CUSTOMER SHOPPING DATASET - PRESENTATION

By Third Team

01 - Introduction

02 - History

03 - Smart Questions

04 - Conclusions

Introduction

The Istanbul Shopping Dataset comprises **99458** records from ten malls in Istanbul (2021-2023), featuring invoice numbers, client IDs, demographics, payment methods, and product details with a total of **10** fields. It is a valuable resource for understanding consumer behavior and decision-making in Istanbul's retail scene.

Variables

Invoice number
Customer ID
Gender
Age
Category
Quantity
Price
Payment Type
Invoice Date
Shopping Mall

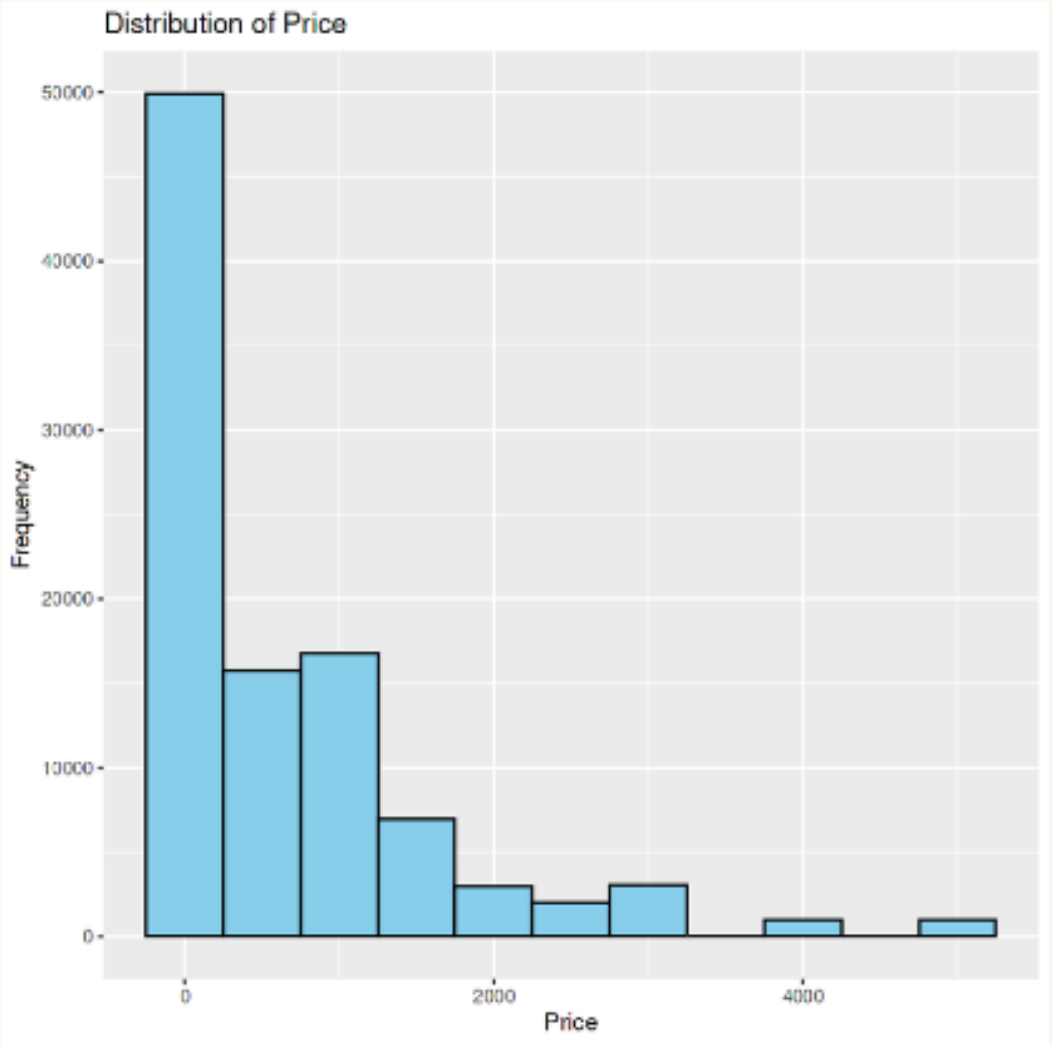
Total Price = *“quantity” * “price”*

The total price is the consequence of multiplying quantity by price.

