WGBH - DCAMM SCRUM Report 5 4/1/21	
Contact	Github accounts: elisa3lopes, rlee99, murtio, carmen-araujo, jenajjedu Email addresses: ● Jena Jordahl jenajj@bu.edu ● Elisa Cordeiro Lopes elisacl@bu.edu ● Richard Lee rlee99@bu.edu ● Murtadha Bahrani murtadha@bu.edu ● Carmen Sabrina Araujo sabrinaa@bu.edu
What have I worked on?	● Created a program that produces a clean dataframe from the PDFs! (!!!!!) ○ Had to read directly from the PDF to a pandas data frame. ○ Downloaded and installed tabula on my local machine so that I could use the template feature that Rashib recommended. Templates are created in Json format and then options parameters for scripts are derived from them. ● Do initial analysis of the data, including initial plots and realize there are no hires for 2019
Have I talked to the client recently? When are we meeting with them next?	● We talked with the client on Thursday 4/1/21 11:30-12:30pm. ● We will meet with the client on Thursday 4/8/21
What will I be working on next?	● With this data, perform more analysis and create clean visualizations ● Answer questions the client has regarding race and gender and government contracts (hours split between trades and ethnicity/gender) ● Generate our own questions to analyze the data ● Finishing up the parser

	● Create presentation
Have I run into any issues? Do I need help?	● Inconsistencies in the data have made reading the data sometimes produce empty numbers ● We get rows that are misaligned with the column headings ● We get two columns for one number split on the thousand’s place comma. ● When we were not using the tabula template configuration generated by the online system, we were getting rows chopped off at the end and entered on a different line. ● The construction trade field had new line characters embedded in the data. ● Column heading fields had ‘\r’ white space characters embedded in them. ● The data that held the project information, the highest level sort on the report was contained in a separate data frame than the detail rows. ● Multiple data frames had to be read per page ● The contractor heading was sometimes left dangling on a page without any data associated with it on the same page. This put it into a separate data frame than the rows containing the contractor’s time. ● New Hire data is 0 for the whole year 2019 ● Creating the parser is taking over 40 hour a week