2018

Strava Data Analysis

INFO 370 ASSIGNMENT 3

DAVID LEE

Exploratory Data analysis

Research Questions:

1. Do men tend to exercise more intensely (taking into account both distance and speed) than women?
2. What 10 countries contain the most Strava users?

Strava's data provides details into the exercise sessions and metrics of users. The dataset contains information about 8, 093 users. Within these users, 3,824 are female and 4,084 are male, a portion of users did not indicate their gender. There are 53 different variables provided about users. However, not all users have data provided for every variable which will require cleaning and removing. All distances and speeds were given in meters and all times were given in seconds, both of these variables are for 1 lap. There are several outliers within the data that might provide some issues in analysis. For example, some users reported an average speed of 1,888 meters per second which might be a data error. While analyzing the data on males and

	Females	Males
Average Speed (1 lap)	4.55 m/s	5.89 m/s
Median Speed	3.86 m/s	6.49 m/s
Mode Speed	2.69 m/s	6.12 m/s
Range Speed	0.003 – 11.151	0.026 – 23.671
Average Meters/S	4 m/s	5.28 m/s
Mode Distance (1 lap)	10,306 meters	7,184.4 meters

Table 1 Descriptive statistics based on relevant variables to first research question

females (table 1), males seem to be faster than females. Within this dataset, the average speed is given for 1 lap. The distance for one lap is different for each user. I had created a variable for

speed based on the distance variable and the time variable. This variable takes the distance and time and divides them to get another speed. This speed differs from the given speed variable as this is not for 1 lap, but for a single activity.

Looking at the distributions for the average speed of one lap (fig. 1 & fig 2), the majority of both female and male distributions are within the 1 – 10 m/s range. However, the male speeds

