Project 1: Wrangling, Exploration, Visualization

SDS322E

Data Wrangling, Exploration, Visualization

Raquel Mejia | rm57578

Introduction **Paragraph or two introducing your datasets and variables, why they are interesting to you, etc.**

```
# read your datasets in here, e.g., with read_csv()
library(tidyverse)
library(usmap)
medincome2019 <- read_csv("/stor/home/rm57578/project1/datasets/medincome2019.csv")
prelimivf2019 <- read_csv("/stor/home/rm57578/project1/datasets/prelimivf2019.csv")
```

Tidying: Reshaping If your datasets are tidy already, demonstrate that you can reshape data with pivot wider/longer here (e.g., untidy and then retidy). Alternatively, it may be easier to wait until the wrangling section so you can reshape your summary statistics. **Note here if you are going to do this. need to wrangle data before I can join, should I add wrangling initial dataset?**

```
# your tidying code (if applicable; can also wait until
# wrangling section)
```

Joining/Merging *Join your datasets into one using a* ***dplyr*** *join function on an ID variable (or ID variables) common to both*

Discuss the process in words, including why you chose the join you did

At a minimum, you should calculate (and your discussion should mention) the number of...

(1) total observations/rows in each dataset
(2) unique IDs in each dataset
(3) IDs that appear in one dataset but not the other (and which those are)
(4) IDs the datasets have common

Discuss the size of the joined dataset and how it relates to the size of the original datasets

In the joined dataset, note how many observations/rows were dropped, and any potential problems with this

```
joinmedincome2019 <- medincome2019 %>% mutate(location = str_replace(Location,
    ".+", toupper)) %>% select(location, "Median Annual Household Income")
```

(1) total observations/rows in each dataset:

```
nrow(joinmedincome2019)  #53 rows: median income of 50 states, USA median income, Puerto Rico median in
```

```
## [1] 53
```

```r
nrow(prelimivf2019)  #448 rows,
```

```
## [1] 448
```

(2) unique IDs in each dataset

```r
n_distinct(joinmedincome2019)  #53 unique locations
```

```
## [1] 53
```

```r
n_distinct(prelimivf2019)  #448 unique clinics that reported IVF data to CDC in 2019
```

```
## [1] 448
```

(3) IDs that appear in one dataset but not the other (and which those are)

```r
anti_join(prelimivf2019, joinmedincome2019, by = c(CurrentClinicState = "location")) %>%
    nrow()  #all clinics in the 'prelimivf2019' dataset are located in the 'joinmedincome2019' dataset.
```

```
## [1] 0
```

```r
anti_join(joinmedincome2019, prelimivf2019, by = c(location = "CurrentClinicState"))  #Four locations i
```

```
## # A tibble: 4 x 2
##   location      `Median Annual Household Income`
##   <chr>         <chr>
## 1 UNITED STATES $65,712
## 2 ALASKA        $75,463
## 3 NEW HAMPSHIRE $77,933
## 4 WYOMING       $65,003
```

