My purchases in S Group companies – Who am I?

Explorative Information Visualization

Kristel Ariko

December 16, 2018

Contents

1 Introduction 2
2 Dataset 2
3 Design 3
3.1 Theme 3
3.2 Time 3
3.3 Space 4
3.4 Visual Encoding 4
3.5 Draft of the Visualization 5
3.6 Draft of the Interactions 6
4 Implementation 6
4.1 Tools 6
4.2 Visualization 6
4.3 Colorblindness 7
4.4 Interactions 8
5 Conclusions 8
5.1 Problems and Improvements 9
5.2 Insights 9
References 10
Attachment 1: Evaluating the colors of the draft visualization 11
Attachment 2: Evaluating the colors of the final visualization 12

1 Introduction

The past 10 years I have had a discount card for S Group companies and during the recent years I have concentrated my purchases to companies which belong to the S Group to gain as much bonus points (€) as possible. The bonus points and transactions of the purchases within last 12 months can be seen in the S Group's Website. The intention is to overview and examine the content of the purchases by summarizing the distribution of the bought ingredients within specific time. Furthermore, using the distribution to recognize which ingredients rule my life and how is time involved in these (e.g. some ingredients are bought only during summer).

2 Dataset

The dataset was fetched from S Group's Website by polling its API interface. The fetched data has about 500 distinct ingredients whereas the name of the ingredient may include 1-5 Finnish words in any order, hence naming does not follow any common rule. To use this dataset, the ingredients needed a hierarchical grouping. A natural approach would be to find an automatic method to do this. It turns out that the task is not as easy as it sounds. Let me introduce few problematic cases:

- "*Apple juice 1l*" is obviously more likely a "juice" than a "fruit"
- Unnecessary words ("20%","violet","kg",...),
- Words have been cut from the end
- Abbreviations
- Manufacturers
- Adjectives

Figure 1. The names of the ingredients presented as word cloud [1].

These problematic cases also pop out from the figure 1 (sorry about the Finnish words). Therefore, within this time limit, I decided to do the hierarchical grouping manually.

The ingredients of each purchase are grouped as following procedure: (1) map all distinct ingredients to a group which best describes the ingredient, (2) map all the groups in previous part to group which best describes the group, (3) repeat previous step until convergence, (4) use the groups of the distinct ingredients to assign the groups for all ingredients. As a result, a hierarchical grouping of the ingredients has been obtained.

Finally, after grouping and other minor modifications, the dataset ended up as table presentation (table 1) whereas the rows are the ingredients (1447 as total) and the columns are the attributes.

Table 1. Dataset in table format.

Name	Year	Month	Date	Hours	Minutes	Cost	Pieces	Group1	Group2	...	Group6	Store
apple juice 1l	2018	3	2	12	24	1.23	1	Liquid	Juice			Prisma Espoo
...												

3 Design

This chapter overviews the process of the design. Starting from the themes that appear around the topic and exploring how time and space are related to the topic. Furthermore, design choices are described in terms of visual encoding until suggesting the first draft of the visualization and its interactions.

3.1 Theme

There are several themes that can be constructed around the content of the purchases. The most obvious themes are picked and introduced as next. Recall that each purchase includes at least one ingredient, whereas ingredient is assigned to specific (1) hierarchical group, (2) store and (3) time. The ingredients which are mapped to same hierarchical group can be used to identify similar ingredients. The availability of the ingredients can be distinguished by the places of the purchases, hence stores. Although, the difference between S Group stores may be quite small. The ingredients bought in specific time period, describes the overall consumption (€ or pieces). In addition, the consumption shows which ingredients rule my life and which are bought more during specific periods, e.g. summer period. The mind map below tries to capture among the others the themes already described above.

