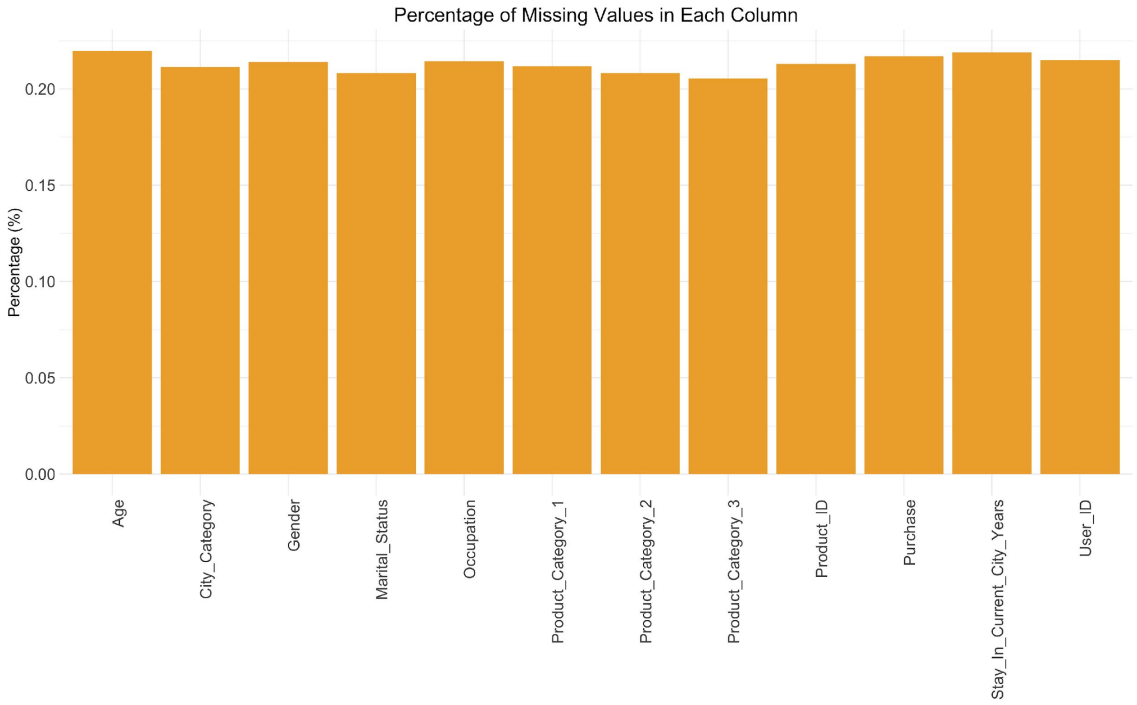
- All variables have a small number of missing values compared to the large amount of raw data.
- Each row represents either a transaction record or product information.

Composition of raw data

- 1 numeric variables
- 9 Factors
- 2 Identifiers/Keys

Data cleansing

Overview



Data cleansing

Investigate Total_Purchase Attribute

Gender <chr>	Mean_Purchase <dbl>	Total_Observations <int>	Missing_Values <int>	Percentage_Missing <dbl>
F	8810.565	131355	280	0.2131628
M	9505.076	403923	885	0.2191012

Age <chr>	Mean_Purchase <dbl>	Total_Observations <int>	Missing_Values <int>	Percentage_Missing <dbl>
0-17	9027.896	14612	24	0.1642486
18-25	9237.494	97017	209	0.2154262
26-35	9315.050	214463	474	0.2210171
36-45	9401.776	106829	241	0.2255942
46-50	9286.189	44233	89	0.2012072
51-55	9620.729	37353	94	0.2516531
55+	9450.086	20771	34	0.1636898

Data cleansing

Investigate Total_Purchase Attribute

City_Category <chr>	Mean_Purchase <dbl>	Total_Observations <int>	Missing_Values <int>	Percentage_Missing <dbl>
A	8958.625	143710	310	0.2157122
B	9201.126	226214	487	0.2152829
C	9844.122	165354	368	0.2225528

Stay_In_Current_City_Years <dbl>	Mean_Purchase <dbl>	Total_Observations <int>	Missing_Values <int>	Percentage_Missing <dbl>
0	9247.119	72287	159	0.2199566
1	9321.933	189151	409	0.2162294
2	9399.509	98784	197	0.1994250
3	9351.493	92675	210	0.2265983
4	9343.845	82381	190	0.2306357

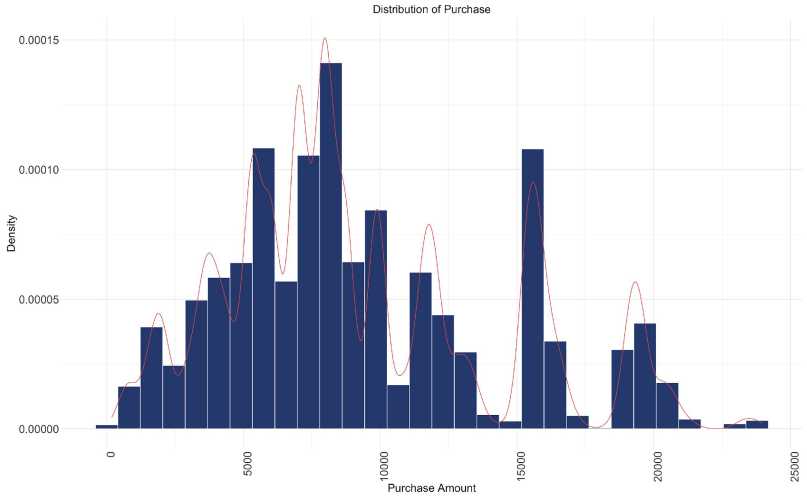
Data cleansing

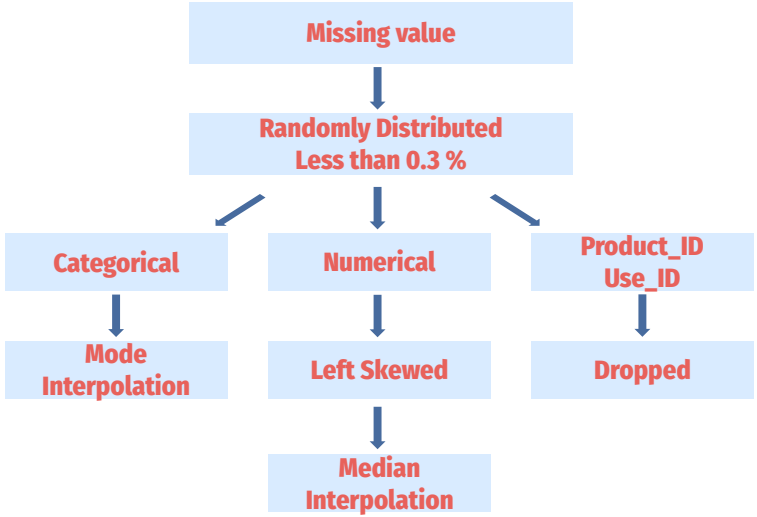
Missing Value Treatment

Chi-square	p-value
751	.489

A MCAR test was performed to investigate if the missing values are randomly located. With a p-value of 0.489, we can consider the missing values are randomly distributed. With a p-value of 0.489, we can consider the missing values are randomly distributed.

Mode was used to interpolate categorical values. It need to mention that due to the skewness of continuous variable “purchase”, median was used to interpolate the missing values.



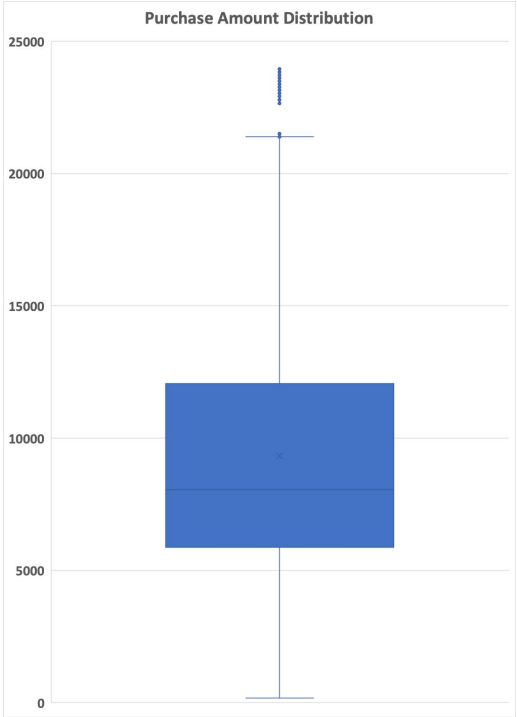


Data cleansing

Outlier Treatment

The distribution of this data over the variable Purchase is not very discrete as judged by the box plot. The Z-score of 3 is used as the cut-off value. Combining the nature of this dataset (purchase records) and the Z-score values for each row, this dataset is considered to have no outlier on variable “Purchase”.

Attribute	Max Z-Score	Min Z-Score
Purchase	2.939	-1.838



The Z-score represents the difference between a data point and the mean of a data set, measured in units of standard deviation.

Analytical Data Review

Initial Observations

Data Quality Overview

	Purchase Table	Product Table	User Table
Unique Identifier	*Product_ID & User_ID*	*Product_ID*	*User_ID*
Output	*Purchase*	*Total Purchase*	*Total Purchase*

Composition of raw data

- 1 numeric variable (response)
- 9 factors

Feature Engineering

For the following analysis, we decided to aggregate the purchase data on both User_ID and Product_ID

User Information

Measures (1)

Outcome:

Total_Purchase

Categories (6)

Demographic Information:

Gender
Age
Occupation
City_Category
Stay_In_Current_City_Years
Marital_Status

Product Information

Measures (1)

Outcome:

Total_Purchase

Categories (3)

Demographic Information:

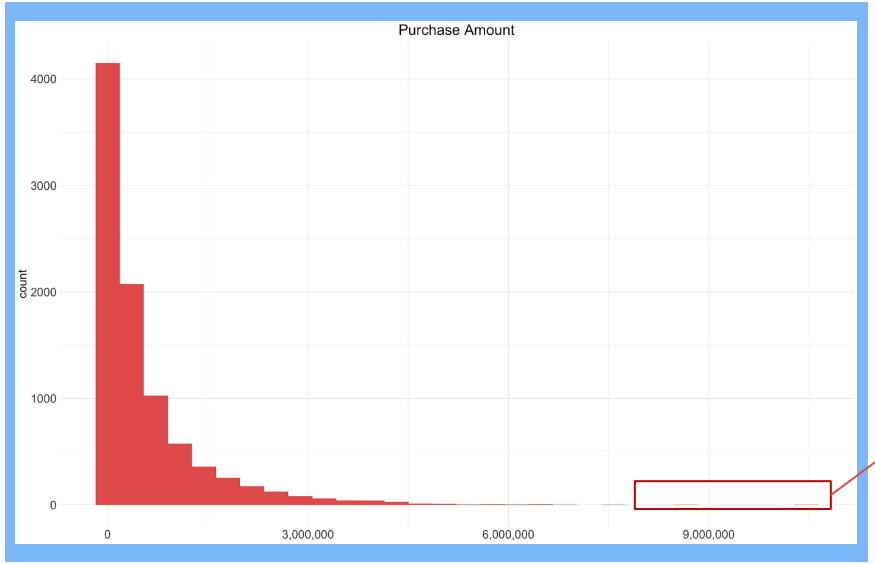
Product_Category_1
Product_Category_2
Product_Category_3

We partitioned the dataset into two distinct segments for analysis: one encapsulating User Information and the other focusing on Product Information.

Univariate Analysis

Target Outcome Variable - User Information

Total_Purchase

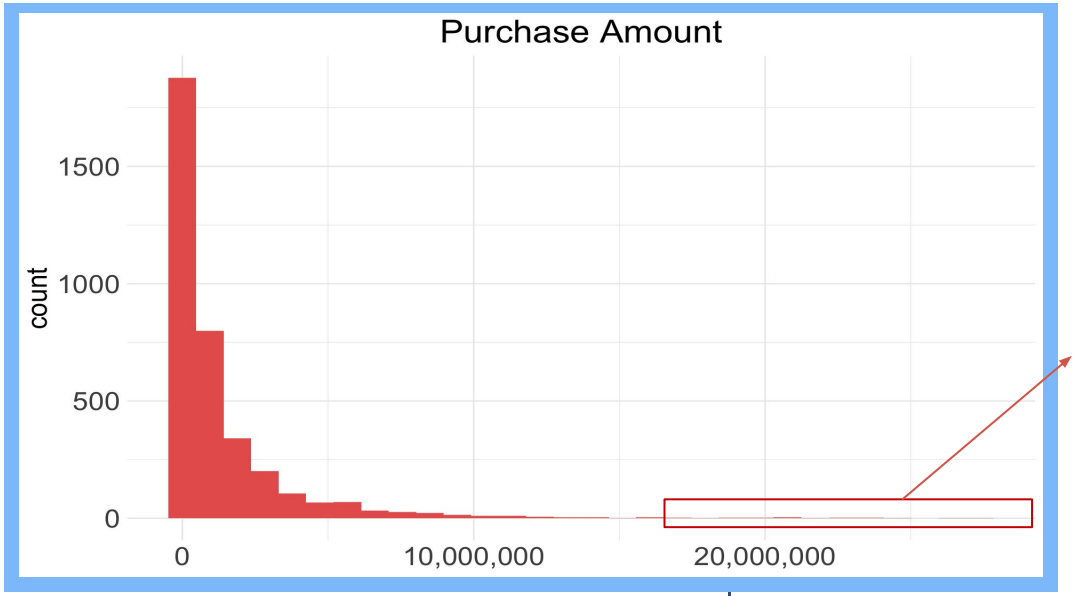


The aggregated user data is more discrete relative to the original Black_Friady data, and in order to ensure that only extreme outliers are filtered out rather than excluding variability in the data, a z-score cut off of 4 was used here, and 0.798% of the data points were removed as outliers.

