Google Data Analytics Certificate Capstone (Bellabeat Case Study)

Gabriel Fernandez

```
# Set default options for code chunks
knitr::opts_chunk$set(
  echo = FALSE,           # Display R code and its output
  comment=NA,            # Suppress code comments in output
  include = FALSE,       # SUpress code outputs
  warning = FALSE,       # Suppress warning messages
  fig.align='center',    # Align figures in the center
  eval = TRUE            # Evaluate R code
)
```

Prepare and preprocess phase

Load datasets

- Datasets:

These datasets originate from a survey distributed on Amazon Mechanical Turk from 03.12.2016 to 05.12.2016. They include personal tracker data from 30 Fitbit users, covering physical activity, heart rate, and sleep monitoring, with differentiation based on Fitbit types and user behavior.

- Metadata: Fitbit data dictionary

Clean datasets

Clean the daily_activity dataset

Let us clean:

- Change column names to lowercase because R is case-sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "ActivityDate" from char to date.

Observations:

- This summary helps us explore each attribute quickly. We notice that some attributes have a minimum value of zero (total_step, total_distance, calories). Let us explore this observation.

Observations:

- We found 77 observations where "total_steps" equal zero. We should delete these observations so they do not affect our mean and median. If the "total_step" is zero, the person did not wear the Fitbit.

Observations:

- From our inspection above, we can see that we just need to delete the entries where "total_steps" equals zero.

Observations:

- We can see that the observations that we removed affected our mean and median.

Clean the daily_sleep dataset

Let us clean:

- Change column names to lowercase because R is case-sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "SleepDay" from char to date. Since the time component of this column is the same for each observation ("12:00:00 AM"), we can remove it. This will help us merge this dataset with "daily_activity" later.
- Delete duplicates (3 observations are duplicates)

Clean the hourly datasets (hourly_calories, hourly_intensities, and hourly_steps)

Join the hourly datasets into a single dataset named "hourly_activity"

- These datasets shared the same "Id" and "Activity_hour". Let us join them into a new dataset ("hourly_activity") before we clean them.

Let us clean:

- Change column names to lowercase because R is case-sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "ActivityHour" from char to datetime.

Note: The default timezone is UTC.

Clean the minute_sleep dataset

Let us clean:

- Change column names to lowercase because R is case-sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "date" from char to datetime.
- Change "value" from double to factor. This variable indicates the sleep state (1 = asleep, 2 = restless, 3 = awake). For for details see: Fitbit data dictionary
- Remove duplicate values: 543.

Clean the seconds_heartrate dataset

Let us clean:

- Change column names to lowercase because R is case-sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "Time" from char to "datetime" and rename it "date_time".
- Rename "Value" to "heart_rate".

For more details, see: Fitbit data dictionary

Clean the weight_logs dataset

Let us clean:

- Change column names to lowercase because R is case-sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "Date" from char to datetime and rename it "date_time".
- Change NA to 0 in the column "fat".

Distribution of ids across datasets

Observations:

- Differences in the number of unique "IDs" between the datasets can imply discrepancies in data collection methods, data incompleteness, or differing levels of user engagement.

Export clean datasets

Analyze phase: Exploratory data analysis

EDA for daily_activity_clean

Univariate analysis for daily_activity_clean

Numerical variables Observations:

- Many variables show a right-skewed distribution: a larger number of data values are located on the left side of the curve.
- The variables "total_steps", "total_distance", "tracker_distance" have a similar distribution. We can explore their correlations later.
- Since the distributions are not normal. The median is a better indicator of central tendency for the numerical variables in these dataset.
- **The variables 'logged_activities_distance' and 'sedentary_active_distance' may not offer valuable insights since most data points are zero. This suggests that users may not be consistently logging their activities.**
- The following variables seem related. We will explore them further in the bivariate analysis section:

  -"sedentary_minutes"; "sedentary_active_distance"

  - "lightly_active_minutes"; "light_active_distance"
  - "fairly_active_minutes"; "moderately_active_distance"
  - "very_active_minutes"; "very_active_distance"
- The variables "calories" and "sedentary_minutes" exhibit a multimodal distribution, indicating the presence of subpopulations within the data. In this dataset, gender could be a potential variable that would result in a bimodal distribution when examining histograms of calories and sedentary minutes. Unfortunately, the gender of the users is not provided, limiting our ability to confirm this hypothesis.

