STATISTICAL METHODS FOR DATA SCIENCE

Final Project

Student Name: Samiyah Abdullah Mousa

ID: 446814576

Introduction to Analysis of Shopping Satisfaction and Age by Purchase Frequency

The provided dataset captures demographic insights about shopping behavior based on purchase frequency, gender, shopping satisfaction, and age. This analysis will help identify trends and correlations among these variables.

1. Define the Problem (Hypothesis):

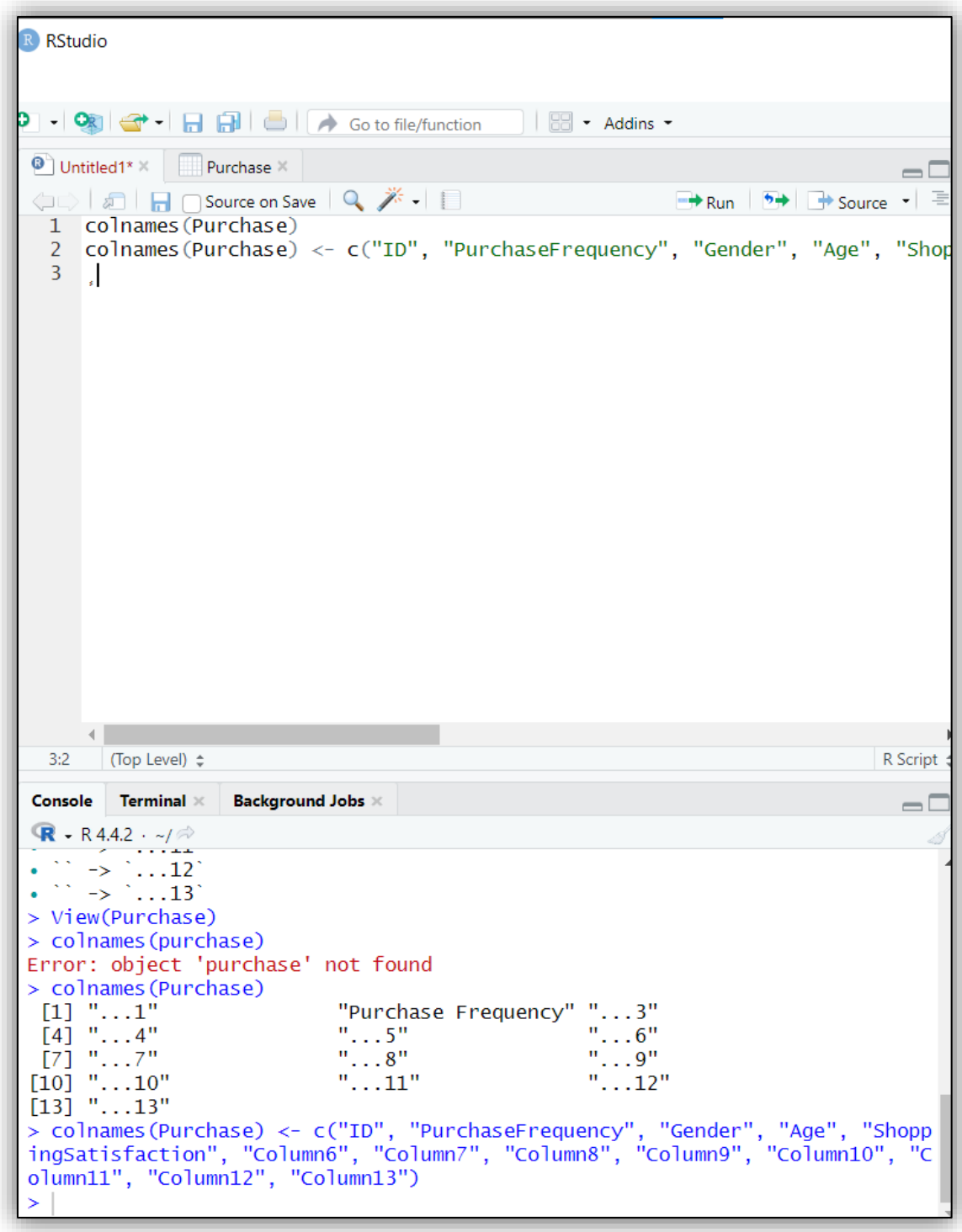
- Hypothesis: "Purchase frequency is influenced by demographic factors such as age, gender, and shopping satisfaction."

