DSCC 462: Computational Introduction to Statistics Final Project

Sixiao Song | Ruilin Zhang | Tan Phan | Raymond Yu

2023-12-14

Introduction

The dataset selected for our final project originates from a Kaggle competition and it belongs to a leading online E-Commerce company. An online retail (E-commerce) company wants to know the customers who are going to churn, so accordingly they can approach customers to offer some promos. The dataset encompasses 20 variables and 5630 rows, providing a rich source of insights into customer behavior and purchase patterns. Key features include customer demographics, purchase history, and customer support interactions.

```
library(readxl)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(stringr)
```

Exploratory data analysis (EDA)

```
data <- read_excel("ECommerce_Dataset.xlsx")
print(data)
```

```
## # A tibble: 5,630 x 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
##  7      50007     1     NA Phone                          3              11
##  8      50008     1     NA Phone                          1               6
##  9      50009     1     13 Phone                          3               9
## 10      50010     1     NA Phone                          1              31
## # i 5,620 more rows
## # 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>
```

```
# Check data structure
str(data)
```

```
## tibble [5,630 x 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 & Access
##  $ 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 ...
```

```
# Descriptive statistics
summary(data)
```

```
##    CustomerID        Churn            Tenure      PreferredLoginDevice
##  Min.   :50001   Min.   :0.0000   Min.   : 0.00   Length:5630
##  1st Qu.:51408   1st Qu.:0.0000   1st Qu.: 2.00   Class :character
##  Median :52816   Median :0.0000   Median : 9.00   Mode  :character
##  Mean   :52816   Mean   :0.1684   Mean   :10.19
##  3rd Qu.:54223   3rd Qu.:0.0000   3rd Qu.:16.00
##  Max.   :55630   Max.   :1.0000   Max.   :61.00
##                                   NA's   :264
##     CityTier     WarehouseToHome  PreferredPaymentMode    Gender
```

```
##  Min.   :1.000   Min.   :  5.00   Length:5630          Length:5630
##  1st Qu.:1.000   1st Qu.:  9.00   Class :character     Class :character
##  Median :1.000   Median : 14.00   Mode  :character     Mode  :character
##  Mean   :1.655   Mean   : 15.64
##  3rd Qu.:3.000   3rd Qu.: 20.00
##  Max.   :3.000   Max.   :127.00
##                  NA's   :251
##  HourSpendOnApp  NumberOfDeviceRegistered PreferedOrderCat   SatisfactionScore
##  Min.   :0.000   Min.   :1.000            Length:5630        Min.   :1.000
##  1st Qu.:2.000   1st Qu.:3.000            Class :character   1st Qu.:2.000
##  Median :3.000   Median :4.000            Mode  :character   Median :3.000
##  Mean   :2.932   Mean   :3.689                               Mean   :3.067
##  3rd Qu.:3.000   3rd Qu.:4.000                               3rd Qu.:4.000
##  Max.   :5.000   Max.   :6.000                               Max.   :5.000
##  NA's   :255
##  MaritalStatus      NumberOfAddress     Complain
##  Length:5630        Min.   : 1.000   Min.   :0.0000
##  Class :character   1st Qu.: 2.000   1st Qu.:0.0000
##  Mode  :character   Median : 3.000   Median :0.0000
##                     Mean   : 4.214   Mean   :0.2849
##                     3rd Qu.: 6.000   3rd Qu.:1.0000
##                     Max.   :22.000   Max.   :1.0000
##
##  OrderAmountHikeFromlastYear   CouponUsed       OrderCount
##  Min.   :11.00               Min.   : 0.000   Min.   : 1.000
##  1st Qu.:13.00               1st Qu.: 1.000   1st Qu.: 1.000
##  Median :15.00               Median : 1.000   Median : 2.000
##  Mean   :15.71               Mean   : 1.751   Mean   : 3.008
##  3rd Qu.:18.00               3rd Qu.: 2.000   3rd Qu.: 3.000
##  Max.   :26.00               Max.   :16.000   Max.   :16.000
##  NA's   :265                 NA's   :256      NA's   :258
##  DaySinceLastOrder CashbackAmount
##  Min.   : 0.000    Min.   :  0.0
##  1st Qu.: 2.000    1st Qu.:145.8
##  Median : 3.000    Median :163.3
##  Mean   : 4.543    Mean   :177.2
##  3rd Qu.: 7.000    3rd Qu.:196.4
##  Max.   :46.000    Max.   :325.0
##  NA's   :307
```

```
colnames(data)
```

```
##  [1] "CustomerID"                "Churn"
##  [3] "Tenure"                    "PreferredLoginDevice"
##  [5] "CityTier"                  "WarehouseToHome"
##  [7] "PreferredPaymentMode"      "Gender"
##  [9] "HourSpendOnApp"            "NumberOfDeviceRegistered"
## [11] "PreferedOrderCat"          "SatisfactionScore"
## [13] "MaritalStatus"             "NumberOfAddress"
## [15] "Complain"                  "OrderAmountHikeFromlastYear"
## [17] "CouponUsed"                "OrderCount"
## [19] "DaySinceLastOrder"         "CashbackAmount"
```

Change Column Names:

```
names(data)[names(data) == "PreferedOrderCat"] <- "PreferredOrderCat"
data$PreferredLoginDevice <- as.factor(str_replace(data$PreferredLoginDevice,
                                                   "Mobile Phone", "Mobile"))
data$PreferredLoginDevice <- as.factor(str_replace(data$PreferredLoginDevice,
                                                   "Phone", "Mobile"))
data$PreferredPaymentMode <- as.factor(str_replace(data$PreferredPaymentMode,
                                                    "CC", "Credit Card"))
data$PreferredPaymentMode <- as.factor(str_replace(data$PreferredPaymentMode,
                                                    "COD", "Cash on Delivery"))
data$PreferredOrderCat <- as.factor(str_replace(data$PreferredOrderCat,
                                                   "Mobile Phone", "Mobile"))
```

Missing Values:

```
sum(is.na(data))
```

```
## [1] 1856
```

```
colSums(is.na(data))
```

```
##                  CustomerID                       Churn
##                           0                           0
##                      Tenure       PreferredLoginDevice
##                         264                           0
##                    CityTier            WarehouseToHome
##                           0                         251
##       PreferredPaymentMode                      Gender
##                           0                           0
##             HourSpendOnApp   NumberOfDeviceRegistered
##                         255                           0
##          PreferredOrderCat          SatisfactionScore
##                           0                           0
##              MaritalStatus            NumberOfAddress
##                           0                           0
##                    Complain OrderAmountHikeFromlastYear
##                           0                         265
##                CouponUsed                  OrderCount
##                         256                         258
##          DaySinceLastOrder              CashbackAmount
##                         307                           0
```

Insight: Some columns have missing values, total of 1856. Some columns, such as "MaritalStatus," "NumberOfAddress," and "Complain," have zero missing values. This information is useful for understanding the completeness of data in certain aspects.

```r
# Imputing missing values under numerical data using median
data$Tenure[is.na(data$Tenure)] <- round(median(data$Tenure,na.rm=TRUE))
data$WarehouseToHome[is.na(data$WarehouseToHome)] <- round(
  median(data$WarehouseToHome,na.rm=TRUE))
data$HourSpendOnApp[is.na(data$HourSpendOnApp)] <- round(
  median(data$HourSpendOnApp,na.rm=TRUE))
data$OrderAmountHikeFromlastYear[is.na(data$OrderAmountHikeFromlastYear)] <- round(
  median(data$OrderAmountHikeFromlastYear,na.rm=TRUE))
data$CouponUsed[is.na(data$CouponUsed)] <- round(
  median(data$CouponUsed,na.rm=TRUE))
data$OrderCount[is.na(data$OrderCount)] <- round(
  median(data$OrderCount,na.rm=TRUE))
data$DaySinceLastOrder[is.na(data$DaySinceLastOrder)] <- round(
  median(data$DaySinceLastOrder,na.rm=TRUE))
```

```r
# Check again missing values
sum(is.na(data))
```

```
## [1] 0
```

Visualize the data using histograms:

a) The CEO would like to gain some insights into the distribution of the time elapsed since customers placed their last orders. How many bins does Sturges' formula suggest we use for a histogram of DaySinceLastOrder?

```r
ceiling(log(length(data$DaySinceLastOrder), 2)) + 1
```

