1994 Census Bureau

Data Story

By Anthony Stachowitz

July 15, 2018

Introduction

There is a lot of conversation about the future of jobs in this country. People are worried about innovation and if it's going to take away jobs. Will self-driving trucks take away good blue-collar jobs? Are algorithms going to take over the financial sector? Should everyone major in robotics, so they can fix all the robots that are surely going to be doing all the factory jobs?

These are conversations that we are going to have to have as technology changes the employment landscape, But, are there things we can do to prepare for the future that has nothing to do with advances in tech? Are the generalities we can make about out type of employment and industry without worrying about specific job choices? Are there basic life decisions that influence how our future will unfold? Is there advice we can give our children that will, if followed, give them an advantage?

I have analyzed a data set extracted from the census bureau by Data Mining and Visualization Silicon Valley (via the UCI machine learning repository) to help answer some of these questions. The data includes two sets. The training set of 32,561 instances of data and the testing set includes 16,281 instances of date. I removed some variables to condense it to five categorical categories, four of the categories were used to predict the fifth variable.

The outcome I am predicting is whether you are going to make more than $50,000 a year. The categorical variables that I am using are type of employment, industry, marital status, and education level completed. Although you would assume going to school will help your chances, what is the best education level? Does your marital status have any effect on your income level, or is it more important then what industry you pick? Do you have a better chance to earn a living by owning your own business or should you try to get a government job… and what government, federal, state, local? These are some questions I will answer in capstone project.

The Data

The data I used comes from the 1994 census bureau by way of Data Mining and Visualization Silicon Graphics. It includes 48,842 instances split into a training set of 32,561 instances and a testing set of 16,281 instances.

Variable	Description
Age	Age of the person
WorkClass	Type S
FnlWgt	Weight based on socio-economic situation
Education	Highest education achieved
Education number	Number associated with education achieved
MaritalStatus	Marital Status or Divorce information
Occupation	Overall class of employment
Relationship	Current status in the family
Race	Race
Sex	Sex
CapitalGain	Capital Gains
CapitalLoss	Capital Losses
Hours per week	Hour worked per week
NativeCountry	Birth country
Salary	Salary amount

Table 2 – Data continued below

Work Class	Education-num	Education
Federal Government	1	Preschool
State Government	2	$1^{th} – 4^{th}$
Local Government	3	$5^{th} – 6^{th}$
Never worked	4	$7^{th} – 8^{th}$
Private	5	9^{th}
Self Employed Corp	6	10^{th}
Self Employed non- Corp	7	11^{th}
Without Pay (unemployed)	8	HS-grad
	9	Some-College
	10	Assoc-voc
	12	Assoc-acdm
	13	Bachelors
	14	Masters
	15	Doctorate
	16	Prof-school

Table 3 – Data continued below

MaritalStatus	Occupation	Relationship	Race
Married-AF-spouse	Adm-clerical	Husband	Black
Married-CIV-spouse	Armed-forces	Wife	White

Married-spouse-absent	Craft-repair	Unmarried	Amer-Indian-Eskimo
Divorced	Exec-managerial	Own-child	Asian-Pac-Islander
Separated	Farming-fishing	Not-in-family	Other
Never-married	Handlers-cleaners	Other-relative	
Widowed	Machine-op-inspct		
	Priv-house-serv		
	Prof-specialty		
	Sales		
	Tech-support		
	Transport-moving		
	Other-service		

Table 4

Sex	Native Country		Salary
Female	Cambodia	Italy	<=50k
Male	Canada	Jamaica	>50k
	China	Japan	
	Columbia	Laos	
	Cuba	Mexico	
	Dominican-Republic	Nicaragua	
	Ecuador	Outlying-US	
	El-Salvador	Peru	
	England	Philippines	
	France	Poland	
	Germany	Portugal	
	Greece	Puerto-Rico	
	Guatemala	Scotland	
	Haiti	South	
	Holand-Netherlands	Taiwan	
	Honduras	Thailand	
	Hong	Trinadad & Tobago	
	Hungary	United-States	
	India	Vietnam	
	Iran	Yugoslavia	
	Ireland		

