BSD2223 DATA SCIENCE PROGRAMMING II

SEMESTER 2 2022/2023

PROJECT TITLE:

EXPLORATORY DATA ANALYSIS OF E-COMMERCE PURCHASES

MATRIC ID	NAME	SECTION
SD21027	AIMAN ZHARFAN BIN EISMA HADIE (Leader)	01G
SD21018	NURATIQAH BINTI MOHD ARIS	01G
SD21032	DHARUMASHAN A/L BATHIBAN	01G
SD21063	TEAN JIN HE	01G

LIST OF CONTENT

1. INTRODUCTION

1.1 What is study about?

The study of this project is about e-commerce transactions. The dataset includes facts about client transactions such as customer IDs, product IDs, quantities, unit pricing, and transaction dates. The dataset enables the examination of numerous elements of e-commerce, including consumer behaviour, product popularity, revenue analysis, and country-specific information.

1.2 Objectives of the study?

The objectives of this project are:

1. To analyse sales performance which evaluates sales performance by using revenue, quantity sold and top-selling products.
2. To understand customer behaviour during making purchases of the products.
3. To investigate seasonal trends where the sales is higher.
4. To gain deeper understanding of e-commerce purchases and provide valuable insights for business decision making.

2. PROJECT DESCRIPTION

2.1 Motivated of this project

The desire to enhance business outcomes through data-driven decision-making led to the selection of this project. Businesses may learn more about different client categories, their traits, and preferences by utilising customer segmentation. With this information, businesses may improve their marketing strategies, raise consumer satisfaction levels, and boost sales.

2.2 Reason of this project to be chosen

The reason of this project to be chosen are, we :

- Can uncover valuable insights into customer preferences, purchasing patterns, and trends.
- Can help enhance decision-making and customer targeting.
- Can gain a better understanding of which marketing strategies and promotions are most effective in driving sales.

2.3 Importance of this project and innovation values

- **Improved Targeted Marketing**: By identifying and comprehending certain client segments, organisations may create specialised marketing efforts. Companies may

improve customer happiness, create loyalty, and increase conversion rates by offering personalised experiences.

- **Resource Allocation Optimisation**: Organisations may more efficiently distribute resources by having a deeper grasp of client groups. Businesses may prioritise marketing expenditures, enhance product development, and streamline customer service by identifying high-value groups, which will increase returns on investment (ROI) and save costs.
- **Competitive Advantage**:Utilising customer segmentation analysis gives organisations a competitive edge by allowing them to set themselves apart from rivals. Organisations may adapt their offers, improve their value propositions, and take a distinctive position in the market by locating niche markets or underserved areas.
- **Predictive analytics**: The study of customer segments is the basis of this discipline. Organisations may predict future trends, spot possible churn risks, and proactively address customer demands by analysing previous customer behaviour. This improves customer retention and boosts long-term profitability.

2.4 Idea on how this project can be extended or improved

- **Dynamic Segmentation**: Building on the current project, dynamic consumer segmentation may be made possible by combining real-time data sources like social media feeds, transactional data, or website interactions. Businesses would have access to the most recent information and be able to react quickly to changing client preferences as a result.
- **Including additional Data**: Adding additional data sources, such as demographic information, economic indicators, or customer surveys, to the customer segmentation dataset can give a more thorough knowledge of the different consumer categories. Businesses would be able to make more exact strategic decisions because of this enhanced analysis.
- **Personalised Recommendation Systems**: Businesses may give tailored product or service suggestions to specific consumers by integrating customer segmentation with recommendation systems, further increasing the customer experience and boosting revenues.
- **Customer Lifetime Value (CLV) Analysis**: Including CLV analysis in the project's scope can assist organisations in determining the long-term profitability of various customer groups. This would make it possible for businesses to deploy resources wisely, improve their acquisition tactics, and spot cross-selling and upselling opportunities

3. DATA DESCRIPTION

The dataset is from Customer segmentation and an E-commerce database that lists purchases made by about 4000 customers over a period of one year (from 2010/12/01 to 2011/12/09). The dataset contained 541909 entries with 8 columns, including ' InvoiceNo ', ' StockCode ', ‘Description ', ‘Quantity ', ‘InvoiceDate ', ‘UnitPrice ', ' CustomerID 'and ' Country '.

