Project: Business Intelligence – Retail

Author: Miro Zilaji, 2.6.2025

 # **Guidelines for Reproducible Steps**

**1. Brief Presentation of the Problem Area / Overview of the Topic**

**1.1 Content**

****Core of the Investigation:****

The project focuses on analyzing and presenting historical data of a retail Business in order to allow basis for later predicting future trends and subsequent decision-making.

****Main Objectives of the Work:****

- Analyze sales data and customer´s reviews and ratings

**1.2 Justification of the Topic**

****Why is the Topic Relevant and Worth Investigating?****

Retail Business is facing (already for years) pretty dramatic challenges. Shift from classical "Brick and Mortar" shops to online sales which dramatically increased number of competitors, which suddenly are not only local stores but rather companies from all around world, competing for each single customer, subsequently pressuring prices down, leading to shrinkage of margins and decrease of profits. These new trends demand new business strategies and processes where knowing and understanding data has become a crucial part of business success.

****Personal Motivation:****

I have managed two online shops and have developed interest in retail dynamics and forces.

**2. Reproducible Steps**

**2.1 State of Research / Literature Review**

Need for understanding the sales data exists as long the trade exists. It is an ongoing process.

Which aspects were examined, and which were not?

- this part of the project focuses on analyzing and understanding historical data and is a basis for further projects (trend predictions, price policy adjustments, seasonal changes in assortment, supply chain adjustments,…) .

Controversies and Methods Used So Far:

- since we are dealing with historical data, there are no known controversies regarding the methodology used.

2.2 Research Question

- which data are more relevant (important) given the rapid changes of global markets and customer´s behaviour?

2.3 State of Research

- Existing studies show **Random Forest and AdaBoost perform well** in obesity classification.
- **Feature importance analysis** often highlights **BMI, diet, and physical activity** as key predictors.

2.4 Knowledge Gap

- Research would benefit from additional data (not available at this moment) :
    - Inventory Data
    - seasonal price (Margin) changes
    - weather conditions per day
    - …

2.5 Methodology

Detailed and Reproducible Description of the Approach:

- Raw Data available for research:
    - Sales.csv
    - Products.csv
    - Customers.csv
    - Reviews.json

Datasets are imported in Python, analyzed data structure and calculated key sales parameters. Those are presented in

number format in tables and in plots for better visualization.

Libraries used:

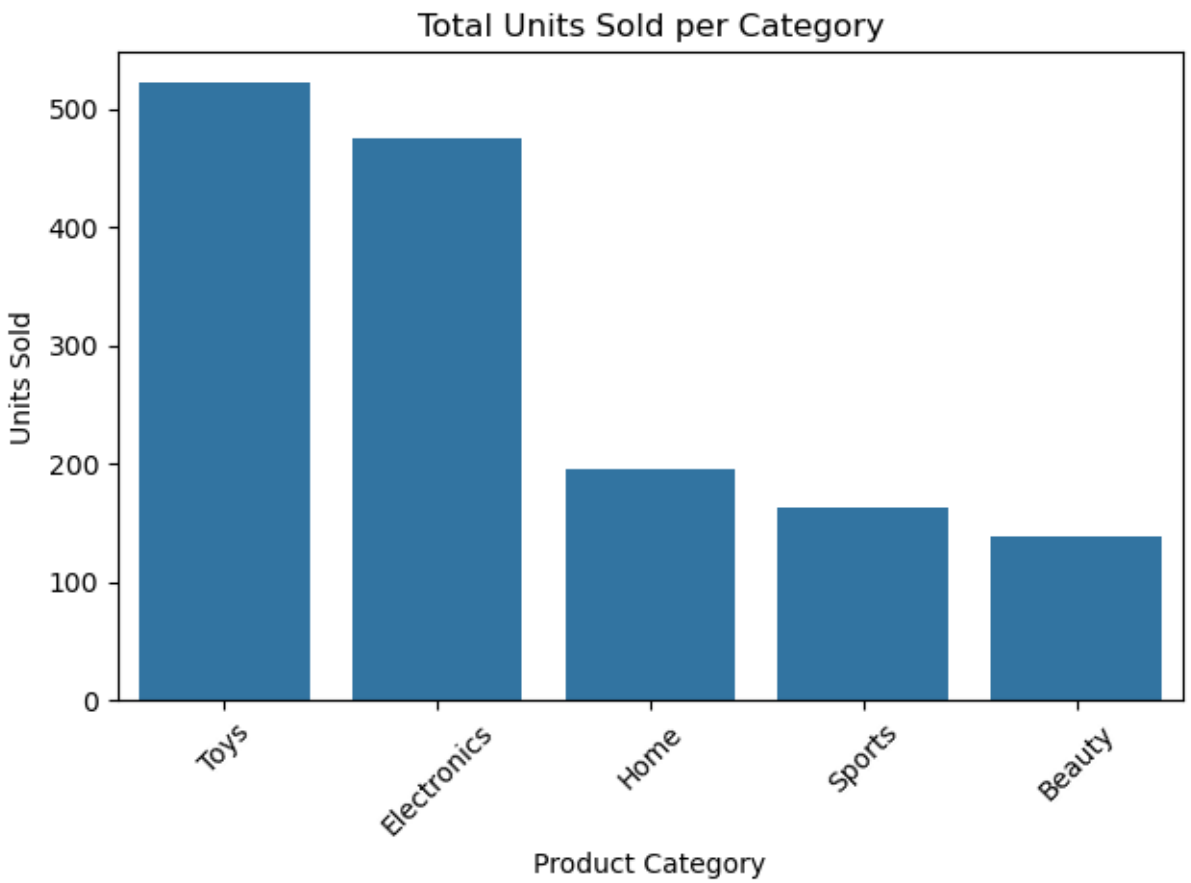
- Pandas
- Seaborn
- Matplotlib

```python
# -*- coding: utf-8 -*-
"""
Created on Mon Jun  2 15:09:38 2025

@author: miroz
"""



import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt



# Load data
customers = pd.read_csv("customers.csv")
products = pd.read_csv("products.csv")
sales = pd.read_csv("sales.csv")

print(customers.columns)

# Merge all three datasets
sales_full = sales.merge(products, on="product_id").merge(customers, on="customer_id")

# Rename customer name column for clarity
sales_full = sales_full.rename(columns={"name_y": "customer_name", "name_x": "product_name"})


print(sales_full.columns)


# ------------------------------
# 1. Total sales by product category
# ------------------------------
category_sales = sales_full.groupby("category")["quantity"].sum().sort_values(ascending=False)
print("\n 🔹 Total units sold by category:\n", category_sales)

sns.barplot(x=category_sales.index, y=category_sales.values)
plt.xticks(rotation=45)
plt.title("Total Units Sold per Category")
plt.ylabel("Units Sold")
plt.xlabel("Product Category")
plt.tight_layout()
plt.show()
```

