Analyze Customer Behaviour in E-Commerce Platforms

Prashant Dhungana-30080130
Prateek Kaushik-30229287
Mariya Mathews-30192182
Nisha Pillai-30158934

2023-10-17

INTRODUCTION

Over the past decade the retail landscape has undergone a profound transformation. Many traditional stores have been supplemented and replaced by online virtual storefronts also known as e-commerce platforms. E-commerce has redefined the way people shop offering convenience and accessibility. The overall objective of this project is to determine which variables are statistically significant in understanding customer behavior on an e-commerce platform. Some of the variables we analyzed are Churn Rate, Hours spent on the app, Purchase Frequency, Gender, Preferred Order Category,CouponUsed, MaritalStatus, Complain, CashbackAmount, OrderCount

Our approach consisted of first exploring the data for any missing values and replacing those values with the median of each column. To answer our guiding questions, we used various statistical approaches like linear regression model, bootstrap techniques, Pearson's Correlation Coefficient Test, chi-squared. The techniques used were appropriate and significant to provide statistical evidence to investigate and answer our questions. It is important to note that the same data set was used for the DATA-601 project and data cleaning part of the project was imported from DATA-601.

Setting Up - Importing Important Libraries

```
library(readxl)
library(dplyr)
library(ggplot2)
library(mosaic)
library(mosaicData)
library(zoo)
```

Part 1:Importing Data and Exploring

```
# Read the Excel file
customer_behaviour_df = read_excel("E Commerce Dataset.xlsx", sheet = "E Comm")

# Print the first few rows of the dataframe
head(customer_behaviour_df)
```

```
## # A tibble: 6 × 20
##   CustomerID Churn Tenure PreferredLoginDevice CityTier WarehouseToHome
##        <dbl> <dbl>  <dbl> <chr>                   <dbl>           <dbl>
## 1      50001     1      4 Mobile Phone                3               6
## 2      50002     1     NA Phone                       1               8
## 3      50003     1     NA Phone                       1              30
## 4      50004     1      0 Phone                       3              15
## 5      50005     1      0 Phone                       1              12
## 6      50006     1      0 Computer                    1              22
## # ℹ 14 more variables: PreferredPaymentMode <chr>, Gender <chr>,
## #   HourSpendOnApp <dbl>, NumberOfDeviceRegistered <dbl>,
## #   PreferedOrderCat <chr>, SatisfactionScore <dbl>, MaritalStatus <chr>,
## #   NumberOfAddress <dbl>, Complain <dbl>, OrderAmountHikeFromlastYear <dbl>,
## #   CouponUsed <dbl>, OrderCount <dbl>, DaySinceLastOrder <dbl>,
## #   CashbackAmount <dbl>
```

```
# Print information about the dataframe structure
str(customer_behaviour_df)
```

```
## tibble [5,630 × 20] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID                 : num [1:5630] 50001 50002 50003 50004 50005 ...
##  $ Churn                      : num [1:5630] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Tenure                     : num [1:5630] 4 NA NA 0 0 0 NA NA 13 NA ...
##  $ PreferredLoginDevice       : chr [1:5630] "Mobile Phone" "Phone" "Phone" "Phone" ...
##  $ CityTier                   : num [1:5630] 3 1 1 3 1 1 3 1 3 1 ...
##  $ WarehouseToHome            : num [1:5630] 6 8 30 15 12 22 11 6 9 31 ...
##  $ PreferredPaymentMode       : chr [1:5630] "Debit Card" "UPI" "Debit Card" "Debit Card" ...
##  $ Gender                     : chr [1:5630] "Female" "Male" "Male" "Male" ...
##  $ HourSpendOnApp             : num [1:5630] 3 3 2 2 NA 3 2 3 NA 2 ...
##  $ NumberOfDeviceRegistered   : num [1:5630] 3 4 4 4 3 5 3 3 4 5 ...
##  $ PreferedOrderCat           : chr [1:5630] "Laptop & Accessory" "Mobile" "Mobile" "Laptop & Accessory" ...
##  $ SatisfactionScore          : num [1:5630] 2 3 3 5 5 5 2 2 3 3 ...
##  $ MaritalStatus              : chr [1:5630] "Single" "Single" "Single" "Single" ...
##  $ NumberOfAddress            : num [1:5630] 9 7 6 8 3 2 4 3 2 2 ...
##  $ Complain                   : num [1:5630] 1 1 1 0 0 1 0 1 1 0 ...
##  $ OrderAmountHikeFromlastYear: num [1:5630] 11 15 14 23 11 22 14 16 14 12 ...
##  $ CouponUsed                 : num [1:5630] 1 0 0 0 1 4 0 2 0 1 ...
##  $ OrderCount                 : num [1:5630] 1 1 1 1 1 6 1 2 1 1 ...
##  $ DaySinceLastOrder          : num [1:5630] 5 0 3 3 3 7 0 0 2 1 ...
##  $ CashbackAmount             : num [1:5630] 160 121 120 134 130 ...
```

```
# Calculate the number of unique values for each column
unique_counts = sapply(customer_behaviour_df, function(x) n_distinct(x, na.rm = TRUE))

# Print the number of unique values for each column
cat("\nNumber of unique values for each column\n")
```

```
## 
## Number of unique values for each column
```

```
print(unique_counts)
```

```
##                  CustomerID                       Churn 
##                        5630                           2 
##                      Tenure        PreferredLoginDevice 
##                          36                           3 
##                    CityTier             WarehouseToHome 
##                           3                          34 
##        PreferredPaymentMode                      Gender 
##                           7                           2 
##              HourSpendOnApp    NumberOfDeviceRegistered 
##                           6                           6 
##            PreferedOrderCat           SatisfactionScore 
##                           6                           5 
##               MaritalStatus             NumberOfAddress 
##                           3                          15 
##                    Complain OrderAmountHikeFromlastYear 
##                           2                          16 
##                  CouponUsed                  OrderCount 
##                          17                          16 
##           DaySinceLastOrder              CashbackAmount 
##                          22                        2586
```

Part 2:Data Cleaning

Part 2.1 Dropping columns not used in the analysis

```
# Dropping columns not used in the analysis
columns_to_drop = c( "CityTier", "WarehouseToHome", "NumberOfAddress", "OrderAmountHikeFromlastYear")
customer_behaviour_df = customer_behaviour_df[, !(names(customer_behaviour_df) %in% columns_to_drop)]
cat("\n Dropped Columns are :\n")
```

```
## 
##  Dropped Columns are :
```

```
columns_to_drop
```

```
## [1] "CityTier"                    "WarehouseToHome"            
## [3] "NumberOfAddress"             "OrderAmountHikeFromlastYear"
```

```
# Print missing values distribution
cat("Missing values distribution: \n")
```

```
## Missing values distribution:
```

```
print(colMeans(is.na(customer_behaviour_df)))
```

```
##               CustomerID                    Churn                   Tenure 
##               0.00000000               0.00000000               0.04689165 
##     PreferredLoginDevice     PreferredPaymentMode                   Gender 
##               0.00000000               0.00000000               0.00000000 
##           HourSpendOnApp NumberOfDeviceRegistered         PreferedOrderCat 
##               0.04529307               0.00000000               0.00000000 
##        SatisfactionScore            MaritalStatus                 Complain 
##               0.00000000               0.00000000               0.00000000 
##               CouponUsed               OrderCount        DaySinceLastOrder 
##               0.04547069               0.04582593               0.05452931 
##           CashbackAmount 
##               0.00000000
```

Part 2.2 Cleaning Categorical Columns

```
# Function to identify categorical columns
is_categorical = function(column) {
  is.factor(column) || is.character(column)
}

# Get column names that are categorical
customer_behaviour_categorical = names(customer_behaviour_df)[sapply(customer_behaviour_df, is_categorical)]

# Print categorical column names
cat("Categorical columns are: \n")
```

```
## Categorical columns are:
```

```
print(customer_behaviour_categorical)
```

```
## [1] "PreferredLoginDevice" "PreferredPaymentMode" "Gender"              
## [4] "PreferedOrderCat"     "MaritalStatus"
```

```
print("Categorical Columns with the unique values and counts")
```

```
## [1] "Categorical Columns with the unique values and counts"
```

```
for (col in customer_behaviour_categorical) {
  cat("\n", col)
  print(table(customer_behaviour_df[[col]]))
}
```

```
## 
##  PreferredLoginDevice
##     Computer Mobile Phone        Phone 
##         1634         2765         1231 
## 
##  PreferredPaymentMode
## Cash on Delivery               CC              COD      Credit Card 
##              149              273              365             1501 
##       Debit Card         E wallet              UPI 
##             2314              614              414 
## 
##  Gender
## Female   Male 
##   2246   3384 
## 
##  PreferedOrderCat
##            Fashion            Grocery Laptop & Accessory             Mobile 
##                826                410               2050                809 
##       Mobile Phone             Others 
##               1271                264 
## 
##  MaritalStatus
## Divorced  Married   Single 
##      848     2986     1796
```

