**--- Data Cleaning ---**

Data was presplit 2/3 and 1/3 from the online repository, after exploring the data further it appears 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 “? “ those were replaced with, “Unknown,” to more accurately describe what they represent. The only variables with that sort of missing data were *workclass* = 1836, *occupation* = 1843, *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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

Class was changed to Income to be more descriptive and easier to explain.

Finally, we removed all before and after whitespace from categorical variables to help with analyzing the data and changed the origin response variable from *class* to *Income*.

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

Capital.gain/ capital.loss variables

There were also 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 it makes more sense when predicting if someone makes more than $50k annually and removed the original numeric variables.

Workclass: Further exploring workclass, there are so few govermental jobs, it looks like it makes sense to merge those together as well as unpaid with unknown. Just to confirm a logical regression analysis obtaining p-values from z-values was used and it indeed made sense to combine these factor levels to end up with just 5 factor levels from the original 9. Proportions between and among levels is below showing more reasonable weight per level.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |

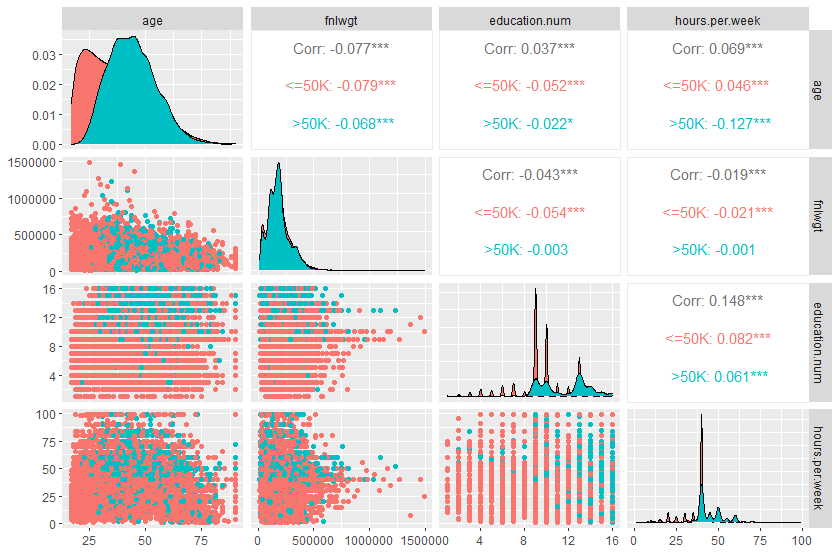
Occupation, reviewing the breakdown of how many observations are within each factor level of occupation, Armed Services represents just 15 of the 48,842 observations or 0.03% of the total, essentially giving it very little predictive power, but after a logistic regression test we see the pvalue = 0.03 from zvalue, 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Overall Proportion | <=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 |

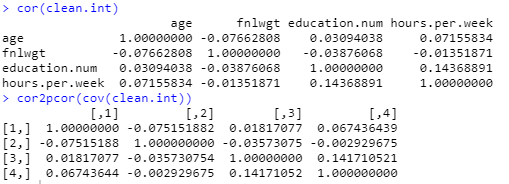
For marital.status, we notice that there are very few Married-AF-spouse (married armed forces spouse) observations at 37, maybe it makes more sense to just combine those with married-civ-spouse (married civilian spouse). We also see from a glm that Married-spouse-absent is not statistically significant 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.

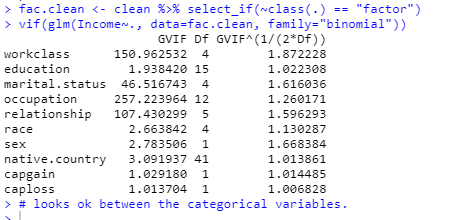


We went ahead and performed PCA next to see what the R would tell us about the continuous variables and how many it thinks we need in our final model. The PCA results show only 2 are necessary, however this only explains about 60% of the variability, so that is something to keep in mind in feature selection.

Chart, line chart

Description automatically generated

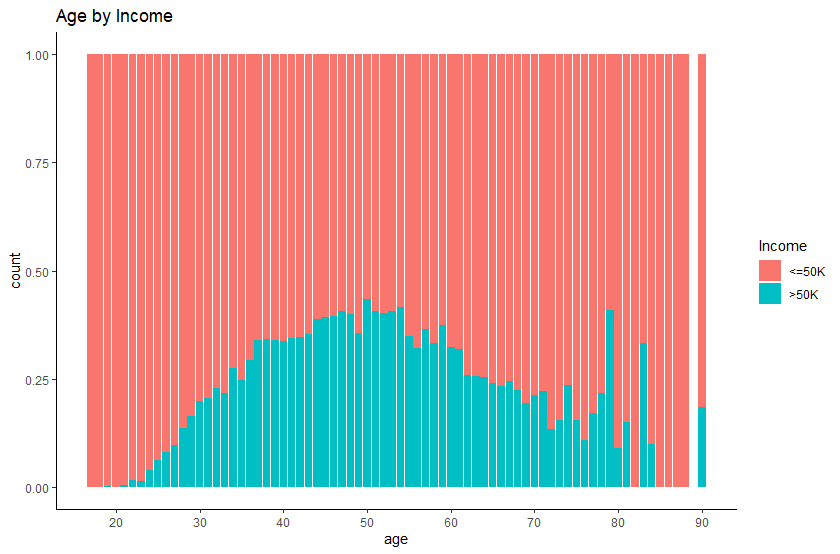
Next we want to see if there is multicollinearity across the dataset using VIFs.

