American Life

An Embedding Study of the ATUS Dataset

Yuan Chen

Chen2243@wisc.edu

Introduction

High-dimension data, such as the ATUS dataset, is challenging to explore. A common practice is mapping the data into low-dimension, called embedding. Embedding can be achieved by dimensionality reduction (DR) techniques. Some popular DR methods mentioned in this class are PCA, t-SNE, and UMAP. I will try all three and compare the results to find the most suitable method in the later sections. In this project, I plan to use the chosen embedding to map the ATUS dataset onto a 2D map. Then this map will be explored to determine what factors (mainly demographic information) affect how Americans spent their time from 2003 to 2021.

Python with Pandas, SciKit-Learn, UMAP, and Matplotlib libraries, is used to preprocess the data and apply the dimension-reduction techniques. Tableau is used to explore the raw data, design visualizations, and construct stories from the data.

Comparison of Dimension Reduction Techniques

In this section, I compare the three dimensionality reduction methods taught in previous classes. I don't see the point of doing all three since the focus is visualization design. One DR method is chosen for the rest of the project.

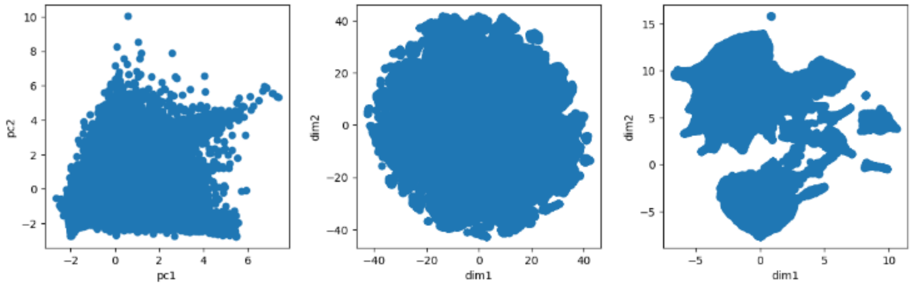

Figure 1 The ATUS18 embeddings using different DR methods. Left: PCA; Middle: t-SNE; Right: UMAP.

I tested the atussum_18 dataset provided by the class (will be called ATUS18 for the rest of the project) with PCA, t-SNE, and UMAP. PCA(figure 1, left) works only well if the first two components preserve most of the information in the data. Otherwise, too much information is lost in the process and renders PCA less effective. In theory, t-SNE should work much better, preserving the global and local structure, yet the first few runs generated no discernible structure(figure 1, middle). Fine-tuning t-SNE may produce a better result. However, t-SNE runs very slowly on my computer, and tuning parameters became too time-consuming. At last, UMAP is chosen for this project. It runs reasonably fast and preserves global and local structures similar to t-SNE. Without too much effort, the embedding shows promising results with a clear structure(figure 1, right).

Visualization Design

After UMAP is chosen for the project, the resulting embedding of ATUS18 is plotted in Tableau to test out design choices. Since demographic information is categorical, color encoding is an effective way to distinguish them. As an example, the employed vs. unemployed plot is shown in figure 2. The default large circle and color encodings(leftmost) used by Tableau don't offer any contrast between groups. With 228K data points, the large circles completely blend in, and the employed group almost covers the unemployed group. I first improved it by changing the size channel. The reduced dot size increased the space between data points and made them more distinguishable.

On top of changing size, the contrast between the default blue and orange is not great either. On a color wheel, colors on opposite sides offer the best visibility. If I choose to keep blue as one of the default colors, the other should be changed to yellow, as in the third picture in figure 2. However, the blue subset is still masked by the yellow subset. I tested several methods to increase the contrast. Filter is one of the choices mentioned several times in the class, but I must be careful about what to filter out. For example, I can choose the dataset by year, age, or random downsampling. A risk is that important information may get lost if the filter is not carefully selected. Eventually, I utilized the opacity setting and changed it to 25%. The groups became very distinguishable. The legend also became opaque and became a little harder to read.

