https://www.youtube.com/watch?v=1keotmoOWo4 (https://www.youtube.com/watch?v=1keotmoOWo4)

Ch4. Exploring data with graphs

The art of presenting data

Packages used in this chapter

```
#install.packages("ggplot2")library(ggplot2)
```

Introducing ggplot2

The anatomy of a plot

In ggplot2, a plot is made up of layers

The anatomy of a graph

Geometric objects (geoms)

https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf (https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf)

http://docs.ggplot2.org/current/ (http://docs.ggplot2.org/current/)

```
geom_.....()
```

Aesthetics

Linetype, Size, Shape, Colour, Alpha

The anatomy of the ggplot() function

myGraph <- ggplot(myData, aes(variable for x axis, variable for y axis))myGraph <- ggplot(myData, aes(variable for x axis, variable for y axis, colour = gender))+ ops(title = "Title") => ggtitle("New Plot Title")

myGraph + geom_bar()myGraph + geom_bar() + geom_point()myGraph <- ggplot(myData, aes(variable for x axis, variable for y axis, colour = gender))myGraph + geom_point(colour = "Blue")myGraph + geom_point(shape = 17, colour = "Blue")myGraph + geom_bar() + geom_point() + labels(x = "Text", y = "Text")

In [ ]:

```
```

Stats and geoms

In [ ]:

```
myHistogram <- ggplot(myData, aes(variable))
```

In [ ]:

```
myHistogram + geom_histogram()
```

In [ ]:

```
myHistogram + geom_histogram(aes(y = ..count..))
```

In [ ]:

```
myHistogram + geom_histogram(aes(y = ..density..))
```

In [ ]:

```
myHistogram + geom_histogram(aes(y = ..count..), binwidth = 0.4)
```

In [ ]:

```
```

Avoiding overplotting

1. position adjustment

2. faceting

In [ ]:

```
+ facet_wrap( ~ y, nrow = integer, ncol = integer)
```

In [ ]:

```
+ facet_grid( x ~ y )
```

In [ ]:

```
facet_grid(gender ~ extroversion)
```

In [ ]:

```
+ facet_wrap( ~ Rating_Type)
```

In [ ]:

```
+ facet_wrap( ~ Rating_Type, ncol = 2)
```

In [ ]:

```
+ facet_wrap( ~ Rating_Type, nrow = 2)
```

Saving graphs

In [ ]:

```
ggsave(filename)
```

In [ ]:

```
ggsave("Outlier Amazon.png")
```

In [ ]:

```
ggsave("Outlier Amazon.tiff")
```

In [ ]:

```
ggsave("Outlier Amazon.tiff", width = 2, height = 2)
```

In [ ]:

In [ ]:

In [1]:

```
library(ggplot2)
```

FacebookNarcissism.dat

In [3]:

```
facebookData <- read.delim("FacebookNarcissism.dat", header = TRUE)
```

In [4]:

```
head(facebookData)
```

Out[4]:

	id	NPQC_R_Total	Rating_Type	Rating
1	1	31	Attractive	2
2	1	31	Fashionable	2
3	1	31	Glamourous	2
4	1	31	Cool	2
5	2	37	Attractive	2
6	2	37	Fashionable	2

In [5]:

```
str(facebookData)
```

```
'data.frame':   776 obs. of  4 variables:
 $ id          : int  1 1 1 1 2 2 2 2 5 5 ...

 $ NPQC_R_Total: num  31 31 31 31 37 ...
 $ Rating_Type : Factor w/ 4 levels "Attractive","Cool",..: 1 3
4 2 1 3 4 2 1 3 ...
 $ Rating      : int  2 2 2 2 2 2 2 2 3 3 ...
```

In [6]:

```
summary(facebookData)
```

Out[6]:

```
       id          NPQC_R_Total        Rating_Type      Rating
 Min.   :  1.0   Min.   :14.00   Attractive :194   Min.   :1.00
0
 1st Qu.: 78.0   1st Qu.:27.44   Cool       :194   1st Qu.:2.00
0
 Median :139.0   Median :33.00   Fashionable:194   Median :3.00
0
 Mean   :141.9   Mean   :33.16   Glamourous :194   Mean   :2.88
8
 3rd Qu.:208.0   3rd Qu.:38.00                     3rd Qu.:3.00
0
 Max.   :275.0   Max.   :52.00                     Max.   :5.00
0
```

contents of the dataframe

1. id : a number indicating from which participant the profile photo came
2. NPQC_R_Total : the total score on the narcissism questionnaire.
3. Rating_Type : whether the rating was for coolness, glamour, fashion or atttractiveness (stored as strings of text)
4. Rating : the rating given (on a scale from 1 to 5)

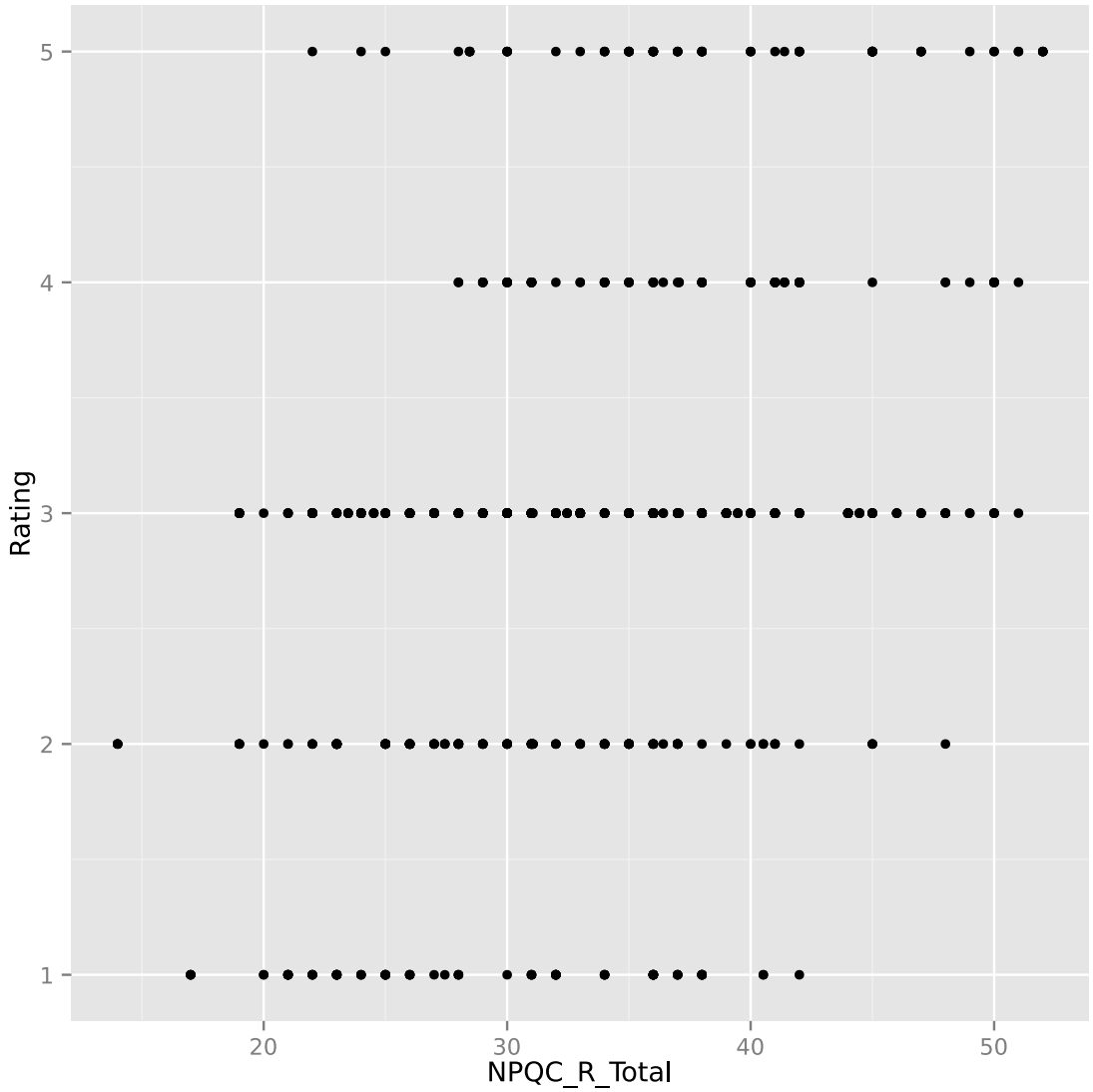
In [10]:

```
graph <- ggplot(facebookData, aes(NPQC_R_Total, Rating))
```

In [11]:

```
graph + geom_point()
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



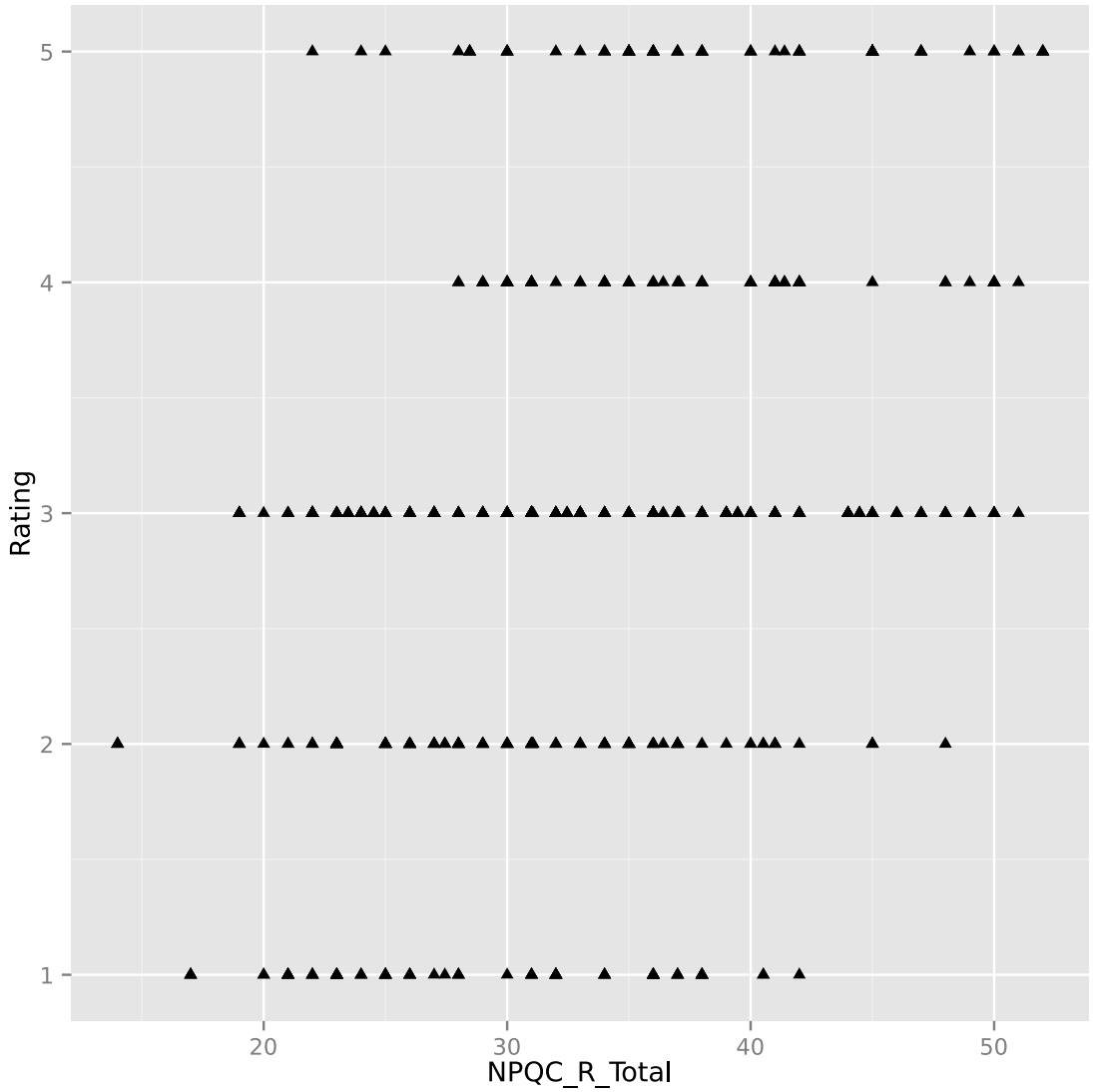
In [13]:

```
# change the shape of point
```

In [15]:

