DATA 606 Data Project

Chunjie Nan

5/10/2020

Contents

Introduction

This project is to investigate if the "American Dream" is still exists for foreigners. In this study, I use the total income of a person as the standard of success, even though there are many ways to present a "success". In this observational study, I want to find out the relationships between the income and how they make different toward non-native Americans. Hypothesis testing is done to determine if there is siggnificant difference in income between native and non-natives.

Data

The data is collected by the IPUMS CPS as part of the Current Population Survey.The Current Population Survey (CPS) database of October 2019, where we obtained demographic and labor information for the US. The raw database has exactly 298,831 observations that contains several variables. From the CPS database, we only considered selecting people whose ages range between 18 and 65 years old, those who are in the labor force, those that work between 40 and 126 hours per week, and those whose wage and income total are greater than zero. With these restrictions, we obtained 55,567 observations and we then chose to work with people who are currently in the labor force.

**INCWAGE: Continue variable with Monthly wage

**AGE:Continue variable with age of people in the database

**FAMSIZE:Continue variable with the family size at home

**NCHILD:Continue variable with numbers of children at home

**educ_sc:Dichotomy variable where "1" are people who have some college and "0" no

**educ_bd:Dichotomy variable where "1" are people who have a bachelor's degree and "0" no

**educ_hs:Dichotomy variable where "1" are people who have high school and "0" no

**educ_md:Dichotomy variable where "1" are people who have master's degree and "0" no

**educ_phd:Dichotomy variable where “1” are people who have a PhD degree and “0” no

**educ_psd:Dichotomy variable where “1” are people who have a Professional School degree and “0” no

**Single:Dichotomy variable where “1” are people whose marital status is single and widowed and “0” no

**Married:Dichotomy variable where “1” are people whose marital status is married and “0” no

**Divorce:Dichotomy variable where “1” are people whose marital status is divorced and “0” no

**Widowed:Dichotomy variable where “1” are people whose marital status is widowed and “0” no

**White:Dichotomy variable where “1” are people whose race is White and “0” no

**Hispanic:Dichotomy variable where “1” are people whose race is Hispanic and “0” no

**Asian:Dichotomy variable where “1” are people whose race is Asian and “0” no

**Black:Dichotomy variable where “1” are people whose race is black and “0” no

**Veteran:Dichotomy variable where “1” are people whose is war veteran and “0” no

**Yearinus:Continue variable that indicate the years that people stay in US

**Usborn:Dichotomy variable where “1” are people who is born in US and “0” people who is Naturalized or Legal residents

Exploratary data analysis

```
