Strava Data Evaluation

Christopher Painton & Mahir Bathija

Assignment 3: Mini Data Science Project - INFO 370 (Autumn 2018)

Exploratory Data Analysis

Our mini data science project will be focusing on data collected from Strava's API, an app that creates unique activity tracking for athletes. Starva's activity dataset contains a wide range of information based on workout activity statistics and athlete's locations. Based on this dataset, we plan on exploring the following two questions:

1. Do men tend to exercise more intensely (taking into account both distance and speed) than women?
2. Which country (with at least 50 bike workouts) has the most efficient bike workouts when considering workout time, moving time, and kilojoules burned?

The current Strava dataset has over eight thousand observations with 53 different variables. Our key variables of interest are: athlete country, athlete sex, average speed/heart rate, moving time, kilojoules, distance, time (total and moving), total elevation gain, and activity type. An initial breakdown of gender (Figure 1) and workout types (Figure 2) shows that the app has a balanced use between men and women with bike rides and running by far the most popular activities.

Male	Female
4,084	3,824

Figure 1: Male vs Female Breakdown. Based on our dataset, there are 260 more men that recorded workouts than females. In addition, there are 185 observations that do not contain any gender information.

Workout Type	Count
Ride (Bike)	4,512
Run	3,001
Walk	194
Swim	178
Workout	66
Hike	37

Figure 2: Top 6 Most Popular Workouts. The app is far more popular for bikers and runners.

For evaluating workout intensity and efficiency, understanding the distributions of our five key variables (average speed, average heart rate, elapsed time, distance traveled, and kilojoules) is crucial to prepare for any analysis. The average speed distribution (Appendix A) increases steadily until the ~2.5 m/s mark, when it begins dipping down until rising again at the ~4.5 m/s point to the ~7 m/s point. This "camel" effect can best be described from the exercise types as the first apex would demonstrate a much more casual running speed/light bike speed and the second spike indicates a more rigorous tempo specifically for biking. The average heart rate distribution (Appendix A) is right skewed, consistent with heart rates during workouts, as they would be much higher with averages centered around ~135 bpm. The distributions for distance, time, and kilojoules (Appendix B) are all left skewed distributions. This can be explained by the popularized idea of an one hour workout as the median for time centers around one hour (3600 seconds) and gives a sense of a standard workout.

One of the most noticeable issues with Strava's dataset is with the missing heart rate values, making it difficult to incorporate heart rate into any of our models. Another issue is with misrecorded/outlier data, as some observations have unrealistic (i.e workout > 30 hours) values that skew the averages of the dataset. In the next section, we detail how the data was prepared to adjust for these issues.

Data Preparation

As mentioned briefly above, the Strava activity dataset contains a significant amount of data that would hinder our analysis. To address our concerns, we developed the following strategies for cleaning our data. We started by removing any variables that were not of interest to our research questions, specifically information that didn't contain some nominal value or were location based (aside from the location of the activity). With the remaining location information, we adjusted the strings to contain no special characters and matched different spellings of countries names to the most common English spelling (i.e the Netherlands to Netherlands). Since we plan to use workout type in our statistical testing, we created dummy variables for the six most popular exercises (from Figure 2) and gender.

One of our primary concerns focused on outliers within many of the exercise variables that skewed our descriptive statistics within the dataset. To prevent this from having a heavy influence on our results, we removed any values that were outside of a range that we deemed to be an "unrealistic" workout result. These ranges include average/max speed (0.1 m/s – 45 m/s), distance (170 m – 80,467.2 m), elapsed time (5 min – 600 min), moving time (5 min – 600 min), max heart rate (37 bpm – 220 bpm), and elevation (under 4000 m). Originally we planned to take the z-scores and remove values that were over four standard deviations away from the mean, but felt it wasn't appropriate based on how our distributions are skewed. Our last step was to create additional columns that converted speed in m/s to mph, meters to miles, and seconds to minutes. The purpose of this final step was to aid our understanding of the values, giving them a imperial comparison.

