Census Income Study

Data Cleaning and EDA

Andrew Graham – Fall 2022

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Contents

Analysis Overview

- Summary and Goals
- Data Overview
  - Meta Data
  - Train and Test
  - Input
  - Output

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL
OF ENGINEERING & COMPUTER SCIENCE

Analysis Overview

Summary and Goals

The purpose of this data is to create a prediction model to determine the income of an individual based on given census data. The goal of this data has been binned into having a salary of > 50k and < 50k. The data is from the US census bureau.

In this section of the study data cleaning and exploratory data analysis was performed. Analysis and recommendation will be provided to assist the next stage of Data Transformation and Model Creation.

Analysis Overview

Data Overview - Meta

- 2 Files were provided containing data split into training and testing data
- A meta file was provided containing column information, such as: name, description, and types (nominal/continuous)
  - This file contained a few tables which were combined using a fuzzy matching algorithm and then the resulting information allowed the data to be labeled
  - The meta document also contained the unique values from the data which was used to assist in data cleaning efforts
- Instructions were given in the document to drop a column (Instance Weight), so those instructions were followed.
- Data consist of 41features (40 input and 1 output) with 299,285 records

Analysis Overview

Train and Test

- Data was split 1/3 into test and 2/3 into train
  - 199,523 training records, 99762 test records
- The Kolmogorov-Smirnov distance was used to test the balance of the sets: the closer to 0 the more similar the sets. Each variable rom one section to the next was tested and the max value for each set was taken
  - Maximum distance before cleaning: 0.004
  - Maximum distance after cleaning: 0.018
  - Test and Train splits are balanced and were not significantly affected by the cleaning efforts
- The training and test sets were merged for the purposes of Data pre-Processing

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL
OF ENGINEERING & COMPUTER SCIENCE

Analysis Overview

Data Overview

- 40 Input features
  - 7 numerical/continuous
  - 33 nominal
- The 7 continuous feature did not have missing data although a majority had 0 values
- 14 nominal features had complete data
- 19 features were incomplete
  - 14 features were dropped for having over 30% missing data that could not be imputed
- 6% of the data was dropped for having missing values
- Target(Salary) is unbalance with more than 90% in the <50k category

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Features Table

	name	num_uniq	type	long_name_t1
0	AAGE	91	continuous	age
1	ACLSWKR	9	nominal	class of worker
2	ADTIND	52	nominal	industry code
3	ADTOCC	47	nominal	occupation code
4	AHGA	17	nominal	education
5	AHRSPAY	1240	continuous	wage per hour
6	AHSCOL	3	nominal	enrolled in edu inst last wk
7	AMARITL	7	nominal	marital status
8	AMJIND	24	nominal	major industry code
9	AMJOCC	15	nominal	major occupation code
10	ARACE	5	nominal	mace
11	AREORGN	10	nominal	hispanic Origin
12	ASEX	2	nominal	sex
13	AUNMEM	3	nominal	member of a labor union
14	AUNTYPE	6	nominal	reason for unemployment
15	AWKSTAT	8	nominal	full or part time employment stat
16	CAPGAIN	132	continuous	capital gains
17	CAPLOSS	113	continuous	capital losses
18	DIVVAL	1478	continuous	divdends from stocks
19	FILESTAT	6	nominal	tax filer statu

	name	num_uniq	type	long_name_t1
21	GRINST	51	nominal	state of previous residence
22	HHDFMX	38	nominal	detailed household and family stat
23	HHDREL	8	nominal	detailed household summary in household
40	INST	0	ignore	Instance Weight
24	MIGMTR1	10	nominal	migration code-change in msa
25	MIGMTR3	9	nominal	migration code-change in reg
26	MIGMTR4	10	nominal	migration code-move within reg
27	MIGSAME	3	nominal	live in this house 1 year ago
28	MIGSUN	4	nominal	migration prev res in sunbelt
29	NOEMP	7	continuous	num persons worked for employer
30	PARENT	5	nominal	family members under 18
31	PEFNTVTY	43	nominal	country of birth father
32	PEMNTVTY	43	nominal	country of birth mother
33	PENATVTY	43	nominal	country of birth self
34	PRCITSHP	5	nominal	citizenship
35	SEOTR	3	nominal	own business or self employed
36	VETQVA	3	nominal	fill inc questionnaire for veteran's admin
37	VETYN	3	nominal	veterans benefits
38	WKSWORK	53	continuous	weeks worked in year
39	YEAR	2	nominal	NaN
41	ZA_TARGET	2	target	Target +- 50k
42	ZZ_SPLIT	2	reference	Test or Train

Data Cleaning

- Data Loading/ Overall
- Numerical Features
- Nominal Features

Data Cleaning

Data Loading/ Overall

- Target feature ZA_TARGET converted to a binary[1,0]
- Train and Test data merged with label to differentiate (ZZ_SPLIT)
- Feature: Instance Weight dropped as per instructions
- 67525 duplicate records were removed

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Data Cleaning

Numerical Features

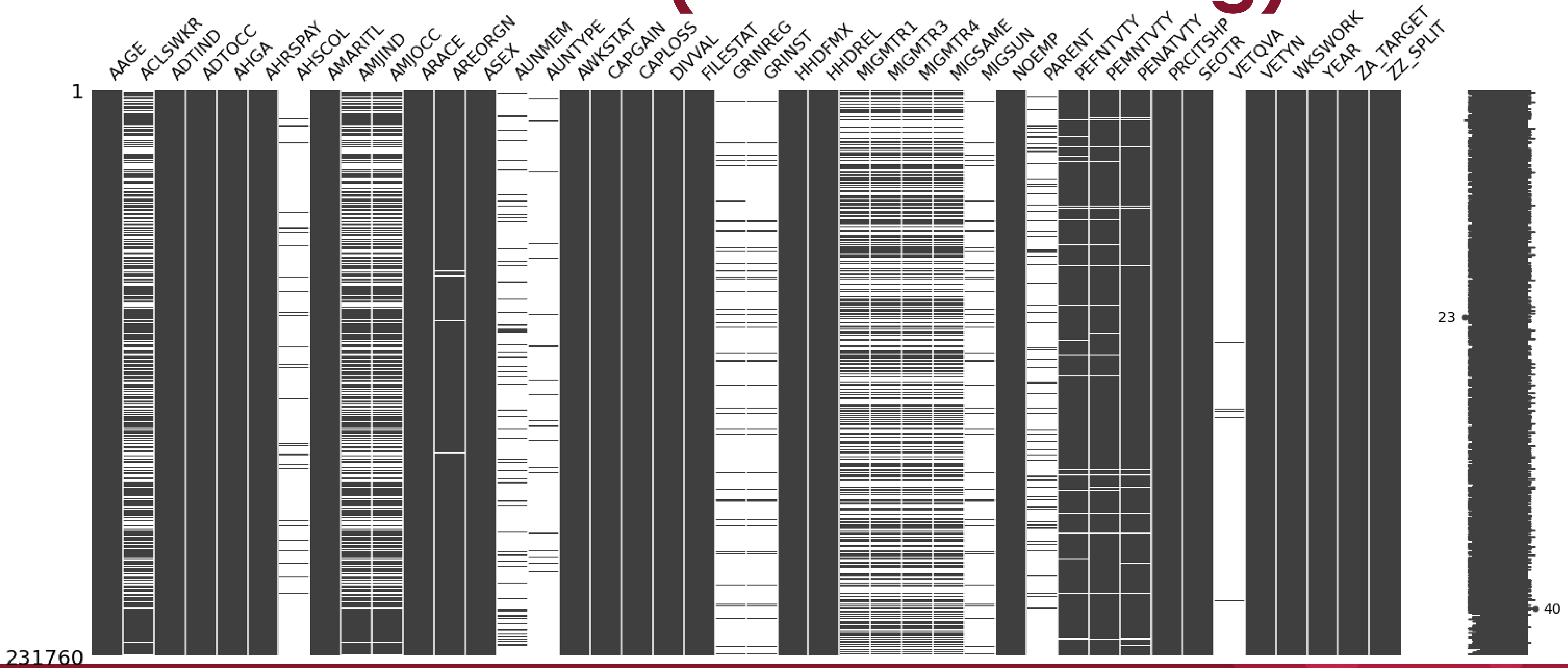
- All numerical features had data
- Values of 9999 were looked at for possible NA, however they seemed to represent a ceiling rather than missing values
- Values of 0 were reasonable, however many of them could have been missing data but there was no way to differentiate, so they were left alone
- Column types verified to be numerical

Data Cleaning

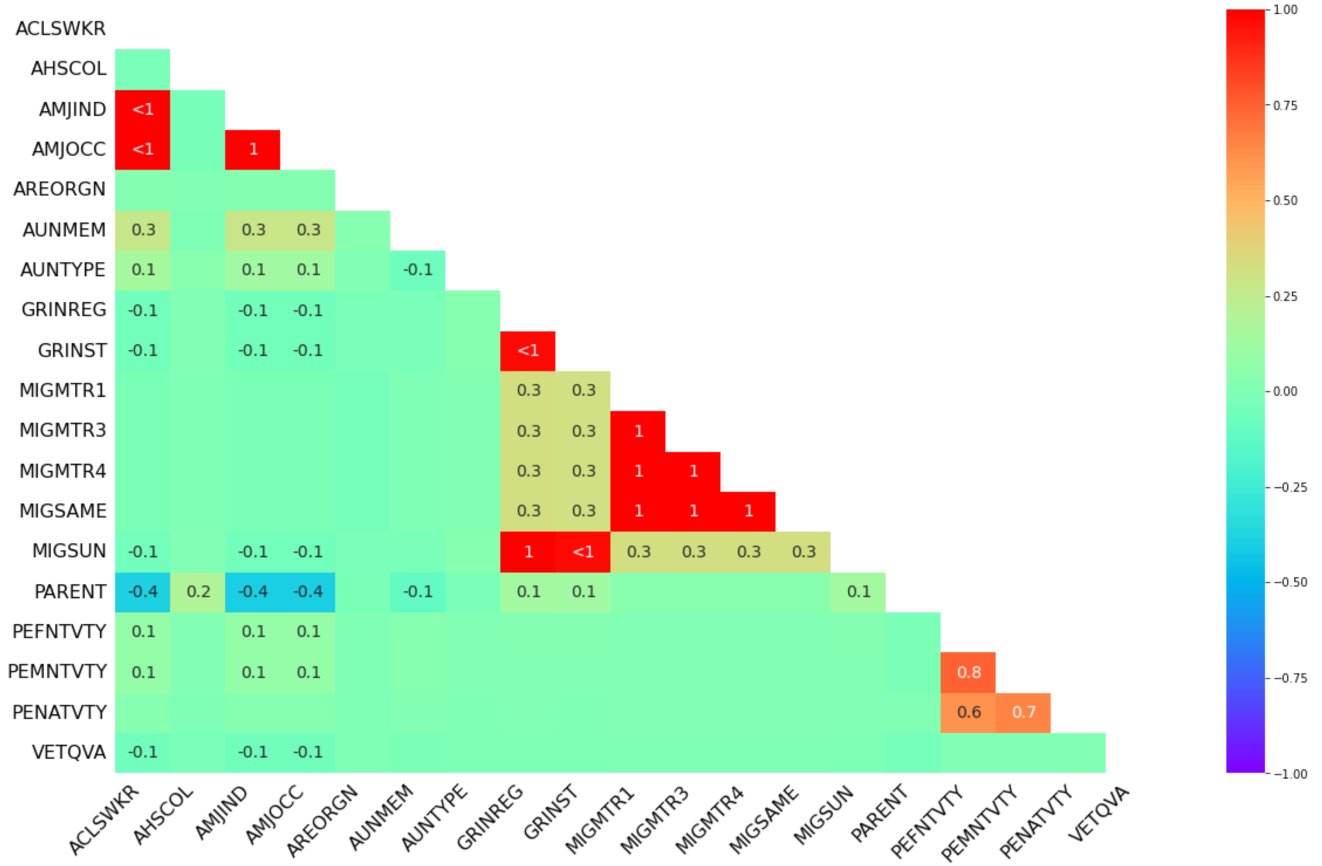
Nominal Features

- Nominal data was converted to lower case and whitespace was trimmed
- The following values were found and converted to NA:
  - 'not in universe','?','do not know','na','not in universe under 1 year old', 'not in universe or children'
- 19 features were found to contain NA values
  - Missing percentages were calculated those over 30% missing were considered for removal
  - Distribution of those were checked with train and test data and found to be randomly distributed between both
  - Correlation between NA containing features were checked and none of the removed features correlated with non- removed features
- Following the four remaining features had missing values totaling 5.9% of the total data and were balanced between train and test so they were dropped as well.

NA Distribution (white is missing)



NA Correlation



- From this and the prior chartwe see that most columns look to be MCAR/MAR with the following exceptions...
- -AMJOCC and AMJIND (Major Industry Code and Major Occupation Code)
- This makes sense as they seem to be referencing the same thing
- -GRINREG and GRINST (region and state of previous residence)
- This makes sense as one is dependent of the other.
- -MIGMTR1,MIGMTR3,MIGMTR4 (Migration code Data)
- -PEFNTVTY and PEMNTVTY (Birth pace of Parents)

NA Feature Dropping

Since the missingness looks to be random and using a threshold of 30%. The following features should be dropped:

Feature	# Missing	Missingness
AMJIND	84080	0.362789
ACLSWKR	83508	0.360321
AMJOCC	84080	0.362789
MIGSAME	114346	0.493381
MIGMTR4	114346	0.493381
MIGMTR3	114346	0.493381
MIGMTR1	114346	0.493381
AUNMEM	203225	0.876877
PARENT	203808	0.879392
GRINREG	208751	0.900721
MIGSUN	208751	0.900721
GRINST	209776	0.905143
AHSCOL	215546	0.930040
AUNTYPE	222633	0.960619
VETQVA	228779	0.987138

With the following to be kept:

Feature	# Missing	Missingness
AREORGN	1672	0.007214
PENATVTY	5057	0.021820
PEMNTVTY	8779	0.037880
PEFNTVTY	9690	0.041810

Exploratory Analysis

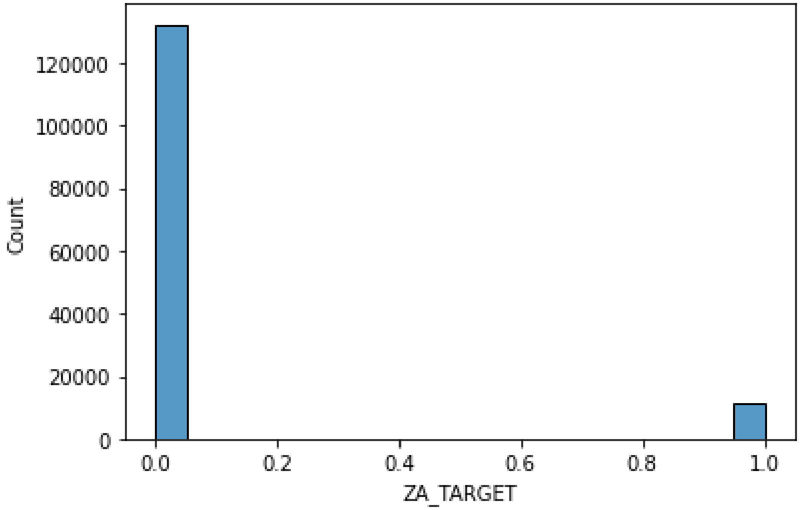
- Dataset
- Target Variable
- Numerical Features
- Nominal Features

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL
OF ENGINEERING & COMPUTER SCIENCE

Exploratory Analysis

Target Variable

- Target Feature is binned at <50k and >50k
- This was converted to 1 for >50 and 0 for <50
- Data was clean and had no missing values



The target variable is highly unbalanced, and this will have to be considered for model creation.

Exploratory Analysis

Numerical Features

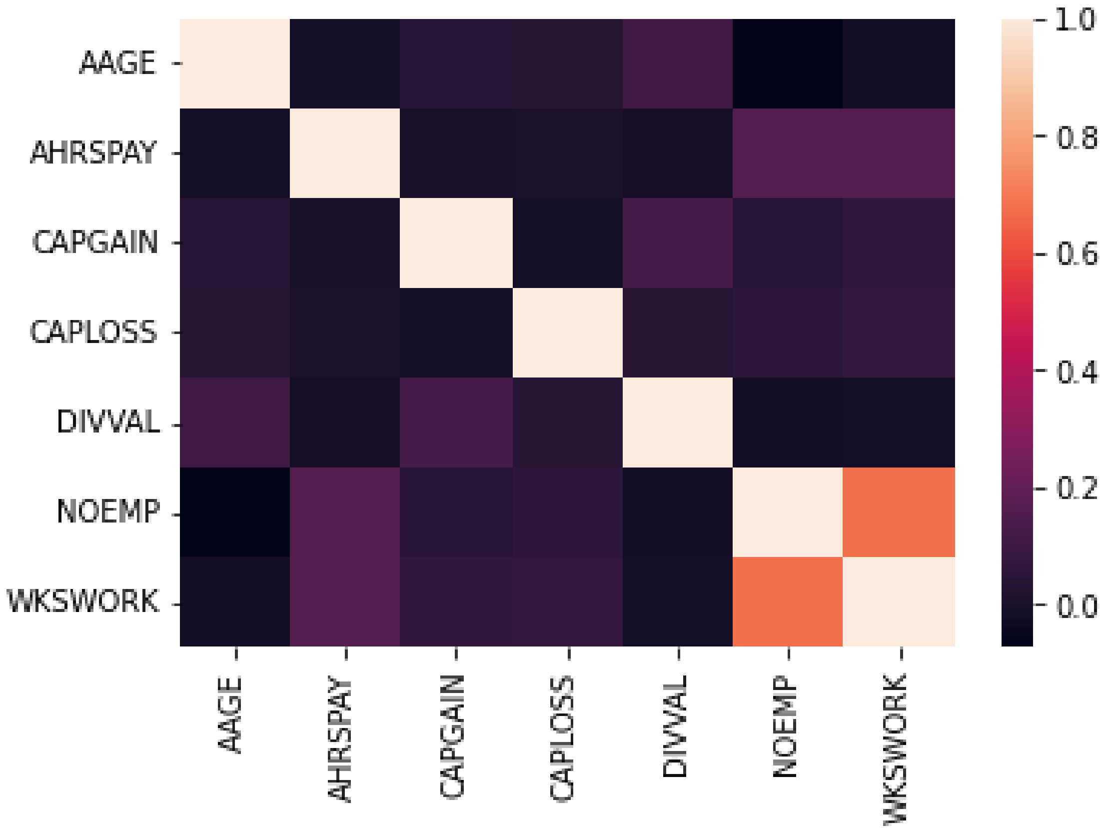
	mean	median	min	max	var	std	skew
AAGE	34.538998	33.0	0.0	90.0	4.981140e+02	22.318468	0.372785
AHRSPAY	55.105027	0.0	0.0	9999.0	7.471515e+04	273.340729	8.878780
CAPGAIN	431.742176	0.0	0.0	99999.0	2.181608e+07	4670.768536	19.090569
CAPLOSS	36.849010	0.0	0.0	4608.0	7.278652e+04	269.789771	7.685924
DIVVAL	195.851259	0.0	0.0	99999.0	3.755251e+06	1937.847082	27.144287
NOEMP	1.956172	1.0	0.0	6.0	5.592548e+00	2.364857	0.752317
WKSWORK	23.178375	8.0	0.0	52.0	5.955560e+02	24.404016	0.210018

- 7 continuous numerical features.
- AHRSPAY (Wage per hour), CAPGAIN, CAPLOSS, and DIVAL are all highly right skewed.
- AHRSPAY, CAPGAIN and DIVAL all have ceiling of 9999