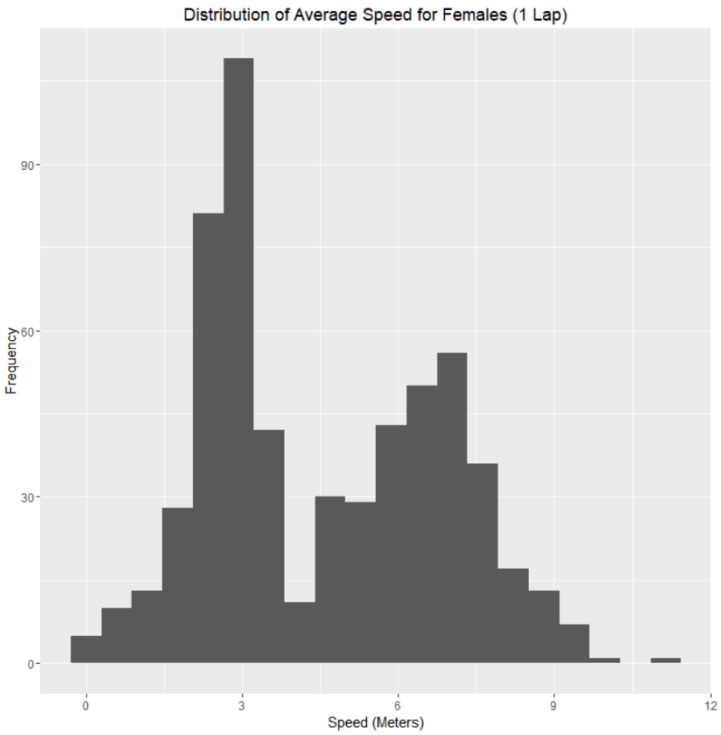

Figure 1 Histogram of female average speed in meters per second based on 1 lap

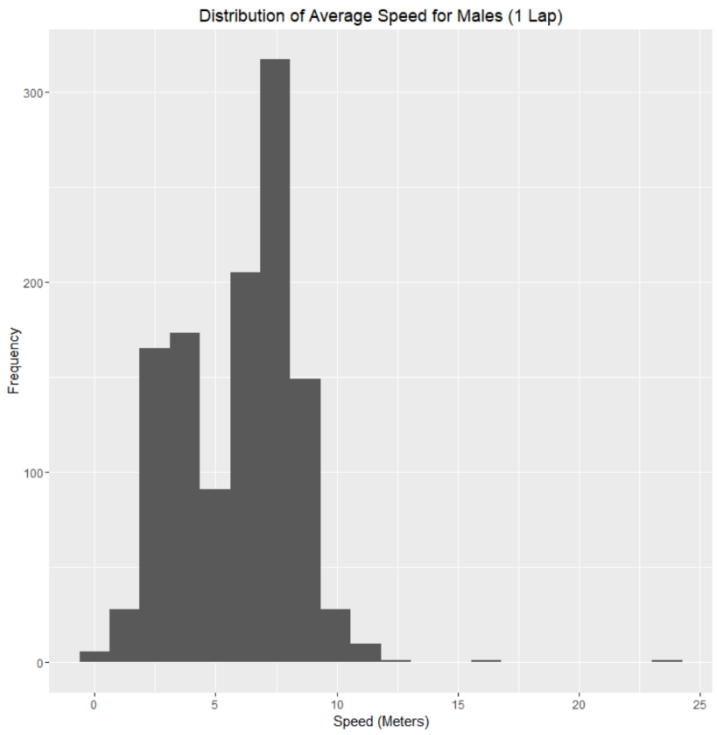

Figure 2 Histogram of male average speed in meters per second based on 1 lap

are clustered at a higher speed than the female speeds. This suggests that males are more active during their lap than females. There are some data points that are considered outliers that would skew the averages. These figures are based on data that has some outliers removed.

Looking at the distributions for average distances of one lap (fig 3 & fig 4), it is difficult to compare the distributions as they are both very similar and have large ranges. The average

