Exploratory Data Analysis (EDA) Project

Martin Frigaard[CSU, Chico]

[a]California State University, Chico, 400 W 1st St, Chico, CA 95929

This version was compiled on October 31, 2021

Your first project is an exploratory data analysis (EDA) of a dataset of your choosing. This document covers set up instructions, functions for checking the structure of your data, basic summary statistics, and visualizations.

EDA | | Visualizations

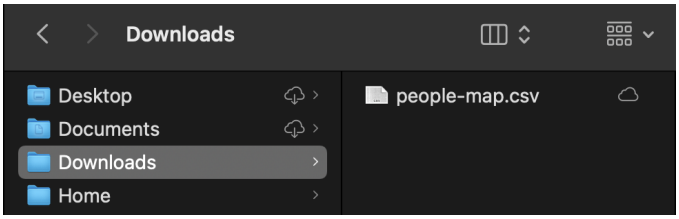
Introduction

This document outlines the requirements for your first project, which is an exploratory data analysis (EDA) of a dataset (or multiple datasets).

Project setup

For this assignment, you've been given a dataset (or datasets) and the EDA template. You should set up your project using the `goodenuffR` package package. You can review the slides for getting started with this package here:

For example, if you've downloaded the `people-map.csv` data into your `Downloads` folder:



We're going to create a new project folder using `goodenuffR`. Assume I have a folder for the course in `Documents/` named `journalism-303/`:

A folder tree for this type of organization is below:

```
Documents/
  |-- journalism-303/
    |-- exercises/
    |-- slides/
```

To create a project folder, I complete the following steps:

1. Use the `getwd()` function to print my current working directory

```
getwd()
# > "/Users/mjfrigaard"
```

2. I want the new project folder to be named `eda-project/` and to be located under the `Documents/journalism-303/` folder, like so:

```
Documents/
  |-- journalism-303/
    |-- exercises/
    |-- slides/
```

```
    |-- eda-project/ <- new project folder!
```

So in the `goodenuffR::goodenuff_project()` function, I enter the **name for the new folder as the `project_name` argument**, and the path to **where I want this folder in the `folder_path` argument.**

```
goodenuffR::goodenuff_project(project_name = "eda-project",
    folder_path = "/Users/mjfrigaard/Documents/journalism-303/")
```

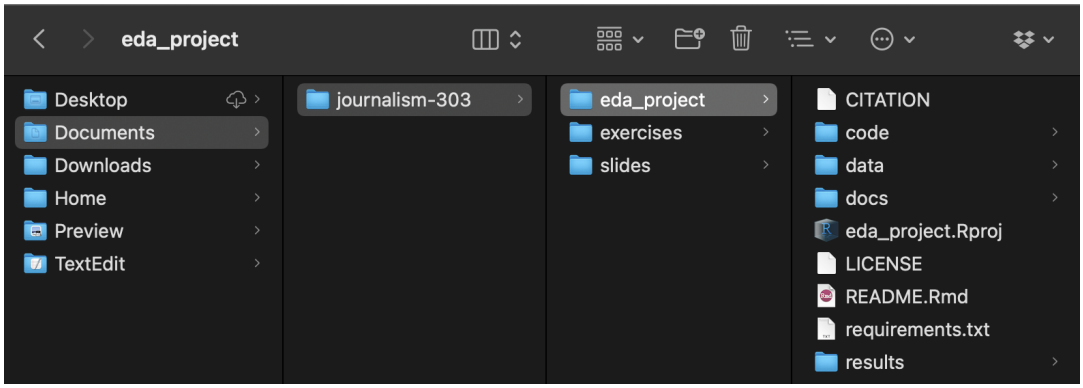
When I execute these commands, a new RStudio session will open and I should see the following folder structure:

```
Documents/
|--journalism-303/
  |-- eda_project/
         |-- eda_project.Rproj
  |-- exercises/
  |-- slides/
```

To create the project files, use the `goodenuffR::goodenuff_files()` function. We can enter this into the console (you should see the following output):

```
goodenuffR::goodenuff_files()
# trying URL 'https://creativecommons.org/publicdomain/zero/1.0/legalcode.txt'
# downloaded 7048 bytes
#
# trying URL 'https://raw.githubusercontent.com/rstudio/rmarkdown/main/inst/rmar
# kdown/templates/github_document/skeleton/skeleton.Rmd'
# Content type 'text/plain; charset=utf-8' length 691 bytes
# ==================================================
# downloaded 691 bytes
```

This will create the following folders and files:



And the folder tree below:

```
Documents/
    journalism-303/
        |-- eda_project/
        |   |-- CITATION
        |   |-- LICENSE
        |   |-- README.Rmd
        |   |-- code/
        |   |   |-- 01-import.R
```

```
|   |   |-- 02-tidy.R
|   |   |-- 03-wrangle.R
|   |   |-- 04-visualize.R
|   |   |-- 05-model.R
|   |   |-- 06-communicate.R
|   |   |-- runall.R
|   |-- data/
|   |   |-- README.md
|   |   |-- raw
|   |-- docs/
|   |   |-- changelog.txt
|   |   |-- manuscript.Rmd
|   |   |-- notebook.Rmd
|   |-- eda_project.Rproj
|   |-- requirements.txt
|   |-- results/
|       |-- figures
|       |-- manuscript
|       |-- tables
|-- exercises/
|-- slides/
```

Outline

After you've set up your project and imported your data into RStudio, you need to answer the following questions:

1. What are the general characteristics of the dataset?

a. *How many rows are in the dataset(s)?*

b. *How many variables are in the dataset(s)?*

2. What are the variable names?

a. *Are the names meaningful?*

b. *Is a data dictionary available (or other info on the dataset)?*

c. *What if the format or type of each variable (i.e. character, numeric, categorical, logical)?*

3. Information about the variables:

a. *For categorical/qualitative variables, what values occurs most frequently?*

b. *Are any of the data missing? If so, how much?*

4. Summary statistics:

a. *Calculate the mean, median, standard deviation, minimum, and maximum for each numerical variable.*

b. *Calculate the counts for each level or unique values for each character variable.*

Dataset characteristics

We'll use the `palmerpenguins` data for this example.

```
# assign to PenguinsRaw
PenguinsRaw <- palmerpenguins::penguins_raw
```

Dimensions. The functions below give you an idea of the dataset shape.

```
nrow(PenguinsRaw) # number of rows
```

```
#>  [1] 344
```

```
ncol(PenguinsRaw) # number of columns
```

```
#>  [1] 17
```

```
dim(PenguinsRaw) # dimensions
```

```
#>  [1] 344  17
```

For information on the data format, use `str()` or `glimpse()`

```
glimpse(PenguinsRaw)
```

```
#>  Rows: 344
#>  Columns: 17
#>  $ studyName             <chr> "PAL0708", "PAL0708", "PAL07~
#>  $ `Sample Number`       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 1~
#>  $ Species               <chr> "Adelie Penguin (Pygoscelis ~
#>  $ Region                <chr> "Anvers", "Anvers", "Anvers"~
#>  $ Island                <chr> "Torgersen", "Torgersen", "T~
#>  $ Stage                 <chr> "Adult, 1 Egg Stage", "Adult~
#>  $ `Individual ID`       <chr> "N1A1", "N1A2", "N2A1", "N2A~
#>  $ `Clutch Completion`   <chr> "Yes", "Yes", "Yes", "Yes", ~
#>  $ `Date Egg`            <date> 2007-11-11, 2007-11-11, 200~
#>  $ `Culmen Length (mm)`  <dbl> 39.1, 39.5, 40.3, NA, 36.7, ~
#>  $ `Culmen Depth (mm)`   <dbl> 18.7, 17.4, 18.0, NA, 19.3, ~
#>  $ `Flipper Length (mm)` <dbl> 181, 186, 195, NA, 193, 190,~
#>  $ `Body Mass (g)`       <dbl> 3750, 3800, 3250, NA, 3450, ~
#>  $ Sex                   <chr> "MALE", "FEMALE", "FEMALE", ~
#>  $ `Delta 15 N (o/oo)`   <dbl> NA, 8.94956, 8.36821, NA, 8.~
#>  $ `Delta 13 C (o/oo)`   <dbl> NA, -24.69454, -25.33302, NA~
#>  $ Comments              <chr> "Not enough blood for isotop~
```

