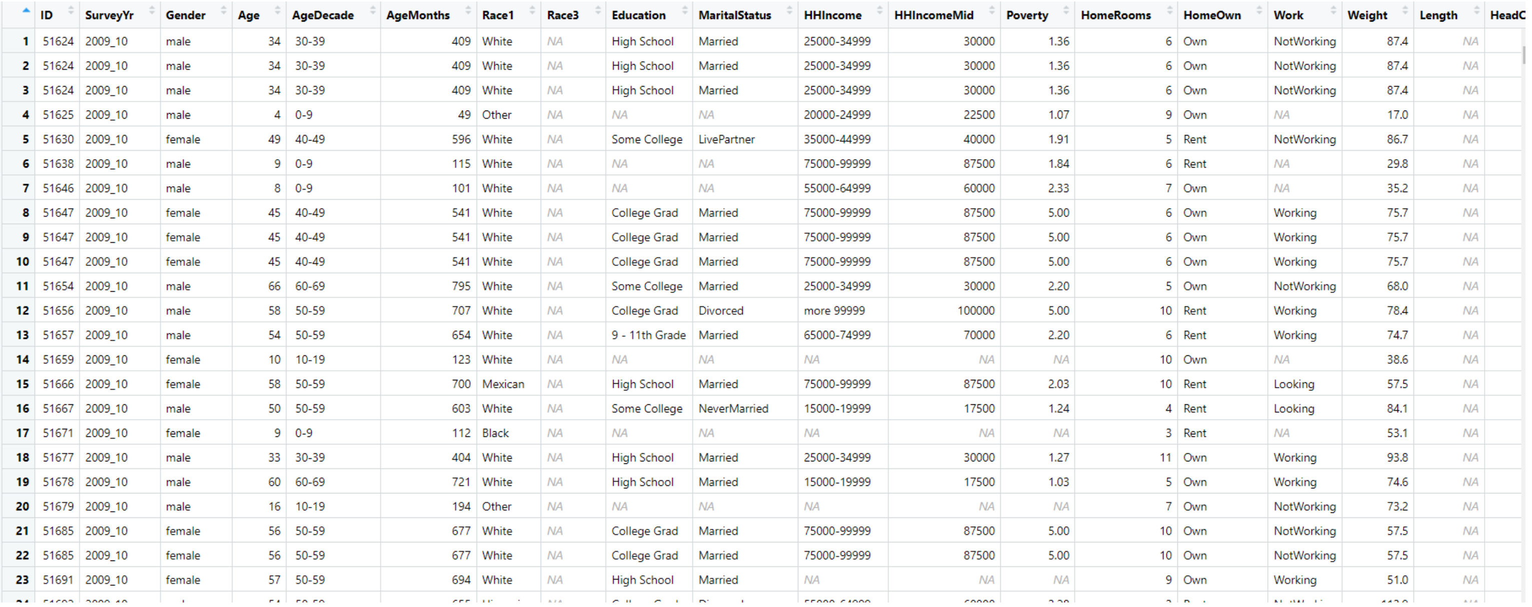
Name: **Varun Sudhir**

Reg No: **21BDS0040**

Exploratory Data Analysis Lab Assignment – 5

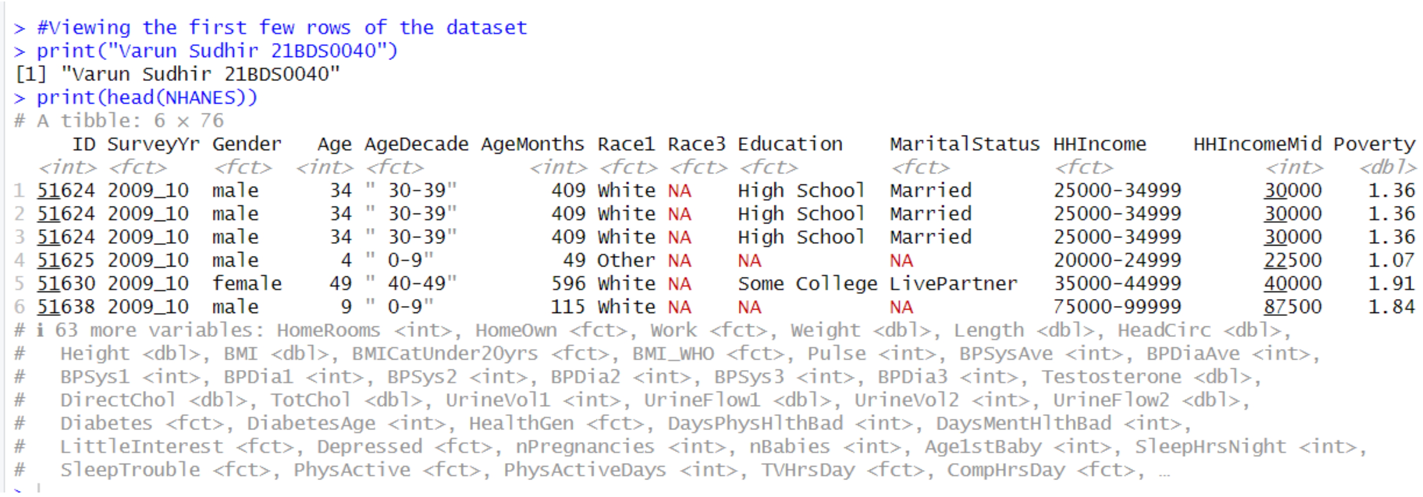
For this experiment,we will utilize the NHANES dataset. The NHANES (National Health and Nutrition Examination Survey) dataset is a large, real-world dataset that comes from a program of studies conducted by the National Center for Health Statistics (NCHS). The NHANES dataset is often used in health and epidemiological studies and contains various missing values across its columns.

```
# Varun Sudhir 21BDS0040

# Install and load the NHANES package
install.packages("NHANES")
library(NHANES)

# Check for missing values
View(NHANES)
summary(NHANES)
```



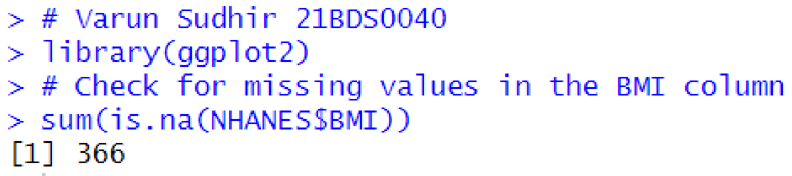
	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1	Race3	Education	MaritalStatus	HHIncome	HHIncomeMid	Poverty	HomeRooms	HomeOwn	Work	Weight	Length	HeadC
1	51624	2009_10	male	34	30-39	409	White	NA	High School	Married	25000-34999	30000	1.36	6	Own	NotWorking	87.4	NA	
2	51624	2009_10	male	34	30-39	409	White	NA	High School	Married	25000-34999	30000	1.36	6	Own	NotWorking	87.4	NA	
3	51624	2009_10	male	34	30-39	409	White	NA	High School	Married	25000-34999	30000	1.36	6	Own	NotWorking	87.4	NA	
4	51625	2009_10	male	4	0-9	49	Other	NA	NA	NA	20000-24999	22500	1.07	9	Own	NA	17.0	NA	
5	51630	2009_10	female	49	40-49	596	White	NA	Some College	LivePartner	35000-44999	40000	1.91	5	Rent	NotWorking	86.7	NA	