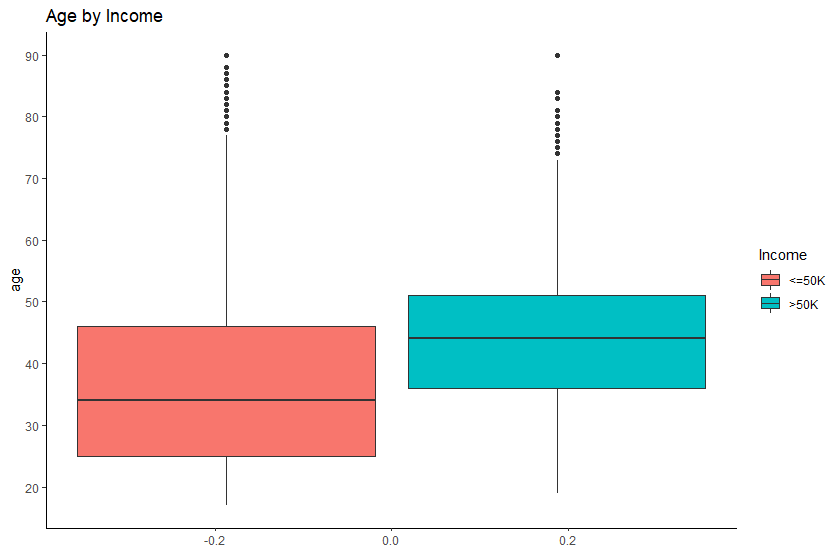


Based on the GVIF^(1/2\*Df) all being relatively small, even when compared to 5 or 10, we should be ok to model with all these variables to start, but curiosity points to whether a continuous variable and a categorical variable might be telling us the same thing, such as the education and education.num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

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

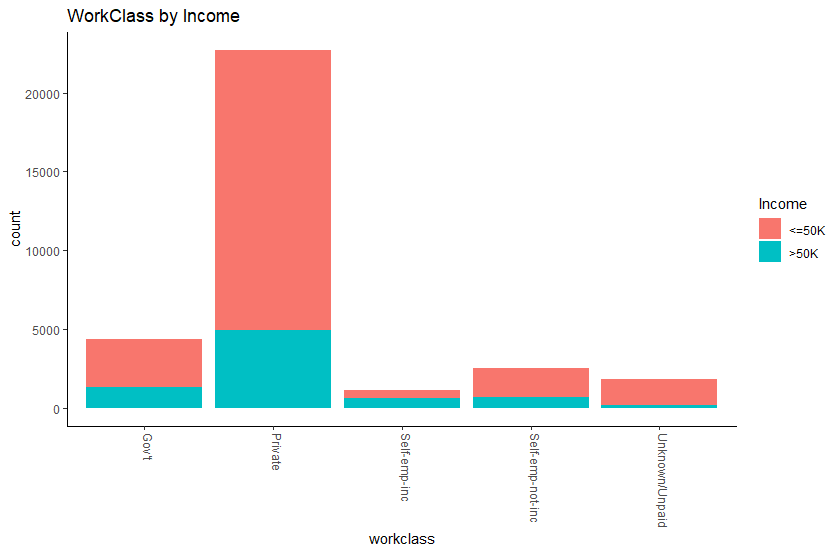
|  |  |  |
| --- | --- | --- |
| Income | "<=50K" | ">50K" |
| age.Min. | "17.00000" | "19.00000" |
| age.1st Qu. | "25.00000" | "36.00000" |
| age.Median | "34.00000" | "44.00000" |
| age.Mean | "36.78374" | "44.24984" |
| age.3rd Qu. | "46.00000" | "51.00000" |
| age.Max. | "90.00000" | "90.00000"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