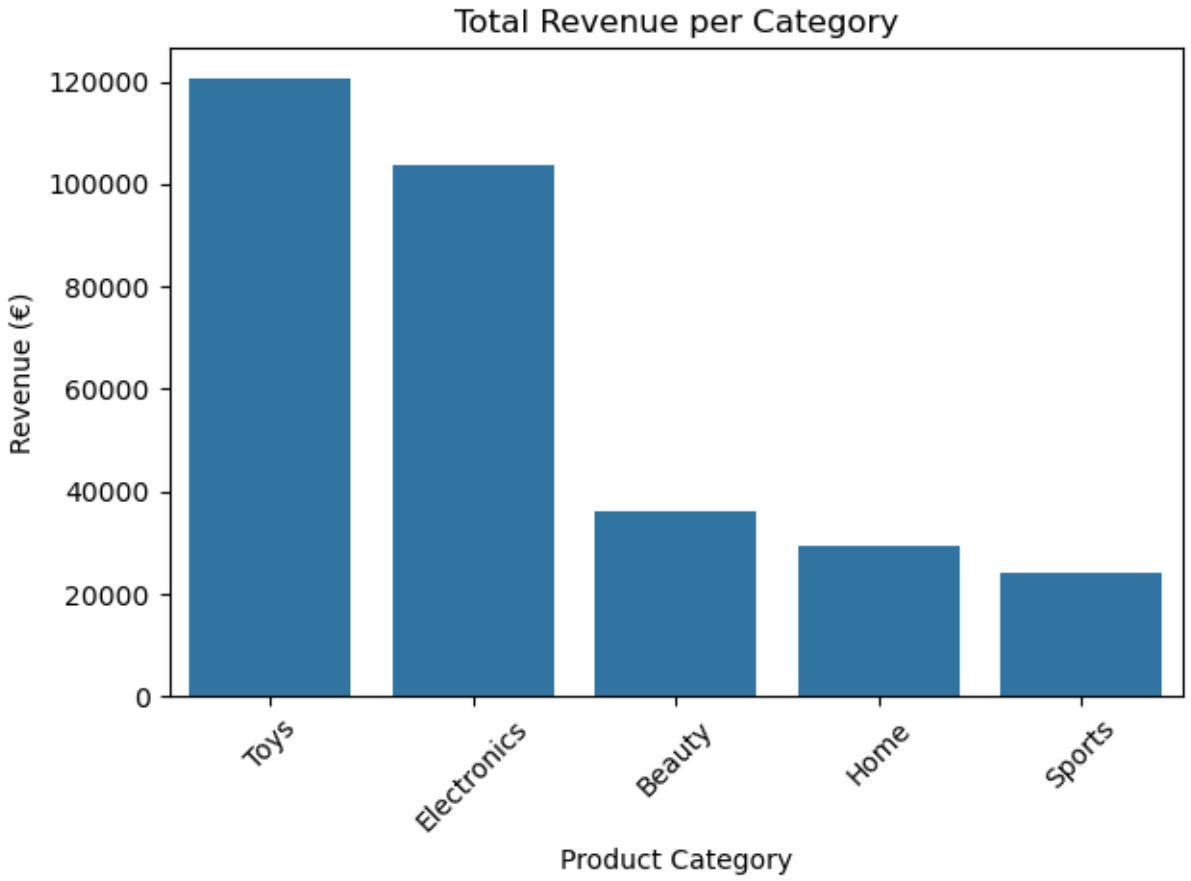
```
Index(['customer_id', 'name', 'age', 'gender', 'country'], dtype='object')
Index(['sale_id', 'customer_id', 'product_id', 'quantity', 'sale_date',
       'product_name', 'category', 'price', 'customer_name', 'age', 'gender',
       'country'],
      dtype='object')
```

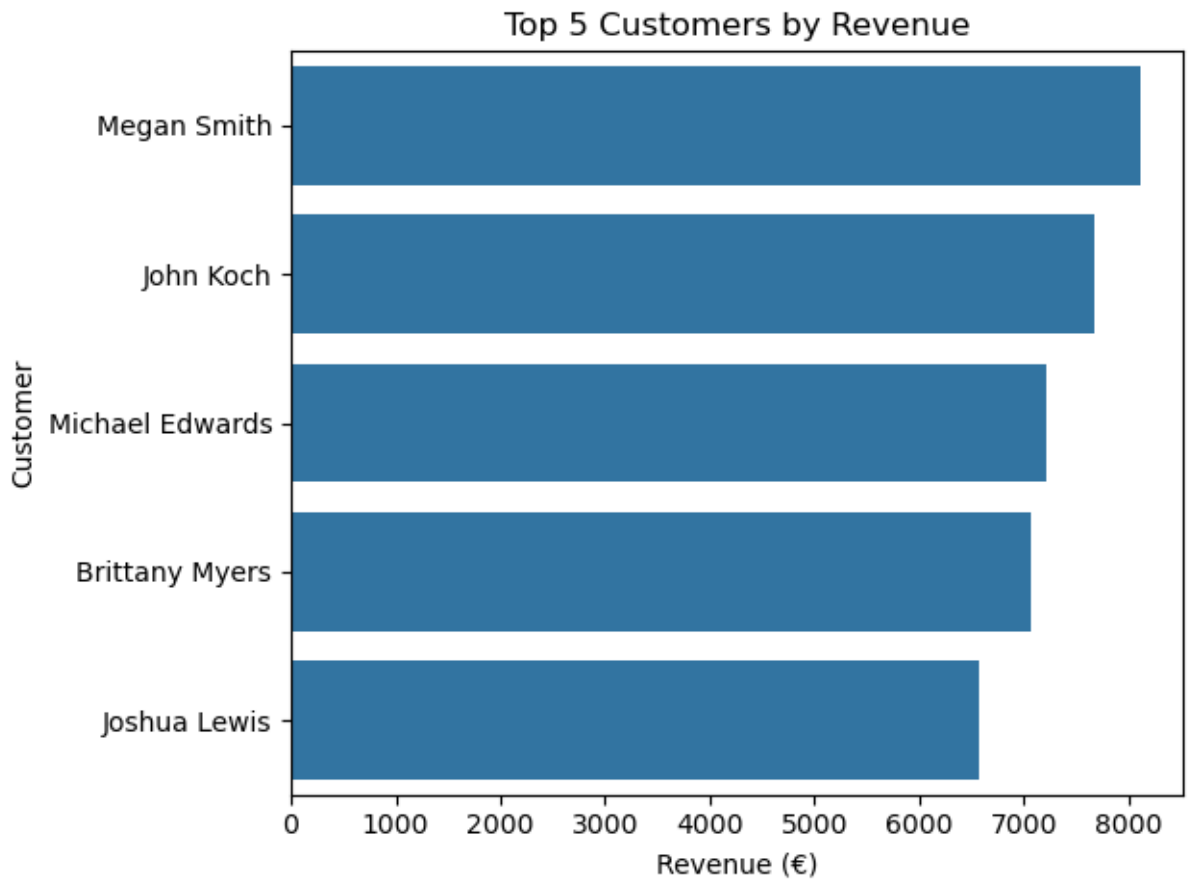
```
In [2]: runfile('C:/Users/miroz/OneDrive/Documents/Miro/Miro/Python/Big Data/Projekt/Big Data - End Project/Big
Data - End Project - Dataset/Sales per Category +.py', wdir='C:/Users/miroz/OneDrive/Documents/Miro/Miro/Python/
Big Data/Projekt/Big Data - End Project/Big Data - End Project - Dataset')
category
Beauty         138
Electronics    475
Home           195
Sports         164
Toys           522
Name: quantity, dtype: int64
```

