Shopping Trends Gender Analysis

Marta Sapizhak

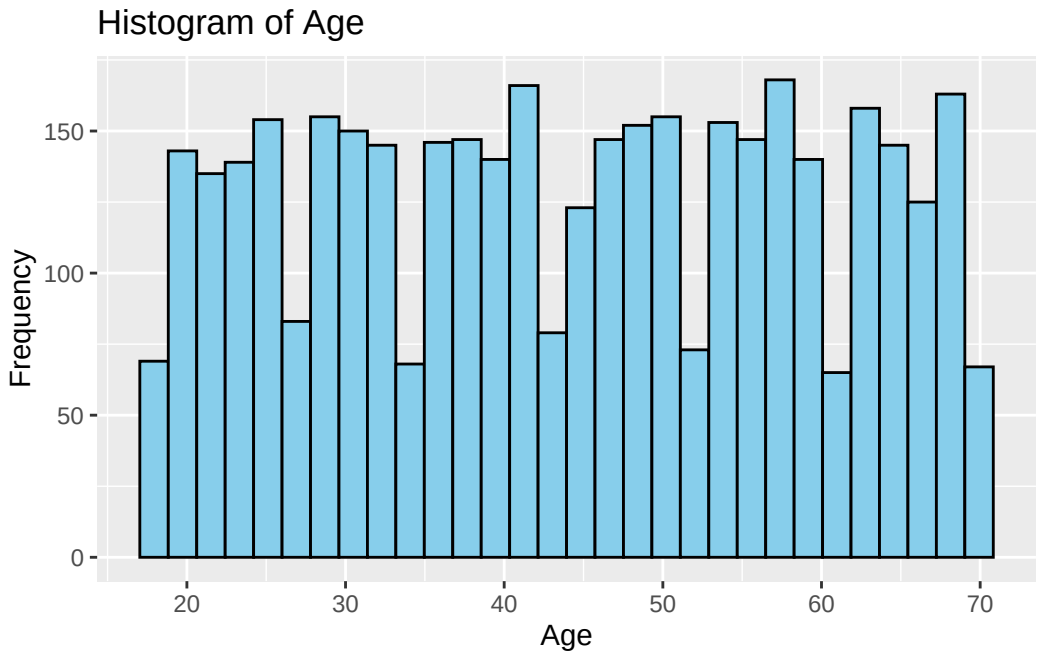
```
# Load raw data
shopping_raw <- read_csv("shopping_trends.csv")

# Display initial data
head(shopping_raw)
```

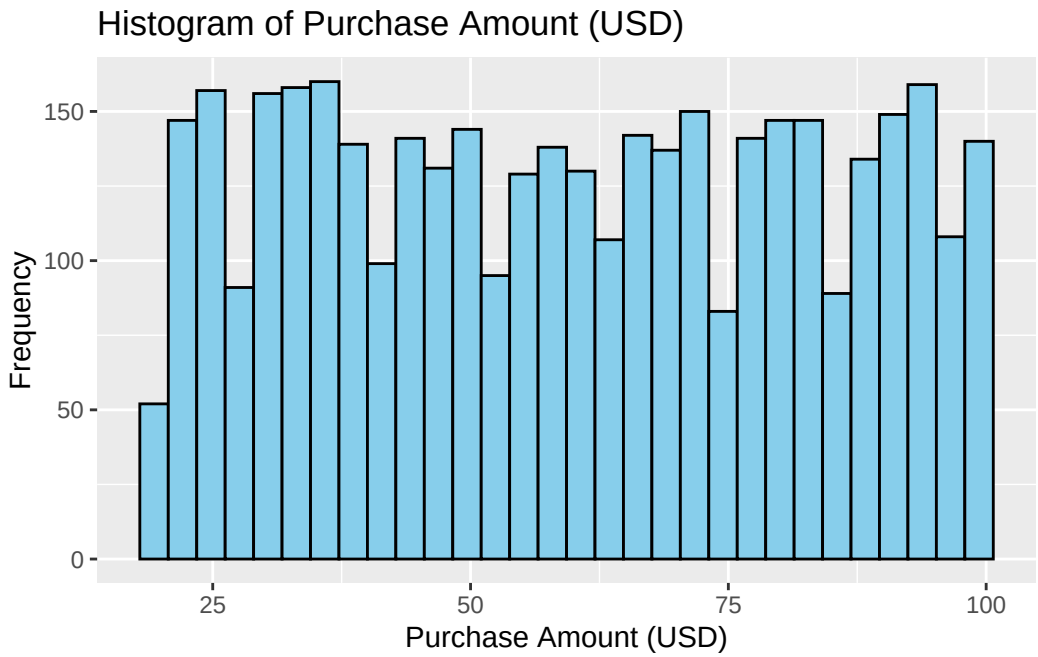
```
# A tibble: 6 x 19
  `Customer ID`   Age Gender `Item Purchased` Category `Purchase Amount (USD)`
          <dbl> <dbl> <chr>  <chr>            <chr>                        <dbl>
1             1    55 Male   Blouse           Clothing                        53
2             2    19 Male   Sweater          Clothing                        64
3             3    50 Male   Jeans            Clothing                        73
4             4    21 Male   Sandals          Footwear                        90
5             5    45 Male   Blouse           Clothing                        49
6             6    46 Male   Sneakers         Footwear                        20
# i 13 more variables: Location <chr>, Size <chr>, Color <chr>, Season <chr>,
#   `Review Rating` <dbl>, `Subscription Status` <chr>, `Payment Method` <chr>,
#   `Shipping Type` <chr>, `Discount Applied` <chr>, `Promo Code Used` <chr>,
#   `Previous Purchases` <dbl>, `Preferred Payment Method` <chr>,
#   `Frequency of Purchases` <chr>
```

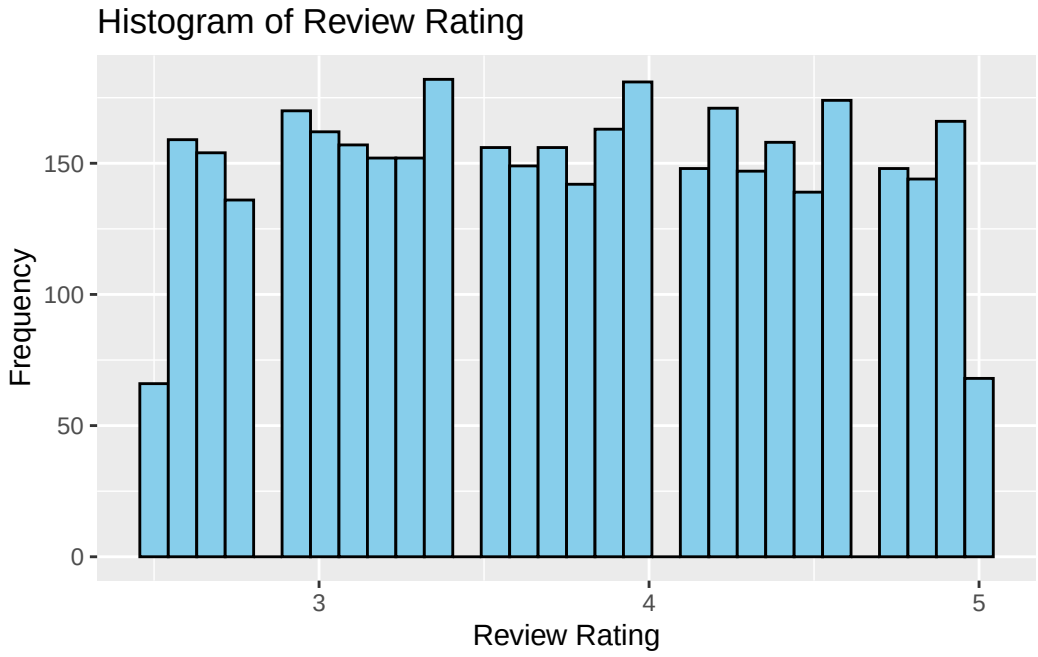
```
# Create binary outcome for gender prediction
shopping_with_outcome <- shopping_raw |>
  mutate(gender_binary = ifelse(Gender == "Female", 1, 0))
```

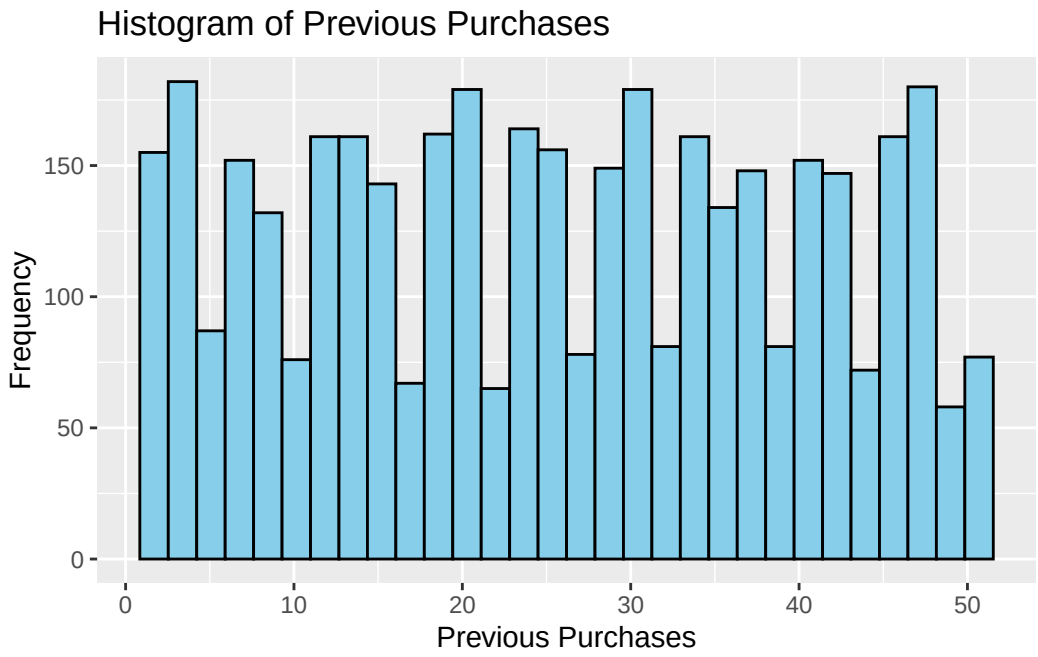
```
# Select key variables
vars_to_plot <- c("Age", "Purchase Amount (USD)", "Review Rating", "Previous Purchases")

# Create histogram for each variable using ggplot2
for (var in vars_to_plot) {
