dataiku

Benoit Gautier

19 february 2016

Contents

Short summary **1**

Descriptive statistics **1**

Quantitative variables 1

Qualitative variables 4

Logistic regression **21**

Decision tree **26**

Short summary

The purpose of this study is to predict whether the income level of a person is over $50,000 a year from the US Census dataset. These data contain 42 social and economic characteristics such as age, sex, taxable income amount. However it was recommended not to use 2 features "instance weight" and "year". Thus there are ignored in this analysis. To perform the prediction task, the data were split into a training set (199523 rows) and a test set (99762 rows). The variable of interest includes two modalities (- 50000., 50000+.). Two supervised analyses were tested in this report: a logistic regression and a decision tree.

Descriptive statistics

First, a descriptive analysis of the variables was conducted. The results are displayed according to the type of the variable (quantitative or qualitative)

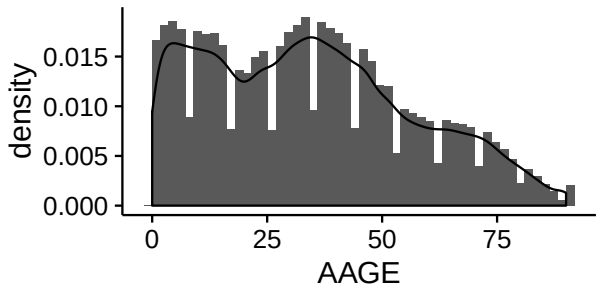
Quantitative variables

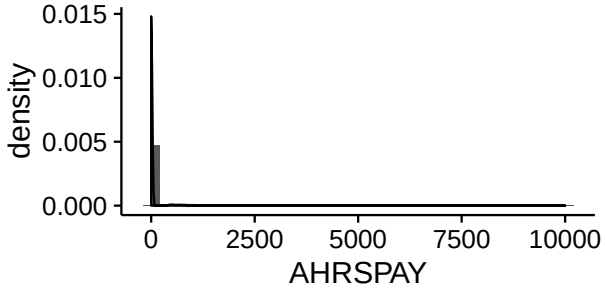
The table below outputs some basic statistics of the quantitative variables.

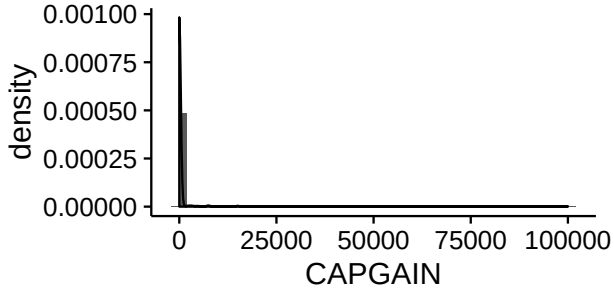
Table 1: Summary statistics

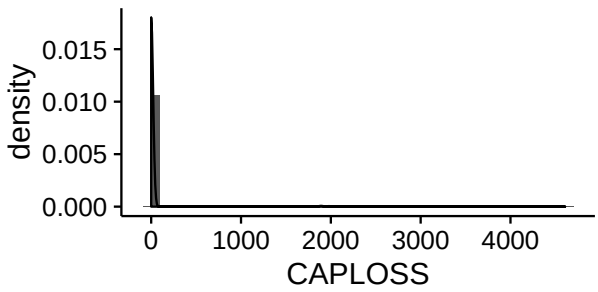
	# NA	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
AAGE	0	0	15	33	34.49	50	90
AHRSPAY	0	0	0	0	55.43	0	9999
CAPGAIN	0	0	0	0	434.7	0	1e+05
CAPLOSS	0	0	0	0	37.31	0	4608
DIVVAL	0	0	0	0	197.5	0	1e+05
NOEMP	0	0	0	1	1.956	4	6
WKSWORK	0	0	0	8	23.17	52	52

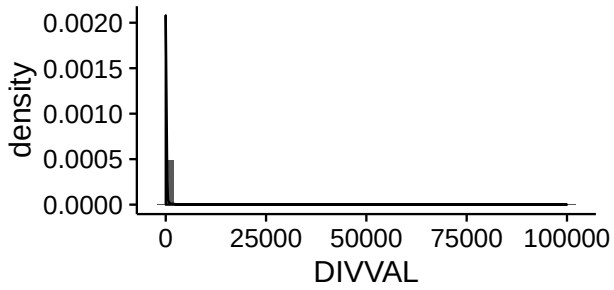
The variables wage per hour (AHRSPAY), capital gains (CAPGAIN), capital losses (CAPLOSS) and dividends from stocks (DIVVAL) have a skewed distribution. A log transformation could have been done. However it was decided to categorize them.

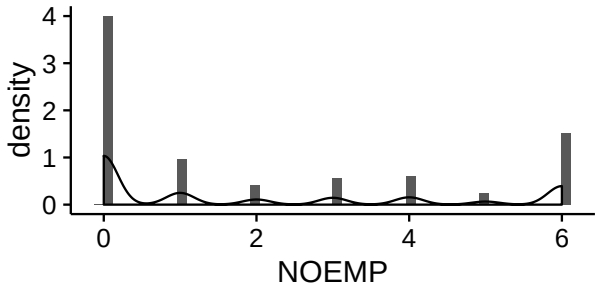


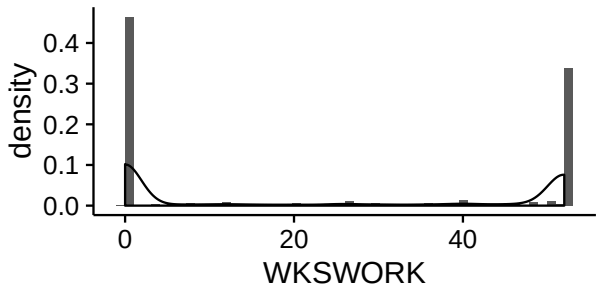


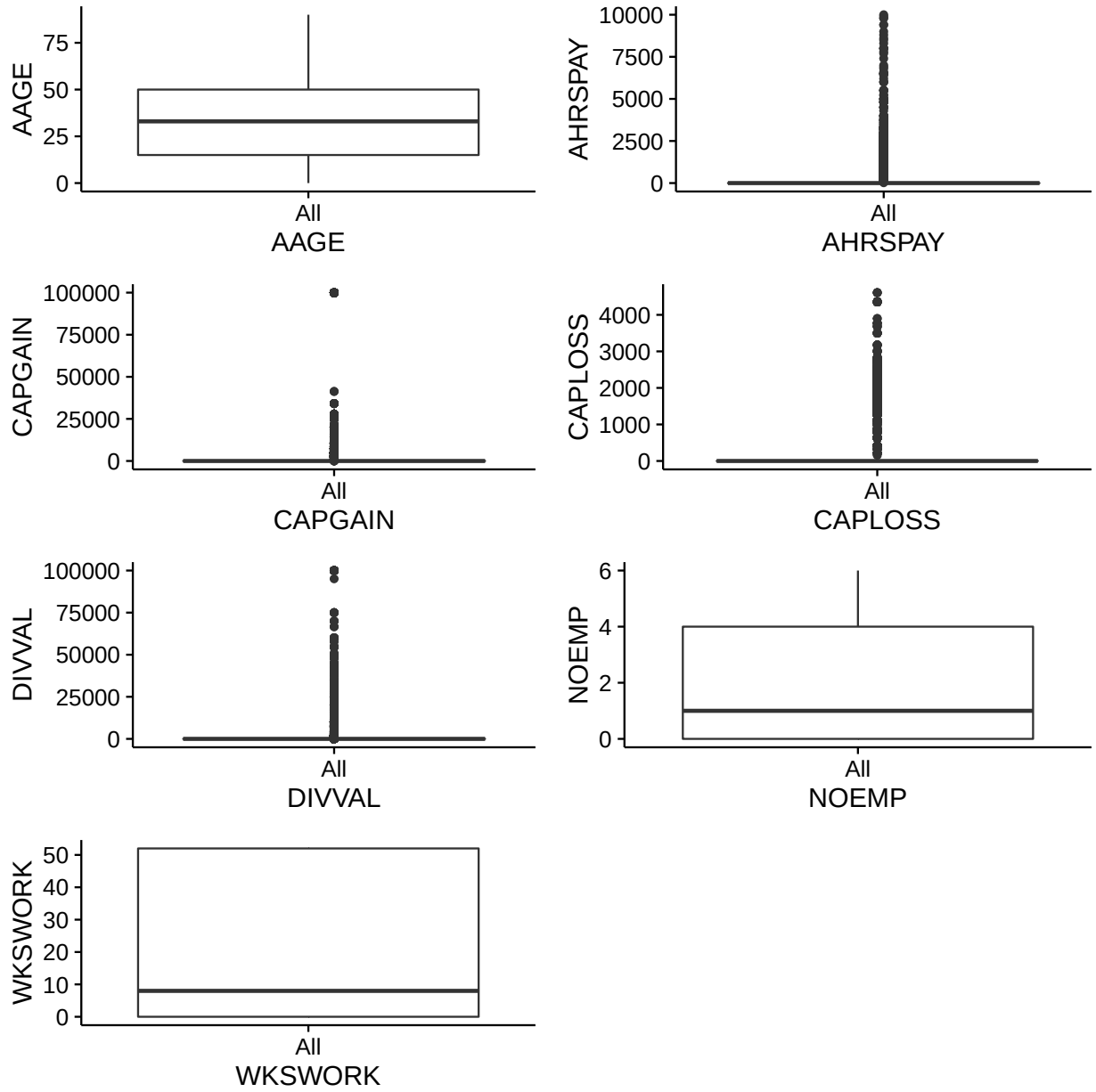












The variables mentioned above (AHRSPAY, CAPGAIN, CAPLOSS, and DIVVAL) with a skewed distribution were categorized into two categories ("= 0" or "> 0").

