Final Project Proposal

Group 11

11/8/2020

Group Members

- James Brennan
- Linh Huynh
- Zachary Palmer
- Mahima Sindhu

Refined Goals of Analysis

Our group is using the `atus` database to investigate whether various demographic variables correlate to family income. Specifically, we will be looking at region, state, educational attainment, home ownership (rent vs. own), household size, race, and time usage. After visualizing the data, we will test the correlations between each of these demographic variables and income using paired t-tests or other related statistical analyses. Using the results from these statistical analyses, we will determine relationships between the demographic variables and how they collectively impact family income and how income potentially affects time usage. This information could be useful to better inform policy regarding issues relevant to specific demographics.

Preliminary Data Exploration

```
time_data = ungroup(atusact)
time_data2 = time_data %>%
  group_by(tucaseid, tiercode) %>%
  summarise(tier1_2Code = tiercode %/% 100,
            tier1 = tiercode %/% 10000,
            tier2 = (tiercode %% 10000) %/% 100,
            tier3 = (tiercode %% 10000) %% 100,
            dur = dur) %>%
  inner_join(atuscps) %>%
  select(tiercode, tier1, dur, region, state, famincome) %>%
  drop_na()
```

Zach

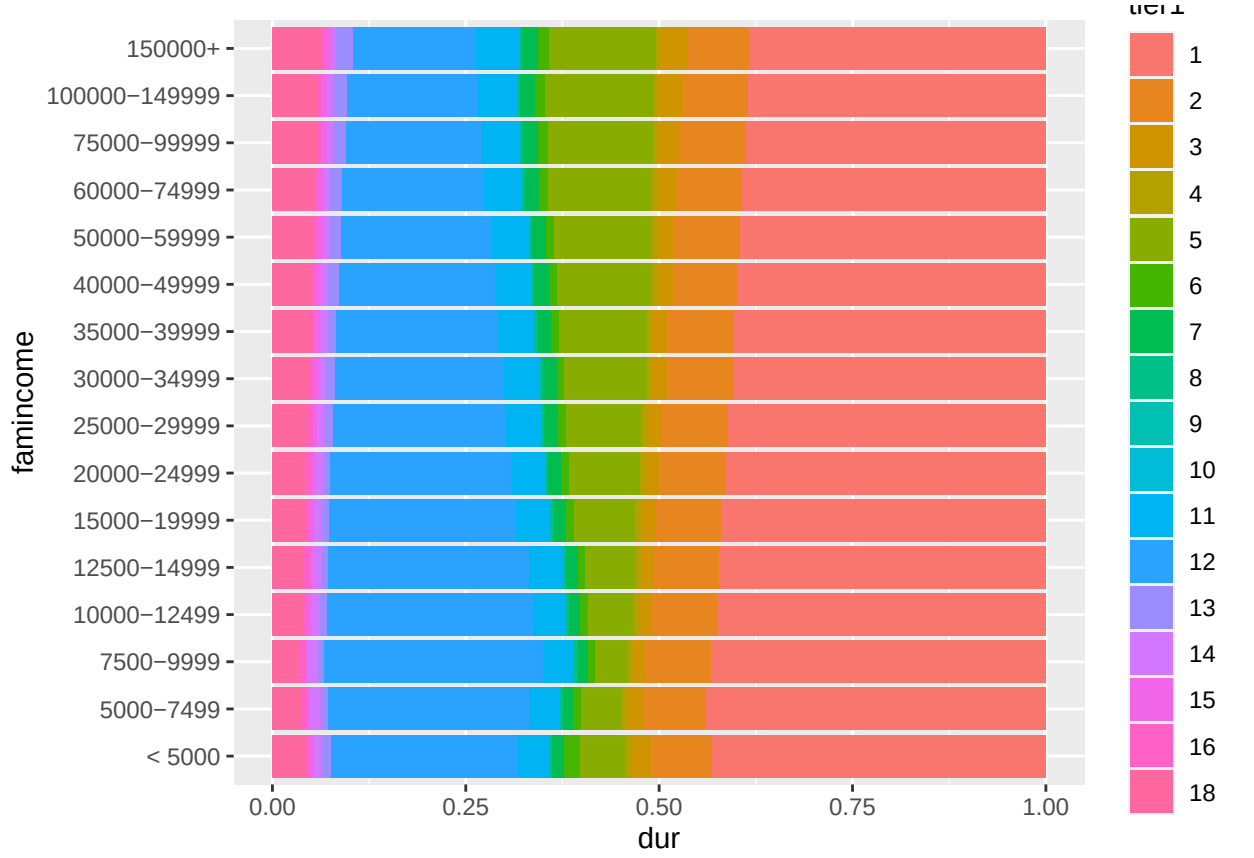
```
## `summarise()` regrouping output by 'tucaseid' (override with `.groups` argument)

## Joining, by = "tucaseid"

## Adding missing grouping variables: `tucaseid`
```

```
time_data2 %>% head(5)
```

```
## # A tibble: 5 x 7
## # Groups:   tucaseid [1]
##   tucaseid tiercode tier1   dur region state famincome
##      <dbl>    <int> <dbl> <int> <fct>  <fct> <fct>
## 1  2.00e13    10101     1   870 west   CA    60000-74999
```

```
## 2  2.00e13    10201     1    40 west   CA    60000-74999
## 3  2.00e13   110101    11     5 west   CA    60000-74999
## 4  2.00e13   120303    12   325 west   CA    60000-74999
## 5  2.00e13   130124    13   200 west   CA    60000-74999
```

```
time_data2$tier1 = as.factor(time_data2$tier1)
ggplot(data = time_data2, mapping = aes(x = famincome, y = dur, fill = tier1)) +
  geom_col(position = "fill") +
  coord_flip()
```




```
# Find average family income by educational attainment
avg_income_by_edu <- atuscps
avg_income_by_edu <- avg_income_by_edu %>%
  separate(famincome, into=c('income_low','income_high'), sep='-',convert=TRUE)%>%
  mutate(income_low = as.integer(income_low)) %>%
  drop_na() %>%
  mutate(fam_income_mid = (income_high+income_low)/2) %>%
  group_by(education=edu) %>%
  summarise(N = n(),
            avg_income = mean(fam_income_mid)) %>%
  arrange(avg_income)
```

