Project Part 3

Minh Tran

Section 1: Write-Up, which will provide your final version of the written steps and justifications for cleaning and preparing the dataset.

The dataset is collected by the American Community Survey in 2019. Originally, it includes the survey results from 3142 counties in the US, with 18 survey questions involving income, age, insurance status, population, household member count, poverty, and unemployment rate. Here are the steps that have taken to clean up this dataset so as it can be made the best use of for our analysis. First, since we are only interested in the results from counties in California and Texas with a population of more than 40,000, the dataset has been filtered to include only counties of equal to and more than 40,000 people. Additionally, to only analyze how the difference in poverty rate for those of 65 and older and under 18 related to other factors, a question involving the median individual income of people aged over 25 has also been removed as it contains people between age 18 and 65. The complete dataset is finalized and saved under "county_2019_final", including 17 questions and inputs from all counties in California and Texas, where the population is 40,000 and over.

Section 2: Responses to feedback, which will provide notes on constructive feedback provided by peers.

oo Feedback1:"Overall, I believe this was a well-executed project. I have a few suggestions for improvement, but the write-up is well-written and grammatically sound. The dataset cleaning plan was thorough and effective. Filtering and selecting the data seems like the appropriate approach. I noticed that there were 18 variables in the dataset instead of the expected 19. You used the variable name "filter" to describe the function. Maybe a better way to describe what the function is doing is to include or choose only the states of California or Texas. If you say you filtered out the states that are less than 40,000 you include states that have a population of 40,000 which I assume is not what you want. At the end after cleaning, a good idea may be to include what the data set looks like after all the changes."

Response: I appreciate the feedback a lot. I will definitely be more careful using the word filter in my write-up. Maybe I was zoning out on the number of questions (variables) but there were initially 18, for some reason I removed median_individual_income_age_25plus (because it was between 18 and 65) and instead of subtracting it I added it to the original number of questions. So at the end there is actually only 17 variables included. And I will make a little summary at the end to show what the final dataset looks like after changes.

oo Feedback2:"The only thing is (it might've only been for me) but the column names were called V1, V2 ... which made it a little more challenging to filter when following along."

Response: I think this has to do with how you uploaded the dataset. It has been a long time since I have seen this problem but I would suggest asking Dr. Burnham.

Section 3: R code used to clean the dataset. Make sure to comment your code.

```
county_2019 <- read.csv("county_2019.csv")

#select cali and texas
county_2019_catx <- county_2019 %>%
  filter(state=="California"|state=="Texas")

#population >40000
```

```r
county_2019_40<-county_2019_catx%>%
  filter(pop>=40000)

#remove na values and remove variable with age 25+
county_2019_final <- county_2019_40 %>%
  filter(!if_any(everything(), is.na))%>%
  select(-median_individual_income_age_25plus)
```

Section 4: Dataset, where you will provide the structure of your final dataset. There should be no code that is printed in this section, just the output provided from the glimpse() function.

```r
glimpse(county_2019_final)
```

```
## Rows: 119
## Columns: 17
## $ state                    <chr> "California", "California", "California", "Ca~
## $ name                     <chr> "Alameda County", "Butte County", "Contra Cos~
## $ mean_household_income    <int> 130710, 75746, 135742, 110041, 74776, 68313, ~
## $ median_age               <dbl> 37.6, 37.1, 39.7, 45.9, 32.2, 38.4, 32.4, 31.~
## $ median_household_income  <int> 99406, 52537, 99716, 83377, 53969, 48041, 476~
## $ median_individual_income <int> 43583, 24067, 42181, 36682, 25238, 25114, 182~
## $ per_capita_income        <int> 47314, 29506, 48178, 42749, 24422, 28769, 180~
## $ persons_per_household    <dbl> 2.82, 2.57, 2.87, 2.63, 3.14, 2.41, 3.81, 3.1~
## $ pop                      <int> 1656754, 225817, 1142251, 188563, 984521, 135~
## $ poverty                  <dbl> 9.9, 19.1, 8.7, 8.4, 22.5, 20.1, 24.1, 21.0, ~
## $ poverty_65_and_over      <dbl> 9.5, 9.7, 6.4, 5.7, 12.8, 9.7, 19.3, 12.3, 11~
## $ poverty_under_18         <dbl> 11.3, 19.6, 10.9, 9.6, 32.2, 20.8, 32.2, 29.1~
## $ unemployment_rate        <dbl> 4.4, 7.7, 5.0, 5.4, 8.2, 6.8, 13.7, 9.4, 7.3,~
## $ uninsured                <dbl> 4.4, 6.2, 5.1, 4.6, 8.3, 8.2, 8.0, 7.9, 7.5, ~
## $ uninsured_65_and_older   <dbl> 1.0, 0.3, 0.9, 0.3, 1.2, 0.2, 1.6, 1.0, 2.0, ~
## $ uninsured_under_19       <dbl> 2.1, 3.2, 2.6, 3.0, 2.9, 5.9, 4.0, 3.2, 3.1, ~
## $ uninsured_under_6        <dbl> 1.9, 3.3, 1.9, 4.9, 2.3, 5.6, 2.0, 2.3, 1.6, ~
```