MSDS 6372 Project 2: Predicting Incomes Under/ Over $50,000

By: Allen Miller, Alex Gilbert, Justin Ehly

--- Introduction ---

The purpose of this study was to create and compare various logistic regression models that are useful for predicting whether a person will make less or more than $50,000 per year, based on predictor variables included in the Census Income Data Set. Our analysis included an data cleaning, explotary data analysis (EDA), modeling with simple logistic regression methods: Stepwise and LASSO, then an additional analysis with more complex models that utilize LDA/ QDA and a revised LASSO model with interactions between explanatory variables to provide more complexity.

At each step our EDA guides us to determine the appropriateness of predictor variables that may be significant in building data models. In our EDA we explore how different or similar variables are related to the response variable individually and in some cases combined.

Finally, we used LASSO and Stepwise to fit simple models and LDA/QDA and added complexity to the LASSO model to explore more complex predictor models. To determine a model's fit we utilized 67/33 split train and test sets and compared the sensitivity, specificity, accuracy and area under the curve (AUC) measurements of each model to guide us to determine the "best fit" model, which we ultimate decided was our final Stepwise model.

--- Data Cleaning ---

Our data came from the Census Income Data Set found on UCIs website here: http://archive.ics.uci.edu/ml/datasets/Census+Income.

The data was initially broken into two parts named adult.data, a training set and adult.test, a test set. The objective of these data sets is to built models to predict whether a person's income exceeds $50K/yr based on census data.

The initial data sets were split 2/3 and 1/3 from the online repository, after careful consideration it appeared that these should be merged and re-split later on in order to reduce the amount of data manipulation needed to clean up the data for processing.

Originally all NA's were represented by "? " that were replaced with, "Unknown," to more accurately describe what they represent. The only variables missing data as "Unknown" were *workclass* = 1,836, *occupation* = 1,843 and *native.country* = 583. It should also be noted that all missing values for workclass were also missing for occupation, with the *workclass* = "Never-worked" that also reported as an "Unknown" in the *occupation* variable.

Below is the initial data summary chart. A full description of each variable can be found here: https://github.com/justinehly/6372---50k-Income-Predictor/blob/main/Data/adult.names.txt

class	age	workclass	fnlwgt	education
<=50K :24720	Min. :17.00	Private :33906	Min. : 12285	HS-grad :15784
<=50K.:12435	1st Qu.:28.00	Self-emp-not-inc: 3862	1st Qu.: 117551	Some-college:10878
>50K : 7841	Median :37.00	Local-gov : 3136	Median : 178145	Bachelors : 8025
>50K. : 3846	Mean :38.64	Unknown : 2799	Mean : 189664	Masters : 2657
	3rd Qu.:48.00	State-gov : 1981	3rd Qu.: 237642	Assoc-voc : 2061
	Max. :90.00	Self-emp-inc : 1695	Max. :1490400	11th : 1812
		(Other) : 1463		(Other) : 7625

education.num	marital.status	occupation	relationship
Min. : 1.00	Divorced : 6633	Prof-specialty : 6172	Husband :19716
1st Qu.: 9.00	Married-AF-spouse : 37	Craft-repair : 6112	Not-in-family :12583
Median :10.00	Married-civ-spouse :22379	Exec-managerial: 6086	Other-relative: 1506
Mean :10.08	Married-spouse-absent: 628	Adm-clerical : 5611	Own-child : 7581
3rd Qu.:12.00	Never-married :16117	Sales : 5504	Unmarried : 5125
Max. :16.00	Separated : 1530	Other-service : 4923	Wife : 2331
	Widowed : 1518	(Other) :14434	

race	sex	capital.gain	capital.loss	hours.per.week	native.country
Amer-Indian-Eskimo: 470	Female: 16192	Min. : 0	Min. : 0.0	Min. : 1.00	United-States:43832
Asian-Pac-Islander: 1519	Male : 32650	1st Qu.: 0	1st Qu.: 0.0	1st Qu.:40.00	Mexico : 951
Black: 4685		Median : 0	Median : 0.0	Median :40.00	Unknown : 857
Other: 406		Mean : 1079	Mean : 87.5	Mean :40.42	Philippines : 295
White: 41762		3rd Qu.: 0	3rd Qu.: 0.0	3rd Qu.:45.00	Germany : 206
		Max. :99999	Max. :4356.0	Max. :99.00	Puerto-Rico : 184
					(Other) : 2517

The response variable “class” was changed to “income” to be more descriptive and easier to explain.

We also removed all whitespace that appeared before and after each categorical variable to help with analyzing the data.

Individual predictor variables that showed some possible need for recombination were capital.gain, capital.loss, workclass, occupation and marital.status.

Capital.gain/ capital.loss: There were a lot of zeros in the capital.gain and capital.loss columns, so we made the decision to change those to “yes” or “no” binary factor columns because we hypothesize that people with capital gains or losses are probably more likely to have higher incomes so the amount of the gain or loss was not as important as if it just existed.

Workclass: There are so few governmental jobs it made sense to merge those together, likewise it made sense to merge the values “unpaid” with “unknown”. We used a glm function call to confirm the significance of these proposed changes based on p-values from z-values, allowing us to reduce the factor levels from 9 to 5. Proportions between and among levels is below showing more reasonable weight per level.

Levels	Proportions	<=50K	>50K
Gov't	0.13362612	0.092442	0.041184
Private	0.69703019	0.544609	0.152422
Self-emp-inc	0.03427413	0.015172	0.019103
Self-emp-not-inc	0.07803814	0.055803	0.022235
Unknown/Unpaid	0.05703142	0.051166	0.005866

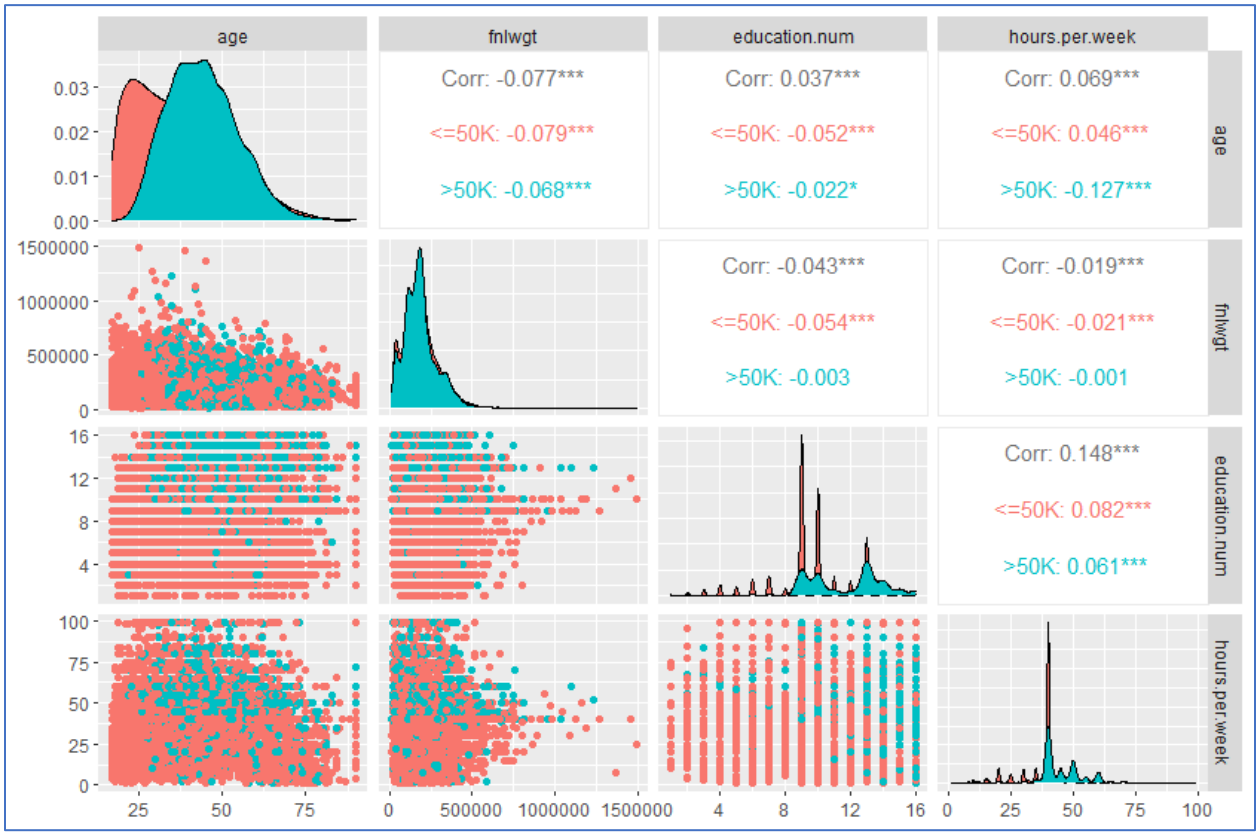
Occupation: Reviewing the breakdown of how many observations are within each factor level of occupation, Armed Services represents just 15 of the 48,842 observations or 0.03% of the total, essentially giving it very little predictive power, but after a logistic regression test we see the pvalue = 0.03 from z-value, we notice the confidence interval (-0.02031461, 2.185816489) crosses zero, so merging it with a similar occupation makes sense. Also notable is that Machine-op-inspct has pvalue=0.07 with CI(-0.25527485, 0.009908866), we merged that with Other-Service. A follow up glm showed the recombined variables are all statistically significant without any confidence intervals crossing zero.

Factor Levels	Proportions	<=50K	>50K
Adm-clerical	0.114880636	0.099156	0.015724
ArmForc/ProtSvc	0.020433234	0.014025	0.006408
Craft-repair	0.125138201	0.096822	0.028316
Exec-managerial	0.124605872	0.065067	0.059539
Farming-fishing	0.030506531	0.026964	0.003542
Handlers-cleaners	0.042422505	0.039597	0.002825
MachOpIns/OthSvc	0.162667376	0.150874	0.011793
Priv-house-serv	0.004954752	0.004893	6.14E-05
Prof-specialty	0.126366652	0.069367	0.057
Sales	0.112689898	0.08249	0.030199
Tech-support	0.029605667	0.021007	0.008599
Transport-moving	0.048216699	0.038369	0.009848
Unknown	0.057511977	0.052086	0.005426

Marital.status: we notice that there are very few Married-AF-spouse (married armed forces spouse) observations at 37, maybe it makes more sense to just combine those with married-civ-spouse (married civilian spouse). We also see from a glm that Married-spouse-absent is not statistically significant when divorced is the reference, after a couple trials with glm, we settled on combining Marrried-spouse-absent with Separated and Married-AF-spouse with Married-civ-spouse, now all yield pvalues below 0.05 based on zvalues and no confidence intervals cross zero.

--- EDA ---

Starting with the remaining continuous predictor variables age, fnlwgt, *education.num*, *hours.per.week* we ran a ggpairs matrix to look for separation by Income and any dependencies.

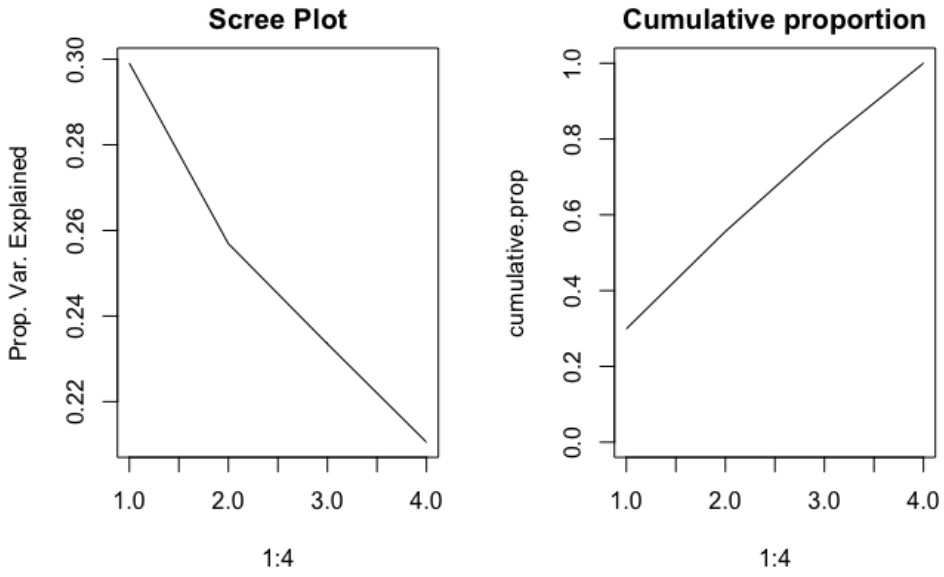


We do show some Income separation between *age vs fnlwgt*, *age* vs *education.num*, *age vs hours.perweek* as well as decent separation between *fnlwgt vs education.num* and *education.num vs hourse.per.week*. Since *education.num* is really just a numerical representation of *education*, it is probably better considered as a categorical variable and in that case is a redundancy. None of the numerical variables appear to have significant correlation with each other.

