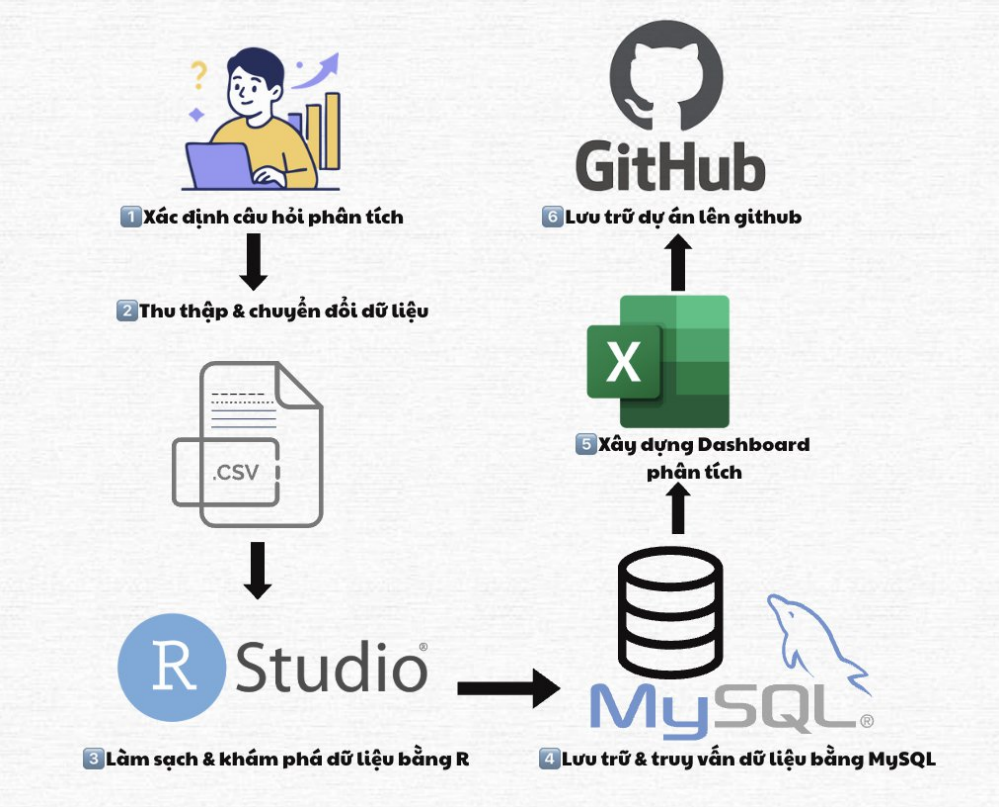
PERSONAL HEALTH DATA ANALYSIS PROJECT

Mô tả tổng quan dự án



1. Tổng quan về dự án

1.1. Nguồn gốc dữ liệu

Dữ liệu trong dự án này được thu thập từ nhiều thiết bị cá nhân trong giai đoạn dài hạn:

- Huawei Smart Band 10 (đeo liên tục cả ngày và ban đêm trong tháng 7/2025)
- iPhone13 ProMax
- iPhone X

Thời gian thu thập dữ liệu: từ 01/2021 đến 30/11/2025

Trong đó:

- Dữ liệu từ Smart Band bao gồm nhiều chỉ số sức khỏe chi tiết như nhịp tim, năng lượng tiêu hao và giấc ngủ.
- Dữ liệu từ iPhone13 ProMax và iPhone X chủ yếu bao gồm:
  - Step Count
  - Ngày ghi nhận (recorded date)

Điều này giúp tạo ra một tập dữ liệu có tính liên tục trong nhiều năm, đặc biệt tập trung mạnh vào Step Count – chỉ số xuyên suốt có mặt ở tất cả các thiết bị.

Dữ liệu được export từ ứng dụng Health của Iphone dưới dạng file thô (raw data) bao gồm: File nén, File XML.

1.2. Mục tiêu dự án

Dự án được thực hiện nhằm phân tích và đánh giá xu hướng vận động và sức khỏe cá nhân trong dài hạn.

Các mục tiêu chính bao gồm:

- Phân tích thói quen vận động cá nhân từ năm 2021 đến 2025
- Nhấn mạnh vào **Step Count** như một chỉ số trung tâm
- Phân tích mối quan hệ giữa:
  - Vận động và Resting Heart Rate (RHR)
  - Vận động và Active Energy Burned
  - Hoạt động cuối tuần và ngày thường
- Đánh giá xu hướng sức khỏe theo thời gian (cải thiện, suy giảm hoặc ổn định)
- Xây dựng dashboard trực quan để theo dõi sức khỏe dài hạn

Đây là một Personal Health Data Case Study, gồm những bước sau:

- Xác định câu hỏi từ dữ liệu để hiểu về sức khoẻ
- Trích xuất dữ liệu thô từ điên thoại và đồng hồ
- Làm sạch và phân tích bằng ngôn ngữ R
- Lưu trữ và truy vấn bằng SQL
- Xây dựng Dashboard bằng Excel
- Lưu trữ và công khai dự án lên Github

1.3. Công cụ sử dụng

Tool	Vai trò
SQL (MySQL)	Làm sạch dữ liệu, xử lý, lọc step cuối tuần, join dữ liệu từ nhiều thiết bị
R	Phân tích xu hướng, trực quan hóa dữ liệu, kiểm tra tương quan
Excel	Pivot Table, Dashboard, tổng hợp báo cáo trực quan
Github	Lưu trữ và công khai dự án

2. Mô tả dữ liệu

2.1. Quy mô dữ liệu

- Tổng số dòng (Rows): 317360 (dòng)
- Tổng số cột (Columns): 8–9 (cột chính)
- Giai đoạn dữ liệu: 01/2021 – 12/2025
- Thiết bị ghi nhận: Smart Band + iPhone 13 Pro Max + iPhone X

2.2. Dữ liệu chính (đáng tin cậy)

- **Step Count (**Số bước đi mỗi ngày là biến trung tâm của dự án vì) :
  - Có mặt xuyên suốt từ 2021 đến 2025
  - Được ghi nhận bởi cả Smart Band và iPhone
  - Đại diện trực tiếp cho mức độ vận động
- **Active Energy Burned**
  Lượng năng lượng tiêu hao trong quá trình vận động (kcal).
  Dùng để đánh giá cường độ hoạt động.

