Homework 4

Kenigbolo Meya Stephen

March 2, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

1. Watch the video presentation by Tamara Munzner: Keynote on Visualization Principles, https://vimeo.com/26205288 and slides - http://www.cs.ubc.ca/~tmm/talks/vizbi11/vizbi11.pdf Summarize the key take-home messages from her presentation.

I believe the major aim of the presentation is to point out that the option to use visualization should depend on the data available to you. When we have different types of data it is easier to visualize relationships and patterns between them by manipulating with the color, shape of the display value etc in order to bring actionable insights to the surface. There is a lot of power in using spatial position to analyze data as in almost every case it works out fine. An important point she also made was concerning "data transformation". Transforming the data into a suitable abstraction is needed in data visualization in order to represent data correctly.

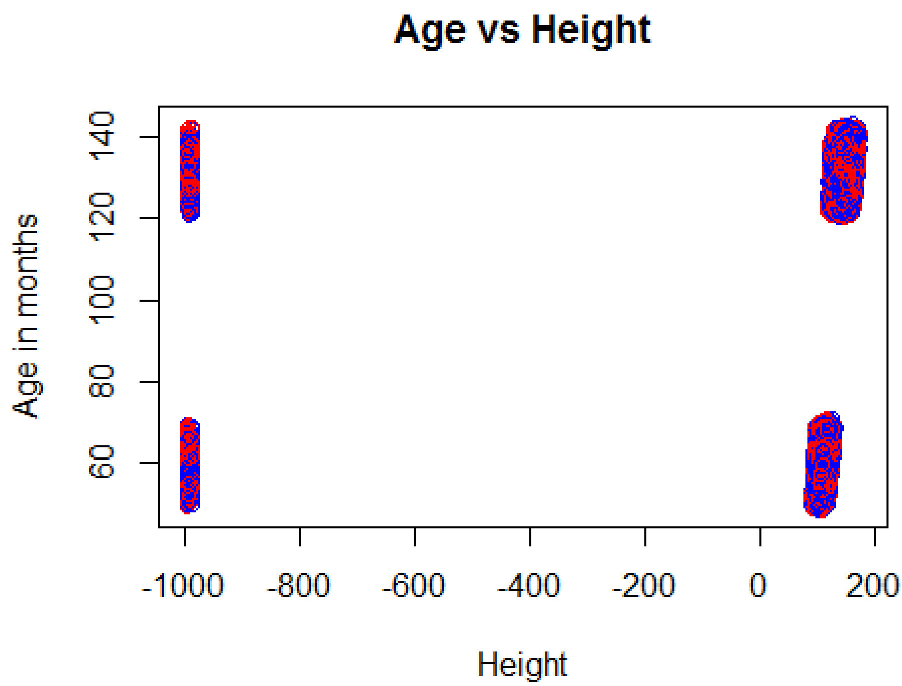
In respect to 3D data visualization, one has to be cautious in this approach because important things could be easily missed (whereas avoidable anyways) and a good example of this is the fact that 3D does distort perspective.Taking a look at text legibility shows the extent to which 3D might be dangerous however 3D visualization is quite good (and mostly adviceable) for true 3D spatial data. Usage of 3D could result to important values from the data to be missing due to an overriding of the value by other values. In majority of the situation, it is easier to draw conclusions from 2-dimensional data visualization as opposed to 3-dimensional.

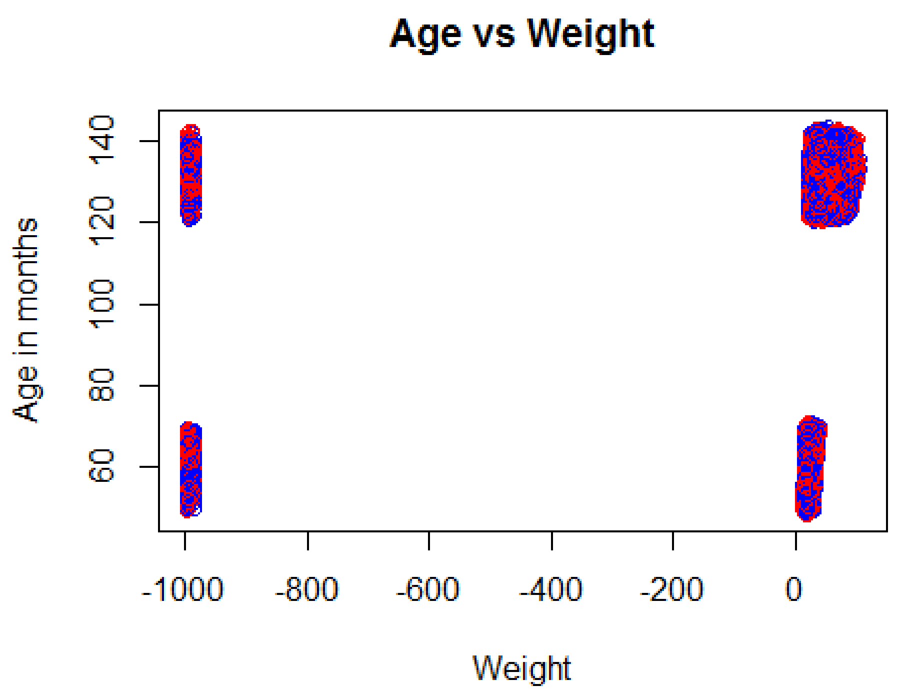
2. Fetch the UK child height/weight measurements data from here. (File).

Study a sufficiently large subset of measurements - plot dotplots comparing age, height, weight, BMI;

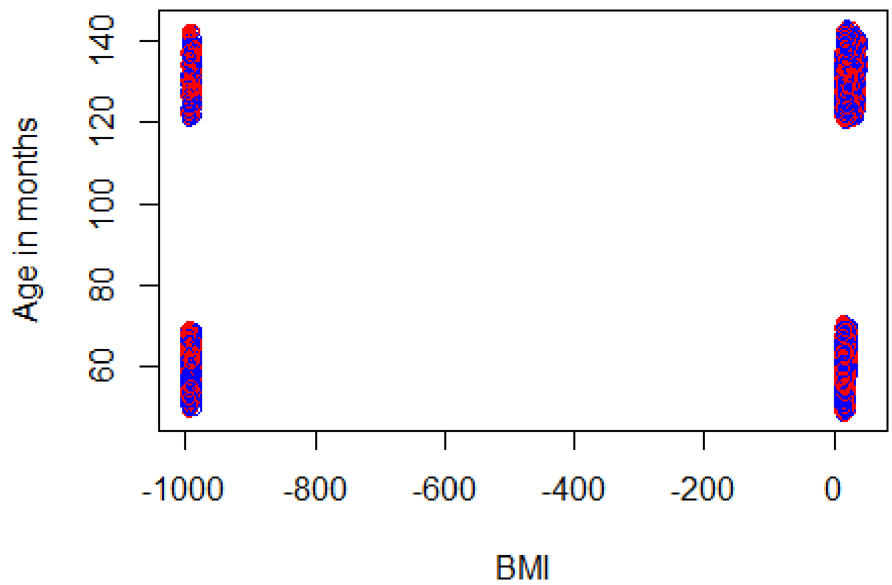
Experiment with color - highlighting gender, BMI clinical score, or age, by different markers, color or color scales.

(hint: you may want to develop ideas first with smaller subsets of data).

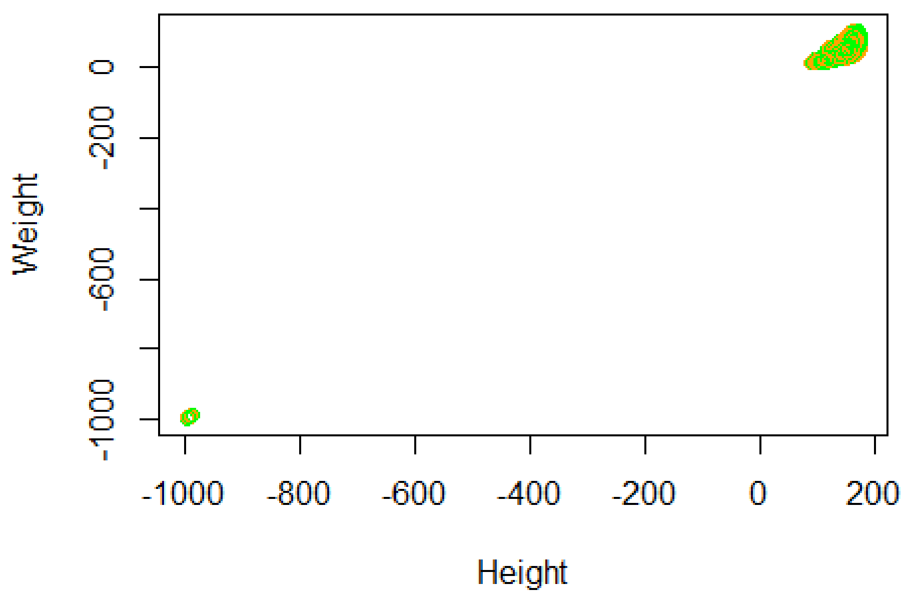

Age vs Height
Age in months
140
120
100
80
60
-1000
-800
-600
-400
-200
0
200
Height

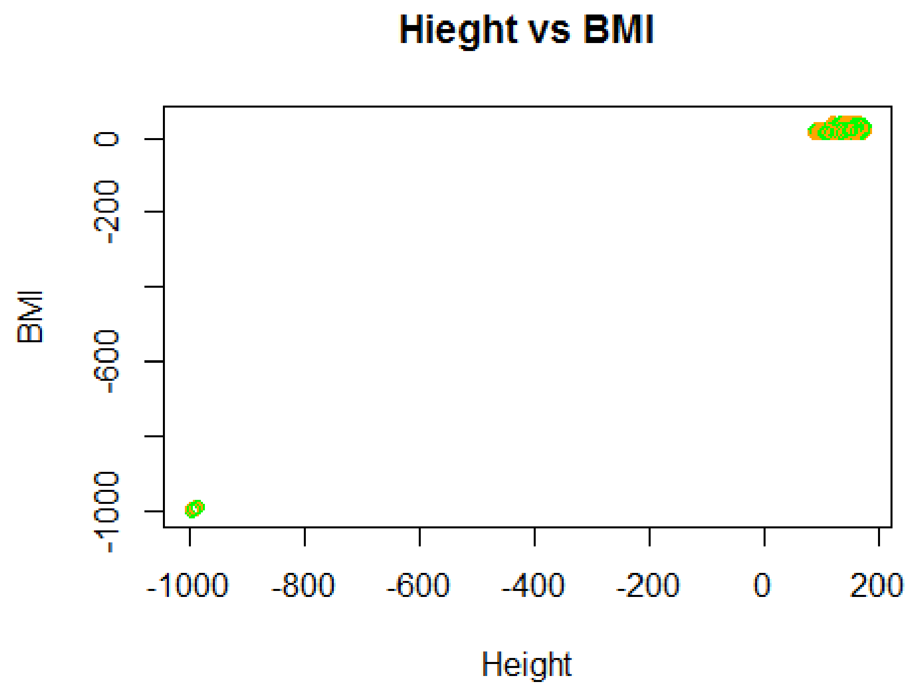

Age vs Weight
Age in months
140
120
100
80
60
-1000
-800
-600
-400
-200
0
Weight

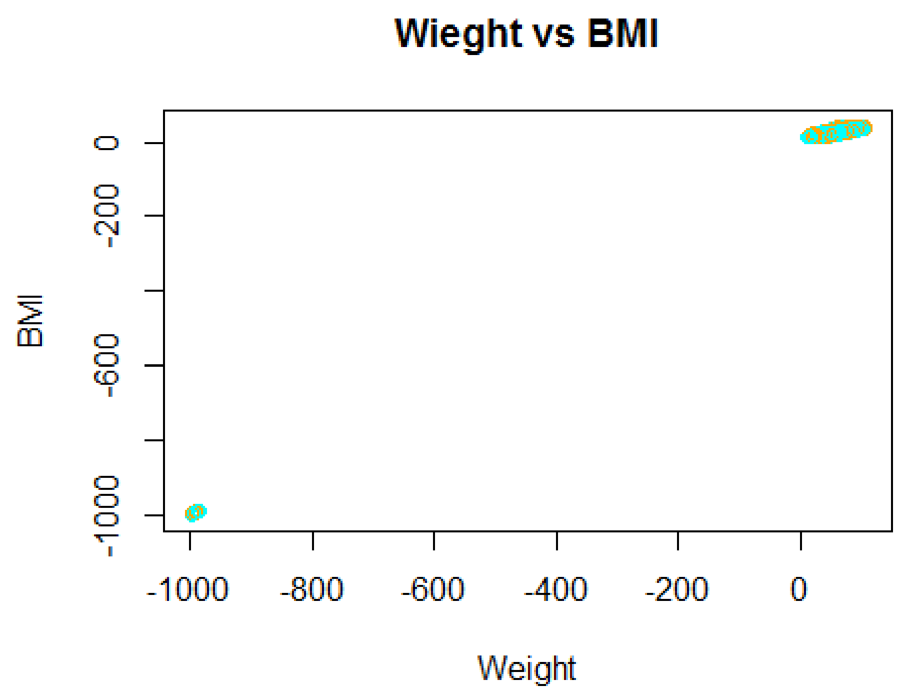
Age vs BMI



Hieght vs weight

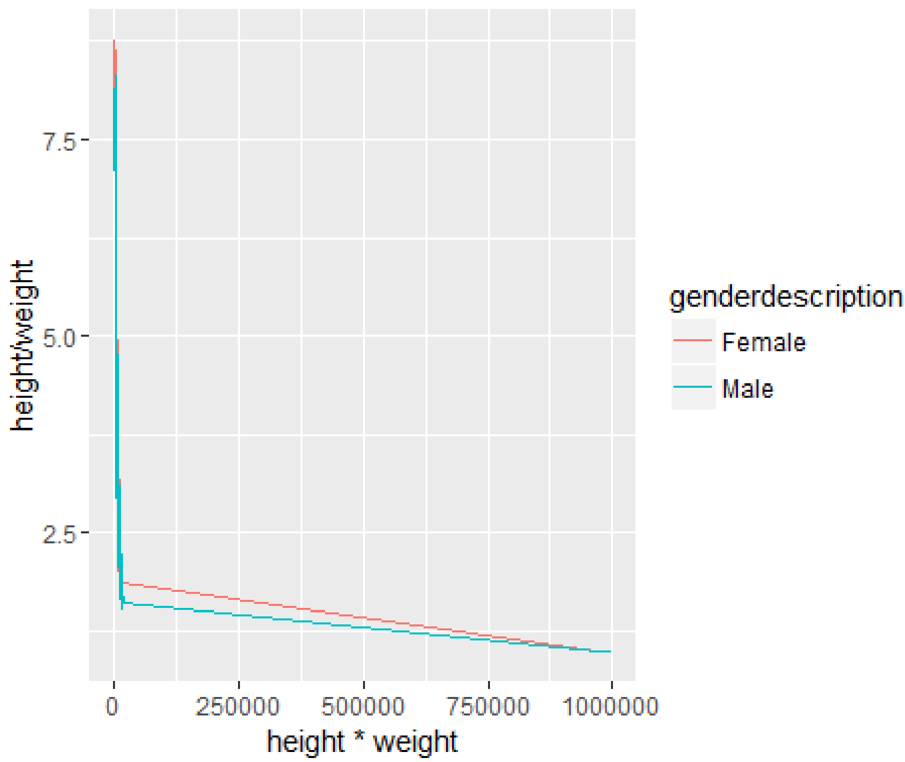




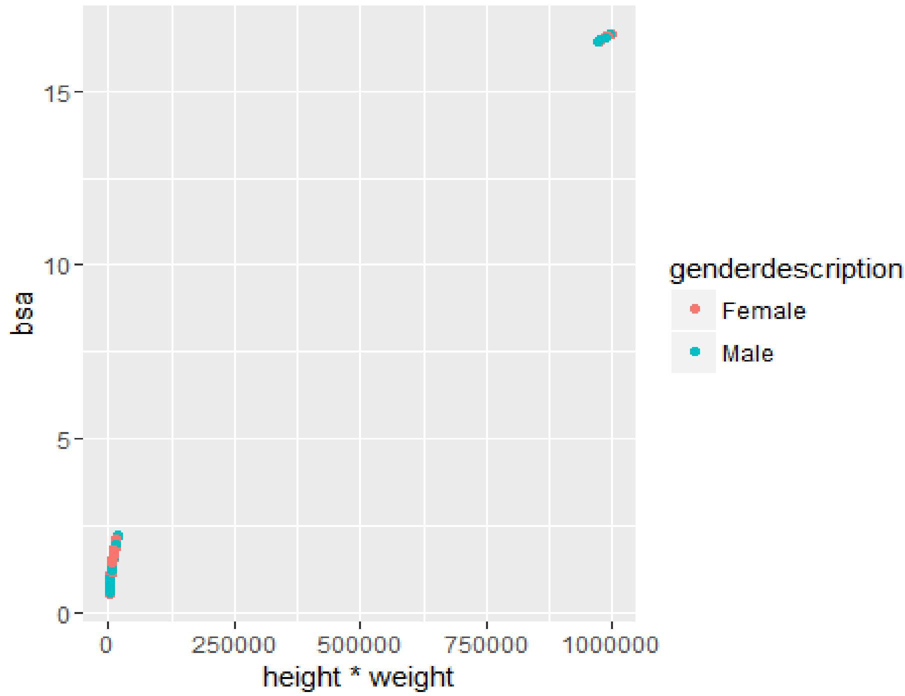


3. Derive new features to plot: height*weight ; height/weight; Body Surface Area ; Plot them against each other; and against BMI, height, or weight. Try to identify interesting meaningful trends or examples, provide some interpretation.

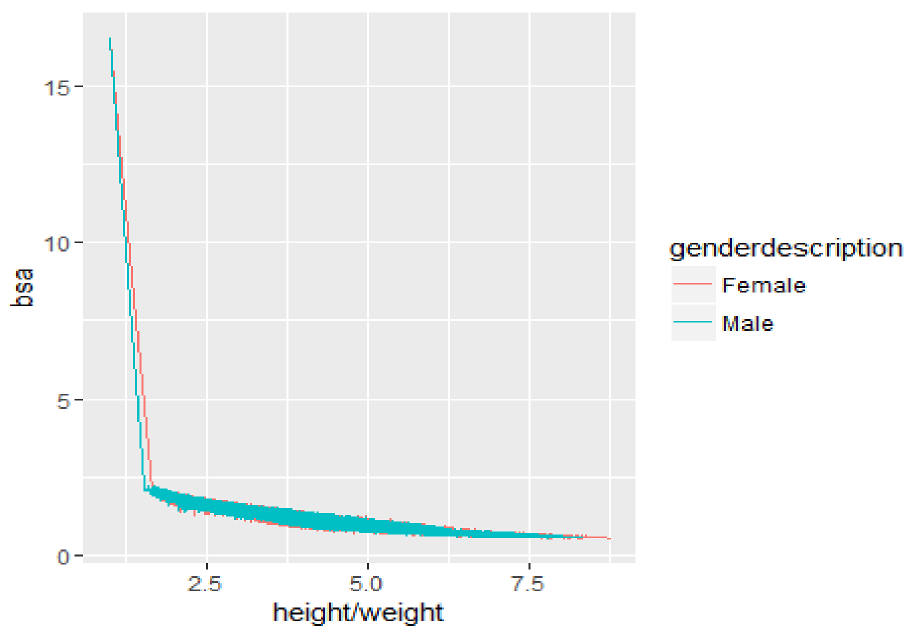
```r
#calculate the Body surface Area using Mosteller formular
childData$bsa <-sqrt((childData$weight*childData$height)/3600)
childData$genderdescription <-factor(childData$genderdescription)
library(ggplot2)

ggplot(childData, aes(height*weight, height/weight, color=genderdescription))
+stat_summary(fun.y=mean, geom="line")
```



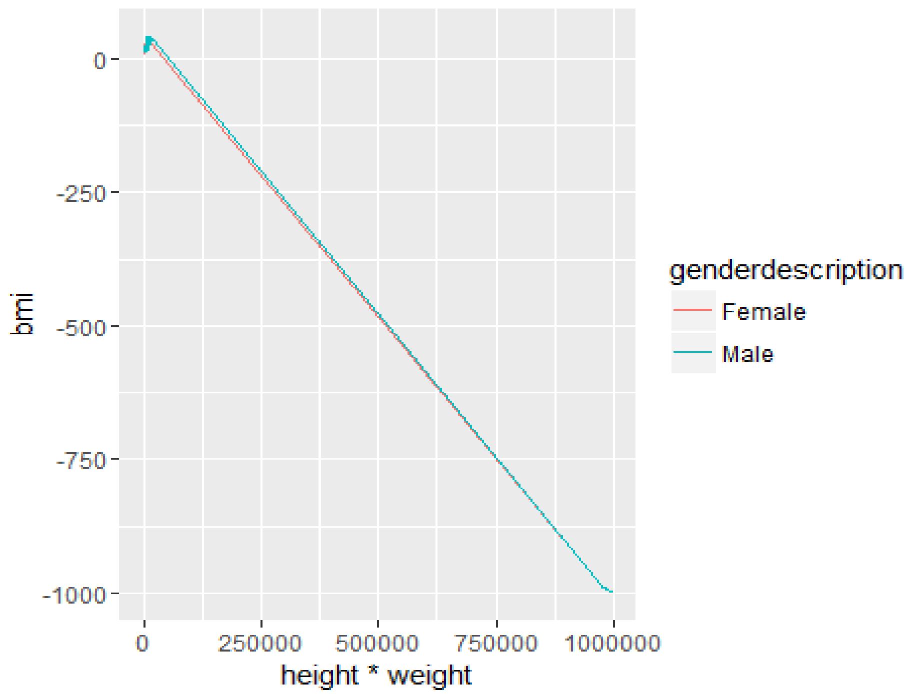
```
ggplot(childData, aes(height*weight, bsa, color=genderdescription))
+geom_point()
```



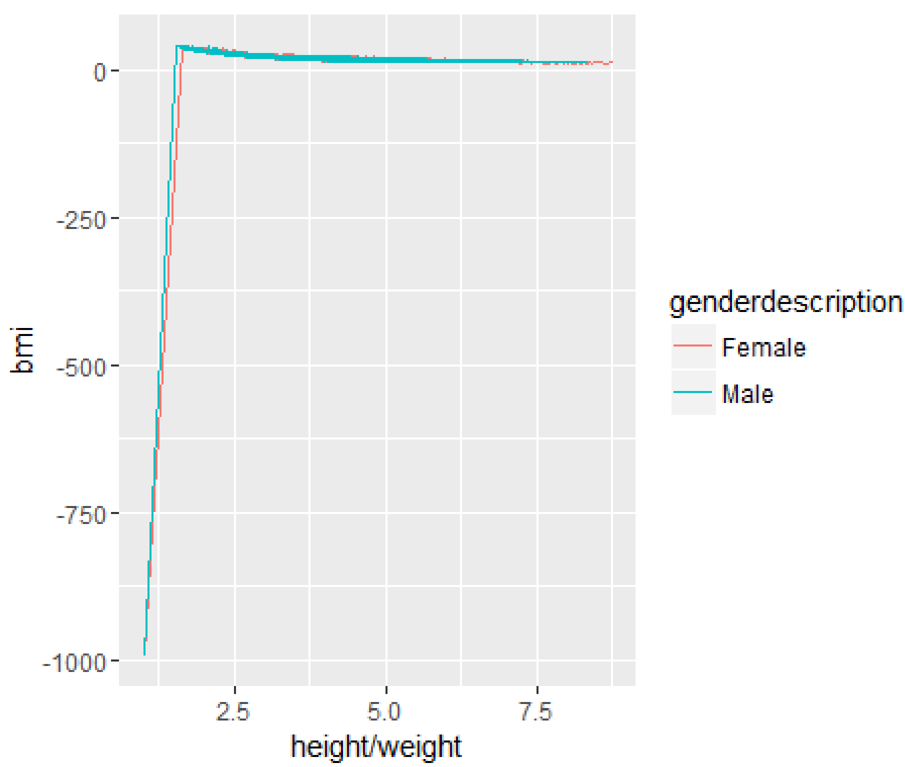

```
ggplot(childData, aes(height/weight, bsa, color=genderdescription))
+stat_summary(fun.y=mean, geom="line")
```



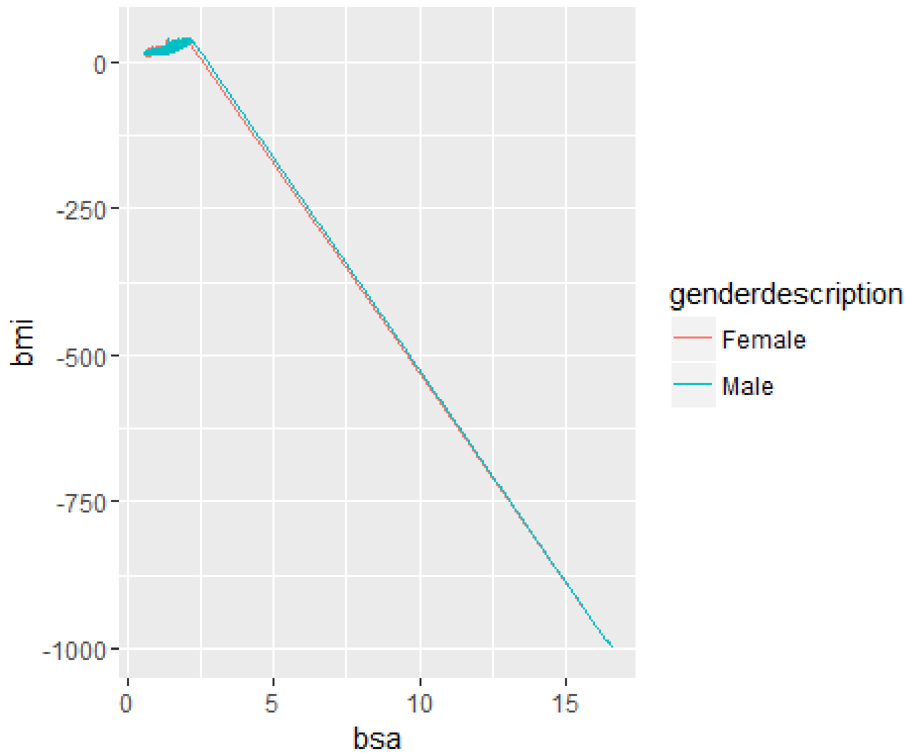

```
ggplot(childData, aes(height*weight, bmi, color=genderdescription))
+stat_summary(fun.y=mean, geom="line")
```




```
ggplot(childData, aes(height/weight, bmi, color=genderdescription))
+stat_summary(fun.y=mean, geom="line")
```




```
ggplot(childData, aes(bsa, bmi, color=genderdescription))
+stat_summary(fun.y=mean, geom="line")
```

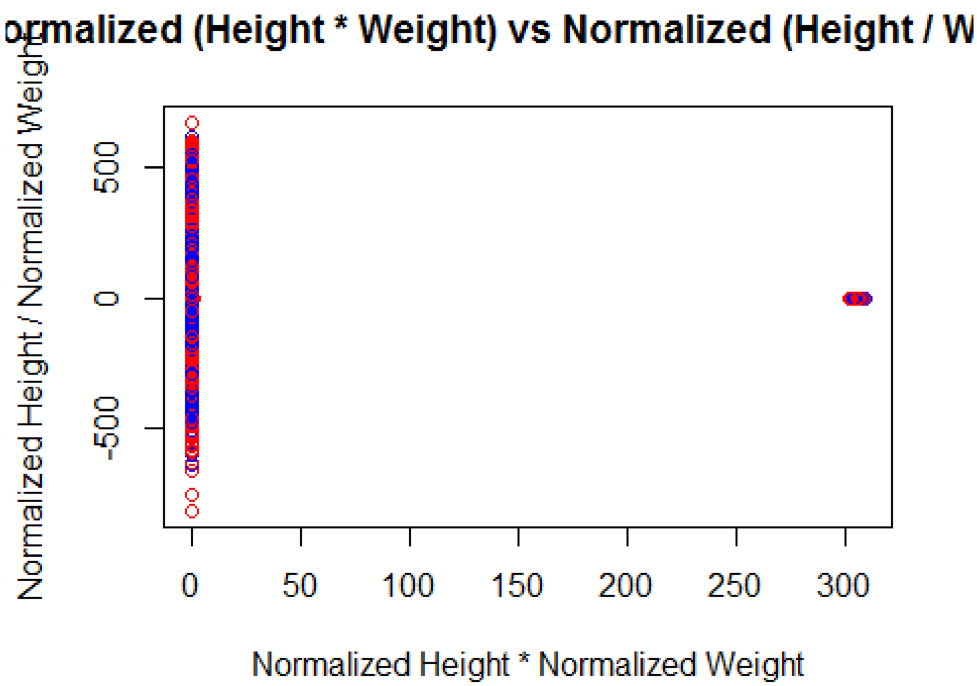


I subsetted a small chunk of the data that included the first 10,000 rows and exported it to excel where I calculated the height*weight, height/weight and Body surface Area and then imported the data back to try our some plots. For the calculation of the body surface area I took the Mosteller formular which calculates the BSA by multiplying the height and weight, dividing it by 3600 and finding the square root of the outcome.

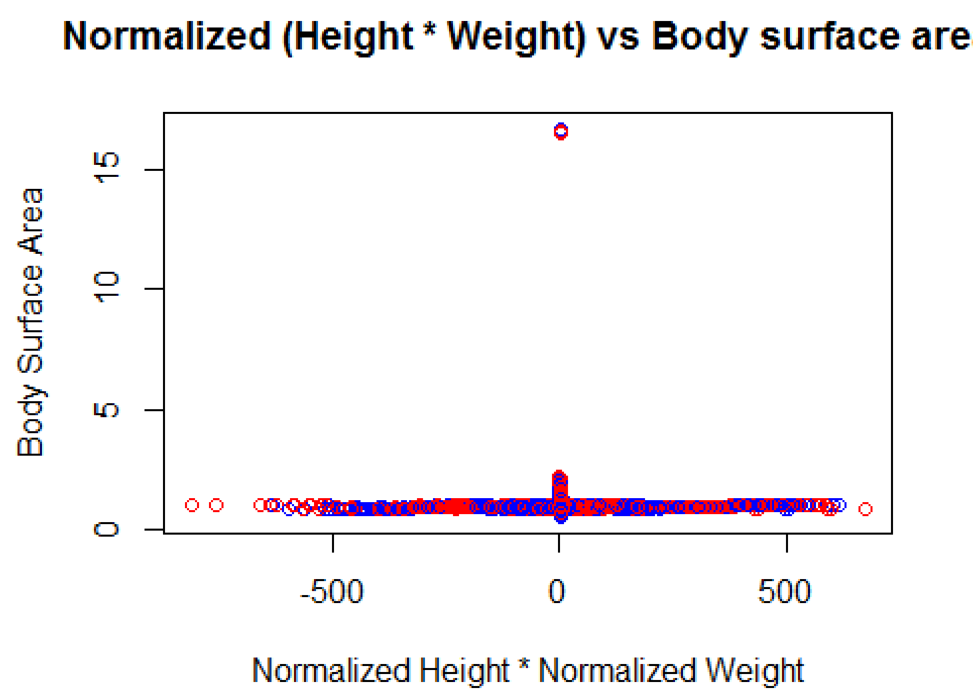
My observation is that the female children seem to grow a bit bigger and faster than the male hence there's a tendency that obesity will be more common in the female children than the male children.

4. Normalise height and weight based on the gender and age, repeat some of the plots from above.

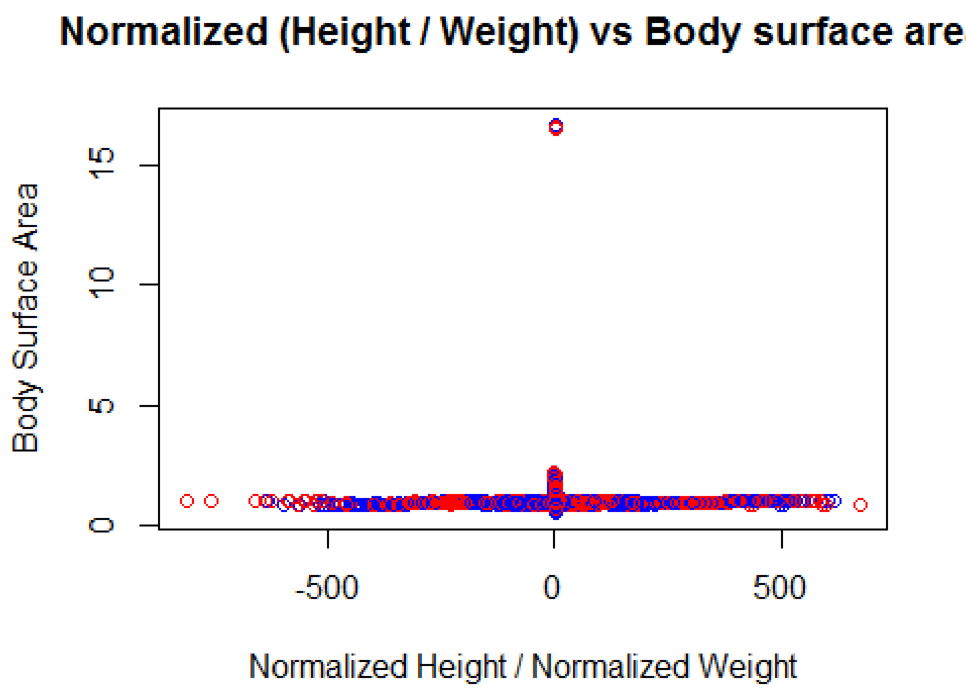
```
library('recommenderlab')

childData$normHeight<-scale(childData$height)
childData$normWeight<-scale(childData$weight)

childData$nHeightMulnWeight <-childData$normHeight*childData$normWeight
childData$nHeightDivnWeight <-childData$normHeight/childData$normWeight


plot(x=childData$nHeightMulnWeight, y=childData$nHeightDivnWeight, xlab
="Normalized Height * Normalized Weight" , ylab ="Normalized Height /
Normalized Weight" , main ="Normalized (Height * Weight) vs Normalized
(Height / Weight)" , col=c("red","blue"))
```



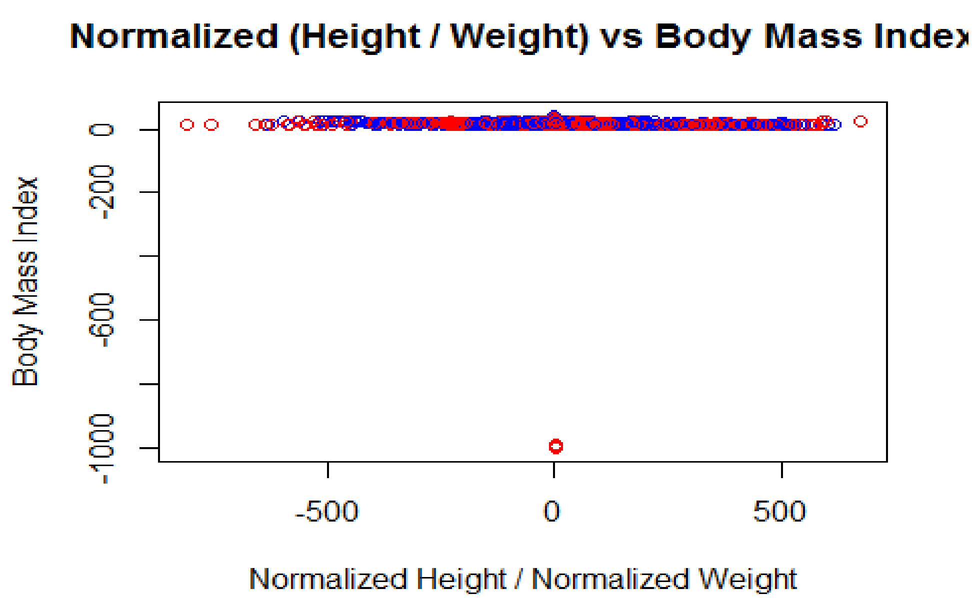

```
plot(x=childData$nHeightDivnWeight, y=childData$bsa, xlab ="Normalized Height
* Normalized Weight" , ylab ="Body Surface Area" , main ="Normalized (Height
* Weight) vs Body surface area" , col=c("red","blue"))
```



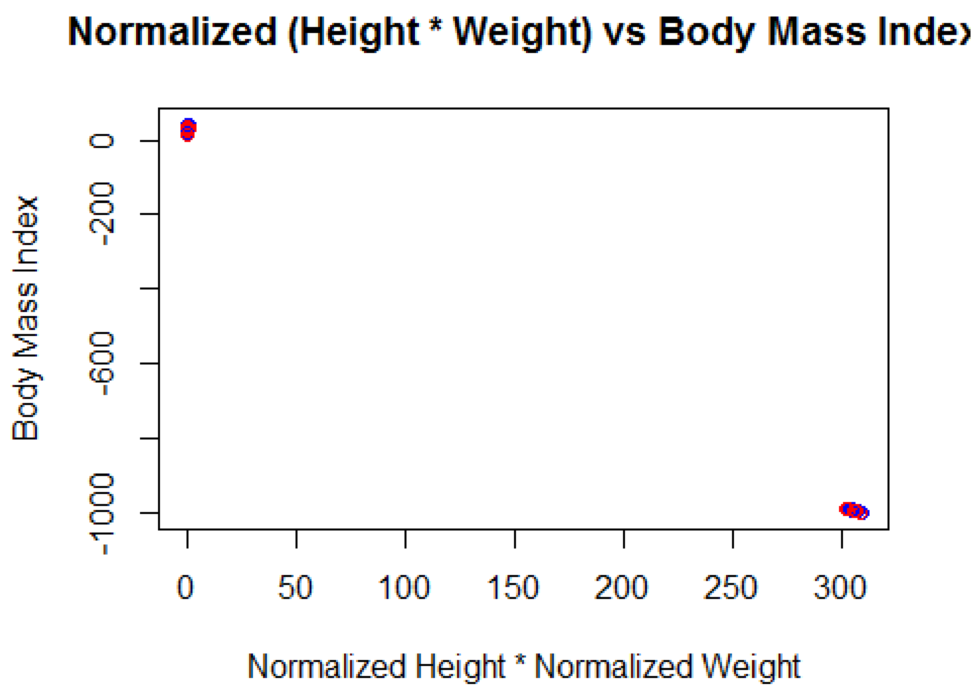
```
plot(x=childData$nHeightDivnWeight, y=childData$bsa, xlab ="Normalized Height
/ Normalized Weight" , ylab ="Body Surface Area" , main ="Normalized (Height
/ Weight) vs Body surface area" , col=c("red","blue"))
```



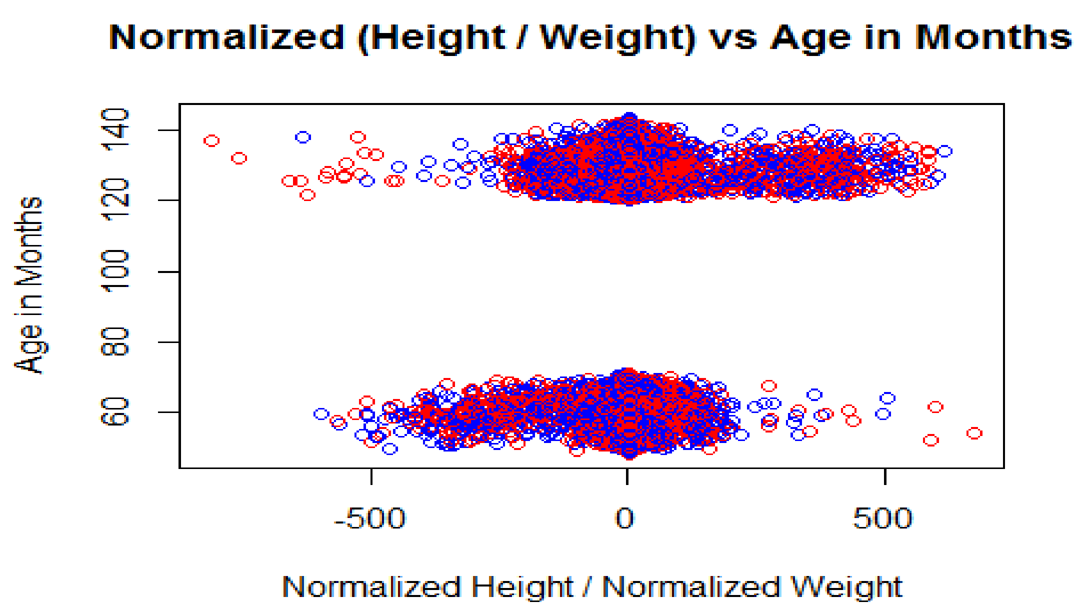

```
plot(x=childData$nHeightDivnWeight, y=childData$bmi, xlab ="Normalized Height
/ Normalized Weight" , ylab ="Body Mass Index" , main ="Normalized (Height /
Weight) vs Body Mass Index" , col=c("red","blue"))
```



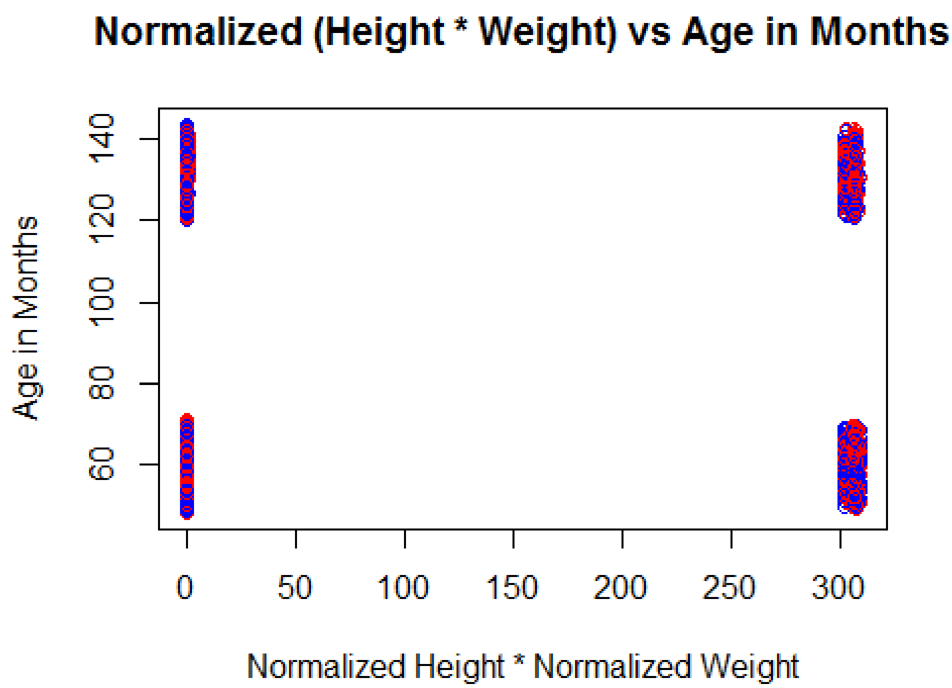
```
plot(x=childData$nHeightMulnWeight, y=childData$bmi, xlab ="Normalized Height * Normalized Weight" , ylab ="Body Mass Index" , main ="Normalized (Height * Weight) vs Body Mass Index" , col=c("red","blue"))
```




```
plot(x=childData$nHeightDivnWeight, y=childData$ageinmonths, xlab ="Normalized Height / Normalized Weight" , ylab ="Age in Months" , main ="Normalized (Height / Weight) vs Age in Months" , col=c("red","blue"))
```



```
plot(x=childData$nHeightMulnWeight, y=childData$ageinmonths, xlab
="Normalized Height * Normalized Weight" , ylab ="Age in Months" , main
="Normalized (Height * Weight) vs Age in Months" , col=c("red","blue"))
```



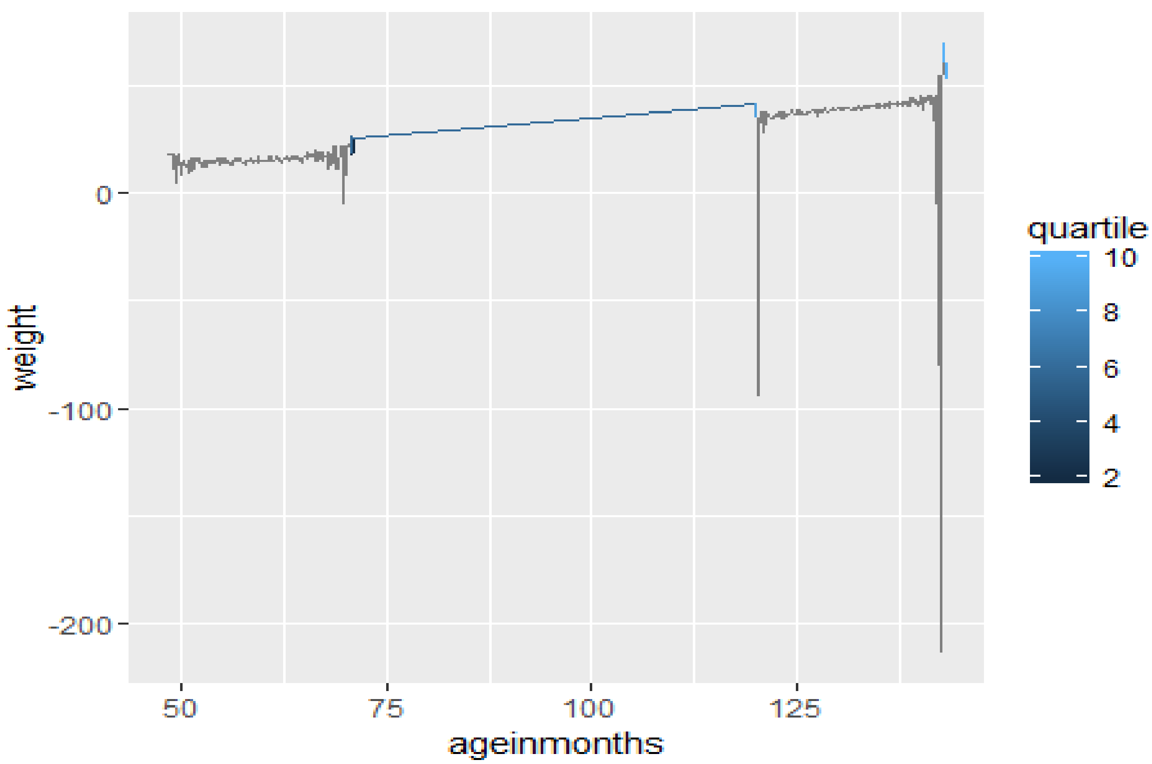
5. Draw approximate growth curves over age. Calculate and plot growth curves of the different deciles (0%-10%, 10%-20%, 20%-30%, ...90%-100%) of obesity categories both by BMI and by weight, for both genders.

```
childData$genderdescription <-factor(childData$genderdescription)

library(nlme)

growth<-lme(height ~ageinmonths, data=childData, random= ~ageinmonths
|genderdescription, method="ML")
print(summary(growth))

## Linear mixed-effects model fit by maximum likelihood
##  Data: childData
##        AIC      BIC   logLik
##   12643623 12643694 -6321805
##
## Random effects:
##  Formula: ~ageinmonths | genderdescription
##  Structure: General positive-definite, Log-Cholesky parametrization
##              StdDev      Corr
## (Intercept)  1.344267109 (Intr)
## ageinmonths  0.007189537 -0.999
```

```
## Residual    61.403857258
##
## Fixed effects: height ~ ageinmonths
##                  Value Std.Error      DF   t-value p-value
## (Intercept) 71.87310 0.9638840 1141856  74.56613       0
## ageinmonths  0.54823 0.0053335 1141856 102.79057       0
##  Correlation:
##             (Intr)
## ageinmonths -0.986
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -18.663167642  -0.007958668   0.058420722   0.120625439   0.431812942
##
## Number of Observations: 1141859
## Number of Groups: 2
```

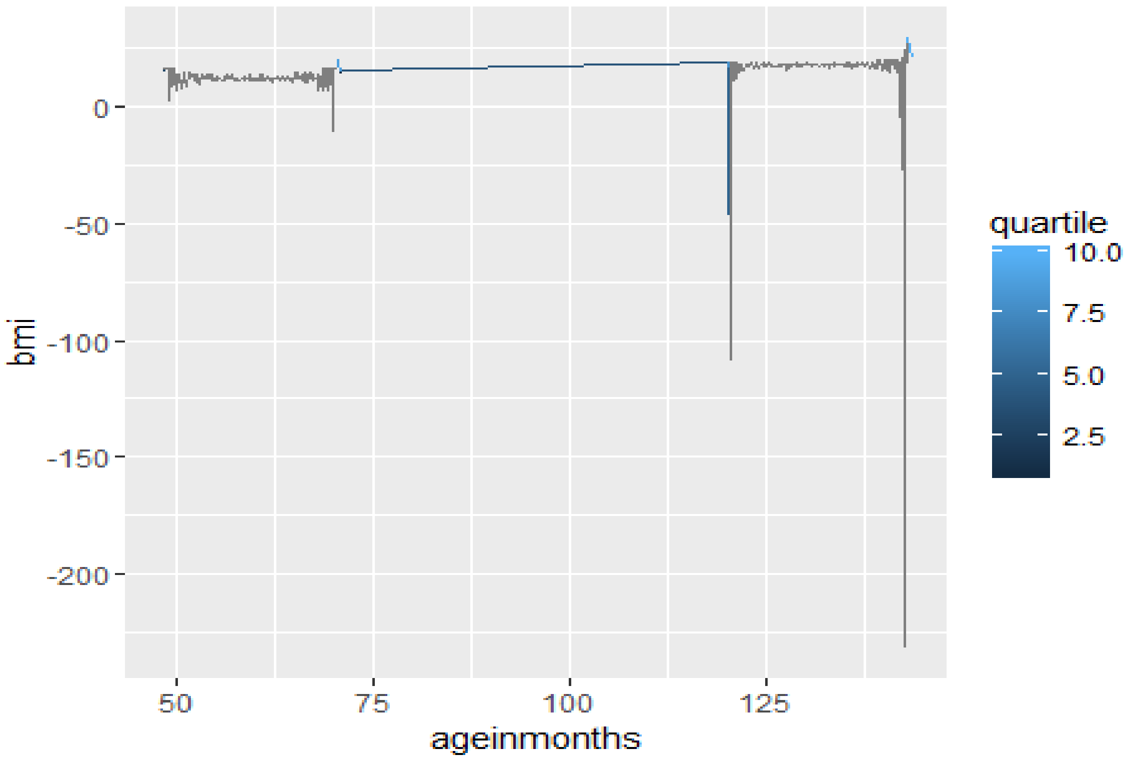
```
library(dplyr)
```

```
obesityByweight<-childData[,c(1,2,4)]
obesityByweight$quartile<-ntile(obesityByweight$weight, 10)
ggplot(obesityByweight, aes(ageinmonths, weight, color=quartile))
+stat_summary(fun.y=mean, geom="line")
```

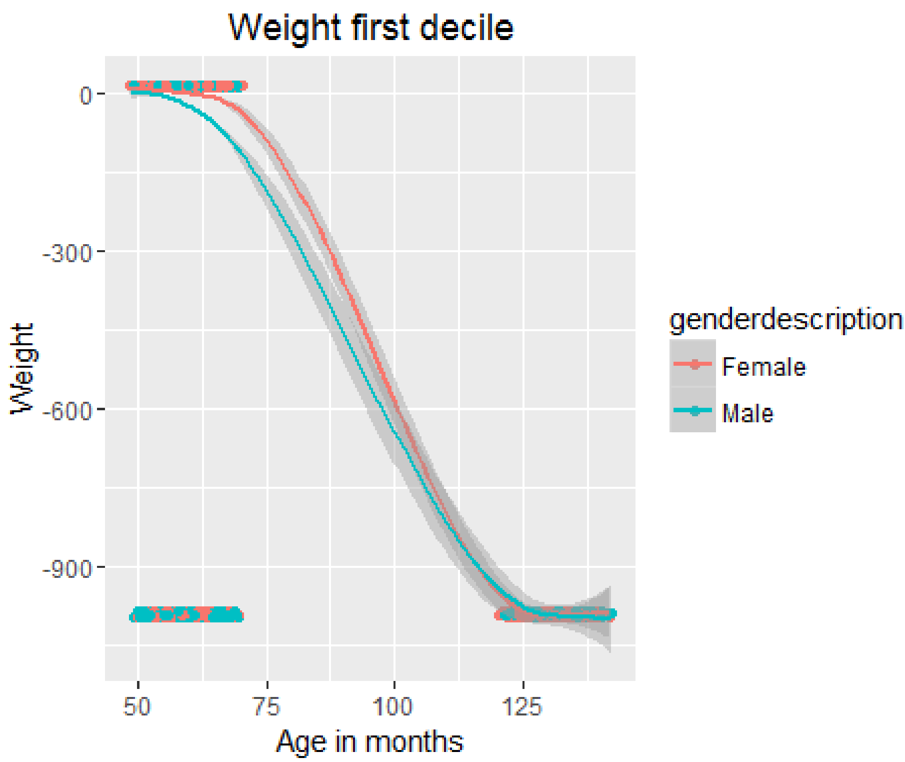



```
obesityBybmi <-childData[,c(1,2,5)]
obesityBybmi$quartile <-ntile(obesityBybmi$bmi, 10)
```

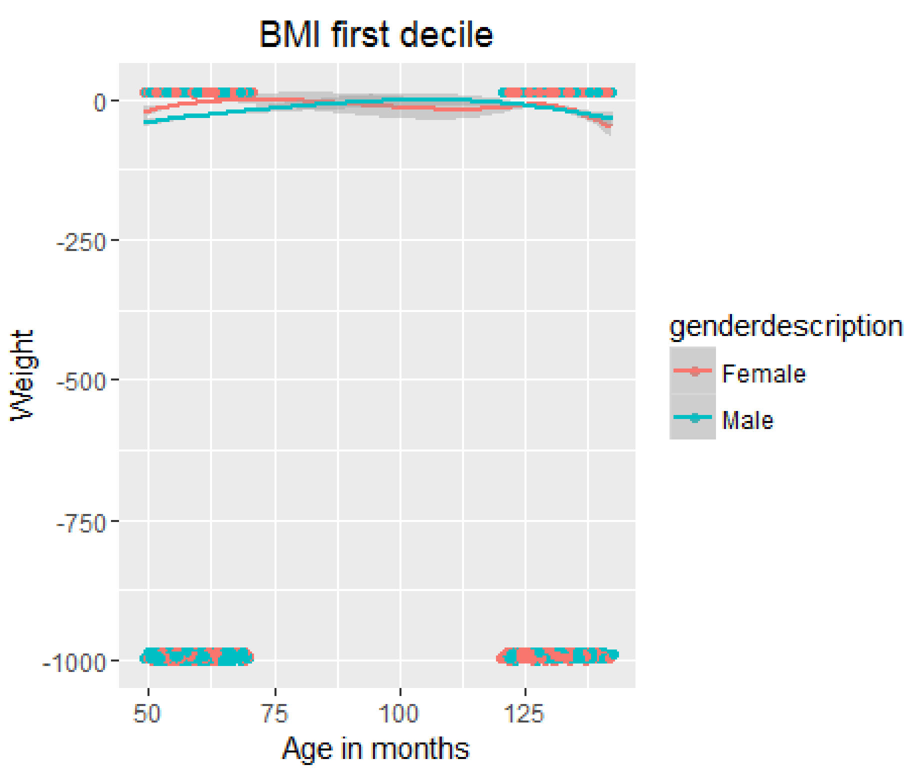
```r
ggplot(obesityBybmi, aes(ageinmonths, bmi, color=quartile))
+stat_summary(fun.y=mean, geom="line")
```



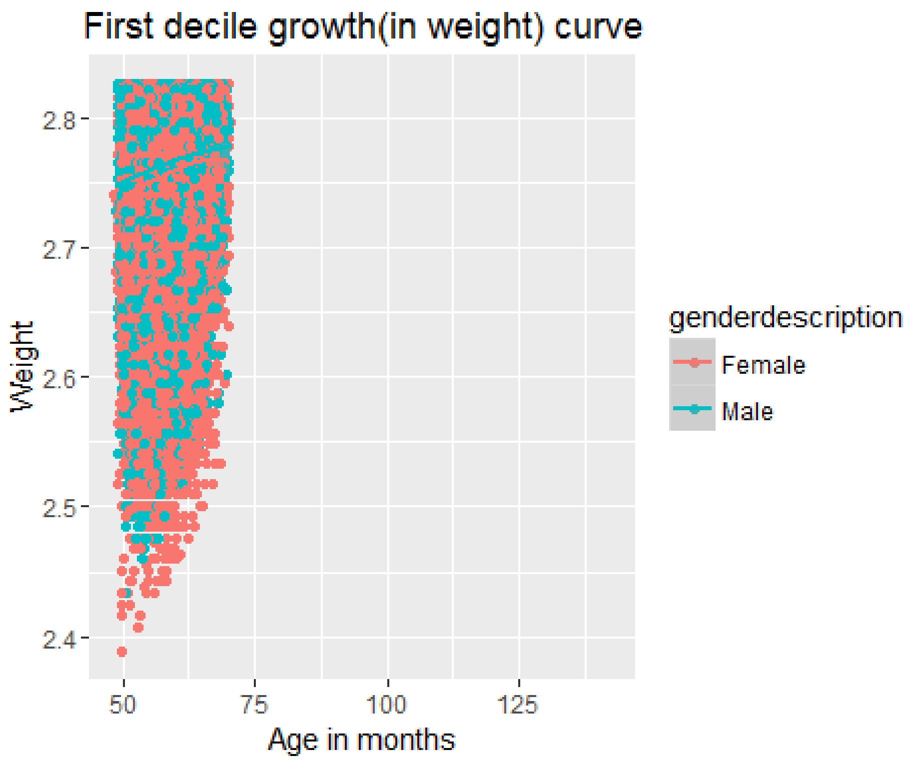

```r
par(mfrow=c(1,2))



#Making subsets to plot based on Male or Female for 1st decile (0% - 10%)
firstdecile_weight <-subset(obesityByweight, quartile ==1)
firstdecile_weight <-na.omit(firstdecile_weight)
firstdecile_bmi <-subset(obesityBybmi, quartile ==1)
firstdecile_bmi <-na.omit(firstdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender

qplot(ageinmonths, weight, data=firstdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight first decile", xlab="Age in
months", ylab="Weight")
```



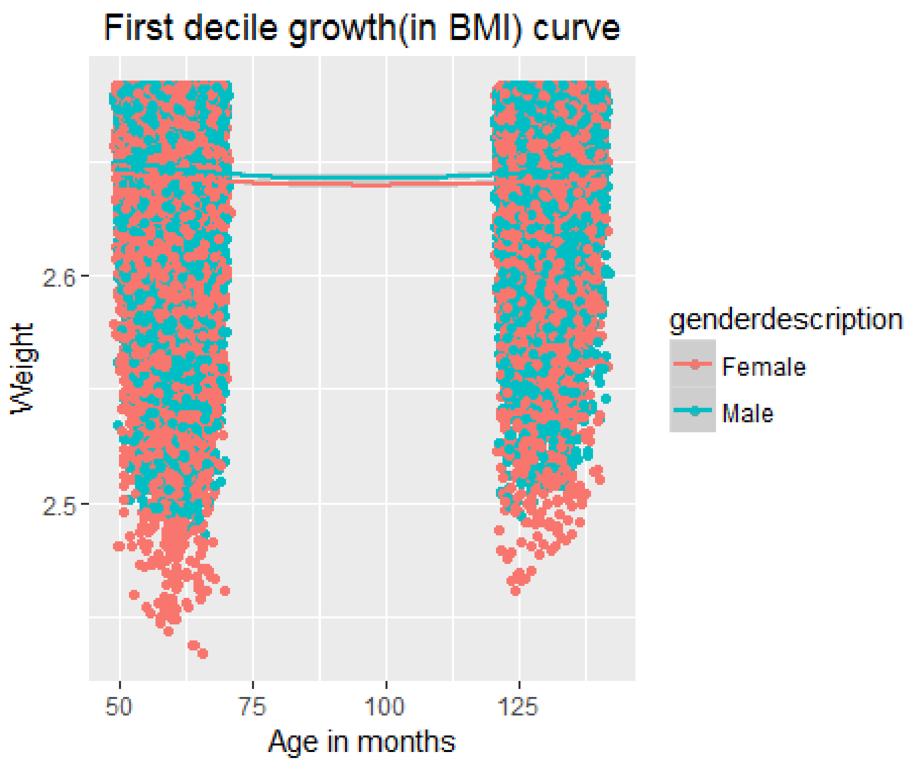

```r
qplot(ageinmonths, bmi, data=firstdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI first decile", xlab="Age in months",
ylab="Weight")
```



```r
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve

qplot(ageinmonths, log(weight), data=firstdecile_weight, geom=c("point", "smooth"), color=genderdescription, main="First decile growth(in weight) curve", xlab="Age in months", ylab="Weight")
```

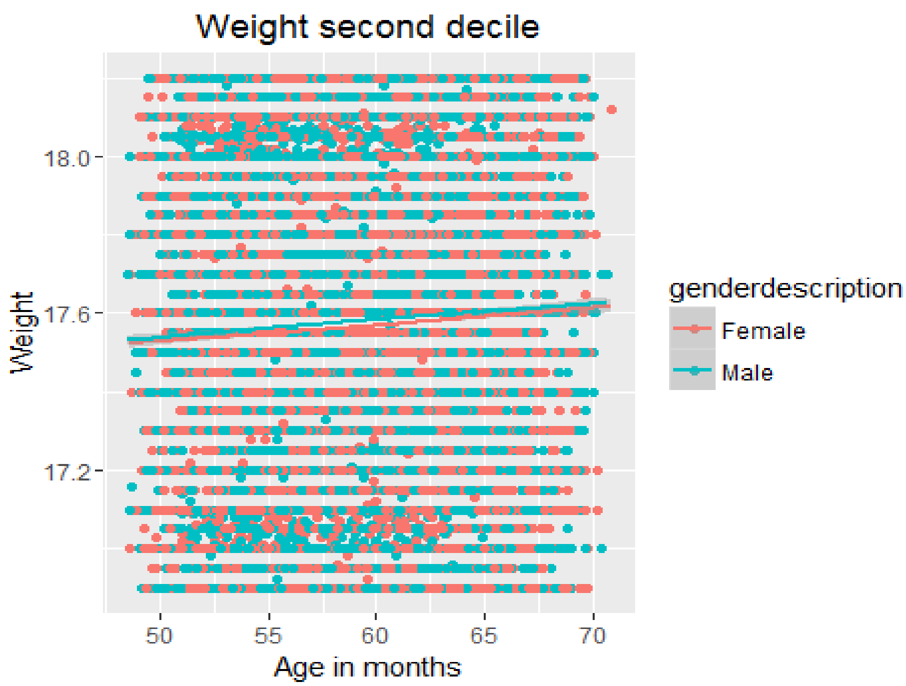



```r
qplot(ageinmonths, log(bmi), data=firstdecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="First decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")

## Warning in log(bmi): NaNs produced

## Warning in log(bmi): NaNs produced

## Warning in log(bmi): NaNs produced

## Warning: Removed 3472 rows containing non-finite values (stat_smooth).

## Warning: Removed 3472 rows containing missing values (geom_point).
```

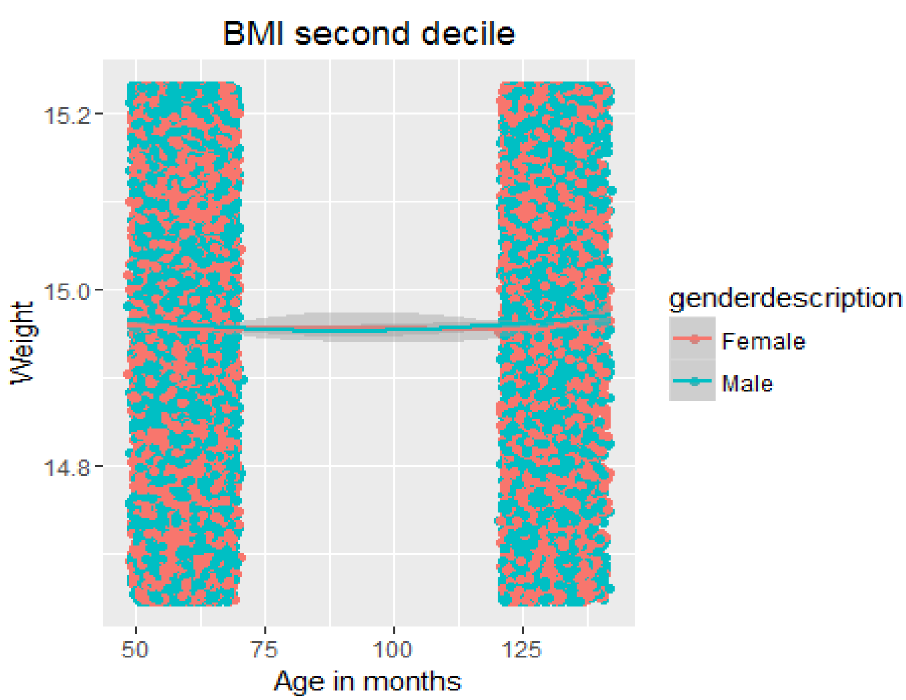



```
#Using Linear modelling function to first perform Regression and get the
intercept dor the growth curve
#print(lm(log(firstdecile_weight[2.5:2.8,"weight"]) ~
firstdecile_weight[2.5:2.8,"ageinmonths"]))
print(lm(log(firstdecile_bmi[2.5:2.8,"bmi"])
~firstdecile_bmi[2.5:3,"ageinmonths"]))

##
## Call:
## lm(formula = log(firstdecile_bmi[2.5:2.8, "bmi"]) ~ firstdecile_bmi[2.5:3,
##     "ageinmonths"])
##
## Coefficients:
##                            (Intercept)
##                                  2.648
## firstdecile_bmi[2.5:3, "ageinmonths"]
##                                     NA

#Making subsets to plot based on Male or Female for 2nd decile (10% - 20%)

seconddecile_weight <-subset(obesityByweight, quartile ==2)
seconddecile_weight <-na.omit(seconddecile_weight)
seconddecile_bmi <-subset(obesityBybmi, quartile ==2)
seconddecile_bmi <-na.omit(seconddecile_bmi)



# Separate regressions of Ageinmonths on weight for each gender
```

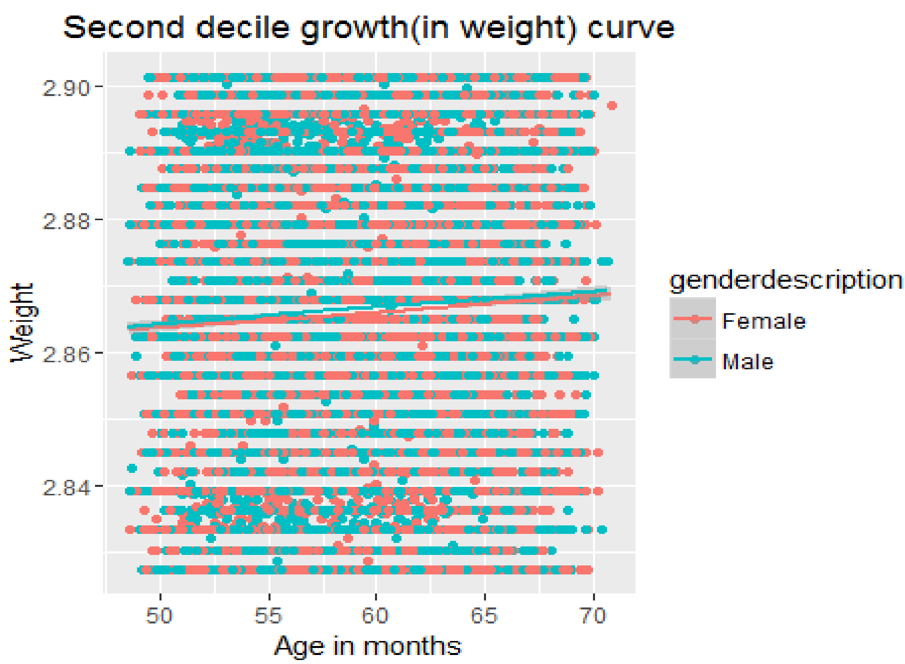
```
qplot(ageinmonths, weight, data=seconddecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight second decile", xlab="Age in
months", ylab="Weight")
```



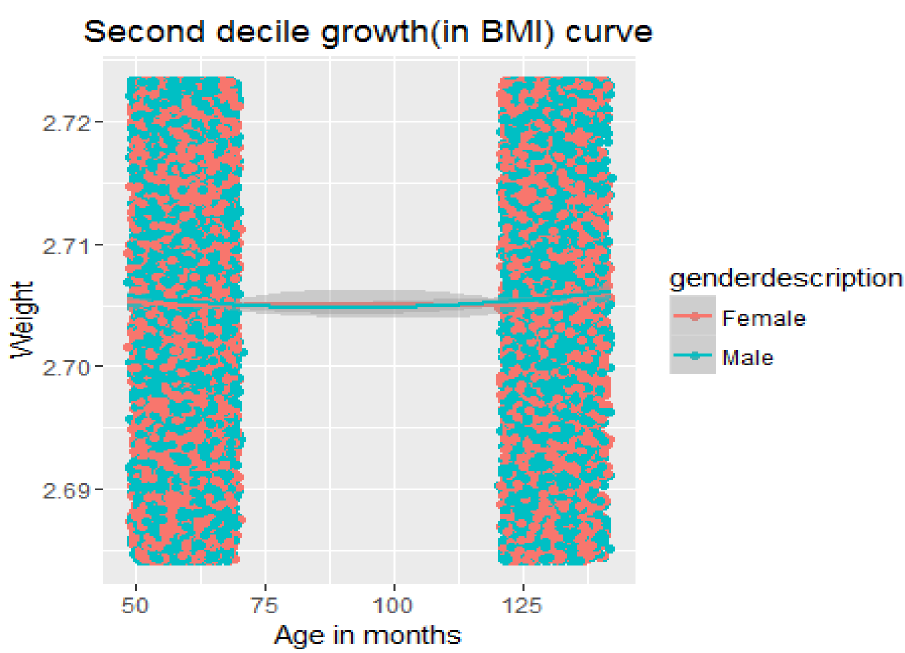

```
qplot(ageinmonths, bmi, data=seconddecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI second decile", xlab="Age in months",
ylab="Weight")
```



```
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve
qplot(ageinmonths, log(weight), data=seconddecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Second decile growth(in weight) curve", xlab="Age in months", ylab="Weight")
```

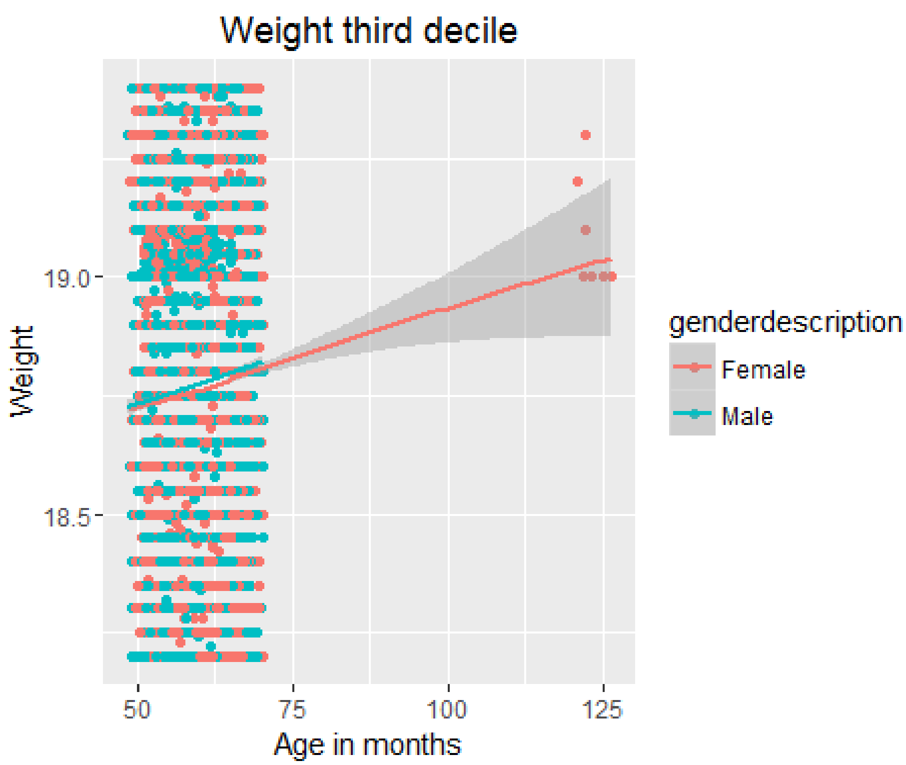



```
qplot(ageinmonths, log(bmi), data=seconddecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="Second decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```

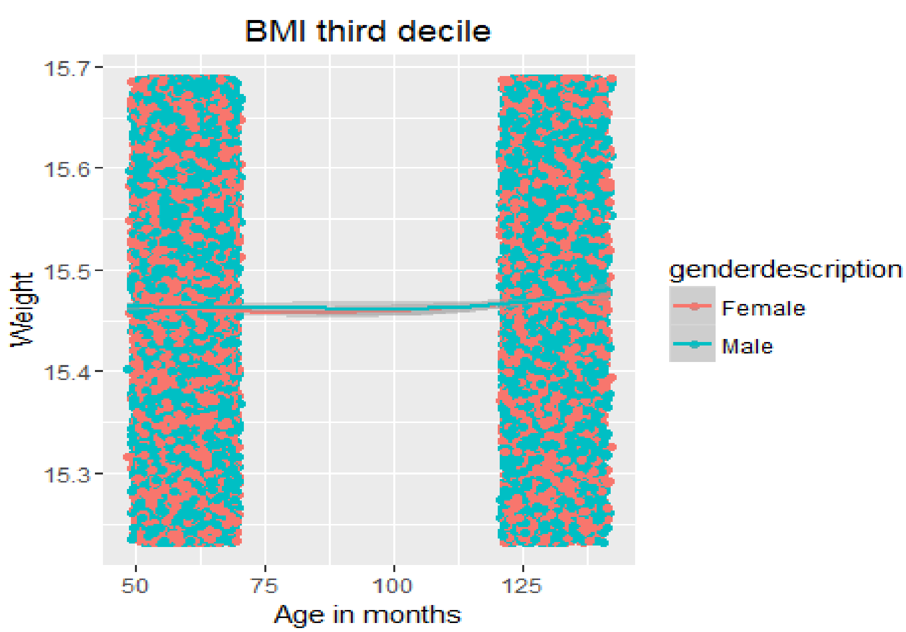


```r
#Using Linear modelling function to first perform Regression and get the
intercept for the growth curve
print(lm(log(seconddecile_weight[2.862:2.87,"weight"])
~seconddecile_weight[2.862:2.87,"ageinmonths"]))
Call:
lm(formula = log(seconddecile_weight[2.862:2.87, "weight"]) ~
seconddecile_weight[2.862:2.87, "ageinmonths"])
Coefficients:
                                   (Intercept)
                                         2.896
 seconddecile_weight[2.862:2.87, "ageinmonths"]
                                             NA

print(lm(log(seconddecile_bmi[2.685 :2.723,"bmi"]) ~seconddecile_bmi[2.685
:2.723,"ageinmonths"]))

##
## Call:
## lm(formula = log(seconddecile_bmi[2.685:2.723, "bmi"]) ~
seconddecile_bmi[2.685:2.723,
##     "ageinmonths"])
##
## Coefficients:
##                                  (Intercept)
##                                        2.693
## seconddecile_bmi[2.685:2.723, "ageinmonths"]
##                                           NA

#Making subsets to plot based on Male or Female for 3rd decile (20% - 30%)

thirddecile_weight <-subset(obesityByweight, quartile ==3)
thirddecile_weight <-na.omit(thirddecile_weight)
thirddecile_bmi <-subset(obesityBybmi, quartile ==3)
thirddecile_bmi <-na.omit(thirddecile_bmi)
```

```r
# Separate regressions of Ageinmonths on weight for each gender
qplot(ageinmonths, weight, data=thirddecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Weight third decile", xlab="Age in months", ylab="Weight")
```

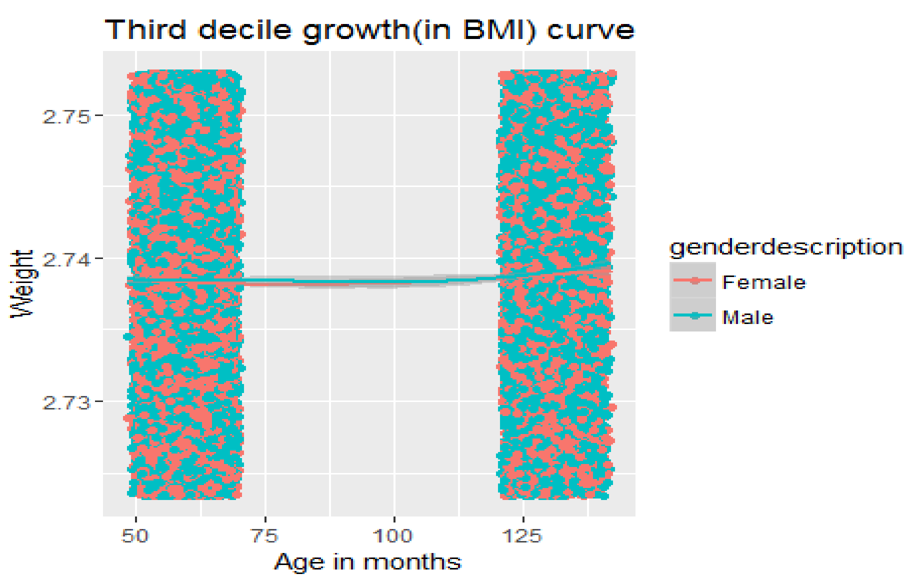



```r
qplot(ageinmonths, bmi, data=thirddecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="BMI third decile", xlab="Age in months", ylab="Weight")
```

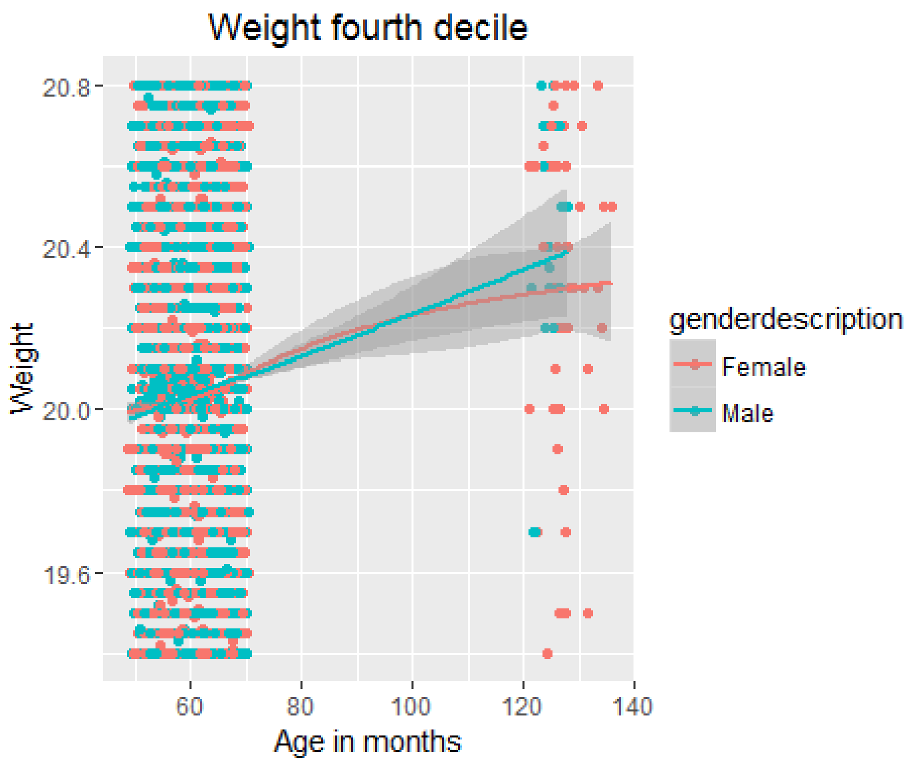


```
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve
qplot(ageinmonths, log(weight), data=thirddecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Third decile growth (in weight) curve", xlab="Age in months", ylab="Weight")
```

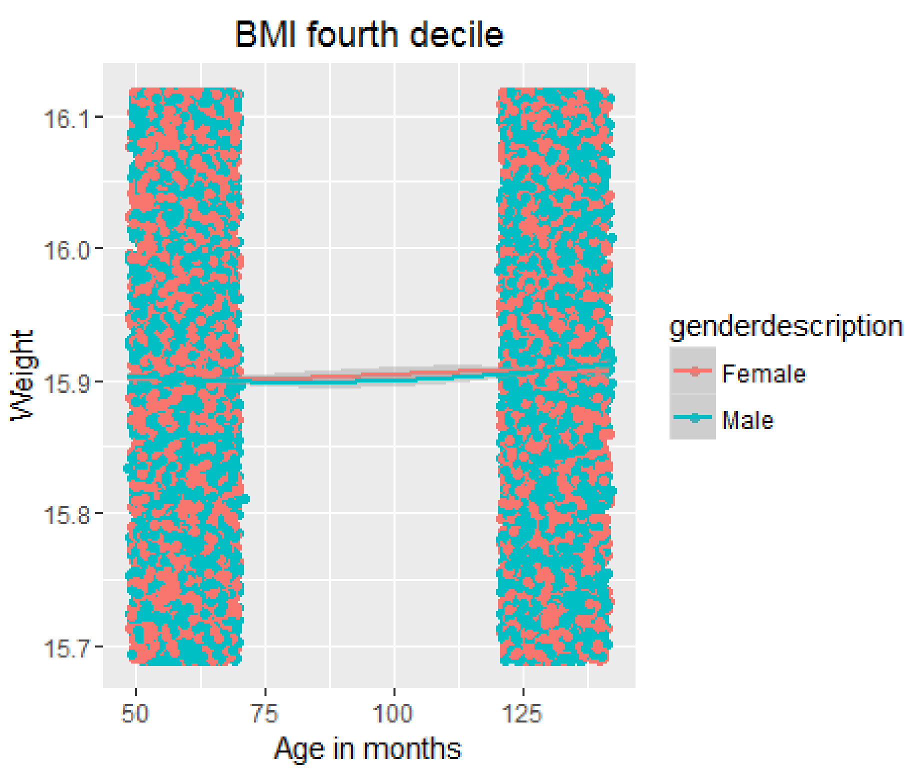

```
qplot(ageinmonths, log(bmi), data=thirddecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="Third decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```



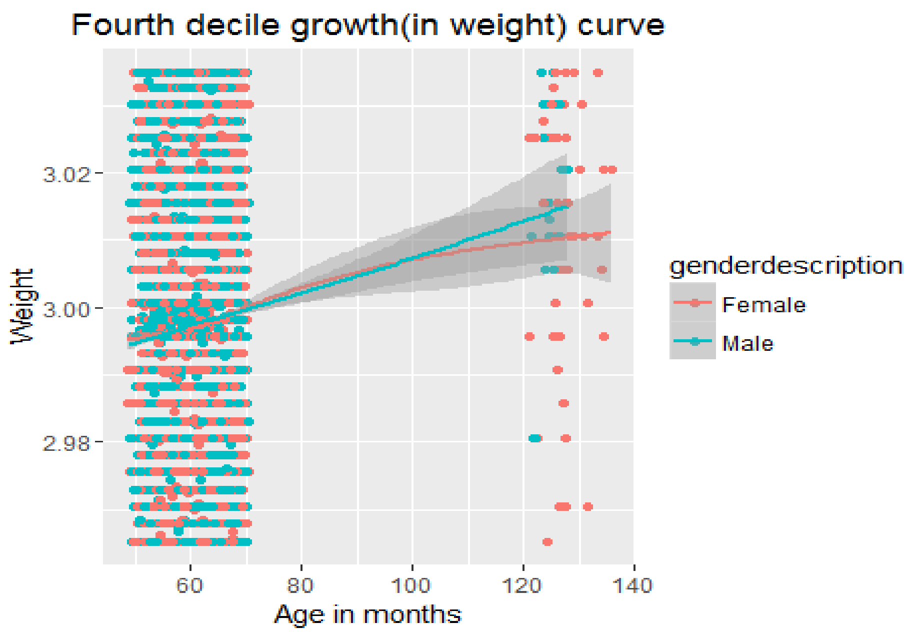
```r
#Using Linear modelling function to first perform Regression and get the
intercept for the growth curve
print(lm(log(thirddecile_weight[2.93:2.95,"weight"])
~thirddecile_weight[2.93:2.95,"ageinmonths"]))

##
## Call:
## lm(formula = log(thirddecile_weight[2.93:2.95, "weight"]) ~
thirddecile_weight[2.93:2.95,
##     "ageinmonths"])
##
## Coefficients:
##                                  (Intercept)
##                                        2.904
## thirddecile_weight[2.93:2.95, "ageinmonths"]
##                                           NA

print(lm(log(thirddecile_bmi[2.723 :2.753,"bmi"]) ~thirddecile_bmi[2.723
:2.753,"ageinmonths"]))

##
## Call:
## lm(formula = log(thirddecile_bmi[2.723:2.753, "bmi"]) ~
thirddecile_bmi[2.723:2.753,
##     "ageinmonths"])
##
## Coefficients:
##                                 (Intercept)
##                                       2.731
## thirddecile_bmi[2.723:2.753, "ageinmonths"]
##                                          NA

#Making subsets to plot based on Male or Female for 4th decile (30% - 40%)

fourthdecile_weight <-subset(obesityByweight, quartile ==4)
fourthdecile_weight <-na.omit(fourthdecile_weight)
fourthdecile_bmi <-subset(obesityBybmi, quartile ==4)
fourthdecile_bmi <-na.omit(fourthdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender

qplot(ageinmonths, weight, data=fourthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight fourth decile", xlab="Age in
months", ylab="Weight")
```



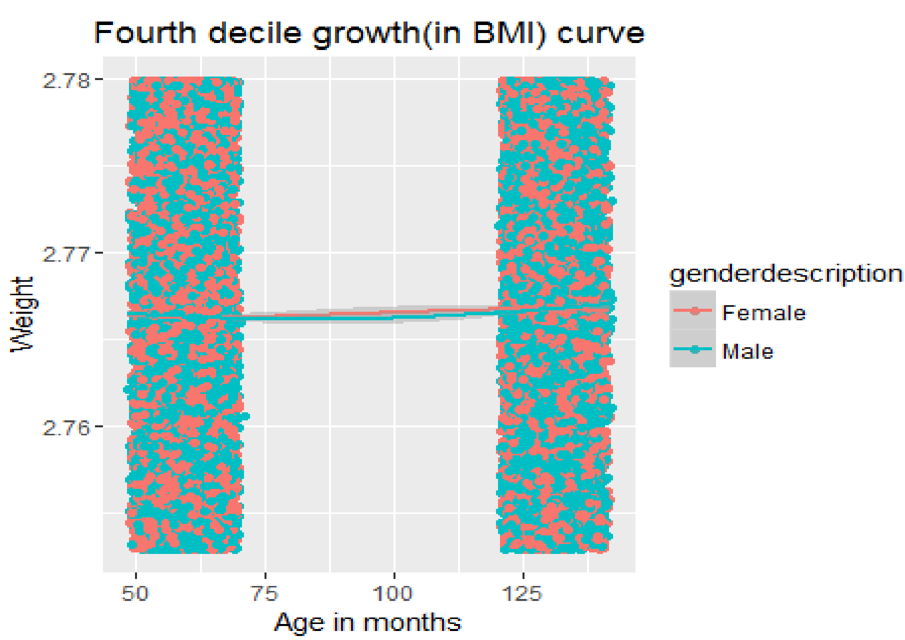

```
qplot(ageinmonths, bmi, data=fourthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI fourth decile", xlab="Age in months",
ylab="Weight")
```



```r
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve

qplot(ageinmonths, log(weight), data=fourthdecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Fourth decile growth(in weight) curve", xlab="Age in months", ylab="Weight")
```

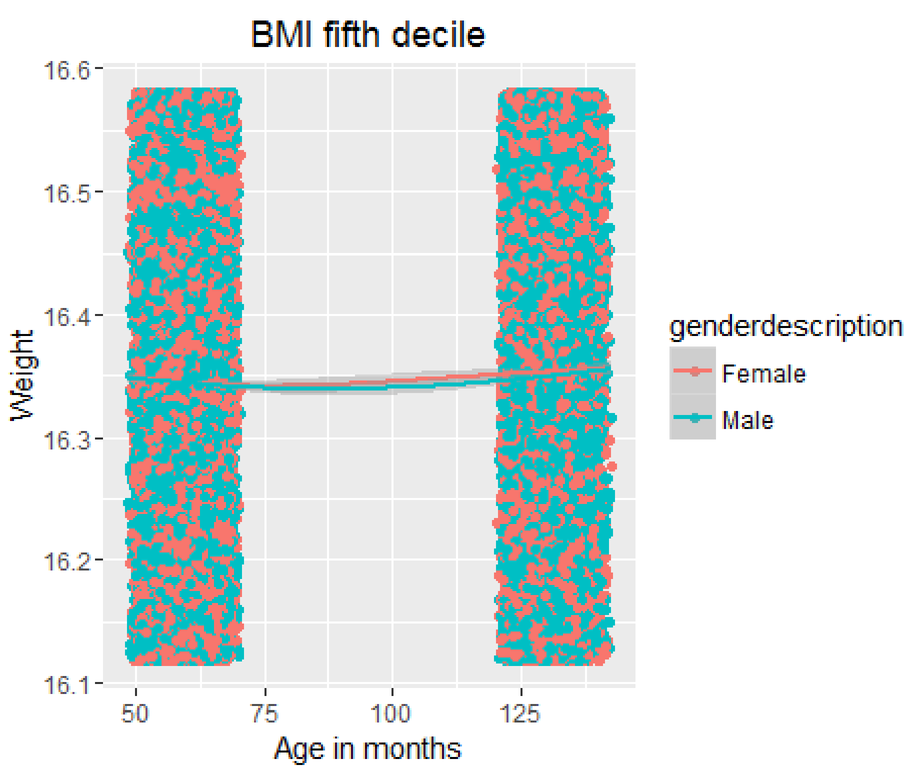



```r
qplot(ageinmonths, log(bmi), data=fourthdecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="Fourth decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```



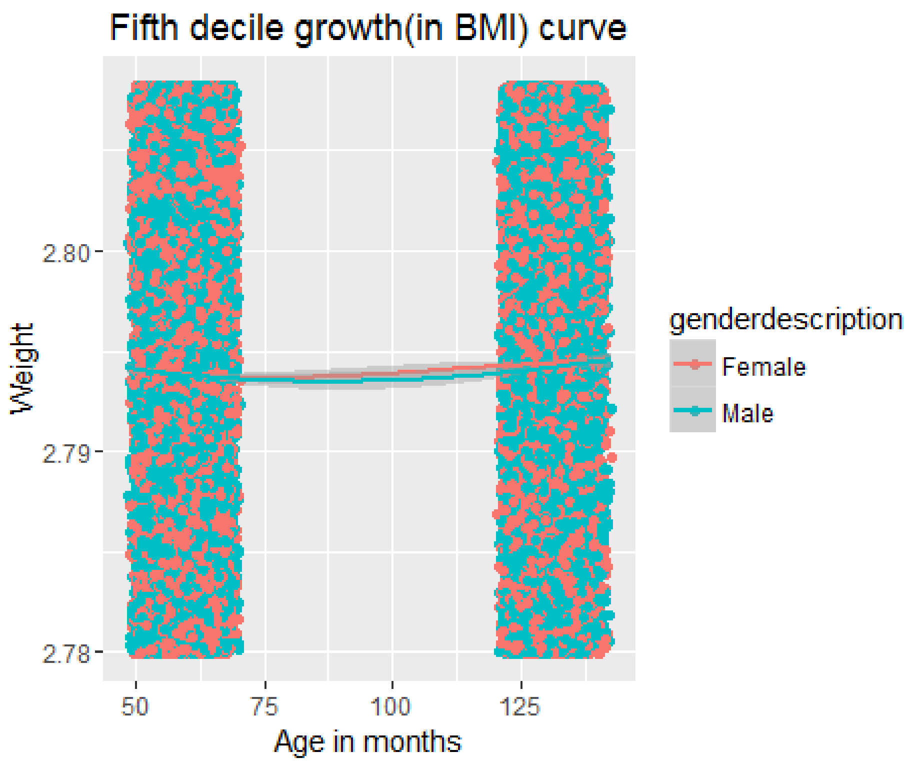
```r
#Using Linear modelling function to first perform Regression and get the intercept dor the growth curve
print(lm(log(fourthdecile_weight[2.5:2.8,"weight"])
~fourthdecile_weight[2.5:2.8,"ageinmonths"]))

##
## Call:
## lm(formula = log(fourthdecile_weight[2.5:2.8, "weight"]) ~
fourthdecile_weight[2.5:2.8,
##     "ageinmonths"])
##
## Coefficients:
##                                 (Intercept)
##                                       2.996
## fourthdecile_weight[2.5:2.8, "ageinmonths"]
##                                          NA

print(lm(log(fourthdecile_bmi[2.5:2.8,"bmi"])
~fourthdecile_bmi[2.5:3,"ageinmonths"]))

##
## Call:
## lm(formula = log(fourthdecile_bmi[2.5:2.8, "bmi"]) ~
fourthdecile_bmi[2.5:3,
##     "ageinmonths"])
##
## Coefficients:
##                            (Intercept)
##                                  2.755
## fourthdecile_bmi[2.5:3, "ageinmonths"]
##                                     NA

#Making subsets to plot based on Male or Female for 5th decile (40% - 50%)

fifthdecile_weight <-subset(obesityByweight, quartile ==5)
fifthdecile_weight <-na.omit(fifthdecile_weight)
fifthdecile_bmi <-subset(obesityBybmi, quartile ==5)
fifthdecile_bmi <-na.omit(fifthdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender
qplot(ageinmonths, weight, data=fifthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight fifth decile", xlab="Age in
months", ylab="Weight")
```


```
qplot(ageinmonths, bmi, data=fifthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI fifth decile", xlab="Age in months",
ylab="Weight")
```



```r
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve
qplot(ageinmonths, log(weight), data=fifthdecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Fifth decile growth(in weight) curve", xlab="Age in months", ylab="Weight")
```


```r
qplot(ageinmonths, log(bmi), data=fifthdecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="Fifth decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```

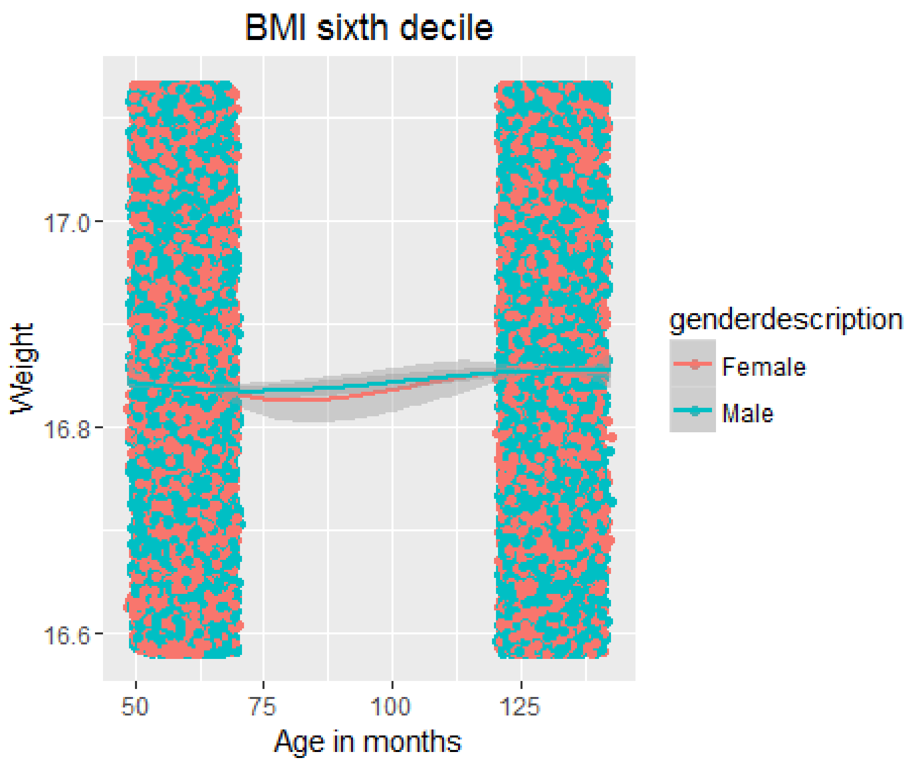



```
#Using Linear modelling function to first perform Regression and get the
intercept for the growth curve
print(lm(log(fifthdecile_weight[2.862:2.87,"weight"])
~fifthdecile_weight[2.862:2.87,"ageinmonths"]))

##
## Call:
## lm(formula = log(fifthdecile_weight[2.862:2.87, "weight"]) ~
##     fifthdecile_weight[2.862:2.87, "ageinmonths"])
##
## Coefficients:
##                                   (Intercept)
##                                         3.054
## fifthdecile_weight[2.862:2.87, "ageinmonths"]
##                                            NA

print(lm(log(fifthdecile_bmi[2.685 :2.723,"bmi"]) ~fifthdecile_bmi[2.685
:2.723,"ageinmonths"]))

##
## Call:
## lm(formula = log(fifthdecile_bmi[2.685:2.723, "bmi"]) ~
fifthdecile_bmi[2.685:2.723,
##     "ageinmonths"])
##
## Coefficients:
##                                 (Intercept)
```

```
##                                                    2.797
## fifthdecile_bmi[2.685:2.723, "ageinmonths"]
##                                                       NA
```

```r
#Making subsets to plot based on Male or Female for 6th decile (50% - 60%)

sixthdecile_weight <-subset(obesityByweight, quartile ==6)
sixthdecile_weight <-na.omit(sixthdecile_weight)
sixthdecile_bmi <-subset(obesityBybmi, quartile ==6)
sixthdecile_bmi <-na.omit(sixthdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender

qplot(ageinmonths, weight, data=sixthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight sixth decile", xlab="Age in
months", ylab="Weight")
```

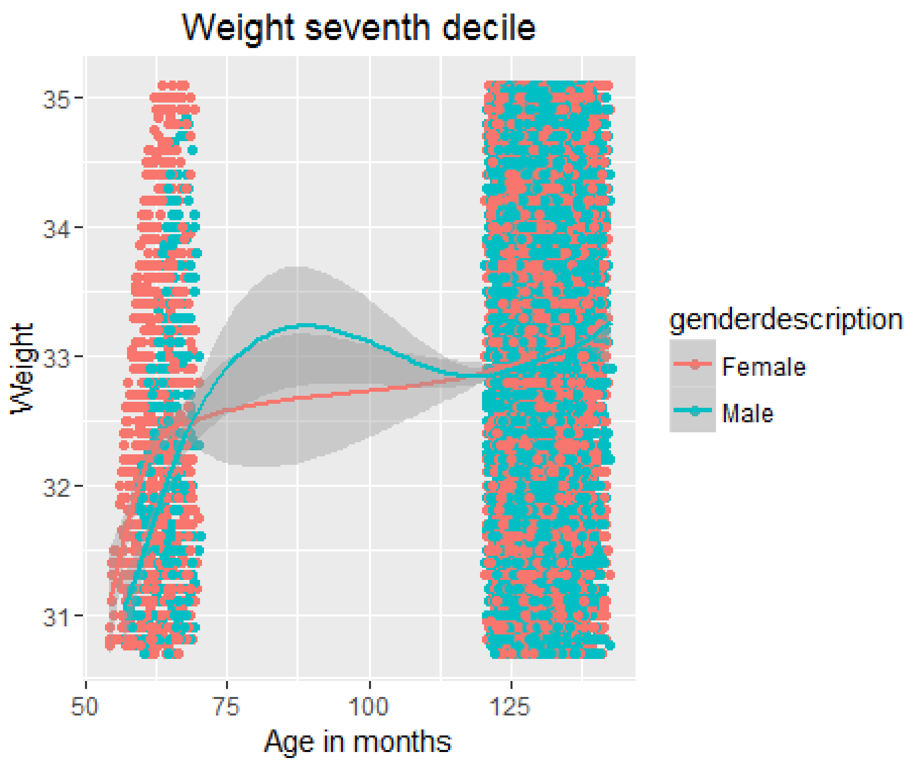

```r
qplot(ageinmonths, bmi, data=sixthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI sixth decile", xlab="Age in months",
ylab="Weight")
```



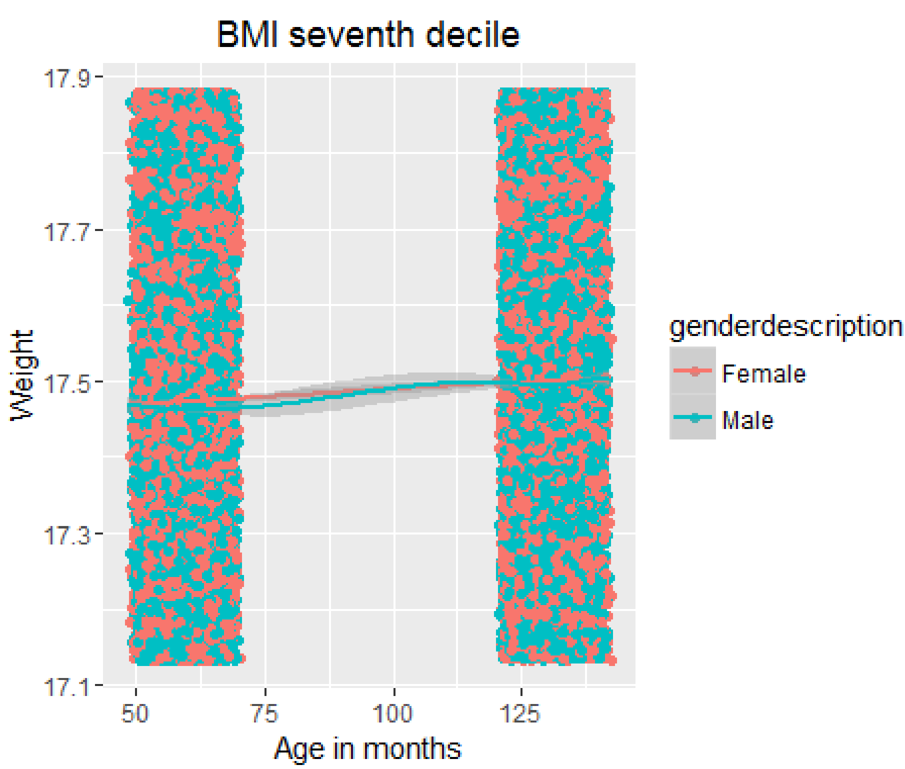

```
#Using the log of variables weight and bmi to plot against time (Age in
month) in order to get the growth curve

qplot(ageinmonths, log(weight), data=sixthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Sixth decile growth(in weight)
curve", xlab="Age in months", ylab="Weight")
```


```
qplot(ageinmonths, log(bmi), data=sixthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="Sixth decile growth(in BMI) curve", xlab="Age
in months", ylab="Weight")
```

```r
#Using Linear modelling function to first perform Regression and get the
intercept dor the growth curve
print(lm(log(sixthdecile_weight[2.5:2.8,"weight"])
~sixthdecile_weight[2.5:2.8,"ageinmonths"]))

##
## Call:
## lm(formula = log(sixthdecile_weight[2.5:2.8, "weight"]) ~
sixthdecile_weight[2.5:2.8,
##     "ageinmonths"])
##
## Coefficients:
##                                (Intercept)
##                                      3.391
## sixthdecile_weight[2.5:2.8, "ageinmonths"]
##                                         NA

print(lm(log(sixthdecile_bmi[2.5:2.8,"bmi"])
~sixthdecile_bmi[2.5:3,"ageinmonths"]))

##
## Call:
## lm(formula = log(sixthdecile_bmi[2.5:2.8, "bmi"]) ~ sixthdecile_bmi[2.5:3,
##     "ageinmonths"])
##
## Coefficients:
##                           (Intercept)
##                                 2.833
## sixthdecile_bmi[2.5:3, "ageinmonths"]
##                                    NA

#Making subsets to plot based on Male or Female for 7th decile (60% - 70%)

seventhdecile_weight <-subset(obesityByweight, quartile ==7)
seventhdecile_weight <-na.omit(seventhdecile_weight)
seventhdecile_bmi <-subset(obesityBybmi, quartile ==7)
seventhdecile_bmi <-na.omit(seventhdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender
qplot(ageinmonths, weight, data=seventhdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight seventh decile", xlab="Age
in months", ylab="Weight")
```




```r
qplot(ageinmonths, bmi, data=seventhdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI seventh decile", xlab="Age in months",
ylab="Weight")
```

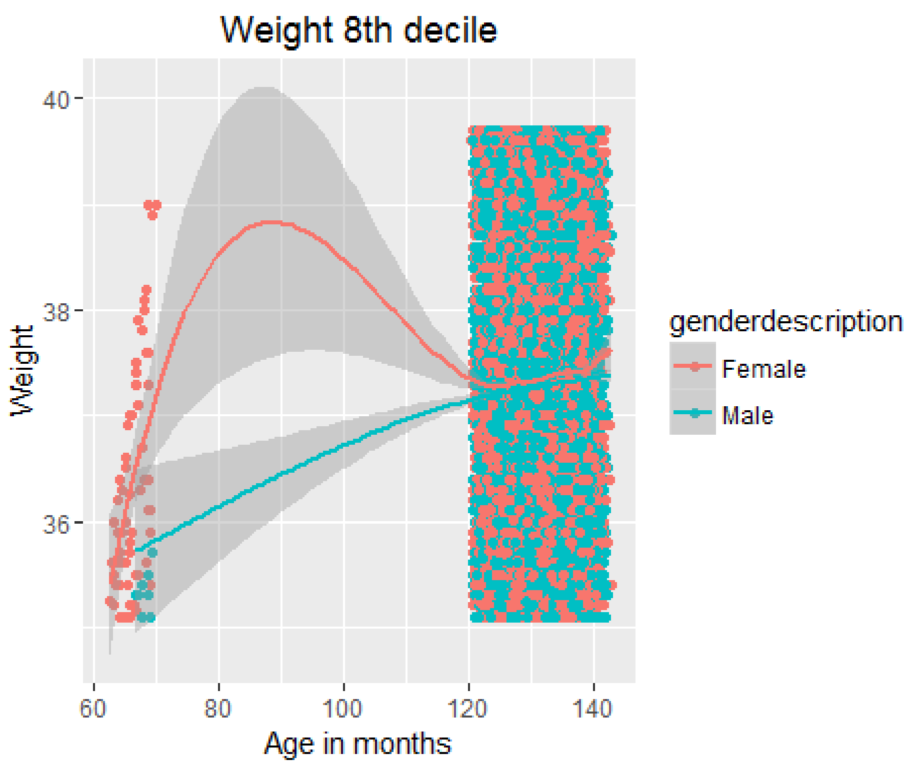


```r
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve
qplot(ageinmonths, log(weight), data=seventhdecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Seventh decile growth(in weight) curve", xlab="Age in months", ylab="Weight")
```

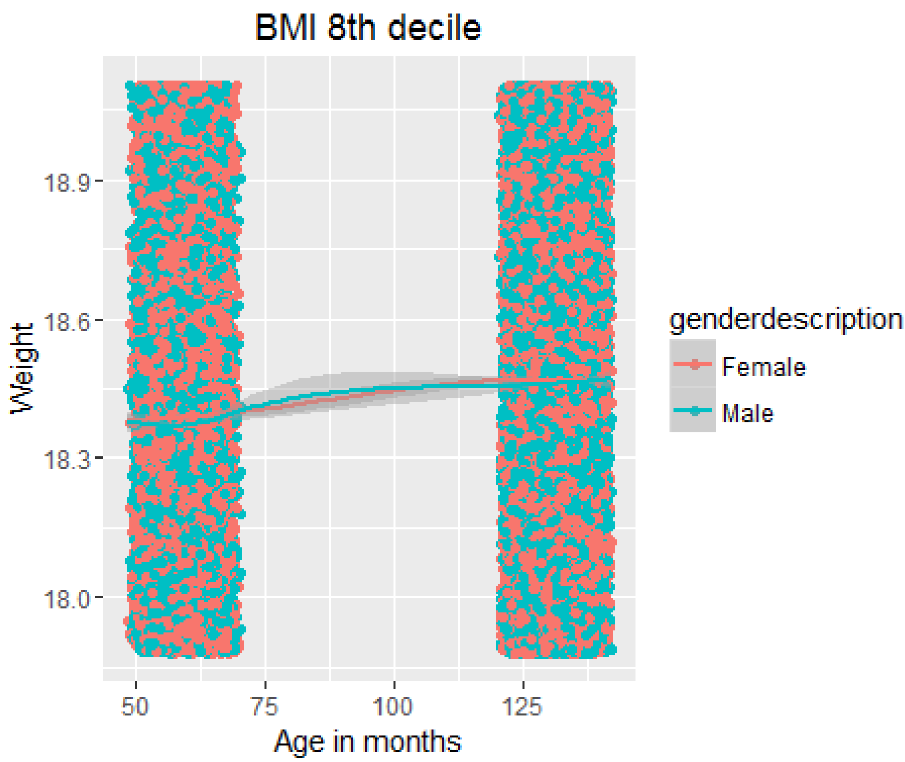

```r
qplot(ageinmonths, log(bmi), data=seventhdecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="Seventh decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```


```
#Using Linear modelling function to first perform Regression and get the
intercept for the growth curve
print(lm(log(seventhdecile_weight[2.862:2.87,"weight"])
~seventhdecile_weight[2.862:2.87,"ageinmonths"]))

##
## Call:
## lm(formula = log(seventhdecile_weight[2.862:2.87, "weight"]) ~
##     seventhdecile_weight[2.862:2.87, "ageinmonths"])
##
## Coefficients:
##                                      (Intercept)
##                                            3.487
## seventhdecile_weight[2.862:2.87, "ageinmonths"]
##                                               NA

print(lm(log(seventhdecile_bmi[2.685 :2.723,"bmi"]) ~seventhdecile_bmi[2.685
:2.723,"ageinmonths"]))

##
## Call:
## lm(formula = log(seventhdecile_bmi[2.685:2.723, "bmi"]) ~
seventhdecile_bmi[2.685:2.723,
##     "ageinmonths"])
##
## Coefficients:
##                                    (Intercept)
```

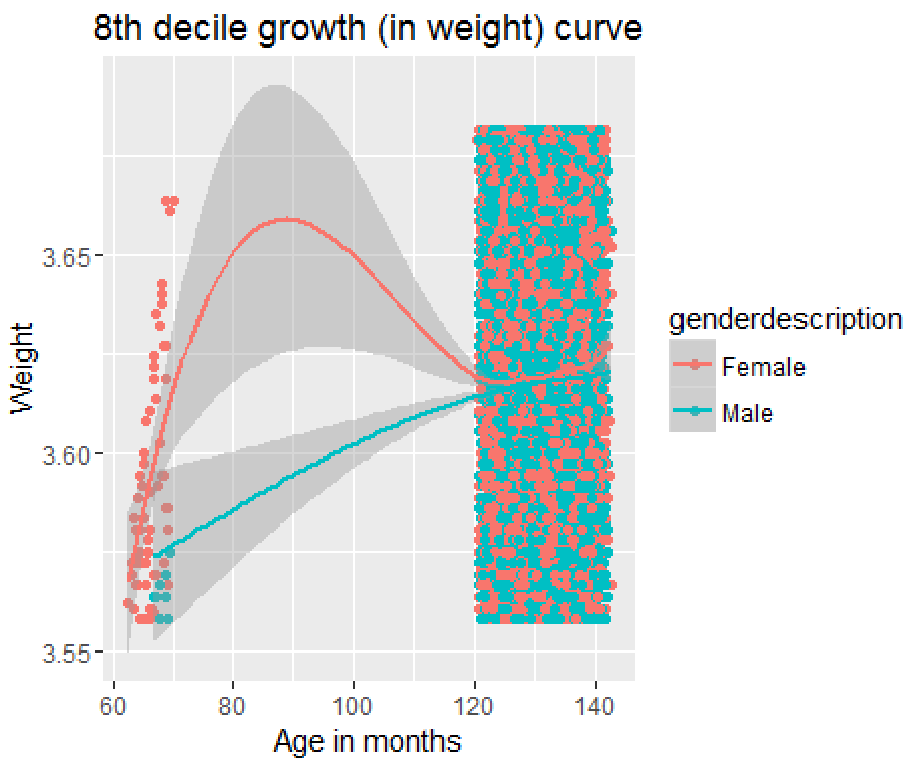
```
##                                                  2.877
## seventhdecile_bmi[2.685:2.723, "ageinmonths"]
##                                                     NA

#Making subsets to plot based on Male or Female for 8th decile (70% - 80%)

eightdecile_weight <-subset(obesityByweight, quartile ==8)
eightdecile_weight <-na.omit(eightdecile_weight)
eightdecile_bmi <-subset(obesityBybmi, quartile ==8)
eightdecile_bmi <-na.omit(eightdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender
qplot(ageinmonths, weight, data=eightdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight 8th decile", xlab="Age in
months", ylab="Weight")
```



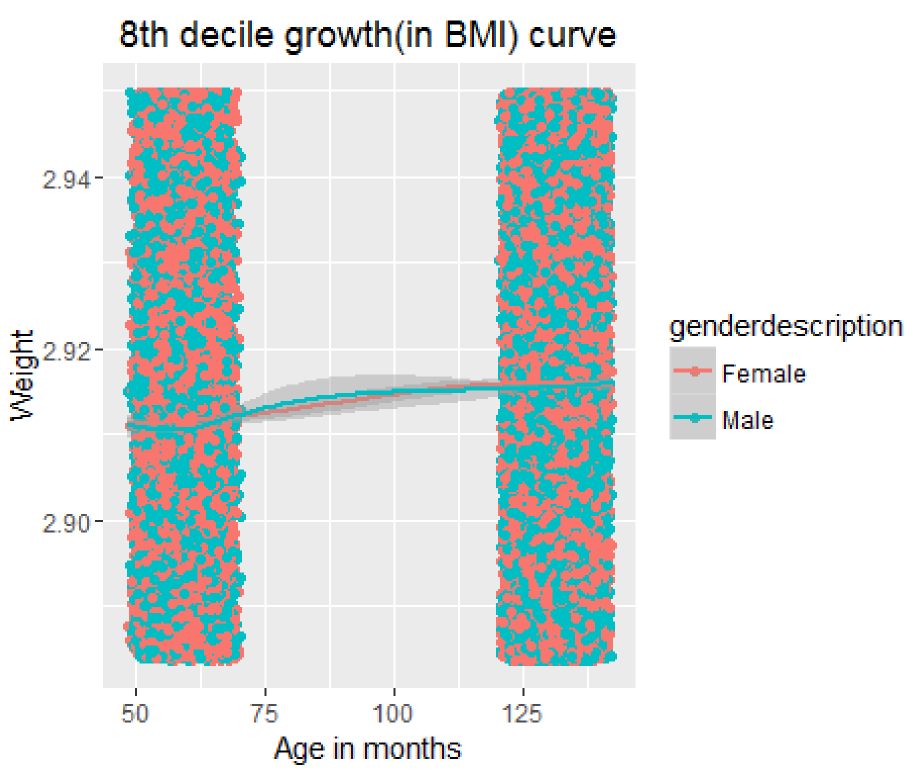

```
qplot(ageinmonths, bmi, data=eightdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI 8th decile", xlab="Age in months",
ylab="Weight")
```




```r
#Using the log of variables weight and bmi to plot against time (Age in
month) in order to get the growth curve
qplot(ageinmonths, log(weight), data=eightdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="8th decile growth (in weight)
curve", xlab="Age in months", ylab="Weight")
```




```r
qplot(ageinmonths, log(bmi), data=eightdecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="8th decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```



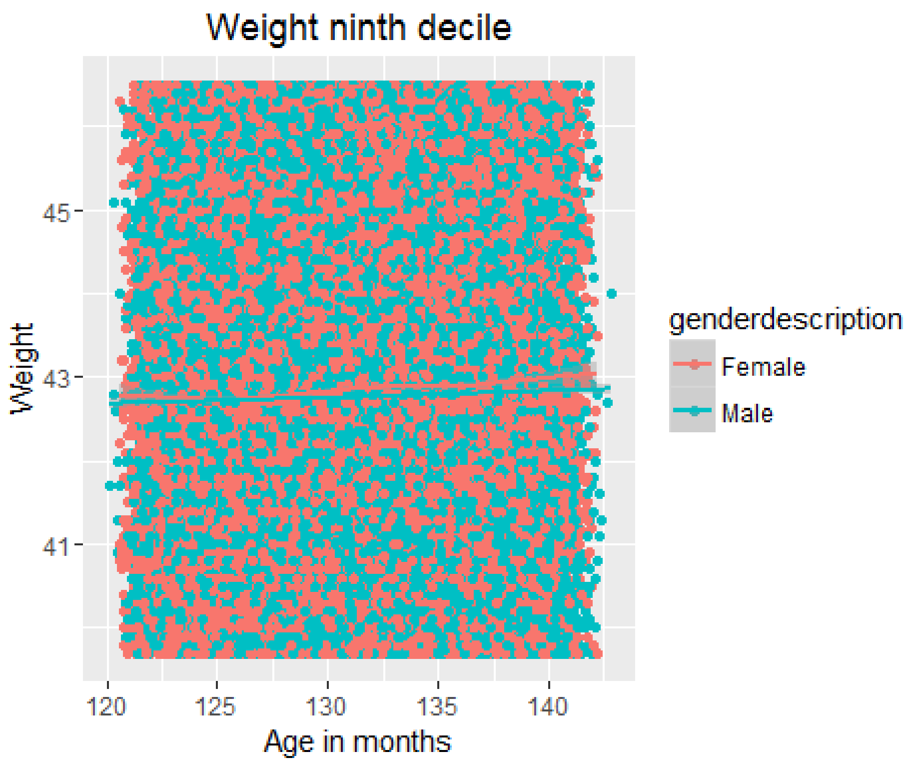
```r
#Using Linear modelling function to first perform Regression and get the intercept for the growth curve
print(lm(log(eightdecile_weight[2.93:2.95,"weight"])
~eightdecile_weight[2.93:2.95,"ageinmonths"]))
```

```
##
## Call:
## lm(formula = log(eightdecile_weight[2.93:2.95, "weight"]) ~
eightdecile_weight[2.93:2.95,
##     "ageinmonths"])
##
## Coefficients:
##                                   (Intercept)
##                                         3.578
## eightdecile_weight[2.93:2.95, "ageinmonths"]
##                                            NA
```

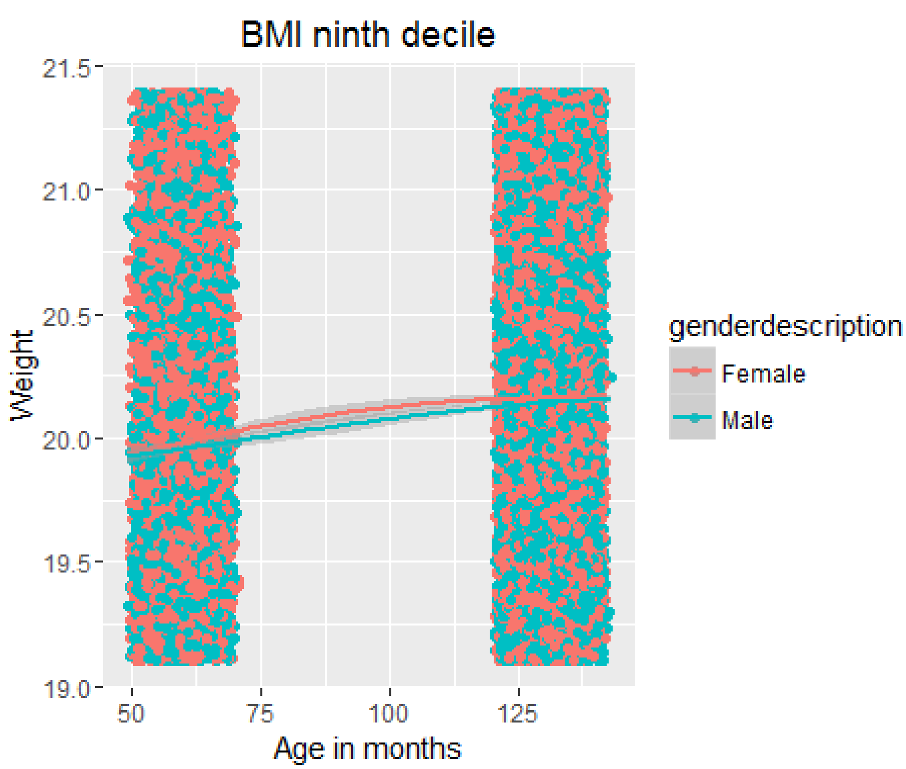
```r
print(lm(log(eightdecile_bmi[2.723 :2.753,"bmi"]) ~eightdecile_bmi[2.723
:2.753,"ageinmonths"]))
```

```
##
## Call:
## lm(formula = log(eightdecile_bmi[2.723:2.753, "bmi"]) ~
eightdecile_bmi[2.723:2.753,
##     "ageinmonths"])
##
## Coefficients:
##                                  (Intercept)
##                                        2.922
## eightdecile_bmi[2.723:2.753, "ageinmonths"]
##                                           NA
```

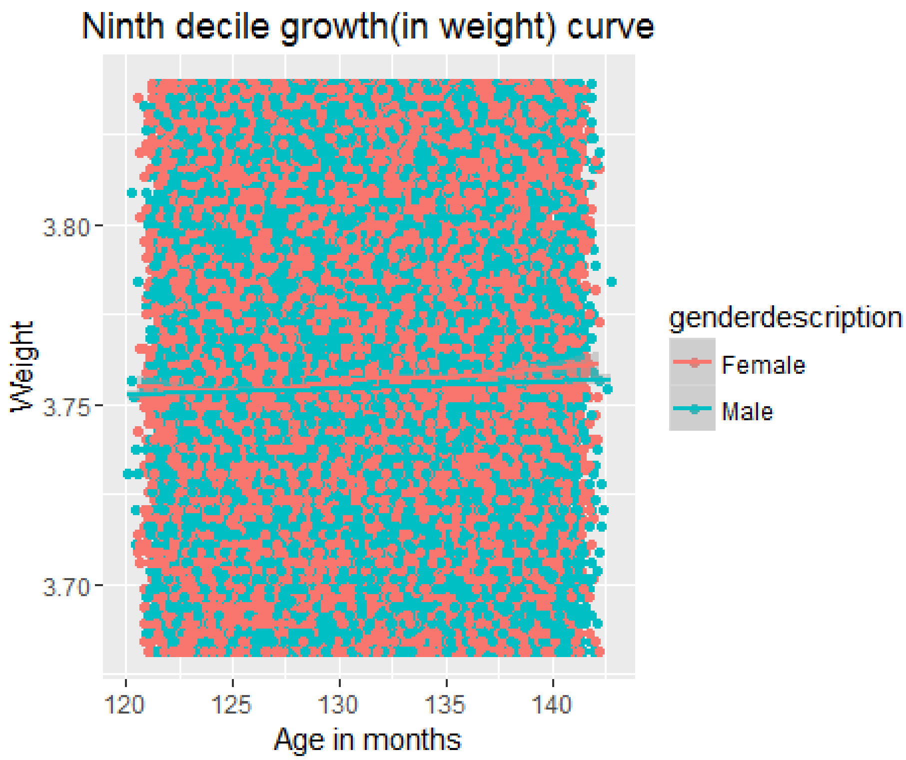
```r
#Making subsets to plot based on Male or Female for 9th decile (80% - 90%)

ninthdecile_weight <-subset(obesityByweight, quartile ==9)
ninthdecile_weight <-na.omit(ninthdecile_weight)
ninthdecile_bmi <-subset(obesityBybmi, quartile ==9)
ninthdecile_bmi <-na.omit(ninthdecile_bmi)



# Separate regressions of Ageinmonths on weight for each gender

qplot(ageinmonths, weight, data=ninthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight ninth decile", xlab="Age in
months", ylab="Weight")
```



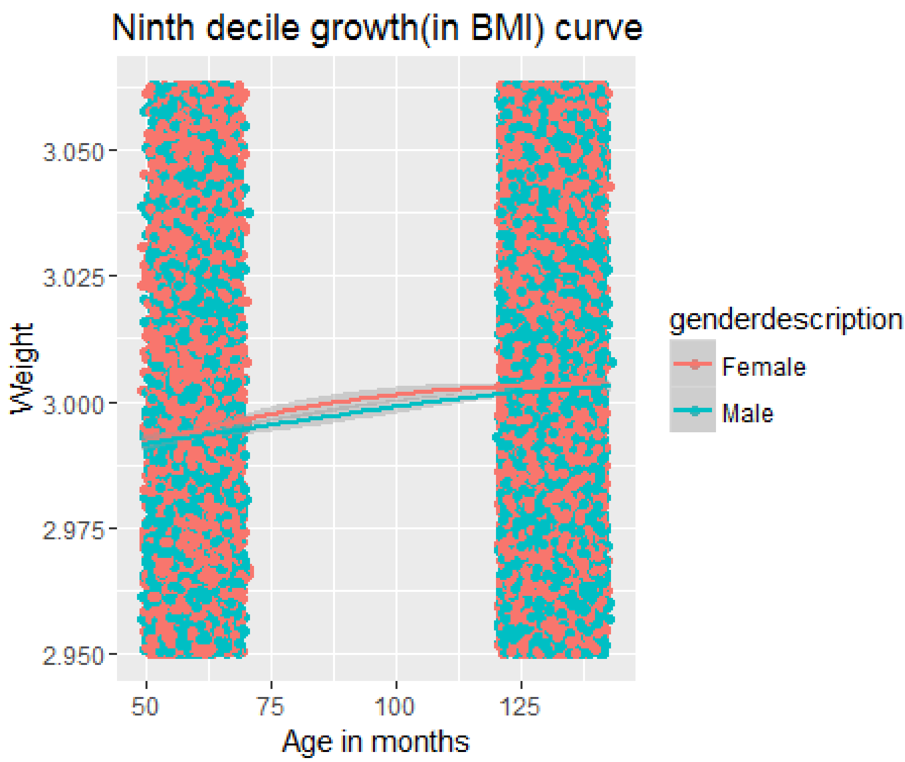

```
qplot(ageinmonths, bmi, data=ninthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI ninth decile", xlab="Age in months",
ylab="Weight")
```



```
#Using the log of variables weight and bmi to plot against time (Age in month) in order to get the growth curve

qplot(ageinmonths, log(weight), data=ninthdecile_weight, geom=c("point", "smooth"), color=genderdescription, main="Ninth decile growth(in weight) curve", xlab="Age in months", ylab="Weight")
```

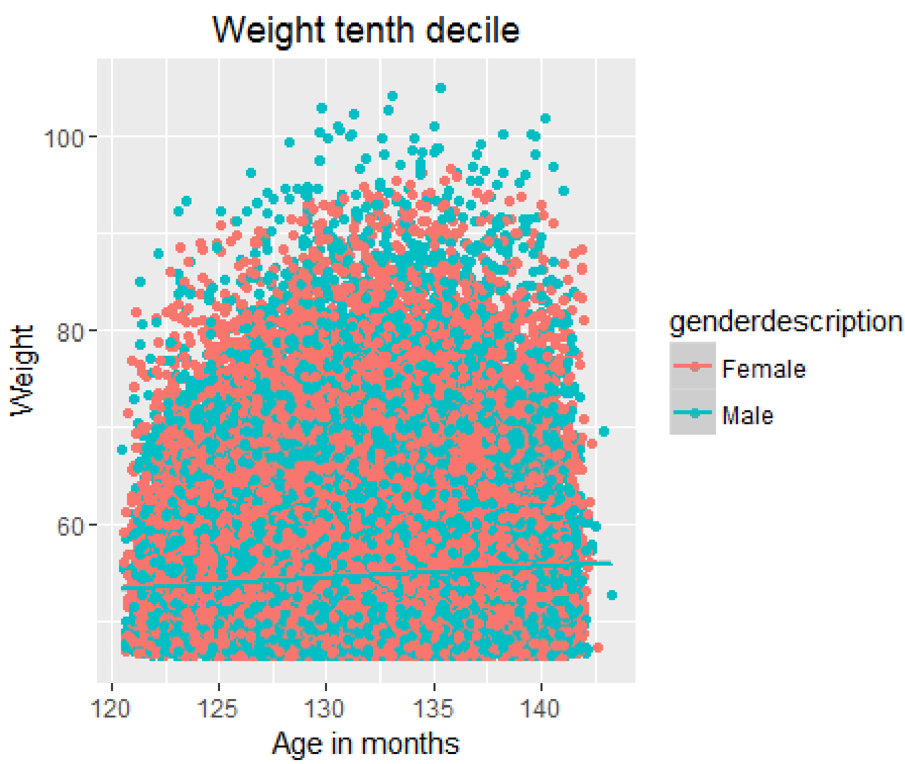



```
qplot(ageinmonths, log(bmi), data=ninthdecile_bmi, geom=c("point", "smooth"), color=genderdescription, main="Ninth decile growth(in BMI) curve", xlab="Age in months", ylab="Weight")
```

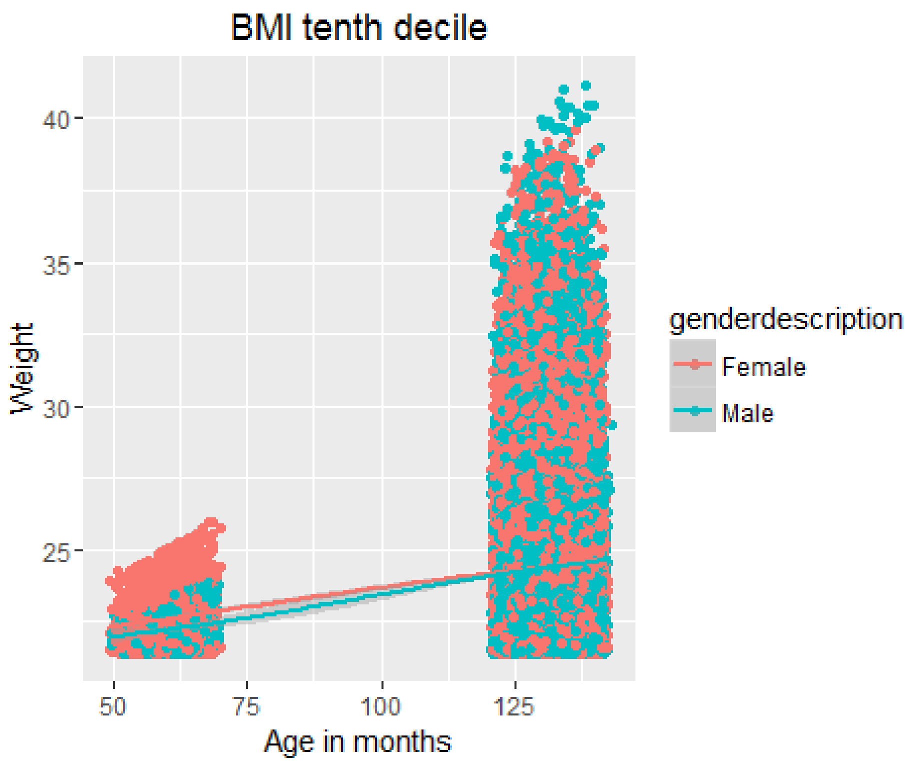



```
#Using Linear modelling function to first perform Regression and get the
intercept dor the growth curve
print(lm(log(ninthdecile_weight[2.5:2.8,"weight"])
~ninthdecile_weight[2.5:2.8,"ageinmonths"]))

##
## Call:
## lm(formula = log(ninthdecile_weight[2.5:2.8, "weight"]) ~
ninthdecile_weight[2.5:2.8,
##     "ageinmonths"])
##
## Coefficients:
##                                (Intercept)
##                                      3.752
## ninthdecile_weight[2.5:2.8, "ageinmonths"]
##                                         NA

print(lm(log(ninthdecile_bmi[2.5:2.8,"bmi"])
~ninthdecile_bmi[2.5:3,"ageinmonths"]))

##
## Call:
## lm(formula = log(ninthdecile_bmi[2.5:2.8, "bmi"]) ~ ninthdecile_bmi[2.5:3,
##     "ageinmonths"])
##
## Coefficients:
##                           (Intercept)
```

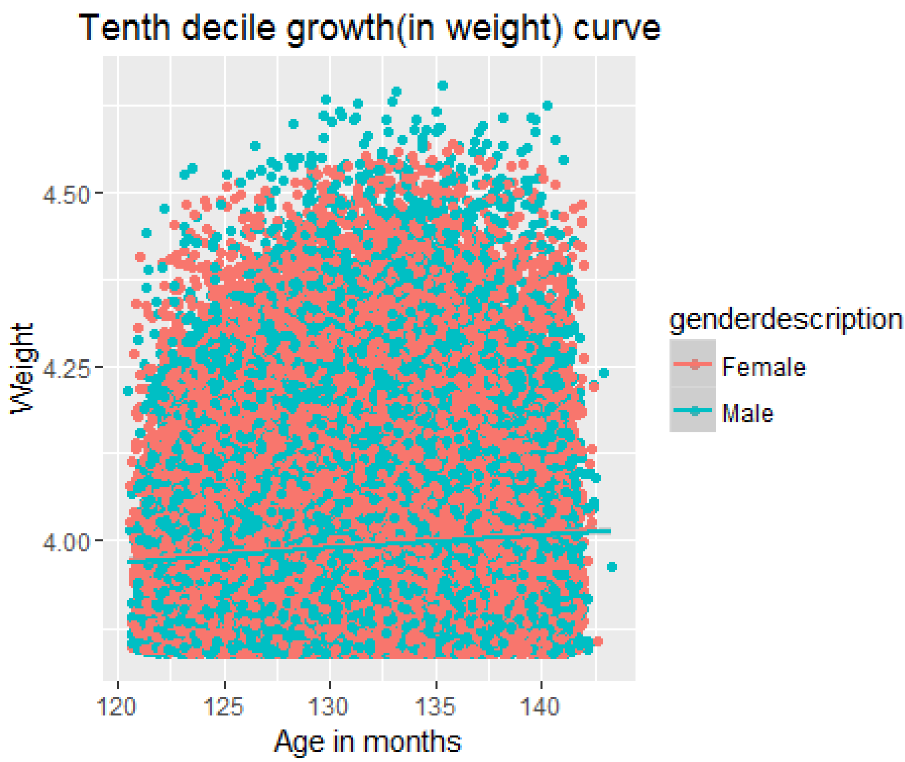
```
##                                    2.984
## ninthdecile_bmi[2.5:3, "ageinmonths"]
##                                       NA

#Making subsets to plot based on Male or Female for 10th decile (90% - 100%)

tenthdecile_weight <-subset(obesityByweight, quartile ==10)
tenthdecile_weight <-na.omit(tenthdecile_weight)
tenthdecile_bmi <-subset(obesityBybmi, quartile ==10)
tenthdecile_bmi <-na.omit(tenthdecile_bmi)


# Separate regressions of Ageinmonths on weight for each gender
qplot(ageinmonths, weight, data=tenthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Weight tenth decile", xlab="Age in
months", ylab="Weight")
```



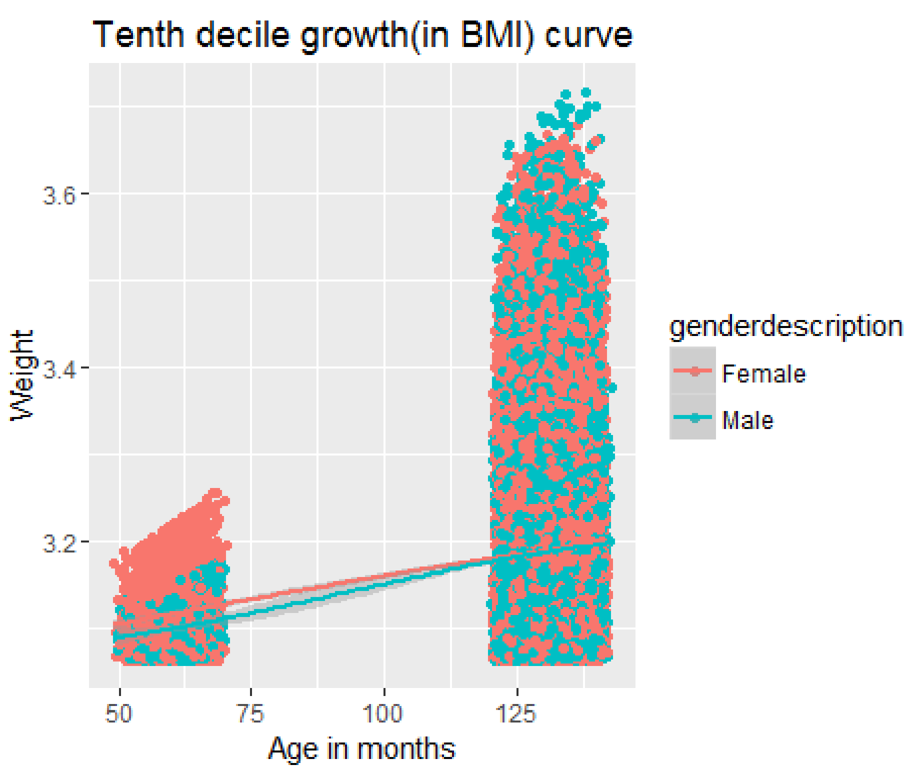

```
qplot(ageinmonths, bmi, data=tenthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="BMI tenth decile", xlab="Age in months",
ylab="Weight")
```




```
#Using the log of variables weight and bmi to plot against time (Age in
month) in order to get the growth curve
qplot(ageinmonths, log(weight), data=tenthdecile_weight, geom=c("point",
"smooth"), color=genderdescription, main="Tenth decile growth(in weight)
curve", xlab="Age in months", ylab="Weight")
```




```r
qplot(ageinmonths, log(bmi), data=tenthdecile_bmi, geom=c("point", "smooth"),
color=genderdescription, main="Tenth decile growth(in BMI) curve", xlab="Age
in months", ylab="Weight")
```



```r
#Using Linear modelling function to first perform Regression and get the
intercept for the growth curve
print(lm(log(tenthdecile_weight[2.862:2.87,"weight"])
~tenthdecile_weight[2.862:2.87,"ageinmonths"]))

##
## Call:
## lm(formula = log(tenthdecile_weight[2.862:2.87, "weight"]) ~
##     tenthdecile_weight[2.862:2.87, "ageinmonths"])
##
## Coefficients:
##                                    (Intercept)
##                                          3.875
## tenthdecile_weight[2.862:2.87, "ageinmonths"]
##                                             NA

print(lm(log(tenthdecile_bmi[2.685 :2.723,"bmi"]) ~tenthdecile_bmi[2.685
:2.723,"ageinmonths"]))

##
## Call:
## lm(formula = log(tenthdecile_bmi[2.685:2.723, "bmi"]) ~
tenthdecile_bmi[2.685:2.723,
##     "ageinmonths"])
##
## Coefficients:
##                                  (Intercept)
##                                         3.27
## tenthdecile_bmi[2.685:2.723, "ageinmonths"]
##                                           NA
```

6. Attend the Tartu city open data event in the Garage48 hub on Saturday (March 5th, 2pm). Participate in some working group. Write a report about the ideas, problem statement, envision the goals of the project and main steps that need to be executed in order to achieve the stated goals.

Kindergartens, Schools, planning…

Group: JaakVilo, OmisakinOluwatobi Samuel, Valdur Kana, MykhailoDorokhov, Kenigbolo Meya Stephen, Darwin Sivalingapandi, JevgeniSavostkin, FortunatMutunda

What is the need for DATA?

- **Population Registry** - all people by precise **age** (of children)**, address,** and language (?) *It can be aggregated for privacy reasons…* (do many siblings affect the need - siblings should go to same (kindergarten/schools)

- **Kindergarten sizes and age distribution of kids currently in** - age groups/cohorts - how many admitted and will be leaving at any given time point (know the number of available places upfront)
- **Waiting list** actual status of the kindergarten (who has signed up - what address, when wants place, how many have already signed up - how many may be missing)
- **Where do kids actually go to kindergarten** now - to understand the real "waste" travel (should siblings go to same school?)
- **New building permits** - how many new families going to move in in short time from now… (match the future demand)
- Maybe some data about **playgrounds, transport,** …
- **Public aggregated data snapshots made available for planning needs**… for businesses to build more efficient services (based on better planning)

WHY

I Planning needs: Predict and Match - demand and supply. Match a demand and supply 2-3 years ahead of actual time … All the time real time changes.

Observe the dynamic changes - observe or predict who moves in, where, how does that change. (Time series, make predictions on trends)

New building permits - how many flats, what type of flats being built, what is GOING TO BE a new demand.

Goal: aid the city and urban planning for Tartu. (new kindergartens, closing, places volume tuning, new teachers to be hired, etc.).

II Teleport like scenario: like companies could take this to the global level examples. - Where and how to move in, what are the costs and options for new people…

Family looks for a place to live, they should know number of places in kindergartens/schools available and some qualitative criteria: Estonian schools vs schools when it is possible to learn Estonian not being a native speaker. (Probably more valuable for international case)

III SWAPPING service: Are there parents/families that might want to switch the places? Find and propose those switches. E.g. after family moved to a new location their travel could have been made too long… And some other may be there available…

Travel distances and SWAPPING: service School catchment areas to calculate actual travel distances to kindergartens, schools.

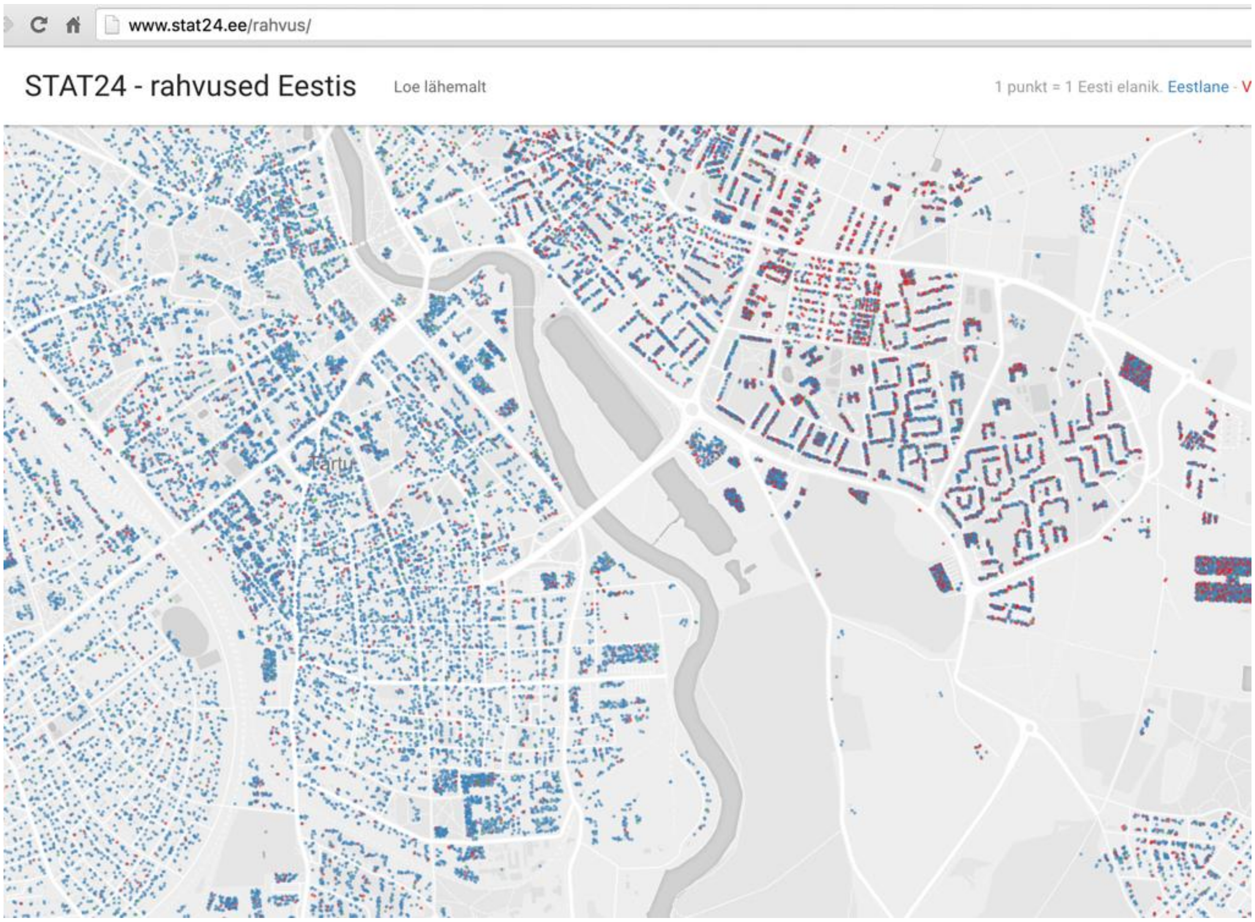
Proximity to school/kindergarten. **What are other factors - e.g. workplaces affecting choice of kindergarten**? Local demand can also be higher if many people work nearby…

Research issues - how to ensure the privacy, differential privacy queries, etc. Play the different scenarios with models and predictions.

Can it be hacked together at Hackathon?

Yes, it can :)

Example of visualisations of Estonians and Russians based purely on open data - census counts per postal code and open street map…



Public access - http://eid.ee/3z7

My personal conclusions

The report above is the final document that summarizes our discussions (barring some editing I made to the doc) however I will like to summarize and highlight some important factors about the topic.

Considering the intricate nature of the data we require, there definitely will be some issues surrounding policy and privacy. In as much as our intentions are good, having such sensitive data publicly could play into the hands of a wrong set of people (for example a pedophile interested in finding out where a large concentration of kids stay or a criminal getting to find out where the newest apartment buildings in the city are located) however as with any data made publicly available the possibility of such questions arising is quite high.

My suggestion will be the use of snapshots that will cover a specific time period. These snapshots could be monthly, quarterly or yearly and will help in comparing the trends in several situations.

Also this data can and most probably will be useful (if used and implemented via statistics) for the city council and government in budgeting as they will be able to properly plan for (if necessary) opening a new Kindergarten or school (the amount of kids to attend to etc.). It could also help marketers to know exactly where to position a business (for example, it makes sense to have an Ice cream shop close to a school or an amusement park close to where there are a large number of children).