MATHS 7107 Data Taming Practical Solutions

Overview

To investigate the effect of fitness, gender, country of origin on red blood cell count (RBC), you have been given an Excel spreadsheet (`rbc.xlsx`) with observations[1] made on 500 subjects.

The variables are

- `gender`: male or female.
- `fitness`: a measure of fitness lying between 0 and 100. Zero corresponds to complete unfit, while 100 is the maximum possible fitness.
- `country`: there are three countries access denoted by 1, 2, or 3.
- `RBC`: red blood cell count in millions of red blood cells in one μL.

In this practical, all you have to do is clean the dataset `rbc.xlsx`. At any time, if you have questions about how to clean the data, you may discuss your decisions in the practical or via Piazza

You can check your cleaning by answering the following quiz questions.

Rule 1: Look at the data

I had a look and did the following:

- Moved the data so it starts at cell `A1`.
- Remove `useless figure`.
- Remove comment on line 360 of original Excel.
- Remove means at end of data.

I suggest you do this first too.

Rule 2: Is there a package to deal with that

So the best way to get Excel data in is the `readxl` package. Notice I saved my cleaned data as the originalname_clean. It is best practice to keep a copy of your original data.

```
rbc <- read_excel("rbc_clean.xlsx", na = "NA")
```

```
## New names:
## * `` -> `...1`
```

```
rbc
```

```
## # A tibble: 500 x 5
##    ...1  gender fitness country   RBC
##    <chr> <chr>    <dbl>   <dbl> <dbl>
##  1 132   male      76.2       1  5.27
##  2 309   male      92.3       2  6.42
##  3 234   M         65.2       1  3.61
##  4 400   F         89.4       3  6.15
```

[1] All data appearing in this work is fictitious. Any resemblance to real data, living or dead, is purely coincidental.

```
##  5 316   F         53.9       2  3.16
##  6 483   F         60.8       3  4.24
##  7 28    M         95.5       1  5.36
##  8 255   F         62.8       2  4.43
##  9 209   M         77.0       2  4.36
## 10 34    M         86.0       3  4.25
## # ... with 490 more rows
```

Rule 3: Break it into pieces

Not needed in this case as already in tabular form.

Rule 4: Get it into a tibble

Luckily for us, `readxl` does this automatically.

Rule 5: Look at each column

`...1`

The first column is not mentioned in the variables description, and so is removed.

```
rbc <-
  rbc %>%
  select(-...1)
rbc
```

```
## # A tibble: 500 x 4
##    gender fitness country   RBC
##    <chr>    <dbl>   <dbl> <dbl>
##  1 male      76.2       1  5.27
##  2 male      92.3       2  6.42
##  3 M         65.2       1  3.61
##  4 F         89.4       3  6.15
##  5 F         53.9       2  3.16
##  6 F         60.8       3  4.24
##  7 M         95.5       1  5.36
##  8 F         62.8       2  4.43
##  9 M         77.0       2  4.36
## 10 M         86.0       3  4.25
## # ... with 490 more rows
```

Gender

So there should be just `F` and `M`, a quick check gives the extra `male`, so we will convert to `M`

```
rbc %>%
  count(gender)
```

```
## # A tibble: 3 x 2
##   gender     n
##   <chr>  <int>
## 1 F        230
## 2 M        251
## 3 male      19
```

```
rbc <-
  rbc %>%
  mutate(
    gender = case_when(
      gender == "male" ~ "M",
      TRUE ~ gender
    )
  )
rbc %>% count(gender)
```

```
## # A tibble: 2 x 2
##   gender     n
##   <chr>  <int>
## 1 F        230
## 2 M        270
```

Fitness

So we are told that fitness should lie between 0 and 100, so lets check.

```
rbc %>%
  filter(!between(fitness, 0, 100))
```

```
## # A tibble: 1 x 4
##   gender fitness country   RBC
##   <chr>    <dbl>   <dbl> <dbl>
## 1 M          105       3  4.56
```

Spoke to the researcher - me - and told myself that this row can be removed, so removed.

```
rbc <-
  rbc %>%
  filter(between(fitness, 0, 100))
```

Country

```
rbc %>%
  count(country)
```

```
## # A tibble: 4 x 2
##   country     n
##     <dbl> <int>
## 1       1   183
## 2       2   164
## 3       3   141
## 4       4     1
```

So we are told in the information that country has only the levels 1, 2, and 3, so that row with country equal to 4 is wrong.

```
rbc %>%
  filter(
    country == 4
  )
```

```
## # A tibble: 1 x 4
##   gender fitness country   RBC
```

```
##   <chr>    <dbl>   <dbl> <dbl>
## 1 M         59.5       4  3.61
```

Spoke to the researchers, and it should be removed.

```
rbc <-
  rbc %>%
  filter(
    country != 4
  )
```

Also this column should be a factor so change.

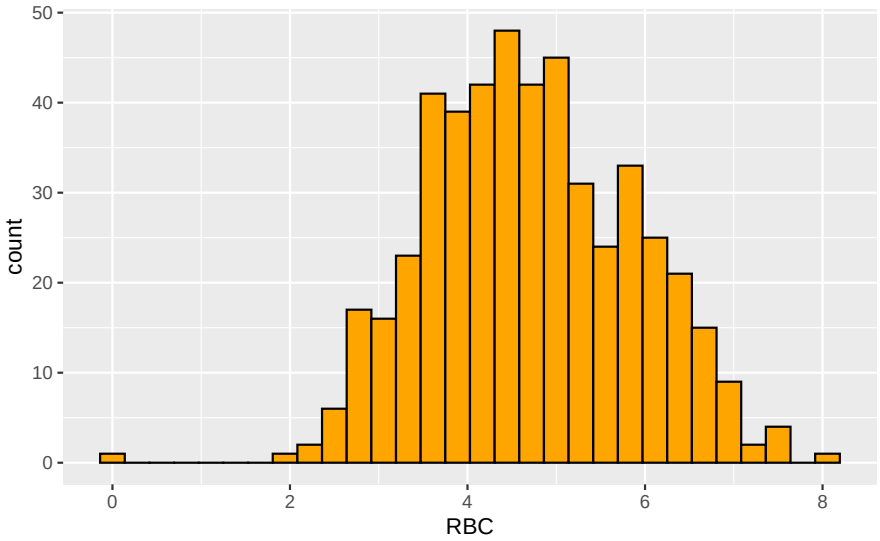
```
rbc <-
  rbc %>%
  mutate(
    country = factor(country)
  )
```

RBC

So for this one, we have not been given a range, so lets have a look to see if the data makes sense.

```
rbc %>%
  ggplot(aes(RBC)) +
  geom_histogram(col = "black", fill = "orange")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



So generally looks good except for that very low one. Let's check it

```
rbc %>%
  filter(RBC < 1)
```

```
## # A tibble: 1 x 4
##   gender fitness country   RBC
##   <chr>    <dbl> <fct>   <dbl>
## 1 M         72.0 1           0
```

A red blood cell count of zero is not good. Check with researcher who says remove.

```
rbc <-
  rbc %>%
  filter(
    RBC > 1
  )
```

Rule 6: You may need to go back

Discovered that missing was indicated as "NA", so have to change the `read_excel` command

Rule 7: create new columns?

No new columns needed.

Rule 8: Save the data and write it up

Save this data and make sure you date it. Keep this Rmd as information on the cleaning.

Questions:

1. How many observations are there in the final dataset?

```
nrow(rbc)
```

```
## [1] 487
```

There are 487 observations

2. How many males are there in the dataset?

```
sum(rbc$gender=="M")
```

```
## [1] 261
```

There are 261 males in the dataset.

3. How many observations are there in country 2?

