|  |
| --- |
| Data Exploration Project :  Retail store analysis  e 1: Tableau Public  Public  e 1: Tableau Public  Data Exploration &  Visualization  FIT5147  FIT5147  Monash UniversityFIT5147  FIT5147 Monash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversitydhMonash University Monash University47  Monash UniversityFIT5147  FIT5147 |
| Monash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversitydhMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversitydhMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversitydhMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversityMonash UniversitydhMonash University Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600  Aryan Jain  31418600 |





Assignment 1

### **Introduction**

With the snowballing growth of globalization, local businesses find it quite challenging competing with global markets. One such area that has suffered significantly is retail. Small and medium scale business may not necessarily have the capital or motivation to hire a team of analyst. Nevertheless, they can still do some elementary analysis, customer segmentation and even predict the growth of their business.

The focus of this project will be on analysing a retail store’s customer database to understand customer behaviour using metrics such as conversion rate and total amount spent in recent transactions.

* The app will also look into which categories of product are more prone to returns and which age groups are more likely to return.
* Trend analysis will be done to identify seasonality and overall growth.
* Finally, factors such as gender, age and location will be used as explanatory variables to analyse the effect on customer’s buying decision.

The analysis done can be used for customer segmentation to provide exclusive offers to existing customers and build a model for customer retention.

### **Data Checking & Wrangling**

The data used in this report is collected from an anonymous retail store and can be accessed through [Kaggle](https://www.kaggle.com/darpan25bajaj).

The data contains 3 csv files from which selected features will be combined to answer the analysis questions.

* Customer: Customer information - 5K row x 4 columns
* Transaction: Transaction of customers - 23K row x 10 columns
* Product Hierarchy: Product information - 23 row x 4 columns

After Initial Examination of Data in R, it was discovered that although relatively clean, there were some issues with the data. The *customer* dataset had some missing values in “*Gender*” and “*city\_code*” columns as shown below.

Figure 1: Missing values visualization in Customer

A screenshot of a cell phone

Description automatically generated

Table 1: Missing Values in Customer

A screenshot of a cell phone

Description automatically generated

Figure 1 shows that the missing data is less than 0.1% and the missing observations can be seen in Table 1. Also, the values in gender columns were changed from abbreviations to full form which is shown later in Table 4.

Furthermore, the *tran\_date* column in the *Transaction* dataset was stored in multiple formats and required tidying up as shown below

Table 2: Glimpse of the transactions

A screenshot of a cell phone

Description automatically generated

The *tran\_date* variable was fixed using *parse\_date\_time* function in R which formatted the variable uniformly and changed the variable to *datetype* so that it can be used for further analysis.

Table 3: Glimpse of the transaction (Cleaned)

A screenshot of a cell phone

Description automatically generated

The Table 3 shows cleaned *tran\_date* column for the same observations.

Finally, the cleaned *Customer* and *Transaction* datasets were merged together (by *customer\_id)* and again merged with the already clean Product dataset (by *prod\_cat\_code*) using the *merge* function and the duplicates were removed with *distinct* function. Two new columns *age* and *age\_group* were also created using *mutate()* function in R. The variable age was created with *age\_calc()* function and *age\_group* was created using *cut()* function. The final dataset displays product and customer information for every transaction as visible in Table 4.

Table 4: Glimpse of Final Merged Dataset

### **A screenshot of a cell phone Description automatically generated**

### **Design**

The five design sheet was referred to for designing the entire app.

**First Sheet :** This sheet explorers the various ideas and brainstorm all the possible questions that can be formed. Then the questions are categorized and properly sorted into 3 main tabs.

Figure 2: First sheet of the FDS

A picture containing text

Description automatically generated

**Second Sheet :** This sheet expands of those ideas formed in the first sheet. Analysis will include best performing products and also which products are often returned.

**Third Sheet :** This sheet will focus on studying the business based on various attributes to get the trend line and sales prediction.

**Fourth Sheet :** This sheet looks at which customers are at risk to allow for a customer retention program.

**Fifth Sheet :** Finally, the last sheet combines the three approaches and pick the best parts out of the individual sections.

For the design of the app, a tabular structure was put into place. All the individual modules were included as tabs of the app with individual UI elements relevant for each tab.

