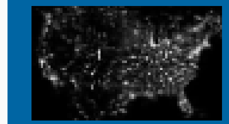
MIDTERM SKILLS EXAM
DATA WRANGLING AND ANALYSIS

De Guzman, Jemuel Endrew C.
CPE22S3

INTRODUCTION TO THE DATASET



Census Income

Donated on 4/30/1996

Predict whether income exceeds $50K/yr based on census data. Also known as Adult dataset.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	48842	14

Dataset Information

Additional Information

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

SHOW LESS

Has Missing Values?

Yes

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  int64
 1   workclass       47879 non-null  object
 2   fnlwgt          48842 non-null  int64
 3   education       48842 non-null  object
 4   education-num   48842 non-null  int64
 5   marital-status  48842 non-null  object
 6   occupation      47876 non-null  object
 7   relationship    48842 non-null  object
 8   race            48842 non-null  object
 9   sex             48842 non-null  object
 10  capital-gain    48842 non-null  int64
 11  capital-loss    48842 non-null  int64
 12  hours-per-week  48842 non-null  int64
 13  native-country  48568 non-null  object
 14  income          48842 non-null  object
dtypes: int64(6), object(9)
```

DATA WRANGLING

```python
# Replacing all NaN and ? values with 'Other'
cols = ['workclass','occupation','native-country']

for col in cols:
  xy[col] = xy[col].fillna('Other')
  xy[col].replace('?','Other',inplace=True)

xy[xy.values == 'Other']

#print(xy[xy.isna().any(axis=1)]) # Shows that there are no more NaN values
#print('\n')
#print(xy.info())
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
14	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	Other	>50K
27	54	Other	180211	Some-college	10	Married-civ-spouse	Other	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
27	54	Other	180211	Some-college	10	Married-civ-spouse	Other	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
38	31	Private	84154	Some-college	10	Married-civ-spouse	Sales	Husband	White	Male	0	0	38	Other	>50K
50	25	Private	32275	Some-college	10	Married-civ-spouse	Exec-managerial	Wife	Other	Female	0	0	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
48812	81	Other	26711	Assoc-voc	11	Married-civ-spouse	Other	Husband	White	Male	2936	0	20	United-States	<=50K.
48812	81	Other	26711	Assoc-voc	11	Married-civ-spouse	Other	Husband	White	Male	2936	0	20	United-States	<=50K.
48826	50	Local-gov	139347	Masters	14	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	40	Other	>50K.
48838	64	Other	321403	HS-grad	9	Widowed	Other	Other-relative	Black	Male	0	0	40	United-States	<=50K.
48838	64	Other	321403	HS-grad	9	Widowed	Other	Other-relative	Black	Male	0	0	40	United-States	<=50K.

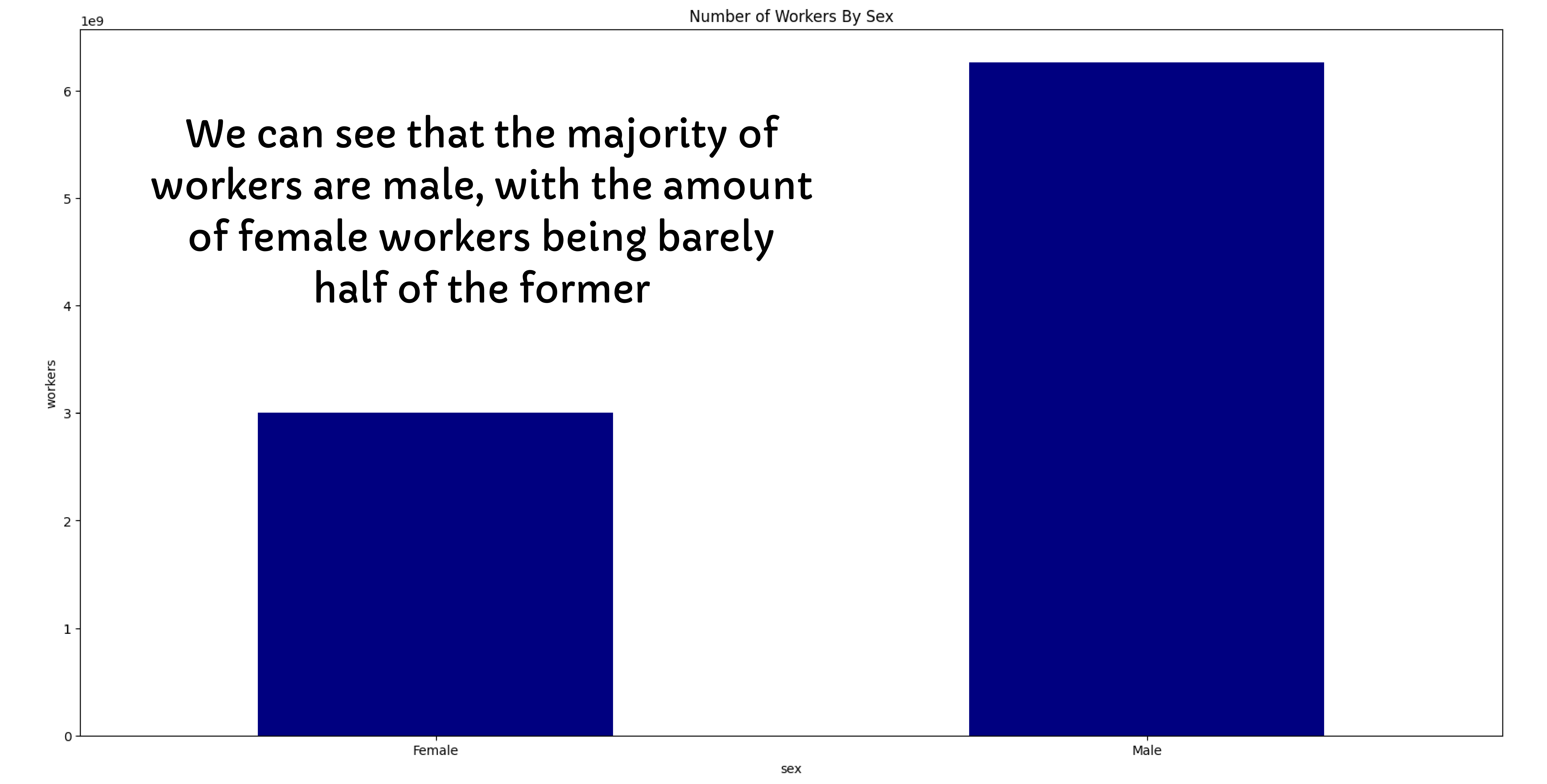
6871 rows × 15 columns

```python
# renaming columns for clarity

xy.rename(columns={'fnlwgt':'record_count',
                   'education-num':'education_num',
                   'marital-status':'marital_status',
                   'capital-gain':'capital_gain',
                   'capital-loss':'capital_loss',
                   'hours-per-week':'hours_per_week',
                   'native-country':'native_country'}, inplace=True)

xy
```

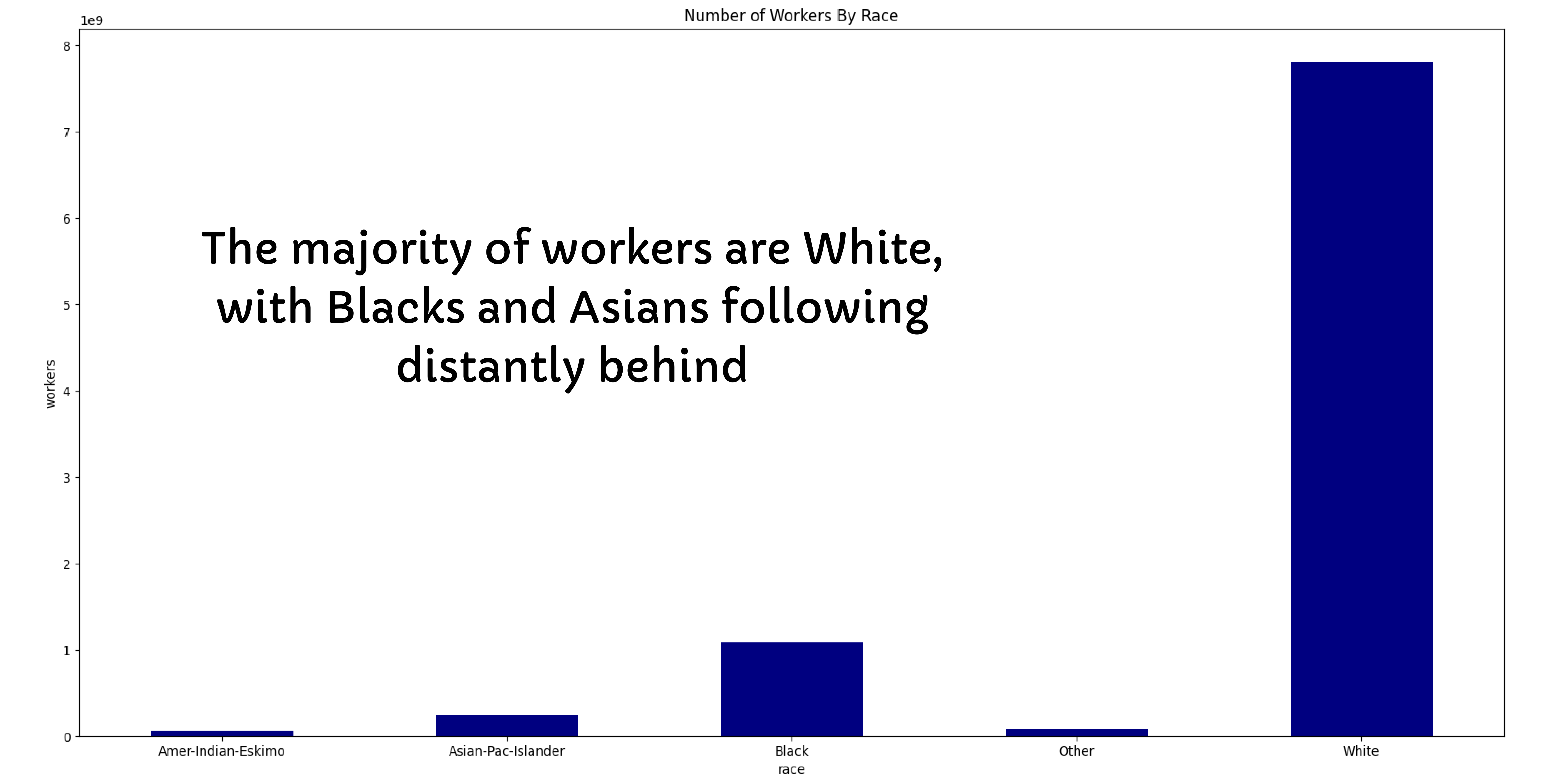
	age	workclass	record_count	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
48837	39	Private	215419	Bachelors	13	Divorced	Prof-specialty	Not-in-family	White	Female	0	0	36	United-States	<=50K.
48838	64	Other	321403	HS-grad	9	Widowed	Other	Other-relative	Black	Male	0	0	40	United-States	<=50K.
48839	38	Private	374983	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	50	United-States	<=50K.
48840	44	Private	83891	Bachelors	13	Divorced	Adm-clerical	Own-child	Asian-Pac-Islander	Male	5455	0	40	United-States	<=50K.
48841	35	Self-emp-inc	182148	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	60	United-States	>50K.

