Dataiku Census Analysis

Descriptive Statistics

The first part of this mini-project consists of doing some quick and basic data visualisation and descriptive statistics to understand the data we are dealing with. From this part to the end, I will assume that the reader of this document has fully read and understood the metadata.txt file, in which we find a lot of basic information concerning the dataset. In order to keep this report light I will only display the interesting parts of the code and graphics. You can find the rest of my code in census.r. The most challenging part of this project is the big skewness of the data, this dataset is indeed really unbalanced.

Import the dataset into R

I used the capabilities of R to clean the whitespaces and set directly the contextual information about the columns (details in the code). I dropped the weight column as it useless for the classifier (cf metadata). I also change the label by 0 and 1 for readability purpuses on my graphs. 0 : -50000 and 1 : 50000+

```
train_df <- read.csv(train_location, header = F, na.strings = '?', col.names = context,
            strip.white = T, colClasses = type_context)

train_df <- subset(train_df, select = -c(instance_weight) )

train_df$income <- ifelse(train_df$income == "- 50000.", "0",
                   ifelse(train_df$income == "50000+.", "1", "other"))
```

Evaluate percentage of missing values

```
incomplete_columns <- sapply(train_df, function(x) (sum(is.na(x)) / nrow(train_df)))*100; incomplete_col
```

```
##    reg_prev_state     migration_msa     migration_reg mig_within_region
##         0.3548463        49.9671717        49.9671717        49.9671717
## migration_sunbelt    country_father    country_mother      country_self
##        49.9671717         3.3645244         3.0668144         1.7005558
```

As you can notice, there is a lot of data missing in the migration columns. After further analysis I decided to drop them but that will be explained next.

Visualising the features

To visualize the categorical features, I will use this kind of graph. That is made very easily by the ggplot2 library

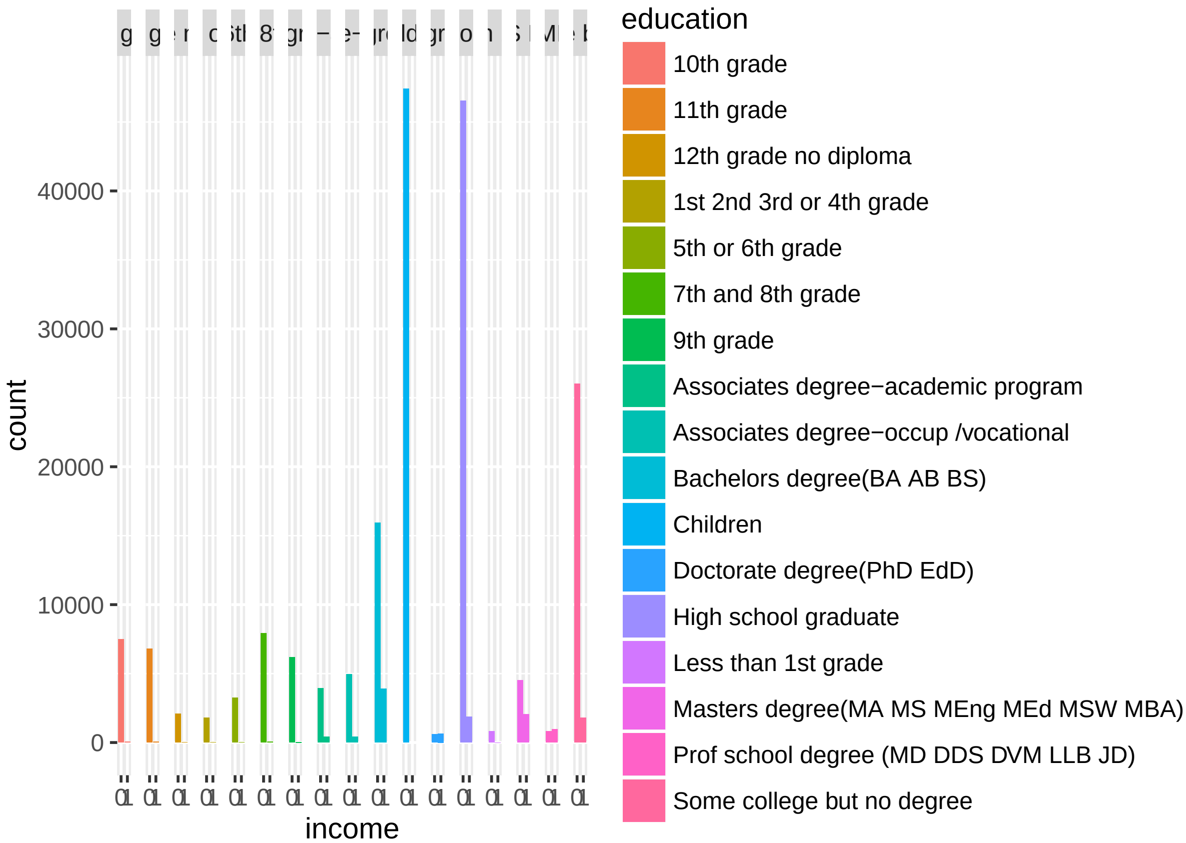
Class of Worker Feature

```r
qplot (income, data = train_df, fill = class_of_worker) + facet_grid (. ~ class_of_worker)
```

People that are self-employed-incorporated are more likely to earn +50K. Self incorporated people along with people from the private sector and the federal governement seems also advantadged.

Education Feature

```r
qplot (income, data = train_df, fill = education) + facet_grid (. ~ education)
```

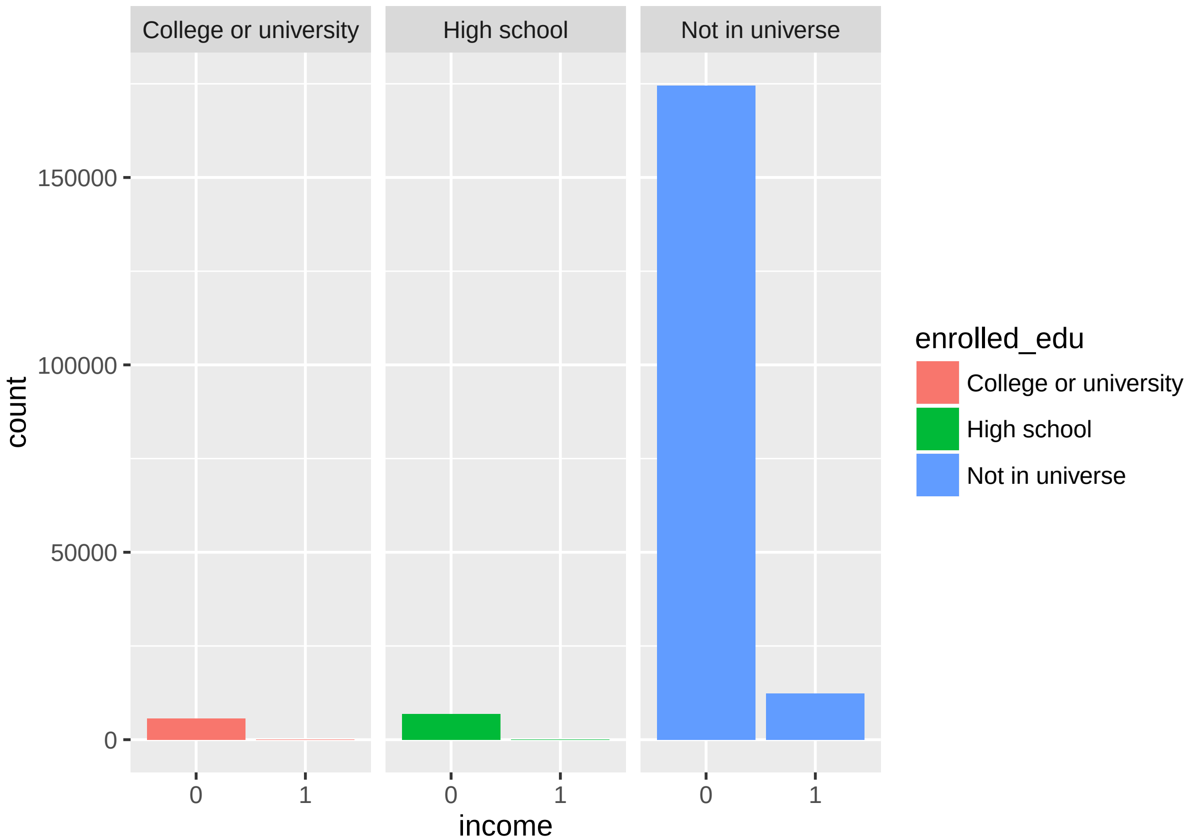


As it was predictable, the more people study the more they are likely to earn +50K. This will be a very important feature for our classifier. To make this task easier I reduced the number of category. The "under 18" for example are very unlikely to make + 50K, I gathered them into one group, and did the same for the different masters and the bachelors.

```
train_df$education_cat<-ifelse(train_df$education == "10th grade", "youth",
                    ifelse(train_df$education == "11th grade", "youth",
                    ifelse(train_df$education == "12th grade no diploma", "youth" ,
                    ifelse(train_df$education == "1st 2nd 3rd or 4th grade", "youth",
                    ifelse(train_df$education == "5th or 6th grade", "youth",
                    ifelse(train_df$education == "7th and 8th grade", "youth",
                    ifelse(train_df$education == "9th grade", "youth",
                    ifelse(train_df$education == "Less than 1st grade", "youth",
                    ifelse(train_df$education == "Children", "youth",
                    ifelse(train_df$education == "Associates degree-academic program", "basicdegree",
                    ifelse(train_df$education == "Associates degree-occup /vocational", "basicdegree",
                    ifelse(train_df$education == "Some college but no degree", "basicdegree",
                    ifelse(train_df$education == "High school graduate", "high school graduate",
                    ifelse(train_df$education == "Bachelors degree(BA AB BS)", "bachelor",
                    ifelse(train_df$education == "Masters degree(MA MS MEng MEd MSW MBA)", "master",
                    ifelse(train_df$education == "Doctorate degree(PhD EdD)", "prof_doct",
                    ifelse(train_df$education == "Prof school degree (MD DDS DVM LLB JD)", "prof_doct",
                    )))))))))))))))))
```

Enrolled in edu last week

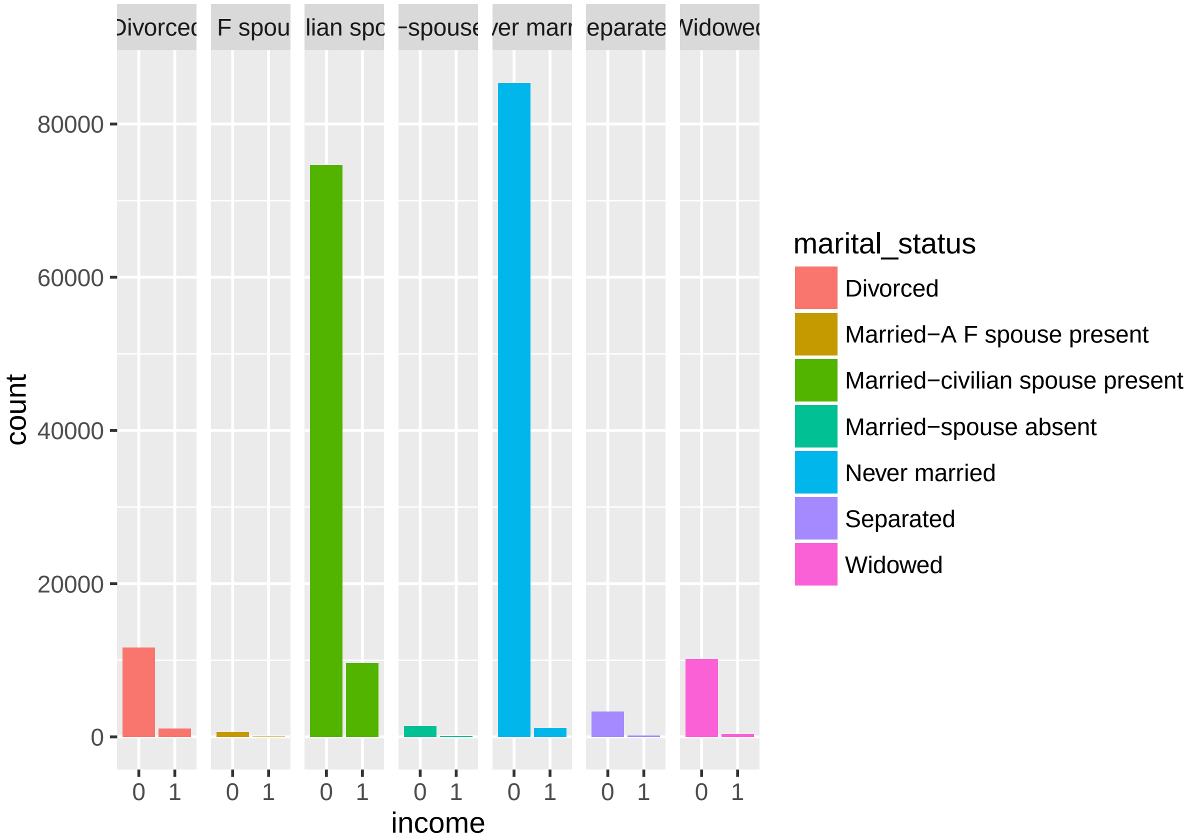
```
qplot (income, data = train_df, fill = enrolled_edu) + facet_grid (. ~ enrolled_edu)
```



As expected, people who are not out of college or high school don't make any money, this is a good feature to discriminate the class - 50K

Marital Status

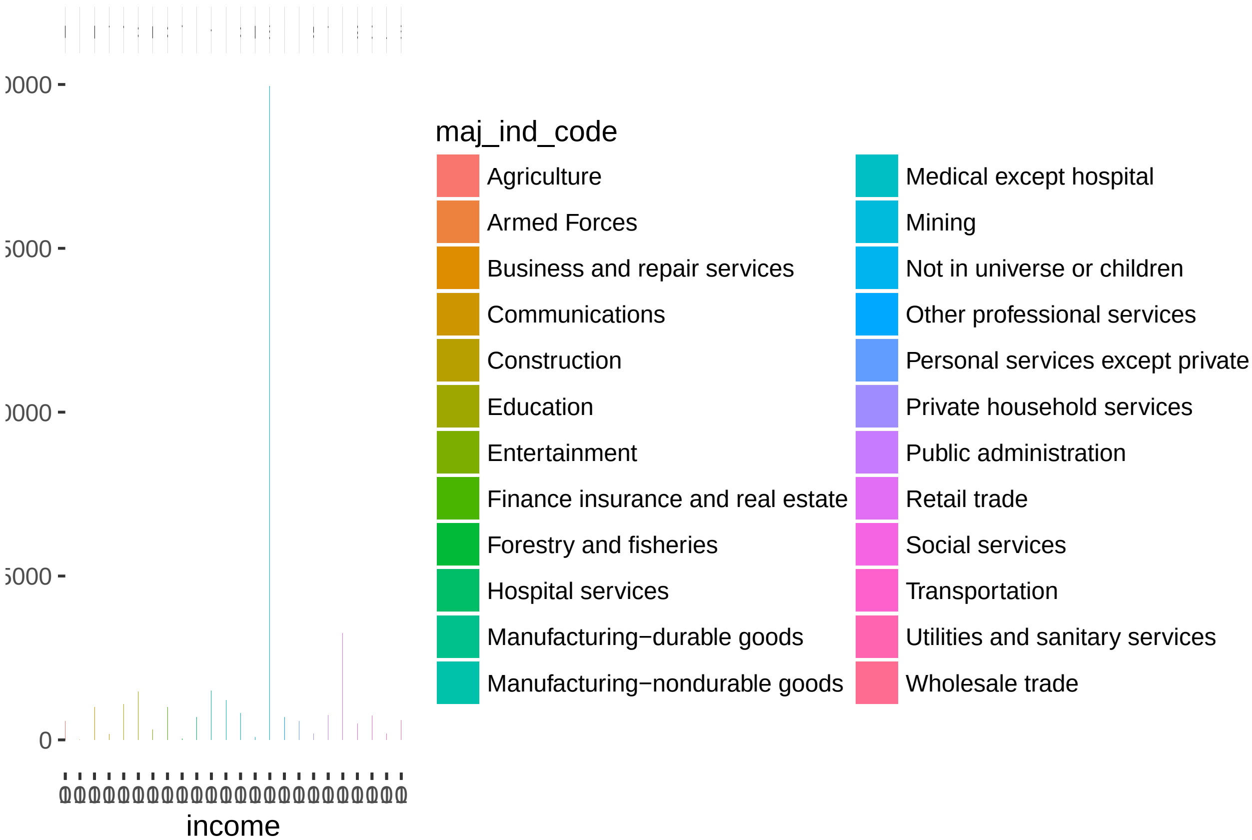
```
qplot (income, data = train_df, fill = marital_status) + facet_grid (. ~ marital_status)
```



There are higher percentages of the +50K earner in the married-civilian spouse present and divorced population than in the other classes.

