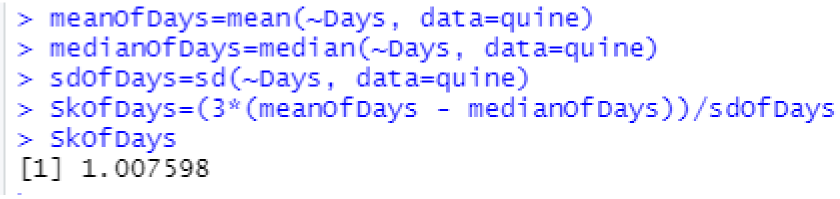
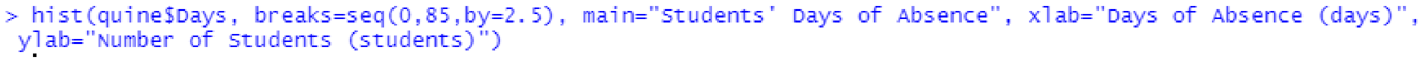
1.



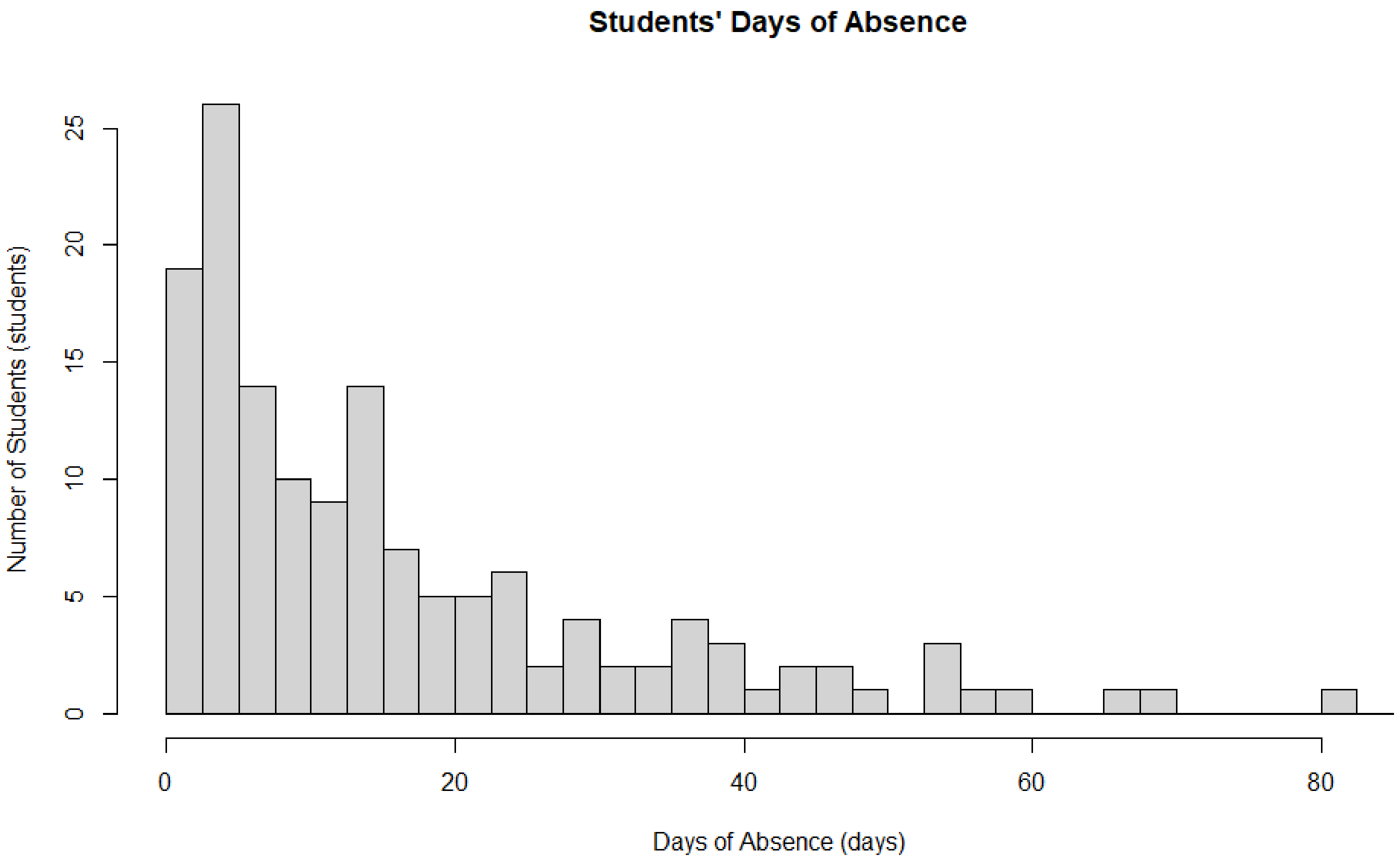
```
> meanOfDays=mean(~Days, data=quine)
> medianOfDays=median(~Days, data=quine)
> sdOfDays=sd(~Days, data=quine)
> skOfDays=(3*(meanOfDays - medianOfDays))/sdOfDays
> skOfDays
[1] 1.007598
```

The Pearsonian coefficient of skewness of this dataset is greater than 1, therefore the data set of the number of days of absence is skewed to the right.

