ALY 6120-EDA-Module3

Introduction

Below is an simple EDA of our e-commerce-dataset to uncover simple patterns before we use predictive analysis.

Loading the dataset

At a glance we can see a quick summary statistics fo each variable for the dataset. Although this isn't ideal, it does give us a brief overview of what we might encounter.

```
# load packages
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.1     ✔ stringr   1.6.0
✔ ggplot2   4.0.0     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.2.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
```

```
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
# load dataset
store <- read.csv("e-commerce-dataset.csv")

# review raw data
glimpse(store) # check structure
```

```
Rows: 5,630
Columns: 20
$ CustomerID                <int> 50001, 50002, 50003, 50004, 50005, 5000
$ Churn                     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ Tenure                    <int> 4, NA, NA, 0, 0, 0, NA, NA, 13, NA, 4,
$ PreferredLoginDevice      <chr> "Mobile Phone", "Phone", "Phone", "Phon
$ CityTier                  <int> 3, 1, 1, 3, 1, 1, 3, 1, 3, 1, 1, 1, 1,
$ WarehouseToHome           <int> 6, 8, 30, 15, 12, 22, 11, 6, 9, 31, 18,
$ PreferredPaymentMode      <chr> "Debit Card", "UPI", "Debit Card", "Deb
$ Gender                    <chr> "Female", "Male", "Male", "Male", "Male
$ HourSpendOnApp            <int> 3, 3, 2, 2, NA, 3, 2, 3, NA, 2, 2, 3, 2
$ NumberOfDeviceRegistered  <int> 3, 4, 4, 4, 3, 5, 3, 3, 4, 5, 3, 4, 3,
$ PreferedOrderCat          <chr> "Laptop & Accessory", "Mobile", "Mobile
$ SatisfactionScore         <int> 2, 3, 3, 5, 5, 5, 2, 2, 3, 3, 3, 3, 3,
$ MaritalStatus             <chr> "Single", "Single", "Single", "Single",
$ NumberOfAddress           <int> 9, 7, 6, 8, 3, 2, 4, 3, 2, 2, 2, 10, 2,
$ Complain                  <int> 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1,
$ OrderAmountHikeFromlastYear <int> 11, 15, 14, 23, 11, 22, 14, 16, 14, 12,
$ CouponUsed                <int> 1, 0, 0, 0, 1, 4, 0, 2, 0, 1, 9, 0, 2,
$ OrderCount                <int> 1, 1, 1, 1, 1, 6, 1, 2, 1, 1, 15, 1, 2,
$ DaySinceLastOrder         <int> 5, 0, 3, 3, 3, 7, 0, 0, 2, 1, 8, 0, 2,
$ CashbackAmount            <int> 160, 121, 120, 134, 130, 139, 121, 123,
```

```r
# clean column names
store <- store %>% clean_names()
```

```r
# check the summary stats
summary(store)
```

```
  customer_id        churn            tenure      preferred_login_device
 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
  city_tier     warehouse_to_home preferred_payment_mode    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
hour_spend_on_app number_of_device_registered prefered_order_cat
Min.   :0.000     Min.   :1.000               Length:5630
1st Qu.:2.000     1st Qu.:3.000               Class :character
Median :3.000     Median :4.000               Mode  :character
Mean   :2.932     Mean   :3.689
3rd Qu.:3.000     3rd Qu.:4.000
Max.   :5.000     Max.   :6.000
NA's   :255
satisfaction_score marital_status     number_of_address    complain
Min.   :1.000      Length:5630        Min.   : 1.000    Min.   :0.0000
1st Qu.:2.000      Class :character   1st Qu.: 2.000    1st Qu.:0.0000
Median :3.000      Mode  :character   Median : 3.000    Median :0.0000
Mean   :3.067                         Mean   : 4.214    Mean   :0.2849
3rd Qu.:4.000                         3rd Qu.: 6.000    3rd Qu.:1.0000
Max.   :5.000                         Max.   :22.000    Max.   :1.0000

order_amount_hike_fromlast_year  coupon_used      order_count
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
day_since_last_order cashback_amount
Min.   : 0.000       Min.   :  0.0
1st Qu.: 2.000       1st Qu.:146.0
Median : 3.000       Median :163.0
Mean   : 4.543       Mean   :177.2
3rd Qu.: 7.000       3rd Qu.:196.0
Max.   :46.000       Max.   :325.0
NA's   :307
```