Major industry code

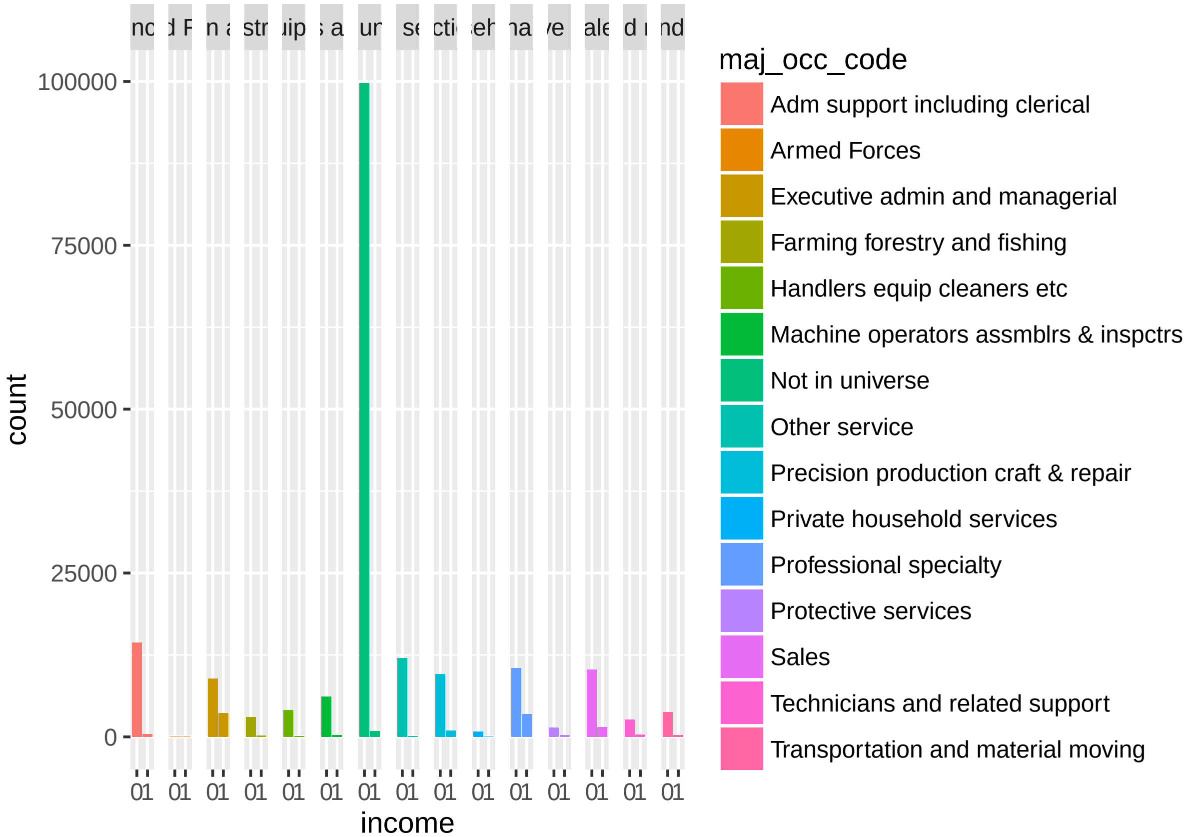
```
qplot (income, data = train_df, fill = maj_ind_code) + facet_grid (. ~ maj_ind_code)
```



People in Trade, Manufacturing, Finance and Communications categories have a bigger proportion to earn +50K

Major Occupation Code

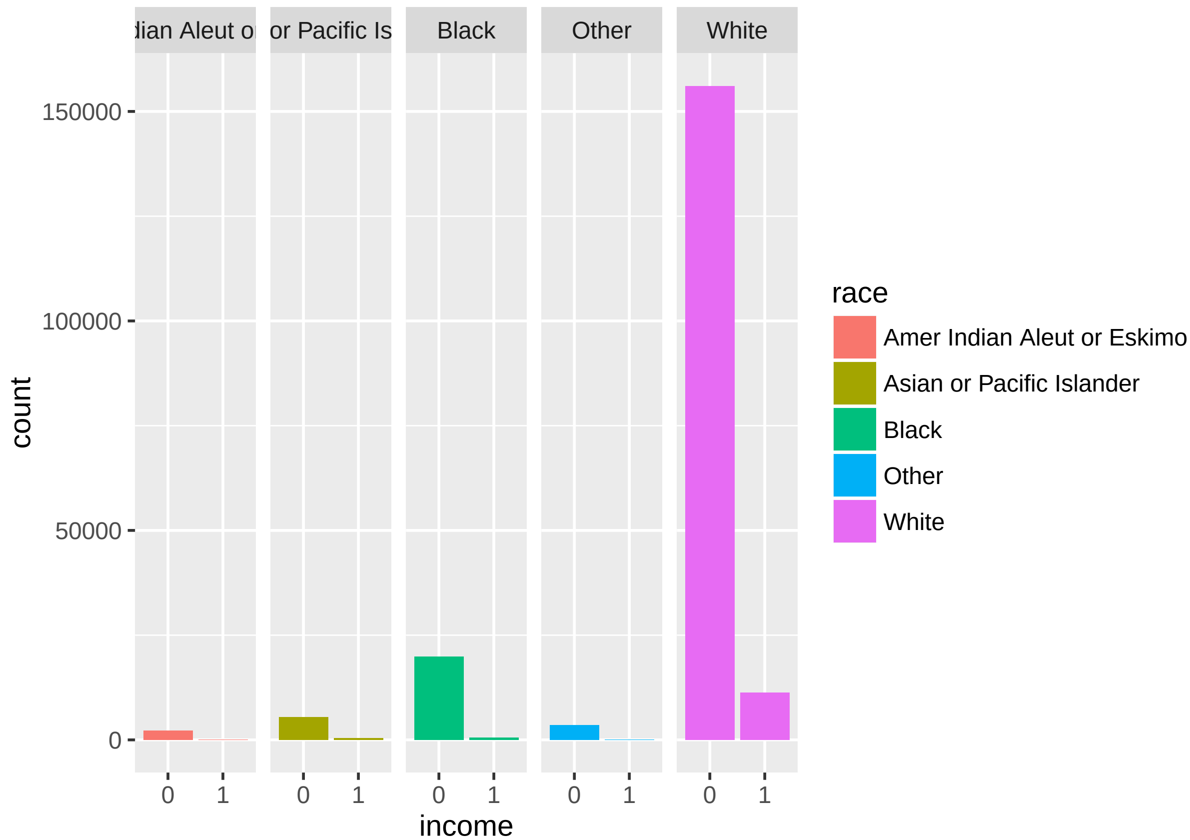
```
qplot (income, data = train_df, fill = maj_occ_code) + facet_grid (. ~ maj_occ_code)
```



Managerial, Professional Special, Protective Services have better proportion to earn +50K

Race

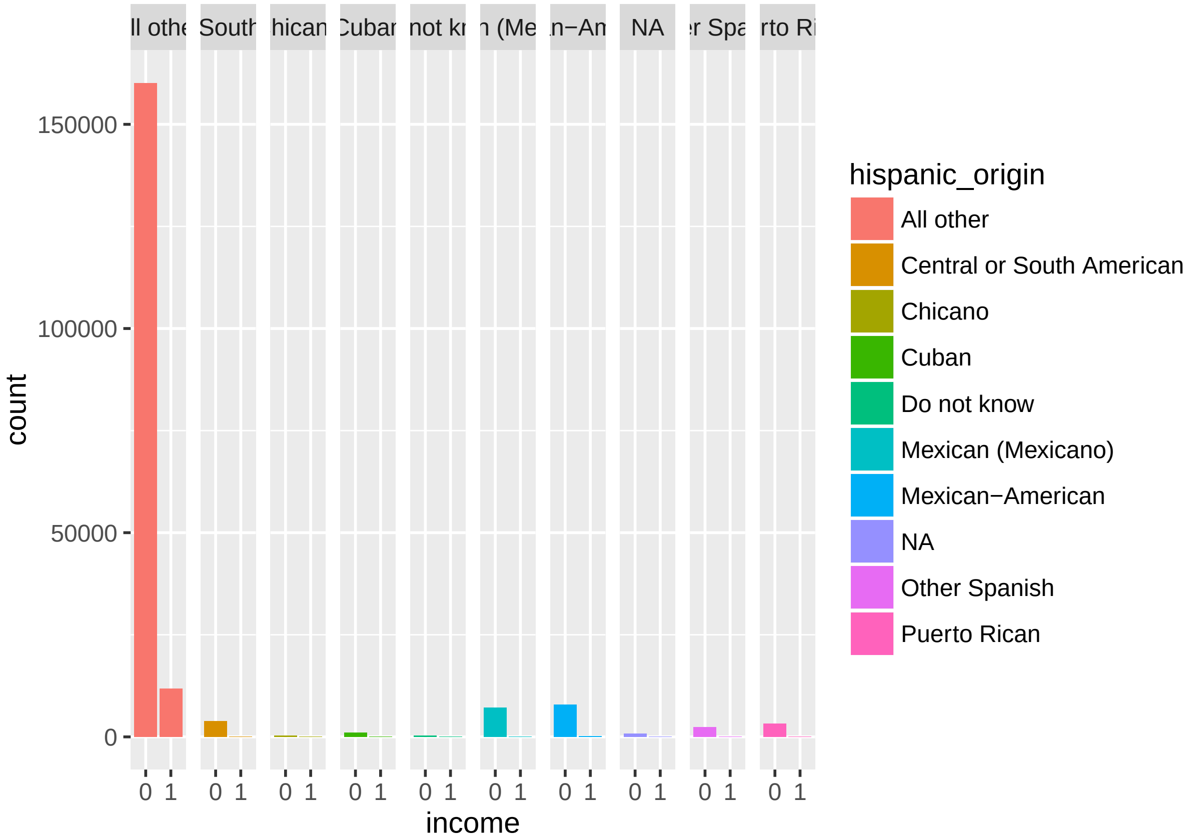
```
qplot (income, data = train_df, fill = race) + facet_grid (. ~ race)
```



White people seem advantaged , maybe transform the other races into a cat “minorities” would help the Random Forest

Hispanic_origin

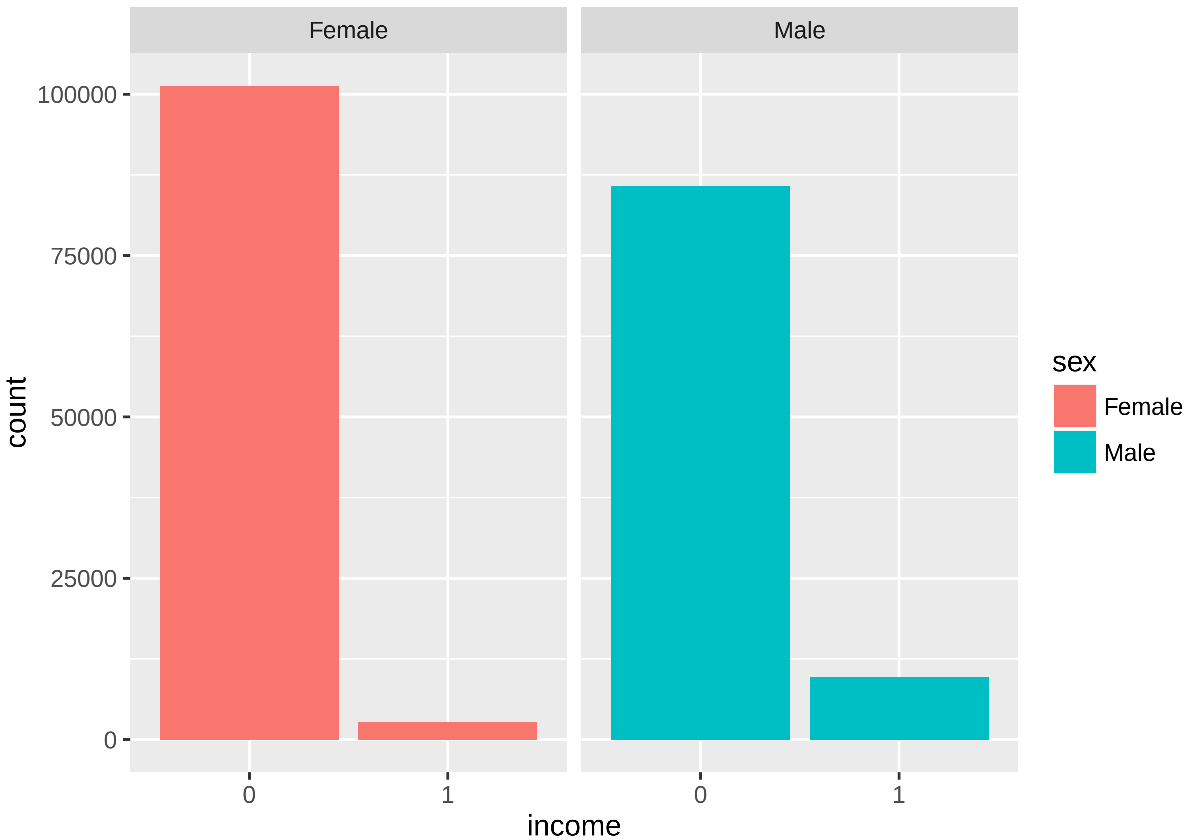
```
qplot (income, data = train_df, fill = hispanic_origin) + facet_grid (. ~ hispanic_origin)
```



Not Relevant "All other"" reprensent white people, and the others, minorities, this is already expressed by the previous feature

Sex

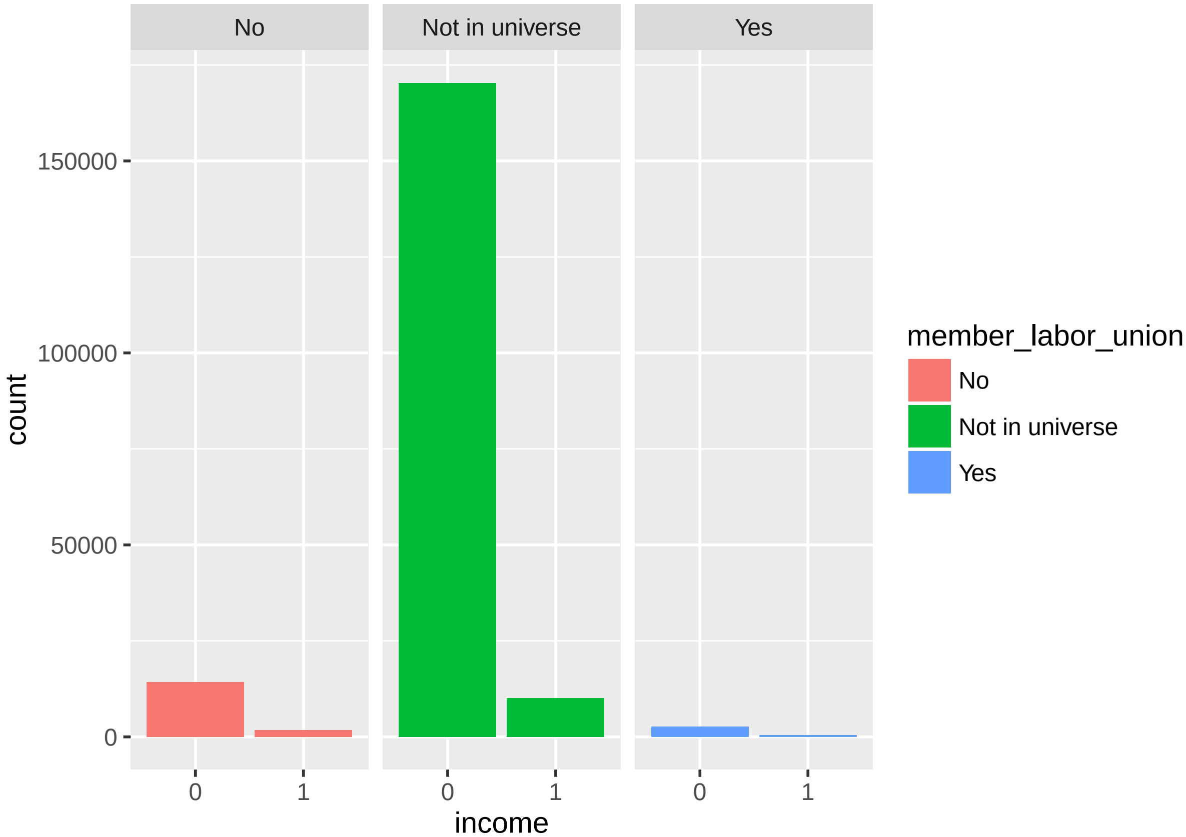
```
qplot (income, data = train_df, fill = sex) + facet_grid (. ~ sex)
```



We can clearly observe that males are advantanged.

Member labor union

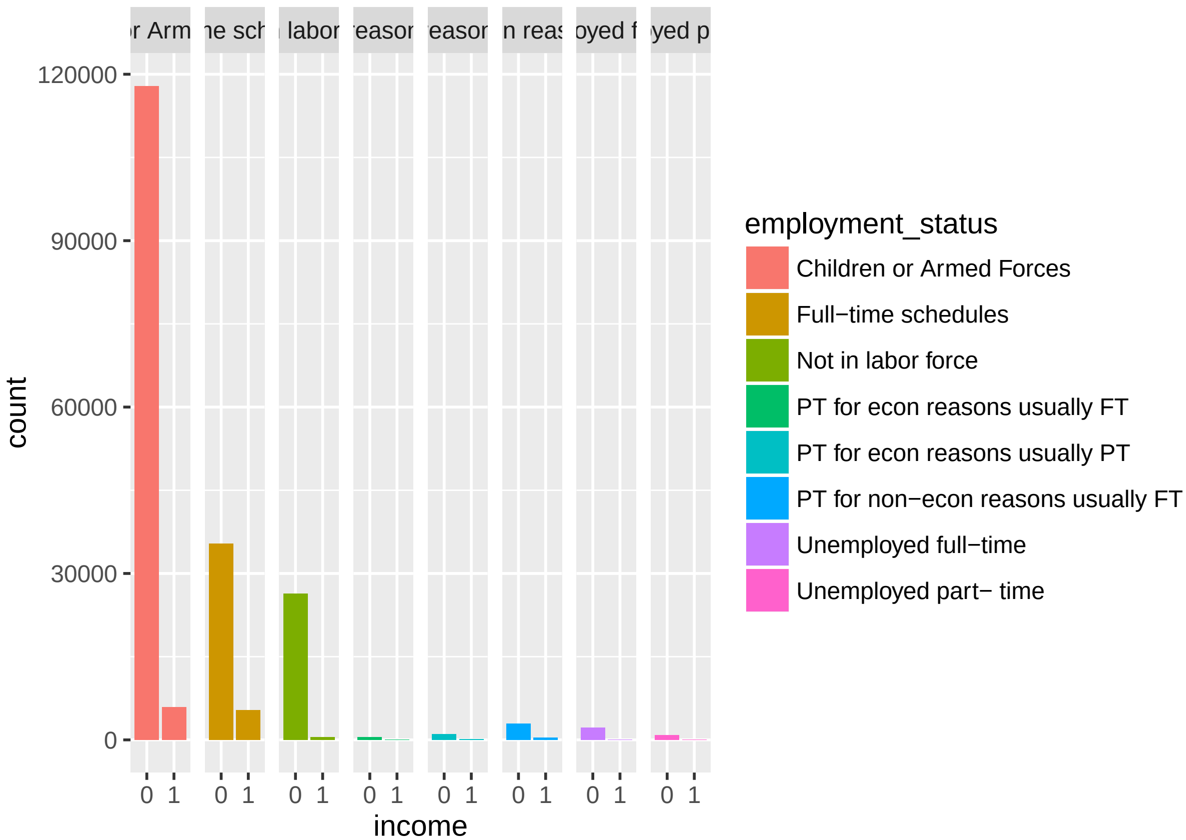
```
qplot (income, data = train_df, fill = member_labor_union) + facet_grid (. ~ member_labor_union)
```



Relevant too

Enployment status

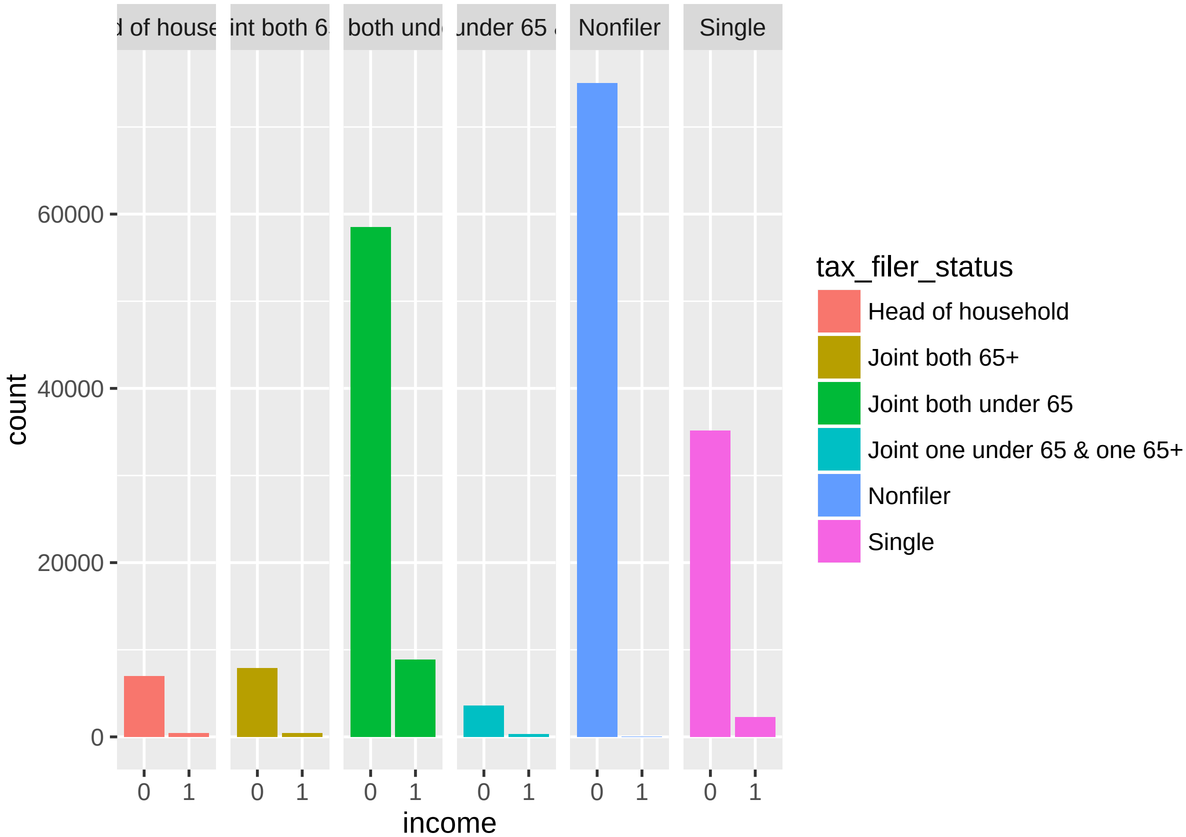
```
qplot (income, data = train_df, fill = employment_status) + facet_grid (. ~ employment_status)
```



Full time schedule more likely to earn +50 as it was also predictable

Tax Filer Status

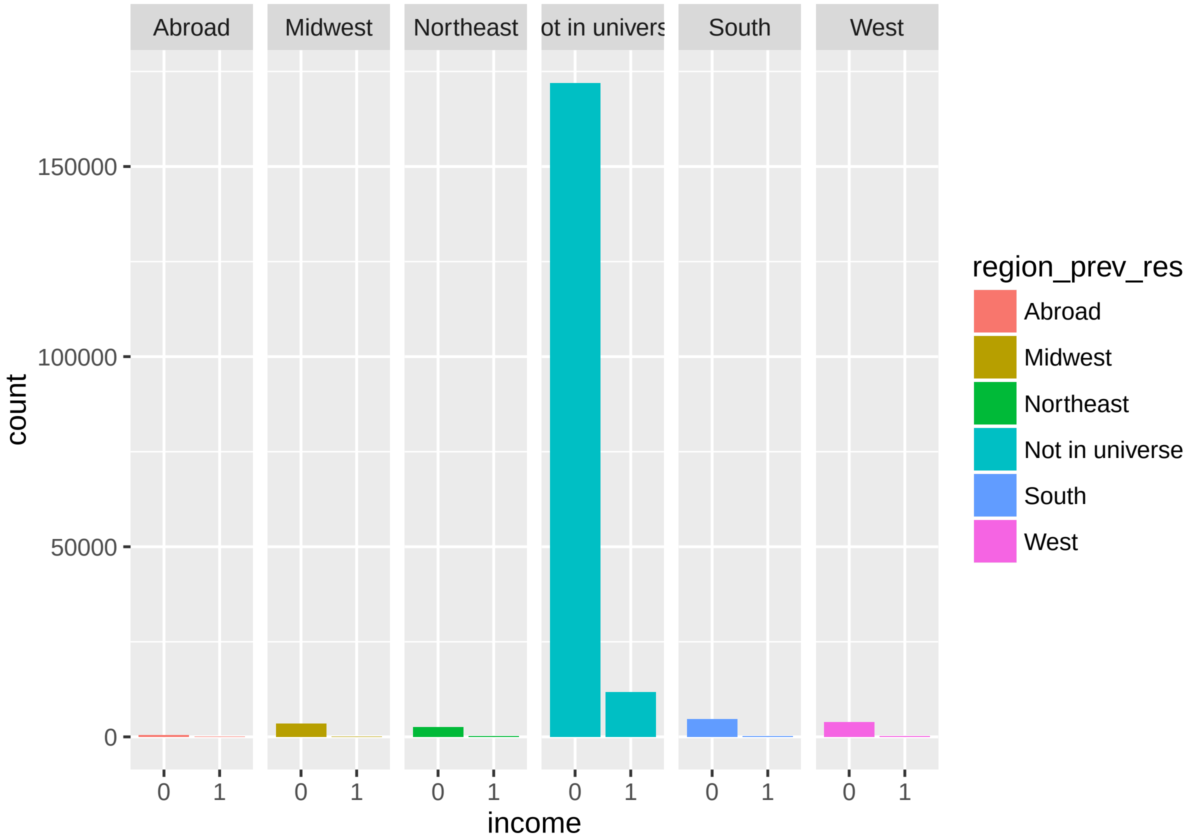
```
qplot (income, data = train_df, fill = tax_filer_status) + facet_grid (. ~ tax_filer_status)
```



Relevant, Joint both under 65 have higher "chance" to make +50K

Region Prev Res

```
qplot (income, data = train_df, fill = region_prev_res) + facet_grid (. ~ region_prev_res)
```

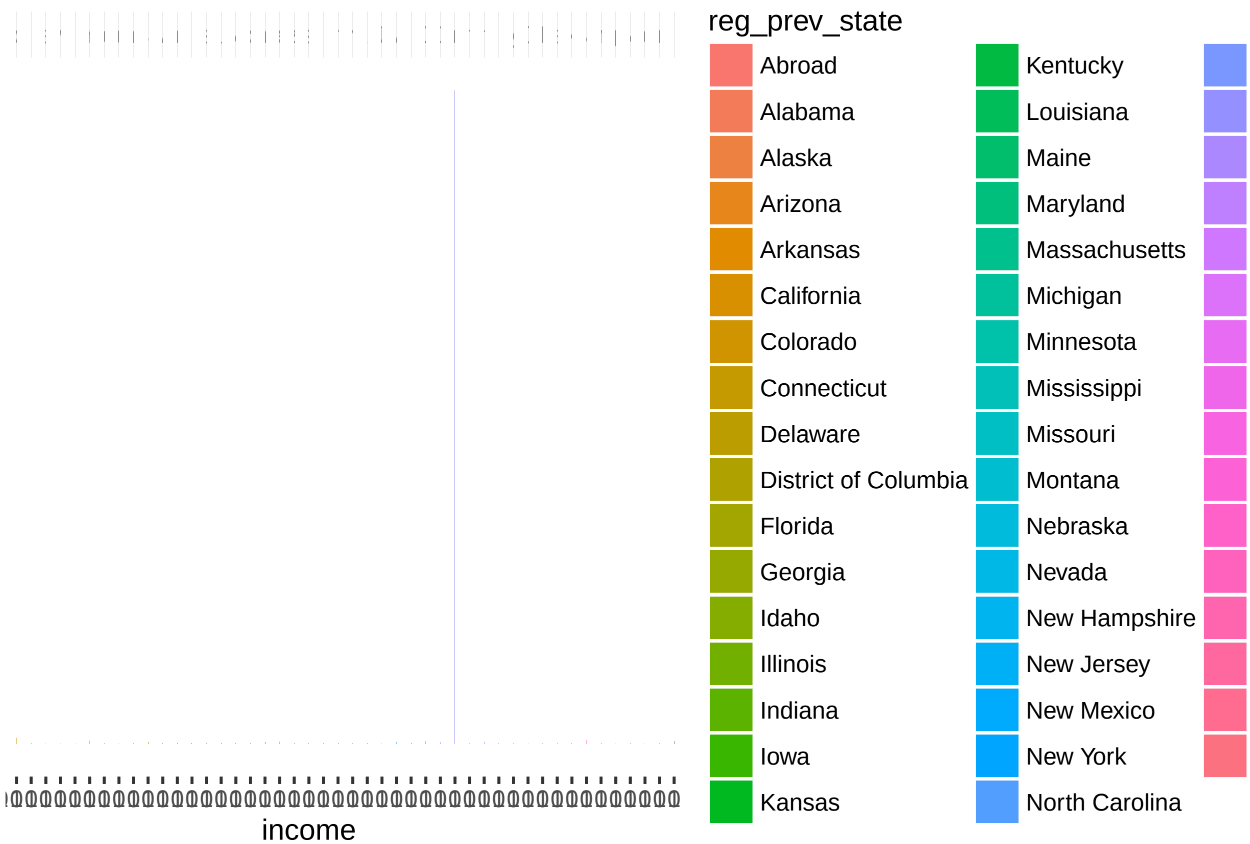


Not

Relevant

Reg Prev State

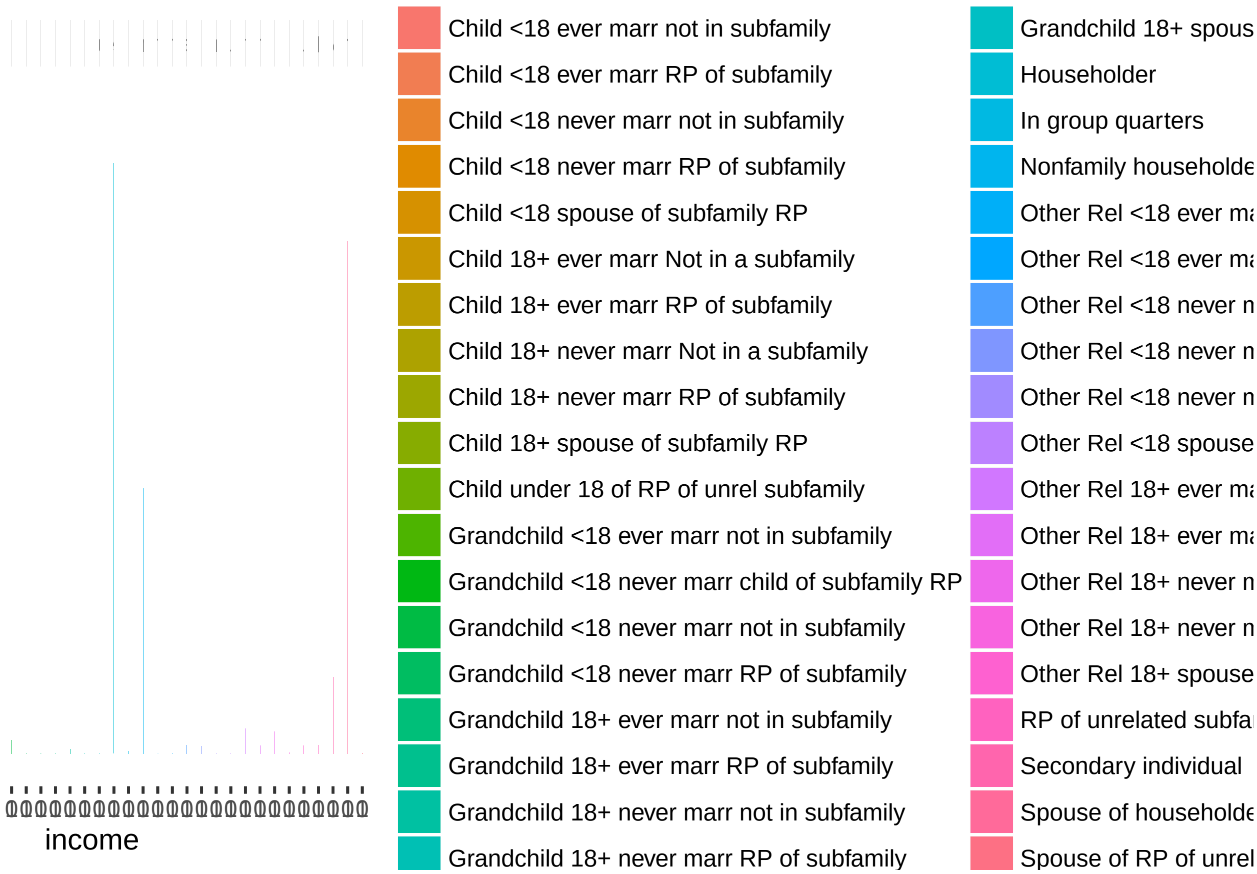
```
qplot (income, data = train_df, fill = reg_prev_state) + facet_grid (. ~ reg_prev_state)
```



Not Relevant

Household_stats

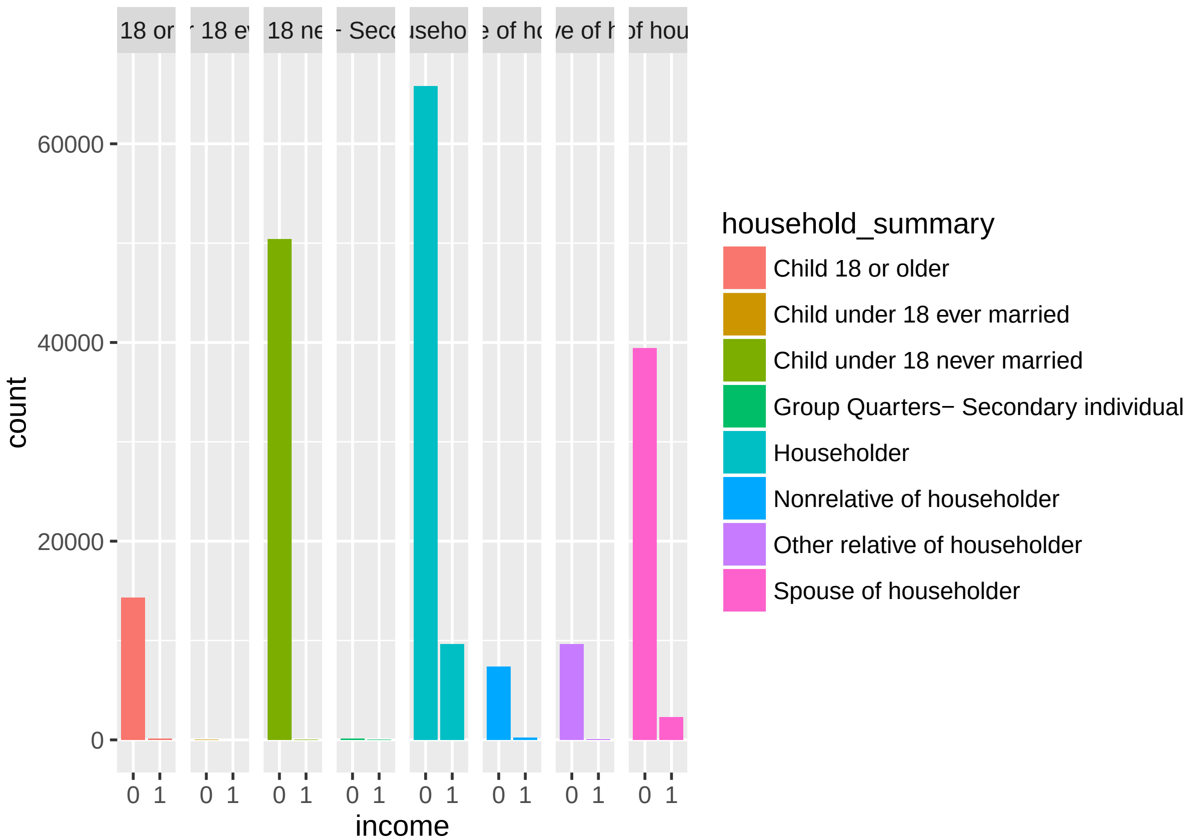
```
qplot (income, data = train_df, fill = household_stats) + facet_grid (. ~ household_stats)
```



Redundant with household_summary

household_summary

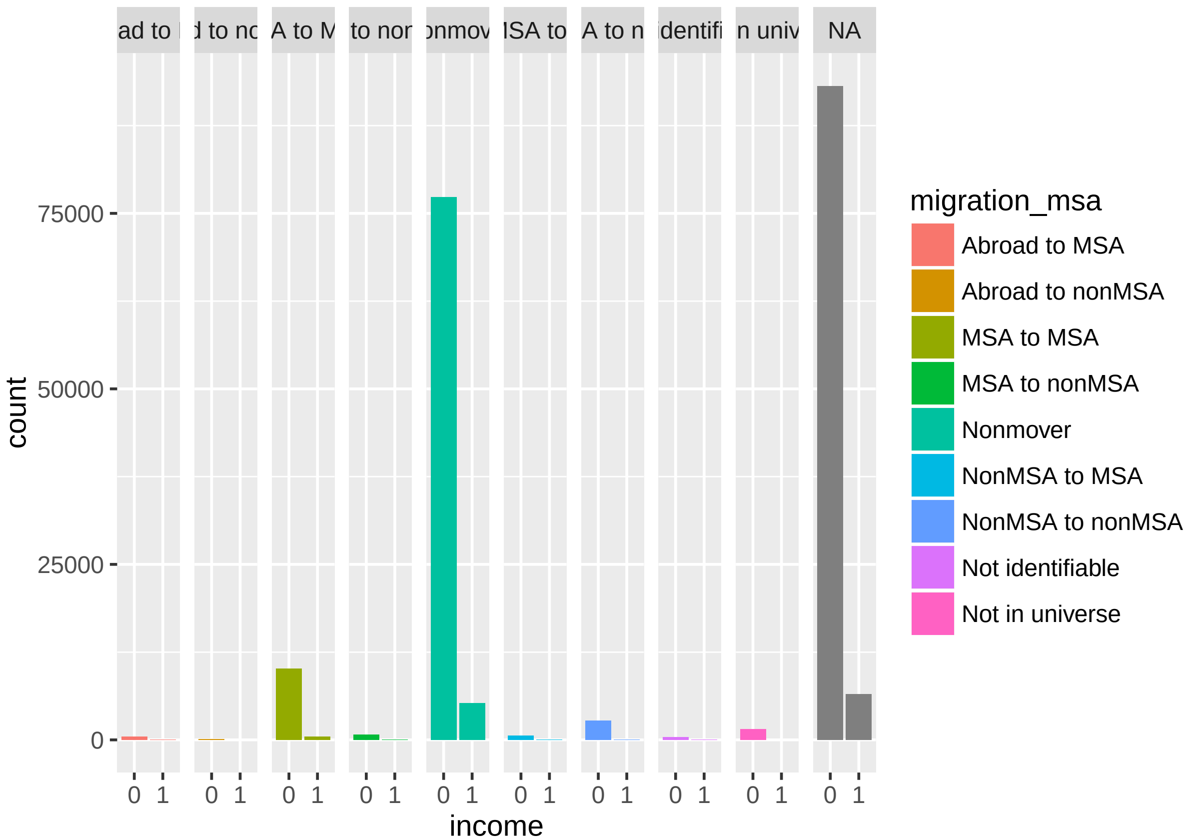
```
qplot (income, data = train_df, fill = household_summary) + facet_grid (. ~ household_summary)
```



Relevant

Migration msa

```
qplot (income, data = train_df, fill = migration_msa) + facet_grid (. ~ migration_msa)
```

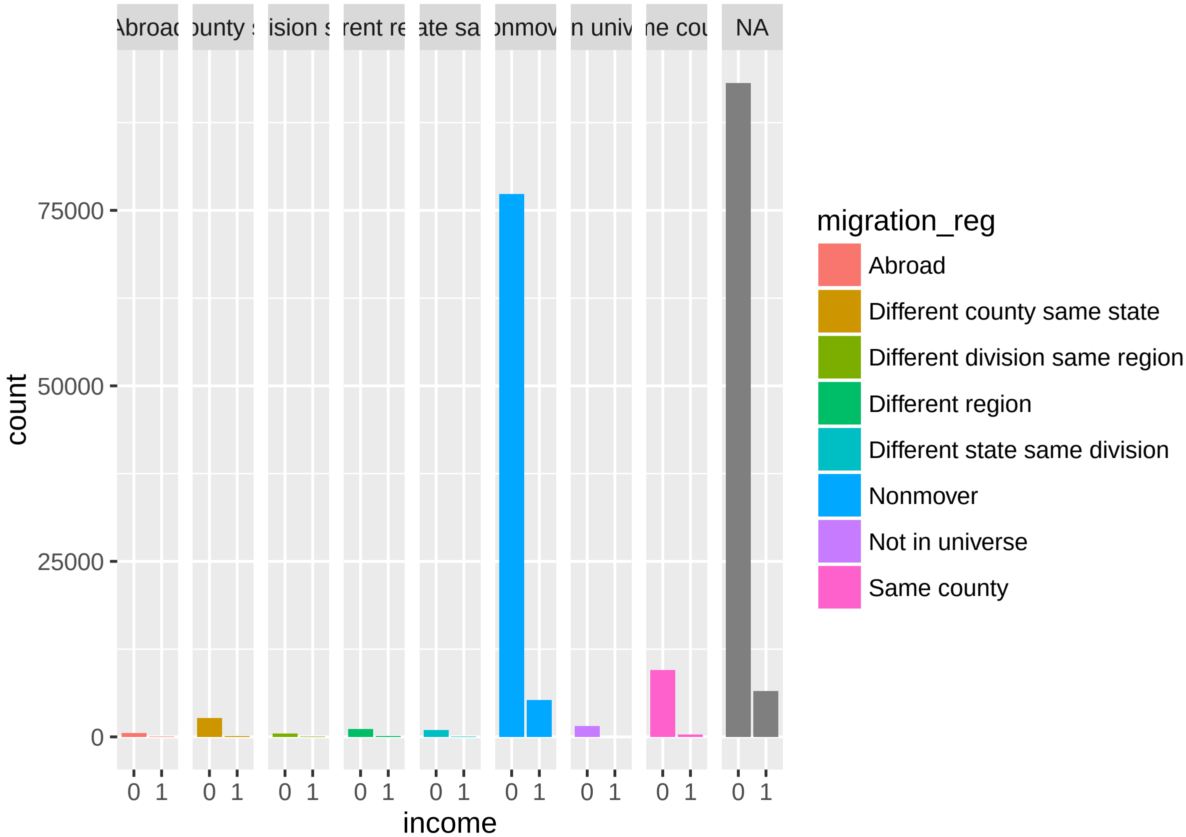


=>

could be relevant but too much missing information

Migration Reg

```
qplot (income, data = train_df, fill = migration_reg) + facet_grid (. ~ migration_reg)
```

