Research Questions:

1. Data Analysis: perform exploratory data analysis, including data cleaning, visualization, and summary statistics using data analysis tool Power BI.

2. Statistical Modeling: Based on the dataset, you will choose an appropriate statistical model to analyze relationships between variables. They will fit the model and interpret the results in the context of the research.

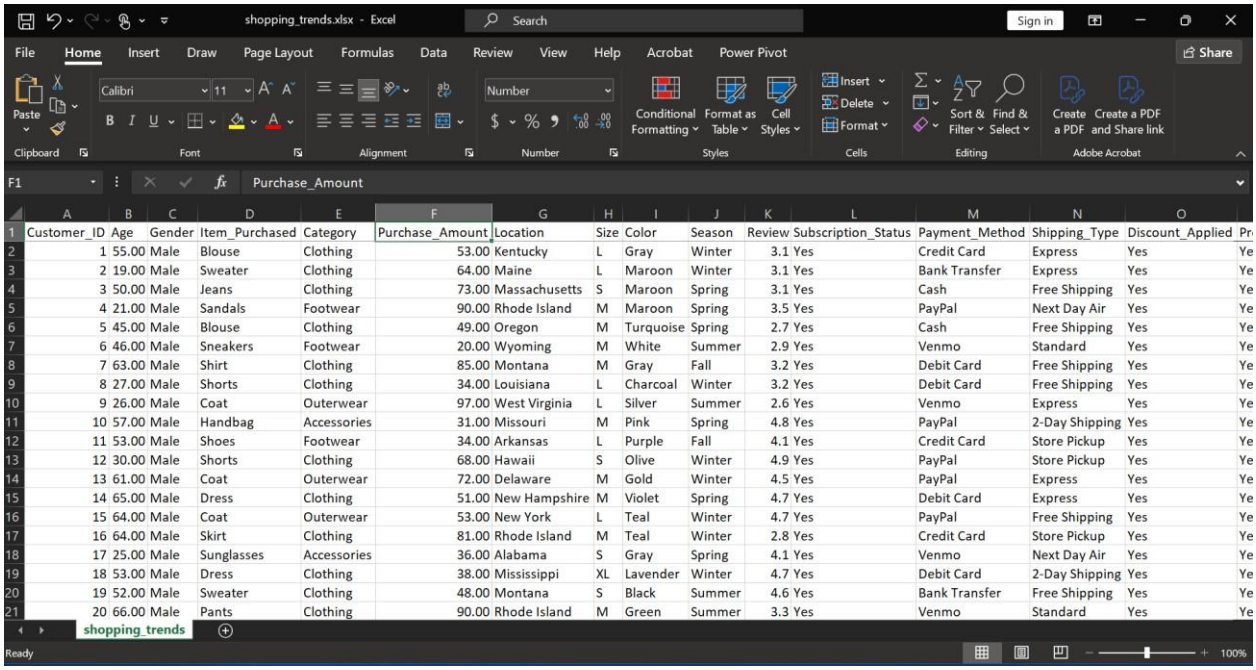
3. Simulation Study: You will design and write a code to simulate the process, generate outcomes, and analyze the results.

4. Report Writing: You will prepare a comprehensive report documenting their findings from the data analysis, statistical modeling, and simulation study. The report should include clear explanations, appropriate visualizations, and interpretations of the results.

5. Presentation: You will deliver a brief presentation summarizing their findings and key insights from the assignment. You will present their analysis, modeling approach, and simulation results to the class.

Response

Data used for this analysis includes Shopping trends for various customers and was downloaded from Kaggle. The link is provided below https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset



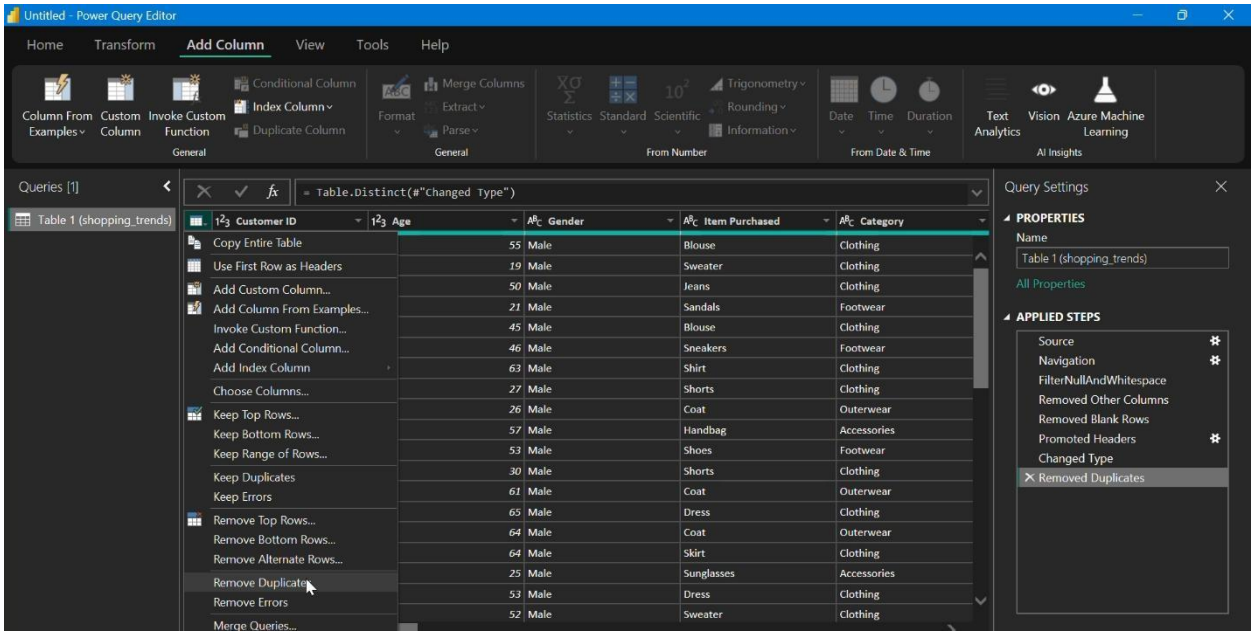
shopping_trends.xlsx - Excel

Customer_ID	Age	Gender	Item_Purchased	Category	Purchase_Amount	Location	Size	Color	Season	Review	Subscription_Status	Payment_Method	Shipping_Type	Discount_Applied	Pr
1	55.00	Male	Blouse	Clothing	53.00	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Express	Yes	Ye
2	19.00	Male	Sweater	Clothing	64.00	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Express	Yes	Ye
3	50.00	Male	Jeans	Clothing	73.00	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Free Shipping	Yes	Ye
4	21.00	Male	Sandals	Footwear	90.00	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Day Air	Yes	Ye
5	45.00	Male	Blouse	Clothing	49.00	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Free Shipping	Yes	Ye
6	46.00	Male	Sneakers	Footwear	20.00	Wyoming	M	White	Summer	2.9	Yes	Venmo	Standard	Yes	Ye
7	63.00	Male	Shirt	Clothing	85.00	Montana	M	Gray	Fall	3.2	Yes	Debit Card	Free Shipping	Yes	Ye
8	27.00	Male	Shorts	Clothing	34.00	Louisiana	L	Charcoal	Winter	3.2	Yes	Debit Card	Free Shipping	Yes	Ye
9	26.00	Male	Coat	Outerwear	97.00	West Virginia	L	Silver	Summer	2.6	Yes	Venmo	Express	Yes	Ye
10	57.00	Male	Handbag	Accessories	31.00	Missouri	M	Pink	Spring	4.8	Yes	PayPal	2-Day Shipping	Yes	Ye
11	53.00	Male	Shoes	Footwear	34.00	Arkansas	L	Purple	Fall	4.1	Yes	Credit Card	Store Pickup	Yes	Ye
12	30.00	Male	Shorts	Clothing	68.00	Hawaii	S	Olive	Winter	4.9	Yes	PayPal	Store Pickup	Yes	Ye
13	61.00	Male	Coat	Outerwear	72.00	Delaware	M	Gold	Winter	4.5	Yes	PayPal	Express	Yes	Ye
14	65.00	Male	Dress	Clothing	51.00	New Hampshire	M	Violet	Spring	4.7	Yes	Debit Card	Express	Yes	Ye
15	64.00	Male	Coat	Outerwear	53.00	New York	L	Teal	Winter	4.7	Yes	PayPal	Free Shipping	Yes	Ye
16	64.00	Male	Skirt	Clothing	81.00	Rhode Island	M	Teal	Winter	2.8	Yes	Credit Card	Store Pickup	Yes	Ye
17	25.00	Male	Sunglasses	Accessories	36.00	Alabama	S	Gray	Spring	4.1	Yes	Venmo	Next Day Air	Yes	Ye
18	53.00	Male	Dress	Clothing	38.00	Mississippi	XL	Lavender	Winter	4.7	Yes	Debit Card	2-Day Shipping	Yes	Ye
19	52.00	Male	Sweater	Clothing	48.00	Montana	S	Black	Summer	4.6	Yes	Bank Transfer	Free Shipping	Yes	Ye
20	66.00	Male	Pants	Clothing	90.00	Rhode Island	M	Green	Summer	3.3	Yes	Venmo	Standard	Yes	Ye

