DSE_200x Income Indicators

Ryan M Lence

Abstract

I found a dataset on Kaggle called Income classification that outlines a list of people who make more than and less than $50,000 per year.

I focused the dataset on just the USA, and the following list of attributes: (workclass, education, marital-status, occupation, face, sex, hours-per-week)

My goal was to find which of these attributes was the single biggest indicator if an individuate will make at least $50,000 per year.

My findings will show that occupation is the single biggest indicator.

Motivation

Education is always pushed as a must for a person who wants to have a decent income above the poverty line. While I believe that education is a big part of it, I also wanted to show that there are other factors in-play that are even more important.

For example we all know people who have a Masters degree and struggle to make a decent living. Also we all know people who do not have any collage experience and make well above $50,000 per year.

With this dataset I wanted to find if education is the main factor in-common or is it something else.

Dataset(s)

This dataset includes a listing of attributes which may influence if a person makes an income of more or less than $50,000 per year.

The dataset is smaller in size, 32561 rows and 15 columns.

I found the data on Kaggle at the following location:

https://www.kaggle.com/lodetomasi1995/income-classification#income_evaluation.csv

Data Preparation and Cleaning

At a high-level I found a few issues with the dataset. Below is a list of issues that I had to address.

- Many of the columns had a leading space in the name make it hard to use df.column notation, so I removed the space. I also had to remove rows from workclass and education that only listed a ‘ ?’.
- Native-country is listed, but 98% of entries are all from USA, so I deleted any non-USA entries not from USA.
- Very few entries have any capital gains or losses so I removed those columns and rows so it would not influence the analysis.
- Workclass is missing for about 10% of entries I removed those rows as I am working to learn the importins of workclass.
- When I finished the clean-up of the data is was 23816 rows and 9 columns

Research Question(s)

- I would like to determine what are the three biggest factors that determines if someone makes more than $50,000 per year from the dataset.
- Which of the following indicators are the greatest prediction if a person is making more than $50,000 per year?
- age, workclass, education, married, occupation, race,
  sex, hours

Methods

After cleaning the data into groups, I created several bar charts to compare the data for each of the indicators. I also did counts and calculations looking at the different methods.

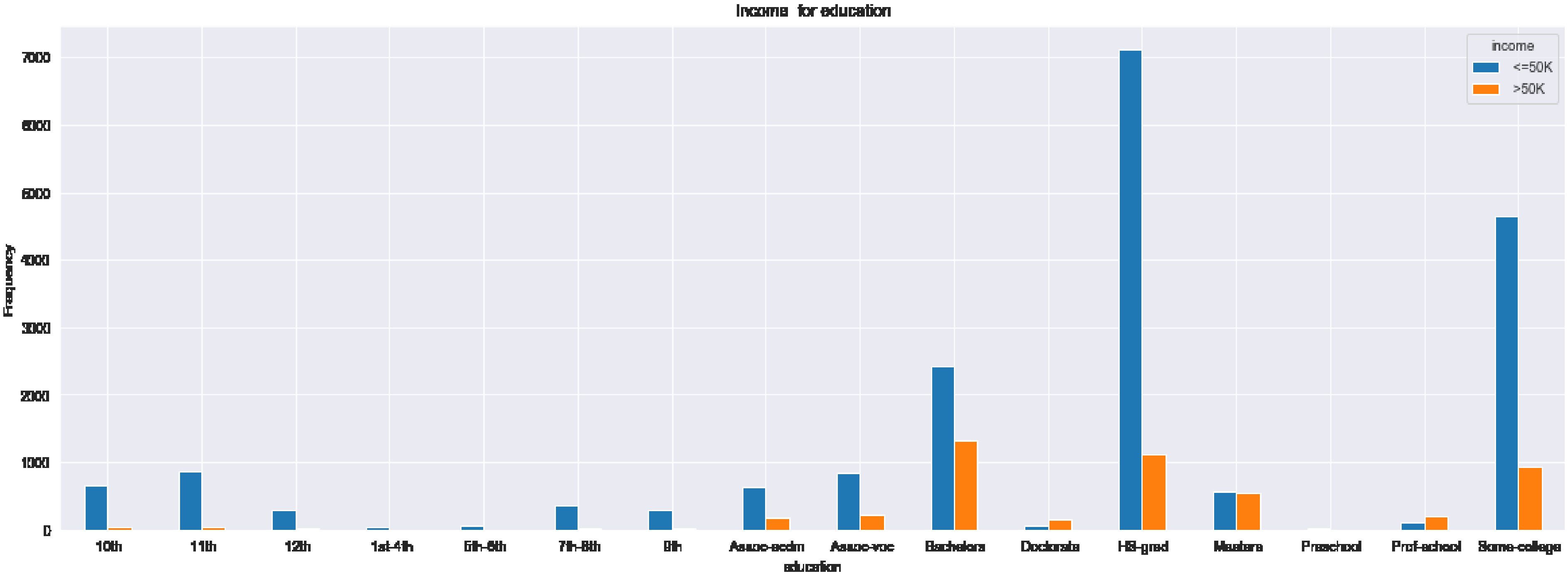
I then looked at stats and trends looking for what are the most common factors for people who make more than $50,000 each year.

