Direct Marketing Group: Final Report

Li Huang, Witawat Daungjaiboon, Rungtham Rodrawangpai, Yingjie Cao, Jennifer Boyce

Introduction

Grocery retailers employ a wide variety of marketing techniques, such as in-store displays, coupons, and weekly sales mailers. To succeed, it's essential that store management understand their customers, their needs and behaviors, and the role that marketing techniques play in supporting sales. Our group sought to better understand these relationships through a variety of visualization techniques.

Data science company Dunnhumby released a dataset called "The Complete Journey" that provided the data for our exploration. "The Complete Journey" provides a comprehensive look at grocery sales and marketing over a 2-year period. The data includes item-level detail for more than 276,484 transactions, as well as data on marketing efforts during the time period, detailed specifications on the 92,260 products available for sale, and demographic data for 2,500 customers. A complete data schema is included in Appendix B.

Specifically, our group asked four questions:

- What sales patterns are evident over time?
- How are coupons offers utilized in marketing products?
- Which products do customers tend to purchase together?
- How product's display location in a store could affect sales?

Data Characteristics and Exploratory Visualizations

To answer our questions, it was necessary to first gain a better understanding of sales, store products, customers, and marketing efforts and .

Products

Conceptually, stores organized individual products into 44 departments (e.g., grocery, produce, deli). These departments were organized into a total of 308 commodity categories (e.g., cheese, bread) and 2,383 sub-commodity categories (e.g., string cheese, bread:Italian/French).

By far, the grocery department was the largest, offering more than 700 individual products. (Appendix C.1)

Sales

Reflecting the large number of products, the grocery department also had the highest total revenue. (Appendix C.2) A breakdown of the top five best-selling items for the top three departments can be found in Appendix C.3. The data seem to begin with the opening of the store, as sales climb rapidly before reaching a consistent revenue of approximately $750,000 - $1,400,000 per week. (Appendix C.4)

While sales in aggregate, demonstrated trends, the sales of individual products also displayed variability, as illustrated in Appendix C.5 which highlights sales trends among seasonal products. Product sales also varied among stores. Appendix C.6 provides an example of store sales rates for soft drinks and isotonic drinks.

Sales revenue and buying habits varied by age, with customers 25-34 years old generating the greatest revenue. (Appendix C.7, C.8) The treemap in Appendix C.9 shows the departments most commonly purchased by various age ranges.   While customers aged 25-34 contributed the greatest total sales revenue, customers aged 45-54 had the highest average transaction amounts. (Appendix C.10) Individual item purchase volume also varied by age, as illustrated in Appendix C.11.

Coupons

Sales promotions relied heavily on coupon offers. During the two years of data, there were 30 individual coupon marketing campaigns, offering 124,811 coupons for 44,133 products. The grocery department offered the greatest number of coupons (Appendix C.12). A more detailed look at coupon offerings is presented in the “Selected Visualizations” section below.

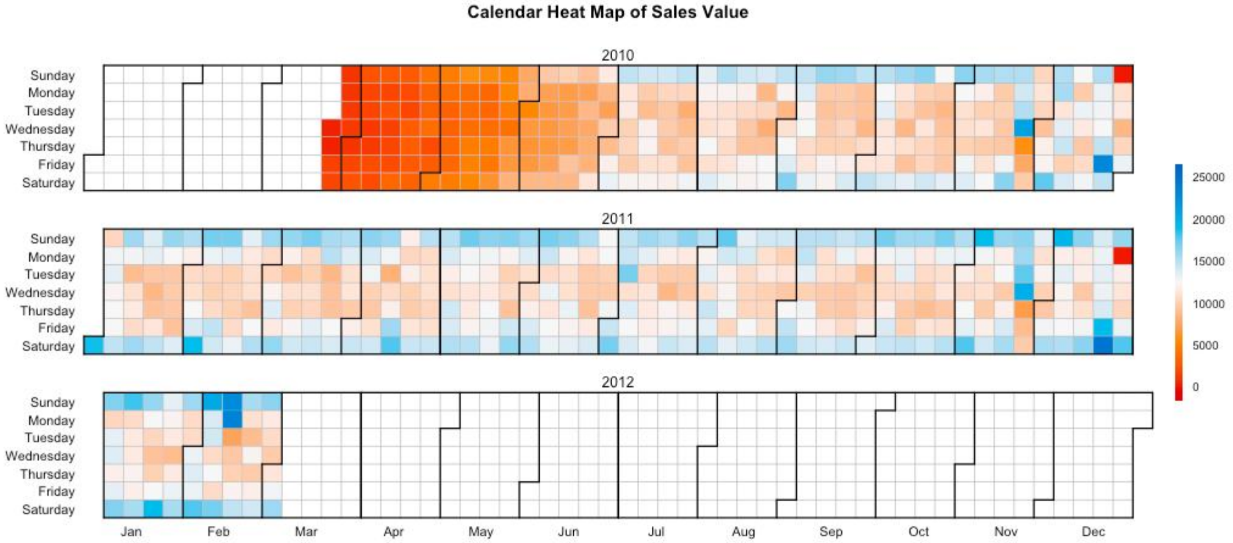
While many coupons were offered, not all were redeemed. Appendix C.13 provides a heatmap that illustrates the timeline in which coupons were redeemed. Although the grocery department offered the greatest number of coupons, the pastry department had the highest redemption rate. (Appendix C.14) Redemption rate varied by age, with shoppers aged 45-54 redeeming the greatest number of coupons. (Appendix C.15) Within each age range, there was also variability in how many coupons from each campaign were redeemed (Appendix C.16, C.17)

Display Location

Imagine when a store manager tries to decide where to put a certain type of product in the available display locations 1~9 inside the store. Is there information he should know? Our group’s exploration of how display location and sales are connected answered this question. A number of different display locations were used to promote products (Appendix C.18), and it seems there is a best display location for many types of products (Appendix C.19). For example, display locations 3 and 5 are hot spots, as a lot of categories of products have top sales at these two locations. For CHEESE, location 2 is the best selling spot. For soft drinks, location 7 is best. The results of visualization provided detailed info for how to arrange product display locations inside a store.

Selected Visualizations

Sales Trends



A calendar heat map is a graphical representation that can visualize values over days in a calendar. Moreover, using this heat map allowed us to see the overall pattern in the data set. In this graph, our group used red color to represent the days that had a high sales value, and used blue to represent the days that had a low sales value.

Weekly pattern of sales value

- Customers usually go shopping on the weekend.

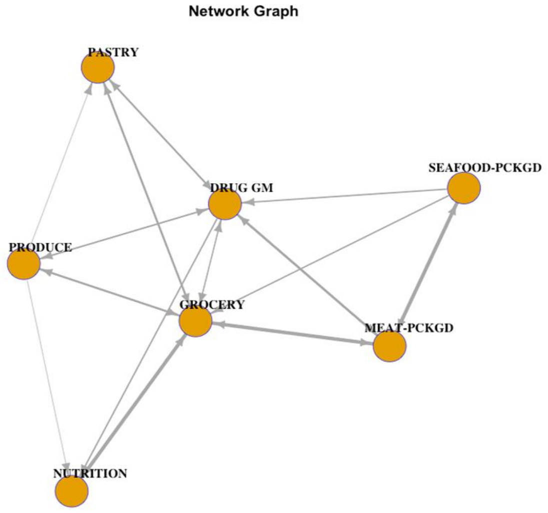
Daily pattern of sales value

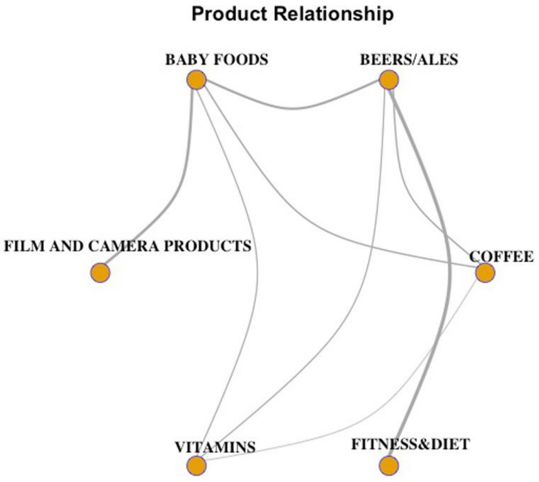
- The customers spent more money on some special days. For example, the customers spent a lot of money at this grocery store on the days before Thanksgiving and Christmas Day.

Ideas for further exploration

- Sales could be plotted against the volume of items sold to see if some days sell more expensive products

Department and Product Buying Associations





Network graphs revealed the relationships between different departments and product categories. With this plot, we used lines with direction, and combined the thickness of lines to represent relationships between collections of data (departments/products) that related to buyers' habits.

Department patterns

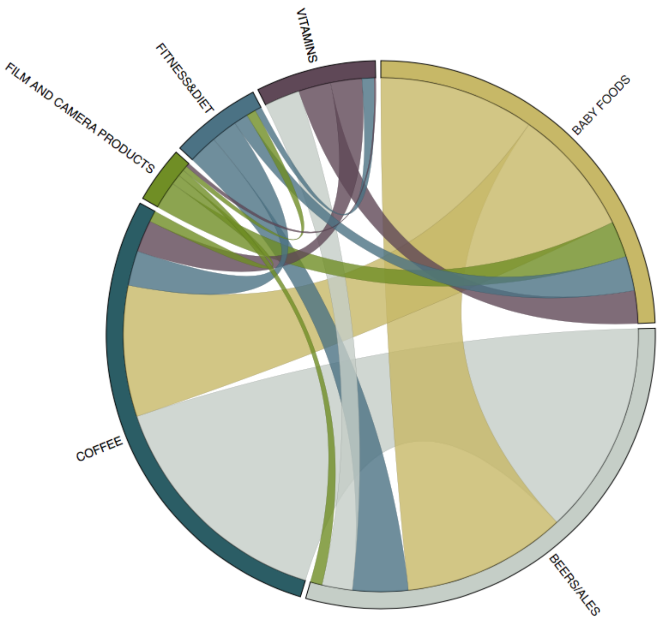
- Grocery is the main department that customers always shop.
- The top three department pairs that have the most shopper traffic are Grocery:Nutrition, Grocery:Meat-Packaged and Meat-Packaged:Seafood-Packaged.

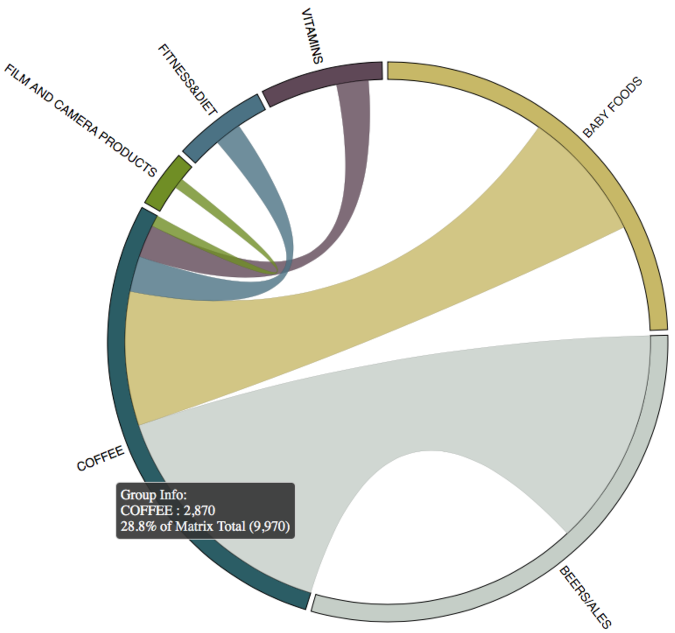
Product relationships

- Shoppers who have a baby are most likely to buy film/camera products and beer, rather than vitamins.
- Customers who usually buy fitness/diet products tend to pick up beers/ales, as well.
- The most coffee buyers possibly have a baby (from the thickness of coffee's edge lines).

Ideas for further exploration

- Network maps could be constructed for different stores to see if similar products are purchased together across all stores.
- Customers could be segmented by age, with a separate network graph for each. This would enable us to see if buying habits change with age.





Group Info:
COFFEE : 2,870
28.8% of Matrix Total (9,970)

Our group used chord diagram as an alternative beyond network graph to display inter-relationships among the products in the grocery data set. Each product is divided and represented on the edge of a circle with different colors. This chord diagram was developed by D3. One important advantage of using D3 is that it is an interactive webpage. It also allows you to hover the mouse over the products to show the inter-relationship of each product. Coffee for example, you will see multiple lines with different colors and thickness clearly show the relationships between coffee and another product.

Link for chord diagram

- http://demo.lanbig.com/CSC465/d3-chord-diagrams-master/pro.html
- http://demo.lanbig.com/CSC465/d3-chord-diagrams-master/dep.html

Product relationships

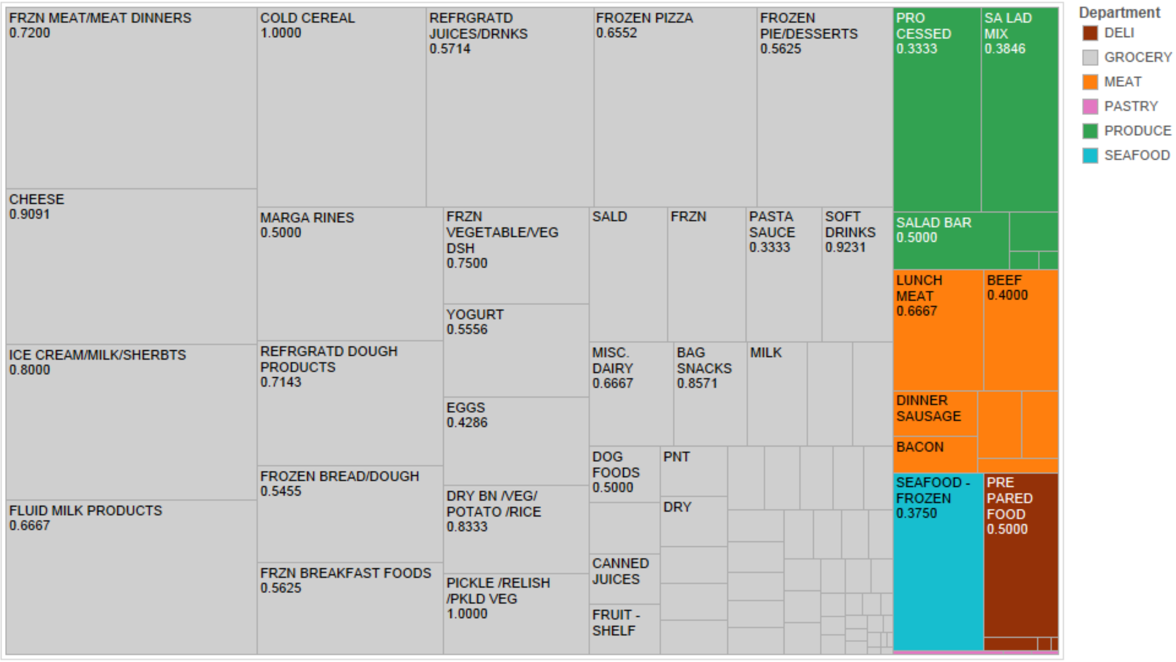
- Customers are most likely buy coffee with beers/ ales, and least likely to buy with film and camera products.
- People who have a baby are more likely to buy baby foods with beers/ ales and coffee

Ideas for further exploration

- All products could be added to the chord diagram and should be divided by department and arranged until the final diagram is not too crowded.

