Sabancı University

CS210 – Introduction to Data Science

Term Project Report

Berke Ayyıldızlı - 31018

1. Introduction

This is my report for the term project of the CS210 – Introduction to Data Science course. The codes and the datasets can be accessed from the respected GitHub page:

https://github.com/berkeayyldzl/CS210_Term_Project.git

a. Motivation

The main hypothesis of the project is: Does my driving affect my heart? This topic is chosen, due to the fact that I had a feeling that I have a faster bpm than others, and I wanted to both see the truth behind it, and to better comprehend the causations behind this thought.

b. Objectives

The main objective in this project is:

- To implement the data science information that we have learned in the lectures on a real-life example.
- Investigate the relationship between driving behavior and heart rate.
- Do exploratory data analysis on the files.
- Successfully visualize the findings.
- Build a predictive model for average bpm using machine learning.

2. Data Collection

a. Data Sources

I have 2 separate data sources for this project.

i. A csv file that contains the trip data, taken from myOpel app.
ii. A csv file that contains all the health data, called export.csv, taken from Apple Health app, which then parsed in to separate files, to just show the relevant data about the project.

b. Data Processing

To process the data, the libraries such as pandas, seaborn and sklearn are used with the jupyter notebook.

3. Data Analysis

a. Exploratory Data Analysis

I began my analysis with some basic statistic data, such as first displaying the information about the databases:

```
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Date                            364 non-null    datetime64[ns]
 1   Departure Time                  364 non-null    object
 2   Arrival Time                    364 non-null    object
 3   Time Amount                     364 non-null    object
 4   Distance (km)                   364 non-null    float64
 5   Odometer Kilometer (km)         364 non-null    int64
 6   Average consumption (l/100km)   364 non-null    float64
 7   Fuel Prices (TRY/l)             363 non-null    float64
 8   Cost (TRY)                      363 non-null    float64
 9   Average Speed (km/h)            364 non-null    int64
 10  averageBPM                      364 non-null    float64
```

These statistics show that the data are cleaned and the nan values are filled with the mean values of the respected columns.

After the correlation matrix are specified, I wanted to dig a little bit deeper about my car usage. For this purpose, I wanted to see which day of the week that I use my car the most? Out of the 364 entries, here is the result:

```
DayOfWeek
Tuesday      84
Wednesday    64
Thursday     62
Monday       55
Friday       49
Sunday       29
Saturday     21
```

An addition to that, I also wanted to see the total driven kilometers per days:

```
Total distance driven on Tuesday: 1085 km
Total distance driven on Friday: 916 km
Total distance driven on Wednesday: 410 km
Total distance driven on Monday: 391 km
Total distance driven on Thursday: 332 km
Total distance driven on Sunday: 216 km
Total distance driven on Saturday: 54 km
```

This statistic matches with my expectation, as I use my car to come to school from Bursa, mainly on Tuesdays.

Continuing with the health Data Frame, I wanted to see which day of the week that I burn the calories most, using the dataset that nearly is from late 2018 until today.

```
Day_of_Week
Saturday     91401
Thursday     86610
Wednesday    83899
Friday       83375
Monday       78985
Tuesday      78342
Sunday       73852
```

Like the trip data, I also wanted to see which day of the week that I walked the most:

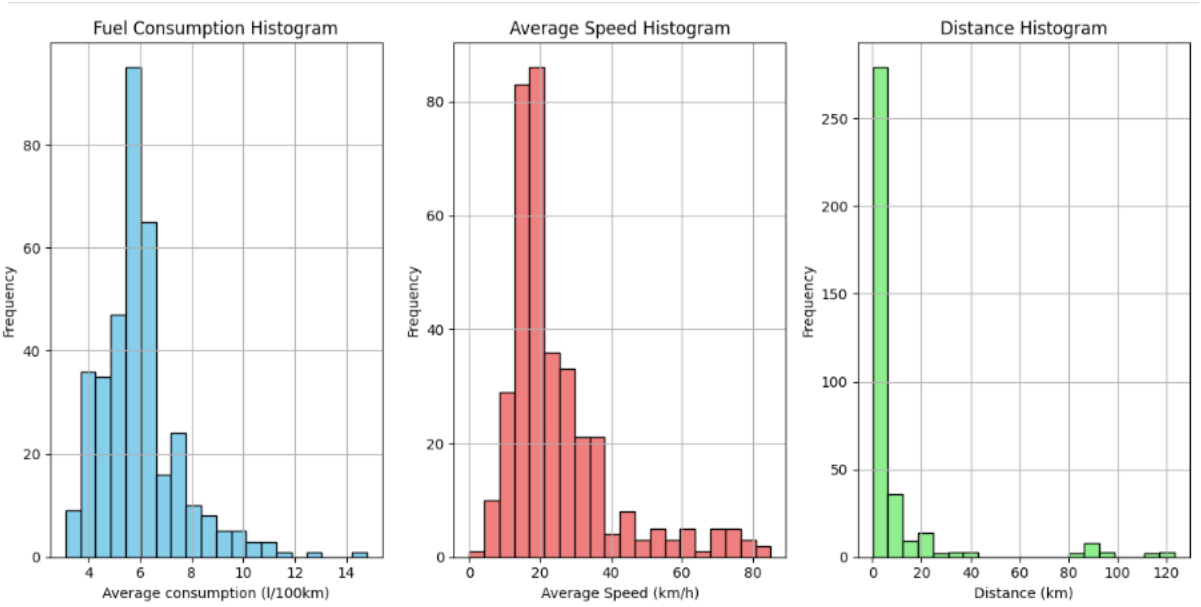
Day_of_Week	
Saturday	2835533
Sunday	2337355
Friday	2114339
Wednesday	2034499
Thursday	1939563
Tuesday	1889264
Monday	1835408

The result also fell within my expectation, as I usually sightsee or wander around on weekends.

b. Visualization

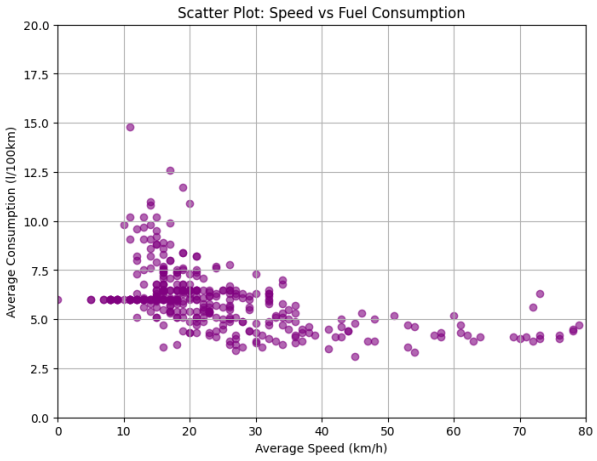
After the statistics, I plotted the information on the relevant plots, to better grasp the concept.

First, 3 histograms on Average Fuel Consumption, Average Speed and Distance:



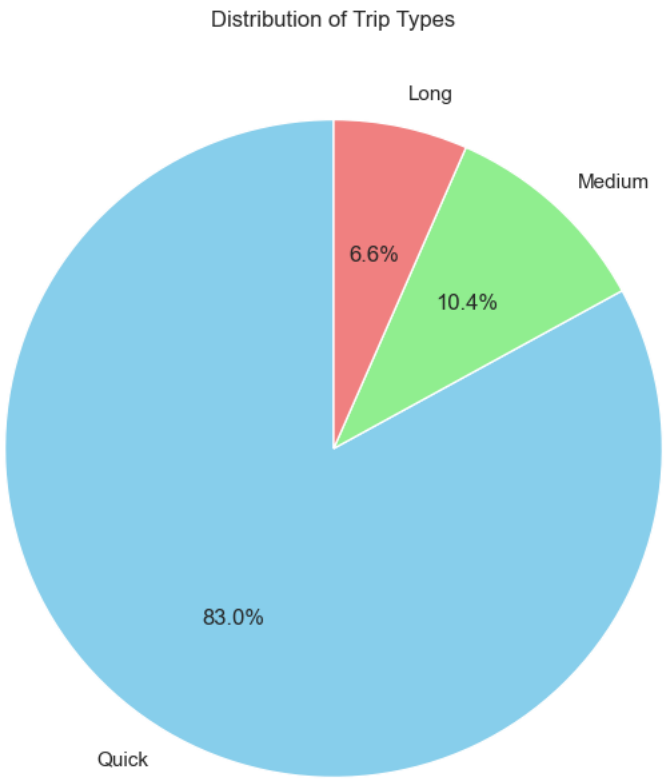
From here, I can see that my fuel consumption is mainly between 5.5 and 6.5 liters, my average speed is mainly below 40, and I usually drive my car for short trips.