Univariate Analysis

Target Outcome Variable - Product Information

Total_Purchase

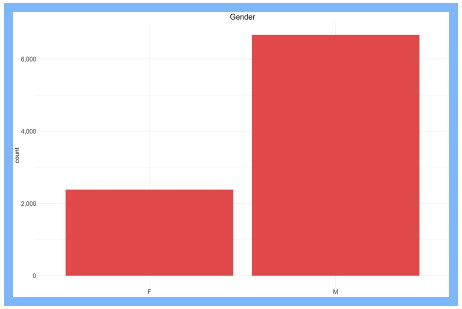


Similarly, since the product data is more discrete compared to the user (which may be due to more Purchases being aggregated into fewer Product IDs), the Z-score Cut-off value used here is 5, and 29 data points out of 3595 are removed as outliers.

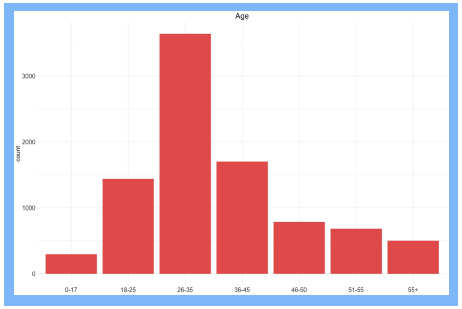
Univariate Analysis

User Demographic Information Categories

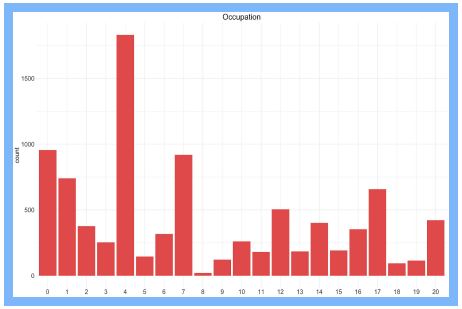
Gender



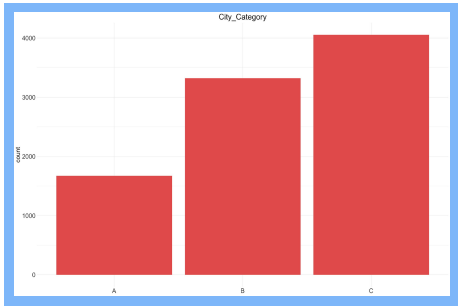
Age



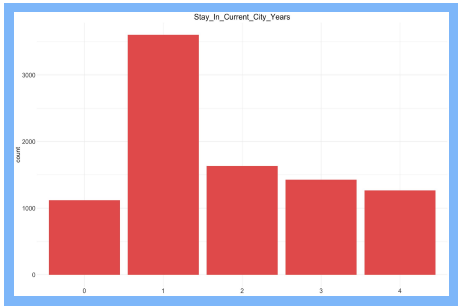
Occupation



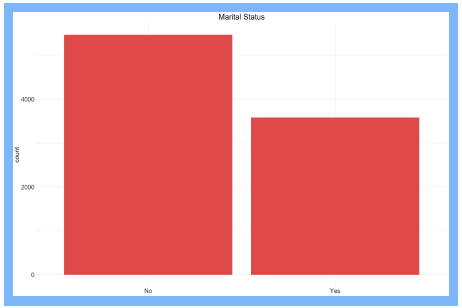
City_Category



Stay_In_Current_City_Years



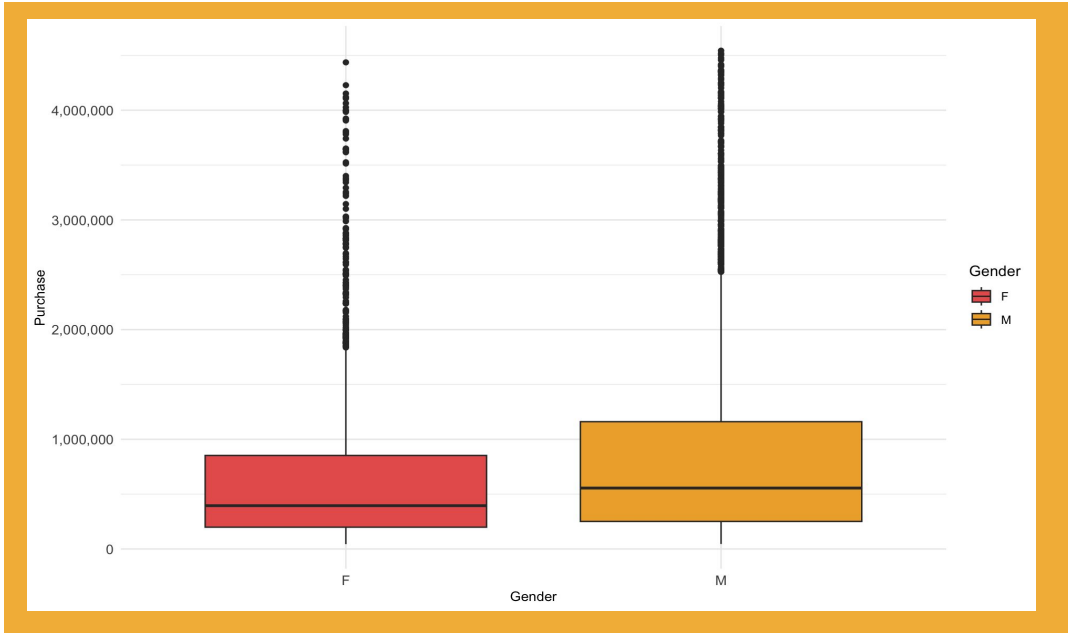
Marital_Status



Bivariate Analysis - User Information

Total Purchase vs Low-Cardinality Categories

Gender vs. Total Purchase

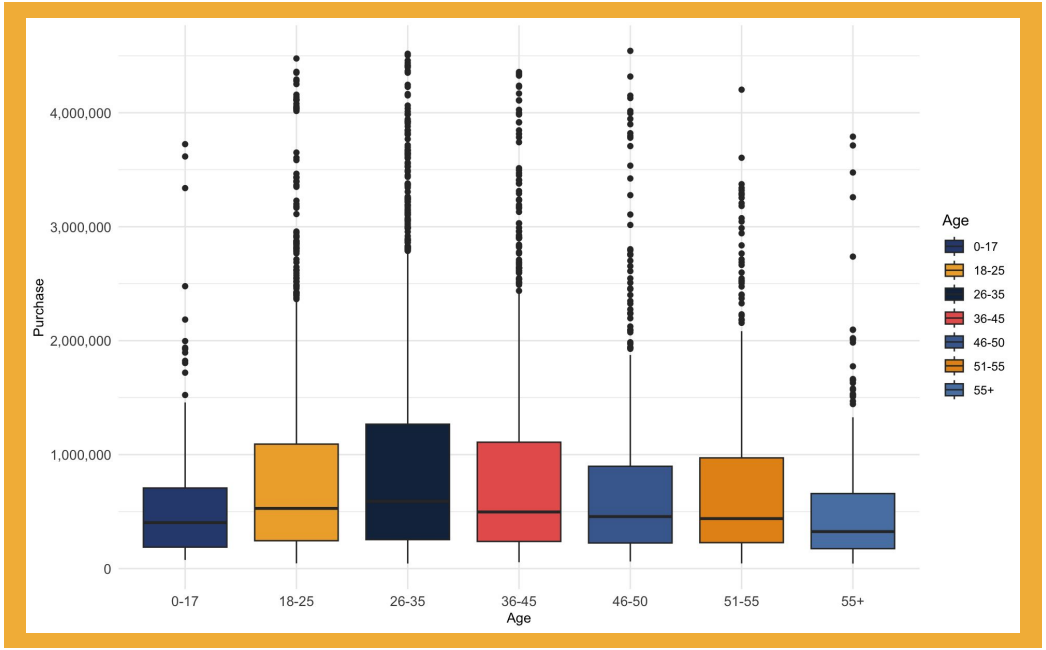


Male customers are associated with higher Total Purchase amount.

Bivariate Analysis - User Information

Total Purchase vs Low-Cardinality Categories

Age vs. Total Purchase

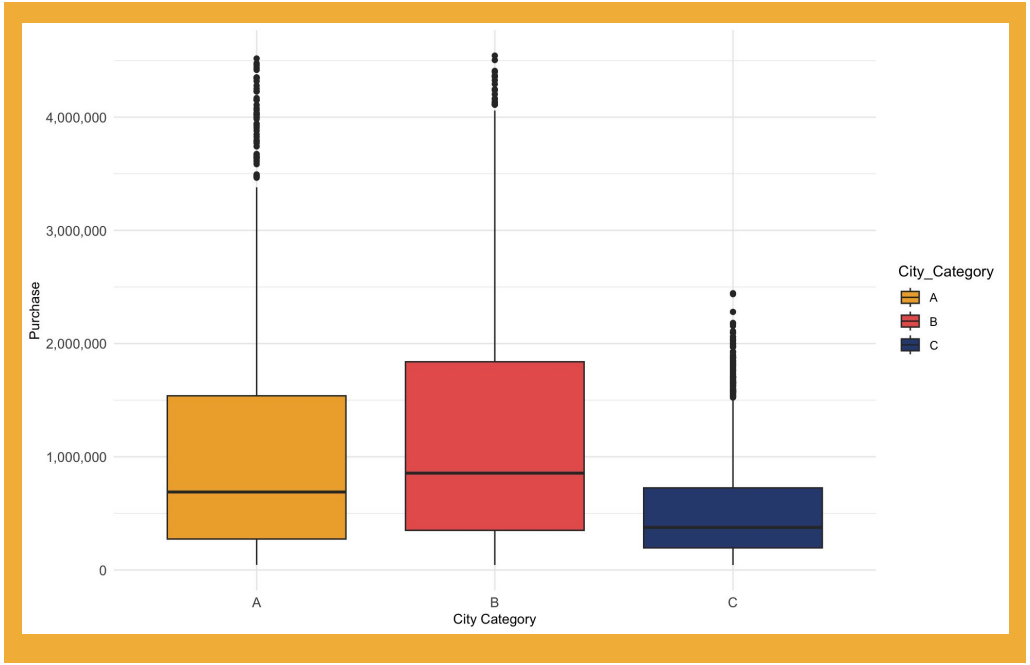


The box diagram indicating a strong correlation between age and total purchase amount. The 26-35 and 35-45 age groups tend to have higher purchase amount.