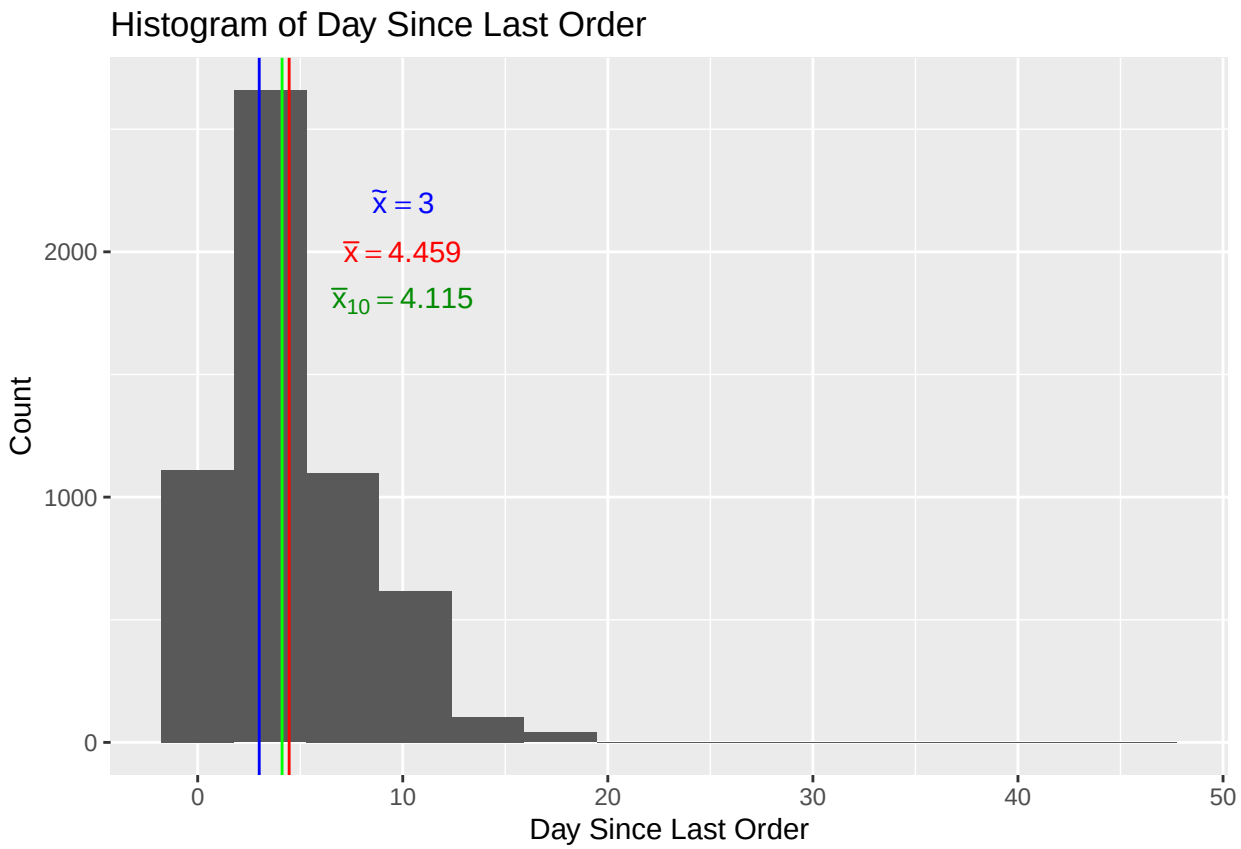
```
## [1] 14
```

b) Create a histogram of DaySinceLastOrder using the number of bins suggested by Sturges' formula. Make sure to appropriately title the histogram and label the axes. Comment on the center, shape, and spread. Calculate the mean, median, and 10% trimmed mean of the DaySinceLastOrder. Report the mean, median, and 10% trimmed mean on the histogram.

```r
x <- mean(data$DaySinceLastOrder)
y <- median(data$DaySinceLastOrder)
z <- mean(data$DaySinceLastOrder, trim = 0.1)

ggplot(data = data, aes(x = DaySinceLastOrder)) +
geom_histogram(bins = 14) +
xlab("Day Since Last Order") +
ylab("Count") +
ggtitle("Histogram of Day Since Last Order") +
  geom_vline(xintercept = x, color = "red") +
  annotate("text", x = 10, y = 2000,
label = paste("bar(x)==",round(x,3)),parse=T, color="red") +
  geom_vline(xintercept = y, color = "blue") +
```

```
  annotate("text", x = 10, y = 2200,
label = paste("tilde(x)==",round(y,3)),parse=T, color="blue") +
  geom_vline(xintercept = z, color = "green") +
  annotate("text", x = 10, y = 1800,
label = paste("bar(x)[10] == ",round(z,3)), parse=T, color="green4")
```



Insight: The histogram is unimodal and positively (right) skewed. The mean of 4.459 indicates the average time elapsed since the last order. The median of 3 suggests that half of the observations fall below 3, indicating a potential right skew in the distribution. And 10% trim mean of day since last order is 4.115.

c) Calculate and report the interquartile range.

```
IQR(data$DaySinceLastOrder)
```

```
## [1] 5
```

Insight: The IQR is 5.

d) Calculate and report the standard span, the lower fence, and the upper fence.

```
standard_span <- 1.5 * IQR (data$DaySinceLastOrder)
lower_fence <- quantile(data$DaySinceLastOrder, 0.25) - standard_span
```

```r
upper_fence <- quantile(data$DaySinceLastOrder, 0.75) + standard_span
cat("The standard span is", standard_span, "\n")
```

```
## The standard span is 7.5
```

```r
cat("The lower fence is", lower_fence, "\n")
```

```
## The lower fence is -5.5
```

```r
cat("The upper fence is", upper_fence, "\n")
```

```
## The upper fence is 14.5
```

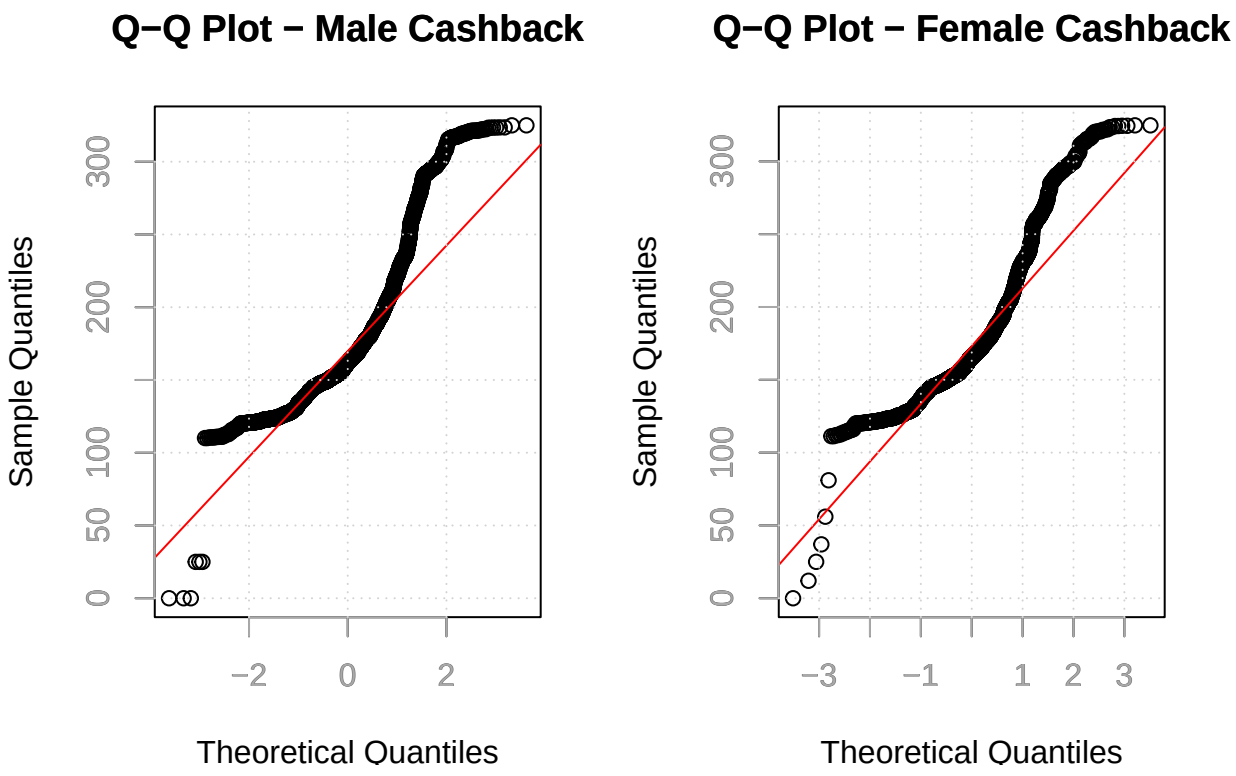
1. Inference about mean(s):

Q1: By using independent samples t-test or Welch's t-test (if variances are not assumed to be equal), the CEO wants to know are the mean values of cashback amount between female and male significantly different.

a) Create a side-by-side quantile-quantile (Q-Q) plots for cashback amounts, comparing the distribution of cashback amounts between male and female customers.

```r
# Filter data for male and female
data_male <- subset(data, Gender == "Male")
data_female <- subset(data, Gender == "Female")
# Create quantile-quantile (Q-Q) plots for each gender
par(mfrow = c(1, 2))

qqnorm(data_male$CashbackAmount, main = "Q-Q Plot - Male Cashback")
qqline(data_male$CashbackAmount, col = "red")
title(main = "Q-Q Plot - Male Cashback", sub = "")
axis(1, col = "darkgray", col.axis = "darkgray")
axis(2, col = "darkgray", col.axis = "darkgray")
grid()
qqnorm(data_female$CashbackAmount, main = "Q-Q Plot - Female Cashback")
qqline(data_female$CashbackAmount, col = "red")
title(main = "Q-Q Plot - Female Cashback", sub = "")
axis(1, col = "darkgray", col.axis = "darkgray")
axis(2, col = "darkgray", col.axis = "darkgray")
grid()
```

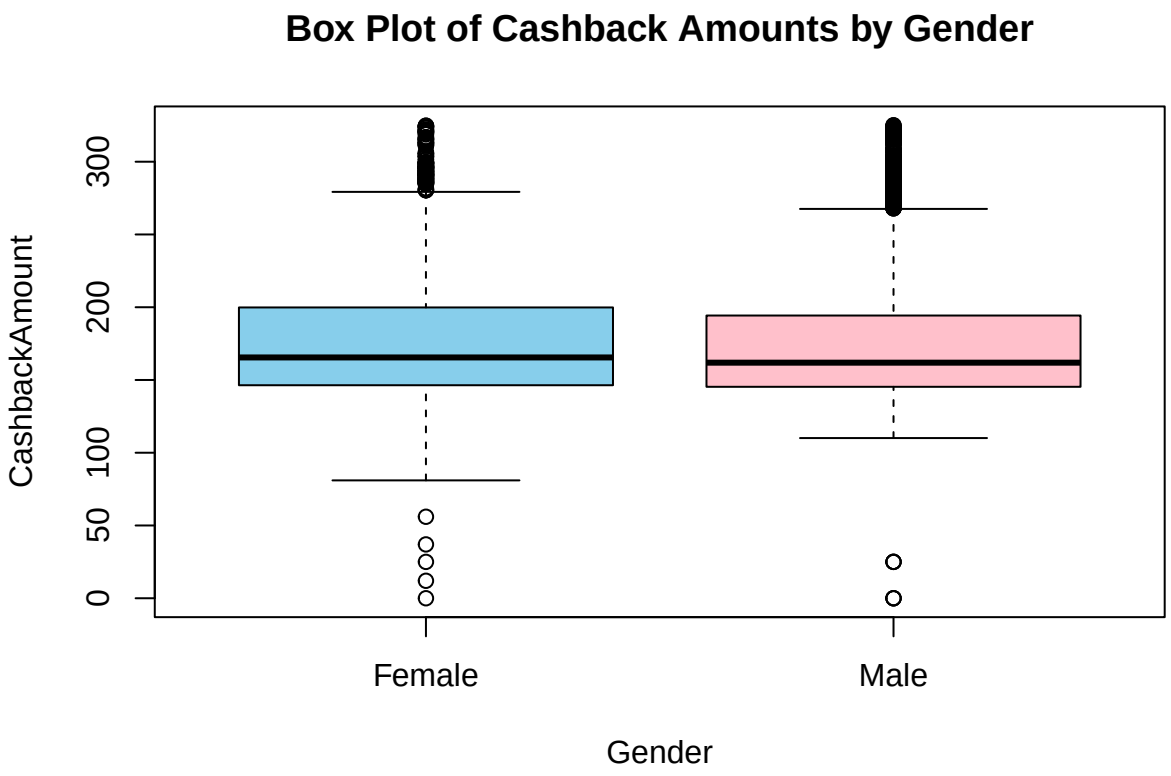



```
par(mfrow = c(1, 1))
```

Insight: It seems that the cashback data is not normally distributed. However, the sample size is large enough (>30). According to the Central Limit Theorem, we can still apply a t-test for mean inference.

b) Create a box plot for Cashback Amounts by Gender, visualize the distribution of cashback amounts for different gender categories.

```
boxplot(CashbackAmount ~ Gender, data = data,
        col = c("skyblue", "pink"),
        main = "Box Plot of Cashback Amounts by Gender",
        xlab = "Gender",
        ylab = "CashbackAmount")
```



c) Are the mean values of cashback amount between female and male significantly different? Perform an appropriate statistical test at the alpha = 0.05 significance level and comment on the results.

