Project Milestone ReadMe

Ringmasters - Simon Kim, Jahleel Murray, Noah Shen, Wen Xi

We want to investigate the dating profiles to see if we can find some fun or surprising relationships between any of the variables that represent different characteristics of the people to whom these profiles belong.

EDA&Exploration: Contains all our Tableau files from our data exploration phase

- Diet-bodytype.twbx - Checks if there's a correlation between diet and body type
- Ringmasters_okcupid_pets_religion.twbx - Explores the data with cleaned pet and religion variables. Checks if there's a correlation between pets and religion as well as religion and ethnicity.
- agecomparison.twbx - Shows the distribution of ages within the dataset
- drugs_smokes_drinks.twbx - Is a heatmap looking to see if there is any correlation between people who do drugs, smoke, and drink.
- ethnicities.ipynb - Shows the distribution of ethnicities in the dataset
- gendercomparion.twbx - Shows the count of each gender within the dataset
- signs.ipynb - Shows the distribution of astrological signs in the dataset.

ML

- decisiontree - Our decision tree implementation attempts to determine whether age, body type, status, drugs, drinks, pets, height, and diet can determine the astrological sign of somebody. It is often said that people within an astrological sign share similar attributes, this decision tree model investigates if this notion is relatively true.
- knn - Folder shows our application of k-nearest neighbors. We use one-hot encoding to separate our categorical data for this application. We look into relationship status, gender, sexuality, school, occupation, and offspring in order to determine two different ordinal variables: age and income.
- regression
  - age-height - Age and Height columns are numerical, hence we use linear regression to age-height to see differences in height between generations.
  - age-pets - We separate people in two types for binary logistic regression: like cats or dogs. We are interested in which is the popular pets in different age groups.
  - height-sex - We use logistic regresstion to see where the line in height between male and female draws.
- nlp
  - Ham_dist_2.py - uses the Levenshtein distance algorithm to generate quantitative distance values between the essays present in the okcupid profiles.
  - nlp_parts_of_speech_tool.ipynb - Is a proof of concept tool we plan to use in our future NLP machine learning investigation. This code allows us to parse specific

parts of speech from the essay portions, we will eventually analyze in seeing what people enjoy doing as activities.