# load data

if (!require("ipumsr")) stop("Reading IPUMS data into R requires the ipumsr package. It can be installe
```

```
## Loading required package: ipumsr
```

```
ddi<- read_ipums_ddi("cps_00003.xml")
ddi
```

```
## An IPUMS DDI for IPUMS-CPS with 39 variables
## Extract 'cps_00003.dat' created on 2019-11-30
## User notes:  User-provided description: Revision of (econometric's project)
##
##   This extract is a revision of the user's previous extract, number 1.
```

```
data <- read_ipums_micro(ddi)
```

```
## Use of data from IPUMS-CPS is subject to conditions including that users should
## cite the data appropriately. Use command `ipums_conditions()` for more details.
```

```
attach(data)
use_variable<-(AGE>=18) & (AGE<=65)&(LABFORCE=2)&(UHRSWORKT>=40 & UHRSWORKT<=126)&(INCWAGE>0)&(INCTOT>0
data_project<-subset(data,use_variable)
detach()
attach(data_project)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     LABFORCE
```

```
data_project$usborn <- ifelse(data_project$CITIZEN<=3,1,0)
data_project$naturalized <- ifelse(data_project$CITIZEN==4,1,0)
data_project$notcitizen <- ifelse(data_project$CITIZEN==5,1,0)
data_project$white<- ifelse(data_project$RACE ==100,1,0)
data_project$hispanic<-ifelse(data_project$HISPAN >0,1,0)
data_project$black<- ifelse(data_project$RACE ==200,1,0)
data_project$asian<- ifelse(data_project$ASIAN <99,1,0)
data_project$educ_ss <- ifelse(data_project$EDUC>=10 & data_project$EDUC<=72, 1,0) #ss = some school
data_project$educ_hs <- ifelse(data_project$EDUC==73, 1,0) #hs = high school
data_project$educ_sc <- ifelse(data_project$EDUC==81, 1,0) #sc = some college
data_project$educ_ad <- ifelse(data_project$EDUC<=92 & data_project$EDUC>=91, 1,0) #ad = associates deg
data_project$educ_bd <- ifelse(data_project$EDUC==111, 1,0) #ba = bachelors degree
data_project$educ_md <- ifelse(data_project$EDUC==123, 1,0) #md = masters degree
data_project$educ_psd <- ifelse(data_project$EDUC==124, 1,0) #pd = professional school degree
data_project$educ_phd <- ifelse(data_project$EDUC==125, 1,0) #phd = PhD degree
data_project$single <- ifelse(data_project$MARST==6, 1,0) #single
data_project$married <- ifelse(data_project$MARST>=1 & data_project$MARST<=2, 1,0) #married
data_project$divorce <- ifelse(data_project$MARST>=3 & data_project$MARST<=4, 1,0) #divorced
data_project$widowed <- ifelse(data_project$MARST==5, 1, 0) #widowed
data_project$veteran <- ifelse(data_project$VETSTAT==2, 1, 0) #veteran
data_project$yearsinus <- ifelse(data_project$YRIMMIG>0, 2019 - data_project$YRIMMIG, 0)
```

```
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(haven)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(ggplot2)
library(magrittr)
library(AER)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:car':
##
##     logit

## The following object is masked from 'package:Hmisc':
##
##     describe

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

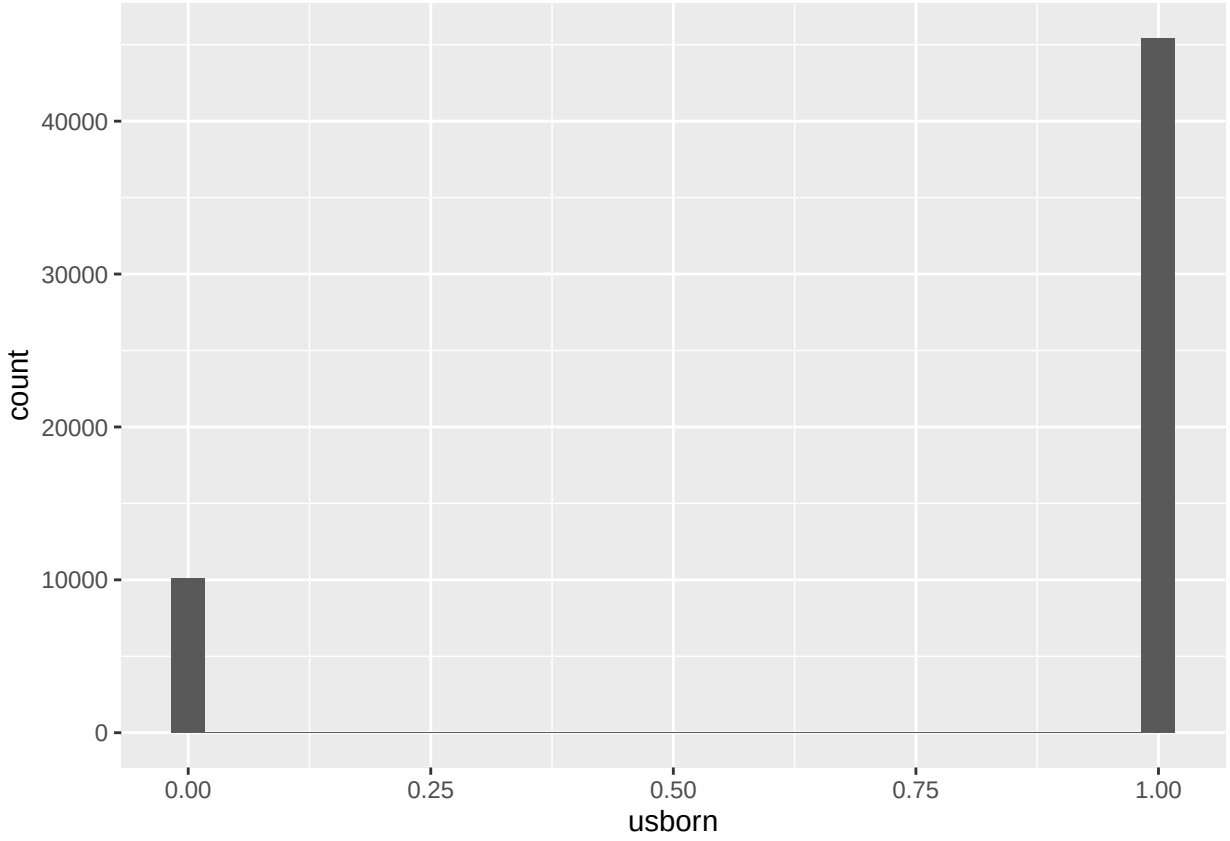
```
usborn.f <- factor(data_project$usborn, labels = c("Naturalized", "US born"))
table(CITIZEN,data_project$usborn)
```