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Exploratory Analysis

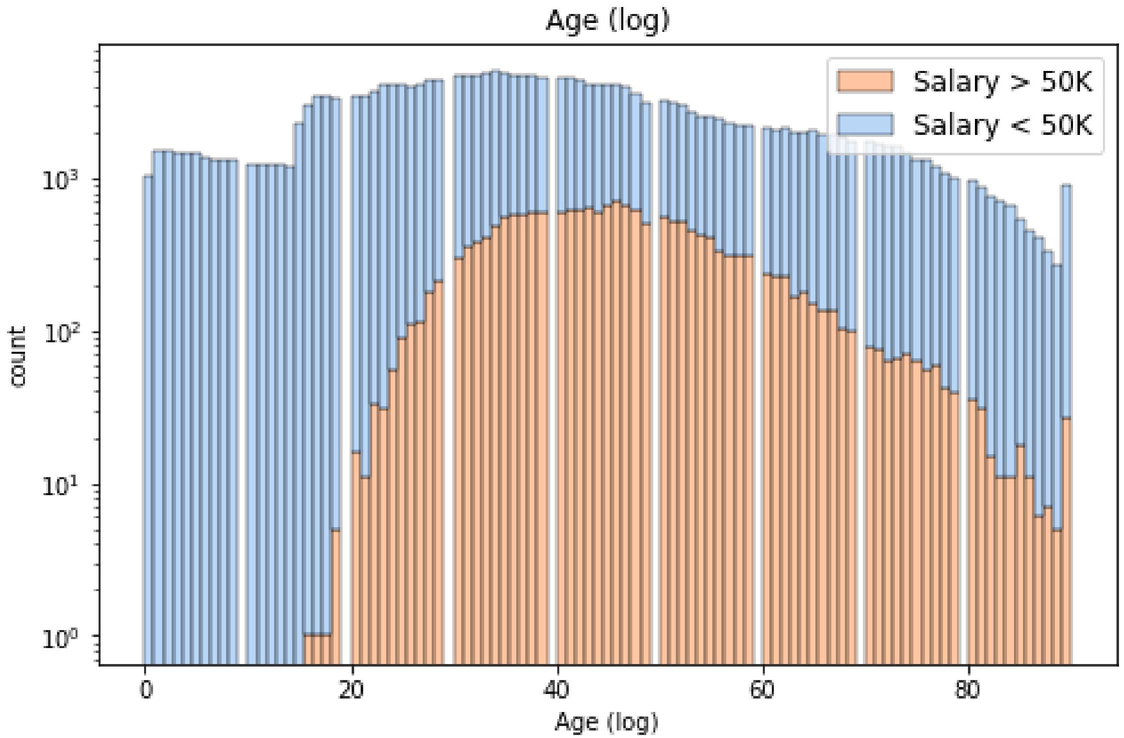
Numerical Features



- Low correlations between most of the variables
- NOEMP and WKSWORK have high correlation
  - This is understandable as someone who has employees likely works a higher amount of weeks

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

AAGE (Age)



*Note: counts showed in log scale do to unbalance data to see what values contained >50k

- Age column seems reasonable.
- 90 years looks like a ceiling value.
- Under 16 falls in line with not having a salary due to US work laws.
- Those with Salary>50k peaks around 35-55.

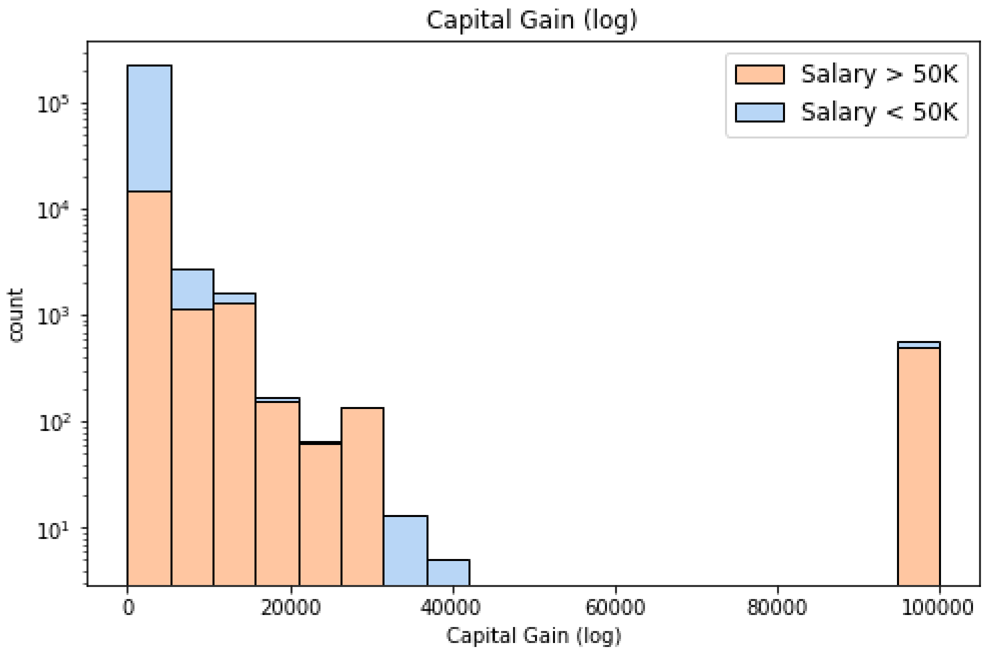
Wage per Hour (AHRSPAY)

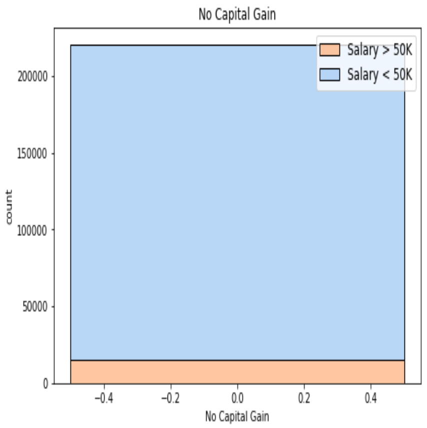
*Note: counts showed in log scale do to unbalance data to see what values contained >50k

Number with No wage: 214874
Percent with No wage: 93%
Minimum wage: 20

- Wage per hour seems suspect. Minimum wage is 20, which is very high for the mid 90's. Also 94% of the data has 0 wages, which indicates many this might be missing data.
- Data does follow a distribution up to about 5000. Over that the data seems incomplete. Possibly at this level income may or may not come from Salary, but other sources. May want consolidate values over 5000.

CAPGAIN (Capital Gains)



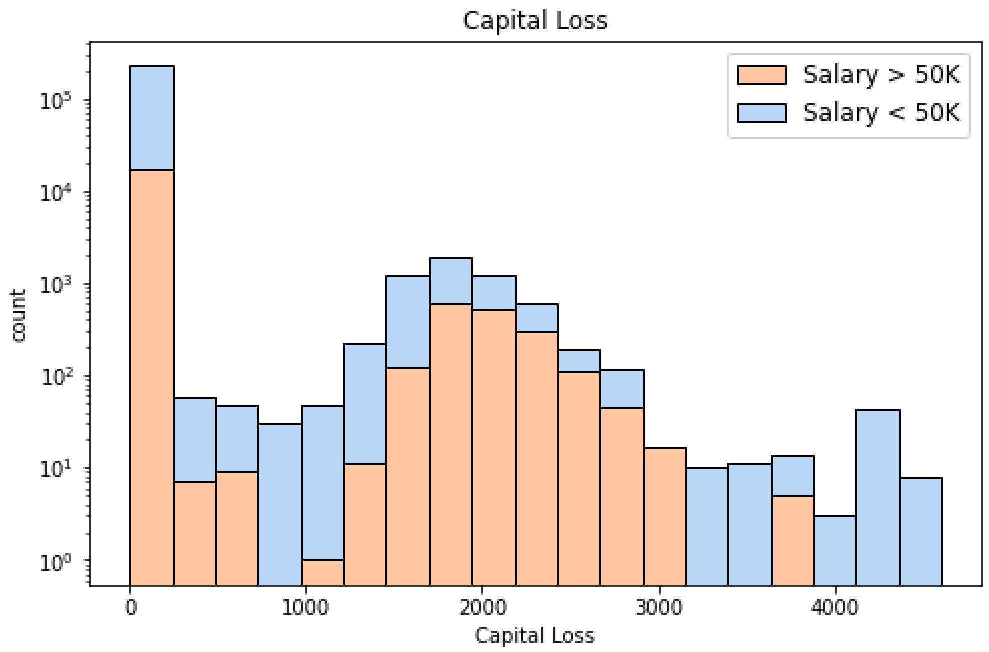


Number with No Capital Gain: 220666
Percent with No Capital Gain: 95%
Number of Capital Gain: 578

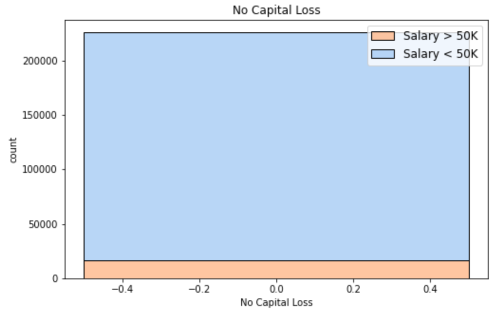
- The amount of capital gain doesn't seem to be that correlated with salary, whether there is capital gains seems to have an effect.
- Most records show No capital Gain.
- Consider switching this to a binary Have/Have no capital Gains.

*Note: counts showed in log scale do to unbalance data to see what values contained >50k

CAPLOSS (Capital Loss)



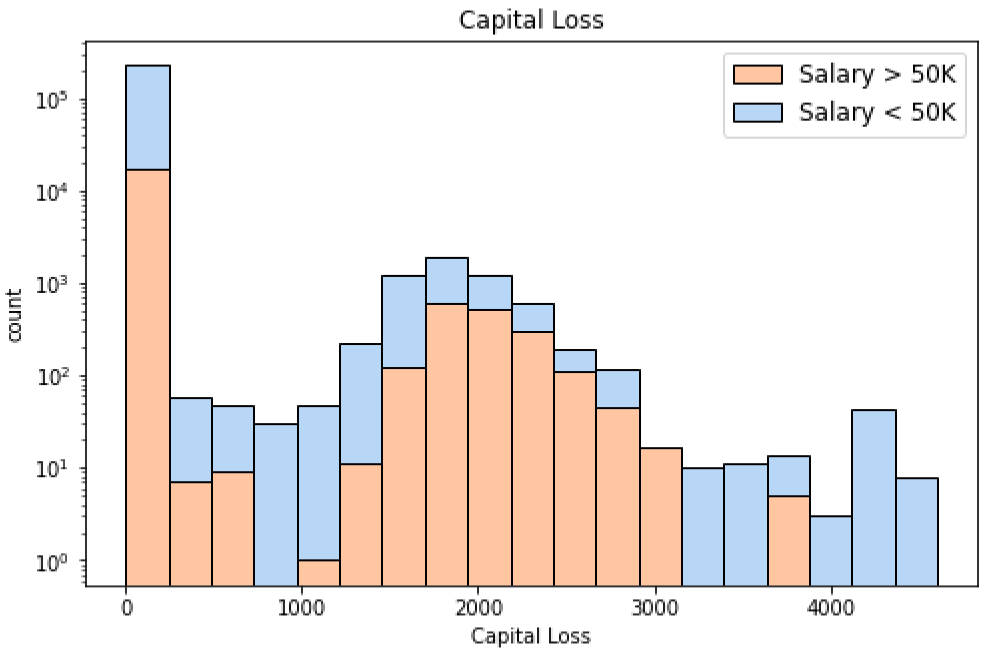
*Note: counts showed in log scale do to unbalance data to see what values contained >50k



Number with No Capital Loss: 212684
Percent with No Capital Loss:97%
Number of Capital Loss: 6

- The amount of capital gain doesn't seem to be that correlated with salary, whether there is capital loss seems to have an effect.
- Most records show No capital Loss.
- Consider switching this to a binary Have/Have no capital Loss.

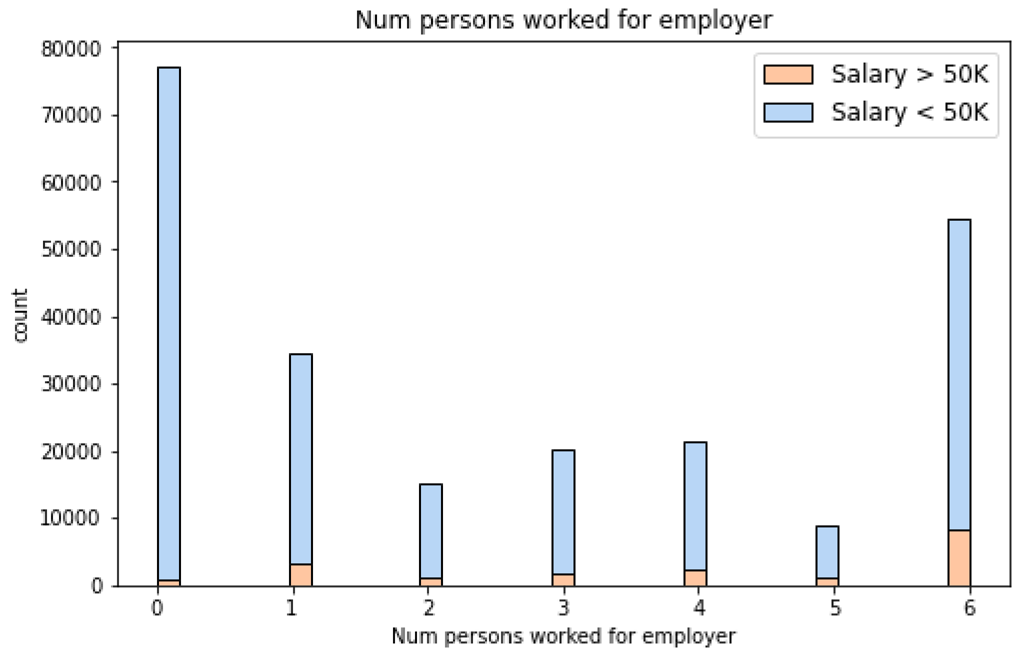
DIVVAL (Dividends from Stocks)



- Same results as Wage and Capital Gains. Possible Binary candidate.

*Note: counts showed in log scale do to unbalance data to see what values contained >50k

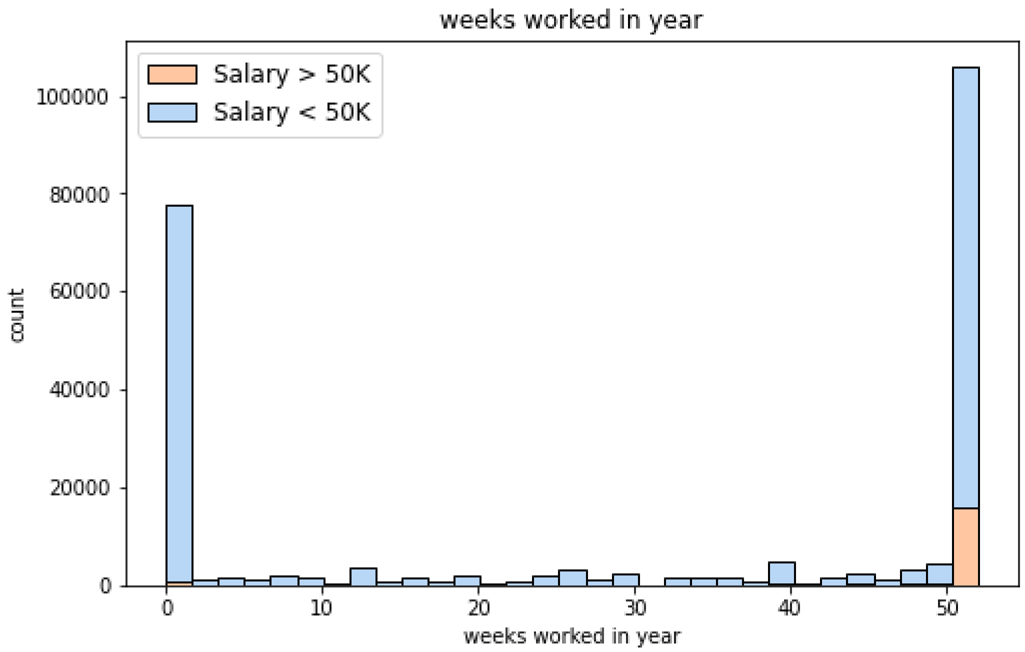
NOEMP (num persons worked for employer)



- Very high imbalance, Vast majority with no person working for employer.

*Note: counts showed in log scale do to unbalance data to see what values contained >50k

WKSWORK (weeks worked in year)



- Majority of values at 0 and 52
- Consider binning this into 0, 1-51, 52

*Note: counts showed in log scale do to unbalance data to see what values contained >50k

Exploratory Analysis

Nominal Features Summary

- Most features can be left as is
- The features with larger number of categories can have them consolidated as many categories only show <50k
- Country of organ for mother, father, and self may consider changing to USA not USA if performance looks to be ian issue
- Year could possibly be deleted
- Education should be updated to Ordinal

Exploratory Analysis

Nominal Features Summary

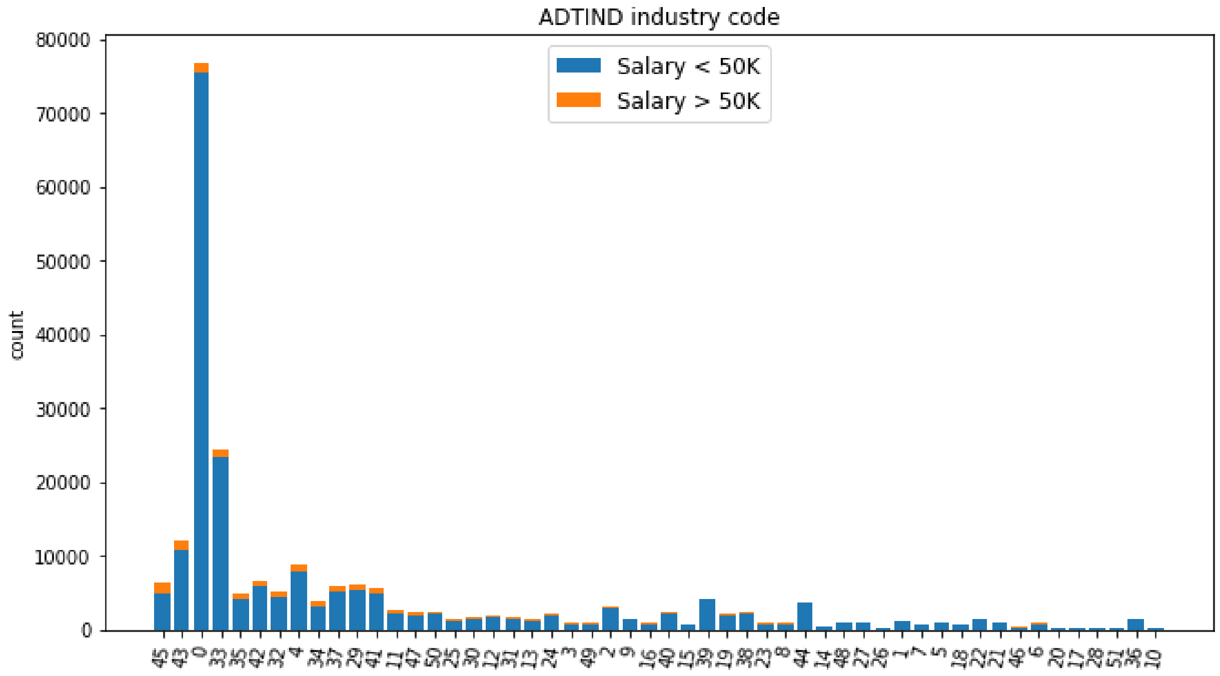
- - ADTIND industry code
- Leave as is. Majority are 0
- - ADTOCC occupation code
- Leave as is. Majority are 0
- - AHGA education
- Group under 1tth grade, reclassify as Ordinal
- - AMARITL marital status
- Leave as is.
- - ARACE race
- Leave as is.
- - AREORGN hispanic Origin
- Bin into Hispanic/ Not Hispanic
- - ASEX sex
- Change to binary. Note that Males are overrepresented in salary >50k

