Exploratory Data Analysis for 1994 US census

Shudi Zhao

October 13, 2019

Contents

Introduction

This paper use one dataset which was extracted from the 1994 US census database. There are 15 variables in this dataset, 6 of them is continuous variables and 9 of them is categorical variables. The variables are age, workclass, fnlwgt, education, education-num, martital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country and salary. Since this paper is about exploratory data analysis, so I have to explore every single variables to see their variation and covariation with other variables. Then, I came up with 14 questions to start my exploratory journey. After some visualizations I answered 4 interesting questions. I also came up with new questions based on the original questions.

Life is tough, we can't live without money. When I was little, my parents told me the importance of higher education, the higher you have achieved, the more money you will get. This memory jump out of my head when I was exploring this dataset. I want to find out if it is true, then this question becomes the first question that I answered. However, time is also precious, we can't spend all our time on working. This dataset we use is a really good tool to understand the mysterious of both time and money.

Data cleaning

The first thing I need to do is clean the data. When I first import this data set, I found that this data set doesn't have column names. I forgot how to fix this problem, so I checked the textbook. I was lucky, because I found the solution. First, I count the number of columns. Then, I passed col_names a character vector for all correct column names which I found from the data set description.

Then, I saw a lot of values in this data set are "?", so I decided to treat these "?" as missing values. First, I checked all variables to see if there is any missing values. I found there are missing values like the question mark in workclass, occupation and native_country.Then I use replace_with_na() replaced these "?" with NA.

Next, I played with the data for a while. I found each education level were assigned with a distinct years of education. This means they are the same thing, but the education number is a continuous representation for education level. In addition, from the table below, we can see that 1st-4th are treated as one year, 5th-6th are treated as one year, and 7th-8th also are treated as one year. Thus, I think this column is not neccessary. I will delete this column from this data set.

```
## # A tibble: 4 x 2
##   education education_num
##   <chr>             <dbl>
## 1 Preschool             1
## 2 1st-4th               2
## 3 5th-6th               3
## 4 7th-8th               4
```

I also deleted the relationship column, because I can find the same information from the sex column and the marital status column. For example, wife and husband are types of relationship, but I can get this information from both sex column and marital-status. Wife is married female and husband is married male, so the relationship column is kind redundent. I also did a lot of exploration with final weight, but it turned out to be nothing. First, I googled the definition for the final weight. It tells me the definition for final weight is, "The continuous variable fnlwgt represents final weight, which is the number of units in the target population that the responding unit represents." I use box plot to see the relationship between the final weight with education, workclass, race and occupation. Since, I can't fully understand the meanning of final weight, I can't say anthing about it. Thus I decided to delete this column.

The mysterious salary

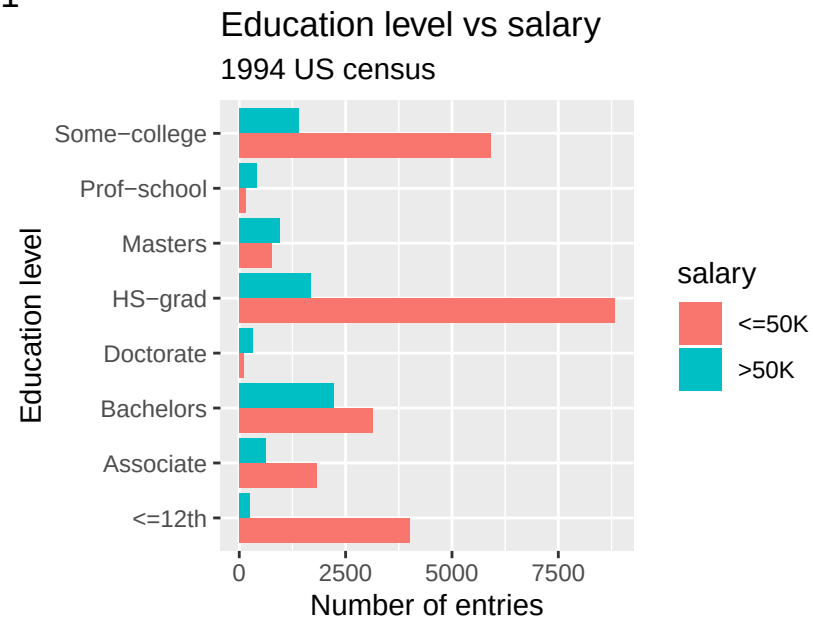
Is it true that the higher education you have achieved, the more money you will earn?