Statistical Modeling

For a quick statistical evaluation, we took the descriptive averages for both male (Figure 3) and females (Figure 4). The resulting tables showed that males tend to have higher average speeds, elevation gains, distance, and exercise/moving time than women. However, women have a higher average heart rate (commonly used to define an “intense” workout). In addition, men favored bike rides as their most popular workout type while women favor running. Since men and women have different “favorite” workouts, we also compared the means of the bike rides and runs separately (Appendix C). The comparison for the bike rides and runs follows a similar breakdown to the two tables below.

Male Descriptive Stats

Average Heart Rate	139.319
Average Speed (MPH)	11.424
Max Speed (MPH)	24.340
Elevation Gain (Meters)	221.447
Distance (Miles)	13.850
Exercise Time (Min)	84.237
Moving Time (Min)	69.757
Most Popular Workout	Bike Ride

Figure 3: Male Descriptive Averages For Workout. Males have higher average speeds, elevation gains, distance, and workout times than females.

Female Descriptive Stats

Average Heart Rate	141.405
Average Speed (MPH)	8.219
Max Speed (MPH)	17.825
Elevation Gain (Meters)	151.595
Distance (Miles)	8.888
Exercise Time (Min)	74.270
Moving Time (Min)	60.364
Most Popular Workout	Run

Figure 4: Female Descriptive Averages For Workout. Females most popular workout is running and have on average a higher exercise heart rate than males.

A similar overview was used to compare countries with at least 50 recorded bike workouts (Appendix D). The chosen variables for these tables were more focused on aspects that define a more efficient workout. From this table, Brazil have the highest average workout time, but were on the bottom half of “efficiency” measures: moving time % (last), average speed (bottom 3), and kilojoules per minute (last). Netherlands, Spain, and France appear to have the most continuous workouts based on the moving time percentages.

For a more statistical test between the differences of the male and female averages, we started with Shapiro tests to see if any of our variables of interest (average heart rate, average/max speed, elevation gain, distance, and exercise/moving time) were normally distributed. The resulting p-values were all low, indicating that the variables were not normally distributed (consistent with our EDA density plots). Since the values were not normally distributed, we conducted an unpaired two-sample Wilcoxon Test to compare the median values. A Wilcoxon Test is an alternative to the two-sample t-test for comparing independent samples medians. Since the descriptive averages displayed that men were higher in most of the exercise categories, our alternative hypothesis for each test was if the male median values were statistically greater than the female median values. The resulting tests were used to compare each of the seven categories of interest across our general, bike, and run datasets (Figure 5).

As a final measure to compare the male and females workouts, we developed an intensity score that is loosely based off the TSS and rTSS scores (training stress score)[1]. The TSS formula consists of:

$$\frac{Time\ (Seconds) * NP\ (Normalized\ Power) * IF\ (Intensity\ Factor = NP/FTP)}{FTP\ (Power\ Threshold) * 3600} * 100$$

The threshold power is traditionally found via watts, but since our model needs to include speed and distance, we measured power based on the average and max speeds of the runner/bike rider and total elevation gain over distance. Our modified formula is:

$$\frac{Elapsed\ Time\ (seconds) * NP\ (\ NP = Average\ Speed\ (m/s) * (Total\ Elevation\ Gain\ (m)\ /\ Total\ Distance\ (m)) * \frac{NP}{FTP}}{Elapsed\ Time\ (seconds) * FTP\ (FTP = Max\ Speed\ (m/s)\ \ * (Total\ Elevation\ Gain\ (m)\ /\ Total\ Distance\ (m))} * Total\ Distance\ (m)$$

This formula rewards athletes that have long distance exercises and higher elevation gains but punishes those whose average speed is a smaller percentage of their max speed (maximum effort). The NP formula is a measure of how the athlete performed while accounting for elevation changes (rewarding higher elevation gain / distance values). The FTP formula is a calculation of an athlete's assumed "maximum" workout, measuring a value if the athlete had maintained their max speed for the entirety of their workout. The NP/FTP value becomes this ratio to measure how the athlete's effort compared to their predicted maximum effort. These scores were then scaled to fit a 0-100 scale where 100 for represented the max score for males (bikes and runs) and 0 represented the minimum score for females on runs and the minimum score for males on bikes. The intensity score values for bike rides and runs were seperated due to the differences in average speeds.

