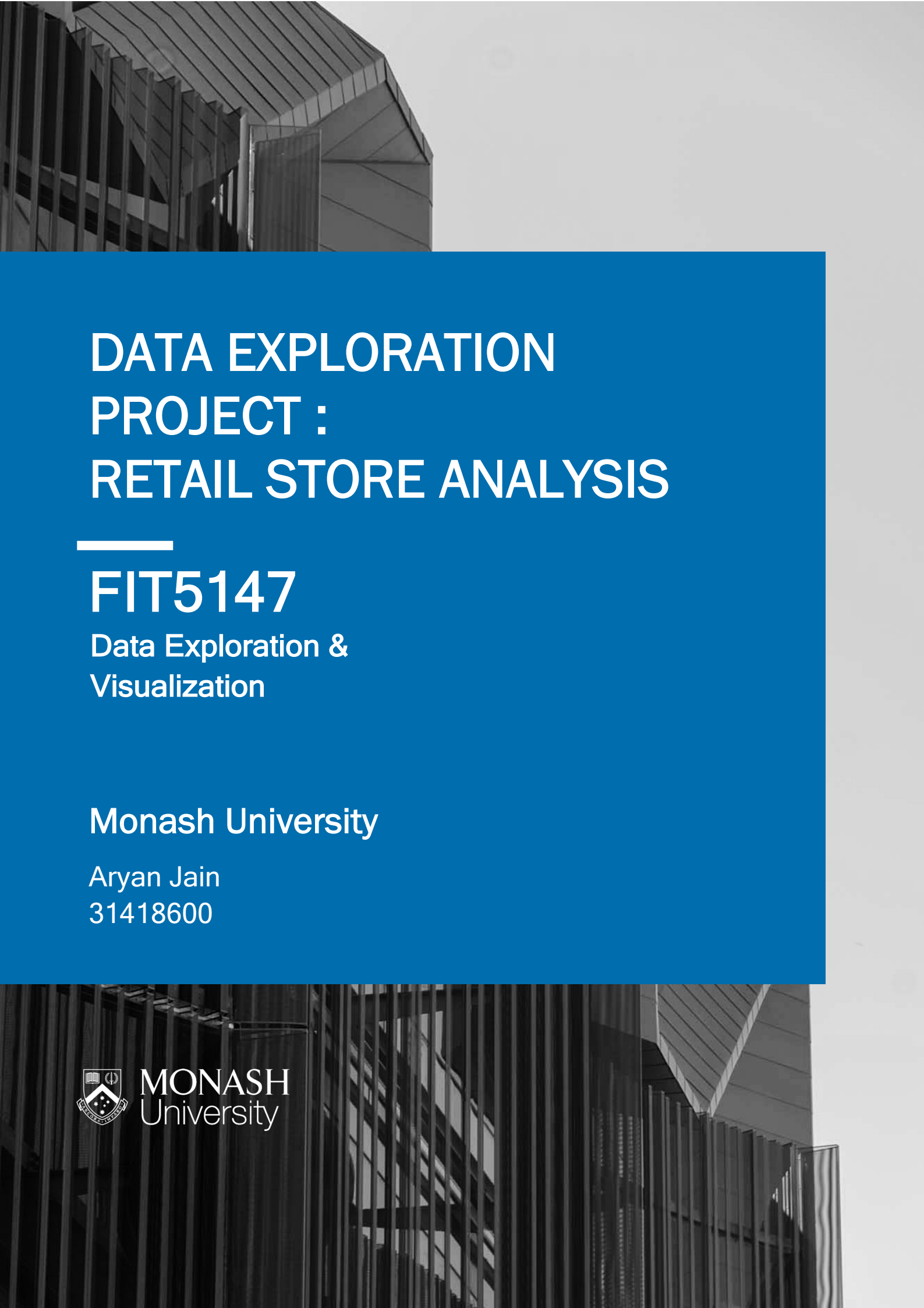


DATA EXPLORATION PROJECT : RETAIL STORE ANALYSIS

FIT5147

Data Exploration & Visualization

Monash University

Aryan Jain
31418600

MONASH University

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.

