GEOG 111A Final Synthesis Lab

Fall 2021

Due Sunday Dec. 5 at 23:00 PST

Your Name: *Matthew Mangawang*

When you knit the .Rmd to a PDF for the first time, you may need to run `tinytex::install_tinytex()` in the console.

Load the packages and data

The packages and data are loaded for you. At this point, you should be quite familiar with the CHTS data. You can look up for variables in the `DataDictionary.xlsx`.

Load the packages

```
library(tidyverse)
library(maps)
```

Load the data

```
PersonData <- read_rds('~/GEOG 111A Data/Data/PersonData_111A.Rds')
HHData <- read_rds('~/GEOG 111A Data/Data/HHData_111A.Rds')
hh_bgDensity <- read_rds('~/GEOG 111A Data/Data/hh_bgDensity.Rds')
personHHData <- left_join(PersonData, HHData) %>% left_join(hh_bgDensity)
```

Questions

Answer the following questions with code and texts (if asked). There are many different ways to write code to achieve the same goal, you are supposed to provide only ONE code solution. Only the first solution will be graded if you include multiple solutions.

1. What are the average PMT and number of trips for students? (5 points) (*Hint: Review Lab 3*)

```
PersonData %>%
  filter(Student == 1) %>%
  summarize(studentPMT = mean(Sum_PMT),
            studentTrips = mean(Sum_Trips)) %>%
  View()
```

Average PMT for students: **20.65**

Average number of trips for students: **3.04**

2. Write the code to return the number of employed males who work from home. (4 points) (*Hint: Review Lab 3*)

```
PersonData %>%
  filter(Male == 1,
         Employed == 1,
         WorkHome == 1) %>%
  summarise(MenWFH = n()) %>%
  View()
```

The number of employed males are working from home is **4293**.

3. Write the code that outputs four columns: household id, person number, number of bike trips, and a column that suggests whether or not the person made at least one bike trip on the survey day. The value would be TRUE if the person made at least one bike trip and would be FALSE otherwise. (4 points) (*Hint: Review Lab 4*)

```
q3Columns <- select(PersonData, hhid, pnum, Bike_trips, WorkMode, workday)
q3Columns %>%
  select(hhid, pnum, Bike_trips, WorkMode) %>%
  mutate(BikedSurveyDay = WorkMode == "Bike / Nonmotorized" && Bike_trips > 0) %>%
  select (-WorkMode)
```

```
## # A tibble: 94,901 x 4
##       hhid  pnum Bike_trips BikedSurveyDay
##      <dbl> <dbl>      <dbl> <lgl>
##  1 1031985     1          0 FALSE
##  2 1031985     2          0 FALSE
##  3 1032036     1          0 FALSE
##  4 1032036     2          0 FALSE
##  5 1032036     3          0 FALSE
##  6 1032036     4          0 FALSE
##  7 1032036     5          0 FALSE
##  8 1032053     1          0 FALSE
##  9 1032053     2          0 FALSE
## 10 1032053     3          0 FALSE
## # ... with 94,891 more rows
```

4. How many people work from home and have a fixed schedule? How many people don't work from home and have a flexible schedule. Your code should return both results in **one output**. (6 points) (*Hint: Review Lab 4*)

```
wfhSchedule <- PersonData %>% select(WorkHome, FlexSched)
wfhSchedule %>%
  group_by(WorkHome, FlexSched) %>%
  summarise(How_Many = n())
```

```
## `summarise()` has grouped output by 'WorkHome'. You can override using the `.groups` argument.
```

```
## # A tibble: 9 x 3
## # Groups:   WorkHome [3]
```

```
##   WorkHome FlexSched How_Many
##      <dbl>     <dbl>    <int>
## 1        0         0    14475
## 2        0         1    22447
## 3        0         9       38
## 4        0        NA      365
## 5        1         0     1007
## 6        1         1     6813
## 7        1         9       22
## 8        1        NA       85
## 9       NA        NA    49649
```

The number of people who work from home and have a fixed schedule: **1007**

The number of people who don't work from home and have a flexible schedule: **22447**

5. Which county has the highest average number of trips by bike? Indicate the county name and its average number of trips by bike in the answer. (5 points) (*Hint: Review Lab 5*)

```
PersonDataQ5 <- left_join(PersonData, HHData)
```

```
## Joining, by = "hhid"
```

```
PersonDataQ5 %>%
  group_by(County) %>%
  summarise(n = n(),
            BikeTrips = mean(BikeTripsWk, na.rm = TRUE)) %>%
  View()
```

Inyo has the highest average number of trips by bike and it is **4.372**.

6. Make a stacked bar chart that shows how mode share based on the number of trips differs between male and female. Make a stacked bar chart of mode share based on the number of trips, with one bar corresponding to male and one to female. Add an appropriate title and legend, and briefly discuss the results. (You do not need to attach your final plots, as long as we can run your code and get the plot ourselves) (10 points) (*Hint: Review Lab 6*)

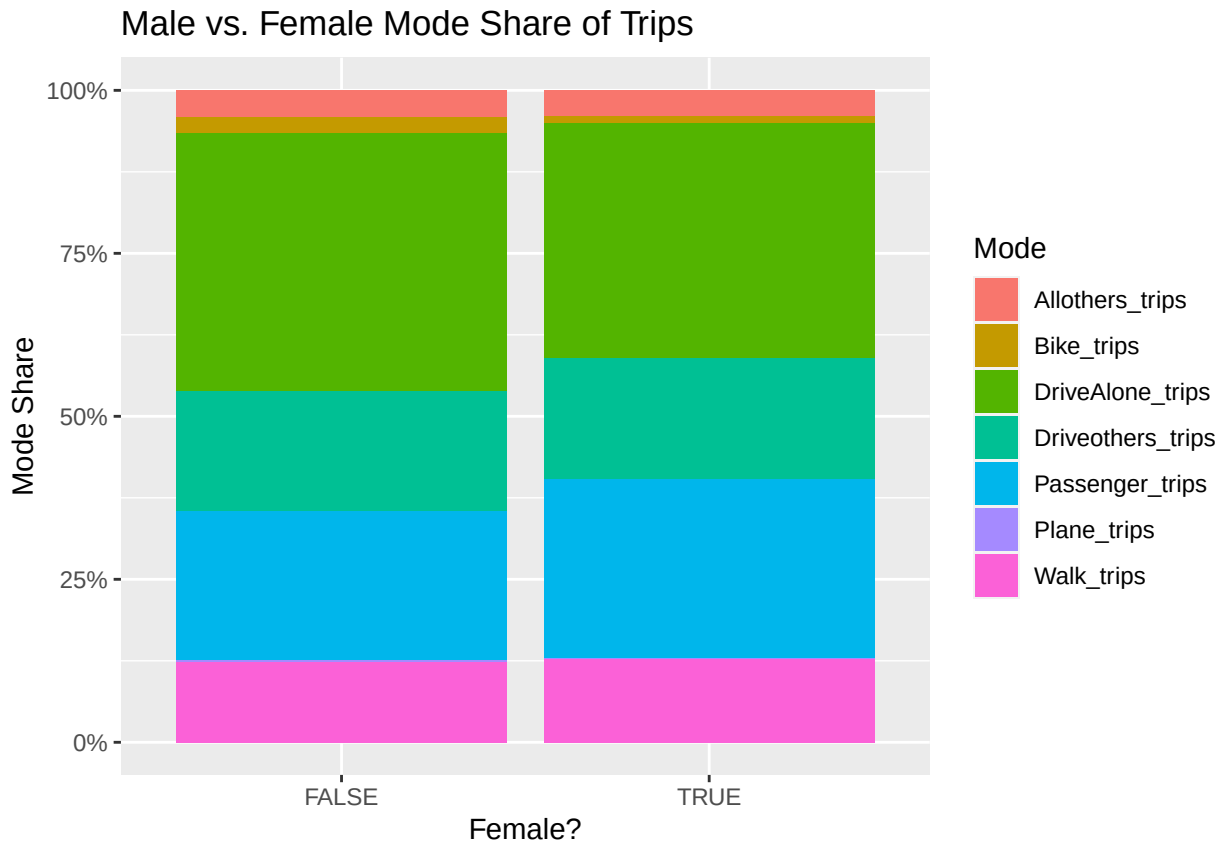
```
plotQ6 <- PersonData %>% select(Male, ends_with('trips'))
plotQ6 <- plotQ6 %>% select(-Sum_Trips)

plotQ6new <- plotQ6 %>% mutate(female = Male == 0)

plotQ6datasum <- plotQ6new %>% group_by(female) %>%
  summarize_at(vars(ends_with('trips')), sum)

plotQ6stacked <- plotQ6datasum %>% gather('Mode', 'trips', -female)

q6Plot <- plotQ6stacked %>% ggplot(aes(x = female, y = trips, fill = Mode))
q6Plot + geom_bar(stat = 'identity', position = 'fill') + scale_y_continuous(labels = scales::percent)
```



