Social Network Outreach Analysis
Capstone 2 Project Step-By-Step
Springboard Data Science Career Track
Eric Cruz

1) Roster download from Web Scraping
    a) Octoparse
        i) Filenames
            (1) CornellW.csv
            (2) CornellM.csv
2) Roster Data Wrangling to produce input file for LinkedIn Search
    a) Groupby() to consolidate same name on multiple rosters
    b) Supplement Roster downloads with All-time Letter Winner lists
    c) Parse Out Names for future use
        i) **Nameparser utility HumanName()**: First, Last, M.I., Suffix
            (1) Future Development: Nickname Database
    d) Output file is “FirstLast” list of names
        i) Future Development: Utilize Married vs. Maiden Name info but not needed in base case due to previous research
        ii) Analyze gathered LinkedIn Profiles not found in Search Result
    e) Program File - Capstone2_DataWrangling_Roster_CornellW.ipynb
        i) Input
            (1) CornellW.csv
            (2) CornellWpdf.csv (letterwinner list if applicable)
        ii) Output
            (1) CornellW_RosterGroup.csv
            (2) NamelistCornellW.csv
            (3) CriteriaListCornellW.csv
    f) Program File - Capstone2_DataWrangling_Roster_CornellM.ipynb
        i) Input
            (1) CornellM.csv
            (2) CornellMpdf.csv (letterwinner list if applicable)
        ii) Output
            (1) CornellM_RosterGroup.csv
            (2) NamelistCornellM.csv
            (3) CriteriaListCornellM.csv
3) LinkedIn Search
    a) **Selenium with Google Chrome Extension** used in **Jupyter Notebook** script
    b) Uses personal login but not counted in LinkedIn limits (only restrict outreach attempts), but applied for whitelist crawler exemption 3/15/20. Applied for PeopleSearch API under Marketing Developer application 2/8/20, received denial response 3/30/20.
    c) Program File - LinkedIn_Selenium_CornellM.ipynb
        i) Input
            (1) NamelistCornellM.csv
            (2) CriteriaListCornellM.csv
        ii) Output

(1) *CornellM_LinkedIn.csv (actual old)*
(2) *CornellMenLinkedIn.csv - to re-rerun and use new file*

d) *Program File - LinkedIn_Selenium_CornellW.ipynb*
i) *Input*
(1) *NamelistCornellM.csv*
(2) *CriteriaListCornellM.csv*
ii) *Output*
(1) *CornellWomenLinkedIn.csv*

4) LinkedIn/Roster Data Wrangling to produce Model Features
a) Gender label
b) FuzzyWuzzy ratios for partial name matching
c) Boolean classification for Tennis, School, SameName, and combination Same/Tennis
d) Decade as function of Year
e) *Program File - Capstone2_DataWrangling_LinkedIn_CornellM.ipynb*
i) *Input*
(1) *CornellM_RosterGroup.csv*
(2) *CornellMenLinkedIn.csv*
ii) *Output*
(1) *RosterLinkedIn_CornellM.csv*
f) *Program File - Capstone2_DataWrangling_LinkedIn_CornellW.ipynb*
i) *Input*
(1) *CornellW_RosterGroup.csv*
(2) *CornellWomenLinkedIn.csv*
ii) *Output*
(1) *RosterLinkedIn_CornellW.csv*

5) Labelled Data from base cases
a) Cornell Women manual search as part of 50th Anniversary Event outreach
b) Cornell Men is base case for collection followed by manual labelling
c) *Program File - Capstone2_Merge_LinkedIn_Labels*
i) *Input*
(1) *RosterLinkedIn_CornellM.csv*
(2) *RosterLinkedIn_CornellW.csv*
(3) *CornellM_match_feedback.csv*
(4) *CWTAMasterRoster50.csv*
ii) *Output*
(1) *RosterLinkedInLabel_CornellM.csv*
(2) *RosterLinkedInLabel_CornellW.csv*
(3) *RosterLinkedInLabel_CornellMW.csv*

6) Model and Feature Selection for Binary Classification *(Iteratively built into LinkedIn/Roster wrangling)*
a) **Random Forest Classifier (RFC)**; KNN, Logistic Regression evaluated
b) **Natural Language Processing with Fuzzy Wuzzy python library**
i) Use for approximate name matching as proxy for nicknames
c) Year, Decade, Fuzzy Scores are numerical features
d) Binary classification features are NameMatch, Tennis Activity, School, and Hybrid NameMatch&Tennis

7) Model Application
    a) Train Models on Men's Data and Predict on Women's Data and vice versa
    b) Combine Men and Women with Gender Classification feature and Re-train
    c) Apply model to unseen data from sample of other schools
    d) Program File - Capstone Project 2-rfc-feature-build-CornellMMWW.ipynb
        i) Input
            (1) RosterLinkedInLabel_CornellMW.csv
        ii) Output
            (1) modelmmww.joblib
            (2) Xdata12_CornellMMWW.csv (Contains predictions)
8) Analysis of Model Results
    a) For each selected feature, use bar charts to get frequency distributions and pie charts to get percentage scores
    b) Compare true labelled results with model predictions in feedback loop to identify important features
    c) Supplement Model Predictions with simple feature sorting to identify likely false predictions
    d) Program File - Capstone2_DataAnalysis-CornellMMW.ipynb
        i) Input
            (1) Xdata12_CornellMMWW.csv (contains predictions)
            (2) RosterLinkedInLabel_CornellMW.csv
        ii) Output
            (1) Note that prediction column is appended to Labelled dataframe
            (2) Only the Cornell Data is used to build the model, and predictions are taken while building - other schools will not build, only run
    e) Program File -
9) Deliverable is Roster List with LinkedIn search results, scores, feature data, and supplemental personal information
    a) Location, Employment, Schools, Activities, Roster Info
    b) Supplemental Letter Winner list downloads