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

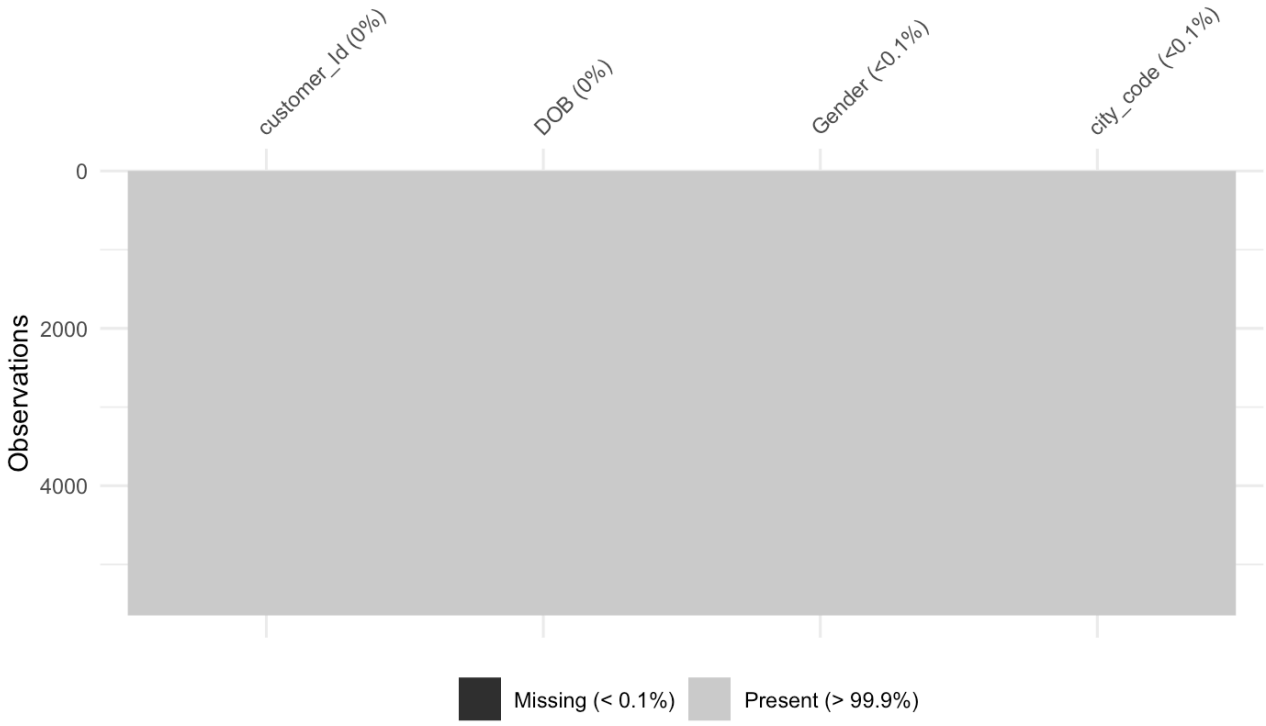


Table 1: Missing Values in Customer

customer_Id <int>	DOB <chr>	Gender <chr>	city_code <int>
267199	14-02-1970	*NA*	2
271626	02-06-1970	*NA*	6
268447	14-07-1970	M	*NA*
268709	09-09-1970	F	*NA*

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

transaction_id <dbl>	cust_id <int>	tran_date <chr>	prod_subcat_code <int>	prod_cat_code <int>	Qty <int>	Rate <int>	Tax <dbl>	total_amt <dbl>	Store_type <chr>
3268991	272172	2/8/2011	11	6	3	91	28.665	301.665	e-Shop
7073244	269640	11/5/2013	12	5	4	1385	581.700	6121.700	MBR
10861359	272671	25-10-2013	4	4	2	103	21.630	227.630	Flagship store
15741026	271544	28-03-2011	12	5	1	299	31.395	330.395	e-Shop
16165359	273203	1/8/2013	4	1	2	580	121.800	1281.800	e-Shop
18629385	273648	16-01-2012	12	6	3	1174	369.810	3891.810	e-Shop
29740699	269719	31-03-2011	7	5	3	1136	357.840	3765.840	TeleShop
33156503	267524	24-09-2011	1	2	3	545	171.675	1806.675	e-Shop
38816402	274437	12/4/2011	12	5	3	543	171.045	1800.045	TeleShop
41453307	269572	4/8/2012	6	5	4	552	231.840	2439.840	e-Shop

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)

transaction_id <dbl>	cust_id <int>	tran_date <S3: POSIXct>	prod_subcat_code <int>	prod_cat_code <int>	Qty <int>	Rate <int>	Tax <dbl>	total_amt <dbl>	Store_type <chr>
3268991	272172	2011-08-02	11	6	3	91	28.665	301.665	e-Shop
7073244	269640	2013-05-11	12	5	4	1385	581.700	6121.700	MBR
10861359	272671	2013-10-25	4	4	2	103	21.630	227.630	Flagship store
15741026	271544	2011-03-28	12	5	1	299	31.395	330.395	e-Shop
16165359	273203	2013-08-01	4	1	2	580	121.800	1281.800	e-Shop
18629385	273648	2012-01-16	12	6	3	1174	369.810	3891.810	e-Shop
29740699	269719	2011-03-31	7	5	3	1136	357.840	3765.840	TeleShop
33156503	267524	2011-09-24	1	2	3	545	171.675	1806.675	e-Shop
38816402	274437	2011-04-12	12	5	3	543	171.045	1800.045	TeleShop
41453307	269572	2012-08-04	6	5	4	552	231.840	2439.840	e-Shop