```
##
## CITIZEN     0     1
##       1     0 44551
##       2     0   352
##       3     0   543
##       4  4870     0
##       5  5251     0
```

In the dataset, we have more than 45k of native Americans which is much more than the observations of naturalized citizens and other legal residents.

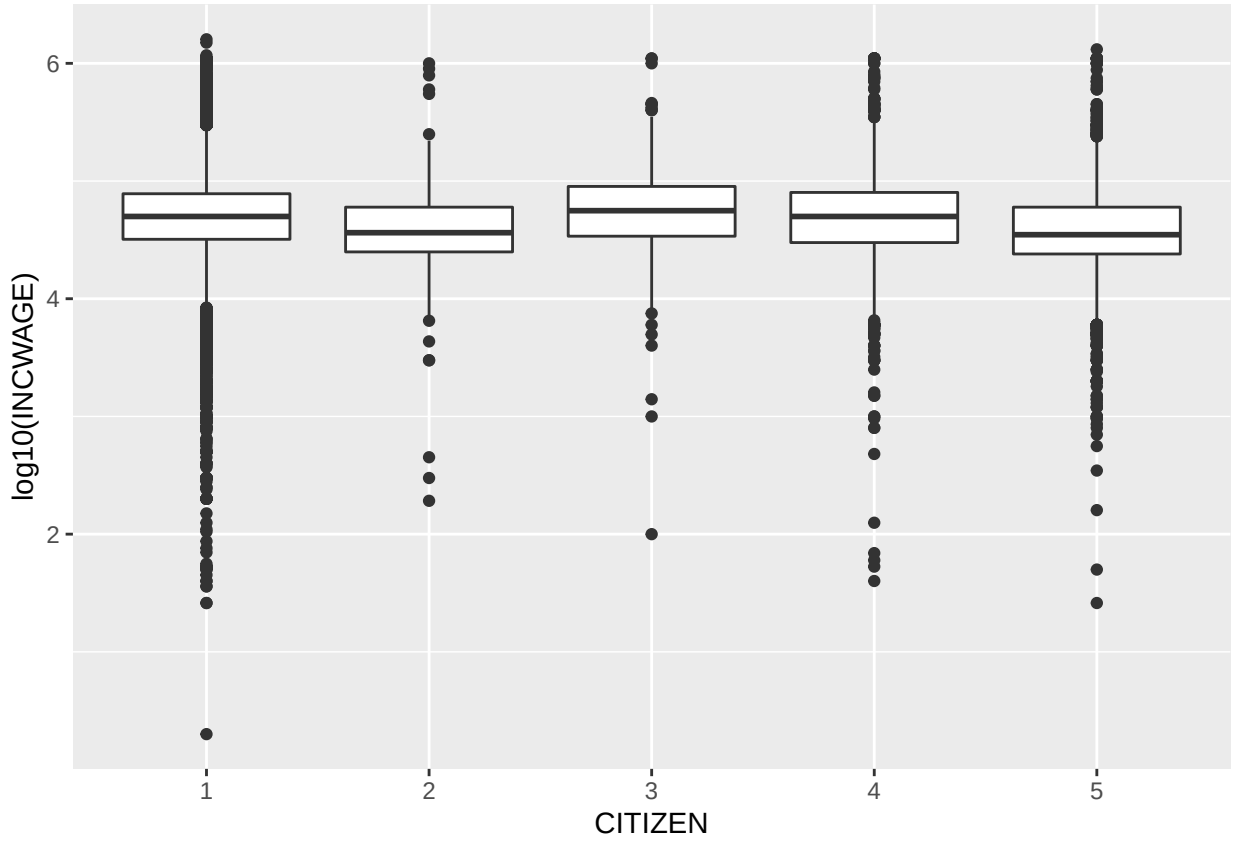
```
ggplot(data = data_project)+
  geom_histogram(aes(usborn))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



