Theme 1: Group Population

Team Members: Chahak Tharani, Amola Hinge

1. Problem Statement

This project aims to design a tool to help explore ways to divide our total population of the ATUS dataset across the various metadata categories into groups with enough samples. Through visualizations, we wanted to show the distribution of samples across multiple subgroupings. Our primary focus is to understand how all the subgroups can be shown effectively as the number of dimensions increases.

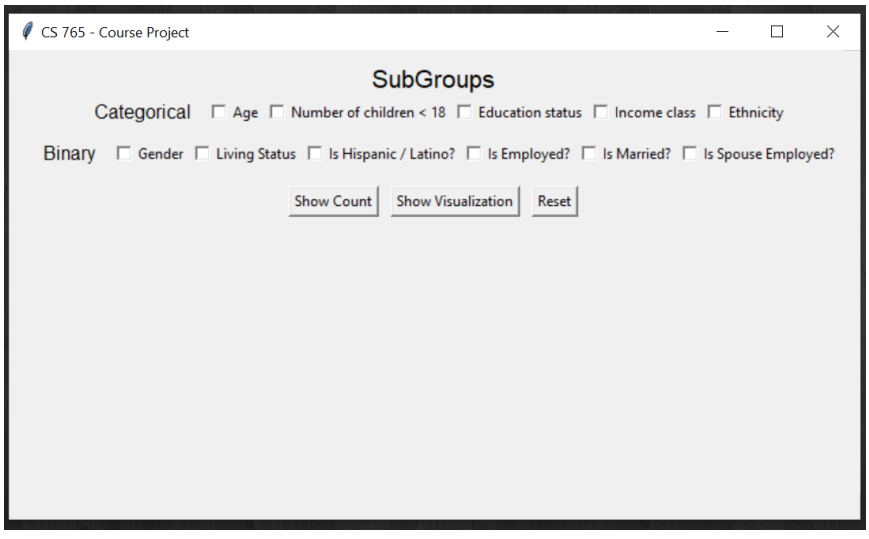
2. Methodology

First, we started exploring the ATUS dataset by understanding all the metadata categories available for dividing the population into groups. Next, we designed 1D histograms and 2D count plot charts through our class learnings to visualize the subgroups. Then, we started researching visualizations for viewing group sizes in multiple dimensions (>= 3). Lastly, we worked on making an interactive Python application that would help select the variables to group the population on and create the visualizations for showing counts of subgroups in real time.

3. Implementation and Visualization

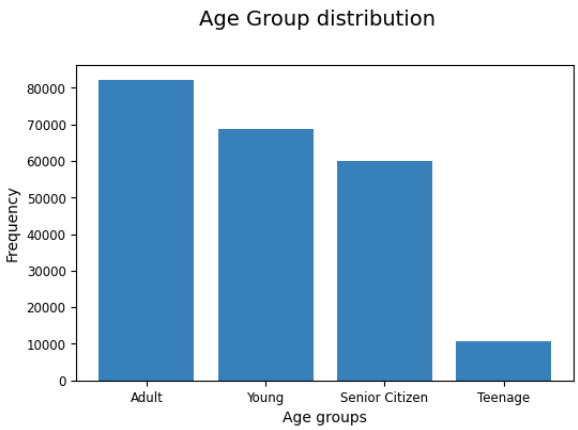
We have worked with the ATUS summary file ("*atussum_0321.csv*"). We preprocessed and cleaned the data according to the metadata description into categorical and binary data.
We have built an interactive tool allowing the user to select variables to group the population. We have divided the variables into two types: Categorical and Binary. Upon selecting the desired variables, we have provided options to see the distribution across subgroupings in a tabular or visualization format.

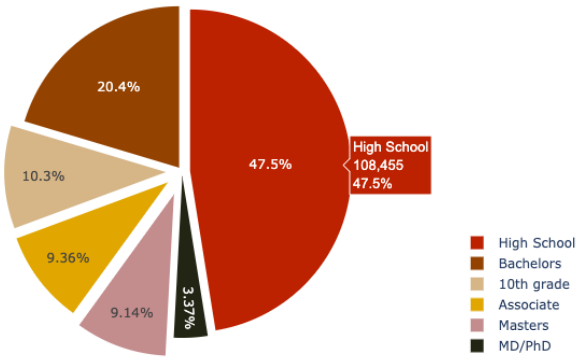
We have primarily used Python for this entire project. We have explored various libraries like pandas, NumPy for data manipulation and seaborn, plotly, matplotlib, upset_plot for data visualization and tkinter for building the interactive tool.



1-Dimension:

Histogram and Pie Chart - These are used to plot the distribution of a categorical variable with frequency shown as bar height or pie percentage, respectively.

