Project Deliverable 1

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.

Logistics

Weekly Meeting with the PM

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

Weekly Meeting 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 Bahrani murtadha@bu.edu, Carmen Sabrina Araujo sabrinaa@bu.edu.

Github accounts

elisa3lopes, rlee99, murtio, carmen-araujo, jenajjedu

Data

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

Since the main goal of this project is to get the data out of the PDF and transform it into a csv or .xlsx file in an organized way, our team is exploring multiple methods. Preliminary, we extracted the data using the online Tabula application, as well as programmatically extracting the data using the Tabula library (https://pypi.org/project/tabula-py/) and the PyPDF2 library. We also used acrobat to save the file in CSV format. Each method produced the same misalignments between the hourly data rows and the company/trades header data. The issue stems from the PDF merging cells to pretty print the data for human readability.

Our methodology will include a means to read through the data file and realign the company and the trade data with the hourly time spent rows. We may leverage Pandas and NumPy libraries when realigning the data. We will need to write a custom python script to complete our task.

When the data is aligned, we will run various machine learning algorithms to predict the number of hours assigned per project for each group. We can use the Scikit-learn library for the predictive analysis step. For visualization we will use Matplotlib and Plotly Dash.

Specific questions that will be answered:

1. How will we extract data from our PDF files?
2. Is there a difference between state-paid contractual hours based on color and/or sex?
3. What are the factors, e.g. location of the project, that fair in hiring working crews?
4. How do state-wise elections affect hiring decisions across projects? Did construction companies hire fewer minorities, people of color and/or women, during the pandemic?
5. Did the companies' work workload change because of COVID-19?

Our project work will have a larger portion of total time devoted to data cleanup vs data exploration due to the condition of the data sources. Subsequently, we will start to look for any patterns or trends to answers questions revolving around race, sex, and opportunities for state contracts.

Discussion and Limitations

Please refer to our screenshots 1-5 below to see our data import. On a preliminary analysis, we concluded that the data is divided by minorities and project type, as well as jobs per month. In header areas, the data is very sparse containing multiple zero rows and columns with nans inserted by Pandas(screenshot 5). Due to the obvious data misalignment and extraneous values, we will continue to build on our custom python parser for this project.

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. Our next step is to align our columns and make our data more organized and accessible to manipulate.

For this deliverable we were able to answer the first question listed in our methods section: How will we extract data from our PDF files? Now we have a new question which is: How to pre-process the data for our later analysis?

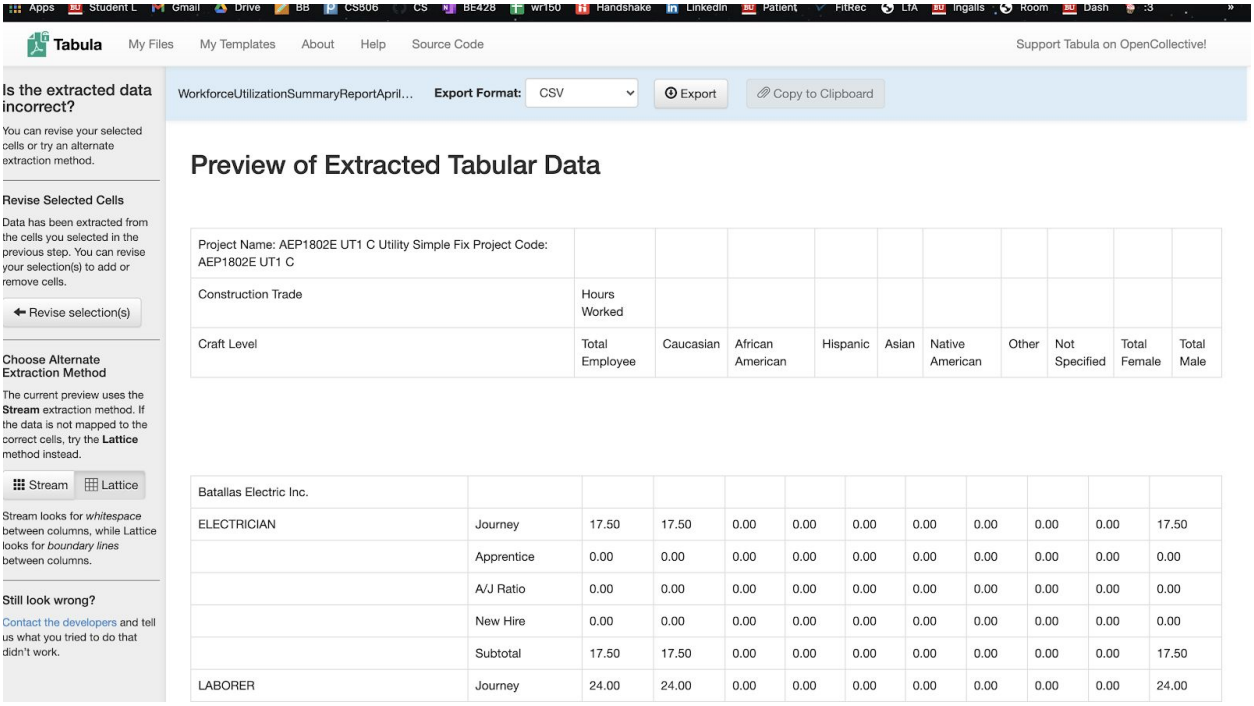
For limitations, when using libraries to transform PDFs into csv or xml, misplacements happen, such as elements in the wrong place and generating the wrong number of columns. To handle these issues, a common practice is to read the file in a series of passes. The first pass will extract the company names and be validated for accuracy by ourselves and our client. The second pass will extract the names of the trades and any

