Data_Man_Main

Steven Carrell

19/11/2019

Head and summary of the Nhanes 2015/2016 dataset

```
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
knitr::kable(head(nhanes.2015.2016, format="latex", booktabs=TRUE)) %>%
  kable_styling(latex_options="scale_down")
```

SEQN	ALQ101	ALQ110	ALQ130	SMQ020	RIAGENDR	RIDAGEYR	RIDRETH1	DMDCITZN	DMDEDUC2	DMDMARTL	DMDHHSIZ	INDFMPIR	BPXSY1	BPXDI1	BPXSY2	BPXDI2	BMXWT	BMXHT	BMXBMI	BMXLEG	BMXARML	BMXARMC	BMXWAIST	HIQ210
83732	1	NA	1	1	1	62	3	1	5	1	2	4.39	128	70	124	64	94.8	184.5	27.8	43.3	43.6	35.9	101.1	2
83733	1	NA	6	1	1	53	3	2	3	3	1	1.32	146	88	140	88	90.4	171.4	30.8	38.0	40.0	33.2	107.9	NA
83734	1	NA	NA	1	1	78	3	1	3	1	2	1.51	138	46	132	44	83.4	170.1	28.8	35.6	37.0	31.0	116.5	2
83735	2	1	1	2	2	56	3	1	5	6	1	5.00	132	72	134	68	109.8	160.9	42.4	38.5	37.7	38.3	110.1	2
83736	2	1	1	2	2	42	4	1	4	3	5	1.23	100	70	114	54	55.2	164.9	20.3	37.4	36.0	27.2	80.4	2
83737	2	2	NA	2	2	72	1	2	2	4	5	2.82	116	58	122	58	64.4	150.9	28.6	34.4	33.5	31.4	92.9	NA

```
summary(nhanes2)
```

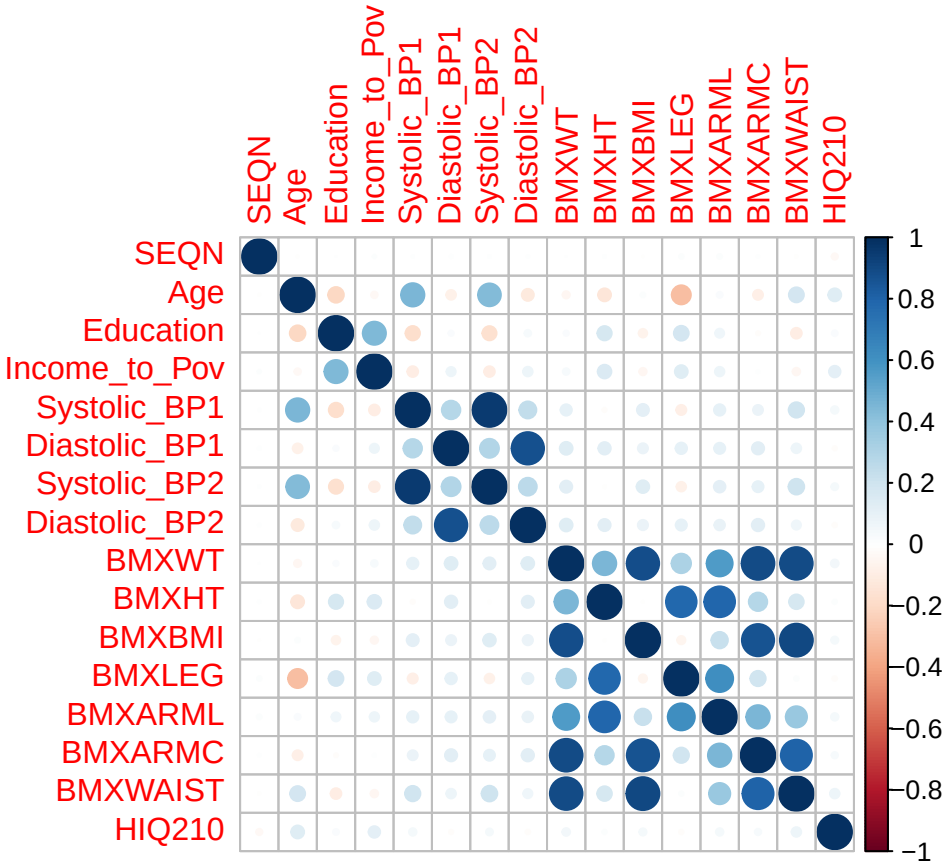
```
##       SEQN           Alcohol_Year       Smoked_100      Gender
##  Min.   :83732   Don't know:   3   Don't know:   8   Female:2976
##  1st Qu.:86164   No        :1728   No        :3406   Male  :2759
##  Median :88668   Yes       :3477   Refused   :   2
##  Mean   :88679   NA's      : 527   Yes       :2319
##  3rd Qu.:91178
##  Max.   :93702
##
##       Age                       Race        Education
##  Min.   :18.00   Black           :1227   Min.   :1.000
##  1st Qu.:32.00   Mexican American:1018   1st Qu.:3.000
##  Median :48.00   Other Hispanic  : 750   Median :4.000
##  Mean   :48.05   Other Race      : 901   Mean   :3.442
##  3rd Qu.:63.00   White           :1839   3rd Qu.:4.750
##  Max.   :80.00                           Max.   :9.000
##                                          NA's   :261
##              Marital_Status Household_Size Income_to_Pov    Systolic_BP1
##  Married            :2780   1: 770         Min.   :0.000   Min.   : 82.0
##  Never Married      :1004   2:1546         1st Qu.:1.060   1st Qu.:112.0
##  Divorced           : 579   3:1037         Median :1.980   Median :122.0
##  Living with partner: 527   4: 936         Mean   :2.403   Mean   :125.1
##  Widowed            : 396   5: 699         3rd Qu.:3.740   3rd Qu.:134.0
##  (Other)            : 188   6: 379         Max.   :5.000   Max.   :236.0
##  NA's               : 261   7: 368         NA's   :601     NA's   :334
```

```
##  Diastolic_BP1     Systolic_BP2   Diastolic_BP2        BMXWT
##  Min.   :  0.00   Min.   : 84.0   Min.   :  0.00   Min.   : 32.40
##  1st Qu.: 62.00   1st Qu.:112.0   1st Qu.: 62.00   1st Qu.: 65.90
##  Median : 70.00   Median :122.0   Median : 70.00   Median : 78.20
##  Mean   : 69.52   Mean   :124.8   Mean   : 69.35   Mean   : 81.34
##  3rd Qu.: 78.00   3rd Qu.:134.0   3rd Qu.: 78.00   3rd Qu.: 92.70
##  Max.   :120.00   Max.   :238.0   Max.   :144.00   Max.   :198.90
##  NA's   :334      NA's   :200     NA's   :200      NA's   :69
##      BMXHT           BMXBMI          BMXLEG         BMXARML
##  Min.   :129.7   Min.   :14.50   Min.   :26.00   Min.   :28.20
##  1st Qu.:158.7   1st Qu.:24.30   1st Qu.:36.00   1st Qu.:35.20
##  Median :166.0   Median :28.30   Median :38.60   Median :37.10
##  Mean   :166.1   Mean   :29.38   Mean   :38.58   Mean   :37.15
##  3rd Qu.:173.5   3rd Qu.:33.00   3rd Qu.:41.20   3rd Qu.:39.00
##  Max.   :202.7   Max.   :67.30   Max.   :51.50   Max.   :47.40
##  NA's   :62      NA's   :73      NA's   :390     NA's   :308
##     BMXARMC         BMXWAIST          HIQ210
##  Min.   :17.10   Min.   : 58.70   Min.   :1.000
##  1st Qu.:29.50   1st Qu.: 87.60   1st Qu.:2.000
##  Median :32.70   Median : 98.30   Median :2.000
##  Mean   :33.11   Mean   : 99.57   Mean   :1.915
##  3rd Qu.:36.20   3rd Qu.:109.30   3rd Qu.:2.000
##  Max.   :58.40   Max.   :171.60   Max.   :9.000
##  NA's   :308     NA's   :367      NA's   :1003
##                            EducationX     agegroup
##  < 9th grade                     : 655   18-29 :1192
##  9-11th grade                    : 643   30-39 : 933
##  High school graduate            :1186   40-49 : 913
##  Some college/Uni                :1621   50-59 : 888
##  College/Uni graduate or above:1366   60-69 : 917
##  9                               :   3   70-80+: 892
##  NA's                            : 261
```

3 functions created (saved in lib file).

1. 'corr_func' to quickly test correlations on the variables whilst exploring the data). The sole purpose using this rather than the built-in function is so the argument 'complete.obs' doesn't have to be typed each time.
2. 'grp_func' to quickly allow the user to input a variable to group by and a variable to analyse which will output the min, max, and mean.
3. 'prop_func' takes the following aruguments: dataframe, catagorical group variable, and binary variable to analyse. Then outputs a new dataframe of proportions based on these arguments (split by gender). This will be used in the shiny dashboard.

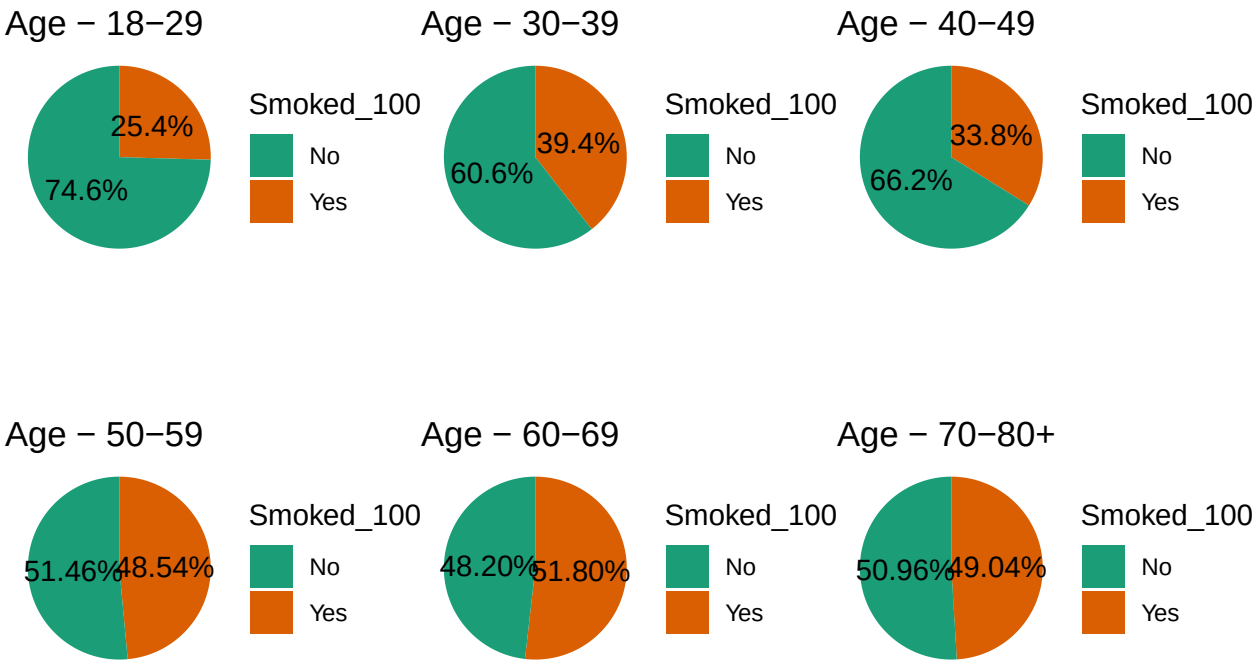
Below code to firstly find out all the columns containing numeric data (numeric_vars. Then a new data table is created with these values (numeric_data). From this table, all the correlations were worked out wit the 'cor' function and saved under a variable names 'correlations'. A correlation plot was created to clearly show any positive or negative correlations between these variables.



