**Assignment 2   
CS2005 -Database Systems**



**Submitted by:**Mishal Fatima [23i-0776]  
Aleena Babar [23i-0628]

Contents

[Data Uploading Process 3](#_Toc193547787)

[1. Database Creation 3](#_Toc193547788)

[2. Importing the CSV File 3](#_Toc193547789)

[Key Uniqueness Enforcement 3](#_Toc193547790)

[1. Identifying Duplicates 3](#_Toc193547791)

[2. Modifying the Primary Key 4](#_Toc193547792)

[3. Ensuring Referential Integrity 4](#_Toc193547793)

[Data Cleaning 4](#_Toc193547794)

[Queries 5](#_Toc193547795)

[1. Order Analysis 5](#_Toc193547796)

[2. Product Analysis 6](#_Toc193547797)

# Data Uploading Process

To upload the dataset into the database, I followed the steps below using SQL Server Management Studio (SSMS):

## **Database Creation**

Before importing the CSV file, I created a new database named BrazilianEcommerce using the following SQL command:

sql

CREATE DATABASE BrazilianEcommerce;

I then ensured that the database was selected for importing the data.

## **Importing the CSV File**

The CSV files were imported using the "Import Flat File" feature in SSMS. The steps involved were:

* + **Navigating to the Import Option**:  
    In SSMS, I selected the database, navigated to *Tasks → Import Flat File…*
  + **Selecting the File**:  
    I browsed and selected the CSV files.
  + **Configuring Column Mapping**:  
    SSMS automatically detected column names and data types, which I reviewed to ensure accuracy.
  + **Finalizing the Import**:  
    I executed the import process, which created a new table and loaded the data.

# Key Uniqueness Enforcement

After importing the data, the dataset was analyzed and it was found that the **primary key for the geolocation table was not unique**. To resolve this issue while maintaining referential integrity, the following modifications were made:

1. Identifying Duplicates  
   The following SQL query was used to check for duplicate values in the geolocation\_id column:

sql

SELECT geolocation\_id, COUNT(\*)

FROM geolocation

GROUP BY geolocation\_id

HAVING COUNT(\*) > 1;

The results confirmed that duplicate entries existed.

1. Modifying the Primary Key  
   To ensure uniqueness, two possible solutions were considered:
   * Using a **composite key**: Combining multiple columns (e.g., latitude and longitude) to create a unique identifier.  
       
     But it turned out that even after composing longitude and latitude there were still many duplicate values being found, hence the idea was dropped.
   * Adding an **auto-incrementing ID column** composed with the existing zip-code prefix column as the new primary key.

It was decided to use the second approach, as it better represented unique geolocation points. I modified the table schema as follows:

sql

ALTER TABLE olist\_geolocation

ADD id INT IDENTITY(1,1);

ALTER TABLE olist\_geolocation

ADD CONSTRAINT PK\_olist\_geolocation PRIMARY KEY (id, geolocation\_zip\_code\_prefix);

This ensured that each geolocation was uniquely identified while preserving existing foreign key relationships.

1. Ensuring Referential Integrity  
   Since the primary key change affected related tables, It was verified that foreign keys referencing the geolocation table were updated accordingly.

# Data Cleaning

~~While importing the data, an observation was made that one of the dataset tables had some redundant columns, namely the olist\_customers table. There were two unique id columns, one was the cutomerid, and the other was customer\_unique\_id. Hence, a decision was made to remove the second column.~~

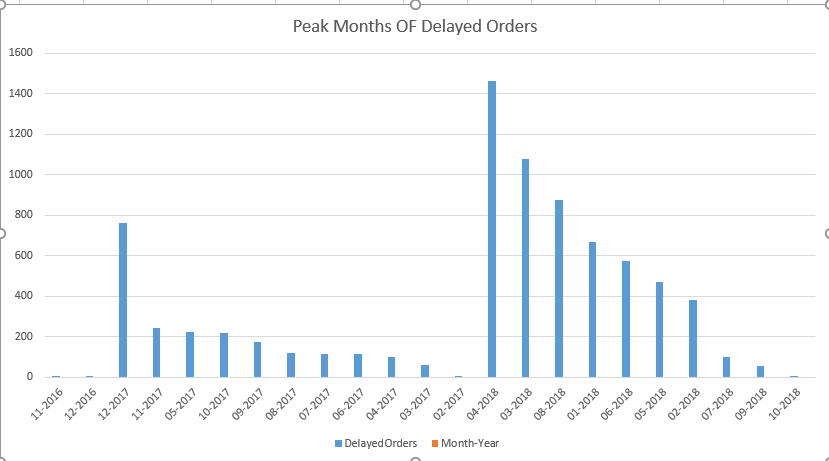
# Queries

## Order Analysis

1. **Percentage of orders that got delayed beyond the estimated date.**

The percentage of orders that got delayed beyond the estimated date was 7.87% (rounded off to two decimal places).

