

# STAT 108: Lab 1

Justin Chan

1/20/2022

## 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("SFO-trails.csv")
```

```
## Rows: 739 Columns: 12
```

```
## -- Column specification -----
```

```
## Delimiter: ","
```

```
## chr (7): county, city, surface, agency, status, year_cmplt, legend
```

```
## dbl (5): objectid, class, seg_num, length, SHAPE_Length
```

```
##
```

```
## i Use 'spec()' to retrieve the full column specification for this data.
```

```
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

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

---

<b>status</b>	Whether the trail is proposed or existing
<b>class</b>	Category for the trail segment (4 types)
<b>length</b>	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 or rows and 12 variables or columns in the 'trails' dataset.

```
glimpse(trails)
```

```
## Rows: 739
## Columns: 12
## $ objectid    <dbl> 2952, 2953, 2954, 2955, 2956, 2957, 2958, 2959, 2960, 296~
## $ county      <chr> "Marin", "Marin", "Marin", "San Mateo", "San Mateo", "San~
## $ city        <chr> "Novato", "Novato", "San Rafael", "Brisbane", "S San Fran~
## $ surface     <chr> NA, NA, NA, NA, "paved", NA, "paved", NA, NA, NA, NA, ~
## $ class       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 2, ~
## $ agency      <chr> "Caltrans", "Sonoma-Marin Area Rail Transit", "San Rafael~
## $ status      <chr> "Proposed", "Proposed", "Proposed", "Proposed", "Existing~
## $ seg_num     <dbl> 9002, 9009, 9024, 2001, 2010, 1032, 2047, 2042, 2089, 206~
## $ length      <dbl> 3.20483759, 2.21318493, 1.47142826, 1.24527351, 0.5966338~
## $ year_cmplt  <chr> NA, NA, NA, NA, "2009", NA, NA, NA, NA, NA, NA, NA, "2014~
## $ legend      <chr> "Planned Bay Trail", "Planned Bay Trail", "Planned Bay Tr~
## $ SHAPE_Length <dbl> 0.052688281, 0.034781330, 0.022816134, 0.018364298, 0.009~
```

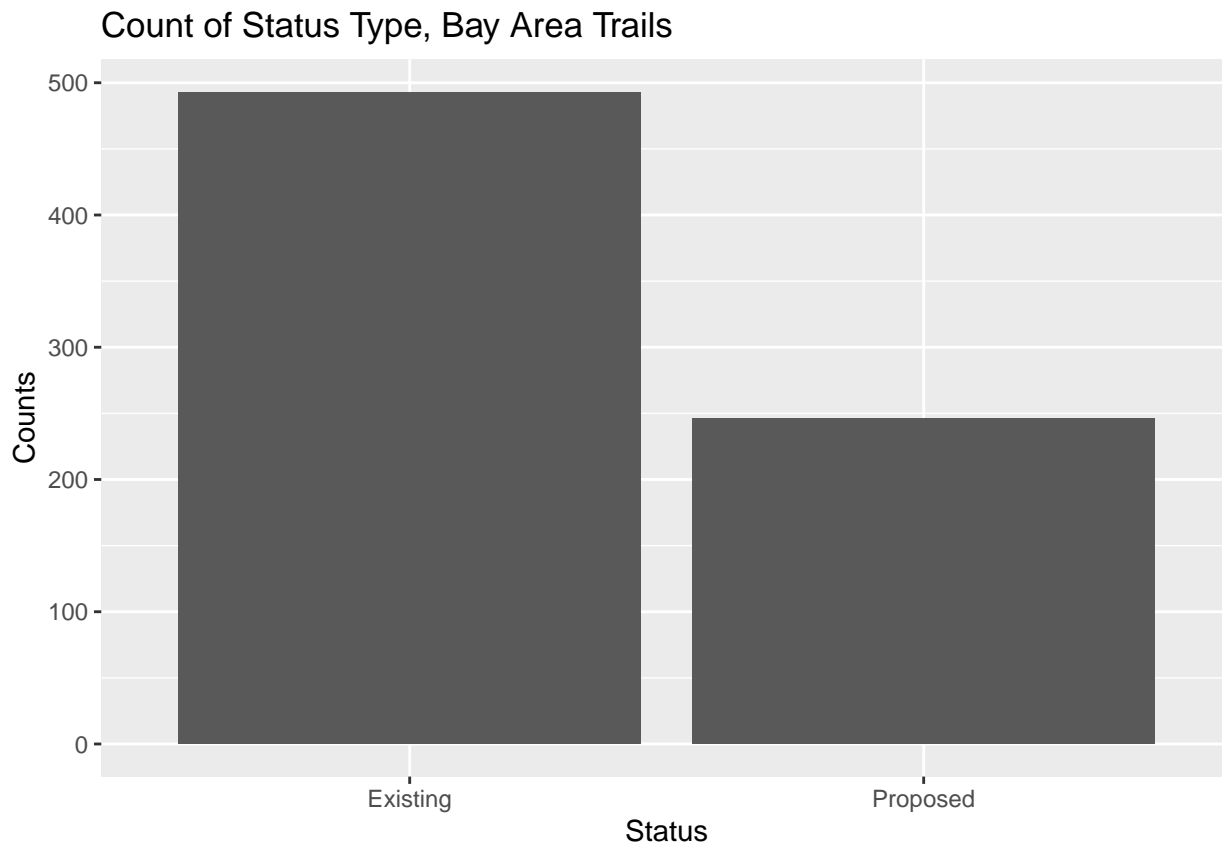
2. Before conducting statistical inference (or eventually fitting regression models), we need to do some exploratory data analysis (EDA). Much of EDA consists of visualizing the data but it also includes calculating summary statistics for the variables in our dataset. Let's begin by examining the distribution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