Group by agegrp to look at proportions of people who have and have not smoked 100 cigarettes in their. Assume that the the older generations will have a larger proportion of smokers than the youngers ones. Displayed in pie charts

```
smoked = prop_func(nhanes2, 'agegroup', 'Smoked_100', percentage = FALSE)

p1 = smoked[1:2,]
p2 = smoked[3:4,]
p3 = smoked[5:6,]
p4 = smoked[7:8,]
p5 = smoked[9:10,]
p6 = smoked[11:12,]

pie1 = ggplot(p1, aes(x='', y=Total_Proportion, fill =Smoked_100 )) +
  geom_bar(stat='identity', width=1) + labs(title = 'Age - 18-29') +
  coord_polar('y', start=0) +
  geom_text(aes(label = percent(Total_Proportion)), position = position_stack(vjust = 0.5), size = 4) +
  scale_fill_brewer(palette = 'Dark2') +
  theme_void()

pie2 = ggplot(p2, aes(x='', y=Total_Proportion, fill =Smoked_100 )) +
  geom_bar(stat='identity', width=1) + labs(title = 'Age - 30-39') +
  coord_polar('y', start=0) +
  geom_text(aes(label = percent(Total_Proportion)), position = position_stack(vjust = 0.5), size = 4) +
  scale_fill_brewer(palette = 'Dark2') +
  theme_void()
```

```
pie3 = ggplot(p3, aes(x='', y=Total_Proportion, fill =Smoked_100)) +
  geom_bar(stat='identity', width=1) + labs(title = 'Age - 40-49') +
  coord_polar('y', start=0) +
  geom_text(aes(label = percent(Total_Proportion)), position = position_stack(vjust = 0.5), size = 4) +
  scale_fill_brewer(palette = 'Dark2') +
  theme_void()

pie4 = ggplot(p4, aes(x='', y=Total_Proportion, fill =Smoked_100)) +
  geom_bar(stat='identity', width=1) + labs(title = 'Age - 50-59') +
  coord_polar('y', start=0) +
  geom_text(aes(label = percent(Total_Proportion)), position = position_stack(vjust = 0.5), size = 4) +
  scale_fill_brewer(palette = 'Dark2') +
  theme_void()

pie5 = ggplot(p5, aes(x='', y=Total_Proportion, fill =Smoked_100)) +
  geom_bar(stat='identity', width=1) + labs(title = 'Age - 60-69') +
  coord_polar('y', start=0) +
  geom_text(aes(label = percent(Total_Proportion)), position = position_stack(vjust = 0.5), size = 4) +
  scale_fill_brewer(palette = 'Dark2') +
  theme_void()

pie6 = ggplot(p6, aes(x='', y=Total_Proportion, fill =Smoked_100)) +
  geom_bar(stat='identity', width=1) + labs(title = 'Age - 70-80+') +
  coord_polar('y', start=0) +
  geom_text(aes(label = percent(Total_Proportion)), position = position_stack(vjust = 0.5), size = 4) +
  scale_fill_brewer(palette = 'Dark2') +
  theme_void()


grid.arrange(
  pie1, pie2, pie3, pie4, pie5, pie6, nrow=2, ncol = 3, top = 'Proportion of people who have smoked 100
)
```



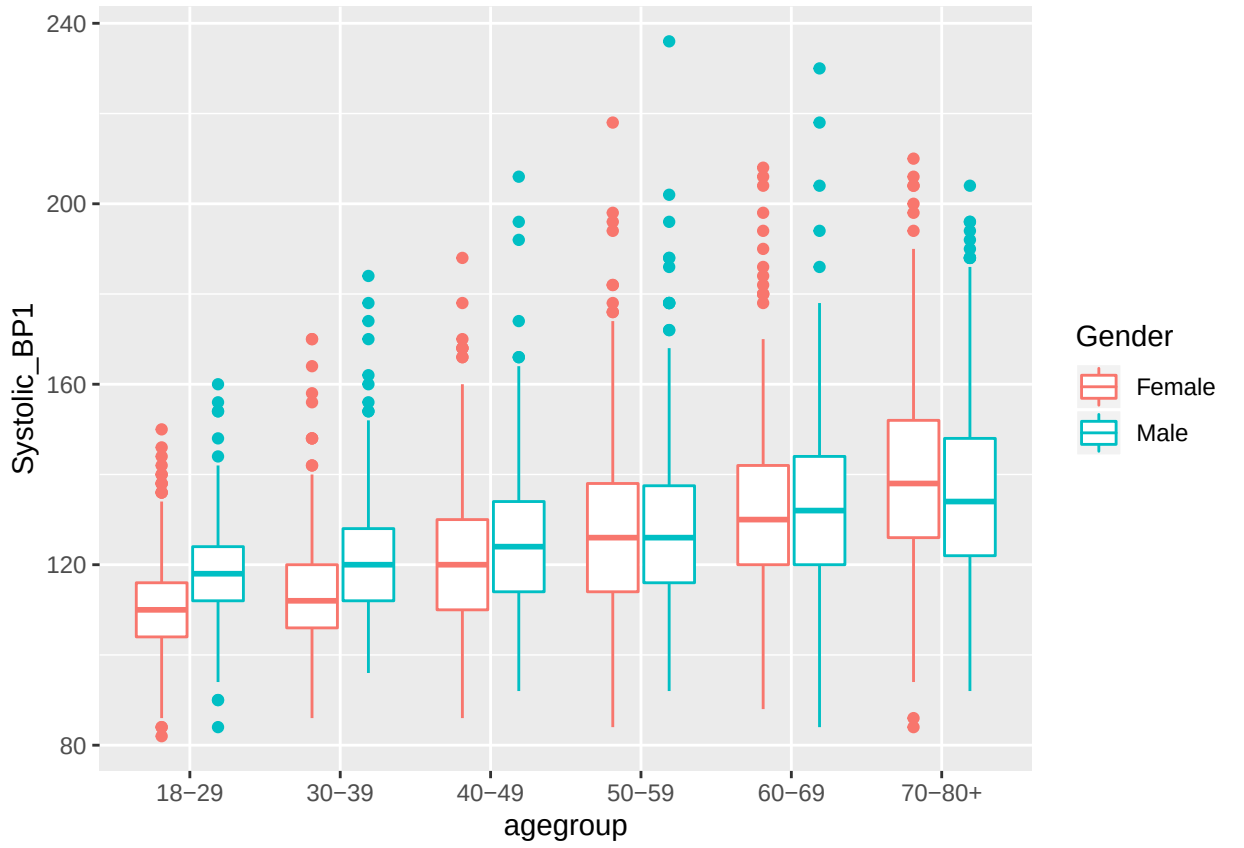
Next test to see if someone with a higher education level is less likely to have smoked 100 cigarettes in their life. From the results we can see that as a total proportion between male and female there is not much difference. However if we look at only the males, we see that a college/uni graduate is less likely to have smoked 100 than someone with a lower education,

```
ed_smoke = prop_func(nhanes2, "EducationX", "Smoked_100")
knitr::kable(ed_smoke)
```

EducationX	Smoked_100	Male_Proportion	Female_Proportion	Total_Proportion
< 9th grade	No	46.71%	79.1%	64.0%
< 9th grade	Yes	53.29%	20.9%	36.0%
9-11th grade	No	33.0%	61.2%	45.79%
9-11th grade	Yes	67.0%	38.8%	54.21%
High school graduate	No	38.6%	62.1%	50.169%
High school graduate	Yes	61.4%	37.9%	49.831%
Some college/Uni	No	44.4%	62.7%	54.66%
Some college/Uni	Yes	55.6%	37.3%	45.34%
College/Uni graduate or above	No	64.4%	78.8%	71.9%
College/Uni graduate or above	Yes	35.6%	21.2%	28.1%