1. **What are the peak months of order delays?**

The peak months for 2018 were: April, March.  
The peak months for 2017 were: December, November.  
The peak months for 2016 were: December, November.  
  
Overall highest number of delayed orders was observed in April 2018.

1. **Which state experiences the highest order delays?**

*Rio de Janeiro (****RJ)*** state experiences the highest order delays.

1. **See how many orders are still in “pending” status for each year.**

None.

1. **What is the average delay duration per seller?**

- The seller with the highest average delay had an average delay of **167** days.

* The overall average delay across all sellers was **8** days.
* **334** sellers had an average delay of more than 10 days, indicating that certain sellers consistently struggled with on-time delivery.

1. **How do shipping costs impact order delays?**

The average shipping cost for delayed orders was: $22.11  
The average shipping cost for On Time orders was: $19.76  
  
**This shows that orders with higher shipping costs tend to experience more delays.** This could indicate that expensive or complex shipments (possibly involving larger distances) are more prone to logistic challenges, resulting in delayed deliveries.

1. **Which product category experience the most order delays.**

Orders involving products from Bed Bath Table **tend to experience the most delays.**

1. **How do number of items per order affect the delays?**

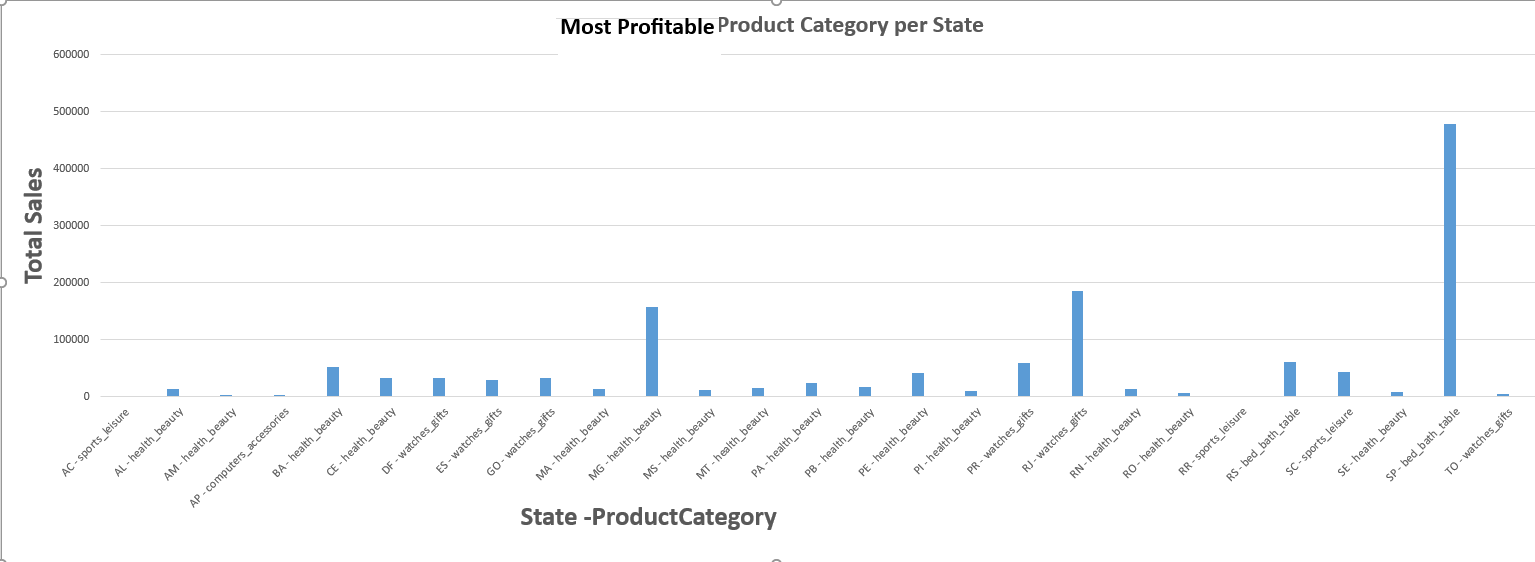
The average amount of items for delayed order is: 1  
The average amount of items for delayed order is: 1  
  
**This shows that the number of items in an order does not have a significant impact on delivery delays.** Most delayed and on-time orders contain only a single item, suggesting that other factors — such as shipping logistics, product category, or location — may play a larger role in delays.