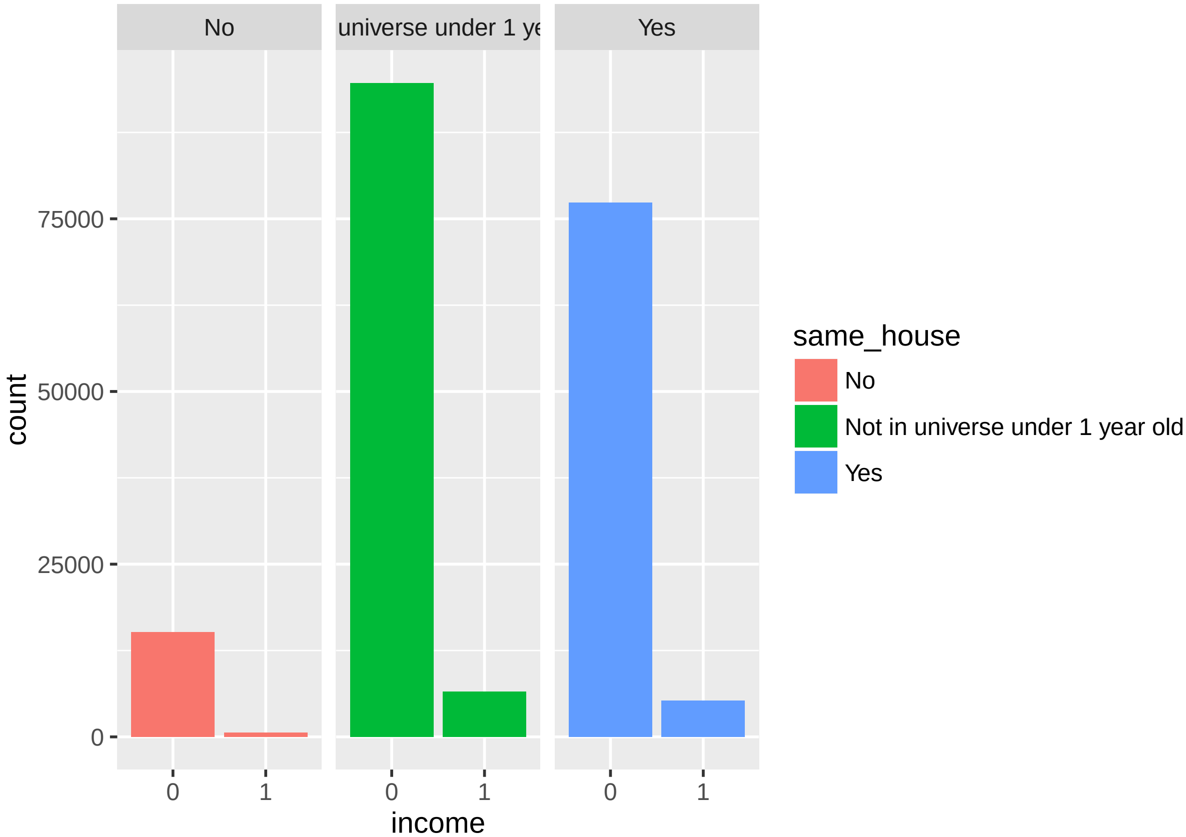


=>

could also be relevant but too much missing information

Same house

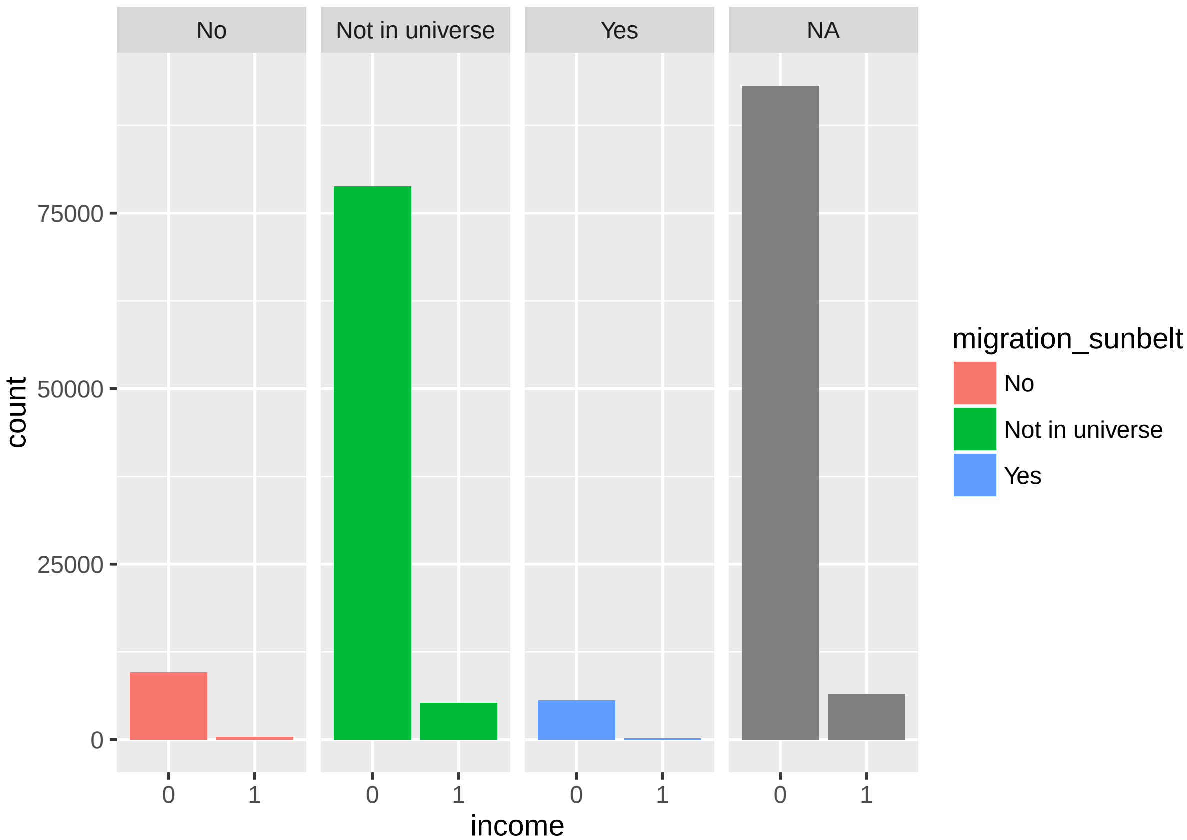
```
qplot (income, data = train_df, fill = same_house) + facet_grid (. ~ same_house)
```



Not relevant

Migration sunbelt

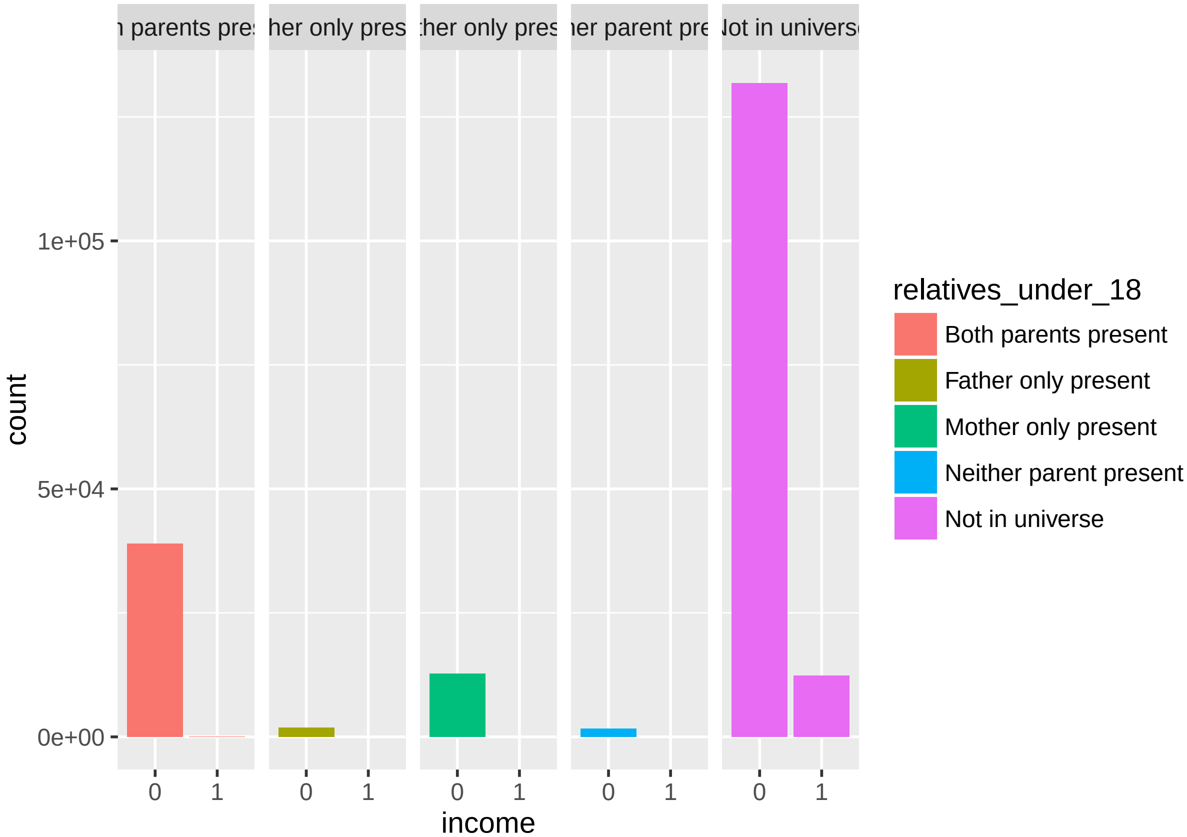
```
qplot (income, data = train_df, fill = migration_sunbelt) + facet_grid (. ~ migration_sunbelt)
```



Relevant

Relatives Under 18

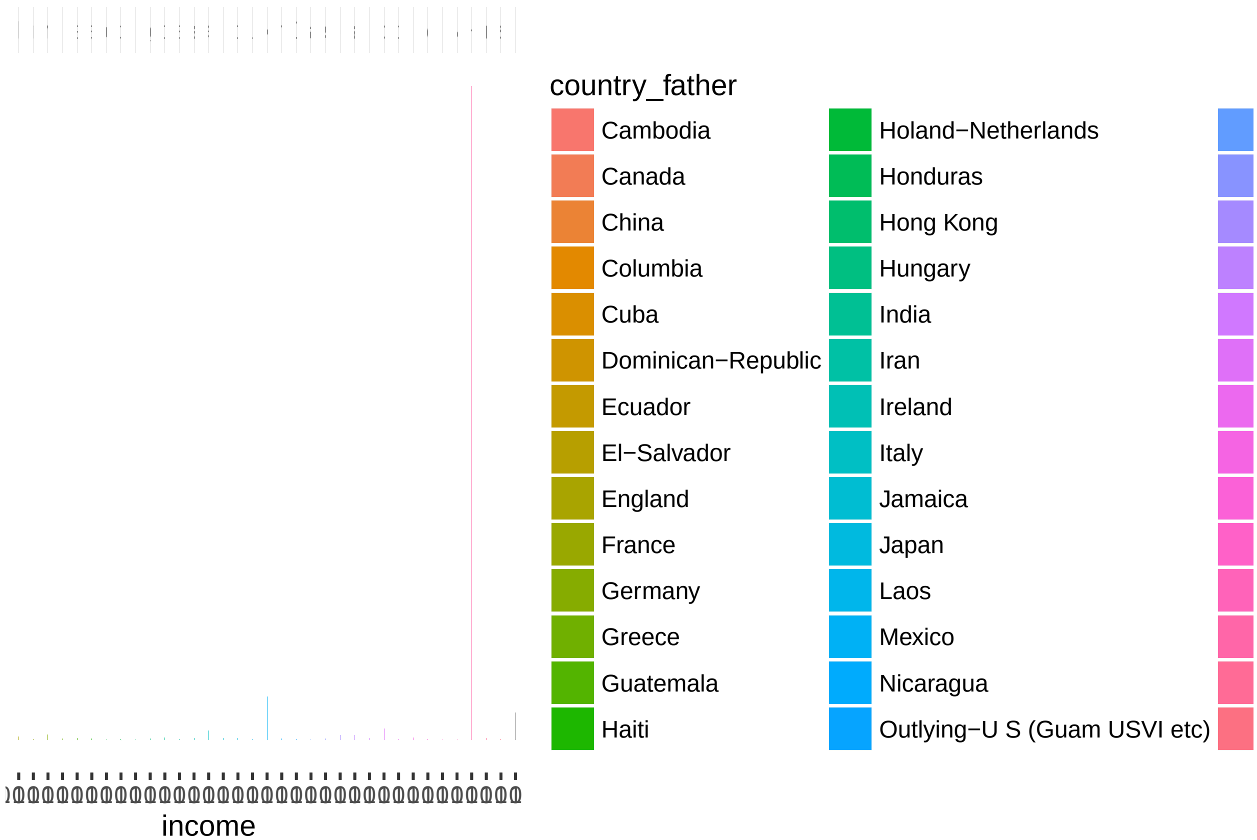
```
qplot (income, data = train_df, fill = relatives_under_18) + facet_grid (. ~ relatives_under_18)
```



Not relevant, seems to apply only to childs

Country Father

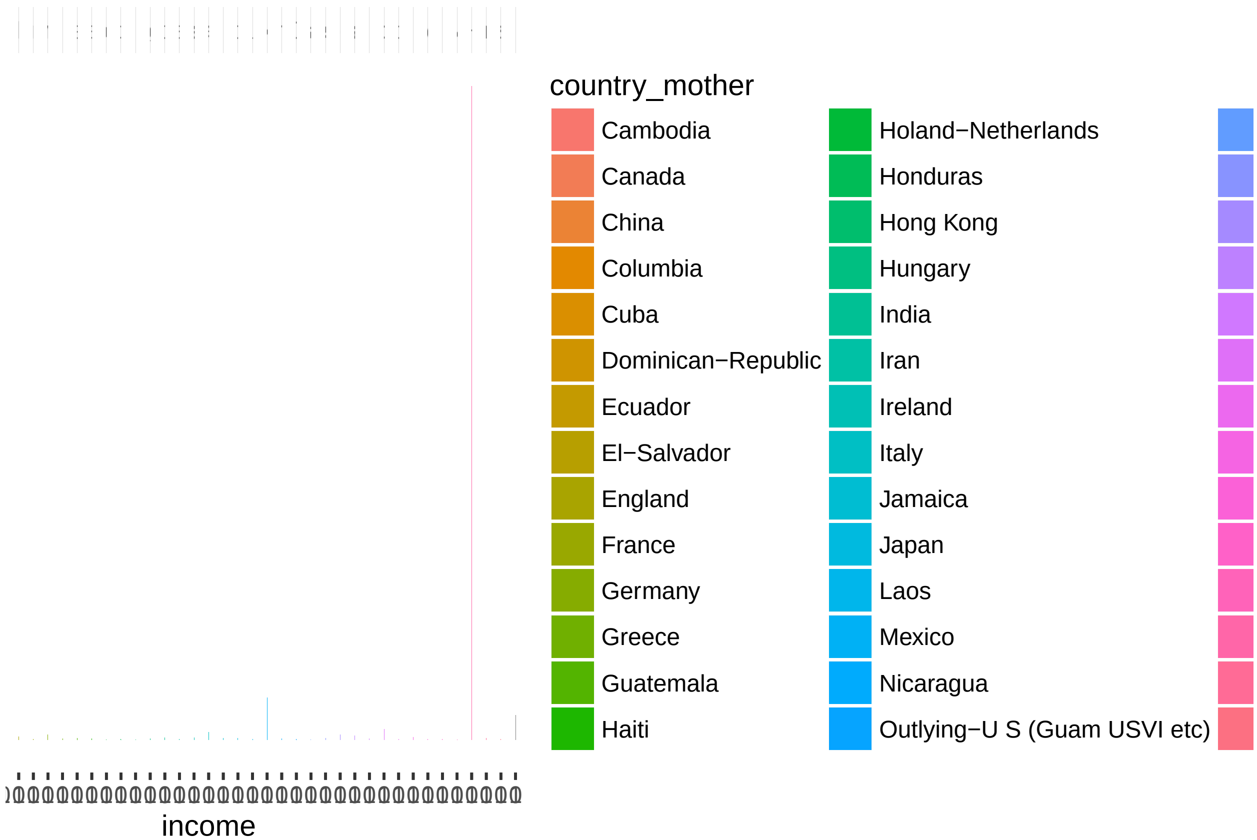
```
qplot (income, data = train_df, fill = country_father) + facet_grid (. ~ country_father)
```



Irrelevant

Country Mother

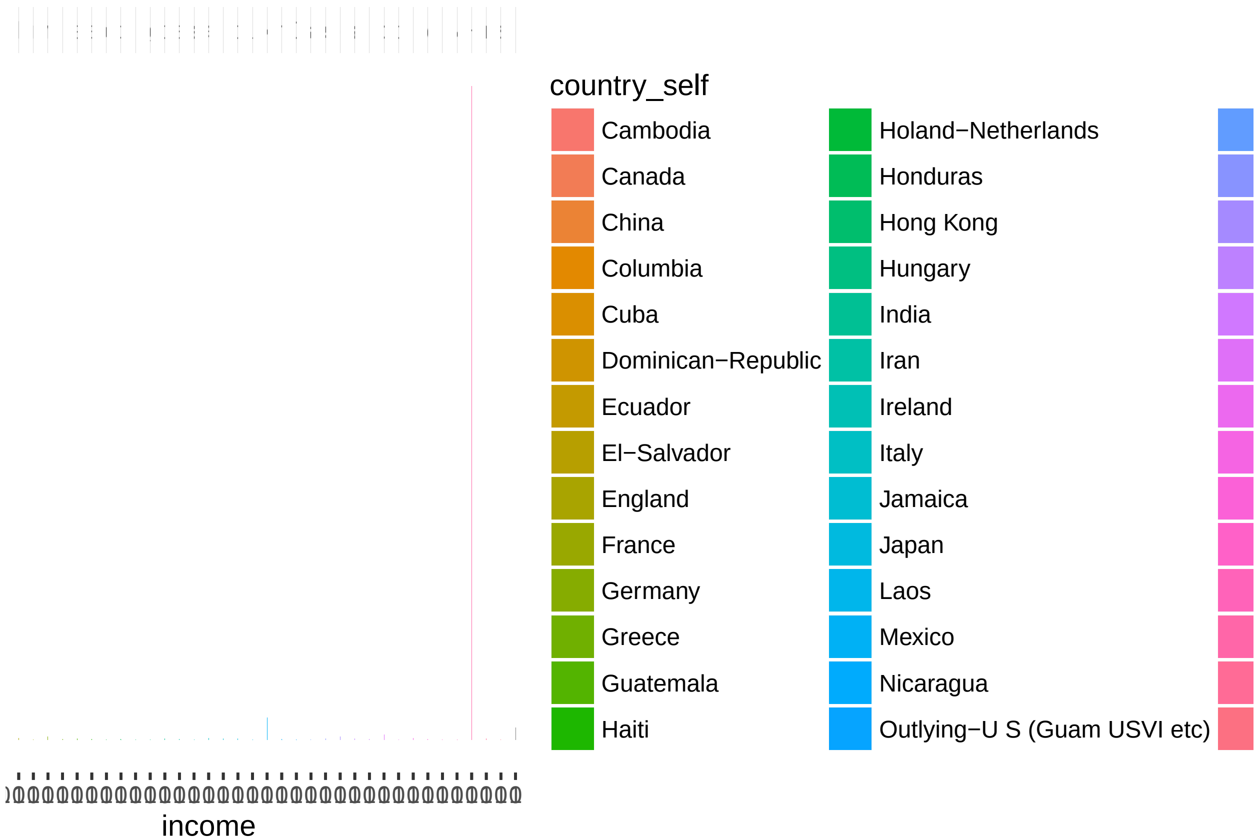
```
qplot (income, data = train_df, fill = country_mother) + facet_grid (. ~ country_mother)
```



