Data Mining
Home work 04
Descriptive Statistics

Aqeel Labash
Supervisor: Jaak Vilo

17 February 2016

First Question

The things that I took home is :) :

1. Each data type has it's own
2. When we want to compare data to each other it's always better for the plot of the groups to be more scalable for people eyes.
3. If we are using line width to show information , there should be a fixed max width.
4. When using color for categories, the max number we should use is 6-12 color , otherwise the reader will get lost between the color's.
5. It's better to use separable channels for different abstract dimensions.
6. There is no bad or good way to show data, but it should match our goal.
7. Using 3d visualization should be accompanied with extra caution.Because what we think it's more clear may delay the information delivery.
8. In 3d visualization the depth take away ability of comparison because of perspective distortion.
9. When we have complex changes it's better to use series of visualization rather than animation because we lose track of global changes and focus on local changes.
10. Validate against the right threat.

Second Question

While trying to plot I found an outliers or damaged data (height , weight less than zero) there was about 3472 row. I cleaned them using the following code:

```
#### Second Question ######
ncmp = read.csv('ncmp_1415_final_non_disclosive.csv',header = TRUE)
nrow(ncmp[ncmp$height <0,])
ncmp <- ncmp[ncmp$height >0,]
```

Note: The data later changed on the homework page with clean data, anyway I sticked to this data since I already started with it. In the following figure I plotted the requested combination (age,height,bmi,age) while showing the gender in colors.

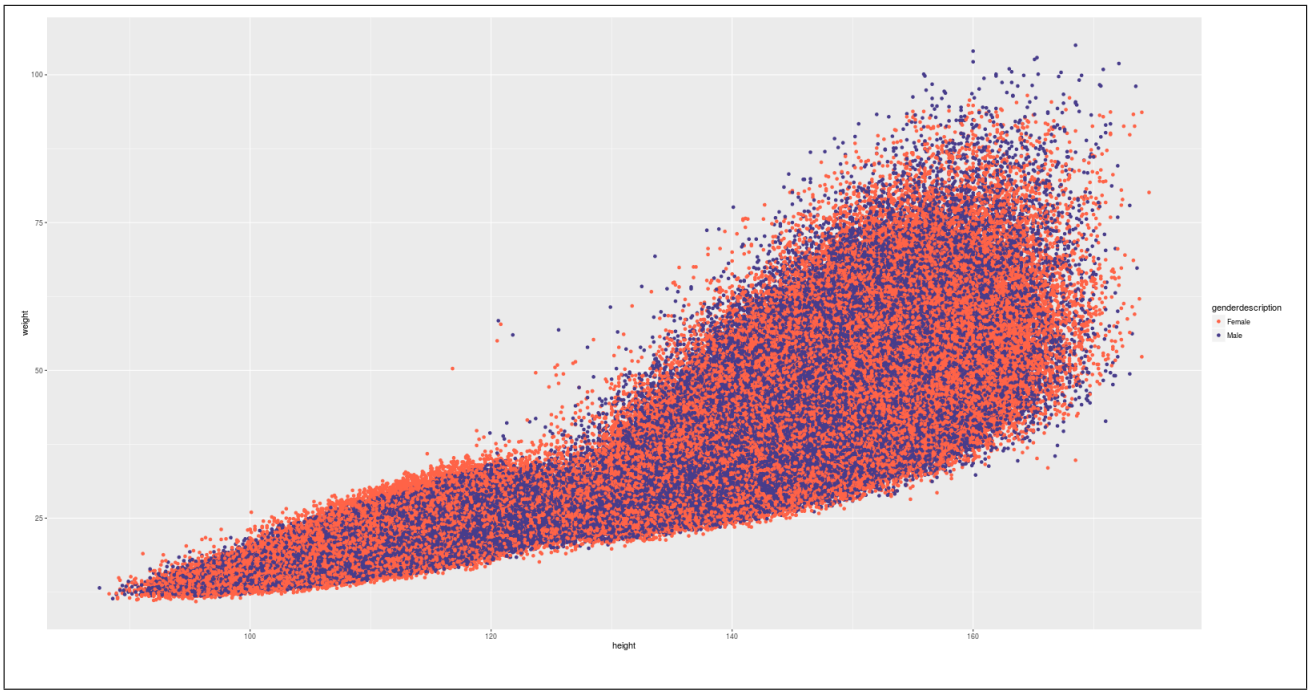

Figure 1: Height vs weight

In the previous plot we notice the increase of height lead to increase of weight as well.

