# WGBH - Color of Money

Jena Jordahl, Murt Bahrani, Carmen Araújo, Richard Lee, Elisa Cordeiro Lopes

CS506 - Spring 2021

# MOTIVATION

**With so much racial tension, has it changed anything?**

**Massachusetts major construction projects**

- Summary of Workforce Utilization

- Work: pandemic, ethnicity, gender, influencing factors, ...

- **Goal: PDF -> CSV-> dataframe or PDF -> dataframe**

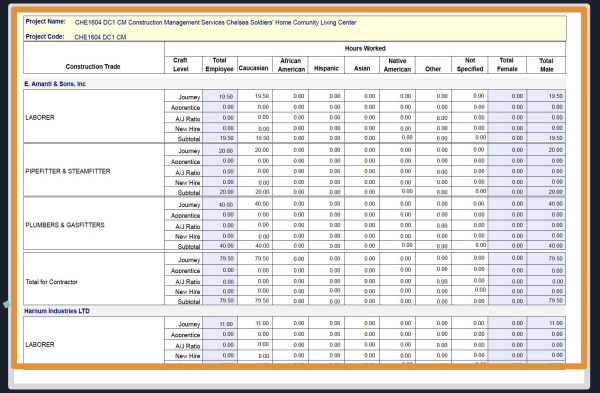
# Current Progress

**Method:** configure bounding box parameters

+ tabula PDF reader -> pandas

+ PyPDF2 page count for tabula

+ custom parser groups report lines per pd.df and per page



Project Name: CHE1604 DC1 CM Construction Management Services Chelsea Soldiers' Home Comunity Living Center

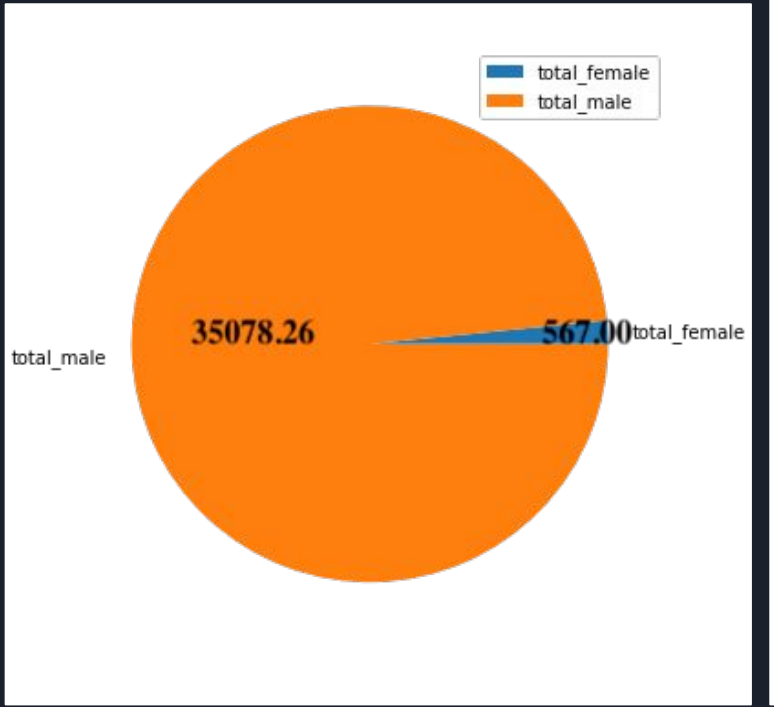
Project Code: CHE1604 DC1 CM

| Construction Trade | | Hours Worked: Craft Level | Total Employee | Caucasian | African American | Hispanic | Asian | Native American | Other | Not Specified | Total Female | Total Male |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E. Amanti & Sons, Inc | | | | | | | | | | | | |
| LABORER | Journey | 19.50 | 19.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 19.50 |
| | Apprentice | 0.00 | 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 | 0.00 |
| | New Hire | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Subtotal | 19.50 | 19.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 19.50 |
| PIPEFITTER & STEAMFITTER | Journey | 20.00 | 20.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.00 |
| | Apprentice | 0.00 | 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 | 0.00 |
| | New Hire | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Subtotal | 20.00 | 20.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.00 |
| PLUMBERS & GASFITTERS | Journey | 40.00 | 40.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 40.00 |
| | Apprentice | 0.00 | 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 | 0.00 |
| | New Hire | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Subtotal | 40.00 | 40.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 40.00 |
| Total for Contractor | Journey | 79.50 | 79.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 79.50 |
| | Apprentice | 0.00 | 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 | 0.00 |
| | New Hire | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Subtotal | 79.50 | 79.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 79.50 |
| Harnum Industries LTD | | | | | | | | | | | | |
| LABORER | Journey | 11.00 | 11.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 11.00 |
| | Apprentice | 0.00 | 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 | 0.00 |
| | New Hire | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