```
graph + geom_point(shape = 17)
```

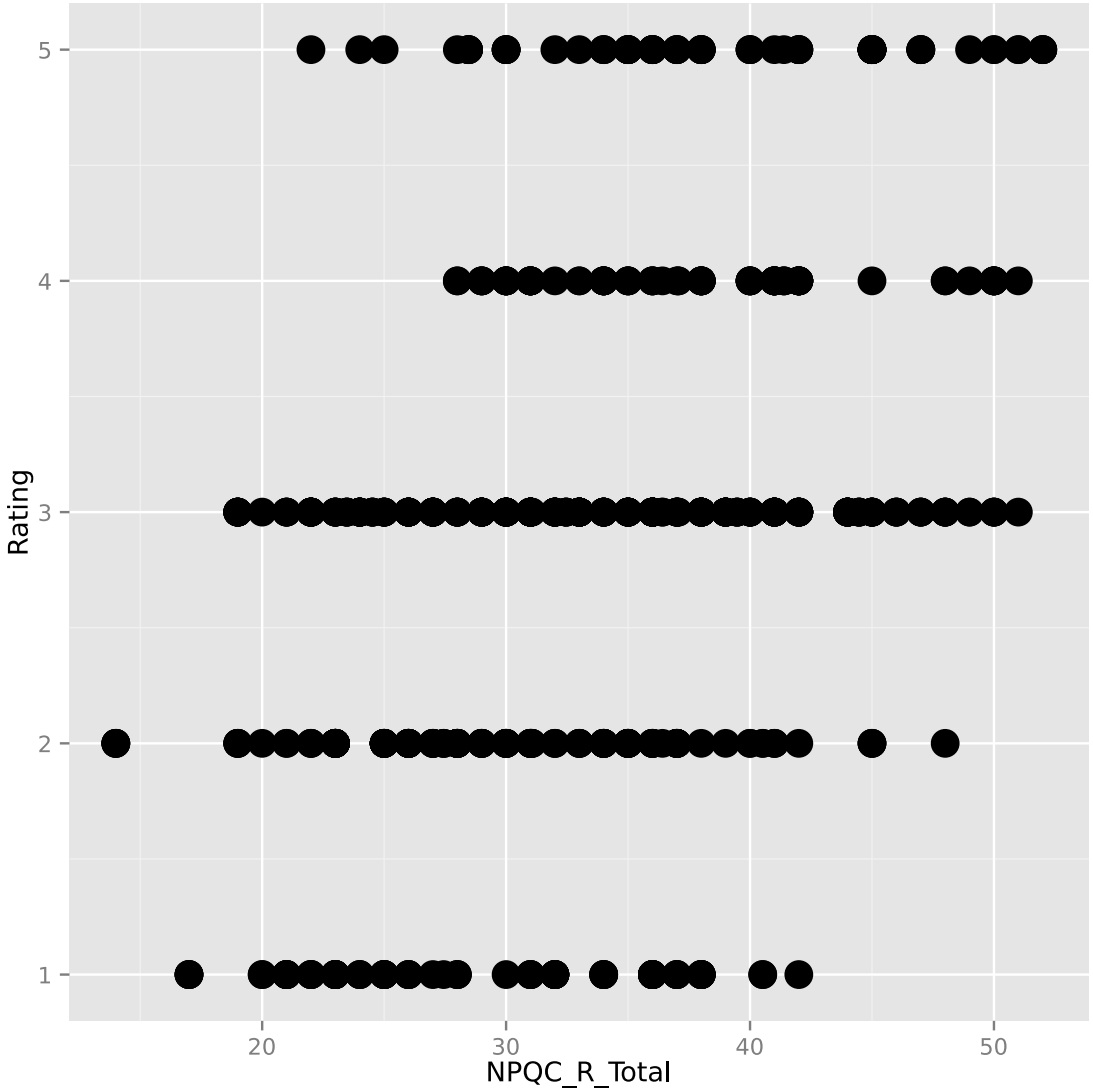
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [16]:

```
graph + geom_point(size = 6)
```

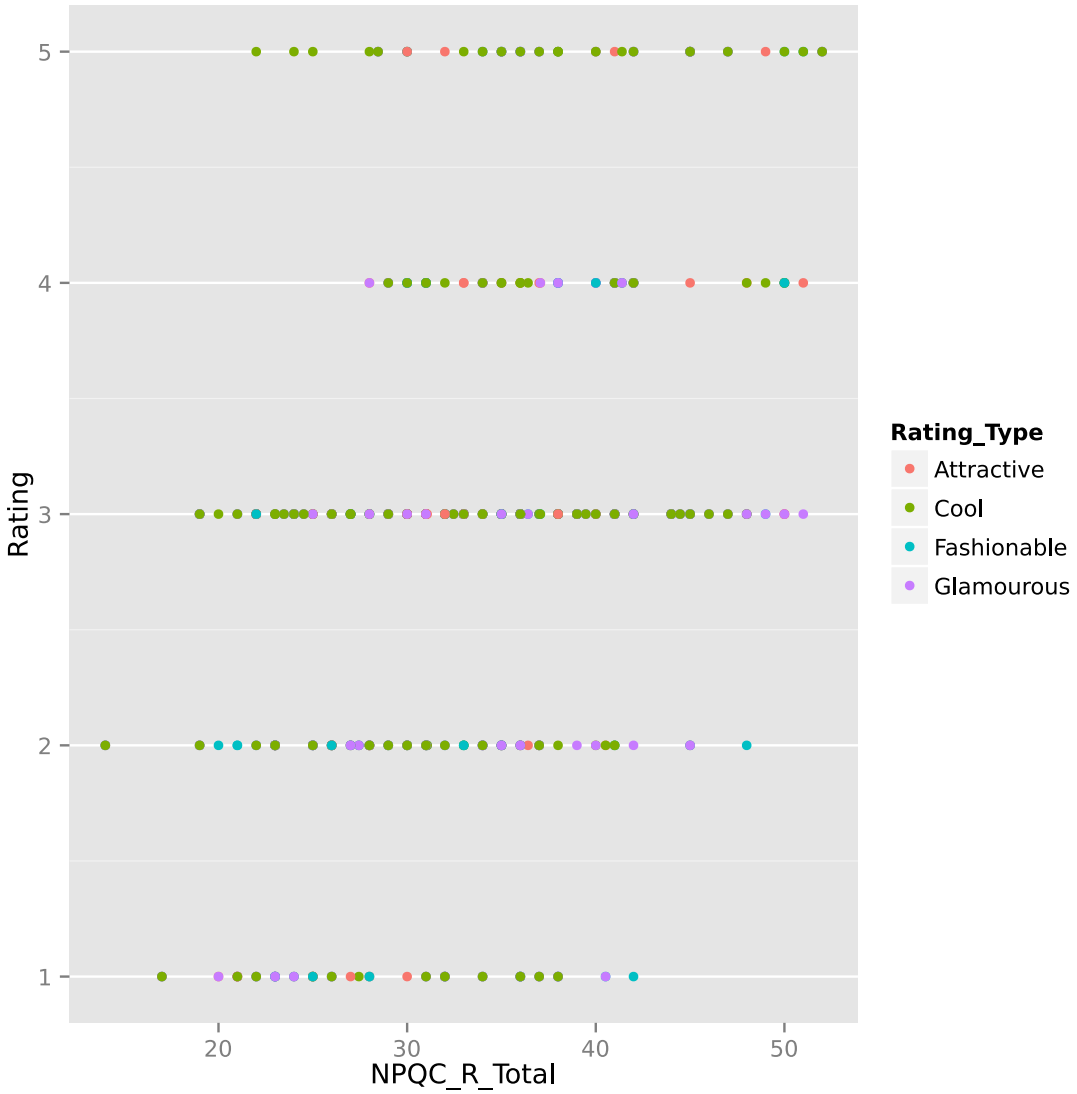
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [17]:

```
graph + geom_point(aes(colour = Rating_Type))
```

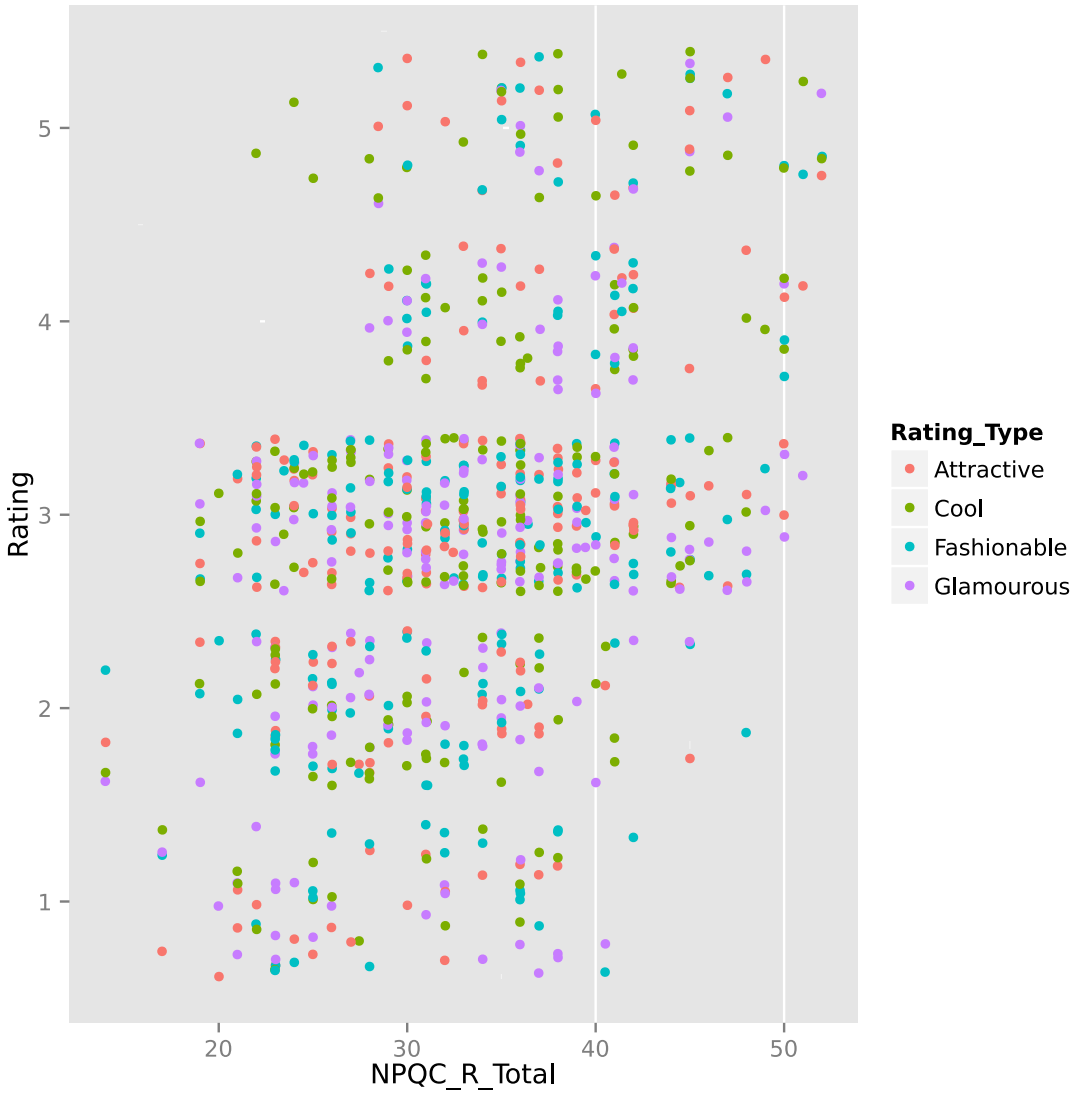
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [18]:

```
graph + geom_point(aes(colour = Rating_Type), position = "jitter")
```

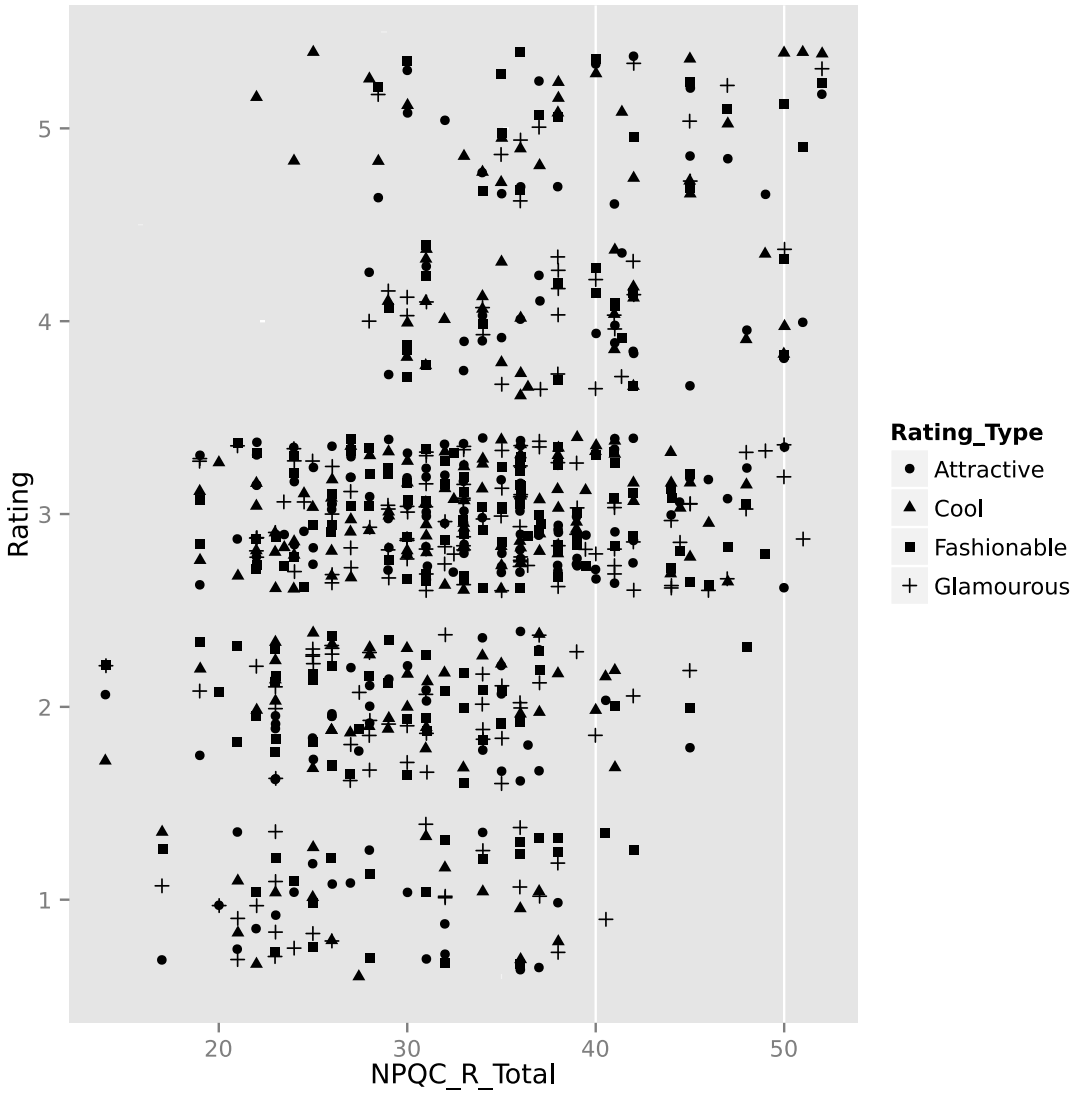
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [19]:

```
graph + geom_point(aes(shape = Rating_Type), position = "jitter")
```

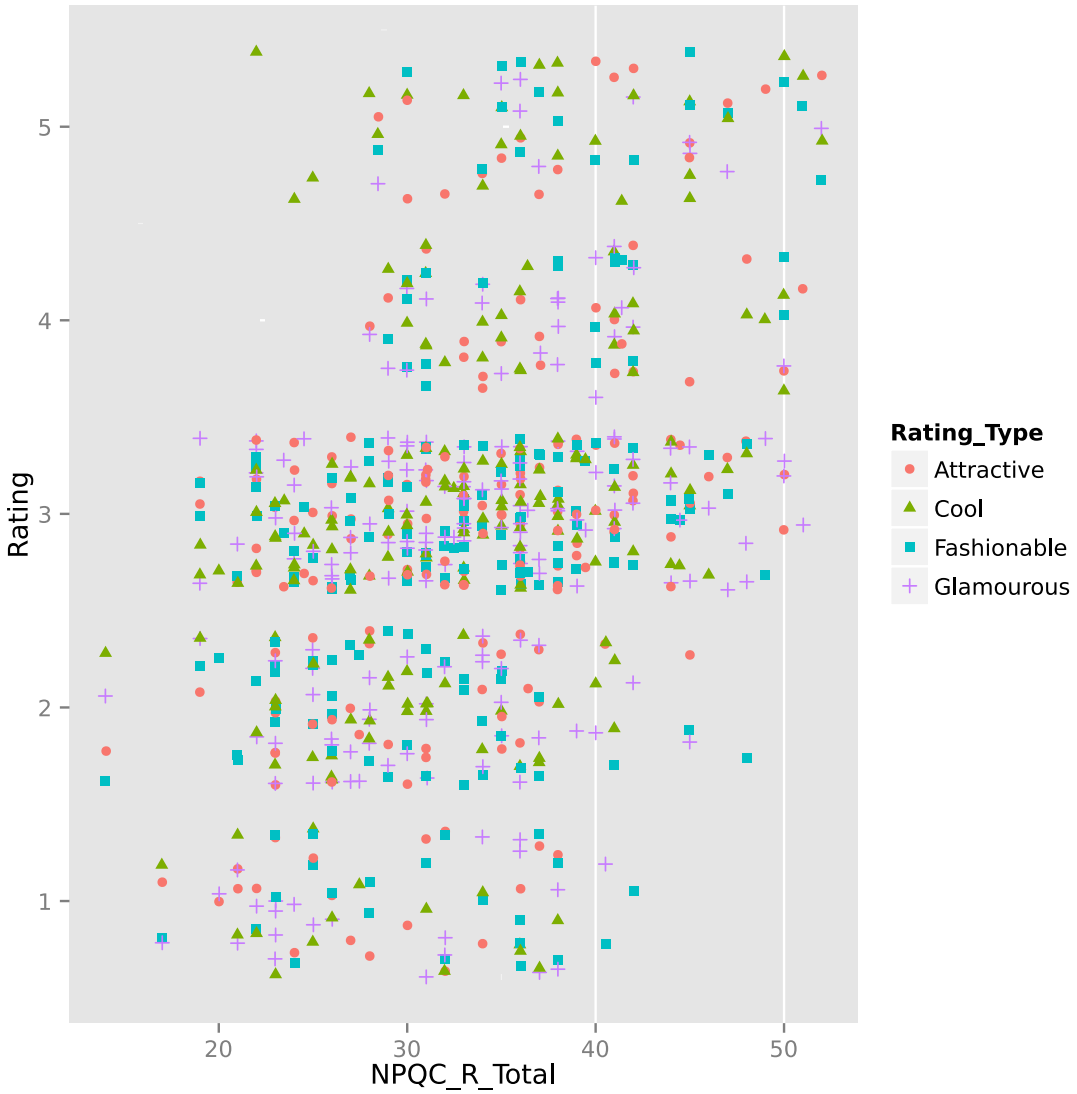
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [20]:

```
graph + geom_point(aes(colour = Rating_Type, shape = Rating_Type), position
= "jitter")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



Graphing relationships : the scatterplot

Simple scatterplot

In [39]:

```
examData <- read.delim("Exam Anxiety.dat", header = TRUE)
```

In [40]:

```
head(examData); str(examData); summary(examData)
```

Out[40]:

	Code	Revise	Exam	Anxiety	Gender
1	1	4	40	86.298	Male
2	2	11	65	88.716	Female
3	3	27	80	70.178	Male
4	4	53	80	61.312	Male
5	5	4	40	89.522	Male
6	6	22	70	60.506	Female

```
'data.frame':   103 obs. of  5 variables:
 $ Code   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Revise : int  4 11 27 53 4 22 16 21 25 18 ...
 $ Exam   : int  40 65 80 80 40 70 20 55 50 40 ...
 $ Anxiety: num  86.3 88.7 70.2 61.3 89.5 ...
 $ Gender : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 1 1 1
1 1 ...
```

Out[40]:

```
      Code            Revise           Exam           Anxiety
Gender
 Min.   :  1.0   Min.   : 0.00   Min.   :  2.00   Min.   : 0.05
6   Female:51
 1st Qu.: 26.5   1st Qu.: 8.00   1st Qu.: 40.00   1st Qu.:69.77
5   Male  :52
 Median : 52.0   Median :15.00   Median : 60.00   Median :79.04
4
 Mean   : 52.0   Mean   :19.85   Mean   : 56.57   Mean   :74.34
4
 3rd Qu.: 77.5   3rd Qu.:23.50   3rd Qu.: 80.00   3rd Qu.:84.68
6
 Max.   :103.0   Max.   :98.00   Max.   :100.00   Max.   :97.58
2
```

1. Code : a number indication from which participant the scores came.
2. Revise : the total hours spent revising.
3. Exam : mark on the exam as a percentage.
4. Anxiety : the score on the EAQ.
5. Gender : whether the participant was male or female (stored as strings of text).

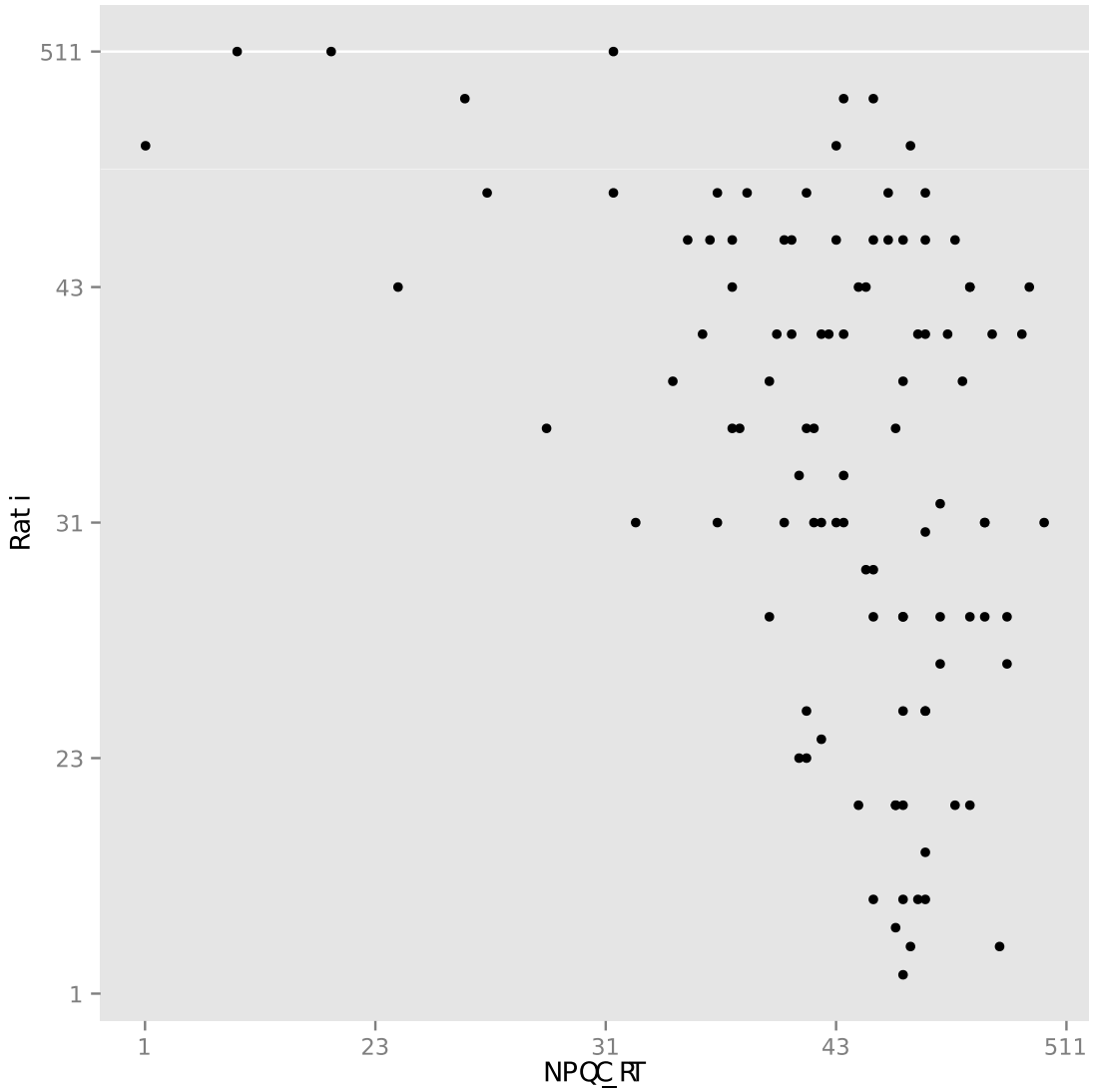
In [41]:

```
scatter <- ggplot(examData, aes(Anxiety, Exam))
```