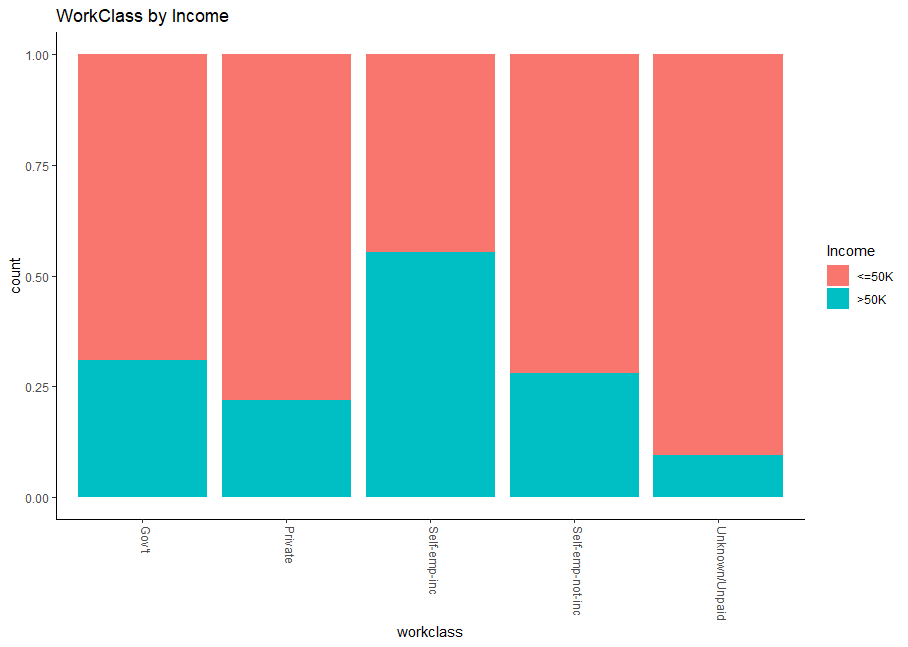




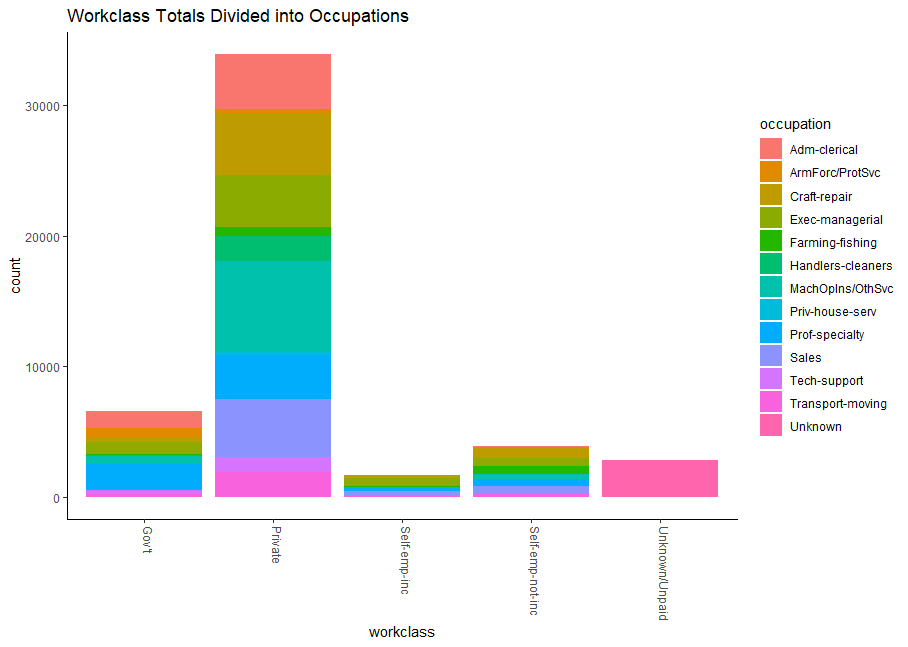
Fnlwgt is what the census from each country assumes is the total number of people meeting all the criteria is each row, it can be used a weighting metric as well, but for prediction we will need to explore how useful it really is when we get to testing via the glm statement, that returns a pvalue=0.0878 from a z-value with a coefficient 95% confidence interval(-4.55e-07, 3.05e-08) that crosses 0, we can safely not worry about this variable moving forward.

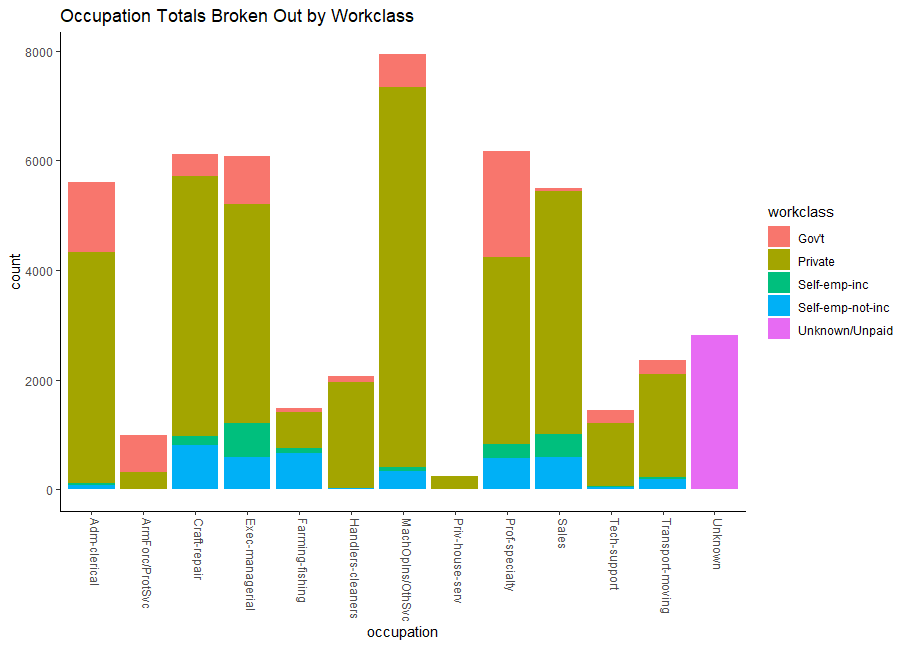
Workclass being reduced to just 5 levels from the original 9 shows that private has the most people with about 70% while unknown/ unpaid has the least. When we graph the variable and break it down by Income, we see that there may very well be some good predictive power with this variable, while most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

