Problem Set 3 Solutions

1.a) 38 individuals
1.b) R code: either `nrow(<epidemiology_data>)` or `summary(<epidemiology_data>)`
1.c) Average infectious period = 48.475 hours
R code: `<infectious_period> <- <epidemiology_data>$infection_end - <epidemiology_data>$infection_begin; mean(<infectious_period>)`

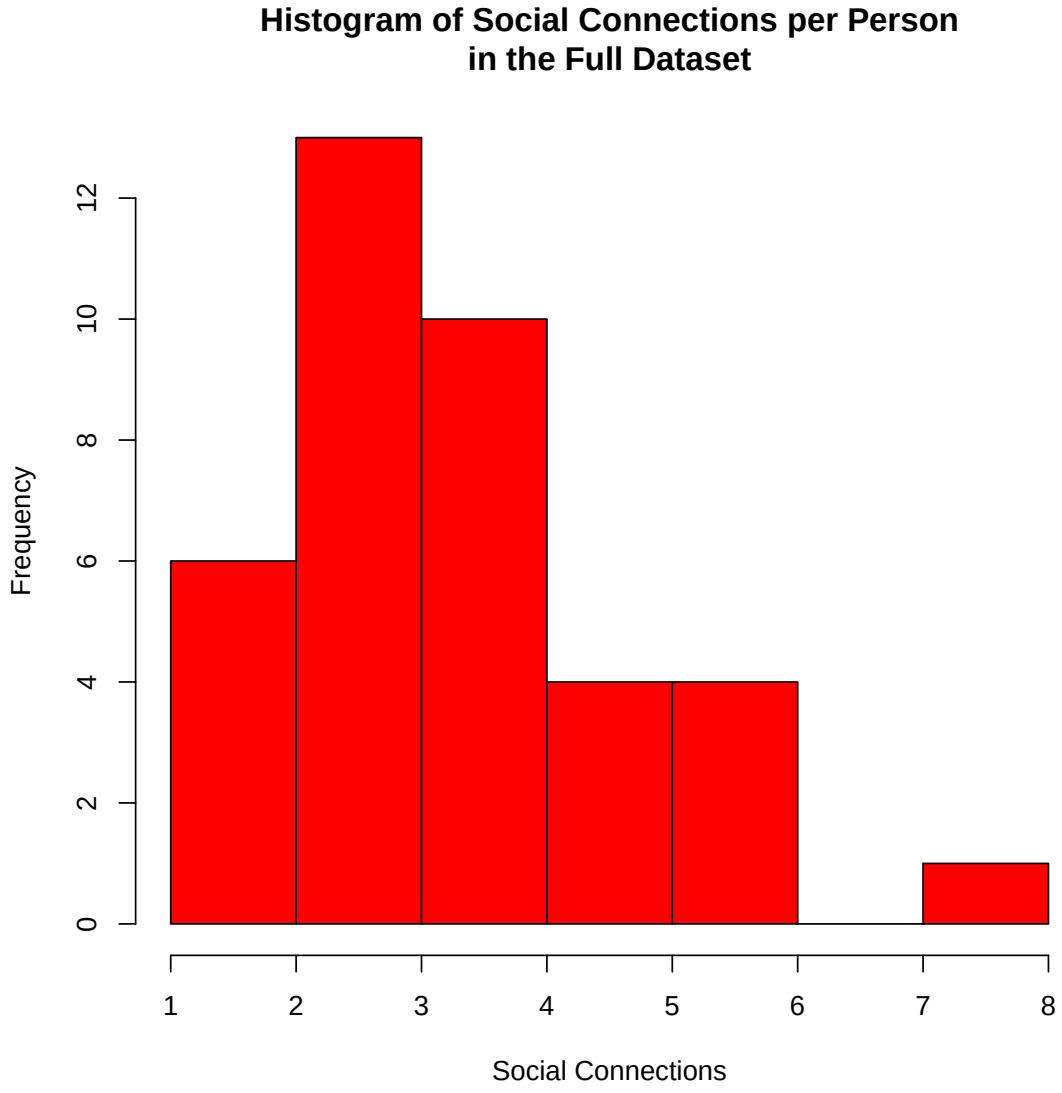
2.a) This simulation doesn't appear to model the observed spread of the parasite very well, because infections are not propagating through the population via social connections, but are rather just appearing randomly in individuals over time.
2.b) This simulation differed from the non-social simulation in that infections only appeared in an individual if that individual shared a social connection with at least one already infected individual. This sort of disease spread appears to be much more in line with what we observed in the actual spread of HEBV in the class, as the disease spread mainly through social connections (which were determined by reported social ties from the HEBV aftermath survey).

