# 152 PROJECT

*Drinking pattern and effect on cardio health*

Siyue Su, Shuya Zhan, Zoey Ruan

*When I read about the evils of drinking,*
*I gave up reading.”*

*– Henny Youngman*

## Introduction

We've been getting a lot of mixed messages about alcohol. Some people claim that moderate amounts of alcohol may bring happiness and benefit. Whereas, some warn that alcohol is addictive and could be harmful, or even affects your body system. Before you "give up" this reading, we promise you that we are not talking about the evils of drinking, instead, we want to examine the alcohol usage pattern in the U.S.A. Is the underage drinking among adults-to-be a widespread phenomenon? Is there any distinct difference in drinking pattern and habit among different race groups and age groups? In addition, is the drinking pattern has association with the gender? Several reports cite that the elderly tend to drink alcohol much more frequently than other age groups, while consuming a much smaller amount each time compared with the others. How much credence could we give to this report based on our analysis on NHANES? After the examination of the overall pattern, we further our analysis in relationship of alcohol drinking with physical health. The truth is that the health effects of alcohol are actually quite complex, so how does alcohol link to our health indeed?

## Questions we are addressing

We are interested in exploring the overall alcohol drinking pattern in the U.S.A, and its relationship with people's physical health. Some of the questions we want to address in the following analysis include:

1) Will different age groups (divided into 5 groups: underage(18-21), millennials(22-36), generation x(37-48), baby boomers(49-67) and the greatest generation(68+)) have different alcohol drinking habits?

2) Will different race groups tend to have different alcohol drinking patterns? How does it vary between different race?

3) Compared with female, will male possess a more frequent alcohol consumption and how it that different when we break down into different age groups?

4) Is there a certain relationship between the frequency of drinking alcohol with the total amount of each alcohol drinking? Would that pattern be the same for male and female?

## Survey Design

### About NHANES

We decided to analyze through National Health and Nutrition Examination Survey (NHANES), which is a program of studies designed to assess the health and nutritional status of adults and children in US. According to this reliable source of information about America's health, we can refine our analysis on nationally representative data about physical activity and fitness levels.

### Design Elements

First Stage PSU: all the United States counties (combine some small counties as one large county).

Second Stage PSU: all census blocks or combination of census blocks.

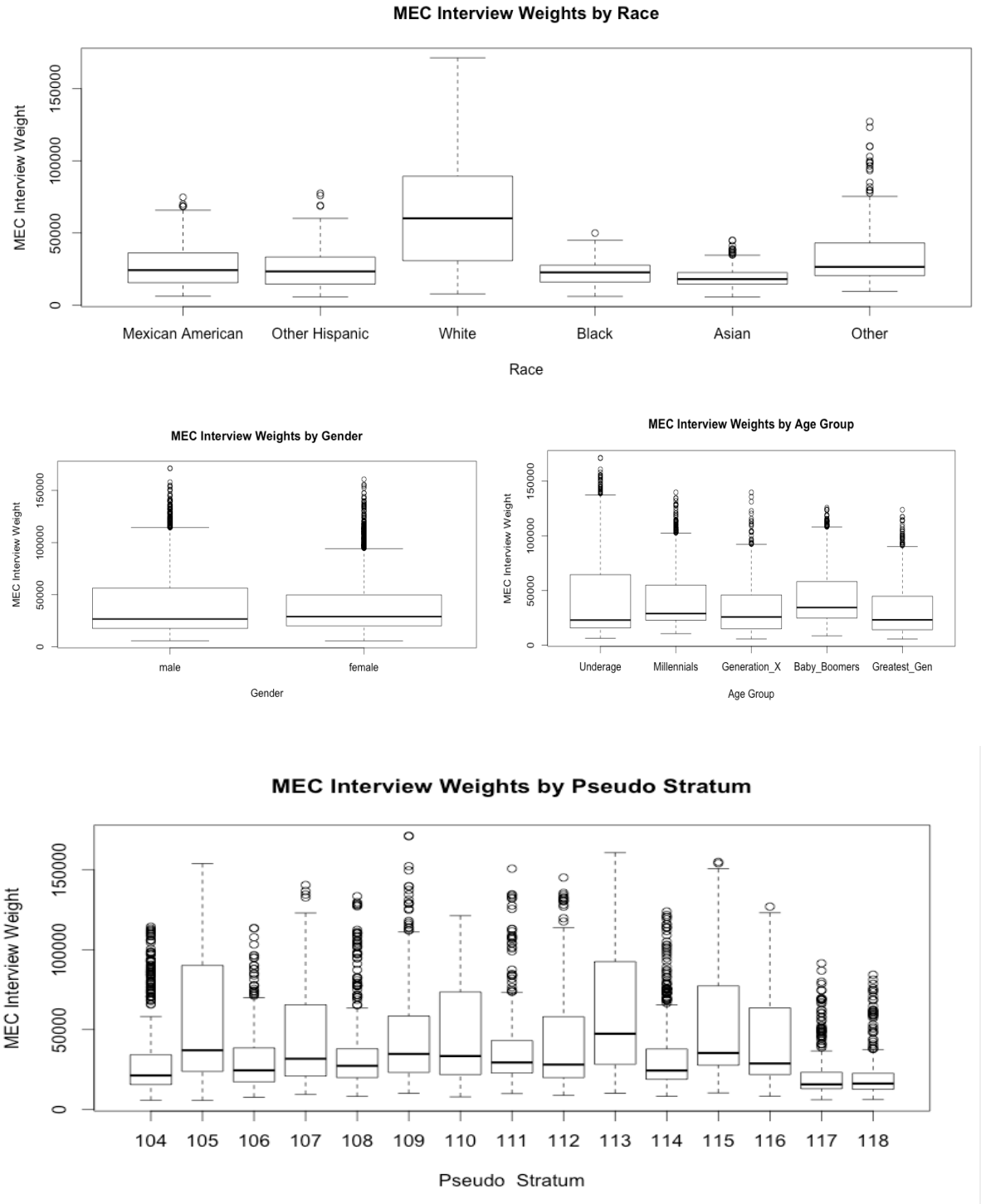
### Exploration of the design elements

Here we want to exam the weight distribution among the stratums. There are two different sets of weights given in the public release data. One is called full_interview_wt and the other is full_exam_wt, each for the interview part and physical examination part respectively. Here since our primary variable of interest “Alcohol Use” comes from the interview section, we will use full_interview_wt as our weights for each individual respondent.