We further confirmed this using a correlation matrix from both the stats and corpcor packages.

```
> cor(clean.int)
                     age      fnlwgt education.num hours.per.week
age           1.00000000 -0.07662808    0.03094038     0.07155834
fnlwgt       -0.07662808  1.00000000   -0.03876068    -0.01351871
education.num 0.03094038 -0.03876068    1.00000000     0.14368891
hours.per.week 0.07155834 -0.01351871    0.14368891     1.00000000
> cor2pcor(cov(clean.int))
            [,1]         [,2]        [,3]         [,4]
[1,]  1.00000000 -0.075151882  0.01817077  0.067436439
[2,] -0.07515188  1.000000000 -0.03573075 -0.002929675
[3,]  0.01817077 -0.035730754  1.00000000  0.141710521
[4,]  0.06743644 -0.002929675  0.14171052  1.000000000
```

We performed PCA next to see what the R would tell us about the continuous variables and if principal components would likely reduce the number of overall variables for the model. The PCA results show 2 PCs make up about 56% of the variability and 3 make up about 80%. Considering there are only 4 continuous variables in the model, it does not seem to make sense that we should use the PCs.



Next we want to see if there is multicollinearity across the dataset using VIFs on a full overfit model.

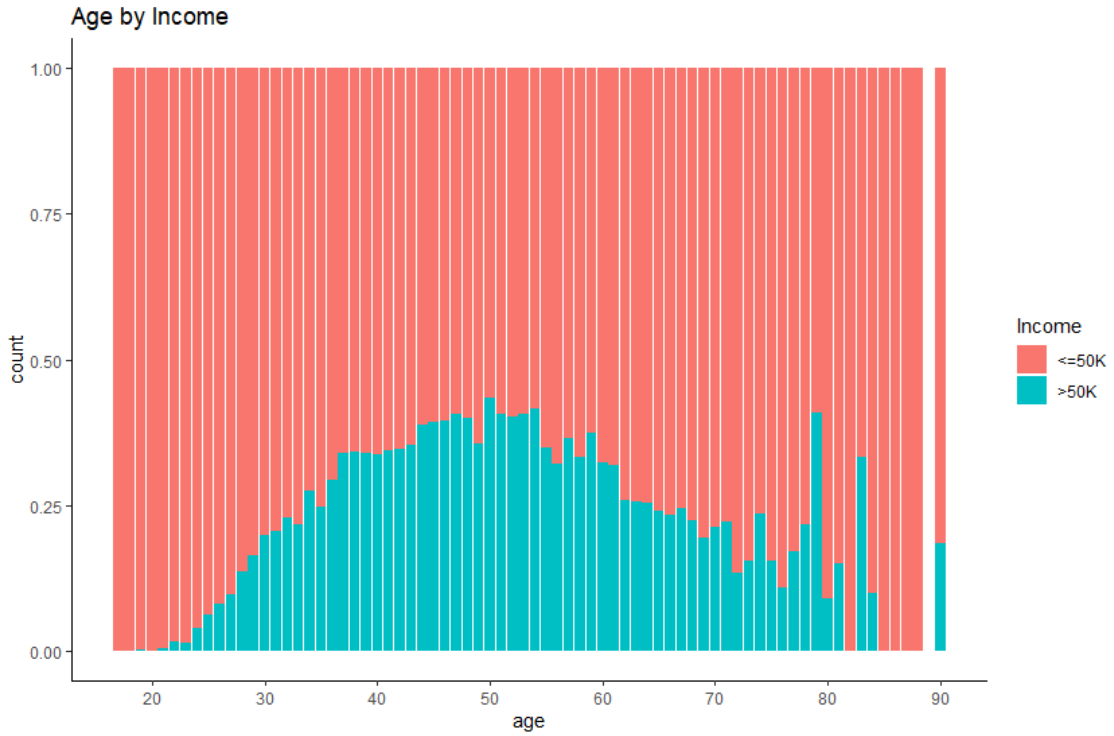
```
> fac.clean <- clean %>% select_if(~class(.) == "factor")
> vif(glm(Income~., data=fac.clean, family="binomial"))
                    GVIF Df GVIF^(1/(2*Df))
workclass      150.962532  4        1.872228
education        1.938420 15        1.022308
marital.status  46.516743  4        1.616036
occupation     257.223964 12        1.260171
relationship   107.430299  5        1.596293
race             2.663842  4        1.130287
sex              2.783506  1        1.668384
native.country   3.091937 41        1.013861
capgain          1.029180  1        1.014485
caploss          1.013704  1        1.006828
> # looks ok between the categorical variables.
```

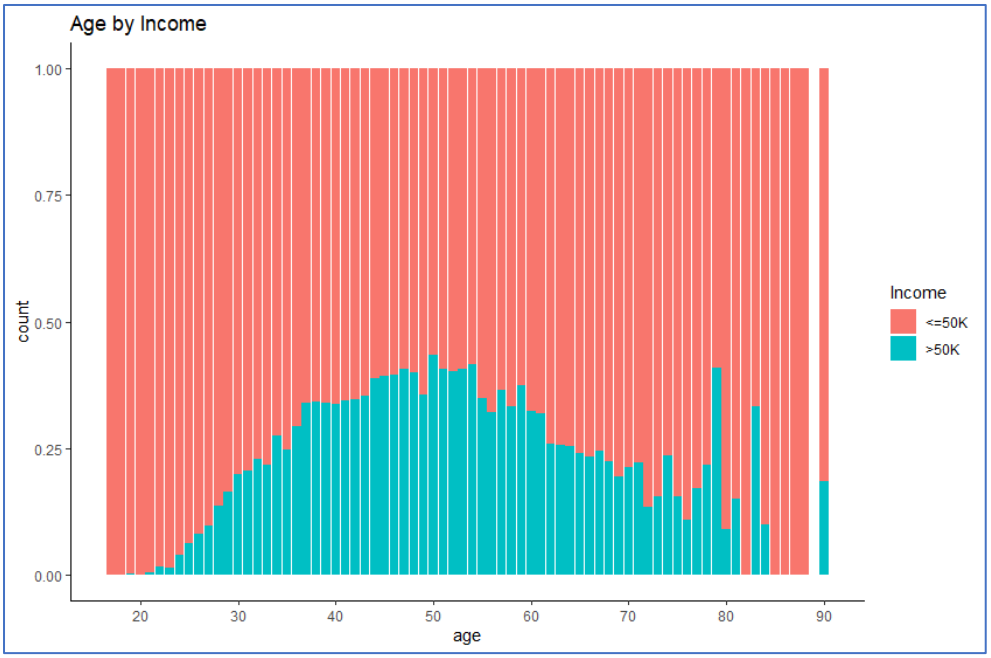
Based on the GVIF^(1/2*Df) all being relatively small, even when compared to 5 or 10, we should be ok to model with all these variables to start, but curiosity points to whether a continuous variable and a categorical variable might be telling us the same thing, such as the education and education.num variables.

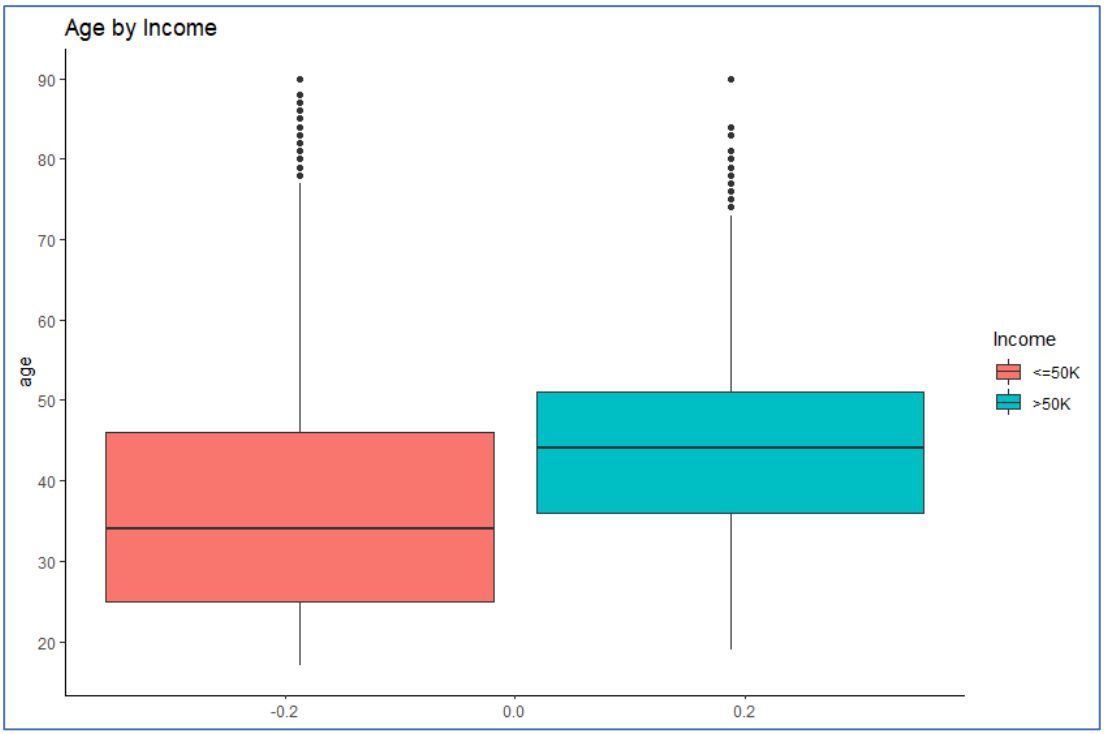
Age: We find that age ranges from 17-90 with people making >50k being on average about 7 years older (see summary statistics)

Income	"<=50K"	">50K"
age.Min.	"17.00000"	"19.00000"
age.1st Qu.	"25.00000"	"36.00000"
age.Median	"34.00000"	"44.00000"
age.Mean	"36.78374"	"44.24984"
age.3rd Qu.	"46.00000"	"51.00000"
age.Max.	"90.00000"	"90.00000"

