ACADEMIC YEAR 2025/2026

DATA ANALYSIS

ANALYSIS OF A RETAIL SALES DATASET

THE DATASET

sale_id	branch	city	customer_ type	gender	product_ name	product_ category	unit_price	quantity	tax	total_price	reward_ points
1	A	New York	Member	Male	Shampoo	Personal Care	5.5	3	1.16	17.66	1
2	B	Los Angeles	Normal	Female	Notebook	Stationery	2.75	10	1.93	29.43	0
3	A	New York	Member	Female	Apple	Fruits	1.2	15	1.26	19.26	1
4	A	Chicago	Normal	Male	Detergent	Household	7.8	5	2.73	41.73	0
5	B	Los Angeles	Member	Female	Orange Juice	Beverages	3.5	7	1.72	26.22	2
6	A	Chicago	Normal	Male	Shampoo	Stationery	11.24	9	7.08	108.24	0

1,000 observations Numerical variable Categorical variable

EXPLORATORY DATA ANALYSIS

01 **Data quality and preparation**

02 Compute **summary statistic**

KPIs

- **Total Revenue**: €118,584.00
- **Units sold**: 10.337
- **Average unit per sale**: 10.3
- **Reward points**: 6,057

EXPLORATORY DATA ANALYSIS

03

Insights to answer **3 questions:**

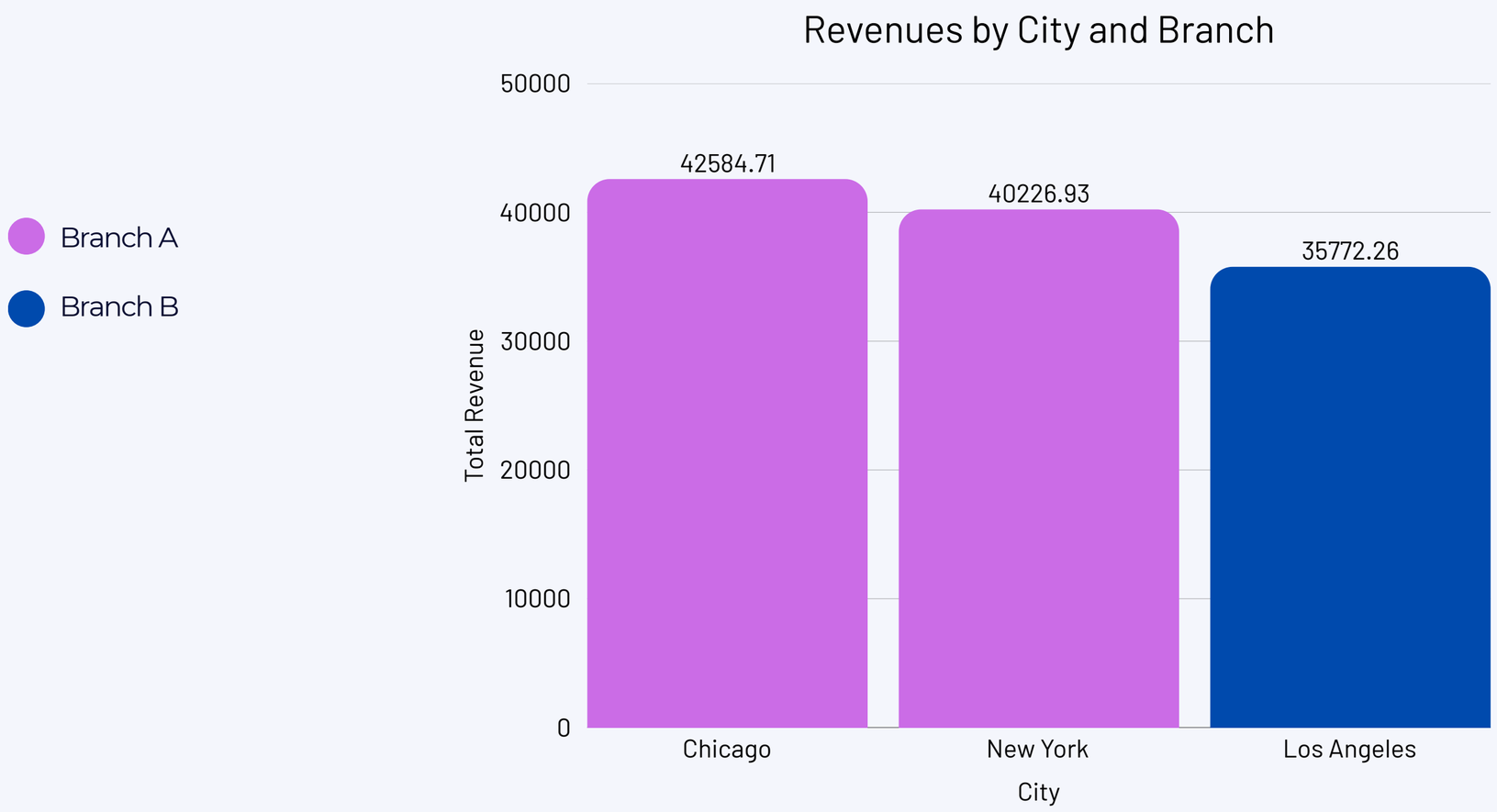
Where are the products sell best?

Who are the customers?

Which are the best-selling categories and what is their value?

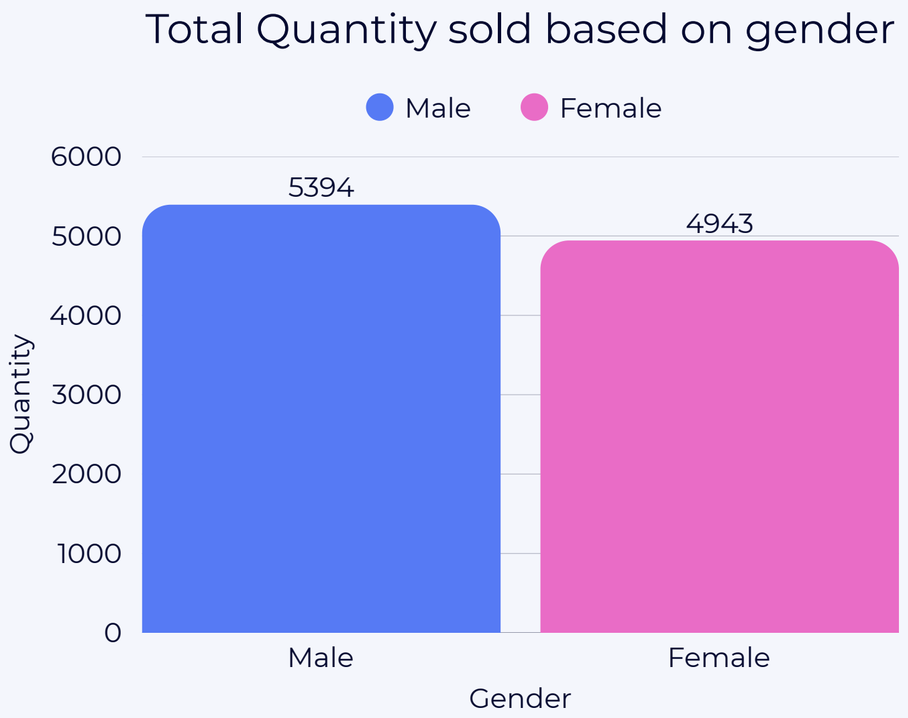
Where are the products sell best?