“0” Null Values

Data Visualization

EDA

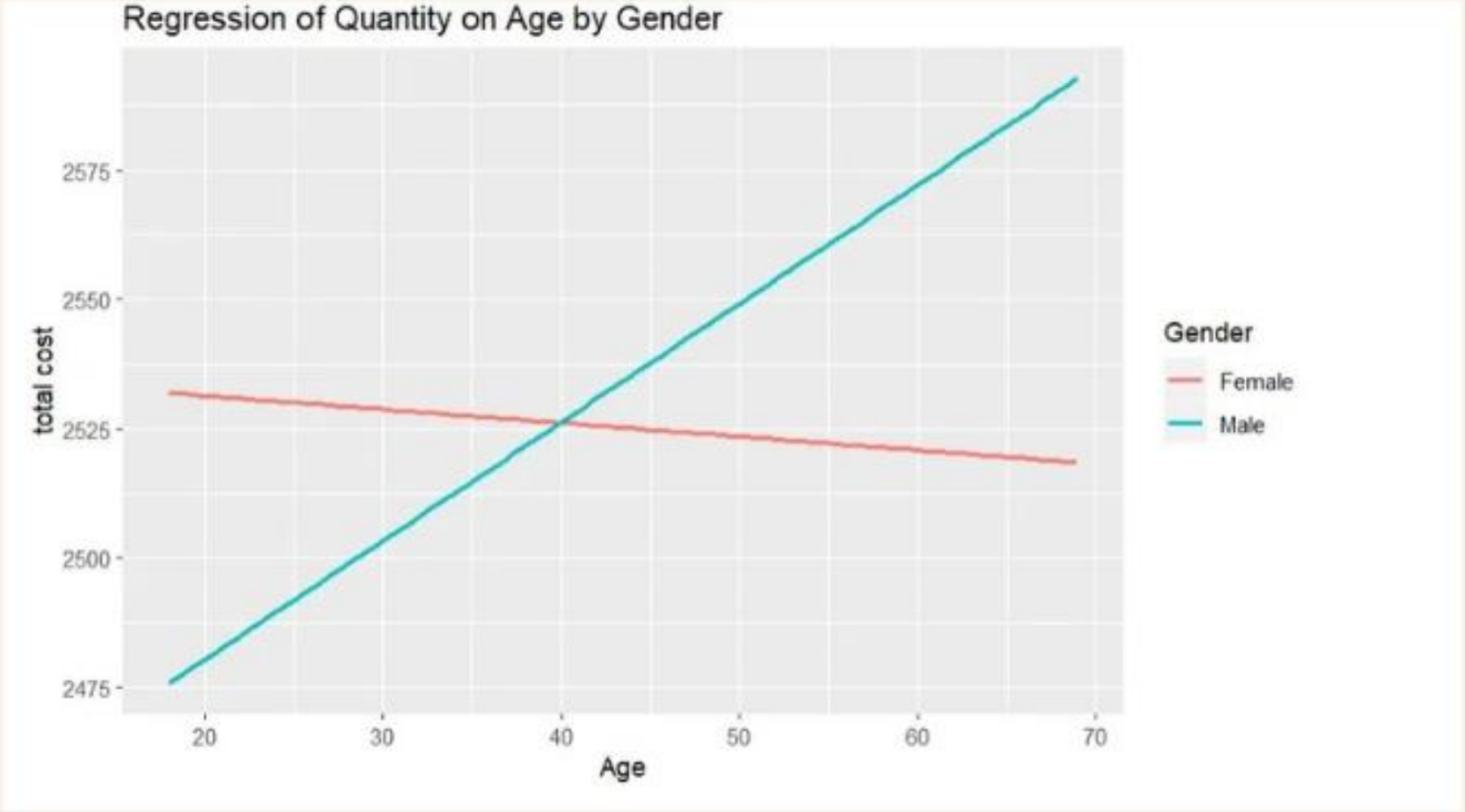


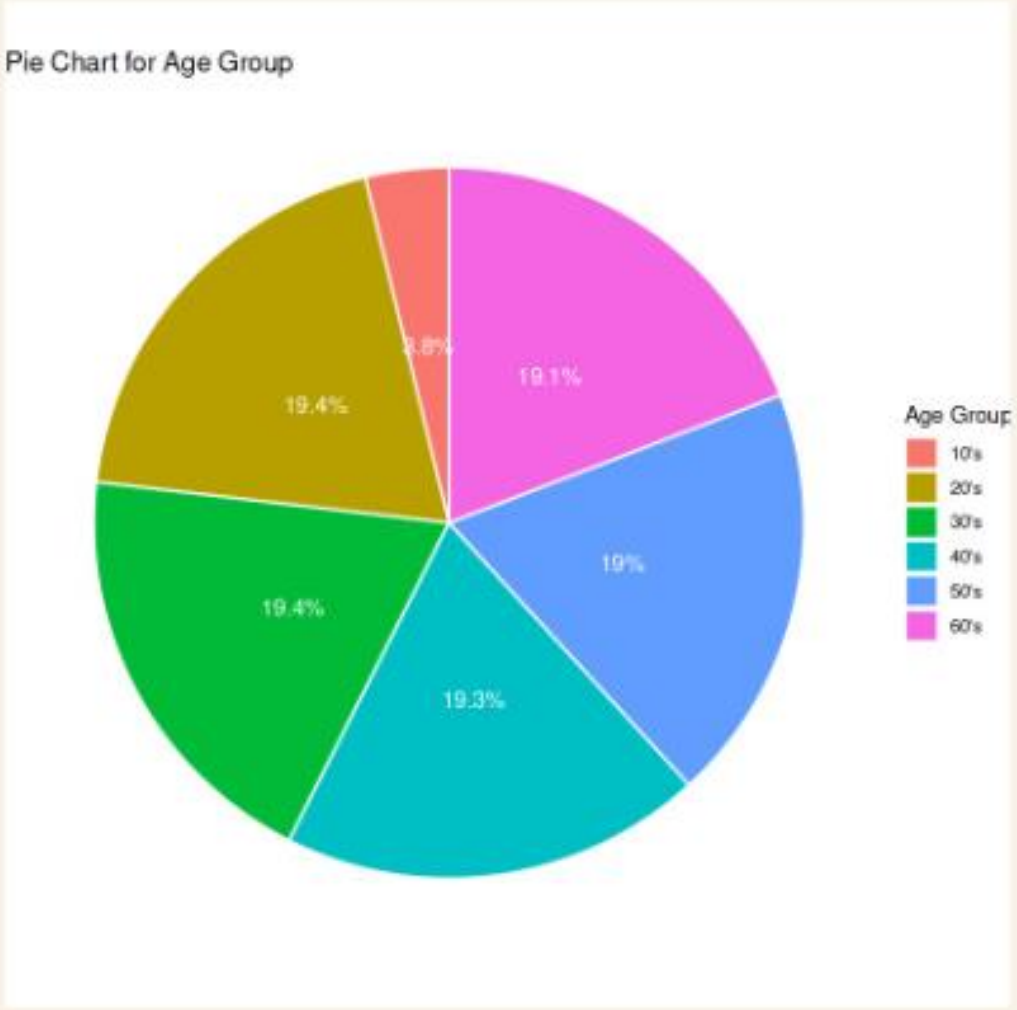
SMART QUESTIONS

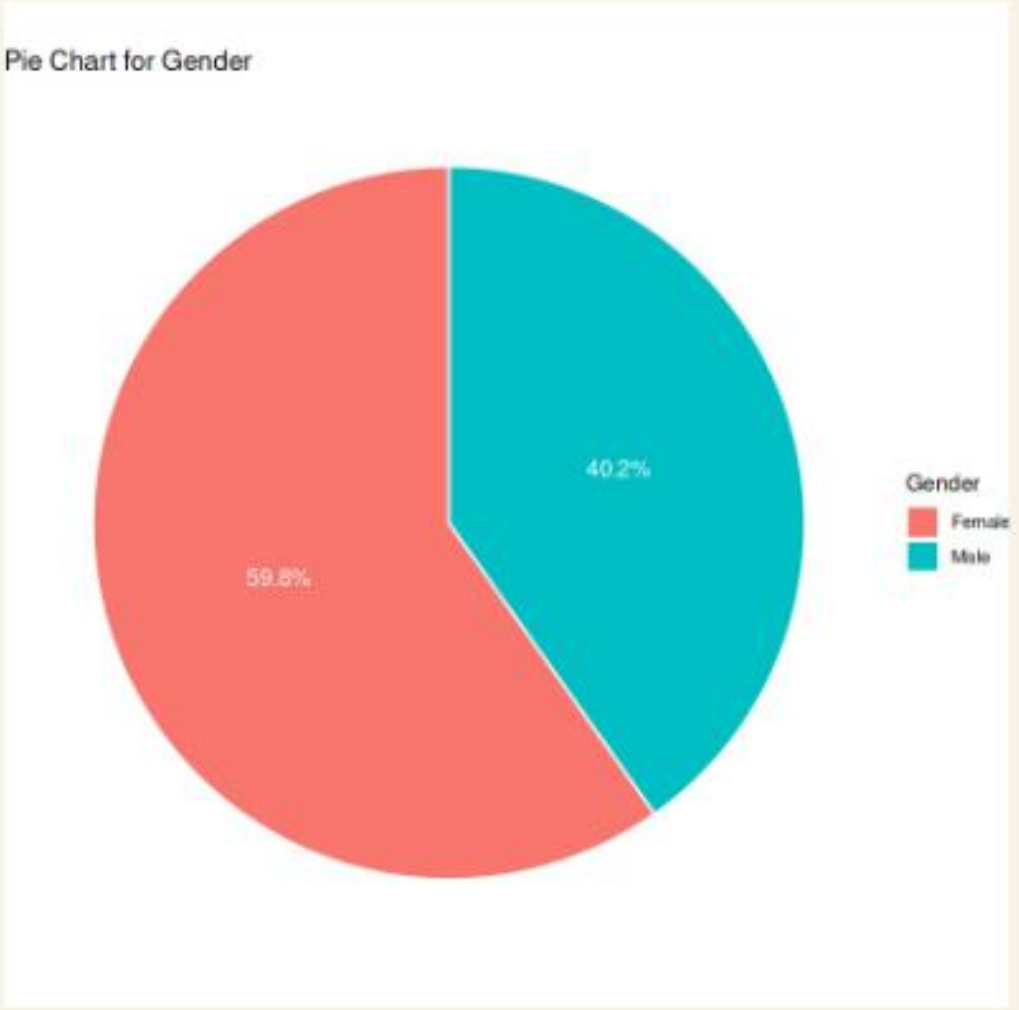
1. Is there a correlation between a customer's age and gender effect on shopping behavior?
2. Does the total purchase amount affect the payment method?
3. Does the shopping mall location affect the purchase frequency?
4. Are certain product categories associated with higher total purchase amounts than others?

Question - 1

Is there a correlation between a customer's age and gender effect on shopping behavior?




Pie Chart for Age Group
19.4%
19.1%
19.4%
19%
19.3%
Age Group
10's
20's
30's
40's
50's
60's


Pie Chart for Gender
40.2%
59.8%
Gender
Female
Male

```
Call:
lm(formula = total_cost ~ age + gender, data = customer_data)

Residuals:
   Min     1Q Median     3Q    Max
 -2549  -2393  -1929    186  23744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2491.798     42.499   58.63   <2e-16 ***
age            0.770      0.893    0.86     0.39
genderMale     8.847     27.309    0.32     0.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4220 on 99454 degrees of freedom
Multiple R-squared:  8.51e-06,  Adjusted R-squared:  -1.16e-05
F-statistic: 0.423 on 2 and 99454 DF,  p-value: 0.655
```

