Trends in Income

Based on US Census Data

Goals:

- Target variable is annual income (over/under 50K)
- Look for trends
- Find variables that best predict income, which can be looked into more deeply in future projects

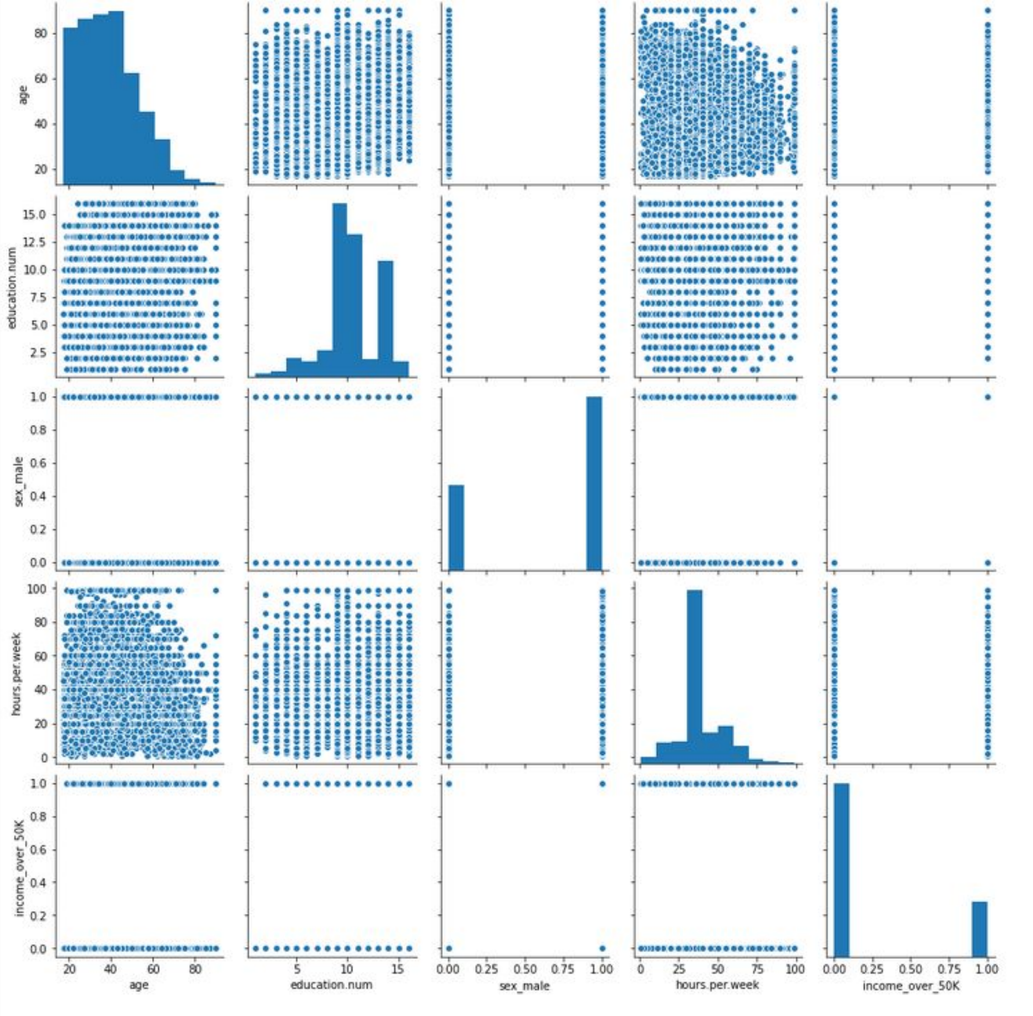
The Data

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

- Income is our target variable. Note that income data was exactly this coarse to start with. No exact incomes were given
- Few numeric features
- education and education.num are 1:1 correspondence, so we can drop one of these columns
- I could not understand the meaning of ‘fnlwgt’ feature, so I dropped it.

First Look

- Age is right skewed
- Education’s two peaks:
  - high school graduated + some college
  - completed Bachelor’s degree
- Sex is heavily skewed towards Male in a 2:1 ratio of Male:Female. Imbalance might indicate some fault in the data gathering method? Could also indicate my own misunderstanding of the assumptions of the survey.
- Income for over/under 50K is about 3:1 ratio of under:over.
- Hard to get much information from the pair plots, since most variables have a very small number of levels.