H0: The means of cashback amount of the two groups (male and female) are equal. H1: The means of cashback amount of the two groups (male and female) are not equal.

```
data_male <- subset(data, Gender == "Male")
data_female <- subset(data, Gender == "Female")
# Perform independent two-sample t-test
t_test_result <- t.test(data_male$CashbackAmount, data_female$CashbackAmount)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_male$CashbackAmount and data_female$CashbackAmount
## t = -1.8953, df = 4850.8, p-value = 0.05812
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.14921725  0.08706856
## sample estimates:
## mean of x mean of y
##  176.2133  178.7444
```

Insight: Given the p-value is 0.05812 which is greater than 0.05, we cannot reject the null hypothesis, indicating no significant difference in the mean Cashback Amounts between male and female groups.

2. Inference about variance(s):

Q2: The CEO wants to know are the variance values of cashback amounts between female and male significantly different? Perform an appropriate statistical test at the alpha = 0.05 significance level and comment on the results.

H0: The variance of cashback amount of the two groups (male and female) are equal. H1: The variance of cashback amount of the two groups (male and female) are not equal.

```
data_male <- subset(data, Gender == "Male")
data_female <- subset(data, Gender == "Female")
# Perform independent two-sample f-test
var_test_result <- var.test(data_male$CashbackAmount,
                            data_female$CashbackAmount)
print(var_test_result)
```

```
##
##  F test to compare two variances
##
## data:  data_male$CashbackAmount and data_female$CashbackAmount
## F = 1.026, num df = 3383, denom df = 2245, p-value = 0.5076
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9511065 1.1060677
## sample estimates:
## ratio of variances
##           1.025964
```

Insight: The test statistic (F) is 1.026. Given the p-value of 0.5076, which is greater than the significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no significant difference in the variance of Cashback Amounts between male and female groups.

3. Inference about proportion(s):

Q3: The CEO wants the team to conduct a hypothesis test to compare the churn rate of customers who use coupons to the average churn rate of all customers in order to determine whether there is a significant difference in churn rates between these two groups.

```
coupon_data = data[data$CouponUsed > 0,]
```

```
# Overall churn rate
overall_churn_rate = sum(data$Churn)  / nrow(data)
cat("Overall churn rate:", overall_churn_rate, "\n")
```

```
## Overall churn rate: 0.1683837
```

```
# Churn rate of customers using coupon
coupon_churn_rate =  sum(coupon_data$Churn, na.rm = TRUE) / nrow(coupon_data)
cat("Coupon churn rate:", coupon_churn_rate, "\n")
```

```
## Coupon churn rate: 0.1656522
```

```
n_coupon_users <- nrow(coupon_data)
cat("Number of coupon users:", n_coupon_users, "\n")
```

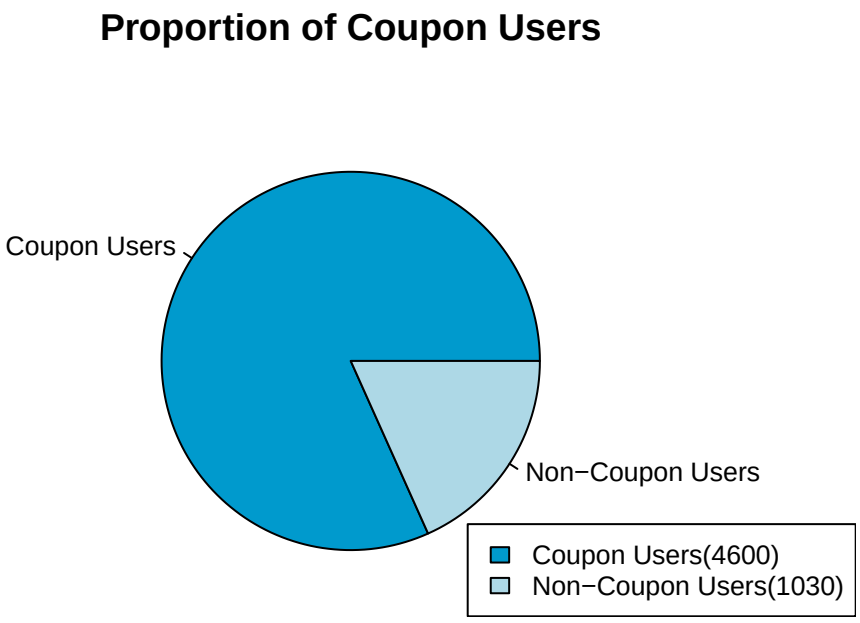
```
## Number of coupon users: 4600
```

```
n_all_users <- nrow(data)
cat("Number of all users:", n_all_users, "\n")
```

```
## Number of all users: 5630
```

a) Create a pie chart to visualize the proportion of coupon users and non-coupon users in the dataset.

```
# Create a data frame for the number of users
user_counts <- data.frame(
  Group = c("Coupon Users", "Non-Coupon Users"),
  Count = c(n_coupon_users, n_all_users - n_coupon_users))
pie(user_counts$Count, labels = user_counts$Group, col = c(
  "deepskyblue3", "lightblue"),
    main = "Proportion of Coupon Users",
    cex = 0.8)
legend("bottomright", legend = paste(user_counts$Group,
                                     "(", user_counts$Count, ")", sep = ""),
       fill = c("deepskyblue3", "lightblue"),cex = 0.8)
```



b) Is the churn rate of customers who uses coupon higher than the average churn rate of all customers? Perform an appropriate statistical test at the alpha = 0.05 significance level and comment on the results.

H0: The churn rate of customers who uses coupon is equal to or smaller than the average churn rate of all customers. H1: The churn rate of customers who uses coupon is higher than the average churn rate of all customers

```
# Perform one sample proportion z-test
standard_error <- sqrt((overall_churn_rate * (1 - overall_churn_rate)
                        / n_coupon_users))
z_stat <- (coupon_churn_rate - overall_churn_rate) / standard_error
p_value <- 1 - pnorm(z_stat)

cat("Coupon Users Churn Rate:", coupon_churn_rate, "\n")
```

```
## Coupon Users Churn Rate: 0.1656522
```

```
cat("Overall Churn Rate:", overall_churn_rate, "\n")
```

```
## Overall Churn Rate: 0.1683837
```

```
cat("Z-statistic:", z_stat, "\n")
```

```
## Z-statistic: -0.4950696
```

```
cat("P-value:", p_value, "\n")
```

```
## P-value: 0.6897245
```

```
#if (p_value < 0.05) {
#  cat("Reject the null hypothesis.")
#} else {
#  cat("Fail to reject the null hypothesis.")}
```

Insight: With a p-value of 0.6897245 (greater than the significance level of 0.05), we fail to reject the null hypothesis. This means there is no significant evidence to conclude that the churn rate for coupon users is higher than the average churn rate for all customers.

4. Inference about two proportions:

Q4: The CFO wants to better understand the relationship between churn rate and gender.

```
female_count <- sum(data$Gender == "Female", na.rm = TRUE)
male_count <- sum(data$Gender == "Male", na.rm = TRUE)
cat("Number of Females:", female_count, "\n")
```

```
## Number of Females: 2246
```

```r
cat("Number of Males:", male_count, "\n")
```

```
## Number of Males: 3384
```

```r
churn <- data$OrderCount[data$Churn == 1]
not_churn <- data$OrderCount[data$Churn == 0]
female_data <- filter(data, Gender == "Female")
male_data <- filter(data, Gender == "Male")
# Count the number of rows where churn is '1' for females
female_data_churn <- sum(female_data$Churn == 1)
cat("Number of Females Churn:", female_data_churn, "\n")
```

```
## Number of Females Churn: 348
```

```r
# Count the number of rows where churn is '1' for males
male_data_churn <- sum(male_data$Churn == 1)
cat("Number of Males Churn:", male_data_churn, "\n")
```

```
## Number of Males Churn: 600
```

a) The CEO wants to know if the percentage of female who churn is higher than the percentage of male who churn. Out of a sample of 2246 females, they found that 348 of them churn. Out of 3384 males, they found that 600 of them churn. We would like to run a test to determine whether their hypothesis is true with a Type I error probability of 0.05.

```r
# Perform two sample proportion z-test
female_count <- sum(data$Gender == "Female", na.rm = TRUE)
male_count <- sum(data$Gender == "Male", na.rm = TRUE)

two_sample_z_test <- prop.test(
  x = c(female_data_churn, male_data_churn),
  n = c(female_count, male_count), p = NULL)
two_sample_z_test
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(female_data_churn, male_data_churn) out of c(female_count, male_count)
## X-squared = 4.6629, df = 1, p-value = 0.03082
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.042469789 -0.002255901
## sample estimates:
##    prop 1    prop 2
## 0.1549421 0.1773050
```

Insight: The analysis conducted on a sample of 2246 females and 3384 males aimed to assess whether there is a significant difference in the churn rates between the two gender groups. The results of a two-sample z-test for equality of proportions revealed a statistically significant difference, with a p-value of 0.03082. The test results yield a p-value of 0.03082, which is less than the Type I error probability of 0.05. Therefore, we reject the null hypothesis, suggesting a gender-based distinction in churn rates within the studied population.

b) Construct a two-sided 95% confidence interval for the proportion of each gender that who churn.

```
n_female <- 2246
n_male <- 3384

ci_female <- prop.test(x = c(female_data_churn),
                       n = c(n_female),
                       alternative = "two.sided",
                       conf.level = 0.95,
                       correct = TRUE)
ci_female
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  c(female_data_churn) out of c(n_female), null probability 0.5
## X-squared = 1068.3, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1403542 0.1707259
## sample estimates:
##         p
## 0.1549421
```

```
ci_male <- prop.test(x = c(male_data_churn),
                       n = c(n_male),
                       alternative = "two.sided",
                       conf.level = 0.95,
                       correct = TRUE)
ci_male
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  c(male_data_churn) out of c(n_male), null probability 0.5
## X-squared = 1408.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1646615 0.1906886
## sample estimates:
##        p
## 0.177305
```

```
cat("95% Confidence Interval for the Proportion of Females who Churn:",
    ci_female$conf.int, "\n")
```

```
## 95% Confidence Interval for the Proportion of Females who Churn: 0.1403542 0.1707259
```

```r
cat("95% Confidence Interval for the Proportion of Males who Churn:",
    ci_male$conf.int, "\n")
```

```
## 95% Confidence Interval for the Proportion of Males who Churn: 0.1646615 0.1906886
```

Insight: For a two-sided confidence interval, we are 95% confident that the interval between 0.1404 and 0.1707 contains the true difference in the proportion of females who churn. We are 95% confident that the interval between 0.1647 and 0.1907 contains the true difference in the proportion of males who churn.