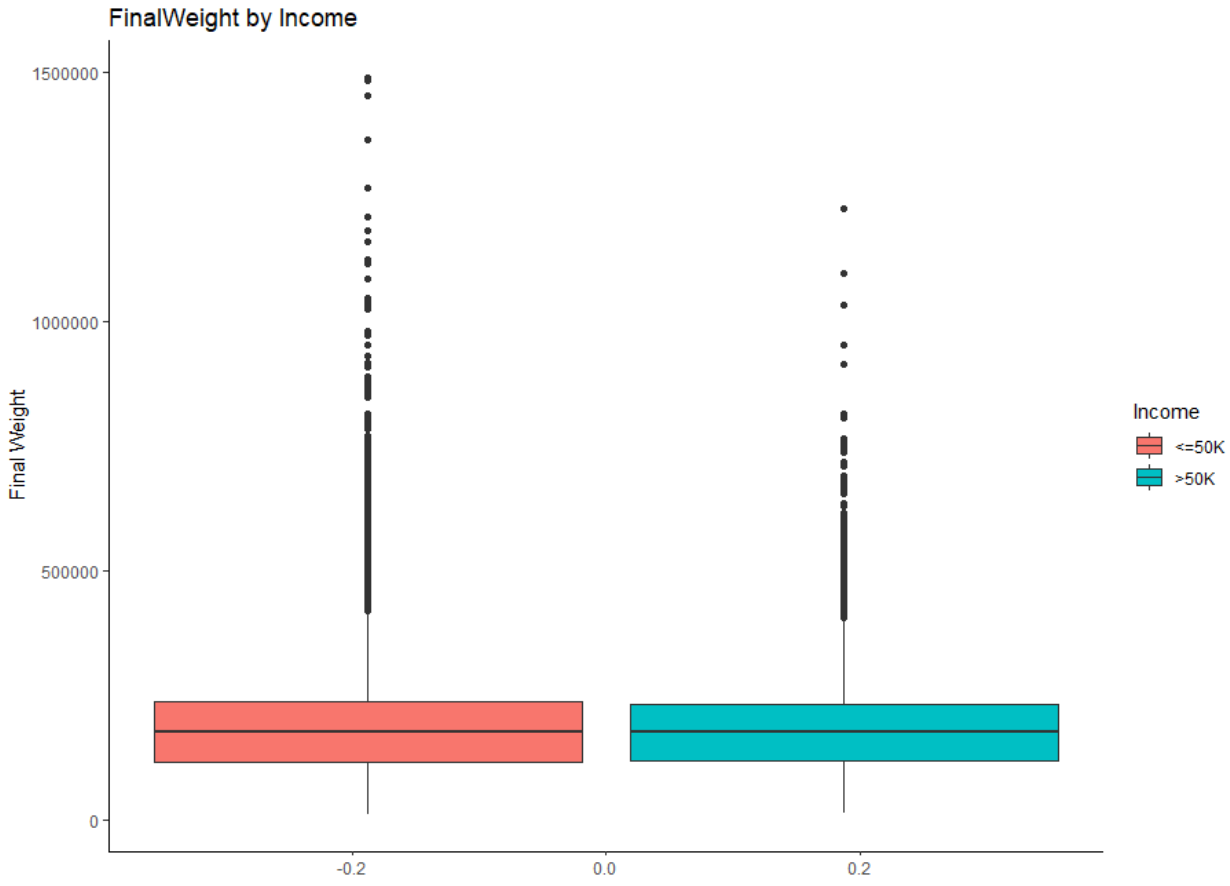
When we plot age and split the age between the under and over $50,000 response variable we can see that there is a range of ages that have a higher probability of earning more than $50,000 from about 35 to 60 years old and then some interesting data points in the 77-79 range.




Age by Income
count
1.00
0.75
0.50
0.25
0.00
20
30
40
50
60
70
80
90
age
Income
<=50K
>50K

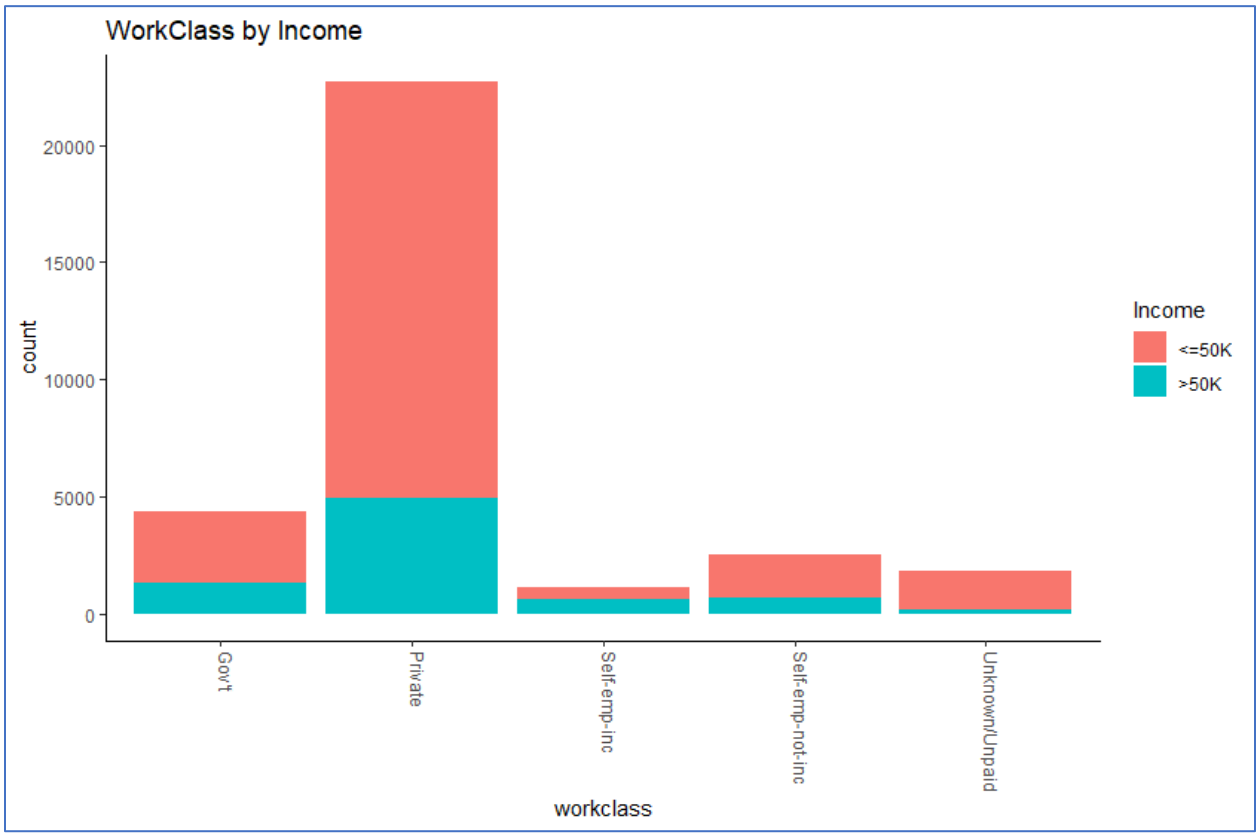

Age by Income
age
90
80
70
60
50
40
30
20
-0.2
0.0
0.2
Income
<=50K
>50K

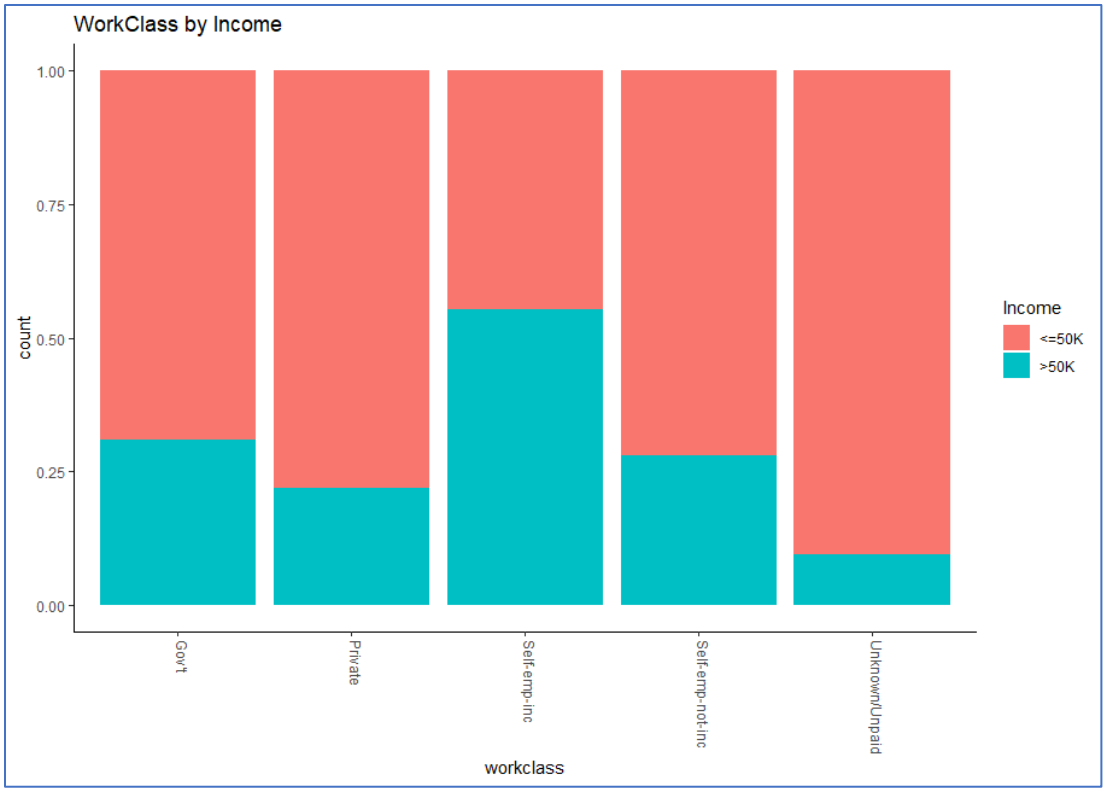
Fnlwgt is what the census from each country assumes is the total number of people meeting all the criteria in each row, it is a weighting metric so it might lend itself useful for prediction in that it can standardize or balance each row of data. We see in the boxplot that there is a pretty good balance between under and over $50,000 income with more variance in the under $50,000 group of data.



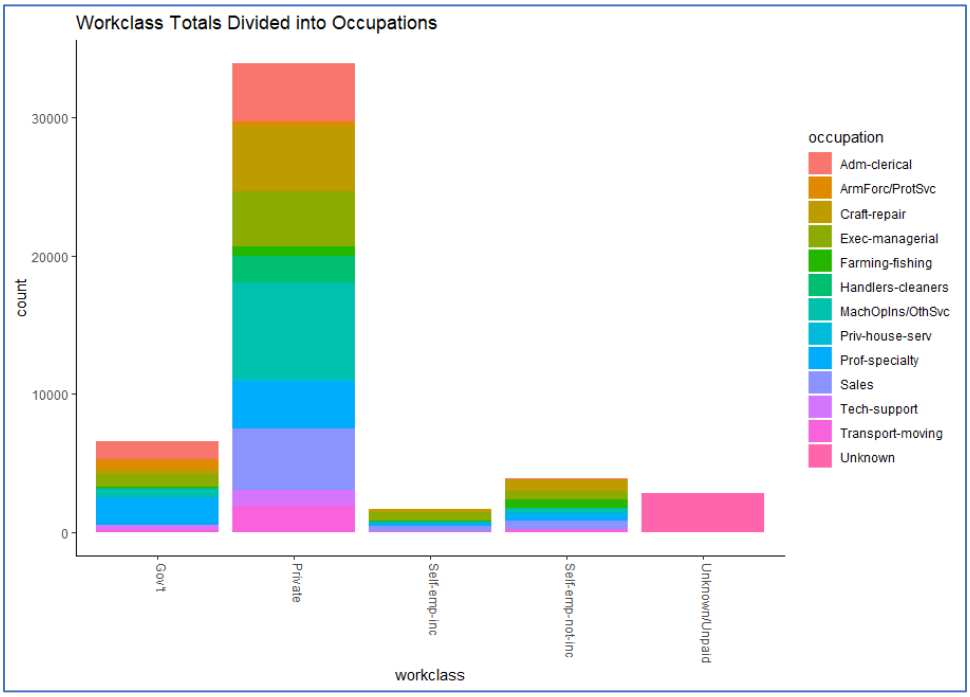
Workclass being reduced to just 5 levels from the original 9 shows that private has the most people with about 70% while unknown/ unpaid has the least. When we graph the variable and break it down by Income, we see that there may very well be some good predictive power with this variable, while most make under $50k annually, over half of the people that work in Self-Emp-Inc make more than $50k.