(4) IDs the datasets have common

```r
semi_join(joinmedincome2019, prelimivf2019, by = c(location = "CurrentClinicState")) %>%
    nrow()  #49 IDs (locations) in common
```

```
## [1] 49
```

```r
prelimivf2019 %>% distinct(CurrentClinicState) %>% semi_join(joinmedincome2019,
    by = c(CurrentClinicState = "location")) %>% as.list  #here is a list of all of the locations they
```

```
## $CurrentClinicState
##  [1] "ALABAMA"              "ARIZONA"              "ARKANSAS"
##  [4] "CALIFORNIA"           "COLORADO"             "CONNECTICUT"
##  [7] "DELAWARE"             "DISTRICT OF COLUMBIA" "FLORIDA"
## [10] "GEORGIA"              "HAWAII"               "IDAHO"
## [13] "ILLINOIS"             "INDIANA"              "IOWA"
## [16] "KANSAS"               "KENTUCKY"             "LOUISIANA"
## [19] "MAINE"                "MARYLAND"             "MASSACHUSETTS"
## [22] "MICHIGAN"             "MINNESOTA"            "MISSISSIPPI"
## [25] "MISSOURI"             "MONTANA"              "NEBRASKA"
## [28] "NEVADA"               "NEW JERSEY"           "NEW MEXICO"
## [31] "NEW YORK"             "NORTH CAROLINA"       "NORTH DAKOTA"
## [34] "OHIO"                 "OKLAHOMA"             "OREGON"
## [37] "PENNSYLVANIA"         "PUERTO RICO"          "RHODE ISLAND"
## [40] "SOUTH CAROLINA"       "SOUTH DAKOTA"         "TENNESSEE"
## [43] "TEXAS"                "UTAH"                 "VERMONT"
## [46] "VIRGINIA"             "WASHINGTON"           "WEST VIRGINIA"
## [49] "WISCONSIN"
```

(5) Discuss the size of the joined dataset and how it relates to the size of the original datasets
(6) In the joined dataset, note how many observations/rows were dropped, and any potential problems with this

```
ivfincome <- left_join(prelimivf2019, joinmedincome2019, by = c(CurrentClinicState = "location"))
ivfincome %>% nrow()
```

```
## [1] 448
```

```
# left join was performed because only interested in median
# incomes in reference to the locations that had IVF data
# reported to the CDC in 2019.  Thus, the number of rows in
# the merged dataset is the same as the number of rows in the
# 'prelimmivf2019' dataset (448 rows representing 448 unique
# clinics).  As a result of this join, four locations from
# the 'joinmedincome2019' dataset were dropped, since United
# States, Alaska, New Hampshire, and Wyoming were not listed
# as the current location of a fertility clinic in the
# 'prelimivf2019' dataset.  Dropping these rows is not
# problematic because they are not relevant to the analysis
# based on clinic location that will be performed below.
```