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

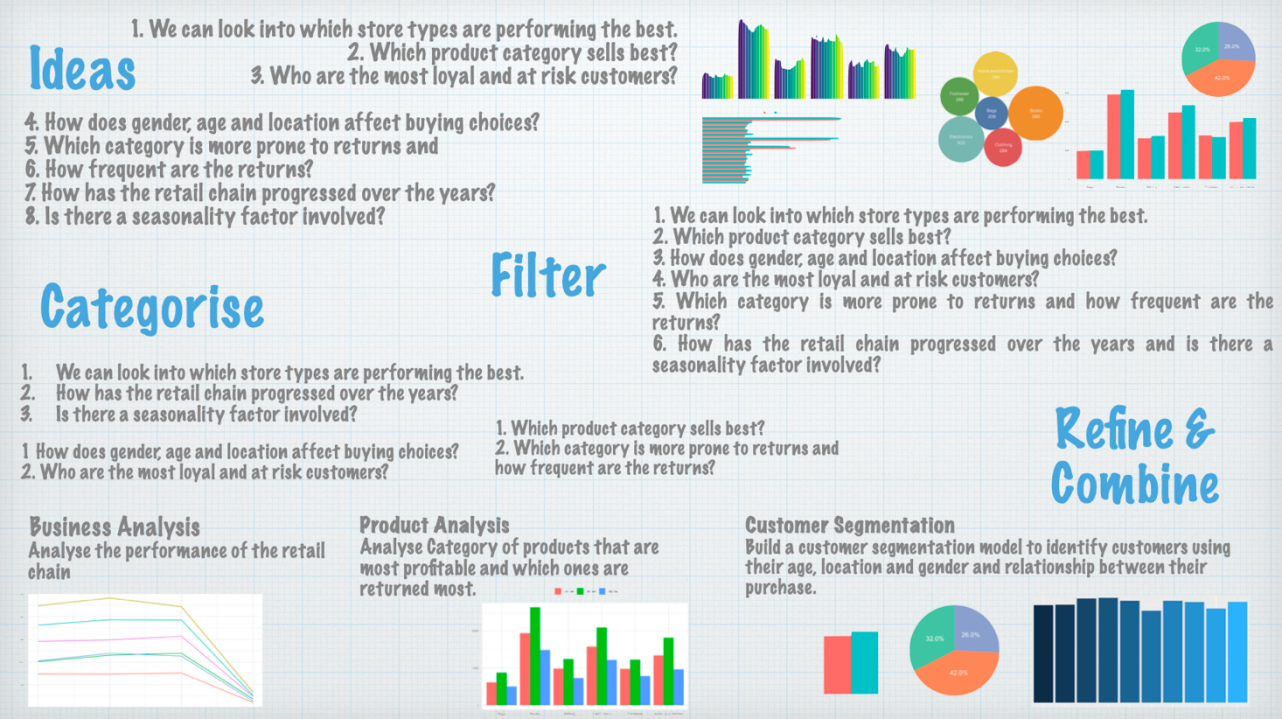
customer_Id <int>	DOB <date>	Gender <chr>	city_code <int>	transaction_id <dbl>	tran_date <S3: POSIXct>	Qty <int>	Rate <int>	Tax <dbl>	total_amt <dbl>	Store_type <chr>	prod_cat <chr>	prod_subcat <chr>	age <dbl>	age_group <fctr>
274066	1976-10-26	Male	8	38110368883	2013-03-05	5	1239	650.475	6845.475	Flagship store	Clothing	Women	43	35-44
268597	1986-05-01	Female	10	76949751078	2013-01-19	4	853	358.260	3770.260	MBR	Clothing	Women	34	25-34
274484	1976-11-13	Female	9	18838561375	2011-08-20	2	324	68.040	716.040	Flagship store	Clothing	Women	43	35-44
268701	1978-02-22	Male	5	10577817962	2012-01-30	5	1342	704.550	7414.550	MBR	Clothing	Women	42	35-44
267134	1992-09-14	Male	2	97764417070	2011-12-09	5	461	242.025	2547.025	MBR	Clothing	Women	27	25-34
268161	1979-06-25	Male	4	44627697269	2012-07-24	3	1309	412.335	4339.335	Flagship store	Clothing	Women	41	35-44
274260	1991-11-10	Male	3	78340973279	2012-01-30	2	630	132.300	1392.300	e-Shop	Clothing	Women	28	25-34
274926	1986-10-20	Male	2	41963699352	2012-10-29	5	224	117.600	1237.600	Flagship store	Clothing	Women	33	25-34
268126	1975-09-12	Female	2	11950520948	2012-08-11	1	1249	131.145	1380.145	TeleShop	Clothing	Women	45	45-54
274618	1992-12-09	Female	7	95974831788	2014-01-13	4	561	235.620	2479.620	Flagship store	Clothing	Women	27	25-34