Missing Values

The dataset does contain missing values.

```r
# check for missing values in each column
colSums(is.na(store))
```

```
                  customer_id                         churn
                            0                             0
                       tenure       preferred_login_device
                          264                             0
                  city_tier             warehouse_to_home
                            0                           251
       preferred_payment_mode                        gender
                            0                             0
            hour_spend_on_app  number_of_device_registered
                          255                             0
          prefered_order_cat           satisfaction_score
                            0                             0
              marital_status            number_of_address
                            0                             0
                    complain order_amount_hike_fromlast_year
                            0                           265
                coupon_used                   order_count
                          256                           258
          day_since_last_order               cashback_amount
                          307                             0
```

Removing missing values

Missing values are now removed for easier EDA. We need to discuss if imputing the values would be better for prediction models.

```r
# drop rows with missing values
store <- store %>% drop_na()  # remove incomplete rows

# confirm values were removed
colSums(is.na(store))
```

```
                  customer_id                         churn
                            0                             0
                       tenure       preferred_login_device
                            0                             0
                  city_tier             warehouse_to_home
                            0                             0
       preferred_payment_mode                        gender
```

```
                            0                                   0
             hour_spend_on_app     number_of_device_registered
                            0                                   0
            prefered_order_cat              satisfaction_score
                            0                                   0
                marital_status               number_of_address
                            0                                   0
                      complain order_amount_hike_fromlast_year
                            0                                   0
                   coupon_used                     order_count
                            0                                   0
          day_since_last_order                 cashback_amount
                            0                                   0
```

Changing structure

The dataset will need to change certain variable's structure.

```r
# convert character columns to factors
store <- store %>% mutate(across(where(is.character), as.factor))

# convert integer columns to numeric
store <- store %>% mutate(across(where(is.integer), as.numeric))

# confirm updated structure
str(store)  # check final structure after cleaning
```

```
'data.frame':   3774 obs. of  20 variables:
 $ customer_id                   : num  50001 50004 50006 50012 50013 ...
 $ churn                         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ tenure                        : num  4 0 0 11 0 0 9 0 0 19 ...
 $ preferred_login_device        : Factor w/ 3 levels "Computer","Mobile P
 $ city_tier                     : num  3 3 1 1 1 1 3 3 1 1 ...
 $ warehouse_to_home             : num  6 15 22 6 11 15 15 11 13 20 ...
 $ preferred_payment_mode        : Factor w/ 7 levels "Cash on Delivery",.
 $ gender                        : Factor w/ 2 levels "Female","Male": 1 2
 $ hour_spend_on_app             : num  3 2 3 3 2 3 3 2 3 3 ...
 $ number_of_device_registered   : num  3 4 5 4 3 4 4 4 5 3 ...
 $ prefered_order_cat            : Factor w/ 6 levels "Fashion","Grocery",
 $ satisfaction_score            : num  2 5 5 3 3 3 2 3 3 4 ...
 $ marital_status                : Factor w/ 3 levels "Divorced","Married"
 $ number_of_address             : num  9 8 2 10 2 1 2 2 2 10 ...
 $ complain                      : num  1 0 1 1 1 1 0 1 1 1 ...
 $ order_amount_hike_fromlast_year: num  11 23 22 13 13 17 16 11 24 18 ...
```

```
 $ coupon_used                : num  1 0 4 0 2 0 0 1 1 1 ...
 $ order_count                : num  1 1 6 1 2 1 4 1 1 4 ...
 $ day_since_last_order       : num  5 3 7 0 2 0 7 3 6 3 ...
 $ cashback_amount            : num  160 134 139 154 134 134 196 157 16
```