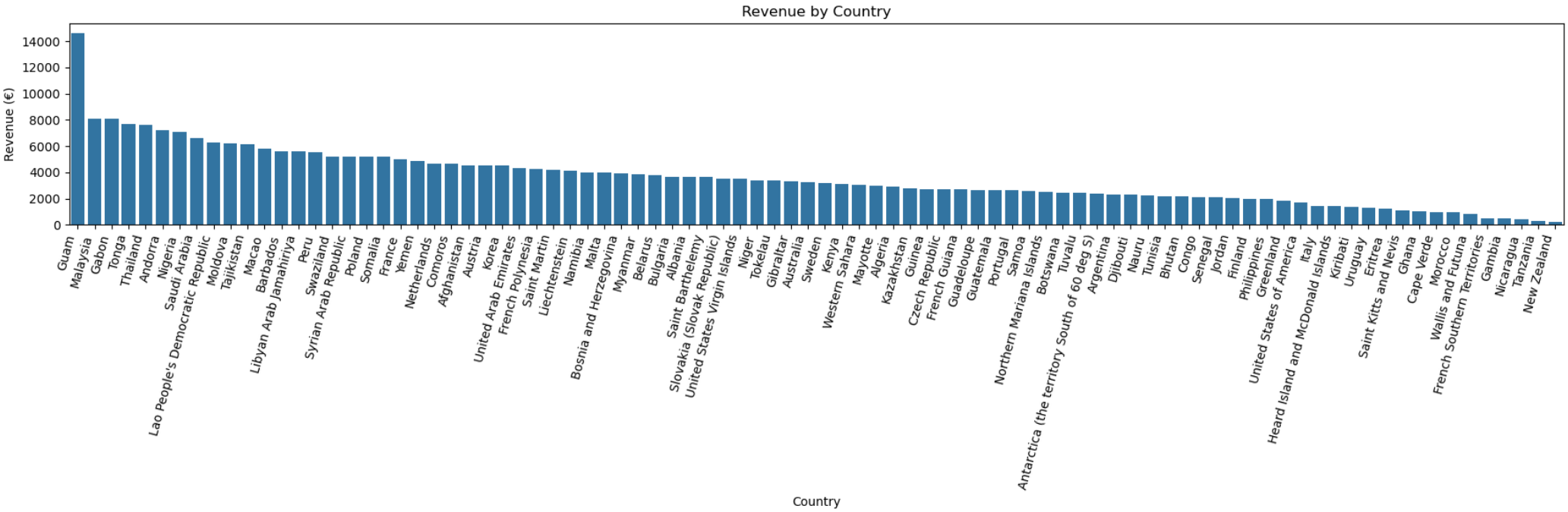
```
Console 1/A ×
In [9]: runfile('C:/Users/miroz/OneDrive/Documents/Miro/Miro/Python/Big Data/Projekt/Big Data - End Project/Big Data - End Project - Dataset/untitled4.py',
wdir='C:/Users/miroz/OneDrive/Documents/Miro/Miro/Python/Big Data/Projekt/Big Data - End Project/Big Data - End Project - Dataset')
Index(['customer_id', 'name', 'age', 'gender', 'country'], dtype='object')
Index(['sale_id', 'customer_id', 'product_id', 'quantity', 'sale_date',
       'product_name', 'category', 'price', 'customer_name', 'age', 'gender',
       'country'],
      dtype='object')

 🔹 Total units sold by category:
 category
Toys           522
Electronics    475
Home           195
Sports         164
Beauty         138
Name: quantity, dtype: int64

 🔹 Total revenue by category:
 category
Toys           120433.26
Electronics    103665.14
Beauty          36258.92
Home            29502.02
Sports          24184.46
Name: revenue, dtype: float64

 🔹 Top 5 customers by revenue:
 customer_name
Megan Smith        8109.39
John Koch          7676.68
Michael Edwards    7207.09
Brittany Myers     7056.27
Joshua Lewis       6574.83
Name: revenue, dtype: float64

 🔹 Revenue by age group:
 age_group
18-29    95719.23
30-44    87331.73
45-59    64700.74
60+      66292.10
Name: revenue, dtype: float64
c:\users\miroz\onedrive\documents\miro\miro\python\big data\projekt\big data - end project\big data - end project - dataset\untitled4.py:108: FutureWarning: The
default of observed=False is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain current behavior or observed=True
to adopt the future default and silence this warning.
  age_revenue = sales_full.groupby("age_group")["revenue"].sum()
```



```python
# ------------------------------
# 2. Revenue by category
# ------------------------------
sales_full["revenue"] = sales_full["quantity"] * sales_full["price"]
revenue_by_category = sales_full.groupby("category")["revenue"].sum().sort_values(ascending=False)
print("\n 🔹 Total revenue by category:\n", revenue_by_category)

sns.barplot(x=revenue_by_category.index, y=revenue_by_category.values)
plt.xticks(rotation=45)
plt.title("Total Revenue per Category")
plt.ylabel("Revenue (€)")
plt.xlabel("Product Category")
plt.tight_layout()
plt.show()

# ------------------------------
# 3. Top 5 customers by total spend
# ------------------------------
top_customers = sales_full.groupby("customer_name")["revenue"].sum().sort_values(ascending=False).head(5)

print("\n 🔹 Top 5 customers by revenue:\n", top_customers)

sns.barplot(x=top_customers.values, y=top_customers.index)
plt.title("Top 5 Customers by Revenue")
plt.xlabel("Revenue (€)")
plt.ylabel("Customer")
plt.tight_layout()
plt.show()

# ------------------------------
# 4. Sales by country
# ------------------------------
country_sales = sales_full.groupby("country")["revenue"].sum().sort_values(ascending=False)
print("\n 🔹 Revenue by country:\n", country_sales)

sns.barplot(x=country_sales.index, y=country_sales.values)
plt.xticks(rotation=90)
plt.title("Revenue by Country")
plt.ylabel("Revenue (€)")
plt.xlabel("Country")
plt.tight_layout()
plt.show()
```