The weights range from 5528.735 to 171395.3, with mean value 40110 and median 27760. We also further discover that the individual with minimum weight is an Asian female aged 19, the individual with the maximum weight is White male aged 61, and the individual with the adjusted median is also Asian female aged 25. The disproportionate sampling probabilities occur in the stratification. If the population is sampled with a higher probability, the weight of the population is smaller. Therefore, from the plot below we can see that the survey designer purposely oversamples areas containing large black and Mexican-American populations. Oversampling these populations allows comparisons of the health of racial and ethnic minorities. Oversampling means that a subgroup forms a small fraction of the total population. By oversampling we can reduce the margin of error.

If we ignore the weights in analyzing data, we are assuming implicitly that whites, blacks and Mexican Americans are largely interchangeable in overall health status and living habit, which is not generally true.

Figure 1: interview weights



## Methodology

The data we obtained is from National Health and Nutrition Examination Survey 2013-2014 (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013). The data is separated by topic and we work on the Demographics Data (DEMO_H Data), Alcohol use questionnaire (ALQ_H Data), and Blood Pressure examination result (BPX_H Data). We try to incorporate both quantitative data and qualitative data, thus we choose data from both the questionnaire and examination. The data was originally in SAS and we use "Hmisc" package to read them in R. Firstly, we get our needed variables from the three datasets (Refer to variable list below), and we merge those variables by id (unique id for responses). We notice that the alcohol use only contains data for adults 18 years old and older. This is reasonable since the legal drinking age is 21 in the United States. And since we are studying everything based on alcohol use, we don't need the data for 17 years old or younger, we simply drop them.

| Variable Name from NHANES | Variable names in our analysis | Description | Values |
|---|---|---|---|
| SEQN | id | Respondent sequence number | Unique id for each respondent |
| RIAGENDER | gender | gender | 1=male<br>2=female |
| RIDAGEYR | age | Age when response | Numeric |
| RIDRETH3 | race | Race/Hispanic origin w/ NH Asian | 1 =Mexican American<br>2=Other Hispanic<br>3=Non-Hispanic White<br>4=Non-Hispanic Black<br>6=Non-Hispanic Asian<br>7=Other, including multi-racial |
| SDMVPSU | pseudo_psu | Masked Variance Pseudo PSU | 1 to 3, nested within each stratum |
| SDMVSTRA | pseudo_stratum | Masked Variance Pseudo Strata | 90-103 representing 14 strata |

| | | | |
|---|---|---|---|
| ALQ120Q | freq | How often drink alcohol over past 1 year | Range 0-365<br>777 for refuse<br>999 for don't know |
| ALQ120U | unit | # days drink alcohol per wk, mo, yr | 1=week<br>2=month<br>3=year<br>7=refuse<br>9=don't know |
| ALQ130 | avg | Avg # alcoholic drinks/day - past 12 mos | Range 1-25<br>777 for refuse<br>999 for don't know |
| ALQ151 | overdrink | Ever have 4/5 or more drinks every day? | 1=yes<br>2=no<br>7=refused<br>9=don't know |
| BPXSY1 | systolic | Systolic: Blood pres (1st rdg) mm Hg | Numeric:<br>Range 66 to 228 |
| BPXDI1 | aiastolic | Diastolic: Blood pres (1st rdg) mm Hg | Numeric:<br>Range 0 to 122 |

**Category:**

We divide the age to five group according to respondents' generation(divided into 5 groups:

| | |
|---|---|
| Underage | 18-21 |
| Millennials | 22-36 |
| Generation X | 37-48 |
| Baby Boomers | 49-67 |
| Greatest Generation | 68+ |

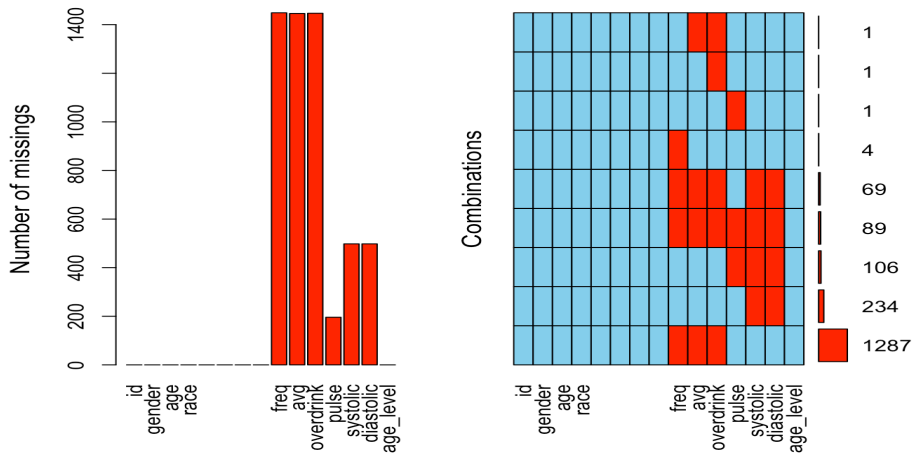
We categorize the frequency of alcohol consumption in a year (2013-2014) by dividing them into 5 level:

| | |
|---|---|
| Never Drink | 0 days |
| Drink Monthly | 1-12 days |
| Drink Weekly | 13-52 days |
| Drink Weekly | 53-200 days |
| Drink Almost Every Day | 200+ days |

**Dealing with nonresponse:**

We noticed that in our alcohol data large number of missing value. We plotted the missing value from all of our data in figure 2.

Figure 2:




From figure 2, we can clearly see that the non-response rate of alcohol use is very high. And those non-responses are not Missing completely at random (MCAR). There is various reason of missing those data, for example, people might be ashamed to answer the average drink question because they drink too much. Taking that into consideration, we decide to use imputation for the missing data. For the frequency of drink, average of drink and blood pressure, we use argImpute to impute the missing data. And for the overdrink data, we use hot deck imputation.

## Result

The analysis focuses on the relationship of alcohol usage and many different other variables such as race, age and gender. The variables in alcohol usage questionnaire including frequency of drinking the alcohol over the past 12 months, and the average number of alcoholic drinks per time over the past 12 months. All analyses were done using R survey package, while some of the graphs we produced also use ggplot2 package. All the graphs consider the influence of full_interview_wt provided by NHANES. First of all, we use the following code to generate a survey package object.