## Product Analysis

1. **What is the most profitable product category per state? Find the most profitable product category in each state based on total sales.**

For each state, the most profitable product category was identified by summing total sales.

* For example, in *São Paulo (SP)*, the most profitable category was **Bed Bath Table**.
* In *Rio de Janeiro (RJ)*, the top category was **Watches gifts**.
* In *Minas Gerais (MG)*, the top category was **Health Beauty**.

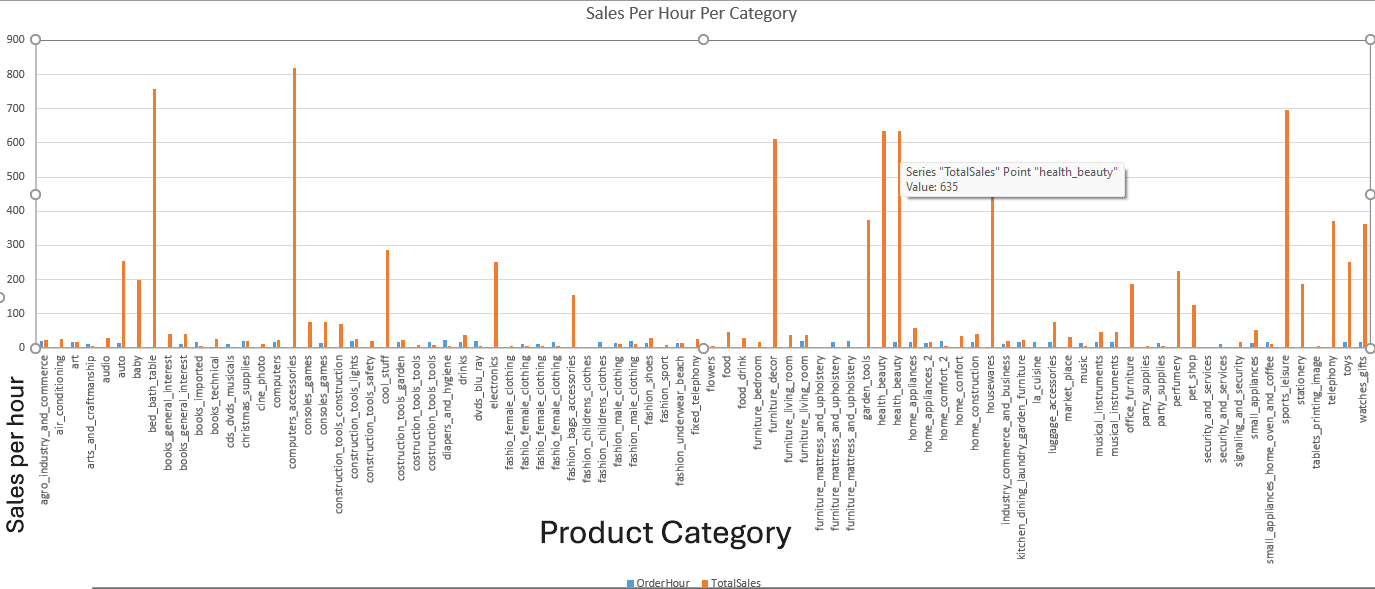
The analysis highlights regional preferences and indicates that categories like **Health Beauty, and Watches gifts** are consistently among the top performers across multiple states.

1. **What are the peak hours for order placements per product category? Find which hours of the day have the highest number of order placements for each product category.**

The peak order placement hours vary across product categories.

* For **Health Beauty**, the highest order placements occurred at **4 PM**.
* For **bed bath table**, peak ordering happened around **2 AM**.
* Watches gifts saw most orders placed around **4 PM**.

Overall, most categories had peak order placements during mid-morning and early evening hours, indicating active shopping times before and after typical working hours.



1. **Which product categories experience the most delays? Find the top 5 product categories with the highest number of delayed orders.**

The top 5 categories are as follows: **Bed Bath Table** with 920 delayed orders.  
**Health Beauty** with 858 delayed orders.  
**Furniture Décor** with 688 delayed orders.  
**Sports Leisure** with 625 delayed orders.  
**Computers Accessories** with 594 delayed orders.