Remove Duplicates in unique Values in Categorical Columns

```
# Replace "Mobile" with "Mobile Phone" in PreferedOrderCat column
customer_behaviour_df = customer_behaviour_df %>%
  mutate(PreferedOrderCat = ifelse(PreferedOrderCat == "Mobile", "Mobile Phone", PreferedOrderCat))

# Replace "Phone" with "Mobile Phone" in PreferredLoginDevice column
customer_behaviour_df = customer_behaviour_df %>%
  mutate(PreferredLoginDevice = ifelse(PreferredLoginDevice == "Phone", "Mobile Phone", PreferredLoginDevice))

# Replace duplicates in PreferredPaymentMode column
customer_behaviour_df = customer_behaviour_df %>%
  mutate(PreferredPaymentMode = case_when(
        PreferredPaymentMode == "CC" ~ "Credit Card",
        PreferredPaymentMode == "COD" ~ "Cash on Delivery",
        TRUE ~ PreferredPaymentMode))
```

Converting two numerical Columns to Categorical for Analysis

```
# Convert 0 to "NO" and 1 to "YES" in a churn column
customer_behaviour_df = customer_behaviour_df %>%
  mutate(Churn = ifelse(Churn == 0, "NO", "YES"))

# Convert 0 to "NO" and 1 to "YES" in a Complain column
customer_behaviour_df = customer_behaviour_df %>%
  mutate(Complain = ifelse(Complain == 0, "NO", "YES"))
```

Part 2.3 Filling NA values of Numerical Data types

```
# Select numerical columns
customer_behaviour_numerical = customer_behaviour_df %>%
  select_if(is.numeric) %>%
  colnames()

# Print numerical column names
cat("Numerical columns are: \n")
```

```
## Numerical columns are:
```

```
print(customer_behaviour_numerical)
```

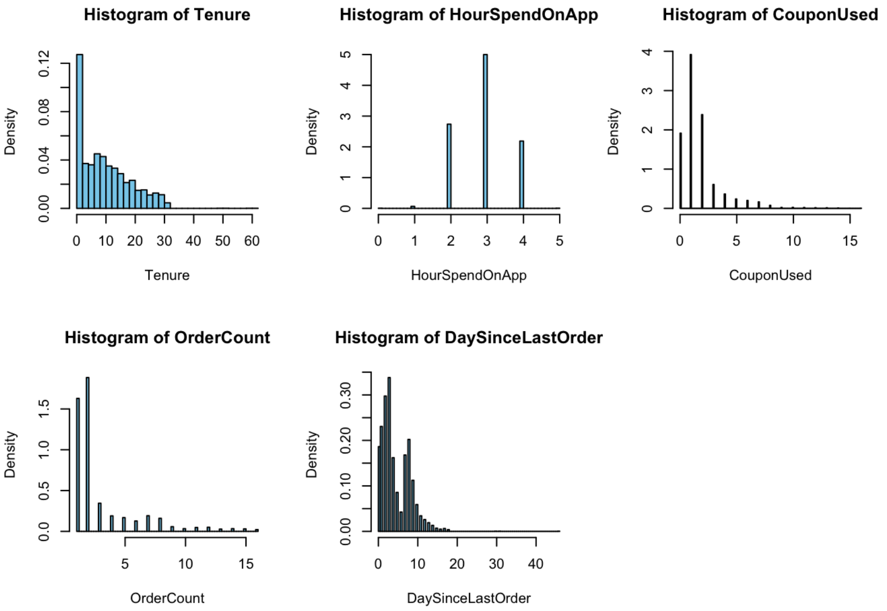
```
## [1] "CustomerID"               "Tenure"
## [3] "HourSpendOnApp"           "NumberOfDeviceRegistered"
## [5] "SatisfactionScore"        "CouponUsed"
## [7] "OrderCount"               "DaySinceLastOrder"
## [9] "CashbackAmount"
```

```
numerical_summary = summary(customer_behaviour_df[c(customer_behaviour_numerical)])

# Print the summary of numerical columns
print(numerical_summary)
```

```
##    CustomerID        Tenure      HourSpendOnApp  NumberOfDeviceRegistered
##  Min.   :50001   Min.   : 0.00   Min.   :0.000   Min.   :1.000
##  1st Qu.:51408   1st Qu.: 2.00   1st Qu.:2.000   1st Qu.:3.000
##  Median :52816   Median : 9.00   Median :3.000   Median :4.000
##  Mean   :52816   Mean   :10.19   Mean   :2.932   Mean   :3.689
##  3rd Qu.:54223   3rd Qu.:16.00   3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :55630   Max.   :61.00   Max.   :5.000   Max.   :6.000
##                  NA's   :264     NA's   :255
##  SatisfactionScore   CouponUsed       OrderCount     DaySinceLastOrder
##  Min.   :1.000     Min.   : 0.000   Min.   : 1.000   Min.   : 0.000
##  1st Qu.:2.000     1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median :3.000     Median : 1.000   Median : 2.000   Median : 3.000
##  Mean   :3.067     Mean   : 1.751   Mean   : 3.008   Mean   : 4.543
##  3rd Qu.:4.000     3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 7.000
##  Max.   :5.000     Max.   :16.000   Max.   :16.000   Max.   :46.000
##                    NA's   :256      NA's   :258      NA's   :307
##  CashbackAmount
##  Min.   :  0.0
##  1st Qu.:145.8
##  Median :163.3
##  Mean   :177.2
##  3rd Qu.:196.4
##  Max.   :325.0
##
```

Understand the distribution of Data

```
# Select the specified numerical columns
columns_to_plot = c("Tenure", "HourSpendOnApp", "CouponUsed", "OrderCount", "DaySinceLastOrder")

# Create a multi-panel plot with histograms for the selected columns
par(mfrow = c(2, 3))  # 2 rows, 3 columns for 5 histograms
for (col in columns_to_plot) {
# Remove NA values, ensure numeric data, and drop invalid entries
  non_na_data = as.numeric(na.omit(customer_behaviour_df[[col]]))

  # Check if there is valid data to plot
  if (length(non_na_data) > 0)
    {
      hist(non_na_data,
       main = paste("Histogram of", col),
       xlab = col,
       col = "skyblue",
       border = "black",
       breaks = "FD",  # "FD" for Freedman-Diaconis rule for bin width
       probability = TRUE)
  }
  else
  {
    print("No data for ",col)
  }

}
```



```
# Since we understood the skewness, fill NA with Median values using na.aggregate function from zoo libarary
for (col in customer_behaviour_numerical) {
  if (any(is.na(customer_behaviour_df[[col]]))) {
    customer_behaviour_df[[col]] = na.aggregate(customer_behaviour_df[[col]], FUN = median)
  }
}
cat("Filling the Null values in numerical columns with Median as the distribution of data is skewed")
```

```
## Filling the Null values in numerical columns with Median as the distribution of data is skewed
```

Part 3: Data Analysis and Visualization

Question-1: Linear Regression

Can the number of hours spent on the app predict the purchase frequency of a user??

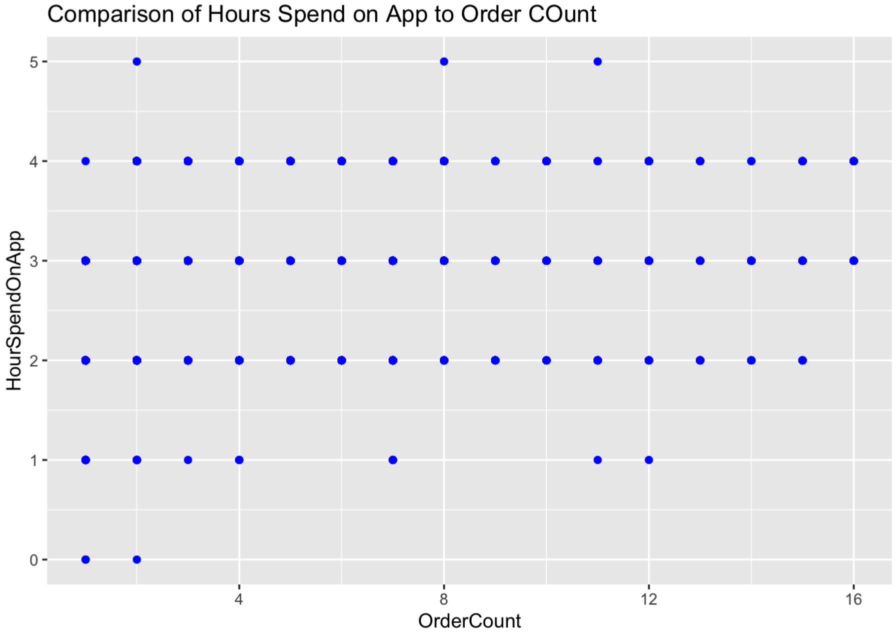
H_0 : There is no relationship between hours spent on the app and the order count

