Activity #1 — Data Mining Scenario Challenge

Presented to the Faculty of College of Computer and Information Sciences

Polytechnic University of the Philippines

Sta. Mesa Manila

In Partial Fulfillment of the Requirements for the

Subject IT Elective 2 (Data Science)

By:

Kenzo C. Ragundiaz

Debbie Kaye Alagao

Trisha Mae Lasanas

Angeli Rivera

Presented to:

Prof. Severino M. Bedis Jr.

March 16, 2025

Activity #1: Data Mining Scenario Challenge

Objective: To reinforce understanding of data mining techniques and their application to different data types by developing realistic data mining scenarios and proposing suitable solutions.

Instructions:

1. **Form Groups:** Divide into groups of 4-5 students.

2. **Data Type Assignment:** Each group will be assigned *one* of the following data types:

   - Database Data
   - Data Warehouses
   - Transactional Data
   - Multimedia Data
   - Spatial Data
   - Time Series Data
   - World Wide Web Data

   Answer: The data type that has been assigned to us is "**World Wide Web Data**."

3. **Scenario Creation:** Your group's task is to invent a realistic scenario where the assigned data type is used by an organization to solve a problem or achieve a goal. The scenario should be detailed and specific. Consider the following:
   - **Industry/Sector:** What type of organization is using the data? (e.g., Retail, Healthcare, Finance, Education, Government, etc.)
   - **Problem/Goal:** What specific problem is the organization trying to solve, or what goal are they trying to achieve? (e.g., Increase sales, reduce costs, improve customer satisfaction, detect fraud, optimize resource allocation, etc.)
   - **Data Description:** Describe the specific data available to the organization. What are the key attributes or features of the data? How is the data collected and stored?

   Answer:

   Fictitious Organization: Move2Go Logistics Inc.

   Move2Go Logistics Inc. (fictitious company as suggested by the instructions to "invent a realistic scenario") is a blooming supply chain company that uses World Wide Web Data to further enhance its services to its customers worldwide.

Data Set

The data set gathered for this group activity is from Move2Go Logistics Inc.'s data, which was gathered from the web. Move2Go Logistics' data set includes the supply chains data used by itself, showcasing the types of payment, sales per customer, countries, product categories, products bought, shipping mode, and country origin of the product from the 53-column-sized data set. The data set covers the operations within Move2Go, which is beneficial for solving its business problems and predicting future user interactions and trends.

Problems

Move2Go Logistics Inc. is facing uncertainty where other departments experience an abundance of sales while others are struggling to gain sales equal to the other departments.

Move2Go Logistics Inc. is also wondering what type of products are being bought the most, what products are mostly left untouched, which countries are their most loyal customers, and which countries don't subscribe to their shipping modes, choosing only the standard class.

Move2Go Logistics Inc. is unsure if its customers are just consumers, belong to a corporation, or are just people working from home — unsure on how they can create innovations to reality to refurbish their products.

Goal

With the problems stated, Move2Go Logistics Inc. wishes to solve these dilemmas to improve its services, gain trust, gain customers, and earn more profit to further assist people worldwide.

Move2Go Logistics Inc. has an idea where it plans on using Power BI to preprocess the data, visualize, and create a dashboard showing the patterns and trends of the worldwide web data set gathered by creating charts and graphs for the data visualizations.

With the assistance of Power BI, Move2Go can clean its data to check specific areas where the business is booming and to investigate the fields where it needs to improve its services. Based on the visualizations in the dashboard, Move2Go can reflect

and take immediate actions on the spotted weaknesses in its business to maintain a well-balanced and successful business.

Data Description

The World Wide Web (WWW) data gathered is structured data that is used by supply chain companies. The data set has both numerical data (quantitative data) and categorical data (qualitative data). The quantitative data revolves around the sales, the customer ID, the days for shipping, the customer coordinates, etc. The qualitative data includes the product bought, customer category, type of shipping mode, department from which the product was bought, country, etc.

The structured data will be beneficial to the company to find its weaknesses, and the strengths to further harness.

The data set was gathered from the web as it is a World Wide Web (WWW) data. We found out that this data set has the potential to expand, and strengthen the market focusing on promoting products to the right people.

Why Was the Data Collected?

The data was collected to analyze customer behavior, product demand, shipping preferences, and overall business performance. It aims to identify trends, address disparities in sales among departments, and improve decision-making based on data-driven insights.

