INTRODUCTION TO MACHINE LEARNING AND DATA MINING

REPORT 1

Group 314	Section 1	Section 2	Section 3	Section 4	Exam questions
Andrea Arieti s225570	45%	20%	20%	45%	33.33%
Andrea Tessarin s221771	40%	50%	20%	25%	33.33%
Stefan Slavkovsky s221931	15%	30%	60%	30%	33.33%

Section 1: INTRODUCTION

1.1 DESCRIPTION OF THE "ADULT" DATASET

The "Adult Income" or simply "Adult" dataset is a common machine learning dataset that will be used in this project by our group. The dataset, which is attributed to Ronny Kohavi and Barry Becker, was derived from the 1994 United States Census Bureau data.

The dataset contains 48842 observations (in the original dataset divided in train=32561, test=16281) of personal information like age, occupation, sex, relationship etc., and whether given individual salary exceeds $50,000 annually.

1.2 DATA SOURCE

All the data contained in dataset used in this project, are obtained via the website (here is the link); https://archive.ics.uci.edu/ml/datasets/Adult

1.3 SUMMARY OF RELEVANT PAPER CONNECTED TO THIS DATASET

Title: "Predicting earning potentials using the Adult dataset"

The Adult dataset is one of the most famous sources in the field, therefore many papers use it as a base for various experiments. One of the most interesting works about this dataset is "Predicting earning potentials using the Adult dataset", by Haojun Zhou. This study compares traditional statistical modelling and machine learning techniques in terms of performance, for predicting if the income is above 50 thousand dollars. Study uses logistic regression as statistical modelling tool and compares it with four different machine learning techniques: neural network, classification and regression tree, random forest, and support vector machine.

Study firstly transformed dataset by excluding certain attributes (fnlwgt) and aggregating certain categories inside work class attribute.

Study continued with explorative analysis by visualisation. Correlations between the income and the number of years in education or the race has been empirically shown in study besides other correlations.

Logistic regression was then applied using income as the response value, to model the probability that an individual makes more than 50k annually. To binarize the results a 0.5 threshold was used. Accuracy od this method was 82.74%.

Then four machine learning techniques were applied to the dataset with the same objective and the following results were reported:

- Neural network: 83.37% accuracy
- Regression tree: 82.94% accuracy
- Random forest: 83.71% accuracy
- Support vector machine: 83.06% accuracy

Given the results obtained, the conclusion is that, even though all machine learning methods provide better accuracy than logistic regression, in this case, we cannot say that the latter is totally outperformed, so the conclusion from the author is that: "Machine learning is no denying a powerful, but it should not be considered as a substitute of traditional statistical modelling."

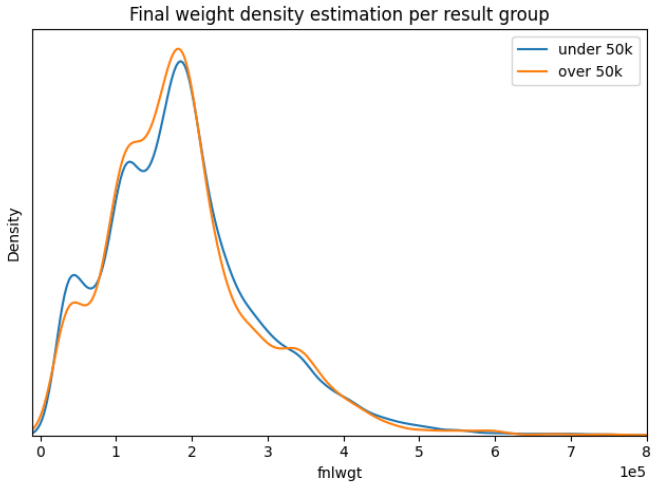
1.4 CONSIDERATIONS FOR FURTHER CLASSIFICATION AND REGRESSION ANALYSIS OF THIS DATASET

We have chosen this dataset because of the substantial number of observations available to conduct the classification and regression analysis. In the following lines it will be explained what we hope to achieve for both, keeping in mind, however, that there may be changes in progress.

- Classification problem: we want to understand, based on the available data, if a specific targeted individual is earning more or less than 50000$ per year.
- Regression problem: here we have two different ideas. Firstly, we can try to predict the age of each person using all the other attributes. Secondly, it could be also interesting to predict the work hours based on all the other attributes.

1.5 DATA TRANSFORMATION

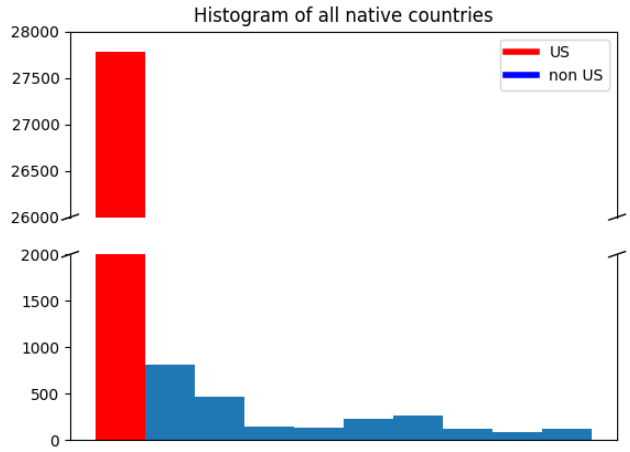
We transformed the data to carry out these tasks. After a thorough analysis of the data, we discovered some issues and opportunities to optimize our dataset to facilitate further analysis.



Primarily, we decided to remove the final weight attribute (fnlwgt), which indicates the number of people in a specific category who share the same attribute. We discovered during the analysis that the density distribution of final weight is the same for people who earn below or more than $50,000. This attribute is thus unlikely to influence our prediction.

Figure 1: Similar density distribution for classes.

Furthermore, we have combined capital gain and capital loss. Both capital loss and capital gain are non-negative integers and are mutually exclusive. There is therefore no reason to keep both capital gain/loss separately. We subtracted capital loss from capital gain.



We also noticed that most of the native countries have small number of observations in them. We aggregated all the countries that differ from the United States to balance number of observations in each category. By doing that we have binarized the attribute.

Figure 2: Disproportionality of native countries.

We also merged the “husband, wife” observations in the relationship category, because it means that they are married, and it makes no sense to consider them separately in our analysis.

Lastly, we applied ‘1 out of K’ encoding for all nominal attributes.

Section 2: DETAILED EXPLANATION OF THE ATTRIBUTES OF THE DATA

2.1 DETAILED EXPLANATION OF THE ATTRIBUTES OF THE DATA

There are 14 input variables in the dataset, which include nominal, ordinal and continuous attributes. The following is a detailed list of all the variables:

- **Age**: continuous and ratio; the age of the person
- **Work class**: discrete and nominal; one of private, self-employed not incorporated, self-employed incorporated, federal government, local government, state government, without pay, never worked
- **Final weight**: continuous and ratio; amount of people estimated to share same observed attributes
- **Education**: discrete and ordinal; the highest achieved education level from preschool to doctorate
- **Education number**: discrete and ordinal; encoded education as number
- **Marital status**: discrete and nominal; one of married to civil spouse, divorced, never married, separated, widowed, married absent spouse, married to armed forces spouse
- **Occupation**: discrete and nominal; one of tech support, craft repair, other services, sales, executive managerial, professional specialty, handlers/cleaners, machine operator/inspector, administrative clerical, farming/fishing, transport/moving, private house services, protective services, armed forces
- **Relationship**: discrete and nominal; one of wife, own child, husband, not in family, other relative, unmarried
- **Race**: discrete and nominal; one of White, Black, Asian-Pac-islander, Amer-Indian-Eskimo, Other
- **Sex**: binary and nominal; male or female
- **Capital-gain**: continuous and ratio; non salary earnings (investments, heritage, etc.)
- **Capital-loss**: continuous and ratio; non salary losses (investments, heritage, etc.)
- **Hours-per-week**: continuous and ratio; number of hours person works weekly
- **Native country**: discrete and nominal; country of origin

2.2 DATA ISSUES AND MISSING VALUES

There were some data issues in the original dataset, which we have corrected. We deleted all rows containing missing values. Some statistics of all of this are summarized in the following lines.