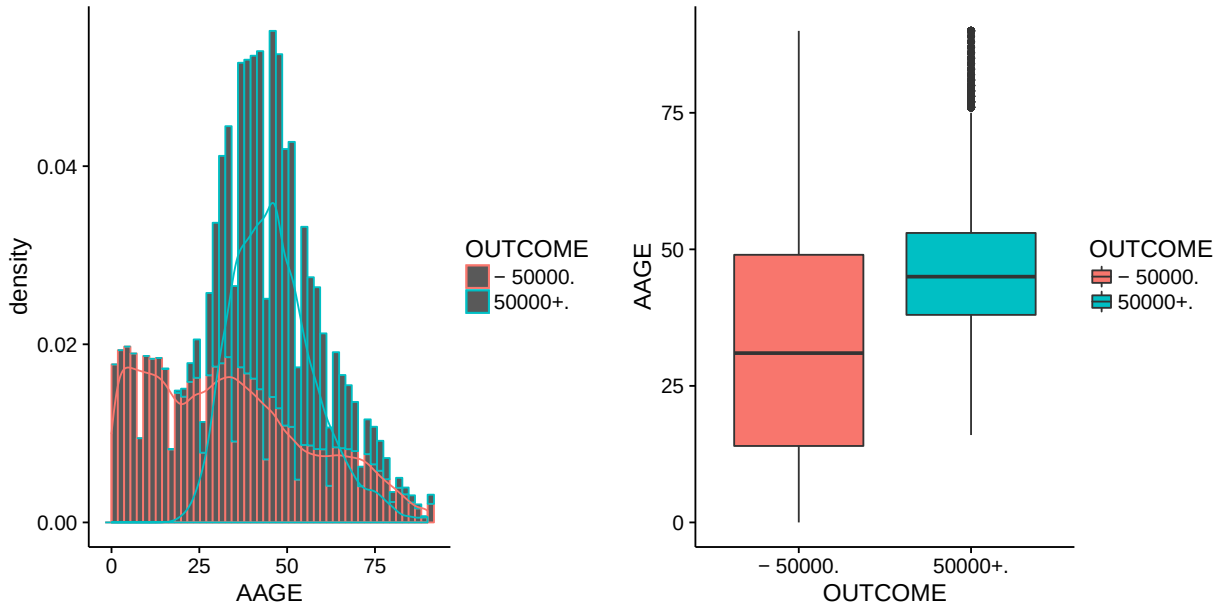
Two additional variables WKSWORK and NOEMP were categorized.

- Weeks work in year (WKSWORK) was categorized into 3 categories: "no week", "less than a year", "a year".
- Num persons worked for employer (NOEMP) was categorized into 2 categories: "no one" and "more than one".

Finally one variable age was kept as continuous variable.

In order to investigate whether there is an association between the variable "age" and the outcome, we looked at the distribution of this variable within each modality of the outcome variable. The plot below shows that a

person between 16yo and 65yo is more likely to earn $50,000 or more a year. In the meantime, no one under 16yo earns more than $50,000 a year.



A parametric test (Student test) was applied to highlight a difference of age between groups. As expected, the test is significant.

Table 2: Welch Two Sample t-test: `training$AAGE` by `training$OUTCOME`

Test statistic	df	P value	Alternative hypothesis
-105.9	19021	0 * * *	two.sided

// Finally, we created a qualitative variable for the variable “AAGE” according to the legal // working age and the retirement age. The variable is defined as “younger than 16 yo”, // “between 17yo and 65yo” and “older than 66yo”

Qualitative variables

The table below reports the number of missing data for each qualitative variable along with the number of modalities.

	# NA	# Modalities
ACLSWKR	0	9
ADTIND	0	52
ADTOCC	0	47
AHGA	0	17
AHSCOL	0	3
AMARITL	0	7
AMJIND	0	24
AMJOCC	0	15
ARACE	0	5
AREORGN	0	10

	# NA	# Modalities
ASEX	0	2
AUNMEM	0	3
AUNTYPE	0	6
AWKSTAT	0	8
FILESTAT	0	6
GRINREG	0	6
GRINST	708	50
HHDFMX	0	38
HHDREL	0	8
MIGMTR1	99696	9
MIGMTR3	99696	8
MIGMTR4	99696	9
MIGSAME	0	3
MIGSUN	99696	3
PARENT	0	5
PEFNTVTY	6713	42
PEMNTVTY	6119	42
PENATVTY	3393	42
PRCITSHP	0	5
SEOTR	0	3
VETQVA	0	3
VETYN	0	3

To facilitate the analysis, we removed all variables with missing data or with too many categories. Besides some of the modalities were brought together and merged into a single modality.

Is there an association between some qualitative variables and the outcome? The Chi-2 test testing for the association between 2 qualitative variables indicates a significant association between the outcome and each qualitative variable present in the data. On the barplot, the horizontal line represents the proportion of the class "-50000" in the data. Picking out the variable ASEX, we observe a less proportion of "50000+" in the modality "Female" than in the modality "Male".

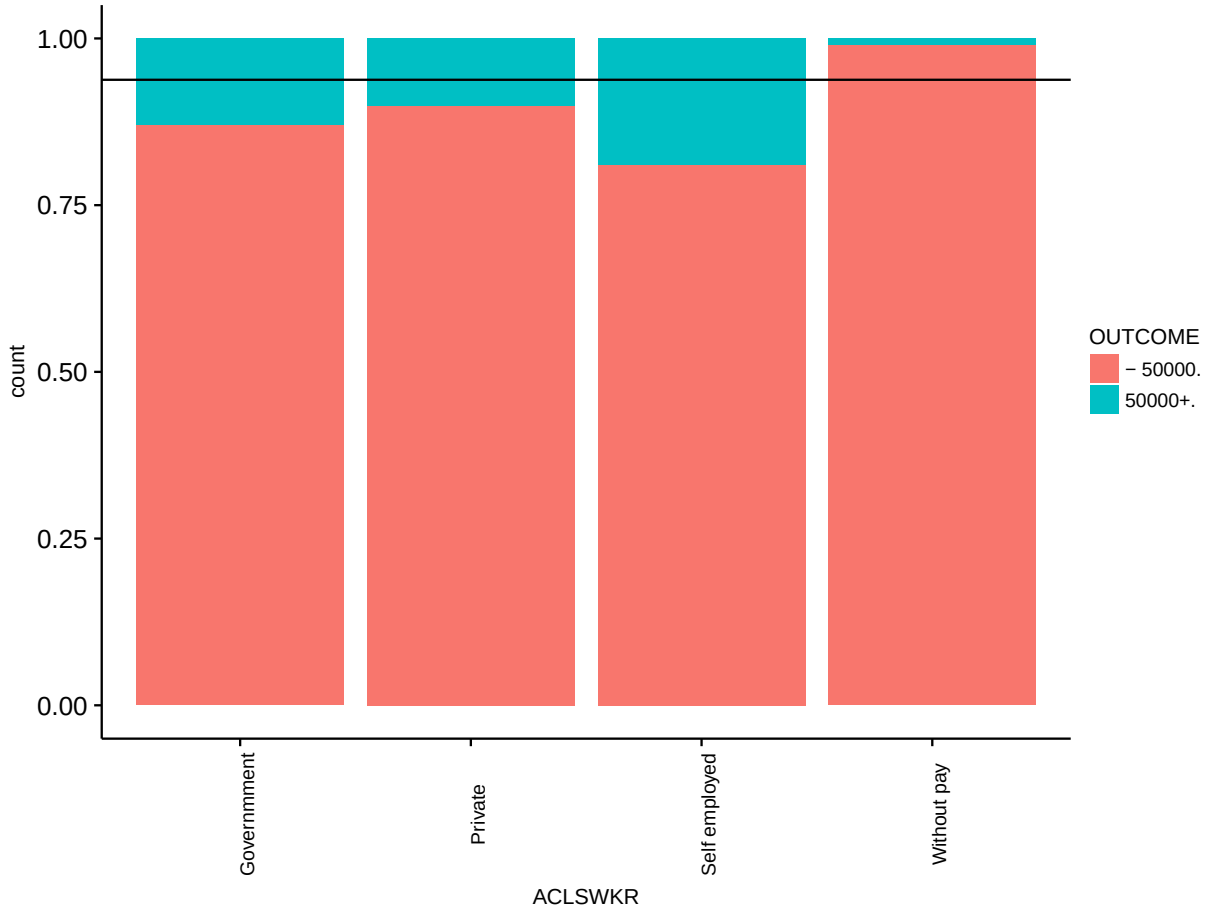


Table 4: Contingency table

	- 50000.	50000+.
Governmment	13007	1929
Private	64706	7322
Self employed	9486	2224
Without pay	99942	907

Table 5: Pearson's Chi-2 test

Test statistic	df	P value
11263	3	0 * * *

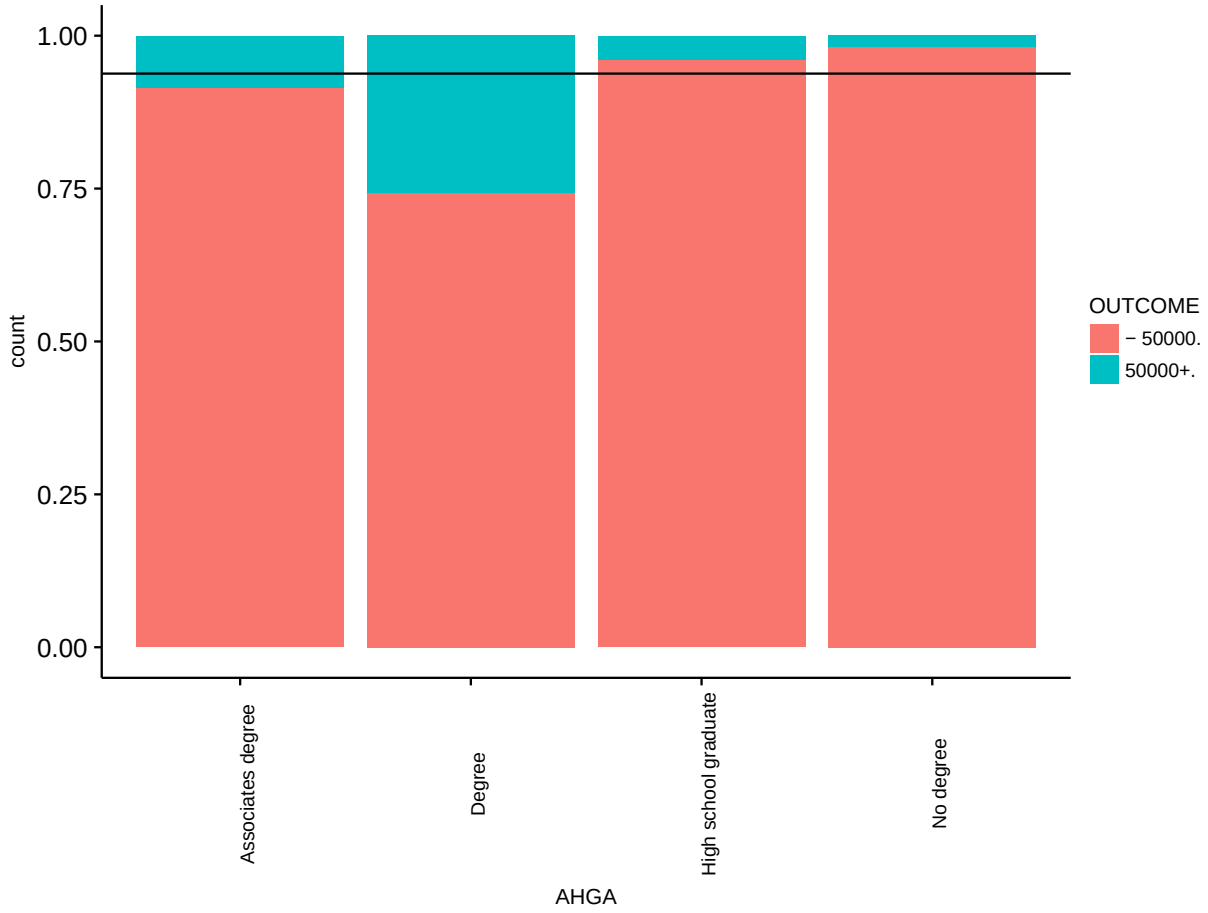


Table 6: Contingency table

	- 50000.	50000+.
Associates degree	8896	825
Degree	21883	7579
High school graduate	46528	1879
No degree	109834	2099

