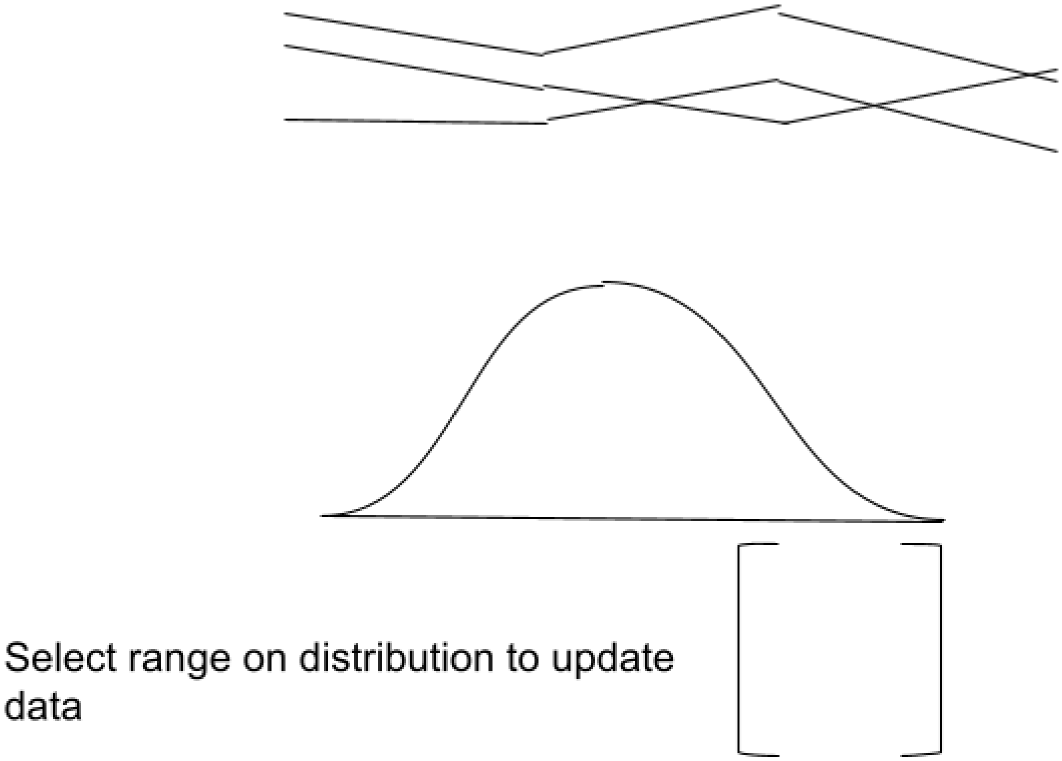
Process Book

Cs 4803 Final Project by: Matt Kiszla, Kathleen Wang, Andrei Bornstein

Project Goals and Motivation:

Initial Project Proposal:

Many of the visualizations we have seen this term have had multiple components to it. There usually is a main visual component paired with a secondary component that complements the main visual; sometimes, it's a secondary table or graph that assists the main visual. For this project, we thought it would be interesting to do this idea with distribution charts. Let's say we have a data set with multiple attributes and we pick one to generate the distribution for that set of the values. Then for the main graph, we could have a variety of different graphs but for this example let's say it's a parallel coordinates chart, then below would be a distribution chart. A user can select portions of the distribution chart and see how it updates on the main chart. It would be interesting to see how a visual updates as you select different sections of the distribution chart.



Initial drawing of possible visual layout with parallel coordinates and distribution chart.

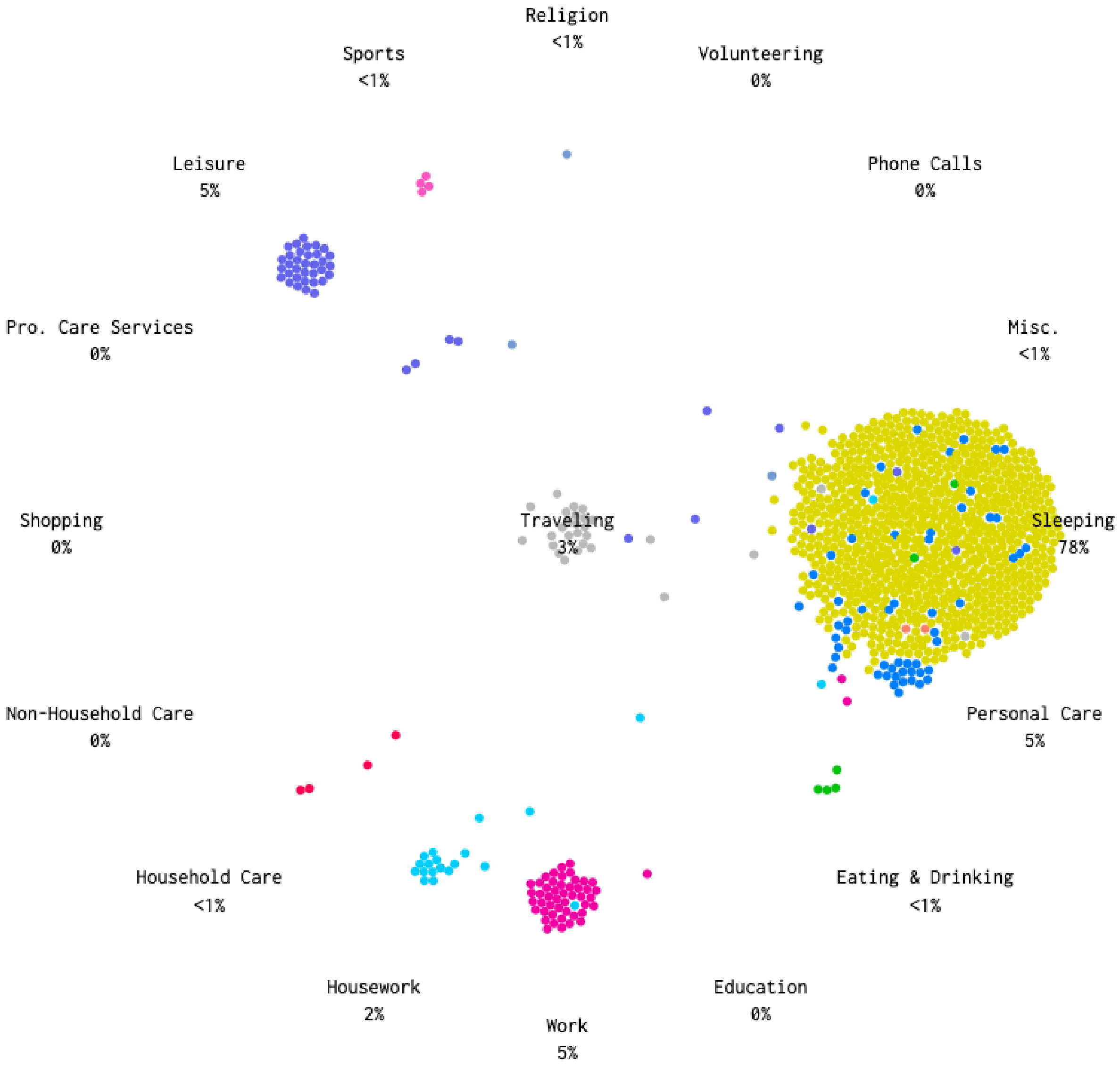
Refined Goals:

After starting the project most of the goals stayed the same, with some minor changes such as the main visual changing into small multiples of various attributes, we were still able to achieve all our goals for distribution charts.

