Strava Data Science Report

Kishore Vasan

11/27/2017

The Social Network for Athletes

Strava is a website and mobile app used to track athletic activity via satellite navigation. The service uses the GPS functionality of mobile phones or other devices such as Garmin navigator devices to record position and time data during athletic activities.
The data was collected randomly using the Strava API. The dataset contains 8093 rows with 53 paramters about athletes, including but not limited to - **Achievement Count, Athlete Country, Athlete Sex, Average Speed, Average Heartrate, Maximum Heartrate, Distance, Elapsed Time, Suffer Score, Total Elevation Gain and Workout Type(Activity).**

The first part of this report will focus on answering the question - **Do men exercise more intensly than women?** and the second part of the report will hope to find out **if the maximum heartrate acheived by athletes is same across different activities** and then hope **to predict the type of activity using different paramters.**
Note: In this report we will also be using the **Kudos Count** parameter. Kudos is similar to a Facebook Like that can be given by fellow athletes to other athletes.

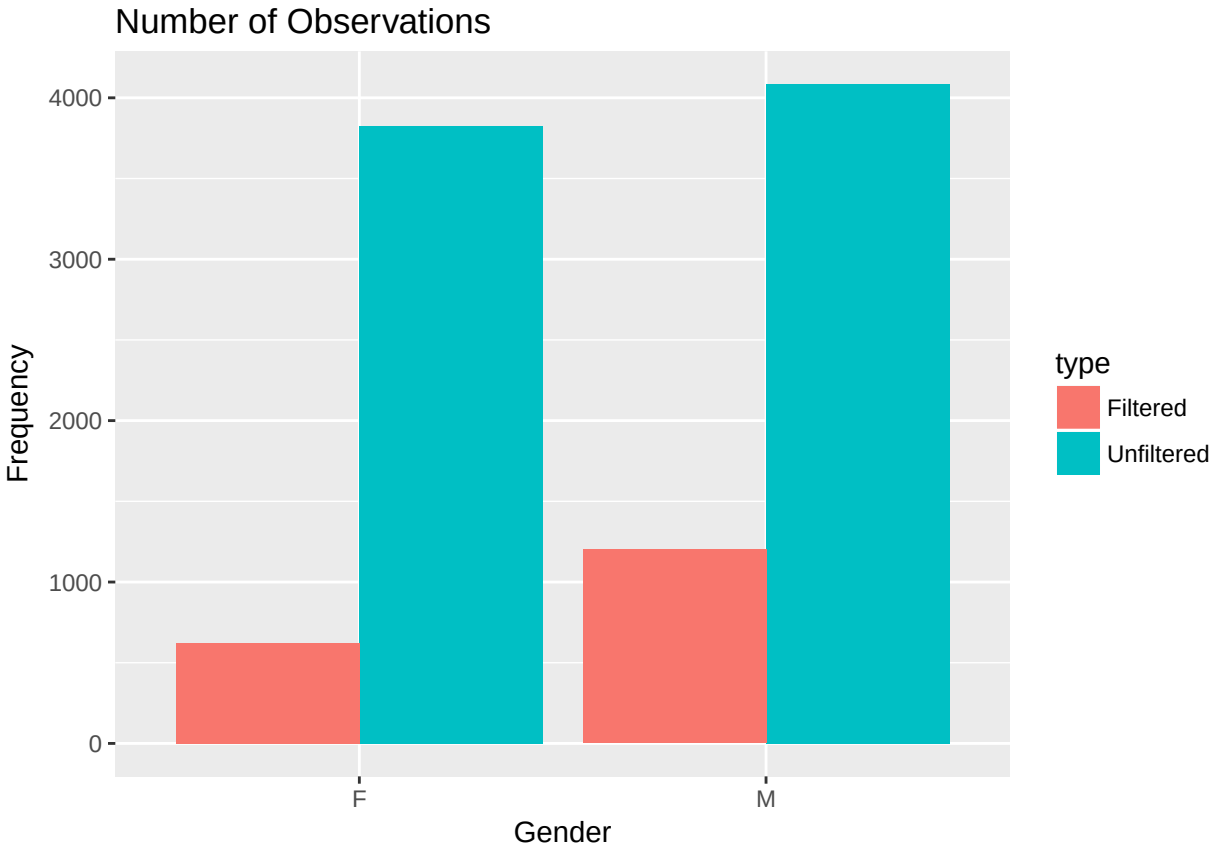
Data Preparation

- Filter rows without a gender value and convert the Male values to 0 and female values to 1. After this we end up with a total dataset row count of 7908.