- **Heart Rate :** Nhịp tim đo theo thời gian thực. Giúp phân tích cường độ hoạt động trong ngày.
- **Resting Heart Rate (RHR) :** Nhịp tim khi nghỉ ngơi – một chỉ số quan trọng phản ánh tình trạng tim mạch và thể lực.
- **Distance Walking/Running :** Quãng đường di chuyển hằng ngày.

3. Phân tích dữ liệu bằng Rstudio và MySQL

3.1. Làm sạch và khám phá dữ liệu qua Rstudio

3.1.1 Import data

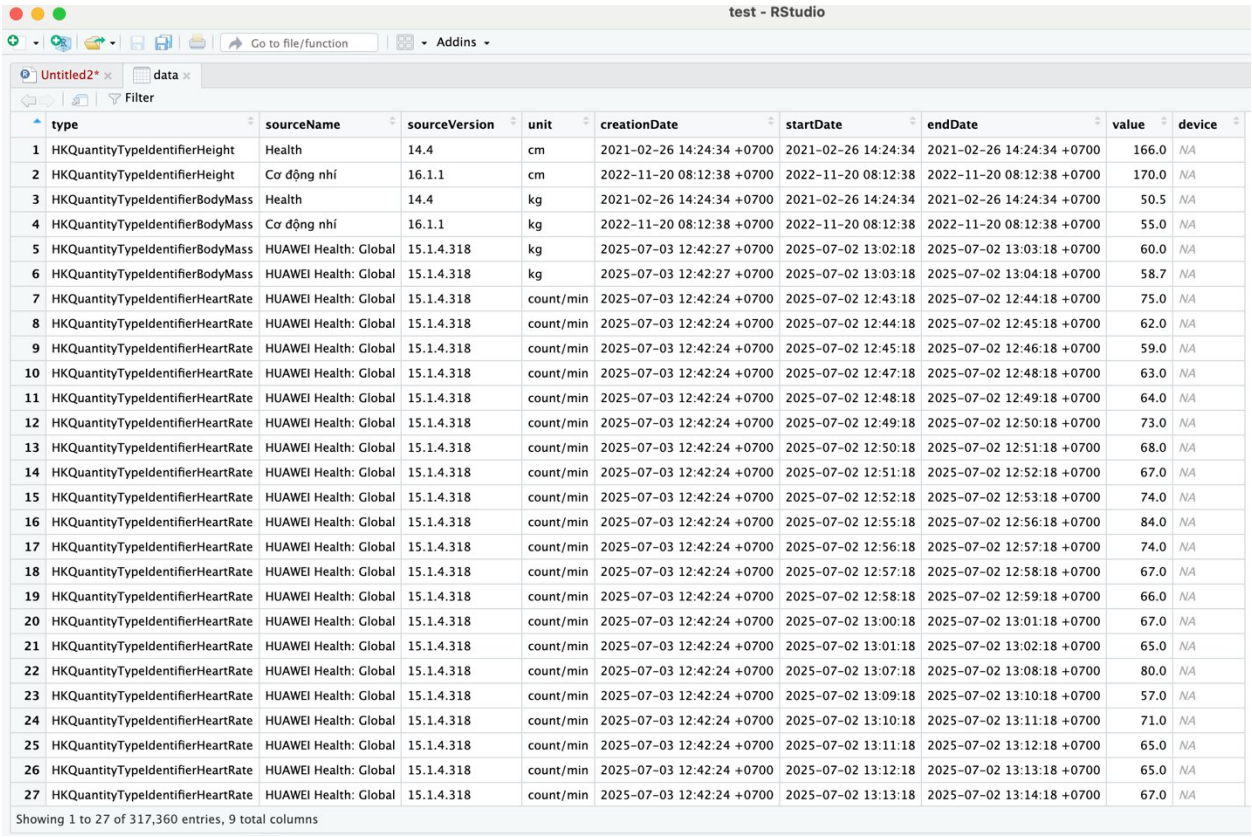
- **Quá trình tải file XML vào Rstudio :**

Code R :

```r
# Install packages
installed.packages("tidyverse")
library(tidyverse)
install.packages(c("xml2", "dplyr", "purrr"))
library(xml2)
library(dplyr)
library(purrr)
install.packages("ggplot2")
library(ggplot2)
library(lubridate)


# Đọc file XML
file_path <- file.choose()  # Thay đường dẫn tới file XML của bạn
doc <- read_xml(file_path)
# Lấy tất cả thẻ <Record> (dữ liệu chính của Apple Health)
records <- xml_find_all(doc, ".//Record")
# Chuyển các attributes của mỗi thẻ <Record> thành danh sách, rồi map sang data frame
data <- map_df(records, function(node) {
  as.list(xml_attrs(node))
})
# Xem data frame trong RStudio
View(data)
# Xem tất cả các giá trị trong 1 cột
unique(data$type)
rm(a,b)
save.image("Workspace_data.RData")
```

Bảng dữ liệu thô được tải vào :