Now another graph to see the correlation between Average Speed and Average Consumption:

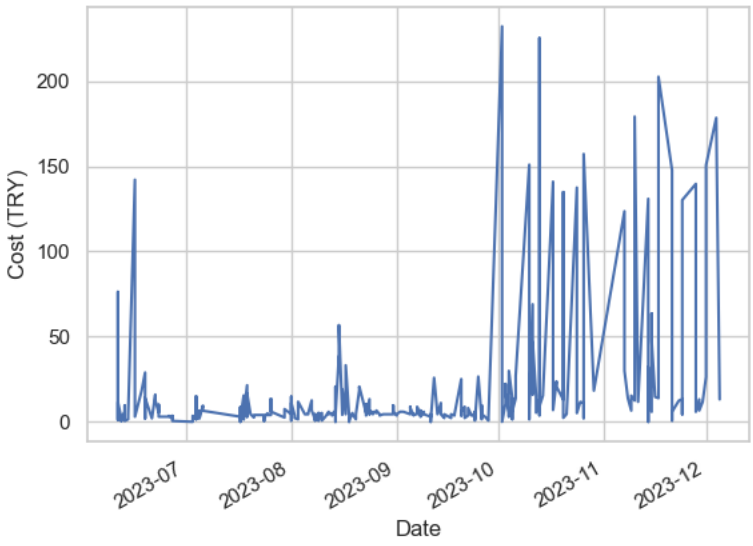


From here, we can see that the fuel consumption, gets higher when the speed becomes slower.

Also, for the trip lengths, I encoded each trip with respect to their kilometers, if a trip is smaller than 10 kilometers, it is a quick trip, if it is between 10 and 30 kilometers, it is a medium trip, and if it is longer than 30 kilometers, It is a long trip. Here is the pie chart distribution:

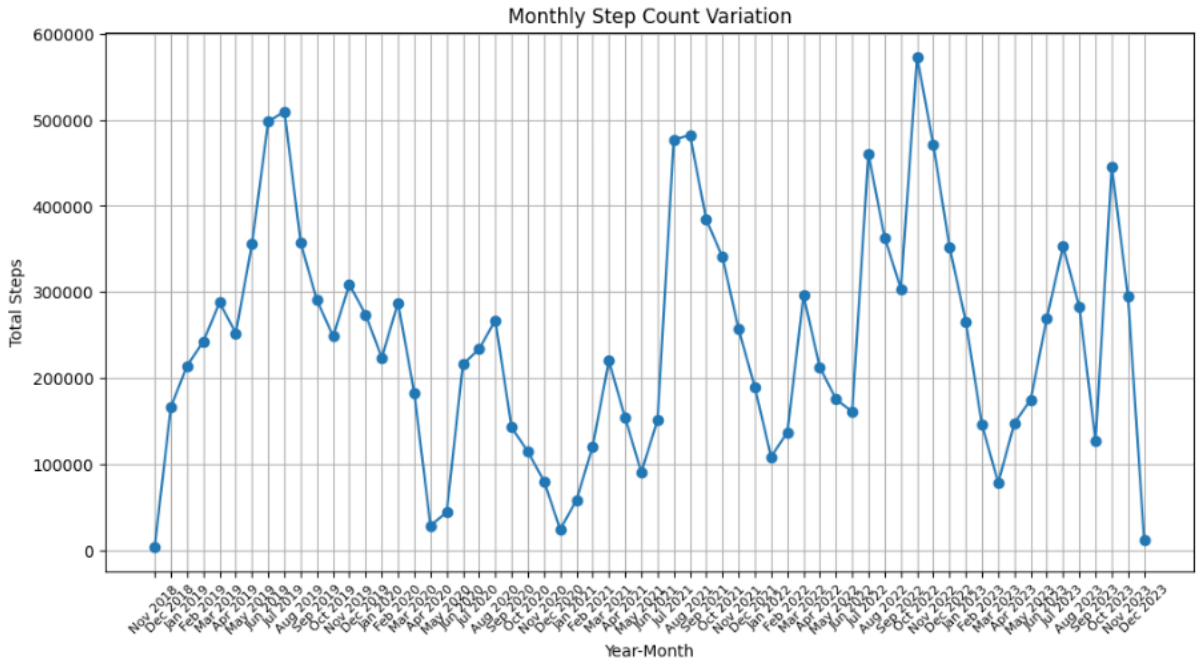


For the end of the trip dataframe, I wanted to show my expenditure of gas through the time:



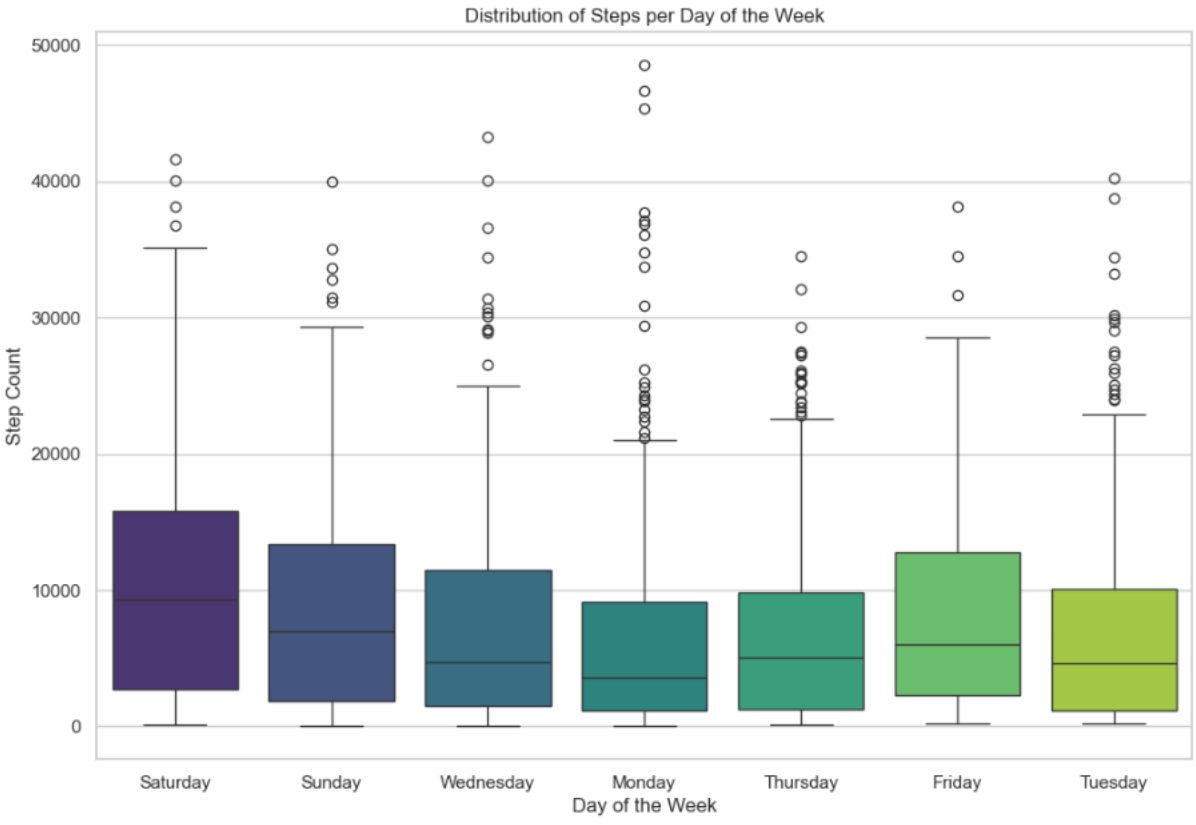
As can be seen from the graph, the cost spikes when the university starts, as I use it mostly to travel back and forth.