4. **Data Mining Solution:** For your scenario, propose a data mining solution that would help the organization achieve its goal. The solution should include:
   ○ **Data Mining Techniques:** What specific data mining techniques would be most appropriate for this scenario? (e.g., Clustering, Classification, Regression, Association Rule Mining, Anomaly Detection, Time Series Analysis, Text Mining, Image Recognition, etc.) Justify your selection.
   ○ **Expected Outcome:** What results would you expect to obtain from the data mining process? What specific insights or knowledge would be revealed?
   ○ **Benefits:** How would the results of the data mining project benefit the organization? How would it help them solve their problem or achieve their goal?
   ○ **Potential Challenges:** What potential challenges or limitations might you encounter when implementing the data mining solution? (e.g., data quality issues, scalability problems, privacy concerns, ethical considerations, etc.)

Data Mining Techniques

Move2Go Logistics Inc. will use various data mining techniques, including:
- **Clustering (K-Means, DBSCAN)**
  - Segment departments based on sales performance (high-performing vs. low-performing).
  - Group products based on purchase frequency to identify popular and unpopular items.
  - Categorize customers into different profiles (consumers, corporations, work-from-home individuals).
  - Identify countries with high and low engagement based on purchase behavior.
- **Regression Analysis (Multiple Linear Regression)**
  - Analyze factors influencing uneven sales across departments (e.g., product category, shipping mode, customer type).
  - Evaluate customer preference for standard vs. other shipping modes based on available attributes.
- **Association Rule Mining (Apriori, FP-Growth)**
  - Identify frequent product purchase patterns to determine which products are often bought together and which are left unsold.
  - Discover customer behavior trends related to shipping mode preferences.
  - Uncover relationships between customer profiles and product categories to enhance targeted marketing strategies.

Expected Outcome

- **Identifies performance gaps among departments**, enabling strategic resource allocation and operational improvements.
- **Predicts customer demand and purchasing trends**, supporting better inventory management and business planning.
- **Reveals purchasing patterns and customer preferences**, allowing for optimized product bundling, marketing strategies, and improved customer engagement.

Benefits

- Uses data-driven insights to support better decision-making.

- Strengthens product promotion and sales strategies.
- Allocates resources efficiently to struggling departments.
- Improves inventory management by preventing shortages and overstocking.
- Boosts customer satisfaction with personalized product recommendations.

By leveraging these insights, Move2Go can refine its marketing strategies, improve operational efficiency, and increase revenue. The ability to predict trends and adjust business strategies accordingly will enhance customer retention, streamline logistics operations, and create a sustainable growth path for the company.

Potential Challenges

- **Data Quality Issues**

    Move2Go might encounter datasets that may contain missing values, inconsistent data, and errors that might affect the result of what the organization is trying to analyze, deeming it an inaccurate and unreliable insight.

- **Scalability Problem**

    Since Move2Go is available to service customers around the world, the data would likely grow rapidly as the organization expands. Millions of data transactions may cause an issue by straining computational resources used and slowing down the processing of transactions.

- **Privacy Concerns**

    The organization collects sensitive personal information from the customers to analyze their preferences and what they might like within the web. However, this may raise privacy concerns with regulations like GDPR.

- **Ethical Considerations**

    Personalization and market targeting using customer data might also be an issue, as it raises ethical questions about manipulation in buying more products on the platform.

- **Results Interpretability**

    Using certain data mining techniques like clustering and regression to make up insights may produce results that are not easy to interpret. The results should be understood by the stakeholders so the can act on the produced results without reluctance.

IS IT COSTLY?

Yes, having to implement a data mining solution can be costly for Move2Go Logistics Inc. It would need to set up infrastructure (e.g., licenses, cloud platform), hire professionals like data scientists and analysts to produce improved insights and decision-making, and maintain the cleaning of missing values, errors, etc. All of that is mentioned would be costly, but it will for sure be for the improvement of the operations that would surely benefit the company in the long run.