Table 7: Pearson's Chi-2 test

Test statistic	df	P value
23427	3	0 * * *

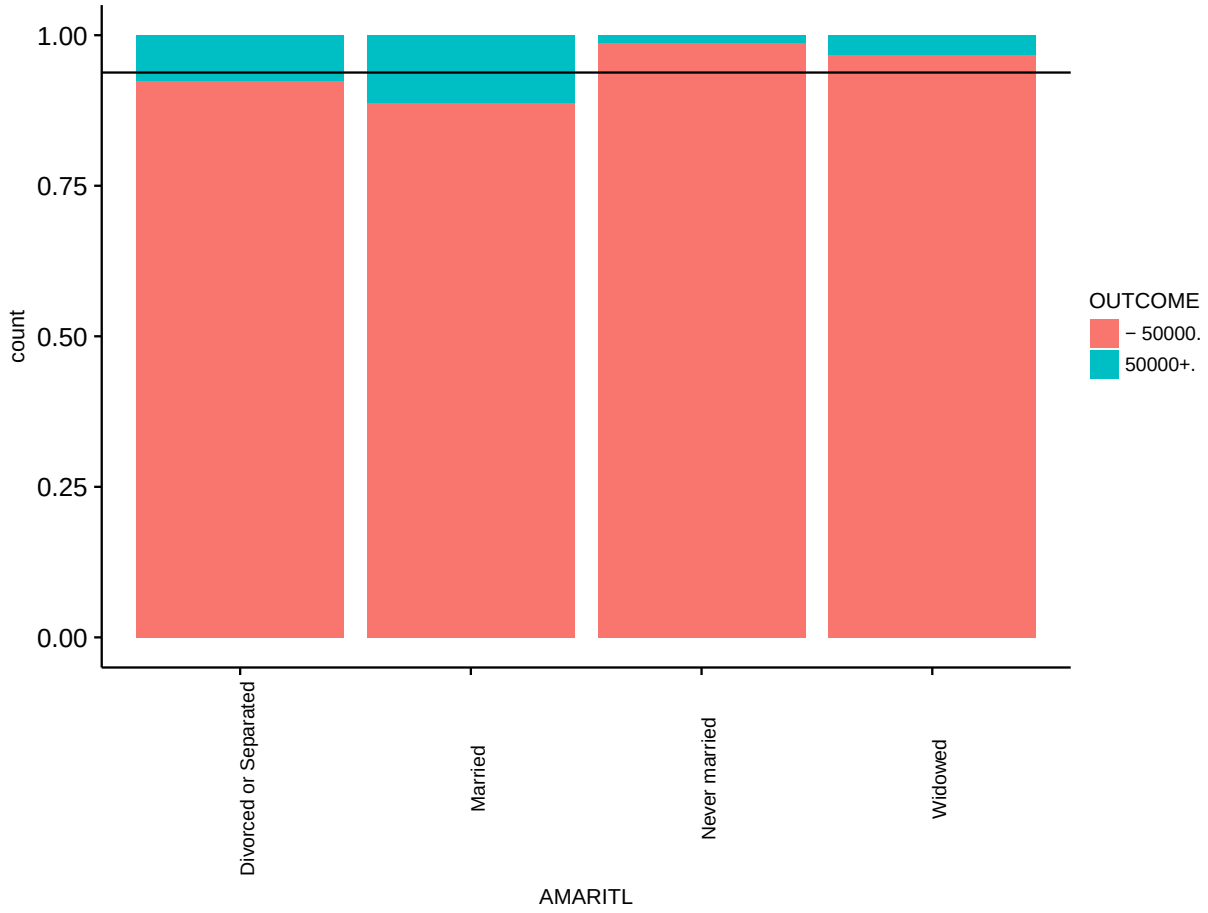


Table 8: Contingency table

	- 50000.	50000+.
Divorced or Separated	14946	1224
Married	76694	9711
Never married	85368	1117
Widowed	10133	330

Table 9: Pearson's Chi-2 test

Test statistic	df	P value
7568	3	0 * * *

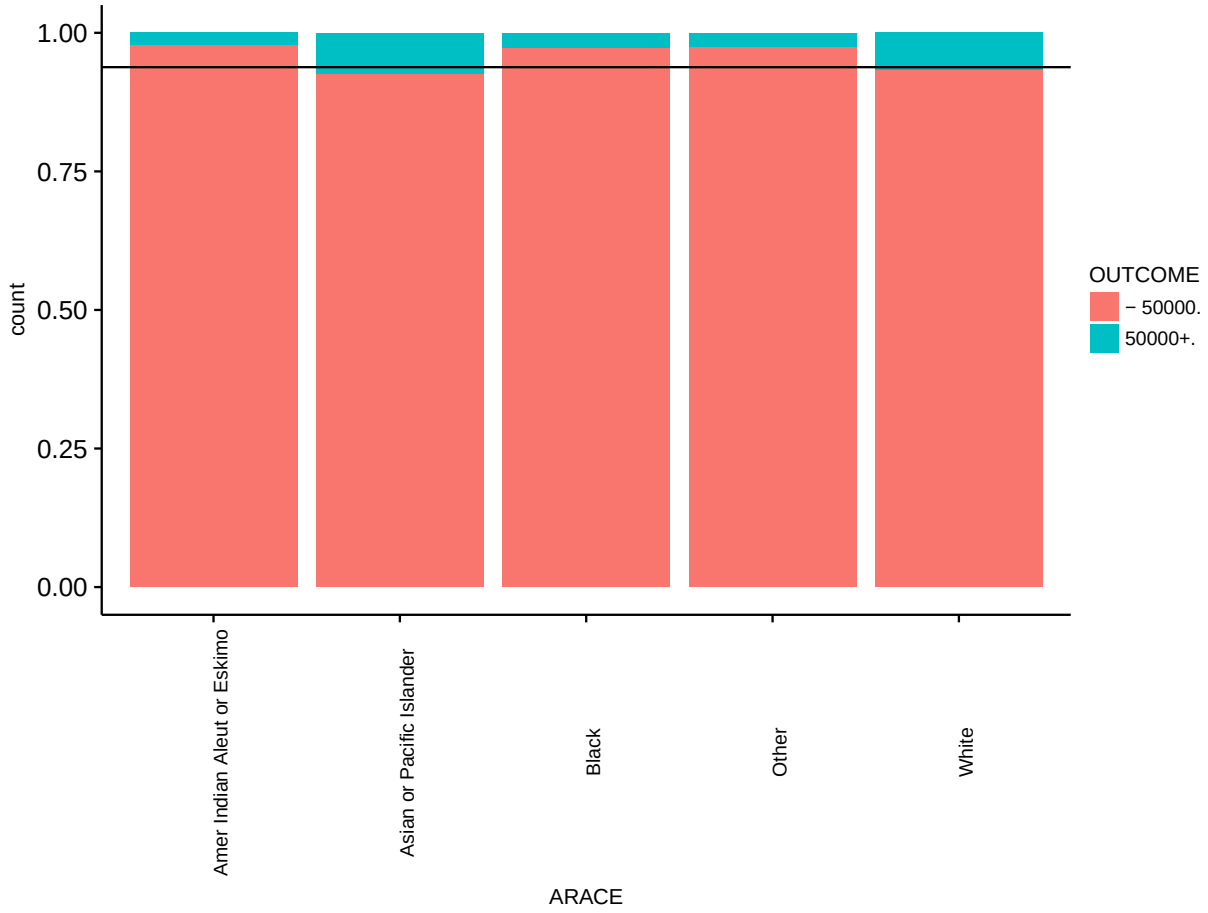


Table 10: Contingency table

	- 50000.	50000+.
Amer Indian Aleut or Eskimo	2202	49
Asian or Pacific Islander	5405	430
Black	19875	540
Other	3566	91
White	156093	11272