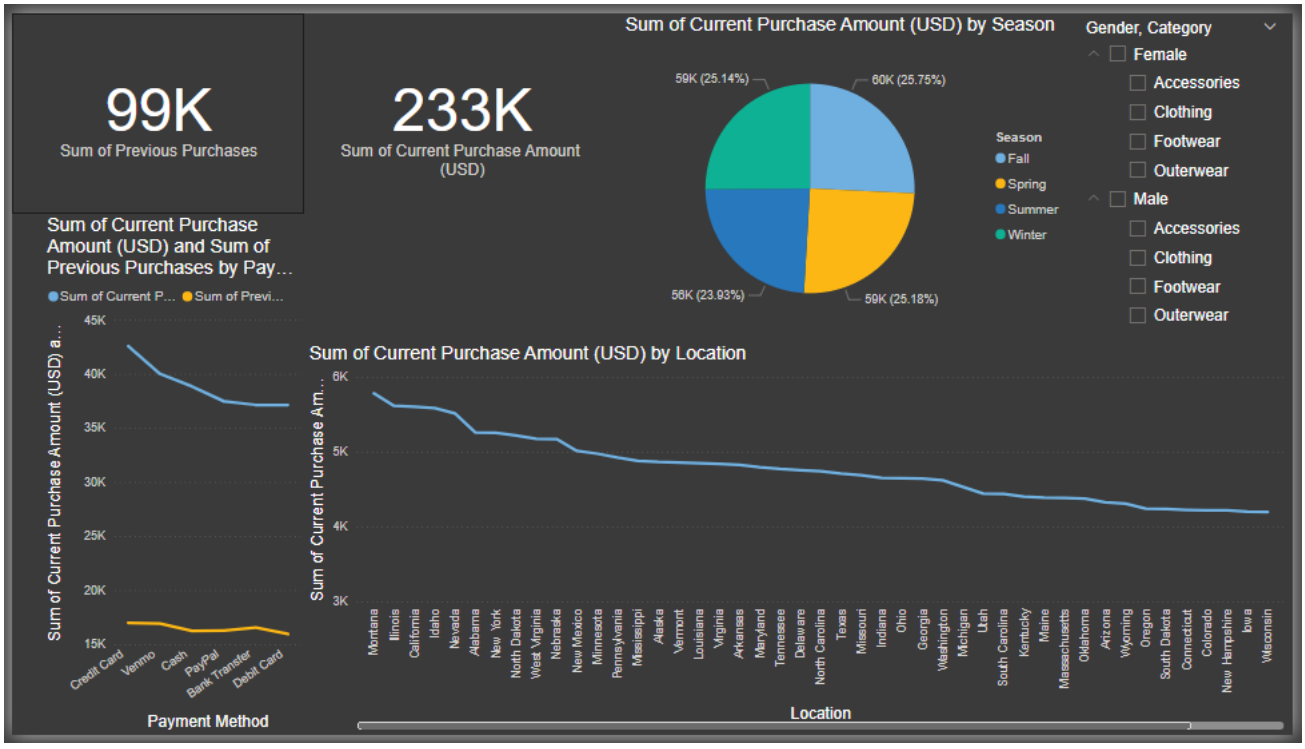
The dataset contains the following columns:

i. Customer_ID: Unique identifier for each customer.
ii. Age: Age of the customer. iii. Gender: Gender of the customer.
iv. Item_Purchased: Name of the item purchased.
v. Category: Category of the item purchased.
vi. Purchase_Amount: Amount spent on the purchase.
vii. Location: Customer's location.
viii. Size: Size of the item purchased. ix. Color: Color of the item purchased.
x. Season: Season of the purchase.
xi. Review Rating: Customer's review rating for the purchase. xii. Subscription_Status: Whether the customer has a subscription.
xiii. Payment_Method: Payment method used. xiv. Shipping_Type: Type of shipping chosen.
xv. Discount_Applied: Whether a discount was applied. xvi. Promo_Code_Used: Whether a promo code was used. xvii. Previous_Purchases: Number of previous purchases made by the customer.
xviii. Preferred_Payment_Method: Customer's preferred payment method.
xix. Frequency_of_Purchases: Purchase frequency.

data cleaning (removing Duplicates and replaces spaces between the names with '_')



data analysis, visualization, and summary statistics using data analysis tool Power BI



statistical model to analyze relationships between variables in R.

codes used in r *library(readxl)*

shopping_trends <- read_excel("D:/notes/MASTERS/Computational Statistics and Programming/ASSIGNMENT/shopping_trends.xlsx")

View(shopping_trends)

#Convert categorical variables to factors shopping_trends$Gender <-

as.factor(shopping_trends$Gender) shopping_trends$Location <-

as.factor(shopping_trends$Location) shopping_trends$Season <-

as.factor(shopping_trends$Season) shopping_trends$Subscription_Status <-

as.factor(shopping_trends$Subscription_Status) shopping_trends$Payment_Method <-

as.factor(shopping_trends$Payment_Method) shopping_trends$Shipping_Type <-

as.factor(shopping_trends$Shipping_Type) shopping_trends$Discount_Applied <-

as.factor(shopping_trends$Discount_Applied) shopping_trends$Promo_Code_Used <-

as.factor(shopping_trends$Promo_Code_Used)