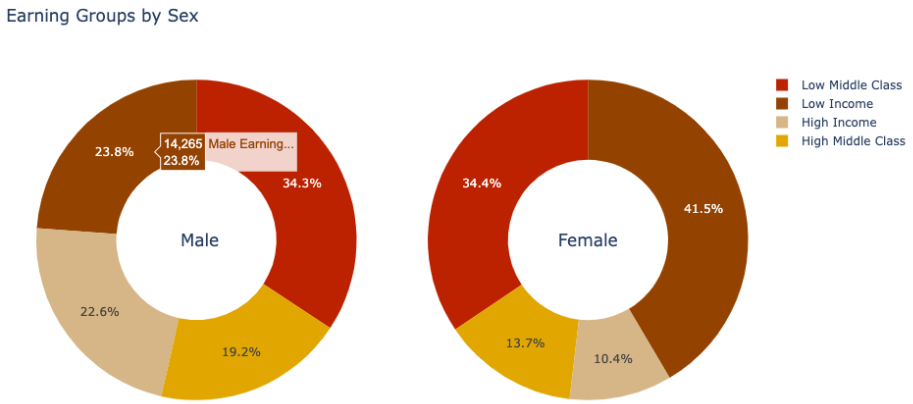


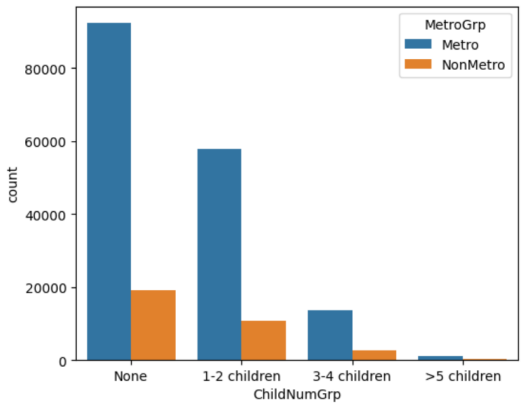


Critique: The Pie chart shows the distribution of education qualifications. The percentages help us identify which group is bigger, and the hover information has the sample size count.

2-Dimension:

1. ***Donut Chart*** - Donut charts are used to show the proportions of categorical data.
2. ***Count Plot*** - A count plot shows the counts of observations in each categorical bin using bars.

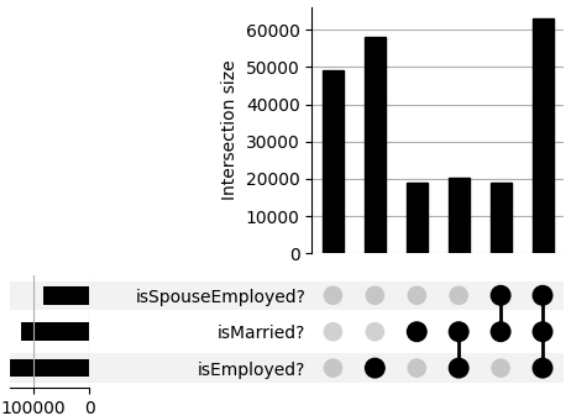




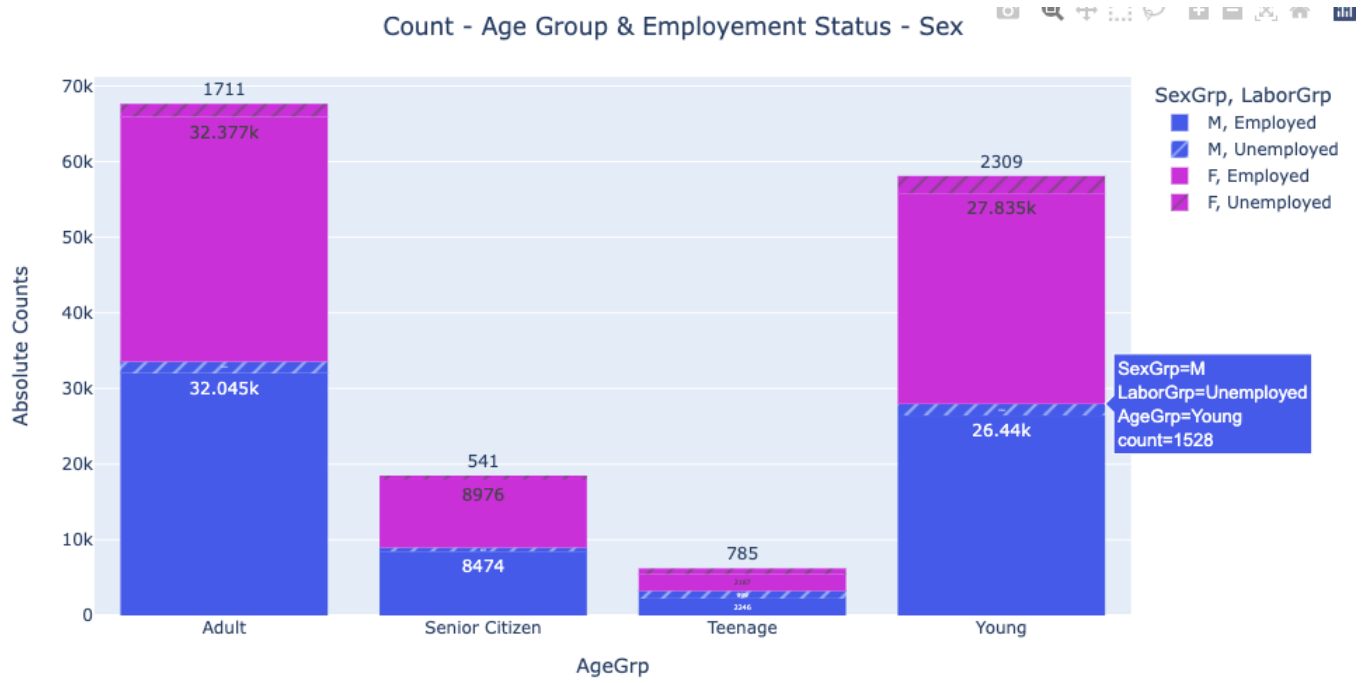
3-Dimension :