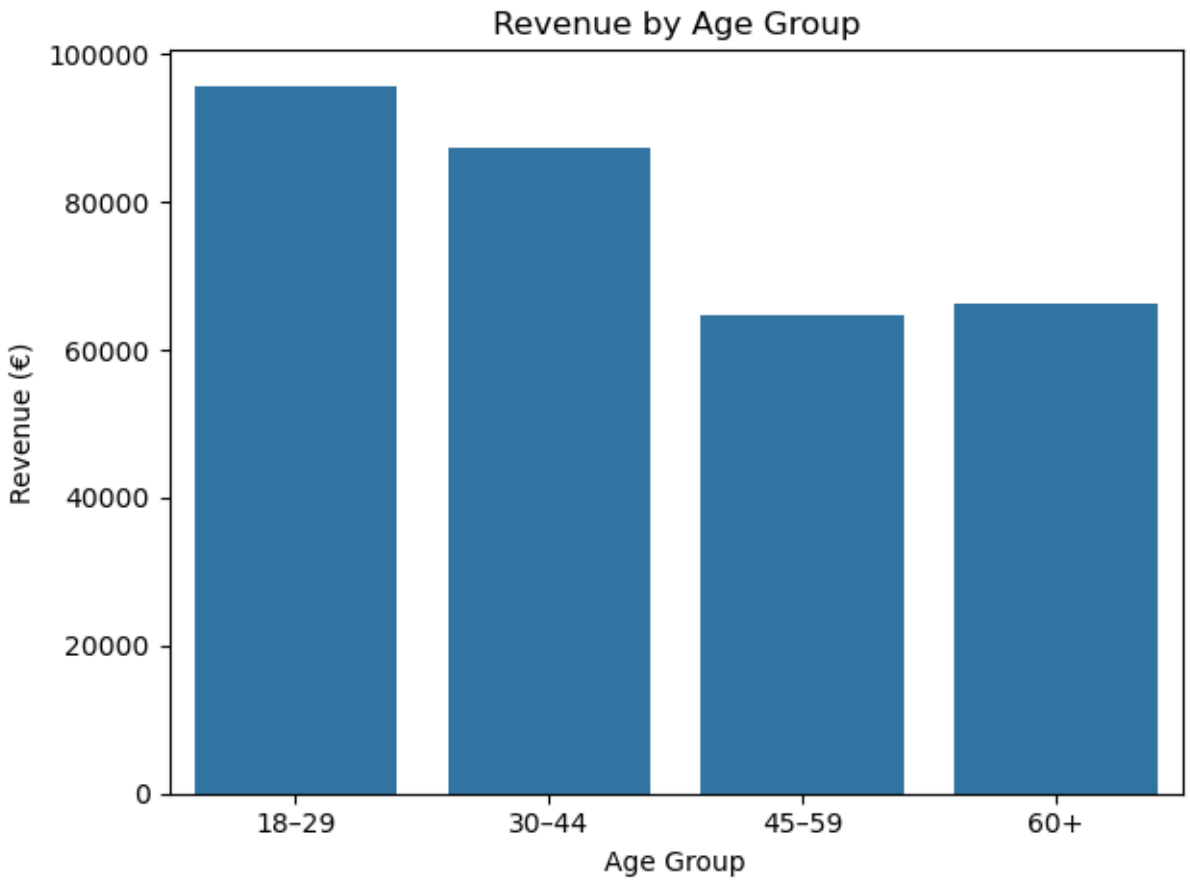

Total Revenue per Category
Revenue (€)
120000
100000
80000
60000
40000
20000
0
Toys
Electronics
Beauty
Home
Sports
Product Category


Top 5 Customers by Revenue
Customer
Megan Smith
John Koch
Michael Edwards
Brittany Myers
Joshua Lewis
0
1000
2000
3000
4000
5000
6000
7000
8000
Revenue (€)


Revenue by Country
Revenue (€)
14000
12000
10000
8000
6000
4000
2000
0
Guam
Malaysia
Gabon
Tonga
Thailand
Andorra
Nigeria
Saudi Arabia
Lao People's Democratic Republic
Moldova
Tajikistan
Macao
Barbados
Libyan Arab Jamahiriya
Peru
Swaziland
Syrian Arab Republic
Poland
Somalia
France
Yemen
Netherlands
Comoros
Afghanistan
Austria
Korea
United Arab Emirates
French Polynesia
Saint Martin
Liechtenstein
Namibia
Malta
Bosnia and Herzegovina
Myanmar
Belarus
Bulgaria
Albania
Saint Barthelemy
Slovakia (Slovak Republic)
United States Virgin Islands
Niger
Tokelau
Gibraltar
Australia
Sweden
Kenya
Western Sahara
Mayotte
Algeria
Kazakhstan
Guinea
Czech Republic
French Guiana
Guadeloupe
Guatemala
Portugal
Samoa
Northern Mariana Islands
Botswana
Tuvalu
Antarctica (the territory South of 60 deg S)
Argentina
Djibouti
Nauru
Tunisia
Bhutan
Congo
Senegal
Jordan
Finland
Philippines
Greenland
United States of America
Heard Island and McDonald Islands
Italy
Kiribati
Uruguay
Eritrea
Saint Kitts and Nevis
Ghana
Cape Verde
Morocco
Wallis and Futuna
French Southern Territories
Gambia
Nicaragua
Tanzania
New Zealand
Country

