Course assignments

This document is a collection of the in-class assignments done during the Intro to Data Analysis with R course offered by the GRAPH Courses. To follow these assignments, please download the relevant folders from our github repo, https://github.com/the-graph-courses/rbp_cohort_0_materials

```
## Workshop exercise 1: Catching errors

# For each question, there is some code with an error
# Either the code doesn't run, or it does something incorrectly
# Your task is to find and fix the error
# The 1st person in the pair should screenshare and answer questions 1 & 2
# Then the 2nd person in the pair should screenshare and answer 3 & 4
# (If there are three people, you can do 2-1-1 or similar)
# When a person is screensharing, they are the "scribe"
# The person watching the scribe is the "reviewer".
# Reviewer(s) should offer comments/help if/when the scribe is stuck or writes problemat
ic code.

# GOOD LUCK!



#  Question 1 ----
# This code tries to calculate a person's BMI, but does so incorrectly.
# Try to correct it.
# First remove the quotes surrounding the code

'
jenny_height_cm <- 150
jenny_weight_kg <- 60
jenny_bmi <- jenny_weight_kg/jenny_height_cm
jenny_bmi
'
```

```
## [1] "\njenny_height_cm <- 150\njenny_weight_kg <- 60\njenny_bmi <- jenny_weight_kg/je
nny_height_cm\njenny_bmi\n"
```

```
#  Question 2 ----
# This code is supposed to create some variable tabulations with the `table()` function
# The dataset is a measles patient linelist from the {outbreaks} package
# Counts are by gender, complications and class
# First remove the quotes surrounding the code

'
if(!require(pacman)) install.packages("pacman")
pacman:p_load(outbreaks)
table()measles_hagelloch_1861$$gender
table()measles_hagelloch_1861$$complications
table()measles_hagelloch_1861$$class
'
```

```
## [1] "\nif(!require(pacman)) install.packages(\"pacman\")\npacman:p_load(outbreaks)\nt
able()measles_hagelloch_1861$$gender\ntable()measles_hagelloch_1861$$complications\ntabl
e()measles_hagelloch_1861$$class\n"
```

```
#  Question 3 ----
# This code is supposed to read in a dataset from the web
# It uses the `read_csv()` function from the {readr} package
# The file is at https://tinyurl.com/diabetes-china
# The code then tries to view the first 5 rows from the dataset
# It has errors that mean it does not run.

"
pacman::p_load(readr)
read_csv(https://tinyurl.com/diabetes-china)
head(rows = 5, x = diabetes_china)
"
```

```
## [1] "\npacman::p_load(readr)\nread_csv(https://tinyurl.com/diabetes-china)\nhead(rows
= 5, x = diabetes_china)\n"
```

```
#  Question 4 ----
# The commands below paste dataset columns together to make sentences like
# "on ____-__-__, xxx cases were reported"
# The first command works but the second command has issues. Try to remedy these.

'
paste("On",
      zika_yap_2007$onset_date,
      zika_yap_2007$value,
      "case(s) were reported in Yap")

paste("On",
      zika_sanandres_2015$cases,)
zika_sanandres_2015$date,
"case(s) were reported in San Andres")
'
```

```
## [1] "\npaste(\"On\",\n      zika_yap_2007$onset_date,\n      zika_yap_2007$value,\n
\"case(s) were reported in Yap\")\n\npaste(\"On\",\n      zika_sanandres_2015$cases,)\nz
ika_sanandres_2015$date,\n\"case(s) were reported in San Andres\")\n"
```

Workshop exercise 2

R for Busy People Workshop 1 Exercise: Data visualization critique

Critique each of the following data visualizations below. They are listed roughly in increasing order of badness!

In what ways are they not ideal? How might they be improved?

1. Zimbabwe New COVID cases

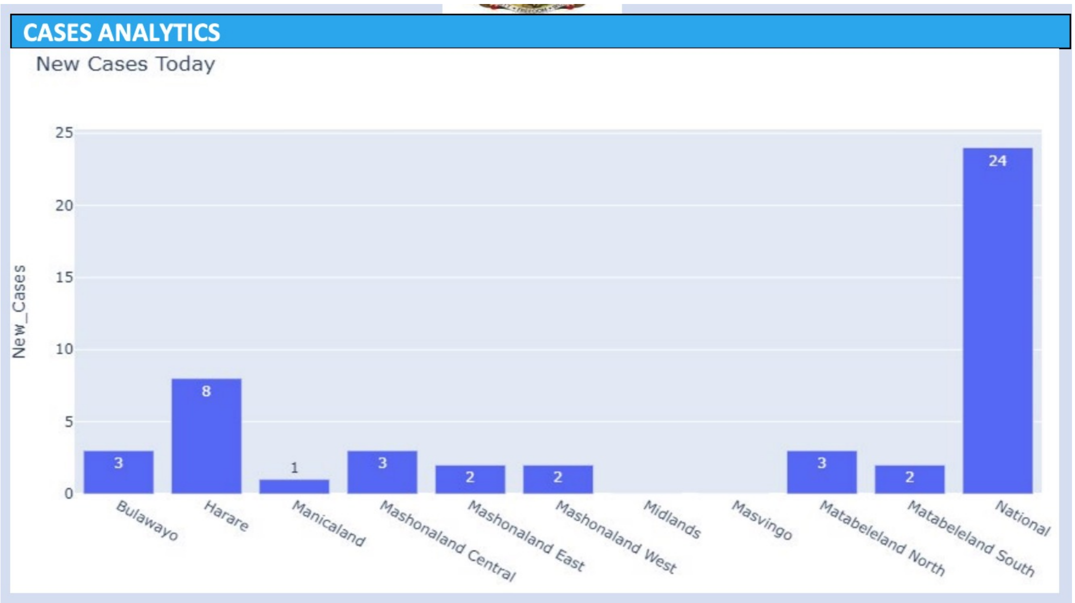


Image source

2. Mean and standard deviation from an experiment

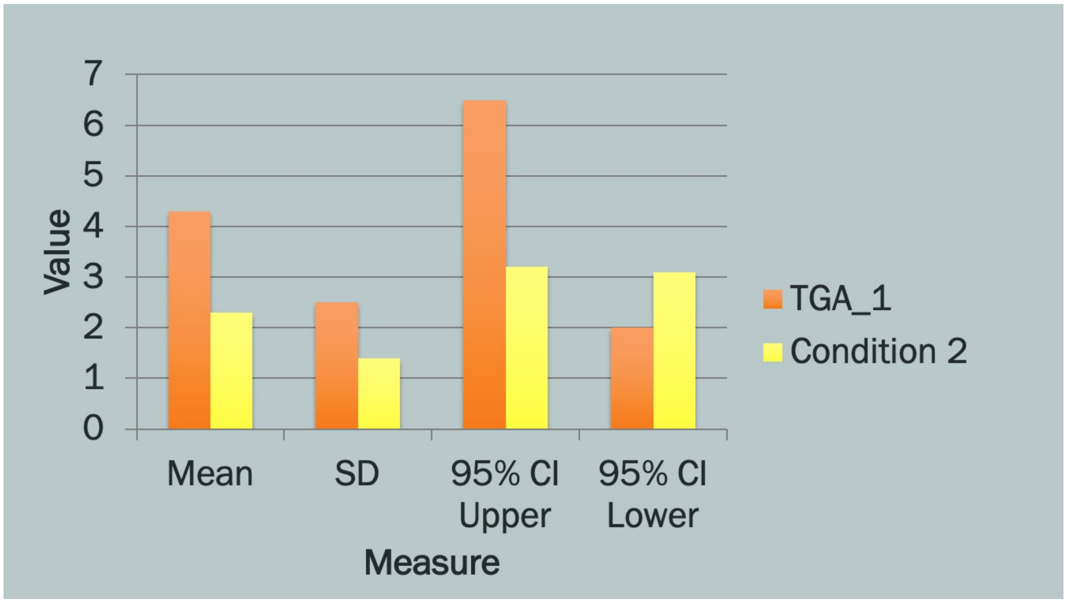
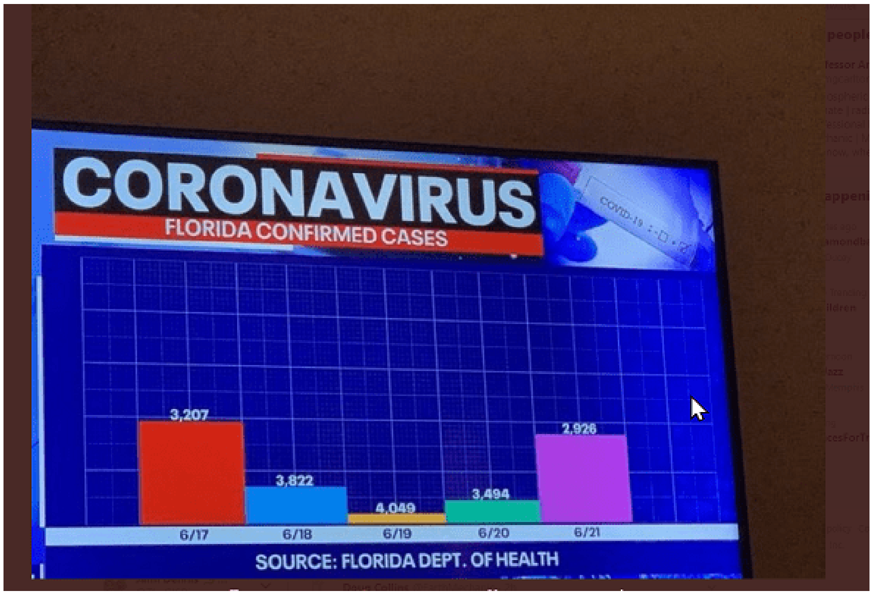


