National Health Data Visualization

Rishab Ravi

This notebook uses datasets obtained from *kaggle*
The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations and includes demographic, socioeconomic, dietary, and health-related questions.
There are six components to the dataset -
1. Demographics
2. Examinations
3. Dietary
4. Laboratory
5. Questionnaire
6. Medication

You may view the same on kaggle as well: *NHDA*
This data analysis and visualization uses R plotting techniques/packages, markdown, and aims to obtain some insights/trends on various health conditions of the volunteers.

For the entire analysis, the following references were used, as supplied with the dataset:
1 2

Initial preprocessing

```
## import necessary libraries
library(plyr)
library(ggplot2)

## reading files
demographic = read.csv("demographic.csv")
diet = read.csv("diet.csv")
examination = read.csv("examination.csv")
labs = read.csv("labs.csv")
medications = read.csv("medications.csv")
questionnaire = read.csv("questionnaire.csv")

## merging files
dfList = list(demographic,examination,diet,labs,questionnaire,medications)
mainDF = join_all(dfList)
```

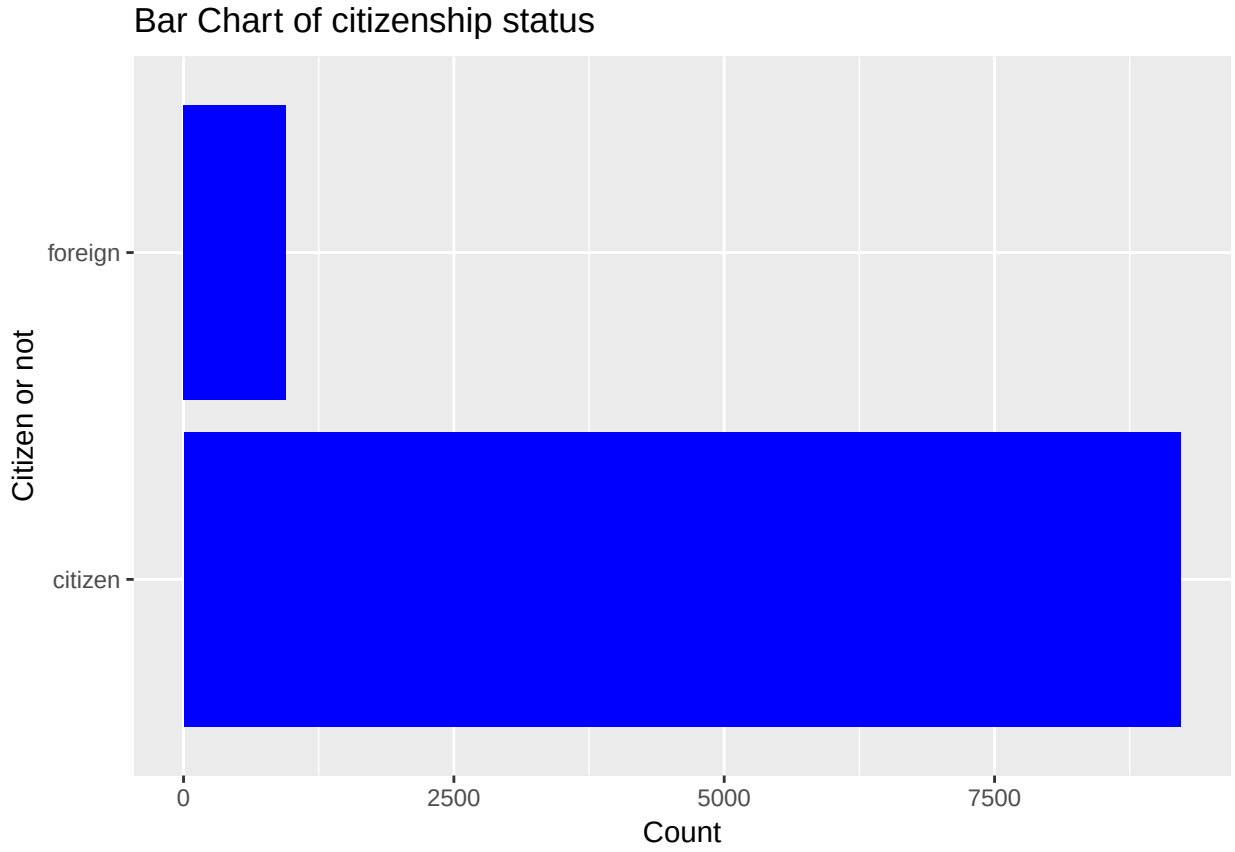
```
## Joining by: SEQN
## Joining by: SEQN
## Joining by: SEQN
## Joining by: SEQN
## Joining by: SEQN
```

We begin our visualization with the *demographics* component which provides individual, family and household level info.