Bivariate Analysis - User Information

Total Purchase vs Low-Cardinality Categories

City Category vs. Total Purchase

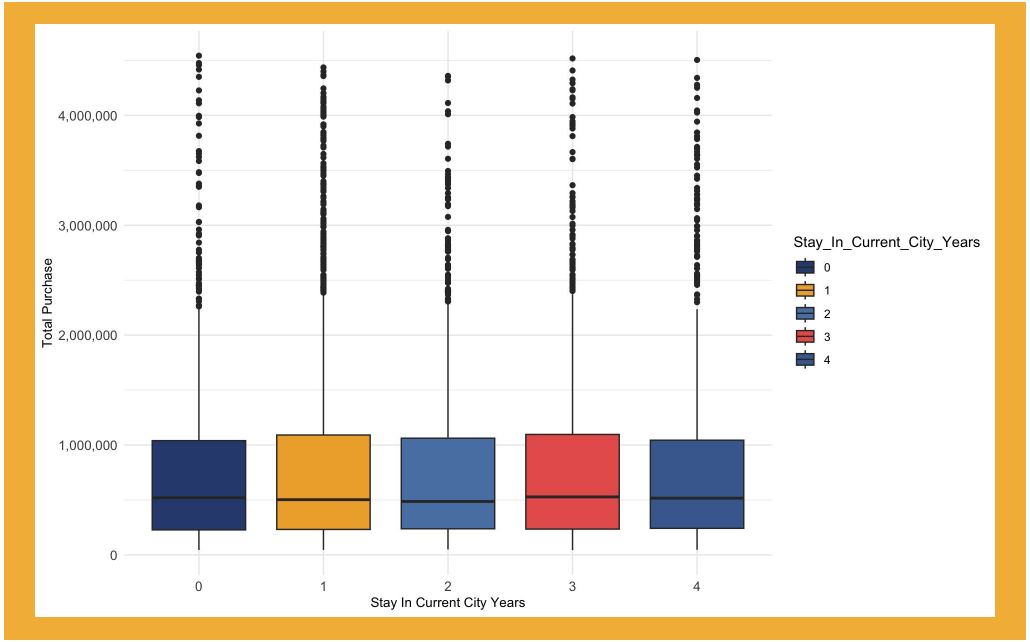


City categories A and B are clearly associated with higher purchase amounts, while the relative values for Category C are lower. This will be further argued in the statistical analysis

Bivariate Analysis - User Information

Total Purchase vs Low-Cardinality Categories

Stay_In_Current_City_Years vs. Total Purchase

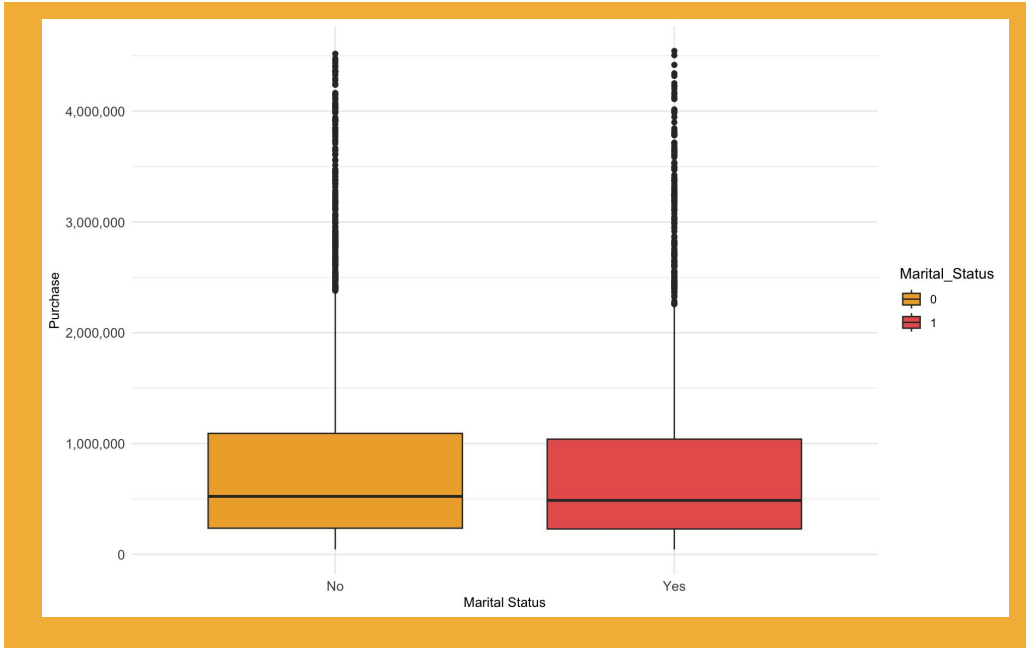


Stay_in_city_years does not seems to have a strong relation with total purchase amount. Further assessment will be shown in hypothesis test.

Bivariate Analysis - User Information

Total Purchase vs Low-Cardinality Categories

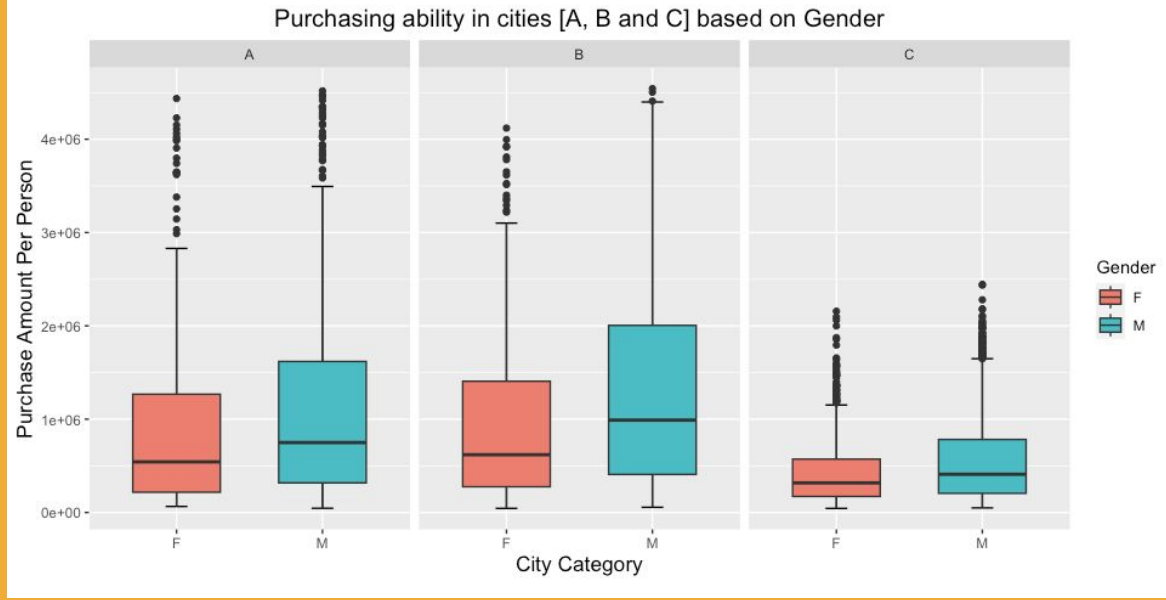
Marital Status vs. Total Purchase



According to box plot, marital status don't seems to have a strong effect on Total Purchase Amount

Bivariate Analysis - User Information

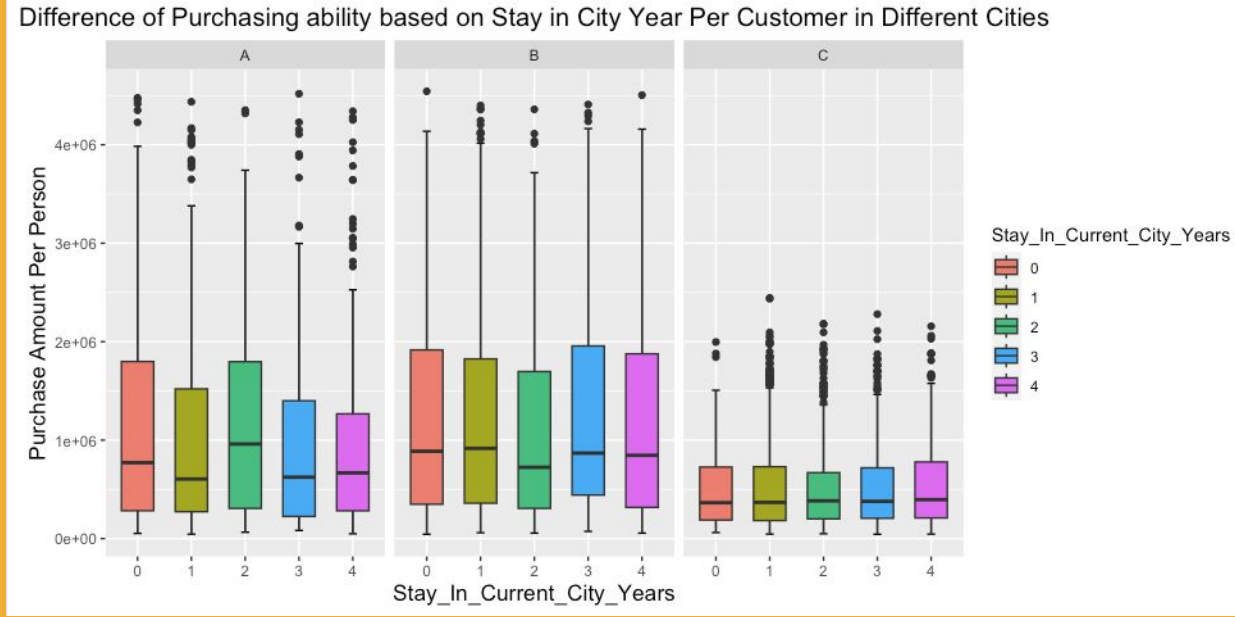
Total Purchase vs Low-Cardinality Categories



According to the box plot, city category B has the largest total purchase while C has the least. Total purchases are slightly higher for males than for females in every city.

Bivariate Analysis - User Information

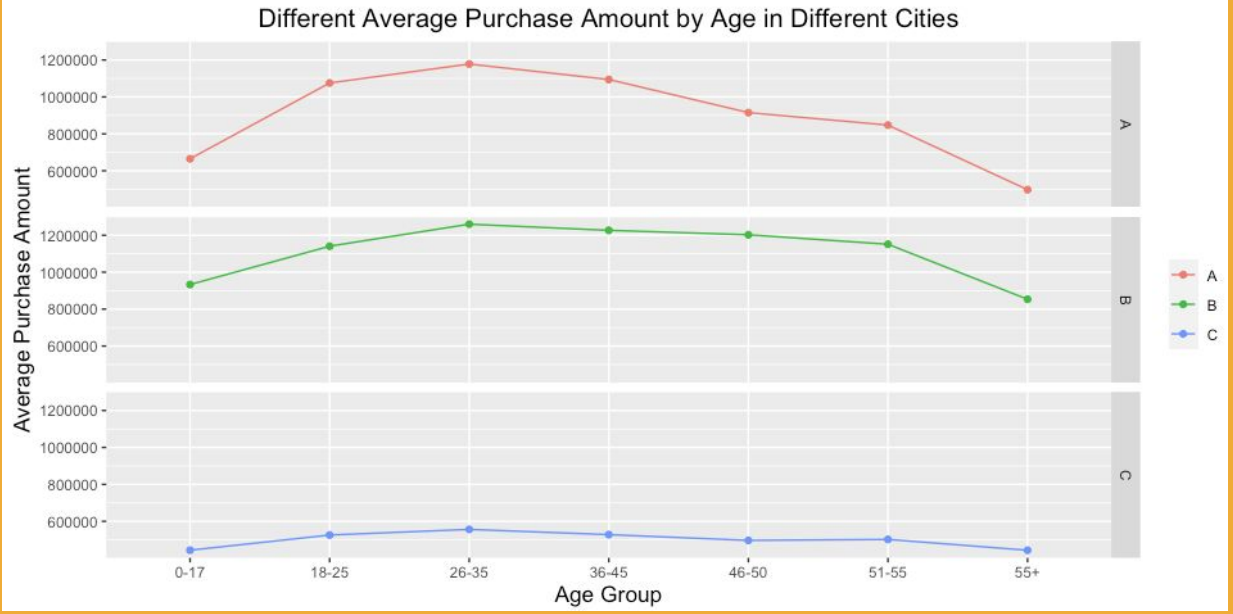
Total Purchase vs Low-Cardinality Categories



According to the box plot, the effect of STAY IN CITY YEARS on total purchases varies across cities. However, the overall difference is not substantial.

Bivariate Analysis - User Information

Total Purchase vs Low-Cardinality Categories



Consistent with the preceding sequence of analyses, City B has the highest total purchasing volume, while consumer purchasing power shows an increasing and then decreasing trend with age.

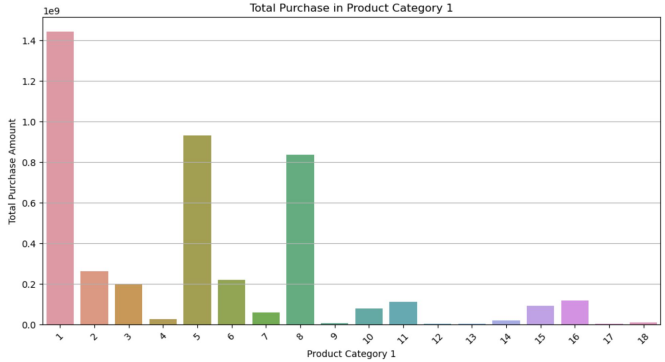
Bivariate Analysis - User Information

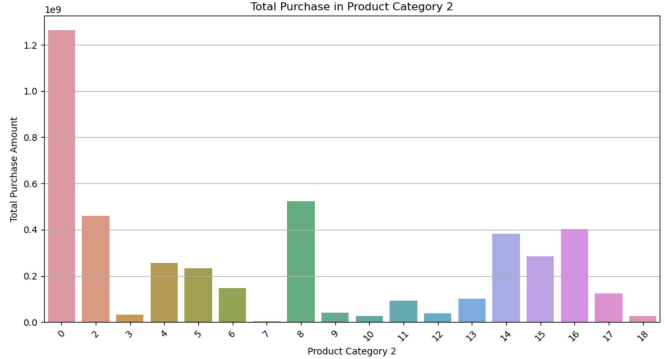
Total Purchase vs Low-Cardinality Categories

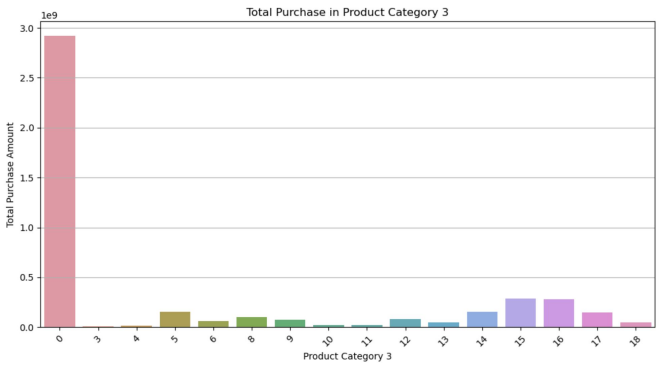
In cities A, B, there are individual occupations that reflect significantly higher/lower total purchases. Other occupation categories are not clearly distinguishable.