Image source

3. Confirmed COVID cases in Florida



4. Pet ownership percentages

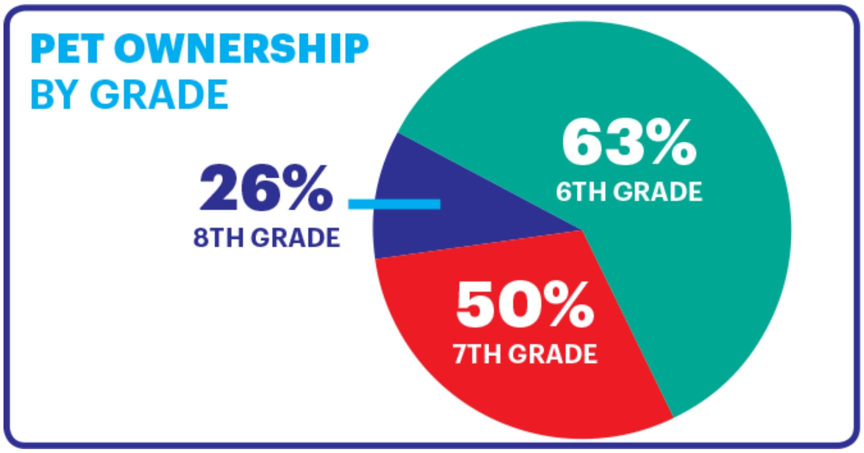


Image source

```r
# DESCRIPTIVE TITLE FOR SCRIPT
# FIRST_NAME LAST_NAME
# Date in YYYY-MM-DD format

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 0. Intro ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# This assignment tests that you can:
# - read in data with `read_csv()`
# - create frequency tables with `janitor::tabyl()` and save those with `write_csv()`
# - create simple plots with `esquisse::esquisser()` and save those plots
# - work well in a group

# Your final grade will be the average of all the work done in your pod.
# So if you finish ahead of time, try to help out your pod members!
# If a pod member is missing from class, you can ignore that script. No need to cover for them.
# You have about 1 hour to complete the task.
# But you can continue to work on it later on. The final deadline is next Tuesday, Oct 25, 23:59 pm CET.
# See section 5 below for details on submission

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 1. Load packages ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
if(!require(pacman)) install.packages("pacman")
pacman::p_load(tidyverse, janitor, here)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 2. Import data ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Here, use `read_csv()` to load in your dataset from the "data" folder in the "week_02" folder
# Which dataset? The dataset you need should have the same name as your script. Again, it is in the "data" subfolder of the "week_02" folder in your pod's RStudio project
# You can find more information about this dataset here:
# https://tinyurl.com/febrile-diseases-burkina-faso







##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 3. Create and export a frequency table ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Using the `tabyl()` function from the {janitor} package,
# make a frequency table of the `ims_final_full_classification` variable
# Then use `write_csv()` to save this table in your "outputs" folder with a descriptive name
```

```
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 4. Visualize the data to illustrate two key points, then export your plots with code
----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Use {esquisse} to generate two {ggplot2} figures that demonstrate each of
the POINTS listed below (If you know how to work with ggplot directly, you can skip esqu
isse)
# Then use the `ggsave()` function to save your plots in the "outputs" folder with descr
iptive names

# POINT A: The most common diagnostic classification was bacterial disease
# POINT B: A majority of children five years and older reported abdominal pain

# HINT: The techniques needed above were covered in the "Data dive" and "RStudio Projec
t" lessons.
# With one exception: for POINT B, where you may need to filter the dataset that you are
plotting
# Do this by clicking on the Data tab of your esquisse window (bottom right).
# You should see some sliders or variable selectors you can use to filter







##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 5. Export the week_02 folder ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: The final step is to export the `week_02` folder from Rstudio cloud.
# Here is an image explaining how to do this: https://imgur.com/a/kbLeIqV
# Then upload the zipped folder as a workshop assignment
# You should only do this AFTER all pod members present in class have finished their own
tasks.
# Your final grade will be the average of all the work done in your pod.
# The final deadline is next Tuesday, Oct 25, 23:59 pm CET.



##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 6. Present ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Someone from your group will be approached by an instructor and asked to
present their work.
# (You will have some time to prepare)
# The selected person will be expected to share their screen, and in about 2 minutes:
# - Say what the given dataset is about. (Show the dataset in your viewer, explain the k
```

```
ey variables)
# - Share one of their figures and explain what it is supposed to show
# - (Optional) Explain their answer to the BONUS question below



##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## BONUS (optional ungraded work): Describing data wrangling steps in words ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# For your assigned dataset, try to describe in words the data manipulation steps you wo
uld use to achieve the DATA REQUEST below.
# Of course, you don't yet know how to data wrangle in R, so how can you do this?
# You can describe how you would achieve the task:
# - Completely manually (with a printed-out spreadsheet, pen, paper and calculator)
# - With a spreadsheet software like Excel
# - With another data tool that you know (STATA, SPSS)
# - With some combination of the above

# If you have time, try to see if you can actually figure out the answer!

# During the presentation session, a GRAPH instructor will demo how to do this in R
# The goal here is simply to start to get you familiar with some data lingo and concepts

#  DATA REQUEST: Count the number of people who had up to four symptoms.
```

```r
# DESCRIPTIVE TITLE FOR SCRIPT
# FIRST_NAME LAST_NAME
# Date in YYYY-MM-DD format

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 0. Intro ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# This assignment tests that you can:
# - read in data with `read_csv()`
# - create frequency tables with `janitor::tabyl()` and save those with `write_csv()`
# - create simple plots with `esquisse::esquisser()` and save those plots
# - work well in a group

# Your final grade will be the average of all the work done in your pod.
# So if you finish ahead of time, try to help out your pod members!
# If a pod member is missing from class, you can ignore that script. No need to cover for them.
# You have about 1 hour to complete the task.
# But you can continue to work on it later on. The final deadline is next Tuesday, Oct 25, 23:59 pm CET.
# See section 5 below for details on submission

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 1. Load packages ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
if(!require(pacman)) install.packages("pacman")
pacman::p_load(tidyverse, janitor, here)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 2. Import data ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Here, use `read_csv()` to load in your dataset from the "data" folder in the "week_02" folder
# Which dataset? The dataset you need should have the same name as your script. Again, it is in the "data" subfolder of the "week_02" folder in your pod's RStudio project
# You can find more information about this dataset here:
# https://tinyurl.com/india-tb-pathways-and-costs






##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 3. Create and export a frequency table ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Using the `tabyl()` function from the {janitor} package,
# make a frequency table of the `first visit location` variable. (You will need backticks to refer to this variable)
# Then use `write_csv()` to save this table in your "outputs" folder with a descriptive name
```

```
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 4. Visualize the data to illustrate two key points, then export your plots with code ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Use {esquisse} to generate two {ggplot2} figures that demonstrate each of the POINTS listed below (If you know how to work with ggplot directly, you can skip esquisse)
# Then use the `ggsave()` function to save your plots in the "outputs" folder with descriptive names

# POINT A: The most common education category was "No Education"
# POINT B: About half of the male respondents do not drink alcohol

# HINT: The techniques needed above were covered in the "Data dive" and "RStudio Project" lessons.
# With one exception: for POINT B, where you may need to filter the dataset that you are plotting
# Do this by clicking on the Data tab of your esquisse window (bottom right).
# You should see some sliders or variable selectors you can use to filter








##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 5. Export the week_02 folder ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: The final step is to export the `week_02` folder from Rstudio cloud.
# Here is an image explaining how to do this: https://imgur.com/a/kbLeIqV
# Then upload the zipped folder as a workshop assignment
# You should only do this AFTER all pod members present in class have finished their own tasks.
# Your final grade will be the average of all the work done in your pod.
# The final deadline is next Tuesday, Oct 25, 23:59 pm CET.




##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 6. Present ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Someone from your group will be approached by an instructor and asked to present their work.
# (You will have some time to prepare)
# The selected person will be expected to share their screen, and in about 2 minutes:
```

```r
# - Say what the given dataset is about. (Show the dataset in your viewer, explain the key variables)
# - Share one of their figures and explain what it is supposed to show
# - (Optional) Explain their answer to the BONUS question below


##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## BONUS (optional ungraded work): Describing data wrangling steps in words ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# For your assigned dataset, try to describe in words the data manipulation steps you would use to achieve the DATA REQUEST below.
# Of course, you don't yet know how to data wrangle in R, so how can you do this?
# You can describe how you would achieve the task:
# - Completely manually (with a printed-out spreadsheet, pen, paper and calculator)
# - With a spreadsheet software like Excel
# - With another data tool that you know (STATA, SPSS)
# - With some combination of the above

# If you have time, try to see if you can actually figure out the answer!

# During the presentation session, a GRAPH instructor will demo how to do this in R
# The goal here is simply to start to get you familiar with some data lingo and concepts

#  DATA REQUEST: Find the average amount of money paid for TB visits by people educated up to Primary level versus people educated up to Secondary level
```

```
# DESCRIPTIVE TITLE FOR SCRIPT
# FIRST_NAME LAST_NAME
# Date in YYYY-MM-DD format

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 0. Intro ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# This assignment tests that you can:
# - read in data with `read_csv()`
# - create frequency tables with `janitor::tabyl()` and save those with `write_csv()`
# - create simple plots with `esquisse::esquisser()` and save those plots
# - work well in a group

# Your final grade will be the average of all the work done in your pod.
# So if you finish ahead of time, try to help out your pod members!
# If a pod member is missing from class, you can ignore that script. No need to cover for them.
# You have about 1 hour to complete the task.
# But you can continue to work on it later on. The final deadline is next Tuesday, Oct 25, 23:59 pm CET.
# See section 5 below for details on submission

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 1. Load packages ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
if(!require(pacman)) install.packages("pacman")
pacman::p_load(tidyverse, janitor, here)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 2. Import data ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Here, use `read_csv()` to load in your dataset from the "data" folder in the "week_02" folder
# Which dataset? The dataset you need should have the same name as your script. Again, it is in the "data" subfolder of the "week_02" folder in your pod's RStudio project
# You can find more information about this dataset here:
# https://tinyurl.com/motorcycle-accidents-colombia The dataset is in Spanish, but you can use Google and common sense to understand it.






##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 3. Create and export a frequency table ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Using the `tabyl()` function from the {janitor} package,
# make a frequency table of the `REC_MUNRES` variable, which stands for municipal residence
# Then use `write_csv()` to save this table in your "outputs" folder with a descriptive
```

```
name








##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 4. Visualize the data to illustrate two key points, then export your plots with code ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Use {esquisse} to generate two {ggplot2} figures that demonstrate each of the POINTS listed below (If you know how to work with ggplot directly, you can skip esquisse)
# Then use the `ggsave()` function to save your plots in the "outputs" folder with descriptive names

# POINT A: The most common age category (`REC_GRUPO EDAD`) was "20-24" (You will need backticks to refer to this variable)
# POINT B: A majority of female victims were passengers, not drivers (Female passengers have REC_SEXO equal to "Femenino" and passengers have `REC_CONDICION` equal to "Pasajero")

# HINT: The techniques needed above were covered in the "Data dive" and "RStudio Project" lessons.
# With one exception: for POINT B, where you may need to filter the dataset that you are plotting
# Do this by clicking on the Data tab of your esquisse window (bottom right).
# You should see some sliders or variable selectors you can use to filter








##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 5. Export the week_02 folder ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: The final step is to export the `week_02` folder from Rstudio cloud.
# Here is an image explaining how to do this: https://imgur.com/a/kbLeIqV
# Then upload the zipped folder as a workshop assignment
# You should only do this AFTER all pod members present in class have finished their own tasks.
# Your final grade will be the average of all the work done in your pod.
# The final deadline is next Tuesday, Oct 25, 23:59 pm CET.




##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 6. Present ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
# INSTRUCTION: Someone from your group will be approached by an instructor and asked to present their work.
# (You will have some time to prepare)
# The selected person will be expected to share their screen, and in about 2 minutes:
# - Say what the given dataset is about. (Show the dataset in your viewer, explain the key variables)
# - Share one of their figures and explain what it is supposed to show
# - (Optional) Explain their answer to the BONUS question below


##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## BONUS (optional ungraded work): Describing data wrangling steps in words ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# For your assigned dataset, try to describe in words the data manipulation steps you would use to achieve the DATA REQUEST below.
# Of course, you don't yet know how to data wrangle in R, so how can you do this?
# You can describe how you would achieve the task:
# - Completely manually (with a printed-out spreadsheet, pen, paper and calculator)
# - With a spreadsheet software like Excel
# - With another data tool that you know (STATA, SPSS)
# - With some combination of the above

# If you have time, try to see if you can actually figure out the answer!

# During the presentation session, a GRAPH instructor will demo how to do this in R
# The goal here is simply to start to get you familiar with some data lingo and concepts

# DATA REQUEST: Find out if the number of accidents in Medellin was increasing or decreasing between 2012 and 2015
```

```
# DESCRIPTIVE TITLE FOR SCRIPT
# FIRST_NAME LAST_NAME
# Date in YYYY-MM-DD format

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 0. Intro ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# This assignment tests that you can:
# - read in data with `read_csv()`
# - create frequency tables with `janitor::tabyl()` and save those with `write_csv()`
# - create simple plots with `esquisse::esquisser()` and save those plots
# - work well in a group

# Your final grade will be the average of all the work done in your pod.
# So if you finish ahead of time, try to help out your pod members!
# If a pod member is missing from class, you can ignore that script. No need to cover for them.
# You have about 1 hour to complete the task.
# But you can continue to work on it later on. The final deadline is next Tuesday, Oct 25, 23:59 pm CET.
# See section 5 below for details on submission

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 1. Load packages ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
if(!require(pacman)) install.packages("pacman")
pacman::p_load(tidyverse, janitor, here)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 2. Import data ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Here, use `read_csv()` to load in your dataset from the "data" folder in the "week_02" folder
# Which dataset? The dataset you need should have the same name as your script. Again, it is in the "data" subfolder of the "week_02" folder in your pod's RStudio project
# You can find more information about this dataset here:
# https://tinyurl.com/covid-19-united-states






##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 3. Create and export a frequency table ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Using the `tabyl()` function from the {janitor} package,
# make a frequency table of the `weekday_admit` variable, which stands for weekday of admission
# Then use `write_csv()` to save this table in your "outputs" folder with a descriptive name
```

```
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 4. Visualize the data to illustrate two key points, then export your plots with code ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Use {esquisse} to generate two {ggplot2} figures that demonstrate each of the POINTS listed below (If you know how to work with ggplot directly, you can skip esquisse)
# Then use the `ggsave()` function to save your plots in the "outputs" folder with descriptive names

# POINT A: The sample is mostly composed of older individuals, aged above 50
# POINT B: A majority of black patients had hypertension

# HINT: The techniques needed above were covered in the "Data dive" and "RStudio Project" lessons.
# With one exception: for POINT B, where you may need to filter the dataset that you are plotting
# Do this by clicking on the Data tab of your esquisse window (bottom right).
# You should see some sliders or variable selectors you can use to filter







##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 5. Export the week_02 folder ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: The final step is to export the `week_02` folder from Rstudio cloud.
# Here is an image explaining how to do this: https://imgur.com/a/kbLeIqV
# Then upload the zipped folder as a workshop assignment
# You should only do this AFTER all pod members present in class have finished their own tasks.
# Your final grade will be the average of all the work done in your pod.
# The final deadline is next Tuesday, Oct 25, 23:59 pm CET.



##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 6. Present ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Someone from your group will be approached by an instructor and asked to present their work.
# (You will have some time to prepare)
# The selected person will be expected to share their screen, and in about 2 minutes:
```

```
# - Say what the given dataset is about. (Show the dataset in your viewer, explain the key variables)
# - Share one of their figures and explain what it is supposed to show
# - (Optional) Explain their answer to the BONUS question below


##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## BONUS (optional ungraded work): Describing data wrangling steps in words ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# For your assigned dataset, try to describe in words the data manipulation steps you would use to achieve the DATA REQUEST below.
# Of course, you don't yet know how to data wrangle in R, so how can you do this?
# You can describe how you would achieve the task:
# - Completely manually (with a printed-out spreadsheet, pen, paper and calculator)
# - With a spreadsheet software like Excel
# - With another data tool that you know (STATA, SPSS)
# - With some combination of the above

# If you have time, try to see if you can actually figure out the answer!

# During the presentation session, a GRAPH instructor will demo how to do this in R
# The goal here is simply to start to get you familiar with some data lingo and concepts

# DATA REQUEST: Which racial groups had higher incidence of diabetes?
```

```r
# DESCRIPTIVE TITLE FOR SCRIPT
# FIRST_NAME LAST_NAME
# Date in YYYY-MM-DD format

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 0. Intro ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# This assignment tests that you can:
# - read in data with `read_csv()`
# - create frequency tables with `janitor::tabyl()` and save those with `write_csv()`
# - create simple plots with `esquisse::esquisser()` and save those plots
# - work well in a group

# Your final grade will be the average of all the work done in your pod.
# So if you finish ahead of time, try to help out your pod members!
# If a pod member is missing from class, you can ignore that script. No need to cover for them.
# You have about 1 hour to complete the task.
# But you can continue to work on it later on. The final deadline is next Tuesday, Oct 25, 23:59 pm CET.
# See section 5 below for details on submission

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 1. Load packages ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
if(!require(pacman)) install.packages("pacman")
pacman::p_load(tidyverse, janitor, here)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 2. Import data ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Here, use `read_csv()` to load in your dataset from the "data" folder in the "week_02" folder
# Which dataset? The dataset you need should have the same name as your script. Again, it is in the "data" subfolder of the "week_02" folder in your pod's RStudio project
# You can find more information about this dataset here:
# https://tinyurl.com/sex-attitudes-survey-uk (The variable dictionary is here: https://tinyurl.com/sex-attitudes-survey-uk-dict)






##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 3. Create and export a frequency table ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Using the `tabyl()` function from the {janitor} package,
# make a frequency table of the `rnssecgp_6` variable, which stands for the respondent's National Statistics Socio-economic classification code
# Then use `write_csv()` to save this table in your "outputs" folder with a descriptive
```

```
name





##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 4. Visualize the data to illustrate two key points, then export your plots with code ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Use {esquisse} to generate two {ggplot2} figures that demonstrate each of the POINTS listed below (If you know how to work with ggplot directly, you can skip esquisse)
# Then use the `ggsave()` function to save your plots in the "outputs" folder with descriptive names

# POINT A: The majority of respondents are in the 25-34 age group (`agrp` variable)
# POINT B: Among respondents aged 65-74, a large proportion consider religion to be "Fairly important" (`religimp` variable)

# HINT: The techniques needed above were covered in the "Data dive" and "RStudio Project" lessons.
# With one exception: for POINT B, where you may need to filter the dataset that you are plotting
# Do this by clicking on the Data tab of your esquisse window (bottom right).
# You should see some sliders or variable selectors you can use to filter






##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 5. Export the week_02 folder ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: The final step is to export the `week_02` folder from Rstudio cloud.
# Here is an image explaining how to do this: https://imgur.com/a/kbLeIqV
# Then upload the zipped folder as a workshop assignment
# You should only do this AFTER all pod members present in class have finished their own tasks.
# Your final grade will be the average of all the work done in your pod.
# The final deadline is next Tuesday, Oct 25, 23:59 pm CET.



##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 6. Present ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# INSTRUCTION: Someone from your group will be approached by an instructor and asked to present their work.
```

```
# (You will have some time to prepare)
# The selected person will be expected to share their screen, and in about 2 minutes:
# - Say what the given dataset is about. (Show the dataset in your viewer, explain the key variables)
# - Share one of their figures and explain what it is supposed to show
# - (Optional) Explain their answer to the BONUS question below



##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## BONUS (optional ungraded work): Describing data wrangling steps in words ----
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# For your assigned dataset, try to describe in words the data manipulation steps you would use to achieve the DATA REQUEST below.
# Of course, you don't yet know how to data wrangle in R, so how can you do this?
# You can describe how you would achieve the task:
# - Completely manually (with a printed-out spreadsheet, pen, paper and calculator)
# - With a spreadsheet software like Excel
# - With another data tool that you know (STATA, SPSS)
# - With some combination of the above

# If you have time, try to see if you can actually figure out the answer!

# During the presentation session, a GRAPH instructor will demo how to do this in R
# The goal here is simply to start to get you familiar with some data lingo and concepts

# DATA REQUEST: Among the "agree with opinion" questions (the variables starting with `sn`, from `snpres` to `snearly`) which have the largest differences between male and female respondents?
```