	type	sourceName	sourceVersion	unit	creationDate	startDate	endDate	value	device
1	HKQuantityTypeIdentifierHeight	Health	14.4	cm	2021-02-26 14:24:34 +0700	2021-02-26 14:24:34	2021-02-26 14:24:34 +0700	166.0	NA
2	HKQuantityTypeIdentifierHeight	Cơ động nhí	16.1.1	cm	2022-11-20 08:12:38 +0700	2022-11-20 08:12:38	2022-11-20 08:12:38 +0700	170.0	NA
3	HKQuantityTypeIdentifierBodyMass	Health	14.4	kg	2021-02-26 14:24:34 +0700	2021-02-26 14:24:34	2021-02-26 14:24:34 +0700	50.5	NA
4	HKQuantityTypeIdentifierBodyMass	Cơ động nhí	16.1.1	kg	2022-11-20 08:12:38 +0700	2022-11-20 08:12:38	2022-11-20 08:12:38 +0700	55.0	NA
5	HKQuantityTypeIdentifierBodyMass	HUAWEI Health: Global	15.1.4.318	kg	2025-07-03 12:42:27 +0700	2025-07-02 13:02:18	2025-07-02 13:03:18 +0700	60.0	NA
6	HKQuantityTypeIdentifierBodyMass	HUAWEI Health: Global	15.1.4.318	kg	2025-07-03 12:42:27 +0700	2025-07-02 13:03:18	2025-07-02 13:04:18 +0700	58.7	NA
7	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:43:18	2025-07-02 12:44:18 +0700	75.0	NA
8	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:44:18	2025-07-02 12:45:18 +0700	62.0	NA
9	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:45:18	2025-07-02 12:46:18 +0700	59.0	NA
10	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:47:18	2025-07-02 12:48:18 +0700	63.0	NA
11	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:48:18	2025-07-02 12:49:18 +0700	64.0	NA
12	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:49:18	2025-07-02 12:50:18 +0700	73.0	NA
13	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:50:18	2025-07-02 12:51:18 +0700	68.0	NA
14	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:51:18	2025-07-02 12:52:18 +0700	67.0	NA
15	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:52:18	2025-07-02 12:53:18 +0700	74.0	NA
16	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:55:18	2025-07-02 12:56:18 +0700	84.0	NA
17	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:56:18	2025-07-02 12:57:18 +0700	74.0	NA
18	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:57:18	2025-07-02 12:58:18 +0700	67.0	NA
19	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 12:58:18	2025-07-02 12:59:18 +0700	66.0	NA
20	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:00:18	2025-07-02 13:01:18 +0700	67.0	NA
21	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:01:18	2025-07-02 13:02:18 +0700	65.0	NA
22	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:07:18	2025-07-02 13:08:18 +0700	80.0	NA
23	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:09:18	2025-07-02 13:10:18 +0700	57.0	NA
24	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:10:18	2025-07-02 13:11:18 +0700	71.0	NA
25	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:11:18	2025-07-02 13:12:18 +0700	65.0	NA
26	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:12:18	2025-07-02 13:13:18 +0700	65.0	NA
27	HKQuantityTypeIdentifierHeartRate	HUAWEI Health: Global	15.1.4.318	count/min	2025-07-03 12:42:24 +0700	2025-07-02 13:13:18	2025-07-02 13:14:18 +0700	67.0	NA

Showing 1 to 27 of 317,360 entries, 9 total columns

Bảng dữ liệu chính (data) bao gồm cấu trúc:

Column	Description
type	Loại chỉ số sức khỏe (Step, Heart Rate, Energy…)
sourceName	Tên thiết bị ghi nhận
sourceVersion	Phiên bản thiết bị
unit	Đơn vị đo
creationDate	Ngày tạo bản ghi
startDate	Thời điểm bắt đầu
endDate	Thời điểm kết thúc
value	Giá trị đo
device	Thiết bị ghi nhận

Mô tả chi tiết cột type : Height, BodyMass, HeartRate, StepCount, WalkingRunningDistance, BasalEnergy, ActiveEnergy, FlightsClimbed, RestingHeartRate, VO2Max, AudioExposure, WalkingDoubleSupport, WalkingSpeed, StepLength, WalkingAsymmetry, SleepGoal, WalkingSteadiness, SleepAnalysis

Dữ liệu tin cậy : HeartRate, StepCount, WalkingRunningDistance, ActiveEnergy, SleepAnalysis

3.1.2 Làm sạch dữ liệu

- **Chuyển Date về đúng định dạng và tạo bảng energy daily và sleep daily :**

Code R :

```
1  # Chuyển cột ngày tạo và ngày kết thúc về đúng định dạng ngày giờ :
```

```r
clean_data$startDate <- as.POSIXct(clean_data$startDate)
clean_data$endDate   <- as.POSIXct(clean_data$endDate)
# Gộp năng lượng vận động theo ngày
energy_daily <- clean_data %>%
  filter(type == "HKQuantityTypeIdentifierActiveEnergyBurned") %>%
  mutate(date = as.Date(startDate)) %>%
  group_by(date) %>%
  summarise(
    active_energy = sum(value, na.rm = TRUE)
  )
# Gộp thời gian ngủ theo ngày
sleep_daily <- clean_data %>%
  filter(type == "HKCategoryTypeIdentifierSleepAnalysis") %>%
  mutate(
    date = as.Date(startDate),
    sleep_hours = as.numeric(difftime(endDate, startDate, units = "hours"))
  ) %>%
  group_by(date) %>%
  summarise(
    total_sleep = sum(sleep_hours, na.rm = TRUE)
  )
```

	date	total_sleep
1	2021-02-26	5.7177778
2	2021-03-20	1.7202778
3	2021-03-21	6.6033333
4	2021-03-22	0.7238889
5	2021-03-23	6.4291667
6	2021-03-25	7.2288889
7	2025-07-02	7.1833333
8	2025-07-03	8.9500000
9	2025-07-04	7.0000000
10	2025-07-05	6.8333333
11	2025-07-06	5.1833333
12	2025-07-07	5.2500000
13	2025-07-08	5.7666667
14	2025-07-09	5.0000000
15	2025-07-10	6.1333333
16	2025-07-11	0.4500000
17	2025-07-12	6.7166667
18	2025-07-13	1.2166667
19	2025-07-14	0.5166667
20	2025-07-15	5.4833333

Showing 1 to 21 of 34 entries, 2 total columns

	date	active_energy
1	2022-11-19	211.158
2	2022-11-20	46.638
3	2022-11-21	42.550
4	2022-11-22	82.751
5	2022-11-23	37.519
6	2022-11-24	28.247
7	2022-11-25	62.043
8	2022-11-26	34.025
9	2022-11-27	51.324
10	2022-11-28	20.911
11	2022-11-29	92.791
12	2022-11-30	40.880
13	2022-12-01	36.096
14	2022-12-02	70.969
15	2022-12-03	109.291
16	2022-12-04	31.364
17	2022-12-05	67.400
18	2022-12-06	62.239
19	2022-12-07	122.579
20	2022-12-08	17.862