H_A : There is a relationship between hours spent on the app and the order count

Data Visualization

Lets visualize the relationship between the two variables

```
ggplot(data=customer_behaviour_df,aes(x=OrderCount,y=HourSpendOnApp))+geom_point(col="blue")+ggtitle("Comparison
of Hours Spend on App to Order COunt")
```



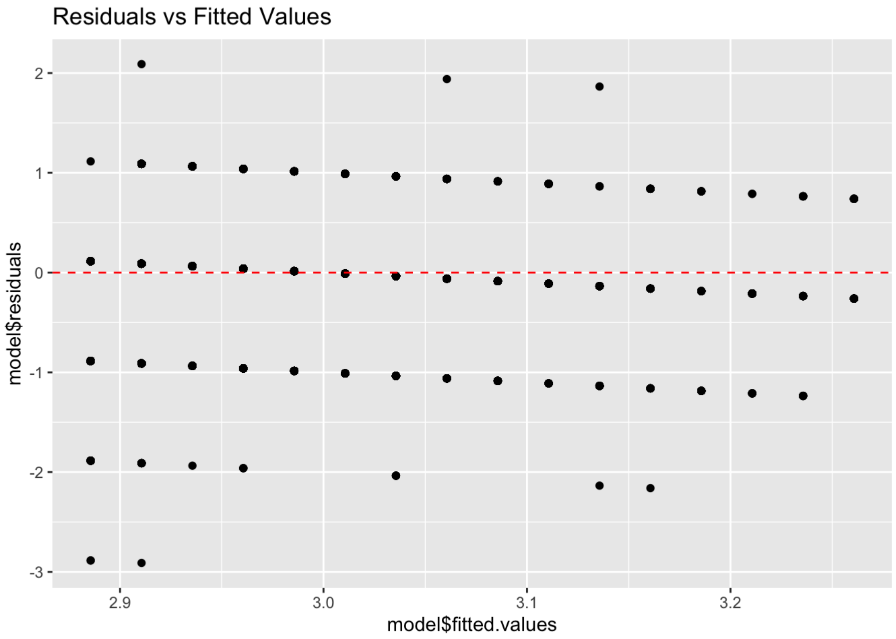
From the scatter plot we can observe there does not appear to be a strong linear relationship between the variables as the data points are spread across the plot without showing a clear upward or downward trend. In order to conclude if these two variables are statistically significant we compute a linear model.

Estimating the model:

```
model=lm(HourSpendOnApp~OrderCount,data=customer_behaviour_df)
```

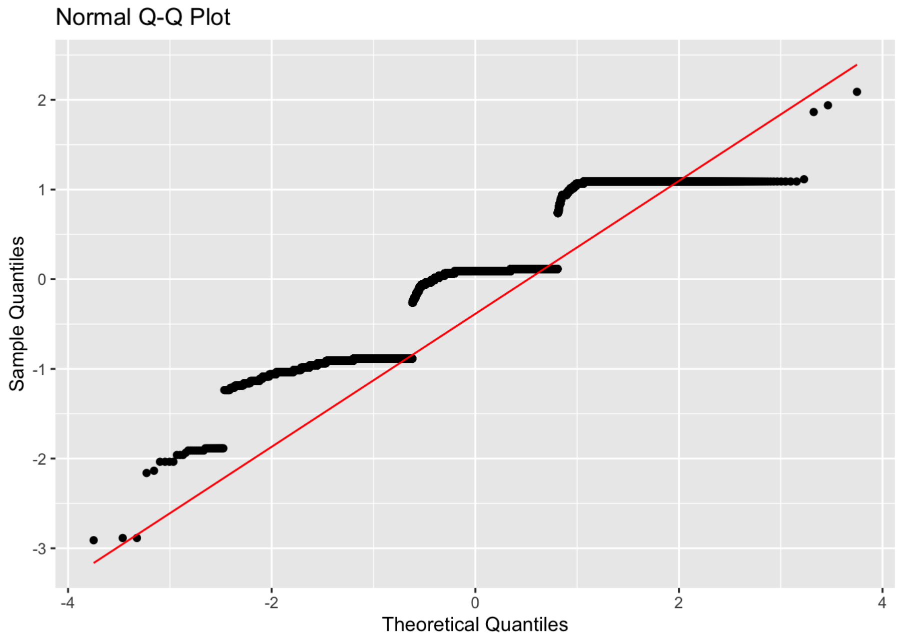
Condition Checking:

```
#First Assumption: Residuals vs Fitted Values
ggplot(model,aes(x=model$fitted.values, y=model$residuals))+geom_point()+geom_hline(yintercept = 0, color = "re
d", linetype = "dashed")+ggtitle("Residuals vs Fitted Values")
```



From the “Residuals vs Fitted Values Plot” we observe that the points are scattered around the horizontal line at zero without any specific pattern suggesting the assumption of linearity might hold. Additionally, the residuals are fairly spread across different fitted values, suggesting homoscedasticity might also be met.

```
#Second Condition: Checking for Normality
ggplot(model,aes(sample=model$residuals))+geom_qq()+geom_qq_line(colour = "red") +ggtitle("Normal Q-Q Plot")+xlab
("Theoretical Quantiles")+ylab("Sample Quantiles")
```



From the normality plot above we observe the point do somewhat follow the red line however, there is a noticeable deviation from normality mainly around the tails.

Analysis:

```
summary(model)
```

```
##
## Call:
## lm(formula = HourSpendOnApp ~ OrderCount, data = customer_behaviour_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91058 -0.88558  0.08942  0.11442  2.08942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.860570   0.013421 213.141  < 2e-16 ***
## OrderCount  0.025007   0.003249   7.696 1.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7019 on 5628 degrees of freedom
## Multiple R-squared:  0.01041,    Adjusted R-squared:  0.01024
## F-statistic: 59.23 on 1 and 5628 DF,  p-value: 1.645e-14
```

From the summary(model) we can observe: HourSpendOnAPP= 2.8606+0.0250*OrderCOunt The intercept Beta(0)=2.8606 indicating the expected hours spend on the app when the order count is 0. The slope Beta(1)=0.0250 suggesting that for every additional order, the hours spent on the app increase by 0.025 hours on average. The R-squared=0.01041 which suggest that only 1% of the variability in hours spent on the app is explained by the order count which implies that other variables might influence the hours spent on the app.
The p-value is less than alpha=0.05, we reject the null hypothesis, indicating that the relationship between hours spent on the app and order count is statistically significant.

Inference:

The scatter plot did not show a strong linear relationship, however the regression coefficient is statistically significant. From the condition check, it was revealed that assumptions are somewhat met. Additionally, the linear regression revealed a small but a significant relationship between the hour spend on the app and order count, with a very low R-squared value. Based on this, we can conclude that there is a statistical evidence to suggest a relationship between order count and hours spend on the app however, other variables not included in the model may provide additional insights into the factors that influences the time spent on the app. Although the scatter plot didn't strongly suggest a linear relationship, the statistical analysis indicated a significant coefficient, indicating a potential link between app usage hours and order count. While our model met some assumptions, there's room for improvement. To strengthen our analysis in the future, we can explore additional variables that may influence app usage. We should consider gathering and including these factors in our model to better explain app usage behavior. Additionally, the low R-squared value highlights the need for a more comprehensive model. In upcoming research, we can work on refining our model, possibly incorporating more data or using more advanced modeling techniques to enhance its explanatory power. By doing so, we aim to gain a deeper understanding of the complex factors driving app usage.

Question-1.1 Bootstrap Analaysis for Gender Differences

Is the average “Hour Spend on the App” different between different “Gender” groups?