2.

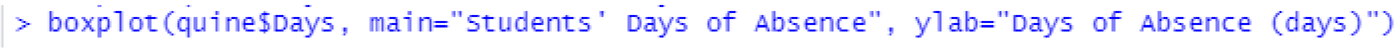


```
> hist(quine$Days, breaks=seq(0,85,by=2.5), main="Students' Days of Absence", xlab="Days of Absence (days)",
 ylab="Number of Students (students)")
```

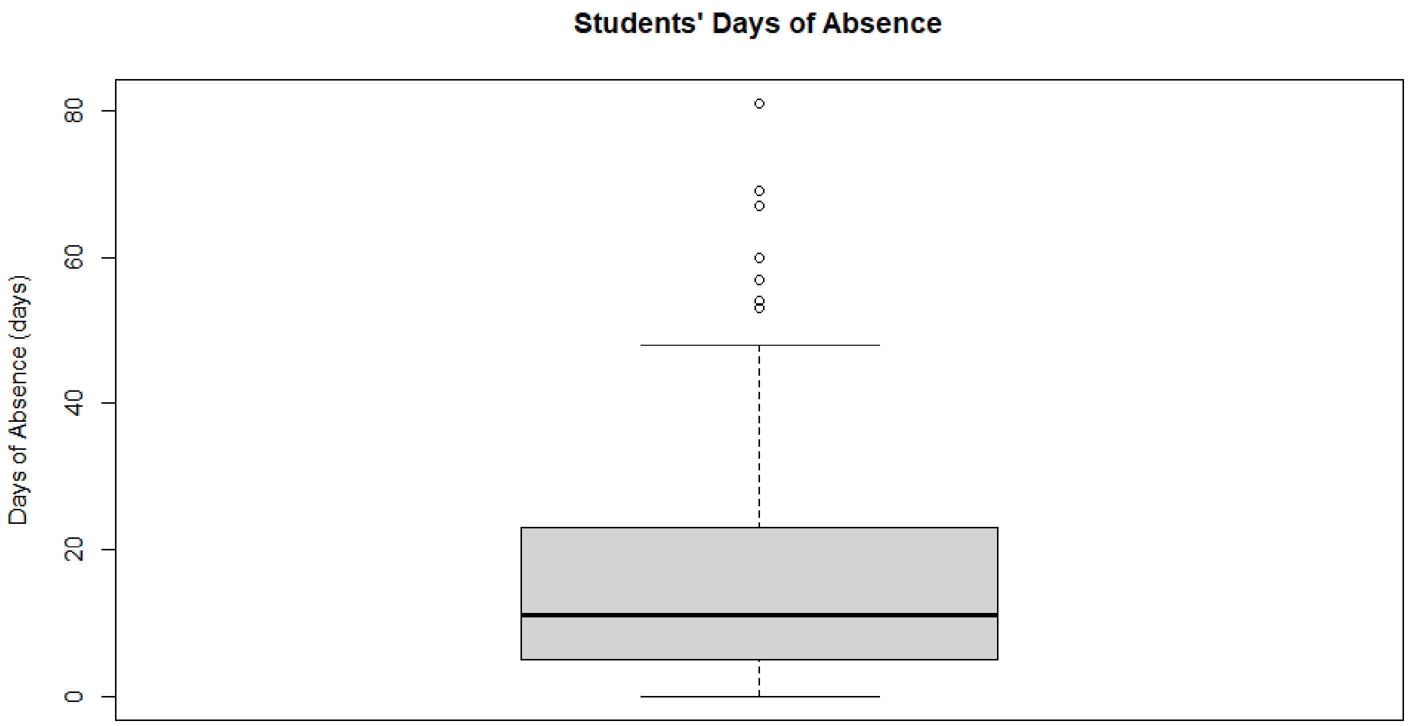


Yes, the histogram of Students' Days of Absence is consistent with the skewness to the right as very few students have number of absent days more than 50 days.

3.



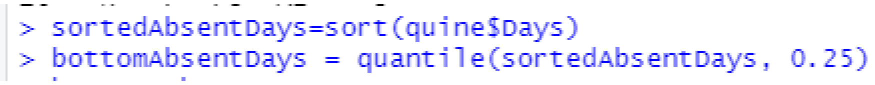
```
> boxplot(quine$Days, main="Students' Days of Absence", ylab="Days of Absence (days)")
```



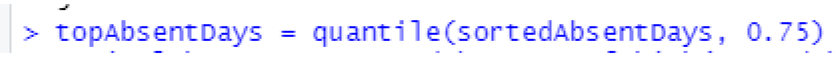
There are 7 outliners in the data based on the boxplot of Students' Days of Absence.

A typical student was absent between 5 and 22 days.

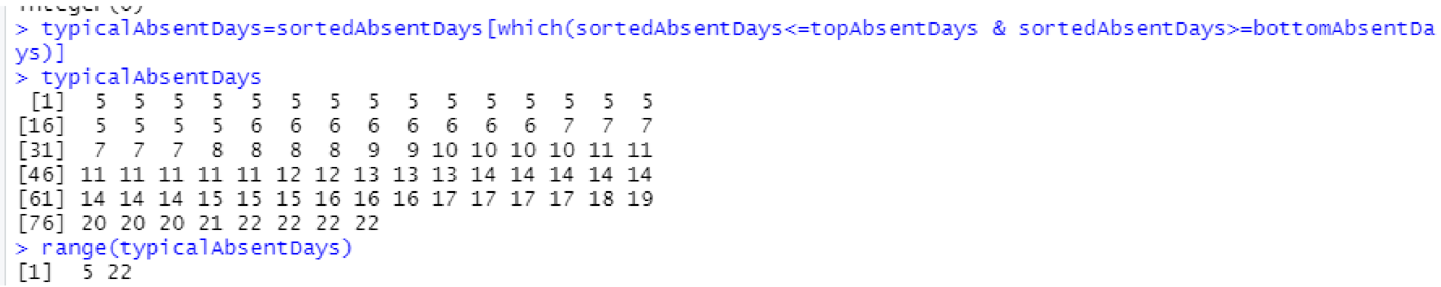
4.



```
> sortedAbsentDays=sort(quine$Days)
> bottomAbsentDays = quantile(sortedAbsentDays, 0.25)
```



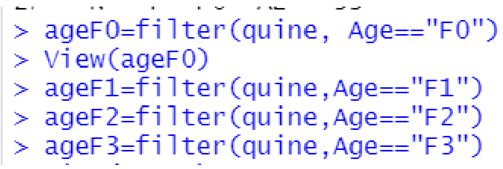
```
> topAbsentDays = quantile(sortedAbsentDays, 0.75)
```



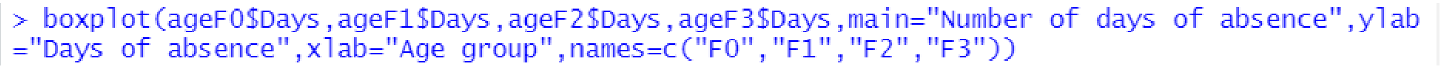
```
> typicalAbsentDays=sortedAbsentDays[which(sortedAbsentDays<=topAbsentDays & sortedAbsentDays>=bottomAbsentDa
ys)]
> typicalAbsentDays
 [1]  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5
[16]  5  5  5  5  6  6  6  6  6  6  6  6  7  7  7
[31]  7  7  7  8  8  8  8  9  9 10 10 10 10 11 11
[46] 11 11 11 11 11 12 12 13 13 13 14 14 14 14 14
[61] 14 14 14 15 15 15 16 16 16 17 17 17 17 18 19
[76] 20 20 20 21 22 22 22 22
> range(typicalAbsentDays)
[1]  5 22
```

It agrees to Question 3.

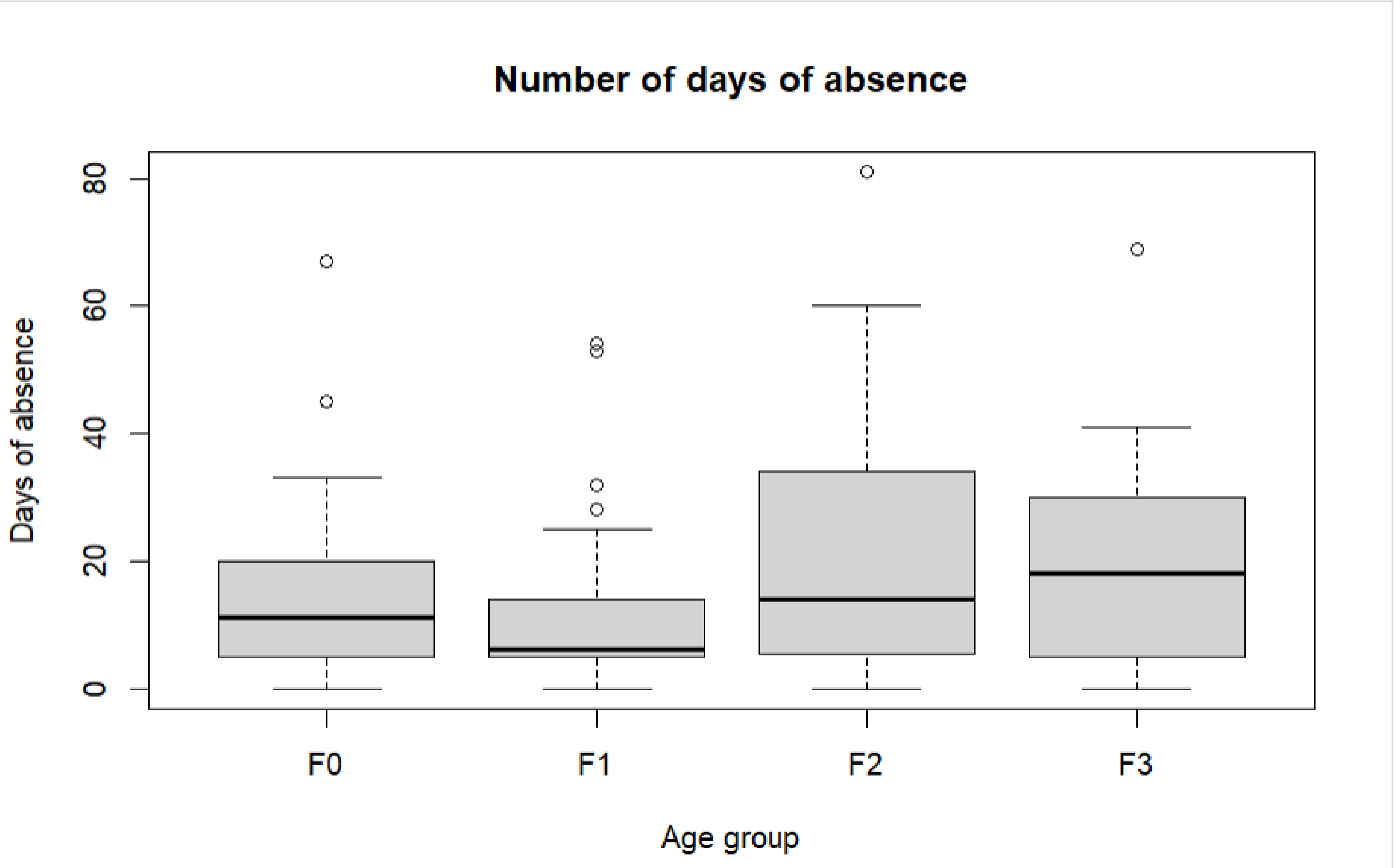
5.



```
> ageF0=filter(quine, Age=="F0")
> View(ageF0)
> ageF1=filter(quine,Age=="F1")
> ageF2=filter(quine,Age=="F2")
> ageF3=filter(quine,Age=="F3")
```

