Income of Adults: Data Storytelling

Paul Jacob

Problem

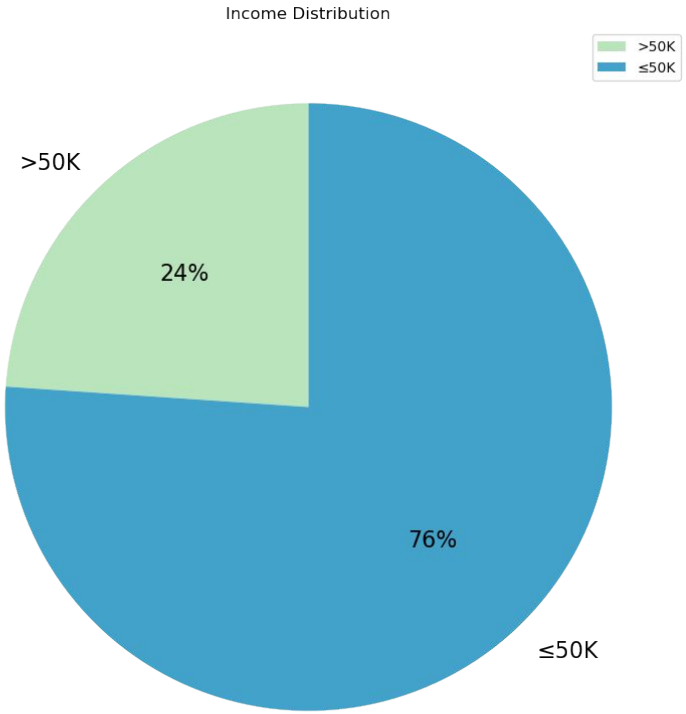
- Problem: Lenders, land lords, and similar entities want to know payment reliability. Aside from having substantial savings or assets, earning an adequate income is a major contributor to payment reliability. From this demographic dataset for adult individuals and their income bracket (i.e. >50K or ≤50K), we want to better understanding variable relationships within the data using visualizations and hypothesis testing.

- Questions: What relationships exist within the income bracket demographic dataset? And what are some of the measurable factors that influence individual annual income in the United States?

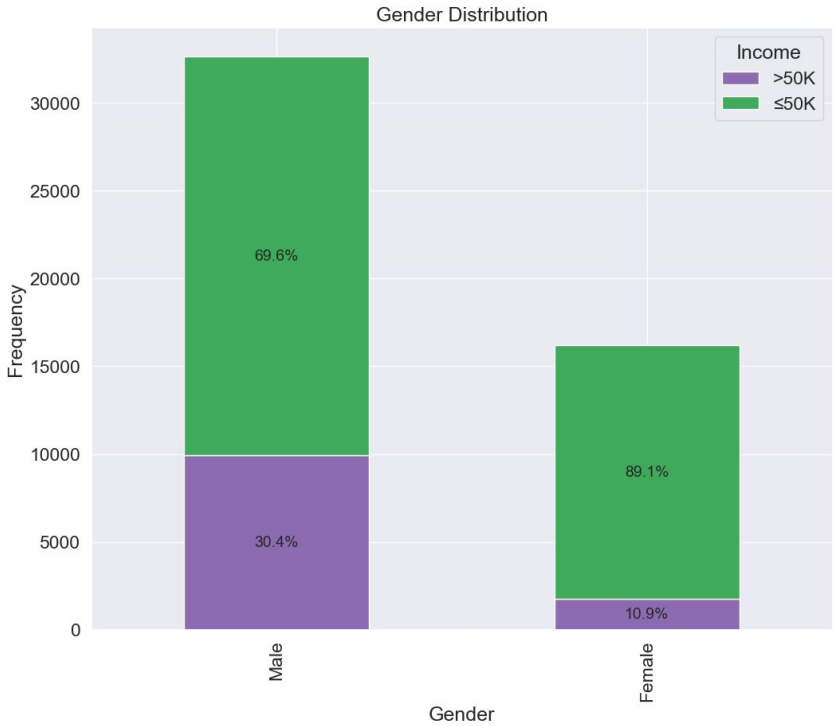
- Solution: We look to measurable factors that influence individual income bracket, namely individual annual income >50K and ≤50K. From the data we had available and explored are demographics, e.g. age, work class, education, marital status, occupation, relationship, race, gender, capital gain, capital loss, and native country. We look at trends, correlations, and glean insights influencing individual income.

Income Distribution

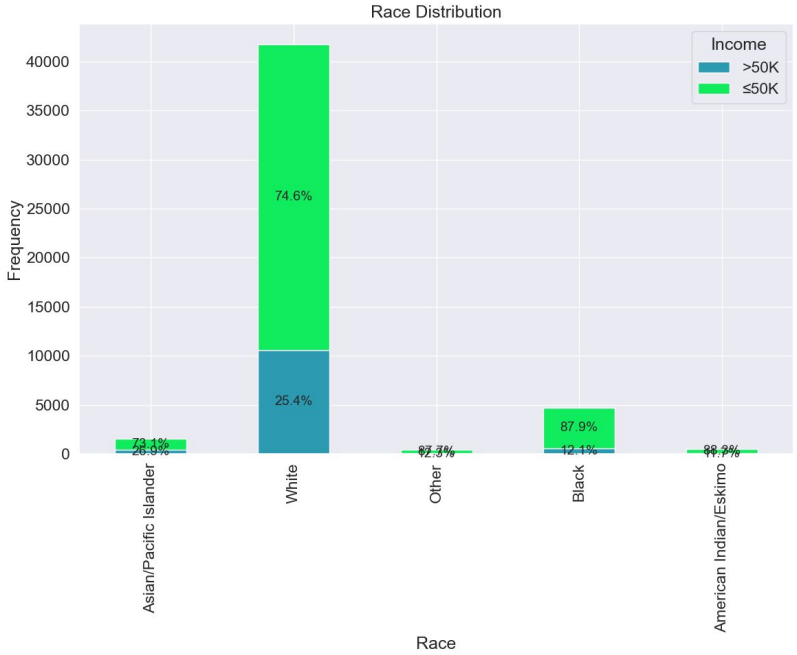
- 3:1 ratio for individuals with ≤50K in annual income to individuals with >50K in annual income.