The type of graph that is appropriate to visualize the distribution of 'status' would be a bar graph since it is a categorical variable and we would like to see the counts of each 'status' type.

- 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 = "Counts",
       title = "Count of Status Type, Bay Area Trails")
```



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

```
## # A tibble: 2 x 3
##   status      n proportion
##   <chr>    <int>     <dbl>
## 1 Existing   493     0.667
## 2 Proposed   246     0.333
```

- 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))
```

*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."*

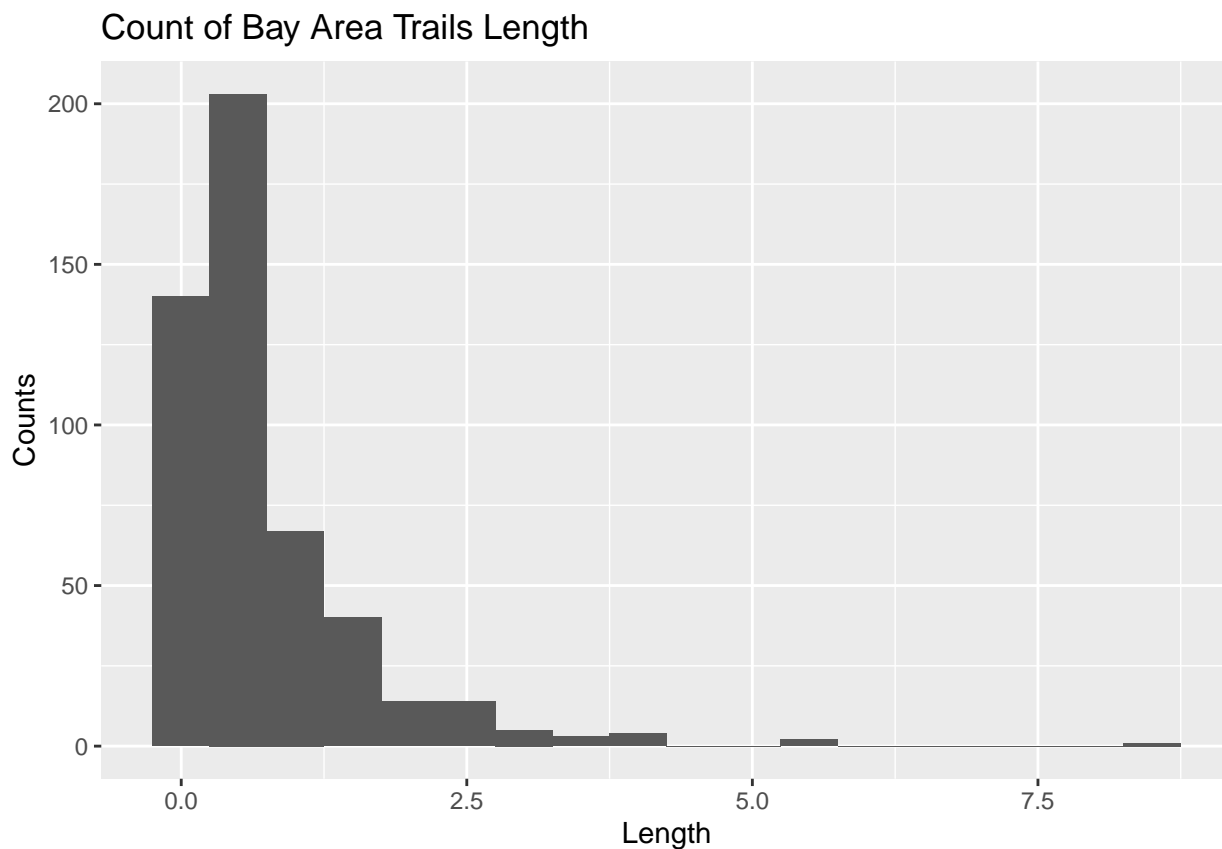
**Use `current_trails` for Exercises 4 - 7.**

- 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

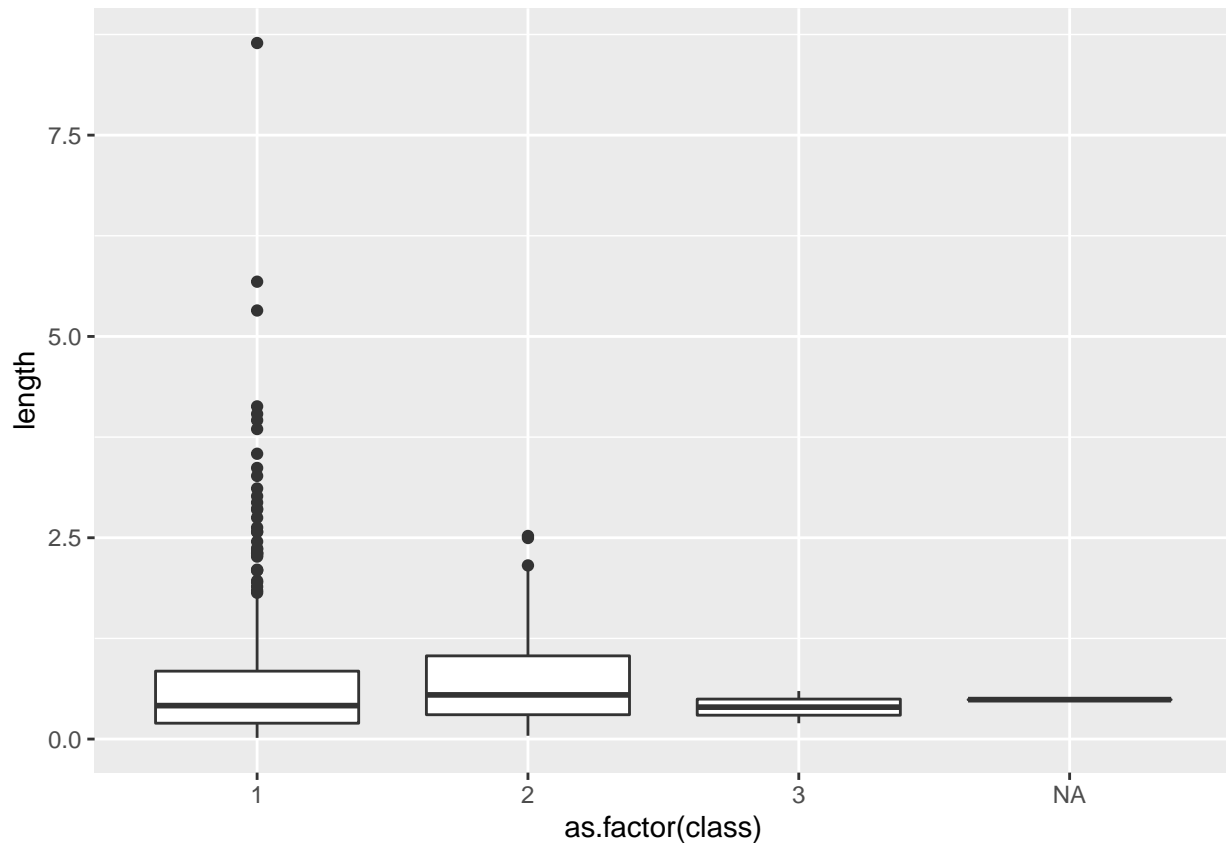
for examining the relationship between multiple variables. There are a lot of ways to make data visualizations in R; we will use the functions available in the `ggplot2` package.