Bivariate Analysis - Product Information

Total Purchase vs Low-Cardinality Categories







According to the histogram, product category has a notable influence on total purchase amount, where each category is focused on a few individual categories with significantly higher total purchase amount. This will be tested statistically in the following analysis.

02

Statistical Analysis

Summarized Statistics

Numeric Variable

	Mean	SD	SE_Classic	SE_Bootstrap
User Information				
Total_Purchase	809,762.86910	823,007.46079	10,765.858529	10,533.898397
Product Information				
Total_Purchase	1,232,411.2890	1,980,737.1126	33,035.234317	32,475.510003

	90%_CI_Classic	90%_CI_Bootstrap
User Information		
Total_Purchase	[788,661.79, 830,863.95]	[792,986.59, 827,190.51]
Product Information		
Total_Purchase	[1,167,662.23, 1,297,160.35]	[1,180,721.39, 1,286,082.41]

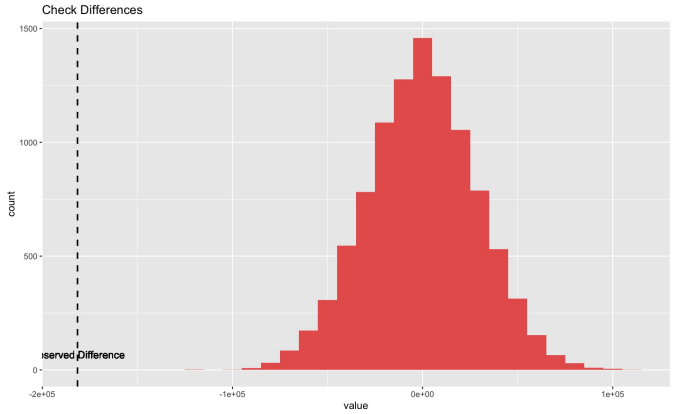
Hypothesis Testing of User Information

Total Purchase – t test

One-tailed hypothesis test to determine if there is a statistically significant difference between the total purchase of the two groups.

Gender

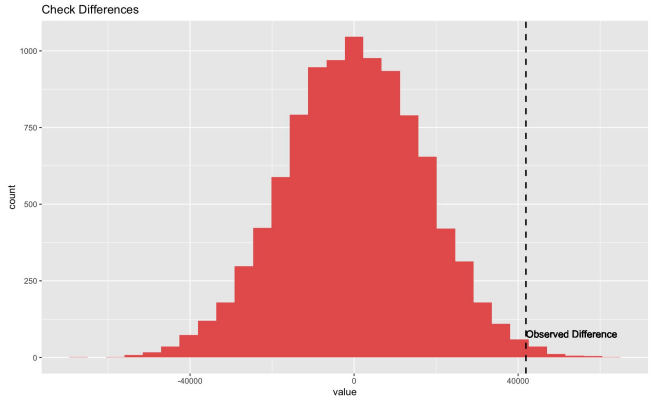
	Classical	Permutation
P-Value	1	1



There is no statistically significant difference between the total purchase of male and female.

Marital Status

	Classical	Permutation
P-Value	.0267 *	.0064 **



The difference between the total purchase of married and unmarried customers is statistically significant at 5% significance level.

One-way ANOVA to determine if there is statistically significant difference among the total purchase's means of more than two group.

Hypothesis Testing of User Information

Total Purchase – ANOVA

	Classical: P-Value	Computational: P-Value
Age	< 2.2e-16 ***	< .001 ***
Occupation	< 2.2e-16 ***	.0104 *
City_Category	< 2.2e-16 ***	< .001 ***
Stay_In_Current_City_Years	1	.8902

The ANOVA test indicates that there is strong statistically significant difference among the total purchase of different age groups, occupations, and different city category. Further investigation could be conducted to understand the buying patterns of these two factors.

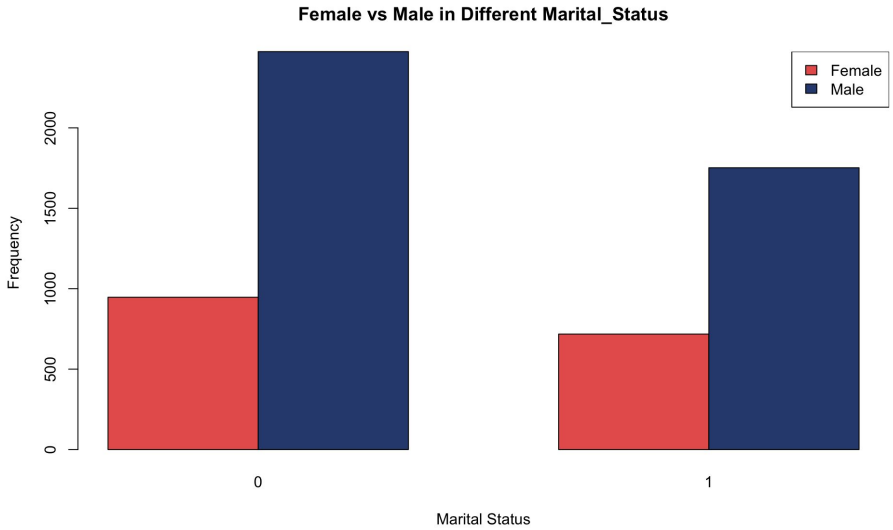
There is no statistically significant difference among the total purchase of different groups of years stay in the current city.

Hypothesis Testing of User Information

Chi-Squared Test

Gender & Marital Status

p-value= .2332



This indicates that there is no statistically significant association or dependence between the two categorical variables, gender and marital status.

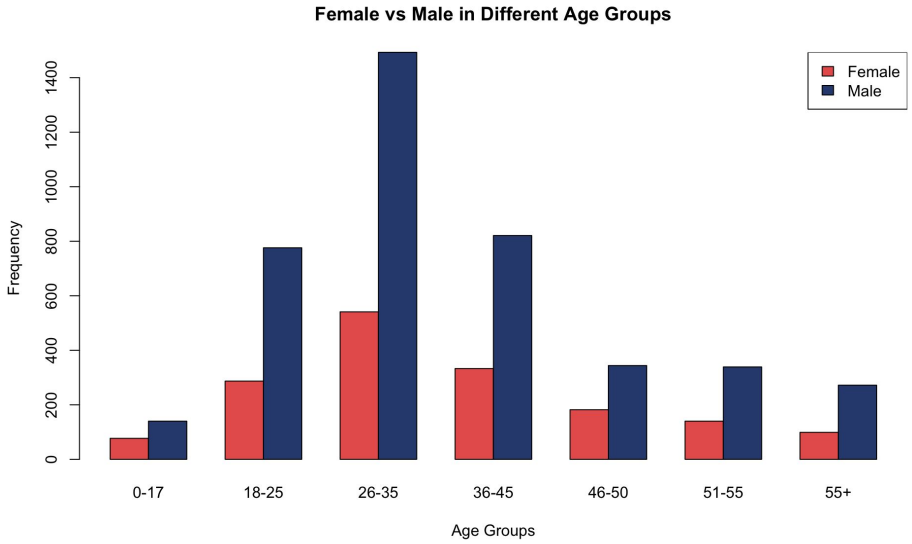
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Gender & Age

p-value = .002353 **



This indicates that there is a statistically significant association or dependence between the two categorical variables, gender and age, at 1% significance level.

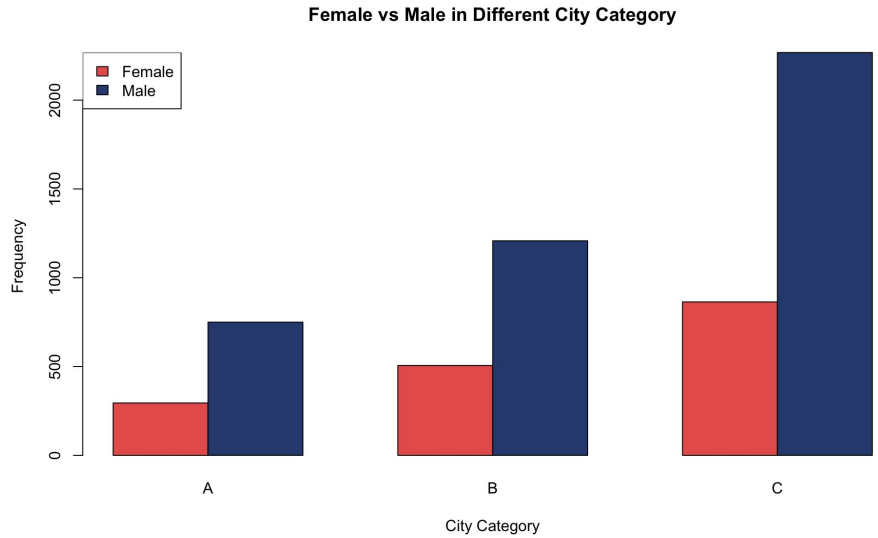
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Gender & City Category

p-value = .3097



This indicates that there is no statistically significant association or dependence between the two categorical variables, gender and city category.

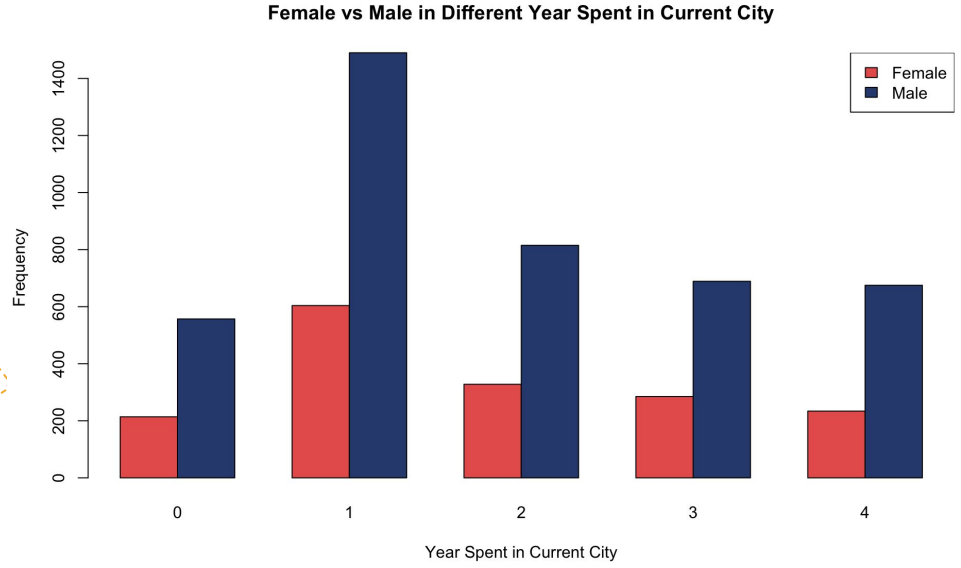
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Gender & Years Stayed in the Current City

p-value = .4226



This indicates that there is no statistically significant association or dependence between the two categorical variables, gender and year spent in current city.

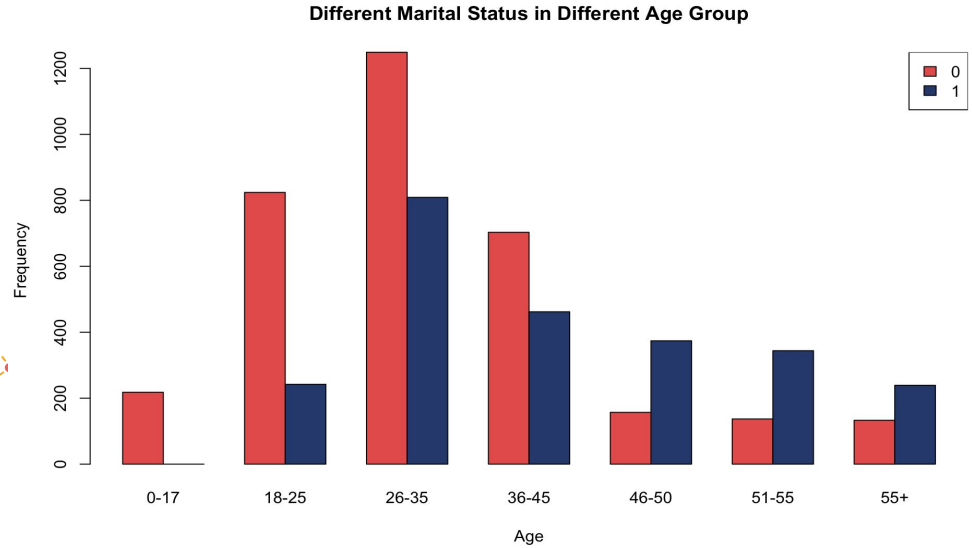
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Marital Status & Age

P-value < 2.2e-16 ***



This indicates that there is a statistically significant association or dependence between the two categorical variables, marital status and age, at 0.1% significance level.