Discussion: **We can see that women have a higher mode share of passenger trips than men, while men have a slightly higher share of bike and drive alone trips versus women. All the other variables are pretty similar between genders.**

7. What is the absolute difference in the average travel distance by walking between the employed and the unemployed? (4 points) (*Hint: Review Lab 5*)

```
PersonData %>% filter(Employed == 1) %>%
  summarise(nPeople = n(),
            avgWalkDist = mean(Walk_Dist))
```

```
## # A tibble: 1 x 2
##   nPeople avgWalkDist
##     <int>       <dbl>
## 1   45252       0.194
```

```
PersonData %>% filter(Employed == 0) %>%
  summarise(nPeople = n(),
            avgWalkDist = mean(Walk_Dist))
```

```
## # A tibble: 1 x 2
##   nPeople avgWalkDist
##     <int>       <dbl>
## 1   32383       0.217
```

The absolute difference is **0.023**.

8. Which day of the week has the highest average miles-per-trip? Write the code to compute average miles-per-trip for each day of the week and find the day that has the highest value. (Reminder: you will exclude the people who didn't travel because they had zero trips) (6 points) (*Hint: Review Lab 6*)

```
personHHData %>%
  filter(Sum_PMT > 0) %>%
  group_by(DOW) %>%
  summarise(n = n(),
            avgMilesTrip = mean(Sum_PMT / Sum_Trips)) %>%
  View()
```

Sunday has the highest average miles-per-trip.

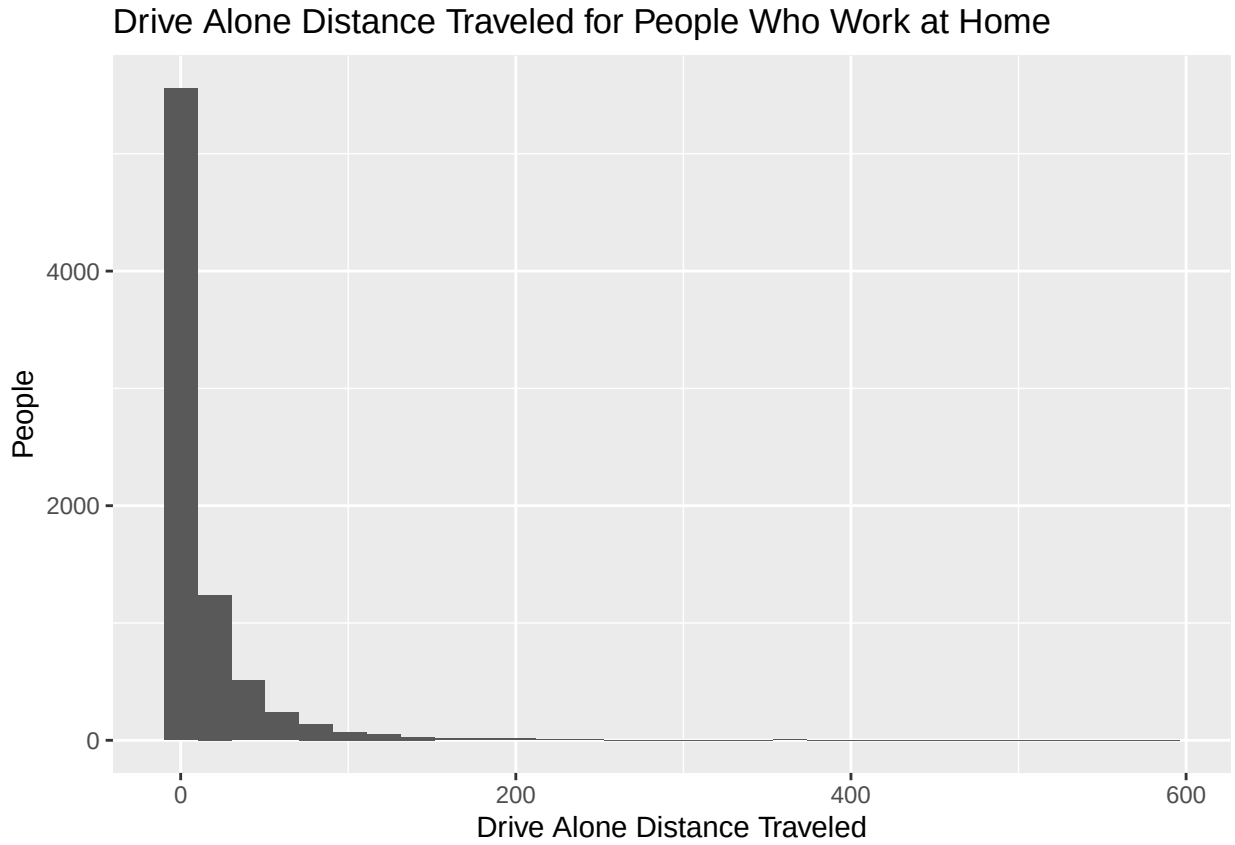
9. Write the code that creates a table showing the average distance traveled by each of the following modes: Drive Alone, Passenger, and Walking on each day of the week for people with disability. (5 points) (*Hint: Review Lab 6*)

```
personHHData %>%
  filter(Disability == 1) %>%
  group_by(DOW) %>%
  summarise(DriveAlone = mean(DriveAlone_Dist),
            Passenger = mean(Passenger_Dist),
            WalkDist = mean(Walk_Dist)) %>%
  View()
```

10. Makes a histogram that shows how drive alone distance traveled varies over the people who work from home. Add an appropriate title, and briefly discuss the results. (You do not need to attach your final plots, as long as we can run your code and get the plot ourselves) (6 points) (*Hint: Review Lab 6*)

```
q10Hist <- personHHData %>% filter(WorkHome == 1) %>% ggplot(aes(x = DriveAlone_Dist))
q10Hist + geom_histogram() + xlab('Drive Alone Distance Traveled') + ylab("People") + ggtitle('Drive Al
```

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



Discussion: **We can see that the vast majority of people who work at home drive less than 25-50 miles alone, with only a small amount of people diving more than 50 miles alone.**

11. Which area (among urban, suburb, exurb, and rural) do you see the lowest average PMT for employed people who are between 35-55 years old (including 35 and 55)? (5 points) (*Hint: Review Lab 6*)

```
PersonDataQ11 <- left_join(PersonData, hh_bgDensity)
```

```
## Joining, by = "hhid"
```

```
PersonDataQ11 %>%
  filter(Age >= 35 & Age <= 55,
         Employed == 1) %>%
  group_by(bg_group) %>%
  summarise(AveragePMT = mean(Sum_PMT)) %>%
  View()
```

Urban has the lowest average PMT for employed people who are between 35-55 years old (including 35 and 55).

