Assignment 1

Welcome to the visualization course! To understand the difficulties and challenges that visualization researchers face today, and the visual design process, we are going to build a visualization tool during the course. Generally, there are various application fields that generate different kinds of data sets, typically recorded over longer time periods. In the lecture, you will see different data types that can be primitive or more complex, but also static or dynamic. Moreover, the data sets can be combinations of several data types. However, we will focus on a specific scenario which is the visual analysis of high-dimensional tabular data.

Assignments will NOT be evaluated, they are meant for you to have a guideline toward the final project result

Exercise 1 – Data Set

This year you can choose one of the below data sets:

- **Australian Shark Incidents**
  https://github.com/cjabradshaw/AustralianSharkIncidentDatabase

  The Australian Shark-Incident Database contains comprehensive reports of 1,196 shark bites that have occurred in Australia over 232 years (1791–2022). Data were collated by the Taronga Conservation Society Australia using purpose designed questionnaires provided to shark-bite victims or witnesses, media reports, and information provided by the department responsible for fisheries in each Australian state (including the Northern Territory). The dataset includes provoked and unprovoked bites from fresh, brackish, and marine waters in Australia. Data span 22 suspected shark species. This dataset will be publicly available, and can be used by analysts to decipher environmental, biological, and social patterns of shark bites in Australia. The information will aid scientists, conservationists, authorities, and members of the public to make informed decisions when implementing or selecting mitigation measures.

- **Railroad Equipment Accident/Incidents**
  https://data.transportation.gov/Railroads/Railroad-Equipment-Accident-Incident-Source-Data-F/aqxq-n5hy/about_data

  Since last August, the Federal Railroad Administration has been rolling out a new portal for its safety data. Through it, you can find datasets on incidents and accidents involving railroad equipment, incidents at grade crossings, and reported injuries and illnesses, as well as dashboards and reports on related topics. The grade crossing dataset, for instance, lists 246,000+ incidents since 1975; it indicates each incident's date, railroad, crossing identifier, nearest station, number of injuries, vehicle and train speeds, and much more.

- **Work-related Injury and Illness**
  https://www.osha.gov/Establishment-Specific-Injury-and-Illness-Data

JBI100 Visualization

Anna Vilanova & Stef van den Elzen

The US Occupational Safety and Health Administration requires many (but not all) businesses to track employees' work-related injuries and illnesses. Larger companies and those in high-risk industries must electronically submit annual counts to the agency. Thanks to freedom-of-information lawsuits by Reveal and Public Citizen, OSHA began to publish business-level data from those electronic submissions in 2020. The records, which go back to 2016, include each business's name, location, industry, employee count, and employee hours worked, plus their reported number of deaths, injuries, skin disorders, respiratory conditions, poisonings, hearing loss, and other illnesses.

You can find the prepared datasets on Canvas. Each dataset is organized in a separate folder and consists of possibly multiple CSV files. An example Jupyter notebook can be found in the 'Notebooks' folder. This notebook can be used for initial data exploration and analysis. Data dictionaries and additional information can be found in the respective folders.

You are welcome to enhance these data with other data sets from other sources to achieve interesting and meaningful analysis. However, they are considered extra and should not be an alternative to choosing one of the proposed data sets.

Our goal is to design a visualization tool for high-dimensional data to achieve specific goals/tasks.

One important choice is to identify the goal/users that will be the focus of your visualization design. Notice that not all possible goals are suitable for a visualization solution.

One of the first jobs is to get familiar with the data set and the domain. Overseeing the structure, size, potential, and challenges of a given data set and the domain is one of the key problems in visualization.

(a) What is the information you can obtain from the data set/ data sets?

(b) What are the attributes in the data and what is their meaning?

(c) Write a small parsing function that can read the data position (column, row) from the file format you selected.

(d) Write another function that outputs the distribution of the attributes, and counts the frequencies of the different values.

(e) Try to describe the data set in just a few sentences. How is the data provided? Which kind of attributes are contained in the data set? How large is the data set in terms of the number of those elements (teams, matches, players, historic data, extra records, and so on)?

(f) Analyze the errors and missing values. Write a function to count how many missing values per attribute and per entry you have. Analyze *what are the most relevant missing values that might hinder the analysis according to you.*

From this exercise on, we give some initial steps and you should start writing the corresponding sections of the interim report. Please, read the final report information provided in canvas. The assignments give a guideline, notice that extra points to what is mentioned in the assignments are needed to have all aspects of the interim report covered.

Exercise 2 – Goal - Data (Domain specific)

We will be following the nested model presented in the lectures for the visualization design process. So the first step is to understand the domain situation and formulate the goal of the visualization. You need to identify by yourself what user and goal you want to work on. We are in a visualization course so the main goal should suit a visualization solution.

(Introduction) Describe what you envision will be the general overall goal and users of the visualization tool. The goal is meant to be from the perspective of the user, which will be the goal/question from the user's perspective to use the visualization tool. Think about the different goals of visualization presented in class and the high-level actions. Define for which users your tool is meant, and which overall goal. The reason why this goal is suitable for the available data and why this is a goal where a visualization tool is the right means to solve it (e.g., visualization vs. an automatic solution).

Exercise 3 – Data (What) Domain specific

(a) Write in section *What (Data)* the description of the data. You can base it on the analysis you have done in exercise 1. What are the general properties of the data you want to use?

(b) Most of the data sets contain noise, missing data values, and relations, or measurement errors. The data of this course is no exception. In exercise 1, you already looked at the missing values. How will you handle missing data values or measurement errors? Think of multiple ways and their pros and cons.

(c) (Data (What)) Choose one of the methods and implement it for the data set. Describe it in the section and mention what is the effect on the data.

Exercise 4 – Data (What) Abstraction

Once the goal and data are understood from the domain point of view. We enter into the abstraction phase such that we can identify the most adequate designs later on.

Make the data abstraction according to the "what" in Munzner's framework. Present it in a summarized version you do not need to present it for each individual attribute. Build a table with a general overview.