Project Deliverable 4

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

This is a draft of your final report that has been reviewed by your client. It includes all visualizations, results, data, and code up to this point, along with proper documentation on how to reproduce your results, compile and use your codebase, and navigate your dataset. Your team will submit this as a PR.

Introduction

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. However, these PDF files are constructed in a way where no regular PDF parser can easily extract data. They were also given in packages of hundreds of pages so that without any form of automation it would be very time consuming to be able to analyze the data.

After working through this project, our team thought of three important questions we wanted to answer. The first question was how we would extract data from our PDF files. Our PDF datasets were formatted in a way where it was extremely difficult to parse any data in a clean and precise manner. Our description of our attempts to parse this data will go into further detail below, however this portion of the project took up most of our time and a lot of time and effort was put into our PDF parser. The second question was if there was a difference between state-paid contractual hours based on color and/or sex. This was the meat of the project that we were fortunately able to get to because of the incredible work our team did to complete our parser. The third and last question we had was why all new hire hours were 0. This is highly unusual as it is quite common for workers to start a job casually whenever they see an opening.

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. On April 12, DCAMM released monthly reports for 2020.

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 and 2020 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

Creating the Parser

Phase 1: We used Python libraries PyPDF2 and Tabula to scrape the data from the PDF files and then used acrobat to save the files in CSV format. Each method produced the same misalignments between the hourly data rows and the company/trades header data. This issue stemmed from the PDF merging cells to pretty print the data for human readability. Initially we had no idea what was necessary so we attempted to set up Grobid on SCC to parse the PDFs into XML files. After our first 2 weeks we had tried 7 different methods (real time, tabula software, tabula library, Grobid, PyPDF2, managing data after tabula, and transforming data back into PDFs)

Phase 2: We were soon able to build a parser that extracted individual project names, project codes, and the contractors/companies involved in each. However, this parser still missed a few contractors/companies in each listing. We had many issues using commas as a splitter as it divided the numbers containing commas into two.

Phase 3: Successfully created a parser that could extract and create a pandas dataframe with columns: project id & name, contractor name, construction trade name, and detail lines per trade level name. Our next issue was extracting numbers with ".". When separating numbers into distinct entries, we were getting arrays like: ['0.000.000.000.000.00', '2.800.00'] , instead of: [0.0, 0.0, 0.0, 0.0, 0.0, , 2.8, 0.0].

Phase 4: After our first parser, the team spent a lot of time trying to find alternate solutions to our problems because nothing was working out. Fortunately, we were finally able to create a program that produces clean dataframes. This was done by reading the data directly from the PDF into pandas dataframes and using a Json format to make a custom shape for our PDFs to be extracted from. However, we still ran into many issues such as: inconsistencies in the data reading empty numbers, misaligned rows and columns, two columns for one number, rows being chopped off and placed in a different row, and white space characters embedded in column heading fields.

Phase 5: All bugs listed in Phase 4 were fixed and our parser successfully worked on different datasets.

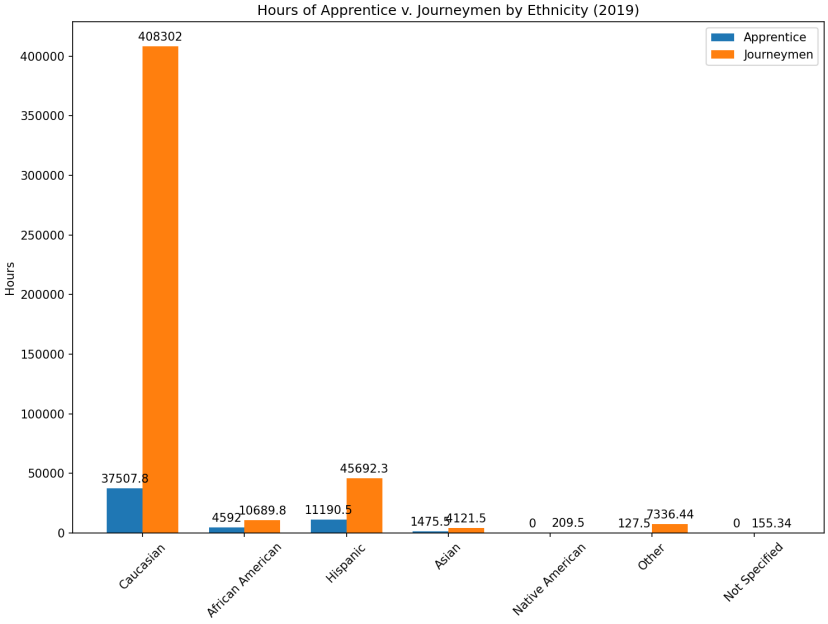
In Depth Description of our Working Parser:

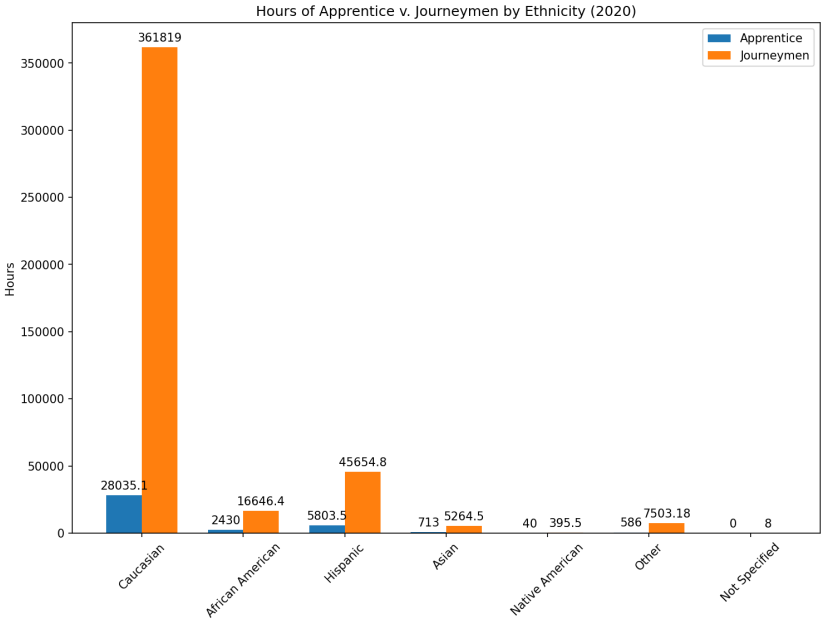
We used PyPDF2 to extract the number of pages, the interactive tabula tool to create bounding box x,y coordinates, and tabula python library which utilizes pandas directly when it chunks out the file page by page. Next, we created a python script to read each pdf file in the input directory and produce 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.

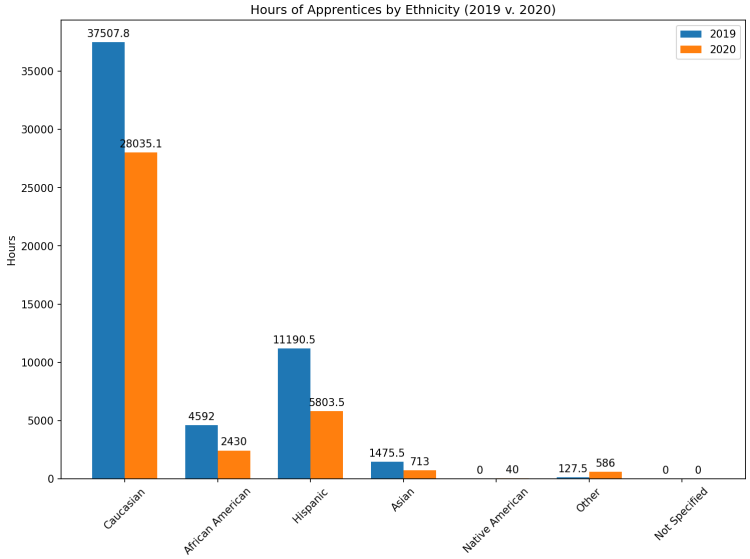
By generating the monthly columnar CSV files, we can build a mini-data mart for querying four different hierarchical trees. One tree for Project/Contractor/Trade/Experience Level, one for time series(month and year) and two others for Ethnicity and Gender. All arms of the tree tie count hours worked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

