Adult Income Study: Milestone 03

Jimmy Liu and Hannah McSorley

2020-03-17

Introduction

The economic well-being of individuals is reliant on their income, where income is defined as the money an individual (or household) receives on a regular basis. In the United States, the Census Bureau uses income (money received before expenses and deductions) to gauge the population's range of poverty, wealth, and financial security (United States Census Bureau, 2016). There are a variety of factors that can influence one's income, including socioeconomic drivers, education and vocation. This project examines some of the variables that are often related to income.

Description of Dataset

This project works with a dataset of adult incomes obtained from the University of California Irvine (UCI) Machine Learning Repository. The data was donated by Ronny Kohavi and Barry Becker (Silicon Graphics) and was originally extracted by Barry Becker from the 1994 Census database and used for machine learning predictions of whether a person makes over $50,000 per year based on personal factors.

This 1994 income census dataset consists of multivariate categorical and integer data that describe socioeconomic and personal classifiers of adults across the USA. Each instance (32,561) is an individual whose annual income was grouped as either above or below $50,000. Table 1 shows an overview of the 15 attributes (variables), including whether each is categorical or integer and a brief interpretation of the variable.

Show 10 ▼ entries   Search: [   ]

Description of adult income dataset variables

	Variable ⬍	Type ⬍	Description ⬍
1	age	continuous	age, range from 17 to 90 (mean 38.58)
2	workclass	categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3	fnlwgt	continuous	"final weight" - demographic weight by US state controlled by three sets of characteristics including race/age/sex, hispanic origin, and proportion of state population over 16 years of age
4	education	categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5	education-num	continuous	years of education. Range from 1 to 16 (mean 10.08)
6	marital-status	categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7	occupation	categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8	relationship	categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9	race	categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10	sex	categorical	Female, Male

Showing 1 to 10 of 15 entries

Notes on original dataset

A couple of assumptions were made about these data based on information on the Census website. It was assumed that "capital gains" indicate non-cash financial benefits (e.g., food stamps, health benefits, subsidized housing or transportation, employer contributions to retirement programs, medical and educational expenses, etc.), and that "capital losses" include non-cash expenses (such as depreciated value of assets). We also assumed that "education number" indicated the number of years allotted to education.

It is of note that these data are from 1994 census, and the income threshold of $50,000 held a different meaning for wealth than it holds today. Additionally, as this dataset includes socioeconomic attributes, it's worth noting that the majority of data instances were dominated by middle-age, white, US-born, male, private-sector employees. Overall, there appeared to be a fairly even distribution of individuals across occupational sectors and the majority of individuals work approximately 40 hours per week.

Project Objectives

This project developed data analysis skills in R and R studio with a strong focus on writing Rscripts and executing script commands via RStudio Terminal or the command line. The ultimate goal was to generate a full report using a pipeline of scripts, run in sequence, and to create a report using 'Make'.

EDA Research Questions

In this study, we explored the relationships between personal attributes and quantitative income-related variables with the goal of identifying relationships and interesting patterns. We focused on addressing the following exploratory research questions:

1. Is there an observable relationship between personal attributes data and income level?
2. Does the number of hours worked per week relate more to occupation, sex, race, age (or is there no clear relationship)?
3. What is the relationship between education and hours worked per week (e.g. does a person work fewer hours if they have completed more schooling)?

Plan of Action

The variables that effect income may be confounding and are unlikely to be direct, therefore these data may not be appropriate for linear regression analyses. We focus on exploring the relationships variables and identifying relationships and patterns through the initial project's exploratory data analysis and first steps of running a data analysis pipeline.

