Kaggle Shopping Data Exploration

Emma Nguyen

2026-02-01

Import Data

text

```
# Use here() to make it dynamic
raw_data <- 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_data
```

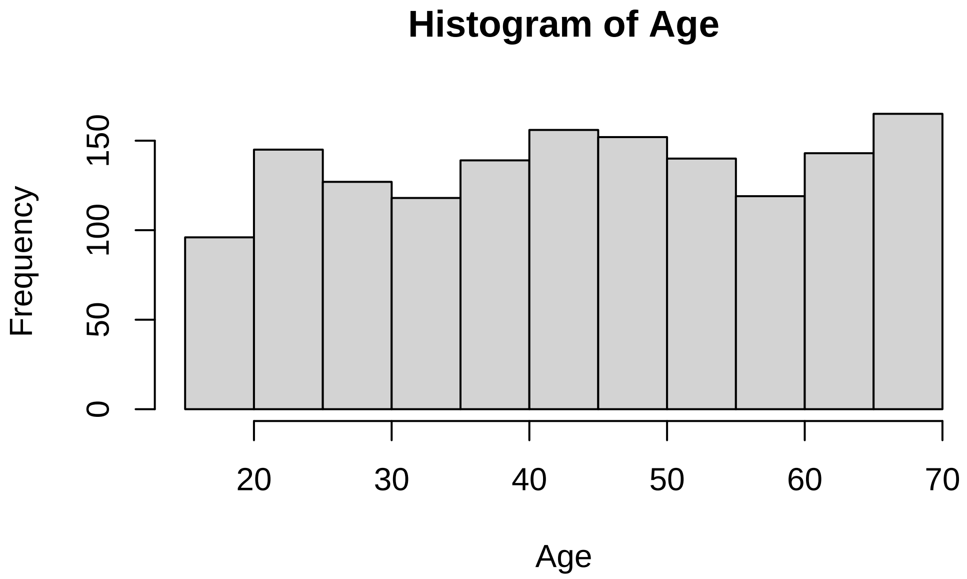
```
# A tibble: 1,500 x 9
     Age Gender AnnualIncome NumberOfPurchases ProductCategory
   <dbl>  <dbl>        <dbl>             <dbl>           <dbl>
 1    40      1       66120.                 8               0
 2    20      1       23580.                 4               2
 3    27      1      127821.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

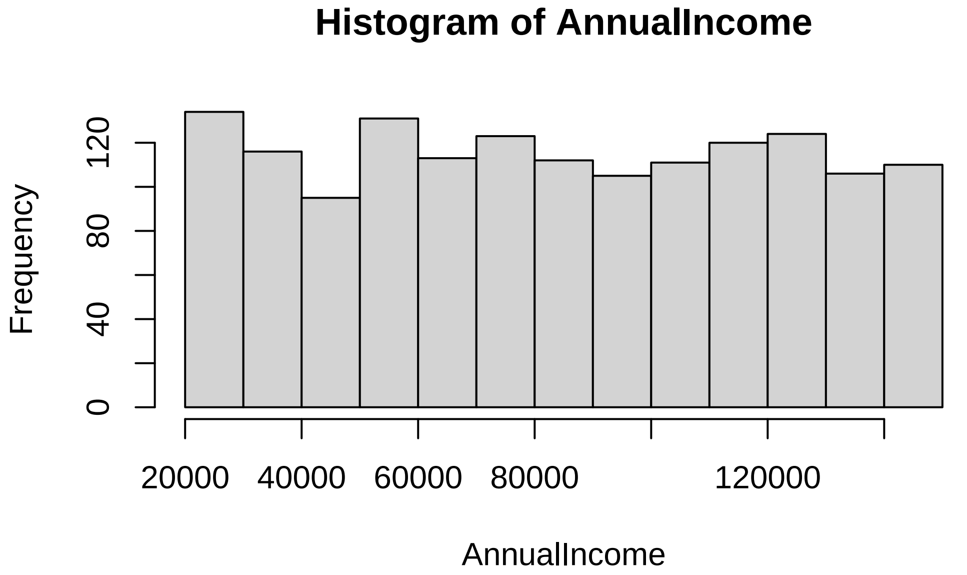
```
10    20      1       63855.                16               0
# i 1,490 more rows
# i 4 more variables: TimeSpentOnWebsite <dbl>, LoyaltyProgram <dbl>,
#   DiscountsAvailed <dbl>, PurchaseStatus <dbl>
```

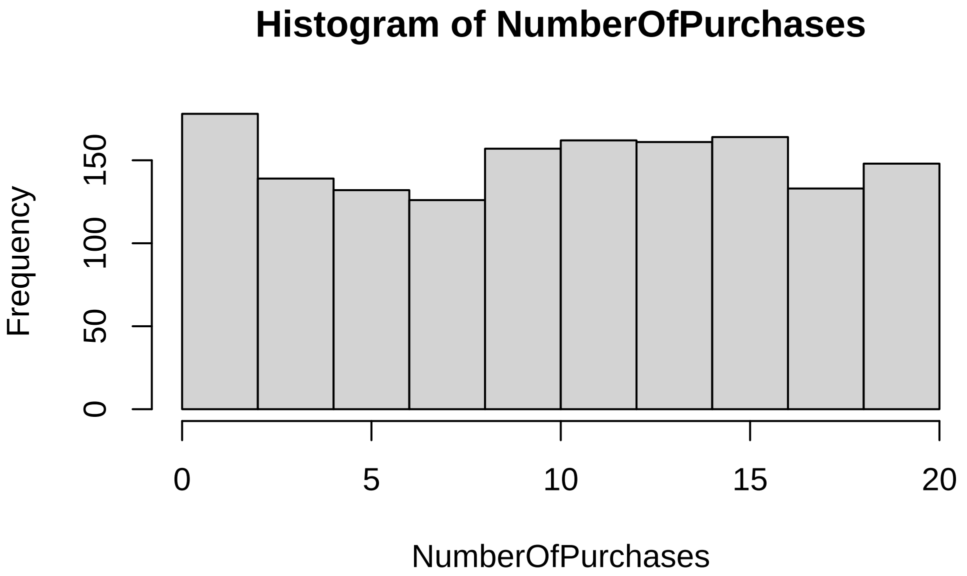
```
raw_data <- raw_data %>%
  mutate(across(c(Gender, ProductCategory, LoyaltyProgram, PurchaseStatus), as.factor))

