Interim Report

ANLY482 Analytics Practicum

Analytics Cafe

Chen Shiqi

Tan Wei Lin Joanna

Table of Content

1 Practicum Recap

The objective of this project is to provide an interactive dashboard of different visualizations of data collected from HoiPOS systems. The project aims to improve HoiPOS' data visualizations as well as add more visualizations of data analysis to value-add their POS system to clients. This project also aims to be able to apply descriptive analytics and effective visualizations to gain insights not only on the sales performance but also operations and marketing campaigns.

The final product will consists of an interactive visualization dashboard which will be designed to provide visualizations that can be easily understood by the layman as well. Intuitive functions such as cross filtering, will be implemented to allow business owners to conduct exploratory data analysis without needing technical knowledge. To ensure that the dashboard is easy to use, the team will also conduct several user testing with stakeholders.

The main objective of the project would be to develop the following functions/visualizations:

A. Filter data by date/time
    - This allows to choose certain periods of the year to analyse as different seasons of the year could have a higher anticipated sales performance e.g restaurants near F1 location during F1 event
    - Being able to view analysis of data at a more micro level also allows the business owner to understand which time periods are peak periods and hence make better human resource deployment decisions quickly
    - Anomalies in the data can also be noticed quickly and will be able to have a reference point of why the anomaly happened using the date and time. E.g Spike in orders during F1 event
B. Dashboard for visualizations
    - Sales Performance over time:
        i. Includes overview of sales, average amount earned a day, number of orders, target analysis and more
        ii. Compare historic data and real-time data to notice gaps and/or trends
    - Association Analysis of items
        i. To allow users to identify popular combination of items. This information is useful when crafting marketing campaigns e.g combo meals of the popular combinations of items

- Visualization of Popular Items
    i. To allow users to identify most popular items and/or identify items that are least popular. This information informs users about items that are not being ordered frequently could be removed from the menu to reduce wastage and costs
    ii. Exploring further for Popular Items - Visualization of Popular Items Modifier
        1. To allow users to identify the most popular item modifier i.e hot/cold options for a cup of tea. This information may prompt users to modify the menu to accommodate for the popular preferences.
    iii. Users will be able to cross-filter the data using time as a variable to see the popular items at different time periods
- Visualization of Productivity Analysis
    i. Users will be able to analyse time taken to prepare different food items and how it varies across the day. This insight will be able to aid business owners in making operational decisions with regards to human resource and productivity in the kitchen.
- Cross-filtering
    i. Allow users to quickly have a glance of different set variables across a few different visualizations.

2 Revised Scope of Work

2.1 Data Preparation

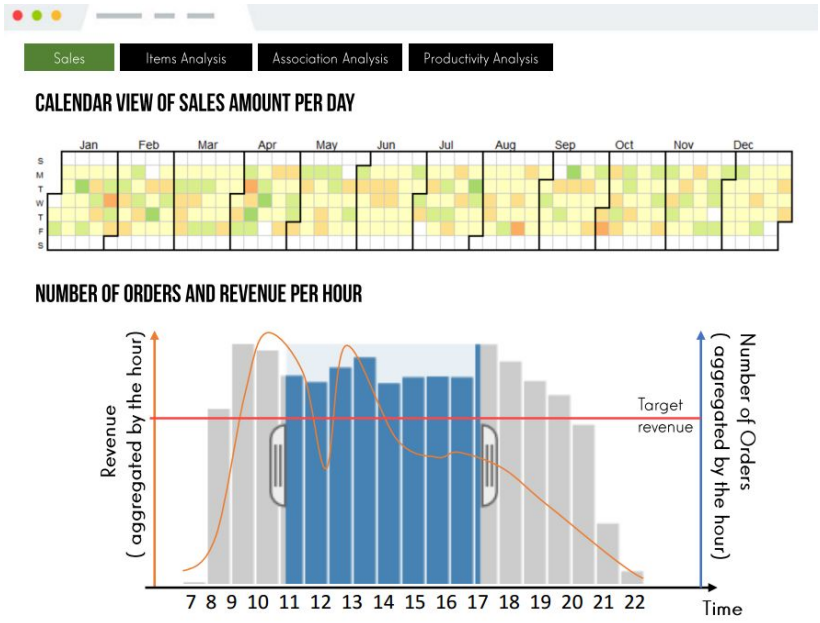
For our data preparation, the team will first create a database diagram to understand the data flow between tables and how variables are related. After which, the team will then remove all unnecessary columns that is not relevant to our project to zoom into the actual data that we need. Next, the team will conduct frequency analysis on all variables to identify outliers. The team then decides whether the outliers were entry errors or actual cases before removing them from the table by either cross-referencing or clarifying with the Sponsor. Following that, tables will be joined in order to be used for data analysis. Lastly, frequency analysis is conducted once again to the joined tables to identify outliers again.

