April Ye, Kelly Yu, and Kenny Zhou

INFO 3300 Project 1: Final Write-Up

Black Friday Consumer Trends:

Data Description:
The first Black Friday dataset was found on kaggle.com (https://www.kaggle.com/mehdidag/black-friday) and is a collection of 550,000 transactions at a retail store on Black Friday, the day after Thanksgiving notoriously known in America as the "best" day to go shopping, as the sales are often emphasized. We decided to utilize the gender, age group, years lived in the city (of the retail store), and the amount spent on purchase variables to create our two of our data visualizations. Our second data visualization was found on finder.com (https://www.finder.com/black-friday-statistics). This dataset had three variables: amount spent per purchase, household income, and likelihood of sales purchase regret (%), which were all used in our third data visualization.

Data Visualizations:
1. Percentage of Consumers per Age Group for Males and Females
Design Rationale: This data visualization broke down the percentages of consumers per age group for males and females. Using the gender and age group variables, we were able to iteratively summate the population count per age group per gender. To best portray this data, we wanted to compare the percentage differences between males and females side by side. We decided that using a gradient colored bar, in which color blocks' heights were scaled to the percentage of their respective age group in the total male/female population , would be the most effective way to visualize the percentage data of the population. We made the gradients with color families indicative of the corresponding gender: blue for males and red/pink for females. To emphasize which data corresponds to which gender, and also to introduce an element of creativity, we also decided to include visual representations of cartoon male and female figures next to their respective bar graph. The bar graphs were shifted towards the center of the page to promote ease for comparison of the graphs' values.

2. Total Amount Spent on Purchases per Age Group vs. Years Lived in City:
Design Rationale: For this data visualization, we looked at the total amount spent on purchases, age group, and years lived in city variables. We decided to make a grouped bar chart based on age group to graph the total amount spent on purchases against the years lived in the city of the retail store. We utilized bars as the visual marks and color and position as the visual channels. Using color as a visual channel, we grouped the different age groups by different colors to make it clear to the audience what each bar represents with the help of the legend. We scaled the color

scale ordinally, the amount spent on purchases linearly, and the age group and years lived in the city by band.

Story: We initially hypothesized that people who have not lived in the city for as long would spend more on Black Friday because they would have traveled to this city with the purpose of spending money on Black Friday. However, this did not seem to be the case, as there really was no correlation between these two variables. For the cluster of bars that lived in the city for 3 years, however, all the bars for each age group was either the highest or the second highest. It still seems as though we cannot make a sound conclusion for that. We also noticed that there seemed to be a common trend among each cluster of bars, which was that the age group of 26-35 spent the most on purchases. This age group seems to be the store’s target audience, considering the large sum of money spent on goods. This was an emphasis on the findings from our first visualization.

3. Household Income vs Amount Spent per Purchase and Likelihood of Purchase Regret
Design Rationale: This final data visualization took in a completely new dataset and showed completely different information. The dataset included 3 columns: 1 for household income, 1 for amount spent per purchase, and 1 for likelihood of regret for that purchase. We concluded that a bar graph, potentially a double bar graph, would be useful to visualize the data, in order to compare the difference in amount spent (in dollars) for each household income bracket. Using these 2 columns, we quickly put the brackets and their respective amount spent per purchase side-by-side, and then we added in another set of bars for their respective likelihood of regret. These next set of bars were placed on top of the previous set and were sized to be a proportion of the previous set, i.e. the bars were scaled using the percentage times the previous bar height. We decided this was not the best way to visualize the change in likelihood of regret from each household income bracket to each bracket. We then tried changing the bar colors depending on the likelihood of regret, scaling from yellow (no regret) to red (all regret), but since the data is so tight (ranging from 40% to 80%), the visualization did not provide a broad difference in color and change. Finally, we decided to pursue a line graph, which helped show the change of likelihood depending on the income bracket. We utilized bars, circles/points, and lines as visual marks and position as the visual channel. We scaled the amount spent and the likelihood of regret linearly and the income bracket by band.
Story: The bar graph visualization tells us that as we approach the $150,000 to $300,000 tax bracket, the amount that people spend increases, and once we hit $300,000+, the amount drops almost half. Additionally, the line graph visualization shows us that the likelihood of purchasing regret steadily increase, again, as we approach the $150,000 to $300,00 tax bracket; once we get to the $300,000+ tax bracket, the likelihood of regret decreases. We were surprised by how much the $150,000 to $300,00 tax bracket regret spending. It’s interesting how the $300,000 bracket group spent a sum significantly less than the previous bracket, as well as their lower likelihood

of regret. This potentially aligns to the notion that those who are better off financially are sometimes the product of smarter money handling.

Team Contributions:
We divided the work up by visualization, so April created the first one, Kelly the second, and Kenny the third.

April: I contributed towards the brainstorming and planning of all 3 visualizations starting from the initial sketch mockups to participating in discussion of how to better portray our data visualizations when overcoming implementation obstacles. I was responsible for the first visualization, and spent the bulk of my time designing the SVG people and figuring out how to use d3 to draw their intricate body parts (especially the paths for the hair). I started the project early on before Feb break, and continuously worked on iterations of the SVG visualizations (and how to best visualize our data using these SVG people) throughout the days leading up to the project deadline. My main obstacle was figuring out how to append existing SVG's to another SVG (which was an obstacle during attempts to iterate the figures) and finding tweaks to bypass the limitations of d3.append(). This entire process, coupled with implementing the two rectangle bar visualizations, took 20+ hours.

Kelly: I took part in the design of all three visualizations by offering my input and making necessary revisions. I also met with our mentor TA to discuss our ideas and ask any questions we had. My main contribution was coding up the entirety of the second data visualization, the grouped bar chart. I spent the majority of my time trying to create a new data array from the first dataset and drawing the actual bars. I spent upwards of 20 hours coding this visualization because my initial data array that I created had nested objects within nested objects. I decided to use that initial data array to create a second data array that parsed through every object, which was much easier to understand what was going on.

Kenny: I was a part of brainstorming potential datasets to visualize. I looked into datasets that aligned with my interests, one in particular being the Spotify dataset, since I enjoy music. Additionally, I participated in sketching out potential visualizations of that dataset. My main contribution was third section/final data visualization with the bar and line graphs. From adding an additional y-axis to creating an effective line graph, I spent nearly 16 hours planning and coding. Fortunately, my dataset was easier to navigate compared to the previous Kaggle dataset we used. Additionally, utilizing the forEach loop, experimenting with different conditions, and trying different visualizations took many hours of debugging, testing, and coding. I enjoyed making the graph more "traditionally" complete by adding rotated axis labels, which also took additional time.