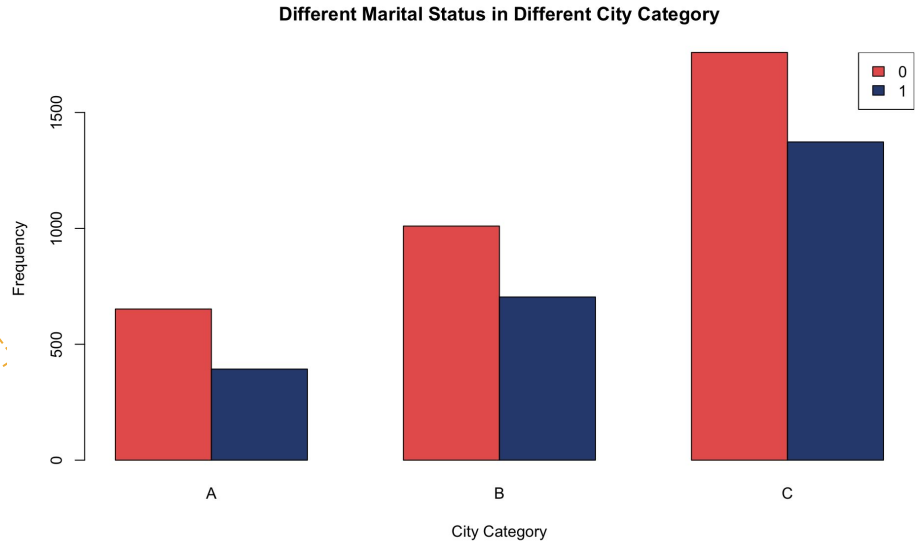
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Marital Status & City Category

P-value = .001261 **



This indicates that there is a statistically significant association or dependence between the two categorical variables, marital status and city category, at 0.1% significance level.

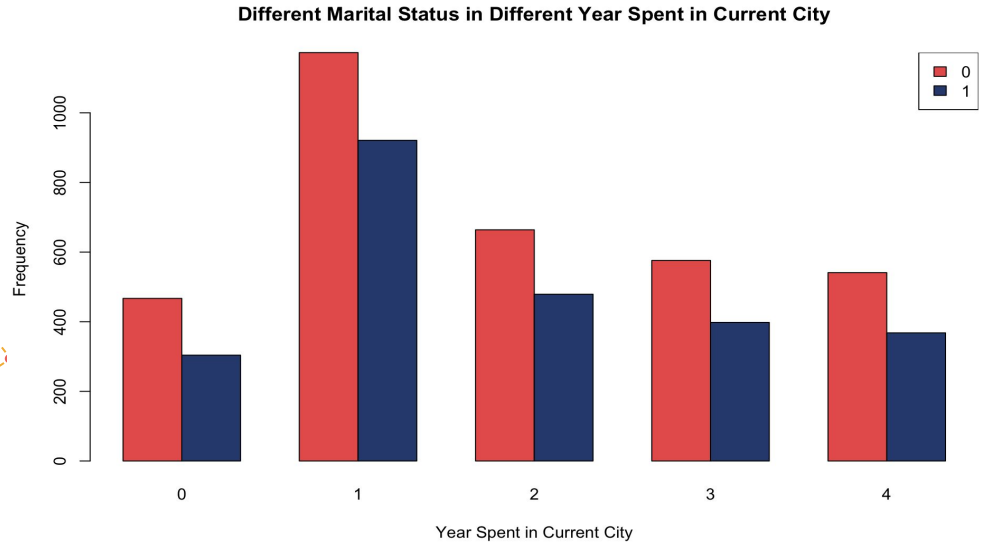
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Marital Status & Years Stayed in the Current City

P-value = .1605



This indicates that there is no statistically significant association or dependence between the two categorical variables, marital status and year spent in current city.

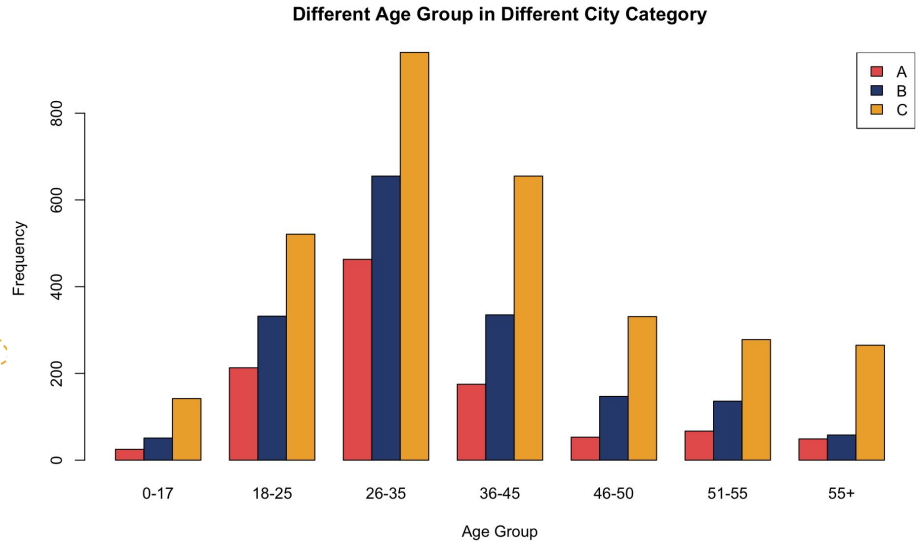
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Age & City Category

P-value < 2.2e-16 ***



This indicates that there is a statistically significant association or dependence between the two categorical variables, age and city category, at 0.1% significance level.

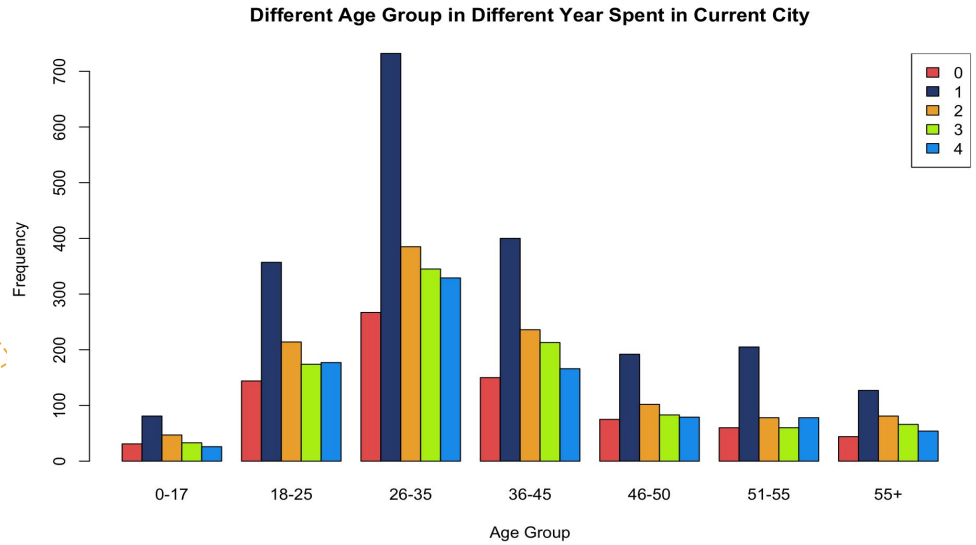
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

Age & Years Stayed in the Current City

P-value = .2387



This indicates that there is no statistically significant association or dependence between the two categorical variables, Age and Years Stayed in the Current City.

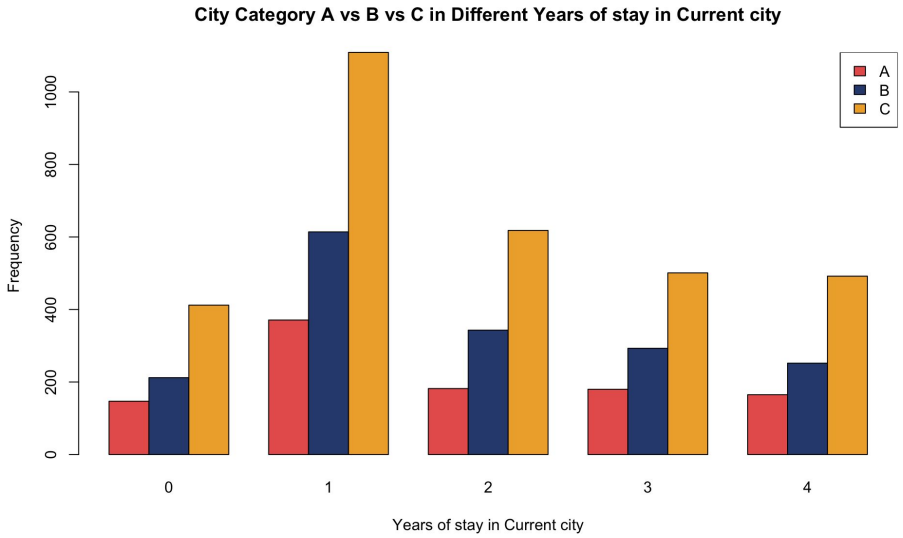
A chi-squared test is used to determine if there is a significant association between two categorical variables.

Hypothesis Testing of User Information

Chi-Squared Test

City Category & Years Stayed in the Current City

p-value = .7594



This indicates that there is no statistically significant association or dependence between the two categorical variables, City Category and Years Stayed in the Current City.

A chi-squared test is used to determine if there is a significant association between two categorical variables.

One-way ANOVA to determine if there is statistically significant difference among the total purchase's means of more than two group.

Hypothesis Testing of Product Category

Total Purchase – ANOVA

	Classical: P-Value	Computational: P-Value
Product_Category_1	< 2.2e-16 ***	.0187
Product_Category_2	< 2.2e-16 ***	.0132
Product_Category_3	< 2.2e-16 ***	.0117

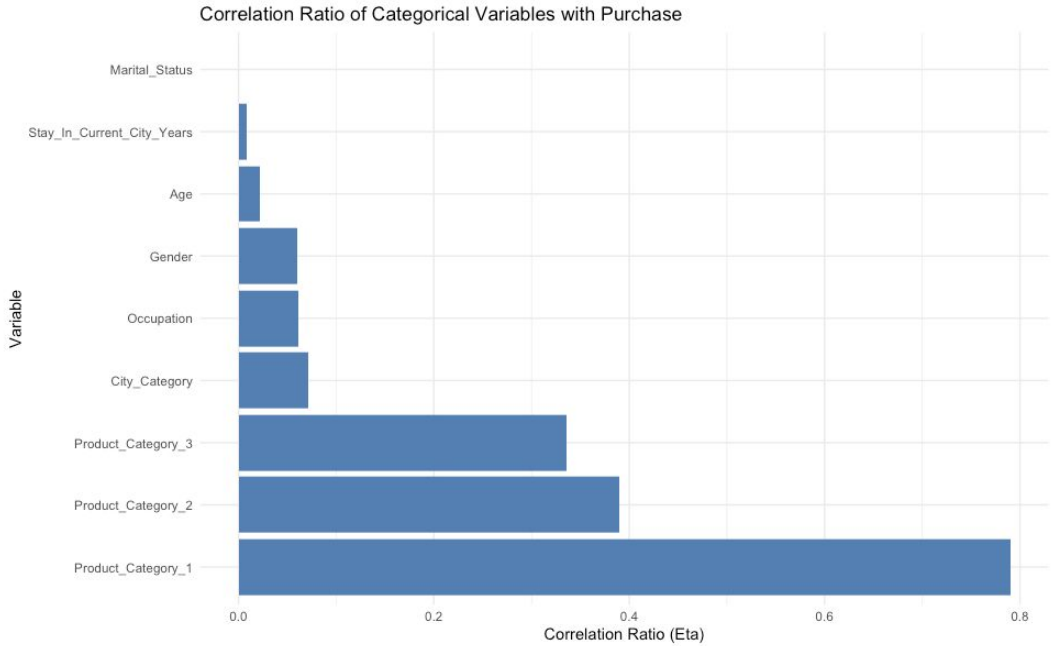
The classical ANOVA test indicates that there is strong statistically significant difference among the total purchase of different product categories. However, the computational ANOVA test shows that there is statistically significant different at the significance level of 5%. Further investigation could be conducted to understand the buying patterns of these three factors.

Chi-square Test

	Category 1 vs 2	Category 2 vs 3	Category 1 vs 3
P-value	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***
X-squared	6,638.6	5,612.6	2,181.2

Chi-square indicates that there is significantly strong association between category 1, 2 and 3. It is worth noting that there are a number of species that have too small a frequency number, which may affect the accuracy of the Chi-square test. However, overall, the correlation between categories should be considered.