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



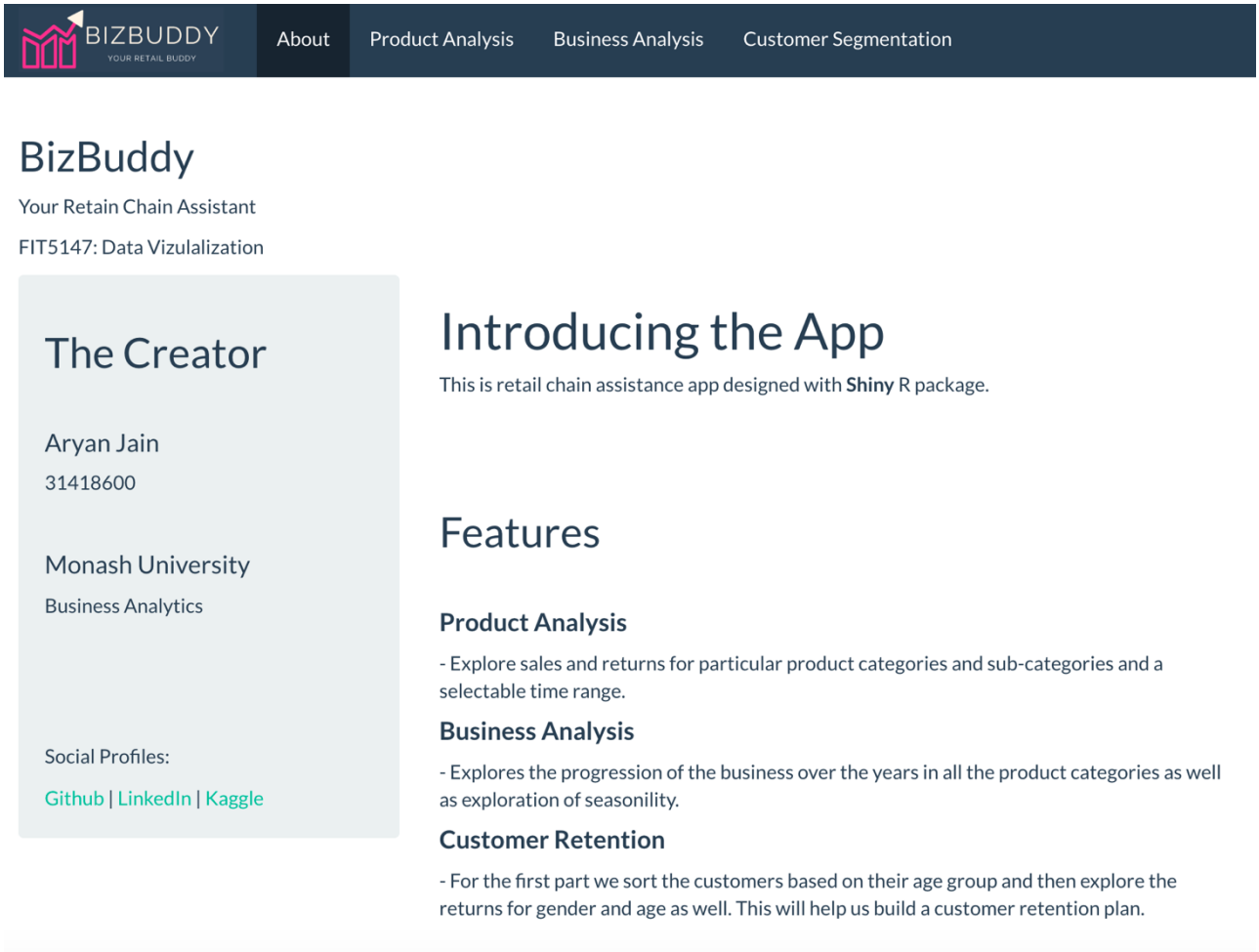
Second Sheet : This sheet expands of those ideas formed in the first sheet. Analysis will include best performing products and also which products are often returned.

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.



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.

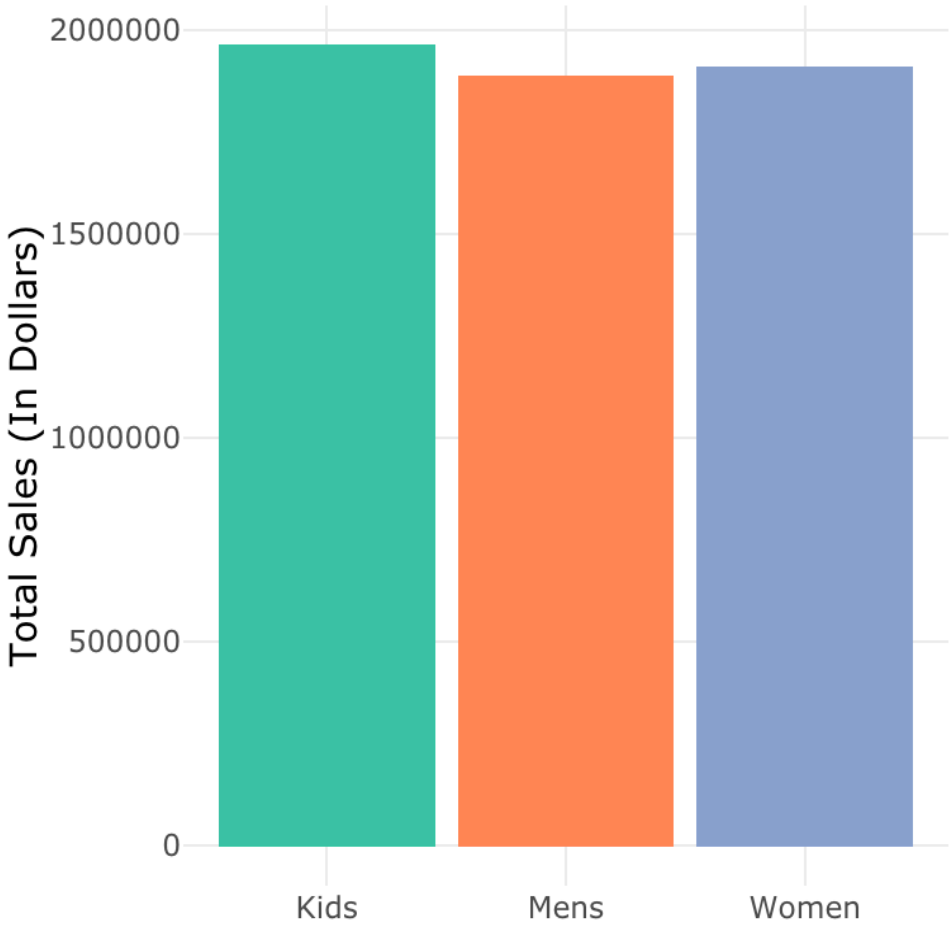
Plotly: This was used to plot interactive bar plots in the first section of the app.

