Project 1: A Dive Into Cardio Activities
Karim Sasa, 9/16/2024

Kaggle Dataset:

https://www.kaggle.com/datasets/deependraverma13/cardio-activities

Introduction to the Problem:

Cardio is an important component of everyone's daily lives. Each person may prefer different types of cardio, but which one stands out among them? As someone trying to burn the most calories efficiently, I'm looking at this Kaggle dataset that introduces three different types of cardio: walking, running, and cycling. I want to also find more correlations like does the climb of each exercise affects the calories burnt over time. Also, I'd like to know if the heart rate is higher for the corresponding cardio exercise. Overall, the main goal of this project is to figure out the best cardio exercise from the ones I'm given that burn the most calories over time.

Introduction to the Data :

The data is from a Kaggle dataset, the reference is listed at the top of this document, and it gathered info from 508 people with information about the date, the activity Id, the type of cardio, the route, the distance in km, duration, average pace, average speed, calories burned, climb in meters, average bpm, friends tagged, notes, and GPX file. So there are 508 rows in total, and there are 14 columns/features included. Overall, this data just shows the stats of each person during their exercise period. Also, all the categories are self-explanatory, but if you don't know what climb is it's essentially uphill resistance.

Pre-processing step:

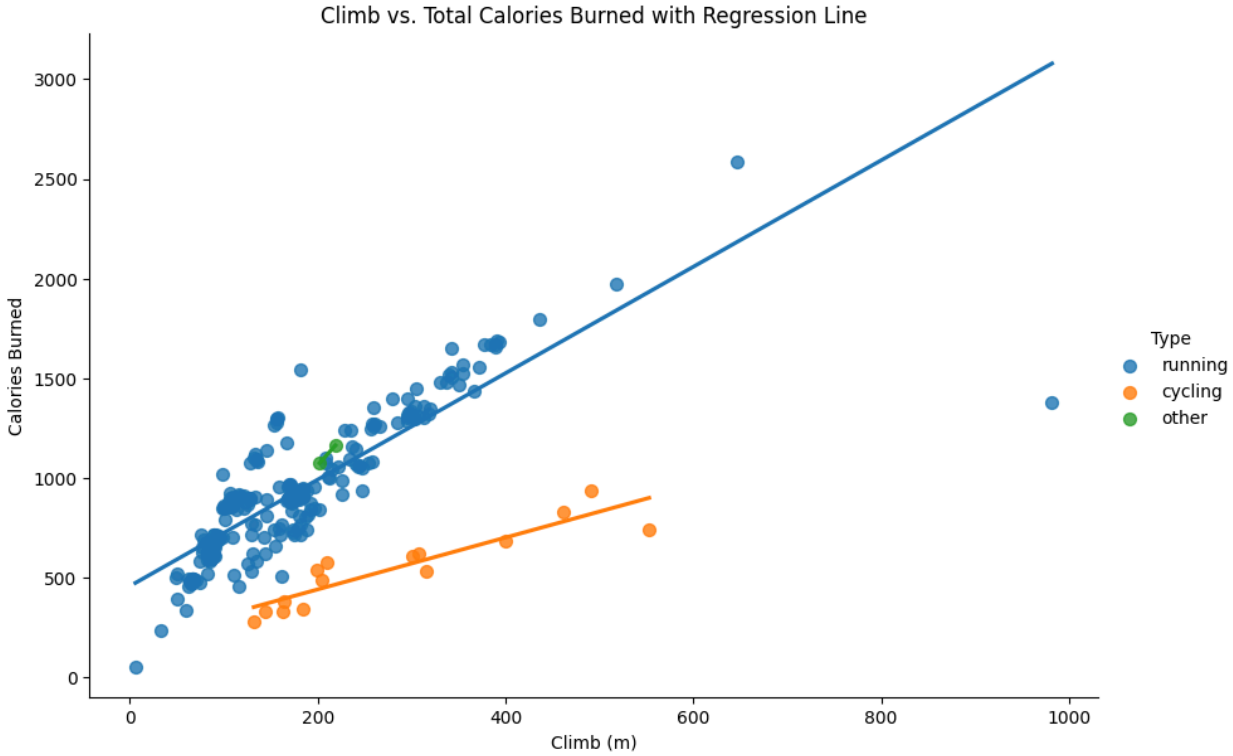
From class, I learned that the preprocessing step is usually the longest in the pipeline; however, my data was pretty clean, but there were some necessary preprocessing steps I had to take. First, I removed irrelevant data that won't contribute to answering my questions such as route name, friends tagged, notes, and the GPX file. Not only were these files irrelevant, but they also had only a few values. I also checked for duplicated values, but none were in my filtered dataset. All the data seems to be correct, so there is no need for type conversion. Now for syntax errors, there are no typos or whitespace. The standardization step required me to convert all my types of cardio names to lowercase. The scaling step isn't essential, as my data doesn't really have large values, and the same goes for normalization. For missing values, I removed all the rows that had null values for average heart rate in bpm, which brought me down to 294 rows.

Data Understanding/Visualization

For visualizing the questions I had for my data I used Seaborn, and I used scatter plots with regression tools. The findings of each visualization for the questions weren't too far from my expectations. Each visualization answers the questions I had were: does the climb affect the total calories burnt, heart rate in relation to cardio type, and calories burnt per minute in relation to the cardio type. The last visualization did surprise me, but I explain why in the visualization story.

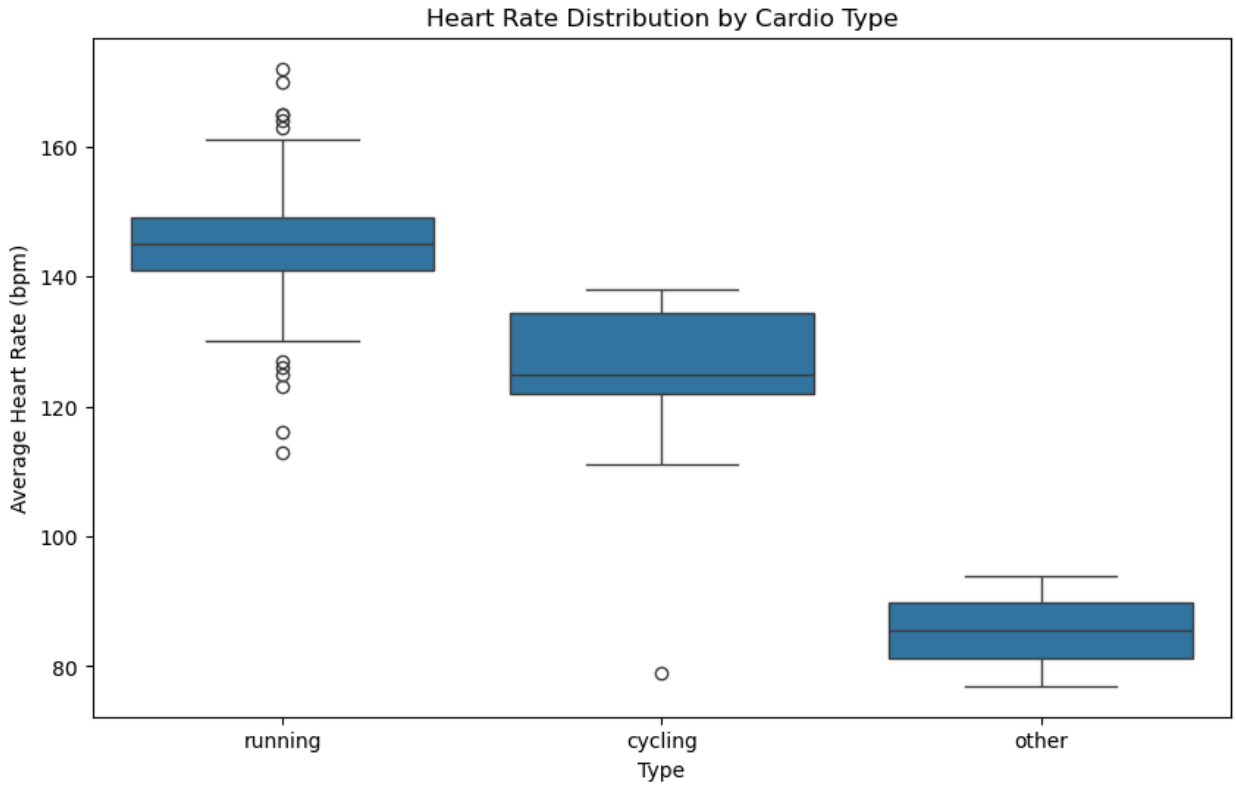
Visualization 1: How Climb Affects Total Calories Burnt: Story

In the figure below, you can see a scatterplot with a regression line showing a positive correlation in the graph. The two categories that are being compared show how as the climb increases, the total calories burnt increase for each different cardio activity. This is not a surprising result because increased vertical climb makes cardio more strenuous, which leads to more calories being burnt. However, even though it is somewhat obvious that this would be the result, it's nice to see the visualization of the graph. You can also observe that increasing the climb on cycling does not have as big of a change as changing the climb in running, it's hard to say for the others category since there is so little data for it. Overall, the visual answers the question of increased climb results an increased amount of calories burnt, and I can conclude that climb does have a significant effect on the value of it.



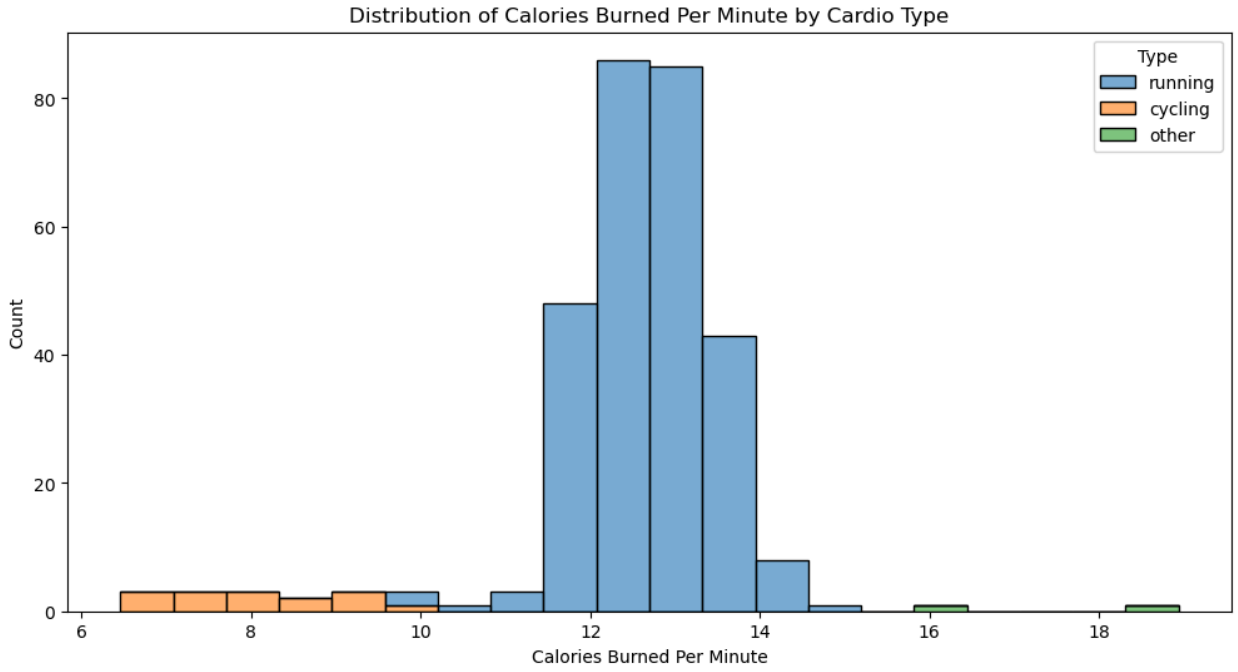
Visualization 2:The Distribution of Heart Rate by Cardio Type: Story

In the second figure below, you can see a boxplot showing the differences in average heart rate in BPM. We can see that running has the highest bpm, then cycling, then other category which is mostly walking. Through further research, it's recommended to reach a higher heart rate to not only burn more calories, but it is just healthy in general. With this information, running has a mean average of about 145 bpm, while cycling has 124 bpm, and other has 85 bpm. I believe that the heart rate of a person depends on their pace and speed, and it adds up since running uses the whole body, cycling only uses legs, and walking doesn't really take that much effort. Overall, coming to this conclusion was also fairly obvious, but you can see all the factors that come together in running affecting other values like climb affects heart rate, which in turn affects the calories burnt.



Visualization 3: Calories Burnt Over Time in Each Cardio Exercise: Story

In the final figure below, you can see a histogram showing the calories burnt overtime for each cardio exercise. It shows that cycling burned about 6.5 to 10 calories per minute for each person in the study, running burned 9 to 16 calories per minute, while “other” burned 16 and 19 calories per minute. I’m very surprised with these results because even though the “other” category has very few data values, it shows that they have a very high calories burned per minute value. This conclusion is very confusing as the “other” category includes walking, and other unknown cardio exercises, so I’m not sure if the other category is reliable in this chart. If we look at the running category, most people burn around 12 to 13.5 calories per minute compared to cycling which doesn’t have a dominant amount of calories burnt in any category. Overall, we can see running provides a consistent calories bpm as the graph has a symmetric shape with those data values.



Impact Section

From my perspective of the project, I saw that there are multiple limitations and some possible consequences to consider about the dataset. To start, it includes data from only 508 people, which definitely does not represent the whole population of active people. Not to mention, some factors like age, gender, fitness level, and health conditions can influence how different exercises affect calorie burn and heart rate, so the findings I found might not be able to be applied to everyone.

Another problem that the dataset has is that it doesn't describe the intensity at which each exercise was performed. Pace and speed could be used to determine that factor, but it could still cause issues. For example, walking could vary from a casual stroll to a fast walk, which can significantly change the calories burned, and affect the results negatively. I believe that this is what happened for the "other" category, which includes walking and other exercises. I think their pace and speed was a lot higher looking at the dataset, which caused it to skew in the last visual.

Another concern I have is, since I was solely focusing on the calories burned, it could encourage people to only exercise for that reason. For example, someone might choose running just because it burns more calories per minute, but they ignore the exercise that has a lower-impact like walking or cycling which might be more beneficial for their fitness level or physical condition.

Lastly, some data that might be missing that would be useful includes details like how often each person exercises and their individual fitness goals. For example, knowing if someone runs every day or just once a week could significantly affect their overall calorie burn and health outcomes. Also, if it included information like age, gender, and health conditions then I could apply my conclusions to more than just the younger generation.

In conclusion, while my visualizations can be helpful for information about how different exercises impact calorie burning, it should be taken with caution given the circumstances. I feel like a better dataset could include a more diverse dataset, including information like exercise enjoyment rating and frequency, and should also be transparent to information about individual health and fitness goals.

References:

Data Pipeline Resource:
https://www.beautiful.ai/player/-NoO5oPOTi1lShJwrDCs/2-The-Pipeline-Defining-the-Problem-and-Understanding-Data

Heart Rate Information Resource:
https://www.hopkinsmedicine.org/health/wellness-and-prevention/understanding-your-target-heart-rate