Findings

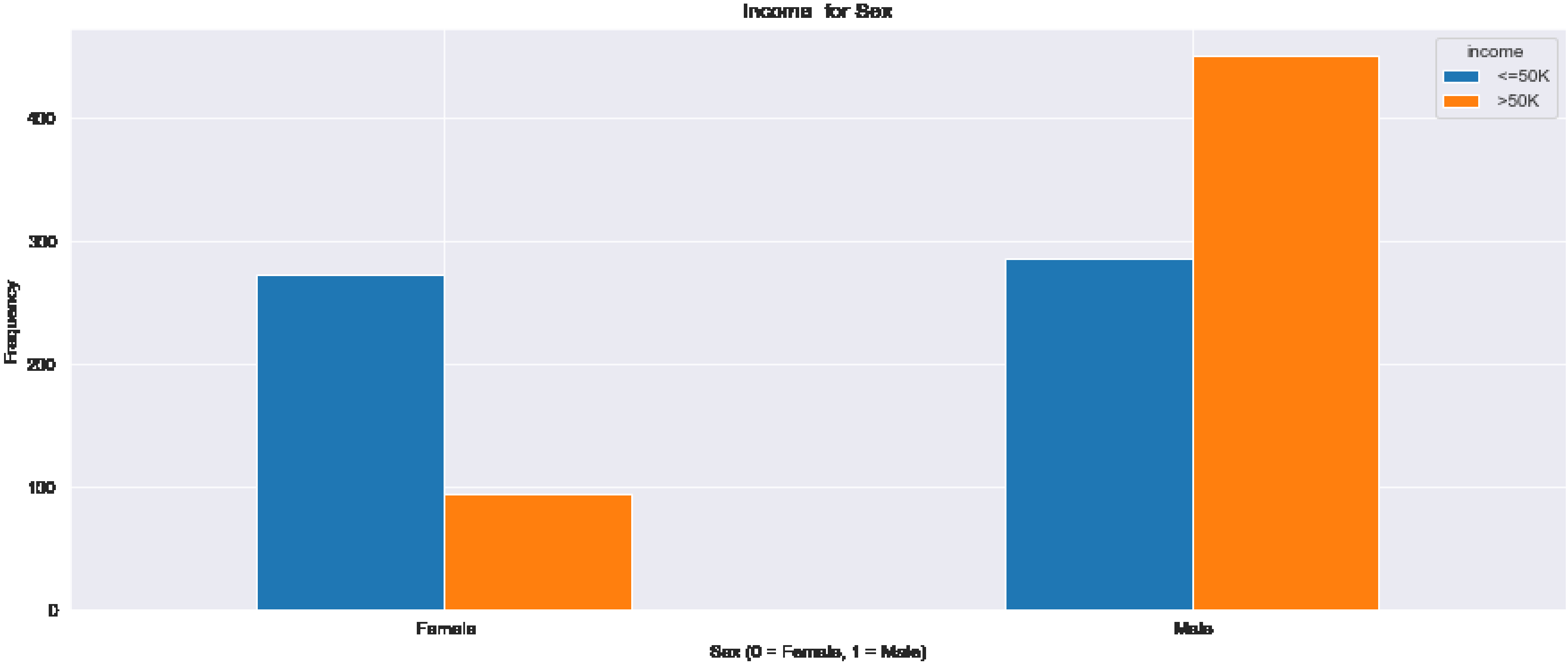
Several factors can be used to predict if a person is making more than $50,000 per year. The top three factors from this dataset are listed below, and the following slides each have a visual to support the findings.