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

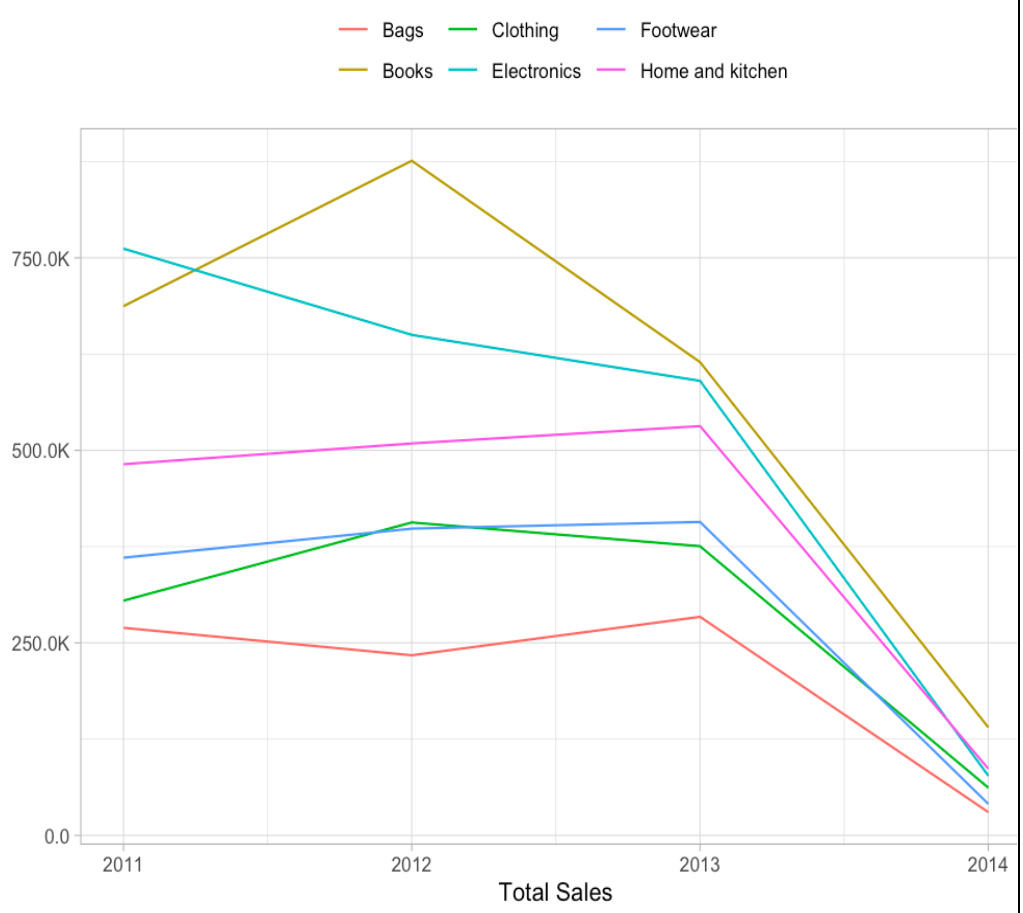
Eeptools: This package is used to calculate the age from DOB.

Scales: This package is used to manipulate the scale in the scale_continuous parameter of the ggplot.

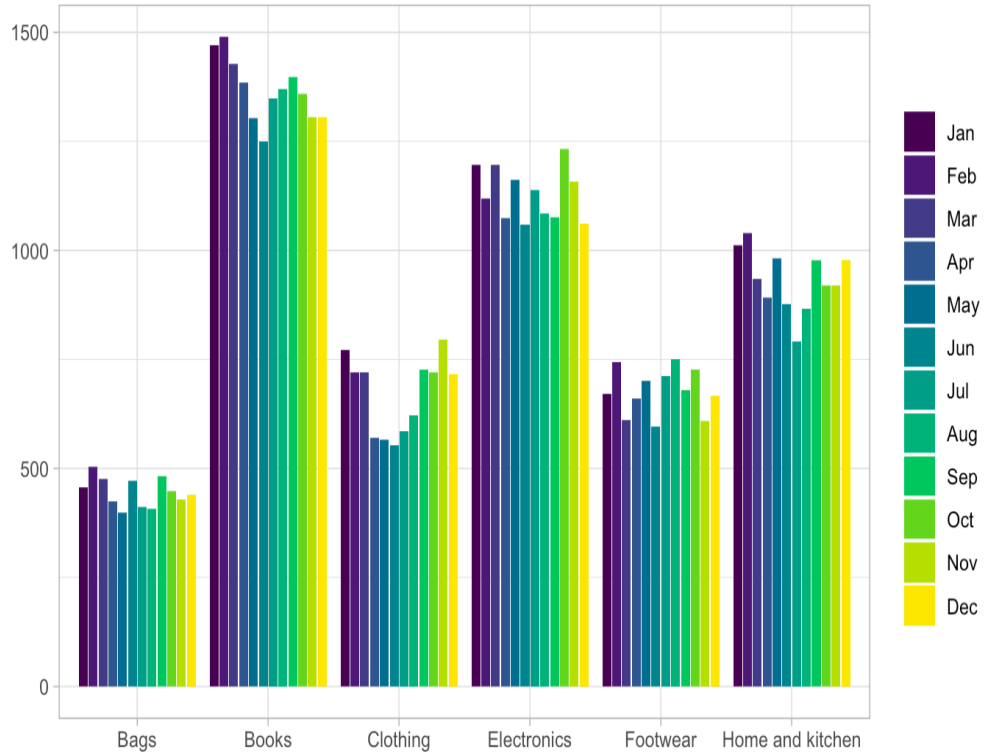
Product Analysis – Plot 1

Visualization graph name	Bar Plot
Purpose	Display total sales based on product category and time period
Visualization Graph	
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	
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