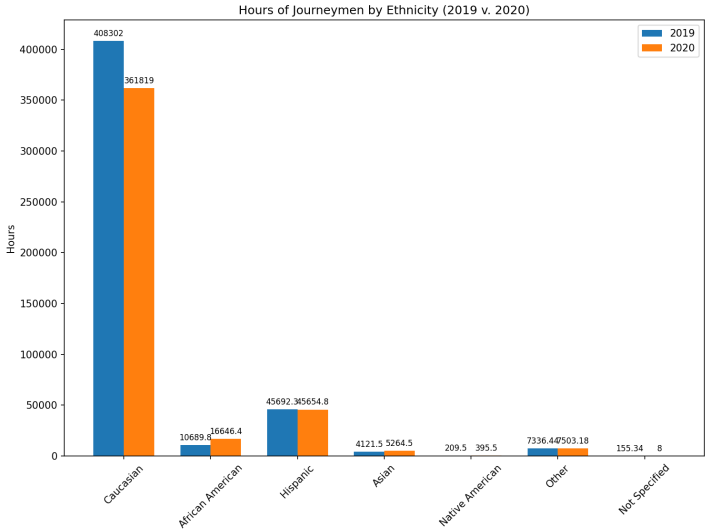
Analysis

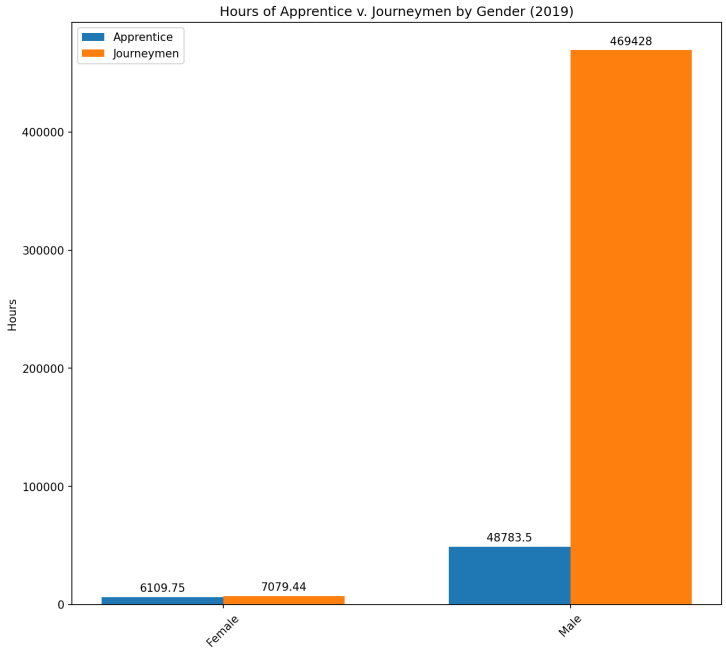
Preliminary, feature extraction, graphs, plots, models

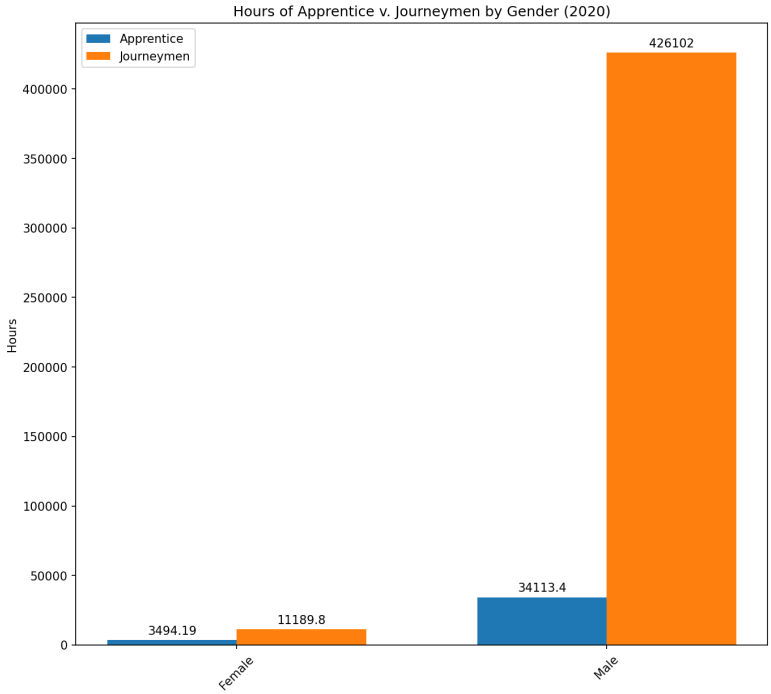

Hours of Apprentice v. Journeymen by Ethnicity (2019)
Apprentice
Journeymen
Hours
0
50000
100000
150000
200000
250000
300000
350000
400000
37507.8
408302
4592
10689.8
11190.5
45692.3
1475.5
4121.5
0
209.5
127.5
7336.44
0
155.34
Caucasian
African American
Hispanic
Asian
Native American
Other
Not Specified

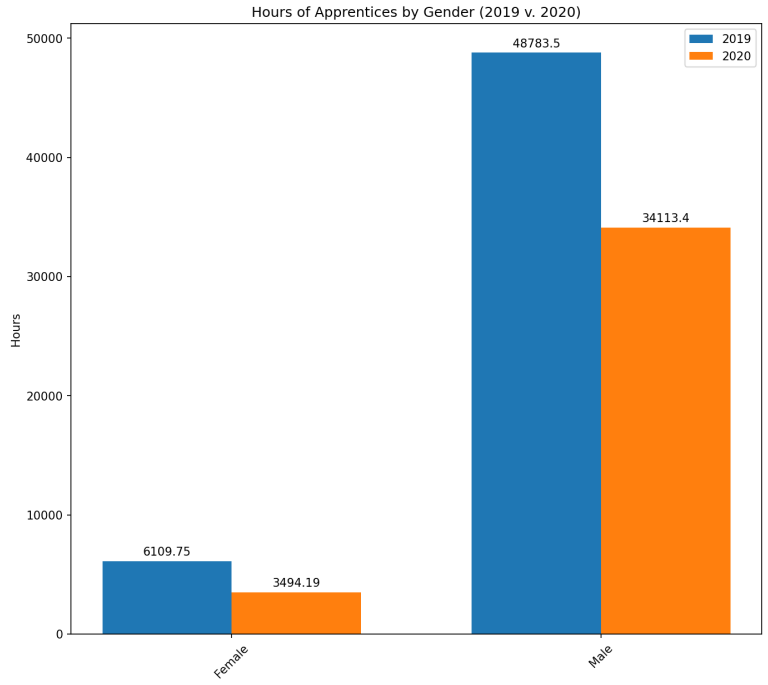

Hours of Apprentice v. Journeymen by Ethnicity (2020)
Apprentice
Journeymen
Hours
0
50000
100000
150000
200000
250000
300000
350000
28035.1
361819
2430
16646.4
5803.5
45654.8
713
5264.5
40
395.5
586
7503.18
0
8
Caucasian
African American
Hispanic
Asian
Native American
Other
Not Specified

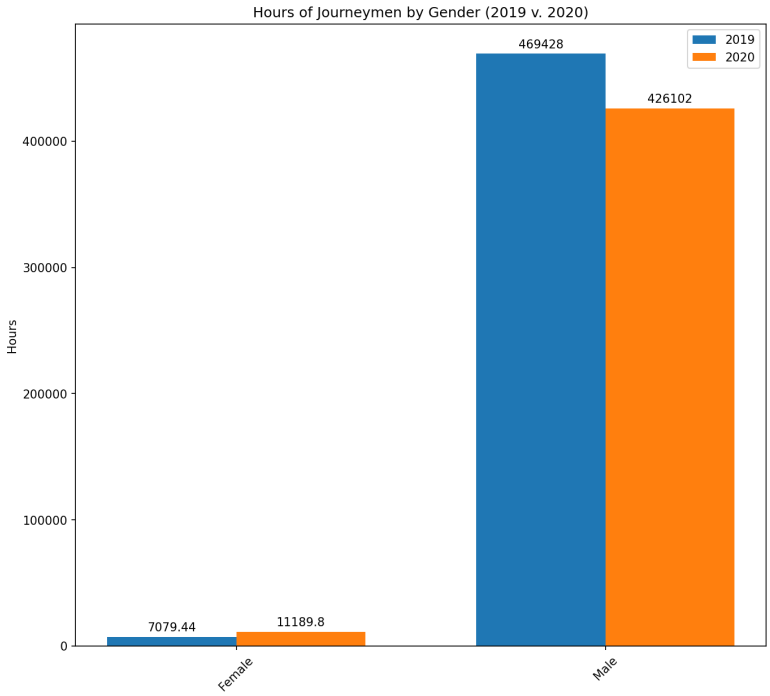

Hours of Apprentices by Ethnicity (2019 v. 2020)
2019
2020
Hours
37507.8
28035.1
4592
2430
11190.5
5803.5
1475.5
713
0
40
127.5
586
0
0
Caucasian
African American
Hispanic
Asian
Native American
Other
Not Specified


Hours of Journeymen by Ethnicity (2019 v. 2020)
2019
2020
Hours
408302
361819
10689.8
16646.4
45692.3
45654.8
4121.5
5264.5
209.5
395.5
7336.44
7503.18
155.34
8
Caucasian
African American
Hispanic
Asian
Native American
Other
Not Specified


Hours of Apprentice v. Journeymen by Gender (2019)
Apprentice
Journeymen
Hours
0
100000
200000
300000
400000
6109.75
7079.44
48783.5
469428
Female
Male


Hours of Apprentice v. Journeymen by Gender (2020)
Apprentice
Journeymen
Hours
0
50000
100000
150000
200000
250000
300000
350000
400000
3494.19
11189.8
34113.4
426102
Female
Male





Conclusions

Three important findings stand out showing the differences in the amount of journeymen work, work year-over-year given the pandemic is one of the years, and the kind of work with the highest percentages given to ethnicities.

