# **Parkinson's Disease Smartwatch dataset (PADS)**

# **A Detailed Report**

As a first step, the PADS dataset was thoroughly analyzed and structured the findings into this visual representation, highlighting data distribution, and insights to ensure a clear and comprehensive understanding of the dataset. (please zoom into 200% for clear visualization of the image below)

### 

### **STEPS FOLLOWED FOR THE DATA EXTRACTION:**

1. **Loading Data from Folder**
   * Used **"Get Data -> Folder"** to load all four tables: *Patient Details, Movement Details, Timeseries Details, and Questionnaire Details*.
   * Clicked on **"Transform Data"**, which opened the **Power Query Editor** for further processing.
2. **Combining JSON Files**
   * The dataset contained JSON files within a **Binary column**.
   * Used **"Combine Files"** to extract and convert JSON data into structured table format.
3. **Parsing JSON Data**
   * Clicked on **"Parse -> JSON"** to convert hierarchical JSON data into a tabular format.
   * Repeated this step for all four tables to ensure consistency.

Successfully extracted and prepared the **PADS dataset** for further analysis in Power BI.

**DATA DEFINITION AND THE CLEANING STEPS OF THE PADS DATASET:**

The sections below outline the data definitions of all columns across tables,the analysis, and changes made during the cleaning process.

**DATA DEFINITION AND DATA CLEANING STEPS OF PATIENT TABLE:**

**Table Name:**

* Renamed from "**PATIENT**" to "**Patient\_Details**" for clarity.

**Column Names:**

**Source.name:** Represents the file name from which the data is getting loaded into Power BI.

**Data Cleaning:**Deleted this column, it's the source filename and does not help in analysis.

**Resource Type :** It defines the data type for the given dataset.

**Data Cleaning:**Deleted this column, it's the source filename and does not help in analysis.

**Id :** It is a unique identifier which is provided to the patient to maintain a medical record.

**Data Cleaning:**Renamed column name id to patient\_id

**study\_id:**This column refers to the study name, as in our dataset it is PADS.

**Data Cleaning:**Renamed it to "study\_name"

**Condition**:It refers to the condition of the patient that impairs the normal structure or function of an organism . For example : Healthy, Parkinson's etc.

**Data Cleaning:**Renamed it to "study\_condition"

**disease\_comment:**It describes a condition of an illness.

**Data Cleaning:**Renamed it to "disease\_description"

**age\_@\_diagnosis:**It is the patient's age at the time of diagnosis.

**Data Cleaning:**Renamed it to "diagnosis\_age"

**Age:**Age is a period of human life which is calculated in years from birth. Here this column refers to the current age of the patient.

**Data Cleaning:**No change required

**Height:**Height refers to the distance from bottom of the feet to top of the body which is measured in cm, as in the dataset it starts from 55cm-203cm.

**Data Cleaning:** height(cm)

**weight:**Weight of the patient which is measured in kilograms(kg).

**Data Cleaning:**weight(kg)

**gender:**Gender refers to the characteristics of a person like Male(1) and Female(0).

**Data Cleaning:**Replaced Male to 1 and Female to 0

**Handedness:**Smartwatch used by a patient either in the right hand or left hand.

**Data Cleaning**:No change required

**appearance\_in\_kinship:** It refers to the patient's family history for PADS, TRUE means patient is having PADS and FALSE means patient is not having PADS which has been updated to True=1 and False=0.

**Data Cleaning:** 1. Renamed to "family\_history"

2. changed the values TRUE to 1 and FALSE to 0

**appearance\_in\_first\_grade\_kinship:**It refers to the medical history of the person's immediate family members including parents, siblings and children. True means PADS and False means no PADS which is updated to True=1 and False=0.

**Data Cleaning:**1.Renamed to "first\_degree\_family\_history"

2.changed the values TRUE to 1 and FALSE to 0

**effect\_of\_alcohol\_on\_tremor**:Alcohol can temporarily improve tremors in patients, but it can also worsen the tremor or can have no effect .

**Data Cleaning:**Renamed it to "alcohol\_effect\_on\_tremor

**(New Column) Patient\_Age\_Group**

* **Purpose:** Categorizes patient records into different age groups.
* **Transformation:**Added to segment large datasets for faster queries and improved processing efficiency.

**DATA DEFINITION AND DATA CLEANING STEPS OF QUESTIONNAIRE TABLE:**

**Table Name:**

* Renamed from "**QUESTIONNAIRE**" to "**Patient\_Questionnaire\_Details**" for clarity.