Data Wrangling

The data I used comes from the 1994 census bureau by way of Data Mining and Visualization Silicon Graphics. It is 48,842 instances split into a training set of 32,561 instances and a testing set of 16,281 instances. I further decreased the number by removing categories and variables that I was not using and removing rows that were missing data in one or more of the variables. The factors I used are WorkingClass, Education, MaritalStatus, and Occupation.

```
knitr::opts_chunk$set(echo = TRUE)
```

```
library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(readxl)
```

1. Load dataset

This loads the individual datasets.

```
income_data <- read_excel("C:/Users/Anthony/Desktop/R studio projects/income/
income_data.xlsx", col_names = FALSE)
```

2. Add headers to the data set.

The data set does not have headers. I will add headers durring this step.

```
names(income_data) <- c("Age", "WorkClass", "FnlWgt", "Education", "Education
-num", "MaritalStatus", "Occupation", "Relationship", "Race", "Sex", "Capital
Gain", "CapitalLoss", "HoursPerWeek", "NativeCountry", "Salary")
```

3. Add column to convert less than $50,000 to 1 and more than $50,000 to 0.

```
income_data <- mutate(income_data, LessThen_50 = ifelse(grepl(">50K", Salary)
, 0, 1))
```

4. Convert characters to factors

Converted WorkClass, Education, MaritalStatus, and Occupation

```
income_data$Occupation <- as.factor(income_data$Occupation)
income_data$WorkClass <- as.factor(income_data$WorkClass)
income_data$Education <- as.factor(income_data$Education)
income_data$MaritalStatus <- as.factor(income_data$MaritalStatus)
```

5. Remove the data that is not being used.

Remove columns to make the easier to work with.

```
income_data <- income_data[, -c(1, 3, 5, 8:15)]
income_data
```

```
## # A tibble: 32,561 x 5
##    WorkClass        Education MaritalStatus         Occupation LessThen_50
##    <fct>            <fct>     <fct>                 <fct>             <dbl>
##  1 State-gov        Bachelors Never-married         Adm-cleri~            1
##  2 Self-emp-not-inc Bachelors Married-civ-spouse    Exec-mana~            1
##  3 Private          HS-grad   Divorced              Handlers-~            1
##  4 Private          11th      Married-civ-spouse    Handlers-~            1
##  5 Private          Bachelors Married-civ-spouse    Prof-spec~            1
##  6 Private          Masters   Married-civ-spouse    Exec-mana~            1
##  7 Private          9th       Married-spouse-absent Other-ser~            1
##  8 Self-emp-not-inc HS-grad   Married-civ-spouse    Exec-mana~            0
##  9 Private          Masters   Never-married         Prof-spec~            0
## 10 Private          Bachelors Married-civ-spouse    Exec-mana~            0
## # ... with 32,551 more rows
```

6. turn missing values into NA's

```
income_data[ income_data == "?" ] <- NA
```

7. Re-check for NA's

```
sapply(income_data,function(x) sum(is.na(x)))
```

```
##     WorkClass     Education MaritalStatus    Occupation   LessThen_50
##          1836             0             0          1843             0
```

8. Remove rows of data that had NA's

```
income_data <- income_data[complete.cases(income_data), ]
```

9. Recheck and data structure

```
sapply(income_data,function(x) sum(is.na(x)))
```

```
##     WorkClass     Education MaritalStatus    Occupation   LessThen_50
##             0             0             0             0             0
```

```
str(income_data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    30718 obs. of  5 variables:
##  $ WorkClass    : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ Education    : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ MaritalStatus: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ Occupation   : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ LessThen_50  : num  1 1 1 1 1 1 1 0 0 0 ...
```

Data Limitations

The limitations of this data set, in my opinion, include the year the data was taken from and the amount that was used for the monetary split point of $50,000.

