I download the both data sets and compared them to decide which data set is more applicable to wrangle in Python. The data file named LLCP2017.XPT is suitable to be manipulated since each row at the file has raw data.

I read the description of the provided data, and then explored it in Python. I made a data frame by filtering New England Area-data that has over 45 thousand rows and 358 columns. Then, I cleaned the data.

The very first step of the data screening is to decide on the columns need to be dropped features with at least 90% missing observations. Before deciding on %95, I tried for different ratios as well. This value is in sweet spot in regards to dropping row-column trade off. Thus, 230 columns are dropped and I have 128 columns left. Then I have kept observations that are complete across all remaining columns /features. After this step, I got 39k rows out of 45k rows. As a result, I lost 6k rows and 230 columns of the original New England data set.

Feature Engineering:

First of all, I identified which columns are categoric or numeric by screening in Python and from the CDC link.

2nd QUESTION:

I decided to apply A/B analysis to find the features that distinguish Rhode Island from other New England states. I have then created data frames covering each one of the states. So, I had 6 different data frames in total.

The method I chose is T-Test. I decided on p-value as 0.05 with %95 confidence level. Later, I checked for all features in Python by comparing RI with the rest New England states one by one. You can find all output values for all combinations in the code. According the result of the T-Test, I have found 64 of 128 features are not likely same at least three states of New England Area. You can find the whole list of the features as following:

['HTIN4', 'SMOKE100', 'HTM4', 'DIFFWALK', '\_RACE\_G1', 'HEIGHT3', '\_LMTSCL1', '\_LMTWRK1', '\_LMTACT1', '\_DRDXAR1', 'INCOME2', 'EMPLOY1', 'MENTHLTH', 'VETERAN3', 'CHECKUP1', 'BPHIGH4', 'CHOLCHK1', 'PHYSHLTH', 'GENHLTH', 'MARITAL', 'FMONTH', 'DISPCODE', '\_PSU', 'RENTHOM1', 'ADDEPEV2', 'HAVARTH3', '\_VEGLT1A', '\_SMOKER3', 'DROCDY3\_', '\_CURECIG', '\_ECIGSTS', '\_VEG23A', '\_VEGETE1', '\_TOTINDA', 'MAXVO2\_', 'FC60\_', '\_PACAT1', '\_PAREC1', '\_RFSEAT2', '\_RFSEAT3', '\_INCOMG', '\_DUALUSE', '\_RFHLTH', '\_LLCPWT', '\_LLCPWT2', '\_IMPRACE', '\_WT2RAKE', '\_RAWRAKE', '\_STRWT', '\_STSTR', 'QSTLANG', 'QSTVER', '\_RFHYPE5', '\_CHOLCH1', '\_PRACE1', '\_MRACE1', '\_HISPANC', '\_RACE', '\_RACEG21', '\_RACEGR3', '\_AGEG5YR', '\_AGE80', '\_AGE\_G', '\_STATE']

3rd QUESTION:

First of all, I assessed the correlations between the columns and the states. I dropped unrelated columns such as SEQNO, IDATE and \_STSTR. I used 123 features (17 numeric, 111 categoric) and handled 578 dummy variables. After visualizing the correlation level of the features, I chose the strongest variables, 31 out of 578 (first 14 inverse correlated and 17 first correlated variables). The highest correlated one is QSTVER\_10 by 0.52 magnitude and highest inverse correlated one is \_PSU -0.32 magnitude in gap of -1 and 1 magnitude.

I chose the logistic regression, KNN, Random Forest and SVM as models. Regarding the results from training and testing data, the best model is KNN because there is no overfitting and bias. There is no significant difference between training and testing scores. It is in sweet spot regarding to trade of between overfitting and bias. Therefore, it is the most consistent method for predictions. Support Vector Classifier (SVC) is the second consistent predictor by the same perspective. Random Forest could mislead due to overfitting. There is a big difference between training and testing scores of it. So, its consistency of prediction is low. Logistic Regression has no overfitting and bias but, its prediction accuracy is not high as much as KNN or SVC.

A screenshot of a cell phone

Description automatically generated