Column names

We want to standardize the column names so they are easier to program with.

```
library(janitor)
Penguins <- PenguinsRaw %>% janitor::clean_names()
glimpse(Penguins)
```

```
#>  Rows: 344
#>  Columns: 17
#>  $ study_name        <chr> "PAL0708", "PAL0708", "PAL0708",~
#>  $ sample_number     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1~
#>  $ species           <chr> "Adelie Penguin (Pygoscelis adel~
#>  $ region            <chr> "Anvers", "Anvers", "Anvers", "A~
#>  $ island            <chr> "Torgersen", "Torgersen", "Torge~
#>  $ stage             <chr> "Adult, 1 Egg Stage", "Adult, 1 ~
#>  $ individual_id     <chr> "N1A1", "N1A2", "N2A1", "N2A2", ~
#>  $ clutch_completion <chr> "Yes", "Yes", "Yes", "Yes", "Yes~
#>  $ date_egg          <date> 2007-11-11, 2007-11-11, 2007-11~
#>  $ culmen_length_mm  <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3~
#>  $ culmen_depth_mm   <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6~
#>  $ flipper_length_mm <dbl> 181, 186, 195, NA, 193, 190, 181~
```

```
#>  $ body_mass_g       <dbl> 3750, 3800, 3250, NA, 3450, 3650~
#>  $ sex               <chr> "MALE", "FEMALE", "FEMALE", NA, ~
#>  $ delta_15_n_o_oo   <dbl> NA, 8.94956, 8.36821, NA, 8.7665~
#>  $ delta_13_c_o_oo   <dbl> NA, -24.69454, -25.33302, NA, -2~
#>  $ comments          <chr> "Not enough blood for isotopes."~
```

Some of these names are a little long, but we can manually change them with `rename()`.

```
Penguins <- Penguins %>%
  rename(
    study = study_name,
    sample = sample_number,
    ind_id = individual_id,
    clutch_cmp = clutch_completion,
    cul_lngth = culmen_length_mm,
    cul_dpth = culmen_depth_mm,
    flip_lngth = flipper_length_mm,
    body_mass = body_mass_g,
    delta15 = delta_15_n_o_oo,
    delta13 = delta_13_c_o_oo
  )
glimpse(Penguins)
```

```
#>  Rows: 344
#>  Columns: 17
#>  $ study      <chr> "PAL0708", "PAL0708", "PAL0708", "PAL07~
#>  $ sample     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ~
#>  $ species    <chr> "Adelie Penguin (Pygoscelis adeliae)", ~
#>  $ region     <chr> "Anvers", "Anvers", "Anvers", "Anvers",~
#>  $ island     <chr> "Torgersen", "Torgersen", "Torgersen", ~
#>  $ stage      <chr> "Adult, 1 Egg Stage", "Adult, 1 Egg Sta~
#>  $ ind_id     <chr> "N1A1", "N1A2", "N2A1", "N2A2", "N3A1",~
#>  $ clutch_cmp <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes~
#>  $ date_egg   <date> 2007-11-11, 2007-11-11, 2007-11-16, 20~
#>  $ cul_lngth  <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9,~
#>  $ cul_dpth   <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8,~
#>  $ flip_lngth <dbl> 181, 186, 195, NA, 193, 190, 181, 195, ~
#>  $ body_mass  <dbl> 3750, 3800, 3250, NA, 3450, 3650, 3625,~
#>  $ sex        <chr> "MALE", "FEMALE", "FEMALE", NA, "FEMALE~
#>  $ delta15    <dbl> NA, 8.94956, 8.36821, NA, 8.76651, 8.66~
#>  $ delta13    <dbl> NA, -24.69454, -25.33302, NA, -25.32426~
#>  $ comments   <chr> "Not enough blood for isotopes.", NA, N~
```

Summary statistics

To calculate summary statistics, I recommend using the `skimr::skim()` function.

```
Penguins %>%
  skim()
```

Table 1. Data summary

Name	Piped data
Number of rows	344
Number of columns	17
Column type frequency:	
character	9
Date	1
numeric	7

Table 1. Data summary

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
study	0	1.00	7	7	0	3	0
species	0	1.00	33	41	0	3	0
region	0	1.00	6	6	0	1	0
island	0	1.00	5	9	0	3	0
stage	0	1.00	18	18	0	1	0
ind_id	0	1.00	4	6	0	190	0
clutch_cmp	0	1.00	2	3	0	2	0
sex	11	0.97	4	6	0	2	0
comments	290	0.16	18	68	0	10	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date_egg	0	1	2007-11-09	2009-12-01	2008-11-09	50

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
sample	0	1.00	63.15	40.43	1.00	29.00	58.00	95.25	152.00
cul_lngth	2	0.99	43.92	5.46	32.10	39.23	44.45	48.50	59.60
cul_dpth	2	0.99	17.15	1.97	13.10	15.60	17.30	18.70	21.50
flip_lngth	2	0.99	200.92	14.06	172.00	190.00	197.00	213.00	231.00
body_mass	2	0.99	4201.75	801.95	2700.00	3550.00	4050.00	4750.00	6300.00
delta15	14	0.96	8.73	0.55	7.63	8.30	8.65	9.17	10.03
delta13	13	0.96	-25.69	0.79	-27.02	-26.32	-25.83	-25.06	-23.79

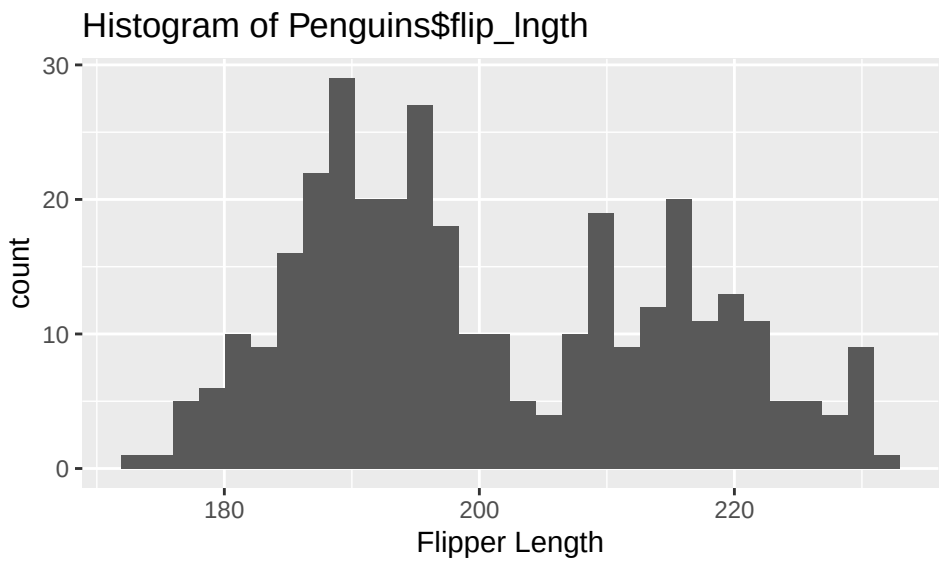
`skim()` output. The `factor` variable is qualitative, so the levels are counted and summarized in the `n_unique` and `top_counts`. The `numeric` variables give us a lot more information, which includes the `n_missing`, `complete_rate`, `mean`, standard deviation (`sd`), minimum (`p0`), 25th percentile (`p25`), median (`p50`), 75th percentile (`p75`), and max (`p100`).

Exploratory visualizations

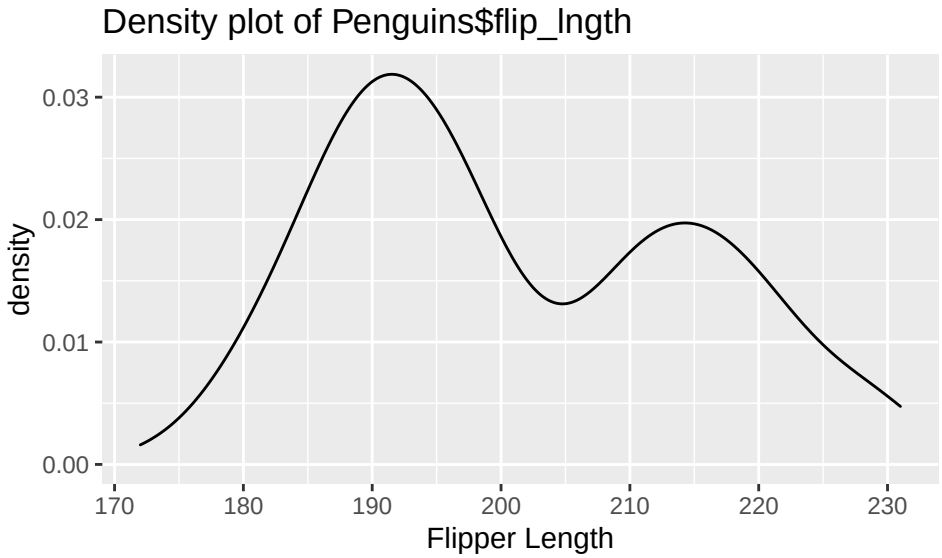
To complete this project, you'll need to create **at least one visualization and explain it's contents.** We've covered how to create visualizations using ggplot2, so feel free to use the code in the exercises or the slides.

Single variable graphs. To view the distribution (or 'shape') of a variable, you can use histograms (`geom_histogram()`), density plots (`geom_density()`), frequency polygons (`geom_freqpoly()`), or box-plots `geom_boxplot()`.

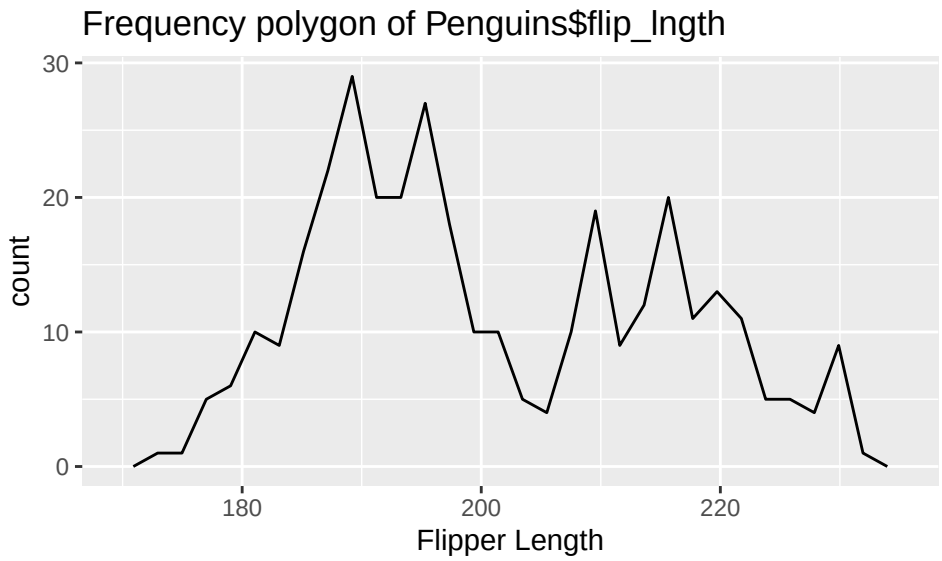
```
# label
labs_hist <- labs(title = "Histogram of Penguins$flip_lngth",
                  x = "Flipper Length")
Penguins %>%
  ggplot(aes(x = flip_lngth)) +
    geom_histogram() +
      labs_hist
```



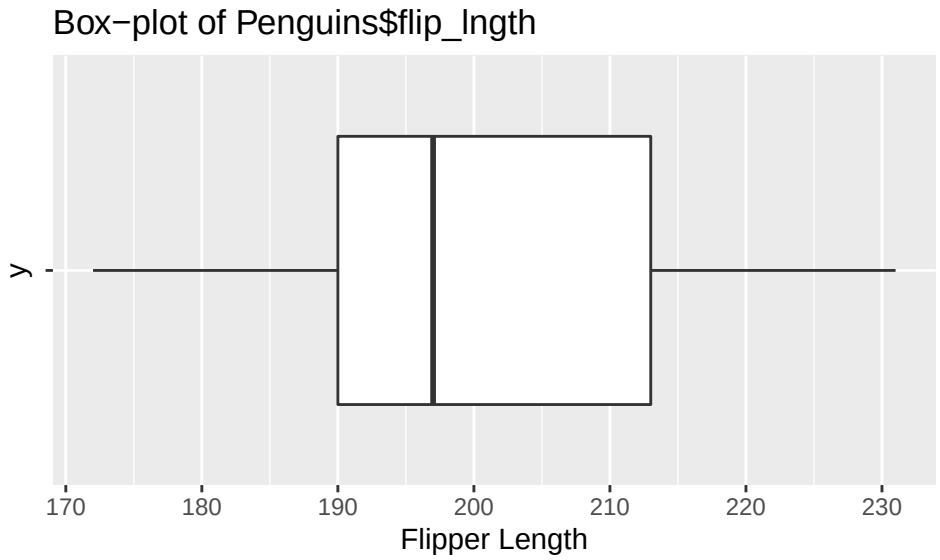

```
# label
labs_dens <- labs(title = "Density plot of Penguins$flip_lngth",
                  x = "Flipper Length")
Penguins %>%
  ggplot(aes(x = flip_lngth)) +
    geom_density() +
      labs_dens
```




```
# label
labs_freq <- labs(title = "Frequency polygon of Penguins$flip_lngth",
                  x = "Flipper Length")
Penguins %>%
  ggplot(aes(x = flip_lngth)) +
    geom_freqpoly() +
      labs_freq
```




```
# label
labs_box <- labs(title = "Box-plot of Penguins$flip_lngth",
                 x = "Flipper Length")
Penguins %>%
  ggplot(aes(x = flip_lngth,
             y = "")) +
  geom_boxplot() +
  labs_box
```

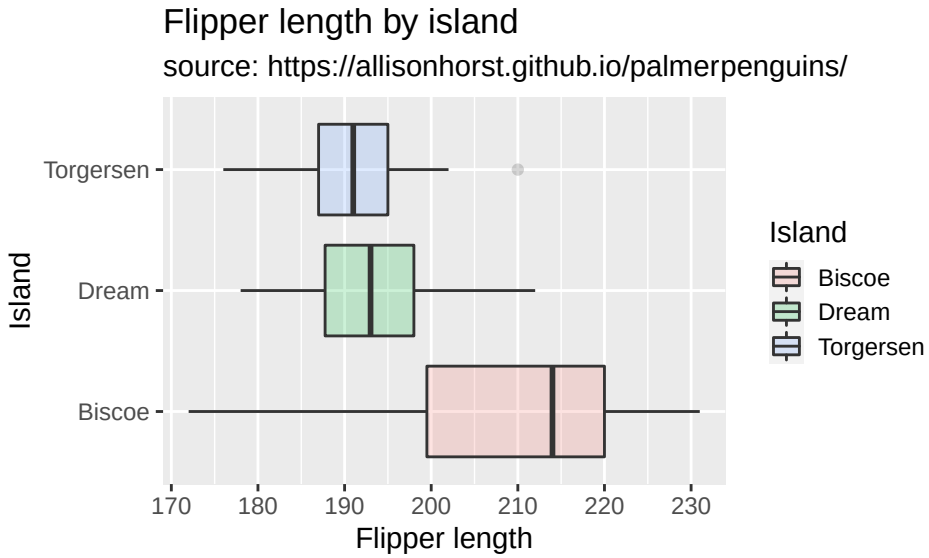


Relationships between variables

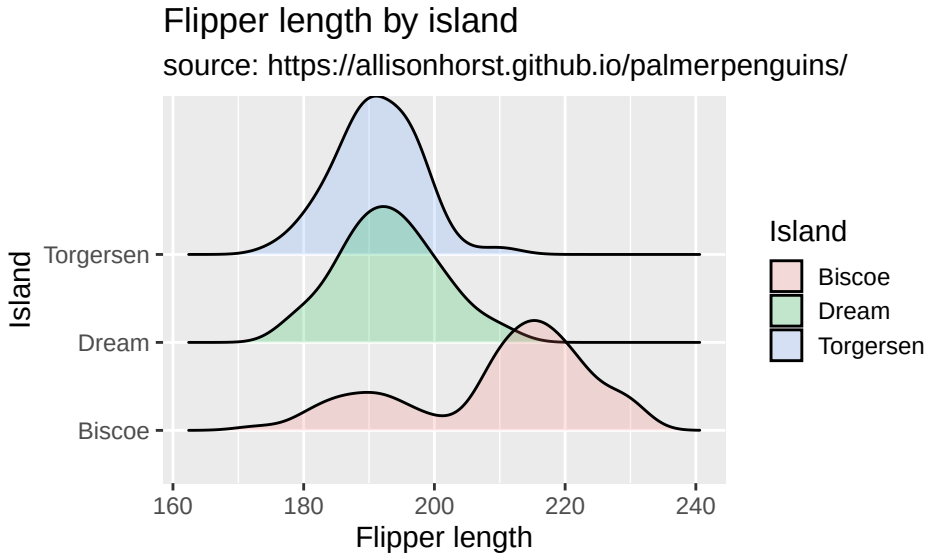
For exploring relationships between variables, we can use box-plots and ridgeline plots (from the `ggridges` package).

```
labs_box2 <- labs(
     title = "Flipper length by island",
     subtitle = "source: https://allisonhorst.github.io/palmerpenguins/",
     fill = "Island",
     x = "Flipper length",
     y = "Island")
Penguins %>%
  ggplot() +
```

```
  geom_boxplot(aes(x = flip_lngth,
                   y = island,
                   fill = island),
                   alpha = 1/5) +
  labs_box2
```




```
library(ggridges)
lab_ridges <- labs(
     title = "Flipper length by island",
     subtitle = "source: https://allisonhorst.github.io/palmerpenguins/",
     fill = "Island",
     x = "Flipper length",
     y = "Island")
Penguins %>%
  ggplot() +
  geom_density_ridges(aes(x = flip_lngth,
                          y = island,
                          fill = island),
                      alpha = 1/5) +
  lab_ridges
```



Data dictionary

To complete this project, you need to turn in your `.html` report and a data dictionary for the dataset which documents each variable and their contents. For examples, see the example for the `palmerpenguins::penguins` data:

```
??palmerpenguins::penguins
```

Example data dictionary.

Size measurements for adult foraging penguins near Palmer Station, Antarctica Description Includes measurements for penguin species, island in Palmer Archipelago, size (flipper length, body mass, bill dimensions), and sex. This is a subset of penguins_raw.

Usage: `penguins`

Format: A tibble with 344 rows and 8 variables:

`species`: a factor denoting penguin species (Adélie, Chinstrap and Gentoo)

`island`: a factor denoting island in Palmer Archipelago, Antarctica (Biscoe, Dream or Torgersen)

`bill_length_mm`: a number denoting bill length (millimeters)

`bill_depth_mm`: a number denoting bill depth (millimeters)

`flipper_length_mm`: an integer denoting flipper length (millimeters)

`body_mass_g`: an integer denoting body mass (grams)

`sex`: a factor denoting penguin sex (female, male)

`year`: an integer denoting the study year (2007, 2008, or 2009)

Source Adélie penguins: Palmer Station Antarctica LTER and K. Gorman. 2020. Structural size measurements and isotopic signatures of foraging among adult male and female Adélie penguins (Pygoscelis adeliae) nesting along the Palmer Archipelago near Palmer Station, 2007-2009 ver 5. Environmental Data Initiative https://doi.org/10.6073/pasta/98b16d7d563f265cb52372c8ca99e60f

Gentoo penguins: Palmer Station Antarctica LTER and K. Gorman. 2020. Structural size measurements and isotopic signatures of foraging among adult male and female Gentoo penguin (Pygoscelis papua) nesting along the Palmer Archipelago near Palmer Station, 2007-2009 ver 5. Environmental Data Initiative https://doi.org/10.6073/pasta/7fca67fb28d56ee2ffa3d9370ebda689

Chinstrap penguins: Palmer Station Antarctica LTER and K. Gorman. 2020. Structural size measurements and isotopic signatures of foraging among adult male and female Chinstrap penguin (Pygoscelis antarcticus) nesting along the Palmer Archipelago near Palmer Station, 2007-2009 ver 6. Environmental Data Initiative https://doi.org/10.6073/pasta/c14dfcfada8ea13a17536e73eb6fbe9e

Originally published in: Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081