# Design Exercise 7: Exploratory Data Analysis with the ATUS data

in [Design Exercises](https://pages.graphics.cs.wisc.edu/765-22/exercises/)

October 12, 2022 (Last Modified: October 15, 2022)

In this assignment (actually, all the design exercises for the rest of the class), we will work with the American Time Use Survey - a data set that surveys thousands of people each year to see how they spend their time. Please read about it here: [**ATUS: American Time Usage Survey**](https://pages.graphics.cs.wisc.edu/765-22/datasets/atus/).

For this design exercise, we ask you to “explore” the data set. Your task is to identify interesting things in the data (that you might want to make visualizations of later). This task is Exploratory Data Analysis - and you will want to do some visualization. It’s just a different kind of visualization. You want to make quick, rough visualizations that show lots of things quickly so you can make judgements about where to did deeper or what stories to tell in more refined visualizations.

You will turn this in on canvas as [**DE07: Exploration (EDA) (due Tue, Nov 1)**](https://canvas.wisc.edu/courses/322540/assignments/1731359).

I’ll be open here: this assignment is an experiment. The idea of having you try to do Exploratory Data Analysis and work with a data set of this scale is important. Exactly how to make this an assignment is a challenge.

## Overview

The ATUS data is big in many ways. Over the years, it has included hundreds of thousands of people. It considers hundreds of categories of time uses. It considers dozens of variables about those people. At the same time, it isn’t too big to work with (you can easily load it into Python or Tableau). We’ve created some reduced versions for you, but you might prefer to work with the data straight from the source (see [**ATUS: American Time Usage Survey**](https://pages.graphics.cs.wisc.edu/765-22/datasets/atus/)).

In the past, we asked students to find “stories” (interesting things) in this data set and make visualizations to show it. This semester, we will break the process down into steps. First, (this assignment) we will ask you to **explore** the data to look for interesting questions to ask/answer (find the stories to tell). In a next phase we will ask you to make “good” visualizations that actually show these stories. Then, in the third phase, we’ll do some “vis research” to see what we might be able to do to address the challenges in working with this kind of data.

Note: For the purposes of this assignment, you do not need to worry about how the results generalize to the broader population. Your work can be descriptive, that is, you are just trying to describe what is in the sample (the set of people surveyed), not to figure out if what we see in the sample may be true across the entire population. The ATUS data actually has information to help do the fancy statistics to test generalizability, but we won’t consider that here.

To put the multiple comparisons problem backwards (see [**Readings 08: Why Does (or doesn't) Vis Work?**](https://pages.graphics.cs.wisc.edu/765-22/all-readings/readings08/)): with hundreds of variables (especially after joins) and tens of thousands of samples the odds of there not being at least some interesting patterns in the data is near zero. It might be a random chance - but we are just trying to describe what we see in the data. Think about this exploration as “let’s see what’s there, we can do good statistics later to see if we can actually draw a conclusion from it.” (I am not suggesting this is good practice, but it does fit this exercise for class).

## Exploration (this exercise)

Your goal in this exercise is to “explore” the data. In class, I was cautious about this as a task. But here, the real goal of exploration is clear: you want to try to better understand the data set to identify what are the interesting questions to ask (stories to tell). Looking ahead, you will want to make visualizations that tell interesting stories from the data: this phase you need to figure out what those stories are, and understand the data well enough that you can make “nice” visualizations later.

This exploration might take different forms… you might just look at the data in different ways that might make patterns stand out; you might want to pick a general question, and look at the data in different ways to see if you can refine it; you might start with a specific question, and look at the data to see if there is a broader picture. You can see an example of me exploring in [**A Quick EDA Example with the ATUS data**](https://pages.graphics.cs.wisc.edu/765-22/datasets/atus-eda/).