Correlation Analysis



Consistent with the hypothesis test , product categories has the greatest relationship with purchase price. And the relationship between user information and purchase price is smaller.

Correlation Ratios (Eta) were used to indicate correlation between numerical variable and categorical variables

03

Modeling

Feature Engineering

Label Encoded & One-Hot Encoded Features

User_ID <dbl>	Gender <dbl>	Age <fctr>	Occupation <fctr>
1000001	0	0-17	10
1000002	1	55+	16
1000003	1	26-35	15
1000004	1	46-50	7
1000005	1	26-35	20
1000006	0	51-55	9

User_ID <dbl>	Gender <dbl>	Age <fctr>	Occupation <fctr>
1000001	0	0-17	10
1000002	1	55+	16
1000003	1	26-35	15
1000004	1	46-50	7
1000005	1	26-35	20
1000006	0	51-55	9

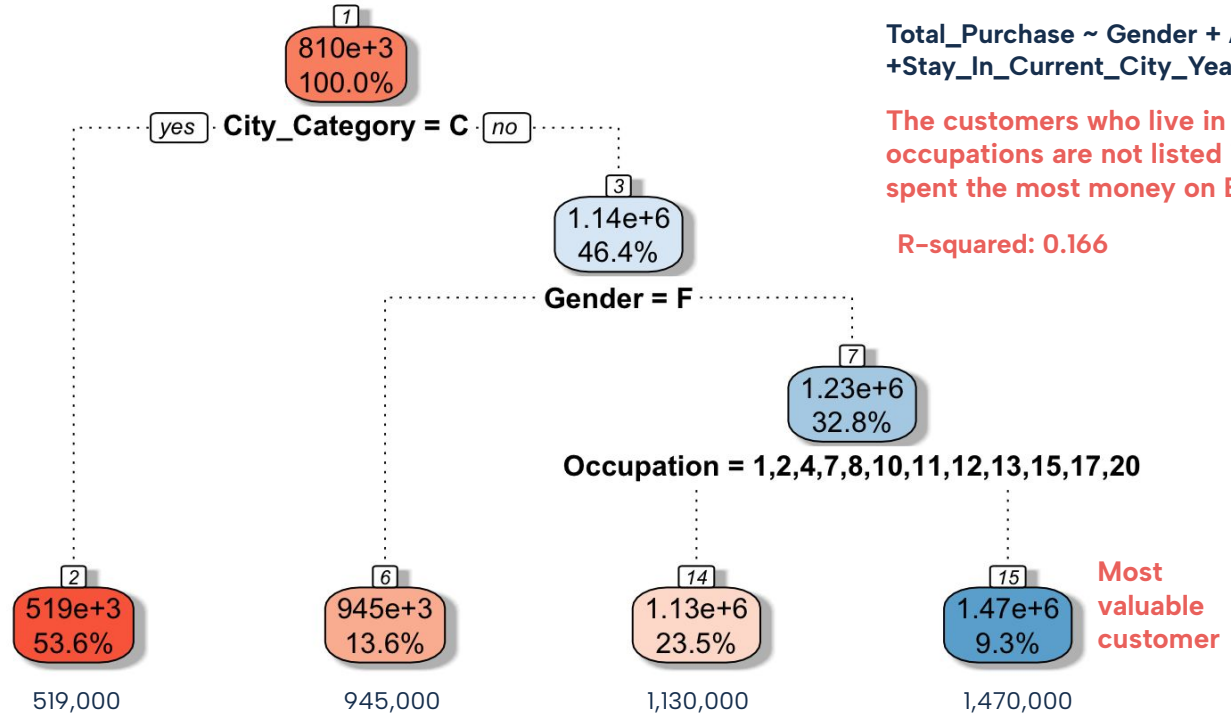
Marital_Status <dbl>	Total_Purchase <dbl>
1	333481
1	810353
1	341635
2	205987
2	822111
1	377371

Age0-17 <dbl>	Age18-25 <dbl>	Age26-35 <dbl>	Age36-45 <dbl>	Age46-50 <dbl>	Age51-55 <dbl>	Age55+ <dbl>
1	0	0	0	0	0	0
0	0	0	0	0	0	1
0	0	1	0	0	0	0
0	0	0	0	1	0	0
0	0	1	0	0	0	0
0	0	0	0	0	1	0

Regression Tree Model

User Information

Pruning is a technique used in decision trees to avoid overfitting. It helps simplify the tree by removing branches that do not contribute significantly to predictive accuracy.



Total_Purchase ~ Gender + Age + Occupation + City_Category +Stay_In_Current_City_Years + Marital_Status

The customers who live in City A or B are male, and whose occupations are not listed in the graph are the ones who spent the most money on Black Friday.

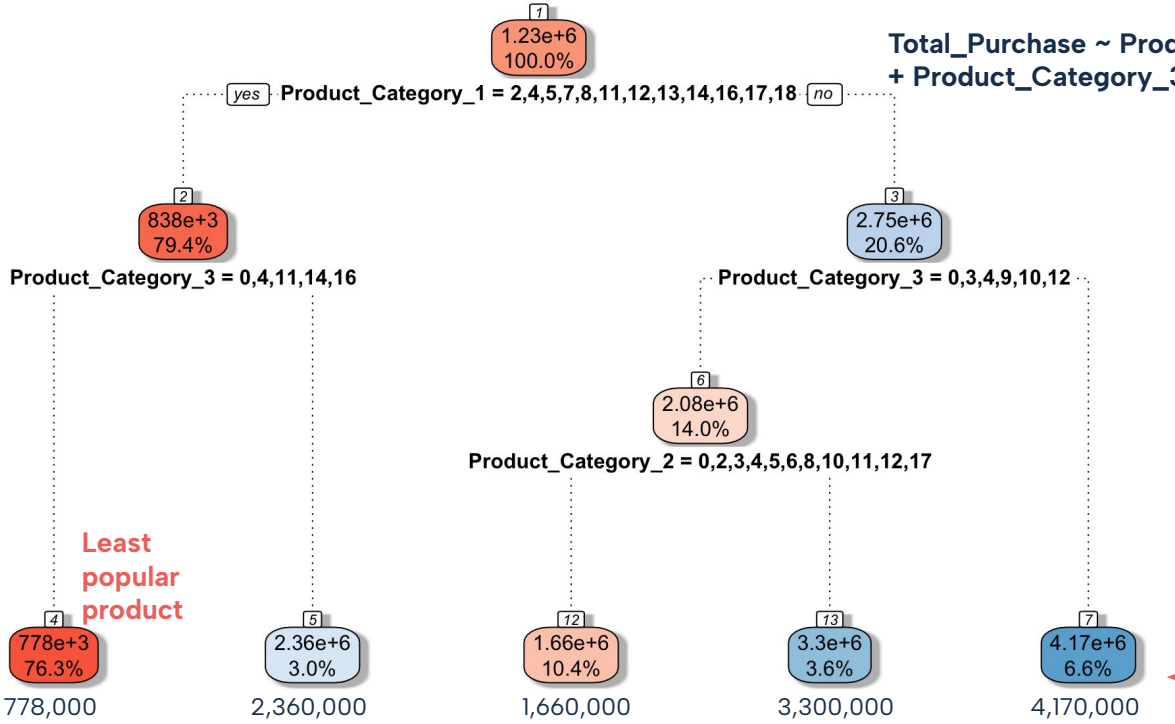
R-squared: 0.166

Variable Importance	
City_Category	78
Occupation	11
Gender	6
Age	5

Regression Tree Model

Product Information

Pruning is a technique used in decision trees to avoid overfitting. It helps simplify the tree by removing branches that do not contribute significantly to predictive accuracy.



Least popular product

Total_Purchase ~ Product_Category_1 + Product_Category_2 + Product_Category_3

R-squared: 0.24

Variable Importance	
Product_Category_1	49
Product_Category_2	30
Product_Category_3	21

Most popular product Revenue is at an average level of $4M

Linear Regression

User Information

Residuals:

Min	1Q	Median	3Q	Max
-1045704	-541920	-264251	262651	3646224

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	606071	100318	6.041	1.62e-09 ***
Occupation12	-152516	52509	-2.905	0.003691 **
Occupation17	-188940	48620	-3.886	0.000103 ***
Occupation7	-121454	44791	-2.712	0.006716 **
Gender	193731	24691	7.846	5.07e-15 ***
Age26-35	255256	92895	2.748	0.006018 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

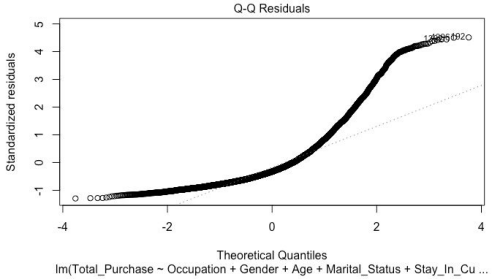
Residual standard error: 809800 on 5811 degrees of freedom

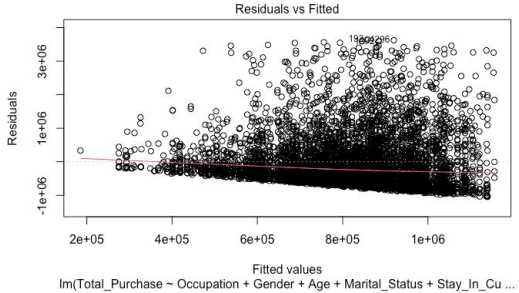
Multiple R-squared: 0.03713, Adjusted R-squared: 0.03182

F-statistic: 7.002 on 32 and 5811 DF, p-value: < 2.2e-16

Marital Status and Stay_In_Current_City_Years

+ Coefficients are not statistically significant (p > 0.05).
+ There is a lack of evidence to support a meaningful effect on the purchase amount.
+ Variability in purchase amount associated with these variables could be due to randomness





Linear Regression

Product Information

According to statistical analysis, all 3 product categories is significant in the model.

Residuals:

Min	1Q	Median	3Q	Max
-4958355	-733411	-385108	304654	13052498

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2767058	688942	4.016	6.03e-05 ***
Product_Category_1_1	1630800	330876	4.929	8.66e-07 ***
Product_Category_1_2	885031	366037	2.418	0.015662 *
Product_Category_1_3	1674043	418107	4.004	6.36e-05 ***
Product_Category_1_6	1070178	360367	2.970	0.003001 **
Product_Category_1_10	2447038	471198	5.193	2.18e-07 ***
Product_Category_1_13	-879462	431032	-2.040	0.041388 *
Product_Category_1_15	1349145	406326	3.320	0.000908 ***
Product_Category_1_16	909834	355300	2.561	0.010486 *
Product_Category_2_3	1778040	829807	2.143	0.032204 *
Product_Category_2_4	-1295573	471427	-2.748	0.006023 **
Product_Category_3_0	-1899334	463244	-4.100	4.22e-05 ***
Product_Category_3_4	-3952250	798165	-4.952	7.70e-07 ***
Product_Category_3_14	-1227248	501341	-2.448	0.014416 *
Product_Category_3_16	-1222099	490944	-2.489	0.012845 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

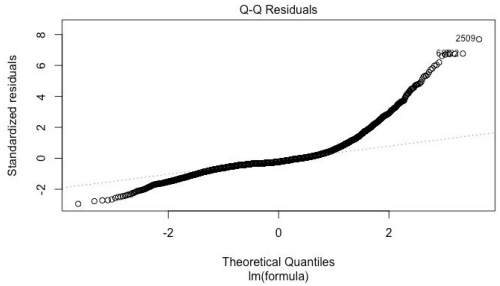
Residual standard error: 1701000 on 3545 degrees of freedom

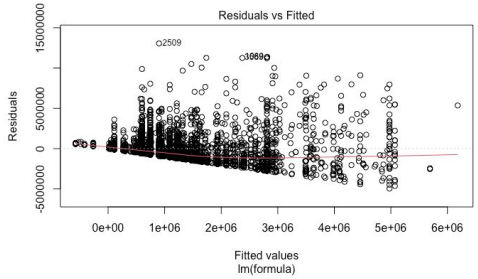
Multiple R-squared: 0.2728, **Adjusted R-squared: 0.2627**

F-statistic: 27.14 on 49 and 3545 DF, p-value: < 2.2e-16

Marital Status and Stay_In_Current_City_Years

+ All three PRODUCT categories were significantly associated with the results (p > 0.05).
+ Some of these product categories are associated with high total purchase values, while some are associated with low total purchase values





Cluster Analysis

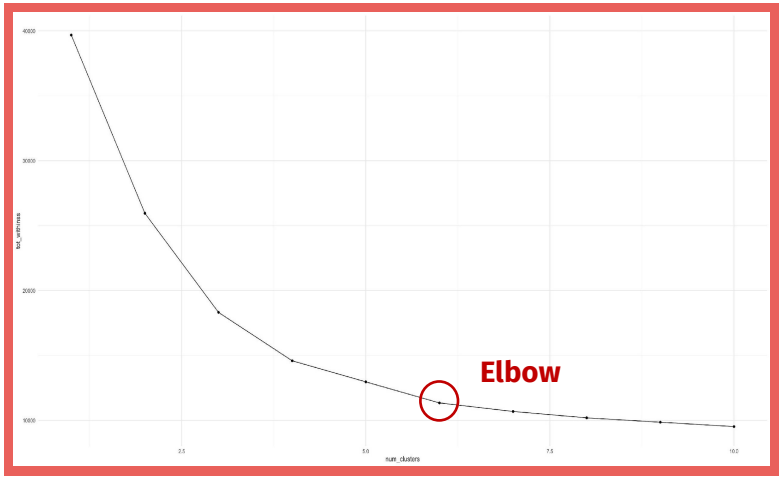
User Information

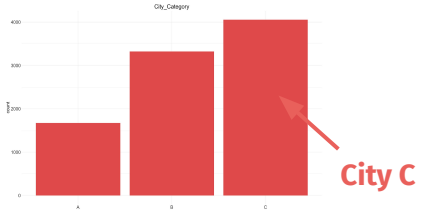
Gender
Age
City_Category
Stay_In_Current_City_Years
Marital_Status

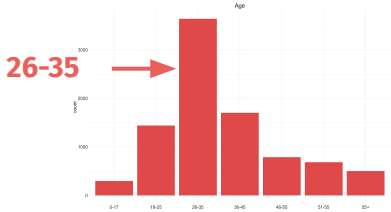
Dummy Variables

The analysis suCi2ggests that the best clustering occurs with K=6, primarily influenced by Gender, Stay in Current City Years, and Marital Status.