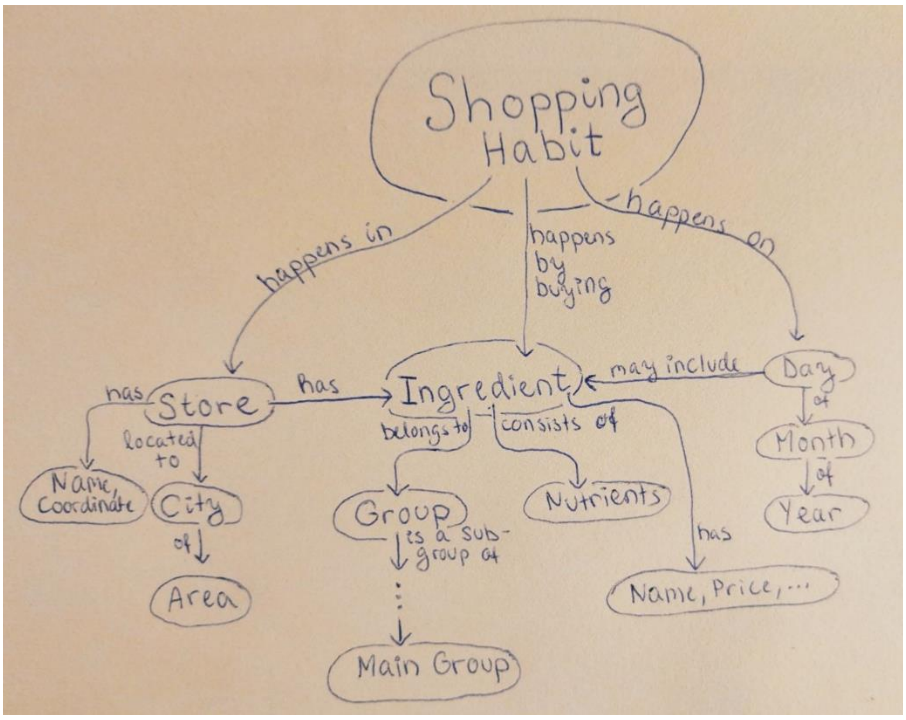


Figure 2. Mind map of the content of the purchases.

This work concentrates on combining themes (1) and (3) using the quantitative value of the ingredient, either € or pieces.

3.2 Time

The time defines when the purchasing happened and as stated in the previous section, this can be used to aggregate consumption within specific time period. In this work, the consumption has been divided to 12 time periods, where each period corresponds to one of the months from the last 12 months. The month presentation was chosen because daily and weekly consumptions do not capture the distribution well. For example, there may be unbalanced weeks and there are probably at most half of the days in a year used to go to stores.

In addition, this work uses time to animate the distribution of each month starting from the oldest month and changing the content according to the next month after specific time. This is used to capture the overall consumption of the distribution and to explore the changes that pop out from the months.

Alternatively, the following time related approaches could be considered to explore the distribution of the purchases:

- Grid calendar view for daily consumption
- Linear chronological multiple timeline view which shows the consumption of the hierarchical groups using month as a time period. On demand, the group could be expanded to smaller groups. Several different views were considered, such as EventDrops [4], The Russian State Duma [5] and NBA Defensive Skills [6], all of them were left out because of the huge amount of the distinct ingredients.

3.3 Space

The space defines the location of the purchases, hence stores. This information combined with the time, could be visualized on a map as dots whereas the size of the dot represents the quantity of the purchases within specific time period. In addition, animation could be included in the same way as in the previous section.

Another alternative map solution is to animate only the occurrences of the visited stores (leaving out the content of the purchases). Starting from time when the store is visited, a large dot appears and starts to mimic a heartbeat. After certain time, the dot will get smaller, colored with grey/black and stops mimicking the heartbeat. If the dots are clearly located to different city, each time when a store is visited, the center point of the visited stores in cities can be computed. This means that currently visited city can be highlighted even though no shopping events are happened. And therefore, once a store from another city is visited, a travelling path can be shown. The animation of the path between two cities, visited stores and current city is managed using time. Even the size or the color of the objects will be changed by keeping track of time.

This work does not implement map because there were only two different cities and six different stores visited within last 12 months. Therefore, the hierarchical grouping can be considered as an aspect of location – ingredients are located to specific group.

3.4 Visual Encoding

Let's recall the task of the visualization - the visualization needs to summarize the distribution of the purchases within specific time using hierarchical groups. Having closer look to attributes, we can notice that there are 3 types of attributes: quantitative, ordinal and nominal. The quantitative values, *Cost* and *Pieces*, are presented by numeric values. The ordering of the ingredients is presented by hierarchical grouping whereas *Group1* is a higher-level group, *Group2* is a subgroup of *Group1* and so on, until ending up to the ingredient (*Name*) itself. In addition, these groups are also nominal, separable by the name of the group.