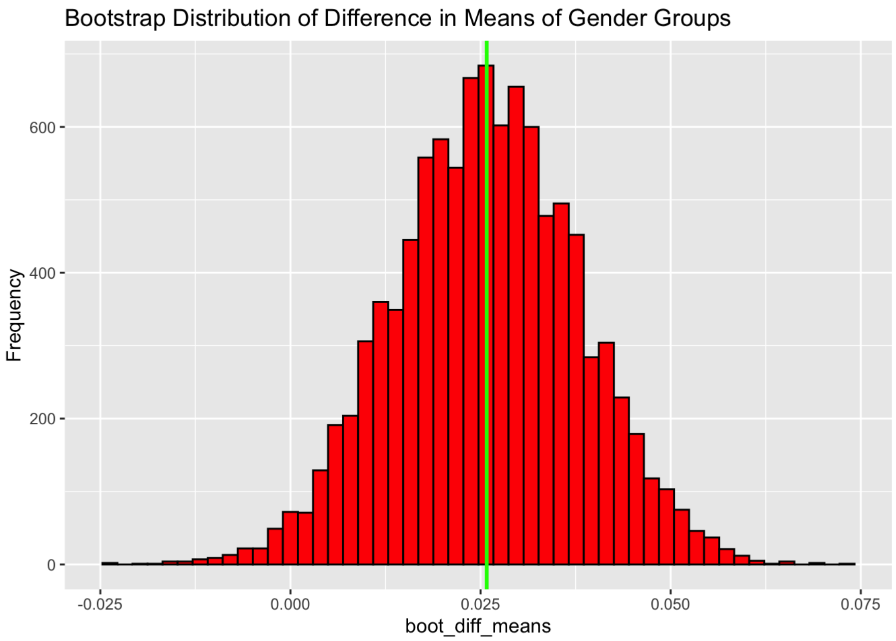
H_0 : The average hour spent on the app is same for both genders.
H_A : The average hour spent on the app is different for both genders.

```
male_data=customer_behaviour_df[customer_behaviour_df$Gender=="Male",]$HourSpendOnApp
female_data=customer_behaviour_df[customer_behaviour_df$Gender=="Female",]$HourSpendOnApp
observed_diff=mean(female_data)-mean(male_data)

N_iterations=10000
boot_diff_means=numeric(N_iterations)

set.seed(42)
for(i in 1:N_iterations){
  sample_male=sample(male_data,length(male_data),replace=TRUE)
  sample_female=sample(female_data,length(female_data),repalce=TRUE)

  boot_diff_means[i]=mean(sample_female)-mean(sample_male)
}
cat("Observed Difference means of hours spend on the app between genders(Female-Male):",observed_diff)
```

```
## Observed Difference means of hours spend on the app between genders(Female-Male): 0.02578369
```

```
boot_mean_data=data.frame(boot_diff_means)
```

```
ggplot(boot_mean_data,aes(x=boot_diff_means))+geom_histogram(col="black",fill="red",bins=50)+
  geom_vline(aes(xintercept=mean(boot_diff_means)),col="green",linewidth=1)+
  ggtitle("Bootstrap Distribution of Difference in Means of Gender Groups")+ylab("Frequency")
```



The green line in the histogram indicates the observed mean difference which is about 0.0257.And the 95% CI intervals is computed below:

```
qdata(boot_diff_means,c(0.025,0.975),data=boot_mean_data)
```

```
##        2.5%       97.5% 
## 0.002143027 0.049424351
```

The 95% CI intervals suggest female spend more time on the app than male. However, there is not a strong indication that the difference in hours spend in the app is significantly higher among females.

```
p_value=sum(abs(boot_diff_means) >= abs(observed_diff)) / length(boot_diff_means)
cat("\n p-value :",p_value)
```

```
## 
##  p-value : 0.5098
```

Since, the p_value is greater than alpha=0.05 we fail to reject the null hypothesis.This suggest that in context to our question that we do not have sufficient evidence to conclude there is a difference in the average hours spent on the app between different gender groups.

Inference

From the analysis above we did not find any significant statistical difference in hours spend on the app between males and females.Therefore, we can infer that gender may not significantly influence the hours spend on the app and understanding customer segments and marketing to these different segments can enhance customer satisfaction. While the effect size is relatively small, to strengthen our future statistical analyses, we will further investigate the underlying factors contributing to this gender-based distinction, such as marketing strategies or user preferences, in order to derive more meaningful insights and make informed decisions.

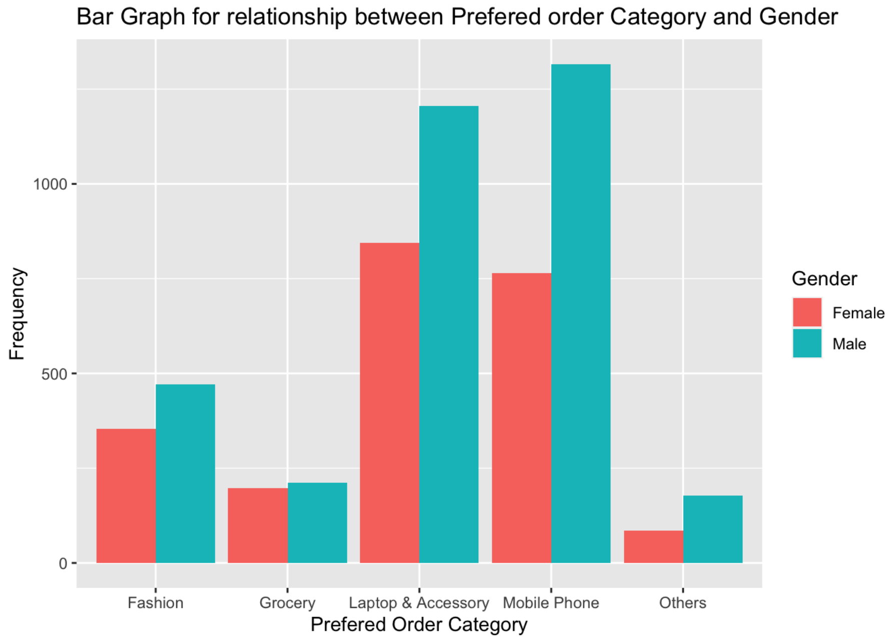
Question-2.1 - Chi-Squared test

Is there enough evidence to suggest Order Category Preference is independent of Gender?

Data Visualization

Let’s first visualize the relationship between two categorical variables

```
ggplot(data = customer_behaviour_df, aes(x = PreferedOrderCat, fill = Gender)) +
  geom_bar(position = "dodge") + xlab("Prefered Order Category") + ylab("Frequency") +ggtitle("Bar Graph for rela
tionship between Prefered order Category and Gender")
```



To illustrate the trend, we created a grouped bar chart for each category. The data reveals that Laptops, Accessories, and Mobile Phones are the top choices for both genders. Furthermore, it is evident from the data that males outnumber females in all categories.

Test Of Independence - Chi-Squared test The chi-squared test is a statistical test used to determine whether there is a significant association between two categorical variables in a data set. In this case, we are trying to find out the association between Preferred Order Category and Gender

Assumption

Assumption 1: Both variables are categorical.
Assumption 2: All observations are independent.(Used random sampling method)
Assumption 3: Cells in the contingency table are mutually exclusive.
Assumption 4: Expected value of cells should be 5 or greater

Step 1: We consider the statistical hypotheses:

$$\begin{aligned} H_0: &\quad \text{Order Category Preferrence is Independent of Gender} \\ H_A: &\quad \text{Order Category Preferrence is NOT Independent of Gender} \end{aligned}$$

Step 2: From the assumed state of the world of independence between these two categorical variables we compute the contingency table:

```
prefercategory_gender_table = table(customer_behaviour_df$PreferedOrderCat,customer_behaviour_df$Gender)
prefercategory_gender_table
```

```
##
##                        Female Male
##    Fashion                354  472
##    Grocery                198  212
##    Laptop & Accessory     844 1206
##    Mobile Phone           764 1316
##    Others                  86  178
```

The contigency table above is the observed frequencies for Male and Female in 5 Order Categories.
In the provided table, you can observe the “Observed frequencies.” The chi-squared test calculates expected frequencies assuming independence between variables. Once these expected frequencies are determined, the chi-squared test statistic measures the disparity between expected and observed frequencies. This statistic quantifies how much the expected values deviate from the actual counts.

Next, the critical value is computed based on the chosen significance level and degrees of freedom. If the calculated chi-squared statistic surpasses this critical value, it leads to the rejection of the null hypothesis.