2.2 Data Exploration

For data exploration, the team used Tableau to create preliminary visualization to understand the data better. The primary goal in data exploration process is to retrieve meaningful information regarding the data that is useful in the business context of the project. This includes sales performance, number of orders ordered and popular products analysis. Using Tableau allows us to retrieve information using the filter functions efficiently and effectively. This step is crucial before we jump into developing visualizations before understand the type of results we should be looking out for.

2.3 Storyboard

2.4 Visualization of Sales Performance

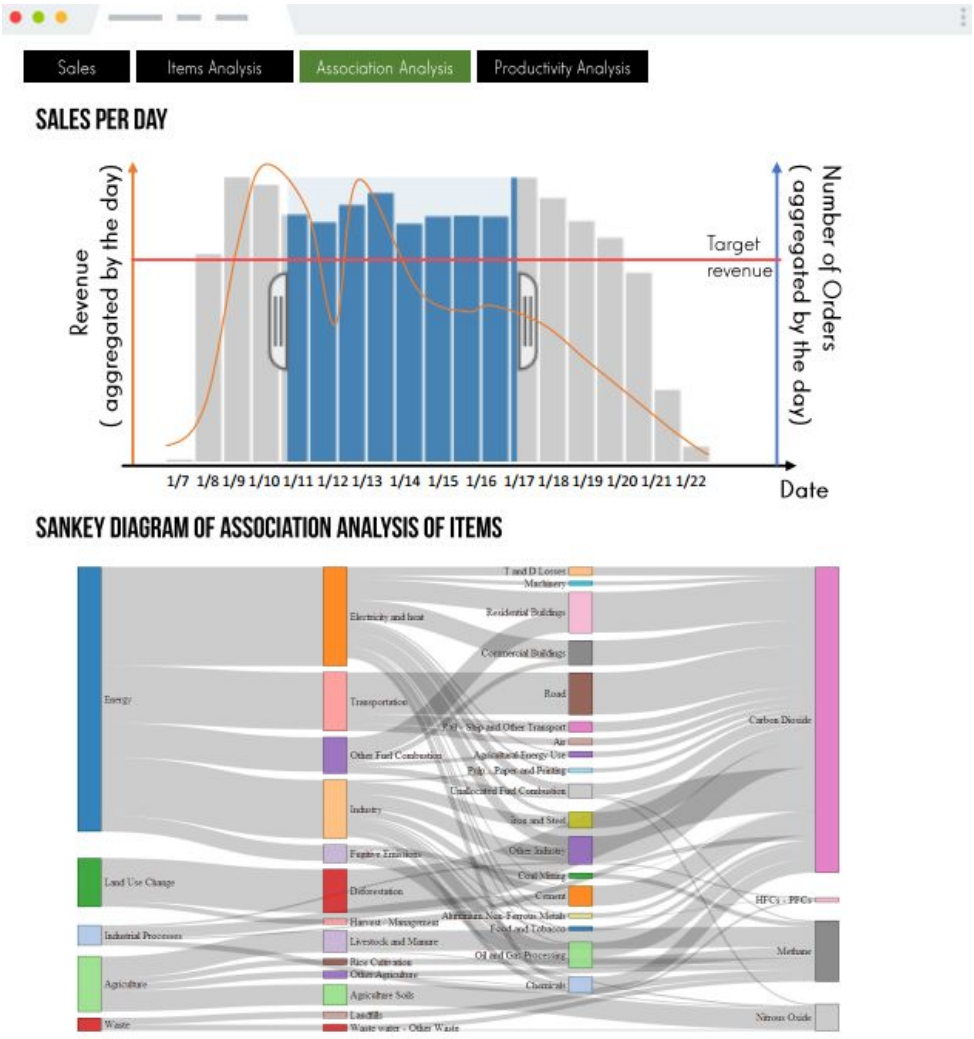


The purpose of this visualization is to allow the user to view sales across time. The user will be able to compare the sales performance of the outlet between the months and years. On the Calendar View, each square represents the total revenue for each day and the color ranges will give an overview of the performance across months. If it is green, it is performing above the target sales for the month, and red if it is performing below the target.

There will also be a reference line on the bar/line chart to indicate target sales each day so that the user is able to quickly measure the performance of the outlet for the day.

With the use of cross-filter function, the user is able to zoom-in into a particular day of the month and see the sales performance of the day itself. This gives flexibility for the user to explore the data and gain more insight.