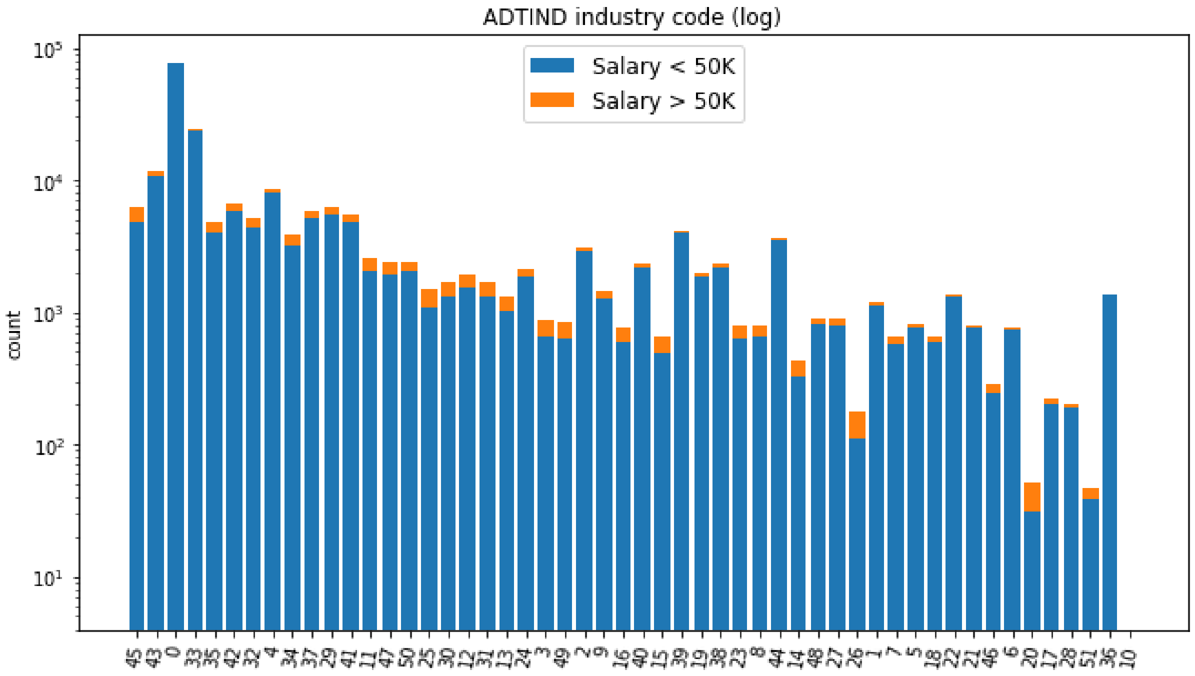
- - AWKSTAT full or part time employment stat
- Consolidate all values relating to unemployed in one category
- - FILESTAT tax filer status
- Leave as is.
- - HHDFMX detailed household and family stat
- Consolidate all categories except for householder, spouse of householder, nonfamily householder, secondary individual.
- Consider removing for HHDREL
- - HHDREL detailed household summary in household
- Consolidate all non householder categories
- - PEFNTVTY country of birth father
- As is, or possible USA and non-USA if performance is an issue
- - PEMNTVTYcountry of birth mother
- As is, or possible USA and non-USA if performance is an issue

- - PENATVTY country of birth self
- As is, or possible USA and non-USA if performance is an issue
- - PRCITSHP citizenship
- Leave as is
- - SEOTR own business or self employed
- - VETYN veterans benefits
- Leave As is
- - YEAR NaN
- Consider Removing, seems to be informational to when data was collected

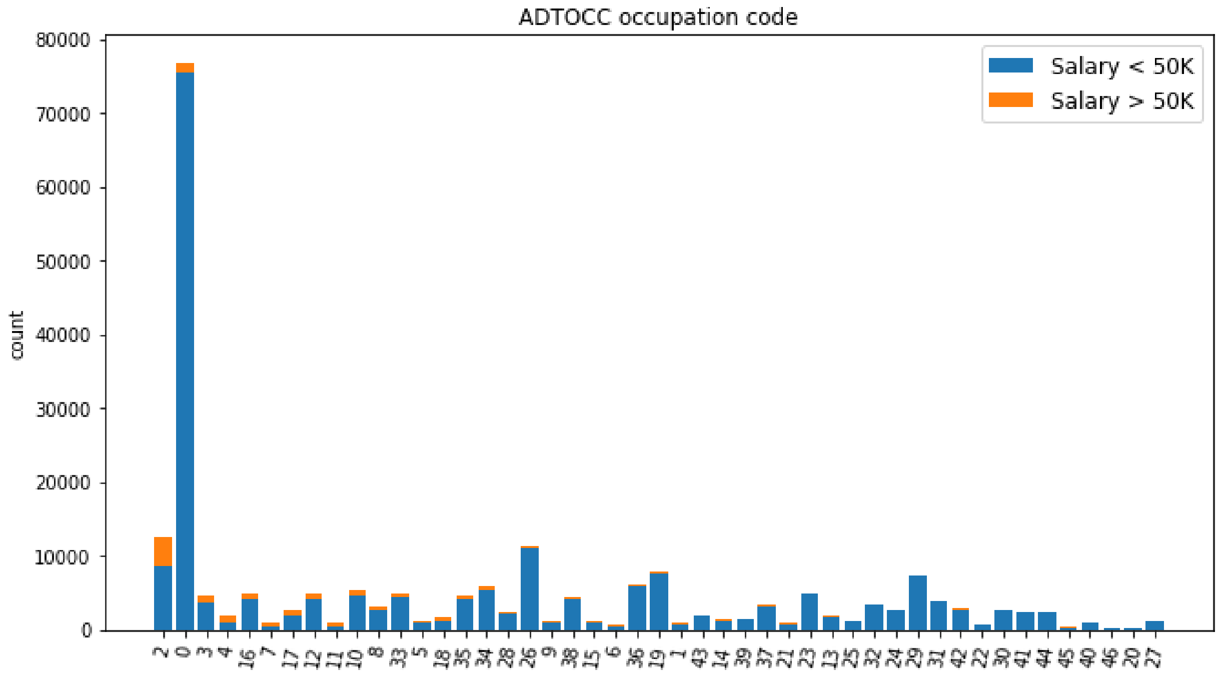
UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

