1ST capstone project

SPEED DATING DATASET

DATA WRANGLING PHASE

Author: Tais Pancier

DATA WRANGLING

1) Data Dictionary:

The following steps were performed to build the data dictionary:

- read csv file: Speed_Dating_Data.csv
- identify column names (95 columns)
- determine data type for each column
- min and max values for each column
- identified NaN in some columns due to the data type defined by Pandas (float instead of integer)

DATA WRANGLING

2) Data Cleaning:

- Dataset initial shape: 8378 rows, 95 columns
    - Each row represents a speed date between 2 participants
- Dataset shape after cleaning: 8378 columns, 69 columns

The following steps were performed to clean the dataset:

- identified columns that wouldn’t be used in the analysis and could be dropped (26 in total)
- changed 17 columns to categorical
- Rating scale was different depending on the date/time of event. These columns were normalized by dividing each rating value by the sum of the related ratings (row-wise)
- Outliers:
    - 1 rating record was higher than the max of the scale for the column. As it was 1 record and just 1 point higher than the top of the scale, no special treatment was performed. This value was normalized as well

DATA WRANGLING

2) Data Cleaning (cont.):

- Missing values:
    - 19 columns with missing values were dropped because they were not relevant for the analysis
    - pid column (partner id): missing data found based on another id field and date/time of event
    - Attribute rating columns: missing data represented attributes that were not important for participants on each dating or they were not sure. NaN replaced with zero to allow correlation analysis
    - NaN in columns that might not be important for the analysis were left as NaN and will be addressed if these columns become necessary

Jupyter Notebook: "Speed Dating Dataset - Data Wrangling"