The missing values in the dataset are indicated by the question mark (?) symbol. So, we started from a total of 48842 rows of data, but 3620 of these have missing values. We ended up with 45222 instances with unknown values removed (train=30162, test=15060).

2.3 BASIC SUMMARY STATISTICS OF ATTRIBUTES

In the original dataset, there is five continuous and one ordinal attribute. After exclusion of final weight, we are left with summary statistics of continuous data:

	Mean	Median	Minimum	Maximum	Standard deviation	Mode
Age	38.40	37.00	17	90	13.13	31
Education number	10.11	10.00	1	16	2.54	9
Capital gain	515.49	0.00	-4356	41310	2630.22	0
Working hours	40.89	40.00	1	99	11.96	40

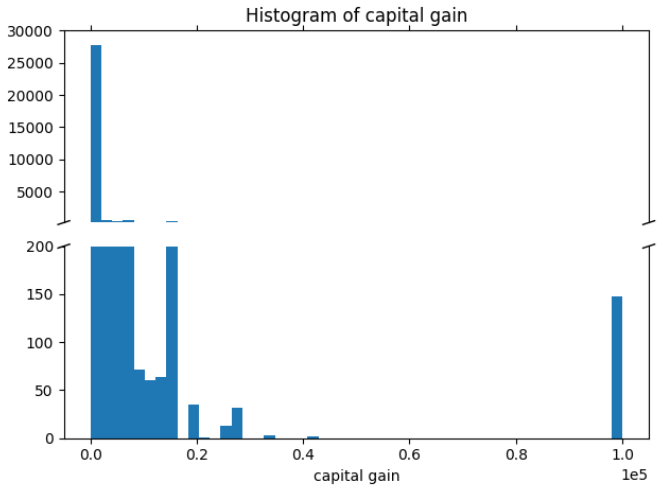
Table 1: Summary statistics of continuous attributes.

Dataset contains another seven nominal/binary attributes: sex, native country, race, work class, occupation, marital status, relationship.

Sex is either male/female. Native country originally contained ten unique values. Race attribute contains five unique values. Work class contains seven unique values. Occupation is the largest category with fourteen unique values. Marital status contains seven unique attributes. Relationship contains five unique values, after joining wife/husband into one category.

Section 3: DATA VISUALIZATION and PRINCIPAL COMPONENT ANALYSIS

3.1 ISSUES WITH OUTLIERS



We were able to identify 148 outliers in capital gain attribute. All 148 data points had capital gain of 99,999. Since next largest capital gain is 41,310 with no large gaps in values till zero, we can confirm all 148 observations as outliers. As a next step of data processing, we dropped all 148 rows of data.

Figure 3: Outliers in capital gain attribute.

3.2 NORMAL DISTRIBUTION

None of the continuous attributes appear to be normal distributed, which in most cases, can be easily explained. Age is not normal distributed in population. Education number is ordinal attribute, so we also do not expect it to be normal distributed. Working hours are also not normal distributed because usual working week contains 40 working hours. Lastly capital gain/loss do not appear to be normal distributed, mostly because majority of people have no capital gain/loss.

3.3 CORRELATION ANALISYS

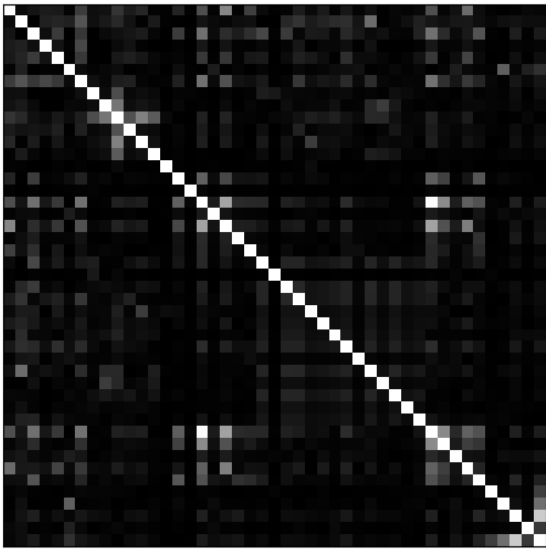


Figure 4: Absolute value of correlation matrix after 1 out of K encoding.

By far the biggest correlation is between married civil spouse and married with correlation of 0.98. We have a lot of strong correlations above absolute value of 0.4 between marital-status and relationship:

- Married civil spouse, married: 0.98
- Married civil spouse, not in family: -0.55
- Never married, married: -0.64
- Never married, own child: 0.50

Figure 5 confirms correlation between relationship and marital status.

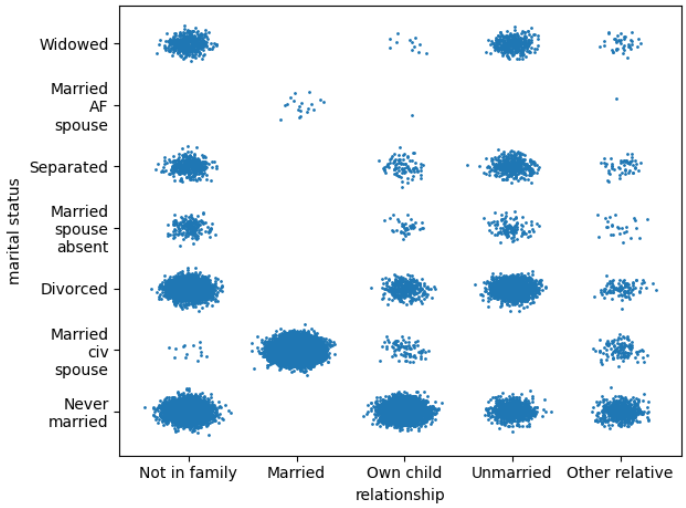


Figure 5: All observations of marital status and relationships plotted against each other.

Another interesting correlation are between attributes in one category:

- Marital status: married civil spouse, never married: -0.64
- Relationship: married, not in family: -0.54
- Work class: Local government, private: -0.46
- Work class: private, self-employed not incorporated: -0.51
- Race: Asian-Pac-Islander, White: -0.43
- Race: Black, White: -0.80

These can be easily explained by the fact than in 1 out of K encoding, one attribute means that all other are not present, meaning negative correlation. The most useful correlations (over 0.4 in absolute values) are between distinct categories:

- Age, never married: -0.52
- Age, own child: -0.42
- Education, professional specialty: 0.42
- Sex, married civil spouse: -0.44
- Sex, married: -0.44
- Result, married civil spouse: 0.44
- Result, married: 0.45

Some of these can be easily explained as for example age and never married, but correlation between sex and relationship status is surprising. We can also note correlation between married and category we aim to predict.

3.4 PREDICTION FEASIBILITY

Our classification problem is feasible to some extent. From all attributes we can observe some bias toward one classification category as we can see in appendix A. On the other hand, it is unlikely to get low error rate because none of the attributes seems to perfectly discriminate one of the classification categories.

Since 75% of data are people who earn under 50k, we could get 25% error rate by simply always predicting lower earnings. Previous attempts by others achieved error rate as low as 15%. We should thus aim for error under 20%.

3.5 VARIATION EXPLAINED

For principal component analysis, we first normalized all data. After applying 1 out of K encoding data contains 44 columns. We can achieve 0.9 variance explained with 30 principal components and 0.95 can be achieved using 33 components. You can see variation explained and cumulative variation explained for each component in Figure B.1 in appendix B.

3.6 PRINCIPAL COMPONENT VISUALIZATION

We have visualised weight of all attributes to four principal components in Figure B.2 in appendix B.

3.7 DATA PROJECTION ONTO THE PRINCIPAL COMPONENTS

We projected our dataset into first principal components and plotted first three components' projections in Figure B.3 in appendix B.

Section 4: SUMMARY DISCUSSION

Thanks to data visualization and correlation analysis we can see that there are clear relationships between some of the attributes in the dataset, of course, we ust keep in mind that the number of instances which presents an annual income of less than 50k is three times higher than the amount of people who earn more. We have also to notice that in some attributes a particular value is dominant with respect to the others, such as the US origin for native country attribute.