Showing 1 to 21 of 1,097 entries, 2 total columns

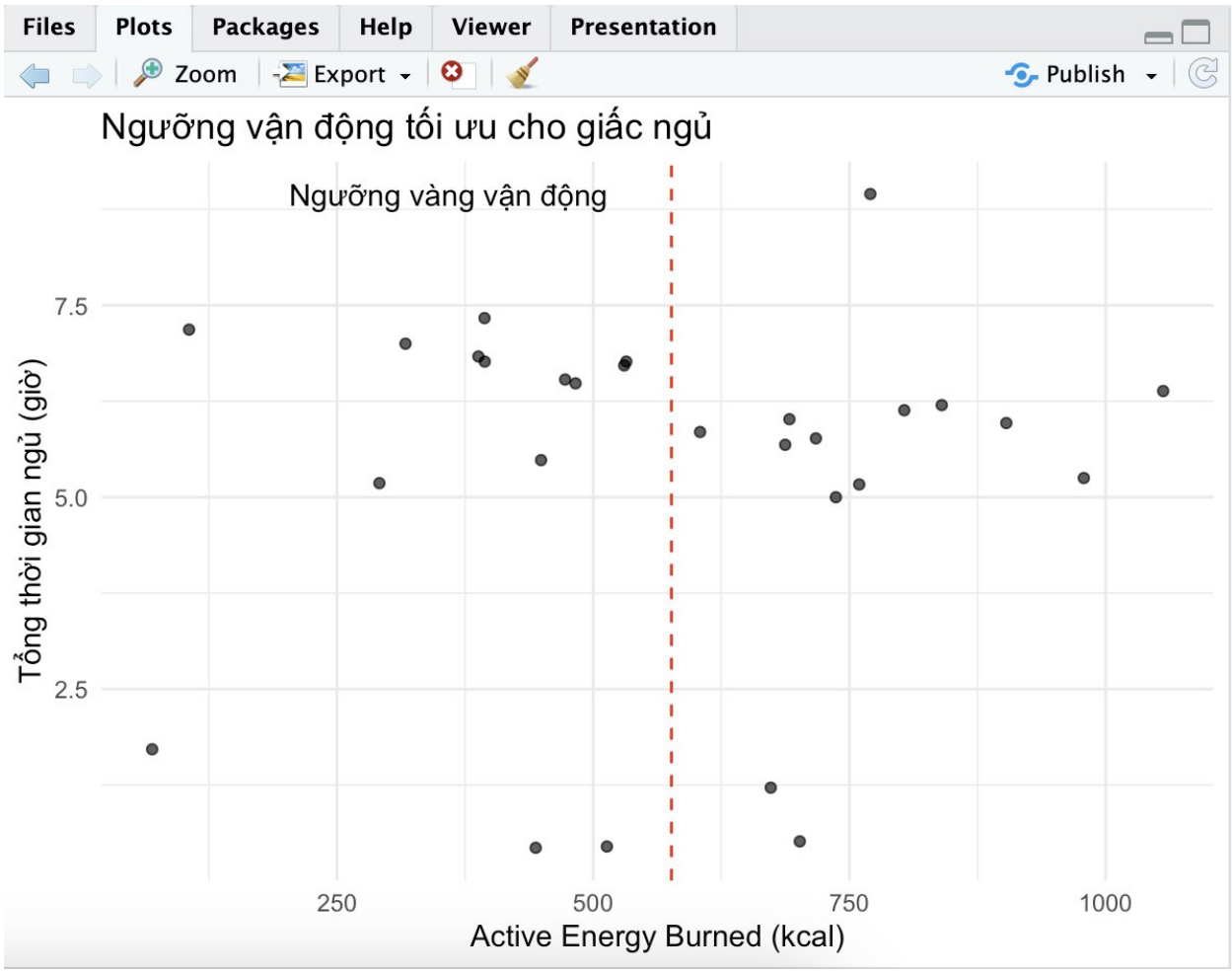
3.1.3. Phân tích mối liên hệ giữa vận động và giấc ngủ

- **Phân tích ngưỡng vận động tối ưu cho giấc ngủ**

Code R :

```r
# Chia thành 3 ngưỡng vận động theo energy (low,moderate,high)
energy_sleep <- energy_sleep %>%
  mutate(
    activity_level = case_when(
      active_energy < quantile(active_energy, 0.33) ~ "Low activity",
      active_energy < quantile(active_energy, 0.66) ~ "Moderate activity",
      TRUE ~ "High activity"
    )
  )
ggplot(energy_sleep, aes(active_energy, total_sleep)) +
  geom_point(alpha = 0.7) +
  geom_vline(xintercept = gold_energy, linetype = "dashed", color = "red") +
  annotate(
    "text",
    x = gold_energy,
    y = max(energy_sleep$total_sleep),
    label = "Ngưỡng vàng vận động",
    hjust = 1.2
  ) +
  labs(
    title = "Ngưỡng vận động tối ưu cho giấc ngủ",
    x = "Active Energy Burned (kcal)",
    y = "Tổng thời gian ngủ (giờ)"
  ) +
  theme_minimal()
```

Biểu đồ :



Phân tích :

Biểu đồ scatter thể hiện mối quan hệ giữa Active Energy Burned (kcal) và tổng thời gian ngủ (giờ**)**. Đường gạch đỏ biểu thị ngưỡng vận động trung tâm (~550 kcal), được xem là "ngưỡng vàng".

Kết quả cho thấy:

- Ở mức vận động trung bình (~400–600 kcal)**,** thời gian ngủ tập trung nhiều trong khoảng 6–7.5 giờ**,** tương đối ổn định và cao hơn so với các mức khác.
- Khi vận động quá thấp (<300 kcal), thời gian ngủ biến động mạnh, xuất hiện cả những ngày ngủ rất ít (<2 giờ), cho thấy vận động thấp không đảm bảo chất lượng phục hồi.
- Ở mức vận động rất cao (>700 kcal)**,** thời gian ngủ không tiếp tục tăng, thậm chí có xu hướng dao động và giảm nhẹ ở một số ngày, gợi ý khả năng quá tải hoặc cơ thể chưa kịp phục hồi.

