**STUDENT ID: 2351044**

**MODULE CODE: CS5803**

**MODULE TITLE: DATA VISUALIZATION**

**ACADEMIC YEAR: 2023/2024**

**TITLE: TABLEAU VISUALIZATION COURSEWORK**

**INTRODUCTION**

Data

The dataset used for this report was retrieved from the following URL, <https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset>. The dataset contains information on customers of a superstore, including their spending habits and response to a previous marketing campaign. The original dataset, “superstore\_data.csv” was modified in RStudio, with 24 rows omitted due to the presence of missing values. This reduced the total number of customers considered from 2240 to 2217. The effects of this reduction are likely to be minimal on overall analysis. Finally, the dataset was renamed “CS5803 Superstore Dataset”. The data was gathered during the campaign from the year prior to the current one. Table 1 below defines the visualization relevant variables within the dataset.

Table 1. Superstore Metadata

|  |  |  |
| --- | --- | --- |
| **Variable Names** | **Variable Type** | **Description** |
| Education | Dimension | Describes the customers level of education, Basic represents those with a basic education level, below any university degree, Graduation represents those who have completed their undergraduate degree, Master represents those who have completed their master’s degree, and PhD represents those who have completed their PhD. 2n cycle for Masters degree as well |
| Kidhome | Dimension | Describes the number of small children in the household. While numerical in nature, the visualization treats the variable as a discrete dimension. 0, 1 and 2 represent the respective number of small children in the household |
| Mnt Meat Products | Measure | The amount spent on meat products in the last 2 years, ranging from zero dollars to $1725 |
| Mnt Wines | Measure | The amount spent on wine products in the last 2 years, ranging from zero dollars to $1493 |
| Num Catalog Purchases | Measure | Number of purchases made using catalog (buying goods to be shipped through the mail), ranging from zero to 28 |
| Response | Dimension | Describes if the customer accepted the offer in the previous campaign. 0 means that they did not and 1 means they did. |
| *Percentage Kidhome (Derived Variable)* | Measure | Describes the percentage of any adjacent sub-category existing within the Kidhome dimension relative to each Kidhome category (after computing using “pane (down)” in shelf). The variable is calculated by dividing the count for each Kidhome category by the total count and multiplying by 100. |

Persona and Questions

The **Marketing Director** for the superstore is preparing for the end-of-year sale. Amongst various marketing schemes including personalized marketing for specific customer demographics, he intends to launch a new gold membership offer, that gives a 20% discount on all purchases, for only $499, $500 less than the usual price. The offer is only available to existing customers and the campaign through phone calls is currently being planned for them. Resulting from insights via other data analytics processes, certain items sold by the store will also be prioritized for augmented advertising. Data visualization executed on the Tableau software will be used to generate valuable insights which will guide the decision making on how to approach the marketing campaign, with considerations focused on how to reduce the cost of and effort spent in executing the campaign plans.

The data visualization in Tableau will attempt to answer the following “simple questions”:

1. Based on the number of children within each household, which type of customer is the least likely to accept the gold membership offer? (Aids in determining which group should require the least focus when marketing)
2. Should customers making more catalog purchases be targeted for meat product advertisements?
3. Amongst the customers, what educational groups should be prioritized when wine is being advertised?

Finally, the visualization will attempt to answer the following “complex question”:

1. Based on the number of children within each household, what type of customer is likely to be interested in all three advertisement efforts in the simple questions, potentially allowing the marketing team to integrate the efforts for this category of customer.

User Requirements Specification

In this section, the choice of representations, relationships and interactions needed to answer the previously defined questions will be considered, accompanied by a brief design idea for implementation.