Coupon Redemption (With distinct coupon redemption rate)

Coupon redemption in Food Department (with distinct coupon redemption rate)



Commodity Desc and Distinct_coup_redem_rate. Color shows details about Department. Size shows count of Coupon Upc. The marks are labeled by Commodity Desc and Distinct_coup_redem_rate. The view is filtered on Department and Exclusions (Commodity Desc,Department). The Department filter keeps 11 of 33 members. The Exclusions (Commodity Desc,Department) filter keeps 345 members.

The main reasons to employ a tree map are to present the hierarchical relationships and show the portion of each subcategory within each category or within the whole picture. In our exploratory stage, we found that the food departments offered the most coupons. We wanted to see how people responded to these coupons. This tree map shows how the redeemed coupons were distributed across different food departments. Because we don’t the total number of coupons that have been sent to customers and we want to take the difference in the number of coupons offered by each departments into account, we introduced the distinct coupon redemption ratio. Although it cannot tell us what are the most popular commodity coupons, it somehow shows people's interest (or preference) on different commodity coupons. So we add the distinct coupon redemption rate as a label in the tree map, which is calculated by (counts of distinct coupons redeemed / counts of distinct coupons offered).

Redeemed coupons came from different food departments

- Most of the redeemed coupons came from the grocery department.
- Only small portions of the redeemed coupons were used in the seafood department and pastry department.

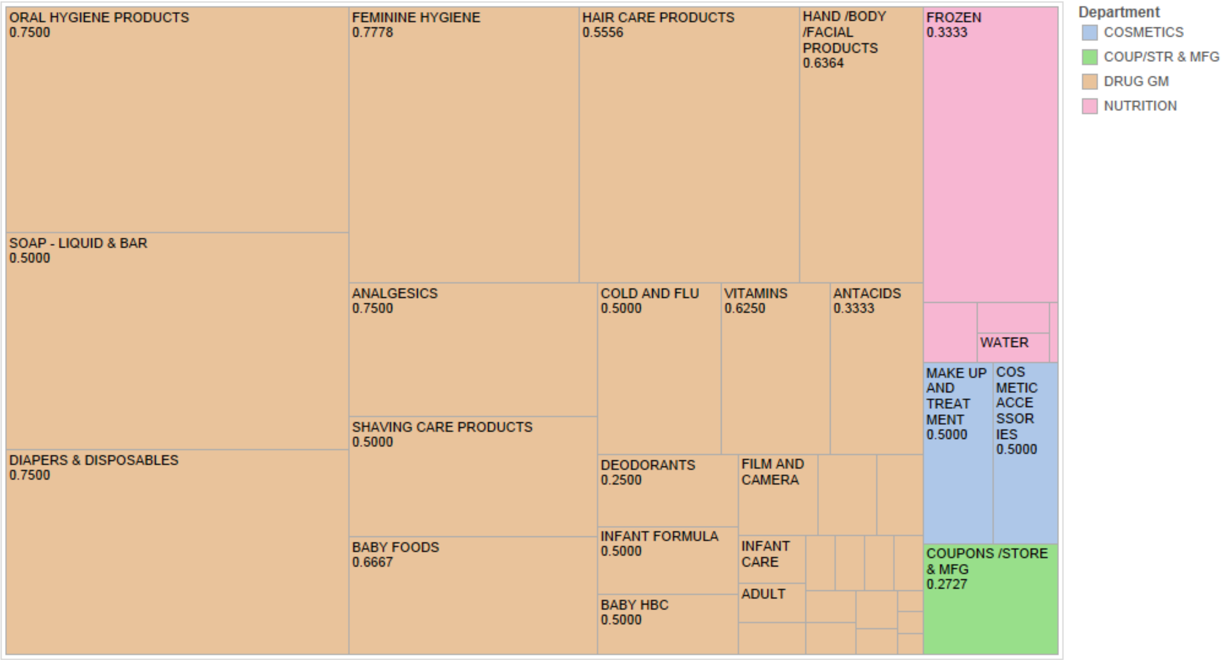
Distribution of redeemed coupons within each department

- Large portions of the redeemed coupons from the produce department were used to purchase processed food and salad. However, these only account for a small proportion of the coupons that were redeemed in the food department.

Distinct coupon redemption rate

- The commodities like cheese, ice cream and cold cereal have the highest distinct coupon redemption rate, which means most of coupons offered for those commodities have been redeemed by at least one customer. These commodities also have a relatively large number of redeemed coupons. This shows the potential of increasing sales by offering more coupons for those commodities.

Coupon redemption in Non-Food Department (with distinct coupon redemption rate)



Commodity Desc and Distinct_coup_redem_rate. Color shows details about Department. Size shows count of Coupon Upc. The marks are labeled by Commodity Desc and Distinct_coup_redem_rate. The view is filtered on Department and Exclusions (Commodity Desc,Department). The Department filter keeps 22 of 33 members. The Exclusions (Commodity Desc,Department) filter keeps 344 members.

This tree map shows how the redeemed coupons were distributed across non-food departments.

Redeemed coupons came from different non-food departments

- Most of the redeemed coupons came from the drug-gm department.
- Only small numbers of the redeemed coupons came from the coup/str & mfg department.

Distribution of redeemed coupons within each department

- Large portions of the redeemed coupons from the drug-gm department were used to purchase oral products, diapers, feminine hygiene products, soap and hair products. The number of coupons redeemed in these subcategories account for half of the coupons redeemed in non-food department.

Ideas for further exploration

- Tree maps could be constructed for a number of different demographics, to see if the coupons redeemed in each department change with customer age, number of children, household income, etc.

Distinct coupon redemption rate

- The commodities like feminine hygiene, oral hygiene and diapers have the highest distinct coupon redemption rate, which means most of coupons offered for those commodities have been redeemed by at least one customer. These commodities also have a relatively large number of redeemed coupons in non-food department. This shows the potential of increasing sales by offering more coupons for those commodities.

Analysis of Results and Discussion

Our group has created interesting visualizations through everyone's exploration of the data. The selected three visualizations revealing interesting patterns from the data, and apply the powerful techniques we have learned.

Heatmap

From the first day of class, we learned that the goal of visualization is to produce visualizations focused on efficiency, little or no distortion, bringing features of data to light, and techniques that don't detract from the message. The heatmap we selected is a demonstration of these goal.

All of our group members tried heatmaps in exploring different aspects of the data. This technique is very powerful and can elegantly reveal the patterns in a timeline. The heatmap we selected in the end is one that was produced in R and demonstrated clarity and accuracy. We found that Tableau is very efficient to produce graphs, and R is very flexible in showing the data with clarity and without clutter. In data visualization, the advantages offered by these different tools are very important.

Network Graph

The advantage of network graphs are being fully applied to our dataset, as we had a huge amount of information on a collections of products. The network graph makes it relatively easy to reveal relationships between two variable in a two dimensional space. To reveal relationships between multiple products, we have tried different techniques like bar charts, treemap, clustering and so on. In the end, the network graph was the one that worked best in this situation.

Our selected network graph visualization clearly shows the correlations among departments products. These revelations could help a retailer improve their sales in many ways; for example, by placing the highly correlated products in close display locations, or by placing departments with close relationships next to each other to improve the customer shopping experience.

Treemap

Treemaps are visualizations that can give us a high-level view of our data, while also showing the line item details. Our eye visually aggregates rectangles in the same group, allowing us to see patterns quickly. Also, we can easily find out the relative sizes of the first-level groups compared to each other, or the second-level groups within each of those groups.Treemaps are good at displaying hierarchical data in an organized manner.

The problem we encountered when using a treemap for our data is that it is hard to choose color combinations that are different enough to be recognized by everyone who read the graph, since we do not interpret precise color values well, or perceive colors in different ways. The result of the chosen colors may not be satisfying for everyone. To make a clear tree map, we dug into the data we had and combined the similar departments into one department to avoid a tree map that was too colorful and might distract the audience's attention. By carefully selecting colors for each rectangle, especially for the largest and smallest departments (rectangles), we tried to not distort the audience's cognition. Additionally, we had a hard time making the plot show the text labels of each rectangle. We made some efforts to edit the alias of the sub-categories so as to display the text labels of the relatively larger rectangles as needed.

Discussion Summary

Besides the visualization techniques we selected, we have also explored other techniques like level plots, univariate group comparisons, violin plots, clustering and so on. We think there is no single best technique, as the visualization results depends on whether a technique is fit for the dataset and how it is applied. The most important thing is to reveal the hidden patterns in the data with clarity and accuracy, without being distracted by techniques.

Appendix A: Personal Summaries

Jennifer Boyce

For this project, I explored pattern of product sales over time, as well as general summaries of the products that were offered. I created graphs to understand both patterns of individual product sales during a 24-hour period, as well as sales during a calendar year. Within our group, I also served as the group liaison.

Graph 1

In my first visualization, I looked at the hourly sales between two top-selling beverages: beer and orange juice. These products were chosen because they represented two separate buying motivations- nutrition and a recreational treat. By graphing the hourly pattern, I sought to discover whether individuals' shopping patterns reflect these differences.

Orange juice sales mirror overall grocery buying behavior, which seems to indicate that juice is a staple commodity purchased during routine shopping trips. While most beer sales seem to occur during routine shopping trips, there is a sharp increase in sales at 8pm, indicates that beer may be an item for which customers will make a special trip to the grocery store. As these special trips occur in the middle of the evening, they seem consistent with beer being an item consumed during after-work relaxation time.

Graph 2

Some foods are highly associated with certain holidays. In my second graph, I looked at three foods commonly associated with U.S. holidays: eggnog, pumpkin pie, and watermelon. By graphing the three items on the same timeline, it was easy to see spikes in sales that occur around each holiday season.

The seasonal trend was especially interesting with watermelon, as it is typically available throughout the summer. Despite its constant availability, there are still spikes in sales that correspond to holiday picnics at Memorial Day and Independence Day.

Conclusions

From these visualizations, I discovered that while most products follow a similar sales pattern, there are items that exhibit strong hourly or seasonal variability. This variability suggests that stores would benefit from sales strategies and promotions that specifically capitalize on these trends.

Over the course of this project, I learned that visualization easily highlights patterns that would otherwise be difficult to discern. For example, in visualizing juice and beer sales, statistical summaries make it difficult to see just how dramatically different the buying patterns are for the 45-54 year olds. To see this trend, it would be necessary to look at hourly summary statistics for each group. Once visualized, the pattern is immediately apparent.

Li Huang

- **Introductory**

Our group has chosen the dataset of a retailer's yearly sales information, including datasets like products, coupon, campaign, transaction, and customer. There are two subgroups in our group, one focus on coupon and marketing, and another focus on product and store sales. As I am in the product group, my personal work starts with two directions: product and store sales. And based on this two direction and the result of further exploration, I have developed this final report here:

- **Basic summary views of the data**

To find out if my two directions would reveal something interesting and what are the proper technique to apply on the dataset, I have made some initial try outs, for example, summary of location and sales relationship, summary of sales by week & day.

My main focus of basic summary was on the display location summary: The display location seems does not affect the income of grocery, one possible reason is that the grocery department includes a variety of products that located everywhere in the store. While for other departments, whether the income has relationship with the display location needs to be further confirmed. For example, when we exclude grocery and plot the rest departments with display locations, from the result, it seems meat-packed department sells best at display 2, and DRUG GM department have the best sale at display7. And we could repeat the process and find out that produce, deli and meat department sell best at display0. Nutrition sells good at display3,6,7 and 9. Pastry at 0,1 and 7. Cosmetics at display 0, seafood packed at 5 and 9.

I am wondering why grocery has the highest sales, much higher than any other department, so I want to look at grocery's location and included number of products. From the graph, I found out the top four sales are display location 3/5/7/2, and with similar number of products, they have different sales. So, it seems the sales have relationship with location and products included. So we could say for department GROCERY, the better display location 5 is better than 7, and display location 6 is better than 1, A is better than 4.

We could explore further about in department grocery, why location 5 is better than 7, does it has to do with the sub-products types, or brand or any promotion going on for each. I was trying the analysis of two stores with similar variety of products and different revenue.

- **Further exploration of data**

Based on the basic summary of the data, I found out the two directions could both provide something interesting. But I still need to confirm which direction has more potential to produce a final visualization that meets the standard of clarity and accuracy.

1. department sales heatmap over time

From the heat map, we could easily find out that many department GROCERY have the highest sales at week 10, the next best is department DRUG GM at week 7. This is something we could explore later to find out the reason.

2. Store sales by tree map

From tree map, we found out store 450 achieve the highest revenue at week12, and the best results mostly come from week 10~12. This coordinate with what we found earlier, it seems in the first month, as times goes by, sales improves.

3. Department sales and display location

From the bubble graph of department vs. display marked by revenue, we found out the department grocery has the biggest income among all the departments. Pastry /deli/meat department has the lowest income.

- **Final visualization formation**

When our group has the basic summary and exploration of the data, we discussed about everyone's possible direction to avoid repeating of the same direction. So my direction now is going to focus on how to bring out the difference between display location and department sales.

Based on the summary of previous work, when we consider display location and department sales, the department GROCERY has much better sales than all the other department that the effect of other department's results are not visible. That's why I am going to use the Commodity Desc. , which is the further development of product category under each department. There are around 400 categories in Commodity Desc., so to make the result clear and readable, I have chosen the 15 which the best sales over year1. And here is the result I have.

1. Find out the top 15 sales sub-categories from Commodity Desc., based on sales of all sub categories over year1.

In this heat map, the 15 sub categories with the best sales are kept only, and in a descending manner. We could also observe that the sales in year1 for the 15 categories are increasing over time, especially for the category FLUID MILK PRODUCTS and SOFT DRINKS.