There appears to be a bias towards hiring caucasian men for journeymen work on MA construction contracts. The evidence shows that ethnicities and women get lower paying apprentice work but not the better and more long-term journeymen positions.

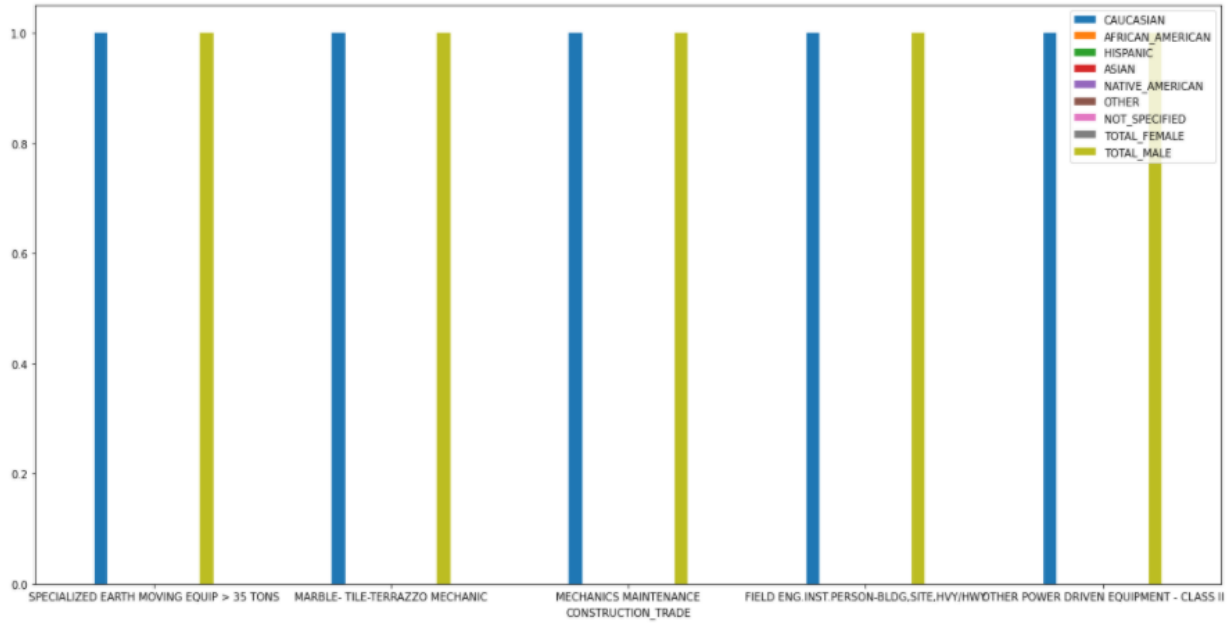
With respect to the pandemic's impact on the workforce, there is a drop in utilization hours in 2020 than 2019, and not as low as one might expect. Surprisingly, the ethnic makeup of the workforce is stable between the years. The consistency in very different economic times shows a tenacity for favoring caucasian males in construction work in the state.

More concerning is the nasty kind of work that pops to the top of the graph when the population is not split out by apprentice and journeymen: **Top 5 trades based on percentage of African Americans, shown in screenshots.** Asphalt Raker and Laborer Hazardous Waste are the top job percentages for African-Americans. For both African-Americans and Hispanics, Laborer Hazardous Waste work is more often tasked to them than caucasions.

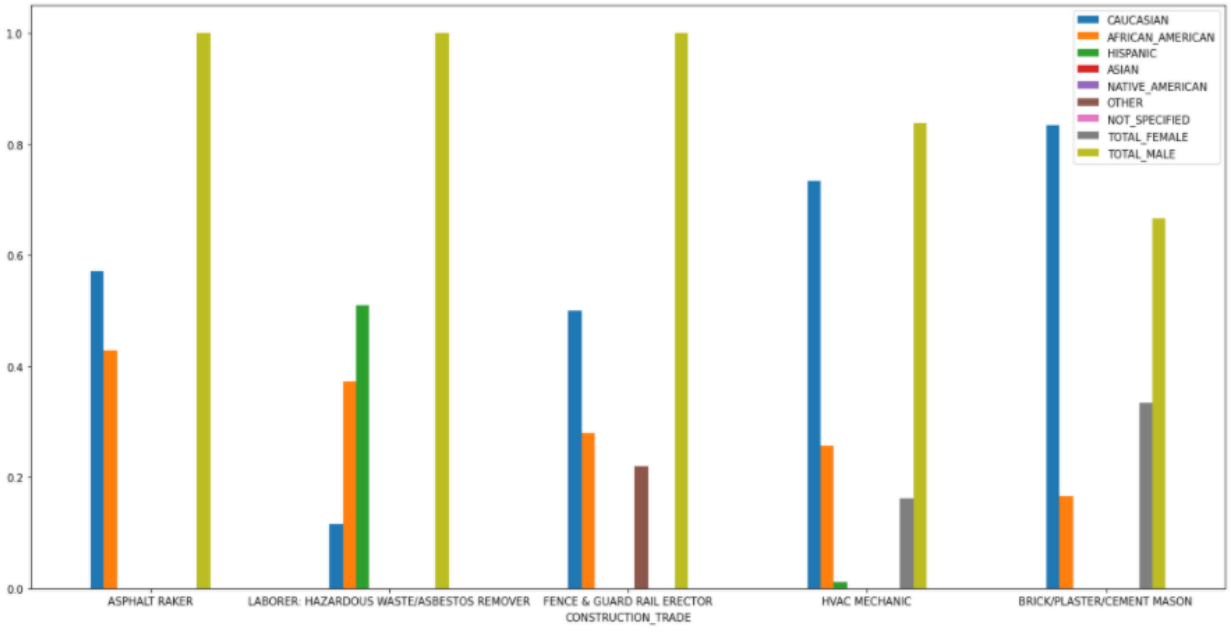
Limitations and Risks

Due to the past two years being very unusual, there is a risk in making assertions about the behavior of the state related to ethnicities and women. With more than 5 years of data, the information will have greater import. Still the data is hard evidence of journeymen work being greatly reduced for the ethnicities and women compared to caucasion men. The kinds of work appear to be more hazardous. The risk involved in making claims of racial or gender bias in the current social climate is that someone is bound to get upset with talking about the subject even if the numbers are accurate.