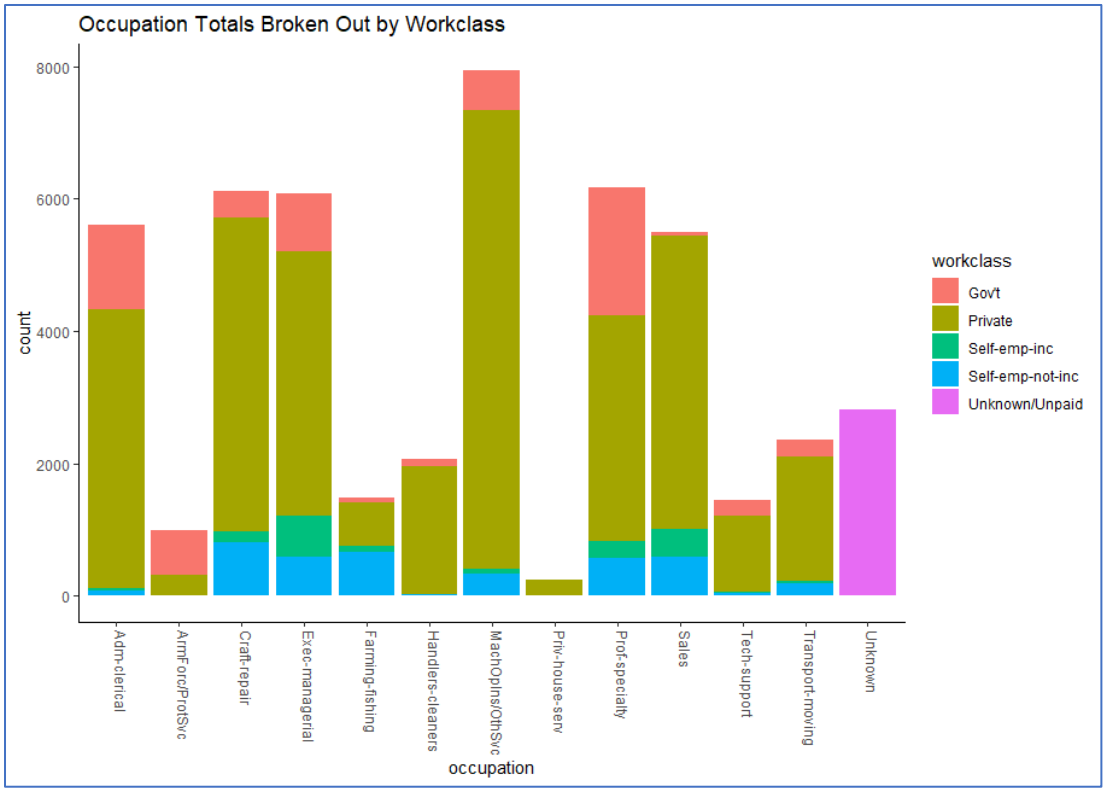
We chart this variable two different ways, one showing the summary of how many people are in each workclass and split it by under/ over $50,000 and then we show distribution of people making under/over $50,000 a year by workclass.

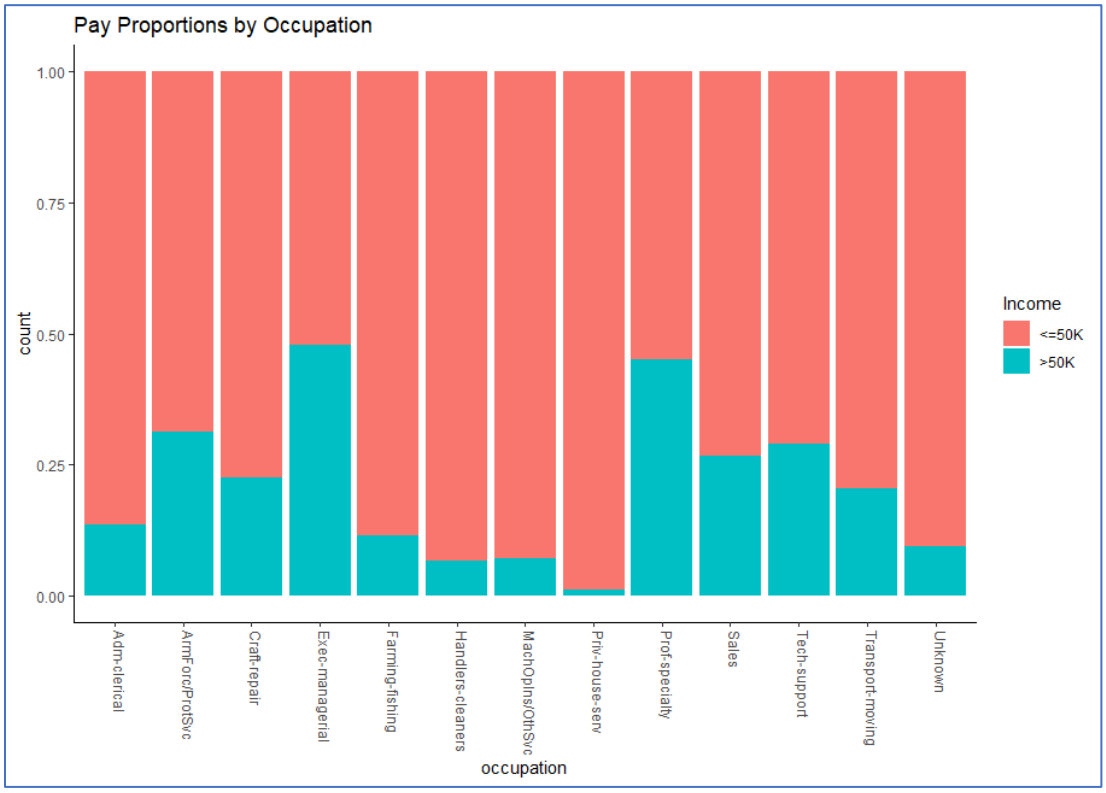




Next we look at Occupation, or a more specific label or category of what each subject does within their workclass, what may be interesting here is how different occupations are paid between working classes or industries.

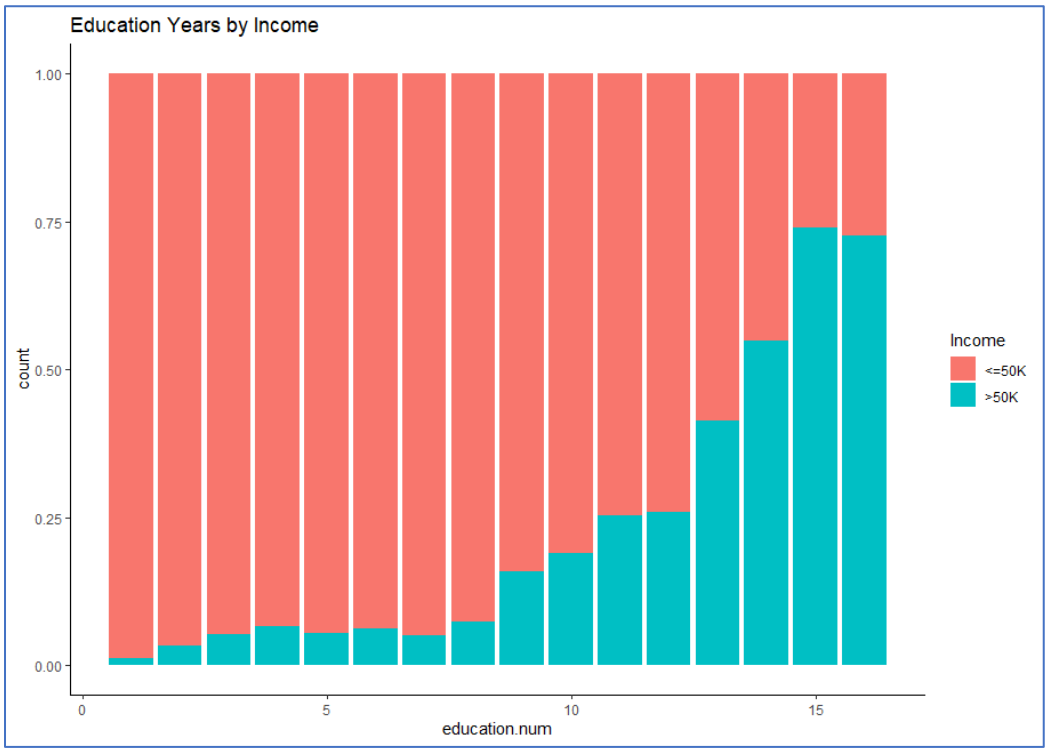


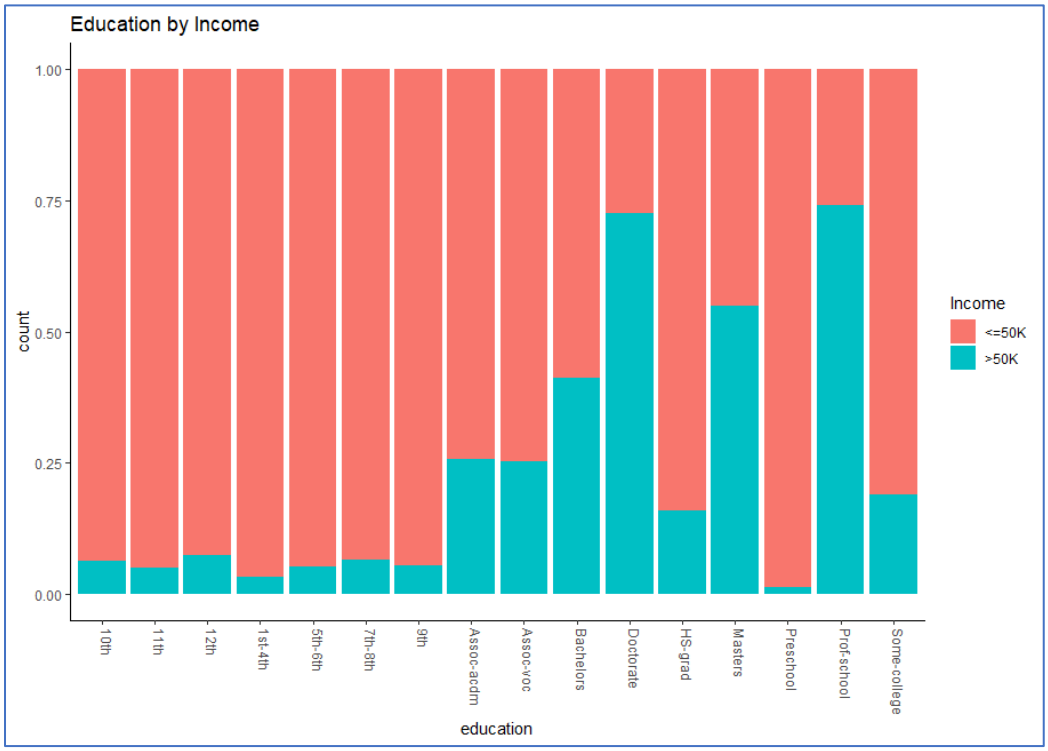


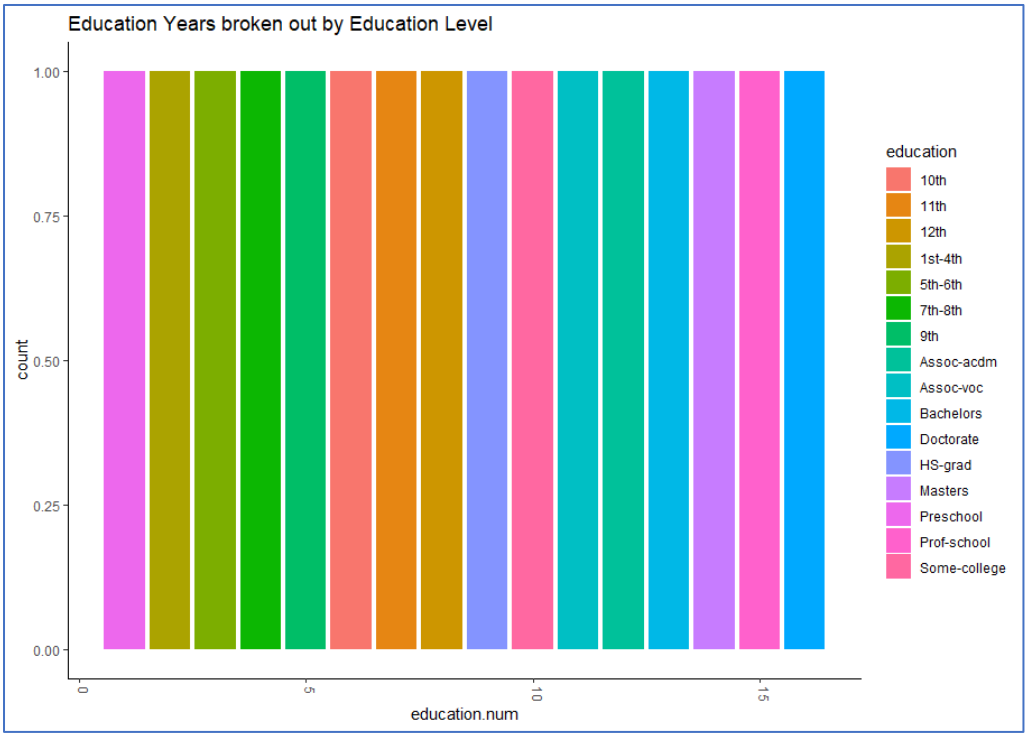


Exec- Managerial and Professional Specialty have the highest proportions of people making over $50k.