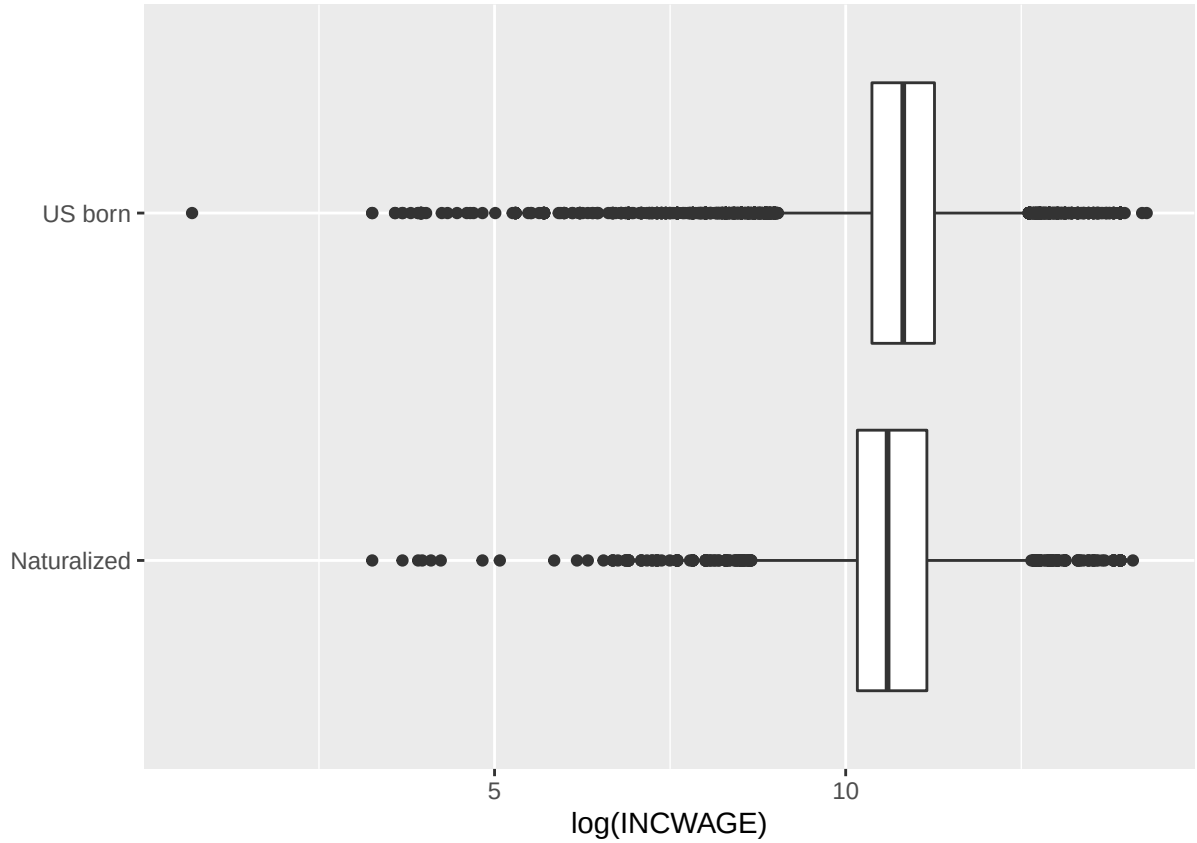
Visualize the complete groups under CITIZEN and compare those 5 groups. There is no much difference among native Americans, naturalized citizens and other legal residents in terms of log income wages.

```
ggplot(data=data_project, mapping=aes(x=as.character(CITIZEN),y=log10(INCWAGE)))+geom_boxplot()+
  xlab("CITIZEN")
```