1. Upset Plot - UpSet plots show set data with more than three intersecting sets. We have used it for visualizing three binary variables. The sample size of the subgroups are shown as bar charts.

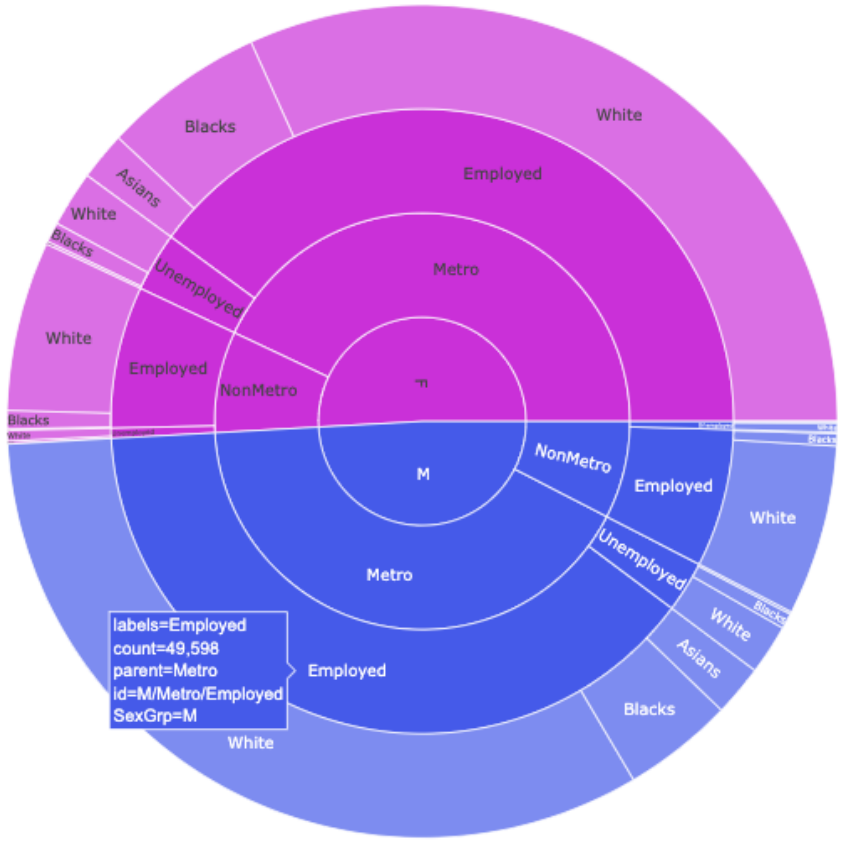
Critique - This plot helps view all combinations of subgroupings of binary variables where the black dot represents *true* value and the gray dot is *false*.



2. Stacked Bar Chart with markings- This stacked bar chart shows three categorical variables. The age groups are shown along the horizontal axis, and sex is represented as stacks within each categorical bar. The employment status can be seen by dashed vs solid color. The hover information provides the sample size.



4-Dimension:



1. Sunburst Plot - The Sunburst Plot is a variant of Doughnut Plot, used to display a hierarchical structure. Each level of the hierarchy is represented by one ring or circle, and all rings show how the outer rings relate to the inner rings.