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 = "Counts",  
       title = "Count of Bay Area Trails Length")
```



```
ggplot(data = current_trails, mapping = aes(x = as.factor(class), y = length)) +  
  geom_boxplot()
```



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

- 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.

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

```
## # A tibble: 1 x 8
##   min    q1 median    q3   max   iqr  mean std_dev
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.0148 0.209 0.448 0.936 8.64 0.727 0.724 0.852
```

- 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.

Based on histogram created in Exercise 4, the variable ‘length’ in the current\_trails dataset is right skewed, uni-modal where there are outliers in the distribution on the right where few trails are really long. There seems to be peak at around 1 which means that a good chunk of Bay Area trails have a ‘length’ of about 1 where they are concentrated from in the histogram. The spread of trail lengths may vary since most trails have short lengths of about 1 but, there are few trails that have longer lengths that play a factor in increasing the spread of ‘length’. The relevant summary statistics from Exercise 5 give further evidence of the findings in Exercise 4. The graph seems to skew to the right since the mean of ‘length’ is 0.724 while the median is much smaller, 0.448, meaning that some larger outliers are increasing the mean of the distribution. The center of the distribution is usually described as the median since it is not affected by outliers while the mean is. The spread of the distribution can be described as a concentration of small trail lengths and a more wider spread of trail lengths as ‘length’ increases shown through the five number summary of min, q1, median, q3, and max. The standard deviation of the ‘length’ distribution is 0.852. The outliers shown in the Exercise 4 histogram can be seen with trails with length > 5 since there are not any trails in neighboring bins for ‘length’.

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.**

