I. Introduction

In the competitive business landscape, understanding customer behavior and market trends are critical for sustaining the business growth and optimizing strategies. Retail businesses often face difficulties in deriving insights from large dataset, which can contain critical information for the business. Hence, this project aims to uncover trends and insights within the dataset, then utilize the data visualisation techniques to explore these trends and insights to enable data-driven decision-making to further improve the profitability and growth of the retail business. The objectives of this project are:

- To analyse the influence of demographic factors such as age and gender on purchasing behavior.
- To identify high-performing locations and product categories that drives revenue.
- To uncover how product color preferences vary and affects sales across different seasons and locations.
- To discover shipping type preference and its impact on the sales performance.

Furthermore, the target audience includes retail business stakeholders, such as the marketing teams who seek insights for targeted campaigns, inventory managers aiming to optimizes the stock based on demand, and strategic decision-makers who plan regional and seasonal sales strategies.

II. Description of Data

This project uses transactional records dataset from a fashion business which has stores in the spread around USA. The dataset contains 3900 entries and 19 attributes, providing comprehensive insights into customer demographics and purchasing behavior. The dataset is obtained from GitHub, with structured and well-organised data. The dataset contains 19 attributes, and the key attributes are described below:

- Age: Represents the age of customers
- Gender: Categorizes customers into male and female
- Category: Categorizes the type of the product purchased
- Purchase Amount (USD): Represents the spending value of each transaction
- Location: Specifies the geographical area where the purchase was made
- Color: Provide additional details of the color of the product
- Season: Indicates the season during which the purchase occurred
- Shipping Type: Indicate the type of shipping used to send the product

The dataset was chosen for its comprehensive and diverse information, covering multiple dimensions such as demographics, transactional, and geographical data for exploration to answer a wide range of analytical questions. By leveraging this dataset, meaningful patterns and insights are aimed to be uncovered to evaluate business strategy and optimize business outcomes.

III. Initial Questions

These questions were formulated to cover diverse aspects of the dataset, making them suitable for exploratory analysis which will guide the shaping of further questions from the visualisations. Based on the dataset's attributes, the following initial questions were formulated:

1. What is the average purchase amount across different age groups and genders?
   - This question aims to explore spending behavior across different age and gender groups, revealing the purchasing behavior trends by demographic segment.
2. Which location contributes the most to overall revenue?
   - This question aims to analyse revenue distribution by location to provide insights into geographical market's performance.
3. How do product color preferences vary across different seasons and locations?
   - This question aims to uncover whether certain colors are more popular in specific seasons or locations, revealing trends which can boost inventory and marketing strategies.
4. Which shipping type is the most popular among the customers?
   - This questions aims to determine the most popular shipping types among the customers to understand the shipping preferences and re-evaluate logistical strategies.

IV. Data Analysis and Visualisation Strategies

Before diving into the initial questions, the data underwent several data cleaning processes to ensure data quality and readiness for analysis. These steps are outlined below:

- Null Value Check and Removal: remove rows with missing value

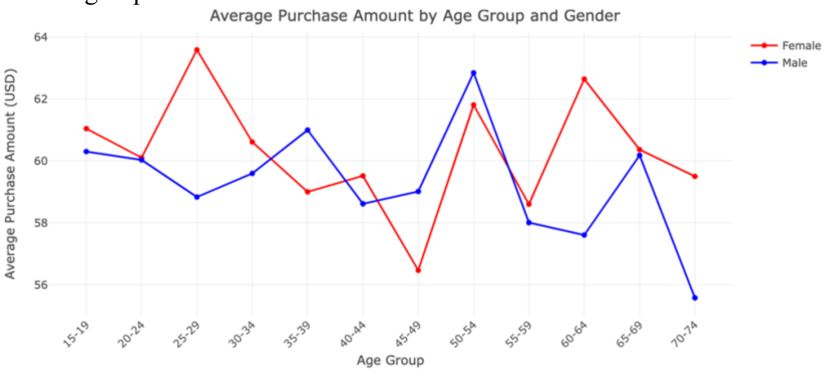
- Column Renaming: for simplicity and visualisation purposes
- Data Type Conversion: convert categorical variables to factors for analysis
- Duplicate Removal: remove duplicate rows
- Ensure Numeric Formatting: ensure the numeric values are properly formatted

These data cleaning steps were performed prior to addressing the initial questions to ensure the dataset was consistent, complete, and ready for further analysis.

A. Initial Question 1: What is the average purchase amount across different age groups and genders?

The subset used is all rows of Age, Gender, and Purchase Amount (USD) which are included to provide a complete view of purchasing behavior across demographic segments. To further prepare the data for analysis, the data underwent data transforming processes as below:

- Age Grouping: Binned the continuous "Age" values into defined ranges for easier comparison
- Averaging Purchase Amounts: Aggregated data by Age and Gender, calculating the mean purchase amount for each group



Line charts were used to visualize the average purchase amount across different age groups, with the following design elements:

- X-axis: Represents different age groups.
- Y-axis: Represents the average purchase amount in USD.
- Visual Encoding: Different genders are represented using distinct colors to allow for easy differentiation and comparison between genders.
- Interactivity: hover pop-ups that display detailed information when the cursor is dragged over a data point, and the ability to choose different age group ranges to enhance comparison and visualization.

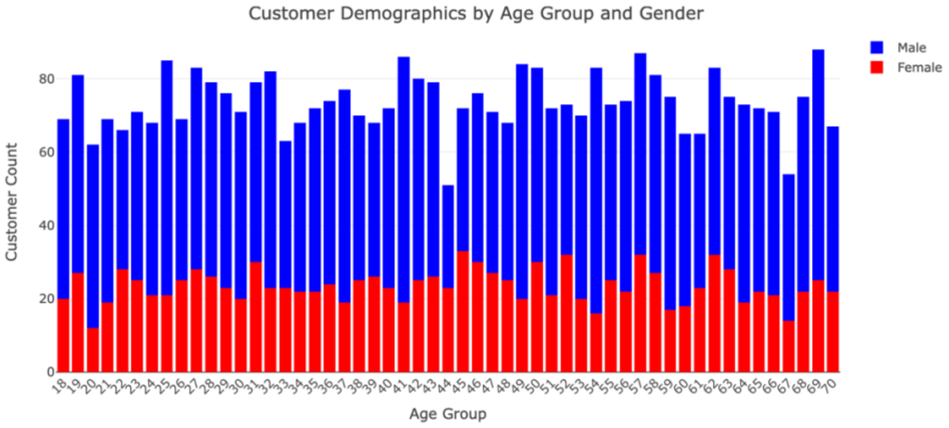
The line chart was chosen instead of other visualisation because it effectively and clearly conveys trends in the average purchase amount across different age groups, making it easy to identify the patterns and disparities between genders over different age ranges. Moreover, the continuous "Age" data further justify the usage of this visualisation. Other visualisation, like boxplots, bar charts, or scatter plots, were less suitable. Boxplots focus on data spread rather than trends, bar charts are better for categorical comparisons, and scatter plots do not clearly show continuous progression.

During this analysis, it can be seen that female consumers at the age range of 18 to 30 and above 55 tend to have higher average purchase amount than male consumers with the same age. This has raised further additional questions such as:

1. Reflective Question 1 for Initial Question 1: Are higher purchases by female customers in certain age groups due to there being more female customers overall?

This question aims to determine if the higher average purchase amounts observed among female customers are influenced by a larger number of female consumers in those age groups. The subset used for this analysis includes all rows of Age and Gender columns. This subset was selected to determine the number of customers for each years and genders. To further prepare the data for analysis, the data underwent these transformation processes:

- Age Grouping: Binned the continuous "Age" values into defined ranges for easier comparison
- Customer Count Aggregation: Data was aggregated by Age and Gender, calculating the total number of customers within each group



Stacked bar charts were used to visualise the total customer count by age group and gender, with the following design elements:

- X-axis: Represents the different age groups.
- Y-axis: Represents the total number of customers in each age group.
- Visual Encoding: Different genders are represented using distinct colors, allowing easier differentiation and comparison.
- Interactivity: Hover pop-ups display specific counts when the cursor is dragged over each segment of the bar, and age groups can be adjusted to facilitate better visualization and analysis.