5. **Presentation Preparation:** Prepare a brief presentation (5-7 minutes) summarizing your scenario, proposed solution, and expected benefits. Include visuals (e.g., diagrams, mockups of dashboards) to illustrate your concepts.
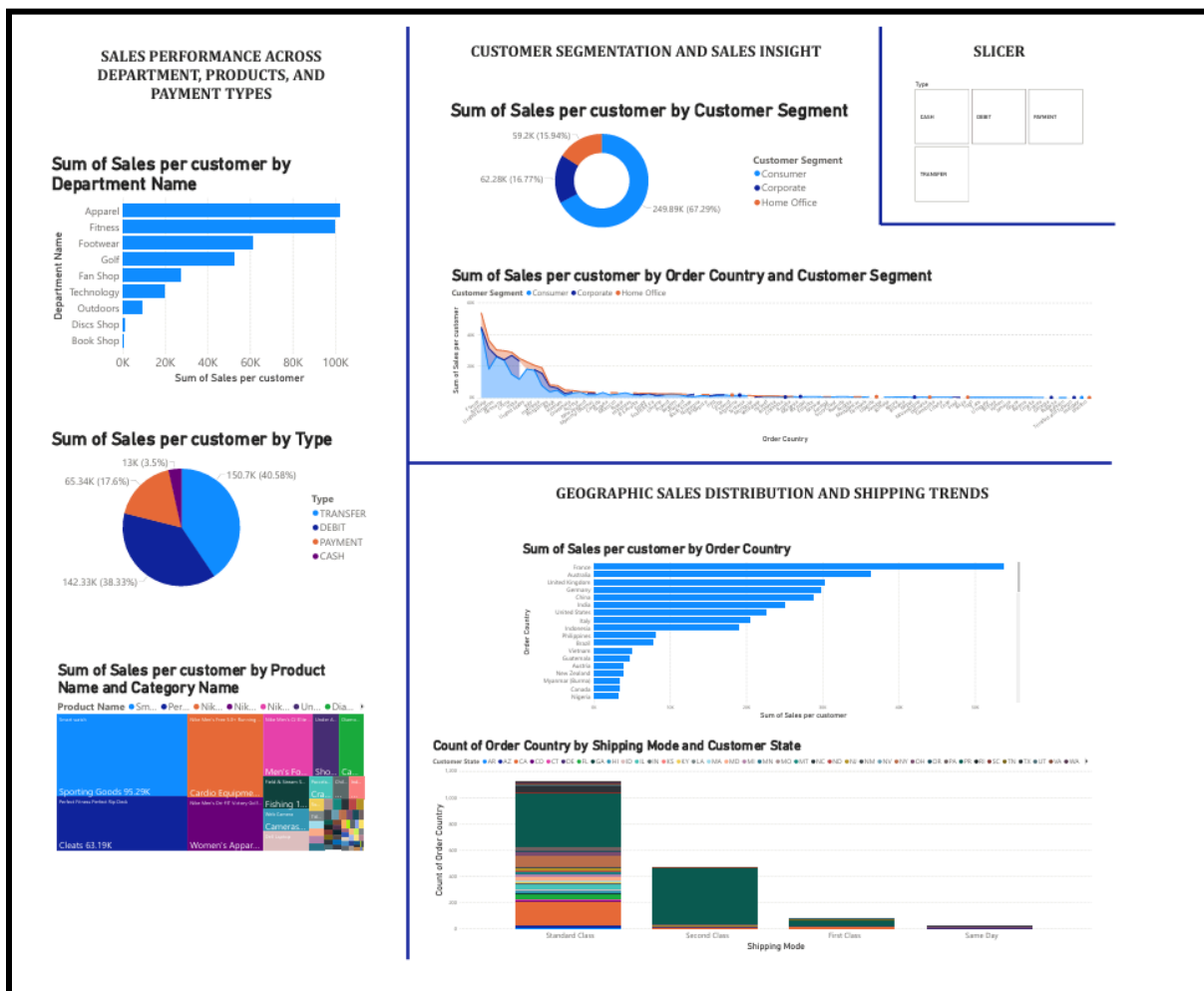
Figure 1: Dashboard