```
no_outliers_trails <- current_trails %>%
  filter(length < 5.0)
```

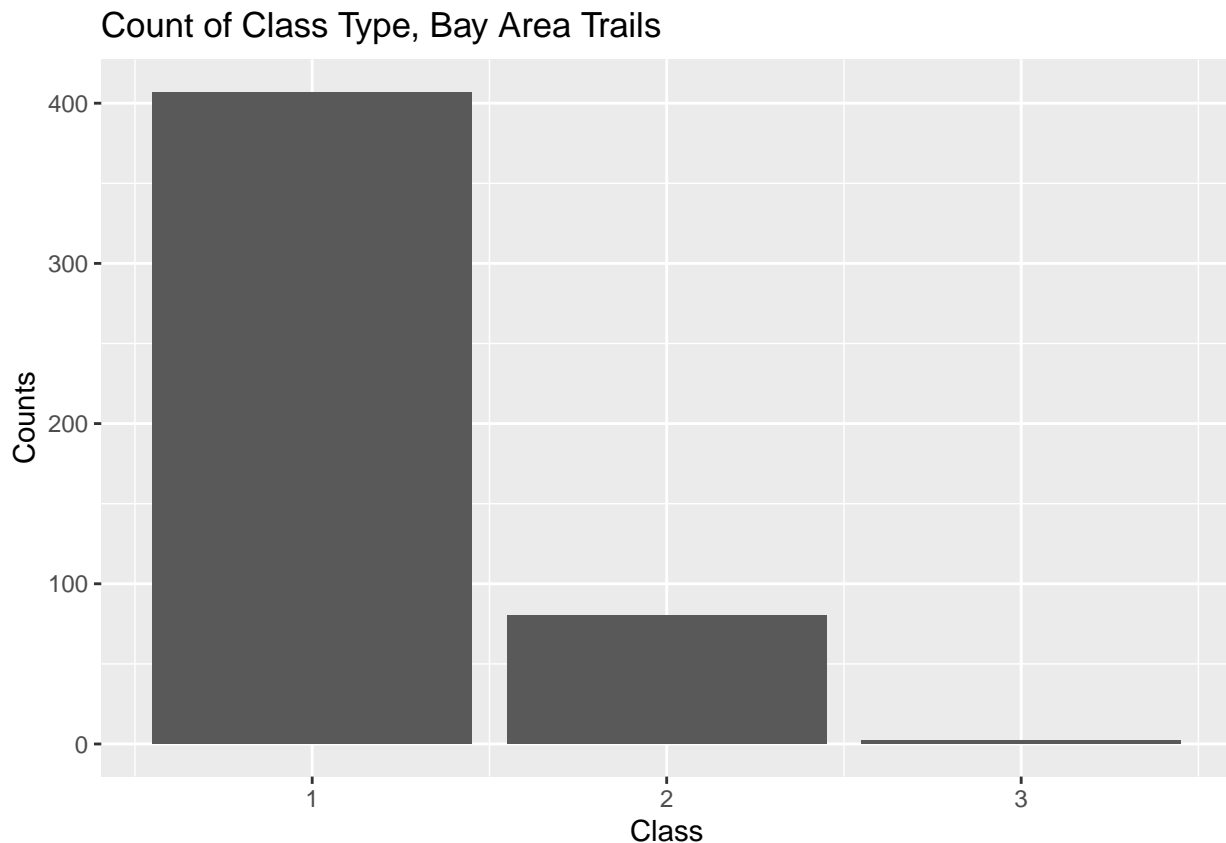
*This is a good place to knit, commit, and push changes to your remote lab-01 repo on GitHub. Be sure to write informative commit message (e.g. “Completed exercises 4 - 7”), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane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

```
ggplot(data = no_outliers_trails, aes(x = class)) +
  geom_bar() +
  labs(x = "Class",
       y = "Counts",
       title = "Count of Class Type, Bay Area Trails")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```