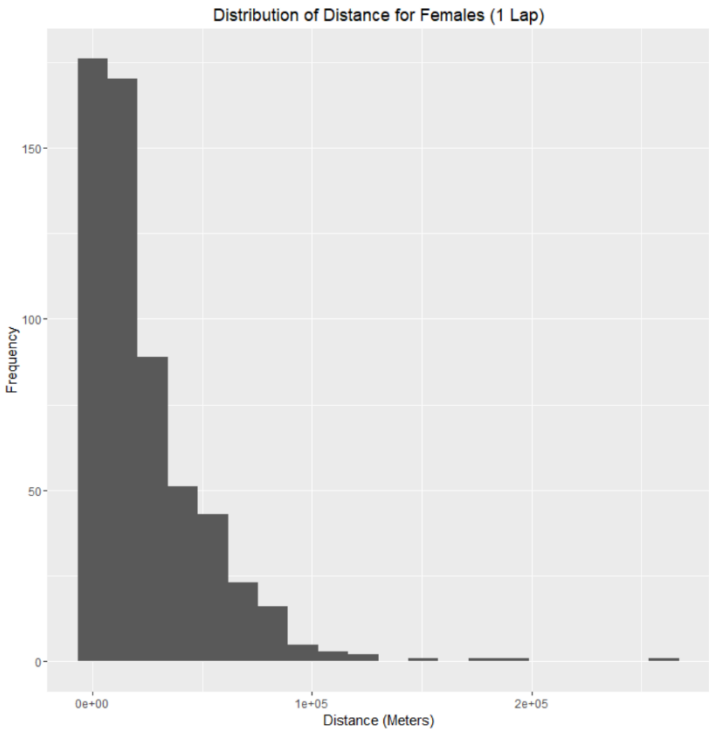

Figure 3 Histogram of female distance in meters based on 1 lap

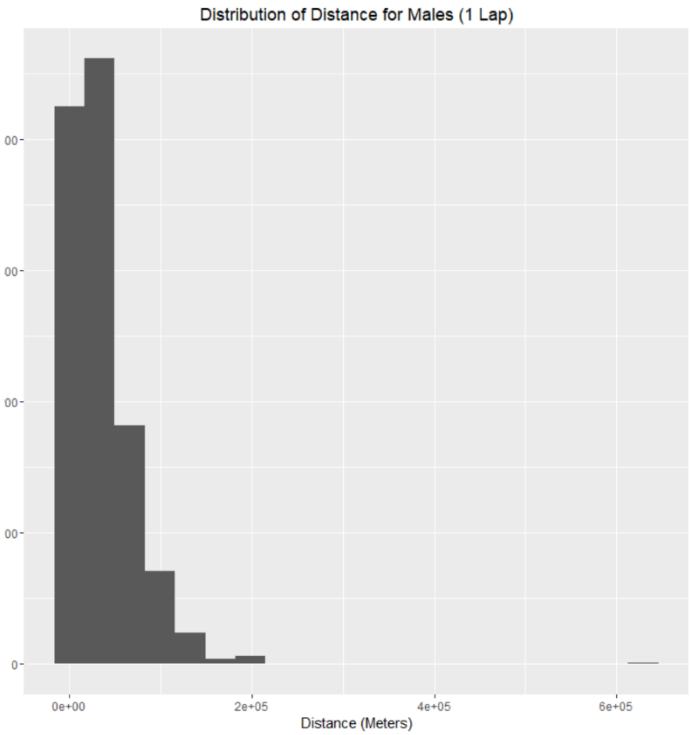

Figure 4 Histogram of male distance in meters based on 1 lap

distance for females is 24, 669.21 meters and for males is 36,011.97 meters. The inclusion of larger values that could be outliers can be causing this average distance.

Looking at the data of interest for the second research question, after filtering data, 1,792 users provided the country which they are from. This data is given in text form “United States” which will need to be changed for mapping. After mapping the data (fig. 5), I had discovered a distribution that the most users are within North America, Europe and Australia. The data mapped shows users from 67 different countries. After further examination of the frequencies (table 2), the top 3 countries with the most users are the United States, Great Britain and Australia.

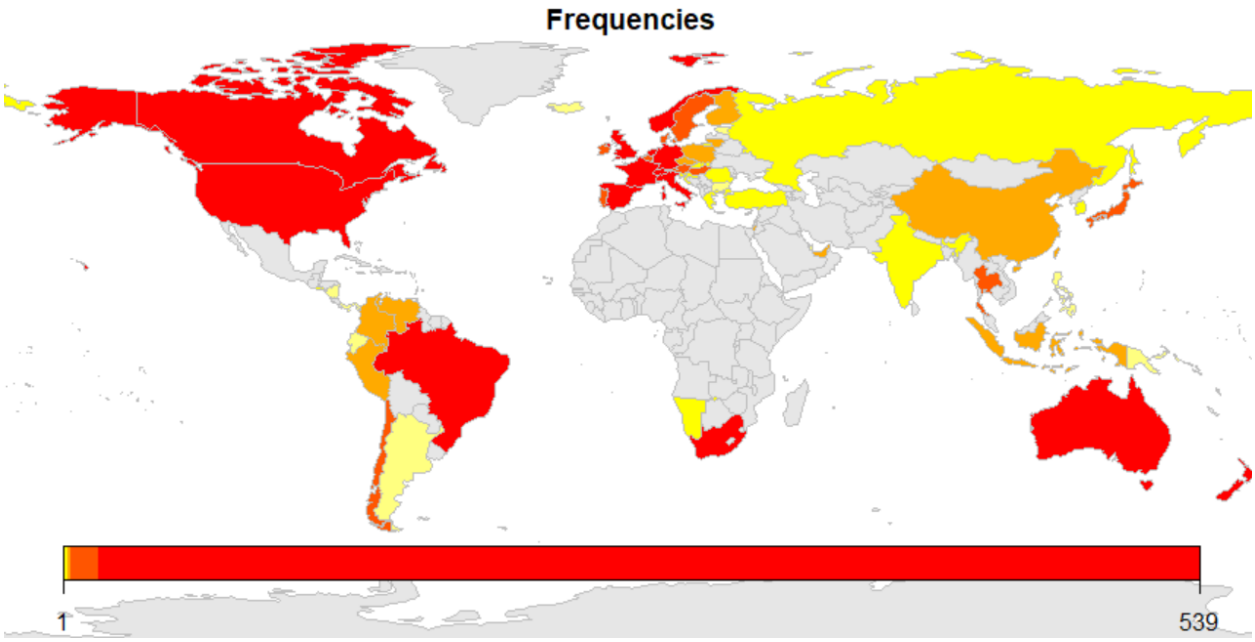

Figure 5 Frequencies of users by country

country	Frequencies
United States of America	539
United Kingdom of Great Britain and Northern Ireland	263
Australia	160
France	80
Spain	47
Netherlands	44
Canada	43
South Africa	36
Italy	28
Norway	26

Table 2 Country frequencies for users

Data Preparation

Given that this dataset contained 53 columns, not all of these variables were needed for my analysis. I had selected columns of interest and rows that did not contain missing values. First, the dataset was selected to include columns of interest such as, “athlete.sex”, “distance”, “elapsed_time”, “average_speed”, “type”, “average_heartrate”, “athlete.counrty”. These select columns were stored in a new dataframe called focus.data. I had also filtered all of the columns to remove missing data values and created a new column that is the result of distance / elapsed_time.

I had noticed that the athlete.country data had contained missing entries and were not removed in the earlier filter so I had to manually remove these rows entirely. There were 13 rows removed due to the lack of country data. I had also removed 2 rows because they contained outliers and were practically impossible values for speeds. These outliers would have caused issues in analysis later on, such as irregularly large average values.

Once the data was complete, I had stored the athlete.country data in its own list. The current state of the data was a full name of the country (“Ecuador”), however, to map this data for the frequencies, I needed to convert this into ISO3 format so they could be mapped. I used the “countrycode” package and was able to to convert the list of country names into ISO3 format and stored it into a list. This column containing 3 letter countries codes was then added into the focus.data dataframe. This column was extracted and used to create a dataframe that contained the country name and the number of athletes from each country. The new dataframe did not have a sum for each country for example, there were multiple rows for USA. This needed to be combined to get an accurate count of users for each country. I had used a group by function to

accomplish this. Afterwards, I removed any NAs from this new dataframe. This dataframe was then used to create a map object using the "rworldmap" package to be mapped.

The focus data was finally filtered into two dataframes, one for female and one for male. As a result, the users that did not provide a gender was filtered out to look at only female and male for the purposes of answering the research question.

Statistical Modeling

I had used several methods to model the data and to develop answers to the research questions. I was able to see the relationships between variables by using linear and multiple regression models.

For the linear regression models, I had looked individually at female and male data separately. I looked at distance as a predictor and speed as a response. Both variables were for one lap. This was significant because it showed the general trend of data for distance as it might play a role in how fast a user goes. Reviewing the data separately allows for a side by side comparison of the data and the linear regression line. Keeping the first research question in mind, I had then created several linear regression models that looked at sex as a predictor. This will show how sex directly plays a role in determining the values for specific variables. One linear regression model that I created was looking at speed of one lap as a response and athlete sex as a predictor. This examines how an individual's sex will determine their speed. The next linear regression model that I created was looking at distance of one lap as a response and athlete sex as a predictor. This looks more closely at how sex plays a role in an individual's distance. The last linear regression model that I created was looking at elapsed time for an activity as a response and athlete sex as a predictor. This will show how long an activity lasts and whether or not one gender spends longer than the other performing that activity.

For the multiple regression models, I modelled how athlete sex combined with other variables would influence the responses. One model looked at speed of one lap as a response and athlete sex and distance of one lap as predictors. This model demonstrates how a combination of sex and distance may make a user faster or slower depending on their gender. This relates to the question of "exercising more intensely" as one gender might have a higher influence on speed than the other, while factoring in distance. Another model I made was looking at speed of one lap as a response and athlete sex and type of activity as predictors. This model shows how a type of activity and gender may influence the speed a user has. If a speed is higher, this may suggest a more intense exercise.