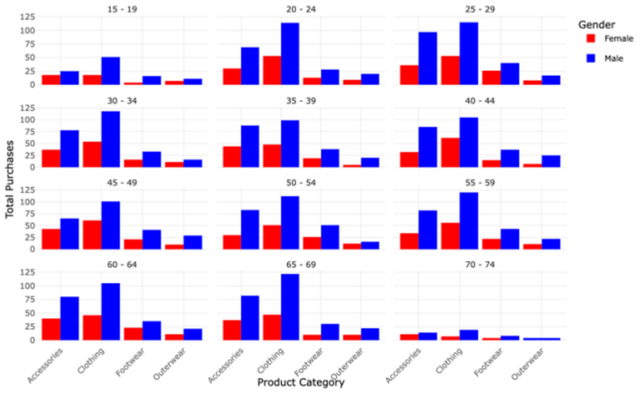
The stacked bar chart was chosen because it effectively shows the composition of customer counts within each age group. Unlike line charts, which are best for showing trends over continuous data, or pie charts, which become cluttered with multiple categories, the stacked bar chart clearly displays proportional data and makes it easy to compare male and female customer numbers across age groups.

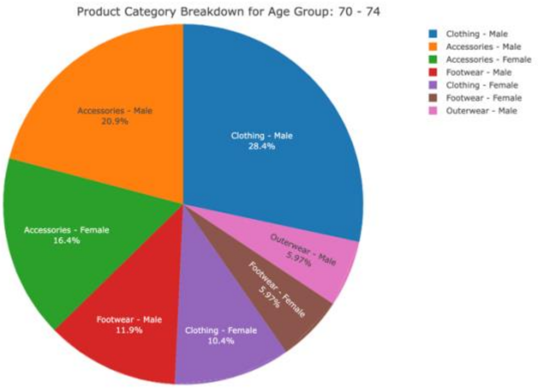
This visualization shows that the number of female customers is always lower than the male customers across all age groups. Despite this, the average spending of the female customers is still higher at certain age groups, indicating that the female customers tend to spend higher amount of spending than male customers. Moreover, factors other than gender may also be contributing to the higher purchase amount.

2. Reflective Question 2 for Initial Question 1: Which product categories are most popular among different age groups and genders?

This question aims to identify the most favoured product categories across demographics and the breakdown of it. The subset used for this analysis includes all rows of Age, Gender, and Product Category columns. This selection of the subset allows an in-depth look into how product preferences vary across demographics, specifically among age groups and genders. To prepare the data for analysis, the following transformation processes were applied:

- Age Grouping: Binned the continuous "Age" values into defined ranges for easier comparison
- Category Aggregation: Data was aggregated by Age, Gender, and Product Category to determine the popularity of each category in each groups.
- Age Range Definition: define age range then create corresponding label for each groups.
- Filtering : Filter data for the selected age group





A grouped bar chart and pie chart were used to visualise the popularity of the product categories, with the following design elements:

- X-axis (Bar Chart): Represents different product categories.
- Y-axis (Bar Chart): Represents the count of purchases within each category.
- Visual Encoding: Different genders are represented using distinct colors in the bar chart and color palettes are used in the pie chart to further represent different categories for each gender
- Interactivity: Users can toggle between a bar chart and a pie chart to enhance visualization. Hover pop-ups will provide the detailed data

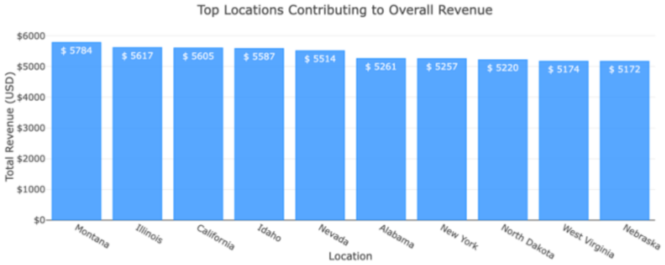
The grouped bar chart was chosen because it can clearly show which categories are the most popular and the least popular, enabling easy comparison across categories in the age groups. On the other hand, the pie chart provides a more intuitive overview of the proportional share of each product categories which is useful for understanding the overall proportion at a glance. The choice of these two charts rather than the other visualisation was based on their specific strengths and focuses. Both options provides different perspectives for better decision-making that the other visualisation can't provide.

This visualisation reveals that clothing categories are the most preferred by all ages and genders, followed by accessories categories. The pie chart further shows that clothing categories alone contributes to almost half of the total sales. This highlights the key differences that could be used to tailor marketing strategies and optimize product offering.

B. Initial Question 2: Which location contributes the most to overall revenue?

The subset used is all rows with the columns Location and Purchase Amount (USD) which are included to provide an overview of the revenue gained across different locations. To further prepare the data for analysis, the following data transformation processes were applied:

- Revenue Aggregation: Data was aggregated by Location to calculate the total revenue gained in each location.
- Location Sorting: Locations were sorted based on their total revenue to identify the top contributors.



A bar chart was used to visualise the total revenue contribution of the top locations, incorporating the following design elements:

- X-axis: Represents different locations, sorted by descending order of revenue.
- Y-axis: Represents the total revenue (USD) for each location.
- Visual Encoding: The highest revenue locations will be represented as a bar, with taller bars indicating higher contributions.
- Interactivity: Users can choose between 5,10,15, or 20 highest revenue locations to be displayed. Hover pop-ups will display detailed insights of the location.

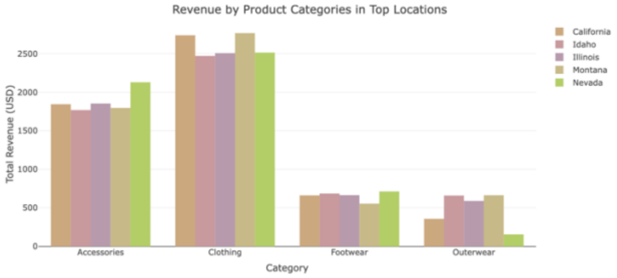
The bar chart was chosen for its effectiveness in displaying ranked categorical data, which supports easy comparison between the revenue contributions across the locations. Unlike pie charts, which can become complex when comparing too many locations, or line charts, which are better suited for time-series data, the bar chart provides a clear and straightforward visual representation of each location's revenue contribution.

During this analysis, the rank of the locations based on their revenue contribution is shown, revealing the top revenue contributor locations and indicating the disparities between different locations. This has raised further additional questions such as:

1. Reflective Question 1 for Initial Question 2: Which product categories generate the highest revenue in the top-performing locations?

This question aims to determine whether specific product category drives the revenue in these locations. The subset used for this analysis includes all rows of Location, Product Category and Purchase Amount (USD) columns. This subset was selected because it provides a complete view of product preferences across the locations. To further prepare the data for analysis, the data underwent these transformation processes:

- Group and Aggregate: calculate total revenues by location and category.
- Filtering : filter on the number of top-performing locations based on user input.
- Location Ranking: To identify the top locations dynamically.



A grouped bar chart was used to visualise the revenue contribution of each product category across the top performing locations, with following design elements:

- X-axis: Represents different product categories within each top-performing location.
- Y-axis: Represents the total revenue (USD) for each product category.
- Visual Encoding: Different locations are represented using distinct colors, allowing for easy differentiation.
- Interactivity: Users can select the number of top-performing locations, with hover pop-ups which can show the detailed data for each bar.

