Transforming and Analyzing Data with SQL

Theo Plessas

Project Description and overview of goals

- Download and import data from the provided CSV files into my local ecommercDB

- -Explore/Study the imported data to gather possible insights, while also keeping an eye for garbage data

- Transform the DB tables, using SQL to clean usable data

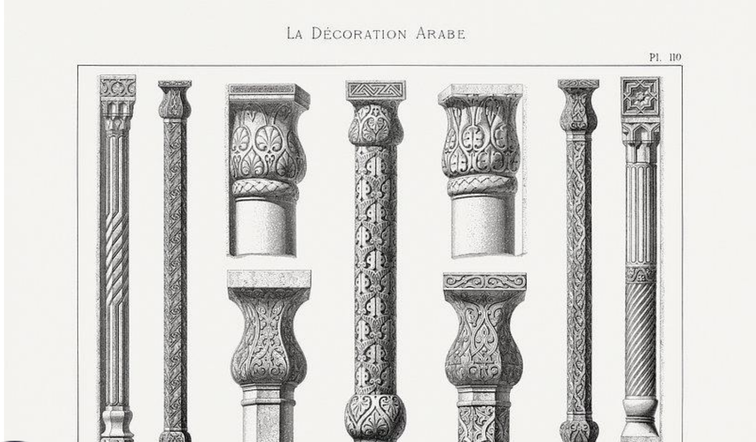
- Chronicle my process for future reference

