

CDS505

Assignment 1

Chuah Kheng Lim | P-COM0079/23

24th November 2023

Dataset Background

- Domain or Area where it was collected:
 - **Clothing Retail Sales**
- What is the intention for its collection:
 - **To provide comprehensive insights into consumers' preferences, tendencies, and patterns during their shopping experiences.**
- Who collected it (data owner), who are the intended users or stakeholders:
 - **User 'ZEE SOLVER' on Kaggle**
 - **Clothing Retail Owner**
 - **Fashion Designer**
- Any other relevant information about your dataset:
 - **This dataset encompasses a diverse range of variables, including demographic information of a retail store based in The United States**

Dataset Attributes

- Dataset type : Table
- 18 original attributes (+3 new attributes)
- List the attribute
 - **Customer ID:(ordered)** A unique identifier assigned to each customer.
 - **Age:(quantitative)** Demographic information indicating the customer's age.
 - **Gender:(categorical)** Identification of the customer's gender(e.g., male, female).
 - **Item Purchased:(categorical)** The specific product selected by the customer during the transaction.
 - **Category:(categorical)** Broad classification of the purchased item (e.g., clothing, electronics, groceries).
 - **Purchase Amount (USD):(quantitative)** Monetary value of the transaction, indicating the cost of purchased items.
 - **Location:(categorical)** Geographical location of the purchase (e.g., Kentucky, Texas, Arkansas, etc.).
 - **Size:(ordered)** Sizes of the purchased item, relevant for every purchases and categories (e.g., S, M, L, XL).
 - **Color:(categorical)** Color associated with the purchased item (e.g., Beige, Black, Blue, Pink, etc.).
 - **Season:(categorical)** Season relevance of the purchased item (e.g., spring, summer, fall, winter).
 - **Review Rating:(quantitative)** Customer's satisfaction assessment.
 - **Subscription Status:(categorical)** Indicates whether the customer has opted for a subscription service (e.g., Yes, No).
 - **Shipping Type:(categorical)** Method used to deliver the purchased item (e.g., standard shipping, express delivery).
 - **Discount Applied:(categorical)** Indicates if any promotional discounts were applied to the purchase (e.g., Yes, No).
 - **Promo Code Used:(categorical)** Notes whether a promotional code or coupon was utilized during the transaction (e.g., Yes, No).
 - **Previous Purchases:(quantitative)** Information on the number or frequency of prior purchases.
 - **Payment Method:(categorical)** Specifies the mode of payment employed by the customer (e.g., credit card, cash).
 - **Frequency of Purchases:(categorical)** Indicates how often the customer engages in purchasing activities (e.g., Biweekly, Fortnightly, Weekly, etc.).
 - **Rating Summary:(ordered)** Binning of Review Rating to provide summary of rating (e.g., Average, Good, Superb).
 - **Shopping Type:(categorical)** Recategorize of Shipping Type to indicate customer's buying method (e.g., Online, Offline).
 - **Customer Loyalty:(ordered)** Binning of Previous Purchases to indicate customer's frequent return (e.g., New, Regular, Loyal).

Attributes Cardinality

- Number of items: 3900
- For each attribute, indicate its cardinality:
 - **Customer ID:** 3900, preserve as ordered
 - **Age:** min:18 max:70 median:44
 - **Gender:** 1 Level, 2 distinct values
 - **Item Purchased:** 1 Level, 25 distinct values
 - **Category:** 1 Level, 4 distinct values
 - **Purchase Amount (USD):** min:20 max:100 median:60
 - **Location:** 1 Level, 50 distinct values
 - **Size:** 1 Level, 4 distinct values
 - **Color:** 1 Level, 25 distinct values
 - **Season:** 1 Level, 4 distinct values
 - **Review Rating:** min:2.5 max:5.0 average:3.75
 - **Subscription Status:** 1 Level, 2 distinct values
 - **Shipping Type:** 1 Level, 6 distinct values
 - **Discount Applied:** 1 Level, 2 distinct values
 - **Promo Code Used:** 1 Level, 2 distinct values
 - **Previous Purchases:** min:1 max:50 median:25
 - **Payment Method:** 1 Level, 6 distinct values
 - **Frequency of Purchases:** 1 Level, 7 distinct values
 - **Rating Summary:** 1 Level, 3 distinct values
 - **Shopping Type:** 1 Level, 2 distinct values
 - **Customer Loyalty:** 1 Level, 3 distinct values

Clothing Retail Business Owner's Questions

- **Question 1** : How many customers exhibit frequent purchase but do not own membership?
- **Question 2** : What are the most frequently bought sizes across 2 genders in every season?
- **Question 3** : What are the shopping preferences among customers from different states?

