Exploring the BRFSS data- SHOBEN

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

Load data

```
load("brfss2013.RData")
```

Part 1: Data

Data contained in the BRFSS (Behavioral Risk Factor Surveillance System) is through an observational study that collects information about non-institutionalized adults (18 years +) across the US by state. This information is collected through telephone interviews of an adult randomly selected from a household. While I was not able to find documentation on whether or not the household called was randomly (was it clustering to help with generaliziability?) selected I was able to find that the defintion of a household and that college housing was included. Cellular phones also stand as a single-person household if called for an interview.

While there may be an attempt for random sampling, I think this study has too many factors that will impact the attempt at being random. First, the non-response bias of a phone interview survey must be substatial. From my personal experiences, I predict more people are going to hang-up rather than answer the questions. Additionally, there is nothing random about the adult of a household who answers the phone. It is likely the household is bias where one adult is more willing or more available to answer. That would ultimately skew these results. Additionally, some adult groups are more likely to physically be home during the call (those who work from home, work unpredictable hours, or are retired) and this again reduces the random nature of the samples. This study does not capture the data from individuals who may have group housing or displaced housing, and possibly with a single shared house phone and no cell phones. These are all biases in the person who answers the phone. Additionally, there are biases in the type of people willing to answer. Some personalities are going to be more willing to coorporate. Any individual whose first language is not English is more likely to refuse or misconceive the question being asked. There are likely sex, gender, prefession, life history, and a number of other factors that will reduce the randomness of these results.

The results of this study can only be used for generalized associations between the samples. And with that, not a robust generalization of the population of interest, but rather simply the population who answers the phone and completes the survey. No causality can be inferred because variables are not controlled and individuals were not assigned to random groups and then studied.

Part 2: Research questions

Research quesion 1: Is there a relationship between sex, race, and number of work hours? Speficially, a relationship between race, sex, and individuals that work overtime (more than 40 hours per week)?

Often individuals from low-income communities work multiple part-time jobs. Their combined hours can amount to well above the average full-time job. Minority races disproportionately make up a large fraction of the lower-class. I'm interested in seeing if data from these phone interviews supports this narrative?

Research quesion 2: Is there an association between race, health status, and health care coverage?

Lacking health care coverage while identifying as having a low health status is a bad situation. Is there a bias across races that idenfiy in this way?

Research quesion 3: Is there a relationship between race and having high blood pressure (hypertension) during pregnancy for females in this study?

Pre-eclampsia is more likely to occur in Black pregnant women than in White pregnant women. This can lead to the death of the child, mother, or both. Does this data show evidence of that? https://www.heart.org/en/news/2019/02/20/why-are-black-women-at-such-high-risk-of-dying-from-pregnancy-complications (https://www.heart.org/en/news/2019/02/20/why-are-black-women-at-such-high-risk-of-dying-from-pregnancy-complications) https://nortonhealthcare.com/news/pregnant-african-american-women-pre-eclampsia/# (https://nortonhealthcare.com/news/pregnant-african-american-women-pre-eclampsia/#):~:text=Pre%2Declampsia%2C%20a%20potentially%20fatal,death%20%E2%80%94%20for%20mother%20and%20baby.

Part 3: Exploratory Data Analysis (Broken down by each question)

Exploratory data analysis Q1

Research quesion 1: Is there a relationship between sex, race, and number of work hours?

Variables: sex: sex of respondent; levels: Male Female scntwrk1: How Many Hours Per Week Do You Work (1-96 are hour, 97 - Don't know/not sure, 98 - zero, 99 - refused)) X_mrace1 - race of individual (excluding ethnicity data)

I am going to limit this analysis to respondants that work more than 40 hours per week, simply because 1) the data is too large otherwise and 2) I think this will still be interesting to consider.