Discussions of joining here. Feel encouraged to break up into more than once code chunk and discuss each in turn.

Wrangling Say which variables I'm focusing on (three numeric: TotNumCyclesAll, Median Annual Household Income, TransPGTAll, SARTmember)

```
# clean state and income columns
ivfincome <- ivfincome %>% rename(state = "CurrentClinicState") %>%
    rename(medianincome = "Median Annual Household Income") %>%
    mutate(medianincome = str_replace_all(medianincome, "[$,]*[$,]*",
        "")) %>% mutate(medianincome = as.integer(medianincome)) %>%
    mutate(ND_NumTrans1 = na_if(ND_NumTrans1, "*")) %>% mutate(ND_NumTrans2 = na_if(ND_NumTrans2,
    "*")) %>% mutate(ND_NumTrans3 = na_if(ND_NumTrans3, "*")) %>%
    mutate(ND_NumTrans4 = na_if(ND_NumTrans4, "*")) %>% mutate(Donor_NumTrans1 = na_if(Donor_NumTrans1,
    "*")) %>% mutate(Donor_NumTrans2 = na_if(Donor_NumTrans2,
    "*")) %>% mutate(Donor_NumTrans3 = na_if(Donor_NumTrans3,
    "*")) %>% mutate(Donor_NumTrans4 = na_if(Donor_NumTrans4,
    "*")) %>% mutate(ND_NumTrans1 = str_replace_all(ND_NumTrans1,
    ",", "")) %>% mutate(ND_NumTrans2 = str_replace_all(ND_NumTrans2,
    ",", "")) %>% mutate(ND_NumTrans3 = str_replace_all(ND_NumTrans3,
    ",", "")) %>% mutate(ND_NumTrans4 = str_replace_all(ND_NumTrans4,
    ",", "")) %>% mutate(Donor_NumTrans1 = str_replace_all(Donor_NumTrans1,
    ",", "")) %>% mutate(Donor_NumTrans2 = str_replace_all(Donor_NumTrans2,
    ",", "")) %>% mutate(Donor_NumTrans3 = str_replace_all(Donor_NumTrans3,
    ",", "")) %>% mutate(Donor_NumTrans4 = str_replace_all(Donor_NumTrans4,
    ",", "")) %>% mutate(ND_NumTrans1 = as.integer(ND_NumTrans1)) %>%
    mutate(ND_NumTrans2 = as.integer(ND_NumTrans2)) %>% mutate(ND_NumTrans3 = as.integer(ND_NumTrans3))
    mutate(ND_NumTrans4 = as.integer(ND_NumTrans4)) %>% mutate(Donor_NumTrans1 = as.integer(Donor_NumTra
    mutate(Donor_NumTrans2 = as.integer(Donor_NumTrans2)) %>%
    mutate(Donor_NumTrans3 = as.integer(Donor_NumTrans3)) %>%
    mutate(Donor_NumTrans4 = as.integer(Donor_NumTrans4)) %>%
```

```
    mutate(TransPGTAll = na_if(TransPGTAll, "*")) %>% mutate(TransPGTAll = str_replace_all(TransPGTAll,
    "%", "")) %>% mutate(TransPGTAll = as.numeric(TransPGTAll))
```