Q1: How many customers exhibit frequent purchase but do not own membership?

- Do you need a chart?
 - Yes, Pie Chart is required to answer.
- List of attributes do you need to use to answer the question:
 - Previous Purchases, Subscription Status, Customer Loyalty
- Transform data? If Yes – provide short description, what data produced:
 - Yes, Perform binning on Previous Purchases and calculate the fractions within the Subscription Status.
- Dataset or attribute type changed? If Yes – list them, what type changed:
 - Dataset type does not change, but Binning of Previous Purchases into Ordinal attribute is required to produce better result.
- All data? Additional data? If Yes – list and provide description:
 - No, Customer Loyalty attribute is required to answer this question therefore it will be derived from Previous Purchases.

Q1: How many customers exhibit frequent purchase but do not own membership?

- Provide high-level task abstraction for your dataset:
 - To consume existing data and deriving new data.
 - To discover and compare the membership percentages among frequent and infrequent customers
- List targets or aspect of the data that would answer your analysis question:
 - The percentages of memberships among 3 frequent types of customers

Q2: What are the most frequently bought sizes across 2 genders in every season?

- Do you need a chart?
 - Yes, Bar Chart is required to answer.
- List of attributes do you need to use to answer the question:
 - Season, Size, Gender
- Transform data? If Yes – provide short description, what data produced:
 - Yes, sum up the sizes count across each season then filter by genders.
- Dataset or attribute type changed? If Yes – list them, what type changed:
 - Dataset type and attribute type do not change.
- All data? Additional data? If Yes – list and provide description:
 - All data is provided in the table.

Q2: What are the most frequently bought sizes across 2 genders in every season?

- Provide high-level task abstraction for your dataset:
 - To consume existing data.
 - to discover the size preferences of each category for both genders over seasonal trends
- List targets or aspect of the data that would answer your analysis question:
 - the distribution of seasonal sales of different sizes in both genders

Q3: What are the shopping preferences among customers from different states?

- Do you need a chart?
 - Yes, Map Chart is required to answer.
- List of attributes do you need to use to answer the question:
 - Shipping Type, Location, Shopping Type, Purchase Amount
- Transform data? If Yes – provide short description, what data produced:
 - Yes, perform recategorize on Shipping Type, sum up the profit based on different locations.
- Dataset or attribute type changed? If Yes – list them, what type changed:
 - Dataset type does not change, though Location attribute type will change to Geo, and recoding of attribute values is required to produce sensible result.
- All data? Additional data? If Yes – list and provide description:
 - No, a new Shopping Method attribute is required, therefore Shipping Type will be recoded into Shopping Method attribute.

Q3: What are the shopping preferences among customers from different states?

- Provide high-level task abstraction for your dataset:
 - To consume existing data and deriving new data.
 - To discover and compare the shopping method preferred by customers across each states
- List targets or aspect of the data that would answer your analysis question:
 - The distribution of shopping method across different states in the map

Clothing Fashion Designer's Questions

- **Question 4** : What are the color preference for customer across 4 seasons?
- **Question 5** : What category emerges as bestseller for each gender for each season?
- **Question 6** : How satisfied are the customers with their purchases across different categories in each season?

Q4: What are the color preference for customer across 4 seasons?

- Do you need a chart?
 - Yes, Heat Map is required to answer.
- List of attributes do you need to use to answer the question:
 - Color, Season
- Transform data? If Yes – provide short description, what data produced:
 - Yes, sum up the purchased item count and compare across each season.
- Dataset or attribute type changed? If Yes – list them, what type changed:
 - Dataset type and Attribute type do not change.
- All data? Additional data? If Yes – list and provide description:
 - All data is provided in the table.

Q4: What are the color preference for customer across 4 seasons?

- Provide high-level task abstraction for your dataset:
 - To consume existing data and deriving new data.
 - To discover and compare the color preferences over seasonal trends
- List targets or aspect of the data that would answer your analysis question:
 - The distribution of shopping method across different states in the map

Q5: What category emerges as bestseller for each gender for each season?

- Do you need a chart?
 - Yes, Bar Chart is required to answer.
- List of attributes do you need to use to answer the question:
 - Gender, Season, Category
- Transform data? If Yes – provide short description, what data produced:
 - Sum up profit and sales count in every category across different season and then filter by gender.
- Dataset or attribute type changed? If Yes – list them, what type changed:
 - Dataset type and attribute type do not change.
- All data? Additional data? If Yes – list and provide description:
 - All data is provided in the table.

