# Doreen Nguyen

# Assignment 3 – INFO370 Data Science

# Exploratory Data Analysis

To commence, Strava is an app design for a physically active audience as it gathers tracking activity data in order to easily analyze a user's performance. It takes away the nuance of building your own analyzation tools and gives the overall and individual workout summary details without any effort on the user's part.

In particular, this Strava dataset contains 53 variables that mostly focus on characteristics of their workout such as type, average speed, distance, time and fewer details about personal information such as sex and location. Their tracked variables heavily rely on workout information in order to support their purpose of analyzing performance.

# Data Preparation/Cleaning

**Question 1: Do men tend to exercise more intensely (taking into account both distance and speed) than women?**

Because this question only asks to take into account these two variables, a simple cleaning just includes these columns where other variables that don't contribute to the definition of "intensity" are excluded.
I initially calculated a time column = distance col/speed col but it was nearly similar to values of moving time.

Initial Cleaning Steps:
1. Group data by gender: male, female
2. Keep columns of speed, distance
3. Remove all entries that contain…

- 0's in either speed, distance, moving time (null entries)
- NaN or Inf rows
- moving time that is greater than elapsed time (doesn't make sense so it must me a data entry error)
- Outliers from the equation 1.5*IQR using a function

Dataset result: 2 Total
1. Males Columns: Speed, Distance, Moving Time

2. Female Columns: Speed, Distance, Moving Time

Further Cleaning Steps (adding/looking at more variables):

While keeping the two previous datasets that look at speed and distance, I wanted to look at other variables that may contribute to the definition of "intense" to see if these averages differ from speed and distance.

- Type of Workout: Allowing all types of workouts in one dataset allow for a large range values, by splitting them up, we have smaller subsets and can compare means of more similar categories. Since ride and runs have significantly the most number of types, I chose these categories.
- Moving Time: The equation for time is distance/speed, which is moving time. I also removed moving time that is greater than elapsed time as mentioned previously because it doesn't make sense to do so.

Dataset result: 4 Total

1. Males/Ride Only Columns: Speed, Distance
2. Males/Run Only Repeat columns above
3. Females/Ride Only Repeat columns above
4. Females/Ride Only Repeat columns above

**Question 2: How do countries vary in their average lengths of workouts? What is the spread of Strava users throughout the world?**

I wanted to practice/challenge myself to visually representing the spread of data throughout the world by making a map graph as well as comparing means of countries to the US.

I used same cleaned dataset base from question 1.

Initial Cleaning Steps:

Visual Data frame:

- Remove all rows where long/lat and country name are blank
- Convert to numeric characters to plot
- Remove all columns except athlete.country, moving_time, start_latitude, start_longitude

Dataset Result: Country, moving_time, long, lat

Analyzing Data:

- Keep columns country
- Add column of counts by country
- Add column of means of workout lengths

Dataset Result: Country, count of country entries, mean of moving distance

# Statistical Modeling

**Question 1:**

We are focusing on two main variables for this test: distance and average speed. By comparing the means of certain columns of Strava dataset, we are able to use a two sided t-test to determine whether or not there is a significant statistical difference in the population means (distance and speed) in the two groups and answer if men exercise more rigorously than women.
Two-Sided t-test is for hypothesis testing where the average difference between two groups (male and female) is really significant or if its due to just random chance. Since we are given a random set of users, a two-sample t-test is the best choice to to attempt to answer this question.

**Two-Sided T-Test - Time, Distance and Time**

- Null Hypothesis: There is no difference between male and female workout intensity. (They workout with the same intensity)
- Alternative Hypothesis: There is a statistically significant difference between male and female workout intensity.

**Distance:**

Null Hypothesis: The means of male and female workout distances are the same.
Alternative Hypothesis: The means of male and female distances are statistically significantly more in distances.

**Speed:**

Null Hypothesis: The means of male and female workout speeds are the same.
Alternative Hypothesis: The means of male and female distances are statistically significantly more in speeds.

Question 2:

After observing our datasets, it is clear that the largest number of users come from the US (1666) and from the UK (1245).

Since we are comparing the average means of time work outs, I would use a t-test in order to see if there is a statistical significance between the means of the US and the UK.

Null Hypothesis: There is no difference between the lengths of workouts between the US and the UK.