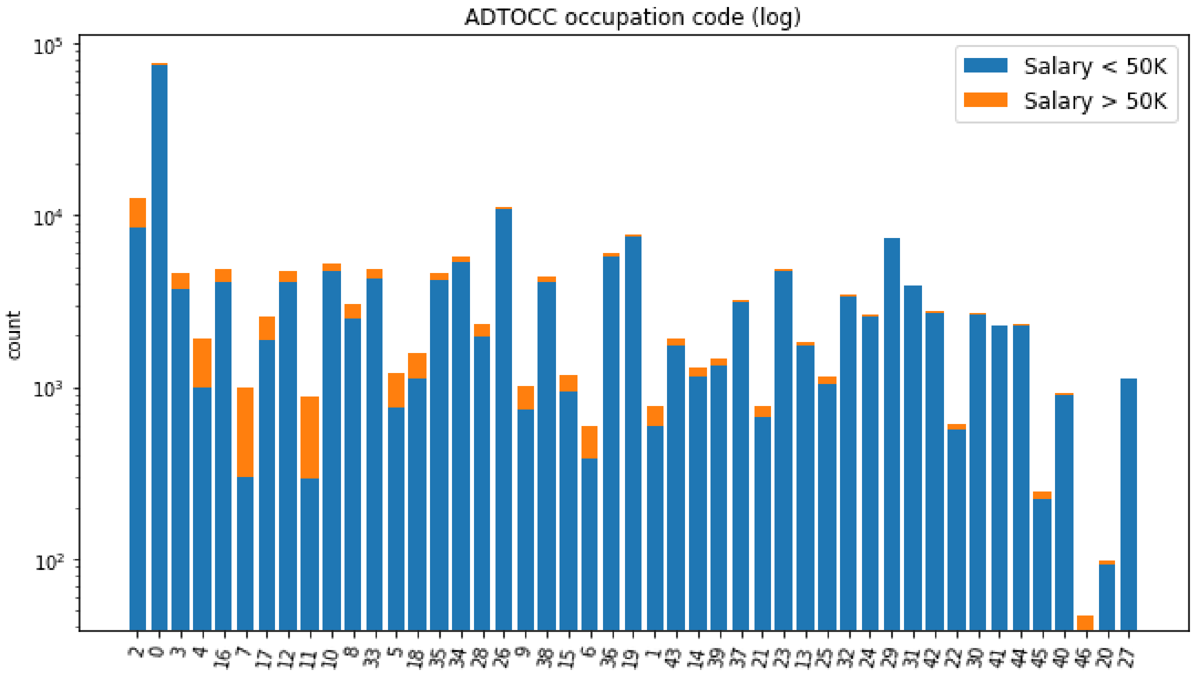
ADTIND   industry code



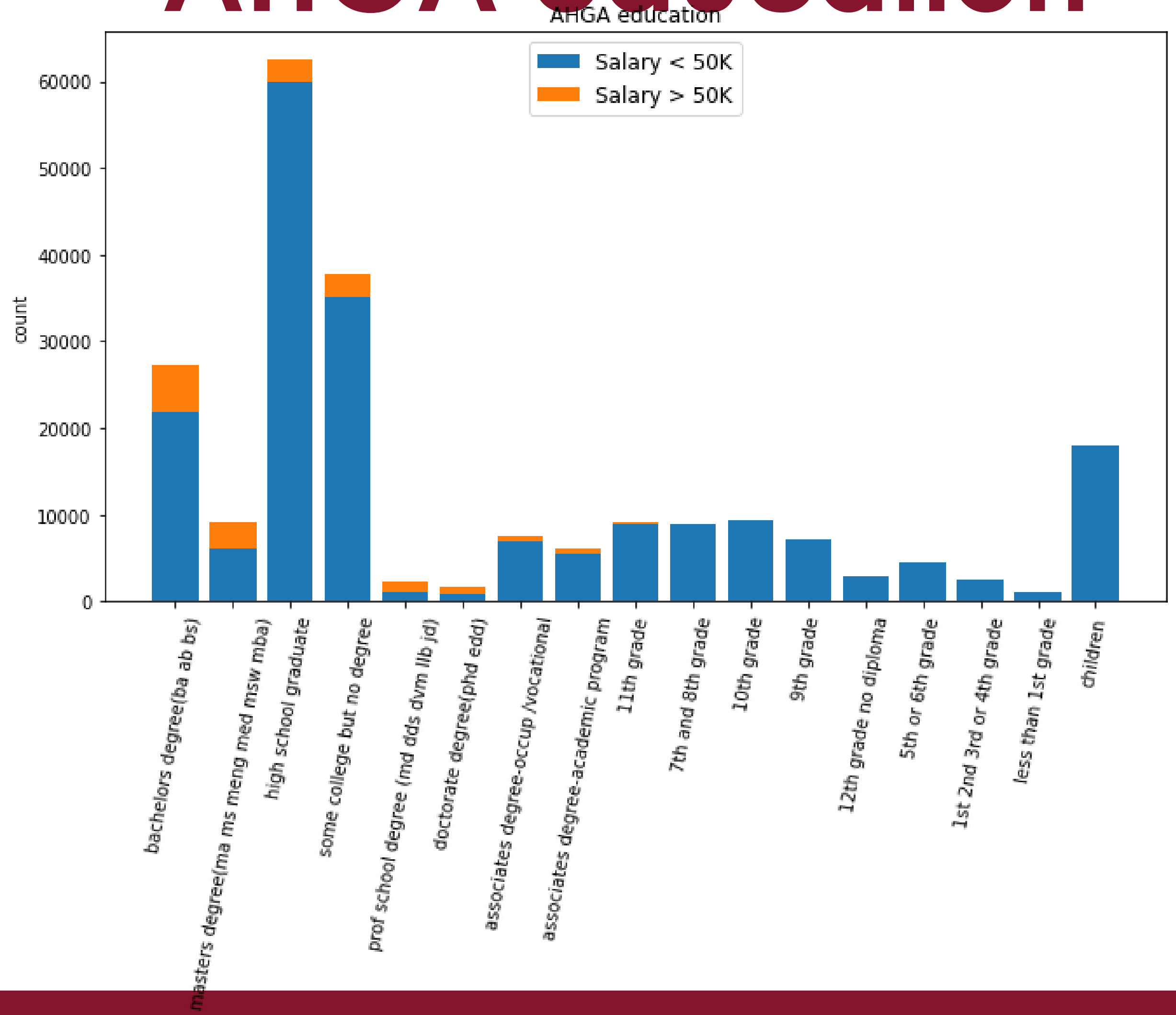


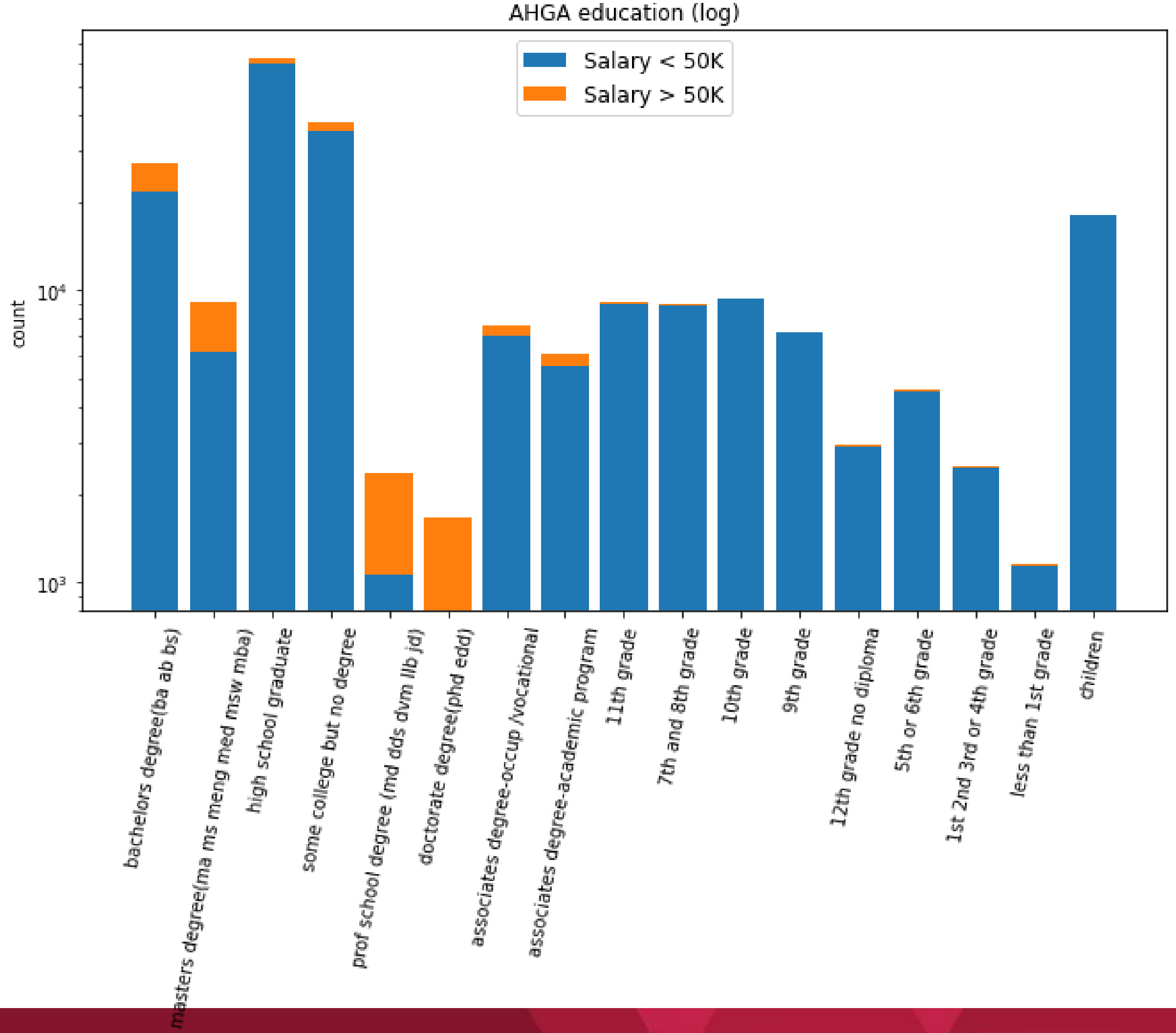
ADTOCC occupation code



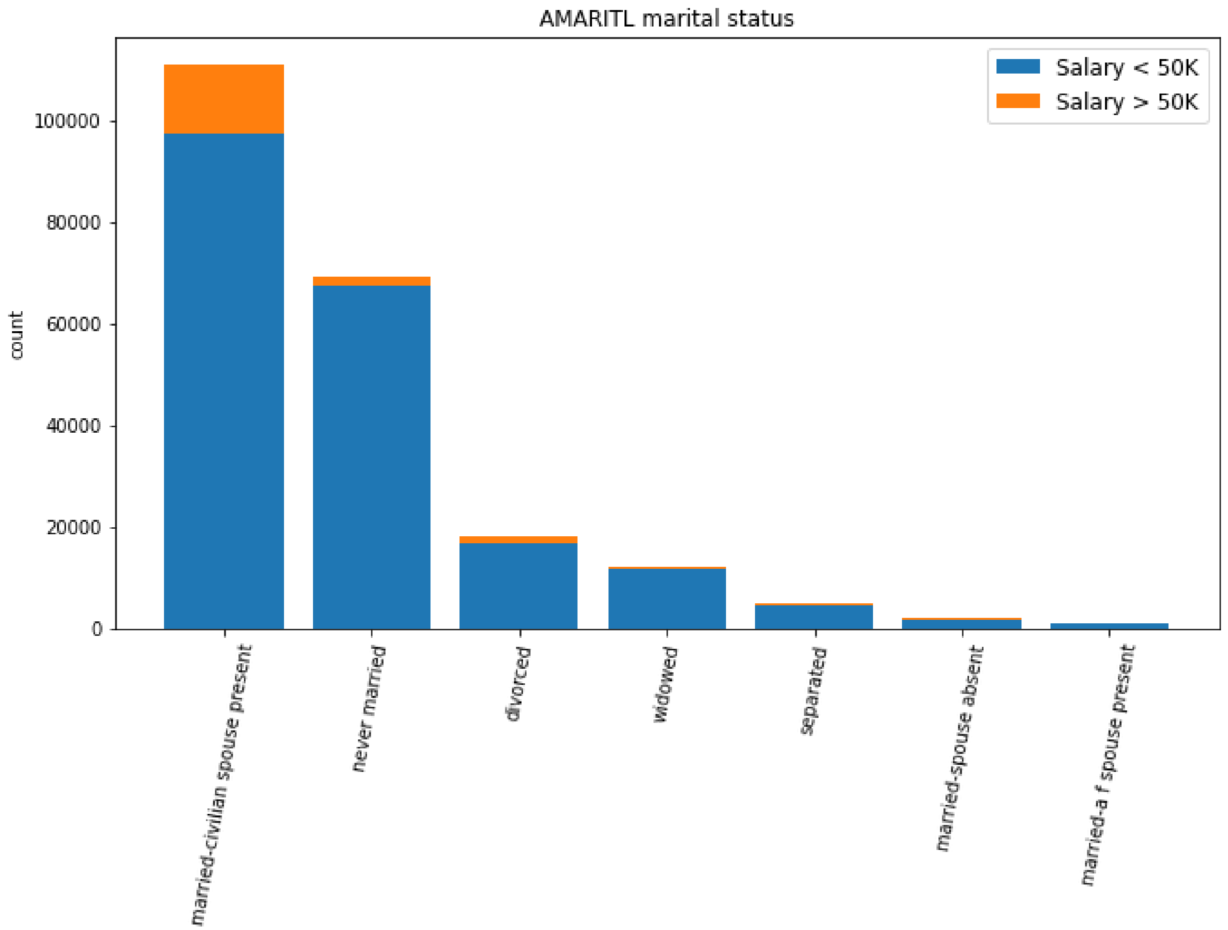


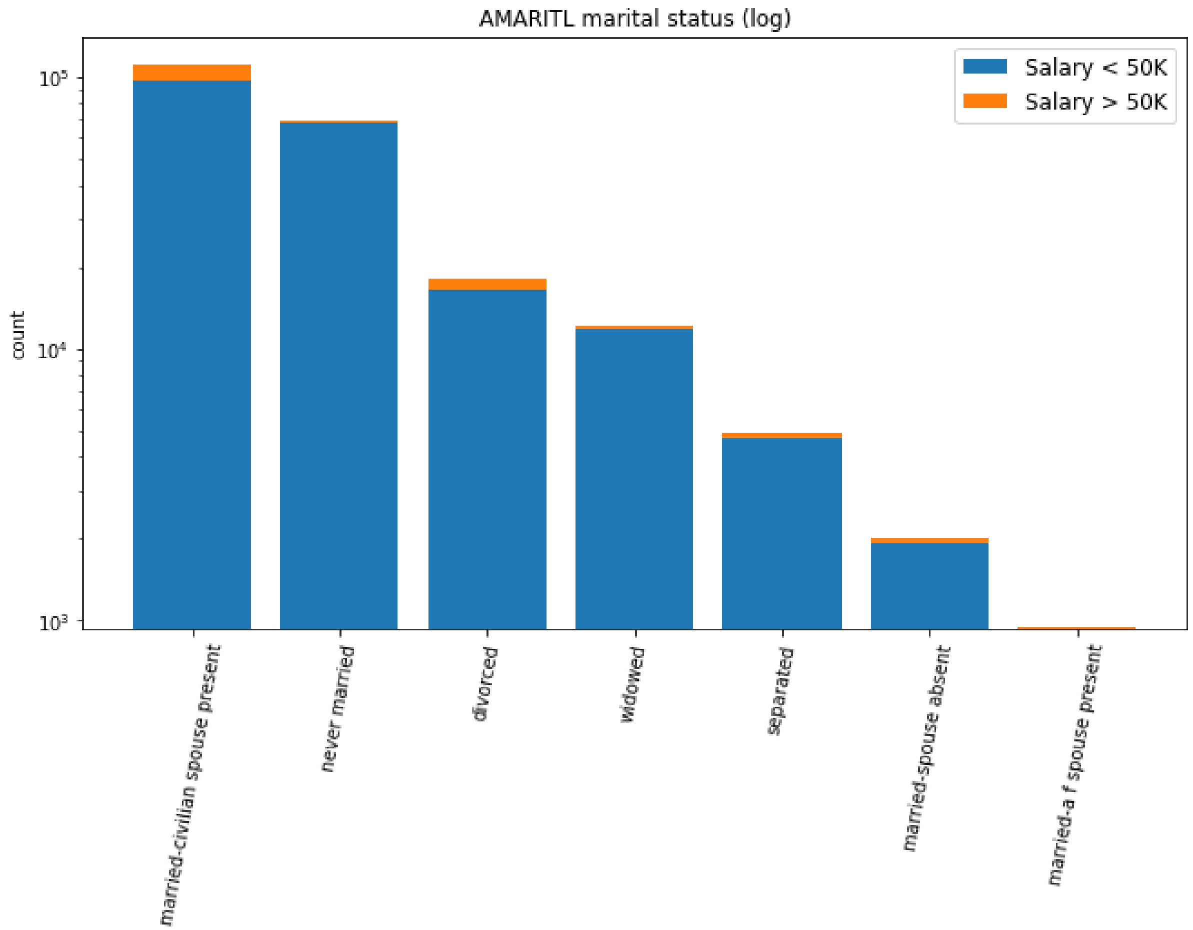
AHGA education



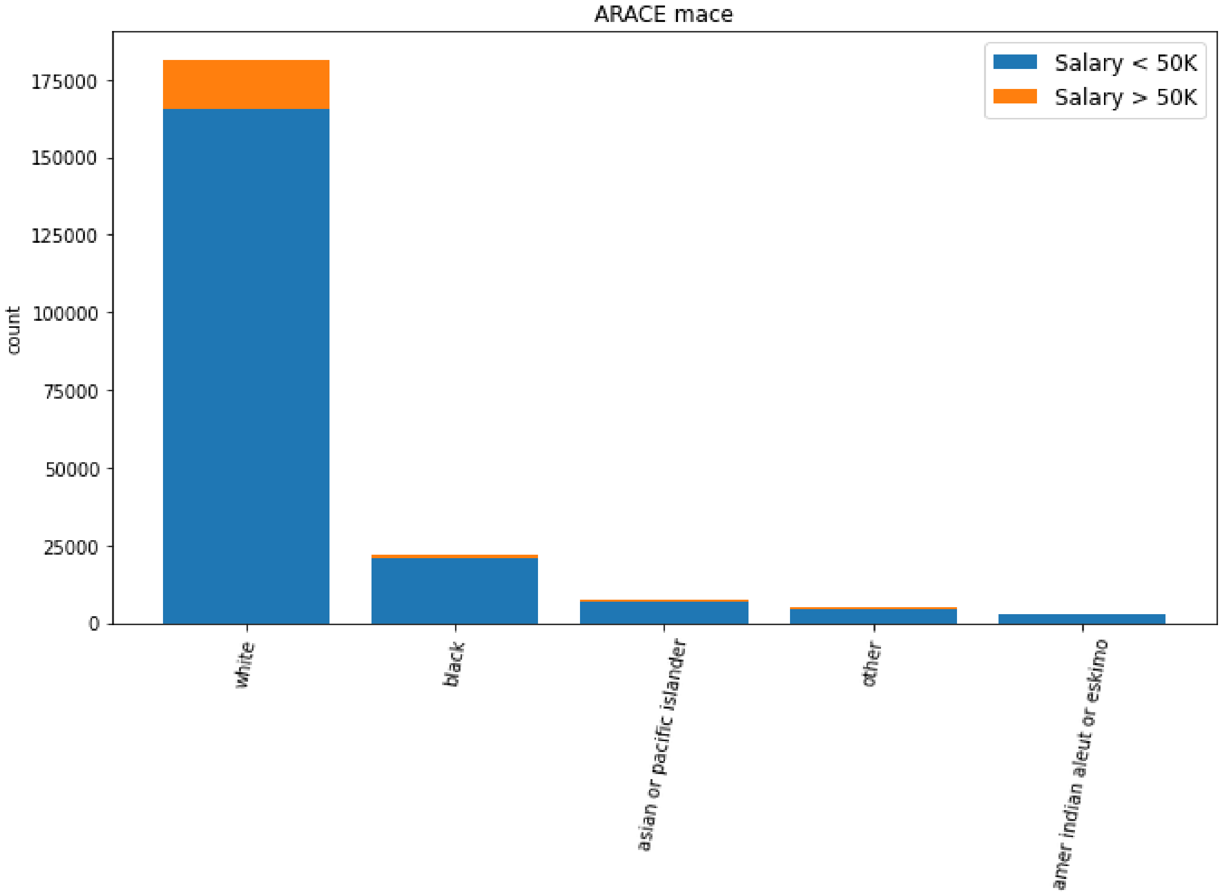


AMARITL marital status

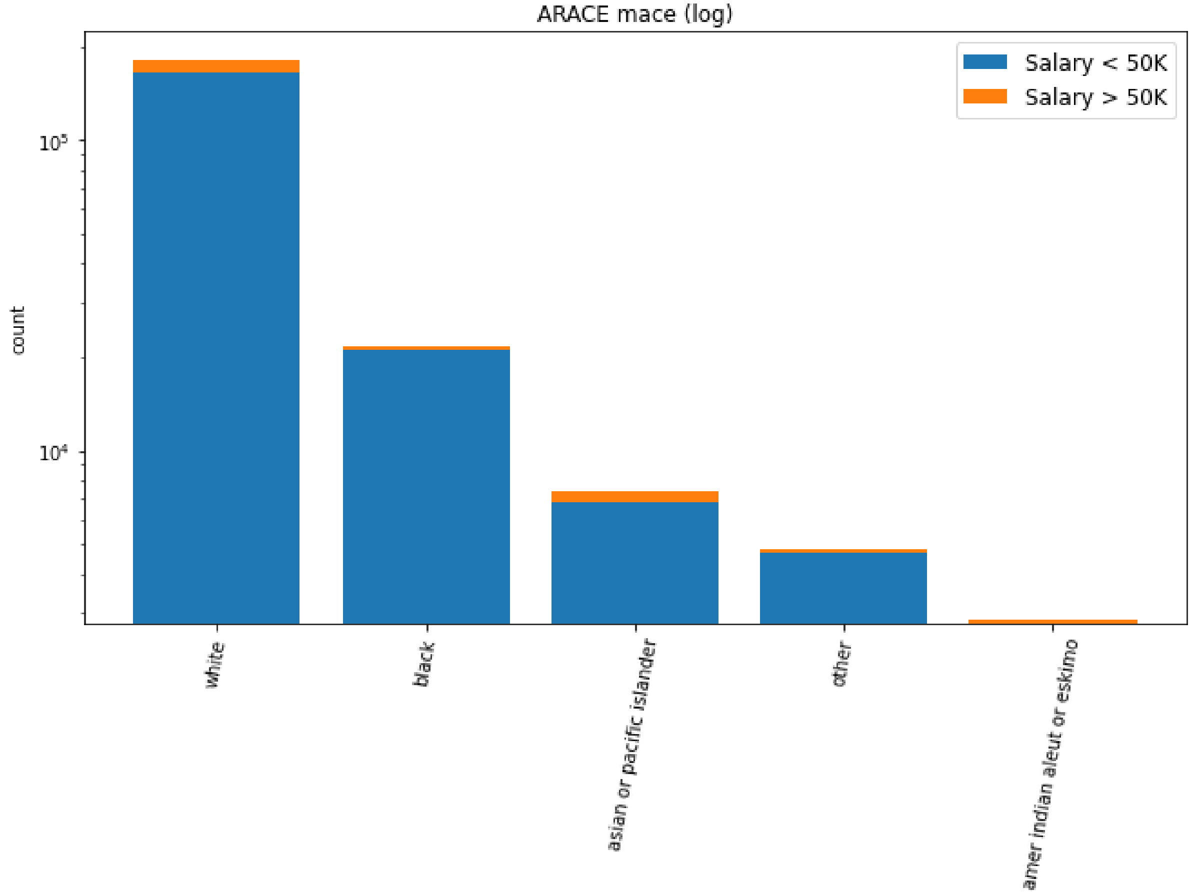




ARACE race

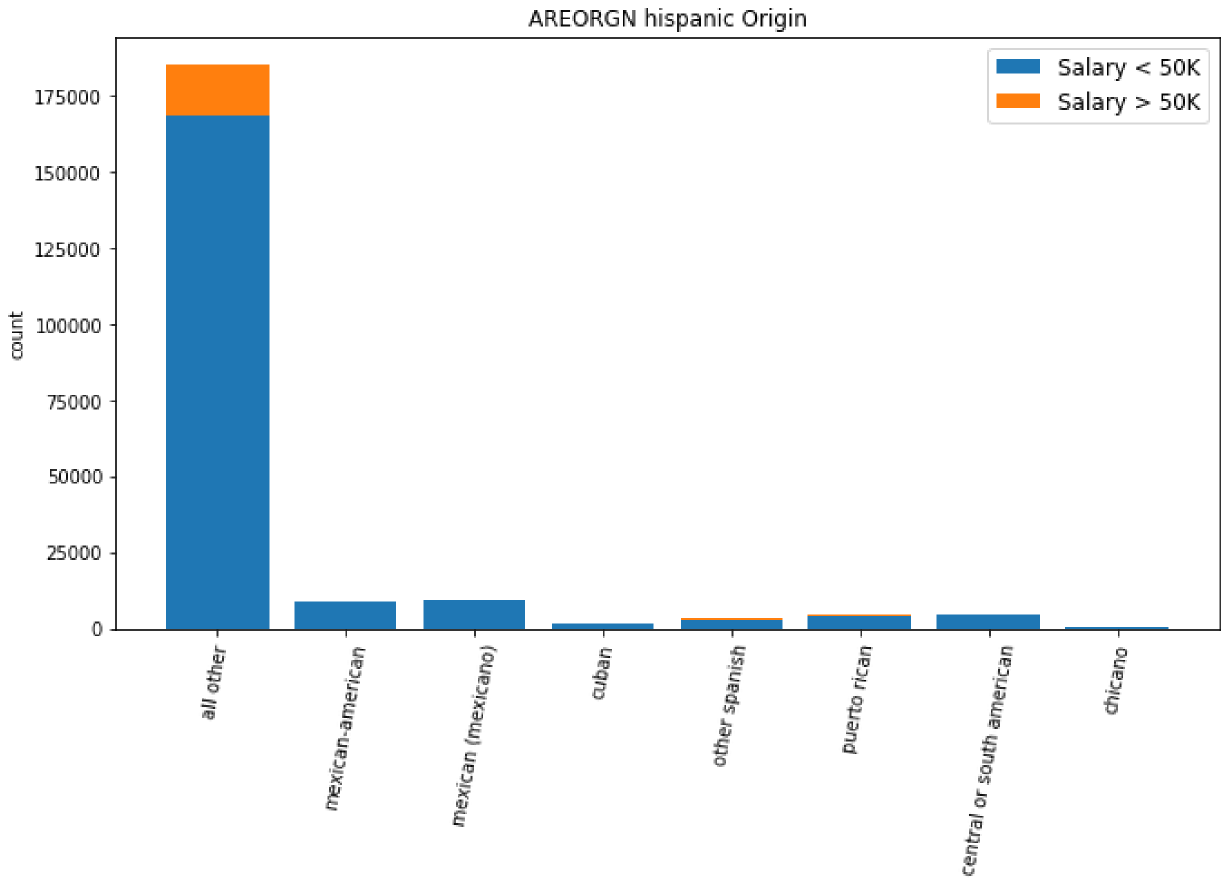


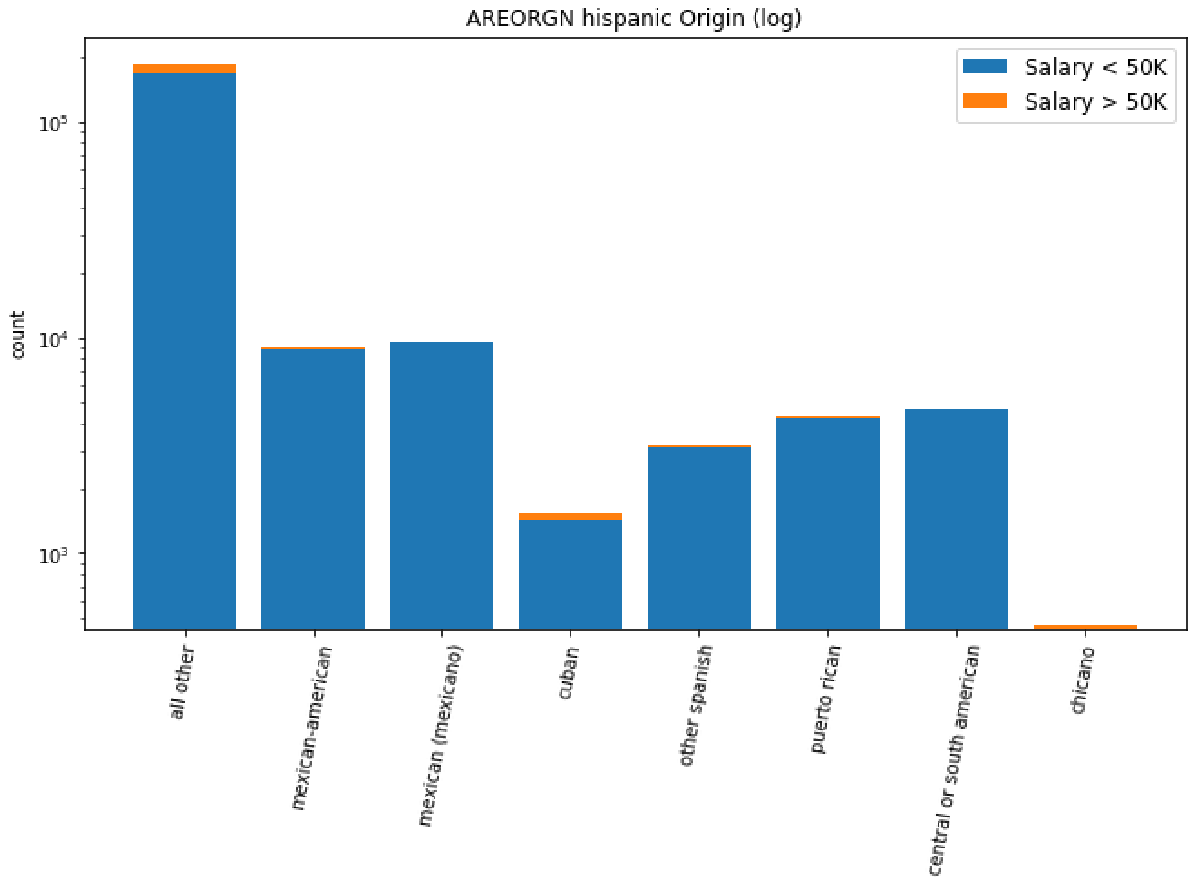
ARACE mace
Salary < 50K
Salary > 50K
count
175000
150000
125000
100000
75000
50000
25000
0
white
black
asian or pacific islander
other
amer indian aleut or eskimo