**Source Name:**Represents the name of the file from which data is being imported.

**Data Cleaning:**Deleted this column, it's the source filename and does not help in analysis.

**Resource Type:**Defines the type of data for this dataset. It stores the questionnaire answers.

**Data Cleaning:**Deleted this column, it's the source filename and does not help in analysis.

**id:**A unique id assigned to each patient. who completed the questionnaire.

**Data Cleaning:**Renamed column name id to patient\_id

**study\_id:**The study name where the questionnaire was used. Example:PADS

**Data Cleaning:**Renamed it to "study\_name"

**id:**The category of the questionnaire. Example:NMS(Non-Motor Symptoms)

**Data Cleaning:**Renamed to"questionnaire\_name"

**questionnaire\_name:**The name of the questionnaire.Example:NMS(Non-Motor Symptoms)

**Data Cleaning:**Deleted this column from the table

**item\_link\_id:**A unique number assigned to each question in the questionnaire.

**Data Cleaning:**Renamed it to "question\_number"

**item.test:**The actual question or symptom description being asked in the questionnaire.

**Data Cleaning:**Renamed it to "question"

**item.answer:**The patient response to the question True= symptom is present, false= symptom is not present.

**Data Cleaning:**Renamed it to "answer"

**(New Column) Unique\_ID**

* **Purpose:** Provides a unique identifier for each row.
* **Transformation:**Added using **Add Column → Index Column → Starting from "1"** since merging other column values for uniqueness resulted in duplicates.

**DATA DEFINITION AND DATA CLEANING STEPS OF OBSERVATION TABLE:**

1. **Source Name**
   * **Definition:** Represents the filename from which data is imported.
   * **Data Cleaning:** **Deleted**, as filenames are not informative for analysis.
2. **Resource Type**
   * **Definition:** Defines the type of data in this dataset, storing questionnaire answers.
   * **Data Cleaning:** **Deleted**, as it does not provide meaningful insights.
3. **id**
   * **Definition:** Unique ID assigned to each patient.
   * **Data Cleaning:** Renamed to **"patient\_id"** for clarity and consistency.
4. **study\_id**
   * **Definition:** Represents the research initiative using wearable sensors to monitor Parkinson’s disease.
   * **Data Cleaning:** Renamed to **"study\_name"** to better describe its purpose.
5. **device\_id**
   * **Definition:** Unique identifier for a wearable device (e.g., Apple Watch) used to track symptoms.
   * **Data Cleaning:** Renamed to **"study\_device\_name"** for improved clarity.
6. **id (Duplicate)**
   * **Definition:** Unique identifier assigned to a specific patient assessment.
   * **Data Cleaning:** Renamed to **"assessment\_type"** to clarify its meaning.
7. **endianness**
   * **Definition:** Defines the byte order of the dataset (e.g., little-endian, big-endian).
   * **Data Cleaning:** Renamed to **"byte\_order"** for standardization.
8. **sampling\_rate**
   * **Definition:** Number of times data is recorded per second (measured in Hz).
   * **Data Cleaning:** Renamed to **"sampling\_rate(Hz)"** to explicitly state units.
9. **data\_type**
   * **Definition:** Specifies the data format (e.g., float, string, whole number).
   * **Data Cleaning:** **No change required**, as the column is useful in analysis.
10. **Bits**

* **Definition:** Specifies the bit size of the dataset (e.g., 32-bit, 16-bit).
* **Data Cleaning:** **No change required**, as it is useful for technical analysis.

1. **Session.record\_name**

* **Definition:** Unique identifier for a recorded session, distinguishing activities by purpose, date, or event.
* **Data Cleaning:** Renamed to **"task\_name"** for clarity.

1. **Session.rows**

* **Definition:** Represents individual recorded time entries (e.g., lap times, split times, event markers).
* **Data Cleaning:** Renamed to **"total\_task\_rows"** for better interpretation.

1. **Session.record.device\_location**

* **Definition:** Indicates the smartwatch placement (left/right wrist, top/bottom orientation).
* **Data Cleaning:** Renamed to **"device\_placement"** for better readability.

1. **Session.records.channel.0**

* **Definition:** Timestamp of recorded movement data.
* **Data Cleaning:** Renamed to **"timestamp\_channel (secs)"** for clarity.

1. **Session.records.channel.1**

* **Definition:** Measures X-axis acceleration (horizontal movement) in g-force.
* **Data Cleaning:** Renamed to **"accelerometer\_x\_channel (g\_force)"** for clarity.

1. **Session.records.channel.2**

* **Definition:** Measures Y-axis acceleration (vertical movement) in g-force.
* **Data Cleaning:** Renamed to **"accelerometer\_y\_channel (g\_force)"** for consistency.

1. **Session.records.channel.3**

* **Definition:** Measures Z-axis acceleration (up-down movement) in g-force.
* **Data Cleaning:** Renamed to **"accelerometer\_z\_channel (g\_force)"** for clarity.