48842 rows × 15 columns

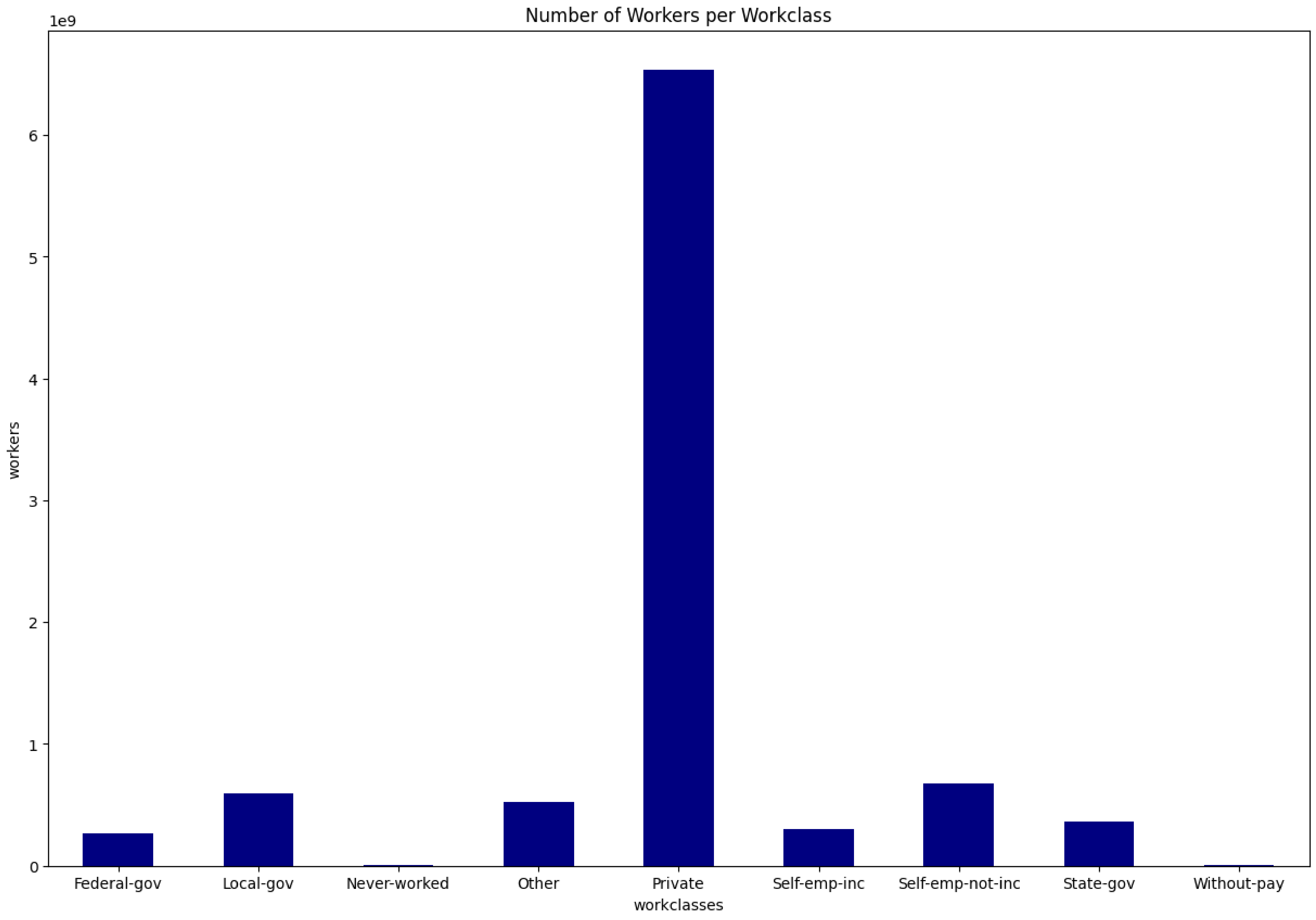
EXPLORATORY DATA ANALYSIS VISUALIZATIONS



We can see that the majority of workers are male, with the amount of female workers being barely half of the former