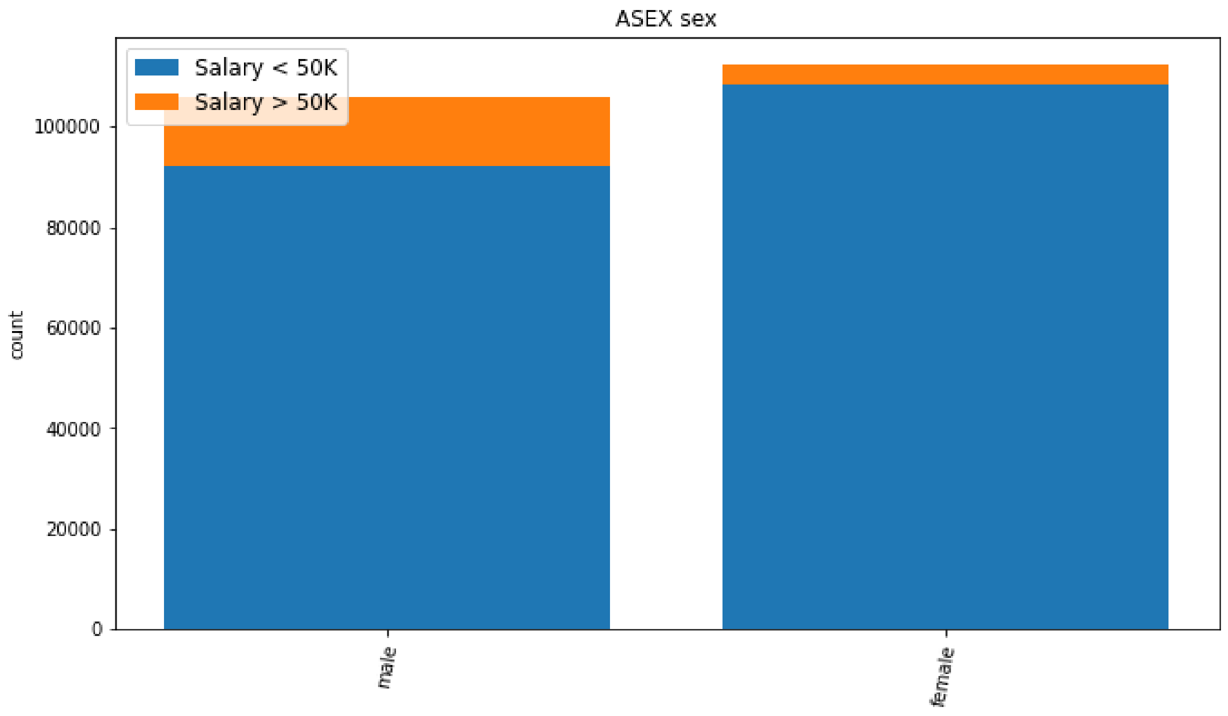
ARACE mace (log)
Salary < 50K
Salary > 50K
count
10^5
10^4
white
black
asian or pacific islander
other
amer indian aleut or eskimo

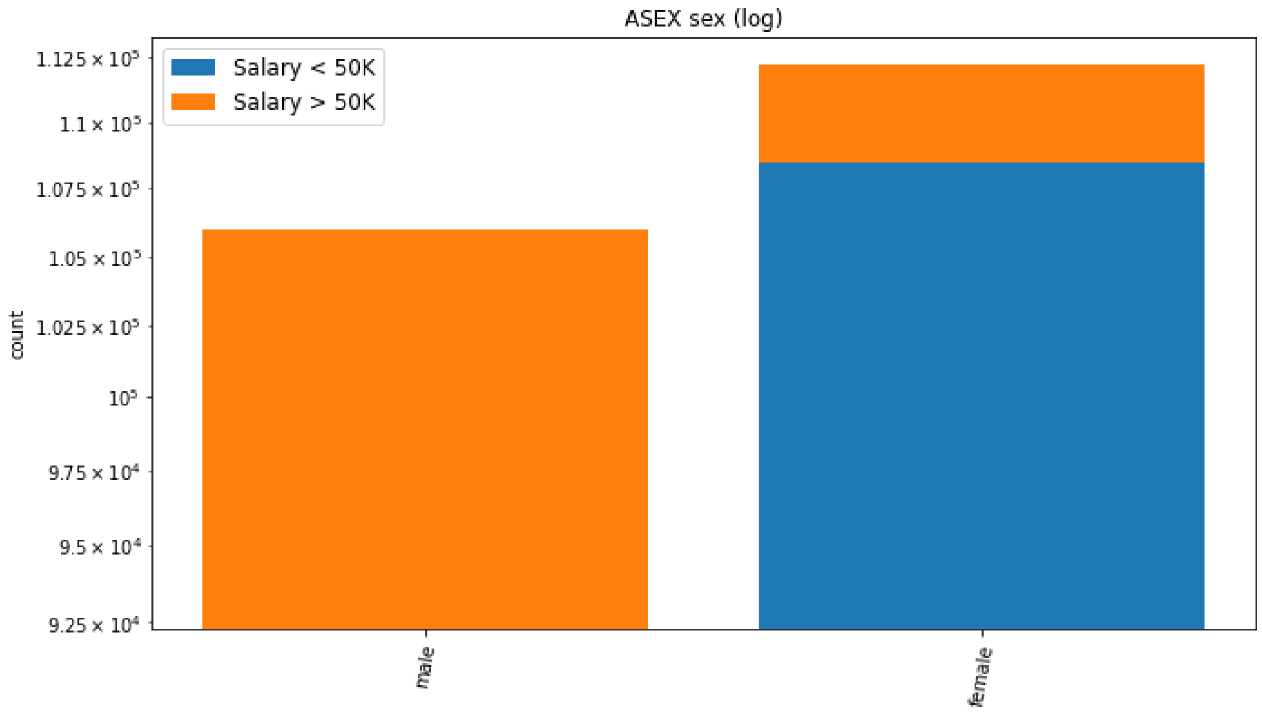
AREORGN Hispanic Origin



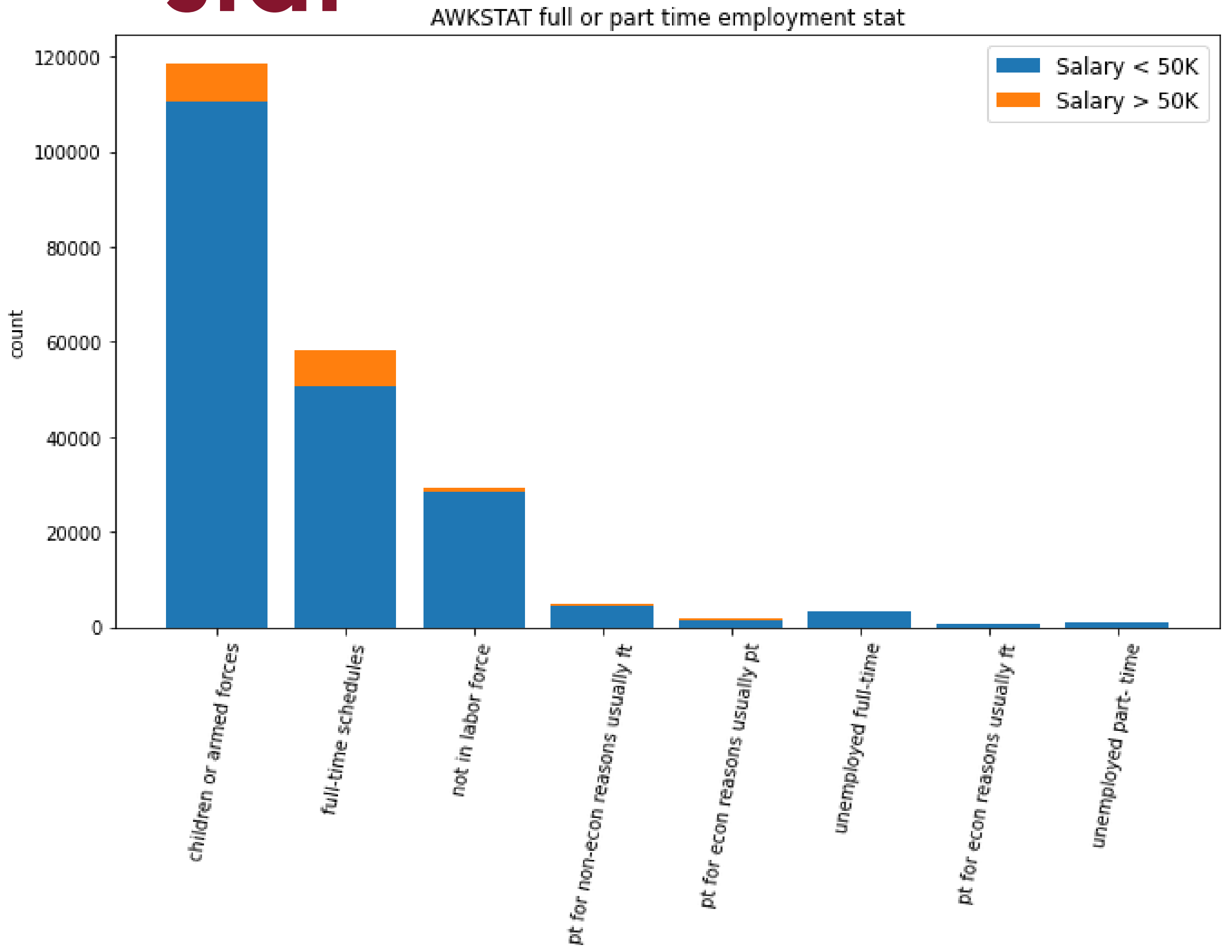


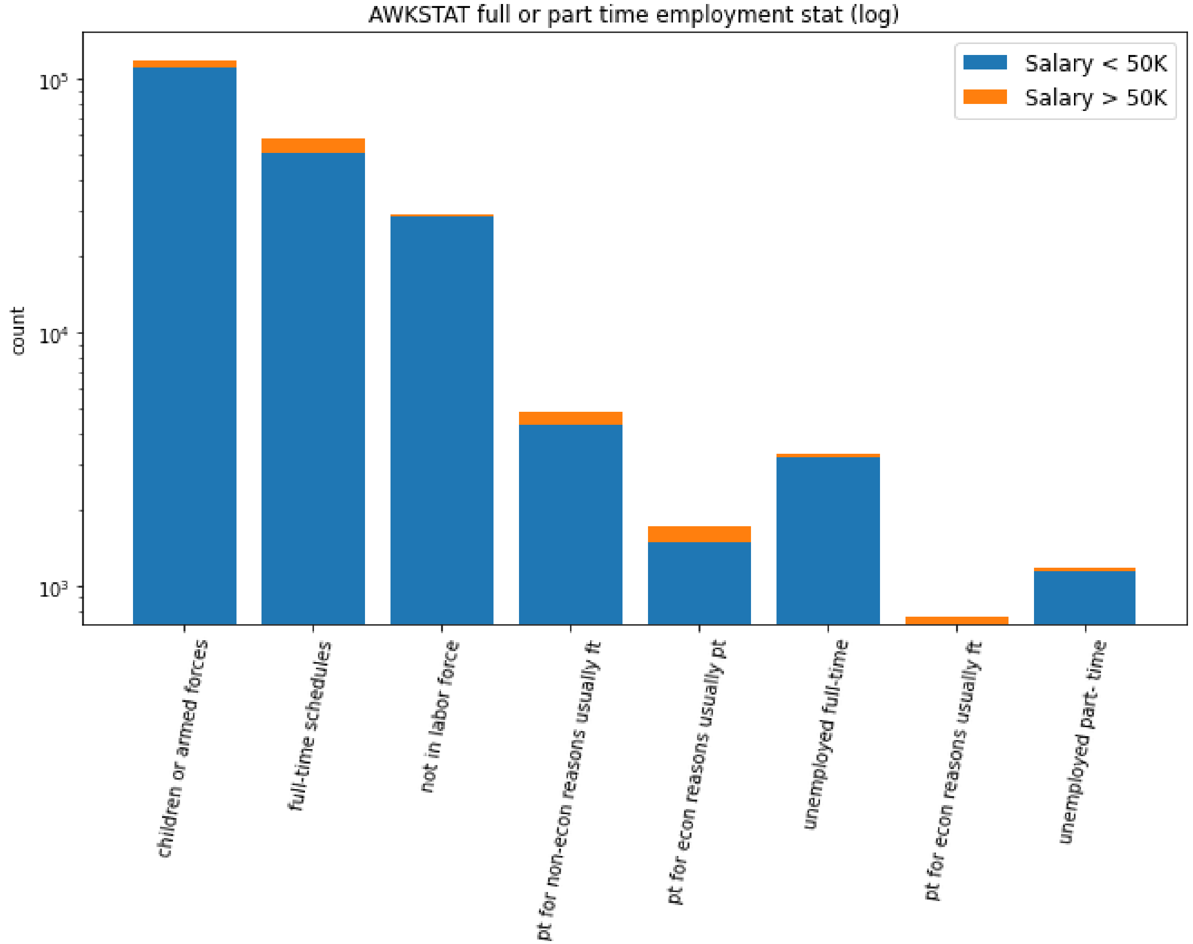
ASEX sex


ASEX sex
Salary < 50K
Salary > 50K
count
100000
80000
60000
40000
20000
0
male
female

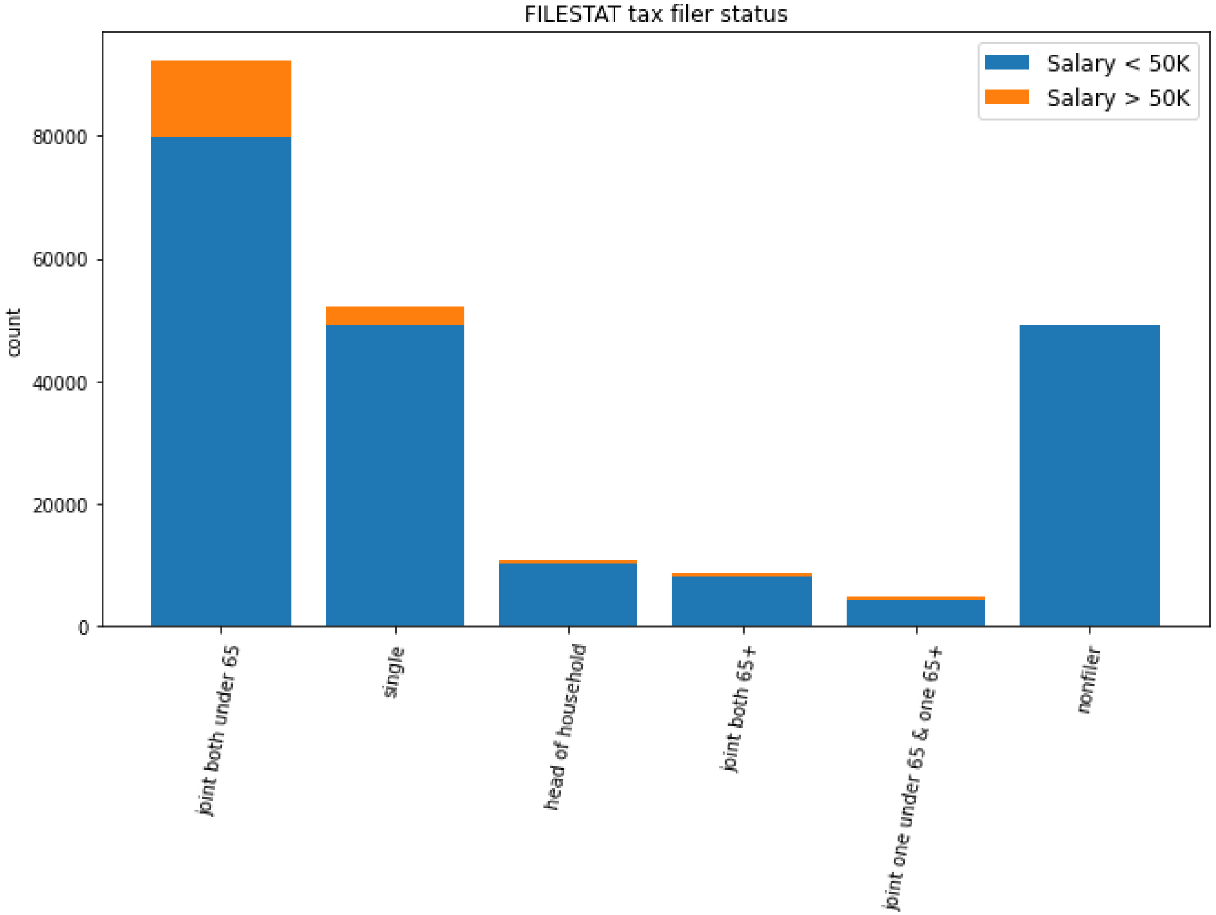

ASEX sex (log)
Salary < 50K
Salary > 50K
count
1.125 × 10^5
1.1 × 10^5
1.075 × 10^5
1.05 × 10^5
1.025 × 10^5
10^5
9.75 × 10^4
9.5 × 10^4
9.25 × 10^4
male
female

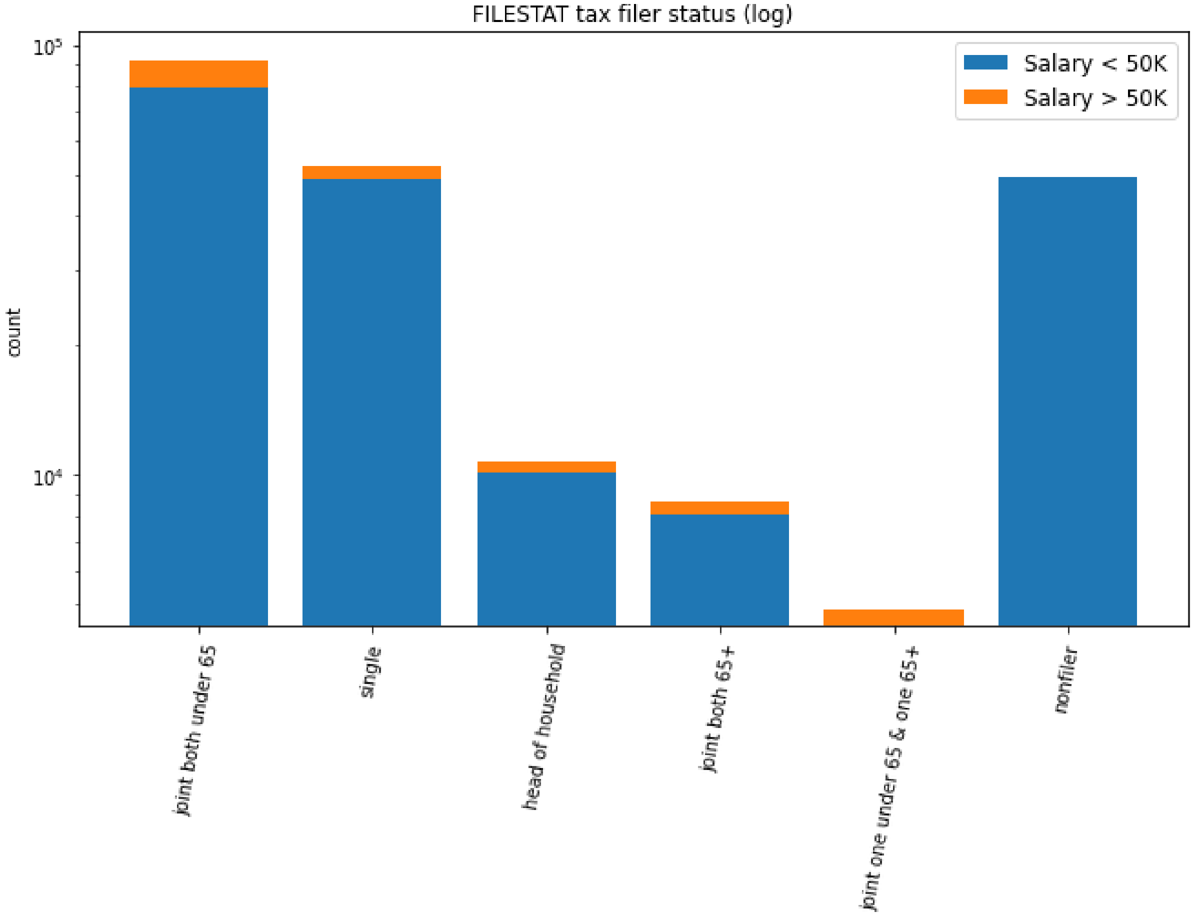
AWKSTAT full or part time employment stat