Sales per Country shows several points that are worth further analysis. All countries but Guam are following almost linear pattern. Guam sales is more than 80 % higher (81,25 %) than second best market (Malaysia). Given the population of Guam (167.777 in 2024) it is clear that available sales data are not reliable and definitely not representative.

```
# ------------------------------
# 5. Age group analysis
# ------------------------------
# Define age groups
bins = [18, 30, 45, 60, 80]
labels = ["18-29", "30-44", "45-59", "60+"]
sales_full["age_group"] = pd.cut(sales_full["age"], bins=bins, labels=labels, right=False)

age_revenue = sales_full.groupby("age_group")["revenue"].sum()
print("\n 🔹 Revenue by age group:\n", age_revenue)

sns.barplot(x=age_revenue.index, y=age_revenue.values)
plt.title("Revenue by Age Group")
plt.xlabel("Age Group")
plt.ylabel("Revenue (€)")
plt.tight_layout()
plt.show()
```



Interesting, somewhat expected result show that the most products are sold in Toys and Electronics categories and most of customers (about 66 %) are 18 – 44 years old. Somewhat unexpected results are that Beauty, Home and Sport categories are sold much less (units and revenue) than Toys and **Electronics. Fast** conclusion is that customers (18 to 44 years old) are spending much more for kids than for themselves (and don´t have much time for sport).

pgAdmin 4

File Object Tools Edit View Window Help

Object Explorer

- Materialized Views
- Operators
- Procedures
- Sequences
- Tables (4)
  - bdmp_table
    - Columns (12)
      - ORDERNUMBER
      - QUANTITYORDERED
      - PRICEEACH
      - ORDERLINENUMBER
      - SALES
      - QTR_ID
      - MONTH_ID
      - YEAR_ID
      - PRODUCTLINE
      - PRODUCTCODE
      - CITY
      - COUNTRY
    - Constraints
    - Indexes
    - RLS Policies
    - Rules
    - Triggers
  - customers
  - products
  - sales
- Trigger Functions
- Types
- Views
- Subscriptions
- Friday
  - Casts
  - Catalogs
  - Event Triggers
  - Extensions
  - Foreign Data Wrappers

MP/postgres@... | BDMP/postgres@... | BDMP/postgres@... | BDMP/postgres@... | public.customers/... | public.sales/BDMP/postgres@PostgreSQL 17

public.sales/BDMP/postgres@PostgreSQL 17

No limit

Query | Query History | Scratch Pad

```
SELECT * FROM public.sales
ORDER BY sale_id ASC
```

Data Output | Messages | Notifications

Showing rows: 1 to 500 Page No: 1 of 1

	sale_id [PK] integer	customer_id integer	product_id integer	quantity integer	sale_date date
97	97	41	1	4	2024-09-07
98	98	44	8	2	2024-11-15
99	99	68	20	2	2024-10-12
100	100	69	4	3	2024-07-24
101	101	69	20	2	2024-09-05
102	102	50	14	2	2024-07-24

Total rows: 500 Query complete 00:00:00.507 CRLF Ln 1, Col 1

Elements Console Sources Network Performance Memory Application Security Lighthouse Recorder

Styles Computed Layout Event Listeners

Search DEU 14:45 02.06.2025

Datasets

Query Query History

```sql
SELECT table_name
FROM information_schema.tables
WHERE table_schema = 'public';
```

Data Output Messages Notifications

	table_name name
1	customers
2	sales
3	products

Sales.csv structure and data type

Data Output Messages Notifications

	sale_id [PK] integer	customer_id integer	product_id integer	quantity integer	sale_date date
1	1	84	20	1	2024-10-05
2	2	72	19	1	2024-12-20
3	3	28	5	5	2024-07-06
4	4	1	3	4	2025-04-10
5	5	24	10	1	2025-05-28
6	6	93	20	3	2025-05-18
7	7	64	4	5	2024-12-07
8	8	31	7	1	2024-10-26
9	9	12	20	4	2025-03-14
10	10	18	3	2	2024-11-01
11	11	88	8	5	2024-06-06
12	12	9	11	2	2024-11-27
13	13	21	2	3	2024-11-05
14	14	57	2	2	2024-07-28
15	15	3	5	4	2024-06-18
16	16	69	2	1	2024-07-22
17	17	15	4	4	2024-08-06

Query Query History

```sql
-- Preview data
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'sales';
```

Data Output Messages Notifications

	column_name name	data_type character varying
1	sale_id	integer
2	customer_id	integer
3	product_id	integer
4	quantity	integer
5	sale_date	date

Products.csv structure and data type

Data Output Messages Notifications

	product_id [PK] integer	name text	category text	price numeric
1	1	At Tool	Home	246.65
2	2	Operation Kit	Toys	473.92
3	3	Bit Gadget	Electronics	294.72
4	4	Set Tool	Electronics	141.45
5	5	Month Gadget	Sports	240.69
6	6	Last Kit	Electronics	123.22
7	7	Six Kit	Sports	22.28
8	8	Let Tool	Toys	209.56
9	9	Development Gadget	Electronics	409.15
10	10	Fine Tool	Beauty	332.59
11	11	Care Kit	Toys	270.26
12	12	Level Item	Electronics	111.36
13	13	Region Item	Toys	25.84
14	14	Not Gadget	Toys	62.98
15	15	Trouble Item	Home	137.86
16	16	Note Item	Beauty	139.82

Query Query History

```sql
-- Preview data
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'products';
```

Data Output Messages Notifications

	column_name name	data_type character varying
1	product_id	integer
2	price	numeric
3	name	text
4	category	text

Customers.csv structure and data type

Data Output Messages Notifications

	customer_id [PK] integer	name text	age integer	gender text	country text
1	1	Karl Contreras	52	Female	Poland
2	2	James Cunningham	69	Female	Kiribati
3	3	Justin Woods	67	Female	Barbados
4	4	Mrs. Kristine Hess	41	Male	Portugal
5	5	Derrick Todd	22	Female	Tanzania
6	6	Rebecca Barrett	62	Female	United States of America
7	7	Eric Rodriguez	26	Female	Saint Kitts and Nevis
8	8	Margaret Silva	40	Male	Italy
9	9	Don Curtis	22	Female	Samoa
10	10	Charles Hall	21	Female	Niger
11	11	David Flores	19	Male	Saint Martin
12	12	Alyssa Rowe	59	Male	Guadeloupe