More specifically we have focused on the attributes which allow us to visualize a relationship with the total annual income attribute that we want to predict. Looking at the graphs in figure A.1, appendix A, we can infer for example that if someone works less than 30 hours a week it is certain that his/her annual income will be under 50k. Same phenomenon happens if we look at the years of education, where we can see that having less than 5-6 years of education implies that the income is lower than 50k.

Another relevant relationship is visualized in figure A.5, where we can clearly see that a high value for capital gain implies that the annual income will be this time higher than 50k.

Other less notable relationships can still be inferred looking at other graphs.

Given what was previously said we are quite confident that it will be possible to obtain a satisfactory performance in the classification task aiming at identifying if the annual income is more or less than 50 thousand dollars, given the others attribute.

Exam problem solutions

Question 1: D

Given that attributes y is ordinal and all attributes from x_2 to x_7 are ratio, we must state the data type of attribute 1 which corresponds to the 30-minute interval time of the day, we say that it is interval since of we look at the set of observation of one specific day it makes sense to calculate the difference between for example 2 and 4 and 7 and 9 and in both case this difference is something in between 30 and 60 minutes.

Question 2: A

If $p = \infty$, by definition, of p_{norm} we have just to choose the max component of the distance vector, in this case, the distance vector is just x14-x18 and so, the max component is 7.

Question 3: A

We can calculate variance explained by single component as component singular value squared over sum of squares of all singular values. We can sum up first four explanations to 0.87 which is more than 0.8.

Question 4: D

A is incorrect because low x_1 would result into negative projection, **B** is incorrect because low x_3 would result in negative projection, **C** is incorrect because low x_1 would result in positive projection. We are left with option **D**, all positive weights have high values, and all negative weights have small values, thus projection is indeed positive.

Question 5: A

First, we look at the equal words that exist between the two texts (intersection) and this will be the numerator: “the”, “words”; we have 2.

Then we look at all the other words (plus these two) and this will be the denominator: “the”, “bag”, “of”, “words”, “representation”, “becomes”, “less”, “parsimonious”, “if”, “we”, “do”, “not”, “stem”; we have 13.

Jaccard similarity of s1 and s2 is 2/13=0,153846.

Question 6: B

Given $y = 2$ we don’t care about x_7 and we only look at the probability of x_2 that needs to be $x_2 = 0$.

So, summing up the two possibilities we have as result:

$$p(x_2 = 0, x_7 = 0 \mid y = 2) + p(x_2 = 0, x_7 = 1 \mid y = 2) = 0.81 + 0.03 = 0.84.$$

Appendix A

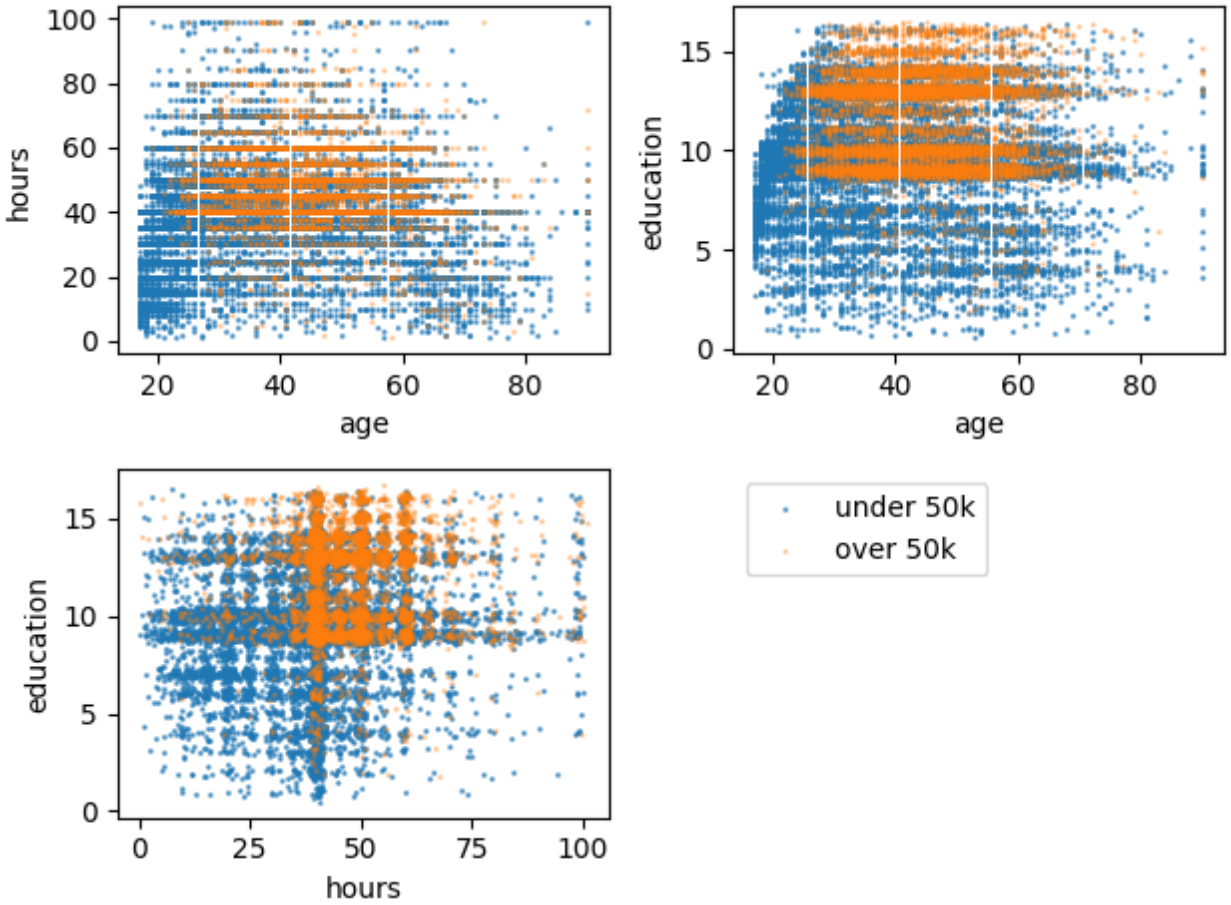


Figure A.1: All observations plotted against three continuous attributes plotted. Education number has random noise added to separate observations.

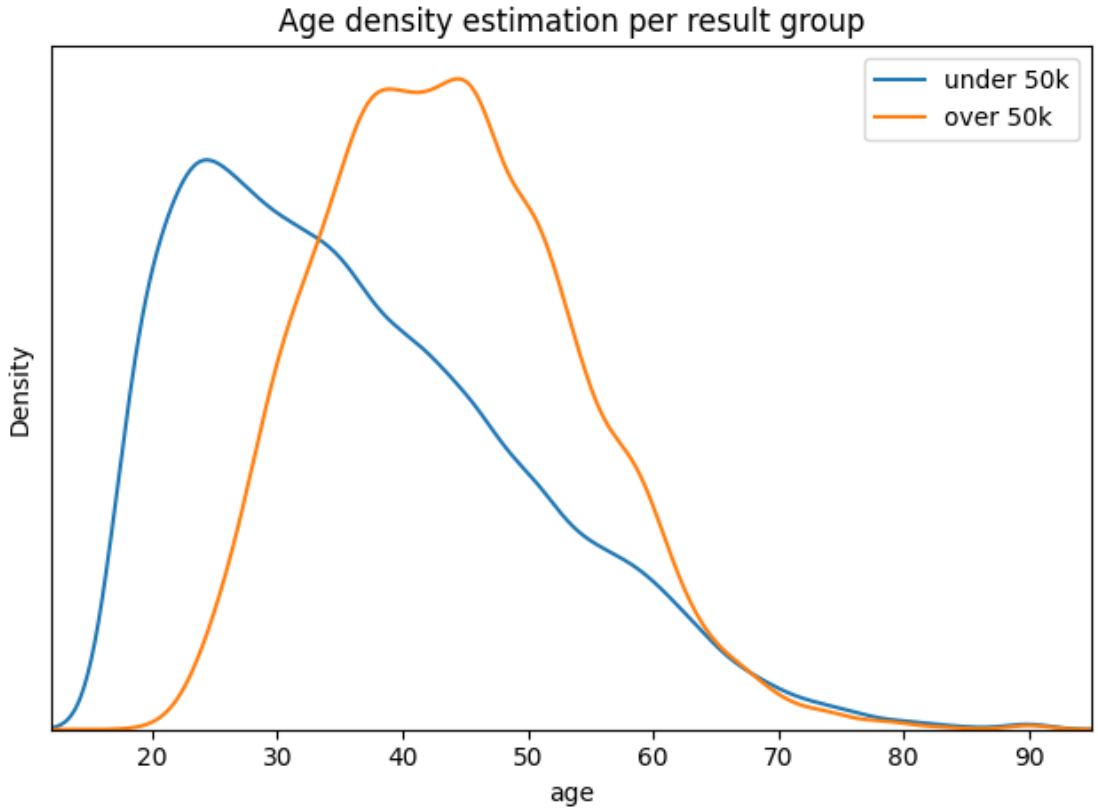


Figure A.2: Age density shows that people who earn more tend to be also older.