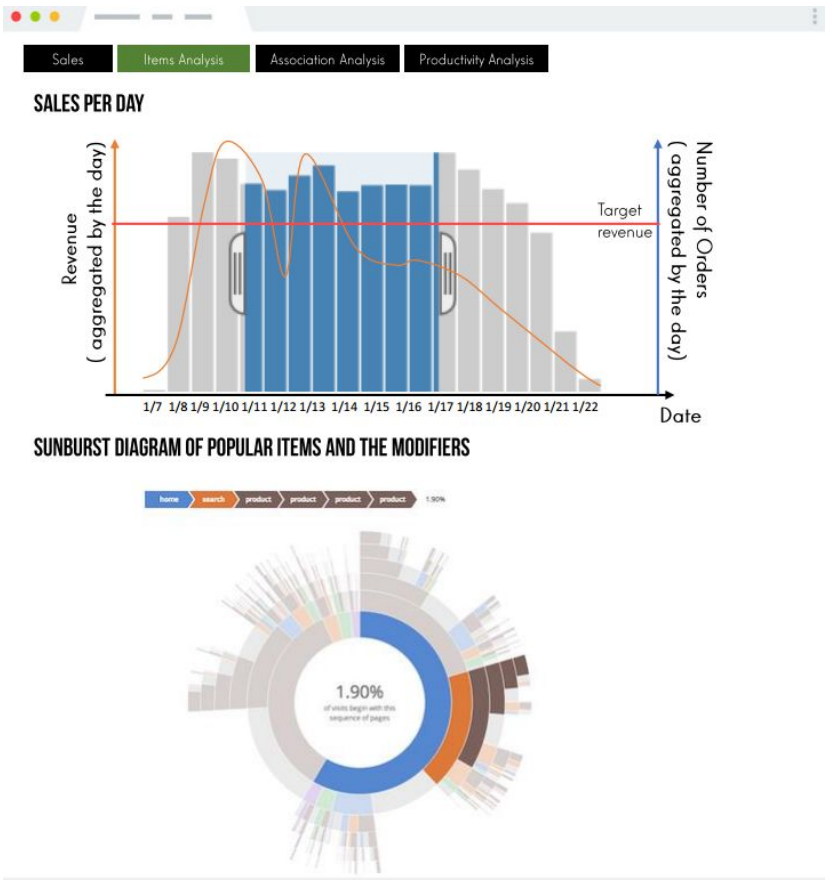
2.5 Association Analysis of Items



The purpose of this graph is to see what items the customer normally purchase together. The bar/line chart above allows users to first have an overview of sales performance as well as number of orders ordered of each day for the selected month.

The Sankey diagram will then show the associated items being ordered according to the range of dates selected using cross-filter on the first graph. The user will then be able to identify the most popular combination of items being ordered for a selected period of time. This information can help the user in deciding promotional bundle or discount to increase the customer return rate or attract more new customers.

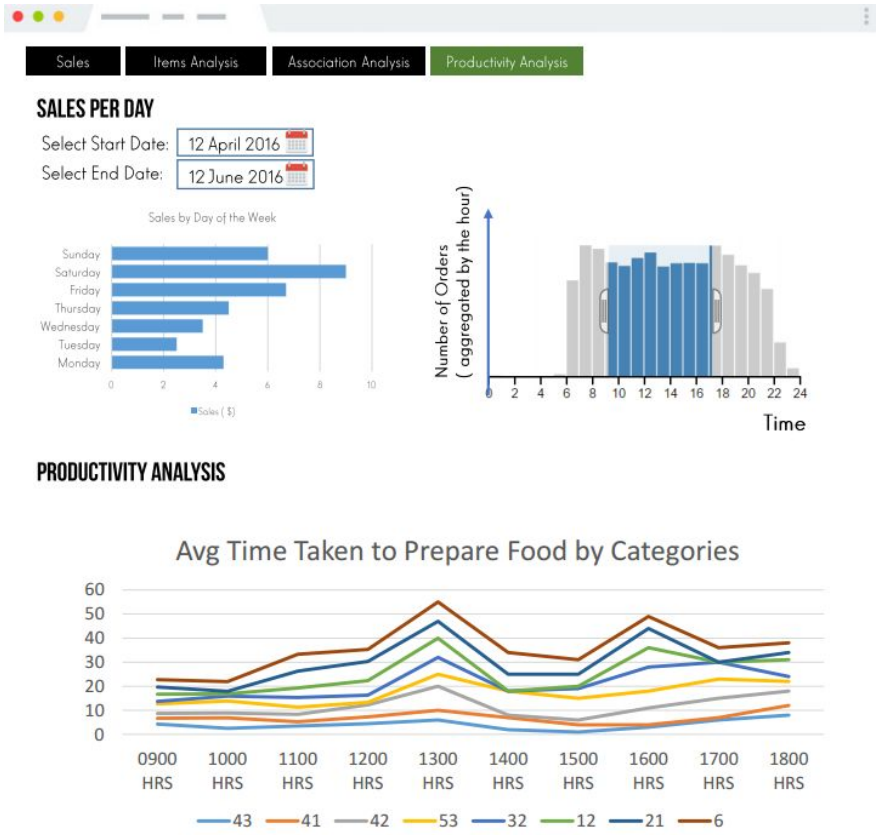
2.6 Visualization of Popular Items and modifier using cross-filter



The purpose of this visualization is to allow the user to view sales of the different product to see which are the most popular item and modifier. The user will be able see what are the popular item across different time. On the bar chart, the bar will show the number of orders and the line will show the revenue aggregated by the day.

With the use of cross-filter function, the user identify could quickly identify the most popular item and modifier across different time. With this insight, the user can make better marketing decision and identify the products that are constantly rank among the lower tier and consider removing the item from the menu to prevent extra cost incurred from making the product and minimize food wastage.