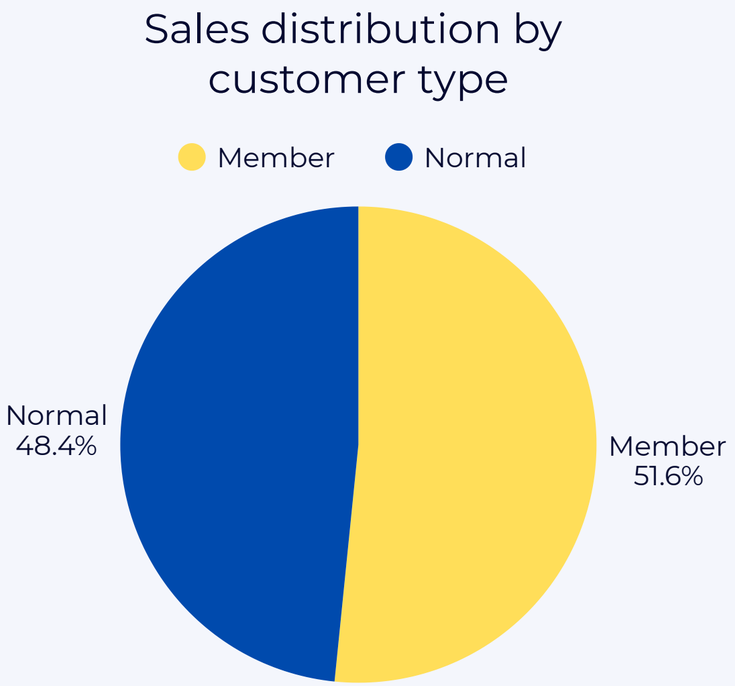
GEOGRAPHIC DISTRIBUTION



Who are the customers?

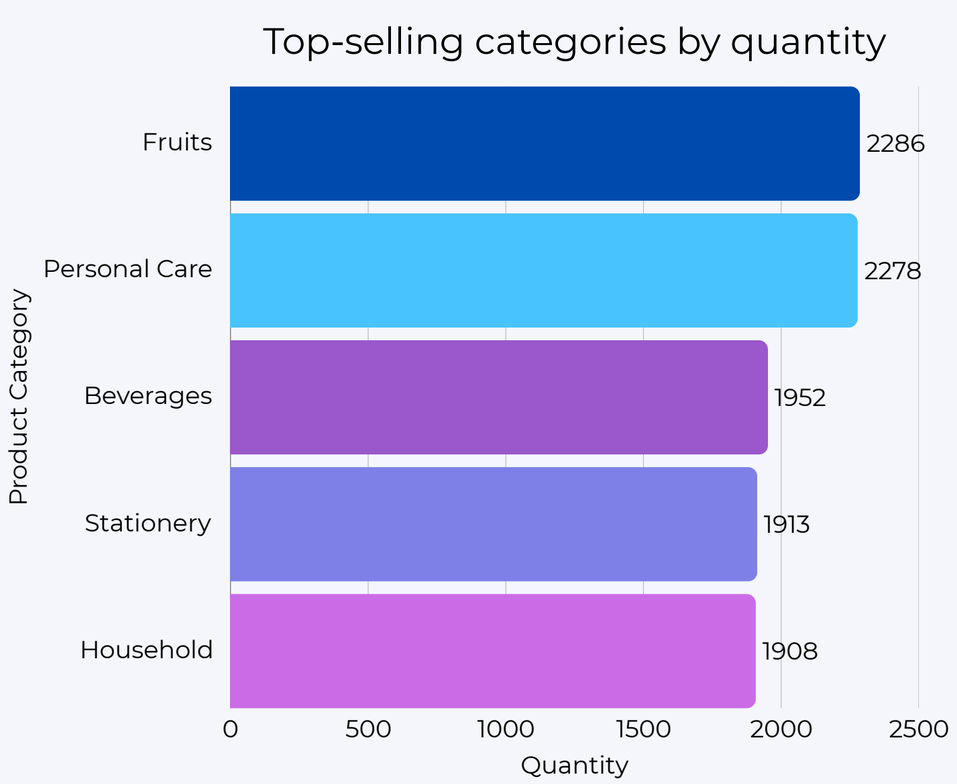
CUSTOMER DEMOGRAPHY AND LOYALTY PROGRAM

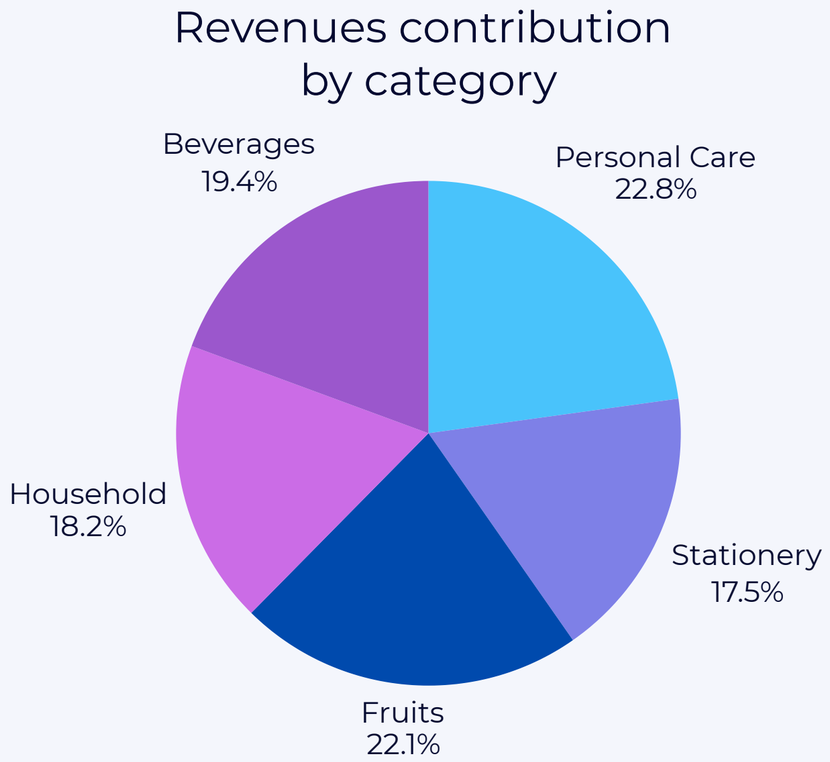




Which are the best-sellers and what is their value?