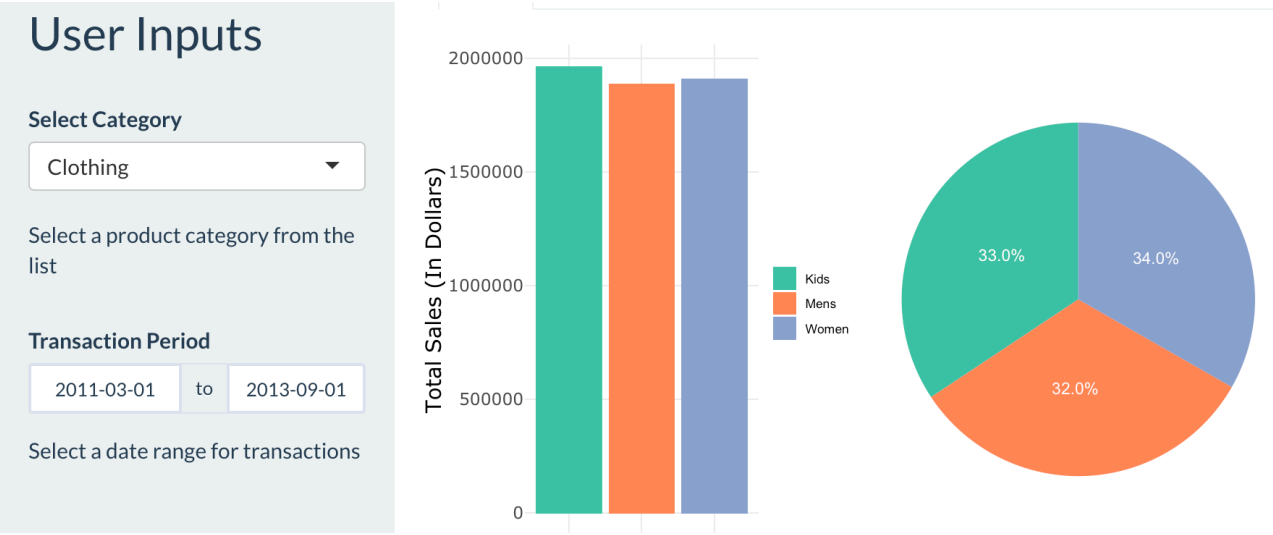
Visualization graph name	Bar Plot
Purpose	Display Seasonality
Visualization Graph	
Implementation	Group the data by product category and month and use the summation of quantity to plot bar plots using ggplot.
Tab Name	Business Analysis
Libraries used	Ggplot, shiny
Filters	Qty > 0

User Guide

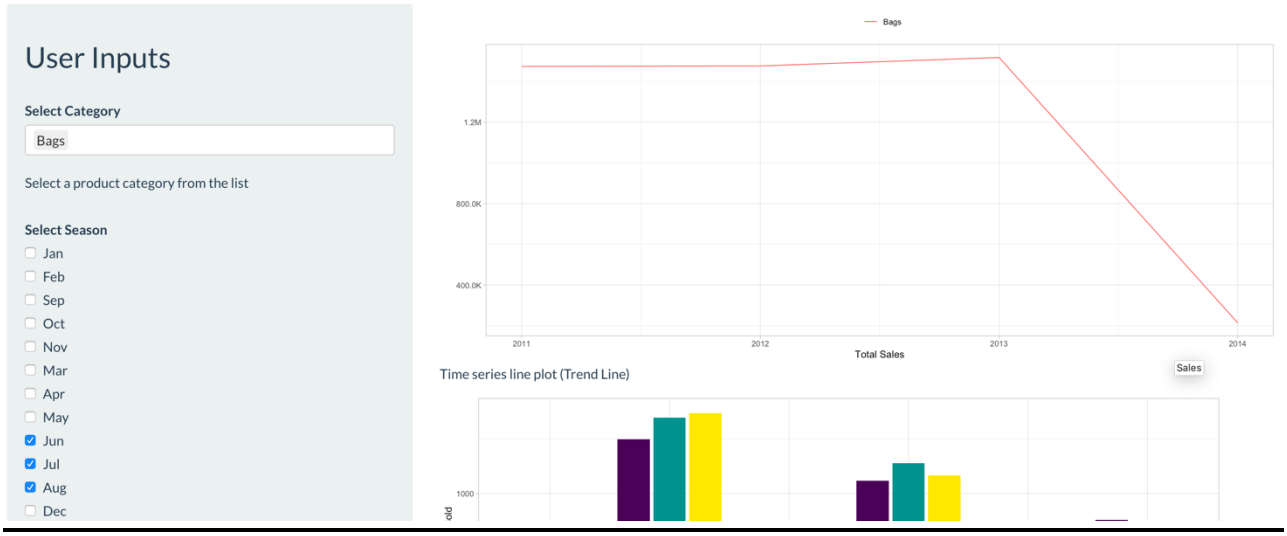
The entire app is based on 4 tabs with the following purpose:

About section (First Tab): This section gives the general information about the app and explains other tabs in a brief. There are no dynamic or UI elements in this tab.