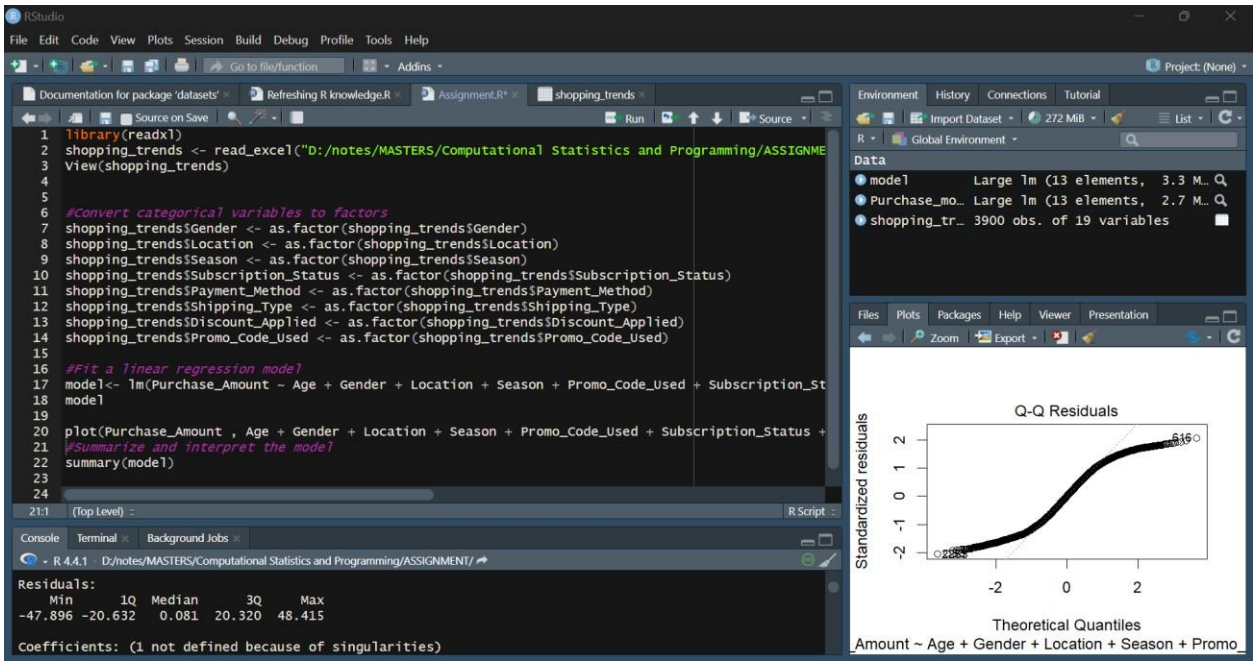
#Fit a linear regression model

model<- lm(Purchase_Amount ~ Age + Gender + Location + Season + Promo_Code_Used + Subscription_Status + Payment_Method + Shipping_Type + Discount_Applied + Promo_Code_Used , data = shopping_trends) model

plot(Purchase_Amount , Age + Gender + Location + Season + Promo_Code_Used + Subscription_Status + Payment_Method + Shipping_Type + Discount_Applied + Promo_Code_Used)

#Summarize and interpret the model summary(model)

plot(model)



Results

Call:

lm(formula = Purchase Amount ~ Age + Gender + Location + Season +

Promo_Code_Used + Subscription_Status + Payment_Method +

Shipping_Type + Discount_Applied + Promo_Code_Used, data = shopping trends)

Residuals:

Min 1Q Median 3Q Max

-47.896 -20.632 0.081 20.320 48.415

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 61.993673 3.088126 20.075 < 2e-16 ***

Age -0.013795 0.025080 -0.550 0.58232

GenderMale -0.169645 1.015718 -0.167 0.86736

LocationAlaska 8.505758 3.753205 2.266 0.02349 * LocationArizona

7.472806 3.856471 1.938 0.05273 .

LocationArkansas 2.310548 3.655593 0.632 0.52739

LocationCalifornia -0.115180 3.487976 -0.033 0.97366

LocationColorado -2.363213 3.712834 -0.636 0.52449

LocationConnecticut -5.063863 3.667923 -1.381 0.16749

LocationDelaware -4.113601 3.577510 -1.150 0.25028

LocationFlorida -3.590174 3.813203 -0.942 0.34650

LocationGeorgia -0.769861 3.657131 -0.211 0.83328

LocationHawaii -1.374635 3.856537 -0.356 0.72153

LocationIdaho 1.301760 3.506267 0.371 0.71046

LocationIllinois 1.942980 3.517194 0.552 0.58069

LocationIndiana 0.055510 3.659495 0.015 0.98790

LocationIowa 1.738326 3.794950 0.458 0.64693

LocationKansas -4.484439 3.895404 -1.151 0.24972

LocationKentucky -3.431461 3.659011 -0.938 0.34840

LocationLouisiana -1.224882 3.599685 -0.340 0.73367

LocationMaine -2.286020 3.685924 -0.620 0.53516

LocationMaryland -3.022450 3.584297 -0.843 0.39914

LocationMassachusetts 2.298504 3.752338 0.613 0.54021

LocationMichigan 2.986077 3.734802 0.800 0.42403

LocationMinnesota -2.505074 3.555304 -0.705 0.48110

LocationMississippi 1.798039 3.644179 0.493 0.62176

LocationMissouri -0.999087 3.634544 -0.275 0.78342

LocationMontana 1.432420 3.485197 0.411 0.68109

LocationNebraska 0.417981 3.563367 0.117 0.90663

LocationNevada 4.729803 3.568500 1.325 0.18511 LocationNew Hampshire -0.003124
3.764503 -0.001 0.99934