Table 11: Pearson's Chi-2 test

Test statistic	df	P value
688.4	4	1.152e-147 * * *

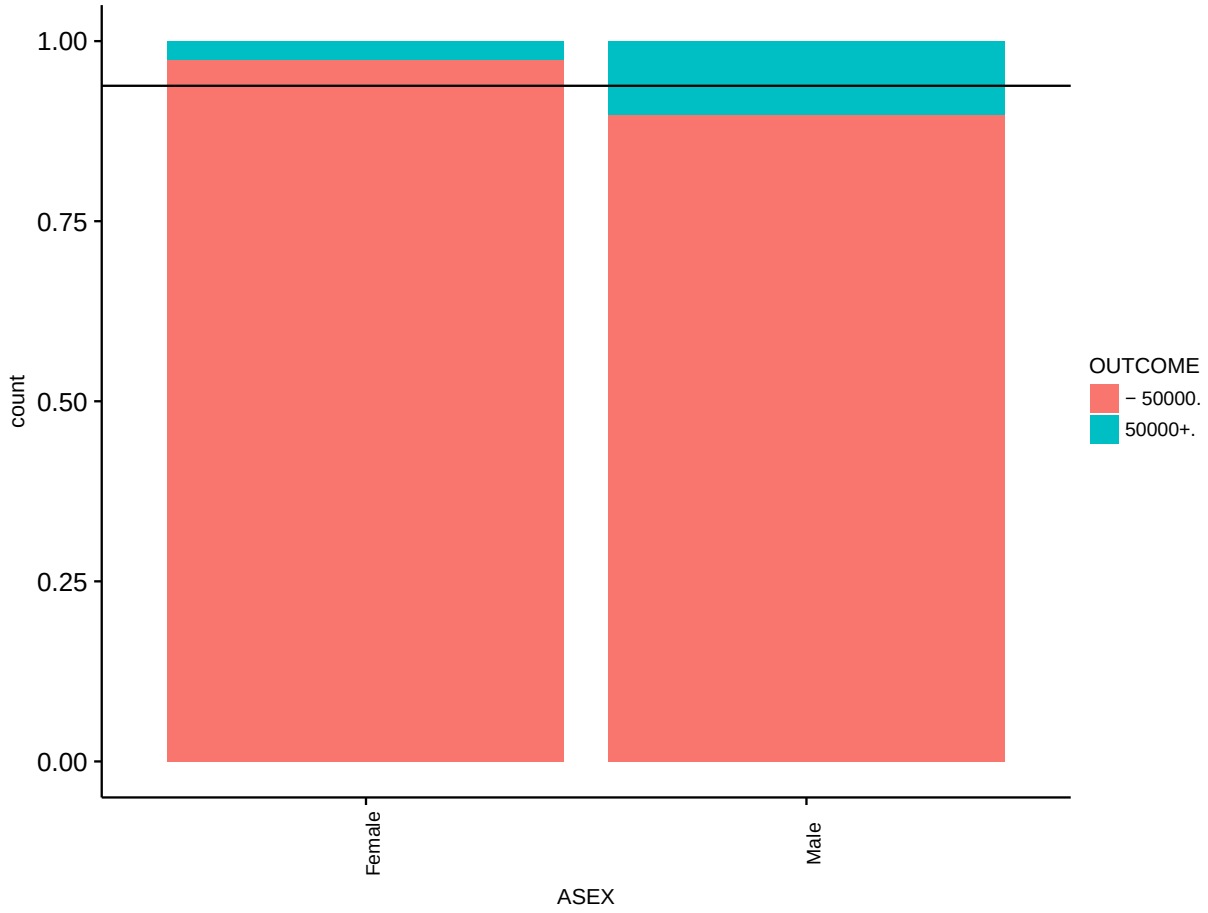


Table 12: Contingency table

	- 50000.	50000+.
Female	101321	2663
Male	85820	9719

Table 13: Pearson's Chi-2 test

Test statistic	df	P value
4955	1	0 * * *

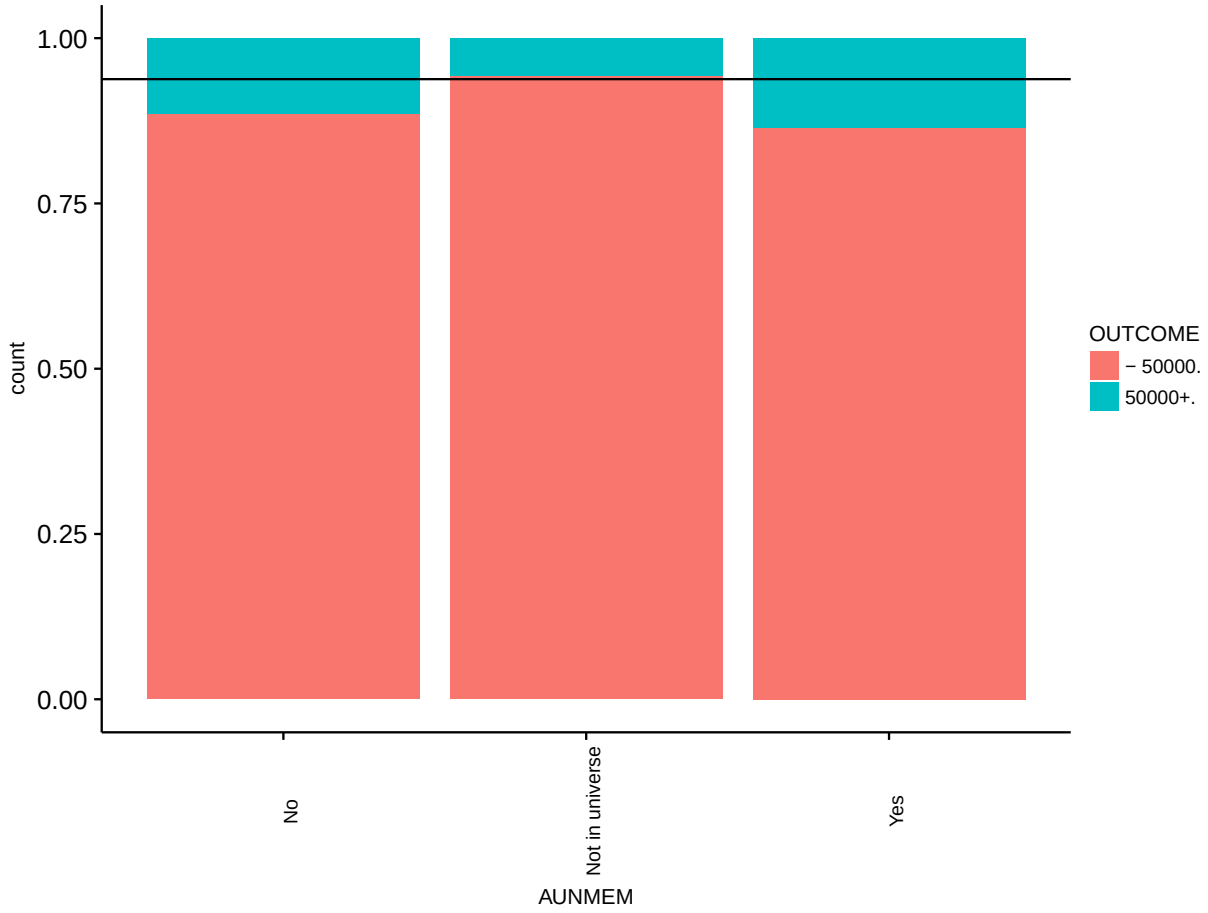


Table 14: Contingency table

	- 50000.	50000+.
No	14212	1822
Not in universe	170311	10148
Yes	2618	412

Table 15: Pearson's Chi-2 test

Test statistic	df	P value
1122	2	1.973e-244 * * *

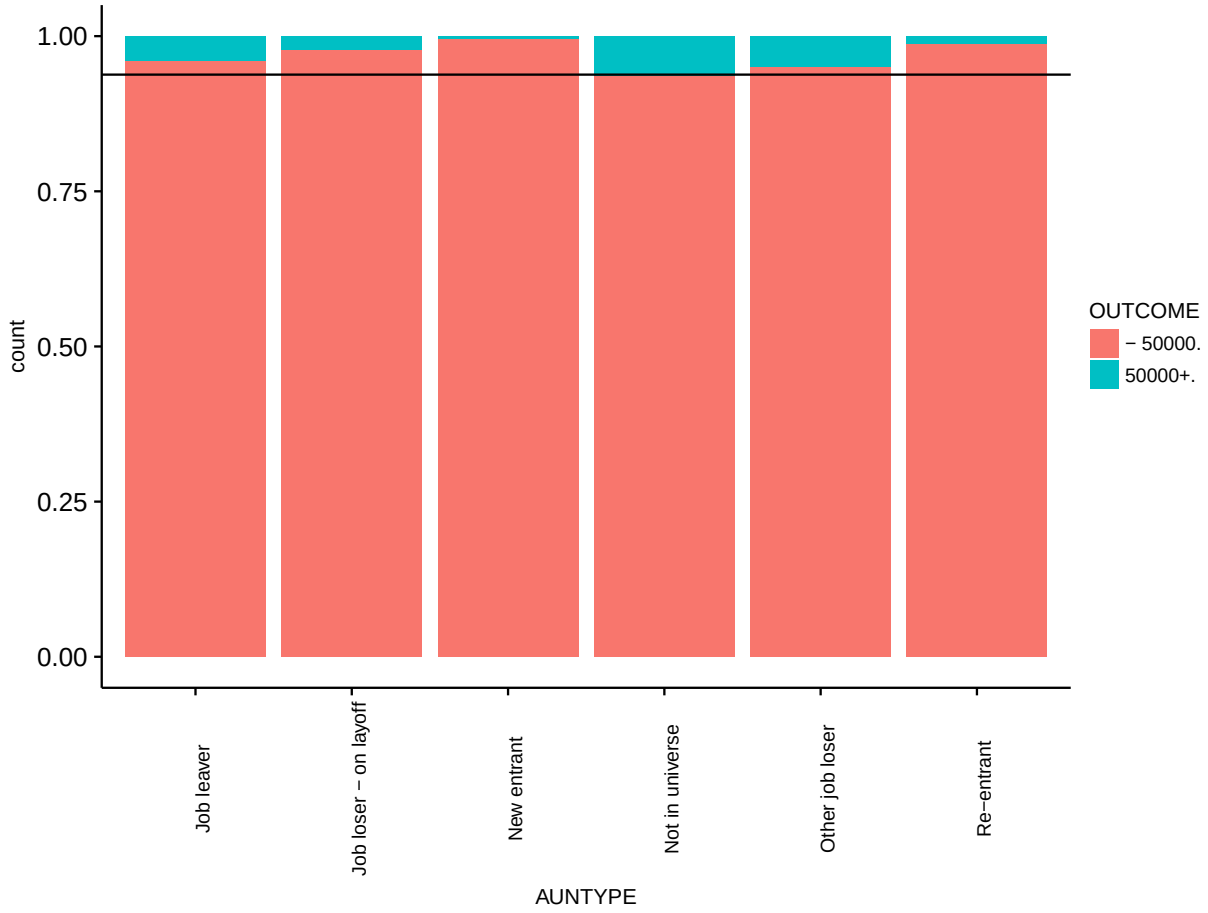


Table 16: Contingency table

	- 50000.	50000+.
Job leaver	574	24
Job loser - on layoff	954	22
New entrant	437	2
Not in universe	181241	12212
Other job loser	1939	99
Re-entrant	1996	23