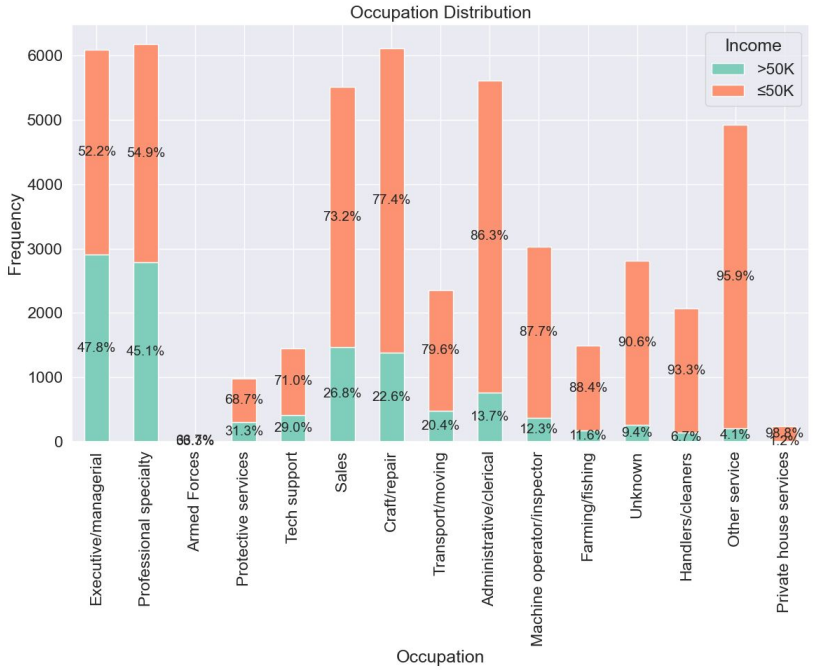
Gender Distribution

- 

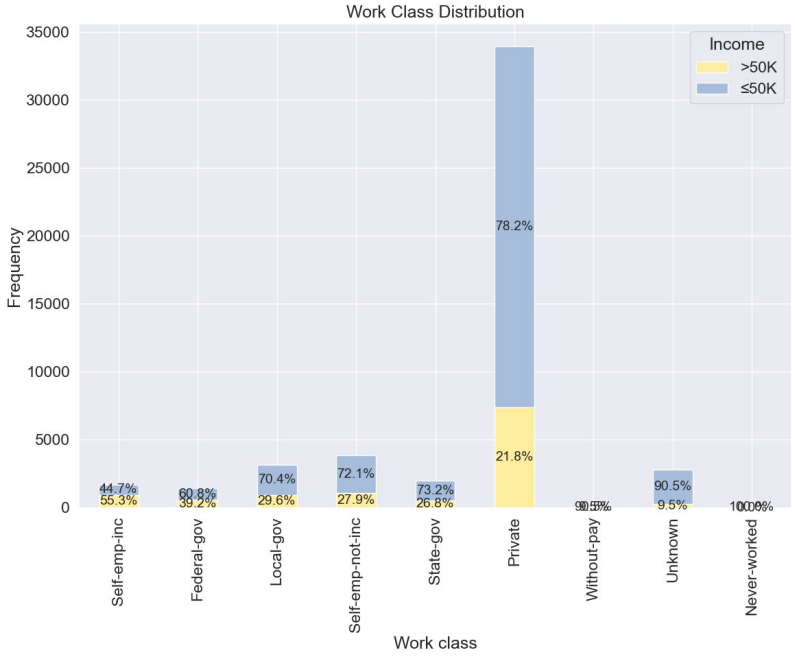
Race Distribution



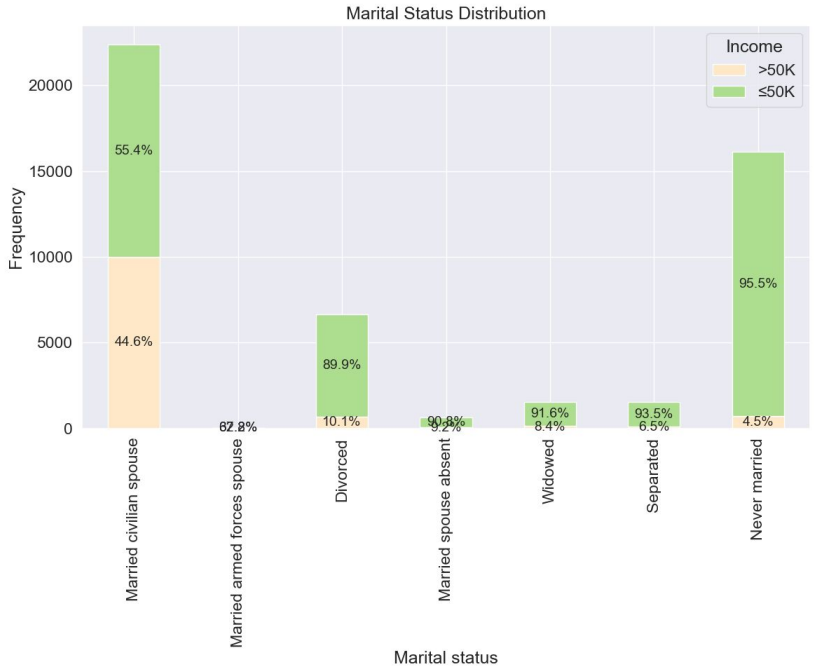
Occupation Distribution



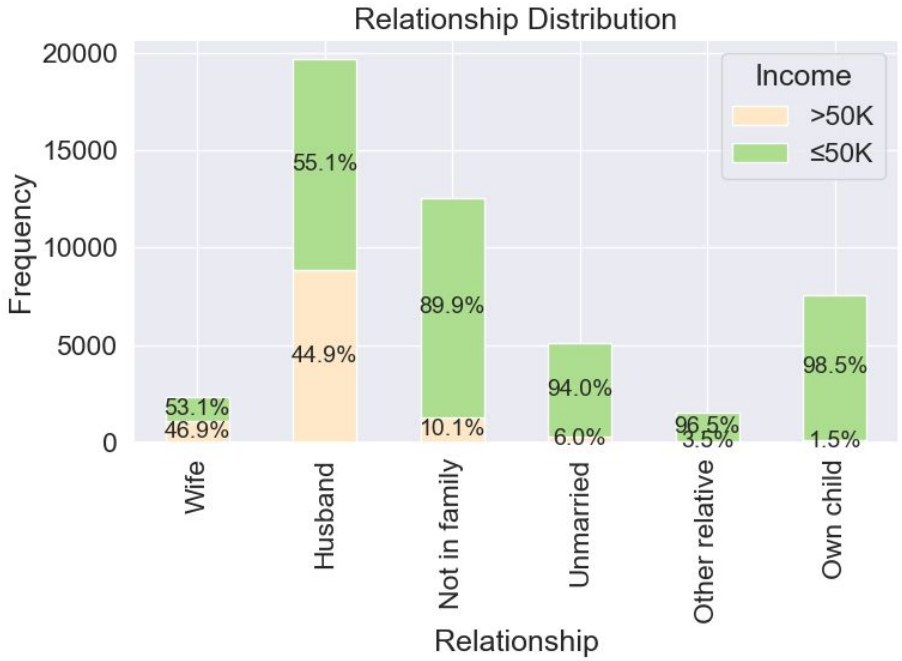
Work Class Distribution



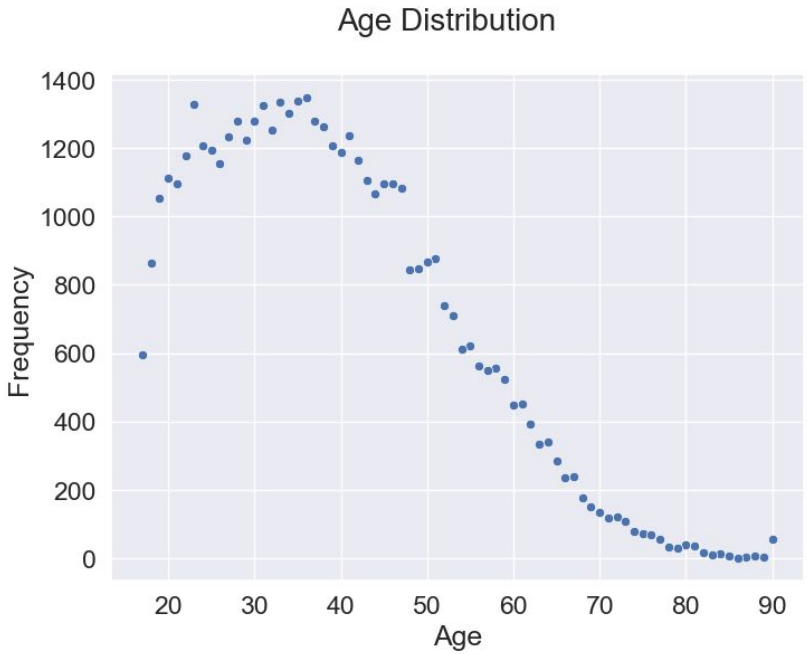
Marital Status Distribution



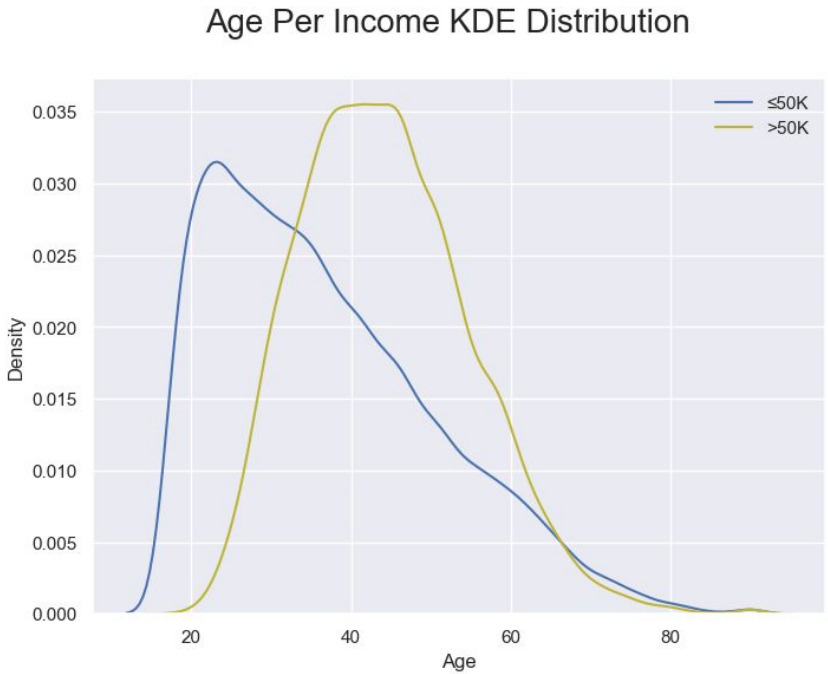