Summary statistics of clinic IVF cycles grouped by state (sum, median, mean, sd)

```
ivfincome %>% group_by(state) %>% summarise(totalivfcycles = sum(TotNumCyclesAll),
    medianivfcycles = median(TotNumCyclesAll), meanivfcycles = mean(TotNumCyclesAll),
    sdivfcycles = sd(TotNumCyclesAll))
```

```
## # A tibble: 49 x 5
##    state                totalivfcycles medianivfcycles meanivfcycles sdivfcycles
##    <chr>                         <dbl>           <dbl>         <dbl>       <dbl>
##  1 ALABAMA                        1407             339          281.        174.
##  2 ARIZONA                        5299             350          408.        266.
##  3 ARKANSAS                        381             381          381          NA
##  4 CALIFORNIA                    57491             405          798.       1141.
##  5 COLORADO                       6322             290.         790.       1215.
##  6 CONNECTICUT                    6278             726         1046.        756.
##  7 DELAWARE                       1258             629          629         362.
##  8 DISTRICT OF COLUMBIA           1229             614.         614.        177.
##  9 FLORIDA                       13042             272          483.        538.
## 10 GEORGIA                        6424             534          803         785.
## # ... with 39 more rows
```

If applicable, at least 1 of these should group by two categorical variables

```
ivfincome %>% group_by(state, SARTmember) %>% summarise(countsclinicSARTmember = n())  #number of clini
```

```
## # A tibble: 75 x 3
## # Groups:   state [49]
##    state       SARTmember countsclinicSARTmember
##    <chr>       <chr>                       <int>
##  1 ALABAMA     No                              1
##  2 ALABAMA     Yes                             4
##  3 ARIZONA     No                              3
##  4 ARIZONA     Yes                            10
##  5 ARKANSAS    Yes                             1
##  6 CALIFORNIA  No                             18
##  7 CALIFORNIA  Yes                            54
##  8 COLORADO    No                              3
##  9 COLORADO    Yes                             5
## 10 CONNECTICUT Yes                             6
## # ... with 65 more rows
```

Used pivot_wider to tidy summary data (counts of clinics in each state, categorized by below/above US median income)

```
ivfincome %>% mutate(aboveUSmedian = ifelse(medianincome > 65712,
    "Above Median U.S. Income", "Below Median U.S. Income")) %>%
    group_by(state, aboveUSmedian) %>% summarise(countaboveUSmedian = n()) %>%
    pivot_wider(names_from = "aboveUSmedian", values_from = "countaboveUSmedian")
```

```
## # A tibble: 49 x 3
## # Groups:   state [49]
##    state                `Below Median U.S. Income` `Above Median U.S. Income`
##    <chr>                                     <int>                       <int>
##  1 ALABAMA                                       5                          NA
```

```
##  2 ARIZONA                                13                      NA
##  3 ARKANSAS                                1                      NA
##  4 CALIFORNIA                             NA                      72
##  5 COLORADO                               NA                       8
##  6 CONNECTICUT                            NA                       6
##  7 DELAWARE                               NA                       2
##  8 DISTRICT OF COLUMBIA                   NA                       2
##  9 FLORIDA                                27                      NA
## 10 GEORGIA                                 8                      NA
## # ... with 39 more rows
```

```
ivfincome %>% filter(medianincome > 65712) %>% nrow()  #231 clinics are located in states with a median
```

```
## [1] 231
```

```
ivfincome %>% filter(medianincome > 65712) %>% summarise(meantotalcycles = mean(TotNumCyclesAll))  #Cli
```

```
## # A tibble: 1 x 1
##   meantotalcycles
##             <dbl>
## 1            938.
```

```
ivfincome %>% filter(medianincome < 65712) %>% nrow()  #217 clinics are located in states with a median
```

```
## [1] 217
```

```
ivfincome %>% filter(medianincome < 65712) %>% summarise(meantotalcycles = mean(TotNumCyclesAll))  #Cli
```

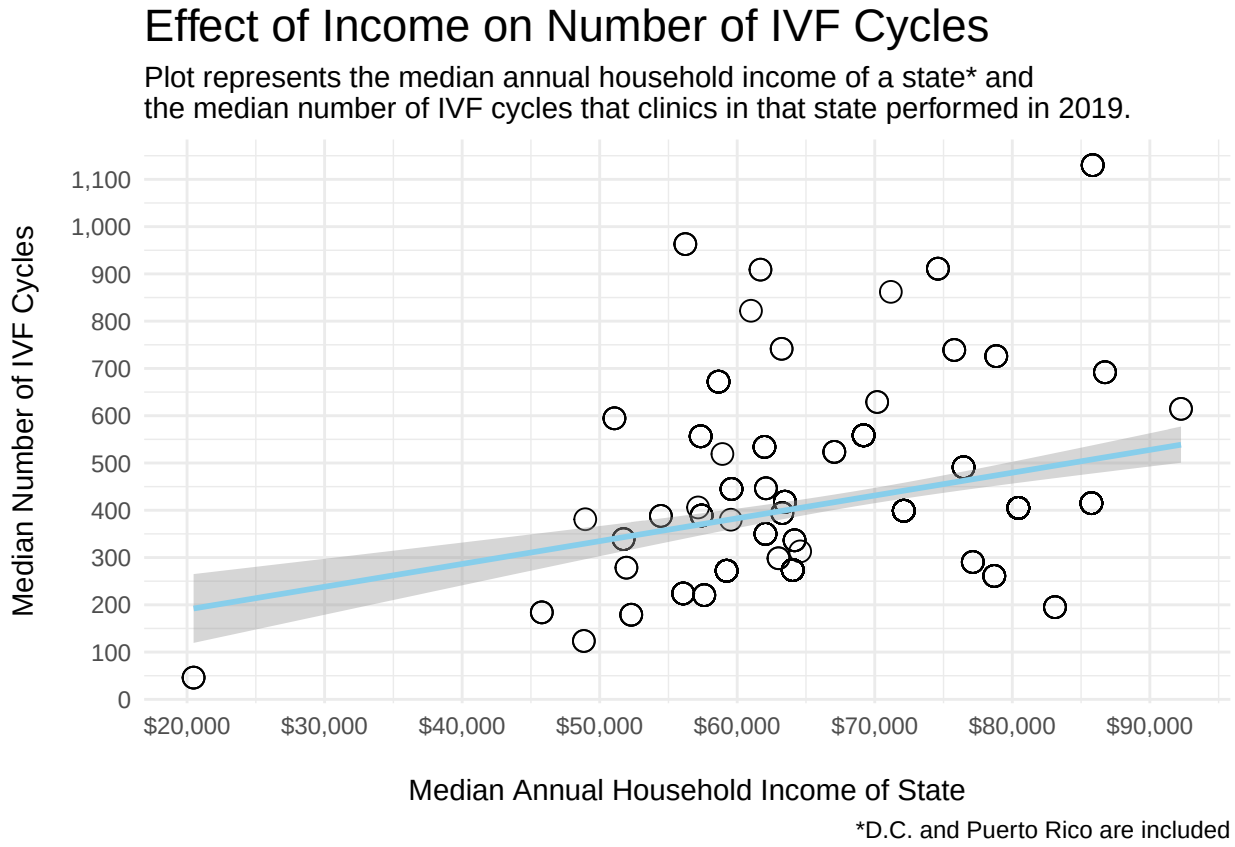
```
## # A tibble: 1 x 1
##   meantotalcycles
##             <dbl>
## 1            526.
```