Q5: What category emerges as bestseller for each gender for each season?

- Provide high-level task abstraction for your dataset:
 - To consume existing data.
 - To discover bestselling category among genders over seasonal trends
- List targets or aspect of the data that would answer your analysis question:
 - The profits and sales count in every season for each category in both genders

Q6: How satisfied are the customers with their purchases across different categories in each season?

- Do you need a chart?
 - Yes, Pie Chart is required to answer.
- List of attributes do you need to use to answer the question:
 - Review Rating, Category, Season, Rating Summary
- Transform data? If Yes – provide short description, what data produced:
 - Yes, perform binning on Review Rating, accumulate the sales count across different season and filter by category.
- Dataset or attribute type changed? If Yes – list them, what type changed:
 - Dataset type does not change, but Binning of Review Rating into Ordinal attribute is required to produce feasible findings.
- All data? Additional data? If Yes – list and provide description:
 - No, Rating Summary attribute is required to answer this question therefore it will be derived from Review Rating

Q6: How satisfied are the customers with their purchases across different categories in each season?

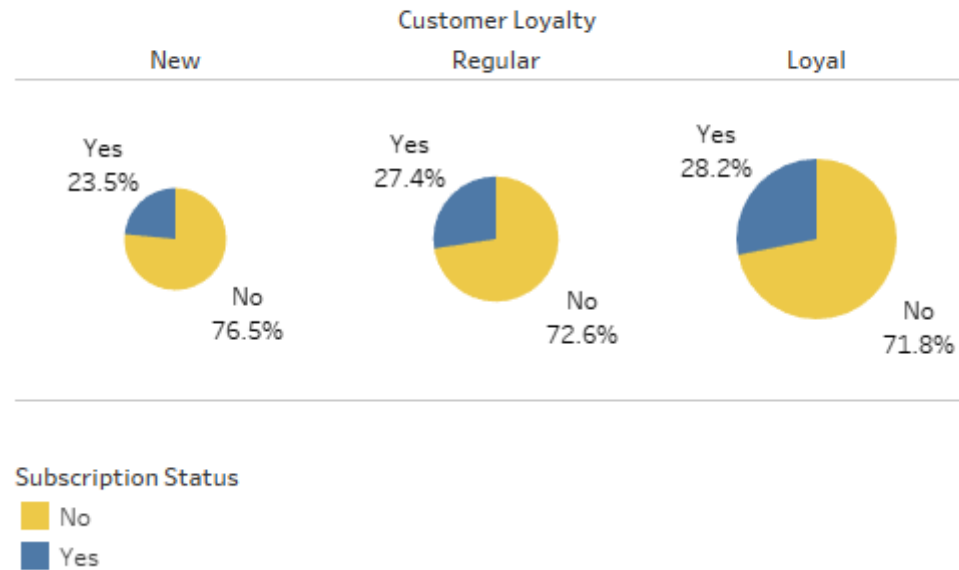
- Provide high-level task abstraction for your dataset:
 - To consume existing data and deriving new data.
 - To discover and compare the categories customers are most satisfied with over seasonal trends
- List targets or aspect of the data that would answer your analysis question:
 - The percentage of ratings across different categories in each season

Choice of Vis Idioms and Explanation

Q1: How many customers exhibit frequent purchase but do not own membership?

Pie Chart

Q1: How many customers exhibit frequent purchase but do not own membership?