First, I use bar chart to see the relationship between education and salary. I saw there are too many types of categories in education variable are redundant, so I renamed the education type that are less and equal to 12th grade as "<=12th", assoc-acdm and assoc-voc as "Associate". This will make my visualization more clear. When I was doing rename, it took me a long time to figure out how to do it. I tried to write a function to rename it, but it failed. Then, found that I can use str_replace_all to do it which is the one we have learned from our class lecture.

After modify the dataset, I found the answer. From Figure 1, we can see that in those who graduated from professional school and the people who got the master and doctorate degrees have more people earn greater than 50K. I also found a positive trend which as the education level increase the amount of people who earn greater than 50K in its propotion will also increase. Thus, it is true that the higher education you have achieved, the more money you will likly to get.

However, education is not the only attributes that will effect the difference in salary. I explored further, I tried to connect salary with other attributes and I want to come up with a model to illustrate the relationship between these attributes with salary. Unfortunatly, It was too complicate for me to finish it, since I only learned the basics for now. At least, I found an interesting question during my exporation. I found that among married and single men, the propotion of those who earn more than 50K in married men is greater than single men.

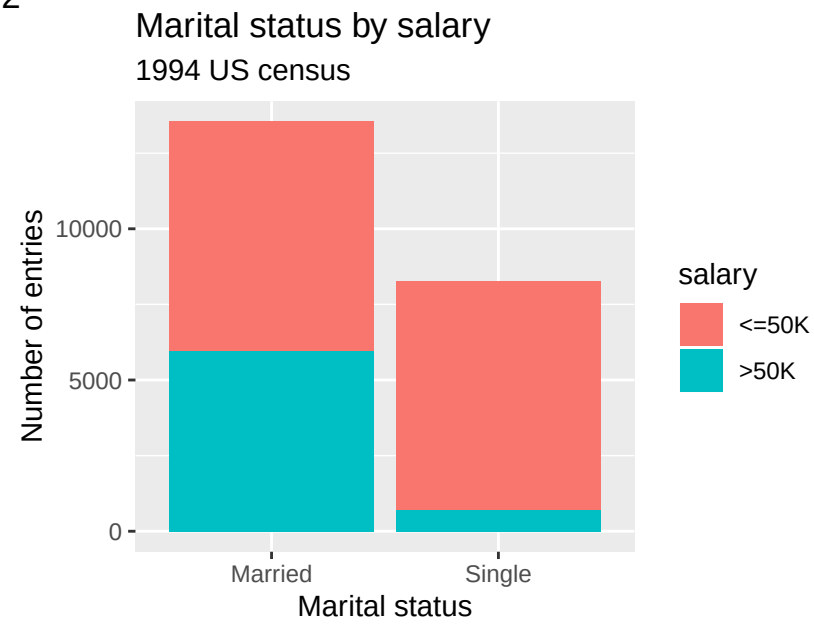
Figure 1



To prove whether my finding is correct or not, I will use a graph to show it. At the beginning, I have to modify my data set for marital_status, because there are seven types of marital status. I changed the name for Married-civ-spouse, Married-spouse-absent and Married-AF-spouse to "Married", then I changed the name for never-married, divorced, separated and widowed to "Single". Next, I use fillter to keep only males in this dataset, then I finished my bar chart.

In Figure 2, there are two bars on the graph. The left one is for married men and the right one is for single men. We can see the propotion of those who earn more than 50K in married men is a lot greater than single men. Thus my finding is correct baesed on this dataset for 1994 US census.

Figure 2



Why the number of married men who earn more than 50K is a lot greater than the number of single men who earn more than 50K? Is it because the married men are much older than singles? I did a calculation for their average age to prove whether my assumption is right. From the table below, we can see there are 5965 Married men who earn more than 50K with average age 44.88. And there are 697 single men who earn more than 50K with average age 42.42. After calculate their difference, there are 5268 more marries men than singles and with average age difference 2.46. Actually, 2.46 average age difference is not very big, so my assumption is wrong. The number of married men who earn more than 50K is a lot greater than the number of single men who earn more than 50K is not because the married men are much older. Probably the reason is because that poor men cannot afford or support a family.