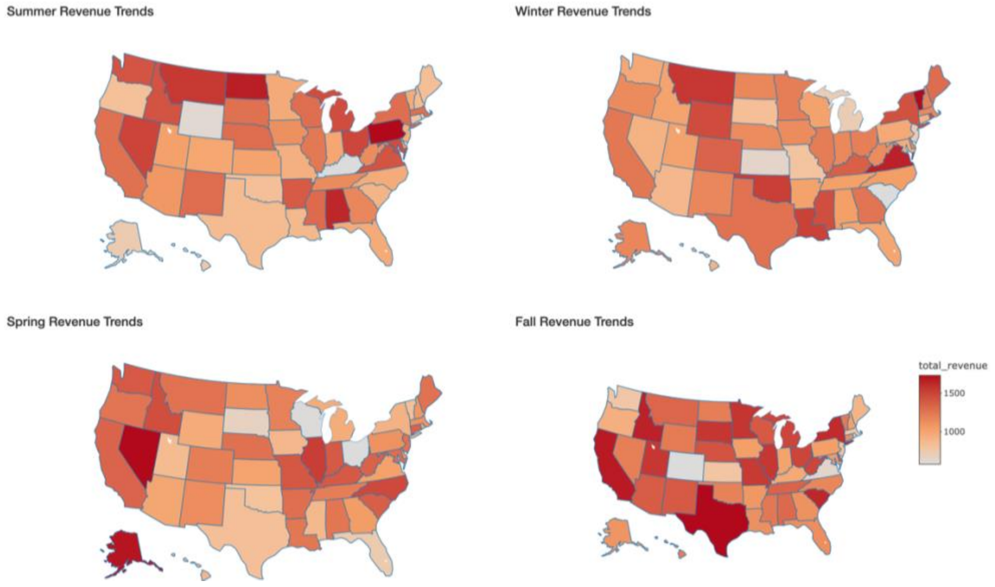
The grouped bar chart was chosen for its effectiveness in showing the variation of revenue contribution of different product categories across multiple locations which pie chart can't provide.

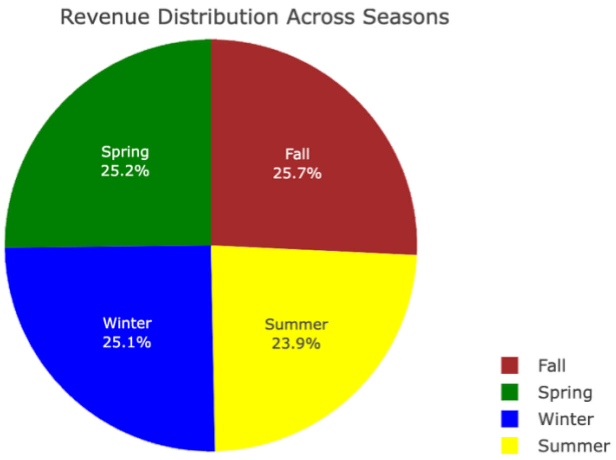
This visualisation reveals that in all of the top-performing location, clothing category contributes the most, followed by accessories. This shows that not only these categories are the most preferred by the customers (as shown in Reflective Question 2 for Initial Question 1), but it also generates the highest revenue in the top-contributing locations. This highlights that clothing and accessories categories have a huge impact on this retail fashion business.

2. Reflective Question 2 for Initial Question 2: Are there significant seasonal variations in revenue for the locations?

This question aims to determine how seasonal revenue trends vary across different seasons in the different locations. The subset used for this analysis includes all rows of Location, Purchase Amount (USD), and Season columns which are selected to provide comprehensive data to evaluate revenue trends across seasons. The data underwent the following transformation processes:

- Revenue Aggregation by Season: Data was aggregated by Location and Season to calculate the total revenue generated in each season.
- Location Mapping: Locations were mapped to state abbreviations to enable geographic visualization.





Faceted Choropleth Map was used to visualise seasonal revenue changes across the locations in USA, incorporating the following design elements:

- Faceting: Each map facet represents one season to show changes across seasons side by side.
- Color Encoding: States are coloured based on total revenue, with darker shades indicating higher revenue.
- Interactivity: Hover pop-ups of the details of the locations and the revenue.

Additionally, a pie chart was used to show the overall revenue distribution across seasons. The Faceted Choropleth Map was chosen for of its ability to display a detailed map of USA, which is required to analyse the revenue changes in the location considering the cardinal directions. This approach will further provide valuable insights of the influence of seasons to the revenue. Furthermore, pie chart was chosen for its straightforward representation of which season generate the highest total revenue.

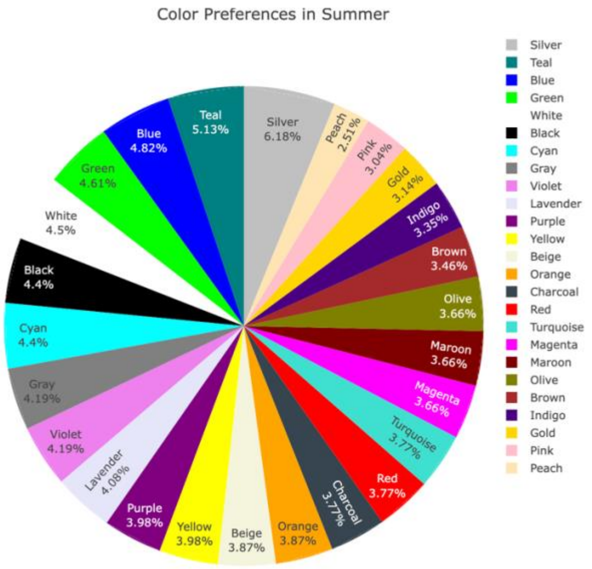
This visualisation reveals that Southern region of USA has the highest revenue during Fall season compared to another seasons. Moreover, the pie chart shows that Fall season has the highest income compared to other seasons. This indicates that the peak revenue for the year is concentrated in the Southern USA during Fall season. Consequently, this insight can inform business strategies, and further analysis of geographical factors is necessary to understand the underlying reasons for this trend.

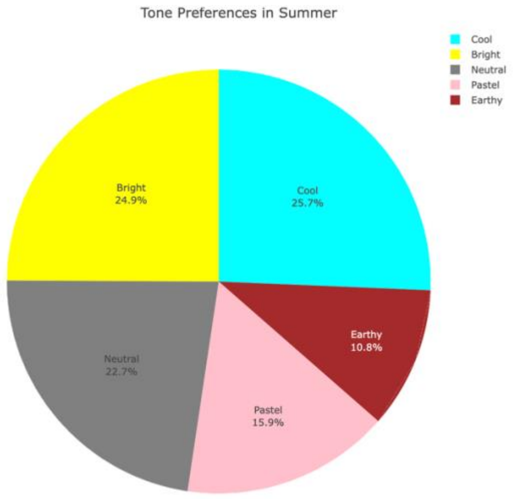
C. Initial Question 3: How do product color preferences vary across different seasons ?

This question aims to analyse the variation of product color preferences across different seasons, providing insights into the seasonal trends. The subset used for this analysis includes all rows of Colors and Seasons columns. The data underwent the following transformation processes:

- Color Grouping: categorized colors into tones for easier comparison (Lee, n.d.)

- Seasonal Aggregation: Grouped data by season, tone , and color to sum up the total purchases for each group





Pie charts were chosen to display the preferences for tones and colors across seasons, incorporating the following design elements:

- Visual Encoding: each tone and color were assigned with their corresponding color in the pie chart
- Interactivity: Toggle between tone and color view for flexible analysis. Hover pop-ups to provide detailed information.