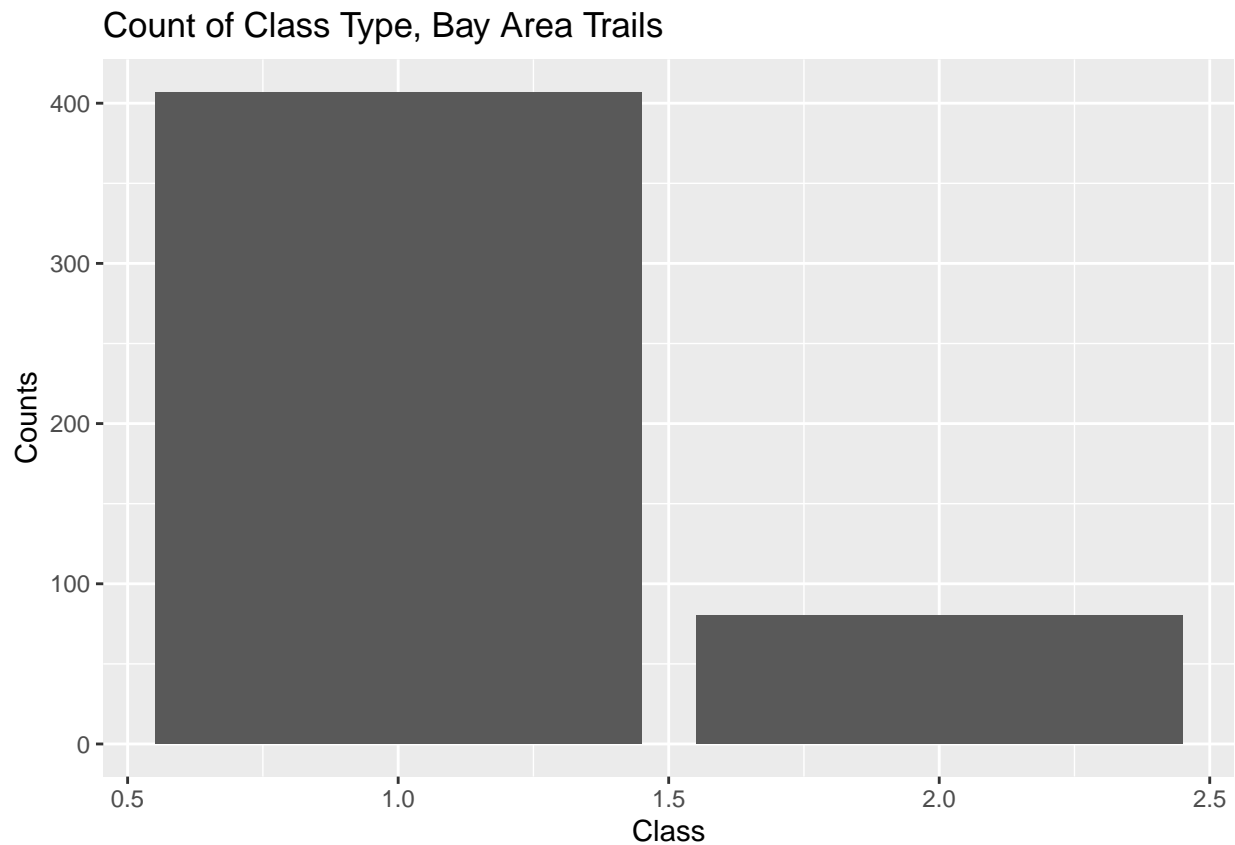
Based on the graph and output from the console, I believe that there are four distinct class variables in the `no_outliers_trails` dataset, 1, 2, 3 like displayed in the bar graph along with an NA class value since the output said 1 row containing non-finite value was removed.

The missing observations of class was likely due to NA being a non-finite value so it was omitted when creating the bar graph. The missing value, NA indicates that some trails do not have a value for 'class' meaning they were not assigned to the trail yet and it is currently undetermined for the class it should belong to.

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.

```
no_class_3_trails <- no_outliers_trails %>% filter(class !=3) %>%
  mutate(class = if_else(is.na(class),0,class))

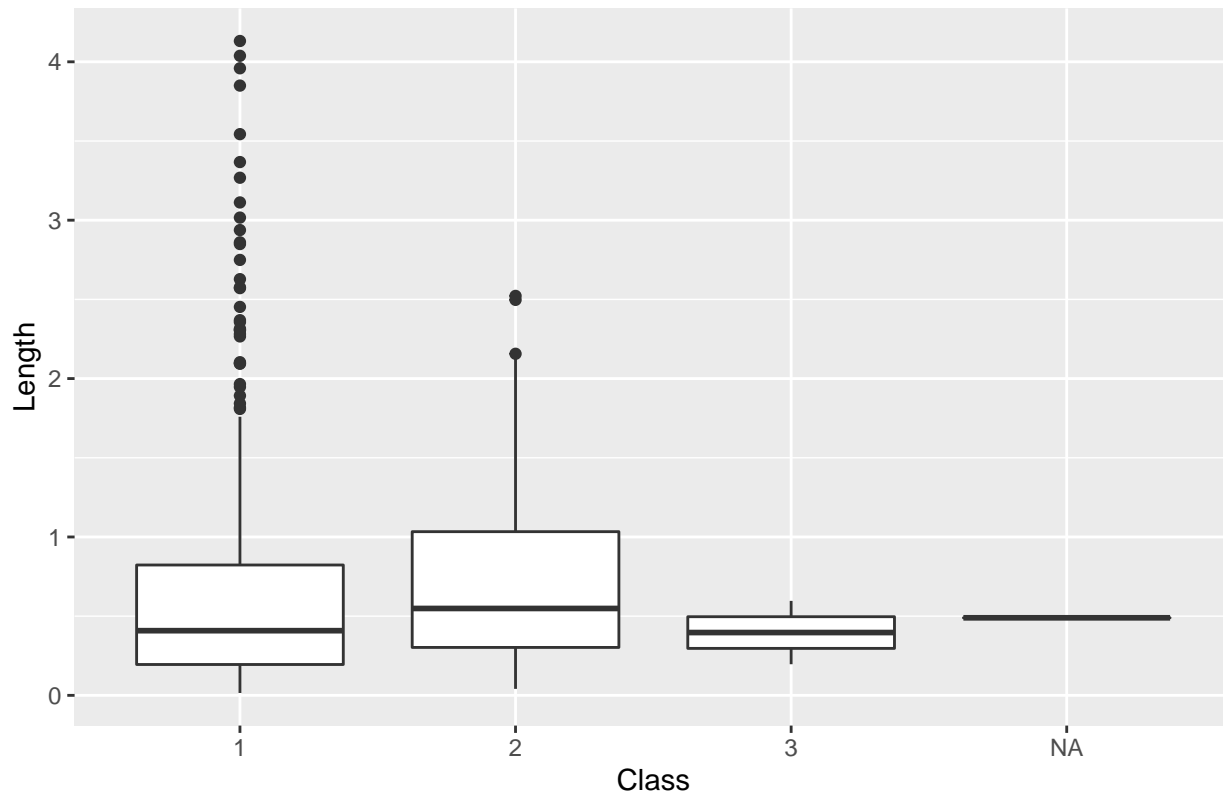
ggplot(data = no_class_3_trails, aes(x = class)) +
  geom_bar() +
  labs(x = "Class",
       y = "Counts",
       title = "Count of Class Type, Bay Area Trails")
```



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 = no_outliers_trails, mapping = aes(x = as.factor(class), y = length)) +  
  geom_boxplot() +  
  labs(x = "Class",  
       y = "Length",  
       title = "Bay Area Trail Length Boxplots per Class")
```

### Bay Area Trail Length Boxplots per Class



```
no_outliers_trails %>%
  group_by(class) %>%
  summarise(min = min(length),
            q1 = quantile(length, probs = c(0.25)),
            median = median(length),
            q3 = quantile(length, probs = c(0.75)),
            max = max(length),
            iqr = IQR(length),
            mean = mean(length),
            std_dev = sd(length)
  )
```