Table 17: Pearson's Chi-2 test

Test statistic	df	P value
155.3	5	1.012e-31 * * *

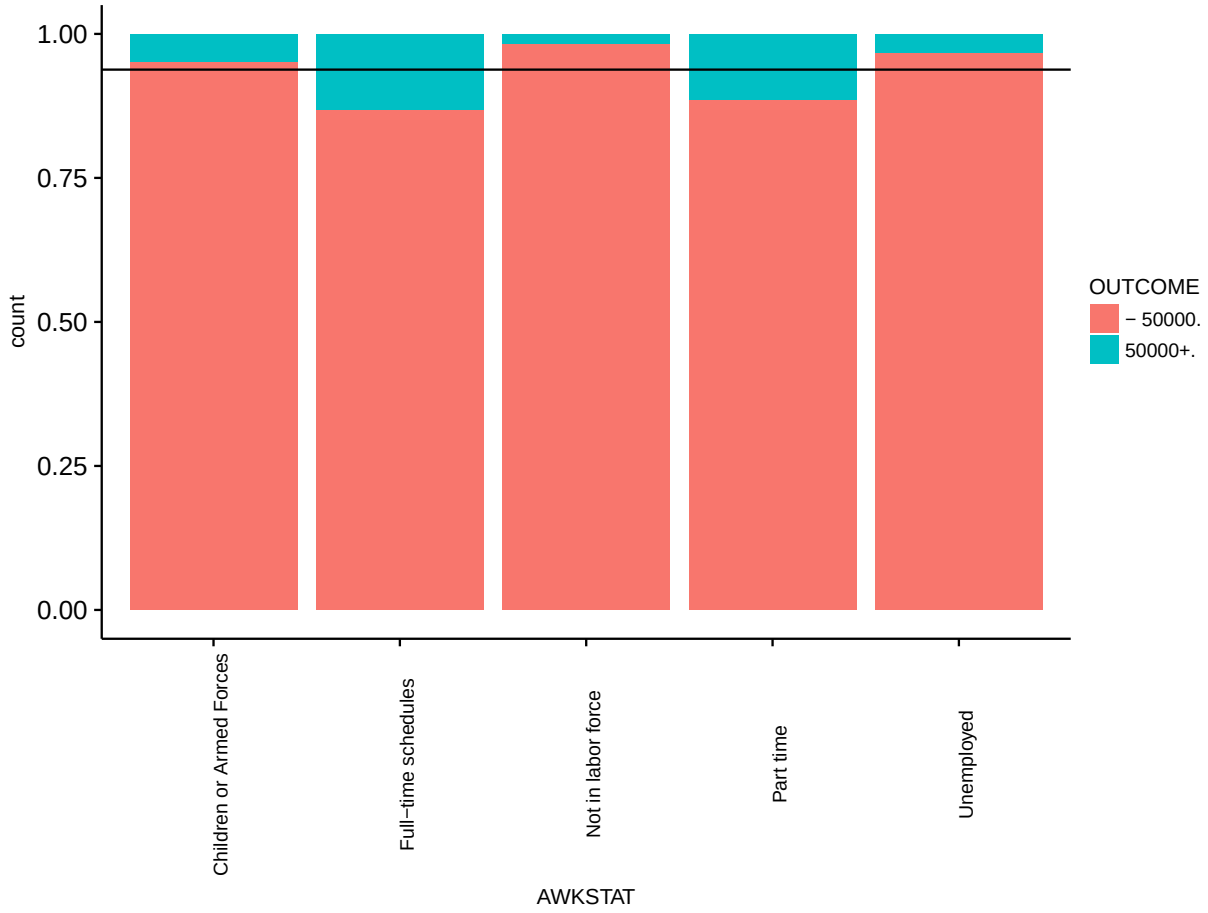


Table 18: Contingency table

	- 50000.	50000+.
Children or Armed Forces	117895	5874
Full-time schedules	35370	5366
Not in labor force	26346	462
Part time	4477	579
Unemployed	3053	101

Table 19: Pearson's Chi-2 test

Test statistic	df	P value
5063	4	0 * * *

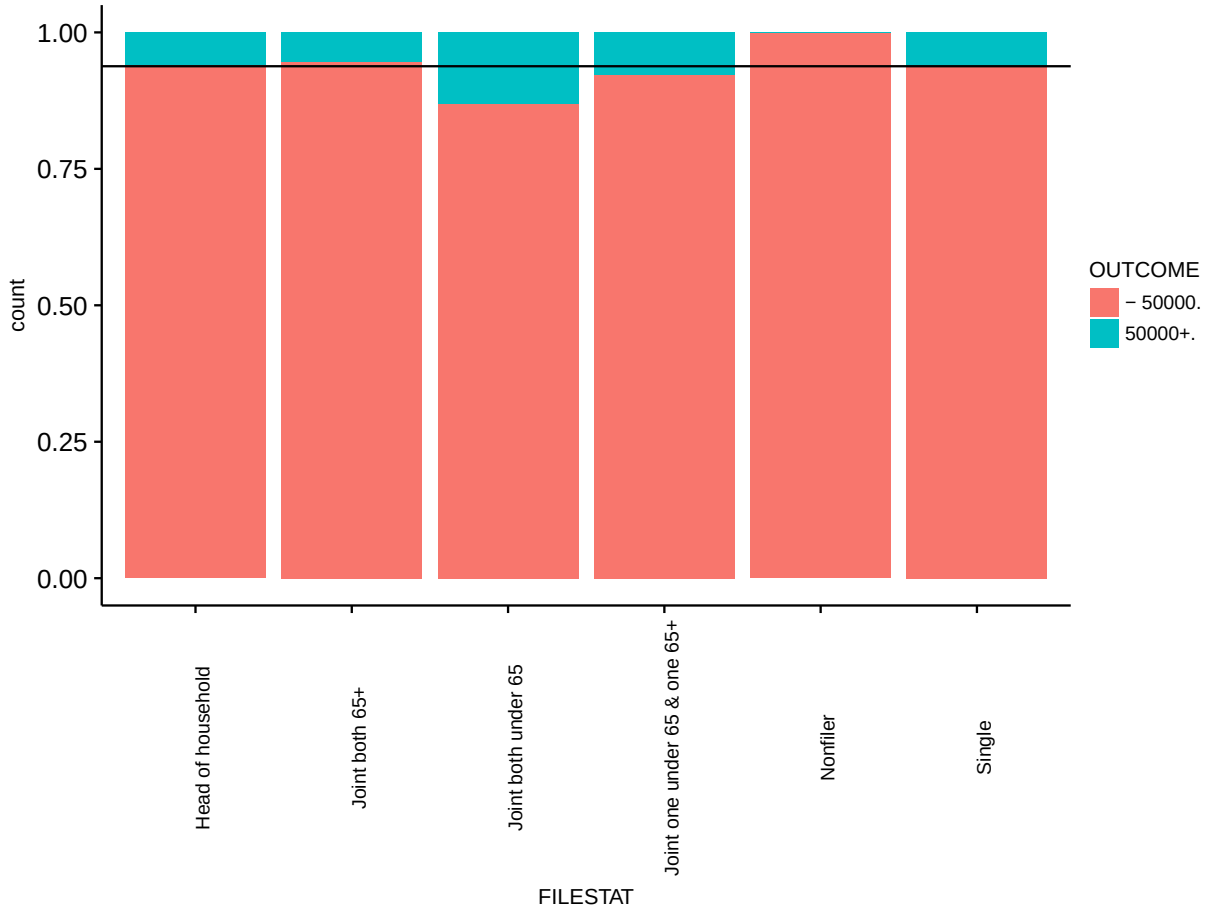


Table 20: Contingency table

	- 50000.	50000+.
Head of household	6978	448
Joint both 65+	7882	450
Joint both under 65	58530	8853
Joint one under 65 & one 65+	3566	301
Nonfiler	75059	35
Single	35126	2295

Table 21: Pearson's Chi-2 test

Test statistic	df	P value
10484	5	0 * * *

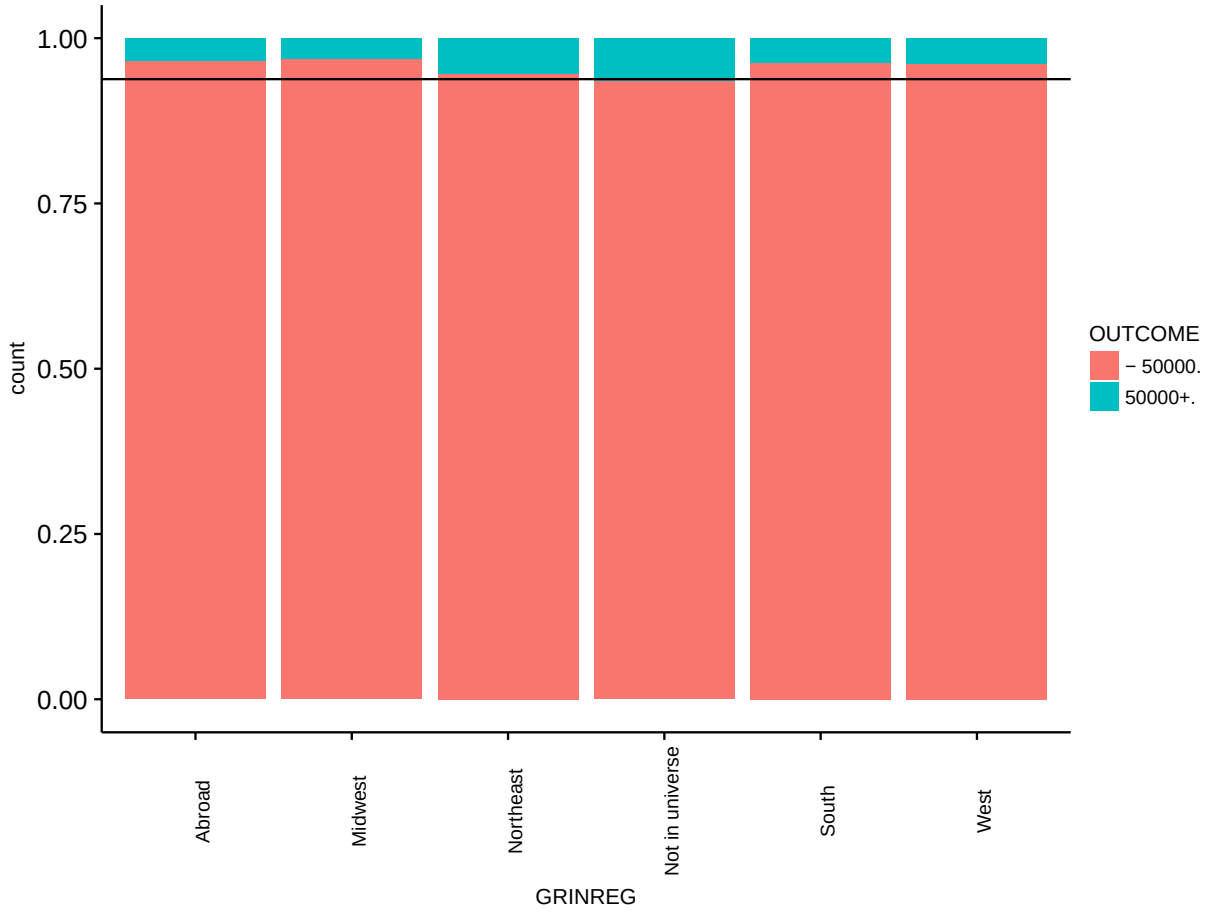


Table 22: Contingency table

	- 50000.	50000+.
Abroad	512	18
Midwest	3466	109
Northeast	2558	147
Not in universe	171986	11764
South	4705	184
West	3914	160

Table 23: Pearson's Chi-2 test

Test statistic	df	P value
169.8	5	8.172e-35 * * *

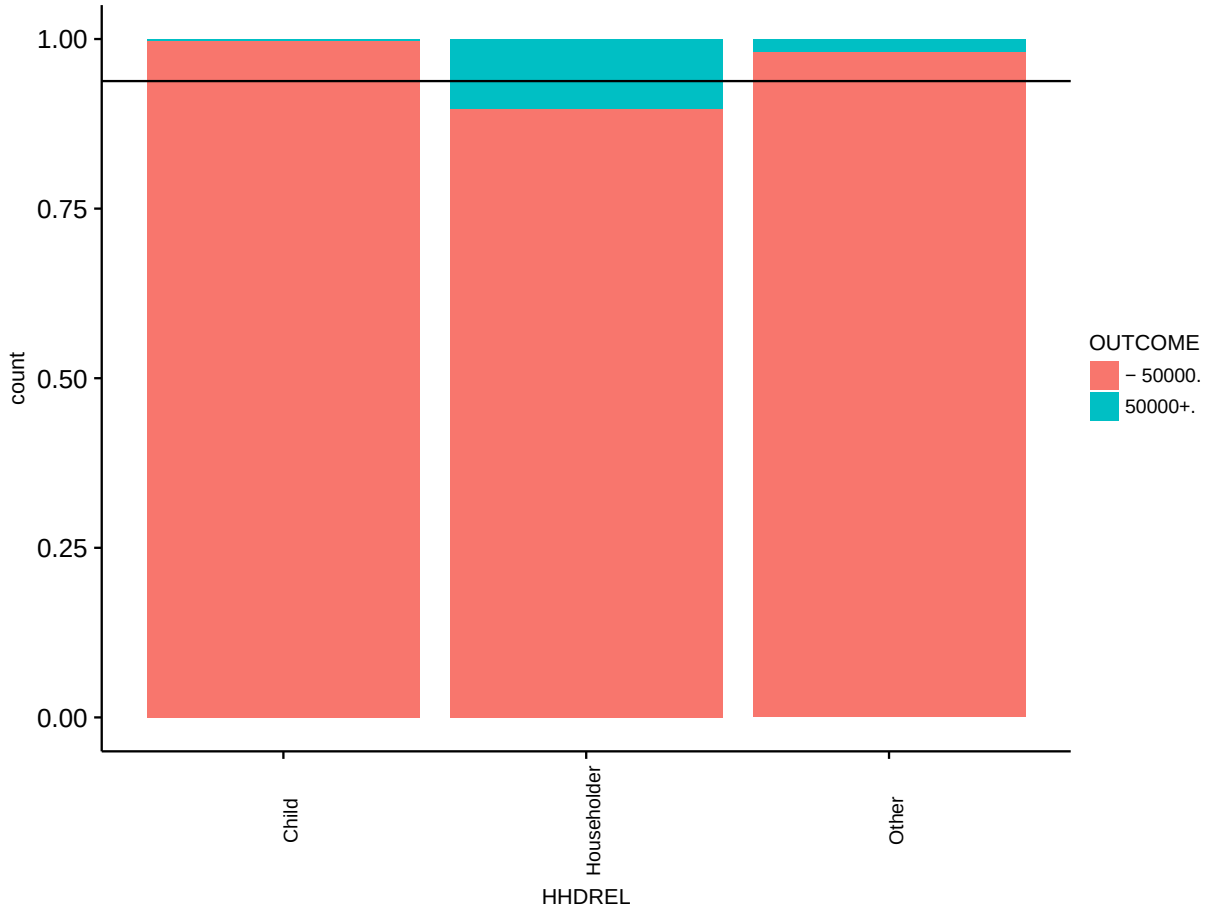


Table 24: Contingency table

	- 50000.	50000+.
Child	64775	128
Householder	105251	11933
Other	17115	321

Table 25: Pearson's Chi-2 test

Test statistic	df	P value
7781	2	0 * * *

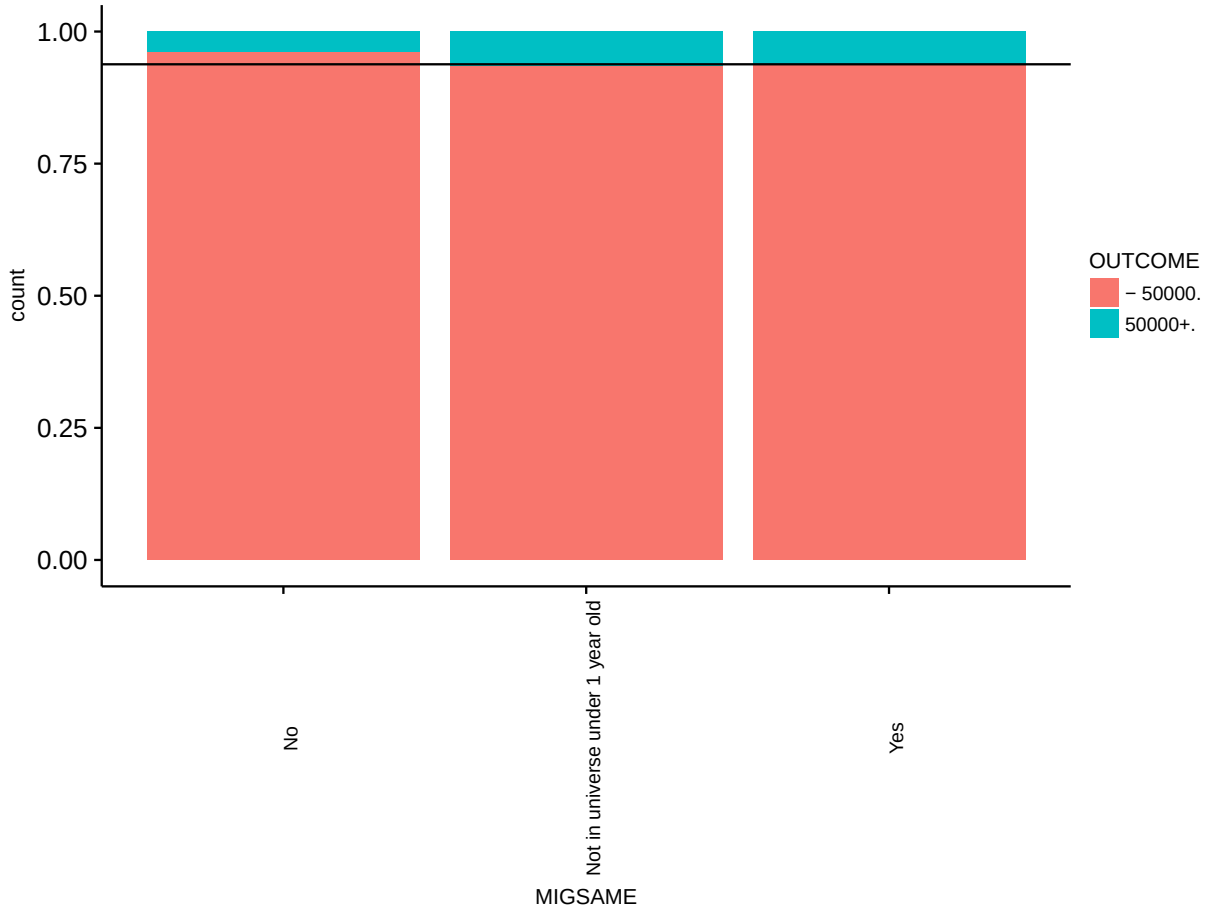


Table 26: Contingency table

	- 50000.	50000+.
No	15155	618
Not in universe under 1 year old	94669	6543
Yes	77317	5221

Table 27: Pearson's Chi-2 test

Test statistic	df	P value
155.5	2	1.707e-34 * * *

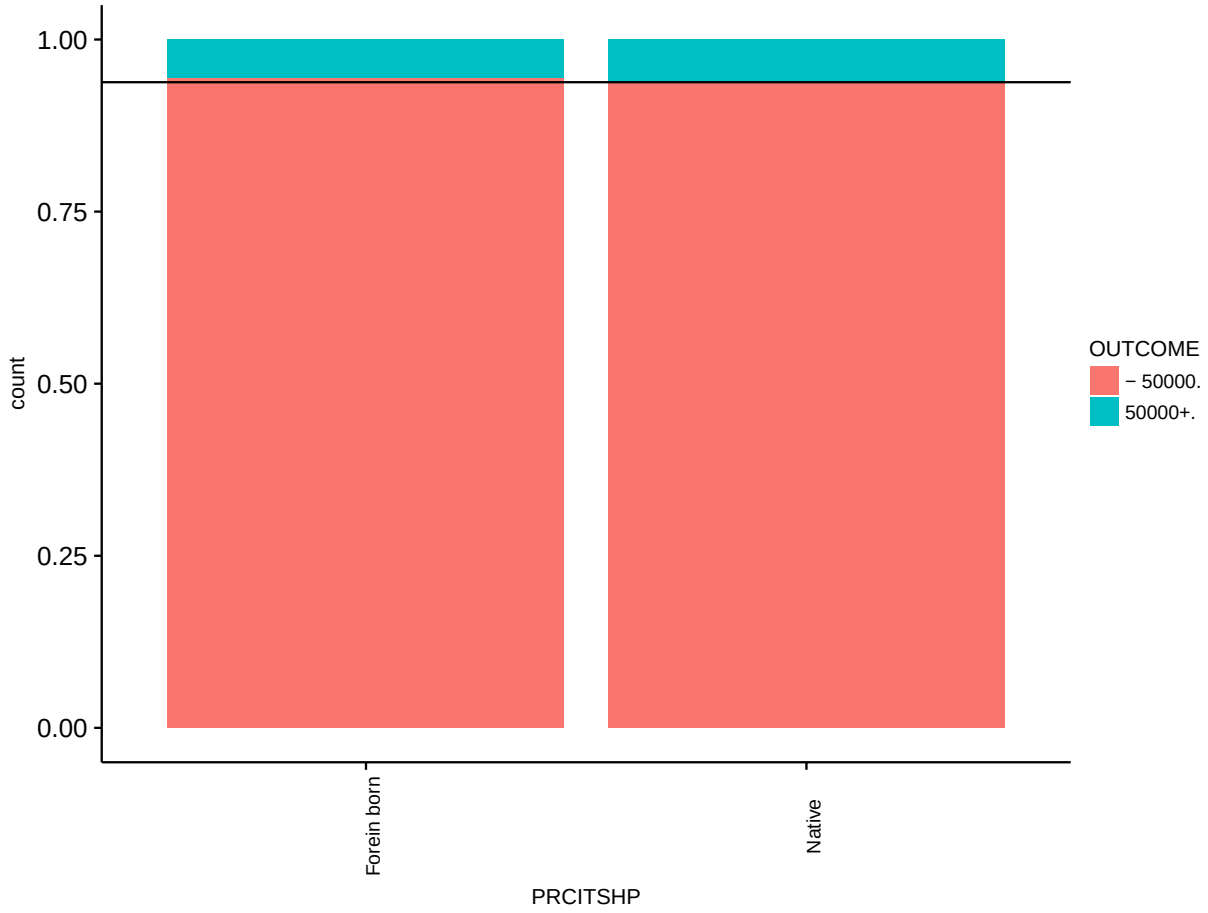


Table 28: Contingency table

	- 50000.	50000+.
Forein born	18184	1072
Native	168957	11310

Table 29: Pearson's Chi-2 test

Test statistic	df	P value
14.82	1	0.0001185 * * *

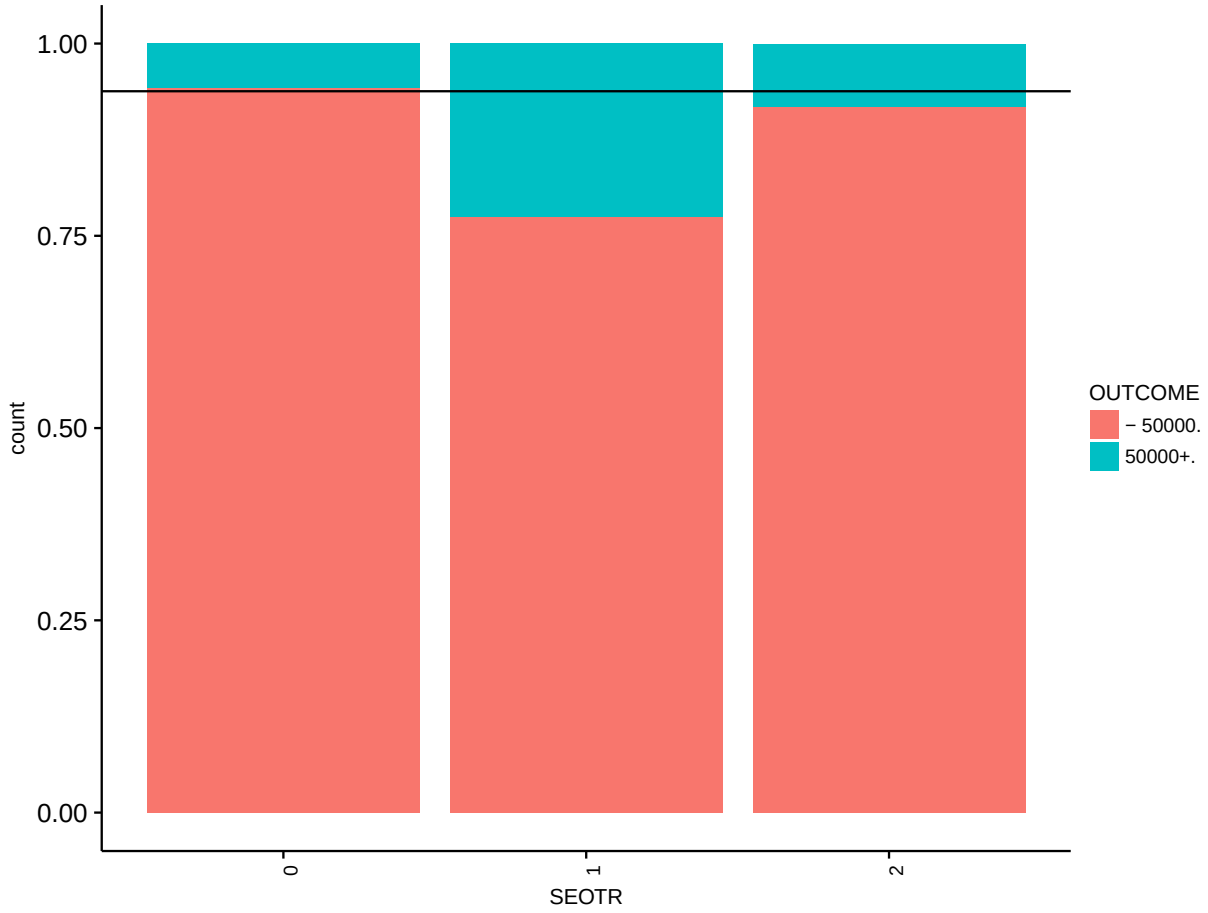


Table 30: Contingency table

	- 50000.	50000+.
0	170220	10452
1	2089	609
2	14832	1321

Table 31: Pearson's Chi-2 test

Test statistic	df	P value
1404	2	1.049e-305 * * *

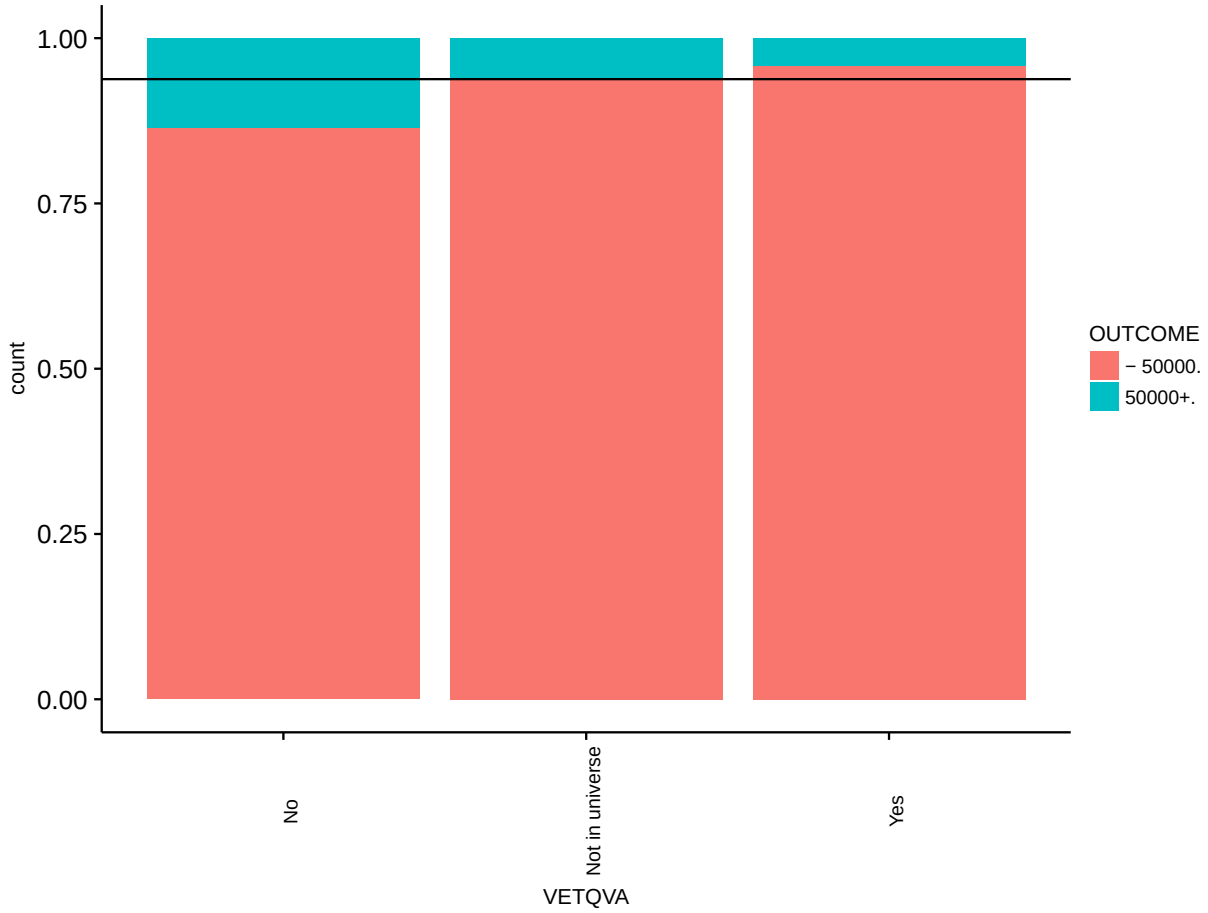


Table 32: Contingency table

	- 50000.	50000+.
No	1378	215
Not in universe	185388	12151
Yes	375	16

Table 33: Pearson's Chi-2 test

Test statistic	df	P value
149.5	2	3.461e-33 * * *

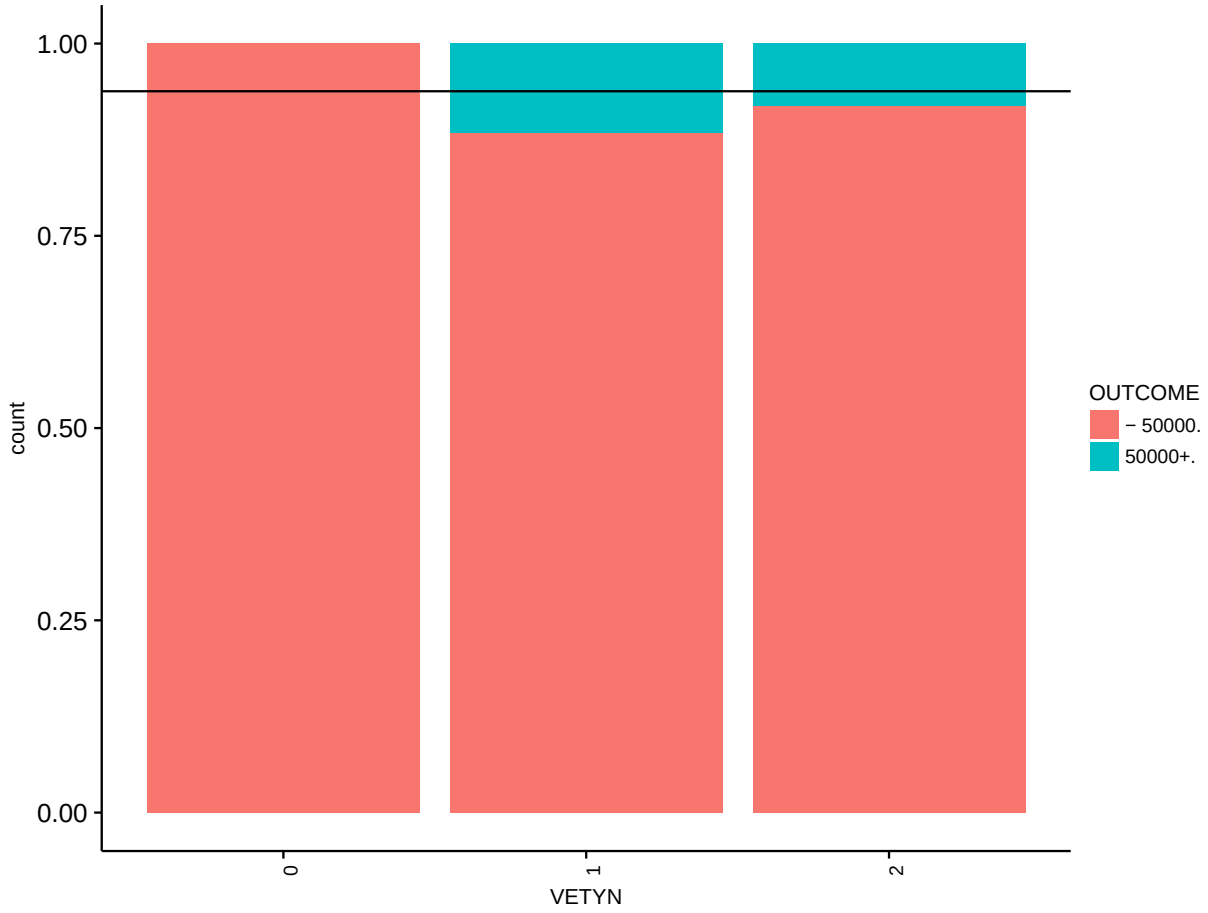


Table 34: Contingency table

	- 50000.	50000+.
0	47409	0
1	1753	231
2	137979	12151

Table 35: Pearson's Chi-2 test

Test statistic	df	P value
4157	2	0 * * *

Logistic regression

A logistic regression was performed with observation weights set at 1 (no weight). In the model only one continuous variable (age) was included. Most of the variables incorporated in the model are significant.