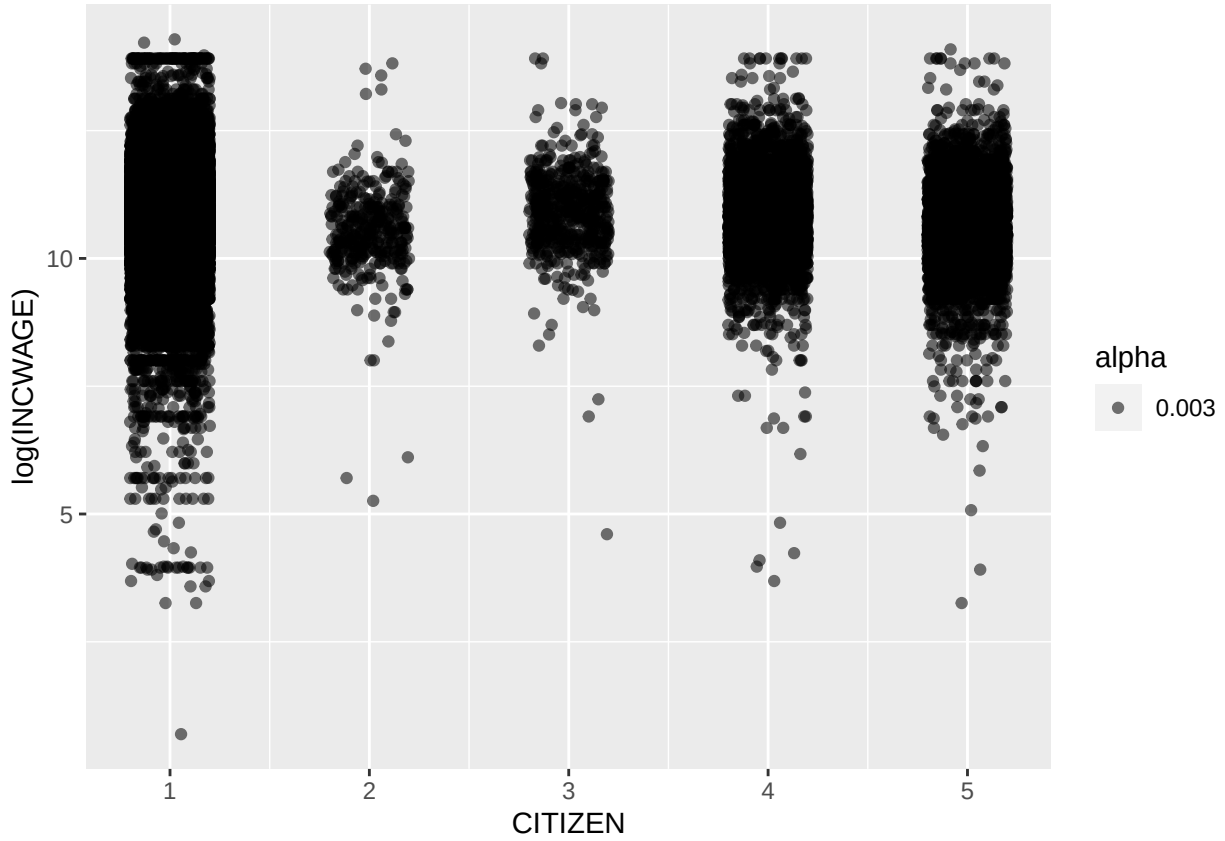
Using boxplot to visualize log income wage difference between native Americans an naturalized Americans. The plot shows slight higher log income wage of native Americans compare to the naturalized americans.

```
ggplot(data_project, aes(x=log(INCWAGE), y=usborn.f))+geom_boxplot()+ylab("")
```