Our definition of efficiency is a combination of two factors: what percentage of the workout was "active" (moving) and how many kilojoules were burned. With these two factors in mind we wanted to create a modified formula to calculate efficiency:

$$\frac{Kilojoules\ (Kj)}{Distance\ (km)} * \frac{Moving\ Time\ (min)}{Elapsed\ Time\ (min)}$$

This efficiency score takes the kilojoules burned per kilometer and multiples it by the percentage of the workout that was continuous, or moving. Since we don't see speed and distance as

[1] https://help.trainingpeaks.com/hc/en-us/articles/204071944-Training-Stress-Scores-TSS-Explained

accurate measures of "efficiency", this formula focuses on the athlete's ability to maintain a continuous pace and to burn kilojoules at a higher ratio per kilometer rather than simply burning the most kilojoules. We then scaled to fit a 0-100 score using this formula:

$$\frac{Country\ (efficiency) - Minimum\ Efficiency\ (efficiency)}{Maximum\ Efficiency\ (efficiency) - Minimum\ Efficiency\ (efficiency)} * 100$$

To determine which statistical hypothesis test to use on our efficiency scores, we used the Shapiro test of normality and got the following results:

Country	Shapiro Test (p value)
United States	$2.595e^{-34}$
United Kingdom	$1.693e^{-19}$
Australia	0.02982
Brazil	$2.0283e^{-9}$
Netherlands	$6.062e^{-5}$
Spain	0.01883
France	0.01986
Italy	0.0002174
Canada	0.0004781
Germany	0.8543

Figure 5: Countries along with p values from Shapiro test of normality.
Ho: The distribution is normal / Ha: The distribution is not normal

Setting our significance level (alpha) to 0.05, we see that the distributions of all the countries are not normal (except for Germany).

Results

Based on the descriptive averages from our simple statistical sets above, we noticed that the males had higher mean averages on the intense workout categories than the females. To statistically measure this for accuracy, we conducted two sample Wilcoxon tests on each variable of interest (for measuring "intensity"). The resulting p-values (Figure 5) show us that we fail to reject the null hypothesis for the average heart rate (for all three datasets) and workout time (for the bike dataset). In contrast, we see strong evidence from the low p-values, that the medians for majority of our variables of interest are consistent with our means in the sense that the males have higher averages than females (aside from the four results that had p-values > 0.05).

Category	General Male vs Female	Bike Male vs Female	Run Male vs Female
Average Heart Rate	0.9957	0.656	0.9412
Average Speed (MPH)	< .00000000000000022	< .00000000000000022	< .00000000000000022
Max Speed (MPH)	< .00000000000000022	< .00000000000000022	.000000001196
Elevation Gain (Meters)	< .00000000000000022	.00000000003428	.000004788
Distance (Miles)	< .00000000000000022	.0000000006565	< .00000000000000022
Exercise Time (Min)	< .00000000000000022	0.4298	0.02444
Moving Time (Min)	< .00000000000000022	0.03421	0.001132

Figure 5: P-Values for Wilcoxon Test for General, Bike and Run Male/Female datasets at 0.95 confidence level. The alternative hypothesis for each test was to see if the male median value was statistically greater than the female median value. The results indicate that we fail to reject the null hypothesis on the average heart rate for all three datasets and exercise time for the bike dataset.

To help visualize these results, we created overlapping density plots for both male and females (Appendix E) displaying the comparison between their average speeds and distances. Based on these plots and support from our Wilcoxon tests, we can see that averages for distance and speed are higher for males than females.

As a final test for comparing workout intensity, we calculated "intensity scores", from the formula above, for males and female (bike rides and runs measured separately). The resulting mean averages (Figure 6) shows that men have higher intensity scores than females on both bike rides and runs (represented visually in Appendix F). Another two sample Wilcoxon test was conducted to identify if the medians for the males runs/bikes is higher than the females runs/bikes. The resulting p-values, 1.539e^{-6} (for bike rides) and 6.437e^{-16} (for runs), leaves us with strong evidence to reject the null and support that the male medians (for bikes and runs) are higher than the female medians This information falls in line with our other findings from the descriptive averages and the Wilcoxon test that males have higher averages than females on our workout variables (listed from Figure 5).

