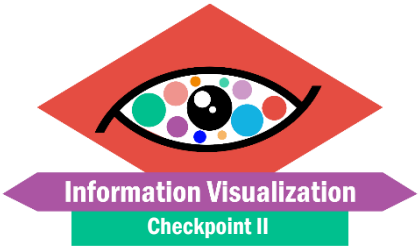

Checkpoint II: Data Cleaning & Processing

Group: G18
Date: 2023/09/22

Initial Dataset

The initial dataset is composed of information collected as part of a survey regarding mental health among gamers worldwide. The data collected is tabular (a single table with 13.464 items x 55 attributes, for a total size of 740520), and its attributes may be organized in the following sections:

4 (survey info) + 8 (demographic) + 9 (general anxiety disorder scoring) + 6 (satisfaction with life) + 17 (social phobia disorder scoring) + 1 (narcissism scoring) + 5 (gaming habits) + 4 (play motivation)

The current dataset is of size (#Items x #Attributes): 13.463*42 (main dataset) + 7*2 (GAD test questions) + 17*2 (SPIN test questions) + 4*2 + 5*2 (GAD and SPIN test options) = 565512

```
(Data sample from "GamingStudy_data.csv")

S.No; Timestamp; GAD1; [...]; GADE; SWL1; Game; Platform; Hours; WhyPlay; League; [...]; Streams; SPIN1; Narcissism; Gender; Age; Work; Degree; Birthplace; Residence; Reference; Playstyle; Accept; GAD_T SWL_T; SPIN_T; Residence_ISO3; Birthplace_ISO3

1; 42052; 0; [...]; Not difficult at all; 3; Skyrim; "Console (PS, Xbox, ...)"; 15; I play for fun;  N/A; [...]; 0; [...]; 1; Male; 25; Unemployed / between jobs; Bachelor(or equivalent); USA; USA; Reddit; Singleplayer; Accept; 1; 23; 5; USA; USA
```

Selected/Derived Data & Data Abstraction

The new dataset is static, partly tabular and partly multidimensional table (for matching attribute codes with their – ratio – string equivalents); its items are players. Averages and medians of the test scores and symptom prevalence (ratio attributes) per age and/or gender and/or playstyle are derived from the original data by d3, via grouping by relevant attributes and calculating the needed value.

Category	Data	Semantics
Ratio	Hours Age GAD_T SPIN_T	Number of hours played weekly on most played game Age of the player Total score for General Anxiety Disorder (GAD) Total score for Social Phobia Inventory (SPIN)
Nominal (categorical)	Gender Work Residence	Gender of the player Employment status (e.g., Employed, Unemployed). Country of residence
Nominal (within domain context)	Playstyle [1-3 attr.] WhyPlay [1-6 attr.]	Style of gameplay (e.g.:, Multiplayer - Yes, Online - No) Reason for playing
Ratio	SPIN1-SPIN17 GAD1-GAD9	Symptoms score for Social Phobia Inventory (SPIN) Symptoms score for General Anxiety Disorder (GAD)

Data Processing

Data selection, sorting, and computation was done through python and its Pandas library. Attributes of sections (comparatively) less interesting/relevant to the problem domain – or whose influence would otherwise likely be superseded by others', like birthplace against residency – were discarded. The attributes deemed most relevant were directly related to general anxiety disorder and social phobia, as well as gaming habits, motivation, and widely used demographic attributes (e.g.: gender).

We have secondary tables, which match GAD and SPIN question/answer ids to their respective string, for better visualizer understanding. A sentinel value of "-1" was used for missing or "NA" numerical ordinal attributes, and "Undefined" for missing or "NA" nominal or ordinal non-numerical values. The standard deviation method was used for identifying and discarding outlier ordinal values (as we are trying to find trends). Playstyle and motivation attributes, as they corresponded to survey portions with an open-ended option (which many responders used), had their values re-computed to fit one or multiple of the available answers in their respective portion of the survey, and had their value options turned into new columns with a simple "Yes"/"No" for possible values to account for multiplicity; in Playstyles case, their options were split into three columns: multiplayer, online, and relation (who they play with).

Mapping (Data sample/Questions)

Question	How to use it (d3)
Do people who play online with friends tend to experience less social phobia than people who do so with strangers?	Regarding the "Playstyle" attributes' section: group the rows where the "Multiplayer" and "online" attributes have value "Yes"; group those into two other groups: one where the "Relations" attribute has value "friends", and another where the value is "strangers". These sections' SPIN scores may then be compared.
How does the prevalence of social phobia symptoms vary throughout countries across different player ages	Group the data by "Residence". Calculate and show the percentage of players median symptom values for each symptom (SPIN1 to SPIN17) and residence within a specified age range and compare
Do players who are unemployed play online multiplayer games with strangers more often than those who are not?	Group the data by attribute "Employed" and the several Playstyle attributes. Sum the amount of hours playing online with strangers for both employed and unemployed players; compare.
Given a similar number of hours played, do victory-motivated players have less general anxiety disorder symptoms than fun-motivated players?	Select the players that have played a number of hours within a specified range; group the data by the Whyplay attributes "Winning" and "Fun" and sum the amount of non-"Not at all" answers on the GED symptoms' attributes for each motivation category. Compare these. We may also calculate and compare the average "GAD_T" for these groups instead.
How do levels of social phobia change between genders, for each playing style?	Group the data by "Gender", and by "Playstyle" attributes' values. Calculate and compare the median "SPIN_T" for each combination of gender and playing style, or order by "SPIN_T" and compare.