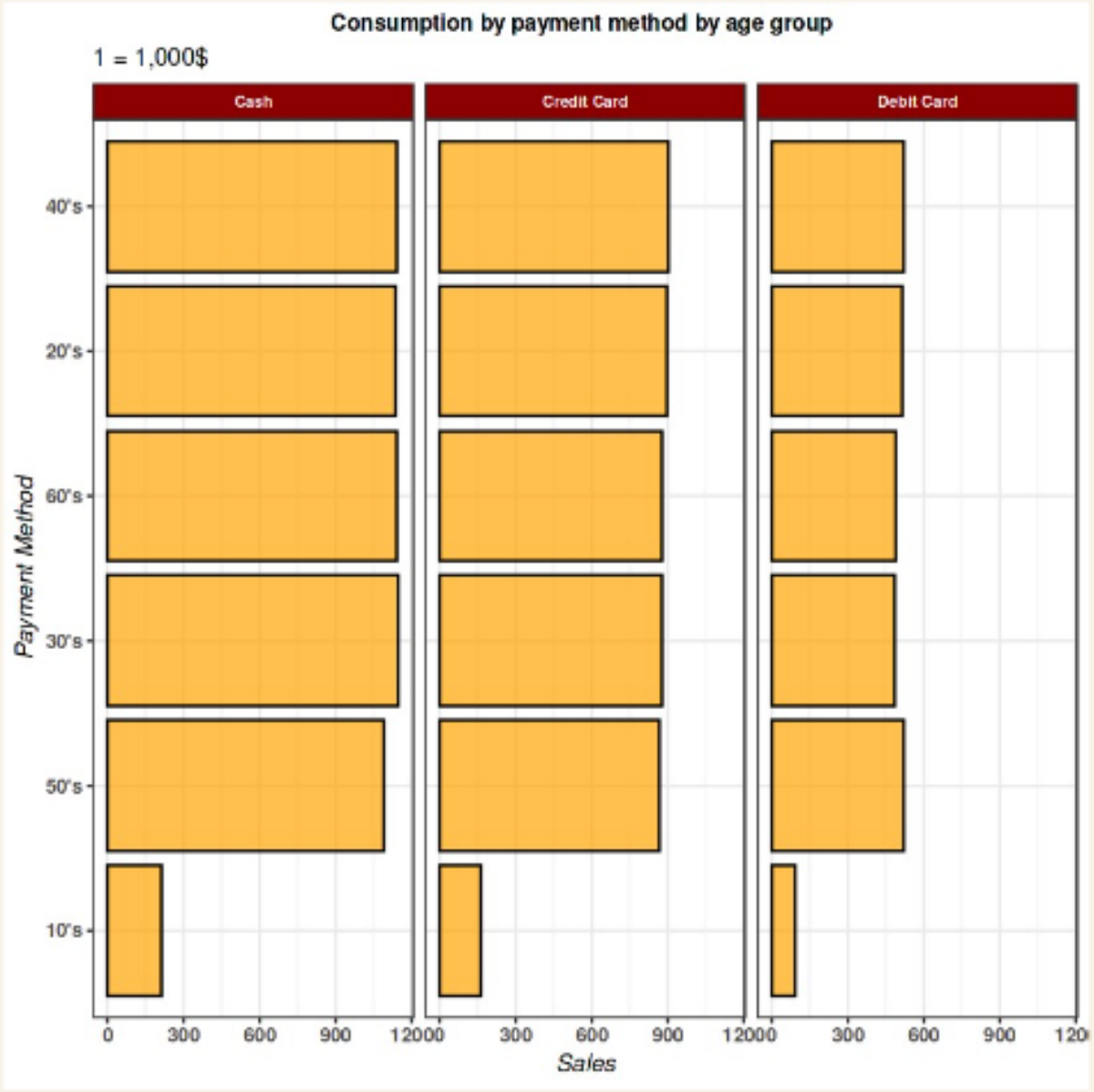
- None of the predictors (age and gender) have a statistically significant impact on total cost. This is because all p-values for the predictors are above common significance levels (e.g., 0.05).
- The model doesn't appear to be a good fit, as the R-squared value is very close to zero, indicating that age and gender together don't explain much of the variability in total cost.
- In summary, the combination of age and gender does not seem to be a significant factor in predicting the total cost, based on the provided model. The model doesn't explain much of the variation in total costs, and neither age nor gender shows a statistically significant impact on total costs.

Conclusion - 1

Since age, gender, and total cost are not statistically significant predictors of total cost, we can conclude from these results that they do not significantly influence shopping behavior in this dataset.

Question - 2

Does the total purchase amount affect the payment method?


Consumption by payment method by age group
1 = 1,000$
Cash
Credit Card
Debit Card
40's
20's
60's
30's
50's
10's
Payment Method
0
300
600
900
1200
Sales

```
Call:
lm(formula = price ~ payment_method, data = customerDF)

Residuals:
   Min     1Q Median     3Q    Max
-685.6 -645.4 -483.7  509.5 4563.0

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                690.823      4.464 154.743   <2e-16 ***
payment_methodCredit Card   -2.281      6.730  -0.339    0.735
payment_methodDebit Card    -3.794      8.003  -0.474    0.635
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 941.2 on 99454 degrees of freedom
Multiple R-squared:  2.572e-06, Adjusted R-squared:  -1.754e-05
F-statistic: 0.1279 on 2 and 99454 DF,  p-value: 0.88
```

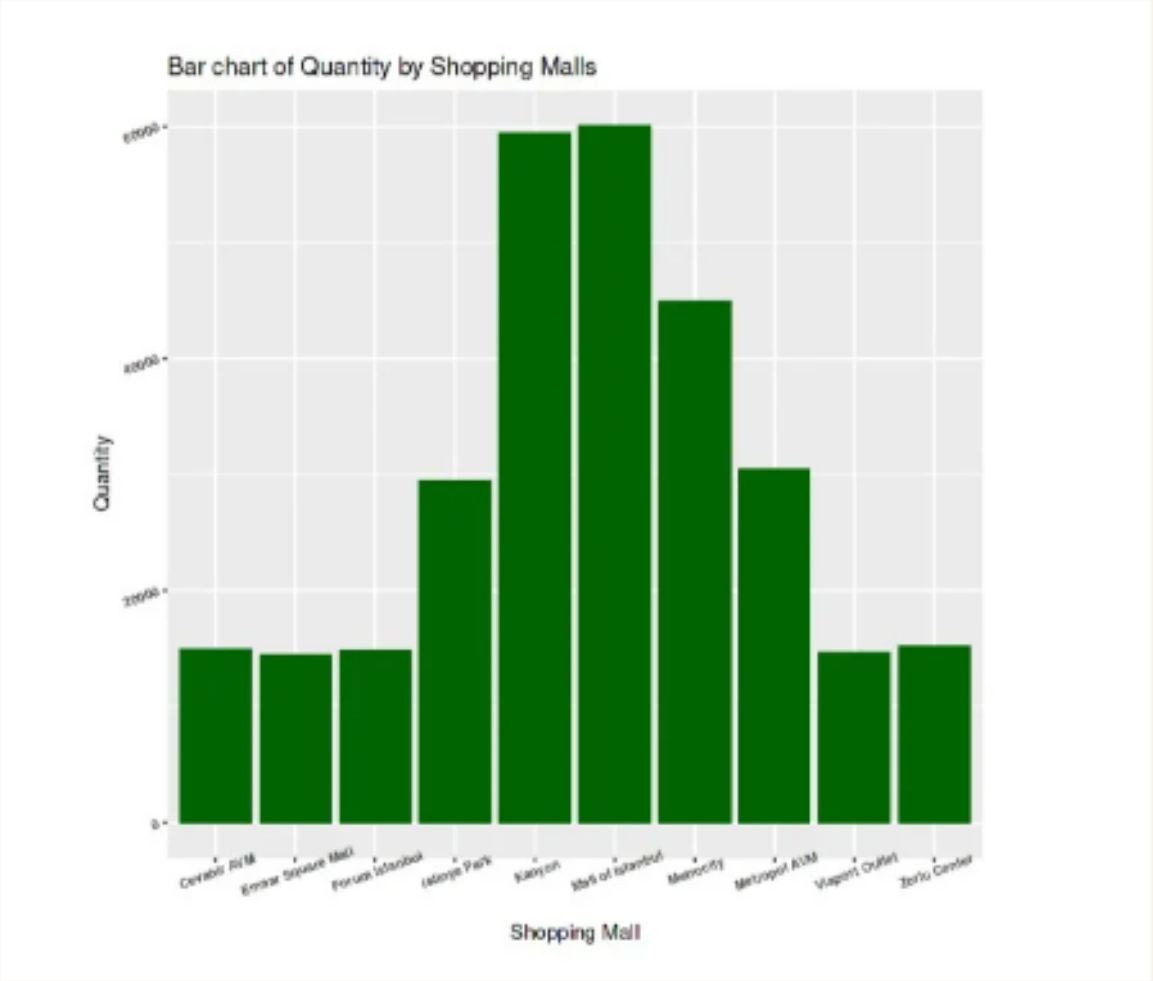
- None of the payment methods have a statistically significant impact on the price. This is because all p-values for the payment methods are above common significance levels (e.g., 0.05).
- The model doesn't seem to be a good fit, as the R-squared value is very close to zero, suggesting that the choice of payment method doesn't explain much of the variability in price.
- In summary, the choice of payment method does not seem to be a significant factor in predicting the price of items, based on the provided model. The model doesn't explain much of the variation in prices, and neither credit card nor debit card payments show a statistically significant impact on prices.