Gender	Average for Bike Ride	Average for Runs
Male	14.250	15.457
Female	12.022	12.315

Figure 6: Average Intensity Scores for Bike Rides and Runs.
(For finding scores, max speeds greater than 40 mph (bike)
and 20 mph (runs) were removed)

In order to test the average values of efficiencies for countries with non normal distributions, we used a one sample Wilcoxon test. For Germany, since it has a normal distribution for efficiency scores, we used the one sample T-test. When setting the significance level (alpha) to 0.05 the resulting p-values are:

Country	**Wilcoxon / T Test (p value)**	**Median Efficiency Score**
United States	0.07462	12.2604
United Kingdom	0.5429	11.8180
Australia	0.6241	12.9540
Brazil	0.07052	11.6726
Netherlands	0.87	10.8264
Spain	0.08465	14.3546
France	0.2922	12.2848
Italy	0.5602	14.7839
Canada	0.7953	12.7042
Germany	0.7052	13.03603

Figure 7: Countries and their median efficiencies along with the p values from hypothesis testing
Ho : The true location of the value is equal to median (country) / Ha : The true location of the value is not equal to median (country)

From the Wilcoxon and T tests, we can support that the countries with the highest efficiency scores are Italy and Spain with scores at 14.78 and 14.35 respectively. By looking at the medians calculated (and only the medians calculated) we can say that Italy has the highest efficiency score for bike workouts. However, since we only have these results from sample data and the fact that Spain is only a few decimal points behind Italy we cannot be certain. When conducting two sample t-tests on the medians of Italy and Spain (Figure 8), we come up with the following results:

Type of T test	P value
Two sample Two tailed T test	0.9079
Two sample Less than T test	0.5461
Two sample Greater than T test	0.4539

Figure 8: Countries and their median efficiencies along with the p values from hypothesis testing
Since the p-values are large values, we fail to reject the null hypothesis: That the medians for Italy and Spain aren't statistically different.

Discussion

Based on our findings from the descriptive stats, Wilcoxon tests and intensity scores, we discovered that males do tend to exercise more intensely than females. With that being said, one of the assumptions of this model relies on the maximum speed being an accurate representation of the athletes maximum effort. In addition, since speed and distance were required for our model, we were limited from finding a proper "power" index needed for TSS and rTSS calculations. Another limitation for our model was in the lack of heart rate data, which can often be a good indicator for workout intensity. As a result, we needed to create a formula that measured intensity without accounting for heart rate. While taking all these limitations into account, our intensity formula placed heavy emphasis on speed, distance and effort. Since our findings from these scores lined up with a comparison of the mean averages and the medians (from the Wilcoxon test), we feel comfortable concluding that men workouts (on bike rides and runs) are more intense than females.

After conducting multiple Wilcoxon and t-tests, we believe that Italy and Spain are the countries with the most efficient workouts. When looking at the efficiency scores we calculated and using statistical testing to measure their significance, we feel this is an accurate representation and these values support basic analysis of the descriptive stats (Appendix D). Although Italy had the higher efficiency score, our two sample t-test with Spain returned large p-values indicating that we cannot reject the null hypothesis that there is no statistical difference between the two. One of the biggest limitations for this model comes from the lack of data. We only compared a total of 10 countries because of the small sample sizes from the other countries, making it difficult to draw conclusive results. Another limitation from our model comes from the fact that we didn't have an efficiency score to measure the run workout types. Countries like Canada and France suffered from this, as a large portion of their sample data were from runs and not bike rides. However, considering the data we were provided with and after thoroughly conducting multiple statistical tests, we feel comfortable stating that Spain and Italy have the most efficient bike workouts.

Our research set out a goal of clearing misconceptions about workouts like men being better athletes or that certain "rich" countries have better workout regimes than others. Although our findings discovered that males workout more "intensely" than females, a key factor, heart rate, was not included in our models and the Wilcoxon tests supported higher heart rates for females. Perhaps more time and data would allow for future investigation in the differences of workout regimes that can properly utilize TSS and rTSS scores. The information gathered from our second research question provided insight into popular biking countries' workout efficiency. Despite countries like the United States having more exercise resources, the efficiency scores helped bring a more representative picture between these countries bike workouts. While our analysis was through and examined closely, it was only a small insight into the activity data and we hope that these findings will inspire further research and discussion into topics of measuring workouts.