Below some box plots and scatter plots have been made influenced by the correlation plot. We hope to find some interetsing results from this. cor_func and grp_func was also used for futher testing. Based on the results some furhter hypothesis testing will be carried out.

```
b1 = ggplot(nhanes2, aes(x=agegroup, y=Systolic_BP1, color=Gender)) +
  geom_boxplot()
b1
```

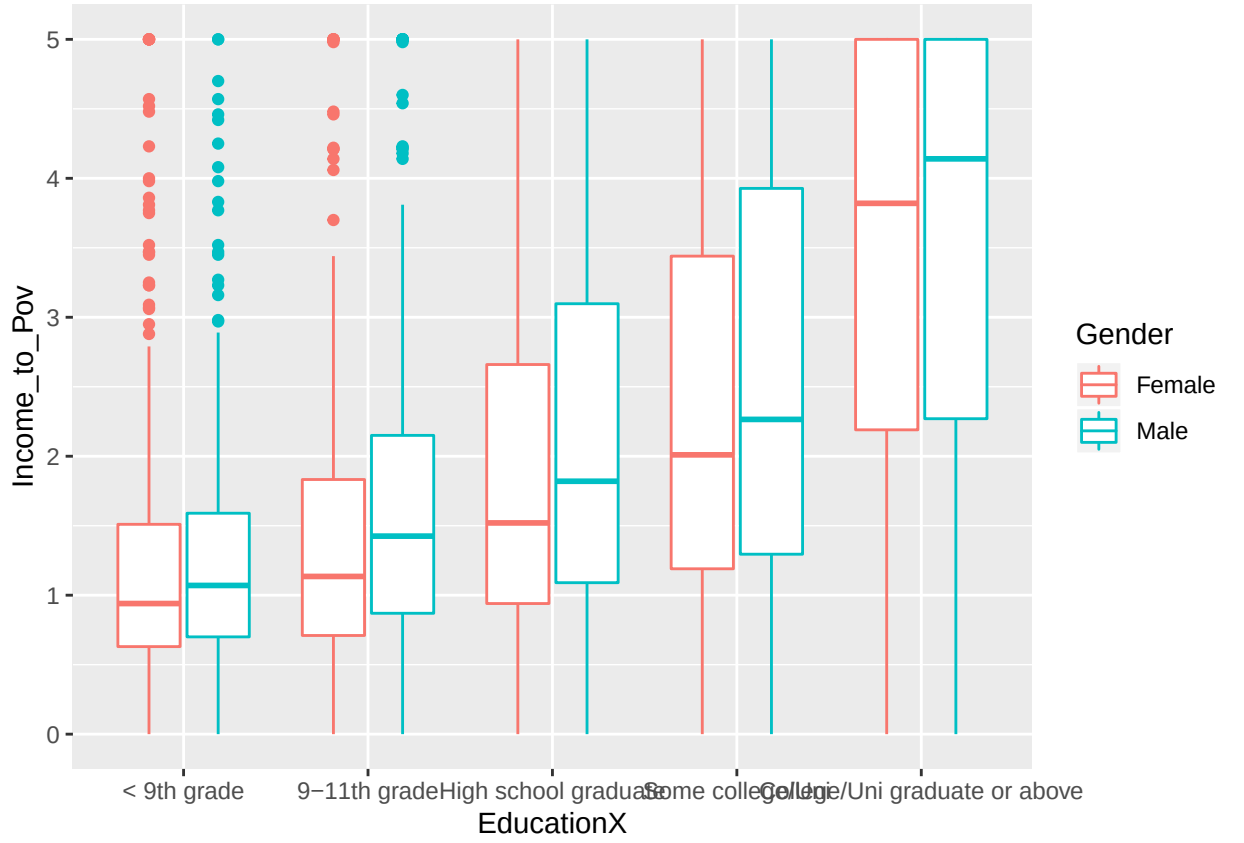



```
b1_test = filter(nhanes2, Age>69)
b1_func = grp_func(b1_test, 'Gender', 'Systolic_BP1')
knitr::kable(b1_func)
```

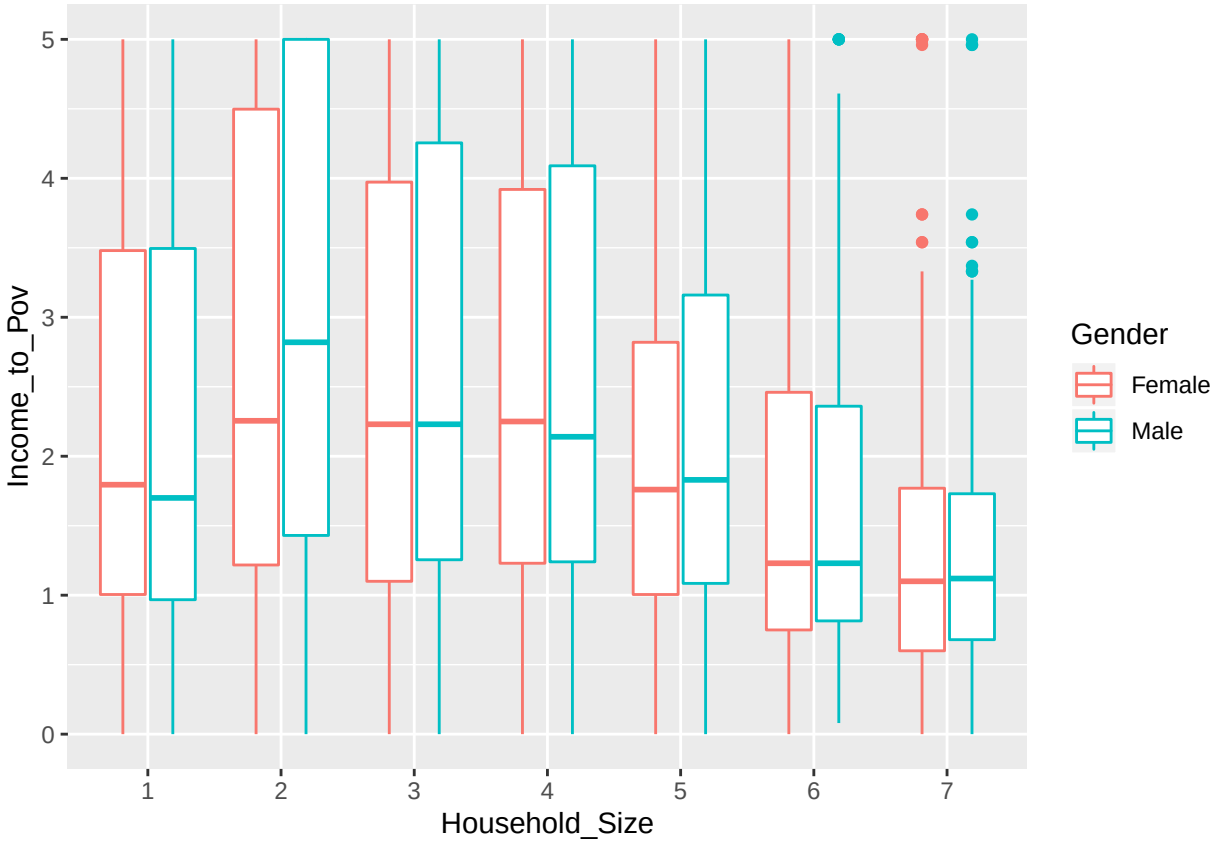
Gender	Mean	Max	Min
Female	139.9763	210	84
Male	136.5012	204	92