5. Chi-Squared Inference (goodness-of-fit or test of independence):

Q5: Based on the data, which involves an online retail (E-commerce) company and includes various customer-related features, we are interested in exploring the relationships between variables. Chi-square tests are appropriate for analyzing such relationships.

a) First of all, the CEO wants to understand how many customers have churn on each preferred order category.

```r
grouped_data <- data %>%
    group_by(PreferredOrderCat) %>%
    summarise(total_count = n(), .groups = "drop") %>%
    as.data.frame()
grouped_data
```

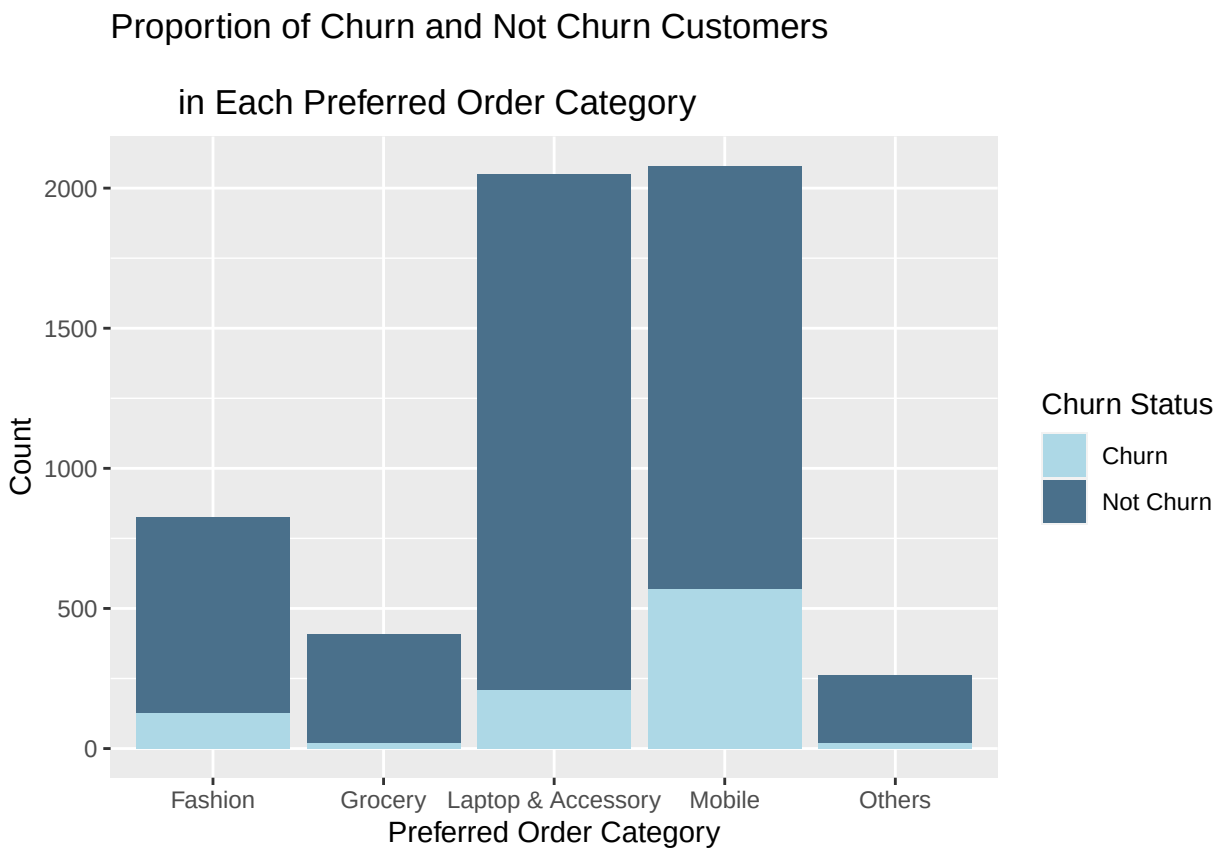
```
##    PreferredOrderCat total_count
## 1            Fashion         826
## 2            Grocery         410
## 3 Laptop & Accessory        2050
## 4             Mobile        2080
## 5             Others         264
```

```r
# Extracting 'OrderCount' for churn and not churn customers
churn <- data$OrderCount[data$Churn == 1]
not_churn <- data$OrderCount[data$Churn == 0]
# Creating a summary data frame
grouped_data <- data %>%
  group_by(PreferredOrderCat) %>%
  summarise(total_churn = sum(Churn == 1), total_not_churn = sum(Churn == 0),
            .groups = "drop") %>%
  as.data.frame()
grouped_data
```

```
##    PreferredOrderCat total_churn total_not_churn
## 1            Fashion         128             698
## 2            Grocery          20             390
## 3 Laptop & Accessory         210            1840
## 4             Mobile         570            1510
## 5             Others          20             244
```

b) Visualize the proportion of churn and not churn customers within each preferred order category. Which preferred order categories have the highest and lowest churn rates?

```
ggplot(data = grouped_data, aes(x = PreferredOrderCat)) +
  geom_bar(aes(y = total_churn + total_not_churn,
               fill = "Not Churn"), stat = "identity") +
  geom_bar(aes(y = total_churn, fill = "Churn"), stat = "identity") +
  scale_fill_manual(values = c("Churn" = "lightblue",
                               "Not Churn" = "skyblue4")) +
  labs(title = "Proportion of Churn and Not Churn Customers\n
       in Each Preferred Order Category",
       x = "Preferred Order Category",
       y = "Count",
       fill = "Churn Status")
```




```
# Calculate churn rates
grouped_data$churn_rate <- (grouped_data$total_churn /
          (grouped_data$total_churn + grouped_data$total_not_churn)) * 100
print(grouped_data)
```

```
##    PreferredOrderCat total_churn total_not_churn churn_rate
## 1            Fashion         128             698  15.496368
## 2            Grocery          20             390   4.878049
## 3 Laptop & Accessory         210            1840  10.243902
## 4             Mobile         570            1510  27.403846
## 5             Others          20             244   7.575758
```

Insight: It was found that the category with the highest churn rate is Mobile, where approximately 27.4% of customers churned. This suggests a relatively higher likelihood of customer attrition within the Mobile category. On the other hand, the category with the lowest churn rate is Grocery, with a churn rate of around 4.88%. This implies a comparatively lower likelihood of customers leaving within the Grocery category.

c) A contingency table with counts for each combination of "Preferred Order Category" and "Total Churn" is presented. The table is formatted for a chi-squared test. Is there a significant association between the preferred order category and customer churn for the online retail company? Conduct an appropriate test at 0.05 significance level to determine whether customers churn is associated with their preferred category.

Prefered Order Category	Total Churn	Total Not Churn	Total
Fashion	128	698	826
Grocery	20	390	410
LaptopAccessory	210	1840	2050
Mobile	570	1510	2080
Others	20	244	264
Total	948	4682	5630

H0: There is no association between the preferred order category and customer churn. H1: There is a significant association between the preferred order category and customer churn.

```
chi.table <- matrix(c(128, 20, 210, 570, 20, 698, 390, 1840, 1510, 244),
                nrow=5, ncol=2)
chisq.test(chi.table, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  chi.table
## X-squared = 288.6, df = 4, p-value < 2.2e-16
```

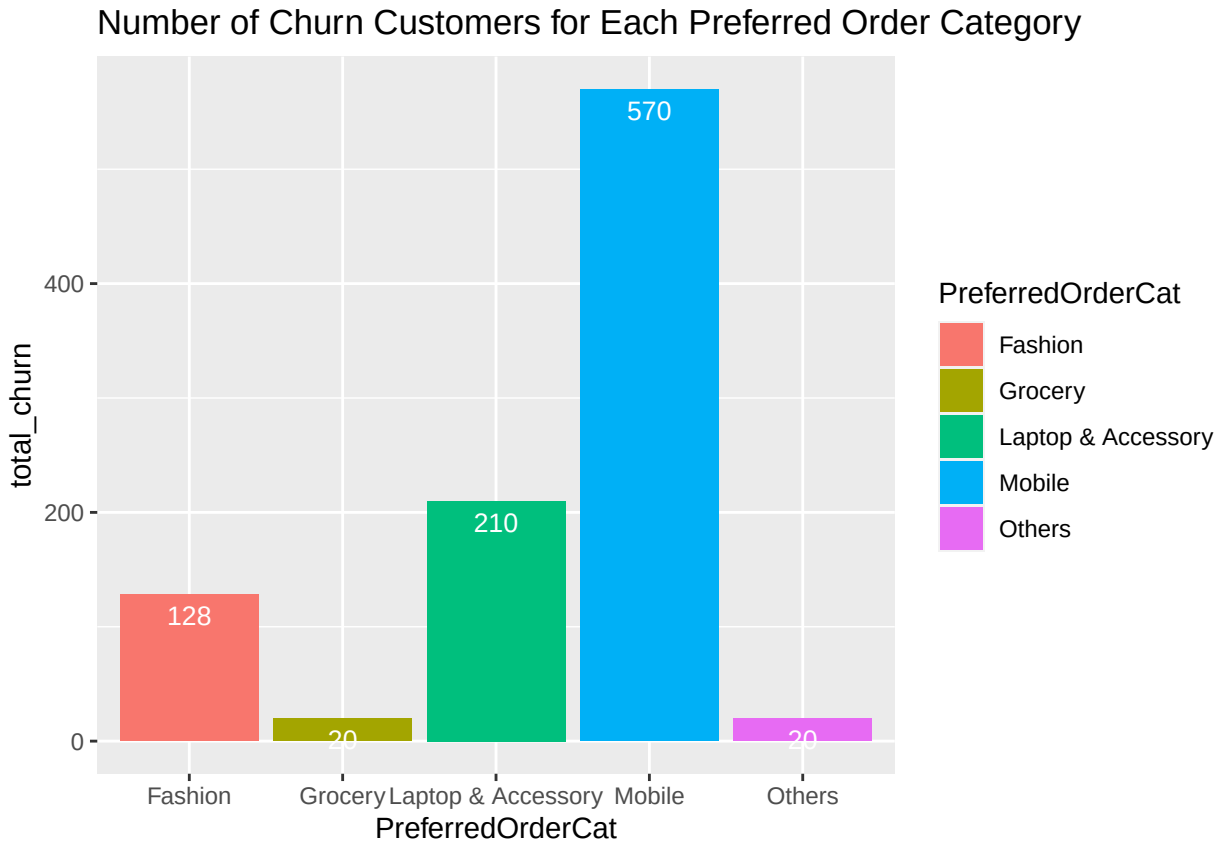
Insight: From the test, chi-squared statistic is 288.6, degree of freedom is 4, the p-value is very close to zero (p-value < 2.2e-16). Given that the p-value is less than the significance level of 0.05, we reject the null hypothesis. Therefore, there is a significant association between the preferred order category and customer churn for the online retail company. This suggests that customer churn is not independent of the preferred order category, and there is evidence that the two variables are associated.