LocationNew Jersey	-1.914175	3.824480	-0.501	0.61675
LocationNew Mexico	2.714185	3.633198	0.747	0.45508
LocationNew York	1.043913	3.567792	0.293	0.76985
LocationNorth Carolina	1.865715	3.673119	0.508	0.61153
LocationNorth Dakota	3.802759	3.609525	1.054	0.29216
LocationOhio	1.132307	3.683244	0.307	0.75854
LocationOklahoma	-0.655295	3.708189	-0.177	0.85974
LocationOregon	-1.942733	3.726236	-0.521	0.60214
LocationPennsylvania	7.453333	3.722677	2.002	0.04534 *
LocationRhode Island	2.559449	3.896997	0.657	0.51137
LocationSouth Carolina	-0.568712	3.696044	-0.154	0.87772
LocationSouth Dakota	1.236671	3.775712	0.328	0.74328
LocationTennessee	2.706750	3.682785	0.735	0.46240
LocationTexas	1.865818	3.682947	0.507	0.61246
LocationUtah	3.097033	3.772312	0.821	0.41170
LocationVermont	-2.244775	3.592312	-0.625	0.53208
LocationVirginia	3.807142	3.687393	1.032	0.30191
LocationWashington	4.258660	3.736552	1.140	0.25447
LocationWest Virginia	4.668712	3.630842	1.286	0.19857
LocationWisconsin	-2.977612	3.711459	-0.802	0.42244
LocationWyoming	1.189606	3.763176	0.316	0.75193
SeasonSpring	-2.977570	1.073092	-2.775	0.00555 **
SeasonSummer	-3.407707	1.081893	-3.150	0.00165 **
SeasonWinter	-1.195842	1.079659	-1.108	0.26810
Promo_Code_UsedYes	-0.880739	1.222105	-0.721	0.47115
Subscription_StatusYes	0.301603	1.205110	0.250	0.80239
Payment_MethodCash	1.186683	1.328127	0.894	0.37164
Payment_MethodCredit Card	2.204655	1.309193	1.684	0.09227 .
Payment_MethodDebit Card	-0.208399	1.337426	-0.156	0.87618

Payment_MethodPayPal 0.046957 1.334459 0.035 0.97193

Payment_MethodVenmo 2.525844 1.328445 1.901 0.05733 .

Shipping_TypeExpress -0.270667 1.335072 -0.203 0.83935

Shipping_TypeFree Shipping -0.422253 1.318439 -0.320 0.74878

Shipping_TypeNext Day Air -2.097325 1.329488 -1.578 0.11475 Shipping_TypeStandard

-2.280367 1.327810 -1.717 0.08599 .

Shipping_TypeStore Pickup -0.731182 1.333779 -0.548 0.58358

Discount_AppliedYes NA NA NA NA ---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.61 on 3833 degrees of freedom

Multiple R-squared: 0.02311, Adjusted R-squared: 0.006285

F-statistic: 1.374 on 66 and 3833 DF, p-value: 0.025

Interpretations

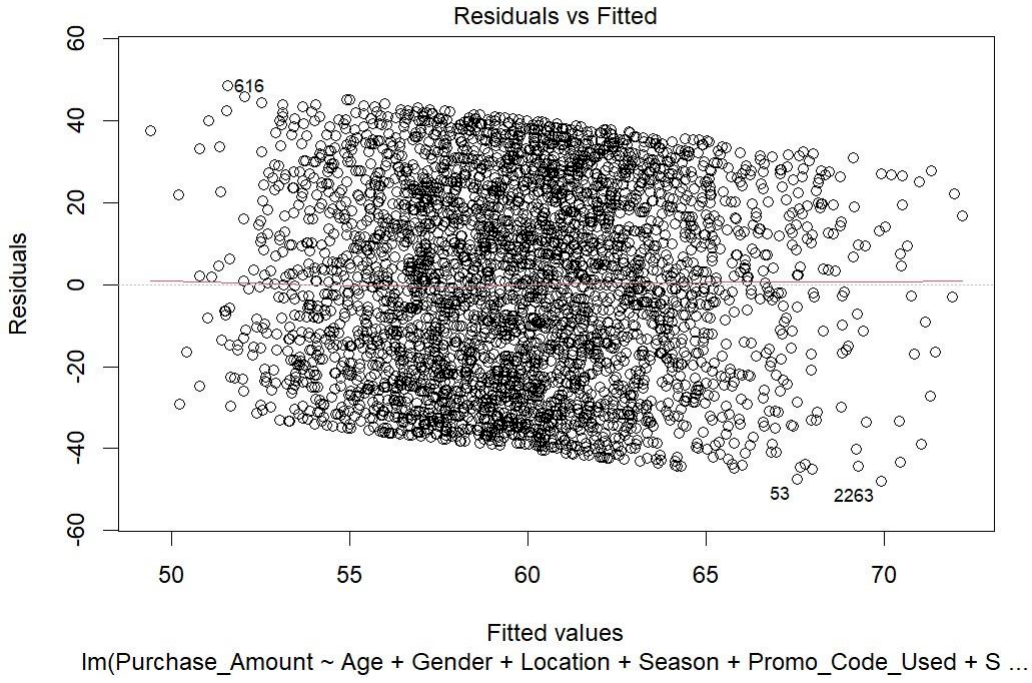
Intercept: When all predictors are zero, the predicted Purchase Amount is 61.993673.