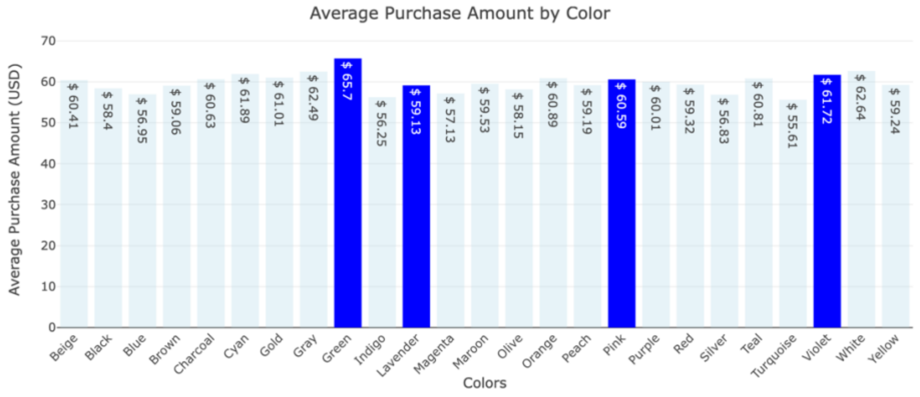
Pie chart was chosen because it is effective to communicate the proportional share of the color preferences within each season. While bar charts are suitable for showing counts, this analysis requires a more intuitive understanding of distribution proportions.

This visualisation shows that bright colors tone is the most popular across all seasons, followed by cool colors tone. This indicates that even though, generally specific color tone is preferred during specific season, there is no significant effect of season to the customer color preferences in this dataset. However, the popularity of bright and cool colours tones still raise further addition question such as:

1. Reflective Question 1 for Initial Question 3: Does color preferences impact the average purchase amount?

This question aims to reveal the specific color preferences that influence the average purchase amount. The subset used for this analysis includes all rows of Color, and Purchase Amount (USD) columns. The data underwent the following transformation processes:

- Color Grouping: Colors were grouped based on their tone by merging the tones with the main dataset.
- Purchase Amount Aggregation: The data was aggregated to calculate the average purchase amount for each color tone by grouping the data by color and tone.



A bar chart was used to represent the visualisation, incorporating the following design elements:

- X-axis: Represents different colors.
- Y-axis: Represents the average purchase amount (USD).

- Visual Encoding: Highlighted tones are represented in bold or distinct colors, enhancing focus.
- Interactivity: Users can select specific tones to highlight, with hover pop-ups displaying detailed purchase amounts for each color.

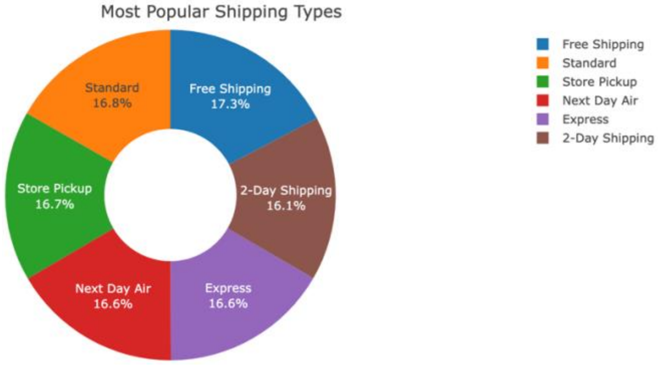
The bar chart was chosen for its effectiveness to compare the average purchase amounts across colors. Moreover, the interactive highlight further improve the comparison. Other chart, such as pie chart and line chart, are not suitable for this analysis because this analysis focus on the discrete data.

This visualisation suggests that while specific colors and tones have higher average purchase amounts, overall the relationship between color and spending behaviour are not strong enough to make any pattern, supported by lack of significant variation (mostly ranging between $55 and $65). Further segmentation are required to possibly reveal more actionable patterns.

D. Initial Question 4: Which shipping type is the most popular among the customers?

This question aims to determine the most popular shipping types among the customers to understand the shipping preferences and re-evaluate logistical strategies. The subset used for this analysis includes all rows of Shipping Type column, which enables clear view of shipping methods choices. The data underwent the following transformation process:

- Aggregation: grouped by Shipping Type to calculate the total number



A doughnut chart was used to visualize the popularity of each shipping type, incorporating the following design elements:

- Visual Encoding: Each segment of the doughnut chart is color-coded to represent the percentage of a specific shipping type.
- Interactivity: Hover pop-ups display detailed information about each shipping type

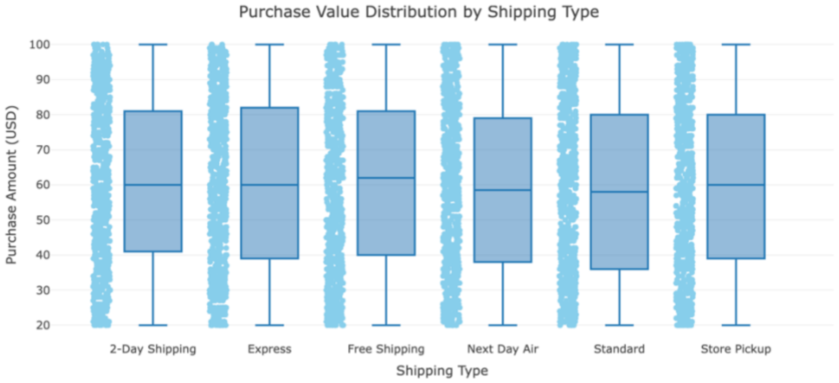
The doughnut chart was chosen for its ability to provide an intuitive visual summary of proportions. While pie chart can provide the same benefit, doughnut chart was chosen just because it has better visual as the proportion of each segment is almost the same.

The visualisation reveals that although Free Shipping is the most popular option, there are no noticeable difference among the shipping type preference. Hence, further additional question is required to make data-driven decision.

1. Reflective Question 1 for Initial Question 4: Does the choice of the shipping type correlate with total purchase value?

This question aims to discover the relationship between the chosen shipping type and the purchase value. The subset used for this analysis includes all rows of Shipping Type and Purchase Amount (USD), which provide detailed information about purchase values across different shipping options. The data underwent the following transformation processes:

- Purchase Value Grouping: distributed the purchase values for each shipping type.



A box plot was used to visualize the distribution of purchase values by shipping type, incorporating the following design elements:

- X-axis: Represents different shipping types.
- Y-axis: Represents the purchase amounts (USD).
- Visual Encoding: Each shipping type is represented by a separate box, illustrating its distribution, median, and variability.
- Interactivity: Hover pop-ups display detailed information for each data point.

The box plot was chosen for its ability to visualise distribution, such as range, median, and quartiles, which is required in this analysis to see what shipping type does the customers prefer. Bar charts are not suitable because it focuses on totals or averages. Scatterplot also shows distribution but it is more suitable to represent continuous value.

This visualisation reveals that although most shipping types have similar median and purchase values, faster shipping options, such as Express and 2-Days Shipping, show slightly higher variability in purchase amounts. This highlights faster shipping are somewhat more preferred even though it costs more. These insights highlight opportunities for enhancing shipping options to improve customer satisfaction and potentially increase profitability through the faster shipping services.