Relationship Distribution



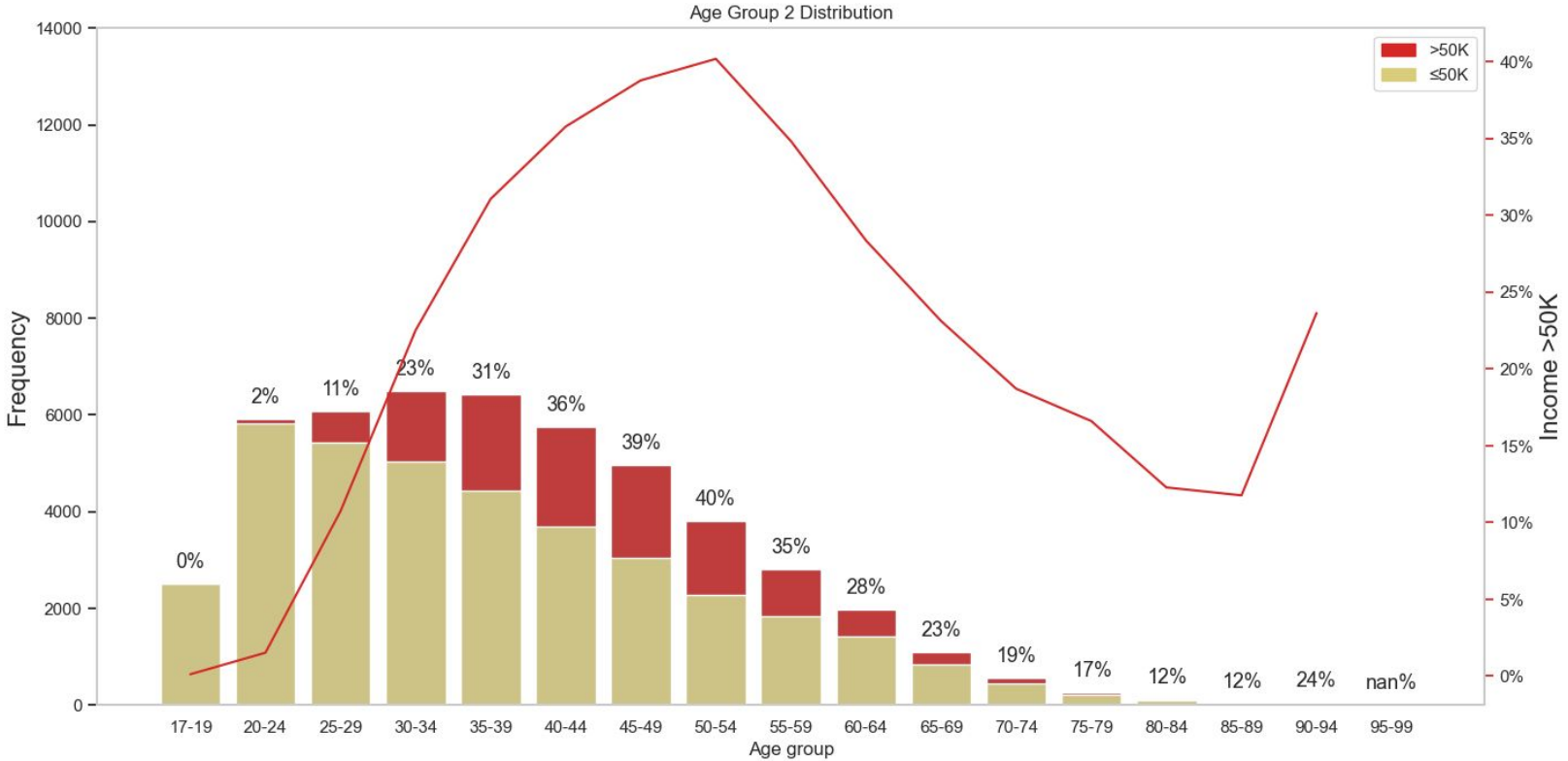
Age Distribution



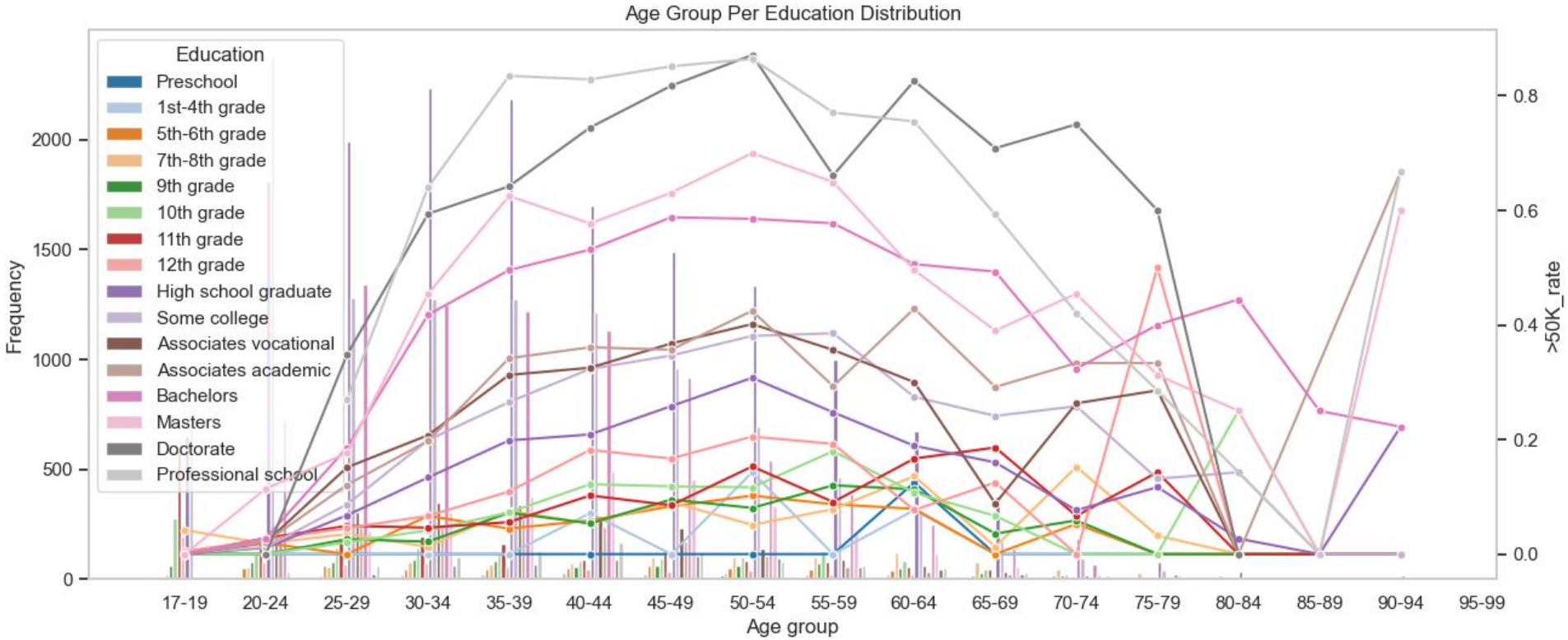
Age Distribution



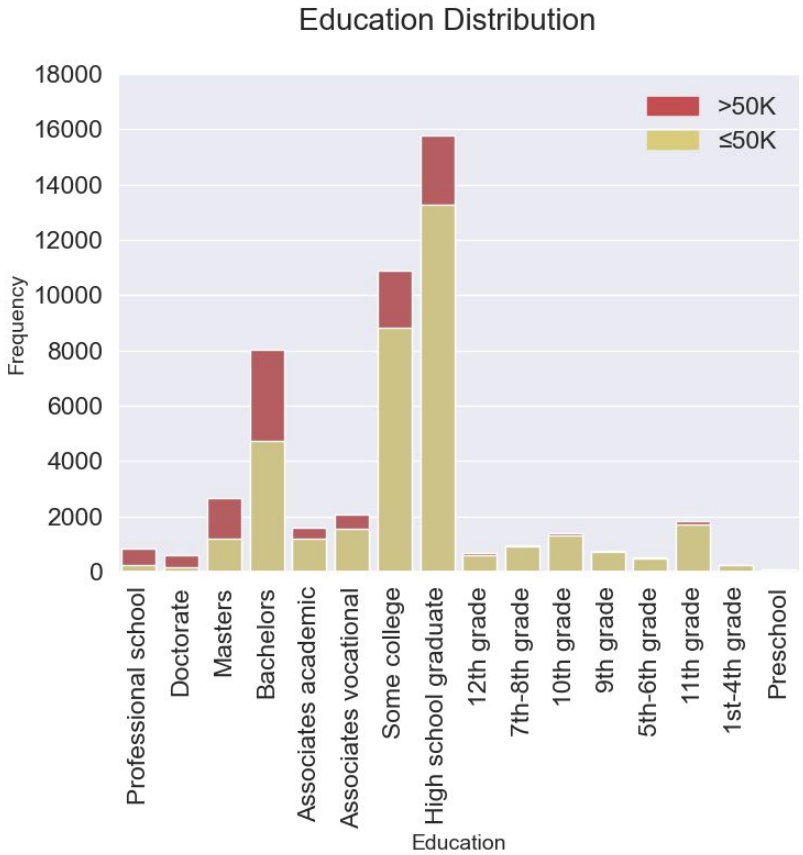
Age Group Distribution



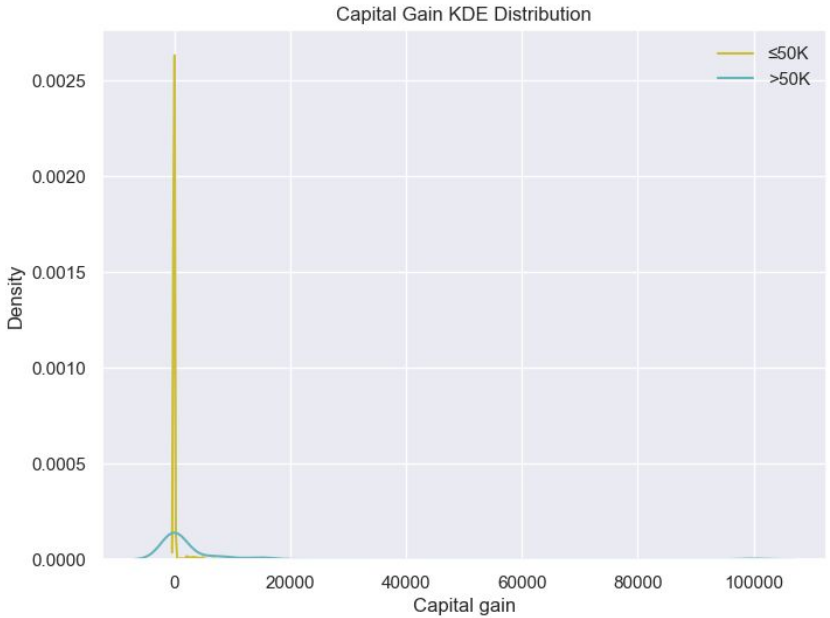
Age Group Per Education Distribution



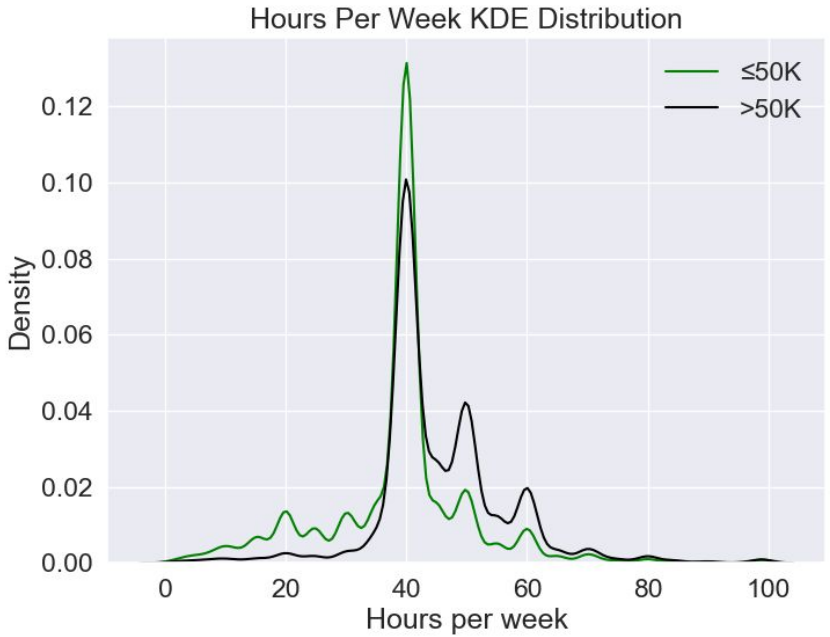
Education Distribution

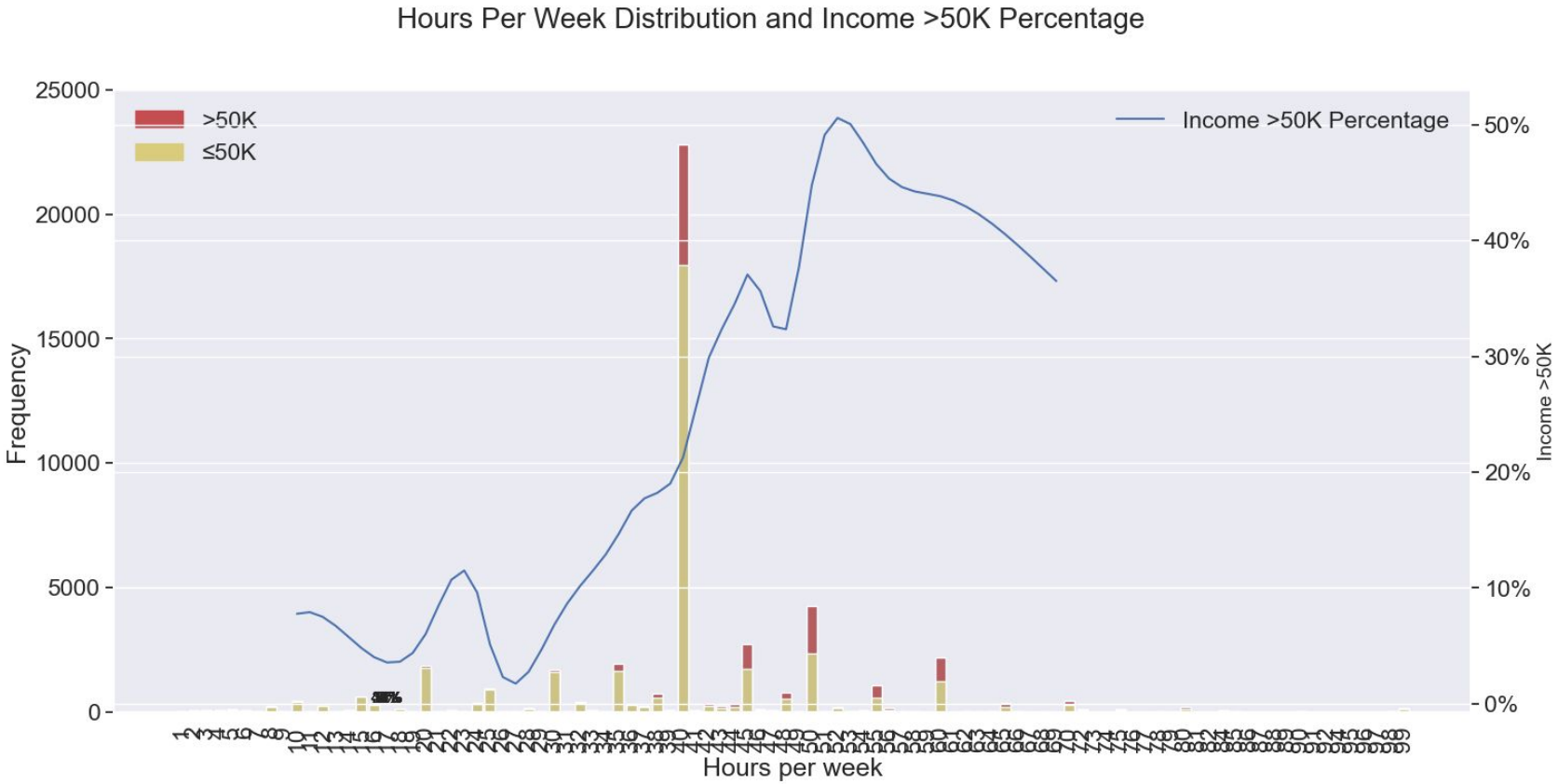