To summarize the distribution, a point marker is chosen whereas each point presents one group aggregated in specific time. The size and color of the points are assigned with the respect of the human perceptual capabilities. The Mackinlay's ranking of perceptual task was taken as a guideline (figure 3). The guideline describes how accurately human can perform a visual

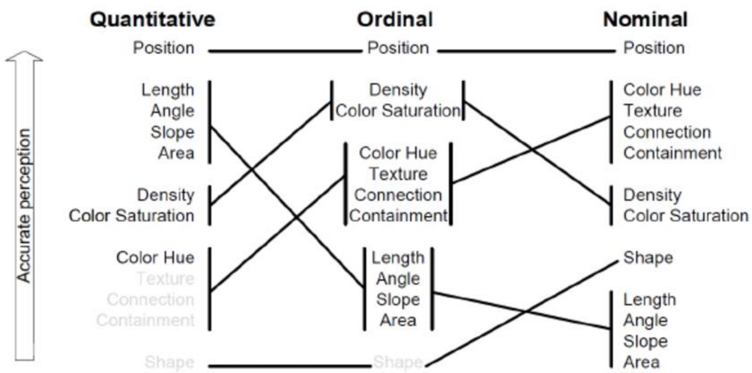


Figure 3. Mackinlay's ranking of perceptual task [2].

task if the data is quantitative, ordinal or nominal. Moreover, if the accuracy level has been decided for each perceptual task, the size and color of the data point is easy to define.

The first perceptual task is to show the quantitative of the point. The interest centers around approximate group size, therefore the area (hence size) of the point is chosen to present this information. To separate groups from each other, the accuracy is considered to be as good as possible. The second option for nominal data is chosen, hence hue coloring. An example of hue colors is shown in the figure 4. To take account of the last perceptual task which is to overview hierarchical view of the groups until ending up to ingredient itself, the position is used for ordinal data. (Although, in the final visualization this is not as accurate as promised, you will see.)

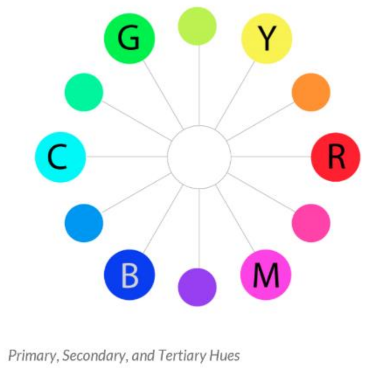


Figure 4. Hue colors [3].

All the accuracy that is missing in this section will be taken account in the interactive part of the visualization (sections 3.7 and 4.4).

3.5 Draft of the Visualization

Let me introduce the draft of the visualization (figure 5) which tries to summarize the distribution of the purchases within specific time using hierarchical grouping. The time accuracy chosen for the visualization is one month and can be selected from a slider. There are seven groups which are separated with different hue colors. The visualization starts from the middle point (not present in the draft) which combines all the groups to one group and presents the overall consumption of the month.

The circles around the middle point are growing from inside to outside. These circles present the hierarchy of the grouping where each subgroup is aligned close to the higher-level group (recall position in the ordinal). Notice also that this position is not accurate. The first circle around the middle point presents the data of the *Group1*, the second circle presents the data of the *Group2* and so on, until ending up to the ingredient itself. The groups are separated by defining a hue color to *Group1* which is inherited to its subgroups. The size of the group is defined by one of the quantities: either the total cost or pieces of the group.

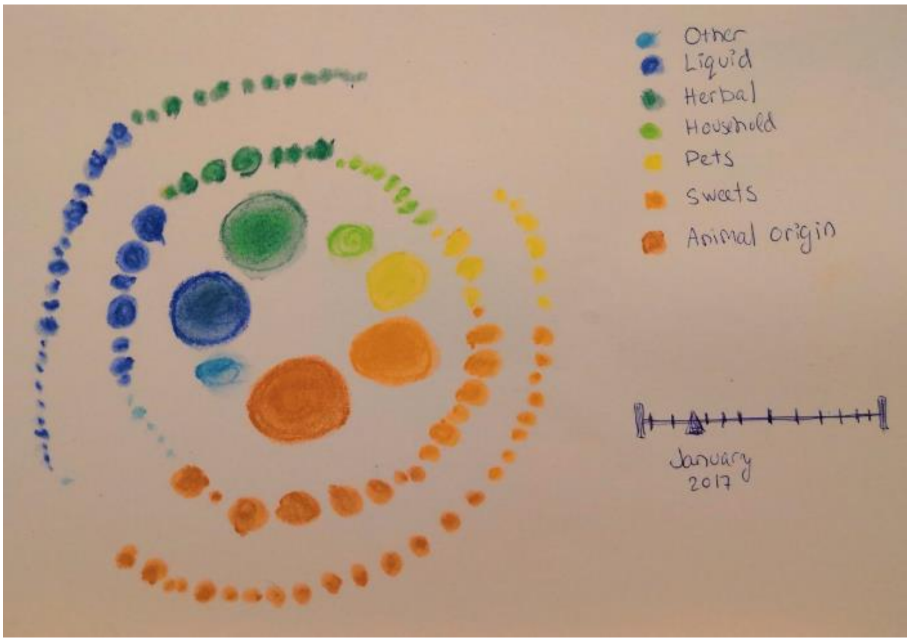


Figure 5. Draft of the visualization.