1. To answer Q1, the user needs to visualize a ranking of the “1” group of the Response variable, within the Kidhome variable, plotted against the *Percentage Kidhome* variable (after flipping axis) to view which type of customer is least likely to accept the membership, this will be represented by the lowest positioned bar. The design will likely require a bar chart with the axis flipped to properly represent the ranking relationship (CS5704 Week 19 Teaching Materials, 2024).
2. To answer Q2, the user needs to visualize a correlation between Mnt Meat Products and Num Catalog Purchases. The correlation relationship can be best visualized using a scatterplot of Mnt Meat Products against Num Catalog Purchases (CS5704 Week 19 Teaching Materials, 2024). A regression line from the bottom left to the top right with points tightly clustered around this line would suggest a strong positive relationship (CS5704 Week 19 Teaching Materials, 2024).
3. To answer Q2, the user needs to make a nominal comparison between the different Education based on the average Mnt Wines. Nominal comparison will be displayed using a bar chart displaying average Mnt Wines against Education (CS5704 Week 19 Teaching Materials, 2024). The comparison will be made based on the relative heights of each bar in the chart.
4. To answer Q4, we need to visualize the R1 bar chart. Two filters linked to the R2 and R3 charts can be added and used to adjust the range for Mnt Meat Products and Mnt Wines. The user will then be able to see which Kidhome category is remaining by applying dynamic filtering through both sliders until only a single category is displayed on the R1 bar chart.

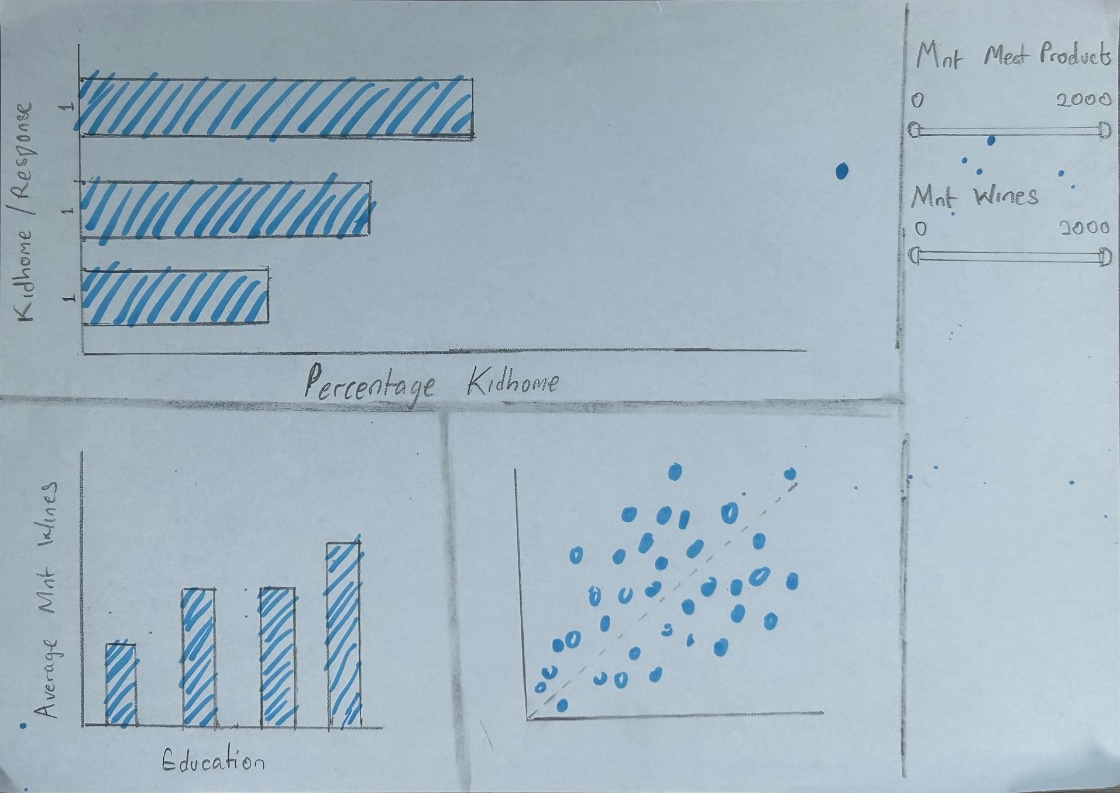
**DESIGN**

Below, the Paper Landscape prototype design for the pre-implementation and the Tableau implementation has been attached.

Figure 1, Paper Landscape Pre-implementation

A bar chart was used to answer Q1 as it encodes the nominal Kidhome/Response with position and quantitative Percentage Kidhome with length in accordance with Mackinlay's guidance.

