Project Deliverable 3

All contractors commissioned by the state for major construction projects need to report their ethnic and gender makeup of the work forces. WGBH would like to understand the data contained in those Summary of Workforce Utilization reports. Furthermore, WGBH is interested in getting data-driven insights of the impact drawn upon specific groups of workers 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
- ❏ 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 by a Massachuttes state office, DCAMM, Division of Capital Asset Management and Maintenance, and reported out to the community via an annual report. WGBH requested additional documentation from the state so they could independently verify the numbers in the annual report. Through a freedom of information act request, WGBH was able to receive monthly construction workforce utilization reports. The reports are kept in PDF format. DCAMM already provided WGBH twelve monthly reports for 2019 and in March they were to provide the data from 2020. So far the 2020 data has not been delivered to WGBH and we are in the second week of April.

Later other data may be of interest to analyze. Our team has only the construction data to analyze not the design data which is also included in the annual report. The annual report includes other data on the contractor's business location where payments were made and the location of the worksite.

Recently, WGBH let us know that we can download the database of WBE and MBE from a state site to validate the volume of contract hours reported to WBE companies and compare to the annual report. By correlating the data, we will assist WGBH in verifying the DCAMM numbers published in the yearly report, and identify new patterns.

The 2019 dataset of construction data is organized as tables of projects summaries by month per contractor, trade and level of experience, such as bridges, buildings, etc. The important statistics per company includes, their types of workers listing the number of hours worked by race, sex, and ethnicty. 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 analyze. 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.

See Image 10 below, the tox testing framework with a couple of tests set up for building a test suite.

1. All data is collected

After a lot of work, the PDF parser is completed. We used PyPDF2, Tabula, and Pandas to read in the report page by page and immediately convert it to a pandas

DataFrame. Once the DataFrames were created, the parser read as many as were found per page. The system read each row like it was parsing a grid map. The code reflects the importance of understanding the exact position of the data on the grid. Though our first attempts to use tabula to pandas did not work out, eventually we learned how to configure tabula to recognize the table boundaries in the report using a template generator. Once we got the appropriate x, y coordinates of the bounding box tabula needed to parse the report, we were no longer getting subtotal lines that are split into two rows. Also, we found a reference online which informed us that we had to use PyPDF2 to count the number of pages in a pdf so that we could control each page parsed by tabula. Only with page by page control were we able to extract the needed rows and drop the rest.

With the two libraries working together, we were able to extract all data from 2019 monthly reports into dataframes meeting with all unique specifications the PDFs had, such as alignment of cells, row break, and absent values. Apart from the data extraction, no further data was required or discussed to be collected since the Summary of Workforce Utilization report already carries a lot of meaningful information that interests the client.

2. Refine the Preliminary Analysis of the Data Performed in PD1&2

In PD1, we performed a few Preliminary Analysis. Our client had not specified analysis of our data because they understood the difficulty of massaging the data, so our project at that time was limited and would not include all four aspects of data science analysis as defined by our instructor. We could however ask some questions by understanding the general type of the data. In PD2, we were still facing issues about parsing the PDFs, including comma split values, row breaks, lost cells, and others. The CSV, however, allowed us to make some comparisons of total hours of work by gender, race, ethnicity, and new hires for April of 2019. From one month's data, we could see that there were very few Apprentice and New Hire hours compared to the Journeyman hours.

With the parser done we could expand our analysis. From the entire 2019 dataset, we could see that finding fewer Apprentice hrs per ethnicity vs Journeymen hours makes sense. The insight is that Apprentices are learning their craft and like other disciplines fewer people are currently studying to join a discipline then are working members of the discipline. However, this is not the case for African Americans and Women. While 8% of Apprentice are African Americans, they make up to only 2% of Journeyman. Similarly, Women are 11% of the overall Apprentice and only 1.5% of

Journeyman. These results showed the client hard evidence that African Americans are not maturing to gain state contract work as Journeymen. Also worth noting, apprentices are cheaper labor than Journeyman so the value of what African Americans and Women get from their participation in the construction industry is smaller.

These results should be further analyzed and explored considering not only these hard numbers but also the cultural, social, historical, and economic part of Massachusetts construction contracts and in light of election results.