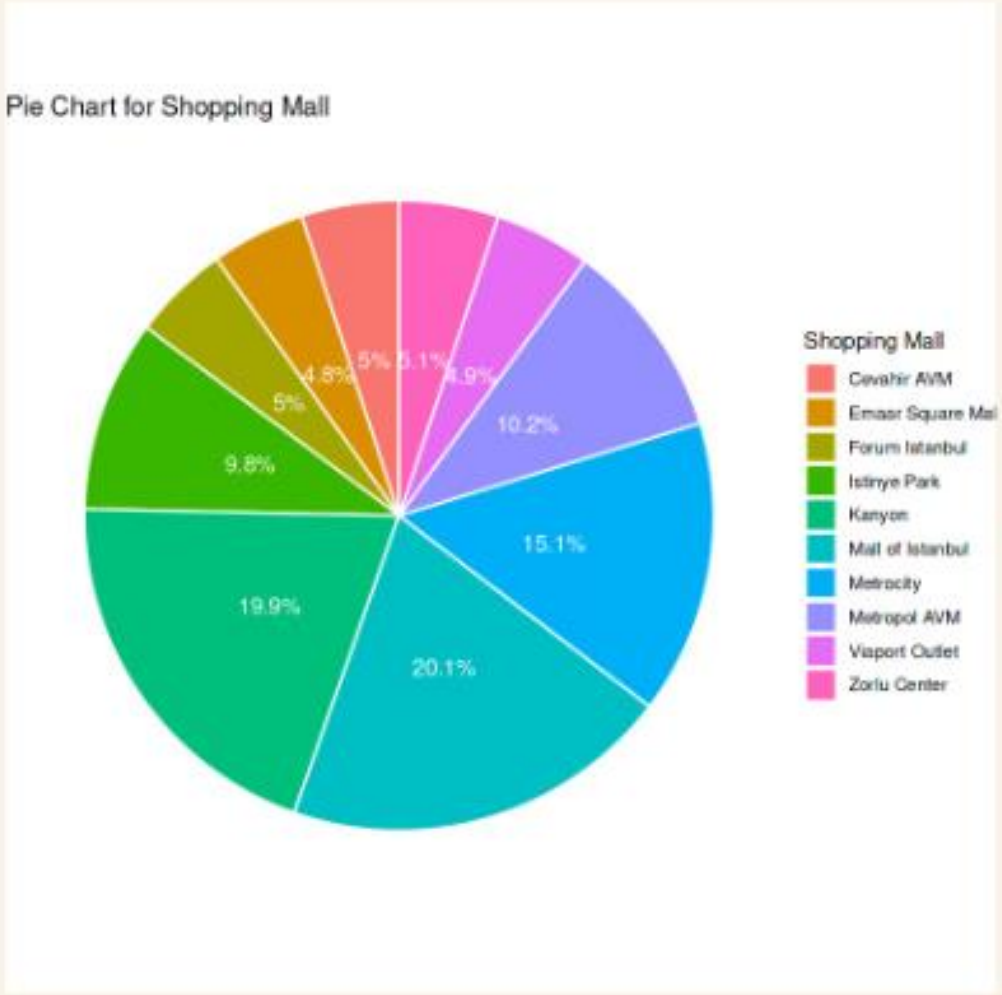
Conclusion - 2

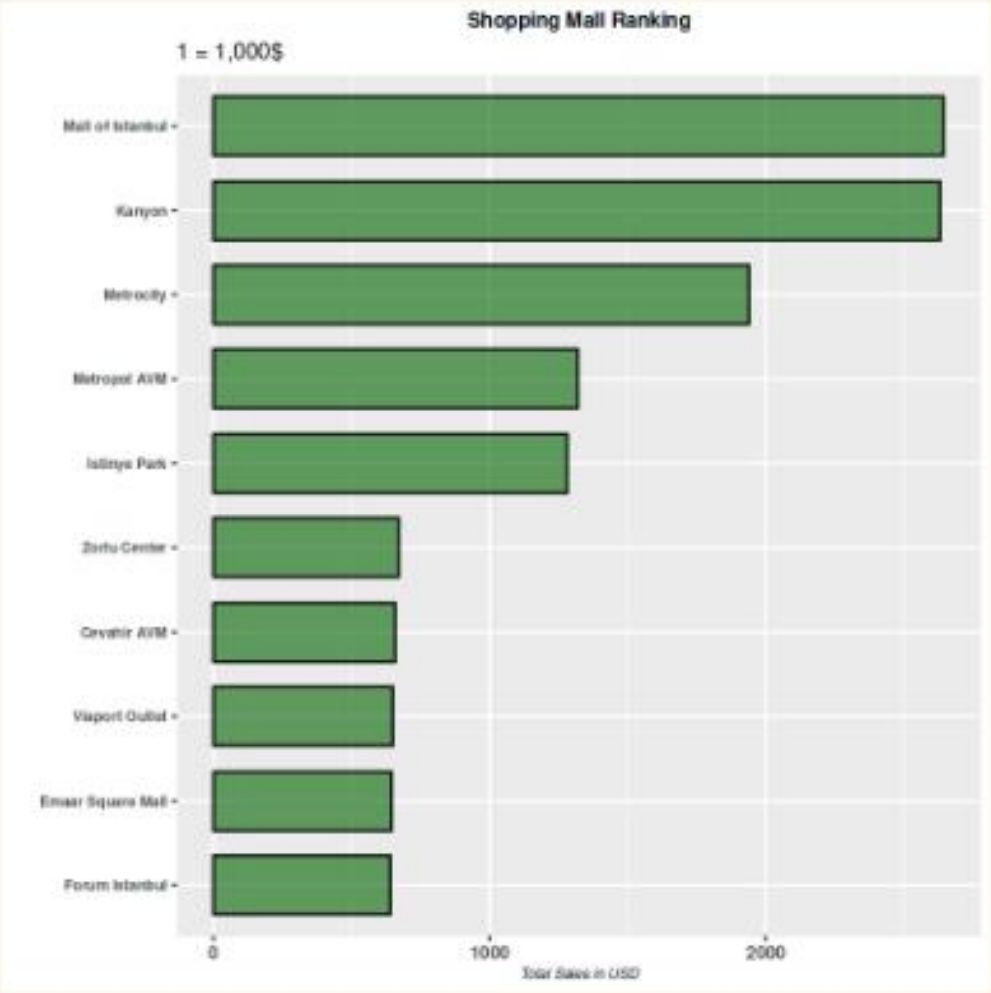
The analysis indicates that the Payment Method (Credit Card or Debit Card) is not a statistically significant predictor and has no discernible or significant effect on Purchase Amount in this dataset.

Question - 3

Does the shopping mall location affect the purchase frequency?




Pie Chart for Shopping Mall
Shopping Mall
Cevahir AVM
Emaar Square Mal
Forum Istanbul
Istinye Park
Kanyon
Mall of Istanbul
Metrocity
Metropol AVM
Viaport Outlet
Zorlu Center
10.2%
15.1%
20.1%
19.9%
9.8%
5%
4.8%
5%
5.1%
4.9%


Shopping Mall Ranking
1 = 1,000$
Mall of Istanbul
Kanyon
Metrocity
Metropol AVM
Istinye Park
Zorlu Center
Cevahir AVM
Viaport Outlet
Emaar Square Mall
Forum Istanbul
0
1000
2000
Total Sales in USD

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.01429 -1.01247 -0.00177  1.00481  2.00926