1. **InvoiceNo**: Invoice number is qualitative data. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
2. **StockCode**: Product (item) code is qualitative data. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. **Description**: Product (item) name is qualitative data. Nominal.
4. **Quantity**: The quantities of each product (item) per transaction is quantitative data. It is discrete .
5. **InvoiceDate**: Invoice Date and time is qualitative data. Nominal, the day and time when each transaction was generated it represents temporal information and does not have quantitative meaning on its own.
6. **UnitPrice**: Unit price is quantitative data. Continuous, Product price per unit in sterling.
7. **CustomerID**: Customer number is qualitative data. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. **Country**: Country name is qualitative data. Nominal, the name of the country where each customer resides.

4. DATA PREPARATION

Data preparation would involve cleaning and preprocessing the data to make it suitable for analysis. This process would be included:

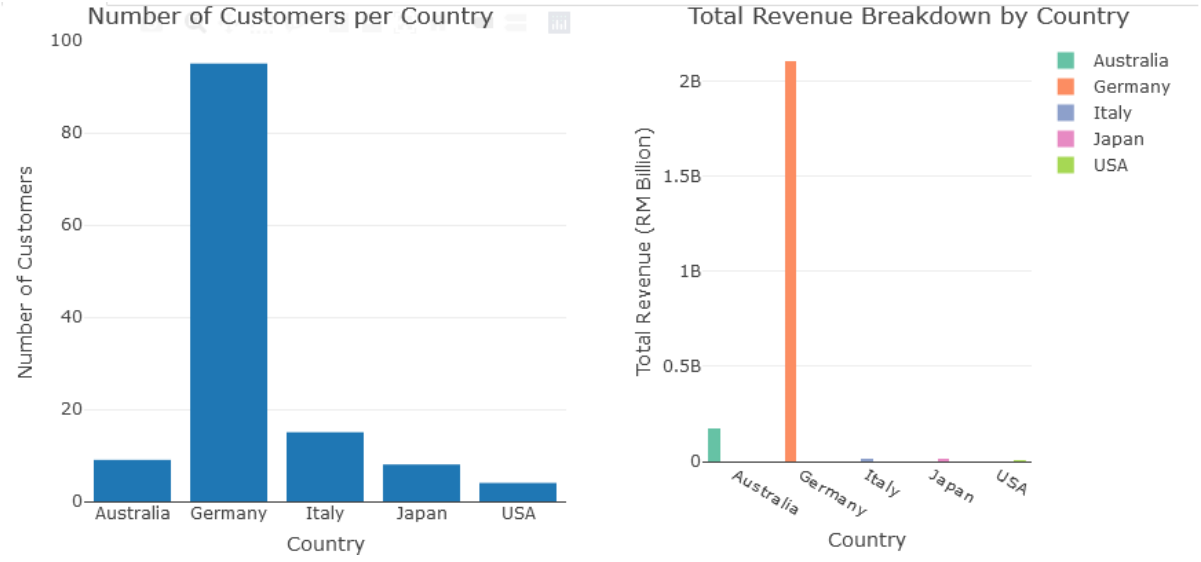
1. Include the specific library to load all the modules.
2. The column names can be converted to a suitable format.
3. The percentage of missing values in each column can be calculated so we know which columns have a high percentage of missing values.
4. Columns with null values or no variation can be dropped.
5. Duplicate entries and arrange by InvoiceNo which can get better data for analysis.