- Summary of the visualizations.
- Made with Power BI.
- Cleaned data set's graphs, and charts.
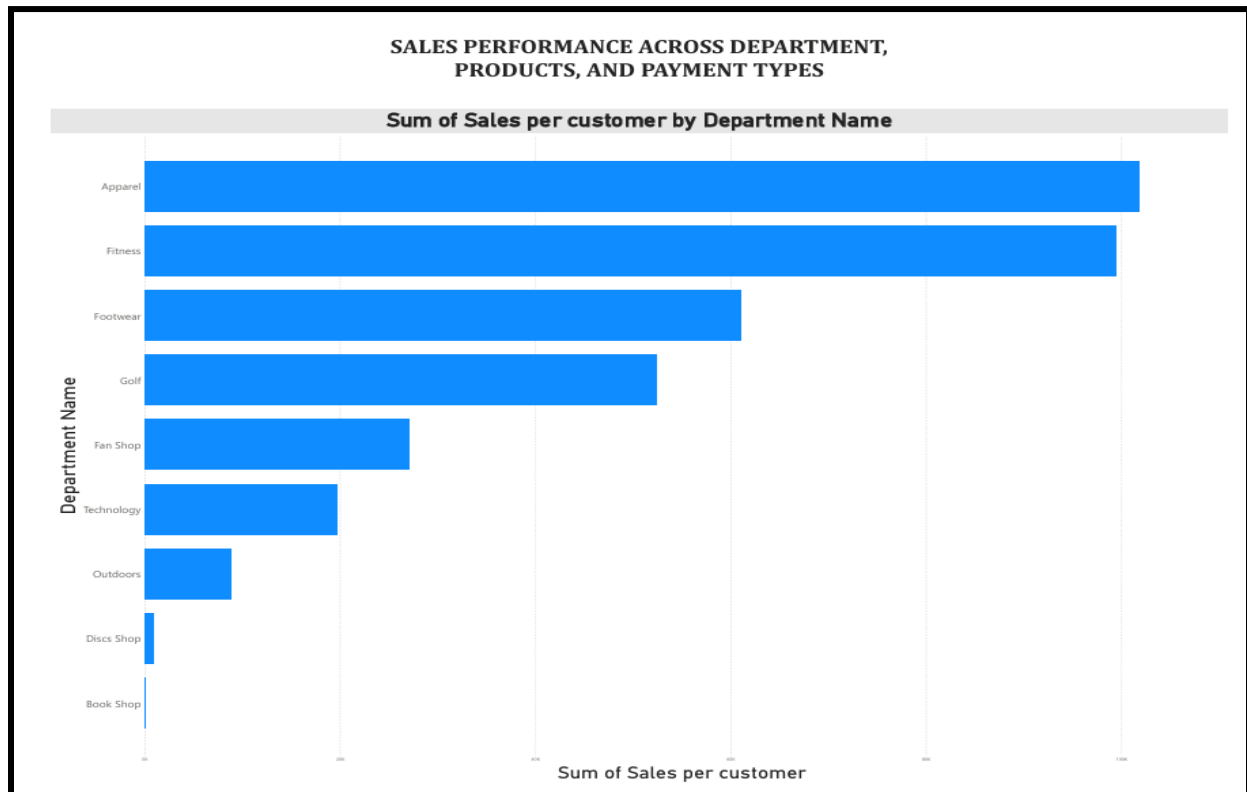- Has a slicer for easier data navigation.

Figure 2: Clustered Bar Chart: Sum of Sales per Customer by Department Name

- To determine which department has the most sales among customers worldwide.
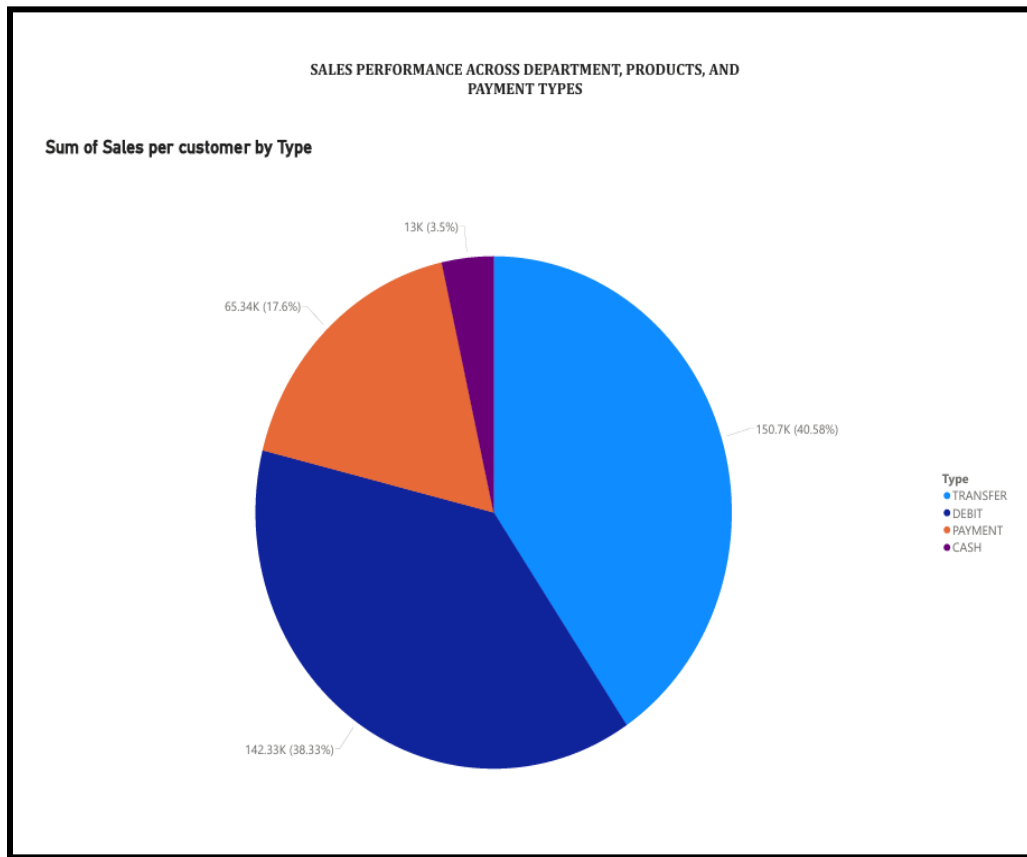- Department Name
- Sales per Customer