James

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 15474 rows [205,
## 211, 248, 292, 311, 377, 400, 431, 504, 535, 554, 555, 581, 596, 611, 616, 655,
```

```
## 662, 710, 766, ...].
```

```
## Warning: Problem with `mutate()` input `income_low`.
## i NAs introduced by coercion
## i Input `income_low` is `as.integer(income_low)`.
```

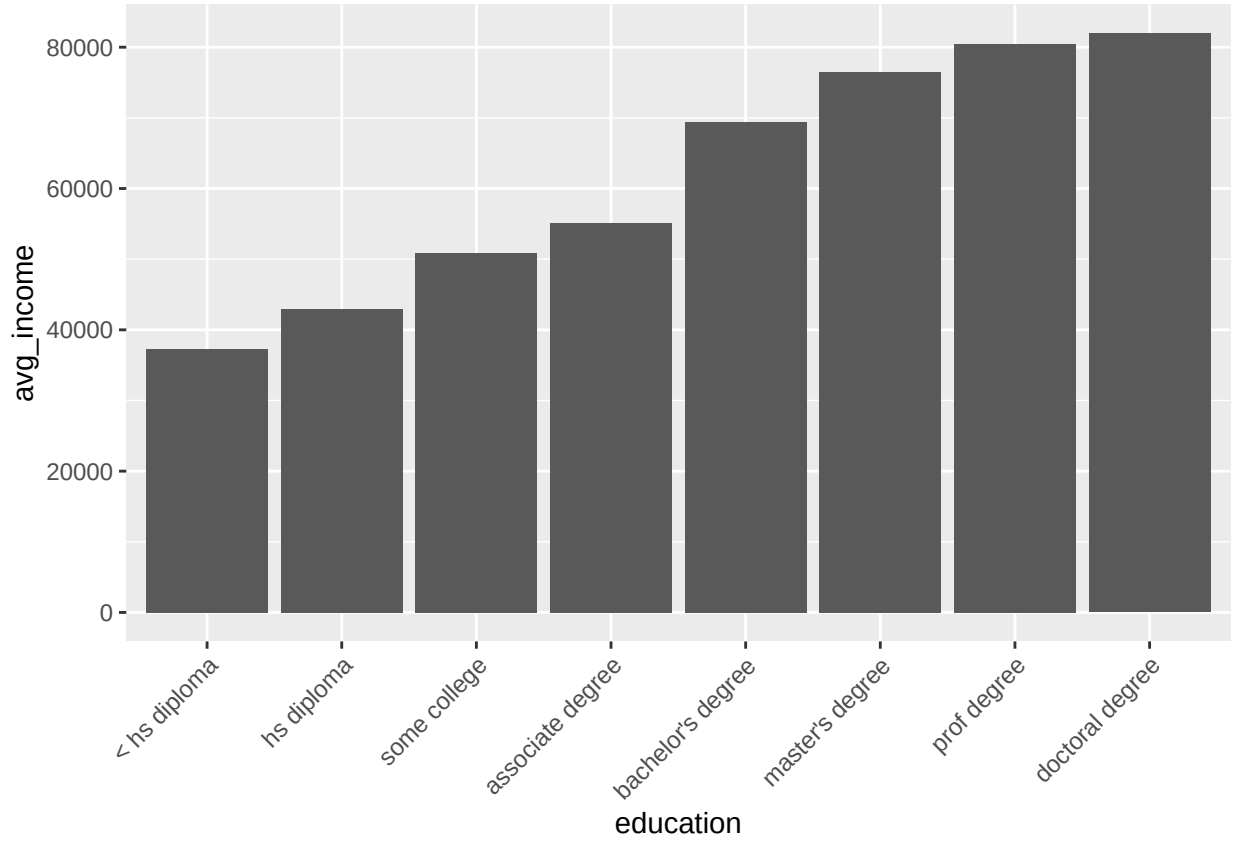
```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
avg_income_by_edu %>% head(5)
```

```
## # A tibble: 5 x 3
##   education             N avg_income
##   <fct>             <int>      <dbl>
## 1 < hs diploma      23982     37274.
## 2 hs diploma        41128     42942.
## 3 some college      27842     50824.
## 4 associate degree  14646     55166.
## 5 bachelor's degree 28930     69471.
```

