Data Analysis and Visualization

Assignment 01
Hands-on Assignment Instructions

Frist, it makes sense to create a private folder structure for the course, e.g., as studies/2023/dav/week02-topic01-intro/ etc. Then, before actually submitting solutions via Moodle, you can write draft solutions and notes in separate files in the weekly folders. This helps keeping your data organized, and it is easy to come back to the old solutions.

Some exercises include very specific tasks. Some other exercises are more open --- with different kinds of solutions --- and require thinking and reflect on certain topics. Note that in a real-world (vague) setting, tasks seldom have clear-cut solutions, so it is important to develop analytical eye (what is this about, what are the essential/relevant bits, what kind of solution is acceptable, and how can I verify my solutions).

It usually helps to consider new topics with respect to applications that interest you, even if not mentioned in the assignment description (e.g., social media applications, computer games, or specific applications at your workplace). Also, many of the exercises are quite small --- the idea is that you can think and try out things by yourself (later on), brains/hands-on.

The point of assignments is to study and learn. Since studying can be more fun and effective with others, **be active and discuss the exercises with others, using the MS Teams group channels**. **It is thus advisable and perfectly ok to collaborate with the assignments, even if everyone must eventually complete the exercises in solo**. For instance, when starting to work with assignments, check your group from Moodle and write "Hi!" for chat for the channel for starters --- or even better, start a video call within the group channel (without a specific "agenda" etc., just to note others that you are currently actively working with the exercises). Recall that you may ping your mate(s) or the teacher, starting with the @ sign, e.g., as @ossi.nykanen (the teacher); this ensures that the person notices your message. Be brave and start discussions, ask questions, and help others. All communications should be constructive, we are here to learn, and it is perfectly ok to make mistakes --- let us study and discuss the solutions together.

Note that the above collaboration strategy works nicely even in a f2f classroom setting, since using Teams, one can more easily share/zoom screen, etc.

Then to the actual exercises:

E01.1. Install free Tableau Public software to your computer:

Unless you already have, get yourself a free Tableau Public account (Sign up) and install Tableau Public from

https://public.tableau.com/en-us/s/

(Notice that there is also the commercial Tableau software version, which we do not use in this course. The commercial version offers a limited trial.)

Do not use any existing, important password for the account.

Note that by default, Tableau Public saves your projects visible to your public Tableau Public profile: Do not process nor save sensitive information with Tableau Public. (You can hide your test projects from plain sight via your profile settings --- but the data still remains in the public server.)

After sigiing in, you have a public Tableau Public profile to which you can save visualizations. You can also load others' visualizations from the Web.

E01.2. Check the Tableau Public tutorial videos "Get Started" and "Connect to Data" from https://public.tableau.com/en-us/s/resources . Following the steps from the video, create a similar visualization using Tableau Public. You can load the Excel file which is the dataset either from the video page, or from Moodle (check the Examples and Data folder in Moodle)

Save your visualization to your public Tableau Public account, and add the link to your solution submission later.

Note: The videos use the online version of Tableau public, not the desktop version installed above. At first, it is convenient to follow the videos and use the online version --- and switch to the desktop version later when appropriate (at least when there are service issues).

E01.3. Check the Tableau Public tutorial videos "Navigate the Workspace Area" and "Map: Profit Ratio by Geography." from https://public.tableau.com/en-us/s/resources . Following the steps from the video, create a similar visualization using Tableau Public.

Save your visualization to your public Tableau Public account, and add the link to your solution submission later.

E01.4. Data analysis is based on the assumption that we know --- or are willing to learn about --- the application and the data we are working with. To get a flavor of the structure of typical datasets, let's get familiar with perhaps the most widely known dataset: Fisher's

iris data.