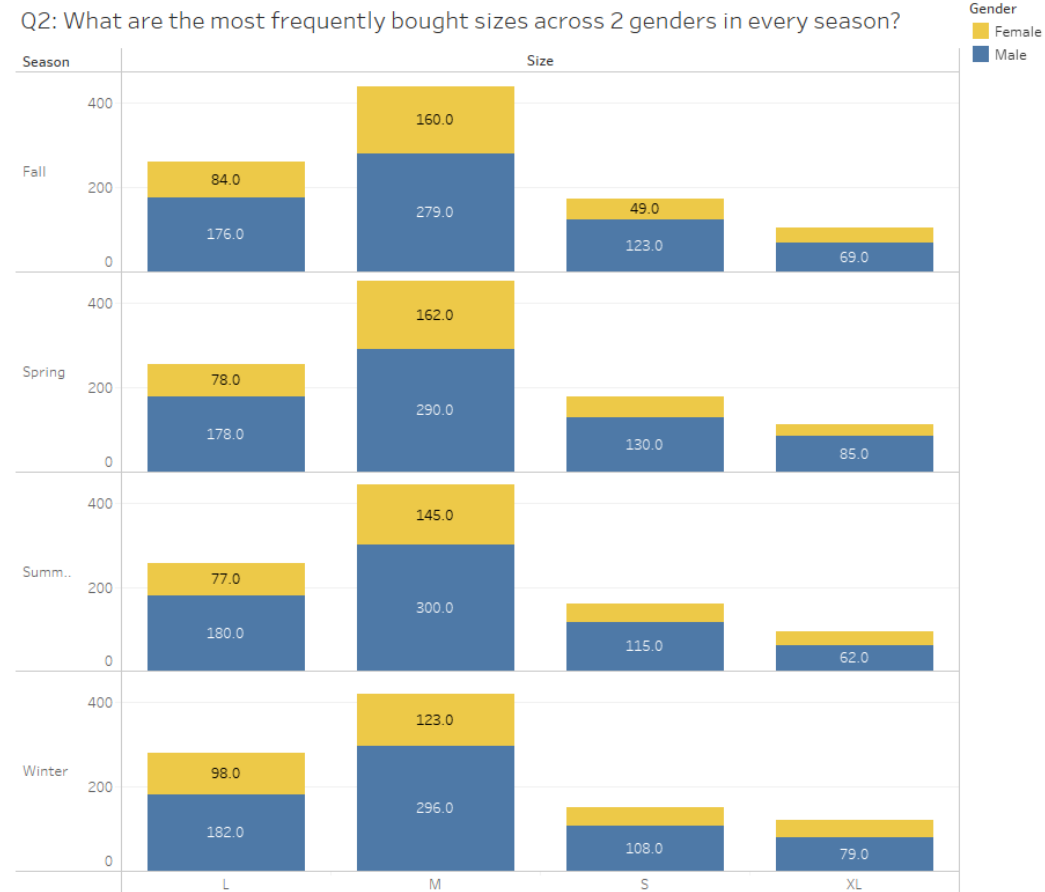


Why is Pie Chart used?

- Pie chart is used as the visual representation for this data due to its quick and effectiveness in distinguishing between non-membership and membership customers, particularly when the data further divided into 3 Customer Loyalty types.
- In this case, displaying the proportions and percentages as a whole allows business owner to immediately get a grasp of how effective the loyalty program reaches out to his/her customers' attention.
- E.g., Business owner can quickly identify how many percentage of customers are non-membership

Q2: What are the most frequently bought sizes across 2 genders in every season?

Bar Chart



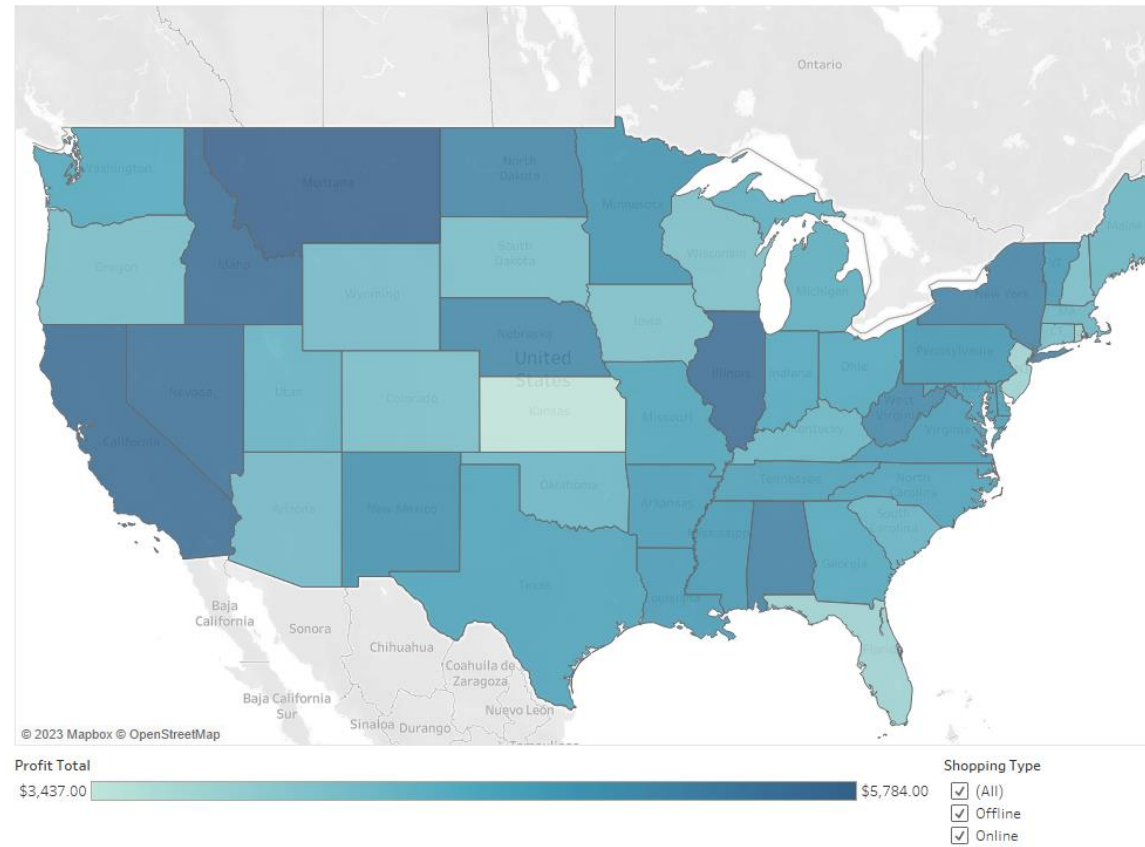
Explanation of why this is the right graph

