Tasks to do

* Data Cleaning/Processing
* Data Analysis
* Visualization
* Create Model

Jonathan Liu

* Organized meeting times, hosted zoom calls, and led discussions
* Created initial drafts for all study docs (proposal, presentation, report, code, task outline)
* Provided framework of key ideas for all steps (cleaning, analysis, visualization, models)
* Met with TA Alan Varkey to obtain project guidance
* Report - abstract, results, discussion
* Presentation - flow, slides, script
* Code - normality tests, correlation tests, revised models, general proofreading

Sameer Mohammad

* Contributed key ideas and writing for initial and final proposals
* Performed data analysis and data visualization

Bhargav Varidi

* Contributed key ideas and writing for initial and final proposals
* Performed data cleaning and preprocessing.

Teja Pavani Jyesta

* Assigned distinct responsibilities to team members based on their original roles in the project proposal.
* Performed machine learning(clustering)

Lavanya Ranganatham

Pallavi Vandanapu

Key Tasks:

Data Cleaning

* Extra considerations for factor scores that are NOT rated 1-9
  + Age
    - fit ages into a 1-9 scale through proportion calculation based on minimum and maximum ages in population
  + Gender
    - determine what genders 1 and 2 represent (male/female?)
  + Level (of risk)
    - to calculate correlations, low/medium/high may be converted to numerical values 1/2/3

Data Analysis

* utilize numpy, scipy.stats, and pandas
* Table - calculate the correlation between each factor and severity
* Matrix - correlation matrix between each factor to identify any associations between factors
* Unweighted cumulative risk score - simple addition of all factor scores
  + special consideration for age/gender during data cleaning
  + 7 (genetic) + 3 (pollution) + 2 (smoking) = unweighted cumulative risk
* Weighted cumulative risk score - weighted addition based on correlation
  + 7\*0.1 + 3\*0.3 + 2\*0.6 = cumulative weighted risk

Visualization

* utilize matplotlib.pyplot
* bar chart of risk factor correlations (descending) - show influence of each factor, from greatest to least
* scatter plot - unweighted cumulative risk VS severity
* scatter plot - weighted cumulative risk VS severity

Model

* linear regression - user inputs an entire novel set of risk factor scores
  + model adds factors together (weighted by correlations determined during data analysis) to yield an overall risk score

—--------------

Presentation Intro Scrip