CATEGORY PERFORMANCE

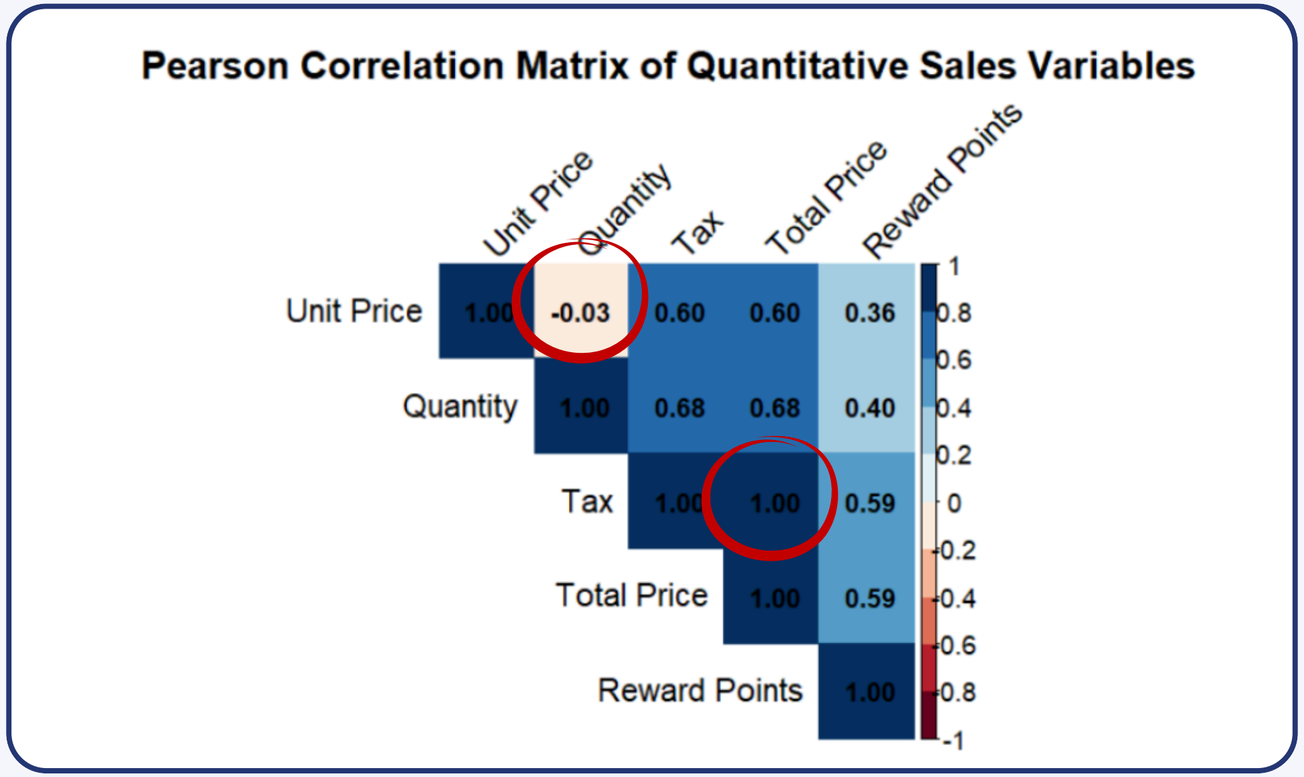




EXPLORATORY DATA ANALYSIS

04

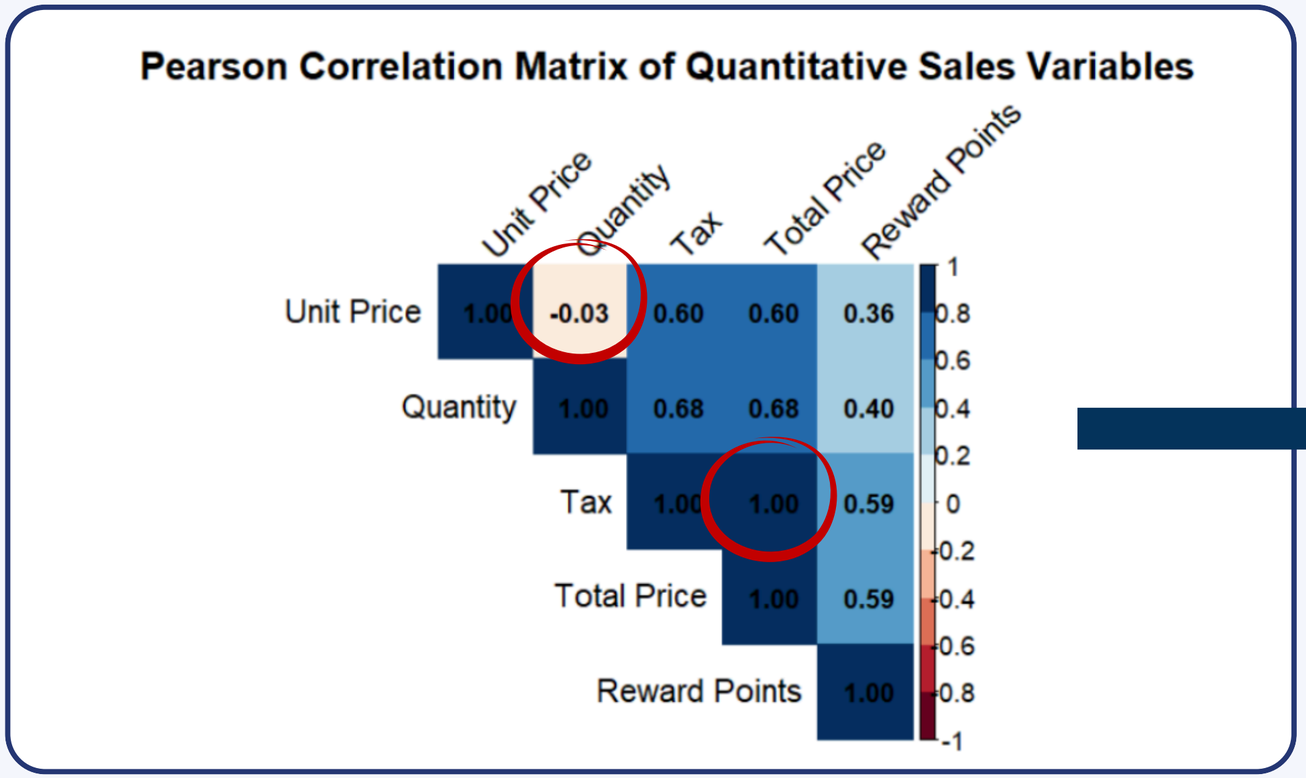
Correlation Analysis on numerical variables