Cluster 1 – Single men with a notably **long stay** in their current city.
Cluster 2 – Married women, characterized by **short-term** city residencies.
Cluster 3 – Single men with the **shortest** average residency.
Cluster 4 – Married men with **lengthy** periods of residence in their city.
Cluster 5 – Married men with a brief **average stay**.
Cluster 6 – Single men with **moderate** lengths of residency.







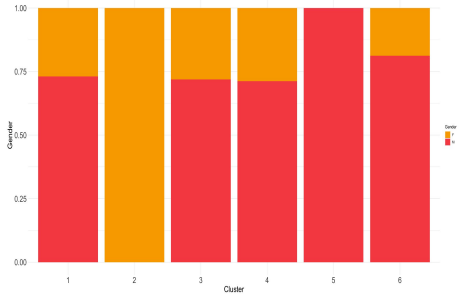
Cluster_km <fctr>	Size <int>	Gender <fctr>	Age <fctr>	City_Category <fctr>	Stay_In_Current_City_Years <dbl>	Marital_Status <fctr>	Total_Purchase <dbl>
1	1107	M	26-35	C	3.4859982	0	842670.0
2	498	F	26-35	C	1.0963855	1	673458.7
3	1627	M	26-35	C	0.7166564	0	820928.4
4	757	M	26-35	C	3.4808454	1	787729.8
5	859	M	26-35	C	0.7566938	1	834227.4
6	996	M	26-35	C	2.0000000	0	818747.9

6 rows

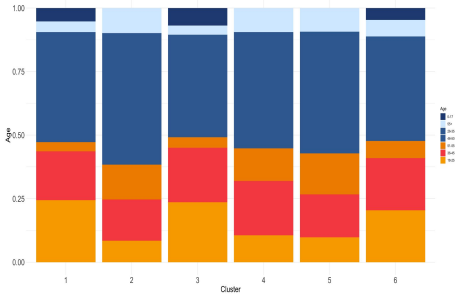
Cluster

User Information

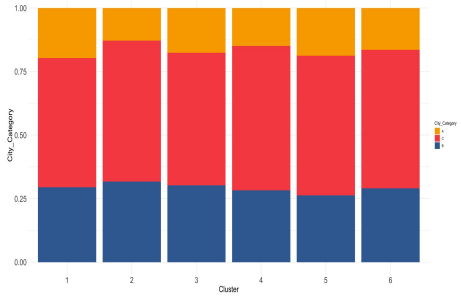
Gender



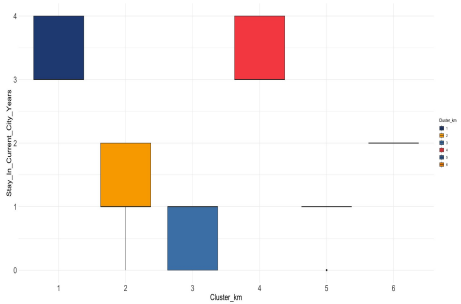
Age



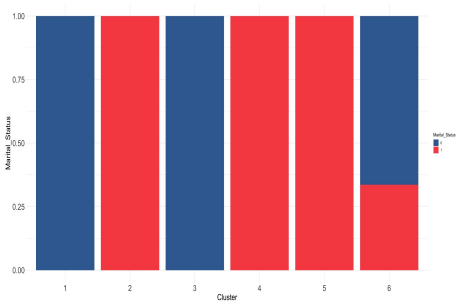
City_Category



Stay_In_Current_City_Years



Marital_Status



Cluster

User Information

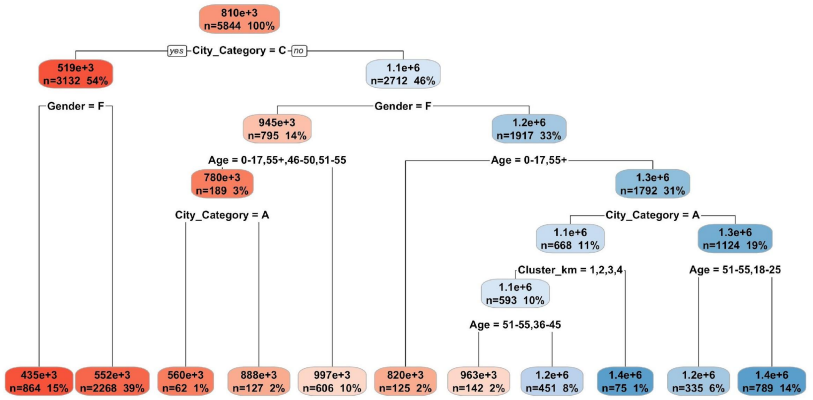
Linear Regression

Total_Purchase ~ Gender + Age + City_Category + Stay_In_Current_City_Years + Marital_Status + Cluster_km

```
Call:
lm(formula = Total_Purchase ~ . - 1, data = Final_data_model)

Residuals:
     Min       1Q   Median       3Q      Max
-1226800  -453518  -174071   309713  3475969

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
Stay_In_Current_City_Years   -21530      21900  -0.983 0.325613
`(Intercept)`                847060      97834   8.658  < 2e-16 ***
Cluster_km2                  -18525      87323  -0.212 0.832005
Cluster_km3                  -70424      67386  -1.045 0.296028
Cluster_km4                   12174      62273   0.195 0.845012
Cluster_km5                  -65558      85466  -0.767 0.443073
Cluster_km6                  -38402      49308  -0.779 0.436108
GenderM                      203563      26094   7.801 7.23e-15 ***
`Age55+`                     -40817      65918  -0.619 0.535805
`Age26-35`                   205324      54696   3.754 0.000176 ***
`Age46-50`                   129929      62687   2.073 0.038247 *
`Age51-55`                    96436      63589   1.517 0.129433
`Age36-45`                   163764      56446   2.901 0.003731 **
`Age18-25`                   130503      56493   2.310 0.020918 *
City_CategoryC              -527014      27599 -19.096  < 2e-16 ***
City_CategoryB               133779      30034   4.454 8.58e-06 ***
Marital_Status1              -20166      51577  -0.391 0.695817
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 753200 on 5827 degrees of freedom
Multiple R-squared:  0.5757,    Adjusted R-squared:  0.5744
F-statistic:   465 on 17 and 5827 DF,  p-value: < 2.2e-16
```

Decision Tree

Total_Purchase ~ Gender + Age + City_Category + Stay_In_Current_City_Years + Marital_Status + Cluster_km



R-squared: 0.155

-The accuracy of the tree model predictions is considerably low.

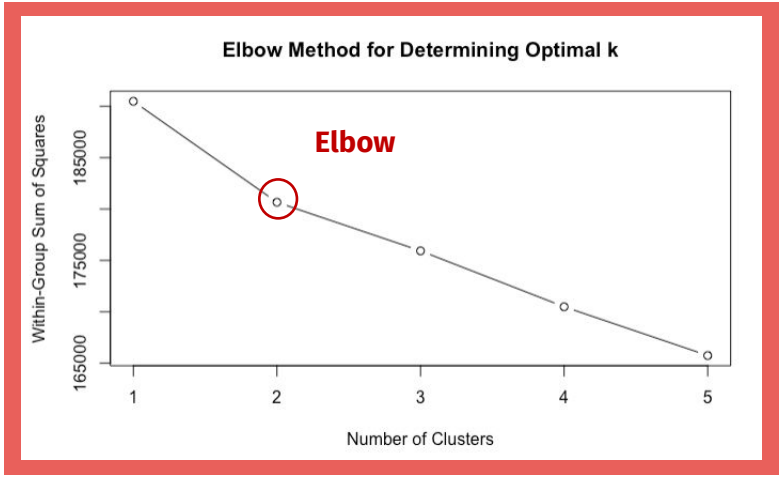
Cluster Analysis

Product Information

Product_Category_1
Product_Category_2
Product_Category_3

Dummy Variables

The analysis suggests that the best clustering occurs with K=2. It is straightforward to interpret this as the group with the higher Total Purchase and the group with lower Total Purchase amount.



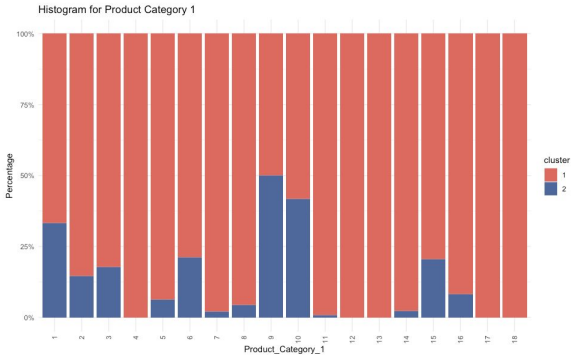
Group.1	Total_Purchase
1	0.9158363
2	-0.1601217

Among them, Product Category 1's 1, Product Category 2's 2, 4 Product Category 3's 5, 8, 9, 12, 14, 15, 16, 17 present and high concentration of purchase amount. Product Category 1's 8, Product Category 1's 0, and Product Category 3's 0 are associated with clusters of low purchase amounts.

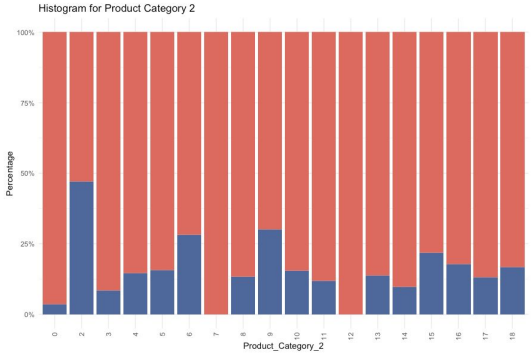
Cluster

Product Information

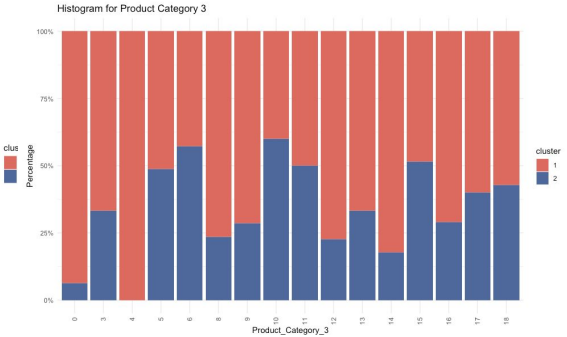
Product_Category 1



Product_Category 2



Product_Category 3



Cluster

Product Information

Linear Regression

Total_Purchase ~Product_Category_1 + Product_Category_2+Product_Category_3 + Distance to Cluster Center 1 + Distance to Cluster Center 2

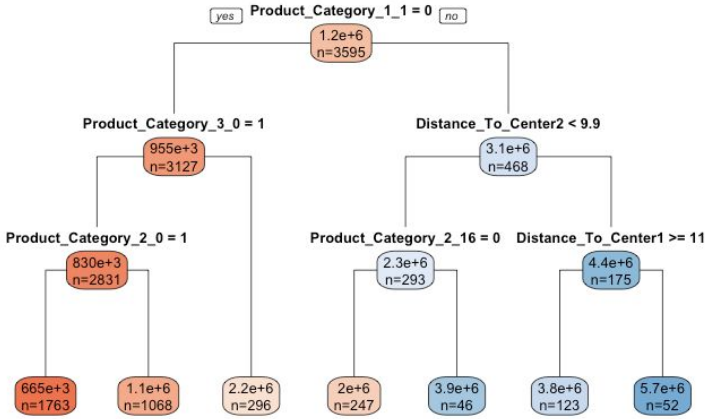
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9658450	3440247	2.807	0.005020 **
Product_Category_1_4	-1811306	655595	-2.763	0.005760 **
Product_Category_1_9	8802565	3425397	2.570	0.010216 *
Product_Category_1_10	2482696	473548	5.243	1.67e-07 ***
Product_Category_1_11	-1584745	747402	-2.120	0.034047 *
Product_Category_1_13	-1190524	456100	-2.610	0.009086 **
Product_Category_2_0	-2815779	1164871	-2.417	0.015689 *
Product_Category_2_4	-3067463	983723	-3.118	0.001834 **
Product_Category_2_8	-2401166	1055403	-2.275	0.022958 *
Product_Category_2_11	-2031318	807268	-2.516	0.011904 *
Product_Category_2_12	-1665760	699992	-2.380	0.017380 *
Product_Category_2_14	-2213977	1056268	-2.096	0.036150 *
Product_Category_3_0	-4238700	1229227	-3.448	0.000571 ***
Product_Category_3_4	-4129679	802442	-5.146	2.80e-07 ***
Product_Category_3_8	-1824864	804823	- 2.267	0.023425 *
Product_Category_3_9	-1943184	806592	-2.409	0.016041 *
Product_Category_3_10	2711707	1332322	2.035	0.041893 *
Product_Category_3_12	-1697845	799836	-2.123	0.033845 *
Product_Category_3_14	-2891726	955324	-3.027	0.002488 **
Product_Category_3_16	-3095546	1036398	-2.987	0.002838 **
Distance_To_Center1	-157902	84262	-1.874	0.061021
Distance_To_Center2	-44578	64525	-0.691	0.489699 .