Appendix

Appendix A: Distribution of Average Speed and Heart Rate

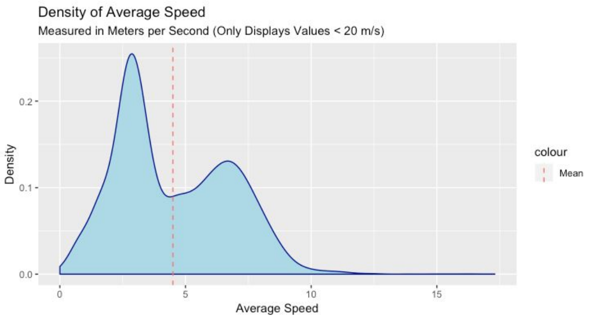

Figure 1: Density Distribution for Average Speed. This distribution has two spikes in density, indicating different workouts. The first spike demonstrates a more casual running speed at ~2.5 m/s or a light biking speed. The second spike indicates a much more intense bike speed, and is a shorter apex because it's easier for a biker to go at slower speeds than a runner to get up to higher speeds.

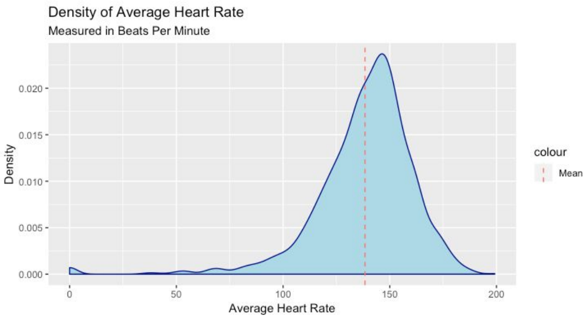

Figure 2: Density Distribution for Average Heart Rate. Our average resting heart rate is anywhere between 60-100 beats per minute and as we begin to workout our heart race increases to an average of 135 beats per minute (depending on the intensity of our workout). As a result, these values will shift the distribution to the right and display an average heart rates well over 100 bpm when working out.

Appendix B: Distribution for Exercise Time, Distance, and Kilojoules

Worth noting that the values that are huddled around the x-axis of zero aren't values that equal zero, but rather quick workouts that recorded a low value greater than 0.

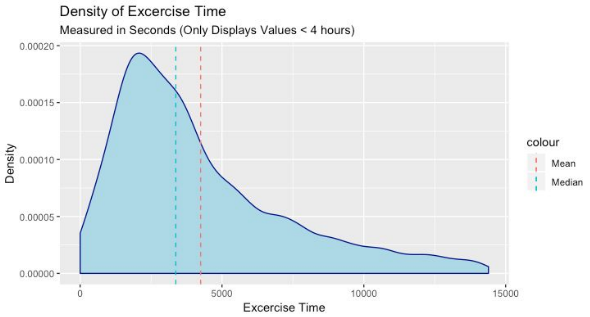

Figure 1: Density of Exercise Time. The popular one hour workout explains why the median value is at ~ 3600 seconds (or one hour). The distribution is left skewed because shorter, quick cardio workouts are much more common than long extended workouts.

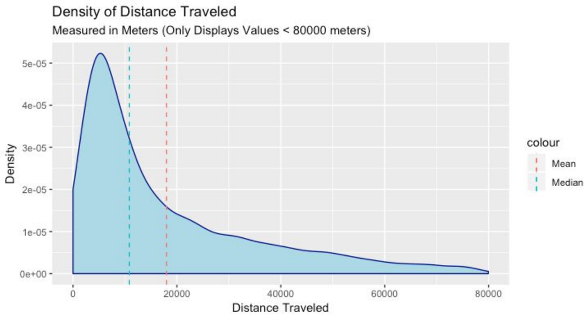

Figure 2: Density of Distance Traveled. Based on the idea from Figure 1, the median for distance traveled can give us a sense of what an average one hour workout would look like. By taking the median at ~11000 meters and assuming that the workout average time is one hour, we can "predict" an average workout speed of ~3 m/s. When comparing this value to Appendix A (Figure 1), we see that the average speed with the highest density closely resembles our workout speed found from this distribution (~2.7 m/s to 3/s).