```
#Visualize results
ggplot(data=avg_income_by_edu, aes(x=education, y=avg_income)) +
  geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Investigate by region
# Find average family income by educational attainment
avg_income_by_edu_region <- atuscps
```

```r
avg_income_by_edu_region <-avg_income_by_edu_region %>%
  separate(famincome, into=c('income_low','income_high'), sep='-',convert=TRUE)%>%
  mutate(income_low = as.integer(income_low)) %>%
  drop_na() %>%
  mutate(fam_income_mid = (income_high+income_low)/2) %>%
  group_by(region,edu) %>%
  summarise(N = n(),
            avg_income = mean(fam_income_mid)) %>%
  arrange(avg_income)
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 15474 rows [205,
## 211, 248, 292, 311, 377, 400, 431, 504, 535, 554, 555, 581, 596, 611, 616, 655,
## 662, 710, 766, ...].
```

```
## Warning: Problem with `mutate()` input `income_low`.
## i NAs introduced by coercion
## i Input `income_low` is `as.integer(income_low)`.
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## `summarise()` regrouping output by 'region' (override with `.groups` argument)
```

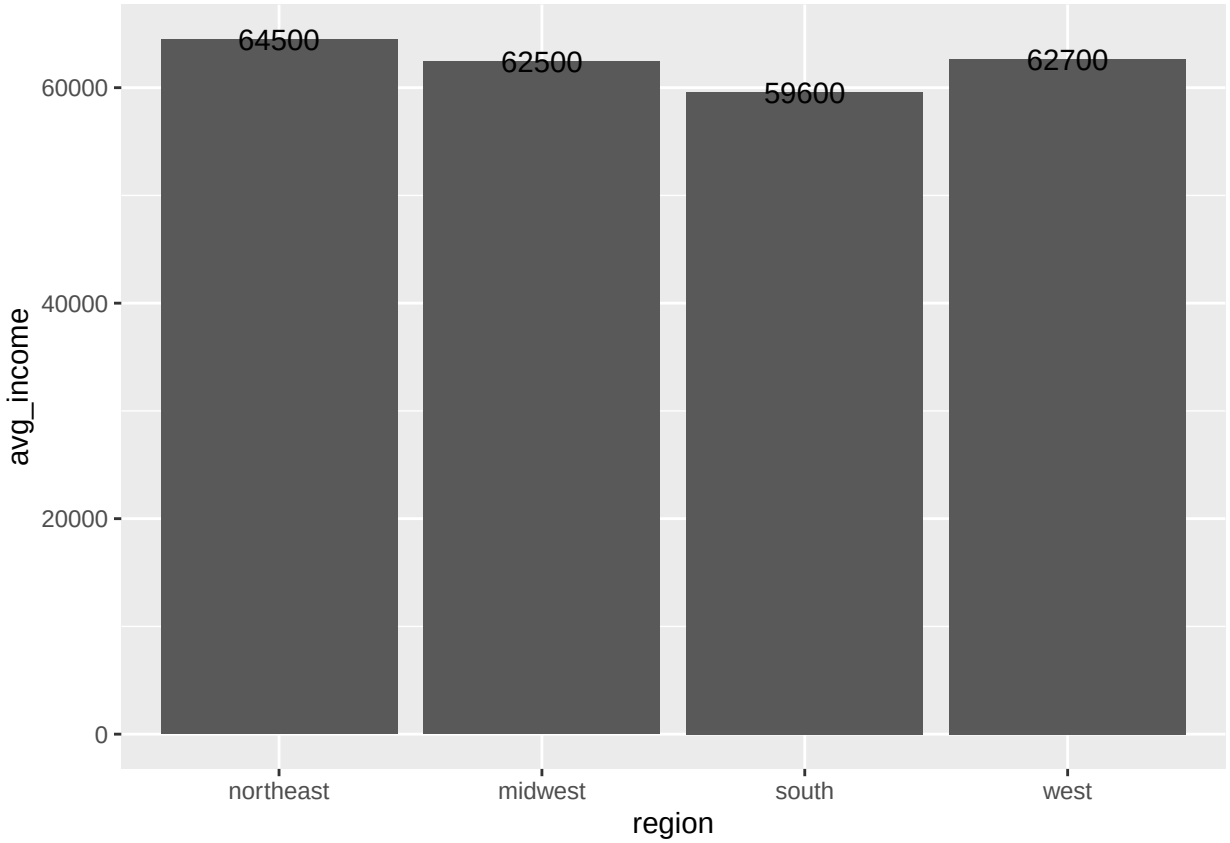
```r
avg_income_by_edu_region %>% head(5)
```