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.9951913  0.0200017 149.747   <2e-16 ***
shopping_mallEmaar Square Mall 0.0189429  0.0285501   0.663    0.507
shopping_mallForum Istanbul    0.0070322  0.0283496   0.248    0.804
shopping_mallIstinye Park      0.0172818  0.0245808   0.703    0.482
shopping_mallKanyon            0.0042033  0.0223785   0.188    0.851
shopping_mallMall of Istanbul  0.0190994  0.0223650   0.854    0.393
shopping_mallMetrocity        -0.0044512  0.0230887  -0.193    0.847
shopping_mallMetropol AVM      0.0094342  0.0244250   0.386    0.699
shopping_mallViaport Outlet   -0.0004823  0.0283973  -0.017    0.986
shopping_mallZorlu Center      0.0065821  0.0281694   0.234    0.815
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.413 on 99447 degrees of freedom
Multiple R-squared:  3.622e-05, Adjusted R-squared:  -5.428e-05
F-statistic: 0.4002 on 9 and 99447 DF,  p-value: 0.9356
```

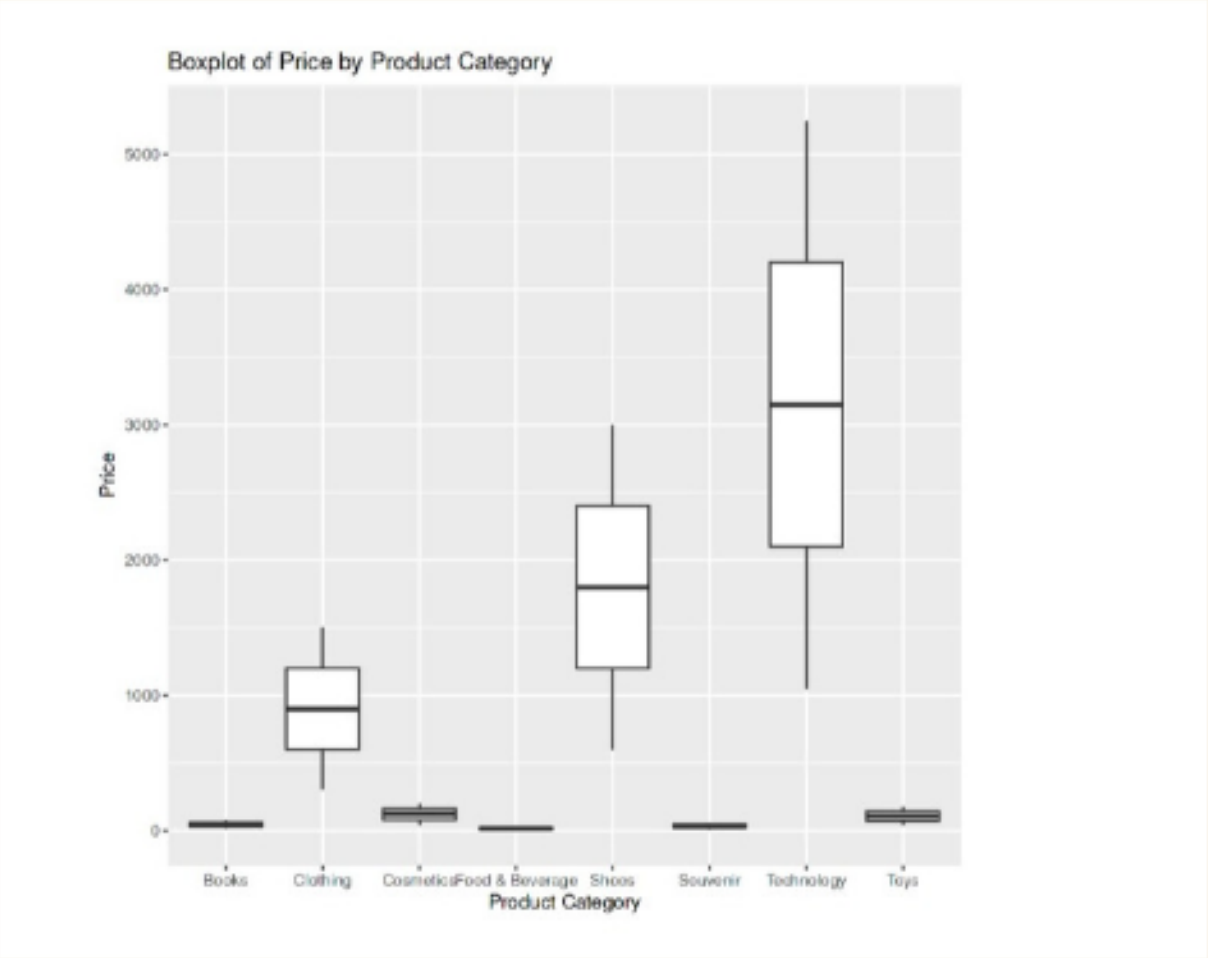
- None of the shopping malls have a statistically significant impact on the quantity of items bought. This is because all p-values for the shopping malls are above the common significance levels (e.g., 0.05).
