REPORT ON SQL DATA CLEANING

- The dataset contains **15 different tables** categorized into daily, hourly, minute-level, and health-related data.
- In total, **5.3+ million records** were processed for a **31-day period (April 12 – May 12, 2016)** across **33 unique users**.
- After cleaning, all tables were validated with **no NULL values or duplicates**, except in sleep and weight tables where specific issues were fixed or documented.
- The data integrity and alignment across tables make it highly reliable for business insights and advanced analytics.

DAILY TABLES

1. daily_activity

- No NULL values or duplicates.
- Covers all **33 users**, with an average of **28.5 days tracked** out of 31 (92% engagement).
- On average, users take **7,638 steps/day** and burn **2,304 calories/day**.
- Business Meaning: Strong user consistency and engagement with the app, but occasional zero-activity days hint at device non-usage periods.

2. daily_calories

- Clean and perfectly consistent with daily_activity.
- Calorie range is **0–4900 per day**, averaging **2304 calories**.
- Same **31-day coverage** and user participation.
- Business Meaning: Confirms calorie burn data is reliable and aligned across tables.

3. daily_intensities

- Data is valid, no duplicates.
- **Highly sedentary behavior**: average **16.5 hours/day** sedentary time.
- **Very low activity**: just **21 minutes/day** of very active time.
- Business Meaning: A major opportunity exists to encourage users to move more frequently and reduce sitting time.

4. daily_steps

- Perfectly clean, no duplicates.
- Steps data is **fully aligned** with daily_activity (0–36,019 steps/day, avg 7,638).
- Business Meaning: Confirms consistency of tracking across daily-level tables.

HOURLY TABLES

5. hourly_calories

- No NULLs or duplicates.
- Users burn **42–948 calories per hour**, averaging **97 cal/hour**.
- Data covers all users and all 31 days.
- Business Meaning: Provides granular view of energy expenditure, useful for **time-based analytics** like identifying peak calorie burn times.

6. hourly_intensities

- Clean, no duplicates.
- Average total intensity is **12**, with average intensity around **0.20**.
- Most hours are sedentary, with only a few high-intensity spikes.
- Business Meaning: Insights into daily rhythm show **long inactivity periods**, which could be used for personalized nudges.

7. hourly_steps

- No NULLs or duplicates.
- Steps range **0–10,554 per hour**, average **320 steps/hour**.
- Business Meaning: Allows identification of **peak walking hours**, which can inform workout scheduling and engagement campaigns.

MINUTE TABLES

8. minute_calories_narrow

- Clean, no duplicates.
- Burn rate averages **1.62 cal/minute (≈97/hour)**.
- Full coverage for all 31 days and 33 users.

- Business Meaning: Enables **real-time tracking of metabolism** for advanced coaching features.

9. minute_intensities_narrow

- Clean and consistent.
- Average intensity is **0.20 (scale 0–3)**, showing most minutes are sedentary.
- Business Meaning: Can power **instant notifications** to prompt activity when inactivity is prolonged.

10. minute_mets_narrow

- Clean dataset.
- Wide metabolic range (0–157), average **14.7 METs**.
- Captures both rest and high-performance exercise.
- Business Meaning: Useful for **personalized training recommendations** and improving calorie burn accuracy.

11. minute_steps_narrow

- Clean, no duplicates.
- Average **5.34 steps/minute (≈320/hour)**.
- Full coverage across all users.
- Business Meaning: Detects **short bursts of activity** and helps track sedentary breaks throughout the day.

HEALTH TABLES

12. heartrate_seconds

- No NULLs or duplicates.
- Heart rate ranges **36–203 bpm**, average **77 bpm**.
- Only **14/33 users (42%)** recorded heart rate data.
- Business Meaning: Opportunity for Strava to promote **wearable device adoption** and premium features based on heart rate insights.

13. minute_sleep

- Initially had **543 duplicates**, which were cleaned using ROW_NUMBER().

- After cleaning, dataset is valid and reliable.
- **24/33 users (73%)** tracked sleep.
- Business Meaning: Opportunity to improve adoption by educating users on benefits of sleep tracking.

14. sleep_day

- Found **3 duplicate records**, removed successfully.
- Matches minute_sleep for coverage.
- Business Meaning: Reliable after cleaning, can support daily-level sleep analysis.

15. weight_log_info

- No duplicates, but **97% missing values in fat% column**.
- Only **8/33 users (24%)** logged weight.
- Business Meaning: Very low adoption – strong potential for Strava to expand into **weight and body composition tracking features**.

INSIGHTS

1. **High Engagement**
    - 92% daily activity tracking – users are consistent.
2. **Low Adoption Gaps**
    - Heart rate: 42% users.
    - Sleep: 73% users.
    - Weight: 24% users.
3. **Lifestyle Patterns**
    - Highly sedentary: 16.5 hrs/day sitting.
    - Very low activity: only 21 minutes/day very active.