2. Find out the location of the top 15 categories products' best selling location over year1:

From the heat map of display location and 15 sub-category from 'Commodity Desc' with the best sales, we could say for FLUID MILK PRODUCTS in GROCERY, location 3&9 is the ones with best sales. The effect of other categories are not so clear, so we apply a whisker plots.

3. Applying different method to find out the location of the top 15 categories products' best selling location over year1:

Based on the heatmap and whisker plots, we could observe easily that for FIUID MILK PRODUCTS, location 3&9 is much better; for Cheese, location 2 is better; for FRZN MEAT, location 5 is better.

- **Reflections on the project**

I have done projects in other courses before this quarter; however, I think our project takes longer time than before. As it is not a question seeking a correct answer, but a process of exploring and refining the visualization that best represent the goal we learned in class, showing the pattern in data with accuracy and clarity.

It is like an open question. I have several potential exploring directions in the beginning, but to reveal and decide which is the most interesting one is not easy. And cooperation with group member is also important, as sometimes we probably going in the same direction, which is something we tried to avoid in the project to brings as much as possible out of the data. And our group's discussion is very beneficial to everyone, ideas keeps coming up and enlightens everyone in their own direction of exploration.

I think overall, doing the project makes everyone of our group learn to value other people's accurate and easy-to-read graph as we could imagine how much work they put into it. And teamwork in visualization is very important, as it will produce a final visualization which is deemed fit for a group, not just yourself.

Rungtham Rodrawangpai

In this project, I've used four different kinds of visualization techniques, which were calendar heat map, chord diagram, network graph, tree map and violin plot.

Graph 1 (Selected Visualizations - Calendar Heat map, Appendix C.15. Coupon Redemption Rate)

First, I used the calendar heatmap to visualize the overall pattern of sales value. I used red color to represent the days that had a high sale value and I used blue color to represent the day that had a low sales value. For weekly pattern of sales value, the customers usually go for shopping on the weekend. For daily pattern of sales value, the customers spent more money on some special events or days. For example, the customers spent a lot of money on this grocery store on the day before thanksgiving and Christmas day. Moreover, I've created another calendar heatmap to visualize the number of coupon redemptions throughout the years. The plot demonstrated that the customers were more likely to redeem the coupon in August and November.

Graph 2 (Selected Visualizations - Chord Diagram)

Second, I used D3 to develop chord diagram to show the relationship among the product. The data matrix of this chord diagram was pre-processed by Witawat. This chord diagram is an interactive webpage which allows you to hover the mouse over the products to show the inter-relationship of each product. I've uploaded this chord diagram to my website. You can check it out from the link below.

Link for chord diagrams
http://demo.lanbig.com/CSC465/d3-chord-diagrams-master/pro.html
http://demo.lanbig.com/CSC465/d3-chord-diagrams-master/dep.html

Graph 3 (Selected Visualizations - Department Association)

Next, I've used network graph to visualization the relationship among product departments. This visualize technique was similar to the chord diagram. For the network graph on the right side, I used the number of edges to show the strong/weak relationships among departments. For example, Drug GM had a strong relationship to grocery and produce departments but it had a weak relationship to seafood-pckgd department. For the network graph on the left side, it was similar to the network graph on the right but it had a direction. I also used the thickness of edges to represent the strong/weak relationships. For example, the customers who shop in seafood-pckgd department tent to go buy grocery or drug for the next items.

Graph 4 (Appendix C.9. Departmental Purchases by Customer Age)

This tree map sought to explore the customer age and what type of products that customs were more likely to purchase at the store. This graph can help the store target their customers more effectively. Size shows sum of sales value. Color shows the number of coupons applied to the product departments. The marks are labeled by customers age and product departments.

Graph 5 (Appendix C.16. Coupon Redemption Rate by Customer Age)

The last visualization that I used was a violin plot. This plot revealed the distribution of the customers' age range and number of coupon redemption. As the violin plot demonstrated, we can see that the number of coupon redemptions by the customer age are not normal distributed. They are skewed to the low number of coupon redemptions.

Project Reflection
In this course project, I have learned new type of graphs and new tools for data visualization. For example, I learned how to use R and I used it to create a calendar heat map. It was very powerful and could reveal the pattern throughout timeline. Furthermore, I've learned the basic of D3 in class. D3 is a good tool for data visualization. I used it to create a chord diagram. It can visualize the department association elegantly and it is interactive. Additionally, I used R to develop network graph. It was a really good visualization that can reveal relationships between multiple products. This can help the retailer improve their sales. They can place departments with close relationship next to each other which can improve the customer's' experience. Last but not least, I used R to graph a violin plot of the distribution of the customer age and number of coupon redemption. I learned that violin plot was one of a good type of graphs that can reveal the shape of the distribution elegantly.

Witawat Daungjaiboon
Throughout the project, I was responsible to investigate the marketing aspect, but in the end I also use the advanced data visualization technique, network graph, to reveal product relationship. I have

discovered the different patterns in coupon redemption, how much different ages customers spend, the most successful campaign and product relationship through multiple data visualization techniques with different applications.

Graph 1: Simple Bar graph

This graph reveals what age range spend the most when transactions occur. I present this chart as a bar graph including zero because it will not be exaggerated if I did not include zero. As we can see, the people whose age between 45-54 spend the most among others.

Graph 2: Univariate Scatter Plot

Graph 2 shows which age range use coupon the most, and which campaign seems to be successful for different age customers. We can see the customers with the age of 45-54 redeem most and almost every campaign, unlike the 19-24 customers use the least coupon. This visualization will be helpful for the company to improve more efficient discount campaigns.

Graph 3: Violin Plot 1

This graph shows customer in which age range redeem the most coupons divided by campaign (Y-axis). The width of violin plot represents number of coupon redemption. We can obviously see Campaign number 18 is a big hit across all customers' age range except 35-44 and 19-24 (can be clearly distinguished in graph 4). People with the age between 45-54 redeem most coupon across all customers, and 19-24 use least coupon.

Graph 4: Tree Map

This is a coherent of graph 3 that helps us to clearly separate which campaign is the most successful (represented by area of squares) divided by age range (presented by different color). Campaign number 18 is the most successful in people with the age of 45-54, 25-34, 55-64 and 65+, and campaign number 13 for 35-44 and 19-24.

Graph 5: Violin 2

This plot shows the distribution of sale values summary by different age range. Having changed from normal scale to logarithmic scale for sale values summary helps the data exhibit itself better rather clumping on the X-axis. Overlaying box/whisker plots on violin plot also helps us to see which group of customers tend to spend the most. From the normal scale plot (left), the box/whiskers plots also help us detect the outliers that may affect analysis result.

Graph 6: Network graph

This graph reveals the relationship between different products being randomly chosen linked together by lines. Linewidth show how much they are correlated to one another. The thicker line, the more highly correlated of products are. For example, from the thickness alone, we can see people who have baby are more likely to buy film/camera products, and beer rather than vitamins. The coffee and fitness & diet products are not correlated at all.

Graph 7: Chord diagram (matrix creation phase)

I created the matrix shows relationship between products for adding to D3 to create a chord diagram from randomly selected some products from 308 types of products.

Having explored and analyzed the data through different visualization techniques, I found the different patterns that are potentially helpful for retailers to understand their customers better in order to make the decision for store improvement and gain more profit. I personally like to see pictures to get better understand on everything, so does everyone else. Indeed, looking at the data alone does not help a non-technical person have an idea what is going on, yet after the proper visualization techniques been applied, the insights/patterns hidden in the data are revealed with easily understandable. Different charts reveal different aspects,network graph for example, it reveals relationship between different products so the stores can improve the layout of the store in order to increase sale volume. For a violin plot, it does not only show the data distribution, but it can also show volume/spike in the data (Coupon redemption by

Age and campaign number). All in all, the data visualization definitely facilitates the data analysis and decision-making processes.

Yingjie Cao

In this project, we have generally two directions. I am in the marketing direction and try to find interesting facts that could help the stores improve sales. I mainly focused on discovering the influence of the coupons on sales of different departments and tried to find the interesting features that might help a store to improve their sales.

Exploratory stage: (Simple Histograms)

At beginning, I did some exploratory plots (histograms) to find the demographic features of the customers who redeemed coupons. For example, I found that that customers fall in age range (45, 54) and (35-44) tends to redeem more coupons comparing to other age groups. Also the customers with one kids accounts for the largest portion of coupon redemption. But these results might simply due to the facts that the customers in those groups dominate the customers of the stores.

Graph 1: (C.12. Number of Distinct Coupons Offered by Department)

Then I wanted to find out how customers responded to the coupons from different departments. I used python to preprocess the raw coupon data and excluded the identical (duplicated) coupons. Then I plotted a histogram to find out how the offered coupons distributed across different departments. And I found that most of the coupons are came from GROCERY (department), which was as I expected, since the grocery department was the largest department.

Graph 2: (C.14. Distinct Coupon Redemption Rate by Department)

Here is to find out the redemption rate of the distinct coupons across different departments. The coupon redemption rate is the ratio between the number of distinct coupons redeemed by customers and the total number of distinct coupons. The reason I introduced this ratio is that I want to minimize the influence of the size of the department when evaluating the popularity of the coupons from different departments. This time I have some interesting findings. The coupons from pastry have the highest redemption rate. The coupons from DELI, GROCERY and COSMETICS are also popular. However, the coupons from MEAT and SEAFOOD have the lowest coupon redemption rate. Although this redemption rate could be affected by other reasons, but it does show some interesting patterns.

Graph 3: (C.21 Coupon Redemption Tree Map)

In this step, I generated a draft of my final graphs, which is the tree map showing the distribution of the number of redeemed coupon across different departments. Because the tree map is one of the few visualizations that can give us a high level view of our data while also showing the line item details, I plot a tree map of the counts of redeemed coupons in this session. The first level showed in this tree map is department level and the second level is the commodity category of each department. At the very first stage, I included all departments in the tree map and colored each department with different colors. The graph did not display well, because there were too many departments in the graph and this made the colors representing the departments unavoidable similar. Also the text labels of showing the commodity categories could not be showed well due to the issue of label length.

Graph 4: (Selected Visualizations - Coupon Redemption Tree Map)

To solve the issues stated in the previous step, I made some adjustments. Since most of coupons are redeemed for purchasing food, I separate the departments into two parts, which are Food Department and Non-Food Department. Then I combined the similar departments into one department in python. For example, I combined Meat and Meat-Packaged department into Meat department. Besides that, I edited the alias of the commodity category labels to make these alias readable in the tree maps. I carefully selected the colors for each department. For example, I avoided using a bright color for the large departments, used attractive colors for the departments that I wanted to highlight and selected distinct colors for each department.

I also add the distinct coupon redemption rate as a label on the tree map, because I think this ratio number could indicate the customer’s preference on coupons in a way. For the commodities that have a high distinct coupon redemption rate and large number of redeemed coupons, we might consider these commodities as good choices for promotion, since customers like the coupons of these commodities, by offering more coupons for these commodities might increase the sales.

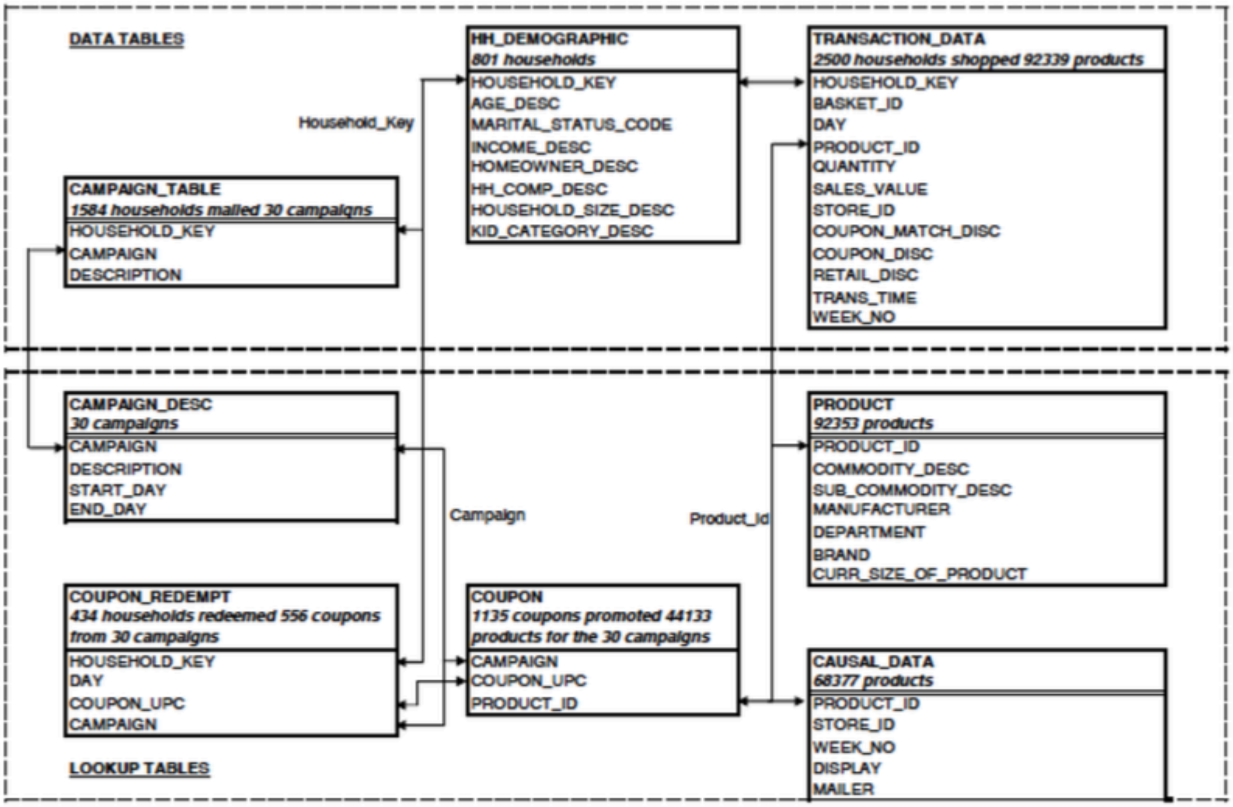
Project Reflection:

In this project, I have a chance to exploit the knowledge I have learnt with a real world data. It’s always important to spend enough time on understanding our data and know what you want to present to your audience.

We can use basic graphs like line graph, bar charts, histograms to explore the data and the distribution of the variables. High-tech graphs usually could show more information, but we also need to be careful about how to present the information clearly, correctly in an easy to understand way. Many times the first graph we came up with is not satisfactory, then we need to redesign or even editing our data to try again and again, it is a recursive process.