```
## # A tibble: 5 x 4
## # Groups:   region [4]
##   region    edu              N avg_income
##   <fct>     <fct>        <int>      <dbl>
## 1 south     < hs diploma  9740     33590.
## 2 west      < hs diploma  5793     38946.
## 3 northeast < hs diploma  3690     39558.
## 4 south     hs diploma   15480     40362.
## 5 midwest   < hs diploma  4759     41010.
```

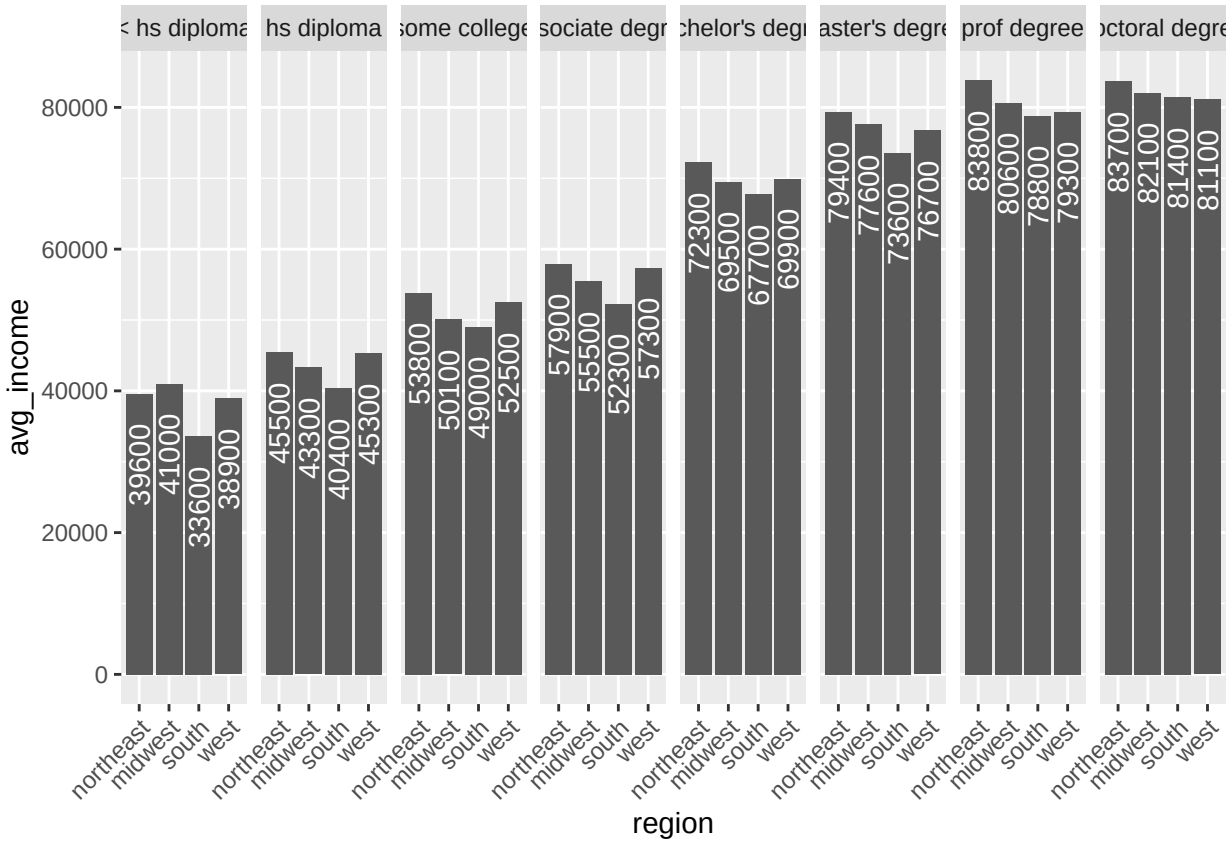
```r
# Compare regions by income
region_df <- avg_income_by_edu_region %>%
  group_by(region) %>%
  summarise(N=n(),
            avg_income = mean(avg_income)) %>%
  arrange(avg_income)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
ggplot(region_df, aes(x=region, y=avg_income, label=avg_income)) +
  geom_bar(stat='identity') +
  geom_text(aes(label = signif(avg_income, digits = 3)))
```




