Data Exploration

12/7/2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
## Warning: package 'ggplot2' was built under R version 3.4.4
## Warning: package 'tibble' was built under R version 3.4.4
## Warning: package 'tidyr' was built under R version 3.4.4
## Warning: package 'readr' was built under R version 3.4.4
## Warning: package 'purrr' was built under R version 3.4.4
## Warning: package 'dplyr' was built under R version 3.4.4
## Warning: package 'stringr' was built under R version 3.4.4
## Warning: package 'forcats' was built under R version 3.4.4
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
## The following object is masked from 'package:purrr':
##
##     set_names
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
##data wrangling ----
df <- read_csv("ipums_time_used.csv") %>%
  clean_names() %>%
  filter(!is.na(wb_resp)) %>%
  rename(
    residence = metro,
```

```
    education = educ,
    employment_status = empstat,
    wellbeing_response = wb_resp,
    hourly_wage = hourwage,
    hours_work_wk = uhrsworkt) %>%
  select(-wt06, -pernum, -lineno, -caseid) %>%
  mutate(sex = case_when(
           sex == 1 ~ "Male",
           sex == 2 ~ "Female"),
         employment_status = case_when(
           employment_status == 1 ~ "Employed",
           employment_status == 2 ~ "Employed",
           employment_status == 3 ~ "Unemployed",
           employment_status == 4 ~ "Unemployed",
           employment_status == 5 ~ "Not in labor force"),
         education = case_when(education < 20 ~ "BelowHS",
                               education == 20 ~ "High School",
                               education == 21 | education == 30 ~ "Some College",
                               education == 31 | education == 32 ~ "Associate Degree",
                               education == 40 ~ "Bachelor's Degree",
                               education == 41 ~ "Master's Degree",
                               education == 42 ~ "Professional Degree",
                               education == 43 ~ "Doctoral Degree"),
         race = case_when(
           race == 100 ~ "White",
           race == 110 ~ "Black",
           race == 120 ~ "American Indian",
           race == 131 ~ "Asian",
           race == 132 ~ "Pacific Islander",
           race == 200 | race == 210 | race == 211 | race == 212 | race == 300 | race == 400 ~ "Black-M
           race == 201 | race == 202 | race == 203 | race == 310 | race == 320 ~ "White-Mixed",
           race == 220 | race == 230 ~ "Other-Mixed"),
         residence = case_when (
           residence == 1 ~ "Metropolitan: Central City",
           residence == 2|residence == 3 ~ "Metropolitan: Others",
           residence == 4 ~ "Nonmetropolitan")) %>%
  filter (!is.na(residence))
```

```
## Parsed with column specification:
## cols(
##   YEAR = col_double(),
##   CASEID = col_double(),
##   METRO = col_double(),
##   PERNUM = col_double(),
##   LINENO = col_double(),
##   WT06 = col_double(),
##   AGE = col_double(),
##   SEX = col_double(),
##   RACE = col_double(),
##   EDUC = col_double(),
##   EMPSTAT = col_double(),
##   UHRSWORKT = col_double(),
##   HOURWAGE = col_double(),
##   WB_RESP = col_double()
```

```
## )
```

Questions:

1. Does location effect happiness of a race?
2. Are people happier when they make more money?

Hypotheses:

1. Location and Race does not have an effect on someone's happiness $\beta_0 = 0$ $\beta_A \neq 0$
2. Someone's hourly wage does not have an effect on someone's happpiness $\beta_0 = 0$ $\beta_A \neq 0$

Models

```
#Does race have an effect on someone's wellbeing?
race <- glm(wellbeing_response ~ race, data = df, family = binomial)
summary(race)
```

```
##
## Call:
## glm(formula = wellbeing_response ~ race, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4535   0.3616   0.3616   0.3616   0.4773
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.61239    0.23768  10.991   <2e-16 ***
## raceAsian               -0.49762    0.25331  -1.964   0.0495 *
## raceBlack               -0.19906    0.24266  -0.820   0.4120
## raceBlack-Mixed          0.34698    0.45471   0.763   0.4454
## raceOther-Mixed          9.95367  102.69325   0.097   0.9228
## racePacific Islander    -0.24527    0.48866  -0.502   0.6157
## raceWhite                0.08249    0.23890   0.345   0.7299
## raceWhite-Mixed         -0.01454    0.32126  -0.045   0.9639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18251  on 36800  degrees of freedom
## Residual deviance: 18193  on 36793  degrees of freedom
## AIC: 18209
##
## Number of Fisher Scoring iterations: 11
```

```
#Does hourly wage have an effect on someone's wellbeing?
wage_logm <- glm(wellbeing_response ~ hourly_wage, data = df, family = binomial)
summary(wage_logm)
```

```
##
## Call:
## glm(formula = wellbeing_response ~ hourly_wage, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3552   0.3595   0.3810   0.3810   0.3810
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.709e+00  4.043e-02  67.013  < 2e-16 ***
## hourly_wage -1.225e-04  4.738e-05  -2.585  0.00973 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18251  on 36800  degrees of freedom
## Residual deviance: 18245  on 36799  degrees of freedom
## AIC: 18249
##
## Number of Fisher Scoring iterations: 5
```

$\text{logit}(\pi) = 2.709\text{e+}00 - -1.225\text{e-}04 * \text{hourly wage}$

```
exp(cbind(OR = coef(wage_logm), confint(wage_logm)))
```

```
## Waiting for profiling to be done...
```

```
##                     OR      2.5 %     97.5 %
## (Intercept) 15.0149843 13.8835525 16.2679316
## hourly_wage  0.9998775  0.9997841  0.9999698
```

```
#Does age status has an effect on someone's happiness?
age <- glm(wellbeing_response ~ age, data = df, family = binomial)
summary(age)
```

```
##
## Call:
## glm(formula = wellbeing_response ~ age, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4374   0.3474   0.3671   0.3929   0.4365
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.049528   0.061547  49.548  < 2e-16 ***
## age         -0.008779   0.001157  -7.586  3.3e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18251  on 36800  degrees of freedom
```

```
## Residual deviance: 18194  on 36799  degrees of freedom
## AIC: 18198
##
## Number of Fisher Scoring iterations: 5
```