##

	Estimate	Std. Error	z value	Pr(>|z|)
AAGE	0.0293	0.001165	25.16	1.086e-139
ACLSWKRPrivate	0.2272	0.03311	6.862	6.798e-12
ACLSWKRSelf employed	0.3784	0.04175	9.065	1.249e-19
ACLSWKRWithout pay	-0.1969	0.07933	-2.482	0.01306
AHGADegree	1.091	0.04315	25.28	5.968e-141
AHGAHigh school graduate	-0.63	0.04712	-13.37	8.814e-41
AHGANo degree	-0.4611	0.04673	-9.866	5.849e-23
AMARITLMarried	-0.2253	0.1233	-1.827	0.06771
AMARITLNever married	-0.4144	0.05311	-7.801	6.137e-15
AMARITLWidowed	-0.1765	0.07828	-2.255	0.02414
ARACEAsian or Pacific Islander	0.4517	0.1725	2.618	0.008841
ARACEBlack	0.07807	0.1651	0.4727	0.6364
ARACEOther	0.05658	0.1993	0.2839	0.7765
ARACEWhite	0.2941	0.1583	1.859	0.06309
ASEXMale	1.256	0.02615	48.03	0
AUNMEMNot in universe	-0.2489	0.03787	-6.571	4.982e-11
AUNMEMYes	0.2555	0.06981	3.66	0.0002522
AUNTYPEJob loser - on layoff	-0.6541	0.3239	-2.019	0.04344
AUNTYPENew entrant	1.952	0.7783	2.508	0.01213
AUNTYPENot in universe	-0.1141	0.2467	-0.4624	0.6438
AUNTYPEOther job loser	-0.1865	0.2621	-0.7117	0.4766
AUNTYPERe-entrant	-0.4303	0.325	-1.324	0.1855
AWKSTATFull-time schedules	0.4924	0.2125	2.317	0.02049
AWKSTATNot in labor force	0.4739	0.2128	2.227	0.02596
AWKSTATPart time	0.3857	0.2183	1.767	0.07723

```
##
##      **AWKSTATUnemployed**         0.2192      0.2541     0.8628     0.3883
##
##    **FILESTATJoint both 65+**     -0.7404      0.1456     -5.086   3.649e-07
##
##    **FILESTATJoint both under     0.07207      0.1295     0.5565     0.5779
##              65**
##
##  **FILESTATJoint one under 65 &   -0.6161      0.149      -4.134   3.558e-05
##           one 65+**
##
##       **FILESTATNonfiler**         -2.105      0.1907     -11.04   2.559e-28
##
##        **FILESTATSingle**         -0.1946     0.06386     -3.047    0.002311
##
##        **GRINREGMidwest**         -0.6136      0.3113     -1.971    0.04873
##
##       **GRINREGNortheast**        0.01666      0.3082     0.05405    0.9569
##
##    **GRINREGNot in universe**     0.04438      0.2921     0.1519     0.8792
##
##         **GRINREGSouth**          -0.1435      0.3037     -0.4725    0.6365
##
##         **GRINREGWest**           -0.08654     0.3058     -0.283     0.7772
##
##      **HHDRELHouseholder**         1.265      0.09779      12.94   2.784e-38
##
##         **HHDRELOther**            0.7072      0.1112      6.361    2.01e-10
##
##  **MIGSAMENot in universe under   -0.3616      0.2114     -1.711    0.08709
##          1 year old**
##
##        **PRCITSHPNative**          0.215      0.04386      4.901   9.515e-07
##
##           **SEOTR1**               0.474       0.0573      8.273   1.308e-16
##
##           **SEOTR2**              -0.1464     0.03629     -4.033    5.5e-05
##
##    **VETQVANot in universe**       -10.2       48.03     -0.2123    0.8319
##
##          **VETQVAYes**            -0.1627      0.3063     -0.5312    0.5953
##
##           **VETYN2**               9.793       48.03      0.2039     0.8385
##
##       **AHRSPAY_qual> 0**         -0.7844     0.05902     -13.29    2.63e-40
##
##       **CAPGAIN_qual> 0**          1.222      0.03353      36.44   9.281e-291
##
##       **CAPLOSS_qual> 0**          1.156      0.04506      25.66   3.62e-145
##
##        **DIVVAL_qual> 0**          1.069      0.02469      43.31        0
##
##    **WKSWORK_qualless than a       -0.874     0.03524      -24.8   8.526e-136
##             year**
```

```
##
##     **WKSWORK_qualno week**        -1.843     0.07992     -23.06   1.199e-117
##
##         **(Intercept)**            -5.382      0.4409     -12.21   2.798e-34
## ------------------------------------------------------------------------------
##
## Table: Fitting generalized (binomial/logit) linear model: OUTCOME ~ .
```

The percentage of well-classified is equal to 95 %. However we observe that only the majority class (-50000.) is correctly predicted by the model (99 % for the class "-50000." and 29 % for the class "50000+") because the design of this study is unbalanced. To improve the performance of the model, a new model with observation weights set at 1 for "-50000." and 8 for "50000+" was processed.

```
##
## ------------------------------------------------------------------------------
##                               Estimate   Std. Error   z value   Pr(>|z|)
## ------------------------------- ---------- ------------ --------- ----------
##             **AAGE**               0.03063    0.0005312     57.67       0
##
##        **ACLSWKRPrivate**           0.1151     0.01525      7.545   4.522e-14
##
##     **ACLSWKRSelf employed**        0.4356     0.02018      21.59   2.383e-103
##
##      **ACLSWKRWithout pay**        -0.2398     0.03228     -7.431   1.079e-13
##
##          **AHGADegree**              1.12      0.01929      58.02       0
##
##   **AHGAHigh school graduate**     -0.6579     0.01959     -33.58   3.023e-247
##
##        **AHGANo degree**            -0.502     0.01952     -25.72   7.205e-146
##
##        **AMARITLMarried**          -0.3924      0.0486     -8.074   6.797e-16
##
##     **AMARITLNever married**       -0.4261     0.02242       -19    1.598e-80
##
##        **AMARITLWidowed**          -0.1545      0.0318     -4.858   1.186e-06
##
##     **ARACEAsian or Pacific         0.5618     0.06905      8.136   4.097e-16
##            Islander**
##
##          **ARACEBlack**             0.1185     0.06414      1.848    0.06467
##
##          **ARACEOther**             0.0389     0.07779       0.5      0.617
##
##          **ARACEWhite**             0.3565     0.06139      5.807   6.362e-09
##
##           **ASEXMale**              1.272      0.01097      115.9       0
##
##    **AUNMEMNot in universe**       -0.2974     0.01877     -15.84   1.543e-56
##
##          **AUNMEMYes**              0.4402     0.03214      13.69   1.089e-42
##
##     **AUNTYPEJob loser - on        -0.8957      0.1167     -7.678   1.614e-14
##            layoff**
```

```
##
##      **AUNTYPENew entrant**        2.274       0.2365      9.617   6.794e-22
##
##    **AUNTYPENot in universe**     -0.4561     0.09335     -4.886   1.028e-06
##
##    **AUNTYPEOther job loser**     -0.3346     0.09872      -3.39   0.0006994
##
##      **AUNTYPERe-entrant**        -0.5833      0.1171      -4.98   6.346e-07
##
##  **AWKSTATFull-time schedules**    0.774      0.09363      8.267    1.37e-16
##
##  **AWKSTATNot in labor force**     0.6511     0.09334      6.976   3.036e-12
##
##       **AWKSTATPart time**         0.6658     0.09637      6.909   4.895e-12
##
##      **AWKSTATUnemployed**         0.3311      0.1087      3.045    0.002328
##
##    **FILESTATJoint both 65+**     -0.5194     0.05806     -8.947   3.667e-19
##
##    **FILESTATJoint both under      0.1866     0.05123      3.642   0.0002706
##                65**
##
##  **FILESTATJoint one under 65 &   -0.4263     0.06008     -7.095   1.291e-12
##             one 65+**
##
##       **FILESTATNonfiler**         -1.951     0.05722     -34.09   9.913e-255
##
##        **FILESTATSingle**         -0.1418     0.02621     -5.408   6.376e-08
##
##        **GRINREGMidwest**         -0.7822      0.1317     -5.938    2.88e-09
##
##       **GRINREGNortheast**        -0.2054      0.1317      -1.56     0.1187
##
##    **GRINREGNot in universe**     -0.1228      0.1244     -0.9866    0.3238
##
##         **GRINREGSouth**          -0.4329      0.1292      -3.35   0.0008092
##
##         **GRINREGWest**           -0.2811      0.1299     -2.163    0.03051
##
##      **HHDRELHouseholder**         1.316      0.03249      40.51       0
##
##         **HHDRELOther**            0.7957     0.03719      21.39   1.501e-101
##
##  **MIGSAMENot in universe under   -0.6286     0.09304     -6.756   1.416e-11
##           1 year old**
##
##        **PRCITSHPNative**          0.2349     0.01848      12.71   5.073e-37
##
##            **SEOTR1**              0.6814     0.02969      22.95   1.465e-116
##
##            **SEOTR2**             -0.09446    0.01601      -5.9    3.634e-09
##
##    **VETQVANot in universe**       -11.89      28.73      -0.4138    0.679
##
```

```
##          **VETQVAYes**            -0.3962      0.1297     -3.056    0.002245
##
##            **VETYN2**              11.37       28.73       0.396     0.6921
##
##       **AHRSPAY_qual> 0**         -0.8183     0.02586     -31.64   1.11e-219
##
##       **CAPGAIN_qual> 0**          1.346       0.0188      71.58          0
##
##       **CAPLOSS_qual> 0**          1.332      0.02524      52.77          0
##
##        **DIVVAL_qual> 0**          1.136      0.01248      90.99          0
##
##    **WKSWORK_qualless than a      -0.8666     0.01434     -60.43          0
##              year**
##
##     **WKSWORK_qualno week**        -1.823     0.03098     -58.82          0
##
##         **(Intercept)**            -2.187      0.1782     -12.27   1.279e-34
## ------------------------------------------------------------------------------
##
## Table: Fitting generalized (binomial/logit) linear model: OUTCOME ~ .
```

On the second model, the performances of the model are: 82 % of well classified for the class "-50000" and 88 for the class "50000+".

These two models were applied to the test set in order to validate them.

-Without weight, 99% are obtained for the class "-50000" and 30 are obtained for the class "50000+"

-With weight, 88% are obtained for the class "-50000" and 88% are obtained for the class "-50000"

The predictions on the test set are close to the predictions obtained on the training set meaning that there is no overfitting.

Decision tree

To complete our analysis, we ran a decision tree. In resubstitution, the following results were obtained:

```
##
## ------------------------------------
##             - 50000.   50000+.
## -------------- ---------- ---------
##  **- 50000.**    185261     8765
##
##  **50000+.**      1880      3617
## ------------------------------------
```

and the prediction on the test set:

```
##
## ------------------------------------
##             - 50000.   50000+.
## -------------- ---------- ---------
```

```
##  **- 50000.**    92563      4370
##
##  **50000+.**      1013      1816
## ------------------------------------
```