5. DATA ANALYSIS, RESULTS AND DISCUSSION

In this section we will show the visualisation and interpret all the analysis in our dashboard, which consists of 3 pages:

1. Page 1:

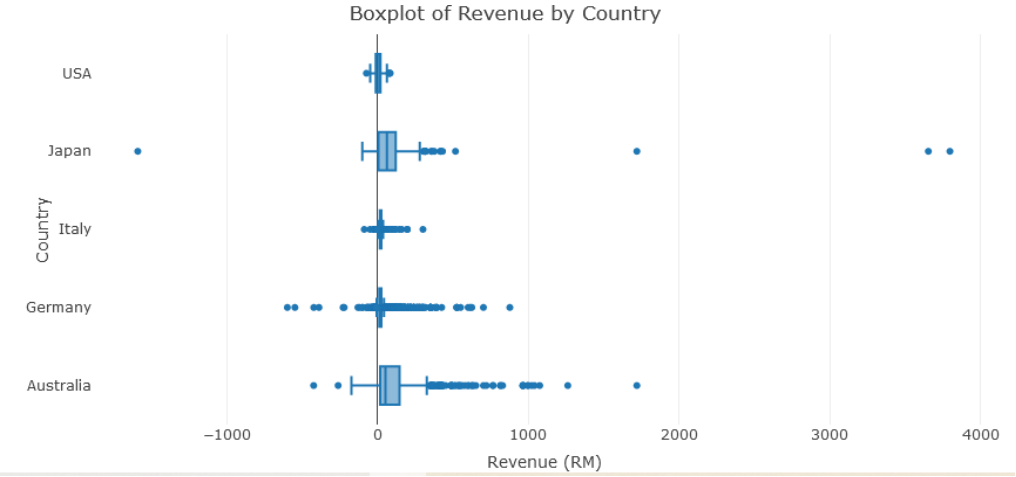
   In page 1, we have bar chart and boxplot.



From the bar chart above, we can see that Germany has the highest number of customers which is 95 people while for the lowest number of customers is 4 which belongs to USA.

The total revenue breakdown by country chart is shows that Germany is the highest one with total revenue is RM 221.69billion while for the lowest total revenue is RM 4.50billion.

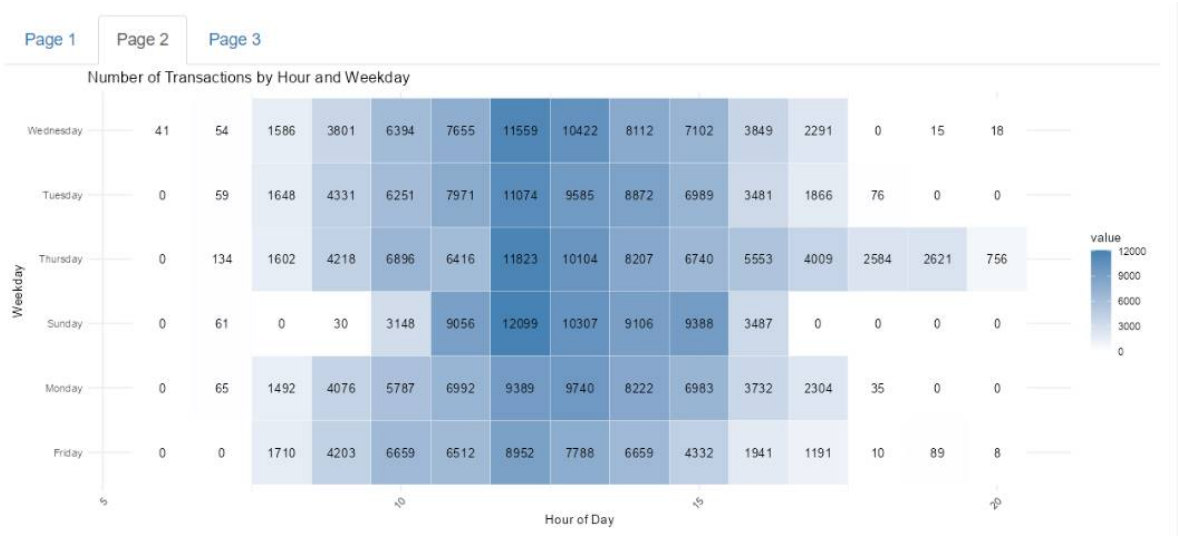
From this two bar plot we can say that, the relationship of number of customers and total revenue is directly proportional which is if the number of customers increase, the total revenue also increases.



