

Data Science: Capstone Project Report

Executive Summary

The goal of this project is to predict the salary given some U.S. Census data. I selected this project in order to attempt to match the prediction accuracy cited in a 1996 paper, Improving the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid.¹ A summary of the results is found in the file adult.names.

This is a classification problem. We don't predict the exact salary. Rather, we predict which class ($> \$50,000$ or $\leq \$50,000$) that each individual belongs.

The key steps are importing the data set, cleaning the data set, creating and tuning a model, and evaluating the model against a validation data set. In the original data set work the authors listed the top 3 performing models: NBTree, C4.5, and Naive-Bayes. I choose a xgbTree, Naive-Bayes, and C5. I split the data set into a train_set and a test_set in order to avoid making any modelling decisions with the validation set. Then model parameters are tuned and evaluated against the test_set. Finally, the models make predictions for the validation set.

Methods

The data required minimal cleaning. I removed the observations with unknown values because the original researchers did.

First I build the data sets. The data files consist of two files: adult.data and adult.validation. I split the adult.data set into a train and test set. I train three models on the training data. When the models are built I use these models to predict the whether the salary of individuals in the validation set are greater than or less than or equal to \$50,000.

Results

The original project used accuracy as evaluation function and so did I. One model, C5, out-performed the original models.

Original Model	Accuracy	My Model	Accuracy
NBTree	0.8590	xgbTree	0.8509
C4.5	0.8446	C5	0.8465
Naive-Bayes	0.8388	Naive-Bayes	0.8208

Conclusion

I've learned quite a bit during this course. I can build models that predict almost as well as others with much more experience than me. Of course, I've only learned how to use the tools that others have built. That counts as success for my first data science course.

A detailed exploratory data analysis follows.

¹Kohavi, Ron, Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. <http://www.aaai.org/Papers/KDD/1996/KDD96-033.pdf>

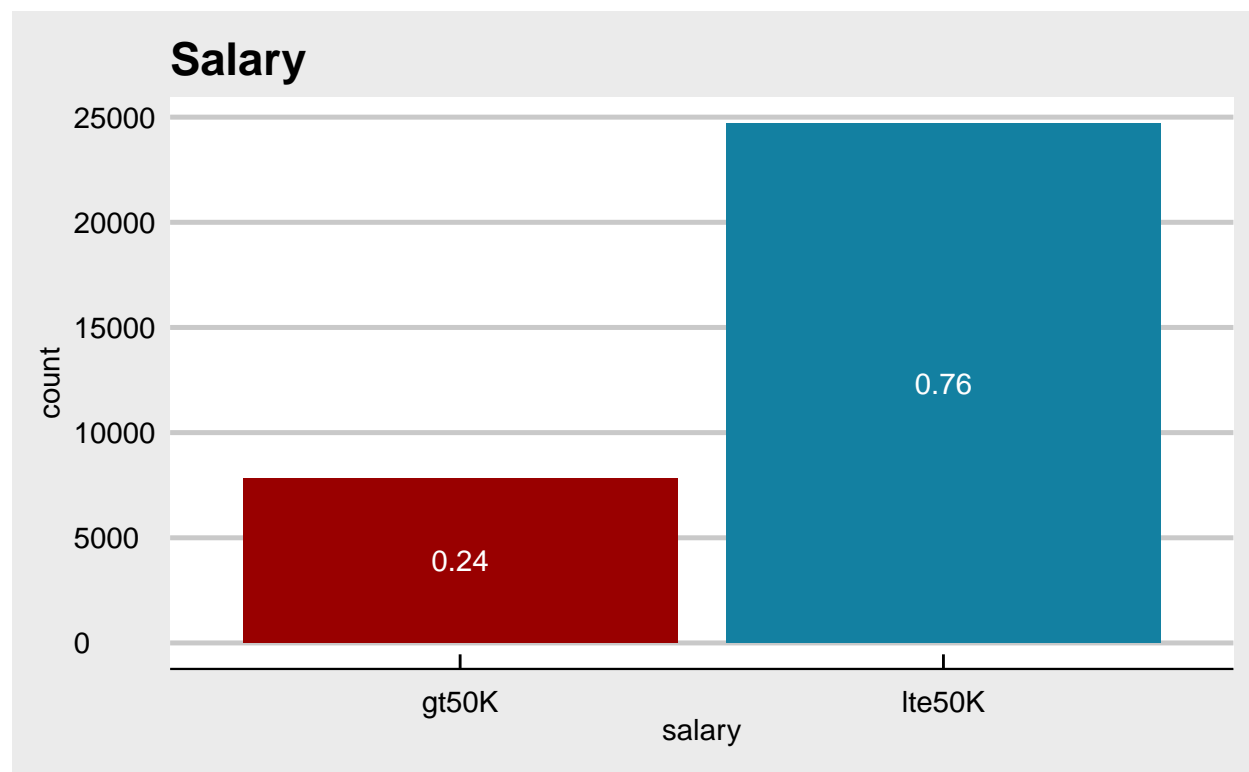
Basic Overview of Data

I obtained the data from <https://archive.ics.uci.edu/ml/datasets/census+income>. This data was originally extracted from the 1994 census bureau database. The data was originally donated to the UCI by Ronny Kohavi and Barry Becker, Data Mining and Visualization Silicon Graphics, e-mail: ronnyk@sgi.com for questions. Extraction was done by Barry Becker from the 1994 Census database.