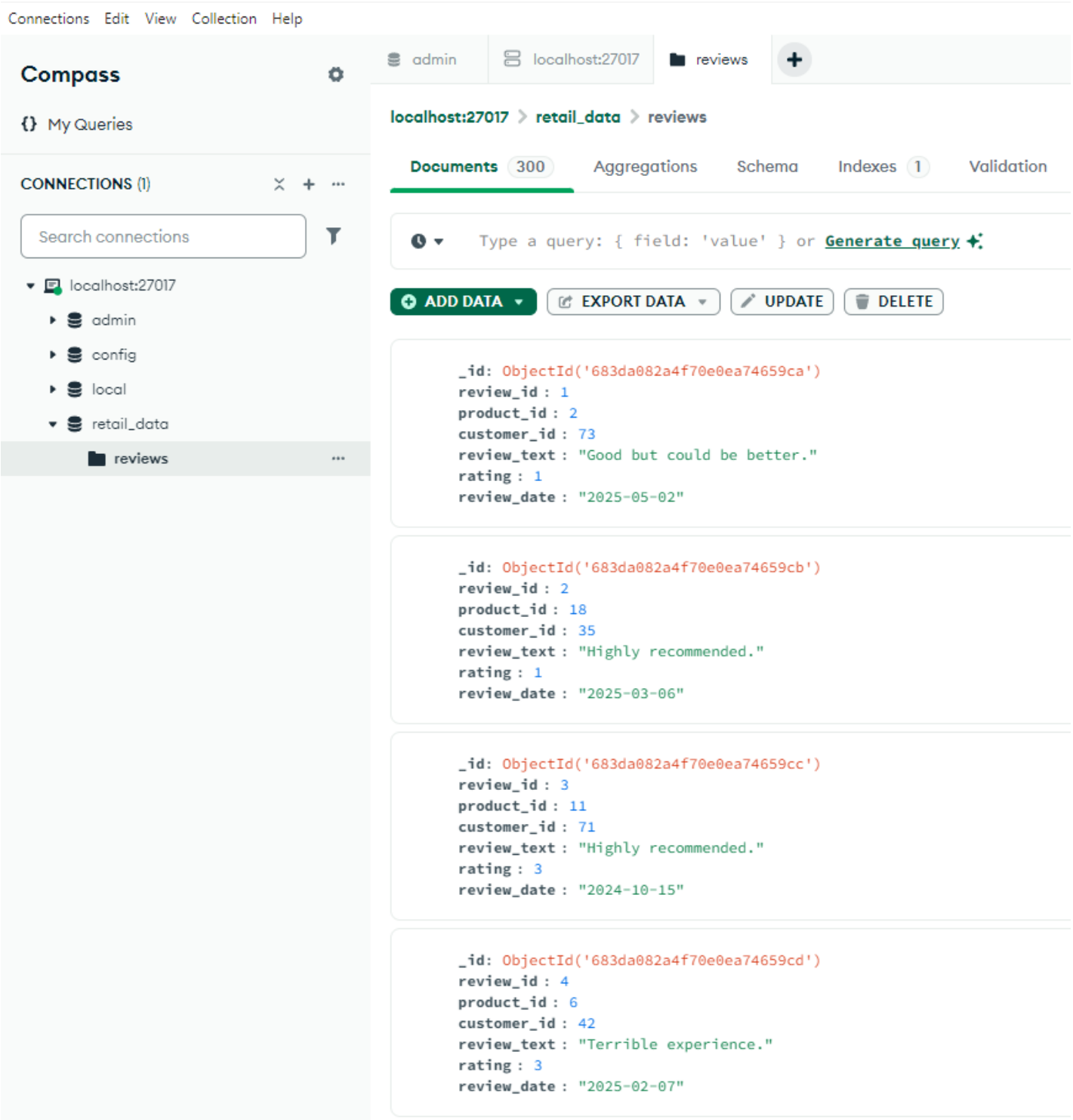
Query Query History

```sql
-- Preview data
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'customers';
```

Data Output Messages Notifications

	column_name name	data_type character varying
1	customer_id	integer
2	age	integer
3	name	text
4	gender	text
5	country	text

Reviews.json structure



Ratings after 1.5.2025

localhost:27017 > retail_data > reviews

Documents 300 | Aggregations | Schema | Indexes 1 | Validation

```
{ "review_date": { "$gte": "2025-05-01" } }
```

Project	`{ field: 0 }`
Sort	`{ field: -1 } or [['field', -1]]`
Collation	`{ locale: 'simple' }`
Index Hint	`{ field: -1 }`

ADD DATA | EXPORT DATA | UPDATE | DELETE | INSIGHT

```
_id: ObjectId('683da082a4f70e0ea74659ca')
review_id : 1
product_id : 2
customer_id : 73
review_text : "Good but could be better."
rating : 1
review_date : "2025-05-02"
```

```
_id: ObjectId('683da082a4f70e0ea74659d5')
review_id : 12
product_id : 1
customer_id : 61
review_text : "Would not buy again."
rating : 1
review_date : "2025-05-19"
```

```
_id: ObjectId('683da082a4f70e0ea74659dd')
review_id : 20
product_id : 20
customer_id : 22
review_text : "Excellent value for money."
rating : 1
review_date : "2025-05-09"
```

Ratings greater than 4

localhost:27017 > retail_data > reviews

Documents 300 | Aggregations | Schema | Indexes 1 | Validation

```
{ "rating": { "$gte": 4 } }
```

Project	`{ field: 0 }`
Sort	`{ field: -1 } or [['field', -1]]`
Collation	`{ locale: 'simple' }`
Index Hint	`{ field: -1 }`

ADD DATA | EXPORT DATA | UPDATE | DELETE | INSIGHT

```
_id: ObjectId('683da082a4f70e0ea74659d0')
review_id : 7
product_id : 9
customer_id : 42
review_text : "Excellent value for money."
rating : 5
review_date : "2024-09-03"
```

```
_id: ObjectId('683da082a4f70e0ea74659d3')
review_id : 10
product_id : 5
customer_id : 8
review_text : "Loved it!"
rating : 4
review_date : "2024-06-03"
```

```
_id: ObjectId('683da082a4f70e0ea74659d4')
review_id : 11
product_id : 20
customer_id : 34
review_text : "Excellent value for money."
rating : 5
review_date : "2024-11-09"
```

```
_id: ObjectId('683da082a4f70e0ea74659d7')
review_id : 14
```

Reviews with word “Great” in text (case insensitive – “$options”: “i”)