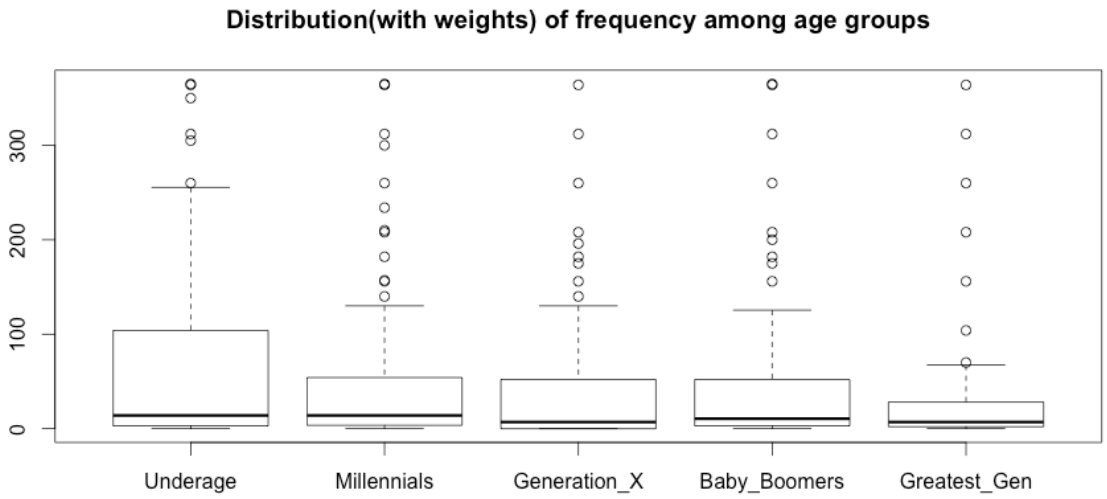
NHANESdesign<-svydesign(ids=~pseudo_psu+id, strata=~pseudo_stratum, nest=T, weights=~full_interview_wt, data=NHANES)

Question 1: Is there any difference in drinking frequency among different age groups?

To examine the pattern, we divide the age range into 5 independent groups. We use the specific definition for Millennials, Baby Boomers, the Greatest Generation and Generation X to assign individuals to the groups (based on the age when they were interviewed). We are doing so because we are inspired by some reports in the newspaper citing that the young male drinks the most, and that the Millennials drinks more than any other age groups. We also define an additional age group "Underage" from 18-21 to see the pattern of underage alcoholic usage.

| | mean | SE | X2.5.. | X97.5.. |
|---|---|---|---|---|
| Underage | 0.30995727 | 0.010399852 | 0.2895739 | 0.33034061 |
| Millennials | 0.21237626 | 0.010102289 | 0.1925761 | 0.23217638 |
| Generation_X | 0.14182801 | 0.006745222 | 0.1286076 | 0.15504840 |
| Baby_Boomers | 0.26061103 | 0.013068342 | 0.2349975 | 0.28622451 |
| Greatest_Gen | 0.07522743 | 0.003895650 | 0.0675921 | 0.08286277 |

In addition, we also categorize the of frequency of alcohol consumption by dividing them into 5 levels: "Never Drink"(0 days), "Drink Monthly"(0-12 days), "Drink Weekly"(12-52 days), "Drink every

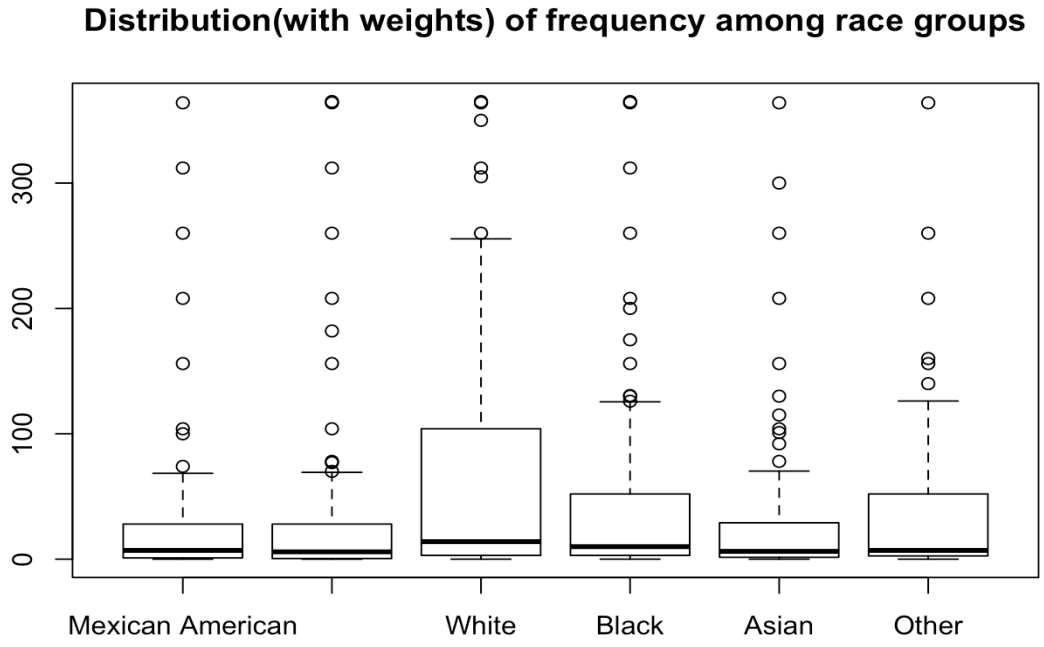



two days"(52-200 days) and "Drink Almost Every Day"(200+ days).

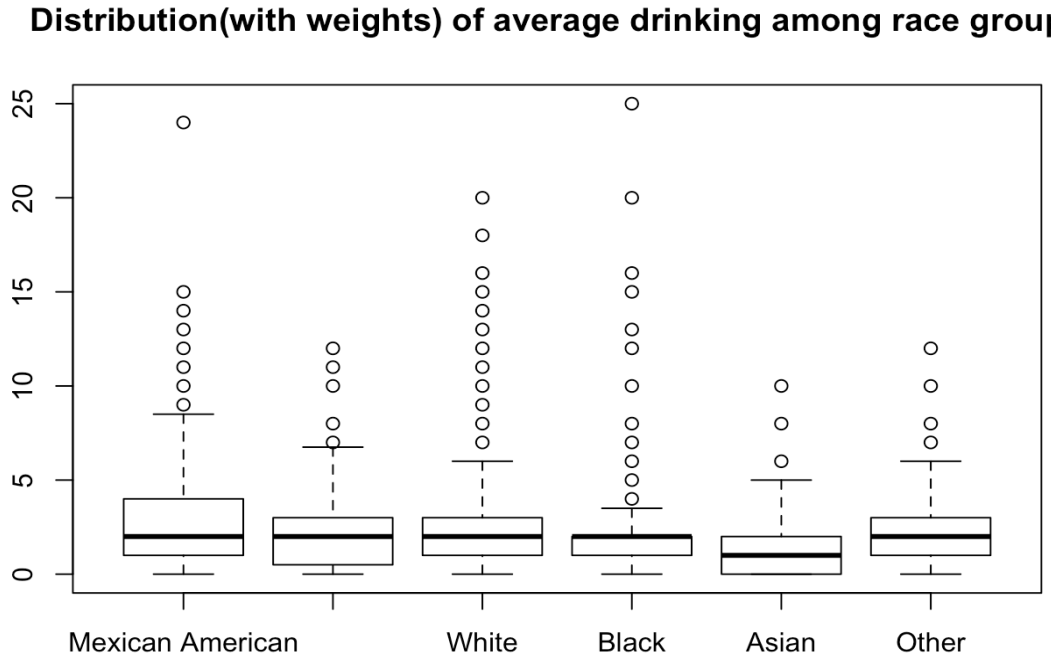
| | mean | SE | X2.5.. | X97.5.. |
|---|---|---|---|---|
| Drink Almost Everyday | 0.09651081 | 0.010430500 | 0.0760674 | 0.1169542 |
| Drink Every Two Days | 0.14930005 | 0.006052574 | 0.1374372 | 0.1611629 |
| Drink Weekly | 0.34898941 | 0.007545790 | 0.3341999 | 0.3637789 |
| Drink Monthly | 0.24353190 | 0.009732955 | 0.2244557 | 0.2626081 |
| Never Drink | 0.16166784 | 0.009732052 | 0.1425934 | 0.1807423 |

By looking at the table and the box-plot that takes weight information into consideration, we could see that teenagers from 18-21 tend to drink the most frequently than the remaining age groups, with the mean value 0.3099 and estimated standard error 0.0104. In addition, people from the Baby Boomers generation also drinks very frequently, with the mean value 0.2606 and estimated standard error 0.0131.

**2) Will different race groups tend to have different alcohol drinking patterns? How does it vary between different race?**



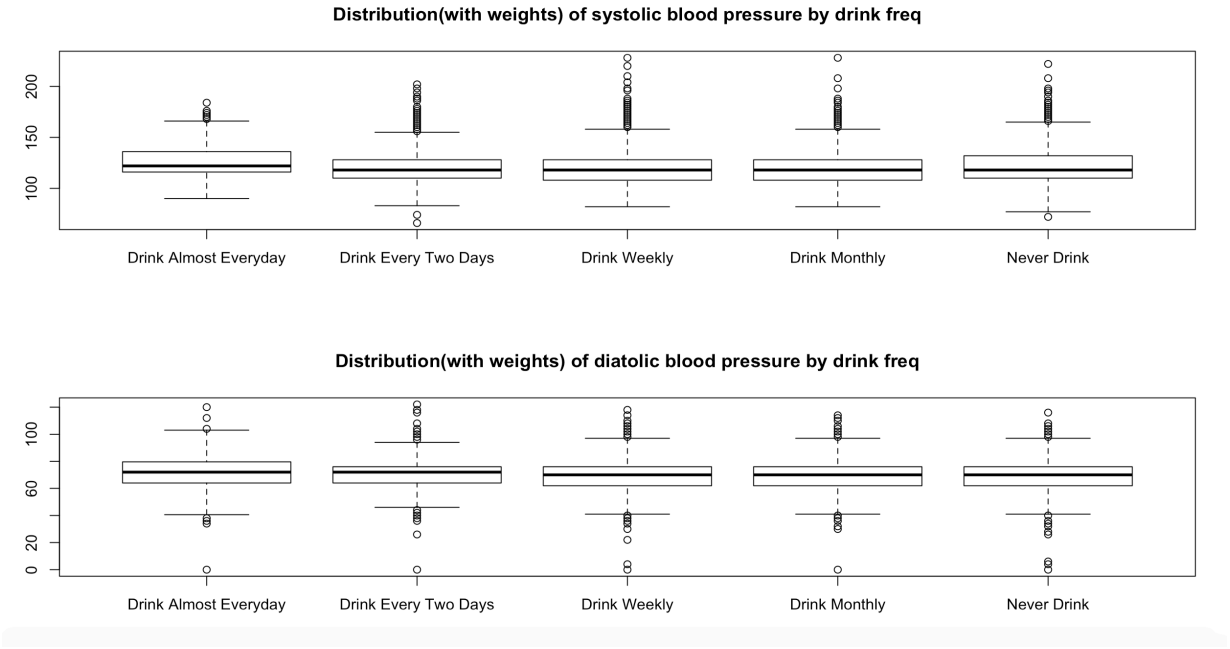

We compare the frequency level in the past 12 months of drinking alcoholic beverage among different race groups firstly. The querioneror have been asked, “how often did you drink any type of alcoholic beverage ”. The result shows in the Figure_X above. Generally speaking,  it turns out that White people tend to drink the most frequency in the past 12 months, whereas Asian people tend to drink least frequency, however, both of these two groups do have that much difference than others. Also, Black people has average level of frequency of drinking, but the group have an outstanding standard error, which suggests that some of Black people tend to drink much more frequency than the ordinary. Other groups, including Mexican American and Other Hispanic, drink normal frequency in past 12 months and with small standard error.



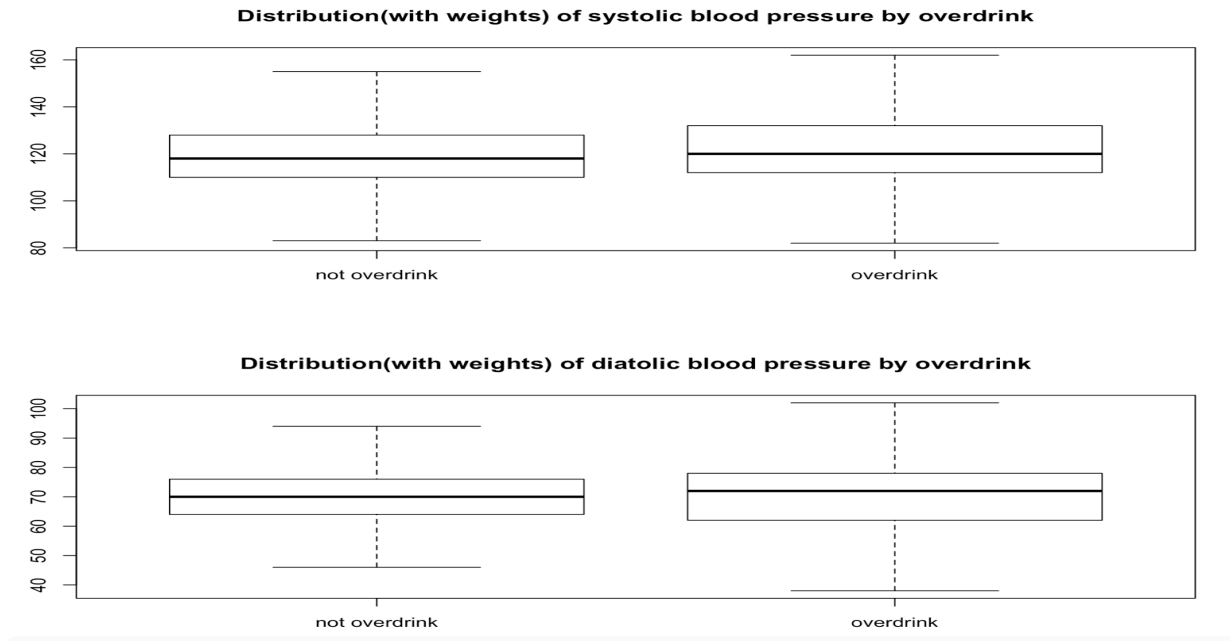

We then compare the average amount of drinking alcoholic beverages on those days they drank in the 12 months among different race groups. The result shows in the Figure_X above. It shows that Asian people have a significant lower amount of average drink than all other groups. It actually follows people's stereotype that Asian people aren't not party animal in general. Additionally, Mexican Americans have a greatest variance, that indicates they have more population drink more amount than other groups. According to the above box plot of Black people, we can tell that their 50% -75% of population drink almost same amount of alcoholic beverages on those days they drank. Other groups, including White and Other Hispanic, drink average amount on those days that they drank alcohol.

Additionally, we also slightly touched on the effect of alcohol to cardio health. We measured the cardiac health by blood pressure. We categorized the blood pressure by "Normal", Prehypertension", "High Blood Pressure Stage 1","High Blood Pressure Stage 2", which is

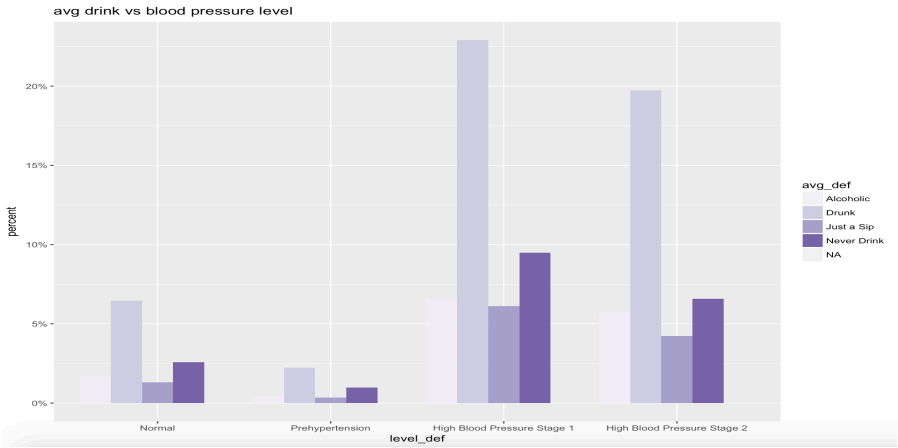
generally how medical field measure the blood pressure. We firstly explore the relationship between drinking frequency and blood pressure (see figure below)




We can clearly see that people who drink alcohol almost every day has slightly higher systolic and diastolic blood pressure than others. However, it seems that there's not very much differences between them, we can guess there is certain threshold when alcohol became harmful. And there is no such significant effect if you drink a reasonable amount.
The plot of overdrink vs blood pressure gives more explicit result. See figure below:



We can clearly see that the median and 75 percentiles and 100 percentiles of overdrink people's blood pressure are significantly higher than those who did not overdrink. We can conclude that overuse of alcohol is associate with high blood pressure. And also from the plot of level of blood pressure and average of drink per day:




We can see that people who are alcoholic has higher percentage of high blood pressure compare to people who don't drink much.

## Discussion/Conclusion

Coming to the conclusion, the health effects of alcohol are indeed quite complex as mentioned previously, consisting of more dimensions than simple straightforward linear relationships.

First of all, regarding to the 5 age groups, the result illustrates a reversed bell-shape distribution. To be more specific, teenagers from 18-21 tend to drink the most frequently than the remaining age groups, with the mean value 0.3099 and estimated standard error 0.0104. Meanwhile, people from the Baby Boomers generation also drinks very frequently, with the mean value 0.2606 and estimated standard error 0.013. According to the research, Since 2000, the high-frequency alcohol drinker segment has more than doubled — from 7.6% of all U.S. LDA (legal drinking age) adults in 2000 to 13% in 2015.From 2000 through 2005, occasional wine drinkers surged from 18% to 26% of all U.S. LDA adults. This was driven by a drop in non-adopter adults from 33% of the legal drinking age population to 24%. More importantly, between 2005 and 2010, there was a surge in high-frequency wine drinkers from 7.9% to

13.9% of the LDA population, driven by the Millennials. This also accounted for a decline in the occasional wine drinker population from 26.2% to 20.3%.

Secondly, in terms of  race group, the graph shows that white people have the highest frequency among all other race groups while Asian people tend to drink the least. Black people has lower average amount of drink than White people, however, associated with a large standard error. This huge volatility suggests that some Black people tend to drink much more than the ordinary which may due to sociological reasons that we will not discuss in this paper.

# stat152_project

## Load data

```
demo<-sasxport.get("~/downloads/DEMO_H.XPT")  blood_pressure<-
sasxport.get("~/downloads/BPX_H.XPT")  alcohol<-
sasxport.get("~/downloads/ALQ_H.XPT")  depression<-sasxport.get("~/downloads/DPQ_H.XPT")
```

## deal with alcohol_freq

we found that the freq that data give us are in different units(per year, per week, or per month). Here we convert it into same unit(per year)

```
alcohol=alcohol[,c(1,4,5,6,9)]  names(alcohol)=c("id","freq","unit","avg","overdrink")  alcohol$unit<-
with(alcohol,ifelse(alcohol$freq==0, 0,alcohol$unit))  alcohol$avg<-
with(alcohol,ifelse(alcohol$freq==0, 0,alcohol$avg))  alcohol$overdrink<-
with(alcohol,ifelse(alcohol$freq==0, 2,alcohol$overdrink))  alcohol$freq <- with(alcohol,
ifelse(alcohol$unit == 1, alcohol$freq*52,  ifelse(alcohol$unit == 2, alcohol$freq*7, alcohol$freq)))
```

## Clean data

```
demo1=demo[,c(1,4,5,8,41,42,43,44)]  names(demo1)=c("id","gender","age","race","full_exam_wt","f
ull_interview_wt","pseudo_psu","pseudo_stratum")  head(demo1)  alcohol=alcohol[,c(1,2,4,5)]  head(
alcohol)  physical=blood_pressure[,c(1,8,12,13)]  names(physical)=c("id","pulse","systolic","diastolic"
)  head(physical)
```

## Merge data together

```
NHANES = merge(demo1, alcohol, by="id")  NHANES = merge(NHANES, physical,
by="id")  head(NHANES)  nrow(NHANES)
```

## Make the table readable by defining the levels

```
NHANES$gender = as.factor(NHANES$gender)  levels(NHANES$gender) = c("male",
"female")  NHANES$race = as.factor(NHANES$race)  levels(NHANES$race) = c("Mexican
American", "Other Hispanic", "White", "Black", "Asian", "Other")  # add a new var "age_level"
```

```
indicating a particular age group the respondent belongs to  NHANES$age_level <- with(NHANES,
ifelse(NHANES$age >=18 & NHANES$age <= 21, "Underage",  ifelse(NHANES$age >=22 &
NHANES$age <= 36, "Millennials",  ifelse(NHANES$age >=37 & NHANES$age <= 48,
"Generation_X",  ifelse(NHANES$age >=49 & NHANES$age <= 67,
"Baby_Boomers","Greatest_Gen")))))  NHANES$age_level =
as.factor(NHANES$age_level)  levels(NHANES$age_level) = c("Underage", "Millennials",
"Generation_X", "Baby_Boomers", "Greatest_Gen")  head(NHANES)  summary(NHANES)
```

## Calculating nonresponse

```
length(which(is.na(NHANES$freq)))/nrow(NHANES)
#0.3934841  length(which(is.na(NHANES$avg)))/nrow(NHANES)  length(which(is.na(NHANES$o
verdrink)))/nrow(NHANES)  length(which(is.na(NHANES$systolic)))/nrow(NHANES)  length(whic
h(is.na(NHANES$diastolic)))/nrow(NHANES)  length(which(is.na(NHANES$sleeping)))/nrow(NH
ANES)  length(which(is.na(NHANES$apetite)))/nrow(NHANES)  length(which(is.na(NHANES$de
pressed)))/nrow(NHANES)
```

## Explorational Graphical Display

```
ggplot(NHANES, aes(freq, systolic))+geom_boxplot(aes(gender))
```

## Preliminary analysis of design elements

```
# exam weights summary  # Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   #
5529   18750   27760   40110   53090  171400   summary(NHANES$full_interview_wt)  #find who is
assigned the lowest, largest and median
weight  NHANES[which(NHANES$full_interview_wt==min(NHANES$full_interview_wt)),]
 NHANES[which(NHANES$full_interview_wt==max(NHANES$full_interview_wt)),]  NHANES[w
hich(NHANES$full_interview_wt==median(NHANES$full_interview_wt[c(2:5924)])),] #adjust
median since total length is even    # visulize exam weight
distribution  boxplot(NHANES$full_interview_wt~NHANES$race,main="MEC Interview Weights by
Race", xlab="Race",  ylab="MEC Interview
Weight")  boxplot(NHANES$full_interview_wt~NHANES$gender,main="MEC Interview Weights by
Gender", xlab="Gender",  ylab="MEC Interview
Weight")  boxplot(NHANES$full_interview_wt~NHANES$age_level,main="MEC Interview Weights
```

by Age Group", xlab="Age Group", ylab="MEC Interview Weight") **boxplot**(NHANES$full_interview_wt~NHANES$pseudo_stratum,main="MEC Interview Weights by Pseudo Stratum",xlab="Pseudo Stratum",ylab="MEC Interview Weight") *# The disproportionate sampling probabilities occur in the stratification. Purposely oversample areas containing large black and Mexican-American populations. Oversampling these populations allows comparision of the health of racial and ethnic minorities. # If we ignore the weights in analyzing data, we are assuming implicitely that whites, blacks and Mexican Americans are largely interchangeable in health status, which is not generally true. # Oversampling means that a subgroup forms a small fraction of tha total population. By oversampling we can reduce the margin of error*

## check missing value

missing = NHANES[!**complete.cases**(NHANES),] **head**(missing) **sum**(**is.na**(missing)) *#10768* **aggr**(missing, prop = F, numbers = T) *#the graph shows the amount of missing value # x <- missing[, 10:12]*

## Deal with nonresponse by using freq-hotdeck imputation

##k-nearest neighbout method for hotdeck imputation ##NHANES$freq[is.na(NHANES$freq)]=mean(NHANES$freq[!is.na(NHANES$freq)]) # #NHANES1=kNN(NHANES,variable=c("freq","avg"),dist_var=c("age","avg","pulse","systolic","diastolic"),k=500) impute_arg <- **aregImpute**(~freq+avg+pulse+age+gender+race, data = NHANES, n.impute = 10) new_values_freq<-**apply**(impute_arg$imputed$freq, 1, median) NHANES$imputed_freq_reg<-NHANES$freq for(i in 1:**length**(new_values_freq)){ NHANES[**rownames**(NHANES)==**names**(new_values_freq)[i],]$imputed_freq_reg<- new_values_freq[i] } **hist**(NHANES$freq) **hist**(NHANES$imputed_freq_reg) *#check imputed value* N<-**nrow**(NHANES) NHANES$N<-N srs_design<-**svydesign**(id=~1, fpc=~N, data = NHANES) **svymean**(~imputed_freq_reg, srs_design) ##impute for "avg" NHANES$avg[NHANES$avg==999]=NA impute_arg <- **aregImpute**(~freq+avg+pulse+age+gender+race, data = NHANES, n.impute = 10) new_values_avg<-**apply**(impute_arg$imputed$avg, 1, median) NHANES$imputed_avg_reg<-NHANES$avg for(i in 1:**length**(new_values_avg)){ NHANES[**rownames**(NHANES)==**names**(new_values_avg)[i],]$impute

```
d_avg_reg<-
new_values_avg[i] } NHANES$imputed_avg_reg[NHANES$imputed_freq_reg==0]=0 hist(NHAN
ES$freq) hist(NHANES$imputed_avg_reg) #check imputed value srs_design<-svydesign(id=~1,
fpc=~N, data = NHANES) svymean(~imputed_avg_reg, srs_design) ##impute for
overdrink NHANES[562, 12] = NA #treat the only "unknow" as NA ##for(j in 1:length(y2)){ for(i in
1:length(x1)){ ## t=data[data$country_code==x1[i]&data$region==y2[j],] #non-response
rate=0.125543 #set.seed(222222) #
t[t$freq_pmt_to_courts<0,]$freq_pmt_to_courts=sample(t[t$freq_pmt_to_courts>0,]$freq_pmt_to_co
urts,length(t[t$freq_pmt_to_courts<0,]$freq_pmt_to_courts),replace=T) #data[data$country_code==
x1[i]&data$region==y2[j],]=t } #} impute.overdrink.hotdeck<-
function(overdrink=NHANES$overdrink,age_level=NHANES$age_level,seed=2333){ na.indices<
-which(is.na(NHANES$overdrink)) #Examining these show that the age_levels of these are only
about right and too thin underage<-na.omit(overdrink[which(age_level=="Underage")])
millennials<-na.omit(overdrink[which(age_level=="Millennials")]) generation_x<-
na.omit(overdrink[which(age_level=="Generation_X")]) baby_boomers<-
na.omit(overdrink[which(age_level=="Baby_Boomers")]) greatest_gen<-
na.omit(overdrink[which(age_level=="Greatest_Gen")]) for(i in
na.indices){ if(age_level[i]=="Underage"){ overdrink[i]<-sample(underage, 1) }else
if(age_level[i]=="Millennials"){ overdrink[i]<-sample(millennials, 1) }else
if(age_level[i]=="Generation_X"){ overdrink[i]<-sample(generation_x, 1) }else
if(age_level[i]=="Baby_Boomers"){ overdrink[i]<-sample(baby_boomers, 1) }else
if(age_level[i]=="Greatest_Gen"){ overdrink[i]<-sample(greatest_gen,
1) } } return(overdrink) } NHANES$overdrink<-impute.overdrink.hotdeck()
```

