Leon Long
INFO 370
Assignment 3

Strava Data Analysis Report

Questions of Interests

Strava is a software that tracks athletic activity via satellite navigation and connects athletes through its social network platform. For this project, we have access to the Strava API, which gives us various user and athletic activity related data. Based on the Strava activity data, we have two questions of interest to be answered through this analysis:

1. Do men tend to exercise more intensely (taking into account both distance and speed) than women?
2. Do athletes tend to exercise more intensely when they are in different countries from their origin countries.

Exploratory Data Analysis

Prior to further data wrangling and analysis, we need to get an overview of our dataset by exploring dimensions and number of observations included: there are **8093 observations** and **53 variables**. By taking a closer look at the list of variables and putting them into two main categories--**user-related profile data** and **activity-related athletic data**--we are able to better navigate ourselves and find the most related ones from other trivial and distracting variables.

Given the questions of interests, I selected **12** mostly-related variables to further explore the datasets: *"athlete.country", "athlete.sex", "average_heartrate", "max_heartrate", "average_speed", "max_speed", "distance", "moving_time", "elapsed_time", "kilojoules", "location_country", and "total_elevation_gain".* **(Figure 1)**

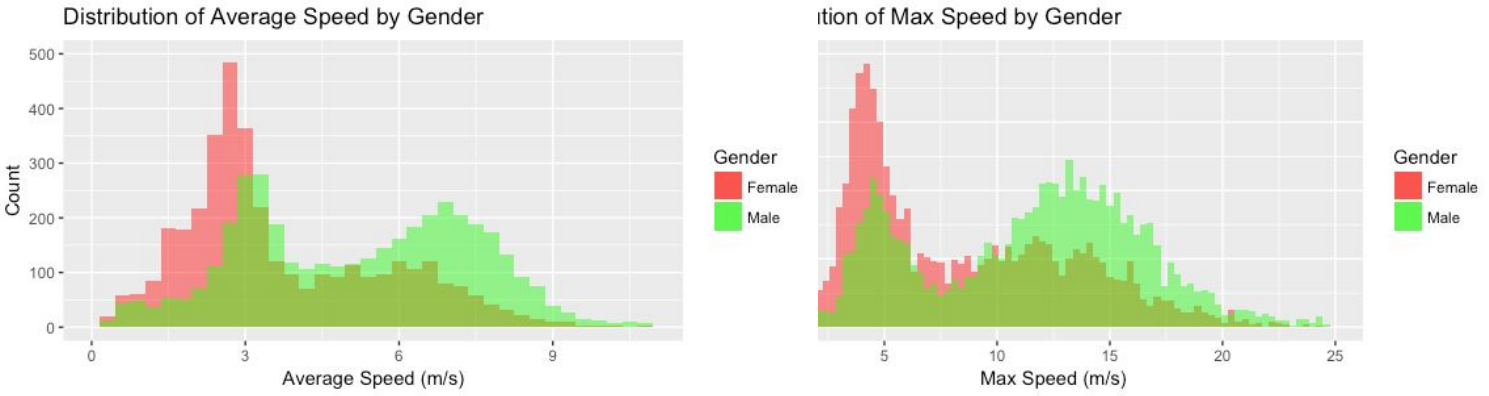
	athlete.country	athlete.sex	average_heartrate	average_speed	distance	elapsed_time	kilojoules	location_country	max_heartrate	max_speed	moving_time	total_elevation_gain
1	Ecuador	F	151.2	3.022	21580.0	8697	641.4	Ecuador	175	10.9	7140	476.3
2	The Netherlands	M	NA	4.062	19092.8	5370	619.5	Spain	NA	12.3	4700	456.9
3	United States	M	NA	5.397	23023.4	4409	658.7	United States	NA	14.4	4266	292.2
4	United Kingdom	F	NA	6.314	101702.0	20117	1905.3	United Kingdom	NA	17.3	16108	1273.3
5	United States	F	NA	1.629	2739.8	1682	NA	United States	NA	4.6	1682	166.6
6	United Kingdom	M	NA	5.868	13056.7	2310	239.4	United Kingdom	NA	12.4	2225	81.4
7	Lithuania	M	135.9	4.675	24438.8	5342	763.1	Lithuania	182	12.8	5228	303.0
8	United States	F	NA	8.173	9447.6	1156	NA	United States	NA	5.0	1156	36.8
9	United Kingdom	F	NA	0.000	0.0	2	0.0	United Kingdom	NA	0.0	2	0.0
10	United Kingdom	F	NA	3.220	4814.0	1543	NA	United Kingdom	NA	7.4	1495	48.8
11	United Kingdom	M	143.0	6.685	33632.8	5722	1114.6	United Kingdom	174	13.9	5031	301.0
12	United Kingdom	F	NA	3.084	15818.6	5286	NA	United Kingdom	NA	14.6	5130	252.9
13	Canada	F	NA	2.573	16014.5	6367	NA	Canada	NA	8.7	6223	533.3
14	United Kingdom	F	NA	2.376	3157.1	1386	NA	United Kingdom	NA	15.6	1329	37.1
15	Australia	M	NA	5.866	10628.8	1812	204.7	Australia	NA	10.5	1812	56.4