6	51638	2009_10	male	9	0-9	115	White	NA	NA	NA	75000-99999	87500	1.84	6	Rent	NA	29.8	NA	
7	51646	2009_10	male	8	0-9	101	White	NA	NA	NA	55000-64999	60000	2.33	7	Own	NA	35.2	NA	
8	51647	2009_10	female	45	40-49	541	White	NA	College Grad	Married	75000-99999	87500	5.00	6	Own	Working	75.7	NA	
9	51647	2009_10	female	45	40-49	541	White	NA	College Grad	Married	75000-99999	87500	5.00	6	Own	Working	75.7	NA	
10	51647	2009_10	female	45	40-49	541	White	NA	College Grad	Married	75000-99999	87500	5.00	6	Own	Working	75.7	NA	
11	51654	2009_10	male	66	60-69	795	White	NA	Some College	Married	25000-34999	30000	2.20	5	Own	NotWorking	68.0	NA	
12	51656	2009_10	male	58	50-59	707	White	NA	College Grad	Divorced	more 99999	100000	5.00	10	Rent	Working	78.4	NA	
13	51657	2009_10	male	54	50-59	654	White	NA	9 - 11th Grade	Married	65000-74999	70000	2.20	6	Rent	Working	74.7	NA	
14	51659	2009_10	female	10	10-19	123	White	NA	NA	NA	NA	NA	NA	10	Own	NA	38.6	NA	
15	51666	2009_10	female	58	50-59	700	Mexican	NA	High School	Married	75000-99999	87500	2.03	10	Rent	Looking	57.5	NA	
16	51667	2009_10	male	50	50-59	603	White	NA	Some College	NeverMarried	15000-19999	17500	1.24	4	Rent	Looking	84.1	NA	
17	51671	2009_10	female	9	0-9	112	Black	NA	NA	NA	NA	NA	NA	3	Rent	NA	53.1	NA	
18	51677	2009_10	male	33	30-39	404	White	NA	High School	Married	25000-34999	30000	1.27	11	Own	Working	93.8	NA	
19	51678	2009_10	male	60	60-69	721	White	NA	High School	Married	15000-19999	17500	1.03	5	Own	Working	74.6	NA	
20	51679	2009_10	male	16	10-19	194	Other	NA	NA	NA	NA	NA	NA	7	Own	NotWorking	73.2	NA	
21	51685	2009_10	female	56	50-59	677	White	NA	College Grad	Married	75000-99999	87500	5.00	10	Own	NotWorking	57.5	NA	
22	51685	2009_10	female	56	50-59	677	White	NA	College Grad	Married	75000-99999	87500	5.00	10	Own	NotWorking	57.5	NA	
23	51691	2009_10	female	57	50-59	694	White	NA	High School	Married	NA	NA	NA	9	Own	Working	51.0	NA	

```
#Viewing the first few rows of the dataset

print("Varun Sudhir 21BDS0040")
print(head(NHANES))
```



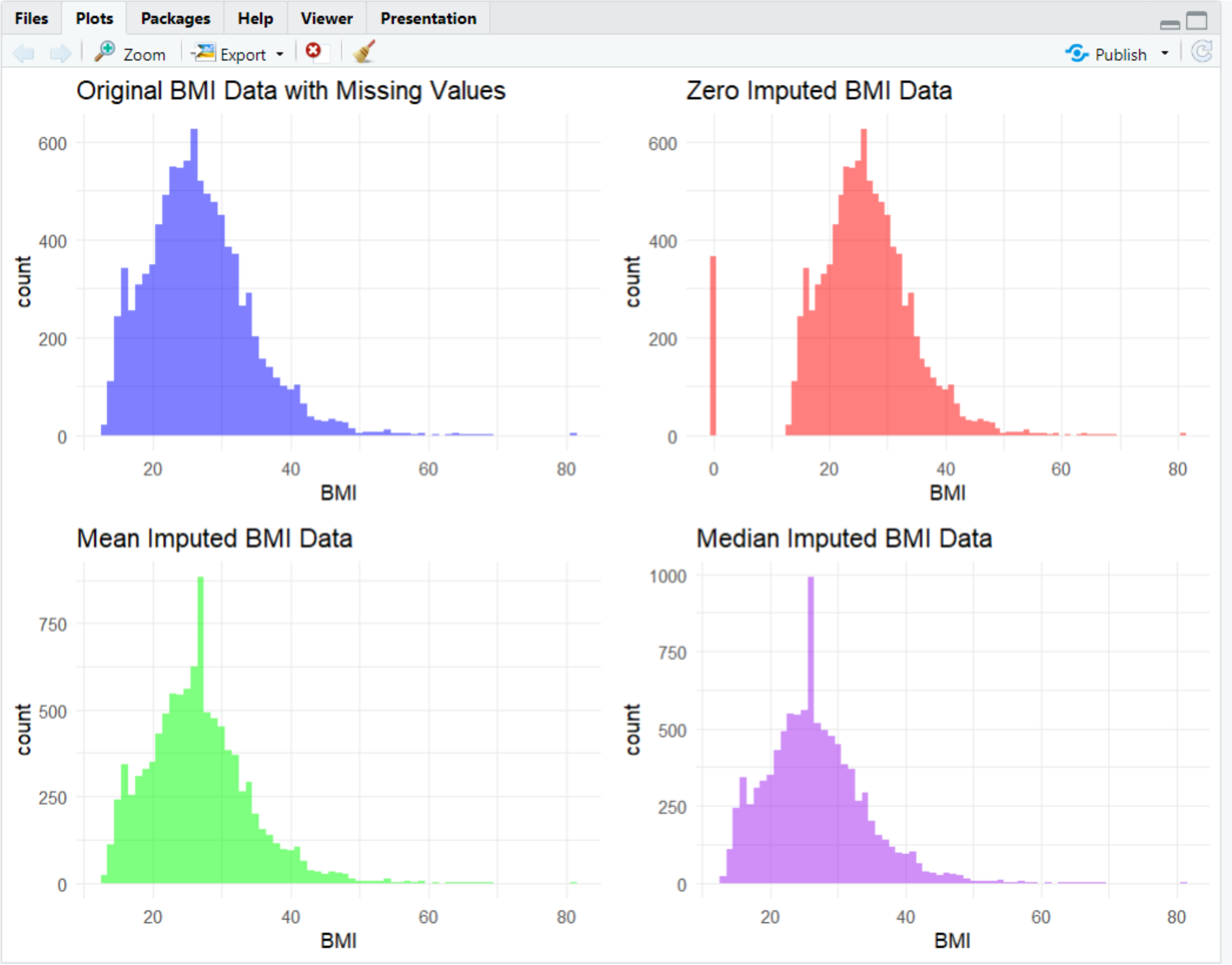
```
> #Viewing the first few rows of the dataset
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
> print(head(NHANES))
# A tibble: 6 × 76
     ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education    MaritalStatus HHIncome    HHIncomeMid Poverty
  <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>        <fct>         <fct>             <int>   <dbl>
1 51624 2009_10  male      34 " 30-39"        409 White NA    High School  Married       25000-34999       30000    1.36
2 51624 2009_10  male      34 " 30-39"        409 White NA    High School  Married       25000-34999       30000    1.36
3 51624 2009_10  male      34 " 30-39"        409 White NA    High School  Married       25000-34999       30000    1.36
4 51625 2009_10  male       4 " 0-9"           49 Other NA    NA           NA            20000-24999       22500    1.07