The data was from 1994 which is over 20 years ago. In my opinion factors such as Marital Status and Occupation have had major changes over the period. For example, gay marriage was not legal or officially counted back in 1994, and just social attitudes towards relationships have changed. With the massive growth of the tech industry and the internet, there have obviously also been major changes in the type of occupations there are in the tech sector. For instance, there was no such thing as a Data Scientist in 1994… at least not officially.

The amount of $50,000 as a spilt point is a little low in 2018. I believe a split of $75,000 or $100,000 would have been better as a point that represents the separation of the middle class in America. This data is also being used to represent a possible future outcome. A $50,000 per year salary in 10 to 20 years will probably not represent an amount much over the poverty rate in some parts of the United States.

Exploratory Data Analysis

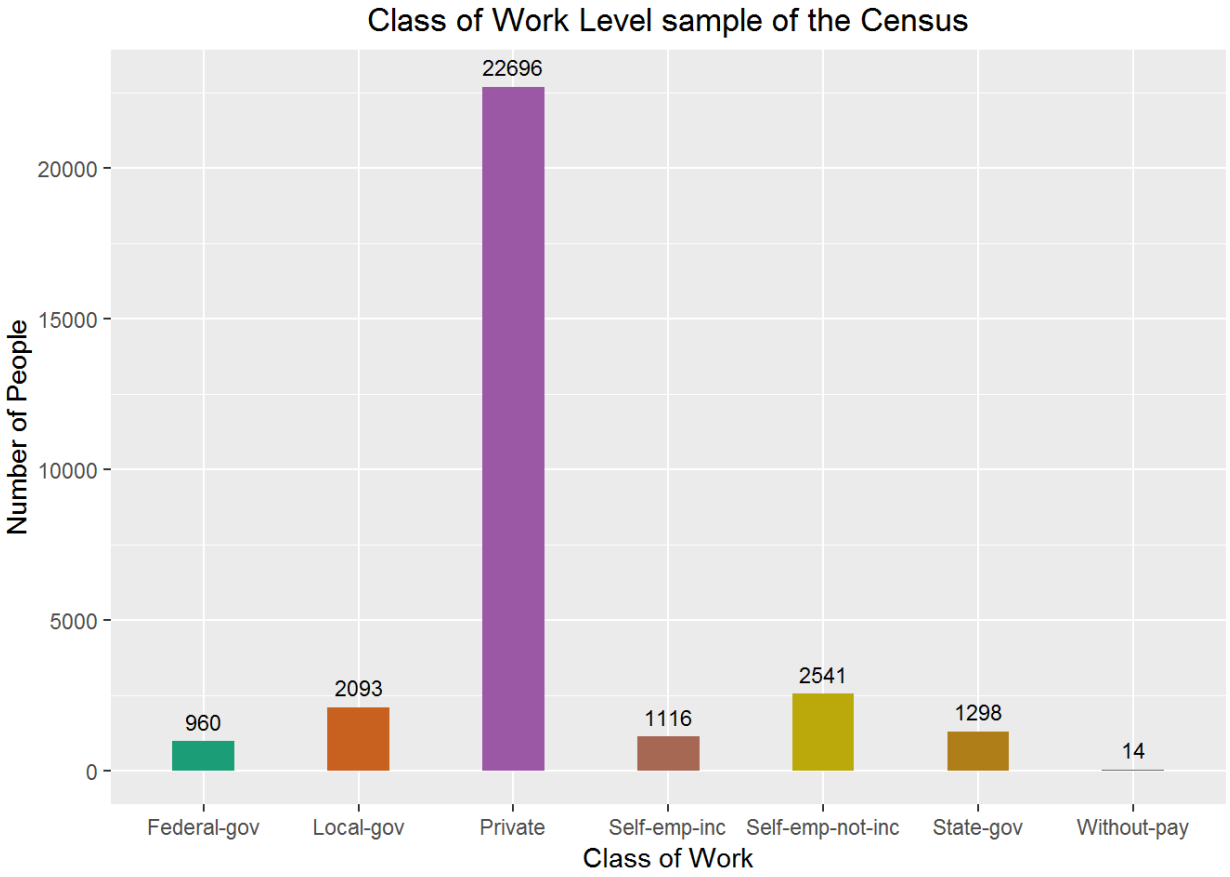
Working Class

```
work_graph <- income_data %>%
  group_by(WorkClass) %>%
```

```
  summarize(work_count = n()) %>%
  arrange(desc(work_count))


work_graph$WorkClass <- factor(work_graph$WorkClass, levels = work_graph$WorkC
lass[order(work_graph$WorkClass)])
colourCount = length(unique(work_graph$WorkClass))
fill_various <- colorRampPalette(brewer.pal(8, "Dark2"))


work_graph %>%
  filter(WorkClass != "NA") %>%
  ggplot(aes(x = WorkClass, y = work_count, fill = WorkClass, width = .4)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = work_count), size = 3, color = "black", hjust = 0.5, v
just = -0.7) +
  labs(x = "Class of Work", y = "Number of People", title = "Class of Work Lev
el sample of the Census") +
  theme(legend.position = "none", plot.title = element_text(hjust = .5)) +
  ylim(0, max(work_graph$work_count + 100)) +
  scale_fill_manual(values = fill_various(colourCount))
```

The class of work category is what type of entity you worked for. Was it the government? And then what level of the government… federal, state, or local. If you didn't work for the government, the choices were working for a company – private employee. Or, working for yourself, either owning a private business or being an equity holder in a corporate structure. The last option was being unemployed, or without pay. The bar graph clearly shows that most people work for an employer by the 22,696 people that are classified as private employees on the table.

This is not a surprise as I would expect most people to work for an employee. I am a little surprised that the number of federal employees is the least populated variable among people with a job. It will be interesting to see which variable shows the greatest likelihood of earning more $50,000 a year.



Occupation

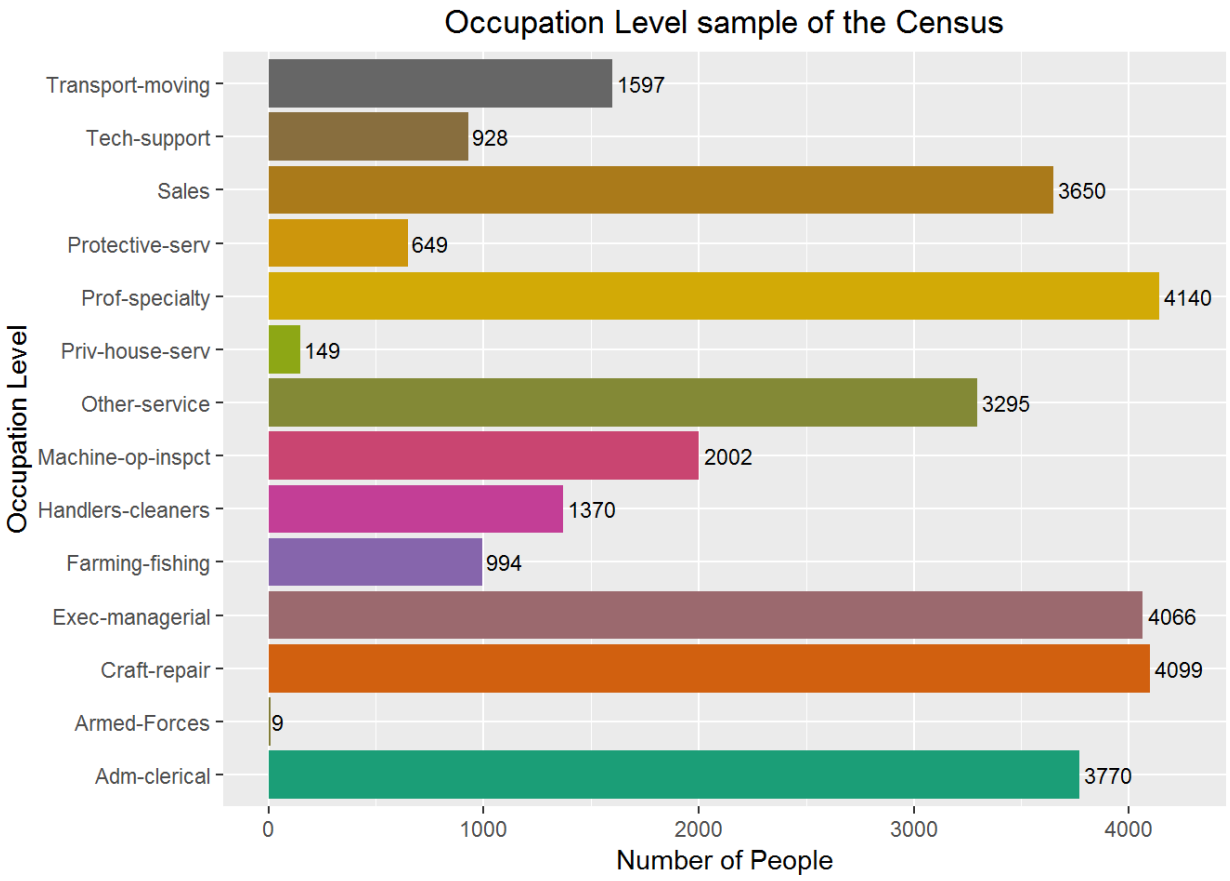
```
Occup_graph <- income_data %>%
  group_by(Occupation) %>%
  summarize(Occupation_count = n()) %>%
  arrange(desc(Occupation_count))


Occup_graph$Occupation <- factor(Occup_graph$Occupation, levels = Occup_graph$
Occupation[order(Occup_graph$Occupation)])
colourCount = length(unique(Occup_graph$Occupation))
fill_various <- colorRampPalette(brewer.pal(8, "Dark2"))
```

```
Occup_graph %>%
  filter(Occupation != "NA") %>%
  ggplot(aes(x = Occupation, y = Occupation_count, fill = Occupation)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  geom_text(aes(label = Occupation_count), size = 3, color = "black", hjust =
-0.1) +
  labs(x = "Occupation Level", y = "Number of People", title = "Occupation Lev
el sample of the Census") +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5)) +
  ylim(0, max(Occup_graph$Occupation_count + 100)) +
  scale_fill_manual(values = fill_various(colourCount))
```

The occupation category has fourteen variables. They are set up so most people can fit their specific job into a variable. This is probably the most subjected category of the four that I chose. It is also probably the most undependable category because of that reason. I struggled a little with the decision to include this category but it the end I figured that in this case more information would not cloud the results and I am individually measuring each of the variables in each category against income, so the occupation category will not have an effect on the relevance of the other variables.

The biggest surprise to me is that sales is not the largest variable. Although, it is in the top four. I am also a little surprised that transportation doesn't represent a larger portion of the total.

Unfortunately, while looking at this data the issue that screams out at me is the amount of errors there probably is in this category. You can easily fit a certain job in more than on variable. For instance, you can work for the armed forces and have almost any job fit into any one of the other variables. You can work in the transportation business answering phones to get clients who need to be move across country. And, in that case, you would work in transportation, sales, and adm-clerical. This data set needs to be further parsed to be able to depend on the results of my final analysis. Although, the final analysis will provide a new starting point for a more in-depth analysis in the future.



Education Level

```
library("RColorBrewer")
library("ggplot2")


ed_graph <- income_data %>%
  group_by(Education) %>%
  summarize(education_count = n()) %>%
  arrange(desc(education_count))
```

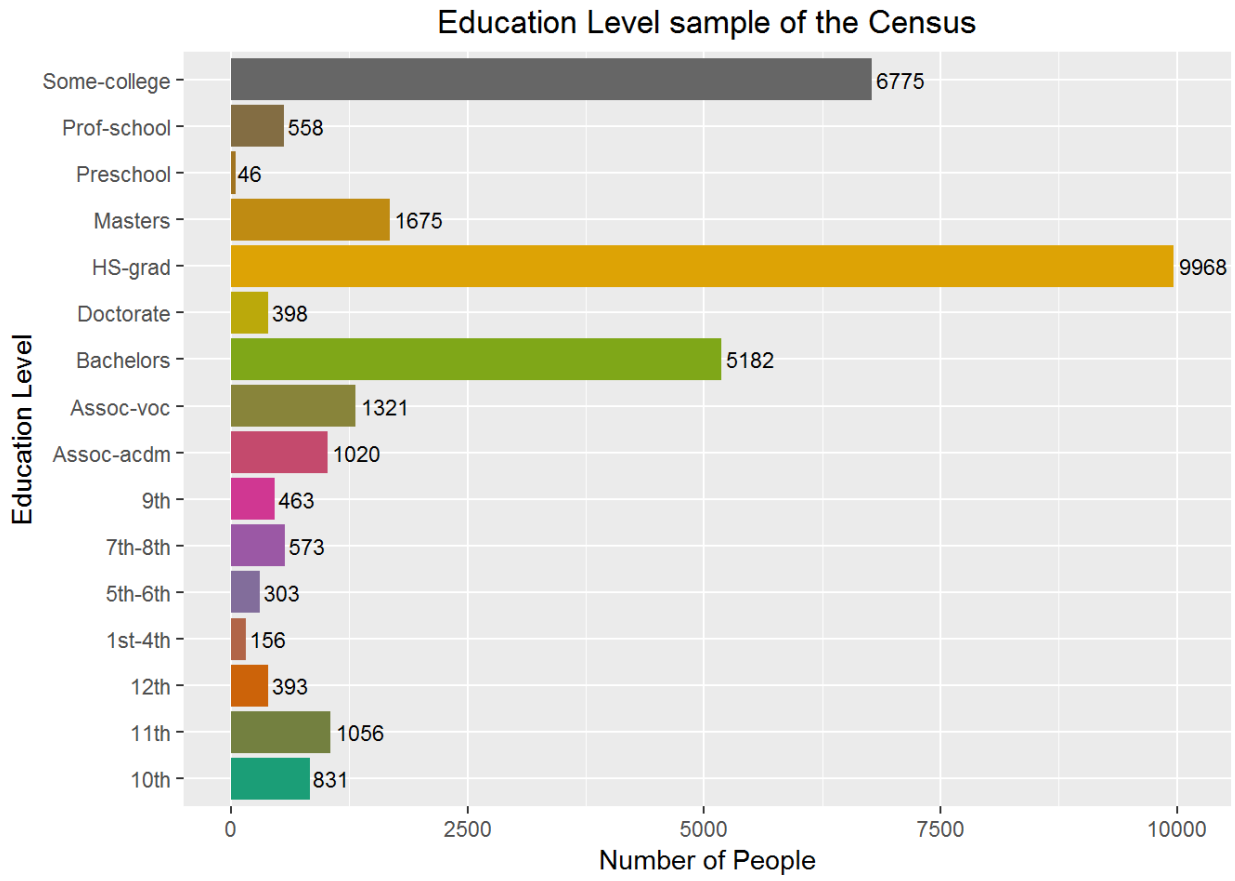
```
ed_graph$Education <- factor(ed_graph$Education, levels = ed_graph$Education[o
rder(ed_graph$Education)])

colourCount = length(unique(ed_graph$Education))

fill_various <- colorRampPalette(brewer.pal(8, "Dark2"))


ed_graph %>%

  filter(Education != "NA") %>%

  ggplot(aes(x = Education, y = education_count, fill = Education)) +

  geom_bar(stat = "identity") +

  coord_flip() +

  geom_text(aes(label = education_count), size = 3, color = "black", hjust = -
0.1) +

  labs(x = "Education Level", y = "Number of People", title = "Education Level
sample of the Census") +

  theme(legend.position = "none", plot.title = element_text(hjust = 0.5)) +

  ylim(0, max(ed_graph$education_count + 100)) +

  scale_fill_manual(values = fill_various(colourCount))
```

The education category has 16 variables. They range from pre-school to doctorate. The lower grades are grouped by two or four until you get to the 9th grade and each level represents one individual year after that.

This category has the least surprising results. The highest variable is high school which stands at 9,968. I would assume that most people finish high school for a few reasons, the most promin ate is because it is against the law to not go to school in some fashion until you are eighteen years old. Most people graduate high school by eighteen.

This is also the category that I think will have the most obvious and predictable relationship to salary. Generally, most people believe the more education you obtain the higher the probability of achieving financial success over your lifetime. It will be interesting to see if the results go up proportionately with the higher level of education.

One of the problems with this data set is the level of salary that was available in this data set. I would like to know if the highest level of education, the doctorate level, has a dip in salary because many doctorate level students go on the teach and do research. But, because most professors would earn over $50,000, it is probably not possible to check that hypothesis with this data set. If the salary variable was below or above $100,000, I think you could better measure that theory.



Marital Status

```
marital_graph <- income_data %>%
  group_by(MaritalStatus) %>%
  summarize(marital_count = n()) %>%
  arrange(desc(marital_count))


marital_graph$MaritalStatus <- factor(marital_graph$MaritalStatus, levels = m
arital_graph$MaritalStatus[order(marital_graph$MaritalStatus)])
colourCount = length(unique(marital_graph$MaritalStatus))
```

```
fill_various <- colorRampPalette(brewer.pal(8, "Dark2"))


marital_graph %>%
  filter(MaritalStatus != "NA") %>%
  ggplot(aes(x = MaritalStatus, y = marital_count, fill = MaritalStatus, widt
h = .4)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  geom_text(aes(label = marital_count), size = 3, color = "black", hjust = 0,
vjust = -1.8) +
  labs(x = "Marital Status", y = "Number of People", title = "Marital Status
Level sample of the Census") +
  theme(legend.position = "none", plot.title = element_text(hjust = .5)) +
  ylim(0, max(marital_graph$marital_count + 100)) +
  scale_fill_manual(values = fill_various(colourCount))
```