Categorical variables Observations:

- It appears that there is missing activity data towards the end of the available period, specifically at the beginning of May.

Observations:

- Users with more than 75% of data consistently report activity dates, while those with less than 75% of data show a decline in reporting starting from the end of April. The decline in Activity Date seems to be primarily due to a lack of data reporting from some users during that period.

Bivariate analysis

Correlation between numerical variables Observations:

- "Total_distance", "tracker_distance", and "total_steps" are highly correlated, so we will retain only "total_distance" and "total_steps" as they provide similar information.
- The following minute and distance variables are highly correlated, indicating they represent different aspects of the same activity, whether related to time or distance
  - "lightly_active_minutes" and "light_active_distance" (corr = 0.85)
  - "fairly_active_minutes" and "moderately_active_distance" (corr = 0.94)
  - very_active_minutes" and "very_active_distance" (corr = 0.82)
- There is a moderately high correlation between the time spent during very active periods and the total number of steps/total distance:
  - The correlation between "very_active_minutes" and "total_distance" is 0.68
  - The correlation between "very_active_minutes" and "total_steps" is 0.66
- There is a moderate correlation of 0.61 between the total duration of very active minutes (very_active_minutes) and the estimated daily calories consumed.
- There is a moderate correlation of 0.62 between the total distance covered and the estimated daily calories consumed.
- There is a moderate correlation coefficient of 0.60 between the distance covered during light activity ("light_active_distance") and the total number of steps taken ("total_steps").

Scatterplots of selected highly correlated variable pairs (>0.60)

User Behavior for the daily activity dataset

Total steps: Total number of steps taken.

Total Distance: Total kilometers tracked.

Sedentary Minutes: Total minutes spent in sedentary activity. Observations:

- These are high values for sedentary minutes. For instance, 1020 minutes equals 17 hours, and 1400 minutes equals 24 hours. After performing a quick search, it seems that the Fitbit uses 1400 a default, sedentary minutes are recorded when the device is not worn, including sleep time.

"SedentaryMinutes" is total minutes spent in sedentary activity according to the data dictionary. See meta data section. Therefore, we need to substract the times sleeping to obtain an more accurate estimate of daily sedentary minutes.

"Sleep time is not considered sedentary time, so it was removed to determine the waking day and to allow the proportion of the day spent sedentary to be calculated." [Reference 1]

Observations:

- The sedentary percentage difference has a median value of 59.95%, indicating a significant distinction between "sedentary_minutes" and "sedentary_min_awake". This suggests that the original column "sedentary_minutes" included the time asleep.

Observation:

- Data inconsistency alert: "sedentary_minutes" is less than "total_minutes_asleep", contrary to our expectations.

Observations:

- By eliminating negative values from "sedentary_min_awake," the resulting values now reflect a more realistic scenario.

Observations:

In a representative sample of U.S. adults, over two-thirds spent 6 + hours/day sitting, and more than half did not meet the recommended 150 min/week of physical activity. The study discovered that prolonged sitting for 6+ hours/day was associated with higher body fat percentages. While exceeding 150 min/week of physical activity was linked to lower body fat percentages, achieving recommended activity levels may not fully offset the increased body fat from prolonged sitting [Reference 2].

Calories: Total estimated energy expenditure (in kilocalories). Observations:

"Females ages 19 through 30 require about 1,800 to 2,400 calories a day. Males in this age group have higher calorie needs of about 2,400 to 3,000 a day. Calorie needs for adults ages 31 through 59 are generally lower; most females require about 1,600 to 2,200 calories a day and males require about 2,200 to 3,000 calories a day."[Reference 3]

Intensity Minutes: Time spent in one of four intensity categories.

- "VeryActiveMinutes": Total minutes spent in very active activity.
- "FairlyActiveMinutes": Total minutes spent in moderate activity.
- "LightlyActiveMinutes": Total minutes spent in light activity.
- "SedentaryMinutes": Total minutes spent in sedentary activity.