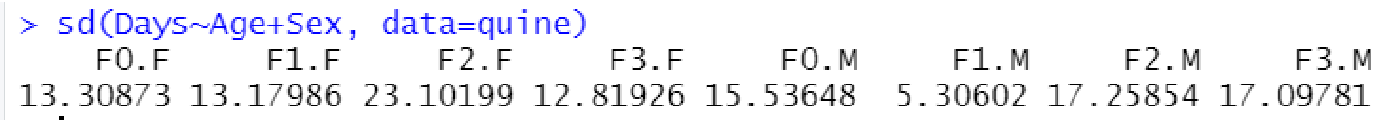


```
> boxplot(ageF0$Days,ageF1$Days,ageF2$Days,ageF3$Days,main="Number of days of absence",ylab
="Days of absence",xlab="Age group",names=c("F0","F1","F2","F3"))
```



- All four age groups had students who had perfect attendance (zero days of absence).
- In the first three quartiles, F1 group had the least absences with 14 days of absence. F2 and F3 had the most days of absence with 33 and 30 days respectively, while F0 had 20 days of absence.
- Following that, the age group that had the maximum of absent days was F2 with 81 days.

6.

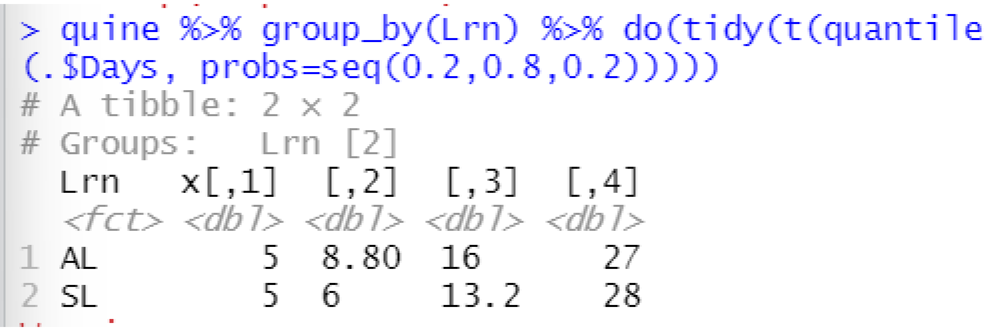


```
> sd(Days~Age+Sex, data=quine)
    F0.F     F1.F     F2.F     F3.F     F0.M     F1.M     F2.M     F3.M
13.30873 13.17986 23.10199 12.81926 15.53648  5.30602 17.25854 17.09781
```

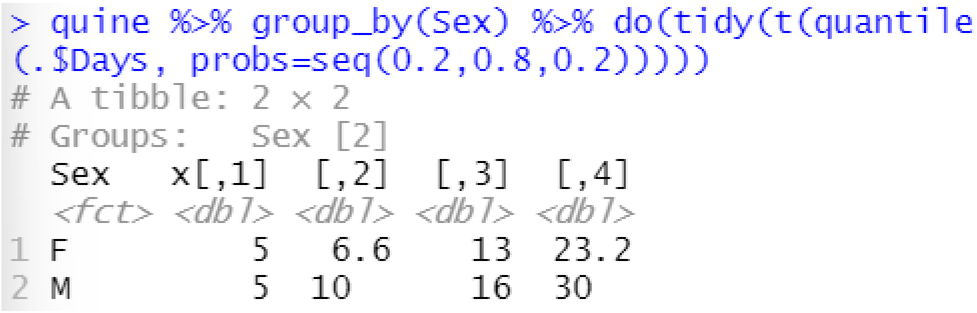
Male F1 age group has the most consistency (5.30602).