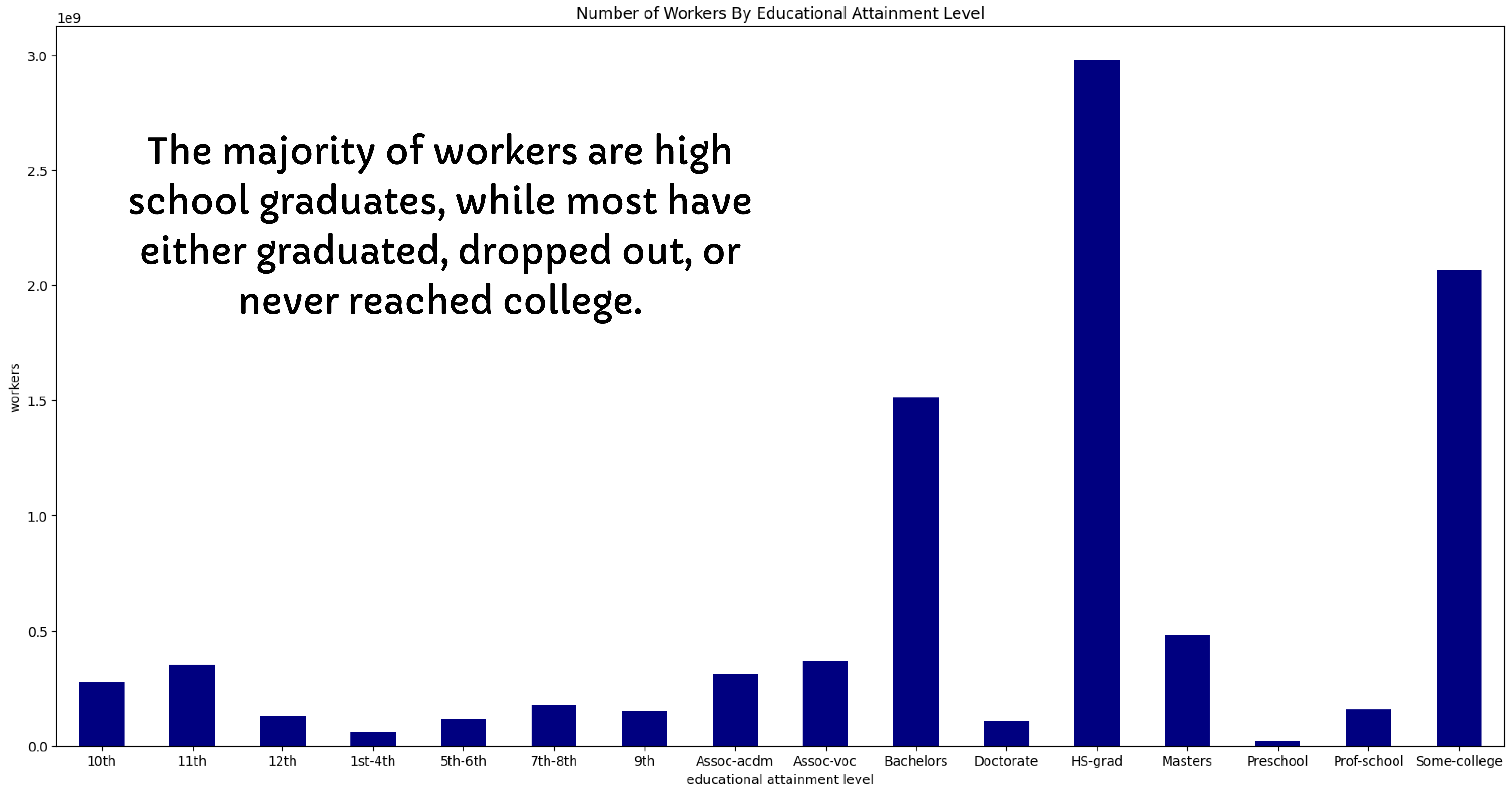


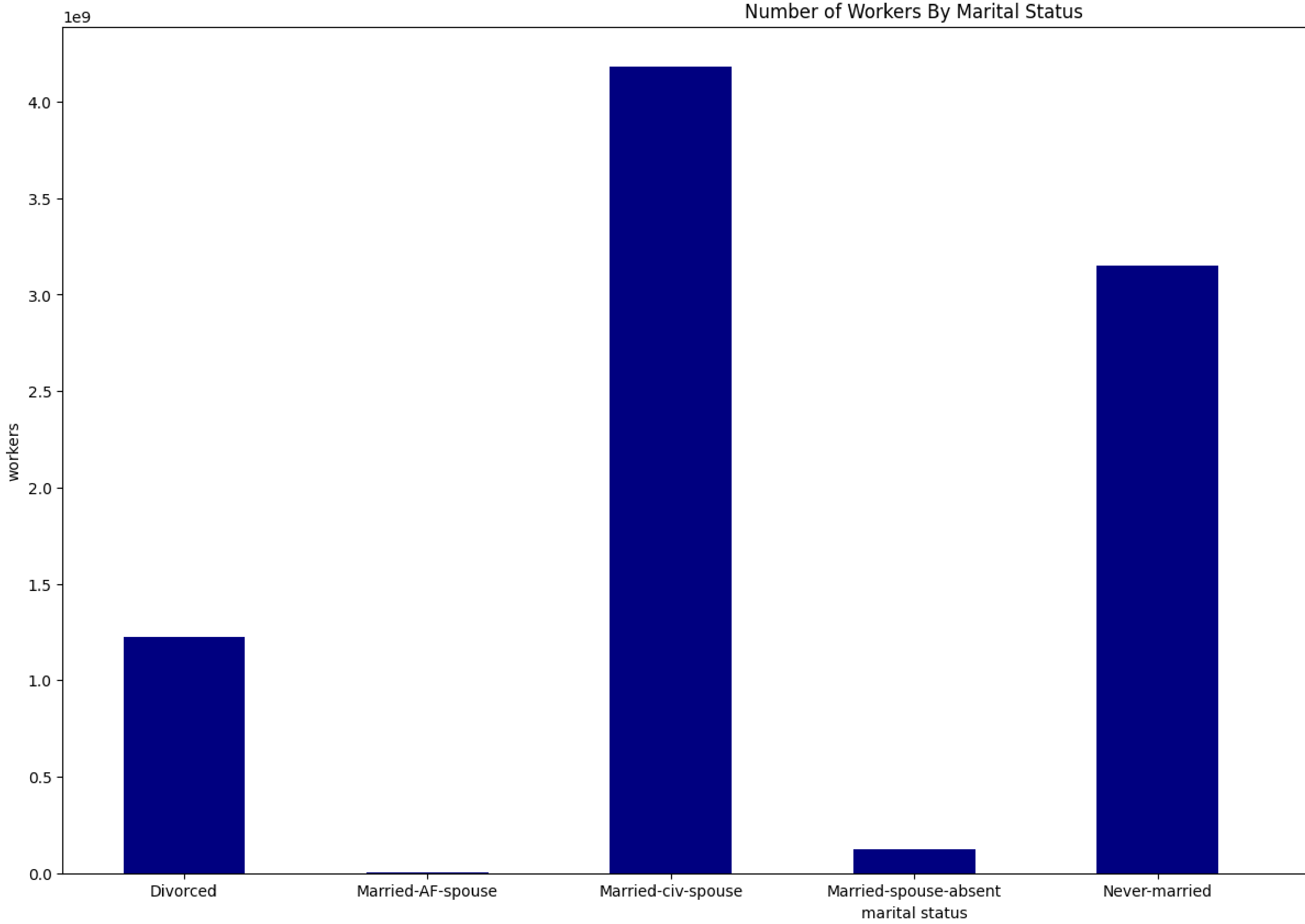
The majority of workers are White, with Blacks and Asians following distantly behind



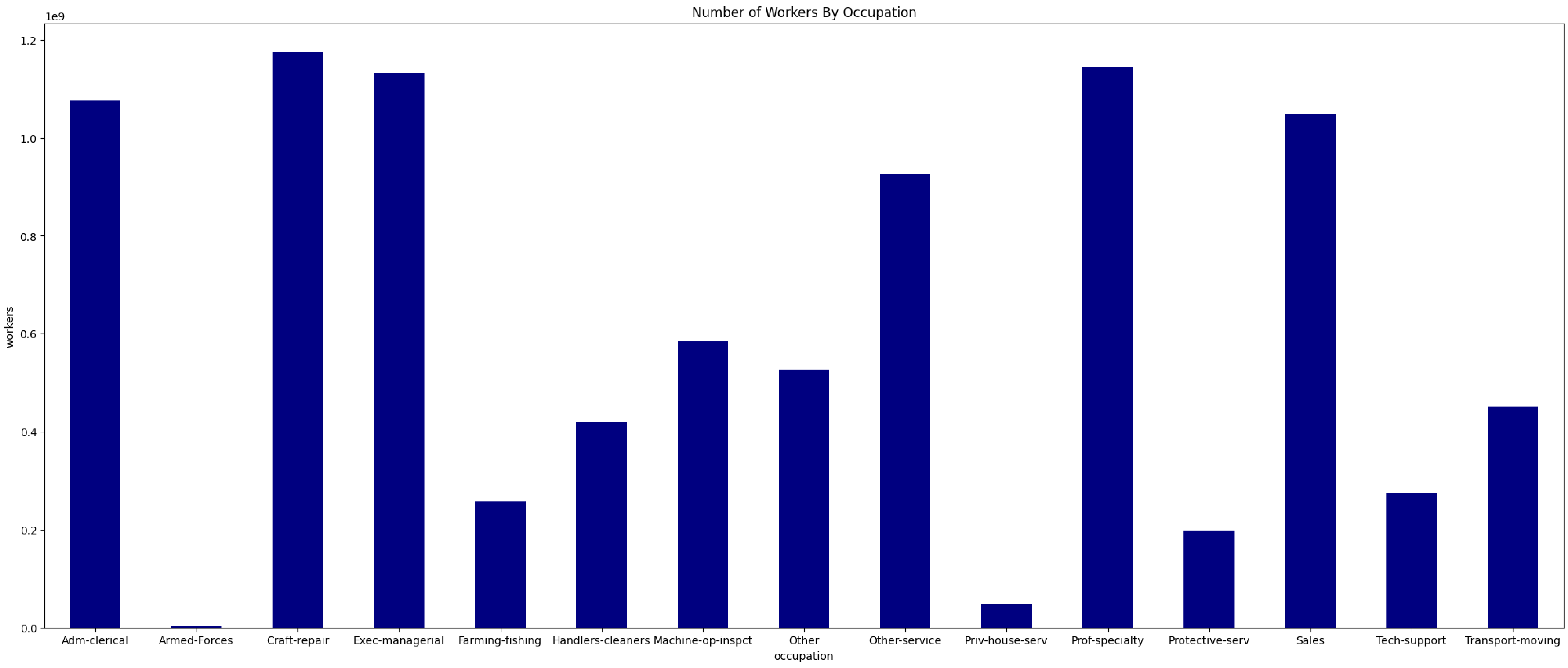
The majority of workers come from the Private sector

The majority of workers are high school graduates, while most have either graduated, dropped out, or never reached college.