- Bar Chart is used in this visual representation due to its simplicity and ease of interpretation to deliver a straightforward insights on the popularity of different sizes across both genders in the four seasons.
- The visual representation makes it easy to identify which sizes are consistently purchased and allows for quick comparisons between genders and seasonal trends.
- E.g., Business Owner can tell right away that M size is mostly preferred in every season.

Q3: What are the shopping preferences among customers from different states?

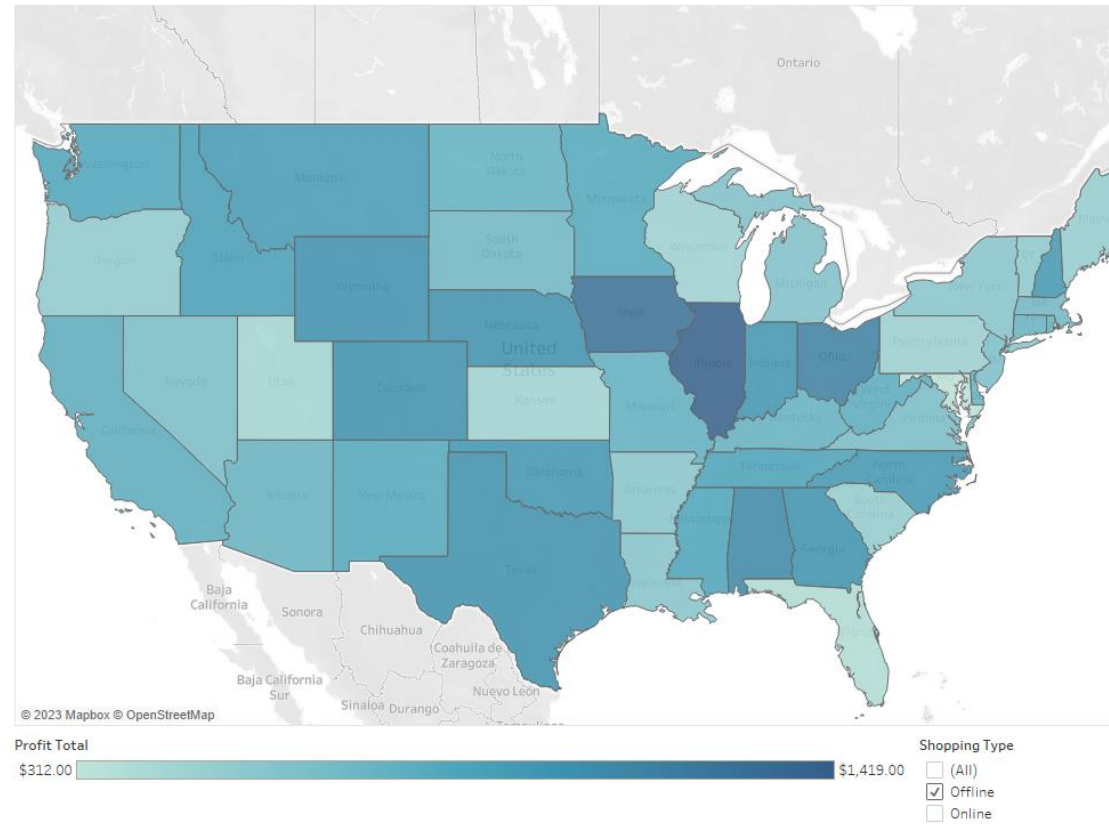
Map Chart

Q3: What are the shopping preferences among customers from different states?