Female F2 age group has the least consistency (23.10199).

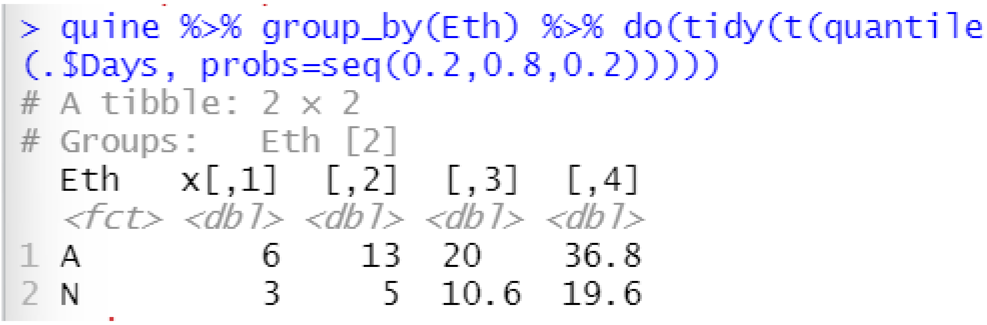
7.



```
> quine %>% group_by(Lrn) %>% do(tidy(t(quantile
(.$Days, probs=seq(0.2,0.8,0.2)))))
# A tibble: 2 x 2
# Groups:   Lrn [2]
  Lrn   x[,1]  [,2]  [,3]  [,4]
  <fct> <dbl> <dbl> <dbl> <dbl>
1 AL        5  8.80  16      27
2 SL        5  6     13.2    28
```



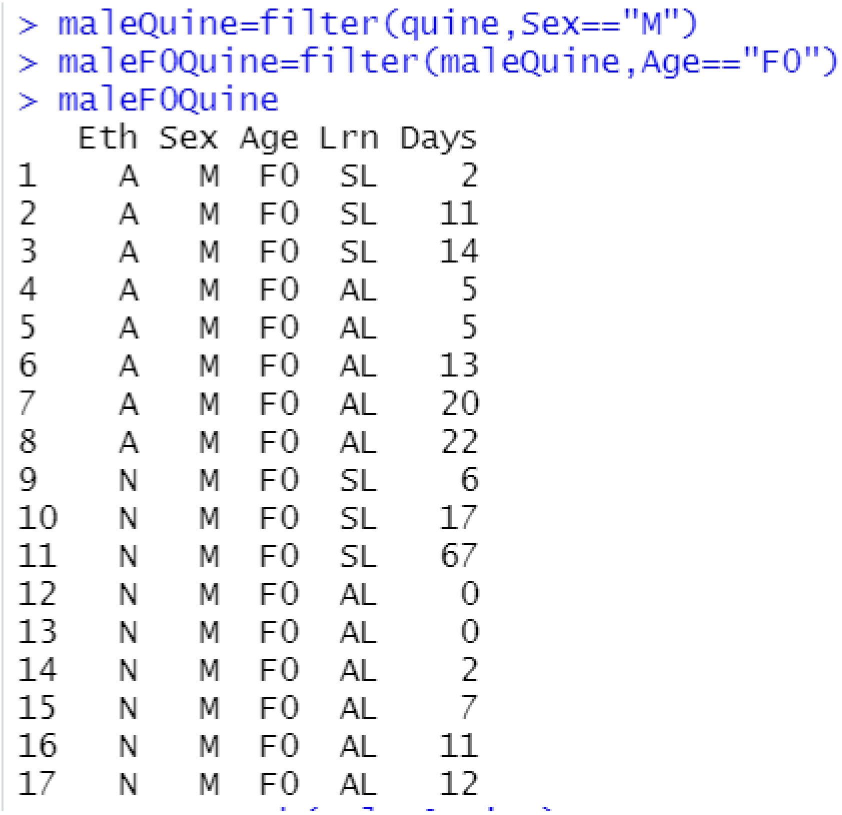
```
> quine %>% group_by(Sex) %>% do(tidy(t(quantile
(.$Days, probs=seq(0.2,0.8,0.2)))))
# A tibble: 2 x 2
# Groups:   Sex [2]
  Sex   x[,1]  [,2]  [,3]  [,4]
  <fct> <dbl> <dbl> <dbl> <dbl>
1 F         5   6.6    13  23.2
2 M         5  10      16  30  
```



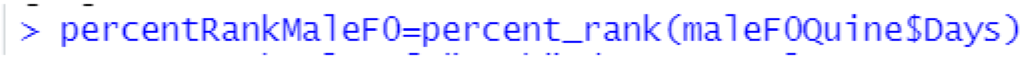
```
> quine %>% group_by(Eth) %>% do(tidy(t(quantile
(.$Days, probs=seq(0.2,0.8,0.2)))))
# A tibble: 2 x 2
# Groups:   Eth [2]
  Eth   x[,1]  [,2]  [,3]  [,4]
  <fct> <dbl> <dbl> <dbl> <dbl>
1 A         6    13  20    36.8
2 N         3     5  10.6  19.6
```

Grouped by Ethnicity produces the two groups that are the most different in terms of absences. If I were to estimate the number of days a student was going to be absent from class, the Learner Status and Sex categorization would allow us to make the most accurate prediction.