2.7 Visualization of Productivity Analysis



The purpose of this visualization is to allow the user to view the preparation time of the different categories of items over the course of the day for the selected date range. The line graph allows the user to see the time taken to prepare the items in the various category at different time of the day. The data points on the line chart takes the average of all orders that was created at the same time as there might be multiple orders that occurred in the same hour. With the use of cross-filter function, the user is able to zoom-in into a particular day of the week or a range of hours to gain more insight.

Understanding this data will allow business owners to identify when there is a spike in the time taken to prepare the dish at a certain time period, and hence might consider improving productivity through increasing human resource or increase kitchen equipment during that time period.

3 Revised Methodology

In this section, we review the chosen visualizations previously proposed for the dashboard and justify why we chose it for our project.

Visualizations	**Advantages**	**Limitations**
Calendar View	• Easy to compare across different month or day of the week • Take up a relative small display area compare to other graph to show the same amount of data	• Need to hover over to see the value of a particular day
Bar/Line Chart	• Dual axis allow for more information to be shown • Show relationship between the two variables	• User might not know which axis to look at
Bar Chart	• Easy to understand • Shows each data category in a frequency distribution • Displays relative numbers of multiple categories	• Requires additional information as it fails to reveal causes or effects
Time-series Line Chart	• Shows data variables and trends clearly • Interim data can be inferred	• Might be hard to read when there are many lines representing each categories
Sankey Diagram	• Able to see relationships of different data categories clearly	• Might be hard to read when there are too many data categories on the chart
Sunburst	• Able to see the hierarchy (item followed by modifier)	• Could be hard to see item with small value

	• Easy to identify the breakdown by looking at the slice size	• Need to hover to see more information

4 Data

Note: Due to the sensitive nature of the content in our data, we will not be showing screenshots of our data. A separate document containing sensitive data will be submitted through the eLearn portal.

4.1 Metadata

Before we begin any form of data cleaning, we first identify the metadata of our data tables.

4.1.1 Category Table

Variable	**Data format**	**Variable type**
id	String	Nominal
main_category_id	String	Nominal
display_name	String	Nominal

4.1.2 Category_item Table

Variable	**Data format**	**Variable type**
id	String	Nominal
category_id	String	Nominal
item_id	String	Nominal

4.1.3 Item Table

Variable	**Data format**	**Variable type**
id	String	Nominal
plu	String	Nominal

Display_name	String	Nominal
price1	Numeric	Continuous
station1	String	Nominal
station2	String	Nominal
station3	String	Nominal
station12	String	Nominal

4.1.4 Modifier_group_item Table

Variable	**Data format**	**Variable type**
id	String	Nominal
item_id	String	Nominal
modifier_group_id	String	Nominal

4.1.5 Item_modifier_group Table

Variable	**Data format**	**Variable type**
id	String	Nominal
parent_item_id	String	Nominal
modifier_group_id	String	Nominal

4.1.6 Modifer_group Table

Variable	**Data format**	**Variable type**
id	String	Nominal
display_name	String	Nominal
price_level	String	Nominal

4.1.7 Order Table

Variable	Data format	Variable type
id	String	Nominal
created_timestamp	Numeric	Continuous
total	Numeric	Continuous
remarks	String	Nominal
Day of Week	Numeric	Ordinal
Day	Numeric	Ordinal
Month	Numeric	Ordinal
Month Year	Numeric	Ordinal
Year	Numeric	Continuous
Date	Numeric	Continuous
Hour	Numeric	Continuous

4.1.8 Order_item_parent Table

Variable	Data format	Variable type
id	String	Nominal
order_id	String	Nominal
quantity	Numeric	Continuous
remarks	String	Nominal
item_name	String	Nominal
modified_timestamp	Numeric	Continuous
item_price	Numeric	Continuous
item_id	String	Nominal
plu	String	Nominal

4.1.9 Order_item_parent_option Table

Variable	Data format	Variable type
id	String	Nominal
item_parent_id	String	Nominal
option_name	String	Nominal
modified_timestamp	Numeric	Continuous
item_id	String	Nominal
plu	String	Nominal