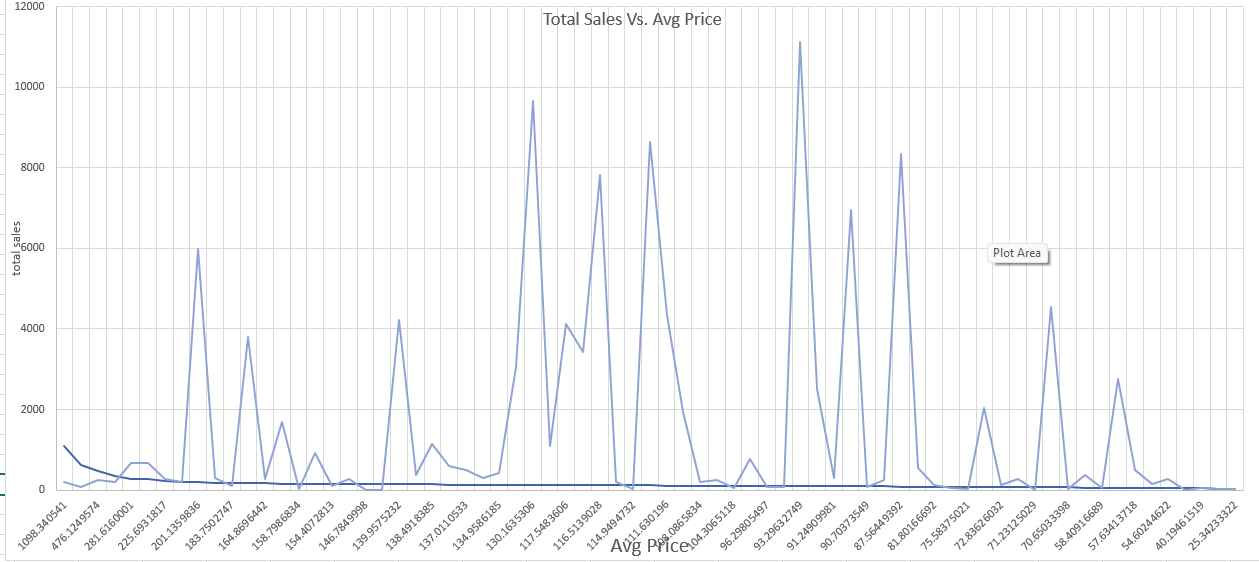
1. **What is the impact of product price on sales volume? Find whether higher-priced products sell less by comparing average price vs. total sales.**

**Negative correlation**: There is a clear negative correlation between average price and sales quantity. Higher-priced products generally sell in lower quantities.

**Price sensitivity varies by category**:   
-Low-priced categories like Bed Bath Table (avg. $93.30) have high sales (11,115 units)  
-High-priced categories like Musical Instruments (avg. $281.62) have lower sales (680 units)

**Sweet spot pricing**: The data suggests a "sweet spot" exists where moderately priced products ($90-130) achieve substantial sales volume.

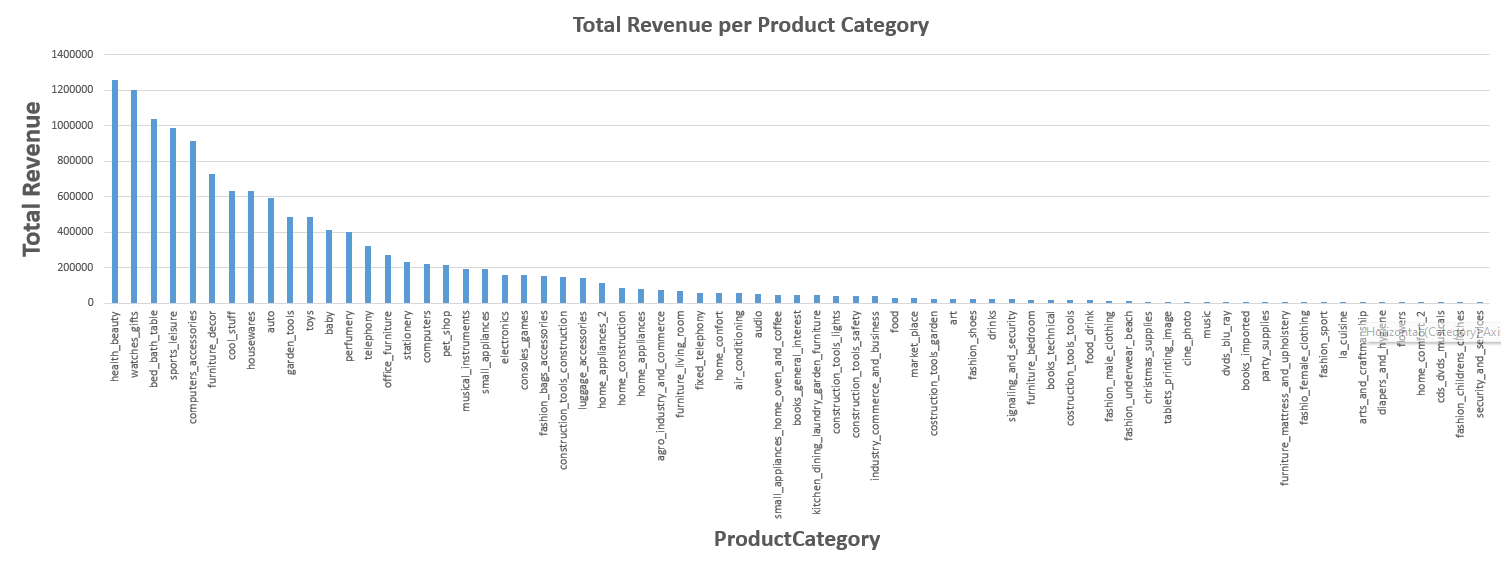
**Technology exception**: Computer Accessories maintains high sales (7,827 units) despite a moderate price point ($116.51), suggesting less price sensitivity in tech categories.

