James Scruggs

M01255052

Today we will explore graphs useful for displaying quantitative data. Recall that quantitative data is date that deals with numbers rather than categories. In the titanic file, some of the columns are qualitative, such as *pclass* and *survived*, others are quantitative, such as *age*.

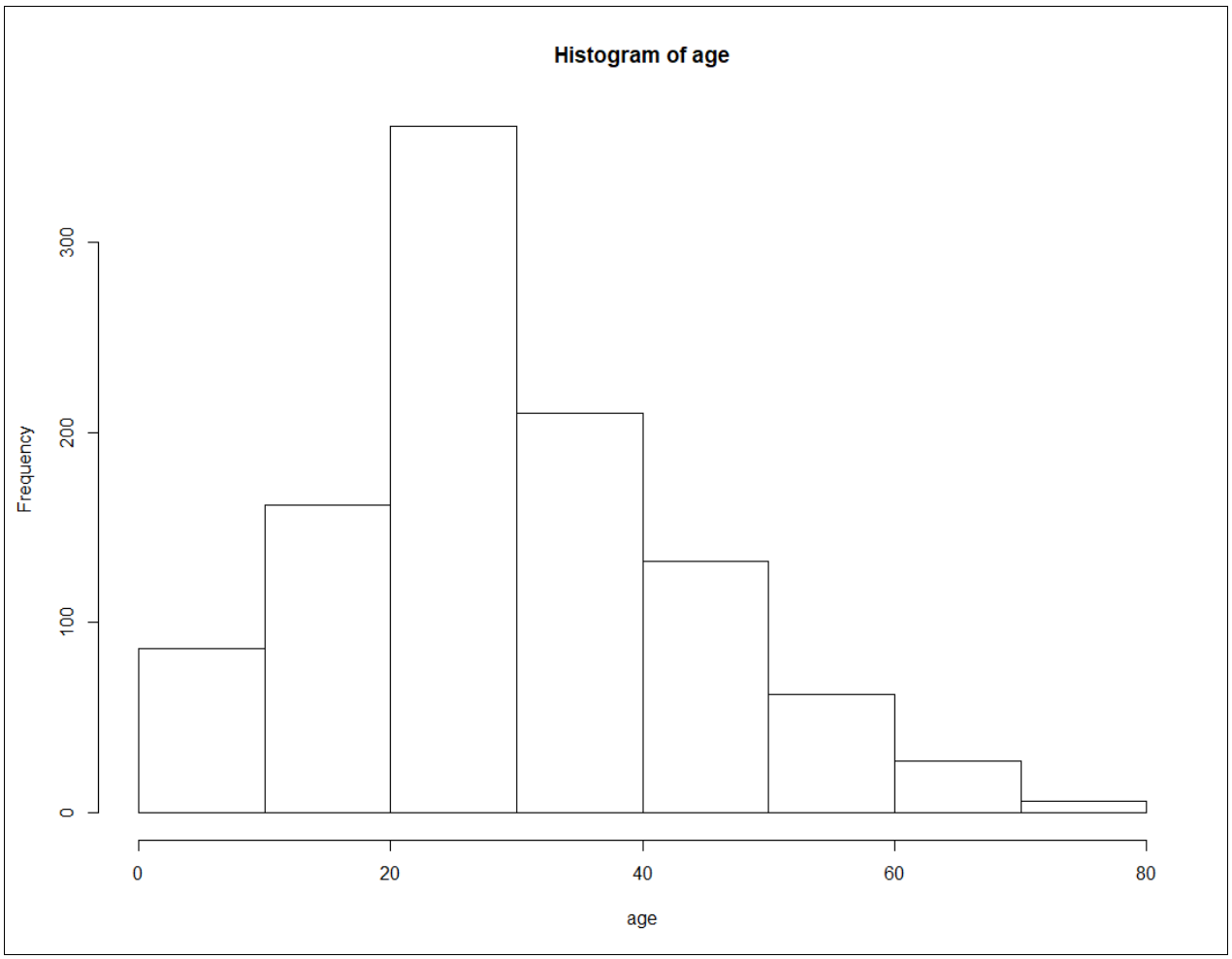
Histograms

1. Load and attach the data and plot the histograms for age column, then insert the graph in this file:

   #plot a histogram

   hist(age)

   hist(age,breaks=c(0,10,20,30,40,50,60,70,80)) #changes the bin sizes



2. Estimate based on the graph you found how many passengers were under age 10.

I would estimate about 90 are under 10 years of age.

