Dickson Aiden DACSS 601 HW5 Bivariate Relationships, Modelling and Functions

Aiden Dickson

5/2/2021

Refinements made to former work after learning more: [Answered HW 5 Questions after Resultant Plot at End]

- I edited code to load dataset so it will have * labels * this time via haven's 'as_factor()' function piped on the end of a read_sav() operation. Before I just used 'mutate()'}, thinking I was required to re-add labels this way.
- I will also now ensure to keep category of answer 'Don't know and refused' in any 'mutate()' per suggestion in various R projects, in this and forthcoming work, if its not combined with another answer to mutate a new column, in which only the other answer is to be taken into consideration, as is the case below.
- I will be attempting to calculate summaries of the data that are now * weighted *, to make sure estimates are representative of the population.

In following I will be attempting a plot of x-tabs table with 3 demographic variables, building on work from HW3 and HW4 and still using the same dataset, now loaded with labels. As a reminder this dataset [name 'us-data' here] represents the results of a survey containing answers to 6 questions regarding attitudes towards globalization and US alliances, with several demographic categories.

In the former analysis I used variables Gender and Q4 (Response to 'Overall has globalization in the past few years been good for the United States?'): '1' ='Good', '2' ='Bad', '3'='Both Good and Bad', and '9'='DK/Refused to Q4'. I erroneously thought I should remove the last category, but will not do that here.

In this analysis I will add Political View: 'polview', Income: 'in_come' and Education: 'edu_cat' instead of Gender for the demographic variables, and use Q4 * and * Q5 to generate a new column as to whether a respondent thinks globalization is good 'In past few years for US' (Q4), 'personally'(Q5), both 'In past few years for Us and personally', 'Neither', or 'Refused/DK' for at least one, if the other is answered it won't be 'Good'. Then I will attempt to create a plot that will show these three demographic categories and their multiple subcategories, side-by-side for comparison of responses based on these. I will also hopefully be able to weight these now.

1. Installing and Imported relevant libraries:

```r
if (!require('tidyverse')) install.packages('tidyverse')
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
if (!require('haven')) install.packages('haven')
```

```
## Loading required package: haven
```

```r
if (!require('dplyr')) install.packages('dplyr')
if (!require('knitr')) install.packages('knitr')
```

```
## Loading required package: knitr
```

```r
if (!require('tinytex')) install.packages('tinytex')
```

```
## Loading required package: tinytex
```

```
if (!require('ggplot2')) install.packages('ggplot2')
if (!require('forcats')) install.packages('forcats')
if (!require('GGally')) install.packages('GGally')
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(tidyverse)
library(haven)
library(dplyr)
library(knitr)
library(tinytex)
library(ggplot2)
library(GGally)
knitr::opts_chunk$set(echo = TRUE)
```

2.Importing Dataset – (from Google Drive as Labelled Now) Viewing real dimensions and top few Rows as Example:

```
us_data<-read_sav("/Volumes/GoogleDrive/My Drive/Classes Spring 2021/Classes -Data Analysis- R/DataSets/
  as_factor()

dim(us_data)
```

```
## [1] 1008   25
```

```
head(us_data, 10)
```

```
## # A tibble: 10 x 25
##    case_id weight state mstatus totper adults parent age   educ  income hispanic
##      <dbl>  <dbl> <chr> <fct>   <fct>  <fct>  <fct>  <fct> <fct> <fct>  <fct>
##  1  5.00e7  0.403 GA    Widowed One    One    <NA>   83    Post~ $30,0~ No
##  2  5.00e7  0.637 WV    Married Three  Two    Yes    34    Some~ $75,0~ No
##  3  5.00e7  0.341 PA    Refused One    One    <NA>   75    Four~ $75,0~ No
##  4  5.00e7  0.258 NC    Divorc~ One    One    <NA>   76    Four~ $30,0~ No
##  5  5.00e7  0.858 NJ    Married Four   Two    Yes    63    Two ~ $100,~ No
##  6  5.00e7  0.442 NY    Widowed Four   Four   <NA>   77    Two ~ $30,0~ No
##  7  5.00e7  1.50  FL    Married Two    Two    <NA>   43    High~ $200,~ No
##  8  5.00e7  1.27  NC    Married Three  Two    Yes    43    Four~ $200,~ No
##  9  5.00e7  1.19  VA    Married Two    Two    <NA>   86    High~ $40,0~ No
## 10  5.00e7  0.786 MI    Single~ Seven  Five   No     19    High~ $100,~ No
## # ... with 14 more variables: race <fct>, partyln <fct>, polview <fct>,
## #   sex <fct>, religion <fct>, date <date>, Q1 <fct>, Q2 <fct>, Q3 <fct>,
## #   Q4 <fct>, Q5 <fct>, Q6 <fct>, country2 <fct>, survey <fct>
```