Workshop 3: Instructions and example report

Group_member_1 and group_member_2

2022-10-25

- Instructions
- Workshop 3 Assignment: Colombia Motorcycle Accidents data
  - Age distribution of fatalities per sex group
    - Areas of improvement
  - Distribution of roles, per sex group

Instructions

Welcome!

Today’s exercise will be done in groups of 2 to 4. You are allowed to pick your own partners. Once you have found a group, you can go into any of the small pair rooms to work.

Each member in the group should follow the instructions below:

1. First download the course repo here. (You will mostly work locally today). Unzip the downloaded folder, then click on the “.Rproj” file to open the project in RStudio. From the Files pane of RStudio, open the “week_03” folder.

2. In the “rmd” sub-folder, the instructions for your exercise are outlined (these are the same instructions you see here).

3. **Each group should pick one of the datasets** in the “data” folder. (You can read through the document titled “00_info_about_each_dataset” to get information about these datasets.)

4. Next, each group member should **select one categorical variable** from their chosen dataset. Their task will be to **create a short R-Markdown-based HTML report showing the frequency distribution of the chosen variable across two sexes.**

   For example, Jane and John pick the India TB dataset. Jane looks at the frequency distribution of the *education* variable for men and women. And John looks at the distribution of the *employment* variable for men and women.