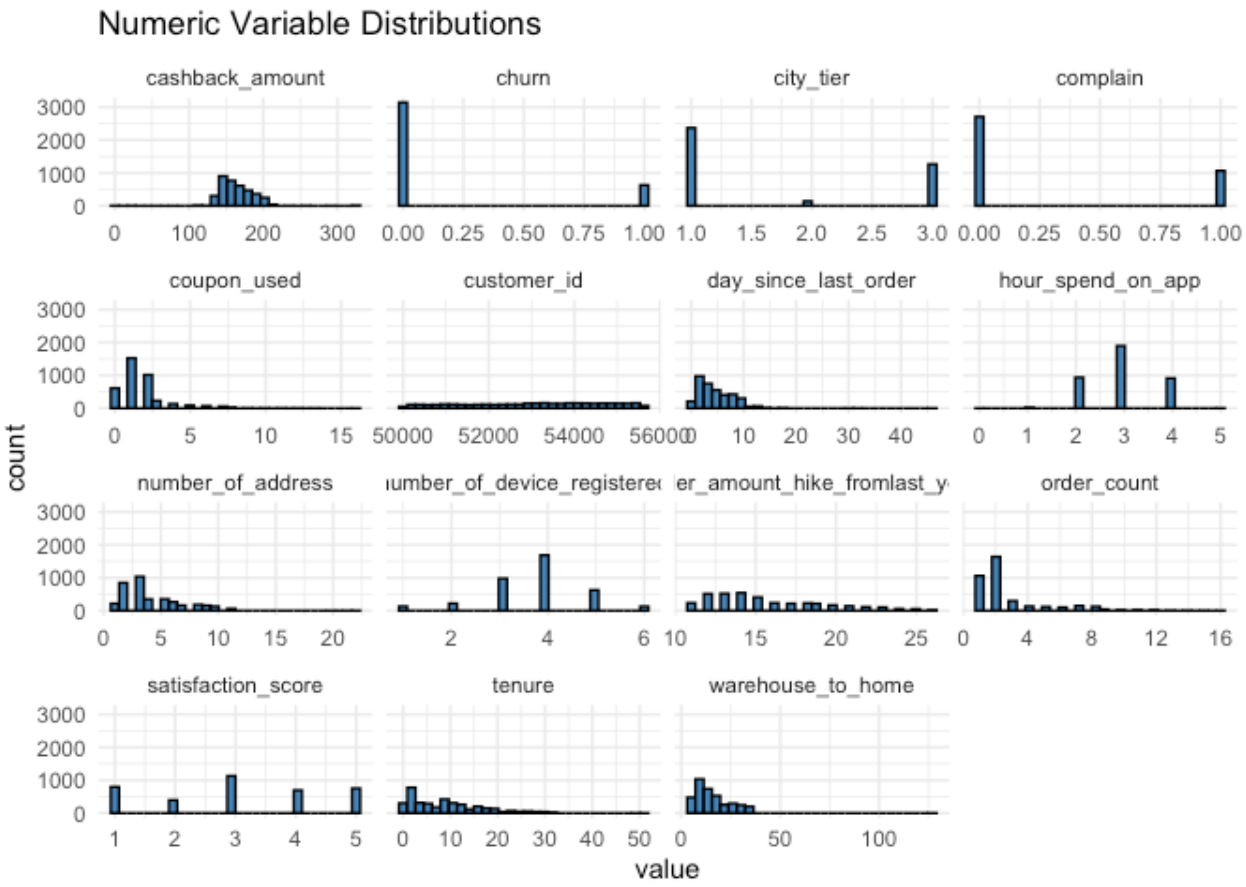
Visualizations and Early Findings

Below are charts that can tell us patters in the dataset.

Distributions

Most variables are leaning toward lower values, so the data isn't evenly spread out. We can see that most customers behave the same way, while only a small group showing higher numbers.