Related Work

Related Work For Data Set:

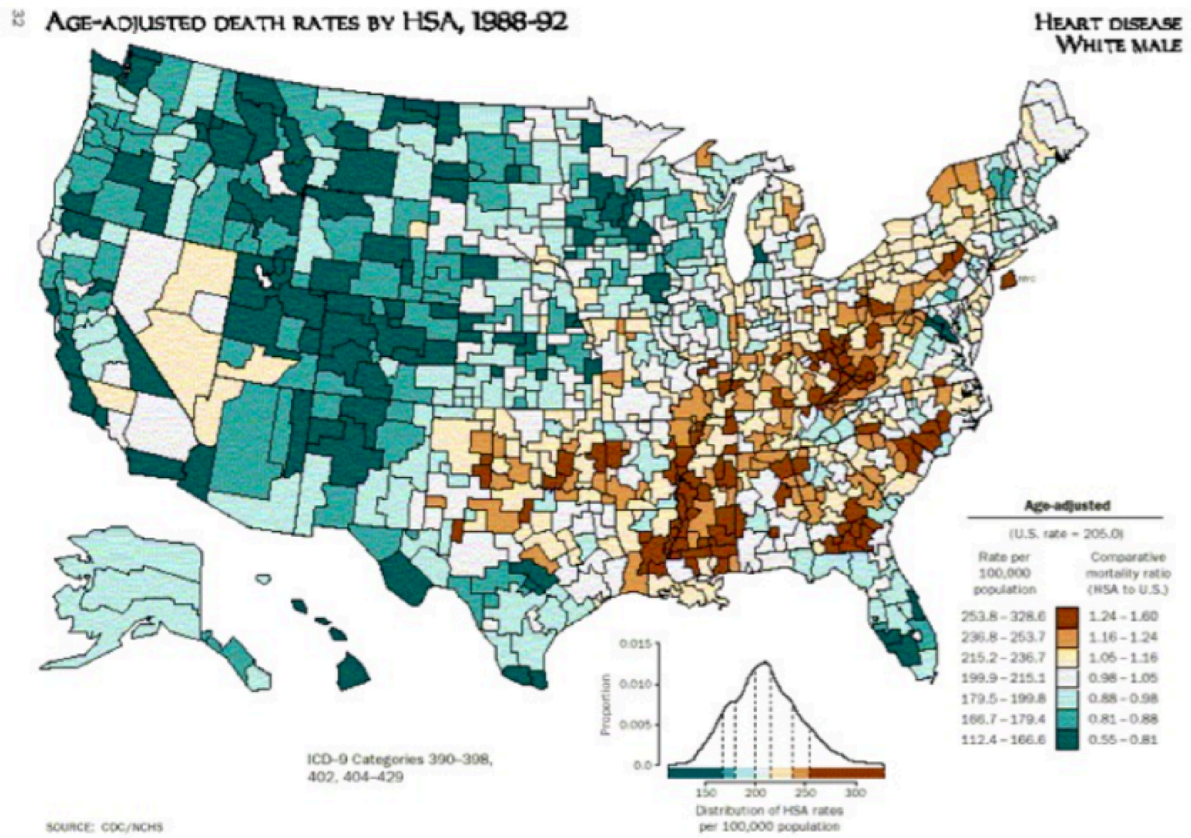
We ended up using the American Time Use Survey (ATUS), one of the more well-known visuals using this survey is *A Day in the Life of Americans This is how America runs* By Nathan Yau.



Screen Shot of Yau's Visualization

Although we did not attempt to replicate his visual nor is our visualization type directly related, his visual still shows what cool things can be done using the ATUS.

Related Work For our Visualization:



Map visualization with distribution chart

This image that was displayed in class was one inspiration for our project. The use of a distribution chart to assist with the main visual was an intriguing idea. This branch into the idea for our project which was making the distraction chart interactive and allow it to update the data for the main visual.

Questions

An important part of the visualization process is what questions the visualization is trying to answer. This can be taken and viewed in different ways, one way is what questions is the visualization trying to answer for the user. Another way is to determine what question is the visualization trying to answer technically or visually. For questions by users being answered, the visualization created will hopefully give insights to user questions about the specific aspects of how Americans spend their time and the characteristics of those time habits. On a more technical aspect, we looked to answer whether distribution charts can be effective at helping query another visual to draw additional or better conclusions from the visual.

Data Source

Determining Which Data to Use:

Before the start of the project, we had to find data that we wanted to use and that worked well for what we wanted to do, which is to have some aspect of a distribution chart.

Here are some other sources for data that we looked at and the reason why they were not selected:

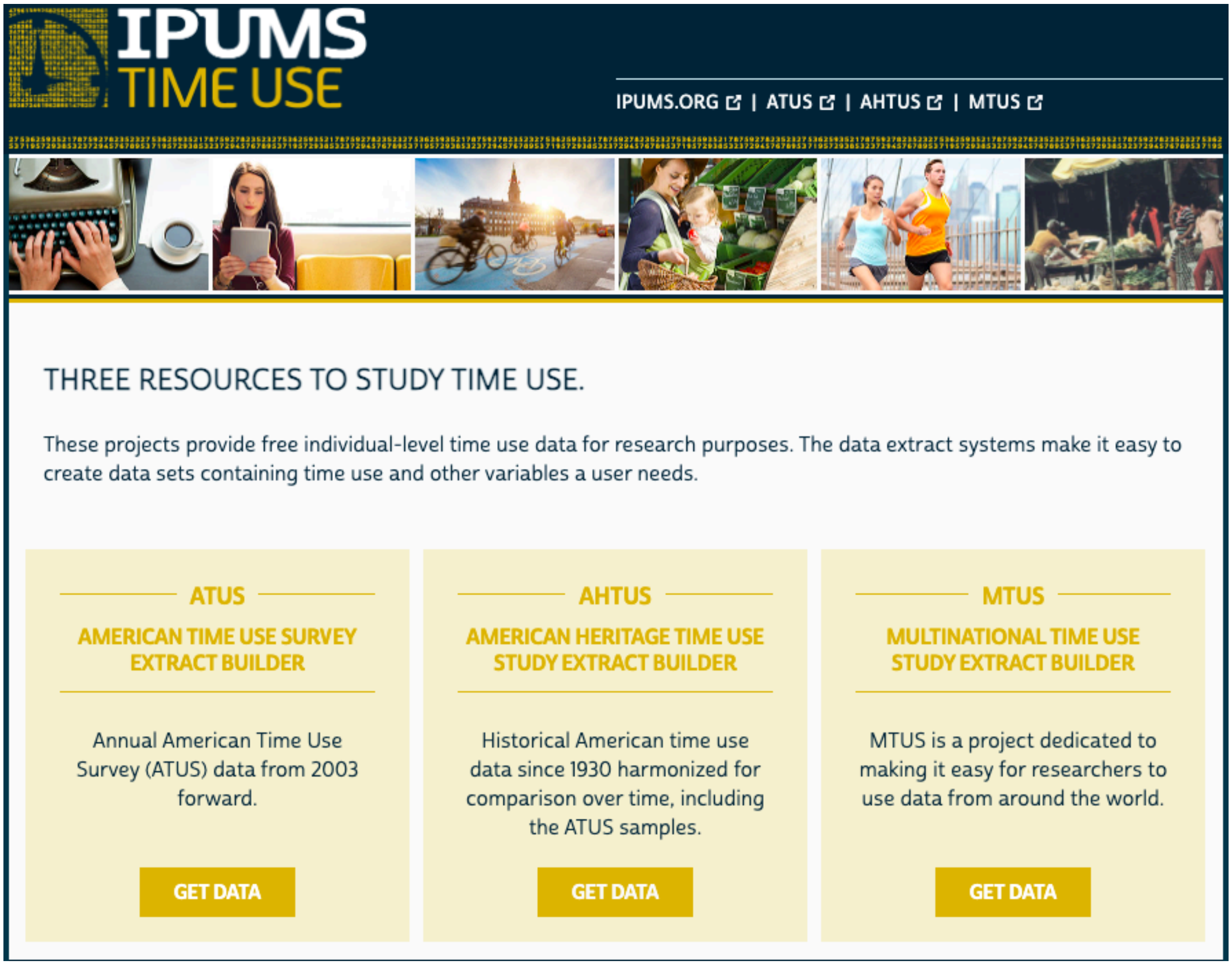
Source	Reason for not Selecting
U.S. Chronic Disease Indicators (CDI)	Values and categories did not seem ideal for what we were trying to achieve
Alzheimer's Disease and Healthy Aging Data	Same reason as above
Global Power Plant Data	Too few categories
1990 Census California Housing Data	Wanted to something more current
NBA Player Data	Terms of Use

Out of all of the ones not selected, the NBA player data would have been the best and could have had some interesting distribution charts. However, it was determined that we would pass on this data as we did not want to possibly violate any terms of use for NBA.com/stats.

It was then decided to use the ATUS data, as it provided various numeric categories that could be used to build distribution charts that would be easy for a user to understand what the data was. The data also had various categories about what type of people were taking the survey which would prove to be useful.

Data Gathering:

It was decided to use the IPUMS Time Use Extractor which is a tool for making interacting with the ATUS easier. This is because they have various Time Use variables that you can create that help group specific categories into broader ones.



THREE RESOURCES TO STUDY TIME USE.

These projects provide free individual-level time use data for research purposes. The data extract systems make it easy to create data sets containing time use and other variables a user needs.

ATUS

AMERICAN TIME USE SURVEY EXTRACT BUILDER

Annual American Time Use Survey (ATUS) data from 2003 forward.

GET DATA

AHTUS

AMERICAN HERITAGE TIME USE STUDY EXTRACT BUILDER

Historical American time use data since 1930 harmonized for comparison over time, including the ATUS samples.

GET DATA

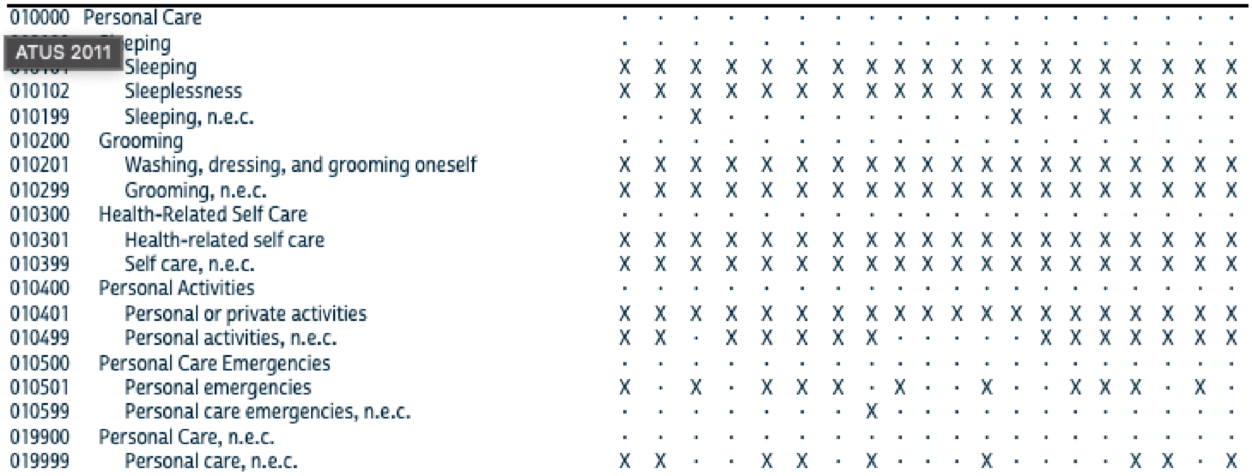
MTUS

MULTINATIONAL TIME USE STUDY EXTRACT BUILDER

MTUS is a project dedicated to making it easy for researchers to use data from around the world.

