Store Sales 2021

2024-08-08

Final Project

Introduction to the Data set: Dataset Link - https://www.kaggle.com/datasets/pipchu/sales-2020-2021

The dataset (sales_06_FY2020-21)contains sales transaction data for the fiscal year 22020–2021. the dataset is quite extensive, with 286,392 entries and 36 columns. Here is a brief overview of the key columns and the data they represent:

order_id: Unique identifier for each order.

order_date: Date when the order was placed.

status: status of the order.

item_id: A unique identifier for each item in the order.

sku: stock-keeping unit, a unique identifier for the product.

qty_ordered: quantity of the item ordered.

price: price per unit of the item.

value: the total value of the item in the order (quantity * price).

discount_amount: Discount applied to the item.

total: the final total for the item after discounts.

category: category of the product.

payment_method: Payment method used for the order.

bi_st: Possibly billing state or a similar identifier.

cust_id: A unique identifier for the customer.

year: year of the transaction.

month: month of the transaction.

full_name: Full name of the customer.

Gender: gender of the customer.

age: age of the customer.

E Mail: Email address of the customer.

SSN: Social Security Number of the customer.

Phone No.: The phone number of the customer.

Place Name: The location name associated with the customer.

County: county associated with the customer.

City: City associated with the customer.

State: State associated with the customer.

Zip: ZIP code associated with the customer.

Region: Region associated with the customer.

User Name: Username of the customer.

Discount_Percent: Percentage of discount applied.

Summary of the Data:

Data Type: The dataset includes a mix of numeric and categorical variables, with some columns (like order_id, sku, and bi_st) being of object (string) type, while others (like price, quantity_ordered, and age) are numeric.

Size: The dataset is large, containing 286,392 rows, which will provide a robust base for any analysis.

```
library(ggplot2) #For visualization
library(dplyr) #Data Manipulation
```

```
##
## 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
```

```
options(scipen = 999) # Helps Scientific notation in numeric outputs


setwd("/Users/saikalyangurumanchi/Library/CloudStorage/OneDrive-CentralMichiganUniversity/Acadamics/Summ
Sales_2020_2021 <- read.csv("sales_06_FY2020-21.csv")
```

