Kaggle Shopping Data Exploration

2026-02-01

Background About Data:

data from where, collected how? , year? , purpose?

Import Data

```
# Use here() to make it dynamic
raw_shopping <- read_csv(here("shopping", "data", "raw_data","kaggle_shopping.csv"))
```

```
Rows: 1500 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (9): Age, Gender, AnnualIncome, NumberOfPurchases, ProductCategory, Time...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
raw_shopping
```

```
# A tibble: 1,500 x 9
     Age Gender AnnualIncome NumberOfPurchases ProductCategory
   <dbl>  <dbl>        <dbl>             <dbl>           <dbl>
 1    40      1       66120.                 8               0
 2    20      1       23580.                 4               2
 3    27      1      127821.                11               2
 4    24      1      137799.                19               3
 5    31      1       99301.                19               1
```

```
 6    66      1       37758.                14               4
 7    39      1      126883.                16               3
 8    64      1       39707.                13               2
 9    43      0      102797.                20               1
10    20      1       63855.                16               0
# i 1,490 more rows
# i 4 more variables: TimeSpentOnWebsite <dbl>, LoyaltyProgram <dbl>,
#   DiscountsAvailed <dbl>, PurchaseStatus <dbl>
```

Change Categorical Variables to Factors

```
shopping <- raw_shopping |>
  mutate(across(c(Gender, ProductCategory, LoyaltyProgram, PurchaseStatus), as.factor))

summary(shopping)
```

```
      Age       Gender   AnnualIncome    NumberOfPurchases ProductCategory
 Min.   :18.0   0:743   Min.   : 20002   Min.   : 0.00     0:289
 1st Qu.:31.0   1:757   1st Qu.: 53029   1st Qu.: 5.00     1:331
 Median :45.0           Median : 83700   Median :11.00     2:273
 Mean   :44.3           Mean   : 84249   Mean   :10.42     3:286
 3rd Qu.:57.0           3rd Qu.:117168   3rd Qu.:15.00     4:321
 Max.   :70.0           Max.   :149785   Max.   :20.00
 TimeSpentOnWebsite LoyaltyProgram DiscountsAvailed PurchaseStatus
 Min.   : 1.037     0:1010         Min.   :0.000    0:852
 1st Qu.:16.157     1: 490         1st Qu.:1.000    1:648
 Median :30.940                    Median :3.000
 Mean   :30.469                    Mean   :2.555
 3rd Qu.:44.370                    3rd Qu.:4.000
 Max.   :59.991                    Max.   :5.000
```

Notes:

- Age Range: [18, 70]
- Gender: almost 50/50; average 750 customers (no gender bias)
- Income: [$20,002 , $149,785]; average $84,249 (slightly right skewed, higher earners high impact)
- NumPurchases: [0, 20]

- ProductCategory: average 300 (pretty balanced)
- TimeSpent: [1, 60); average is 30 minutes (decent engagement)
- Loyalty: ~33% enrolled
- Discounts: [1,5]; average 3 (slightly left skewed, customers will little/no usage pull down average)
- PurchaseStatus: ~43% purchase

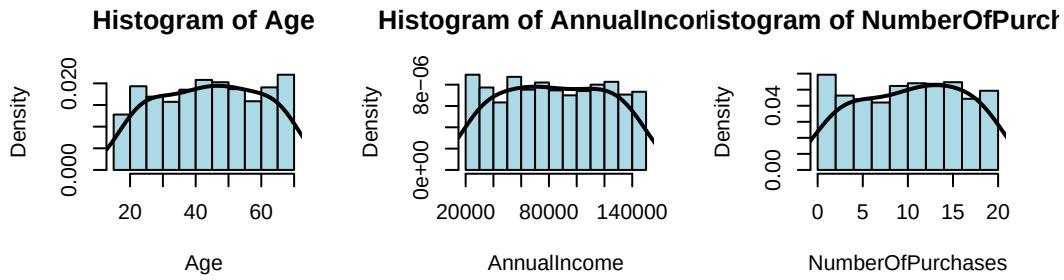
Histograms of the Continuous Variables

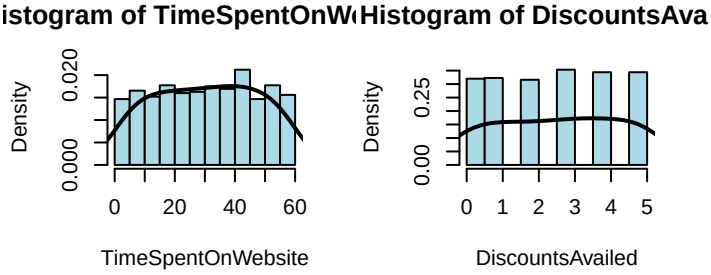
```
## Histogram of continuous variables
continuous_var <- c("Age", "AnnualIncome", "NumberOfPurchases", "TimeSpentOnWebsite", "Discou

# set up a 2 x 3 grid to display all histograms together
par(mfrow = c(2, 3))

for (i in 1:length(continuous_var)) {
  current = shopping[[continuous_var[i]]] # get rows of that variable
  hist(current,
       xlab = continuous_var[i],
       main = paste("Histogram of", continuous_var[i]),
       col = "lightblue",
       freq = FALSE)
  lines(density(current, adjust = 2), col = "black", lwd = 2) # visualize shape of distributi
  # note: adjust > 1 will smooth out the line (adjust < 1 will hug the data more)
}

# reset back to default
par(mfrow = c(1, 1))
```





Takeaway: The shopping data has a very even distribution for all variables. Since our data is very balanced, this can prevent bias and over-fitting of a particular dominant class.

Boxplots - Distribution of Variables based on Purchase Status

```
## Box plots
time <- ggplot(data = shopping, mapping = aes(x = PurchaseStatus, y = TimeSpentOnWebsite)) +
  geom_boxplot()

income <- ggplot(data = shopping, mapping = aes(x = PurchaseStatus, y = AnnualIncome)) +
  geom_boxplot()

age <- ggplot(data = shopping, mapping = aes(x = PurchaseStatus, y = Age)) +
  geom_boxplot()

discount <- ggplot(data = shopping, mapping = aes(x = PurchaseStatus, y = DiscountsAvailed))
  geom_boxplot()

# showcase all boxplots using gridExtra library
grid.arrange(time, income, age, discount, ncol = 2)
```


```
## Frequency tables
prop.table(table(shopping$LoyaltyProgram, shopping$PurchaseStatus))
```

```
           0         1
  0 0.4546667 0.2186667
  1 0.1133333 0.2133333
```

Gender on Purchase Status

```
gender_x_purchase <- shopping |>
  group_by(Gender) |>
  summarise(
    TotalPurchases = sum(NumberOfPurchases),
    AveragePurchases = mean(NumberOfPurchases),
    Purchase_Prob = sum(as.integer(PurchaseStatus))/n())
gender_x_purchase
```

```
# A tibble: 2 x 4
  Gender TotalPurchases AveragePurchases Purchase_Prob
```

```
  <fct>           <dbl>            <dbl>         <dbl>
1 0                7736             10.4          1.43
2 1                7894             10.4          1.43
```

Checking Purchase Status with Number Of Purchases?

Question: If PurchaseStatus is 0, why is there a nonempty NumberOfPurchases category for those rows?

- NumberOfPurchases is historical, which PurchaseStatus is the most current.

```
## NEXT STEP: create a new column that introduces censoring (1 if customer bought something,
## and 0 if the customer didn't buy something and if time spent is longer than cut off perio
```