sub-categories like “journeymen”. The second pass data will again be vetted by ourselves and the client. The third pass will integrate the two previously cleaned sources and produce a data frame sufficient for our analysis.

Our client has not specified analysis of our data because they understand the difficulty of massaging the data. Due to these limitations, our project will be limited and will not include all four aspects of data science analysis as defined by our instructor.

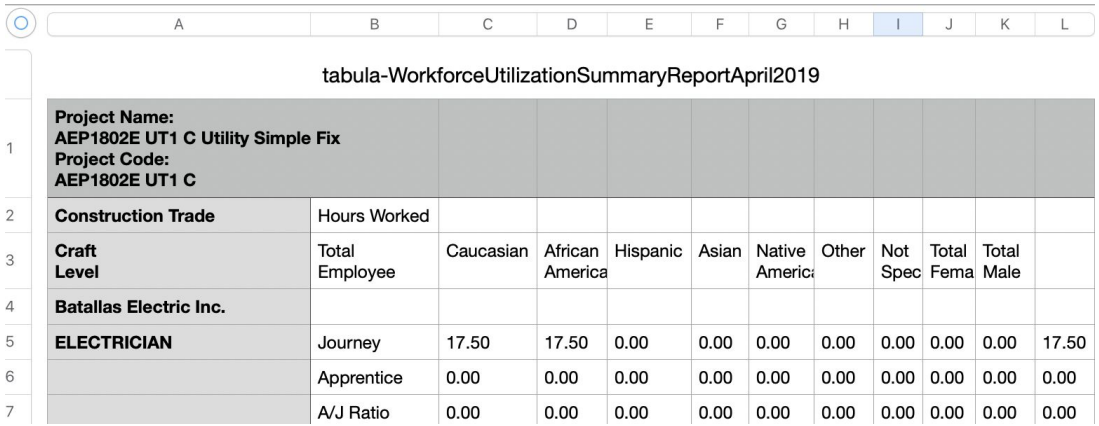
Screenshots

1. Tabula software example. Note the Company name row is empty for all other columns. This is the cause of the unalignment.