Your discussion of wrangling section here. Feel encouraged to break up into more than once code chunk and discuss each in turn.

==========

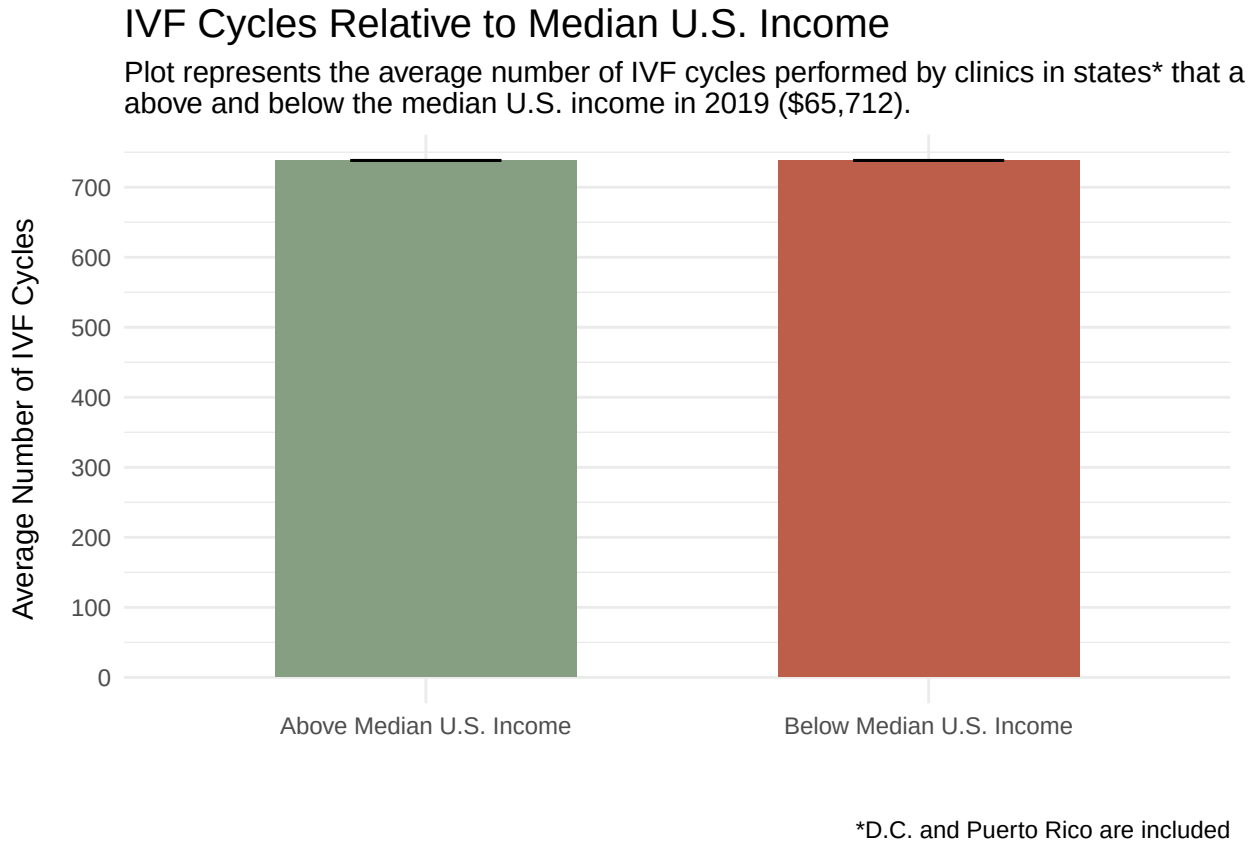
Visualizing

```
ivfincome %>% group_by(state) %>% mutate(medianivfcycles = median(TotNumCyclesAll)) %>%

ggplot(aes(x = medianincome, y = medianivfcycles)) + geom_point(shape = 1,
    size = 3.5) + geom_smooth(method = "lm", col = "skyblue") +
    theme_minimal() + theme(plot.title = element_text(size = 17),
    axis.title.x = element_text(margin = margin(15, 0, 0, 0)),
    axis.title.y = element_text(margin = margin(0, 12, 0, 0)),
    ) + labs(title = "Effect of Income on Number of IVF Cycles",
    subtitle = "Plot represents the median annual household income of a state* and \nthe median number
    caption = "*D.C. and Puerto Rico are included") + scale_x_continuous(name = "Median Annual Household
    labels = scales::dollar_format(), breaks = seq(20000, 1e+05,
        by = 10000)) + scale_y_continuous(name = "Median Number of IVF Cycles",
    labels = scales::comma, breaks = seq(0, 1200, by = 100))
```



^^Your discussion of plot 1

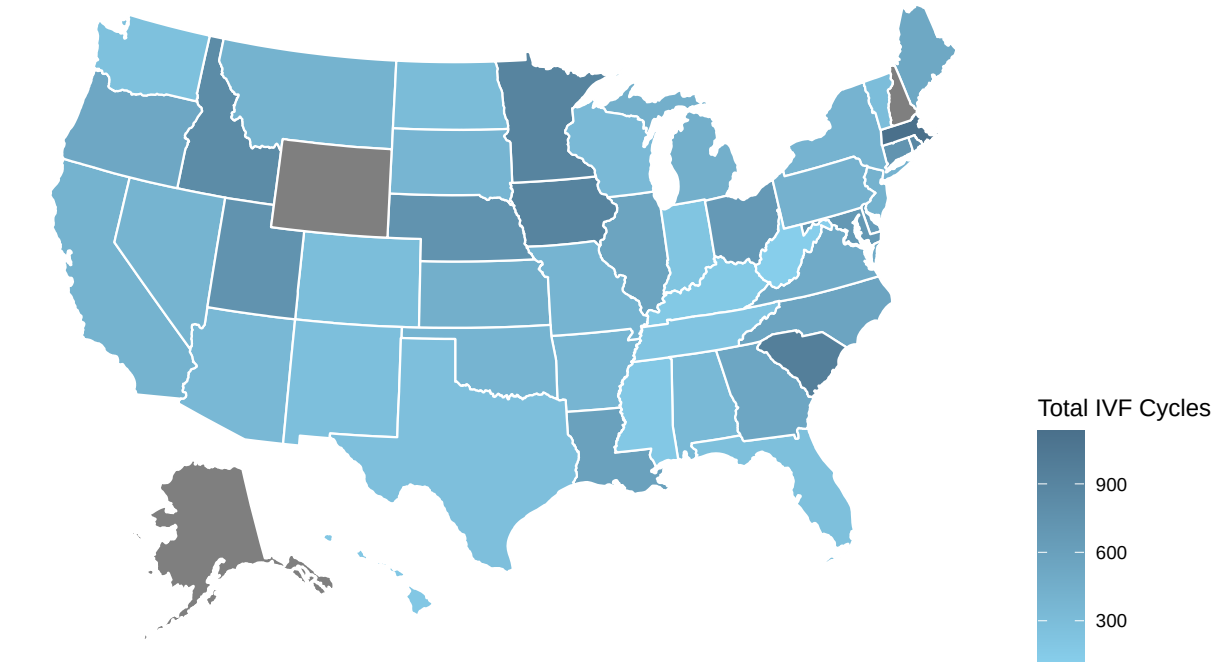
```
medianUSincome <- ivfincome %>% mutate(aboveUSmedian = ifelse(medianincome >
    65712, "Above Median U.S. Income", "Below Median U.S. Income")) %>%
    mutate(meanivfcycles = mean(TotNumCyclesAll))

medianUSincome %>% ggplot(aes(x = aboveUSmedian, y = meanivfcycles,
    fill = aboveUSmedian)) + geom_bar(stat = "summary", fun = mean,
    width = 0.6) + geom_errorbar(stat = "summary", fun.data = mean_se,
    width = 0.3) + theme_minimal() + theme(plot.title = element_text(size = 15),
    axis.title.x = element_text(margin = margin(15, 0, 0, 0)),
    axis.title.y = element_text(margin = margin(0, 12, 0, 0)),
    legend.position = "none") + labs(title = "IVF Cycles Relative to Median U.S. Income",
    subtitle = "Plot represents the average number of IVF cycles performed by clinics in states* that a
    caption = "*D.C. and Puerto Rico are included") + xlab(" ") +
    scale_y_continuous(name = "Average Number of IVF Cycles",
        labels = scales::comma, breaks = seq(0, 1200, by = 100)) +
    scale_fill_manual(values = c("#869E81", "#BD5E4B"))
```



