Checkpoint II: Data Cleaning & Processing

Group: G18
Date: 2023/10/02

Initial Dataset

The data is tabular (a table with 13.464 items x 55 attributes + the SPIN survey questions, possible answers, and scoring correspondence at 17 items + 5 items x 2 attributes + 5 items x 2 attributes, for a total size of 740.574), and its attributes may be organized in the following sections:

4 (survey info) + 8 (demographic) + 9 (GAD scoring) + 6 (satisfaction with life) + 17*2+5*2 (SPIN scoring) + 1 (narcissism scoring) + 5 (gaming habits) + 4 (play motivation)

The current dataset is of size (#Items x #Attributes): 13.463*34 (main dataset) + 17*2 (SPIN test questions) + 5*2 (SPIN test options) + 5 * 3 (SPIN test result interpretation) = 457.801

```
(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; US
```

Selected/Derived Data & Data Abstraction

The new dataset is static, partly tabular and partly multidimensional table (for matching attribute codes with their string equivalents); its items are players. Averages and medians of the test scores and symptom prevalence (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	Number of hours played weekly on most played game Age of the player
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
Continuous, sequential	SPIN1-SPIN17 SPIN_T Answer Id, Meaning Min/Max Score (Static)	Symptoms score for Social Phobia Inventory (SPIN) Total score for Social Phobia Inventory (SPIN) Result for SPIN1-17 and SPIN_T Minimum and maximum intervals for a given SPIN_T score

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 SPIN question/answer ids, and test result intervals to their respective string, for better visualizer understanding, and which were derived from both the survey and a psychology website. 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. The motivation data had its value options turned into new columns with a simple “Yes”/”No” for possible values to account for multiplicity. In the Playstyles case, their options were split into three columns: multiplayer, online, and relation (who they play with). Relation has the options “alone”/”friends”/”strangers”/”online acquaintances” and the option “various” in case more than one of the previous options is applicable while Multiplayer and Online have as possible answers “Yes”/”No”/”Both”; “Both” is applicable when the player plays online games as well as offline games for example.

Mapping (Data sample/Questions)

(Data sample from “clean_data.csv”)

```
Hours; SPIN1; [...]; SPIN17; Gender; Age; Work; Residence; SPIN_T; Multiplayer;
Online; Relation; Whyplay_winning; [...]; Whyplay_habit;

18; 2; 1; Male; 19;  Student at college; Sweden; -1;no; no;alone; No; [...]; No
```

Table	spin_questions.csv	spin_answers.csv
Attr.	Question Id; Meaning	Answer Id; Meaning
Sample	SPIN1; I am afraid of people in authority.	0; Not at all

Question	Attributes/Tables used
Do people who play online with friends tend to experience less social phobia than people who do so with strangers, in most countries?	Relations, SPIN_T, Online, Residence, *spin_results*
How does the prevalence of social phobia ailments vary throughout countries across different player ages?	SPIN1-17, SPIN_T, Residence, Age, *spin_results*
Do players who are unemployed play online multiplayer games with strangers more often than those who are not?	Work, Online, Relations, Multiplayer, Residence
Given a similar number of hours played, do victory-motivated players have less social phobia incidence than fun-motivated players?	Hours, Whyplay_winning, Whyplay_fun, SPIN_T, SPIN1-17, Residence
How do levels of social phobia change between genders, for each playing style?	SPIN_T, *spin_results*, Gender, Whyplay[all], Residence