Originally, I had considered using type of activity or heartrate as a variable in my analysis but it did not seem relevant. The type of activity is something that will vary depending on the individual whether male or female. Users also have different levels of activity and that could be based on the nature of the activity itself rather than based on sex. Although this is something an analysis may answer, I did not believe it to be needed. Heartrate is also something that is not really different between male and female. The differences are based on the type of activity or how healthy an individual is. This may factor in with gender to determine how intensely a user exercises, but it is also based on how much an individual exercises.

Results

After performing an initial linear regression for speed and distance on females and males (fig 6), a majority of both female and male points lie within the speeds of 1 – 4 meters. However, males have more points above 4 meters leading to a higher linear regression line. Outliers not displayed on the graphs could also lead to a difference in where the linear regression line is

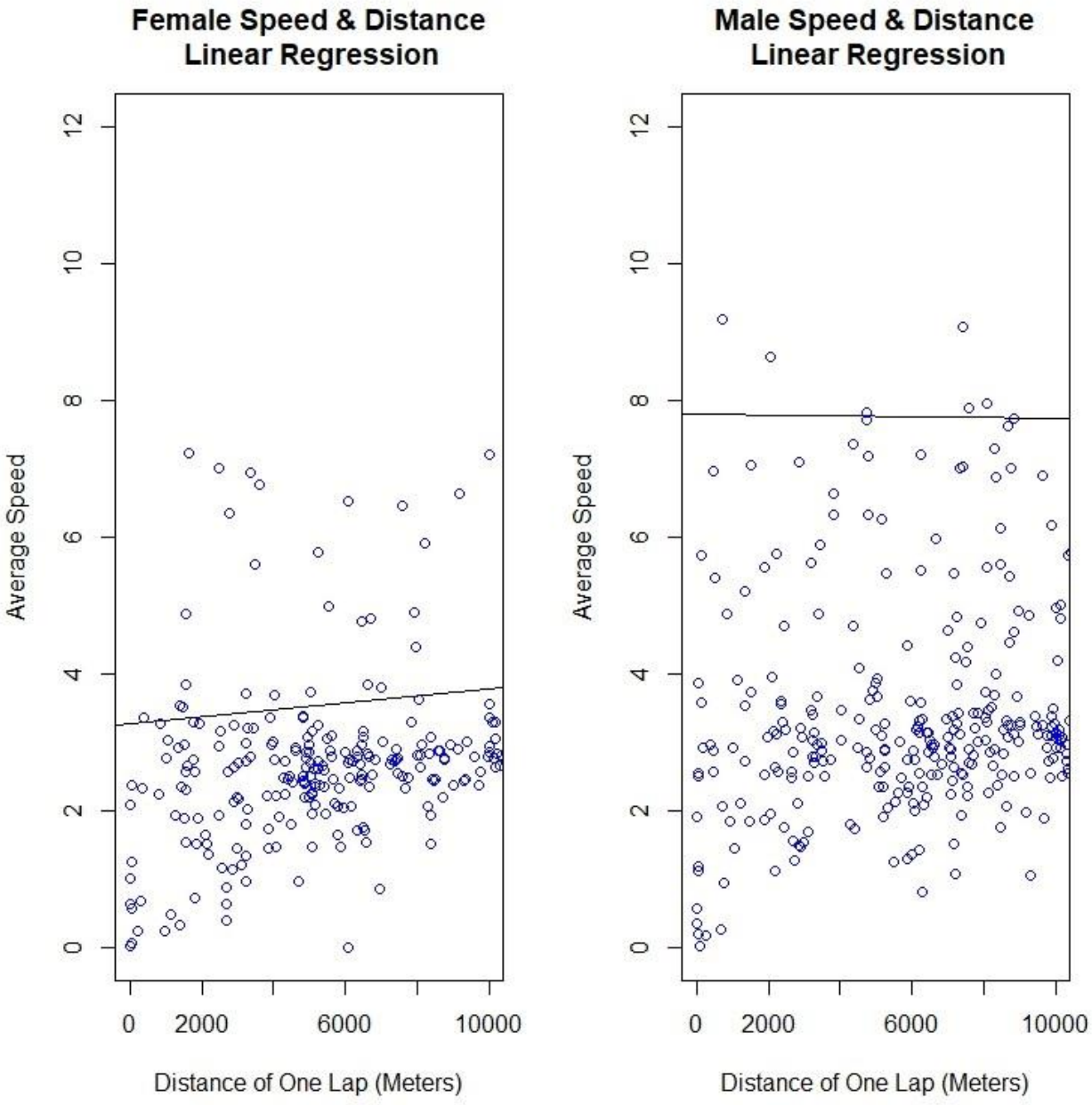

Figure 6 Linear regression model showing speed and distance for 1 lap in meters for both female and male

placed. Since the regression line is fairly flat for both genders, it suggests that distance is not a determining factor in the average speed.

Another linear regression that I did was looking at speed for one lap as response and athlete sex as a predictor. I found that the coefficients for sex (female) was -0.1495 and sex (male) was 1.1885. I had made another model looking at distance as response and athlete sex as a predictor. I found that the coefficients for sex (female) was -4,119 and sex (male) was 7,224. Another model was looking at elapsed time for an activity as response and athlete sex as a predictor. I found that the coefficients for sex (female) was 1,730 and sex (male) was 5, 092.

I also used a multiple regression model looking at speed for one lap as a response, and athlete sex and distance as predictors. I found that the coefficients for sex (female) was -1.387e-03 and sex (male) was 9.288e-01. Distance was 3.596e-05. Performing another multiple regression with speed as response and athlete sex and activity type as predictors, I found that sex (female) was -0.2410 and sex (male) was 0.3242. Each activity was given a coefficient that was higher than 1.0 or less than -1.0.

Through these models, testing athlete sex as a predictor will help determine the answer to the research question of if men exercise more intensely than women. These models demonstrate how gender plays a role in the values of other variables and in what magnitude.

Discussion