3.a) The pure non-social transmission model should have been best supported, most likely with an Akaike weight of 1. You should be extremely confident in this fit, since the output is telling you that there is a 100% probability that the non-social model is the best model.
3.b) The pure social transmission model should have been best supported, most likely with an Akaike weight of nearly 1. You should be extremely confident in this fit, since the output is telling you that there is an almost 100% probability that the social model is the best model.
3.c) You should predict that pure social transmission should be the most strongly supported model for the actual observed spread of HEBV in the classroom population because the simulated social transmission of HEBV most closely resembled what we actually observed in the graphs on the HEBV Case File.
3.d) The pure non-social transmission model was the best supported, with an Akaike weight of 1.
3.e) If you chose the logical prediction, that the social transmission model would be best supported, then your prediction was not correct. This may have been for a number of different reasons, but the most likely explanation is that some of the infections may have occurred at random, when one person without a social tie to another just by chance infected that individual.

4.a) The social and non-social transmission model was the best supported, with an Akaike weight of 0.9997844743. This result was probably found for the same reason as explained in 3.e, because HEBV spread through the population in a mostly social, but sometimes, random pattern. Allowing for the possibility of some non-social transmissions makes this model much more likely than a completely random (non-social) infection model.
4.b) Randomness probably plays a vital role in the spread of diseases in the real world. While most of the individuals that a person infects will be direct social connections, there are always chance interactions in the world, which allow diseases to break out of tight-knit, but small, social clusters.
4.c) The pure non-social transmission model was the best supported, with an Akaike weight of 1. These results are identical to those found in 3.d, and this is because, as always, if there is even one infection that does not occur through a single social connection (via random/non-social transmission), then the AIC value for the pure social transmission model will be infinite (this is what the "Inf" under the AIC values indicates), defaulting to a selection of the pure non-social transmission model.
4.d) The pure non-social transmission model was the best supported, with an Akaike weight of 0.7571955, the exact opposite of what was found in 4.a. This is because, after removing key infections from the population (since they were not symptomatic or "reported"), the spread of the disease would appear to be random. Therefore, if we use self-reporting of illness (which is generally the only means we have of tracking an outbreak), then we will not be able to properly determine the means of spread of a given disease.
4.e) If we use the results from 4.a to develop a public health protocol, then we might suggest social means of containing the outbreak (hand washing, covering coughs/sneezes, self-quarantine), whereas if we use the results from 4.d, we might suggest further research into possible non-social means of transmission (environmental, water, vectors) and control strategies for these types of pathogens. If we were to use the data from 4.d (which would likely be all that we have), we would likely misjudge the spread of the outbreak and waste valuable time and resources by following dead-end leads for control. Therefore, the strategy concluded from 4.a would be more effective (albeit, less realistically obtained) than that from 4.d.