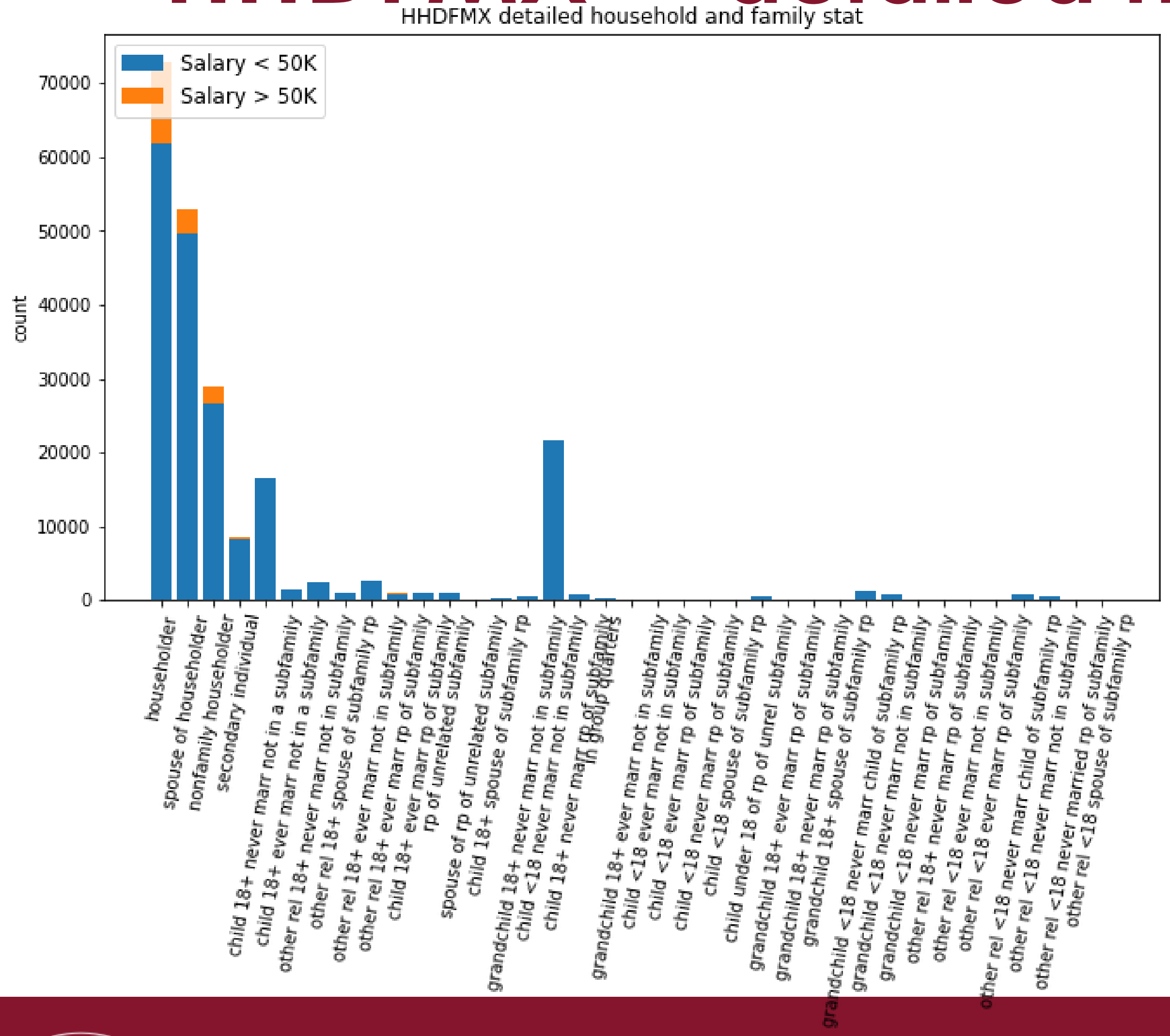


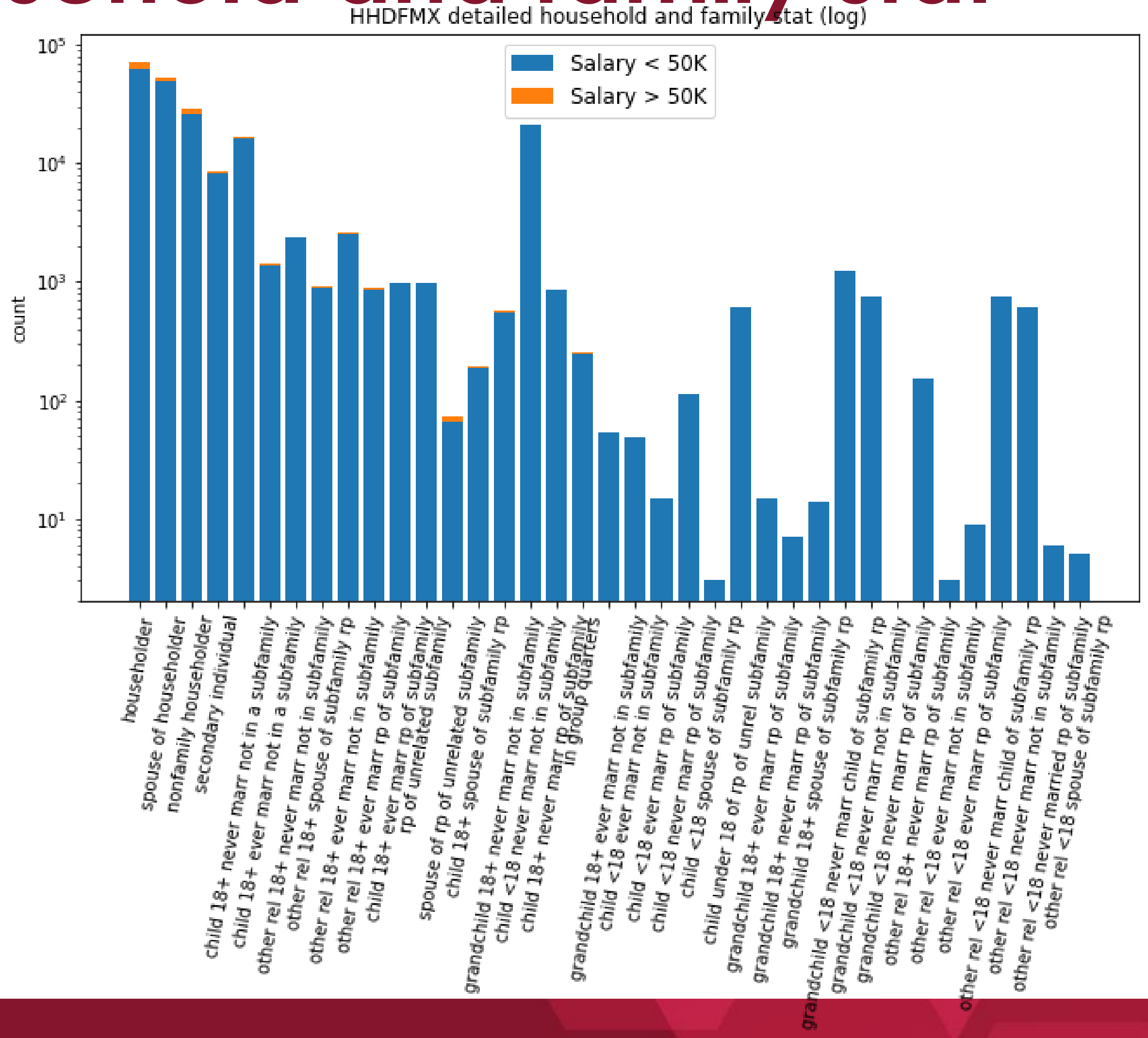
FILESTAT tax filer status



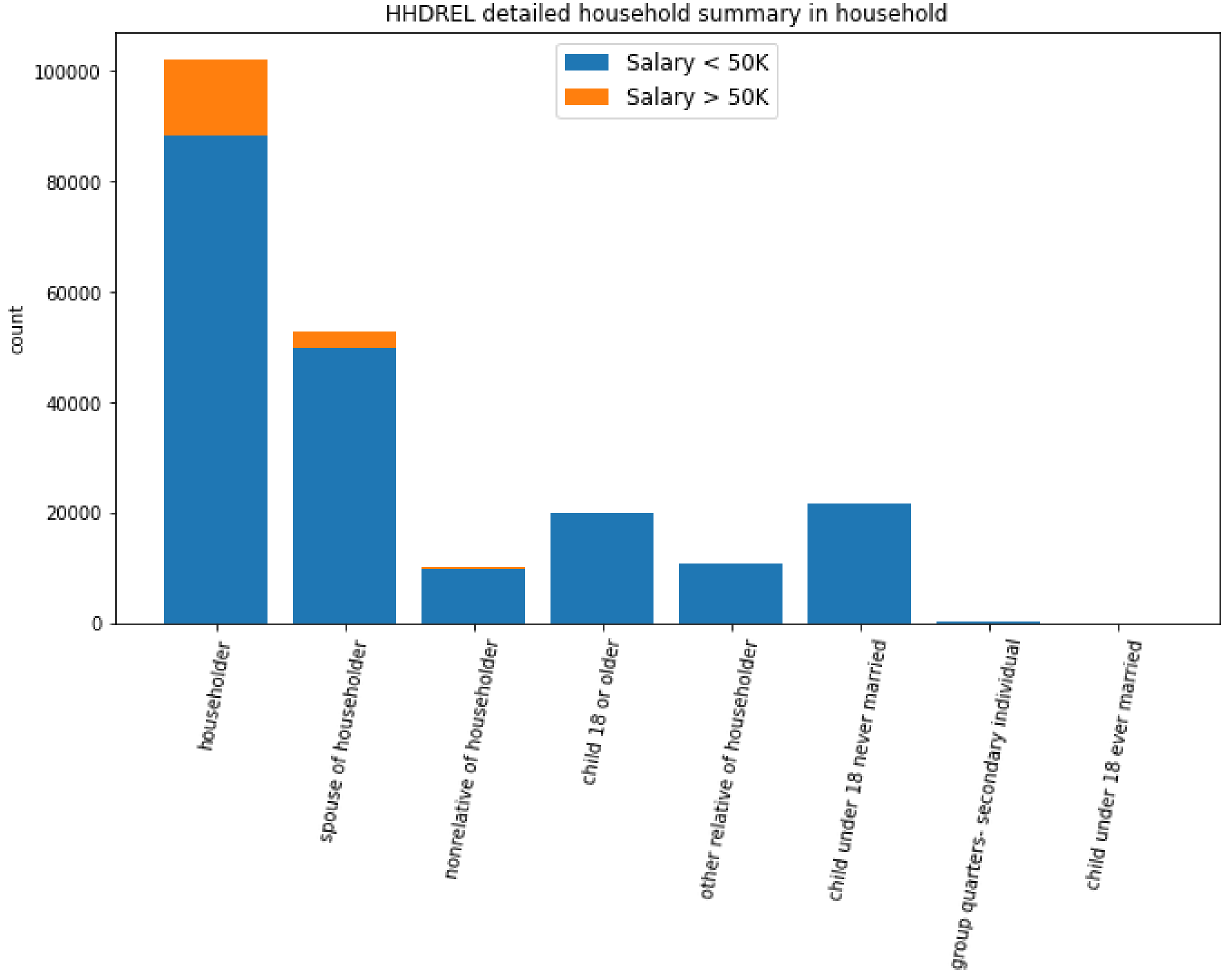


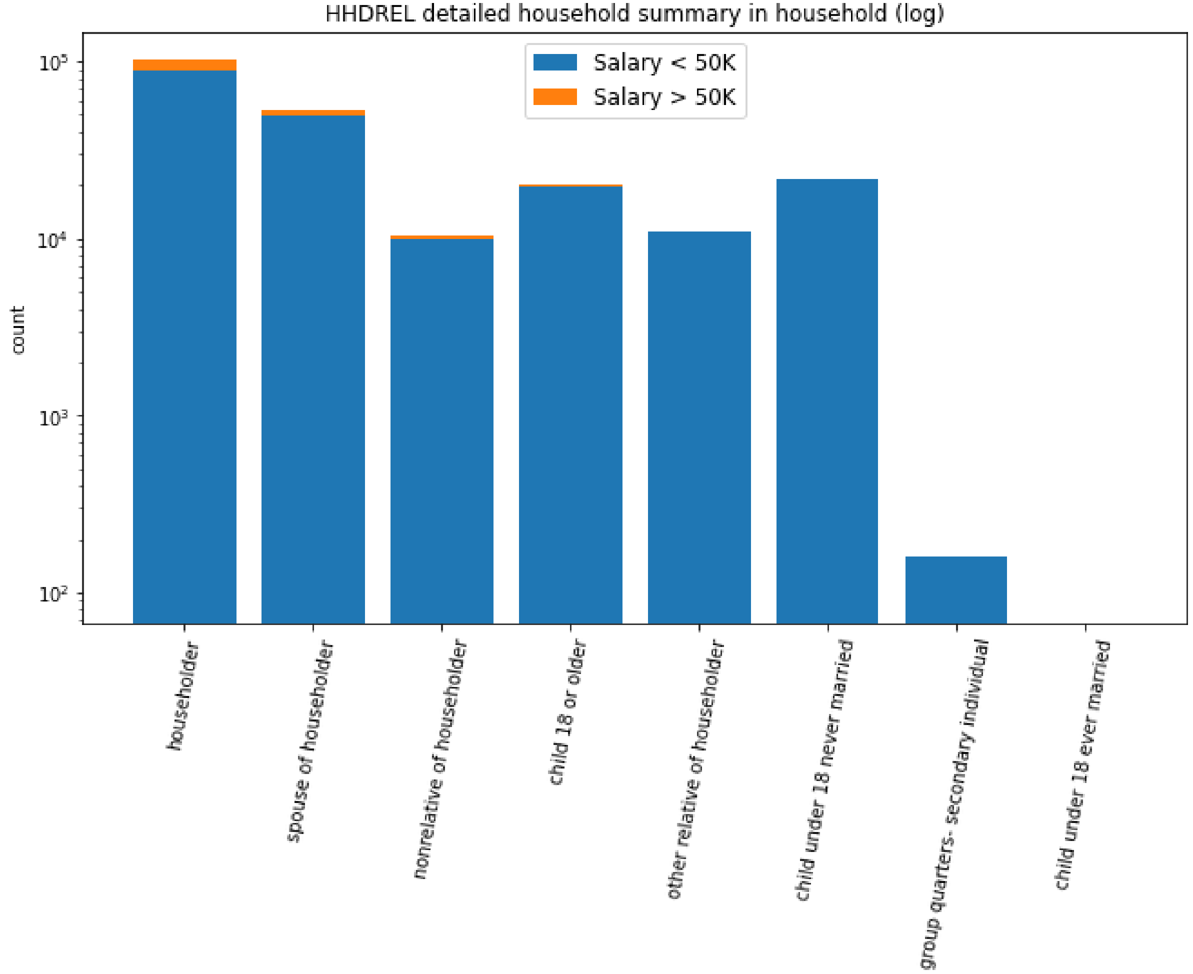
HHDFMX detailed household and family stat



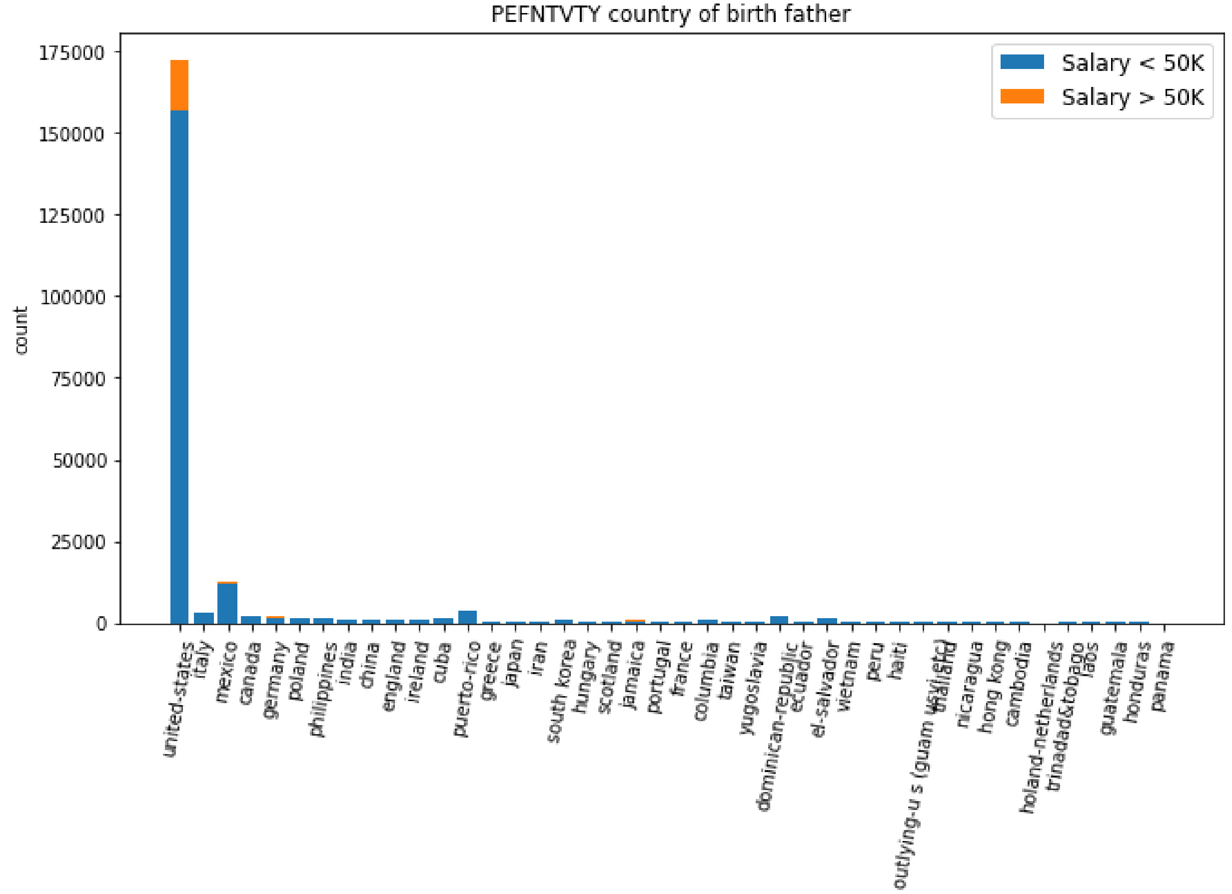


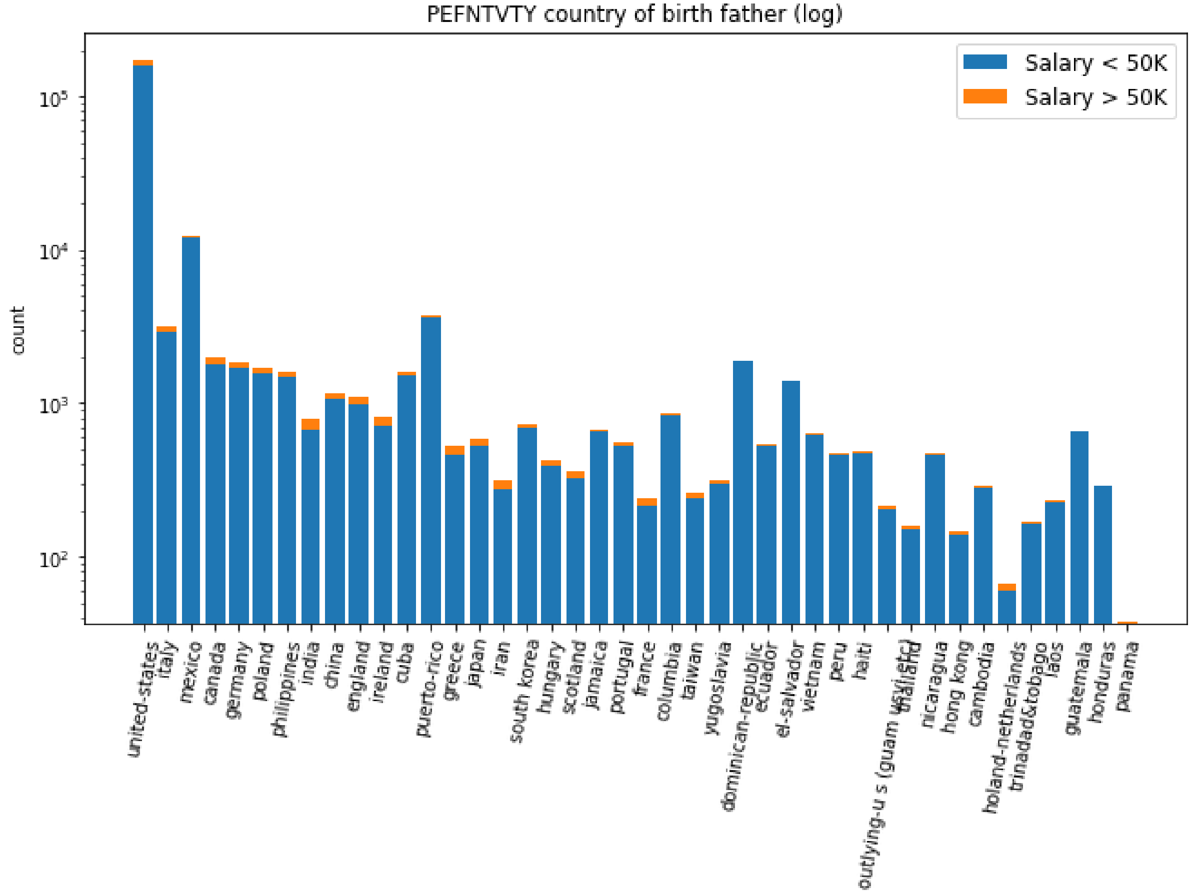
HHDREL detailed household summary in household