  p <- ggplot(shopping_raw, aes(x = .data[[var]])) +
```

```
    geom_histogram(bins = 30, fill = "skyblue", color = "black") +
    labs(title = paste("Histogram of", var),
         x = var,
         y = "Frequency")

  print(p)
}
```




Histogram of Purchase Amount (USD)
Frequency
150
100
50
0
25
50
75
100
Purchase Amount (USD)


Histogram of Review Rating
Frequency
150
100
50
0
3
4
5
Review Rating

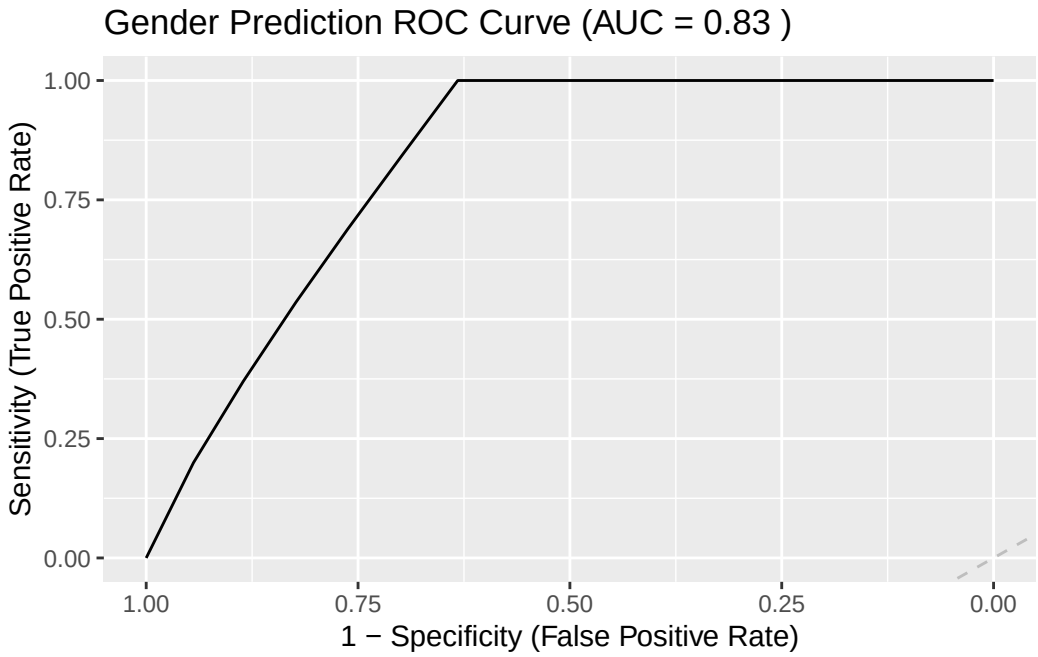



```
# Calculate VIF for the selected variables
vif_data <- shopping_raw |>
  select(all_of(vars_to_plot))

vif_values <- vif(lm(Age ~ ., data = vif_data))

# Print VIF values
print(vif_values)
```

```
`Purchase Amount (USD)`         `Review Rating`     `Previous Purchases`
               1.001011                1.000964                 1.000081
```

```
# Full model with all potential predictors
gender_full_model <- glm(gender_binary ~
                      `Discount Applied` +
                      `Promo Code Used` +
                      `Subscription Status` +
                      Location +
                      `Item Purchased` +
                      Color +
                      `Shipping Type`,
                      family = binomial(link = "logit"),
```

```
                    data = shopping_with_outcome)

# Perform stepwise selection
gender_stepwise_model <- step(gender_full_model, direction = "both", trace = 0)

# Get predictions using augment
gender_stepwise_predictions <- augment(gender_stepwise_model, type.predict = "response")

# Calculate ROC and AUC for stepwise model
gender_roc_stepwise <- roc(shopping_with_outcome$gender_binary, gender_stepwise_predictions$