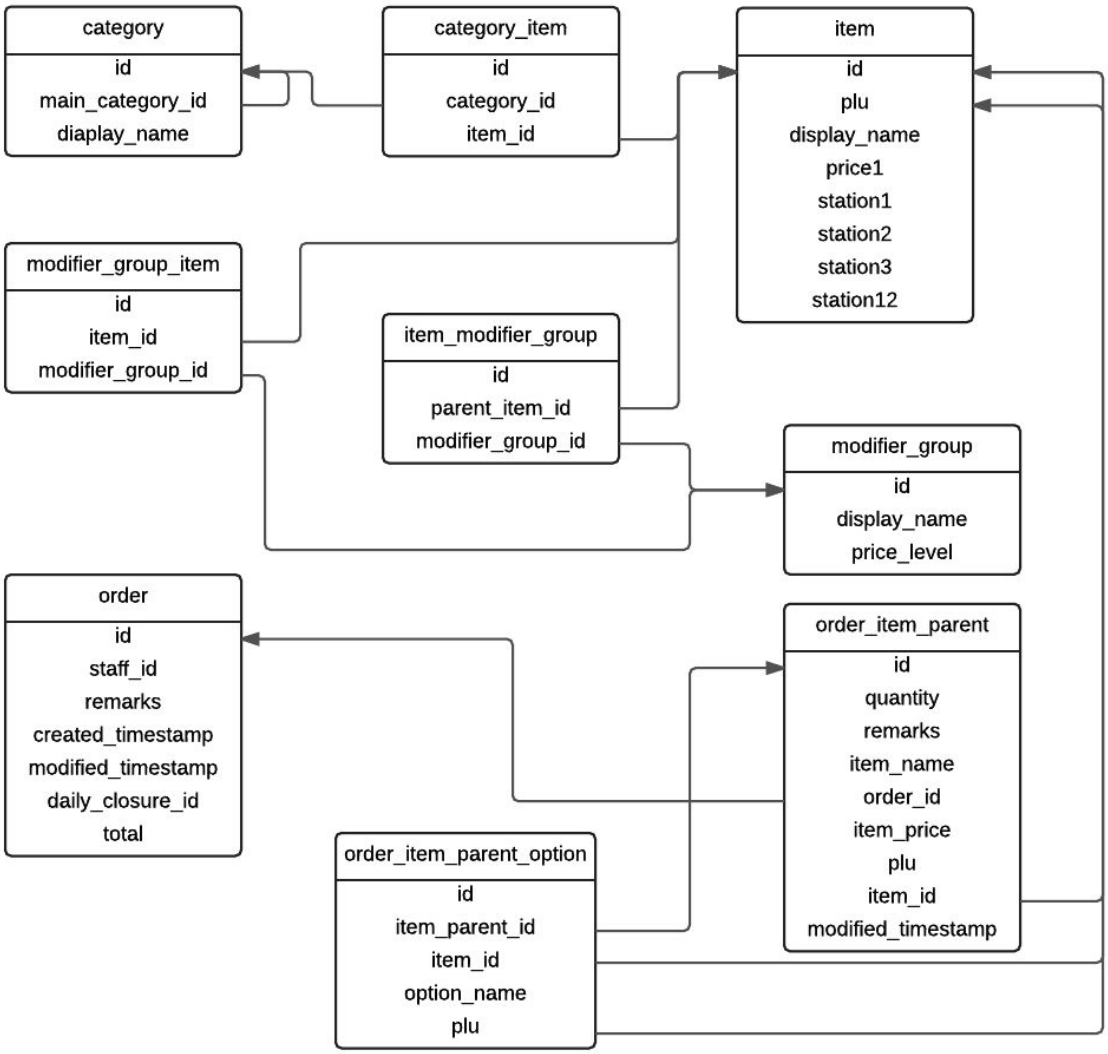
4.1.10 Full List Table

The team joined all tables into a single table to ensure data accuracy and consistency e.g order_id is present in all tables. First, the data is left-outer joined with Order Table being the left table and order_item_parent table being the right table. This resulted joined table will then be used as the left table to be joined with order_item_parent_option table.The same method is applied for the remaining tables to be joined.

Variable	Data format	Variable type
order_id	String	Nominal
item_parent_id	String	Nominal
option_id	String	Nominal
item_parent_name	String	Nominal
option_name	String	Nominal
item_parent_itemID	String	Nominal
option_itemID	String	Nominal
item_parent_plu	String	Nominal
option_plu	String	Nominal
quantity	Numeric	Continuous

item_price	Numeric	Continuous
item_parent_modified_timestamp	Numeric	Continuous
option_modified_timestamp	Numeric	Continuous
created_timestamp	Numeric	Continuous
total	Numeric	Continuous
category_id	String	Nominal
main_category_id	String	Nominal
display_nameofcategory	String	Nominal
Date[created_timestamp]	Numeric	Continuous
Time of Day[created_timestamp]	Numeric	Continuous
Date[item_parent_modified_timestamp]	Numeric	Continuous
Time of Day[item_parent_modified_timestamp]	Numeric	Continuous
Date[option_modified_timestamp]	Numeric	Continuous
Time of Day[option_modified_timestamp]	Numeric	Continuous
food_prep_duration	Numeric	Continuous

4.2 Database Diagram



4.3 Data cleaning and processing

4.3.1 Item Table

Plu are supposed to be unique but duplicate plu was found in the item table. The team clarify this with the sponsor and understand that this happen because the store side added in the plu without checking that the plu already exist in the system. As there are only 8 of such instance, the sponsor said that the team could manually edit the plu to some other value.