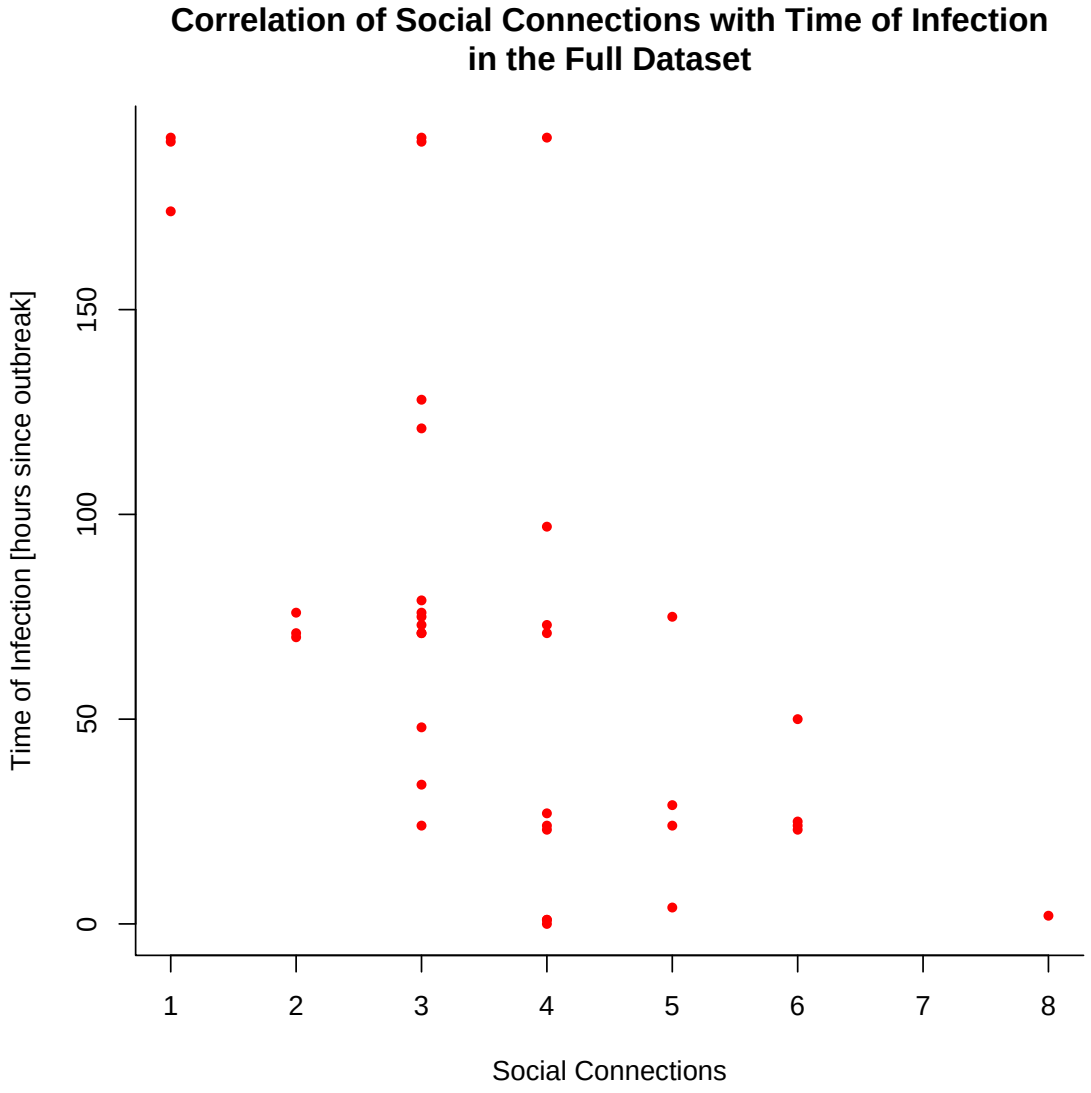
5.a) The variables are id, sex, symptomatic, heb_classes, connections, inf_time, and inf_group. `head()` shows you the first couple of lines of the data with each of the column headers, while `summary()` shows general summary stats for each variable, including mean, median, minimum, maximum, and quartiles.

5.b) Although the mean appears to be approximately 3 connections per individual, there is a considerable amount of variation in the number of connections per person, ranging from 1 to 8.



R code (to reproduce above graph): `hist(survey$connections, col="red", xlab="Social Connections", main="Histogram of Social Connections per Person\nin the Full Dataset")`