```
# Compare income by region and edu
ggplot(avg_income_by_edu_region, aes(x=region, y=avg_income)) +
    geom_bar(position='stack',stat='identity') +
    scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    geom_text(aes(label = signif(avg_income, digits = 3)), angle = 90,
              nudge_y = -10000, color = 'white') +
    facet_grid(~ edu)
```



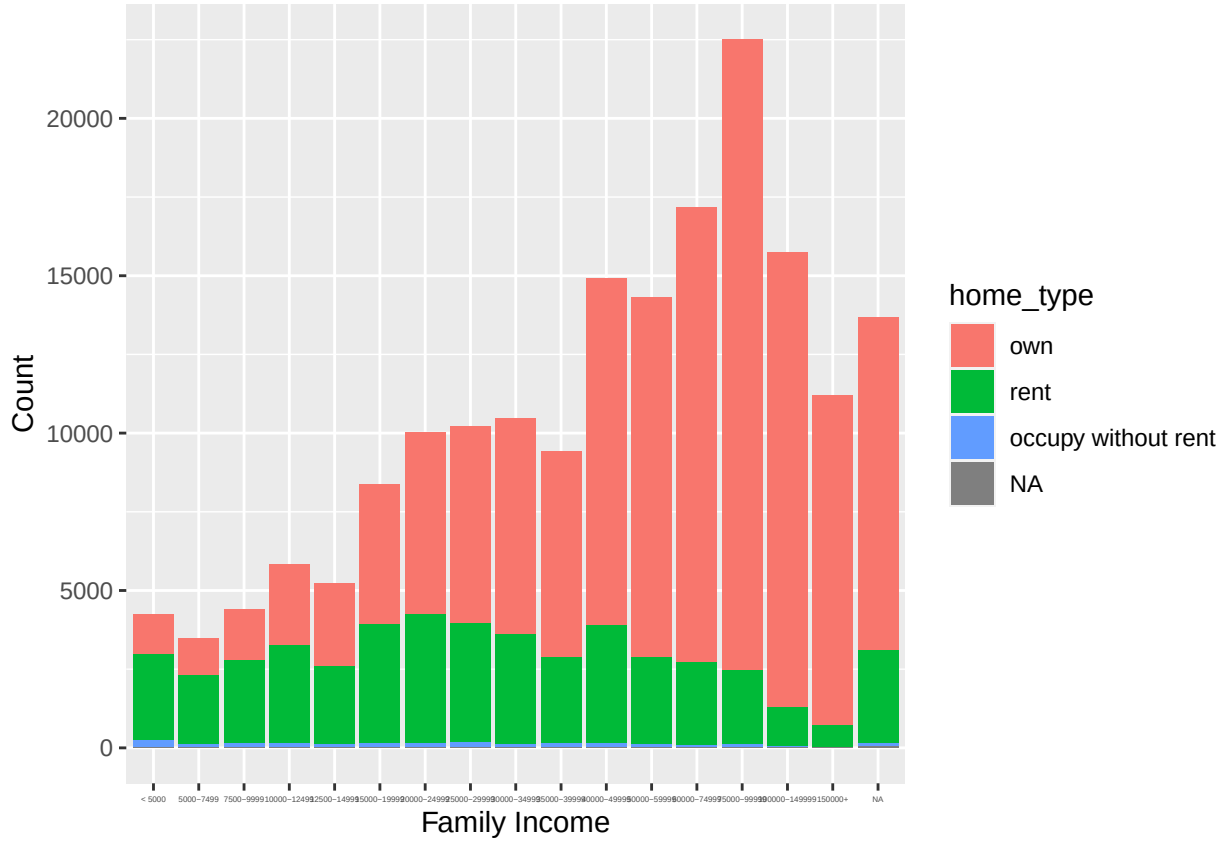

```
data("atuscps")
cps <- as_tibble(atuscps)
cps
```

Mahima

```
## # A tibble: 181,335 x 13
##    tucaseid region state sex     age edu   race  hispanic country_born citizen
##       <dbl> <fct>  <fct> <fct> <int> <fct> <fct> <fct>    <fct>        <fct>
##  1  2.00e13 west   CA    male     60 mast~ Blac~ no       US           yes
##  2  2.00e13 west   CA    fema~    41 some~ Whit~ no       US           yes
##  3  2.00e13 west   CA    fema~    25 asso~ Whit~ no       US           yes
##  4  2.00e13 south  GA    fema~    36 hs d~ Blac~ no       US           yes
##  5  2.00e13 south  KY    male     50 prof~ Whit~ no       US           yes
##  6  2.00e13 south  KY    fema~    32 bach~ Whit~ no       US           yes
##  7  2.00e13 south  LA    fema~    43 hs d~ Whit~ no       US           yes
##  8  2.00e13 midwe~ MI    fema~    20 some~ Whit~ no       US           yes
##  9  2.00e13 midwe~ MN    fema~    33 asso~ Whit~ no       US           yes
## 10  2.00e13 north~ NJ    fema~    38 asso~ Blac~ no       non-US       no
## # ... with 181,325 more rows, and 3 more variables: marital <fct>,
## #   home_type <fct>, famincome <fct>
```

```
library(ggplot2)
plotcps <- ggplot(data=atuscps) +
  geom_bar(mapping=aes(x=famincome, fill=home_type)) +
  xlab("Family Income") +
```

```
  ylab("Count")
plotcps + theme(axis.text.x = element_text(size = 3))
```



Linh Preparing Data Set

```
reg_income <- atuscps %>%
  select(region, famincome) %>%
  separate(famincome, into = c("low_famincome", "high_famincome"), sep = "-") %>%
  mutate(
    low_famincome = as.integer(low_famincome),
    high_famincome = as.integer(high_famincome)
  ) %>%
  filter(
    !low_famincome %in% NA,
    !high_famincome %in% NA
  )
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 15474 rows [205,
## 211, 248, 292, 311, 377, 400, 431, 504, 535, 554, 555, 581, 596, 611, 616, 655,
## 662, 710, 766, ...].
```

```
## Warning: Problem with `mutate()` input `low_famincome`.
## i NAs introduced by coercion
## i Input `low_famincome` is `as.integer(low_famincome)`.
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```r
reg_income %>% head(5)
```

```
##   region low_famincome high_famincome
## 1   west         60000          74999
## 2   west         75000          99999
## 3  south         20000          24999
## 4  south         75000          99999
## 5  south         40000          49999
```

Finding Original Income Brackets

```r
levels(atuscps$famincome)
```

```
##  [1] "< 5000"        "5000-7499"     "7500-9999"     "10000-12499"
##  [5] "12500-14999"   "15000-19999"   "20000-24999"   "25000-29999"
##  [9] "30000-34999"   "35000-39999"   "40000-49999"   "50000-59999"
## [13] "60000-74999"   "75000-99999"   "100000-149999" "150000+"
```

Bounds of Family Income

```r
bounds_reg_income <- reg_income %>%
  group_by(region) %>%
  summarise(
    mean_low_fincome = mean(low_famincome),
    mean_high_fincome = mean(high_famincome)
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
bounds_reg_income %>% head(5)
```