Alternative Hypothesis: There is a significant statistical difference in length of workout between the US and the UK.

# Results

**Question 1.**
Results of the two-sided t-test using t.test in R using a critical value of .05:

**Two-Sample T-Tests**
**Distance:**
Results:

All Workout Types: p-value = 2.2e-16.

Rides: p-value = 1.148e-07

Runs: p-value = 7.053e-15

Since the p-value < critical value, we reject the null hypothesis and claim that the distance means of males are statistically significantly greater than females.

**Speed:**
Results:

All Workout Types: p-value = 2.2e-16.

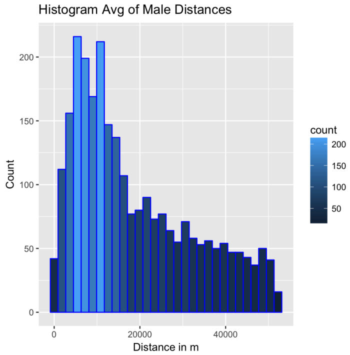
Rides: p-value = 2.2e-16

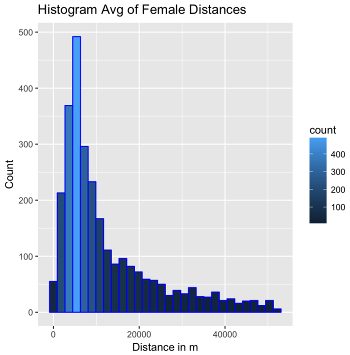
Runes: p-value = 2.2e-16

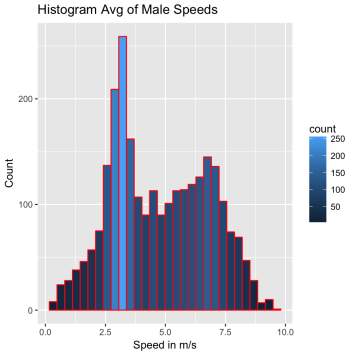
Since the p-value < critical value for all types, rides and runs, we reject the null hypothesis and claim that the speed means of males are statistically significantly greater than females.

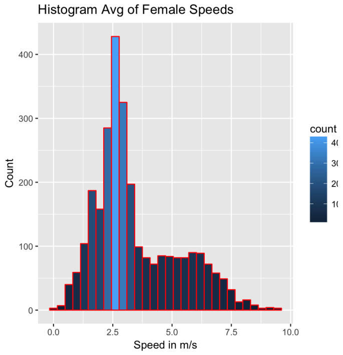
**Visual Representation**
The visual representation shows the spread and distributions of the average speeds and distances with both genders together and individually.

The histogram gives a great idea representation of the distribution, while the box plot gives a better representation of the range and highlights the quantiles. One noticeable feature is that females distances are much more right skewed than males. For speeds, males seem to have a more normal distribution than females with less of a jump going down in frequencies on the right.

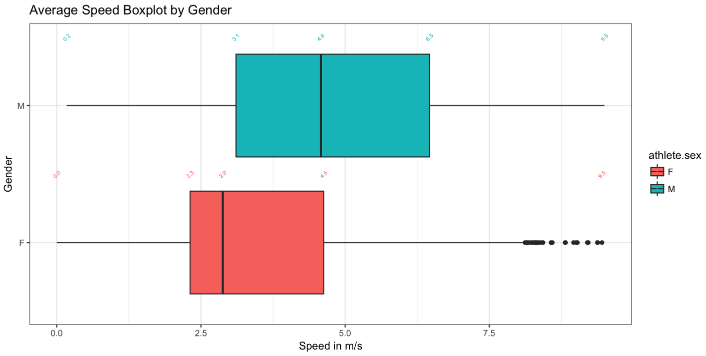


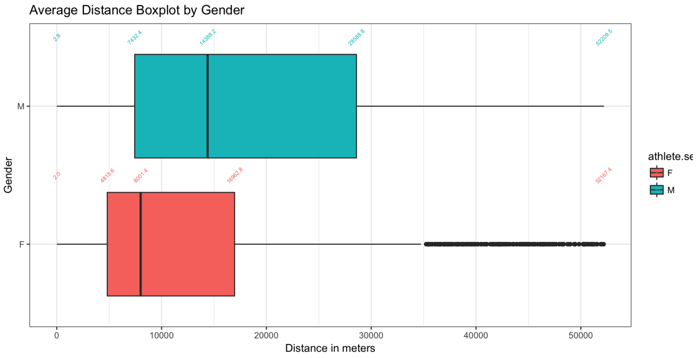






In the box plots, you can visually see that for both speeds and distances, the max/mins, means, and overall quantiles are higher for males than females.