## Analysis of alchohol consumption pattern and habit

```
#Exploratory Plots not adjusted by weights
mosaicplot(table(NHANES$freq,NHANES$gender),main="Frequency of Alchohol Use vs
Gender",las=1) mosaicplot(table(NHANES$freq,NHANES$age_level),main="Frequency of
Alchohol Use vs Age Group",las=1)
```

```
mosaicplot(table(NHANES$freq,NHANES$race),main="Frequency of Alchohol Use vs Race",las=1)
```

## create an addtional column about the level of one's drinking frequency

```
#define freq as never occasionally sometimes oftern almost everyday NHANES$freq_def<-c() NHANES$freq_def[NHANES$imputed_freq_reg==0] <- "Never Drink" NHANES$freq_def[NHANES$imputed_freq_reg<=12 & NHANES$imputed_freq_reg!=0] <- "Drink Monthly" NHANES$freq_def[NHANES$imputed_freq_reg<=52 & NHANES$imputed_freq_reg>12] <- "Drink Weekly" NHANES$freq_def[NHANES$imputed_freq_reg<=200 & NHANES$imputed_freq_reg>52] <- "Drink Every Two Days" NHANES$freq_def[NHANES$imputed_freq_reg>200] <- "Drink Almost Everyday" NHANES$freq_def<-as.factor(NHANES$freq_def) ##order levels NHANES$freq_def <- factor(NHANES$freq_def, labels=c("Drink Almost Everyday","Drink Every Two Days", "Drink Weekly","Drink Monthly","Never Drink"))
# survey package NHANESdesign<-svydesign(ids=~pseudo_psu+id, strata=~pseudo_stratum, nest=T, weights=~full_interview_wt, data=NHANES) summary(NHANESdesign) # proportion for freq svytable(~freq_def, design = NHANESdesign) freqPro <- svymean(~freq_def, design = NHANESdesign, na.rm = TRUE) freqConfit <- confint(svymean(~freq_def, design=NHANESdesign, na.rm = TRUE)) freqPro <-data.frame(freqPro) freqConfit <- data.frame(freqConfit) freqTable <- bind_cols(freqPro,freqConfit) rownames(freqTable) <- c("Drink Almost Everyday","Drink Every Two Days", "Drink Weekly","Drink Monthly","Never Drink") #proportion for race svytable(~race, design = NHANESdesign) racePro <- svymean(~race, design = NHANESdesign, na.rm = TRUE) raceConfit <- confint(svymean(~race, design=NHANESdesign, na.rm = TRUE)) racePro <- data.frame(racePro) raceConfit <- data.frame(raceConfit) raceTable <- bind_cols(racePro,raceConfit) rownames(raceTable) <- c("Mexican American","Hispanic", "White","Black","Asian","Other") #proportion for age svytable(~age_level, design = NHANESdesign) agePro <- svymean(~age_level, design = NHANESdesign, na.rm = TRUE) ageConfit <- confint(svymean(~age_level, design=NHANESdesign, na.rm = TRUE)) agePro <-data.frame(agePro) ageConfit <- data.frame(ageConfit) ageTable <- bind_cols(agePro,ageConfit) rownames(ageTable) <-
```

```
c("Underage","Millennials","Generation_X","Baby_Boomers","Greatest_Gen")  # graphic display of drinking pattern and habit among different age groups  #ggplot(NHANES, aes(y=imputed_freq_reg, x=age_level)) + #geom_boxplot(aes(color=age_level, fill=age_level, group=age_level), alpha = 0.5, #outlier.size=3, notch = FALSE)  svyboxplot(imputed_freq_reg~age_level, NHANESdesign, all.outliers = TRUE, main = "Distribution(with weights) of frequency among age groups")  #ggplot(NHANES, aes(y=imputed_freq_reg, x=race)) + geom_boxplot(aes(color=race, #fill=race, group=race), alpha = 0.5, outlier.size=3, notch = FALSE)  svyboxplot(imputed_freq_reg~race, NHANESdesign, all.outliers = TRUE, main = "Distribution(with weights) of frequency among race groups")  # cannot include weight information  ggplot(NHANES, aes(y=imputed_freq_reg, x=age_level)) + geom_boxplot(aes(color=age_level, fill=age_level, group=age_level),alpha = 0.5, outlier.size=0.1, notch = FALSE) + facet_wrap(~race) + labs(title ="Drinking Pattern and Habit between Different Age Group and Race", x = "Age Group", y = "Number of Days Per Year That You Drink")  # further examination of those who drinks the most  # consider weights when plotting  most_frequently <- NHANES%>%filter(freq_def == "Drink Every Two Days" | freq_def =="Drink Almost Everyday")%>%group_by(age_level) %>% summarise(sum = sum(full_interview_wt))  ggplot(most_frequently, aes(age_level))+geom_bar(aes(weight = sum, fill=age_level))
```

