SQL ANALYSIS REPORT

1. Project Title

Strava Fitness Data Analysis - Strategic Business Intelligence Through SQL

2. SQL Objective

Data Cleaning, Transformation, and Exploratory Analysis

3. Tools Used

- **PostgreSQL & pgAdmin4**: Database Management & Query Execution
- **VSCode**: Code Writing & Version Control

4. Workflow

• Database Connection Setup

Established PostgreSQL connection through pgAdmin4 with dedicated 'strava' connection using postgres authentication and port 5432.

• Database Creation

Created project-specific 'strava' database using CREATE DATABASE command to isolate fitness tracking data analysis.

• Table Design and Schema Creation

Designed star schema with 15 tables using composite primary keys (id, date/time) and foreign key relationships connecting all tables via user ID for optimal join performance.

• Data Import via pgAdmin4's PSQL Tool

Loaded 15 CSV datasets using \copy commands through PSQL Tool with UTF8 encoding and comma delimiters, handling over 5.3 million total records.

• Data Cleaning, Validation & QA

Implemented systematic NULL validation, duplicate removal using ROW_NUMBER() window functions, and data type standardization across time-series fitness data.

• Analytical Questions

1. What are the overall user activity patterns and engagement levels across the platform?
2. How do the most active users differ from the least active users in behavior and consistency?
3. What percentage of users achieve daily step goals and what drives success vs failure?

4. When are users most active throughout the day for optimal campaign timing?
5. How do weekend activity patterns compare to weekday patterns for strategic planning?
6. How many users track sleep and what's their average sleep duration?

5. Key Business Insights and Analysis

A. Data Cleaning, Validation & QA Summary

Validation Status: Data successfully validated against expected schema. All 15 tables (83% of available datasets) loaded successfully with proper data types and relationships established.

Summary of Cleaning Steps: Removed 546 duplicate sleep records using ROW_NUMBER() partitioning, handled empty string values in weight data using NULL parameters, standardized date formats using datestyle settings.

Key Issues Resolved: Sleep data duplicates (minute_sleep: 543 removed, sleep_day: 3 removed), weight data missing body fat values (97% NULL), column naming inconsistencies (logid vs log_id standardization).

B. Exploratory Data Analysis (EDA)

Question 1: What are the overall user activity patterns and engagement levels across the platform?

Thought Process: I started with the daily_activity table since it contains all key metrics in one place. Used COUNT(DISTINCT id) to get unique users and AVG() functions to calculate platform-wide averages for steps, calories, and activity minutes.

SQL Query:

```sql
SELECT
    COUNT(DISTINCT id) as total_users,
    ROUND(AVG(total_steps)) as avg_daily_steps,
    ROUND(AVG(calories)) as avg_daily_calories,
    ROUND(AVG(very_active_minutes)) as avg_very_active_minutes,
    ROUND(AVG(sedentary_minutes)) as avg_sedentary_minutes
FROM daily_activity;
```

Insights:

- **Step Goal Gap**: Average 7,638 steps vs 10K target represents 24% shortfall, indicating major opportunity for motivation features
- **Sedentary Crisis**: 991 minutes (16.5 hours) sedentary daily reveals users need intervention for break reminders and activity prompts

- **Low Intensity Problem**: Only 35 minutes combined active time shows users require coaching for more vigorous exercise engagement

Question 2: How do the most active users differ from the least active users in behavior and consistency?

Thought Process: I used GROUP BY id to aggregate per user, then applied ORDER BY with LIMIT to get extremes. Included COUNT(*) to check tracking consistency alongside step averages to understand both activity and engagement patterns.

SQL Query:

```sql
-- Top 5 Most Active Users
SELECT
    id as user_id,
    ROUND(AVG(total_steps)) as avg_daily_steps,
    ROUND(AVG(calories)) as avg_daily_calories,
    COUNT(*) as tracking_days
FROM daily_activity
GROUP BY id
ORDER BY avg_daily_steps DESC
LIMIT 5;

-- Bottom 5 Least Active Users
SELECT
    id as user_id,
    ROUND(AVG(total_steps)) as avg_daily_steps,
    ROUND(AVG(calories)) as avg_daily_calories,
    COUNT(*) as tracking_days
FROM daily_activity
GROUP BY id
ORDER BY avg_daily_steps ASC
LIMIT 5;
```

Insights:

- **Power User Profile**: Top user averages 16K steps (60% above 10K goal) with perfect 31-day tracking consistency, representing ideal user archetype
- **Calorie Burn Leaders**: Most active users burn 1,816-3,420 calories daily vs 2,304 average, indicating significantly higher metabolic activity
- **Engagement Paradox**: Low-step users (916-2,580 avg) still maintain 26-31 day tracking consistency, showing platform engagement despite low activity

Question 3: What percentage of users achieve daily step goals and what drives success vs failure?

Thought Process: I used CASE WHEN with conditional aggregation to count days meeting 10K steps. Applied the same logic with DISTINCT to count unique users. Used nested calculations to get percentages for both daily success rate and user capability.

SQL Query:

```sql
SELECT
    COUNT(*) as total_activity_days,
    COUNT(CASE WHEN total_steps >= 10000 THEN 1 END) as days_with_10k_steps,
    ROUND(COUNT(CASE WHEN total_steps >= 10000 THEN 1 END) * 100.0 / COUNT(*), 1) as percentage
    COUNT(DISTINCT CASE WHEN total_steps >= 10000 THEN id END) as users_who_hit_10k,
    COUNT(DISTINCT id) as total_users,
    ROUND(COUNT(DISTINCT CASE WHEN total_steps >= 10000 THEN id END) * 100.0 / COUNT(DISTINCT i
FROM daily_activity;
```

Insights:

- **Goal Achievement Gap**: Only 32.2% of days meet 10K step target, meaning 68% of days fall short, representing massive motivation opportunity
- **User Capability**: 75.8% of users (25/33) have achieved 10K at least once, proving most users can reach goals but lack consistency
- **Inconsistency Problem**: Users hit goals only 1 in 3 days on average, making daily motivation and streak features critical for improvement

Question 4: When are users most active throughout the day for optimal campaign timing?