V. Reflection

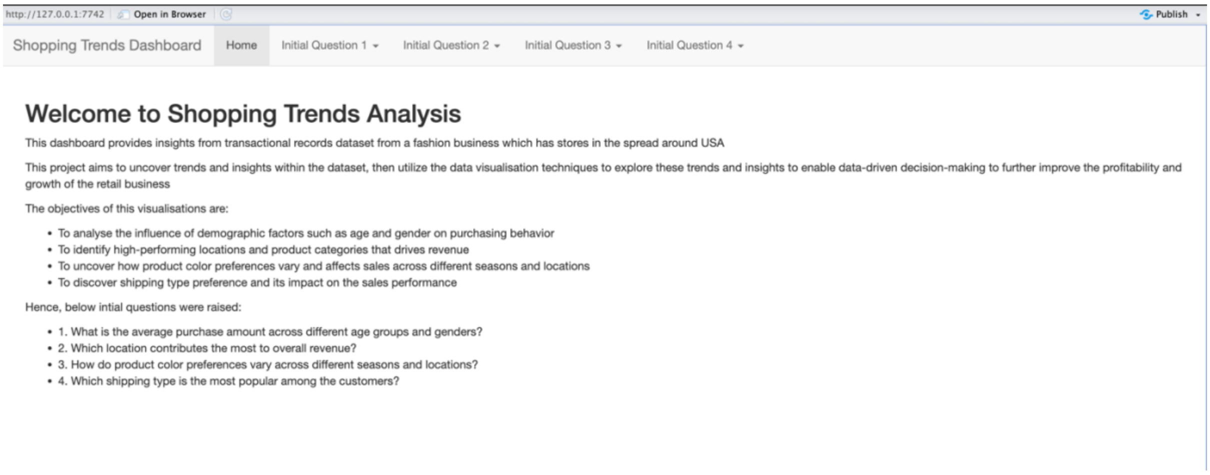
This project has provided valuable insights into the development process of creating effective data visualisations. The importance of selecting the appropriate visualisation techniques to answer specific analytical questions has been learnt through this process. This indicates the need to align visualisations with the data and objectives. Furthermore, throughout the project, data cleaning and preprocessing skills have been learnt. Those were essential to ensure the reliability of the analysis. Moreover, the experience of implementing interactivity has been acquired which made the visualisations more engaging and user-friendly. However, the development process revealed the need for better time management, particularly in refining the visualisations. Furthermore, the dataset used for this project is too simple, which can be a lesson to improve in the next project. Additionally, the importance of both functionality and aesthetic appeal of the visualisation has been understood through this process. Overall, this project enhanced both technical skills and data-driven problem approach. By focusing of these areas for improvement, more impactful visualisations can be achieve in future projects.

VI. References

Lee, R. (n.d.). *Seasonal tonal palettes: A comprehensive guide to colour analysis*. Retrieved December 8, 2024, from https://robertastylelee.co.uk/seasonal-tonal-palettes-a-comprehensive-guide-to-colour-analysis/

VII. Appendix

Appendix A : Home Page Dashboard



Welcome to Shopping Trends Analysis

This dashboard provides insights from transactional records dataset from a fashion business which has stores in the spread around USA

This project aims to uncover trends and insights within the dataset, then utilize the data visualisation techniques to explore these trends and insights to enable data-driven decision-making to further improve the profitability and growth of the retail business

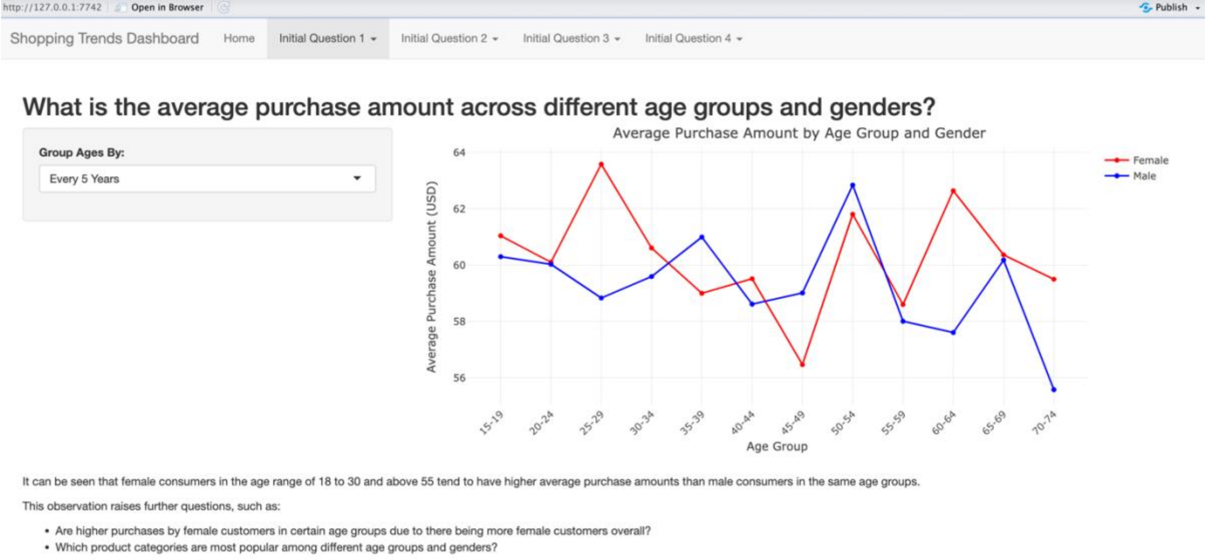
The objectives of this visualisations are:

- To analyse the influence of demographic factors such as age and gender on purchasing behavior
- To identify high-performing locations and product categories that drives revenue
- To uncover how product color preferences vary and affects sales across different seasons and locations
- To discover shipping type preference and its impact on the sales performance

Hence, below intial questions were raised:

- 1. What is the average purchase amount across different age groups and genders?
- 2. Which location contributes the most to overall revenue?
- 3. How do product color preferences vary across different seasons and locations?
- 4. Which shipping type is the most popular among the customers?

Appendix B : Initial Question 1 Dashboard



What is the average purchase amount across different age groups and genders?

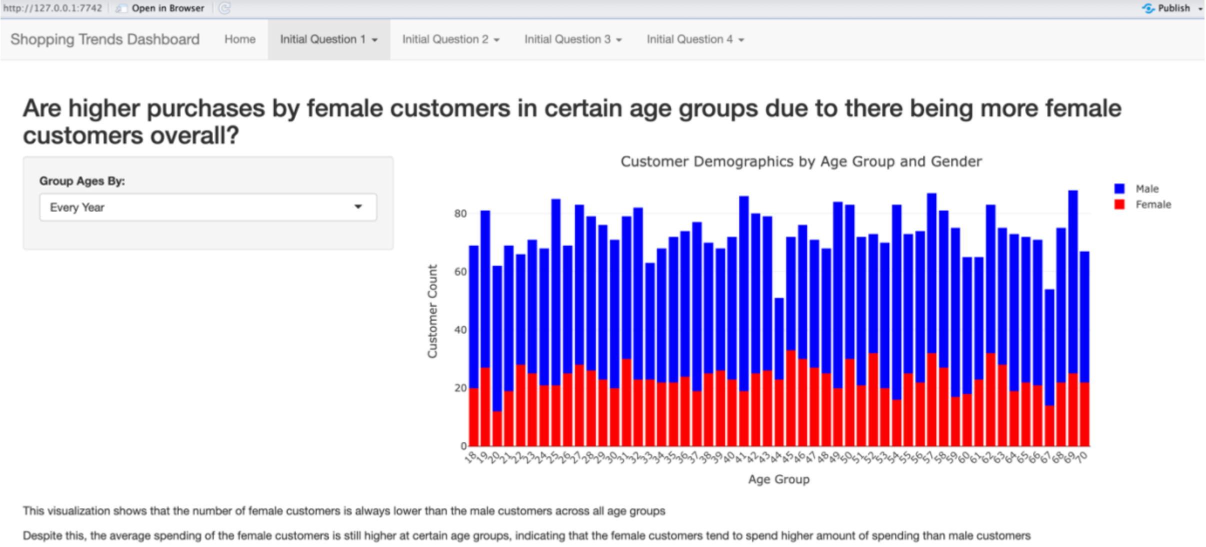
Group Ages By: Every 5 Years

It can be seen that female consumers in the age range of 18 to 30 and above 55 tend to have higher average purchase amounts than male consumers in the same age groups.

This observation raises further questions, such as:

- Are higher purchases by female customers in certain age groups due to there being more female customers overall?
- Which product categories are most popular among different age groups and genders?