The native Americans are more widely spred in terms of income wage compare to naturalized citizen and other legal residents.

```
ggplot(data_project,aes(jitter(CITIZEN),y=log(INCWAGE),alpha=0.003))+geom_point()+xlab("CITIZEN")
```

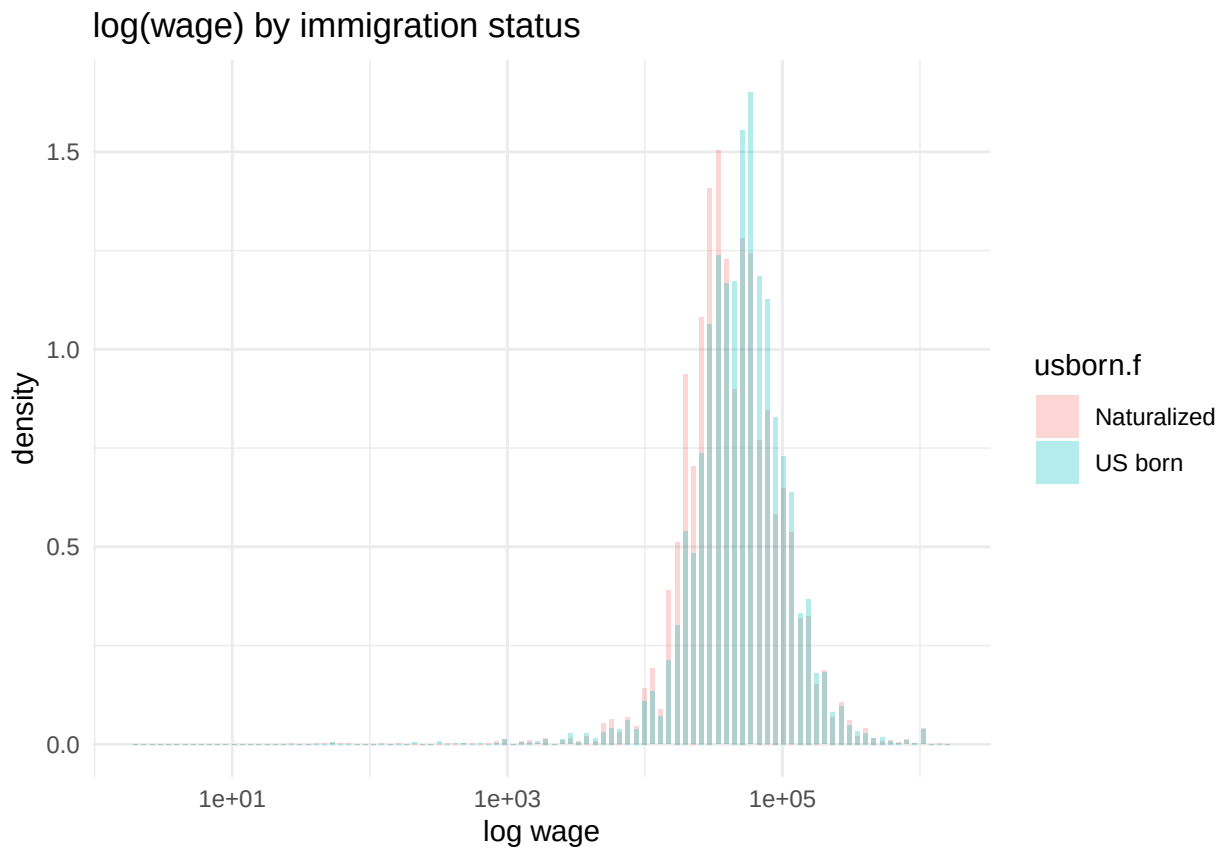


The histogram shows the native Americans are slightly on the right side of the naturalized citizens, which means the native Americans make slightly more income wage than native Americans.

```
usborn.f <- factor(data_project$usborn, labels = c("Naturalized", "US born"))
table(usborn.f)
```

```
## usborn.f
## Naturalized     US born
##       10121       45446
```

```
data_project %<>%
    mutate(usborn.f <- factor(usborn, labels = c("Naturalized or Resident","US born")))
ggplot(data_project, aes(x = INCWAGE, y = ..density.., fill = usborn.f)) +
    geom_histogram(bins = 100, position = position_dodge(width = 0), alpha = 0.3) +
    labs(x = "log wage", y = "density", title = "log(wage) by immigration status") +
    scale_x_log10() +
    theme_minimal()
```



Inference

The Inference is trying to find out the difference in average income wage between native Americans and foreigners in the United States. In order to determine if there is significance difference between native Americans and foreigners, using the T-distribution to perform the hypothesis test. According to the histogram and boxplot, I believe there is a slight difference in income wages between these two groups.

Null hypothesis: There is no siginificant difference in average income wage between native Americans and foreginers. Alternative hypothesis: There is a significant difference in the average income wage between native Americans and foreginers.

```
describeBy(data_project$INCWAGE,data_project$usborn)
```

```