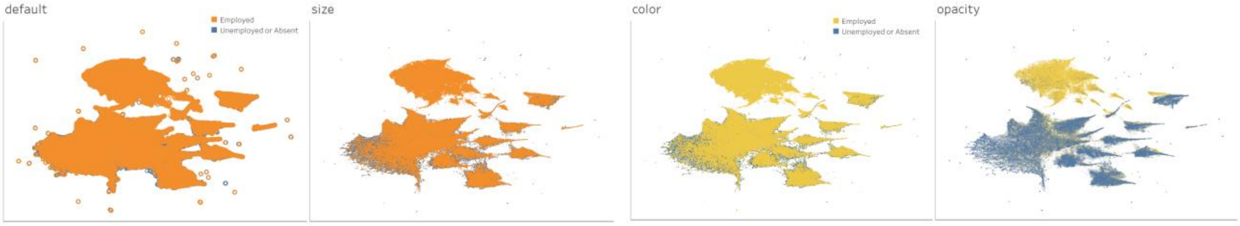

Figure 2 Embedding visualization. From left to right: Tableau default, smaller circle size, high contrast colors, and higher opacity.

Some other plot types are tried and compared to scatter plots, such as density map (figure 2, left) and bin map(figure 2, right). They are better at displaying the data point density at given positions. However, the density map doesn't deal with the overlapping data points well and gives a foggy visual effect. The gap between bins in the bin map significantly weakens the visibility.

As a result, I chose the opaque scatterplot with color encoding for this project since my objective is to separate groups assembled by UMAP and use the information to tell what factors are essential when the algorithm groups people in the same cluster. Data point density at given positions is good to know but not crucial.

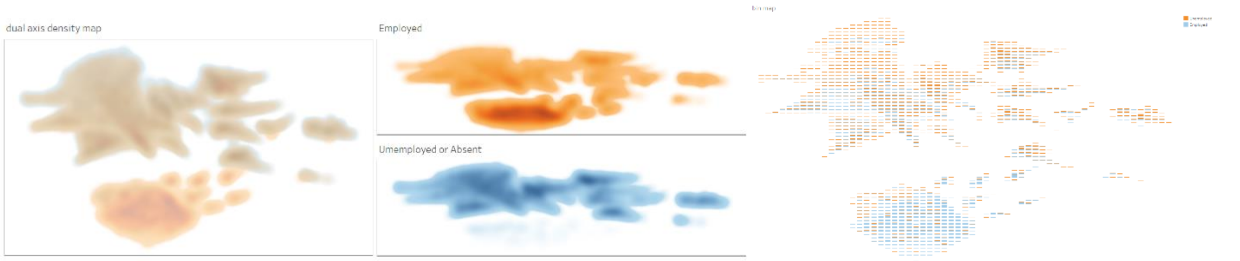

Figure 3 other designs: density map (left, middle), bin map (right).

Storys

Patterns Finding Using the Embedding

UMAP transforms a high-dimensional dataset into a low dimension by calculating the similarity of the data points. Similar data points are close to each other in the resulting embedding if such a structure exists. The 18 time-usage attributes in the ATUS18 dataset are scaled by the StandardScaler module from the SciKit-Learn library and then fitted using the UMAP library with the default setting. The resulting embedding is tested with several color-encoded demographic attributes provided in the dataset to look for patterns. I found that the most significant one is employment status(TEFLS), as shown in figure 4(top left). TEFLS contains five labels, but 2-5 have negligible differences, so they are grouped into "unemployed or absent." Clearly, the working population(yellow) occupies the top cluster, and the population that is not working(blue) contributes the most to the surrounding groups. Working people spend more than 6 hours working at the cost of everything else, especially social and relaxing (figure 4, bottom left).