5. You can do the initial work on their own, but the final document for submission will be a single HTML file containing a section for each chosen categorical variable.

   For example, Jane and John will submit an HTML document with two sections: the first section (primarily done by Jane) on the distribution of the education variable for both sexes, and the second section (primarily done by John) covering the distribution of the employment variable for both sexes.

6. **Each section of the report must contain these four things:**

   a. A plot created with {ggplot2}/{esquisse}

   b. A table created with {flextable} (See the flextable book for tips)

   c. At least one use of *inline R code* within the Rmd.

   d. At least one possible *area of improvement* mentioned.

7. As noted above, since RStudio cloud caused some problems in the last session (specifically with {esquisse}), it is recommended that you first work on your local RStudio, then when it is time to combine your work, you can:

   a. both copy your Rmd code into a document in your pod folder on RStudio Cloud, and do the final render from there;

   b. copy the material onto one of the group member's computers, and perform the final render from there.

8. **To submit your work** when you are done, you should share your document online, using the Rpubs service. This can be done as described in the video here. The link to this Rpubs page should be posted as a comment on our lesson page.

9. *Around 7:20pm UTC+2, a single person from the group will present the work done so far. Your work does not have to be finished by this time. You'll simply present what you succeeded at doing and what you struggled with.

10. The final due date is **Tuesday, Nov 1 at 23:59pm UTC+2**. You are encouraged to visit one of our study halls if you need assistance with this later.*

Finally, note that your work will be judged simply on whether you have met the four requirements mentioned above; it does not have to an amazing document. Just follow the instructions and you'll get full marks!

The rest of this document is an example of what one section of a report might look like.

Workshop 3 Assignment: Colombia Motorcycle Accidents data

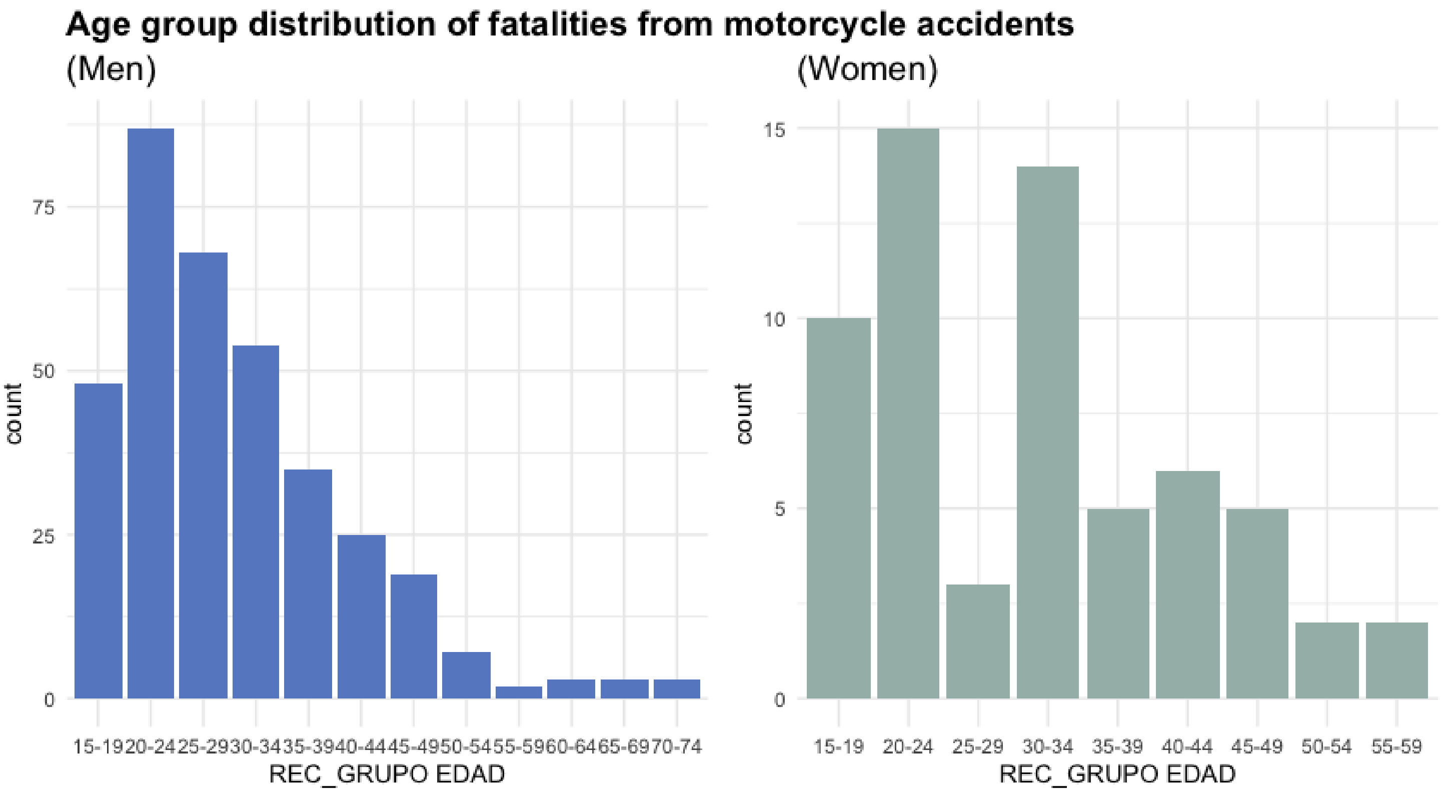
Age distribution of fatalities per sex group

Work primarily done by group member 1, Jane Doe

The dataset we chose provides information on deaths due to motorcycle accidents in Medellín, a Colombian city, as recorded in medical and police certificates.

I chose to look at the difference in the age distribution for male and female victims.

The plot below shows the age distributions for each sex:



Teacher commentary: We have not yet learned how to create dodged/faceted bar charts, which would be best for showing differences between genders. So for now, you should use the + and / syntax from the patchwork package to combine two plots, one for each sex. This is shown in the code chunk above. (Note that if you are reading this from the HTML document, you cannot see the code chunks. You need to look at the source Rmd file for this.)

And the table below shows similar information:

Age group distribution of
fatalities from motorcycle

accidents

REC_GRUPO EDAD	Femenino	Masculino
15-19	10	48
20-24	15	87
25-29	3	68
30-34	14	54
35-39	5	35
40-44	6	25
45-49	5	19
50-54	2	7
55-59	2	2
60-64	0	3
65-69	0	3
70-74	0	3

For both sexes, the age group with the most deaths was the 20 to 24 age group. In women, there were 15 deaths in this age group, and in men there were 87 deaths.

Teacher commentary: We have not yet learned how to extract specific slices of data from a data frame. So for now, you should use the syntax demonstrated above: within a pair of square brackets, place the row number, a comma, then the column number. (Note that if you are reading this from the HTML document, you cannot see the code chunks. You need to look at the source Rmd file for this.)

Areas of improvement

- The two bar charts have different numbers of bars. This is not ideal for comparisons.
- The age group labels are too horizontally compressed.
- I do not yet know how to change the variable names, so I left in the `REC_GRUPO EDAD` name in the table and plot.

Distribution of roles, per sex group

Work primarily done by group member 2, John Doe…

Teacher commentary: As already mentioned, your final document should contain all the sections from the different group members. Note that once the work has been done for one variable, it is easy to copy and paste the code to reproduce the analyses on another variable. This makes it easy to help out your group members who are struggling!

Honors Assignment: Extracting and Exporting Data Subsets

Author's name here

2022-11-01

- 1 Intro
- 2 Get the assignment repo
- 3 Load and clean the data
- 4 Create and export data subsets
  - 4.1 **Data Subset 1:** Extract demographic information
  - 4.2 **Data Subset 2:** Extract all male adults from the dataset
  - 4.3 **Data Subset 3:** Extract at-risk individuals
  - 4.4 **Data Subset 4:** Extract the knowledge-evaluation survey question variables
  - 4.5 **Data Subset 5:** Extract respondents with "ideal" knowledge, attitudes and practices (KAPs) towards rabies
- 5 Submission: Upload zipped folder

1 Intro

Welcome!

For this Honor's assignment, you will be carrying out a new task as a data analyst: preparing data subsets for **someone else to use**. It is a hands-on approach to using the `select()` and `filter()` verbs.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for the assignment is Tuesday, November 28th at 23:59 PM UTC+2.

2 Get the assignment repo

To get started, you should download and look through the assignment folder.

1. First download the repo to your local computer here.

You should ideally work on your local computer, but if you would rather work on RStudio Cloud, you can upload the zip file to RStudio Cloud through the Files pane. Consult one of the instructors for guidance on this.

2. Unzip/Extract the downloaded folder.

   If you are on macOS, you can simply double-click on a file to unzip it.

   If you are on Windows and are not sure how to "unzip" a file, see this image. You need to right-click on the file and then select "extract all".

3. Once done, click on the RStudio Project file in the unzipped folder to open the project in RStudio.

4. In RStudio, navigate to the Files tab and open the "rmd" folder. The instructions for your exercise are outlined there (these are the same instructions you see here).

5. Open the "data" folder and observe its components. You will work with the "rabies_dataset.csv" file. (You can also open the "00_info_about_the_dataset" file to learn more about this dataset.)

6. In the same folder, open the metadata (data dictionary) file, "rabies_metadata.pdf". This data dictionary is necessary because the data for today is numerically encoded; that is, all variables are stored as numbers which correspond to categories. Without the variable dictionary, you wouldn't know what the numeric categories correspond to!

3 Load and clean the data

Now that you understand the structure of the repo, you can load in and clean your dataset.

In the code section below, **load in the needed packages** (hint: load {tidyverse} and {here}).

Pro tip: Use `p_load()` to load in your packages, since it both loads and installs packages as needed.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Now, **read the dataset into R**. The data frame you import should have 1466 rows and 23 columns. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Next, perform the following two cleaning tasks on the imported dataset, then store the cleaned dataset in a new object.

- **Bring the respondent ID to the front**. The respondent ID is the 23rd column of your dataset. It should be more visible so that someone who opens up the CSV knows immediately that each row corresponds to a respondent. Move it to the first position in the data frame.

- **Remove the `Education` variable, which has not been properly encoded.** (To notice the encoding issues, look at both the data frame and the metadata file). We are considering this variable unusable, and hence removing it.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4 Create and export data subsets

In each "Data subset" section below, you should

- decide whether to use the `filter()` or `select()` function, then apply that function to create the required extract of the dataset;
- export the data subset into an appropriately-named CSV file in the "data_exports" folder.
- NOTE: For all data subsets, always keep the respondent ID variable! Without it, you cannot link back your data to the original dataset and you lose crucial information.