Appendix C: Reflective Question 1 for Initial Question 1 Dashboard



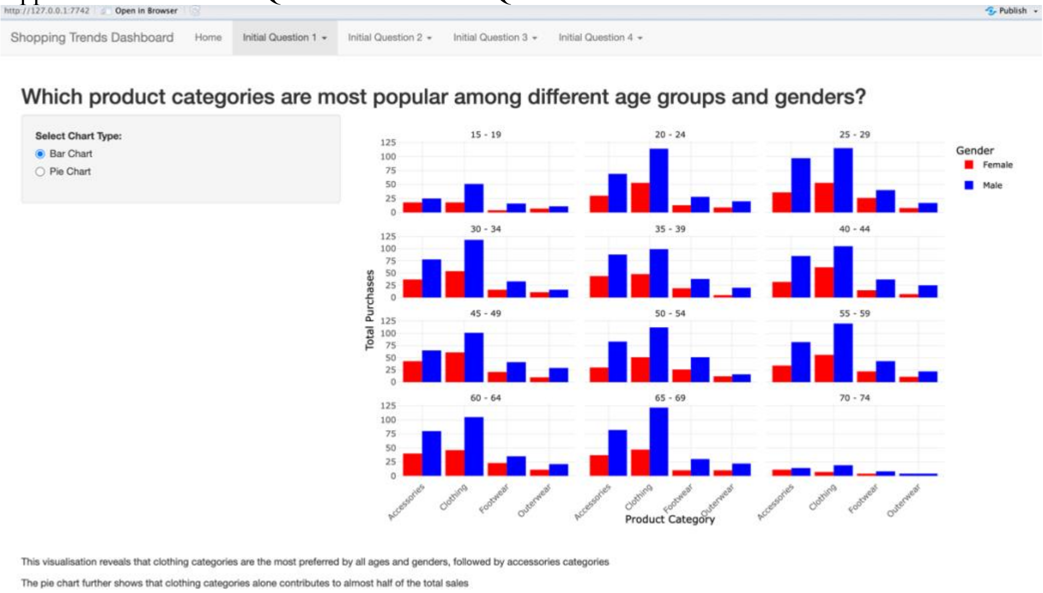
Are higher purchases by female customers in certain age groups due to there being more female customers overall?

Group Ages By: Every Year

This visualization shows that the number of female customers is always lower than the male customers across all age groups

Despite this, the average spending of the female customers is still higher at certain age groups, indicating that the female customers tend to spend higher amount of spending than male customers

Appendix D: Reflective Question 2 for Initial Question 1 Dashboard



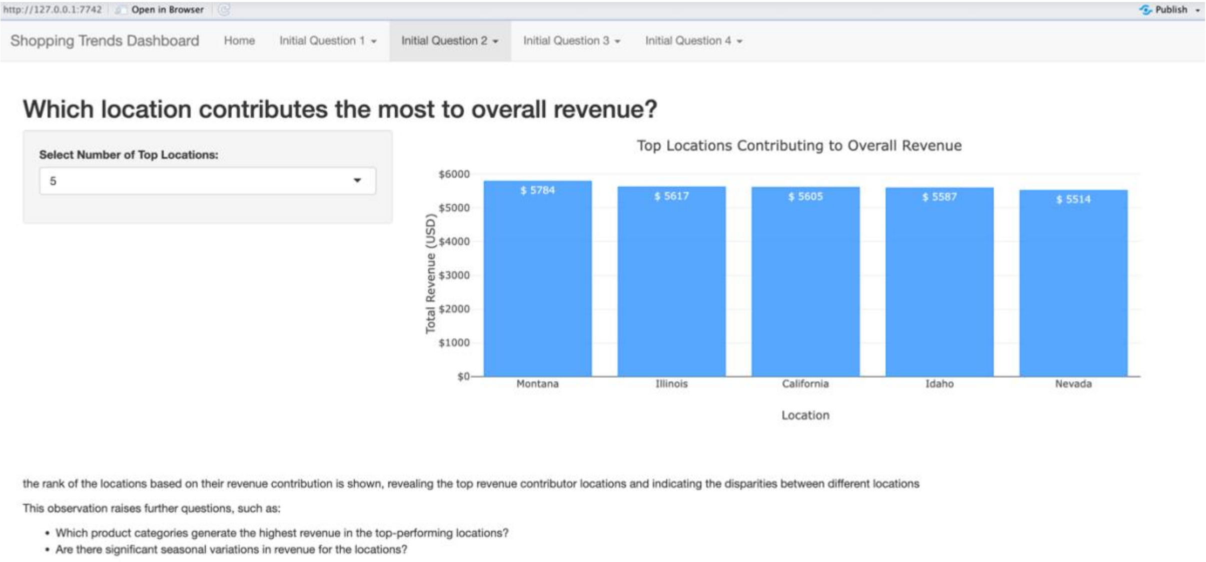
Which product categories are most popular among different age groups and genders?

Select Chart Type:
● Bar Chart
○ Pie Chart

This visualisation reveals that clothing categories are the most preferred by all ages and genders, followed by accessories categories

The pie chart further shows that clothing categories alone contributes to almost half of the total sales

Appendix E: Initial Question 2 Dashboard



Which location contributes the most to overall revenue?

Select Number of Top Locations:

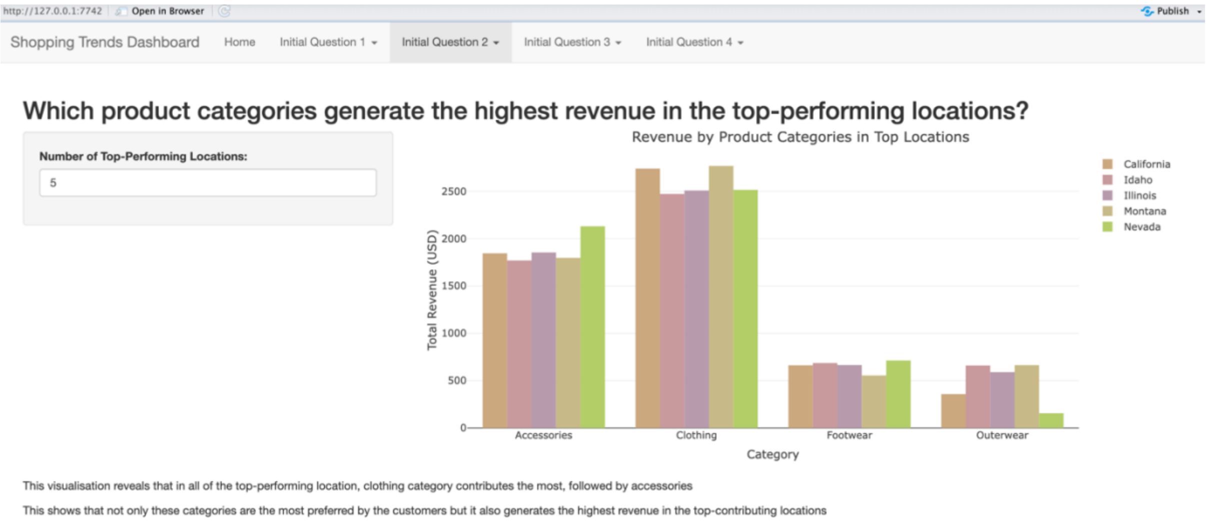
5

the rank of the locations based on their revenue contribution is shown, revealing the top revenue contributor locations and indicating the disparities between different locations

This observation raises further questions, such as:

- Which product categories generate the highest revenue in the top-performing locations?
- Are there significant seasonal variations in revenue for the locations?

Appendix F: Reflective Question 1 for Initial Question 2 Dashboard



Which product categories generate the highest revenue in the top-performing locations?