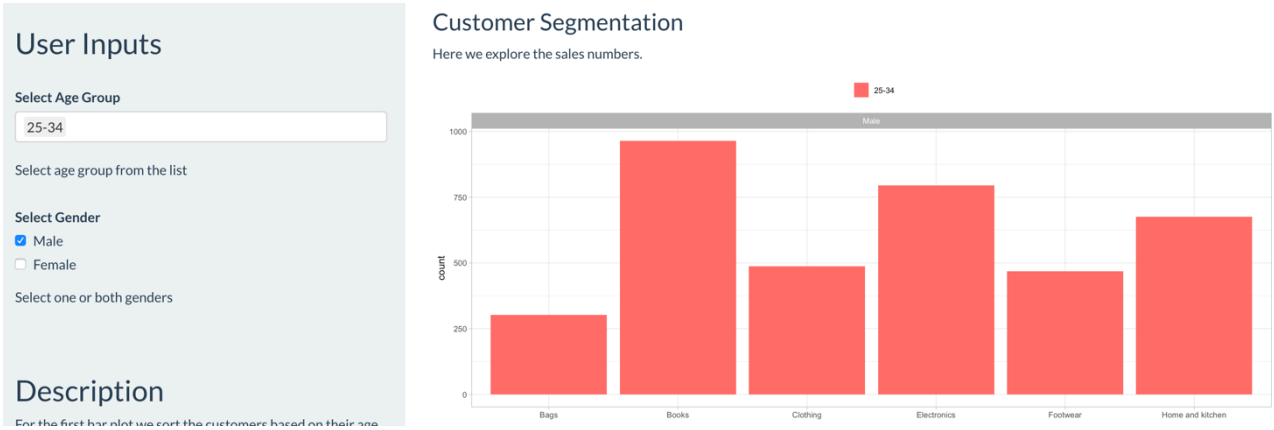
Product Analysis (Second Tab): This tab explores the product category and its performance. The user can interact with input box on the left to select product category and also a time period for exploration.



Business Analysis (Third Tab): This tab explores the business performance with a trend line and seasonal bar plots with user selectable category and months.



Customer Segmentation (Fourth Tab): This tab looks into the customer characteristics such as gender, age and location to identify at risk customers. The user will interact with age-groups filter as well as gender to subset the data.



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.

The product analysis module revealed that certain product sub-category sells better than others. For example, it showed that kids wear sell more than

Men's and Women's wear individually. Also, mobile phones are the biggest sellers in electronics and children and fiction books give the most business. The analyses also revealed that there is a bias in product category and age when it comes to returning product. And categories like books and electronics are more prone to return.

Going in the business analysis tab, we realised that after performing quite well in the first year the business really went downhill, especially during the last year in the dataset. Also, the second plot reveals that seasonality does exist in this retain store and should be kept in mind when introducing discounts and limited time offers.

The customer segmentation module showed that middle-aged group not only shops the most but are also returns products the most. Although, a little stereotypical, the dataset also reveals that women are more likely to shop for bags and footwear than men. 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.

Overall, the purpose of the app was to able to study the sales and returns of a business and identifying patterns in customer purchase decisions. Bizbuddy with its simple three modules can be easily adapted for any other business with minimal data manipulation.

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.

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/.

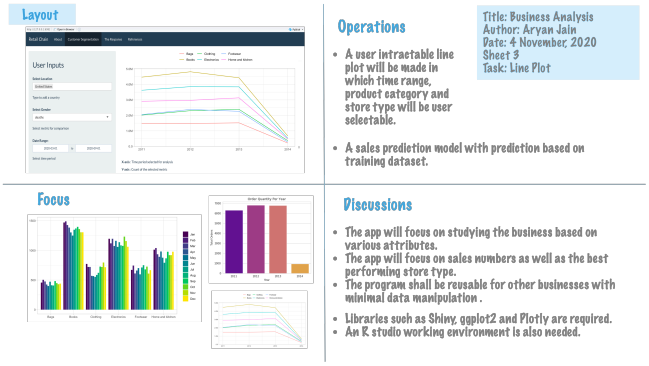
Jared E. Knowles (2020). eeptools: Convenience Functions for Education Data. R package version 1.2.4. https://CRAN.R-project.org/package=eeptools

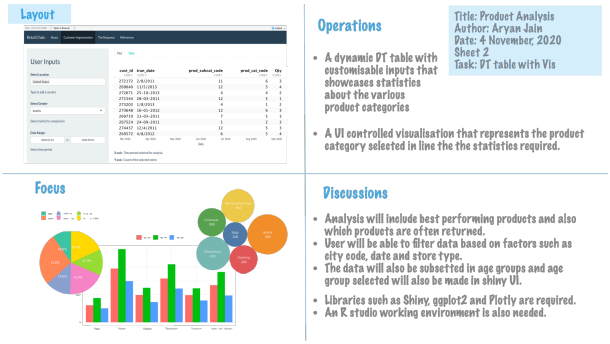
Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

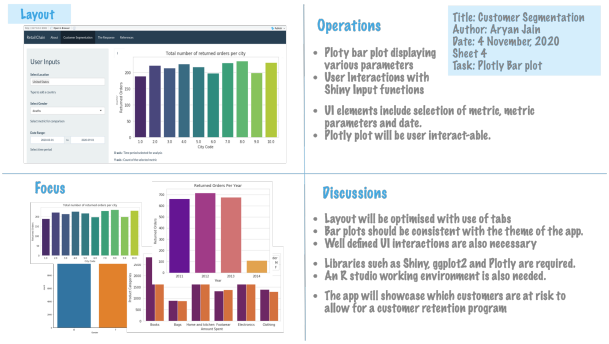
Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