GET DATA

IPUMS site page



010000	Personal Care	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
ATUS 2011	eping	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
[illegible]	Sleeping	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010102	Sleeplessness	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010199	Sleeping, n.e.c.	·	·	X	·	·	·	·	·	·	·	·	·	X	·	·	X	·	·	·	·
010200	Grooming	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
010201	Washing, dressing, and grooming oneself	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010299	Grooming, n.e.c.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010300	Health-Related Self Care	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
010301	Health-related self care	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010399	Self care, n.e.c.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010400	Personal Activities	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
010401	Personal or private activities	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
010499	Personal activities, n.e.c.	X	X	·	X	X	X	X	X	·	·	·	·	·	X	X	X	X	X	X	X
010500	Personal Care Emergencies	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
010501	Personal emergencies	X	·	X	·	X	X	X	·	X	·	·	X	·	·	X	X	X	·	X	·
010599	Personal care emergencies, n.e.c.	·	·	·	·	·	·	·	X	·	·	·	·	·	·	·	·	·	·	·	·
019900	Personal Care, n.e.c.	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
019999	Personal care, n.e.c.	X	X	·	·	X	X	·	X	·	·	·	X	·	·	·	·	X	X	·	X

Example of IPUMS time use variable Personal Care and its many sub categories

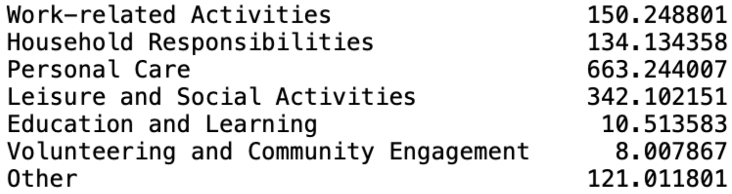
The data from the site was given as a .dat file with an .sas file for labels and headers. SAS Studio was then used to convert the data to a .csv format.

Data Cleaning/Modification

The dataset consisted of 8135 entries and 46 variables. These variables give demographic information about an individual and how they allocate their time across categories, such as ‘Working’, ‘SocializingAndLesiure’, and more. The dataset had no missing values. However, the data needed to be modified for our uses as we felt 17 variables for time allocation was too much for what we were trying to show. We combined certain features under broader names for 7 time-related categories. Data was consolidated by adding the variables. The following table shows the new, or recycled column name, and the features we combined for the new column:

New Feature Name	Features from Original Dataset Used
Work-related Activities	Working
Household Responsibilities	HouseActivities, HouseholdServices
Personal Care	Personal_Care, EatingDrinking, ProfessionalPersonalCareServices
Leisure and Social Activities	SocializingAndLesiure, SportsExercisRecreation, ReligionSpiritualActivities
Education and Learning	Education
Volunteering and Community Engagement	Volunteering
Other	Caring, CaringNonHouseHold, ConsumerPuchases, Telephone, Traveling, GovernmentServicesCivicDuties

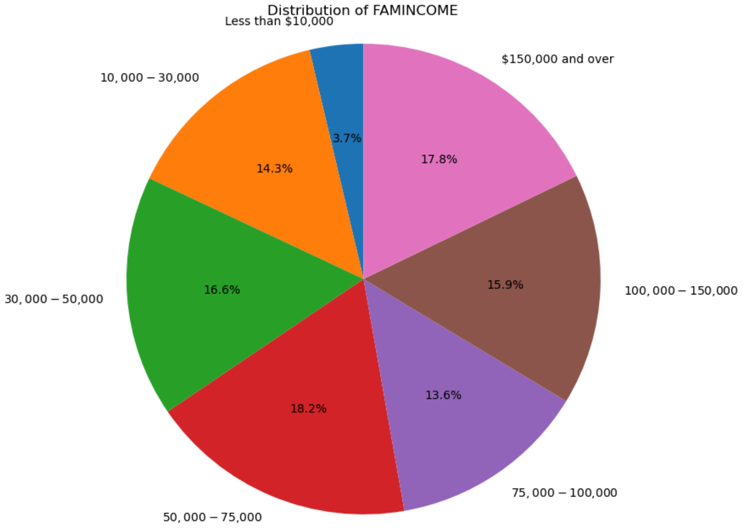
Based on these categories, the average time spent looks like the following:



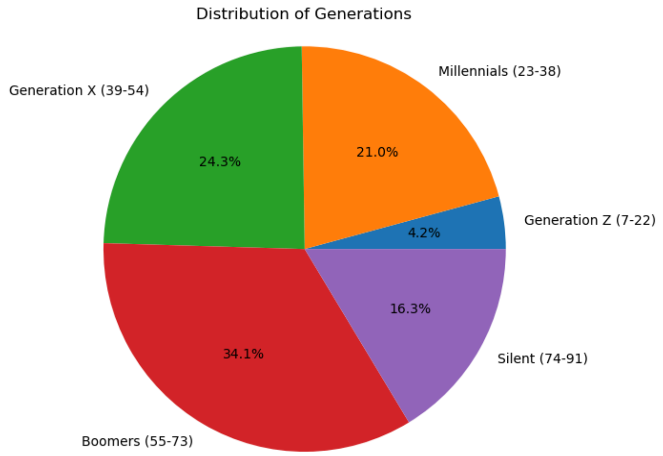
```
Work-related Activities                    150.248801
Household Responsibilities                 134.134358
Personal Care                              663.244007
Leisure and Social Activities              342.102151
Education and Learning                      10.513583
Volunteering and Community Engagement        8.007867
Other                                      121.011801
```

Additionally, we felt it was necessary to map certain functions to get more succinct bins to increase the interpretability of specific variables. This included 'FAMINCOME', 'AGE', 'RACE', 'GENHEALTH', 'HEIGHT', 'WEIGHT', 'EDUC'.

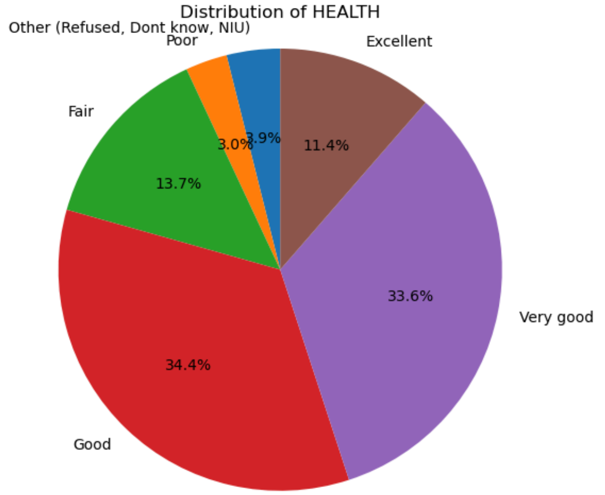
The 'FAMINCOME' variable represents the family income of respondents. However, this variable had 16 unique entries, and was stored as a strings (e.g., '$60,000 to $74,999'). We felt that 16 was too many and needed to bin the data. Ideally this should have been done along some defined guidelines, but given varying income brackets based on family size, we felt it was enough just to cut down on the number of bins. Therefore, we binned the data into the following categories:



A similar process was done with 'AGE', 'WEIGHT', and 'HEIGHT'. 'AGE' was mapped to a generation name, including age range, based on generational boundaries defined by the Pew Research Center as seen in the following:

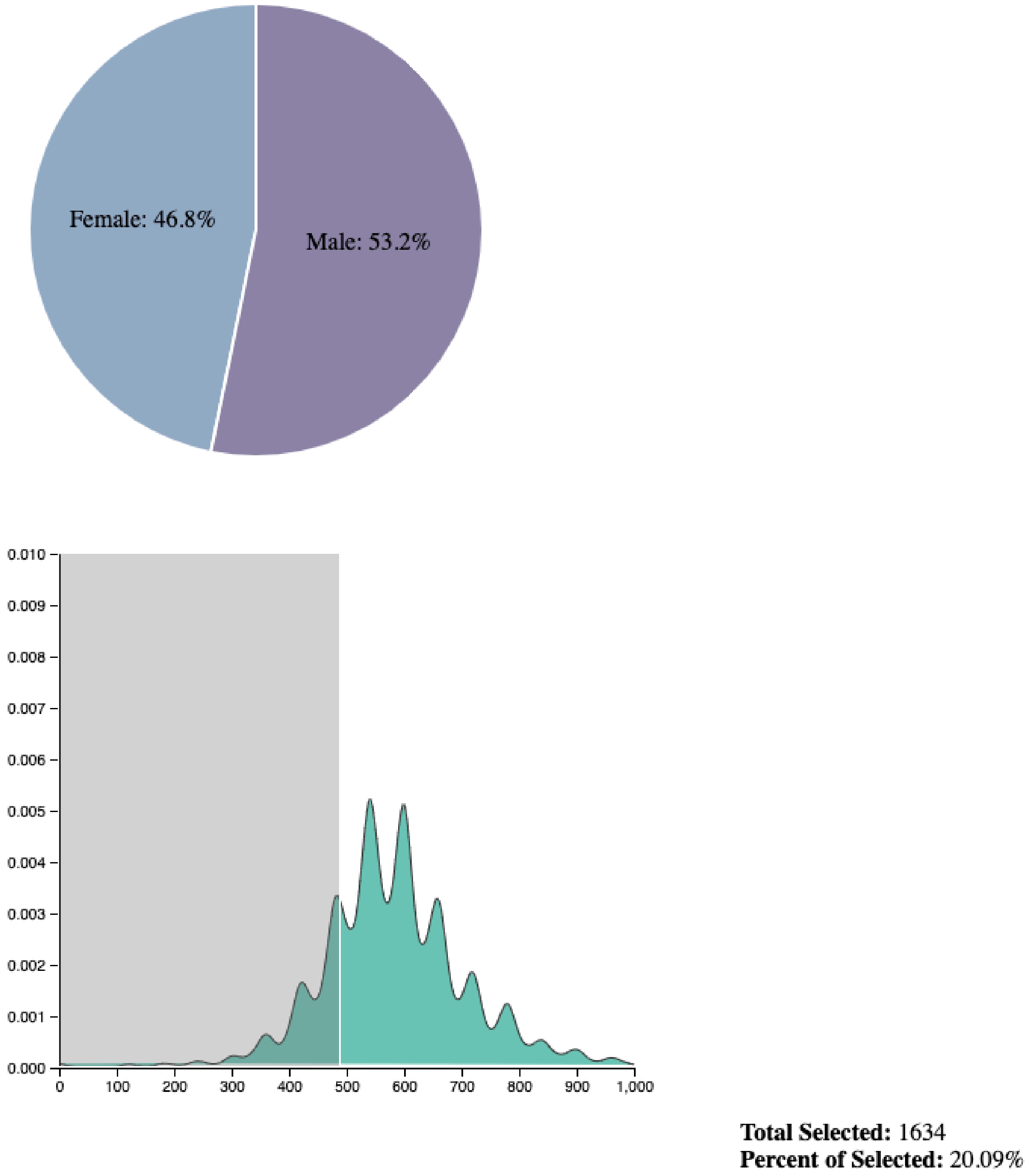


‘RACE’ and ‘GENHEALTH’ were mapped in a similar manner, though both had multiple unique entries that were negligible in frequency. To address this, we consolidated rare categories into a broader ‘Other’ category:



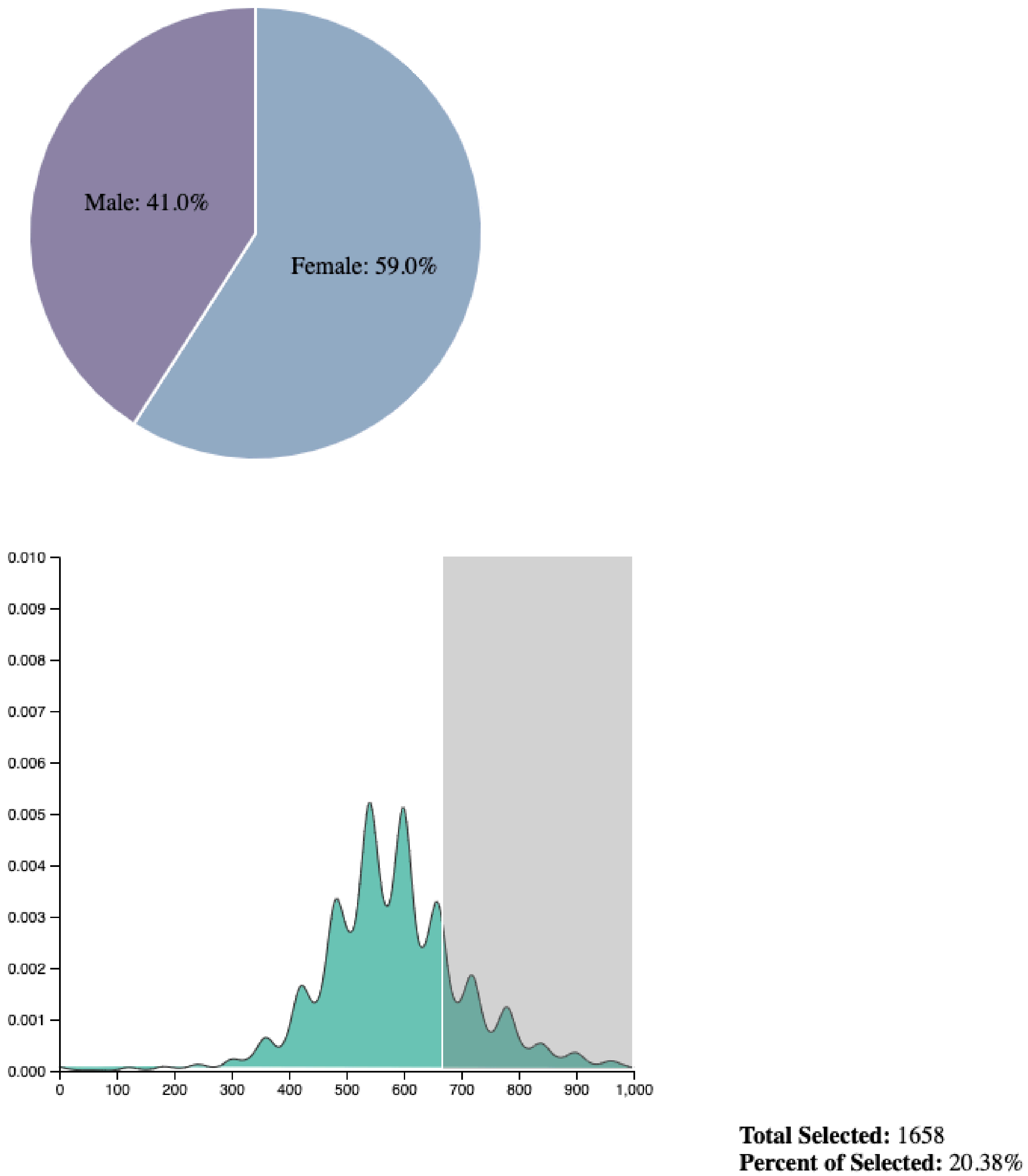
Design Evolution and Implementation

Our First Initial test run looked like this:



Total Selected: 1634
Percent of Selected: 20.09%

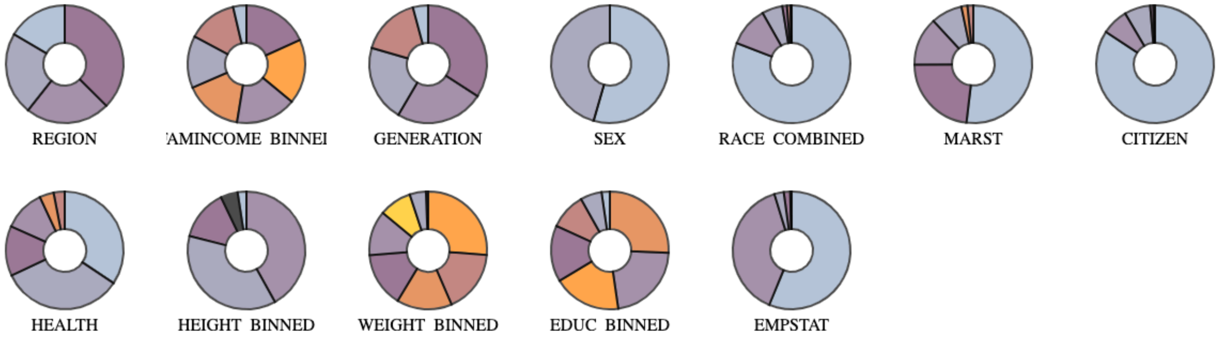
Selection of a portion of time use variable distribution chart and corresponding pie chart about the sex of the data selected



Total Selected: 1658
Percent of Selected: 20.38%

Different selections showing the updated pie chart

From there it was discussed that it would be better to have small multiples of various characteristics



We also decided to add interactive legends and tooltips

For legends:

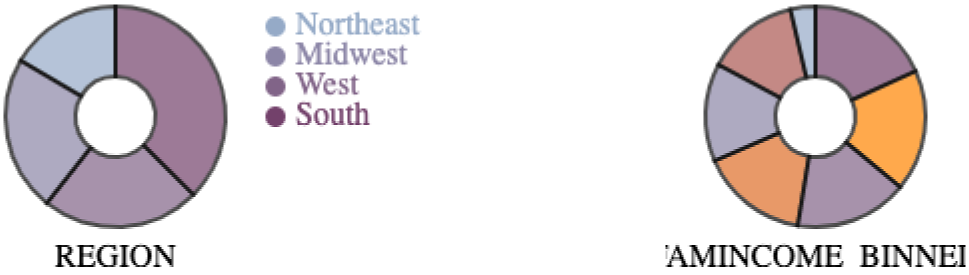
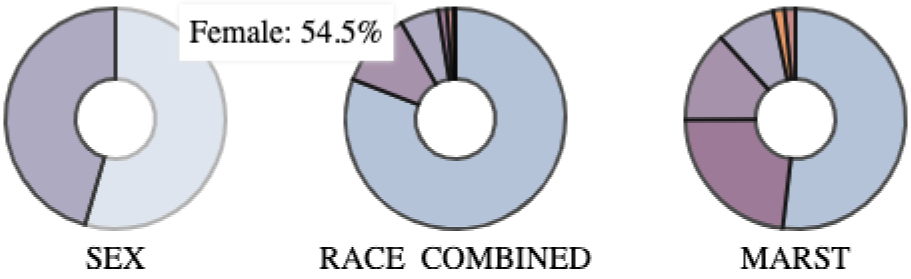


Image of legend example we used

We added the ability that when a user clicks on a donut chart it will expand to show them the legend breakdown.