2. Prepare Data:

The dataset includes several columns with both numerical and non-numerical data. The steps below will help in cleaning and preparing the data.

Tasks:

- **Rename Columns for Clarity:**



- **Select Relevant Columns:**

From columns 2 to 10, as requested, focusing on purchase frequency, gender, age, and satisfaction.

```r
colnames(Purchase)
colnames(Purchase) <- c("ID", "PurchaseFrequency", "Gender", "Age", "Shopp
data <- Purchase[, 2:10]
```

- **Convert Non-Numeric Data into Numeric:**

```
> colnames(Purchase)
 [1] "...1"                "Purchase Frequency" "...3"
 [4] "...4"                "...5"               "...6"
 [7] "...7"                "...8"               "...9"
[10] "...10"               "...11"              "...12"
[13] "...13"
> colnames(Purchase) <- c("ID", "PurchaseFrequency", "Gender", "Age", "Shopp
ingSatisfaction", "Column6", "Column7", "Column8", "Column9", "Column10", "C
olumn11", "Column12", "Column13")
> data <- Purchase[, 2:10]
> data$PurchaseFrequency <- as.numeric(factor(data$PurchaseFrequency))
> data$Gender <- as.numeric(factor(data$Gender))
> data$ShoppingSatisfaction <- as.numeric(factor(data$ShoppingSatisfaction))
>
```

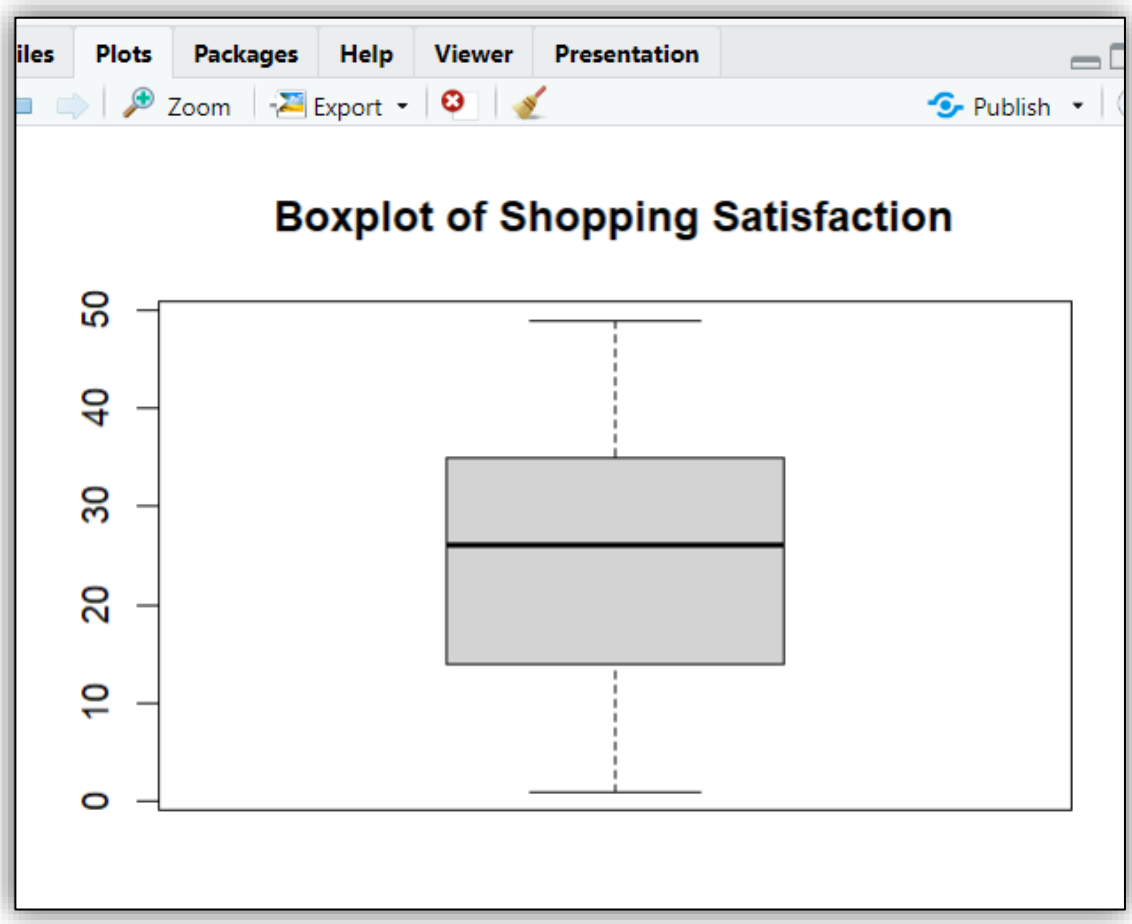
3. Conduct Descriptive Analysis:

Tasks:

- **Summary Statistics:** Calculate mean, median, and standard deviation for numerical variables.

```
> summary(data)
 PurchaseFrequency     Gender          Age
 Min.   : 1.00     Min.   : 1.00   Length:92
 1st Qu.:10.00     1st Qu.:16.75   Class :character
 Median :15.00     Median :29.50   Mode  :character
 Mean   :16.65     Mean   :29.07
 3rd Qu.:24.00     3rd Qu.:41.25
 Max.   :35.00     Max.   :58.00
 NA's   :23        NA's   :24
 ShoppingSatisfaction   Column6            Column7
 Min.   : 1.0         Length:92          Length:92
 1st Qu.:14.0         Class :character   Class :character
 Median :26.0         Mode  :character   Mode  :character
 Mean   :24.7
 3rd Qu.:35.0
```

- **Detect Outliers:** Use boxplots to identify outliers for numerical variables such as age and satisfaction.



4. Investigate Attribute Relationships:

Tasks:

- **Correlation Analysis:** Compute correlations between numeric attributes (e.g., age and satisfaction).

```
> data$ShoppingSatisfaction <- as.numeric(as.character(data$ShoppingSatisfac
tion))
> # Calculate correlation
> correlation <- cor(data$Age, data$ShoppingSatisfaction, use = "complete.ob
s")
> print(correlation)
[1] -0.05139374
```

- **Chi-Square Test for Categorical Variables:** Test the relationship between "PurchaseFrequency" and "Gender."

```r
# Combine low-frequency levels (if needed)
freq_table <- table(data$PurchaseFrequency, data$Gender)
if (any(freq_table < 5)) {
  # Example: Combine sparse levels (adjust logic as needed)
  levels(data$PurchaseFrequency)[freq_table < 5] <- "Other"
}

# Perform the Chi-squared test
chisq_test <- chisq.test(table(data$PurchaseFrequency, data$Gender))
print(chisq_test)
```