d) Create a visualization that represents the total churn in each preferred order category.

```
# Extracting a dataframe with only "PreferredOrderCat" and "total_churn"
extracted_data <- grouped_data[, c("PreferredOrderCat", "total_churn")]
extracted_data
```

```
##    PreferredOrderCat total_churn
## 1            Fashion         128
## 2            Grocery          20
## 3 Laptop & Accessory         210
## 4             Mobile         570
## 5             Others          20
```

```
ggplot(data = extracted_data, aes(x = PreferredOrderCat,
                                  y = total_churn, fill = PreferredOrderCat)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_churn), vjust = 1.6, color = "white",
            position = position_dodge(width = 0.5), size = 3.5) +
  ggtitle("Number of Churn Customers for Each Preferred Order Category")
```



Insight: A significant number of customers in the Mobile category have churn, while the proportions of churn in the Fashion, Laptop & Accessory, Mobile categories appear to be relatively uniform. The number of customers churn in the category Grocery and Others appear to be the same.

e) Based on previous knowledge, the team believe that 30% of customers have churn on Fashion, 10% of customers have churn on Grocery, 20% on Laptop & Accessory, 30% on Mobile, and 10% on others. To see if this is correct, run an appropriate statistical test at the alpha = 0.05 significance level. Is there evidence of a significant difference in churn rates between preferred order categories? Use the chi-square goodness-of-fit test.

H0: Fashion = 0.3, Grocery = 0.1, Laptop & Accessory = 0.2, Mobile = 0.3, Others = 0.1 H1: At least one of these equities does not hold.

```
chisq_test <- chisq.test(c(128, 20, 210, 570, 20),
                         p = c(0.3, 0.1, 0.2, 0.3, 0.1))

print(chisq_test)
```

```
##
```

```
##  Chi-squared test for given probabilities
##
## data:  c(128, 20, 210, 570, 20)
## X-squared = 493.05, df = 4, p-value < 2.2e-16
```

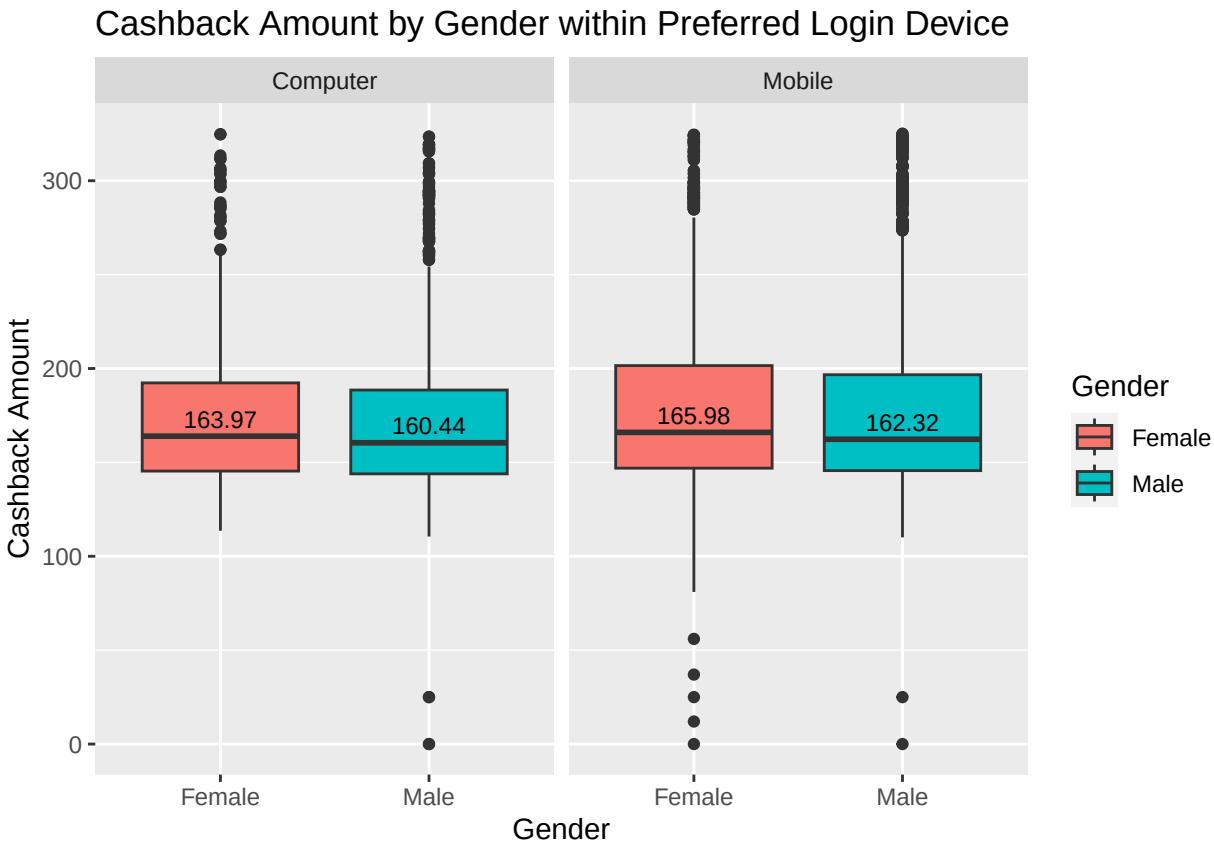
Insight: We conduct a chi-square test for given probabilities to evaluate whether the observed proportions of churn in each preferred order category align with the team's beliefs. The assumed proportions are specified. From the test, chi-squared statistic is 493.05, degree of freedom is 4, the p-value is very close to zero (p-value < 2.2e-16). Given that the p-value is less than the significance level of 0.05, we reject the null hypothesis. Therefore, there is a significant difference in churn rates between preferred order categories. This suggests that the observed distribution of churn customers is significantly different from what would be expected based on the given probabilities.

6. ANOVA:

Q6: The CEO thinks no matter which login device used that the female spend more on as they get more cash back amount on each login device.

a) Create a side-by-side boxplot to visualize the cash back amount of each preferred login device and group by Gender. Make sure to label the plot (title, axes), and comment on trends based on observation.

```r
# Convert CashbackAmount to numeric
data$CashbackAmount <- as.numeric(as.character(data$CashbackAmount))
# Calculate median values for each boxplot
medians <- data %>%
  group_by(PreferredLoginDevice, Gender) %>%
  summarise(Median = median(CashbackAmount), .groups = 'drop')
# Create a side-by-side boxplot
ggplot(data, aes(x = Gender, y = CashbackAmount, fill = Gender)) +
  geom_boxplot() +
  geom_text(data = medians, aes(label = Median, y = Median),
            position = position_dodge(width = 0.75),
            vjust = -0.5, size = 3, color = "black") +
  facet_grid(.~PreferredLoginDevice) +
  labs(title = "Cashback Amount by Gender within Preferred Login Device",
       x = "Gender",
       y = "Cashback Amount")
```



b) Are the means of multiple populations, which are equal by Preferred Login Device and Gender? The CEO wants to compare mean for different preferred login device and gender. Run a two-way ANOVA test at the alpha = 0.05 significance level and comment on the results.

i. Main Effect of PreferredLoginDevice:

H0: There is no significant difference in the means of the dependent variable across the different levels of PreferredLoginDevice.

H1: There is a significant difference in the means of the dependent variable across the different levels of PreferredLoginDevice.

ii. Main Effect of Gender:

H0: There is no significant difference in the means of the dependent variable across different genders.

H1: There is a significant difference in the means of the dependent variable across different genders.

iii. Interaction Effect between PreferredLoginDevice and Gender:

H0: The effect of PreferredLoginDevice on the dependent variable is the same for all levels of Gender (i.e., there's no interaction).

