# Feature Observations

## Age

|  |  |
| --- | --- |
| Figure 1: Histogram of the age feature | * Uniform distribution with people making >50K   + Goes up at 20s and peaks at people in their 30s and 40s * <=50K income is heavily skewed right, goes down as age increases   + However people in there 60s and above could just be because of population |

## Workclass

* Self-emp-inc is higher in the population of >50K than <=50K (approximately 24% difference)
* Never-worked doesn’t appear in the >50K population but only has 10 entries

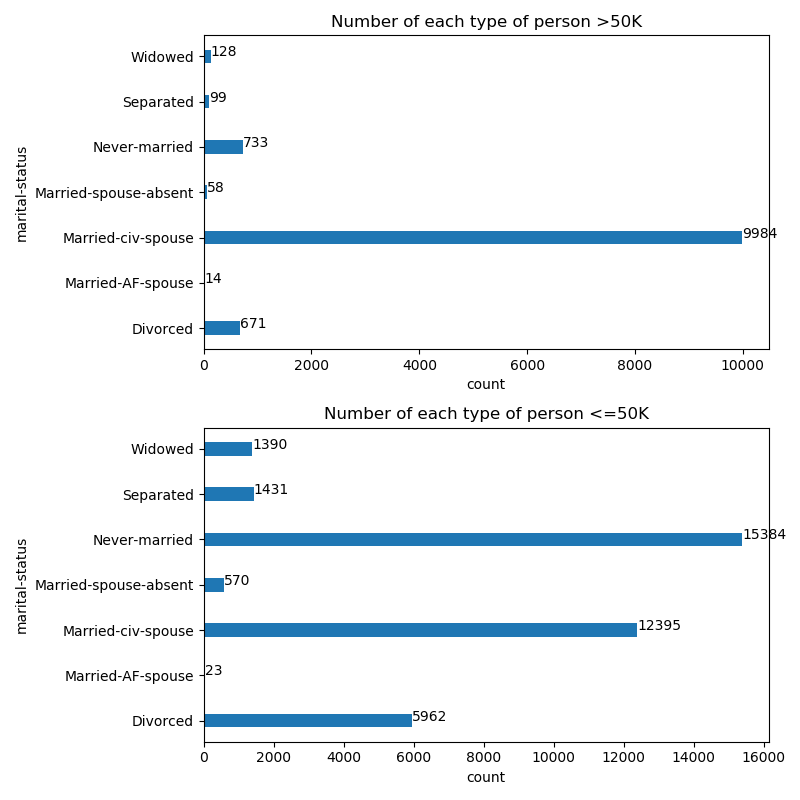
## 

## Education

* 3/4’s of people with prof-school education make >50K
* Only 1 person with their highest as preschool has an income of >50K
* Most people with highest education as high-school makes less than 50K (16% makes more)
  + Even less with a highest education below high-school make >50k
* Most people with doctorate make more than 50K