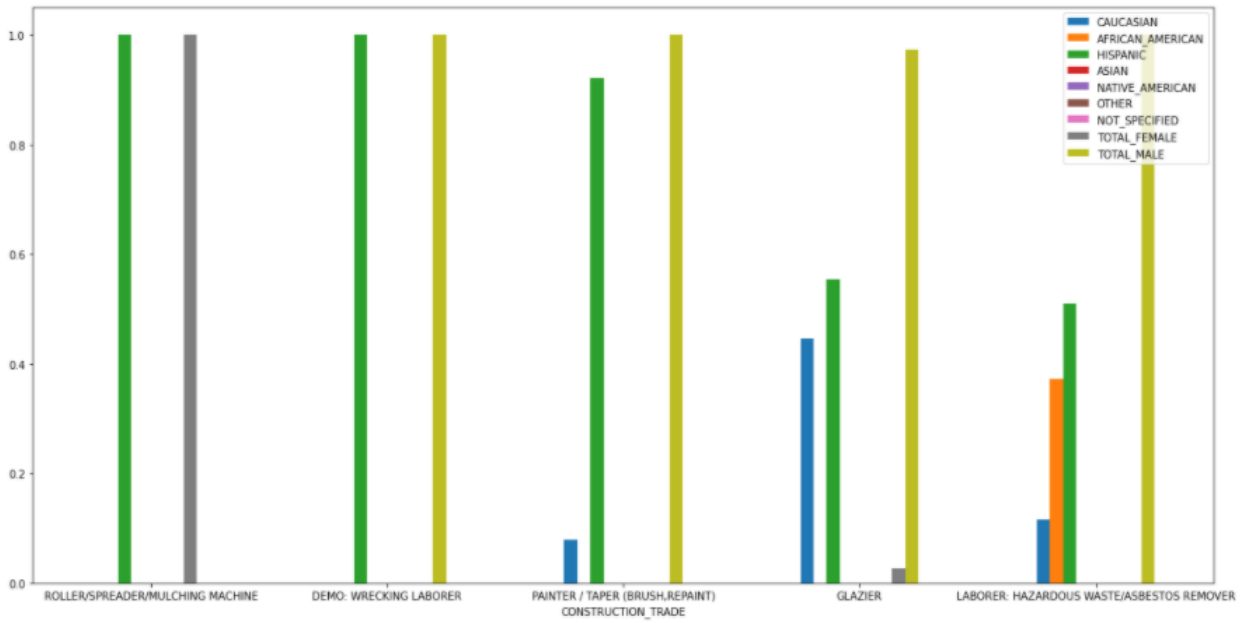
SCREENSHOTS



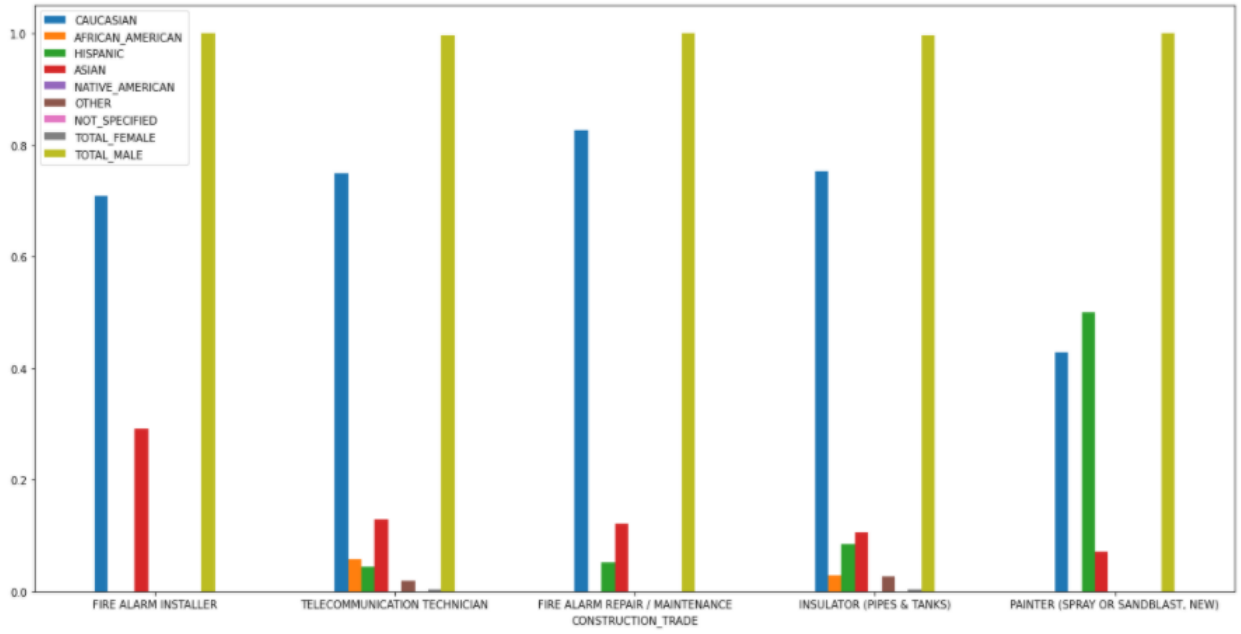
Top 5 trades ranked based on percentage of Caucasians (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



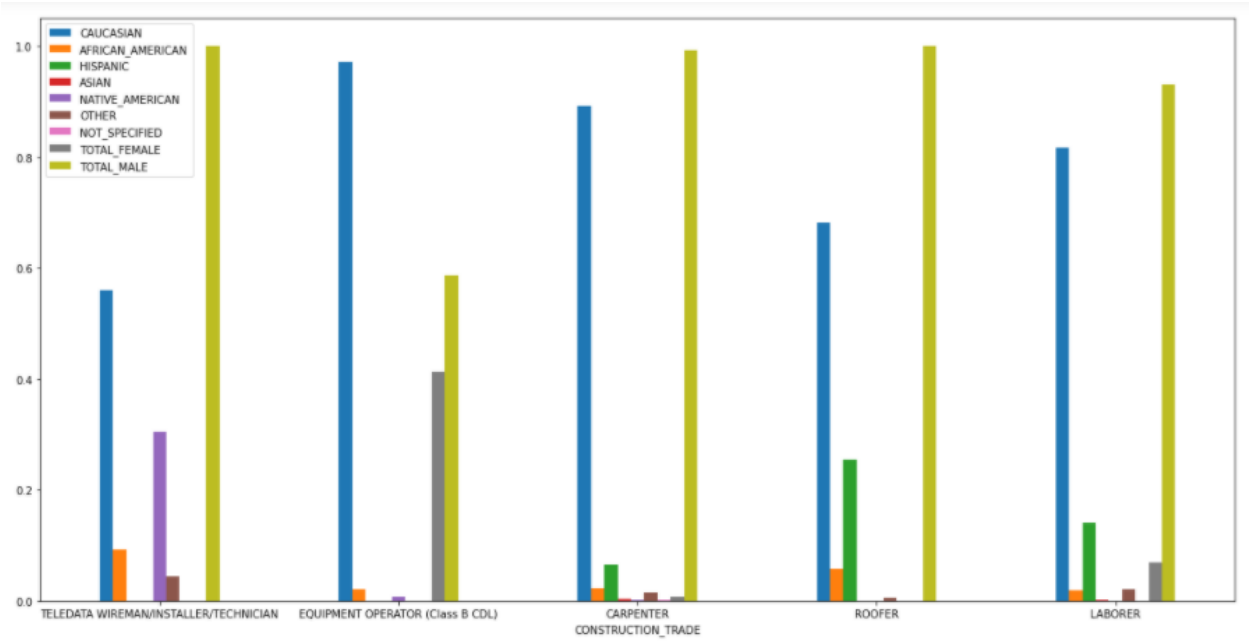
Top 5 trades based on percentage of African Americans (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



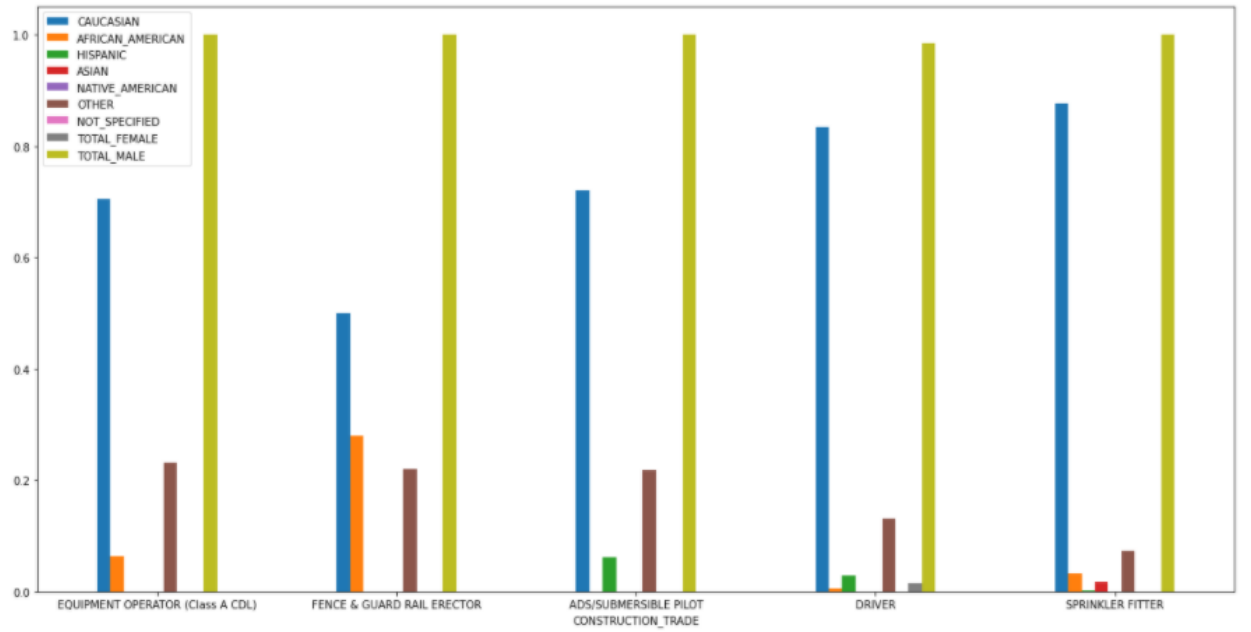
Top 5 trades based on percentage of Hispanics (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