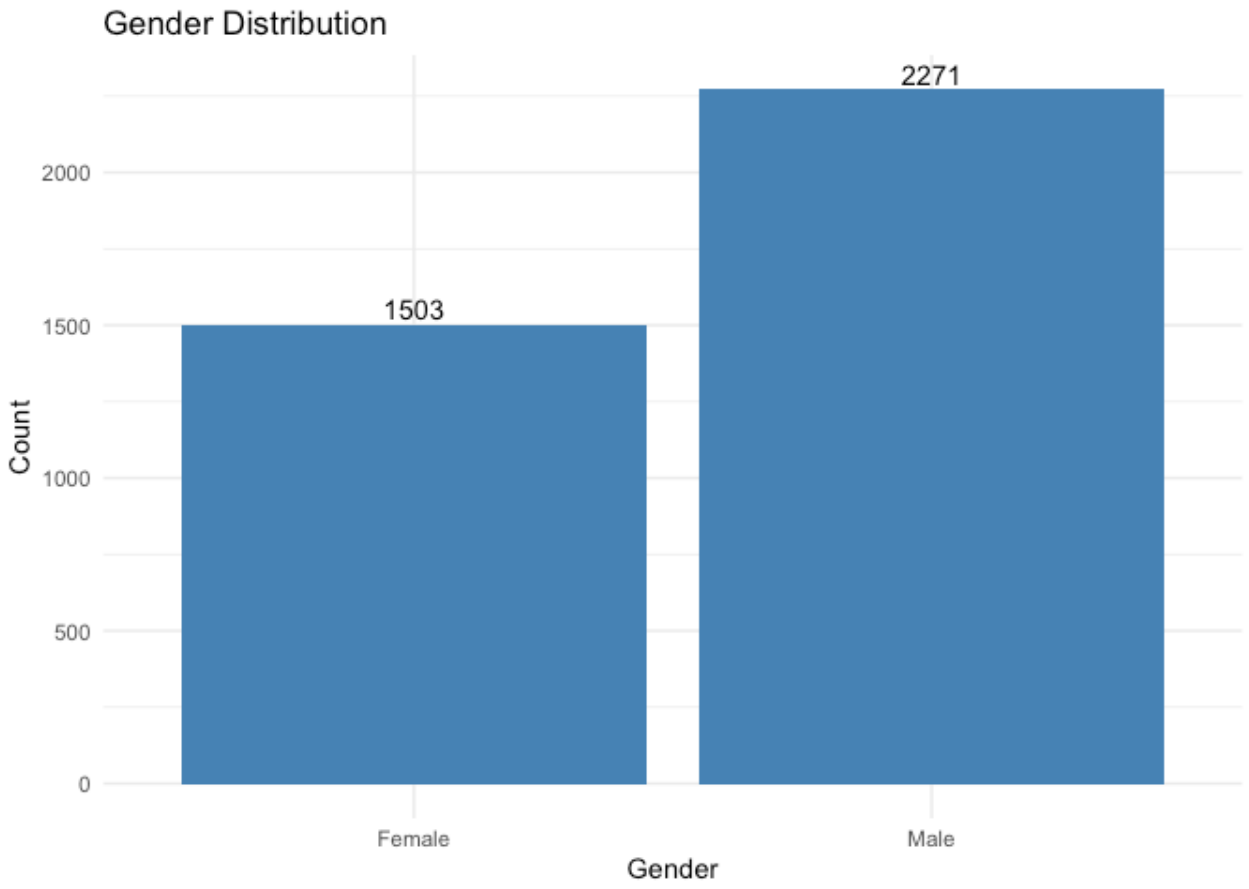
```r
# numeric distributions
store %>%
  select(where(is.numeric)) %>%
  gather(variable, value) %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black") +
  facet_wrap(~ variable, scales = "free_x") +
  labs(title = "Numeric Variable Distributions") +
  theme_minimal()
```



Bar charts

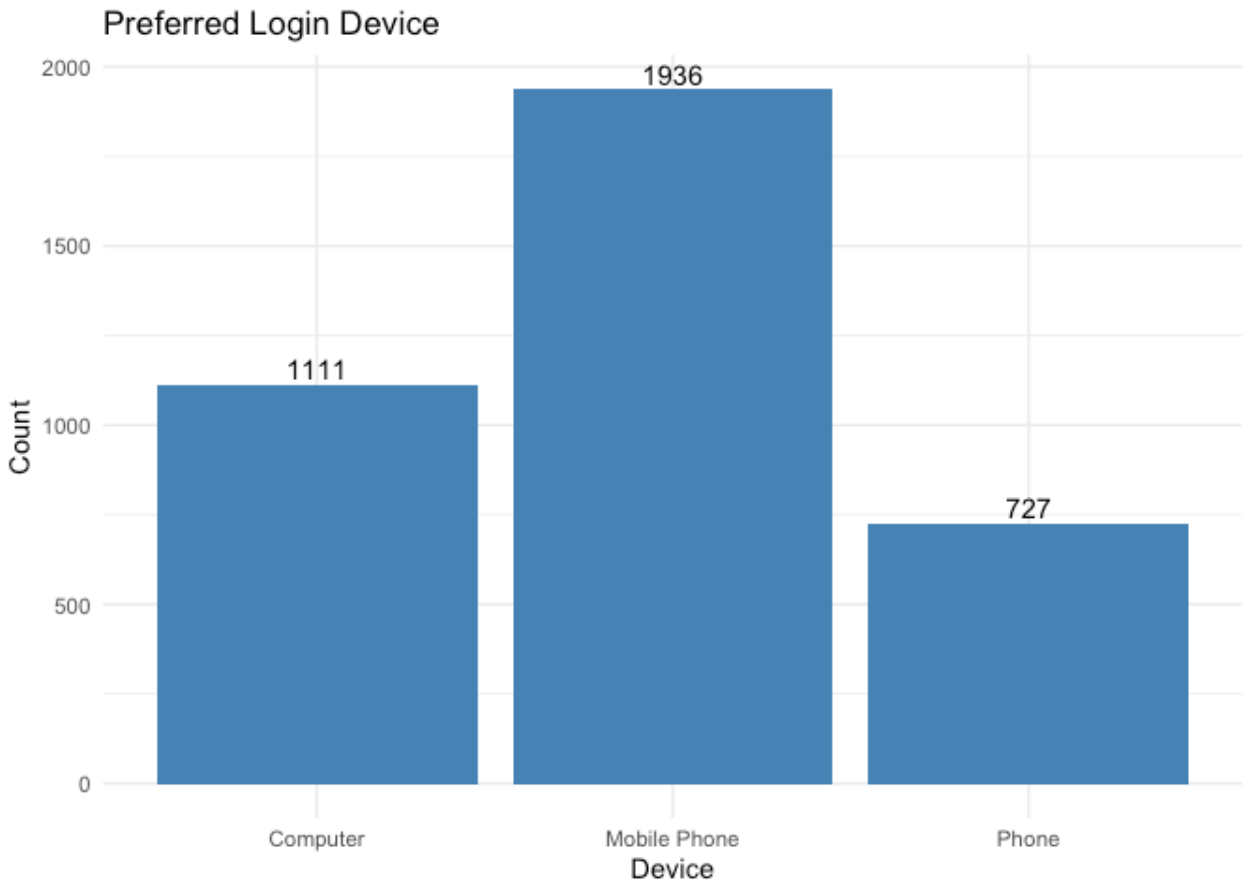
Male customers make up the larger share of the dataset, with 2271 entries. Female customers are fewer at 1503. This means the store's customer base skews more male.

```
#  Gender Distribution
ggplot(store, aes(x = gender)) +
  geom_bar(fill = "steelblue") +
    geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.3)
  labs(title = "Gender Distribution",
       x = "Gender",
       y = "Count") +
  theme_minimal()
```



Most customers log in with a mobile phone, which has the highest count at 1936. computers come next with 1111 users. Regular phones are the least used at 727. This shows that mobile is the main way people access the store.

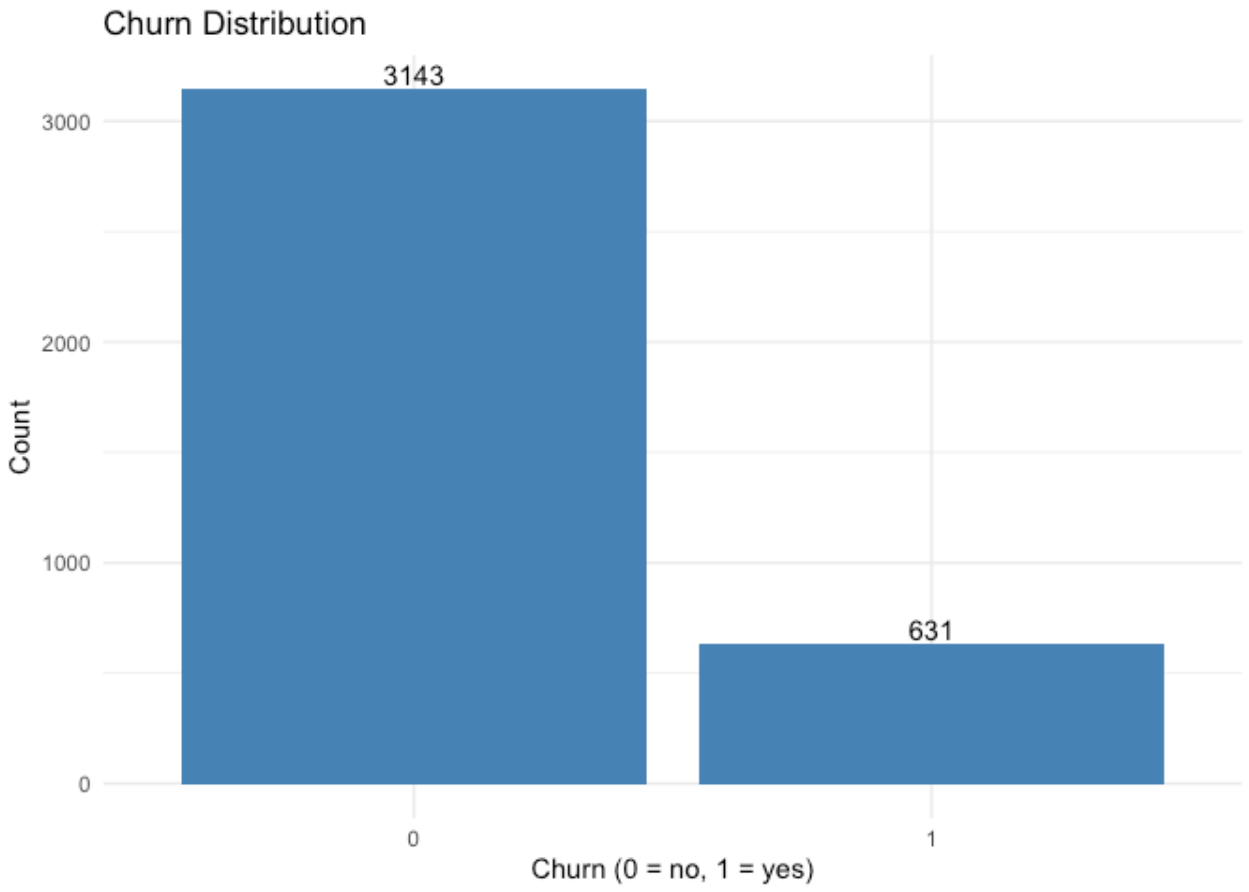
```
# preferred login device
ggplot(store, aes(x = preferred_login_device)) +
  geom_bar(fill = "steelblue") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.3) +
  labs(title = "Preferred Login Device",
       x = "Device",
       y = "Count") +
  theme_minimal()
```