The chosen hue colors are shown in a legend. Moreover, they are presented in specific order across the visualization, whereas the moving from one color to another is visually fluent: dark green – light green – yellow – orange – brown – light blue – dark blue.

In addition, the quantitative value of the point will be shown as text in the middle of the point. This value will be only shown for the points with bigger size, hence at most in the second level of the hierarchical grouping.

3.6 Draft of the Interactions

The interactions of this visualization will give more information about the groups. If any of the groups is hovered, a summary of the group is shown. The summary includes at least the name and quantitative value of the group.

In addition, the time window of the distribution can be manipulated using the time slider shown in figure 5 (bottom-right) whereas the slider includes 12 ticks, one for each month. Furthermore, automatic time animation could be included.

4 Implementation

This chapter describes the process of the implementation. First, introducing the tools and the template of the visualization. Second, producing a visualization according to the design specified in the previous chapter. As well, justifying why some properties of the visualization were not implemented as planned. Third, improving the visualization by choosing colorblind friendly colors. In addition, introducing two different tools: *Coblis* to test different color disabilities and *ColorBrewer* to select colors from colorblind friendly palette. Fourth, describing the implemented interactions. The final visualization is available from https://letsirk.github.io/EIV/ .

4.1 Tools

JavaScript is used to collect, preprocess, aggregate and visualize the data set. The preprocessed data is turned into JSON array which is aggregated using lodash library [10]. The aggregated data is visualized using a circular layout of a scatter graph (CircosJS, based on 3D.js) [11]. The graph was modified heavily to show the data similarly as described in the previous chapter.

4.2 Visualization

The implemented visualization (figure 6) ended up almost as planned whereas some properties of the draft visualization is either left out or changed. Starting from the center of the visualization, the middle point which was reserved for describing the total consumption of the month, was left out. This decision choice simplifies the visualization and therefore, the visualization will not look crowded. As well, the huge size of the middle point would capture the viewer's attention which makes harder to notice any other differences between the distributions.

The size of the point in each group is computed according to the total pieces instead of the total cost. This decision choice supports better the goals described in the first chapter. For example, if cost is used, the information about the ingredients will be hidden behind the price and it makes harder to recognize which ingredients rule my life.

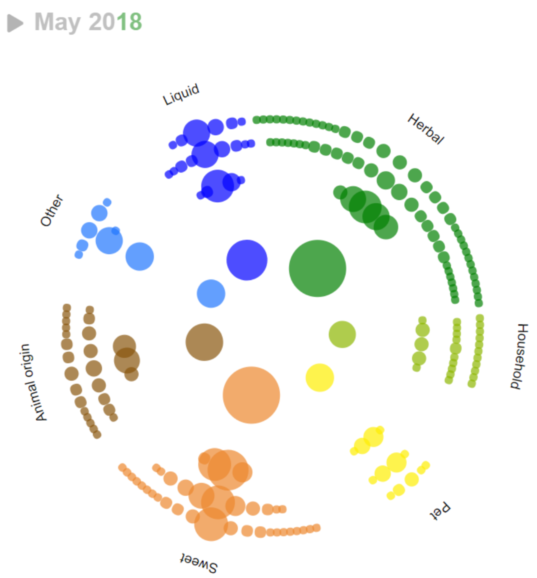


Figure 6. Implemented draft visualization.

In addition, the text in the middle of the point is not included because the same information can be seen once the viewer hovers the point. The legend of the graph is not included, instead all the labels are positioned close to the group itself. Notice that the time slider has been replaced with a text on top of the graph. This slider can be manipulated when mouse is dragged (more details in the 4.4 section).

In the next section, the colors of the visualization will be further analyzed.

4.3 Colorblindness

Almost 10% of people have a disability to see colors or their differences (figure 7). Therefore, to take account of the colorblindness, at least the problematic colors will be replaced with colors which are suitable for colorblind people.

