Ashwin Krishna, Hamish Nicholson, Ryan Plunkett

We chose to collect information on incumbency status for every congressional election determining seats in the House of Representatives since 2010, using binary indicators where a True value implied that the given candidate was running for re-election. In order to procure these features, we wrote a Python script to scrape the Wikipedia pages corresponding to each election, where all of the needed information was conveniently stored in HTML tables. Because page structure was not consistent across elections, we scraped each election separately, tinkering with our BeautifulSoup code as necessary to produce Pandas DataFrames storing the correct results. After ensuring that our DataFrames were formatted properly and manually removing some erroneous results, we wrote said DataFrames to CSV files and then concatenated them to make a master DataFrame. Our attached CSV should contain the requested columns (Year, District, Candidate, and Incumbency), and it can be used to help train our machine learning methods for the upcoming project.