## 

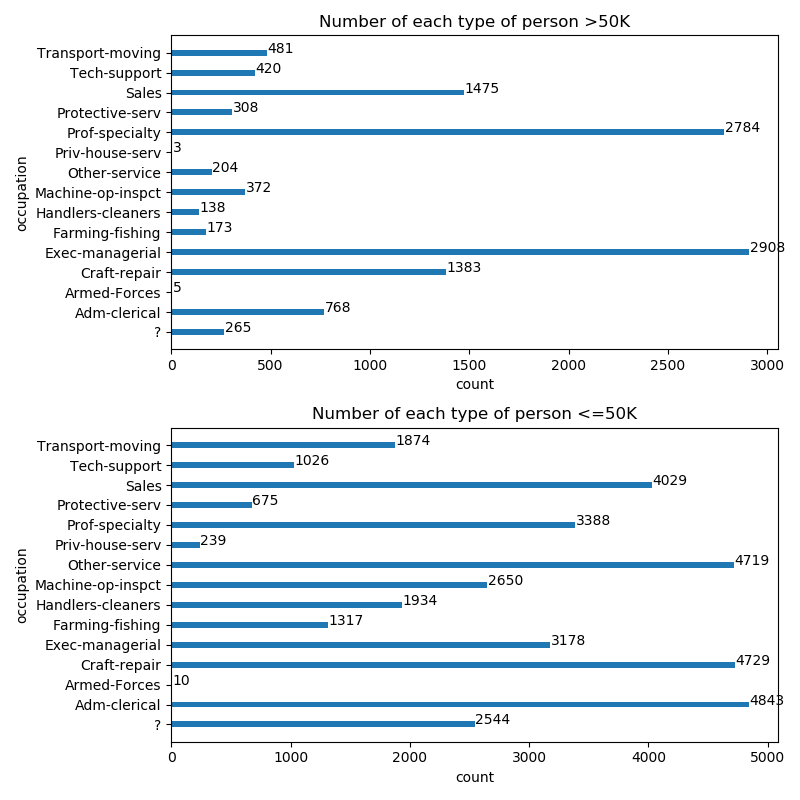
## Marital-status

* Majority of people who make >50K annually are married-civ-spouse
* Only 4.5% of people who never-married make more than 50K annually
* Very few outside of married makes more than 50K annually

## 

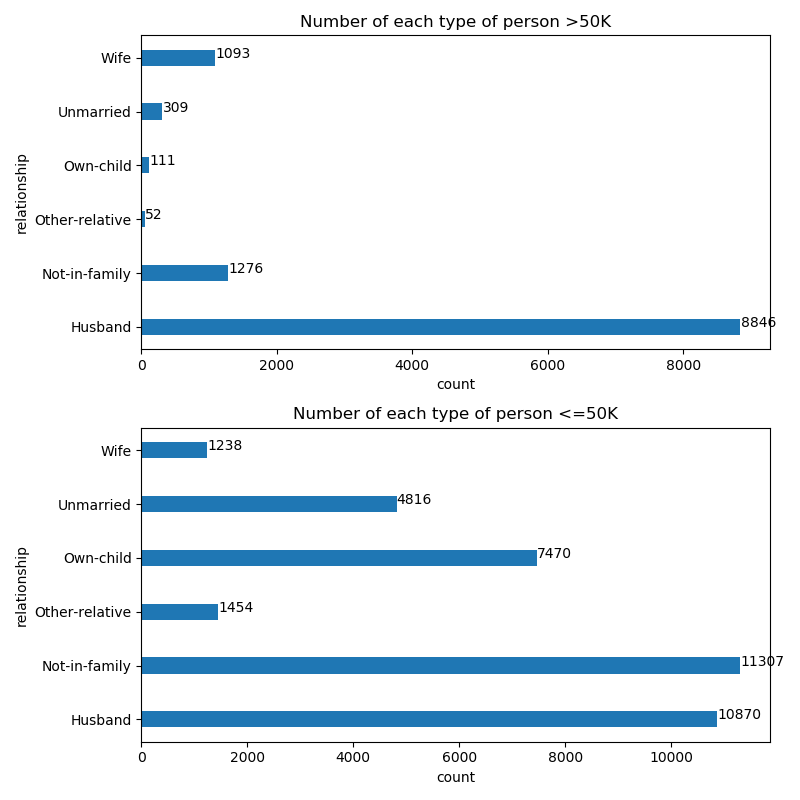
## Occupation

* >50K and <=50K annually almost have the same amount of people who occupation is Exec-managerial
* Only 3 priv-house-serv make more than 50K annually out of 242 with the occupation
* ? does appear to have people >50K but it is very few (approximately 10%)