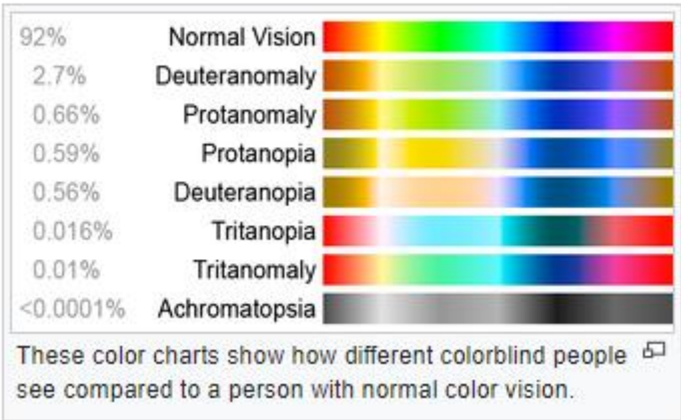


Figure 7. Normal vision versus colorblind vision [7].

To notice easily the problematic colors, a separate tool called *Coblis* (Color Blindness Simulator) was used [8]. A picture of the implemented visualization was uploaded into *Coblis* and three different colorblind (Anomalous Trichromacy, Dichromatic and Monochromatic) tests were executed. The results are indicating that the implemented visualization includes four colors which are not distinguishable by the colorblind people: green, light-green, orange and brown. The full results are shown in the *Attachment 1: Evaluating the colors of the draft visualization*.

Based on these results, the colors of the visualization should be changed. A tool called *ColorBrewer* was used to select colors from a colorblind friendly palette. Recall, that in the visual encoding (section 3.4) the hue colors were selected because the groups need to be separated as accurately as possible. Now, this assumption will not anymore hold because it is impossible to select hue colors which do not contradict with colorblind people. Therefore, two main colors were picked and extended to more colors (to cover all the groups) by changing the density and color saturation. In visual encoding point of view, this

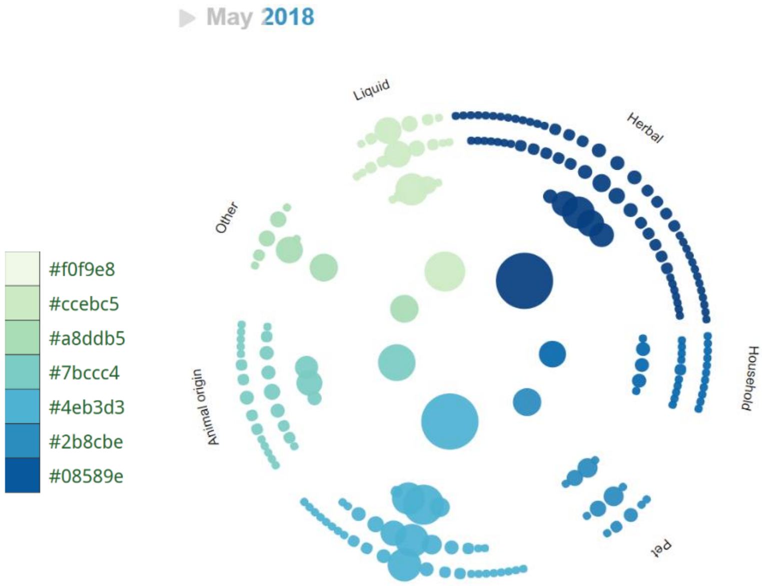


Figure 8. The colors recommended by *ColorBrewer* on the left are taken into use in the visualization on the right.

decision choice will decrease the accuracy of the nominal data (groups separable by the name of the group). The decreasing of the accuracy does not have a bad affect in the final visualization (figure 8).

Similarly, these new colors were tested in *Coblis* and better results were obtained. Only people with Monochromatic disability (<0.0001, Achromatopsia in figure 7) cannot distinguish all the groups. The full results are shown in the *Attachment 2: Evaluating the colors of the final visualization*.

4.4 Interactions

The final visualization includes three interactions: tooltip text, timeline scrolling manually and automatically. The first interaction appears if the user hovers any of the groups: the summary of the group is shown as a tooltip text. Furthermore, the user can navigate through the whole group until ending up to the ingredient itself (figure 9).