4.3.2 Order Table

In the Order table, we discovered only 1 outlier in the created_timestamp column where the data is from the year 2019. Since this is obviously a data entry error, we removed it from our table. There were 2 other outliers from the food_preparation_duration column where the time taken to prepare the food took more than 600 minutes. In the business context of our dataset, that is not possible. Hence, these rows are removed from the table.

In the total column, there were 2 outliers with the values 500. To verify if this is valid, the team cross-referenced the data to order_item_parent table using the order_id variable. The data rows are missing from the latter table and hence, deeming these data rows to be invalid. These rows could be created as testing data.

4.3.3 Order_item_parent Table

In the order_item_parent table, there were 2 outliers from the quantity column with values 15 and 20. Upon further investigation by checking the item_name, the team decided that this is relevant and should be retained in the data due to the nature of the item that is purchased.

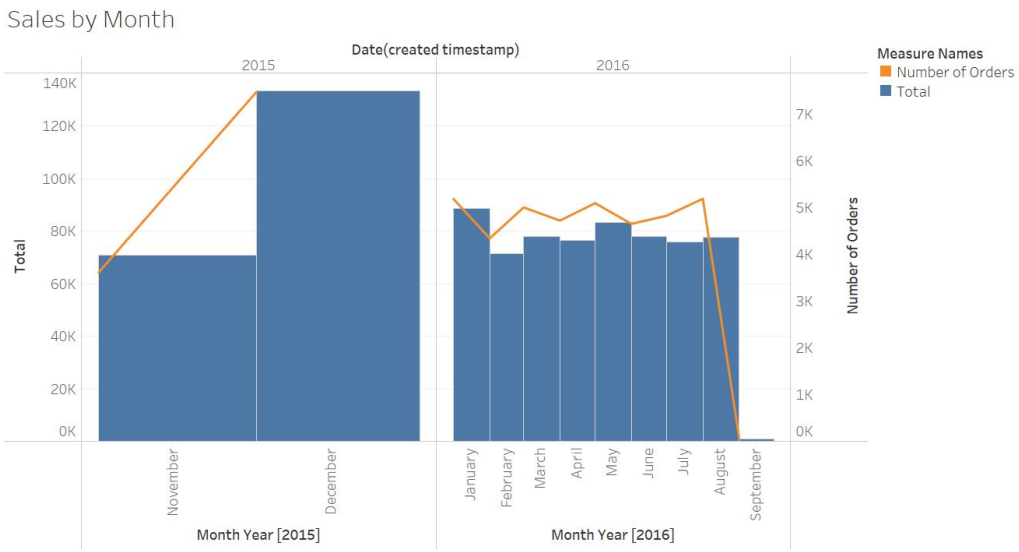
There were 1465 missing PLU codes in this table. After discussing with the sponsor, the team understands that some of the data rows with missing PLU codes are open items that are created on an ad-hoc basis. To verify if all of these data were open items, the team cross-referenced these data rows with their item_id and item_name. Open items should not have item_id as well. However, the team found that only 125 of these selected data are open items. The other data rows with item_ids are cross-referenced and updated with the other present data in the table by matching their item_ids.

4.3.4 Order_item_parent_option Table

In the last table, the team removed 5 rows with null item_id as upon discussion with the sponsor, the team decided that these rows are data entry errors and are anomalies. There were also 213 missing PLU values. Again, to verify whether these data rows are still relevant, the team cross-referenced the other variables. We find that the item_ids are not missing and hence update theses rows' PLU codes by matching the item_ids with existing data rows in the table. After cleaning, there are no missing values for PLU column.