gender_auc_stepwise <- auc(gender_roc_stepwise)

# Plot ROC curve using ggplot2
ggroc(gender_roc_stepwise) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray") +
  labs(title = paste("Gender Prediction ROC Curve (AUC =", round(gender_auc_stepwise, 3), ")
       x = "1 - Specificity (False Positive Rate)",
       y = "Sensitivity (True Positive Rate)")
```




```
# Look at coefficients of significant predictors
gender_coef_summary <- tidy(gender_stepwise_model) |>
  mutate(
```

```
    odds_ratio = exp(estimate),
    ci_lower = exp(estimate - 1.96 * std.error),
    ci_upper = exp(estimate + 1.96 * std.error)
  ) |>
  arrange(p.value)

# Display significant coefficients
gender_coef_summary |>
  filter(p.value < 0.05) |>
  select(term, estimate, odds_ratio, ci_lower, ci_upper, p.value)
```

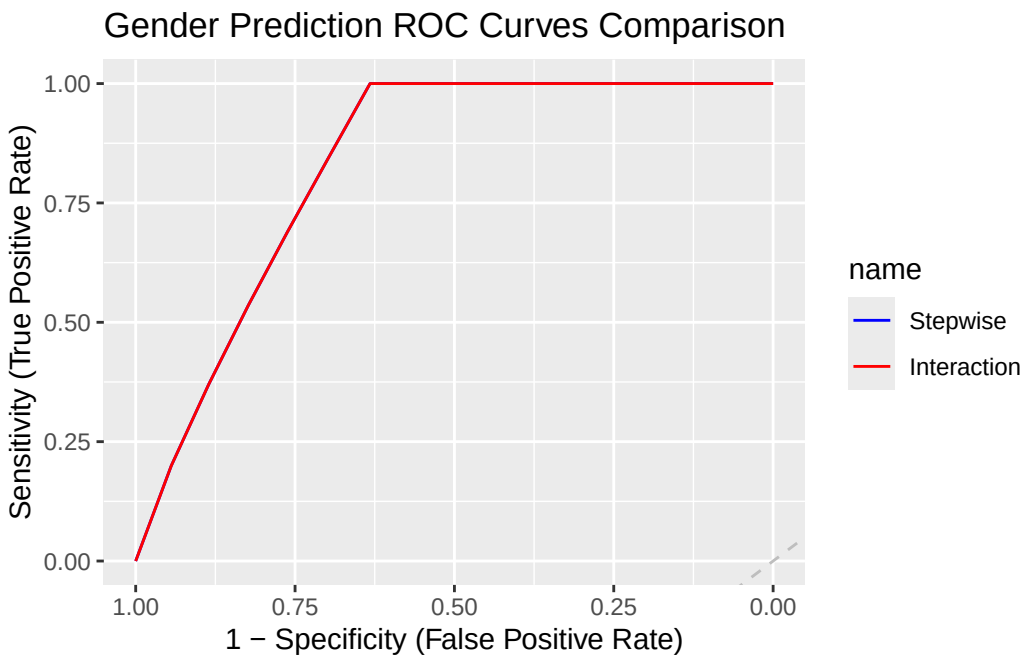
```
# A tibble: 2 x 6
  term                          estimate odds_ratio ci_lower ci_upper p.value
  <chr>                            <dbl>      <dbl>    <dbl>    <dbl>   <dbl>
1 (Intercept)                      0.227       1.25     1.02     1.54  0.0298
2 `Shipping Type`Free Shipping     0.293       1.34     1.00     1.79  0.0462
```

```
# Fit interaction model for gender prediction
gender_interaction_model <- glm(gender_binary ~
                           `Discount Applied` * `Promo Code Used` +
                           `Subscription Status` * `Discount Applied` +
                            `Shipping Type`,
                           family = binomial(link = "logit"),
                           data = shopping_with_outcome)

# Get predictions for interaction model
gender_interaction_predictions <- augment(gender_interaction_model, type.predict = "response