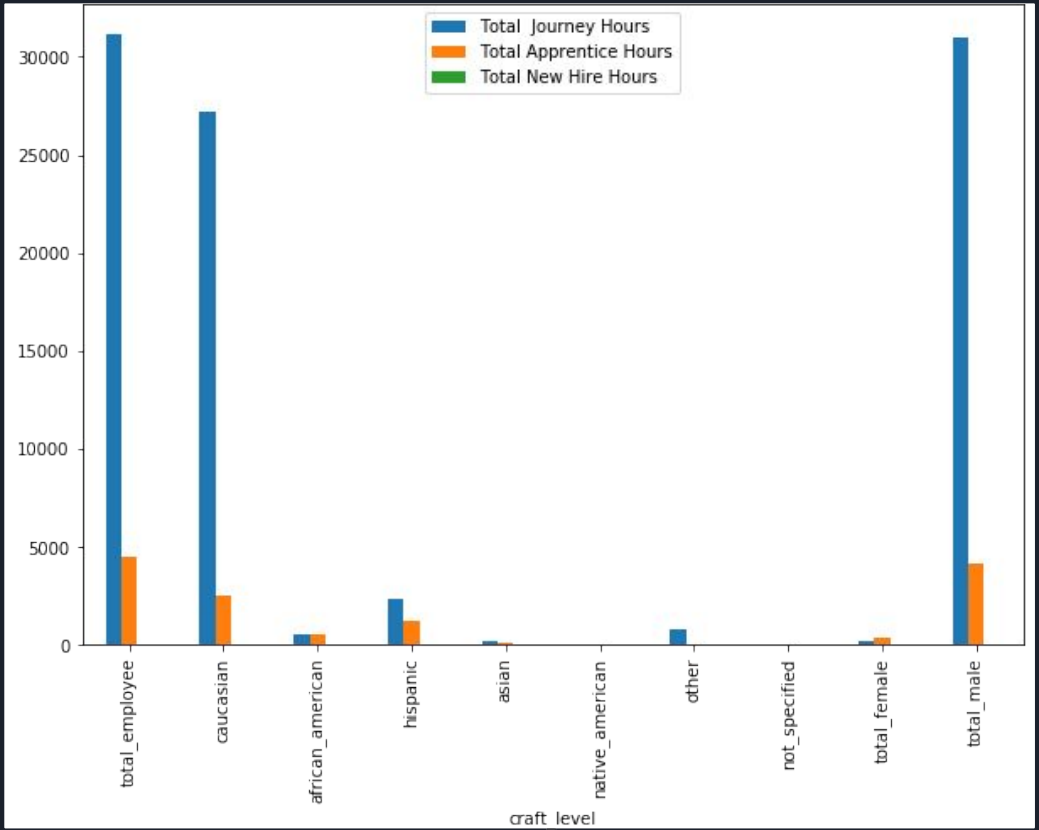
| | OR | CONSTRUCTION_TRADE | CRAFT_LEVEL | TOTAL_EMPL... | CAUCASIAN | AFRICAN_... | HISPANIC | ASIAN | NATIVE_A... | OTHER | NOT_SPEC... | TOTAL_FE... | TOTAL_MALE | HOUR... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 197 | | GLAZIER | Apprentice | 64.00000 | 0.00000 | 0.00000 | 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... |

