R Workshop Session 4 Exercise

Exercise 1:

Check out the following resources

a) R graph gallery (https://www.r-graph-gallery.com/). Which graph sparks the most joy for you?

b) Bioconductor course materials (https://bioconductor.org/help/course-materials/). Can you find anything of interest to you?

c) R-bloggers site (https://www.r-bloggers.com/). This site aggregates some of the best blogs about R. You can start by checking out the "most visited articles of the week" section on right hand side. You can also sign up and you will get a daily email with some of the best blogs. This covers a very wide range of topics so don't worry if 99% of them are irrelevant. We are looking for the needle in the haystack.

There is nothing for you to return back to us.

Exercise 2:

The data is a simulation of a longitudinal birth cohort. The two attached files contain the child length (i.e. height when the baby is laid down) and weight at two timepoints: delivery and month 3.

1) Download both files and eyeball scan for anything suspicious. Report anything that is odd.

SOLUTION:

Eyeballing the data in Excel shows:

- Length for Subject 0110-08107 in the delivery file has unusual number of decimals.
- Missing values are coded as NA in delivery file and as -999 in month 3 files.
- Weight is recorded in kg in delivery file and in g in month 3 file.
- Two subjects (0110-08010, 0110-08035) in month 3 are missing for both weight and length. Length is missing for subject 0110-18020.

and the subtler differences that are hard to spot:

- The values for length column in month 3 file is mostly recorded to nearest 0.5cm. Data with other decimals are potential recording errors that could be checked further.
- Subject "0110-18020" is mistyped as "0110- 18020" (extra white space) in the month 3 file. This is difficult to spot by eye but you will see it in the setdiff() command in step 4.

2) Set the working directory, load the packages and read in both files.

3) Make the necessary changes (e.g. changing column names and units) to merge the two files.

SOLUTION: Solving steps 2 - 3 in one.

```
# setwd("C:/Users/aramasamy/Desktop/R_workshop/")
pacman::p_load(tidyverse, readxl, gridExtra, reshape2, janitor)

m0 <- read_csv("anthropometry_delivery.csv") %>%
  rename(wt_m0 = weight_g, len_m0=length)

m3 <- read_excel("anthropometry_month3.xlsx", na="-999") %>%
  rename(wt_m3 = Weight_kg, len_m3=length) %>%
  mutate(wt_m3 = 1000*wt_m3)
```

4) Check for overlap in Subject IDs. Fix any errors in R (hint: `gsub()` might be useful) or Excel (in which case you need to re-read the data).

SOLUTION: There is an extra space in one of the IDs in month 3.

```
setdiff( m0$SubjectID, m3$SubjectID )
```

[1] "0110-10016" "0110-10029" "0110-18020"

```
setdiff( m3$SubjectID, m0$SubjectID ) # New subject? Unlikely in a birth cohort.
```

[1] "0110- 18020"

```
## One of the ID in month 3 has extra space. Fix it in Excel or using the following R code:
m3 <- m3 %>% mutate(SubjectID = gsub(" ", "", SubjectID))

## Let's check again after this correction
setdiff( m3$SubjectID, m0$SubjectID ) # All found
```

character(0)

```
setdiff( m0$SubjectID, m3$SubjectID ) # Two possible dropouts
```

[1] "0110-10016" "0110-10029"

5) Merge the two files. Calculate the gain in weight and length.

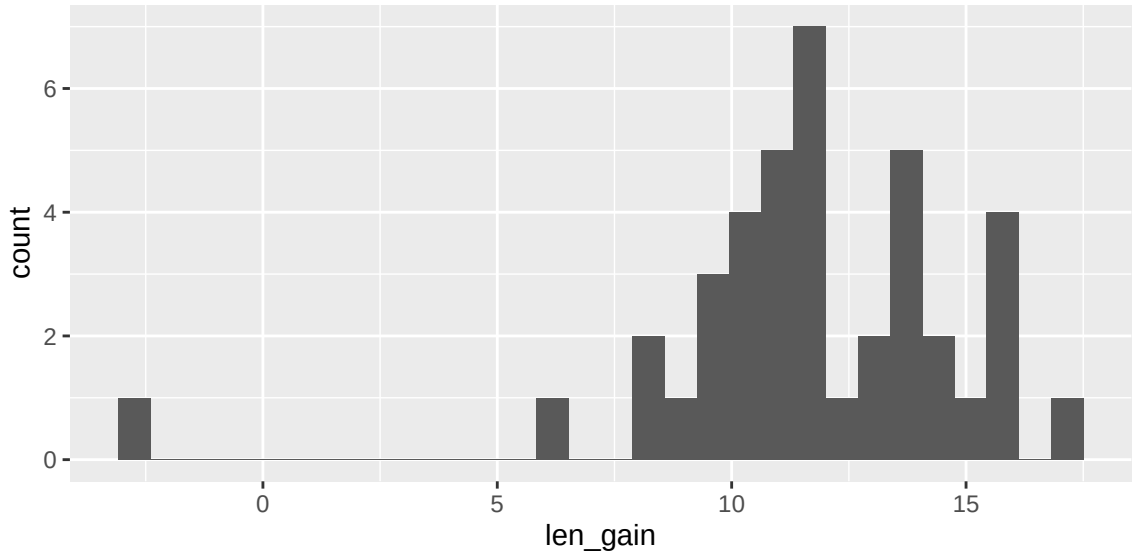
SOLUTION: I am using left join as all subjects should exist in the delivery file. You can also use outer join here.

```
comb <- left_join(m0, m3) %>%
  mutate(len_gain = len_m3 - len_m0,
         wt_gain  = wt_m3  - wt_m0)
```

6) Investigate the gain in length. Identify anything suspicious. You don't need to fix the suspicious data point here.

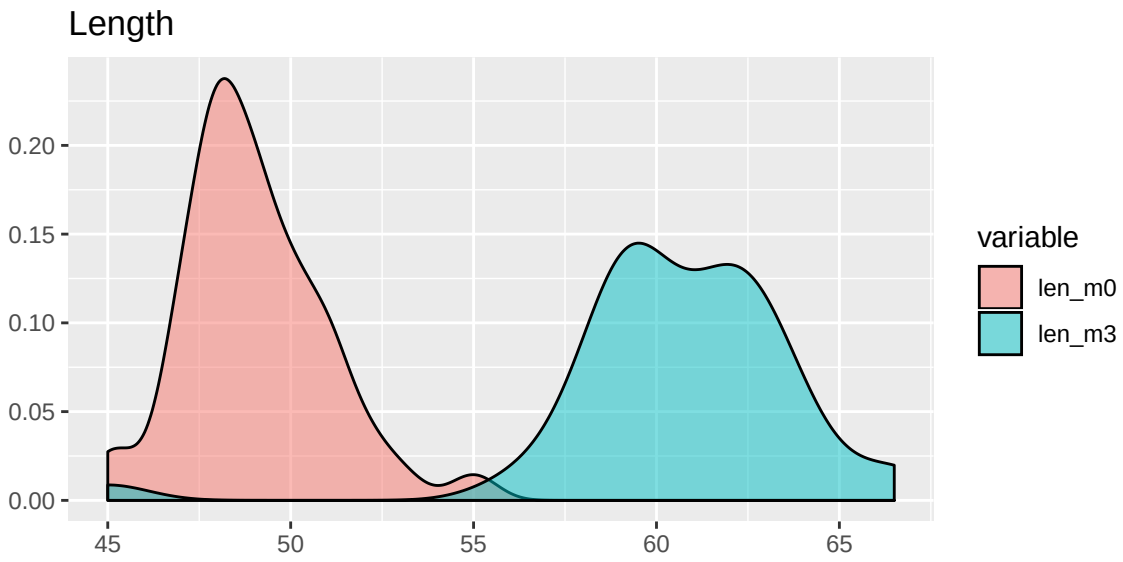
SOLUTION: Baby 0110-18007 appears to have shrunk which is impossible, so we need to check if the error is in the delivery or month 3 dataset. For this, we look at the length distribution cross-sectionally using density plots. So a length of 45cm is unlikely for a 3 month child. You can delete this data point in the Excel file and repeat if you like or check with the data provider to fix this problem if possible. For the purposes of this exercise, we will ignore it.

```
tmp <- comb %>%
  select(SubjectID, contains("len")) %>%
  arrange(len_gain)

ggplot(tmp, aes(x=len_gain)) + geom_histogram()
```




```
head(tmp)
## # A tibble: 6 x 4
##   SubjectID  len_m0 len_m3 len_gain
##   <chr>       <dbl>  <dbl>    <dbl>
## 1 0110-18007     48   45        -3
## 2 0110-08012     51   57.5       6.5
## 3 0110-08087     48   56         8
## 4 0110-08102     51   59.5       8.5
## 5 0110-10013     50   59         9
## 6 0110-08002     50   59.5       9.5
tail(tmp)
## # A tibble: 6 x 4
##   SubjectID  len_m0 len_m3 len_gain
##   <chr>       <dbl>  <dbl>    <dbl>
## 1 0110-08125     47   63.9      16.9
## 2 0110-08010     47   NA        NA
## 3 0110-08035     45   NA        NA
## 4 0110-10016     52   NA        NA
## 5 0110-10029     55   NA        NA
```

```
## 6 0110-18020     NA    59       NA

## Check the ranges at each timepoint
mdf <- comb %>% select(SubjectID, len_m0, len_m3) %>% melt()

ggplot(mdf, aes(x=value, fill=variable)) +
  geom_density(alpha=0.5) +
  labs(x=NULL, y=NULL, title="Length")
```

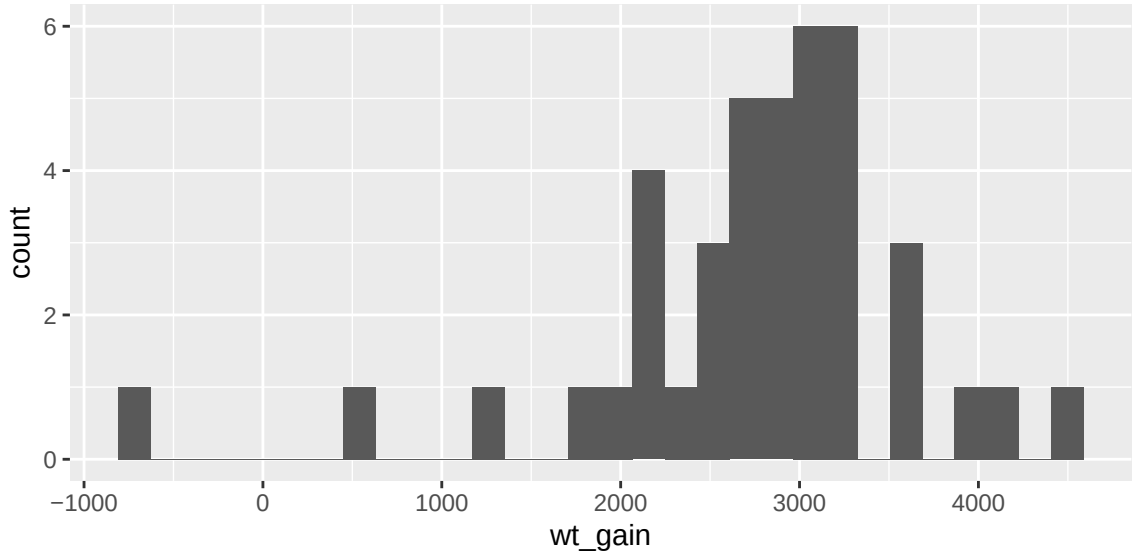



```
rm(tmp, mdf)
```

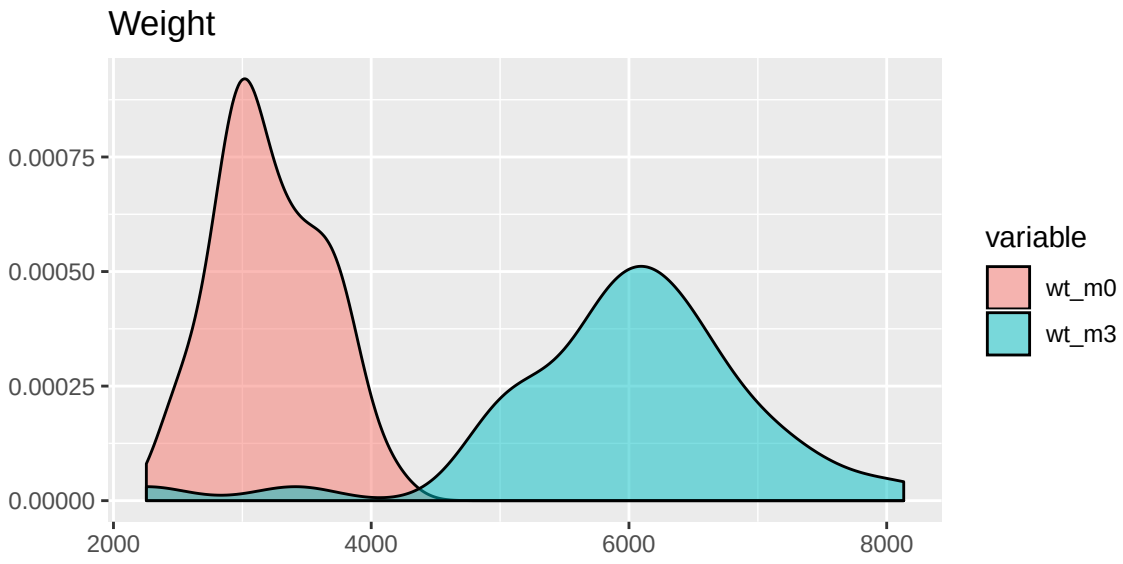
7) Investigate the gain in weight. Identify anything suspicious. You don't need to fix the suspicious data point here.

SOLUTION: Baby 0110-08088 appears to have lost weight which may possible due to illness. But checking the cross sectional data, it appears the month 3 value is very small. Checking with the data provider and hospital records might be useful. Again, for the purposes of this exercise, I will ignore it.

```
tmp <- comb %>%
  select(SubjectID, contains("wt")) %>%
  arrange(wt_gain)

ggplot(tmp, aes(x=wt_gain)) + geom_histogram()
```



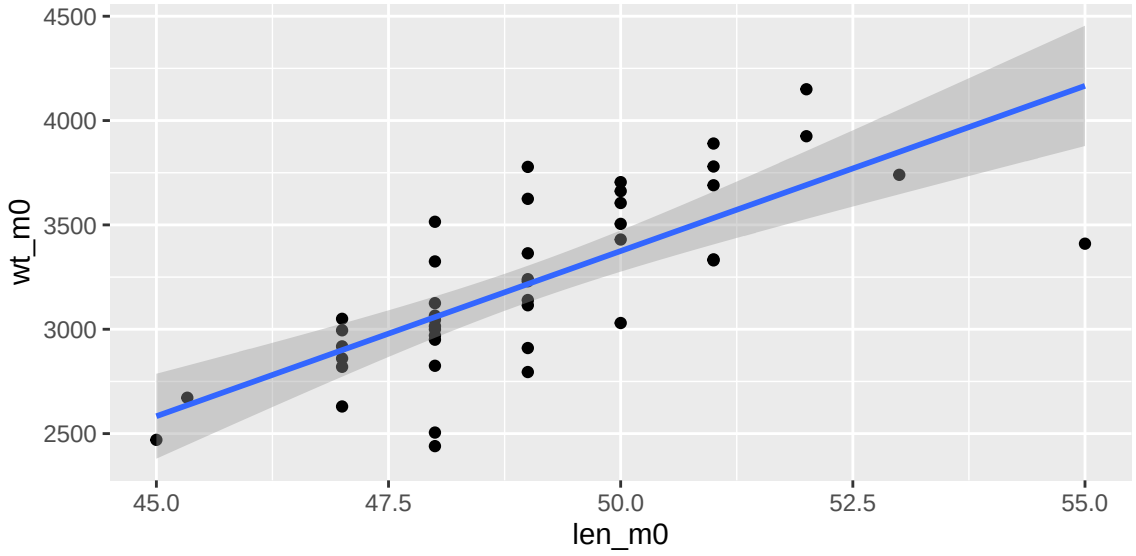

```
head(tmp)
## # A tibble: 6 x 4
##   SubjectID  wt_m0 wt_m3 wt_gain
##   <chr>      <dbl> <dbl>   <dbl>
## 1 0110-08088  3000  2255    -745
## 2 0110-08125  2918  3412     494
## 3 0110-08012  3690  4875    1185
## 4 0110-08076  3364  5250    1886
## 5 0110-18007  3065  5112    2047
## 6 0110-08102  3890  6060    2170
tail(tmp)
## # A tibble: 6 x 4
##   SubjectID  wt_m0 wt_m3 wt_gain
##   <chr>      <dbl> <dbl>   <dbl>
## 1 0110-08123  2970  7045    4075
## 2 0110-08026  3662  8132    4470
## 3 0110-08010  2630    NA      NA
## 4 0110-08035  2470    NA      NA
## 5 0110-10016  4150    NA      NA
## 6 0110-10029  3410    NA      NA

## Check the ranges at each timepoint
mdf <- comb %>% select(SubjectID, wt_m0, wt_m3) %>% melt()

ggplot(mdf, aes(x=value, fill=variable)) +
  geom_density(alpha=0.5) +
  labs(x=NULL, y=NULL, title="Weight")
```



