Python Script Instructions

The data processing and analysis occurs sequentially in the three following python files:

data_filter.py
Input:

ACS_NSQIP_PUF_05_06_vr1.txt
ACS_NSQIP_PUF07_TXT.txt
ACS_NSQIP_PUF08_TXT.txt
ACS_NSQIP_PUF09_TXT.txt
ACS_NSQIP_PUF10_TXT.txt
ACS_NSQIP_PUF11_TXT.txt
acs_nsqip_puf12.txt
acs_nsqip_puf13.txt
acs_nsqip_puf14.txt
acs_nsqip_puf15_v2.txt
acs_nsqip_puf16.txt

All of these files can be found in the box folder.

This script takes in the raw text files and removes entries that do not contain certain CPT codes, generating filtered csv files. These CPT codes can be adjusted in the variable CPT_CODES. Run the script to generate the csv files.

Output:

A processed csv file for each input file (ex: "filtered_csv16.csv" represents the filtered NSQIP file from 2016 data)

csv_processor.py
Input:

The filtered csv files that were outputted from data_filter.py

This script converts each patient's medical record into a feature vector of specified comorbidities and outcome variables. Run the script to generate the csv file. The name of the output csv file can be adjusted using the variable FILE_NAME.

Output:

testingProcessor.csv

data_regression.py
Input:

testingProcessor.csv (output of csv_processor.py)

This script runs regression analysis on a specific outcome variable. To select which features are included in the analysis (ex: statistically significant comorbidities), place them in the feature_cols list. To select the outcome variable, set the variable y = dataset.outcome_variable. For example, if the outcome variable of interest is pneumonia, y = dataset.Pneumonia. The variable must be identical in casing and spacing to how it appears in col_names at the top of the file.

To examine inpatient or outpatient populations, adjust the X and y variables. Ex: To look specifically at the inpatient population,

```
X = datasetInpatient[feature_cols]
X = sm.add_constant(X)
y = datasetInpatient.Pneumonia
```

This script produces a console output. A sample output for the relationship between diabetes and pneumonia with features = ["Diabetes", "COPD", "Hypertension", "Independent Functional Health Status"] is as follows:

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              Pneumonia   No. Observations:                 1199
Model:                          Logit   Df Residuals:                     1194
Method:                           MLE   Df Model:                            4
Date:                Mon, 18 Feb 2019   Pseudo R-squ.:                  0.1111
Time:                        19:41:48   Log-Likelihood:                -67.751
converged:                       True   LL-Null:                       -76.221
                                        LLR p-value:                 0.001986
==============================================================================

                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         -4.0643      1.032     -3.939      0.000      -6.087      -2.042
Diabetes                       1.6291      0.624      2.609      0.009       0.405       2.853
COPD                           0.9906      0.925      1.070      0.284      -0.823       2.804
Hypertension                   0.4741      0.632      0.750      0.453      -0.764       1.713
Independent Functional Health Status    -1.1963      1.001     -1.195      0.232      -3.158       0.766
==============================================================================

                                                 2.5%      97.5%       OR
const                                        0.002273   0.129800  0.017175
Diabetes                                     1.499659  17.339631  5.099366
COPD                                         0.439061  16.517130  2.692959
Hypertension                                 0.465592   5.544155  1.606647
Independent Functional Health Status         0.042495   2.150641  0.302309
```

p-value for Diabetes = 0.00907491063352
p-value for COPD = 0.288143420389
p-value for Hypertension = 0.462009623049
p-value for Independent Functional Health Status = 1.30049462122

The key information for analysis is highlighted. This is the 95% confidence interval (0.464492 - 5.544155), the odds ratio (1.606647), and the p-value (0.009).