Figure 1

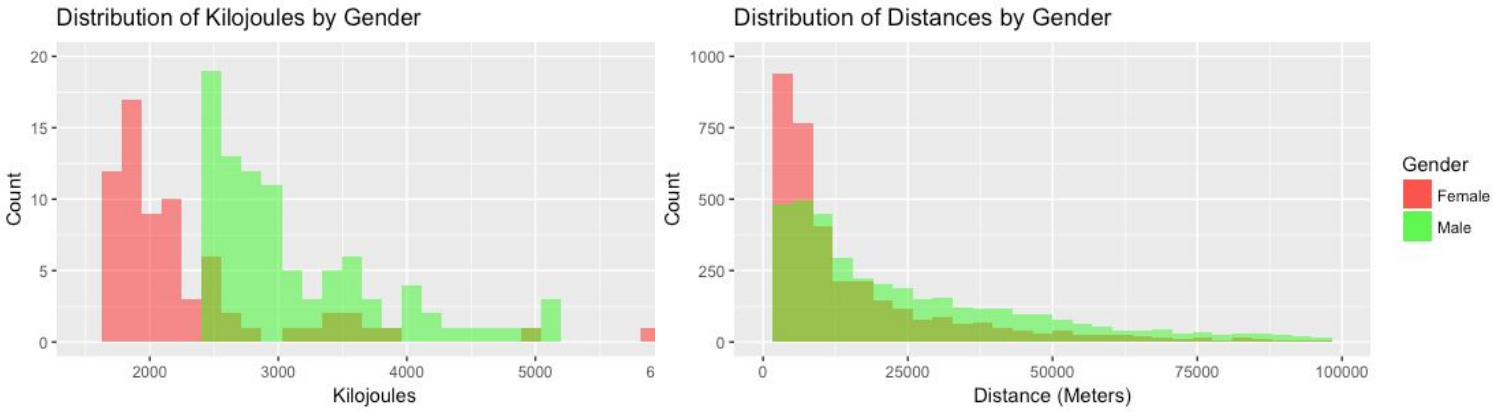
Figure 1 is a snippet of **15** rows of data. As it shows, there are multiple "NA" values in the dataset, most of which are in ***"average_heartrate", "max_heartrate",* and *"kilojoules"***. This makes sense since most of the mobile devices have no built-in heart rate tracking feature. In order to figure out how would the missing values affect our further analysis, I found that there were only **1120 /8093** rows of data with all variables in place, and by removing ***"average_heartrate", "max_heartrate"***, we got 3824 complete rows of observations.

To answer the questions of interests, I specifically assessed the key variables--there are 4084 male athletes and 3824 females athletes. These relatively even population sizes ensure the accuracy of later analysis. In contrast, only **856 /8093** activities occurred in different countries from the athletes' origin countries.