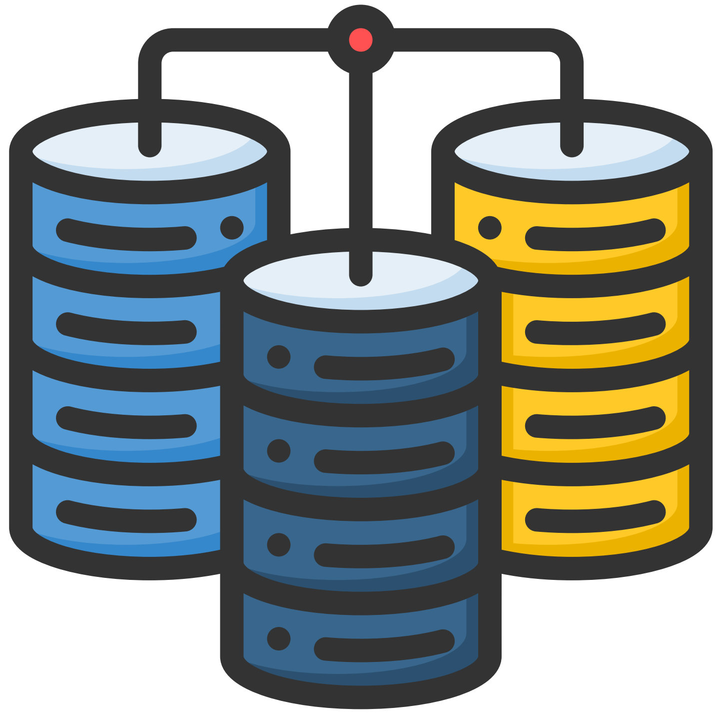


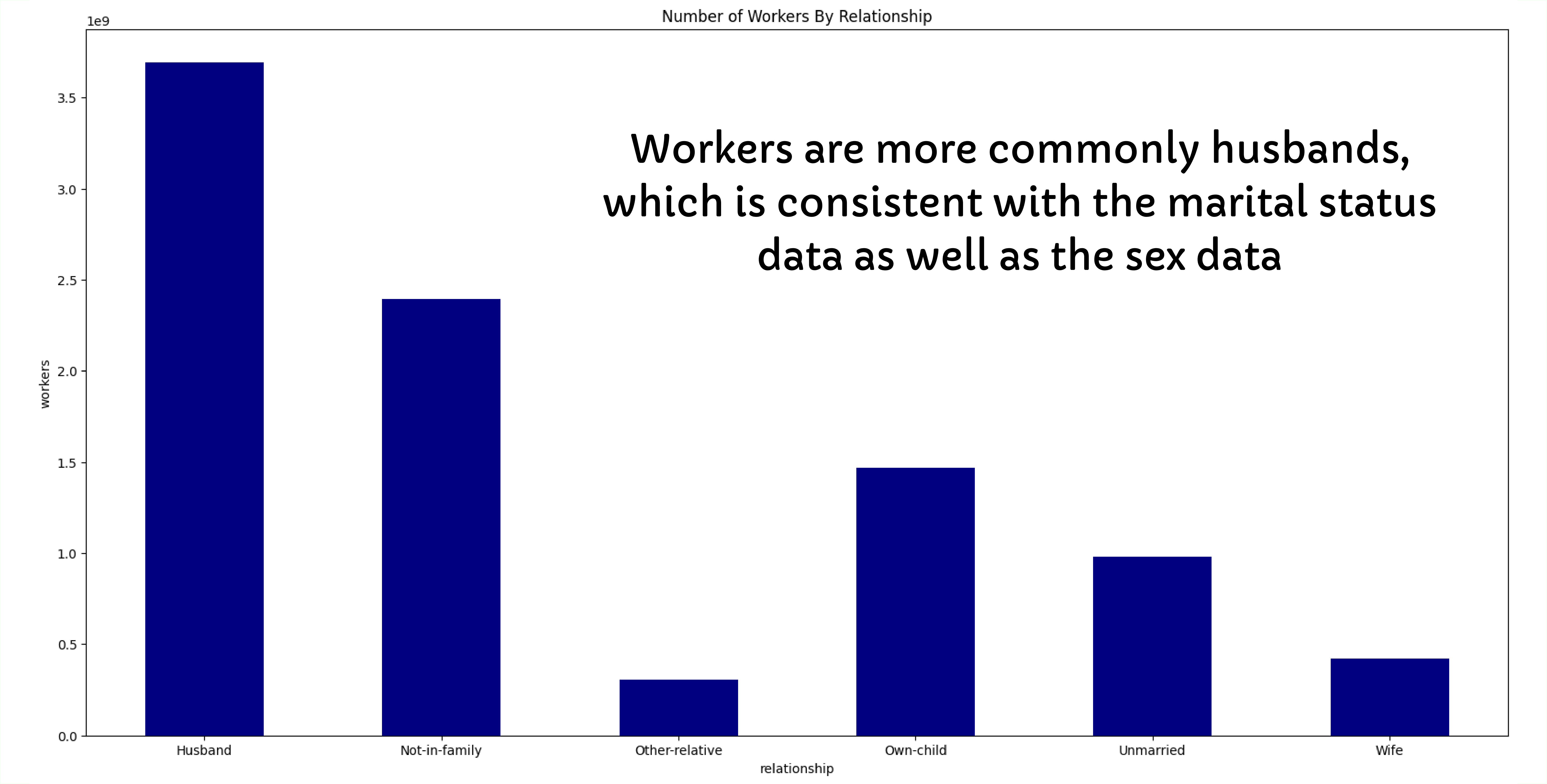
Most workers are either married or have never been married, while some are from dysfunctional families

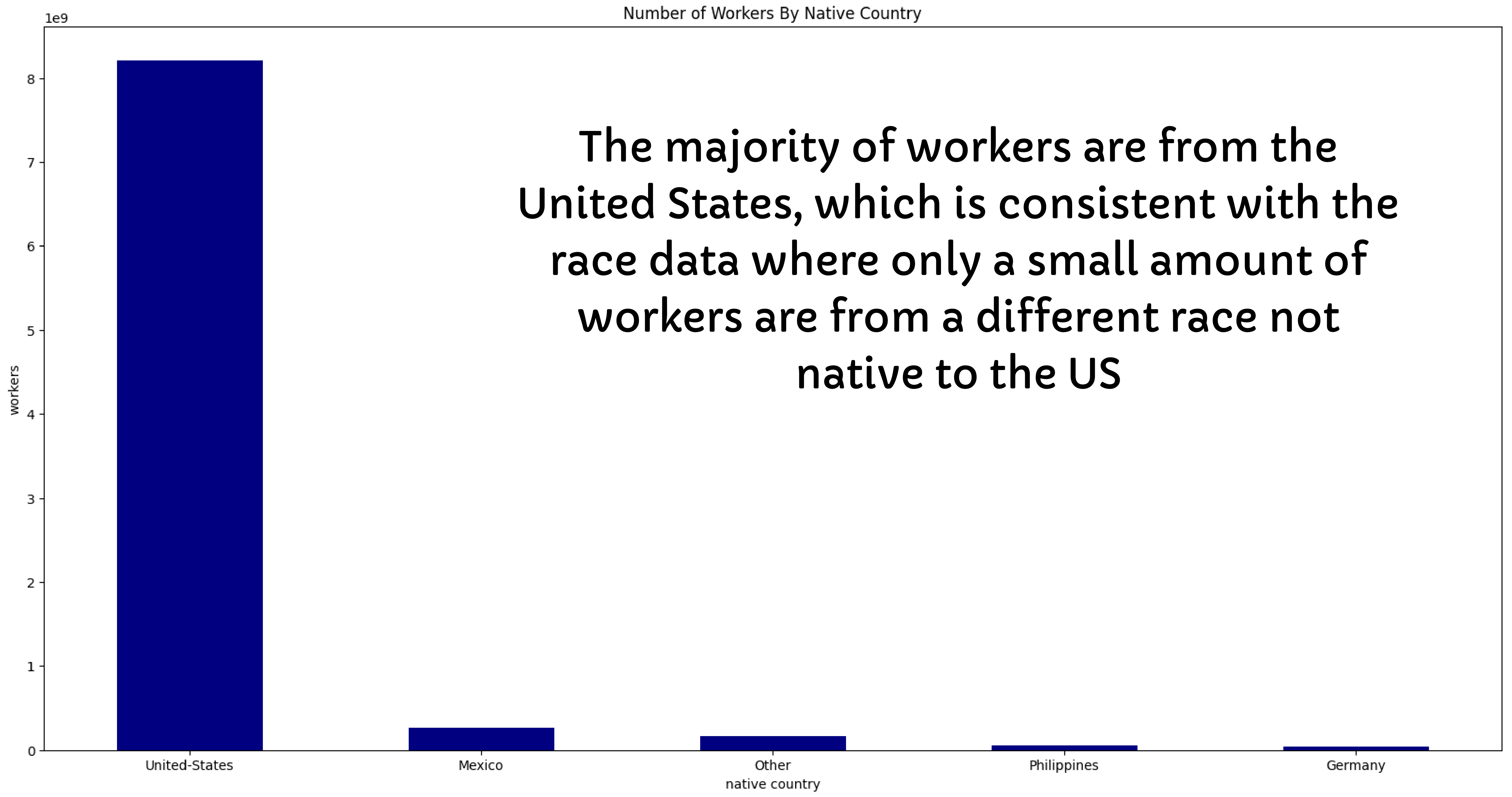


Number of Workers By Occupation
1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0
workers
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
occupation

