Lab 01: Review R + inference

Isaac Plotkin

1/18/2022

Data: Trails in San Francisco, CA.

Today's data comes from the Metropolitan Transportation Commission (MTC) Open Data Catalog an Open Data program managed by the MTC and the Association of Bay Area Governments to provide local agencies and the public with their data needs.

In this lab, we will focus on data about the existing and planned segments of the San Francisco Bay trail. The data is located in the *SFO_trails.csv* file located in the *data* folder. Use the code below to read in the .csv file and save it in the RStudio environment as a data frame called `trails`.

```
trails <- read_csv("data/SFO-trails.csv")
```

A full list of the variables in the dataset is available here. For today's analysis, we will primarily focus on the following variables:

`status`	Whether the trail is proposed or existing
`class`	Category for the trail segment (4 types)
`length`	Length of the trail segment in miles

Exercises

Write your answers in complete sentences and show all code and output.

Before doing any analysis, we may want to get quick view of the data. This is a useful thing to do after importing data to see if the data imported correctly. One way to do this, is to look at the actual dataset. Type the code below in the **console** to view the entire dataset.

```
View(trails)
```

Exploratory Data Analysis

1. Now that we've had a quick view of the dataset, let's get more details about its structure. Sometimes viewing a summary of the data structure is more useful than viewing the raw data, especially if the dataset has a large number of observations and/or rows. Run the code below to use the `glimpse` function to see a summary of the `trails` dataset.

   How many observations are in the `trails` dataset? How many variables? There are 739 observations and 12 variables.

```
glimpse(trails)
```

2. Before conducting statistical inference (or eventually fitting regression models), we need do some exploratory data analysis (EDA). Much of EDA consists of visualizing the data but it also includes calculating summary statistics for the variables in our dataset. Let's begin by examining the distribution of `status` with a data visualization and summary statistics.

   - What is a type of graph that's appropriate to visualize the distribution of `status`? Fill in the `ggplot` code below to plot the distribution of `status`. Include informative axis labels and title on the graph.
   - Then, calculate the proportion of observations in each category of `status` by completing the code below.

```
ggplot(data = trails, aes(x = status)) +
  geom_bar()  +
  labs(x = "Status",
       y = "Number of Trails",
       title = "Status of trails")
```

```
trails %>%
  count(status) %>%
  mutate(proportion = n / sum(n))
```

3. Since we want to analyze characteristics for trails in the Bay Area, we will just use data from currently existing trails for the remainder of the analysis. Complete the code below to use the `filter` function to create a subset consisting only of trails that currently exist and have a value reported for `length`. Assign the subset the name `current_trails`. (*Hint: There should be 493 observations in current_trails.*)

```
current_trails <- trails %>%
  filter(status == "Existing", !is.na(length))
current_trails
```

This is a good place to knit, commit, and push changes to your remote lab-01 repo on GitHub. Be sure to write an informative commit message (e.g. "Completed exercises 1 - 3"), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane in RStudio should be empty."

Use `current_trails` for Exercises 4 - 7.

4. Let's examine the distribution of `length`. One important part of EDA is creating data visualizations to see the shape, center, spread, and outliers in a distribution. Data visualizations are also useful for examining the relationship between multiple variables. There are a lot of ways to make data visualizations in R; we will use the functions available in the `ggplot2` package.

   Make a graph to visualize the distribution of `length`. Include an informative title and axis labels.

```
ggplot(data = current_trails) +
  geom_histogram(mapping = aes(x = length), binwidth = 0.5) +
labs(x = "Length",
       y = "Count",
       title = "Length Histogram")
```

```r
ggplot(data = current_trails, mapping = aes(x = as.factor(class), y = length)) +
  geom_boxplot() +
  labs(x = "Class",
       y = "Length",
       title = "Length and Class Boxplot")
```

See Section 7.3.1 "Visualizing Distributions" or the ggplot2 reference page for details and example code.

5. Next, fill in the code below to use the `summarise` function to calculate various summary statistics for the variable `length`. You can use the summarise reference page for more information about the function and example code.

```r
current_trails %>%
  summarise(min = min(length),
            q1 = quantile(length, 0.25),
            median = median(length),
            q3 = quantile(length, 0.75),
            max = max(length),
            iqr = IQR(length),
            mean = mean(length),
            std_dev = sd(length)
            )
```

6. Describe the distribution of `length`. Your description should include comments about the shape, center, spread, and any potential outliers. Use the graph from Exercise 4 and relevant summary statistics from Exercise 5 in your description.

There are both small and large outliers in the distribution of length. The histogram is right skewed with most of the data points centered around the mean. There are quite a few points that are much larger than the mean and 3rd quantile. Most of the longer trails are class 1 trails.

7. We want to limit the analysis to trails that are more likely intended for day hikes, rather than multi-day hikes and camping. Therefore, let's remove the extreme outliers from the data for this analysis and only consider those trails that are 5 miles or shorter.

   Filter the dataset to remove the extreme outliers. **Be sure to save the updated dataset, so you can use it for the remainder of the lab.**

```r
day_trails <- trails %>%
  filter(length <= 5)
day_trails
```

This is a good place to knit, commit, and push changes to your remote lab-01 repo on GitHub. Be sure to write informative commit message (e.g. "Completed exercises 4 - 7"), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane in RStudio should be empty."

8. Consider the distribution of `class`.

   - What are the values of `class` in the dataset? Show the code and output to support your answer.