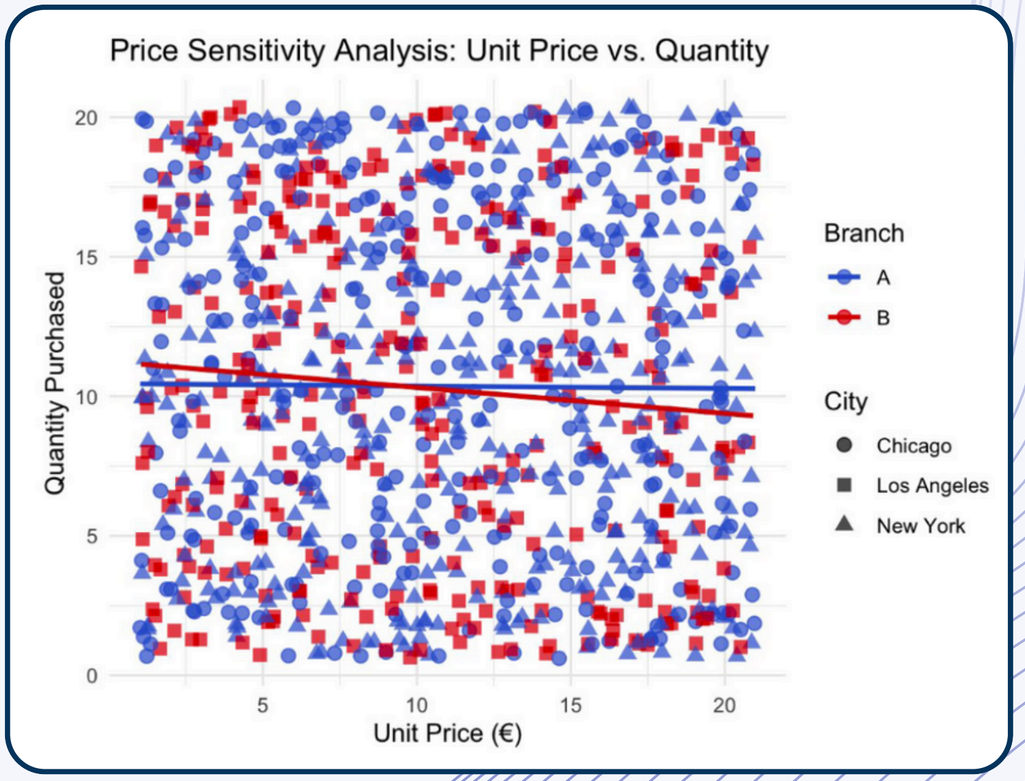


EXPLORATORY DATA ANALYSIS

04

Correlation Analysis on numerical variables





CLUSTERING

unsupervised learning technique

AIM: organize data into groups (*clusters*) based on similarity

Hierarchical Clustering

K-Means Clustering

01 Select **numerical attributes**: *unit_price*, *total_price*, *quantity*, *tax* and *reward_points*

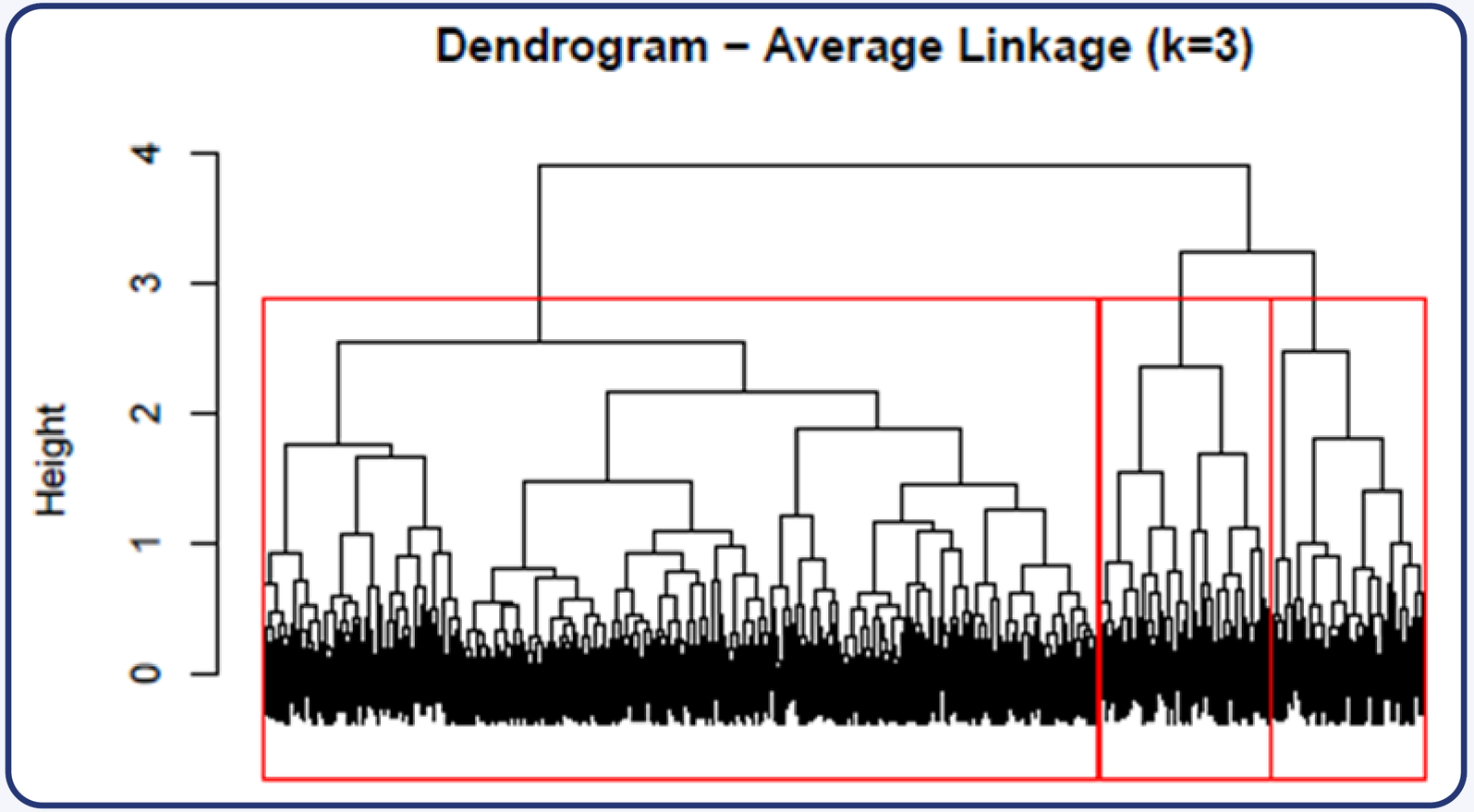
02 **Standardize**: *scale()* function

03 **Distance metric**: *Euclidean*

HIERARCHICAL CLUSTERING

average linkage

cutree() function → k = 3



Observations

- Cluster 1: 718
- Cluster 2: 134
- Cluster 3: 148

SCATTERPLOT quantity vs. total price

city	1	2	3
Chicago	228	47	55
Los Angeles	243	43	40
New York	247	44	53

gender	1	2	3
Female	350	54	68
Male	368	80	80

K-MEANS CLUSTERING

```
data_km <- kmeans(data_scaled, 3, nstart = 50)
```

number of clusters defined *a priori* → $k = 3$

Observations

- Cluster 1: 256
- Cluster 2: 309
- Cluster 3: 435

The clustering quality is evaluated through **variance decomposition**

```
Within cluster sum of squares:
896.85, 442.90, 707.92
```