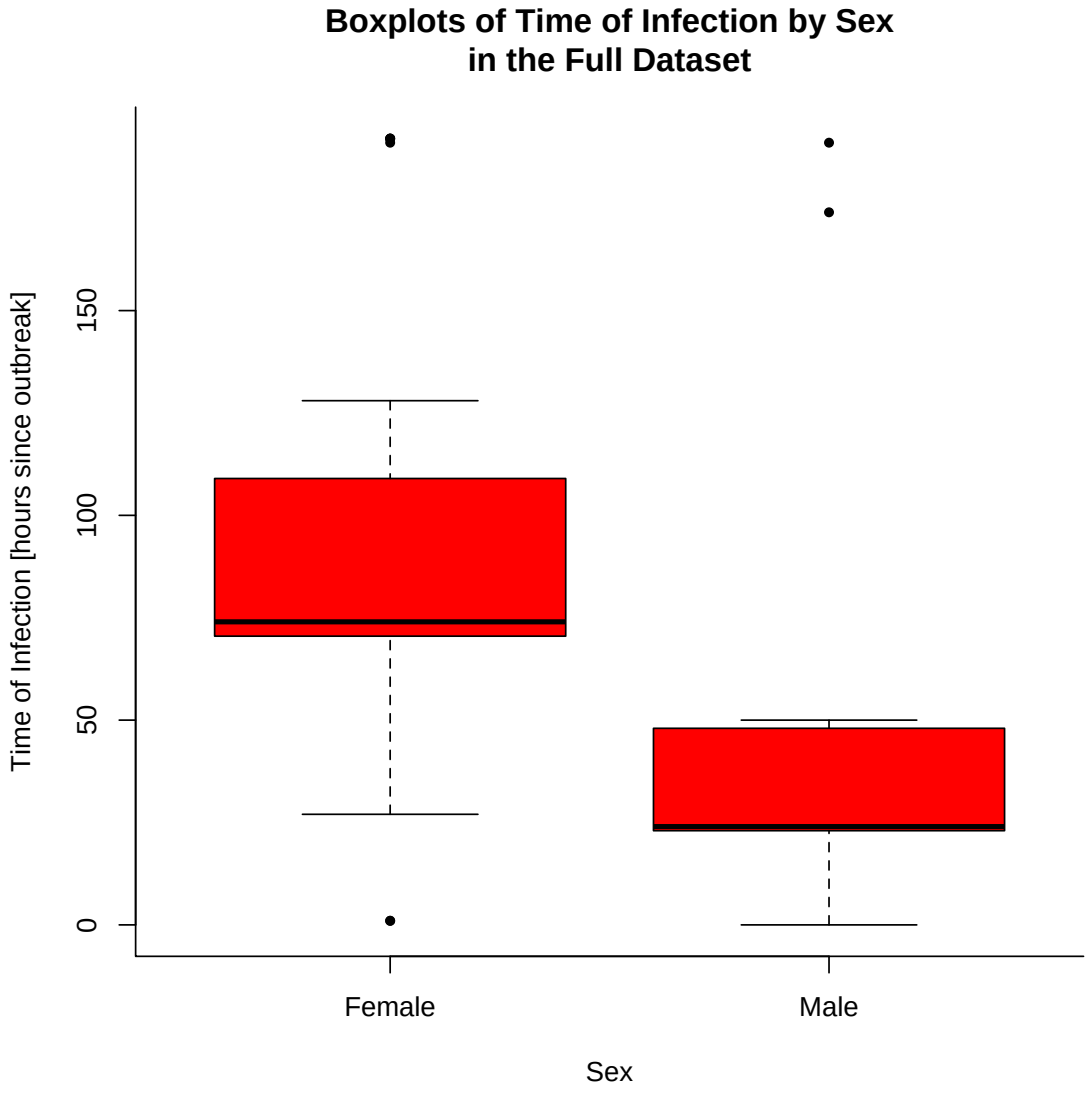
5.c) There appears to be a negative relationship between the number of social connections an individual has and the time at which that individual was infected (more connections = earlier infection).



R code (to reproduce above graph):

```
par(bty="l");
plot(survey$inf_time ~ survey$connections, xlab="Social
Connections", ylab="Time of Infection [hours since
outbreak]", main="Correlation of Social Connections with
Time of Infection\nin the Full Dataset", pch=20, col="red")
```