3. Answer another key question

a. How will we extract data from our PDF files?

   i. As mentioned, we used PyPDF2 to extract the number of pages, the interactive tabula tool to create the bounding box x,y coordinates, and tabula python library which utilizes pandas directly when it chunks out the file page by page.

   ii. Next, we created a python script to read each pdf file in the input directory and produce a CSV file into a second directory. The file contains a denormalized model of the monthly project and contractor workforce hours performed per ethnicity and gender. One output CSV file is created per input PDF file. To keep the data appropriately marked, we added month and year data to the dataframe from the filename being parsed.

   iii. By generating the monthly columnar CSV files, we can build a mini-data mart for querying four different hierarchical trees. One tree for Project/Contractor/Trade/Experience Level, one for time series(month and year) and two others for Ethnicity and Gender. All arms of the tree tie count hours worked per contract per month. The structure of organizing the data is commonly called a data cube and the schema strucalled a star schema. See **Figure 3 Final DataMart DataFrame below.**

   iv. Finally, to do the analysis work, we read our file-based data cube into a single pandas DataFrame again using a script. From the combined dataset, we could easily execute the group-by statements to compute percentages of the money received per ethnicity and per gender. See **Figure 4, Distribution of Apprentice vs Journeymen hours annually.**

b. Is there a difference between state-paid contractual hours based on color and/or sex?

i. Based on our analysis from the datamart, we can see that caucasian males have accumulated the most state-paid contractual hours.
ii. We can also see that African Americans, Hispanics and Women are under represented as Journeymen compared to their numbers as apprentices.
iii. We also see that the annual report chooses to give statistics for the ethnicities as fractions compared to total ethnicity hours rather than compared to total employee hours worked. We noticed this discrepancy when we tried to tie our computed values to the published values.

c. We also noticed that all new hire hours were zero, 0. This seems like an error in the report produced by DCAMM and warrants a question to be posed to their office: **Phone** 617-727-4050 .

4. Attempt to answer overarching project question

While we weren't originally given a specific analysis question to answer, after successfully building our parser, we did gather that the overarching project question was how hours were dispersed based on ethnicity, gender, and by project, contractor, trade, level of experience (Apprentice/Journeymen), and new hires. While we were able to produce some results and visualizations, we have yet to parse the data from 10 years back.

Some more questions that we will try to answer in the coming days with our working parser is how many different companies get contracts? How many distinct companies are all one race? How much bigger are the big companies? Are people being blocked in earning enough hours to become a journeyman?

With one year of data we found some pretty remarkable results. 86% of journeymen were caucasian americans and 68% of apprentices were caucasian americans. 2% of journeymen were african americans and 8% of apprentices were african americans. Even just from one year we can see some drastic trends. On the gender side, 98% of journeymen were male americans while 88% of apprentices were male americans.

So can we say that there is some type of descrimination in state-paid contractual hours? Our one year analysis says yes, but we must find the 10 year trend and dive much deeper into the data we will receive.

5. Create a draft of your Final Report

One finding of our analysis highlighted a hiring practice pattern significantly skewed against minorities. We found that more apprentice hours were worked by several ethinic groups compared to the total number of journeymen hours. Though African Americans are known for their talent as laborers, they have roughly as much opportunity to work journeymen hours as a woman does, 2% vs 1.5% respectively. Women are not traditionally thought of as good construction laborers; hence the rough equivalency of representation as journeymen in construction work for women and African Americans stood out as deserving more attention.

Since 86% of the total journeymen hours were performed by Caucasian Americans and 68% apprentice hours, Caucsian's dominance in the MA construction contract work is obvious. On the other hand, male or female African Americans fulfill 8% of the apprentice hours; hispanics, 20%. So the lower paying jobs are the ones made available to minorities in state construction work during 2019. Similar observations can also be made for women. Once more data is available, a trend might indicate that African Americans and Women are not gaining journeymen work positions or that there is a push to add ethnicities and women so they are starting out at the first level of the trade and will eventually move up to journeymen hours.. See **Figure 0, Representation of African Americans and Women as Apprentice and Journeymen**.