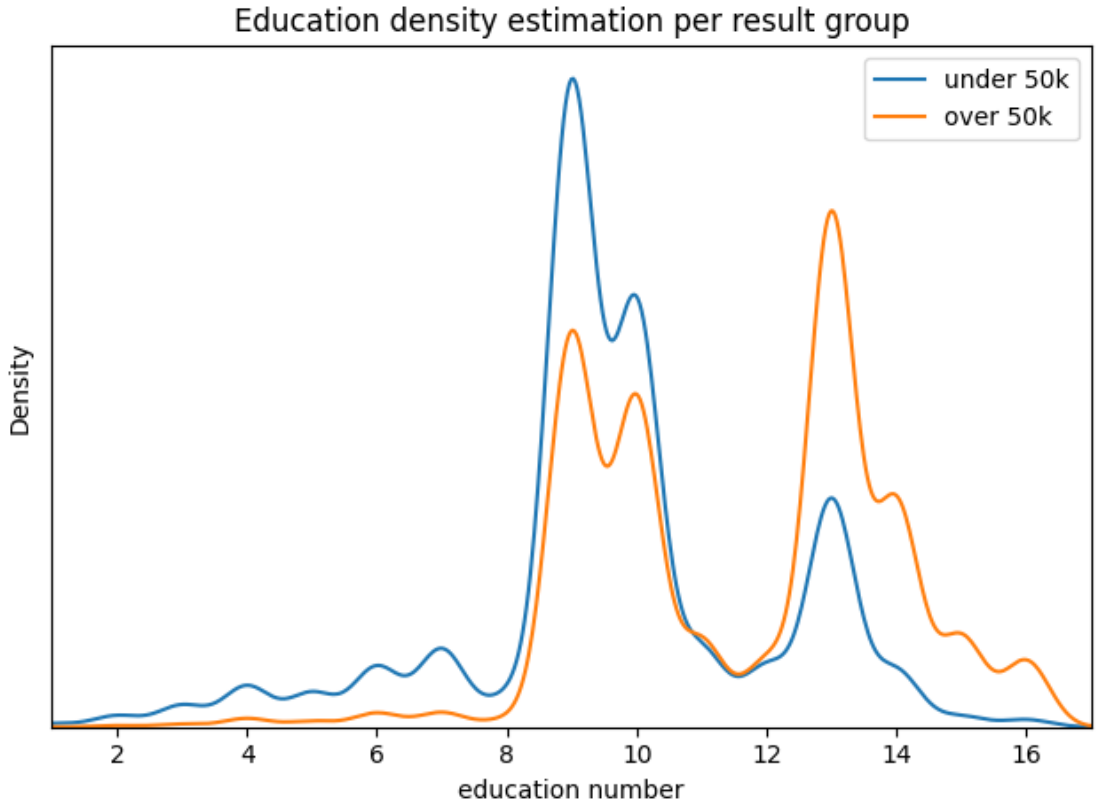


Figure A.3: Education density shows that higher education increases chances of higher salary.

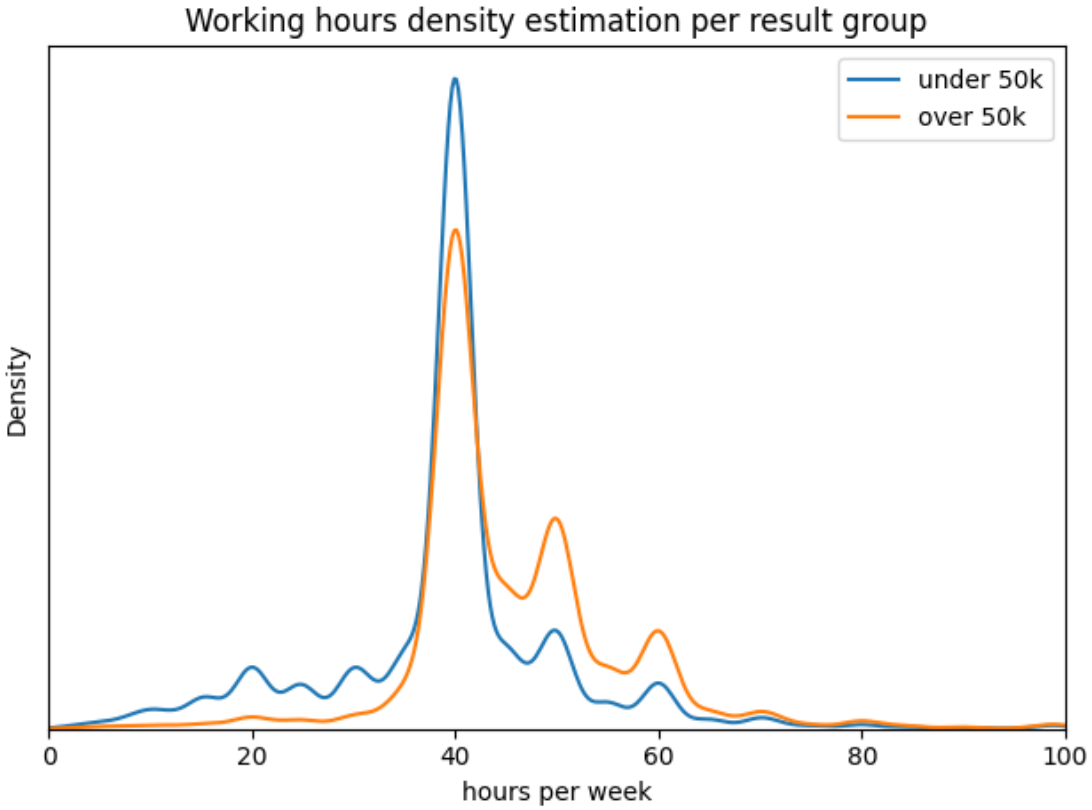


Figure A.4: Working hours shows correlation between working hours and salary.