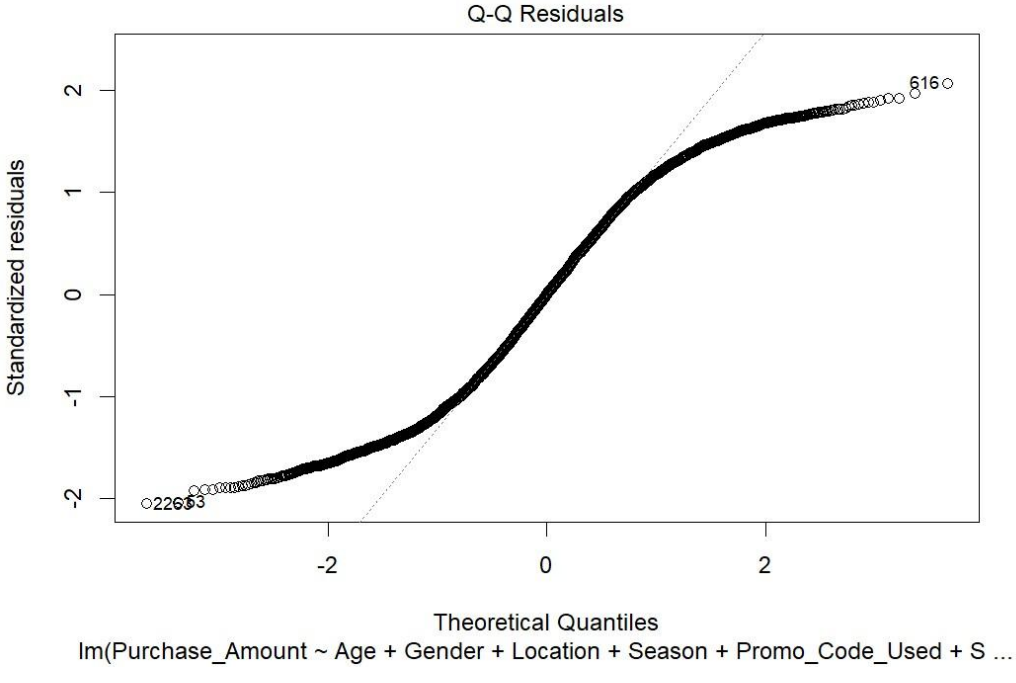
Age: For each additional year of age, the purchase amount decreases by 0.013795 units (statistically significant, at 5% level of significance).

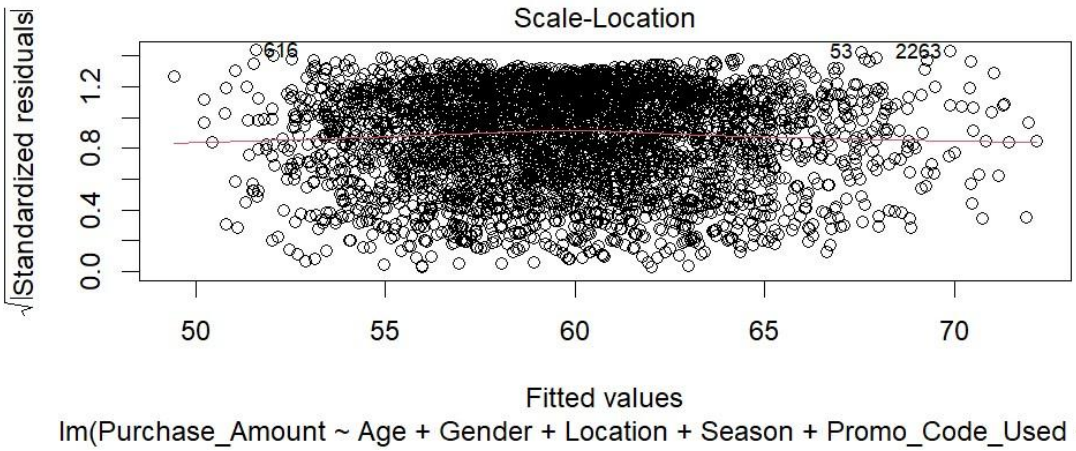
GenderMale: Male customers spend 0.169645 units less compared to the baseline gender (e.g., Female), this value is statistically significant at 5%.

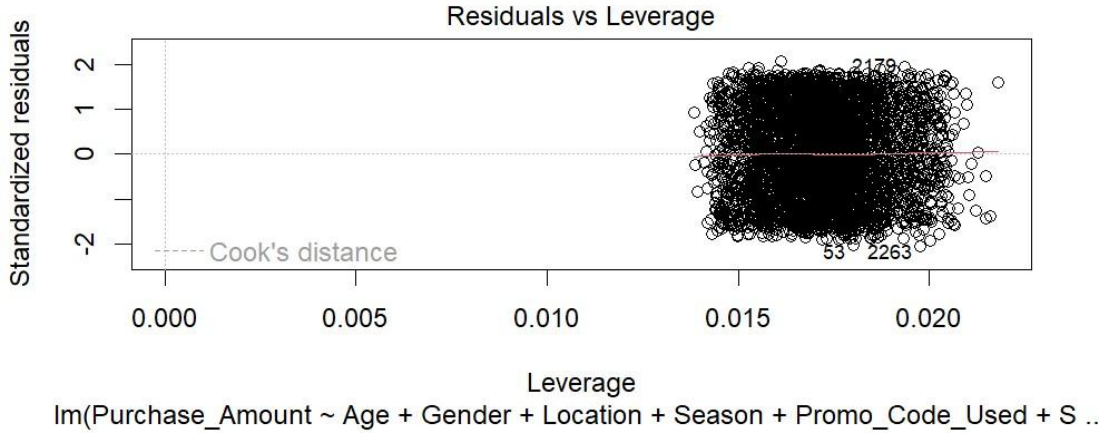
Season Spring: Purchases in spring reduce the amount by 2.977570 units compared to the baseline season, statistically significant at 5%.

Subscription status: Subscription status had a positive impact on the sales & increased it by 0.301603 units (statistically significant, at 5% level of significance).