This chunk shows the citizenship status of the respondents.

```
citiStatus<-subset(demographic,demographic$DMDCITZN<=2) ##ignore missing val
citiStatus$DMDCITZN[citiStatus$DMDCITZN==1]<-"citizen" ## rename
citiStatus$DMDCITZN[citiStatus$DMDCITZN==2]<-"foreign"

ggplot(citiStatus, aes(DMDCITZN)) + geom_bar(fill = "blue") + theme_bw() +
 xlab("Citizen or not") + ylab("Count") + coord_flip() +
 labs(title = "Bar Chart of citizenship status") + theme_gray()
```



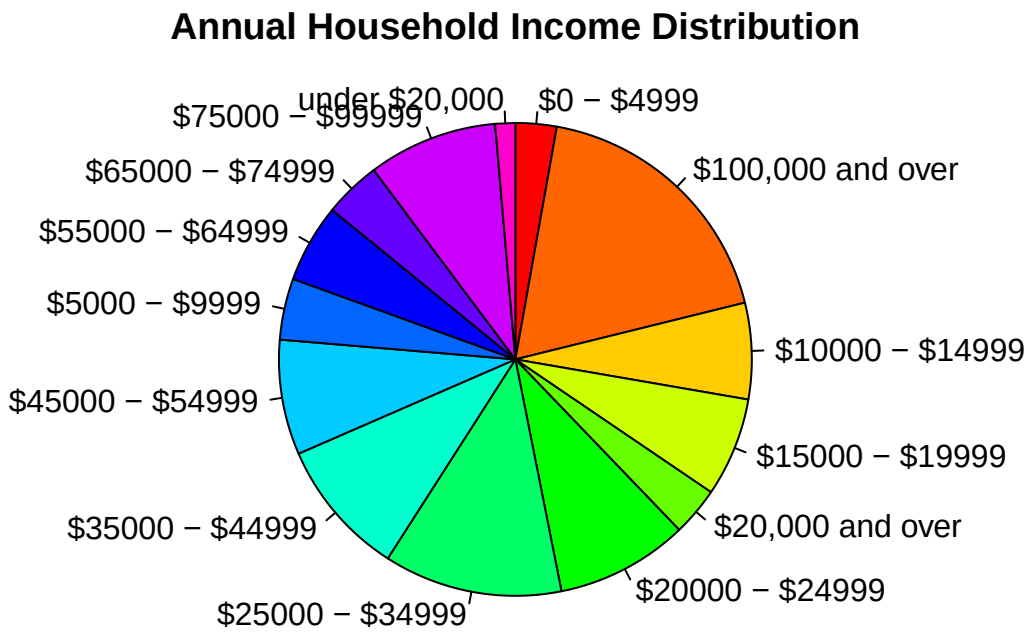
Now, we plot the annual household income of the volunteers.

```
annual_Hinc<-demographic$INDHHIN2
annual_Hinc_rep<-subset(annual_Hinc,annual_Hinc<=15) ##ignore unreported/missing val

##renaming observations:
annual_Hinc_rep[annual_Hinc_rep==1]<-"$0 - $4999"
annual_Hinc_rep[annual_Hinc_rep==2]<-"$5000 - $9999"
annual_Hinc_rep[annual_Hinc_rep==3]<-"$10000 - $14999"
annual_Hinc_rep[annual_Hinc_rep==4]<-"$15000 - $19999"
annual_Hinc_rep[annual_Hinc_rep==5]<-"$20000 - $24999"
annual_Hinc_rep[annual_Hinc_rep==6]<-"$25000 - $34999"
annual_Hinc_rep[annual_Hinc_rep==7]<-"$35000 - $44999"
annual_Hinc_rep[annual_Hinc_rep==8]<-"$45000 - $54999"
annual_Hinc_rep[annual_Hinc_rep==9]<-"$55000 - $64999"
annual_Hinc_rep[annual_Hinc_rep==10]<-"$65000 - $74999"
annual_Hinc_rep[annual_Hinc_rep==12]<-"$20,000 and over"
annual_Hinc_rep[annual_Hinc_rep==13]<-"under $20,000"
```

```
annual_Hinc_rep[annual_Hinc_rep==14]<-"$75000 - $99999"
annual_Hinc_rep[annual_Hinc_rep==15]<-"$100,000 and over"

##Piechart
pie(table(annual_Hinc_rep),
    clockwise = TRUE, main="Annual Household Income Distribution",
    radius = 1,col=rainbow(15))
```