Each within-cluster sum of squares measures the **compactness of a cluster**

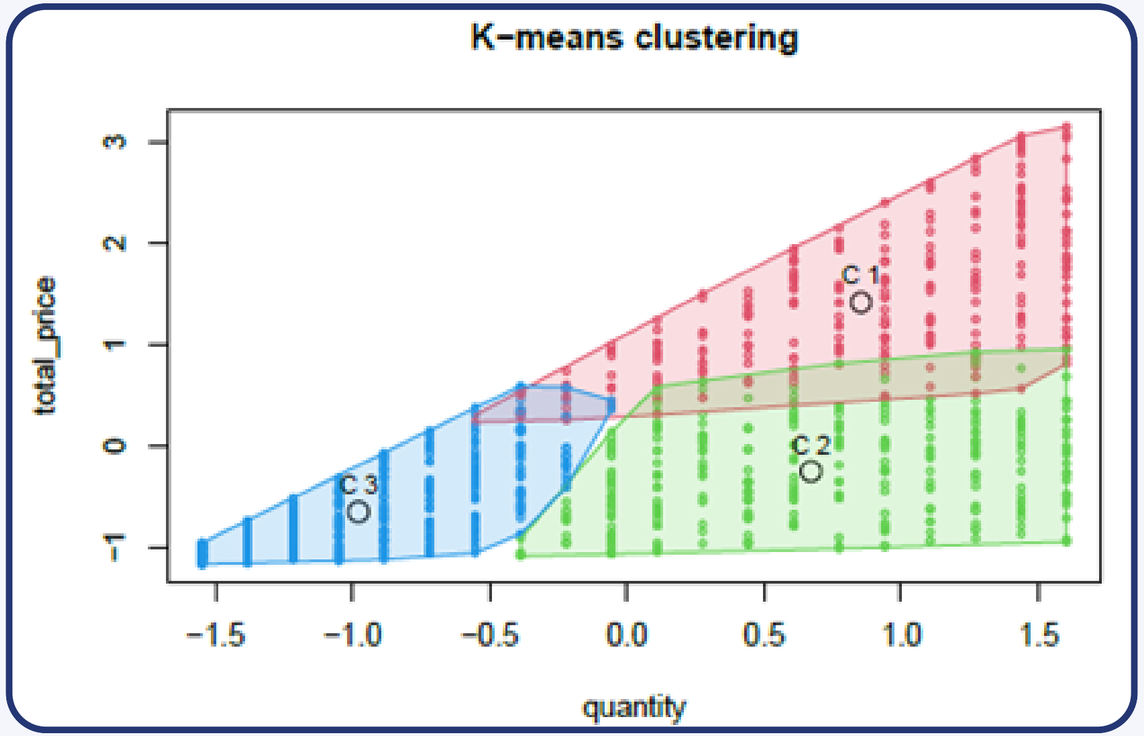
```
Between_SS / Total_SS = 59.0%
```

The **between-cluster variance**

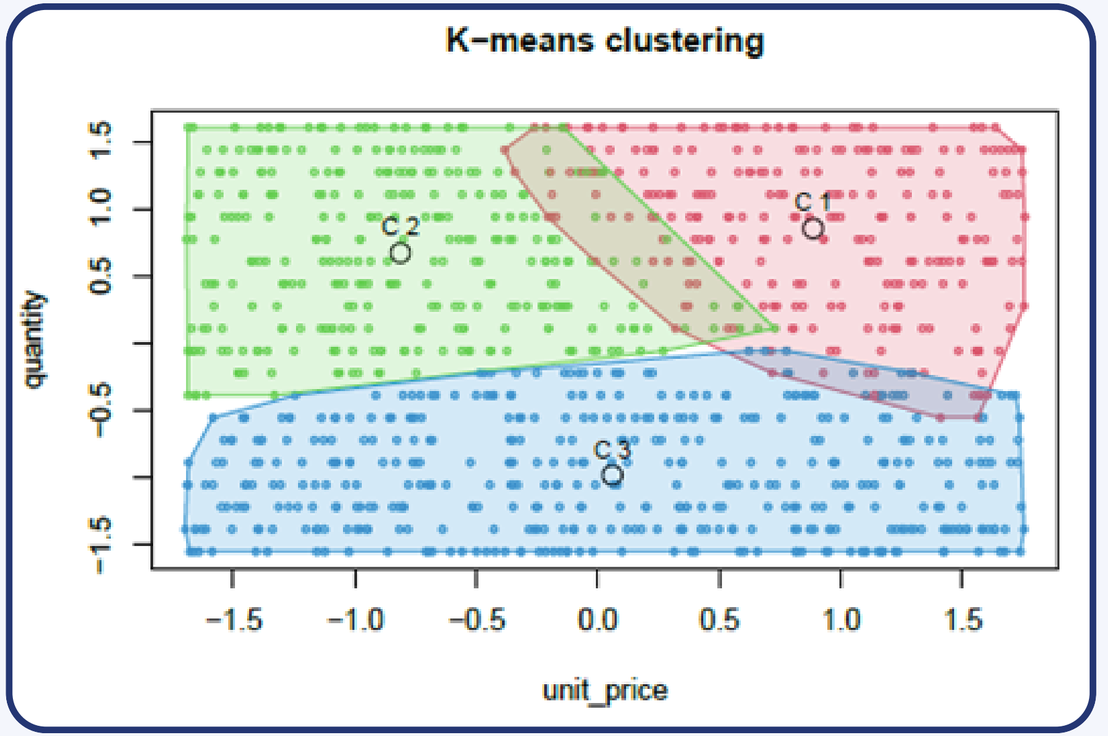
HULL PLOT

	unit_price	quantity	tax	total_price	reward_points
Cluster 1	0.884	0.853	1.418	1.418	1.012
Cluster 2	-0.814	0.671	-0.254	-0.254	-0.249
Cluster 3	0.058	-0.978	-0.654	-0.654	-0.419

quantity vs. total price



quantity vs. unit price



PRINCIPAL COMPONENT ANALYSIS

AIM: reduce dimensionality of the dataset while preserving as much as possible of the original variance of the data

01

Select variables

Extract numerical attributes only: *unit_price*, *total_price*, *quantity*, *tax* and *reward_points*