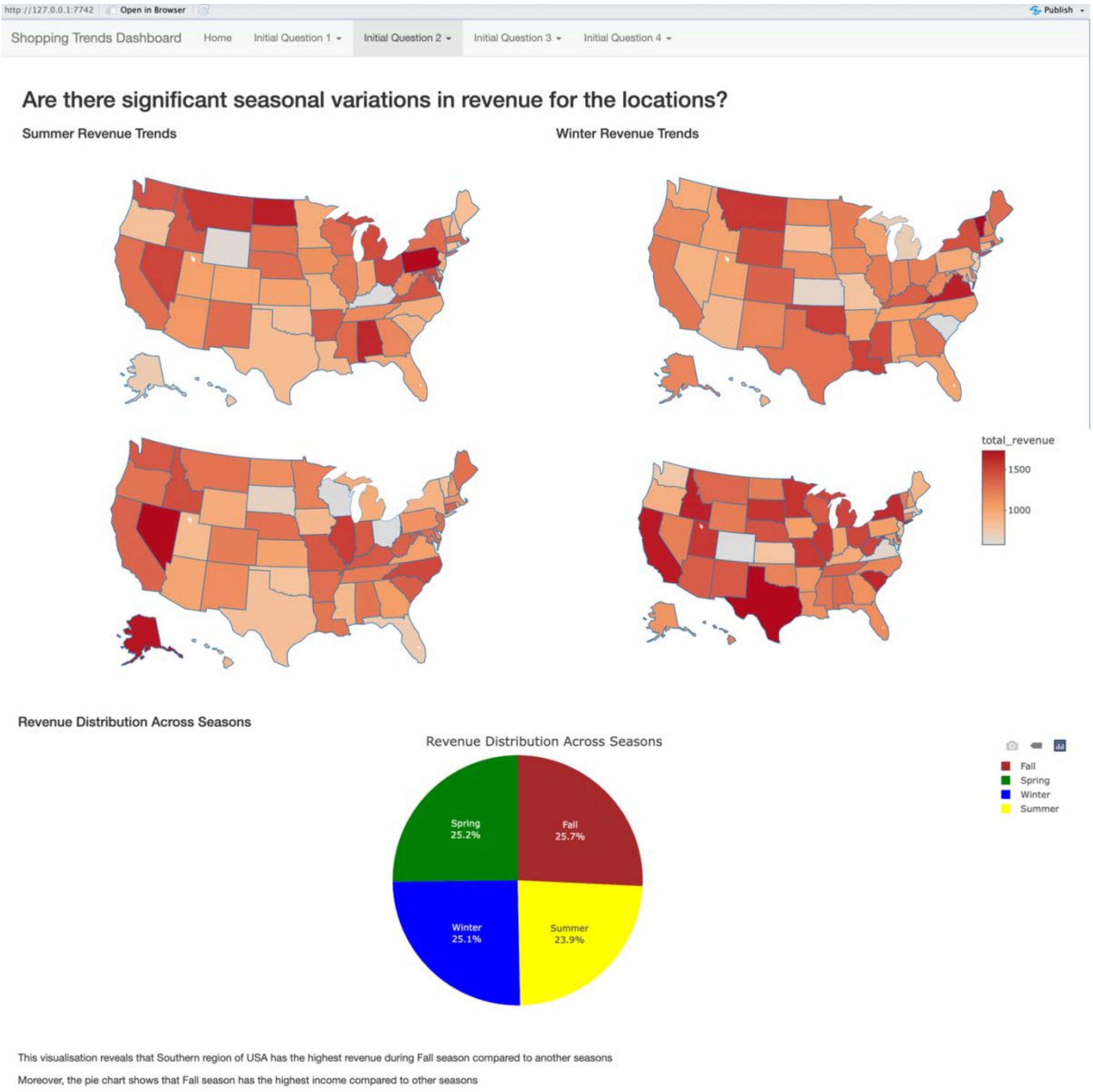
Number of Top-Performing Locations:

5

This visualisation reveals that in all of the top-performing location, clothing category contributes the most, followed by accessories

This shows that not only these categories are the most preferred by the customers but it also generates the highest revenue in the top-contributing locations

Appendix G: Reflective Question 2 for Initial Question 2 Dashboard



Are there significant seasonal variations in revenue for the locations?

Summer Revenue Trends

Winter Revenue Trends

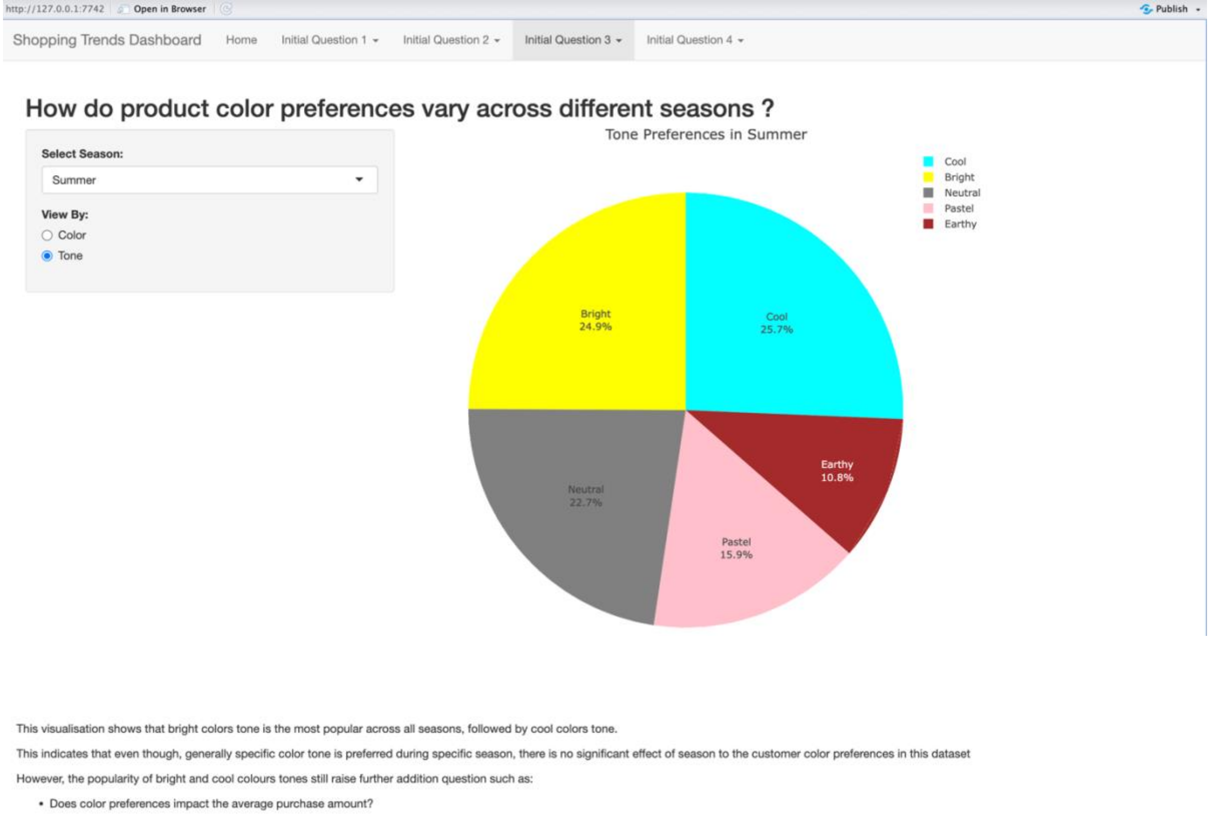
Revenue Distribution Across Seasons

Revenue Distribution Across Seasons

This visualisation reveals that Southern region of USA has the highest revenue during Fall season compared to another seasons

Moreover, the pie chart shows that Fall season has the highest income compared to other seasons

Appendix H: Initial Question 3 Dashboard



How do product color preferences vary across different seasons ?