Figure 3: Pie Chart: Sum of Sales per Customer by Type

- To determine which type of payment is used frequently
- Sales per customer
- Type/mode of payment

Figure 4: Heat Map: Sum of Sales per Customer by Product Name and Category Name

- To determine what product has the most sales among products and Department name
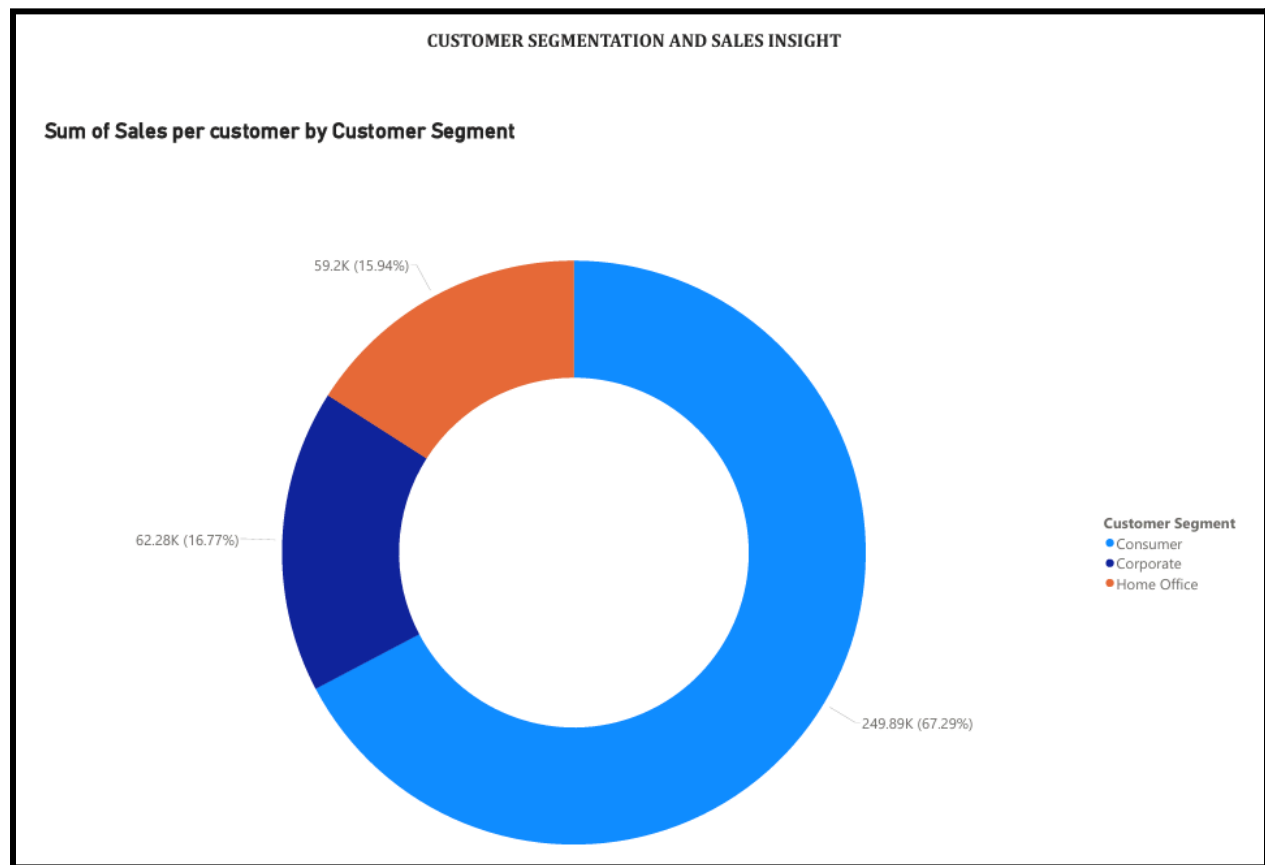- Product name
- Category name
- Sales per customer

