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. 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 towards the final project result

You have to choose one of the following two data sets for your project. You can find the data sets in canvas Files/Data Sets 2022 FIFA world cup and Credit Score:

2022 FIFA world cup.

During the 2022 World Cup, 32 teams from around the world competed for the trophy. The series of data sets capture most of the action, from player statistics, team standings, game scores, and match performances, to Twitter data, player images, and penalty shootouts.

- FIFA World Cup 2022 Player Data (https://www.kaggle.com/datasets/swaptr/fifa-world-cup-2022-player-data)
- FIFA World Cup 2022 Match Data (https://www.kaggle.com/datasets/swaptr/fifa-world-cup-2022-match-data)
- FIFA World Cup 2022 Team Data (https://www.kaggle.com/datasets/swaptr/fifa-world-cup-2022-statistics)
- FIFA World Cup 2022 Twitter Dataset (https://www.kaggle.com/datasets/kumari2000/fifa-world-cup-twitter-dataset-2022)
- FIFA World Cup 2022 Prediction (https://www.kaggle.com/datasets/shilongzhuang/soccer-world-cup-challenge)
- FIFA World Cup 2022 Player Images (https://www.kaggle.com/datasets/soumendraprasad/fifa-2022-all-players-image-dataset)
- FIFA World Cup Historic (https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup)
- FIFA World Cup Penalty Shootouts (https://www.kaggle.com/datasets/pablollanderos33/world-cup-penalty-shootouts, https://www.kaggle.com/datasets/jandimovski/world-cup-penalty-shootouts-2022)

Each data set is organized in a separate folder and consists of multiple CSV files. An example Jupyter notebook can be found in the 'Notebook' folder. This notebook can be used for initial data exploration and analysis. Data dictionaries and additional information can be found in the respective folders.

JBI100 Visualization

Anna Vilanova

Credit score

https://www.kaggle.com/datasets/parisrohan/credit-score-classification

This data set is supplied by *Rohan Paris* and contains a data set with credit related information of people that ask for a credit. The credit score is an indication of how likely it is that somebody returns a credit. The data set is meant for the development of machine learning models for credit score prediction. However, our goal is to analyze the data from various aspects which are suitable for a visualization solution. The data set has multiple aspects that can be explored.

- 'all_data.csv' contains all data set items together
- A small python script to load the data in a data frame: 'load_data.py'

See Kaggle webpage for details on the attributes and values. If there are ambiguous or unclear values/meanings, you can give it a meaning that you find most probable. Just describe your assumptions in the report.

Exercise 1 – Data Set

You are welcome to enhance the data proposed 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 (listings, reviews, vehicles, geographic regions and locations, 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.

(g) Think about possible solutions for the missing/error values and propose a solution arguing your choice.

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. 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 the general overview.