Data Cleaning

```
#Converting the date into "%Y-%m-%d" formate
Sales_2020_2021$order_date <- as.Date(Sales_2020_2021$order_date, format="%Y-%m-%d")

#filtering the sales data of 2021 year
Sales_2021 <- Sales_2020_2021 %>%
  filter(format(order_date, "%Y") == "2021")

#View(Sales_2021)
no_of_rows <- nrow(Sales_2021)
print(no_of_rows) #177213 rows
```

```
## [1] 177213
```

```
#Viewing Column names and datatype
glimpse(Sales_2021)
```

```
## Rows: 177,213
## Columns: 36
## $ order_id         <chr> "100441393", "100533723", "100533724", "100533726", "~
## $ order_date       <date> 2021-02-01, 2021-06-25, 2021-06-25, 2021-06-25, 2021~
## $ status           <chr> "canceled", "canceled", "canceled", "canceled", "canc~
## $ item_id          <dbl> 723634, 855999, 856000, 856002, 856003, 856005, 85601~
## $ sku              <chr> "APPORI5A433F2870FC6", "MEFVIT5A9D50E8177A1", "MEFVIT~
## $ qty_ordered      <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 3, 2, 3, 3,~
## $ price            <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9, 274~
## $ value            <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9, 274~
## $ discount_amount  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ total            <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9, 274~
## $ category         <chr> "Appliances", "Men's Fashion", "Men's Fashion", "Men'~
## $ payment_method   <chr> "Payaxis", "Payaxis", "Payaxis", "Payaxis", "Payaxis"~
## $ bi_st            <chr> "Gross", "Gross", "Gross", "Gross", "Gross", "Gross",~
## $ cust_id          <dbl> 56449, 56449, 56449, 56449, 56449, 56449, 56449, 5644~
## $ year             <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021,~
## $ month            <chr> "Feb-2021", "Jun-2021", "Jun-2021", "Jun-2021", "Jun-~
## $ ref_num          <int> 849200, 849200, 849200, 849200, 849200, 849200, 84920~
## $ Name.Prefix      <chr> "Ms.", "Ms.", "Ms.", "Ms.", "Ms.", "Ms.", "Ms.", "Ms.~
## $ First.Name       <chr> "Elizbeth", "Elizbeth", "Elizbeth", "Elizbeth", "Eliz~
## $ Middle.Initial   <chr> "E", "E", "E", "E", "E", "E", "E", "E", "E", "E", "E"~
## $ Last.Name        <chr> "Raminez", "Raminez", "Raminez", "Raminez", "Raminez"~
## $ Gender           <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F"~
## $ age              <dbl> 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 7~
## $ full_name        <chr> "Raminez, Elizbeth", "Raminez, Elizbeth", "Raminez, E~
## $ E.Mail           <chr> "elizbeth.raminez@gmail.com", "elizbeth.raminez@gmail~
## $ Customer.Since   <chr> "3/31/2017", "3/31/2017", "3/31/2017", "3/31/2017", "~
## $ SSN              <chr> "023-92-2382", "023-92-2382", "023-92-2382", "023-92-~
## $ Phone.No.        <chr> "217-405-0048", "217-405-0048", "217-405-0048", "217-~
## $ Place.Name       <chr> "Brownstown", "Brownstown", "Brownstown", "Brownstown~
## $ County           <chr> "Fayette", "Fayette", "Fayette", "Fayette", "Fayette"~
## $ City             <chr> "Brownstown", "Brownstown", "Brownstown", "Brownstown~
## $ State            <chr> "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL",~
## $ Zip              <int> 62418, 62418, 62418, 62418, 62418, 62418, 62418, 6241~
## $ Region           <chr> "Midwest", "Midwest", "Midwest", "Midwest", "Midwest"~
## $ User.Name        <chr> "eeraminez", "eeraminez", "eeraminez", "eeraminez", "~
## $ Discount_Percent <dbl> 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 0.00000,~
```

```
#Removing the column names that are not required
Sales_2021 <- Sales_2021[, !colnames(Sales_2021) %in% c(
  'item_id', 'sku', 'bi_st', 'order_date', 'ref_num',
  'Middle.Initial', 'Name.Prefix', 'full_name', 'price', 'qty_ordered', 'Discount_Percent', 'County', '

glimpse(Sales_2021)
```

```
## Rows: 177,213
## Columns: 23
## $ order_id        <chr> "100441393", "100533723", "100533724", "100533726", "1~
## $ status          <chr> "canceled", "canceled", "canceled", "canceled", "cance~
## $ value           <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9, 274.~
## $ discount_amount <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total           <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9, 274.~
## $ category        <chr> "Appliances", "Men's Fashion", "Men's Fashion", "Men's~
```

```
## $ payment_method  <chr> "Payaxis", "Payaxis", "Payaxis", "Payaxis", "Payaxis",~
## $ cust_id         <dbl> 56449, 56449, 56449, 56449, 56449, 56449, 56449, 56449~
## $ year            <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, ~
## $ month           <chr> "Feb-2021", "Jun-2021", "Jun-2021", "Jun-2021", "Jun-2~
## $ First.Name      <chr> "Elizbeth", "Elizbeth", "Elizbeth", "Elizbeth", "Elizb~
## $ Last.Name       <chr> "Raminez", "Raminez", "Raminez", "Raminez", "Raminez",~
## $ Gender          <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F",~
## $ age             <dbl> 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71~
## $ E.Mail          <chr> "elizbeth.raminez@gmail.com", "elizbeth.raminez@gmail.~
## $ Customer.Since  <chr> "3/31/2017", "3/31/2017", "3/31/2017", "3/31/2017", "3~
## $ SSN             <chr> "023-92-2382", "023-92-2382", "023-92-2382", "023-92-2~
## $ Phone.No.       <chr> "217-405-0048", "217-405-0048", "217-405-0048", "217-4~
## $ City            <chr> "Brownstown", "Brownstown", "Brownstown", "Brownstown"~
## $ State           <chr> "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL", ~
## $ Zip             <int> 62418, 62418, 62418, 62418, 62418, 62418, 62418, 62418~
## $ Region          <chr> "Midwest", "Midwest", "Midwest", "Midwest", "Midwest",~
## $ User.Name       <chr> "eeraminez", "eeraminez", "eeraminez", "eeraminez", "e~
```

```
#Changing the column names
new_names <- c('order_id', 'status', 'price', 'discount_amount', 'price_after_discount', 'category', 'pa
               , 'year', 'month', 'first_name', 'last_name', 'sex', 'age', 'E-mail', 'customer_since',
               , 'zip', 'region', 'user_name')
names(Sales_2021) <- new_names #Assigning the new column name to the Sales_2021 dataset

glimpse(Sales_2021)
```

```
## Rows: 177,213
## Columns: 23
## $ order_id             <chr> "100441393", "100533723", "100533724", "100533726~
## $ status               <chr> "canceled", "canceled", "canceled", "canceled", "~
## $ price                <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9,~
## $ discount_amount      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ price_after_discount <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9,~
## $ category             <chr> "Appliances", "Men's Fashion", "Men's Fashion", "~
## $ payment_type         <chr> "Payaxis", "Payaxis", "Payaxis", "Payaxis", "Paya~
## $ customer_id          <dbl> 56449, 56449, 56449, 56449, 56449, 56449, 56449, ~
## $ year                 <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2~
## $ month                <chr> "Feb-2021", "Jun-2021", "Jun-2021", "Jun-2021", "~
## $ first_name           <chr> "Elizbeth", "Elizbeth", "Elizbeth", "Elizbeth", "~
## $ last_name            <chr> "Raminez", "Raminez", "Raminez", "Raminez", "Rami~
## $ sex                  <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F",~
## $ age                  <dbl> 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 7~
## $ `E-mail`             <chr> "elizbeth.raminez@gmail.com", "elizbeth.raminez@g~
## $ customer_since       <chr> "3/31/2017", "3/31/2017", "3/31/2017", "3/31/2017~
## $ SSN                  <chr> "023-92-2382", "023-92-2382", "023-92-2382", "023~
## $ phone_num            <chr> "217-405-0048", "217-405-0048", "217-405-0048", "~
## $ city                 <chr> "Brownstown", "Brownstown", "Brownstown", "Browns~
## $ state                <chr> "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL", "~
## $ zip                  <int> 62418, 62418, 62418, 62418, 62418, 62418, 62418, ~
## $ region               <chr> "Midwest", "Midwest", "Midwest", "Midwest", "Midw~
## $ user_name            <chr> "eeraminez", "eeraminez", "eeraminez", "eeraminez~
```

```r
#Reordering the columns in the order
Sales_2021 <- Sales_2021 %>% select('user_name','customer_id','first_name', 'last_name', 'sex', 'age',
                                    'price_after_discount', 'payment_type', 'month', 'year', 'status',
                                    'zip', 'region','E-mail')



glimpse(Sales_2021)
```

```
## Rows: 177,213
## Columns: 23
## $ user_name            <chr> "eeraminez", "eeraminez", "eeraminez", "eeraminez~
## $ customer_id          <dbl> 56449, 56449, 56449, 56449, 56449, 56449, 56449, ~
## $ first_name           <chr> "Elizbeth", "Elizbeth", "Elizbeth", "Elizbeth", "~
## $ last_name            <chr> "Raminez", "Raminez", "Raminez", "Raminez", "Rami~
## $ sex                  <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F",~
## $ age                  <dbl> 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 71, 7~
## $ order_id             <chr> "100441393", "100533723", "100533724", "100533726~
## $ category             <chr> "Appliances", "Men's Fashion", "Men's Fashion", "~
## $ price                <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9,~
## $ discount_amount      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ price_after_discount <dbl> 2625.0, 299.9, 299.9, 299.9, 299.9, 1600.0, 59.9,~
## $ payment_type         <chr> "Payaxis", "Payaxis", "Payaxis", "Payaxis", "Paya~
## $ month                <chr> "Feb-2021", "Jun-2021", "Jun-2021", "Jun-2021", "~
## $ year                 <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2~
## $ status               <chr> "canceled", "canceled", "canceled", "canceled", "~
## $ customer_since       <chr> "3/31/2017", "3/31/2017", "3/31/2017", "3/31/2017~
## $ SSN                  <chr> "023-92-2382", "023-92-2382", "023-92-2382", "023~
## $ phone_num            <chr> "217-405-0048", "217-405-0048", "217-405-0048", "~
## $ city                 <chr> "Brownstown", "Brownstown", "Brownstown", "Browns~
## $ state                <chr> "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL", "~
## $ zip                  <int> 62418, 62418, 62418, 62418, 62418, 62418, 62418, ~
## $ region               <chr> "Midwest", "Midwest", "Midwest", "Midwest", "Midw~
## $ `E-mail`             <chr> "elizbeth.raminez@gmail.com", "elizbeth.raminez@g~
```

```r
#Changing the values in payment_type and status column which seems same

#Checking for unique values in the payment_type column
unique_payment_type <- unique(Sales_2021$payment_type)
print(unique_payment_type)
```

```
##  [1] "Payaxis"         "jazzwallet"      "Easypay_MA"      "cod"
##  [5] "Easypay"         "easypay_voucher" "jazzvoucher"     "customercredit"
##  [9] "bankalfalah"     "apg"
```

```r
# Transform the payment_types column into categories
Sales_2021 <- Sales_2021 %>%
  mutate(payment_type = case_when(
    payment_type %in% c('jazzwallet', 'Easypay', 'customercredit') ~ 'Digital Wallets',
    payment_type %in% c('Payaxis', 'Easypay_MA', 'apg', 'Easypay') ~ 'Payment Gateways',
    payment_type %in% c('easypay_voucher', 'jazzvoucher') ~ 'Voucher Systems',
    payment_type == 'bankalfalah' ~ 'Banking Services',
```

```
    TRUE ~ payment_type  # Keep the original value if none of the conditions match
  ))

# View the updated data set
head(Sales_2021, 10)
```

```
##    user_name customer_id first_name last_name sex age  order_id      category
## 1  eeraminez       56449   Elizbeth   Raminez   F  71 100441393    Appliances
## 2  eeraminez       56449   Elizbeth   Raminez   F  71 100533723 Men's Fashion
## 3  eeraminez       56449   Elizbeth   Raminez   F  71 100533724 Men's Fashion
## 4  eeraminez       56449   Elizbeth   Raminez   F  71 100533726 Men's Fashion
## 5  eeraminez       56449   Elizbeth   Raminez   F  71 100533727 Men's Fashion
## 6  eeraminez       56449   Elizbeth   Raminez   F  71 100533729 Men's Fashion
## 7  eeraminez       56449   Elizbeth   Raminez   F  71 100533733 Men's Fashion
## 8  eeraminez       56449   Elizbeth   Raminez   F  71 100533735 Men's Fashion
## 9  eeraminez       56449   Elizbeth   Raminez   F  71 100533739 Men's Fashion
## 10 eeraminez       56449   Elizbeth   Raminez   F  71 100560796 Men's Fashion
##     price discount_amount price_after_discount     payment_type    month year
## 1  2625.0               0                2625.0 Payment Gateways Feb-2021 2021
## 2   299.9               0                 299.9 Payment Gateways Jun-2021 2021
## 3   299.9               0                 299.9 Payment Gateways Jun-2021 2021
## 4   299.9               0                 299.9 Payment Gateways Jun-2021 2021
## 5   299.9               0                 299.9 Payment Gateways Jun-2021 2021
## 6  1600.0               0                1600.0 Payment Gateways Jun-2021 2021
## 7    59.9               0                  59.9 Payment Gateways Jun-2021 2021
## 8   274.9               0                 274.9 Payment Gateways Jun-2021 2021
## 9    39.9               0                  39.9 Payment Gateways Jun-2021 2021
## 10    0.0               0                   0.0 Payment Gateways Sep-2021 2021
##      status customer_since         SSN    phone_num       city state   zip
## 1  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 2  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 3  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 4  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 5  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 6  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 7  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 8  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 9  canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
## 10 canceled      3/31/2017 023-92-2382 217-405-0048 Brownstown    IL 62418
##     region                     E-mail
## 1  Midwest elizbeth.raminez@gmail.com
## 2  Midwest elizbeth.raminez@gmail.com
## 3  Midwest elizbeth.raminez@gmail.com
## 4  Midwest elizbeth.raminez@gmail.com
## 5  Midwest elizbeth.raminez@gmail.com
## 6  Midwest elizbeth.raminez@gmail.com
## 7  Midwest elizbeth.raminez@gmail.com
## 8  Midwest elizbeth.raminez@gmail.com
## 9  Midwest elizbeth.raminez@gmail.com
## 10 Midwest elizbeth.raminez@gmail.com
```

```
# Categorize the status column
unique_status <- unique(Sales_2021$status)
```

```r
print(unique_status)
```

```
##  [1] "canceled"       "complete"       "received"       "order_refunded"
##  [5] "refund"         "cod"            "paid"           "processing"
##  [9] "pending"        "closed"         "payment_review" "holded"
```

```r
Sales_2021 <- Sales_2021 %>%
  mutate(status = case_when(
    status %in% c("complete", "received", "closed", "paid") ~ "Completed or Fulfilled",
    status %in% c("processing", "pending", "payment_review", "holded") ~ "In Progress or Awaiting Actio
    status %in% c("canceled", "order_refunded", "refund") ~ "Refund or Cancellation",
    TRUE ~ status  # Fallback category for any unexpected values
  ))

# View the updated data set
head(Sales_2021, 10)
```

```
##    user_name customer_id first_name last_name sex age  order_id      category
## 1  eeraminez       56449   Elizbeth   Raminez   F  71 100441393    Appliances
## 2  eeraminez       56449   Elizbeth   Raminez   F  71 100533723 Men's Fashion
## 3  eeraminez       56449   Elizbeth   Raminez   F  71 100533724 Men's Fashion
## 4  eeraminez       56449   Elizbeth   Raminez   F  71 100533726 Men's Fashion
## 5  eeraminez       56449   Elizbeth   Raminez   F  71 100533727 Men's Fashion
## 6  eeraminez       56449   Elizbeth   Raminez   F  71 100533729 Men's Fashion
## 7  eeraminez       56449   Elizbeth   Raminez   F  71 100533733 Men's Fashion
## 8  eeraminez       56449   Elizbeth   Raminez   F  71 100533735 Men's Fashion
## 9  eeraminez       56449   Elizbeth   Raminez   F  71 100533739 Men's Fashion
## 10 eeraminez       56449   Elizbeth   Raminez   F  71 100560796 Men's Fashion
##     price discount_amount price_after_discount     payment_type    month year
## 1  2625.0               0               2625.0 Payment Gateways Feb-2021 2021
## 2   299.9               0                299.9 Payment Gateways Jun-2021 2021
## 3   299.9               0                299.9 Payment Gateways Jun-2021 2021
## 4   299.9               0                299.9 Payment Gateways Jun-2021 2021
## 5   299.9               0                299.9 Payment Gateways Jun-2021 2021
## 6  1600.0               0               1600.0 Payment Gateways Jun-2021 2021
## 7    59.9               0                 59.9 Payment Gateways Jun-2021 2021
## 8   274.9               0                274.9 Payment Gateways Jun-2021 2021
## 9    39.9               0                 39.9 Payment Gateways Jun-2021 2021
## 10    0.0               0                  0.0 Payment Gateways Sep-2021 2021
##                    status customer_since         SSN    phone_num       city
## 1  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 2  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 3  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 4  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 5  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 6  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 7  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 8  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 9  Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
## 10 Refund or Cancellation      3/31/2017 023-92-2382 217-405-0048 Brownstown
##    state   zip  region                    E-mail
## 1     IL 62418 Midwest elizbeth.raminez@gmail.com
```

```
## 2     IL 62418 Midwest elizbeth.raminez@gmail.com
## 3     IL 62418 Midwest elizbeth.raminez@gmail.com
## 4     IL 62418 Midwest elizbeth.raminez@gmail.com
## 5     IL 62418 Midwest elizbeth.raminez@gmail.com
## 6     IL 62418 Midwest elizbeth.raminez@gmail.com
## 7     IL 62418 Midwest elizbeth.raminez@gmail.com
## 8     IL 62418 Midwest elizbeth.raminez@gmail.com
## 9     IL 62418 Midwest elizbeth.raminez@gmail.com
## 10    IL 62418 Midwest elizbeth.raminez@gmail.com
```

```r
#Changing the customer_since column to date datatype and changing the format
Sales_2021$customer_since <- as.Date(Sales_2021$customer_since, format = "%m/%d/%Y")
Sales_2021$customer_since <- format(Sales_2021$customer_since, "%Y")

# Function to mask data, showing only the last 4 digits
mask_data <- function(x) {
  if (is.character(x)) {
    masked <- sub(".+(....)$", "XXX-XX-\\1", x)
    return(masked)
  }
  return(x)
}

#Masking the SSN data and showing the last 4 digits of the data
Sales_2021$SSN <- sapply(Sales_2021$SSN, mask_data)

#Changing the data in month column from month-YYYY to Month formate
Sales_2021$month <- sub("([A-Za-z]+)-\\d+", "\\1", Sales_2021$month)

head(Sales_2021, 10)
```

```
##    user_name customer_id first_name last_name sex age  order_id      category
## 1  eeraminez       56449   Elizbeth   Raminez   F  71 100441393    Appliances
## 2  eeraminez       56449   Elizbeth   Raminez   F  71 100533723 Men's Fashion
## 3  eeraminez       56449   Elizbeth   Raminez   F  71 100533724 Men's Fashion
## 4  eeraminez       56449   Elizbeth   Raminez   F  71 100533726 Men's Fashion
## 5  eeraminez       56449   Elizbeth   Raminez   F  71 100533727 Men's Fashion
## 6  eeraminez       56449   Elizbeth   Raminez   F  71 100533729 Men's Fashion
## 7  eeraminez       56449   Elizbeth   Raminez   F  71 100533733 Men's Fashion
## 8  eeraminez       56449   Elizbeth   Raminez   F  71 100533735 Men's Fashion
## 9  eeraminez       56449   Elizbeth   Raminez   F  71 100533739 Men's Fashion
## 10 eeraminez       56449   Elizbeth   Raminez   F  71 100560796 Men's Fashion
##     price discount_amount price_after_discount     payment_type month year
## 1  2625.0               0                2625.0 Payment Gateways   Feb 2021
## 2   299.9               0                 299.9 Payment Gateways   Jun 2021
## 3   299.9               0                 299.9 Payment Gateways   Jun 2021
## 4   299.9               0                 299.9 Payment Gateways   Jun 2021
## 5   299.9               0                 299.9 Payment Gateways   Jun 2021
## 6  1600.0               0                1600.0 Payment Gateways   Jun 2021
## 7    59.9               0                  59.9 Payment Gateways   Jun 2021
## 8   274.9               0                 274.9 Payment Gateways   Jun 2021
## 9    39.9               0                  39.9 Payment Gateways   Jun 2021
## 10    0.0               0                   0.0 Payment Gateways   Sep 2021
##                    status customer_since         SSN    phone_num       city
```

```
## 1  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 2  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 3  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 4  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 5  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 6  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 7  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 8  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 9  Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
## 10 Refund or Cancellation           2017 XXX-XX-2382 217-405-0048 Brownstown
##    state   zip  region                    E-mail
## 1     IL 62418 Midwest elizbeth.raminez@gmail.com
## 2     IL 62418 Midwest elizbeth.raminez@gmail.com
## 3     IL 62418 Midwest elizbeth.raminez@gmail.com
## 4     IL 62418 Midwest elizbeth.raminez@gmail.com
## 5     IL 62418 Midwest elizbeth.raminez@gmail.com
## 6     IL 62418 Midwest elizbeth.raminez@gmail.com
## 7     IL 62418 Midwest elizbeth.raminez@gmail.com
## 8     IL 62418 Midwest elizbeth.raminez@gmail.com
## 9     IL 62418 Midwest elizbeth.raminez@gmail.com
## 10    IL 62418 Midwest elizbeth.raminez@gmail.com
```

```r
# Find duplicates based on all columns
duplicates <- Sales_2021[duplicated(Sales_2021), ]
number_of_duplicatesrecords <- nrow(duplicates)
print(number_of_duplicatesrecords)
```

```
## [1] 7192
```

```r
# Remove duplicates
Sales_2021 <- Sales_2021[!duplicated(Sales_2021), ]

#Checking No of rows after removing duplicate data
no_of_rows <- nrow(Sales_2021)
print(no_of_rows) #170021 rows
```

```
## [1] 170021
```

```r
# Check for null values in each column
null_values <- sapply(Sales_2021, function(x) sum(is.na(x)))

# Print the result
print(null_values) # There is no missing data in the Data set
```

```
##            user_name          customer_id           first_name
##                    0                    0                    0
##            last_name                  sex                  age
##                    0                    0                    0
##             order_id             category                price
##                    0                    0                    0
##      discount_amount price_after_discount         payment_type
##                    0                    0                    0
```

```
##                month                 year               status
##                    0                    0                    0
##        customer_since                  SSN            phone_num
##                    0                    0                    0
##                 city                state                  zip
##                    0                    0                    0
##               region               E-mail
##                    0                    0
```