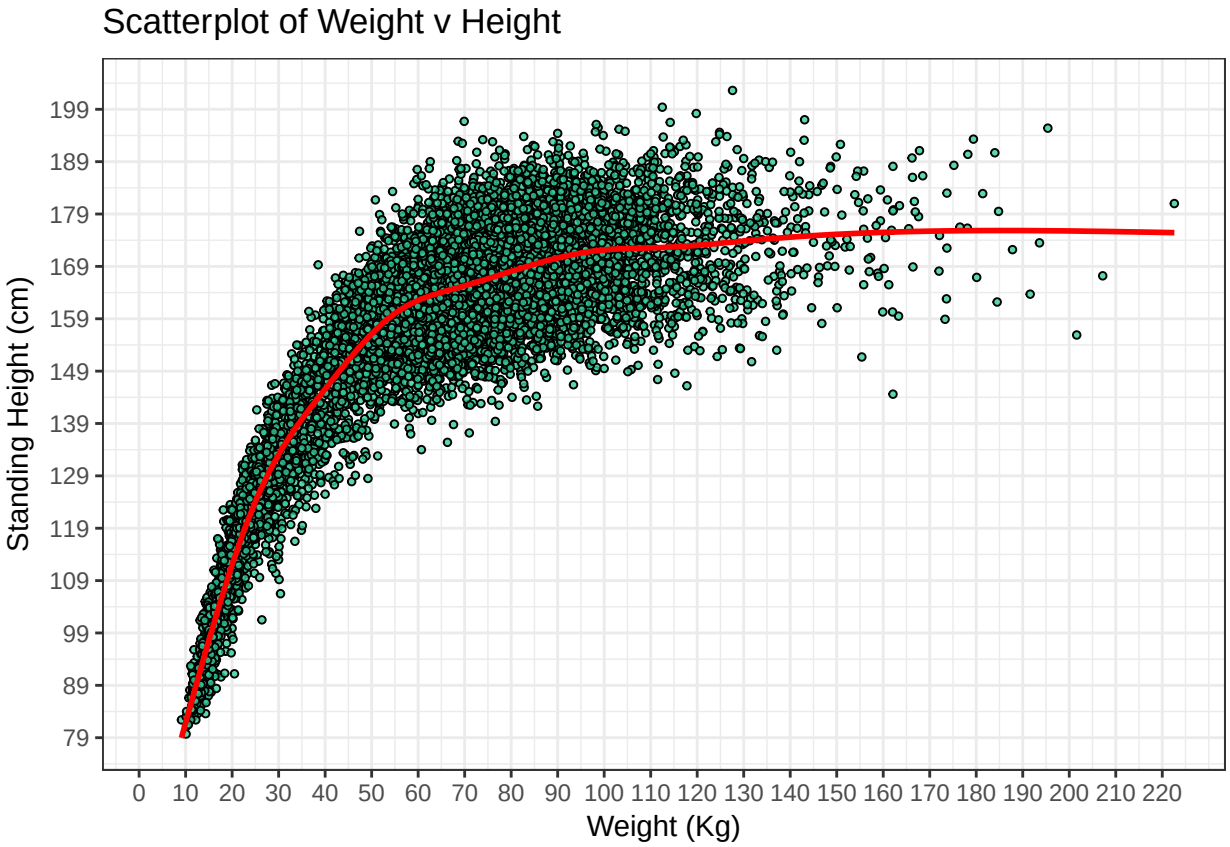
Moving to the *examinations* component, we have a scatterplot of the weigths of the individuals and their corresponding standing heights.

```
#scatterplot:
ggplot(examination, aes(BMXWT, BMXHT)) +
  geom_point(pch = 21, size = 1, fill = rgb(0.2,0.8,0.6,0.8)) +
  geom_smooth(method = "auto", color = "red", se = FALSE) +
  scale_x_continuous("Weight (Kg)", breaks = seq(0,225,10)) +
  scale_y_continuous("Standing Height (cm)", breaks = seq(79,210,by=10)) +
  labs(title="Scatterplot of Weight v Height") + theme_bw()
```

```
## `geom_smooth()` using method = 'gam'
```

