**“Will you have an annual income over $50,000?” – An analysis of the 1994 US Census**

**Authors: Aleph Aseffa, Wagih Henawi, and Haruko Okada.**

**Project Title** - Will you have an annual income over $50,000? – An analysis of the 1994 US Census.

**Project Goals**

For our project, we took a dataset which was a modified version of the US Census of 1994 which had information on the characteristics of 32500 individuals and for each individual, a prediction of whether their annual income would be over $50,000.

Initially, our reach goal was to create a procedure that would calculate the probability of a certain hypothetical individual earning over $50,000 a year (when data about the individual is entered by a user). Our safe goal was to find the characteristics that make a person most likely to earn over $50,000 annual income.

However, we found both goals to be rather easy to reach, so we ended up ignoring the safe goal, implementing the reach goal, and going one step further. We selected education level, race, and sex as the three most significant factors that affect earning potential and wanted to explore how each of these factors contribute to the chance of an individual earning over $50k. This became the new reach goal of our project.

**Description of original dataset**

Link to dataset: <https://www.kaggle.com/uciml/adult-census-income/home>

The original data for this dataset was the 1994 Census (provided by the census bureau). Ronny Kohavi and Barry Becker from the UCI Machine Learning Institute extracted data from this census database and used the relevant data and mathematical techniques to predict whether the income of the person from the census was over $50,000 or not. This prediction of income was then added to the dataset. For this project, we are using the database that has been edited by Kohavi and Becker.

This dataset includes 32562 rows and 15 columns. The first row of the dataset is a description of what each column represents. The columns of the database are as follows: age, workclass, fnlwgt, education, education.num, marital.status, occupation, relationship, race, sex, capital.gain, capital.loss, hours.per.week, native.country, income.

Explanation of each column (some parts taken from scg.sdsu.edu/dataset-adult\_r):

Note: if the data held in the column is not described, it is because we will not be using the data in our analysis.

Age = the age of the individual (range from 17 to 90)

Workclass = The classification of the organization the individual is working at (can be either “Private”, “Self-emp-not-inc” (self-employed but not incorporated), “Self-emp-inc” (self-employed and incorporated), “Local-gov”, “State-gov”, “Without-pay”, “Never-worked”).

Fnlwgt = “The number of people the census takers believe that observation represents”.

Education = the highest level of education the individual has (can be either “Bachelors”, “Some-college”, “11th”, “HS-grad”, “Prof-school”, “Assoc-acdm”, “Assoc-voc”, “9th”, “7th-8th", “12th” “Masters”, “1st-4th”, “10th", “Doctorate”, “5th-6th", “Preschool”)

These categories are fairly self-explanatory.

Education.num = “Highest level of education in numerical form”

Occupation = The occupation of the individual.

Marital.status = The marital status of the individual (can be either “Married-civ-spouse” (married civilian spouse), “Divorced”, “Never-married”, “Separated”, “Widowed”, “Married-spouse-absent”, “Married-AF-spouse” (married, armed forces spouse))

Explanations of civ and AF taken from fr.answers.yahoo.com/question/index?qid=20111128051933AAAb8hh

Relationship = What the individual represents in the family (can be either “Wife”, “Own-child”, “Husband”, “Not-in-family”, “Other-relative”, “Unmarried”)

Race = The race of the individual (can be either “White”, “Asian-Pac-Islander”, “Amer-Indian-Eskimo”, “Other”, “Black”)

Sex = The sex assigned at birth of the individual (can be either “Female” or “Male”)

Capital-gain = “Capital gains recorded”

Capital-loss = “Capital losses recorded”

Hours-per-week = The number of hours per week the individual works (range from 1 to 99)

Native-country = The country that the individual is a national of.

Income = The income of the individual (either “<=50k” or “>50k”)

In these categories, if the required data is not available, a “?” is present.

**Description of wrangled dataset**

While this dataset had a vast amount of useful information, we did not need all of that for our analysis.

We selected age, work-class, education-level, race, sex, hours-per-week, and income as the relevant categories for each individual in the dataset, and removed all other columns. We also removed all instances of data for individuals that were not from the United States. Additionally, we changed the age and hours-per-week from continuous to discrete categories.

The age categories became: 17-22, 23-45, 46-65, and 65+.  
The hours-per-week categories became: 1-10, 11-25, 26-45, 45+

We then defined the variable wrangled-dataset as the dataset to be used for the rest of the analysis.

We also defined another variable no-income which is the same as wrangled-dataset but has the income field removed. This was done as it conflicted with one of our procedures.