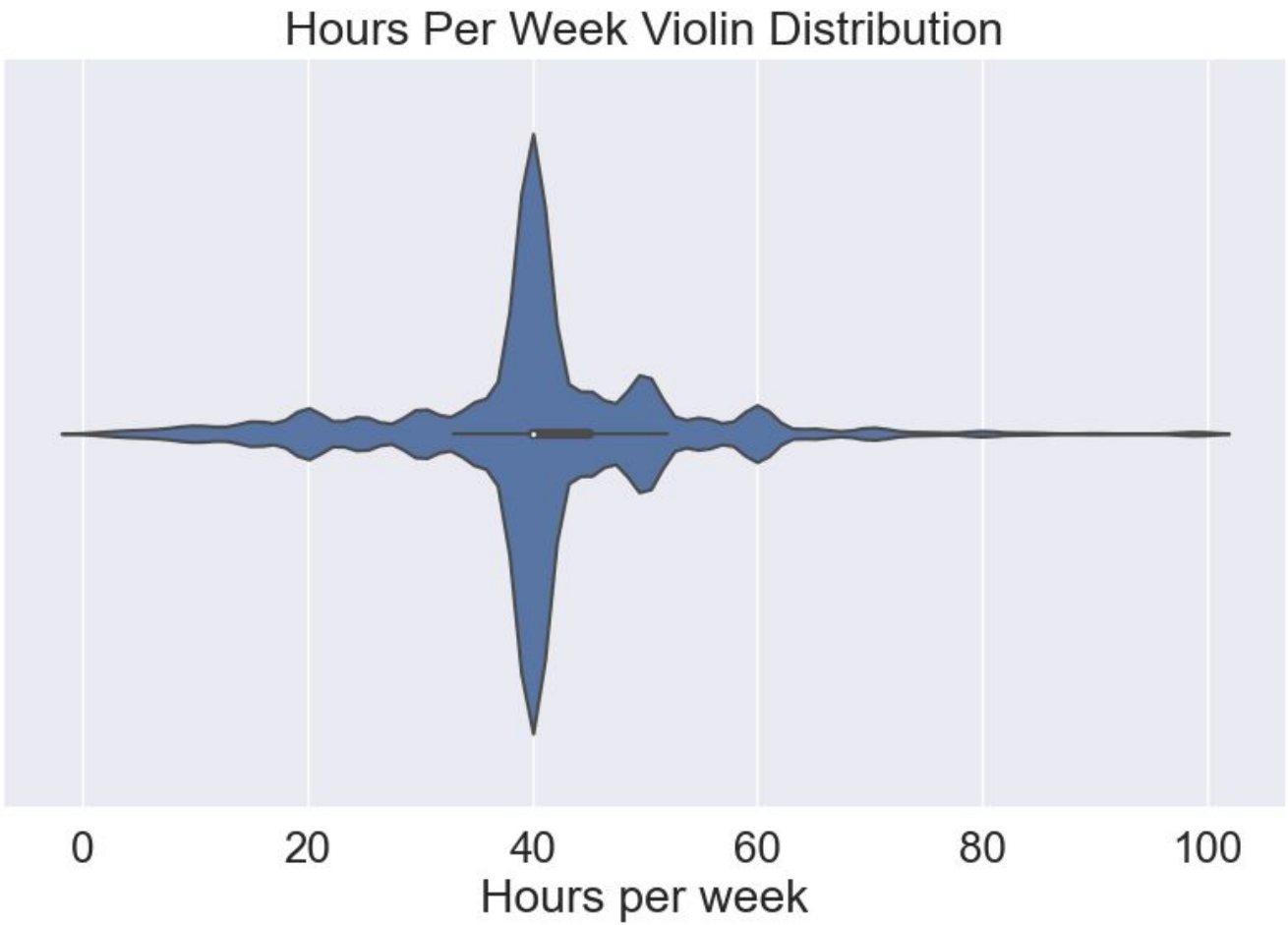
Capital Gain KDE Distribution

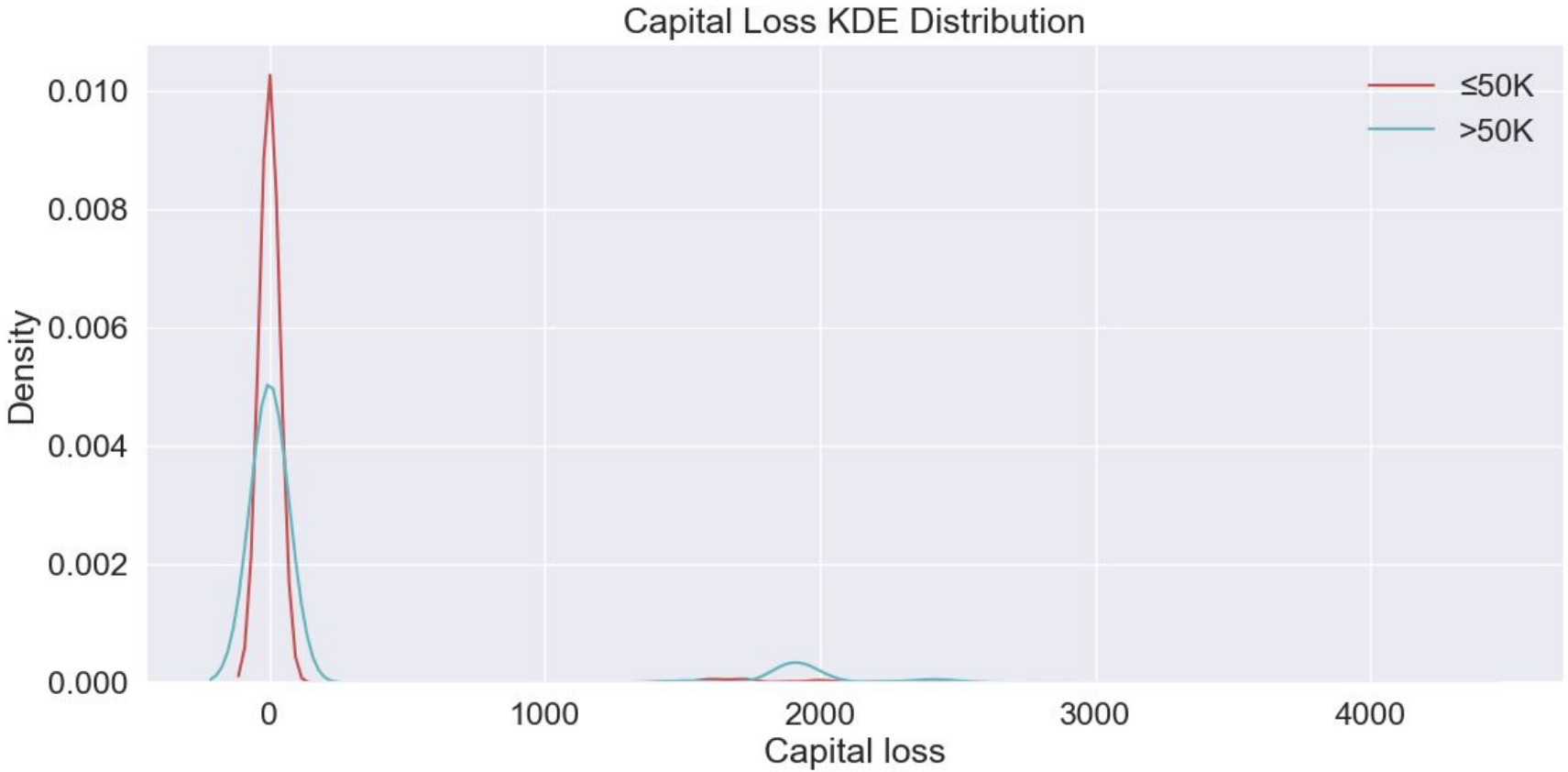
- 

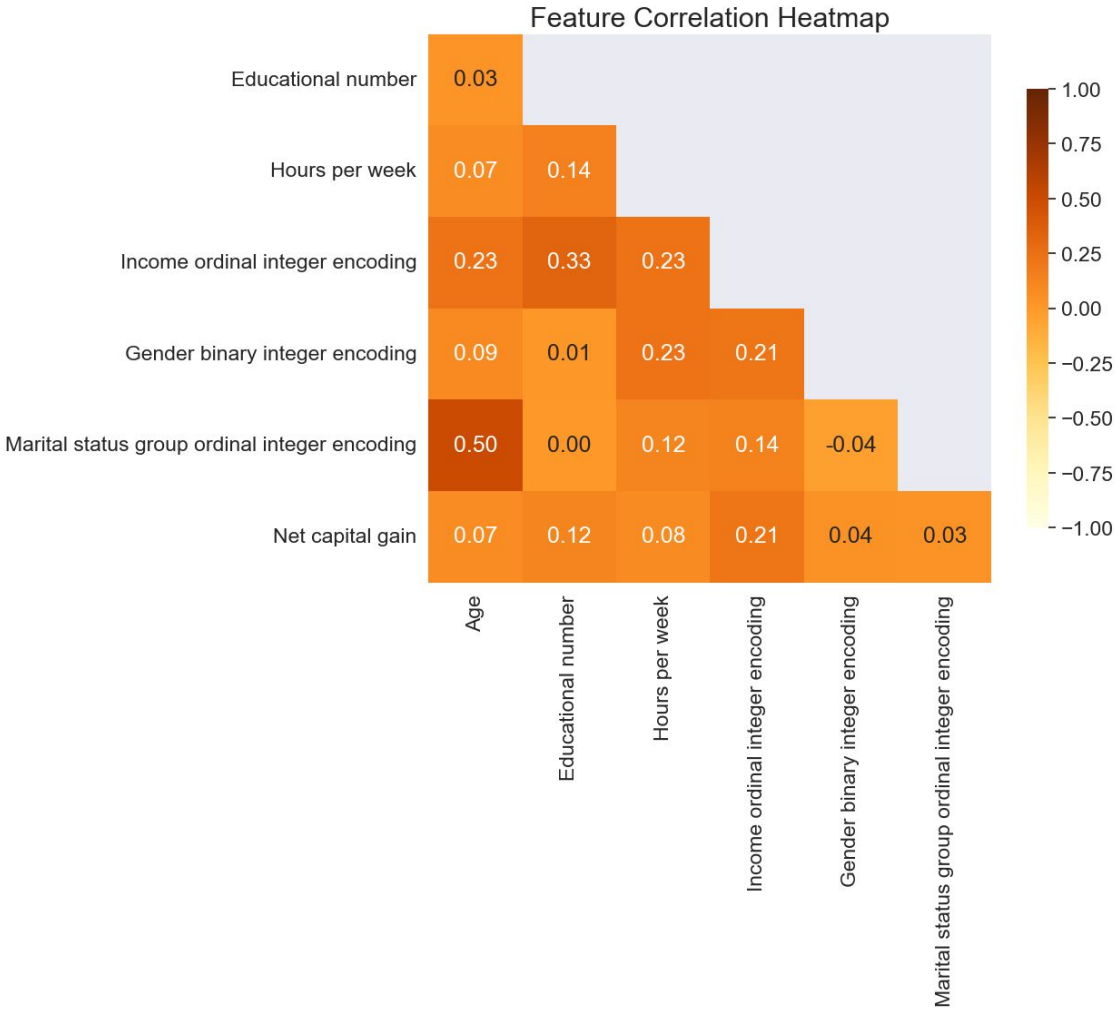
Hours Per Week KDE Distribution

- 


Hours Per Week Distribution and Income >50K Percentage
>50K
≤50K
Income >50K Percentage
Frequency
Income >50K
Hours per week
25000
20000
15000
10000
5000
0
50%
40%
30%
20%