I strongly recommend you use Tableau for this. Yes, there’s a learning curve, but I believe that if you spend 2 hours with Tableau (and a lot of web searching to pick up specific tricks), then the next 2 hours you will be more productive, and ultimately get more done in the 4 hours than you would with 4 hours of trying to do this programmatically with Python (assuming you are already good with Python and some high-level plotting library). I will not require you to use Tableau.

Exploring data like this is a big, open ended exercise. Everytime I sit down to do it, I find myself getting sucked in to looking around more. (you can see some of my meanderings in [**A Quick EDA Example with the ATUS data**](https://pages.graphics.cs.wisc.edu/765-22/datasets/atus-eda/) - trying to write while doing it slowed me down a bit).

To make the exercise concrete, I’d like you to make sure you achieve the following (which you will need to hand in):

1. A list of good “questions” (things to show, stories to tell) that you would want to make visualizations from this data set. These should be things that you’ve found in exploring, so you know there’s something interesting, and sufficient data to show it.
2. Examples of where a visualization helped you find a good story to tell - even if the visualization doesn’t tell the story itself well.

I recommend that you do this with the following plan:

1. Come up with some initial questions - without looking at the data. (you should read the descriptions, and the meta-data to know what is in the data)
2. Think of what visualizations (or other data analyses) you might want to look at that would help you identify more and/or better questions.
3. Actually do some of exploration as you thought about in #2. There is a good chance that when you actually start looking at things, you will change what you want and create different things. You make make visualizations (or other analyses) that are based on convenience (it was easy to try X, or “while trying Y, I ended up making Z”) or interest (“when I saw X, I decided to look at Y”).
4. Come up with an additional set of questions based on #3.

I am separating #2 and #3 because I am assuming we arent’ expert data explorers. I think it is better to start with some ideas of what exploring we want to do, and then to see what is actually practical to do. It is also interesting to see how what you actually end up looking at deviates from what you initially thought you might look for.

Steps #3 and #4 probably happen together - you come up with questions while doing the exploration.

You will turn in the results of your exploration as [**DE07: Exploration (EDA) (due Tue, Nov 1)**](https://canvas.wisc.edu/courses/322540/assignments/1731359).

## What is a “Good Question”

For the purposes of this exercise, a “Good Question” has a non-standard (but still hand-wavy) definition: it will make for a good visualization.

I should also say that I am using the terms “Good Question” and “Stories in the Data” in a very similar way. The idea is to find something interesting in the data: you might phrase it as a question the data answers or a fact/pattern in the data (that would answer a question).

A good question/answer should be:

1. Interesting (non obvious, ideally it is something that the course staff who has looked at this data hasn’t already seen; uniqueness is valued)
2. Multi-variate (involve bringing together many variables)
3. Taking advantage of visualization (Generally, this means the answer is complex. But it could be a simple yes no, but the visualization would have value to give context, or support the answer, or help address followup questions, …)
4. Can actually be answered from the data (you could actually make a good visualization to answer the question)

A simple way to look at it would be to consider [**last year’s assignment**](https://pages.graphics.cs.wisc.edu/765-21/assigns/dc1/). You goal is to come up with questions/stories that would lead to good responses to this assignment (with the criteria it presents). You don’t need to make the resulting visualizations (yet).

Or, next week, you will be asked to make some interesting visualizations from the ATUS data set (where “interesting” visualization is defined as in last year’s assignment). What are the questions you want these visualizations to answer / the stories you want them to tell. The criteria for the **next phase** will be something like:

* Is the question/story interesting and clear?
* Is it multi-variate?
* Is the design effective? (is it well adapted to the story/task?)
* Do the details represent good choices?
* Is the design appropriate for the data?
* Is the rationale properly stated (in the documentation)
* Is the design complete (it has enough of a caption that it stands alone)?

One trick to doing well on last year’s assignment (it wasn’t a secret - it is stated in the assignment) was to do good exploration to find the good stories. This year, we have the current design exercise to make you do that.