```
## # A tibble: 4 x 3
##   region    mean_low_fincome mean_high_fincome
##   <fct>                <dbl>             <dbl>
## 1 northeast           48936.            64740.
## 2 midwest             46743.            61227.
## 3 south               43389.            56696.
## 4 west                47499.            62535.
```

New Income Brackets

```r
i_brackets <- bounds_reg_income %>%
  mutate(
    mean_low_fincome = round(mean_low_fincome),
    mean_high_fincome = round(mean_high_fincome)
  ) %>%
  unite(income_bracket, mean_low_fincome, mean_high_fincome, sep = "-")
i_brackets %>% head(5)
```

```
## # A tibble: 4 x 2
##   region    income_bracket
##   <fct>     <chr>
## 1 northeast 48936-64740
## 2 midwest   46743-61227
## 3 south     43389-56696
## 4 west      47499-62535
```

Racial Breakdown

```r
num_resp <- nrow(atuscps); num_resp
```

```
## [1] 181335
```

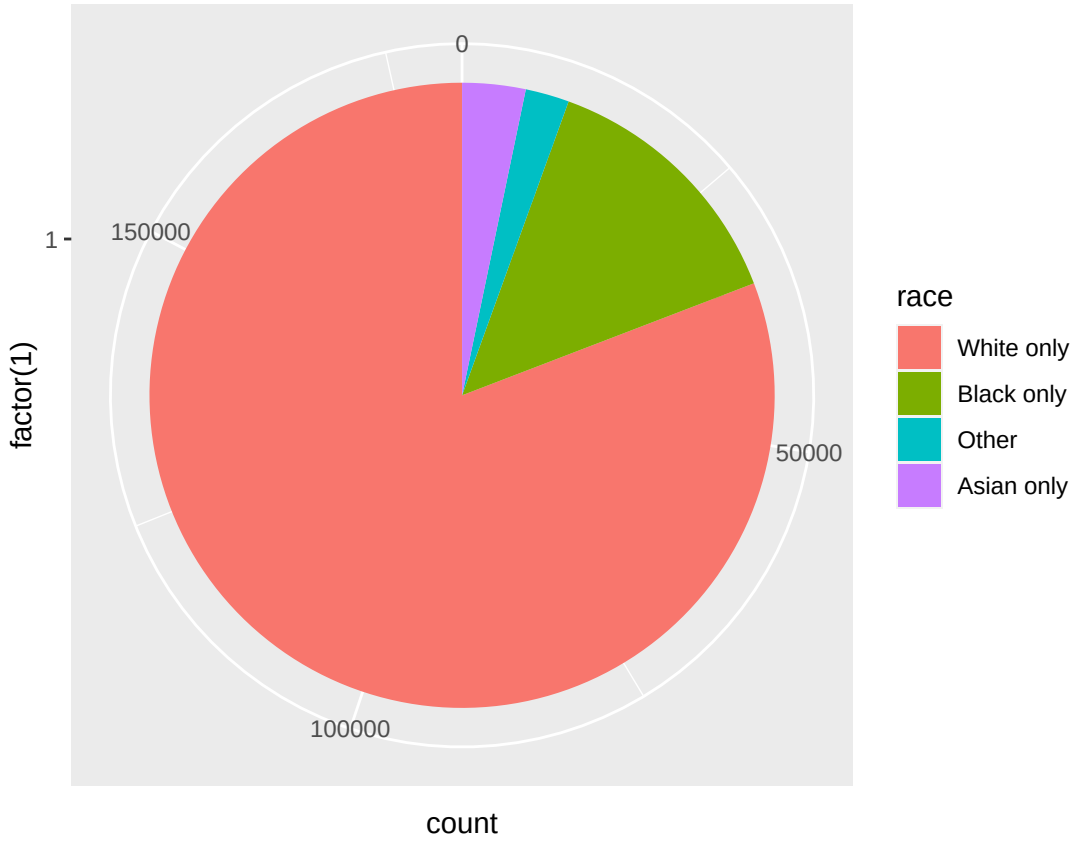
```r
race <- atuscps %>%
  select(race) %>%
  group_by(race) %>%
  summarise(
    num_race = n()
  ) %>%
  mutate(percent_race = percent((num_race/num_resp)))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
race %>% head(5)
```

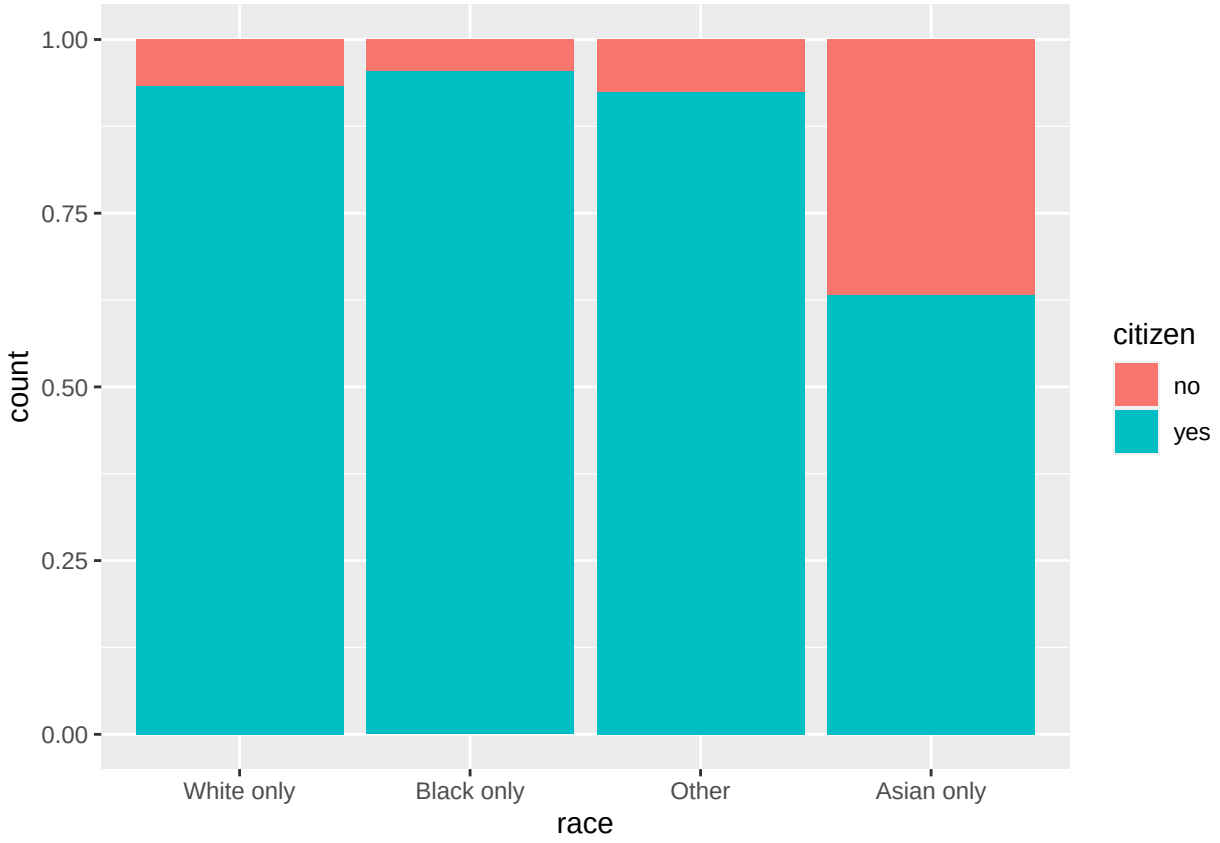
```
## # A tibble: 4 x 3
##   race       num_race percent_race
##   <fct>         <int> <chr>
## 1 White only   146578 80.8%
## 2 Black only    24719 13.6%
## 3 Other          4107 2.3%
## 4 Asian only     5931 3.3%
```

