# Project WGBH DCAMM

Project Deliverable 0

All contractors commissioned by the state for major construction projects need to report their ethnic and gender makeup of the work forces. The WGBH would like to understand the data contained in those Summary of Workforce Utilization reports. Furthermore, the WGBH is interested in getting data-driven insights of the impact drawn upon specific groups of working forces between 2019 to 2020. The data is given in PDF format and organized by hours spent per project per organization. Our goal is to first extract data in proper formats from the PDF files and then run some analysis.

## Weekly Meeting with the PM

Lingyan Jiang is Thurs 11:30 AM - 1:00 PM

## With WGBH

Paul Singer, - every other Thurs 11:30 AM - 1:00 PM
Second meeting with the client on Thurs March 4th
Spark Liason - Greta Bruce

### Contact List:

Client Paul Singer paul_singer@wgbh.org,
Spark Liason Greta Bruce gretab@bu.edu,
PM Lingyan Jiang lingyanj@bu.edu,
Students Rep Jena Jordahl jenajj@bu.edu,
Elisa Cordeiro Lopes elisacl@bu.edu, Richard Lee rlee99@bu.edu Murtadha - Ahmad M Al Bahranimurtadha@bu.edu Carmen - Sabrina Araujosabrinaa@bu.edu

### Github accounts:

elisa3lopes, rlee99, murtio, carmen-araujo, jenajjedu

## Data Sources

The data is collected weekly by DCAMM. They sort it by months and keep it in PDF form. DCAMM already provided WGBH the work force from 2019 and will provide in March the data from 2020. The data is organized as tables of projects (such as bridges, buildings, etc) containing the companies included, their types of workers, and the hour rate separated by race, sex, and ethnicy. For this project, no additional datasets are required to be extracted, but our team is open to get any other information as it seems relevant to analysis. An example of a file is April 2019: https://drive.google.com/file/d/1brxGTjfkhwKRXPAbzDwHl4bP6J08Xwtz/view?usp=sharing

We have been given a file folder with files for each month Jan - Dec 2019, e.g. WorkforceUtilizationSummaryReportApril2018.pdf

## Methods

Use Python library PyPDF2 to scrape the data from the PDFs files into comma separated documents. For some PDFs we manually import them into CSV files. Clean the extracted data from unrecognized characters and missing values. We use Pandas and NumPy libraries for this step. We explore the data by performing basing exploratory data analysis using Matplotlib library. Finally, We run various machine learning algorithms to predict the number of hours assigned per project for each group. We use the Scikit-learn library for the predictive analysis step.

## Discussion and Limitations

Note: The client was only interested in the extraction of the data from PDF files and said we would discuss specific questions for analysis after this was completed. Fortunately, we were able to convert PDF files into Excel Spreadsheets and then to CSV files. Our next step is to align our columns and make our data more organized and accessible to manipule. Subsequently, we will start to look for any patterns or trends to answers questions revolving around race, sex, and opportunities for state contracts. Specific questions that will be answered: How will we extract data from our PDF files? Is there a difference between state-paid contractual hours based on color and/or sex? What are the factors, e.g. location of the project, that fair in hiring working crews? How state-wise elections affect hiring decisions across projects?