Now switching to health dataframe, I wanted to see the number of steps I took from November 2018 to today:

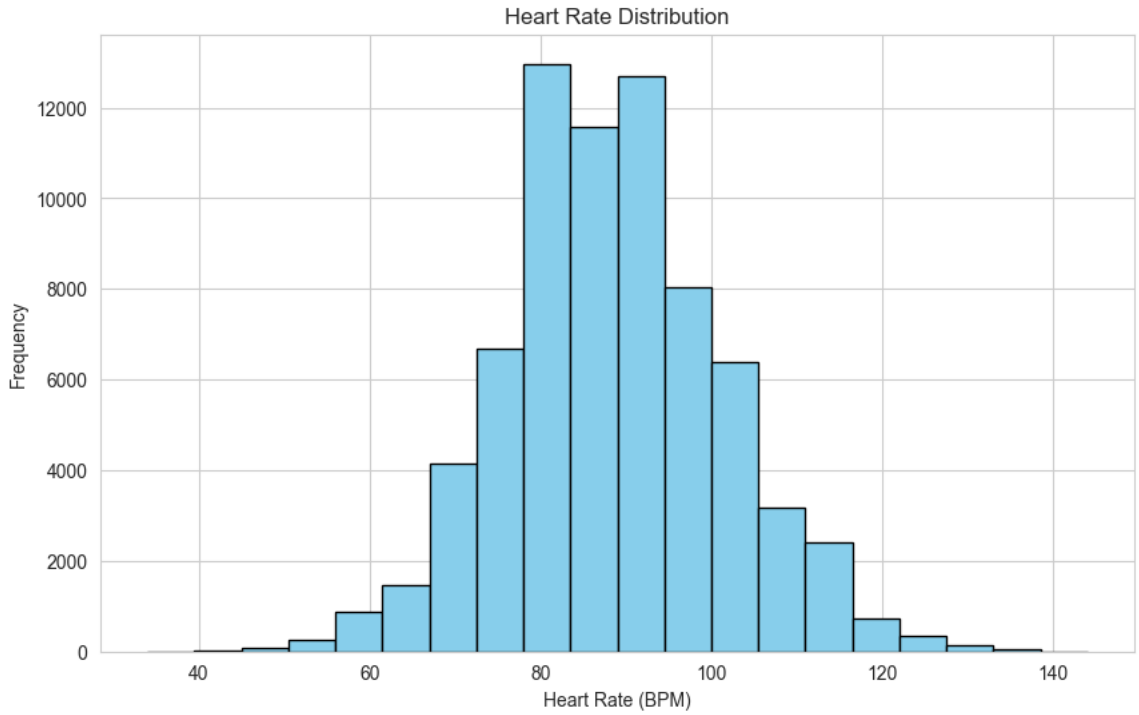


As it can be seen in this graph, my steps count spikes, when the university starts.

Let's see the distribution of steps per day of the week on this box chart:

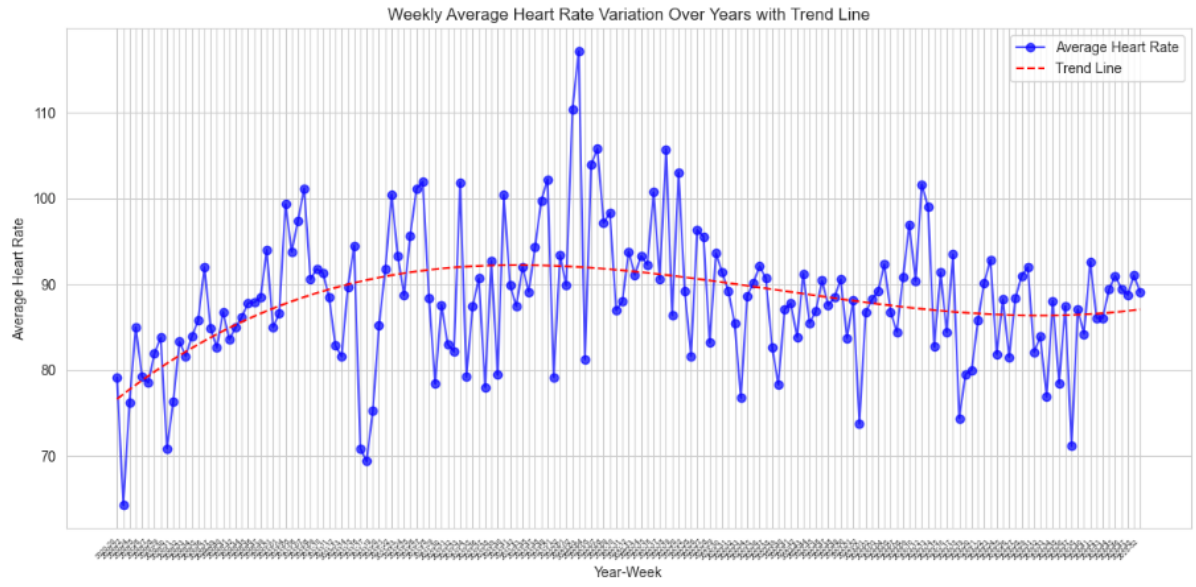


Before we correlate driving and bpm, let's look at my heart rate distribution on this histogram.



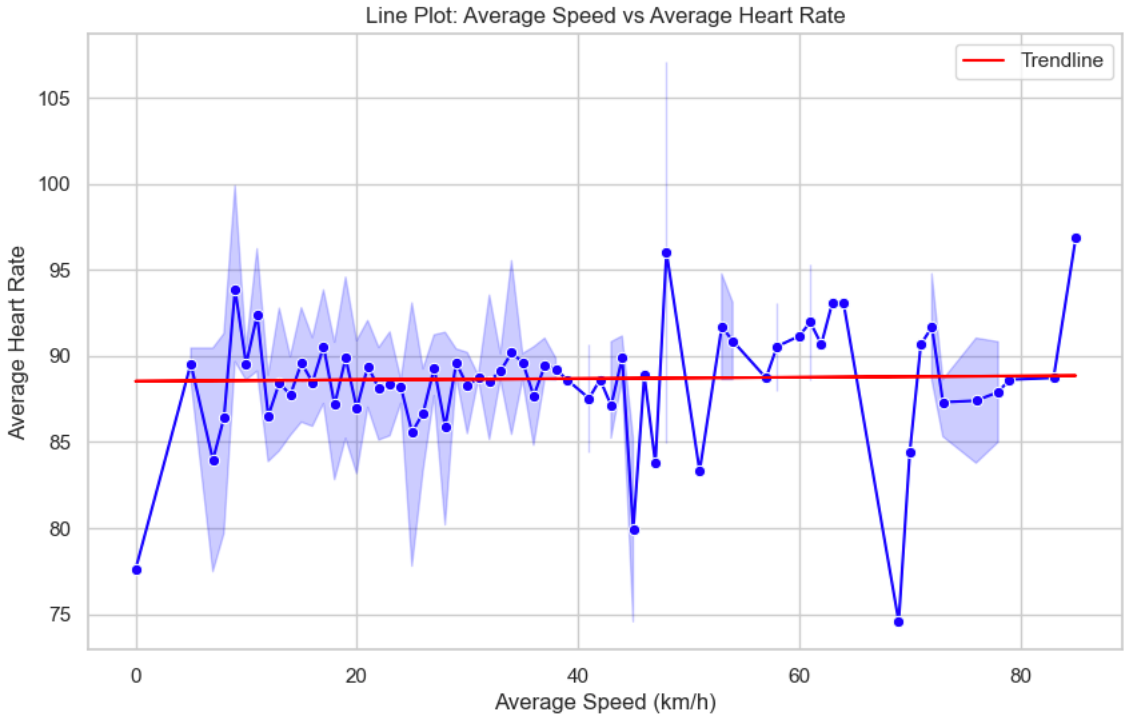
As can be interpreted from this chart, my average heart rate differs between 80 and 100.

Now let's look at this data with a trend through the years:



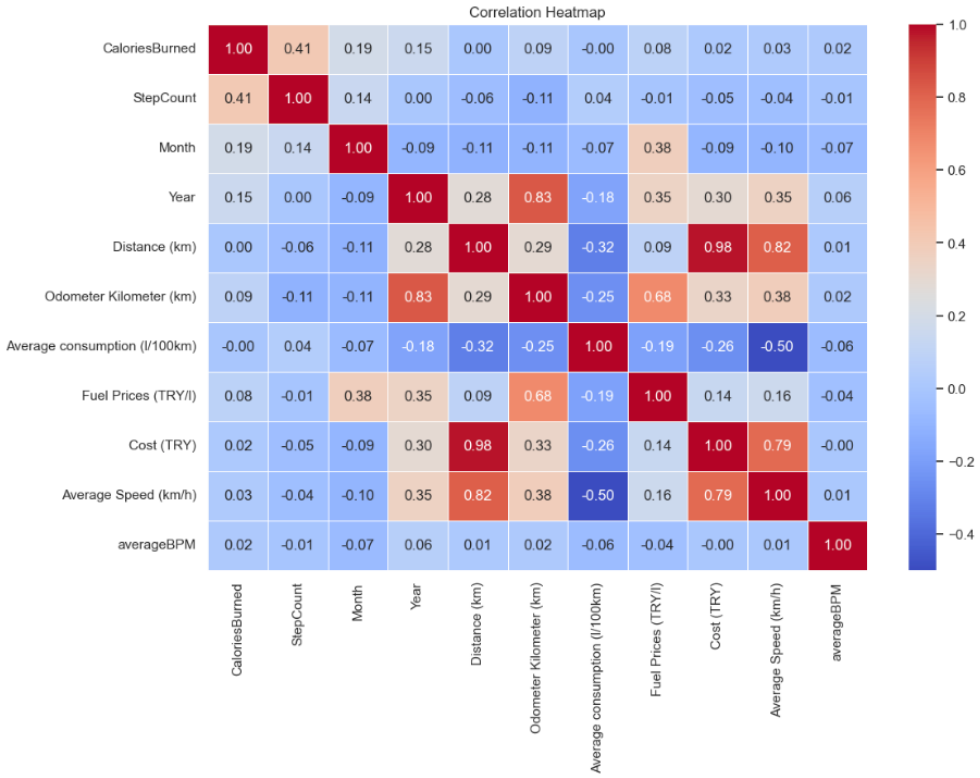
This graph shows that I have an increased hearth rate on the time of the pandemic, which, my number of steps also contribute to that.

Now for the hypothesis, here is the line plot with Average speed and Average Heart Rate:



From this graph, although, we can see the slight increase in average heart rate, the change is so small. The reasons for this conclusion problem will be addressed later.

To end the visualization part, here is the correlation matrix:



c. Machine Learning

For the machine learning part, I first shuffled the data and split it in a 80-20 perspective. After split, here is the results:

```
X_train shape: (291, 11)
X_test shape: (73, 11)
y_train shape: (291,)
y_test shape: (73,)
```

Now, to see the correlations of each category with the averageBPM:

```
Correlations with 'averageBPM':
averageBPM                      1.000000
Arrival Time                    0.139340
Arrival Minute                  0.139340
Arrival Hour                    0.136698
Departure Minute                0.135636
Departure Time                  0.135636
Departure Hour                  0.133644
Type                            0.068374
Time Amount                     0.042804
Odometer Kilometer (km)         0.021773
Distance (km)                   0.013613
Fuel Prices (TRY/l)            -0.036296
Average consumption (l/100km)  -0.059998
DayOfWeek                      -0.178134
```