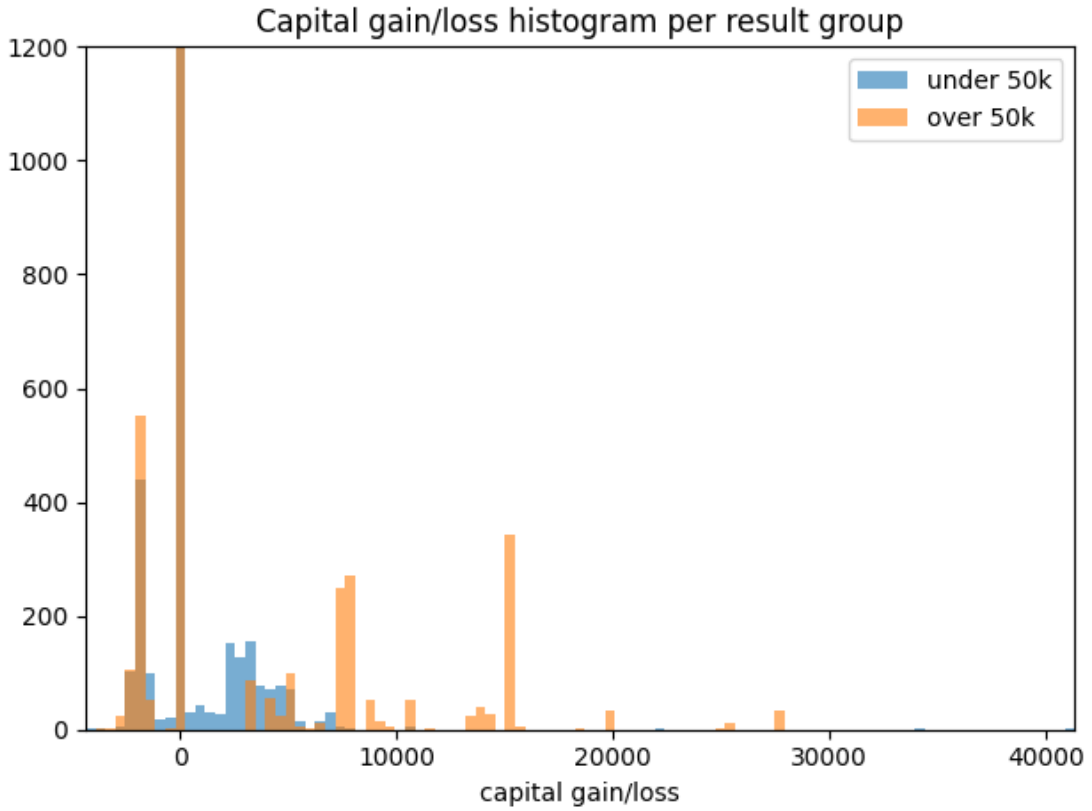


Figure A.5: We can notice that, even though we have less observations with smaller salaries they have bigger spread.

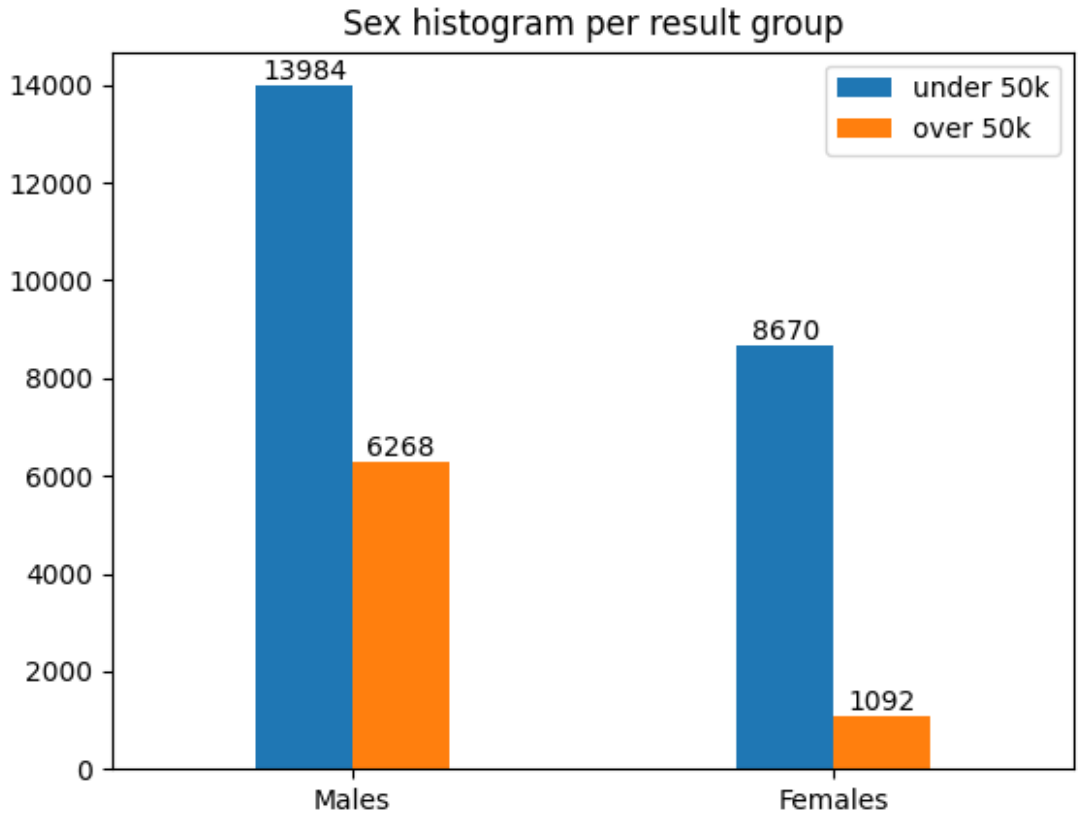


Figure A.6: We can notice that females have lower representation and that they are paid less.

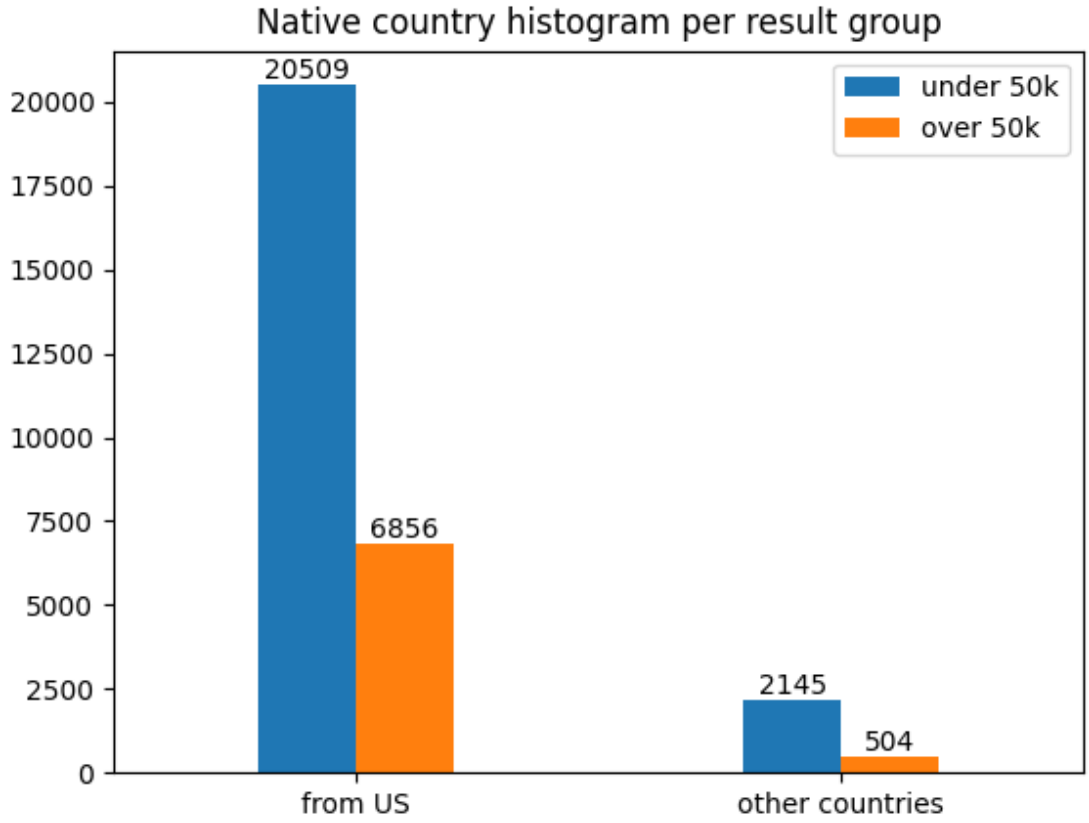


Figure A.7: We can notice that other countries are still in minority and are less paid.