Load the dataset from: https://archive.ics.uci.edu/ml/datasets/Iris (notably the files iris.data and iris.names)

Check the main points from: https://en.wikipedia.org/wiki/Iris_flower_data_set

Open the files in your favorite text editor (such as Notepad++; https://notepad-plus-plus.org/downloads/), and write brief answers to the following questions:

1. What is the data all about?
2. How big is the dataset? (Number of rows or instances.)

3. How was the data collected?
4. What does an individual dataset row mean?
5. What do the attributes (columns) mean?
6. (Based on this example:) How is data encoded in .csv files?
7. Which use cases you can find for the dataset, from the wikipedia article?
8. How precise you think the data is? What kinds of errors might occur in a dataset like this, collected like this?

Let us then take a deep dive with Tableau Public, to get a feel how it works in principle. Don't worry if you can't yet fully complete the next exercise E01.5: We will come back to the basics of Tableau next week, and you can come back to this exercise and complete it still next week.

E01.5. Using Tableau Public, create a new visualization using the iris dataset. Create simple viz which looks like the following:
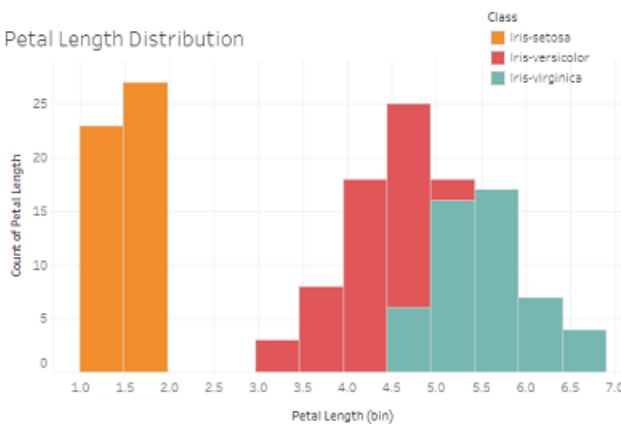
Example

Visual Tests with Iris Data

The following visualizations demonstate comparing the average values of attributes and illustrating the value frequencies of a single attribute.

Average Attribute Values

Petal Length Distribution

Tip: You might wish to use the iris.csv data file available via Moodle. (This includes the attribute names as headers, making things more intuitive at first.) Note that additional technical tips may be found below.

Save your visualization to your public Tableau Public account, and add the link to your solution submission later.

What do you think the visualization tells about the data? What kind of challenges you had when creating the viz? What kind of issues you had when completing the exercise? How did you overcome the issues?

Do not worry if you do not understand all Tableau functionality or know how to handle datasets in general --- this first exercise is a "teaser" where we take the leap of faith and jump directly to making visualizations. In the following sessions and assignments, we'll then consider the different areas of work in more detail.

Also, note also that now, we just "played around" with visualizations to get familiar with Tableau --- without any specific question or business problem about the Fisher's iris data set. Quite obviously, **in real-world applications, analysis work begins with actual domain-specific business objectives and the biggest task is compiling and pre-processing the data**. Well, we just had to start from somewhere: We'll come back to the big picture of analysis objectives and tasks soon, too.