Once you make a lot of money, what do you want to do with these money? Some people choose to make an investment. Next question, I want to explore if it is true for those who make more than 50K also have more capital gain than those who earn less than 50K.

```
## # A tibble: 4 x 4
## # Groups:   marital_status, salary [4]
##   marital_status salary average_age     n
##   <chr>          <chr>        <dbl> <int>
## 1 Married        <=50K         42.7  7576
## 2 Married        >50K          44.9  5965
## 3 Single         <=50K         31.6  7552
## 4 Single         >50K          42.4   697
```

The power of money

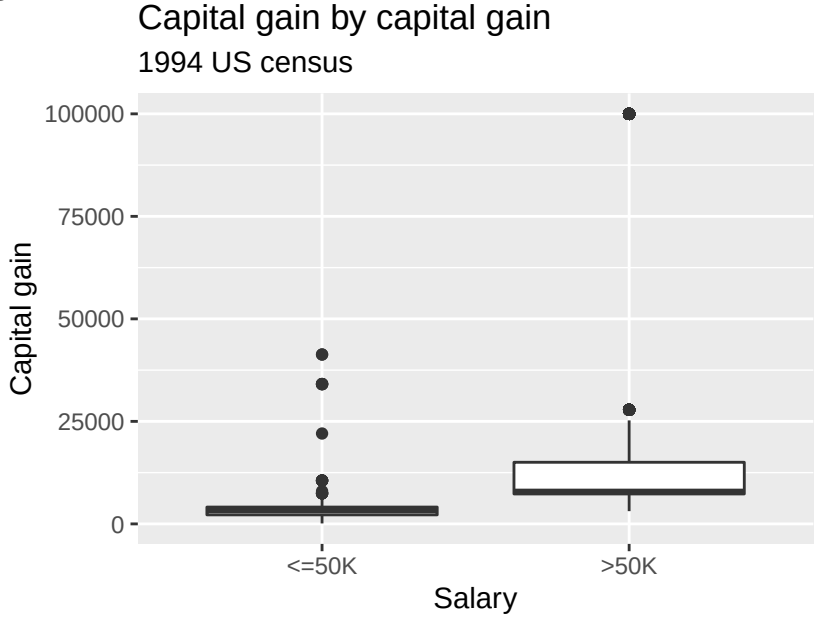
Is it true for those who make more than 50K also have more capital gain than those who earn less than 50K ?

First, I searched the defination for capital gain and capital loss. The capital-gain and capital loss are income from investment sources other than salary. Then, I found a lot of zeroes in this column, so I have to check whether my data is big enough or not. I filter out of the capital gain that are equal to zero in order to make my visualization accurate. I found there are 2712 matched observations which is 8.32 percent of the dataset, so it is big enough to move on.

Figure 3 is a box plot for salary by capital gain. By looking at the graph, I found there is a positive trend in this graph, as the salary increased the capital gain also increased. Thus, people who make more than 50K also have more capital gain than those who earn less than 50K, beacuse they have more many to make an investment.

I also spend a lot of time to see the relationship between capital loss and capital gain with sex. I want to know if it is true that male make more investment than female. From the graph that I ploted, I can see it is true, but I forgot one condition which is the number of males are twice bigger than females in this dataset. Thus, I went to a dead end.

Figure 3



The mysterious time

Count the average time of work for those who earn a little and a lot for each country. Which country stands out ?

In this question, I want to explore the relationship between native country, salary and work hours per week. First, I need to take out the NAs, because there are a lot of NAs in native country column. Next, I groupd by salary and native country, and count their average hours of work per week. Then, I arranged them in a descending order to see which country have the highest average hours. However, there are many countries have a very small numbers of n. Thus, I used n $>$ 30 to filter out the countries that with small numbers of n. The reason I use n $>$ 30 is because 30 is a magic number in statistics. I looked up online, it saids, The central limit theorem states that if you have a population with mean and standard deviation and take sufficiently large random samples frome the population with replacement text annotation indicator, then the distribution of the sample mean will be approximately normally distributed. This will hold true regarless of whether the source population is normal or skewed, provided the sample size is sufficiently large(usually n $>$ 30)."

```
## # A tibble: 6 x 4
## # Groups:   salary, native_country [6]
##   salary native_country Average_hour     n
##   <chr>  <chr>                 <dbl> <int>
## 1 >50K   Mexico                46.6    33
## 2 >50K   India                 46.5    40
## 3 >50K   Canada                45.6    39
## 4 >50K   United-States          45.5  7171
## 5 >50K   Germany               45.0    44
## 6 >50K   Philippines           43.0    61
```

From the table above, we can see that the Mexico stands out with average 46.57576 hours. The interesting things is that these six native countries have a common relationship which is their salary is all graater than 50K. Then, which country has the highest average hours in those salary less and equal to 50K? I ploted one

more table to illustrate this question. From the table below, we see that the Dominican Republic has the highest average hours among those salary less and equal to 50K.