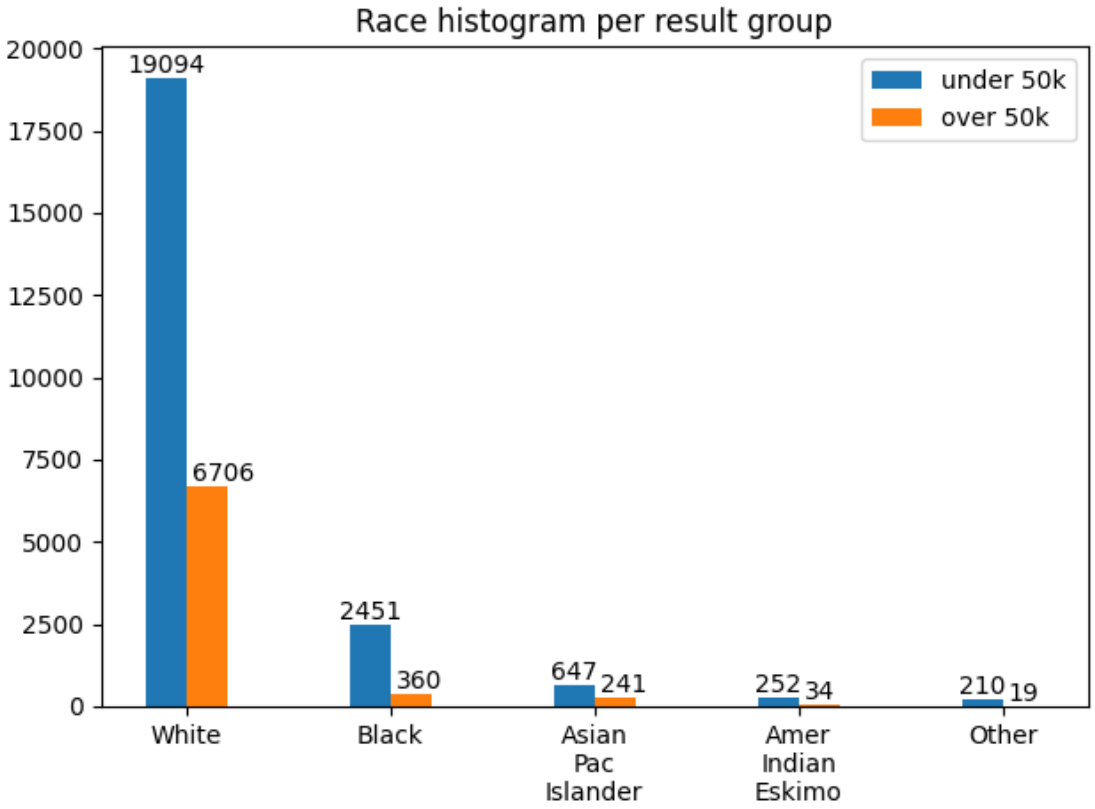


Figure A.8: We can notice that races other than white are less paid.

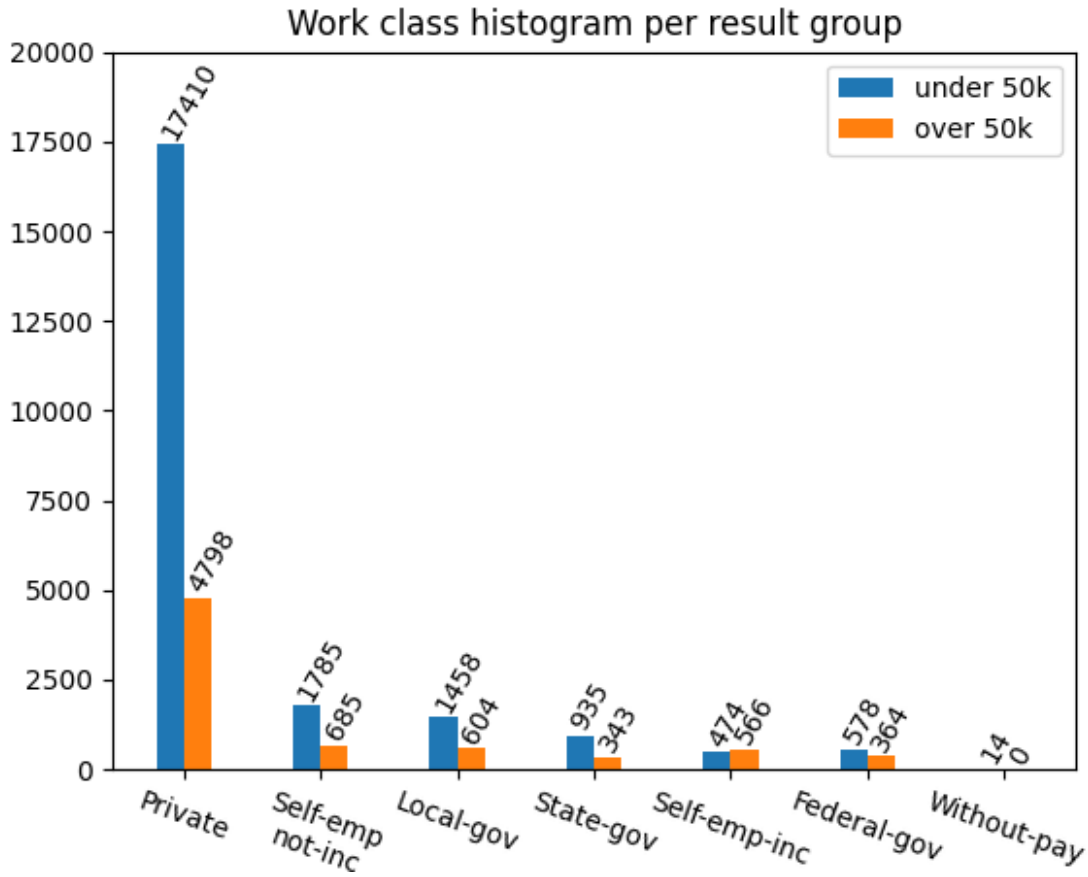


Figure A.9: We can notice that self-employed people are only category with more people earning over $50k than less.

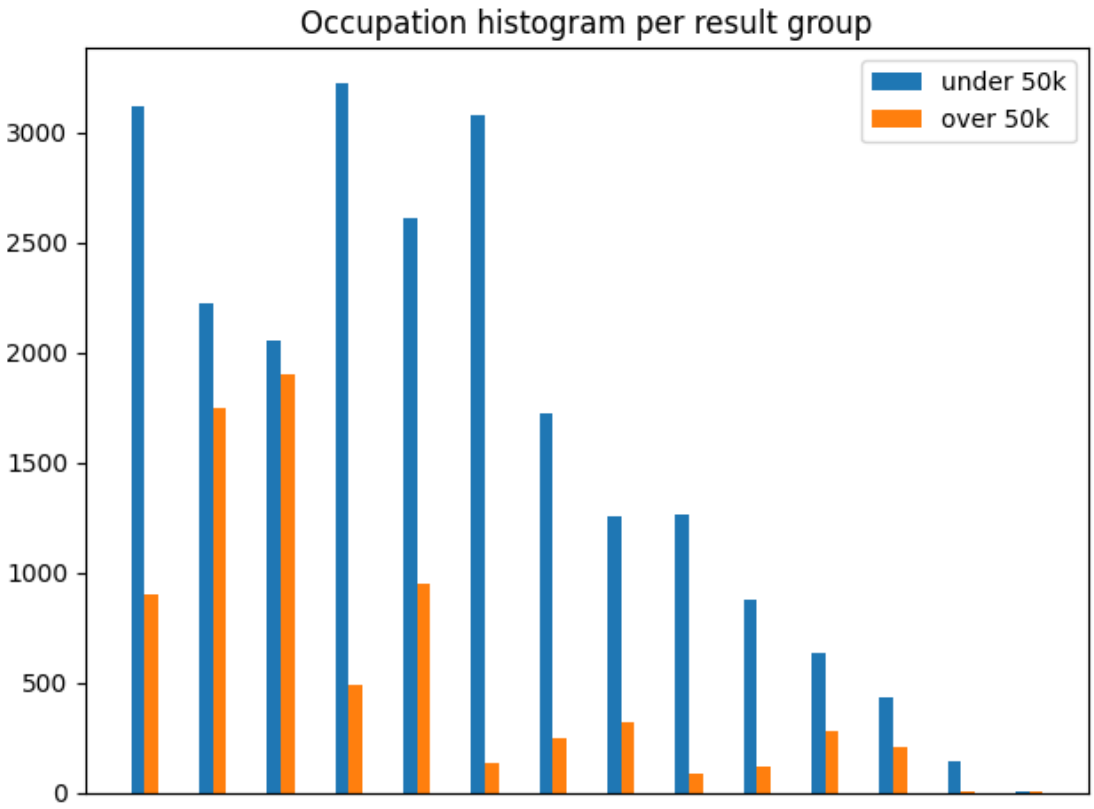


Figure A.10: Histogram of occupation, see table A.1 below for values and category names.