Based on the above information 2 hypothesis tests will be carried out. 1 to see if men of all ages have higher Systolic BP than females, and another to test if females aged 70+ have a higher Systolic BP than men. (teststing at bottom of repoort)

```
#Filter dataframe to remove below value
df2 = filter(nhanes2, EducationX !='9')
b2 = ggplot(df2, aes(x=EducationX, y=Income_to_Pov, color=Gender)) +
  geom_boxplot()
b2
```

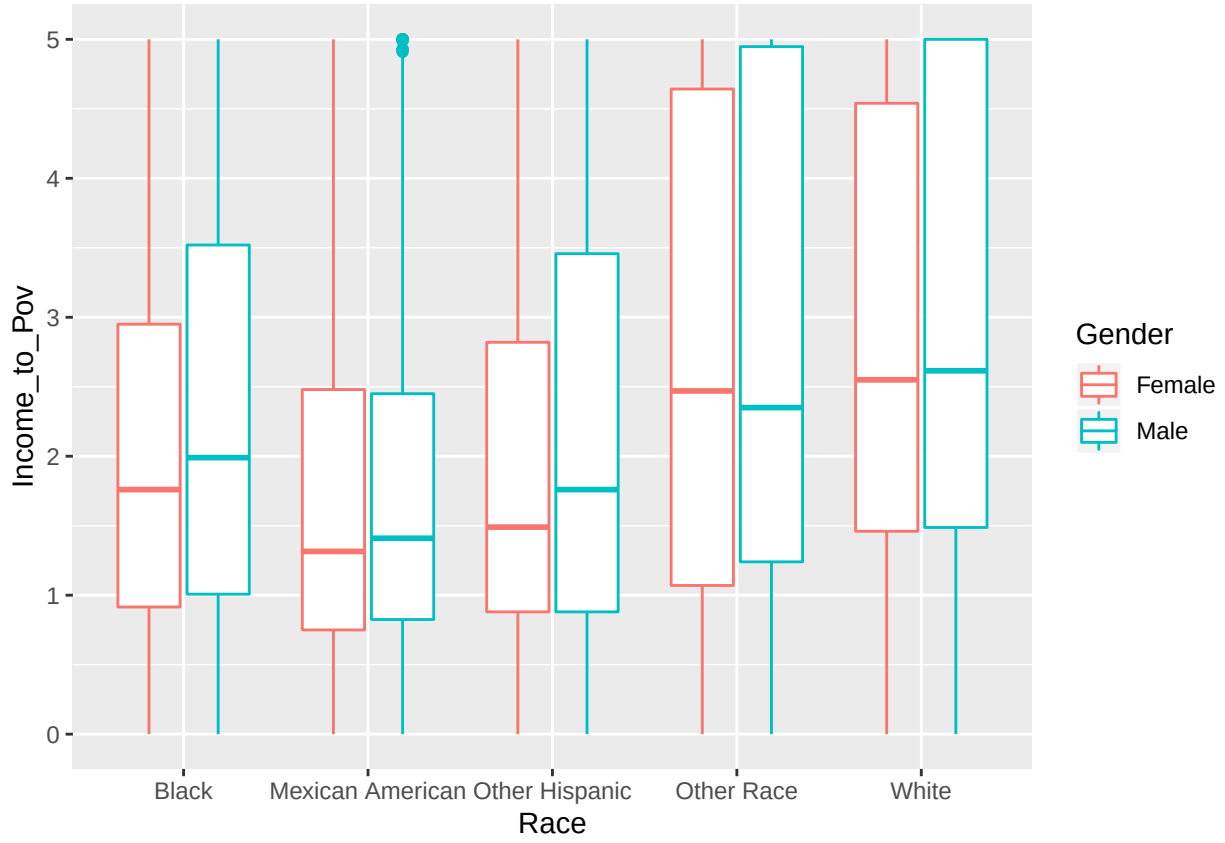



```
nhanes2$Household_Size = as.factor(nhanes2$Household_Size)
b3 = ggplot(nhanes2, aes(x=Household_Size, y=Income_to_Pov, color=Gender)) +
  geom_boxplot()
b3
```