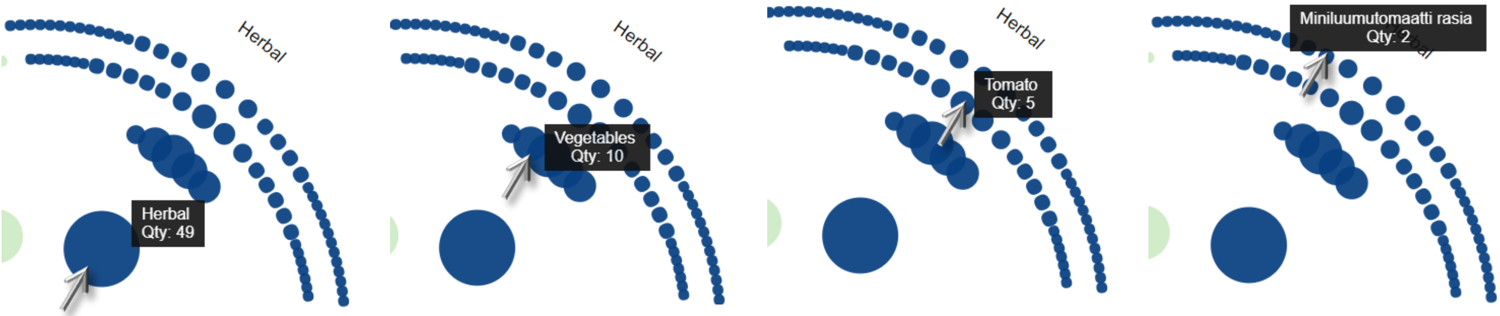


Figure 9. Navigating through the hierarchical group.

The second interaction can be activated by going on the top of the visualization (May 2018 in the figure 8), pressing down the left-button of the mouse and keeping it down while mouse is moved either to the left or to the right. If the direction is right, the timeline will be scrolled forwards (from May to June). If the direction is opposite, the timeline is scrolled backwards (from June to May). The time when the month is changed from one to another is visualized smoothly by changing the coloring of the text (May 2018): increasing the grey color and decreasing the blue color until the specific time has been elapsed (figure 10).

Figure 10. Changing from one month to next using coloring.

The last interaction can be activated by pressing a play button (triangle on the left of the May 2018 text in the figure above). If automatic timeline scrolling is active, manual scrolling is disabled and the time when the month is changed to the next one is visualized similarly as previously.

5 Conclusions

This chapter discusses about problems encountered during exploring visualization and proposes improvements to solve them. As well, overviews the insights gained from the visualization.

5.1 Problems and Improvements

The visualization does not keep the relationship between groups and their subgroups meaning that there is no trackable path from main group to ingredient. This can be improved by adding connecting lines between groups. Furthermore, then the grouping could be collapsed and expanded on demand.

In addition, the visualization does not show the progress of all the months in one view when manual or automated scrolling is used. This means that the viewer cannot know the number of all the months, the starting/ending month before scrolling will start from the beginning/end and the current position of the month. This can be solved by adding a small progress bar below *May 2018* text in the figure 8.

5.2 Insights

Once scrolling the timeline of the visualization (going through the months) five insights popped out. First, all the months over the year have similar consumption and distribution. There are three big groups which are dominating all the months: herbal, animal origin and sweets. The consumption has two exceptional months: the consumption in February 2018 was below average consumption fixed by the next month, May 2018, having consumption above average consumption.

Second, the visualization can capture the differences between summer periods and other periods. During summer period, the increasing consumption of the sweets gained the attention and once a closer look was taken, the ice cream seemed to be the main reason of the high number of sweets.

Third, every month includes constant consumption of ingredients related to pets. This means that I must have my own pets. Moreover, if the ingredient names are examined, the pets can be distinguished to cat and dog.

Fourth, liquid group reveals the high consumption of the blueberry juice. As well, not buying any other flavor of juice. This can be concluded that the blueberry juice is my favorite juice.

Fifth, every month at least one plastic bag is bought for the groceries. This is a sad information and can be used to make deliberate lifestyle change, hence not buying the plastic bags at all.

References

[1] Zygomatic.nl, (2018). Free online Wordcloud generator. Retrieved November 10, 2018 from https://www.wordclouds.com/.

[2] Causse, M., Hurter, C., (2009). The physiological user's response as a clue to assess visual variables. 2.2 Visual variables, Fig 2. Retrieved November 10, 2018 from https://www.researchgate.net/publication/221098028_The_Physiological_User's_Response_as_a_Clue_to_Assess_Visual_Variables_Effectiveness

[3] Digital Art, Design, and Communication Education, (2018). Hue,Value,Saturation. Retrieved November 10, 2018 from http://learn.leighcotnoir.com/artspeak/elements-color/hue-value-saturation/

[4] Marmelab, (2018). EventDrops - time based / event series interactive visualization tool powered by D3.js. Retrieved November 10, 2018 from https://github.com/marmelab/EventDrops .

[5] Devaikin, P., (2016). The Russian State Duma. Retrieved November 10, 2018 from http://petrdevaikin.com/duma/en/ .

[6] Peel, L., (2015). Sports skills. Retrieved November 10, 2018 from https://piratepeel.github.io/sportskills.html

[7] Wikipedia, (2018). Color blindness. Retrieved December 15, 2018 from https://en.wikipedia.org/wiki/Color_blindness .

[8] Wickline M., HCIRN, (2000-2001). Coblis – Color Blindness Simulator. Retrieved December 15, 2018 from https://www.color-blindness.com/coblis-color-blindness-simulator/ .

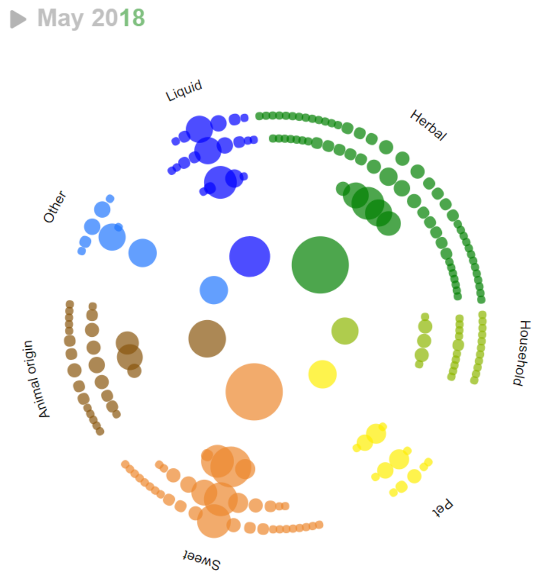
[9] Brewer, C., Harrower, M., The Pennsylvania State University, (2018). ColorBrewer 1.0 – color advice for cartography. Retrieved December 15, 2018 from http://colorbrewer2.org/#type=sequential&scheme=GnBu&n=7 .

[10] JS Foundation. Lodash. A modern JavaScript utility library delivering modularity, performance & extras. Retrieved November 10, 2018 from https://lodash.com/ .

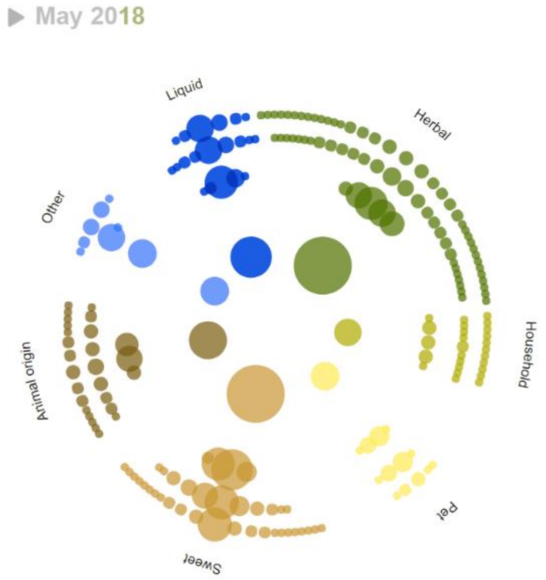
[11] Girault, N., Clayburn, R., Wang, J., Daniel, (2018). CircosJS – JavaScript library to build interactive grapsh in a circular layout. Retrieved November 10, 2018 from https://github.com/nicgirault/circosJS#scatter .

