Final Report

due November 16, 2021 by 11:59 PM

Grace Lee and Jiyun Hyo

November 16, 2021

Load Packages

```
library(dplyr)
library(tidyverse)
library(sf)
library(viridis)
library(tidyverse)
library(tidymodels)
library(ggspatial) #for scale annotation
library(ggplot2)
```

Load Data

```
data <- read.csv(file = '../data/COVID_raw_12.8.csv')
tidy_data <- select(data, c('Participant_ID', 'age',"usres", "state", "race", "sex", "localsip

tidy_data <- tidy_data %>%
  filter(is.na(tidy_data$race)== FALSE & is.na(tidy_data$localsiphours)== FALSE)
```

Introduction and Data, including Research Questions

In response to the COVID-19 pandemic, 42 states and territories issued mandatory stay-at-home orders between March 1 to May 31, 2020, affecting 2,355 (73%) of 3,233 U.S. counties (CDC, 2020). These stay-at-home policies reduced both population movement and person-to-person contact, which slowed the spread of COVID-19. In a study published by Cambridge University Press in May 2020, the total number of infections was projected to reach 287 million in the absence of stay-at-home and social distancing policies and 188 million with the enforcement of these policies, translating to 1.24 million lives saved (Thunström et al., 2020).

Due to the importance of stay-at-home orders in slowing the spread of COVID in the United States, we want to ask if the average number of hours spent at home differed between different populations. For example, we want to ask if people of different races and income levels, among other variables, differed significantly in their mean number of hours spent at home. We also wanted to ask if different

demographic characteristics affected the probability that the participant had left the home at all.

To do so, we used the dataset, "Associations of Urbanicity and Sociodemographic Characteristics with Protective Health Behaviors and Reasons for Leaving the Home during COVID-19," found on the Harvard Dataverse (Burford, 2020). The data was collected between April 15-May 5, 2020 through a 15-minute self-completed online questionnaire of U.S. adults (N = 2,441). Participants were recruited through social media platforms such as Twitter, Instagram, and Facebook, were aged over 18 and currently residing in the U.S., and did not include essential service workers, who were excluded due to their need to leave the home for employment.

The dataset had 66 variables corresponding to the questionnaire questions. We chose to focus on the survey responses pertaining to (1) age, (2) country & (3) state of residence, (4) race, (5) sex, (6) if local stay-at-home orders existed, (7) if the participant stayed home even if no order existed or (8) even if they didn't know if an order existed, (9) how the participant protected themselves in public, (10) reasons for leaving home during the order, (11) average hours per day spent at home during the pandemic, (12) if the participant had contracted COVID, (13) if anyone in the household had contracted COVID, (14) if any close friends had contracted COVID, (15) if the participant lived in an urban, suburban, or rural area, (16) whether the participant had been tested for COVID, (17) educational attainment, and (18) annual income. Each participant/observation was identified by a unique participant ID.

Glimpse

```
glimpse(tidy_data)
```

```
## Rows: 1,863
## Columns: 31
## $ Participant_ID      <int> 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16~
## $ age                 <int> 27, 26, 27, 23, 24, 40, 36, 35, 28, 36, 31, 31, 55~
## $ usres               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ state               <int> 44, 44, 44, 38, 44, 34, 44, 7, 44, 26, 48, 44, 44,~
## $ race                <int> 5, 4, 4, 5, 1, 4, 5, 4, 4, 4, 4, 4, 4, 1, 6, 4, 4,~
## $ sex                 <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1,~
## $ localsip            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ localsip2           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ localsip3           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ leavehomeact___1    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,~
## $ leavehomeact___2    <int> 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0,~
## $ leavehomeact___3    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ leavehomeact___4    <int> 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1,~
## $ leavehomeact___5    <int> 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,~
## $ leavehomeact___6    <int> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1,~
## $ leavehomeact___7    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ leavehomereason___1 <int> 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0,~
## $ leavehomereason___2 <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,~
## $ leavehomereason___3 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1,~
## $ leavehomereason___4 <int> 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1,~
## $ leavehomereason___5 <int> 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0,~
```

```
## $ leavehomereason___6 <int> 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1,~
## $ leavehomereason___7 <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0,~
## $ localsiphours       <int> 14, 23, 24, 14, 24, 24, 23, 24, 24, 22, 24, 20, 22~
## $ covidsick           <int> 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ hhcovidsick         <int> 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ ffcovidsick         <int> 2, 3, 1, 2, 3, 1, 1, 4, 3, 4, 2, 2, 4, 1, 4, 2, 2,~
## $ Classification      <chr> "Urban", "Urban", "Suburban", "Rural", "Urban", "R~
## $ covidtest           <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ educ                <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 4, 6, 6, 6, 6,~
## $ hhincome            <int> 12, 11, 11, 5, 3, 7, 3, 6, 12, 12, 12, 12, 12, 12,~
```

Data Analysis Plan

We excluded people who did not respond to race from the dataset. In addition, none of the participants indicated that they had completed no schooling or had not completed grades 1-8. We created no new variables.

In order to explore the relationships between certain demographic characteristics and hours stayed at home during the pandemic, we will conduct multiple two-sample t-tests comparing the mean number of hours spent at home during the pandemic between different races (e.g. Asian vs. non-Asian), levels of education, and income levels, among others. We also plan on constructing 95% confidence intervals regarding the number of hours spent at home for populations with and without formal stay-at-home orders. In addition, we plan on visualizing the most frequently cited methods of protection from COVID used when in public as well as reasons for leaving home during a stay-in-place order.

At present, we hypothesize that Asian people will differ in their mean hours spent at home compared to white people, as they had 0.7 times the hospitalization rate of white people (CDC, 2020). We also hypothesize that people with incomes over $150,000 will have different/greater average hours spent at home due to having access to grocery and meal delivery services such as InstaCart, which would decrease the number of hours needed to be spent outside. This is further supported by the data, where grocery shopping was the highest cited reason for leaving the home and by reports that meal delivery services had increased by approximately 70% in March 2020 (Hobbs, 2020). To achieve these results, we would need significant p-values of under 0.05 from our t-tests. The 95% confidence intervals for mean hours spent at home should also completely overlap infrequently. The table and graph below give us a preliminary idea of the differences in average hours spent home among different populations such as urban vs. rural and between different races.

References:

Burford, K. G., 2020, “Replication Data for: Associations of Urbanicity and Sociodemographic Characteristics with Protective Health Behaviors and Reasons for Leaving the Home during COVID-19”, https://doi.org/10.7910/DVN/7FA07D, Harvard Dataverse, V3

Centers for Disease Control and Prevention. (2020, September 3). Timing of state and territorial COVID-19 stay-at-home orders and changes in population movement - United States, March 1–May 31, 2020. Centers for Disease Control and Prevention.

Hobbs, J. E., 2020, “Food supply chains during the COVID-19 pandemic”, https://doi.org/10.1111/cjag.12237, Canadian Journal of Agricultural Economics,

Thunström, L., Newbold, S. C., Finnoff, D., Ashworth, M., & Shogren, J. F. (2020, May 21). The

benefits and costs of using social distancing to flatten the curve for covid-19: Journal of Benefit-Cost Analysis. Cambridge.

```
tidy_data$race[tidy_data$race == 0] <- "Native American"
tidy_data$race[tidy_data$race == 1] <- "Asian"
tidy_data$race[tidy_data$race == 2] <- "Hawaiian"
tidy_data$race[tidy_data$race == 3] <- "African American"
tidy_data$race[tidy_data$race == 4] <- "White"
tidy_data$race[tidy_data$race == 5] <- "Mixed"
tidy_data$race[tidy_data$race == 6] <- "Unknown"

number_of_hours <- tidy_data %>%
  group_by(race) %>%
  summarise_at(vars(localsiphours), list(hours = mean),na.rm = TRUE) #to summarize count

number_of_hours_two <- tidy_data %>%
  group_by(Classification) %>%
  summarise_at(vars(localsiphours), list(hours = mean),na.rm = TRUE) %>% #to summarize count
  print()
```

```
## # A tibble: 4 x 2
##   Classification hours
##   <chr>          <dbl>
## 1 Rural           21.6
## 2 Suburban        21.3
## 3 Urban           21.2
## 4 <NA>            17.5
```

```
tidy_data$Classification[is.na(tidy_data$Classification)== TRUE] <- "Urban"

reasons <-c("Work", "Provide Care for Others", "Grocery Shopping", "Essential Shopping", "Exer
freq_reasons <- c(sum(tidy_data$leavehomereason___1, na.rm = TRUE), sum(tidy_data$leavehomerea
reasons_for_leaving = data.frame(reasons, freq_reasons)

print(reasons_for_leaving)
```

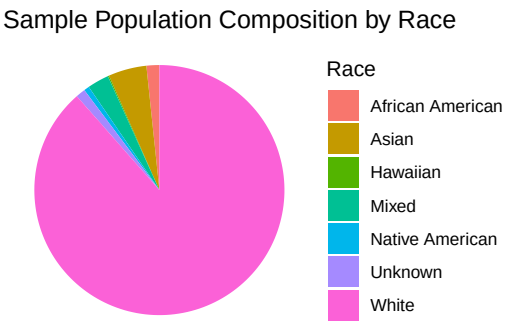
```
##                   reasons freq_reasons
## 1                    Work          534
## 2 Provide Care for Others          212
## 3        Grocery Shopping         1620
## 4      Essential Shopping          729
## 5                Exercise         1241
## 6                Walk Dog          781
## 7                   Other          223
```

```
pie_chart <- tidy_data %>%
  group_by(race) %>%
  count() %>%
  ungroup() %>%
```

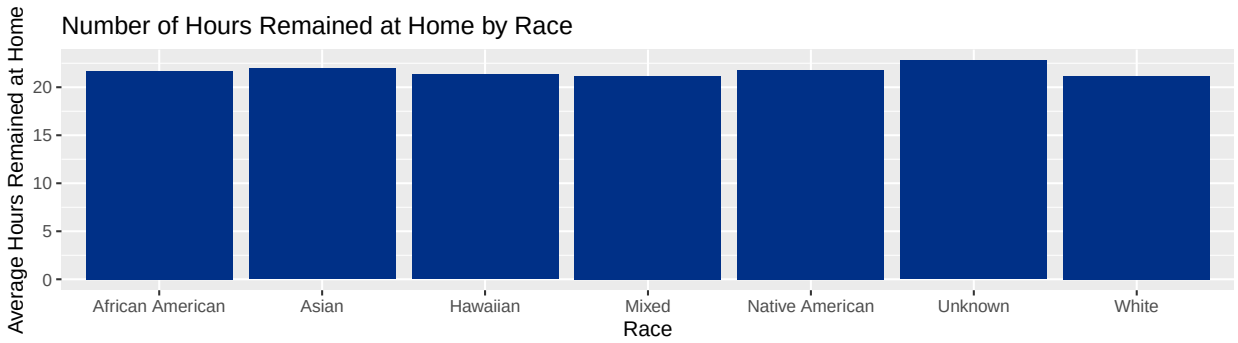
```
  mutate(perc = n / sum(n))

print(pie_chart)
```

```
## # A tibble: 7 x 3
##   race                n    perc
##   <chr>           <int>   <dbl>
## 1 African American    31 0.0166
## 2 Asian               92 0.0494
## 3 Hawaiian             3 0.00161
## 4 Mixed               52 0.0279
## 5 Native American     13 0.00698
## 6 Unknown             23 0.0123
## 7 White             1649 0.885
```

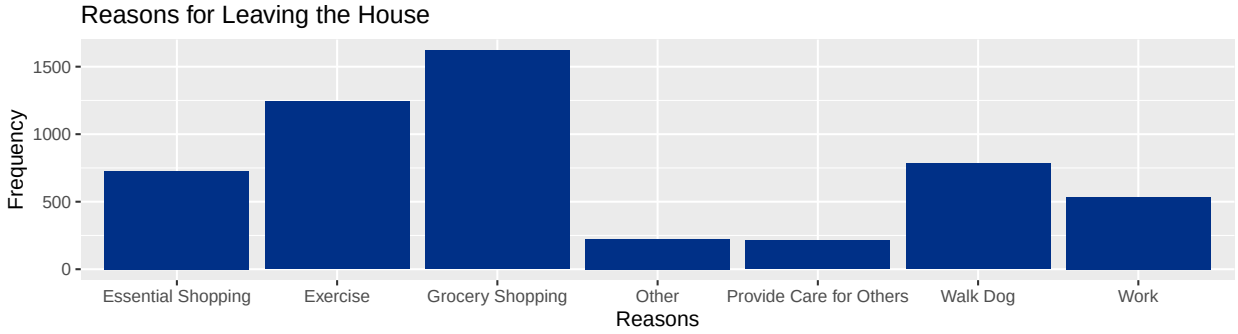
```
# pie chart
ggplot(pie_chart, aes(x="", y=perc, fill=race)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)+
  labs (
    fill = "Race",
    title = "Sample Population Composition by Race",
    ) +
  theme_void()
```




```
# bar graphs for race
ggplot(data=number_of_hours, aes(x=race, y=hours)) +
  geom_bar(stat="identity", fill = "#003087") +
  labs (
    y = "Average Hours Remained at Home",
    x = "Race",
    title = "Number of Hours Remained at Home by Race",
    )
```




```
ggplot(data=reasons_for_leaving, aes(x=reasons, y=freq_reasons)) +
  geom_bar(stat="identity", fill = "#003087") +
  labs (
    y = "Frequency",
    x = "Reasons",
    title = "Reasons for Leaving the House",
    )
```




```
# preprocess data for t-tests
tidy_data <- tidy_data %>%
  mutate(asian = ifelse(race == "Asian", 1, 0)) %>%
  mutate(white = ifelse(race == "White", 1, 0)) %>%
  mutate(unknown = ifelse(race == "Unknown", 1, 0)) %>%
  mutate(africanamerican = ifelse(race == "African American", 1, 0)) %>%
  mutate(americanindian = ifelse(race == "Native American", 1, 0)) %>%
  mutate(mixed = ifelse(race == "Mixed", 1, 0)) %>%
  mutate(hawaiian = ifelse(race == "Hawaiian", 1, 0)) %>%
  mutate(noschool = ifelse(educ == 1, 1, 0)) %>%
  mutate(g1_8 = ifelse(educ == 2, 1, 0)) %>%
  mutate(g9_11 = ifelse(educ == 3, 1, 0)) %>%
  mutate(g12 = ifelse(educ == 4, 1, 0)) %>%
  mutate(technical_college = ifelse(educ == 5, 1, 0)) %>%
  mutate(four_years_college = ifelse(educ == 6, 1, 0)) %>%
  mutate(poorest = ifelse(hhincome == 1, 1, 0)) %>%
  mutate(richest = ifelse(hhincome == 12, 1, 0))

# t-test by EDUCATION insignificant p-value (do not reject null hypothesis)
```

```r
# t.test(tidy_data$localsiphours~tidy_data$noschool, var.equal=FALSE)
# t.test(tidy_data$localsiphours~tidy_data$g1_8, var.equal=FALSE)
t.test(tidy_data$localsiphours~tidy_data$g9_11, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$g9_11
## t = 0.92639, df = 1.0016, p-value = 0.5241
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -123.2403  142.7021
## sample estimates:
## mean in group 0 mean in group 1
##        21.23089        11.50000
```

```r
t.test(tidy_data$localsiphours~tidy_data$g12, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$g12
## t = -0.73531, df = 44.873, p-value = 0.466
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -2.185161  1.016422
## sample estimates:
## mean in group 0 mean in group 1
##        21.20975        21.79412
```

```r
t.test(tidy_data$localsiphours~tidy_data$technical_college, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$technical_college
## t = 2.3857, df = 957.77, p-value = 0.01724
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  0.1994046 2.0485164
## sample estimates:
## mean in group 0 mean in group 1
##        21.38056        20.25660
```

```r
t.test(tidy_data$localsiphours~tidy_data$four_years_college, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$four_years_college
```

```
## t = -2.1467, df = 1198.5, p-value = 0.03202
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -1.89997377 -0.08543864
## sample estimates:
## mean in group 0 mean in group 1
##        20.38944        21.38215
```

```
# t-test by RACE insignificant p-value (do not reject null hypothesis)
t.test(tidy_data$localsiphours~tidy_data$asian, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$asian
## t = -1.4506, df = 218.19, p-value = 0.1483
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -1.7639892  0.2682241
## sample estimates:
## mean in group 0 mean in group 1
##        21.18690        21.93478
```

```
t.test(tidy_data$localsiphours~tidy_data$white, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$white
## t = 1.5607, df = 1257.6, p-value = 0.1189
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -0.1629826  1.4310749
## sample estimates:
## mean in group 0 mean in group 1
##        21.78505        21.15100
```

```
t.test(tidy_data$localsiphours~tidy_data$africanamerican, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$africanamerican
## t = -0.8345, df = 59.915, p-value = 0.4073
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -1.5669406  0.6444163
## sample estimates:
## mean in group 0 mean in group 1
##        21.21616        21.67742
```

```
t.test(tidy_data$localsiphours~tidy_data$americanindian, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$americanindian
## t = -0.52001, df = 14.14, p-value = 0.6111
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -2.812414  1.713952
## sample estimates:
## mean in group 0 mean in group 1
##        21.22000        21.76923
```

```
t.test(tidy_data$localsiphours~tidy_data$mixed, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$mixed
## t = 0.091563, df = 98.419, p-value = 0.9272
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -1.079356  1.183782
## sample estimates:
## mean in group 0 mean in group 1
##        21.22529        21.17308
```

```
t.test(tidy_data$localsiphours~tidy_data$hawaiian, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$hawaiian
## t = -0.04088, df = 2.0492, p-value = 0.971
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -11.39162  11.17227
## sample estimates:
## mean in group 0 mean in group 1
##        21.22366        21.33333
```

```
# t-test by INCOME LEVEL
t.test(tidy_data$localsiphours~tidy_data$poorest, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$poorest
## t = -0.30494, df = 44.42, p-value = 0.7618
```

```
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -1.452853  1.070890
## sample estimates:
## mean in group 0 mean in group 1
##        21.21643        21.40741
```

```
t.test(tidy_data$localsiphours~tidy_data$richest, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$richest
## t = 0.63165, df = 1381.7, p-value = 0.5277
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -0.6866933  1.3389351
## sample estimates:
## mean in group 0 mean in group 1
##        21.34941        21.02329
```

```
# t-test by RACE significant p-value (reject null hypothesis)
t.test(tidy_data$localsiphours~tidy_data$unknown, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$unknown
## t = -4.2165, df = 158.94, p-value = 4.153e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -2.3175201 -0.8390017
## sample estimates:
## mean in group 0 mean in group 1
##        21.20435        22.78261
```

```
# t-test by SEX significant p-value (reject null hypothesis)
t.test(tidy_data$localsiphours~tidy_data$sex, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  tidy_data$localsiphours by tidy_data$sex
## t = -3.5035, df = 1809.3, p-value = 0.0004703
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to
## 95 percent confidence interval:
##  -2.5981696 -0.7332444
## sample estimates:
## mean in group 1 mean in group 2
##        20.09075        21.75646
```

```r
# ANOVA with UNKOWN
summary(aov(localsiphours~race,data=tidy_data))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race           6    122   20.26   0.125  0.993
## Residuals   1856 300760  162.05
```

```r
summary(aov(localsiphours~state,data=tidy_data))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## state          1    198   198.1   1.216   0.27
## Residuals   1845 300564   162.9
## 16 observations deleted due to missingness
```

```r
# filter out unkonwn and observe results
tidy_data_without_unknown <- tidy_data %>%
  filter(unknown ==0)

# ANOVA test
summary(aov(localsiphours~race,data=tidy_data_without_unknown))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race           5     65   12.99   0.079  0.995
## Residuals   1834 300734  163.98
```

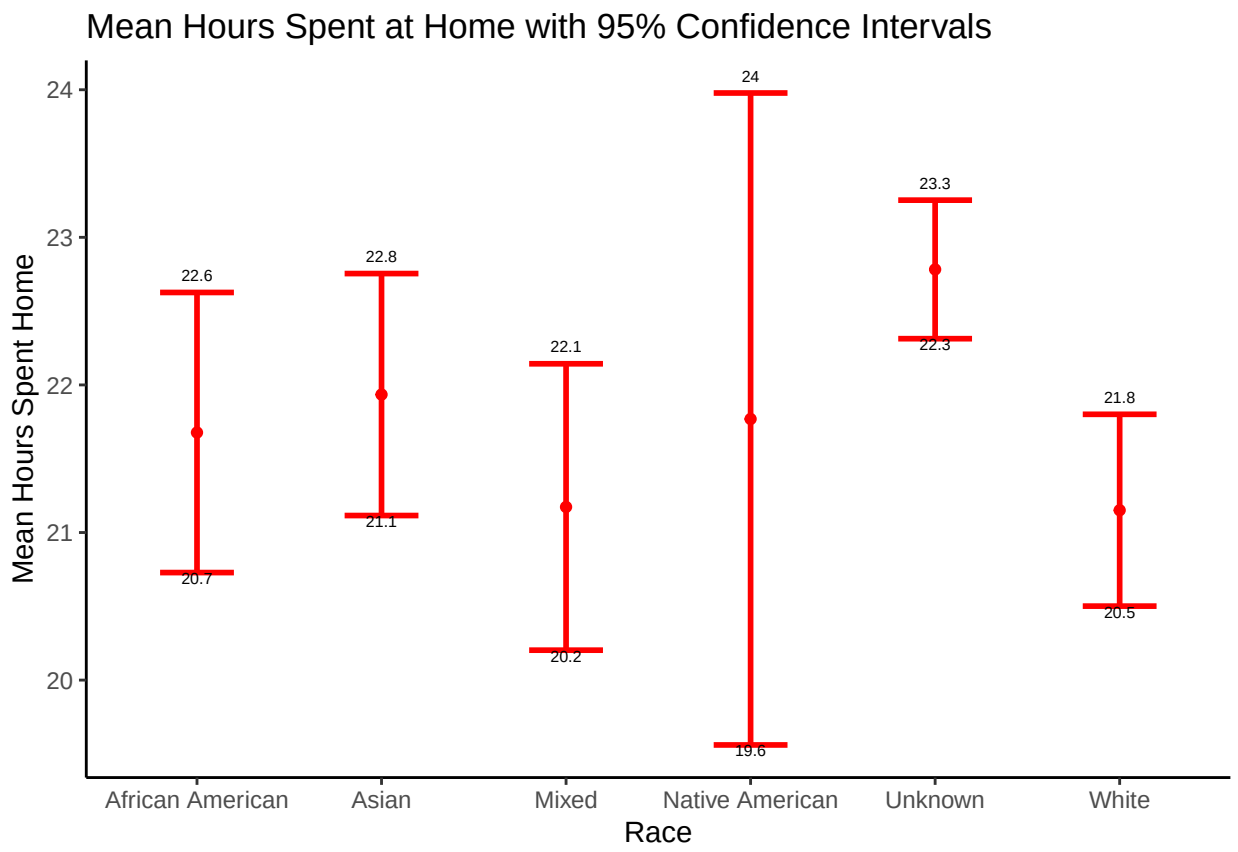
```r
summary(aov(localsiphours~state,data=tidy_data_without_unknown))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## state          1    188   187.5   1.137  0.286
## Residuals   1822 300491   164.9
## 16 observations deleted due to missingness
```

```r
# 95% confidence interval by race
dt <- tidy_data%>%
  group_by(race)%>%
  filter(race != "Hawaiian")%>%
  summarise(
    mean = mean(localsiphours),
    lci = t.test(localsiphours, conf.level = 0.95)$conf.int[1],
    uci = t.test(localsiphours, conf.level = 0.95)$conf.int[2])
dt
```

```
## # A tibble: 6 x 4
##   race              mean   lci   uci
##   <chr>            <dbl> <dbl> <dbl>
## 1 African American  21.7  20.7  22.6
## 2 Asian             21.9  21.1  22.8
## 3 Mixed             21.2  20.2  22.1
## 4 Native American   21.8  19.6  24.0
## 5 Unknown           22.8  22.3  23.3
## 6 White             21.2  20.5  21.8
```

```
pl2 <- ggplot(data = dt)
pl2 <- pl2 + geom_point(aes(x=race, y=mean), color= "red")
pl2 <- pl2 + geom_errorbar(aes(x=race, ymin=lci, ymax= uci), width = 0.4, color ="red", size =
pl2 <- pl2 + geom_text(aes(x=race, y=lci, label = round(lci,1)), size= 2, vjust = 1)
pl2 <- pl2 + geom_text(aes(x=race, y=uci, label = round(uci,1)), size= 2, vjust = -1)
pl2 <- pl2 + theme_classic()
pl2 <- pl2 + labs(title = "Mean Hours Spent at Home with 95% Confidence Intervals")
pl2 <- pl2 + labs(x= "Race", y = "Mean Hours Spent Home")
pl2
```

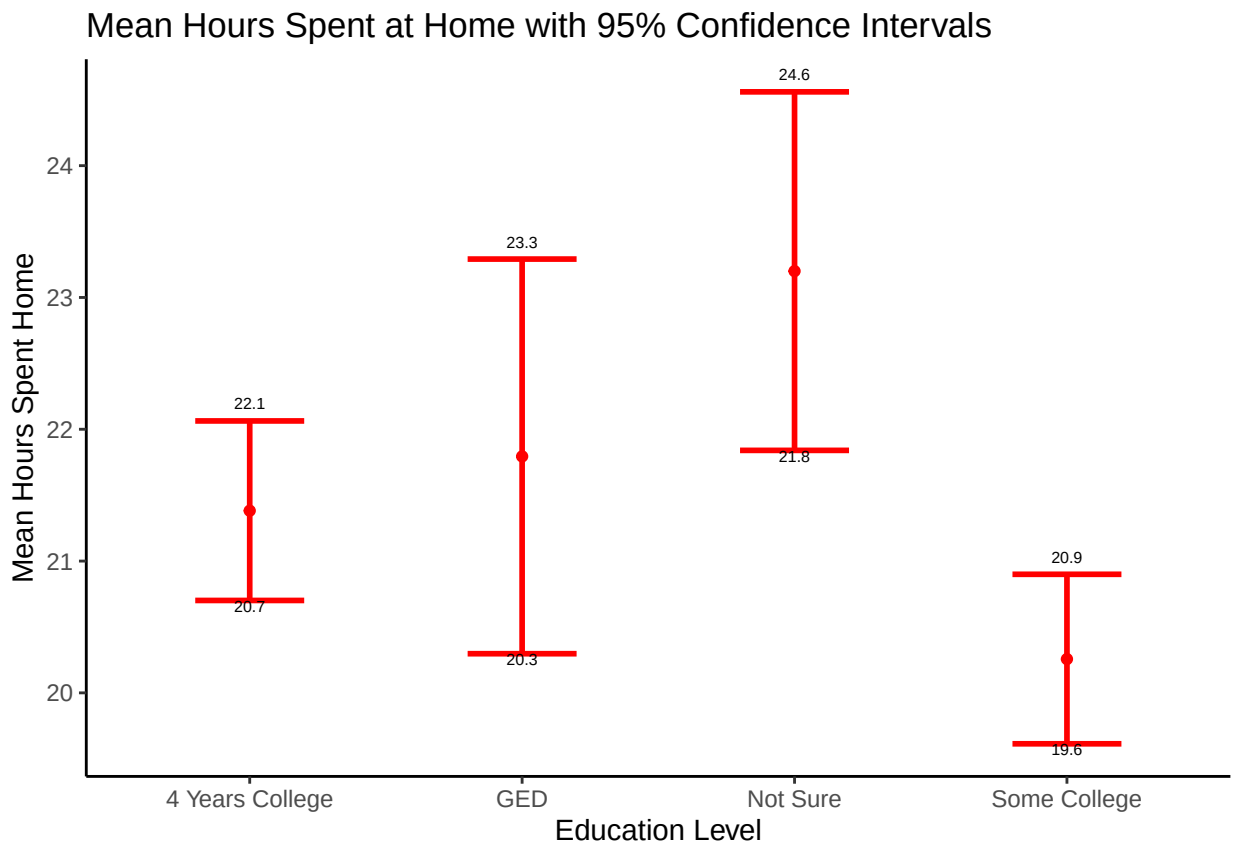



```
# 95% confidence interval by education
tidy_data$educ[tidy_data$educ == 1] <- "No School"
tidy_data$educ[tidy_data$educ == 2] <- "G1-8"
tidy_data$educ[tidy_data$educ == 3] <- "G9-12"
tidy_data$educ[tidy_data$educ == 4] <- "GED"
tidy_data$educ[tidy_data$educ == 5] <- "Some College"
tidy_data$educ[tidy_data$educ == 6] <- "4 Years College"
tidy_data$educ[tidy_data$educ == 7] <- "Not Sure"
tidy_data$educ[is.na(tidy_data$educ)] <- "Not Sure"

dt <- tidy_data%>%
  group_by(educ)%>%
  filter(educ != "G9-12")%>%
```

```
  summarise(
    mean = mean(localsiphours),
    lci = t.test(localsiphours, conf.level = 0.95)$conf.int[1],
    uci = t.test(localsiphours, conf.level = 0.95)$conf.int[2])
dt
```

```
## # A tibble: 4 x 4
##   educ             mean   lci   uci
##   <chr>           <dbl> <dbl> <dbl>
## 1 4 Years College  21.4  20.7  22.1
## 2 GED              21.8  20.3  23.3
## 3 Not Sure         23.2  21.8  24.6
## 4 Some College     20.3  19.6  20.9
```

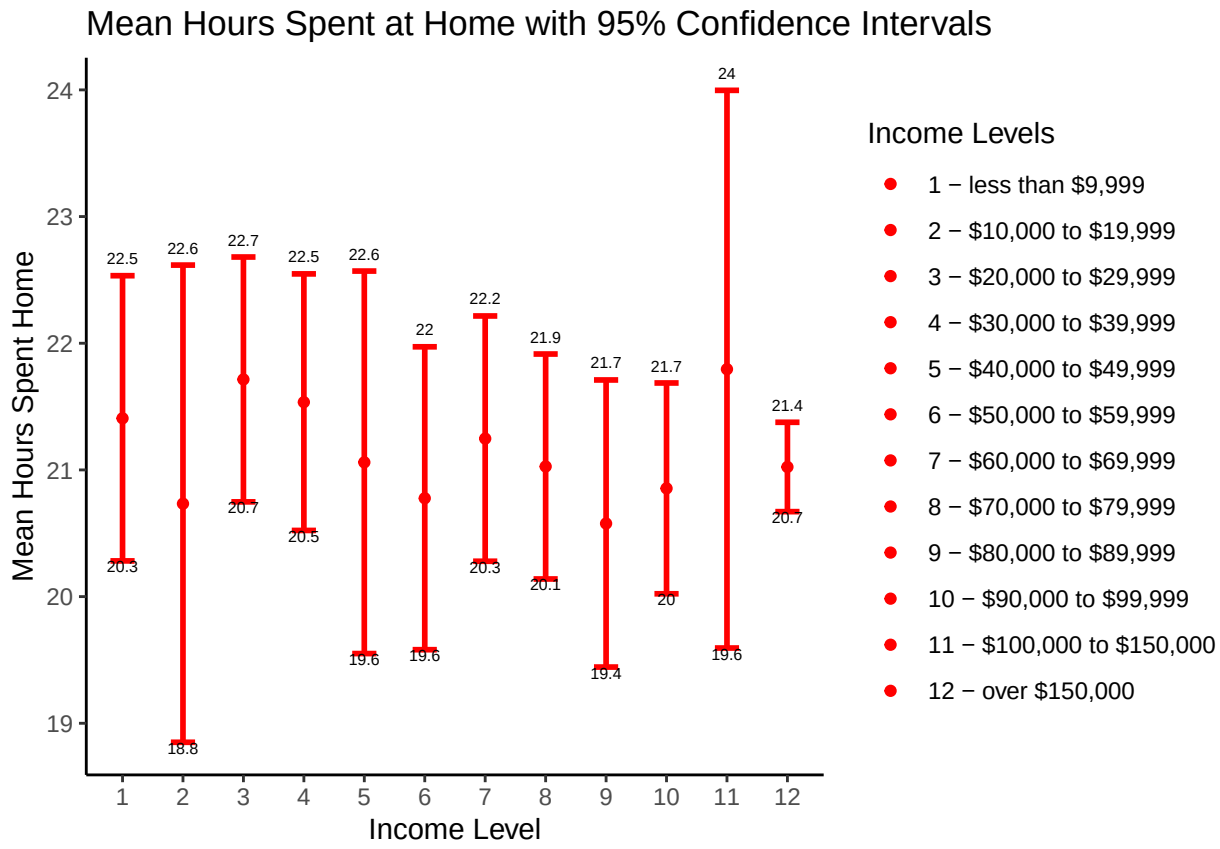
```
pl2 <- ggplot(data = dt)
pl2 <- pl2 + geom_point(aes(x=educ, y=mean), color= "red")
pl2 <- pl2 + geom_errorbar(aes(x=educ, ymin=lci, ymax= uci), width = 0.4, color ="red", size =
pl2 <- pl2 + geom_text(aes(x=educ, y=lci, label = round(lci,1)), size= 2, vjust = 1)
pl2 <- pl2 + geom_text(aes(x=educ, y=uci, label = round(uci,1)), size= 2, vjust = -1)
pl2 <- pl2 + theme_classic()
pl2 <- pl2 + labs(title = "Mean Hours Spent at Home with 95% Confidence Intervals")
pl2 <- pl2 + labs(x= "Education Level", y = "Mean Hours Spent Home")
pl2
```



```r
# 95% confidence interval by income level
dt <- tidy_data%>%
  group_by(hhincome)%>%
  filter(is.na(hhincome) == FALSE)%>%
  summarise(
    mean = mean(localsiphours),
    lci = t.test(localsiphours, conf.level = 0.95)$conf.int[1],
    uci = t.test(localsiphours, conf.level = 0.95)$conf.int[2])
dt
```

```
## # A tibble: 12 x 4
##    hhincome  mean   lci   uci
##       <int> <dbl> <dbl> <dbl>
##  1        1  21.4  20.3  22.5
##  2        2  20.7  18.8  22.6
##  3        3  21.7  20.7  22.7
##  4        4  21.5  20.5  22.5
##  5        5  21.1  19.6  22.6
##  6        6  20.8  19.6  22.0
##  7        7  21.2  20.3  22.2
##  8        8  21.0  20.1  21.9
##  9        9  20.6  19.4  21.7
## 10       10  20.9  20.0  21.7
## 11       11  21.8  19.6  24.0
## 12       12  21.0  20.7  21.4
```

```r
pl2 <- ggplot(data = dt)
pl2 <- pl2 + geom_point(aes(x=as.factor(hhincome), y=mean, fill = as.factor(hhincome)), color=
pl2 <- pl2 + geom_errorbar(aes(x=hhincome, ymin=lci, ymax= uci), width = 0.4, color ="red", si
pl2 <- pl2 + geom_text(aes(x=hhincome, y=lci, label = round(lci,1)), size= 2, vjust = 1)
pl2 <- pl2 + geom_text(aes(x=hhincome, y=uci, label = round(uci,1)), size= 2, vjust = -1)
pl2 <- pl2 + theme_classic()
pl2 <- pl2 + labs(title = "Mean Hours Spent at Home with 95% Confidence Intervals")
pl2 <- pl2 + labs(x= "Income Level", y = "Mean Hours Spent Home", fill = "okay")
pl2 <- pl2 + scale_fill_discrete(name = "Income Levels", labels = c("1 - less than $9,999", "2
pl2
```

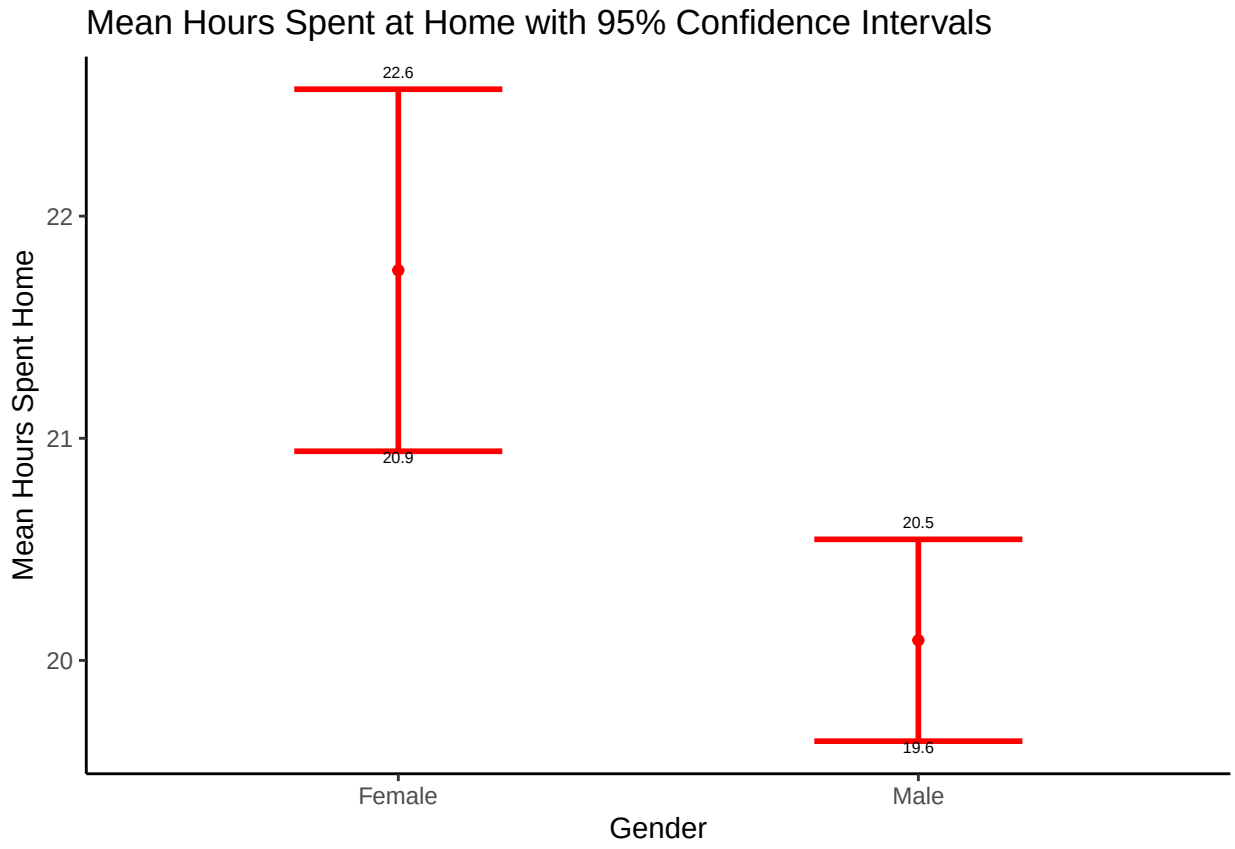



```
# 95% confidence interval by gender
tidy_data$sex[tidy_data$sex == 1] <- "Male"
tidy_data$sex[tidy_data$sex == 2] <- "Female"
dt <- tidy_data%>%
  filter(is.na(sex) == FALSE) %>%
  group_by(sex)%>%
  summarise(
    mean = mean(localsiphours),
    lci = t.test(localsiphours, conf.level = 0.95)$conf.int[1],
    uci = t.test(localsiphours, conf.level = 0.95)$conf.int[2])
dt
```

```
## # A tibble: 2 x 4
##   sex     mean   lci   uci
##   <chr>  <dbl> <dbl> <dbl>
## 1 Female  21.8  20.9  22.6
## 2 Male    20.1  19.6  20.5
```

```
pl2 <- ggplot(data = dt)
pl2 <- pl2 + geom_point(aes(x=as.factor(sex), y=mean), color= "red")
pl2 <- pl2 + geom_errorbar(aes(x=sex, ymin=lci, ymax= uci), width = 0.4, color ="red", size = 1
pl2 <- pl2 + geom_text(aes(x=sex, y=lci, label = round(lci,1)), size= 2, vjust = 1)
pl2 <- pl2 + geom_text(aes(x=sex, y=uci, label = round(uci,1)), size= 2, vjust = -1)
pl2 <- pl2 + theme_classic()
```

```
pl2 <- pl2 + labs(title = "Mean Hours Spent at Home with 95% Confidence Intervals")
pl2 <- pl2 + labs(x= "Gender", y = "Mean Hours Spent Home")
pl2
```




```
# fit linear regression model with Education
localsiphours_fit_education <- linear_reg() %>%
  set_engine("lm") %>%
  fit(localsiphours ~ g9_11 + technical_college + four_years_college, data = tidy_data)

# fit linear regression model with UNKNOWN
localsiphours_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(localsiphours ~ asian + white + africanamerican + americanindian + mixed + hawaiian, data

#fit linear regression model without UNKNOWN
localsiphours_fit_without_unknown <- linear_reg() %>%
  set_engine("lm") %>%
  fit(localsiphours ~ asian + white + africanamerican + americanindian + mixed + hawaiian, data

# fit linear regressin model
localsiphours_fit_age <- linear_reg() %>%
  set_engine("lm") %>%
  fit(localsiphours ~ age, data = tidy_data)
```

```
#fit logistic regression model without Association Terms
localsiphours_fit_without_unknown_association <- linear_reg() %>%
  set_engine("lm") %>%
  fit(localsiphours ~ asian + white + unknown + africanamerican + americanindian + mixed + hawa

tidy(localsiphours_fit, conf.int=TRUE, exponentiate = TRUE)
```

```
## # A tibble: 7 x 7
##   term            estimate std.error statistic  p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)       22.8        2.65     8.58  1.92e-17    17.6      28.0
## 2 asian             -0.848      2.97    -0.286 7.75e- 1    -6.67      4.97
## 3 white             -1.63       2.67    -0.610 5.42e- 1    -6.87      3.61
## 4 africanamerican   -1.11       3.50    -0.315 7.52e- 1    -7.98      5.77
## 5 americanindian    -1.01       4.42    -0.229 8.19e- 1    -9.68      7.65
## 6 mixed             -1.61       3.19    -0.505 6.14e- 1    -7.86      4.64
## 7 hawaiian          -1.45       7.81    -0.185 8.53e- 1   -16.8      13.9
```

```
tidy(localsiphours_fit_without_unknown, conf.int=TRUE, exponentiate = TRUE)
```

```
## # A tibble: 7 x 7
##   term            estimate std.error statistic  p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)       21.3        7.39    2.89    0.00395     6.83      35.8
## 2 asian              0.601      7.51    0.0801  0.936     -14.1       15.3
## 3 white             -0.182      7.40   -0.0246  0.980     -14.7       14.3
## 4 africanamerican    0.344      7.74    0.0444  0.965     -14.8       15.5
## 5 americanindian     0.436      8.20    0.0531  0.958     -15.7       16.5
## 6 mixed             -0.160      7.60   -0.0211  0.983     -15.1       14.8
## 7 hawaiian          NA         NA      NA      NA          NA         NA
```

```
tidy(localsiphours_fit_education, conf.int=TRUE, exponentiate = TRUE)
```

```
## # A tibble: 4 x 7
##   term               estimate std.error statistic  p.value conf.low conf.high
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)          21.9        2.12    10.3   2.81e-24    17.7      26.0
## 2 g9_11               -10.4        9.24    -1.12  2.62e- 1   -28.5       7.76
## 3 technical_college    -1.60       2.26    -0.710 4.78e- 1    -6.04      2.83
## 4 four_years_college   -0.479      2.14    -0.223 8.23e- 1    -4.69      3.73
```

```
tidy(localsiphours_fit_age, conf.int=TRUE, exponentiate = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  20.1       0.971      20.7  1.08e-85  18.2      22.0
## 2 age           0.0271    0.0220      1.23 2.18e- 1  -0.0161    0.0703
```

```
tidy(localsiphours_fit_without_unknown_association, conf.int=TRUE, exponentiate = TRUE)
```

```
## # A tibble: 8 x 7
##   term            estimate std.error statistic  p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)       21.3        7.35    2.90    0.00374     6.92      35.7
## 2 asian              0.601      7.47    0.0805  0.936     -14.0       15.2
## 3 white             -0.182      7.36   -0.0248  0.980     -14.6       14.2
## 4 unknown            1.45       7.81    0.185   0.853     -13.9       16.8
## 5 africanamerican    0.344      7.70    0.0447  0.964     -14.8       15.4
## 6 americanindian     0.436      8.15    0.0535  0.957     -15.6       16.4
## 7 mixed             -0.160      7.56   -0.0212  0.983     -15.0       14.7
## 8 hawaiian          NA         NA      NA      NA          NA         NA
```

Conclusion

Limitations The sample over-represented Hispanic non-white individuals while under-representing other races such as Black and Asian people. Thus, our results may be skewed due to having small sample sizes for certain demographics.