Mag. Markus Hofer
FT Software Solution GmbH.
Schottenring 33, 1010 Vienna
Subject: Application to the Data Science position

Georgi Dinev
BSc Student at TUWien, Engineering Physics
Adelheid-Popp-Gasse 24, 1220 Vienna

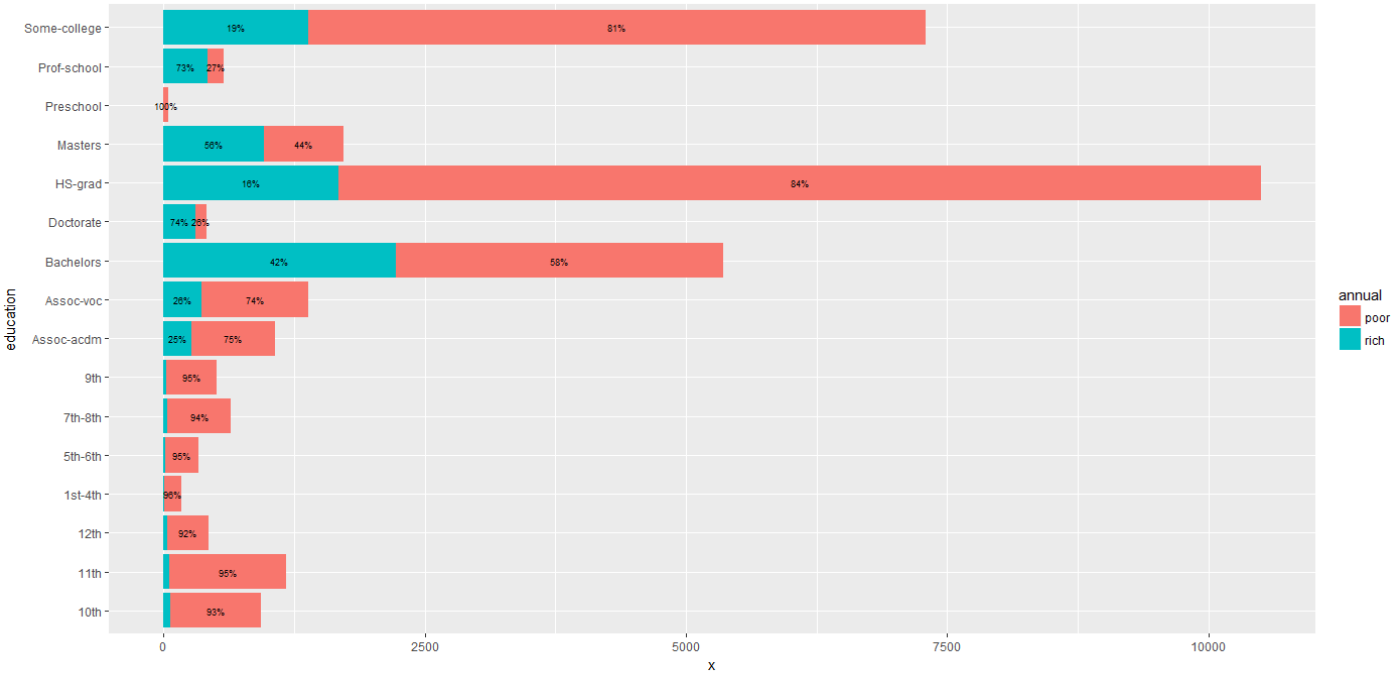
Predicting the income

(Challenge 3.)

Given the data on the site: http://archive.ics.uci.edu/ml/datasets/Adult
I had to subtract useful infromation regarding whether a person's annual salary exceeds *$50 000* a year, and if so, which attributes have the most effect on it.
The original dataset contains 14 different attributes, however some of them does not provide crucial information (for example: the properties **education** and **education-num** are almost identical, **education** however has more information, than **edu-num**).
In the followings I will only consider the attributes:

1. Age (from 0-90 by steps of 1)
2. Education (Discrete, type of highest education)
3. Annual (Binary, >50K,<=50K or 1,0 or Rich,Poor)

I used **R** with **ggplot2** to analyze the given data, and to find correlation between these properties.
Firstly I wanted to see the distribution of the salaries, when we group people up by their highest level of education, the results are on the following diagramm:

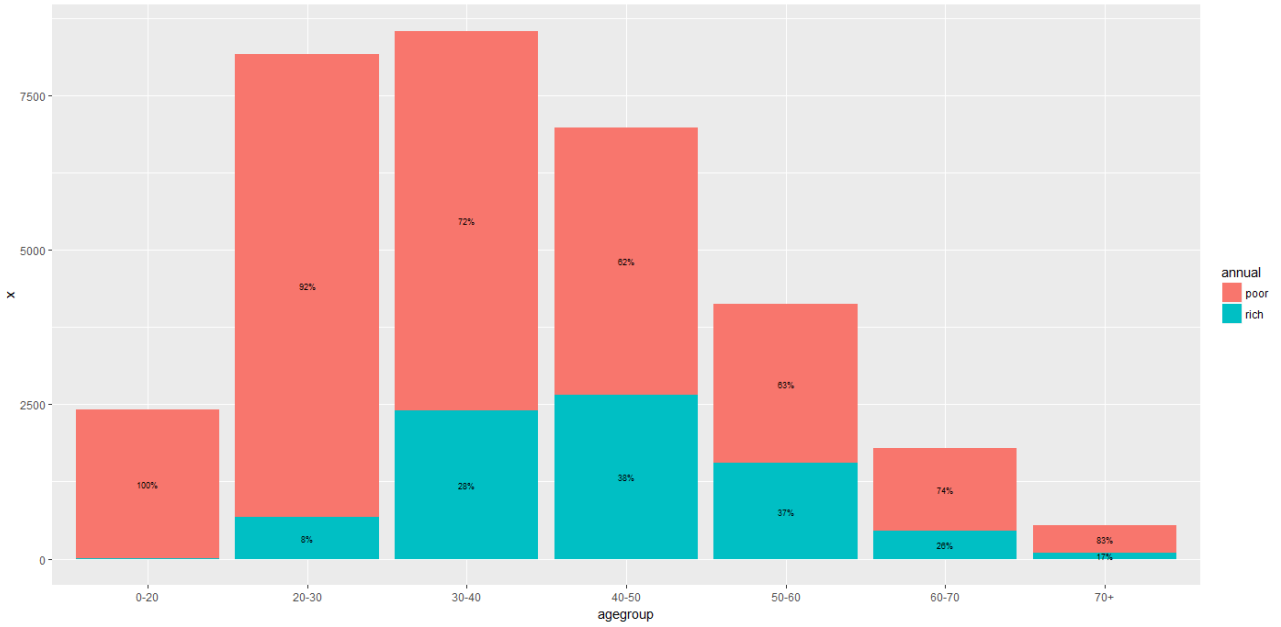


A strong correlation is visible between the annual salary and the type of education (as expected), thus I would say, that the dataset is not biased when it comes to represent the value of education in this context.

Mag. Markus Hofer
FT Software Solution GmbH.
Schottenring 33, 1010 Vienna
Subject: Application to the Data Science position

Georgi Dinev
BSc Student at TUWien, Engineering Physics
Adelheid-Popp-Gasse 24, 1220 Vienna

For the second test, I grouped all the data/individuals in 7 groups considering their age, and plotted the annual income:



As with age, a person's experience grows, so shall the salary too.
Again we can clearly see that in the age group 40-50 years is the peak of the number of „rich" individuals, which also fits well with the expectations.

These two statisctics could be used as a base for a linear regression model to predict whether a person makes over *$50 000* given his/her age and educational background.

If we only consider the correlation between **age** and **salary**, the coefficients in the linear model are surely the probabilities that are visible on the chart, with a neglegable offset (only 2 people earn more than $50K and 2408 individuals earn less than that, in the ages between 0-20)

```
lm(formula = data1$annualbinary ~ data1$agegroup)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38035 -0.28165 -0.08331 -0.00083  0.99917

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.0008299  0.0082852   0.100     0.92
data1$agegroup20-30 0.0824830  0.0094294   8.747   <2e-16 ***
data1$agegroup30-40 0.2808224  0.0093810  29.935   <2e-16 ***
data1$agegroup40-50 0.3795224  0.0096091  39.496   <2e-16 ***
data1$agegroup50-60 0.3739279  0.0104269  35.862   <2e-16 ***
data1$agegroup60-70 0.2541924  0.0126871  20.035   <2e-16 ***
data1$agegroup70+   0.1732442  0.0193650   8.946   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4067 on 32554 degrees of freedom
Multiple R-squared:  0.09545,   Adjusted R-squared:  0.09529
F-statistic: 572.5 on 6 and 32554 DF,  p-value: < 2.2e-16
```

Mag. Markus Hofer
FT Software Solution GmbH.
Schottenring 33, 1010 Vienna
Subject: Application to the Data Science position

Georgi Dinev
BSc Student at TUWien, Engineering Physics
Adelheid-Popp-Gasse 24, 1220 Vienna

In order to create a somewhat more precise model, I also checked the linear correlation between annual salary and education as well as age (not grouped).
The coefficients are:

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.1733108  0.0140869 -12.303  < 2e-16
data1$age                     0.0064057  0.0001621  39.521  < 2e-16
data1$education 11th          0.0171143  0.0170486   1.004 0.315458
data1$education 12th          0.0445417  0.0225984   1.971 0.048731
data1$education 1st-4th      -0.0865511  0.0325772  -2.657 0.007893
data1$education 5th-6th      -0.0533542  0.0248049  -2.151 0.031487
data1$education 7th-8th      -0.0750979  0.0199566  -3.763 0.000168
data1$education 9th          -0.0371790  0.0213395  -1.742 0.081472
data1$education Assoc-acdm    0.1822173  0.0174060  10.469  < 2e-16
data1$education Assoc-voc     0.1875650  0.0164556  11.398  < 2e-16
data1$education Bachelors     0.3392244  0.0137787  24.619  < 2e-16
data1$education Doctorate     0.6086662  0.0230120  26.450  < 2e-16
data1$education HS-grad       0.0831616  0.0132687   6.268 3.71e-10
data1$education Masters       0.4477287  0.0158212  28.299  < 2e-16
data1$education Preschool    -0.1006260  0.0558510  -1.802 0.071604
data1$education Prof-school   0.6210541  0.0206120  30.131  < 2e-16
data1$education Some-college  0.1345023  0.0135052   9.959  < 2e-16
```

Surprisingly, in this model, the coefficient for the age variable is almost neglegable.

Observations:

- As I first sorted the list for workclass, I noticed that there are several (around 1800) entries with unknown occupations.
- For a more complex analysis of the dataset we should consider implementing the wighting factor from the attribute **fnlwgt**.
- If we want to search for the attribute, which has the most effect on the annual salary, we could use the ID3 algorithm.
- To further increase the accuracy of our model, we could calculate the correlation coefficient (I would use Spearman correlation, as we assume that the annual salary is in linear relation to the given attribute)
- As some attributes are discrete (e.g. **Education**-Bachelors, Masters,...), and converting them into numeric would cause false weighting, a Chi-Square Test could be used to determine the accuracy of the model.