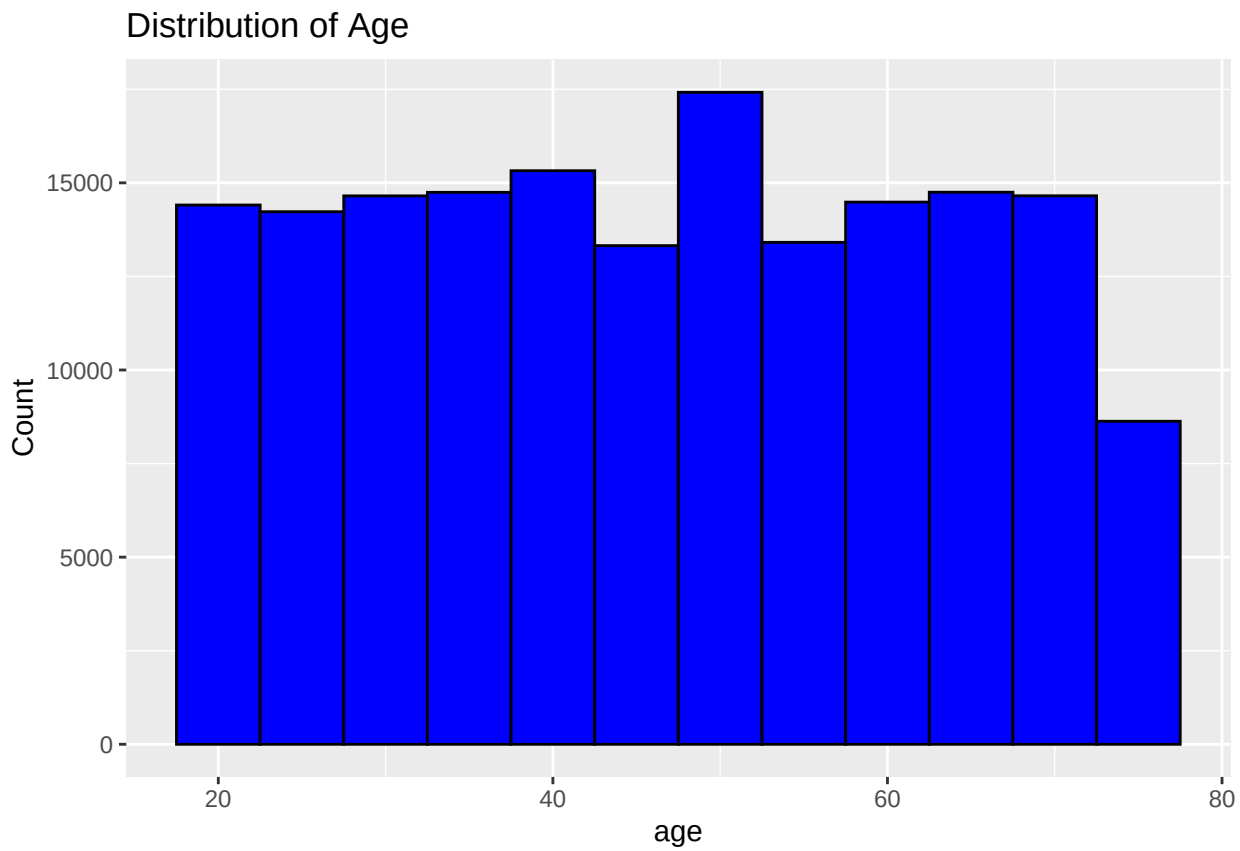
```
head(Sales_2021, 20)
```

```
##    user_name customer_id first_name last_name sex age  order_id        category
## 1  eeraminez       56449   Elizbeth   Raminez   F  71 100441393      Appliances
## 2  eeraminez       56449   Elizbeth   Raminez   F  71 100533723   Men's Fashion
## 3  eeraminez       56449   Elizbeth   Raminez   F  71 100533724   Men's Fashion
## 4  eeraminez       56449   Elizbeth   Raminez   F  71 100533726   Men's Fashion
## 5  eeraminez       56449   Elizbeth   Raminez   F  71 100533727   Men's Fashion
## 6  eeraminez       56449   Elizbeth   Raminez   F  71 100533729   Men's Fashion
## 7  eeraminez       56449   Elizbeth   Raminez   F  71 100533733   Men's Fashion
## 8  eeraminez       56449   Elizbeth   Raminez   F  71 100533735   Men's Fashion
## 9  eeraminez       56449   Elizbeth   Raminez   F  71 100533739   Men's Fashion
## 10 eeraminez       56449   Elizbeth   Raminez   F  71 100560796   Men's Fashion
## 11 eeraminez       56449   Elizbeth   Raminez   F  71 100560807   Men's Fashion
## 12 eeraminez       56449   Elizbeth   Raminez   F  71 100560817   Men's Fashion
## 13 eeraminez       56449   Elizbeth   Raminez   F  71 100560829   Men's Fashion
## 14 eeraminez       56449   Elizbeth   Raminez   F  71 100560841   Men's Fashion
## 15    lzmayo       31655     Lauran      Mayo   F  38 100445653 Health & Sports
## 16    lzmayo       31655     Lauran      Mayo   F  38 100445653   Men's Fashion
## 17    lzmayo       31655     Lauran      Mayo   F  38 100468670 Health & Sports
## 18    lzmayo       31655     Lauran      Mayo   F  38 100468676 Health & Sports
## 19    lzmayo       31655     Lauran      Mayo   F  38 100468677 Health & Sports
## 20    lzmayo       31655     Lauran      Mayo   F  38 100471838 Health & Sports
##     price discount_amount price_after_discount     payment_type month year
## 1  2625.0               0               2625.0 Payment Gateways   Feb 2021
## 2   299.9               0                299.9 Payment Gateways   Jun 2021
## 3   299.9               0                299.9 Payment Gateways   Jun 2021
## 4   299.9               0                299.9 Payment Gateways   Jun 2021
## 5   299.9               0                299.9 Payment Gateways   Jun 2021
## 6  1600.0               0               1600.0 Payment Gateways   Jun 2021
## 7    59.9               0                 59.9 Payment Gateways   Jun 2021
## 8   274.9               0                274.9 Payment Gateways   Jun 2021
## 9    39.9               0                 39.9 Payment Gateways   Jun 2021
## 10    0.0               0                  0.0 Payment Gateways   Sep 2021
## 11    0.0               0                  0.0 Payment Gateways   Sep 2021
## 12    0.0               0                  0.0 Payment Gateways   Sep 2021
## 13    0.0               0                  0.0 Payment Gateways   Sep 2021
## 14  119.8               0                119.8 Payment Gateways   Sep 2021
## 15   78.8               0                 78.8  Digital Wallets   Feb 2021
## 16   22.5               0                 22.5  Digital Wallets   Feb 2021
## 17  130.0               0                130.0 Payment Gateways   Apr 2021
## 18  130.0               0                130.0 Payment Gateways   Apr 2021
## 19  130.0               0                130.0 Payment Gateways   Apr 2021
## 20   65.0               0                 65.0  Digital Wallets   Apr 2021
##                    status customer_since         SSN    phone_num       city
```

```
## 1  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 2  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 3  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 4  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 5  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 6  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 7  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 8  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 9  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 10 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 11 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 12 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 13 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 14 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 15 Completed or Fulfilled          2003 XXX-XX-5318 480-518-4474    Phoenix
## 16 Completed or Fulfilled          2003 XXX-XX-5318 480-518-4474    Phoenix
## 17 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
## 18 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
## 19 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
## 20 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
##    state   zip  region                    E-mail
## 1     IL 62418 Midwest elizbeth.raminez@gmail.com
## 2     IL 62418 Midwest elizbeth.raminez@gmail.com
## 3     IL 62418 Midwest elizbeth.raminez@gmail.com
## 4     IL 62418 Midwest elizbeth.raminez@gmail.com
## 5     IL 62418 Midwest elizbeth.raminez@gmail.com
## 6     IL 62418 Midwest elizbeth.raminez@gmail.com
## 7     IL 62418 Midwest elizbeth.raminez@gmail.com
## 8     IL 62418 Midwest elizbeth.raminez@gmail.com
## 9     IL 62418 Midwest elizbeth.raminez@gmail.com
## 10    IL 62418 Midwest elizbeth.raminez@gmail.com
## 11    IL 62418 Midwest elizbeth.raminez@gmail.com
## 12    IL 62418 Midwest elizbeth.raminez@gmail.com
## 13    IL 62418 Midwest elizbeth.raminez@gmail.com
## 14    IL 62418 Midwest elizbeth.raminez@gmail.com
## 15    AZ 85015    West      lauran.mayo@gmail.com
## 16    AZ 85015    West      lauran.mayo@gmail.com
## 17    AZ 85015    West      lauran.mayo@gmail.com
## 18    AZ 85015    West      lauran.mayo@gmail.com
## 19    AZ 85015    West      lauran.mayo@gmail.com
## 20    AZ 85015    West      lauran.mayo@gmail.com
```

Data Analysis

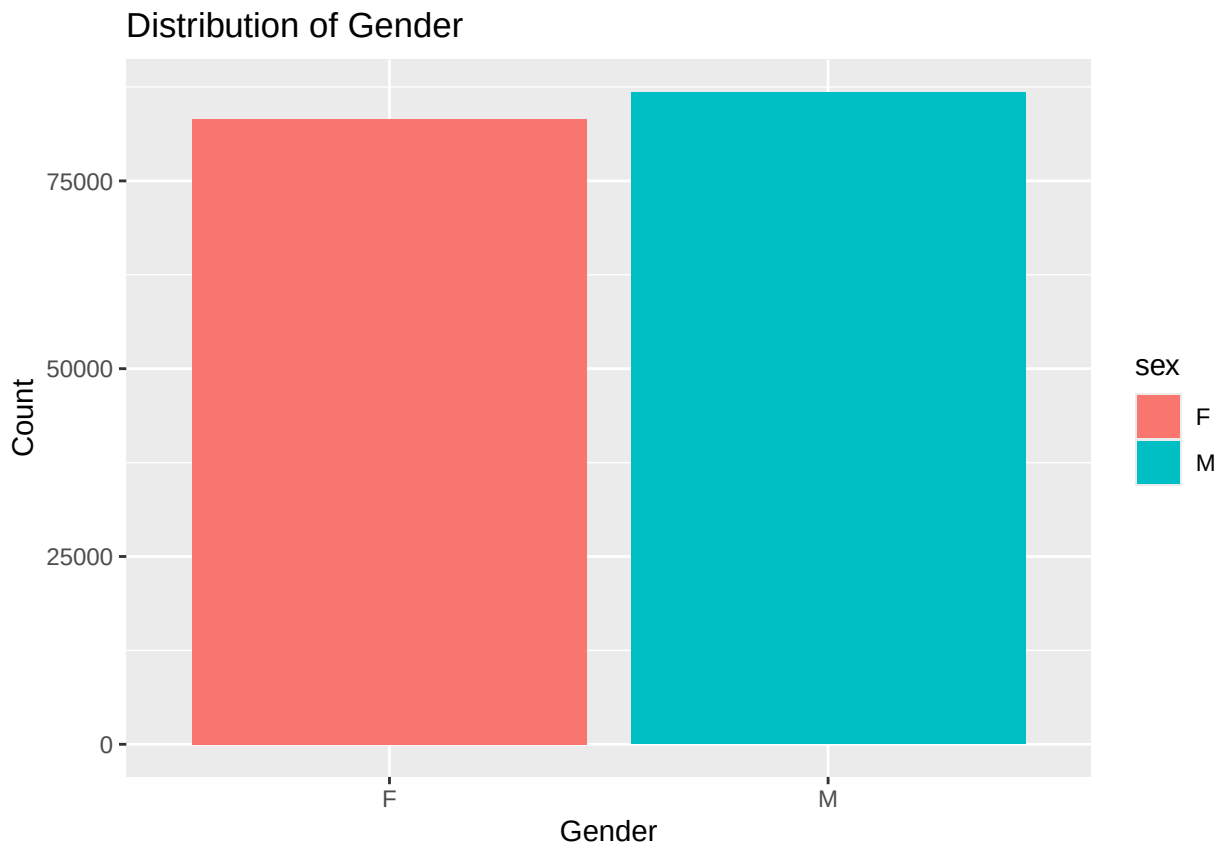
```r
#Demographics
# Age distribution
ggplot(Sales_2021, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Age", x = "age", y = "Count")
```



"Customers are equally distributed across various age groups, with a notable concentration of individuals between 45-55 years."

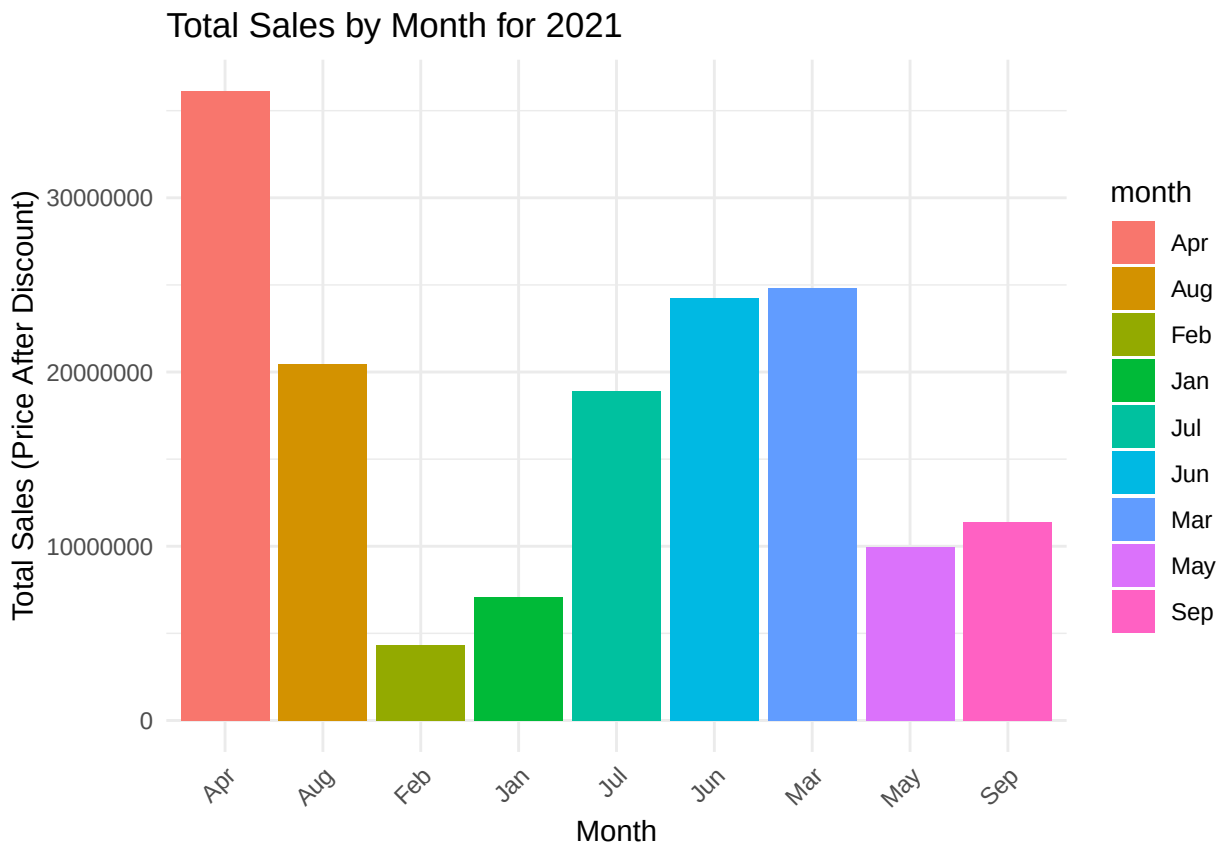
To increase the score, implement targeted marketing campaigns, gather and act on customer feedback, offer exclusive promotions, enhance customer experience, create engaging activities, introduce loyalty programs, and strengthen digital marketing efforts specifically for the 45-55 age group.

```
# Gender distribution
ggplot(Sales_2021, aes(x = sex, fill = sex)) +
  geom_bar() +
  labs(title = "Distribution of Gender", x = "Gender", y = "Count")
```



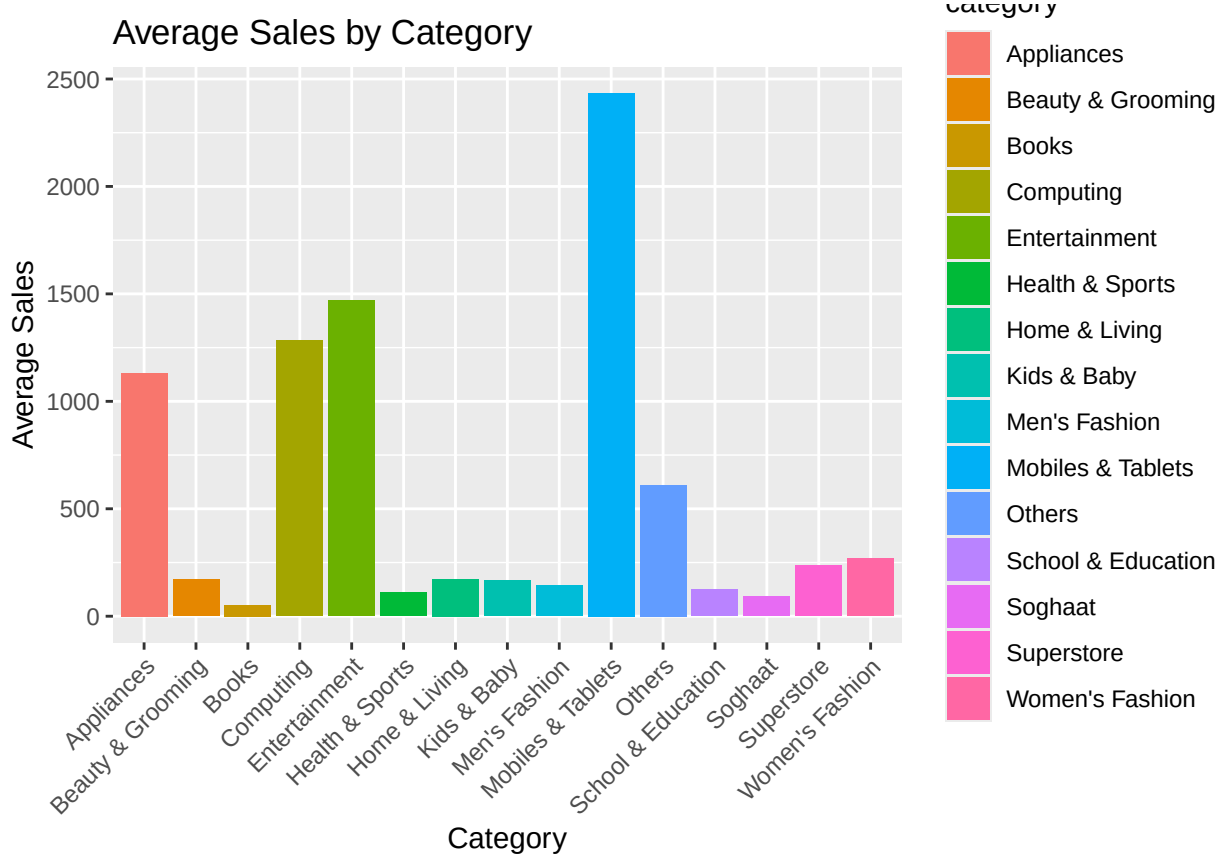
Male customers slightly outnumber female customers its good to enhancing product offerings to better appeal to both genders.

```
#Monthly Sales Analysis of 2021 using bar chart

monthly_sales <- Sales_2021 %>%
  filter(year == 2021) %>%
  group_by(month) %>%
  summarize(total_sales = sum(price_after_discount))

# Plotting
ggplot(monthly_sales, aes(x = month, y = total_sales, fill = month)) +
  geom_col() +
  labs(title = "Total Sales by Month for 2021", x = "Month", y = "Total Sales (Price After Discount)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x labels for clarity
```



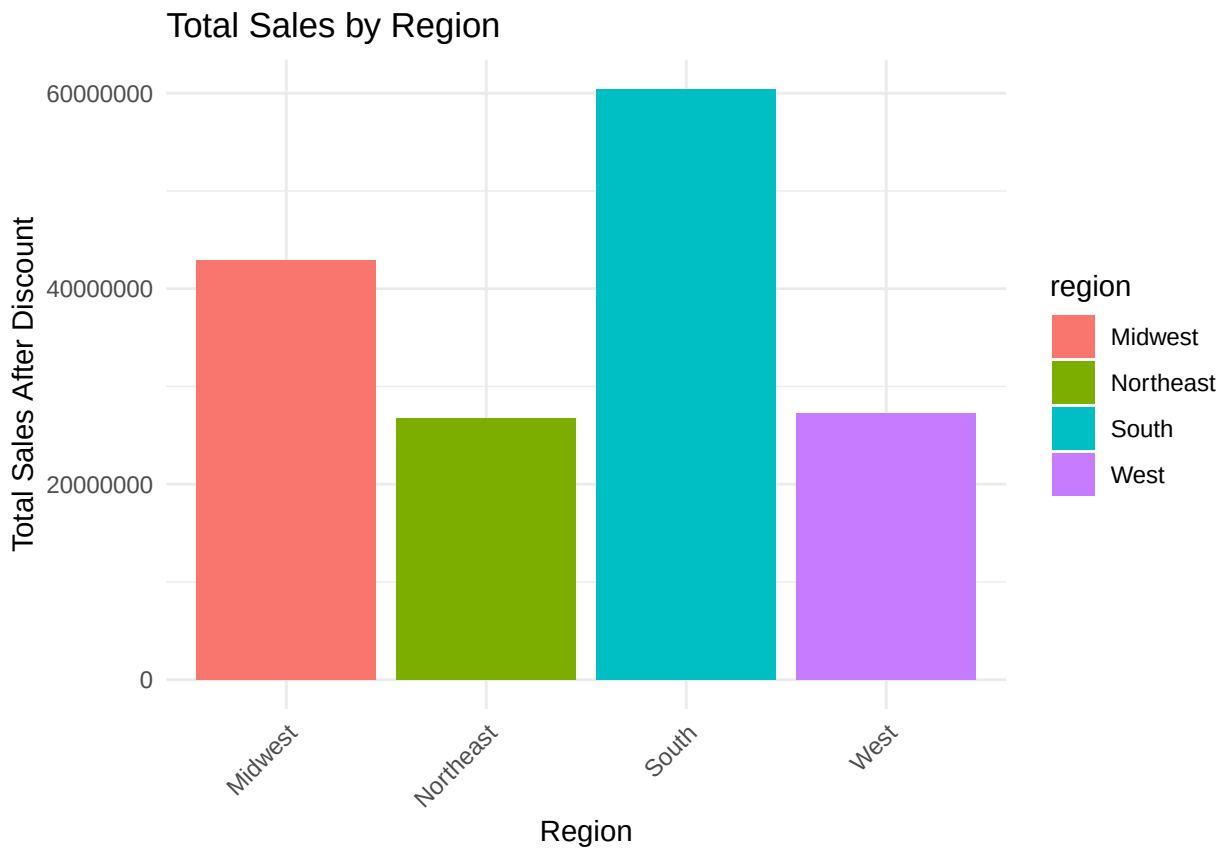
Focusing on replicating successful strategies from April, March, and June by offering special promotions, launching new products, enhancing marketing campaigns, and engaging customers through targeted advertising.

```
# Category-wise sales
ggplot(Sales_2021, aes(x = category, y = price_after_discount, fill = category)) +
  geom_bar(stat = "summary", fun = mean) +
  labs(title = "Average Sales by Category", x = "Category", y = "Average Sales") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



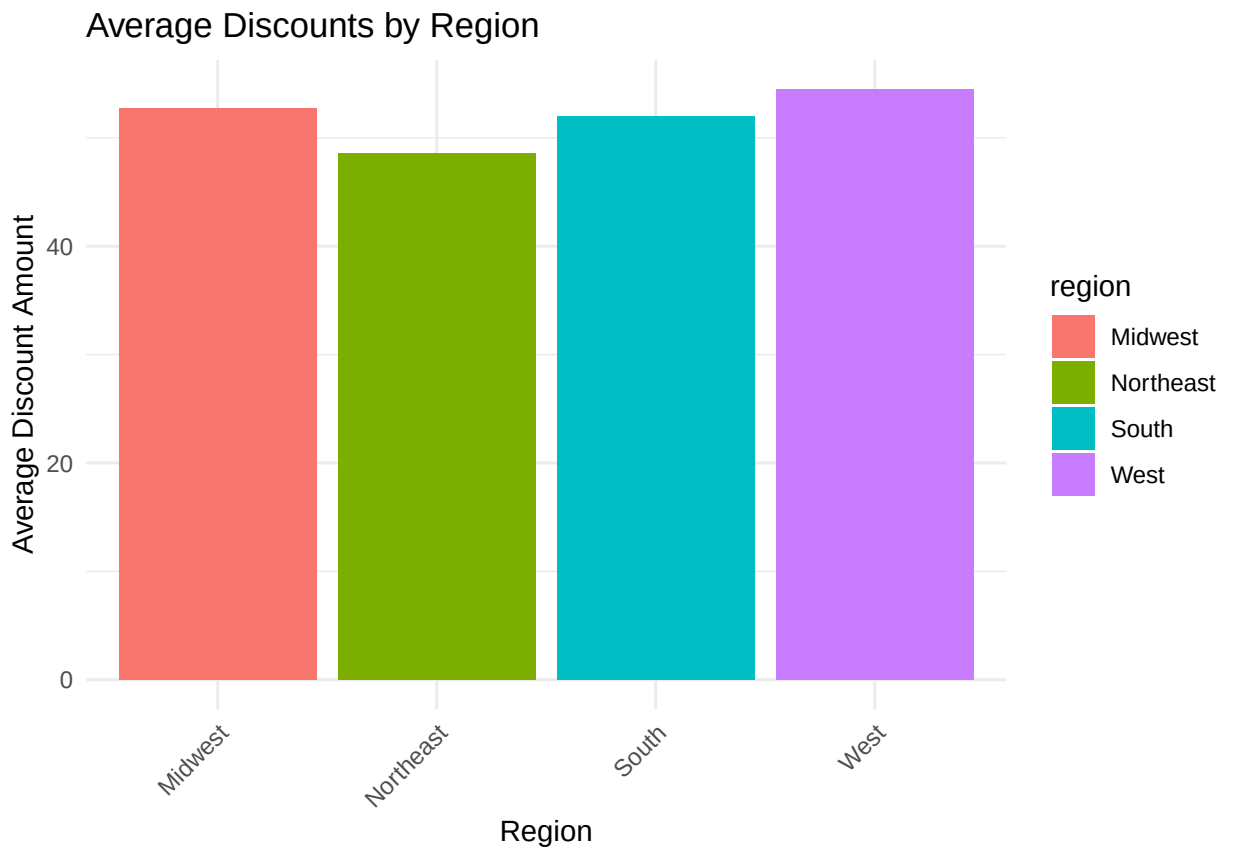
More than 50% of the sales are coming from mobiles and tablets, entertainment, computing, and appliances.By this analysis stores can also track expected sale and maintain the inventory.

```
# Total sales by region
ggplot(Sales_2021, aes(x = region, y = price_after_discount, fill = region)) +
  geom_bar(stat = "summary", fun = sum) +
  labs(title = "Total Sales by Region", x = "Region", y = "Total Sales After Discount") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



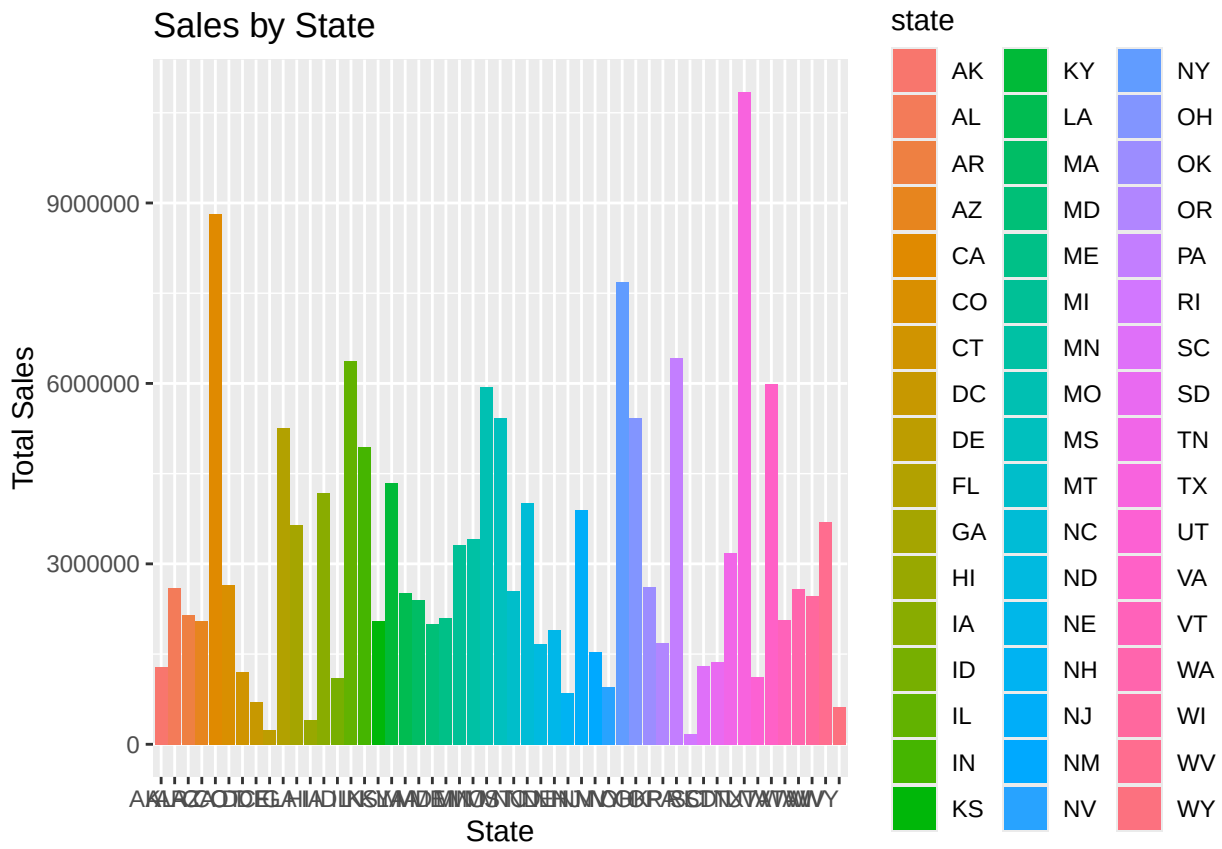
South region has the highest sales.

```
# Summarize data to find the Average discount by region
highest_discounts <- Sales_2021 %>%
  group_by(region) %>%
  summarize(highest_discount = mean(discount_amount), .groups = 'drop')

# Optional: Create a bar plot of the highest discounts by region
ggplot(highest_discounts, aes(x = region, y = highest_discount, fill = region)) +
  geom_col() +
  labs(title = "Average Discounts by Region", x = "Region", y = "Average Discount Amount") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



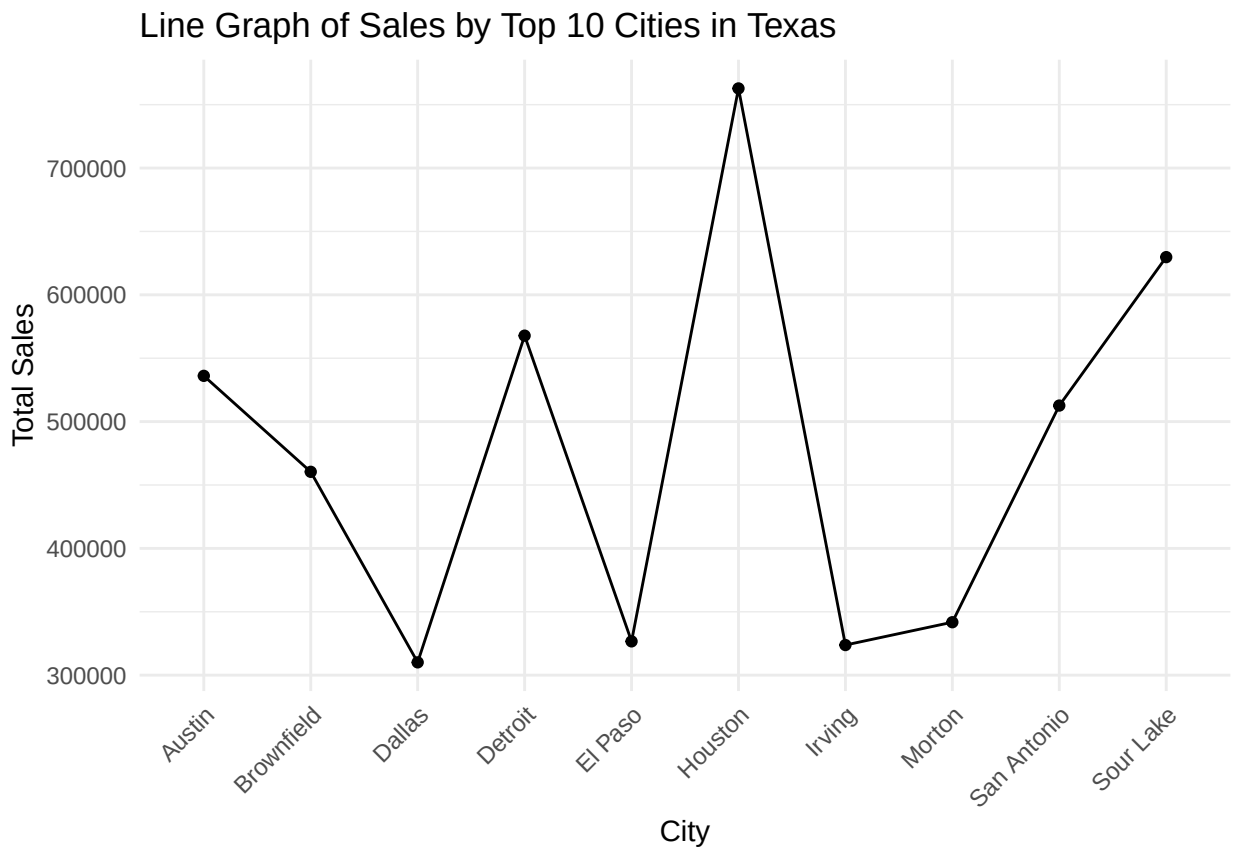
Despite offering equal discounts across all regions, sales in the Northeast and West are lagging, to increase sales in these regions, considering implementing region-specific promotions, enhancing local marketing efforts, understanding and addressing unique regional preferences.

```
#Geographic Distribution of Sales by state
ggplot(Sales_2021, aes(x = state, y = price_after_discount, fill = state)) +
  geom_bar(stat = "identity") +
  labs(title = "Sales by State", x = "State", y = "Total Sales") +
  theme(axis.text.x = element_text(angle = 0, hjust = 1))
```



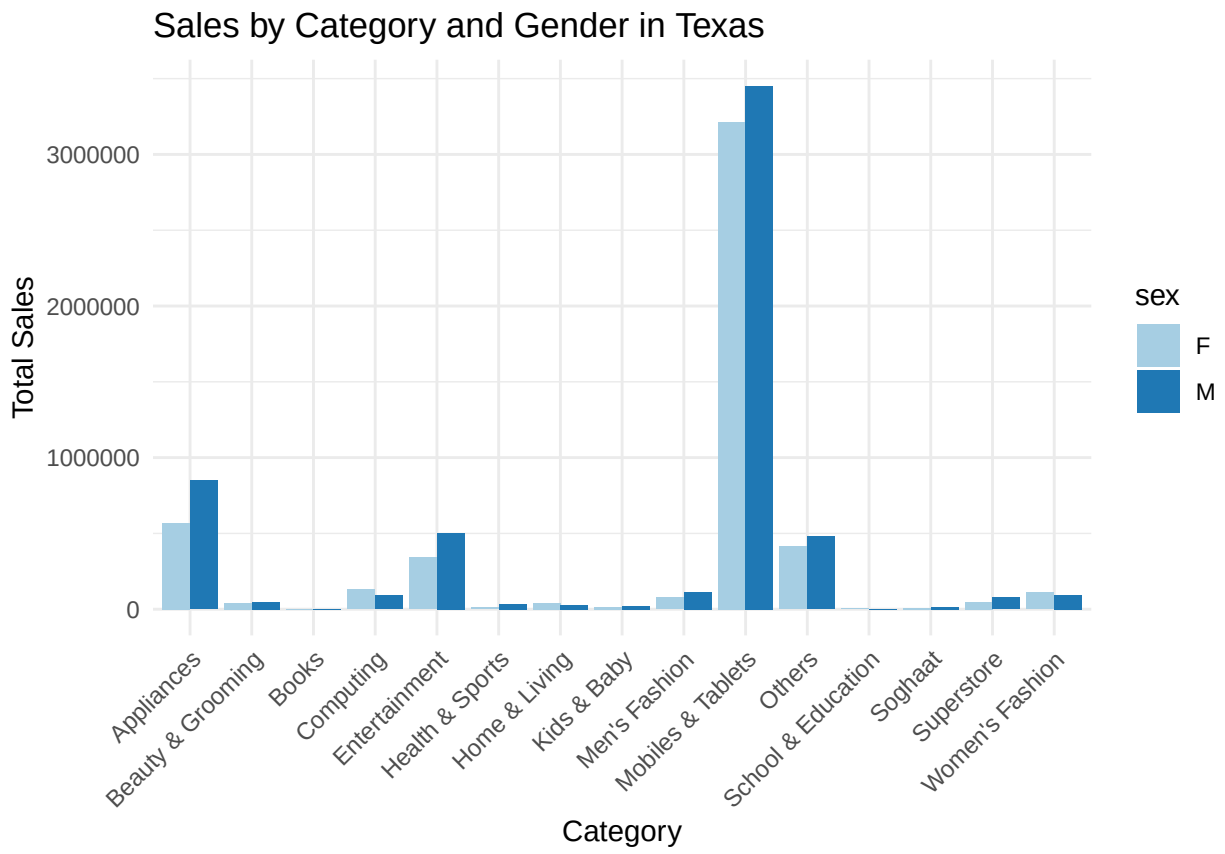
TX, CA, and NY cities have the highest sales overall.

```
#We can see the texas state has the highest sales, lets check which top 10 cities in texas has the high
#Filtering top 10 cities data in texas
texas_sales_10_cities <- Sales_2021 %>%
  filter(state == "TX") %>%
  group_by(city) %>%
  summarize(total_sales = sum(price_after_discount)) %>%
  arrange(desc(total_sales)) %>%
  slice_head(n = 10)

# Create a line graph of sales for the top 10 cities in Texas
ggplot(texas_sales_10_cities, aes(x = city, y = total_sales, group = 1)) +
  geom_line() +
  geom_point() +  # Add points for better visibility of each city
  labs(title = "Line Graph of Sales by Top 10 Cities in Texas", x = "City", y = "Total Sales") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Houston, Detroit, and San Antonio are the top three highest-selling cities in Texas.

```r
#Filtering records of texas city and calculating the sum of sales and find out the highest sold categor
texas_sales <- Sales_2021 %>%
  filter(state == "TX") %>%
  group_by(sex, category) %>%
  summarize(total_sales = sum(price_after_discount), .groups = 'drop')

# Create a bar plot of sales by category and gender in Texas
ggplot(texas_sales, aes(x = category, y = total_sales, fill = sex)) +
  geom_col(position = "dodge") +
  labs(title = "Sales by Category and Gender in Texas", x = "Category", y = "Total Sales") +
  scale_fill_brewer(palette = "Paired") +  # Color palette for distinction
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



As expected, Mobiles & Tablets are the highest sold category in Texas, with male customers slightly outnumbering females; however, it's beneficial to concentrate on both demographics and expand focus to different categories such as School & Education products, Books, and Soghaat to diversify sales.

```
# Add a categorization column based on customer_since
sales_data <- Sales_2021 %>%
  mutate(customer_group = ifelse(customer_since < 2000, "Before 2000", "2000 and After"))

# Calculate average price_after_discount for each group
average_prices <- sales_data %>%
  group_by(customer_group) %>%
  summarize(average_price_after_discount = mean(price_after_discount), .groups = 'drop')