Explanation of each column in wrangled-dataset (each column contains a string):

Age = The age range of the individual (can be either “17-22”, “23-45”, “46-65”, or “65+”)

Work-class = The classification of the organization the individual is working at (can be either “Private”, “Self-emp-not-inc” (meaning self-employed but not incorporated), “Self-emp-inc” (meaning self-employed and incorporated), “Local-gov”, “State-gov”, “Without-pay”, “Never-worked”).

Education-level = The highest level of education the individual has (can be either “Bachelors”, “Some-college”, “11th”, “HS-grad”, “Prof-school”, “Assoc-acdm”, “Assoc-voc”, “9th”, “7th-8th", “12th” “Masters”, “1st-4th”, “10th", “Doctorate”, “5th-6th", “Preschool”)

Race = The race of the individual (can be either “White”, “Asian-Pac-Islander”, “Amer-Indian-Eskimo”, “Other”, “Black”)

Sex = The sex assigned at birth of the individual (can be either “Female” or “Male”)

Hours-per-week = The range for number of hours per week the individual works (can be either “1-10”, “11-25”, “26-45”, or “45+”)

Income = The income of the individual (either “<=50k” or “>50k”)

Note: the no-income dataset is exactly the same as wrangled-dataset but with the income column removed. This was needed to avoid a conflict with a different procedure.

**Description of algorithms**

We used quite a few procedures for our data analysis.

The procedures we used for the data-wrangling were:

* change-age-single
* change-hours-single
* remove-columns
* remove-income-single
* remove-foreigners

These procedures are explained in the documentation.

We have two main algorithms, plot-lst-helper and percentage.

percentage was used for our initial reach goal which was to create a procedure that would calculate the probability of a certain hypothetical individual earning over $50,000 a year (when data about the individual is entered by a user).

How percentage works:

1. First filter-table is run
   1. Filter-table takes in a table (which in this case is wrangled-dataset), and the characteristics of a hypothetical individual (age, wc, ed, r, s, hrs) [these parameters are explained in the documentation].
   2. Filter-table runs through the inputted table (wrangled-dataset) and creates a new list of each individual in wrangled-dataset that matches the characteristics of the hypothetical individual.
   3. This list is passed into percentage.
2. Percentage contains a recursive let function called over-50k-tally which takes in the table passed into it by filter-table and:
   1. For each list in the table, the 7th element of the list (the income column) is inspected.
   2. If that element is equal to “>50k” (meaning the individual with those characteristics earns over 50k), 1 is added to the tally.
   3. Else (meaning the element is equal to “<=50k” and the individual with those characteristics is earning less than or equal to 50k), nothing is added to the tally.
   4. This procedure recurses over the table until the table is empty.
   5. It then passes the tally to percentage.
3. Percentage then calculates the length of table (which is the number of people in wrangled-dataset who match the characteristics entered by the user).
4. Then, the tally of people in the table who earn over 50k are divided by the length of the table.
5. This number is then multiplied by 100.
6. This is the output of percentage (a number between 0 and 100 inclusive) and represents the percentage of people with the characteristics the user entered who earn over 50k. Thus, our initial reach goal has been met.  
   However, as the database is not nationally representative, this is not the real chance of an individual with such characteristics earning over 50k – it is more the percentage of individuals in our database with such characteristics that earn over 50k. This is an important limitation to note.

Our reach goal was then changed to a more advanced one: We selected education level, race, and sex as the three most significant factors that affect earning potential and wanted to explore how each of these factors contribute to the chance of an individual earning over $50k.

The main procedure which achieves this is plot-lst-helper:

How plot-lst-helper works:

1. plot-lst-helper takes in individuals (a table) and lst-so-far (a list).
2. It then goes through each individual list in individuals and for each list (where each list contains characteristics of an individual in the format explained in the documentation):
   1. Runs chance-of-over-50k on it
      1. This produces a number that represents the percentage of individuals with such characteristics that earn over 50k
   2. Runs get-desc on the list
      1. This outputs a string in the format education-level, race, sex
   3. These two are put into a single list in the following format:
      1. ‘(desc prob)
         1. Desc is the description of the individual in terms of education-level, race, sex. It is a single string.
         2. Prob is a number between 0 and 100 inclusive
3. The final output is a table where each list in the table contains a description of the individual and the percentage of people in individuals that have those percentage and earn over 50k.
4. This format for the data puts it in the ideal format for plotting a discrete histogram. We can now run the procedure (plot discrete-histogram) on the output of this procedure to visualize our analysis. This is defined as the variable final-plot.
5. This gives us the way to visualize how different combinations of education level, race, sex affect chances of earning over 50k.

**Instructions for running code**

**To determine the percentage of individuals with such characteristics that earn over 50k:**

Step 1: (filter-table wrangled-dataset age wc ed r s hrs)

Conditions:

* age is an element of age-list
* wc is an element of wc-list
* ed is an element of ed-list
* r is an element of race-list
* s is an element of sex-list
* hrs is an element of hrs-list

This creates a list of individuals with the same characteristics that you entered.

Step 2: Run (percentage table) on the table you just created. Note: table must not be empty.

This will give you the proportion of individuals in table (the list of individuals with the same characteristics you entered) that earn over 50k.

This is your answer.

Alternatively, you can do the following:

(percentage (filter-table wrangled-dataset age wc ed r s hrs))

Sample input and output:

> (percentage (filter-table wrangled-dataset "46-65" "Private" "Bachelors" "White" "Male" "26-45"))

68 47/71

**To investigate how the combinations of education level, race, and sex contribute to the percentage of individuals earning over $50,000**

Step 1: Create all combinations of education level, race, and sex:

(combine\* ed-list race-list sex-list)

Step 2: Format the previous table into a format for filter-table to work.

This has already been done and defined as a variable: correct-format-combs

Step 3: Filter out the combinations that are not present in wrangled-dataset

(filter-combs correct-format-combs)

Or alternatively, as it has been defined as a variable: filtered-combs

Step 4: Create a list that is plottable in a discrete histogram

(plot-lst-helper filtered-combs null)

Step 5: Plot the previous table

(plot (discrete-histogram (plot-lst-helper filtered-combs null))

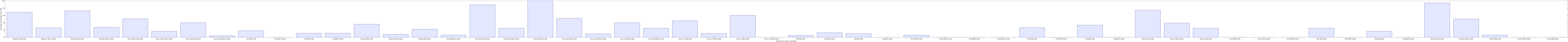
#:x-label "Education level, Race, and Sex"

#:y-label "Percentage in dataset that earn >$50k"

#:width 15000)

Or alternatively, it has been defined as the variable final-plot so you can just type that.

**Description of what the analysis shows**



The above figure is small extract of what final-plot produces.

For our analysis, we set these factors as constant:

* Age = “46-65” as we believed individuals would be nearing their maximum earning potential during this age range.
* Work-class = “Private”
* Hours-per-week = “26-45”

Final-plot shows us the different combinations of education level, race, and sex, and the corresponding percentage of individuals with such characteristics that earn over 50k. A simple visual examination of these results show that in most cases, the biggest factor that affects probability of earning over 50k is sex. In almost each case, keeping education level and race the same while changing sex from male to female, dramatically reduces the percentage of individuals with such characteristics earning over 50k. This indicates that the sex pay gap is real and significant, at least in this dataset.

We had also expected race to disadvantage earning potential, and this hypothesis is supported by our analysis. From the graph and from the plot-lst-helper procedure, we can see how the characteristics which had the highest percentage of individuals earning over 50k were educated white males.

Another interesting analysis is how for black males, education level significantly impacted the percentage of individuals with such characteristics earning over 50k – especially the jump between HS-grad and Bachelors. In fact, this coincides with sociological studies which show how educated black males have significantly higher employment opportunities than those with only high school degrees or high school dropouts (as explained by Elijah Anderson in his book *Against the Wall*, 2009).

No analysis of data is complete without noting limitations. The main limitation of this analysis is the dataset itself. One of the main issues is that there is a sex-bias in terms of data-collection: the dataset itself is approximately 60% male and 40% female. Furthermore, there aren’t enough pieces of data from individuals with races other than white and black. Another problem with the dataset is that it was too small to be nationally representative; to generalize our findings over the entire country, we would need significantly more data.

Another significant limitation is that some combinations of education-level, race, and sex did not have enough entries in our dataset. For example:

> (percentage (filter-table wrangled-dataset "46-65" "Private" "Prof-school" "Black" "Male" "26-45"))

100

Indicates that black males who have gone to professional school have a 100% chance of earning over 50k (at least in this dataset). However, if we check the number of individuals with such characteristics in our dataset, we see that there is only one:

>(filter-table wrangled-dataset "46-65" "Private" "Prof-school" "Black" "Male" "26-45")

'(("46-65" "Private" "Prof-school" "Black" "Male" "26-45" ">50K"))

This indicates how the extreme percentages (those very close or equal to 0 or 100) are likely due to a lack of sufficient data.

Furthermore, this data is rather outdated (from 1994) and it would be interesting to do the same analysis on more recent census data. Perhaps even a longitudinal study could be done to see how the factors that affect earning potential have changed over time. Another limitation is that the way the algorithm is implemented has an implicit assumption that the data sample is nationally representative, which as we have noted, is not the case.

However, at least for this dataset, we can conclude that our data indicates how sex is the most significant determining factor when it comes to percentage of individuals earning over 50k, suggesting that the sex pay gap is real and significant. Race however, does not have as big of an effect. Education levels did have a large effect, but this was true for all races and sexes, so did not indicate any sort of discrimination.