**Group Name**: Solo

**Name**: Armel Moumbe

**Email** : armel.moumbe@aivancity.education

**Country** : France

**College**: Aivancity school for Technology Business & Society

**Specialization**: Data Analyst

**NB: This project is done by me alone due to not having any group members. Thank you for your time and understanding.**

## Problem description

XYZ company is collecting the data customer using google forms/survey monkey and they have floated n number of forms on the web.

The company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard. The company wants clean data and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data).

## File Overview: Fitness consumer survey

This file contains answers to research made to study the impact of fitness wearables on consumer behavior. The data was collected using a survey by the researcher. The dataset consists of 30 responses from 30 respondents and 21 questions that were asked along with the timestamp.

The problem with this dataset is its small number of responses. There are no missing values in the dataset, and it is well organized.

**File Content and Structure:**

**Columns and Data Types**: The data type of each column is string. These are the columns and what they mean.

These are the names of the columns, some of their function and choice of answers from the survey.

- Timestamp

It tells the time

- What is your age?
  It tells the age

- What is your gender?
  It tells the gender
- What is your highest level of education?
  Education
- What is your current occupation?
  Profession
- How often do you exercise in a week?
  Frequency of exercise in a week

- How long have you been using a fitness wearable?
- How frequently do you use your fitness wearable?
- How often do you track fitness data using wearable?
- How has the fitness wearable impacted your fitness routine?

- Has the fitness wearable helped you stay motivated to exercise? Strongly agree, Agree, Neutral.

- Do you think that the fitness wearable has made exercising more enjoyable? Strongly agree, Agree, Neutral.

- How engaged do you feel with your fitness wearable? Somewhat engaged, Very engaged, Neutral.

- Does using a fitness wearable make you feel more connected to the fitness community? Agree, somewhat agree, neutral.

- How has the fitness wearable helped you achieve your fitness goals? No impact on achieving my goals, helped me achieve my goal somewhat more quickly, helped me achieve my goals much more quickly.

- How has the fitness wearable impacted your overall health? No impact on my overall health, improved my overall health somewhat, improved my overall health significantly.

- Has fitness wearable improved your sleep patterns? Agree, somewhat agree, neutral.

- Do you feel that the fitness wearable has improved your overall well-being? Agree, somewhat agree, neutral.

- Has using a fitness wearable influenced your decision? [To exercise more?] Agree, somewhat agree, neutral.
- Has using a fitness wearable influenced your decision? [To purchase other fitness-related products?] Agree, somewhat agree, neutral.
- Has using a fitness wearable influenced your decision? [To join a gym or fitness class?] Agree, somewhat agree, neutral.
- Has using a fitness wearable influenced your decision? [To change your diet?] Agree, somewhat agree, neutral.

**Rows**: Each row in the files represents a unique individual.

**Header**: The first row of these CSV files serves as a header, clearly labeling each column to denote the corresponding data field.

## File Overview: Fitness analysis

This file contains dataset from a survey data for the type of fitness practices that people follow. It has answers from participants which are friends, family and coworkers of the researcher who made the survey.

**File Content and Structure:**

These are the features/columns of the dataset, starting from the first feature to the last with what they do.

Name of the person attending the survey

Gender of the person attending the survey

Age of the person attending the survey

How important is an exercise to you on the scale of 1 to 5

How do you describe your current level of fitness? - Perfect, very good, Good, Average, Unfit

How often do you exercise? - Every day, 1 to 2 times a week, 2 to 3 times a week, 3 to 4 times a week, 5 to 6 times a week, never

What barriers, if any, prevent you from exercising more regularly? (Select all that applies) - I don't have enough time, I can't stay motivated, I'll become too tired, I have an injury, I don't really enjoy exercising, I exercise regularly with no barriers

What forms of exercise do you currently participate in? (Select all that applies) - Walking or jogging, gym, swimming, yoga, Zumba dance, lifting weights, team sport, I don't really exercise

Do you exercise __? - Alone, with a friend, With a group, Within a class environment, I don't really exercise

What time of the day do you prefer to exercise? - Early morning, afternoon, evening

How long do you spend exercising per day? - 30 min, 1 hour, 2 hours, 3 hours and above, I don't really exercise

Would you say you eat a healthy balanced diet? - Yes, No, not always

What prevents you from eating a healthy balanced diet, if any? (Select all that applies) - Lack of time, Cost, Ease of access to fast food, Temptation, and cravings, I have a balanced diet

How healthy do you consider yourself on a scale of 1 to 5?

Have you recommended your friends to follow a fitness routine? - Yes, No

Have you ever purchased fitness equipment? - Yes, No

What motivates you to exercise? (Select all that applies) - I want to be fit, I want to increase muscle mass and strength, I want to lose weight, I want to be flexible, I want to relieve stress, I want to achieve a sporting goal, I'm not really interested in exercising.

> **Columns and Data Types**: Two of these columns are Boolean, fourteen are string and two are integer.
> **Rows**: Each row in the files represents a unique individual.
> **Header**: The first row of these CSV files serves as a header, clearly labeling each column to denote the corresponding data field.

## File Overview: Fitness trackers

This is a fitness tracker product dataset consisting of different products from various brands with their specifications, ratings and reviews for the Indian market. The data has been collected from an e-commerce website (Flipkart) using web scraping technique.

**File Content and Structure:**

This dataset contains 565 samples with 11 attributes. There are some missing values in this dataset. Here are the columns in this dataset

**Columns and Data Types**: The data types of this file are string, decimal, integer. With twelve string columns, eight decimal and two integers. These are their names and meaning.

- Brand Name
  Name of the Brand
- Device Type
  Type of Fitness Tracker
- Model Name
  Name of the device model
- Color
  Color of the device
- Selling Price
  Selling Price or Discounted Price of the product
- Original Price
  Original Price of the product
- Display
  Type of screen
- Rating (Out of 5)
  user rating
- Strap Material
  Material of the strap
- Average Battery Life (in days)
  Average battery life in days
- Reviews
  count of product reviews received

**Rows**: Each row in the files represents a unique individual.

**Header**: The first row of these CSV files serves as a header, clearly labeling each column to denote the corresponding data field.

For the missing values, I'll be using python to clean the data and remove missing values. Create a Yaml file to make the validation process more effective. And try to make the process more dynamic based on the work done during week 6.

**GitHub Repo link**: https://github.com/m-armel/Data-glacier-Internship.git