The second most important factor is whether a person has children (Trchildnum). The population with children(green) occupies two horn-like areas in the middle (figure 4, middle top). The number of children doesn't impact the clustering significantly, so they are grouped into "one or more children." People with children spend 69 min more on "care for household members" than those who don't (figure 4, middle bottom). And again, people squeeze the time mainly from Socializing and Relaxing. If I combine these two patterns above with age, there is a rough shift from the right side to the left as age increases. When people are in adolescence (light green), they spend more time in school, play more sports, and sleep more. They don't do much housework nor need to care for others (figure 4, bottom right). Life becomes more burdensome when they grow older and become a young adult (dark green). They graduate from school, look for jobs and start to form families. Some may even have had one or two kids. The following two groups, marked greenish blue (25-34 yrs old) and cyan (35-44 yrs old), are the most challenging stage of an individual's life. They sleep the least, relax the least, but work the most. These two groups largely overlap with the work cluster and the children clusters. Life moves forward to the age between 45 and 59(dark blue), and the burden starts to fade. Their children grow into adulthood, and the data points move out of the two horn-like regions in the middle. Finally, they begin to retire at age 60(orange). The points advance to the bottom left portion of the embedding. Few of them are still working, and none are left to care for, so they can start to sleep and relax once again.

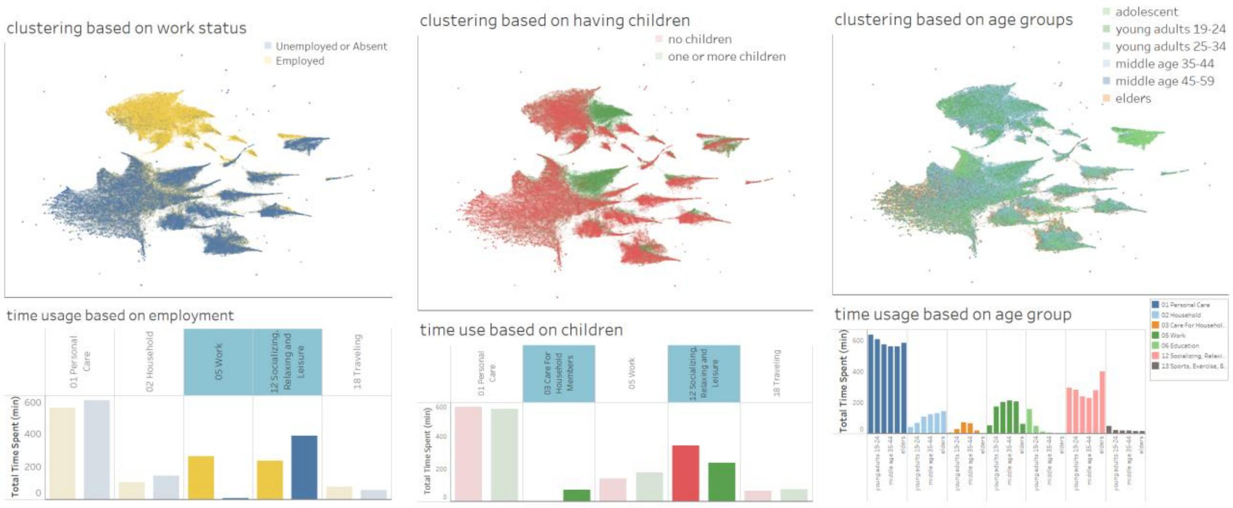

Figure 4 Time use grouping: working status(left), children(middle), age(right)