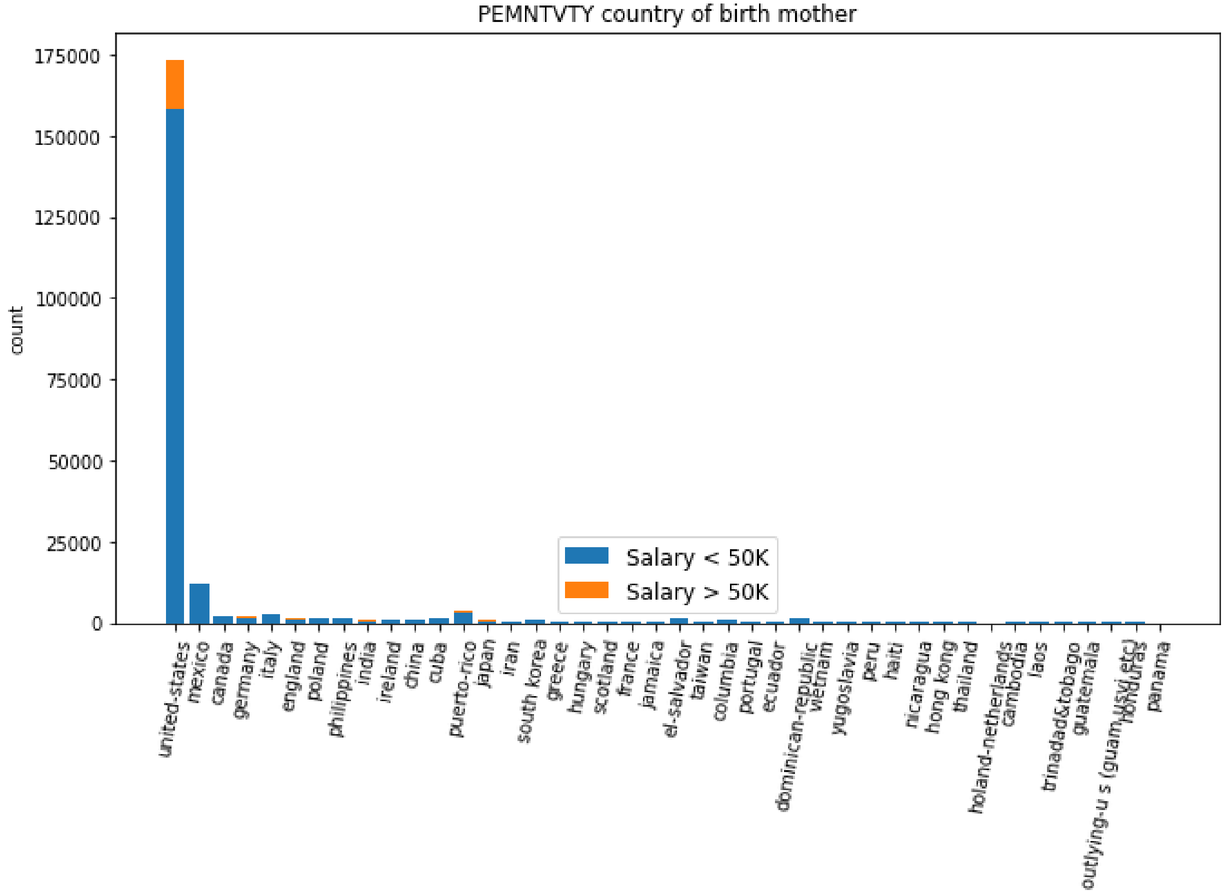


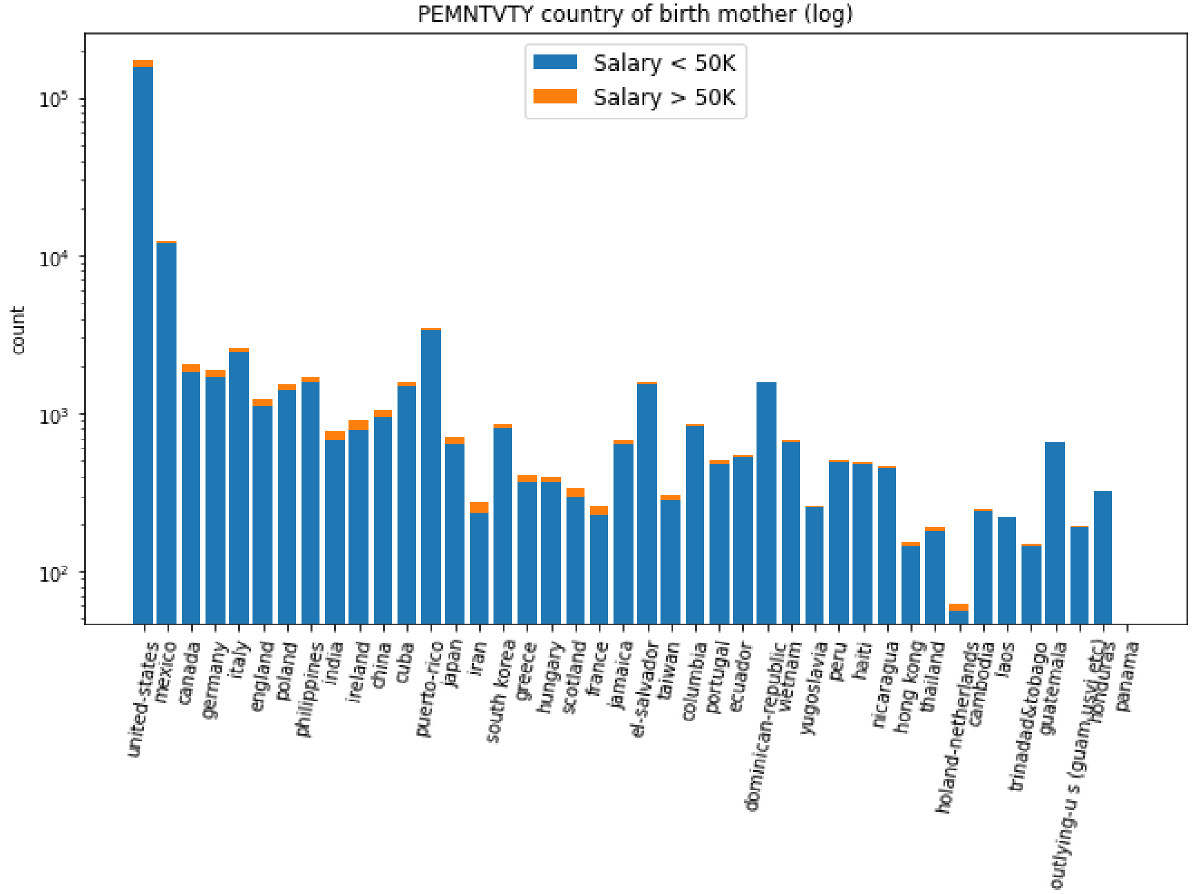
PEFNTVTY country of birth father



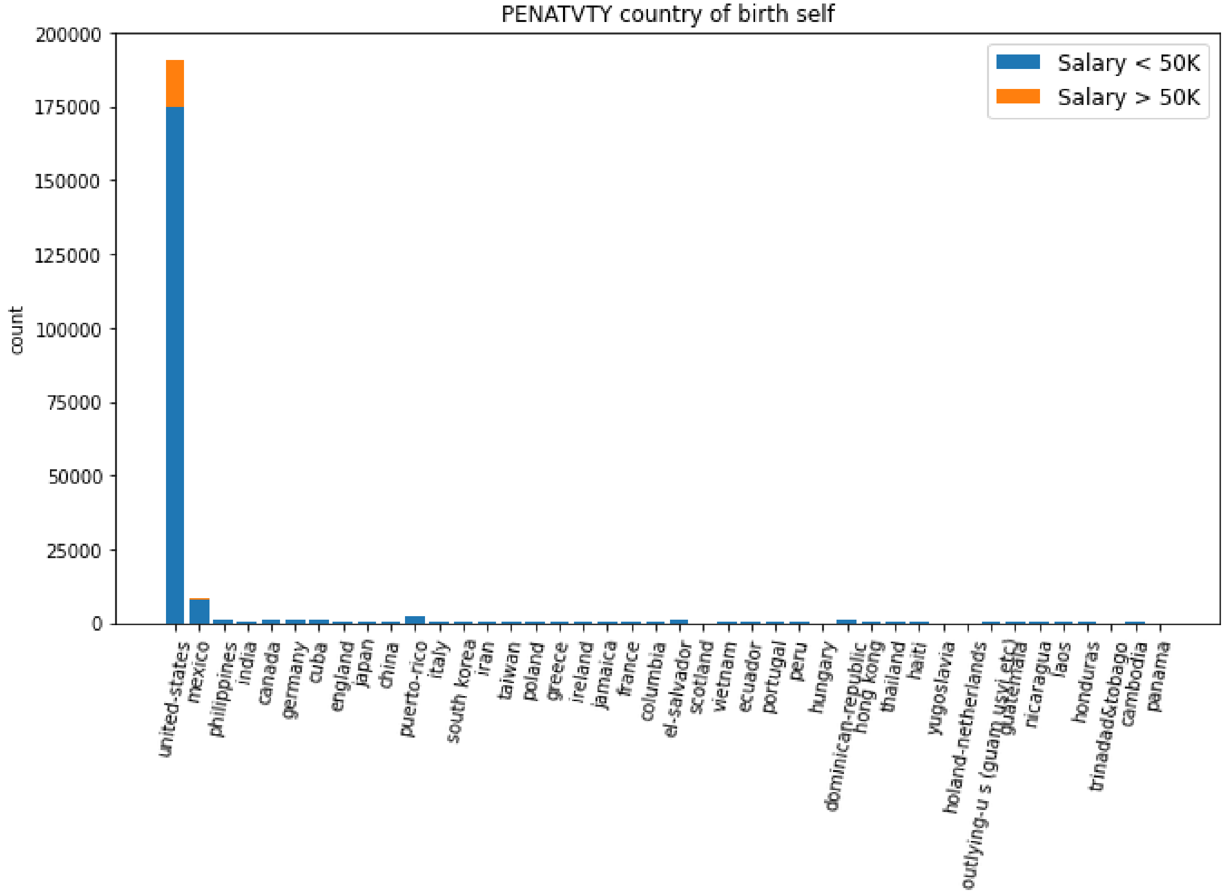


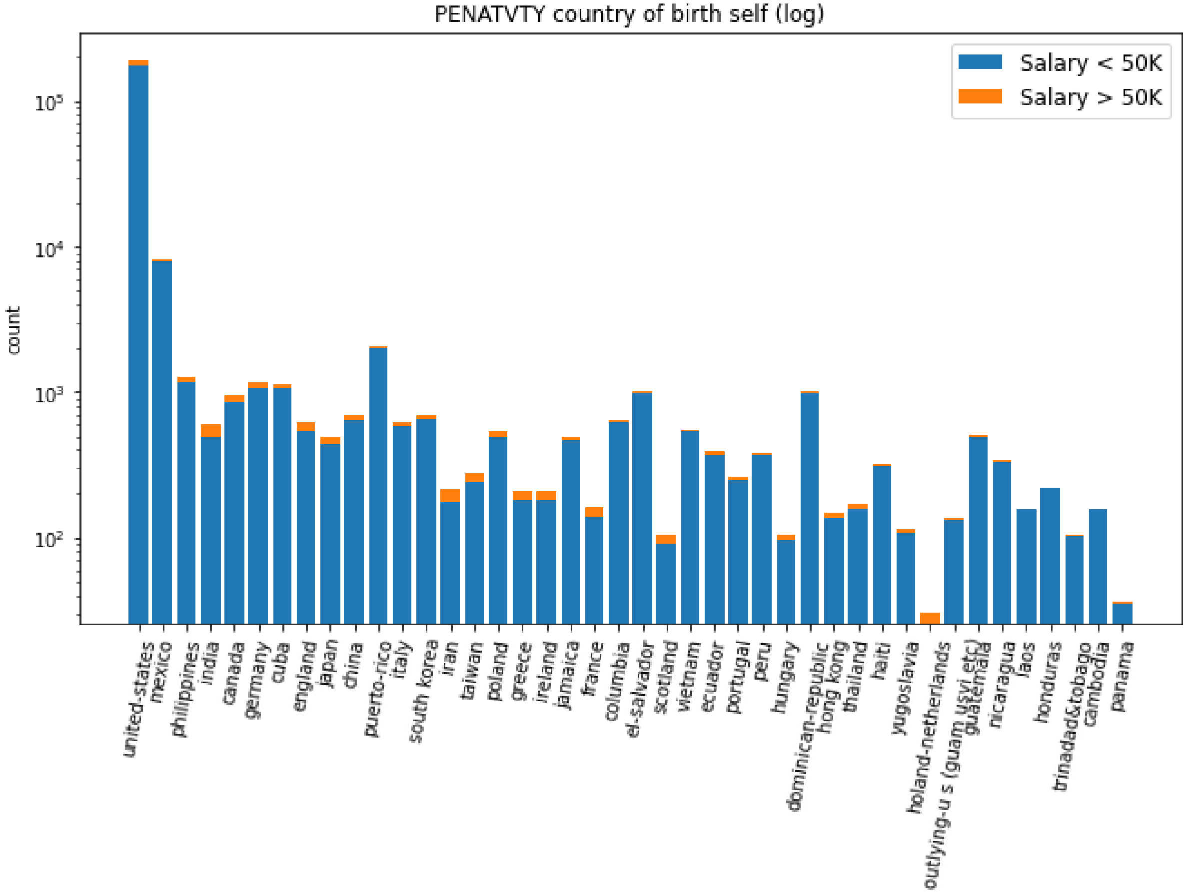
PEMNTVTYcountry of birth mother



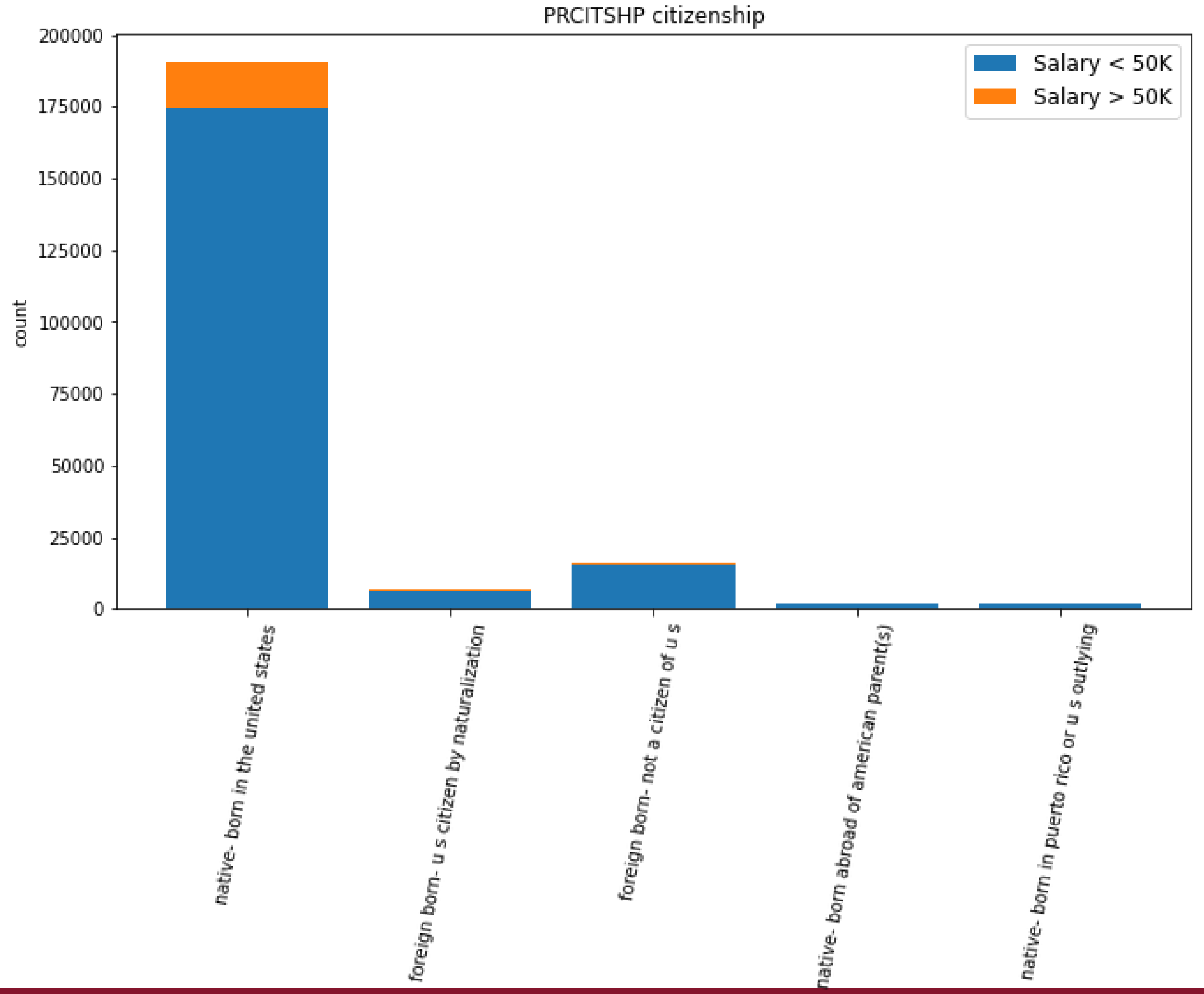


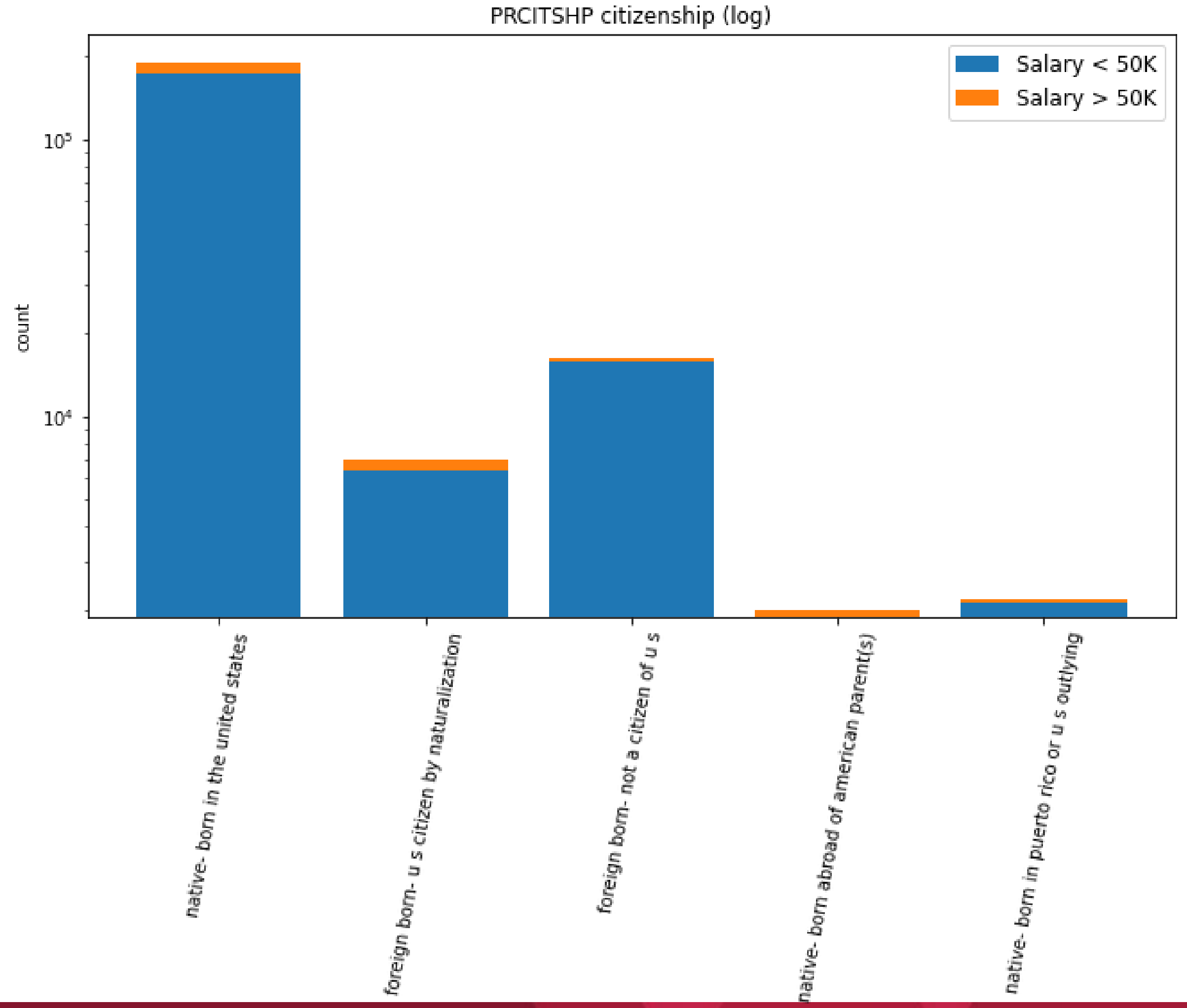
PENATVTY country of birth self



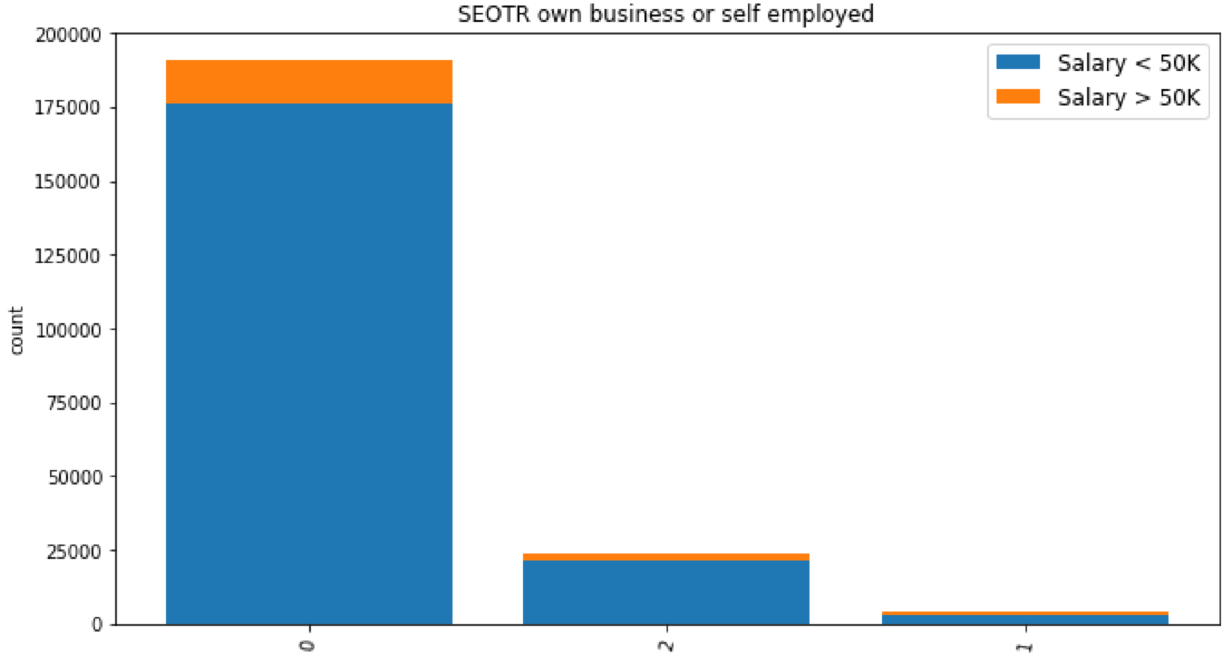


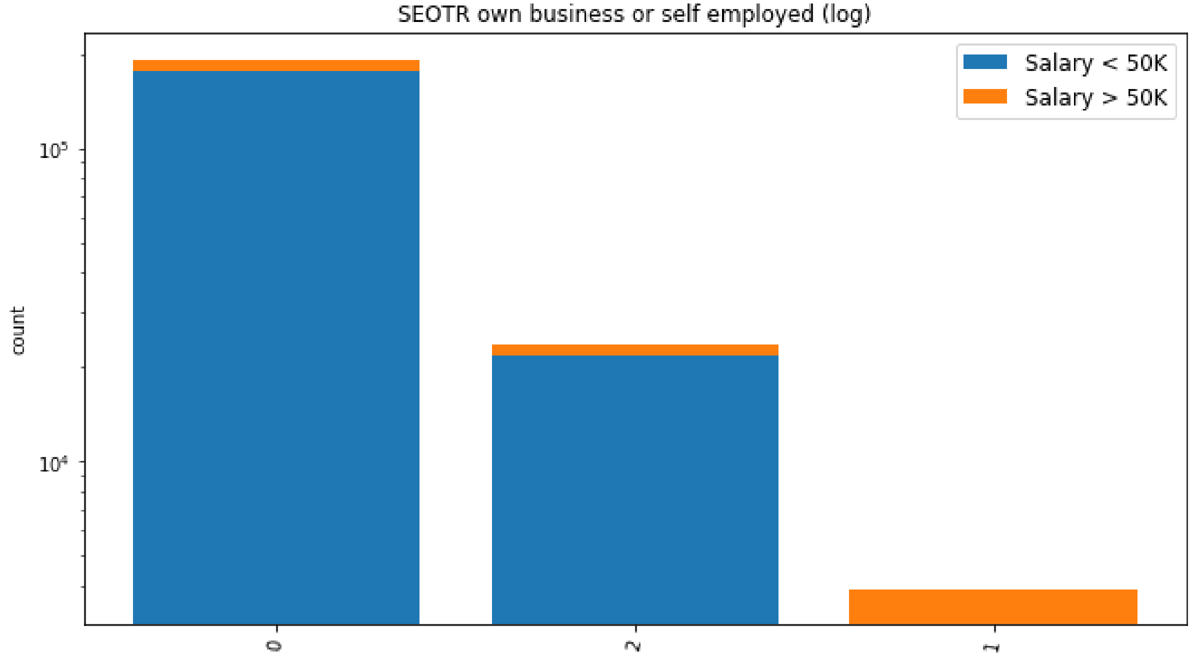
PRCITSHP citizenship