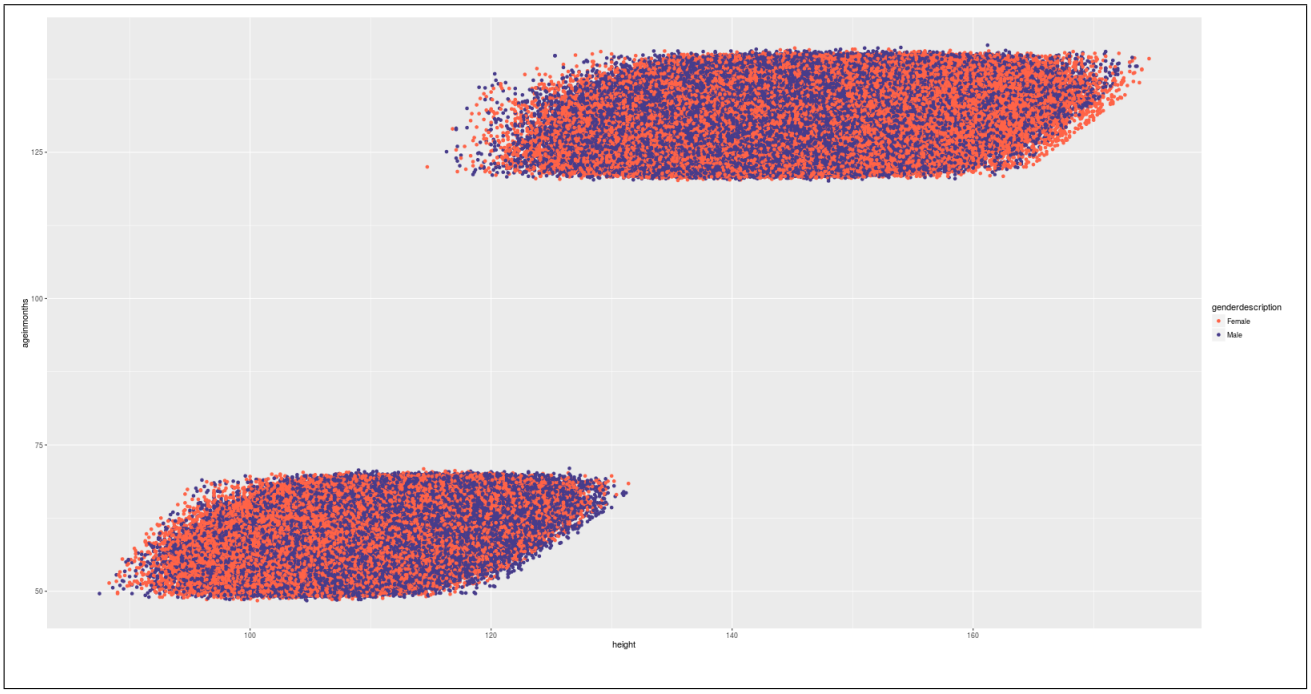

Figure 2: Height vs Age

In this figure we notice that there is a break in the data which cause this empty area.