The boxplot of revenue by country is including maximum, minimum, upper boundary, lower boundary, q1, median, q3 and outliers which are selected by countries. Outliers are data points that fall outside the whiskers of the boxplot. These points are significantly higher or lower than the rest of the data. From this visualization, we can see that Japan has the highest outlier which is RM3794.40 and this means that Japan has a revenue value that is much higher than the rest of the countries in the dataset. By analyzing the boxplot, we can visually compare the revenue distributions across different countries, identify the range of values, and detect any extreme values (outliers) that may indicate unique characteristics or anomalies in a specific country's revenue.

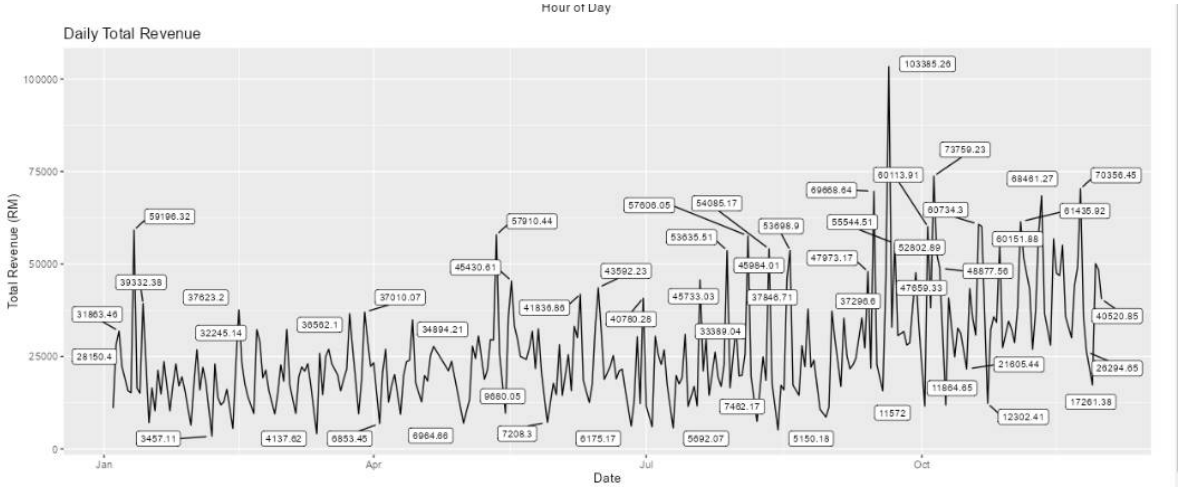
2. Page 2

   In page 2, we have a heatmap chart and line graph.



Heatmap chart of number of transaction by hour and day for January until December

Based on the heatmap chart above,it shows that Sunday and 12 PM have the highest number of transactions which is 12099 which suggests some interesting insights about customer behaviour, business strategy and also efficiency of the platform on that day such as weekend shopping, lunchtime shopping, strategic timing, user experience considerations and customer preferences.

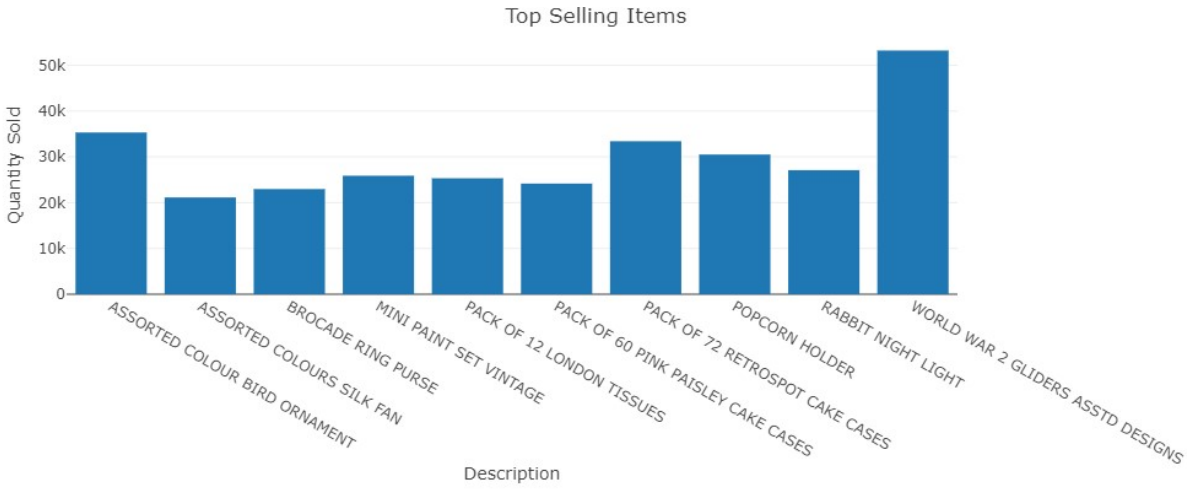


Line graph for daily total revenue and date from january to december

Based on the line graph above, we can see that in the month of September has the highest total revenue which is RM103385.26 billion compare to other months. From here, there are several factors that contribute to this which are seasonal factors, promotions and sales, timing for paychecks, products releases or events and holiday preparation.

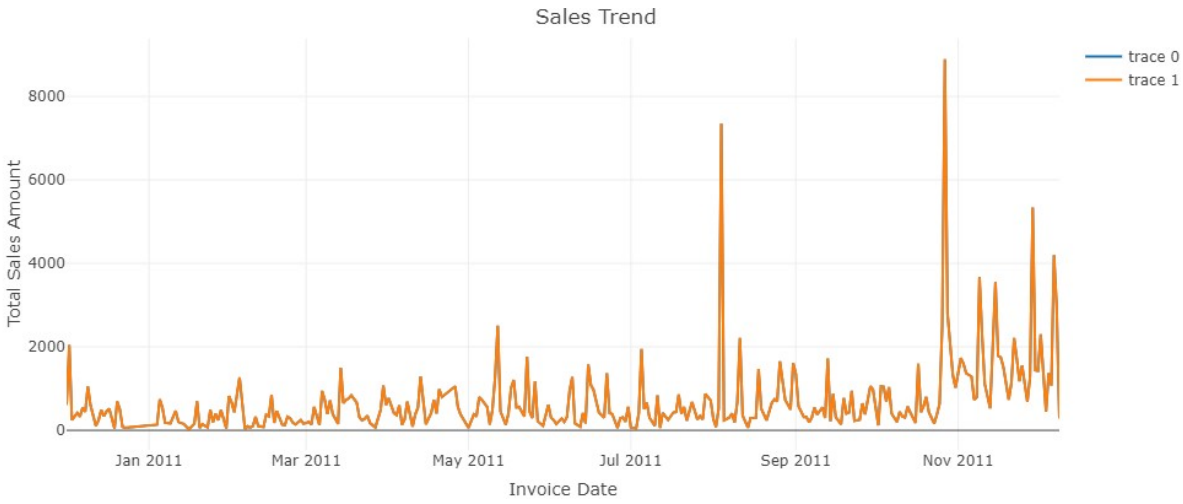
3. Page 3

In page 3, we have a bar chart, line chart, and a scatter plot.



Bar Chart of top selling items

The bar chart represents the top selling items based on the quantity sold over the period of one year. Each bar represents a different item, and its height represents the number of units sold for that particular item. By looking at the bar heights, we can identify the items with the highest sales volume or popularity. The items displayed above are the top 10 most sold items over the period of a year. In the chart above, the item "World War 2 Gliders Asstd Designs" is the best-selling item with a quantity sold of 53.215k units.



Line graph for sales trend

The line chart represents the sales trend over time. The x-axis represents the invoice dates, which have been converted to a datetime format. The y-axis displays the total sales amount for each date. There is a significant increase in sales from October to December. This indicates a period of higher sales performance or increased customer demand during those months. The highest point on the line chart, is on October 27, stands out as a notable peak in sales with a total sales amount of RM 8888.88. This is because October is known for various holidays and events such as Halloween or the beginning of the holiday shopping season. These occasions often lead to increased consumer spending and higher sales for businesses that offer relevant products or services.