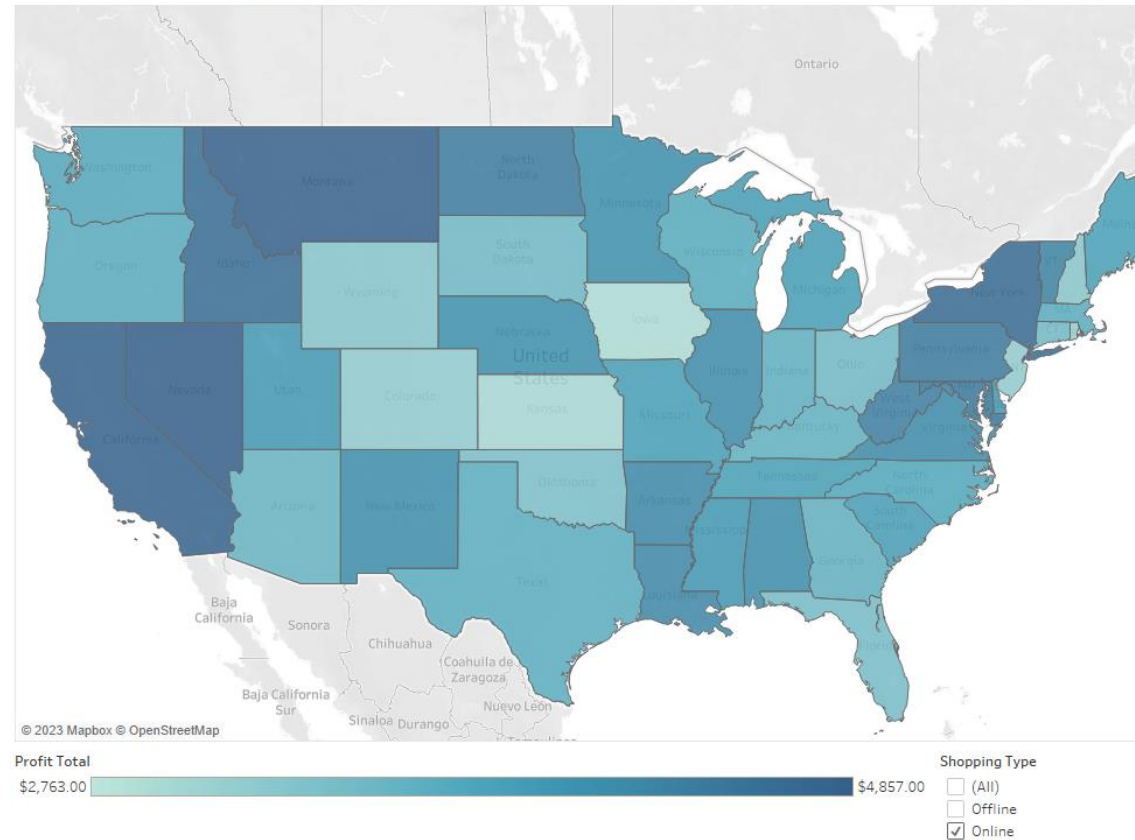
Map Chart (Offline)

Q3: What are the shopping preferences among customers from different states?



Map Chart (Online)

Q3: What are the shopping preferences among customers from different states?