Besides, we also plot out the distributions of key variables across genders:



Based the the above two overall distribution, we can see that across all activities, male athletes tend to have large **average and maximum speeds** compared to female athletes.



The Kilojoules graph shows a clear trend that the male athletes are further distributed along the x-axis (**Kilojoules**) and the Distance graph shows a slight shift between the male and female distributions as well. Solely based on these four distributions of key variables, we can see a potential trend that males tend to exercise more intense than women.

Data Preparation

To prepare for further statistical modeling, I cleaned and transformed the data so it best serves our later analysis and choice of statistical modeling method.

Since there are only **856 /8093** fully-complete observations, and heart rate related data are missing the most, I decided to remove ***"average_heartrate", and "max_heartrate"*** from our data frame. Furthermore, since ***"kilojoules"*** seemed to be a very fair indicator of the activity intensity, I decided to include it in our analysis even though we need to give up some portion of our dataset. I also made sure that other missing-value and duplicate rows were removed from our clean dataset.

When it comes to intensity, I realized that ***moving_time*** and ***elapsed_time*** themselves, among the other variables, seemed to be a little irrelevant. I decided to create another aggregated variable named ***moving_ratio*** that, which is defined as ***moving_time / elapsed time***, which indicated how active the athlete is in each activity session. Moreover, I created another variable ***same_country***, which indicates whether the activity occurred in the same country from the athlete's origin country.

On top of that, I also transformed two categorical variables into factors as ***gender.code*** and ***same_country.code*** for future convenience when it comes to statistical modeling. Here is the clean data after preparation (**Figure 2**):

	athlete.country	athlete.sex	average_speed	distance	elapsed_time	kilojoules	location_country	max_speed	moving_time	total_elevation_gain	moving_ratio	same_country	gender.code	same_country.code
1	Ecuador	F	3.022	21580.0	8697	641.4	Ecuador	10.9	7140	476.3	0.8209727	TRUE	F	TRUE
2	The Netherlands	M	4.062	19092.8	5370	619.5	Spain	12.3	4700	456.9	0.8752328	FALSE	M	FALSE
3	United States	M	5.397	23023.4	4409	658.7	United States	14.4	4266	292.2	0.9675663	TRUE	M	TRUE
4	United Kingdom	F	6.314	101702.0	20117	1905.3	United Kingdom	17.3	16108	1273.3	0.8007158	TRUE	F	TRUE
6	United Kingdom	M	5.868	13056.7	2310	239.4	United Kingdom	12.4	2225	81.4	0.9632035	TRUE	M	TRUE
7	Lithuania	M	4.675	24438.8	5342	763.1	Lithuania	12.8	5228	303.0	0.9786597	TRUE	M	TRUE
9	United Kingdom	F	0.000	0.0	2	0.0	United Kingdom	0.0	2	0.0	1.0000000	TRUE	F	TRUE
11	United Kingdom	M	6.685	33632.8	5722	1114.6	United Kingdom	13.9	5031	301.0	0.8792380	TRUE	M	TRUE
15	Australia	M	5.866	10628.8	1812	204.7	Australia	10.5	1812	56.4	1.0000000	TRUE	M	TRUE
18	United Kingdom	M	6.124	28765.1	4855	450.6	United Kingdom	17.1	4697	241.8	0.9674562	TRUE	M	TRUE
19	Australia	M	9.249	69559.8	7946	1547.7	Australia	16.8	7521	383.8	0.9465140	TRUE	M	TRUE
22	United States	F	6.750	43609.3	6680	904.8	United States	18.1	6461	556.0	0.9672156	TRUE	F	TRUE
23	Germany	F	4.481	25781.1	6639	863.2	Germany	14.6	5754	279.4	0.8666968	TRUE	F	TRUE
25	United States	M	7.180	48415.5	52457	1532.7	United States	9.3	6743	103.8	0.1285434	TRUE	M	TRUE
27	United Kingdom	M	7.531	34499.3	4793	629.2	United Kingdom	15.4	4581	258.9	0.9557688	TRUE	M	TRUE

Figure 2

Statistical Modeling

The major question that we want to answer involves how intense each athletic activity is, which brings us to defining the word "intensity". As we discussed previously, we consider more than one variables as our indicators of intensity including ***average and max speed, distance, kilojoules, moving_ratio, and etc,*** which means that we have multiple dependant variables corresponding to the independent variable--***athelete.sex***. In this situation, after some researches, I decided to use **MANOVA** as our statistical modeling method.

MANOVA is a procedure for comparing multivariate sample means. As a multivariate procedure, it is used when there are two or more dependent variables, and is typically followed by significance tests involving individual dependent variables separately. (Warne, R. T. 2014, Stevens, J. P. 2002)

In an MANOVA, we examine for statistical differences on multiple continuous dependent variables, bundled together into a weighted linear combination or composite variable, by an independent grouping variable. In this case, the multiple continuous dependent variables are those potential indicators of activity intensity. We were examining whether the ***athelete.sex*** determines the activity intensity, which is represented by those dependent variables.

However, since MANOVA test only yields whether there's a difference between the two gender groups, to get the direction of correlation, in other words, do men tend to exercise more intensely than women, I would fit linear regression model on each dependent variable.

Results

As we performed the MANOVA tests on ***athelete.sex*** and ***same_country*** for Question 1 and Question 2, we found that both **p-values are very significant**, which mean that **there are statistical differences** on **activity intensity** across **two gender groups** and **whether the activity occurred in the same country with the athlete's origin country**.

```
> summary(same_country.manova)
                   Df   Pillai approx F num Df den Df    Pr(>F)
same_country.code   1 0.021663   13.986      6   3790 8.377e-16 ***
Residuals        3795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(gender.manova)
             Df   Pillai approx F num Df den Df    Pr(>F)
gender.code   2 0.065244   21.301     12   7580 < 2.2e-16 ***
Residuals  3794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

I further evaluated each indicator of activity intensity to see how strong each of them is related to ***athelete.sex*** and ***same_country.***

```
> summary.aov(gender.manova)
 Response average_speed :
              Df  Sum Sq Mean Sq F value Pr(>F)
gender.code    2    2362  1180.8  1.2434 0.2885
Residuals   3794 3603170   949.7

 Response distance :
              Df     Sum Sq    Mean Sq F value   Pr(>F)
gender.code    2 4.4815e+10 2.2407e+10  19.038 5.93e-09 ***
Residuals   3794 4.4654e+12 1.1770e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response kilojoules :
              Df     Sum Sq  Mean Sq F value    Pr(>F)
gender.code    2   67777060 33888530  54.214 < 2.2e-16 ***
Residuals   3794 2371567493   625084
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response max_speed :
              Df Sum Sq Mean Sq F value    Pr(>F)
gender.code    2   1565  782.65   31.74 2.136e-14 ***
Residuals   3794  93553   24.66
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response moving_ratio :
              Df Sum Sq Mean Sq F value    Pr(>F)
gender.code    2   0.67  0.3352   8.156 0.0002921 ***
Residuals   3794 155.93  0.0411
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response total_elevation_gain :
              Df     Sum Sq Mean Sq F value  Pr(>F)
gender.code    2    4148731 2074365  4.4568 0.01166 *
Residuals   3794 1765862812  465436
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary.aov(same_country.manova)
 Response average_speed :
                   Df  Sum Sq Mean Sq F value Pr(>F)
same_country.code   1       0    0.13   1e-04 0.9906
Residuals        3795 3605532  950.07