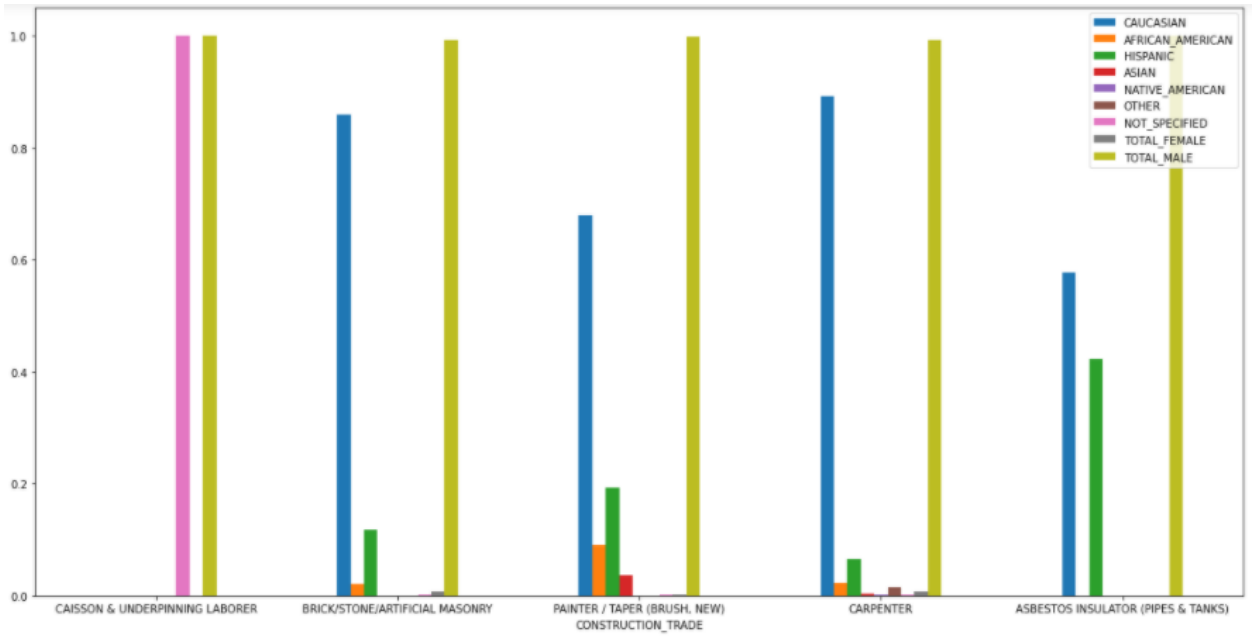
Top 5 trades based on percentage of Asians (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



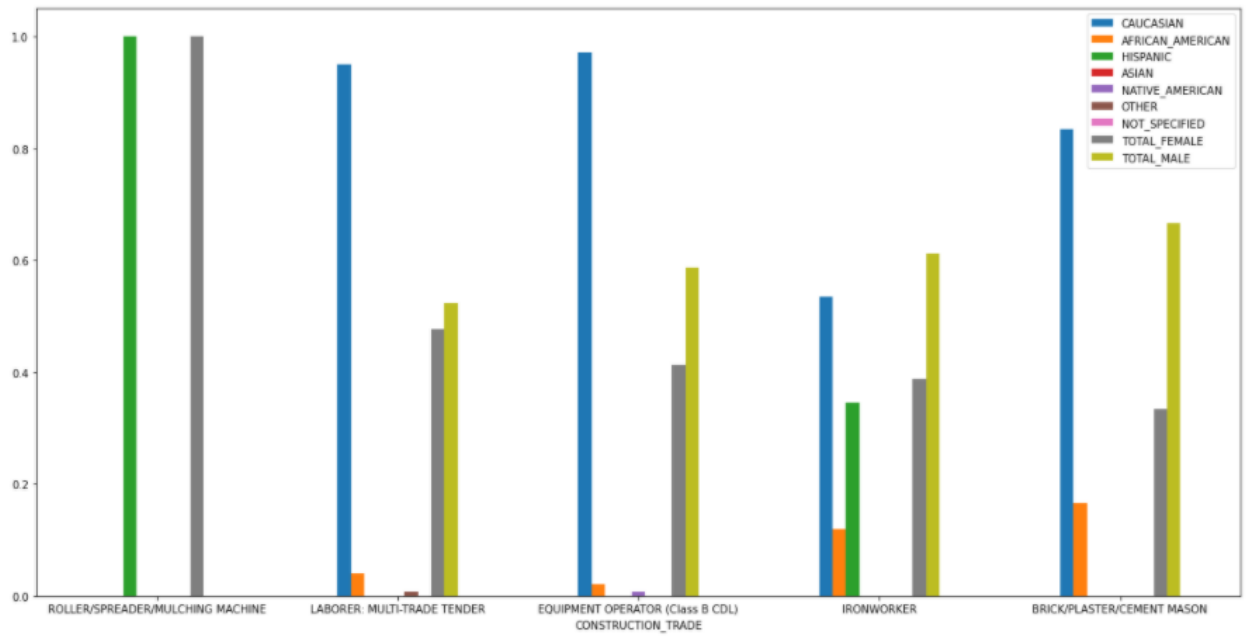
Top 5 trades based on percentage of Native Americans (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



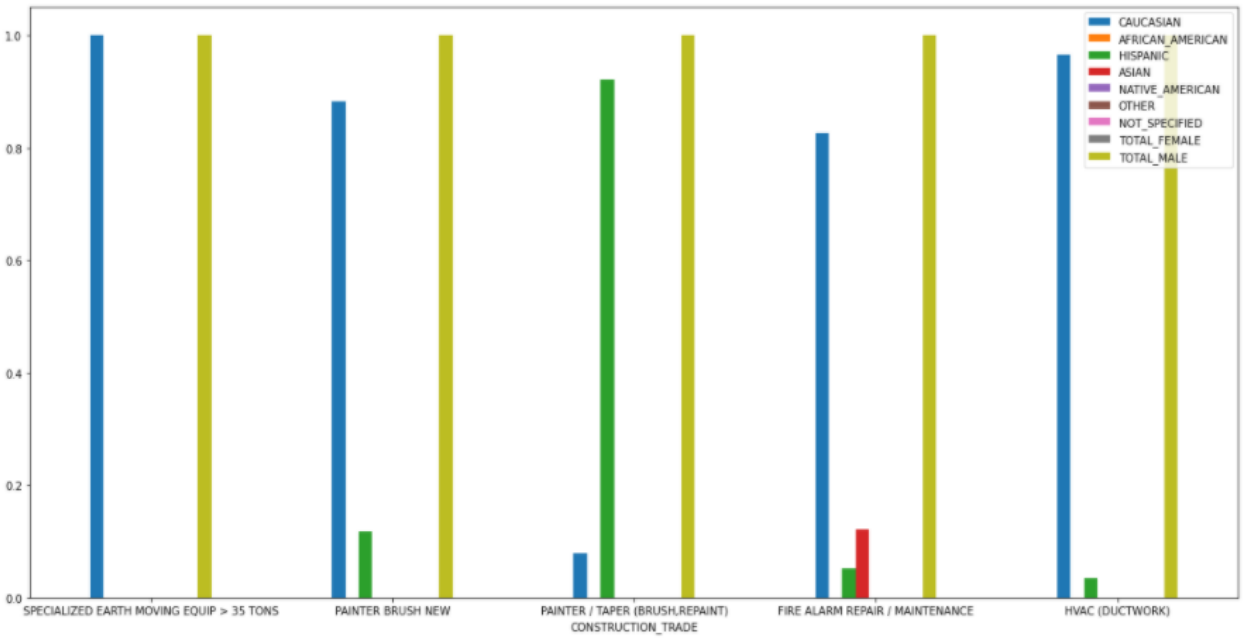
Top 5 trades based on percentage of Other (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



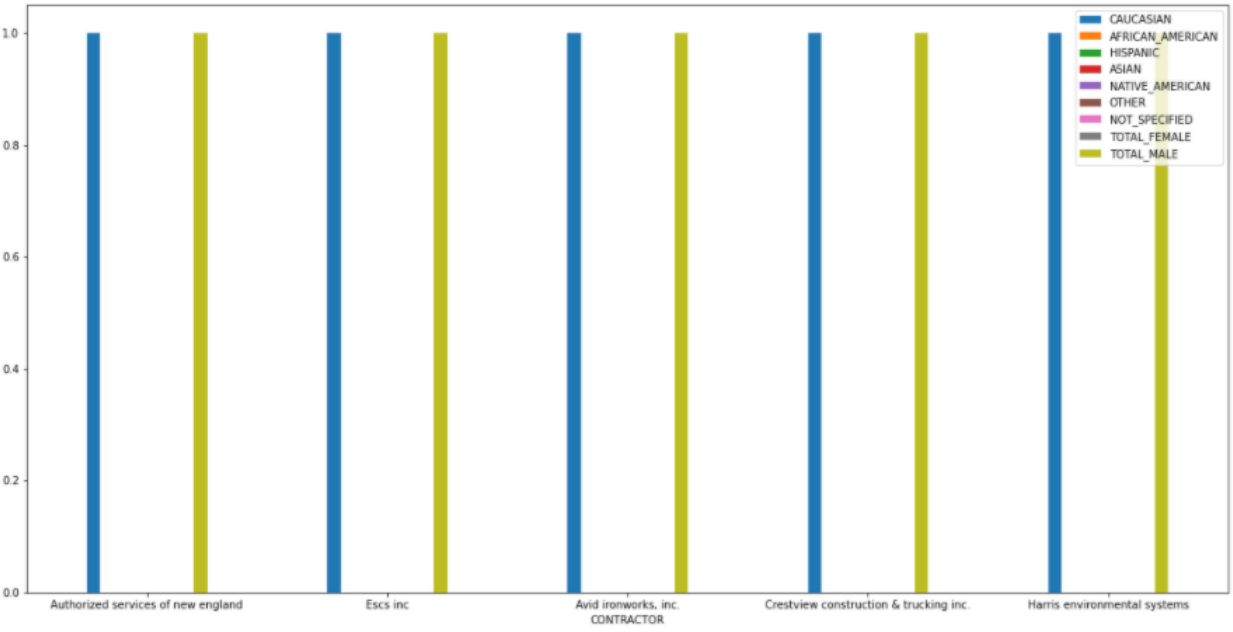
Top 5 trades based on percentage of Not Specified (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