Structure

```
## 'data.frame':    32561 obs. of  15 variables:
## $ age           : int   39 50 38 53 28 37 49 52 31 42 ...
## $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",...: 8 7 5 5 5 5 7 5 5 ...
## $ fnlwt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
## $ education     : Factor w/ 16 levels " 10th"," 11th",...: 10 10 12 2 10 13 7 12 13 10 ...
## $ education.num : int   13 13 9 7 13 14 5 9 14 13 ...
## $ marital.status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",...: 5 3 1 3 3 3 4 3 5 3 ...
## $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",...: 2 5 7 7 11 5 9 5 11 5 ...
## $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",...: 2 1 2 1 6 6 2 1 2 1 ...
## $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",...: 5 5 5 3 3 5 3 5 5 5 ...
## $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
## $ capital.gain  : int   2174 0 0 0 0 0 0 0 14084 5178 ...
## $ capital.loss  : int    0 0 0 0 0 0 0 0 0 0 ...
## $ hours.per.week: int   40 13 40 40 40 40 16 45 50 40 ...
## $ native.country: Factor w/ 42 levels " ?"," Cambodia",...: 40 40 40 40 6 40 24 40 40 40 ...
## $ salary        : Factor w/ 2 levels "gt50K","lte50K": 2 2 2 2 2 2 2 1 1 1 ...
```

Distribution of the response variable. We see this is a fairly unbalanced data set.



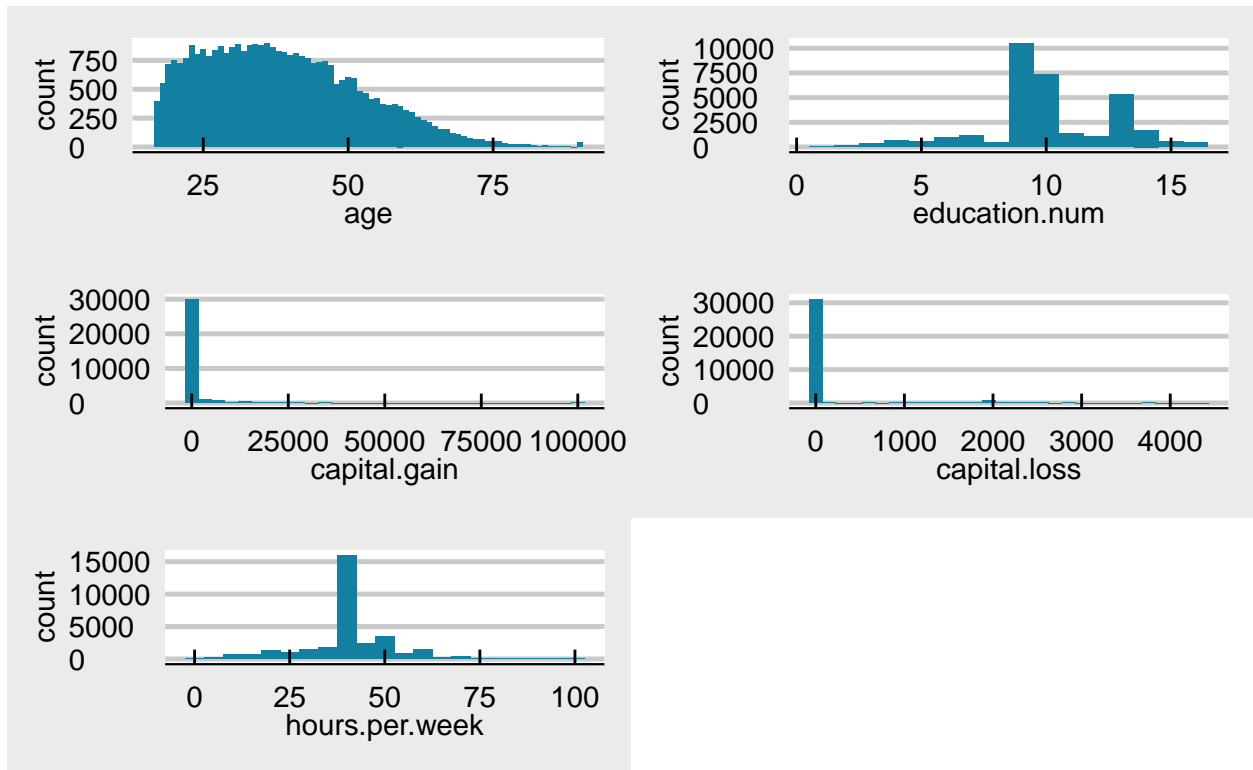
Univariate Distribution

Histograms

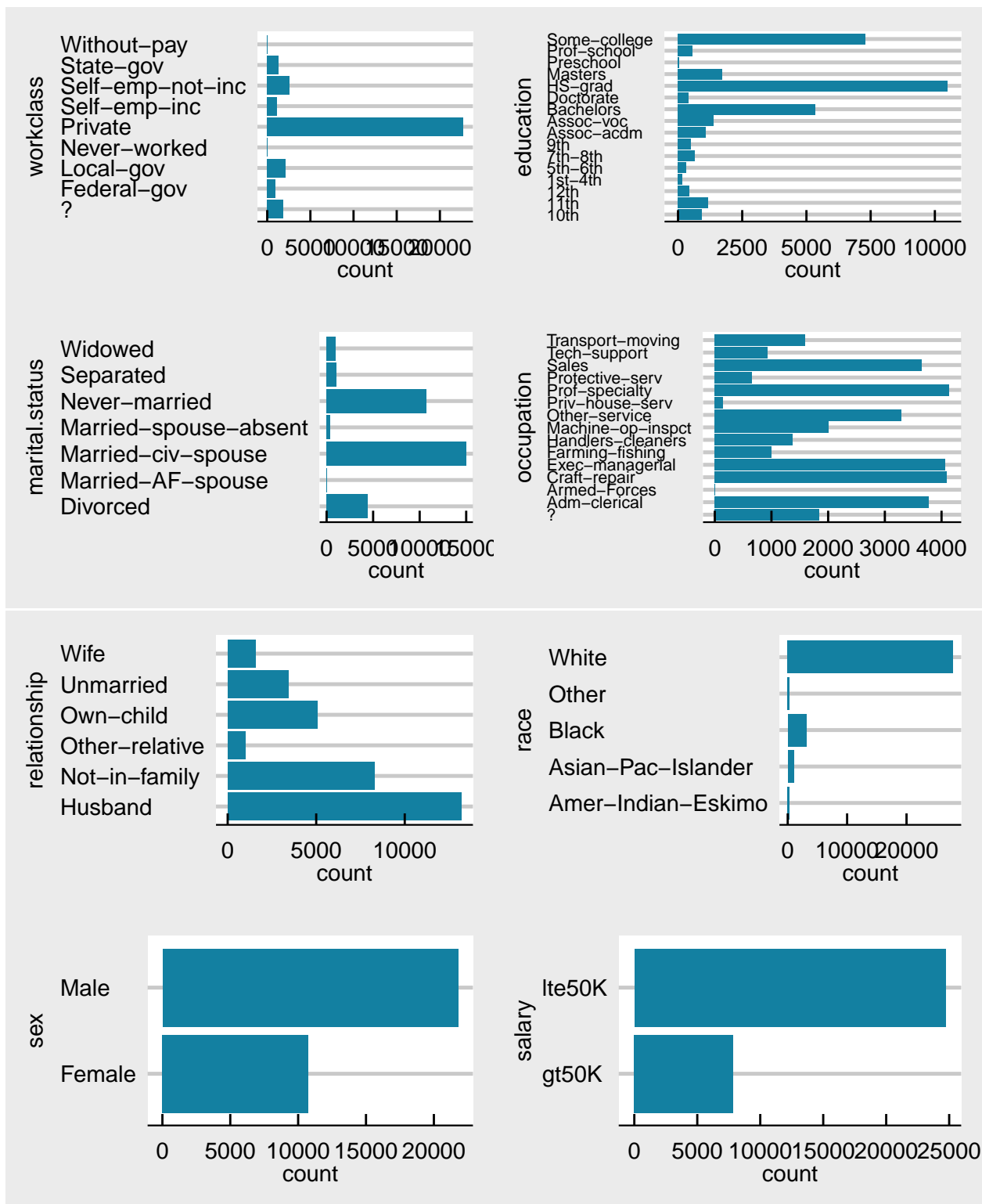
Let's take a look at the numeric variables. If we use AGE as a predictor we would likely center and normalize.