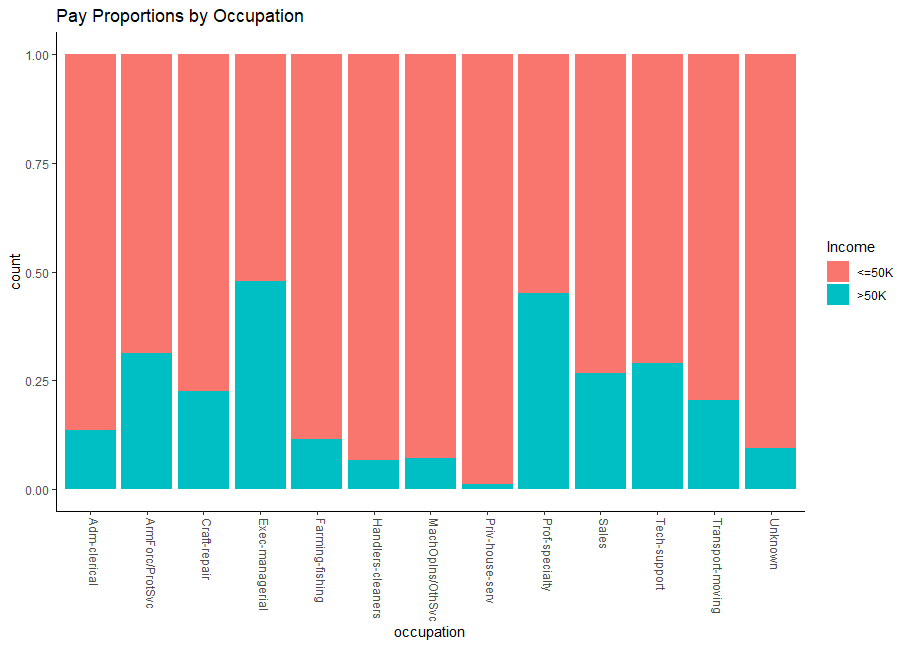




Next we look at Occupation, or a more specific label or category of what each subject does within their workclass, what may be interesting here is how different occupations are paid between working classes or industries.