Reviewing education, the education and education.num are essentially telling us the same thing and we confirm that with both visuals and in a table…the hypothesis is that people with more education make more money.







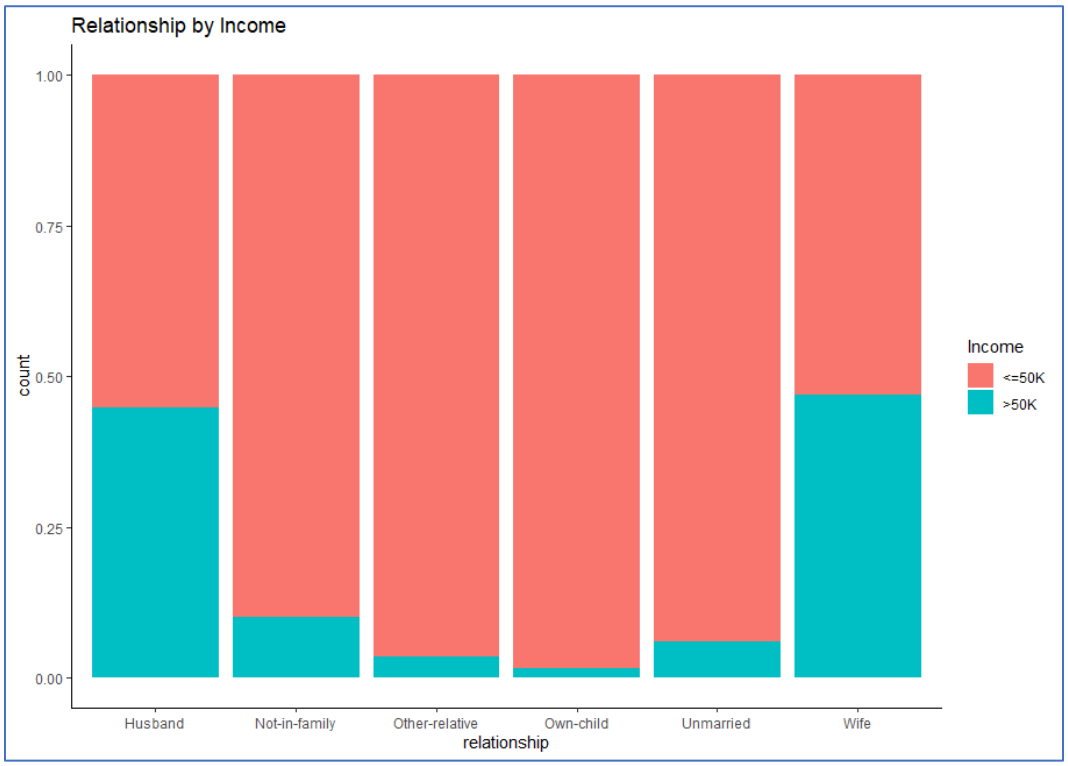
A table further confirms this.

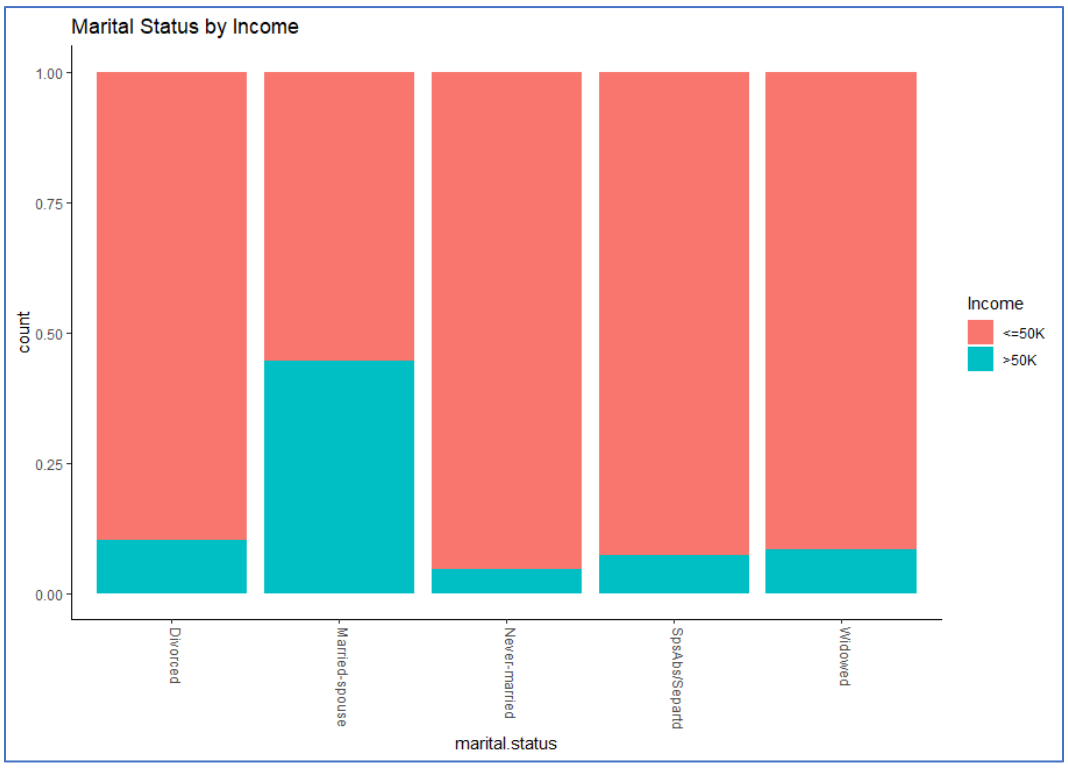
```
> table(clean$education.num, clean$education)

     10th  11th  12th 1st-4th 5th-6th 7th-8th   9th Assoc-acdm Assoc-voc Bachelors Doctorate HS-grad Masters
  1     0     0     0       0       0       0     0          0          0         0         0       0       0
  2     0     0     0     247       0       0     0          0          0         0         0       0       0
  3     0     0     0       0     509       0     0          0          0         0         0       0       0
  4     0     0     0       0       0     955     0          0          0         0         0       0       0
  5     0     0     0       0       0       0   756          0          0         0         0       0       0
  6  1389     0     0       0       0       0     0          0          0         0         0       0       0
  7     0  1812     0       0       0       0     0          0          0         0         0       0       0
  8     0     0   657       0       0       0     0          0          0         0         0       0       0
  9     0     0     0       0       0       0     0          0          0         0         0   15784       0
  10    0     0     0       0       0       0     0          0          0         0         0       0       0
  11    0     0     0       0       0       0     0          0       2061         0         0       0       0
  12    0     0     0       0       0       0     0       1601          0         0         0       0       0
  13    0     0     0       0       0       0     0          0          0      8025         0       0       0
  14    0     0     0       0       0       0     0          0          0         0         0       0    2657
  15    0     0     0       0       0       0     0          0          0         0         0       0       0
  16    0     0     0       0       0       0     0          0          0         0       594       0       0

   Preschool Prof-school Some-college
1         83           0            0
2          0           0            0
3          0           0            0
4          0           0            0
5          0           0            0
6          0           0            0
7          0           0            0
8          0           0            0
9          0           0            0
10         0           0        10878
11         0           0            0
12         0           0            0
13         0           0            0
14         0           0            0
15         0         834            0
16         0           0            0
```

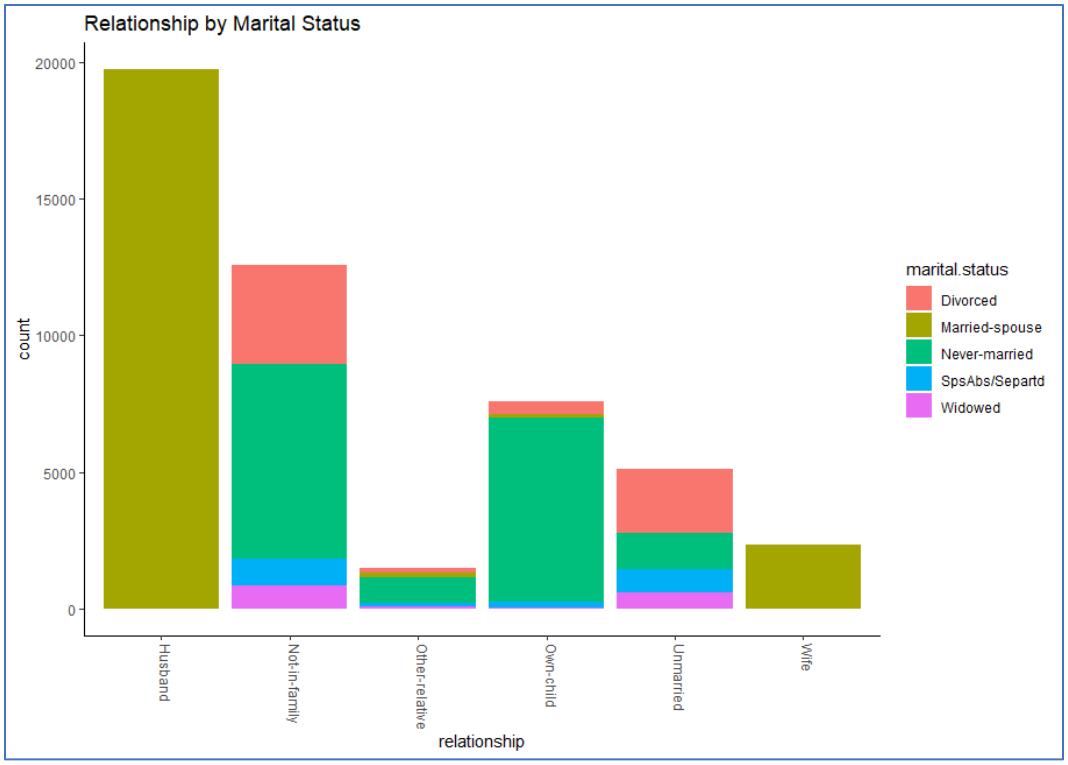
Again, just reconfirming these variables are telling us the same thing, while the continuous variable seems to make more sense comparing number of years of education to earning potential, but it may be useful to also have a categorical label on it as well. Here we can see there isn't any overlap between

education years and the category of education. Probably a good idea to let the software selection feature choose which variable makes the most sense.

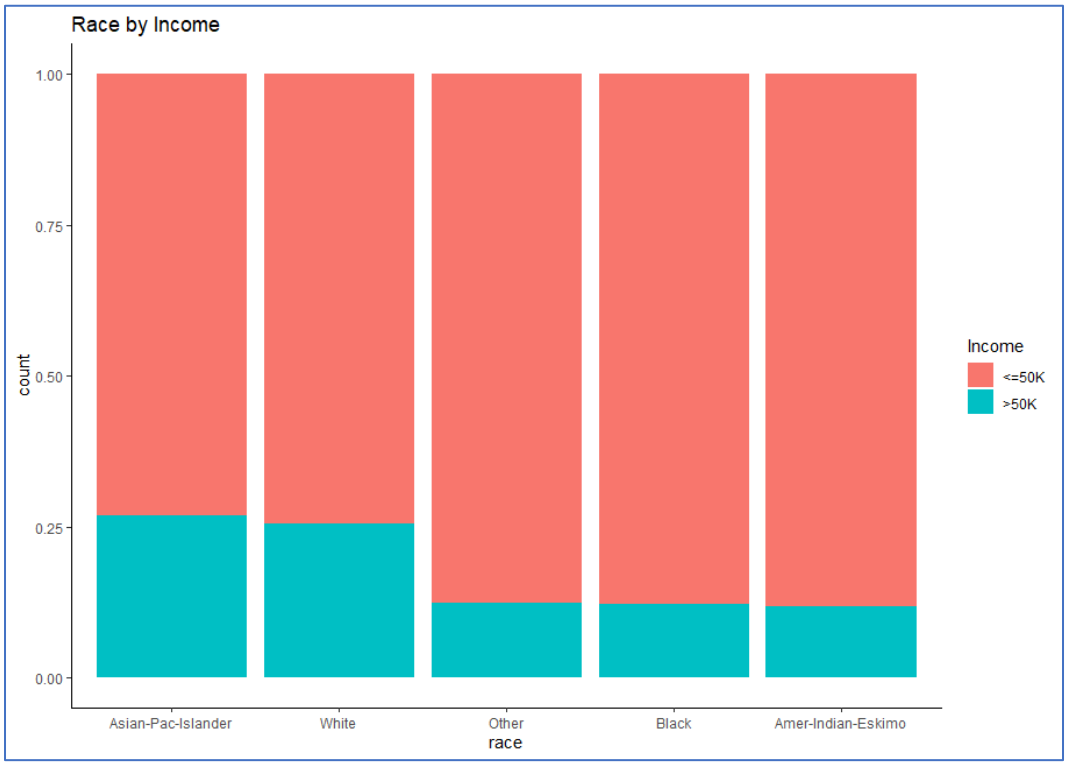
Marital.status and Relationship variables are just about telling us the same things as well, we can see from the graphs that married people have a higher propensity to earn more than 50K.