Graphical user interface, application, Teams

Description automatically generated

**Product Analysis :** This is the first module of the app which used bar plots to explore the retail chains data for patterns in products sold based on their category. A shiny sidebar layout was used for the UI part of the app where the user could enter the time period and category that they are interested in exploring. Sub-tab panels were included for this part to reduce the overall visual bulk of the app. The tabs explored sales and returns with bar plots and pie charts.

**Business Analysis :** This module explores the performance of the business in various product categories with a time series line plot. The line-plot has user selectable categories and showcases the overall sales throughout the years. The second plot explores the seasonality using user selectable months which will help the user understand the sales for various seasons and months.

**Customer Segmentation :** This tab was designed to be used as a customer segmentation and retention module which could be used to identify at risk customers and loyal customers and create a focused marketing plan. The tab used shiny’s radio buttons and selectizeInput() function for the user intractability. The user would be able to subset the data based on age and gender. The use of bar plots made the most sense as the data was mostly categorical. The facet by city code will also help businesses identify the impact of location on sales and create marketing plans accordingly.

### **Implementation**

**Libraries Used:**

**R Shiny**: This is the main library used to create the layout the user interactive elements of the app.

**ShinyThemes**: This package was used to incorporate the flatly theme for the shiny app.

**Tidyverse**: This package contains various libraries including but not limited to:

* **Ggplot2**: Used to create the visualisations such as bar plots and line plots
* **Dplyr**: Used for data manipulation and filtering.
* **Tidyr**: Included a set of functions used to tidy the data.
* **Readr**: included function fuch as read\_csv() for reading the data.

**Plotly**: This was used to plot interactive bar plots in the first section of the app.

**Lubridate**: This package is used to perform wrangling on datetype data.

**Eeptools**: This package is used to calculate the age from DOB.

**Scales**: This package is used to manipulate the scale in the scale\_continuous parameter of the ggplot.

### Product Analysis – Plot 1

|  |  |
| --- | --- |
| Visualization graph name | Bar Plot |
| Purpose | Display total sales based on product category and time period |
| Visualization Graph | **Chart, bar chart  Description automatically generated** |
| Implementation | Group the data by product category and use the summation of total sales to plot using ggplot. |
| Tab Name | Product Analysis |
| Libraries used | Ggplot, plotly, shiny |
| Filters | Qty > 0 |

### Business Analysis – Plot 2

|  |  |
| --- | --- |
| Visualization graph name | Line Plot |
| Purpose | Time series line-plot to display progression of business |
| Visualization Graph | **Chart, line chart  Description automatically generated** |
| Implementation | Group the data by year calculated by lubridate and use the summation of total sales to plot using ggplot. |
| Tab Name | Business Analysis |
| Libraries used | Ggplot, shiny |
| Filters | Qty > 0 |

### Customer Segmentation – Plot 3

|  |  |
| --- | --- |
| Visualization graph name | Bar Plot |
| Purpose | Display Seasonality |
| Visualization Graph | A picture containing implement, stationary, pencil, computer  Description automatically generated |
| Implementation | Group the data by product category and month and use the summation of quantity to plot bar plots using ggplot. |
| Tab Name | Business Analysis |
| Libraries used | Ggplot, shiny |
| Filters | Qty > 0 |

Figure 4: Popularity of product categories for different age groups

A screenshot of a cell phone

Description automatically generated

Figure 4 shows that *Books*, *Electronics* and *Home and Kitchen* are the most popular category for all age groups. It also shows that people in the aged between *35 to 44* spends the most followed by people aged from *25 to 34* and people older than *44* spends the least in this group.