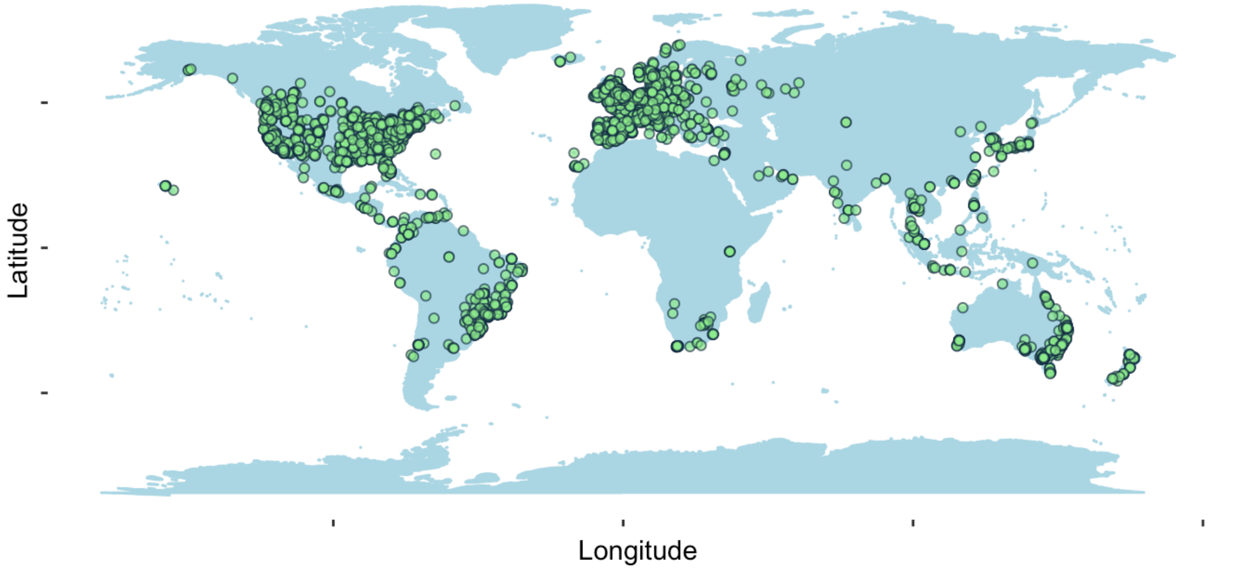




## Question 2

Calculated p-value = 0.0003372.
Since we are using a critical value of .05, we reject the null hypothesis and accept the alternative hypothesis that the length of workouts between the UK and the US are different.

Rather than doing a histogram similar to the first question, I used ggplot’s map feature in order to plot all the users coordinates to better grasp the spread of Strava users throughout the world.



# Discussion

**Question 1:**

Given that all the p-values were less than the critical values for all two-sided t-tests, I would conclude that if given similar data and similarly constructed hypothesis, males do exercise statistically significantly more intensely than females. However, there are many factors that may skew this data. While this is a randomly chosen dataset, types of workouts may drastically change the average. For example, if one gender has a higher number of rides than runs, it is most likely that the average distance and speeds will be higher as well.
Given our result, I would assume that men who are more physically stronger and bigger built would inevitably workout more intensely than women.

**Question 2:**

The attempt the group countries and find averages of how data vary resulted in only having two countries with large enough counts to compare. Nearly half of countries only had a count of 10 or less, which was insufficient for our analysis to compare Strava users worldwide. With the two groups of US and the UK, we can conclude that the two do not have statistically significant similar length workouts.

As for spread of users, you can see that the majority of the users are in heavy use in the North America, Europe, and South America. This would assume that because the app itself caters to English speaking users, there is less data obtained from other users who do not.

References:

Plotting maps: https://sarahleejane.github.io/learning/r/2014/09/21/plotting-data-points-on-maps-with-r.html