6. Refine project scope and list of limitations with data and potential risks of achieving project goal

Project Scope:

- This project's goal was to solely make the parser and transform the PDFs into CSVs/DataFrames. Now that we finished the parser, we have space to actually do more analysis.
- Besides looking for ways to parse PDFs, to break rows, or to adjust misalignments, our team can think outside the box and actually analyze the data.

Limitations and Risks:

- The parser may not work for later years. It works perfectly on 2019 data, but if the PDF format changes the parser will have to be adapted.

- Our data is time limited. At most we will only have 2 years of data and making assumptions of past years is something we need to be careful about asserting. For example, as mentioned in question B section 3.
- African Americans and Women are represented as a larger percentage of apprentices than journeymen, Figure 0 Apprentice vs Journeymen hours. This doesn't make sense considering because an Apprentice should be part of the trade community “in-training” vs trained workers. This can be interpreted as a disadvantage for them, being harder for them to actually get the job. On the other hand, the companies actually might be looking to achieve more diversity and hiring more Women and African Americans than they ever did, still resulting in the abnormality of having more Apprentice than Journeyman.
- Since the PDFs are long, it is hard to check if the parser is actually acquiring all the data according to the original. Also, if WGBH wants to use the information we identify then they would need proof that we parsed all that data according to the PDFs they received. To overcome this limitation, our team successfully made the parser create a proof.txt file for use in a random sample test plan to confirm that each row in the received PDF file was accurately categorized and processed in the final pandas DataFrame. Proof.txt is 115,304 lines long.Figure 2, Proof.txt excerpt from the top and bottom of the file.

SCREENSHOTS

Figure 0, Representation of African Americans and Women as Appretice and Journeymen

Figure 1, Apprentice vs Journeymen Hours

WorkforceCraftSummaryReport2019

CRAFT_LEVEL	TOTAL_EMPLOYEE	CAUCASIAN	AFRICAN_AMERICAN	HISPANIC	ASIAN	NATIVE_AMERICAN	OTHER	NOT_SPECIFIED	TOTAL_FEMALE	TOTAL_MALE
Apprentice	54893	37508	4592	11191	1476	0	128	0	6110	48784
Journeymen	476507	408302	10690	45692	4122	210	7336	155	7079	469428