02

Standardize

Apply *scale()* function

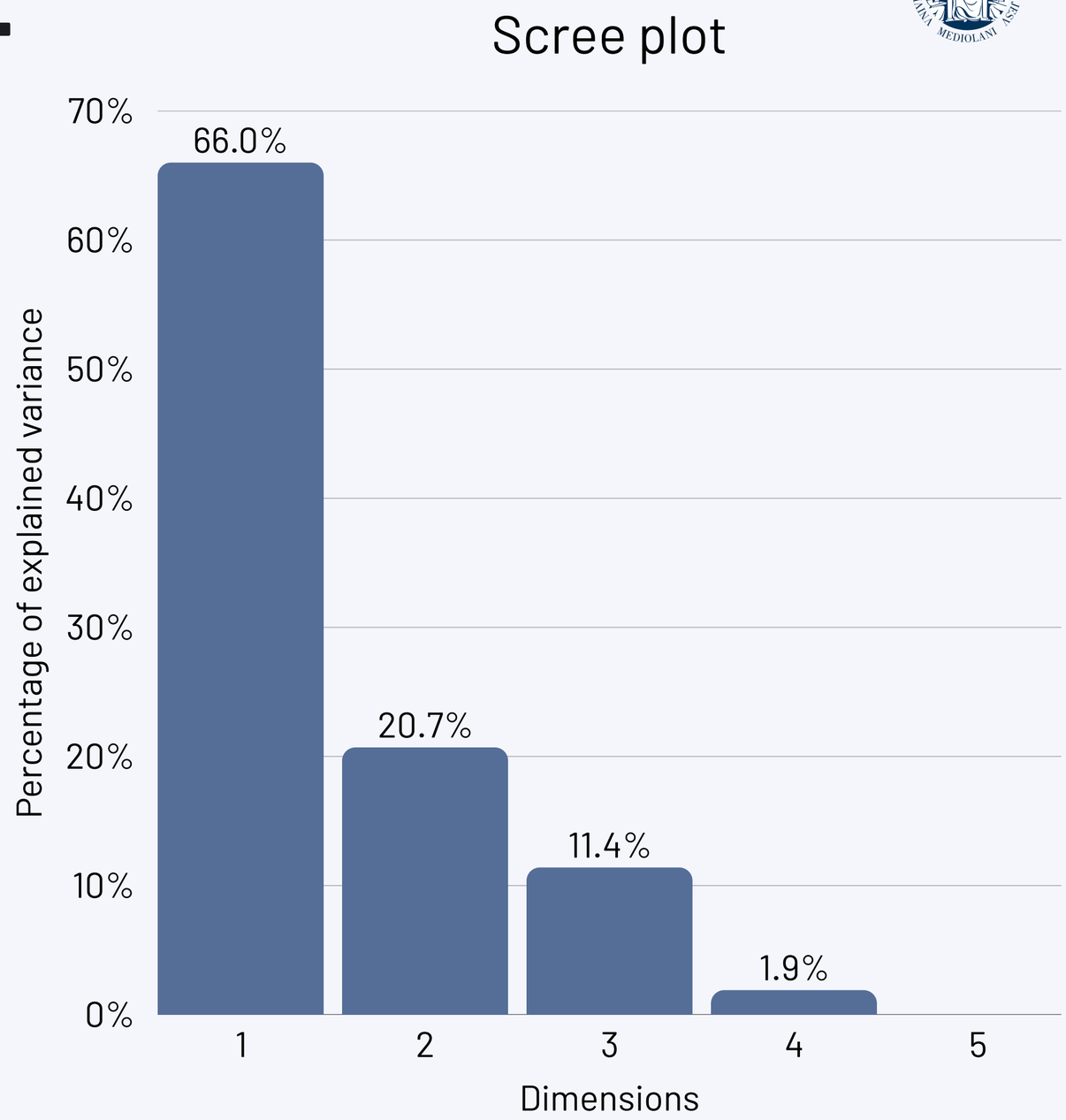
03

Run PCA

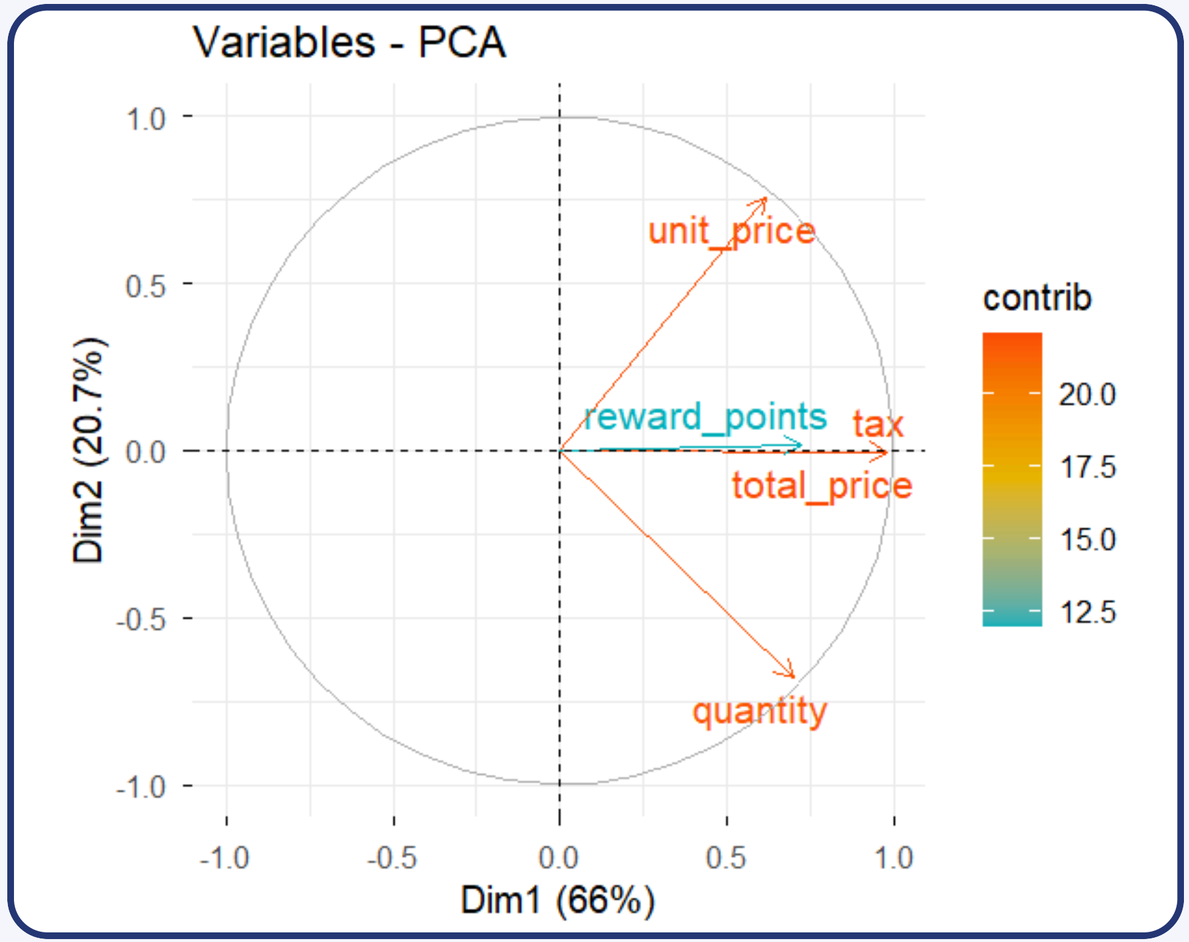
Execute *prcomp()* function

EXPLAINED VARIANCE

	eigenvalue	variance	cumulative variance
Dim.1	3.301643e+00	6.603287e+01	66.03287
Dim.2	1.034011e+00	2.068021e+01	86.71308
Dim.3	5.713521e-01	1.142704e+01	**98.14012**
Dim.4	9.299369e-02	1.859874e+00	100.00000
Dim.5	8.702953e-08	1.740591e-06	100.00000