```
## # A tibble: 4 x 9
##   class    min    q1 median    q3    max    iqr    mean std_dev
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1  0.0148 0.195  0.408 0.823  4.13  0.628 0.682  0.740
## 2     2  0.0409 0.302  0.548 1.03  2.52  0.731 0.733  0.570
## 3     3  0.197  0.297  0.397 0.497 0.597 0.200 0.397  0.283
## 4    NA  0.490  0.490  0.490 0.490 0.490  0     0.490  NA
```

- 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.

Originally, I used the `no_class_3_trails` dataset since I thought we always used the new filtered dataset. I used the `no_outliers_trails` after looking at Exercise 11 which wants to compare the different classes (1 =

shared use bicycle and pedestrian path, 2 = bike lane, and 3 = bike route). I created box plots for the length distribution for each class variable. The class variable 1 representing the shared used bicycle and pedestrian path had the same median as class variable 3 but, a lower median than class variable 2. Class variable 1 had the largest range for 'length' where it had seem to the most right skewed since it had the most outliers for largest trail lengths. Class variable 2 which represents the bike lane has the highest median in comparison to class variable 1 and 3 which still had a wide spread but, had much fewer outliers than class variable 1. Class variable 3 which represents the bike route had the median similar to class variable 1 but, had a much smaller spread where the lengths had a much smaller spread. The summary statistics provide more evidence where class 3 (0.283) had the smallest standard deviation or spread followed by class 2 (0.570) then class 1 (0.740) and the difference between the mean and the median following the same pattern: class 3 (about 0.000) having the smallest difference followed by class 2 (0.185) and class 1 (0.274) to depict the number of possible outliers each 'length' distribution has for each class variable.

*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 line (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 sampling in the study would be independent because the trails chosen in the study should be randomly selected since there does not seem to be any other reason to choose a trail other than randomly selecting a trail. Also, since trails are being selected without replacement, it seems safe to say that there is more than 739 trails is more than 10 times less than the total trails in the Bay Area.
  - **Sample Size** The sample size for both bicycle and pedestrian path, class variable 1 and bike line, class variable 2 are greater than 30 as shown by the bar graph in Exercise 9 making the sample size for each category large enough.
  - **Independent Groups** The two groups, bicycle/pedestrian paths and bike lines are independent from one another which simply means that they are not associated with each other. From the information given about the study, there is not a correlation between these two group so we can assume the groups are independent.
13. While we have observed a small difference in the mean length in trails with bike lanes (class = 2) and trials 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.

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

In words, the null hypothesis is saying the difference between mean length of trails that are for bikes/pedestrians and the mean length of trails that have bike lanes are 0. In other words, the average length of trails between both different types of trails are about the same.

The alternative hypothesis is saying that there is difference between the mean length of trails that are for bikes/pedestrians and the mean length of trails that have bike lanes which is not 0. In other words, the average length of trails between both different types of trails are different from one another.

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.

```
?t.test # to see the help page from the function
t.test(length ~ class, data = no_class_3_trails,
       alternative = "two.sided",
       conf.level = 0.99) #less, greater, or two.sided

##
## Welch Two Sample t-test
##
## data: length by class
## t = -0.69986, df = 137.16, p-value = 0.4852
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 99 percent confidence interval:
## -0.2434412 0.1405578
## sample estimates:
## mean in group 1 mean in group 2
## 0.6819343 0.7333760
```

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 in this case is the t-score of -0.69986 which is how far from the estimated mean difference of two means or the estimated difference of means between trail lengths of bicycle/pedestrian and bike lanes in terms of the sample distribution standard deviation. In this context, it is saying were about 0.700 sample distribution standard deviations below the expected value 0 for difference of trail length between the two classes.
- Write the definition of the p-value in the context of this analysis. The p-value is the probability of observing the dataset which represents a sample given that the null hypothesis is true. In this case, our null hypothesis stated that there should not be a mean or average difference of lengths between trails for bicycles/pedestrians versus bike lanes. The p-value, 0.4852, means that there is a 48.52% chance that we observe this data if we assumed that there is no difference in the average length of both types of trails is true.
- State your conclusion in the context of this analysis. Use a significance level of  $\alpha = 0.01$ . It can be concluded that we fail to reject to the null hypothesis which states that the average trail length for bicycles/pedestrians is the same as the average trail length for bike lanes. Since the p-value is 0.4852 which is much greater than 0.05 with a two.sided t-test, we can say with 99% statistical confidence that we can not reject the null hypothesis.

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.

One way of interpreting the confidence interval is saying that with 99% confidence, the difference of average trail lengths from class 1, the bicycle/pedestrian trails from class 2, the bike lane trails lies between -0.243 and

0.141. However, we understand this as if we were able to take repeated random samples of the population, 99% of the confidence interval calculated from samples will contain the real population mean difference in average trail lengths. Since 0 is in the confidence interval, we can not determine that there is a significant difference between the mean trail lengths shown above.

*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 that all 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.*