##
##  Descriptive statistics by group
## group: 0
##    vars     n    mean       sd median trimmed     mad min     max   range skew
## X1    1 10121 60848.9 77626.93  40000 60848.9 26686.8  26 1314999 1314973 7.62
##    kurtosis     se
## X1    85.98 771.62
## ------------------------------------------------------------
## group: 1
##    vars     n    mean       sd median trimmed   mad min     max   range skew
## X1    1 45446 66701.6 75949.95  50000 66701.6 29652   2 1599999 1599997 7.75
##    kurtosis     se
```

```
## X1    90.84 356.27
```

```
wage<-data_project[ ,c("INCWAGE","CITIZEN")]
wage
```

```
## # A tibble: 55,567 x 2
##      INCWAGE         CITIZEN
##    <dbl+lbl>       <int+lbl>
##  1     12000 1 [Born in U.S]
##  2     55000 1 [Born in U.S]
##  3     32621 1 [Born in U.S]
##  4     24002 1 [Born in U.S]
##  5     50000 1 [Born in U.S]
##  6     41500 1 [Born in U.S]
##  7     52000 1 [Born in U.S]
##  8     50000 1 [Born in U.S]
##  9     46701 1 [Born in U.S]
## 10     10700 1 [Born in U.S]
## # ... with 55,557 more rows
```

```
q<-tapply(wage$INCWAGE,wage$CITIZEN,mean)
diff.mean<-sum(q[1:3])/3 - sum(q[4:5])/2
diff.mean
```

```
## [1] 5344.841
```

```
with(data_project, t.test(INCWAGE[CITIZEN<=3],INCWAGE[CITIZEN==4]),paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  INCWAGE[CITIZEN <= 3] and INCWAGE[CITIZEN == 4]
## t = -1.7666, df = 5792.6, p-value = 0.07734
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4592.5296   238.7372
## sample estimates:
## mean of x mean of y
##   66701.6   68878.5
```

Conclusion

Based on the hypothesis test above, the p-value of 0.07734 which is greater than 0.05. Therefore, I cannot reject my null hypothesis in 95% of confidence interval. The mean difference between native Americans and naturalized citizen with other legal residents are $5344.841. In other words, there is no significant difference in average income wage between native Americans and foreigners. It is a surprise and a good news for me because as a foreigner, I still believe "American Dream" still exists. At least, I may make incomes as much as the native Americans do.