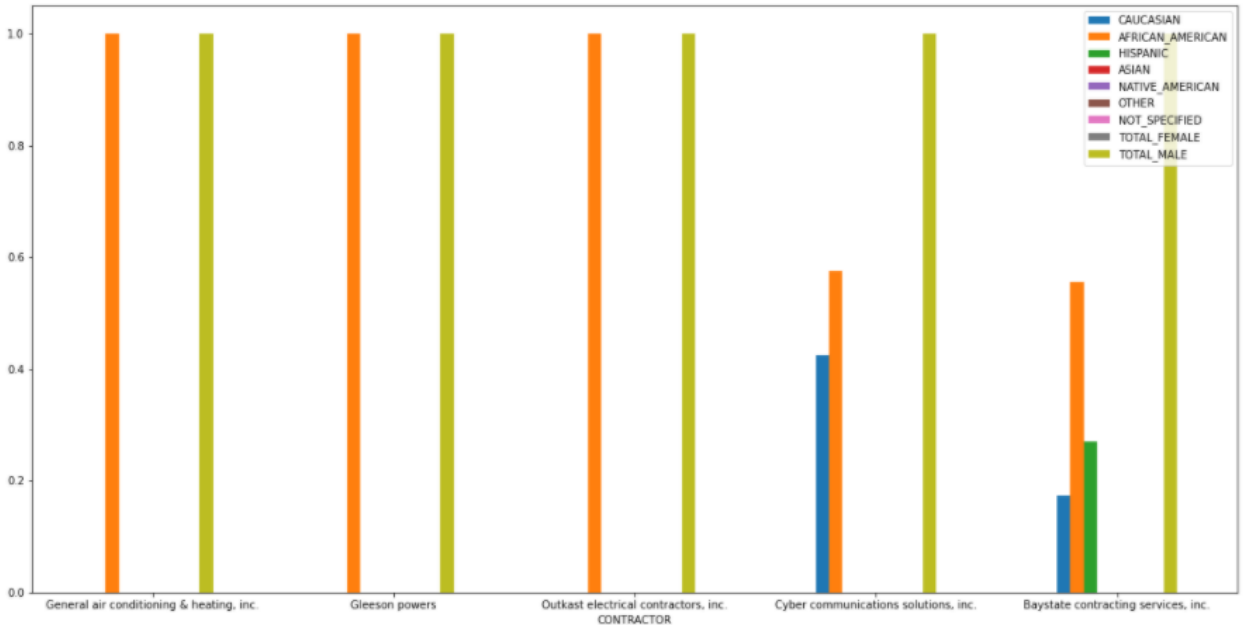
Top 5 trades based on percentage of Females (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



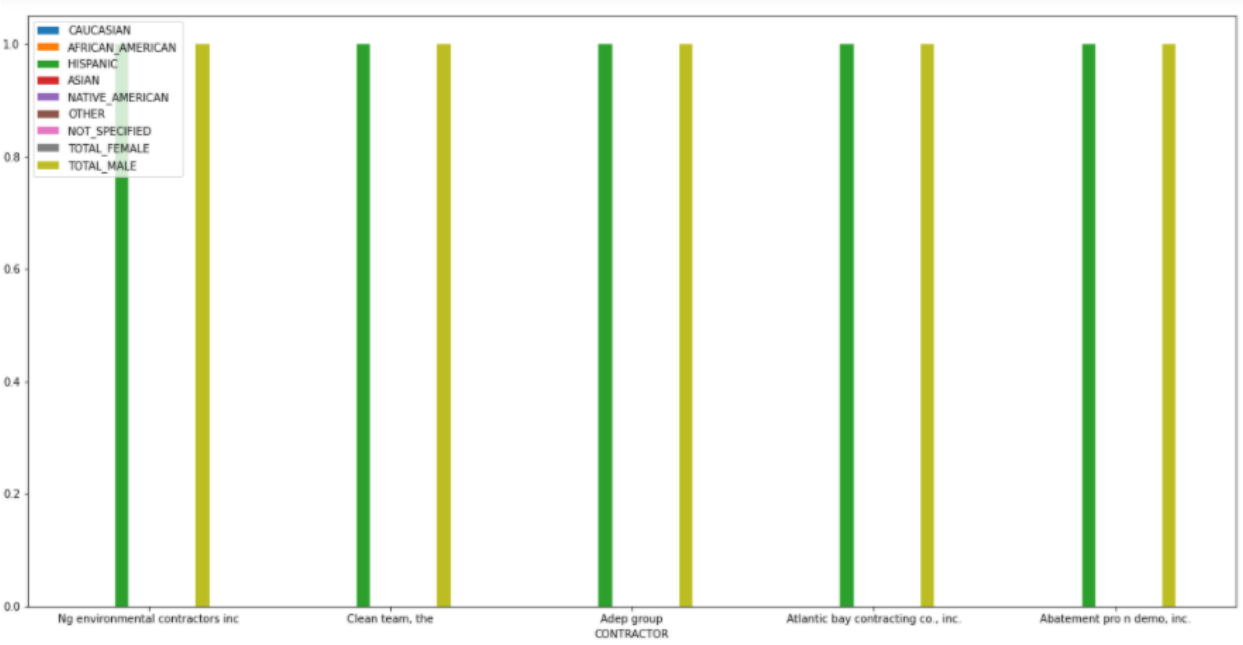
Top 5 trades based on percentage of Males (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



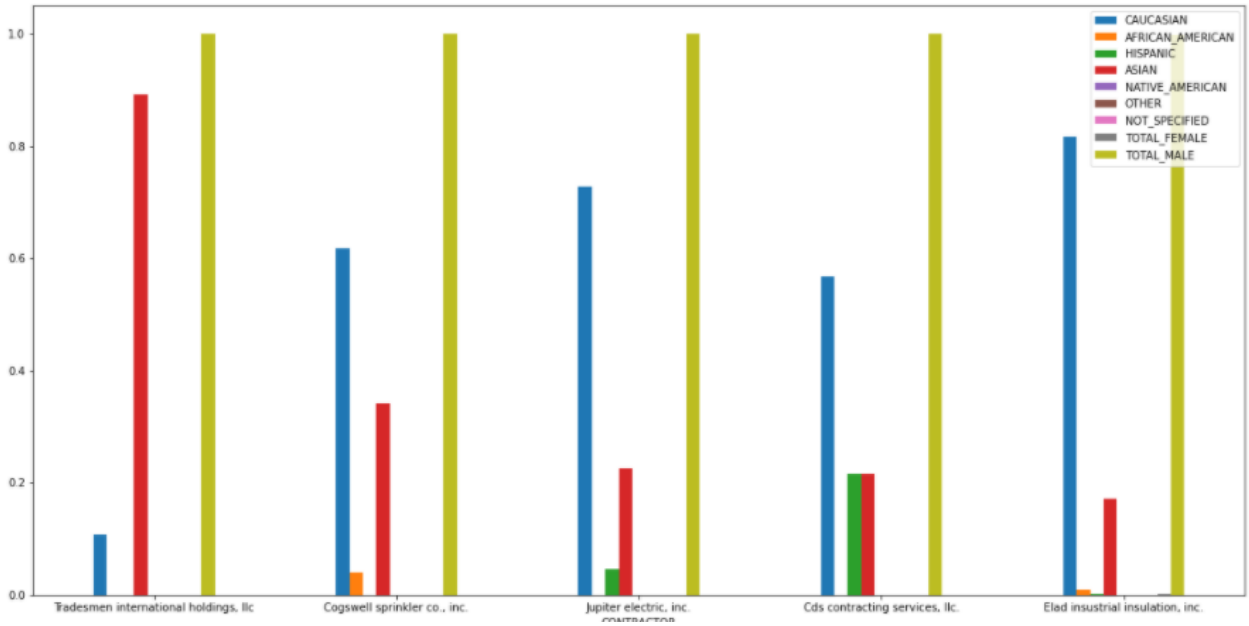
Top 5 companies based on percentage of Caucasians (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