| 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 |

# Early Results




**Total number of hours by ethnicity split up by Journey, Apprentice, and New Hires**

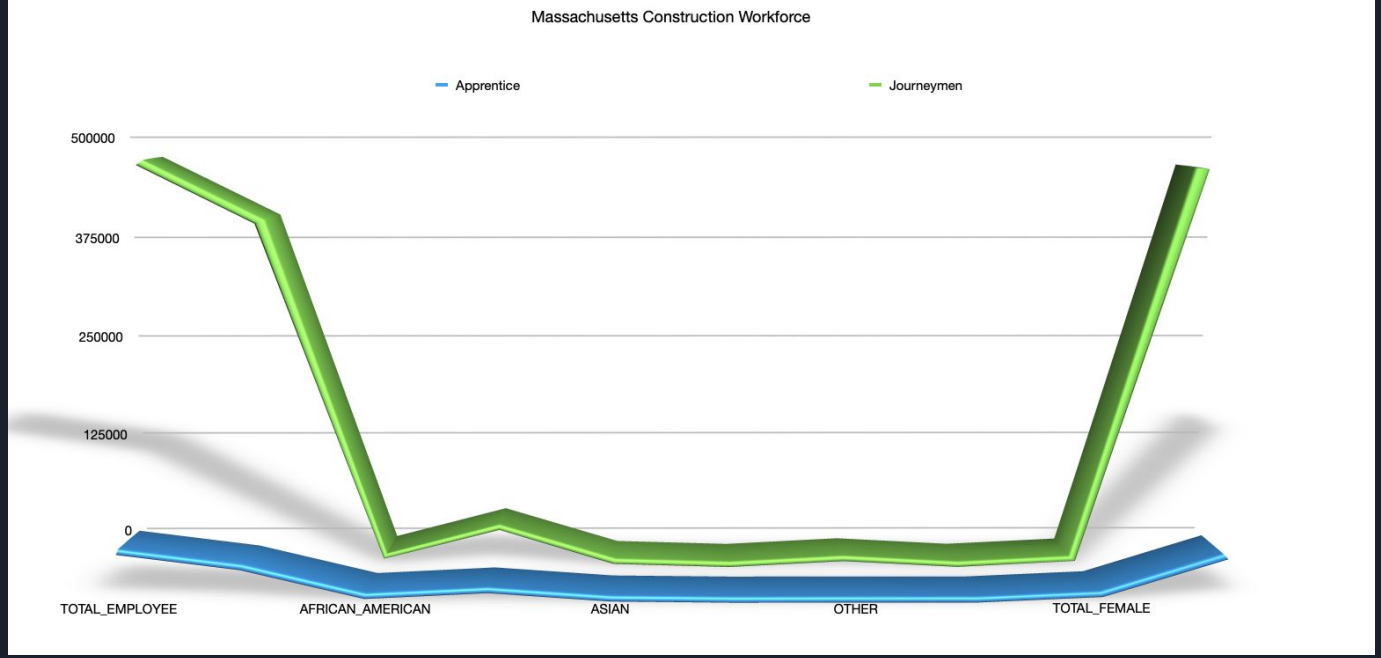


# **Findings** Apprentice vs Journeymen

## If you don't apprentice you cannot become a Journeymen

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 |



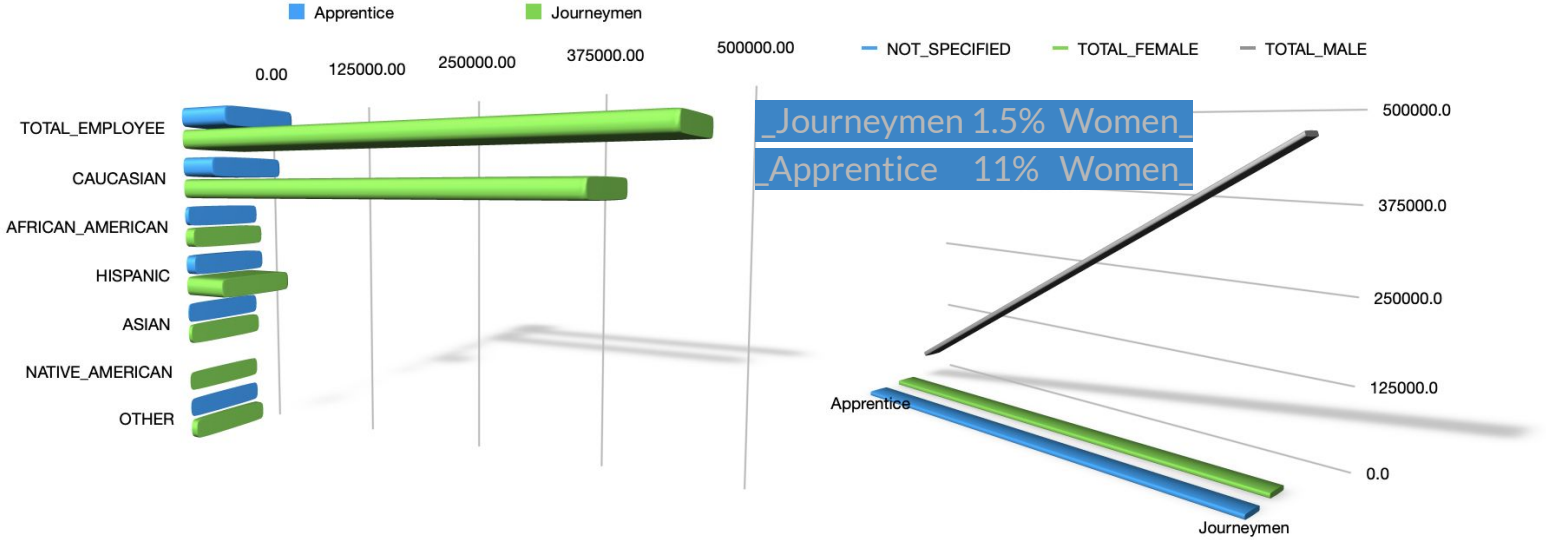
# KEY FINDINGS

Journeymen 86% Caucasian Americans
Apprentice 68% Caucasion Americans
98% Male Americans
88% Male Americans
Journeymen 2% African Americans
Apprentice 8% African Americans

The color of money is significant in 2019, and so is its gender.

WorkforceCraftSummaryReport2019

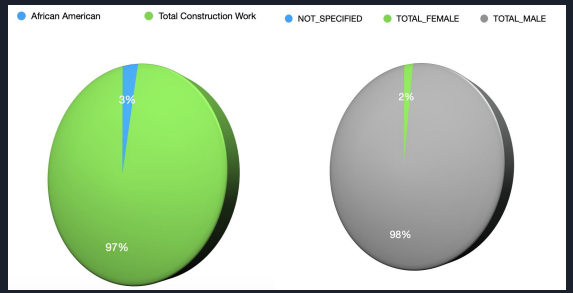
| CRAFT_LEVEL | TOTAL_EMPLOYEE | CAUCASIAN | AFRICAN_AMERICAN | HISPANIC | ASIAN | NATIVE_AMERICAN | OTHER | NOT_SPECIFIED | TOTAL_FEMALE | TOTAL_MALE |
|---|---|---|---|---|---|---|---|---|---|---|
| **Apprentice** | 54893.25 | 37507.75 | 4592.0 | 11190.5 | 1475.5 | 0.0 | 127.5 | 0.0 | 6109.75 | 48783.5 |
| **Journeymen** | 476507.39 | 408302.44 | 10689.83 | 45692.34 | 4121.5 | 209.5 | 7336.44 | 155.34 | 7079.44 | 469427.95 |



# Next Steps

- Discussing on possible analysis to do

ONE YEAR IS NOT AS IMPORTANT AS A 10 YR TREND




SOME PEOPLE FEEL BLOCKED IN EARNING

ENOUGH HOURS TO BECOME A JOURNEYMEN

How many different companies get contracts?

How many distinct companies are all one race?

How much bigger are the big companies?

PG 1 of 82
for APRIL

# Data Format Challenges

DF0

Report Header
Separate DF:
(3 rows & two categories)
Project Information,