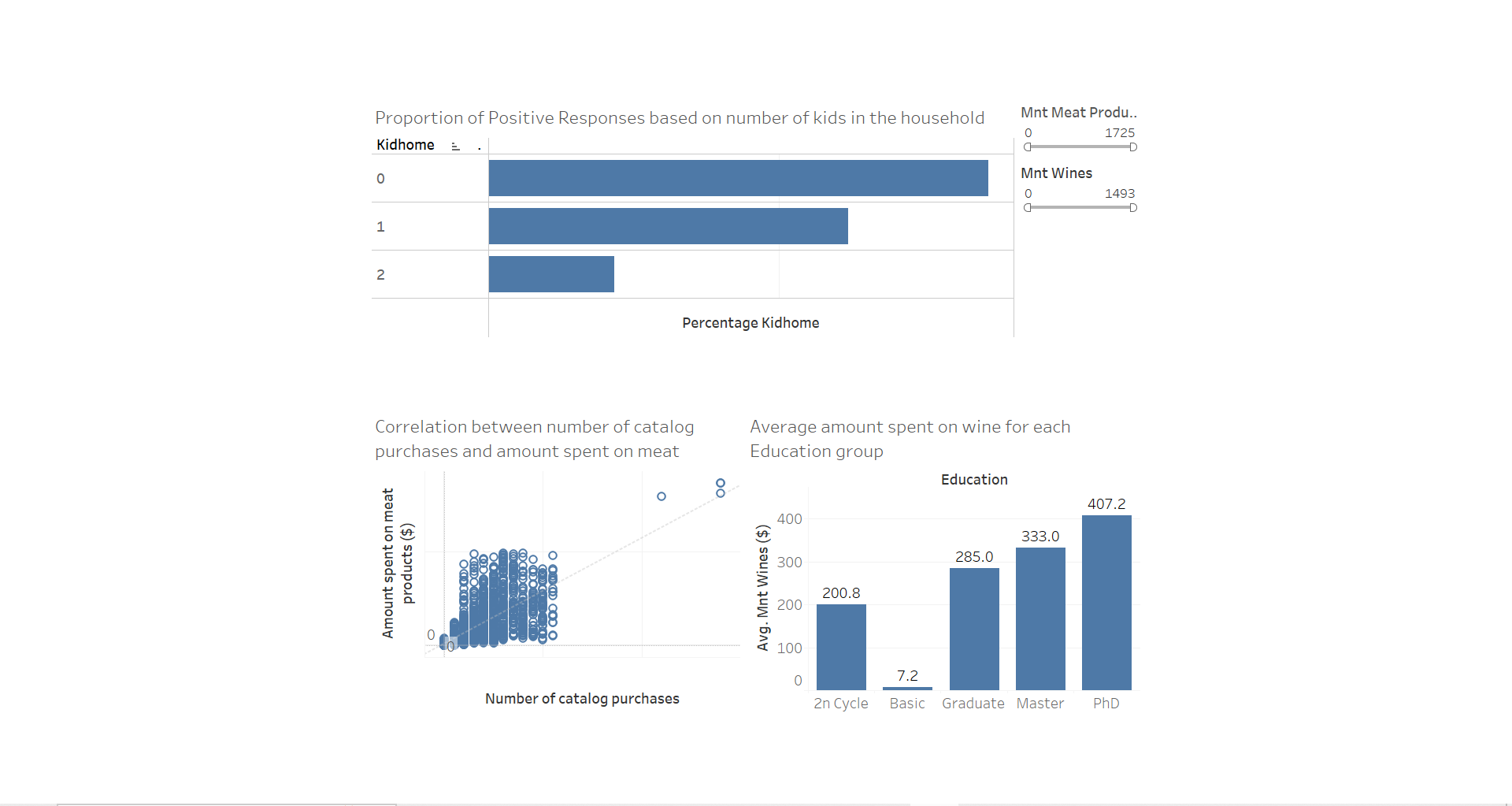
Dynamic filtering to determine which Kidhome group will be interested in all adcertising efforts.

****

A scatterplot was used to answer Q1 as it encodes the quantiative Num Catalog Purchases and Mnt Meat Products with position in accordance with Mackinlay's guidance.

A bar chart was used to answer Q3, with similar reasons to Q1

Minimized emphasis on the axis to amplify cognition.

****

2n Cycle group I was not aware of initially and labelled Average amount spent on wines for clarity, and comparison to a standard.

Displayed axis label as I forgot to do so during pre-implementation.

**IMPLEMENTATION**

In this section, the implementation of the Tableau visualization design solution will be discussed, including the problems faced in implementation and attempted solutions to said problems.

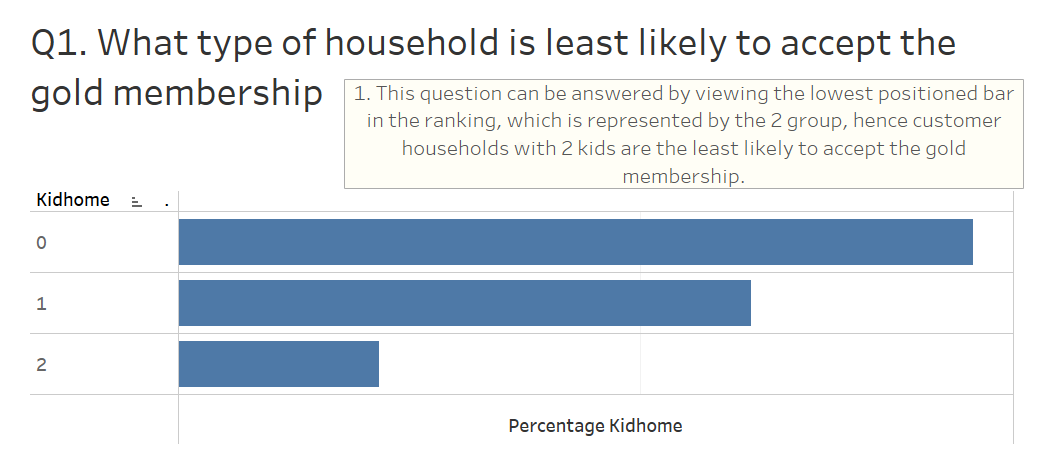
First, the CS5803 Superstore Dataset.csv file was connected to. “Sheet 1” was then selected where I could view all dimensions and measures on the Data pane.

