Megan Pokal

Statistical / Hypothetical Questions

DSC 530 -T301 Data Exploration and Analysis (2253-1)

Matthew Metzger

Statistical Analysis of Fitness Data

Hypothesis

Are the number of steps a day significantly correlated with the number of calories spend, as measured by Fitbits?

The analysis indicated a moderate positive correlation between number of steps per day and calories burned with Pearson Correlation Coefficient of 0.59.

Outcome of your EDA.

Pearson Correlation is 0.59, which is a moderate positive correlation. T-Statistic showered 31.83 and P-Value <0.001, which is a significant relationship. Some extreme outliers were identified in total distance with 23 and calories burned with 16, indicating variability in physical activity patterns. The exploratory Data Analysis, EDA, showed the most calories burned fell within a predictive range, with only a few outliers contributing to higher energy expenditure.

What do you feel was missed during the analysis?

There were a variety of lifestyle factors that were not put into the dataset, that could have influenced the data. With more details like age, weight or fitness level, could impact the data

significantly. There is a lack of information of seasonal or temporal data, which might help with defining if the outliers are seasonally related or their normal behavior?

Were there any variables you felt could have helped in the analysis?

Some additional variables that could have made an impact are heart rate, age, dietary and sleep studies. Heart rate could help with the analysis because it is a more direct indicator of energy expenditure. Age can help with can determine a user's physical activity level and help determine one metabolism, which can affect calorie burning. Dietary intake information can better calculate caloric consumption. Sleep duration can contribute to the user's overall energy expenditure.

Were there any assumptions made you felt were incorrect?

From the results of the Durbin-Watson Statistic (0.560), suggested that potential model refinement was needed, and the residuals may not be independent. It could be from the lack of other defining variables that were not included in the study.

What challenges did you face, what did you not fully understand?

In this study, there was a large number of outliers in the variables like Total Steps with 12 and Very Active Minutes with 65 outliers. These outliers may have skewed the data, making the

analysis a little less reliable. I used a single source data, from the Fitbit Tracker, which made the data consistent but could introduce source bias.