summary(raw_data)
```

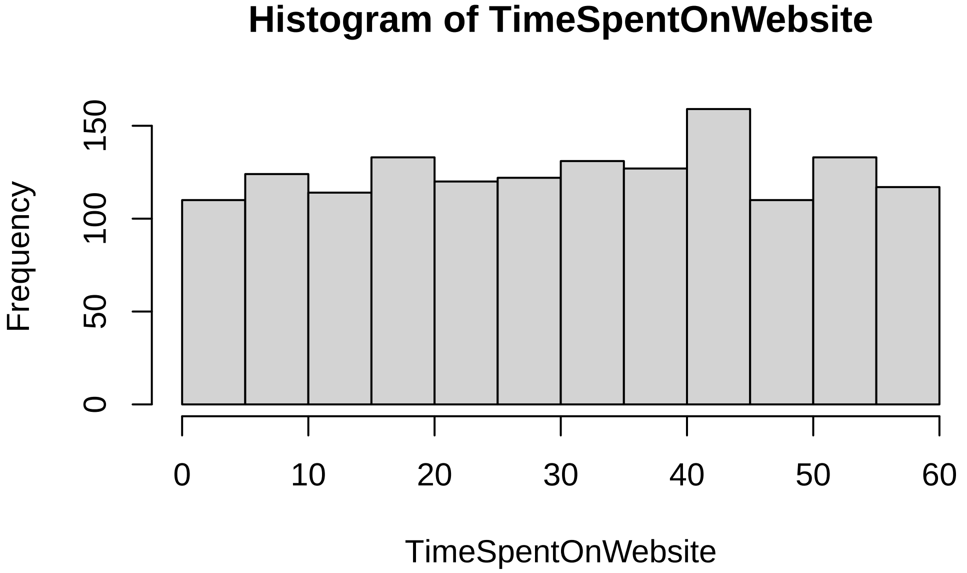
```
      Age         Gender    AnnualIncome    NumberOfPurchases ProductCategory
 Min.   :18.0   0:743   Min.   : 20002   Min.   : 0.00     0:289
 1st Qu.:31.0   1:757   1st Qu.: 53029   1st Qu.: 5.00     1:331
 Median :45.0           Median : 83700   Median :11.00     2:273
 Mean   :44.3           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
```

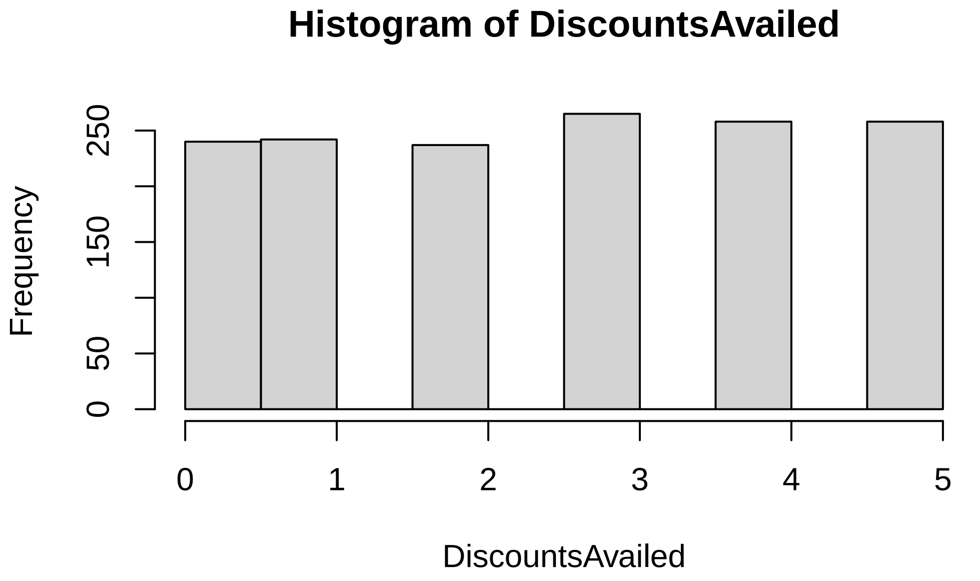
```
## Histogram of continuous variables
continuous_var <- c("Age", "AnnualIncome", "NumberOfPurchases", "TimeSpentOnWebsite", "Disco

for (i in 1:length(continuous_var)) {
  hist(raw_data[[continuous_var[i]]], xlab = continuous_var[i],
       main = paste("Histogram of", continuous_var[i]))
}
```


Histogram of Age
Frequency
0
50
100
150
20
30
40
50
60
70
Age

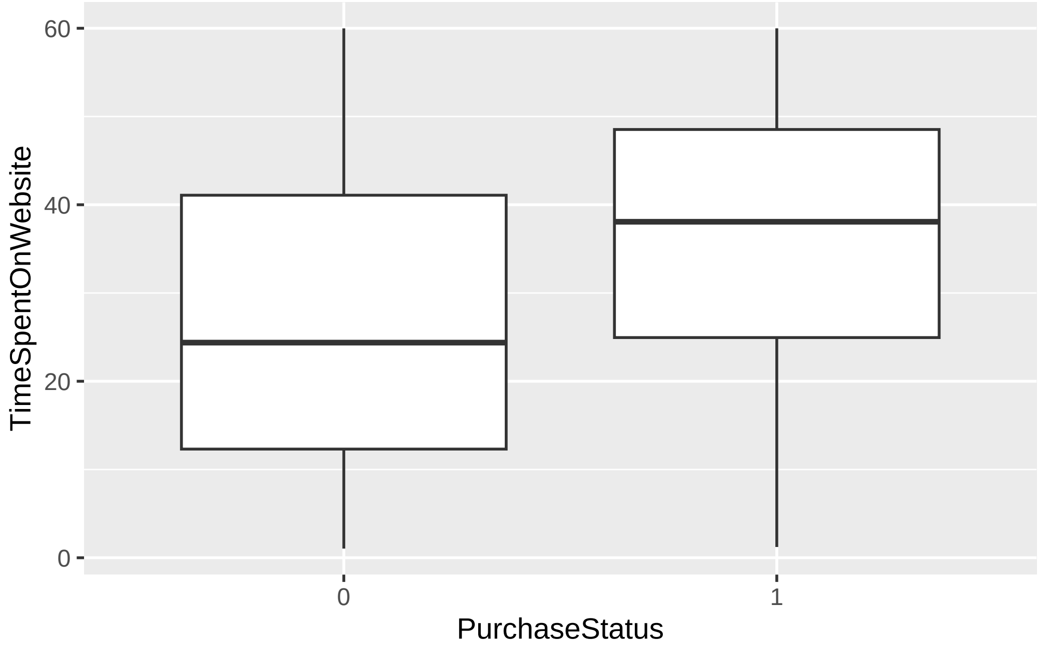

Histogram of AnnualIncome
Frequency
0
40
80
120
20000
40000
60000
80000
120000
AnnualIncome


Histogram of NumberOfPurchases
Frequency
0
50
100
150
0
5
10
15
20
NumberOfPurchases

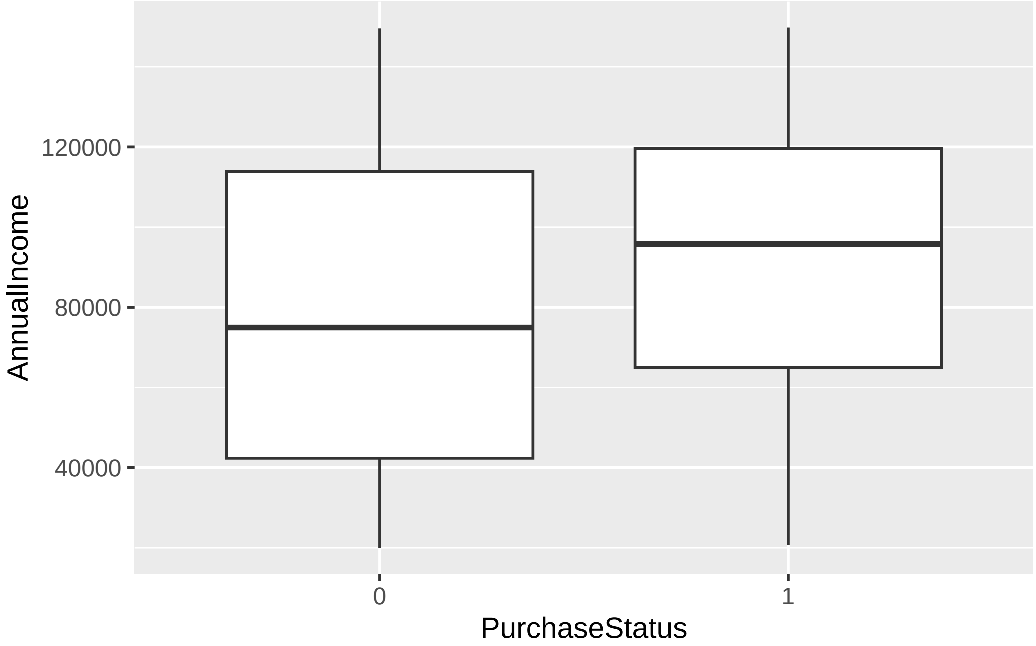

Histogram of TimeSpentOnWebsite
Frequency
0
50
100
150
0
10
20
30
40
50
60