Churn Findings

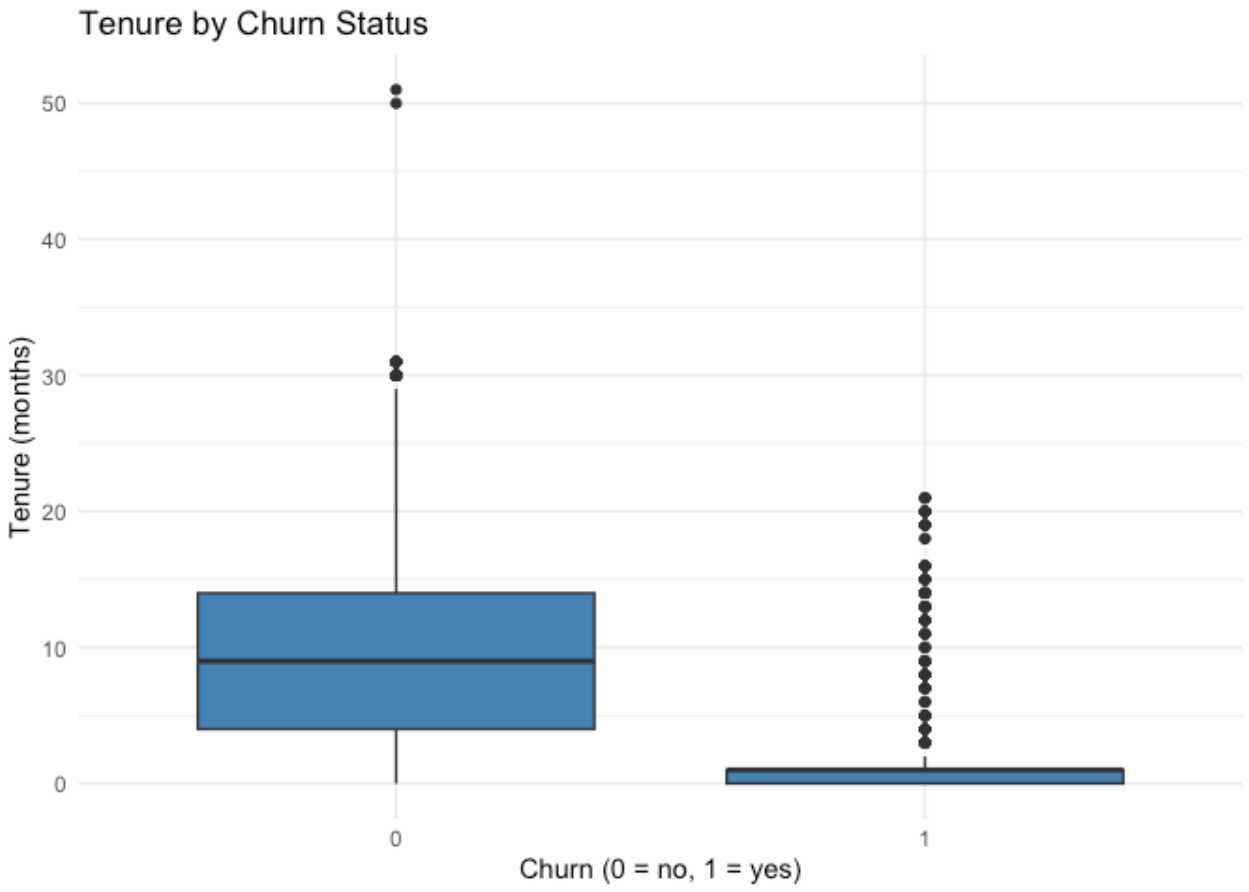
Most customers stay with the store, with 3143 not churning. only 631 customers churned. That means churn is much lower than retention in this dataset.

```
# churn distribution
ggplot(store, aes(x = factor(churn))) +
  geom_bar(fill = "steelblue") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.3) +
  labs(title = "Churn Distribution",
       x = "Churn (0 = no, 1 = yes)",
       y = "Count") +
  theme_minimal()
```



Customers who churn have very short tenure, almost all close to zero. customers who stay have much higher tenure and a wider spread.

```
# tenure by churn status boxplot
ggplot(store, aes(x = factor(churn), y = tenure)) +
  geom_boxplot(fill = "steelblue") +
  labs(title = "Tenure by Churn Status",
       x = "Churn (0 = no, 1 = yes)",
       y = "Tenure (months)") +
  theme_minimal()
```



Correlation Martix

Variables like cashback amount, order count, coupon used, and spending all show high correlations, meaning customers who spend more tend to do all of these together. Churn and tenure show a strong negative correlation, meaning people with low tenure are more likely to leave. The rest of the variables have low correlations, so they don't move much with anything else.