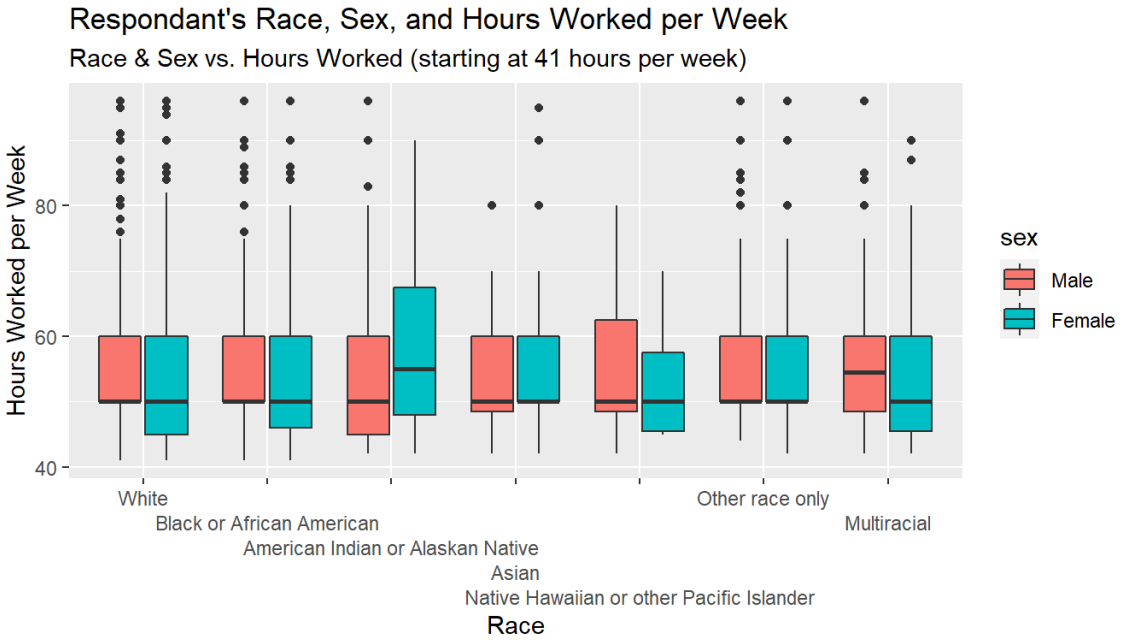
Q1 Selecting and cleaning data

```
# 1) Make new data frame with only variables of interest
q1.work =
  brfss2013 %>%
select(sex, scntwrk1, X_mrace1)

# 2) Remove rows with NAs.
q1.work = q1.work %>%
  drop_na()

# 3) Drop levels for easier manipulation later.
q1.work <- droplevels(q1.work)

# 4) Rename the work column to make it easier to remember.
q1.work =
  q1.work %>%
  rename(hoursworked = scntwrk1)

#5) Remove value of 97, 98, 99 from the work columns which refer to "dont know", "zero", or "refused" re
spectively.
q1.work = q1.work[!(q1.work$hoursworked==97),]
q1.work = q1.work[!(q1.work$hoursworked==98),]
q1.work = q1.work[!(q1.work$hoursworked==99),]

# 6) Filter to only include those who worked over 40 hours per week.
q1.filter <- filter(q1.work, hoursworked > 40)

# 7) Overall summary for the reader
summary(q1.filter)
```

```
##      sex         hoursworked
##  Male  :7469   Min.   :41.00
##  Female:5137   1st Qu.:48.00
##                Median :50.00
##                Mean   :54.17
##                3rd Qu.:60.00
##                Max.   :96.00
##
##                                      X_mrace1
##  White                                   :10566
##  Black or African American               : 1212
##  American Indian or Alaskan Native       :  160
##  Asian                                   :  167
##  Native Hawaiian or other Pacific Islander:   24
##  Other race only                         :  304
##  Multiracial                             :  173
```

Graphs Q1

```
ggplot(q1.filter, aes(x = X_mrace1, y = hoursworked, fill = sex))+
  geom_boxplot(position = "dodge2") +
  scale_x_discrete(guide = guide_axis(n.dodge=5)) +
  labs(title="Respondant's Race, Sex, and Hours Worked per Week", subtitle = "Race & Sex vs. Hours Worke
d (starting at 41 hours per week)") +
  xlab("Race") +
  ylab("Hours Worked per Week"
  )
```



Conclusion Q1

Considering those that work more than 40 hours per week as "over time". It appears that mean number of hours worked per week (for those who work overtime), even when stratified by sex, are similar between White, Black, Asian, Native Hawaiian/Pacific Islander, and other race. A difference in sexes is seen for American Indian or Alaskan Native, with female average work hours being higher. Female American Indian or Alaskan Native seem to have the highest skew of overtime hours, with all responses being captured in the whiskers of the plot (no outliers). In other words, the third quartile range for Native American or Alaskan Native females spans well into the range of outliers for all other groups, indicating that this high number of work hours is more common for this group. Average work hours for multiracial males that worked overtime was higher than their female counterparts.

I think that it would be useful if the survey asked if individuals worked more than one job and I expect we would have seen a trend across both race and sex. I also wonder how many people this missed, especially those that work overtime as I imagine they would have been working when these interviews were conducted. I also wonder about how this data is skewed by priviledged races that are more likely to have a salary job or a job with flexibility where they could step away for a call during the day. I personally have worked multiple jobs in food and did not have the ability to just walk away when needed. These types of jobs (essential workers) are more likely to be filled by more marginalized races and thus more likely to have been missed during this data collection.

Other limitations to this interpretation include what I have already expressed at the beginning of this doc (sample bias, response bias, use of contact/data collection bias, etc.)

I do not believe these results can be generalized to the American people as a whole, and further data would need to be collected (randomized sampling) to speculate further.

Exploratory data analysis Q2

Research quesion 2:

Is there an association between race, health status, and health care coverage?

X_mrace1: race of individual GENHLTH: 1 excellent, 2 very good, 3 good, 4 fair, 5 poor, 7 dont know, 9 refused. HLTHPLN1: health care coverage. 1- Yes, 2- No, 7- dont know 9- refused

Is there a bias by race for having poor health and no health care coverage?

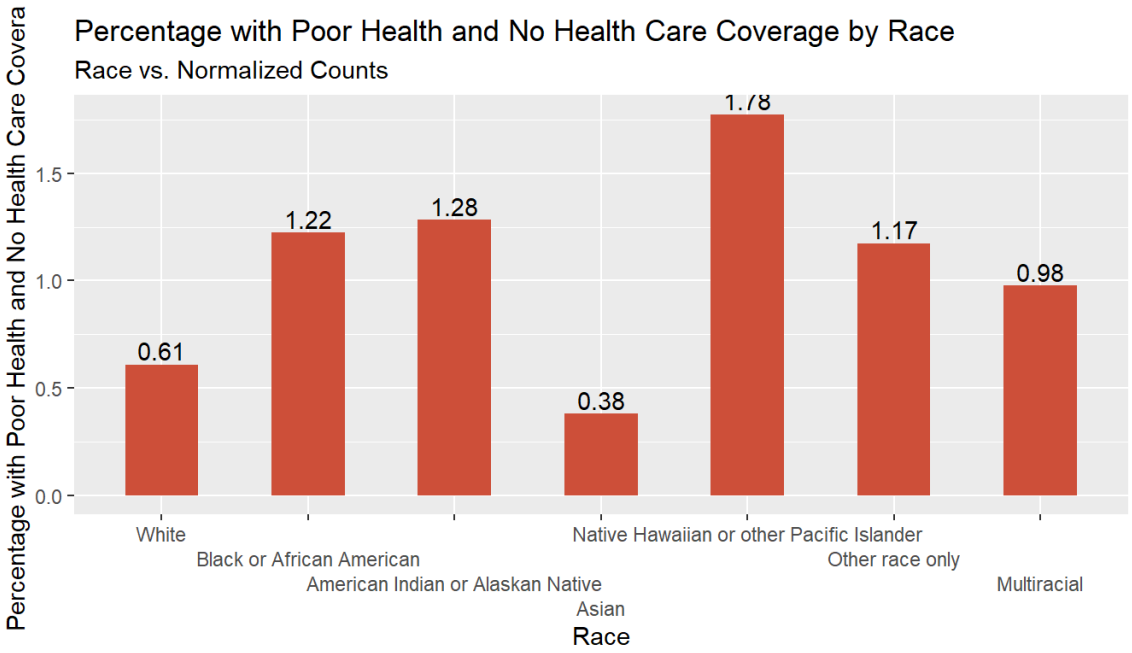
Q2 Selecting and cleaning data

```
# 1) Select for the three variables of interest, remove NAs, and drop levels
q2.data =
  brfss2013 %>%
select(X_mrace1, genhlth, hlthpln1)
q2.data = q2.data %>%
  drop_na()

# 2) Filter for the respondents that meet the criteria of interest (low health and no health insurance).
keep <- c("Poor")
q2.filter <- filter(q2.data, genhlth %in% keep)
q2.filter <- filter(q2.filter, hlthpln1 == "No")

# 3) Need to normalize if I am going to look across races. My goal is to divide the total number of indi
viduals from each race in q2.filter by the total number of individuals for each race that answered both
 the question on general health and health insurance (q2.data). Then multiply that value by 100 to get t
he percentage of respondents fitting my criteria from each race category. I will round these percentages
for easy graphing later.
totalRaceQ2 <- count(q2.data, X_mrace1)
races <- data.frame(totalRaceQ2$X_mrace1)
filterRaceQ2 <- count(q2.filter, X_mrace1)
tot.race <- data.frame(Race = races$totalRaceQ2.X_mrace1, Total = totalRaceQ2$n, Filter = filterRaceQ2$
n)
tot.race$Normalized <- tot.race$Filter/tot.race$Total
tot.race <- tot.race %>%
  mutate(Percentage = Normalized*100)
tot.race$PercentageRound = round(tot.race$Percentage, digit = 2)

# 4) Summary of final data frame
summary(tot.race)
```

```
##                                        Race       Total
##  White                                   :1   Min.   :  1915
##  Black or African American               :1   1st Qu.:  9366
##  American Indian or Alaskan Native       :1   Median : 10127
##  Asian                                   :1   Mean   : 68546
##  Native Hawaiian or other Pacific Islander:1   3rd Qu.: 25676
##  Other race only                         :1   Max.   :397697
##  Multiracial                             :1
##      Filter         Normalized         Percentage     PercentageRound
##  Min.   :  34.0   Min.   :0.003788   Min.   :0.3788   Min.   :0.380
##  1st Qu.:  68.0   1st Qu.:0.007927   1st Qu.:0.7927   1st Qu.:0.795
##  Median : 115.0   Median :0.011741   Median :1.1741   Median :1.170
##  Mean   : 475.0   Mean   :0.010600   Mean   :1.0600   Mean   :1.060
##  3rd Qu.: 311.5   3rd Qu.:0.012532   3rd Qu.:1.2532   3rd Qu.:1.250
##  Max.   :2417.0   Max.   :0.017755   Max.   :1.7755   Max.   :1.780
##
```

Q2 graph

```
ggplot(tot.race, aes(x=Race, y=Percentage)) +
  geom_bar(stat="identity", width=.5, fill="tomato3") +
  labs(title="Percentage with Poor Health and No Health Care Coverage by Race", subtitle = "Race vs. Nor
malized Counts") +
  xlab("Race") +
  ylab("Percentage with Poor Health and No Health Care Coverage"
  ) +
  scale_x_discrete(guide = guide_axis(n.dodge=4)) +
  geom_text(aes(label=PercentageRound), position=position_dodge(width=0.9), vjust=-0.25)
```



Conclusion Q2

I was interested in seeing if identifying as having poor health and no health care coverage varied by racial groups.

The graph produced shows that Native Hawaiian/Pacific Islander's were most likely to report having poor health and no health care coverage, followed by a close triple tie between American Indian/Alaskan Native, other race, and Black. Lowest was amongst the Asian respondants. I am not totally surprised by these results, although the fact that the percentage of Native Hawaiians/Pacific Islanders is triple that of White and 4.7 times that of Asian is rather surprising. I'm curious about the make of the multiracial group. And curious as to why the those of Asian race reported to have poor health/no coverage the least. In addition to variation in racial priviledges, I wonder how much these responses are varied by trust in the respondants. Is it

possible that even more are poor health/no coverage but they responded to the survey falsely/incorrectly? Another data point that would have been good to compare is if some of their racial differences are due to poverty/income. However, I suspect not all racial differences would go away, as there are many systemic reasons for why the medical system fails communities of color.

We can NOT place a casual connection and we can only generalize these results to the respondents, which is not to extend to the wider population for reasons mentioned at the beginning of this document. We would need to do more research, ideally a randomized sampling of the population. It would be unethical to do any type of experiment in this realm.

Exploratory data analysis Q3

Research quesion 3: Is there a relationship between race and having high blood pressure (hypertension) during pregnancy for females in this study? Women of color, specifically Black American women, are at higher risk for pregnancy complications when compared to other groups. Some reportings have pointed at the stress of a being a women of color in America as the problem. Pre-eclampsia is more likely to occur in Black women, which can be fatal for the mother and baby. A symptom of pre-eclampsia is high blood pressure. High blood pressure is commonly referred to as hypertension, both terms will be used in this analysis. Do the data in this study show a relationship between race and high blood pressure during pregnancy? https://www.heart.org/en/news/2019/02/20/why-are-black-women-at-such-high-risk-of-dying-from-pregnancy-complications (https://www.heart.org/en/news/2019/02/20/why-are-black-women-at-such-high-risk-of-dying-from-pregnancy-complications) https://nortonhealthcare.com/news/pregnant-african-american-women-pre-eclampsia/# (https://nortonhealthcare.com/news/pregnant-african-american-women-pre-eclampsia/#):~:text=Pre%2Declampsia%2C%20a%20potentially%20fatal,death%20%E2%80%94%20for%20mother%20and%20baby.

Do we see a pattern to support this claim from the article above? : "White women and Hispanic women had substantially the same rate of the disease. Asian and Pacific Island women had the lowest rate of any ethnic group" (Except hispanic is ethnicitiy and not race. I will only be looking at race.)

I am limiting my data to women who only had hypertension during pregnany.

Variables:
sex : 1- male, 2-female bphigh4 : "Have you EVER been told by a doctor, nurse or other health professional that you have high blood pressure? (If "Yes" and respondent is female, ask "Was this only when you were pregnant?".) X_mrace1 : Reported race (1- White only, 2- Black only, 3- American Indian or Alaskan Native only, 4- Asian only, 5- Native Hawaiian or other Pacific Islander only, 6- other race only, 7- multiracial)

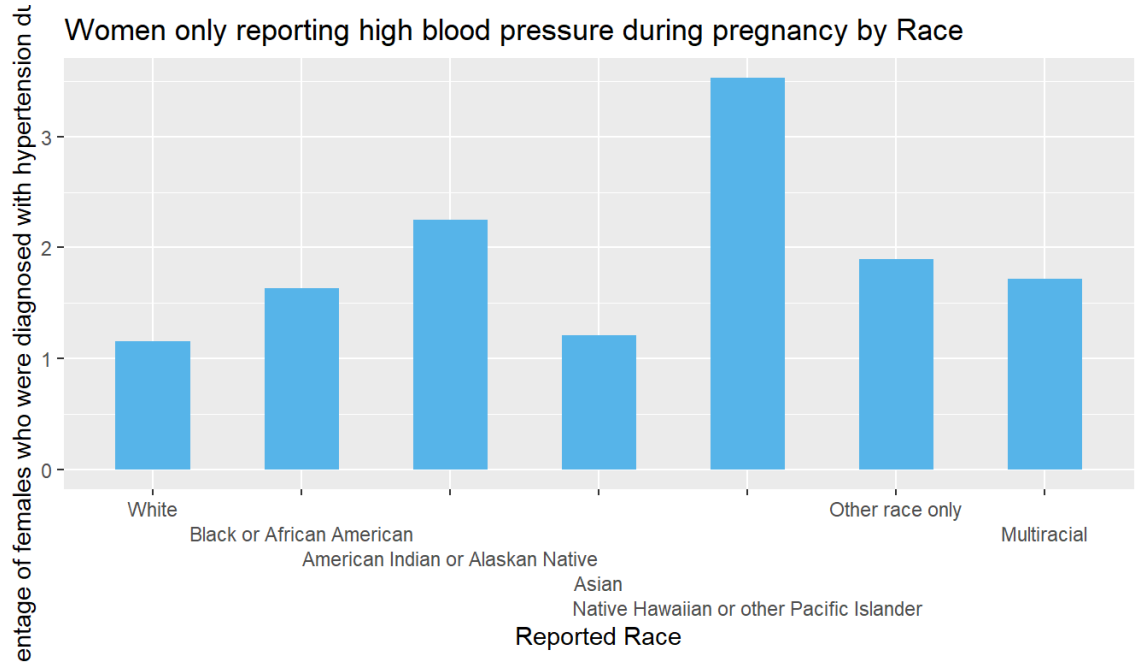
Q3 Selecting and clearning data

```
# 1) Filter to only sex, bphigh4, and X_mrace1 variables. Clean data (remove NAs and refused to respon
d). Drop levels.
q3.data <-
  brfss2013 %>%
  select(X_mrace1, sex, bphigh4) %>%
  drop_na()

q3.data <- droplevels(q3.data)

# 2) Filter only females ($sex = female). Filter only those who reported "yes" to high blood pressure du
ring pregnancy. ($bphigh4" = 2) (2 - Yes, but female told only during pregnancy)
q3.filterfemale <-
  q3.data %>%
  filter(sex == "Female")
q3.filterBpHigh <-
  q3.filterfemale %>%
  filter(bphigh4 == "Yes, but female told only during pregnancy")

# 3) sum total respondents by race, and divide by total respondents by race. (X_mrace1) And determine th
e percentage of female that had higg blood pressure during pregnancy.
percentageByRace <- data.frame(race=count(q3.data, X_mrace1)$X_mrace1, total = count(q3.filterfemale, X_
mrace1)$n, filtered = count(q3.filterBpHigh, X_mrace1)$n)
percentageByRace$normalized <- percentageByRace$filtered/percentageByRace$total
percentageByRace$percentage <- (percentageByRace$normalized * 100)


# 4) summary stats of final data
summary(percentageByRace)
```

```
##                                       race       total
##  White                                  :1   Min.   :  1076
##  Black or African American              :1   1st Qu.:  5098
##  American Indian or Alaskan Native      :1   Median :  5696
##  Asian                                  :1   Mean   : 40750
##  Native Hawaiian or other Pacific Islander:1   3rd Qu.: 16401
##  Other race only                        :1   Max.   :235482
##  Multiracial                            :1
##     filtered        normalized        percentage
##  Min.   :  38.0   Min.   :0.01156   Min.   :1.156
##  1st Qu.:  79.5   1st Qu.:0.01423   1st Qu.:1.423
##  Median : 110.0   Median :0.01721   Median :1.721
##  Mean   : 512.6   Mean   :0.01915   Mean   :1.915
##  3rd Qu.: 279.0   3rd Qu.:0.02074   3rd Qu.:2.074
##  Max.   :2723.0   Max.   :0.03532   Max.   :3.532
##
```

Q3 graph

```
ggplot(percentageByRace, aes(x = race, y = percentage)) +
  geom_bar(stat="identity", width=.5, fill="#56B4E9") +
  labs(title="Women only reporting high blood pressure during pregnancy by Race") +
  xlab("Reported Race") +
  ylab("Percentage of females who were diagnosed with hypertension during pregnancy"
  ) +
  scale_x_discrete(guide = guide_axis(n.dodge=5))
```



Conclusion Q3

Is there a correlation between a woman's race and her having hypertension during pregnancy? An article by Norton Health Care, cited above, stated "White women and Hispanic women had substantially the same rate of the disease. Asian and Pacific Island women had the lowest rate of any ethnic group", while expanding on data that supports that Black women are most at risk for pre-eclampsia. A symptom of pre-eclampsia is high blood pressure.

I'm pretty shocked by the final graph. The highest group from respondants was among Native Hawaiian and Pacfic Islanders. Opposite to the data cited in the article I included in this analysis. Maybe this is skewed due to the Native Hawaiian population being combined with it? Maybe it is due to the form of data collection used with the BRFSS? Or maybe Native Hawaiian and Pacific Islanders have always been at more risk but we tend to talk more about Black versus White health care? I am shocked that the Black percentage was not more stikingly different than the White percentage.

Lowest was among Asian and White Americans. In all reported races, at least 1% of women report to develop hypertension during pregnancy. These data do support that non-white, non-Asian American's are at most risk. Of course, we are only observing among those surveyed and this is not predictive of the American population as a whole. None the less, it is interesting to see if this data supports claims by more rigorously done science.

We can NOT draw any causal conclusions and again, this is an area where conducting an experiment would be un reasonable. However, future research could be (and probably has been) done to ensure that the sampling is more random that this data probably offers.