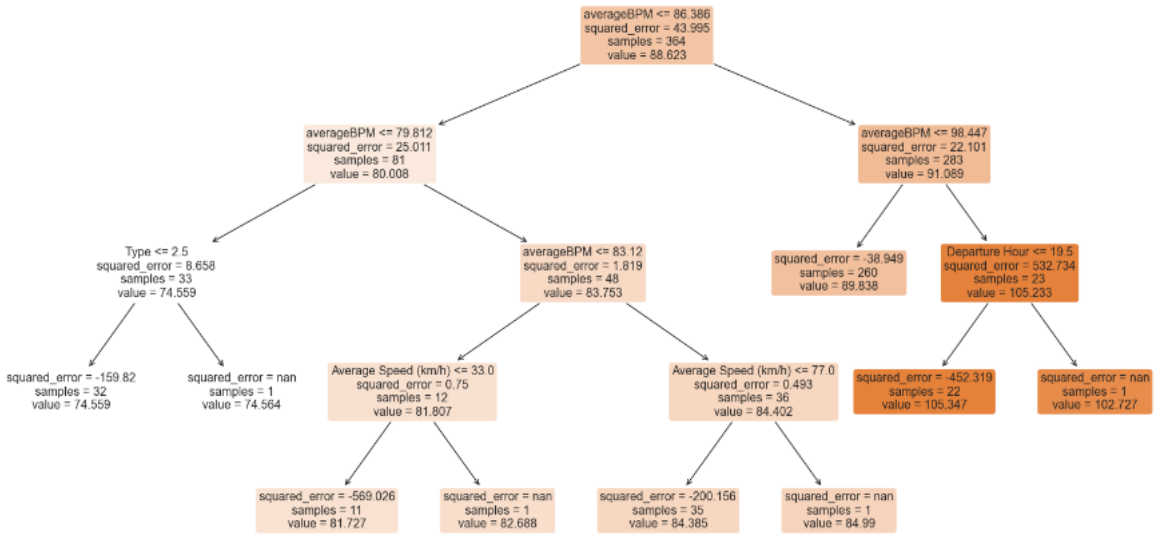
Here, I can see the beats per minute is mainly correlated with the time of the day.

To further investigate this, I created 2 hypothetical features, Time of day category (Morning, Afternoon, Evening and Night), and Weekday vs. Weekend.

The correlation of time of day is: 0.12806860295186653
The correlation of Weekday vs. Weekend is: 0.1731921362748717

Here, again I can see it is similar to the Time values.

After this part, I chose max_depth and min_samples_split parameters to tune for my decision tree. And the plot of the tree is like this:



To conclude the machine learning model, I calculated the MSE: which is 5.45

This value is a measure of how well the Decision Tree Regressor model performs in predicting the average BPM values. But while the MSE provides a measure of the average squared deviation of predictions from the true values, it does not provide an intuitive sense of the scale of the errors.

For this, I also calculated RMSE, which is 2.33. RMSE provides a measure of the average magnitude of the errors in the same units as the target variable (average BPM). Smaller values of RMSE indicate better model performance, as they represent lower average prediction errors. Therefore, an RMSE of 2.33 suggests that, on average, the model's predictions deviate by approximately 2.33 units from the actual average BPM values in the test set.

4. Limitations and Future Work

   a. The main limitation in this project, is the lack of variety in the trip data. As I only had less than 400 entries, the results and the decision tree lacked the proper details. Another limitation is the lack of consistency in the bpm data, since I use apple watch to measure it, I have the data of only for a period, versus the whole trip time. For this problem, proper heart rate monitors may be used.

   b. For the future work, further analysis with the bpm, using the activities section on the apple watch can be used. Also, for the trip data, different whether and speed conditions can also be implemented.

5. Conclusion

   In this project to find the possible correlation between driving culture and the heart health, the results show a little increase in bpm, when the speed of the car increases. To better understand this effect though, one must further analyze the causations and factors that may also be the cause of this. Although I am hesitant to give a direct conclusion, at this point I should express my gratitude on selecting this topic as my research. While doing this process, I have learned new insights about myself, as well as better understood the basics of data science.

   I want to thank you for your time in your interest in my project.