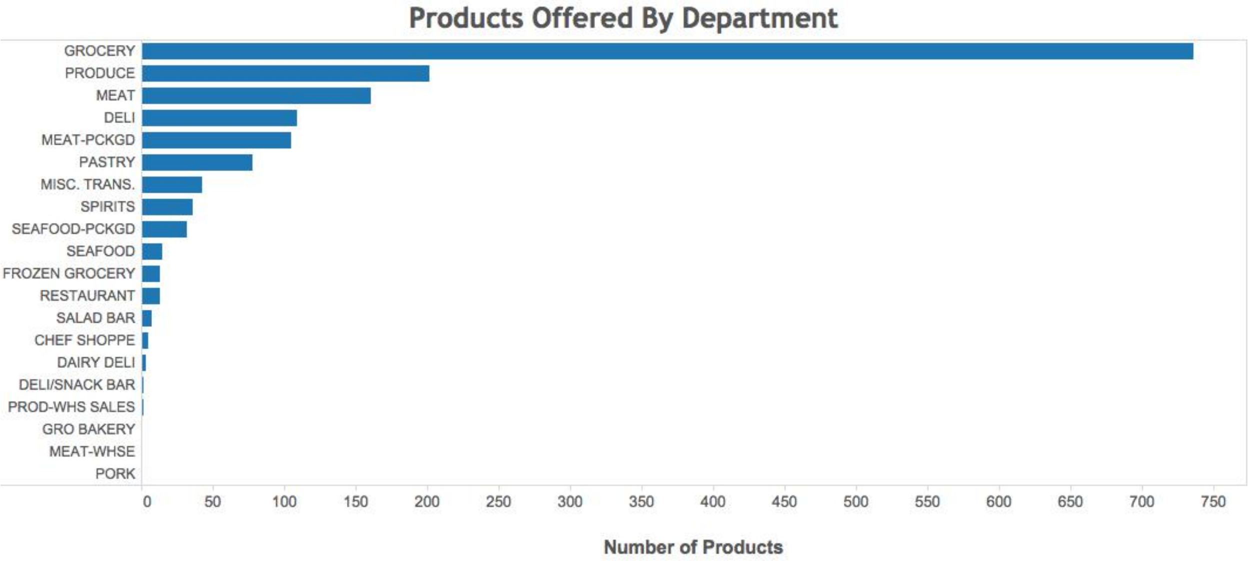
Also since graphs are regarding to perceptions, we need to put a lot of efforts on how to balance people’s perception, especially we need to tailor our graphs based on the background of the audiences. Also we need to choose the right scales, symbols and measurements to present the data to make sure the data is not distorted.

Appendix B: Data Source Schema



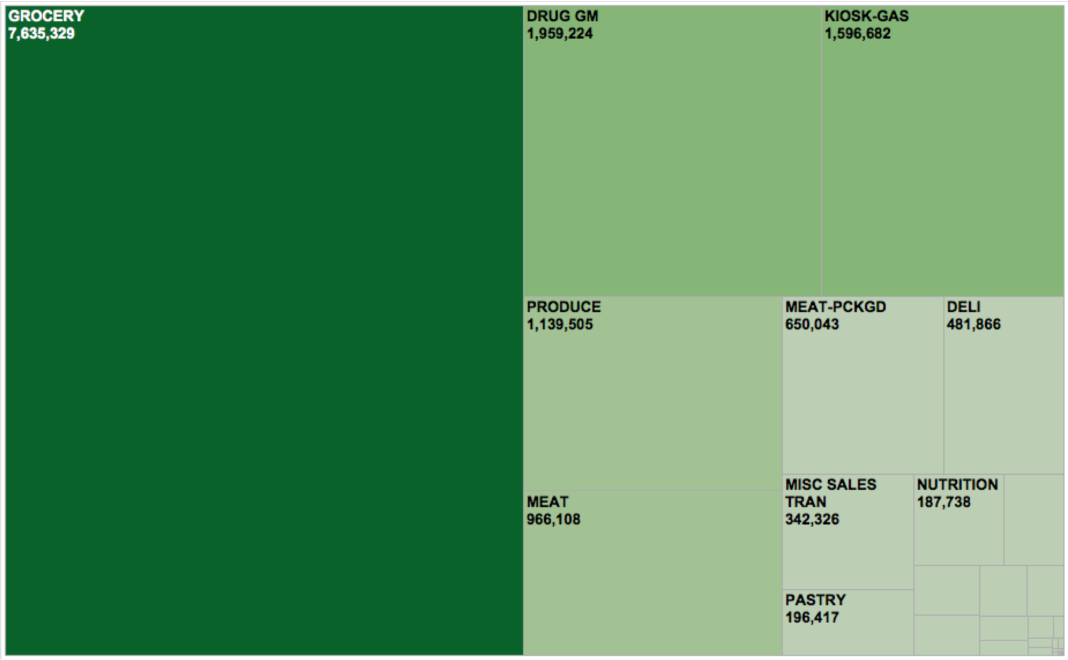
Appendix C: Exploratory Visualizations

C.1. Number of Products Top Departments

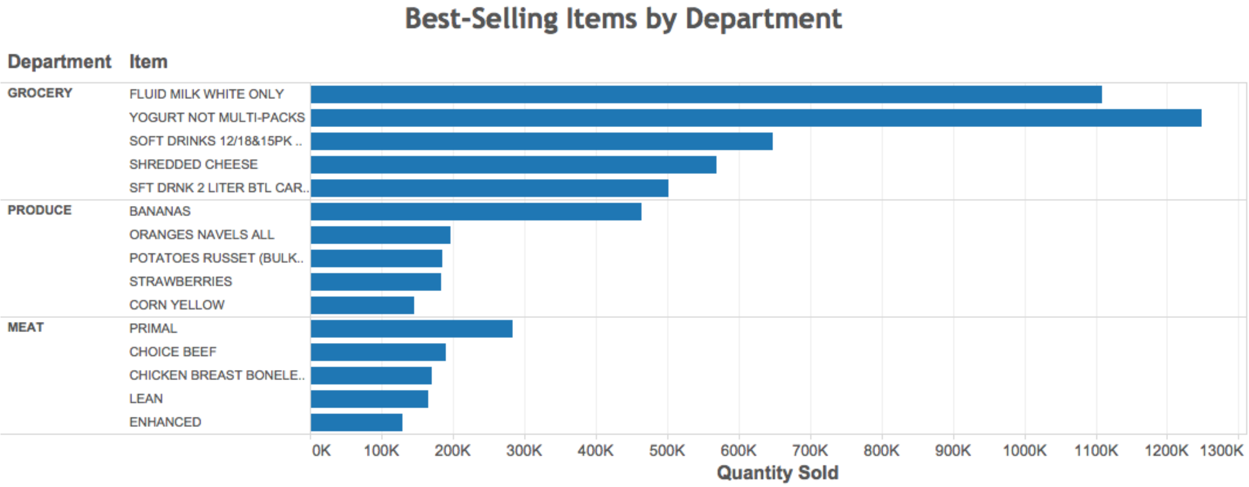


C.2. Sales Revenue by Department

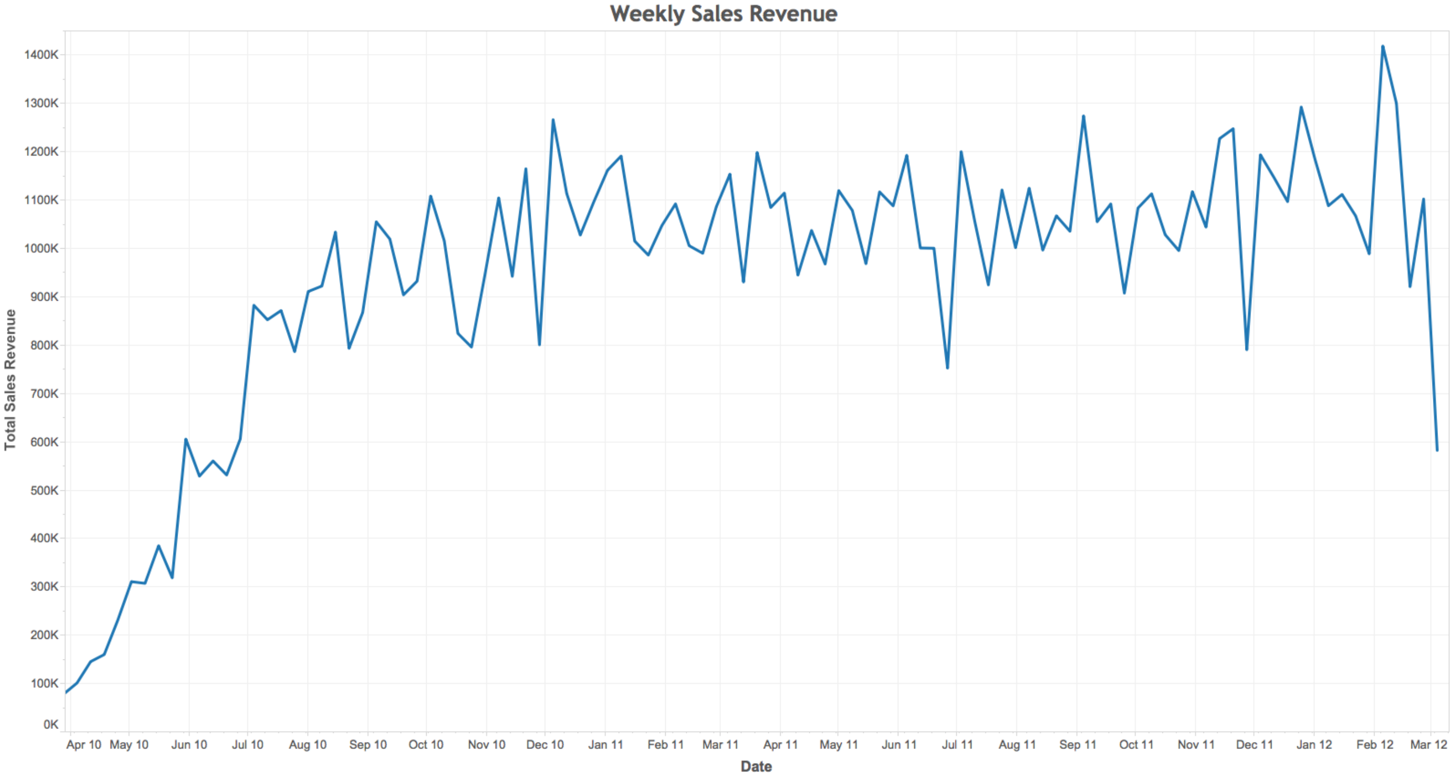
Sales were highest in the grocery department, followed by drug-gm and kiosk-gas.



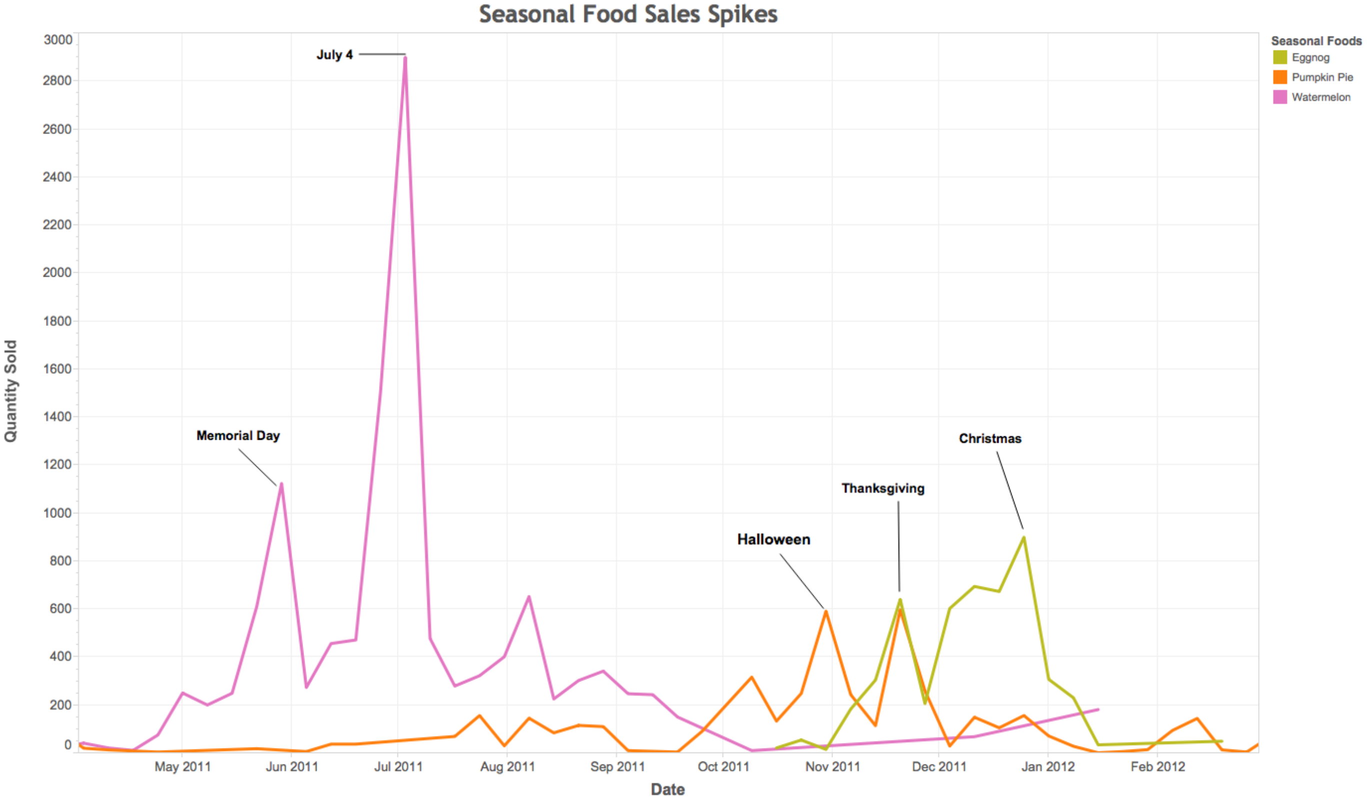
C.3. Best-Selling Products by Department