Step 3: Compute the complete contingency table, χ2 test statistic, p-value

```
xchisq.test(prefercategory_gender_table, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 31.055, df = 4, p-value = 2.983e-06
##
##    354      472
## ( 329.52) ( 496.48)
##  [1.82]   [1.21]
## < 1.35>  <-1.10>
##
##    198      212
## ( 163.56) ( 246.44)
##  [7.25]   [4.81]
## < 2.69>  <-2.19>
##
##    844     1206
## ( 817.82) (1232.18)
##  [0.84]   [0.56]
## < 0.92>  <-0.75>
##
##    764     1316
## ( 829.78) (1250.22)
##  [5.22]   [3.46]
## <-2.28>  < 1.86>
##
##     86      178
## ( 105.32) ( 158.68)
##  [3.54]   [2.35]
## <-1.88>  < 1.53>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

```
cat("critical value of chi-squared for 4 degrees of freedom and 5% level of significance\n")
```

```
## critical value of chi-squared for 4 degrees of freedom and 5% level of significance
```

```
criticalvalue = qchisq(0.95,4)
criticalvalue
```

```
## [1] 9.487729
```

Inference

Expected frequencies (shown within() in the above summary), is calculated based on independence and all of these expected frequencies are above 5 and so it meets the criteria for Chi-squared test. Degrees of freedom is 4 which refer to the number of values in the final calculation of a statistic that are free to vary.In other words,there are 4 independent frequencies in the contigency table and other frequencies depend on these 4 independent frequencies.

Conclusion

chi-Squared test statistic is 31.055 which is higher than the critical value of chi-squared for 4 degrees of freedom.(9.487729) (Also, p-value is too low, “0.000002983” less than 0.05 (at 5 % level of significance)) So,we reject the null hypothesis at the 0.05 significance level, indicating that there is a significant association between Preferred Order Category and Gender in the population. The chi-squared statistic of 31.055 with 4 degrees of freedom provides strong evidence against the null hypothesis. Therefore, we conclude that Preferred Order Category and Gender are dependent variables. The p-value of 0.000002983 indicates an extremely low probability of observing such an extreme association between the variables by chance alone. These findings suggest a significant relationship between Preferred Order Category and Gender.

Question-2.2 - Chi-Squared test

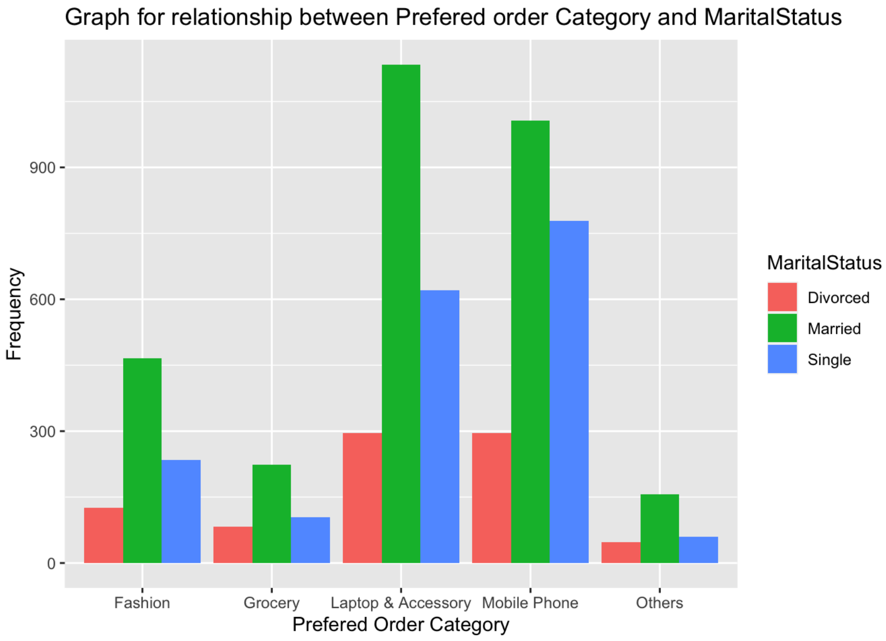
Is there enough evidence to suggest Order Category Preference is independent of

Marital Status?

Data Visulaization

Let’s first visualize the relationship between two categorical variables

```
ggplot(data = customer_behaviour_df, aes(x = PreferedOrderCat, fill = MaritalStatus)) +
  geom_bar(position = "dodge") + xlab("Prefered Order Category") + ylab("Frequency") +ggtitle("Graph for relation
ship between Prefered order Category and MaritalStatus")
```



To illustrate the trend, we created a grouped bar chart for each category. The data reveals that Laptops, Accessories, and Mobile Phones are the top choices for all marital status. Furthermore, it is evident from the data that married outnumber single and divorced in all categories.

Test Of Independence - Chi-Squared test The chi-squared test is a statistical test used to determine whether there is a significant association between two categorical variables in a data set. In this case, we are trying to find out the association between Preferred Order Category and Marital Status

Assumption

Assumption 1: Both variables are categorical.
Assumption 2: All observations are independent.(Used random sampling method)
Assumption 3: Cells in the contingency table are mutually exclusive.
Assumption 4: Expected value of cells should be 5 or greater

Step 1: We consider the statistical hypotheses:

$$\begin{array}{ll} H_0: & \text{Order Category Preferrence is Independent of Marital Status} \\ H_A: & \text{Order Category Preferrence is NOT Independent of Marital Status} \end{array}$$

Step 2: From the assumed state of the world of independence between these two categorical variables we compute the contigency table:

```
prefercategory_maritalStatus_table = table(customer_behaviour_df$PreferedOrderCat,customer_behaviour_df$MaritalSt
atus)
prefercategory_maritalStatus_table
```

```
##
##                        Divorced Married Single
##    Fashion                  126     466    234
##    Grocery                   82     224    104
##    Laptop & Accessory       296    1134    620
##    Mobile Phone             296    1006    778
##    Others                    48     156     60
```

The contigency table above is the observed frequencies for Divorced,Married and Single in 5 Order Categories. In the provided table, you can observe the “Observed frequencies.” The chi-squared test calculates expected frequencies assuming independence between variables. Once these expected frequencies are determined, the chi-squared test statistic measures the disparity between expected and observed frequencies. This statistic quantifies how much the expected values deviate from the actual counts.

Next, the critical value is computed based on the chosen significance level and degrees of freedom. If the calculated chi-squared statistic surpasses this critical value, it leads to the rejection of the null hypothesis.

Step 3: Compute the complete Contigency table, χ2 test statistic, p-value

```
xchisq.test(prefercategory_maritalStatus_table, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 61.48, df = 8, p-value = 2.386e-10
##
##    126      466      234
## ( 124.41) ( 438.09) ( 263.50)
## [ 0.02]  [ 1.78]  [ 3.30]
## < 0.14>  < 1.33>  <-1.82>
##
##     82      224      104
## (  61.75) ( 217.45) ( 130.79)
## [ 6.64]  [ 0.20]  [ 5.49]
## < 2.58>  < 0.44>  <-2.34>
##
##    296     1134      620
## ( 308.77) (1087.26) ( 653.96)
## [ 0.53]  [ 2.01]  [ 1.76]
## <-0.73>  < 1.42>  <-1.33>
##
##    296     1006      778
## ( 313.29) (1103.18) ( 663.53)
## [ 0.95]  [ 8.56]  [19.75]
## <-0.98>  <-2.93>  < 4.44>
##
##     48      156       60
## (  39.76) ( 140.02) (  84.22)
## [ 1.71]  [ 1.82]  [ 6.96]
## < 1.31>  < 1.35>  <-2.64>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

```
cat("critical value of chi-squared for 8 degrees of freedom and 5% level of significance \n")
```

```
## critical value of chi-squared for 8 degrees of freedom and 5% level of significance
```

```
criticalvalue = qchisq(0.95,8)
criticalvalue
```

```
## [1] 15.50731
```

Inference

Expected frequencies (shown within() in the above summary), is calculated based on independence and all of these expected frequencies are above 5 and so it meets the criteria for Chi-squared test. Degrees of freedom is 8 which refer to the number of values in the final calculation of a statistic that are free to vary.In other words,there are 8 independent frequencies in the contingency table and rest frequencies depend on these 8 independent frequencies.

Conclusion

At 5% level of significance, chi-Squared test statistic is 61.48 which is higher than the critical value of chi-squared for 8 degrees of freedom. (15.50731) (Also, p-value is too low, “0.0000000002386” less than 0.05) So, we reject Null hypothesis. in other words, we do not have enough evidence to suggest that Order Category Preference is Independent of MaritalStatus

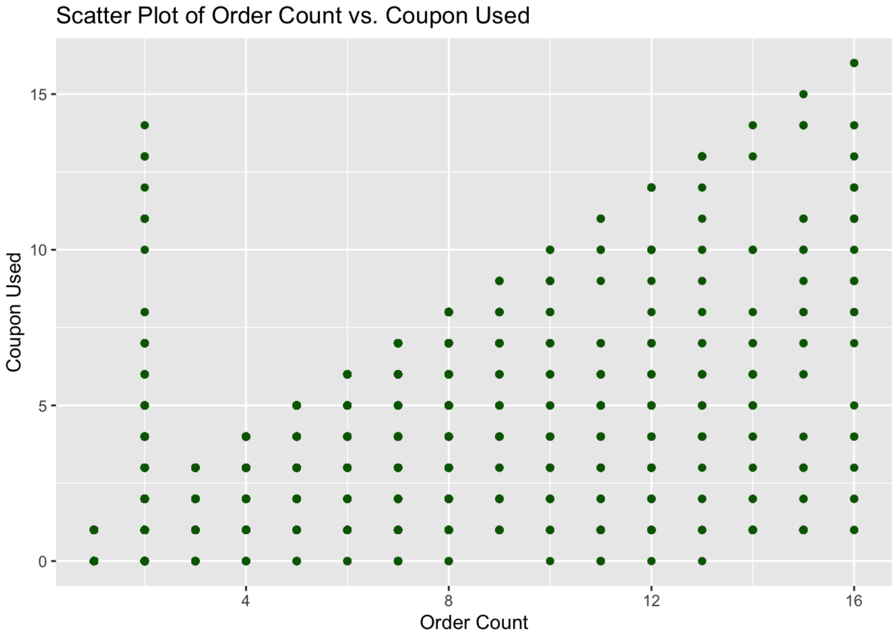
Question-3.1 - Pearson’s Correlation Coefficient Test

Do we have any correlation with the number of orders and number of coupons?