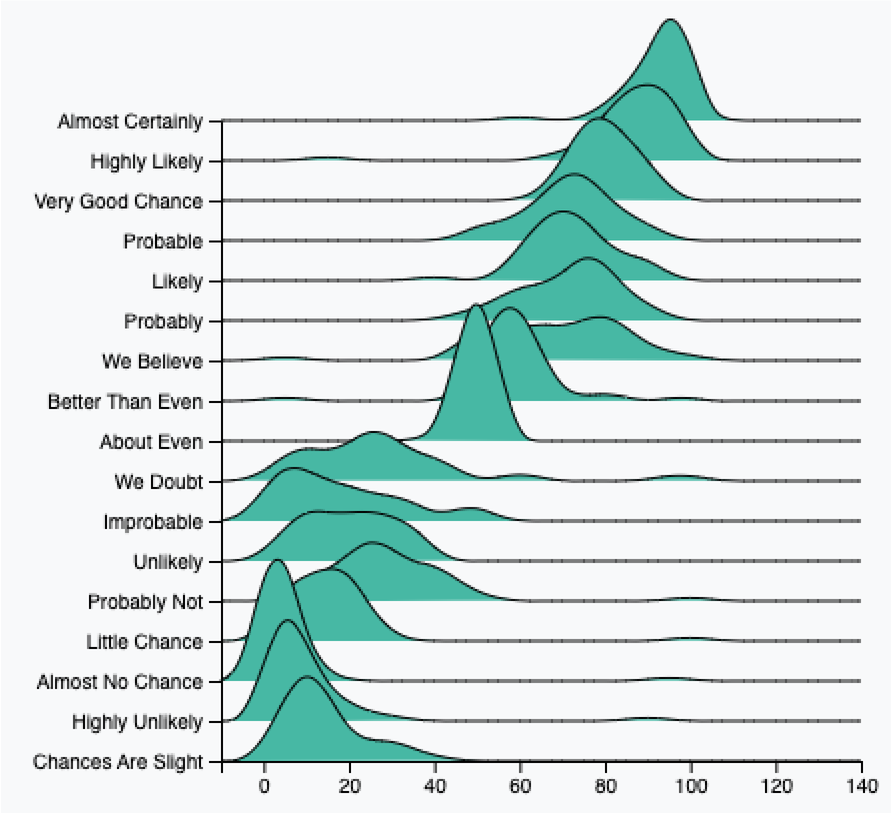
For Tooltips:

Along with updating donut charts with multi-colored categorical slices, we decided to add tooltips to easily see what slice corresponded to what category and what percentage of the whole of each category was. This is useful for debugging as well as easy interpretation of the data for users.



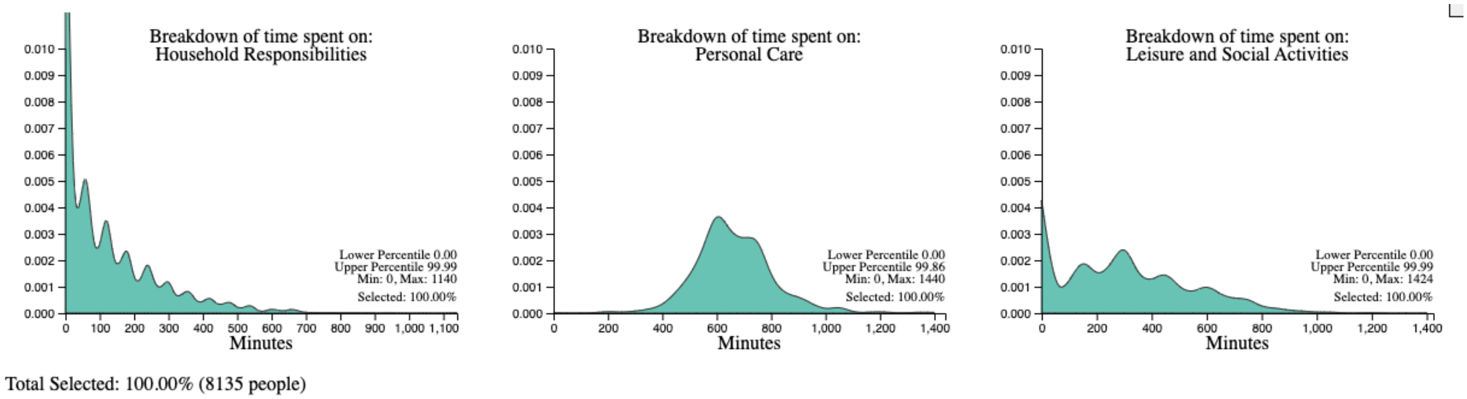
Tooltip example after hovering over donut slice.

Additionally, it was discussed how we could incorporate different density charts for each of our time-related variables. One idea was to stack density charts on top of each other as seen in this example in D3-Gallery



Ridgechart example from D3 Gallery

However, it was determined that since some of the time variables had most of their data close to zero it was better to just allow the user to select up to three variables to look at a single time.

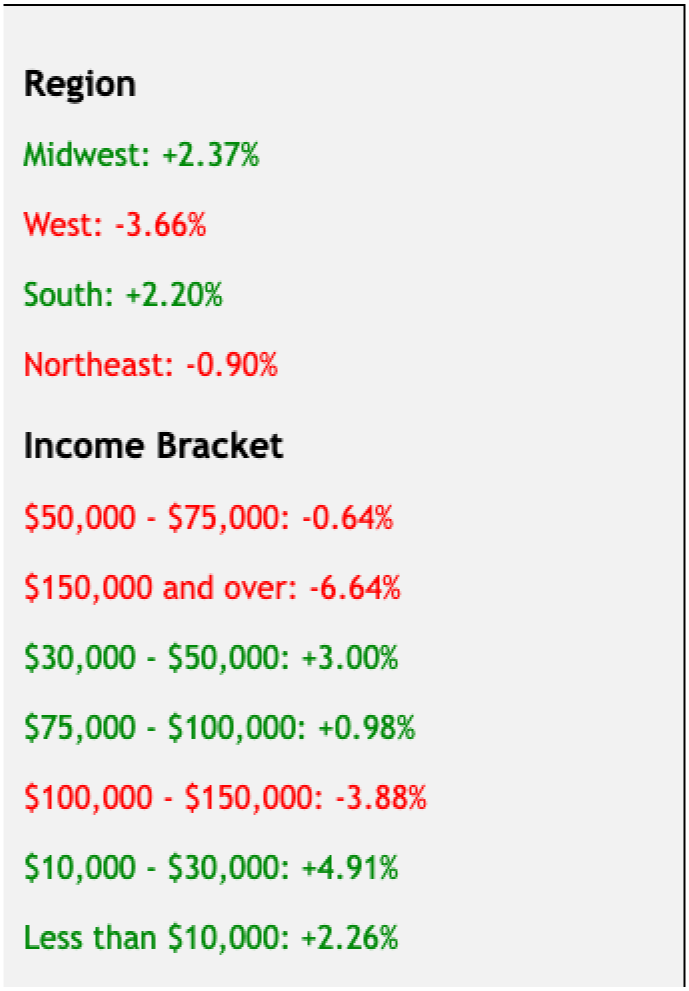


Total Selected: 100.00% (8135 people)

Example of three time-use density charts next to each other

In addition, we added info that allowed the user to see what portions of the graph they are selecting and some information about the data selected along with the total amount of data selected between all the density charts the users are currently displaying.