In terms of Race, it appears that Asian-Pacific-Islander have the highest propensity to earn more while Amer-Indian-Eskimo have the lowest propensity.

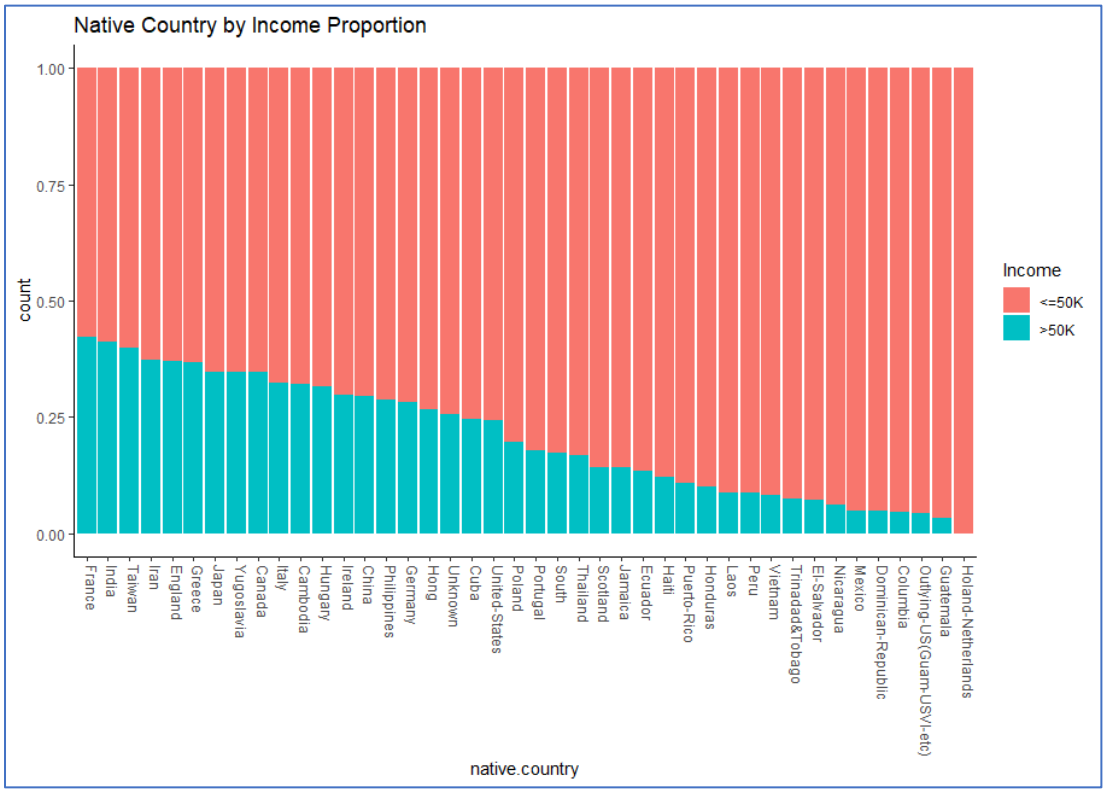


In terms of Native Country, we first want to make sure there is enough observations from each country.

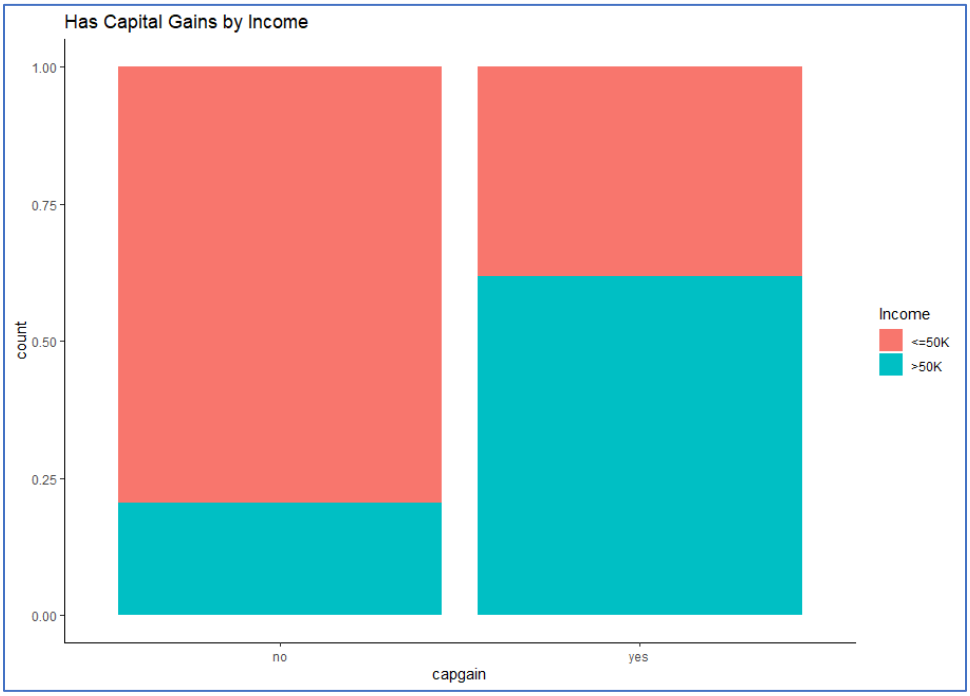
Income	"<=50K"	">50K"
native.country.Cambodia	" 19"	" 9"
native.country.Canada	" 119"	" 63"
native.country.China	" 86"	" 36"
native.country.Columbia	" 81"	" 4"
native.country.Cuba	" 104"	" 34"
native.country.Dominican-Republic	" 98"	" 5"
native.country.Ecuador	" 39"	" 6"
native.country.El-Salvador	" 144"	" 11"
native.country.England	" 80"	" 47"
native.country.France	" 22"	" 16"
native.country.Germany	" 148"	" 58"
native.country.Greece	" 31"	" 18"
native.country.Guatemala	" 85"	" 3"
native.country.Haiti	" 66"	" 9"
native.country.Holand-Netherlands	" 1"	" 0"
native.country.Honduras	" 18"	" 2"
native.country.Hong	" 22"	" 8"
native.country.Hungary	" 13"	" 6"
native.country.India	" 89"	" 62"
native.country.Iran	" 37"	" 22"
native.country.Ireland	" 26"	" 11"
native.country.Italy	" 71"	" 34"
native.country.Jamaica	" 91"	" 15"
native.country.Japan	" 60"	" 32"
native.country.Laos	" 21"	" 2"
native.country.Mexico	" 904"	" 47"
native.country.Nicaragua	" 46"	" 3"
native.country.Outlying-US(Guam-USVI-etc)	" 22"	" 1"
native.country.Peru	" 42"	" 4"
native.country.Philippines	" 210"	" 85"
native.country.Poland	" 70"	" 17"
native.country.Portugal	" 55"	" 12"
native.country.Puerto-Rico	" 164"	" 20"
native.country.Scotland	" 18"	" 3"
native.country.South	" 95"	" 20"
native.country.Taiwan	" 39"	" 26"
native.country.Thailand	" 25"	" 5"
native.country.Trinadad&Tobago	" 25"	" 2"
native.country.United-States	"33138"	"10694"
native.country.Unknown	" 637"	" 220"
native.country.Vietnam	" 79"	" 7"
native.country.Yugoslavia	" 15"	" 8"

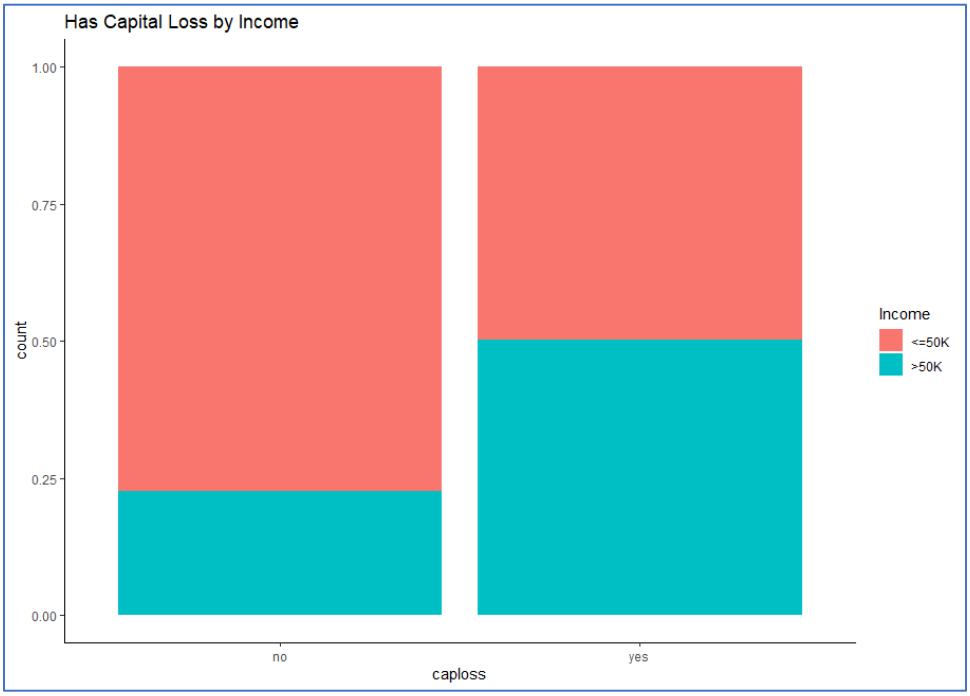
What we find is that Holand only has 1 observation – we should drop that altogether because it won't work with a train/ test split. Other countries that might cause errors are the ones we see with single digit numbers in either column, something we will keep in mind during model building.

France has the highest proportion of people making over $50K, while Guatemala has the lowest, again we are removing Holand-Netherlands.