1. Kidhome and Response, which were intended to function as dimensions had been interpreted by Tableau as measures due to the numerical values within them, hence the pill drop down was selected and “convert to dimension”. After this, both pills were added to the rows shelf, Kidhome first, then Response to the right. This was done to capture the customers who responded positively to the gold membership offer. Next the *Percentage Kidhome* measure needed to be created and the following formula was used, (COUNT([Kidhome]) / TOTAL(COUNT([Kidhome])))\*100. The formula helps calculate the percentage of each Kidhome count, relative to total count. The *Percentage Kidhome* pill is then placed in the column shelf. The visualization at this point displayed the percentage count relative to the entire dataset, but what was needed was each percentage of Response instances within Kidhome to apply only relative to each Kidhome category, not the entire dataset. This is the first major issue that was encountered and eventually the solution was found after selecting the *Percentage Kidhome* pill dropdown on the shelf and noticing the “Compute Using” option. After selecting “Pane (down)”, the percentages for each bar in the chart adjusted as needed. Next the 0 Response for all customers who did not accept was to be excluded as only the 1 Response bar was needed for the visualization. This caused a minor issue as initially, the “Exclude” option was selected which recalculated each percentage for the 1 Responses as 100 percent, but the “Hide” option was sufficient for the visualization to display as intended. Tick marks for the x-axis were removed as only the length of the bars were needed for the insight from the visualization to be acquired, this also falls in line with the need to amplify cognition and manage ink ratio(CS5704 Week 20 Teaching Materials, 2024) . Also, problems were faced in limiting the ink on the y-axis, hence the axis was dragged to the left as much as possible to hide the Response label as a solution and the fonts of the y-axis labels were also made bold to support the user (CS5704 Week 20 Teaching Materials, 2024). Cognition was further amplified by hiding the “1” labels for the Response measure on the y-axis of the chart (CS5704 Week 20 Teaching Materials, 2024). This was accomplished by formatting the font as white to stop it from appearing in view. Issues arose when renaming the Kidhome label on the x-axis to “Number of kids in household” and this was not resolved. The chart was then enlarged and the title was changed from “Sheet 1” to “Proportion of Positive Responses based on number of kids in the household”. Finally, the worksheet was renamed to Q1.
2. A new worksheet was opened. Mnt Meat Products and Num Catalog Products pills were placed on the rows and columns shelf respectively. A scatterplot with one point was automatically generated. “Aggregate measures” option in the Analysis menu was unticked and all datapoints were displayed as expected. A trend line was then added to aid the user in the visualization process. The line was made grey and dotted in order to not distract from the primary information (CS5704 Week 19 Teaching Materials, 2024). Both axis were given a lighter grey colour as well to avoid distracting the user, and tick marks were removed from both axis as only the trend line is really needed to assess the correlation (CS5704 Week 20 Teaching Materials, 2024). Num Catalog Purchases and Mnt Meat Products were renamed to Number of catalog purchases and Amount spent on meat products respectively for viewer clarity and professionalism. Mnt Meat Products pill was used to create a filter and the filter was made visible. Finally, the chart was enlarged, and the title was renamed to “Correlation between number of catalog purchases and amount spent on meat products” and the worksheet was renamed to Q2.
3. Education and Mnt Wines pills were added to the columns and rows shelf respectively. The dropdown for the latter on the shelf was then used to plot as average rather than sum. Mnt Wines pill was used to create a filter, which was consistent in appearance to the one utilized in Q2 in accordance with interaction design principles (CS5704 Week 21 Teaching Materials, 2024), and the “All Values” option was selected. The chart was then enlarged for clarity. The title of the chart was renamed to “Average amount spent on wine for each Education group” and the worksheet was renamed as Q3.
4. A new dashboard was created, and all charts were added with similar positioning to that used in the landscape design. The Q1 chart was reduced in size in order to fit into view without scrolling as all the important information needed to be within viewer eyeline (CS5704 Week 22 Teaching Materials, 2024). The title fonts were all reduced to 11 and the colour was changed to a slightly lighter grey in order to not distract from the visualization, while still maintaining legibility based on the dashboard design tips in CS5704 Week 23 Teaching Materials (2024). The “Graduation” group in the “Education” dimension was renamed “Graduate” in order to ensure the entire text was visible in the dashboard. The filters in Q2 and Q3 were then set to apply to Q1 in order for the effects to be reflected in the Dashboard.