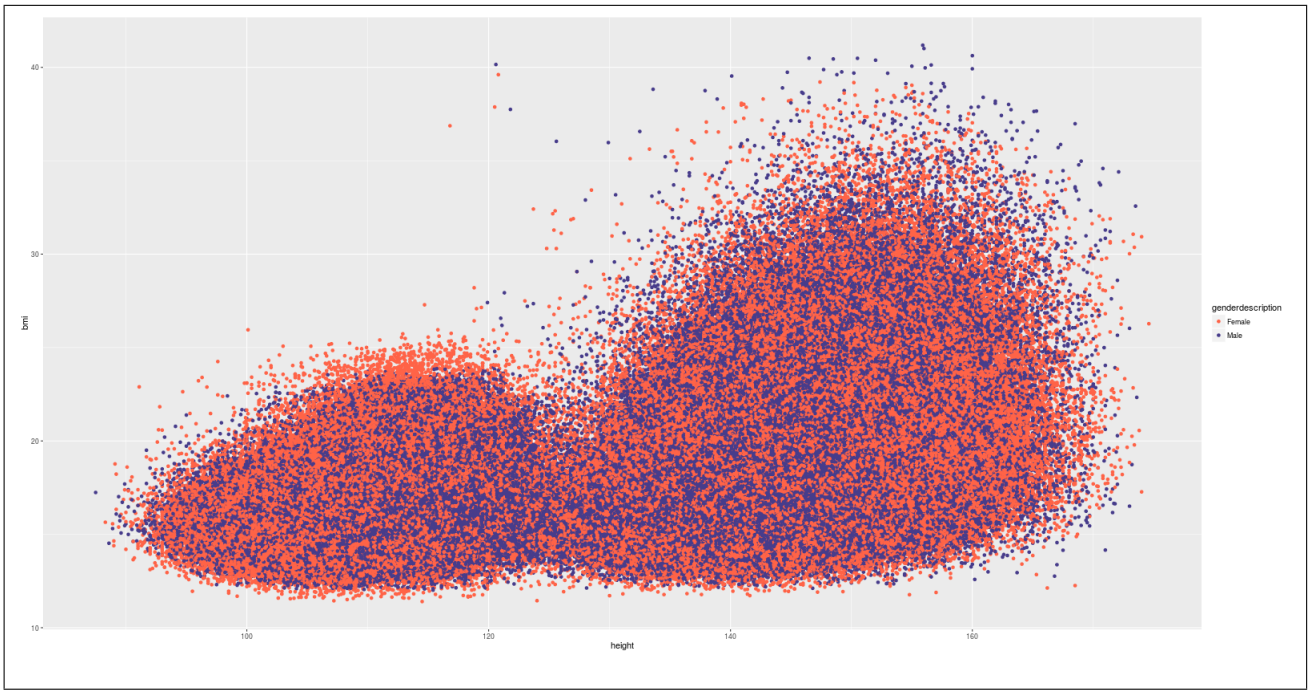

Figure 3: Height vs BMI

In the previous figure we notice the higher the higher BMI.

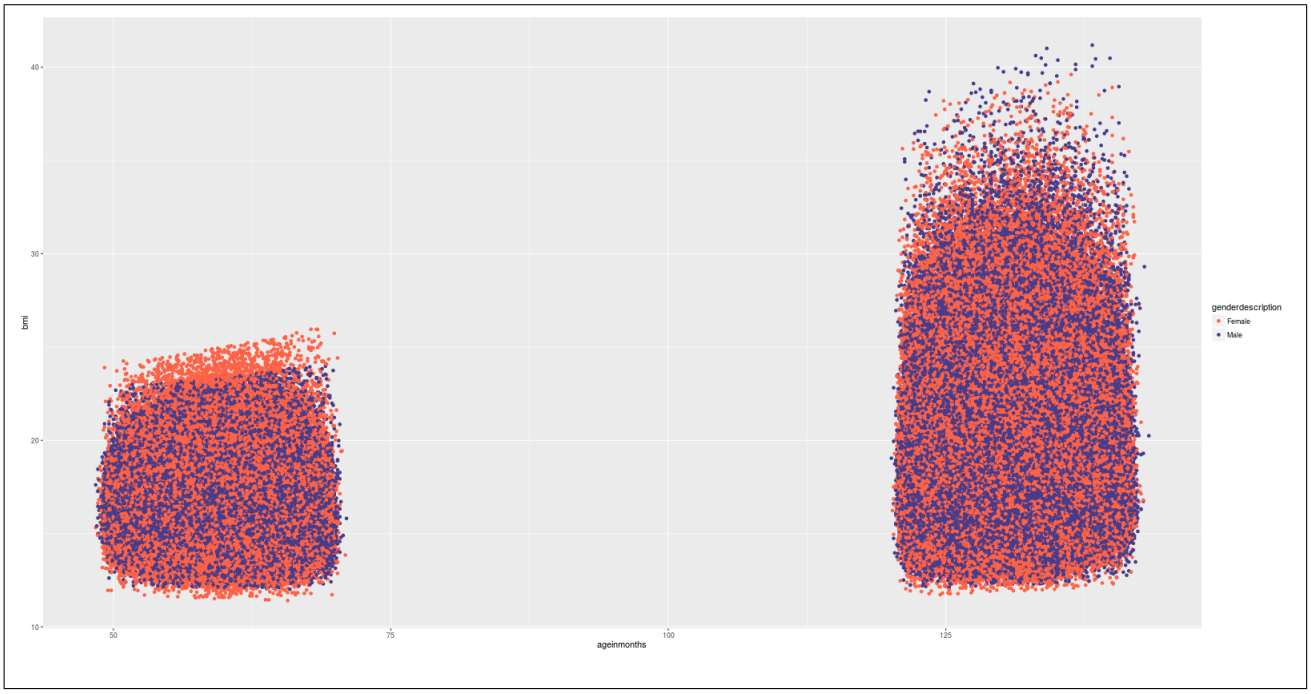

Figure 4: Age vs BMI

In Figure 4 we can see that we have two quantiles of ages. And the older the more BMI.

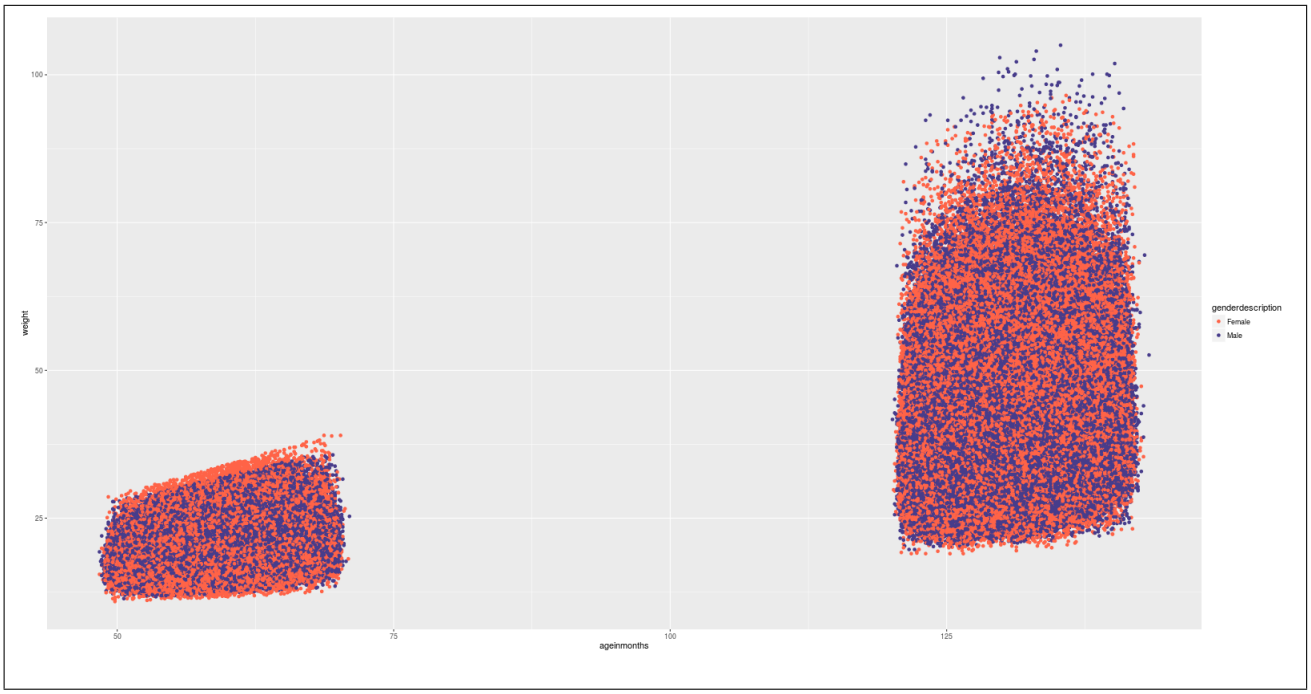

Figure 5: Age vs weight

It's expected here where the age increase , the weight increase as well.

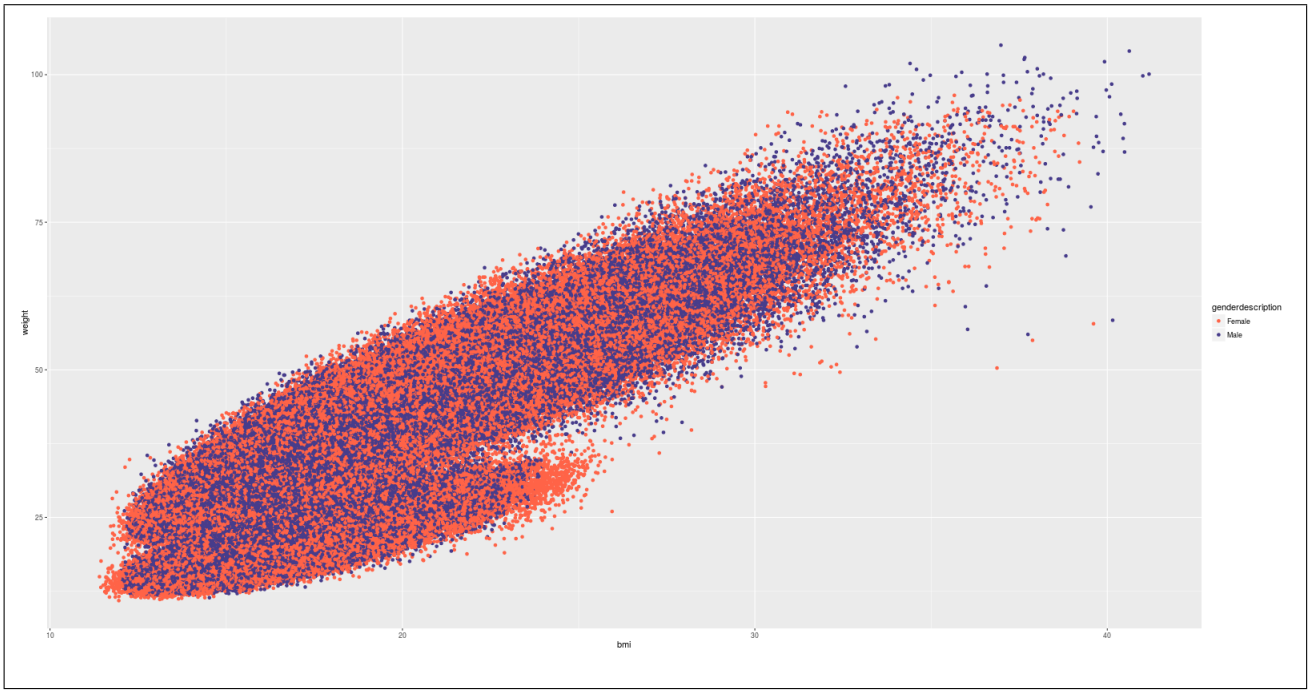

Figure 6: BMI vs weight

In figure 6 we can see that increasing in weight mean increase in BMI. $BMI = \frac{weight}{height^2}$ **Note:**Although the previous plots looks cool (for me at least :)) but I think there isn't much information we can get from. To plot the previous figures I used the following code:

```
library(ggplot2)
#Height Weight
png('heightweight.png',height = 800,width = 1600)
qplot(height,weight,colour =genderdescription,data = ncmp)+
scale_color_manual(values=c("tomato", "slateblue4"))
dev.off()
# Height Age
png('heightage',height = 800,width = 1600)
qplot(height,ageinmonths,colour =genderdescription,data = ncmp)+
scale_color_manual(values=c("tomato", "slateblue4"))
```

```
dev.off()
# Height BMI
png('heightBMI',height = 800,width = 1600)
qplot(height,bmi,colour =genderdescription,data = ncmp)+
scale_color_manual(values=c("tomato", "slateblue4"))
dev.off()
# age BMI
png('ageBMI',height = 800,width = 1600)
qplot(ageinmonths,bmi,colour =genderdescription,data = ncmp)+
scale_color_manual(values=c("tomato", "slateblue4"))
dev.off()

# age weight
png('ageweight',height = 800,width = 1600)
qplot(ageinmonths,weight,colour =genderdescription,data = ncmp)+
scale_color_manual(values=c("tomato", "slateblue4"))
dev.off()

# BMI weight
png('BMIweight',height = 800,width = 1600)
qplot(bmi,weight,colour =genderdescription,data = ncmp)+
scale_color_manual(values=c("tomato", "slateblue4"))
dev.off()
```

In the following figure we can see the caterogical bmi of the kids depending on there age,and BMI :

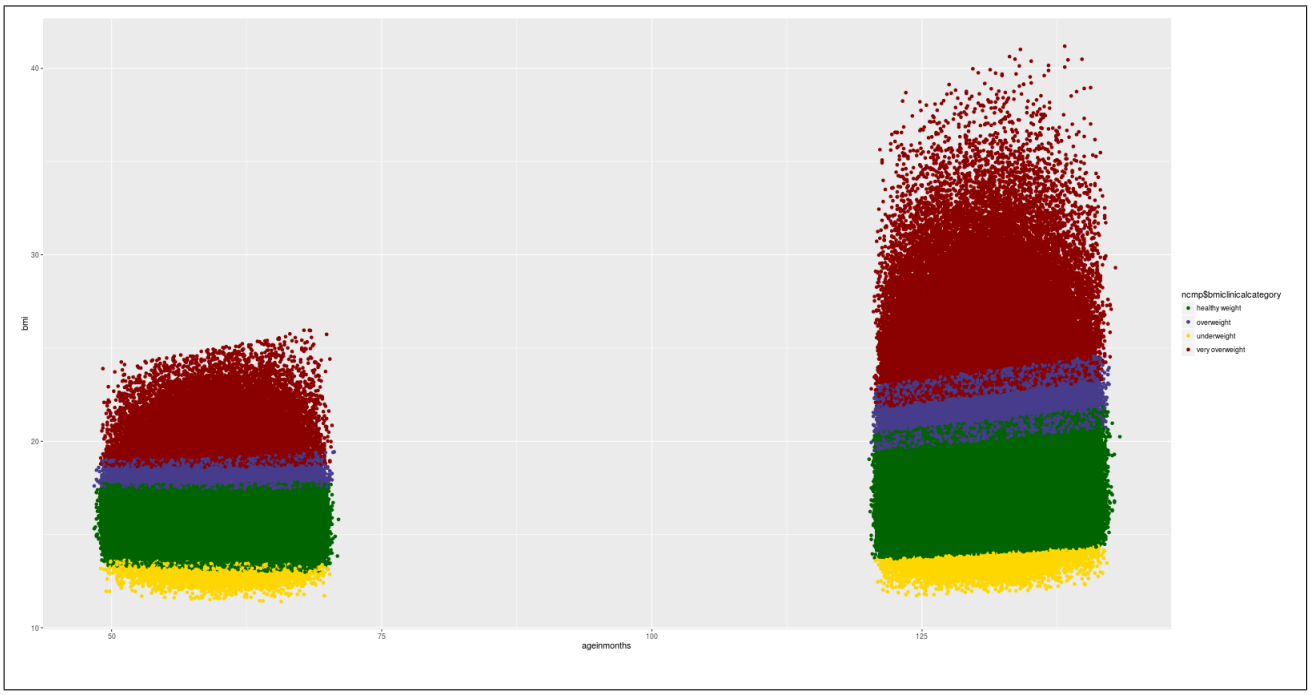

Figure 7: age vs BMI

And here is the code that generated the previous figure.

```
#bmi category with age
png('bmicatage',height = 800,width = 1600)
qplot(ageinmonths,bmi,colour =ncmp$bmiclinicalcategory,data = ncmp)+
scale_color_manual(values=c("darkgreen", "slateblue4","gold","darkred"))
dev.off()
```

Third Question

For this question I used Mosteller formula to calculate Body Surface Area $BSA = \frac{\sqrt{W \times H}}{60}$ In the following plots I colored them depending to BMI clinical category because it was continuous at some ranges.

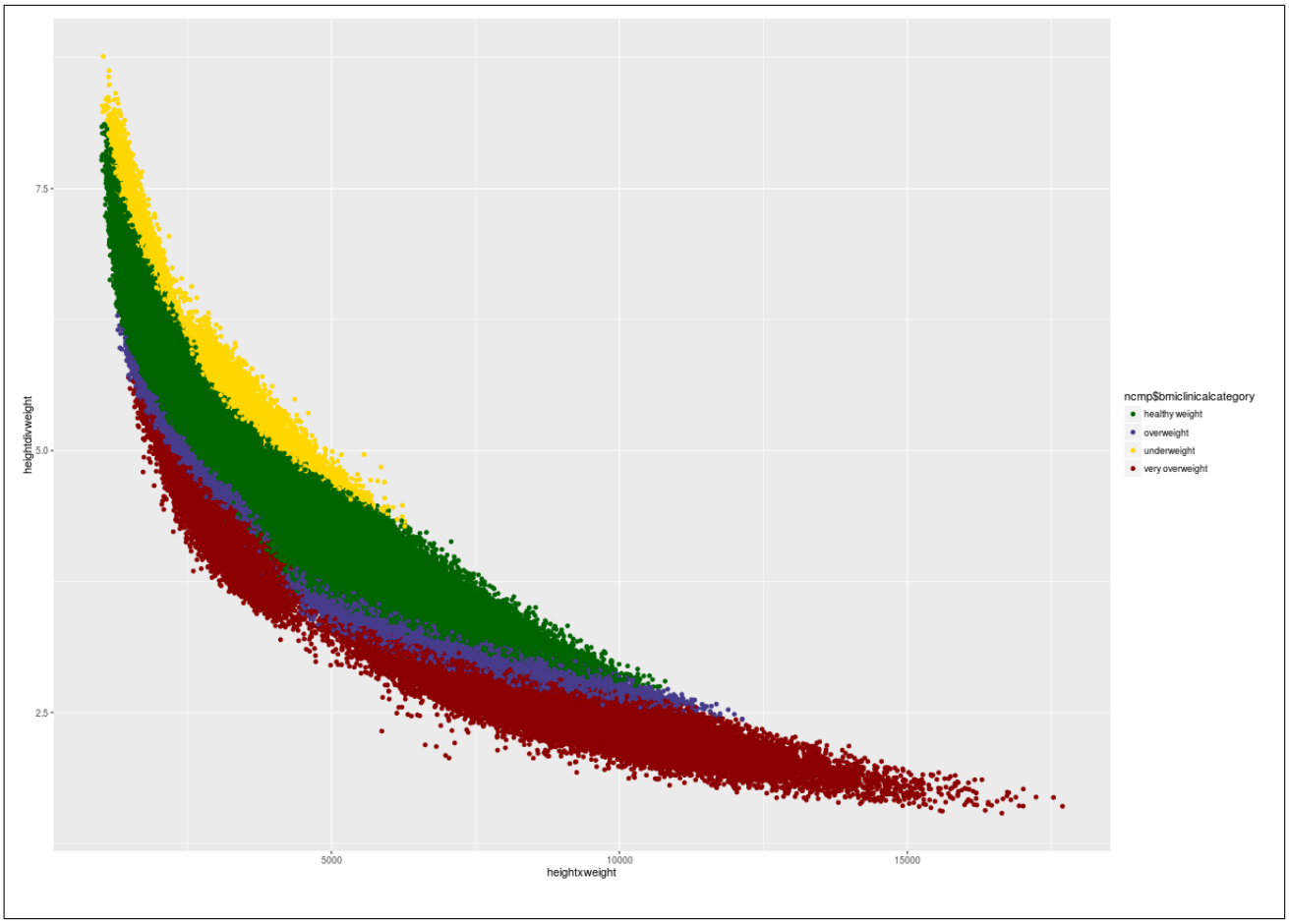

Figure 8: Multiplication vs Devision of height and weight

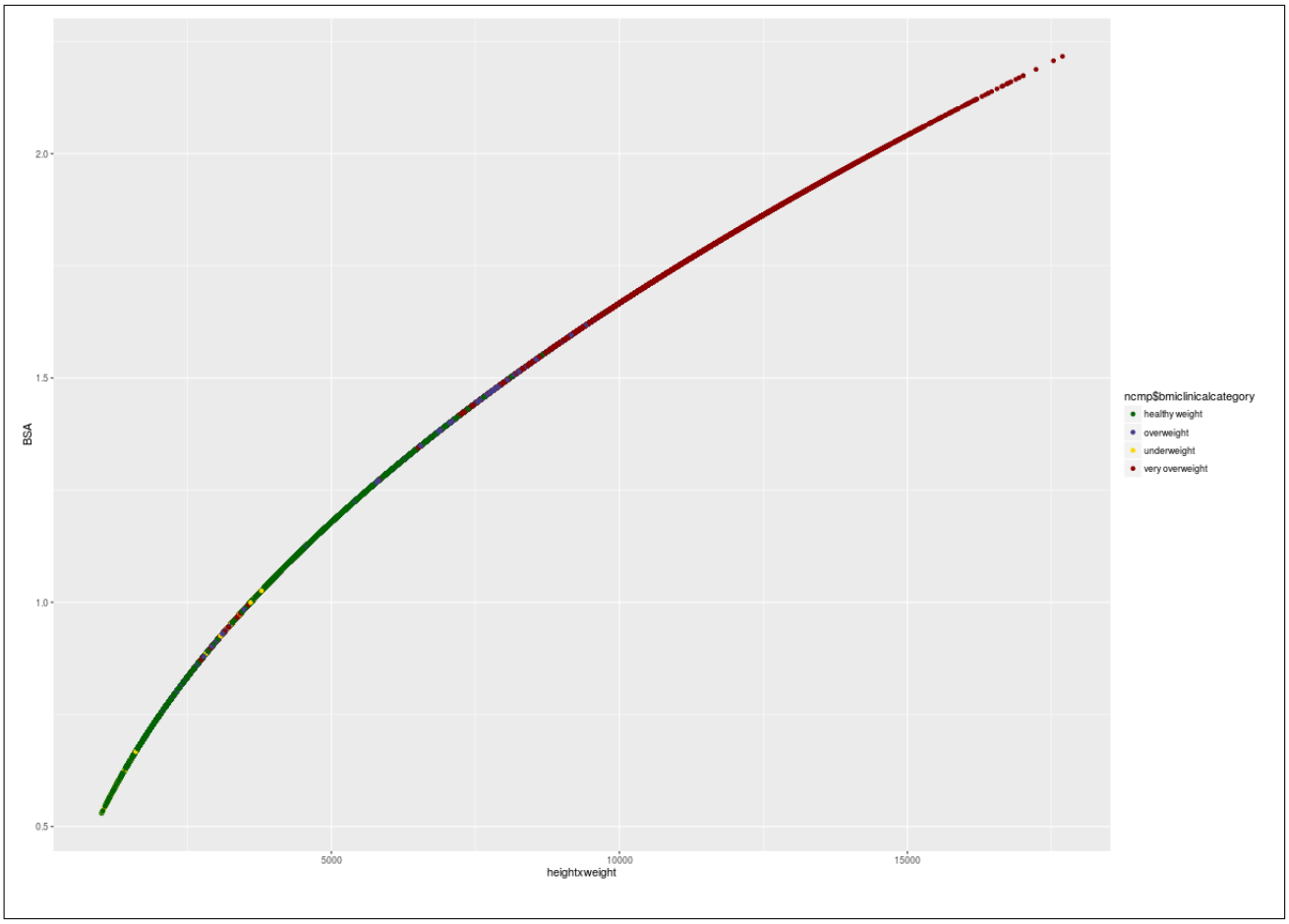