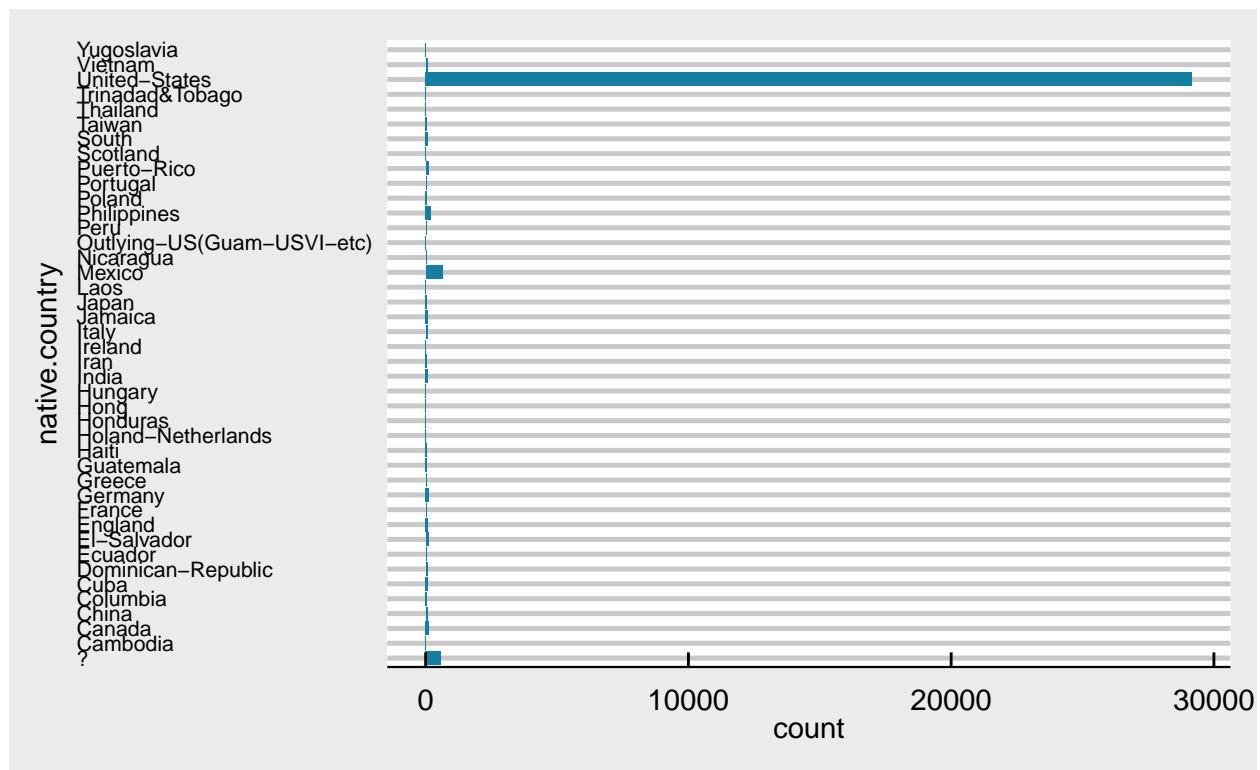
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Next we look at the categorical data.





We remove those observations with an unknown value. We could have imputed the values, perhaps considering the distribution of known values. We did not because the original authors removed the data.