5 51630 2009_10  female    49 " 40-49"        596 White NA    Some College LivePartner   35000-44999       40000    1.91
6 51638 2009_10  male       9 " 0-9"          115 White NA    NA           NA            75000-99999       87500    1.84
# ℹ 63 more variables: HomeRooms <int>, HomeOwn <fct>, Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>,
#   Height <dbl>, BMI <dbl>, BMICatUnder20yrs <fct>, BMI_WHO <fct>, Pulse <int>, BPSysAve <int>, BPDiaAve <int>,
#   BPSys1 <int>, BPDia1 <int>, BPSys2 <int>, BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,
#   DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>, UrineVol2 <int>, UrineFlow2 <dbl>,
#   Diabetes <fct>, DiabetesAge <int>, HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,
#   LittleInterest <fct>, Depressed <fct>, nPregnancies <int>, nBabies <int>, Age1stBaby <int>, SleepHrsNight <int>,
#   SleepTrouble <fct>, PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>, …
```

```
# Varun Sudhir 21BDS0040

librar(ggplot2)

# Check for missing values in the BMI column
sum(is.na(NHANES$BMIy))
```

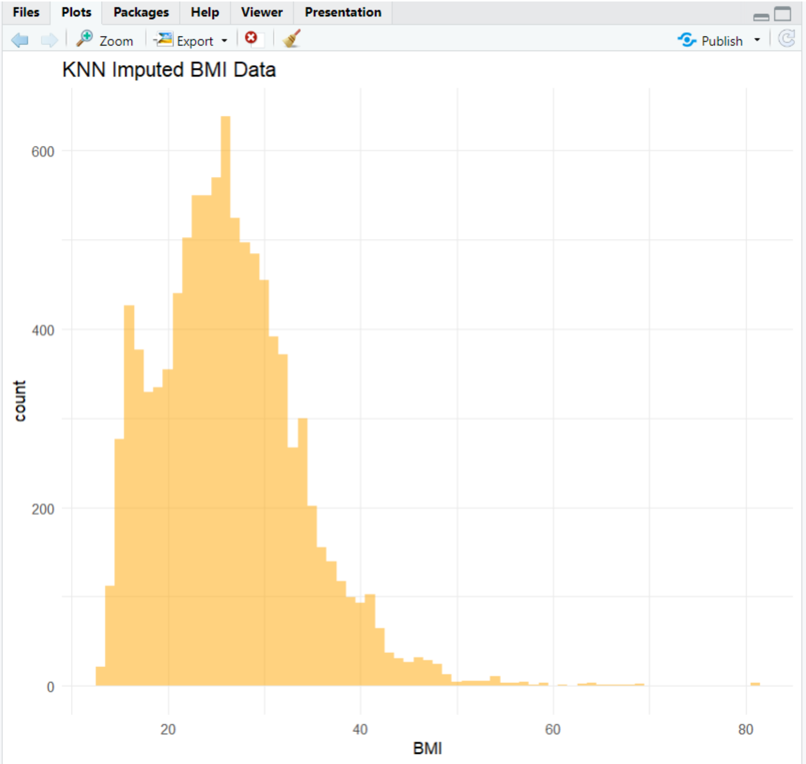


```
> # Varun Sudhir 21BDS0040
> library(ggplot2)
> # Check for missing values in the BMI column
> sum(is.na(NHANES$BMI))
[1] 366
```

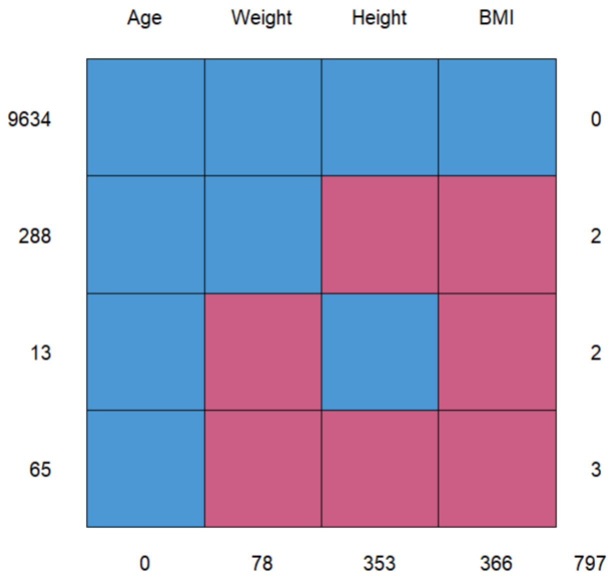
```
# Varun Sudhir 21BDS0040

# Zero imputation for BMI
BMI_zero_imputed <- NHANES
BMI_zero_imputed$BMI[is.na(BMI_zero_imputed$BMI)] <- 0

# Mean imputation for BMI
BMI_mean_imputed <- NHANES
BMI_mean_imputed$BMI[is.na(BMI_mean_imputed$BMI)] <-
mean(BMI_mean_imputed$BMI, na.rm = TRUE)

# Median imputation for BMI
BMI_median_imputed <- NHANES
BMI_median_imputed$BMI[is.na(BMI_median_imputed$BMI)] <-
median(BMI_median_imputed$BMI, na.rm = TRUE)
```



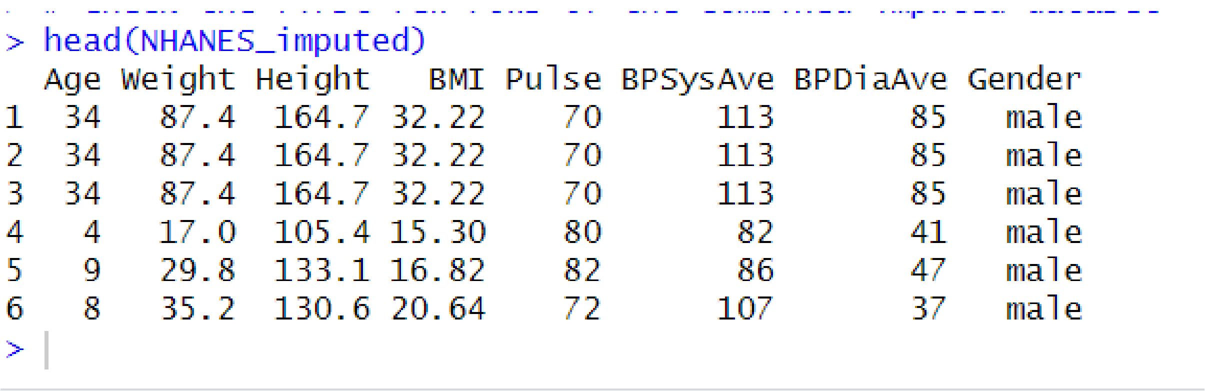
```r
# KNN Imputation
# Varun Sudhir 21BDS0040

library(VIM)

# Perform KNN imputation (k = 5) for BMI
BMI_knn_imputed <- NHANES
BMI_knn_imputed <- kNN(BMI_knn_imputed, variable = "BMI", k = 5)

# Plot the KNN-imputed BMI data
p_knn <- ggplot(BMI_knn_imputed, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = "orange", alpha = 0.5) +
  ggtitle("KNN Imputed BMI Data") +
  theme_minimal()

# Print the plot
p_knn
```



```
# Varun Sudhir 21BDS0040

# Varun Sudhir 21BDS0040

# Subsetting only numeric columns for imputation
NHANES_numeric <- NHANES %>%
  select(Age, Weight, Height, BMI)

# Check the missing data pattern
md.pattern(NHANES_numeric)

# Perform multiple imputation using the 'mice' package
# Default method: pmm (Predictive Mean Matching)
mice_imputation <- mice(NHANES_numeric, m = 5, method = 'pmm', seed = 123)

# Check the summary of imputed data
summary(mice_imputation)

# Get the completed dataset after imputation
imputed_data <- complete(mice_imputation, 1)

# Print the first few rows of the imputed data
head(imputed_data)
```



```r