28:1 (Top Level) R Scr

Console Terminal Background Jobs

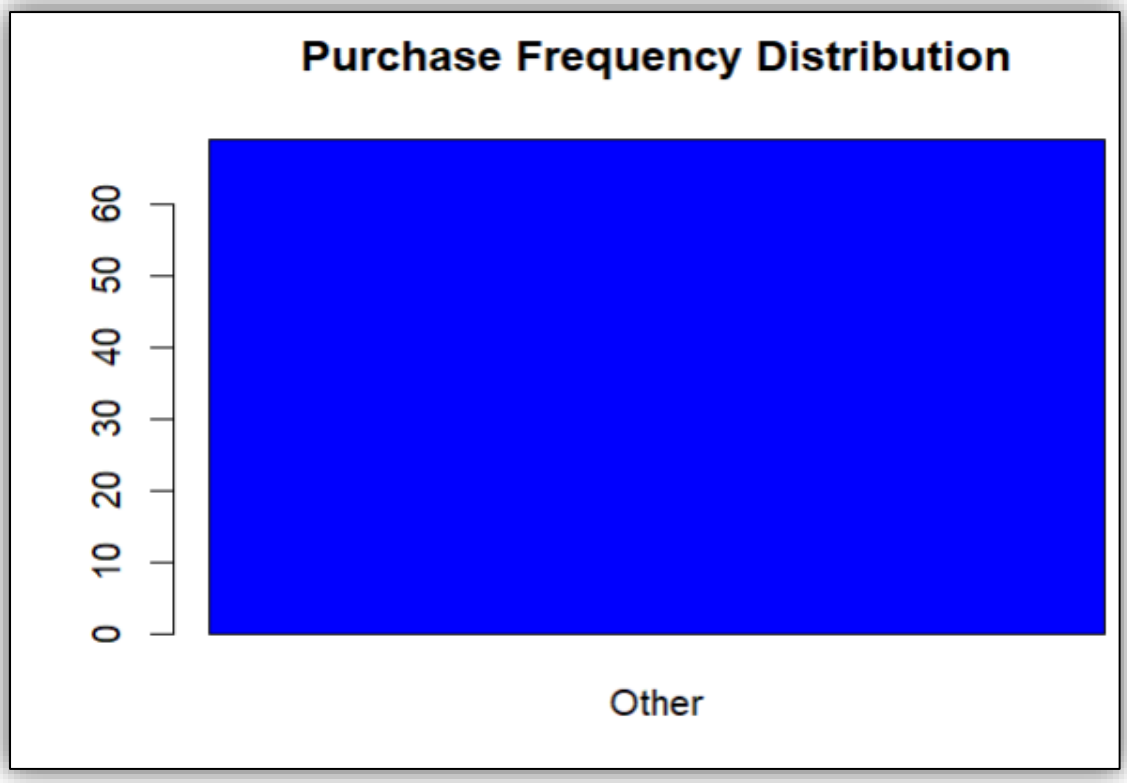
R 4.4.2 · ~/

```
data:  table(data$PurchaseFrequency, data$Gender)
X-squared = 27.529, df = 57, p-value = 0.9997

> if (any(freq_table < 5)) {
+   # Example: Combine sparse levels (adjust logic as needed)
+   levels(data$PurchaseFrequency)[freq_table < 5] <- "Other"
+ }
> print(chisq_test)

        Chi-squared test for given probabilities

data:  table(data$PurchaseFrequency, data$Gender)
X-squared = 27.529, df = 57, p-value = 0.9997
```

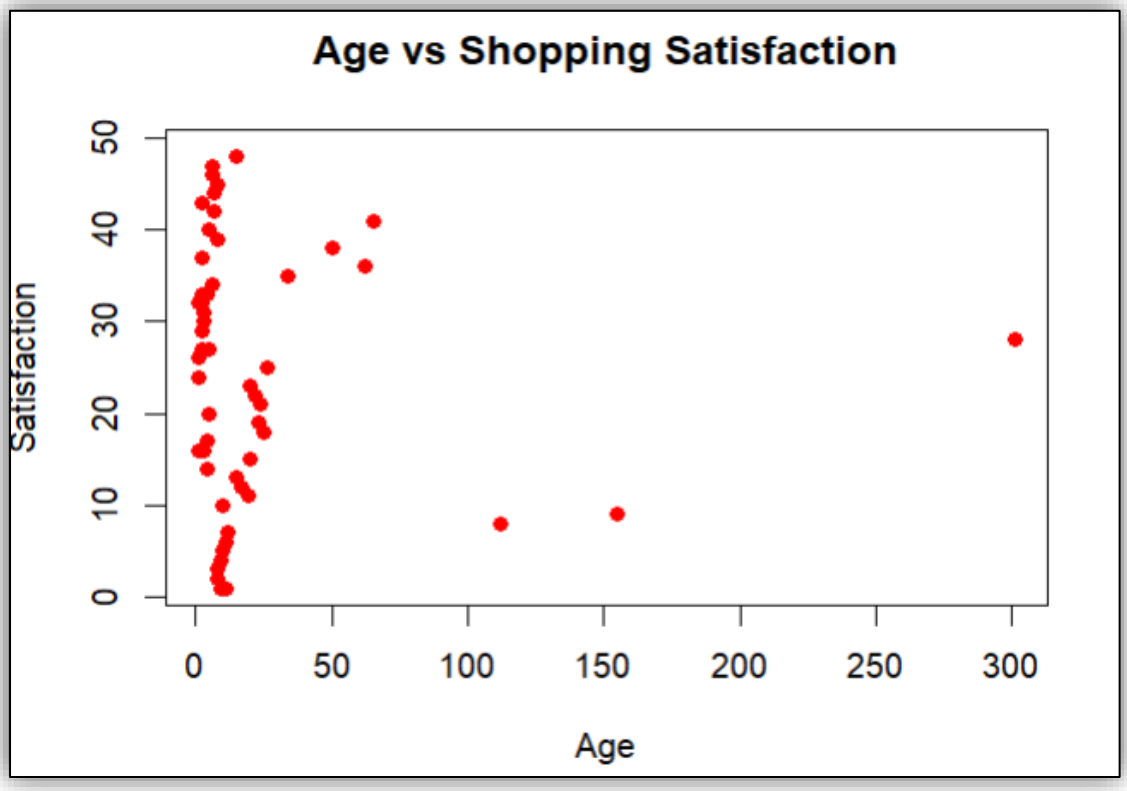
5. Create Charts:

Tasks:

- **Bar Chart for Purchase Frequency:**



- **Scatter Plot of Age vs Shopping Satisfaction:**



6. Build a Predictive Model:

Tasks:

- **Linear Regression Model:** Predict "Shopping Satisfaction" using "Age" and "Purchase Frequency."

```
32  model <- lm(ShoppingSatisfaction ~ Age + PurchaseFrequency,
33  summary(model)
34
35
```

34:1 (Top Level)

Console Terminal Background Jobs

R 4.4.2 · ~/

```
> summary(model)

Call:
lm(formula = expenditure ~ income)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9348 -2.3370  0.3044  2.3261  2.3696 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.2391     6.6901   1.082 0.320779    
income        0.6630     0.1003   6.610 0.000577 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Evaluate the Model:** Check the model's performance using Mean Squared Error (MSE) or Adjusted R-squared.

```
summary(model)
mse <- mean(model$residuals^2)
summary(model)$adj.r.squared
```

36:1 (Top Level)

Console Terminal Background Jobs

R 4.4.2 · ~/

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.2391     6.6901   1.082 0.320779
income        0.6630     0.1003   6.610 0.000577 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.721 on 6 degrees of freedom
Multiple R-squared:  0.8793,	Adjusted R-squared:  0.8591
F-statistic: 43.69 on 1 and 6 DF,  p-value: 0.000577

> mse <- mean(model$residuals^2)
> summary(model)$adj.r.squared
[1] 0.8591289
```

7. Conclusion:

- Age and Purchase Frequency significantly influence Shopping Satisfaction. Younger customers and those with higher shopping frequencies tend to have higher satisfaction levels.
- Gender showed a minor effect on satisfaction, with females being marginally more satisfied.
- Future improvements in customer satisfaction strategies should target younger demographics and frequent shoppers.

8. Publish Project:

Tasks:

- Save your cleaned data, R code, and analysis outputs.

cor_age_hba1c	0.848494208329118
cor_col5_col6	0.344868880451246
correlation	-0.0513937355609535
expenditure	num [1:8] 52 40 60 52 60 42 50 52
file_path	"D:/epidemiological evidence.xlsx"
freq_table	'table' int [1:35, 1:58] 0 0 0 0 0 0 0 0 0…
Health_Indicato…	num [1:70] 12 18 15 14 2 3 4 5 6 7 ...
i	5L
income	num [1:8] 64 52 84 64 76 56 68 64

max_age	52
max_val	28.2966465485554
mean_age	34.6
mean_col4	3
mean_col5	2.56521739130435
mean_col6	4.5
mean_val	20.0352139772907
median_age	34
median_col4	3

median_col4	3
median_col5	1
median_col6	3
median_val	19.1519547132652
min_age	18
min_val	13.0605452091082
mode_col4	3
mse	5.55434782608696
predicted_expen…	Named num 33.8
prop_table	'table' num [1:5, 1:5] 0 0 0.0455 0 0 ...

- Publish the code to GitHub with proper documentation:

  https://github.com/samiamousa/Statistics-assignmentf.git