Select Season: Summer

View By:
○ Color
● Tone

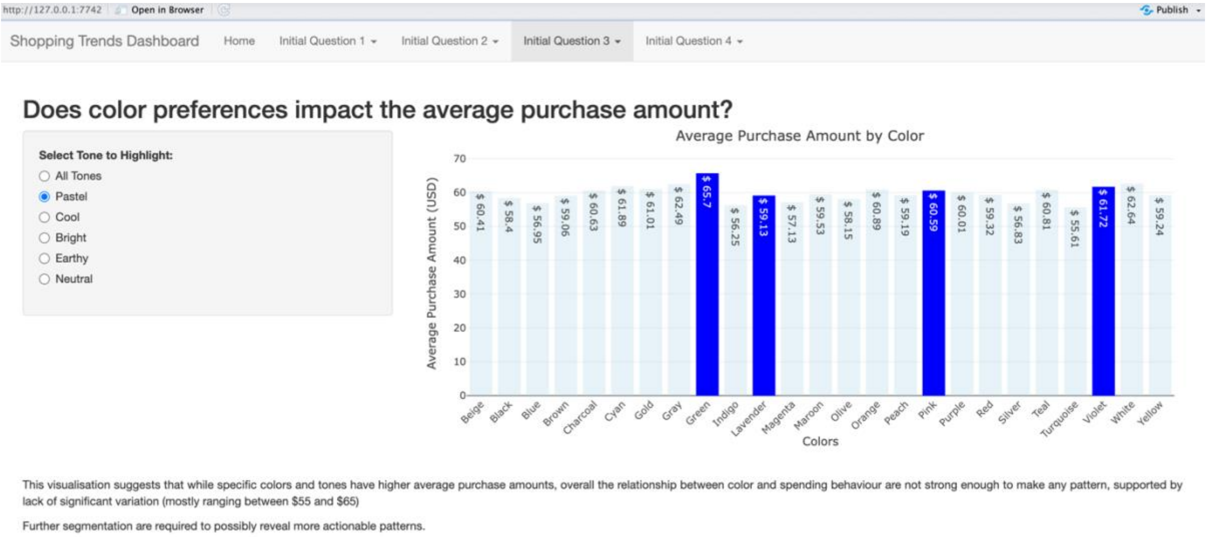
This visualisation shows that bright colors tone is the most popular across all seasons, followed by cool colors tone.

This indicates that even though, generally specific color tone is preferred during specific season, there is no significant effect of season to the customer color preferences in this dataset

However, the popularity of bright and cool colours tones still raise further addition question such as:

- Does color preferences impact the average purchase amount?

Appendix I : Reflective Question 1 for Initial Question 3 Dashboard



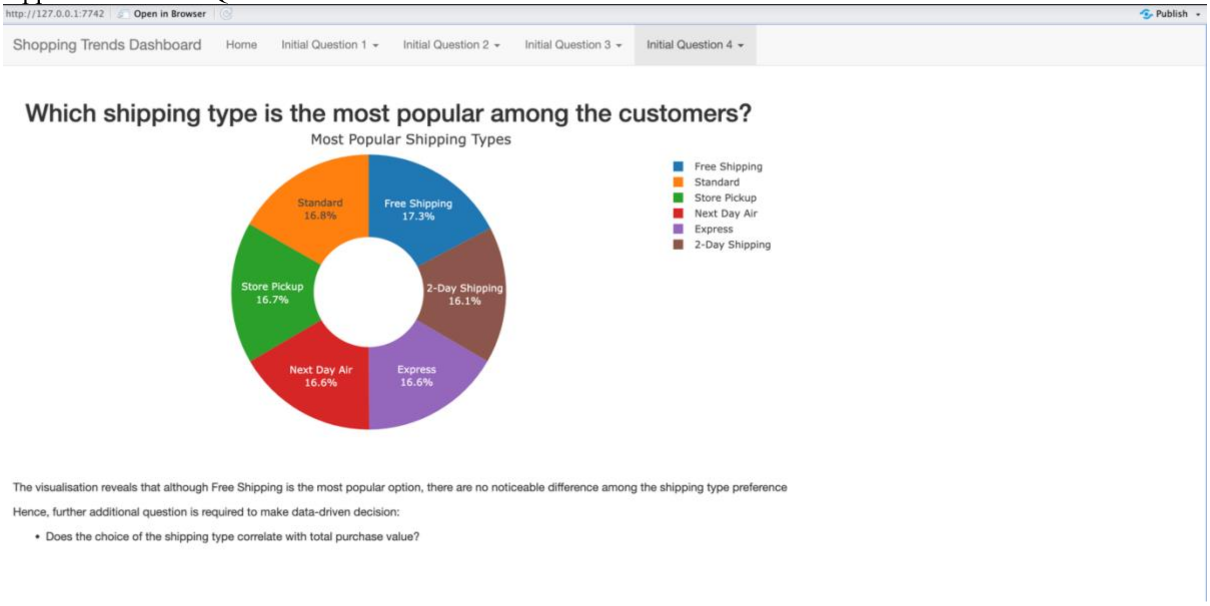
Does color preferences impact the average purchase amount?

Select Tone to Highlight:
○ All Tones
● Pastel
○ Cool
○ Bright
○ Earthy
○ Neutral

This visualisation suggests that while specific colors and tones have higher average purchase amounts, overall the relationship between color and spending behaviour are not strong enough to make any pattern, supported by lack of significant variation (mostly ranging between $55 and $65)

Further segmentation are required to possibly reveal more actionable patterns.

Appendix J: Initial Question 4 Dashboard



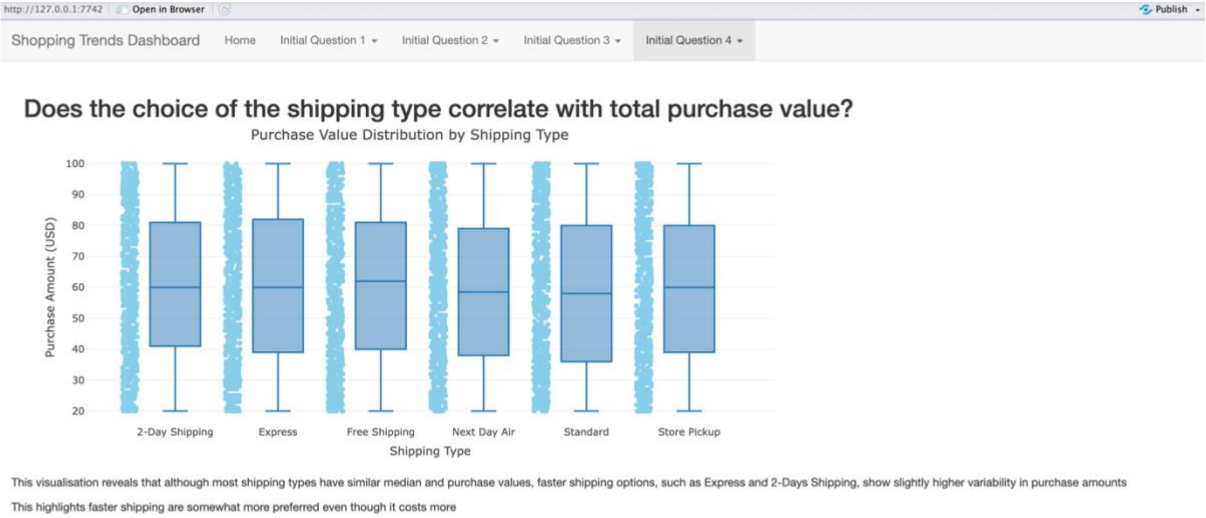
Which shipping type is the most popular among the customers?

The visualisation reveals that although Free Shipping is the most popular option, there are no noticeable difference among the shipping type preference

Hence, further additional question is required to make data-driven decision:

- Does the choice of the shipping type correlate with total purchase value?

Appendix K : Reflective Question 1 for Initial Question 4 Dashboard



Does the choice of the shipping type correlate with total purchase value?

This visualisation reveals that although most shipping types have similar median and purchase values, faster shipping options, such as Express and 2-Days Shipping, show slightly higher variability in purchase amounts

This highlights faster shipping are somewhat more preferred even though it costs more