Importing Process

- For each CSV file , create a matching table in the ecommerce DB

- Every table needs to have identical column structure

- Appropriate data types must be chosen per column

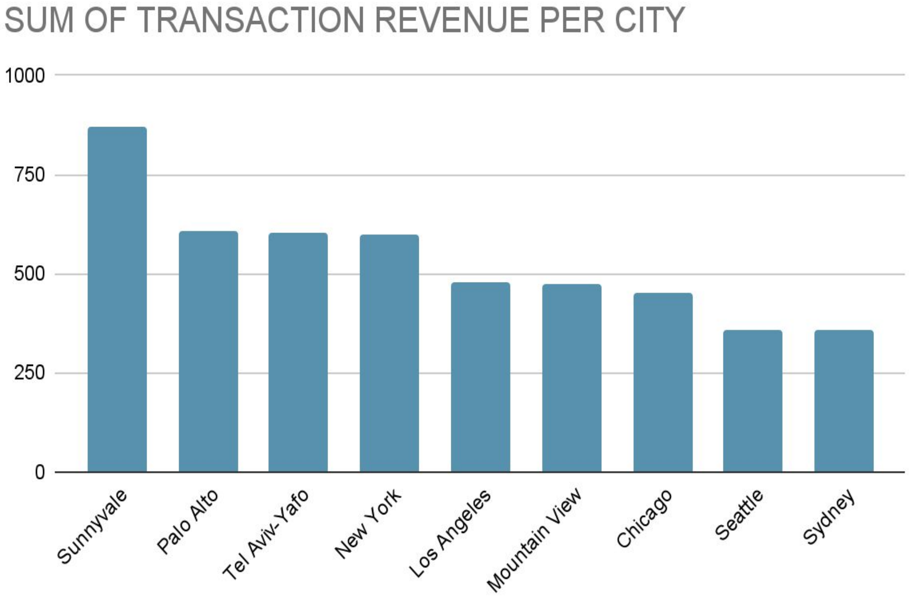


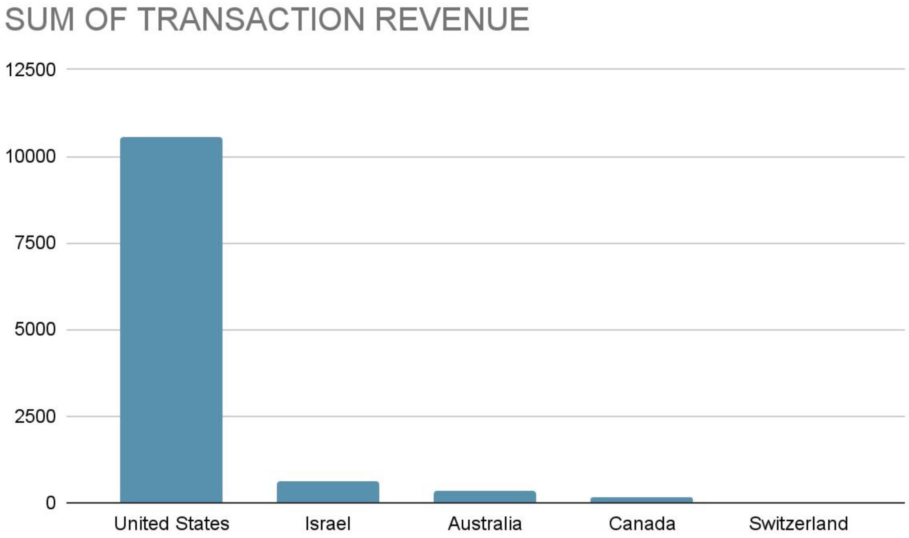
Data cleaning

- Identify Data Quality Issues
  - Before cleaning the data, it's essential to identify any data quality issues such as missing values, duplicate records, incorrect data types, outliers, and inconsistencies
- Handle Missing Values
  - Decide how to handle missing values
- Remove Duplicates
  - Identify and remove duplicate records to ensure data integrity. This typically involves comparing records based on key attributes and deleting duplicates while retaining one instance of each unique record.

Starting with Questions

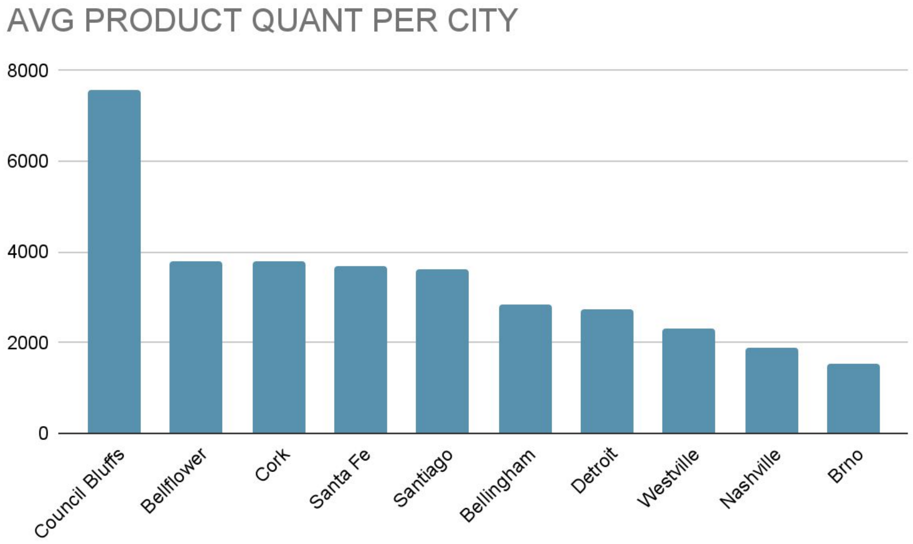
Question 1: Which cities and countries have the highest level of transaction revenues on the site?

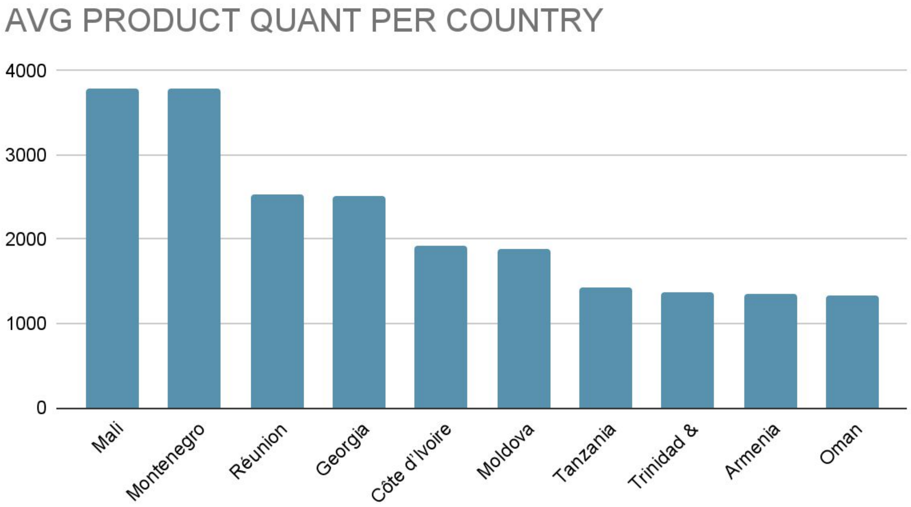




Starting with Questions

Question 2: What is the average number of products ordered from visitors in each city and country?