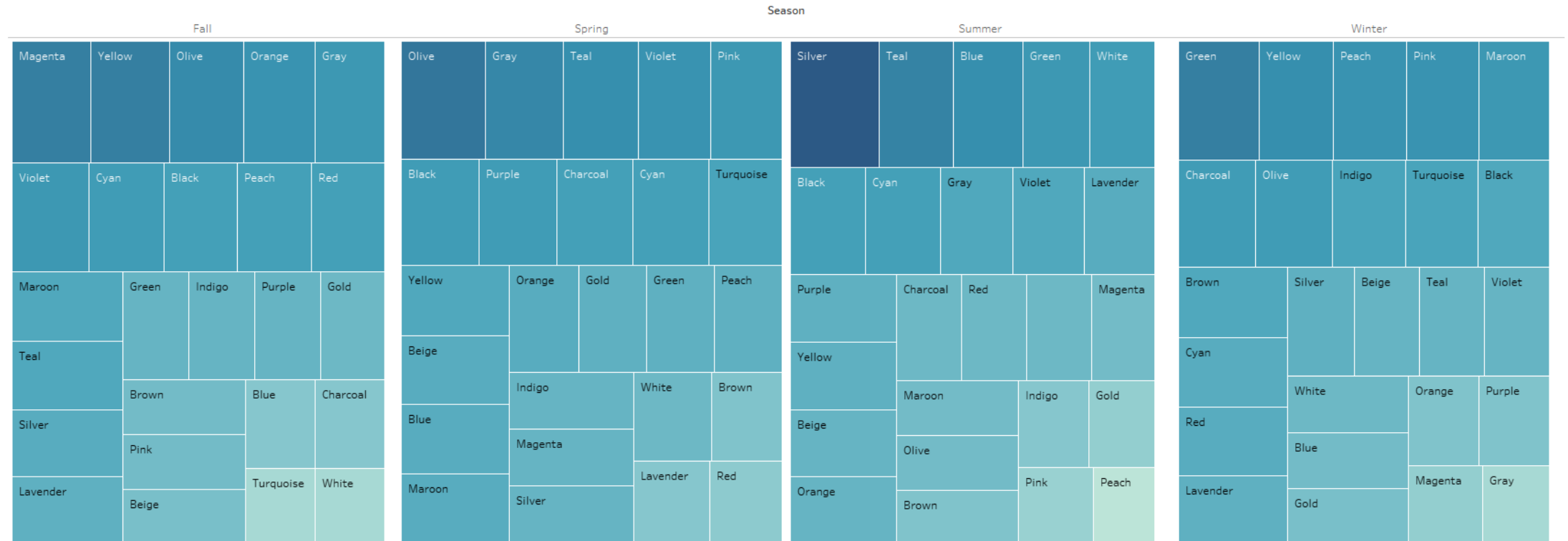
Explanation of why this is the right graph

- Map Chart is used as the visual representation for representation to understand the shopping preferences among customers from different states. The unique characteristics of a map chart provide a powerful way to display regional variations in shopping behavior.
- In this case, the use of color gradients on the map, where deeper colors signify higher profits, allows for a quick and intuitive understanding of the economic performance in each state.
- E.g., Business owners can readily compare the shades to assess the potential profitability of different states.

Q4: What are the color preference for customer across 4 seasons?

Heat Map

Q4: What are the color preference for customer based across 4 seasons?



