Session 2 - Reading Data Manipulation

Alex Mounsey

30/09/20

The Student Questionnaire Dataset

The file `MATH513_Questionnaire_Data.csv` contains information regarding a group of University of Plymouth students. A transcription error has been introduced for illustration purposes. Your task is to read in the data, perform data manipulation, and produce appropriate plots of specific variables.

Read in the data from the file `MATH513_Questionnaire_Data.csv` using the function `read_csv()` from the `readr` package, and take a look at the data:

```
q_data <- read_csv('../data/MATH513_Questionnaire_Data.csv')
head(q_data)
```

```
## # A tibble: 6 x 19
##   Height   Age Sex   BirthPlace SiblingsNo EatMeat DrinkCoffee LikeBeer Sports
##    <dbl> <dbl> <chr> <chr>           <dbl> <chr>   <chr>       <chr>    <chr>
## 1    170  23   Fema~ essex               1 Yes     Yes         No       Yes
## 2    188  22.4 Male  London              1 Yes     Yes         No       No
## 3    180  30.1 Male  Athens              0 Yes     Yes         Yes      Yes
## 4    185  21   Male  China               0 Yes     Yes         Yes      Yes
## 5    170  22.1 Fema~ Plymouth            2 Yes     Yes         No       No
## 6    182  25   Male  Nigeria             4 Yes     No          No       Yes
## # ... with 10 more variables: Driver <chr>, LeftHanded <chr>, Abroad <chr>,
## #   Sleep <dbl>, Rent <dbl>, Happy_accommodation <chr>, Distance <dbl>,
## #   Travel_time <dbl>, Mode_of_transport <chr>, Safe <chr>
```

Suppose you are helping a team of health scientists that are studying the eating and drinking habits of the group of students. **Show the height, age, sex, and sports habits of the students who eat meat, drink coffee, and like beer:**

```
filtered_q_data <- q_data %>%
  filter(EatMeat == 'Yes', DrinkCoffee == 'Yes', LikeBeer == 'Yes')

select(filtered_q_data, Height, Age, Sex, Sports)
```

```
## # A tibble: 6 x 4
##   Height   Age Sex    Sports
##    <dbl> <dbl> <chr>  <chr>
## 1    180  30.1 Male   Yes
## 2    185  21   Male   Yes
## 3    187  24.8 Male   Yes
```

```
## 4    165  28   Female No
## 5    158  24.2 Female Yes
## 6    177  22.2 Male   No
```

The company which manages student accommodation is interested in analysing feedback about the quality of its services. Suppose now that you are helping the accommodation company. **Show interesting summary statistics about the interviewed students, such as the average and minimum `Sleep` time, and the median and maximum `Rent`. Split the results by students who are happy/not happy with their accommodation and those who do/don't feel safe. Additionally, show the number of students in each category. Comment on these results:**

```
q_data_by_happiness <- q_data %>%
  group_by(Happy_accommodation, Safe)

summarise(q_data_by_happiness,
          avg_sleep = mean(Sleep), min_sleep = min(Sleep),
          med_rent = median(Rent), max_rent = max(Rent),
          n_students = n())
```

```
## # A tibble: 4 x 7
## # Groups:   Happy_accommodation [2]
##   Happy_accommodation Safe  avg_sleep min_sleep med_rent max_rent n_students
##   <chr>               <chr>     <dbl>     <dbl>    <dbl>    <dbl>      <int>
## 1 No                  No          9           9      600      600          1
## 2 No                  Yes         7           7      600      600          1
## 3 Yes                 No          8           8      100      100          1
## 4 Yes                 Yes         7.6         5      450     5000         15
```