12. Complete the following code to make a static map of the average number of cars in a household at county level. Remove background grid, add an appropriate title and choose appropriate color palette, and briefly discuss the results. (7 points) (*Hint: Review Lab 7*)

```
require(maps) || install.packages("maps", dependencies = TRUE)
```

```
## [1] TRUE
```

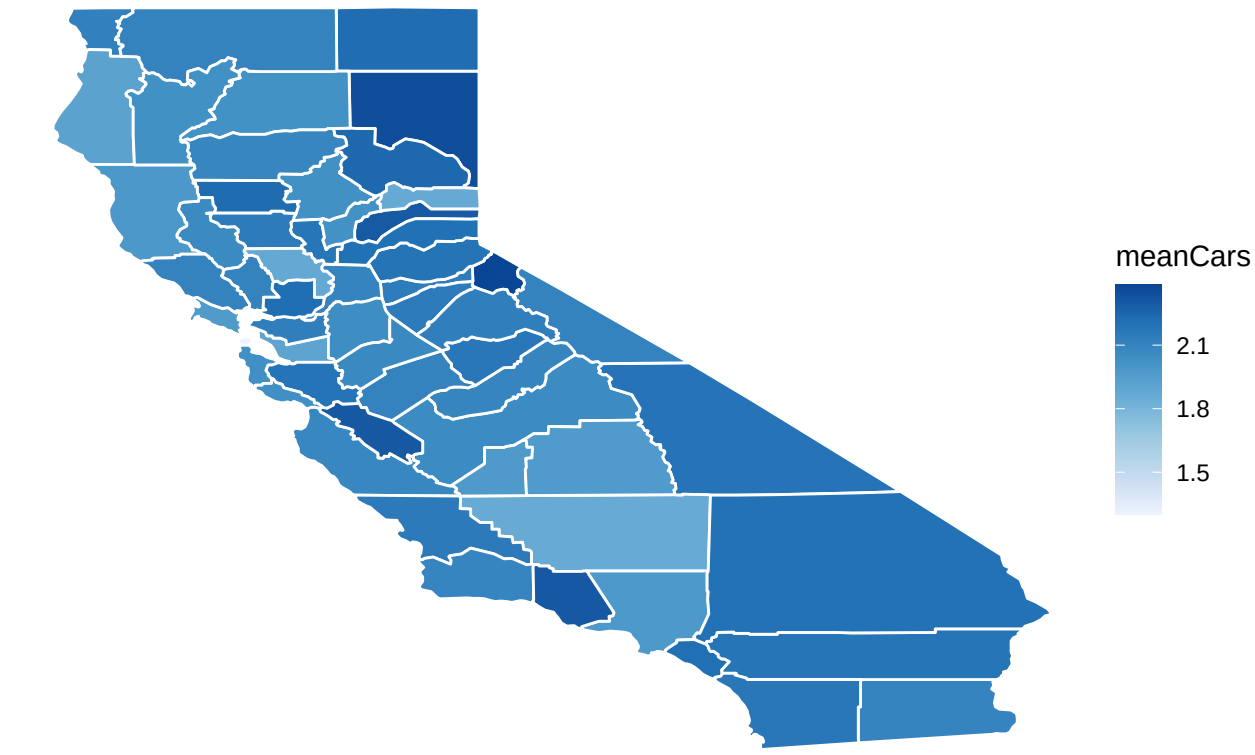
```
library(maps)

county <- ggplot2::map_data("county", region = "california")

prhh_aggreg <- personHHData %>%
  group_by(County, CTFIP) %>%
  summarise(meanCars = mean(HH_nCars))

county_prhh <- prhh_aggreg %>%
  mutate(subregion = tolower(County)) %>%
  full_join(county, by = "subregion")

ggplot(county_prhh) + geom_polygon(aes(x = long, y = lat, group = subregion, fill = meanCars), colour =
  scale_fill_distiller(palette = "Blues", direction = 0) +
  ggtitle("Average Number of Cars per Household at County-level") +
  theme_void()
```



Discussion: **We can see that the distribution of average number of cars per household is pretty evenly spread at the county level. The urban areas of SF, LA, and SD counties have average values near the median, while suburban counties like Ventura, Santa Clara, and San Benito counties have higher than average amounts of cars per household.**

13. Do not include ANY irrelevant code! You can only provide ONE code solution to each question. (1 point)