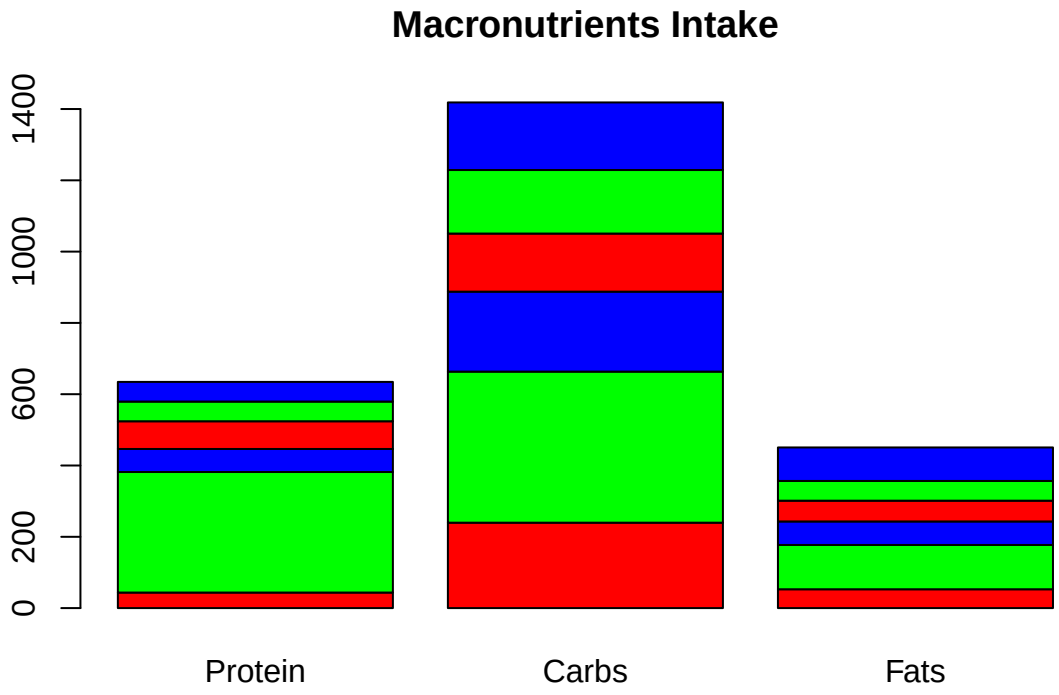
```
## Warning: Removed 758 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 758 rows containing missing values (geom_point).
```



The chunk below plots the daily macronutrients intake, as part of the *dietary* component. (Proteins, carbs and fats make up macronutrients)

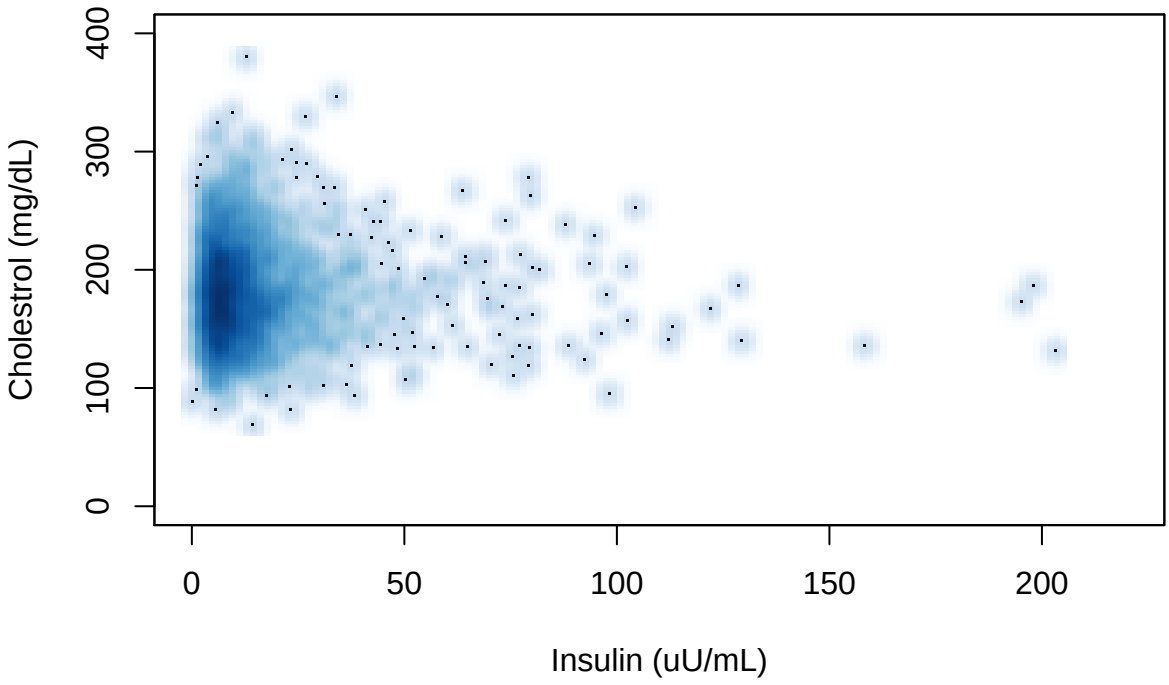
```
macros<-diet$DR1TPROT
macros<-cbind(macros,diet$DR1TCARB)
macros<-cbind(macros,diet$DR1TTFAT)
colnames(macros)<-c("Protein","Carbs","Fats")
cols<-c("red","green","blue")
barplot(macros,main="Macronutrients Intake",col=cols) #barplot
```



We have below a smooth scatterplot of cholestrol levels v insulin levels of individuals, from the *laboratory* component.

```
chin<-subset(labs,labs$LBXTC!="NA"&labs$LBXIN!="NA")

smoothScatter(chin$LBXIN,chin$LBXTC,xlim=c(0,220),
ylim=c(0,400),xlab = "Insulin (uU/mL)",ylab="Cholestrol (mg/dL)")
```



Now, let's try and correlate the effects of certain psychoactive drugs (cocaine/heroin/meth) on memory loss, obtained from the *questionnaire* component.

```
# cocaine/heroin/meth v difficulty thinking/remembering
drgUse <- subset(questionnaire,questionnaire$MCQ084<=2)
drgUse <- subset(drgUse,drgUse$DUQ240<=2)

exmp1 <- cbind(drgUse$SEQN,drgUse$DUQ240,drgUse$MCQ084)
colnames(exmp1) <- c("ID","Drug Use","Memory Loss")
exm2 <- data.frame(exmp1)
exm2 <- arrange(exm2,by=ID)
drgUsers <- subset(exm2,exm2$Drug.Use==1)
nonUse <- subset(exm2,exm2$Drug.Use==2)

as.data.frame(table(drgUsers$Memory.Loss))
```

```
##   Var1 Freq
## 1    1   28
## 2    2   87
```

```
as.data.frame(table(nonUse$Memory.Loss))
```

```
##   Var1 Freq
## 1    1   72
## 2    2  642
```

We find that **24.34%** reported memory loss among drug users while **10.09%** reported memory loss among those who haven't.