TimeSpentOnWebsite



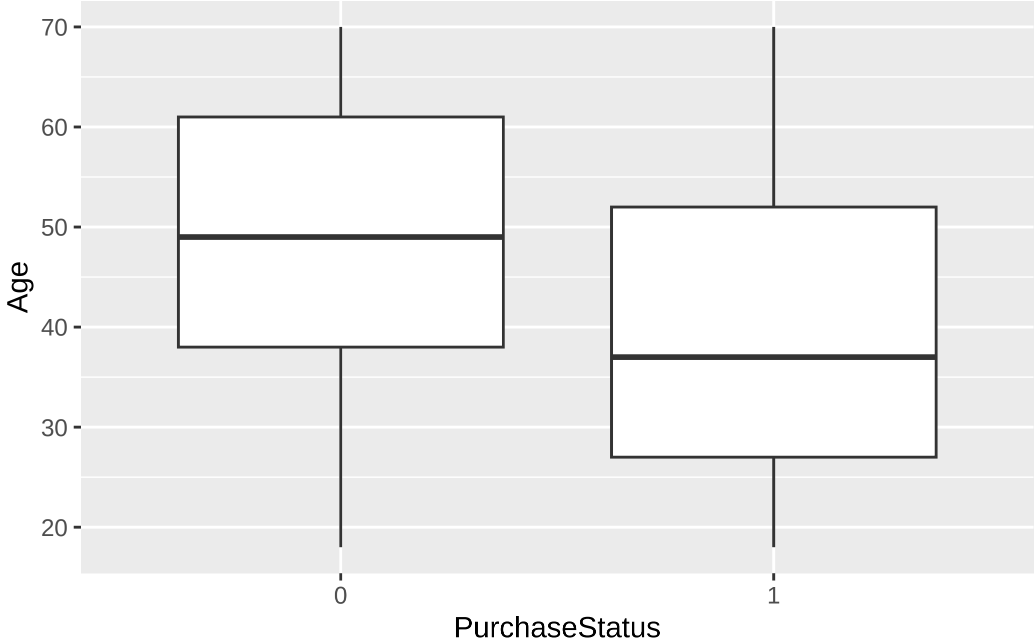
```
## Box plots
ggplot(data = raw_data, mapping = aes(x = PurchaseStatus, y = TimeSpentOnWebsite)) +
  geom_boxplot()
```



```
ggplot(data = raw_data, mapping = aes(x = PurchaseStatus, y = AnnualIncome)) +
  geom_boxplot()
```



```
ggplot(data = raw_data, mapping = aes(x = PurchaseStatus, y = Age)) +
  geom_boxplot()
```



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

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

```r
## 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
```