Mối quan hệ giữa vận động và giấc ngủ có dạng phi tuyến (inverted-U shape)**.**

Mức vận động trung bình (~500–600 kcal/ngày) là vùng tối ưu giúp duy trì thời gian ngủ ổn định và cao hơn. Vận động quá ít hoặc quá nhiều đều không mang lại lợi ích tối đa cho giấc ngủ.

Phát hiện này cho thấy tối ưu sức khỏe không nằm ở việc "tập càng nhiều càng tốt", mà ở việc duy trì cường độ vận động phù hợp với khả năng phục hồi của cơ thể.

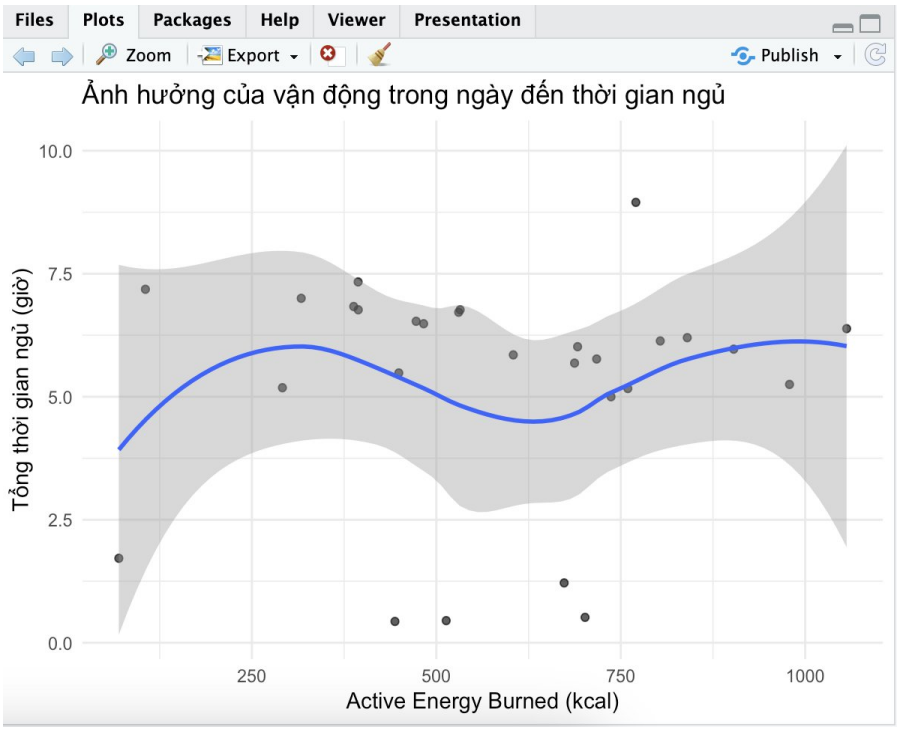
- **Phân tích mối quan hệ giữa mức vận động và chất lượng giấc ngủ**

Code R :

```r
# Dùng Inner_join để kết hợp 2 bảng energy và sleep lại với nhau
energy_sleep <- energy_daily %>%
  inner_join(sleep_daily, by = "date")

# Tạo biểu đồ
ggplot(energy_sleep, aes(x = active_energy, y = total_sleep)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "loess", se = TRUE) +
  labs(
    title = "Ảnh hưởng của vận động trong ngày đến thời gian ngủ",
    x = "Active Energy Burned (kcal)",
    y = "Tổng thời gian ngủ (giờ)"
  ) +
  theme_minimal()
```

Biểu đồ :



Phân tích :

Biểu đồ thể hiện mối quan hệ giữa Active Energy Burned (kcal) và Tổng thời gian ngủ (giờ), với đường xu hướng làm mượt (LOESS) và vùng tin cậy 95%.

- Mối quan hệ phi tuyến (non-linear relationship)

  Đường xu hướng cho thấy quan hệ giữa vận động và giấc ngủ không tuyến tính.

Thay vì "vận động càng nhiều ngủ càng lâu", dữ liệu cho thấy một dạng dao động hình chữ U nhẹ.

- Vùng vận động trung bình (~250–400 kcal)
  Thời gian ngủ tăng dần và đạt khoảng 6 giờ, cho thấy vận động vừa phải có tác động tích cực đến giấc ngủ.
- Vùng vận động trung cao (~500–650 kcal)
  Thời gian ngủ có xu hướng giảm nhẹ (~4.5–5 giờ). Điều này có thể phản ánh:
  - Cơ thể mệt nhưng chưa phục hồi hoàn toàn
  - Hoặc cường độ vận động chưa tối ưu cho giấc ngủ
- Vận động cao (>750 kcal)
  Thời gian ngủ tăng trở lại (~6 giờ), tuy nhiên vùng tin cậy mở rộng cho thấy độ biến động lớn → hiệu quả không ổn định giữa các ngày.

Kết luận : Dữ liệu cho thấy vận động có ảnh hưởng đến thời gian ngủ, nhưng mối quan hệ không tuyến tính và phụ thuộc vào cường độ. Mức vận động vừa phải (khoảng 300–400 kcal/ngày) có xu hướng mang lại thời gian ngủ ổn định và tối ưu hơn. Tăng vận động quá mức không đảm bảo cải thiện giấc ngủ và có thể gây dao động lớn trong thời gian nghỉ ngơi.