Scatter Plot of Quantity and Unit Price

The scatter plot represents the relationship between the quantity sold and the unit price for each item. The x-axis represents the quantity sold, while the y-axis represents the unit price. The majority of data points are in the lower range of quantity values. This can be because of volume discounts or bulk pricing, where customers tend to purchase larger quantities at a lower price per unit. There are a few points that deviate from the graph. This could represent a bulk purchase or a special pricing arrangement for a specific customer or order.

6. CONCLUSION

In conclusion, the analysis of the E-commerce database has provided valuable business insights and achieved its goals. The project has allowed the business to adapt strategies, optimize operations, and increase customer satisfaction by visualizing trends, revenue, and popular products. By analyzing sales performance, the business has gained a thorough understanding of revenue generation, quantity sold, and top-selling products, which has enabled them to focus on high-demand products, modify pricing policies, and allocate resources efficiently. The project has also provided insightful information about customer preferences and purchasing practices, allowing the business to personalize marketing initiatives, enhance customer targeting, and improve the shopping experience. By analyzing seasonal trends, the business can foresee changes in demand, modify inventory levels, and optimize resource allocation. Overall, this analysis has deepened the understanding of consumer behavior, seasonal trends, and purchasing patterns, empowering the business to make data-driven decisions that foster expansion, enhance customer comprehension, and maintain a competitive edge in the e-commerce sector.

7. LIMITATIONS OF THE STUDY

The E-commerce analysis project has limitations that should be considered. The one-year duration may not capture long-term trends and seasonality patterns, necessitating a longer analysis period. The impact of cancellation transactions on revenue and consumer behavior is not fully explored, which could affect revenue and customer retention. The analysis's geographic constraints limit the broad applicability of findings, emphasizing the need for a wider representation of regions. The focus on overall revenue and top-selling products overlooks the potential insights from product characteristics and categorizations. These limitations highlight the importance of conducting further research to enhance the depth and applicability of the analysis.

8. REFERENCE

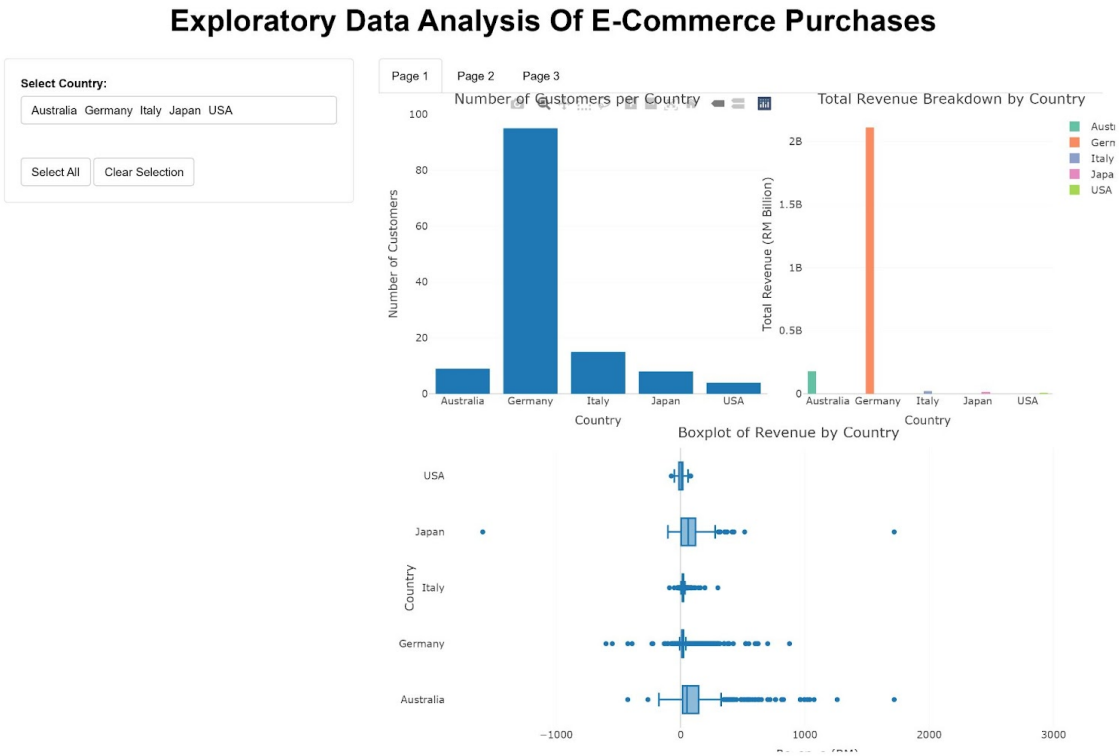
a. RS, A. (2021, January 29). *Build your first web app dashboard using Shiny and R*. Medium. https://medium.com/free-code-camp/build-your-first-web-app-dashboard-using-shiny-and-r-ec433c9f3f6c
b. Lee, A. (2022, August 27). *Exploratory Data Analysis on E-Commerce Data*. Medium. https://towardsdatascience.com/exploratory-data-analysis-on-e-commerce-data-be24c72b32b2
c. *Build Interactive Dashboards With R Shiny - Tilburg Science Hub*. (n.d.). Build Interactive Dashboards With R Shiny - Tilburg Science Hub. https://tilburgsciencehub.com/building-blocks/collaborate-and-share-your-work/publish-on-the-web/shiny-apps/
d. *Ecommerce Analytics: How to Use Data to Grow Sales*. (2021, August 6). The BigCommerce Blog. https://www.bigcommerce.com/blog/ecommerce-analytics/