# Calculate ROC and metrics for interaction model
gender_roc_interaction <- roc(shopping_with_outcome$gender_binary, gender_interaction_predict
gender_auc_interaction <- auc(gender_roc_interaction)

# Calculate CIs for both models
gender_ci_stepwise <- ci.auc(gender_roc_stepwise)
gender_ci_interaction <- ci.auc(gender_roc_interaction)

# Compare models comprehensively
model_comparison <- tibble(
  Metric = c("AUC", "95% CI Lower", "95% CI Upper"),
  Stepwise = c(gender_auc_stepwise, gender_ci_stepwise[1], gender_ci_stepwise[3]),
  Interaction = c(gender_auc_interaction, gender_ci_interaction[1], gender_ci_interaction[3]
```

```
)

# Display model comparison
model_comparison
```

```
# A tibble: 3 x 3
  Metric       Stepwise Interaction
  <chr>           <dbl>       <dbl>
1 AUC             0.830       0.830
2 95% CI Lower    0.818       0.818
3 95% CI Upper    0.842       0.842
```

```
# Plot ROC curves for both models using ggplot2
ggroc(list(Stepwise = gender_roc_stepwise, Interaction = gender_roc_interaction)) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray") +
  labs(title = "Gender Prediction ROC Curves Comparison",
       x = "1 - Specificity (False Positive Rate)",
       y = "Sensitivity (True Positive Rate)") +
  scale_color_manual(values = c("Stepwise" = "blue", "Interaction" = "red"))
```



```r
# Calculate metrics for stepwise model
stepwise_metrics <- gender_stepwise_predictions |>
  mutate(predicted = if_else(.fitted > 0.5, 1, 0)) |>
  summarize(
    Sensitivity = sum(predicted == 1 & gender_binary == 1) / sum(gender_binary == 1),
    Specificity = sum(predicted == 0 & gender_binary == 0) / sum(gender_binary == 0),