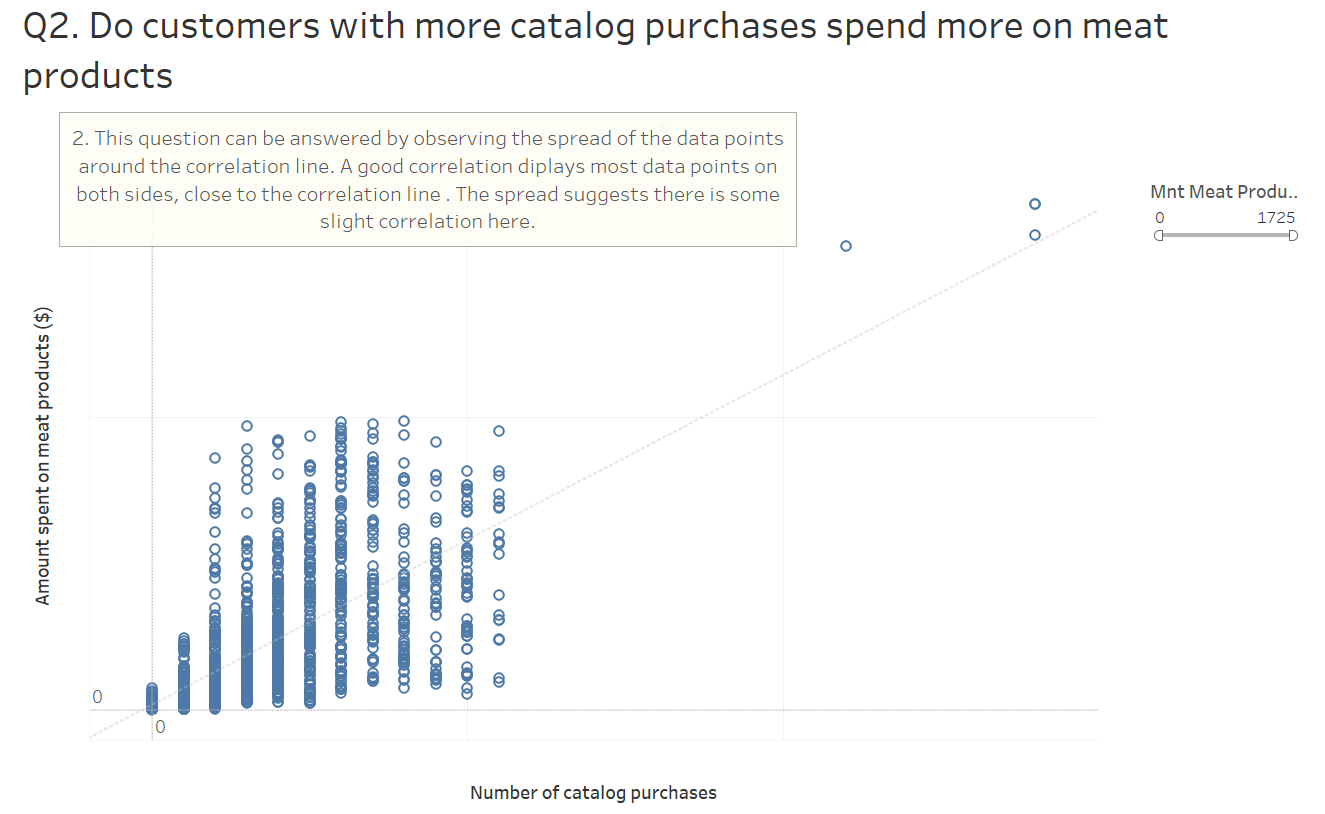
**WALKTHROUGH**

In this section, the insights from each visualization were discussed.

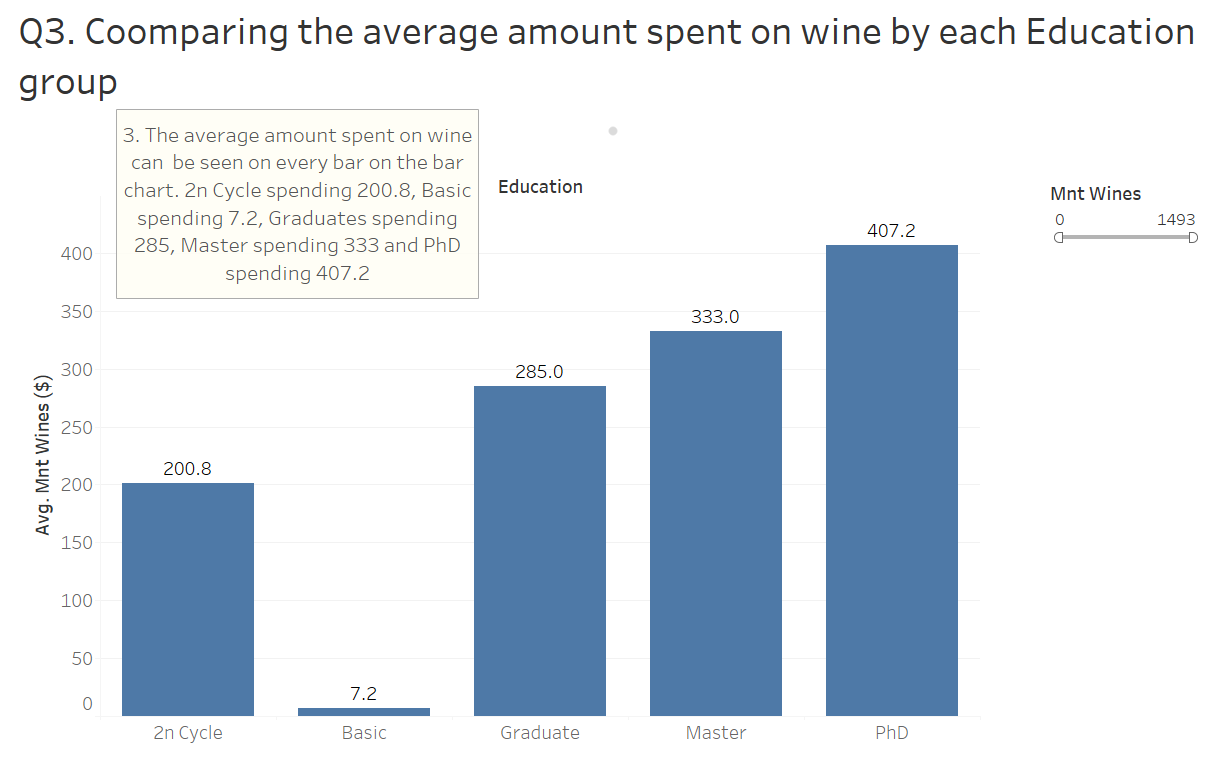
The horizontal bar chart ranked the Kidhome customer groups by the likelihood they will accept the gold membership offer. Customers with two kids were seen to be the least likely to accept the offer. This was clear to see as their bar was the lowest position in the ranking.