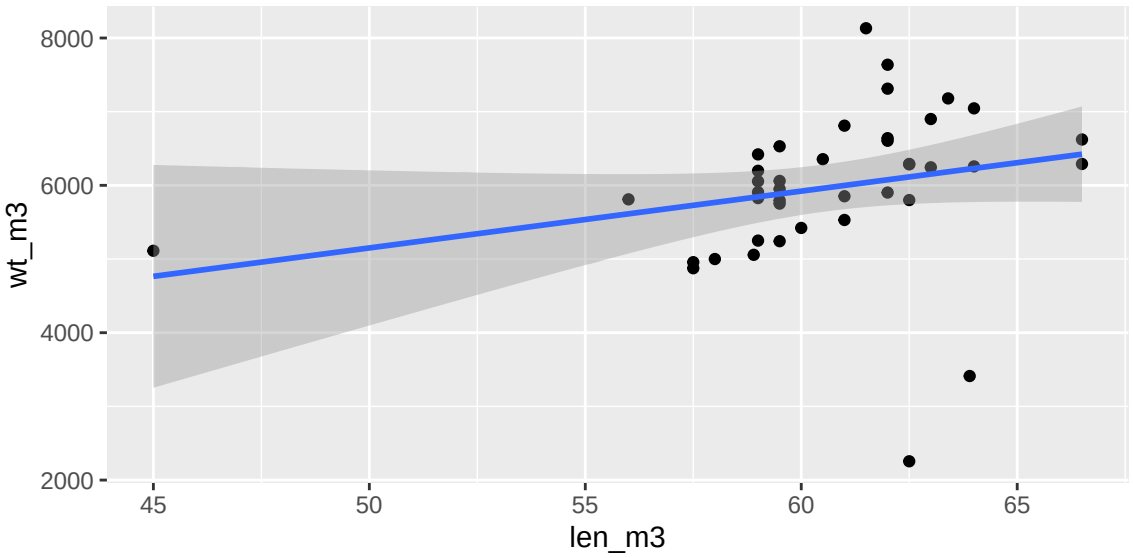

```
rm(tmp, mdf)
```

EXPANSION: You can also check the relationship between the length and weight at each time point to help clarify and identify further issues.

```
ggplot(comb, aes(x=len_m0, y=wt_m0)) + geom_point() + geom_smooth(method="lm")
```




```
ggplot(comb, aes(x=len_m3, y=wt_m3)) + geom_point() + geom_smooth(method="lm")
```



8) It is known in literature that babies who are born small for their gestational age tend to have a faster growth in the first few months. Can you able to test this with the available information?

SOLUTION: This is a trick question. You need to have the gestational age and cutoff definition to answer this. Be aware of the limitations in your dataset and avoid overinterpretations.