Measure Values

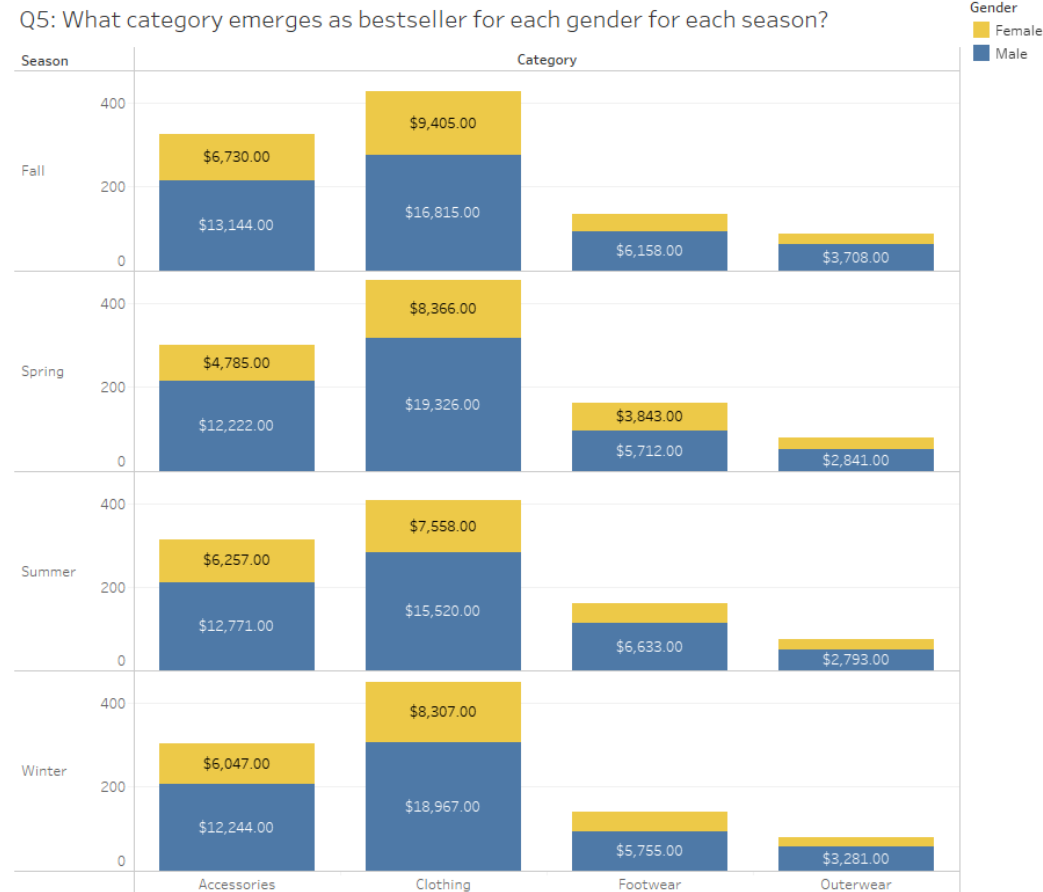


Explanation of why this is the right graph

- Heat Map is used as the visual representation to facilitate quick identification of trends in color preferences. The characteristics of a heat map provide a visually intuitive and informative way to convey the popularity of colors.
- In this case, the heat map proves capable of capturing the color preferences over the 4 seasons. Like the Map Chart, the color gradient showcases the popularity in color preference in the seasons.
- E.g., Fashion designer can easily spot silver being the top preference during Summer or spot consistent preferences in other seasons.

Q5: What category emerges as bestseller for each gender for each season?

Bar Chart



Explanation of why this is the right graph

- Bar Chart is used as the visual representation for understanding the best-selling category for both genders across different seasons. This method is straightforward and effective when comparing data across different attributes to provide insights.
- In this case, the bar charts are segmented to represent different genders and seasons, providing a comprehensive overview of sales performance across these dimensions.
- E.g., Fashion designer can understand right away that Clothing is always the dominant product within the clothing retail store.

Q6: How satisfied are the customers with their purchases across different categories in each season?

Pie Chart

Q6: How satisfied are the customers with their purchases across different categories in each season?



Explanation of why this is the right graph

- Pie chart is used in the visual representation for this data due to its effectiveness in providing insights on the ratings of the apparel products from different categories and seasons.
- In this case, the customer's satisfaction ratings are represented in terms of percentages. Each slice of the pie represents a percentage of customers falling into different satisfaction categories across 4 categories and seasons, providing a quick and insightful overview.
- E.g., Fashion Designer can easily understand what the customers think of the product, giving idea of the product's survivability in the market.

Reflection and Discussion

How might this analysis be useful in visualization design?

To summarize, the analysis of visualization design can significantly contribute to decision-making, understanding customers, and identifying trends over time.

Before the prominence of visualization, businesses found it challenging to uncover trends and market demands solely from sales data. Over time, businesses realized that there are other crucial factors impacting business performance. Subsequently, businesses began to employ statistical survey methods to collect data and study human behavior. However, this method alone proved to be expensive, inefficient, and not conducive to a business's health.

In the current era of visualization, businesses now rely on available data or collect their own data through sales. This data undergoes various processes before being visualized to understand customer behavior. During analysis, businesses can identify patterns in customer spending behavior and gather insights into customer perceptions of their products. Following the analysis, businesses gain a comprehensive understanding of their position in the market, enabling them to take necessary actions.

From this analysis, it is evident that both the intended user and different users must first identify their problems. Subsequently, appropriate visualization types are used to analyze these problems, providing valuable insights for decision-making. After analyzing the identified problems, business owners can fully comprehend their business performance and trends across all seasons. Finally, both users can plan strategies through the analysis to address and eliminate the identified problems.

Additional Notes

- Github Repository to notebook
<https://github.com/abilimckl1/ShoppingBehaviour>
- Percentage representation in pie chart
https://help.tableau.com/current/pro/desktop/en-us/calculations_percentages_options.htm#:~:text=To%20calculate%20percentages%20in%20your,then%20select%20a%20percentage%20option.
- Types of Charts
<https://eazybi.com/blog/data-visualization-and-chart-types>
- History of survey
[https://en.wikipedia.org/wiki/Survey_\(human_research\)](https://en.wikipedia.org/wiki/Survey_(human_research))
- History of Visualization
<https://insightsoftware.com/blog/a-brief-history-of-data-visualization/>