- The model doesn't seem to be a great fit since the R-squared value is very close to zero, suggesting that the choice of shopping mall doesn't explain much of the variability in quantity.
- In summary, the shopping mall choice does not seem to be a significant factor in predicting the quantity of items bought, based on the provided model.

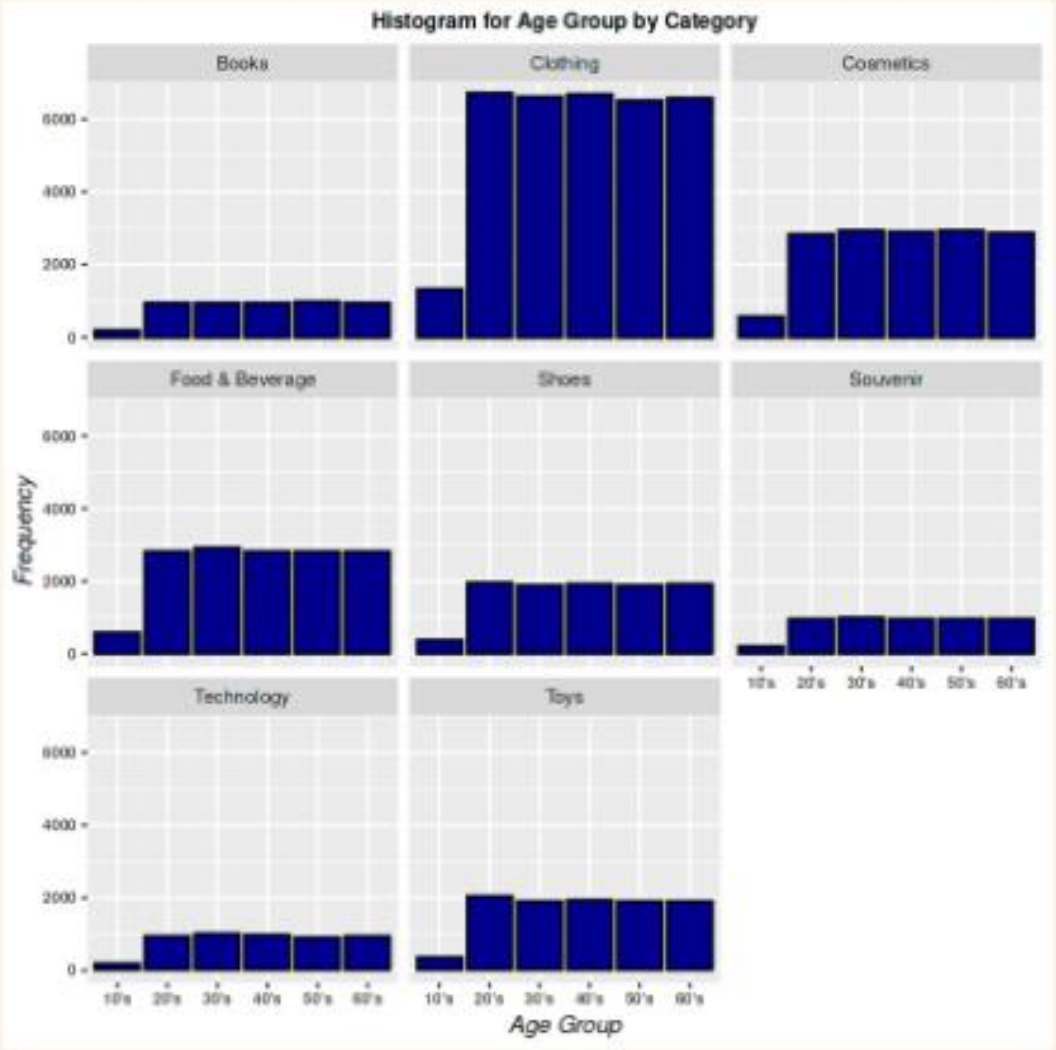
Conclusion - 3

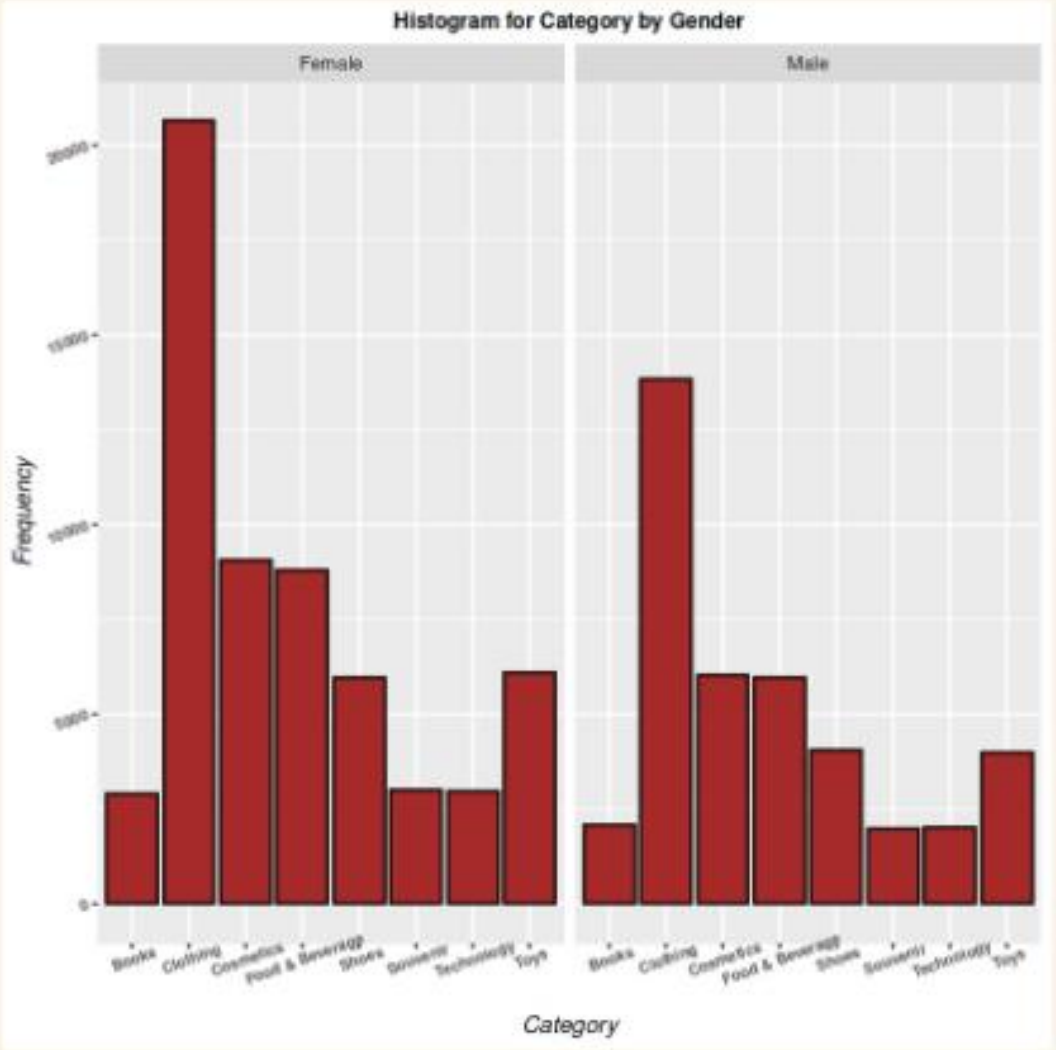
The regression analysis indicates no statistically significant relationship between shopping mall location and quantity purchased. It's crucial to recognize that a bar plot illustrates data distribution and doesn't establish causality or relationships between variables.

Question - 4

Are certain product categories associated with higher total purchase amounts than others?




Histogram for Age Group by Category
Books
Clothing
Cosmetics
Food & Beverage
Shoes
Souvenir
Technology
Toys
Frequency
0
2000
4000
6000
10's
20's
30's
40's
50's
60's
Age Group


Histogram for Category by Gender
Female
Male
Frequency
0
5000
10000
15000
Books
Clothing
Cosmetics
Food & Beverage
Shoes
Souvenir
Technology
Toys
Category