Figure 5: Donut Chart: Sum of Sales Per Customer by Customer Segment

- To determine where the consumer belongs based on customer segment
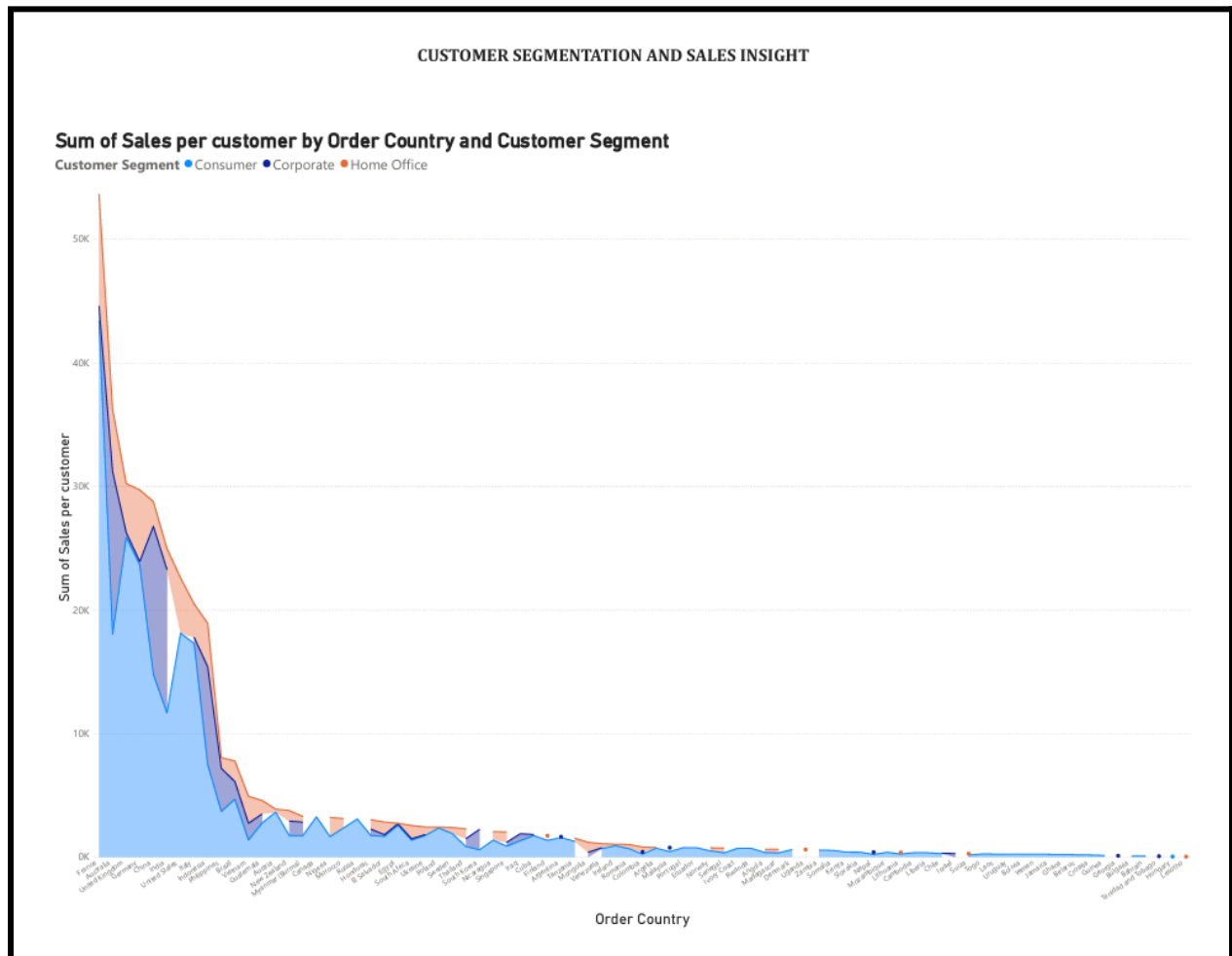- Customer Segment
- Sales per Customer

Figure 6: Stacked Area Chart: Sum of Sales Per Customer by Order Country and Customer Segment