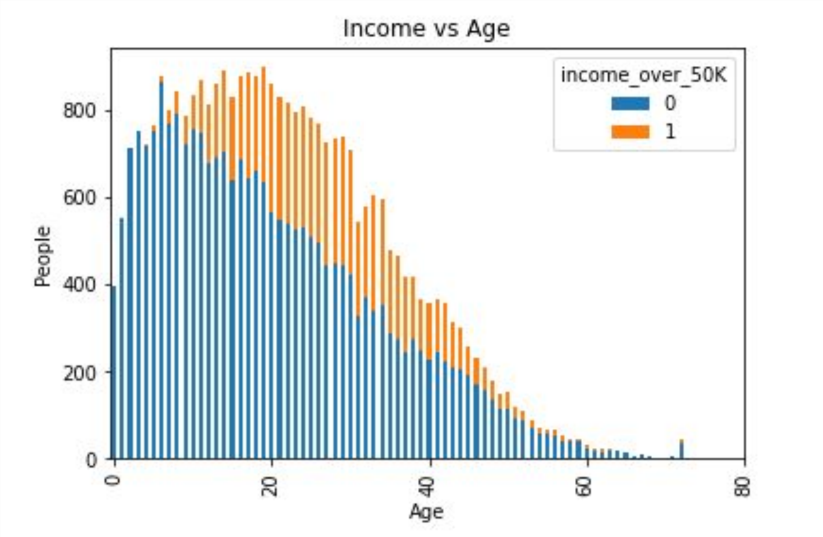


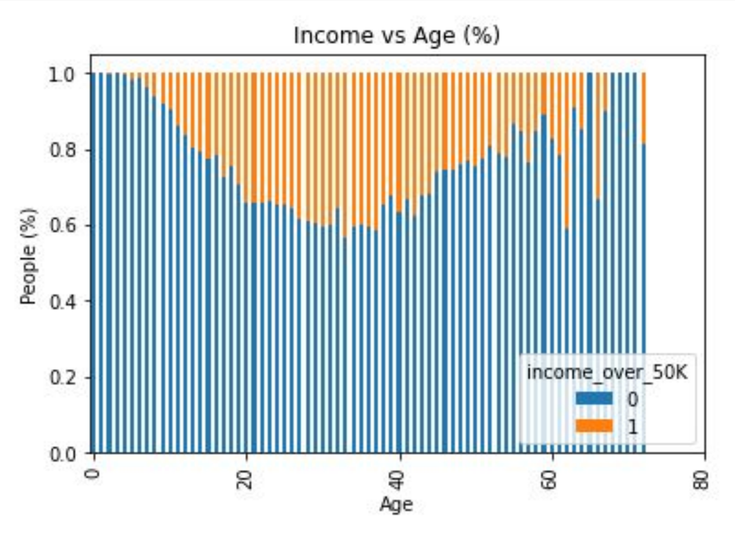
Next Steps: Fit a Random Forest model, and use feature importances to get best predicting features.

```
Feature ranking:
1. feature age (0.276803)
2. feature education.num (0.154152)
3. feature hours.per.week (0.135545)
4. feature marital.status_Married-civ-spouse (0.072537)
5. feature relationship_Husband (0.049222)
6. feature occupation_Exec-managerial (0.023039)
7. feature marital.status_Never-married (0.020708)
8. feature occupation_Prof-specialty (0.020001)
9. feature sex_male (0.017956)
10. feature relationship_Own-child (0.014019)
11. feature relationship_Wife (0.013327)
12. feature relationship_Not-in-family (0.012184)
13. feature workclass_Private (0.011814)
14. feature occupation_Other-service (0.010067)
15. feature workclass_Self-emp-not-inc (0.009500)
16. feature occupation_Craft-repair (0.008205)
17. feature marital.status_Divorced (0.008093)
18. feature occupation_Sales (0.007924)
19. feature race_White (0.007918)
20. feature native.country_United-States (0.007780)
```