3. Create a new histogram in which survived = 0.

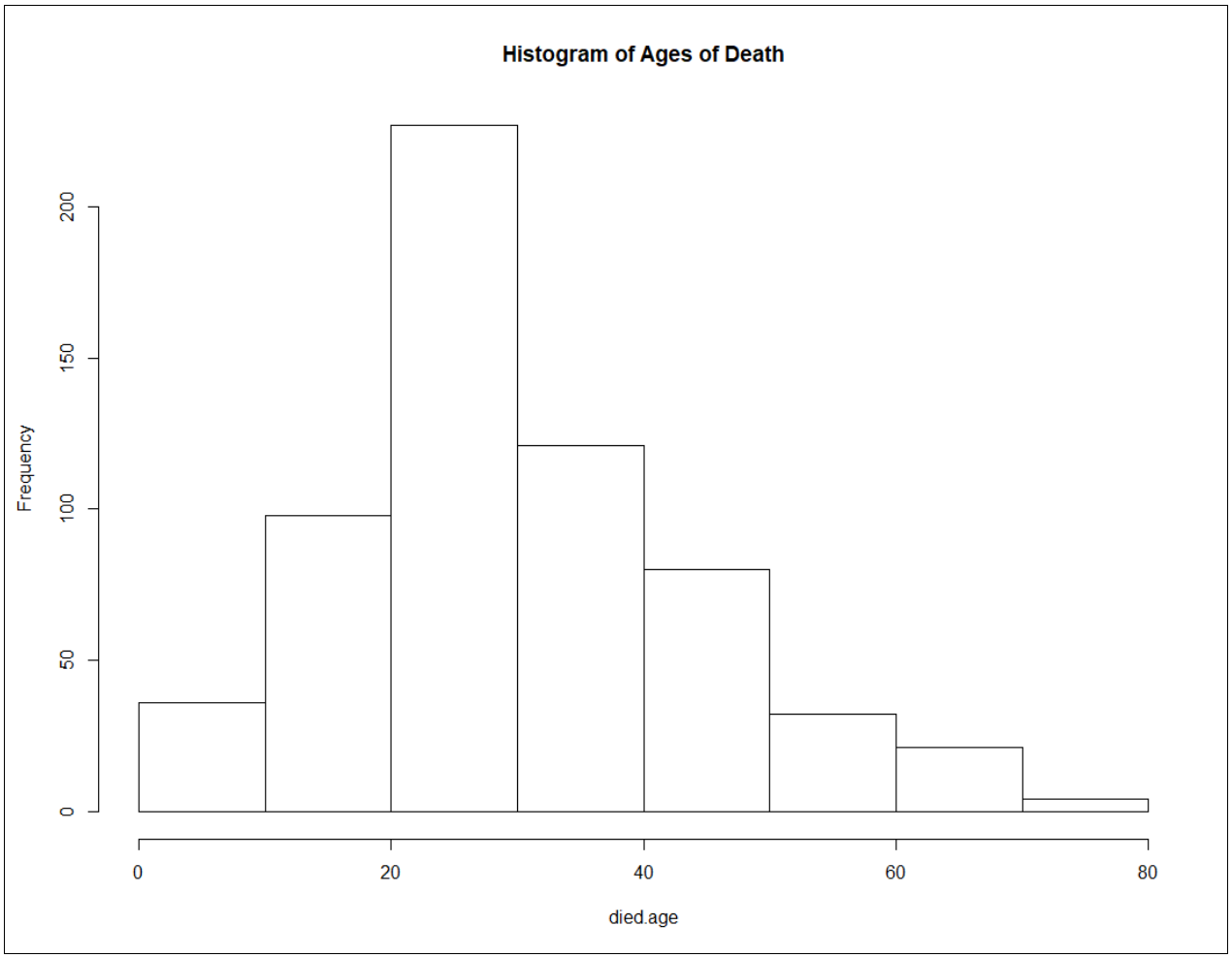
#Choose some observations from a variable by another variable

L <- survived==0

died.age <- age[L]

hist(died.age,breaks=c(0,10,20,30,40,50,60,70,80))

Insert the graph here:

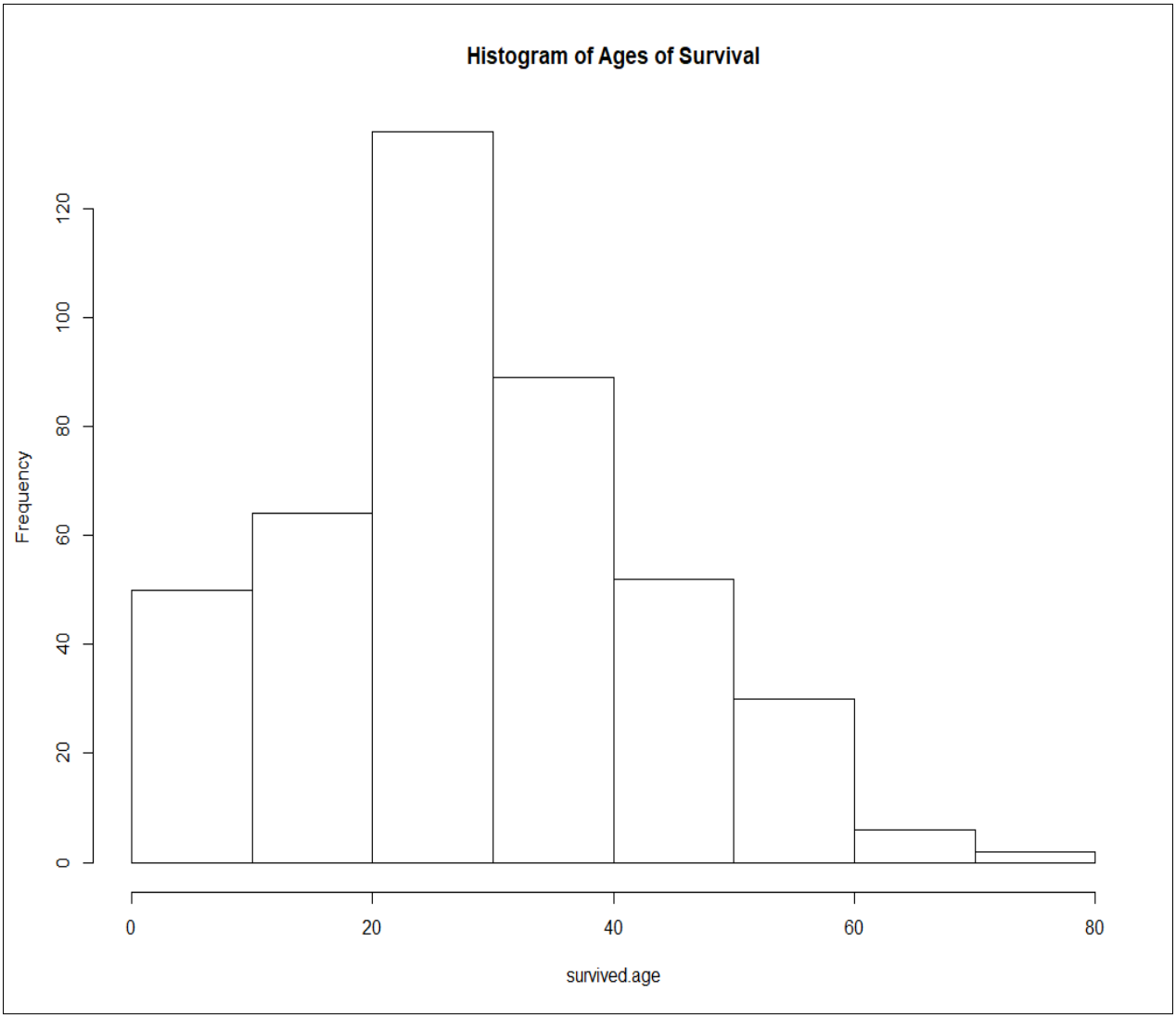


4. Create another histogram, in which survived = 1

survived.age=age[!L]

hist(survived.age,breaks=c(0,10,20,30,40,50,60,70,80))

Insert it here:



5. Compare the histograms of the ages of those that survived vs. those that died. Especially look at the number of children. What do these graphs imply?

It appears that far more people died than survived EXCEPT for those under 10 years of age. Those 10 years or younger seem to have the same if not more survivors than deaths.

Dot plots

6. Load the file babyboom.csv. This data set comes from an article in the Journal of Statistics Education, v7n3. It contains the time of birth, sex (1=girl, 2=boy), and birth weight for 44 babies born in a 24 hour period at a particular hospital. Create dot plots for birth weight.