# Output the results
print(average_prices)
```

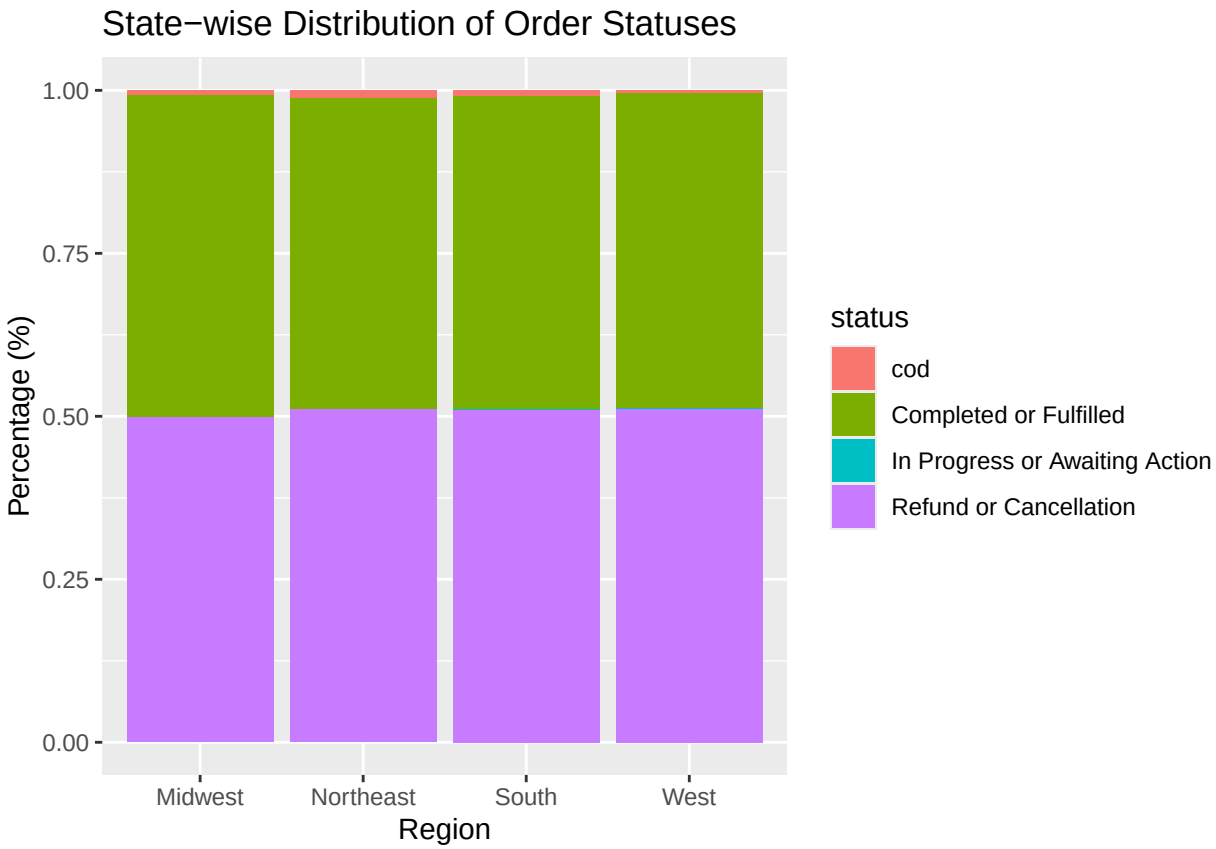
```
## # A tibble: 2 x 2
##   customer_group average_price_after_discount
##   <chr>                                <dbl>
## 1 2000 and After                        894.
## 2 Before 2000                          1059.
```

```
# Create a bar chart to visualize average prices
ggplot(average_prices, aes(x = customer_group, y = average_price_after_discount, fill = customer_group))
  geom_col() +  # geom_col is used for bar charts where bars start at 0
  labs(title = "Average Price After Discount by Customer Group",
```

```
       x = "Customer Group",
       y = "Average Price After Discount") +
  scale_fill_brewer(palette = "Pastel1") +  # Aesthetic color setting
  theme_minimal() +  # Minimal theme for a clean look
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5))  # Adjust text alignment if needed
```

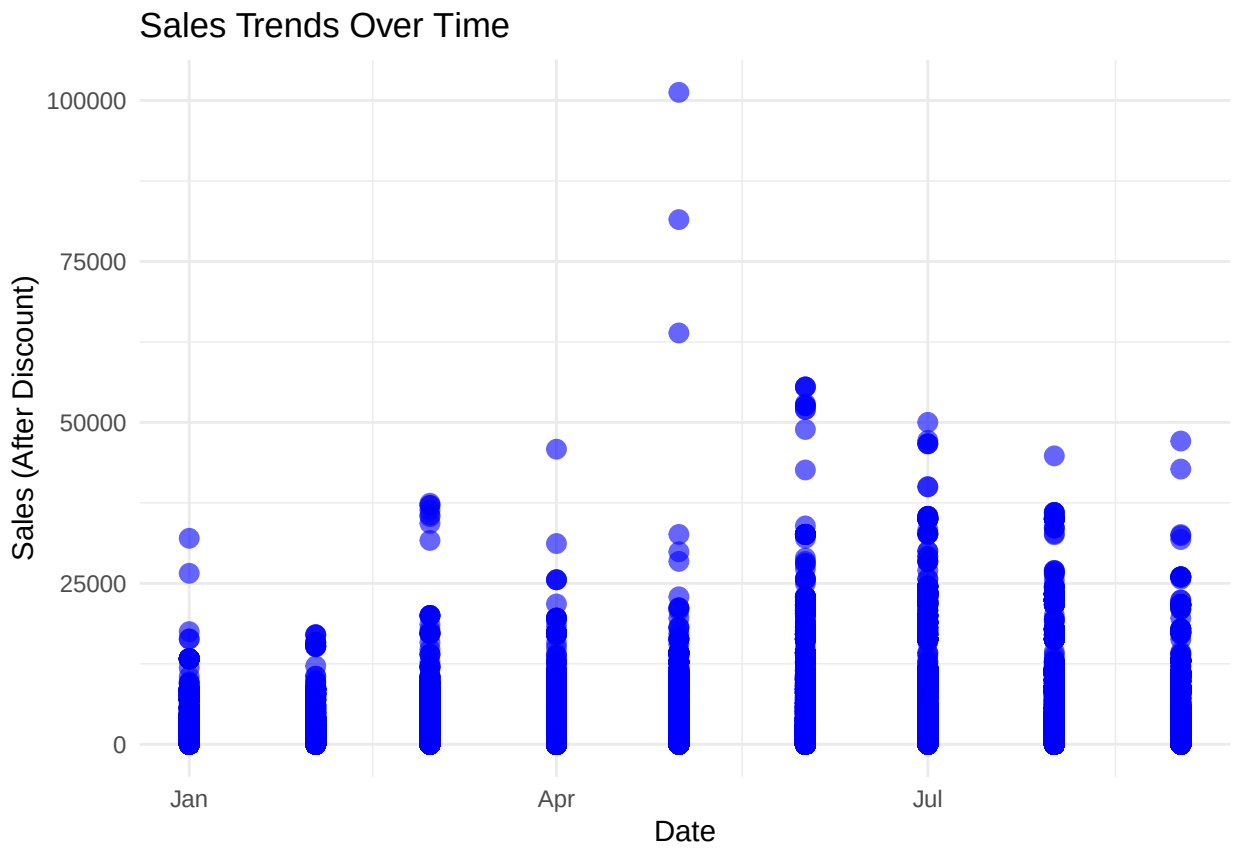
Old customers from before 2000 are generating more sales than those who joined later; to boost sales, implementing Personalized Anniversary Gifts, Personalized Communication, Recognition Programs and Referral Rewards for Old customers

```
# State-wise Distribution of Order Statuses
status_distribution <- Sales_2021 %>%
  group_by(region, status) %>%
  summarize(count = n(), .groups = 'drop') %>%
  mutate(percentage = count / sum(count) * 100)

# Visualization Proportional Bar Chart
ggplot(status_distribution, aes(x = region, y = percentage, fill = status)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "State-wise Distribution of Order Statuses", x = "Region", y = "Percentage (%)")
```



Nearly 50 % of the orders are refunds and cancellations, To decrease refunds and cancellations, we should focus on improving product quality, enhancing customer support, providing clear product descriptions, implementing flexible return policies, gathering customer feedback, sending pre-order confirmations, optimizing delivery experiences, improving the user experience are to be done.

```
# Monthly Sales Analysis of 2021 using point plot

Sales_2021$date <- as.Date(with(Sales_2021, paste(year, month, "01", sep = "-")), "%Y-%b-%d")

library(ggplot2)
ggplot(Sales_2021, aes(x = date, y = price_after_discount)) +
  geom_point(color = "blue", size = 3, alpha = 0.6) +
  labs(title = "Sales Trends Over Time", x = "Date", y = "Sales (After Discount)") +
  theme_minimal()
```



In May, June, and July, people tend to buy more.

Modelling:

we're building a logistic regression model to predict whether an order will be a "Refund or Cancellation" based on factors like price, discount, category, payment type, and more. After creating the model, we calculate the predicted probabilities for each order, showing how likely it is to be refunded or canceled. Finally, we plot these probabilities against the price to see how the price affects the likelihood of a refund or cancellation.

We're finding how different factors, especially the price, influence the chances of an order being refunded or canceled.

Modeling

```
# Convert 'status' to a binary factor
Sales_2021$status_binary <- ifelse(Sales_2021$status == "Refund or Cancellation", 1, 0)

# Logistic regression model to predict order status
model <- glm(status_binary ~ price + discount_amount + category + payment_type + month + year + age + r
             data = Sales_2021,
             family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# Summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = status_binary ~ price + discount_amount + category +
##     payment_type + month + year + age + region, family = binomial,
##     data = Sales_2021)
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate   Std. Error z value
## (Intercept)                     0.652805938  0.032397513  20.150
## price                           0.000157975  0.000004506  35.056
## discount_amount                -0.001757227  0.000048103 -36.531
## categoryBeauty & Grooming      -0.259497542  0.026973590  -9.620
## categoryBooks                  -0.554243791  0.104241653  -5.317
## categoryComputing               0.183685438  0.035279981   5.207
## categoryEntertainment          -0.089611659  0.026493321  -3.382
## categoryHealth & Sports        -0.508799466  0.033554783 -15.163
## categoryHome & Living          -0.224519889  0.031369593  -7.157
## categoryKids & Baby            -0.098511288  0.038531549  -2.557
## categoryMen's Fashion           0.195043827  0.023044691   8.464
## categoryMobiles & Tablets       0.182211189  0.019811957   9.197
## categoryOthers                  0.560540908  0.023946240  23.408
## categorySchool & Education      0.161180798  0.077954467   2.068
## categorySoghaat                -0.210967360  0.035521354  -5.939
## categorySuperstore             -0.696299624  0.027258897 -25.544
## categoryWomen's Fashion         0.019571136  0.024089364   0.812
## payment_typecod                -1.710883029  0.020084406 -85.185
## payment_typeDigital Wallets    -0.500390957  0.020796065 -24.062
## payment_typePayment Gateways   -0.335054484  0.021700894 -15.440
## payment_typeVoucher Systems    -0.915756194  0.025303607 -36.191
## monthAug                        0.416923266  0.026808510  15.552
## monthFeb                        0.369650649  0.025931939  14.255
## monthJan                        0.260068741  0.021998366  11.822
## monthJul                        0.363385596  0.023250321  15.629
## monthJun                        0.282280747  0.019422132  14.534
## monthMar                        0.110571368  0.018047500   6.127
## monthMay                        0.311228675  0.022826833  13.634
## monthSep                        0.032369615  0.025031225   1.293
## year                                     NA           NA      NA
## age                            -0.000226015  0.000315345  -0.717
## regionNortheast                -0.031400630  0.015979337  -1.965
## regionSouth                    -0.028753613  0.013242226  -2.171
## regionWest                     -0.043601187  0.015852953  -2.750
##                                            Pr(>|z|)
## (Intercept)                   < 0.0000000000000002 ***
## price                         < 0.0000000000000002 ***
## discount_amount               < 0.0000000000000002 ***
## categoryBeauty & Grooming     < 0.0000000000000002 ***
## categoryBooks                    0.000000105542549 ***
## categoryComputing                0.000000192429267 ***
## categoryEntertainment                     0.000718 ***
## categoryHealth & Sports       < 0.0000000000000002 ***
## categoryHome & Living            0.000000000000823 ***
## categoryKids & Baby                       0.010569 *
## categoryMen's Fashion         < 0.0000000000000002 ***
```