In [42]:

```
scatter + geom_point()
```

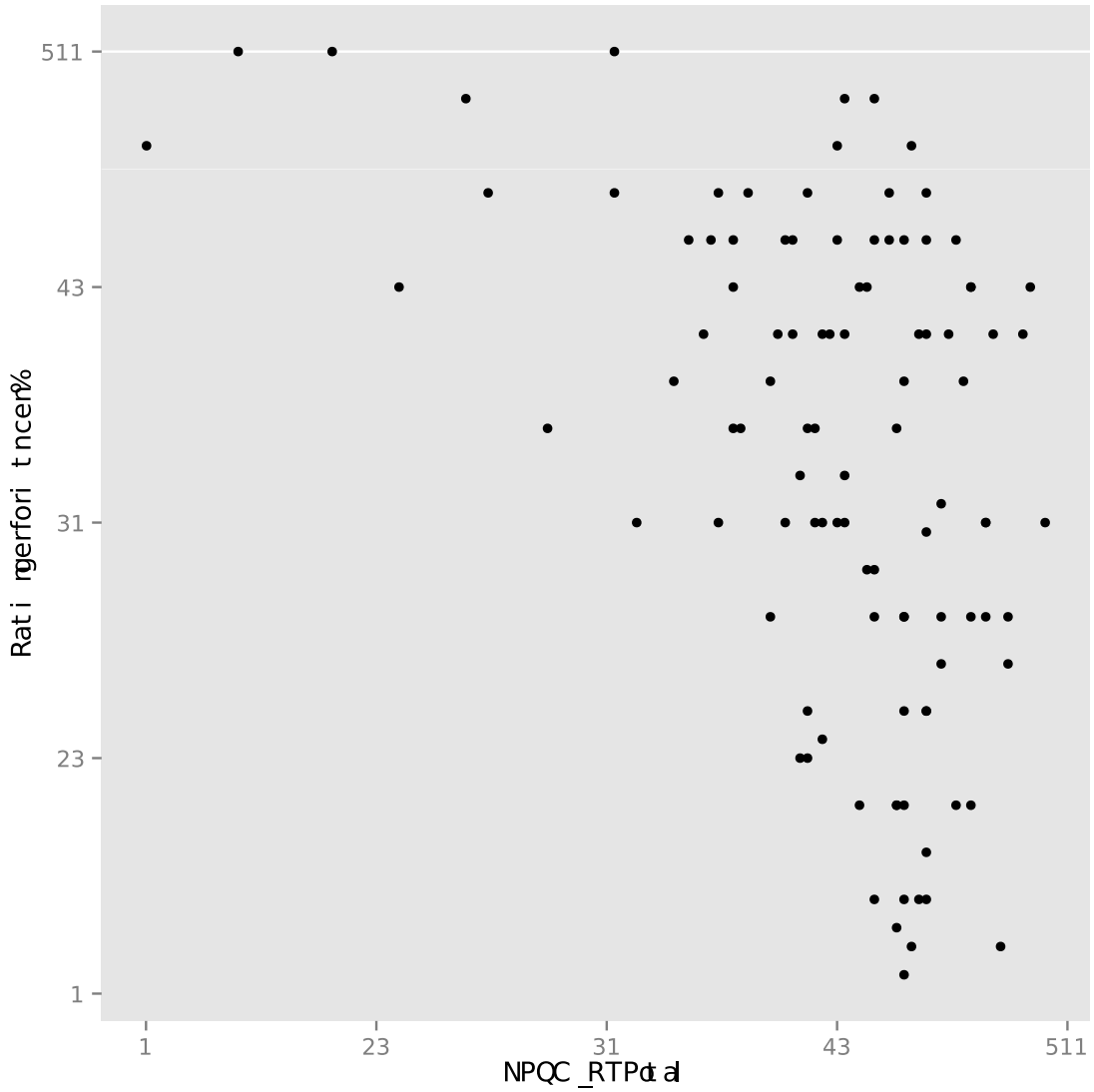
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [43]:

```
scatter + geom_point() + labs(x = "Exam Anxiety", y = "Exam Performance %")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



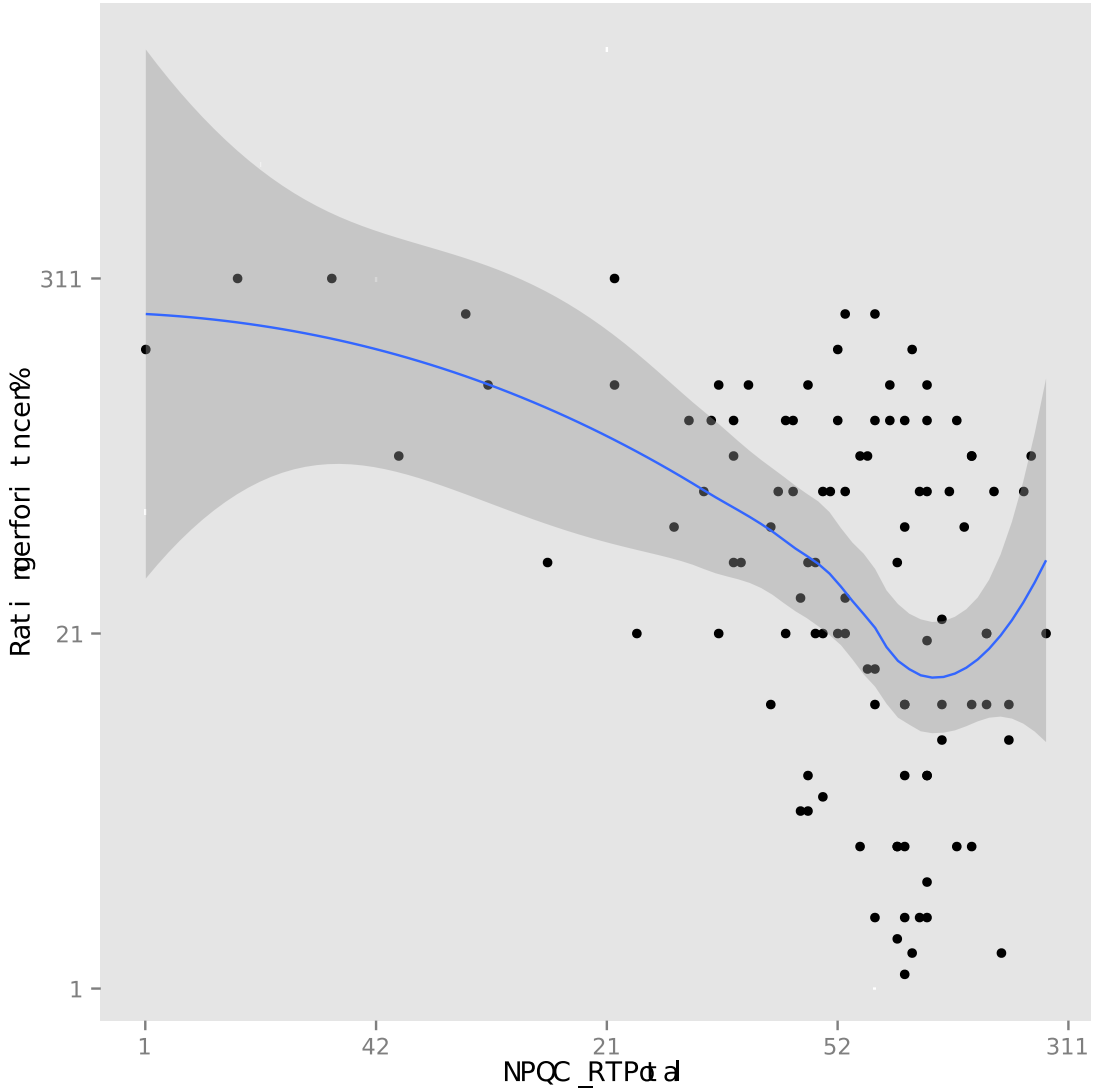
Adding a funky line : regression line

In [44]:

```
scatter + geom_point() + geom_smooth() + labs(x = "Exam Anxiety", y = "Exam
Performance %")
```

```
geom_smooth: method="auto" and size of largest group is <1000,
so using loess. Use 'method = x' to change the smoothing metho
d.
```

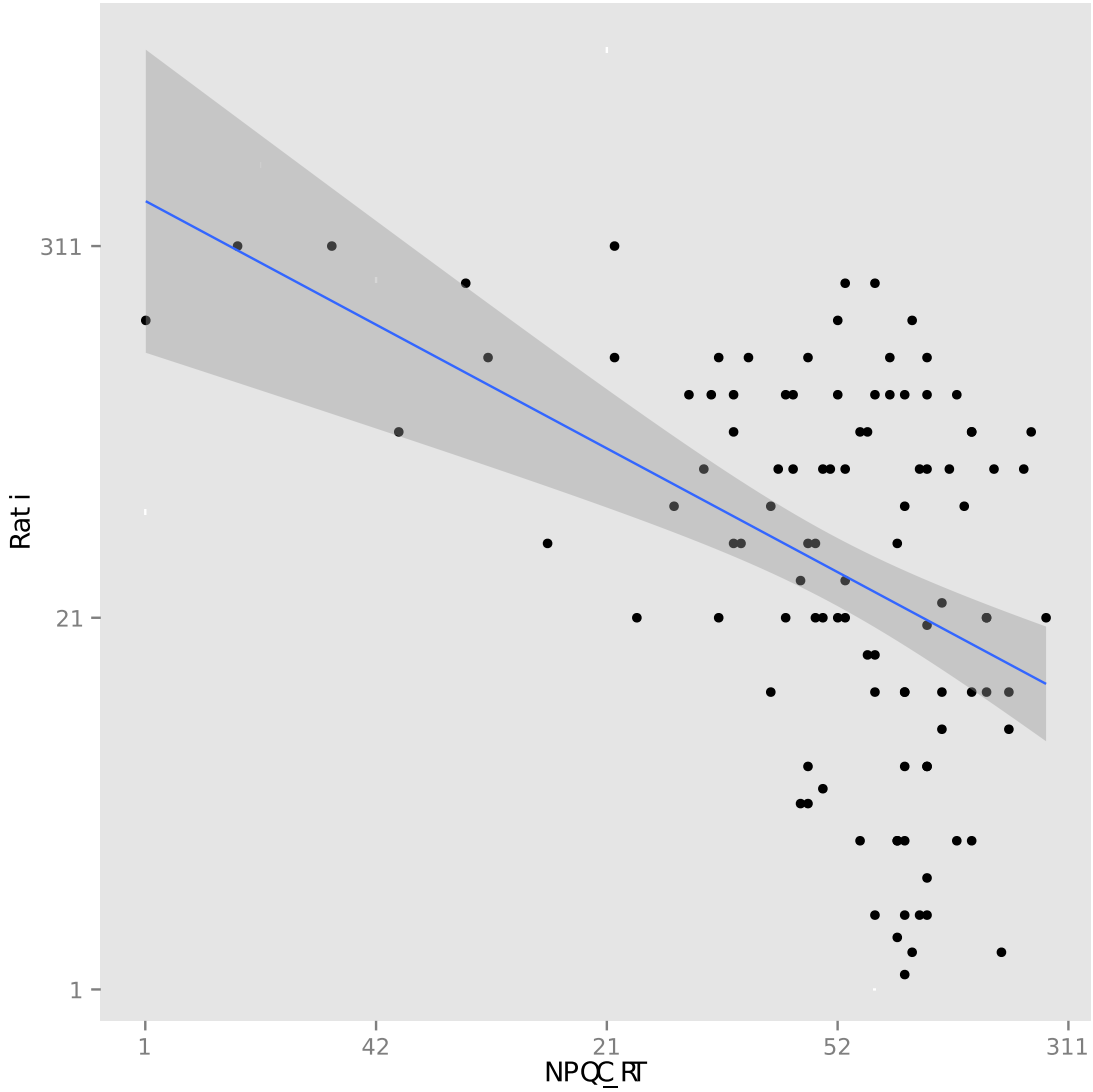
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [45]:

```
scatter + geom_point() + geom_smooth(method = "lm")
```

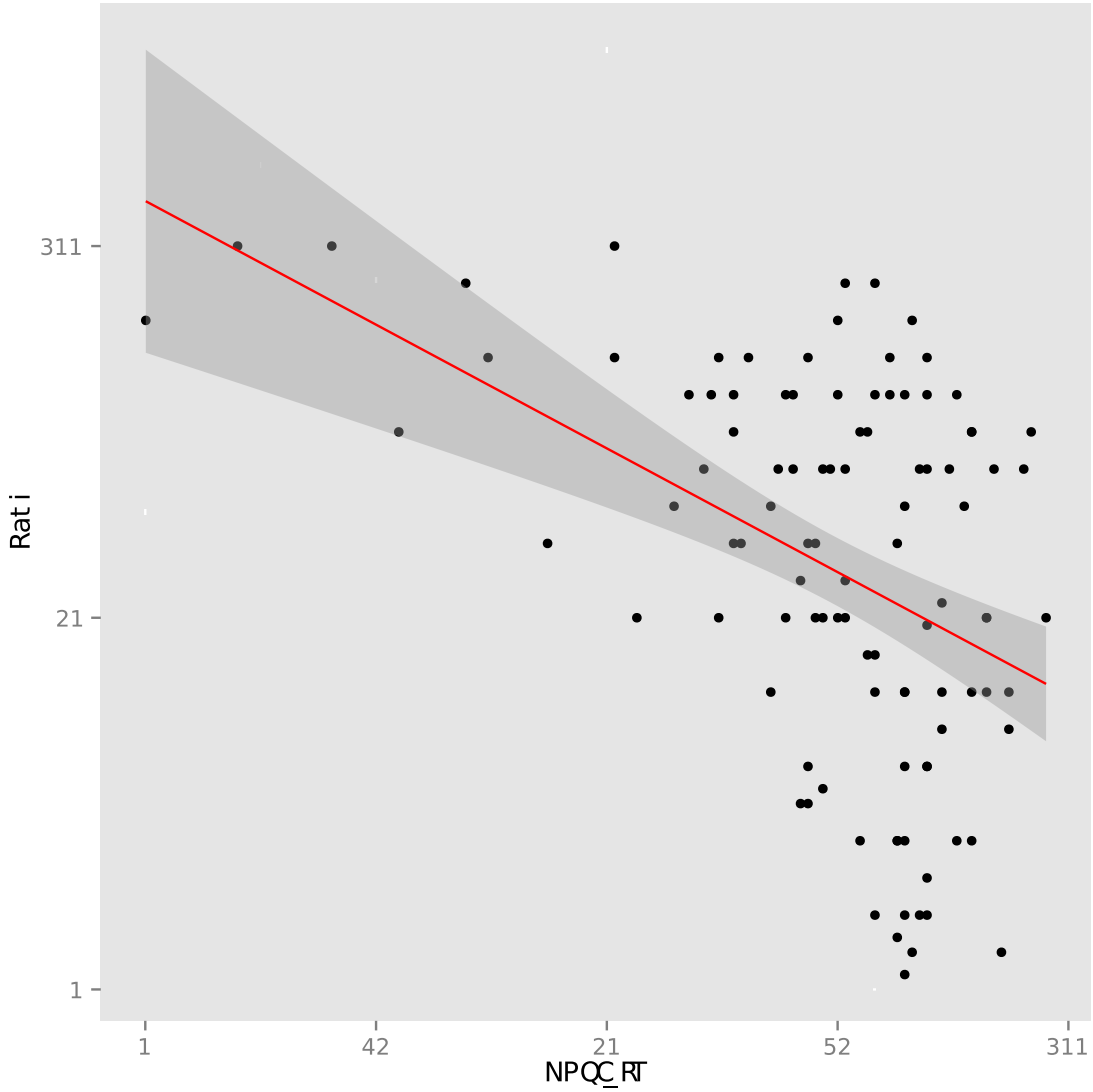
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [46]:

```
scatter + geom_point() + geom_smooth(method = "lm", colour = "Red")
```

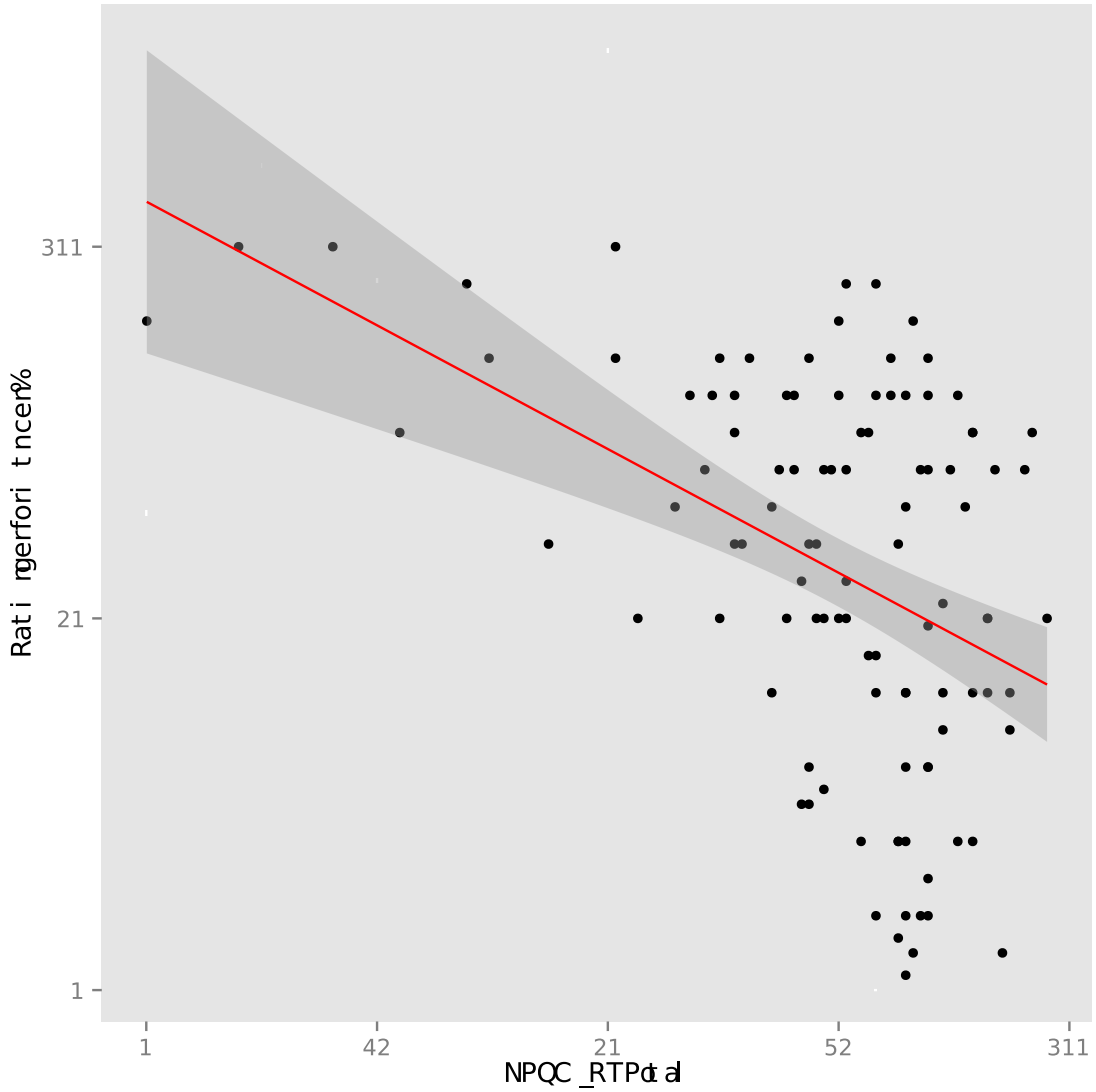
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [47]:

```
scatter <- ggplot(examData, aes(Anxiety, Exam))
scatter + geom_point() + geom_smooth(method = "lm", colour = "Red") + lab
s(x = "Exam Anxiety")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



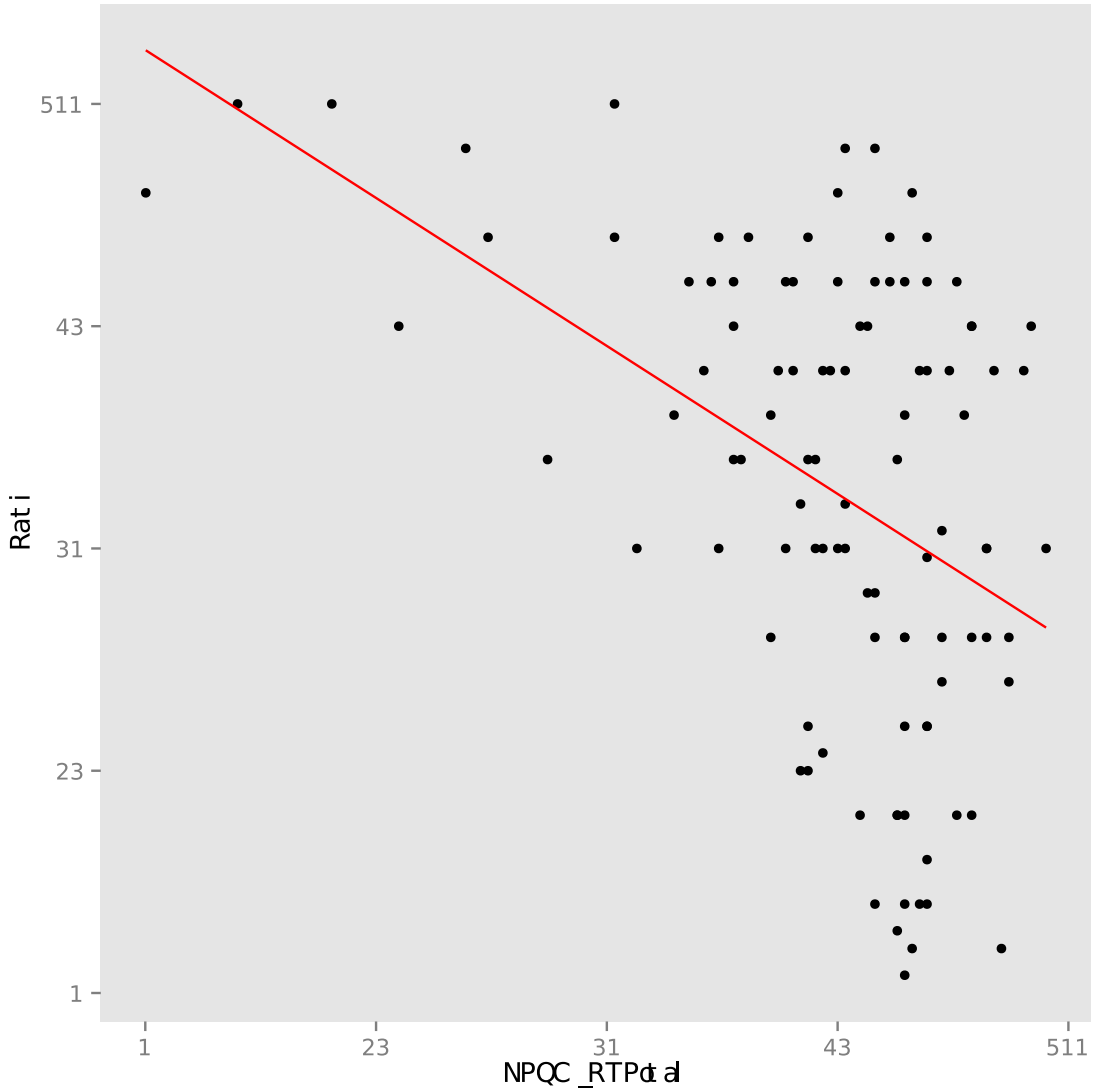
In [49]:

```
# se = F  : standard error = false
```

In [48]:

```
scatter <- ggplot(examData, aes(Anxiety, Exam))
scatter + geom_point() + geom_smooth(method = "lm",se = F, colour = "Red")
+ labs(x = "Exam Anxiety")
```

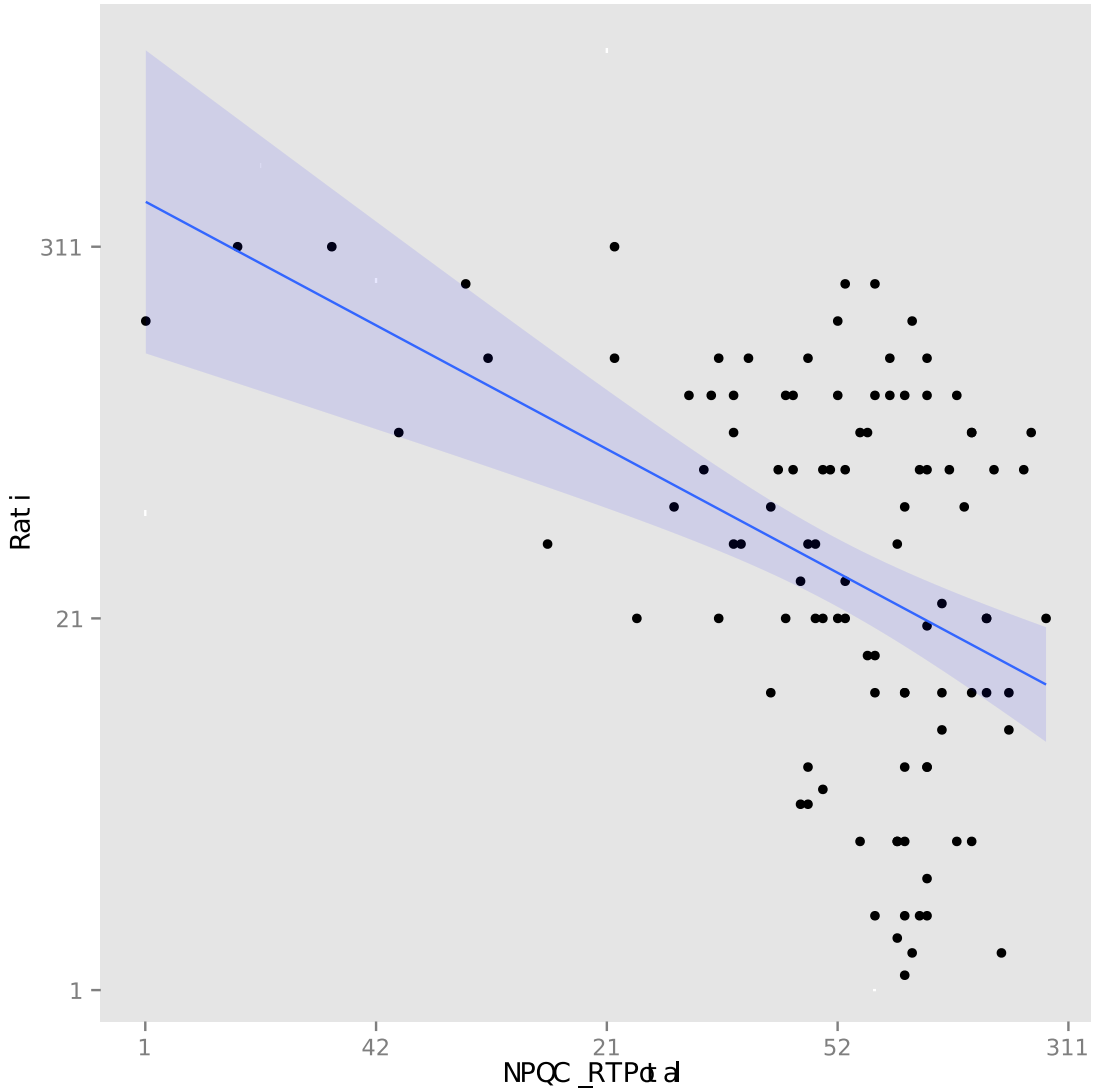
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [50]:

```
scatter <- ggplot(examData, aes(Anxiety, Exam))
scatter + geom_point() + geom_smooth(method = "lm",alpha = 0.1, fill= "Blu
e") + labs(x = "Exam Anxiety")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



Grouped scatterplot

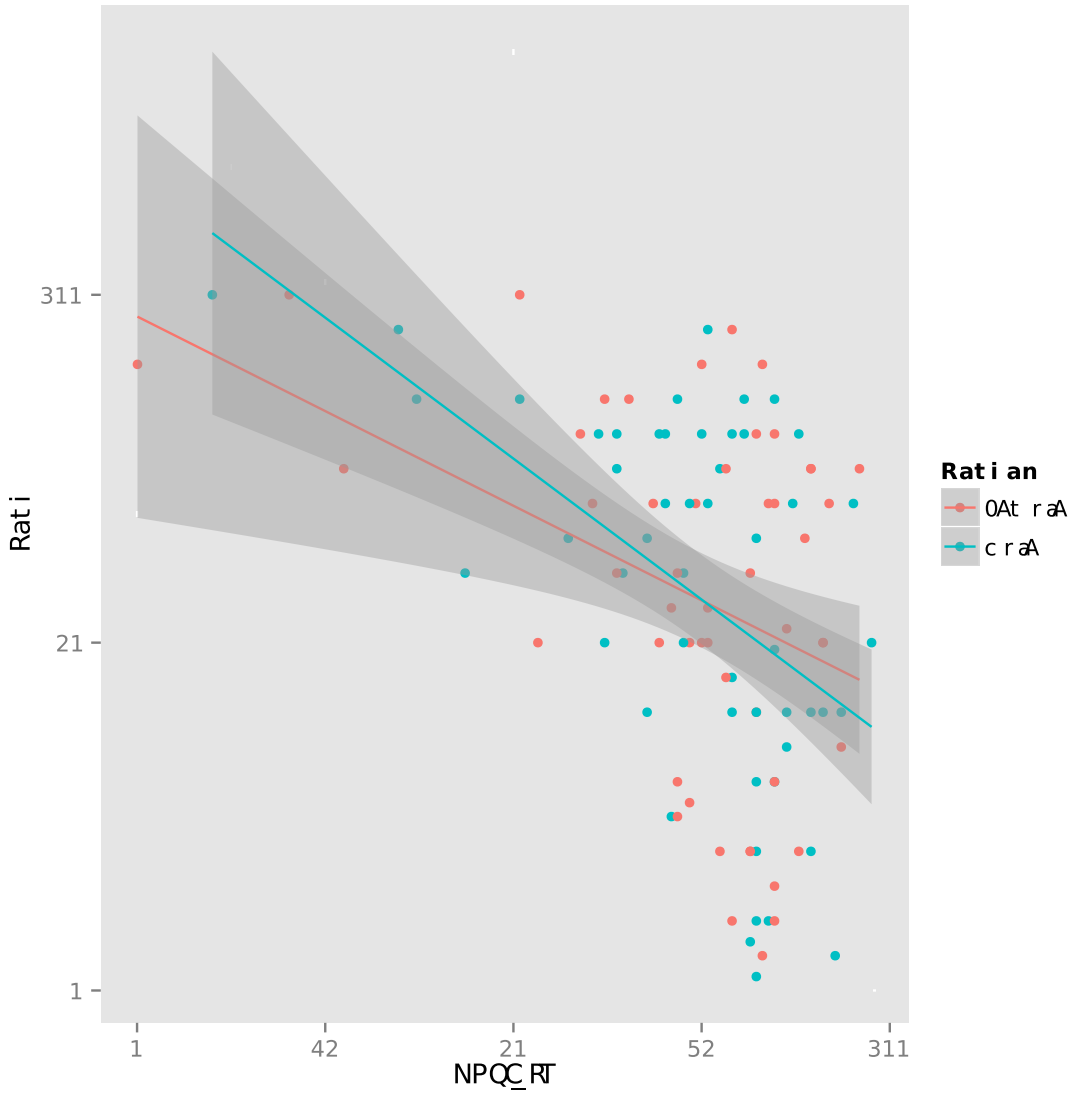
In [51]:

```
scatter <- ggplot(examData, aes(Anxiety, Exam, colour = Gender))
```

In [53]:

```
scatter + geom_point() + geom_smooth(method = "lm")
```

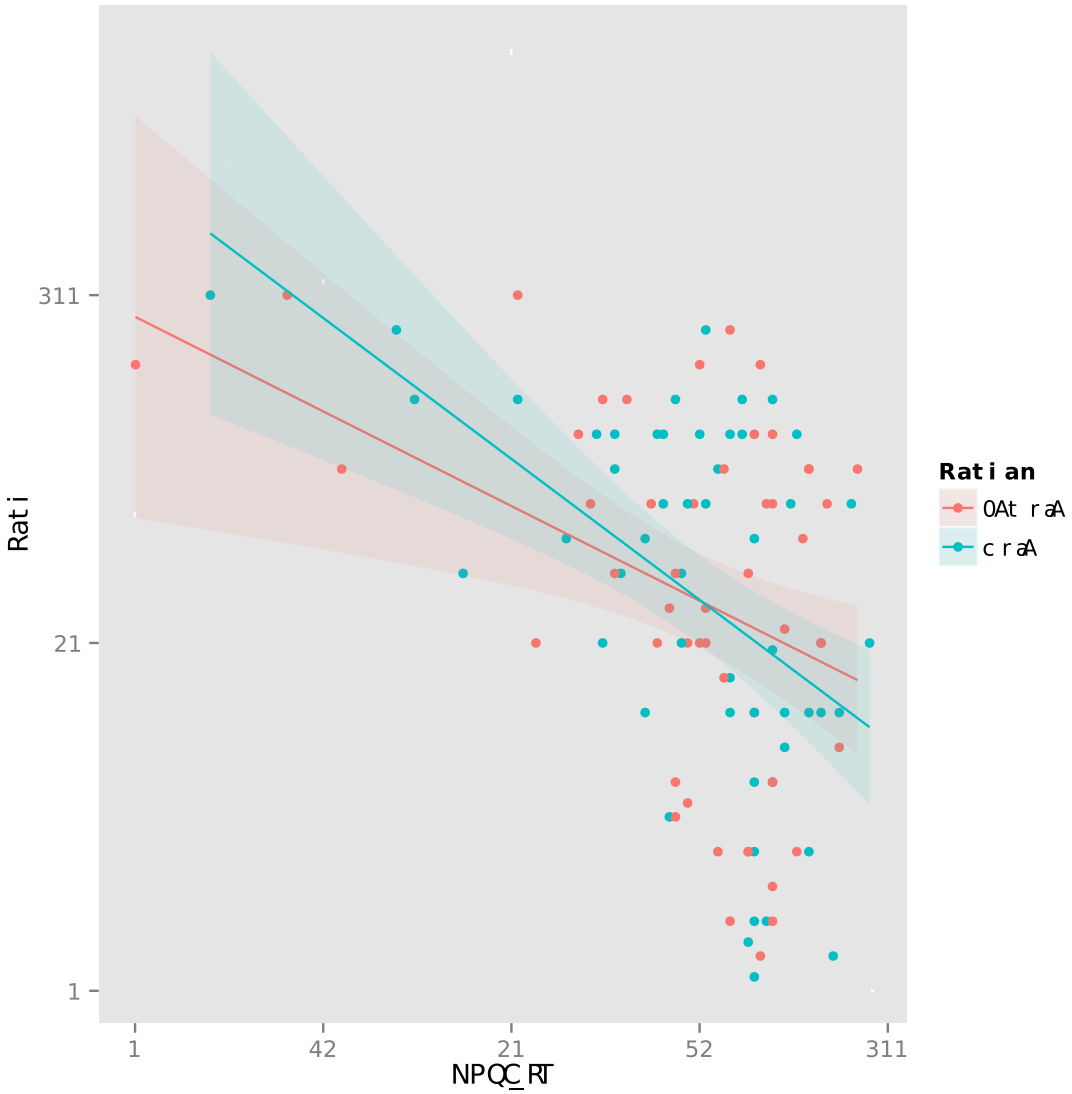
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [54]:

```
scatter + geom_point() + geom_smooth(method = "lm", aes(fill = Gender), alp
ha = 0.1)
```

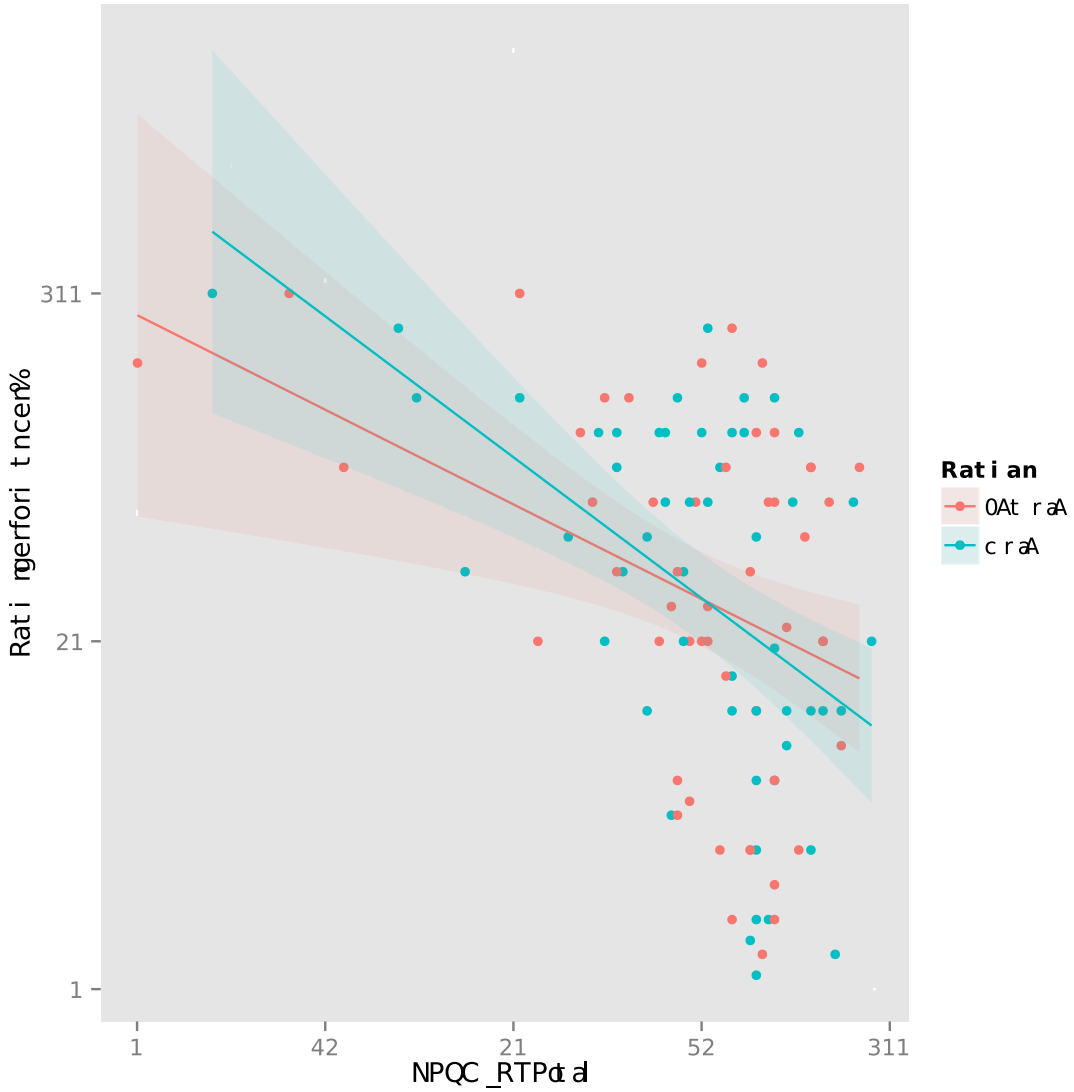
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [55]:

```
scatter + geom_point() + geom_smooth(method = "lm", aes(fill = Gender), alp
ha = 0.1) + labs(x = "Exam Anxiety", y = "Exam Performance %", colour = "Ge
nder")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



Histograms : a good way to spot obvious problems

In [56]:

```
festivalData <- read.delim("DownloadFestival.dat", header = TRUE)
```

In [57]:

```
festivalHistogram <- ggplot(festivalData, aes(day1)) + opts(legend.positio
n="none")  # not working
```

```
Error: Use 'theme' instead. (Defunct; last used in version

0.9.1)
```

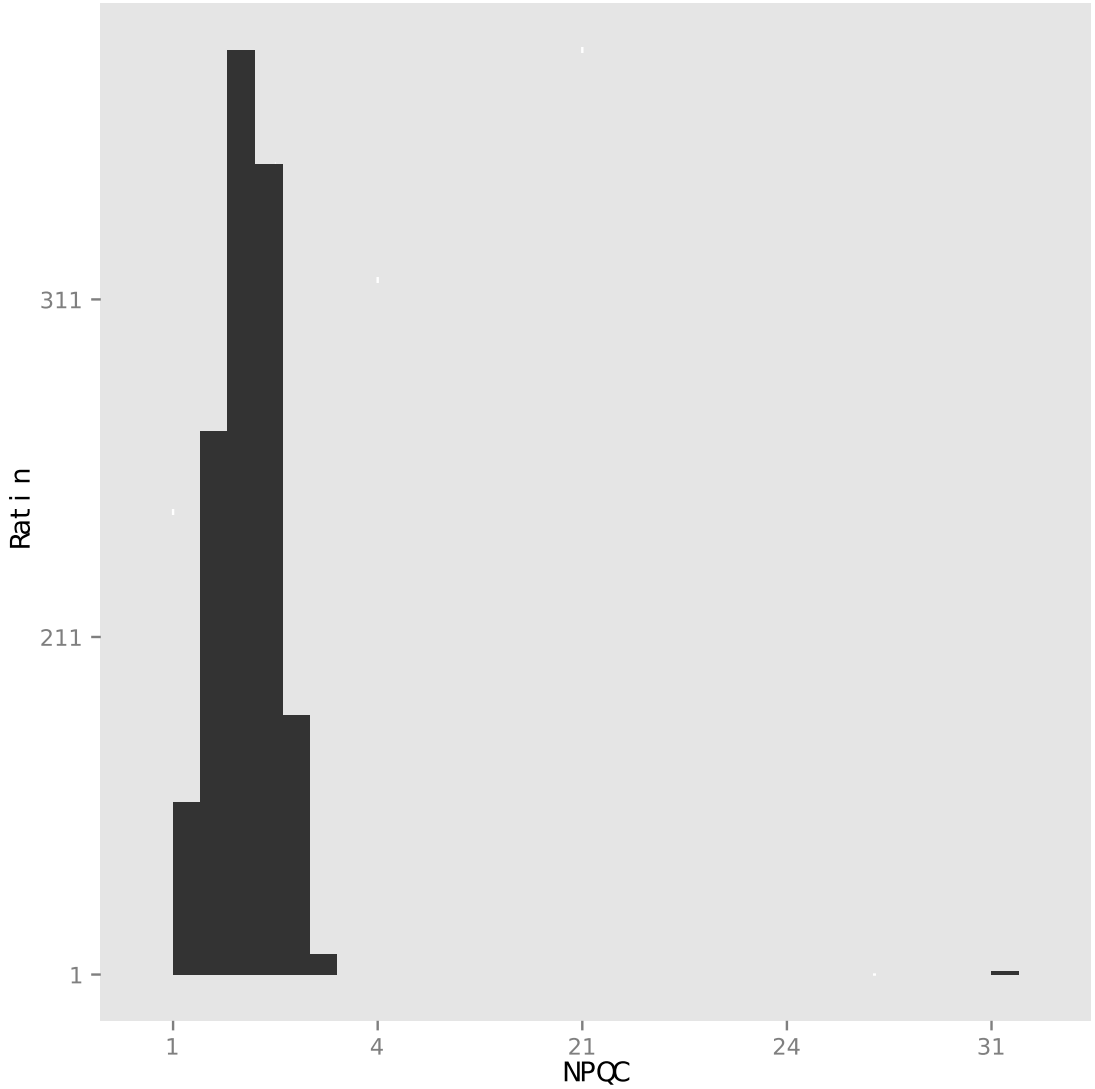
In [58]:

```
festivalHistogram <- ggplot(festivalData, aes(day1)) + theme(legend.positio
n="none")
```

In [59]:

```
festivalHistogram + geom_histogram()
```

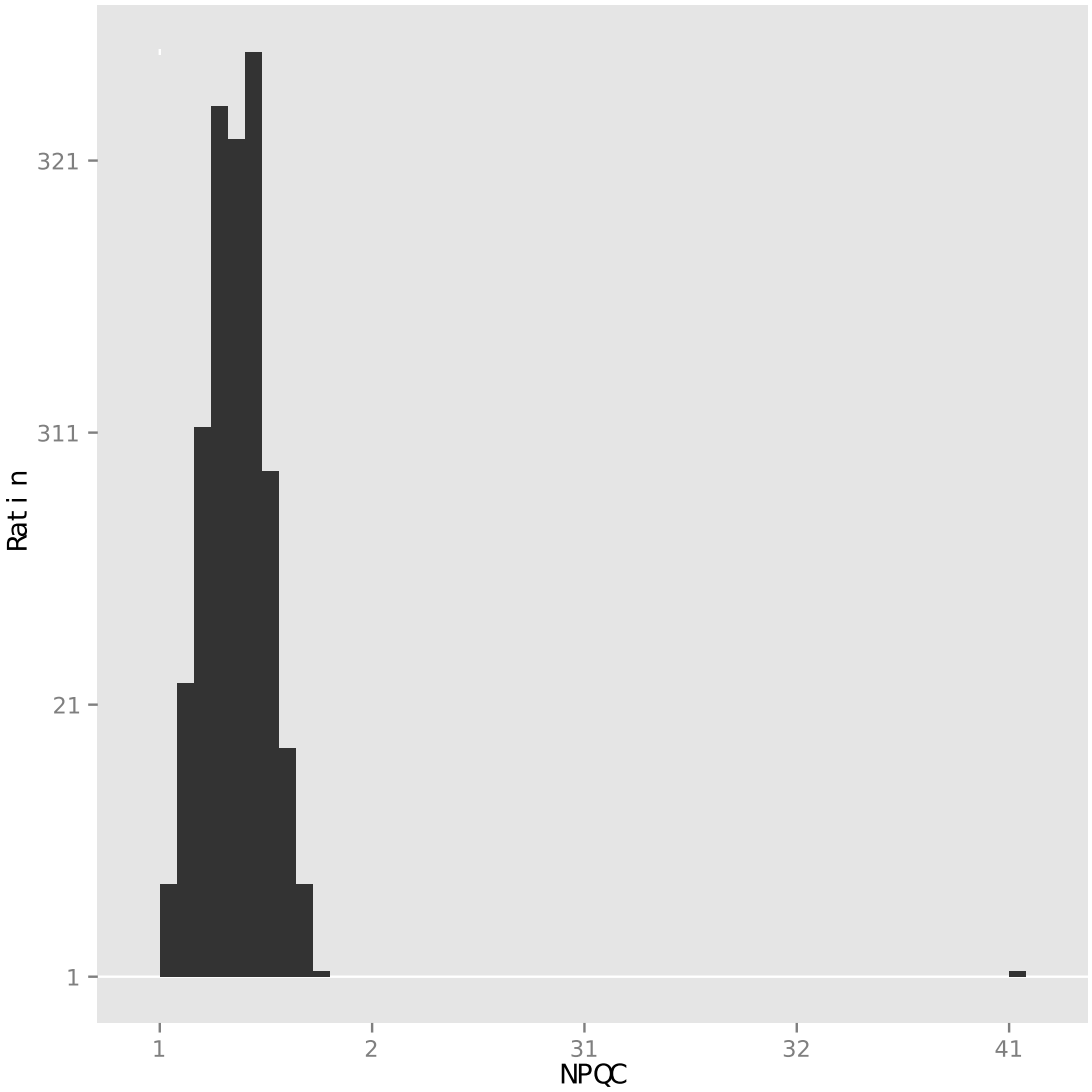
```
stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to
adjust this.

Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [61]:

```
festivalHistogram + geom_histogram(binwidth = 0.4)
```

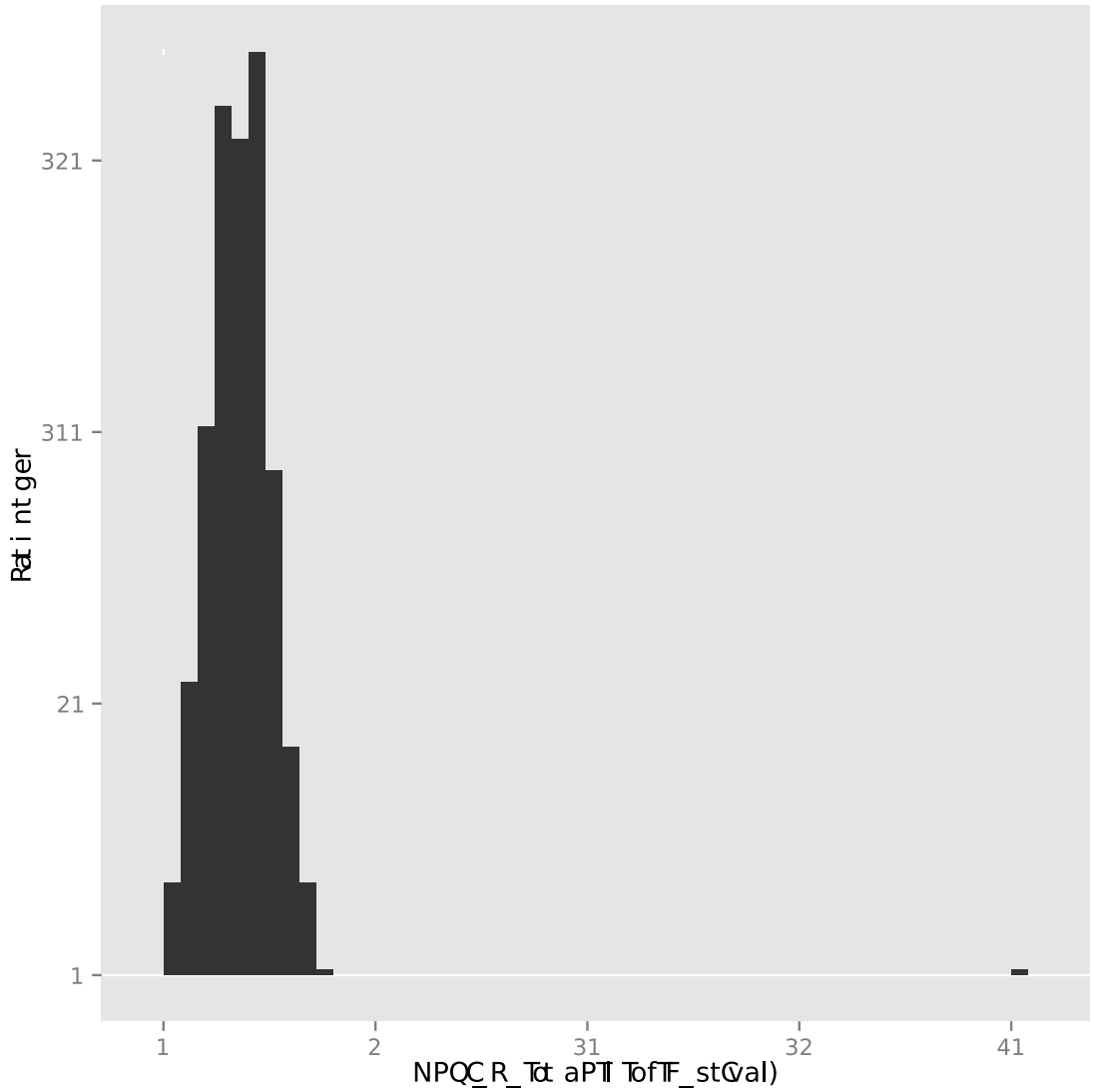
```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [62]:

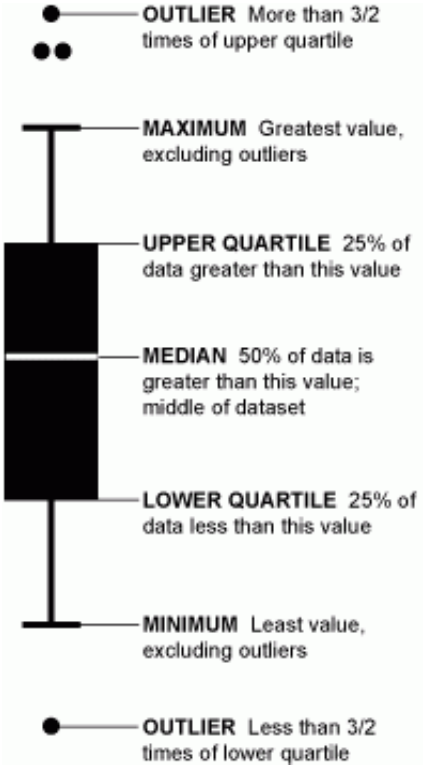
```
festivalHistogram + geom_histogram(binwidth = 0.4) + labs(x = "Hygiene (Day
1 of Festival)", y = "Frequency")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



Boxplots (box-whisker disgrams)

Box plot 정확히 상자와 수염 그림(box and whisker plot)은 두 개 이상의 집단의 상대적 비교를 위해서 각 집단의 최대값(max)과 최소값(min) 그리고 중앙값(자료를 크기순으로 나열했을 때 가운데 위치하는 값: median) 및 사분위수(자료를 크기 순서에 따라 늘어놓은 자료를 4등분 했을 때 위치하는 값을 의미함) 제 1사분위수(아래에서 25% 백분위점에 위치하는 수: Q1), 제 3사분위수(아래에서 75% 백분위점에 위치하는 수: Q3)등 다섯 숫자를 요약하여 그래프로 나타내는 방법으로 John W. Tukey가 제안한 탐색적 데이터 분석 방법입니다. 출처 : http://wsyang.com/2013/07/add-more-info-to-the-boxplot/ (http://wsyang.com/2013/07/add-more-info-to-the-boxplot/)



http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/ (http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/)

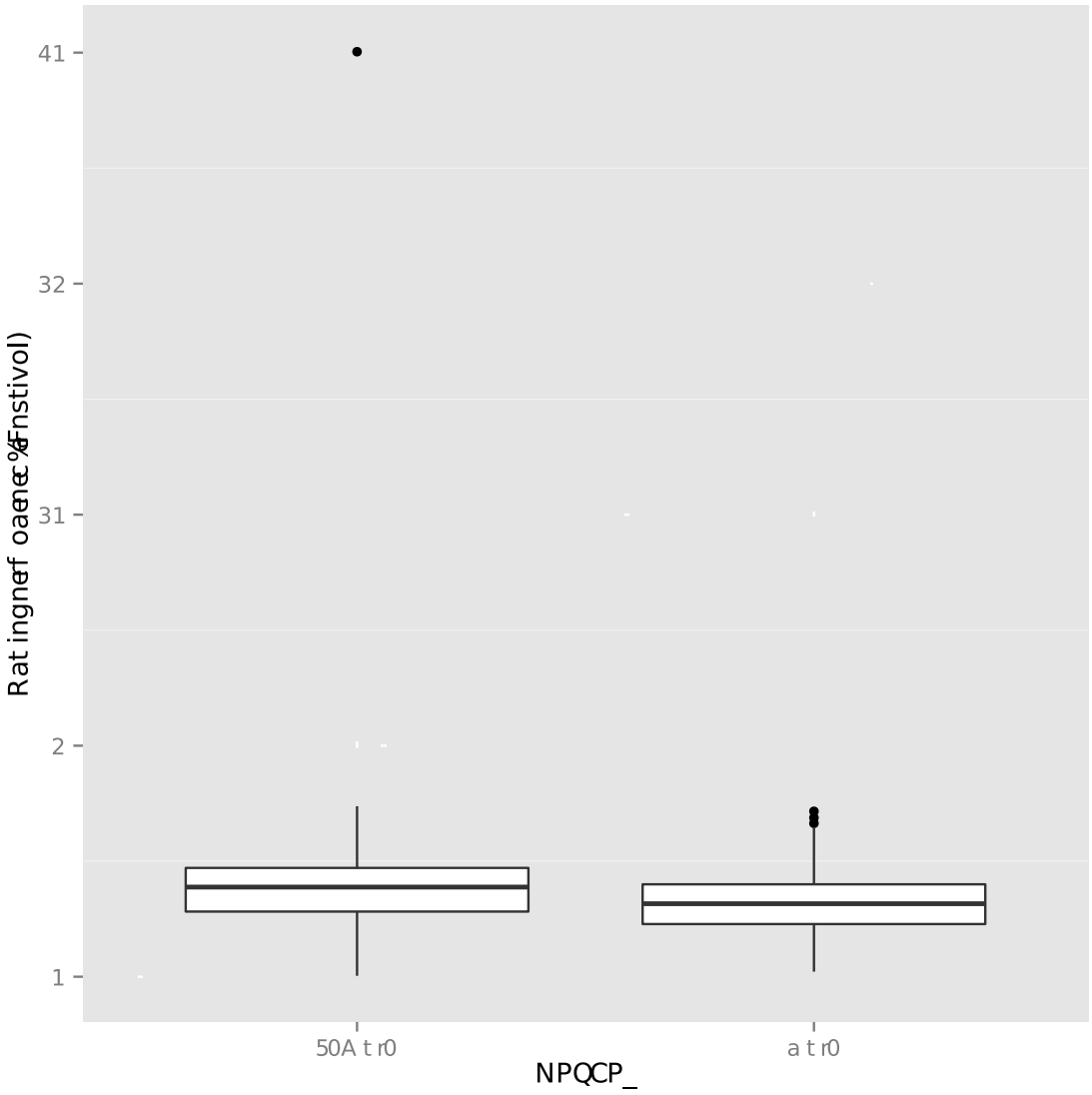
In [66]:

```
festivalBoxplot <- ggplot(festivalData, aes(gender, day1))
```

In [67]:

```
festivalBoxplot + geom_boxplot() + labs(x = "Gender", y = "Hygiene (Day 1 o
f Festival)")
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



In [68]:

```
festivalData <- festivalData[order(festivalData$day1),]
```

In [71]:

```
tail(festivalData)
```

Out[71]:

	ticknumb	gender	day1	day2	day3
774	4564	Female	3.38	3.44	3.41
300	3371	Female	3.41	NA	NA
657	4264	Male	3.44	NA	NA
303	3374	Male	3.58	3.35	NA
574	4016	Female	3.69	NA	NA
611	4158	Female	20.02	2.44	NA

Density plots

In [77]:

```
festivalData <- read.delim("DownloadFestival(No Outlier).dat", header = TRUE)
```

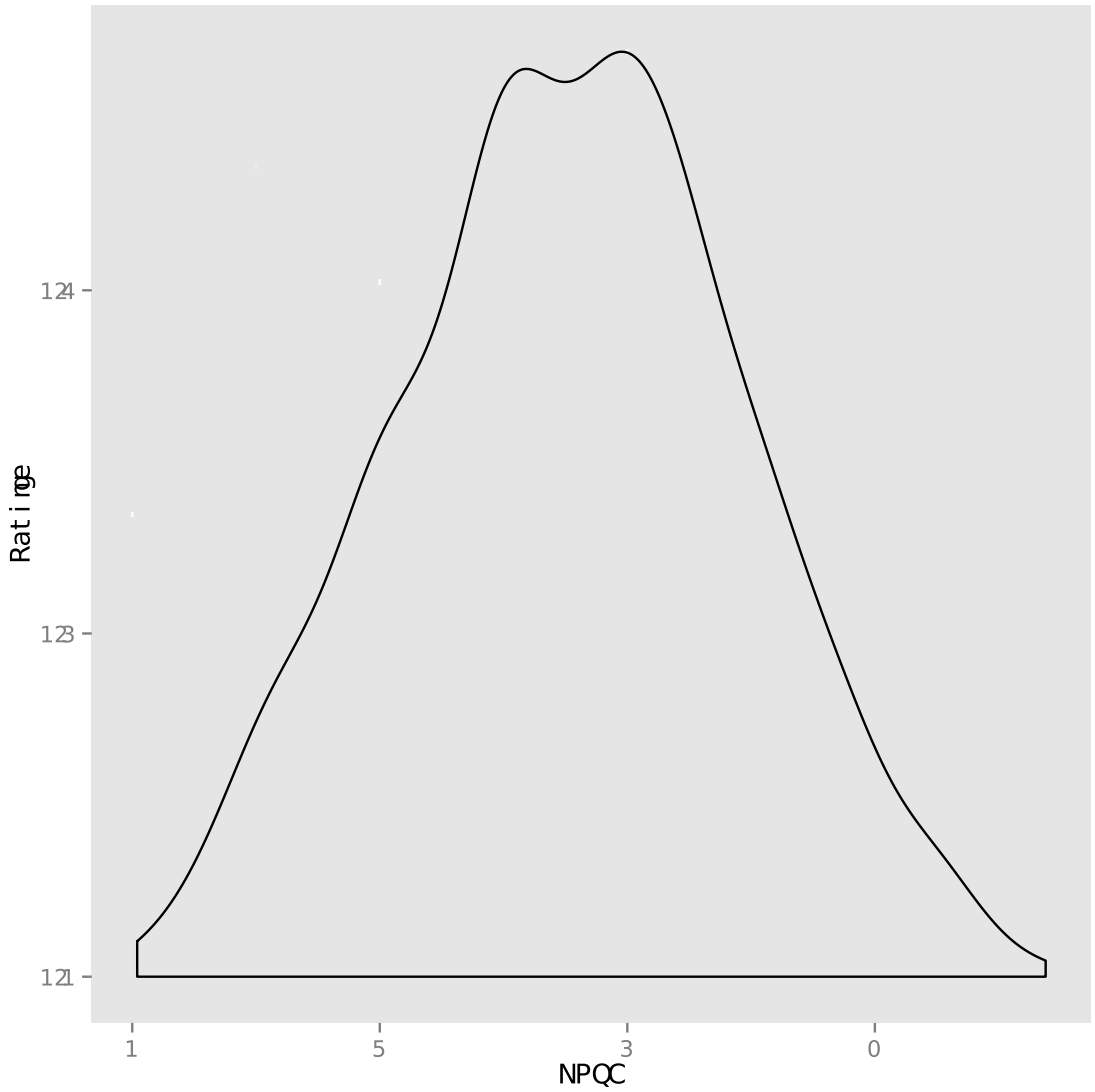
In [78]:

```
density <- ggplot(festivalData, aes(day1))
```

In [79]:

```
density + geom_density()
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(arg
s[[2]])) {: missing value where TRUE/FALSE needed
```



Graphing means

In [81]:

```
chickFlick <- read.delim("ChickFlick.dat",  header = TRUE)
```

In [82]:

```
head(chickFlick); summary(chickFlick);str(chickFlick)
```

Out[82]:

	gender	film	arousal
1	Male	Bridget Jones' Diary	22
2	Male	Bridget Jones' Diary	13
3	Male	Bridget Jones' Diary	16
4	Male	Bridget Jones' Diary	10
5	Male	Bridget Jones' Diary	18
6	Male	Bridget Jones' Diary	24

Out[82]:

```
    gender                      film       arousal     
 Female:20   Bridget Jones' Diary:20   Min.   : 3.00  
 Male  :20   Memento             :20   1st Qu.:14.00  
                                       Median :19.50  
                                       Mean   :20.02  
                                       3rd Qu.:24.25  
                                       Max.   :37.00  

'data.frame':   40 obs. of  3 variables:
 $ gender : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2
2 2 ...
 $ film   : Factor w/ 2 levels "Bridget Jones' Diary",..: 1 1 1
1 1 1 1 1 1 1 ...
 $ arousal: int  22 13 16 10 18 24 13 14 19 23 ...
```

Bar charts for one independent variable

In [83]:

```
bar <- ggplot(chickFlick, aes(film, arousal))
```

In [84]:

```
bar + stat_summary(fun.y = mean, geom = "bar", fill = "White", colour = "Bl
ack") + stat_summary(fun.data = mean_cl_normal, geom = "pointrange") + lab
s(x = "Film", y = "Mean Arousal")
```

```
Error: Hmisc package required for this functionality. Please in
stall and try again.
```

In [ ]:

```

```