^^Your discussion of plot 2

Median or mean? What am i trying to show?

```
medivfincome <- ivfincome %>% group_by(state) %>% mutate(medianivfcycles = median(TotNumCyclesAll))

plot_usmap(data = medivfincome, values = "medianivfcycles", regions = "states",
    col = "white") + labs(title = "Median Number of IVF Cycles Across the U.S.",
    subtitle = "This map shows the median number of IVF cycles that clinics in that state performed in 
    caption = "Alaska, New Hampshire, and Wyoming did not report any IVF cycles in 2019 (shown as grey)
    scale_fill_continuous(low = "skyblue", high = "skyblue4",
        name = "Total IVF Cycles", label = scales::comma) + theme(legend.position = "right",
    plot.title = element_text(hjust = 0.5, size = 15))
```

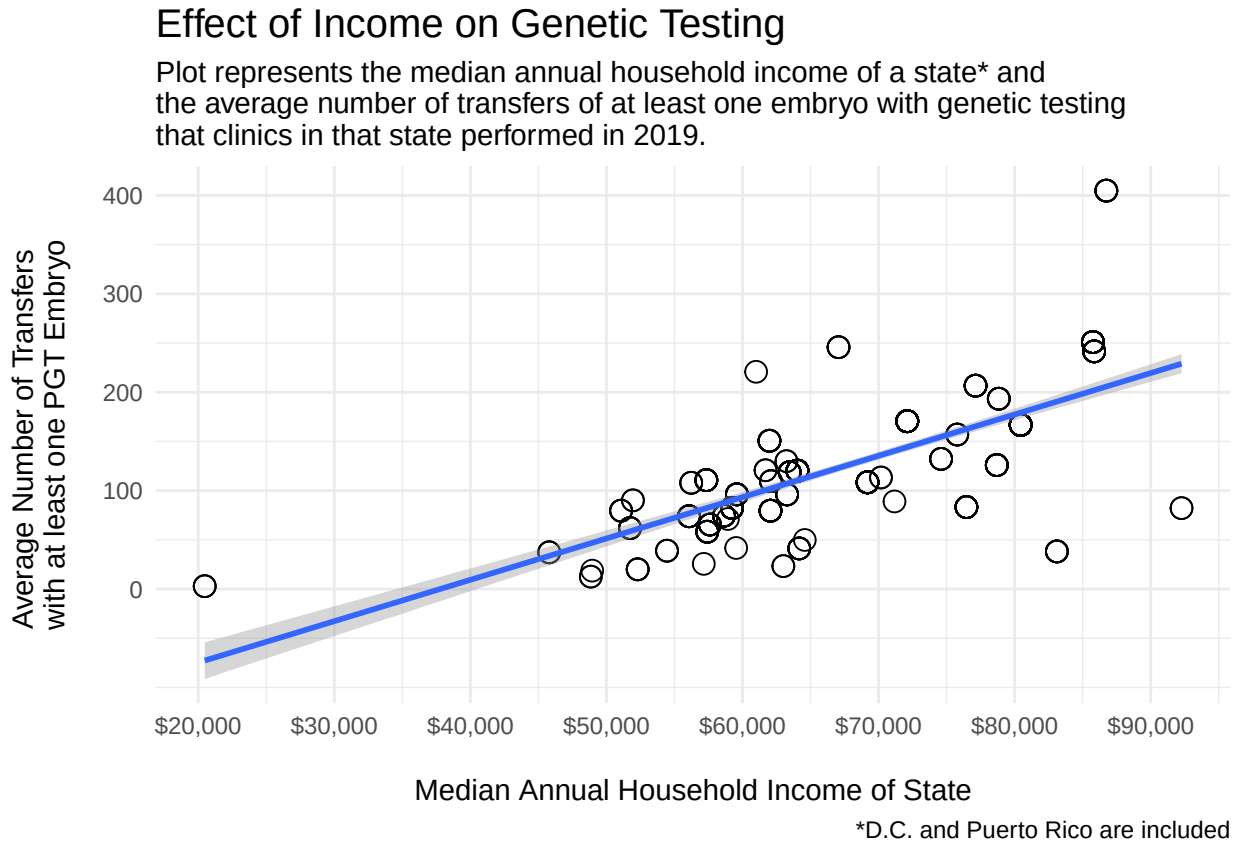


Your discussion of plot 2

For plot: total number of cycles per state with at least one PGT vs. median income

includes variable thats a function of other variable

```
ivfincome %>% group_by(CurrentClinicName1) %>% mutate(totaltransfers = sum(Donor_NumTrans1,
    Donor_NumTrans2, Donor_NumTrans3, Donor_NumTrans4, ND_NumTrans1,
    ND_NumTrans2, ND_NumTrans3, ND_NumTrans4, na.rm = T)) %>%
    mutate(cyclesatleastonePGT = (TransPGTAll/100) * totaltransfers) %>%
    group_by(state) %>% mutate(statecycPGT = mean(cyclesatleastonePGT,
    na.rm = T)) %>%
ggplot(aes(x = medianincome, y = statecycPGT)) + geom_point(shape = 1,
    size = 3.5) + geom_smooth(method = "lm") + theme_minimal() +
    theme(plot.title = element_text(size = 15), axis.title.x = element_text(margin = margin(15,
        0, 0, 0)), axis.title.y = element_text(margin = margin(0,
        12, 0, 0)), ) + labs(title = "Effect of Income on Genetic Testing",
    subtitle = "Plot represents the median annual household income of a state* and \nthe average number
    caption = "*D.C. and Puerto Rico are included") + scale_x_continuous(name = "Median Annual Household
    labels = scales::dollar_format(), breaks = seq(20000, 1e+05,
        by = 10000)) + scale_y_continuous(name = "Average Number of Transfers \nwith at least one PGT E
    labels = scales::comma, breaks = seq(0, 500, by = 100))
```



Your discussion of plot 3

old vs young patients vs income

Your discussion of plot 4

Concluding Remarks If any! wealth disparity! infertility should be covered under insurance!