Documents 300 | Aggregations | Schema | Indexes 1 | Validation

```
{ "review_text": { "$regex": "great", "$options": "i" } }
```

Project	`{ field: 0 }`
Sort	`{ field: -1 } or [['field', -1]]`
Collation	`{ locale: 'simple' } `
Index Hint	`{ field: -1 }`

ADD DATA | EXPORT DATA | UPDATE | DELETE | INSIGHT

```
_id: ObjectId('683da082a4f70e0ea74659d9')
review_id : 16
product_id : 2
customer_id : 72
review_text : "Great product!"
rating : 5
review_date : "2025-03-14"
```

```
_id: ObjectId('683da082a4f70e0ea74659de')
review_id : 21
product_id : 11
customer_id : 27
review_text : "Great product!"
rating : 1
review_date : "2025-01-19"
```

```
_id: ObjectId('683da082a4f70e0ea74659e0')
review_id : 23
product_id : 12
customer_id : 56
review_text : "Great product!"
rating : 4
review_date : "2025-01-26"
```

All reviews for Product 5

localhost:27017 > retail_data > reviews

Documents 300 | Aggregations | Schema | Indexes 1 | Validation

```
{ "product_id": 5 }
```

Project	`{ field: 0 }`
Sort	`{ field: -1 } or [['field', -1]]`
Collation	`{ locale: 'simple' }`
Index Hint	`{ field: -1 }`

ADD DATA | EXPORT DATA | UPDATE | DELETE | INSIGHT

```
_id: ObjectId('683da082a4f70e0ea74659d1')
review_id : 8
product_id : 5
customer_id : 18
review_text : "Five stars!"
rating : 2
review_date : "2024-09-06"
```

```
_id: ObjectId('683da082a4f70e0ea74659d3')
review_id : 10
product_id : 5
customer_id : 8
review_text : "Loved it!"
rating : 4
review_date : "2024-06-03"
```

```
_id: ObjectId('683da082a4f70e0ea74659da')
review_id : 17
product_id : 5
customer_id : 49
review_text : "Excellent value for money."
rating : 1
review_date : "2024-07-31"
```

```
_id: ObjectId('683da082a4f70e0ea74659dc')
```

Anomaly: Reviews with word “great” but Rating 1 or 2

localhost:27017 > retail_data > reviews

Documents 300 | Aggregations | Schema | Indexes 1 | Validation

```
{
  "$and": [
    { "review_text": { "$regex": "great", "$options": "i" } },
    { "rating": { "$lte": 2 } }
  ]
}
```

Project	`{ field: 0 }`
Sort	`{ field: -1 } or [['field', -1]]`
Collation	`{ locale: 'simple' }`
Index Hint	`{ field: -1 }`

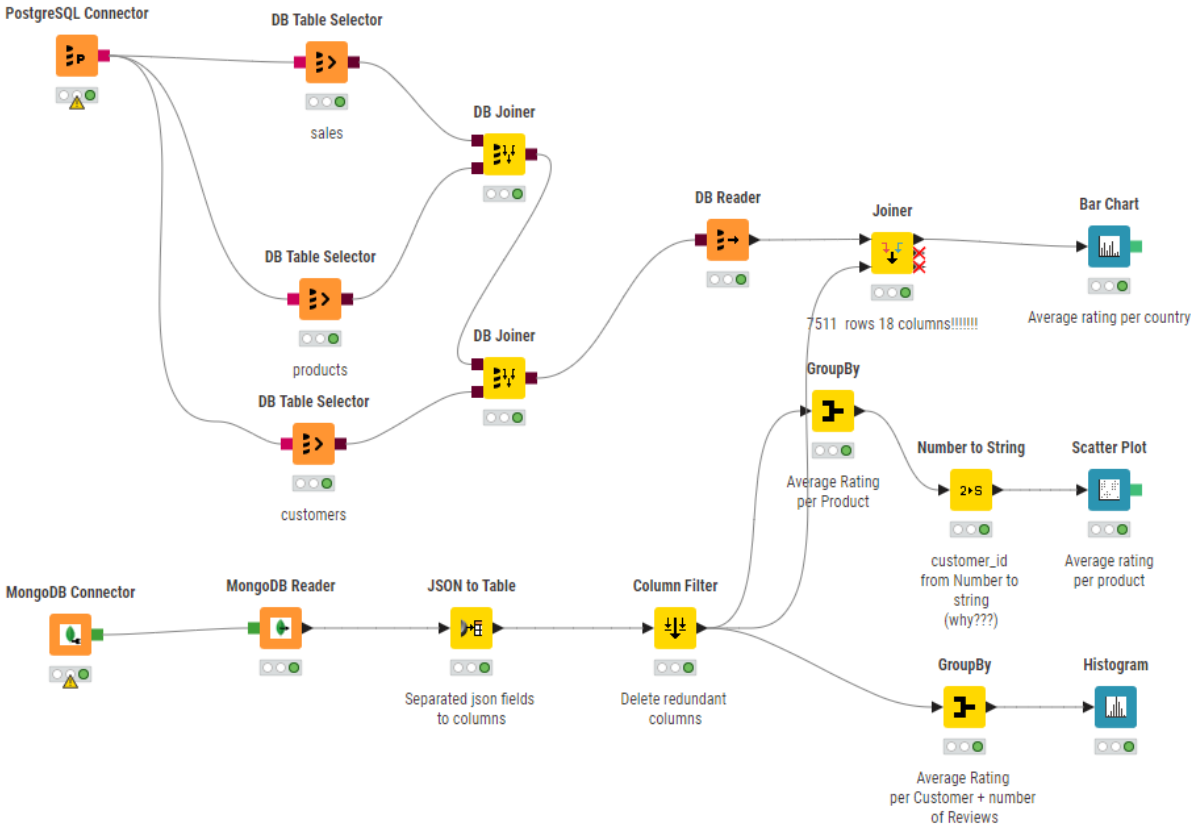
ADD DATA | EXPORT DATA | UPDATE | DELETE | INSIGHT