Again - the visualizations you make for this exercise are the ones for exploration. They are things that are useful to you, the explorer trying to find the interesting story, rather than to the potential viewer that you will ulitimately try to tell the story to.

## What to turn in

The Canvas Hand-In [**DE07: Exploration (EDA) (due Tue, Nov 1)**](https://canvas.wisc.edu/courses/322540/assignments/1731359) will ask you for:

1. A list of initial questions that you have to start your exploration. These can be broad, vague, unconfirmed, …
2. A list of the visualizations that you want to make for exploring (to start with)
3. Some evidence from your exploration. I don’t expect you to make a full “log” like I did on [**A Quick EDA Example with the ATUS data**](https://pages.graphics.cs.wisc.edu/765-22/datasets/atus-eda/). A few pictures is probably OK, but it is better to put them into a document and say something about what they are. You can turn in a PDF or a ZIP file of images.
4. A list of new “questions” (or stories, or observations) that you think you would want to make visualizations for in the upcoming parts of the assignment. One challenge for you: try to come up with good questions that are unique - that your classmates will say “I wish I thought of that”.
5. An example of a visualization you made in your exploration that lead you to a question. This will be a picture, and an explanation of how the visualization lead you to something else.

We’ll also ask some other questions about the process.

## What will happen with this

Canvas will give you a check/no check for turning something in. We will do evaluation separately.

We will have some form of “class discussion” to share questions and refine things before we send you off to make visualizations.

Because we might discuss the results of this in class, the deadline for handing in the questions may be strict.

## Coming Attractions

The idea is that we will use this one data set for the rest of the semester. You will work with it in many ways. So, it is useful for you to get used to it.

The general strategy: I want you to experience working with a hard data set, and then we can do projects that try to address some of the challenges. Until you’ve seen the challenges, its harder to come up with solutions.

So the plan: (1) explore the data, (2) make visualization with it, (3) propose projects that address some of the challenges we encountered, (4) do a project where you push the boundaries of the state of the art.

It may be useful for you to see how this assignment fits in with the rest of the semester:

* [**DE07: Exploration (EDA) (due Tue, Nov 1)**](https://canvas.wisc.edu/courses/322540/assignments/1731359) - You will do some explorations with the ATUS data set.
* [**DE08: Story Building (due Tue, Nov 8)**](https://canvas.wisc.edu/courses/322540/assignments/1731360) - You will decide some visualizations to make with the ATUS data.
* [**DE09: Story Telling with Data (due Fri, Nov 18)**](https://canvas.wisc.edu/courses/322540/assignments/1731361) - You will make (good) visualizations showing the things you found.
* [**DE10: Project Proposals (due Tue, Nov 22)**](https://canvas.wisc.edu/courses/322540/assignments/1731362) - You will critique visualizations, and think of how tools might make working with this data better.
* [**DE11: Projects and Re-Submissions (due Tue, Nov 29)**](https://canvas.wisc.edu/courses/322540/assignments/1731364) - You will propose a project that supports working with this kind of data (and improve your DE9 visualizations based on the critiques)/
* [**DE12: Evaluation Exercise (due Tue, Dec 6)**](https://canvas.wisc.edu/courses/322540/assignments/1731365) - You will do an exercise about how we evaluate visualizations - this is meant to reinforce some class content, but we’ll apply it to the project.
* [**DE13: Project Handin (due Tue, Dec 13)**](https://canvas.wisc.edu/courses/322540/assignments/1731367) - You turn in your project.