```
sum(rbc$country == 2)
```

```
## [1] 164
```

There are 164 observations from country 2.

4. What is the mean `RBC` to 1 decimal place?

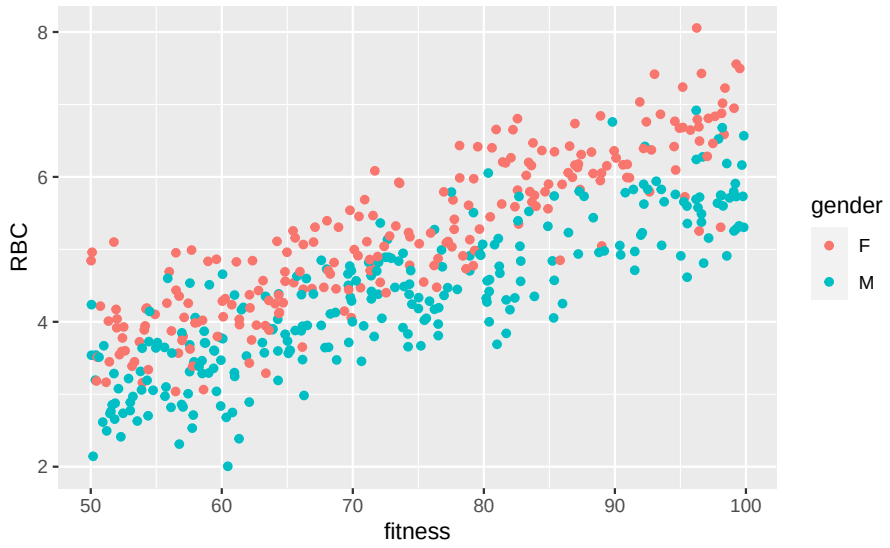
```
round(mean(rbc$RBC), 1)
```

```
## [1] 4.7
```

The mean RBC is 4.7

5. Produce a scatter point of `fitness` on `RBC` and colour the points differently for males and females. Describe the relationship between `RBC` and `colour`. Also, what do you notice about males and females?

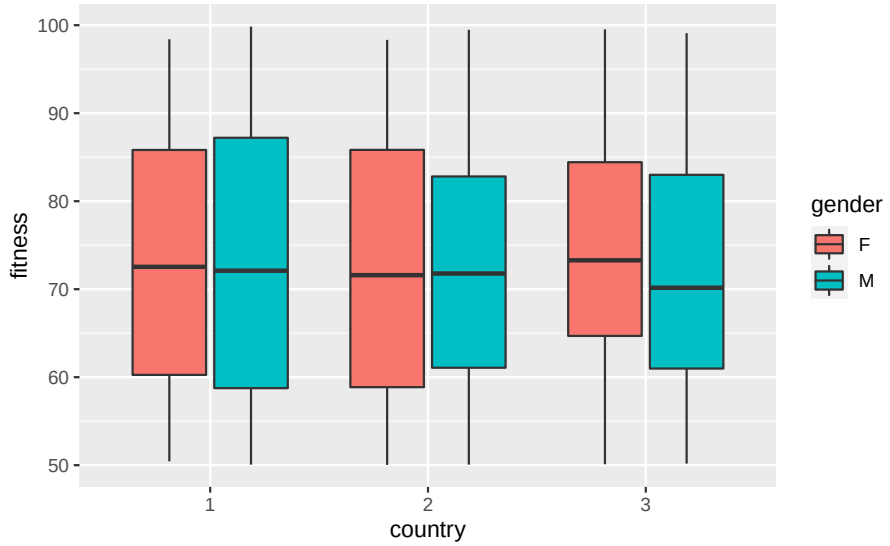
```
ggplot(rbc, aes(x = fitness, y = RBC, colour = gender)) +
  geom_point()
```



There is a moderate, positive, linear relationship between `fitness` and `RBC`. On average, it appears that females have a higher `RBC` than males.

6. Produce a boxplot of `fitness` and `country` with separate boxes for males and females. Describe the distribution.

```
ggplot(rbc, aes(x = country, y = fitness, fill = gender)) +
  geom_boxplot()
```




```
aggregate(fitness~country, data=rbc, median)
```

```
##   country  fitness
## 1       1 72.46631
## 2       2 71.76432
## 3       3 72.47801
```

```
aggregate(fitness~country, data=rbc, IQR)
```

```
##   country  fitness
## 1       1 26.89300
## 2       2 24.18242
```

```
## 3       3 20.35925
```

- Shape: (look at the histogram for each country to describe this).
- Location: The median for each country is extremely similar. Country 3 has the highest median, whilst country 2 has the lowest.
- Spread: Country 1 has the largest IQR, whilst country 3 has the lowest IQR.
- Outliers: There are no outliers in any country.