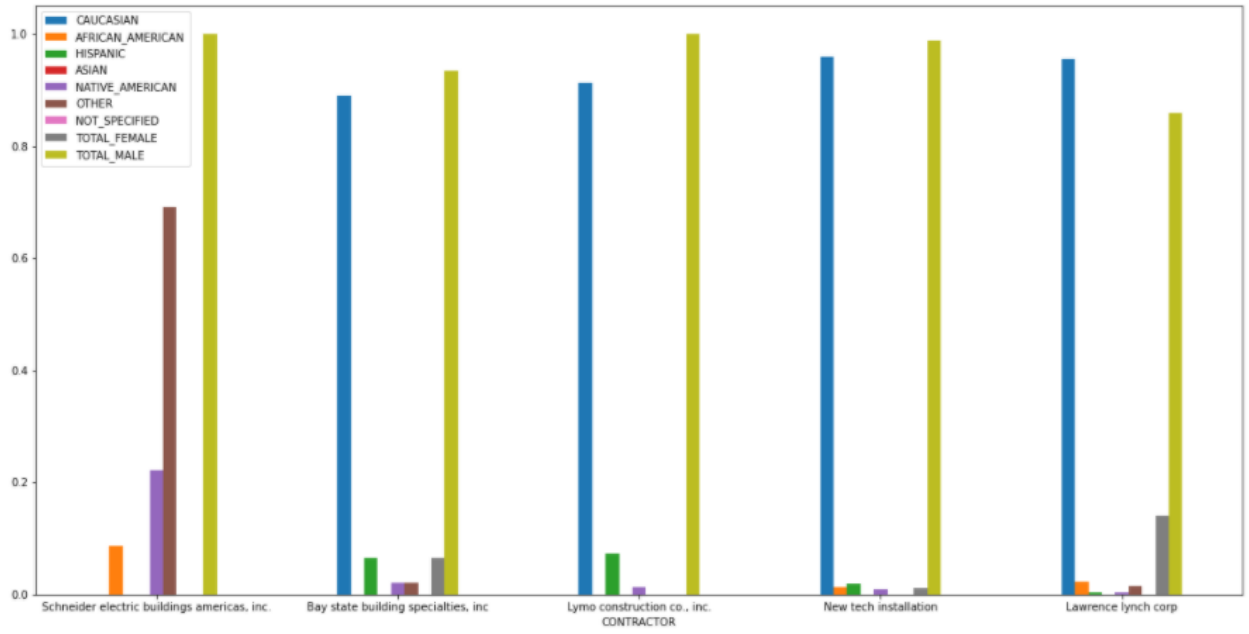
Top 5 companies based on percentage of African Americans (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



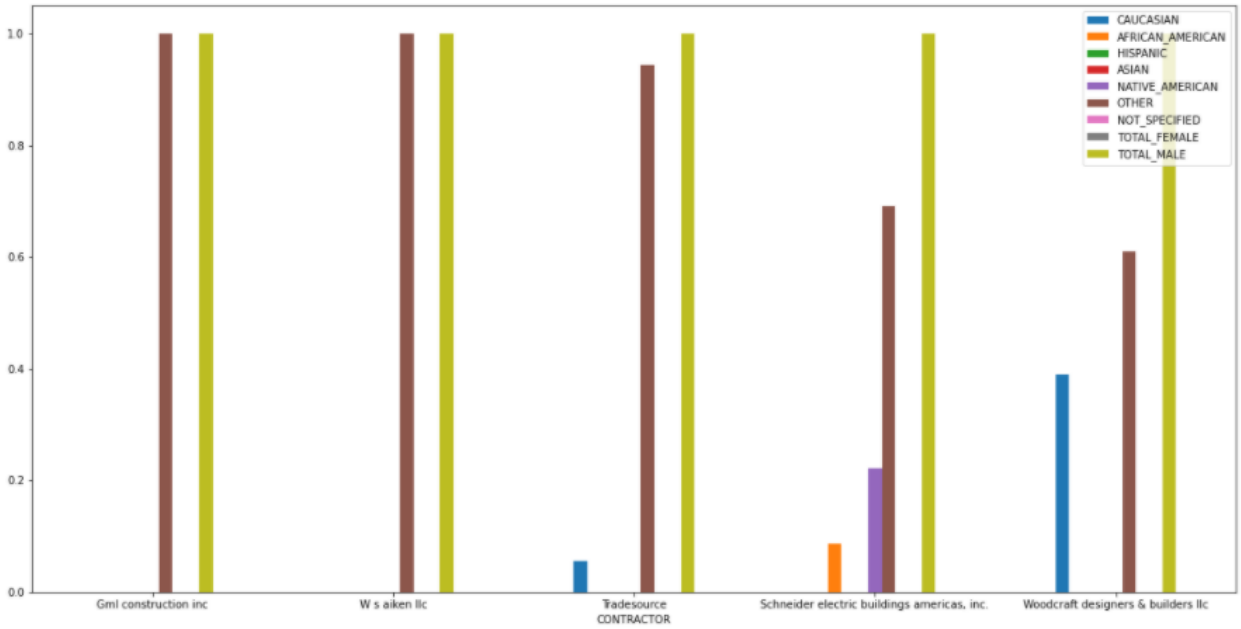
Top 5 companies based on percentage of Hispanics (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



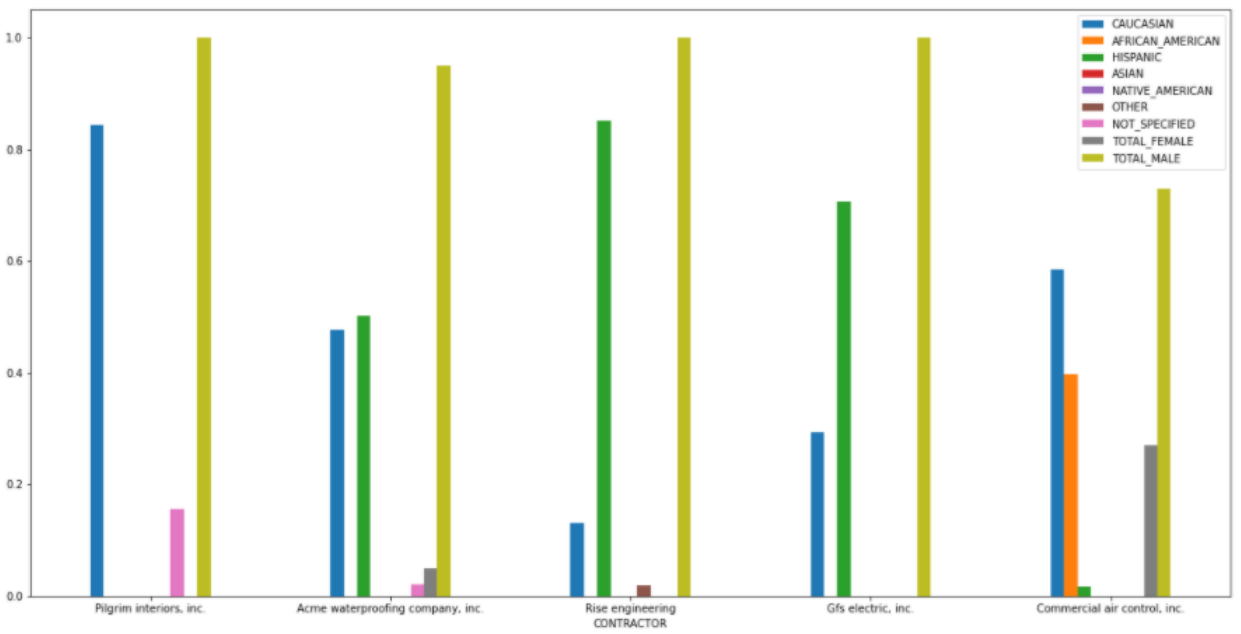
Top 5 companies based on percentage of Asians (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



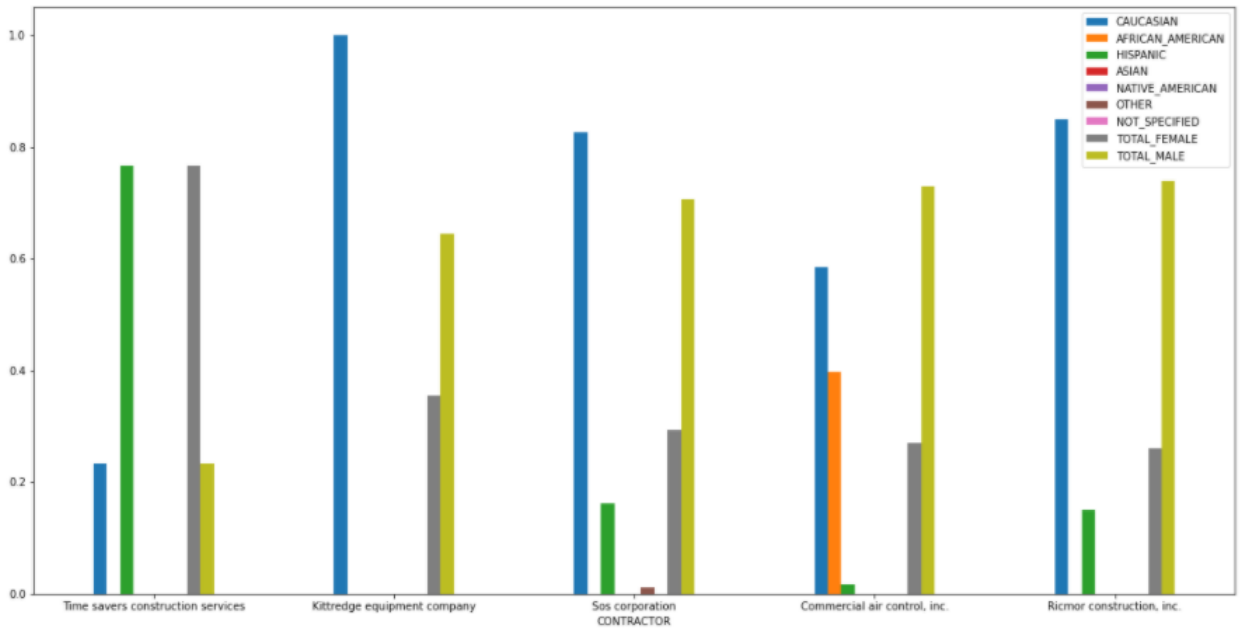
Top 5 companies based on percentage of Native Americans (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