Let us first understand that we tend towards finding relationship between the Coupon Used Vs Order Count. If any particular customer is given coupons to make a purchase then is it possible that the Order Count will increase or decrease or stays the same? To determine this, lets first draw a scatter plot graph to check and we’ll deduce our interpretation of it afterwards.

Data Visualization

```
# Create a scatter plot
ggplot(data = customer_behaviour_df, aes(x = OrderCount, y = CouponUsed)) +
  geom_point(color='darkgreen') +
  labs(x = "Order Count", y = "Coupon Used") +
  ggtitle("Scatter Plot of Order Count vs. Coupon Used")
```



As per the scatterplot designed above, we can infer that there is a direct and a very strong relationship between the OrderCount and CouponUsed because as we move upwards and right, the order count increases exponentially and is a very strong indication that if businesses implements a strategy to provide coupons to customers to make a purchase then the Order Volume is expected to grow significantly.

Let us check the correlation between CouponUsed Vs OrderCount

Pearson's Correlation Coefficient Test

Step 1: We consider the statistical hypotheses:

H_0 : There is no significant correlation between the number of OrderCount and the number of CouponUsed
H_A : There is a significant correlation between the number of OrderCount and the number of CouponUsed

Before proceeding with Correlation test, we have ensured the following Assumptions:

Assumption 1: Data Type: The variables under consideration are numeric and continuous, suitable for Pearson correlation analysis.

Assumption 2: Absence of Extreme Outliers: The data does not contain extreme outliers that could unduly influence the results or distort the correlation coefficient.

Assumption 3: Approximate Normal Distribution: The data follows an approximately normal distribution, particularly when dealing with small sample sizes. This is important as deviations from normality can impact the validity of the Pearson correlation test.

Statistical Data Analysis

```
# Perform Pearson's correlation test
correlation_check <- cor.test(customer_behaviour_df$OrderCount, customer_behaviour_df$CouponUsed)

cat('Correlation Test Results: ')
```

```
## Correlation Test Results:
```

```
print(correlation_check)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 62.681, df = 5628, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6255325 0.6563075
## sample estimates:
##       cor
## 0.6411777
```

Inference

The statistical test using Pearson's Correlation test, we can infer that the t-statistic value is 62.681 stating very strong relationship between OrderCount and CouponUsage. We have degree of freedom, df = 5628 and p-value is significantly small and below the significance level proving we have strong evidence against the null hypothesis and we can reject the null hypothesis. Also, we have computed 95 Percent Confidence Interval stating that estimated correlation coefficient falls within the range of 0.6255325 to 0.6563075 with 95% confidence. The given interval is very narrow and we can be 95% confident to generate precised results. The coefficient of correlation is 0.6411777. The closer the absolute value is to 1, the stronger the correlation.

Conclusion

As per our Pearson's correlation test, we can confidently say that there's a very strong relationship between Order Count and Coupon Used. With a very small p-value which is 2.2e-16, stating that we can reject the null hypothesis (we reject the null hypothesis if the significance value is smaller than 0.05) because we have high correlation between the Order Count and Coupon Used. On the other hand,the 95 percent confidence interval for the correlation coefficient falls between approximately 0.6255 and 0.6563. This interval provides a range of values within which the true population correlation is likely to lie.

Question -3.2 - Pearson's Correlation Coefficient Test

Can we find any relationship between the number of orders and cashback amount received?

To determine whether we have any correlation between the Order Count vs Cash back Amount, we need to understand if there's any possible correlation between the two given columns. If any particular e-commerce website is offering certain amount of Cash back Amount on every purchase they've made from their website, then it is possible that there could be a linear relationship between the two. For i.e if the Cash back Amount increases, then the Order Count is also expected to grow significantly. Let us first understand this through Data Visualization

Pearson's Correlation Coefficient Test

Step 1: We consider the statistical hypotheses:

H_0 : There is no significant correlation between the number of orders ("OrderCount") and the cashback amount received ("CashbackAmount"

H_A : There is a significant correlation between the number of orders ("OrderCount") and the cashback amount received ("CashbackA

Before proceeding with statistical tests, we have ensured the following Assumptions

Assumption 1: Data Type: The variables under consideration are numeric and continuous, suitable for Pearson correlation analysis.

Assumption 2: Absence of Extreme Outliers: The data does not contain extreme outliers that could unduly influence the results or distort the correlation coefficient.

Assumption 3: Approximate Normal Distribution: The data follows an approximately normal distribution, particularly when dealing with small sample sizes. This is important as deviations from normality can impact the validity of the Pearson correlation test.

Statistical Data Analysis

```
# Perform Pearson's correlation test
correlation_check1 <- cor.test(customer_behaviour_df$OrderCount, customer_behaviour_df$CashbackAmount)

cat('Correlation Test Results: ')
```

```
## Correlation Test Results:
```

```
print(correlation_check1)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  x and y
## t = 25.543, df = 5628, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2987055 0.3455259
## sample estimates:
##       cor 
## 0.3223128
```

Inference

The Pearson correlation test with a t-value of 25.543 and 5628 degrees of freedom indicates a strong positive correlation between two variables. The minuscule p-value (< 2.2e-16) strongly rejects the null hypothesis, confirming the presence of a significant correlation. The 95 percent confidence interval of 0.2987055 to 0.3455259 further emphasizes this finding. With a sample estimate of approximately 0.322, it is evident that there is a substantial positive correlation between the variables under examination.

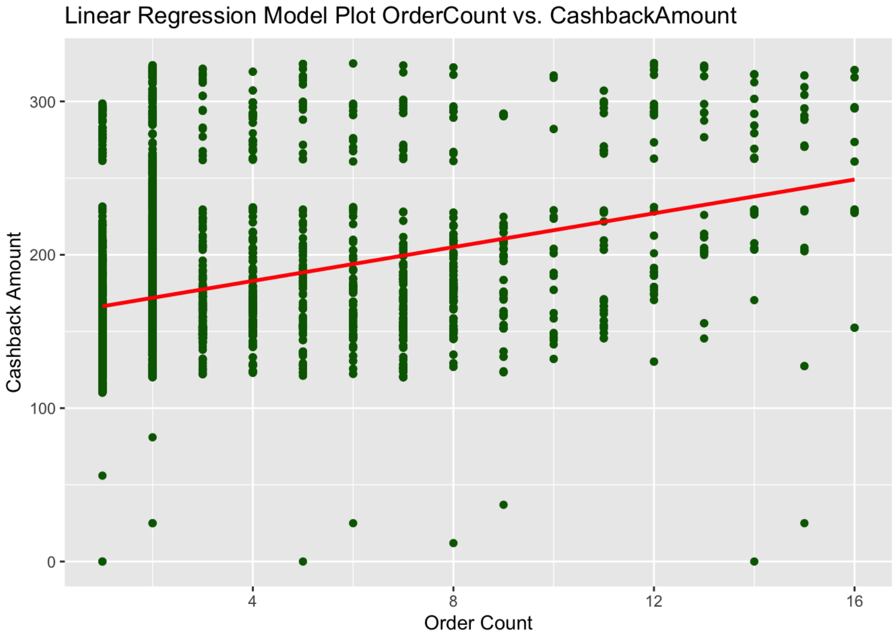
Conclusion

We would reject the null hypothesis (H0) in favor of the alternative hypothesis (H1). This suggests that there is a significant and strong correlation between the number of orders ("OrderCount") and the cashback amount received ("CashbackAmount"). The p-value of 2.2e-16 indicates that the correlation is highly statistically significant. In summary, a p-value of 2.2e-16 provides very strong evidence against the null hypothesis and supports the conclusion that there is a significant relationship between "OrderCount" and "CashbackAmount." The correlation between these two variables is likely to be highly significant.