10%
0%

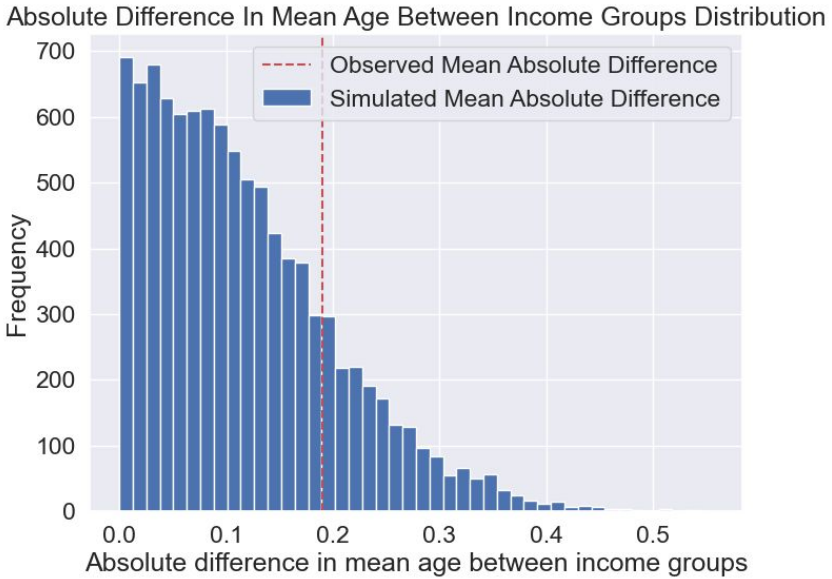
- 
Hours Per Week Violin Distribution
0
20
40
60
80
100
Hours per week


Capital Loss KDE Distribution
≤50K
>50K
Density
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
Capital loss


Feature Correlation Heatmap
Educational number
Hours per week
Income ordinal integer encoding
Gender binary integer encoding
Marital status group ordinal integer encoding
Net capital gain
0.03
0.07
0.14
0.23
0.33
0.23
0.09
0.01
0.23
0.21
0.50
0.00
0.12
0.14
-0.04
0.07
0.12
0.08
0.21
0.04
0.03
Age
Educational number
Hours per week
Income ordinal integer encoding
Gender binary integer encoding
Marital status group ordinal integer encoding
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Hypothesis Test: Permutation Test

- **Null hypothesis:** The observed difference in the mean age of ≤50K income and >50K income individuals is due to chance (and thus not due to income bracket).
- **Alternate hypothesis:** The observed difference in the mean age of ≤50K income and >50K income individuals is not due to chance (and is actually due to income bracket).
- **Finding:** We reject the null hypothesis. Using a significance level of 5%, the number of absolute differences greater than the observed absolute difference is 18.92% (i.e. >5%).



Conclusion

- There is a skew for the number of individuals making >50K compared to ≤50K in the dataset at 1:3 ratio.
- The top correlations with income bracket were education, age, and hours worked per week at 33%, 23%, and 23% respectively.

Appendix