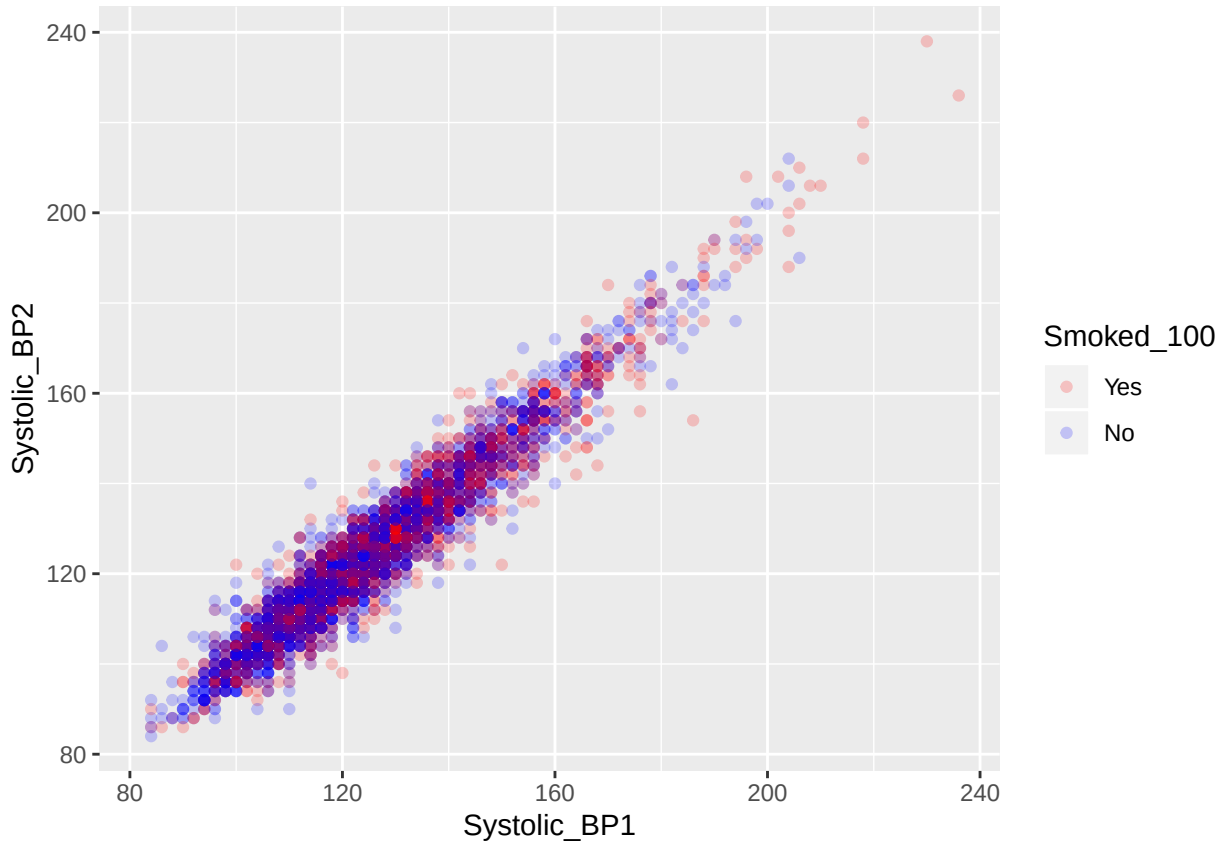


Based on the above plot, a hypothesis test will be carried out to see if females living alone earn more than men living on their own. (test below)

```
b4 = ggplot(nhanes2, aes(x=Race, y=Income_to_Pov, color=Gender)) +
  geom_boxplot()
b4
```



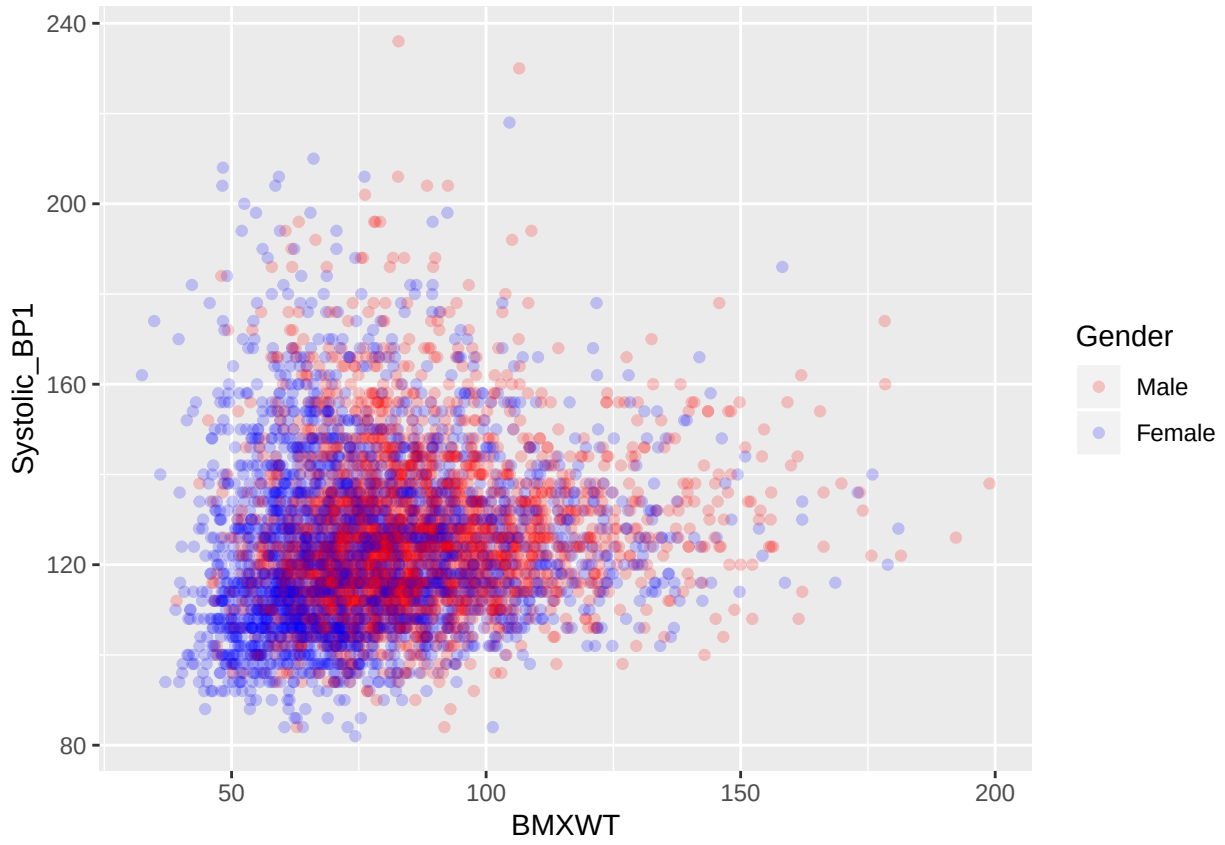

```
#Scatter plots
#Filter dataframe to remove below values
df1 = filter(nhanes2, Smoked_100 !="Don't know" & Smoked_100 !='Refused')

s1 = ggplot(df1, aes(x=Systolic_BP1, y=Systolic_BP2, color = Smoked_100)) +
  geom_point(alpha=0.2) +
  scale_color_manual(breaks = c('Yes', 'No'), values = c('blue', 'red'))
s1
```




```
corr_func('Systolic_BP1', 'Systolic_BP2')
```

```
## [1] 0.9622873
```

```
s2 = ggplot(nhanes2, aes(x=BMXWT, y=Systolic_BP1, color=Gender)) +
  geom_point(alpha=0.2) +
  scale_color_manual(breaks = c('Male', 'Female'), values = c('blue', 'red'))
s2
```