Statistical Data Analysis

```
#Performing linear regression model
lm_model <- lm(CashbackAmount ~ OrderCount, data = customer_behaviour_df)
```

```
# Overlay the regression line on the scatter plot
ggplot(data = customer_behaviour_df, aes(x = OrderCount, y = CashbackAmount)) +
  geom_point(color='darkgreen') +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "red") +
  labs(x = "Order Count", y = "Cashback Amount") +
  ggtitle("Linear Regression Model Plot OrderCount vs. CashbackAmount")
```



As per the scatter plot given above with a fitted linear regression model, we have seen an upward trend in comparison of Order Count Vs Cash back Amount represented by the red line. As the maximum proportion of orders lies within 100 to 300 for Cash back Amount and the Order Count displayed here is also significantly higher.

```
summary(lm_model)
```

```
##
## Call:
## lm(formula = CashbackAmount ~ OrderCount, data = customer_behaviour_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.03  -29.91  -12.96   17.61  151.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 160.9082     0.8907  180.64   <2e-16 ***
## OrderCount    5.5084     0.2157   25.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.59 on 5628 degrees of freedom
## Multiple R-squared:  0.1039, Adjusted R-squared:  0.1037
## F-statistic: 652.4 on 1 and 5628 DF,  p-value: < 2.2e-16
```

Inference/Conclusion

The presented result is from a linear regression model analysis. The “Residuals” section shows the distribution of the prediction errors, ranging from -238.03 to 151.66, with a median error of -12.96. The “Coefficients” section provides estimates for the model’s intercept and “OrderCount” coefficient. A one-unit increase in “OrderCount” leads to a 5.5084-unit change in “CashbackAmount,” and both coefficients are highly significant. The “Residual standard error” represents the average error magnitude. The “Multiple R-squared” and “Adjusted R-squared” indicate the model’s goodness of fit, with 0.1039 and 0.1037, respectively. The “F-statistic” assesses the model’s significance, and the very low p-value (< 2.2e-16) suggests the model’s overall significance and value in explaining the relationship between the variables.

Statistical Data Analysis

```
# Define a function to compute the correlation coefficient
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:mosaic':
##
##     logit
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
correlation_fn <- function(data, indices) {
  sampled_data <- data[indices, ]
  correlation <- cor(sampled_data$OrderCount, sampled_data$CashbackAmount)
  return(correlation)
}

# Set the number of bootstrap resamples
num_resamples <- 1000

# Perform bootstrapping
set.seed(123)  # For reproducibility
boot_results <- boot(data = customer_behaviour_df, statistic = correlation_fn, R = num_resamples)

# Calculate the confidence interval
boot_ci <- boot.ci(boot_results, type = "basic")

# View the confidence interval
print(boot_ci)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_results, type = "basic")
##
## Intervals :
## Level      Basic
## 95%   ( 0.2923,  0.3511 )
## Calculations and Intervals on Original Scale
```

Inference/Conclusion

The bootstrap analysis indicates that the 95% confidence interval for the correlation coefficient (Pearson’s r) between “OrderCount” and “CashbackAmount” is between 0.2923 and 0.3511. This means that we can be 95% confident that the true correlation in the population falls within this interval.

In simpler terms, it suggests that there is a statistically significant positive correlation between the number of orders and the cashback amount received, and the correlation is estimated to be within the specified interval. This information provides a measure of the relationship’s strength and the uncertainty associated with the estimate.

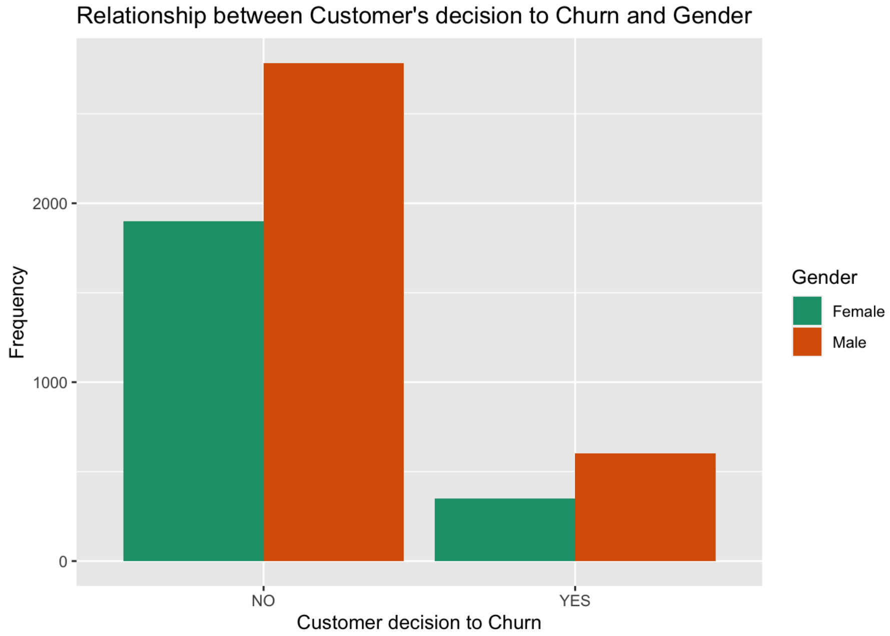
Question-4.1 - Chi-Squared test

Is there enough evidence to suggest Churn is independent of Gender?

Data Visulaization

Let’s first visualize the relationship between two categorical variables

```
ggplot(data = customer_behaviour_df, aes(x = Churn, fill = Gender)) +geom_bar(position ="dodge") +scale_color_brewer(palette="Dark2") +scale_fill_brewer(palette="Dark2") + xlab("Customer decision to Churn") + ylab("Frequency") +ggtitle("Relationship between Customer's decision to Churn and Gender")
```



To illustrate the trend, we created a grouped bar chart for each category. The data reveals that of the population who have decided to churn, it is evident that male gender is more likely to churn than females.

Test Of Independence - Chi-Squared test The chi-squared test is a statistical test used to determine whether there is a significant association between two categorical variables in a data set. In this case, we are trying to find out the association between Churn and Gender

Assumption

Assumption 1: Both variables are categorical.
Assumption 2: All observations are independent.(Used random sampling method)
Assumption 3: Cells in the contingency table are mutually exclusive.
Assumption 4: Expected value of cells should be 5 or greater

Step 1: We consider the statistical hypotheses:

$$
\begin{aligned}
H_0 &: \quad \text{(There is NO relationship between Gender and Decision to churn)} \\
H_A &: \quad \text{(There is a relationship between Gender and Decision to churn)}
\end{aligned}
$$

Step 2: From the assumed state of the world of independence between these two categorical variables we compute the contingency table:

```
tableofcounts=tally(~ Churn | Gender, data=customer_behaviour_df)
tableofcounts
```

```
##      Gender
## Churn Female Male
##   NO    1898 2784
##   YES    348  600
```

The contingency table above is the observed frequencies for Male and Female between Churn Yes & No.
In the provided table, you can observe the “Observed frequencies.” The chi-squared test calculates expected frequencies assuming independence between variables. Once these expected frequencies are determined, the chi-squared test statistic measures the disparity between expected and observed frequencies. This statistic quantifies how much the expected values deviate from the actual counts.

Next, the critical value is computed based on the chosen significance level and degrees of freedom. If the calculated chi-squared statistic surpasses this critical value, it leads to the rejection of the null hypothesis.

Step 3: Compute the complete contingency table, χ2 test statistic, p-value

```
chisq.test(tableofcounts, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tableofcounts
## X-squared = 4.8213, df = 1, p-value = 0.02811
```

#critical value of chi-squared for 1 degrees of freedom and 5% level of significance

```
criticalvalue = qchisq(0.95,1)
criticalvalue
```

```
## [1] 3.841459
```

Inference/Conclusion Using the Chi-square test we can see that a higher Chi-square value shows a strong association between the categorical variables. The P-value is 0.02811. Level of significance is 0.05. Since the p- value is less than the level of significance we reject the null Hypothesis. The statistical evidence shows that there is a relationship between gender and decision to churn. It can be seen from the histogram as well that in the population who decide to churn, Males decide to churn more than females. To ascertain why males churn more than females, it's essential to conduct a comprehensive analysis and we will consider it as a future scope of work