Irrelevant

Country Self

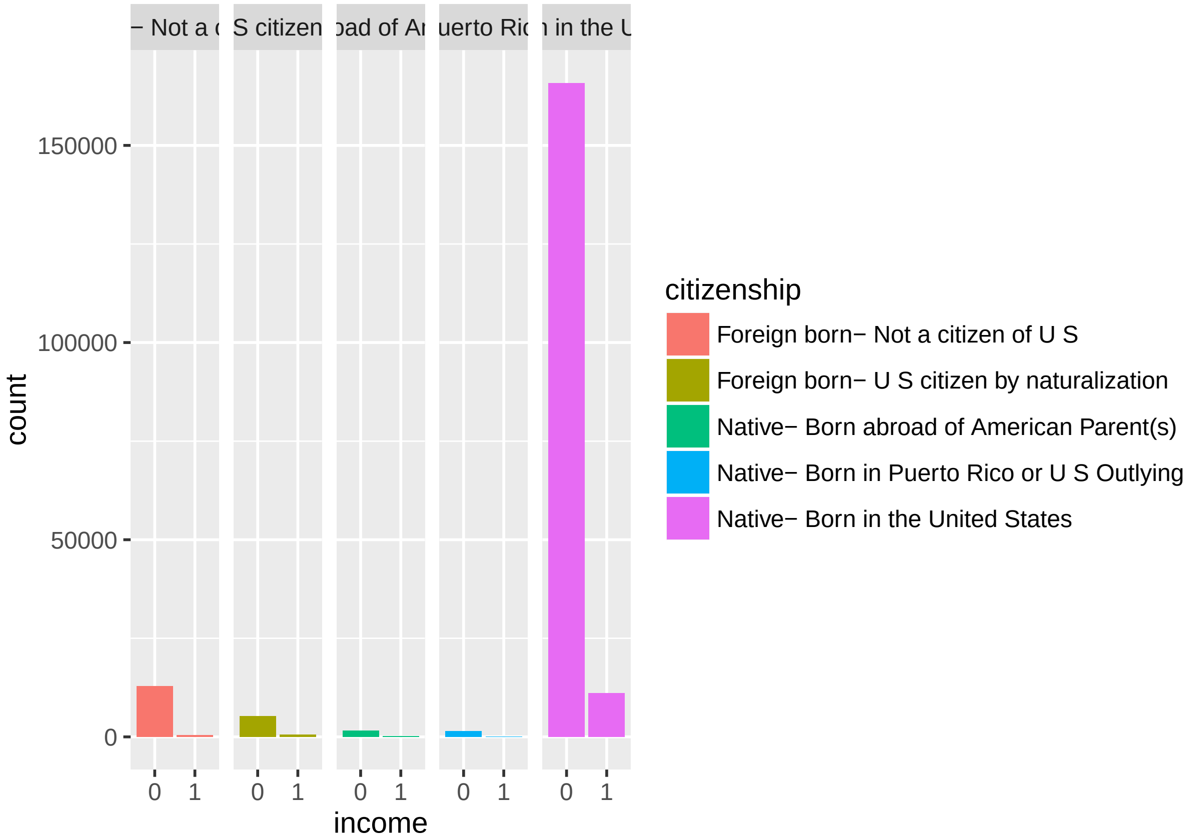
```
qplot (income, data = train_df, fill = country_self) + facet_grid (. ~ country_self)
```



May be relavant but redundant with citizenship

citizenship

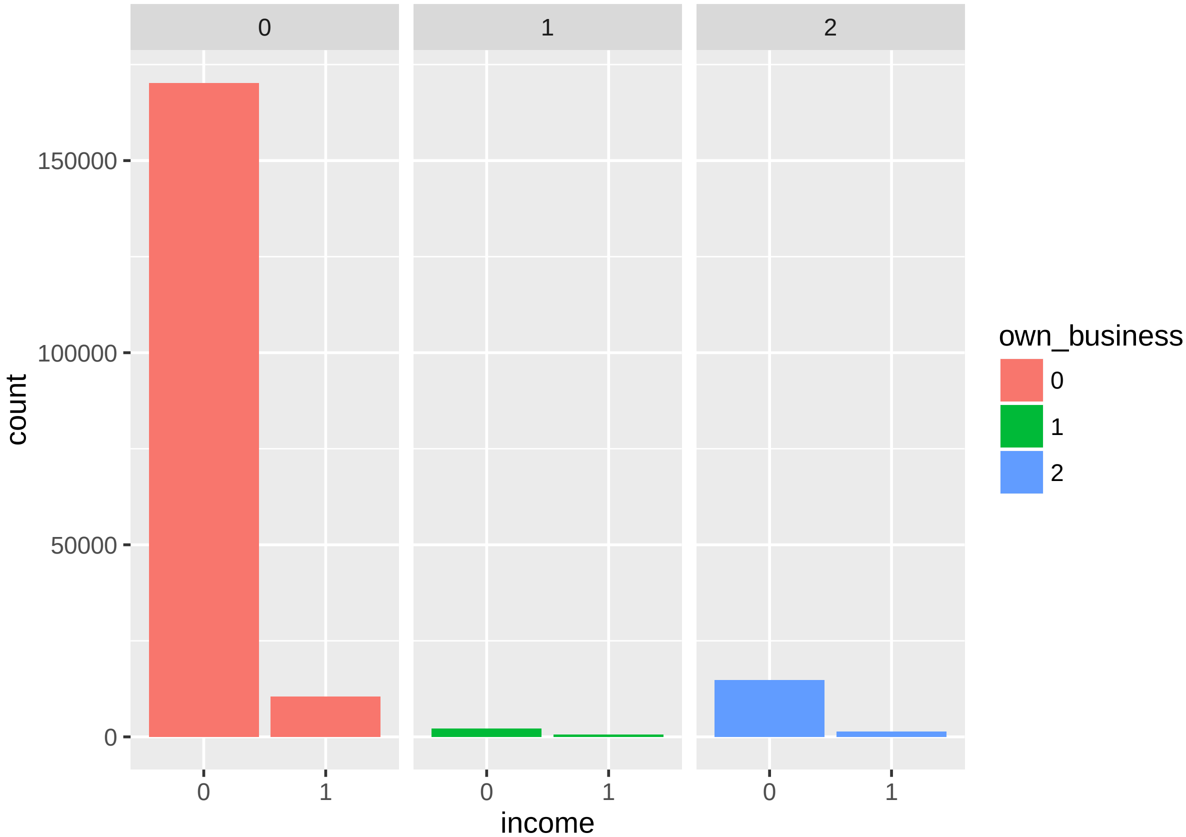
```
qplot (income, data = train_df, fill = citizenship) + facet_grid (. ~ citizenship)
```



May be relevant.

Own Business

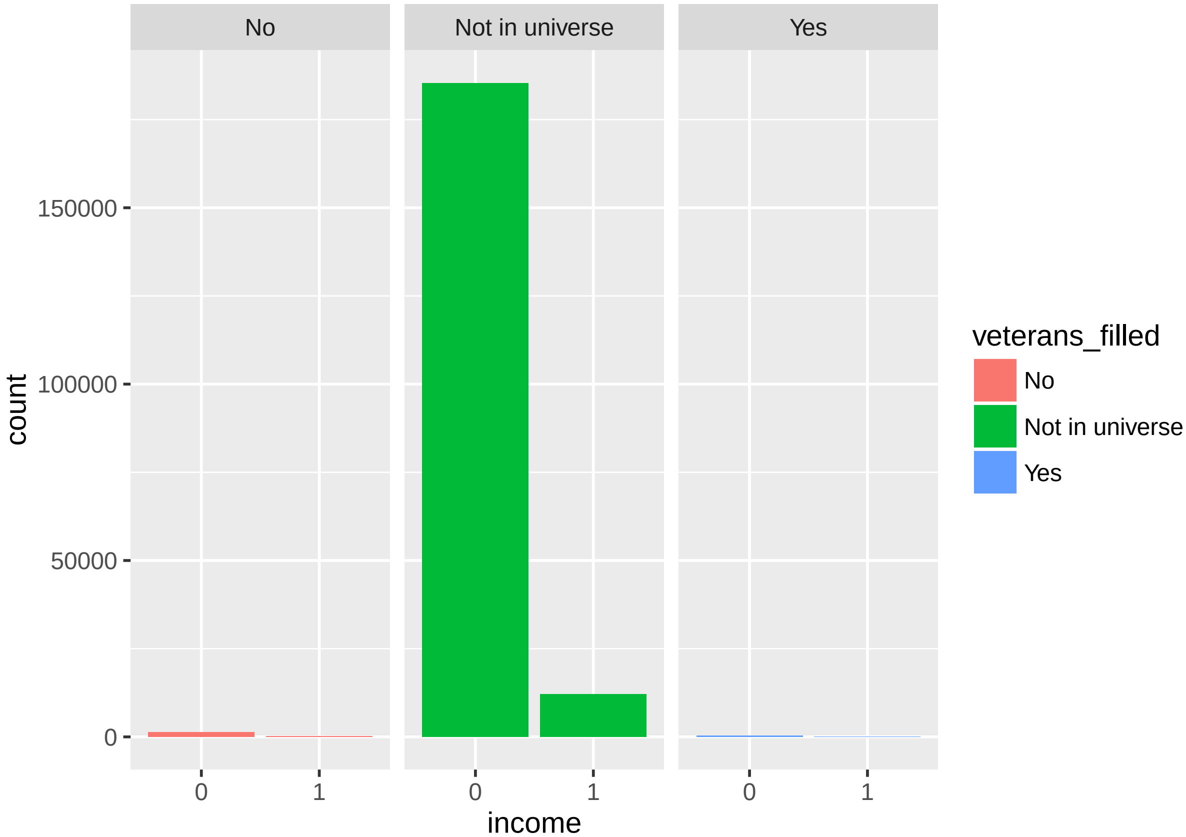
```
qplot (income, data = train_df, fill = own_business) + facet_grid (. ~ own_business)
```



Creator of business advantaged, meaning bosses are more likely to earn +50K

Veterans filled

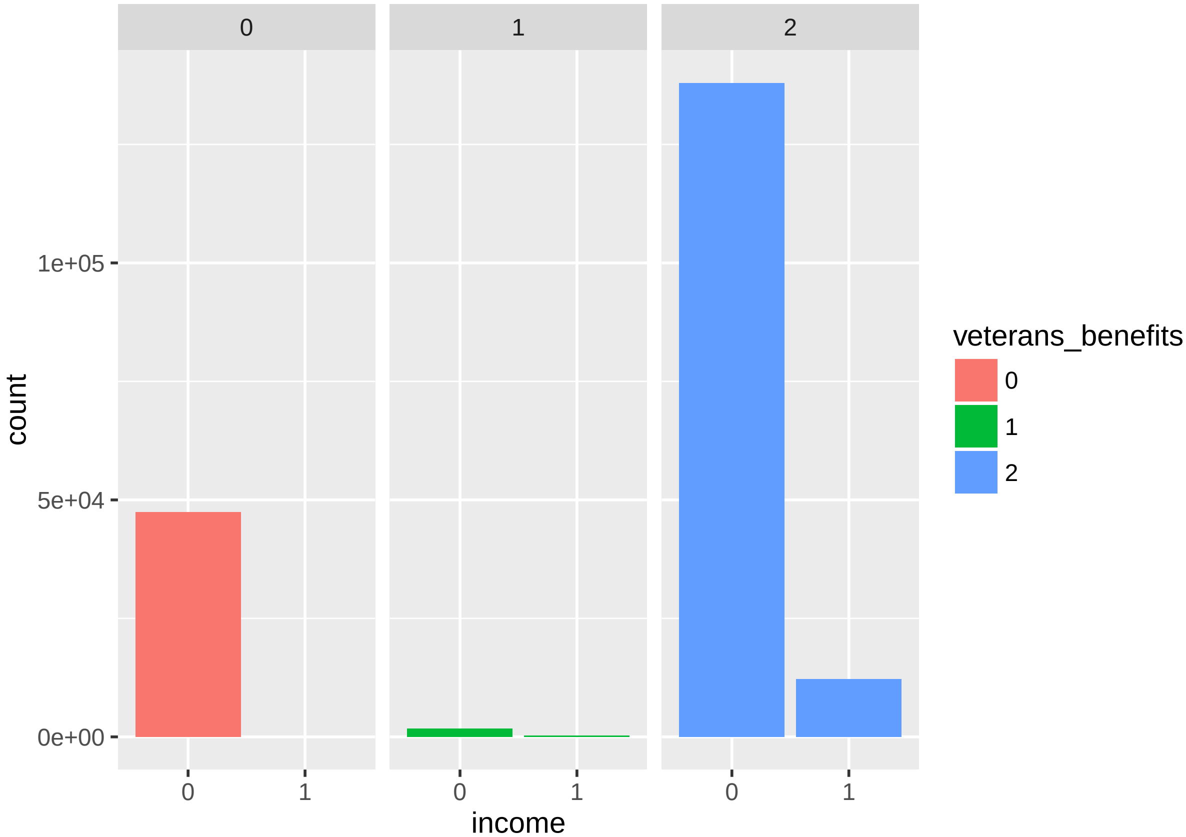
```
qplot (income, data = train_df, fill = veterans_filled) + facet_grid (. ~ veterans_filled)
```



Not relevant, all the data is in the not in univers class and there is not enough data for veterans

Veterans Benefits

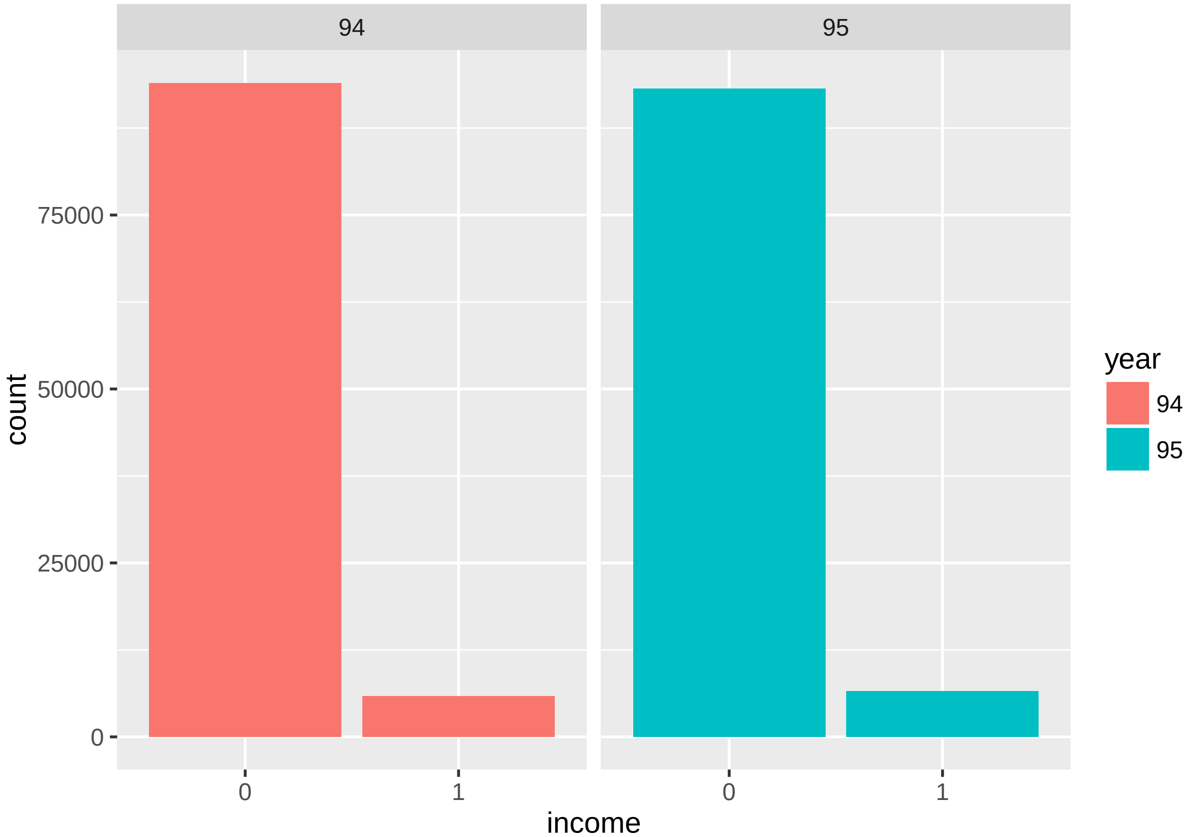
```
qplot (income, data = train_df, fill = veterans_benefits) + facet_grid (. ~ veterans_benefits )
```



Maybe relevant

Year

```
qplot (income, data = train_df, fill = year) + facet_grid (. ~ year )
```

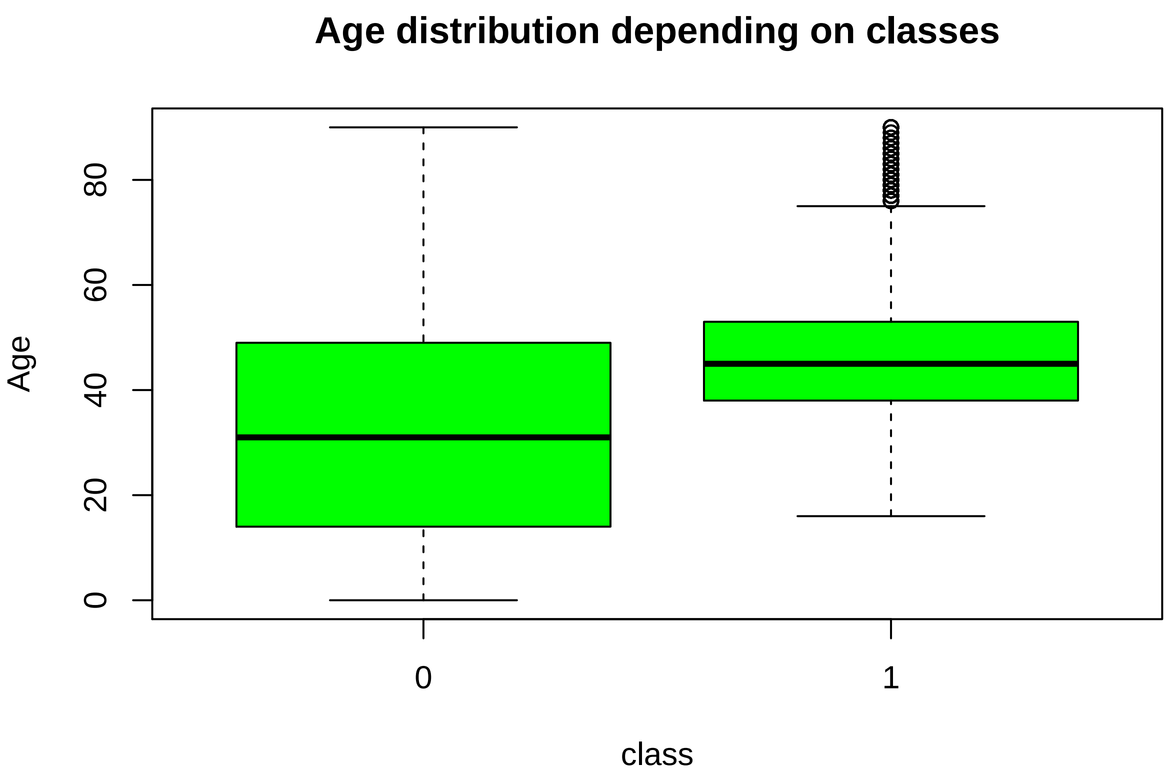


As predicted, useless

Visualising numerical variables

BoxPlot of the age

```
boxplot (age ~ income, data = train_df, main = "Age distribution depending on classes",
         xlab = "class", ylab = "Age", col = c("green") )
```

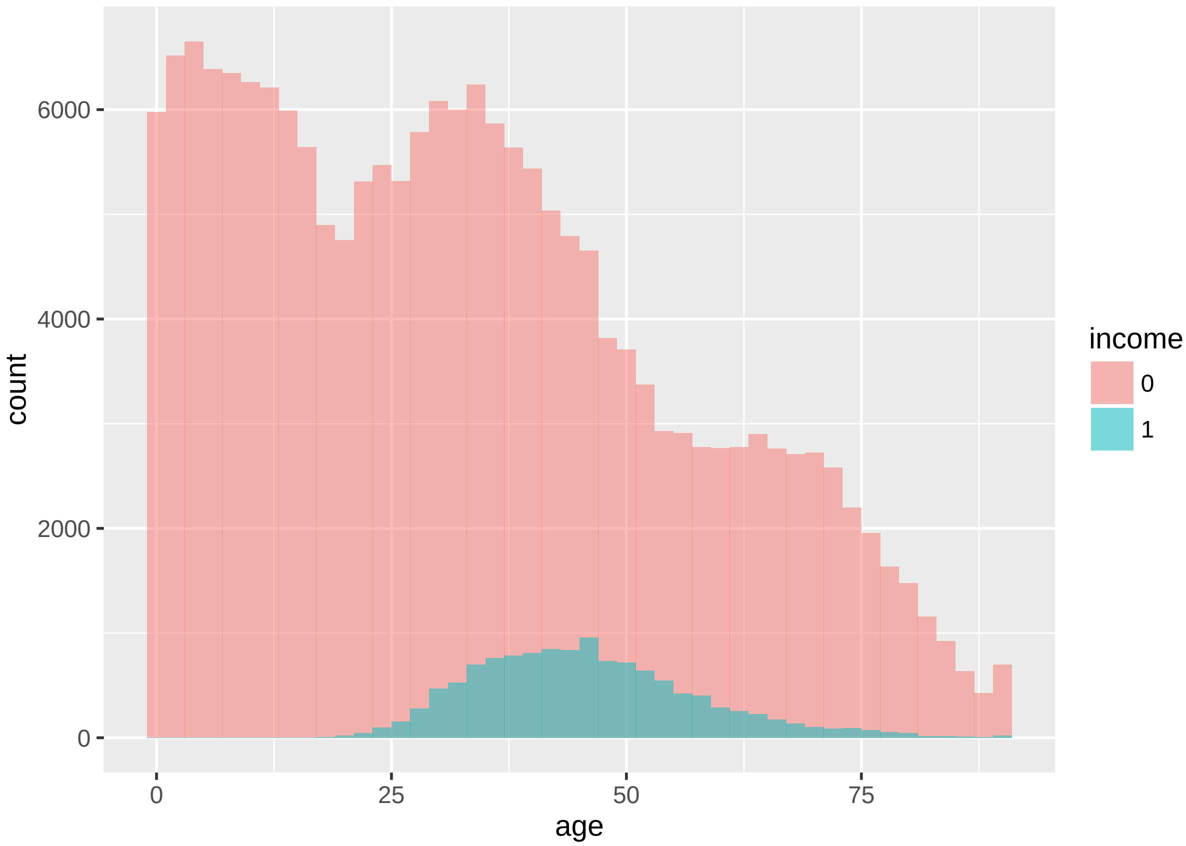


=>

Relevant

Distribution of the variable age

```
ggplot(train_df, aes(x=age, fill=income)) +
  geom_histogram(binwidth=2, alpha=0.5, position="identity")
```

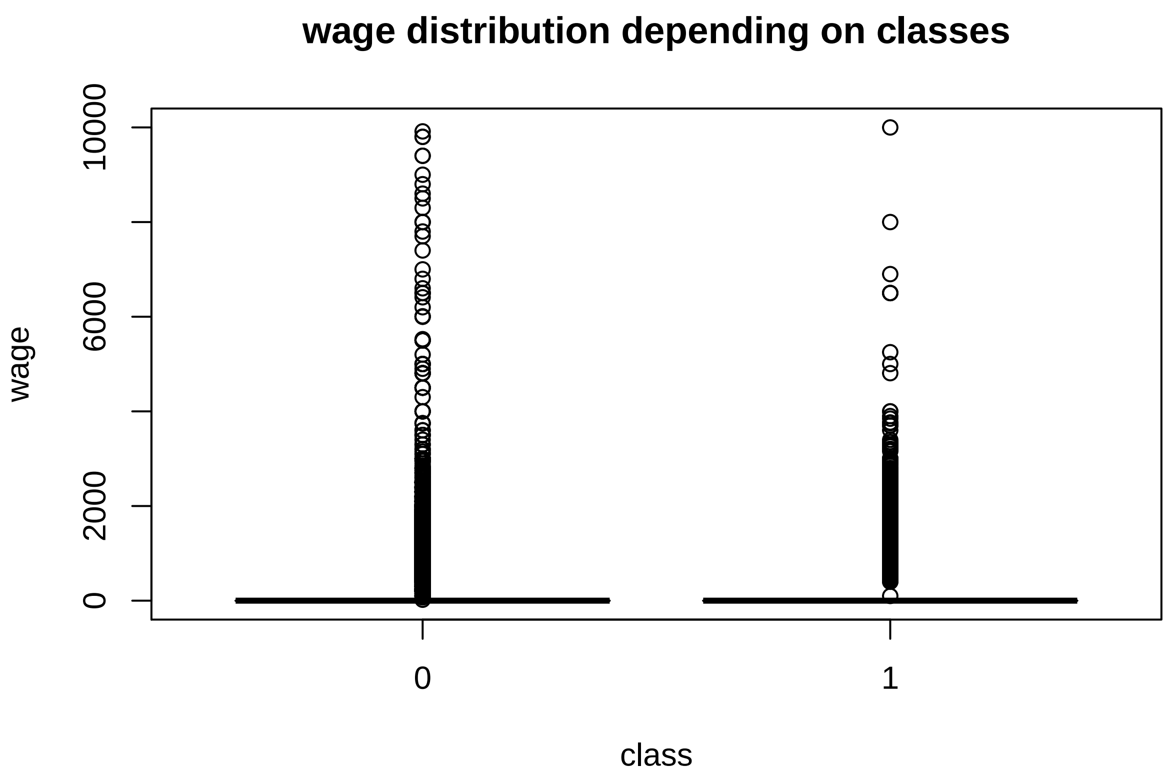


=>

Different distribution

Wage boxplot

```
boxplot (wage_per_hour ~ income, data = train_df, main = "wage distribution depending on classes",
         xlab = "class", ylab = "wage", col = c("green") )
```

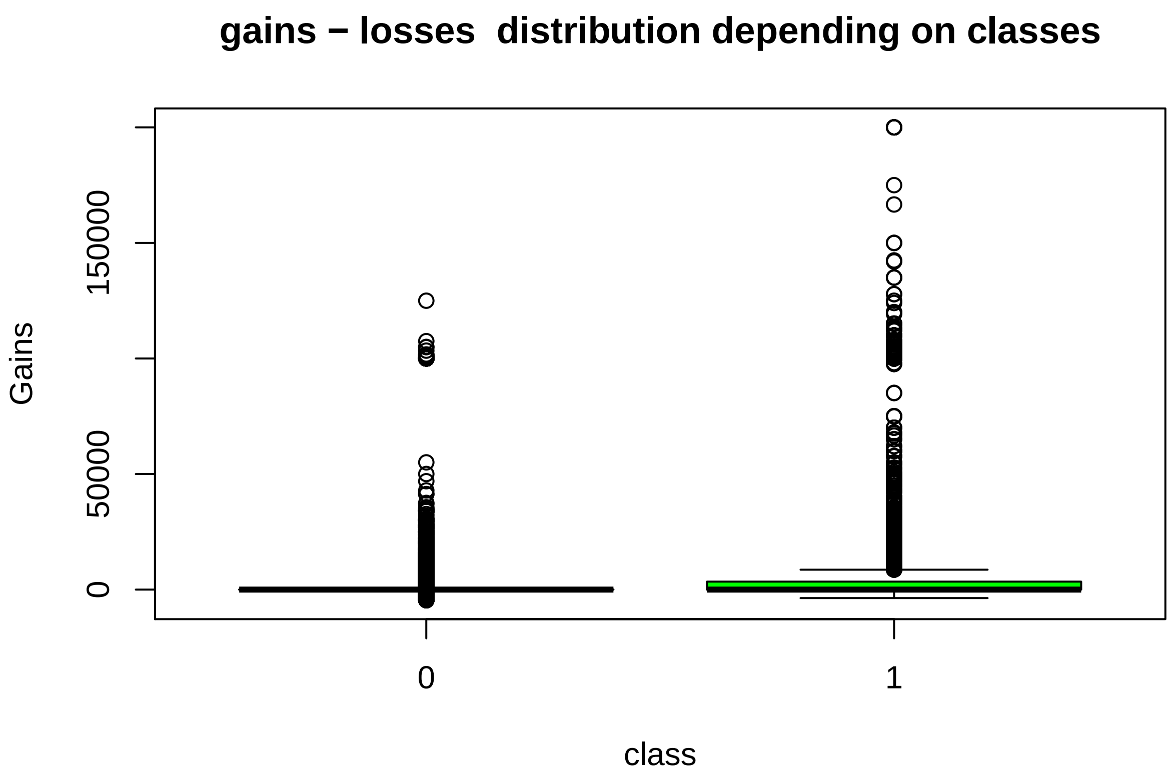


Too much zeros in the dataset, doesn't seem to be useful, a simple overview of the first lines shows a individual having 1200 of hourly wage and still under 50

Capital Gains Losses & Dividends.

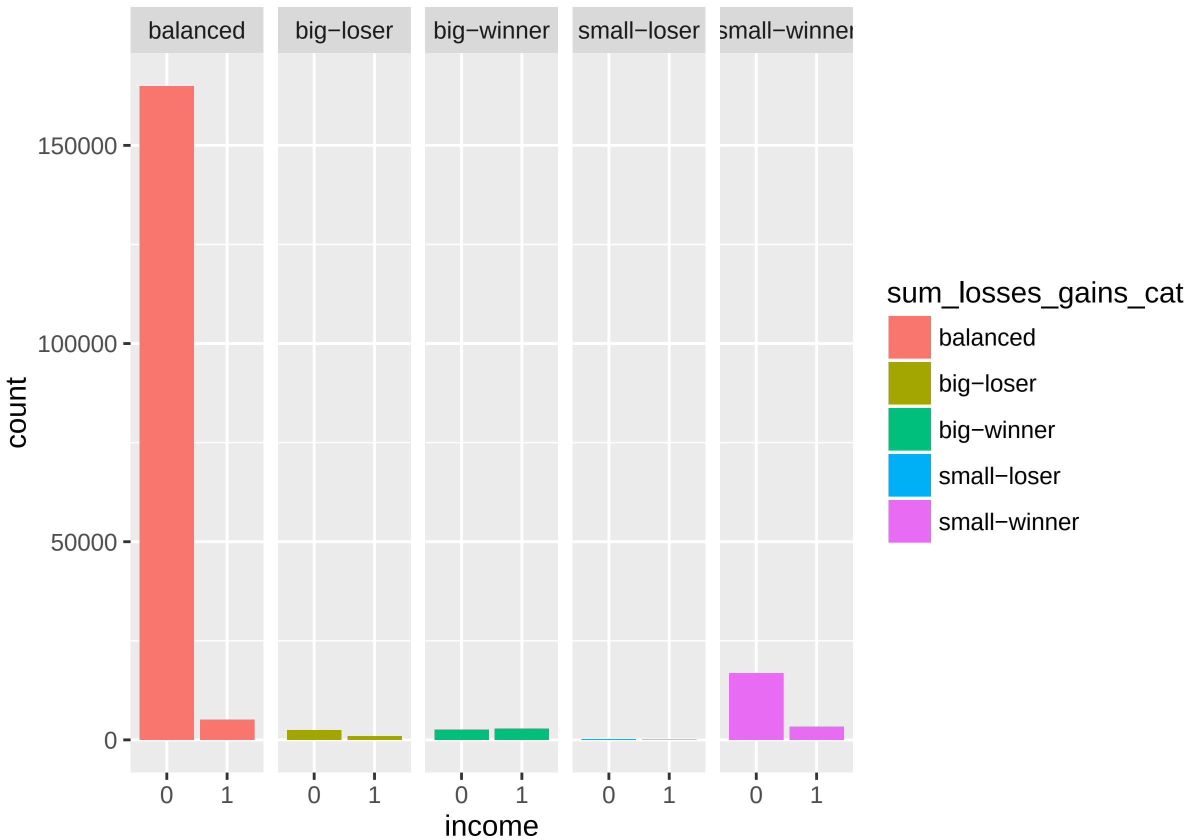
There are a lot of zeros in that columns. Indeed not everybody has stocks in companies or any capital invested in something. Nevertheless we can think that wealthy people are in position to invest money in capitals so the non-zeros values should help to classify the +50K Class. I decided to sum up the gains and dividends minus the losses to obtain a new variable.

```
train_df$sum_losses_gains = train_df$capital_gains - train_df$capital_losses + train_df$dividends

boxplot (sum_losses_gains ~ income, data = train_df, main = "gains - losses  distribution depending on c
         xlab = "class", ylab = "Gains", col = c("green") )
```



As predicted we see that people who gain money from capital are usually likely to make +50K. To show this to a classifier I decided to cut these people into 5 categories and make a new feature

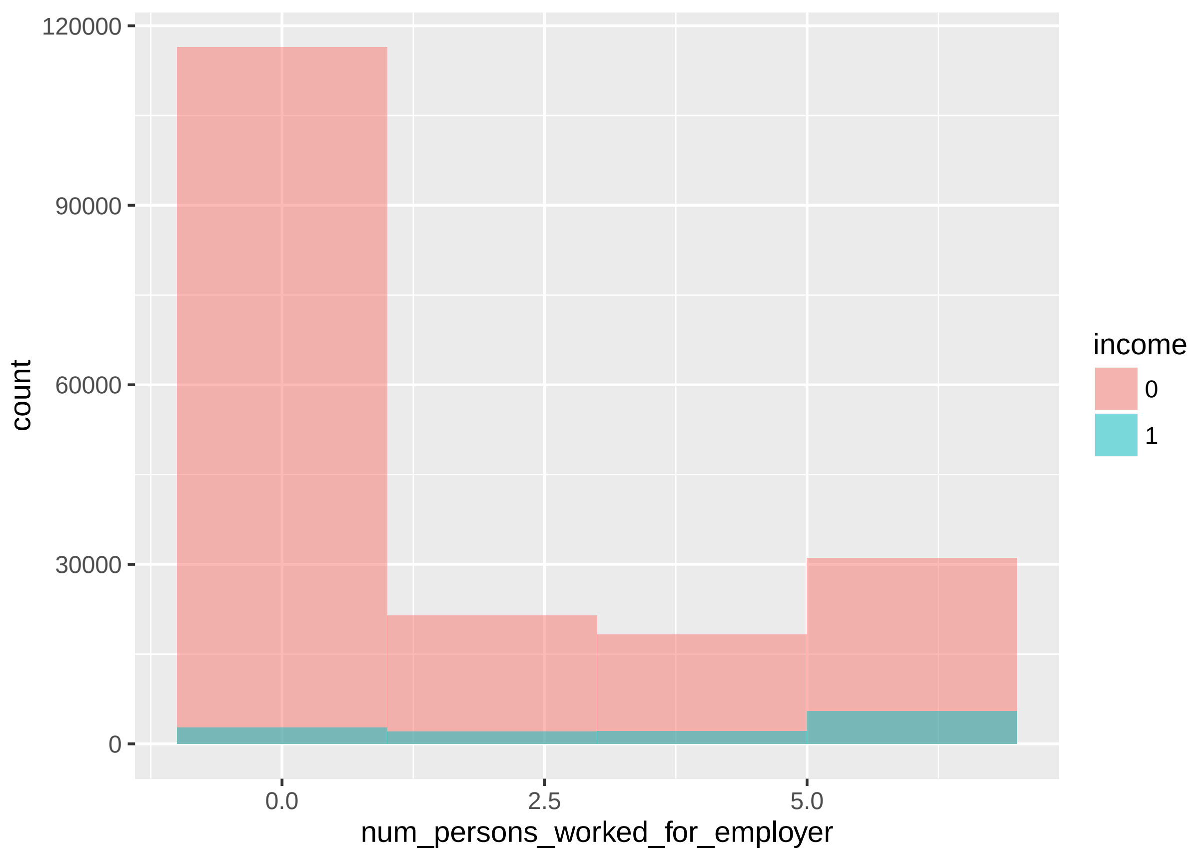
```
train_df$sum_losses_gains_cat<-ifelse(train_df$sum_losses_gains< -1000,"big-loser",
             ifelse(train_df$sum_losses_gains >= -1000 & train_df$sum_losses_gains < 0, "small-loser",
             ifelse(train_df$sum_losses_gains == 0 , "balanced",
             ifelse(train_df$sum_losses_gains> 0 & train_df$sum_losses_gains <= 5000 , "small-winner",
             ifelse(train_df$sum_losses_gains > 5000 , "big-winner","other"
             )))))

# Plot the result
qplot (income, data = train_df, fill = sum_losses_gains_cat) + facet_grid (. ~ sum_losses_gains_cat)
```



Num persons worked for employer

```
boxplot (num_persons_worked_for_employer ~ income, data = train_df, main = "Num persons worked for emplo
```

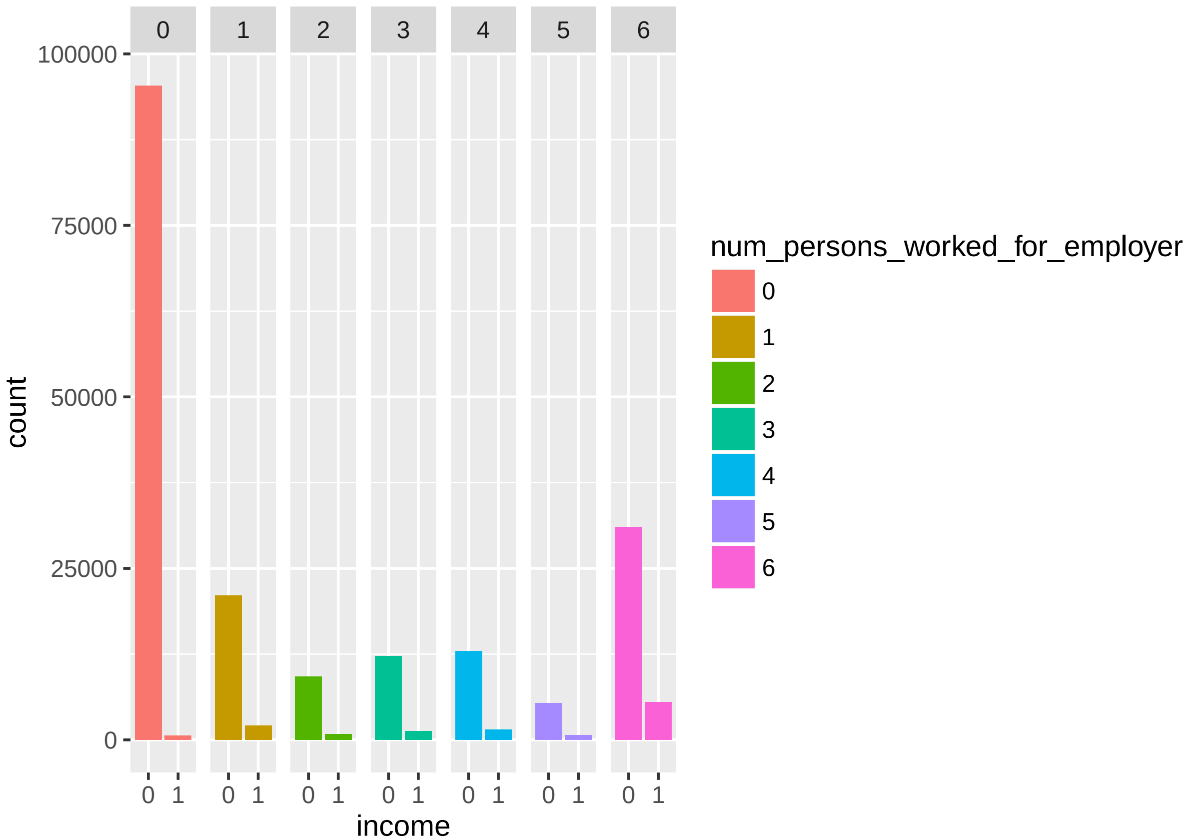

```
ggplot(train_df, aes(x = num_persons_worked_for_employer, fill=income)) +
  geom_histogram(binwidth = 2, alpha = 0.5, position="identity")
```



We

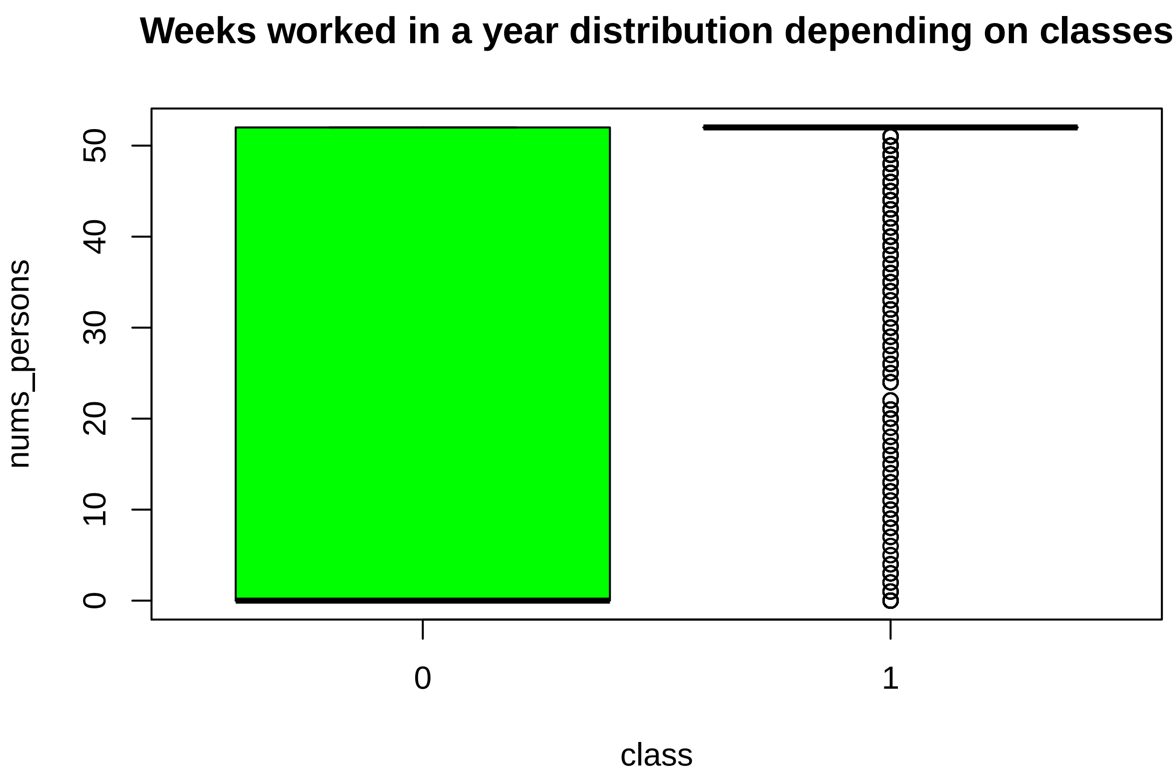
doesn't have that much information about this features, metadata stipulates that this is a continous variable but it rather seems categorical. We can see that people in the last value possible, 6, are more likely to earn +50K. Thus I turned this feature into a factor.

```
train_df$num_persons_worked_for_employer <- as.factor(train_df$num_persons_worked_for_employer)
qplot (income, data = train_df, fill = num_persons_worked_for_employer) + facet_grid (. ~ num_persons_wo
```



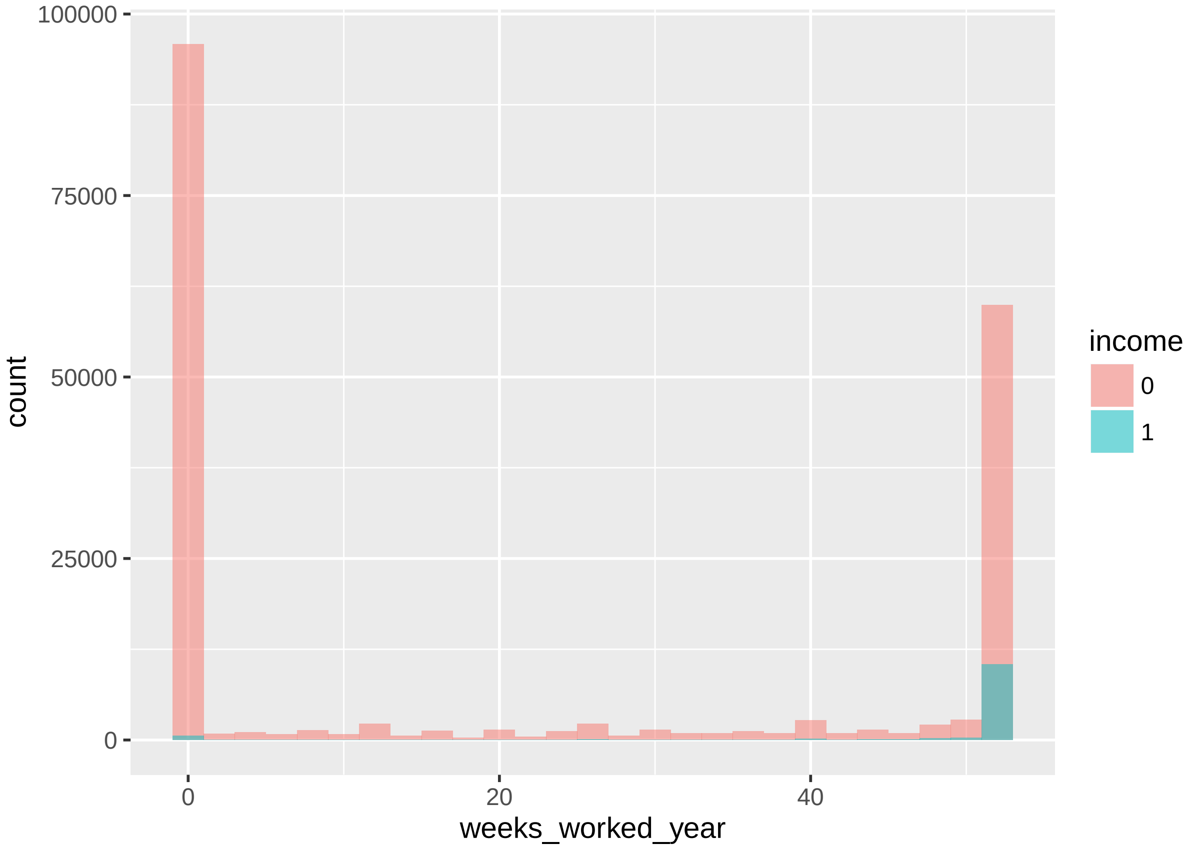
Weeks worked in a year

```
boxplot (weeks_worked_year ~ income, data = train_df, main = "Weeks worked in a year distribution depend
```



Very Relevant Feature, the tendency is clear, the more you work the more you are likely to earn. People who earn +50K work in average the full number of weeks in a year, and this graph is even more explicit

```
ggplot(train_df, aes(x = weeks_worked_year, fill=income)) +
  geom_histogram(binwidth = 2, alpha = 0.5, position="identity")
```



Nevertheless we are biaised by the big amount of children and students that are still not on the job market

Cleaning

Based on these observations I decided to remove the columns that don't give us information and that could mislead the classifier. I also removed the columns that I used to create new and more insightful variables.

```
train_clean_df <- subset(train_df, select = -c(age,industry_code,occupation_code,education,hispanic_orig
migration_msa, migration_reg, mig_within_region, migration_sunbelt, country_father, country_mother,
country_self, veterans_filled, year, income, sum_losses_gains))

# Re append the income at the end for esthetic purpose
train_clean_df$income <- train_df$income
```

We are now all set to train our classifiers on the data. In the next part I will train a Random Forest Classifier and then try to find improvements. The feature that I was not sure about will be kept or removed according to their importance in the random Forest