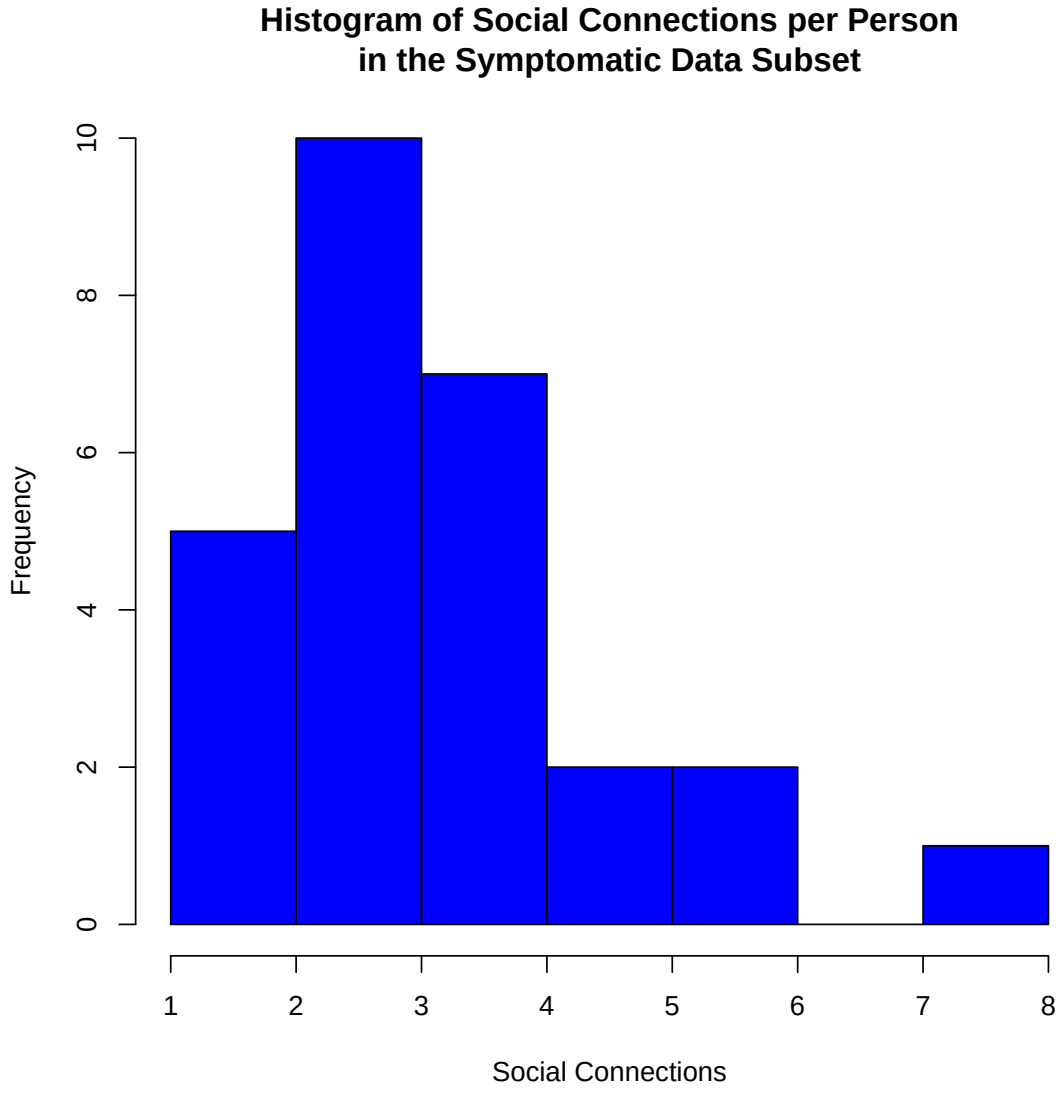
5.d) Males appear to have gotten HEBV earlier on average than females.



R code (to reproduce above graph):

```
survey$sex <- gsub("F", "Female", survey$sex); survey$sex <- gsub("M", "Male", survey$sex); survey$sex <- as.factor(survey$sex); par(bty="l"); plot(survey$inf_time ~ survey$sex, xlab="Sex", ylab="Time of Infection [hours since outbreak]", main="Boxplots of Time of Infection by Sex\nin the Full Dataset", pch=20, col="red")
```

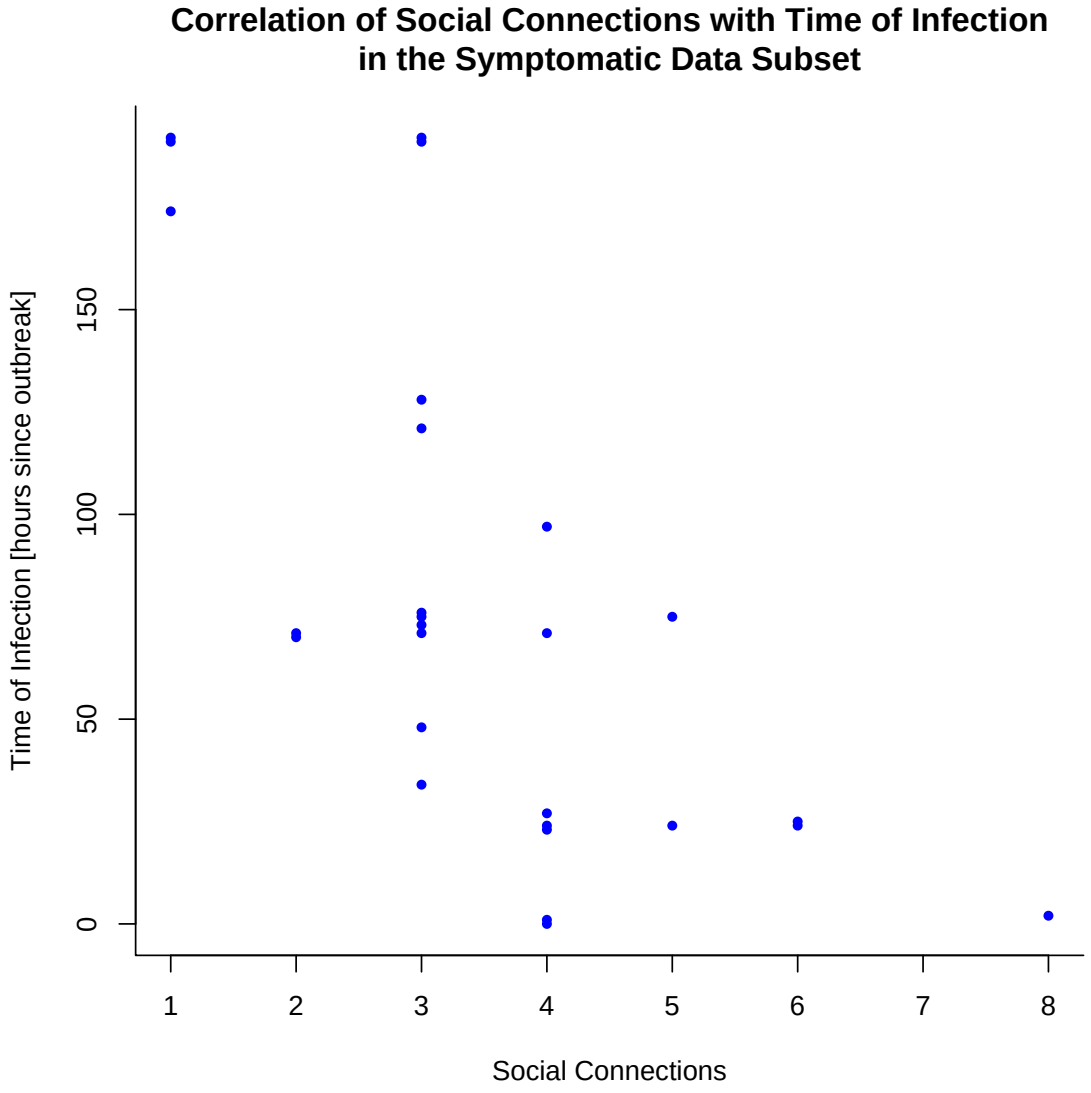
6.a) This histogram doesn't appear much different from that reported in 5.b, the same basic distribution still appears.



R code (to reproduce above graph):

```
hist(symptomatic.Y$connections, col="blue", xlab="Social
Connections", main="Histogram of Social Connections per
Person\nin the Symptomatic Data Subset")
```