Project Name: AEP1802E UT1 C Utility Simple Fix Project Code: AEP1802E UT1 C											
Construction Trade	Hours Worked										
Craft Level	Total Employee	Caucasian	African American	Hispanic	Asian	Native American	Other	Not Specified	Total Female	Total Male	
Batallas Electric Inc.											
ELECTRICIAN	Journey	17.50	17.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.50
	Apprentice	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	A/J Ratio	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	New Hire	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Subtotal	17.50	17.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.50
LABORER	Journey	24.00	24.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.00

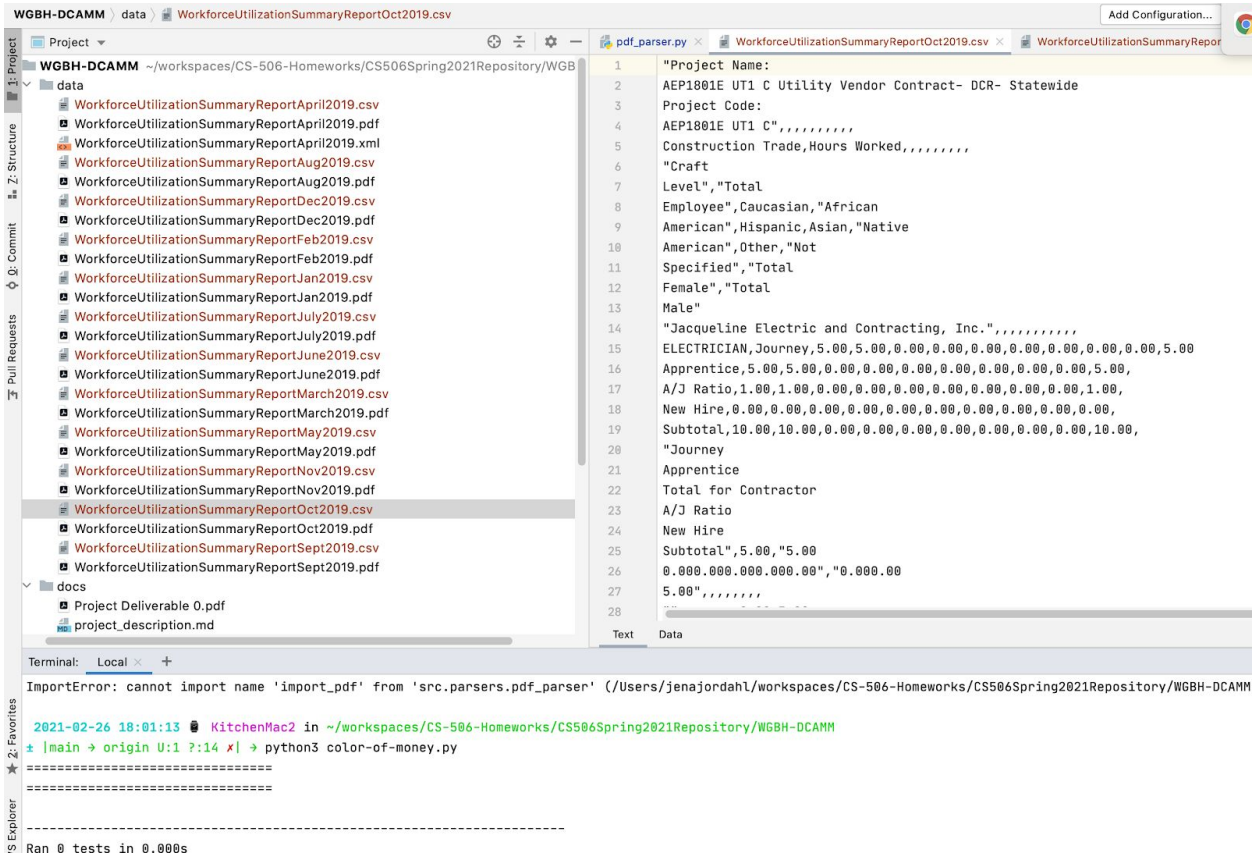
2. Tabula software CSV example. In this format, there is no visual problem, however when transforming into a Pandas Dataframe the row below the company name will merge with it the job name, and have one extra column.