## Q2

```
#define average as Never Drink(0),just a sip(1-3), drunk(4-10), alcoholic(11-30)  NHANES$avg_def<- c()  NHANES$avg_def[NHANES$imputed_avg_reg==0] <- "Never Drink"  NHANES$avg_def[NHANES$imputed_avg_reg<=3.0 & NHANES$imputed_freq_reg!=0] <- "Just a Sip"  NHANES$avg_def[NHANES$imputed_avg_reg<=10.0 & NHANES$imputed_freq_reg>3.0] <- "Drunk"  NHANES$avg_def[NHANES$imputed_freq_reg<=30.0 & NHANES$imputed_freq_reg>10.0] <- "Alcoholic"  NHANES$avg_def<- as.factor(NHANES$avg_def)  # proportion for avg  NHANESdesign<- svydesign(ids=~pseudo_psu+id, strata=~pseudo_stratum, nest=T, weights=~full_interview_wt, data=NHANES)  svytable(~avg_def, design = NHANESdesign)  avgPro <- svymean(~avg_def, design
```

```
= NHANESdesign, na.rm = TRUE)  avgConfit <- confint(svymean(~avg_def,
design=NHANESdesign, na.rm = TRUE))  avgPro <-data.frame(avgPro)  avgConfit <-
data.frame(avgConfit)  avgTable <- bind_cols(avgPro,avgConfit)  rownames(avgTable) <-
c("Alcoholic","Drunk", "Just a Sip","Never Drink")    svyboxplot(imputed_avg_reg~age_level,
NHANESdesign, all.outliers = TRUE, main = "Distribution(with weights) of average drinking among
age groups")    # cannot include weight information  ggplot(NHANES, aes(x=age_level,
y=imputed_avg_reg, fill=gender)) + geom_bar(stat = "identity", position = "dodge") +
scale_fill_manual(breaks = c("male", "female"), values=c("#27b1e5", "pink")) + labs(title ="Amounts
of Alcohol Consumption Every Time vs Age", x = "Age Group", y = "Avg Amounts of Alcoholic
Drinks per Day")    # Consider weights when ploting  most_frequently_2 <-
NHANES%>%filter(avg_def == "Alcoholic" | freq_def =="Drunk")%>%group_by(age_level) %>%
summarise(sum = sum(full_interview_wt))  ggplot(most_frequently_2,
aes(age_level))+geom_bar(aes(weight = sum, fill=age_level))    # average drinking vs
race  svyboxplot(imputed_avg_reg~race, NHANESdesign, all.outliers = TRUE)
```

## Q3

```
ggplot(NHANES, aes(x=imputed_freq_reg, y=imputed_avg_reg, color=gender, size =
full_interview_wt)) + geom_point(alpha=0.5)+scale_color_manual(breaks = c("male", "female"),
values=c("#27b1e5", "#ef9bba"))+geom_smooth(method=lm,se=FALSE,fullrange=TRUE)
```

## imputation of blood pressure data

```
impute_arg <- aregImpute(~pulse+age+gender+race+systolic, data = NHANES, n.impute =
10)  new_values_systolic<-apply(impute_arg$imputed$systolic, 1,
median)  NHANES$imputed_systolic_reg<-NHANES$systolic    for(i in
1:length(new_values_systolic)){  NHANES[rownames(NHANES)==names(new_values_systolic)[i],]
$imputed_systolic_reg<-new_values_systolic[i]  }  impute_arg <-
aregImpute(~pulse+age+gender+race+diastolic, data = NHANES, n.impute =
10)  new_values_diastolic<-apply(impute_arg$imputed$diastolic, 1,
median)  NHANES$imputed_diastolic_reg<-NHANES$diastolic    for(i in
1:length(new_values_diastolic)){  NHANES[rownames(NHANES)==names(new_values_diastolic)[i]
,]$imputed_diastolic_reg<-new_values_diastolic[i]  }
```

## categorize blood pressure

```
NHANES$level_def<-
c() NHANES$level_def[NHANES$imputed_systolic_reg<120&NHANES$imputed_diastolic_reg<80]
<- "Normal" NHANES$level_def[NHANES$imputed_systolic_reg>=120 &
NHANES$imputed_systolic_reg<140 | NHANES$imputed_diastolic_reg>=80 &
NHANES$imputed_diastolic_reg<90] <-
"Prehypertension" NHANES$level_def[NHANES$imputed_systolic_reg>=140 &
NHANES$imputed_systolic_reg<160 | NHANES$imputed_diastolic_reg>=90 &
NHANES$imputed_diastolic_reg<100] <- "High Blood Pressure Stage
1" NHANES$level_def[NHANES$imputed_systolic_reg>=160
|NHANES$imputed_diastolic_reg>=100] <- "High Blood Pressure Stage 2" NHANES$level_def<-
as.factor(NHANES$level_def) NHANES$level_def <- factor(NHANES$level_def,
labels=c("Normal","Prehypertension", "High Blood Pressure Stage 1","High Blood Pressure Stage 2"))
```

## define overdrink

```
NHANES$overdrink_def<-c() NHANES$overdrink_def[NHANES$overdrink==1] <-
"overdrink" NHANES$overdrink_def[NHANES$overdrink==2] <- "not overdrink"
```

## analysis

```
#propotion of blood pressure level NHANESdesign<-svydesign(ids=~pseudo_psu+id,
strata=~pseudo_stratum, nest=T,data=NHANES) svytable(~level_def, design =
NHANESdesign) level_defPro <- svymean(~level_def, design = NHANESdesign, na.rm =
TRUE) level_defConfit <- confint(svymean(~level_def, design=NHANESdesign, na.rm =
TRUE)) level_defPro <-data.frame(level_defPro) level_defConfit <-
data.frame(level_defConfit) level_defTable <-
bind_cols(level_defPro,level_defConfit) rownames(level_defTable) <-
c("Normal","Prehypertension", "High Blood Pressure Stage 1","High Blood Pressure Stage 2")
```

## graph

```
##boxplot for Distribution(with weights) of blood pressure by drink
freq attach(mtcars) par(mfrow=c(2,1)) svyboxplot(imputed_systolic_reg~freq_def,
NHANESdesign, all.outliers = TRUE, main = "Distribution(with weights) of systolic blood pressure by
```

```
drink freq")  svyboxplot(imputed_diastolic_reg~freq_def, NHANESdesign, all.outliers = TRUE, main
= "Distribution(with weights) of diatolic blood pressure by drink freq")
attach(mtcars)  par(mfrow=c(2,1))  svyboxplot(imputed_systolic_reg~overdrink_def,
NHANESdesign, all.outliers = TRUE, main = "Distribution(with weights) of systolic blood pressure by
overdrink")  svyboxplot(imputed_diastolic_reg~overdrink_def, NHANESdesign, all.outliers = TRUE,
main = "Distribution(with weights) of diatolic blood pressure by overdrink")
ggplot(NHANES, aes(x=level_def, y = (..count..)/sum(..count..), fill=avg_def, size =
full_interview_wt)) + geom_bar(alpha=0.9,position = "dodge")+ scale_y_continuous(labels =
scales::percent)+labs(title="avg drink vs blood pressure
level",y="percent")+scale_fill_brewer(palette="Purples")
```