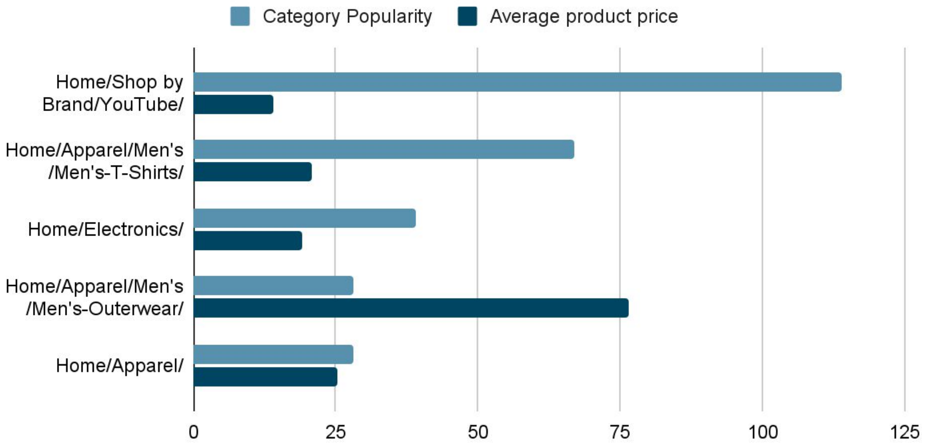




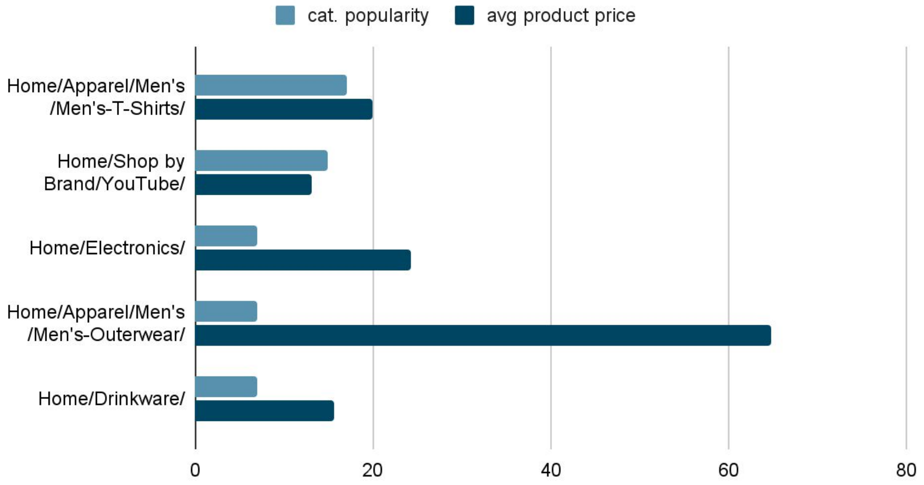
Starting with Questions

Question 3: Is there any pattern in the types (product categories) of products ordered from visitors in each city and country?

Product popularity vs Avg Product price (Where country = canada)



Cat. popularity vs Avg Product price (Where city= toronto)