Critique: The different colors show different sex categories, but if we incorporate color in other variables, too like Employment status and Ethnicity categories, it would be easier to link the sample group counts.

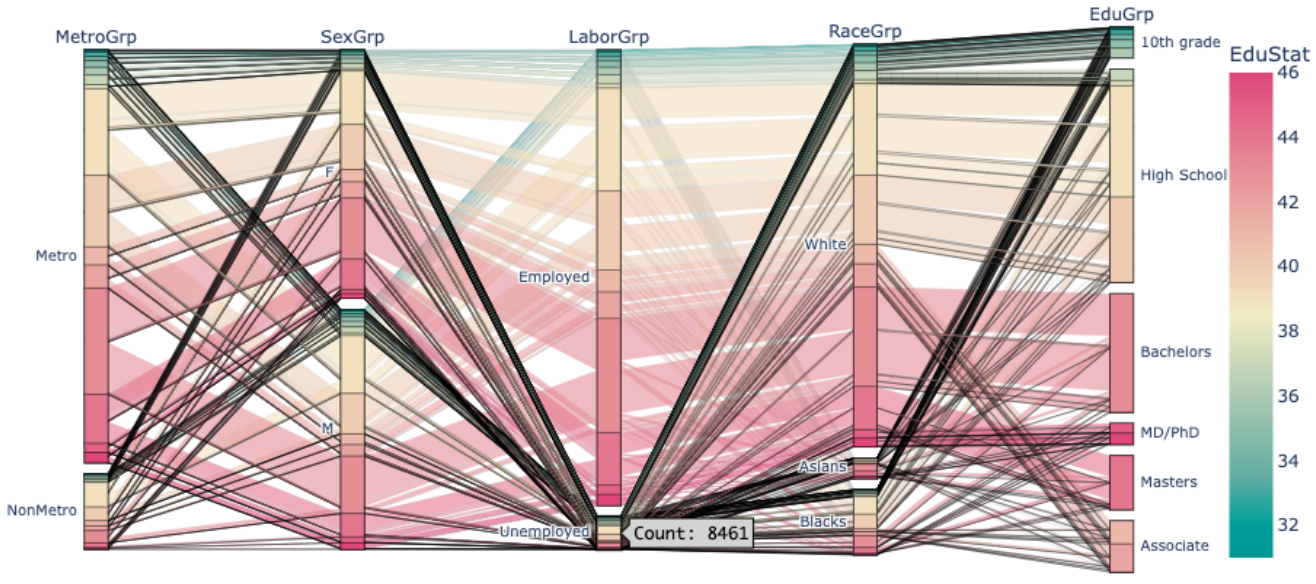
Fig: Sex Group with Living status, employment status and ethnicity

2. TreeMap: This also shows hierarchical groupings. The plot below shows how males and female population is divided into employed vs unemployed. Then among those groups, what is the distribution of people having partners, and among them, how many partners are unemployed vs employed?

Critique: The male and female group colors are both dark purple which do not give an estimate of the group size if we do not see the hover information. Employed males who have unemployed spouses are more in number than employed females with unemployed partners (dark blue represents a bigger count than black)

5-Dimension :

1. Parallel Coordinates - Each of the measures corresponds to a vertical axis, and each data element is displayed as a series of connected points along the measure/axes. This shows living status, sex, race, employment status, and educational qualification.



Critique: We can visualize that unemployed people (especially males) primarily belong to non-metro categories and have lower educational qualifications than employed people. The interaction in this chart helps in viewing multivariate categories at once and their sample size, and whether it is so small that the population cannot be divided on that axis.

5. References

[1] https://medium.com/swlh/effective-visualization-of-multi-dimensional-data-a-hands-on-approach-b48f36a56ee8

[2] https://medium.com/@narayanmahto/visualizing-intersecting-sets-upset-chart-in-python-cf72e4cad5b1

[3] https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57

[4] https://towardsdatascience.com/visualizing-multidimensional-categorical-data-using-plotly-bfb521bc806f

[5] https://towardsdatascience.com/processing-and-visualizing-multiple-categorical-variables-with-python-nbas-schedule-challenges-b48453bff813

[6] https://www.analyticsvidhya.com/blog/2021/11/visualize-data-using-parallel-coordinates-plot/

[6] https://towardsdatascience.com/parallel-coordinates-plots-with-plotly-dffe3f526c6b

[7] https://chartio.com/learn/charts/histogram-complete-guide/#:~:text=A%20histogram%20is%20a%20chart,value%20within%20the%20corresponding%20bin

[8] https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/

[9] https://www.originlab.com/doc/Origin-Help/Sunburst-Plot#:~:text=The%20Sunburst%20Plot%20is%20a,relate%20to%20the%20inner%20rings

[10] https://plotly.com/python/parallel-coordinates-plot/