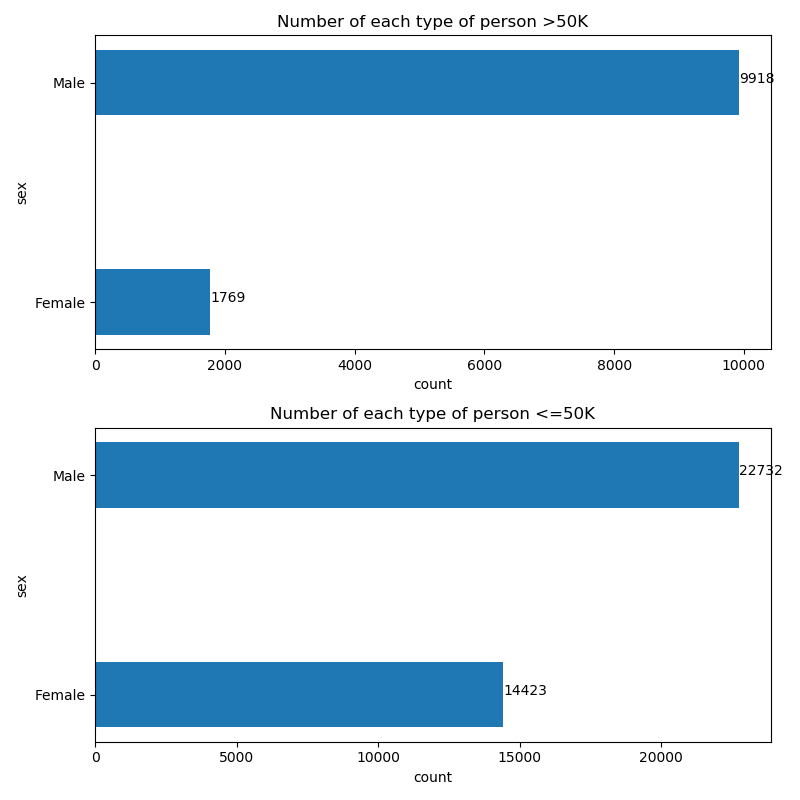
## Relationship

* Majority of people making >50K are a wife or husband
* Very few unmarried, own-child, other-relative make >50K



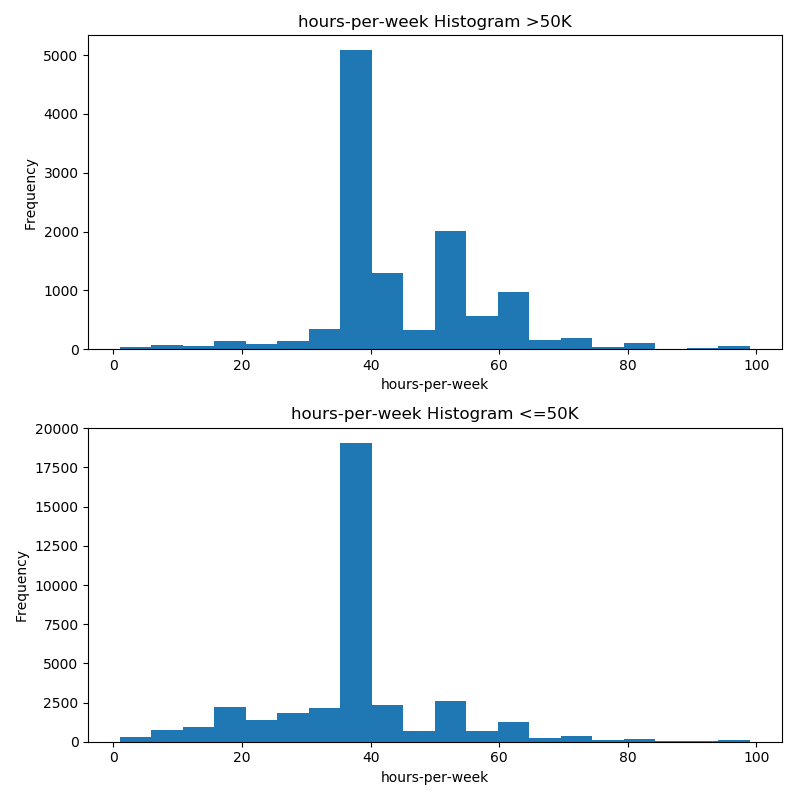
## Sex

* Males are more likely to make >50K a year (30% of males and 11% females make >50K annually)



## Hours-per-week

* A majority of people worked 40 hours but not much working less than 40 hours make >50K a year



## Features that didn’t really show much

* Fnlwgt, Race, Native-country, Capital-Gain, Capital-loss, Education-num (same as education but less information

# Results

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Training Error | Testing Error | F1\_Score |
| SVC | 0.2408 | 0.2362 | N/A |
| Decision Tree | 0.1428 | 0.1430 | 0.6568174 |
| Random Forest | 0.1397 | 0.1388 | 0.6522137 |

Note: F1\_Score is another measure of accuracy, 1 means the model has perfect prediction while 0 means the opposite

# Chosen Model (Decision Tree)

Decision Tree proved to work quite well with the dataset given, giving a strong error rate of 0.1430 on the test set at only depth 10. This is relevant as normally the higher the depth of the tree is, the more likely you are to overfit to the dataset. It also consistently got a better F1\_Score than Random Forest, despite a slight increase in error.