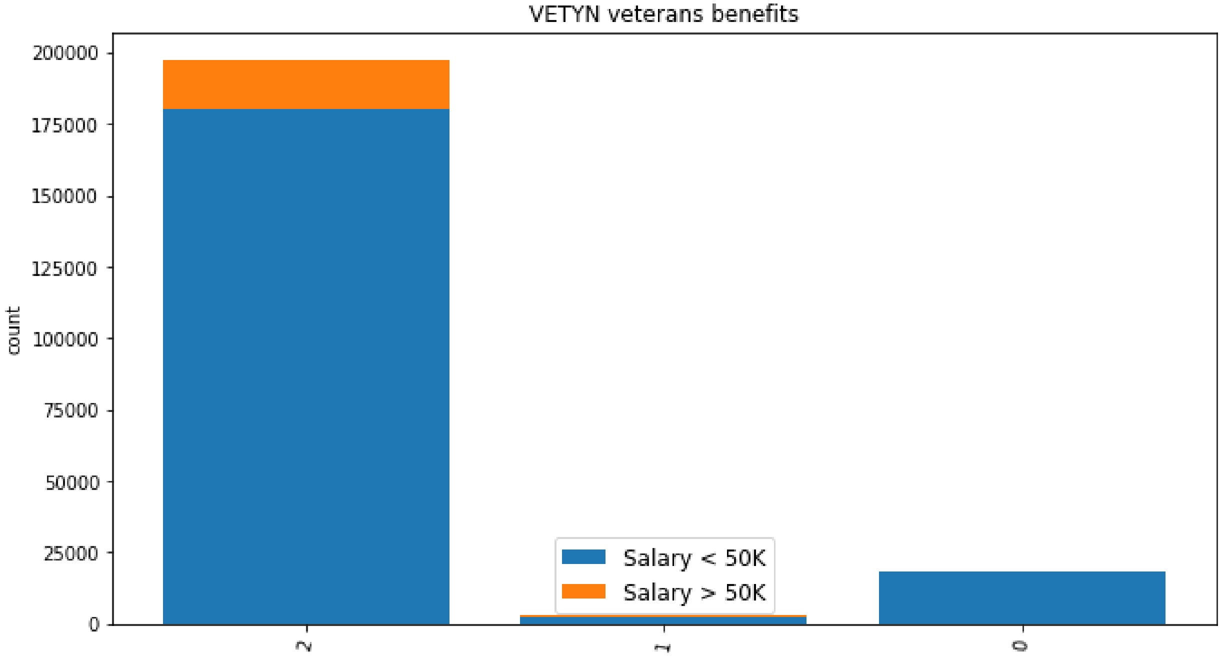


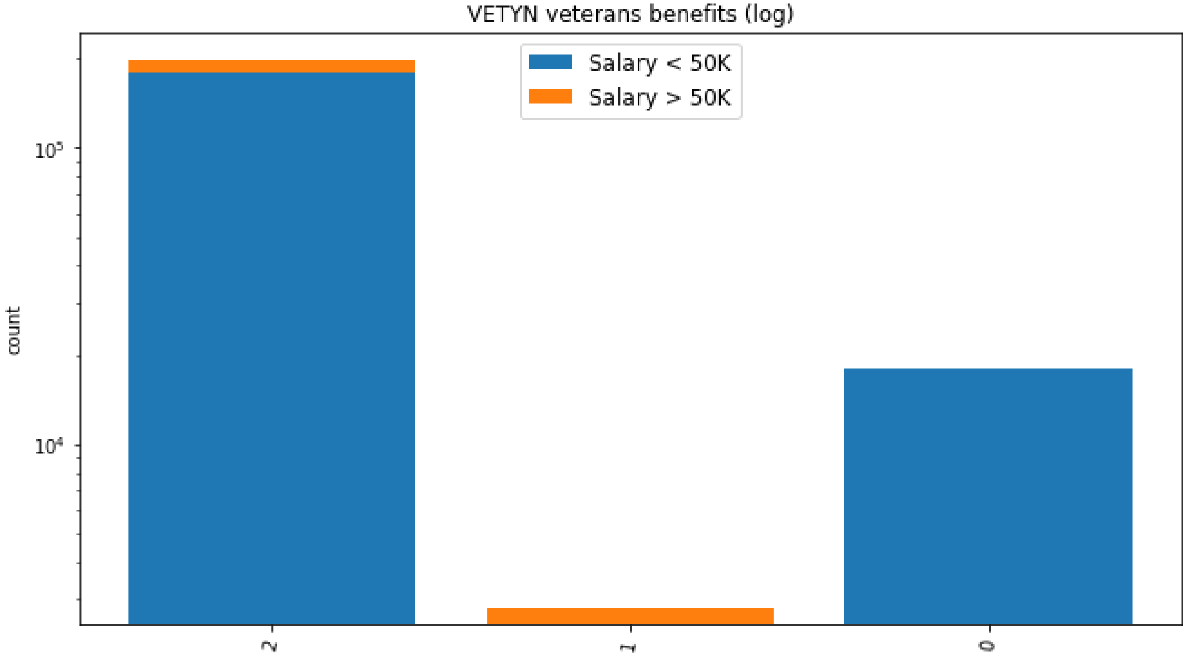
SEOTR own business or self employed



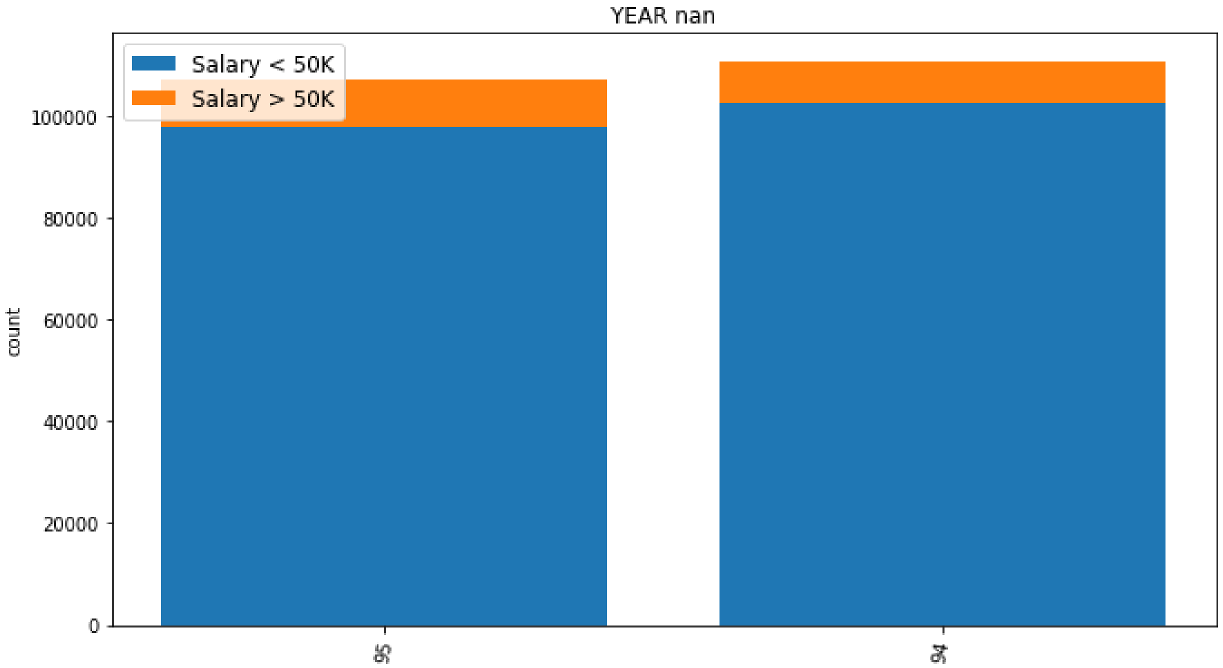


VETYN Veteran's benefits

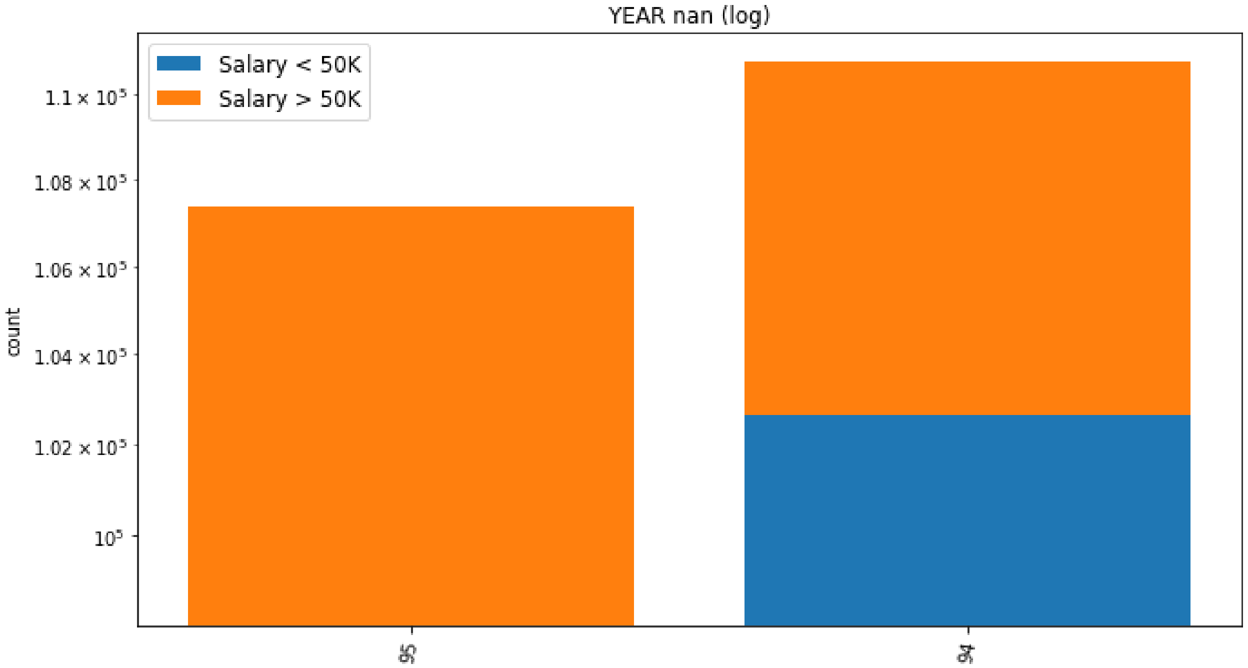




YEAR



YEAR nan
Salary < 50K
Salary > 50K
count
100000
80000
60000
40000
20000
0
95
94



YEAR nan (log)
Salary < 50K
Salary > 50K
count
1.1 × 10^5
1.08 × 10^5
1.06 × 10^5
1.04 × 10^5
1.02 × 10^5
10^5
95
94