****

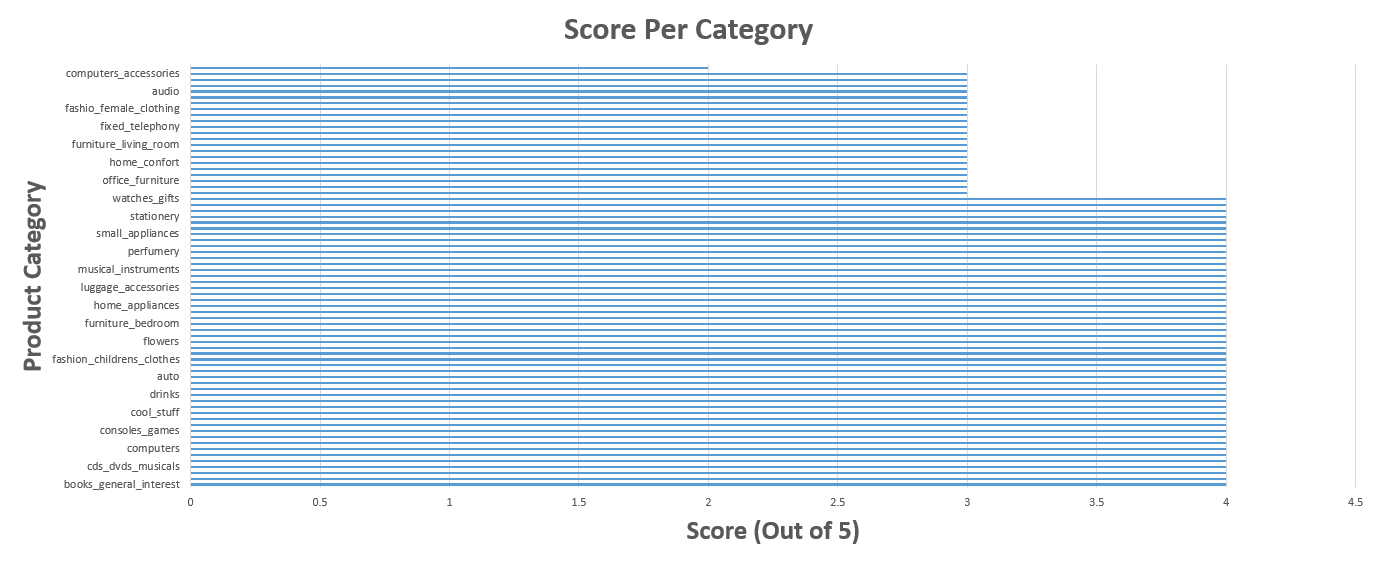
1. **Which products are most frequently bought together? Find the most frequently purchased product pairs.**

Due to the dataset size and complexity, identifying the most frequently purchased product pairs would require advanced self-joins and large computational resources.  
However, based on overall product frequency analysis, it was observed that categories like **furniture decor**, **bed bath table**, and **watches gifts** are among the most frequently purchased products, and are likely to appear together in orders containing multiple items.  
This suggests that home decor and personal gift items are popular combinations among customers.

1. **Calculate the total revenue per product category by summing order item prices.**

****

1. **Compute the average review score for each product category.**

****

1. **Retrieve the top 5 products based on total sales revenue.**

The top five products based on sales revenue are as following:  
1. Health Beauty with 63885 sales revenue.  
2. Computers with 48899.3 sales revenue.  
3. Computer Accessories with 47214.5 sales revenue.  
4. Bed Bath Table with 43025.5 sales revenue.  
5. Baby with 38907.3 sales revenue