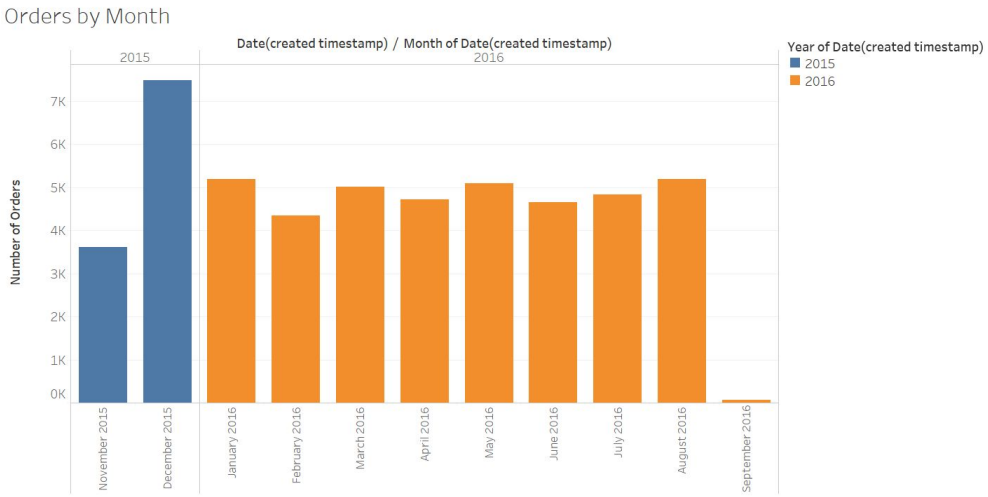
4.4 Data Exploration Results

4.4.1 Sales by Month



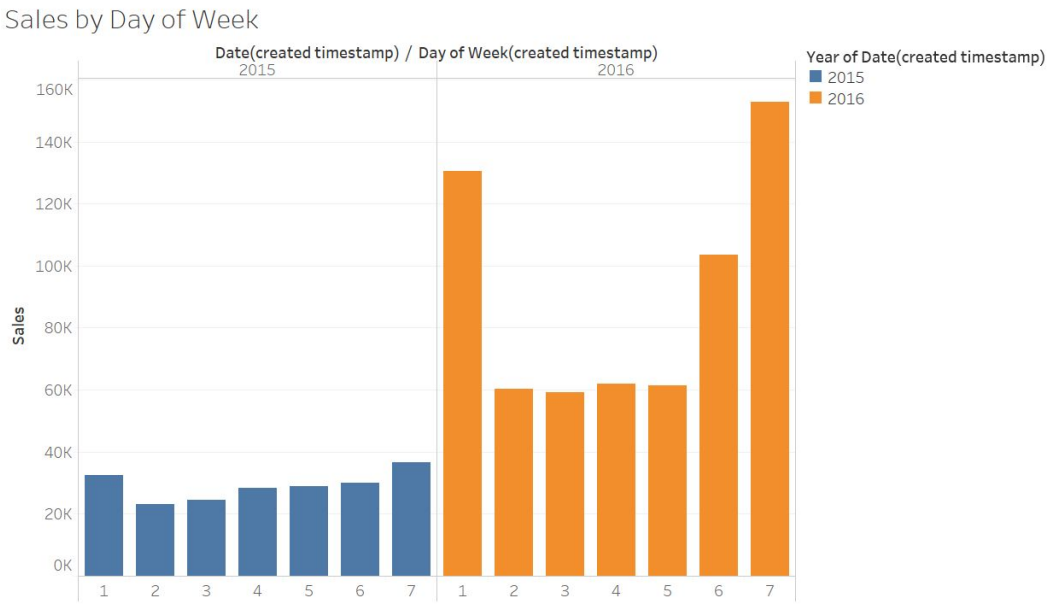
From this visualization, the sales performance on December 2015 performs considerably much better as compared to the rest of the months. Also, in August 2016, there the number of orders seem to peaked but the sales amount did not increase as much as expected. As the team only received the data in early September, there are only a few data entries in that month, therefore explaining the height of the bar for September 2016.

4.4.2 Order Quantity by Month



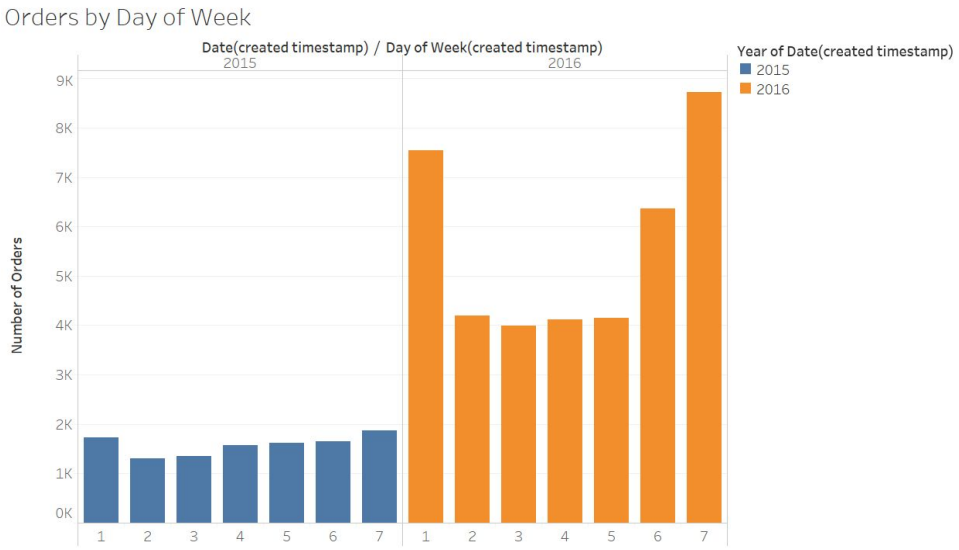
Similarly to the first visualization, we can see that December 2015 is the best performing month in the history of this outlet.

4.4.3 Sales by Day of Week



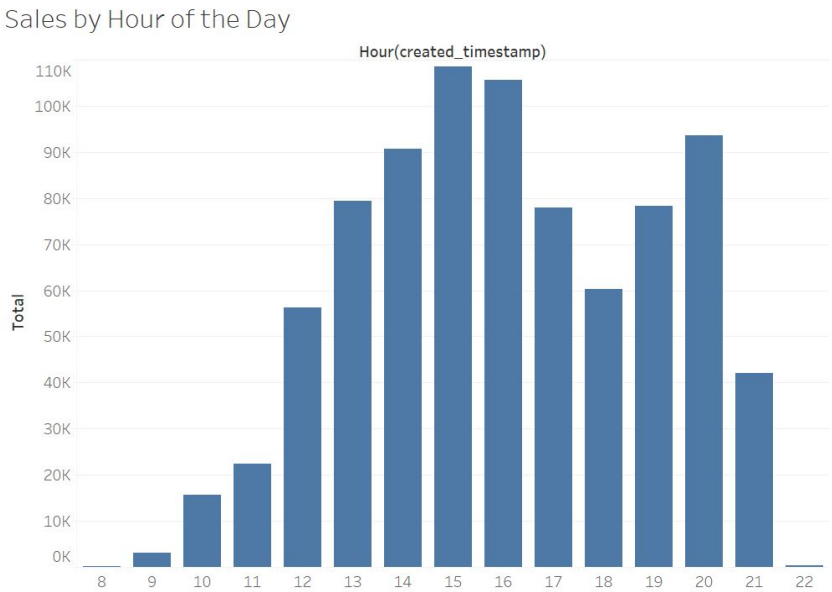
One would expect that the weekends will perform better in terms of sales, however interestingly, in 2016, the sales on Mondays is the second highest after Sunday. The height of the bars in 2015 is much lower as compared to the bars in 2016 is due to the fact that the dataset only contains data from November and December in 2015.

4.4.4 Order Quantity by Day of Week



Similar to the previous visualization, the top 3 best performing days in 2016 are Sunday, Monday and Saturday in order.

4.4.5 Sales by Hour of the Day



From this graph, we see that the highest performing hour of the day is from 3pm to 4pm. This tells the user that the peak period is during those times of the day. This

information can help the manager of the outlet to create special promotions at the hour in order to attract more customers.

4.4.6 Number of times an item is ordered

Item Parent Plu. Color shows sum of Quantity. Size shows sum of Quantity. The marks are labeled by Item Parent Plu. Details are shown for Item Parent Plu. The data is filtered on Date(created timestamp), which ranges from 15/11/2015 to 1/9/2016.

Note: Due to the NDA, item names are excluded from the visualizations.

From this graph, we can clearly see the most popular item which is item 102. This information is useful as the managers of the outlet can utilize this knowledge in creating marketing campaigns such as bundle deals or set meals that include this item.

4.4.7 Category Ranking

Category Id. Color shows count of Order Id. Size shows count of Order Id. The marks are labeled by Category Id. Details are shown for Category Id and Main Category Id. The data is filtered on Date(created timestamp), which ranges from 15/11/2015 to 1/9/2016.

In this graph, the most popular category of items is Category 41. The manager of the outlet can use this information in deciding which items to advertise more or spend more time and effort in marketing it to the customers. Similarly, for the least popular

categories, the manager might consider removing the category from the menu itself to reduce costs.

5 Revised Work Plan

Description of Tasks:

Note: Green rows are completed tasks

Name of Task	**Week**	**Any Changes?**	**Remarks**
Gather requirements from client	2	No	
Tools exploration	2	No	
Draft Proposal	2	No	
Submit Final Proposal	3	No	
Update wiki for Proposal Phase	3	No	
Obtain datasets	3	No	
Data exploration	4-5	Yes	Received dataset late in the previous week

Scope refining after data exploration	4-5	Yes	More uncertainties of data than expected. Met up with client twice to resolve uncertainties.
Research and brainstorm for most suitable visualizations for data and context	5-6	Yes	Due to delay in previous task, there was an overflow in delay.
Design mock-up of prototype	6	Yes	Due to delay in previous task, there was an overflow in delay.
Present mock-up to sponsor	6	Yes	Due to delay in previous task, there was an overflow in delay.
Modify mock-up	6	Yes	Due to delay in previous task, there was an overflow in delay.
Update wiki with finalized mock-up	8	Yes	Due to delay in previous task, there was an overflow in delay.
Update and refine wiki for Midterm	9	No	After attending briefing by Supervisor, team has updated work plan.
Midterm report write-up	9	No	
Begin Coding	10	No	
Review and Revise Prototype with Sponsor and Supervisor	10	No	
Visualizations for association analysis of items ordered	10-11	Yes	Prioritising tasks to suit Sponsor's requirements
Visualizations for productivity analysis	10-11	Yes	Prioritising tasks to suit Sponsor's requirements
Improve application based on sponsor/supervisor feedback	10-11	No	
Visualizations for sales performance (over time)	11-12	Yes	Prioritising tasks to suit Sponsor's requirements
Visualizations for popular items and modifier	11-12	Yes	Prioritising tasks to suit Sponsor's requirements
User Testing with Stakeholders	13	No	
Modify and improve based on feedback	13-14	No	
Final Report Write-up	14-15	No	
Update and refine Wiki for Finals	14-15	No	
Prepare presentation slides for final	14-15	No	

presentation			
Final Presentation	14-15	No	
Final Submission	14-15	No	