- **Having a Masters degree**
- **Working in an occupation in management or a professional**
- **Being a Man (Even with a masters men have a much higher chance at making at least 50K)**

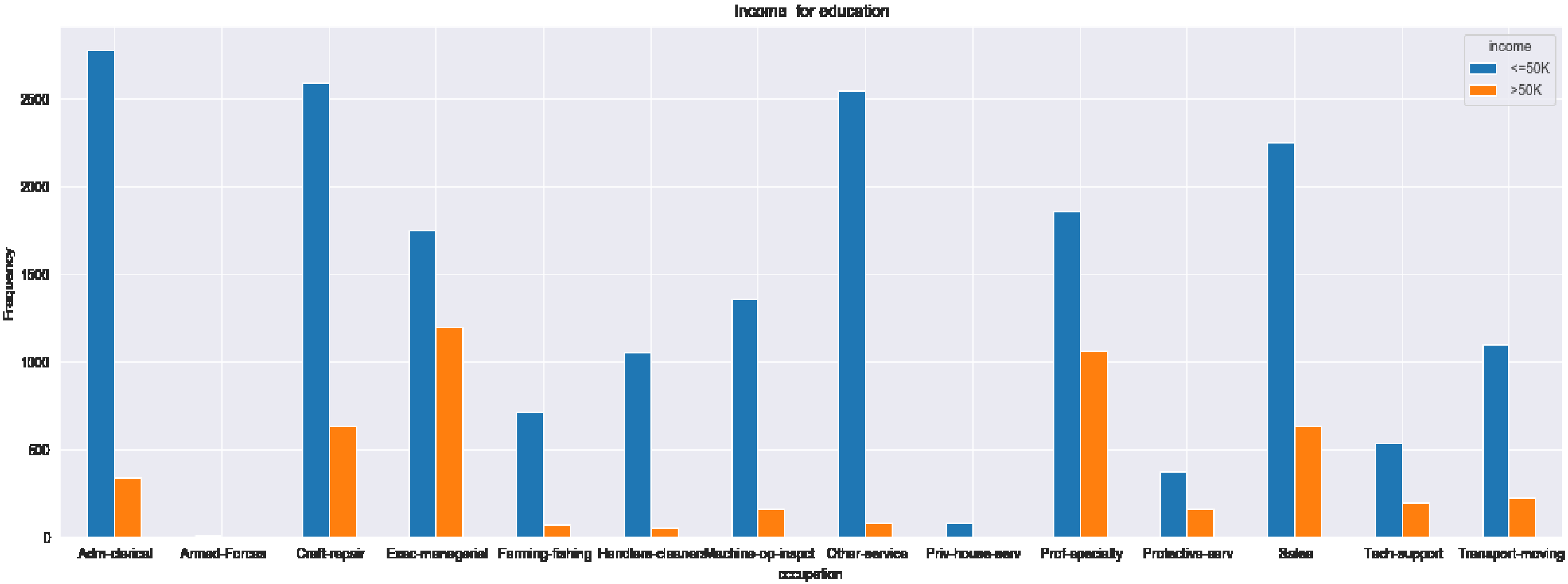
Having a Masters degree



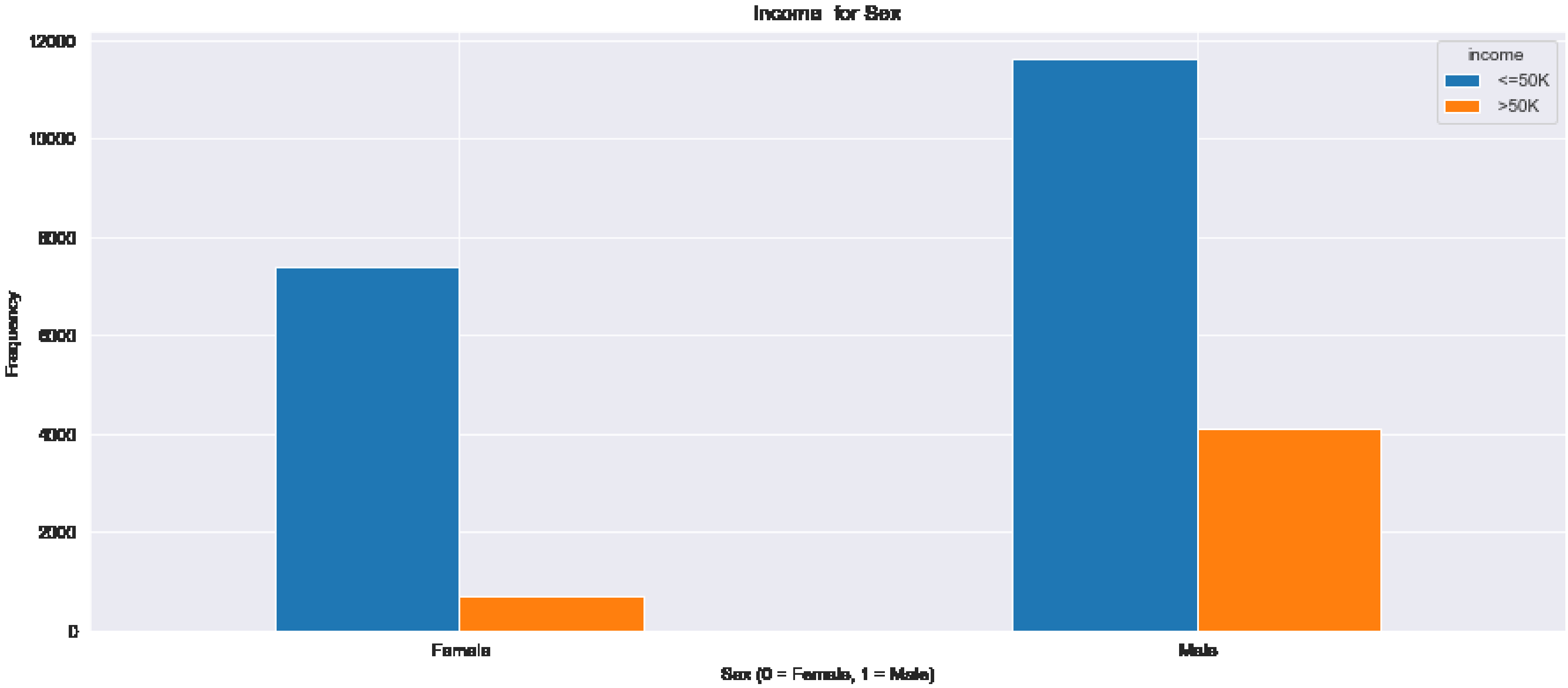
Masters degree by sex



Occupation


Income for education
Frequency
0
500
1000
1500
2000
2500
income
<=50K
>50K
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
occupation

50K by sex



Income for Sex
income
<=50K
>50K
Frequency
0
2000
4000
6000
8000
10000
12000
Female
Male
Sex (0 = Female, 1 = Male)

Limitations

This database is limited by several factors, including a focus on USA only, people who have work class listed, and people who do not have capital gains of losses. If for example you wanted to see if being from another country than the USA influences your income, then you would need to add back in those entries, or find a database with more details around that attribute.

Conclusions

- Education, Occupation, and sex are the three greatest factors that determine if someone makes more than $50,000 per year from the dataset.
- Having a Masters degree gives over a 50% chance of making more than 50K
- Men are 5 times more likely to make 50K than women however; if a women gets a Masters degree that ration falls in half.
- Working 60 hours a week or more greatly increases the chance of making more than 50K

Acknowledgements

I go the data from Kaggle in the following location: https://www.kaggle.com/lodetomasi1995/income-classification#income_evaluation.csv

I also reviewed a few of the Kernels that were run on the dataset for some ideas of what to focus on or code examples. I also used examples from the class notes from this class.

I was the only reviewer of this project due to time limitations.

References

I do not have any references for this work other than what I have reported from the Kaggle dataset. https://www.kaggle.com/lodetomasi1995/income-classification#income_evaluation.csv

All the rest of the work from my own research. Programing ideas and concepts where from the class and from the kernels from Kaggle.