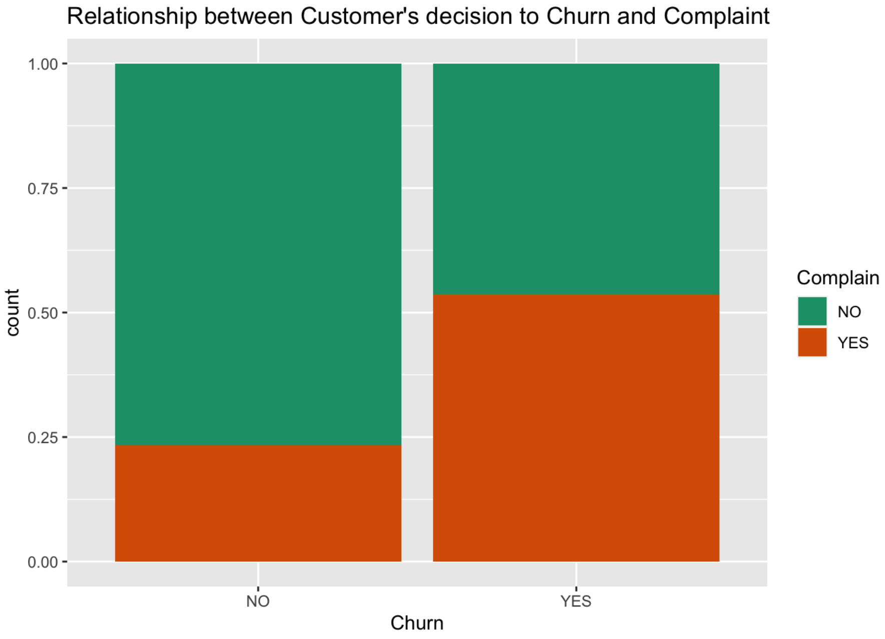
Question-4.2 - chi-Squared test

Is there enough evidence to suggest churn is independent of Complaints registered or not?**

Data Visualization

Let's first visualize the relationship between two categorical variables

```
ggplot(data=customer_behaviour_df) + geom_bar(aes(x = Churn, fill = Complain), position = "fill") + scale_color_b
rewer(palette="Dark2") + scale_fill_brewer(palette="Dark2")+ggtitle("Relationship between Customer's decision to
Churn and Complaint")
```



To illustrate the trend, we created a grouped bar chart for each category. The data reveals that customer's who decide to churn have higher number of complaints registered. Furthermore, it is evident from the data that Customers who register a complaint are more likely to churn.

Test Of Independence - Chi-Squared test The chi-squared test is a statistical test used to determine whether there is a significant association between two categorical variables in a data set. In this case, we are trying to find out the association between Churn and Complaints

Assumption

Assumption 1: Both variables are categorical.
Assumption 2: All observations are independent.(Used random sampling method)
Assumption 3: Cells in the contingency table are mutually exclusive.
Assumption 4: Expected value of cells should be 5 or greater

Step 1: We consider the statistical hypotheses:

$$\begin{aligned} H_0 &: \text{There is NO relationship between customer registering Complaint and decision to Churn} \\ H_A &: \text{There is a relationship between customer registering Complaint and decision to Churn} \end{aligned}$$

Step 2: From the assumed state of the world of independence between these two categorical variables we compute the contingency table:

```
tableofcounts1=tally(~ Churn | Complain, data=customer_behaviour_df)
tableofcounts1
```

```
##      Complain
## Churn   NO  YES
##   NO  3586 1096
##   YES  440  508
```

The contingency table above is the observed frequencies for Complaint Yes/No against Churn Yes/No.
In the provided table, you can observe the "Observed frequencies." The chi-squared test calculates expected frequencies assuming independence between variables. Once these expected frequencies are determined, the chi-squared test statistic measures the disparity between expected and observed frequencies. This statistic quantifies how much the expected values deviate from the actual counts.

Next, the critical value is computed based on the chosen significance level and degrees of freedom. If the calculated chi-squared statistic surpasses this critical value, it leads to the rejection of the null hypothesis.

Step 3: Compute the complete contingency table, χ2 test statistic, p-value

```
chisq.test(tableofcounts1, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tableofcounts1
## X-squared = 352.41, df = 1, p-value < 2.2e-16
```

#critical value of chi-squared for 1 degrees of freedom and 5% level of significance

```
criticalvalue = qchisq(0.95,1)
criticalvalue
```

```
## [1] 3.841459
```

Inference/Conclusion

Using the Chi-square test we can see that a higher Chi-square value shows a strong association between the categorical variables. The P-value is 0.00000000000000022 ~ 0. Level of significance is 0.05. Since the p- value is very less than the level of significance we reject the null Hypothesis. There is enough statistical evidence that shows that there is a relationship between customer registering a complain and decision to churn.It can be seen in the histogram well that Customers who decide to churn have registered more complaints.

Conclusion

Question 1 & 1.1

From question 1 we can conclude that there is a statistical evidence to suggest that there is a linear relationship between the “Order Count” and “Hours Spend on the App” however, the low R-Squared suggests that other variables may influence the “Hour Spend on the App’. This is due to only 1% variability of”Hour Spend on the App” is explained by the “Order Count” variable. Additionally from the bootstrap method we did not find any significant statistical difference in hours spend on the app between males and females. Females spend more time on the app in comparison to males however,there is not a strong indication that the difference in hours spend in the app is significantly higher among females. Other variables not included in the model likely play a role in influencing app usage hours. For a more comprehensive grasp of the underlying factors, future research should delve deeper into customer segments and their particular preferences, while also considering the inclusion of additional variables and the application of advanced modeling techniques to enrich our comprehension of app usage behavior. This approach will enable us to make more informed decisions, fine-tune marketing strategies, and, in the end, enhance overall customer satisfaction.

Question 2.1

In conclusion, the results of the Pearson’s Chi-squared test conducted to analyze the relationship between preferred order category and gender in the given data(X-squared = 31.055, df = 4, p-value = 0.000002983) indicate a statistically significant association between these two variables. The low p-value (p = 0.000002983) suggests that the observed distribution of preferred order categories is unlikely to have occurred by chance alone. Therefore, there is substantial evidence to support the hypothesis that there is a significant relationship between preferred order category and gender in the studied population.

Question 2.2

The results of the Chi-squared test clearly show that there is a strong connection between preferred order category and marital status. The extremely low p-value (0.0000000002386) indicates that this relationship is not random chance; it’s a meaningful and significant finding. In practical terms, it means that people’s marital status significantly influences the way they prefer to order things. This information is valuable for businesses, social researchers, and policymakers, as it provides concrete evidence that marital status plays a significant role in shaping consumer preferences and behaviors.

Question 3.1

The Pearson’s Correlation test strongly suggests a significant relationship between OrderCount and CouponUsage, with a high t-statistic of 62.681 and an extremely low p-value of 2.2e-16, allowing us to confidently reject the null hypothesis. The narrow 95 percent confidence interval, spanning from 0.6255 to 0.6563, adds further precision, and the correlation coefficient of 0.6411777 signifies a robust correlation. In conclusion, this analysis provides compelling evidence that OrderCount and CouponUsage are indeed strongly correlated in the dataset.

Question 3.2

The Pearson correlation test confirms a strong positive correlation between two variables with a high t-value (25.543) and extremely low p-value (< 2.2e-16). The 95% confidence interval (0.2987 to 0.3455) supports this finding, as does the sample estimate (0.322). In summary, the p-value (2.2e-16) provides robust evidence against the null hypothesis, signifying a strong link between “OrderCount” and “CashbackAmount.” This result from a linear regression model, where “Residuals” show error distribution, “Coefficients” provide insights, and the “Residual standard error” reflects error magnitude. “Multiple R-squared” and “Adjusted R-squared” gauge model fit, with a low p-value emphasizing significance. The bootstrap analysis suggests a statistically significant positive correlation between OrderCount and CashbackAmount, within the specified confidence interval, enhancing our understanding of correlation strength and uncertainty. Hence we reject the null hypothesis.

Question 4.1

Based on the results of the chi-squared test conducted, we reject the null hypothesis. This implies that there is a statistically significant relationship between the variables ‘Churn’ and ‘Gender.’ The data provides evidence to suggest that gender has an influence on a customer’s decision to churn

Question 4.2

Based on the results of the chi-squared test conducted we reject the null Hypothesis. This implies that there is a statistically significant relationship between the variables ‘Churn’ and ‘Complaint’. There is enough statistical evidence that shows that there is a relationship between customer registering a complain and decision to churn.

References

Verma,(2021, January 26) A. Ecommerce customer churn analysis and prediction. Kaggle. https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction (https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction)

Countants. (2020, January 5). Why consumer behavior analysis is so relevant to the ecommerce business? Medium https://medium.datadriveninvestor.com/why-consumer-behavior-analysis-is-so-relevant-to-the-ecommerce-business-8f49c250ca9c (https://medium.datadriveninvestor.com/why-consumer-behavior-analysis-is-so-relevant-to-the-ecommerce-business-8f49c250ca9c)

Zanzana, Salim, and Jessica Martin. (2023, February 21). Retail e-commerce and COVID-19: How online sales evolved as in-person shopping resumed. https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2023002-eng.htm (https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2023002-eng.htm).

Verma, Ankit.(2023, July 6). “E-commerce Dataset.” (CC BY-NC-SA 4.0) creativecommon.org https://creativecommons.org/licenses/by-nc-sa/4.0/ (https://creativecommons.org/licenses/by-nc-sa/4.0/)