    Accuracy = sum(predicted == gender_binary) / n(),
    Odds = Sensitivity / (1 - Specificity),
    `Odds Ratio` = 1
  )

interaction_metrics <- gender_interaction_predictions |>
  mutate(predicted = if_else(.fitted > 0.5, 1, 0)) |>
  summarize(
    Sensitivity = sum(predicted == 1 & gender_binary == 1) / sum(gender_binary == 1),
    Specificity = sum(predicted == 0 & gender_binary == 0) / sum(gender_binary == 0),
    Accuracy = sum(predicted == gender_binary) / n(),
    Odds = Sensitivity / (1 - Specificity),
    `Odds Ratio` = Odds / stepwise_metrics$Odds
  )

# Combine metrics into a single table
metrics_table <- bind_rows(
  stepwise_metrics |> mutate(Model = "Stepwise"),
  interaction_metrics |> mutate(Model = "Interaction")
)

# Display the metrics table
metrics_table
```

```
# A tibble: 2 x 6
  Sensitivity Specificity Accuracy  Odds `Odds Ratio` Model
        <dbl>       <dbl>    <dbl> <dbl>        <dbl> <chr>
1           1       0.632     0.75  2.72            1 Stepwise
2           1       0.632     0.75  2.72            1 Interaction
```

```r
gender_interaction_model
```

```
Call:  glm(formula = gender_binary ~ `Discount Applied` * `Promo Code Used` +
    `Subscription Status` * `Discount Applied` + `Shipping Type`,
```

```
    family = binomial(link = "logit"), data = shopping_with_outcome)

Coefficients:
                                       (Intercept)
                                          0.226773
                             `Discount Applied`Yes
                                         -19.817533
                              `Promo Code Used`Yes
                                                 NA
