In [2]:
```
library(NHANES)
library(ggplot2)
df1 = NHANESraw
df = df1[,c("BPSysAve","Age","Weight","Height", "WTMEC2YR","Gender")]
```

In [9]:
```
df = df[complete.cases(df, df$BPSysAve),]
colSums(is.na(df))
dim(df)
```

BPSysAve	0
Age	0
Weight	0
Height	0
WTMEC2YR	0
Gender	0

14720 6

We have dropped around 25% of our data because we didn't have BPSysAve value in it.

In [10]:
```
s = sum(df$WTMEC2YR)
df$WTMEC2YR = df$WTMEC2YR/s
```

In [11]:
```
ggplot(df, aes(x=Age,y=BPSysAve,color=Gender,size=WTMEC2YR)) +
  geom_point()  +
  geom_smooth(method.args= list(degree=1),span=0.75) +
  facet_wrap(~(cut_number(Height, n = 2)) + ~(cut_number(Weight,n = 2))) +
ylab("Average Systolic Blood Pressure") +
ggtitle("Average BP vs Age- Weight increases from left to right and Height inc
reases from Top to Bottom")
```

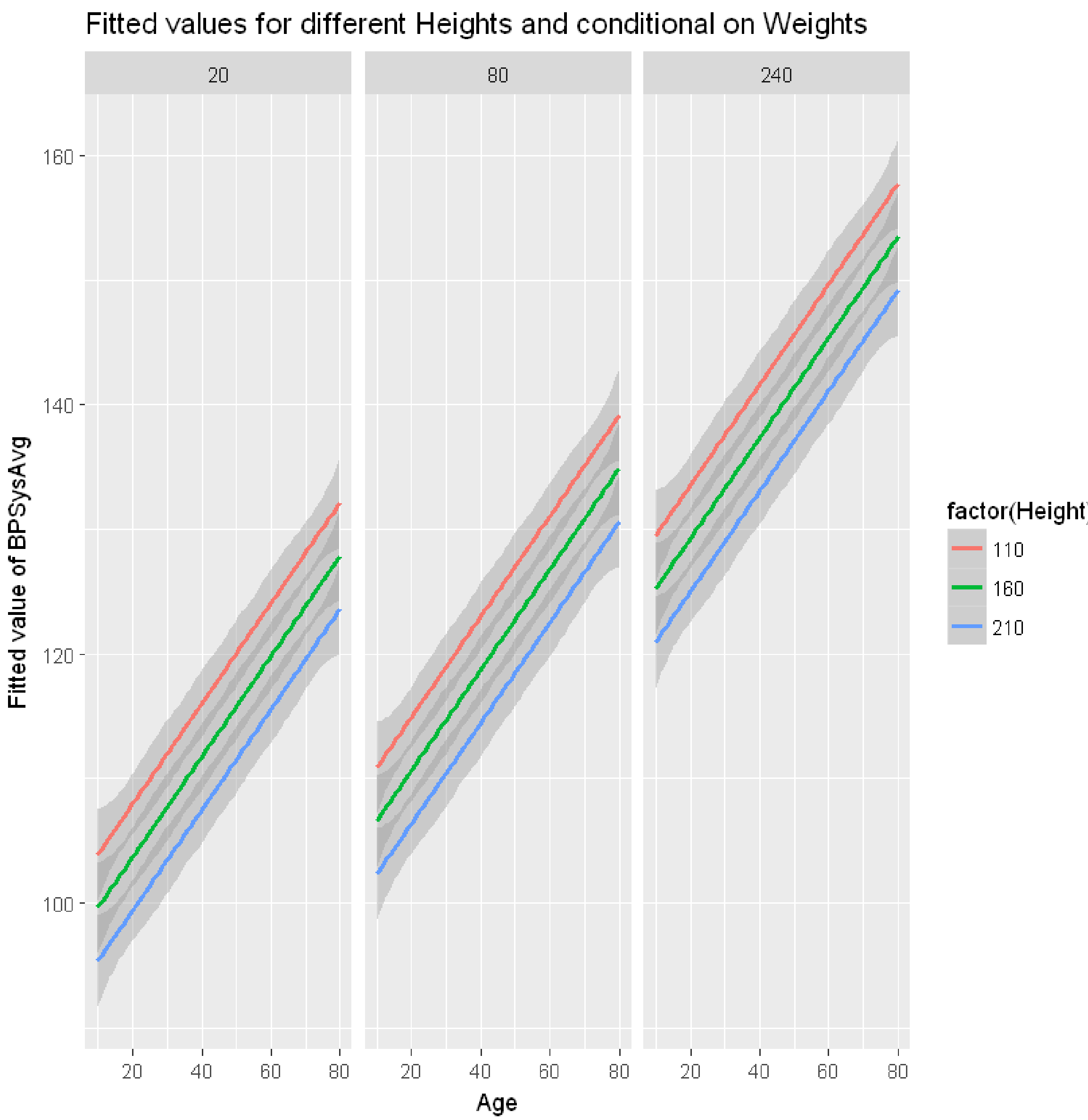
```
`geom_smooth()` using method = 'gam'
```

So there is an age and gender factor but the BPSysAve remains the same in all the heighs and weights group and there is an additive shift within gender.

In [17]:
```
df.lm1 = lm(BPSysAve ~ Age + Gender + Height + Weight  , data = df, weights =W
TMEC2YR )
new.grid = expand.grid(Age = seq(10,80,10), Gender = c("female", "male"), Heig
ht = c(110,160,210), Weight = c(20,80,240))
new.predict = predict(df.lm1, newdata = new.grid)
new.total = data.frame(new.grid, fit = as.vector(new.predict))
```

In [18]:
```
ggplot(new.total, aes(x=Age,y=fit, group = Height, color = factor(Height))) +
   geom_smooth() + facet_wrap(~Weight) + ylab("Fitted value of BPSysAvg") +
ggtitle("Fitted values for different Heights and conditional on Weights")
```

```
`geom_smooth()` using method = 'loess'
```



We know that there is just an additive shift for Male and Female BPSysValue. So, we didn't show that in this plot.

We can observe that as the age increases the BPSysValue increases and the increase in value(slope) is constant for different Heights and Different Weights. Also, people who are tall or heavy weighted have their BPSysValue much higher than that of their same aged peers. This makes sense as for taller and Heavy person it takes much more pressure(Systolic Pressure from heart) to pump the blood to their farthest organ from heart.

Q2

In [23]:
```
library(NHANES)
library(ggplot2)
df2 = NHANESraw
dfd = df2[,c("Diabetes","Age","Weight","Height", "WTMEC2YR","Gender")]
dim(dfd)
colSums(is.na(dfd))
dfd = dfd[complete.cases(dfd, dfd$Height),]
colSums(is.na(dfd))
dim(dfd)
```

20293 6

Diabetes	833
Age	0
Weight	888
Height	2258
WTMEC2YR	0
Gender	0

Diabetes	0
Age	0
Weight	0
Height	0
WTMEC2YR	0
Gender	0

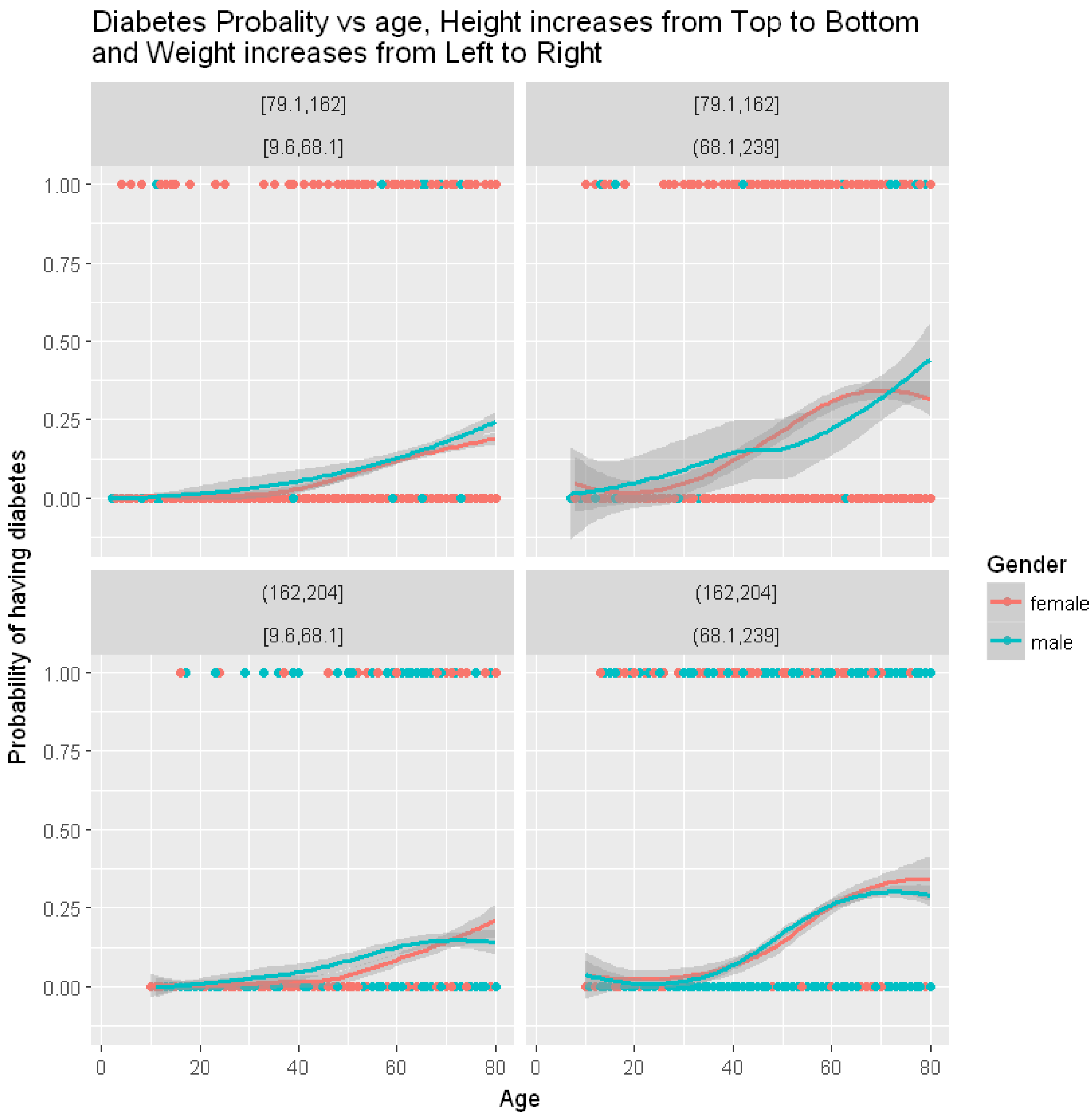
18005 6

We have dropped around 10% of our data as we don't have Height or Diabetes or weight values

In [24]:
```
dfd$Diabetes = as.numeric(dfd$Diabetes)

dfd$Diabetes = dfd$Diabetes - 1
```

In [25]:
```
ggplot(dfd,aes(x=Age, y= Diabetes, color = Gender)) + geom_point() +
geom_smooth(method = "loess")+
facet_wrap(~(cut_number(Height, n = 2)) + ~(cut_number(Weight,n = 2))) +
ylab("Probability of having diabetes")+
ggtitle("Diabetes Probality vs age, Height increases from Top to Bottom
and Weight increases from Left to Right")
```

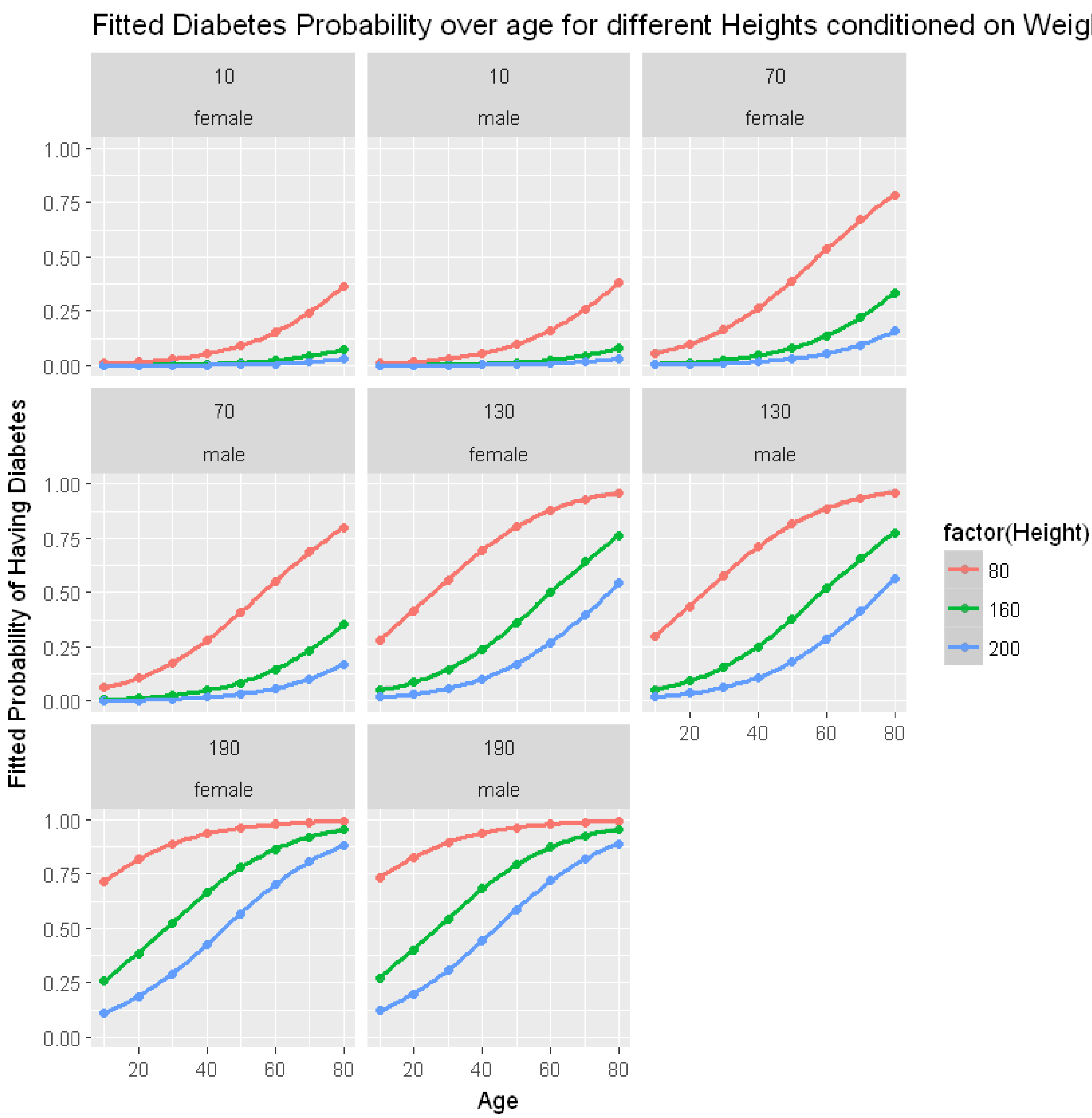


Less number of observations for female are there in the upper right corner that's why the curve is somewhat distorted for different genders. So, we just need a model which has age, Height and Weight interaction with Gender.

In [26]:
```
dfd.logit = glm(Diabetes ~ Age + Gender + Height + Weight, family = "binomial"
, data = dfd)
new.grid = expand.grid(Age = seq(10,80,10), Gender = c("female", "male"), Heig
ht = c(80,160,200), Weight = seq(10,240,60))
dfd.pred = predict(dfd.logit, type = "response", newdata = new.grid)
dfd.pred.df = data.frame(new.grid, fit = as.vector(dfd.pred))
```

In [28]:
```
ggplot(dfd.pred.df, aes(x=Age,y=fit,color = factor(Height))) + geom_point() +
   geom_smooth() +
facet_wrap(~(Weight) + ~(Gender)) +
ylab("Fitted Probability of Having Diabetes") +
ggtitle("Fitted Diabetes Probability over age for different Heights conditione
d on Weights")
```

```
`geom_smooth()` using method = 'loess'
```



We can observe that Diabetes Probability is not depending upon Gender as for Female and Male the distribution is similar.

Over the age the Diabetes Probability of a person increases.

The taller person's Diabetes Probability is more than that of his/her same aged and same weighted person.

The more the weight of a person the higher the Diabetes Probability.

Q3

In [7]:
```
df2 =  NHANESraw
df3 = df2[,c("Diabetes","Age","Weight","Height", "WTMEC2YR","Gender", "HHIncom
eMid", "Poverty", "Pulse", "DirectChol", "TotChol")]
```

In [8]:
```
df3 = df3[,c("Diabetes","Age","Weight","Height","Gender", "Pulse")]
```

In [9]:
```
dim(df3)
colSums(is.na(df3))
```

20293 6

Diabetes	833
Age	0
Weight	888
Height	2258
Gender	0
Pulse	5397

In [10]:
```
df3 = df3[complete.cases(df3, df3$Pulse),]
```

In [11]:
```
dim(df3)
```

14737 6

Taken Pulse as one more explanatory variable as it is more Biological variable.

As we know that we are dropping around 25% (5397/20293) of data we are loosing so much of the information as compared to that of the above model where we dropped only 10% of the data.

In [12]:
```
df3$Diabetes = as.numeric(df3$Diabetes)

df3$Diabetes = df3$Diabetes - 1

summary(df3)
dim(df3)
```

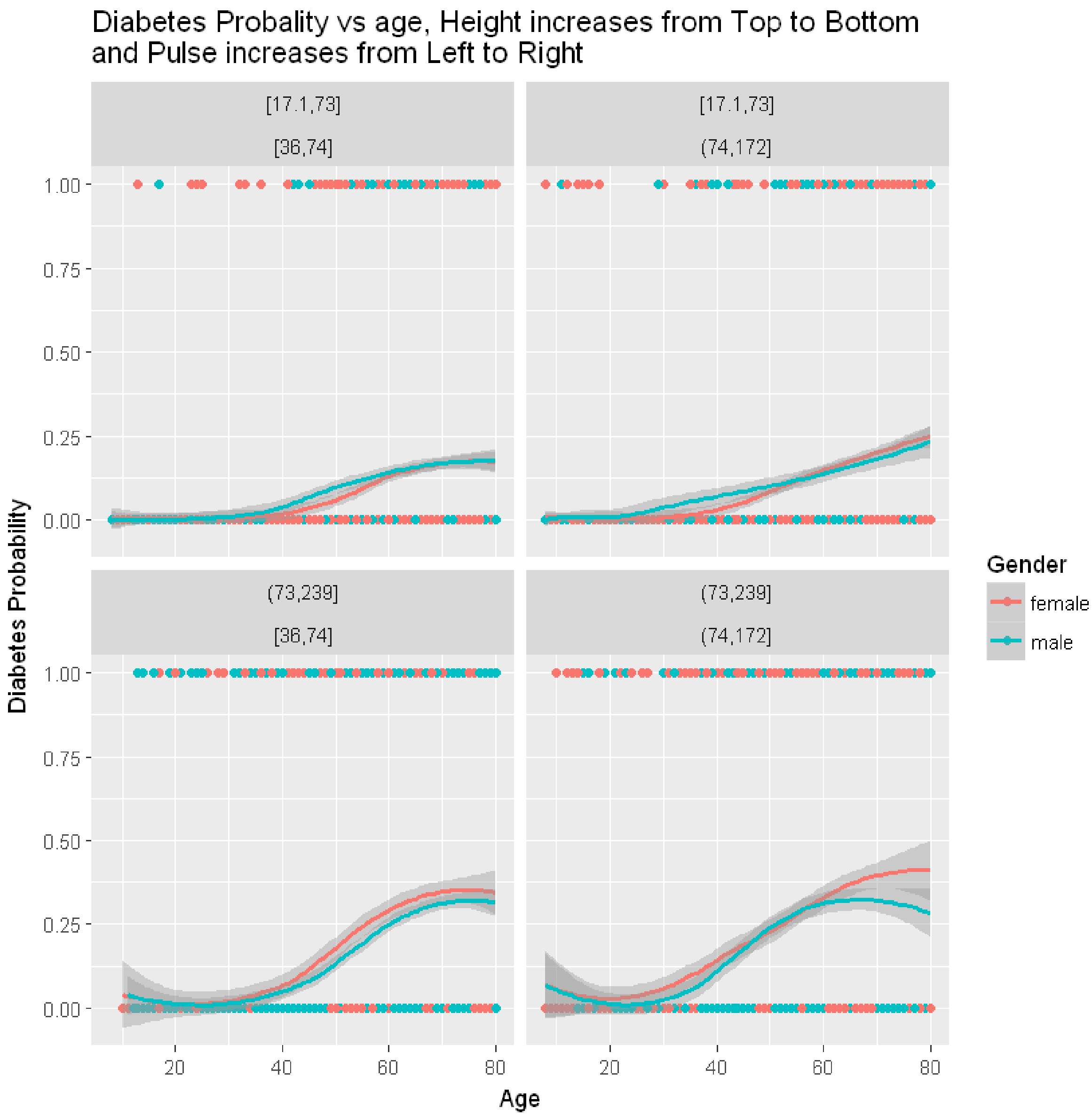
```
    Diabetes           Age            Weight           Height     
 Min.   :0.0000   Min.   : 8.00   Min.   : 17.10   Min.   :112.5  
 1st Qu.:0.0000   1st Qu.:18.00   1st Qu.: 58.70   1st Qu.:156.7  
 Median :0.0000   Median :38.00   Median : 73.00   Median :165.1  
 Mean   :0.1065   Mean   :39.39   Mean   : 74.49   Mean   :164.2  
 3rd Qu.:0.0000   3rd Qu.:58.00   3rd Qu.: 88.40   3rd Qu.:173.2  
 Max.   :1.0000   Max.   :80.00   Max.   :239.40   Max.   :204.5  
    Gender          Pulse       
 female:7383   Min.   :  0.00  
 male  :7354   1st Qu.: 66.00  
               Median : 74.00  
               Mean   : 74.04  
               3rd Qu.: 82.00  
               Max.   :172.00  
```

14737 6

Drop those rows in which the pulse rate is zero. As it pulse of a living person cannot be zero.

In [13]:
```
df3 = df3[df3["Pulse"] != 0,]
```

In [16]:
```
ggplot(df3,aes(x=Age, y= Diabetes, color = Gender)) + geom_point() + geom_smoo
th(method = "loess")+
facet_wrap(~(cut_number(Weight, n = 2)) + ~(cut_number(Pulse,n = 2))) +
ylab("Diabetes Probability") +
ggtitle("Diabetes Probality vs age, Height increases from Top to Bottom
and Pulse increases from Left to Right")
```



There is just an additive shift with in a gender.

Similarly there is an additive shift for different pulse sections for same weight range.

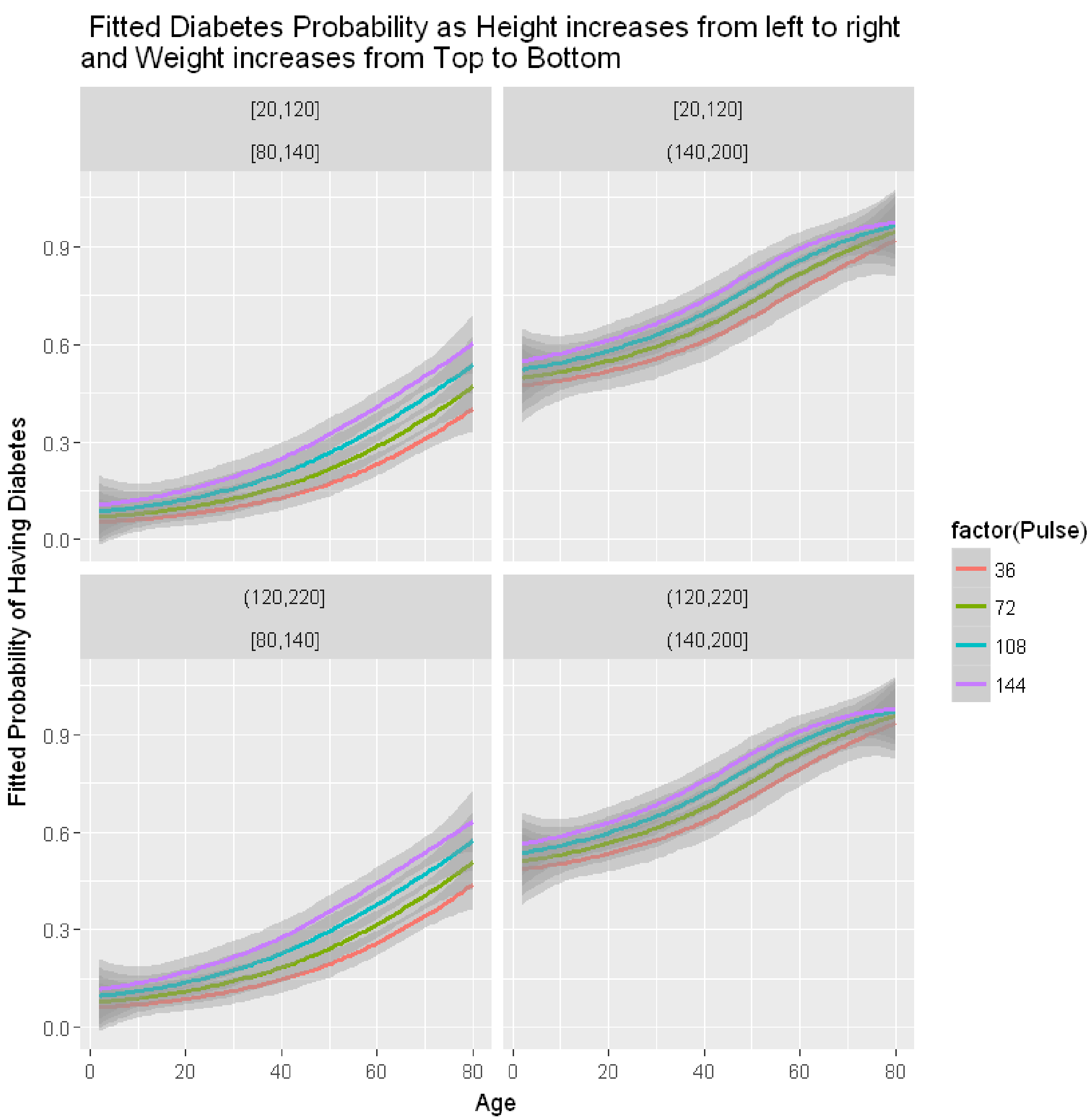
In [30]:
```
df3.logit = glm(Diabetes ~ (Age*Weight*Height) +Pulse + Gender, family = "bino
mial", data = df3)
```

In [18]:
```
df3.g = expand.grid(Age = seq(2, 80, 2), Pulse = c(36, 72, 108, 144),  Height
= seq(80,200,40),
                    Weight = seq(20, 240, 40) , Gender = c("female","male"))
```

In [19]:
```
fit = predict(df3.logit, type = "response", newdata = df3.g)
df3.pred.df = data.frame(df3.g, Diabetes = as.vector(fit))
```

In [29]:
```
ggplot(df3.pred.df, aes(x=Age, y = fit, color = factor(Pulse))) + geom_smooth
() +
facet_wrap(~(cut_number(Weight, n = 2))+ ~(cut_number(Height, n = 2)))+
ylab("Fitted Probability of Having Diabetes") +
ggtitle(" Fitted Diabetes Probability as Height increases from left to right
and Weight increases from Top to Bottom")
```

```
`geom_smooth()` using method = 'loess'
```



We can say that weight does not have much effect on the Diabetes probability where as the Height has.

There is an additive shift for different pulse values the higher the pulse the more the probability of having Diabetes.

This is the major thing that the interaction with Pulse is explaining.

Also, there is linear increase in the Diabetes Percentage over the age.