Column Header Info

Trade Totals (5 lines):
Journeymen, Apprentice
A/J ratio, New Hire,
Trade SubTotal

Contractor Totals (5 lines):
Journeymen, Apprentice
A/J ratio, New Hire,
Contractor SubTotal

**Project Name:** CHE1604 DC1 CM Construction Management Services Chelsea Soldiers' Home Comunity Living Center

**Project Code:** CHE1604 DC1 CM

| Construction Trade | Craft Level | Hours Worked: Total Employee | Caucasian | African American | Hispanic | Asian | Native American | Other | Not Specified | Total Female | Total Male |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **E. Amanti & Sons, Inc** | | | | | | | | | | | |
| LABORER | Journey | | | | | | | | | | |
| | 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 | | | | | | | | | | |
| PIPEFITTER & STEAMFITTER | Journey | 20.00 | 20.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.00 |
| | 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 | 20.00 | 20.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.00 |
| PLUMBERS & GASFITTERS | Journey | | | | | | | | | | |
| | 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 | | | | | | | | | | |
| Total for Contractor | Journey | 79.50 | 79.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 79.50 |
| | 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 | 79.50 | 79.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 79.50 |
| **Harnum Industries LTD** | | | | | | | | | | | |
| LABORER | Journey | | | | | | | | | | |
| | 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 | 11.00 | 11.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 11.00 |