Let's now look at some mean samples of the dataset.

```
##mean height of representative sample
standingHeight <- mainDF$BMXHT ##Standing height (in cm)
bad <- is.na(standingHeight)
reportedHeight <- standingHeight[!bad]
meanHeight <- mean(reportedHeight)

##mean weight of representative sample
standingWeight <- mainDF$BMXWT ##Standing Weight (in kg)
bad2 <- is.na(standingWeight)
reportedWeight <- standingWeight[!bad]
meanWeight <- mean(reportedWeight)

##mean calorie intake of participants
cal <- mainDF$DR1TKCAL
bad4 <- is.na(cal)
calRep <- cal[!bad4] ##Reported participants
meanCal <- mean(calRep)

print("Mean Height of Participants (cm) - "); print(meanHeight)
```

```
## [1] "Mean Height of Participants (cm) - "
```

```
## [1] 160.2417
```

```
print("Mean Weight of Participants (kg) - "); print(meanWeight)
```

```
## [1] "Mean Weight of Participants (kg) - "
```

```
## [1] NA
```

```
print("Mean Daily Calorie Intake of Participants (kcal) - "); print(meanCal)
```

```
## [1] "Mean Daily Calorie Intake of Participants (kcal) - "
```

```
## [1] 1922.761
```

Below is the gender distribution of the participants.

```
##Gender distribution of participants
gender <- mainDF$RIAGENDR
maleParticipants <- subset(gender,gender=="1")
femaleParticipants <- subset(gender,gender=="2")
length(gender)
```

```
## [1] 20194
```

```
length(maleParticipants)
```

```
## [1] 9423
```

```
length(femaleParticipants)
```

```
## [1] 10771
```

```
mPercent = (length(maleParticipants)/length(gender))*100 ## % Male
fPercent = (length(femaleParticipants)/length(gender))*100 ## % Female

print("Male % - "); print(mPercent)
```

```
## [1] "Male % - "
```

```
## [1] 46.66237
```

```
print("Female % - "); print(fPercent)
```

```
## [1] "Female % - "
```

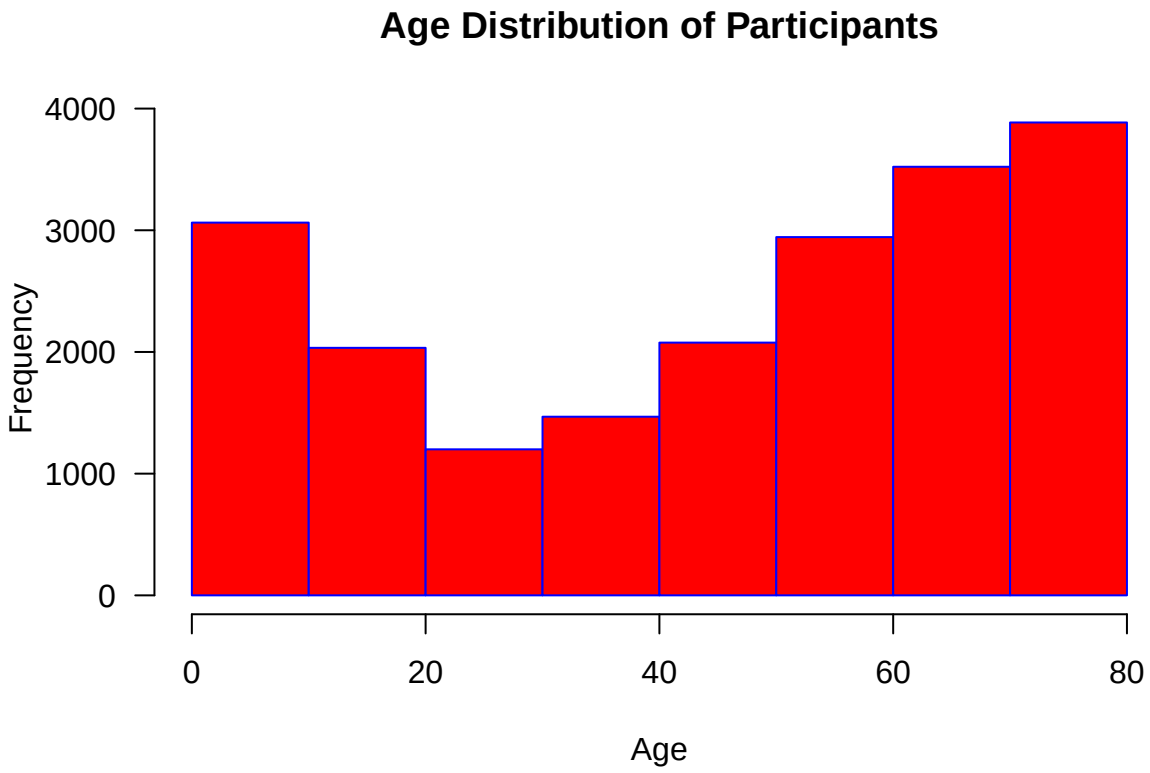
```
## [1] 53.33763
```

A histogram of the age distribution is shown below.

```
partAge <- mainDF$RIDAGEYR
mean(partAge) ##mean Age of participants
```

```
## [1] 45.37333
```