Observations:

- Users' overall average intensity minutes consist primarily of sedentary and lightly active time, comprising 97%.
- We can use average user activity patterns to develop indicators to provide insights into activity levels. They can help track progress, set goals, and evaluate user behavior over time.
- This is a concern since the Physical Activity Guidelines for Americans recommend 150 minutes of moderate-intensity or 75 minutes of vigorous-intensity aerobic activity weekly, along with muscle-strengthening exercises for adults. Additionally, it's advised to reduce sedentary time and break up long periods of inactivity.[Reference 4]

EDA for daily_sleep_clean

- activity_date (sleep_day): Date on which the sleep event started.
- total_sleep_records: Number of recorded sleep periods for that day. Includes naps > 60 min.
- total_minutes_asleep: Total number of minutes classified as being "asleep".
- total_time_in_bed: Total minutes spent in bed, including asleep, restless, and awake, that occurred during a defined sleep record.

Univariate analysis

Bivariate analysis

Correlation between numerical variables

Scatterplots of total_minutes_asleep vs total_time_in_bed

User Behavior for daily_sleep dataset

Total minutes asleep

Sleep duration consistency Observations:

- Regular sleepers have higher median sleep hours than irregular sleepers, indicating they get more sleep on average.
- Additionally, the "average_sleep_hours" spread for irregular sleepers appears to be wider, indicating more variability in their sleep duration. In contrast, the violin plot for regular sleepers shows a narrower spread, suggesting that their sleep duration is more consistent.
- Regular sleepers exhibit a slightly higher median average awake-in-bed duration compared to irregular sleepers.

Summary: Regular sleepers get more sleep on average, have a more consistent sleep duration, and slightly higher median awake-in-bed duration than irregular sleepers.

EDA minute_sleep_clean

This data seems to come from the Classic Sleep Log (1 minute)

Value indicating the sleep state. 1 = asleep, 2 = restless, 3 = awake

For more detail, check : Fitbit data dictionary

EDA for hourly_activity_clean

Observations:

- "Calories": Total number of estimated calories burned.
- "TotalIntensity": Value calculated by adding all the minute-level intensity values that occurred within the hour.
- "AverageIntensity": intensity state exhibited during that hour ("TotalIntensity" for that "ActivityHour" divided by 60).
- "StepTotal": Total number of steps taken.

For more detail, check : Fitbit data dictionary

Observations:

- Intensity: On average, users engage more actively at 5:00 AM, 8:00 AM, 5:00 PM, and 7:00 PM.
- Step Count: On average, users record more steps at 8:00 AM and 7:00 PM.
- These observations suggest that user activity may be influenced by their daily routines and responsibilities, with higher activity levels before or after typical workday hours.

EDA for seconds_heartrate_clean

Observations:

- Users' average heart rate is within the normal range. Nothing remarkable here.
- We could suggest adding a feature to the app that sends notifications to users whose average resting heart rate falls outside the normal range.

EDA for weight_logs_clean

- Fat: Body fat percentage recorded.
- "BMI": Measure of body mass index based on the height and weight in the participant's Fitbit.com profile.
- "isManualReport": If the data for this weigh-in was done manually (TRUE), or if data was measured and synced directly to Fitbit.com from a connected scale (FALSE).

For more detail, check : Fitbit data dictionary

Observation:

- Only two users reported fat percentage.

Observation:

- 61% of users log their weight manually, while 39% sync their weight from other devices.

Observations:

- It appears that users who log their weight data manually have a lower median weight than users who sync their weight from other devices.
- The previous observation should be viewed as exploratory and could benefit from additional data. The weight log dataset only has 68 entries; more data would be needed to evaluate these hunches.
- The lack of completeness in the weight log dataset could indicate a lack of user engagement.

References

Guidelines and research articles

- 1. Handling sedentary time: A Comparison of Sedentary Behavior as Measured by the Fitbit and ActivPAL in College Students
- 2. Danger of prolonged sitting(sedentary time): Association of daily sitting time and leisure-time physical activity with body fat among U.S. adults. Journal of Sport and Health Science
- 3. Dietary Guidelines for Americans, 2020-2025
- 4. Physical Activity Guidelines for Americans (2nd ed.)

Links

- Projects Datasets:
- EDA guide
- Metadata: Fitbit data dictionary
- Plotting histograms with ggplot2
- Histograms article
- Error bars vs CI
- Add density line to histogram
- Categorical, ordinal, and interval variables

Appendix: Interesting sites for further investigation

- Adult Physical Inactivity Prevalence Maps by Race/Ethnicity
- Physical activity among adults aged 18 and over: United States, 2020