PCA VARIABLE BIPLOT ANALYSIS



Dimension 1 - 66%

total_price - 29%

tax - 29%

reward_points - 15.7%

Dimension 2 - 20.7%

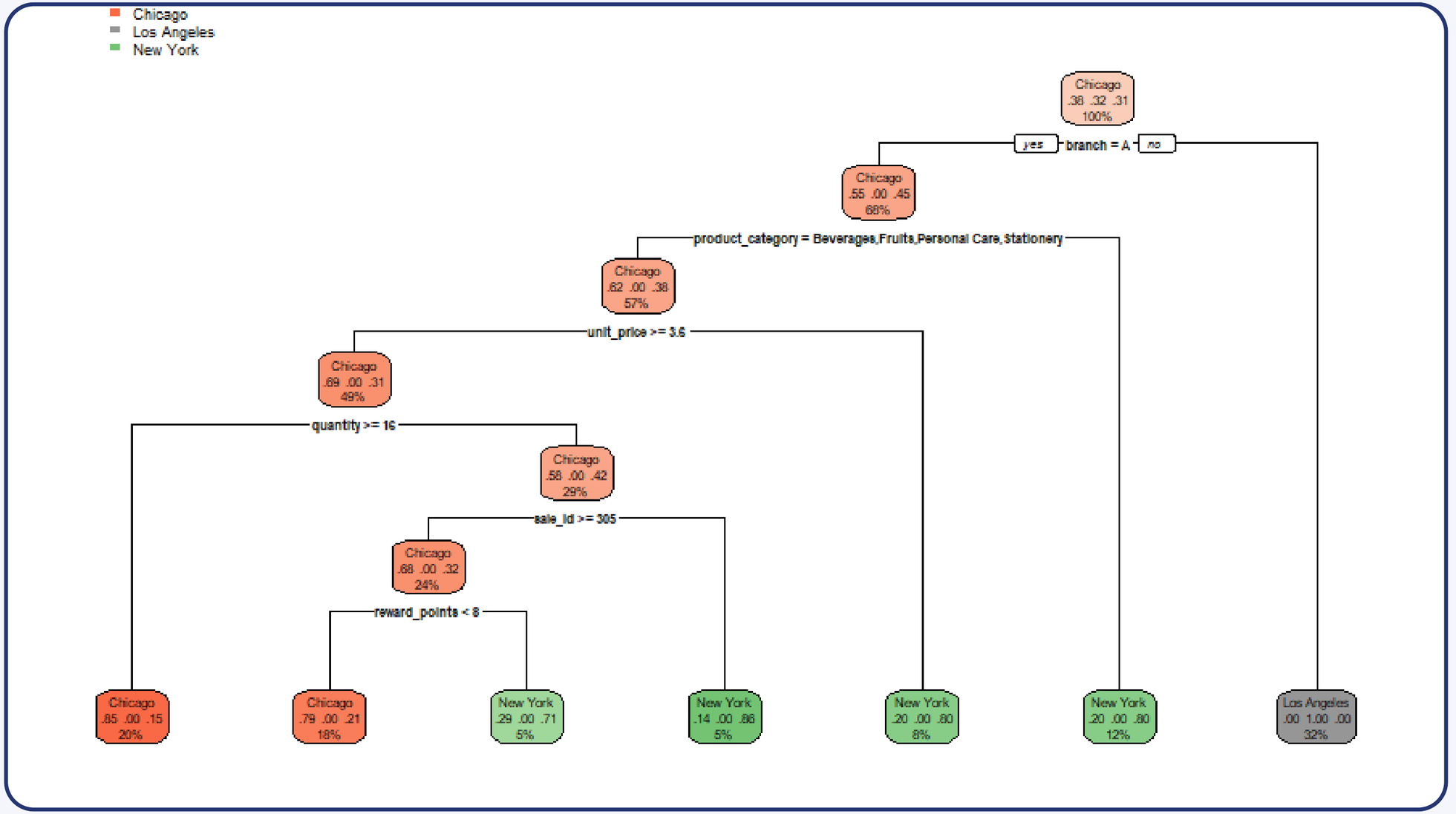
unit_price - 55.5%

quantity - 44.4%

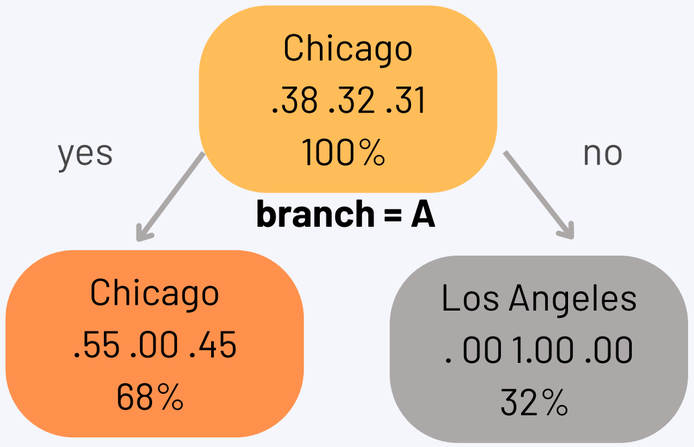
Dimension 3 - 11.4%

reward_points - 84.1%

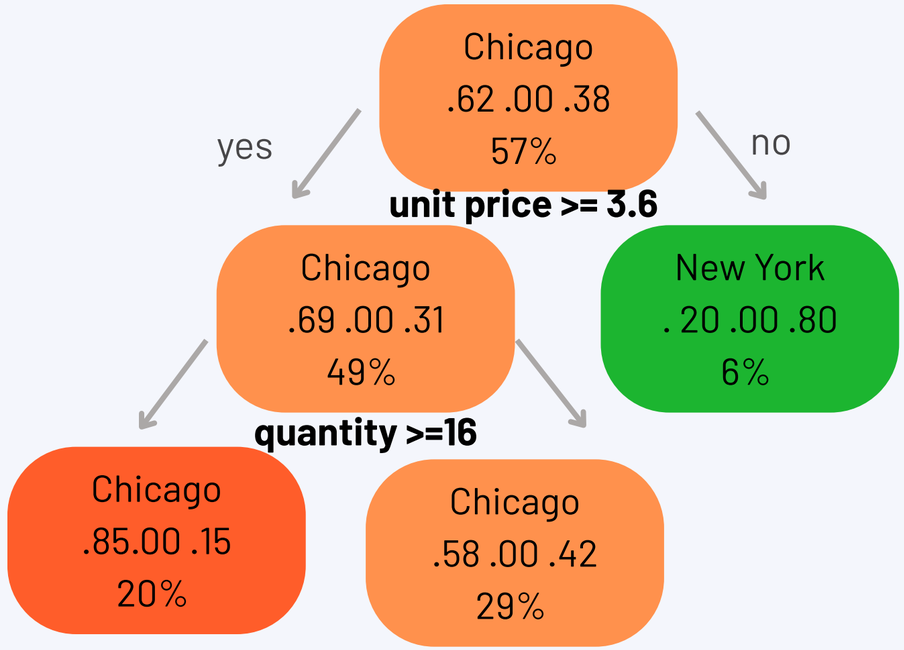
DECISION TREE CLASSIFICATION



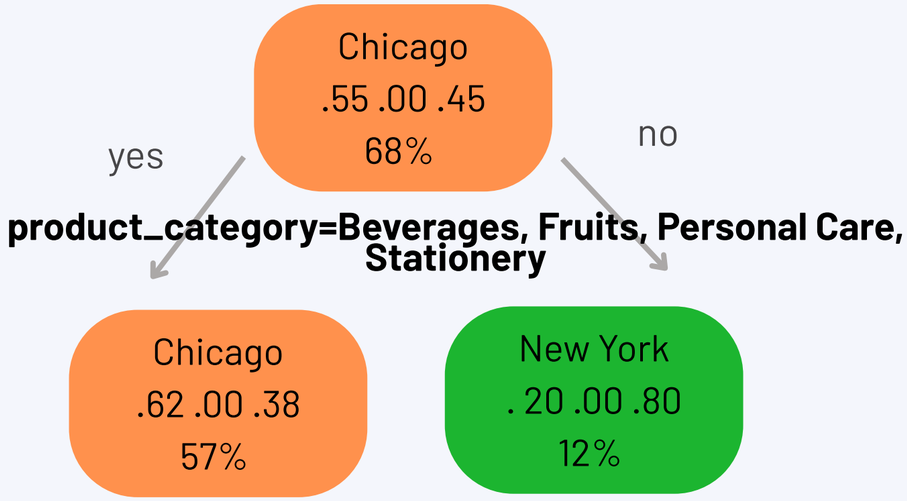
First level



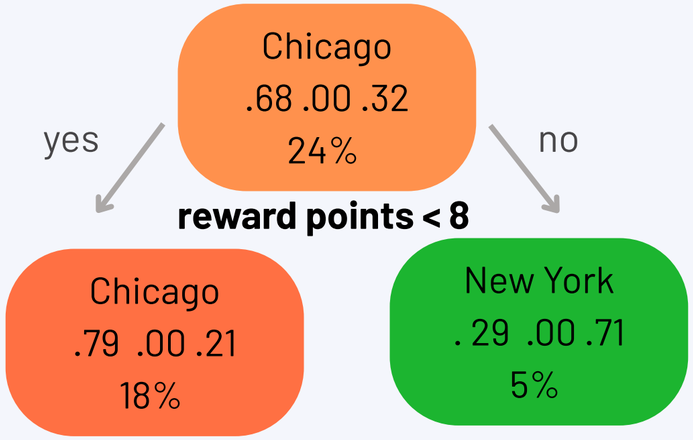
Intermediate node



Intermediate node



Leaf node



MODEL PERFORMANCE ON TEST SET

64.25%

Overall accuracy

Correctly classified 559 of 870 test observations

100%

Los Angeles

Perfect classification: 285/285 correct

42%

Chicago

117 correct, 164 misclassified as NY

52%

New York

157 correct, 147 misclassified as Chicago

LINEAR REGRESSION

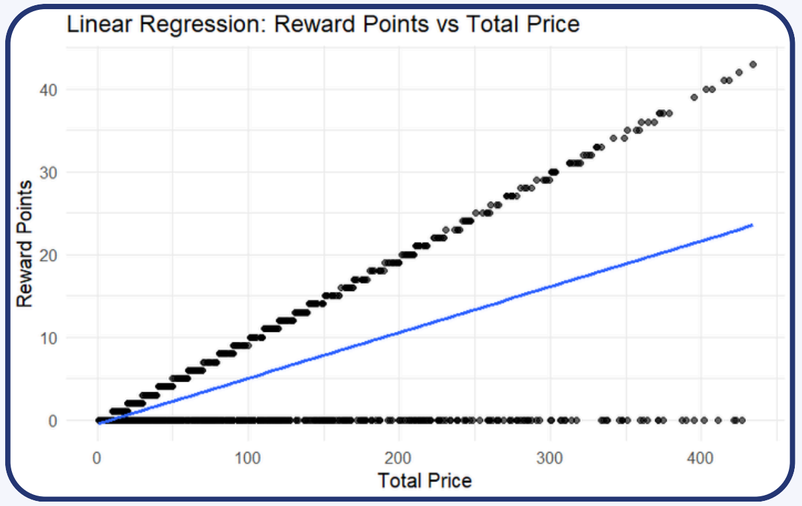