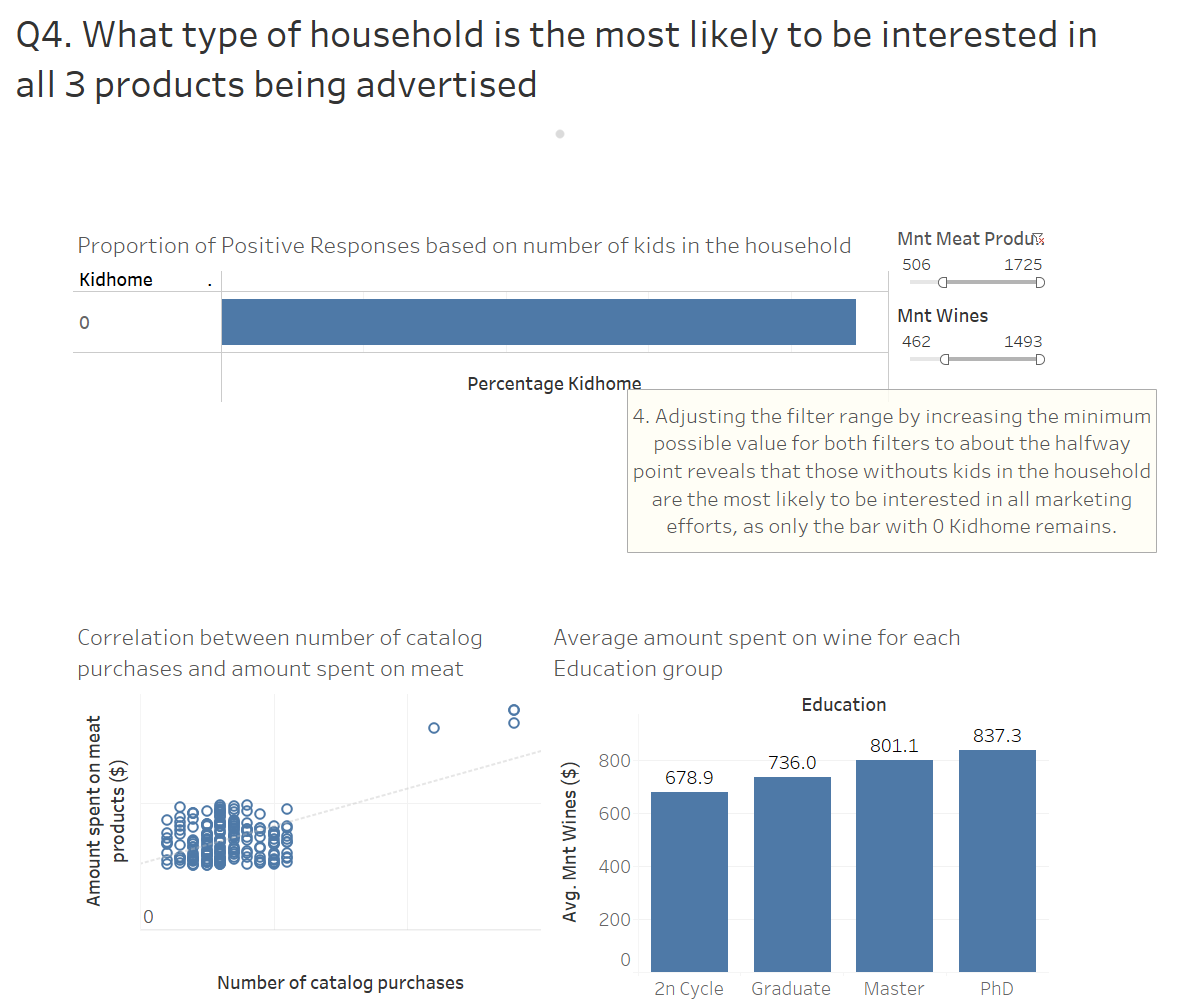
The scatterplot displayed the correlation between the amount spent on meat products by each customer against the number of catalog purchases made by each customer. The graph displayed a slight correlation, and this could be seen as most data points were situated in close proximity to the correlation line running through them.