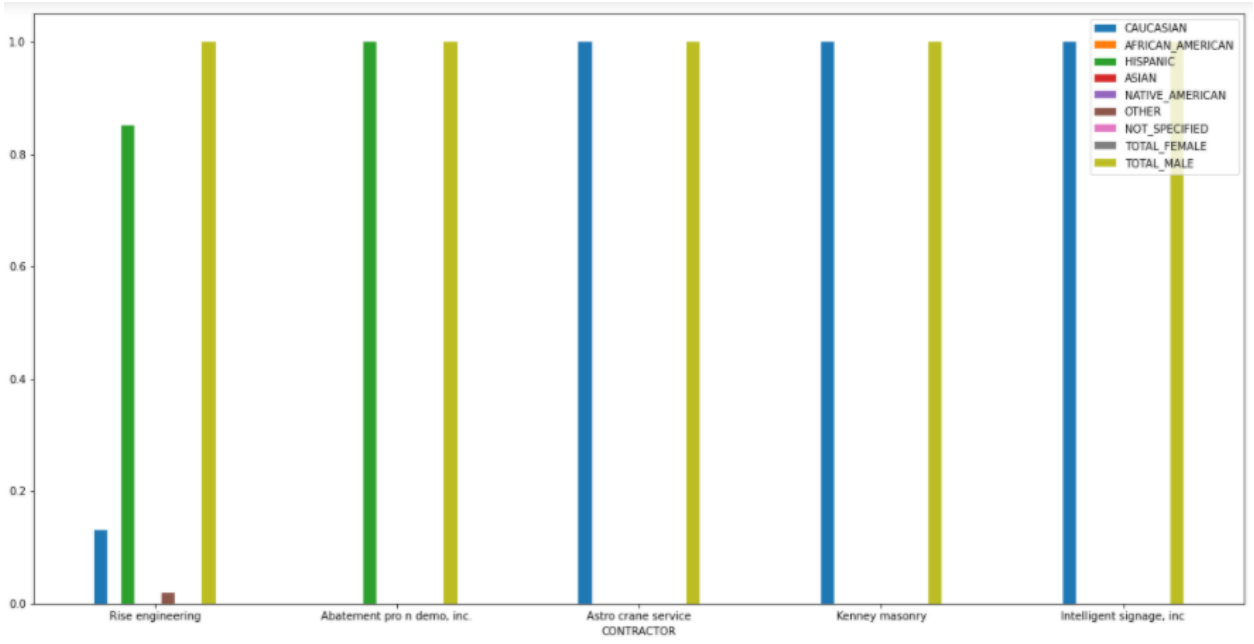
Top 5 companies based on percentage of Other (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



Top 5 companies based on percentage of Not Specified (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)

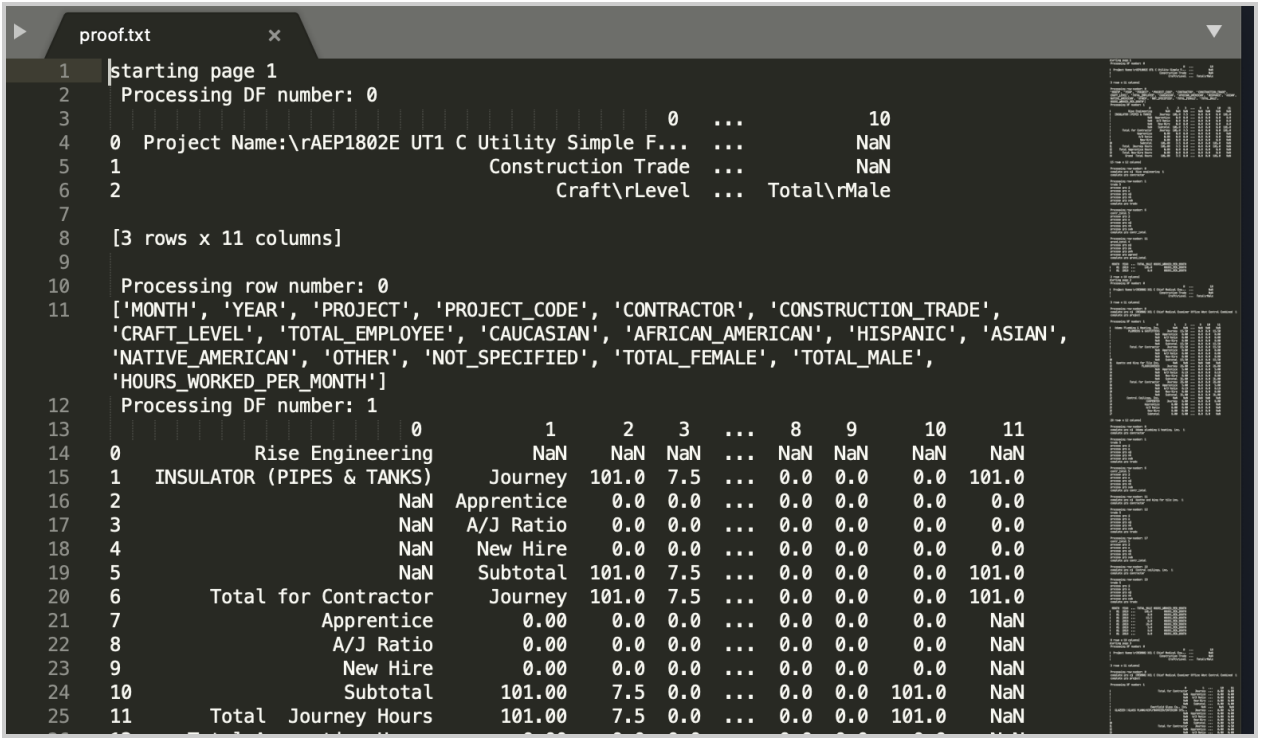


Top 5 companies based on percentage of Females (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)



Top 5 companies based on percentage of Males (Percentage aggregate of total hours worked per month for year for that ethnicity/gender divided by total hours worked for all ethnicities/genders)

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
```

```
115269
115270   Processing row number: 0
115271   complete grp cj  WSC1902 DB1 C WSC-Campus-wide-Redundant Steam and Condensate Loop  1
115272   complete grp project
115273
115274   Processing DF number: 1
115275                     0         1      2    3   ...   7    8      9      10
115276   0    Total  Journey Hours  1,056.00  904.5  0.0  ...  0.0  0.0  143.5  912.5
115277   1  Total Apprentice Hours      7.00    7.0  0.0  ...  0.0  0.0    0.0    7.0
115278   2    Total New Hire Hours      0.00    0.0  0.0  ...  0.0  0.0    0.0    0.0
115279   3      Grand  Total Hours  1,063.00  911.5  0.0  ...  0.0  0.0  143.5  919.5
115280
115281   [4 rows x 11 columns]
115282
115283   Processing row number: 0
115284   grand_total 4
115285   process grp pj
115286   process grp pa
115287   process grp pnh
115288   process grp pgrand
115289   complete grp grand_total
115290
115291     MONTH  YEAR  ... TOTAL_MALE HOURS_WORKED_PER_MONTH
115292   0     12  2019  ...       16.0        HOURS_PER_MONTH
115293   0     12  2019  ...        0.0        HOURS_PER_MONTH
115294   0     12  2019  ...        0.0        HOURS_PER_MONTH
115295   0     12  2019  ...      165.0        HOURS_PER_MONTH
115296   0     12  2019  ...     1632.0        HOURS_PER_MONTH
115297   ..   ...   ...  ...        ...                    ...
115298   0     12  2019  ...        7.0        HOURS_PER_MONTH
115299   0     12  2019  ...      393.0        HOURS_PER_MONTH
115300   0     12  2019  ...        0.0        HOURS_PER_MONTH
115301   0     12  2019  ...      512.5        HOURS_PER_MONTH
115302   0     12  2019  ...        0.0        HOURS_PER_MONTH
115303
115304   [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

	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...
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...