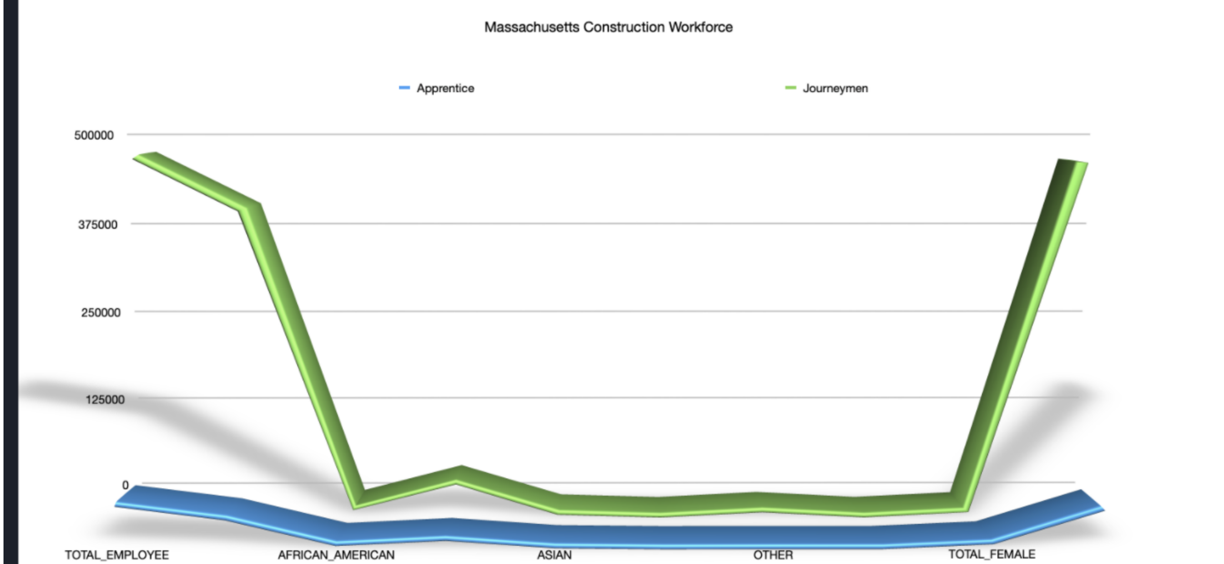

Figure 2, Proof.txt file

```
proof.txt

starting page 1
 Processing DF number: 0
                                              0   ...          10
0  Project Name:\rAEP1802E UT1 C Utility Simple F...  ...         NaN
1                               Construction Trade  ...         NaN
2                                     Craft\rLevel  ...  Total\rMale

[3 rows x 11 columns]

 Processing row number: 0
['MONTH', 'YEAR', 'PROJECT', 'PROJECT_CODE', 'CONTRACTOR', 'CONSTRUCTION_TRADE',
'CRAFT_LEVEL', 'TOTAL_EMPLOYEE', 'CAUCASIAN', 'AFRICAN_AMERICAN', 'HISPANIC', 'ASIAN',
'NATIVE_AMERICAN', 'OTHER', 'NOT_SPECIFIED', 'TOTAL_FEMALE', 'TOTAL_MALE',
'HOURS_WORKED_PER_MONTH']
 Processing DF number: 1
                            0           1      2    3   ...   8    9     10     11
0            Rise Engineering         NaN    NaN  NaN   ...  NaN  NaN    NaN    NaN
1   INSULATOR (PIPES & TANKS)     Journey  101.0  7.5   ...  0.0  0.0    0.0  101.0
2                         NaN  Apprentice    0.0  0.0   ...  0.0  0.0    0.0    0.0
3                         NaN   A/J Ratio    0.0  0.0   ...  0.0  0.0    0.0    0.0
4                         NaN    New Hire    0.0  0.0   ...  0.0  0.0    0.0    0.0
5                         NaN    Subtotal  101.0  7.5   ...  0.0  0.0    0.0  101.0
6        Total for Contractor     Journey  101.0  7.5   ...  0.0  0.0    0.0  101.0
7                  Apprentice        0.00    0.0  0.0   ...  0.0  0.0    0.0    NaN
8                   A/J Ratio        0.00    0.0  0.0   ...  0.0  0.0    0.0    NaN
9                    New Hire        0.00    0.0  0.0   ...  0.0  0.0    0.0    NaN
10                   Subtotal      101.00    7.5  0.0   ...  0.0  0.0  101.0    NaN
11       Total  Journey Hours      101.00    7.5  0.0   ...  0.0  0.0  101.0    NaN
```

```
 Processing row number: 0
 complete grp cj  WSC1902 DB1 C WSC-Campus-wide-Redundant Steam and Condensate Loop  1
 complete grp project

 Processing DF number: 1
                        0         1      2    3   ...    7    8      9     10
0    Total  Journey Hours  1,056.00  904.5  0.0   ...  0.0  0.0  143.5  912.5
1  Total Apprentice Hours      7.00    7.0  0.0   ...  0.0  0.0    0.0    7.0
2    Total New Hire Hours      0.00    0.0  0.0   ...  0.0  0.0    0.0    0.0
3      Grand  Total Hours  1,063.00  911.5  0.0   ...  0.0  0.0  143.5  919.5

[4 rows x 11 columns]

 Processing row number: 0
 grand_total 4
 process grp pj
 process grp pa
 process grp pnh
 process grp pgrand
 complete grp grand_total

   MONTH  YEAR  ... TOTAL_MALE HOURS_WORKED_PER_MONTH
0     12  2019  ...       16.0        HOURS_PER_MONTH
0     12  2019  ...        0.0        HOURS_PER_MONTH
0     12  2019  ...        0.0        HOURS_PER_MONTH
0     12  2019  ...      165.0        HOURS_PER_MONTH
0     12  2019  ...     1632.0        HOURS_PER_MONTH
..   ...   ...  ...        ...                    ...
0     12  2019  ...        7.0        HOURS_PER_MONTH
0     12  2019  ...      393.0        HOURS_PER_MONTH
0     12  2019  ...        0.0        HOURS_PER_MONTH
0     12  2019  ...      512.5        HOURS_PER_MONTH
0     12  2019  ...        0.0        HOURS_PER_MONTH

[654 rows x 18 columns]
```