Residuals vs Fitted
Residuals
Fitted values
lm(Purchase_Amount ~ Age + Gender + Location + Season + Promo_Code_Used + S ...
616
53
2263
60
40
20
0
-20
-40
-60
50
55
60
65
70


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Purchase_Amount ~ Age + Gender + Location + Season + Promo_Code_Used + S ...
616
2263
53
2
1
0
-1
-2
-2
0
2


Scale-Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
616
53
2263
50
55
60
65
70
Fitted values
lm(Purchase_Amount ~ Age + Gender + Location + Season + Promo_Code_Used + S ..


Residuals vs Leverage
Standardized residuals
-2
0
1
2
2179
Cook's distance
53
2263
0.000
0.005
0.010
0.015
0.020
Leverage
lm(Purchase_Amount ~ Age + Gender + Location + Season + Promo_Code_Used + S ..

A Monte Carlo simulation study to investigate a probabilistic problem

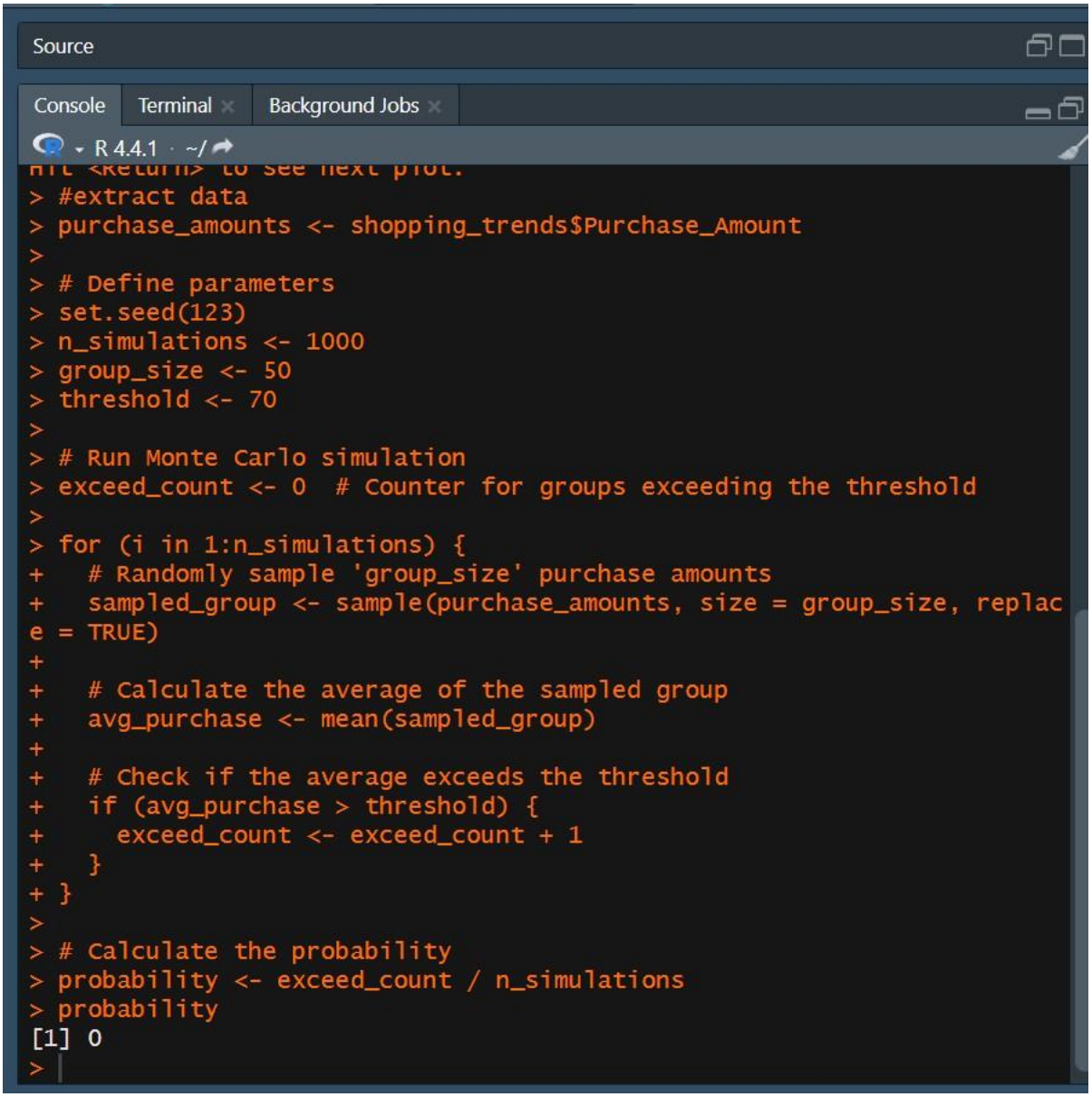
We want to calculate the probability that the average purchase amount in a randomly selected group of customers exceeds $70.

R codes

```
#extract data purchase_amounts <-
shopping_trends$Purchase_Amount

# Define parameters
set.seed(123)
n_simulations <- 1000
group_size <- 50       threshold
<- 70

# Run Monte Carlo simulation exceed_count <- 0  # Counter for
groups exceeding the threshold

for (i in 1:n_simulations) {
  # Randomly sample 'group_size' purchase amounts   sampled_group <-
sample(purchase_amounts, size = group_size, replace = TRUE)

  # Calculate the average of the sampled group
avg_purchase <- mean(sampled_group)

  # Check if the average exceeds the threshold
```

```
  if (avg_purchase > threshold) {
exceed_count <- exceed_count + 1
  }
}

# Calculate the probability probability <-
exceed_count / n_simulations probability
```

Results



Interpreting Results

The value of probability gives the likelihood that the average purchase amount in a randomly selected group exceeds $70.Hence there is no probability of the average purchase amount in a randomly selected group exceeding $70 in 1000 simulations.