```
## categoryMobiles & Tablets    < 0.0000000000000002 ***
## categoryOthers               < 0.0000000000000002 ***
## categorySchool & Education              0.038675 *
## categorySoghaat                0.000000002864688 ***
## categorySuperstore           < 0.0000000000000002 ***
## categoryWomen's Fashion                 0.416540
## payment_typecod              < 0.0000000000000002 ***
## payment_typeDigital Wallets  < 0.0000000000000002 ***
## payment_typePayment Gateways < 0.0000000000000002 ***
## payment_typeVoucher Systems  < 0.0000000000000002 ***
## monthAug                     < 0.0000000000000002 ***
## monthFeb                     < 0.0000000000000002 ***
## monthJan                     < 0.0000000000000002 ***
## monthJul                     < 0.0000000000000002 ***
## monthJun                     < 0.0000000000000002 ***
## monthMar                       0.000000000897282 ***
## monthMay                     < 0.0000000000000002 ***
## monthSep                                0.195952
## year                                         NA
## age                                     0.473545
## regionNortheast                         0.049405 *
## regionSouth                             0.029904 *
## regionWest                              0.005953 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 235667  on 170020  degrees of freedom
## Residual deviance: 208959  on 169988  degrees of freedom
## AIC: 209025
##
## Number of Fisher Scoring iterations: 6
```

```r
# Add predicted probabilities to the data frame
Sales_2021$predicted_probabilities <- predict(model, type = "response")
head(Sales_2021, 20)
```

```
##    user_name customer_id first_name last_name sex age  order_id        category
## 1  eeraminez       56449   Elizbeth   Raminez   F  71 100441393      Appliances
## 2  eeraminez       56449   Elizbeth   Raminez   F  71 100533723   Men's Fashion
## 3  eeraminez       56449   Elizbeth   Raminez   F  71 100533724   Men's Fashion
## 4  eeraminez       56449   Elizbeth   Raminez   F  71 100533726   Men's Fashion
## 5  eeraminez       56449   Elizbeth   Raminez   F  71 100533727   Men's Fashion
## 6  eeraminez       56449   Elizbeth   Raminez   F  71 100533729   Men's Fashion
## 7  eeraminez       56449   Elizbeth   Raminez   F  71 100533733   Men's Fashion
## 8  eeraminez       56449   Elizbeth   Raminez   F  71 100533735   Men's Fashion
## 9  eeraminez       56449   Elizbeth   Raminez   F  71 100533739   Men's Fashion
## 10 eeraminez       56449   Elizbeth   Raminez   F  71 100560796   Men's Fashion
## 11 eeraminez       56449   Elizbeth   Raminez   F  71 100560807   Men's Fashion
## 12 eeraminez       56449   Elizbeth   Raminez   F  71 100560817   Men's Fashion
## 13 eeraminez       56449   Elizbeth   Raminez   F  71 100560829   Men's Fashion
## 14 eeraminez       56449   Elizbeth   Raminez   F  71 100560841   Men's Fashion
## 15    lzmayo       31655     Lauran      Mayo   F  38 100445653 Health & Sports
```

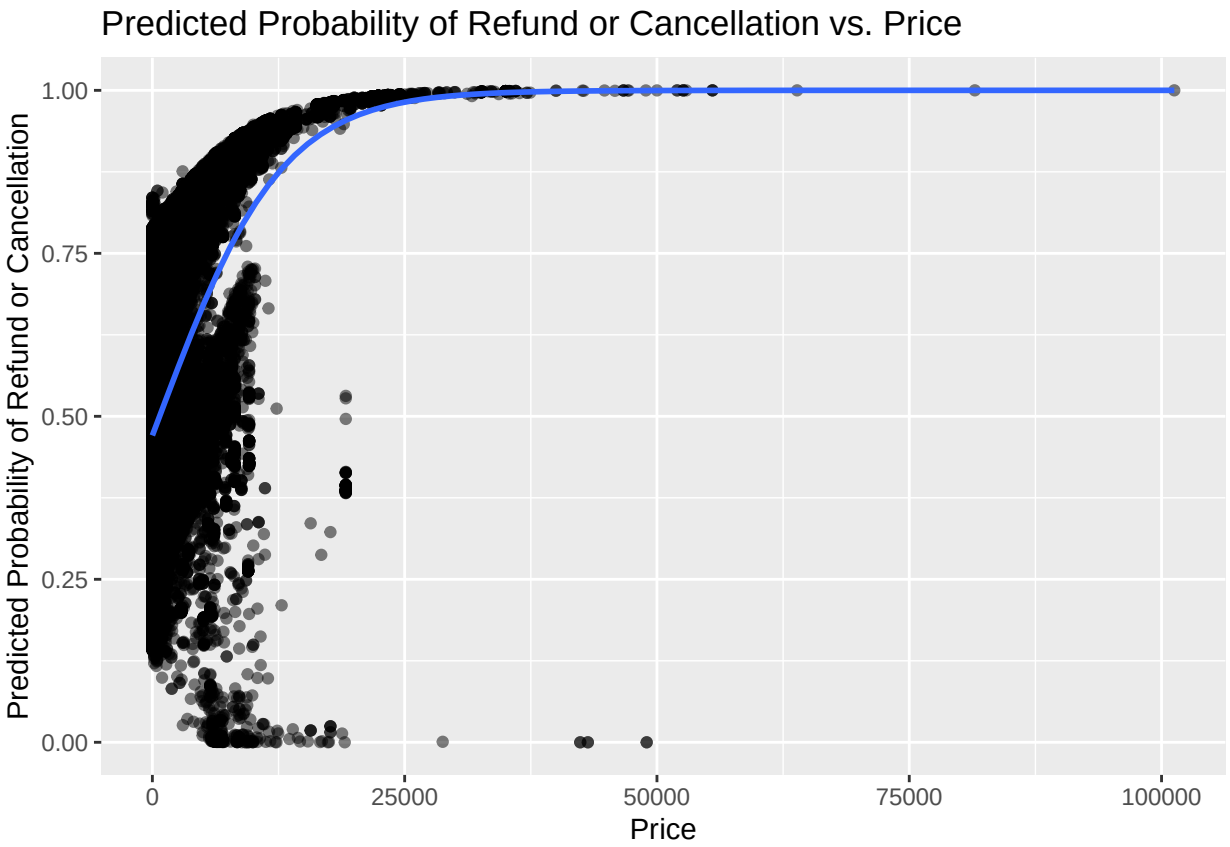
```
## 16    lzmayo       31655     Lauran      Mayo   F  38 100445653   Men's Fashion
## 17    lzmayo       31655     Lauran      Mayo   F  38 100468670 Health & Sports
## 18    lzmayo       31655     Lauran      Mayo   F  38 100468676 Health & Sports
## 19    lzmayo       31655     Lauran      Mayo   F  38 100468677 Health & Sports
## 20    lzmayo       31655     Lauran      Mayo   F  38 100471838 Health & Sports
##     price discount_amount price_after_discount     payment_type month year
## 1  2625.0               0                2625.0 Payment Gateways   Feb 2021
## 2   299.9               0                 299.9 Payment Gateways   Jun 2021
## 3   299.9               0                 299.9 Payment Gateways   Jun 2021
## 4   299.9               0                 299.9 Payment Gateways   Jun 2021
## 5   299.9               0                 299.9 Payment Gateways   Jun 2021
## 6  1600.0               0                1600.0 Payment Gateways   Jun 2021
## 7    59.9               0                  59.9 Payment Gateways   Jun 2021
## 8   274.9               0                 274.9 Payment Gateways   Jun 2021
## 9    39.9               0                  39.9 Payment Gateways   Jun 2021
## 10    0.0               0                   0.0 Payment Gateways   Sep 2021
## 11    0.0               0                   0.0 Payment Gateways   Sep 2021
## 12    0.0               0                   0.0 Payment Gateways   Sep 2021
## 13    0.0               0                   0.0 Payment Gateways   Sep 2021
## 14  119.8               0                 119.8 Payment Gateways   Sep 2021
## 15   78.8               0                  78.8  Digital Wallets   Feb 2021
## 16   22.5               0                  22.5  Digital Wallets   Feb 2021
## 17  130.0               0                 130.0 Payment Gateways   Apr 2021
## 18  130.0               0                 130.0 Payment Gateways   Apr 2021
## 19  130.0               0                 130.0 Payment Gateways   Apr 2021
## 20   65.0               0                  65.0  Digital Wallets   Apr 2021
##                    status customer_since         SSN    phone_num       city
## 1  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 2  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 3  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 4  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 5  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 6  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 7  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 8  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 9  Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 10 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 11 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 12 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 13 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 14 Refund or Cancellation          2017 XXX-XX-2382 217-405-0048 Brownstown
## 15 Completed or Fulfilled          2003 XXX-XX-5318 480-518-4474    Phoenix
## 16 Completed or Fulfilled          2003 XXX-XX-5318 480-518-4474    Phoenix
## 17 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
## 18 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
## 19 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
## 20 Refund or Cancellation          2003 XXX-XX-5318 480-518-4474    Phoenix
##    state   zip  region                    E-mail       date status_binary
## 1     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-02-01             1
## 2     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 3     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 4     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 5     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 6     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
```

```
## 7     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 8     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 9     IL 62418 Midwest elizbeth.raminez@gmail.com 2021-06-01             1
## 10    IL 62418 Midwest elizbeth.raminez@gmail.com 2021-09-01             1
## 11    IL 62418 Midwest elizbeth.raminez@gmail.com 2021-09-01             1
## 12    IL 62418 Midwest elizbeth.raminez@gmail.com 2021-09-01             1
## 13    IL 62418 Midwest elizbeth.raminez@gmail.com 2021-09-01             1
## 14    IL 62418 Midwest elizbeth.raminez@gmail.com 2021-09-01             1
## 15    AZ 85015    West      lauran.mayo@gmail.com 2021-02-01             0
## 16    AZ 85015    West      lauran.mayo@gmail.com 2021-02-01             0
## 17    AZ 85015    West      lauran.mayo@gmail.com 2021-04-01             1
## 18    AZ 85015    West      lauran.mayo@gmail.com 2021-04-01             1
## 19    AZ 85015    West      lauran.mayo@gmail.com 2021-04-01             1
## 20    AZ 85015    West      lauran.mayo@gmail.com 2021-04-01             1
##    predicted_probabilities
## 1                0.7476352
## 2                0.6955944
## 3                0.6955944
## 4                0.6955944
## 5                0.6955944
## 6                0.7372626
## 7                0.6875074
## 8                0.6947575
## 9                0.6868282
## 10               0.6292773
## 11               0.6292773
## 12               0.6292773
## 13               0.6292773
## 14               0.6336815
## 15               0.4933816
## 16               0.6611614
## 17               0.4445537
## 18               0.4445537
## 19               0.4445537
## 20               0.4017194
```

```r
# Plot the predicted probabilities against one of the predictors, e.g., price
library(ggplot2)

ggplot(Sales_2021, aes(x = price, y = predicted_probabilities)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) + #Adds the regressi
  labs(title = "Predicted Probability of Refund or Cancellation vs. Price",
       x = "Price",
       y = "Predicted Probability of Refund or Cancellation")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



Potential Insights:

Refund/Cancellation Sensitivity: There is a strong correlation between price and the likelihood of a refund or cancellation, with higher prices leading to a greater chance of these outcomes.