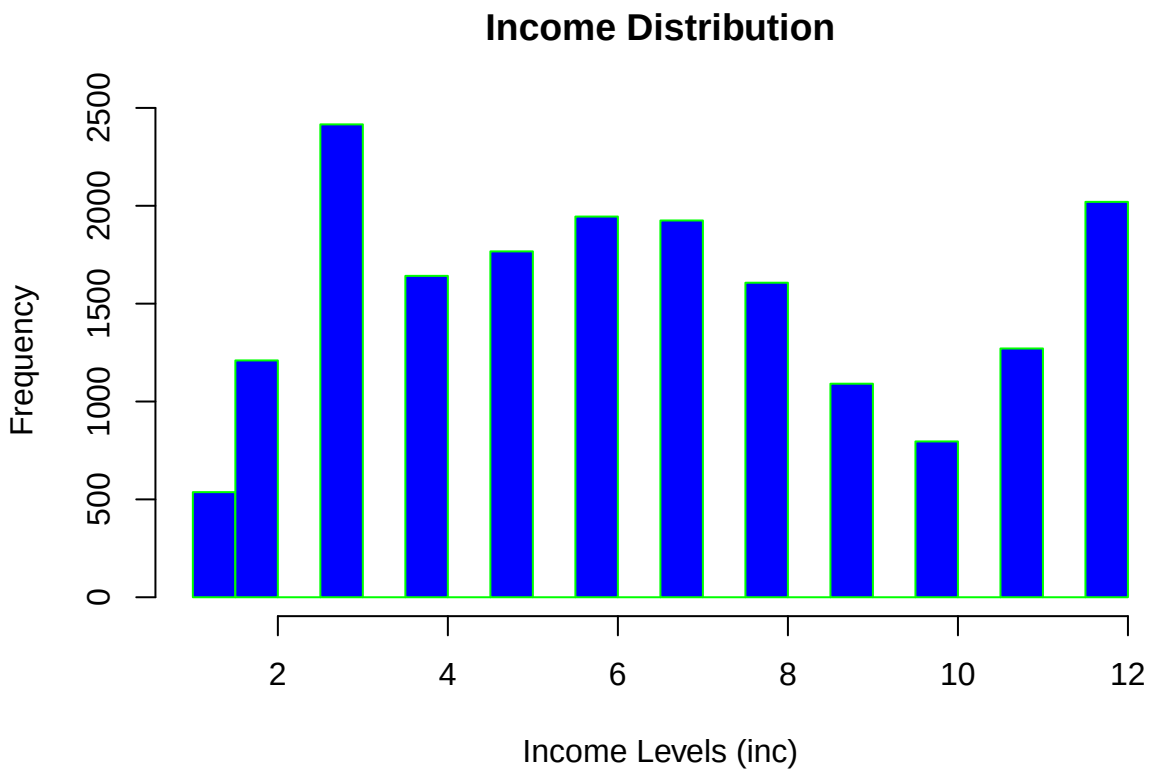
```
hist(partAge,main="Age Distribution of Participants", xlab="Age",
     border="blue", col="red", las=1, breaks=8)
```



Now, we try and correlate annual family income to drug (Marijuana/Hashish) use.