 Response distance :
                   Df     Sum Sq    Mean Sq F value    Pr(>F)
same_country.code   1 7.6935e+10 7.6935e+10  65.858 6.482e-16 ***
Residuals        3795 4.4333e+12 1.1682e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response kilojoules :
                   Df     Sum Sq  Mean Sq F value    Pr(>F)
same_country.code   1   51630119 51630119   82.06 < 2.2e-16 ***
Residuals        3795 2387714433   629174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response max_speed :
                   Df Sum Sq Mean Sq F value  Pr(>F)
same_country.code   1     82  81.521  3.2553 0.07127 .
Residuals        3795  95037  25.043
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response moving_ratio :
                   Df  Sum Sq  Mean Sq F value Pr(>F)
same_country.code   1   0.002 0.001834  0.0444  0.833
Residuals        3795 156.599 0.041265

 Response total_elevation_gain :
                   Df     Sum Sq  Mean Sq F value    Pr(>F)
same_country.code   1   13995047 13995047  30.245 4.058e-08 ***
Residuals        3795 1756016496   462718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

5 observations deleted due to missingness
>
```

Then, I only looked at the variables with **'***' < 0.001 p-value** and fit linear regression model on each of them with **athelete.sex** and **same_country**.

```
> kilojoules.same_city.fit

Call:
lm(formula = kilojoules ~ same_country.code, data = clean_data)

Coefficients:
          (Intercept)  same_country.codeTRUE
               1058.4                 -341.6

> distance.same_city.fit

Call:
lm(formula = distance ~ same_country.code, data = clean_data)

Coefficients:
          (Intercept)  same_country.codeTRUE
                45742                 -13182

> max_speed.same_city.fit

Call:
lm(formula = max_speed ~ same_country.code, data = clean_data)

Coefficients:
          (Intercept)  same_country.codeTRUE
              13.8776                -0.4185

> moving_ratio.same_city.fit

Call:
lm(formula = total_elevation_gain ~ same_country.code, data = clean_data)

Coefficients:
          (Intercept)  same_country.codeTRUE
                502.1                 -177.9
```

```
> kilojoules.gender.fit

Call:
lm(formula = kilojoules ~ gender.code, data = clean_data)

Coefficients:
 (Intercept)  gender.codeF  gender.codeM
       748.9        -176.8         108.6

> distance.gender.fit

Call:
lm(formula = distance ~ gender.code, data = clean_data)

Coefficients:
 (Intercept)  gender.codeF  gender.codeM
     30186.8        -725.1        6599.9

> max_speed.gender.fit

Call:
lm(formula = max_speed ~ gender.code, data = clean_data)

Coefficients:
 (Intercept)  gender.codeF  gender.codeM
     13.3333       -0.7438        0.6446

> moving_ratio.gender.fit

Call:
lm(formula = moving_ratio ~ gender.code, data = clean_data)

Coefficients:
 (Intercept)  gender.codeF  gender.codeM
     0.80961       0.02005       0.04751

>
```

We can see that for all intensity indicators, **same_country.codeTrue** is always negative, which means that if the athlete is in the same country, he or she tends to exercise less intensively. For all intensity indicators, in general, **gender.codeM** is positive and **gender.codeF** is negative, which means that males tend to exercise more intensely than females.

Discussion

Based on our statistical modeling and analysis, we can conclude that men do tend to exercise more intensely than women and athletes do tend to exercise more intensely when they are in different countries from their origin countries. This results align with the assumptions we made based on the initial distributions. I think the statistical model is very accurate as we performed other tests. One of the limitation I was concerned with was that there are different types of activities, distance, elevation_gain, speed, vary from activity to activity and it's very hard to define and measure "intensity". However, since the number of males and females are very close, by including those related variables, we can just neglect the activity types, and that's why we employed MANOVA. This finding may be interesting to sociologist, nutritionist, and athletes in general to further explore the underlying principles and implications.

Works Cited

Warne, R. T. (2014). "A primer on multivariate analysis of variance (MANOVA) for behavioral scientists". Practical Assessment, Research & Evaluation. 19 (17): 1–10.

Stevens, J. P. (2002). Applied multivariate statistics for the social sciences. Mahwah, NJ: Lawrence Erblaum.