- To determine the sum of sales by country and specified by customer segment
- Customer segment
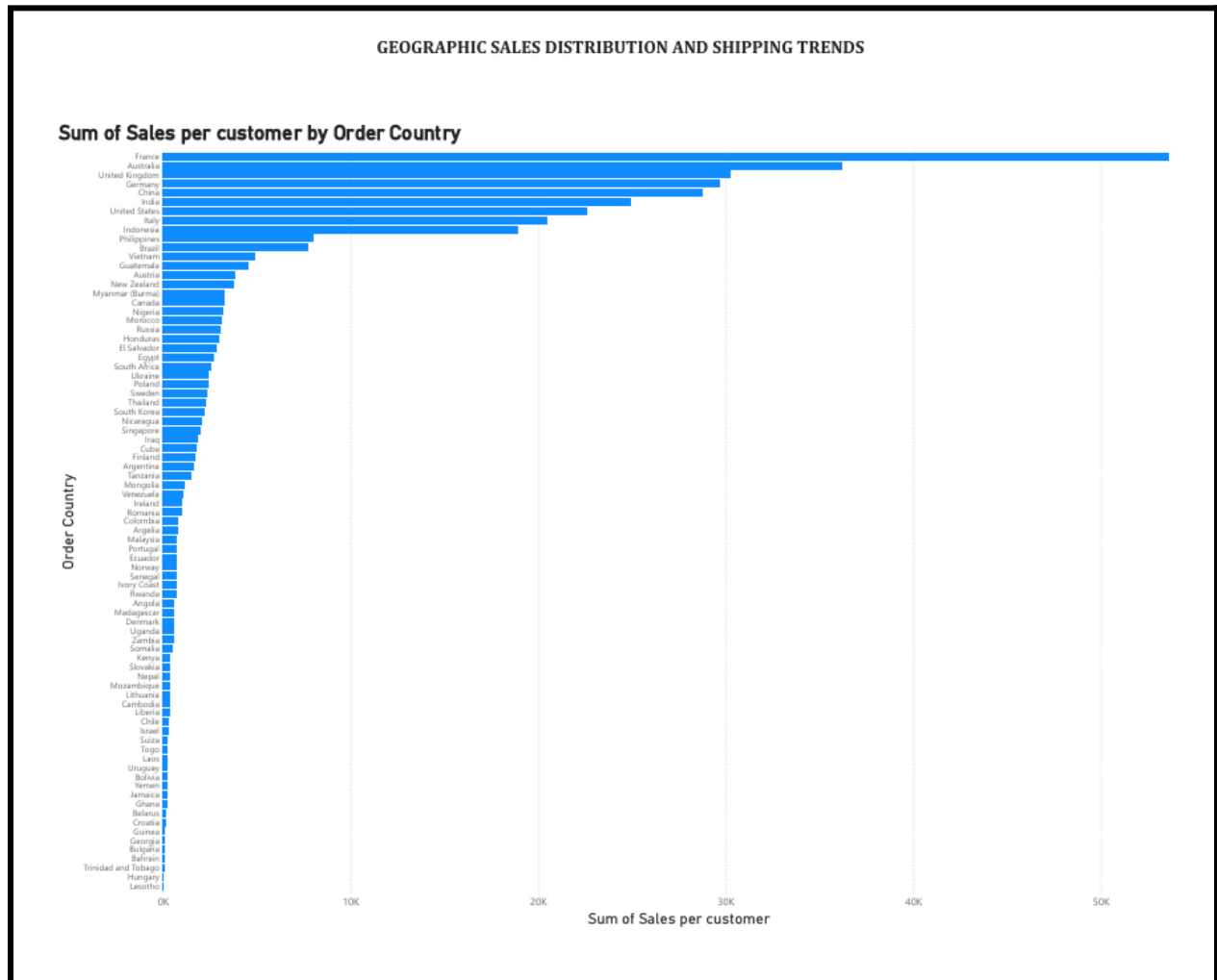- Order country
- Sales per customer