- Filter rows that dont have a maximum heartrate and also rows with maximum heartrate as 0 or 1(which are obviously false). Here we end up with a filtered dataset row count of 1823

Exploratory Data Analysis

One of the initial explorations that is esssential is to see the representation of different groups in the dataset.



As you can see from the above bar plot, the dataset is pretty evenly split between Male and Female observation in the original data collection, thus removing the problem of bias with data representation. But looking at the **gender representation in the maximum heartrate filtered dataset calls for concern when doing analysis.**
Now that we have a basic idea of the dataset, we can now move on to answer specific questions about the Strava dataset.

Do men tend to exercise more intensely than women?

This is a very subjective question that can be answered to a certain level using Data Science. We can **use a Linear model to see how the achievement count(which ideally portrays workout intensity) relates with average speed, distance and athlete sex.** We can then look at the coefficients corresponding to the athlete sex parameter and formulate an opinion about how the achievement count(or intensity level) correlates with gender.
Note: We will not be able to use maximum heartrate value in the linear model due to bias in the sample size as shown above.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9822818	0.1519143	6.4660250	0.0000000
distance	0.0001228	0.0000031	39.9199427	0.0000000
average_speed	-0.0003069	0.0038496	-0.0797191	0.9364627
athlete.sex	0.4339803	0.1834638	2.3654823	0.0180306

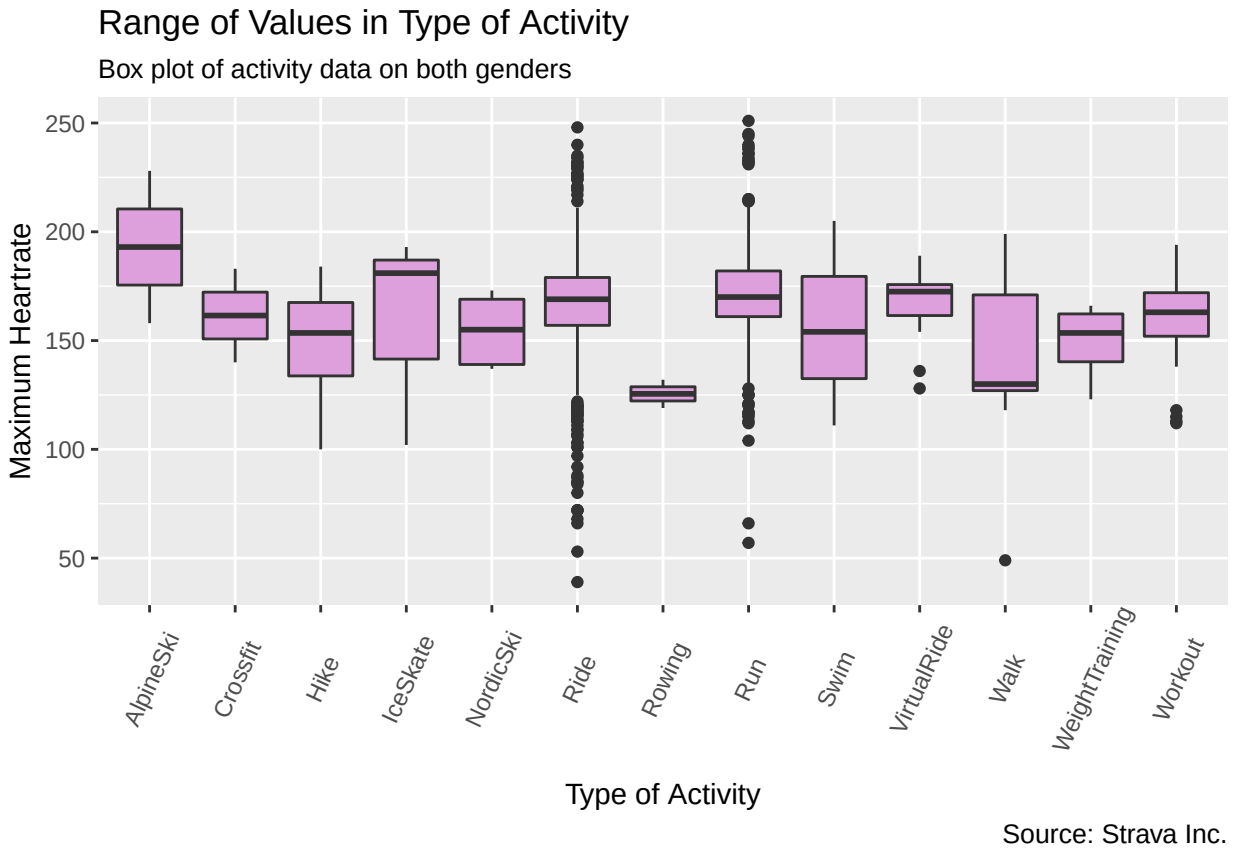
As you can see from the coefficient value of athlete.sex, though small there is an increase in achievement count