   1 2 3 and NA (which could mean 4). There are no trails in class 4.

- What do you think is the most likely reason for the missing observations of `class`? In other words, what does a missing value of `class` indicate?

Maybe there are not any trails developed far enough to be considered in class 4. Might be difficult to achieve a class 4 trail.

```
print(unique(trails$class))
ggplot(data = trails, aes(x = class)) +
  geom_bar()  +
  labs(x = "Class",
       y = "Number of Trails",
       title = "Classes of trails")
```

9. Complete the code below to impute (i.e. fill in) the missing values of `class` with the appropriate value. After that, eliminate all the observations from class = 3, since we are not going to use the. Then, display the distribution of `class` to check that the missing values were correctly imputed.

```
updated_trails <- trails %>% filter(class !=3) %>%
  mutate(class = if_else(is.na(class),3,class))

updated_trails
```

10. Now that we've completed the univariate EDA (i.e. examining one variable at a time), let's examine the relationship between the length of the trail and its class variable. Make a graph to visualize the relationship between `length` and `class` and calculate the appropriate summary statistics. Include informative axis labels and title on your graph.

```
ggplot(data = trails, mapping = aes(x = as.factor(class), y = length)) +
 geom_boxplot() +
 labs(title = "Class vs. Length",
      x = "Class",
      y = "Length")

print(current_trails %>%
  summarise(min = min(class),
            median = median(class),
            max = max(class),
            mean = mean(class),
            std_dev = sd(class)
            ))

print(current_trails %>%
  summarise(min = min(length),
            q1 = quantile(length, 0.25),
            median = median(length),
            q3 = quantile(length, 0.75),
            max = max(length),
            iqr = IQR(length),
            mean = mean(length),
            std_dev = sd(length)
            ))
```

11. Describe the relationship between `length` and `class`. In other words, describe how the distribution of `length` compares between trails that have different classes (1 = shared use bicycle and pedestrian

path, 2 = bike lane, and 3 = bike route). Include information from the graph and summary statistics from the previous exercise in your response.

Trails of class 1 (shared use bicycle and pedestrian path) tend to have the longest trails; however, trails from class 3 (bike route) have the longest trails on average. Most of the trails lie within class one or two which means there are not many trails with just a bike route.

This is a good place to knit, commit, and push changes to your remote lab-01 repo on GitHub. Be sure to write informative commit message (e.g. "Completed exercises 8 - 11"), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane in RStudio should be empty."

Statistical Inference

We'd like to use the data from the trails in SFO to make more general conclusions about trails in urban areas in California, United States. We will reasonably consider the trails in SFO representative of the trails in other urban areas in the West Coast of United States.

Over the next few questions, will use statistical inference to assess whether there is a difference in the mean length of trails that share use bicycle and pedestrian path (class = 1) and those that only have a bike lane (class = 2).

12. The following conditions must be met when we conduct statistical inference on the difference in means between two groups. For each condition, specify whether it is met and a brief explanation of your reasoning.

    - **Independence**

    The sampled observations are random and less than 10% of the population size.

    - **Sample Size**

    There are over 30 samples

    - **Independent Groups**

    Trails that share bicycle and pedestrian path and those that only have a bike lane do not depend on each other at all.

13. While we have observed a small difference in the mean length in trails with bike lanes (class = 2) and trails that share bikes with pedestrians (class = 1), let's assess if there is enough evidence to consider the difference "statistically significant" or if it appears to be due to random chance.

The null and alternative hypotheses are written in statistical notation below. State the hypotheses in words in the context of this analysis.

$$H_0 : \mu_1 - \mu_2 = 0 H_a : \mu_1 - \mu_2 \neq 0$$

The null hypothesis is that there is no statistically significant difference between the mean length in trails with bike lanes and trails that share bikes with pedestrians.

The alternative hypothesis is that there is statistically significant difference between the mean length in trails with bike lanes and trails that share bikes with pedestrians.

14. Fill in the code below to use the `t.test` function to calculate the test statistic and p-value. Replace `response` with the variable we're interested in drawing conclusions about and `group_var` with the variable used to define the two groups.

```r
?t.test # to see the help page from the function
t.test(length ~ class, data = updated_trails,
       alternative = "two.sided",
       conf.level = 0.99) #less, greater, or two.sided
```

15. Use the output from the previous exercise to answer the following:

    - Write the definition of the test statistic in the context of this analysis.

    The test statistic means there is a 0.60392 difference in t value between the length in trails with bike lanes and trails that share bikes with pedestrians.

    - Write the definition of the p-value in the context of this analysis.

    There is a 54% chance that the null hypothesis is true given this data.

    - State your conclusion in the context of this analysis. Use a significance level of $\alpha = 0.01$.

    There is not statistically significant evidence to prove that there is a difference between the mean length in trails with bike lanes and trails that share bikes with pedestrians. The p value is 0.54 which is much larger than the significance level of 0.01.

16. Notice the confidence interval for the difference in mean trail length printed in the output from Exercise 14. Interpret this confidence interval in the context of this analysis.

With repeated sampling, there is a 99% chance that the true population difference in group 1 mean and group 2 mean is between -0.1254090 and 0.2017853.

You're done and ready to submit your work! Knit, commit, and push all remaining changes. You can use the commit message "Done with Lab 1!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and thatall documents are updated in your repo on GitHub. Then submit the pdf for your assignment on Gradescope. Include your repo name, so I can check your commits.