4.1 Data Subset 1: Extract demographic information

Create and export a data subset of respondents' demographic information—their age, gender and geographic background.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Don't forget that you were asked to include the respondent ID variable in all subsets.

Remember to use the `here()` function to allow your Rmd to use project-relartive paths.

4.2 Data Subset 2: Extract all male adults from the dataset

Create and export a data subset containing only records for males aged over 18.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

(Hint: this is a row-filtering question, so there is no need to select or drop columns.)

4.3 Data Subset 3: Extract at-risk individuals

Create and export a subset with “at-risk” individuals. These are people who a) have a pet at home, b) have no access to health facilities, and c) consider that the rabies vaccine is not affordable for them.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4.4 Data Subset 4: Extract the knowledge-evaluation survey question variables

Reading the metadata, you will see that some variables correspond to “knowledge-evaluation” questions about rabies. Create and export a subset that includes these variables.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4.5 Data Subset 5: Extract respondents with “ideal” knowledge, attitudes and practices (KAPs) towards rabies

People with *ideal* KAPs are defined as people who answered that they:

- vaccinate their pets,
- know the clinical signs for rabies, and
- visit a doctor after being bitten by an animal

Create and export a data subset that includes just these individuals.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

5 Submission: Upload zipped folder

Once you have finished the tasks above, you should **create a zipped file** of your project folder, containing all the work you have done, and **upload it** on the assignment page.

Workshop 5: Transforming, selecting, filtering, and plotting

Author’s name here

2022-11-08

- 1 Intro
- 2 Get the assignment repo
- 3 Load and clean the data
  - 3.1 **Step 1:** Verify the type of your variables
  - 3.2 **Step 2:** Convert the physical activity variables
- 4 **Plot 1:** BMI distribution by sex
  - 4.1 **Extract:** Make a subset
  - 4.2 **Plot with Esquisse:** Violin Plot
- 5 **Plot 2:** Male respondents’ Light Physical Activity (LPA, in minutes per week)
  - 5.1 **Extract:** Make a data subset
  - 5.2 **Plot with Esquisse:** Histogram of Light Physical Activity (LPA) of Male Respondents
- 6 **Plot 3:** Adults complying to OMS/WHO recommendations’ Moderate to Vigorous Physical Activity (MVPA, minutes per week)
  - 6.1 **Extract:** Make a data subset
  - 6.2 **Plot with Esquisse:** Boxplots of Moderate to Vigorous Physical Activity per Age Group
- 7 Submission: Upload HTML

1 Intro

Welcome!

For this workshop, we will be cleaning a dataset. It is a hands-on approach to using the `select()`, `filter()`, and `mutate()`.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for the assignment is Tuesday, November 15th at 23:59 PM UTC+2.

2 Get the assignment repo

To get started, you should download and look through the assignment folder.

1. First download the repo to your local computer here.

   You should ideally work on your local computer, but if you would rather work on RStudio Cloud, you can upload the zip file to RStudio Cloud through the Files pane. Consult one of the instructors for guidance on this.

2. Unzip/Extract the downloaded folder.

   If you are on macOS, you can simply double-click on a file to unzip it.

   If you are on Windows and are not sure how to "unzip" a file, see this image. You need to right-click on the file and then select "extract all".

3. Once done, click on the RStudio Project file in the unzipped folder to open the project in RStudio.

4. In RStudio, navigate to the Files tab and open the "rmd" folder. The instructions for your exercise are outlined there (these are the same instructions you see here).

5. Open the "data" folder and observe its components. You will work with the "obesity.csv" file. (You can also open the "00_info_about_the_dataset" file to learn more about this dataset.)

3 Load and clean the data

Now that you understand the structure of the repo, you can load in and clean your dataset.

In the code section below, **load in the needed packages**.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Now, **read the dataset into R**. The data frame you import should have 142 rows and 9 columns. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

3.1 Step 1: Verify the type of your variables

Before jumping into wrangling or plotting, you should take the time to know what data types you are working with. You can look at these types with `summary()` or `typeof()`.

Some of your variables are numeric but they should be factors (i.e. categories), some are characters but should be factors, and some are characters but should be numeric. Having them in the correct type will be essential for the next manipulations and for plotting !

Use `mutate()` to convert your variables into the right type.

3.2 Step 2: Convert the physical activity variables

Currently, the variables of physical activity are in seconds per day. There are 3 types of physical activity variables: sedentary physical activity (`sedentary_ap_s_day`), light physical activity (`light_ap_s_day`), and moderate to vigorous physical activity (`mvpa_s_day`).

Please convert these numerical variables in **seconds/day** to **minutes/week**. As a kind reminder, 60 seconds = 1 minute and 7 days = 1 week.

(Hint: use `mutate()` to create new variables that are in minutes per week. If you feel more comfortable changing the variabless in-place, that's also acceptable.)

Why do we perform this conversion? The WHO (known as OMS in French) recommendations are in minutes per week, so we want to align with these measures.

You will use this `obesity_data_cleaned` for all the following subset making and plotting.

4 Plot 1: BMI distribution by sex

4.1 Extract: Make a subset

1. Make a subset with only the variables of interest for your plot. **This is good practice to make a subset with the variables you need for plotting.**
2. Print this subset in an elegant manner for your HTML (hint: use `reactable`).

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4.2 Plot with Esquisse: Violin Plot

Using esquisse and the subset you just made above, plot BMI distributions by sex, as a violin plot.

Violin plots are interesting because you can compare the density curves' peaks, valleys, and tails to see where the groups are similar or different.

```
## [1] "PASTE_THE_ESQUISSE_CODE_HERE"
```

5 Plot 2: Male respondents’ Light Physical Activity (LPA, in minutes per week)

5.1 Extract: Make a data subset

1. To make this subset, you will only male respondents.
2. Then keep only the variables useful for the plot.
3. Print this subset in an elegant manner for your HTML (hint: use `reactable`).

```
## [1] "WRITE_YOUR_CODE_HERE"
```

5.2 Plot with Esquisse: Histogram of Light Physical Activity (LPA) of Male Respondents

Using esquisse and the subset you just made above, plot LPA distribution (minutes/week) for male respondents, as a histogram.

```
## [1] "PASTE_THE_ESQUISSE_CODE_HERE"
```

6 Plot 3: Adults complying to OMS/WHO recommendations’ Moderate to Vigorous Physical Activity (MVPA, minutes per week)

6.1 Extract: Make a data subset

1. To make this subset, you will only keep individuals in the dataset who have complied to OMS/WHO recommendations

(Hint 1: `oms_recommendation` should be equal to `Yes`. Side-note: OMS is Organisation Mondiale de la Santé, French for WHO.)

(Hint 2: The variable `status` is encoded in French as well. “Adulte” means “Adult” and “Enfant” means “Child”.)

2. Then keep only the variables useful for the plot.

3. Print this subset in an elegant manner for your HTML (hint: use `reactable`).

```
## [1] "WRITE_YOUR_CODE_HERE"
```

6.2 Plot with Esquisse: Boxplots of Moderate to Vigorous Physical Activity per Age Group

Using esquisse and the subset you just made above, plot MVPA distributions (minutes/week) by age groups, as boxplots.

```
## [1] "PASTE_THE_ESQUISSE_CODE_HERE"
```

7 Submission: Upload HTML

Once you have finished the tasks above, you should **knit this Rmd into an HTML** and **upload it** on the assignment page.

Workshop 9: Lines, Scales, Labels, and Themes

Author’s name here

2022-12-06

- 1 Intro
- 2 Instructions
- 3 Prepare the data
  - 3.1 **Step 1:** Pivot from wide to long
  - 3.2 **Step 2:** Rename variables
- 4 Recreate desired plot
- 5 Submission: Upload Rmd HTML

1 Intro

Welcome!

For this workshop, we have two activities. You will be completing Activity 1 in this Rmd.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for Activity 1 is Tuesday, December 6th at 23:59 PM UTC+2.

2 Instructions

Your assignment for Activity 1 is to recreate a plot from the EpiGraphHub COVID-19 Switzerland dashboard in R using {ggplot2}.

1. First choose one of the four line graphs shown on the dashboard. This is the one you will recreate.

2. Download the data associated with your chosen plot by clicking the three dots on the top right on the plot and selecting “Export CSV”

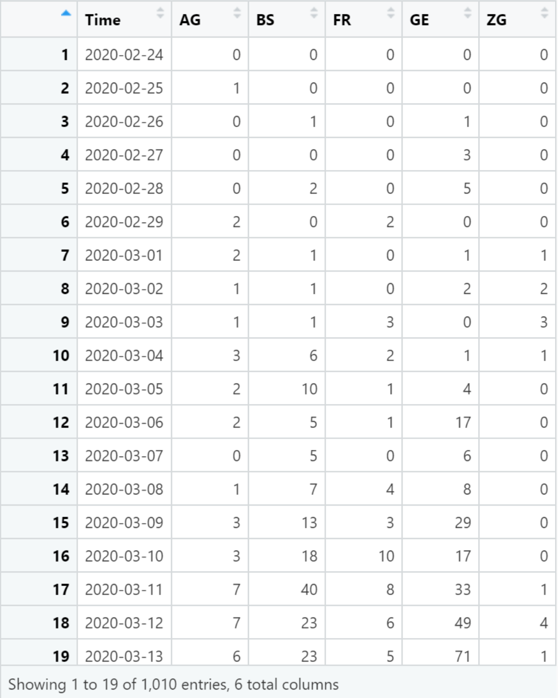
3. Once downloaded, move the CSV to the “data” folder of this R project.

3 Prepare the data

Now, **read the dataset into R**. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

This data needs some cleaning before we can plot it.