Contractor Header (1 lines):
Name of Contractor without any labeling

DF1

# Data Format Challenges

33,000 lines of text and 1,097 pages

2019 Data

- 5,540 Detail lines

DF2

## LIMITATIONS

We need more years for trend analysis

Report Header
Separate DF:
(3 rows & two categories)
Project Information
After 1st, throw away the column heading info

Project Totals (4 lines):
Journeymen, Apprentice
New Hire,
Contractor SubTotal

Project End (1 lines):
Project Description

**Project Name:** CHE1604 DC1 CM Construction Management Services Chelsea Soldiers' Home Comunity Living Center

**Project Code:** CHE1604 DC1 CM

| Construction Trade | Hours Worked | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Craft Level | Total Employee | Caucasian | African American | Hispanic | Asian | Native American | Other | Not Specified | Total Female | Total Male |
| Total Journey Hours | | 1,291.00 | 1,272.00 | 11.00 | 8.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1,291.00 |
| Total Apprentice Hours | | 48.00 | 8.00 | 40.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 48.00 |
| Total New Hire Hours | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grand Total Hours | | 1,339.00 | 1,280.00 | 51.00 | 8.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1,339.00 |

*End of data for project:* CHE1604 DC1 CM Construction Management Services Chelsea Soldiers' Home Comunity Living Center

PG 2 of 82
for APRIL

```
Jan2019 pages 61
Feb2019 pages 63
Mar2019 pages 75
Apr2019 pages 82
May2019 pages 89
Jun2019 pages 90
Jul2019 pages 81
Aug2019 pages 85
Sep2019 pages 103
Oct2019 pages 108
Nov2019 pages 129
Dec2019 pages 131
Total pages processed by the parser: 1097
```

DF3

# CHALLENGES

- Directly read the PDF file into a Pandas DataFrame, each pg .. df
- Manually transformed into CSV
- PyPDF2, Tabula, Camelot, Pandas, PDFplumber, and batch Tabula-CSV

```
[],
[],
[16.0, 8.0, 0.0, 0.0, 0.0, 0.0, 8.0, 0.0, 0.0, 16.0],
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0],
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0],
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0],
[16.0, 8.0, 0.0, 0.0, 0.0, 0.0, 8.0, 0.0, 0.0, 16.0],
[],
[],
[],
[],
[],
[16.0, '8.00'],
['0.000.000.000.008.00', '0.000.00'],
['16.00'],
[0.0, 16.0],
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0],
```

Hindsight:
The core process of a parser is identify and process, “IP”.
Around the “IP” ,
Is logic to get the
Pages and
Data Frames

- CUSTOM Parser

```
# To test for a smaller number of pages set m to a small number
while proc_pages <= m:
    print(f'starting page {proc_pages}')
    df0 = tabula.read_pdf(filename, pages=proc_pages, lattice=True, area=(11, 26, 582, 829),
                          pandas_options={'header': None})
    rows = 0
    proc_dfs = 0
    while proc_dfs < len(df0):
        print(f' Processing DF number: {proc_dfs}')
        print(df0[proc_dfs])
        print()
        proc_rows = 0
        while proc_rows < len(df0[proc_dfs]):
            print(f' Processing row number: {proc_rows}')
            if beginning:
                rows,COL,PROJCT,PROJCT_CD= start_process(df0[proc_dfs])
                beginning = False
            else:
                grp = identify_grp(df0[proc_dfs],proc_rows)
                rows, remaining,STOREIT, NEWHIRE, PROJCT, PROJCT_CD, CONTRACTOR, CONSTR_TRAD
                    process_grp(remaining, df0[proc_dfs],proc_rows,grp, STOREIT, NEWHIRE, CO
            proc_rows += rows
        proc_dfs += 1
    proc_pages += 1
    if debug:
        print(STOREIT)
#         "./data/WorkforceUtilizationSummaryReportApril2019.pdf")
outfile1 =  outfile + MONTHYR[3:7] + MONTH + '.csv'
STOREIT.to_csv(outfile1, index=False)
```