Actually, hours per week is a pretty good attribute to explore, so I did a further explore for hours per week in the next question.

```
## # A tibble: 6 x 4
## # Groups:   salary, native_country [6]
##   salary native_country     Average_hour     n
##   <chr>  <chr>                     <dbl> <int>
## 1 <=50K  Dominican-Republic         42.3    68
## 2 <=50K  Portugal                   41.9    33
## 3 <=50K  Japan                      41      38
## 4 <=50K  England                    40.5    60
## 5 <=50K  South                      40.2    64
## 6 <=50K  Mexico                     40.0   610
```

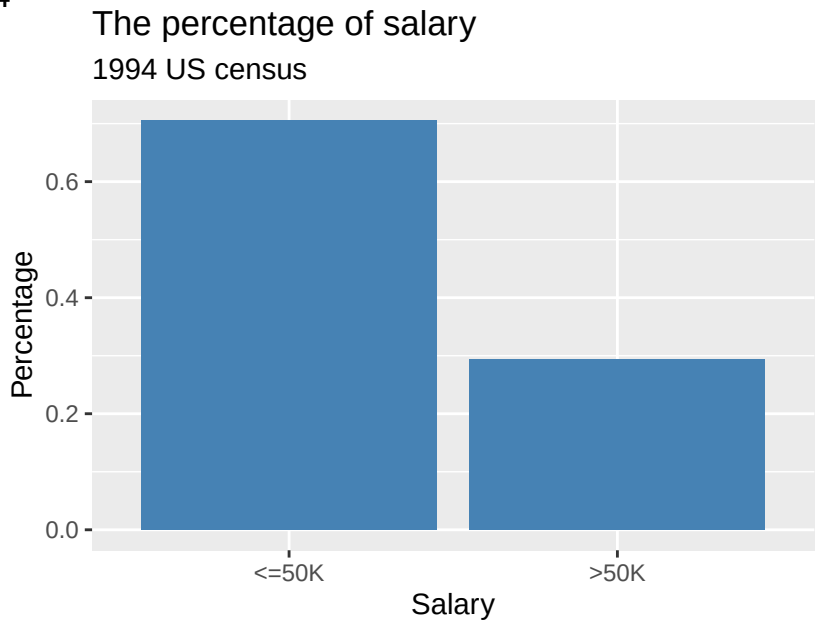
Time is money

What is the maximum number of hours a person works per week? How many people work such a number of hours, and what is the percentage of those who earn a lot among them ?

To answer this question I have to arrange the hours per week column in a descending order, then I can see the maximum number from the table which is 99 hours. Wow, that is surprising me. Let's explore it further. I filtered the rows that its hours per week is equal to 99. Then, I found there are 85 people who worked 99 hours per week. That is unbelieveable.

Next, I want to draw a graph to see the percentage of those who earn a lot among the them. By calculation, I found that there are only 29.4 percent of peolpe who work 99 hours a week earn greater than 50K. It was ridiculous, that 80.4 percent of people who work 99 hours a week earn less than 50K. See Figure 4.

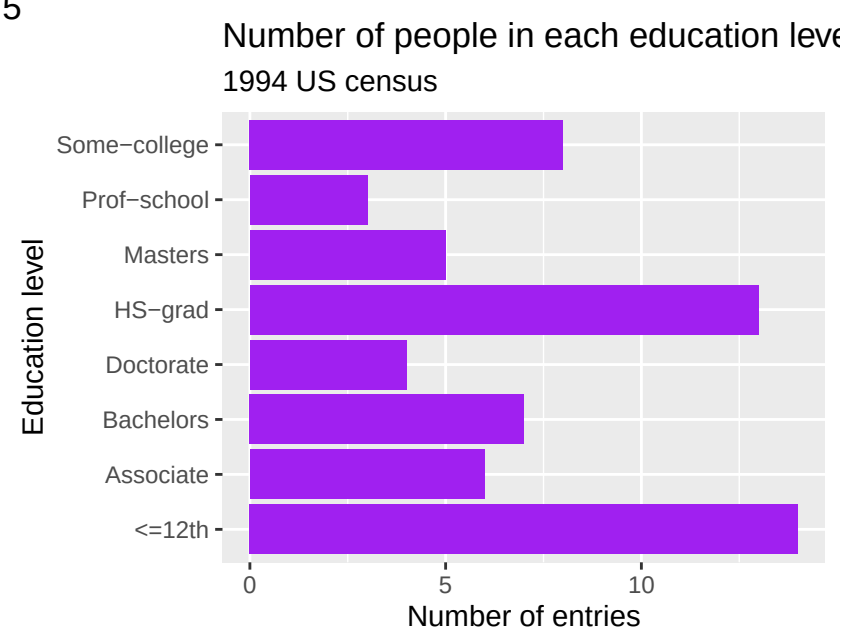
Figure 4



I want to explore further to see the relationship between education level and the number of people who work 99 hours per week and earn a little. Although, my observations are small, but it is intersting. From the