```r
race_plot <- ggplot(data = atuscps, aes(x = factor(1), fill = race)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y")
race_plot
```



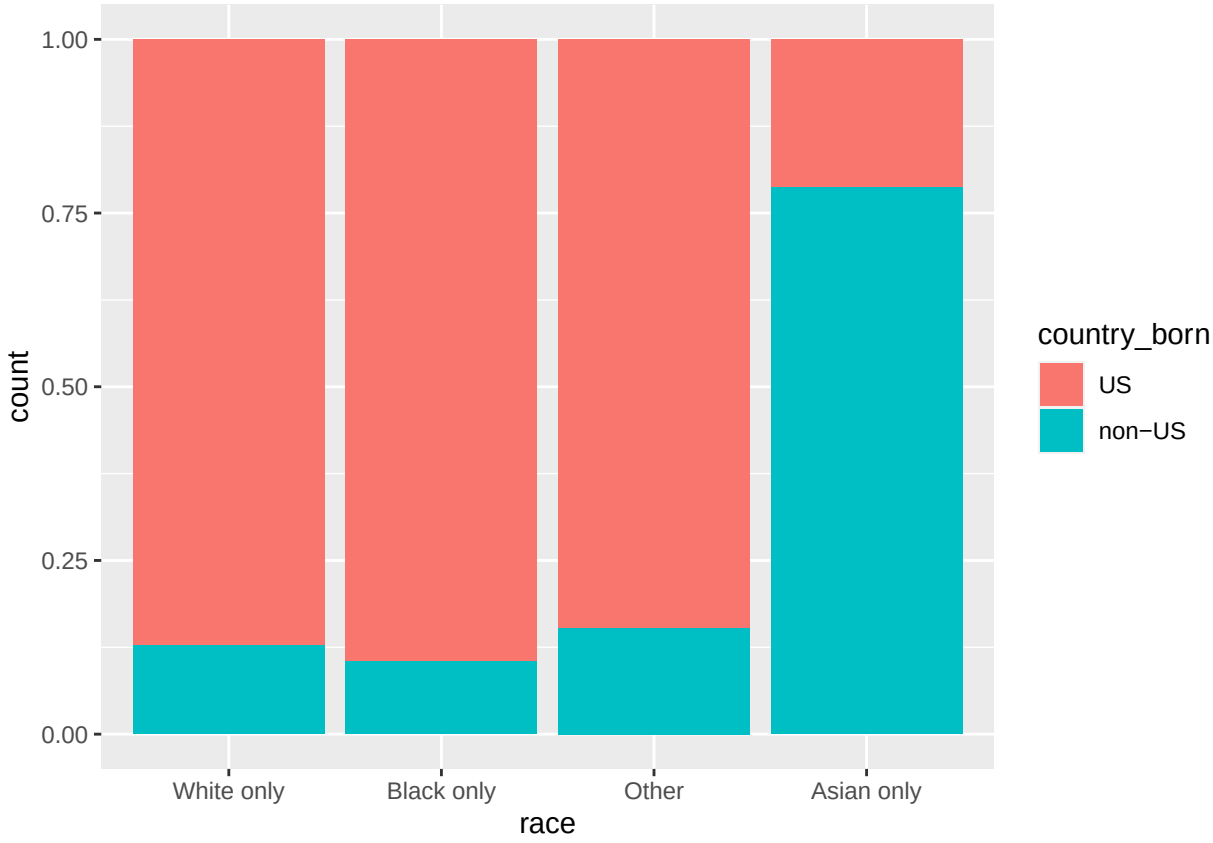
Citizenship Status Across Races