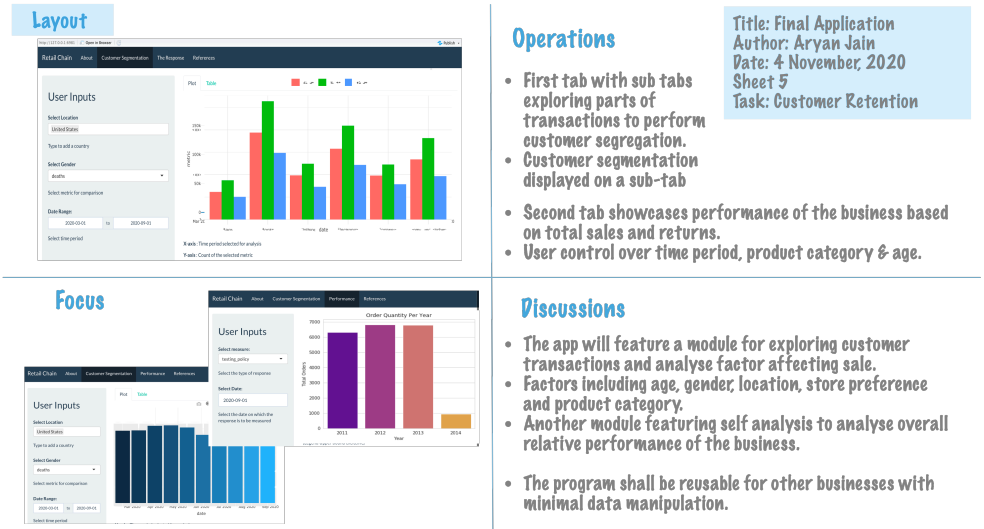
Appendix


Ideas
1. We can look into which store types are performing the best.
2. Which product category sells best?
3. Who are the most loyal and at risk customers?
4. How does gender, age and location affect buying choices?
5. Which category is more prone to returns and
6. How frequent are the returns?
7. How has the retail chain progressed over the years?
8. Is there a seasonality factor involved?
Categorise
1. We can look into which store types are performing the best.
2. How has the retail chain progressed over the years?
3. Is there a seasonality factor involved?
1 How does gender, age and location affect buying choices?
2. Who are the most loyal and at risk customers?
Filter
1. Which product category sells best?
2. Which category is more prone to returns and how frequent are the returns?
Refine & Combine
1. We can look into which store types are performing the best.
2. Which product category sells best?
3. How does gender, age and location affect buying choices?
4. Who are the most loyal and at risk customers?
5. Which category is more prone to returns and how frequent are the returns?
6. How has the retail chain progressed over the years and is there a seasonality factor involved?
Business Analysis
Analyse the performance of the retail chain
Product Analysis
Analyse Category of products that are most profitable and which ones are returned most.
Customer Segmentation
Build a customer segmentation model to identify customers using their age, location and gender and relationship between their purchase.


Layout
Operations
A user intractable line plot will be made in which time range, product category and store type will be user selectable.
A sales prediction model with prediction based on training dataset.
Title: Business Analysis
Author: Aryan Jain
Date: 4 November, 2020
Sheet 3
Task: Line Plot
Focus
Discussions
The app will focus on studying the business based on various attributes.
The app will focus on sales numbers as well as the best performing store type.
The program shall be reusable for other businesses with minimal data manipulation .
Libraries such as Shiny, ggplot2 and Plotly are required.
An R studio working environment is also needed.


Layout
Operations
A dynamic DT table with customisable inputs that showcases statistics about the various product categories
A UI controlled visualisation that represents the product category selected in line the the statistics required.
Title: Product Analysis
Author: Aryan Jain
Date: 4 November, 2020
Sheet 2
Task: DT table with Vis
Focus
Discussions
Analysis will include best performing products and also which products are often returned.
User will be able to filter data based on factors such as city code, date and store type.
The data will also be subsetted in age groups and age group selected will also be made in shiny UI.
Libraries such as Shiny, ggplot2 and Plotly are required.
An R studio working environment is also needed.


Layout
Operations
Ploty bar plot displaying various parameters
User interactions with Shiny input functions
UI elements include selection of metric, metric parameters and date.
Plotly plot will be user interact-able.
Title: Customer Segmentation
Author: Aryan Jain
Date: 4 November, 2020
Sheet 4
Task: Plotly Bar plot
Focus
Discussions
Layout will be optimised with use of tabs
Bar plots should be consistent with the theme of the app.
Well defined UI interactions are also necessary
Libraries such as Shiny, ggplot2 and Plotly are required.
An R studio working environment is also needed.
The app will showcase which customers are at risk to allow for a customer retention program


Layout
Operations
First tab with sub tabs exploring parts of transactions to perform customer segregation.
Customer segmentation displayed on a sub-tab
Second tab showcases performance of the business based on total sales and returns.
User control over time period, product category & age.
Title: Final Application
Author: Aryan Jain
Date: 4 November, 2020
Sheet 5
Task: Customer Retention
Focus
Discussions
The app will feature a module for exploring customer transactions and analyse factor affecting sale.
Factors including age, gender, location, store preference and product category.
Another module featuring self analysis to analyse overall relative performance of the business.
The program shall be reusable for other businesses with minimal data manipulation.