Project Proposal

due October 11, 2021 by 11:59 PM

Grace Lee and Ji Yun Hyo, TBD3

October 11th, 2021

Load Packages

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

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 (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. 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. The data was collected through an online questionnaire of U.S. adults (N = 2,441) recruited through social media platforms such as Twitter, Instagram, and Facebook. 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

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. White), 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 essential workers, whose incomes are lower on average, being required to leave home for work. 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: 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.

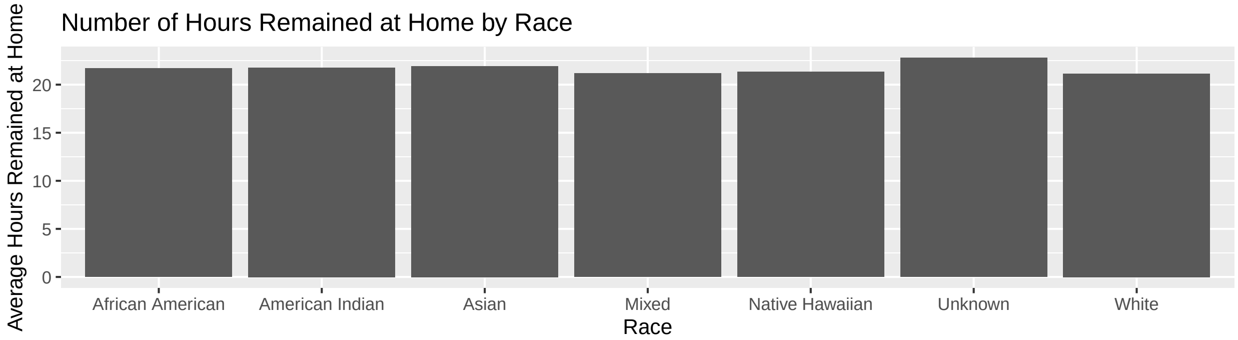
Thunström, L., Newbold, S. C., Finnoff, D., Ashworth, M., & Shogren, J. F. (2020, May 21). The benefits and costs of using social distancing to flatten the curve for covid-19: Journal of Benefit-Cost Analysis. Cambridge.

```r
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
```

```r
tidy_data$Classification[is.na(tidy_data$Classification)== TRUE] <- "Urban"
number_of_hours$race[number_of_hours$race == 0] <- "American Indian"
```

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

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



```r
tidy_data <- tidy_data %>%
  mutate(asian = ifelse(race == 1, 1, 0)) %>%
  mutate(white = ifelse(race == 4, 1, 0)) %>%
  mutate(unknown = ifelse(race == 6, 1, 0)) %>%
  mutate(africanamerican = ifelse(race == 3, 1, 0)) %>%
  mutate(americanindian = ifelse(race == 0, 1, 0)) %>%
  mutate(mixed = ifelse(race == 5, 1, 0)) %>%
  mutate(hawaiian = ifelse(race == 2, 1, 0))

# 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 t
## 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 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
```

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

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race           1     24   24.46   0.151  0.697
## Residuals   1861 300857  161.66
```

```
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
```

```
# 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           1     57   57.08   0.349  0.555
## Residuals   1838 300742  163.62
```

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

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

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
```