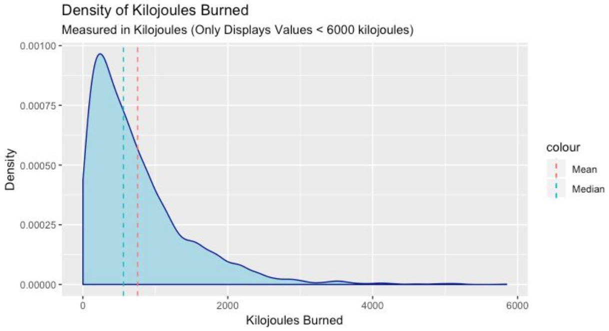

Figure 3: Kilojoules Output Density Plot. This specific example displays a heavily left skewed distribution, influenced by short bike workouts with low kilojoules values. These low values display more casual workouts as the median is ~1000 kilojoules .

Appendix C: Male vs Female Descriptive Stats for Bike Rides and Runs

Male Descriptive Bike Ride Stats

Average Heart Rate	136.4
Average Speed (MPH)	13.862
Max Speed (MPH)	30.05
Elevation Gain (Meters)	283.7
Distance (Miles)	17.902
Exercise Time (Min)	96.48
Moving Time (Min)	78.32

Figure 3: Male Descriptive Averages For Bike Ride. Males have higher averages in six of the seven categories for bike rides and have a higher percentage of moving time in their exercises.

Female Descriptive Bike Ride Stats

Average Heart Rate	135.9
Average Speed (MPH)	12.219
Max Speed (MPH)	27.02
Elevation Gain (Meters)	236.5
Distance (Miles)	15.566
Exercise Time (Min)	100.383
Moving Time (Min)	76.606

Figure 4: Female Descriptive Averages For Bike Ride. Females only have a higher average exercise time than men for bike rides, but spend less of that time moving (on average)

Male Descriptive Run Stats

Average Heart Rate	147.7
Average Speed (MPH)	6.670
Max Speed (MPH)	12.663
Elevation Gain (Meters)	92.29
Distance (Miles)	5.471
Exercise Time (Min)	56.148
Moving Time (Min)	50.857

Figure 3: Male Descriptive Averages For Runs. Males have higher averages in five of the seven categories for runs except for heart rate and elevation gain.

Female Descriptive Run Stats

Average Heart Rate	149.0
Average Speed (MPH)	5.823
Max Speed (MPH)	11.80
Elevation Gain (Meters)	98.08
Distance (Miles)	4.4731
Exercise Time (Min)	54.105
Moving Time (Min)	48.013

Figure 4: Female Descriptive Averages For Runs. Although the male statistics have higher averages in majority of the categories, females have higher averages in the more "intense" aspects of the workout: heart rate and elevation gain.

Appendix D: Averages for Countries with at least 100 recorded workouts

Note that kilojoules were only recorded for the bike rides.

Country	Exercise Time	Moving Time and %	Elevation Gain	Average Speed	Kilojoules Burned Per Workout and Minute	Top Workout
United States	74.059	61.366 (82.9%)	164.837 m	4.131mph	591.457 (7.97)	Bike Ride
United Kingdom	68.521	55.462 (80.9%)	130.449 m	4.546 mph	461.955 (6.74)	Bike Ride
Australia	78.660	61.376 (78.0%)	215.646 m	4.756 mph	631.341 (8.03)	Bike Ride
Brazil	118.900	90.389 (76.0%)	339.785 m	4.382 mph	762.007 (6.41)	Bike Ride
France	85.416	74.316 (87.0%)	230.033 m	4.230 mph	770.604 (9.02)	Run
Canada	68.171	56.522 (82.9%)	135.875 m	3.867 mph	499.087 (7.32)	Run
Spain	97.473	86.074 (88.3%)	353.558 m	4.508 mph	971.316 (9.96)	Bike Ride
Italy	113.142	96.368 (85.2%)	438.241 m	4.733 mph	1066.594 (9.43)	Bike Ride
Netherlands	80.542	73.623 (91.4%)	52.336 m	4.534 mph	711.625 (8.84)	Bike Ride
Germany	85.070	70.584 (83.0%)	180.730 m	4.456 mph	634.344 (7.46)	Bike Ride

Figure 1: Display of Descriptive Averages for each Country with at least 100 workouts. Although Brazil has the highest average exercise time, their average kilojoules output and moving time percentage are lower compared to the other countries. France and Canada are the only two countries that have runs as their most popular workout type (explaining why their average speeds make up two of the bottom three). Spain and Italy have the highest kilojoules burned per minute count.