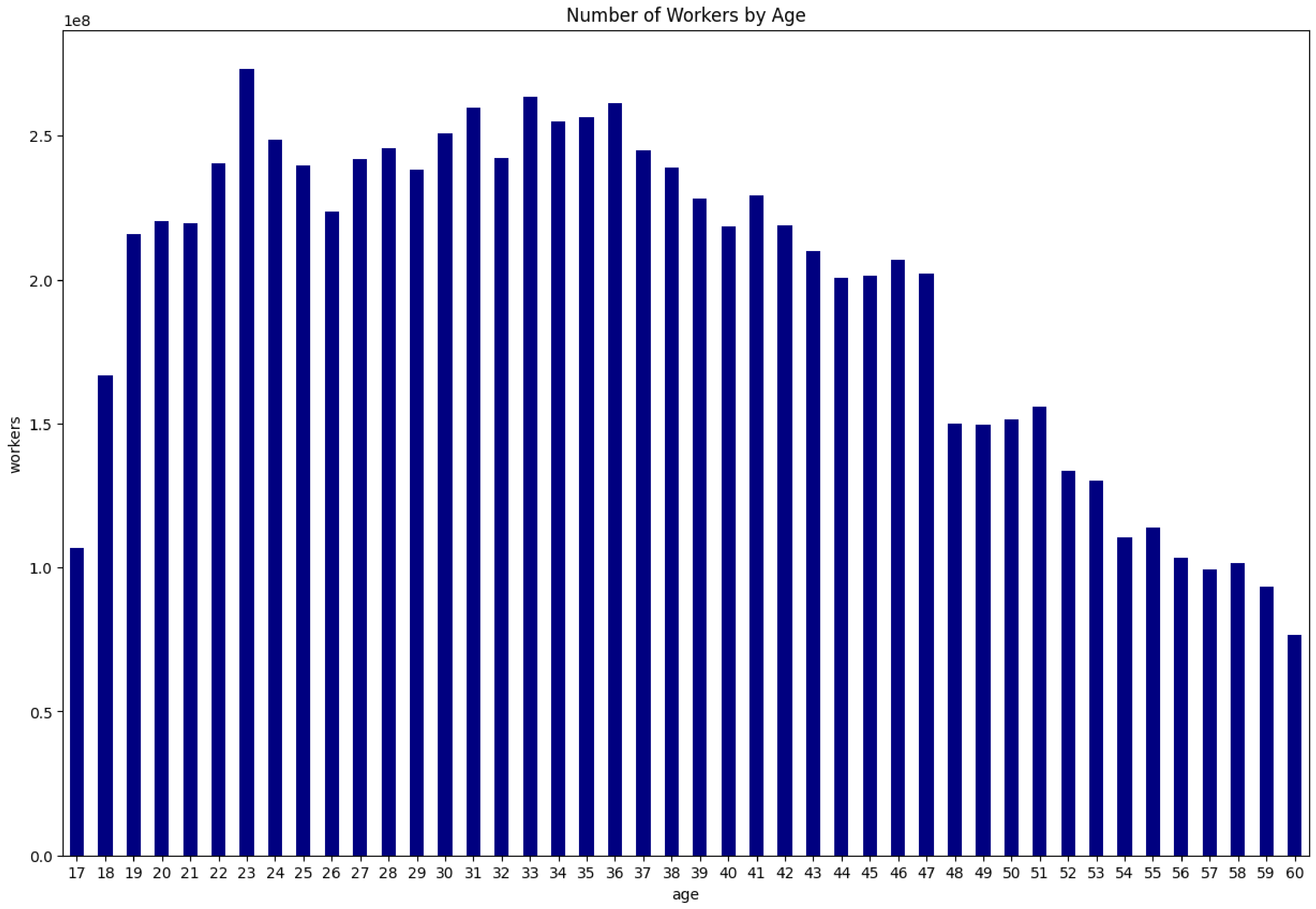
It can be said from the data that the distribution between white-collar and blue-collar workers is fairly even. While it does seem like the white-collar jobs are of higher count based on the graph, the same can be said for the blue-collar jobs, but they are more distributed into more columns while the latter are condensed into lesser columns.



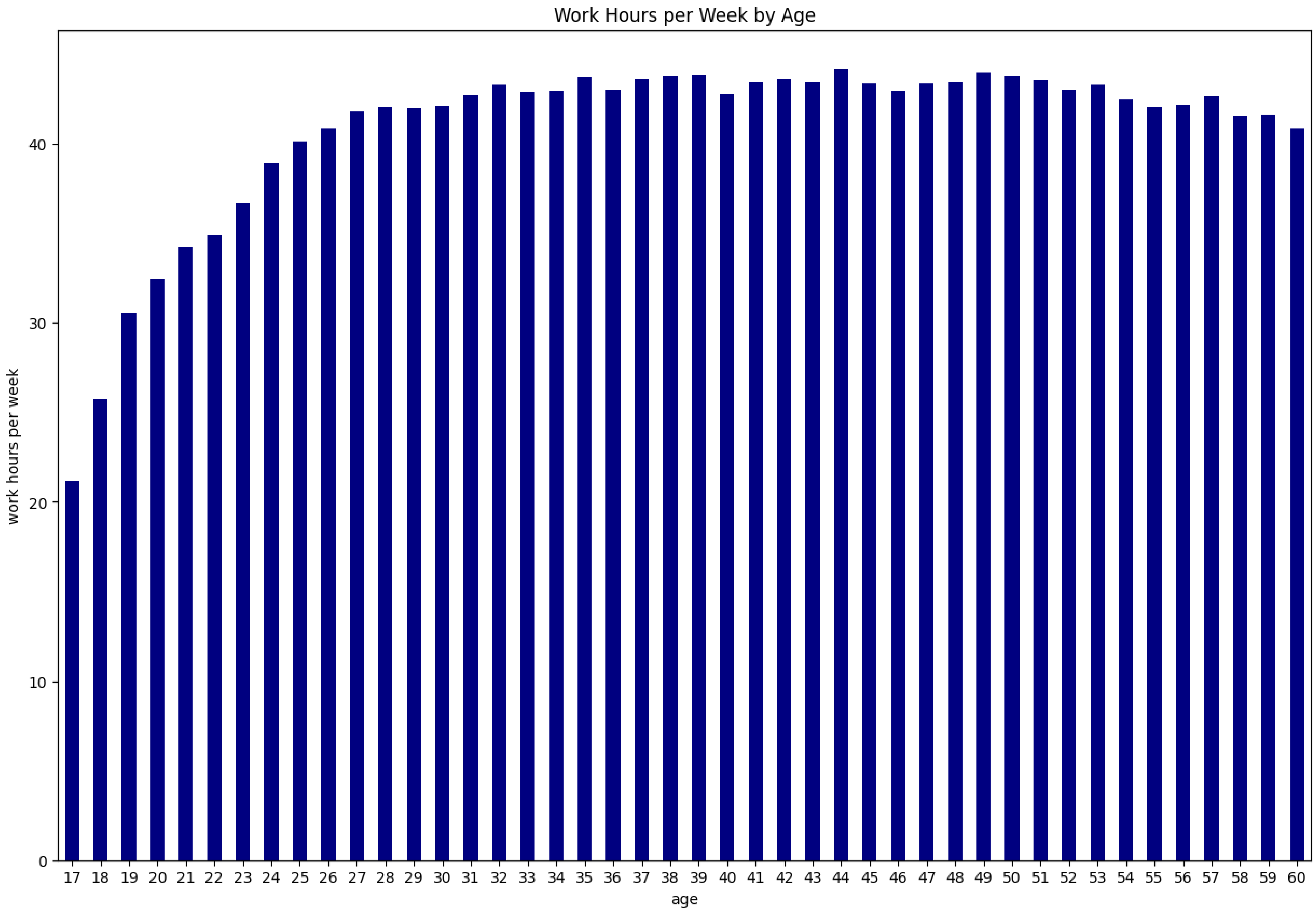

Number of Workers By Relationship
1e9
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
workers
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife
relationship
Workers are more commonly husbands, which is consistent with the marital status data as well as the sex data



The majority of workers are from the United States, which is consistent with the race data where only a small amount of workers are from a different race not native to the US