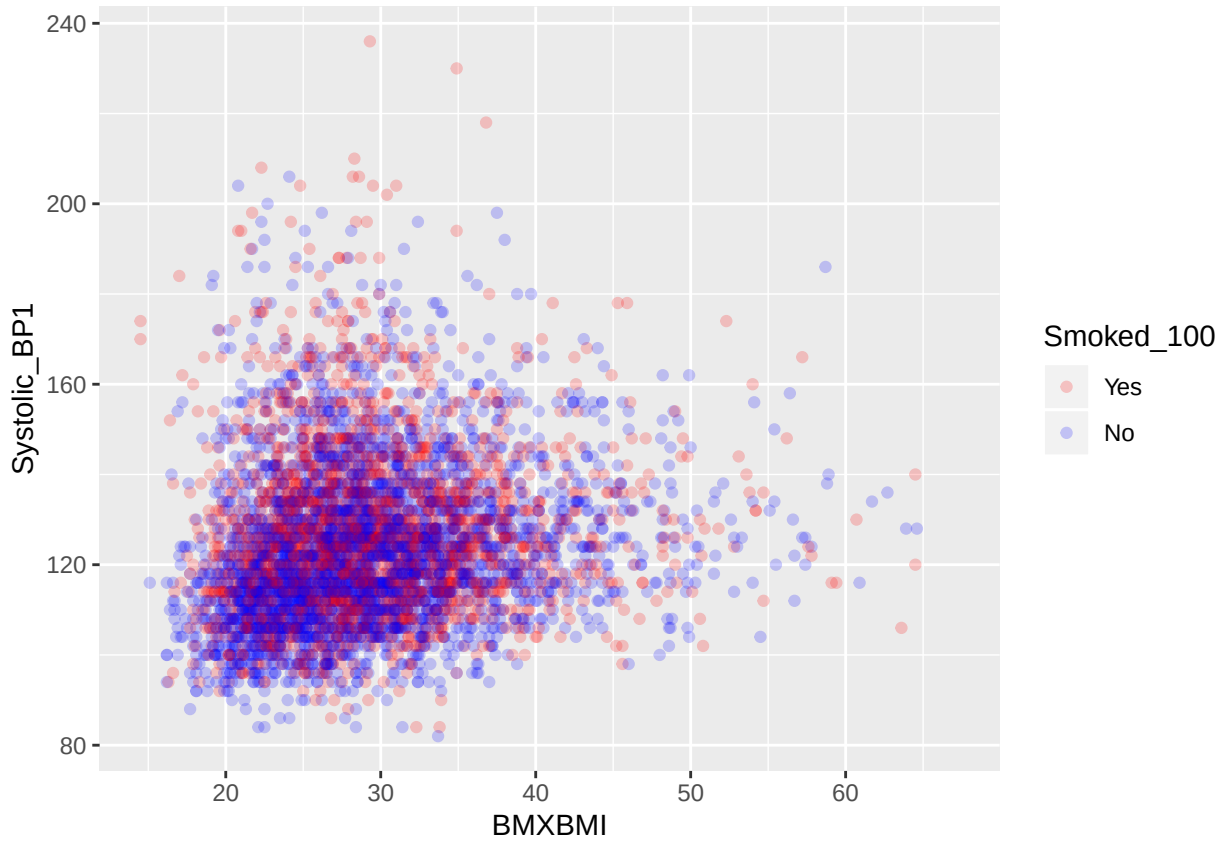

```
corr_func('BMXWT', 'Systolic_BP1')
```

```
## [1] 0.1225117
```

```
grp_func(nhanes2, 'Gender', 'BMXBMI')
```

```
## # A tibble: 2 x 4
##   Gender  Mean   Max   Min
##   <fct>  <dbl> <dbl> <dbl>
## 1 Female  29.9  67.3  14.5
## 2 Male    28.8  58.8  15.1
```

```
s3 = ggplot(df1, aes(x=BMXBMI, y=Systolic_BP1, color = Smoked_100)) +
  geom_point(alpha=0.2) +
  scale_color_manual(breaks = c('Yes', 'No'), values = c('blue', 'red'))
s3
```




```
corr_func('Systolic_BP1', 'BMXBMI')
```

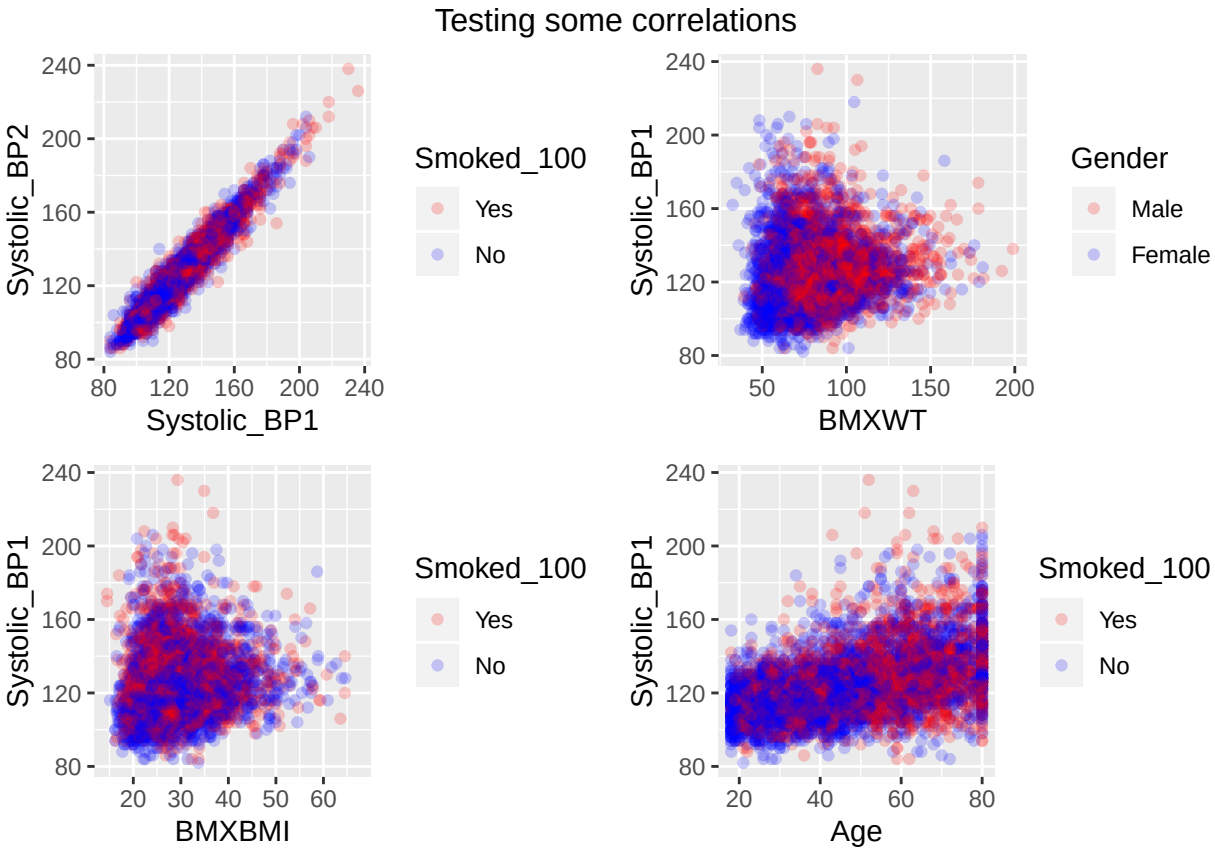
```
## [1] 0.1352012
```

```
s4 = ggplot(df1, aes(x=Age, y=Systolic_BP1, color = Smoked_100)) +
  geom_point(alpha=0.2) +
  scale_color_manual(breaks = c('Yes', 'No'), values = c('blue', 'red'))
s4
```


```
corr_func('Age', 'Systolic_BP1')
```

```
## [1] 0.4692335
```

```
grid.arrange(
  s1, s2, s3, s4, nrow=2, ncol = 2, top = 'Testing some correlations'
)
```