3. Mutating the combination of Q4 and Q5 for new column, 'Globalization_Good_US_and_Pers'

```
library(forcats)

us_data <- us_data %>%
  mutate(Globalization_Good_US_and_Pers = case_when(

 Q4 == "Good" & (Q5 == "Good") ~ "Both",
 Q4 == "Good" & (Q5 != "Good") ~ "Good US",
 Q4 != "Good" & (Q5 == "Good") ~ "Good Prsnlly",
 Q4 != "Good" & (Q5 != "Good") & (Q4 != "DK/Refused" & Q5 != "DK/Refused") ~ "Neither",
 Q4 == "DK/Refused" |
 Q5 == "DK/Refused" ~ "Refused " )%>%


 fct_relevel("Both",
             "Good US",
             "Good Prsnlly",
             "Neither",
             "Refused ")
  )

view(us_data)
```

4. Using Frequency of Occurrence Table to Double-check the Conditional Logic in Mutate

```
tbl <-table(us_data$Q4, us_data$Q5, us_data$Globalization_Good_US_and_Pers)

view(tbl)
```

5. Viewing Categories of Interest, Pol View, Educ, Income, Collapsing Educ and Income into fewer categories. (Next 3)

Political View (left alone)

```
levels(us_data$polview)
```

```
## [1] "Very conservative"     "Somewhat conservative" "Moderate"
## [4] "Somewhat liberal"      "Very liberal"          "Don't know"
## [7] "Refused"
```

Income - Collapsing Categories [in_come]:

```
us_data <- us_data %>%
  mutate(in_come = fct_collapse(income,
              "Less than $50,000" = c(
"Less than $15,000",
"$15,000 but less than $25,000",
"$25,000 but less than $30,000",
"$30,000 but less than $40,000",
"$40,000 but less than $50,000",
"Less than $50,000 (Unspecified)"
                                    ),

"At least $50,000 but less than $100,000" = c(
"$50,000 but less than $75,000",
"$75,000 but less than $100,000",
"$50,000 but less than $100,000 (Unspecified)" ),

"Equal or Greater than $100,000" = c(
"Over $100,000",
"$100,000 to under $150,000",
"$150,000 to under $200,000",
"$250,000 or more",
"$100,000 and over (Unspecified)",
"$200,000 to under $250,000")),

 fct_relevel("Less than $50,000",
             "At least $50,000 but less than $100,000",
             "Equal or Greater than $100,000",
             "Don't know",
             "Refused")
      )
```

```
## Warning: Unknown levels in ‘f‘: At least $50,000 but less than $100,000, Equal
## or Greater than $100,000, Don’t know, Refused
```

```
levels(us_data$in_come)
```

```
## [1] "Less than $50,000"
## [2] "At least $50,000 but less than $100,000"
## [3] "Equal or Greater than $100,000"
## [4] "Don’t know"
## [5] "Refused"
```

Education - Collapsing Categories [edu_cat]:

```
us_data <- us_data %>%
  mutate(educ_cat = fct_collapse(educ,
          "High school grad or less" = c(
"Less than high school (Grades 1-8 or no formal schooling)",
"High school incomplete (Grades 9-11 or Grade 12 with NO diploma)",
"High school graduate (Grade 12 with diploma or GED certificate)"
                                        ),

                      "Some college" = c(
"Some college, no degree (includes community college)",
"Two year associate degree from a college or university"
                                        ),

                     "College grad+" = c(
"Four year college or university degree/Bachelor.s degree (e.g., BS, BA, AB)",
"Some postgraduate or professional schooling, no postgraduate degree",
"Postgraduate or professional degree, including master's, doctorate, medical or law degree (e.g., MA, M
                                       )
                                  )
      )

levels(us_data$educ_cat)
```

```
## [1] "High school grad or less" "Some college"
## [3] "College grad+"            "Don’t know"
## [5] "Refused"
```

```
view(us_data)
```

Just Example Here of Grouping by Globalization Good US and Personal, then obtaining weighted Summaries. ‘Sampling weights, also known as survey weights, are positive values associated with the observations (rows) in your dataset (sample), used to ensure that metrics derived from a data set are representative of the population (the set of observations).’ [1]

```
globalization_Good_US_and_Pers <- us_data %>%
 group_by(Globalization_Good_US_and_Pers) %>%
summarize(weighted_n = sum(weight)) %>%
mutate(weighted_group_size = sum(weighted_n),
         weighted_estimate = weighted_n / weighted_group_size
         )

globalization_Good_US_and_Pers
```

```
## # A tibble: 5 x 4
##   Globalization_Good_US_and_Pers weighted_n weighted_group_size weighted_estima~
##   <fct>                               <dbl>               <dbl>            <dbl>
```

[1] https://community.alteryx.com/

```
## 1 "Both"                              383.                1008.           0.380
## 2 "Good US"                            93.8               1008.           0.0931
## 3 "Good Prsnlly"                      111.                1008.           0.110
## 4 "Neither"                           369.                1008.           0.366
## 5 "Refused "                           50.5               1008.           0.0501
```

6. Selecting only Columns Needed for Analysis: case_id, weight, Globalization_Good_US_and_Pers, Political_View =polview, Income =in_come, Education=educ_cat

```
Globalization_Good_US_and_Pers_Tibble <- us_data %>%
  select(case_id, weight, Globalization_Answers= Globalization_Good_US_and_Pers, Political_View =polvie

Globalization_Good_US_and_Pers_Tibble
```

```
## # A tibble: 1,008 x 6
##     case_id weight Globalization_Ans~ Political_View  Income        Education
##       <dbl>  <dbl> <fct>              <fct>           <fct>         <fct>
##  1 50000004  0.403 Neither            Somewhat conse~ Less than $5~ College gra~
##  2 50000006  0.637 Good US            Moderate        At least $50~ Some college
##  3 50000008  0.341 Good US            Somewhat conse~ At least $50~ College gra~
##  4 50000009  0.258 Both               Somewhat liber~ Less than $5~ College gra~
##  5 50000012  0.858 Neither            Very liberal    Equal or Gre~ Some college
##  6 50000013  0.442 Neither            Somewhat conse~ Less than $5~ Some college
##  7 50000014  1.50  Both               Somewhat conse~ Equal or Gre~ High school~
##  8 50000015  1.27  Good US            Moderate        Equal or Gre~ College gra~
##  9 50000016  1.19  Neither            Very conservat~ Less than $5~ High school~
## 10 50000018  0.786 Both               Somewhat liber~ Equal or Gre~ High school~
## # ... with 998 more rows
```

```
view(Globalization_Good_US_and_Pers_Tibble)
```

7. Running a Gather Function to Make it Easier to Plot:

```
Globalization_Good_US_and_Pers_Tibble_Long<- Globalization_Good_US_and_Pers_Tibble %>%
  gather(key = subgroup_variable, value = subgroup, Political_View, Income, Education)
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
Globalization_Good_US_and_Pers_Tibble_Long
```

```
## # A tibble: 3,024 x 5
##     case_id weight Globalization_Answers subgroup_variable subgroup
##       <dbl>  <dbl> <fct>                 <chr>             <chr>
##  1 50000004  0.403 Neither               Political_View    Somewhat conservative
##  2 50000006  0.637 Good US               Political_View    Moderate
##  3 50000008  0.341 Good US              Political_View    Somewhat conservative
##  4 50000009  0.258 Both                 Political_View    Somewhat liberal
##  5 50000012  0.858 Neither              Political_View    Very liberal
##  6 50000013  0.442 Neither              Political_View    Somewhat conservative
##  7 50000014  1.50  Both                 Political_View    Somewhat conservative
##  8 50000015  1.27  Good US              Political_View    Moderate
##  9 50000016  1.19  Neither              Political_View    Very conservative
## 10 50000018  0.786 Both                 Political_View    Somewhat liberal
## # ... with 3,014 more rows
```

```
view(Globalization_Good_US_and_Pers_Tibble_Long)
```

8. Calculating weighted estimates of these answers.

```
weighted_tibble_for_plot <- Globalization_Good_US_and_Pers_Tibble_Long %>%

  group_by(subgroup_variable, subgroup, Globalization_Answers) %>%

  summarise(weighted_n = sum(weight)) %>%

  group_by(subgroup) %>%

  mutate(weighted_group_size = sum(weighted_n),
         weighted_estimate = weighted_n/weighted_group_size,

         )
```

```
## ‘summarise()‘ has grouped output by ’subgroup_variable’, ’subgroup’. You can override using the ‘.gr
```

```
  weighted_tibble_for_plot$subgroup<-factor(weighted_tibble_for_plot$subgroup, levels = c("High school
"Some college", "College grad+",
"Refused","Less than $50,000",
"At least $50,000 but less than $100,000","Equal or Greater than $100,000","Don't know", "Very conserva

weighted_tibble_for_plot <- weighted_tibble_for_plot %>%
  select(-weighted_n, -weighted_group_size)




weighted_tibble_for_plot
```

```
## # A tibble: 77 x 4
## # Groups:   subgroup [13]
##    subgroup_variable subgroup             Globalization_Answe~ weighted_estima~
##    <chr>             <fct>                <fct>                           <dbl>
##  1 Education         College grad+        "Both"                         0.502
##  2 Education         College grad+        "Good US"                      0.0873
##  3 Education         College grad+        "Good Prsnlly"                 0.0985
##  4 Education         College grad+        "Neither"                      0.290
##  5 Education         College grad+        "Refused "                     0.0225
##  6 Education         Don’t know           "Good US"                      0.0733
##  7 Education         High school grad or ~ "Both"                          0.308
##  8 Education         High school grad or ~ "Good US"                       0.0861
##  9 Education         High school grad or ~ "Good Prsnlly"                  0.132
## 10 Education         High school grad or ~ "Neither"                       0.398
## # ... with 67 more rows
```

```
dim(weighted_tibble_for_plot)
```

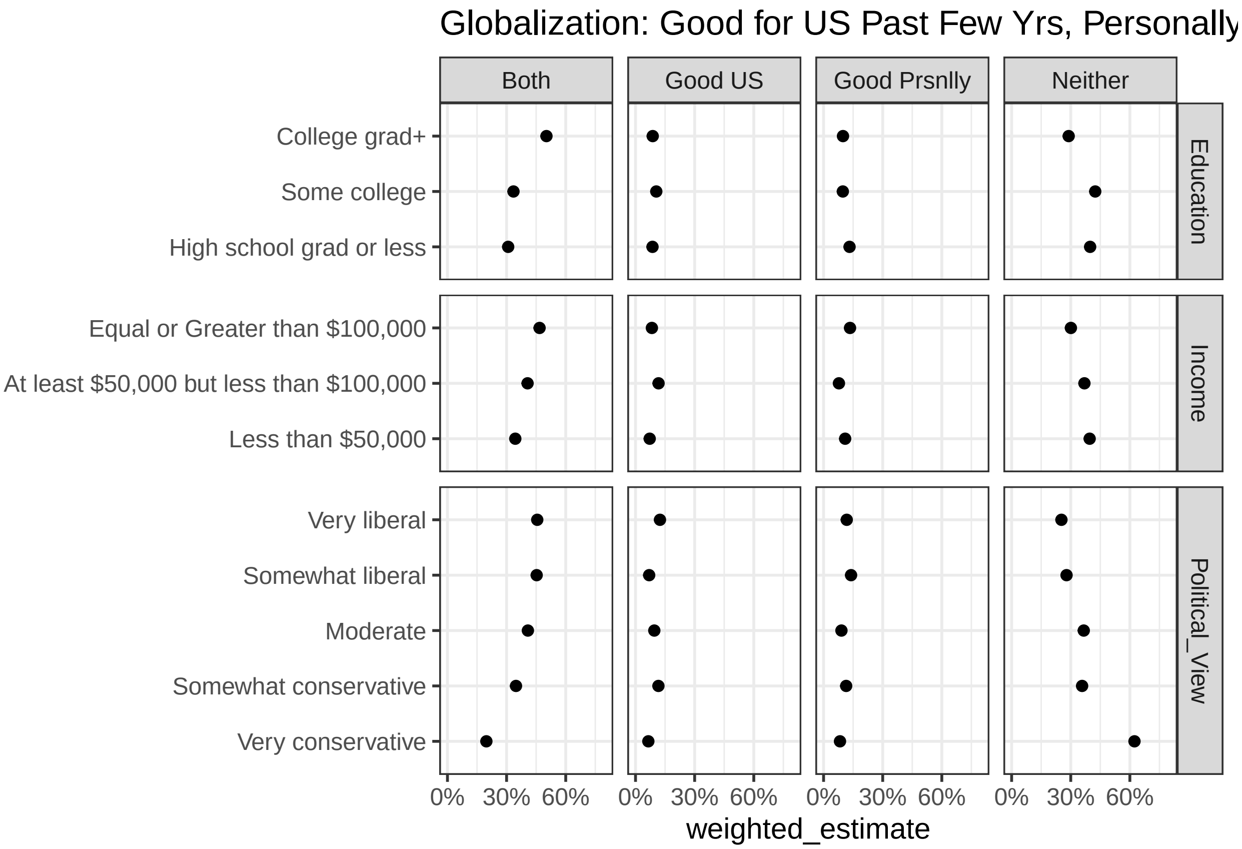
```
## [1] 77  4
```

```
view(weighted_tibble_for_plot)
```

9. Plotting the Weighted Results, First Removing the "Refused"

```
plot <- weighted_tibble_for_plot %>%
  filter(!(Globalization_Answers %in% c("Refused "))) %>%
  filter(!(subgroup %in%
                c("Don't know", "Refused"))) %>%
  ggplot(
    aes(
      x = weighted_estimate,
      y = subgroup
    )
  ) + ggtitle("Globalization: Good for US Past Few Yrs, Personally") +
  geom_point() +
  scale_x_continuous(limits = c(0, .8),
                     breaks = seq(0, .6, by = .3),
                     labels = scales::percent(
                              seq(0, .6, by = .3), accuracy = 1)
                     ) +
  facet_grid(cols = vars(Globalization_Answers),
             rows = vars(subgroup_variable),
             scales = "free_y",
             space = "free"
             ) +
  theme_bw() +
  theme(axis.title.y = element_blank())

plot
```



1. Descriptions of the variables - how they were collected, any missing values, etc

How Collected: This analysis of attitudes regarding globalization, based on gender, is taken from a sample of 1008 survey participants, 18 years of age or older, living in the United States, (304 respondents were interviewed on a land-line telephone, and 704 were interviewed on a mobile phone, including 469 who had no landline telephone). The survey was conducted under the direction of SSRS. It is a study of overall attitudes towards globalization with a larger scope of demographic features such as religion, race and political persuasion.

Missing Values: I encountered lot of 'grey-area' judgement call on how to handle these categories, and I need to have time to read up on how they are handled by protocol. For eg: I had to make a choice on category so if they answered either "Good" and one "DK / Refused" It counts the "Good' in created column because its a measure of answers"Good" so I harvested the 'Good'. If one were "Bad" and the other "DK/ Refused" I put "Refused". But it's a grey zone, should I have put "Neither"? 'Neither' to me is they did answer and it wasn't 'Good'. I didn't encounter however an 'N/A' in the data.

2.and 3. How you cleaned and coded the data, including a before/after comparison as needed and summary descriptives of the recoded variables

BEFORE: 1008 row, 25 column dataset: The initial Pew Dataset.

In this analysis I added Political View: 'polview', Income: 'in_come' and Education: 'edu_cat' instead of Gender for the demographic variables, and use Q4 * and * Q5 to generate a new column 'Globalization_Good_US_and_Pers' (eventually renamed 'Globalization Answers'), as to whether a respondent thinks globalization is good 'In past few years for US' (Q4 Good), 'personally'(Q5 Good), both 'In past few years for Us and personally (Q4 & Q5 Good)', 'Neither (Not Good & Not Good, excluding DK/Refused), or 'Refused'(Not Good & DK/Refused, or DK/Refused & DK Refused) Then I weighted these and created a plot shows these three demographic categories and their multiple subcategories, side-by-side for comparison of responses based on these.

a. Installed and Imported relevant libraries.

b. Imported Dataset

c. Mutated the combination of Q4 and Q5 for new column, 'Globalization_Good_US_and_Pers'

d. Frequency of Occurance Table to Double-check the Conditional Logic in Mutate

e. Viewed Categories of Interest, Pol View, Educ, Income, Collapsing Educ and Income into fewer categories.

f. Selecting only Columns Needed for Analysis: case_id, weight, Globalization_Good_US_and_Pers, Political_View, Income, Education

g. Ran a Gather() to make it easier to plot, consolidating into 'subgroup' column (formerly column label on above 3 demographics, to column contents), and also the original column contents as values in second column 'subgroup_variable'.

h. Calculating weighted estimates of these answers.

i. Plotting results [With an attempt to remove 'refused']

AFTER: 77 rows 4 column dataset.

Plot created shows these three demographic categories and their multiple subcategories, side-by-side for comparison of responses based on these.

4.and 5. Appropriate visualizations (not required). Description of the relationship between the variables, including a hypothesis (or hypotheses) about the relationship. As well as initial demonstration of the relationship, which could include correlation, visualization, or statistical model.

Political View: In the resultant plot 'Globalization, Good for US Past Few Yrs, Personally' for the facet of Globalization being 'Both' Good personally and for US in past few years, there is a deviation from all the rest for the 'Very Conservative' category in that it is distinctly under 30% hovering around 20% and separating itself markedly from the other political view facets. This hovers around 45% for liberal persuasions, a little less but still over 30%. This is exciting for a data sci beginner because it it showing concretely the more 'isolationist' views of Conservatives in aggregation. It would be what I predicted, because it is a match with online writing in conservative, pro-Trump outlets. Liberal outlets do not key in vehemently on isolationism and not 'getting involved' in foreign affairs, thought it's not that they are touting globalism either in my perspective. This more 'middle of the road' stance is shown in the 45%s for the liberals. Ultra-Conservatives that went so far as to answer 'Neither Good Personally or for US in Past Few years' also of course are markedly greater as these are the majority of the answers for this group, pro-Isolationist: at over 60%, while the liberal portion is under 30%.

Educational Level: Interesting that the higher educated bracket, college and post-college educated, were in the lead of rest of group for virtues of Globalism at almost 45% while rest ['High School Grad or less', 'some College'] hovered around 30%. And again this was mirrored for Isolationism 'Neither' as the highest bracket was least Isolationist at 30% and the rest hovered around 45%.

Income Bracket: Income bracket shows a direct relationship between level and positive sentiment towards Globalism as Educational level did. If the respondent made equal or greater than 100k then it exceeded the other lower income brackets at 45% but not much exceeding the between 50k and 100k as this was at around 40%. Less than 50k a year was a little more than 30%, so not as wide a difference for Income Bracket but this might be due to the grouping I did of the Income Categories. As for Isolationism 'Neither' it likewise is a mirror for this, where at 30% highest earners are least isolationist, and at ~35% for 50-<100k earners, and ~35% for <50k earners, shows a slightly more isolationist sentiment.

Synopsis: There appears to be a direct relationship between positive sentiment towards Globalization [level: Low to High] and the continuums of Income and Education Level, as well as the continuum of Political View: (level: Very Conservative to Very Liberal). And is an interesting and potentially rich area of further inquiry that these 3 demographics should parallel each others viewpoint in this way for all 3 demographics.

---

6. (Optional Advanced) Try creating a function that will allow you to easily and accurately implement a repetitive recoding task. (If you are cutting and pasting, use a function!) I looked into this but ran out of time.

7. (Optional Advanced) If you are working on a model, play around with some of the visual diagnostic tools described in RDS. Ran out of time.