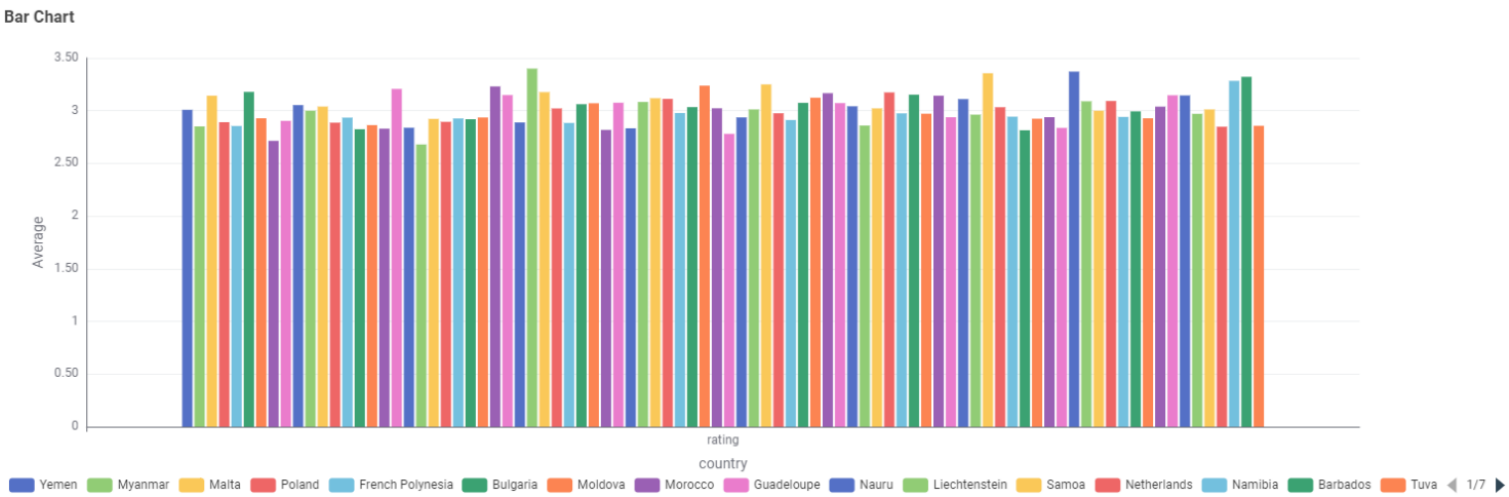
```
_id: ObjectId('683da082a4f70e0ea74659de')
review_id : 21
product_id : 11
customer_id : 27
review_text : "Great product!"
rating : 1
review_date : "2025-01-19"
```

```
_id: ObjectId('683da082a4f70e0ea74659e3')
review_id : 26
product_id : 17
customer_id : 5
review_text : "Great product!"
rating : 1
review_date : "2025-03-26"
```

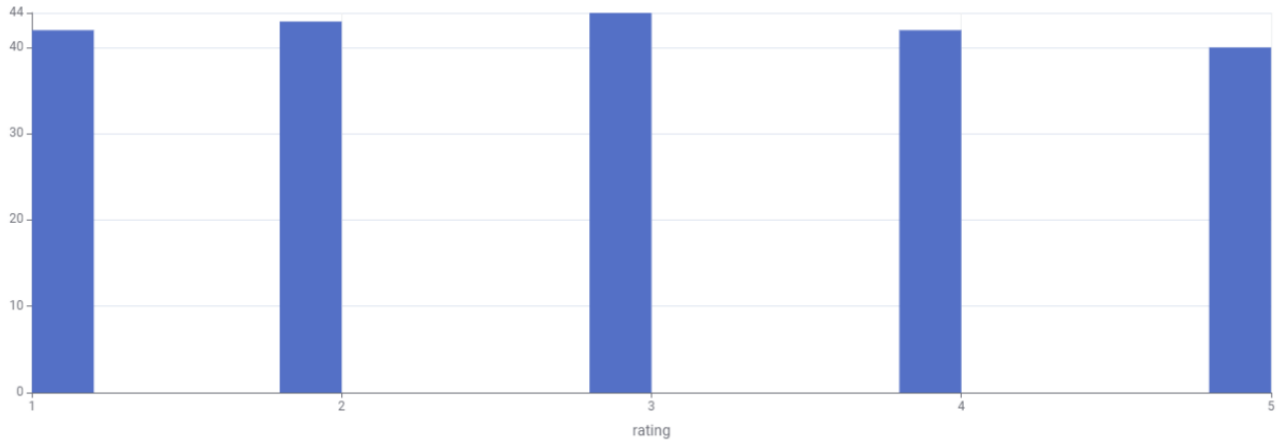
```
_id: ObjectId('683da082a4f70e0ea74659e7')
review_id : 30
product_id : 19
customer_id : 82
review_text : "Great product!"
rating : 2
review_date : "2024-06-30"
```

Visualisation with KNIME

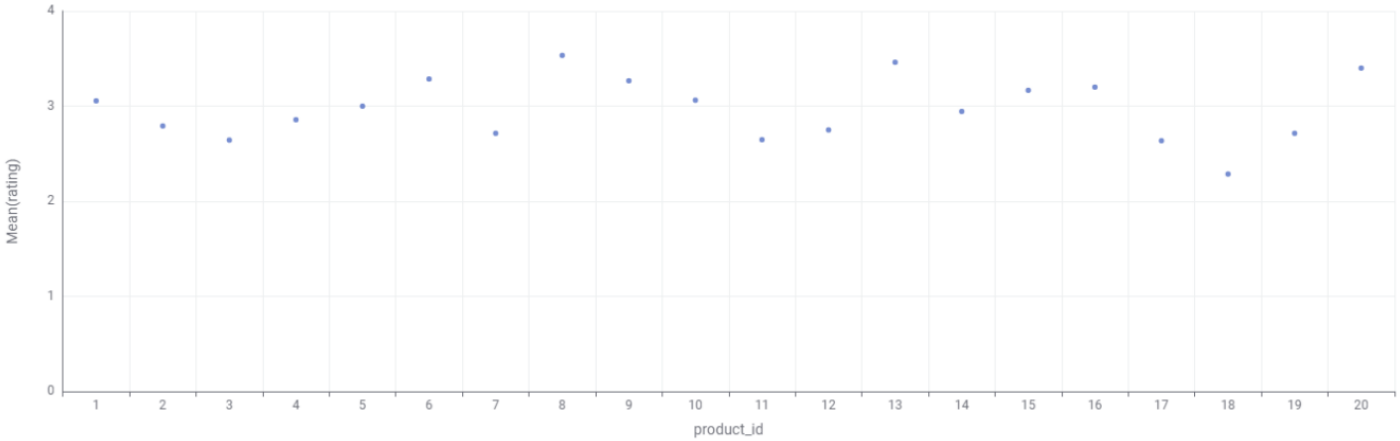




Histogram



Scatter Plot



Conclusion

The Project has processed raw data and provided necessary analysis, so that management can make their decisions.