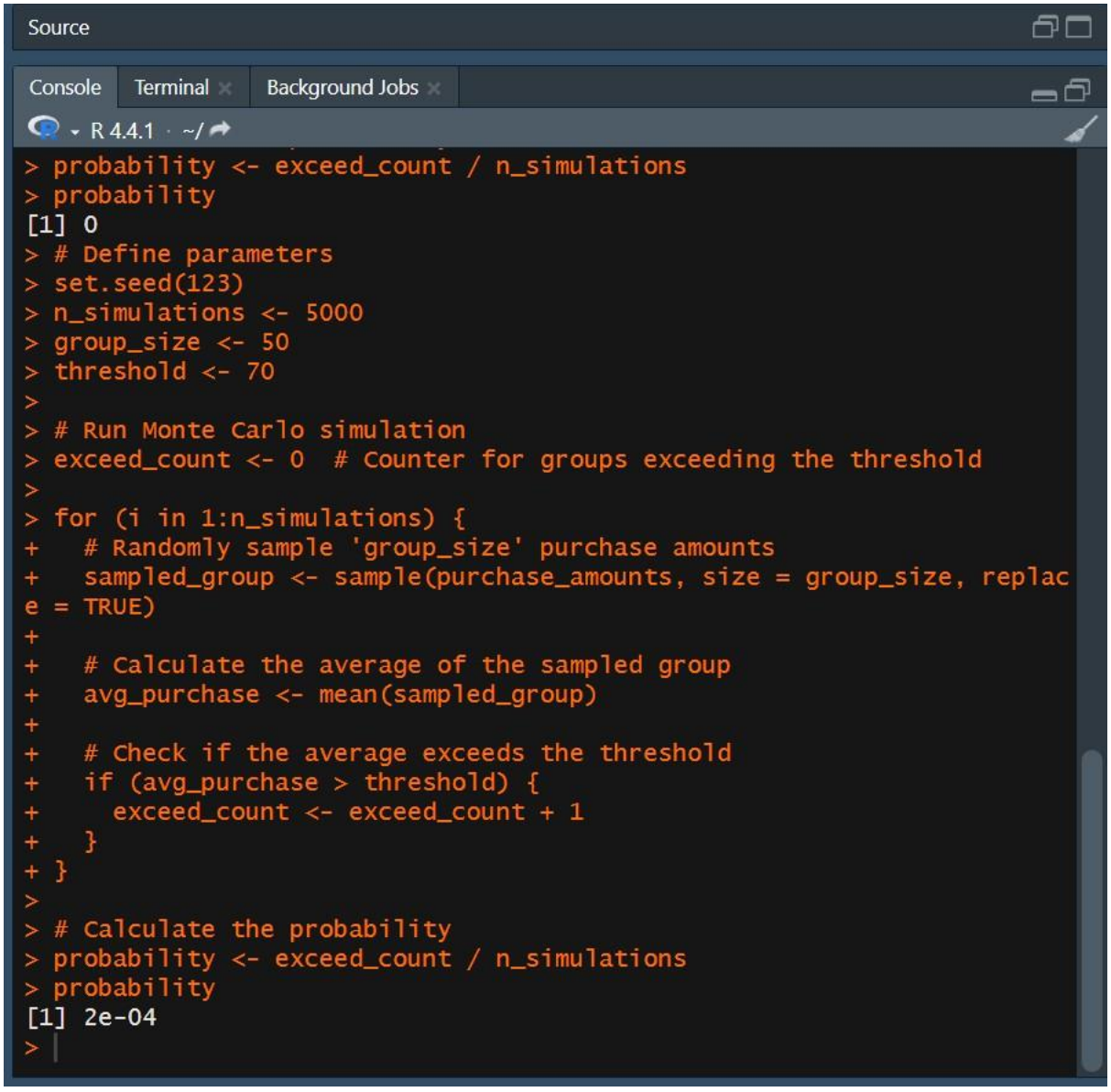
However, generating more simulations eg 5000 gives a higher probability

Codes in R

 # Define parameters

```
set.seed(123)
n_simulations <- 5000
group_size <- 50       threshold
<- 70

# Run Monte Carlo simulation exceed_count <- 0  # Counter for
groups exceeding the threshold

for (i in 1:n_simulations) {
  # Randomly sample 'group_size' purchase amounts   sampled_group <-
sample(purchase_amounts, size = group_size, replace = TRUE)

  # Calculate the average of the sampled group
avg_purchase <- mean(sampled_group)

  # Check if the average exceeds the threshold
if (avg_purchase > threshold) {
exceed_count <- exceed_count + 1
  }
}

# Calculate the probability probability <-
exceed_count / n_simulations probability
```

Results



This implies that there is a 2% probability of the average purchase amount in a randomly selected group exceeding $70 in 5000 simulations.

Conclusion

The higher the simulations the better the chances of average purchase amount exceeding $70.