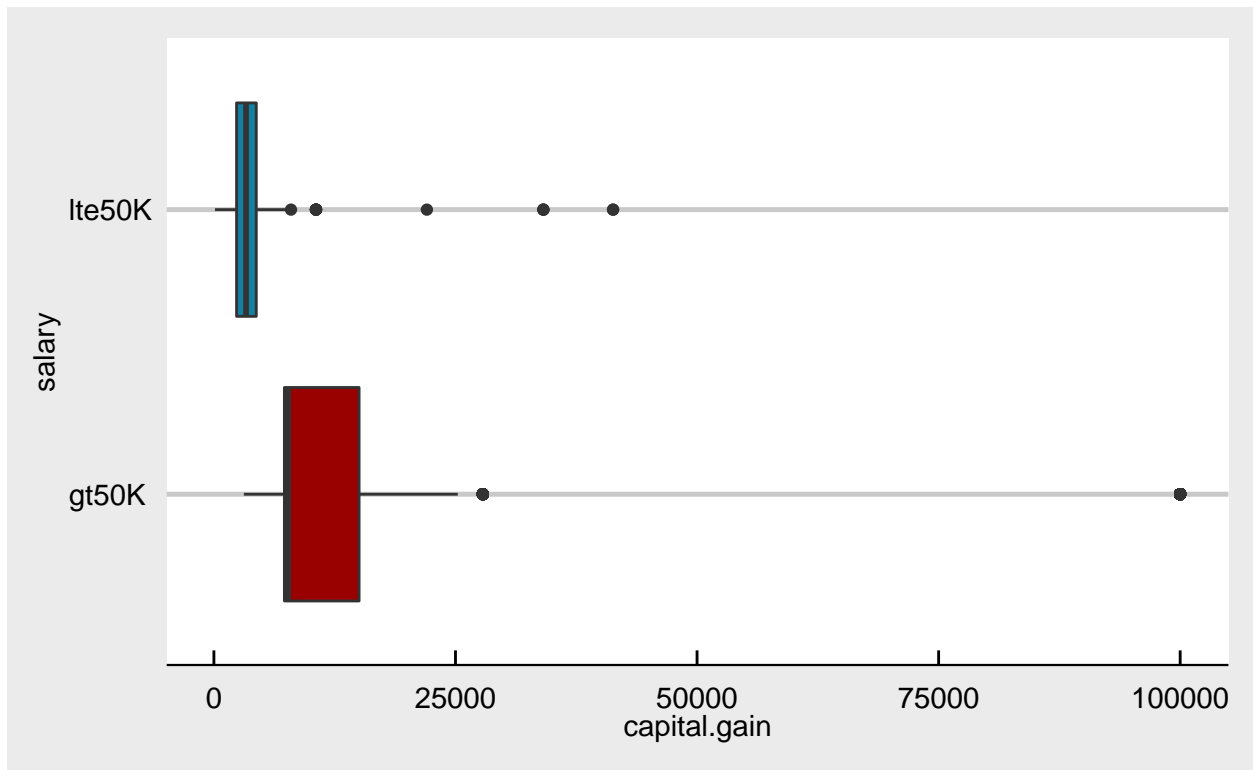
Bivariate

Response and predictor relationships

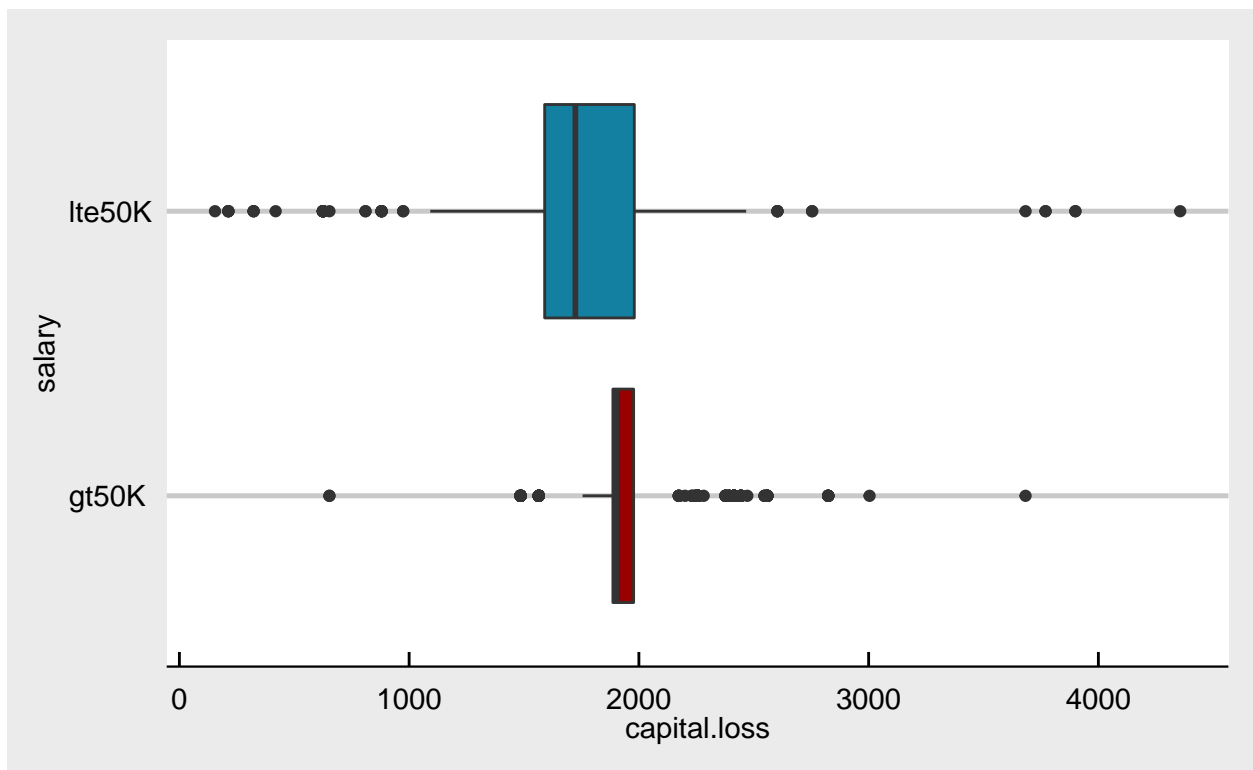
Continuous Predictors

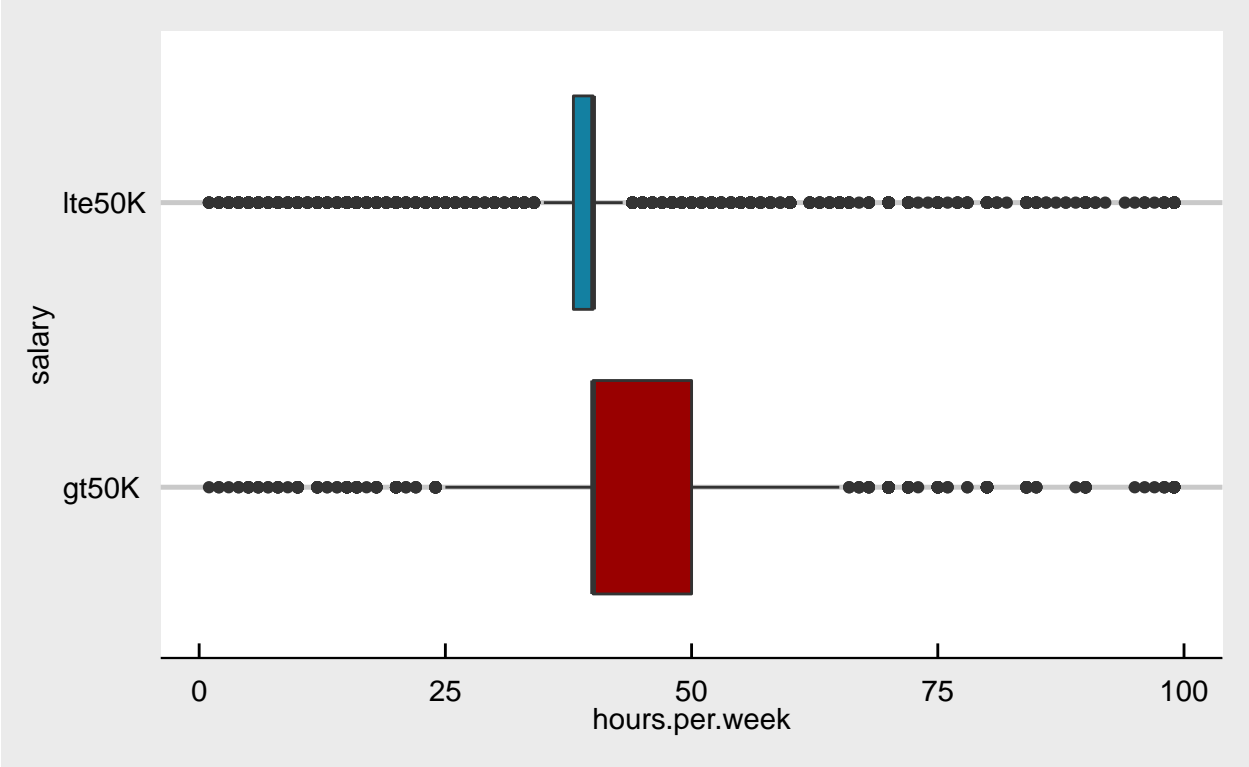
We see that age can predict low salary for young adults.





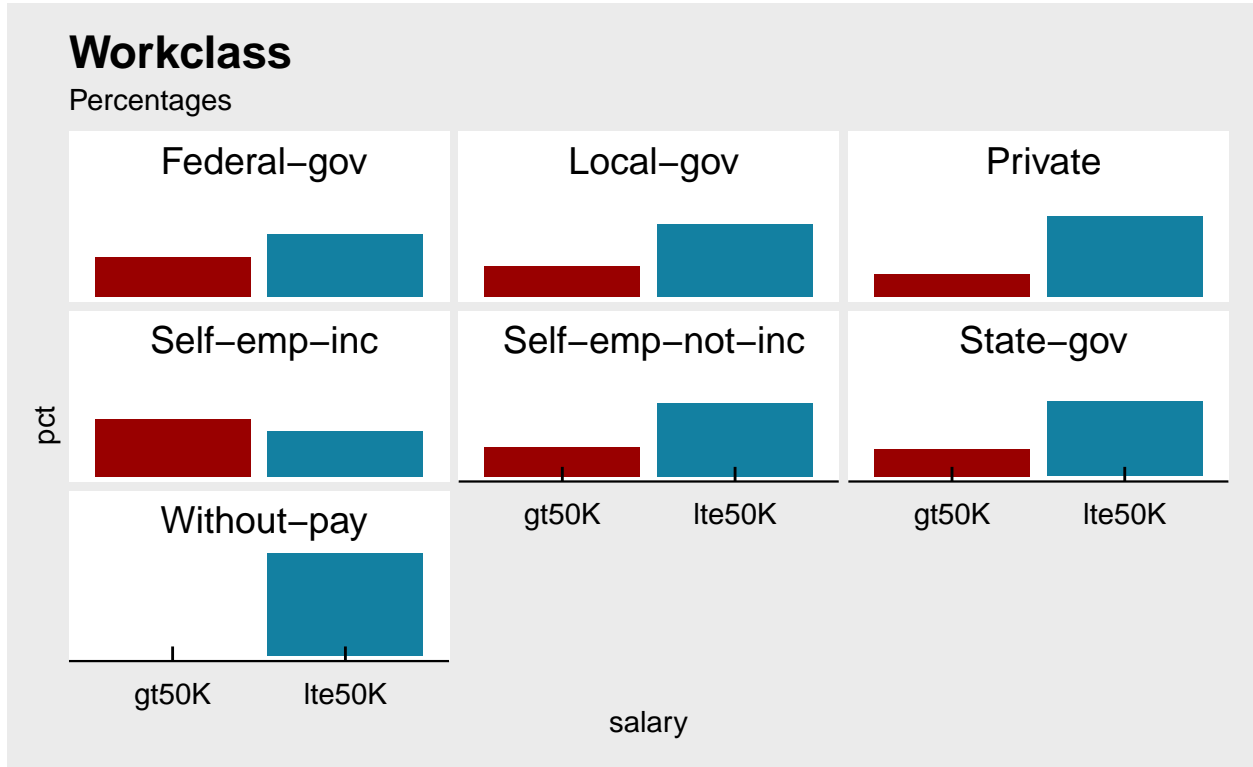
It appears that Capital Gain can also help predict salary. Capital Loss is not as helpful.