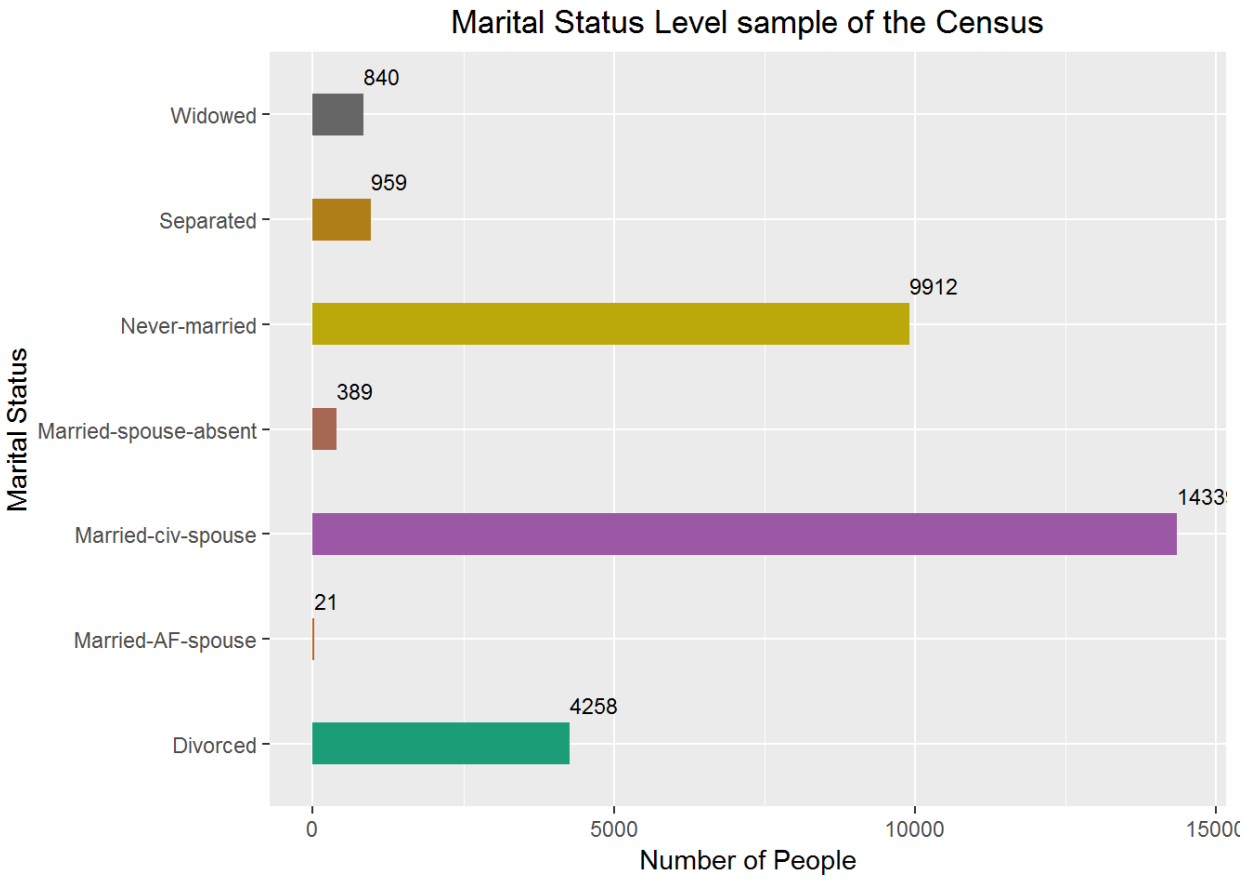
Marital status has six variables. They range from being married, divorced, and never married, with different classifications within those. This is the most interesting category for the question I am answering with this set of data.

Most people wouldn't look to a person's marriage status as an indicator to what type of salary you would earn. But marriage is such an emotionally charged union, as most types of serious relationships are. Emotion, mental health and general wellbeing influences your day to day life, so you would probably expect that to spill over into your work life.

There are two problems with this data that jump out at me. The first is that according to this census from 1994, the divorce rate is low. The second is that this data doesn't include civil unions or gay marriage.

The divorce rate now is generally understood to be about 50% and the rate give visually by this data looks to be around 23%. At the least, this data may not relate correctly to the data for 2018 and would have to be updated in another study.

Gay marriage wasn't legal until recently so there was no official information on those unions. They were not included in this data. Therefore, information involving gay marriage cannot in any way be estimated. It has a negative effect on the other results in this set because it will change the percentages of the results by changing the number of every variable except the never-married variable.



Conclusion

This data set from the Census and Data Mining and Visualization Silicon Graphics provides a good example of the way Logistic regression can help look at data in a different way. Our society is always looking for ways to explain, or formulas to teach, a way to better ourselves. Modern data science gives us a way to take data that is not obvious and explore the relationship it has with level of salary someone makes. We are able to compare something obvious, like the level of education, to something less obvious, like marriage status, and to see what has more of an effect.

I am glad we are exploring this data and believe this may be information we can pass on to our children to give them more knowledge on how life decisions effect their future. The data I used was from 1994. I would like to look at more recent data and apply some of the same techniques and compare them and to the data we will find in this study.