Based on the results of the exploratory data analysis and statistical modelling, the answer to the first research question "Do men tend to exercise more intensely (taking into account both distance and speed) than women?" is that men exercise more intensely than women. For all descriptive statistics other than distance for one lap, men had higher statistics than women (table 1). After reviewing the distributions of average speed for 1 lap, both distributions were relatively similar but the male distribution was clustered around a slightly higher area (fig. 1 & 2). It is difficult to determine which sex had a longer distance as the distributions included outliers which affected the averages given in the descriptive statistics. The results of the statistical modelling show the coefficients for male in every model was higher than females. This shows that the male gender variable had a greater effect in increasing the response variable. This means including a male coefficient made speed, distance and elapsed time larger than adding a female coefficient. As a result, the male coefficient made the user faster, exercised a longer distance, and for a longer time. This shows that men do exercise more intensely than women.

Based on the exploratory data analysis and distributions of countries, the answer to the second research question "What 10 countries contain the most Strava users?", is the USA, Britain, Australia, France, Spain, Netherlands, Canada, South Africa, Italy and Norway. These distributions are also described by (fig. 5 & table 2). The number of users from these countries are the highest out of the 67 countries looked at. There is a heavy skew with the USA, Britain and Australia as the highest. The large presence within the countries might be influenced by Strava entities or advertising.

There are some limitations to my analysis and my approach. Not all conclusions are certain as there are many different statistical models and factors that influence the results. Some

models might be better and more accurate at showing relationships than what I have used here because the data had been cleaned and filtered, which meant that not all of the data was used. This could have had an impact on the results of the analysis or on the country data. The data provided was only a subset of the entire user-base of Strava. If I had more data the answers to the research questions here could be very different.

This data might be useful to Strava, as it gives insight into their customers. In this analysis, it was found that men exercised more intensely than women. This might change the way Strava interacts with users by added more features to its platform to record metrics or to compare data with others. Introducing a method to compare with other users is a way to motivate users to exercise more. This data might also be used to target specific countries to attract more users. As seen by the map of users (fig. 5), Strava does not need to focus as much advertising effort in countries with already high user numbers. Extra advertising costs could be used on countries with a smaller user-base to increase users and revenue in an area.