1. A list of initial questions that you have to start your exploration. These can be broad, vague, unconfirmed, …
   1. Do people with household children tend to sleep less? If so, does it differ across sex? For example, if people with children indeed sleep less, is the difference in sleeping time varies across fathers and mothers?
   2. More generally, is there a difference in who people spend their time between those who have household children and those who don’t? Does such a difference vary across sex?
   3. For those who have household children, does the number of household children also make a difference on how people spend their time?
   4. Is there a difference in how people spend their time during the peak of the pandemic (e.g., 04/2020-12/2020) compared to the year prior to the pandemic (e.g., 04/2019-12/2019)?
   5. Does the impact of having household children change after the pandemic happened? For example, if having household children reduce the sleeping time even more compared to before the pandemic?
2. A list of the visualizations that you want to make for exploring (to start with)
   1. A stacked bar plot with two bars (the bar area encodes the amount of time for each activity), where one is for those with household children and the other one for those without.
   2. One bar plot with N bars (the bar area encodes the difference in time between those who have children and those who don’t) where each bar is for one activity.
   3. A stacked bar plot with four bars (the bar area encodes the amount of time for each activity) for the four subgroups – male/female with/without children
   4. A (unstacked) bar plot with four bars (encoding the “sleeping time”) for the four subgroups – male/female with/without children
   5. A line plot with two lines (x-axis: number of children, y-axis: sleeping time) for two groups – male and female
   6. A stacked bar plot with three bars (the bar area encodes the amount of time for each activity), where one is for 2019 (pre-pandemic), another one 2020 (pandemic), and another for 2021 (pandemic)
3. Some evidence from your exploration. I don’t expect you to make a full “log” like I did on [**A Quick EDA Example with the ATUS data**](https://pages.graphics.cs.wisc.edu/765-22/datasets/atus-eda/). A few pictures is probably OK, but it is better to put them into a document and say something about what they are. You can turn in a PDF or a ZIP file of images.

See “Design Exercise 7 Exploratory Data Analysis with the ATUS data\_Q3.docx”

For #4 in the description (the list of questions): give a list of the "good questions/stories" you came up with. 5-10 is the target (if you have more than 10, give the 10 best).

If you have "less good questions" - please put them in the next type-in.

For these questions, we want questions/stories that will lead to interesting visualizations for the later assignment. (see the notion of "good question" in the description). Additional measures of goodness are "unique" (no one else thought of it), "surprising" (you'd expect most of us in class to be surprised by the story), and well supported (it actually something that you can find in the data).

Please use the text format as described at the top.

1. To my surprise, the main difference between those who have children and those who don't is not in the sleeping time. In fact, they have similar sleeping time.

2. The key differences between those who have children and those who don't are in 1) "Television and movies (not religious)" (less time for those with children), and 2) "Work, main job" (more time for those with children).

3. Based on the previous discovery, I further split the data by sex and found that those the effect is stronger for male than female. That is, if a man had children, then he will spend much less time on TV and much more time on work than a woman would change.

4. Does having more children make the effect I found above even stronger? Roughly yes, as shown in the line plot, having more children seem to reduce the TV time and increase the work time. However, when one has more than 3 children, having more children seem to reduce men's work time but somehow continue increasing women's work time.

5. To my surprise, the time spending pattern of the entire population does not seem to change too much across year 2019, 2020, 2021, which corresponds to the pre-pandemic period and pandemic period.

6. However, if we break down the population by sex and whether having children, there emerges an interesting pattern. First, for men who do not have children, the working time seems to decline after the start of pandemic, whereas for men with children, the working time seems to increase. Second, for women, regardless of whether they have children or not, the working time steadily increase from 2019 to 2021.

Give an example of an "exploratory" visualization that lead you to a question. Give both the picture, and the question that lead to it (some explanation might be helpful). This is similar to part 3, but it is asking for a specific instance. Please embed an image in the type-in box (as you do for seek and finds).

I am interested in the difference between those who have children and those who don't. Therefore, I create this plot, showing the difference in time spent (minutes) for each activity between those who have children and those who don't (those who have minus those who don't). The figure shows that the top two activities are 1) working and 2) watching TV. I therefore focus on these two activities (aside from sleeping, which was my initial guess) when doing subsequence visualization (e.g., looking into the gender difference, looking into how it changes before and after the pandemic)