1. **Session.records.channel.4**

* **Definition:** Measures X-axis angular velocity (wrist rotation) in rad/s.
* **Data Cleaning:** Renamed to **"gyroscope\_x\_channel (rad/s)"** for consistency.

1. **Session.records.channel.5**

* **Definition:** Measures Y-axis angular velocity (side-to-side wrist tilt) in rad/s.
* **Data Cleaning:** Renamed to **"gyroscope\_y\_channel (rad/s)"** for clarity.

1. **Session.records.channel.6**

* **Definition:** Measures Z-axis angular velocity (vertical wrist motion) in rad/s.
* **Data Cleaning:** Renamed to **"gyroscope\_z\_channel (rad/s)"** for consistency.

1. **session.records.units.0, session.records.units.1, session.records.units.2, session.records.units.3, session.records.units.4, session.records.units.5, session.records.units.6**

* **Definition:** Measurement units for each recorded data point.
* **Data Cleaning:** **Deleted**, as metadata already contains unit information, and merging it enhances data cleaning and organization.

1. **(New Column) Patient\_Age\_Group**

* **Definition:** Categorizes patients into different age groups for easier segmentation.
* **Data Cleaning:** Added to **improve query performance**, enabling **faster filtering** and **better analysis of patterns across age groups**.

**DATA DEFINITION AND DATA CLEANING STEPS OF TIMESERIES TABLE:**

**Table Name:**

* Renamed from "TIMESERIES" to "Patient\_Task\_Timeseries\_Details" for clarity.

**Column Names:**

**Source.name:**Represents the file name from which the data is getting loaded into Power BI.

**Data Cleaning:**Renamed to **"timeseries\_id"**.Replace **".txt"** in values with blank to match the **timeseries\_id** in the observations table for relationship establishment.

**Column1:**Represents the exact time when motion signals were captured by the smartwatch. Used to track movement changes over time during PADS tasks.

**Data Cleaning:**Renamed to **"timestamp\_channel\_value"** for clarity.Rounded off values to four decimal places to retain necessary detail for accurate motion analysis.**(**Transform → Rounding → Round → Up to 4 decimal places)

**Column2:**Motion signal of **Accelerometer along the X** **axis** captured at the timestamp value.

**Data Cleaning:**Renamed to **"Accelerometer\_x\_channel\_value"** for clarity.Rounded off values to **f**our decimal places to improve motion analysis accuracy.

**Column3:**Motion signal of **Accelerometer along the Y** **axis** captured at the timestamp value.

**Data Cleaning:**Renamed to **"Accelerometer\_y\_channel\_value"** for clarity.Rounded off values to four decimal places to improve motion analysis accuracy.

**Column4:**Motion signal of **Accelerometer along the Z** **axis** captured at the timestamp value.

**Data Cleaning:**Renamed to **"Accelerometer\_z\_channel\_value"** for clarity.Rounded off values to four decimal places to improve motion analysis accuracy.

**Column5:**Motion signal of **Gyroscope along the X axis** captured at the timestamp value.

**Data Cleaning:**Renamed to **"Gyroscope\_x\_channel\_value"** for clarity.Rounded off values to **four decimal places** to improve motion analysis accuracy.

**Column6:**Motion signal of **Gyroscope along the Y axis** captured at the timestamp value.

**Data Cleaning:**Renamed to **"Gyroscope\_y\_channel\_value"** for clarity.Rounded off values to four decimal places to improve motion analysis accuracy.

**Column7:**Motion signal of **Gyroscope along the Z** **axis** captured at the timestamp value.

**Data Cleaning:**Renamed to **"Gyroscope\_z\_channel\_value"** for clarity.Rounded off values to four decimal places to improve motion analysis accuracy.

**(New Column) Unique\_ID**

* **Purpose:** Provides a unique identifier for each row.
* **Transformation:**Added using **Add Column → Index Column → Starting from "1"** since merging other column values for uniqueness resulted in duplicates.

**(New Column) Patient\_Age\_Group**

* **Purpose:** Categorizes patient records into different age groups.
* **Transformation:**Added to segment large datasets for faster queries and improved processing efficiency.

**STEPS FOLLOWED FOR THE DATA TRANSFORMATION OF PADS DATASET:**

1. Once all data cleaning transformations were completed, **DAX Studio** was used to export the cleaned and structured tables into CSV format without any data loss since the "Timeseries" table was extremely large (~13.5 million records),
2. In DAX Studio, used the "**Advanced Options -> Export Data**" feature to export all tables as CSV files.This option ensures that all records, including the 13.5 million rows from Timeseries, were extracted without truncation.

The transformation process ensured that the PADS dataset was well-structured, optimized for performance, and ready for accurate analysis. By systematically refining the data, it created a strong foundation for meaningful insights and reliable interpretations.