                          `Subscription Status`Yes
                                          0.002987
                            `Shipping Type`Express
                                         -0.112207
                     `Shipping Type`Free Shipping
                                          0.293467
                      `Shipping Type`Next Day Air
                                         -0.062096
                          `Shipping Type`Standard
                                          0.084663
                      `Shipping Type`Store Pickup
                                         -0.100579
    `Discount Applied`Yes:`Promo Code Used`Yes
                                                 NA
`Discount Applied`Yes:`Subscription Status`Yes
                                                 NA

Degrees of Freedom: 3899 Total (i.e. Null);  3892 Residual
Null Deviance:      4890
Residual Deviance: 3037     AIC: 3053
```

```
gender_stepwise_model
```

```
Call:  glm(formula = gender_binary ~ `Discount Applied` + `Shipping Type`,
    family = binomial(link = "logit"), data = shopping_with_outcome)

Coefficients:
                  (Intercept)          `Discount Applied`Yes
                      0.22677                       -19.81566
      `Shipping Type`Express  `Shipping Type`Free Shipping
                     -0.11221                        0.29347
 `Shipping Type`Next Day Air        `Shipping Type`Standard
```

```
                   -0.06210                        0.08466
 `Shipping Type`Store Pickup
                   -0.10058

Degrees of Freedom: 3899 Total (i.e. Null);  3893 Residual
Null Deviance:      4890
Residual Deviance: 3037     AIC: 3051
```