Thought Process: I joined hourly_calories and hourly_steps tables on id and activity_hour. Used EXTRACT(HOUR) to get hour numbers from timestamps, then GROUP BY hour and ORDER BY calories to find peak activity periods.

SQL Query:

```sql
SELECT
    EXTRACT(HOUR FROM h.activity_hour) as hour_of_day,
    ROUND(AVG(h.calories)) as avg_calories_per_hour,
    ROUND(AVG(s.step_total)) as avg_steps_per_hour
FROM hourly_calories h
JOIN hourly_steps s ON h.id = s.id AND h.activity_hour = s.activity_hour
GROUP BY EXTRACT(HOUR FROM h.activity_hour)
ORDER BY avg_calories_per_hour DESC
LIMIT 10;
```

Insights:

- **Evening Peak**: 5-7 PM (hours 17-19) show highest activity with 123 avg calories/hour, representing prime time for workout notifications and social challenges
- **Lunch Break Activity**: 12-2 PM demonstrates strong midday activity (115-117 calories/hour), ideal for targeting office workers with movement prompts
- **After Work Surge**: 6 PM absolute peak at 599 steps/hour indicates optimal timing for group challenges and community engagement features

Question 5: How do weekend activity patterns compare to weekday patterns for strategic planning?

Thought Process: I used EXTRACT(DOW) to get day of week numbers, then applied CASE WHEN to classify Saturday/Sunday (0,6) as weekends vs weekdays (1-5). Grouped by this classification and compared averages for steps, calories, and activity minutes.

SQL Query:

```sql
SELECT
    CASE
        WHEN EXTRACT(DOW FROM activity_date) IN (0, 6) THEN 'Weekend'
        ELSE 'Weekday'
    END as day_type,
    COUNT(*) as total_days,
    ROUND(AVG(total_steps)) as avg_steps,
    ROUND(AVG(calories)) as avg_calories,
    ROUND(AVG(very_active_minutes)) as avg_very_active_minutes
FROM daily_activity
GROUP BY CASE
    WHEN EXTRACT(DOW FROM activity_date) IN (0, 6) THEN 'Weekend'
    ELSE 'Weekday'
END
ORDER BY avg_steps DESC;
```

Insights:

- **Minimal Activity Difference**: Only 118 steps more on weekdays (7,669 vs 7,551) indicates users maintain consistent activity patterns regardless of day type
- **Weekend Calorie Spike**: Slightly higher weekend calories (2,310 vs 2,302) despite fewer steps suggests longer duration, leisure-based activities
- **Consistent Intensity**: Identical very active minutes (21) across both day types shows users aren't "weekend warriors" but maintain steady moderate activity levels

Question 6: How many users track sleep and what's their average sleep duration?

Thought Process: I used the sleep_day table with COUNT(DISTINCT id) to find sleep tracking users, then created a subquery to get total users from daily_activity for percentage calculation. Applied AVG() to sleep duration and converted minutes to hours for readability.

SQL Query:

```sql
SELECT
    COUNT(DISTINCT id) as sleep_tracking_users,
    (SELECT COUNT(DISTINCT id) FROM daily_activity) as total_users,
    ROUND(COUNT(DISTINCT id) * 100.0 / (SELECT COUNT(DISTINCT id) FROM daily_activity), 1) as s
    ROUND(AVG(total_minutes_asleep)) as avg_minutes_asleep,
    ROUND(AVG(total_minutes_asleep) / 60.0, 1) as avg_hours_asleep,
    ROUND(AVG(total_time_in_bed)) as avg_minutes_in_bed,
    ROUND(AVG(total_time_in_bed) / 60.0, 1) as avg_hours_in_bed
FROM sleep_day;
```

Insights:

- **Good Sleep Adoption**: 72.7% of users track sleep (24/33) showing strong feature adoption compared to other health metrics
- **Healthy Sleep Duration**: 7.0 hours average sleep meets recommended guidelines, indicating users prioritize good sleep habits
- **Sleep Efficiency Gap**: 39 minutes difference between bed time and sleep time reveals opportunity for sleep optimization coaching features

6. Conclusion

The SQL analysis reveals Strava users demonstrate strong platform engagement with 92% daily tracking consistency (28.5 out of 31 days average), indicating solid product-market fit. However, significant opportunities exist in user motivation systems, with 68% of days missing step goals despite 76% of users proving capable of achieving 10K steps.

Trends Discovered:

- Peak activity occurs during 5-7 PM window, providing optimal timing for engagement campaigns
- User behavior remains consistent across weekdays and weekends, enabling universal campaign strategies
- Clear user segmentation emerges between power users (16K+ steps) and coaching-needed users (under 5K steps)

Key Outliers and Patterns:

- Extreme sedentary behavior averages 16.5 hours daily across all user types
- Sleep tracking shows strong adoption (73%) compared to heart rate (42%) and weight tracking (24%)
- Activity consistency exceeds step goal achievement, indicating engagement without performance optimization

Business Implications: The analysis identifies three strategic opportunities: implementing evening-timed motivation campaigns leveraging the 5-7 PM peak activity window, developing tiered user experiences for power users versus coaching-needed segments, and expanding health feature adoption through the proven engagement foundation. The data foundation supports immediate campaign optimization while revealing longer-term opportunities in health tracking expansion and advanced user segmentation strategies.