H1: The effect of PreferredLoginDevice on the dependent variable is different for at least one level of Gender (i.e., there's interaction).

```
fit <- aov(data$CashbackAmount~data$PreferredLoginDevice * data$Gender)
summary(fit)
```

```
##                                        Df    Sum Sq Mean Sq F value    Pr(>F)
```

```
## data$PreferredLoginDevice                1    31253   31253  12.938 0.000325
## data$Gender                              1     8137    8137   3.369 0.066507
## data$PreferredLoginDevice:data$Gender    1      766     766   0.317 0.573267
## Residuals                             5626 13589525    2415
##
## data$PreferredLoginDevice             ***
## data$Gender                           .
## data$PreferredLoginDevice:data$Gender
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Insight: As p-value of the combination of column PreferredLoginDevice and column Gender is 0.934, which is not statistically significant, it means that the interaction effects is not significant. The non-significant p-values for Gender and the interaction suggest that they don't have a significant effect on the dependent variable in this analysis. Only The significant p-value for PreferredLoginDevice indicates that it significantly influences the dependent variable.

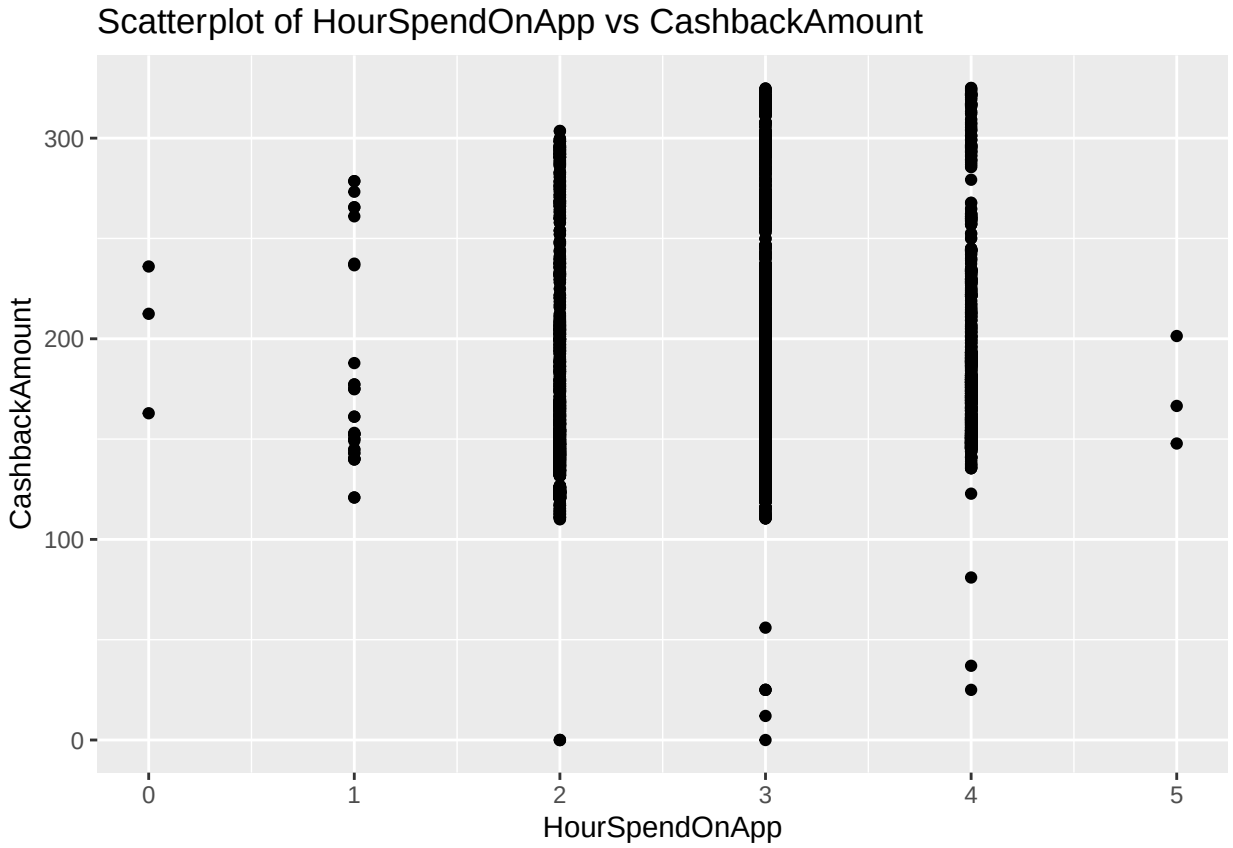
The CEO concludes that gender and the interaction of gender and preferred login devices do not statistically significantly affect the average cashback amount. Only the preferred login device is identified as a significant factor influencing the dependent variable.

7. Inference about correlation:

Q7: It's time to start understanding the correlation between the number of hours spent on the app and the cashback amount.

a) Create a scatterplot that visualizes the relationship between two variables: 'HourSpendOnApp' (representing the number of hours spent on the app) and 'CashbackAmount' (representing the cashback amount).

```
ggplot(data, aes(x = HourSpendOnApp, y = CashbackAmount)) +
  geom_point() +
  labs(title = "Scatterplot of HourSpendOnApp vs CashbackAmount",
       x = "HourSpendOnApp",
       y = "CashbackAmount")
```



b) Is number of hour spent on app correlated with cashback amount? Perform an appropriate statistical test using Pearson correlation at the alpha = 0.05 significance level and comment on the results.

H0: A significant linear relationship does not exist between Hour Spent On App and Cashback Amount.
H1: A significant linear relationship exists between Hour Spent On App and Cashback Amount.