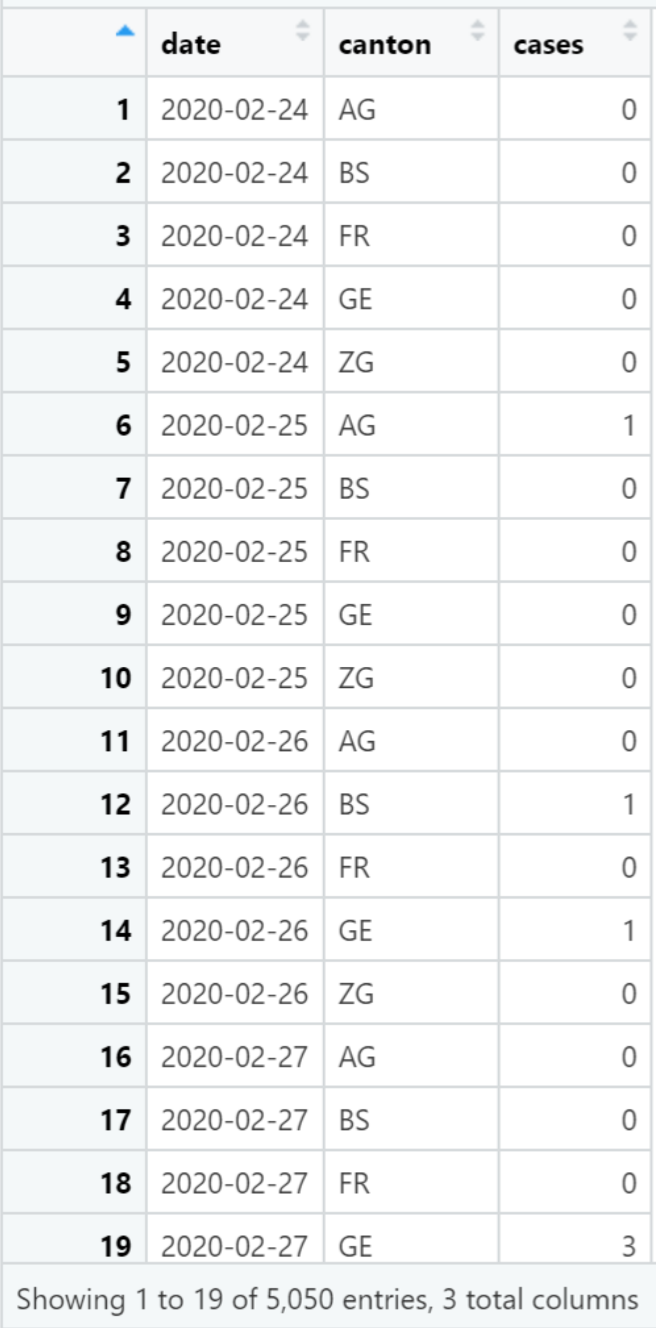
This is what it looks like now:



	Time	AG	BS	FR	GE	ZG
1	2020-02-24	0	0	0	0	0
2	2020-02-25	1	0	0	0	0
3	2020-02-26	0	1	0	1	0
4	2020-02-27	0	0	0	3	0
5	2020-02-28	0	2	0	5	0
6	2020-02-29	2	0	2	0	0
7	2020-03-01	2	1	0	1	1
8	2020-03-02	1	1	0	2	2
9	2020-03-03	1	1	3	0	3
10	2020-03-04	3	6	2	1	1
11	2020-03-05	2	10	1	4	0
12	2020-03-06	2	5	1	17	0
13	2020-03-07	0	5	0	6	0
14	2020-03-08	1	7	4	8	0
15	2020-03-09	3	13	3	29	0
16	2020-03-10	3	18	10	17	0
17	2020-03-11	7	40	8	33	1
18	2020-03-12	7	23	6	49	4
19	2020-03-13	6	23	5	71	1

Showing 1 to 19 of 1,010 entries, 6 total columns

And this is what we want it to look like:



	date	canton	cases
1	2020-02-24	AG	0
2	2020-02-24	BS	0
3	2020-02-24	FR	0
4	2020-02-24	GE	0
5	2020-02-24	ZG	0
6	2020-02-25	AG	1
7	2020-02-25	BS	0
8	2020-02-25	FR	0
9	2020-02-25	GE	0
10	2020-02-25	ZG	0
11	2020-02-26	AG	0
12	2020-02-26	BS	1
13	2020-02-26	FR	0
14	2020-02-26	GE	1
15	2020-02-26	ZG	0
16	2020-02-27	AG	0
17	2020-02-27	BS	0
18	2020-02-27	FR	0
19	2020-02-27	GE	3

Showing 1 to 19 of 5,050 entries, 3 total columns

3.1 Step 1: Pivot from wide to long

Think about which columns need to be pivoted and give these to the `cols` argument of `pivot_longer()`

3.2 Step 2: Rename variables

Use `rename()` to change the column names to match the image above.

Now it's ready for plotting!

4 Recreate desired plot

We want a plot that looks like the one on the website!

This can be done in several steps. See the demo Rmd for more :)

```
## [1] "PLOT!"
```

5 Submission: Upload Rmd HTML

Once you have finished the tasks above, you should **knit this Rmd into an HTML** and **upload both files** on the assignment page.

Workshop 7: Pivoting then grouping

Author’s name here

2022-11-22

- 1 Intro
- 2 Get the assignment repo
- 3 Load packages
- 4 Easy Pivoting
- 5 Intermediate Pivoting
- 6 Bonus: Hard Pivoting (You can do it!)
- 7 Submission: Upload HTML

1 Intro

Welcome!

For this workshop, we will be cleaning a dataset. It is a hands-on approach to using pivoting and groupings

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for the assignment is Tuesday, November 29th at 23:59 PM UTC+2.

2 Get the assignment repo

To get started, you should download and look through the assignment folder.

1. First download the repo to your local computer here.

   You should ideally work on your local computer, but if you would rather work on RStudio Cloud, you can upload the zip file to RStudio Cloud through the Files pane. Consult one of the instructors for guidance on this.

2. Unzip/Extract the downloaded folder.

   If you are on macOS, you can simply double-click on a file to unzip it.

> *If you are on Windows and are not sure how to “unzip” a file, see this image. You need to right-click on the file and then select “extract all”.*

3. Once done, click on the RStudio Project file in the unzipped folder to open the project in RStudio.

4. In RStudio, navigate to the Files tab and open the “rmd” folder. The instructions for your exercise are outlined there (these are the same instructions you see here).

5. Open the “data” folder and observe its components. You will work with the “diet_diversity_vietnam_wide_EASY.csv”, “diet_diversity_vietnam_wide_INTERMEDIATE.csv” and “diet_diversity_vietnam_wide_HARD.csv” files. (The data is from the same source, remodelled for the exercise: you can also open the “00_info_about_the_dataset” file to learn more about this dataset.)

3 Load packages

Now that you understand the structure of the repo, you can load in and clean your dataset.

In the code section below, **load in the needed packages**.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4 Easy Pivoting

For this pivoting, please import the “diet_diversity_vietnam_wide_EASY.csv”. The data frame you import should have 61 rows and 3 columns. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

Using the lesson on pivoting you prepared for today: pivot this dataset into long format. Print the pivoted dataframe as a reactable table.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

5 Intermediate Pivoting

For this pivoting, please import the “diet_diversity_vietnam_wide_INTERMEDIATE.csv”. The data frame you import should have 61 rows and 5 columns.

Using what you learnt in the code demo about some more advanced pivoting, pivot the data to long format. Print the pivoted dataframe as a reactable table.

Hint: Remember to use the neat separator in your column names.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

6 Bonus: Hard Pivoting (You can do it!)

For this pivoting, please import the “diet_diversity_vietnam_wide_HARD.csv”. The data frame you import should have 61 rows and 9 columns. **This is the original data.**

There is no neat separator, think about how you could make one to then pivot the dataframe into long format. Print the pivoted dataframe as a reactable table.

Hint: Think about the `rename()` function of {dplyr}

```
## [1] "WRITE_YOUR_CODE_HERE"
```

7 Submission: Upload HTML

Once you have finished the tasks above, you should **knit this Rmd into an HTML** and **upload it** on the assignment page.

Workshop 6: Intro ggplot & scatter plots

Joy Vaz

2022-11-29

- 1 Introduction
- 2 Get the assignment repo
- 3 Load the data
  - 3.1 **Step 1:** Inspect the data types of your variables
  - 3.2 **Step 2:** Choose your countries.
- 4 Creating a ggplot in layers
  - 4.1 Time series plot
  - 4.2 Extra: explore scatter plots
- 5 Submission: Upload HTML

1 Introduction

Hello!

For the first half of this workshop, we will be plotting datasets about HIV prevalence or incidence from countries around the world.

They number of persons that are estimated to be infected by HIV, including those without symptoms, those sick from AIDS and those healthy due to treatment of the HIV infection

Prevalence is a measure of the number of total cases in a year, and incidence is a measure of the number of new cases in a given time period.

The data also gives you the population for a given country-year.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for the assignment is Tuesday, November 29th at 23:59 PM UTC+2.

2 Get the assignment repo

To get started, you should download and look through the assignment folder.

1. First download the repo to your local computer here.

   You should ideally work on your local computer, but if you would rather work on RStudio Cloud, you can upload the zip file to RStudio Cloud through the Files pane. Consult one of the instructors for guidance on this.

2. Unzip/Extract the downloaded folder.

   If you are on macOS, you can simply double-click on a file to unzip it.

   If you are on Windows and are not sure how to "unzip" a file, see this image. You need to right-click on the file and then select "extract all".

3. Once done, click on the RStudio Project file in the unzipped folder to open the project in RStudio.

4. In RStudio, navigate to the Files tab and open the "rmd" folder. The instructions for your exercise are outlined there (these are the same instructions you see here).

5. Open the "data" folder and observe its components. Chose one of the two HIV datasets - would you like to analyze incidence or prevalence? Discuss with your partner and choose.

3 Load the data

Now that you understand the structure of the repo, you can load your chosen dataset.

In the code section below, **load in the needed packages**.

Now, **read the dataset into R**. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

3.1 Step 1: Inspect the data types of your variables

The kinds of plots you can make is dependent on the data classes of your variable. Take a look at these types with `summary()` or `typeof()`.

Note that this dataset is structured very similarly to the `nigerm` dataset we used in Lesson 1.

3.2 Step 2: Choose your countries.

In this exercises you will be comparing patterns between countries. Since there are dozens of countries in the data, this may be too much to visualize at once. Choose 3-5 countries you would like to analyze.

(Hint: use `dplyr::filter()` to subset your data.)

You will be using `hiv_mini` to plot.

4 Creating a ggplot in layers

4.1 Time series plot

Using the subset you just made above, plot the number of cases over time as line graph. Line graphs are great for visualizing time series.

Think about what x and y aesthetic you need, and then which `geom_*()` function.

```
## [1] "ggplot(data = hiv_mini)"
```

Now, add color to this plot so that you can distinguish between countries. Then, increase the line width. Think about whether these aesthetic are mapping or fixed aesthetics!

Chaining together functions to add layers is a key feature of {ggplot2}. In the scatter plots lesson, we added `geom_smooth()` to `geom_point()` .

Think about what other types of plots besides line graphs can visualize the relationship between two continuous numeric variables.

You can check the *Cheatsheets* tab in your RStudio Help menu and browse the different geoms.

Next, add your another `geom_*()` layer to this plot.

Hint: there's a very common plot type that works for this!

Now, add at least 2 more aesthetics (fixed or mapped) to your plot.

The aesthetic we have covered so far are: - `color` - point color or point outline color

- `size` - point size
- `alpha` - point opacity
- `shape` - point shape
- `fill` - point fill color (only applies if the point has an outline)

Lastly, you can add a title to your plot with the `ggtitle()` function. See `?ggtitle` for more information.

4.2 Extra: explore scatter plots

Next you get to exercise your creativity! You task for this section is to use one of the datasets in the “extra” folder under “data”. Both these datasets contain continuous data that excellent for scatter plots.

1. **Create a scatter plot** with `geom_point()`, and show a relationship between two variables.
2. **Group or scale the points by a third variable**. You can use color, shape, are any other aesthetic to display information about an additional variable.
3. **Add a smoothing line!**

5 Submission: Upload HTML

Once you have finished the tasks above, you should **knit this Rmd into an HTML** and **upload it** on the assignment page.

Workshop 9: Scales and Themes demo

Joy Vaz

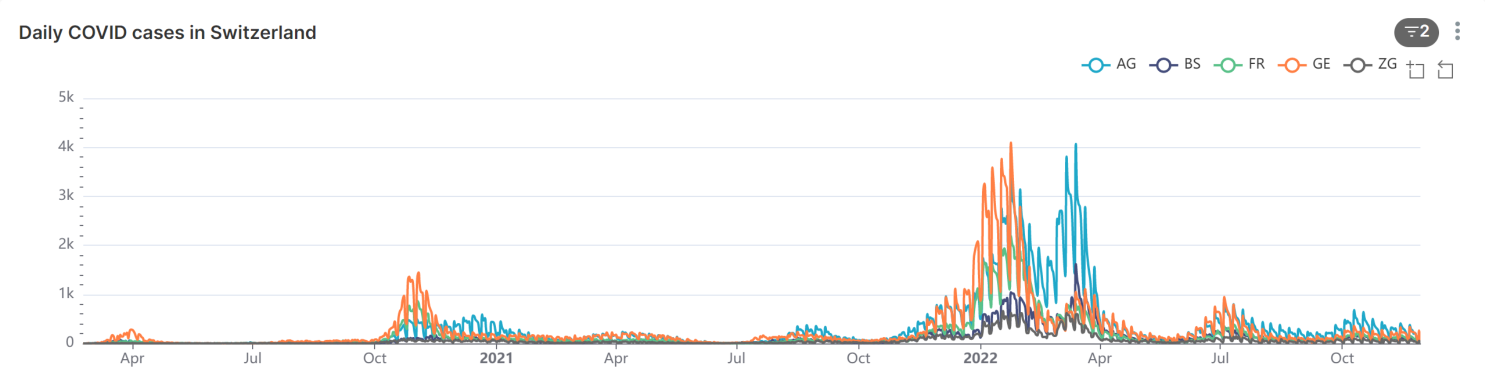
2022-12-06

- 1 Activity 1 demo: Scales and Themes

1 Activity 1 demo: Scales and Themes

Your assignment for Activity 1 is to recreate a plot from the EpiGraphHub COVID-19 Switzerland dashboard in R using {ggplot2}.

The graph we are aiming for looks like this:

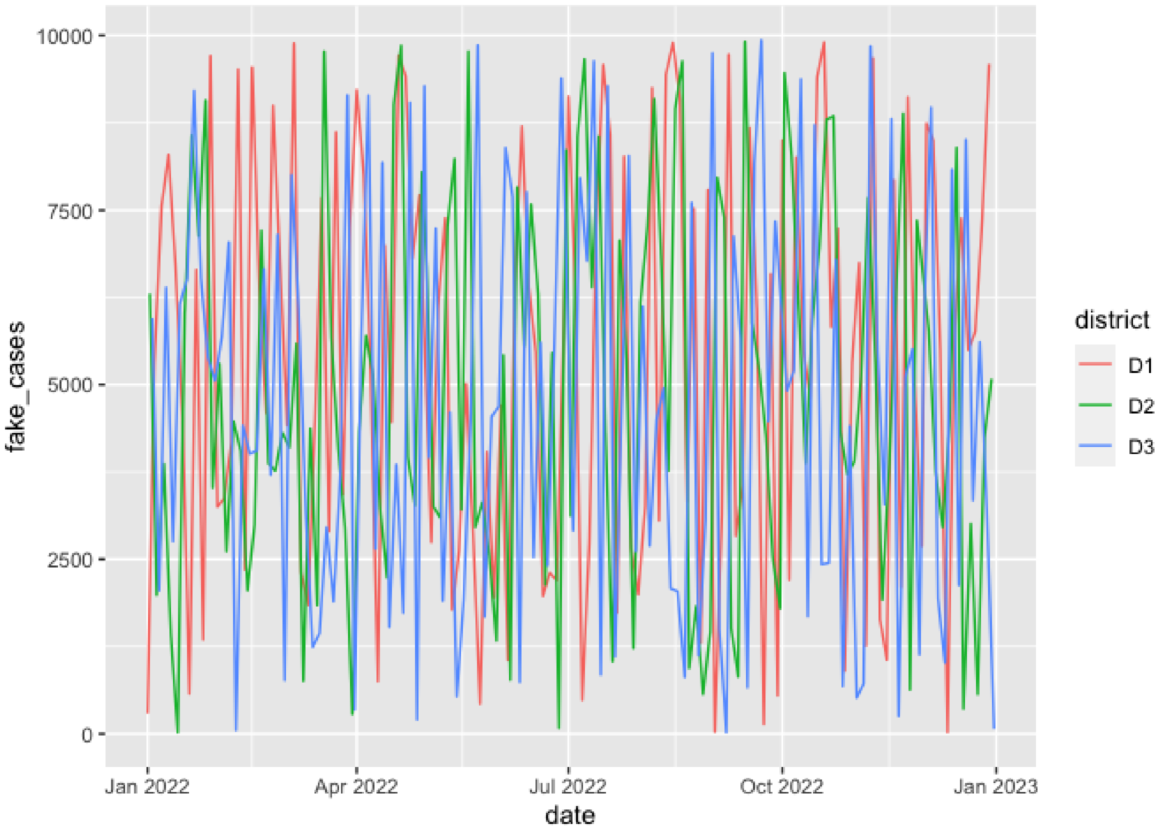


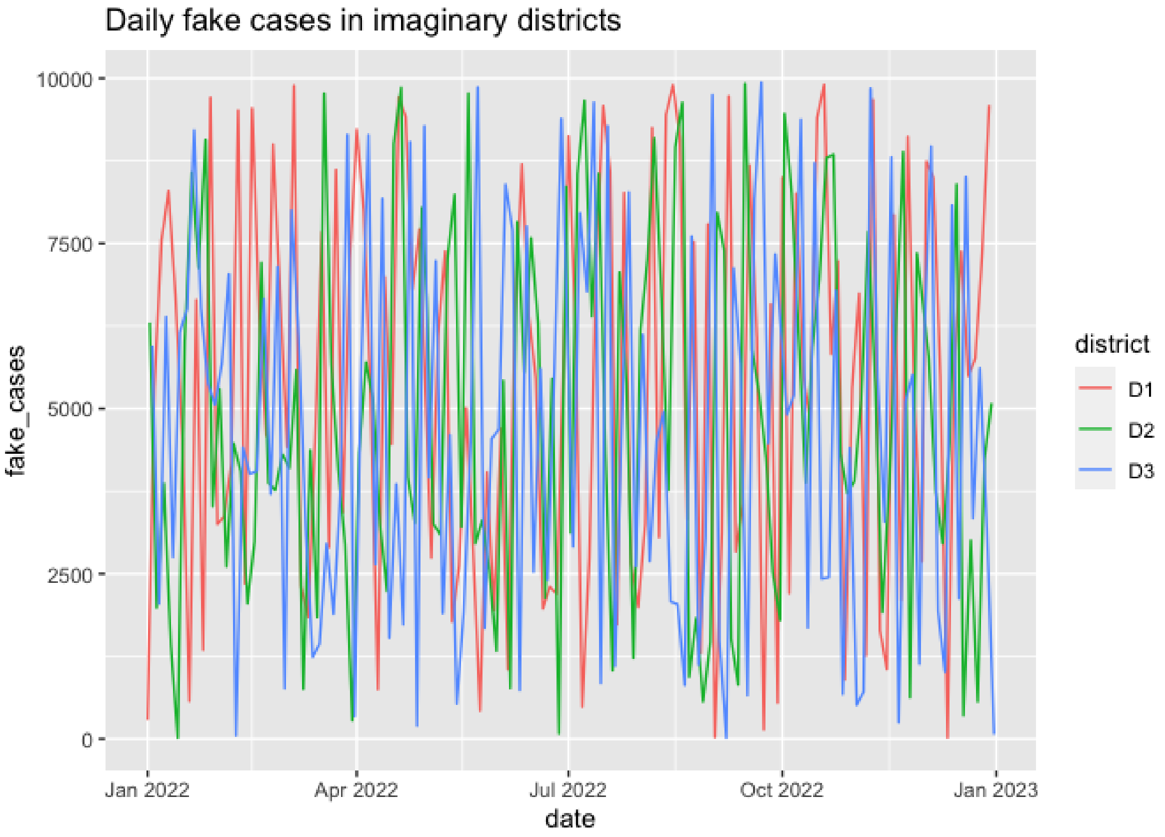
Now most of the elements shown here you have already learned how to create: line graph, group into regions by color, add labels.

However, there are additional modifications we want to make to the scales and themes.

I will demonstrate how to do this using a fake dataset

```
## Warning: 7 failed to parse.
```





To-do

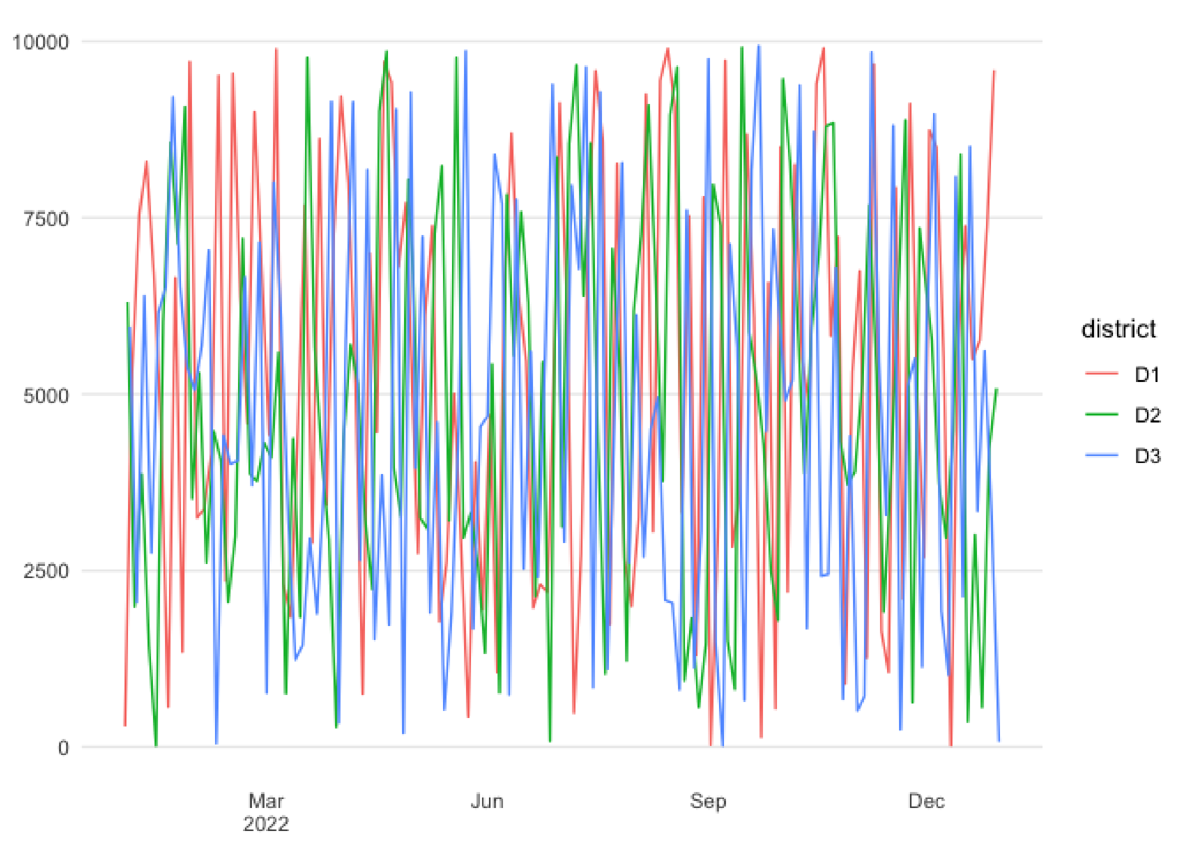
1. Change background from gray to white

2. Relabel x-axis scale breaks to month abbreviations

3. Remove axis titles

4. Remove most grid lines

5. Relabel y-axis scale breaks to shorten 1000s to “k”



Four out of five tasks done!

Now, I leave the last and final step up to you. Do some web searches to learn how to shorten the y-axis labels.

If you have a hard time finding a good answer, check out this page.

Workshop 9: Lines, Scales, Labels, and Themes

Author’s name here

2022-12-06

- 1 Intro
- 2 Instructions
- 3 Prepare the data
  - 3.1 **Step 1:** Pivot from wide to long
  - 3.2 **Step 2:** Rename variables
- 4 Recreate desired plot
- 5 Submission: Upload Rmd and HTML

1 Intro

Welcome!

For this workshop, we have two activities. You will be completing Activity 1 in this Rmd.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for Activity 1 is Tuesday, December 6th at 23:59 PM UTC+2.

2 Instructions

Your assignment for Activity 1 is to recreate a plot from the EpiGraphHub COVID-19 Switzerland dashboard in R using {ggplot2}.

1. First choose one of the four line graphs shown on the dashboard. This is the one you will recreate.

2. Download the data associated with your chosen plot by clicking the three dots on the top right on the plot and selecting “Export CSV”

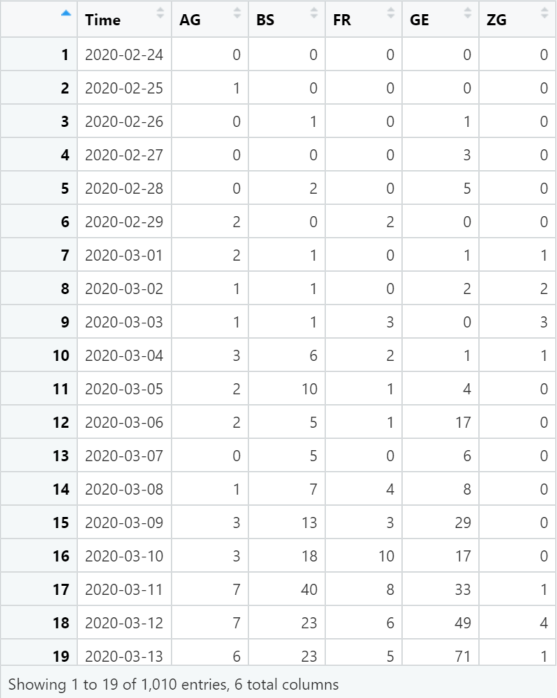
3. Once downloaded, move the CSV to the “data” folder of this R project.

3 Prepare the data

Now, **read the dataset into R**. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

This data needs some cleaning before we can plot it.

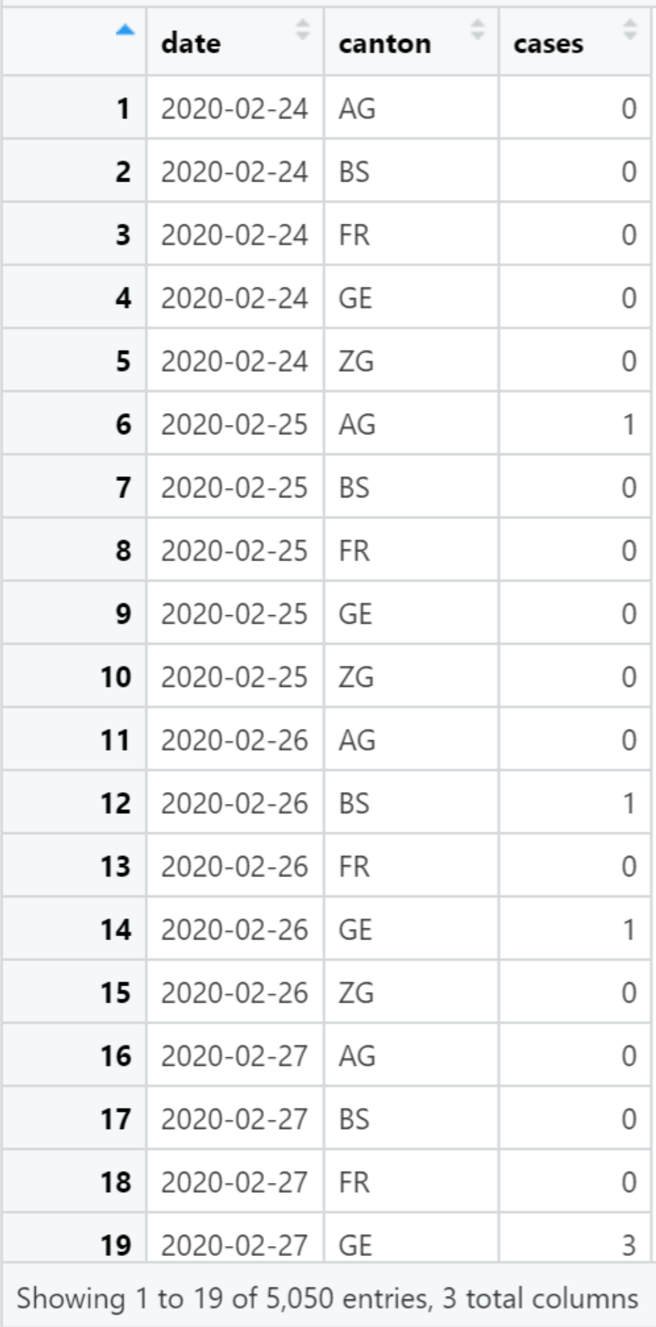
This is what it looks like now:



	Time	AG	BS	FR	GE	ZG
1	2020-02-24	0	0	0	0	0
2	2020-02-25	1	0	0	0	0
3	2020-02-26	0	1	0	1	0
4	2020-02-27	0	0	0	3	0
5	2020-02-28	0	2	0	5	0
6	2020-02-29	2	0	2	0	0
7	2020-03-01	2	1	0	1	1
8	2020-03-02	1	1	0	2	2
9	2020-03-03	1	1	3	0	3
10	2020-03-04	3	6	2	1	1
11	2020-03-05	2	10	1	4	0
12	2020-03-06	2	5	1	17	0
13	2020-03-07	0	5	0	6	0
14	2020-03-08	1	7	4	8	0
15	2020-03-09	3	13	3	29	0
16	2020-03-10	3	18	10	17	0
17	2020-03-11	7	40	8	33	1
18	2020-03-12	7	23	6	49	4
19	2020-03-13	6	23	5	71	1

Showing 1 to 19 of 1,010 entries, 6 total columns

And this is what we want it to look like:



	date	canton	cases
1	2020-02-24	AG	0
2	2020-02-24	BS	0
3	2020-02-24	FR	0
4	2020-02-24	GE	0
5	2020-02-24	ZG	0
6	2020-02-25	AG	1
7	2020-02-25	BS	0
8	2020-02-25	FR	0
9	2020-02-25	GE	0
10	2020-02-25	ZG	0
11	2020-02-26	AG	0
12	2020-02-26	BS	1
13	2020-02-26	FR	0
14	2020-02-26	GE	1
15	2020-02-26	ZG	0
16	2020-02-27	AG	0
17	2020-02-27	BS	0
18	2020-02-27	FR	0
19	2020-02-27	GE	3

Showing 1 to 19 of 5,050 entries, 3 total columns

3.1 Step 1: Pivot from wide to long

Think about which columns need to be pivoted and give these to the `cols` argument of `pivot_longer()`

3.2 Step 2: Rename variables

Use `rename()` to change the column names to match the image above.

Now it's ready for plotting!

4 Recreate desired plot

We want a plot that looks like the one on the website!

This can be done in several steps. See the demo Rmd for more :)

```
## [1] "PLOT!"
```

5 Submission: Upload Rmd and HTML

Once you have finished the tasks above, you should **knit this Rmd into an HTML** and **upload both files** on the assignment page in a ZIP folder.