Occupation statistics

	Under 50k	Over 50k	Total
Craft repair	3122	900	4022
Professional specialty	2227	1746	3973
Executive/Managerial	2055	1899	3954
Administrative/Clerical	3223	492	3715
Sales	2614	946	3560
Other services	3080	130	3210
Machine operator/Inspector	1721	244	1965
Transportation	1253	318	1571
Handlers/Cleaners	1267	82	1349
Farming/Fishing	874	115	989
Tech support	634	277	911
Protective services	434	209	643
Private house services	142	1	143
Armed forces	8	1	9

Table A.1: All occupations and observations.

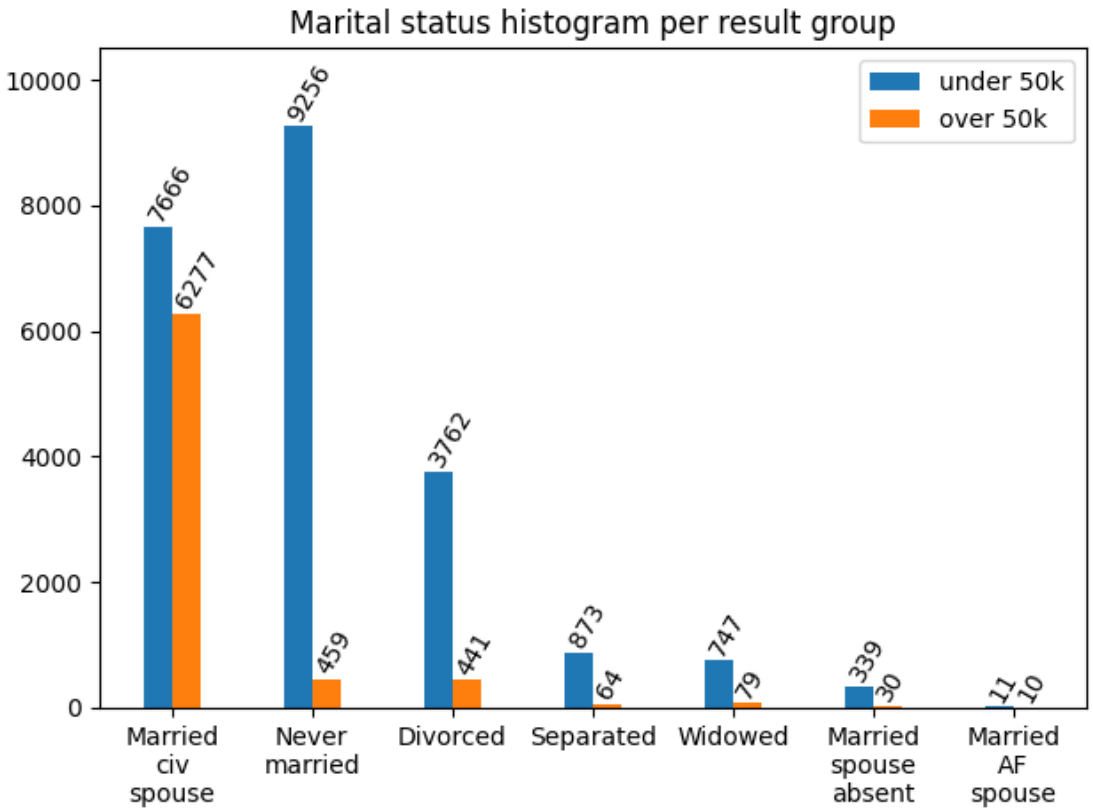


Figure A.11: Clear correlation of result and marriage.

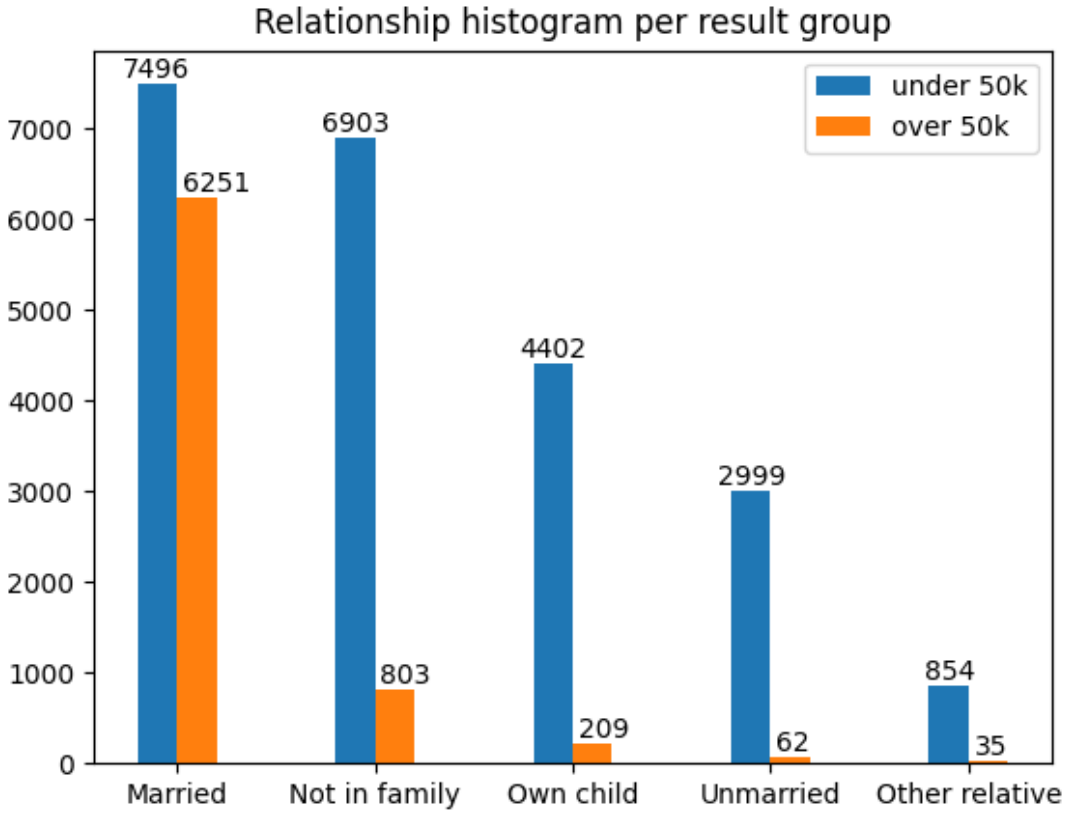


Figure A.12: Clear correlation of result and marriage.