### Influence based on **Location**

Figure 5: Location codes vs Sales

A picture containing drawing

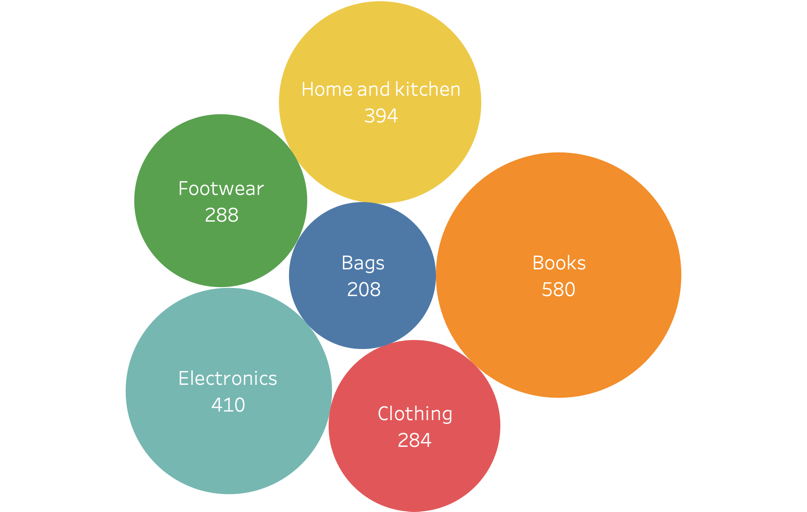
Description automatically generated

Figure 5 shows the variance in sales for different cities. It does reveal that the city codes 3,4,5,7,8 and 10 sell a lot more than 1,2,6 and 9. It can also be seen that the city code 4 is the city with the highest sales city while 6 is the one with the lowest.

**How does the category of product and age group affect returns?**

For looking at returns, the data had to be filtered to only show observations with negative quantity (*Qty*) which was done using *filter()* in R.

Figure 6: Return for all Categories

A picture containing stereo

Description automatically generated

Figure 7: Return for all age groups

Figure 6 reveals that *Books*, *Electronics* and *Home & Kitchen* are returned most frequently and figure 7 shows that people aged between 35 to 44 returns products the most at 42%.

**What is the overall trend and is there seasonality?**

Figure 8: Time series graph of the retail chain

A close up of a map

Description automatically generated

Figure 8 shows the average sales for all categories throughout the period of 2011 – 2014. Unfortunately for the store, the sales have dropped significantly and linearly in 2013. For books alone the sales have dropped from 4.5M to approx. 700K in just 1 year.

Figure 9: Analysing seasonality for all categories

A picture containing implement, stationary, pencil, computer

Description automatically generated

Figure 9 reveals a lot of very interesting insights. Sales for books, clothing electronics, home & kitchen and even bag snowballs around the end and start of the year with some exceptions. This may suggest holiday surge around Christmas and New Year.

Clothing and books show clear peaks around the start and end of the year which may be partially due to the holidays but in the case of clothing it may also be due to the change of season. Electronics sales peaks many times throughout the year which may be due to the huge discounts during back to school season, Christmas, cyber Monday and other such events.

### **Conclusion**

The retail store analysis really showed some interesting results. It’s safe to assume from the analysis that age, gender and location does have an effect on buying decision. Although, a little stereotypical, the dataset clearly reveals that would are more likely to shop for bags and footwear. Similarly, age and location are also important factors while doing customer segmentation as it clearly shoes that certain categories are more popular among certain age groups and locations. This can be very useful in customer retention and personalised advertisements.

The analyses also revealed that there is a bias in product category and age when it comes to returning product. Middle-aged group not only shops the most but are also returns products the most. And categories like books and electronics are more prone to return.

It was also revealed from figure 9 that seasonality does exist in this retain store and should be kept in mind when introducing discounts and limited time offers. Finally figure 8 showed the calamitous decline of the retail store which could be due to a number of reasons which are beyond the scope of this analyses.

### **Reflection**

The project was quite informative and a good learning experience for working with realistic data. It challenged me to look at data with an analytical mindset identify patterns with relatively simple visualizations. It can be used as a base for customer segmentation and build models for customer retention.

Although, in hindsight I would have changed the visualization in figure 3 to show proportions of gender population rather than showing discrete counts as the sample may be biased and have higher number of observations for male customers which makes it slightly more difficult to understand the graph and draw inferences from it.

Also, in figure 6, I would like to change the counts to percentage as it is easier to observe from a glance. Lastly, I would have also liked to explore the returns against other subcategories and explore seasonality using cumulative data for all categories using a line plot.

### **Bibliography**

### Dataset:

Bajaj, D. (2019, November). Retail Case Study Data, Version 1. Retrieved August 21, 2020 from <https://www.kaggle.com/darpan25bajaj/retail-case-study-data>

### R Packages:

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, <https://doi.org/10.21105/joss.01686>

Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2020). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.5.2. https://CRAN.R-project.org/package=naniar

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25.

URL <http://www.jstatsoft.org/v40/i03/>.

Jared E. Knowles (2020). eeptools: Convenience Functions for Education Data. R package version 1.2.4. <https://CRAN.R-project.org/package=eeptools>

Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. <https://CRAN.R-project.org/package=scales>

Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. <https://CRAN.R-project.org/package=kableExtra>

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot