IPUMS Economic data

MATH 301 Data Confidentiality

Henrik Olsson

February 18, 2020

```
library(readr)
library(LearnBayes)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(imputeTS)
ipumsdata<- read.csv("usa_00010.csv")
head(ipumsdata, 10)
```

	Y...	SAM...	SERIAL	CBSERIAL	...	R...	RA...	HCOVA...	SCH...	▸
	<int>	<int>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>	
1	2018	201801	1	2.01801e+12	2	1	100	2	2	

	Y…	SAM…	SERIAL	CBSERIAL	…	R…	RA…	HCOVA…	SCH…
	<int>	<int>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
2	2018	201801	2	2.01801e+12	2	2	200	2	2
3	2018	201801	3	2.01801e+12	1	1	100	1	1
4	2018	201801	4	2.01801e+12	1	1	100	1	1
5	2018	201801	5	2.01801e+12	2	1	100	2	1
6	2018	201801	6	2.01801e+12	2	1	100	2	1
7	2018	201801	7	2.01801e+12	2	1	100	2	1
8	2018	201801	8	2.01801e+12	1	1	100	1	1
9	2018	201801	9	2.01801e+12	1	1	100	2	1
10	2018	201801	10	2.01801e+12	2	1	100	2	2

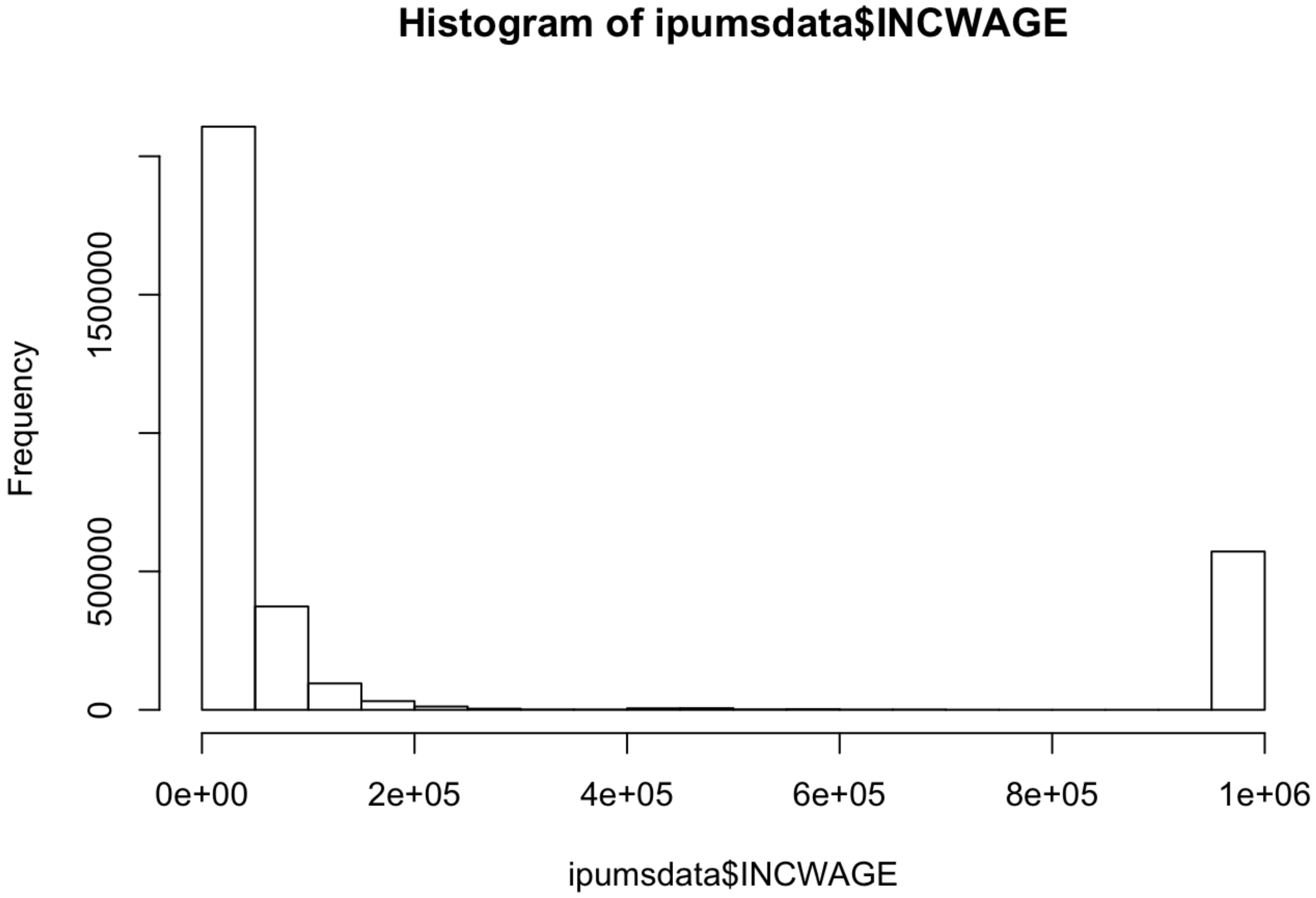
1-10 of 10 rows | 1-10 of 15 columns

The IPUMS data holds U.S. census microdata for social, economic, and health research. The dataset only looks at the year 2018. I extracted a total of 12 variables. There a total of approximately 3 million data points. Year, sample, IPUMS sample identifier, household serial number, and Census Buereau household serial number will not be modified and are in the dataset for reference. Sex (binary), race (categorical), healthcare coverage (binary), school attendance (categorical), field of degree (categorical), employment status (categorical), and total salary/wage income (continuous). There are also variables labeled RACED, DEFIELDD, and EMPSTATD, which are more detailed versions of race, degree of field, and empoyment status. We will choose to ignore these variables in this project.

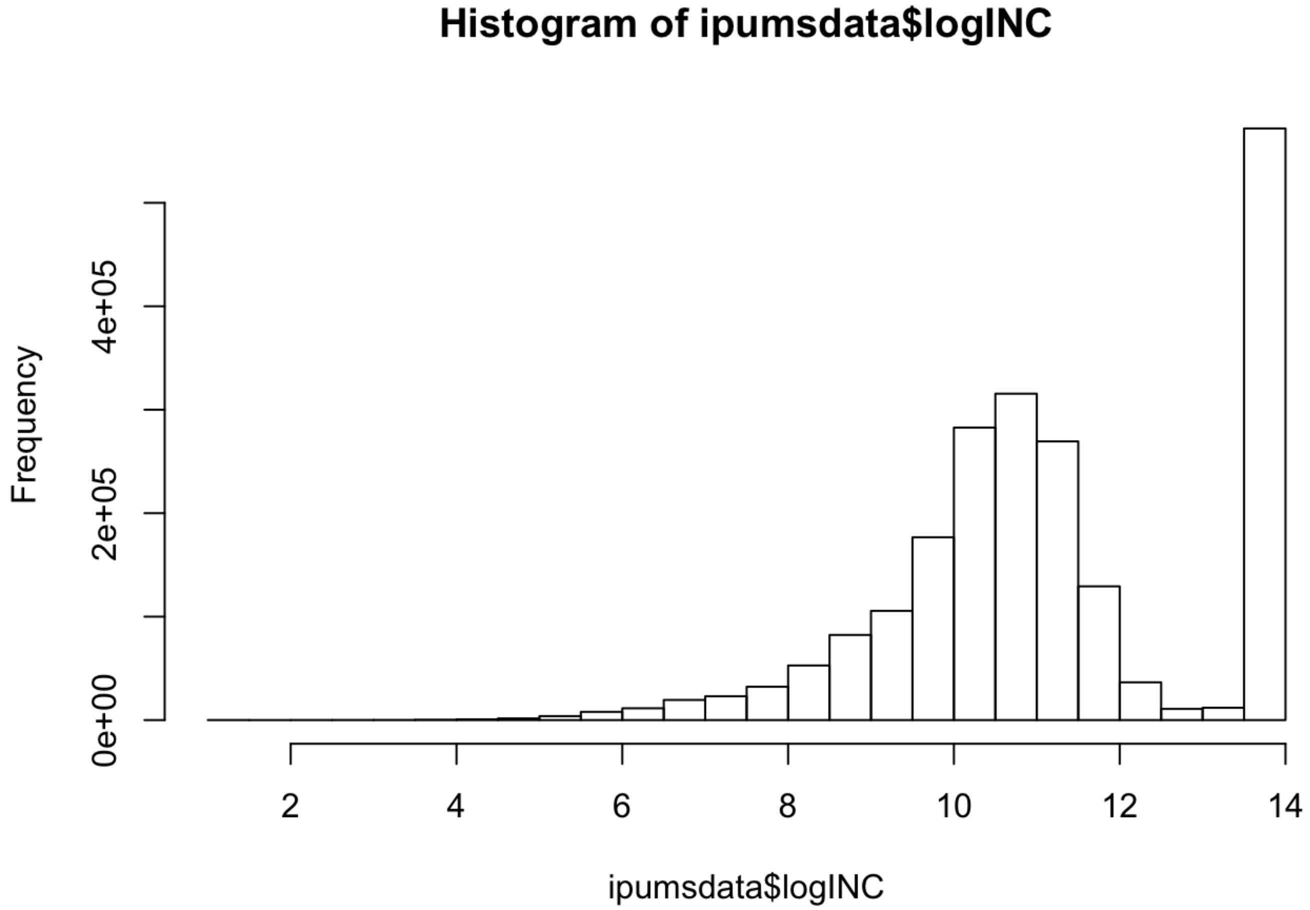
We will look at potential disclosure risks in the original data. First, we will clean up the data and remove any observations that are missing.

Next, we will assess which variables are the most sensitive.

```r
## Salary income
ipumsdata$logINC <-log(ipumsdata$INCWAGE)
hist(ipumsdata$INCWAGE)
```

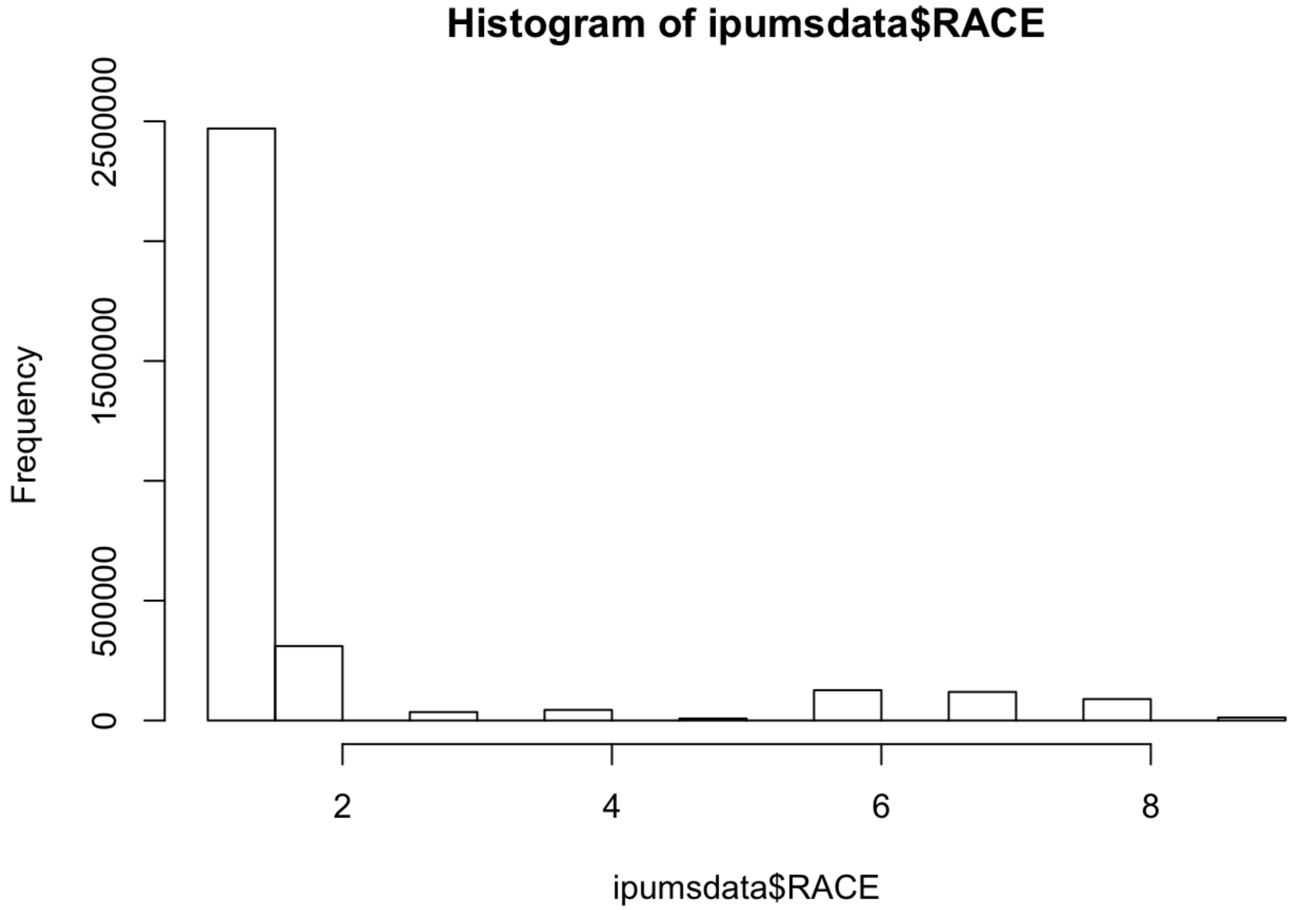


```
hist(ipumsdata$logINC)
```

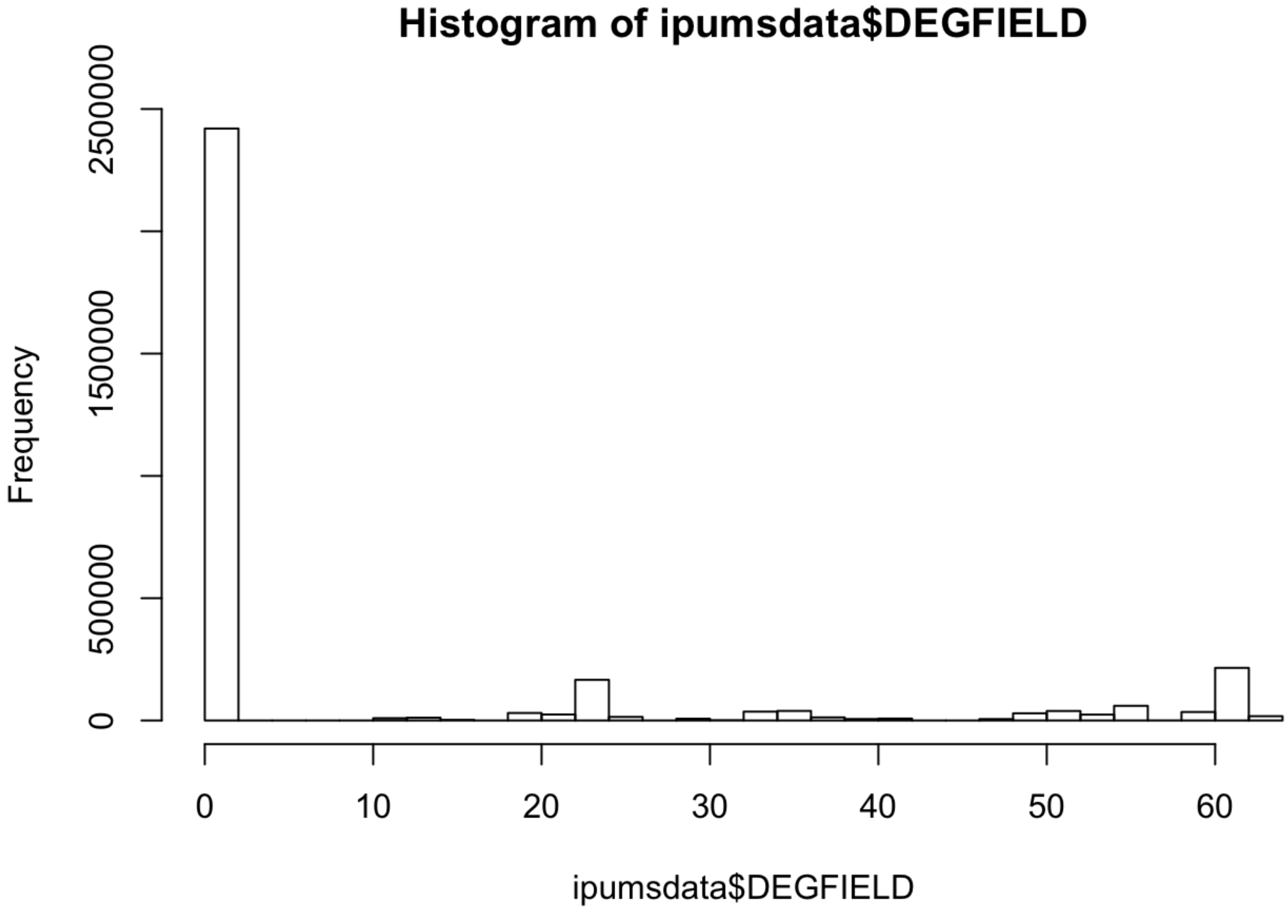


The most sensitive variable is Income, which is a continuous variable. If an intruder were to know one’s income then they can obtain the person’s information with much greater probability than if they had access to another variable.