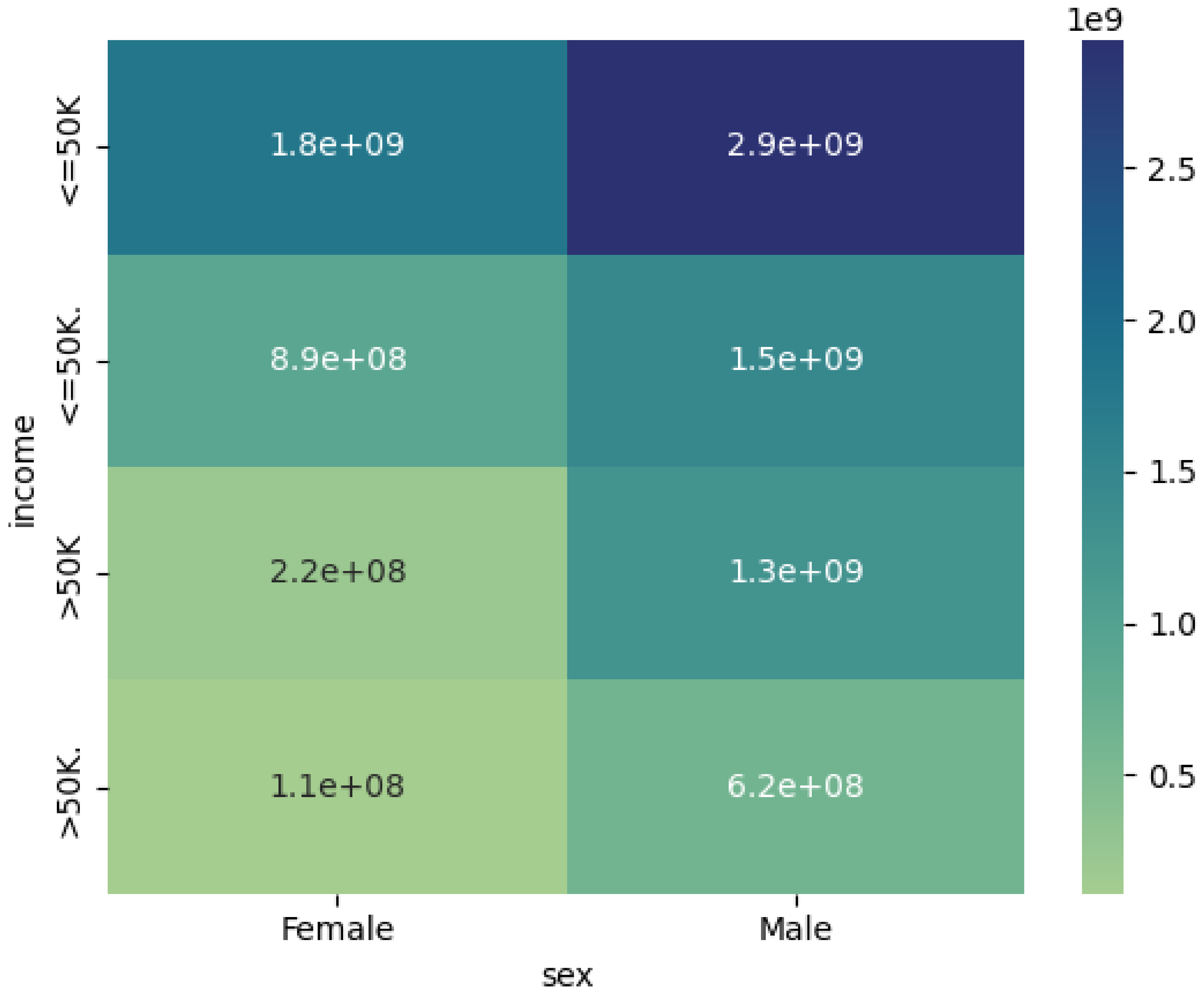


The small number of workers from the minority age group as well as from the 40s and above age group shows how middle-aged workers are more favored over the previous 2 age groups.



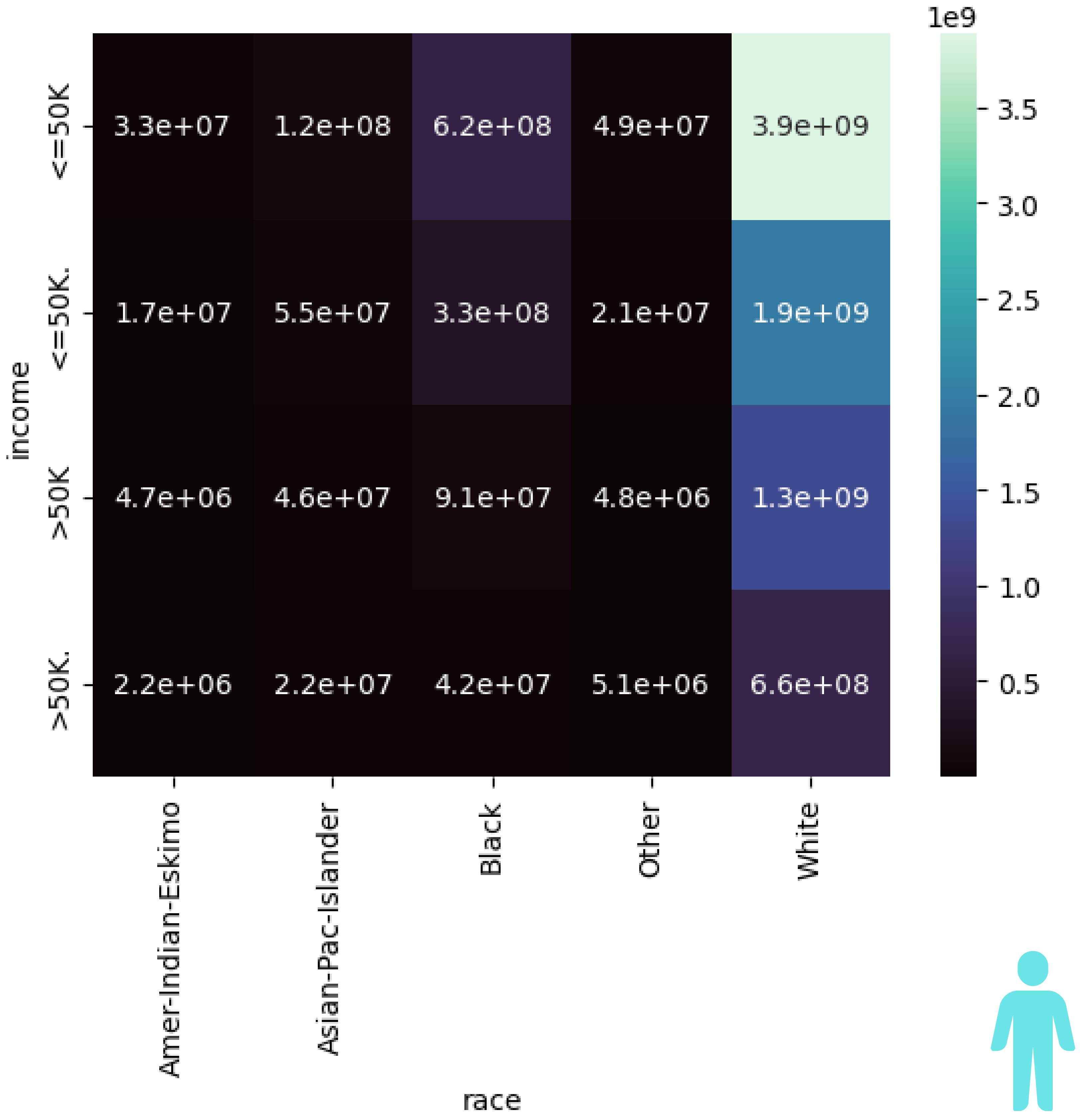
However, when the work hours are compared, only the minority have lower work hours. It can be inferred that these workers are probably interns or part-time workers, explaining the shorter work hours

CORRELATION ANALYSIS



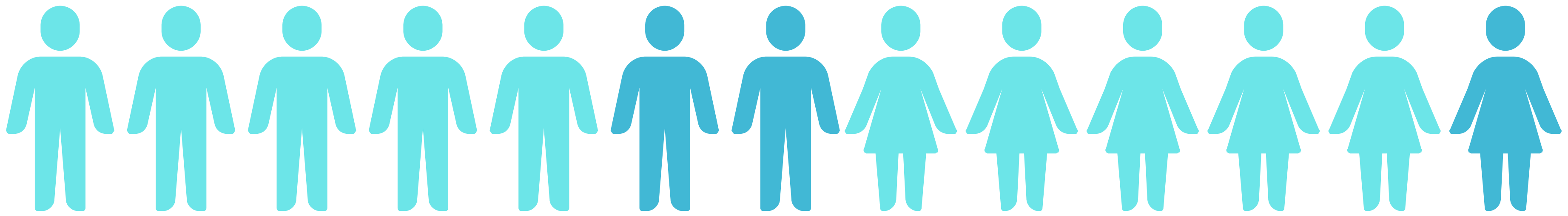
Sex vs Income

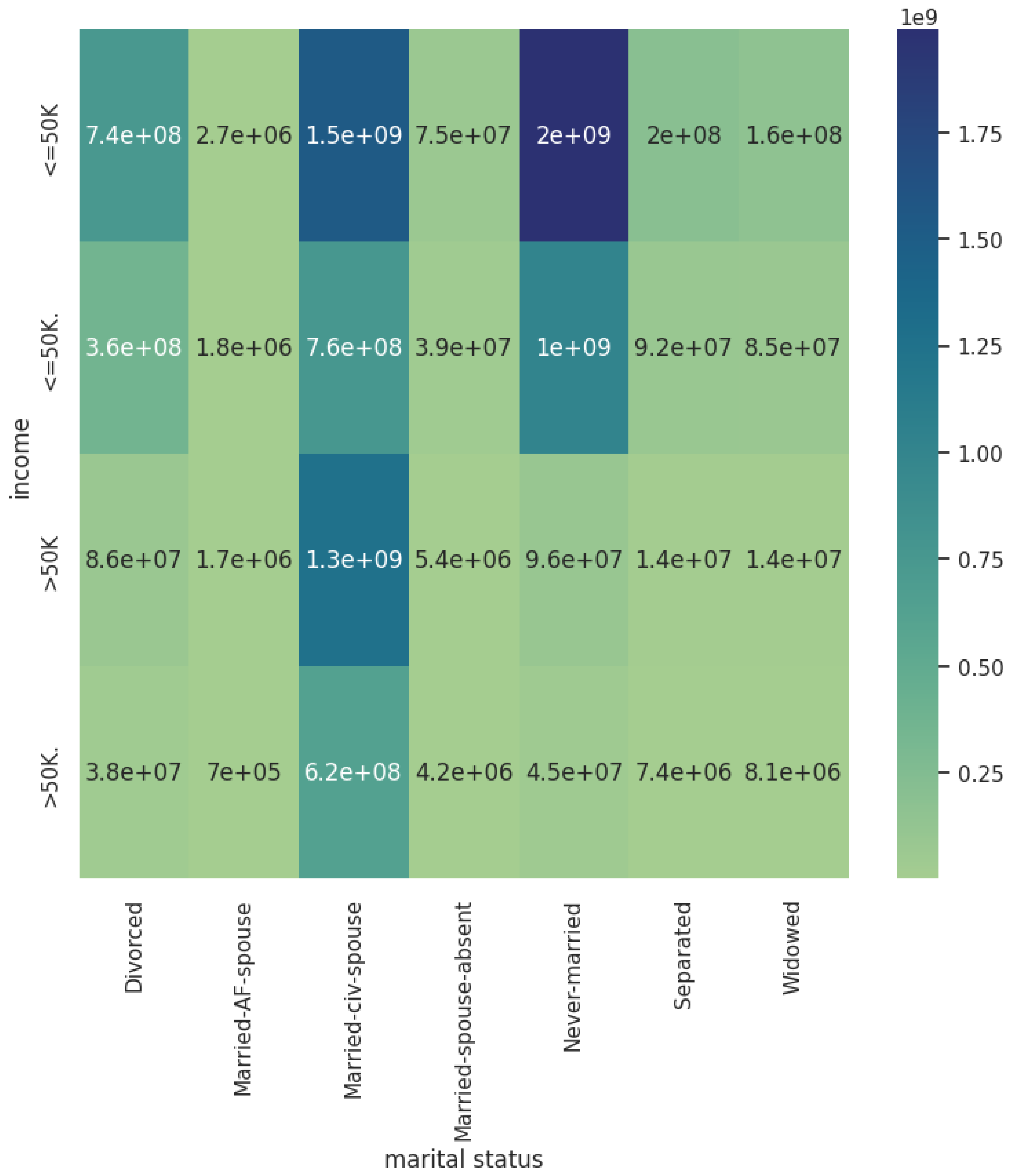
When comparing the income ranges for each gender, we can see that the majority of males earn less than 50K, as well as for females but at a lesser volume. This is consistent with the sex data shown previously



Race vs Income

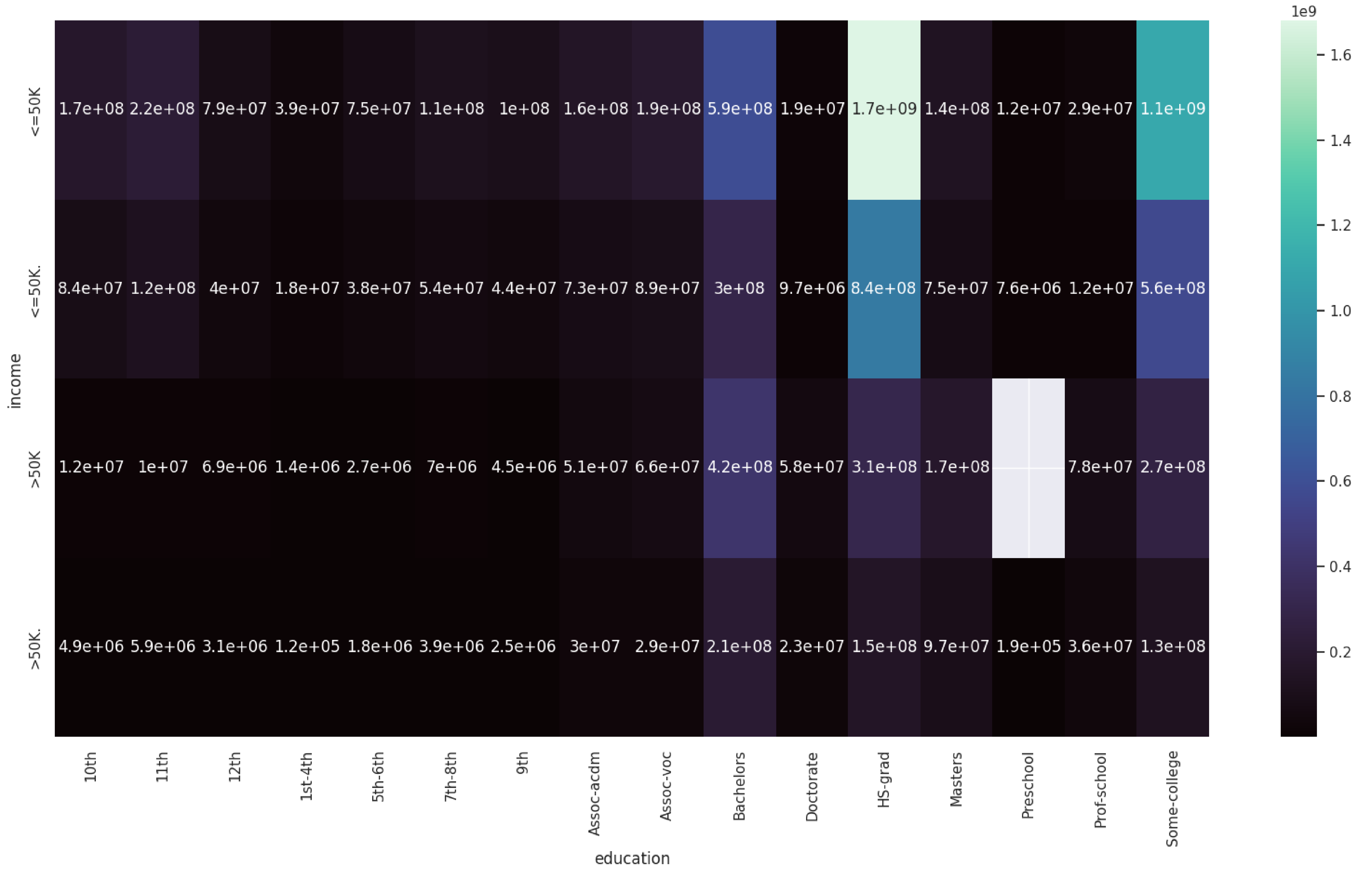
It can be seen that a majority of workers who are white earn less than 50K, as well as more than 50K. This is because the majority of workers are White in comparison to other races





Marital Status vs Income

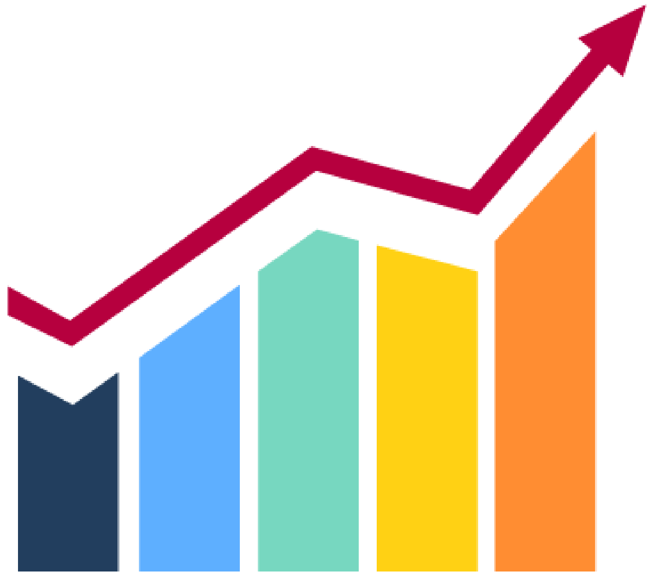
The majority of workers who are either never married or have married a civilian earn more or less than 50K. It is more distributed for married civilians in comparison to the ones who have never married.