We also added a section that shows the change from the base percentage for each category in the donut charts when you make a selection on the density chart:



Region

Midwest: +2.37%

West: -3.66%

South: +2.20%

Northeast: -0.90%

Income Bracket

$50,000 - $75,000: -0.64%

$150,000 and over: -6.64%

$30,000 - $50,000: +3.00%

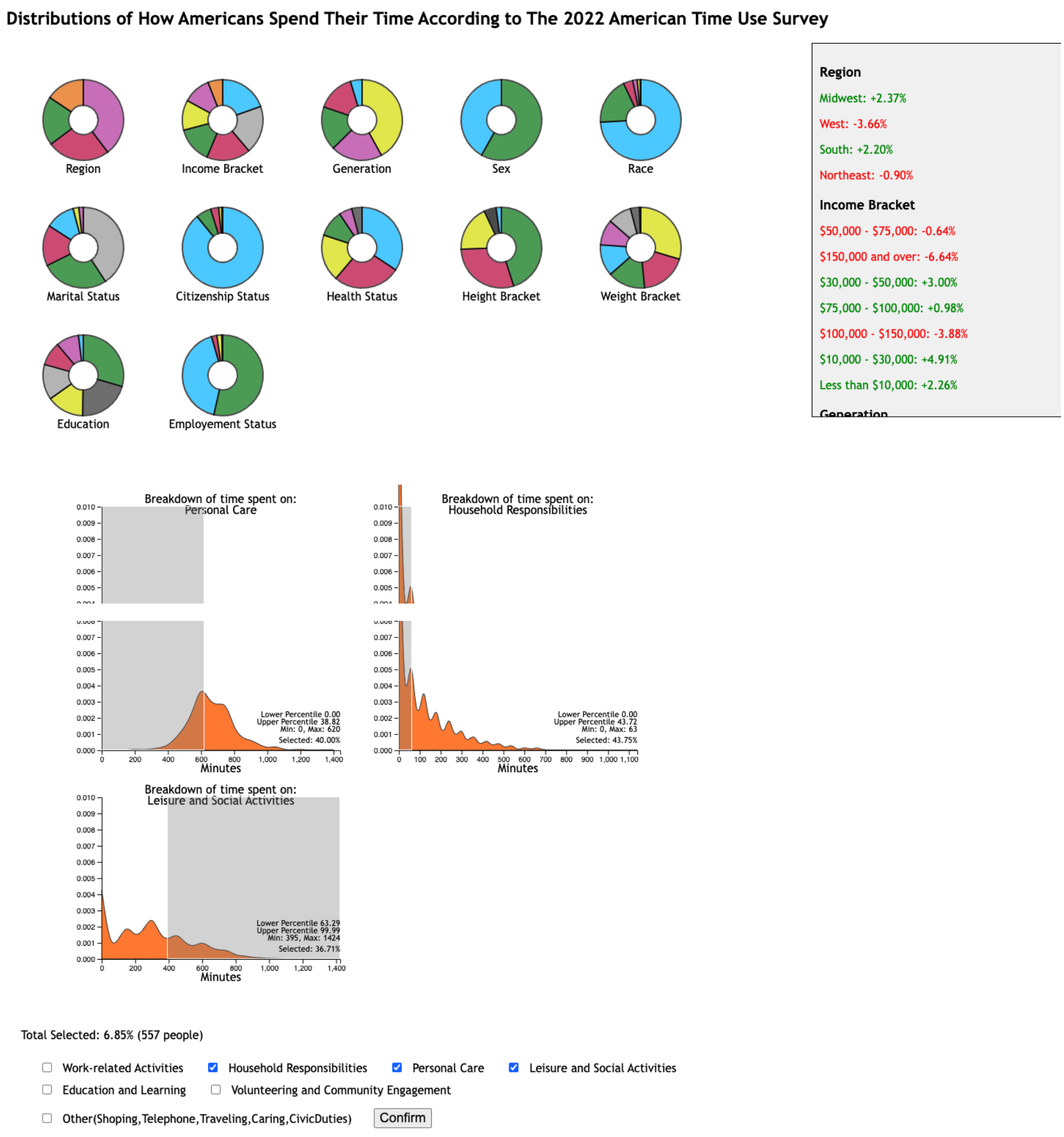
$75,000 - $100,000: +0.98%

$100,000 - $150,000: -3.88%

$10,000 - $30,000: +4.91%

Less than $10,000: +2.26%

Lastly, we made styling changes to the whole document:



Distributions of How Americans Spend Their Time According to The 2022 American Time Use Survey

Region

Midwest: +2.37%

West: -3.66%

South: +2.20%

Northeast: -0.90%

Income Bracket

$50,000 - $75,000: -0.64%

$150,000 and over: -6.64%

$30,000 - $50,000: +3.00%

$75,000 - $100,000: +0.98%

$100,000 - $150,000: -3.88%

$10,000 - $30,000: +4.91%

Less than $10,000: +2.26%

Generation

Lower Percentile 0.00
Upper Percentile 38.82
Min: 0, Max: 620
Selected: 40.00%

Lower Percentile 0.00
Upper Percentile 43.72
Min: 0, Max: 63
Selected: 43.75%

Lower Percentile 63.29
Upper Percentile 99.99
Min: 395, Max: 1424
Selected: 36.71%

Total Selected: 6.85% (557 people)

☐ Work-related Activities ☑ Household Responsibilities ☑ Personal Care ☑ Leisure and Social Activities

☐ Education and Learning ☐ Volunteering and Community Engagement

☐ Other(Shoping,Telephone,Traveling,Caring,CivicDuties) Confirm

Images showing the final design of the page

Evaluation

Using our visualization we were able to draw additional conclusions from the data by quarrying subsets of the data through the distribution chart. You can see how the types of people change when you select different portions of the data. These visualizations also helped answer our question.

"For questions by users being answered, the visualization created will hopefully give insights to user questions about the specific aspects of how Americans spend their time and the characteristics of those time habits."

Are visualization does provide more specific aspects that would help users explore the data more

"On a more technical aspect, we looked to answer whether distribution charts can be effective at helping query another visual to draw additional or better conclusions from the visual"

From our visualization it can be determined that distributions can be an effective way to help query a data set to explore further and draw more conclusions then looking at the data set as a whole.

Overall, are visual works fairly well. Some ways to improve it might be to have more clear labels of the different sections and find a better way to fit everything on the page on the page. In addition, maybe allow users to switch between years of the ATUS might be a good idea.

Citations:

Sarah M. Flood, Liana C. Sayer, Daniel Backman, and Annie Chen. American Time Use

Survey Data Extract Builder: Version 3.2 [dataset]. College Park, MD: University of Maryland and Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D060.V3.2

Yau, N. (2023, October 21). *A day in the life of Americans*. FlowingData. https://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/

Holtz, Y. (n.d.). *The D3 Graph Gallery – Simple charts made in d3.js*. https://d3-graph-gallery.com/