Starting with Questions

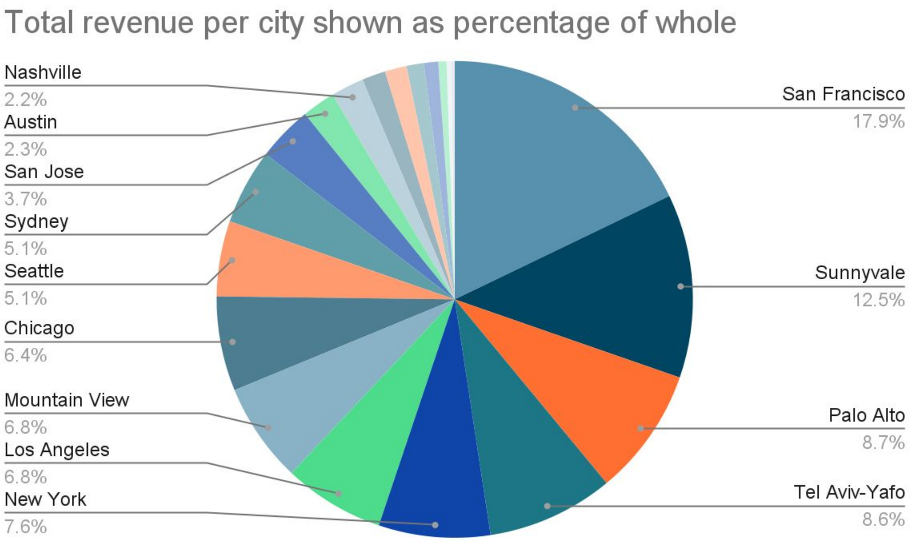
Question 4: What is the top-selling product from each city/country? Can we find any pattern worthy of noting in the products sold?

Mountain View	Indoor Security Camera	22
New York	100% Cotton Short Sleeve Hero Tee White	13
San Francisco	100% Cotton Short Sleeve Hero Tee White	10
London	YouTube Twill Cap	9
Palo Alto	Outdoor Security Camera	7

United States	100% Cotton Short Sleeve Hero Tee White	148
United Kingdom	22 oz YouTube Bottle Infuser	21
India	YouTube Custom Decals	21
Germany	YouTube Twill Cap	13
Canada	22 oz YouTube Bottle Infuser	13

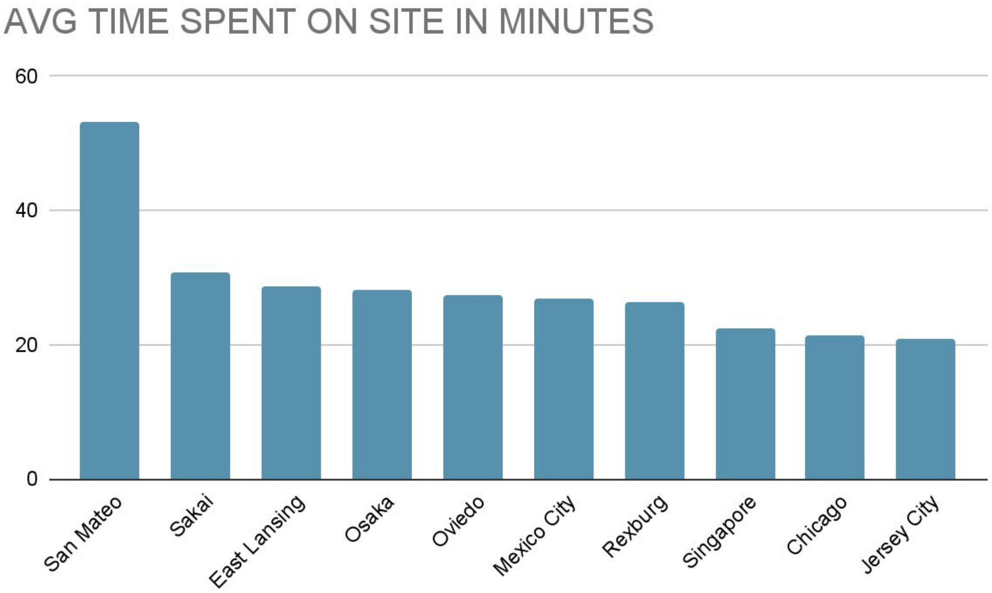
Starting with Questions

Question 5: Can we summarize the impact of revenue generated from each city/country?



Starting with Data

My Question 1: Which city country combo has the highest avg time on a website?