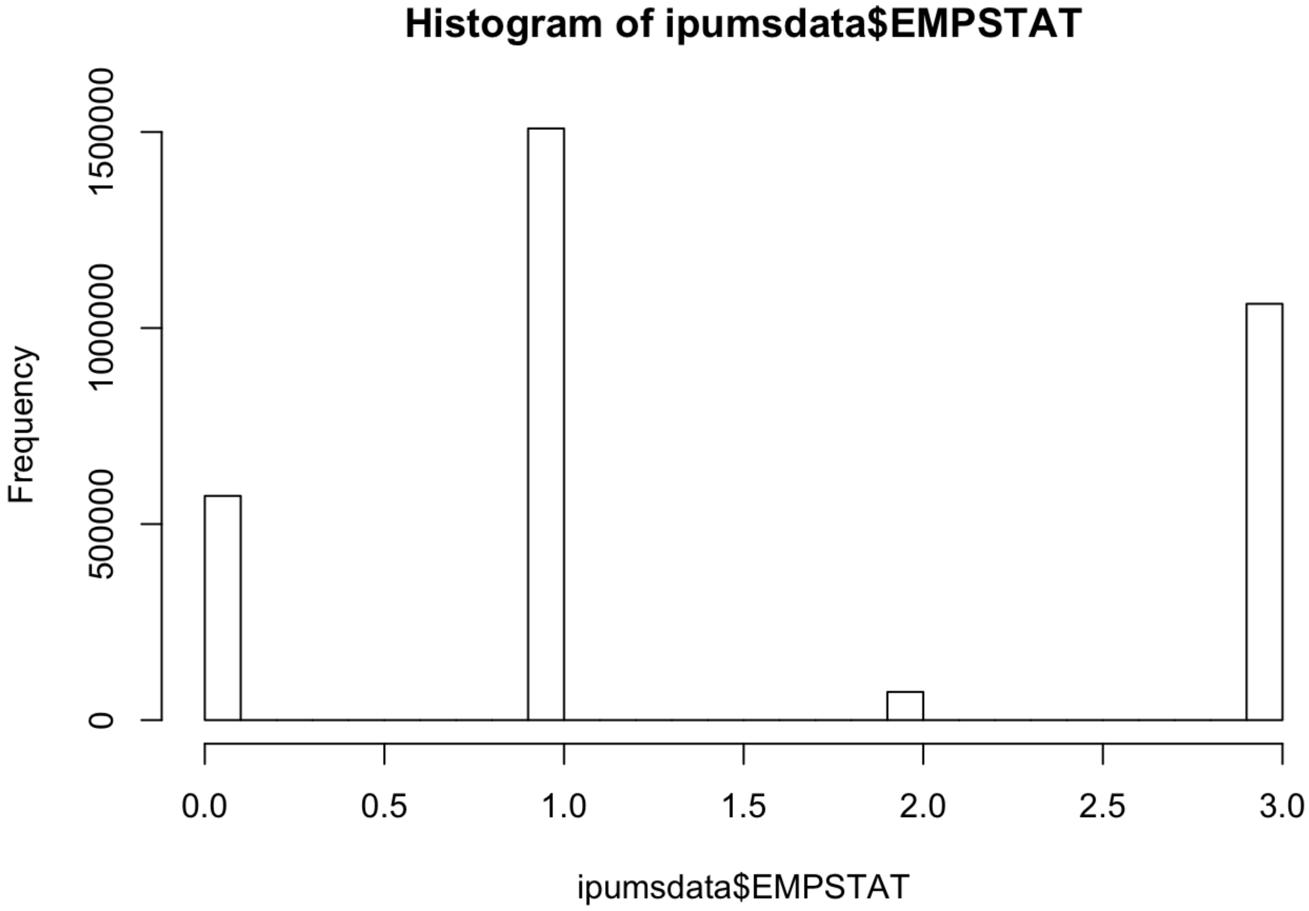
```
## Race
hist(ipumsdata$RACE)
```



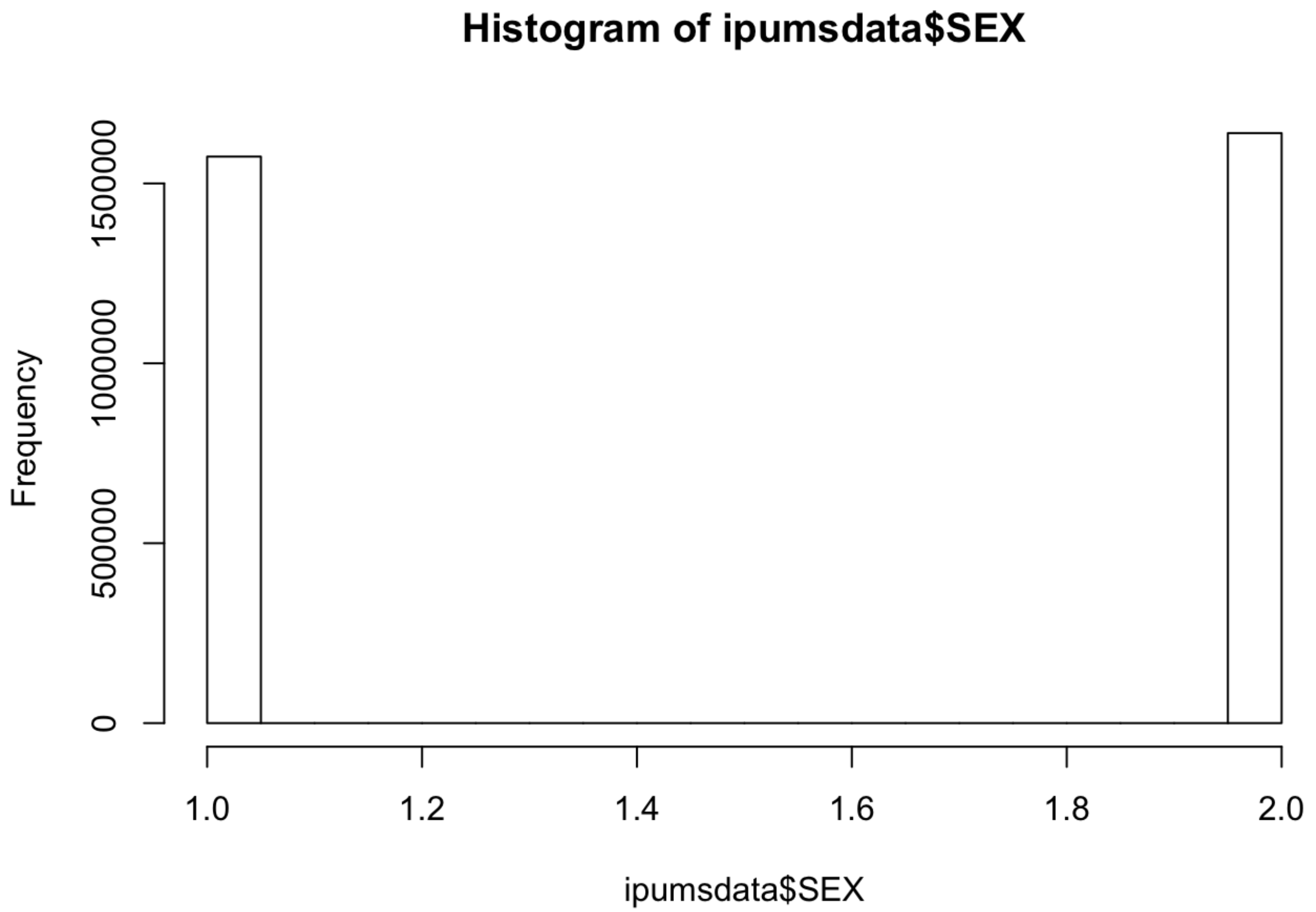
```
## Field of degree
hist(ipumsdata$DEGFIELD)
```



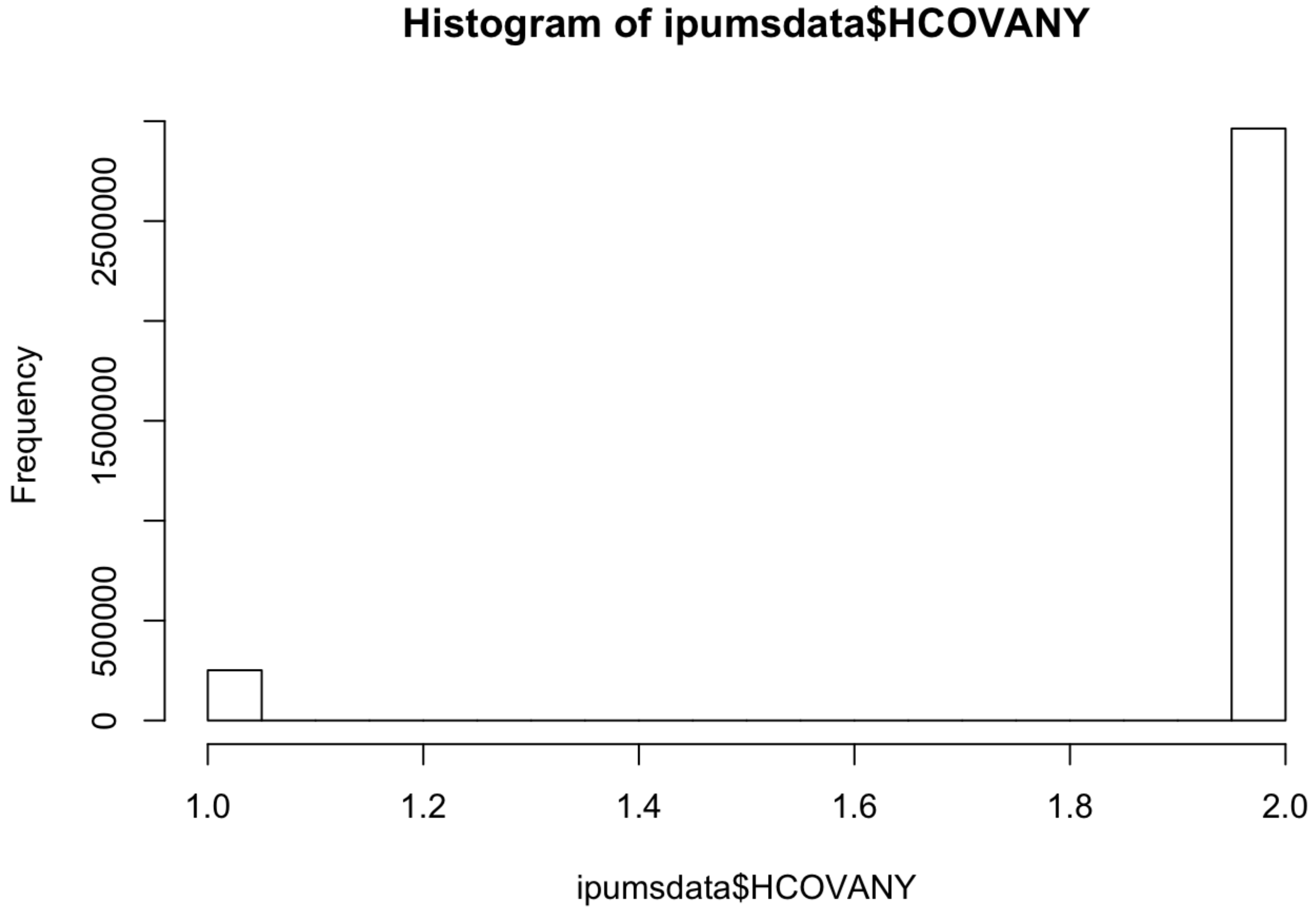
```
## Employment Status
hist(ipumsdata$EMPSTAT)
```



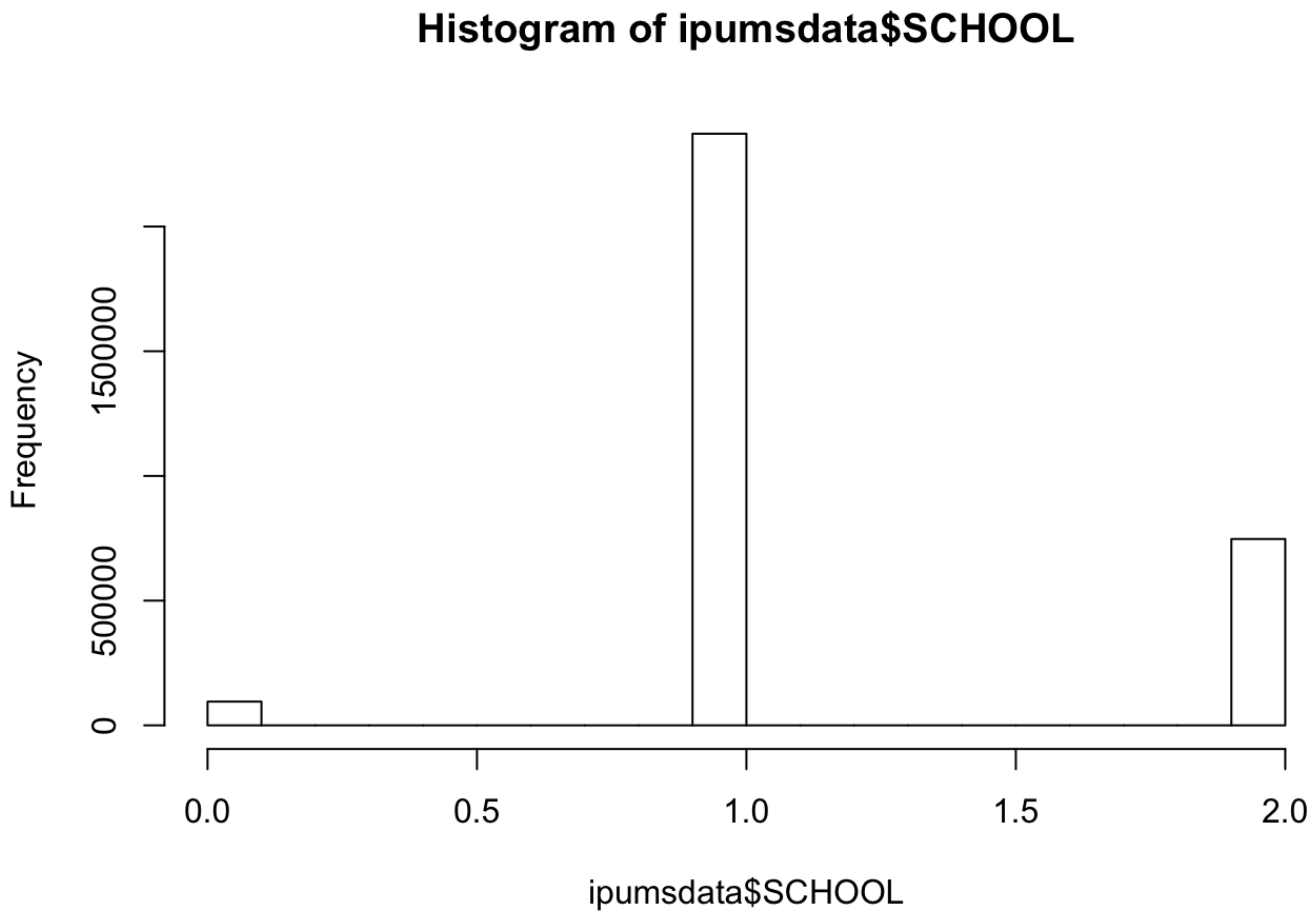
```
## Sex
hist(ipumsdata$SEX)
```



```
## Healthcare coverage
hist(ipumsdata$HCOVANY)
```



```
## School attendance
hist(ipumsdata$SCHOOL)
```



Since field of degree contains 64 categories, this variable is the second most sensitive variable. The next most sensitive is Race, with a total of 9 categories.

The least sensitive varaible is Sex due to the bimodal distribution shape of the binary variable.

Type 1: Identification disclosure

```
## SAMPLE
NeighborSet <- ipumsdata %>%
  filter(SEX == 1 & RACE == 1 & SCHOOL == 1)
dim(NeighborSet)
```

```
## [1] 915091     15
```

```
## Random guess which one is your neighbor- risk as a probability
1/dim(NeighborSet)[1]
```

```
## [1] 1.092787e-06
```

The biggest potential disclosure risk will be with regards to Income. Thus, we will test to see what relationships we want to preserve. The income is top coded at the 99.5th percentile in State.

```
## Income at $401,000
NeighborSet <- ipumsdata %>%
  filter(INCWAGE == 401000)
dim(NeighborSet)
```

```
## [1]  1 15
```

```
## Random guess which one is your neighbor- risk as a probability
1/dim(NeighborSet)[1]
```

```
## [1] 1
```

```
## Income at $402,000
NeighborSet2 <- ipumsdata %>%
  filter(INCWAGE == 402000)
dim(NeighborSet2)
```

```
## [1] 119  15
```

```
## Random guess which one is your neighbor- risk as a probability
1/dim(NeighborSet2)[1]
```

```
## [1] 0.008403361
```

By randomizing potential salary income values, an intruder can determine the individual if they knew their income was $401,000. However, if the intruder was $1,000 off by mistake, they would have to search through a total of 119 possible data points, with a 0.8% probability of finding the right individual.

```
## Relationship between Male sex, Japanese race, No Health insurance, Unem
ployed, and Not in school
NeighborSet <- ipumsdata %>%
  filter(SEX == 1 & RACE == 5 & SCHOOL == 1 & HCOVANY == 1 & EMPSTAT == 2)
dim(NeighborSet)
```

```
## [1] 10 15
```

```
## Random guess which one is your neighbor- risk as a probability
1/dim(NeighborSet)[1]
```

```
## [1] 0.1
```

If an intruder were to know that the individual is a Japanese male who is unemployed, not insured, and not in school. Then they would have a 10% probability of identifying the person. This relationship does not include knowledge of the two most sensitive variables, income and field of degree.

```
## Degree of field: Nuclear and Biological Technologies and Native America
n race
NeighborSet <- ipumsdata %>%
  filter(DEGFIELD == 51 & RACE == 3)
dim(NeighborSet)
```

```
## [1]  1 15
```

```
## Random guess which one is your neighbor- risk as a probability
1/dim(NeighborSet)[1]
```

```
## [1] 1
```

If an intruder were to know that an individual's degree of field is in Nuclear and Biological Technologies, and that they were from a Native American descent then they would be able to identify the individual.

```
## Henrik profile disclosure risk
HenrikSet <- ipumsdata %>%
  filter(SEX == 1 & DEGFIELD == 37 & RACE == 8 & EMPSTAT == 1 & HCOVANY ==
2 & SCHOOL == 2)
dim(HenrikSet)
```

```
## [1] 18 15
## Random guess which one is your neighbor- risk as a probability
1/dim(HenrikSet)[1]
```

```
## [1] 0.05555556
```

For fun, lets assume I was in the IPUMS dataset. Of the over 3 million samples, a total of 18 individuals have the same characteristic as I do. This is excluding income, which will certainly give away the individual’s whole profile.

Type 2: Attribute disclosure

An intruder could correctly infer the true value of one unknown variable/attribute of an individual. We are not looking in-depth at attribute disclosure, but just starting the conversation.

```
## For a uniquely identified person:
NeighborSet %>% count(HCOVANY) %>% group_by(HCOVANY)
```

HCOVANY	n
<int>	<int>
2	1

1 row