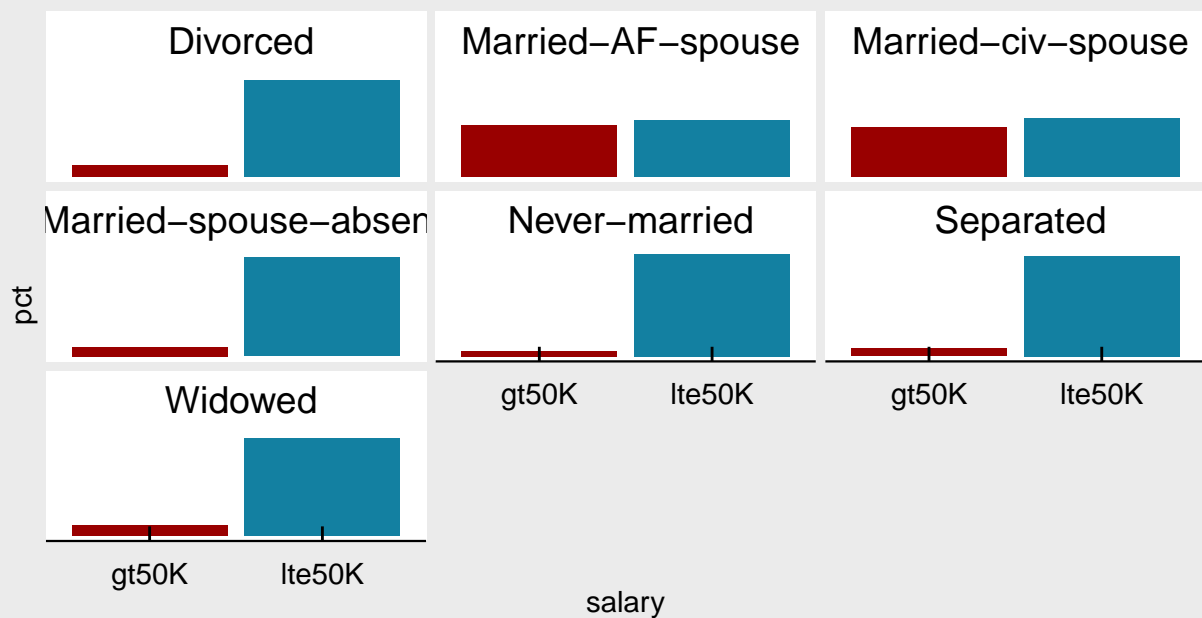
Categorical Predictors

Looking at the categorical predictors we see several that appear useful. We will want to verify our intuition by considering independence using a Chi-Square test.