```
cor.test(data$HourSpendOnApp, data$CashbackAmount, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$HourSpendOnApp and data$CashbackAmount
## t = 8.6302, df = 5628, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.08842724 0.13998985
## sample estimates:
##       cor
## 0.1142855
```

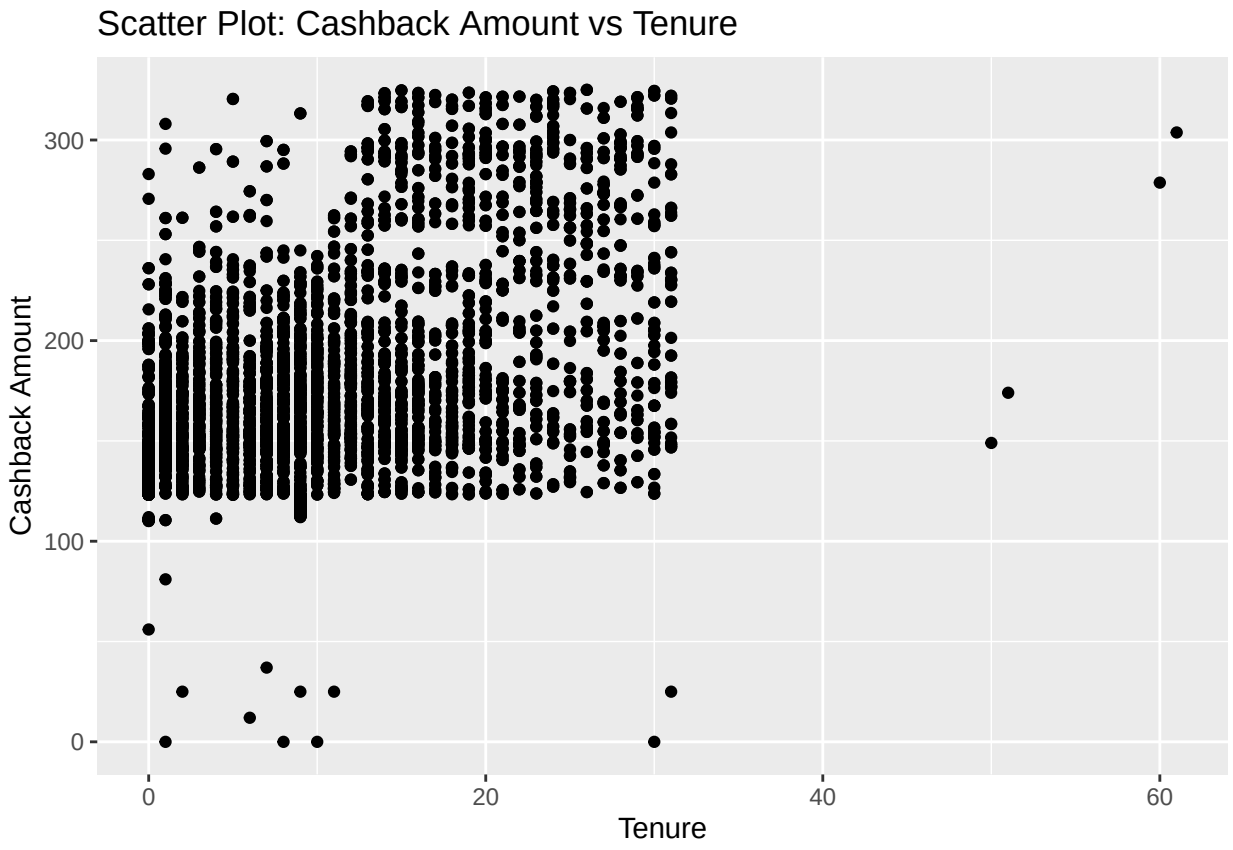
Insight: The correlation coefficient (0.12) is positive, indicating there is nearly no correlation. The p-value is very small, indicating that we can reject the null hypothesis, suggesting that there is a statistically significant correlation between the number of hours spent on the app and the cashback amount.

8. Regression:

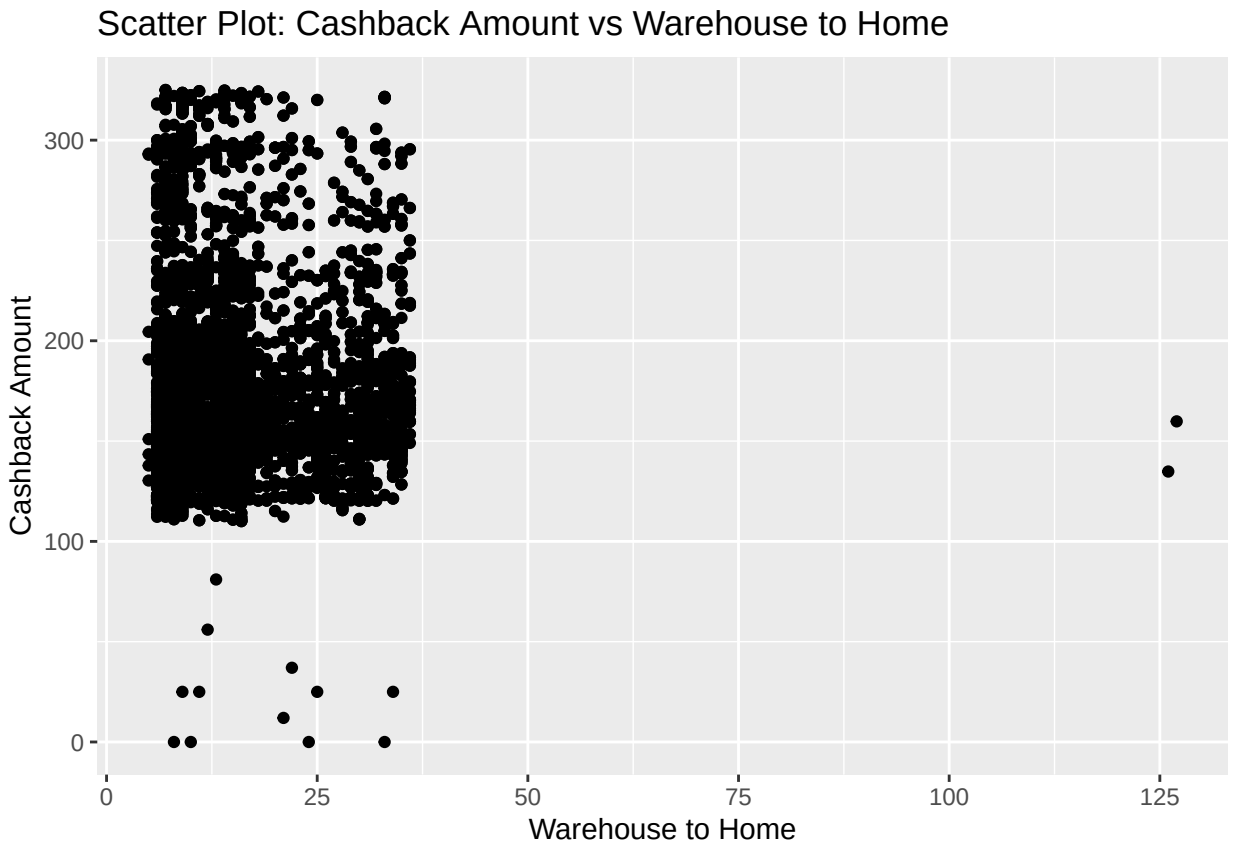
Q8: The CEO wants to know what related attributes affect the cash back amount, so he choose some of the columns, including tenure, distance of warehouse to home, hour spend on app, satisfaction score, how many coupons are used and how many order count there are, which he thinks might be the factors. In addition, he wants to know which column of attribute is statistically significant to the cash back amount.

a) Create scatterplots to visualize the relationship of cash back amount and selected column. Make sure to label the plot (title, axes), and comment on trends you observe.

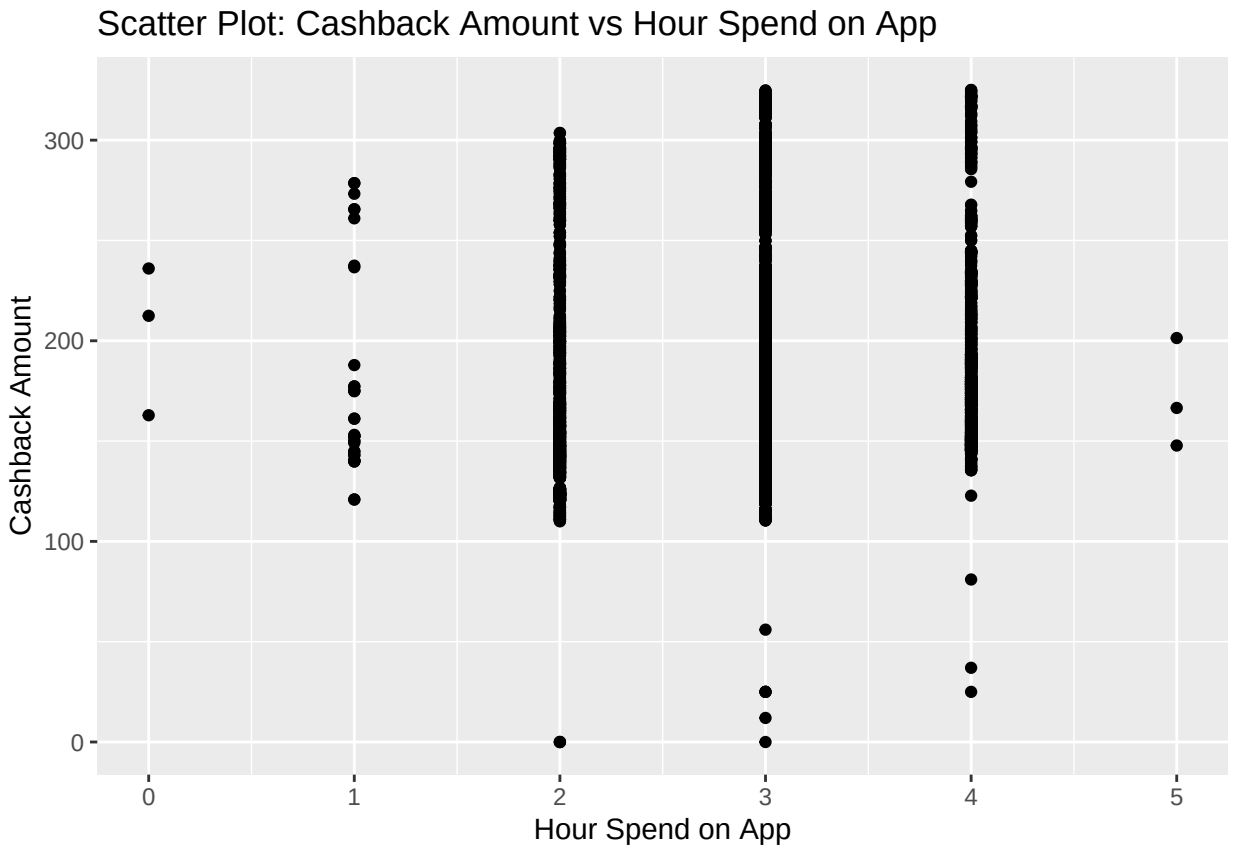
```
ggplot(data = data, aes(x = Tenure, y = CashbackAmount)) +
  geom_point() +
  labs(x = "Tenure", y = "Cashback Amount") +
  ggtitle("Scatter Plot: Cashback Amount vs Tenure")
```



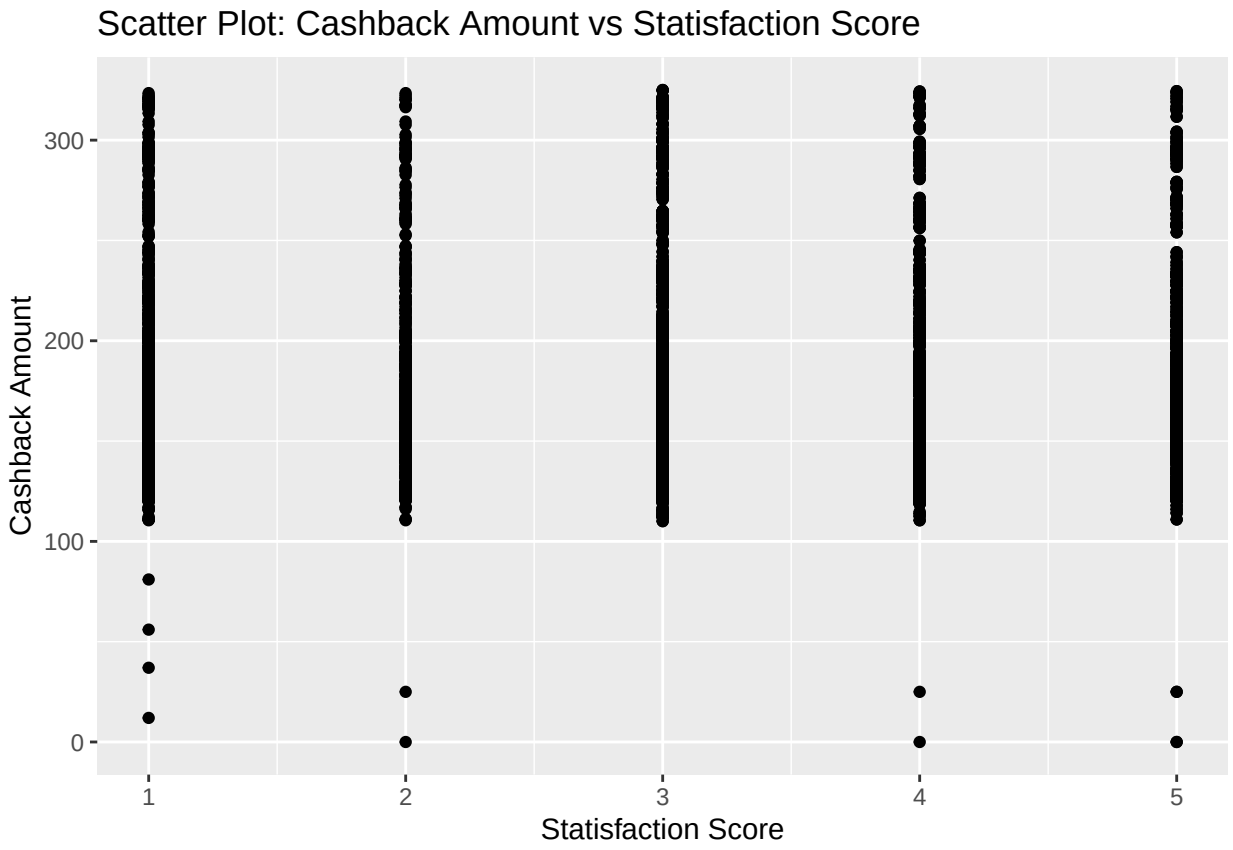