We begin by plotting the annual family income distribution:

```
famInc <- mainDF$IND235
bad4 <- is.na(famInc)
repFamInc <- famInc[!bad4]
repFamInc2 <- subset(repFamInc,repFamInc<=12) ##excluding missing, refused to report
sortInc<-sort(repFamInc2)
hist(sortInc,main="Income Distribution",xlab="Income Levels (inc)",col="blue",border = "green")
```

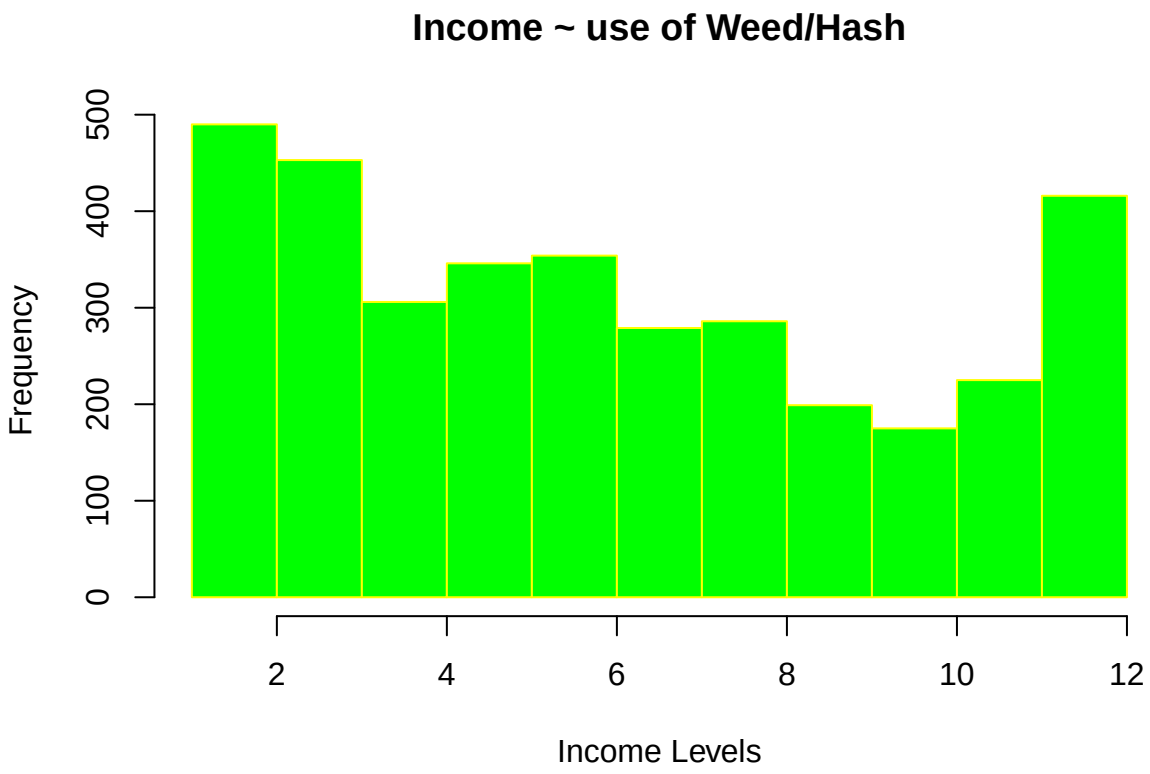



```
incLevel <- list("0-399","400-799","800-1249","1250-1649",
                 "1650-2099","2100-2899","2900-3749","3750-4599",
                 "4600-5399","5400-6249","6250-8399","8400+")  ## Family Income Level in $ per month
lev<-c(1:12)
cbind(lev,incLevel)
```

```
##       lev incLevel
##  [1,] 1   "0-399"
##  [2,] 2   "400-799"
##  [3,] 3   "800-1249"
##  [4,] 4   "1250-1649"
##  [5,] 5   "1650-2099"
##  [6,] 6   "2100-2899"
##  [7,] 7   "2900-3749"
##  [8,] 8   "3750-4599"
##  [9,] 9   "4600-5399"
## [10,] 10  "5400-6249"
## [11,] 11  "6250-8399"
## [12,] 12  "8400+"
```

The plot below shows the comparison:

```
drugInc <- subset(mainDF,mainDF$DUQ200==1) ## Represents those who've tried Marijuana or Hash
famDrugInc <- drugInc$IND235
repFamDrugInc <-subset(famDrugInc,famDrugInc<=12)
repFamDrugInc<-sort(repFamDrugInc)
hist(repFamDrugInc,main="Income ~ use of Weed/Hash",xlab="Income Levels",col="green",border="yellow")
```



So, this report has tried to look at some insights based on various health factors of the volunteers in the survey.