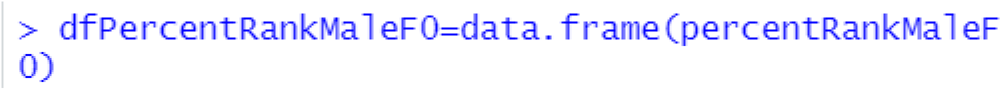
8.



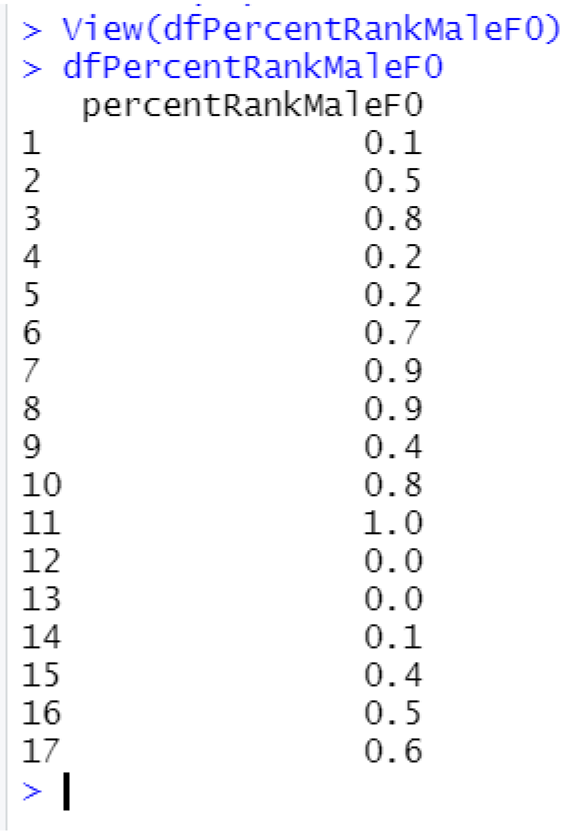
```
> maleQuine=filter(quine,Sex=="M")
> maleF0Quine=filter(maleQuine,Age=="F0")
> maleF0Quine
   Eth Sex Age Lrn Days
1    A   M  F0  SL    2
2    A   M  F0  SL   11
3    A   M  F0  SL   14
4    A   M  F0  AL    5
5    A   M  F0  AL    5
6    A   M  F0  AL   13
7    A   M  F0  AL   20
8    A   M  F0  AL   22
9    N   M  F0  SL    6
10   N   M  F0  SL   17
11   N   M  F0  SL   67
12   N   M  F0  AL    0
13   N   M  F0  AL    0
14   N   M  F0  AL    2
15   N   M  F0  AL    7
16   N   M  F0  AL   11
17   N   M  F0  AL   12
```



```
> percentRankMaleF0=percent_rank(maleF0Quine$Days)
```

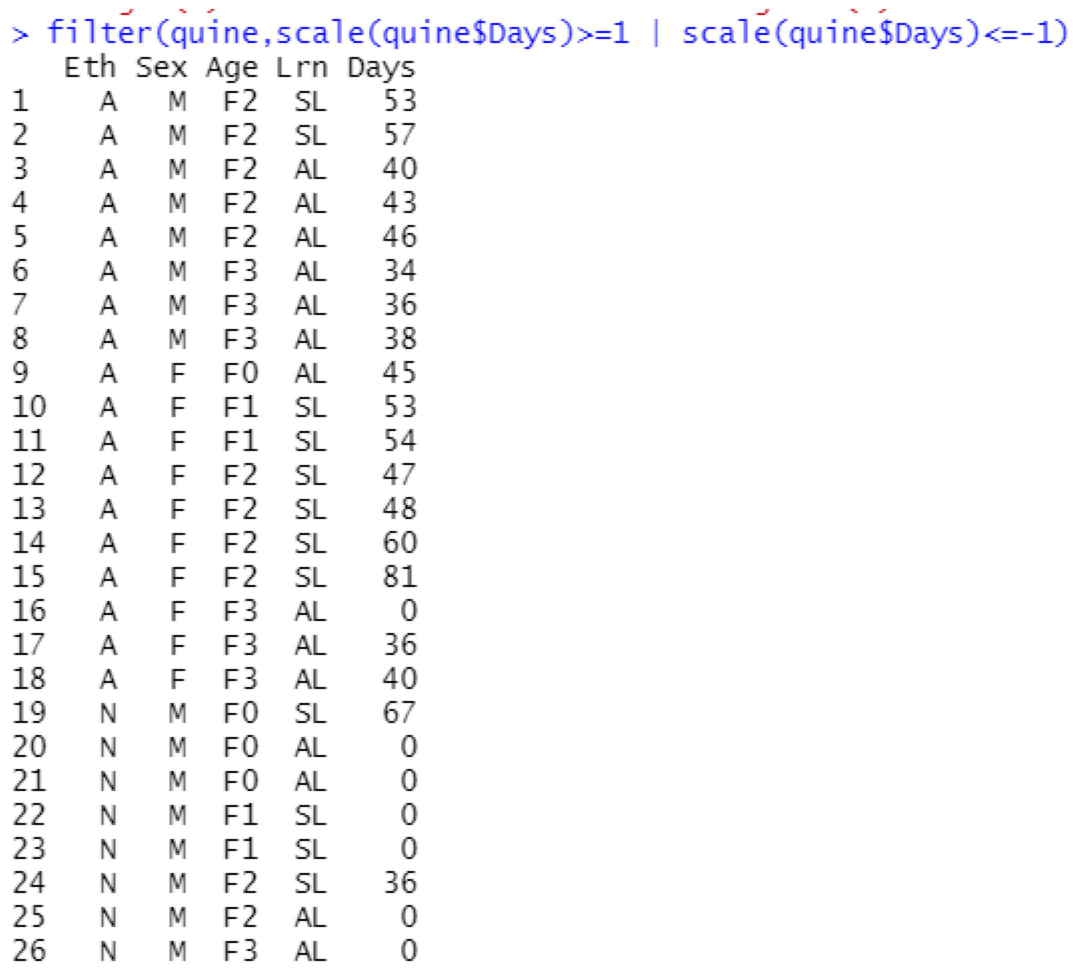


```
> dfPercentRankMaleF0=data.frame(percentRankMaleF
0)
```

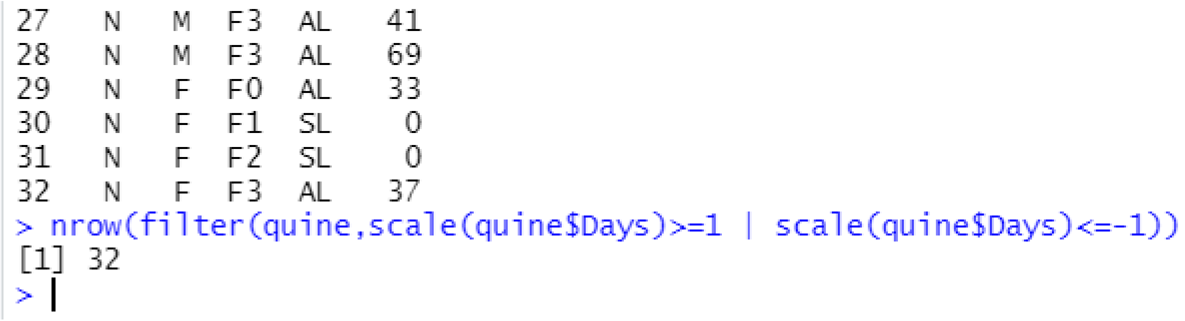


```
> View(dfPercentRankMaleF0)
> dfPercentRankMaleF0
   percentRankMaleF0
1                0.1
2                0.5
3                0.8
4                0.2
5                0.2
6                0.7
7                0.9
8                0.9
9                0.4
10               0.8
11               1.0
12               0.0
13               0.0
14               0.1
15               0.4
16               0.5
17               0.6
> 
```

9.

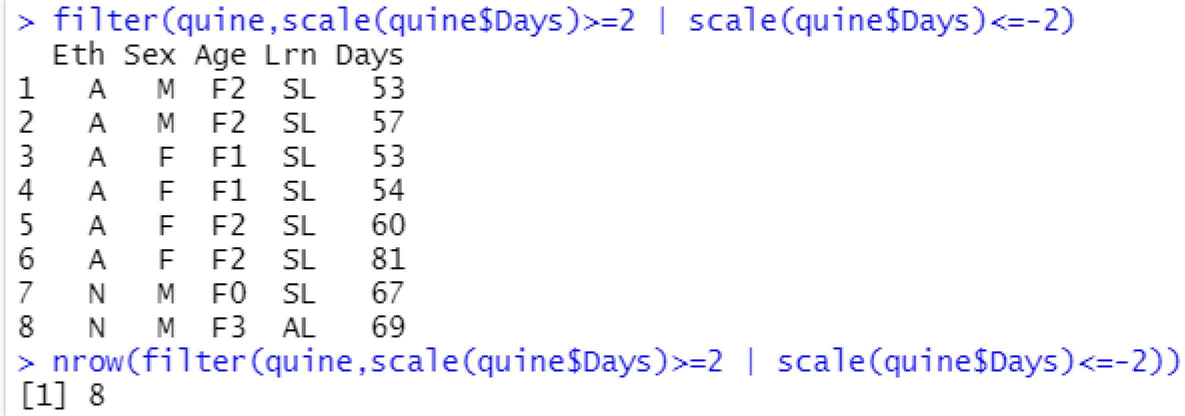


```
> filter(quine,scale(quine$Days)>=1 | scale(quine$Days)<=-1)
   Eth Sex Age Lrn Days
1    A   M  F2  SL   53
2    A   M  F2  SL   57
3    A   M  F2  AL   40
4    A   M  F2  AL   43
5    A   M  F2  AL   46
6    A   M  F3  AL   34
7    A   M  F3  AL   36
8    A   M  F3  AL   38
9    A   F  F0  AL   45
10   A   F  F1  SL   53
11   A   F  F1  SL   54
12   A   F  F2  SL   47
13   A   F  F2  SL   48
14   A   F  F2  SL   60
15   A   F  F2  SL   81
16   A   F  F3  AL    0
17   A   F  F3  AL   36
18   A   F  F3  AL   40
19   N   M  F0  SL   67
20   N   M  F0  AL    0
21   N   M  F0  AL    0
22   N   M  F1  SL    0
23   N   M  F1  SL    0
24   N   M  F2  SL   36
25   N   M  F2  AL    0
26   N   M  F3  AL    0
```



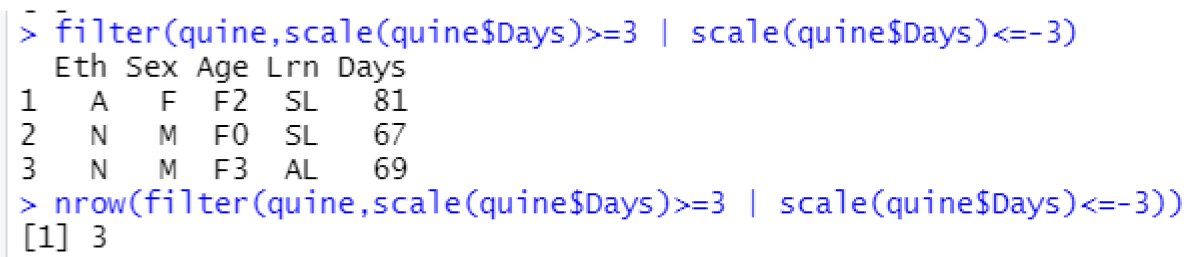
```
27   N   M  F3  AL   41
28   N   M  F3  AL   69
29   N   F  F0  AL   33
30   N   F  F1  SL    0
31   N   F  F2  SL    0
32   N   F  F3  AL   37
> nrow(filter(quine,scale(quine$Days)>=1 | scale(quine$Days)<=-1))
[1] 32
> |
```

There are 32 students who were absent for a number of days that were at least 1 standard deviation from the mean.



```
> filter(quine,scale(quine$Days)>=2 | scale(quine$Days)<=-2)
  Eth Sex Age Lrn Days
1   A   M  F2  SL   53
2   A   M  F2  SL   57
3   A   F  F1  SL   53
4   A   F  F1  SL   54
5   A   F  F2  SL   60
6   A   F  F2  SL   81
7   N   M  F0  SL   67
8   N   M  F3  AL   69
> nrow(filter(quine,scale(quine$Days)>=2 | scale(quine$Days)<=-2))
[1] 8
```

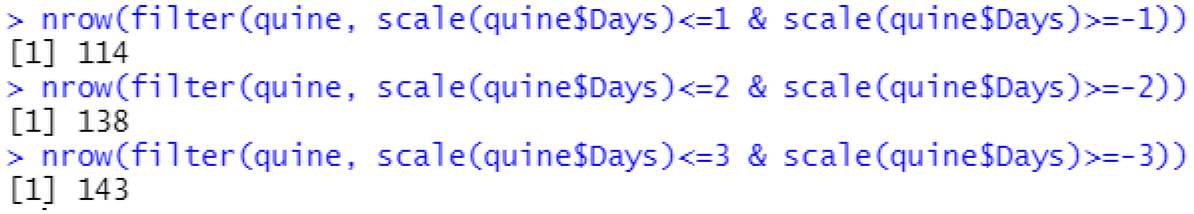
There are 8 students who were absent for a number of days that were at least 2 standard deviations from the mean.



```
> filter(quine,scale(quine$Days)>=3 | scale(quine$Days)<=-3)
  Eth Sex Age Lrn Days
1   A   F  F2  SL   81
2   N   M  F0  SL   67
3   N   M  F3  AL   69
> nrow(filter(quine,scale(quine$Days)>=3 | scale(quine$Days)<=-3))
[1] 3
```

There are 3 students who were absent for a number of days that were at least 3 standard deviations from the mean.

10.



```
> nrow(filter(quine, scale(quine$Days)<=1 & scale(quine$Days)>=-1))
[1] 114
> nrow(filter(quine, scale(quine$Days)<=2 & scale(quine$Days)>=-2))
[1] 138
> nrow(filter(quine, scale(quine$Days)<=3 & scale(quine$Days)>=-3))
[1] 143
```

There are 114 students (78%) who were absent for a number of days that were within 1 standard deviation from the mean.

There are 138 (95%) students who were absent for a number of days that were within 2 standard deviations from the mean.

There are 143 (97.95%) students who were absent for a number of days that were within 3 standard deviations from the mean.

This data set doesn't satisfy the Empirical Rule because the data is not normal and there are too few observations.

According to Chebyshev's Theorem, with k = 2, 1-1/k^2=1-1/4=0.75, means at least 75% of the data values will be found within 2 standard deviations of the mean. With k =3, it will be 0.85 meaning at least 85% of the data values will be found within 2 standard deviations of the mean. This is off by more than 10% of the actual percentage we got from the data(95% for k=2 and 97.95% for k=3). However, it is satisfied the Chebyshev's Theorem because this theorem is true no matter how the shape of the data is.