Exploratory Data Analysis

The original data set was loaded into RStudio where we ran a summary and performed initial exploratory data analysis. A key discovery was that the variable 'income' was not (as might be expected) annual income values for each data instance, it was instead a categorical variable that distinguished whether that instance (row value, person) had earned more than or less than $50,000 USD. Below is a summary of the initial dataset.

```
##      age                  workclass          fnlwgt
##  Min.   :17.00    Private         :22696   Min.   :  12285
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00    Local-gov       : 2093   Median : 178356
##  Mean   :38.58    ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
##                   (Other)         :  981
##         education    education.num             martial_status
##   HS-grad     :10501   Min.   : 1.00   Divorced            : 4443
##   Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse   :   23
##   Bachelors   : 5355   Median :10.00   Married-civ-spouse  :14976
##   Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##   Assoc-voc   : 1382   3rd Qu.:12.00   Never-married       :10683
##   11th        : 1175   Max.   :16.00   Separated           : 1025
##  (Other)      : 5134                   Widowed             :  993
##          occupation          relationship              race
##   Prof-specialty :4140    Husband      :13193    Amer-Indian-Eskimo:  311
##   Craft-repair   :4099    Not-in-family : 8305    Asian-Pac-Islander: 1039
##   Exec-managerial:4066    Other-relative:  981    Black             : 3124
##   Adm-clerical   :3770    Own-child     : 5068    Other             :  271
##   Sales          :3650    Unmarried     : 3446    White             :27816
##   Other-service  :3295    Wife          : 1568
##  (Other)         :9541
##      sex          capital.gain    capital.loss    hours.per.week
```

3

```
##   Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##   Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                  Median :    0   Median :   0.0   Median :40.00
##                  Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                  3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                  Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##       native.country      label
##   United-States:29170   <=50K:24720
##   Mexico       :  643   >50K : 7841
##   ?            :  583
##   Philippines  :  198
##   Germany      :  137
##   Canada       :  121
##   (Other)      : 1709
```

Data handling

Summary of the original data showed that 'capital-gains' and 'capital-losses' were not categorical values like
the 'income' variable, and while these variables were numeric, there were many 'zero' values for capital gains
and losses. Because the 'income' variable in this dataset was a binary category (above or below $50K) the
capital gains and losses appeared to be a more interesting metric in gauging wealth for the individuals in
the Census. Therefore, we filtered the dataset to include only instances when there was a non-zero value for
capital gains or losses, then combined values of gains and losses to create a 'net' capital gain variable.

To create a data analysis pipeline, data filtering was performed with Rscripts executed via command line
arguments in the RStudio Terminal. The resulting filtered dataframe (augmented by filtering for instances
that included capital gains and losses) the data demographics were shifted to slightly older individuals
represented by more men than women. Below is a summary of the filtered dataframe.

```
##       age          workclass             fnlwgt          education
##  Min.   :17.00   Length:4231        Min.   :  19302   Length:4231
##  1st Qu.:34.00   Class :character   1st Qu.: 118346   Class :character
##  Median :42.00   Mode  :character   Median : 175669   Mode  :character
##  Mean   :43.18                      Mean   : 187152
##  3rd Qu.:51.00                      3rd Qu.: 234292
##  Max.   :90.00                      Max.   :1033222
##  education_num   martial_status      occupation         relationship
##  Min.   : 1.00   Length:4231        Length:4231        Length:4231
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :11.03
##  3rd Qu.:13.00
##  Max.   :16.00
##      race             sex             capital_gain    capital_loss
##  Length:4231        Length:4231        Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median : 3137   Median :   0.0
##                                        Mean   : 8293   Mean   : 671.9
##                                        3rd Qu.: 7688   3rd Qu.:1740.0
##                                        Max.   :99999   Max.   :4356.0
##  hours_per_week  native_country        label              net
##  Min.   : 1.00   Length:4231        Length:4231        Min.   :-4356
##  1st Qu.:40.00   Class :character   Class :character   1st Qu.:-1740
```

```
##  Median :40.00    Mode  :character    Mode  :character    Median : 3137
##  Mean   :43.42                                            Mean   : 7622
##  3rd Qu.:50.00                                            3rd Qu.: 7688
##  Max.   :99.00                                            Max.   :99999
```