The bar chart compared the average amount spent on wine by customers, based on differing levels of education. The chart showed that the PhD students spent the most on wine while Basic education students spent the least. The differences in spending behavior could be seen based on the height of each bar or the mark label above them.



The dashboard was designed to filter for the Kidhome category which was most likely to be interested in all three (gold membership, meat products and wine products) advertising efforts. This was executed via dynamic filtering, using two filter widgets for the Mnt Wines and the Mnt Meat Products measures. The filters were used to reduce the range of amount spent on each product by increasing the minimum amount. The visualization process revealed that the 0 Kidhome group, which was the only bar remaining, were the most likely to be interested in all products.



**DISCUSSION**

The visualization prioritized functionality over aesthetics, because the user did not require a presentation that went beyond functional purposes. The ability to hide certain aspects of a visualization which were not needed to answer the questions lends itself to the functionality of tableau as an effective visualization tool which optimizes on human cognition. However, the process of representing the percentage of each Response dimension group within the Kidhome dimension group required the creation of a new variable, which was one weakness I observed as such a function is quite common in visualization hence there should be a simpler method to implement it, or if there was, it should have been more apparent given the commonality with which it is used. In the process of answering the questions within the coursework, I came across multiple alternative, at times ineffective yet somewhat insightful means of visualizing data and would be considering applying tableau visualizations within my dissertation, rather than the more traditional excel applications. I will also be looking to explore other visualization software such as power bi in order to determine which software is simpler to use and offers more functionality and design options.

**CONCLUSION**

The four questions answered within this coursework included applying a horizontal bar chart to determine the lowest ranked Kidhome category based on percentage of positive responses within each, revealing that those with 2 kids are the least likely to be interested in the gold membership marketing. Second the correlation between Num Meat Products and Num Catalog Purchases was established with a scatterplot, then the average Mnt Wines spent by each Education dimension was determined via nominal comparison using a bar chart to figure out who might be interested in wine and finally, dynamic filtering using Mnt Wines and Mnt Meat Products revealed that those without kids are likely to be interested in all advertising efforts.

**REFERENCES**

CS5703 Week 19 Teaching Materials (2024). Lecture 3: Representation. Available at https://brightspace.brunel.ac.uk, accessed 15 April 2024.

CS5703 Week 20 Teaching Materials (2024). Lecture 4: Presentation. Available at https://brightspace.brunel.ac.uk, accessed 15 April 2024.

CS5703 Week 21 Teaching Materials (2024). Lecture 4: Interaction. Available at https://brightspace.brunel.ac.uk, accessed 15 April 2024.

CS5703 Week 22 Teaching Materials (2024). Lecture 4: Dashboards and Infographics. Available at https://brightspace.brunel.ac.uk, accessed 15 April 2024.

CS5703 Week 23 Teaching Materials (2024). Lecture 4: Analysis and Design. Available at https://brightspace.brunel.ac.uk, accessed 15 April 2024.