of **0.4339802595** when athlete sex is 1(or female) or the acheivement count of a female is 0.4339803 higher than that of a male when all the other parameters are the same. We also get a correlation value(R squared value) of **0.1705972**. This shows that the variables used in the linear model have a positive correlation with the Achievement Count variable.
Hence, we can say that women exercise more intensly than men.

Is the maximum heartrate achieved independent of the type of activity?

Motivation:

Most people workout for heart reasons- in particular to increase the heart capacity(or maximum heartrate). But **is the maximum heartrate achieved same across all activities?**
To answer this question, we can use either maximum heartrate or average heartrate values from the dataset. The problem with average heartrate is that it gets affected by the duration of workout. There could be people who work out for a long time less intensly compared to some who work out for a short time but intensly.
The goal here is to **see if the average maximum heartrate people achieve doing different activities using Strava the same at 0.05 level of significance?**



As you can see from the box plot, there are quite a few outliers in the Ride and Run activity types. There is also a difference in the average maximum heartrate among each activities and quite a difference in means within the groups(activities). To check if the grand mean(mean of means) different for each activity. **To do this ideally we would use Analysis of Variance(ANOVA) test.**

But first, lets look at the sample size across each activity.

Type	Freq
AlpineSki	2
Crossfit	2
Hike	4
IceSkate	3
NordicSki	5
Ride	1200
Rowing	2
Run	550
Swim	3
VirtualRide	18
Walk	9
WeightTraining	4
Workout	21

Though there is a huge difference in the sample size between groups. **ANOVA will still be able to deal with unequal sample sizes across groups to a certain level.**
The ANOVA test, however requires that variances be equal across groups. The Bartlett test can be used to verify this assumption.

Bartlett's Test:

Ho: The population variance is the same across all activities
Ha: The population variance is not the across all activities

```
##
##  Bartlett test of homogeneity of variances
##
## data:  max_heartrate by type
## Bartlett's K-squared = 29.285, df = 12, p-value = 0.003573
```

Since the p value is <0.05, we reject the Null hypothesis and hence with a 95% confidence **we can conclude that the variance among groups is not the same.**

We cannot assume normality of the population each group was sampled from and as concluded from the test above, we also cannot conclude that the variance is equal among the groups. Hence, **we cannot use one-way ANOVA test but instead we will use Kruskal-Wallis rank sum test**. This is a non-parametric test that is less sensitive to the normality assumption.

Kruskal-Wallis Test:

Ho: The mean of each group is same
Ha: The mean of each group is not the same

```
##
##  Kruskal-Wallis rank sum test
##
## data:  max_heartrate by type.activity
## Kruskal-Wallis chi-squared = 31.823, df = 12, p-value = 0.001473
```

Since p value is <0.05 we reject the Null and accept the Alternate. Hence with a 95% confidence **we can say that the means among populations is not the same.**

Can we predict Type of Activity?

Now that we have showed that maximum heartrate is not the same across activities at 0.05 level of significance, we can use this fact to create a K-Nearest-Neighbor model to predict the type of activity.

K-Nearest Neighbors Model:

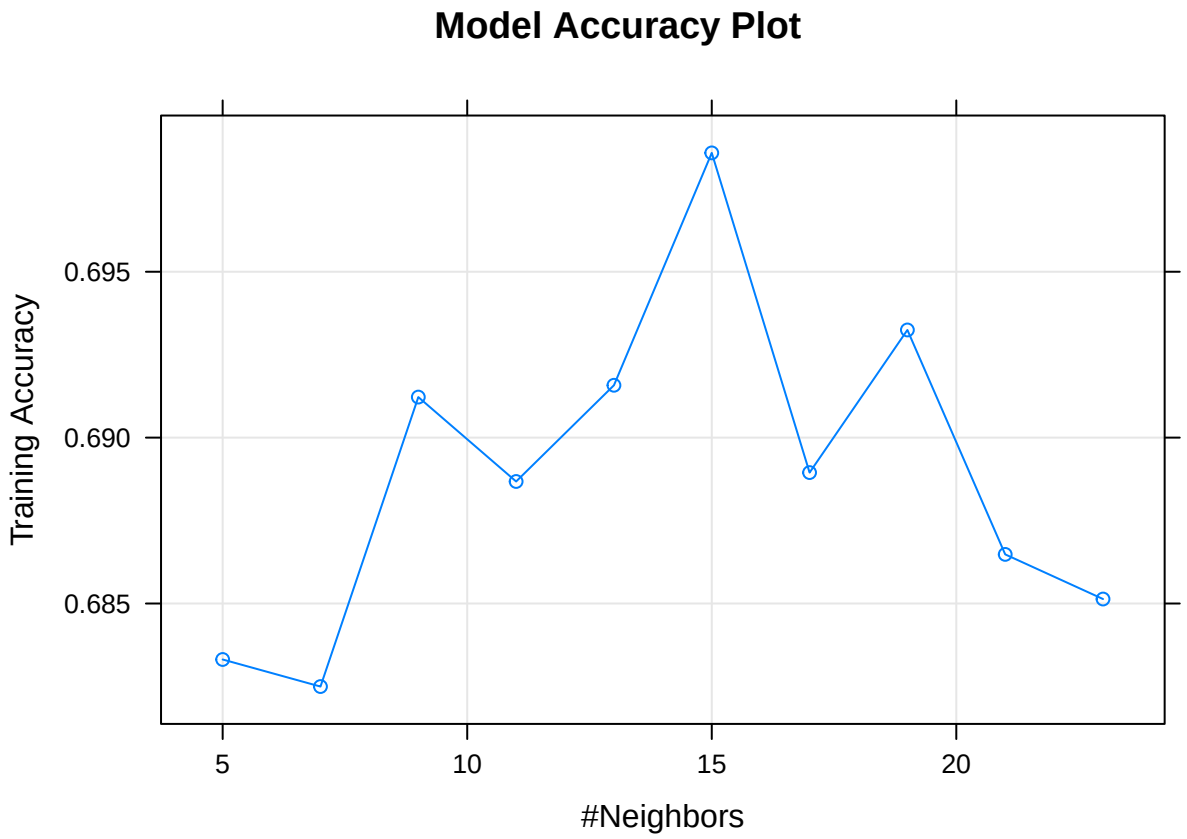
K-Nearest Neighbors(k-NN) is a type of **instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.** In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. A commonly used distance metric for continuous variables is Euclidean distance.
We can use this method to predict the type of activity, we will use **Maximum Heartrate, Kudos Count and Elapsed Time** values from the dataset.

Splitting the dataset into 75% training data and 25% test data. Using repeated Cross Validation method of 10 splits running 3 times on each epoch to further train the model better.

Given below are the model results for test data.

```
##                   Result
## Accuracy       0.6791209
## Kappa          0.2237219
## AccuracyLower  0.6340670
## AccuracyUpper  0.7218271
## AccuracyNull   0.6461538
## AccuracyPValue 0.0768014
```

We can also see how the training accuracy had improved over increase in number of neighbors.



Conclusion

The few questions that I was able to get to in this report using Strava data are just pebbles in an ocean of questions that could be asked from this dataset and answered using Data Science.
The initial question about comparing intensity of workout between Men and Women was done by looking at the coefficient of gender variable in a linear model. This is just one way to go about answering the question. It is important to note that the coeffecient of gender was very low and even close to 0. So there is room for improvement in terms of building the model, which **could possibly increase the favorability towards women or even reverse the answer towards Men.** Hence we cannot say with utmost confidence that women exercise more intensly than men.

The second question was to see if the maximum heartrate achieved by people is same across different activities. It goes without saying that maximum heartrate is not a precise value and is bound to some error in calculation as we can see from the outliers in the box plot. The ideal way to compare heartrates across different groups(activities) would be to use ANOVA test. The most important assumption in ANOVA is that the variances are same across groups. We used Bartlett's test to make sure that the variances are same. Since we weren't able to **conclude that the variances are same with a 95% confidence** and there is a variation in sample size across groups, we were not able to use the standard ANOVA test. Instead we used Kruskal Wallis test that does not require that each group be sampled from a normal population. This test **concluded that the maximum heartrate is not the same across all activities at a 0.05 level of significance.**

The conclusion in the second part was the motivation to see if using this conclusion we can create a Nearest Neighbors model to identify the type of activity. We also decided to include Kudos count and Elapsed time to better fit the model. While we were able to get close to a 70% accuracy in training data, the test data accuracy was close to 65%. This is a **good result given the huge dataset bias in number of data for each category.** These values can be increased given good representation of different activities in the dataset.