```
citizenship_plot <- ggplot(data = atuscps, mapping = aes(x = race, fill = citizen)) +
  geom_bar(position = "fill")
citizenship_plot
```



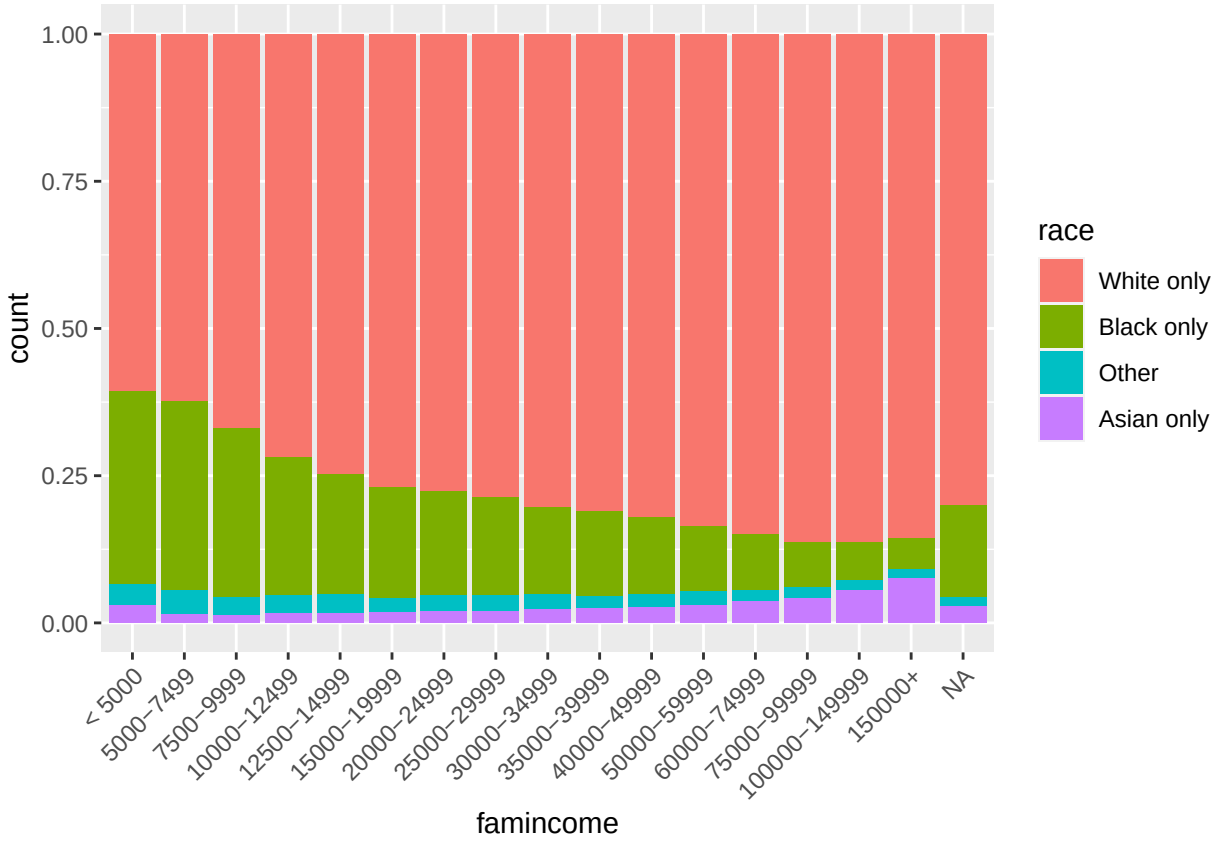
Country of Origin Across Races

```
country_plot <- ggplot(data = atuscps, mapping = aes(x = race, fill = country_born)) +
  geom_bar(position = "fill")
country_plot
```



Family Income Based on Race

```
race_income <- ggplot(data = atuscps, mapping = aes(x = famincome, fill = race)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
race_income
```


1.00
0.75
0.50
0.25
0.00
count
< 5000
5000–7499
7500–9999
10000–12499
12500–14999
15000–19999
20000–24999
25000–29999
30000–34999
35000–39999
40000–49999
50000–59999
60000–74999
75000–99999
100000–149999
150000+
NA
famincome
race
White only
Black only
Other
Asian only