EDA: Relationship between education attainment and annual net gain

As part of exploratory data analysis, we visualized some relationships among the data. Here we are visualizing the annual net gain across education levels.
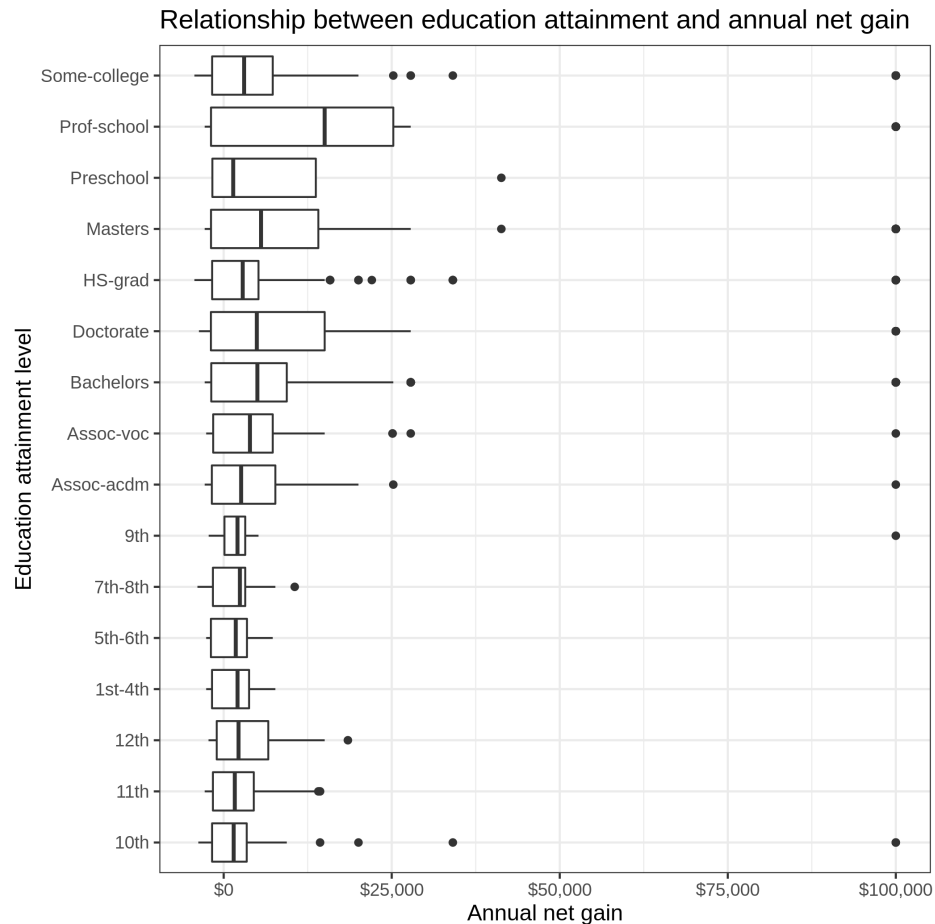
Figure 1: Figure 1: Boxplot of annual net capital gain across education levels

Figure 1 shows that minimal correlation between annual net gain and education attainment. However, there appeared to be a greater spread in annual net gain for individuals with at least a high school diploma, and persons with professional school education demonstrated the highest median in annual net gain.

EDA: Relationship between race, gender and annual net gain

We were interested to see if there were any visible patterns between annual net gain and individucal's race and or gender. In the inital exploratory data analysis, we examined the relationships between annual net gain across race and gender.
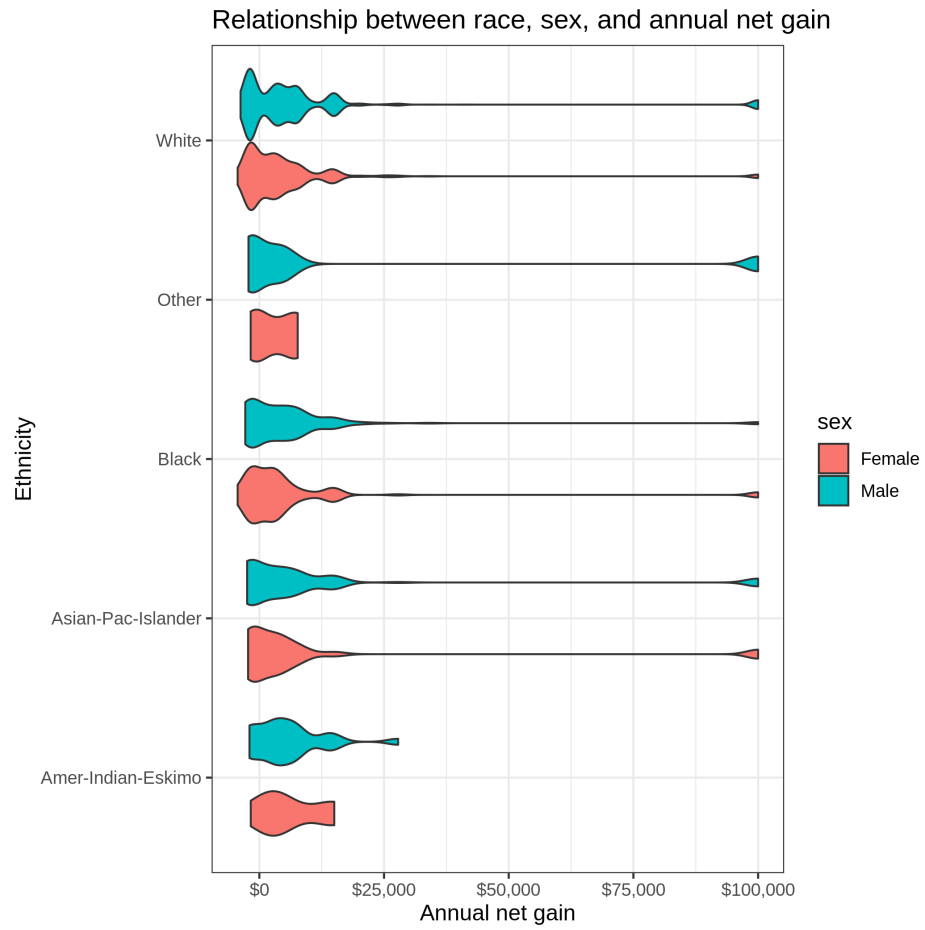
Figure 2: Figure 2: Violin plot of annual net capital gain by ethnicity and sex

There did not appear to be any significant differences in annual net gain between sex across all ethnic groups (Figure 2). Moreover, no obvious correlation between ethnicity and annual net gain was observed.

EDA: Correlation between work hours per week and annual net gain

We generated an additional script as part of the data analysis pipeline, to plot the relationship between annual net gain and hours worked per week. We categorized work hours per week as being short (under 25 hrs), medium (25-50 hrs/wk), long (50-75 hrs/wk) and very long (over 75 hrs/wk).
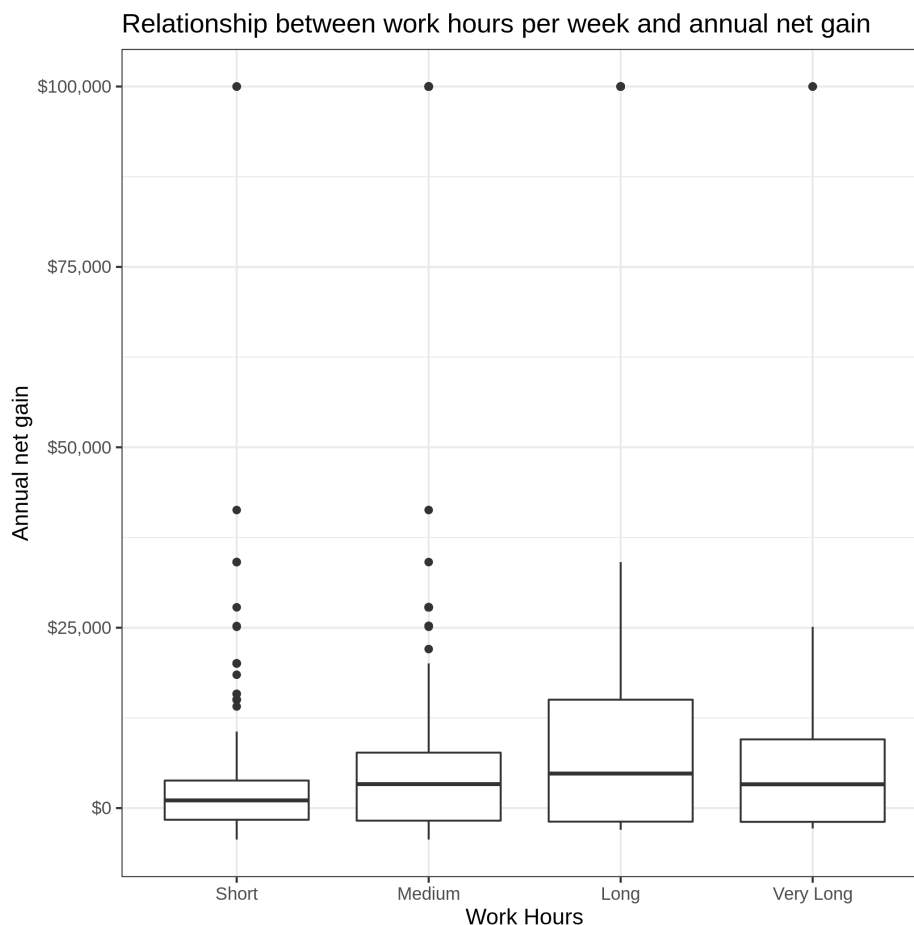

Figure 3: Figure 3: Plot of annual net capital gain by hours worked per week

From the above boxplot, there appears to be an increase in annual net gain from short to long work hours. However, the differences may not be significant because greater variance in annual net gain is observed for individuals with long work hours.

Linear Regression

A goal of this project was to generate a linear model for later use. The 'adult income' dataset did not have many linear relationships, so we isolated a section of data which showed a more direct relationship between variables in order to accomplish the scripting task of creating a linear model. Figure 4 shows scatter plots of the relationship between hours worked per week and age of worker, including the full dataset and a subsection with a near-linear trend.
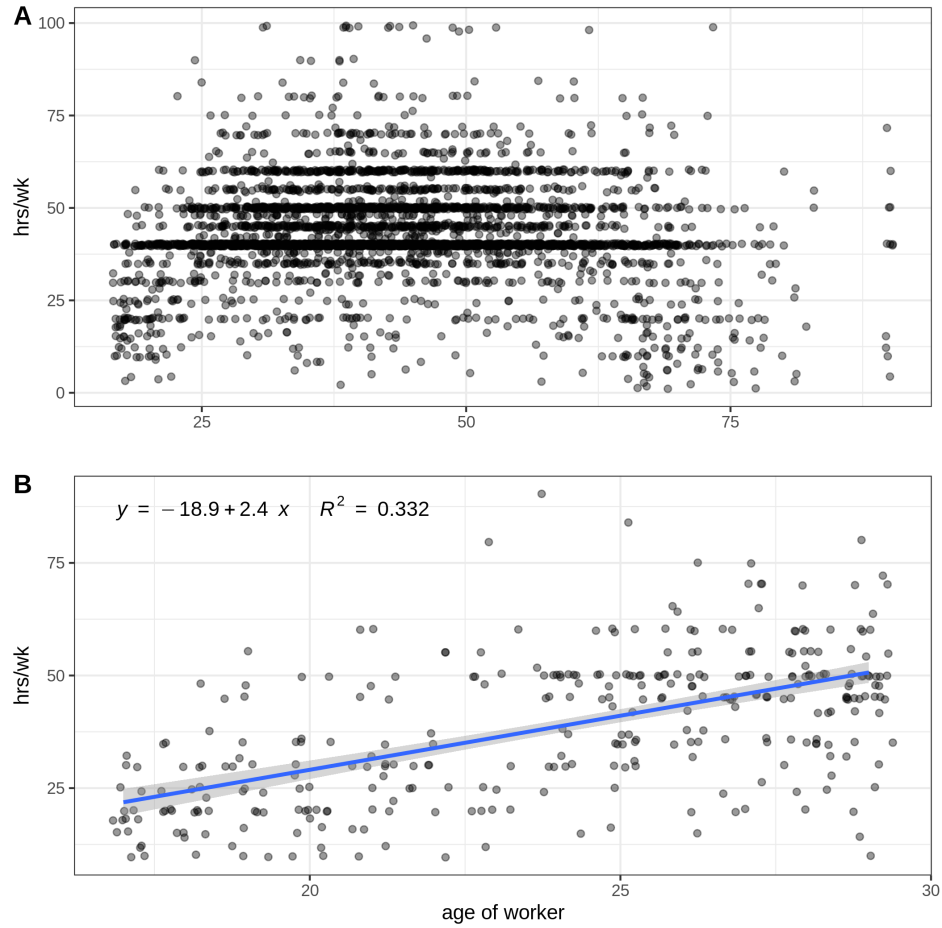
Figure 4: Figure 4: Hours worked per week by age where plot 'A' shows the full dataset and plot 'B' shows a more linear section of filtered data (for workers under 30)

The relationship between age and hours worked appeared to be loosely parabolic. Work hours increased with age until approximately age 30-50, at which time work hours stabilize before beginning to decrease with age up to approximately 80 years old (Figure 4-A). In order to perform a linear regression, we isolated the earlier part of this data, for workers under the age of 30 who worked more than 10 hours per week and not standard full time of 40 hours per week (Figure 4-B).

We wrote a script to generate a linear model based on the filtered data shown in Figure 4-B, as well as a script to generate that sub-plot as a stand-alone graphic (Figure 5).

hours worked per week for workers under 30 yrs of age

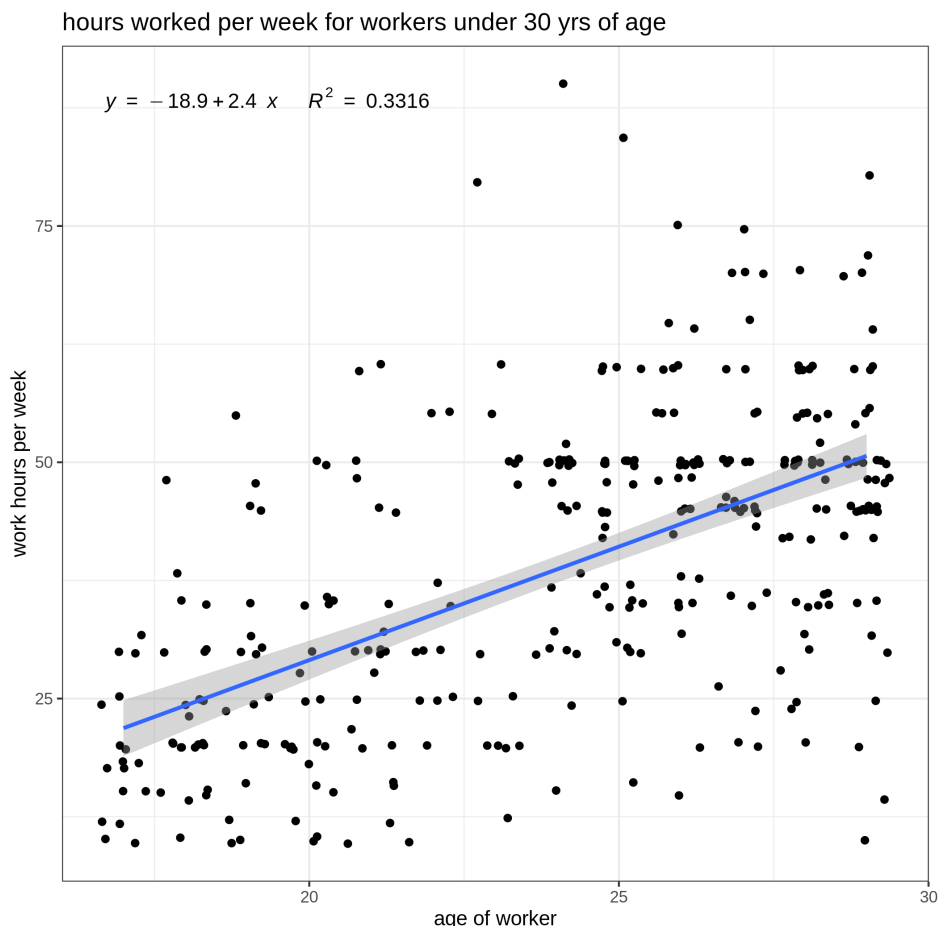

$y = -18.9 + 2.4\,x$ $R^2 = 0.3316$

Figure 5: Figure 5: Plot of linear regression model data for hours worked by those under age 30

The linear model based on this section of data is summarized here:

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.332         0.330  3.16      164. 8.29e-31     2  -854. 1714. 1726.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

References

United States Census Bureau, 2016. Income and Poverty, 'about income'. https://www.census.gov/topics/income-poverty/income/about.html

University of California Irvine, Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult.