Week2-Data Management

Matthew Onimus

2020-09-18

Contents


```
library(tidyverse)
library(gtsummary)
library(gt)
library(haven)
library(readxl)
library(rlang)
library(corrplot)
```

1 Reading in the Data

The first thing I needed to do was read the data in join as per the SAS script.

```
nhanesdemo <- read_sav("assignments/NHANES_2015_DEMO.sav")
nhanesbmx <- read_csv("assignments/NHANES_2015_BMX_I.csv")
nhaneshosp <- read_excel("assignments/NHANES_2015_HUQ.xlsx")

# note that all the data is not the same legnth

nhanes <- nhanesdemo %>%
  left_join(nhanesbmx, by = "SEQN") %>%
  left_join(nhaneshosp, by = "SEQN")

tbl_summary(nhanes)
```

Characteristic	**N = 9,971**
Respondent sequence number	88,717 (86,224, 91,210)
Data release cycle	
9	9,971 (100%)
Interview/Examination status	
1	427 (4.3%)
2	9,544 (96%)
Gender	
1	4,892 (49%)
2	5,079 (51%)
Age in years at screening	27 (9, 53)
Age in months at screening - 0 to 24 mos	10 (5, 17)
Unknown	9,276
Race/Hispanic origin	
1	1,921 (19%)
2	1,308 (13%)
3	3,066 (31%)
4	2,129 (21%)
5	1,547 (16%)
Race/Hispanic origin w/ NH Asian	
1	1,921 (19%)
2	1,308 (13%)
3	3,066 (31%)
4	2,129 (21%)
6	1,042 (10%)
7	505 (5.1%)
Six month time period	
1	4,594 (48%)
2	4,950 (52%)
Unknown	427
Age in months at exam - 0 to 19 years	100 (41, 162)
Unknown	5,911
Served active duty in US Armed Forces	
1	527 (8.6%)
2	5,622 (91%)
Unknown	3,822
Served in a foreign country	
1	258 (49%)
2	267 (51%)
7	2 (0.4%)
Unknown	9,444
Country of birth	
1	7,733 (78%)
2	2,236 (22%)
99	2 (<0.1%)
Citizenship status	
1	8,785 (88%)
2	1,168 (12%)
7	9 (<0.1%)
9	7 (<0.1%)
Unknown	2
Length of time in US	5.00 (3.00, 7.00)
Unknown	7,735

Characteristic	**N = 9,971**
Education level - Children/Youth 6-19	5.0 (2.0, 9.0)
Unknown	7,324
Education level - Adults 20+	
1	688 (12%)
2	676 (12%)
3	1,236 (22%)
4	1,692 (30%)
5	1,422 (25%)
9	5 (<0.1%)
Unknown	4,252
Marital status	
1	2,886 (50%)
2	421 (7.4%)
3	614 (11%)
4	192 (3.4%)
5	1,048 (18%)
6	555 (9.7%)
77	2 (<0.1%)
99	1 (<0.1%)
Unknown	4,252
Pregnancy status at exam	
1	70 (5.4%)
2	1,125 (87%)
3	93 (7.2%)
Unknown	8,683
Language of SP Interview	
1	8,584 (86%)
2	1,387 (14%)
Proxy used in SP Interview?	
1	3,689 (37%)
2	6,281 (63%)
Unknown	1
Interpreter used in SP Interview?	
1	457 (4.6%)
2	9,514 (95%)
Language of Family Interview	
1	8,430 (87%)
2	1,212 (13%)
Unknown	329
Proxy used in Family Interview?	
1	9 (<0.1%)
2	9,633 (100%)
Unknown	329
Interpreter used in Family Interview?	
1	405 (4.2%)
2	9,237 (96%)
Unknown	329
Language of MEC Interview	
1	6,382 (91%)
2	595 (8.5%)
Unknown	2,994
Proxy used in MEC Interview?	

Characteristic	**N = 9,971**
1	57 (0.8%)
2	6,921 (99%)
Unknown	2,993
Interpreter used in MEC Interview?	
1	346 (5.0%)
2	6,632 (95%)
Unknown	2,993
Language of ACASI Interview	
1	5,218 (88%)
2	638 (11%)
3	106 (1.8%)
Unknown	4,009
Total number of people in the Household	
1	828 (8.3%)
2	1,723 (17%)
3	1,719 (17%)
4	2,061 (21%)
5	1,672 (17%)
6	994 (10.0%)
7	974 (9.8%)
Total number of people in the Family	
1	1,305 (13%)
2	1,510 (15%)
3	1,634 (16%)
4	2,011 (20%)
5	1,635 (16%)
6	961 (9.6%)
7	915 (9.2%)
# of children 5 years or younger in HH	
0	6,298 (63%)
1	2,147 (22%)
2	1,199 (12%)
3	327 (3.3%)
# of children 6-17 years old in HH	
0	4,715 (47%)
1	1,990 (20%)
2	1,833 (18%)
3	822 (8.2%)
4	611 (6.1%)
# of adults 60 years or older in HH	
0	7,151 (72%)
1	1,663 (17%)
2	1,099 (11%)
3	58 (0.6%)
HH ref person's gender	
1	5,053 (51%)
2	4,918 (49%)
HH ref person's age in years	44 (34, 57)
HH ref person's country of birth	
1	6,359 (66%)
2	3,207 (33%)
77	5 (<0.1%)

Characteristic	**N = 9,971**
99	4 (<0.1%)
Unknown	396
HH ref person's education level	
1	1,087 (11%)
2	1,200 (13%)
3	2,015 (21%)
4	2,908 (30%)
5	2,331 (24%)
9	34 (0.4%)
Unknown	396
HH ref person's marital status	
1	5,681 (57%)
2	524 (5.3%)
3	977 (9.9%)
4	353 (3.6%)
5	1,305 (13%)
6	1,017 (10%)
77	44 (0.4%)
99	8 (<0.1%)
Unknown	62
HH ref person's spouse's education level	
1	619 (12%)
2	511 (9.8%)
3	980 (19%)
4	1,462 (28%)
5	1,629 (31%)
7	2 (<0.1%)
9	23 (0.4%)
Unknown	4,745
Full sample 2 year interview weight	20,160 (12,879, 33,257)
Full sample 2 year MEC exam weight	20,281 (12,551, 33,708)
Masked variance pseudo-PSU	
1	5,127 (51%)
2	4,844 (49%)
Masked variance pseudo-stratum	126.0 (123.0, 130.0)
Annual household income	8.0 (6.0, 14.0)
Unknown	345
Annual family income	8.0 (5.0, 14.0)
Unknown	329
Ratio of family income to poverty	1.82 (0.97, 3.48)
Unknown	1,052
BMDSTATS	
1	8,687 (91%)
2	366 (3.8%)
3	411 (4.3%)
4	80 (0.8%)
Unknown	427
BMXWT	65 (37, 84)
Unknown	526
BMIWT	
1	14 (3.2%)
3	406 (92%)

Characteristic	N = 9,971
4	23 (5.2%)
Unknown	9,528
BMXRECUM	82 (70, 93)
Unknown	8,898
BMIRECUM	33 (100%)
Unknown	9,938
BMXHEAD	41.70 (39.70, 43.20)
Unknown	9,756
BMIHEAD	0 (NA%)
Unknown	9,971
BMXHT	161 (149, 170)
Unknown	1,202
BMIHT	
1	37 (35%)
3	68 (65%)
Unknown	9,866
BMXBMI	25 (20, 31)
Unknown	1,215
BMDBMIC	
1	86 (2.6%)
2	2,041 (61%)
3	561 (17%)
4	652 (20%)
Unknown	6,631
BMXLEG	38.2 (35.3, 41.0)
Unknown	2,861
BMILEG	402 (100%)
Unknown	9,569
BMXARML	36 (30, 38)
Unknown	995
BMIARML	420 (100%)
Unknown	9,551
BMXARMC	30 (22, 34)
Unknown	995
BMIARMC	421 (100%)
Unknown	9,550
BMXWAIST	89 (72, 104)
Unknown	1,658
BMIWAIST	489 (100%)
Unknown	9,482
BMXSAD1	20.9 (17.5, 24.6)
Unknown	2,988
BMXSAD2	20.9 (17.5, 24.6)
Unknown	2,988
BMXSAD3	22.4 (19.2, 25.9)
Unknown	9,618
BMXSAD4	22.7 (19.3, 26.0)
Unknown	9,618
BMDAVSAD	20.9 (17.5, 24.6)
Unknown	2,988
BMDSADCM	
1	436 (98%)

Characteristic	**N = 9,971**
2	2 (0.4%)
3	3 (0.7%)
4	1 (0.2%)
5	4 (0.9%)
Unknown	9,525
HUQ010	
1	2,561 (26%)
2	2,658 (27%)
3	3,001 (30%)
4	1,424 (14%)
5	317 (3.2%)
7	3 (<0.1%)
9	7 (<0.1%)
HUQ020	
1	1,965 (21%)
2	653 (6.8%)
3	6,954 (73%)
9	3 (<0.1%)
Unknown	396
HUQ030	
1	8,563 (86%)
2	1,340 (13%)
3	68 (0.7%)
HUQ041	
1	2,979 (35%)
2	5,093 (59%)
3	277 (3.2%)
4	148 (1.7%)
5	104 (1.2%)
6	30 (0.3%)
Unknown	1,340
HUQ051	2.00 (1.00, 3.00)
HUQ061	
1	114 (8.0%)
2	104 (7.3%)
3	508 (36%)
4	409 (29%)
5	221 (16%)
6	59 (4.2%)
99	6 (0.4%)
Unknown	8,550
HUQ071	
1	869 (8.7%)
2	9,098 (91%)
9	4 (<0.1%)
HUD080	
1	630 (72%)
2	141 (16%)
3	55 (6.3%)
4	18 (2.1%)
5	14 (1.6%)
6	10 (1.2%)

Characteristic	N = 9,971
99999	1 (0.1%)
Unknown	9,102
HUQ090	
1	788 (9.0%)
2	8,000 (91%)
9	6 (<0.1%)
Unknown	1,177

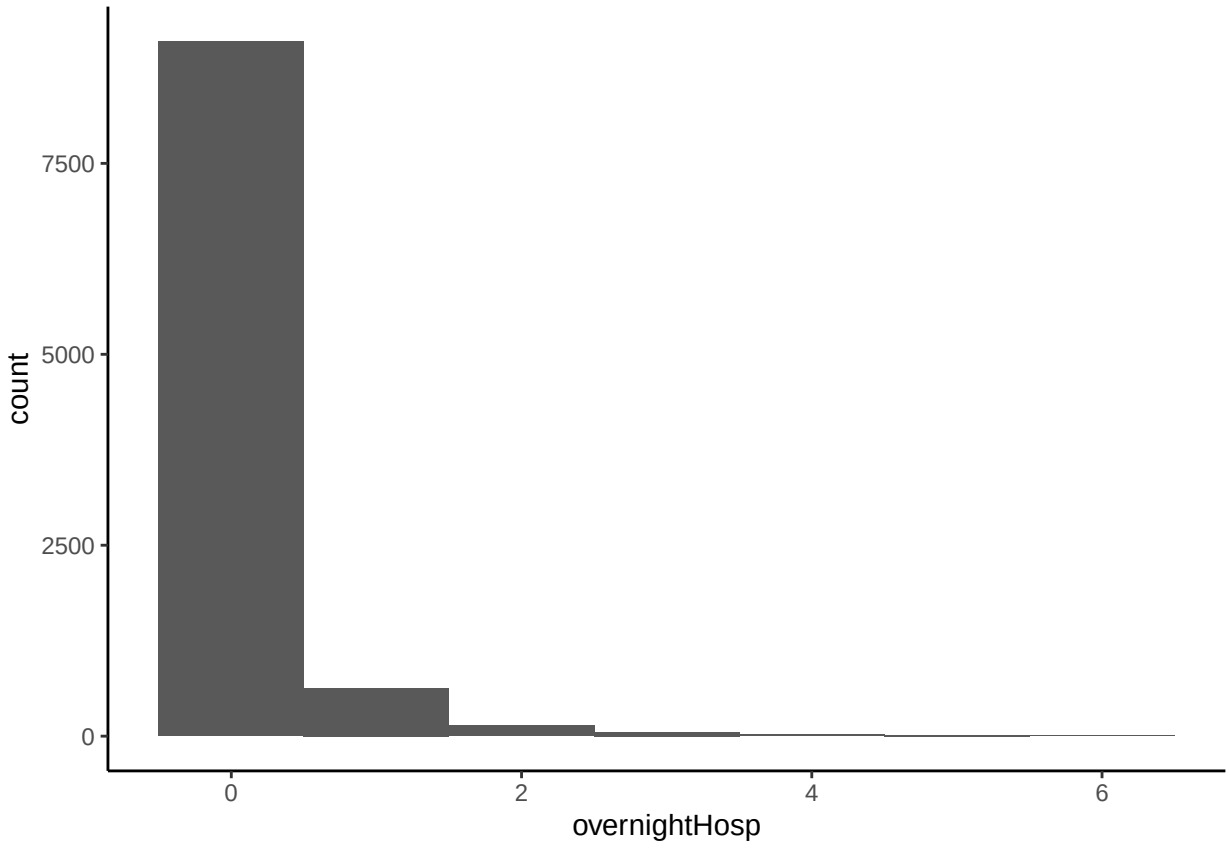
The final, joined data had 81 variables and 9971 observations. A pretty massive dataset.

1.1 1. Review the frequency table for the number of overnight hospital visits and a histogram, what is this data and distribution telling you?

```
nhanes2 <- nhanes %>%
  #select(HUD080, HUQ071) %>%
  mutate(overnightHosp = HUD080,
         overnightHosp = ifelse(HUQ071==7 | HUQ071==9 | is.na(HUQ071), NA, overnightHosp),
         overnightHosp = ifelse(HUQ071 == 1 & (HUD080 == 77777 | HUD080 == 99999), NA, overnightHosp),
         overnightHosp = ifelse(HUQ071 == 2, 0, overnightHosp),
         waistin = BMXWAIST / 2.54,
         children = 0,
         children = ifelse(DMDHHSZA > 0 | DMDHHSZB > 0, 1, children)

  )

nhanesChildren <- nhanes2 %>%
  filter(children == 1)

nhanes2 %>%
  select(overnightHosp) %>%
  group_by(overnightHosp) %>%
  summarise(count = n()) %>%
  gt()
```

overnightHosp	count
0	9098
1	630
2	141
3	55
4	18
5	14
6	10
NA	5

```
ggplot(nhanes2, aes(x = overnightHosp)) +
  geom_histogram(bins = 7) +
  theme_classic()
```



1.1.1 Answer:

The frequency table shows most of the data do not have an overnight hospital stay. The histogram tells the same story. There is a large tail to the data showing the skewness.

1.2 2. What is the median age of respondents? What are the median ages for those who do and do not have a child in the home?

```
median(nhanes2$RIDAGEYR)
```

```
## [1] 27
```

```
nhanes2 %>%
  group_by(children) %>%
  summarise(count = n(),
            median = median(RIDAGEYR))
```

```
## # A tibble: 2 x 3
##   children count median
##      <dbl> <int>  <dbl>
## 1        0  3330     59
## 2        1  6641     14
```

1.2.1 Answer:

The median age for all respondents is 27. The median age for those with children is 14 and without children is 59.

1.3 3. What is the average waist circumference (in inches) for those who had 6+ overnight hospital stays in the past year?

```
nhanes2 %>%
  group_by(overnightHosp) %>%
  summarize(count = n(),
            waist = mean(waistin, na.rm = TRUE)) %>%
  gt()
```

overnightHosp	count	waist
0	9098	34.30683
1	630	38.39725
2	141	40.62992
3	55	39.03864
4	18	39.01247
5	14	34.54068
6	10	36.84055
NA	5	35.27559

1.3.1 Answer:

The average waist size (in inches) for 6 overnight stays is 36.841 inches.

1.4 4. Create a cross-tabulation for the number of overnight hospital stays with race/ethnicity (RIDRETH3). Which group had the lowest proportion on inpatients? The greatest proportion of high utilizers?

```
nhanes2 %>%
  select(overnightHosp, RIDRETH3) %>%
  tbl_summary(by = RIDRETH3, percent = 'cell') %>%
  add_overall() %>%
  as_gt() %>%
  #tab_header(title = "Table of overnighthosp by RIDRETH3") %>%
  tab_spanner(label = "RIDRETH3(Race/Hispanic origin w/ NH Asian)", columns = c(5:10))
```

		RIDRETH3(Race/Hispanic origin w/ NH Asian)				
Characteristic	**Overall**, N = 9,971	**1**, N = 1,921[1]	**2**, N = 1,308[1]	**3**, N = 3,066[1]	**4**, N = 2,129[1]	**6**, N = 1,042[1]
overnightHosp						
0	9,098 (91%)	1,782 (18%)	1,178 (12%)	2,802 (28%)	1,909 (19%)	981 (9.8%)
1	630 (6.3%)	103 (1.0%)	93 (0.9%)	194 (1.9%)	153 (1.5%)	46 (0.5%)
2	141 (1.4%)	24 (0.2%)	21 (0.2%)	44 (0.4%)	35 (0.4%)	10 (0.1%)
3	55 (0.6%)	6 (<0.1%)	9 (<0.1%)	12 (0.1%)	21 (0.2%)	3 (<0.1%)
4	18 (0.2%)	1 (<0.1%)	4 (<0.1%)	6 (<0.1%)	4 (<0.1%)	1 (<0.1%)
5	14 (0.1%)	2 (<0.1%)	0 (0%)	8 (<0.1%)	3 (<0.1%)	0 (0%)
6	10 (0.1%)	2 (<0.1%)	3 (<0.1%)	0 (0%)	2 (<0.1%)	0 (0%)
Unknown	5	1	0	0	2	1

[1]Statistics presented: n (%)

1.4.1 Answer:

The group with the lowest proportion of inpatients is group 7 (505 total). The greatest proportion of was group 3 (3066 total).

2 Exploring Data

Here is the code to replicate the EDA section before the questions.

```
hs15 <- read_sas("assignments/hs15ar1.sas7bdat")

tbl_summary(hs15)
```

Characteristic	**N = 10,048**
Respondent interviewer number	300,787 (105,920, 400,783)
S1 County of residence	
1	1,464 (15%)
2	1,377 (14%)
3	1,425 (14%)
4	1,742 (17%)
5	4,040 (40%)
S2 Zip Code	19,128 (19,055, 19,153)
S2 Zip Code based regions	15 (10, 23)
Unknown	12
S9 Relationship of proxy to respondent	
1	15 (31%)
2	8 (17%)
3	17 (35%)
4	7 (15%)
5	1 (2.1%)
Unknown	10,000
S11.3 Adult respondent's sex	
1	3,891 (39%)
2	6,157 (61%)
S12 Adult respondent's age	53 (43, 64)
S12a Adult respondent's age by 3 categories	
1	6,784 (68%)
2	2,154 (21%)
3	1,110 (11%)
S12a Adult respondent's age by 4 categories	
1	1,163 (12%)
2	2,907 (29%)
3	3,512 (35%)
4	2,466 (25%)
S13 # of related children in household	0.00 (0.00, 1.00)
Unknown	27
S15 Child's Sex	
1	1,892 (53%)
2	1,699 (47%)
Unknown	6,457
S15a Child Hispanic or Latino origin?	
1	314 (8.8%)
2	3,255 (91%)

Characteristic	N = 10,048
Unknown	6,479
S16 Child's Race	
1	2,439 (69%)
2	663 (19%)
3	98 (2.8%)
4	8 (0.2%)
5	236 (6.6%)
6	8 (0.2%)
7	104 (2.9%)
Unknown	6,492
S17 Child's age	11.0 (6.0, 15.0)
Unknown	6,510
S21 Child proxy same as respondent	
1	3,243 (90%)
2	348 (9.7%)
Unknown	6,457
S23 Relationship of proxy to child	
1	2,303 (64%)
2	1,092 (30%)
3	19 (0.5%)
4	118 (3.3%)
5	40 (1.1%)
6	1 (<0.1%)
7	14 (0.4%)
8	1 (<0.1%)
9	3 (<0.1%)
Unknown	6,457
S25 # of unrelated children in household	0.0000 (0.0000, 0.0000)
Unknown	20
S13&S25 # of children in household	0.00 (0.00, 1.00)
Unknown	15
S27 # of adults in household	2.00 (1.00, 2.00)
Unknown	29
S28 # of adults 60+ in household	
0	7,399 (74%)
1	1,168 (12%)
2	1,401 (14%)
3	54 (0.5%)
4	10 (<0.1%)
5	4 (<0.1%)
9	1 (<0.1%)
12	1 (<0.1%)
15	1 (<0.1%)
Unknown	9
S29 # of related adult household members	1.00 (0.00, 1.00)
Unknown	24
Q1 Health status	
1	2,244 (22%)
2	3,239 (32%)
3	2,804 (28%)
4	1,307 (13%)
5	422 (4.2%)

Characteristic	**N = 10,048**
Unknown	32
Q1 Health status, binary	
1	8,287 (83%)
2	1,729 (17%)
Unknown	32
Q2a Ever had Asthma	
1	1,621 (16%)
2	8,409 (84%)
Unknown	18
Q2b Ever had Diabetes	
1	1,368 (14%)
2	8,609 (86%)
7	48 (0.5%)
Unknown	23
Q3 Ever told by doctor have high BP	
1	3,544 (35%)
2	6,444 (64%)
7	31 (0.3%)
Unknown	29
Q4 Now taking meds for high BP	
1	3,085 (87%)
2	457 (13%)
Unknown	6,506
Q5 Taking high BP meds as prescribed	
1	2,693 (87%)
2	240 (7.8%)
3	102 (3.3%)
4	24 (0.8%)
5	23 (0.7%)
Unknown	6,966
Q6 Has regular source of care	
1	9,109 (91%)
2	902 (9.0%)
Unknown	37
Q7 Where go for regular source of care	
1	7,780 (85%)
2	527 (5.8%)
3	429 (4.7%)
4	127 (1.4%)
97	215 (2.4%)
98	29 (0.3%)
99	2 (<0.1%)
Unknown	939
Q8 # of visits to healthcare provider in past yr	2.0 (1.0, 4.0)
Q8 # of visits to healthcare provider in past yr (2 cat)	8,791 (89%)
Unknown	173
Q8 # of visits to healthcare provider in past yr (3 cat)	
0	1,084 (11%)
1	4,160 (42%)
2	4,631 (47%)
Unknown	173
Q9 # of visits to emergency room in past yr	0.00 (0.00, 1.00)

Characteristic	**N = 10,048**
Unknown	32
Q9 Any visits to emergency room in past year	2,695 (27%)
Unknown	32
Q10 Didn't go to needed dr appt due to transportation problems	
1	861 (8.6%)
2	9,171 (91%)
Unknown	16
Q11a Insured by work, school, union	
1	5,890 (59%)
2	4,091 (41%)
Unknown	67
Q11b Insured by self or family (including w gov assist)	
1	5,140 (51%)
2	4,763 (47%)
8	132 (1.3%)
9	13 (0.1%)
Q11c Insured by Medicare A	
1	3,110 (31%)
2	6,687 (67%)
8	241 (2.4%)
9	10 (<0.1%)
Q11d Insured by Medicare B	
1	2,950 (29%)
2	6,794 (68%)
8	295 (2.9%)
9	9 (<0.1%)
Q11e Insured by Medicaid	
1	1,318 (13%)
2	8,445 (84%)
8	278 (2.8%)
9	7 (<0.1%)
Q11f Insured by Champus or Tricare	
1	218 (2.2%)
2	9,697 (98%)
Unknown	133
Q11g Insured by other group	
1	2,404 (24%)
2	7,493 (76%)
Unknown	151
Adult respondent insurance status	
1	9,611 (96%)
2	437 (4.3%)
Insurance status of those 18-64	
1	7,149 (94%)
2	433 (5.7%)
Unknown	2,466
Q12 Currently have health insurance	
1	22 (4.9%)
2	429 (95%)
Unknown	9,597
Q13 Of insured, uninsured in past yr	
1	571 (6.0%)

Characteristic	N = 10,048
2	9,024 (94%)
Unknown	453
Q14 Of insured, how long uninsured	
1	308 (55%)
2	89 (16%)
3	55 (9.9%)
4	105 (19%)
Unknown	9,491
Q15 Name of insurance company	21 (16, 53)
Unknown	437
Q16 Looked into buying ins through healthcare.gov	
1	1,565 (16%)
2	8,438 (84%)
Unknown	45
Q17 Find plan with monthly premiums could afford	
1	598 (41%)
2	389 (26%)
3	489 (33%)
Unknown	8,572
Q18 Find plan could afford to USE, copays and deductibles	
1	611 (41%)
2	385 (26%)
3	477 (32%)
Unknown	8,575
Q19 Enrolled in plan through healthcare.gov	
1	563 (37%)
2	976 (63%)
Unknown	8,509
Q20 Why no insurance now	5 (5, 10)
Unknown	9,635
Q21 Of uninsured, how long no coverage	
1	87 (20%)
2	34 (7.9%)
3	89 (21%)
4	218 (51%)
Unknown	9,620
Q22 Of uninsured, Medicaid past year	
1	47 (39%)
2	74 (61%)
Unknown	9,927
Q23 Have prescription med coverage	
1	8,763 (87%)
2	1,164 (12%)
8	116 (1.2%)
9	5 (<0.1%)
Q25 Sick but not seek care due to cost	
1	864 (8.6%)
2	9,170 (91%)
Unknown	14
Q26 No prescription med due to cost	
1	1,255 (13%)
2	8,770 (87%)

Characteristic	N = 10,048
Unknown	23
Q27 Cut meal due to lack of money	
1	223 (6.9%)
2	3,032 (93%)
Unknown	6,793
Q28a Time since last visit to dentist	
1	7,264 (73%)
2	1,093 (11%)
3	476 (4.8%)
4	410 (4.1%)
5	334 (3.3%)
6	388 (3.9%)
7	39 (0.4%)
Unknown	44
Q28b Time since last blood pressure reading	
1	9,244 (93%)
2	436 (4.4%)
3	98 (1.0%)
4	78 (0.8%)
5	40 (0.4%)
6	29 (0.3%)
7	38 (0.4%)
Unknown	85
Q28c Time since last pap test	
1	3,302 (55%)
2	1,057 (17%)
3	445 (7.3%)
4	366 (6.0%)
5	296 (4.9%)
6	402 (6.6%)
7	187 (3.1%)
Unknown	3,993
Q28c Had a Pap smear in past yr	
1	3,302 (55%)
2	2,753 (45%)
Unknown	3,993
Q28d Time since last breast exam by doctor	
1	4,165 (68%)
2	816 (13%)
3	296 (4.9%)
4	227 (3.7%)
5	173 (2.8%)
6	165 (2.7%)
7	240 (3.9%)
Unknown	3,966
Q28d Had a breast exam in past yr	
1	4,165 (68%)
2	1,917 (32%)
Unknown	3,966
Q28e Time since last mammogram	
1	3,197 (63%)
2	734 (15%)

Characteristic	N = 10,048
3	266 (5.3%)
4	250 (4.9%)
5	179 (3.5%)
6	144 (2.8%)
7	289 (5.7%)
Unknown	4,989
Q28e Had a mammogram in past yr	
1	3,197 (63%)
2	1,862 (37%)
Unknown	4,989
Q28f Time since last prostate exam	
1	1,409 (52%)
2	333 (12%)
3	165 (6.1%)
4	169 (6.2%)
5	107 (3.9%)
6	62 (2.3%)
7	467 (17%)
Unknown	7,336
Q28f Had a prostate exam in past yr	
1	1,409 (52%)
2	1,303 (48%)
Unknown	7,336
Q28g Time since last sigmoidoscopy or colonoscopy	
1	1,033 (18%)
2	747 (13%)
3	655 (11%)
4	928 (16%)
5	815 (14%)
6	255 (4.3%)
7	1,453 (25%)
Unknown	4,162
Q28g Had a sigmoidoscopy or colonoscopy in past yr	
1	1,033 (18%)
2	4,853 (82%)
Unknown	4,162
Q29 Ever been tested for HIV	
1	4,541 (45%)
2	5,015 (50%)
8	482 (4.8%)
9	10 (<0.1%)
Q29a Main reason for HIV test	6 (5, 13)
Unknown	5,639
Q29b Where HIV test received	8 (6, 9)
Unknown	5,507
Q31 Weight without shoes	170 (145, 200)
Unknown	193
Q32. Height in Feet	
4	222 (2.2%)
5	8,412 (84%)
6	1,343 (13%)
Unknown	71

Characteristic	N = 10,048
Q32. Height: Inches Added to Feet	5.0 (2.0, 8.0)
Unknown	74
Q32 Body Mass Index level	26.9 (23.7, 31.0)
Unknown	229
Q32 Obesity level	
1	135 (1.4%)
2	3,292 (33%)
3	3,499 (35%)
4	3,051 (31%)
Unknown	71
Q33 Shop at store that stopped selling tobacco products	
1	1,215 (13%)
2	405 (4.2%)
3	8,011 (83%)
Unknown	417
Q34 Smoked at least 100 cigarettes	
1	4,374 (44%)
2	5,618 (56%)
Unknown	56
Q35 How often smoke	
1	973 (9.7%)
2	507 (5.1%)
3	2,885 (29%)
4	5,618 (56%)
Unknown	65
Q35 Smoke cigarettes, binary	
1	1,480 (15%)
2	8,503 (85%)
Unknown	65
Q36 # of cigarettes per day	10 (8, 20)
Unknown	9,535
Q37 Use coupons to buy cigarettes	
1	410 (52%)
2	93 (12%)
3	199 (25%)
4	93 (12%)
Unknown	9,253
Q38 Tried to quit smoking in the past yr	
1	863 (59%)
2	609 (41%)
Unknown	8,576
Q39 How tried to quit smoking	1.0 (1.0, 3.0)
Unknown	9,193
Q40 Anyone smoke inside home	
1	962 (9.6%)
2	9,075 (90%)
Unknown	11
Q41 Use other tobacco products	
1	328 (3.3%)
2	9,717 (97%)
Unknown	3
Q42 How many times used e-cigarette in past month	

Characteristic	N = 10,048
1	9,519 (95%)
2	160 (1.6%)
3	163 (1.6%)
4	21 (0.2%)
5	70 (0.7%)
6	76 (0.8%)
Unknown	39
Q43 Days per week exercise > 30 mins	
1	1,592 (16%)
2	677 (6.8%)
3	2,639 (26%)
4	5,071 (51%)
Unknown	69
Q44 # of servings of fruits & vegetables	2.00 (2.00, 4.00)
Unknown	273
Q45 # of times drink soda (Phila only)	
1	300 (7.5%)
2	319 (8.0%)
3	684 (17%)
4	702 (18%)
5	1,998 (50%)
Unknown	6,045
Q49 # of times drink fruit drinks or bottled teas (Phila only)	
1	363 (9.1%)
2	348 (8.7%)
3	642 (16%)
4	442 (11%)
5	2,197 (55%)
Unknown	6,056
Q49 # of times drink soda, fruit drinks, bottled teas (Phila only)	
1	583 (14%)
2	535 (13%)
3	854 (21%)
4	636 (16%)
5	1,422 (35%)
Unknown	6,018
Q53 How hard to find fruit and vegetables	
1	6,415 (64%)
2	3,071 (31%)
3	335 (3.4%)
4	127 (1.3%)
Unknown	100
Q54 Quality of groceries in neighborhood	
1	5,065 (51%)
2	3,643 (37%)
3	943 (9.5%)
4	244 (2.5%)
7	42 (0.4%)
Unknown	111
Q55 Currently watching or reducing salt intake	
1	5,648 (57%)
2	4,327 (43%)

Characteristic	N = 10,048
Unknown	73
Q56 Buy items labeled 'low salt' or 'low sodium'	
1	1,593 (16%)
2	2,302 (23%)
3	2,978 (30%)
4	1,480 (15%)
5	1,511 (15%)
6	48 (0.5%)
Unknown	136
Q57 Think too much salt harmful to health	
1	773 (7.7%)
2	1,611 (16%)
3	3,818 (38%)
4	3,545 (35%)
8	240 (2.4%)
9	13 (0.1%)
Unknown	48
Q58 Think too much salt affects risk of stroke	
1	903 (9.0%)
2	1,307 (13%)
3	3,048 (30%)
4	4,087 (41%)
8	633 (6.3%)
9	22 (0.2%)
Unknown	48
Q59 # of times per week eat fast food	0.00 (0.00, 1.00)
Unknown	25
Q60 # of hrs per day screen time	
0	49 (1.2%)
0.5	223 (5.6%)
1	415 (10%)
2	876 (22%)
3	750 (19%)
4	511 (13%)
5	1,184 (30%)
Unknown	6,040
Q61 Has diagnosed mental health condition	
1	1,619 (16%)
2	8,378 (84%)
Unknown	51
Q62 Receive treatment for mental health condition	
1	1,015 (63%)
2	598 (37%)
Unknown	8,435
Q64 # of organizations currently participating in	1.00 (0.00, 2.00)
Unknown	48
Q65 People in my neighborhood can be trusted	
1	2,277 (23%)
2	5,186 (52%)
3	1,397 (14%)
4	522 (5.2%)
8	592 (5.9%)

Characteristic	N = 10,048
9	26 (0.3%)
Unknown	48
Q66 Neighbors willing to help each other	
1	2,560 (26%)
2	3,128 (31%)
3	2,696 (27%)
4	824 (8.2%)
5	517 (5.2%)
8	254 (2.5%)
9	21 (0.2%)
Unknown	48
Q67 Neighbors ever worked together	
1	5,859 (59%)
2	3,712 (37%)
8	415 (4.2%)
9	14 (0.1%)
Unknown	48
Q68 I feel that I belong in my neighborhood	
1	3,134 (31%)
2	5,457 (55%)
3	910 (9.1%)
4	283 (2.8%)
8	196 (2.0%)
9	20 (0.2%)
Unknown	48
Q69 Nearby park or outdoor space you're comfortable visiting	
1	7,702 (78%)
2	2,169 (22%)
Unknown	177
Q71 Past month, provide care to family member or friend with illness or disability	
1	3,448 (34%)
2	6,576 (66%)
Unknown	24
Q72 Ever attended CPR training	
1	6,472 (65%)
2	3,501 (35%)
3	3 (<0.1%)
Unknown	72
Q 73 When did you last attend CPR training	
1	1,926 (30%)
2	1,152 (18%)
3	1,193 (19%)
4	2,175 (34%)
Unknown	3,602
Q74 CPR level of certification	
1	2,397 (37%)
2	1,433 (22%)
3	2,368 (37%)
4	211 (3.3%)
Unknown	3,639
Q76 Past three months, used social media	

Characteristic	N = 10,048
1	6,022 (60%)
2	3,957 (40%)
Unknown	69
Q77 Used social media to connect with community orgs providing services	
1	2,582 (43%)
2	3,421 (57%)
Unknown	4,045
Q78 Preferred way to recieve health or social service information	2.00 (1.00, 4.00)
Unknown	48
Q124 Adult respondent is main wage earner	
1	6,338 (64%)
2	3,539 (36%)
Unknown	171
Q125 Employment status of main wage earner	
1	6,025 (61%)
2	634 (6.4%)
3	248 (2.5%)
4	79 (0.8%)
5	2,213 (22%)
6	636 (6.4%)
7	64 (0.6%)
8	57 (0.6%)
Unknown	92
Q126 Adult respondent's education status	
1	582 (5.8%)
2	2,658 (27%)
3	276 (2.8%)
4	1,892 (19%)
5	2,512 (25%)
6	2,068 (21%)
Unknown	60
Q127 Adult respondent's employment status	
1	4,576 (46%)
2	1,195 (12%)
3	444 (4.4%)
4	149 (1.5%)
5	2,286 (23%)
6	824 (8.3%)
7	370 (3.7%)
8	141 (1.4%)
Unknown	63
Q128 Rent or own home	
1	2,017 (20%)
2	7,454 (75%)
3	504 (5.1%)
Unknown	73
Q130 How difficult to afford housing costs in past yr	
1	1,032 (11%)
2	2,867 (29%)
3	2,372 (24%)
4	3,397 (35%)
5	125 (1.3%)

Characteristic	**N = 10,048**
Unknown	255
Q131 Adult respondent's marital status	
1	5,311 (53%)
2	419 (4.2%)
3	1,030 (10%)
4	784 (7.9%)
5	240 (2.4%)
6	2,047 (21%)
7	118 (1.2%)
Unknown	99
Q132 Sex identity	
1	9,412 (94%)
2	132 (1.3%)
3	102 (1.0%)
4	63 (0.6%)
8	105 (1.0%)
9	234 (2.3%)
Q133 Adult respondent of Hispanic or Latino origin	
1	539 (5.5%)
2	9,337 (95%)
Unknown	172
Q134 Original question about adult respondents race	
1	6,942 (69%)
2	2,135 (21%)
3	173 (1.7%)
4	54 (0.5%)
5	302 (3.0%)
7	208 (2.1%)
97	12 (0.1%)
98	31 (0.3%)
99	191 (1.9%)
Q134 Race categories combined	
1	6,942 (71%)
2	2,135 (22%)
3	173 (1.8%)
4	576 (5.9%)
Unknown	222
Q133&Q134 Race categories with Latino separate	
1	6,789 (69%)
2	2,086 (21%)
3	539 (5.5%)
4	170 (1.7%)
5	210 (2.1%)
6	45 (0.5%)
Unknown	209
Q136 Country of origin	1 (1, 1)
Unknown	97
Q137 Speak other language at home	
1	1,183 (12%)
2	8,778 (88%)
Unknown	87
Q138 What other language speak at home	2 (1, 11)

Characteristic	N = 10,048
Unknown	8,876
Q139 Veteran status	
1	883 (8.9%)
2	9,067 (91%)
Unknown	98
Q140a Receive SSI	
1	1,016 (10%)
2	8,800 (90%)
Unknown	232
Q140b Receive SSDI	
1	1,201 (12%)
2	8,650 (88%)
Unknown	197
Q140d Receive food stamps	
1	1,390 (14%)
2	8,518 (86%)
Unknown	140
Q140f Receive WIC	
1	254 (2.6%)
2	9,638 (97%)
Unknown	156
Q140g Receive TANF	
1	97 (1.0%)
2	9,760 (99%)
Unknown	191
Q140x # of people living in household	3.00 (2.00, 4.00)
S13&S29 Family size	2.00 (1.00, 4.00)
Unknown	2
Q146 Family income by category	21 (13, 24)
Unknown	1,926
Less than 50% poverty	
1	329 (3.3%)
2	9,717 (97%)
Unknown	2
Less than 100% poverty	
1	1,132 (11%)
2	8,914 (89%)
Unknown	2
Less than 150% poverty	
1	1,918 (19%)
2	8,128 (81%)
Unknown	2
Less than 200% poverty	
1	2,710 (27%)
2	7,336 (73%)
Unknown	2
Q148 Has personal cell phone	
1	6,838 (87%)
2	1,054 (13%)
Unknown	2,156
Q149 Cell phone is only phone	
1	1,006 (51%)

Characteristic	N = 10,048
2	980 (49%)
Unknown	8,062
Q150 Cell phone calls being received	
1	3,142 (40%)
2	2,465 (32%)
3	2,171 (28%)
Unknown	2,270
Language of interview	
1	9,885 (98%)
2	162 (1.6%)
Unknown	1
Landline or cell phone interview	
1	8,019 (80%)
2	2,029 (20%)
Service Area	27 (13, 40)
Weighting Area	
1	1,431 (14%)
2	1,403 (14%)
3	1,420 (14%)
4	1,750 (17%)
5	1,452 (14%)
6	1,308 (13%)
7	1,284 (13%)
Adult Projection Weight	149 (53, 465)
Adult Balance Weight	0.47 (0.17, 1.47)
Household Projection Weight	80 (27, 248)
Household Balance Weight	0.51 (0.17, 1.59)

```
hs15 %>%
  arrange(WSHOES) %>%
  head(n = 10) %>%
  gt()
```

INTNUM	COUNTY	ZIPCODE	ZIPREGION	APROXY	SEX01	RESPAGE	AGER1	AGER2	NUMKIDS
202844	1	19053	16	NA	1	51	1	3	1
105349	5	19128	8	NA	2	80	3	4	0
104675	4	19046	26	NA	2	26	1	1	0
201234	4	19464	31	NA	2	75	3	4	2
403173	5	19144	9	NA	2	56	1	3	0
105380	5	19139	4	NA	2	89	3	4	0
404614	5	19115	12	NA	2	92	3	4	0
301047	4	19525	30	NA	2	49	1	2	0
100643	4	19096	27	NA	2	71	2	4	0
100511	5	19135	11	NA	2	82	3	4	0

```
hs15 %>%
  arrange(desc(WSHOES)) %>%
  head(n = 10) %>%
  gt()
```

INTNUM	COUNTY	ZIPCODE	ZIPREGION	APROXY	SEX01	RESPAGE	AGER1	AGER2	NUMKIDS
305701	5	19120	10	NA	2	50	1	3	0
203872	2	19335	20	NA	1	47	1	2	0
306271	5	19151	4	NA	2	38	1	2	0
404562	5	19146	2	NA	2	52	1	3	0
400582	5	19129	9	NA	2	33	1	1	0
301939	5	19140	6	NA	1	48	1	2	0
201604	5	19131	4	NA	2	29	1	1	0
102265	5	19104	4	NA	1	37	1	2	0
103058	2	19355	17	NA	1	52	1	3	0
204175	1	18920	14	NA	1	33	1	1	2

2.1 5. Run the proc means and interpret the descriptive statistics for weight without shoes and height in feet by gender (1=male and 2= female)

```
hs15 %>%
  select(SEX01, WSHOES, HFEET) %>%
  group_by(SEX01) %>%
  summarize(across(everything(), list(mean = mean, sd = sd,
                                      min = min, max = max,
                                      median = median),
                   na.rm = TRUE)) # can add move if needed
```

```
## # A tibble: 2 x 11
##   SEX01 WSHOES_mean WSHOES_sd WSHOES_min WSHOES_max WSHOES_median HFEET_mean
##   <dbl>       <dbl>     <dbl>      <dbl>      <dbl>         <dbl>      <dbl>
## 1     1        198.      40.9         65        475           190       5.33
## 2     2        163.      41.1         79        490           155       4.97
## # ... with 4 more variables: HFEET_sd <dbl>, HFEET_min <dbl>, HFEET_max <dbl>,
## #   HFEET_median <dbl>
```

2.1.1 Answer:

For the stats I was able to generate quickly, Sex 1 has a higher mean weight, smaller sd. Sex 1 also has a higher mean for height but a larger sd, meaning the data may be more spread. The min and max for weight and height are pretty similar.

2.2 6. What are differences between 95% and 99% Confidence Intervals? Report the CI's for both variables. What does this mean, interpret the results

```
hs15 %>%
  select(SEX01, WSHOES, HFEET) %>%
  group_by(SEX01) %>%
  summarize(lower95CI_WSHOES = gmodels::ci(WSHOES, na.rm = TRUE)[[2]],
            upper95CI_WSHOES = gmodels::ci(WSHOES, na.rm = TRUE)[[3]],
            lower99CI_WSHOES = gmodels::ci(WSHOES, confidence = 0.99, na.rm = TRUE)[[2]],
            upper99CI_WSHOES = gmodels::ci(WSHOES, confidence = 0.99, na.rm = TRUE)[[3]],
            lower95CI_HFEET = gmodels::ci(HFEET, na.rm = TRUE)[[2]],
            upper95CI_HFEET = gmodels::ci(HFEET, na.rm = TRUE)[[3]],
            lower99CI_HFEET = gmodels::ci(HFEET, confidence = 0.99, na.rm = TRUE)[[2]],
```

```
            upper99CI_HFEET = gmodels::ci(HFEET, confidence = 0.99, na.rm = TRUE)[[3]]) %>%
  gt()
```

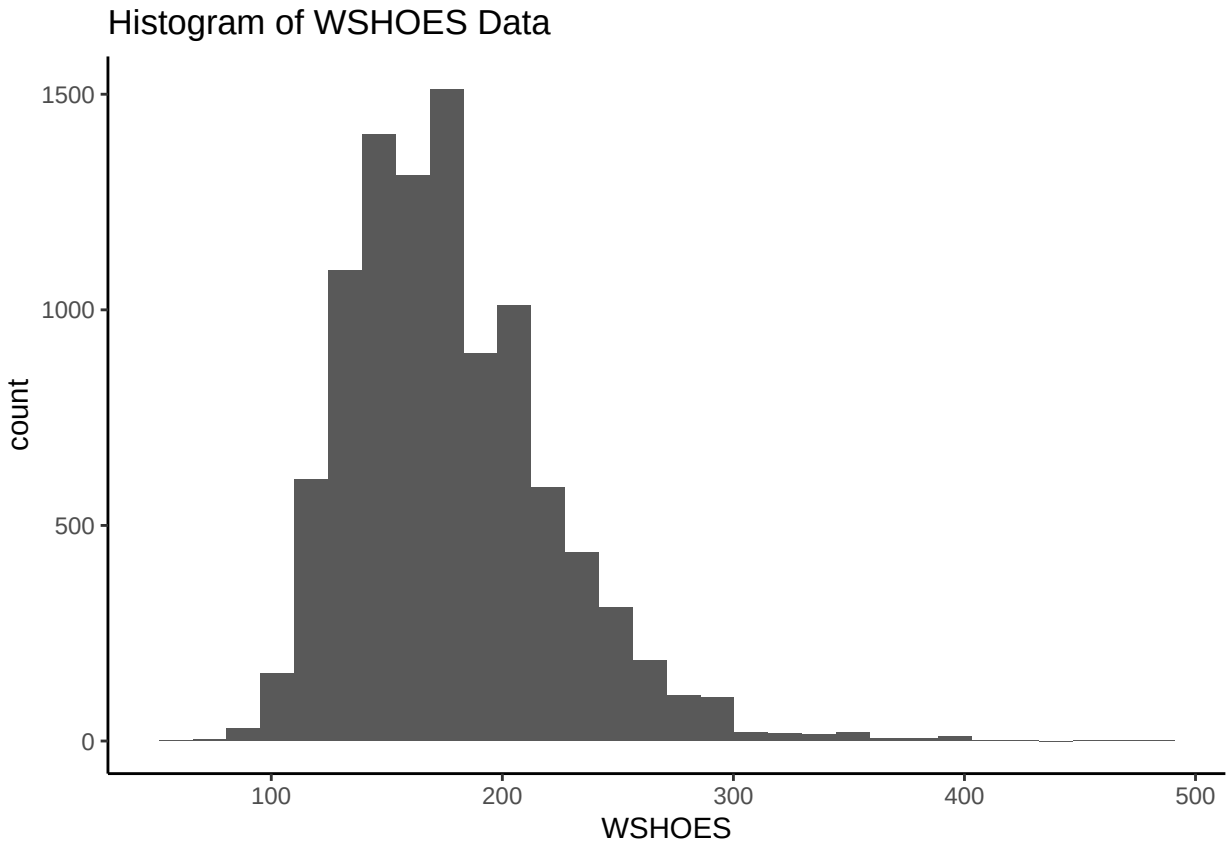
SEX01	lower95CI_WSHOES	upper95CI_WSHOES	lower99CI_WSHOES	upper99CI_WSHOES	lower95CI_HF
1	196.6610	199.2417	196.2553	199.6475	5.31
2	162.1389	164.2221	161.8114	164.5496	4.90

2.2.1 Answer:

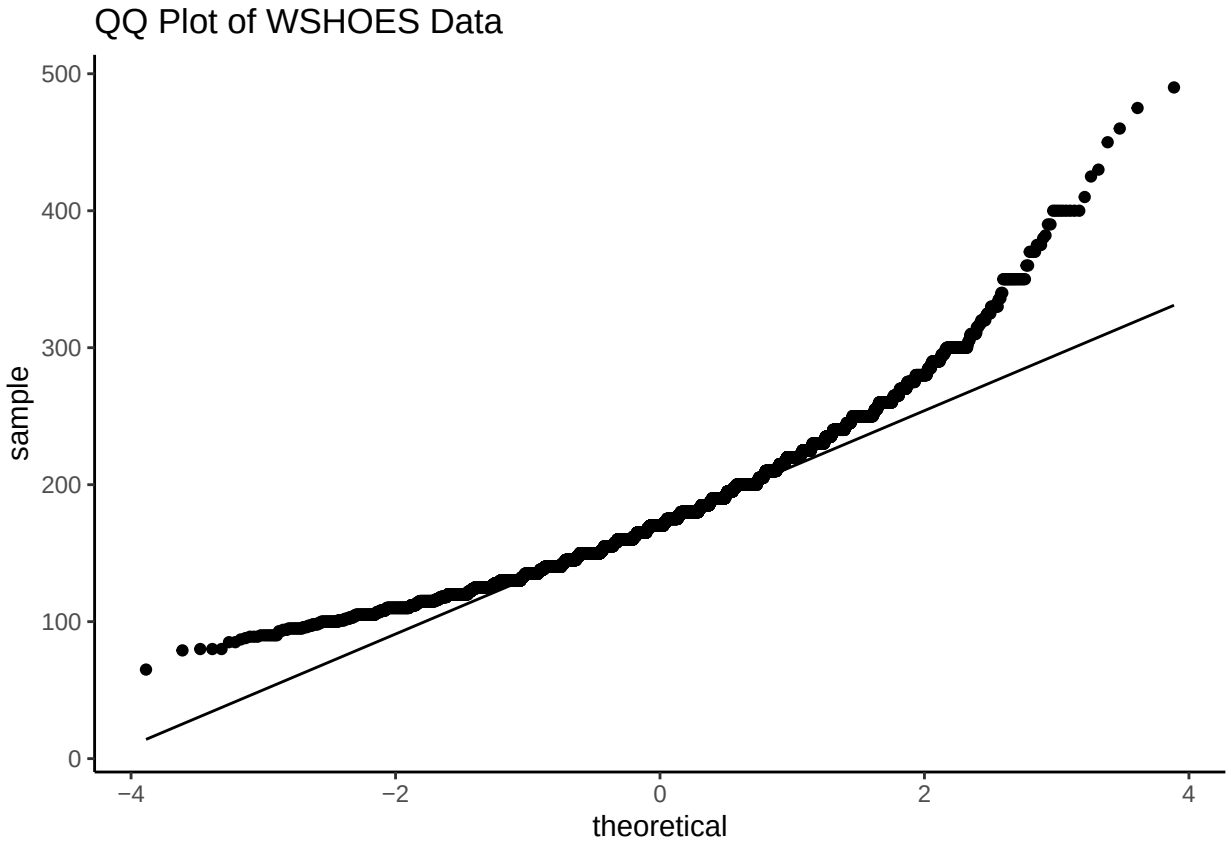
The difference between a 95% and 99% confidence interval is how wide the interval is and thus how likely you are to capture the data in the interval. A 95% interval will, on average, exclude ~5% of the data and is thus narrower than a 99% interval which will only exclude ~1% of the data.

2.3 7. Interpret the histogram, qqplot, ppplot, and boxplot on weight without shoes and gender. In addition a scatterplot on weight without shoes and BMI

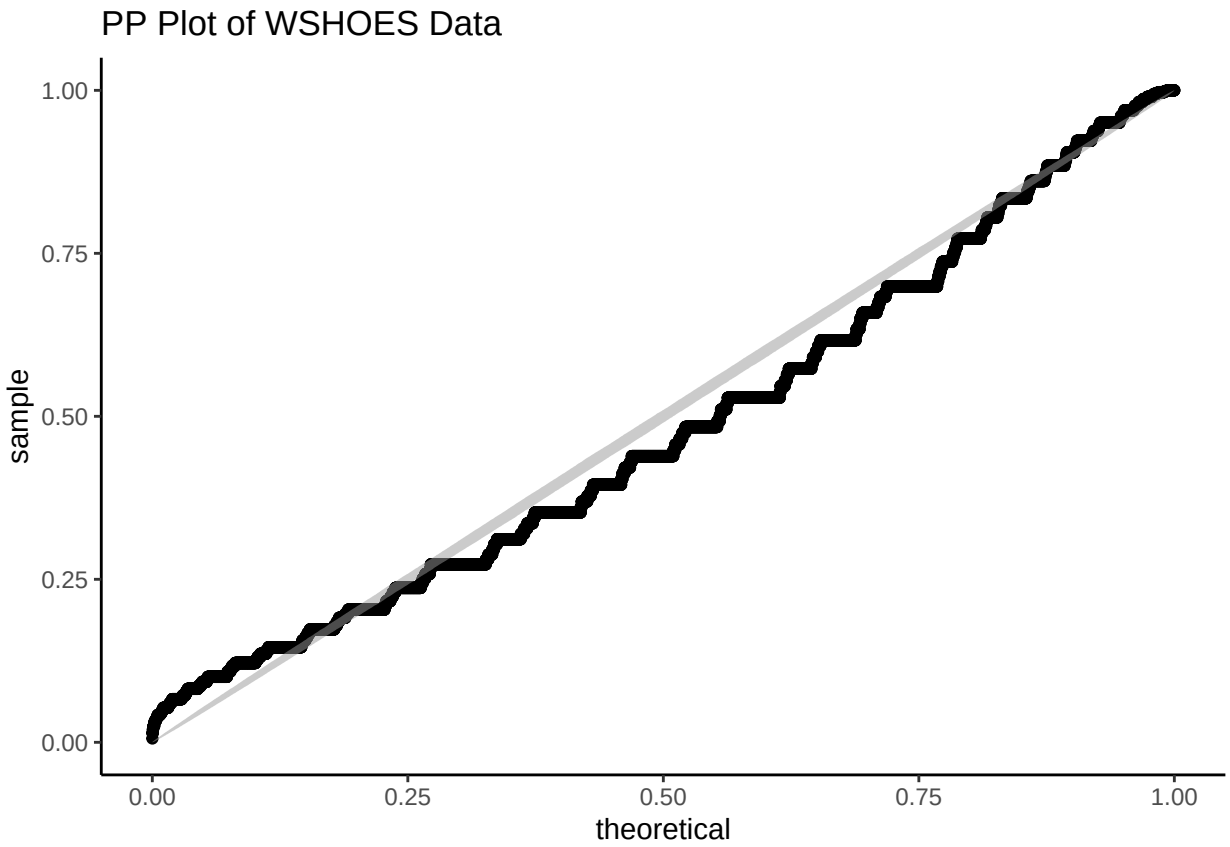
```
ggplot(hs15, aes(x = WSHOES)) +
  geom_histogram() +
  theme_classic() +
  labs(
    title = "Histogram of WSHOES Data"
  )
```



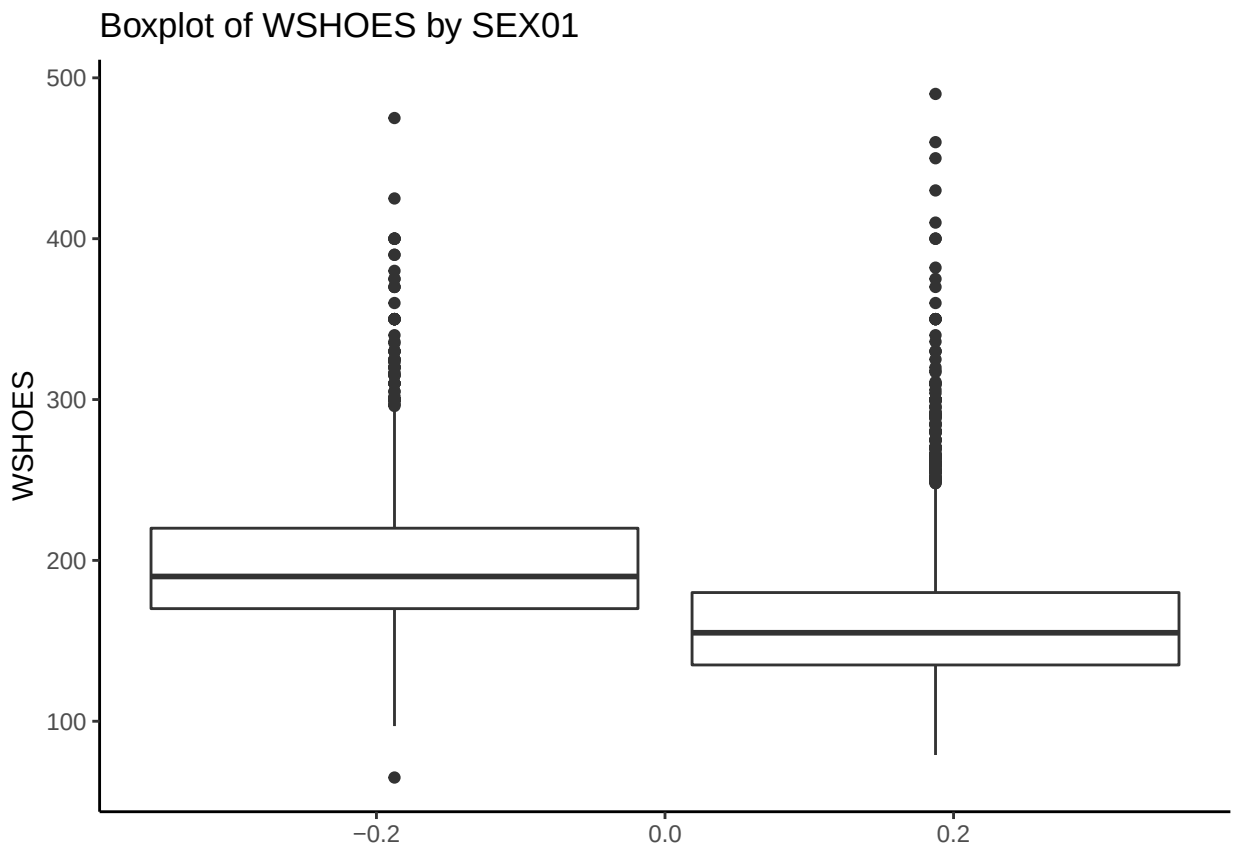
```r
ggplot(hs15, aes(sample = WSHOES)) +
  geom_qq() +
  geom_qq_line() +
  theme_classic() +
  labs(
    title = "QQ Plot of WSHOES Data"
  )
```



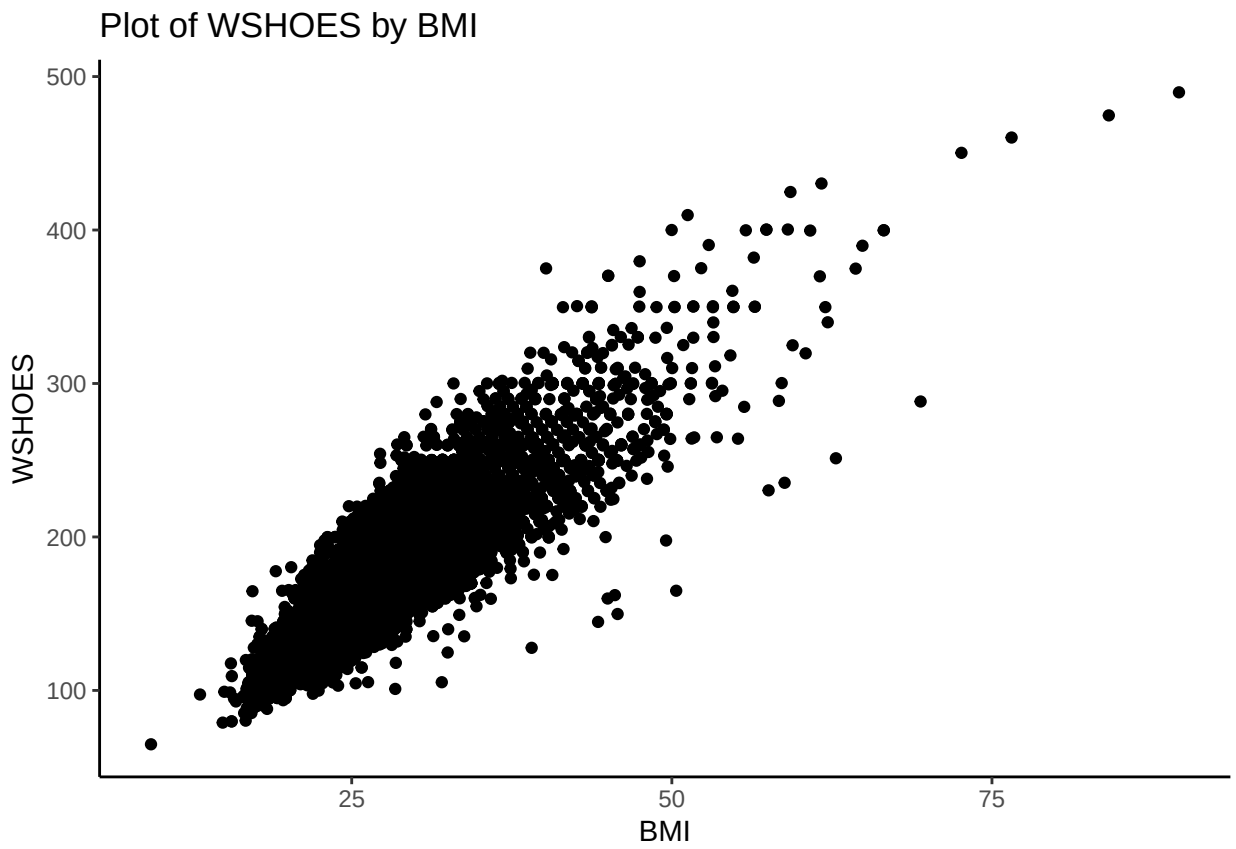

```r
ggplot(hs15, aes(sample = WSHOES)) +
  qqplotr::stat_pp_point() +
  qqplotr::stat_pp_band() +
  theme_classic() +
  labs(
    title = "PP Plot of WSHOES Data"
  )
```