9. APPENDIX

Dataset:

https://www.kaggle.com/datasets/carrie1/ecommerce-data?datasetId=1985&sortBy=voteCount

Dashboard:



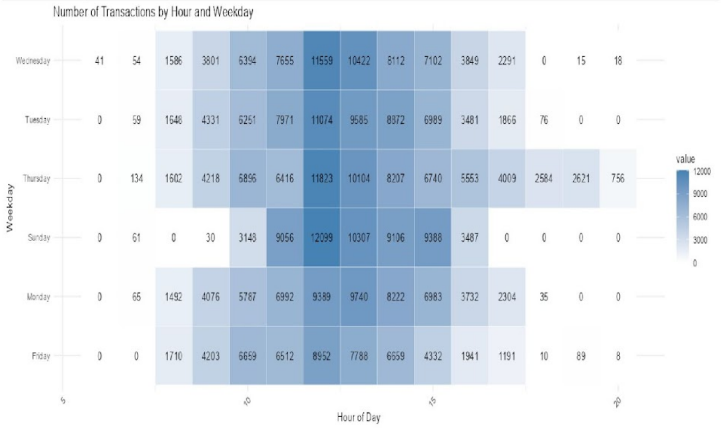
Exploratory Data Analysis Of E-Commerce Purchases

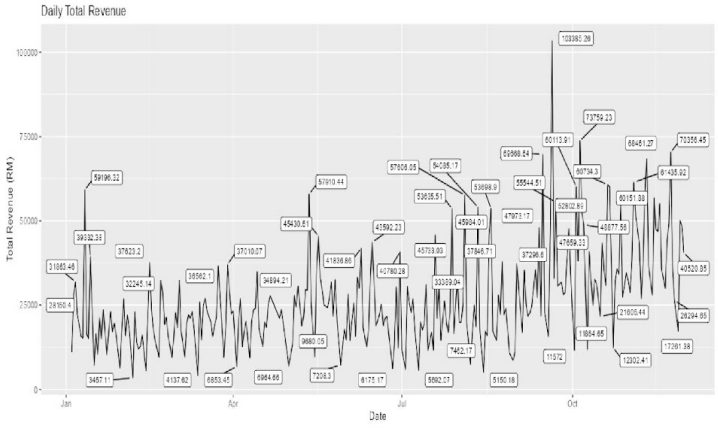
Select Month:

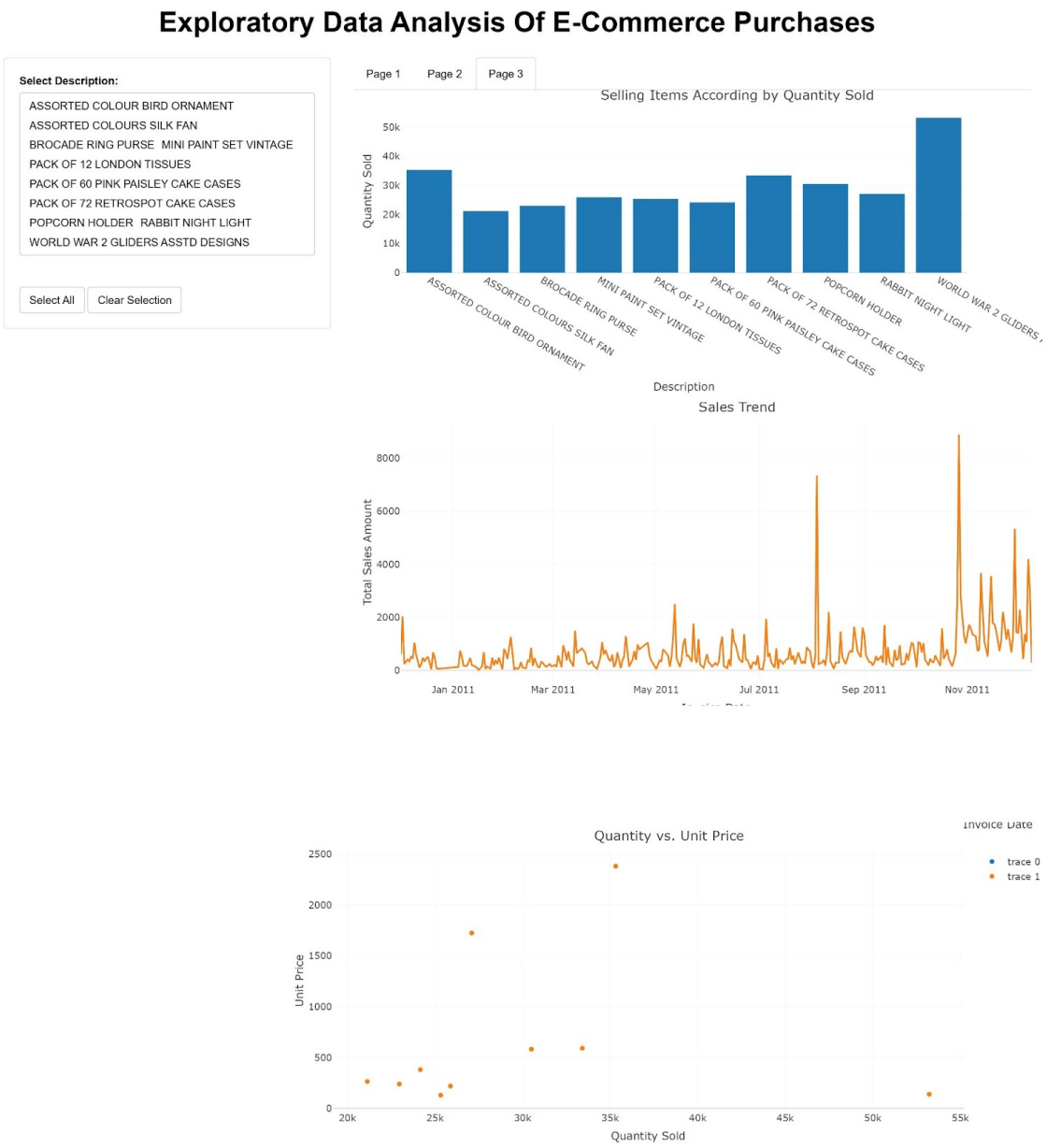
January February March April May
June July August September
October November

Select All | Clear Selection

Page 1 | Page 2 | Page 3







Source Code:

https://drive.google.com/file/d/1wlpUfrBnY4jJ_J2nZrlDsUJenJyhga37/view?usp=sharing