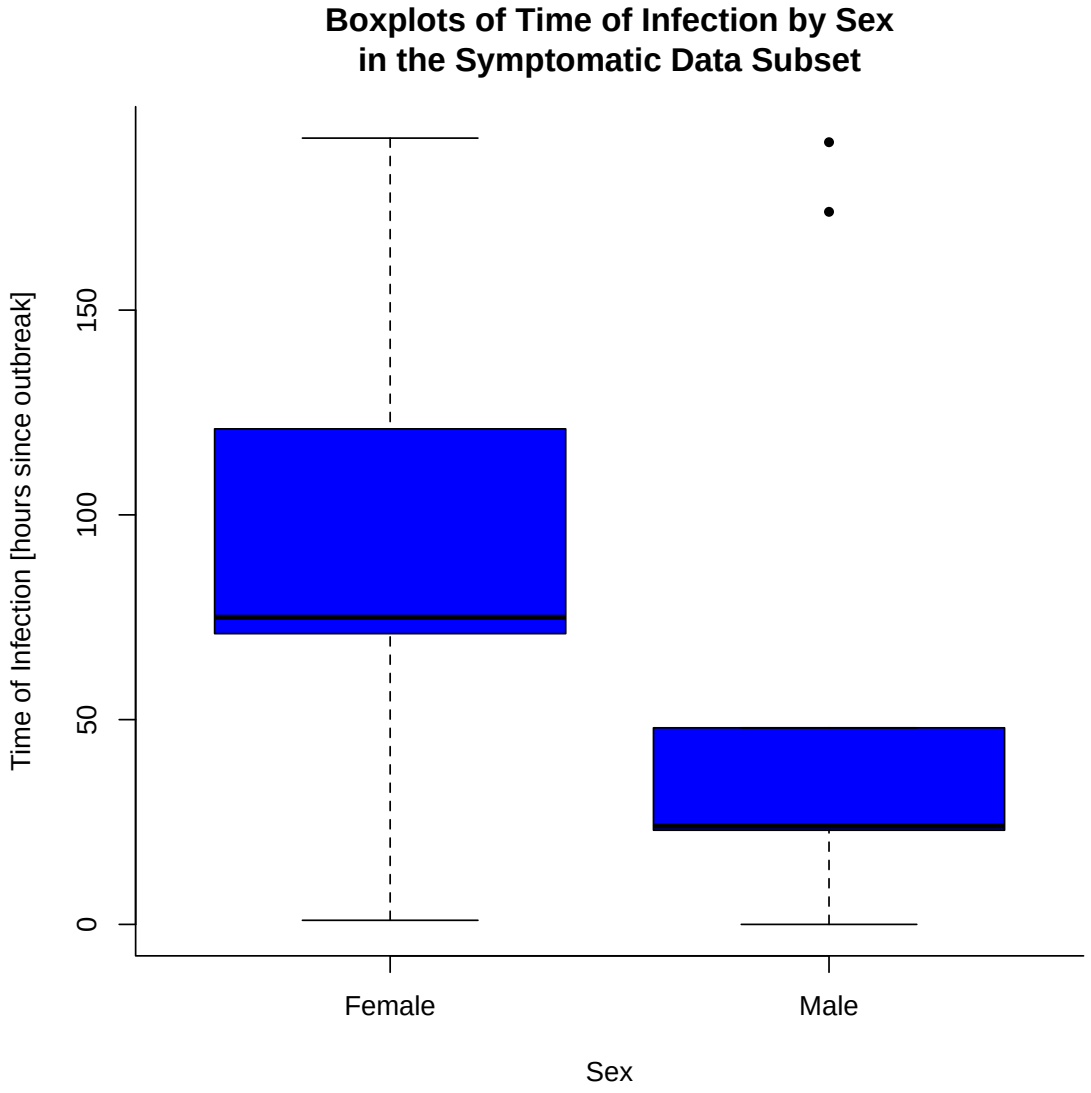
6.b) This plot also doesn't appear to be much different from the one provided in 5.c, as there is still a marked negative relationship between time of infection and connections.



R code (to reproduce above graph):

```
par(bty="l");
plot(symptomatic.Y$inf_time ~ symptomatic.Y$connections,
xlab="Social Connections", ylab="Time of Infection [hours
since outbreak]", main="Correlation of Social Connections
with Time of Infection\nin the Symptomatic Data Subset",
pch=20, col="blue")
```

6.c) Just like the other two graphs, the original relationship of males becoming infected earlier than females also appears.



R code (to reproduce above graph):

```
symptomatic.Y$sex <- gsub("F", "Female", symptomatic.Y$sex); symptomatic.Y$sex <- gsub("M", "Male", symptomatic.Y$sex); symptomatic.Y$sex <- as.factor(symptomatic.Y$sex); par(bty="l"); plot(symptomatic.Y$inf_time ~ symptomatic.Y$sex, xlab="Sex", ylab="Time of Infection [hours since outbreak]", main="Boxplots of Time of Infection by Sex\nin the Symptomatic Data Subset", pch=20, col="blue")
```

6.d) Symptomatic reporting did not appear to have a substantial effect on our conclusions about risk factors in this situation. This result was in stark contrast to our conclusions on the transmission mode of HEBV using either the full dataset or the symptomatic subset. The reason for this was that holes in our knowledge of the transmission pattern had a much larger effect on conclusions of our tests in NBDA than they did in risk factor analyses (as explained in more detail in the answer to 4.d).