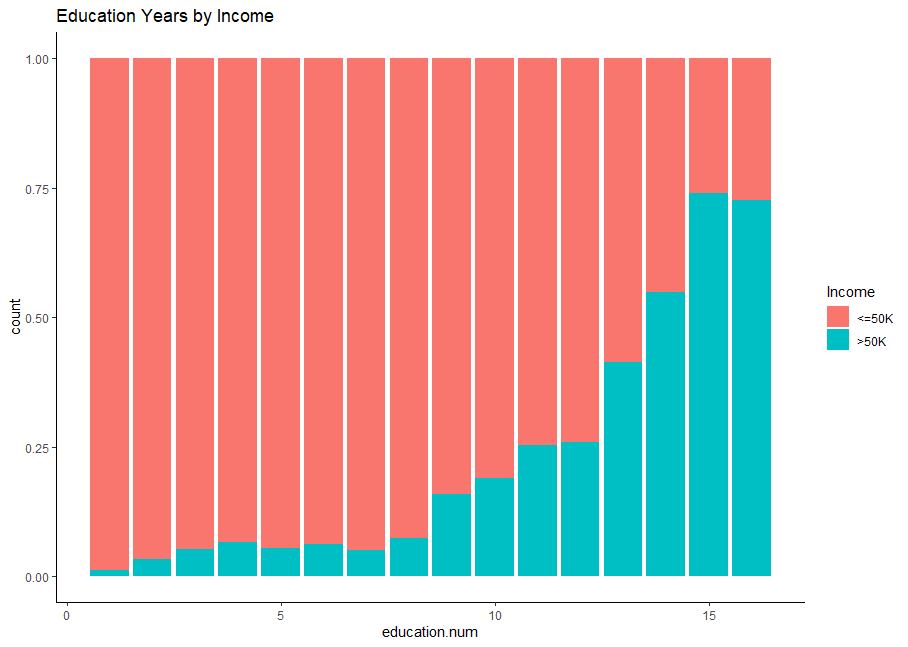


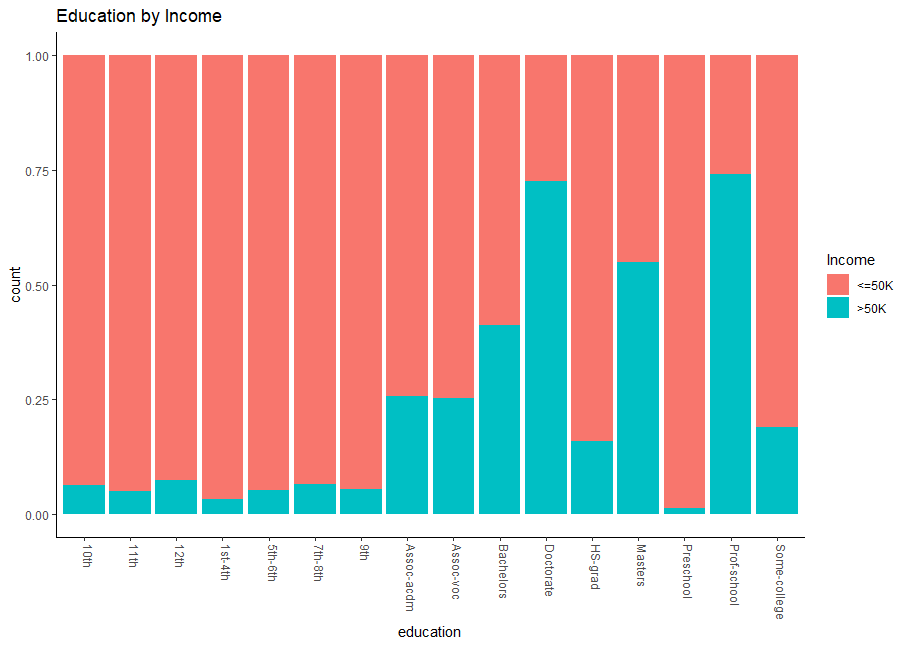


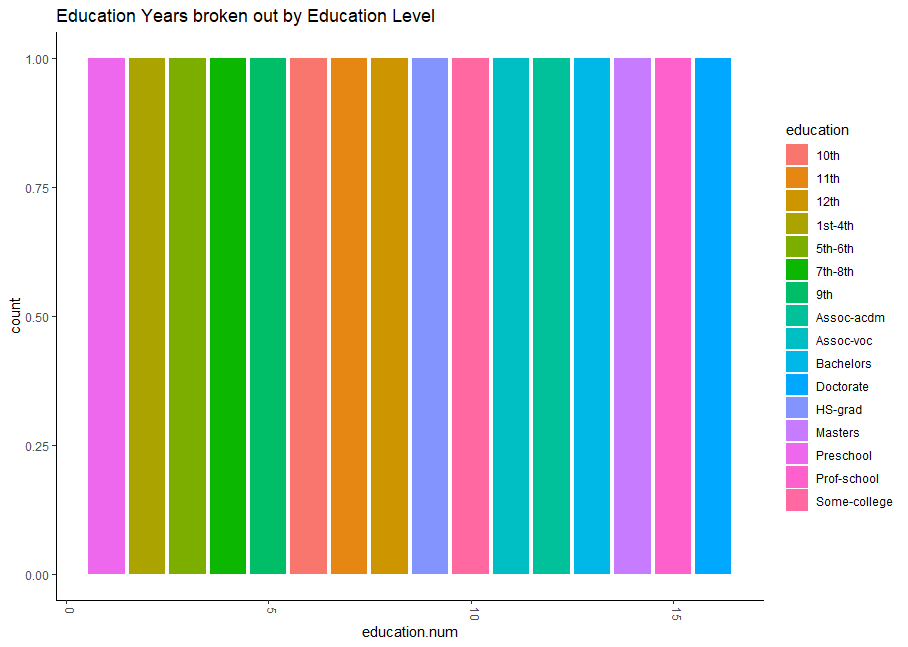


Exec- Managerial and Professional Specialty have the highest proportions of people making over $50k.

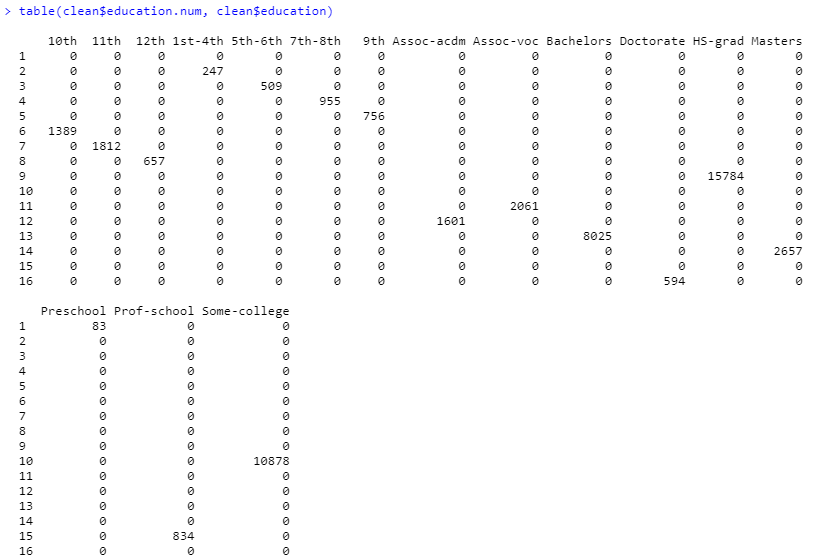
Reviewing education, the education and education.num are essentially telling us the same thing and we confirm that with both visuals and in a table…the hypothesis is that people with more education make more money.





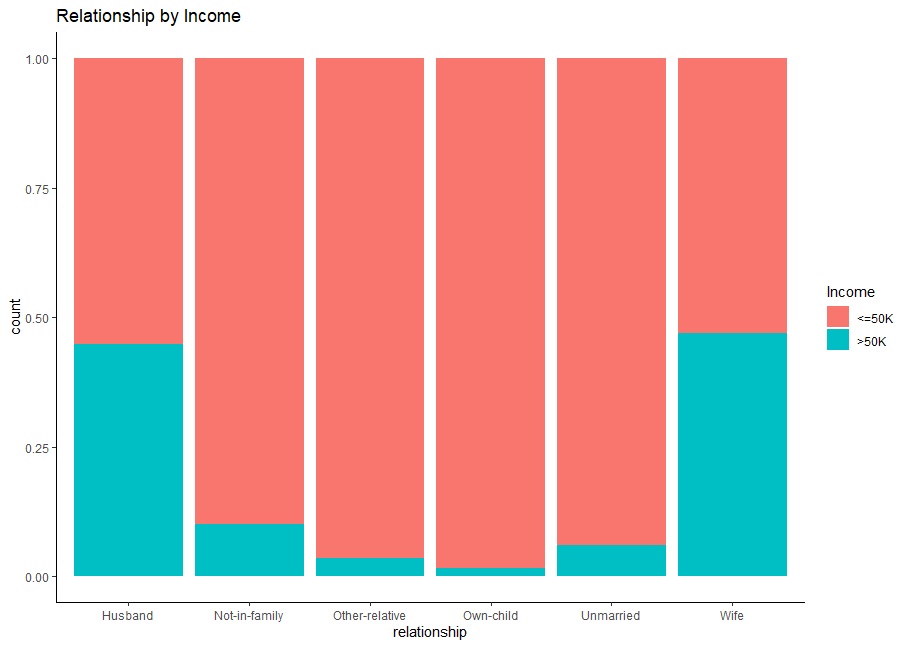


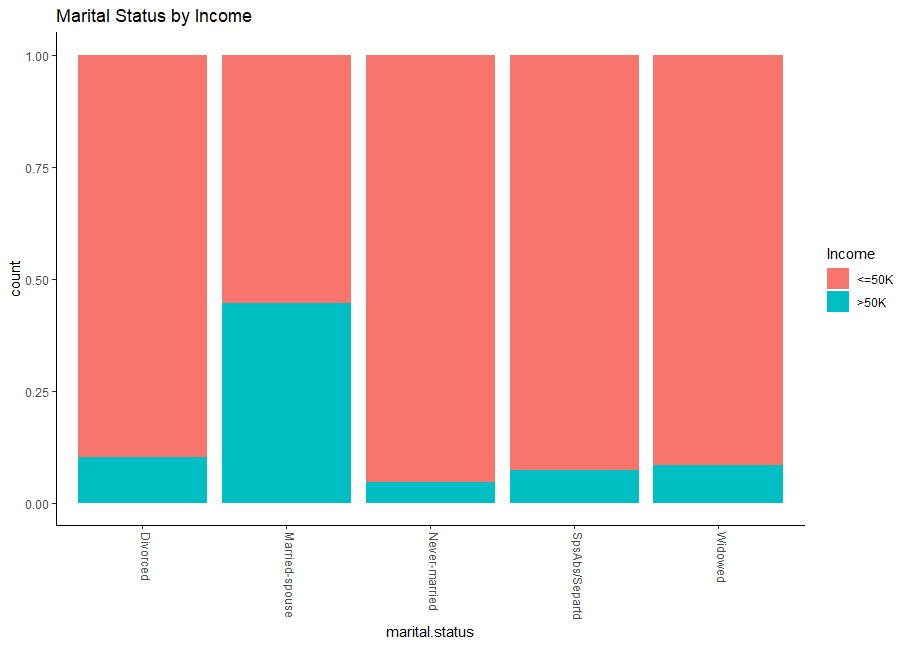
A table further confirms this.

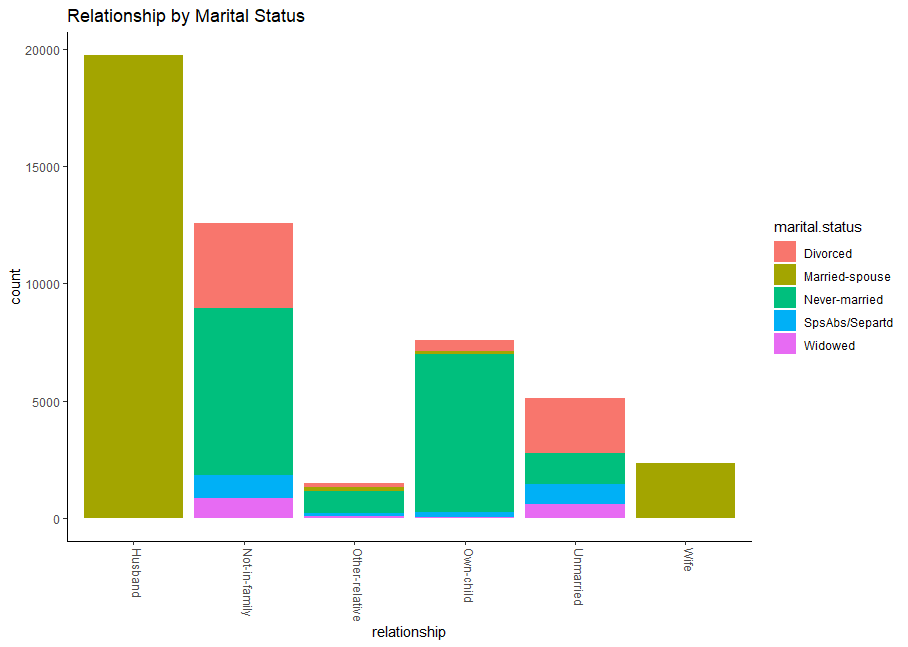


They are telling us the same thing, while the continuous variable seems to make more sense comparing number of years of education to earning potential, but it may be useful to also have a categorical label on it as well. Here we can see there isn’t any overlap between education years and the category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

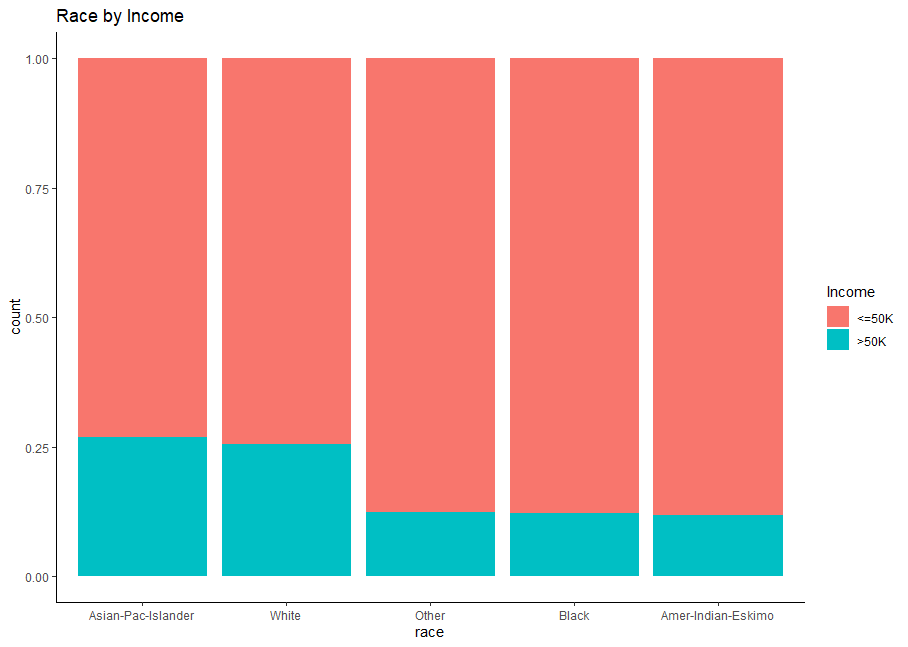
Marital.status and Relationship variables are just about telling us the same things as well, we can see from the graphs that married people have a higher propensity to each more than 50K.