Shiny dashboard with interactive scatter plot and box plot. Interactive table using the prop_func

Hypothesis testing - all below tests based on significance level of 0.05

T-test to see if male who has smoked 100 cigarettes in his live and had at least 12 alcoholic drinks in 1 year has a higher Systolic blood pressure than a male who has not. H0 != H1

```
test1_1 = filter(nhanes2, Gender=='Male' & Smoked_100=='Yes' & Alcohol_Year=='Yes')
test1_2 = filter(nhanes2, Gender=='Male' & Smoked_100=='No' & Alcohol_Year=='No')

#Test using the t.test function
t.test(test1_1$Systolic_BP1, test1_2$Systolic_BP1, alternative = "two.sided",
       var.equal = FALSE, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  test1_1$Systolic_BP1 and test1_2$Systolic_BP1
## t = 4.6317, df = 692.33, p-value = 4.333e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.701071 6.676093
## sample estimates:
## mean of x mean of y
##  128.8670  124.1784
```

```r
#Test manually. As the std_dev on each sample are not similar, I have used the
#unpooled approach
t1_n1 = nrow(test1_1)
t1_n2 = nrow(test1_2)
t1_u1 = mean(test1_1$Systolic_BP1, na.rm = TRUE)
t1_u2 = mean(test1_2$Systolic_BP1, na.rm = TRUE)
t1_sig1 = sd(test1_1$Systolic_BP1, na.rm = TRUE)
t1_sig2 = sd(test1_2$Systolic_BP1, na.rm = TRUE)

t_test = (t1_u1-t1_u2)/sqrt(((t1_sig1^2)/t1_n1)+((t1_sig2^2)/t1_n2))
t_test
```

```
## [1] 4.730432
```

```r
#t-score for inbuilt_funtion = 4.6317, t-score worked out manually is 4.73. Both give a
#p-value of <0.00001 which provides us with sufficient evidence to reject the null hypothesis and say t

#T-test to see if somoen who has at least 12 alcoholic drinks in a year has a higher Systolic
#BP than someone who has not.

test2_1 = filter(nhanes2, Gender=='Male' & Alcohol_Year=='Yes')
test2_2 = filter(nhanes2, Gender=='Male' & Alcohol_Year=='No')

#As the variance is similar (testing below), we will use thr pooled approach
sd(test2_1$Systolic_BP1, na.rm = TRUE)
```

```
## [1] 17.61412
```

```r
sd(test2_2$Systolic_BP1, na.rm = TRUE)
```

```
## [1] 17.8955
```

```r
t.test(test2_1$Systolic_BP1, test2_2$Systolic_BP1, alternative = 'two.sided', var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  test2_1$Systolic_BP1 and test2_2$Systolic_BP1
## t = 2.275, df = 2446, p-value = 0.02299
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2770041 3.7363882
## sample estimates:
## mean of x mean of y
##  127.4336  125.4269
```

```r
#P-value of 0.02299 still provides us with enough evidence to reject the null and conclude that 'on ave

#Hypothesis test 3 - Do men of all ages have higher Systolic BP than women.
test3_1 = filter(nhanes2, Gender == 'Male')
```

```r
test3_2 = filter(nhanes2, Gender == 'Female')

#Testing variancwe below. As the variance differs, we will use the unpooled approach
sd(test3_1$Systolic_BP1, na.rm = TRUE)
```

```
## [1] 17.64247
```

```r
sd(test3_2$Systolic_BP1, na.rm = TRUE)
```

```
## [1] 19.06579
```

```r
t.test(test3_1$Systolic_BP1, test3_2$Systolic_BP1, alternative = 'two.sided', var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  test3_1$Systolic_BP1 and test3_2$Systolic_BP1
## t = 7.4453, df = 5397.1, p-value = 1.12e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.739755 4.698244
## sample estimates:
## mean of x mean of y
##  126.9989  123.2799
```

```r
#Based on this low p-value we reject the null hypothesis.

#Hypothesis test to see if women >69 have a higher Systolic BP than men of the similar age
test4_1 = filter(nhanes2, Gender == 'Female' & Age > 69)
test4_2 = filter(nhanes2, Gender == 'Male' & Age > 69)

#Testing variancwe below. As the variance is similar, the pooled approach will be used
sd(test4_1$Systolic_BP1, na.rm = TRUE)
```

```
## [1] 20.7647
```

```r
sd(test4_2$Systolic_BP1, na.rm = TRUE)
```

```
## [1] 20.46368
```

```r
t.test(test4_1$Systolic_BP1, test4_2$Systolic_BP1, alternative = 'two.sided', var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  test4_1$Systolic_BP1 and test4_2$Systolic_BP1
## t = 2.4618, df = 851, p-value = 0.01402
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  0.7044388 6.2458476
## sample estimates:
## mean of x mean of y
##  139.9763  136.5012
```

```
#Hypothesis test to see if females living alone earn more than males living alone.
test5_1 = filter(nhanes2, Gender == 'Female' & Household_Size == '1')
test5_2 = filter(nhanes2, Gender == 'Male' & Household_Size == '1')

#Signifincant difference in variance, so we will used the unpooled approach
sd(test5_1$Income_to_Pov, na.rm = TRUE)
```

```
## [1] 1.550615
```

```
sd(test5_2$Income_to_Pov, na.rm = TRUE)
```

```
## [1] 1.598473
```

```
t.test(test5_1$Income_to_Pov, test5_2$Income_to_Pov, alternative = 'two.sided', var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  test5_1$Income_to_Pov and test5_2$Income_to_Pov
## t = 0.11781, df = 685.79, p-value = 0.9063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2201615  0.2482691
## sample estimates:
## mean of x mean of y
##  2.279837  2.265783
```

```
#With a p-value of 0.9063, we fail to reject the null hypothesis based on a significance
#level of 0.05. Confidence level also includes 0.
```