3.2 Lưu trữ và truy vấn dữ liệu qua MySQL

3.2.1 Nhập dữ liệu

- Code SQL :

```sql
LOAD DATA INFILE '/path/to/health.csv'
INTO TABLE health
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

3.2.2 Làm sạch dữ liệu

- Code SQL :

```sql
/* 3.1 Chuẩn hóa định dạng ngày giờ
   Dữ liệu gốc dạng: '2/7/25 13:02'
   Format đúng: %m/%d/%y %H:%i
*/
UPDATE health_raw
SET activity_date = STR_TO_DATE(Attribute_startDate, '%m/%d/%y %H:%i')
WHERE Attribute_startDate IS NOT NULL;

/* 3.2 Xử lý dữ liệu NULL hoặc rỗng */

DELETE FROM health_raw
WHERE Attribute_startDate IS NULL
   OR Attribute_startDate = ''
   OR Attribute_value IS NULL
   OR Attribute_value = '';


```

```sql
/* 3.3 Chuẩn hóa kiểu dữ liệu số */

UPDATE health_raw
SET cleaned_value = CAST(Attribute_value AS DECIMAL(10,2))
WHERE Attribute_value REGEXP '^[0-9]+(\\.[0-9]+)?$';


/* 3.4 Loại bỏ dữ liệu trùng */

DELETE t1 FROM health_raw t1
INNER JOIN health_raw t2
WHERE t1.id > t2.id
AND t1.activity_date = t2.activity_date
AND t1.cleaned_value = t2.cleaned_value
AND t1.Attribute_type = t2.Attribute_type;


/* 3.5 Tạo bảng dữ liệu sạch cuối cùng */

DROP TABLE IF EXISTS health_cleaned;

CREATE TABLE health_cleaned AS
SELECT
    id,
    Attribute_type,
    activity_date,
    cleaned_value
FROM health_raw
WHERE activity_date IS NOT NULL
  AND cleaned_value IS NOT NULL;
```

3.2.3 Truy vấn dữ liệu bằng SQL

- **Lọc dữ liệu vận động xu hướng vào ngày cuối tuần :**

Code SQL và kết quả :

```sql
-- Import file.csv vào Database (MySQL)
-- Cleaning data: kiểm tra data types, data format, xoá cột và giá trị Null không cần thiết
-- Kiểm tra bảng:
select distinct Attribute_type from health;
describe health;
-- Lọc dữ liệu xu hướng vận động vào ngày cuối tuần:
SELECT
   DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS step_date,
   SUM(Attribute_value) AS total_steps
FROM health
WHERE Attribute_type = 'HKQuantityTypeIdentifierStepCount'
 AND DAYOFWEEK(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) IN (1,
7)
GROUP BY step_date
```

```
ORDER BY step_date;
```

step_date	total_steps
2025-07-05	5269
2025-07-06	6281
2025-07-12	10767
2025-07-13	9059
2025-07-19	12595
2025-07-20	18023
2025-07-26	14691
2025-07-27	19109
2025-08-02	2881
2025-08-03	1213
2025-08-10	1082
2025-09-27	8883
2025-10-04	2339
2025-10-05	10752
2025-10-11	3180
2025-10-12	12021
2025-10-18	2324
2025-10-19	14840
2025-10-25	2531
2025-10-26	9356

Phân tích :

Dữ liệu cho thấy mức độ vận động vào cuối tuần có sự dao động rõ rệt. Một số cuối tuần ghi nhận tổng số bước cao hơn trung bình ngày thường, cho thấy xu hướng hoạt động ngoài trời hoặc tham gia các hoạt động thể chất nhiều hơn vào thời gian rảnh.

Tuy nhiên, cũng có cuối tuần mức vận động thấp → cho thấy hành vi vận động chưa thực sự ổn định.

Nhận định: Thói quen vận động phụ thuộc vào lịch sinh hoạt cá nhân, chưa duy trì đều đặn.

- **Ngày vận động nhiều nhất & ít nhất :**

Code SQL và kết quả :

```
-- Lọc ngày vận động nhiều nhất:
SELECT
    step_date,
    total_steps
FROM (
    SELECT
        DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS step_date,
        SUM(Attribute_value) AS total_steps
```

```sql
    FROM health
    WHERE Attribute_type = 'HKQuantityTypeIdentifierStepCount'
    GROUP BY step_date
) t
ORDER BY total_steps DESC
LIMIT 1;
-- Ngày vận động ít nhất:
SELECT
    step_date,
    total_steps
FROM (
    SELECT
        DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS step_date,
        SUM(Attribute_value) AS total_steps
    FROM health
    WHERE Attribute_type = 'HKQuantityTypeIdentifierStepCount'
    GROUP BY step_date
) t
ORDER BY total_steps ASC
LIMIT 1;
```

step_date	total_steps
2025-07-27	19109

step_date	total_steps
2025-09-10	123

- **Những ngày đạt trên 8.000 bước :**
  Code SQL và kết quả :

```sql
-- Lọc hiển thị những ngày vận động nhiều (steps>8000):
SELECT
    step_date,
    total_steps
FROM (
    SELECT
        DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS step_date,
        SUM(Attribute_value) AS total_steps
    FROM health
    WHERE Attribute_type = 'HKQuantityTypeIdentifierStepCount'
    GROUP BY step_date
) t
WHERE total_steps > 8000
ORDER BY  step_date ASC;
```

step_date	total_steps
2025-07-07	18507
2025-07-08	15766
2025-07-09	14342
2025-07-10	14138
2025-07-11	12291
2025-07-12	10767
2025-07-13	9059
2025-07-14	11649
2025-07-19	12595
2025-07-20	18023
2025-07-21	11015
2025-07-26	14691
2025-07-27	19109
2025-07-28	16586
2025-07-29	9816
2025-07-30	11785
2025-07-31	14378
2025-08-01	16623
2025-09-11	9890
2025-09-18	12037

```sql
-- Lọc và hiển thị có bao nhiêu ngày vận động ít (steps<1000):
SELECT COUNT(*) AS number_of_low_active_days
FROM (
   SELECT
      DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS step_date,
      SUM(Attribute_value) AS total_steps
   FROM health
   WHERE Attribute_type = 'HKQuantityTypeIdentifierStepCount'
   GROUP BY step_date
) t
WHERE total_steps < 1000;
```

number_of_low_active_days
7

Phân tích :

Tần suất ngày vận động cao (> 8.000 bước)

Trong giai đoạn theo dõi, ghi nhận 30 ngày đạt trên 8.000 bước**.**

Phần lớn các ngày này nằm trong tháng 7 và đầu tháng 8, với nhiều ngày đạt trên 14.000–19.000 bước.

Đặc biệt:

- Ngày cao nhất: **27/7/25 – 19.109 bước**

- Một số ngày duy trì mức rất cao liên tiếp (7/7 → 14/7)

Kết luận hành vi sức khoẻ

- Có nhiều ngày đạt mức khuyến nghị (≥ 8.000 bước) → lối sống tương đối năng động.
- Tuy nhiên vẫn tồn tại các ngày ít vận động → ảnh hưởng đến tính bền vững.
- Giai đoạn tháng 7 là thời điểm vận động cao nhất.
- Cần cải thiện tính ổn định thay vì tập trung vào các ngày đạt đỉnh.

˜ **Tìm nhịp tim trung bình qua step và heart rate :**

Code SQL và kết quả :