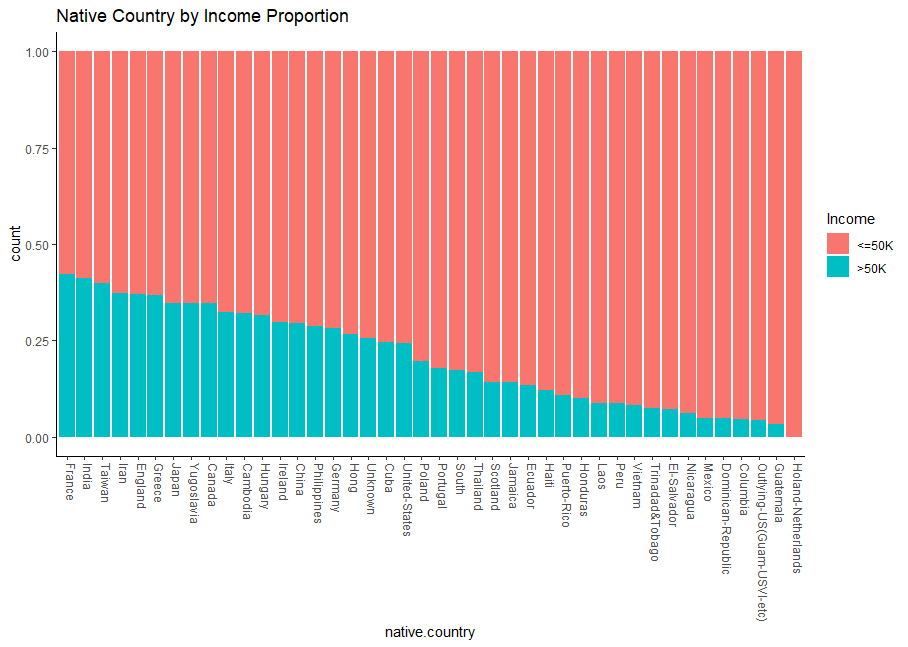




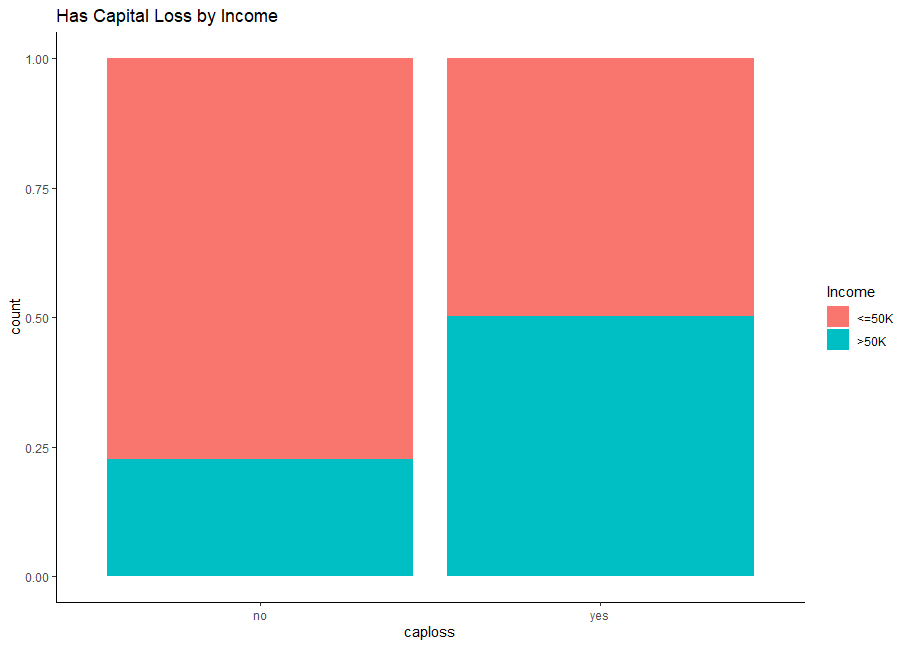
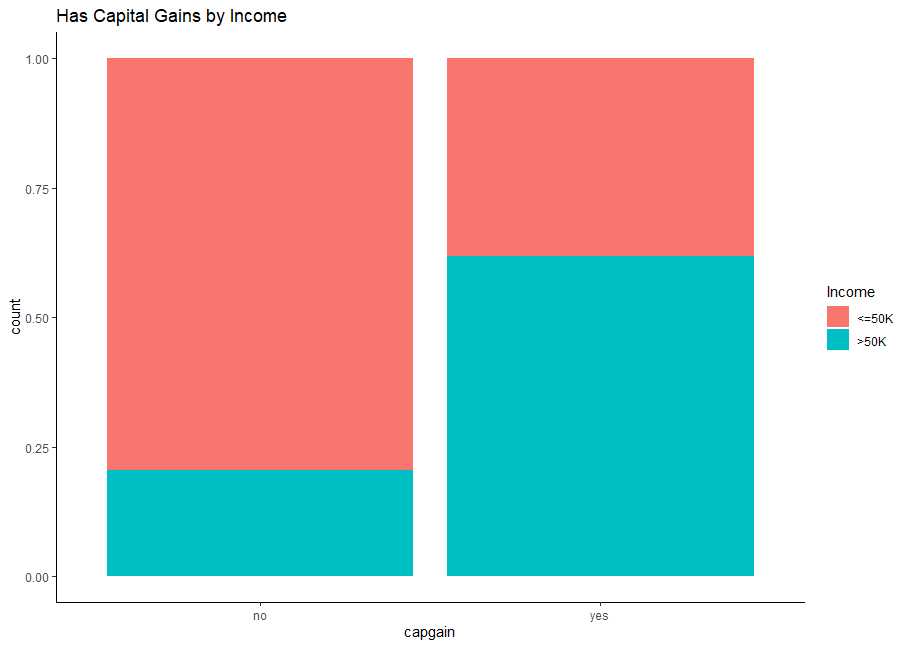
In terms of Race, it appears that Asian-Pacific-Islander have the highest propensity to earn more while Amer-Indian-Eskimo have the lowest propensity.



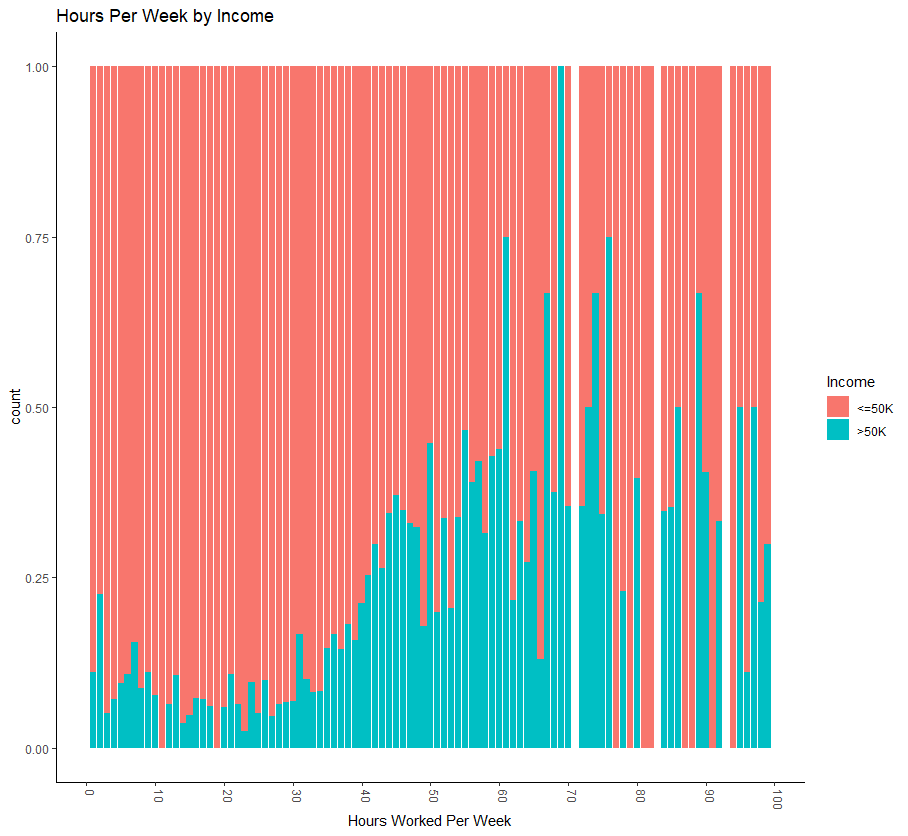
In terms of Native Country, France has the highest proportion of people making over $50K, while Holand-Netherlands has the smallest.



People with capital gains or losses seemingly have to 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 --- #**

Recall we did PCA in the exploratory data analysis and it determined only 2 principal components were necessary.

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, caploss*.

Using stepwise and the StepAIC call in R we show that age, workclass, flnwgt, education, marital.status, occupation, relationship, race, hours.per.week, native.country, capgain and caploss are all included in the model. Notice that both LASSO and Stepwise returned the same results in terms of which independent variables to use.

When we check the residuals from the LASSO glm we find there are