Figure 9: Multiplication vs BSA

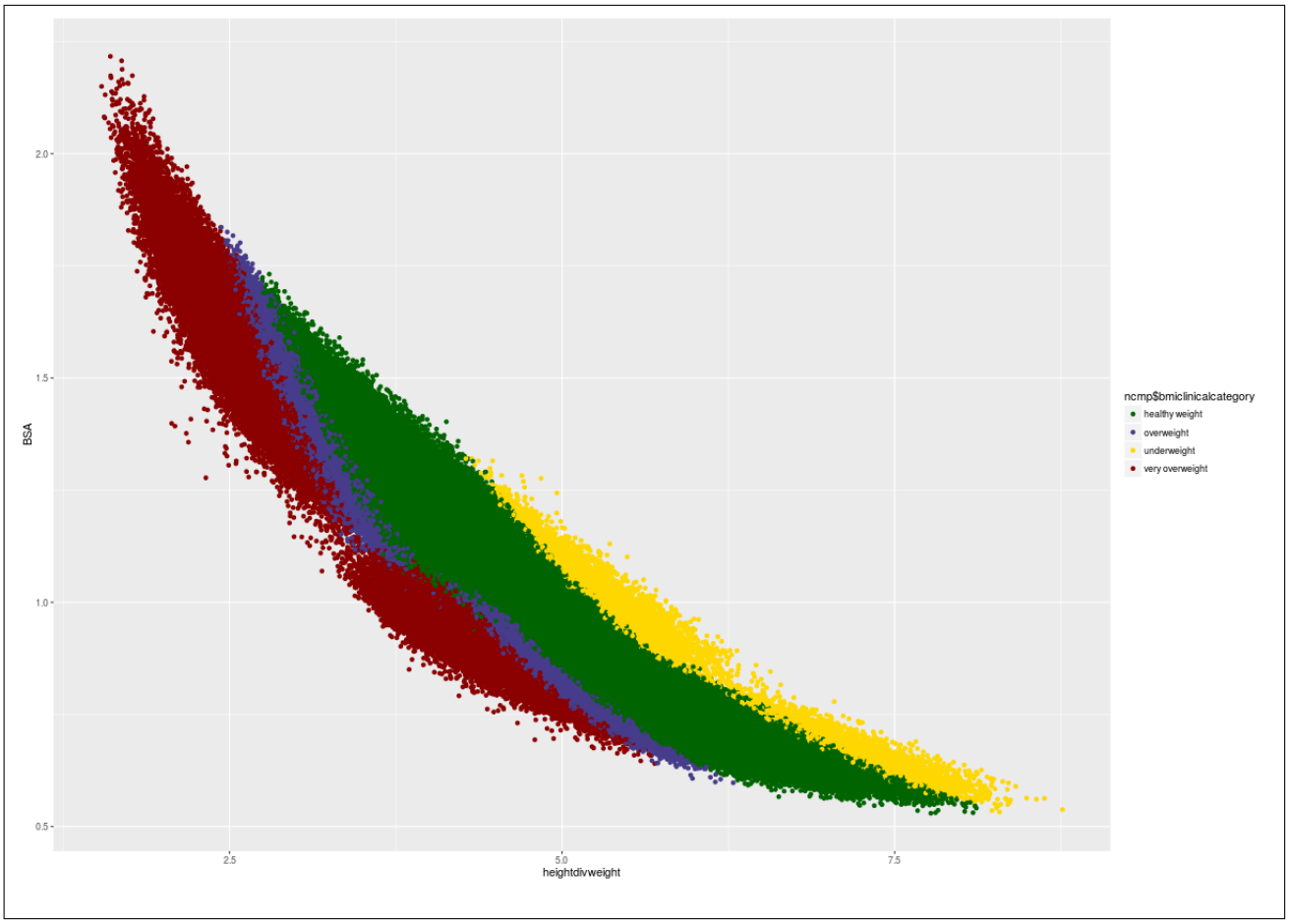

Figure 10: Devision vs BSA

Fourth Question

Fifth Question

Sixth Question