#load the data

babyboom<-read.csv(file.choose(),header=TRUE)

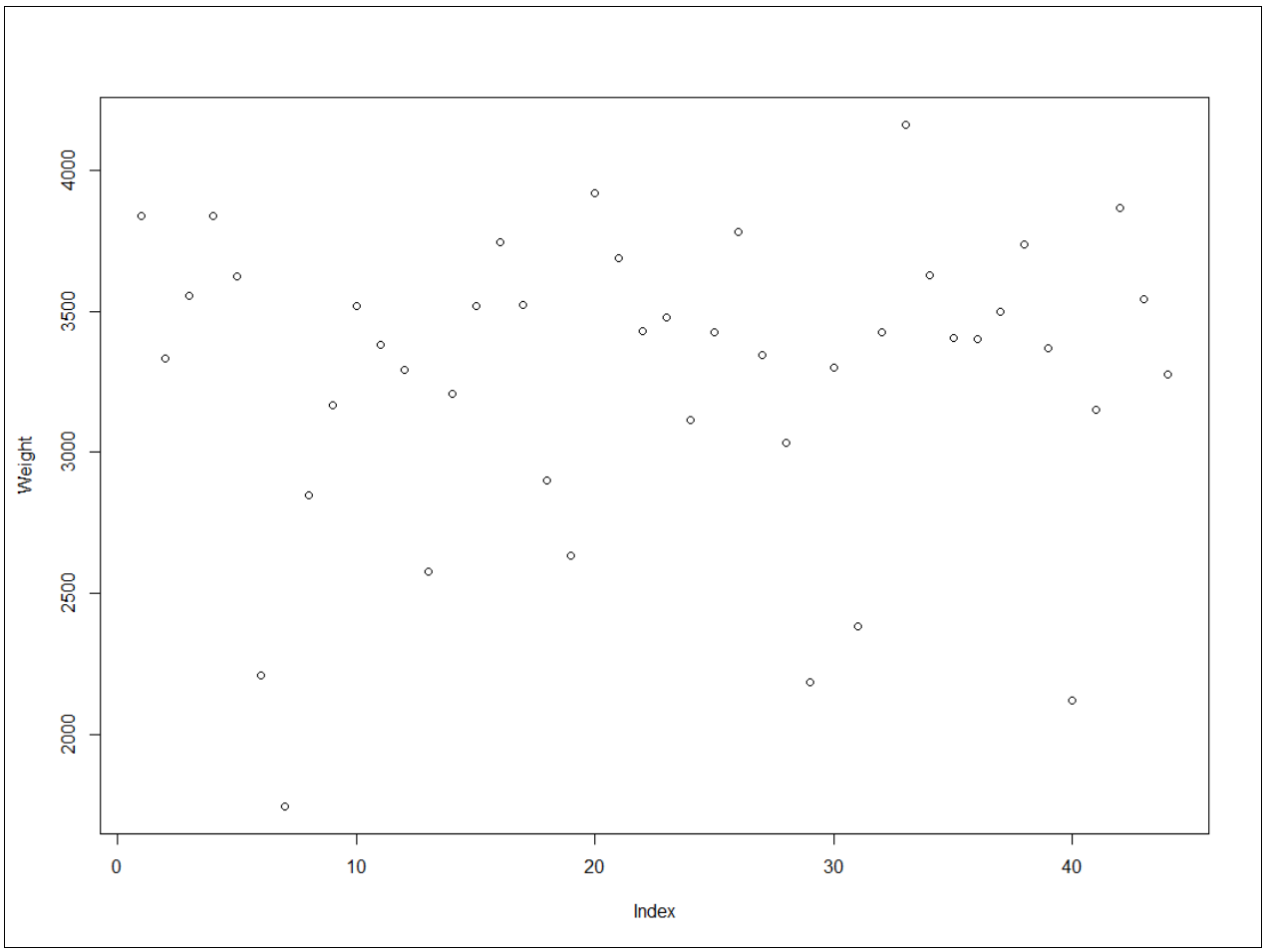
attach(babyboom)

#dot plot for birth weight

stripchart(Weight,method="stack",pch=19) #try leaving off the "method" and "pch" and see what happens

 # plot(Weight) Try this and compare the graphs.

Insert the graph below.

