Case Study: Investigation of exercise behavior

Jamilah Foucher

July 25, 2023

Contents

1 Project Summary/Executive Summary

The objective of this work was to gain insight into how consumers are using their smart devices.

The Baseline Solution is to statistically understand how each lifestyle group is behaving using five numerical features (mean steps, mean total distance, mean calories, mean heart rate, sleep duration). A two-step method is used to identify two categorical features: lifestyle, type of exercise.

The methodology used to implement the baseline solution, is to use basic statistics to find user trends in smart watch data; kmeans was used to make distinct labels for exercise lifestyle and type of behavior performed.

The code and results for this case study are at
https://github.com/j622amilah/Case_Studies/tree/main/2_case_study_exercise and
https://github.com/j622amilah/GCP_ingestion_analysis_tools. I hope to publish the GCP bigquery statistic library and the GCP bigquery case study library, such that others can use the Google Cloud Platform tools in an automated manner.

2 Business Objective/Ask

The business objective provides a list of problems that need to be solved, with measurable objectives of successful for each solved problem. The first problem should correspond with the last problem (Act), because it shows that the main global problem was indeed solved by the specific problems. The problems that need to be solved for the exercise case study, from general to specific, are :

- Bellabeat needs to gain more insight into how customers use their smart watches. Measurable success for this task is identifying trends and user behavior that could be used for future marketing.
- In order to understand user trends, a lifestyle label was created from existing data via kmeans, using total mean distance. The reliability of a given data collected lifestyle label using mean distance data, was compared with the kmeans label. The most reliable label, that identified lifestyle groups independently, was used for group comparative statistical analysis. Measurable success for this task is to obtain a list ranking weekday and weekend lifestyle groups with respect to the numerical features, and the statistical significance per group for each feature.
- The type of behavior and/or exercise that users performed was investigated via kmeans using four statistically relevant numerical features (mean steps, total mean distance, mean calories, and mean heartrate). Measurable success for this task is to obtain a list ranking type of behavior groups with respect to the numerical features.
- Act: Marketing can use weekday and weekend lifestyle and type of behavior categorical features to market exercise interest related material; sedentary users could be given encouragement to exercise more during .

3 Project Background

Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. They would like to gain insight into how consumers are using their smart devices, such that they can develop new ways to make profits.

4 Baseline Solution

The Baseline Solution (top three recommendations) is to :

1. use mean_total_distance to make a lifestyle label with kmeans
2. identify which numerical features are important with respect to the kmeans lifestyle label
3. use the statistically significant features to predict type_of_exercise

5 Project Deliverables

5.1 Prepare

The data was assumed to have data integrity, credibility, with minimal errors because it was given directly from a trustworthy source (the company). Null values were removed.

5.2 Methodology/Process

An automatic GCP ingestion program was written in bash that:

1. Downloaded the data from Kaggle,
2. Joined a main activity table with a sleep table,
3. Partitioned the joined table by Id and ActivityDate, such that the mean numerical feature values were calculated.
4. Created the categorical lifestyle label using two methods: by hand using the given labels, and using kmeans using mean_total_distance as a feature.
5. Group comparative statistical analysis was used to determine which features were most important for identifying lifestyle groups
6. Created the type_of_exercise label using mean steps, total mean distance, mean calories, and mean heartrate as features

5.3 Analyze

Statistical analysis (one-sample z-score) and one-way ANOVA were used to evaluate the following relationships:

- The quantity of data per lifestyle category: from most to least quantity of data was Sedentary, Moderate_Active, Light_Active, Active. Thus, the majority of people who contributed data were Sedentary and the least number of contributors were Active.
- Validity of the handmade lifestyle label in comparison to the kmeans label: The handmade/manual lifestyle label with given data categories gave conflicting results when groups were compared using numerical features. The manual label created overlapping ambiguous categories, giving non-conclusive results for statistical analysis. Using clustering techniques like kmeans created distinct clusters, thus allowing for group specific trends to be found statistically.
- Lifestyle group means per numerical feature were observed, without statistical comparison. Active people have the most mean steps, mean total distance, and mean calories. Light_Active, Moderate_Active, and Sedentary groups followed in respective order where the most to least statistical significance was observed, for the three features. Heart rate and sleep duration were similar for all groups
- Statistical comparison of lifestyle group means, per corresponding population mean, for each numerical feature. The ratio of the category group means with respect to the population mean was calculated, followed by calculating the probability of occurrence. The z-score was used to calculate which mean probability of occurrence was more or less significant than the other categories. Active weekend users have statistically higher mean_steps than any other lifestyle category (z=1.75). Whereas, Sedentary weekday users have statistically the least number of mean_steps than any other category (z=-1.12). Similar results of means_steps, were found for mean_total_distance and mean_calories. Mean_hr and sleep_duration had no significant differences between lifestyle group means.
- One-way ANOVA of the lifestyle groups showed that is a significant difference between the lifestyle group means.
- Finally, kmeans used the three statistically relevant features means_steps, mean_total_distance, and mean_calories along with mean_hr to predict type_of_exercise. Two main clusters were found: Run/Walk (high steps, high distance, high calories, low hr), and Sedentary/low aerobic (low steps, low distance, low calories, high hr).

6 Recommendations for improving the project in the future

The project can be improved by investigating the data related to exercise and non-exercise, instead of activity behavior averaged per day. We could identify periods of exercise per ActivityDate and average the exercise and non-exercise windows respectively, thus giving a more accurate perspective of group behavior during exercise only. If we treated the time-series data as mentioned, it is likely

that more distinct type_of_exercise groups could be identified like Running, Walking, Gym/Aerobic, Sedentary.