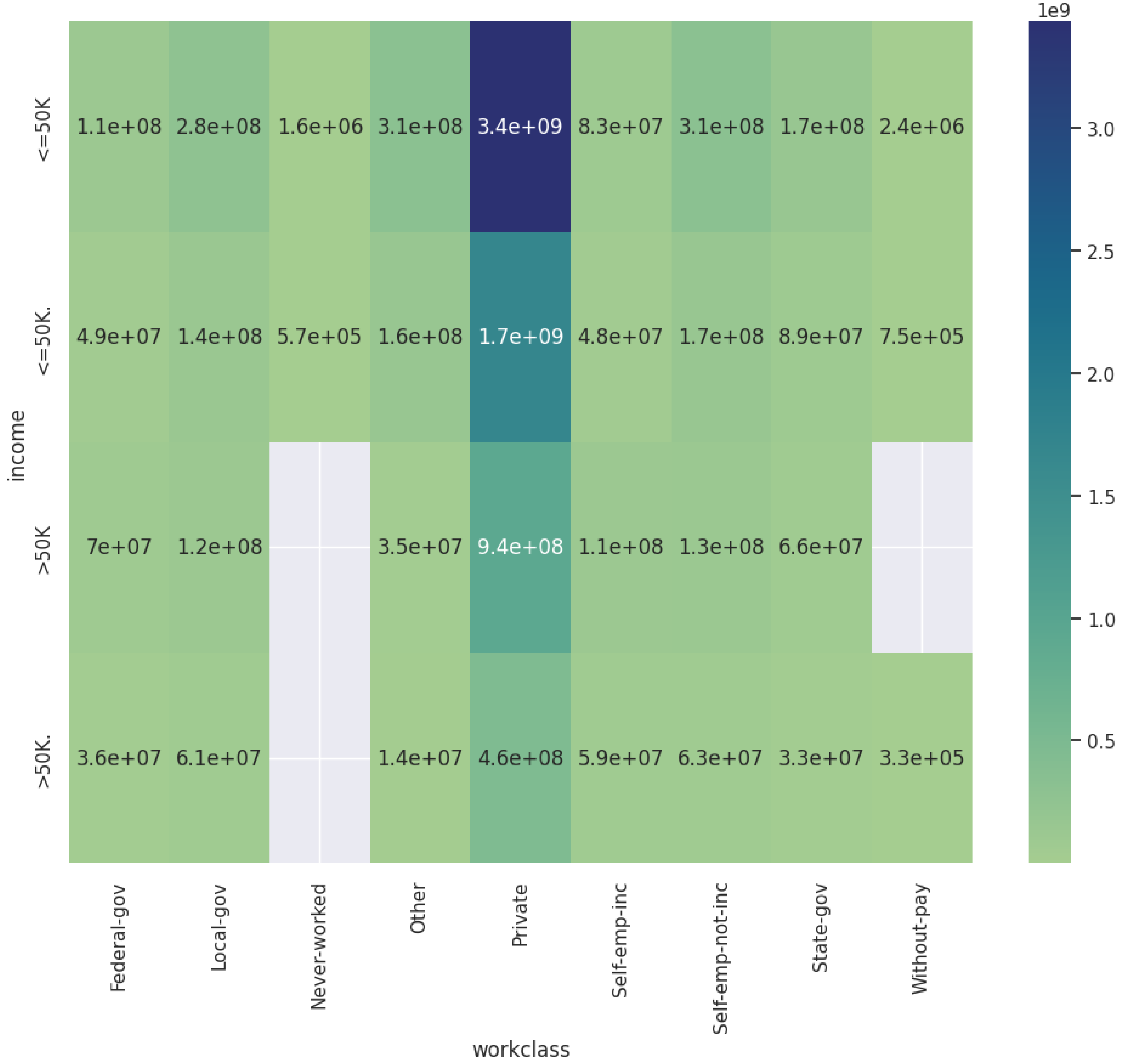
income
education
<=50K
<=50K.
>50K
>50K.
10th
11th
12th
1st-4th
5th-6th
7th-8th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Preschool
Prof-school
Some-college
1.7e+08 2.2e+08 7.9e+07 3.9e+07 7.5e+07 1.1e+08 1e+08 1.6e+08 1.9e+08 5.9e+08 1.9e+07 1.7e+09 1.4e+08 1.2e+07 2.9e+07 1.1e+09
8.4e+07 1.2e+08 4e+07 1.8e+07 3.8e+07 5.4e+07 4.4e+07 7.3e+07 8.9e+07 3e+08 9.7e+06 8.4e+08 7.5e+07 7.6e+06 1.2e+07 5.6e+08
1.2e+07 1e+07 6.9e+06 1.4e+06 2.7e+06 7e+06 4.5e+06 5.1e+07 6.6e+07 4.2e+08 5.8e+07 3.1e+08 1.7e+08 7.8e+07 2.7e+08
4.9e+06 5.9e+06 3.1e+06 1.2e+05 1.8e+06 3.9e+06 2.5e+06 3e+07 2.9e+07 2.1e+08 2.3e+07 1.5e+08 9.7e+07 1.9e+05 3.6e+07 1.3e+08
1e9
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2

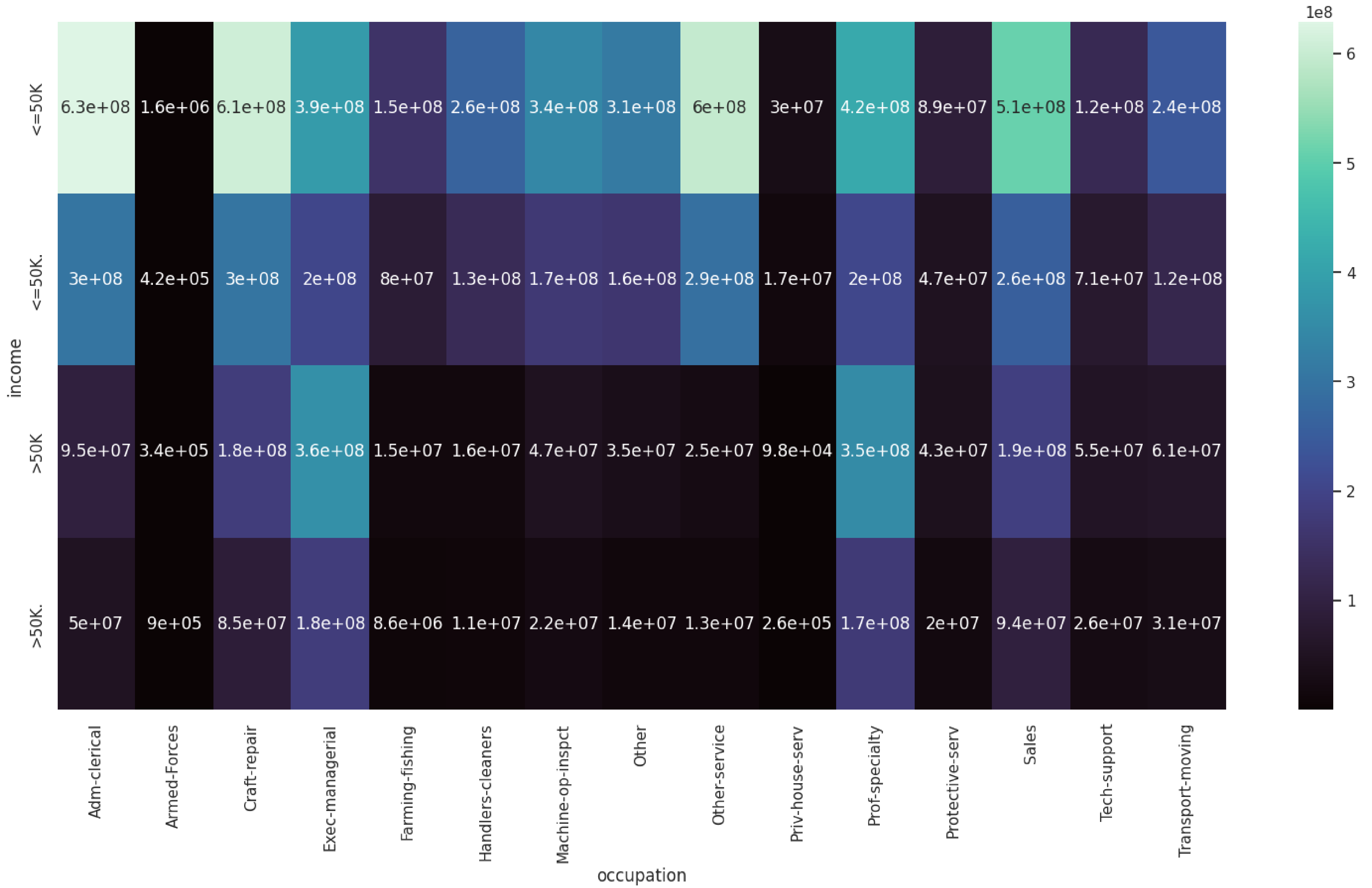
Educational Attainment vs Income

It can be said that a majority of those who are earning less than 50K are those who have dropped out of college or are high-school graduates. This is also consistent with the education data, which shows that these 2 are the majority of workers in the US.



Workclass vs Income




1e8
6
5
4
3
2
1
income
<=50K
<=50K.
>50K
>50K.
6.3e+08 1.6e+06 6.1e+08 3.9e+08 1.5e+08 2.6e+08 3.4e+08 3.1e+08 6e+08 3e+07 4.2e+08 8.9e+07 5.1e+08 1.2e+08 2.4e+08
3e+08 4.2e+05 3e+08 2e+08 8e+07 1.3e+08 1.7e+08 1.6e+08 2.9e+08 1.7e+07 2e+08 4.7e+07 2.6e+08 7.1e+07 1.2e+08
9.5e+07 3.4e+05 1.8e+08 3.6e+08 1.5e+07 1.6e+07 4.7e+07 3.5e+07 2.5e+07 9.8e+04 3.5e+08 4.3e+07 1.9e+08 5.5e+07 6.1e+07
5e+07 9e+05 8.5e+07 1.8e+08 8.6e+06 1.1e+07 2.2e+07 1.4e+07 1.3e+07 2.6e+05 1.7e+08 2e+07 9.4e+07 2.6e+07 3.1e+07
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
occupation

Conclusion

From the dataset, I was able to clean up some missing values as well as rearrange the dataset to make it easier to analyze. As I was analyzing the dataset, I was able to determine that the record_count column (previously the fnlwgt column) was going to be the most important data in the dataset. This is because it is the most consistent quantity present in the dataset. Most columns in the dataset, as well as the income column, are mostly qualitative, which doesn't give much space for any computations and accurate analysis. With the record_count column, I was able to visualize and compare the amount of people represented by the records per specific demographic, as well as correlate the income category in specific columns through the use of the record_count column. I think that being able to analyze and identify how to work with a dataset is key to being able to interpret it. While I was only able to visualize multiple relationships in the dataset, I think that I have done enough of it that some form of interpretation and conclusion can be drawn from some of them combined.

THANK YOU FOR LISTENING!