Figure 3, Final DataMart DataFrame

MONTH	YEAR	PROJECT	PROJECT_CODE	CONTRACTOR	CONSTRUCTION_TRADE	CRAFT_LEVEL	TOTAL_EMPL...	CAUCASIAN	AFRICAN_...	HISPANIC
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	North shore steel company, inc	IRONWORKER/WELDER	Journeymen	48.00000	0.00000	0.00000	24.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	North shore steel company, inc	IRONWORKER/WELDER	Apprentice	0.00000	0.00000	0.00000	0.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	North shore steel company, inc	LABORER	Journeymen	48.00000	0.00000	0.00000	24.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	North shore steel company, inc	LABORER	Apprentice	0.00000	0.00000	0.00000	0.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	Stanley roofing company, inc	ROOFER	Journeymen	620.25000	307.25000	0.00000	313.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	Stanley roofing company, inc	ROOFER	Apprentice	0.00000	0.00000	0.00000	0.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	Stanley roofing company, inc	SHEETMETAL WORKER	Journeymen	268.50000	100.00000	0.00000	168.50000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	Stanley roofing company, inc	SHEETMETAL WORKER	Apprentice	0.00000	0.00000	0.00000	0.00000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	Zap electric	ELECTRICIAN	Journeymen	523.00000	252.50000	0.00000	270.50000
12	2019	TRC1407 FC1 C Ex...	TRC1407 FC1 C	Zap electric	ELECTRICIAN	Apprentice	0.00000	0.00000	0.00000	0.00000
12	2019	TRC1702 HC1 C S...	TRC1702 HC1 C	3 phase elevator	ELEVATOR CONSTRUCTOR	Journeymen	447.50000	447.50000	0.00000	0.00000

	ɔR	CONSTRUCTION_TRADE	CRAFT_LEVEL	TOTAL_EMPL...	CAUCASIAN	AFRICAN_...	HISPANIC	ASIAN	NATIVE_A...	OTHER	NOT_SPEC...	TOTAL_FE...	TOTAL_MALE	HOUR...
197		GLAZIER	Apprentice	64.00000	0.00000	0.00000	64.00000	0.00000	0.00000	0.00000	0.00000	0.00000	64.00000	HOUR...
198		GLAZIER (GLASS PLANK/AIR BA...	Journeymen	264.00000	264.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	264.00000	HOUR...
199		GLAZIER (GLASS PLANK/AIR BA...	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
200		IRONWORKER	Journeymen	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
201		IRONWORKER	Apprentice	93.00000	0.00000	0.00000	93.00000	0.00000	0.00000	0.00000	0.00000	0.00000	93.00000	HOUR...
202		IRONWORKER/WELDER	Journeymen	304.00000	304.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	304.00000	HOUR...
203		IRONWORKER/WELDER	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
204	ng	CARPENTER	Journeymen	2344.50000	2160.50000	0.00000	184.00000	0.00000	0.00000	0.00000	0.00000	0.00000	2344.50000	HOUR...
205	ng	CARPENTER	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
206	ng	LABORER	Journeymen	608.50000	608.50000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	608.50000	HOUR...
207	ng	LABORER	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
208		ELEVATOR CONSTRUCTOR	Journeymen	350.00000	350.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	350.00000	HOUR...
209		ELEVATOR CONSTRUCTOR	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
210		ELEVATOR CONSTRUCTOR HEL...	Journeymen	281.25000	281.25000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	140.00000	141.25000	HOUR...
211		ELEVATOR CONSTRUCTOR HEL...	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
212	any	CARPENTER	Journeymen	120.00000	120.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	120.00000	HOUR...
213	any	CARPENTER	Apprentice	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	HOUR...
214	any	LABORER	Journeymen	575.00000	463.00000	0.00000	112.00000	0.00000	0.00000	0.00000	0.00000	0.00000	575.00000	HOUR...
215	any	LABORER	Apprentice	157.00000	0.00000	0.00000	157.00000	0.00000	0.00000	0.00000	0.00000	157.00000	0.00000	HOUR...

Figure 4, Distribution of Apprentice vs Journeymen hours annually