Figure 7: Stacked Bar Chart: Sum of Sales Per Customer by Order Country

- To determine which country has the most sales among the countries.
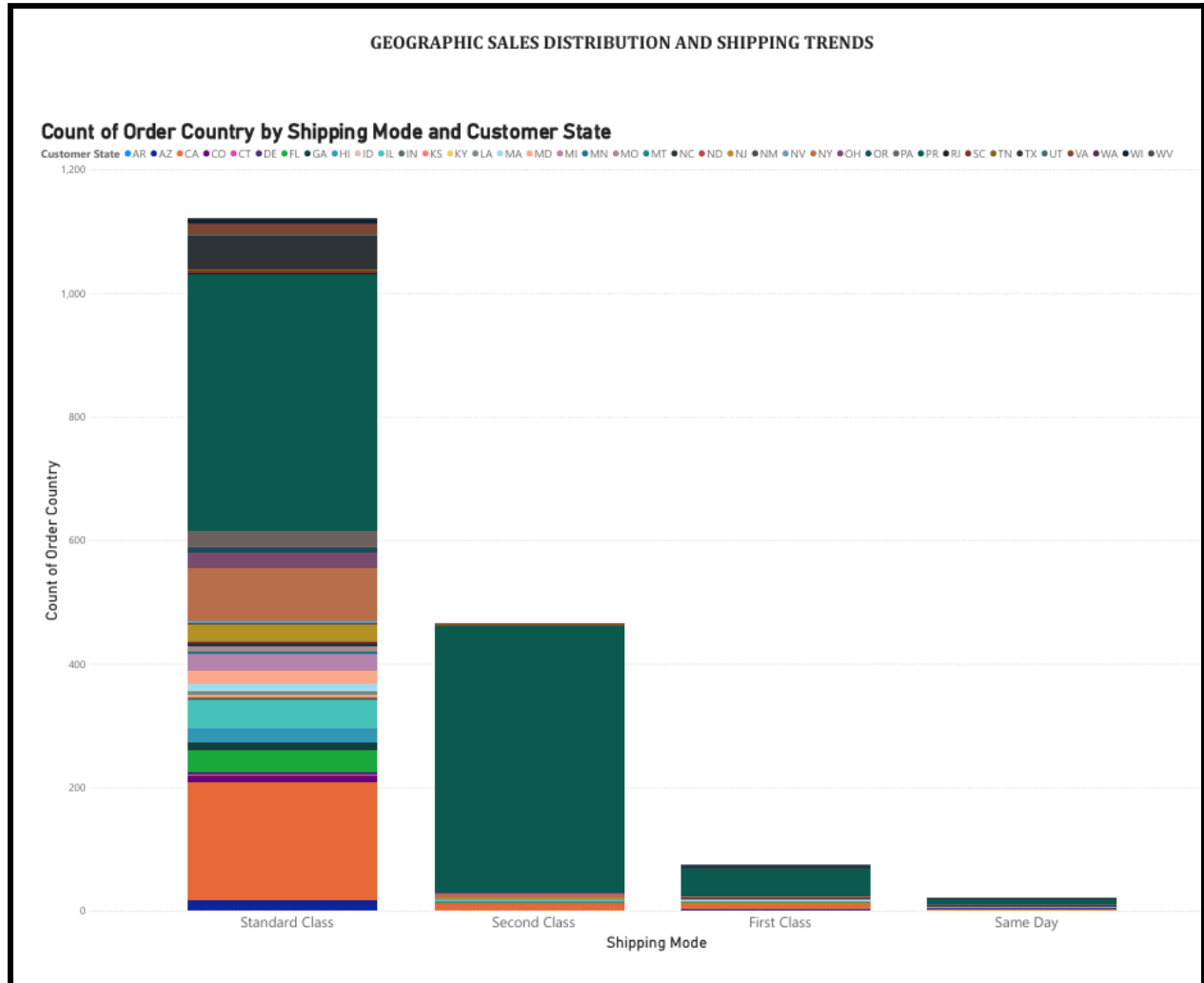- Order country
- sales per customer

Figure 8: Stacked Column Chart: Count of Order Country by Shipping Mode and Customer State

- To determine which country avails the most used shipping mode among standard class, second, first, and same day.
- Customer State
- Order country
- Shipping mode

Reference

Constante, F., Silva, F., & Pereira, A. (2019). DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS. *Mendeley Data*. https://doi.org/10.17632/8gx2fvg2k6.3

Grading Rubric:

- **Scenario Realism (25%):** Is the scenario realistic and plausible?
- **Solution Appropriateness (30%):** Are the proposed data mining techniques appropriate for the scenario and the data available? Is the justification for the selection clear and convincing?
- **Expected Benefits (20%):** Are the expected benefits clearly defined and realistically achievable?
- **Presentation Clarity (15%):** Is the presentation well-organized, clear, and engaging? Are visuals used effectively?
- **Consideration of Challenges (10%):** Does the presentation acknowledge potential challenges and limitations?

Example Scenario (for Database Data):

- **Industry:** Banking
- **Problem:** High customer churn rate (customers closing their accounts).
- **Data:** Customer database with demographics, account balances, transaction history, customer service interactions, and product usage.

- **Data Mining Solution:**
  - **Technique:** Classification (e.g., decision trees, logistic regression) to predict which customers are most likely to churn.
  - **Expected Outcome:** A model that identifies key factors contributing to churn, such as low account balances, infrequent transactions, and negative customer service interactions.
  - **Benefits:** The bank can target at-risk customers with personalized retention offers (e.g., fee waivers, bonus rewards) to reduce churn.
  - **Challenges:** Data privacy concerns, ensuring fairness and avoiding bias in the churn prediction model.

Key Considerations:

- **Focus on Application:** The goal is to apply the concepts of data mining to real-world situations, not to implement complex algorithms.
- **Encourage Creativity:** Encourage students to think creatively and develop innovative solutions.
- **Promote Collaboration:** Emphasize the importance of teamwork and collaboration.
- **Ethical Awareness:** Remind students to consider the ethical implications of data mining.