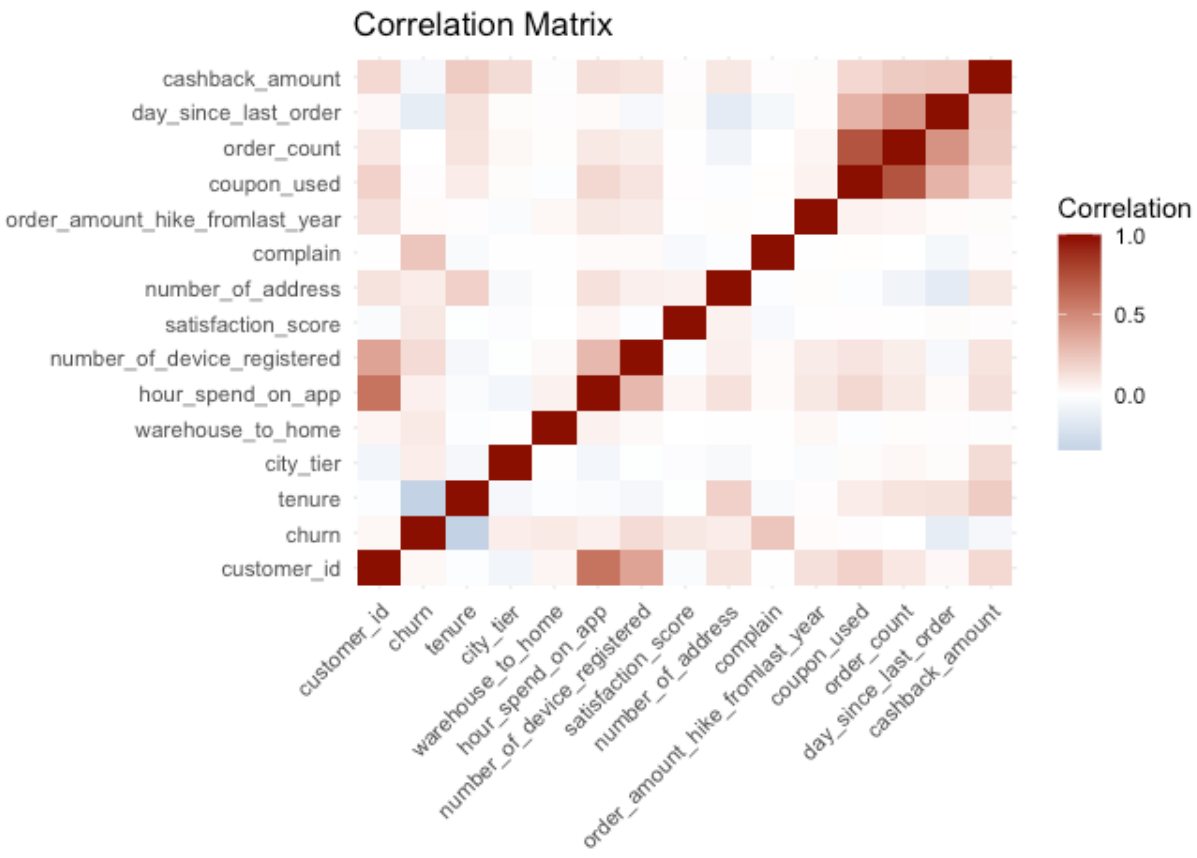
```
# prep for correlation but selecing numeric columns only
numeric_store <- store %>% select(where(is.numeric))  # keep only numeric v

# create correlation matrix
cor_matrix <- cor(numeric_store, use = "complete.obs")

# convert matrix to long format for plotting
melted_cor <- as.data.frame(as.table(cor_matrix))

# plot correlation martix
ggplot(melted_cor, aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient2(low = "steelblue", high = "darkred") +
  labs(title = "Correlation Matrix",
```

```
        x = "",
        y = "",
        fill = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Conclusion

The patterns in this dataset give a good base for data mining because some variables clearly separate customer behavior. The strong links between spending-related features can support clustering, while the churn-tenure relationship is a solid starting point for prediction models. Next steps would be building a churn model, testing which features actually drive churn, and creating customer segments to target high-value or high-risk groups more effectively.