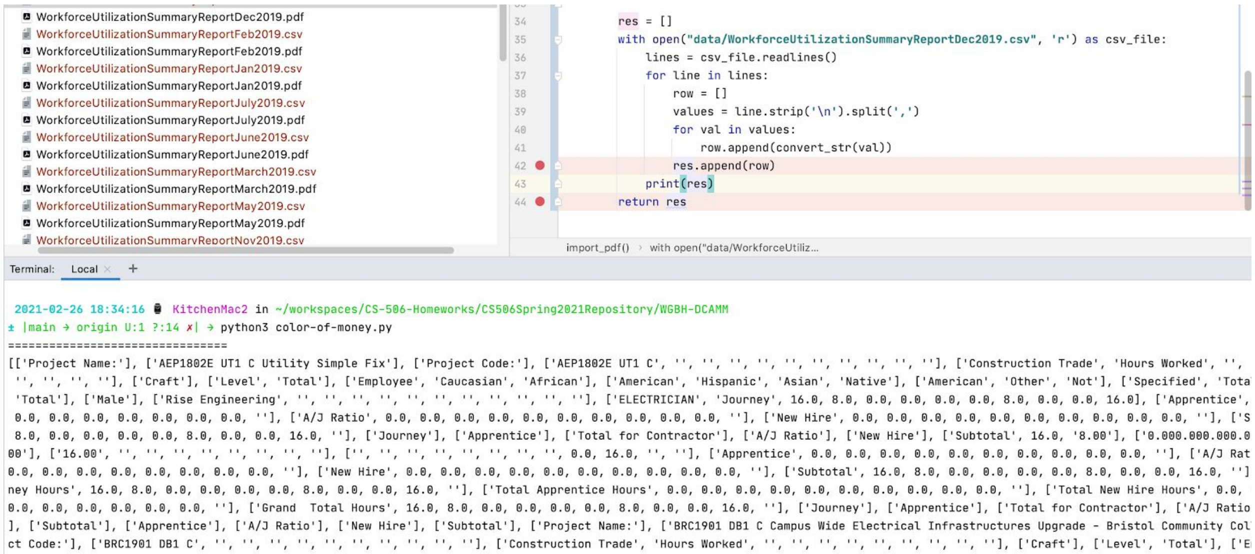
tabula-WorkforceUtilizationSummaryReportApril2019

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Project Name: AEP1802E UT1 C Utility Simple Fix Project Code: AEP1802E UT1 C**											
2	**Construction Trade**	Hours Worked										
3	**Craft Level**	Total Employee	Caucasian	African America	Hispanic	Asian	Native America	Other	Not Spec	Total Fema	Total Male	
4	**Batallas Electric Inc.**											
5	**ELECTRICIAN**	Journey	17.50	17.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.50
6		Apprentice	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7		A/J Ratio	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3. Tabula library example showing how Tabula reformats all PDFs in a directory into like named csv files. The raw CSV output is shown on the right side panel:



4. An example of the first try at creating a custom python script to read the csv output is shown on the right side panel. The output is printed in the terminal window at the bottom.:



5. Pandas Dataframe example from Tabula. Here in row 14 it is possible to see one of the merging problems. Also, rows 7 to 12 were supposed to carry the race and gender, but they also have a misalignment.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	Project Name:	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
1	AEP1802E UT1 C Utility Simple Fix	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
2	Project Code:	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
3	AEP1802E UT1 C											None	None	None	None	None
4	Construction Trade	Hours Worked										None	None	None	None	None
5	Craft	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
6	Level	Total	None	None	None	None	None	None	None	None	None	None	None	None	None	None
7	Employee	Caucasian	African	None	None	None	None	None	None	None	None	None	None	None	None	None
8	American	Hispanic	Asian	Native	None	None	None	None	None	None	None	None	None	None	None	None
9	American	Other	Not	None	None	None	None	None	None	None	None	None	None	None	None	None
10	Specified	Total	None	None	None	None	None	None	None	None	None	None	None	None	None	None
11	Female	Total	None	None	None	None	None	None	None	None	None	None	None	None	None	None
12	Male	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
13	Rise Engineering												None	None	None	None
14	ELECTRICIAN	Journey	16.0	8.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0	16.0	None	None	None	None
15	Apprentice	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		None	None	None	None
16	A/J Ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		None	None	None	None
17	New Hire	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		None	None	None	None