graph below, we can see there are more people from those who only gradurated from high school and not gradurated from high school. Does it mean the higher education you received, the less hours you are going to work? In Figure 5, I found a positive trend that is the higher education you have received the less hours you will work. However, this conclusion is not accruate, because the sample size is too small. Next, I want to use a larger sample size to see whether it is correct or not.

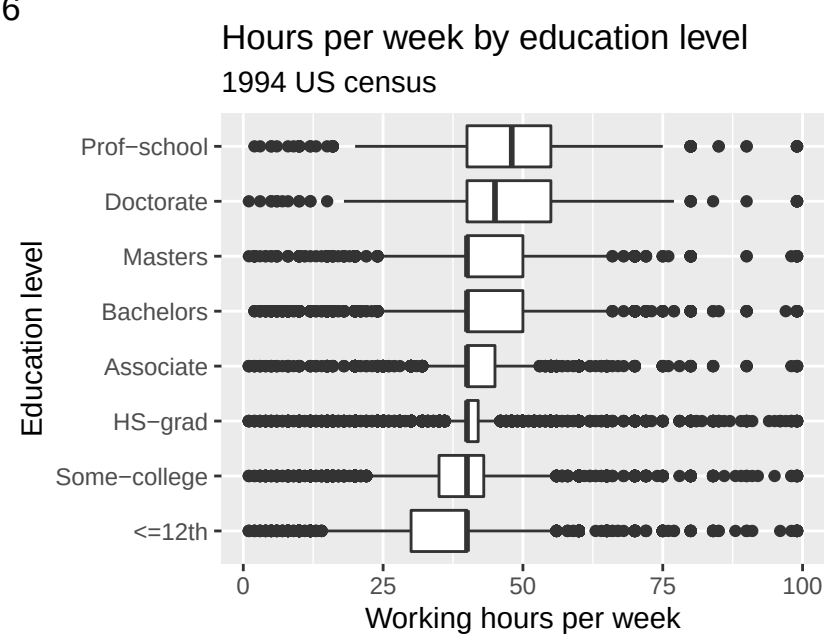
Figure 5



Is it true that the higher education you received, the less hours you are going to work??

First, I choose the box plot to answer this question. At the beginning, I found a problem after I ploted the box plot. I can't see their trend, because the variables are not odered. Then, I used reorder() to arrange them in an accending order by its average hours. And reorder() is what I have learned in class, but that is the first time I use it. Surprisingly, there is a positive trend in Figure 6. We can see that the education level is increased as the average work hours is increased. Thus, I can conclude that the higher education you received, the more hours you might need to work, so my previous assumption is wrong.

Figure 6



Conclusion

The main conclusion of my analysis is in four part. The first one is that the higher education will increase your possibility to earn more money. That's one of the reason that I'm going to college and pursue for higher education. In fact, this data analysis proved that my decision is correct. The second one is that the number of married men who earn a lot is greater than the number of single men who earn a lot. This is not because of the age difference, is probably because the married men have more stress, since they have to take care of their families. If you want to earn more money, go get married. I'm just kidding, because there are more ways to do it. The third one is that the richer people will gain more investment than those who earn a little. This is reasonable, because the people who earn a little will spend most of their money on every day life like, housing, foods, taxes and etc... The last one is that, the education level have a positive impact on the hours of work per week. Which as the education level increases the hours of working also increases. This conclusion is surprising me, because I think it should be the opposite. Perhaps, as your education level and knowlage increased your value to the socialty and companies also increased, so you will most like to work more hours.

Some of my data analysis might not be 100 percent correct, but I have tried my best to do it and it was based on my current knowlage. Overall, this exploratory data analysis is very hard than what I was expected. It was very challenging when I was working on it. I spend a lot of time to put my questions and findings together, because I don't really know their relationship before I wrote this paper. I started to connecting them while I was writting this paper. Another challenging task is to convert this R markdown into pdf which I spend a lot of time on it. I googled online to see how to create a pdf document from R markdown. First, I installed the tinytex package. Then, I found out that I have to specify the pdf document output format in the very top left corner. I think this is a really good practice for me to understand what I'm going to do in the really world. I still have a lot of things to learn in order to become a real data scientist.