Dealing with Anomalous Points: Two Alternatives using `dplyr`

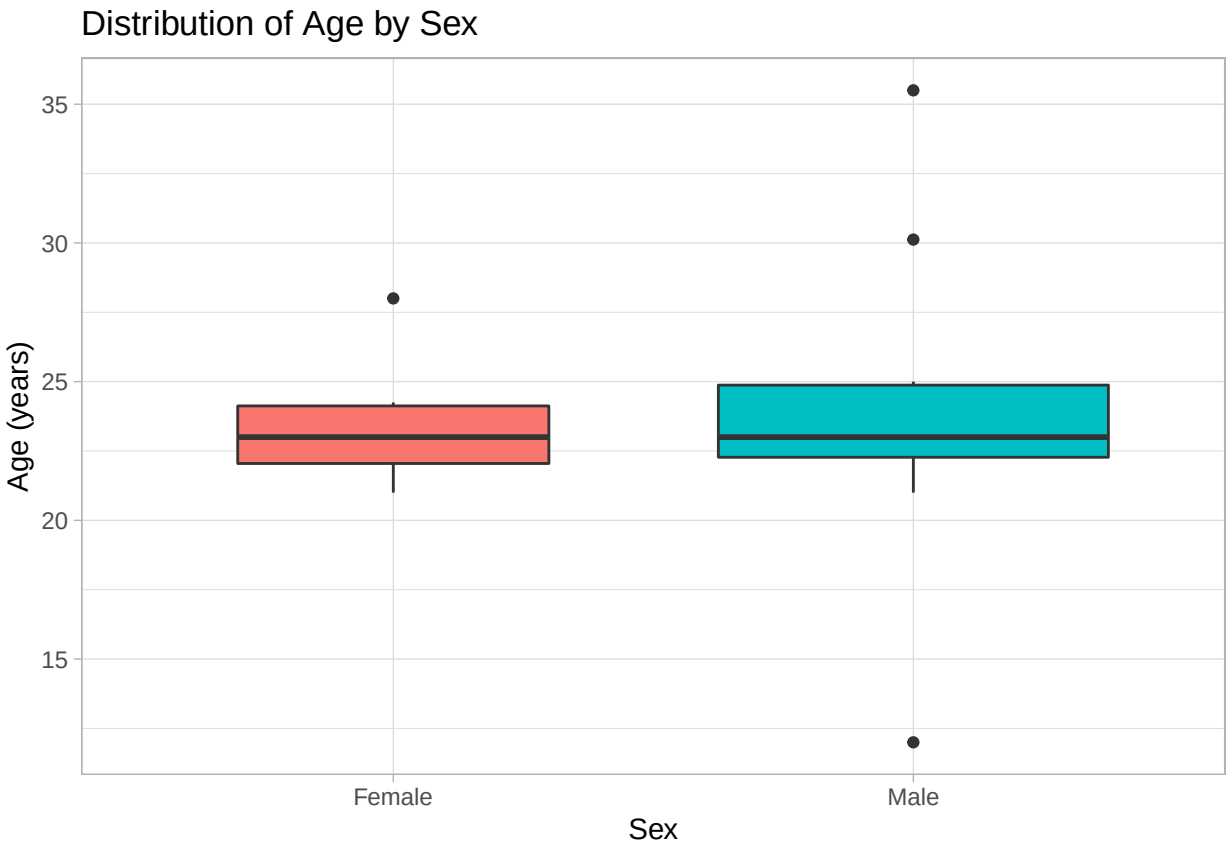
Consider the entire questionnaire dataset. How many females and how many males are there?

```
count(q_data, vars = Sex)
```

```
## # A tibble: 2 x 2
##   vars       n
##   <chr>  <int>
## 1 Female     7
## 2 Male      11
```

Produce boxplots of `Age`, stratified by `Sex`, using `geom_boxplot()`:

```
ggplot(q_data, aes(x = Sex, y = Age, fill = Sex)) +
  theme_light() + geom_boxplot(varwidth = 1) +
  labs(x = "Sex", y = "Age (years)",
       title = "Distribution of Age by Sex") +
  theme(legend.position = 'none')
```



There is clearly a problem with the data in that no student is under 15 years old. **Use `dplyr` to work out the minimum `Age`, to be called `min_age`, for each gender:**

```
min_age_by_sex <- q_data %>%
  group_by(Sex) %>%
  summarise(min_age = min(Age))

min_age_by_sex
```

```
## # A tibble: 2 x 2
##   Sex    min_age
##   <chr>    <dbl>
## 1 Female      21
## 2 Male        12
```

Extract the minimum `Age` for males, using the `filter()` method followed by `select()`:

```
min_age_male <- min_age_by_sex %>%
  filter(Sex == 'Male') %>%
  select(min_age) %>%
  as.numeric()

min_age_male
```

```
## [1] 12
```

First Way of Handling the Anomalous Point

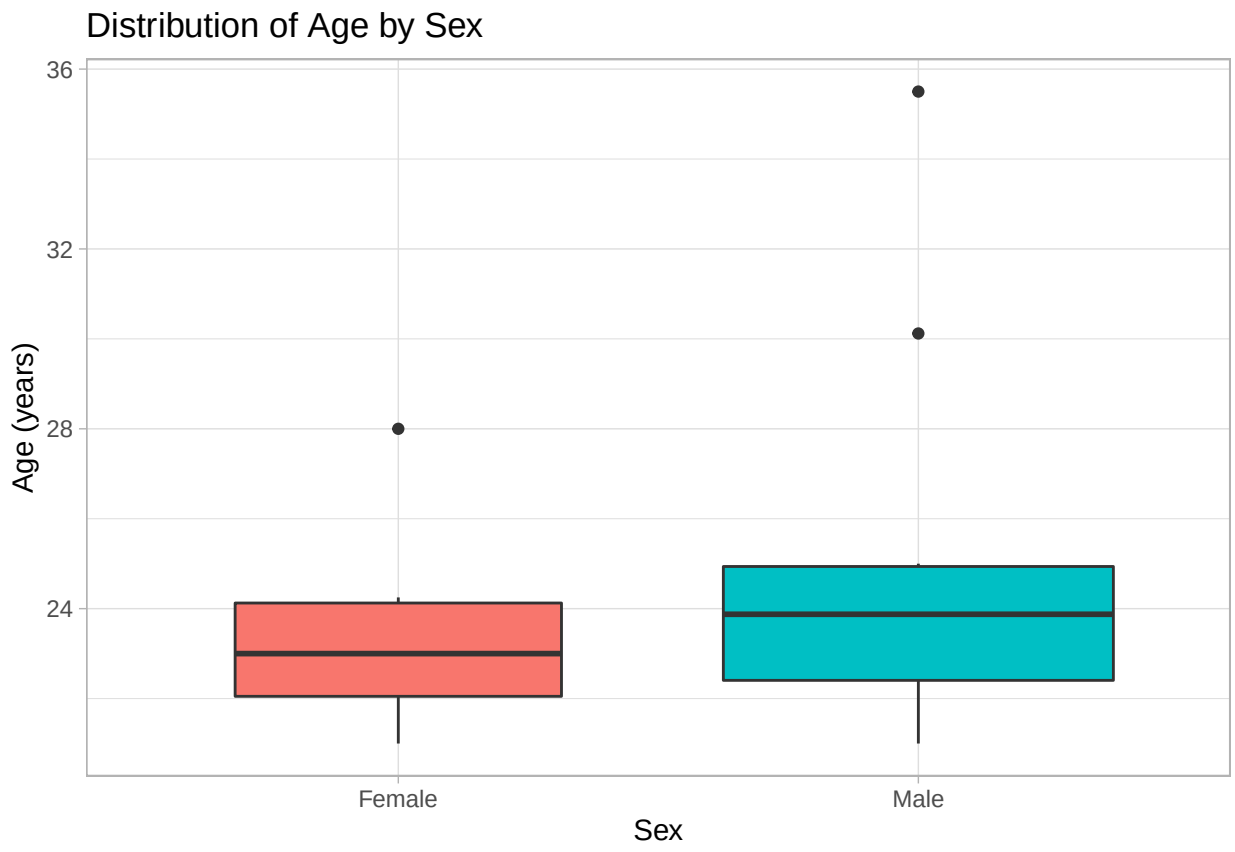
**Omit the data corresponding to the person with this anomalous age value by using `filter()` to include only rows where `Age >= 17` *(as a sensible lower bound)*, saving the result in `q_data_omitted`:

```
q_data_omitted <- q_data %>%
  filter(Age >= 17)

q_data_omitted$Age
```

```
##  [1] 23.00 22.40 30.12 21.00 22.10 25.00 22.42 24.75 23.00 24.00 22.00 28.00
## [13] 24.25 35.50 21.00 24.75 22.15
```

Now produce a boxplot of `Age`, stratified by `Sex`, using the data in `q_data_omitted`, from which the data corresponding to the person with the anomalous `Age` value has been omitted:

```
ggplot(q_data_omitted, aes(x = Sex, y = Age, fill = Sex)) +
  theme_light() + geom_boxplot(varwidth = 1) +
  labs(x = "Sex", y = "Age (years)",
       title = "Distribution of Age by Sex") +
  theme(legend.position = 'none')
```



Second Way of Handling the Anomalous Point

Calculate the median `Age` of all males who didn't provide an unrealistic age, producing a numerical result using `as.numeric()`:

```r
median_age_male <- q_data_omitted %>%
  group_by(Sex) %>%
  summarise(median_age = median(Age)) %>%
  filter(Sex == 'Male') %>%
  select(median_age) %>%
  as.numeric()

median_age_male
```

```
## [1] 23.875
```

Replace the unrealistic values for `Age` (`min_age_male`) with `median_age_male`:

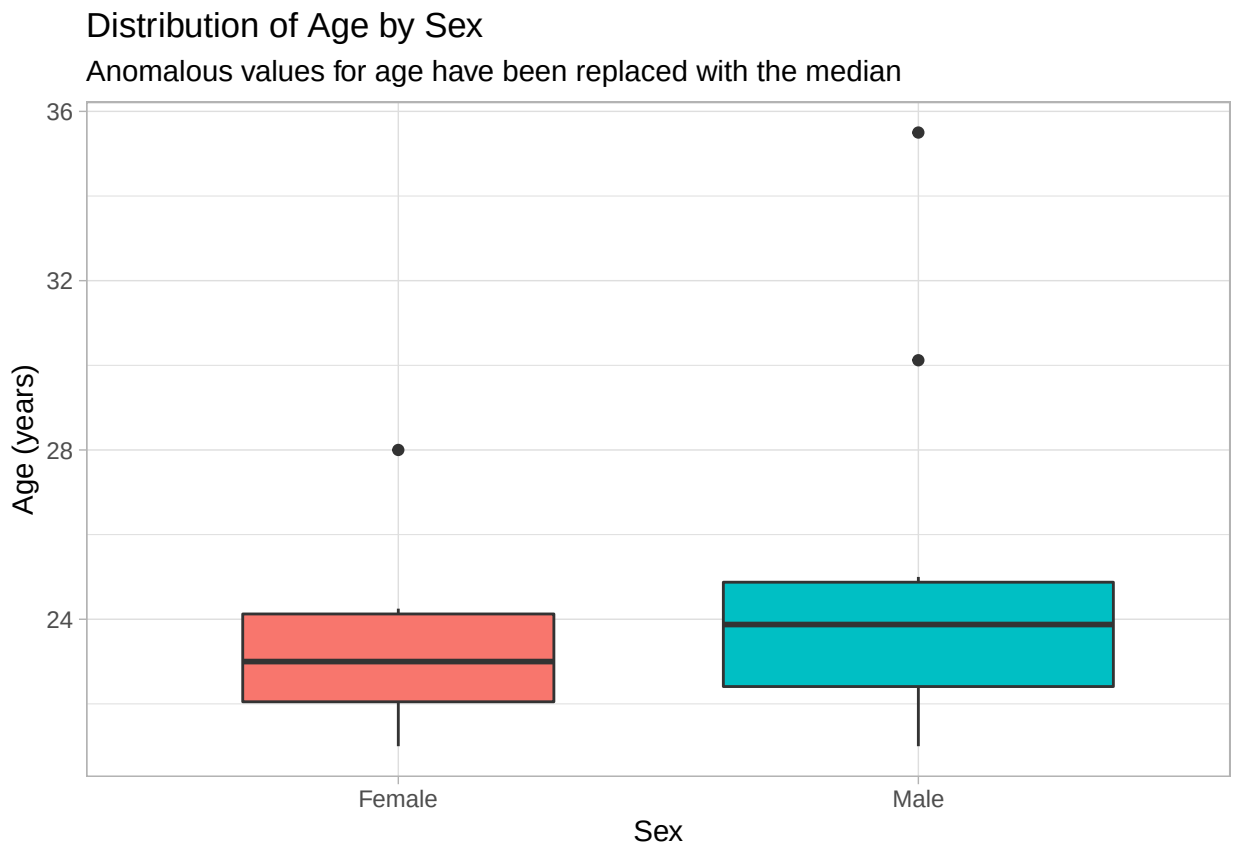
```r
q_data_corrected <- q_data %>%
  mutate(Age_corrected = ifelse(Age == min_age_male, median_age_male, Age))

# Check that the minimum age for males is no longer an unrealistic value:
q_data_corrected %>%
  group_by(Sex) %>%
  summarise(min_age = min(Age_corrected))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   Sex    min_age
##   <chr>    <dbl>
## 1 Female      21
## 2 Male        21
```

Produce a boxplot of `Age`, stratified by `Sex`, using the data in `q_data_corrected`:

```r
ggplot(q_data_corrected, aes(x = Sex, y = Age_corrected, fill = Sex)) +
  theme_light() + geom_boxplot(varwidth = T) +
  labs(x = "Sex", y = "Age (years)",
       title = "Distribution of Age by Sex",
       subtitle = "Anomalous values for age have been replaced with the median") +
  theme(legend.position = 'none')
```



The interquartile range is a measure of spread which is more robust to outliers than the standard deviation, and can be obtained in R using `IQR()`. **Compare the means, medians, standard deviations, and interquartile ranges of `Age` across `Sex`:**

```
q_data_corrected %>%
  group_by(Sex) %>%
  summarise(mean_age = mean(Age_corrected), median_age = median(Age_corrected),
            sd_age = sd(Age_corrected), iqr_age = IQR(Age_corrected))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 5
##   Sex    mean_age median_age sd_age iqr_age
##   <chr>     <dbl>      <dbl>  <dbl>   <dbl>
## 1 Female     23.5       23     2.30    2.07
## 2 Male       25.0       23.9   4.24    2.46
```