Technical tips:
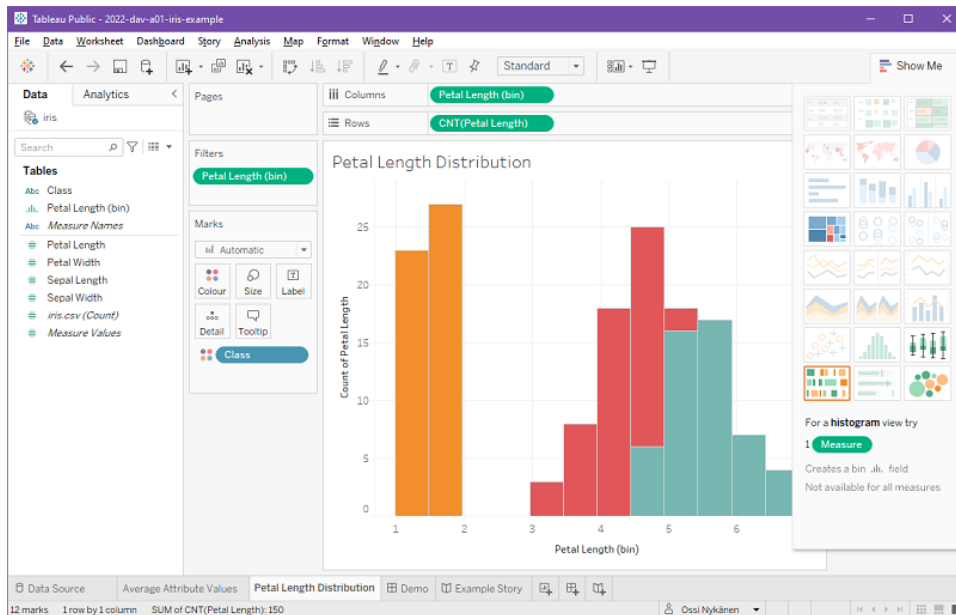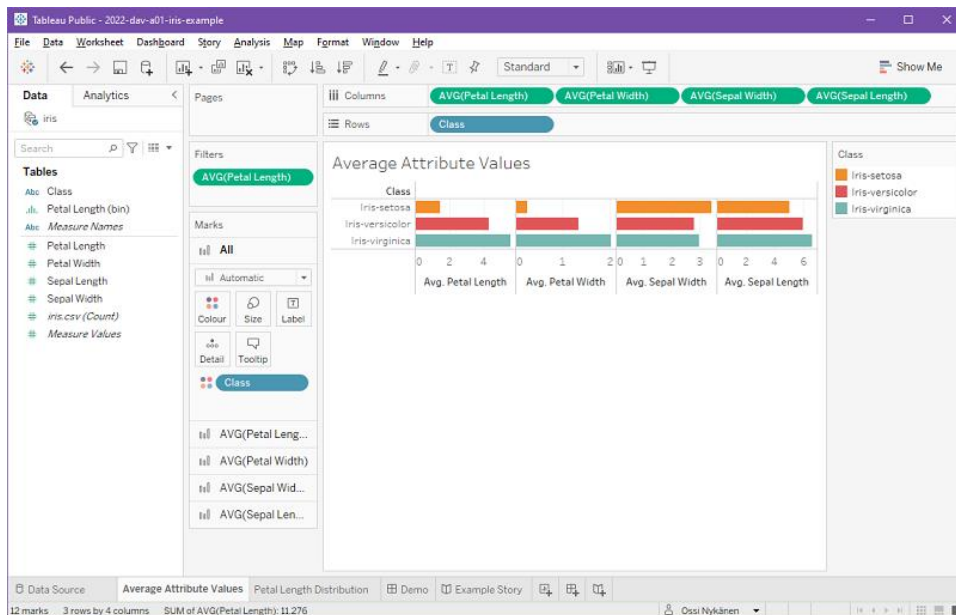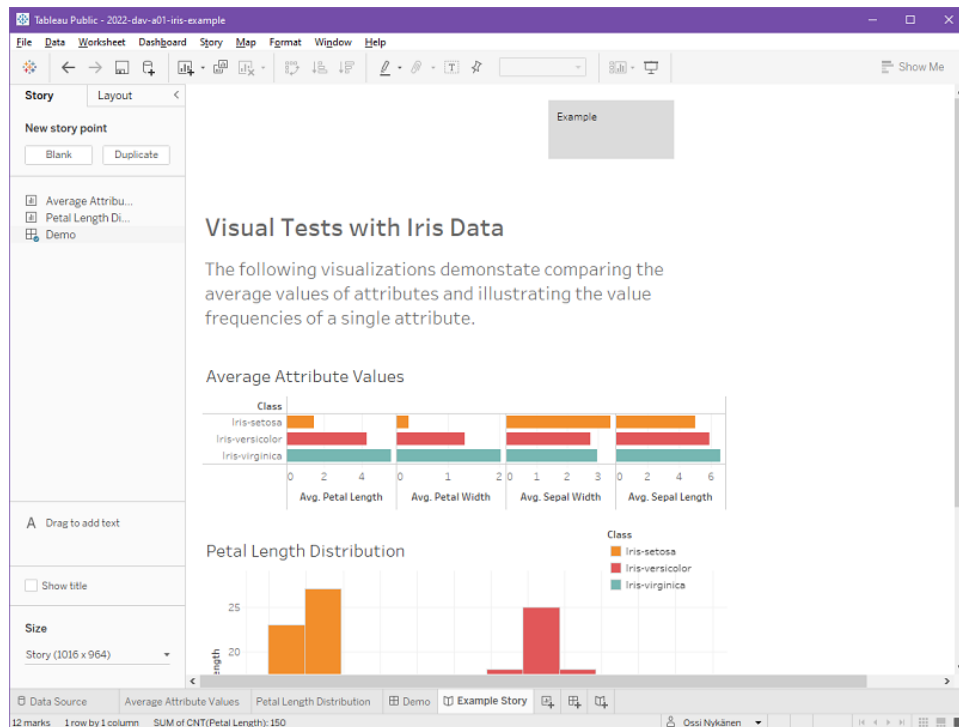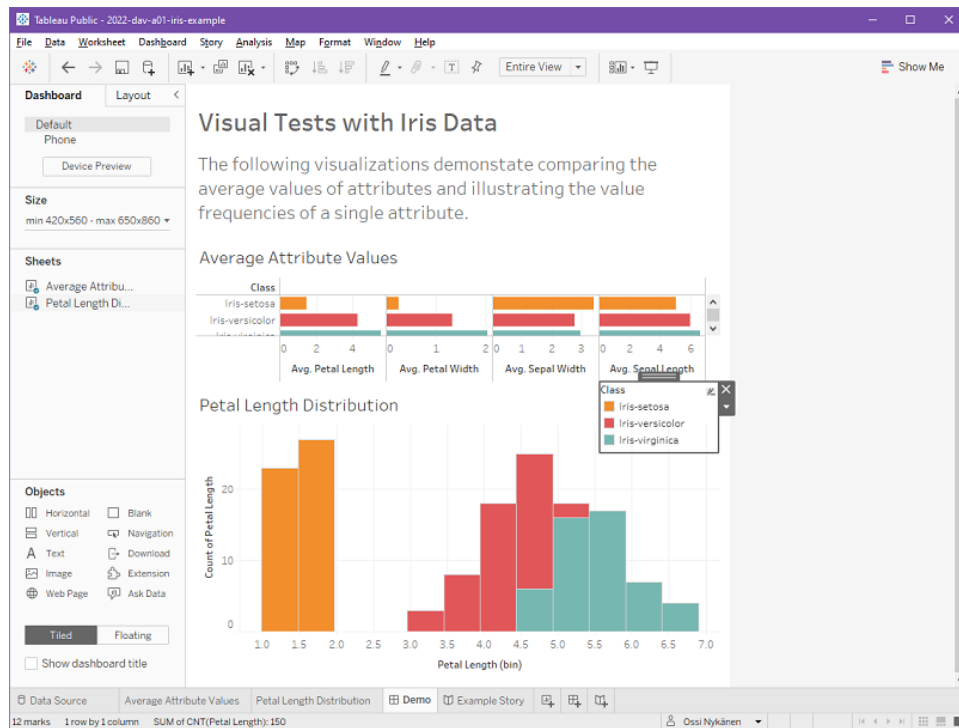
Once you have completed the exercises of this assignment:

- **Submit** (brief notes, code(s) or link(s) about) your solutions to the correct Moodle discussion forum (start a new discussion; check the instructions for more details and how to update already submitted solutions, etc.); and
- **Write** peer feedback to two other students' solutions in the correct Moodle discussion forum (check their solutions and add a brief comment as a reply, written in a constructive tone).