Attachment 1: Evaluating the colors of the draft visualization

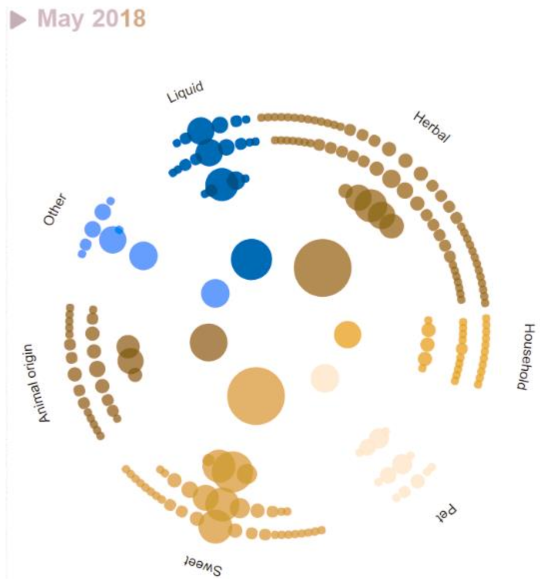
Normal vision



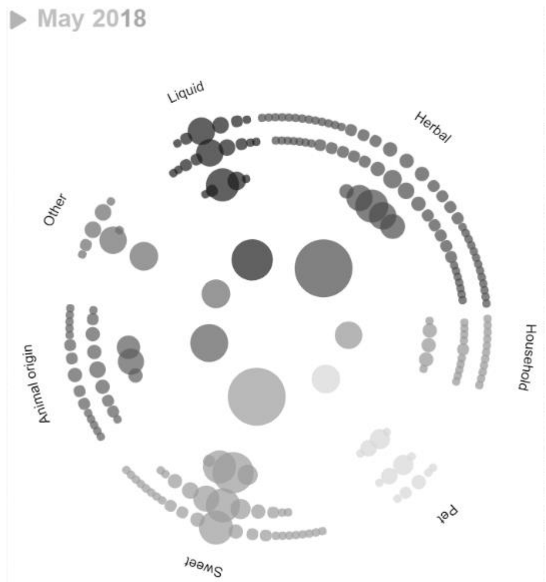
Anomalous Trichromacy vision –
Red-Weak /Protanomaly



Dichromatic vision–
Green-Blind/Deuteranopia

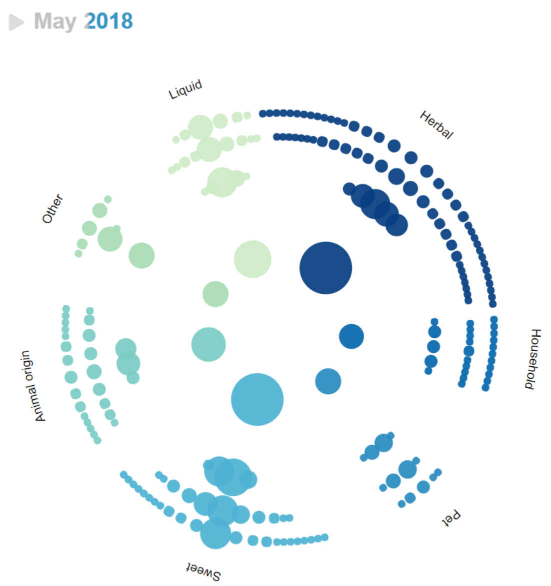


Monochromatic vision –
Monochromacy/Achromatopsia

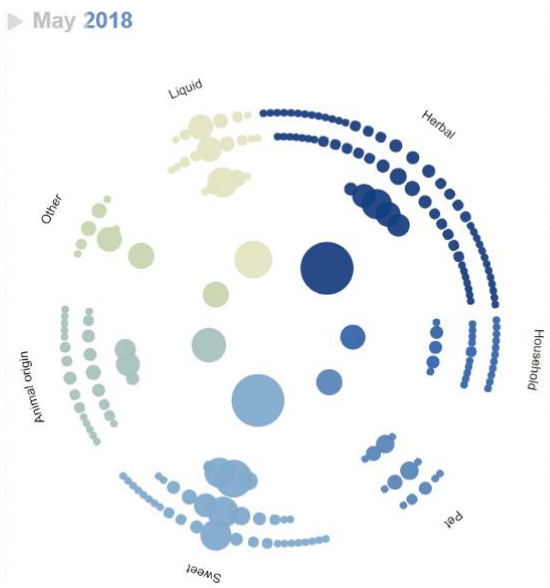


Attachment 2: Evaluating the colors of the final visualization

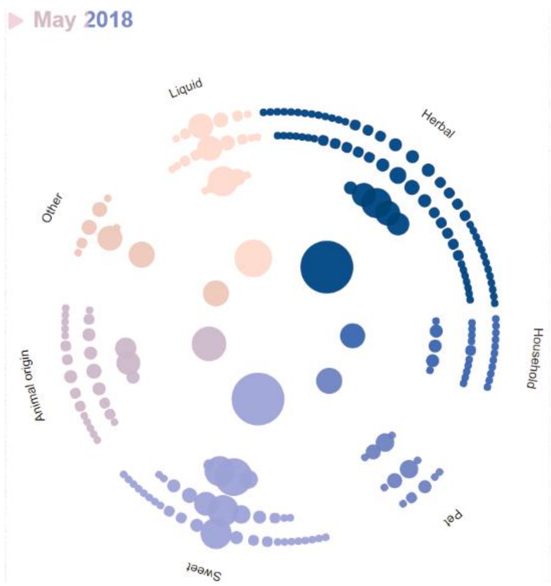
Normal vision



Anomalous Trichromacy vision – Red-Weak /Protanomaly



Dichromatic vision – Green-Blind/Deuteranopia



Monochromatic vision – Monochromacy/Achromatopsia