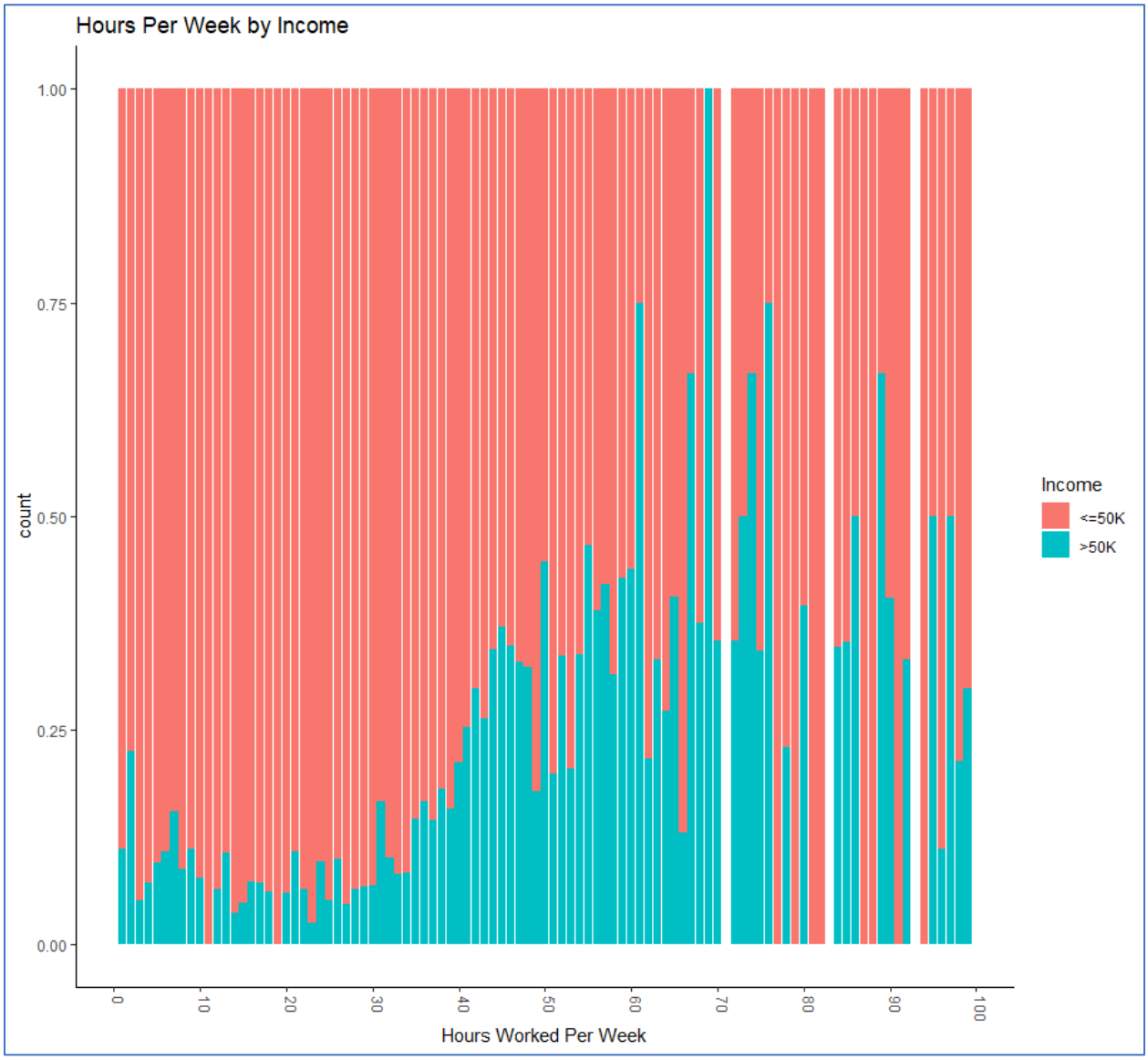


People with capital gains or losses seemingly must have expendable income to be able to invest in order to report on capital gains or losses, that is not to say people making less than $50K don't have capital gains or losses to report since it can include real estate sales, investment accounts, inheritance and a myriad of other savings sources.





Hours.per.week may be a good predictor as well since we can see some separation between number of hours worked and whether or not the observation made over 50K. There appears to be a sweet spot in the graph as well.



---Model Building Part 1/ Objective 1---

Problem of Interest:

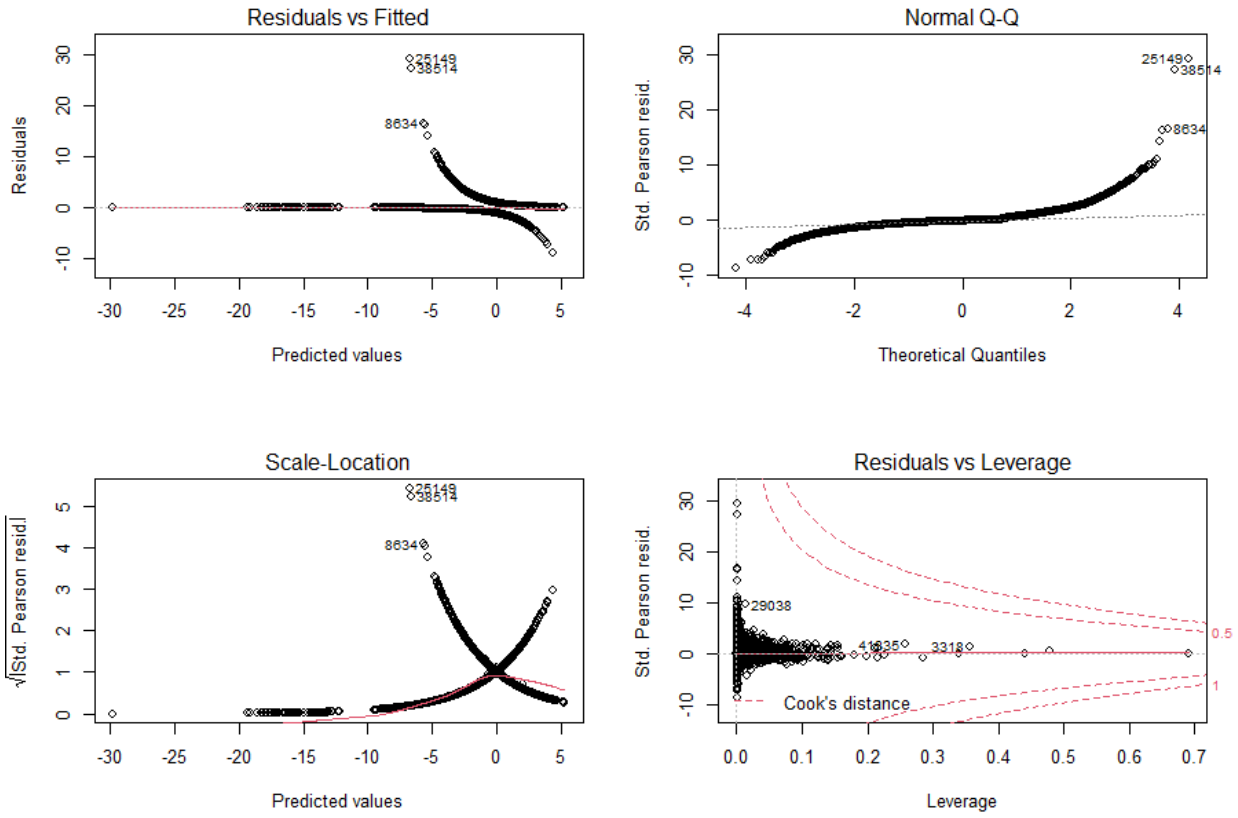
We are exploring models to determine the predictability of someone making over $50,000 a year.

Recall we did PCA in the exploratory data analysis and it determined 2 principal components would explain about 56% of the variation in the data, so we kept that in mind and included PCs as an option for LASSO and Stepwise model building.

LASSO:

We start out our initial model building with LASSO using the glmnet package and the software determined the following variables are the most important in making predictions if someone makes more than $50K annually: *age, workclass, fnlwgt, education, marital.status, occupation, relationship, race, hours.per.week, native.country, capgain and caploss*.

When checking the Cook's D plot we don't see any observations that are over a 1, so we are confident this model meets the assumptions.



We also observe a large GVIF^(½*Df)) value for workclass, something we will explore as well.

```
> car::vif(mod.lasso)
                      GVIF Df GVIF^(1/(2*Df))
age            1.300644e+00  1        1.140458
workclass      9.139114e+06  4        7.415032
fnlwgt         1.041972e+00  1        1.020770
education      2.025908e+00 15        1.023813
marital.status 5.288108e+01  6        1.391901
occupation     1.553451e+07 12        1.993597
relationship   5.457023e+01  5        1.491748
race           2.567736e+00  4        1.125107
hours.per.week 1.147803e+00  1        1.071356
native.country 3.061424e+00 40        1.014084
capgain        1.033741e+00  1        1.016730
caploss        1.015650e+00  1        1.007794
```

When we remove the workclass variable the GVIF^(½*Df)) table looks much better with nothing over a 5.

```
                     GVIF Df GVIF^(1/(2*Df))
age              1.265498  1        1.124944
fnlwgt           1.041546  1        1.020562
education        1.964348 15        1.022760
marital.status  52.669011  6        1.391435
occupation       2.061814 12        1.030609
relationship    54.419106  5        1.491335
race             2.545249  4        1.123871
hours.per.week   1.137041  1        1.066321
native.country   3.009012 40        1.013865
capgain          1.033453  1        1.016589
caploss          1.015758  1        1.007848
```

In comparing the two LASSO models using and not using workclass we get the following results using a 0.265 cutoff, showing that workclass wasn't contributing very much to the model and we can move forward without it. We played with the cutoffs using a variety of different values starting with the standard 0.5. 0.265 actually makes sense when we consider the overall data was split 76/24 for <=$50k/ >$50k.

	Cutoff	Sensitivity	Specificity	Accuracy	AUC
LASSO	0.265	0.809	0.808	0.8094	0.894
LASSO (w/o workclass)	0.265	0.807	0.816	0.8097	0.894

Interpretation of odds ratios

	Odds ratio	2.5 %	97.5 %
(Intercept)	1.316411e-03	3.030498e-04	5.718332e-03
age	1.024086e+00	1.020912e+00	1.027270e+00
fnlwgt	1.000001e+00	1.000001e+00	1.000001e+00
education11th	8.605901e-01	5.723713e-01	1.293942e+00
education12th	1.103515e+00	6.424140e-01	1.895577e+00
education1st-4th	4.164627e-01	1.476985e-01	1.174292e+00
education5th-6th	5.783587e-01	3.097090e-01	1.080042e+00
education7th-8th	4.442053e-01	2.803425e-01	7.038475e-01
education9th	7.364365e-01	4.474377e-01	1.212099e+00
educationAssoc-acdm	3.924576e+00	2.817422e+00	5.466805e+00
educationAssoc-voc	3.297920e+00	2.400899e+00	4.530083e+00
educationBachelors	6.357284e+00	4.735589e+00	8.534325e+00
educationDoctorate	1.713379e+01	1.141995e+01	2.570649e+01
educationHS-grad	2.070232e+00	1.553986e+00	2.757979e+00
educationMasters	9.125046e+00	6.673957e+00	1.247633e+01
educationPreschool	1.793638e-05	1.345373e-94	2.391260e+84
educationProf-school	1.629612e+01	1.117312e+01	2.376809e+01
educationSome-college	3.084019e+00	2.305409e+00	4.125590e+00

We are including just one example of each continuous and categorical variable with this interpretation.

Interpretation of age: When all other predictors are held constant it is estimated that the odds of a person making more than $50,000 a year increase 1.024% for every year a person ages (95% CI(1.02, 1.03)).

Interpretation of education: When all other predictors are held constant, it is estimated that the odds of a person in with a Bachelors degree are 6.4% better than a person that only completed 10th grade (95% CI(4.7, 8.5)) of making an income of more than $50,000.

Stepwise:

Using stepwise and the StepAIC call in R, we show that age, workclass, fnlwgt, education, marital.status, occupation, relationship, race, sex, hours.per.week, native.country, capgain and caploss are all included in the model. The main difference from LASSO's original model is the inclusion of the sex predictor. When we examine the GVIF chart we still notice that workclass has GVIF^(1/(2*Df)) of 7.4, so that tells us it may not be necessary to include.