Format is rank, feature name, feature relative importance

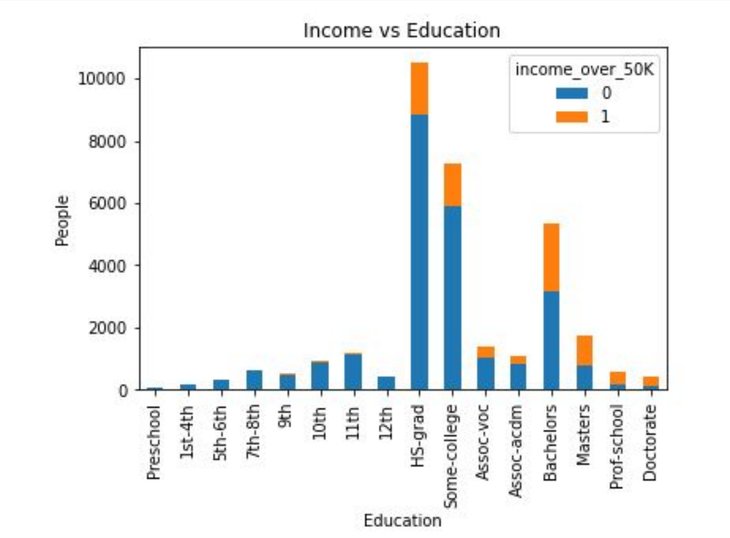
Age

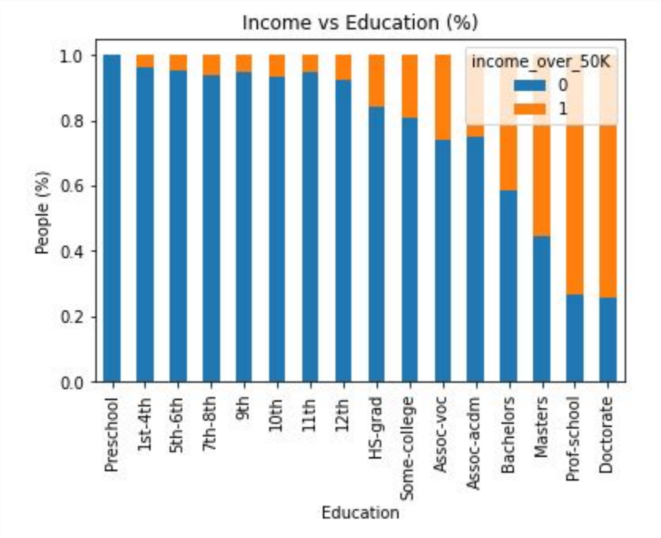




The above graphs show the same information, with the left graph having absolute amounts while the right is percentage based

Education



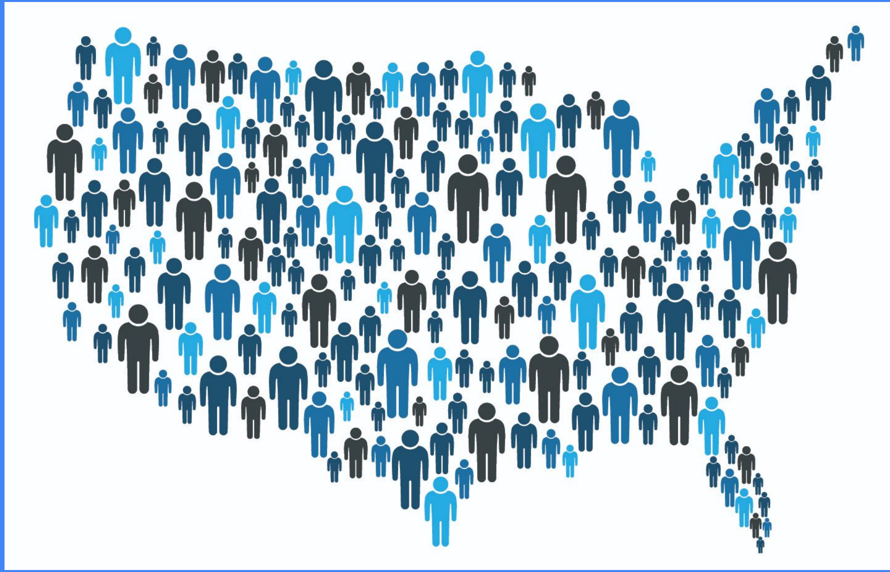


The above graphs show the same information, with the left graph having absolute amounts while the right is percentage based

Conclusion

Some important features in predicting income:

1. Age
2. Education
3. Hours per week
4. Marital status
5. Occupation



Further work:

- Investigate interactions between features- e.g. how does marital status affect the income distribution for age, or race, or education?
- Fit other models, maximize predictive accuracy
- If the income was numerical, rather than only being over/under 50K, I would especially like to see the cost vs benefit of various education levels on income