Appendix E: Density Comparison of Average Speed and Distance for Male and Females

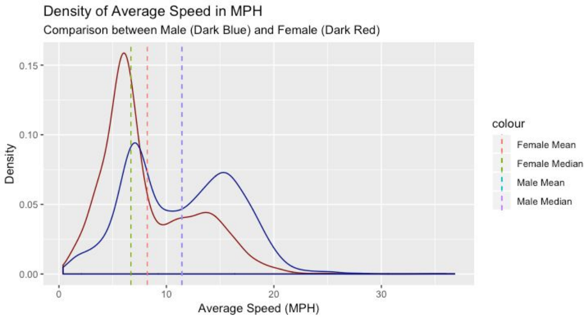

Figure 1: Density of Average Speed for Male and Female. After considering the mean averages and the p-values from the Wilcoxon test for average speed, we can see that males have both higher means and medians when comparing the average speed of workouts.

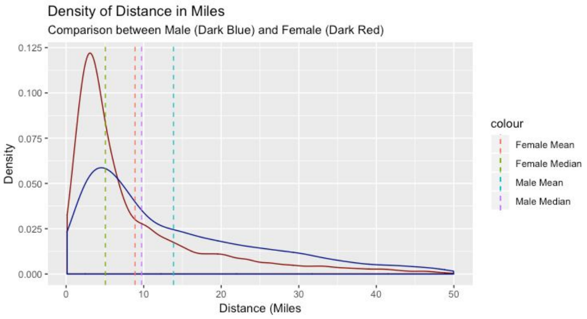

Figure 2: Density of Distance for Male and Female. After considering the mean averages and the p-values from the Wilcoxon test for the distance, we can see that males have both higher means and medians when comparing the distance covered of the workouts.

Appendix F: Density Comparison of Intensity Scores for Males and Females

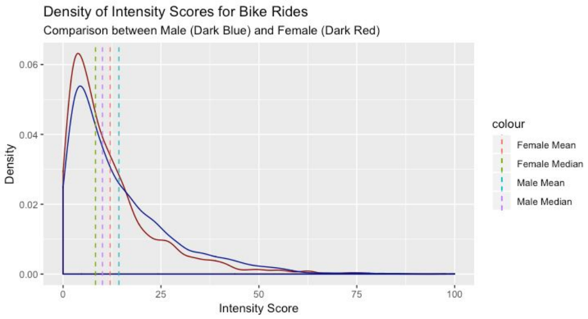

Figure 1: Intensity Score for Bike Rides. Based on our intensity metric, men have higher mean and median intensity scores than women for bike rides. These values were scaled to a 0-100 scale where 100 represents the maximum male bike ride intensity score and 0 represents the minimum male bike ride intensity score.

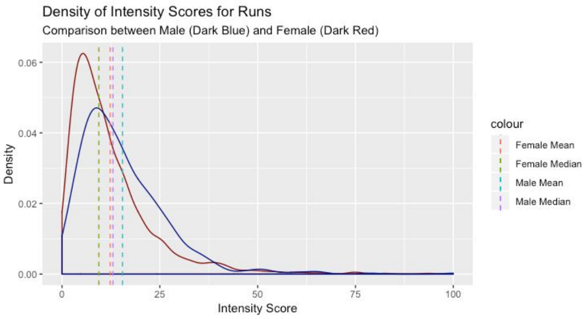

Figure 2: Intensity Score for Runs. Based on the formula for our intensity scores, men have higher mean and median intensity scores for runs, although the difference between the means are much closer than bike rides. These values were scaled to a 0-100 scale where 100 represents the maximum male bike ride intensity score and 0 represents the minimum female bike ride intensity score.