```
lm(formula = price ~ category, data = customerDF)

Residuals:
     Min       1Q   Median       3Q      Max
-2106.94   -41.13    -0.47    40.19  2093.06

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              45.569      7.021   6.491 8.59e-11 ***
categoryClothing        855.515      7.511 113.908  < 2e-16 ***
categoryCosmetics        76.880      8.096   9.496  < 2e-16 ***
categoryFood & Beverage -29.897      8.118  -3.683 0.000231 ***
categoryShoes          1761.820      8.588 205.142  < 2e-16 ***
categorySouvenir        -10.674      9.920  -1.076 0.281905
categoryTechnology     3111.367      9.921 313.604  < 2e-16 ***
categoryToys             62.165      8.581   7.245 4.37e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 495.5 on 99449 degrees of freedom
Multiple R-squared:  0.7229,    Adjusted R-squared:  0.7228
F-statistic: 3.706e+04 on 7 and 99449 DF,  p-value: < 2.2e-16
```

- The model is pretty good at predicting prices based on item categories, with about 72% of price variations explained by the chosen categories.
- Most categories significantly affect prices except for "Souvenir." This means the type of item you're looking at tends to influence its price.
- The differences between predicted and actual prices follow a normal pattern, suggesting the model is doing a good job.

Conclusion - 4

Product category significantly influences purchase amount, with technology, shoes, and clothing having positive effects, while food and beverage negatively impact purchases. Souvenirs show no significant impact on purchase amount.

CONCLUSIONS

After a comprehensive analysis, it seems that Age and Gender don't greatly affect how much individuals spend on items, and product category significantly affects purchase amount, there's no apparent correlation between these criteria and the amount paid.

- The Elements of Statistical Learning
- Categorical Data Analysis
- Kaggle

REFERENCES

Thank You