```sql
-- "Trong những ngày đi bộ nhiều (step > 8000), nhịp tim nghỉ trung bình (Resting Heart Rate) là bao nhiêu?"
SELECT
    s.step_date,
    s.total_steps,
    r.avg_resting_hr
FROM (
    SELECT
        DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS step_date,
        SUM(Attribute_value) AS total_steps
    FROM health
    WHERE Attribute_type = 'HKQuantityTypeIdentifierStepCount'
    GROUP BY step_date
) s
JOIN (
    SELECT
        DATE(STR_TO_DATE(Attribute_startDate, '%d/%m/%y %H:%i')) AS hr_date,
        AVG(Attribute_value) AS avg_resting_hr
    FROM health
    WHERE Attribute_type = 'HKQuantityTypeIdentifierRestingHeartRate'
    GROUP BY hr_date
) r
ON s.step_date = r.hr_date
WHERE s.total_steps > 8000
ORDER BY s.total_steps DESC;
```

step_date	total_steps	avg_resting_hr
2025-07-07	18507	48
2025-07-08	15766	49
2025-07-09	14342	51
2025-07-10	14138	49
2025-07-03	9943	46

Phân tích :

Trong những ngày vận động nhiều, nhịp tim nghỉ duy trì ở mức thấp và ổn định.
Điều này cho thấy:

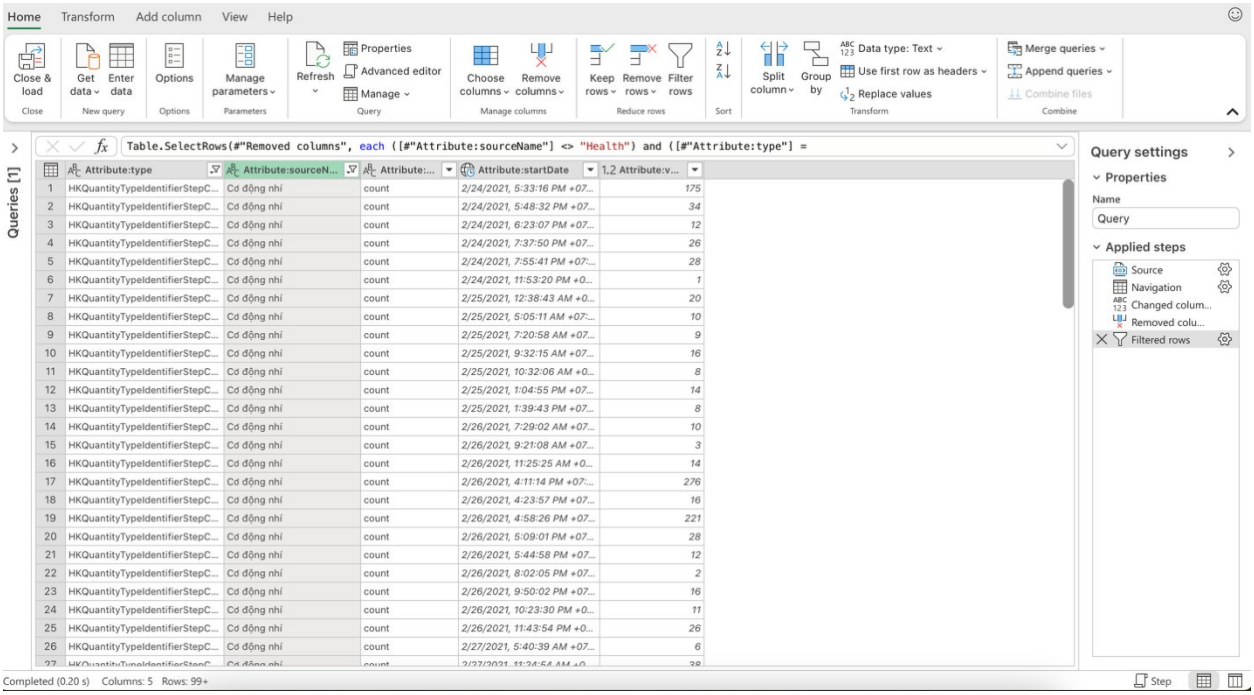
- Cơ thể có khả năng phục hồi tốt
- Hệ tim mạch thích nghi tích cực với mức vận động cao
- Không có dấu hiệu quá tải sinh lý trong giai đoạn này

Kết luận phân tích
Dữ liệu bước đi và nhịp tim nghỉ cho thấy mối quan hệ tích cực giữa vận động và tình trạng tim mạch:

- Vận động cao không làm tăng nhịp tim nghỉ
- Nhịp tim nghỉ thấp phản ánh thể trạng tương đối tốt
- Giai đoạn đầu tháng 7 có thể được xem là thời điểm thể lực tối ưu

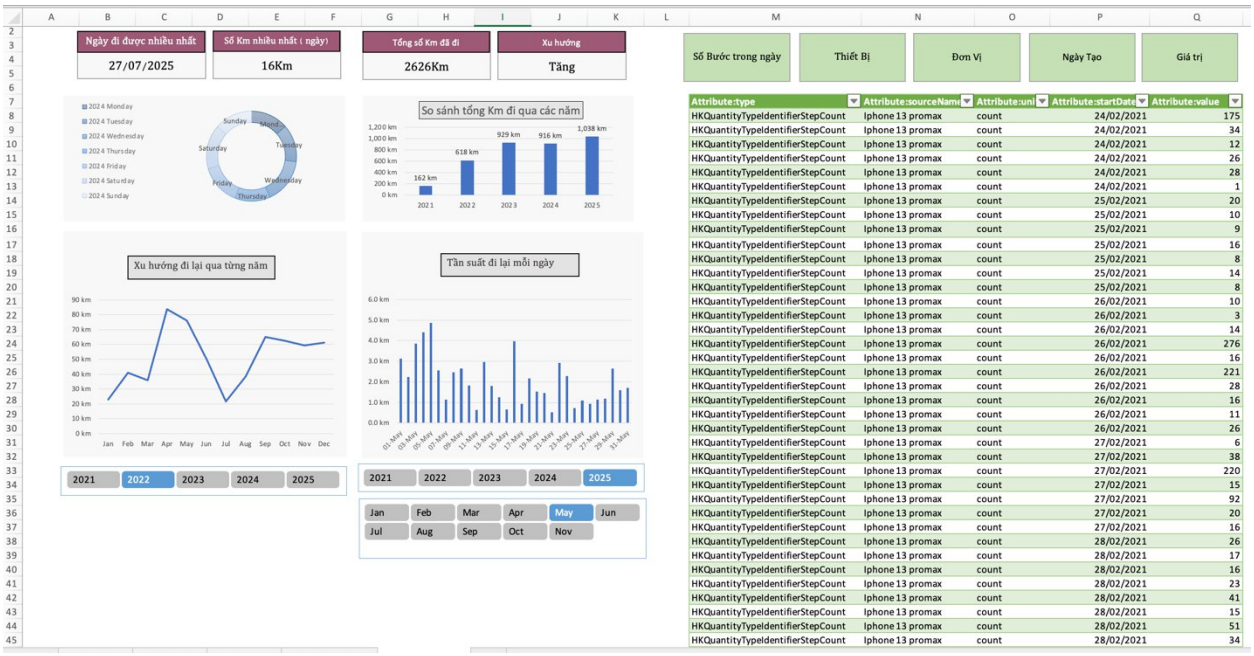
3.3 Tạo Dashboard qua Excel

3.3.1. Import file lên Excel



3.3.2. Tạo Dashboard qua Pivot table trong Excel

- Dashboard :



- Chi tiết Dashboard :

BÁO CÁO PHÂN TÍCH DỮ LIỆU VẬN ĐỘNG GIAI ĐOẠN 2021 - 2025

1. Tổng quan hiệu suất vận động dài hạn (Long-term Performance)

- Chỉ số tăng trưởng: Chúng ta ghi nhận một xu hướng tăng trưởng tích cực và bền vững. Tổng quãng đường đã tăng từ 162 km (2021) lên mức kỷ lục 1,038 km (2025).
- Kết luận báo cáo: Mức độ cam kết vận động năm 2025 tăng 640% so với năm cơ sở 2021. Điều này cho thấy sự thay đổi hoàn toàn về lối sống hoặc tính chất công việc theo hướng năng động hơn qua từng năm.

