Data Cleaning

Data was presplit 2/3 and 1/3 from the online repository, so we kept those splits and mirrored any changes made to the training dataset to the test dataset.

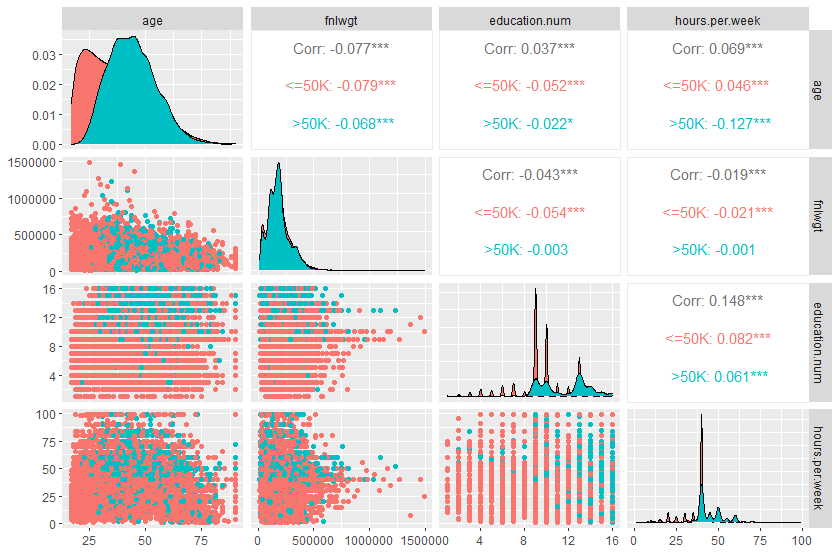
Originally all NA’s were represented by “? “ those were replaced with, “Unknown,” to more accurately describe what they represent. The only variables with that sort of missing data were *workclass* = 1836, *occupation* = 1843, *native.country* = 583. It should also be noted that all missing values for workclass were also missing for occupation, with the *workclass* = “Never-worked” that also reported as an “Unknown” in the *occupation* variable.

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.

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*.

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.

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"

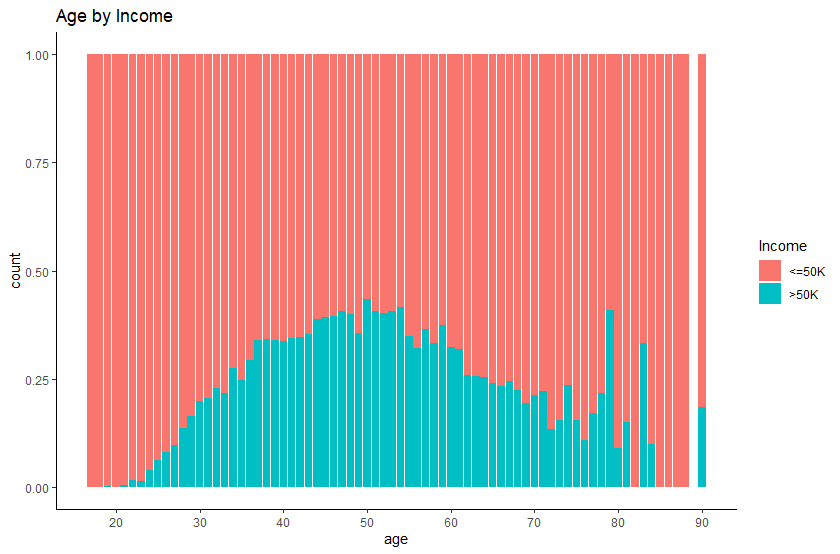
age.1st Qu. "25.00000" "36.00000"

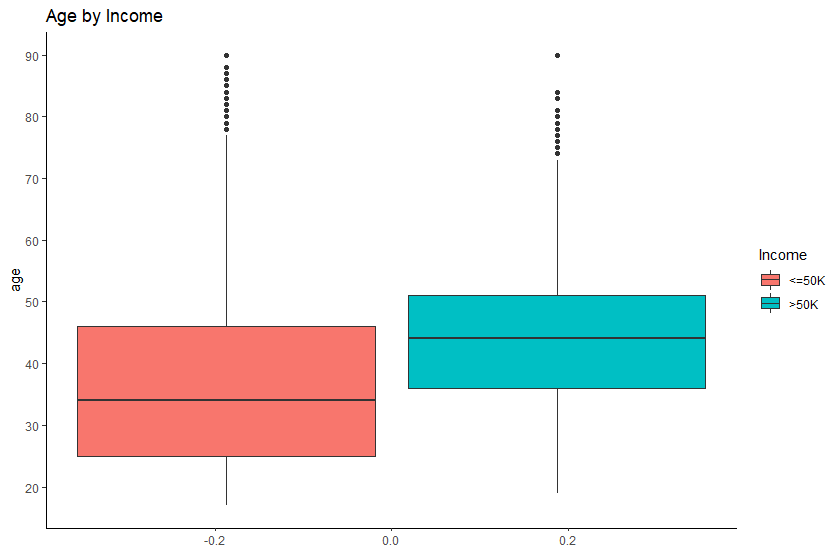
age.Median "34.00000" "44.00000"

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 further exploring workclass, there are so few govermental jobs, it looks like it makes sense to merge those together, also we can probably merge unpaid with unknown.

Workclass proportions

Federal-gov 0.0294831240

Local-gov 0.0642793526

Private 0.6970301895

Self-emp-inc 0.0342741316

Self-emp-not-inc 0.0780381438

State-gov 0.0398636406

Unknown 0.0563864746

Unpaid 0.0006449433

After reworking the data – we get the following proportins between factors and within factors

workclass proportions

Gov't 0.13362612

Private 0.69703019

Self-emp-inc 0.03427413

Self-emp-not-inc 0.07803814

Unknown/Unpaid 0.05703142

<=50K >50K

Gov't 0.092441878 0.041184239

Private 0.544608581 0.152421609

Self-emp-inc 0.015171524 0.019102607

Self-emp-not-inc 0.055802954 0.022235189

Unknown/Unpaid 0.051165505 0.005865913