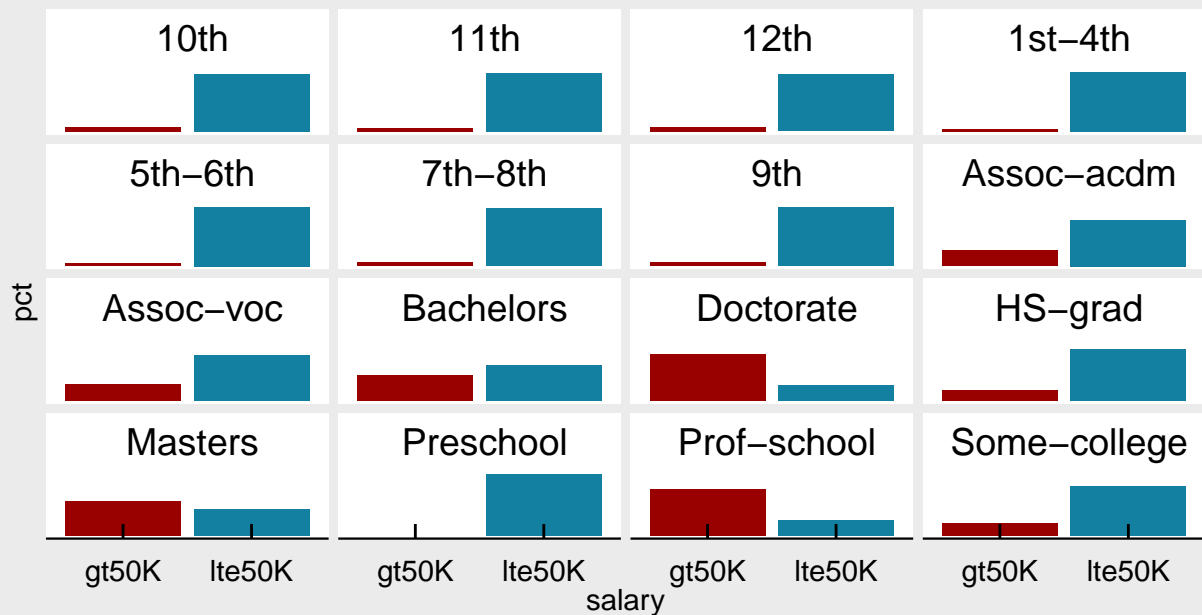
Marital Status

Percentages



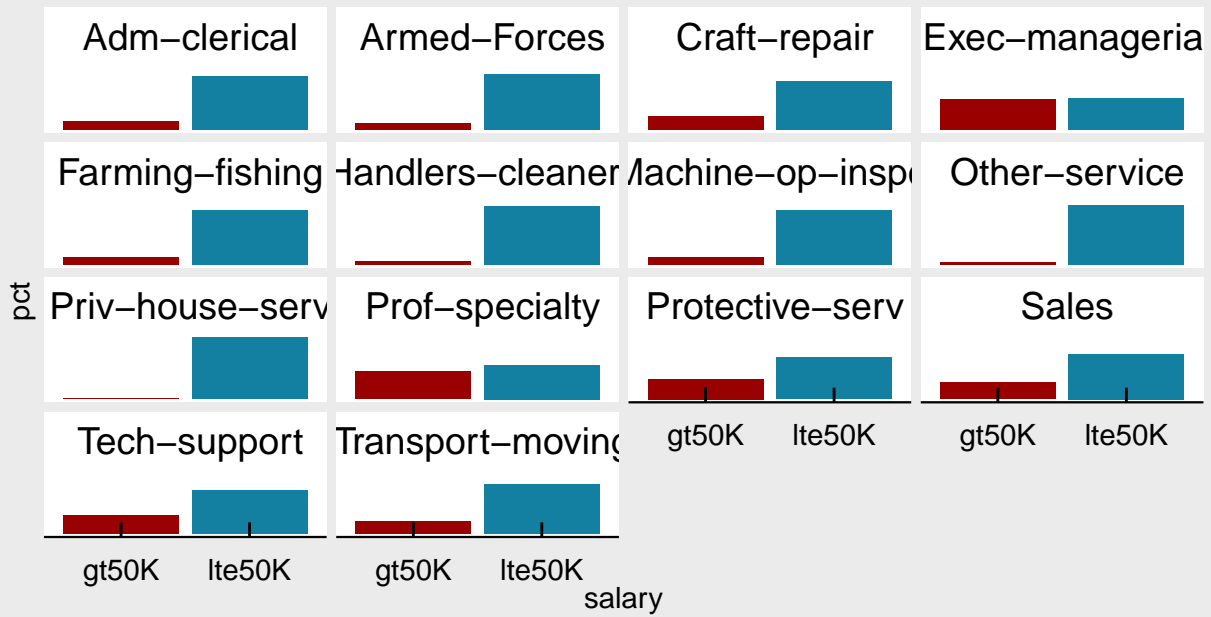
Education

Percentages



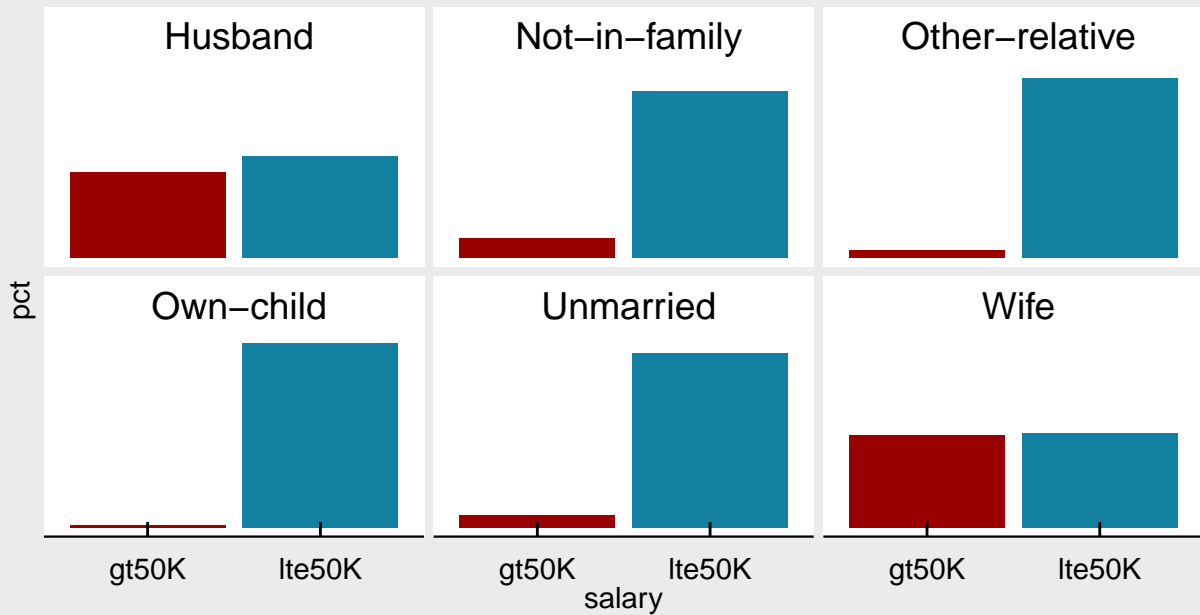
Occupation

Percentages



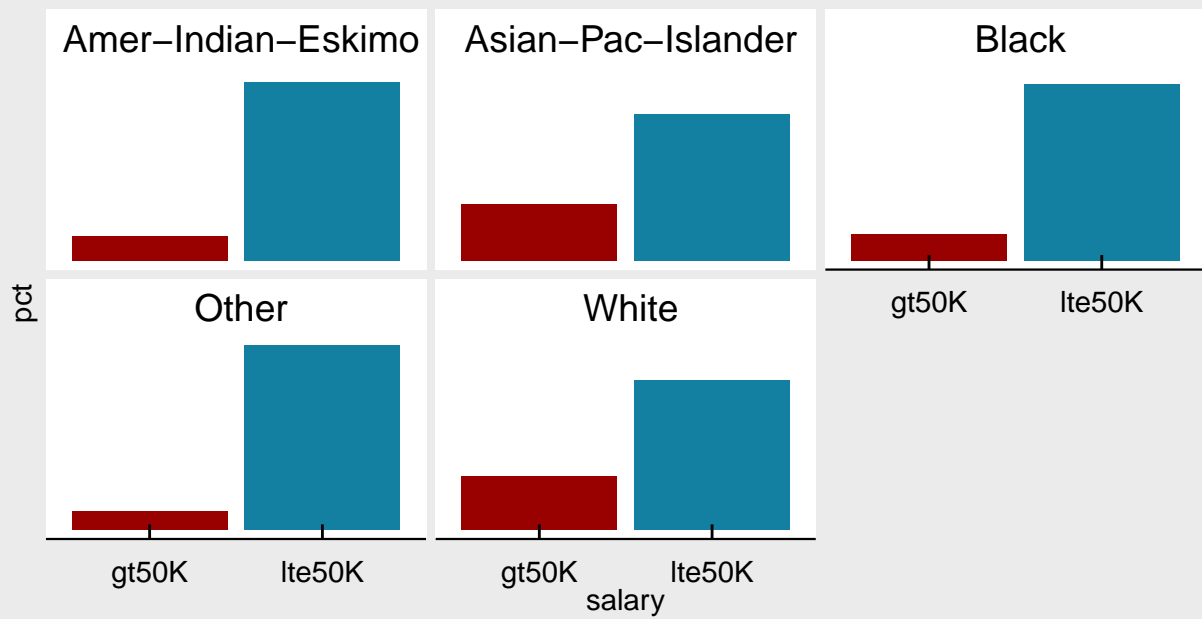
Relationship

Percentages



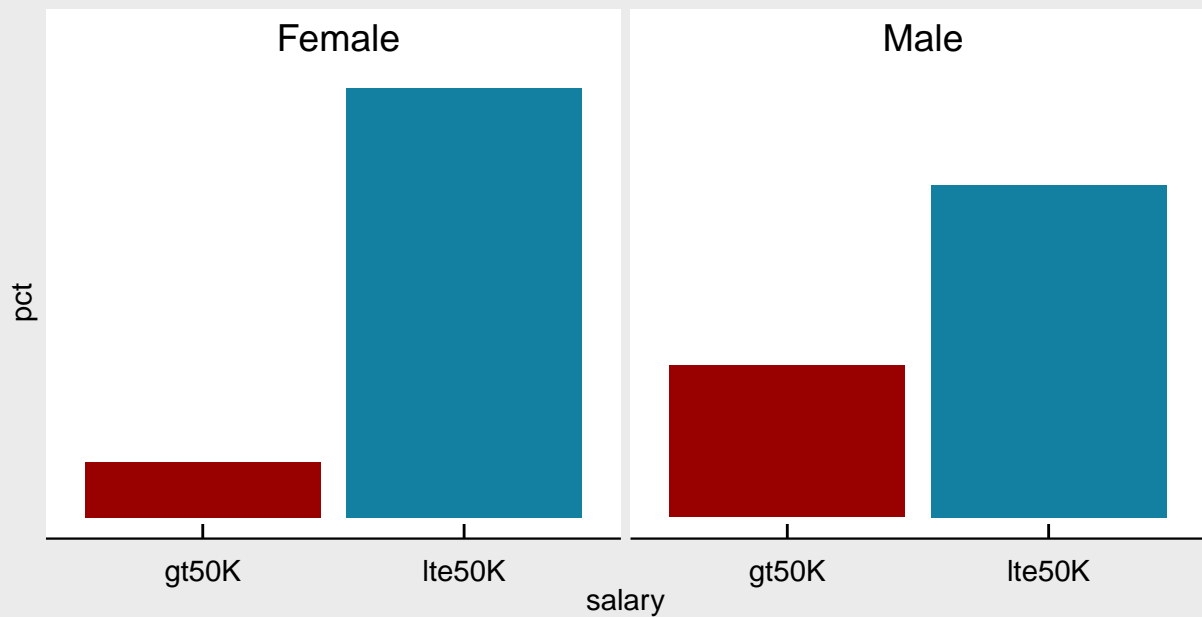
Race

Percentages



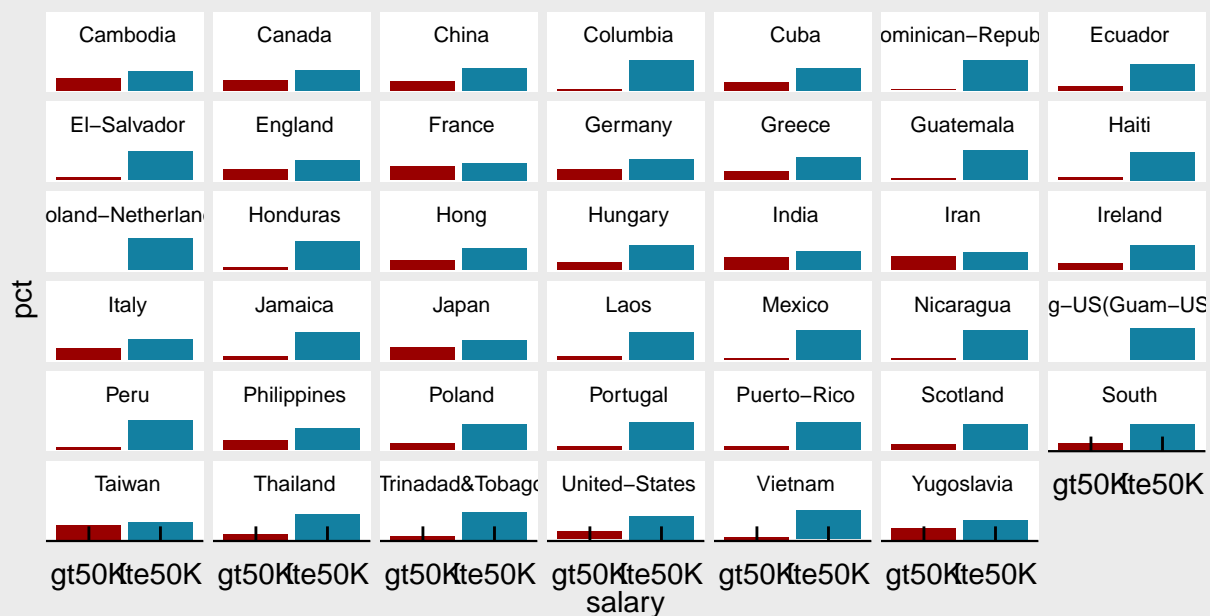
Sex

Percentages



Native Country

Percentages



We apply a Chi-Square test for independence for each variable against `salary`. We calculate the Cramer's V value for each variable against `salary`. We see that `workclass`, `occupation`, and `native.country` do not get a valid Cramer's V value since they fail to produce a correct Chi-Square test.

Those variables that are important to predicting salary are `education`, `marital.status`, `relationship`, and `sex`. If we run into model performance constraints we can reduce the categorical predictors to these.

		workclass								
education	?		education							
marital.status	?	0.09		marital.status						
occupation	?	?	?		occupation					
relationship	?	0.12	0.49	?		relationship				
race	?	0.08	0.08	?	0.1		race			
sex	?	0.09	0.47	?	0.65	0.12		sex		
native.country	?	?	?	?	?	?	?		native.country	
salary	?	0.37	0.45	?	0.45	0.1	0.22	?		