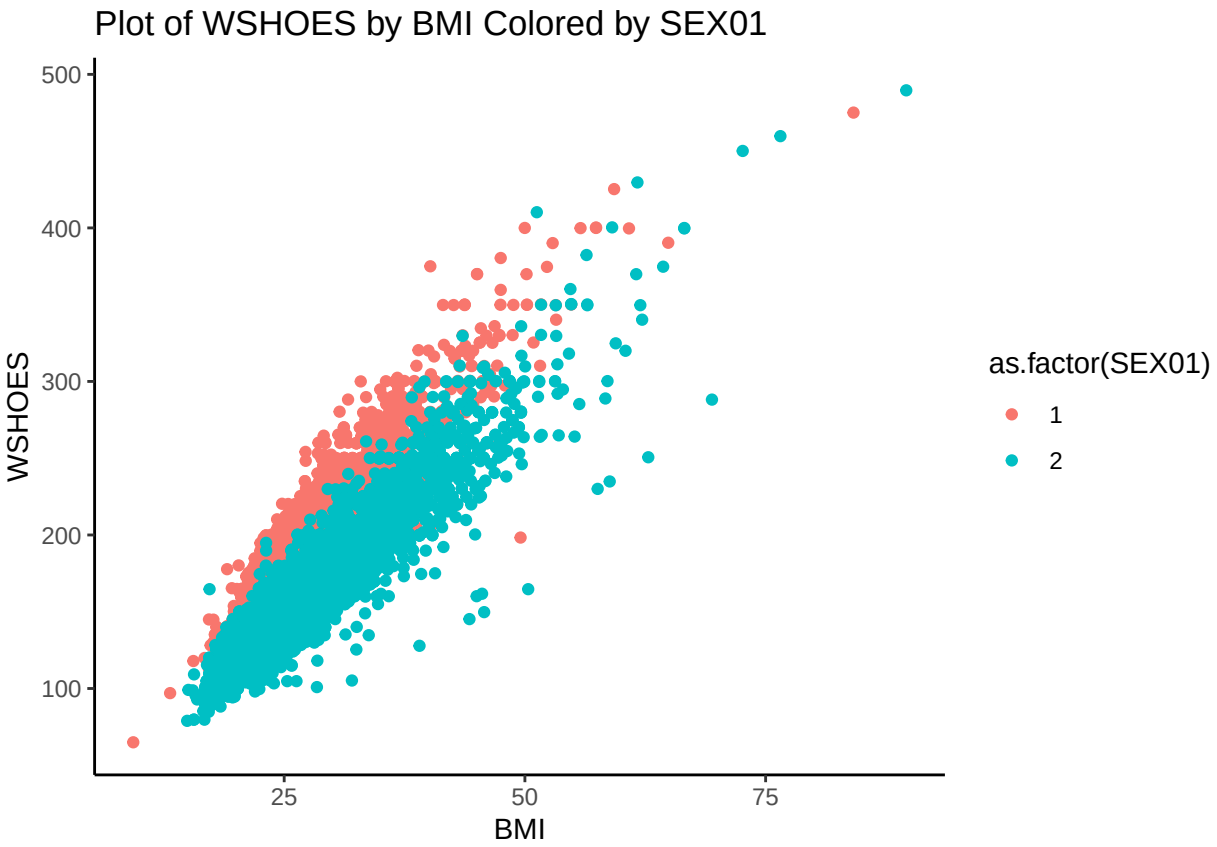

```
ggplot(hs15, aes(y = WSHOES, group = as.factor(SEX01))) +
  geom_boxplot() +
  theme_classic() +
  labs(
    title = "Boxplot of WSHOES by SEX01"
  )
```




```
ggplot(hs15, aes(x = BMI, y = WSHOES)) +
  geom_jitter() +
  theme_classic() +
  labs(
    title = "Plot of WSHOES by BMI"
  )
```




```
ggplot(hs15, aes(x = BMI, y = WSHOES, color = as.factor(SEX01))) +
  geom_jitter() +
  theme_classic() +
  labs(
    title = "Plot of WSHOES by BMI Colored by SEX01"
  )
```



2.3.1 Answer:

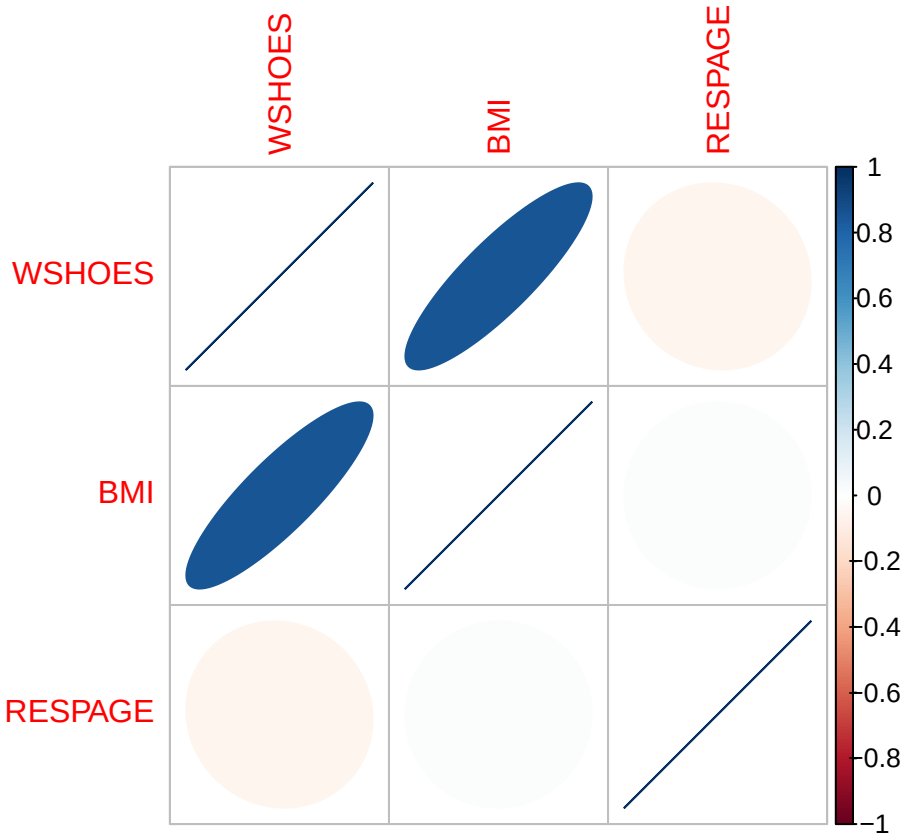
The histogram, qq, and pp plots show data that are approximately normally distrubted, with a slight tail. The boxplot shows the distrubtion and spread of the data about the same. The scatter plots show a data that seems to be correlated together both for the entire population and when grouped by sex.

2.4 8. Run a correlation on weight without shoes, BMI, and age

```
q8Data <- hs15 %>%
  select(WSHOES, BMI, RESPAGE) %>%
  drop_na()

cor(q8Data)
```

```
##              WSHOES        BMI     RESPAGE
## WSHOES   1.00000000 0.85844185 -0.05209302
## BMI      0.85844185 1.00000000  0.01085119
## RESPAGE -0.05209302 0.01085119  1.00000000
```

```
corrplot(cor(q8Data), method = 'ellipse')
```



2.4.1 Answer:

As can be inferred from the plot or taken directly from the matrix, it is clear there is a strong correlation with BMI and weight without shoes. There is a very week correlation with weight without shoes and age and no correlation with BMI and age.