# Varun Sudhir 21BDS0040
# Class-based imputation

# Subset numeric columns and include 'Gender' for class-based imputation
NHANES_class_based <- NHANES %>%
  select(Age, Weight, Height, BMI, Pulse, BPSysAve, BPDiaAve, Gender)

# Split the data by Gender
NHANES_male <- NHANES_class_based %>% filter(Gender == "male")
NHANES_female <- NHANES_class_based %>% filter(Gender == "female")

# Perform multiple imputation for males
mice_male <- mice(NHANES_male %>% select(-Gender), m = 5, method = 'pmm', seed
= 123)
NHANES_male_imputed <- complete(mice_male, 1)
NHANES_male_imputed$Gender <- "male"

# Perform multiple imputation for females
mice_female <- mice(NHANES_female %>% select(-Gender), m = 5, method = 'pmm',
seed = 123)
NHANES_female_imputed <- complete(mice_female, 1)
NHANES_female_imputed$Gender <- "female"

# Combine the imputed datasets for males and females
NHANES_imputed <- bind_rows(NHANES_male_imputed, NHANES_female_imputed)

# Check the first few rows of the combined imputed dataset
head(NHANES_imputed)
```



```
> head(NHANES_imputed)
  Age Weight Height   BMI Pulse BPSysAve BPDiaAve Gender
1  34   87.4  164.7 32.22    70      113       85   male
2  34   87.4  164.7 32.22    70      113       85   male
3  34   87.4  164.7 32.22    70      113       85   male
4   4   17.0  105.4 15.30    80       82       41   male
5   9   29.8  133.1 16.82    82       86       47   male
6   8   35.2  130.6 20.64    72      107       37   male
> 
```