Lab report 1

Comments and tips regarding the coursework are given in italics.

The data

The dataset was compiled by Columbia Business School professors Ray Fisman and Sheena Iyengar for their paper Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment. The data comprises responses to questionnaires given to participants of speed dating events. The description of the data can be found in this data key document.

Note: When using data not collected by ourselves we usually acknowledge the researchers who collected the data and provide information on how to obtain access to the data. However, the acknowledgement should be written in your own words, not simply copying word by word sentences from the practical (this is poor academic style, if not plagiarism).

Note: Later in the report we will provide an interpretation of some plots. Interpretations should be embedded in the application context, so it is important to provide some background on the relevant data. This also helps us treat the data appropriately in our analysis.

Data import, subsetting and export

To import the data we use the command:

```
SpeedRawData <- read.csv("SpeedDatingRawData.csv")
```

Note: we defined a relative directory path.

We then select a small number of variables from the data set:

```
SpeedData <- SpeedRawData[c("id", "attr", "gender", "like")]
```

The variables are:

- `id`: participant number within a speed dating event;
- `gender`: gender of participant;
- `attr`: a rating of the date partner's attractiveness on a scale from 1 to 10 with 1 = awful and 10 = great;
- `like`: responses to the question "Overall, how much do you like this person?" on a scale from 1 to 10 with 1 = don't like at all and 10 = like a lot.

For the latter two variables, the participants were given the option of awarding a rating to indicate that they were unable to form an opinion based on their conversation with the speed data partner.

We can save the data frame `SpeedData` in a csv file as follows.

```
write.table(SpeedData, file = "LabReport1Data.csv", sep = ",", row.names=FALSE)
```

Boxplots

Before we produce boxplots for the likeability scores grouped by attractiveness scores we check whether all participants have used the prescribed scale and, if some have not, how many.

```
library(knitr)
Table1 <- table(SpeedData$attr, exclude = NULL)
Table2 <- table(SpeedData$like, exclude = NULL)
kable(t(as.matrix(Table1)), caption = "Frequency table of attractiveness scores")
```

Table 1: Frequency table of attractiveness scores

0	1	2	3	3.5	4	5	6	6.5	7	7.5	8	8.5	9	9.5	9.9	10	NA
8	109	244	390	1	749	1260	1658	7	1646	3	1231	1	540	3	1	325	202

```
kable(t(as.matrix(Table2)), caption = "Frequency table of likeability scores")
```

Table 2: Frequency table of likeability scores

0	1	2	3	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	9.7	10	NA
8	110	223	396	645	3	1319	2	1709	20	1816	6	1274	9	412	3	1	182	240

We note that some participants were unable to form an opinion and thus responded with the NA option. Some used non-integer responses and others used zero which is outside the range of the prescribed scale. There are various options we may consider in addressing these but will not discuss here. For simplicity, we round up any non-integer responses and remove observations with response equal to zero or NA. Given we have 8378 observations in total, this amounts to removing just over 3% of observations.

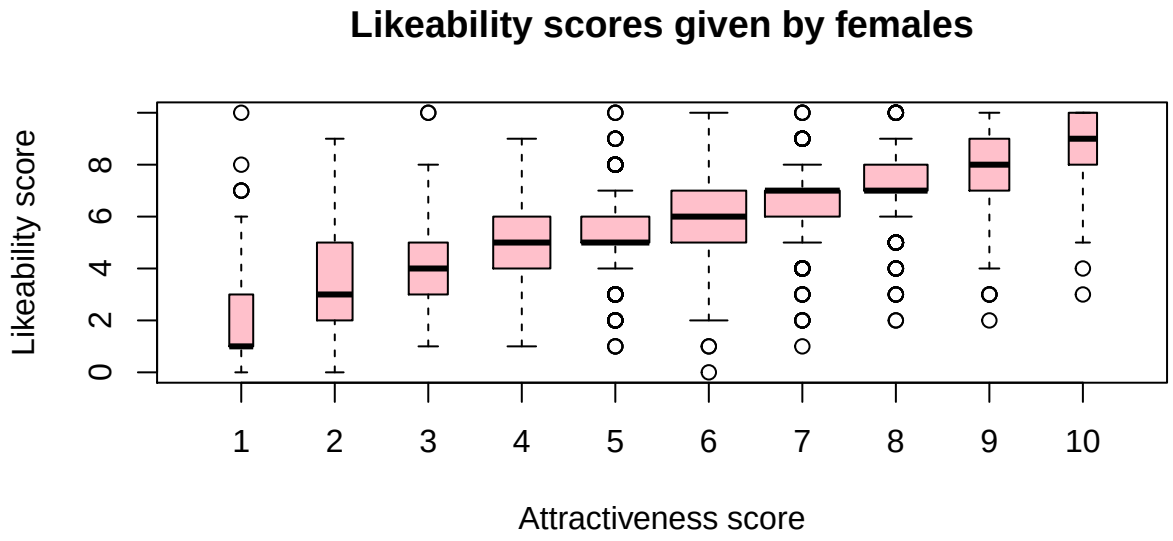
Note: in the context of the lab report you were expected to ***at least*** *comment on the zero, non-integer and NA values. If you remove observations it is good practice at least to state how many observations are being removed.*

From the frequency plot we also observe that the size of groups varies from just over 100 to just under 1900. The R function `boxplot` has an option 'varwidth' which, if set to TRUE, will adjust the widths of the boxplots to being proportional to the square-roots of the group sizes.

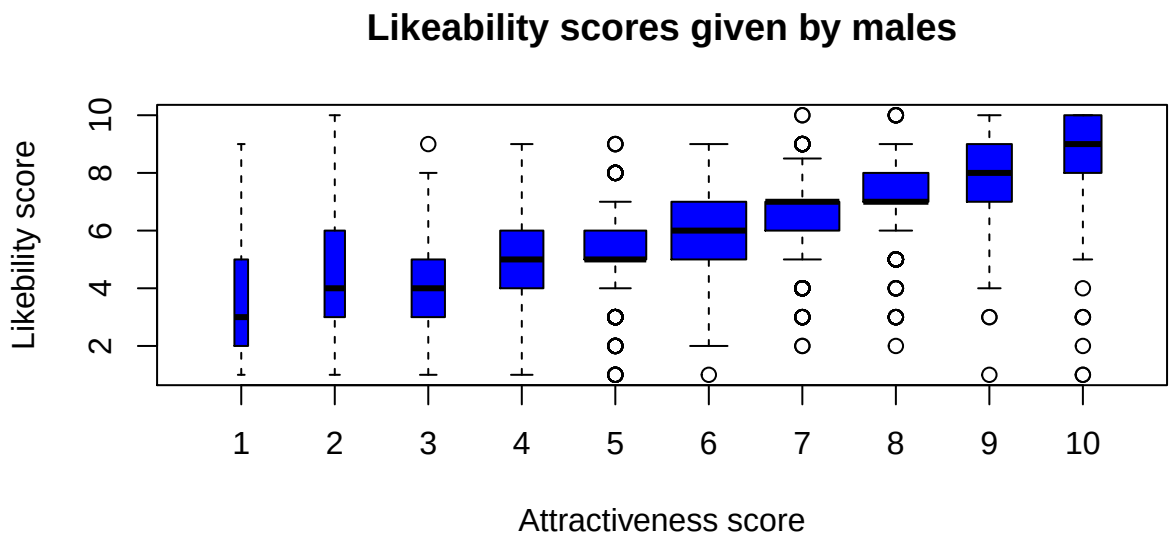
Note: you may not have been aware of the variable width option for boxplot but should have made some comment about the range of group sizes.

The following code produces for each gender boxplots of likeability scores grouped by attractiveness scores.

```
boxplot(like~round(attr), data = SpeedData, subset = (gender == 0) & (attr != 0),
        varwidth = TRUE, col = "pink", main="Likeability scores given by females",
        xlab = "Attractiveness score", ylab = "Likeability score" )
```




```
boxplot(like~round(attr), data = SpeedData, subset = (gender == 1) & (attr != 0),
        varwidth = TRUE, col = "blue", main="Likeability scores given by males",
        xlab = "Attractiveness score", ylab = "Likebility score" )
```



Note: in Practical 1 the title of the boxplot was "Importance of Attractiveness". This was because participants were asked how important they felt attractiveness was in their dating decisions. So using the same title for the above boxplots is somewhat misleading.

Interpretation

The boxplots show that the median score for likability increases with increasing attractiveness score. This can be observed both for female and male participants. The boxplots also show relatively large

interquartile ranges and the occurence of outliers, meaning that even given the attractiveness score we observe substantial variation in the likeability score.

Note: some students discussed averages which suggests the use of means. However, boxplots display medians.