Confirm Data Exploration Results Using the Embedding

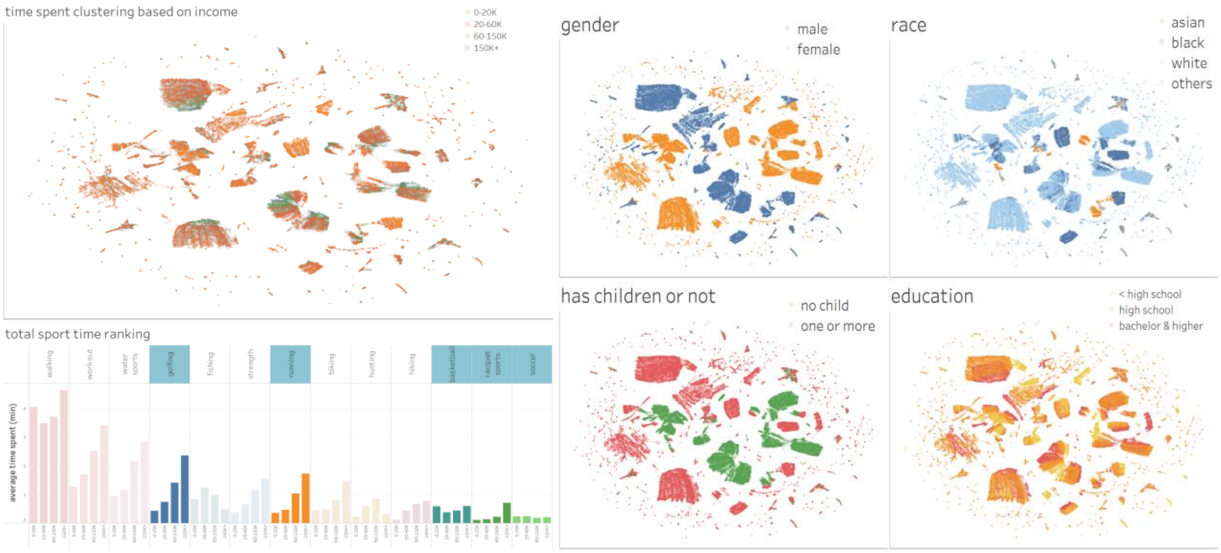

Figure 5 Time use clustering based on income (top left). Top 10 sports played, plus three sports chosen by the creator of the original visualization (bottom). And clusters labeled by gender, race, children, and education (right).

In the critique practice, I found that #37_3 tells a fascinating story about how wealthy and poor people play different sports. To test this story, I extracted T130101 to T130199 from atussum_0321 and mapped the subset to 2D space (figure 5 top left). Surprisingly, there is no clear separation of groups, indicating income is not one of the main factors when UMAP grouping people. A deeper investigation into the subset reveals some problematic choices from the creator. For example, more than half of the chosen sports in #37_3 are not even in the top 10 played ones except golf and running, which ranked #4 and #7, respectively. In an alternative story, the high-income group dominates most of the top 10 popular sports, but in the most popular sport – walking, poor people are not far behind the riches. It's safer to say that wealthy people play more sports than low-income people in general. Are there some factors that can separate people who play different sports better? In figure 5(right), I demonstrate that labels such as gender, race, and children are much more impactful in grouping the data points. Labeling based on these three labels has little overlap and forms distinct groups. Some other factors, such as education, is also more important than income, although less than the three other factors. Based on these labels, we can find a specific description of individuals who play similar sports. For example, the top left cluster can be described as a white male with no children. And these characteristics determine what sports he plays.

Summary

In this work, I studied how embedding can be used to find patterns in the ATUS dataset and whether it can be used to confirm the patterns found during exploration. Three dimensionality reduction techniques, PCA, t-SNE, and UMAP, are compared, and UMAP appears to produce the best results for my purpose. In the first story, UMAP is used to explore what factors affect Americans' time usage the most. Some clear patterns point to work and children as the most impactful ones. With the age label, they can show how average Americans live through their lives. In the second story, the embedding of sports participation tells a different story about the relationship between income and sports from the original visualization. Income is not a good indicator of what kinds of sports are played. Gender, race, and children decide sport participation. This study is an excellent exercise to showcase how embedding can be used to assist data analysis on high-dimension data.