C.4. Weekly Sales Revenue

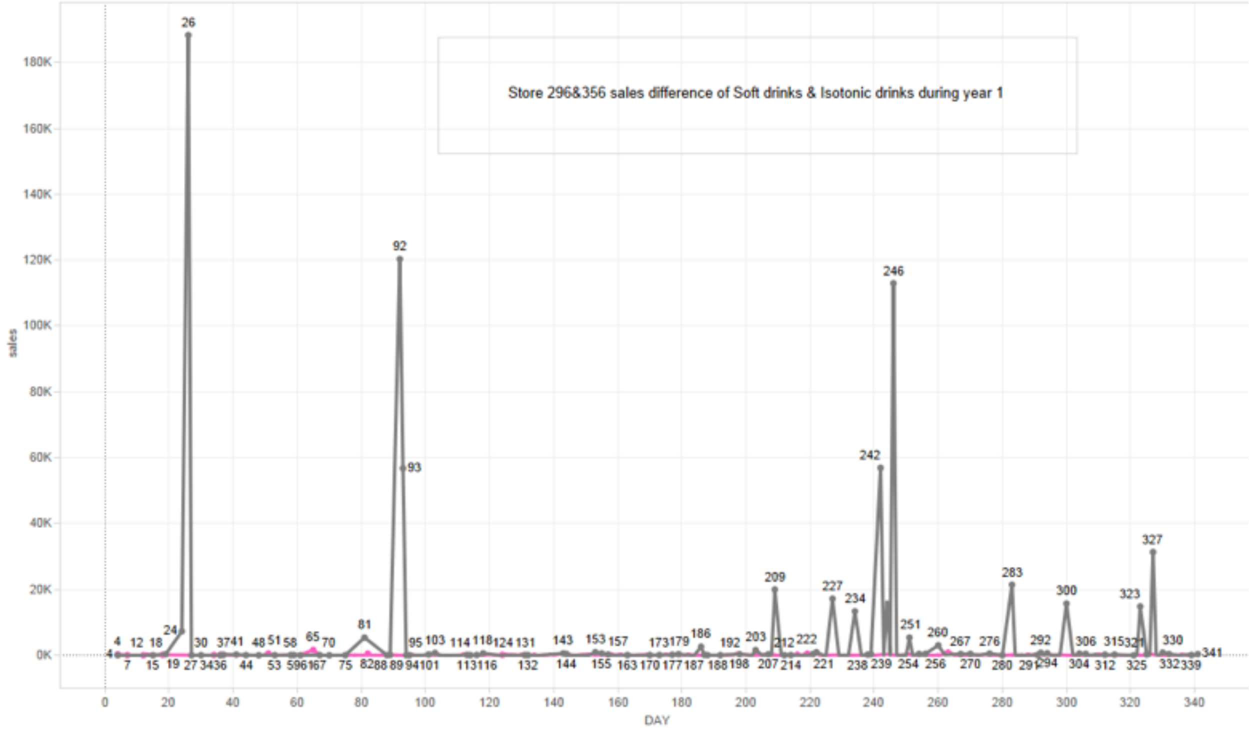


C.5 Sales Volume of Seasonal Products

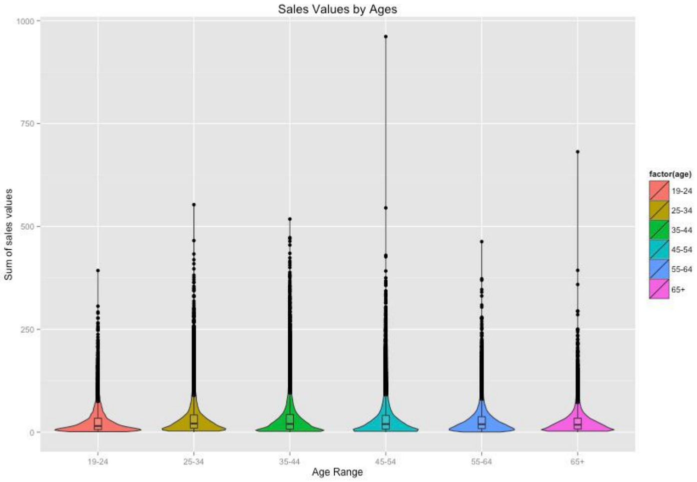


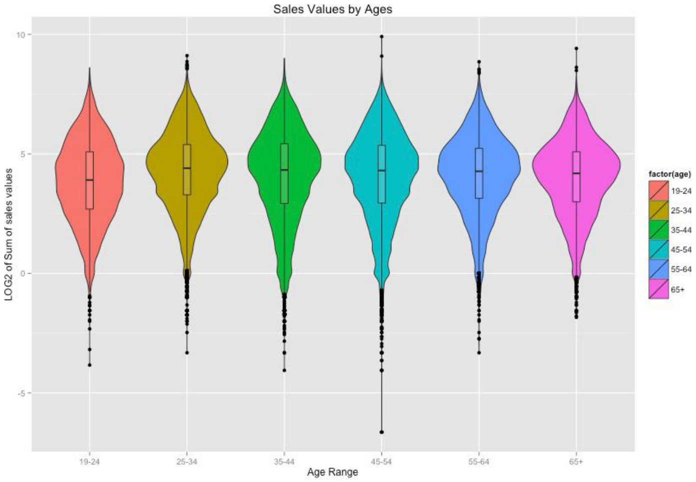
C.6. Compare sales of two stores with similar variety of products during year 1:

There is one observation we made that Store 296’s sales is much more than store 356. By displaying the sales line graph, we could infer that there is some specific days in year 1, days like 26,92,246 and etc., when store 296’s sales is much more than store 356. If we could further analyze what happened for store 296 in these days, for example, if there is promotion activity, then we could find out ways to improve sales for stores like 356.

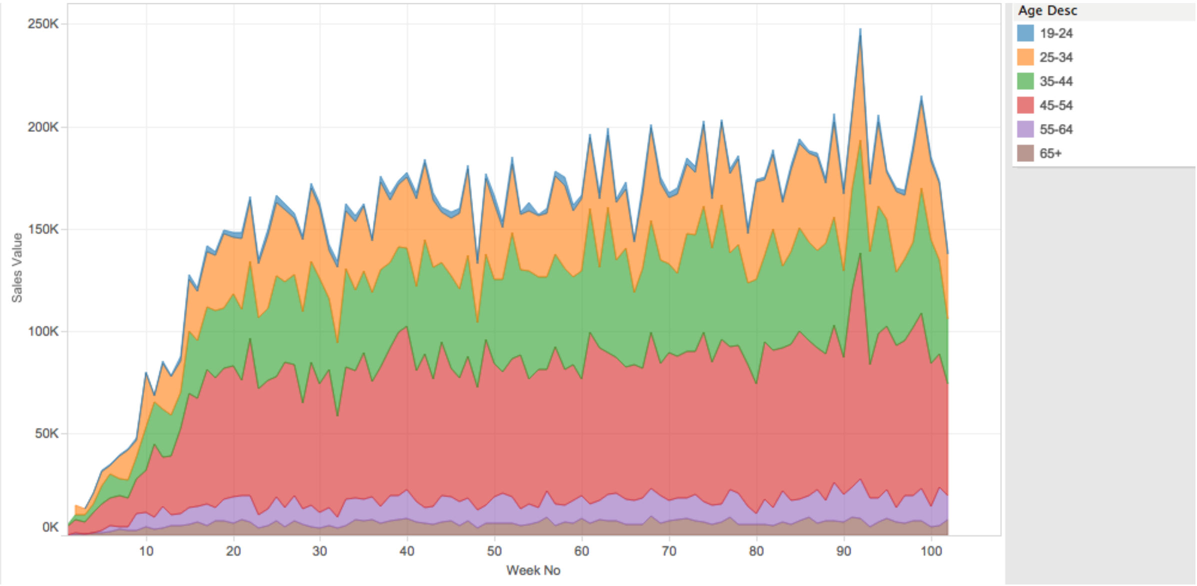


C.7. The Summary of Sale Values by Customer Age

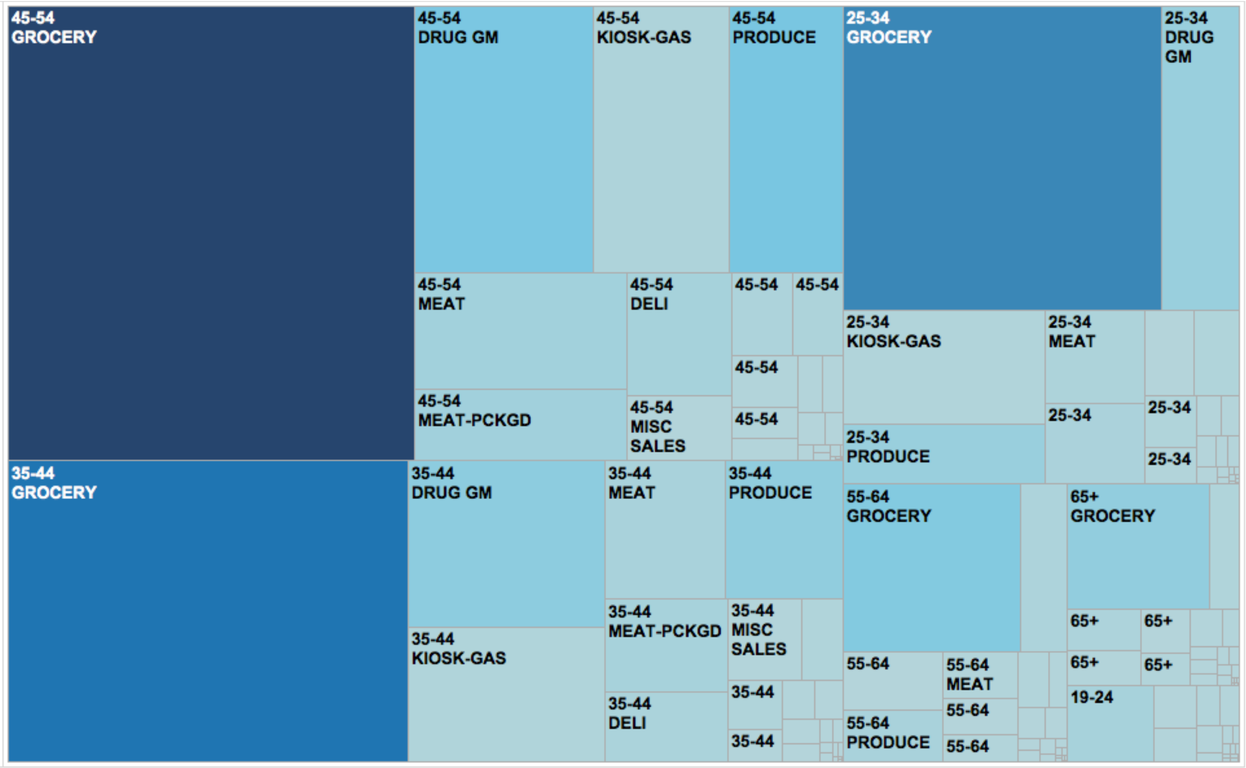




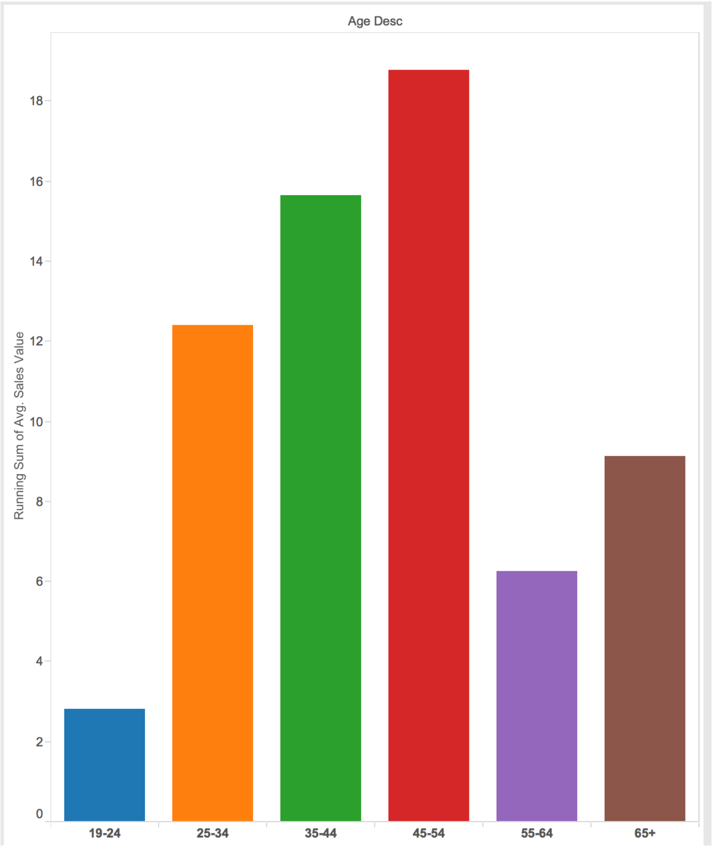
C.8. Sales Revenue by Customer Age



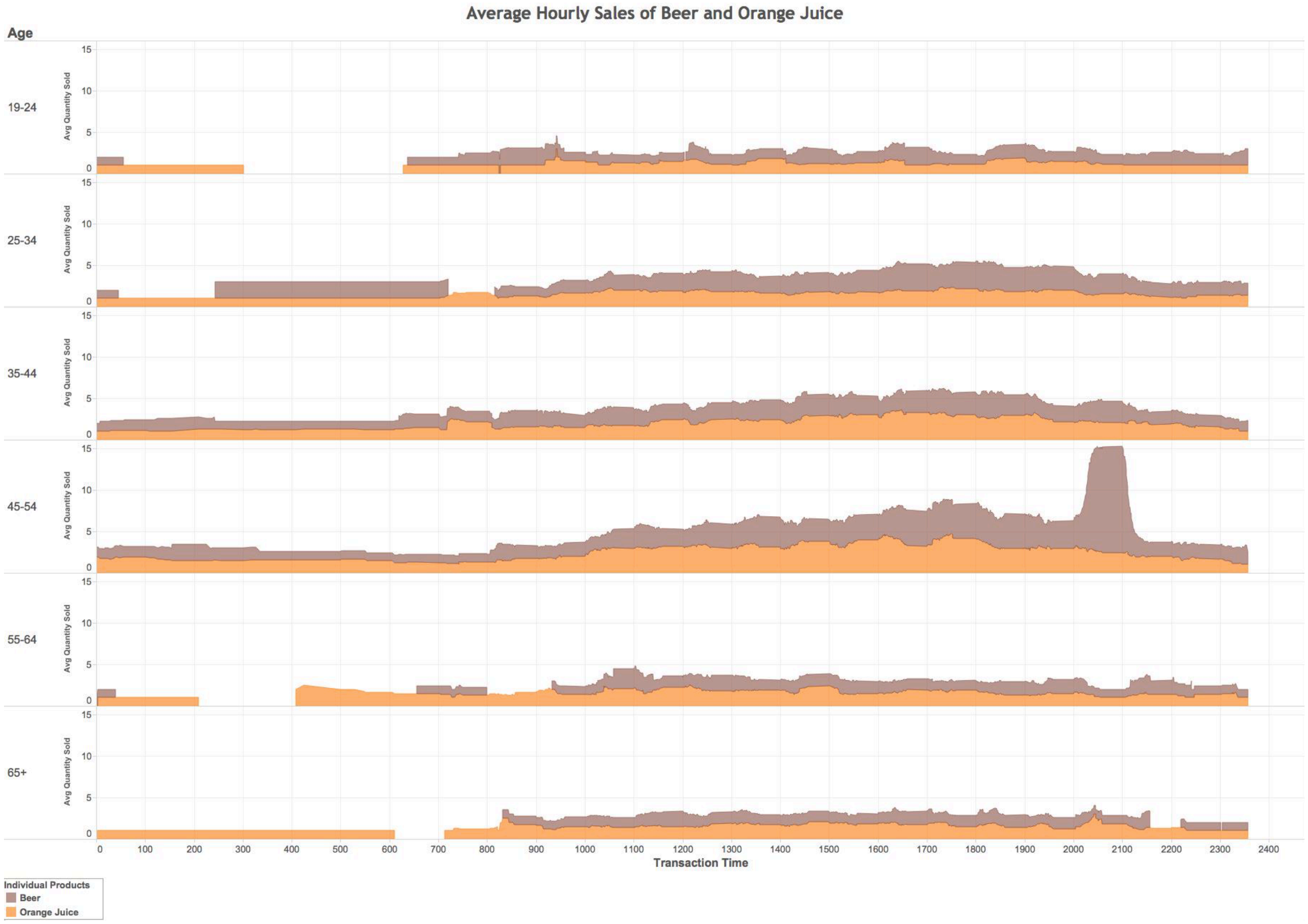
C.9. Departmental Purchases by Customer Age



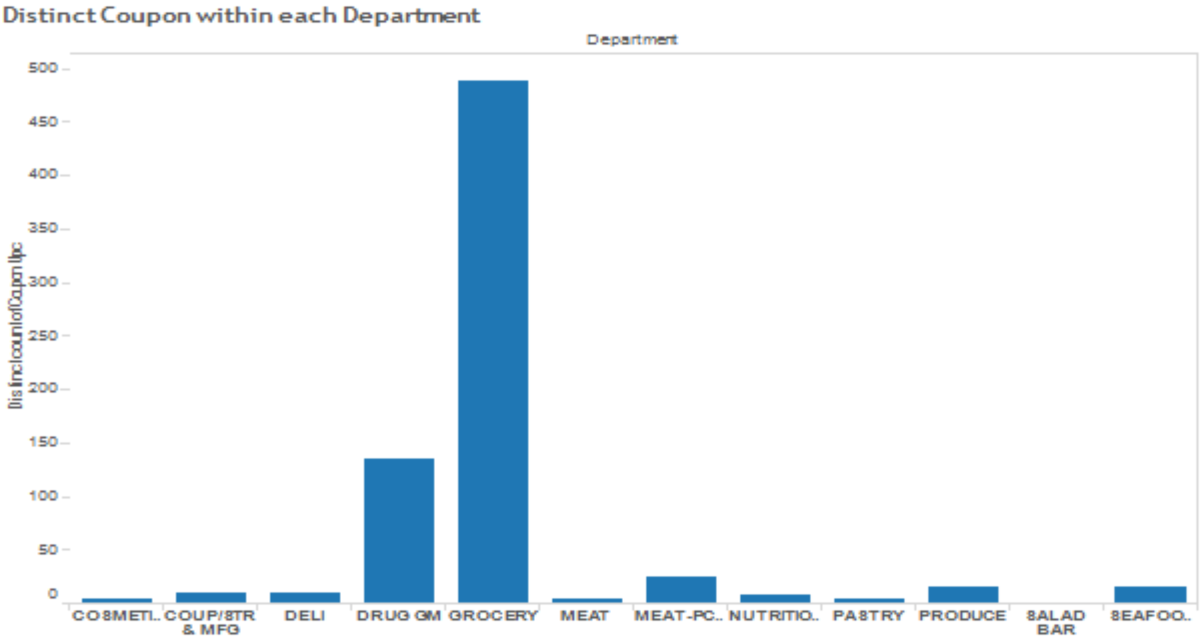
C.10. Average Transaction Amount by Customer Age



C.11. Hourly Purchases of Beer and Orange Juice by Customer Age

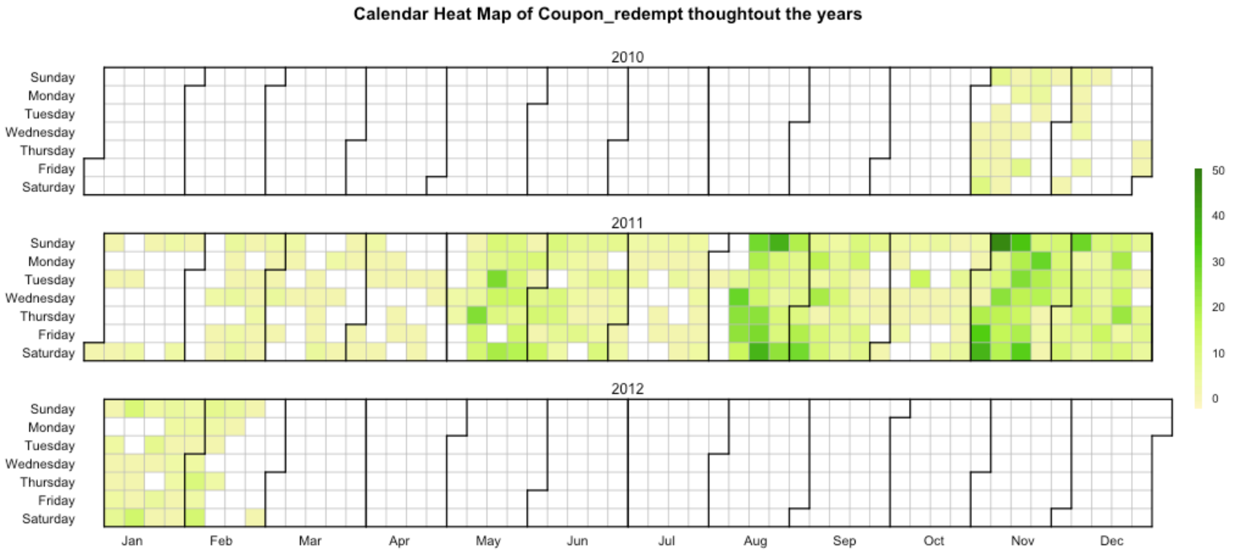


C.12. Number of Distinct Coupons Offered by Department

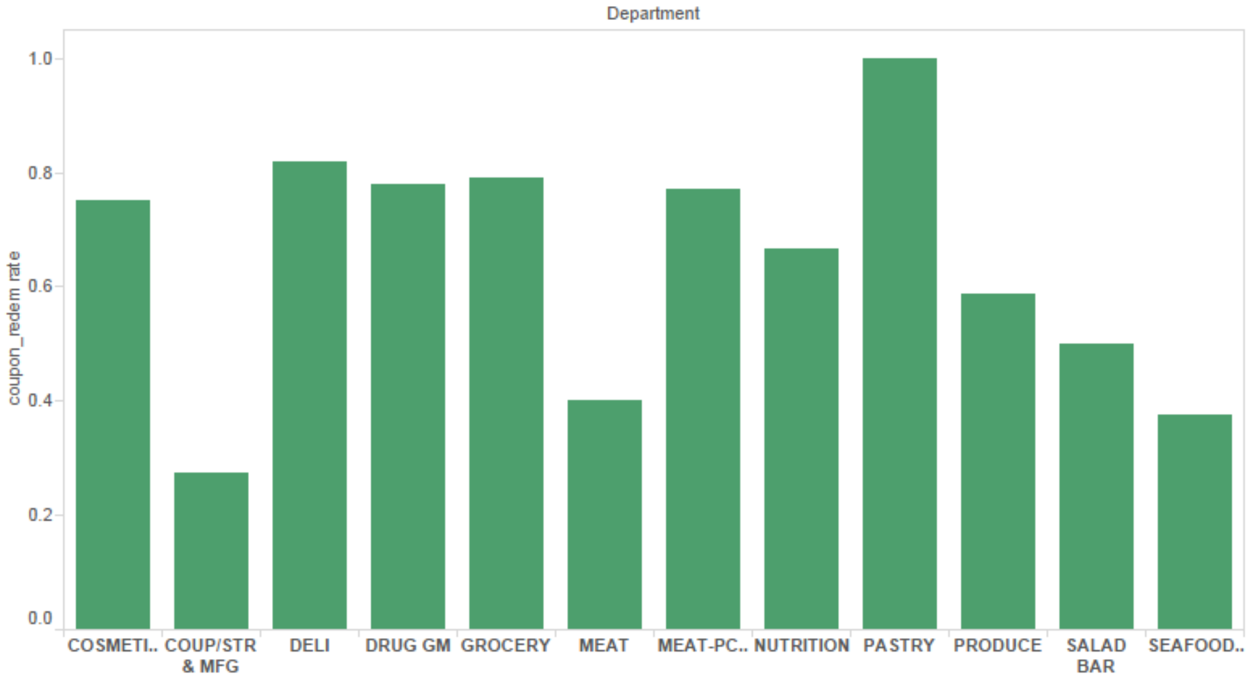

Distinct Coupon within each Department

Distinct count of Coupon Upc for each Department.

C.13. Coupon Redemption Rate


Calendar Heat Map of Coupon_redempt thoughtout the years

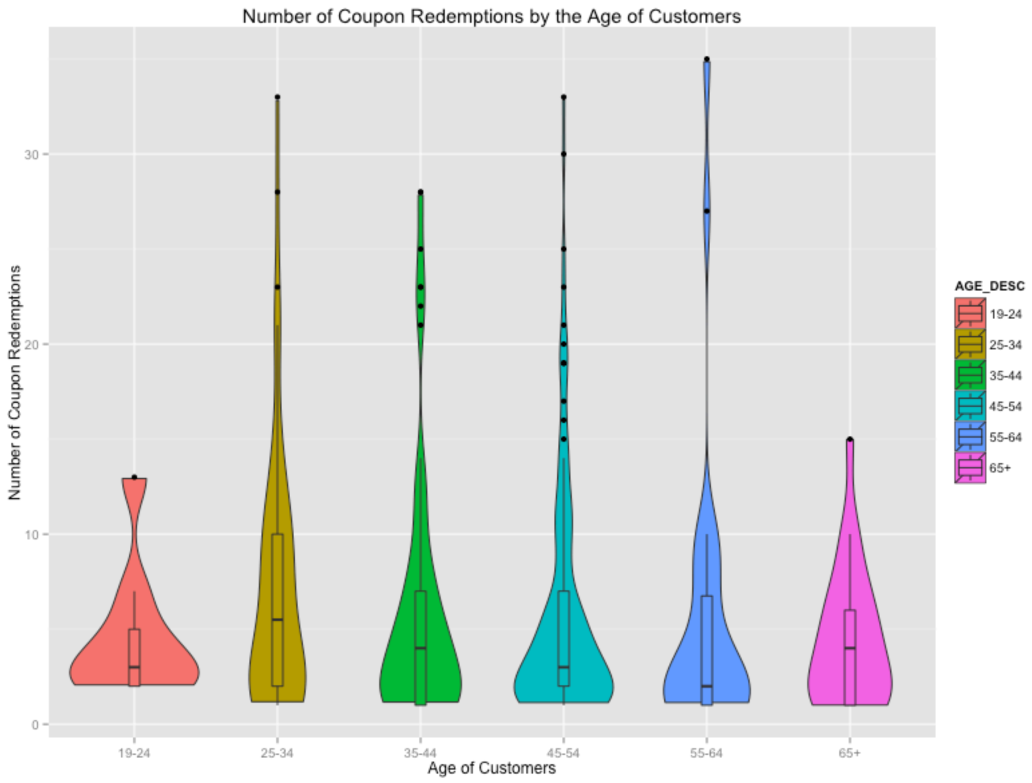
C.14. Distinct Coupon Redemption Rate by Department



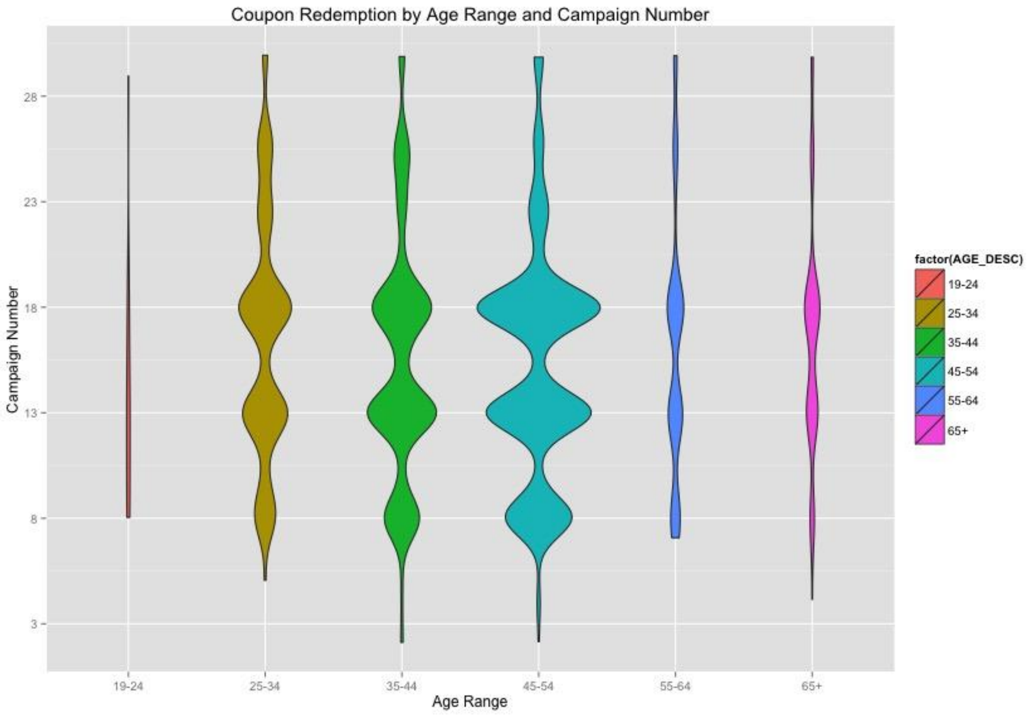
Distinct Coupon Redemption Rate

Coupon_redem rate for each Department. The view is filtered on Department, which keeps 44 of 44 members.

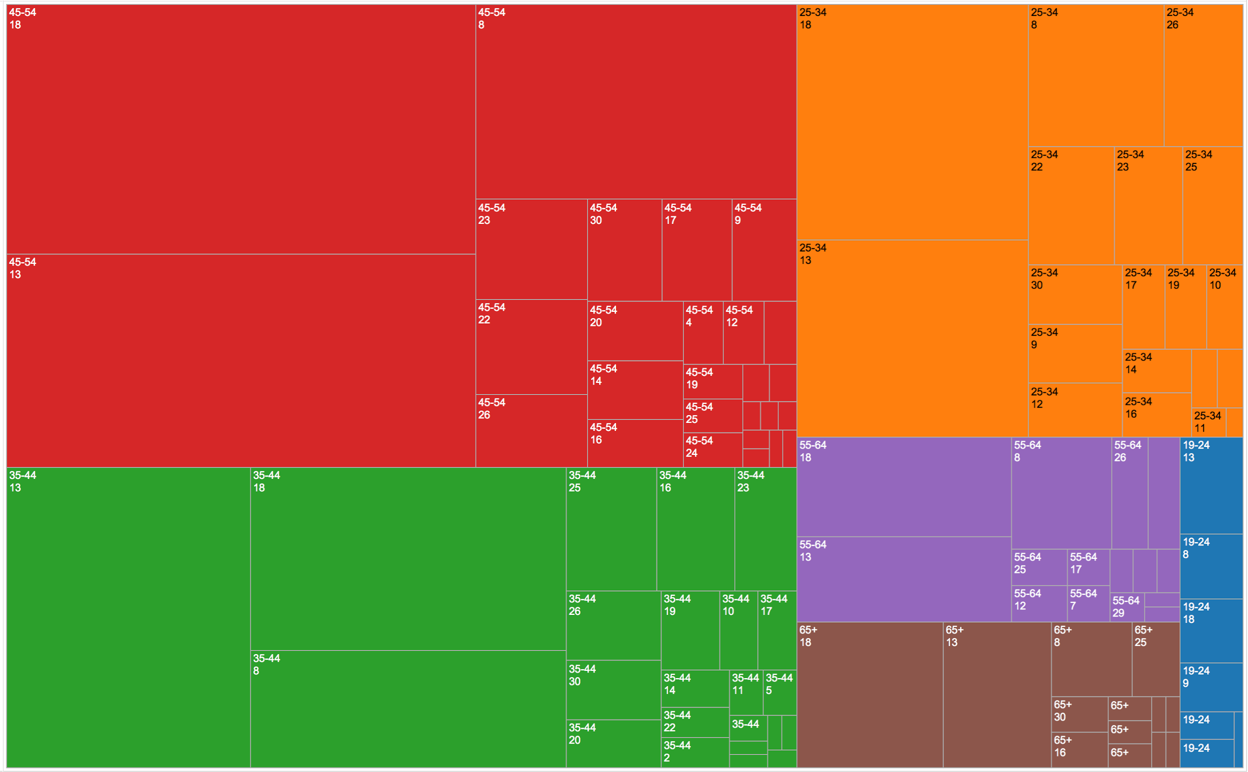
C.15. Coupon Redemption Rate by Customer Age



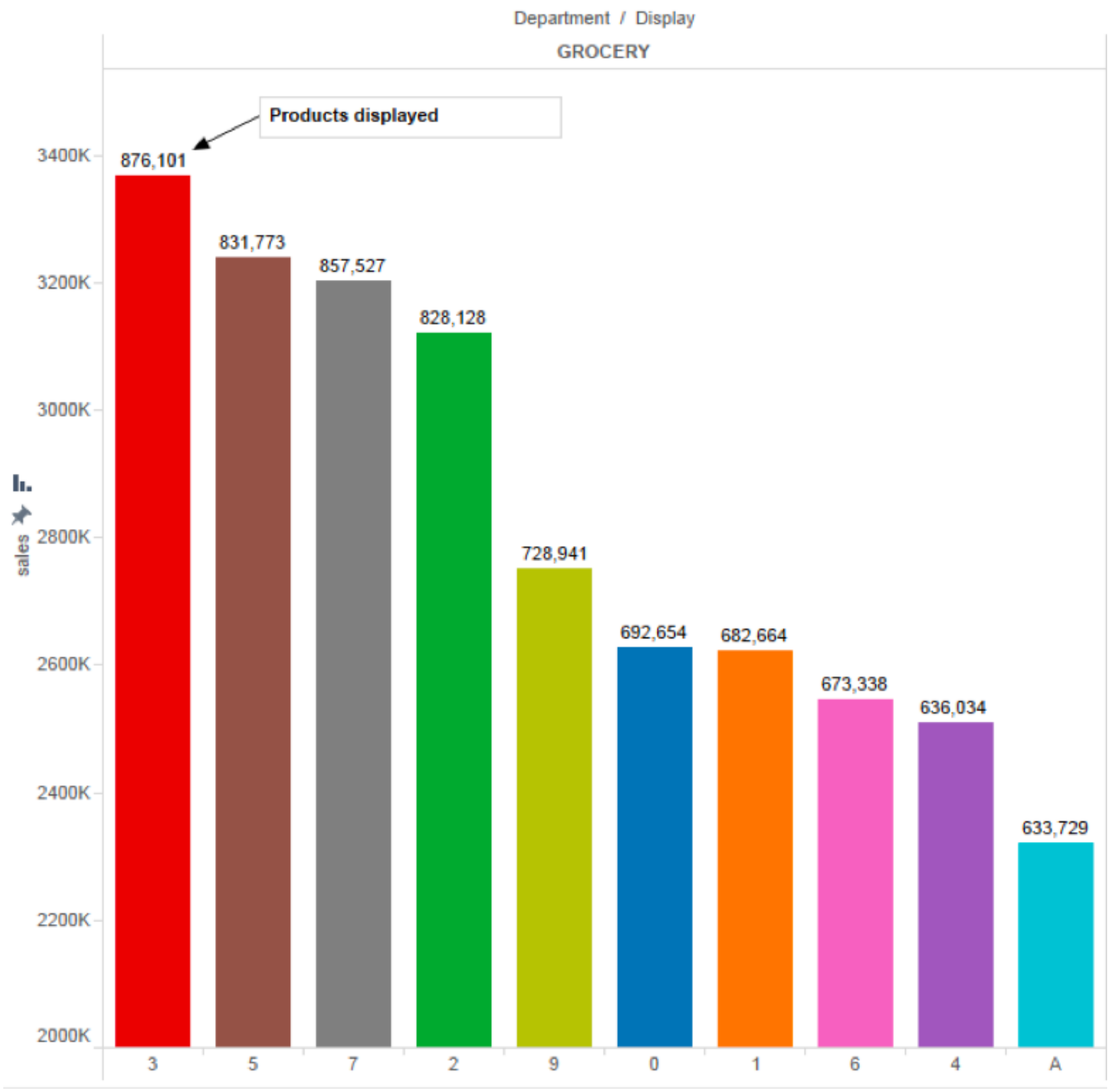
C.16. Coupon Redemption Rate by Customer Age and Campaign Number (Violin Plot)



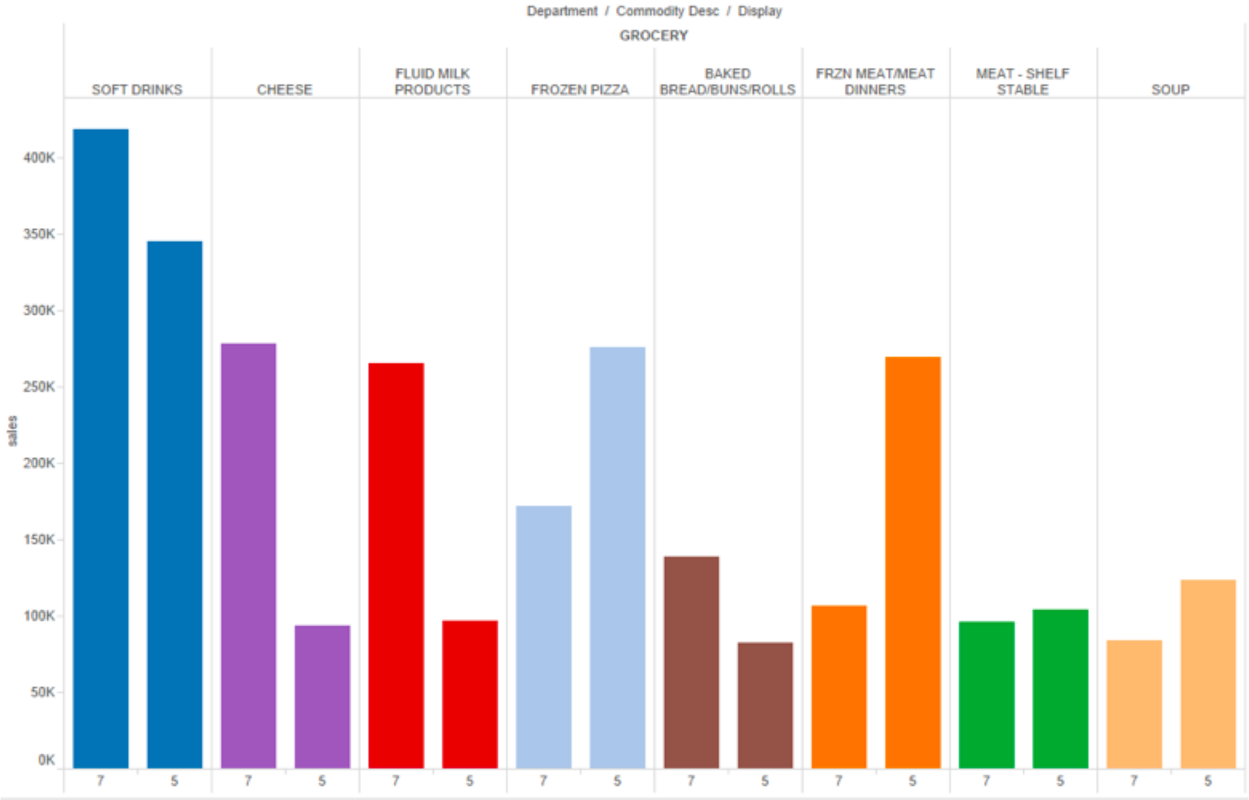
C.17. Coupon Redemption Rate by Customer Age and Campaign Number (Tree Map)



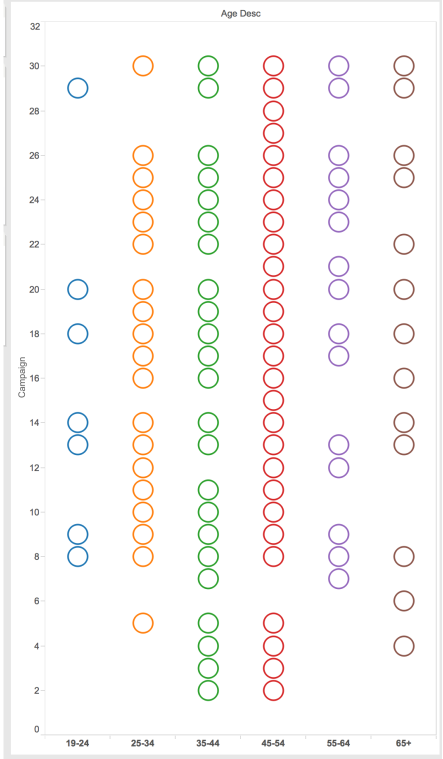
C.18. Year 1 products sales in the grocery department with different display location



C.19. Top-5 commodity sales with display location 5 compared with location 7 in year 1
By comparing the two specific display locations, we see there is room for improvement in sales just by changing the location of different products.

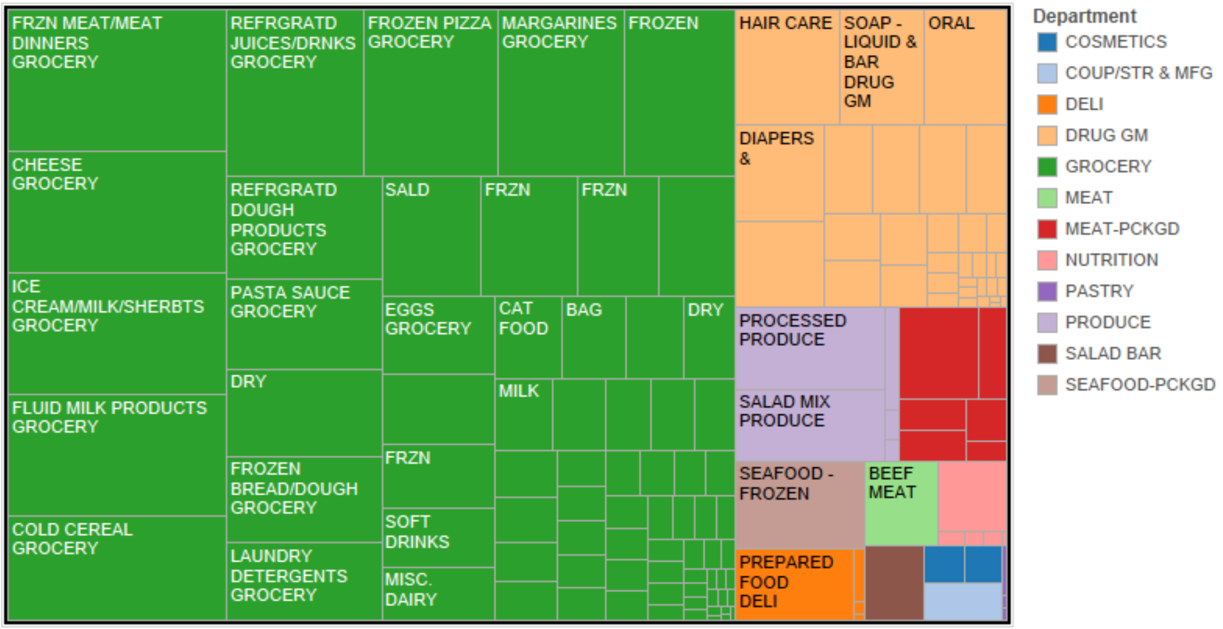


C.20. Coupon Redemption Rate by Customer Age and Campaign Number



C.21 Coupon Redemption Tree Map

Coupon Redemption Tree Map



Commodity Desc and Department. Color shows details about Department. Size shows count of Coupon Upc. The marks are labeled by Commodity Desc and Department. The view is filtered on Department, which keeps 44 of 44 members.

Appendix D: Visualization Source Code

R Source Code - Calendar Heat Map

```
library(plyr)
setwd("~/Dropbox/CSC465/CSC465-Project/")
raw_transaction = read.table("./Dataset/transaction_data.csv", sep=",", header=T ,na.strings = " ")

#remove NA
raw_transaction_nNA <- raw_transaction[complete.cases(raw_transaction),]

# the calendar heat for sales value
DateVol <- raw_transaction[,c("DAY","SALES_VALUE")]
DateVol_agg <- ddply(DateVol,.(DAY),numcolwise(sum))
DateVol_agg$Date <- as.Date(DateVol_agg$DAY -1, origin = "2010-03-24")

p1 <- calendarHeat(DateVol_agg$Date, DateVol_agg$SALES_VALUE, varname="Sales Value", color = 'r2b')
p1

dev.copy(device = png, filename = 'CalendarHeat-Sales.png', width = 1024, height = 768)
dev.off()

#####################################

# the calendar heat for number of coupon_redempt thoughtout the years
raw_coupon_redempt = read.table("./Dataset/coupon_redempt.csv", sep=",", header=T ,na.strings = " ")
raw_coupon_redempt <- cbind(raw_coupon_redempt,1)
colnames(raw_coupon_redempt)[5] <- "nCoupon"
library(plyr)
coupon_redempt_agg <- ddply(raw_coupon_redempt,.(DAY),numcolwise(sum))
coupon_redempt_agg$Date <- as.Date(coupon_redempt_agg$DAY -1, origin = "2010-03-24")

p2 <- calendarHeat(coupon_redempt_agg$Date, coupon_redempt_agg$nCoupon, varname="Coupon_redempt thoughtout the
years", color = 'w2g')
```

```
p2

dev.copy(device = png, filename = 'CalendarHeat-Coupon_redempt.png', width = 1024, height = 768)
dev.off()



##############################################################################
#                      Calendar Heatmap                                #
#                              by                                      #
#                      Paul Bleicher                                   #
# an R version of a graphic from:                                      #
# http://stat-computing.org/dataexpo/2009/posters/wicklin-allison.pdf         #
#  requires lattice, chron, grid packages                              #
##############################################################################

## calendarHeat: An R function to display time-series data as a calendar heatmap
## Copyright 2009 Humedica. All rights reserved.

## This program is free software; you can redistribute it and/or modify
## it under the terms of the GNU General Public License as published by
## the Free Software Foundation; either version 2 of the License, or
## (at your option) any later version.

## This program is distributed in the hope that it will be useful,
## but WITHOUT ANY WARRANTY; without even the implied warranty of
## MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
## GNU General Public License for more details.

## You can find a copy of the GNU General Public License, Version 2 at:
## http://www.gnu.org/licenses/gpl-2.0.html

calendarHeat <- function(dates,
                         values,
                         ncolors=99,
                         color="r2g",
                         varname="Values",
                         date.form = "%Y-%m-%d", ...) {
  require(lattice)
  require(grid)
  require(chron)
  if (class(dates) == "character" | class(dates) == "factor" ) {
    dates <- strptime(dates, date.form)
  }
  caldat <- data.frame(value = values, dates = dates)
  min.date <- as.Date(paste(format(min(dates), "%Y"),
                          "-1-1",sep = ""))
  max.date <- as.Date(paste(format(max(dates), "%Y"),
                          "-12-31", sep = ""))
  dates.f <- data.frame(date.seq = seq(min.date, max.date, by="days"))

  # Merge moves data by one day, avoid
  caldat <- data.frame(date.seq = seq(min.date, max.date, by="days"), value = NA)
  dates <- as.Date(dates)
  caldat$value[match(dates, caldat$date.seq)] <- values

  caldat$dotw <- as.numeric(format(caldat$date.seq, "%w"))
  caldat$woty <- as.numeric(format(caldat$date.seq, "%U")) + 1
  caldat$yr <- as.factor(format(caldat$date.seq, "%Y"))
  caldat$month <- as.numeric(format(caldat$date.seq, "%m"))
  yrs <- as.character(unique(caldat$yr))
  d.loc <- as.numeric()
  for (m in min(yrs):max(yrs)) {
    d.subset <- which(caldat$yr == m)
    sub.seq <- seq(1,length(d.subset))
    d.loc <- c(d.loc, sub.seq)
  }
  caldat <- cbind(caldat, seq=d.loc)
```

```
#color styles
r2b <- c("#0571B0", "#92C5DE", "#F7F7F7", "#F4A582", "#CA0020") #red to blue
r2g <- c("#D61818", "#FFAE63", "#FFFFBD", "#B5E384")   #red to green
w2b <- c("#045A8D", "#2B8CBE", "#74A9CF", "#BDC9E1", "#F1EEF6")   #white to blue
w2g <- c("#FFF8C6", "#CCFB5D", "#52D017", "#347C17")   #white to green

assign("col.sty", get(color))
calendar.pal <- colorRampPalette((col.sty), space = "Lab")
def.theme <- lattice.getOption("default.theme")
cal.theme <-
  function() {
    theme <-
      list(
        strip.background = list(col = "transparent"),
        strip.border = list(col = "transparent"),
        axis.line = list(col="transparent"),
        par.strip.text=list(cex=0.8))
  }
lattice.options(default.theme = cal.theme)
yrs <- (unique(caldat$yr))
nyr <- length(yrs)
print(cal.plot <- levelplot(value~woty*dotw | yr, data=caldat,
                            as.table=TRUE,
                            aspect=.12,
                            layout = c(1, nyr%%7),
                            between = list(x=0, y=c(1,1)),
                            strip=TRUE,
                            main = paste("Calendar Heat Map of ", varname, sep = ""),
                            scales = list(
                              x = list(
                                at= c(seq(2.9, 52, by=4.42)),
                                labels = month.abb,
                                alternating = c(1, rep(0, (nyr-1))),
                                tck=0,
                                cex = 0.9),
                              y=list(
                                at = c(0, 1, 2, 3, 4, 5, 6),
                                labels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",
                                           "Friday", "Saturday"),
                                alternating = 1,
                                cex = 0.9,
                                tck=0)),
                            xlim =c(0.4, 54.6),
                            ylim=c(6.6,-0.6),
                            cuts= ncolors - 1,
                            col.regions = (calendar.pal(ncolors)),
                            xlab="" ,
                            ylab="",
                            colorkey= list(col = calendar.pal(ncolors), width = 0.6, height = 0.5),
                            subscripts=TRUE
) )
panel.locs <- trellis.currentLayout()
for (row in 1:nrow(panel.locs)) {
  for (column in 1:ncol(panel.locs))  {
    if (panel.locs[row, column] > 0)
    {
      trellis.focus("panel", row = row, column = column,
                    highlight = FALSE)
      xyetc <- trellis.panelArgs()
      subs <- caldat[xyetc$subscripts,]
      dates.fsubs <- caldat[caldat$yr == unique(subs$yr),]
      y.start <- dates.fsubs$dotw[1]
      y.end   <- dates.fsubs$dotw[nrow(dates.fsubs)]
      dates.len <- nrow(dates.fsubs)
      adj.start <- dates.fsubs$woty[1]

      for (k in 0:6) {
        if (k < y.start) {
          x.start <- adj.start + 0.5
        } else {
          x.start <- adj.start - 0.5
        }
        if (k > y.end) {
          x.finis <- dates.fsubs$woty[nrow(dates.fsubs)] - 0.5
        } else {
          x.finis <- dates.fsubs$woty[nrow(dates.fsubs)] + 0.5
        }
        grid.lines(x = c(x.start, x.finis), y = c(k -0.5, k - 0.5),
                  default.units = "native", gp=gpar(col = "grey", lwd = 1))
      }
      if (adj.start <  2) {
        grid.lines(x = c( 0.5,  0.5), y = c(6.5, y.start-0.5),
                  default.units = "native", gp=gpar(col = "grey", lwd = 1))
        grid.lines(x = c(1.5, 1.5), y = c(6.5, -0.5), default.units = "native",
                  gp=gpar(col = "grey", lwd = 1))
        grid.lines(x = c(x.finis, x.finis),
                  y = c(dates.fsubs$dotw[dates.len] -0.5, -0.5), default.units = "native",
                  gp=gpar(col = "grey", lwd = 1))
        if (dates.fsubs$dotw[dates.len] != 6) {
          grid.lines(x = c(x.finis + 1, x.finis + 1),
                    y = c(dates.fsubs$dotw[dates.len] -0.5, -0.5), default.units = "native",
                    gp=gpar(col = "grey", lwd = 1))
        }
        grid.lines(x = c(x.finis, x.finis),
                  y = c(dates.fsubs$dotw[dates.len] -0.5, -0.5), default.units = "native",
                  gp=gpar(col = "grey", lwd = 1))
      }
      for (n in 1:51) {
        grid.lines(x = c(n + 1.5, n + 1.5),
                  y = c(-0.5, 6.5), default.units = "native", gp=gpar(col = "grey", lwd = 1))
      }
      x.start <- adj.start - 0.5

      if (y.start > 0) {
        grid.lines(x = c(x.start, x.start + 1),
                  y = c(y.start - 0.5, y.start -  0.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
        grid.lines(x = c(x.start + 1, x.start + 1),
                  y = c(y.start - 0.5 , -0.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
        grid.lines(x = c(x.start, x.start),
                  y = c(y.start - 0.5, 6.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
        if (y.end < 6  ) {
          grid.lines(x = c(x.start + 1, x.finis + 1),
                    y = c(-0.5, -0.5), default.units = "native",
                    gp=gpar(col = "black", lwd = 1.75))
          grid.lines(x = c(x.start, x.finis),
                    y = c(6.5, 6.5), default.units = "native",
                    gp=gpar(col = "black", lwd = 1.75))
        } else {
          grid.lines(x = c(x.start + 1, x.finis),
                    y = c(-0.5, -0.5), default.units = "native",
                    gp=gpar(col = "black", lwd = 1.75))
          grid.lines(x = c(x.start, x.finis),
                    y = c(6.5, 6.5), default.units = "native",
                    gp=gpar(col = "black", lwd = 1.75))
        }
      } else {
        grid.lines(x = c(x.start, x.start),
                  y = c( - 0.5, 6.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
      }

      if (y.start == 0 ) {
        if (y.end < 6  ) {
          grid.lines(x = c(x.start, x.finis + 1),
                    y = c(-0.5, -0.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
            grid.lines(x = c(x.start, x.finis),
                  y = c(6.5, 6.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
           } else {
            grid.lines(x = c(x.start + 1, x.finis),
                  y = c(-0.5, -0.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
            grid.lines(x = c(x.start, x.finis),
                  y = c(6.5, 6.5), default.units = "native",
                  gp=gpar(col = "black", lwd = 1.75))
           }
          }
          for (j in 1:12)  {
           last.month <- max(dates.fsubs$seq[dates.fsubs$month == j])
           x.last.m <- dates.fsubs$woty[last.month] + 0.5
           y.last.m <- dates.fsubs$dotw[last.month] + 0.5
           grid.lines(x = c(x.last.m, x.last.m), y = c(-0.5, y.last.m),
                 default.units = "native", gp=gpar(col = "black", lwd = 1.75))
           if ((y.last.m) < 6) {
            grid.lines(x = c(x.last.m, x.last.m - 1), y = c(y.last.m, y.last.m),
                  default.units = "native", gp=gpar(col = "black", lwd = 1.75))
            grid.lines(x = c(x.last.m - 1, x.last.m - 1), y = c(y.last.m, 6.5),
                  default.units = "native", gp=gpar(col = "black", lwd = 1.75))
           } else {
            grid.lines(x = c(x.last.m, x.last.m), y = c(- 0.5, 6.5),
                  default.units = "native", gp=gpar(col = "black", lwd = 1.75))
           }
          }
         }
        }
        trellis.unfocus()
       }
       lattice.options(default.theme = def.theme)
      }
```

R Source Code - Network Graph For Product Departments

```
library(igraph)
setwd("~/Dropbox/CSC465/CSC465-Project/")

raw_transaction = read.table("./Dataset/transaction_data.csv", sep=",", header=T ,na.strings = " ")
raw_product = read.table("./Dataset/product.csv", sep=",", header=T, na.strings = " ")

raw_total <- merge(raw_transaction,raw_product,by="PRODUCT_ID", all.x=TRUE)

#remove NA
raw_total_nNA <- raw_total[complete.cases(raw_total),]

head(raw_transaction)
head(raw_product)


raw_Day1 = raw_total_nNA[raw_total_nNA$DAY == 2, ]
raw_Day1

edges = data.frame(p0=rep(0, 10000), p1=rep(0, 10000))

nEdges = 0
for (household in unique(raw_Day1$household_key))
{
 hPurchases = raw_Day1[raw_Day1$household_key == household, ]
 for (i in 1:nrow(hPurchases))
 {
  row = hPurchases[i, ]
  row
  hPurchases$DEPARTMENT[i]
  prodID = hPurchases$DEPARTMENT[i]
  for (j in i:nrow(hPurchases))
  {
   prodID2 = hPurchases$DEPARTMENT[j]
   if (prodID != prodID2)
   {
    nEdges = nEdges + 1
    edges$p0[nEdges] = as.character(prodID)
    edges$p1[nEdges] = as.character(prodID2)
   }
  }
 }
}
edges = edges[1:nEdges, ]
#####################
g = graph.data.frame(edges, directed=T)
g = simplify(g, remove.multiple = T)
#V(g)$color <- sample(rainbow(7, alpha=1))
V(g)$number <- sample(1:50, vcount(g), replace=TRUE)

plot(g,  layout=layout.fruchterman.reingold,            # the layout method. see the igraph documentation for details
     main='Network Graph',       #specifies the title
     vertex.label.dist=0.5,                             #puts the name labels slightly off the dots
     vertex.frame.color='blue',                #the color of the border of the dots
     vertex.label.color='black',               #the color of the name labels
     vertex.label.font=2,                               #the font of the name labels
     vertex.label=V(g)$name,                   #specifies the lables of the vertices. in this case the 'name' attribute is used
     edge.width=edge.betweenness(g),
     edge.arrow.size=0.3,
     vertex.color=V(g)$color,
     vertex.label.cex=1)

dev.copy(device = png, filename = 'NetworkGrp-ProductDep-D.png', width = 1024, height = 768)
dev.off()
```