```
ggplot(data = data, aes(x = WarehouseToHome, y = CashbackAmount)) +
  geom_point() +
  labs(x = "Warehouse to Home", y = "Cashback Amount") +
  ggtitle("Scatter Plot: Cashback Amount vs Warehouse to Home")
```



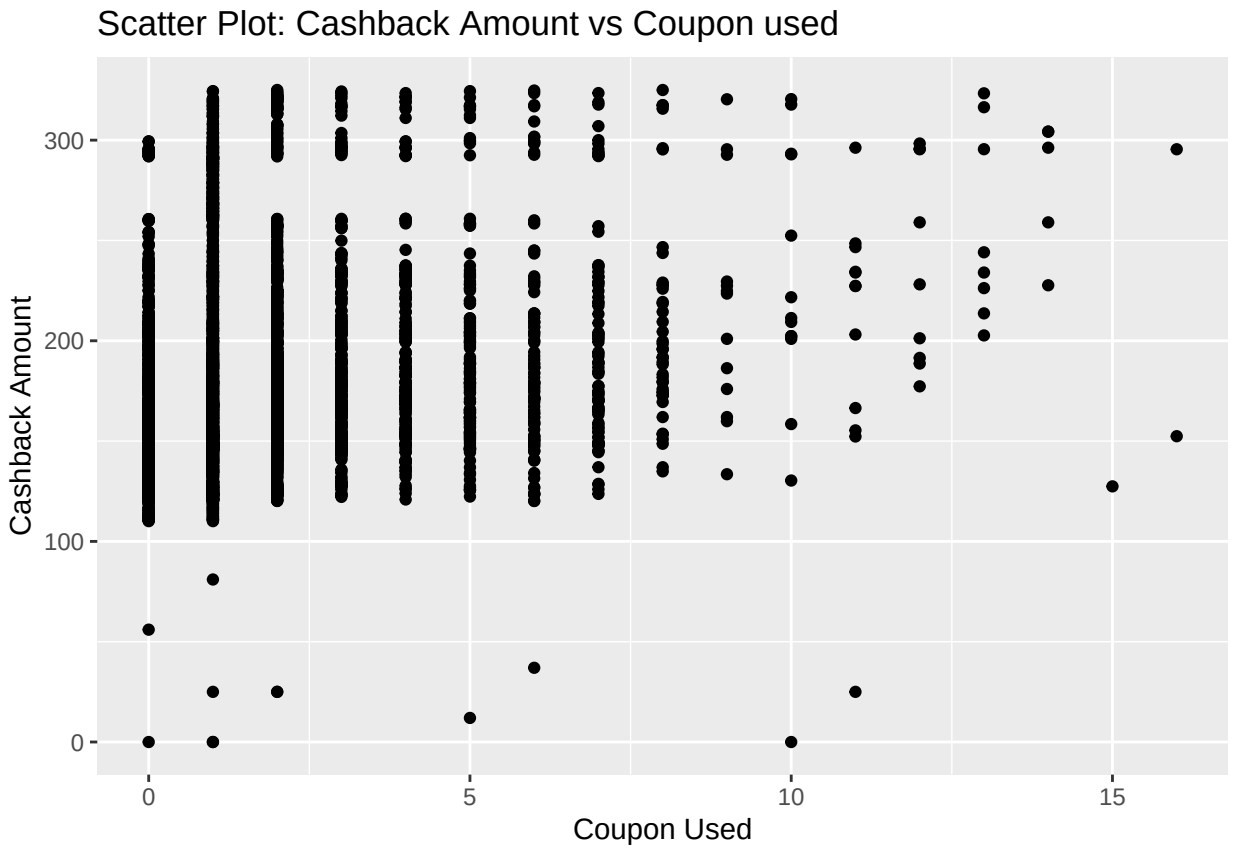

```
ggplot(data = data, aes(x = HourSpendOnApp, y = CashbackAmount)) +
  geom_point() +
  labs(x = "Hour Spend on App", y = "Cashback Amount") +
  ggtitle("Scatter Plot: Cashback Amount vs Hour Spend on App")
```



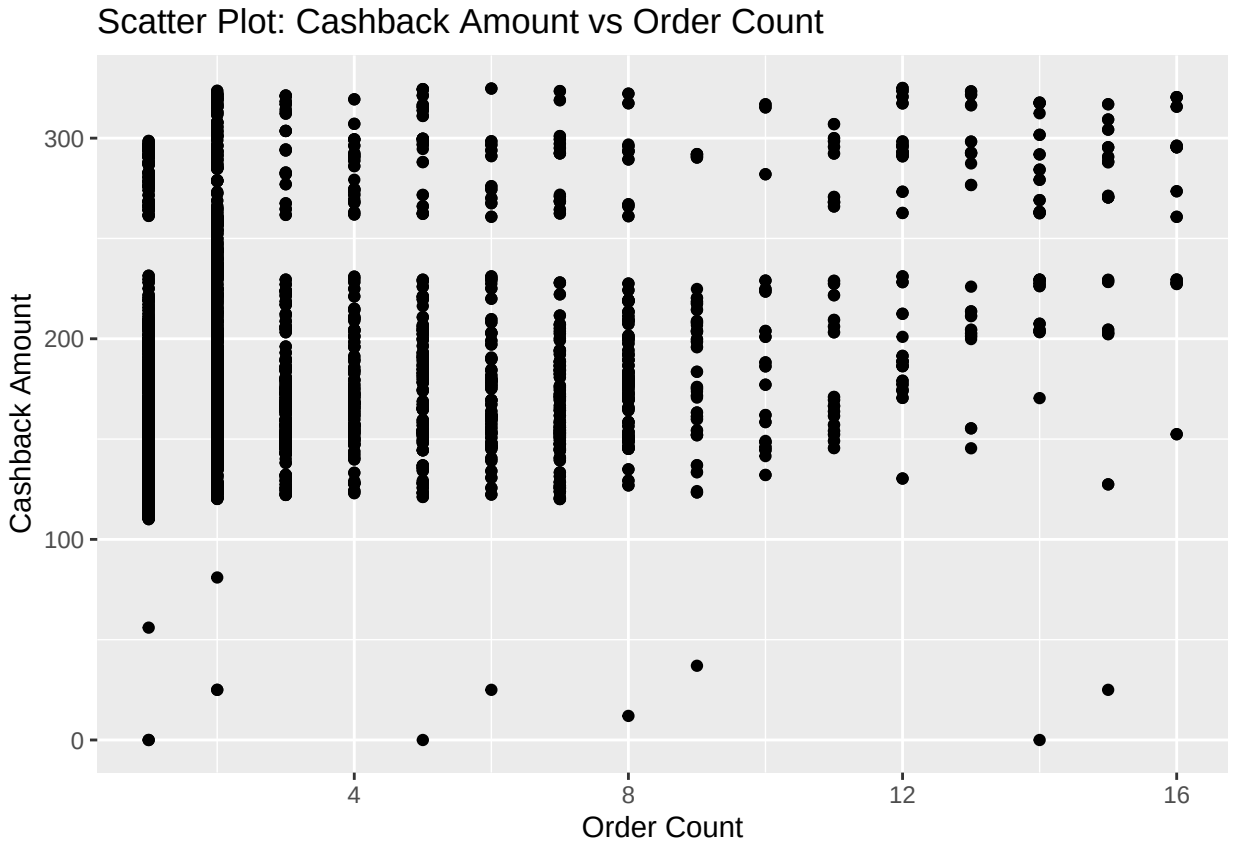

```
ggplot(data = data, aes(x = SatisfactionScore, y = CashbackAmount)) +
  geom_point() +
  labs(x = "Statisfaction Score", y = "Cashback Amount") +
  ggtitle("Scatter Plot: Cashback Amount vs Statisfaction Score")
```




```
ggplot(data = data, aes(x = CouponUsed, y = CashbackAmount)) +
  geom_point() +
  labs(x = "Coupon Used", y = "Cashback Amount") +
  ggtitle("Scatter Plot: Cashback Amount vs Coupon used")
```




```
ggplot(data = data, aes(x = OrderCount, y = CashbackAmount)) +
  geom_point() +
  labs(x = "Order Count", y = "Cashback Amount") +
  ggtitle("Scatter Plot: Cashback Amount vs Order Count")
```




```
model <- lm(data$CashbackAmount ~ data$Tenure + data$WarehouseToHome +
              data$HourSpendOnApp + data$SatisfactionScore  +
              data$CouponUsed + data$OrderCount)
summary(model)
```

```
##
## Call:
## lm(formula = data$CashbackAmount ~ data$Tenure + data$WarehouseToHome +
##     data$HourSpendOnApp + data$SatisfactionScore + data$CouponUsed +
##     data$OrderCount)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -267.327  -26.964   -5.747   20.146  157.481
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             119.36498    2.88455  41.381   <2e-16 ***
## data$Tenure               2.53039    0.06713  37.693   <2e-16 ***
## data$WarehouseToHome     -0.00377    0.06639  -0.057    0.955
## data$HourSpendOnApp       6.72694    0.80045   8.404   <2e-16 ***
## data$SatisfactionScore    0.08173    0.40091   0.204    0.838
## data$CouponUsed           0.13071    0.39291   0.333    0.739
## data$OrderCount           4.07094    0.25250  16.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 41.48 on 5623 degrees of freedom
## Multiple R-squared:   0.29,  Adjusted R-squared:  0.2893
## F-statistic: 382.8 on 6 and 5623 DF,  p-value: < 2.2e-16
```

Insight: Among the independent columns, Tenure, WarehouseToHome, and HourSpendOnApp seem to have a statistically significant relationship with CashbackAmount. However, SatisfactionScore and CouponUsed don't appear to have a significant linear relationship with CashbackAmount based on their p-values. However, this model explains about 23.53% of the variance in CashbackAmount, which means even the statistically significant columns don't explain much of the variance in cash back amount. The CEO might need to gather other columns of data to have a more linear and better relationship of cash back amount.

b) Report the regression equation, a 90% confidence interval for the coefficient of tenure, and the coefficient of determination.

```
# Report the regression equation
cat("Regression Equation:\n")
```

```
## Regression Equation:
```

```
cat("Cash back amount =", coef(model)[1], "+", coef(model)[2], "* data$Tenure +",
    coef(model)[3], "* data$WarehouseToHome +", coef(model)[4], "* data$HourSpendOnApp +",
    coef(model)[5], "* data$SatisfactionScore +", coef(model)[6], "* data$CouponUsed +",
    coef(model)[7], "* data$OrderCount\n")
```

```
## Cash back amount = 119.365 + 2.530388 * data$Tenure + -0.003769901 * data$WarehouseToHome + 6.726937
```

```
# 90% confidence interval for the coefficient of data$Tenure
conf_interval <- confint(model, "data$Tenure", level = 0.9)
cat("90% Confidence Interval for the Coefficient of data$Tenure:\n")
```

```
## 90% Confidence Interval for the Coefficient of data$Tenure:
```

```
cat(paste("(", conf_interval[1], ",", conf_interval[2], ")\n\n"))
```

```
## ( 2.41994755373134 , 2.64082849317187 )
```

```
# Coefficient of determination
cat("Coefficient of Determination (R-squared):", summary(model)$r.squared, "\n")
```

```
## Coefficient of Determination (R-squared): 0.2900264
```

c) Construct a 95% confidence interval for the slope of the estimated regression equation and interpret the results.

```r
confint(model, level=0.95)
```

```
##                               2.5 %      97.5 %
## (Intercept)            113.7101540 125.0198130
## data$Tenure              2.3987834   2.6619926
## data$WarehouseToHome    -0.1339281   0.1263883
## data$HourSpendOnApp      5.1577421   8.2961320
## data$SatisfactionScore  -0.7042175   0.8676717
## data$CouponUsed         -0.6395429   0.9009619
## data$OrderCount          3.5759504   4.5659290
```

Insight: The 95% confidence intervals for the regression coefficients provide a range of plausible values for the impact of each variable on the cashback amount. For instance, there is 95% confidence that a one-unit increase in data$Tenure is associated with an increase in cashback amount between 2.40 and 2.66.