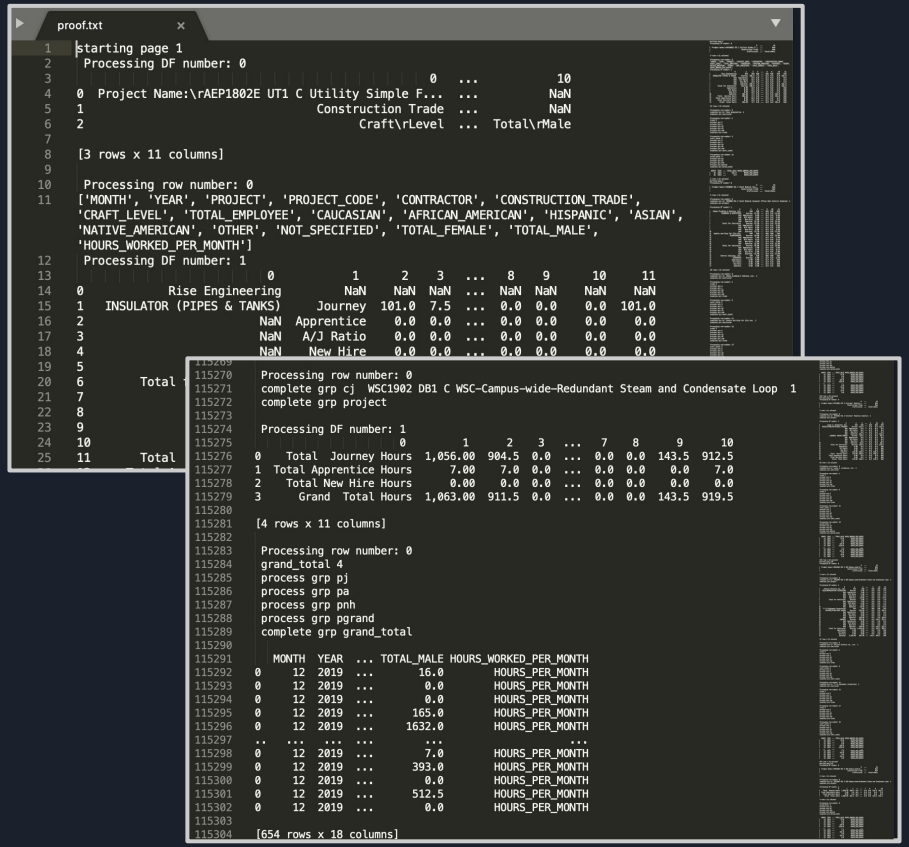
| | Project_Code | Project_Name | Contractor_Name | Construction_Trade |
|---|---|---|---|---|
| 0 | AEP1802E UT1 C | Utility Simple Fix | Batallas Electric Inc. | ELECTRICIAN |
| 1 | AEP1802E UT1 C | Utility Simple Fix | Batallas Electric Inc. | LABORER |
| 2 | AEP1802E UT1 C | Utility Simple Fix | Rise Engineering | INSULATOR (PIPES & TANKS) |
| 3 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | E. Amanti & Sons Inc | LABORER |
| 4 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | E. Amanti & Sons Inc | PIPEFITTER & STEAMFITTER |
| 5 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | E. Amanti & Sons Inc | PLUMBERS & GASFITTERS |
| 6 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | Harnum Industries LTD | LABORER |
| 7 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | S&F Concrete Contractors Inc. | EQUIPMENT OPERATOR (Class B CDL) |
| 8 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | S&F Concrete Contractors Inc. | LABORER |
| 9 | CHE1604 DC1 CM | Construction Management Services Chelsea Soldi... | W.L. French Excavating Corp. | ADS/SUBMERSIBLE PILOT |

# CHALLENGES:
# News stories need the facts to be in evidence

If WGBH wants to use the information we identify then they need proof that we parsed all that data according to the PDFs they received.

For use in a random sample test plan, the parser creates a proof.txt file to confirm that each row in the received PDF file was accurately categorized and processed in the final pandas DataFrame. On the right, please see excerpts from proof.txt:

FYI, proof.txt is 115,304 lines long.



Thank you!

# Reference

1min motivation / background (why does your project matter?)

1min30s current progress, results, findings (what are the key findings of your project?)

30s next steps, limitations, challenges