2. Phân tích biến động theo chu kỳ thời gian (Seasonality & Trends)

- Biểu đồ xu hướng năm (2025): Dữ liệu cho thấy sự mất cân đối rõ rệt giữa các tháng.
    - Điểm bùng nổ: Tháng 7 ghi nhận mức đỉnh cao nhất (vượt mốc 200 km/tháng), tăng đột biến so với trung bình các tháng khác.
    - Điểm trũng: Các tháng 2, 5 và 6 có xu hướng giảm sâu (chỉ duy trì quanh mức 50 - 70 km).
- Kết luận báo cáo: Cần xác định nguyên nhân gây ra sự sụt giảm trong quý 2 để có các biện pháp duy trì phong độ ổn định hơn, tránh tình trạng "vận động dồn nén" chỉ vào một vài thời điểm trong năm.

3. Phân tích thói quen vận động hằng tuần (Weekly Distribution)

- Phân bổ theo thứ: Biểu đồ Donut cho thấy sự phân bổ cực kỳ đồng đều giữa tất cả các ngày trong tuần (từ Thứ Hai đến Chủ Nhật).
- Kết luận báo cáo: Đối tượng không có thói quen "tập bù" vào cuối tuần. Ngược lại, việc vận động dàn trải đều cho thấy đây là hoạt động phát sinh từ việc di chuyển hằng ngày (đi làm, đi học) hơn là các hoạt động thể thao tự nguyện vào ngày nghỉ.

4. Phân tích cường độ hằng ngày (Daily Intensity - Data Jan 2025)

- Tần suất: Trong tháng 1/2025, cường độ đi lại trung bình dao động từ 2.0 km đến 4.0 km/ngày.
- Các điểm đột biến: Có ít nhất 4 ngày trong tháng vượt mức 5.0 km, cá biệt có ngày lên tới 6.0 km.
- Kết luận báo cáo: Các mốc đột biến này thường rơi vào giữa và cuối tháng, có thể tương ứng với các lịch trình di chuyển ngoại khóa hoặc công tác đặc biệt.

5. Đánh giá chỉ số kỷ lục & Đề xuất (Key Metrics & Recommendations)

- Kỷ lục cá nhân: Ngày 27/07/2025 ghi nhận mức vận động tối đa lên đến 16 km. Đây là ngưỡng vận động cường độ rất cao (gần bằng một nửa cự ly bán marathon).
- Tổng quãng đường tích lũy: 2,626 km trong vòng 5 năm là con số rất ấn tượng đối với dữ liệu từ thiết bị di động.

Đề xuất của người làm báo cáo:

1. Duy trì đà tăng trưởng: Với xu hướng hiện tại, mục tiêu cho năm 2026 có thể đặt ở mức 1,200 km.
2. Cân bằng quý: Cần tập trung cải thiện chỉ số của các tháng "thấp điểm" (tháng 2, tháng 6) để đảm bảo sức bền lâu dài.
3. Theo dõi nồng độ Lactate: Dựa trên các ngày đạt 16 km, tôi khuyến nghị kiểm tra thêm chỉ số Aerobic Threshold (Ngưỡng hiếu khí) để tối ưu hóa hiệu quả vận động mà không gây quá tải cho cơ thể.