Residual standard error: 1700000 on 3543 degrees of freedom

Multiple R-squared: 0.2737, Adjusted R-squared: 0.2632

F-statistic: 26.18 on 51 and 3543 DF, p-value: < 2.2e-16

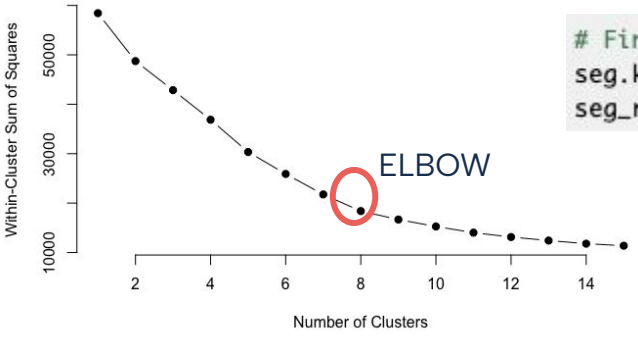
Decision Tree



R-squared: 0.2843

Model Improvement

Feature Engineering: K-Means Clustering Technique




```r
# Find the K-means groups
seg.k7 <- kmeans(normalizedData, centers=7, nstart = 50, iter.max = 20 )
seg_result <- calculateClusterMeans(clusterDataMatrix, seg.k7$cluster)
```

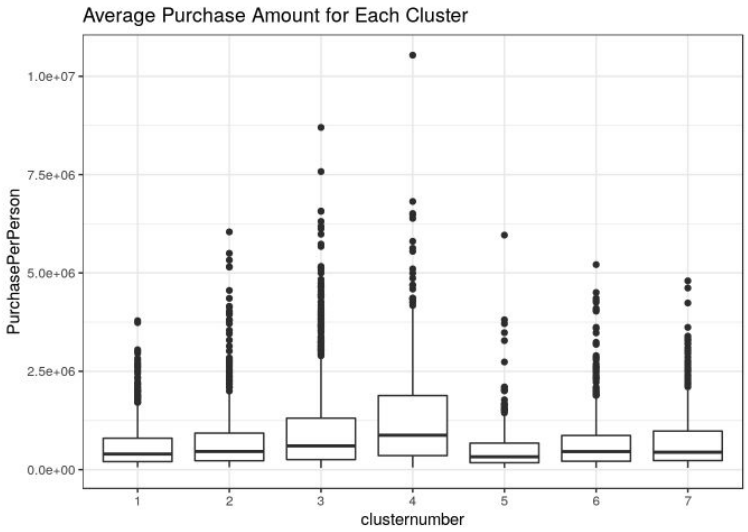
```
'data.frame':   5844 obs. of  17 variables:
 $ User_ID                   : num  1e+06 1e+06 1e+06 1e+06 1e+06 ...
 $ Gender                    : num  0 1 1 1 1 0 1 1 1 0 ...
 $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 7 3 5 3 6 4 3 3 4 ...
 $ Occupation                : Factor w/ 21 levels "0","1","10","11",..: 3 9 8 19 14 21 2 5 10 2 ...
 $ City_Category             : Factor w/ 3 levels "A","B","C": 1 3 1 2 1 1 2 3 3 2 ...
 $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 5 4 3 2 2 2 5 1 5 ...
 $ Marital_Status            : num  1 1 1 2 2 1 2 2 1 2 ...
 $ Total_Purchase            : num  333481 810353 341635 205987 822111 ...
 $ clusternumber             : Factor w/ 8 levels "1","2","3","4",..: 7 1 7 5 7 2 6 7 7 6 ...
```

Based on WSS plot, we are selecting the cluster size to be **7**

Cluster Number has been added as a feature to our User_Info Dataset to enrich the dataset with more information about users.

Model Improvement

Feature Engineering: K-Means Clustering Technique



Answering Business Questions

- The average purchase amount in group 3 and 4 are much more than the other groups (9,70,523 for group 3 and 12,94,549 for group 4).
- The customers in group 3 from all 3 cities, with similar percentage, and most of them are male and 26-35 years old; Most customers in group 4 are from city B, and most of them are 18-25 years old or 36-45 years old.
- This conclusion reminds us that our main customers can be those who from city B, with age range from 18-25 and 36-45 years old.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.57708    0.30349   1.901  0.05724 .
GenderM                    0.67802    0.07386   9.179  < 2e-16 ***
Age18-25                   0.38551    0.28133   1.370  0.17058
Age26-35                   0.59332    0.28126   2.109  0.03490 *
Age36-45                   0.51822    0.28606   1.812  0.07006 .
Age46-50                   0.21222    0.29710   0.714  0.47505
Age51-55                   0.20014    0.29988   0.667  0.50452
Age55+                     0.02558    0.30618   0.084  0.93341
Occupation1                0.34367    0.14669   2.343  0.01914 *
Occupation2                0.08768    0.18131   0.484  0.62869
Occupation3                0.37661    0.21618   1.742  0.08148 .
Occupation4                0.43808    0.15113   2.899  0.00375 **
Occupation5                0.21675    0.26446   0.820  0.41244
Occupation6                0.15065    0.18616   0.809  0.41837
Occupation7                0.27234    0.13912   1.958  0.05027 .
Occupation8                0.03530    0.58963   0.060  0.95226
Occupation9               -0.26191    0.24619  -1.064  0.28739
Occupation10               0.40884    0.30145   1.356  0.17502
Occupation11               0.53399    0.27451   1.945  0.05174 .
Occupation12               0.42638    0.17655   2.415  0.01573 *
Occupation13              -0.26628    0.22379  -1.190  0.23409
Occupation14               0.15119    0.17444   0.867  0.38611
Occupation15               0.67288    0.27383   2.457  0.01400 *
Occupation16               0.16361    0.18850   0.868  0.38544
Occupation17               0.34454    0.15656   2.201  0.02776 *
Occupation18               0.07835    0.31928   0.245  0.80615
Occupation19               1.03599    0.41549   2.493  0.01265 *
Occupation20               0.29770    0.18825   1.581  0.11378
City_CategoryB             0.11010    0.11315   0.973  0.33054
City_CategoryC            -0.51520    0.09911  -5.198 2.01e-07 ***
Marital_Status1           -0.11628    0.07335  -1.585  0.11289
Stay_In_Current_City_Years1  0.02484    0.10807   0.230  0.81820
Stay_In_Current_City_Years2 -0.02640    0.11898  -0.222  0.82440
Stay_In_Current_City_Years3  0.05871    0.12450   0.472  0.63724
Stay_In_Current_City_Years4+ 0.20809    0.12905   1.612  0.10686
---
```

Logistic Regression: Analysis of customer purchase of top 10 products

- **Logistic Regression Variables:**
  - Independent Variables: Gender, Age, Occupation, Marital Status, City Category, Stay in Current City Years.
  - Dependent Variable: Best Selling Product.
- **Key Findings from Regression Analysis:**
  - Significant Variables at 5% Level: Gender (Male), Age 26-35, Occupations 1, 4, 12, 15, 17, 19, City Category 'C'.
  - Example Insight: Male users have 67.8% higher odds (log odds: 0.678) of buying the top 10 best selling products compared to female users.
- **Customer Profile for Bestselling Products:**
  - Likely Buyers: Males, aged 26-35, in specific occupations (1, 4, 12, 15, 17, 19).
  - Spending Pattern: Higher spending correlates with a greater likelihood of purchasing these products.

Linear Modelling Results

city_name	adjusted_R_Squared
All_cities	0.18500
A	0.47950
B	0.04309
C	0.02079

After adding the cluster number feature, there was a better fit on the data. Resulting R_squared value increased for regression_all and regression_cityA

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.77831    1.14564  12.027  < 2e-16 ***
Occupation1    -0.08720    0.14274  -0.611   0.5414
Occupation2     0.08394    0.15652   0.536   0.5919
Occupation3     0.14987    0.20141   0.744   0.4570
Occupation4     0.03185    0.13028   0.244   0.8069
Occupation5    -0.17911    0.29687  -0.603   0.5464
Occupation6     0.14162    0.23248   0.609   0.5425
Occupation7    -0.06187    0.13570  -0.456   0.6485
Occupation8    -0.91322    0.53062  -1.721   0.0856 .
Occupation9    -0.44610    0.40898  -1.091   0.2756
Occupation10   -0.33095    0.37277  -0.888   0.3749
Occupation11    0.07052    0.25886   0.272   0.7853
Occupation12   -0.25937    0.15302  -1.695   0.0904 .
Occupation13   -0.76427    0.33774  -2.263   0.0239 *
Occupation14    0.07495    0.17091   0.439   0.6611
Occupation15    0.01907    0.22964   0.083   0.9338
Occupation16    0.39899    0.20308   1.965   0.0497 *
Occupation17   -0.05935    0.16340  -0.363   0.7165
Occupation18    0.35441    0.40532   0.874   0.3821
Occupation19    0.14264    0.29878   0.477   0.6332
Occupation20    0.40272    0.16533   2.436   0.0150 *
GenderM         0.31876    0.07565   4.214 2.74e-05 ***
Age18-25       -0.54114    1.52135  -0.356   0.7221
Age26-35        0.20845    0.35340   0.590   0.5554
Age36-45       -0.61216    1.14105  -0.536   0.5917
Age46-50       -0.69919    1.15337  -0.606   0.5445
Age51-55       -0.89597    1.15191  -0.778   0.4369
Age55+         -1.07848    1.15668  -0.932   0.3514
clusternumber2        NA         NA      NA       NA
clusternumber3 -0.61417    1.10778  -0.554   0.5794
clusternumber4  2.09595    0.29415   7.125 1.97e-12 ***
clusternumber5        NA         NA      NA       NA
clusternumber6 -0.10031    1.08605  -0.092   0.9264
clusternumber7        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.042 on 1014 degrees of freedom
Multiple R-squared:  0.1326,  Adjusted R-squared:  0.107
F-statistic: 5.168 on 30 and 1014 DF,  p-value: < 2.2e-16
```

Answers / Findings:

We can get insights of different cities through our regressions.

For City A, male customers whose occupation is 16 and 20 will purchase more.

For City B, male customers whose occupation is 8 will purchase more.

For City C, male customers tend to purchase more than females.

Summary

- **Business Objective**: The aim of this project is to analyse consumer buying behaviour and product categories to improve profitability and efficiency on Black Friday.
- **EDA & Statistical Analysis**: The EDA indicates significant relationships between the total purchase amount and variables such as age, city category, and product category. These relationships were verified in the statistical analysis.
- **Model & Evaluation**
  - **Regression Tree:** In regression tree, the most valuable customer identified with a total estimated revenue of $ 1,470,000, and the most popular products were reveal to have $4,170,000 total purchase.
  - **Linear Regression:** Factors affecting total purchases are identified and their positive and negative impact on them can be inferred from the parameters
  - **Clustering:** Age, Gender and City was identified as key influencer. The Age group **26-35** emerged as a significant factor, suggesting a pivotal demographic for targeted strategies.
  - **Model Optimization**: Cluster-Based Feature was experimented to be used to improve model accuracy and have a significanlyy improvment in linear & tree models.

Marketing strategies for customers can be tailored for 26-35 years old male customer who live in A/B-type cities with specific occupations.

Marketing strategies for products can be tailored for the most and least popular products that fall under certain product categories based on the analysis results.