Appendix B

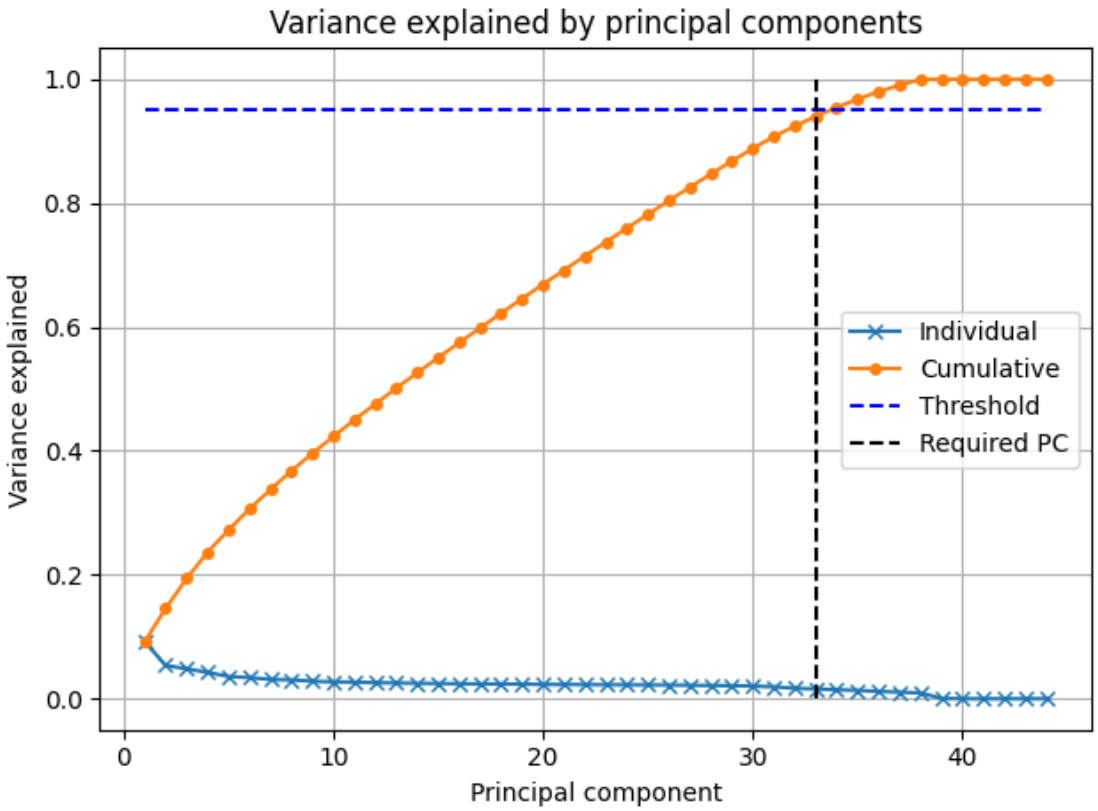


Figure B.1: Variance and cumulative variance explained by principal components.

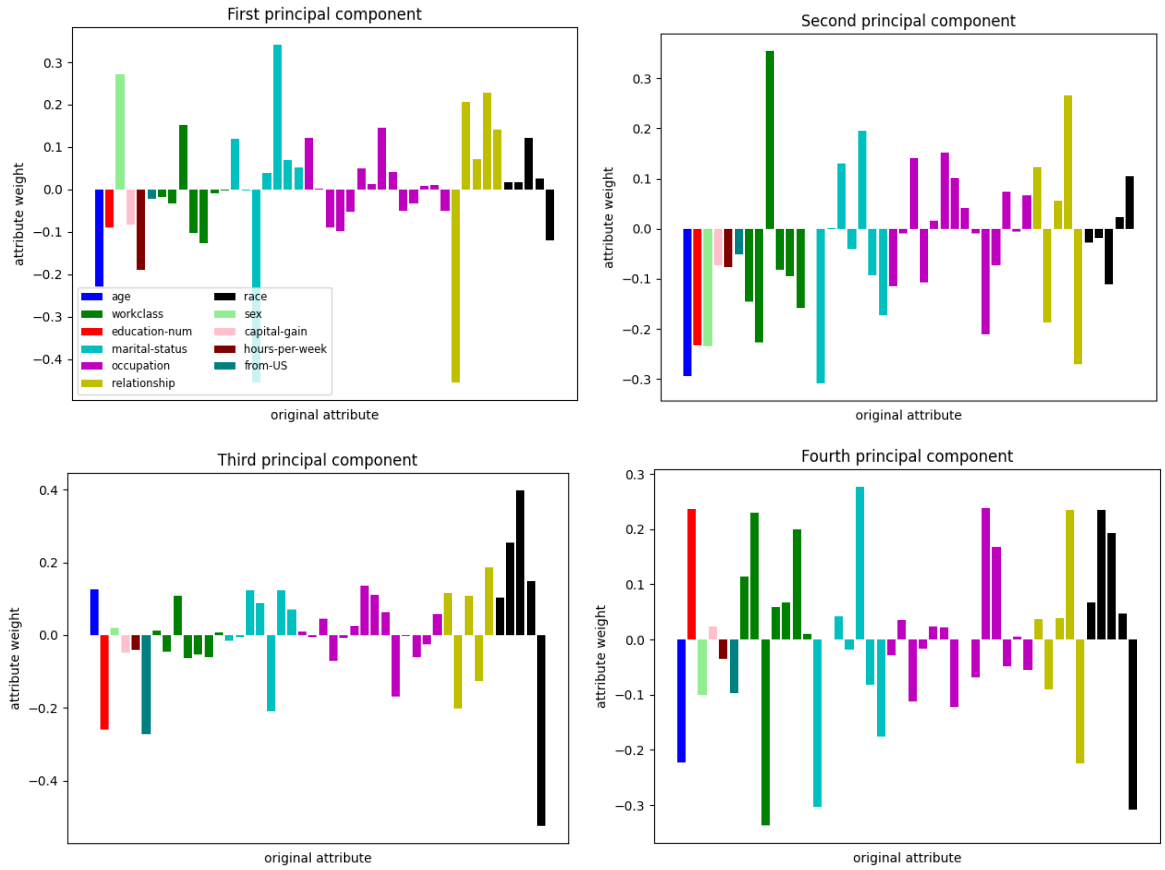


Figure B.2: Weight of attributes towards principal components. First principal component might represent young females who never married.

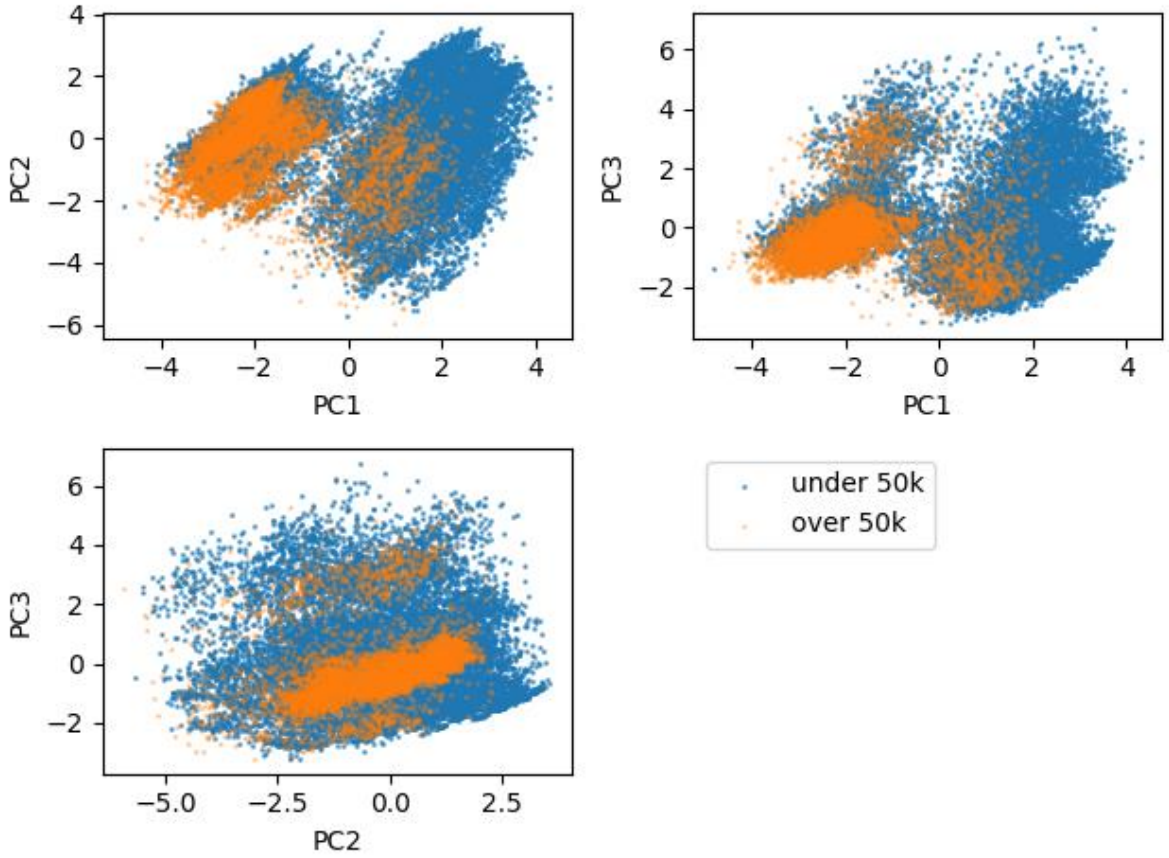


Figure B.3: Dataset projected into principal components.