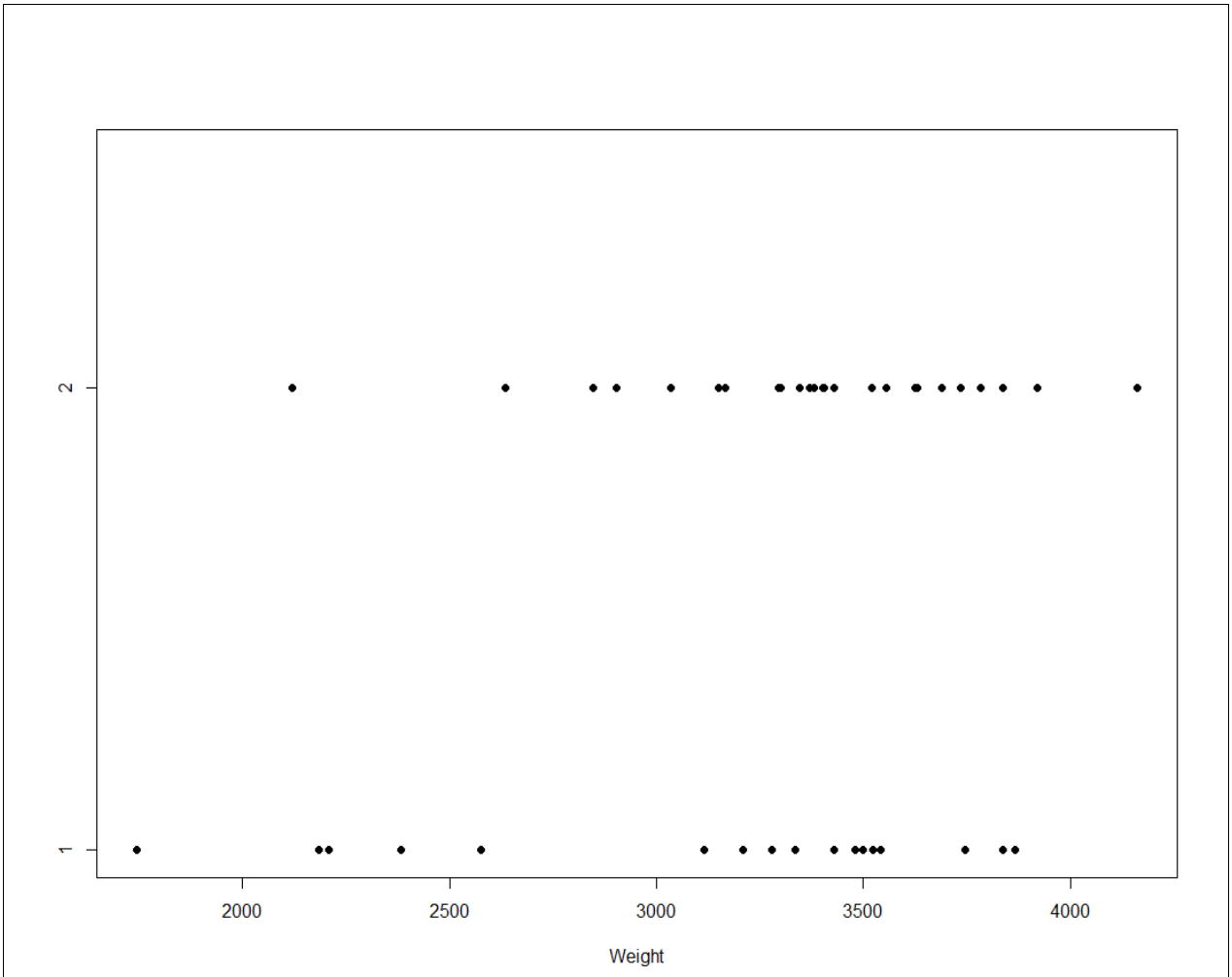


A dotplot displays a dot for each data point. If two data points are the same, then two dots will be stacked on top of each other.

7. Create dot plots again, but this time group the data by sex.

stripchart(Weight~Sex, method="stack", pch=19)

Insert your graph below. Does it look like there is a big difference in birth weight for boys vs. girls?



Yes, there does appear to be a significant difference. Boy's weight appears more distributed, the girl's weight seems concentrated at about 3500.

Comparing the two graphs

8. Clearly there are situations in which both histograms and dot plots can be used. When is it better to use a histogram, and when is it better to use a dot plot?

I would say that a **dot plot** is good if you are wanting to get an idea on how the data is spread at every point. However, if you need to group your data and need to get an idea of the distribution of the data then I would say to use a **histogram.**

Percentiles and Boxplots

9. *#summarize birth weight*
   summary(Weight)

Look at the values reported, and list the ones requested here:

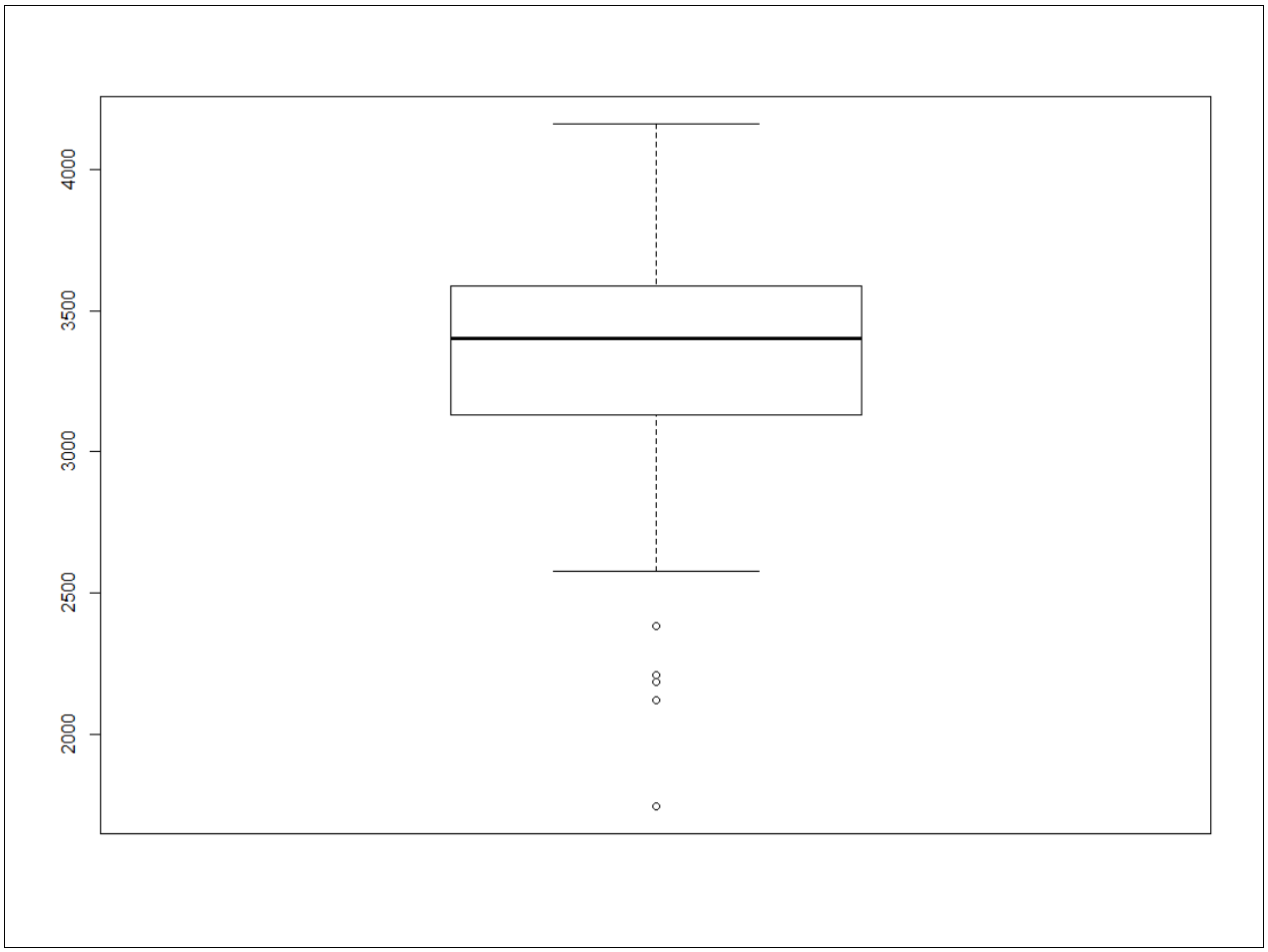
Min: 1745 Q1: 3142 Med: 3404 Q3: 3572 Max: 4162

These five numbers (Minimum, First Quartile, Median, Third Quartile, and Maximum) divide the data set into four pieces, with 25% of the data between each neighbor pair. Together, they are referred to as the five number summary

10. *#create boxplots*

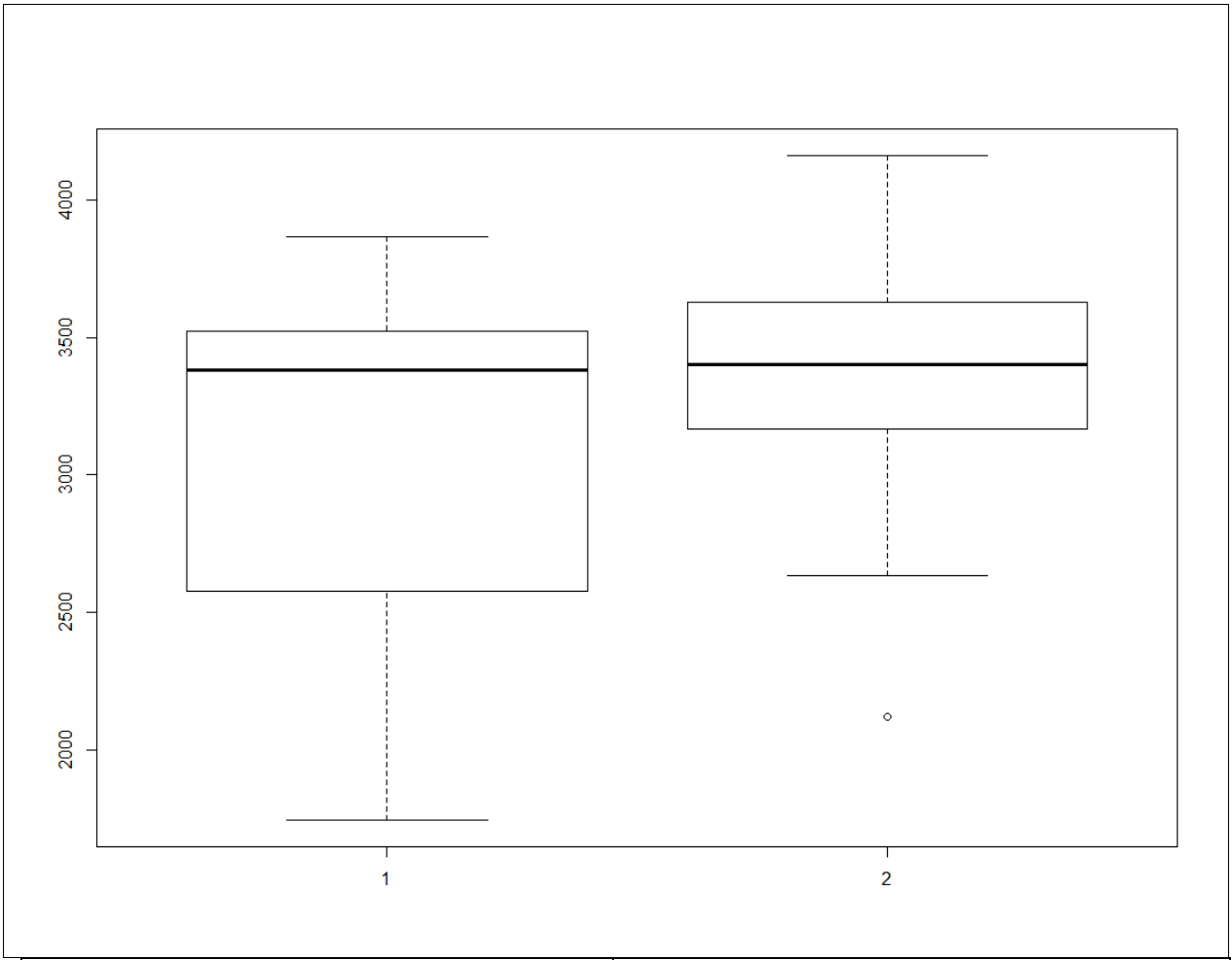
    boxplot(Weight)

Insert the graph below.



The boxplot is no more than a visualization of the five number summary. The location of the bottom of the graph matches the minimum, the location of the bottom of the box is the first quartile, the middle of the box is at the median, the top of the box is at the third quartile, and the top of the graph is at the maximum. Occasionally, as in this case, you will see a boxplot that has extra dots outside the "whiskers." These dots represent values that have been deemed unusual compared to the rest of the data set. These values are called outliers.

11. Create boxplots again, this time grouped by sex. (*#boxplots grouped by sex boxplot(Weight~Sex)*) Also find the five number summaries of birth weights grouped by sex.

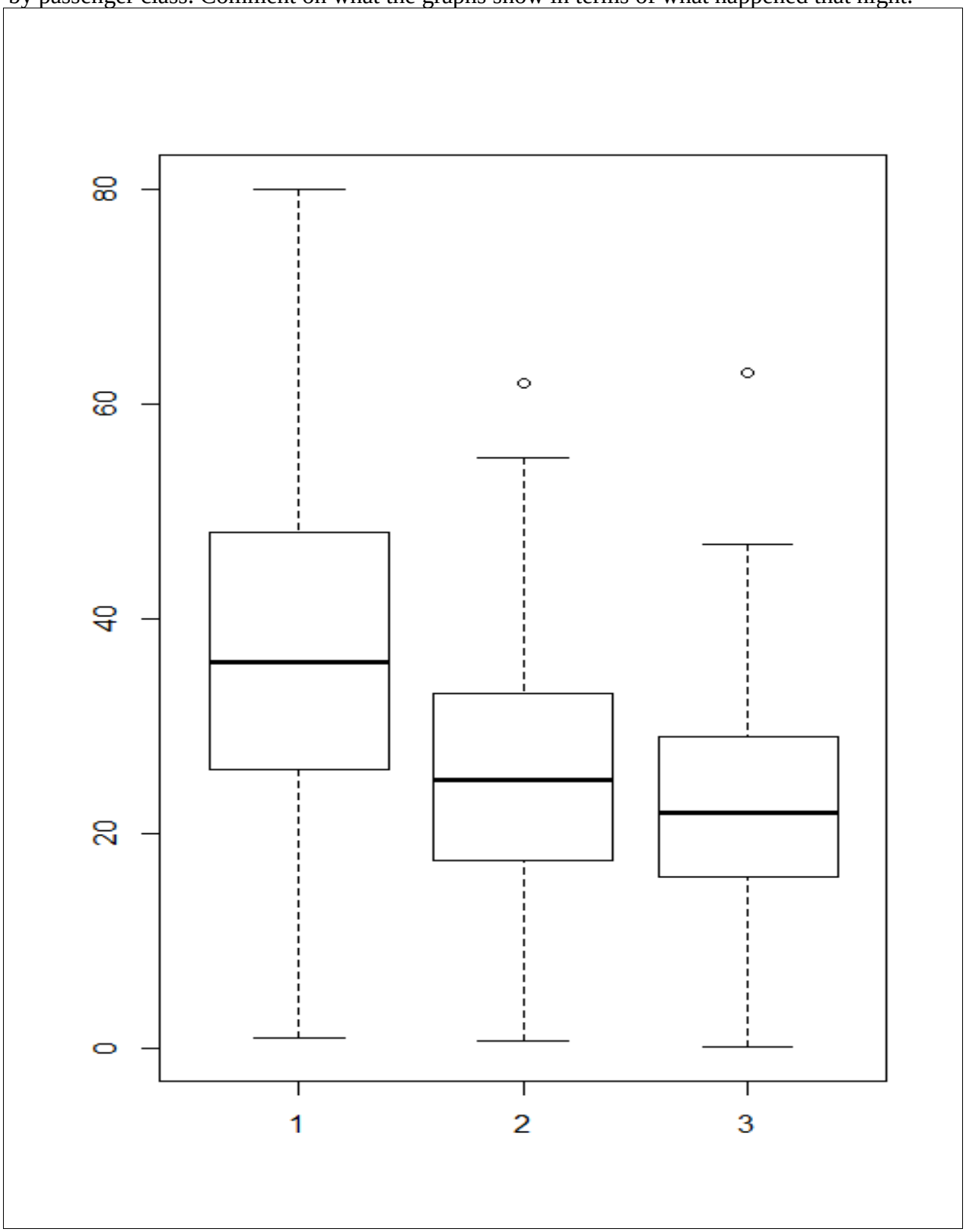


Boys: 26	Girls: 18
Min: 2121	Min: 1745
Q1: 3198	Q1: 2711
Med: 3404	Med: 3381
Q3: 3629	Q3: 3517
Max: 4162	Max: 3866

12. Describe any difference you observe between the boxplots for the boys' weights and the girls' weights.

It appears that boys box plot is more evenly drawn compared to the girls. It also appears that the boys have higher weights. Despite these differences, both of their medians appear to be about the same.

13. **Back to the Titanic**. Create appropriate graphs that show the ages of the survivors, grouped by passenger class. Comment on what the graphs show in terms of what happened that night.



It is quite clear from the graphs above that 1st class took priority followed by 2nd, and then 3rd class. This is obvious since the median is higher in the 1st class. The median in the 2nd class is still higher than 3rd class but not by much. The same can be said about the upper and lower quartiles.