Target variable: reward_points

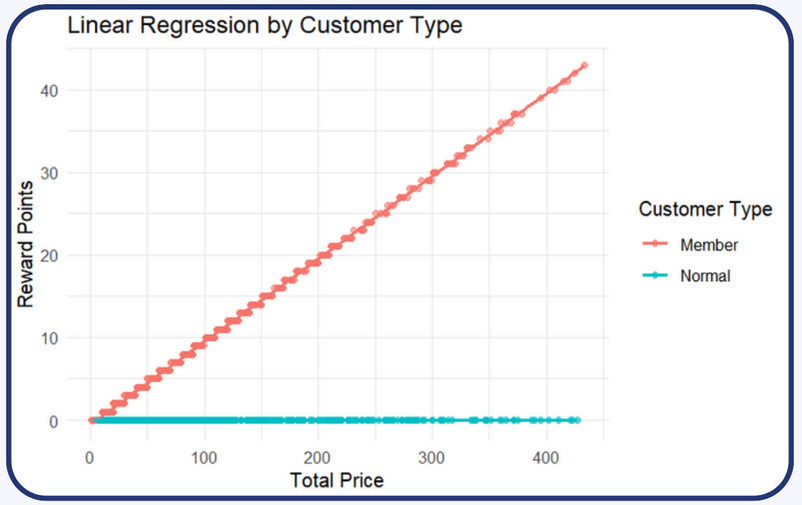
Predictors: total_price, customer_type and product_category

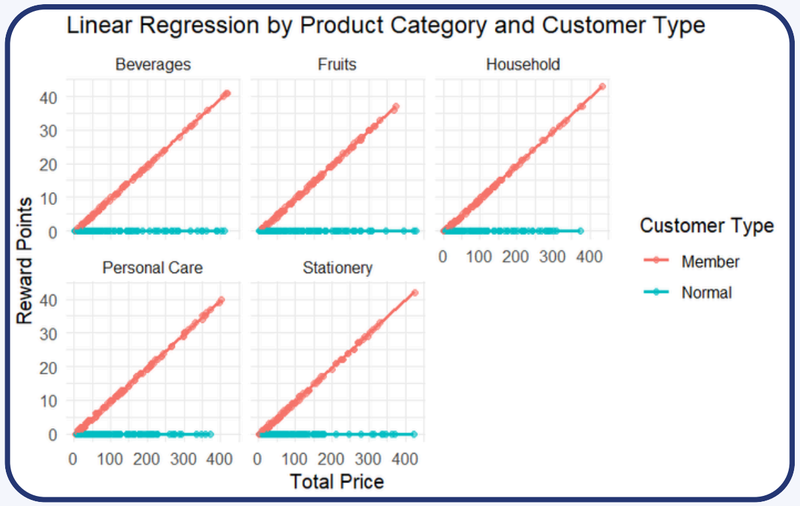
```
model1 = lm(reward_points ~ total_price)
```

```
model2 = lm(reward_points ~ total_price + customer_type)
```

```
model3 = lm(reward_points ~ total_price + customer_type + product_category)
```







β Coefficient 0.055

slope estimate

R^2 Value 0.35

Explained variance

p-value < 0.001

Highly significant

β Coefficient -11.31

slope estimate

R^2 Value 0.715

Explained variance

p-value < 0.001

Highly significant

β Coefficient from -0.3 to 0.6

slope estimate

R^2 Value 0.713

Explained variance

p-values > 0.05

NOT significant

Our team

Vittoria Calonghi

Claudia Cristofolini

Benedetta Di Palma

THANK YOU FOR YOUR ATTENTION