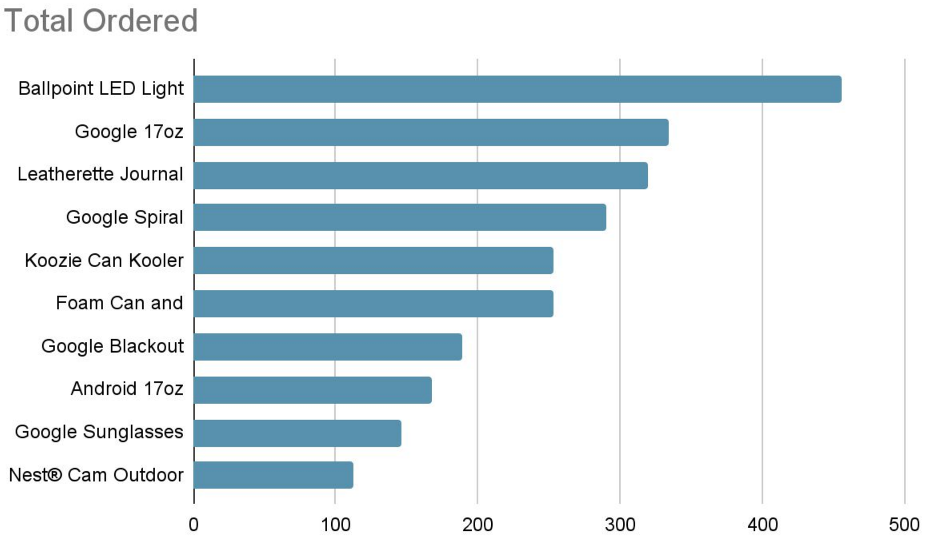
Starting with Data

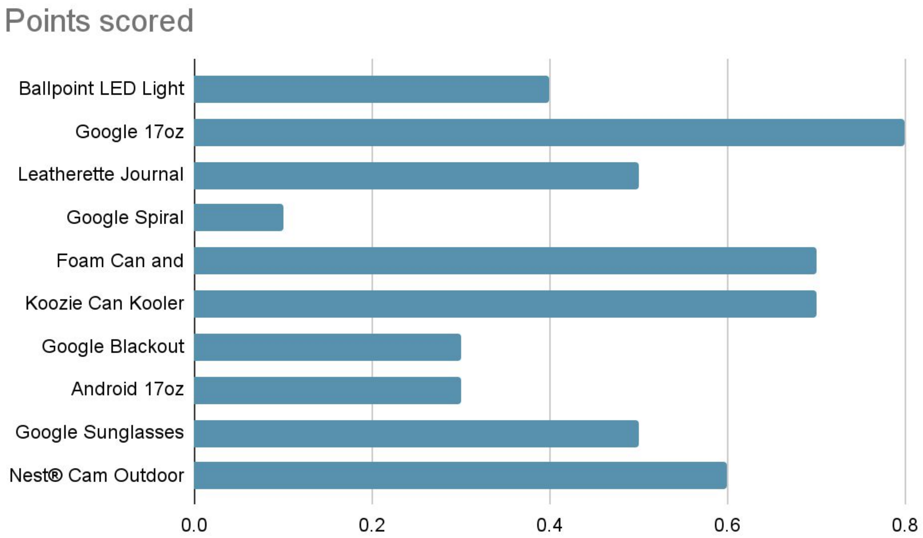
My Question 2: Which CITY COUNTRY COMBO had the highest sentiment score FOR WHICH PRODUCT?

Fremont	United States	Google Stylus Pen w/ LED Light	0.9
Istanbul	Turkey	Google G Noise-reducing Bluetooth Headphones	0.9
Chennai	India	Google Hard Cover Journal	0.9
Edmonton	Canada	Google Stylus Pen w/ LED Light	0.9
Chicago	United States	Google Pocket Bluetooth Speaker	0.9
Chicago	United States	Google G Noise-reducing Bluetooth Headphones	0.9
Austin	United States	Google G Noise-reducing Bluetooth Headphones	0.9
Ann Arbor	United States	Metal Texture Roller Pen	0.9
Detroit	United States	Google 22 oz Water Bottle	0.9
Mountain View	United States	Google G Noise-reducing Bluetooth Headphones	0.9

Starting with Data

Question 3: How does the total amount ordered relate to its sentiment score?





QA

Risk areas

- Performance
- Data integrity

Process

- Check for and address dupes for all tables in columns that should have unique entries
- Set up relationships between tables and ensure proper key constraints
- Check that naming conventions for non data in non null columns follow the same rules
- Check that data is stored on a proper scale
- Double check that all queries run in a reasonable time and return the expected output

Thank you