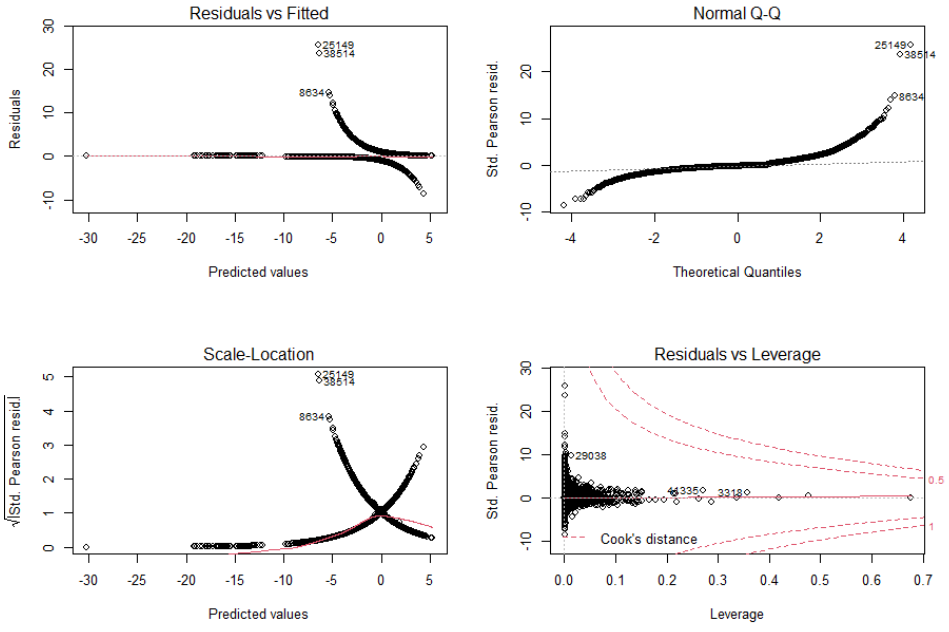
```
> car::vif(mod.step)
                     GVIF Df GVIF^(1/(2*Df))
age          1.304132e+00  1        1.141986
workclass    9.081662e+06  4        7.409189
fnlwgt       1.043736e+00  1        1.021634
education    2.022827e+00 15        1.023761
marital.status 5.413058e+01  6        1.394613
occupation   1.563334e+07 12        1.994124
relationship 1.167935e+02  5        1.609689
race         2.560739e+00  4        1.124723
sex          2.798824e+00  1        1.672969
hours.per.week 1.149810e+00  1        1.072292
native.country 3.058792e+00 40        1.014073
capgain      1.033922e+00  1        1.016819
caploss      1.015851e+00  1        1.007894
```

So we rerun the model without workclass and compare the outcomes, the GVIF table no longer shows any values above 5 in the GVIF^(1/(2*Df)) column.

```
> car::vif(mod.step2)
                     GVIF Df GVIF^(1/(2*Df))
age              1.268450  1        1.126255
fnlwgt           1.043448  1        1.021493
education        1.961381 15        1.022709
marital.status  53.915511  6        1.394150
occupation       2.090908 12        1.031210
relationship   116.476055  5        1.609251
race             2.540550  4        1.123611
sex              2.798710  1        1.672935
hours.per.week   1.139353  1        1.067405
native.country   3.009234 40        1.013866
capgain          1.033660  1        1.016691
caploss          1.015916  1        1.007927
>
```

We also don't see much change in the fit measurements between the 2 models, so it is safe to use the smaller model that excludes workclass.

	Cutoff	Sensitivity	Specificity	Accuracy	AUC
Stepwise	0.265	0.81	0.81	0.81	0.895
Stepwise w/o workclass	0.265	0.81	0.82	0.81	0.895



When checking the Cook's D plot we don't see any observations that are over a 1, so we are confident this model meets the assumptions. With both LASSO and Stepwise, the observations being called out are more than likely from countries that did not have a large enough group of observations, but with the Cook's D being below 1, we remain confident they are not doing too much harm.

The coefficients are very similar to the LASSO model with sex being the only additional predictor in the Stepwise model.

Interpretation of odds ratios

	Odds ratio	2.5 %	97.5 %
(Intercept)	6.594002e-04	1.509108e-04	2.881229e-03
age	1.024118e+00	1.020930e+00	1.027315e+00
fnlwgt	1.000001e+00	1.000000e+00	1.000001e+00
education11th	8.595063e-01	5.718491e-01	1.291864e+00
education12th	1.077421e+00	6.270437e-01	1.851285e+00
education1st-4th	4.174211e-01	1.482233e-01	1.175526e+00
education5th-6th	5.837343e-01	3.126401e-01	1.089898e+00
education7th-8th	4.474164e-01	2.824618e-01	7.087027e-01
education9th	7.367694e-01	4.478434e-01	1.212096e+00
educationAssoc-acdm	3.928047e+00	2.819470e+00	5.472501e+00
educationAssoc-voc	3.333484e+00	2.426601e+00	4.579291e+00
educationBachelors	6.272628e+00	4.672823e+00	8.420147e+00
educationDoctorate	1.659855e+01	1.105480e+01	2.492237e+01
educationHS-grad	2.071624e+00	1.555166e+00	2.759595e+00
educationMasters	9.181626e+00	6.712384e+00	1.255921e+01
educationPreschool	1.808468e-05	5.069916e-95	6.450907e+84
educationProf-school	1.556746e+01	1.067167e+01	2.270926e+01
educationSome-college	3.084949e+00	2.306292e+00	4.126498e+00

Again, we are only going to interpret one of each continuous and categorical variable.

Interpretation of age: When all other predictors are held constant it is estimated that the odds of a person making more than $50,000 a year increase 1.024% for every year a person ages (95% CI(1.021, 1.027)).

Interpretation of education: When all other predictors are held constant, it is estimated that the odds of a person in with a Bachelors degree are 6.27% better than a person with just a 10th grade education (95% CI(4.67, 8.42)).

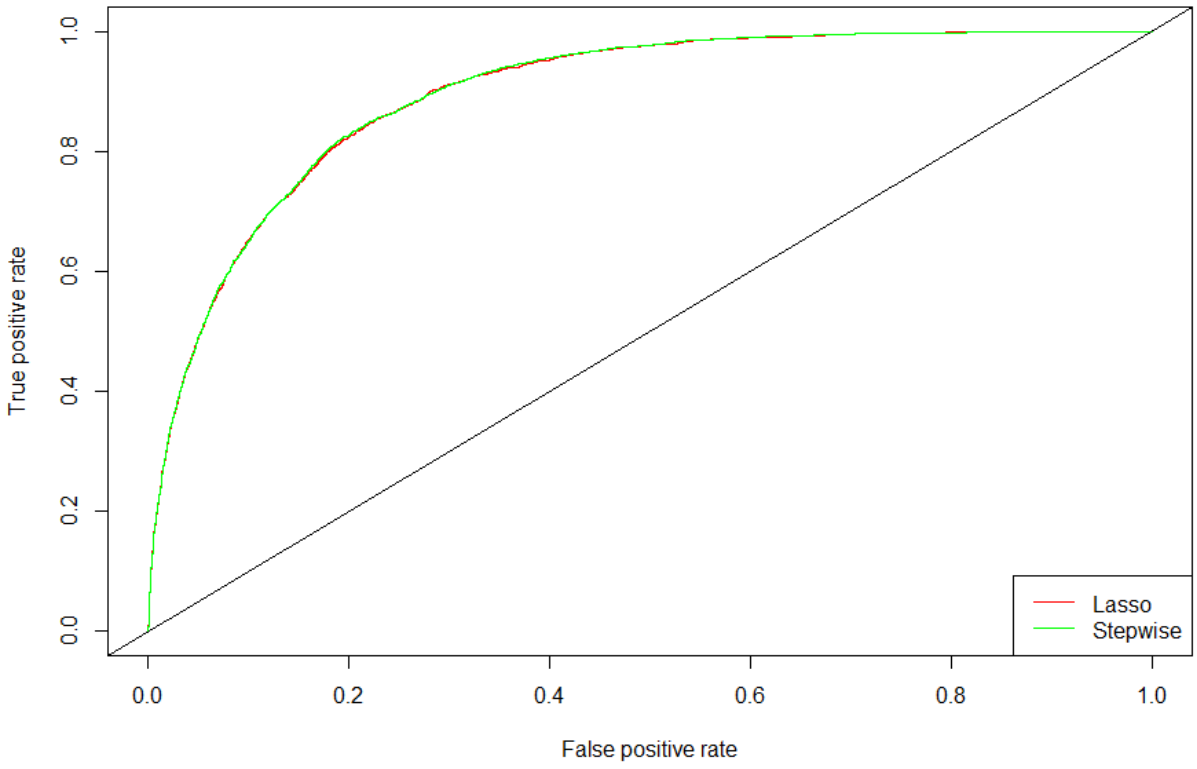
Summary for Objective 1:

Using confusion matrices, accuracy tests and ROC curves, we see that either model performs about the same with Stepwise performing slightly better (note neither final model uses the workclass variable). After adjusting the cutoffs for each model, we found the best was a cutoff at 0.265 and that is pretty close to the 76: 24 split ratio between the incomes of under $50,000 and over $50,000.

```
[1] "Confusion matrix for LASSO w/o workclass"
> conf.lasso

class.lasso <=50K  >50K
      <=50K 10018   710
      >50K   2388  3165
[1] "Confusion matrix for Stepwise w/o workclass"
> conf.step2

class.step2 <=50K  >50K
      <=50K 10027   702
      >50K   2379  3173
```

	Cutoff	Sensitivity	Specificity	Accuracy	AUC
LASSO	0.265	0.81	0.82	0.81	0.894
Stepwise	0.265	0.81	0.82	0.81	0.895

The ROC curves nearly perfectly align with each other.



---Model Building Part 2/ Objective 2---

Overview for Objective 2:

We would like to build and compare models that will compete against our original Stepwise and LASSO models. We will sacrifice the ability to interpret coefficients in the hopes that some interactions we include can help yield better accuracy. We will also build an LDA and a QDA model to see if a different approach can help answer our question of interest better.

Complex Model:

Seeing the performance of the LASSO model, we decided to build a more complex model based on the original predictors in that model. In addition to the original predictors we added an interaction term of age*native.country along with 5 polynomials for age (age^2 * through age^6). After some trial and error, we decided to use a cutoff of 0.25 for this model since the unbalanced data could cause our accuracy for prediction to suffer. With all predictors testing as significant we ran the model to test its performance. Overall, this model seemed to perform nearly as well as both Stepwise and LASSO from part 1.

	Cutoff <dbl>	Sensitivity <dbl>	Specificity <dbl>	Accuracy <dbl>	AUC <dbl>
LASSO.Complex	0.25	0.7990488	0.8291613	0.8062158	0.8982669

LDA without PCA:

For our second competing model we created an LDA model which only contained our original (non-PCA) predictors. These predictors were *age, fnlwgt, education.num,* and *hours.per.week.* This model's overall accuracy was close to what the previous models were, however, the specificity (the ability to accurately classify observations over 50K) was only at 31.56% which means that a large amount of our over 50K observations were not classified correctly. This severely hurts our predictive power and won't be the model we will decide to use in the end.

LDA with PCA:

We decided to try an LDA model that only contained PCA predictors to see how it would perform against the other models. We decided to use all 4 PCs that were generated to get the most predictive power possible. We found that using PCA with LDA performed exactly the same as our LDA model that only used the continuous predictors. Like the other LDA model we won't select this as our final model due to its inability to classify over 50K accurately.

QDA with PCA:

Our final model we created was a QDA model using PCA as predictors in the model. We found that this model performed nearly the same as our LDA models, it even suffered from inaccuracy in predicting the over 50K class. With the model performing worse in overall accuracy than the Logistic models and having a low specificity, there is no reason to consider this model for our final selection.

Overall results:

	Cutoff <dbl>	Sensitivity <dbl>	Specificity <dbl>	Accuracy <dbl>	AUC <dbl>
LASSO	0.265	0.8076737	0.8167742	0.8098397	0.8940707
Stepwise	0.265	0.8081573	0.8185806	0.8106382	0.8949643
Complex LASSO	0.250	0.7990488	0.8291613	0.8062158	0.8982669
LDA	*NA*	0.9383363	0.3156129	0.7901235	*NA*
LDA.PCA	*NA*	0.9383363	0.3156129	0.7901235	*NA*
QDA.PCA	*NA*	0.9397872	0.2993548	0.7873595	*NA*

Overall, we decided that either Stepwise or LASSO would be acceptable models. Ultimately though we think that Stepwise is the slightly better option and would perform all future analysis based on this model. This model performed exceptionally well with an accuracy of 81%, all while accurately classifying both the over 50K and the below 50K observations.

 # --- Appendix --- #

R Markdown:

The final R Markdown file can be found at: https://github.com/justinehly/6372---50k-Income-Predictor/blob/main/R_Markdown/Income%20Predictor.Rmd

Data files:

Data descriptions can be found here: https://github.com/justinehly/6372---50k-Income-Predictor/blob/main/Data/adult.names.txt

Data files adult.data.csv and adult.test.csv can be found here: https://github.com/justinehly/6372---50k-Income-Predictor/tree/main/Data