Women in
Data Science
Worldwide

WiDS Datathon 2025
#WiDSDatathon

Kylie Cancilla
Caterina Ponti
Liana Mendoza

Women in
Data Science
Worldwide

WiDS Long-term Vision

We envision a future in which women are fully integrated and represented in all areas of Data Science, and share equally in Decision Making, Economic Prosperity, and Opportunities.

Workshop Goals

- Exploratory Data Analysis
- Encoding categorical variables
- Merging Metadata and Functional Connectivity Matrix
- NaN values

Women in
Data Science
Worldwide

WiDS Mission

To change the field of data science across the globe by elevating, educating, and empowering women to achieve 30% representation of women in data science by 2030.

WiDS Datathon Challenge Task

This datathon challenge aims to answer this question:

What brain activity patterns are associated with ADHD; are they different between males and females, and, if so, how?

To work towards the answer to this question, participants will be tasked with building a **multi outcome model to predict both an individual's sex and their ADHD diagnosis** using functional brain imaging data of adolescents and their socio-demographic, emotions, and parenting information.

A multi-outcome model is designed to predict multiple target variables simultaneously using a single machine learning model.

(female:1, male:0, ADHD: 1 ; no ADHD 0)

Loading Jupyter Notebook

STEP 1: Download Python

- Anaconda installment

STEP 2: Set-up by creating a virtual environment

- Open your Terminal (for Windows, open Anaconda Prompt)

To create the environment: In your terminal, run the command:

```
conda create -n wids-datathon python=3 anaconda
```

To activate it:

- on Mac or Linux:
  ```
  source activate wids-datathon
  ```
- on Windows:
  ```
  activate wids-datathon
  ```

STEP 3: Create your notebook

Loading Jupyter Notebook

STEP 3: Create your notebook

- Open the terminal or Anaconda Prompt
- Navigate to desired directory or folder
- Activate your virtual environment
- Run this command: `jupyter notebook`

This will activate Jupyter Notebook and open the program itself.

More resources are available at the Community Hub under the **technical resources section.**

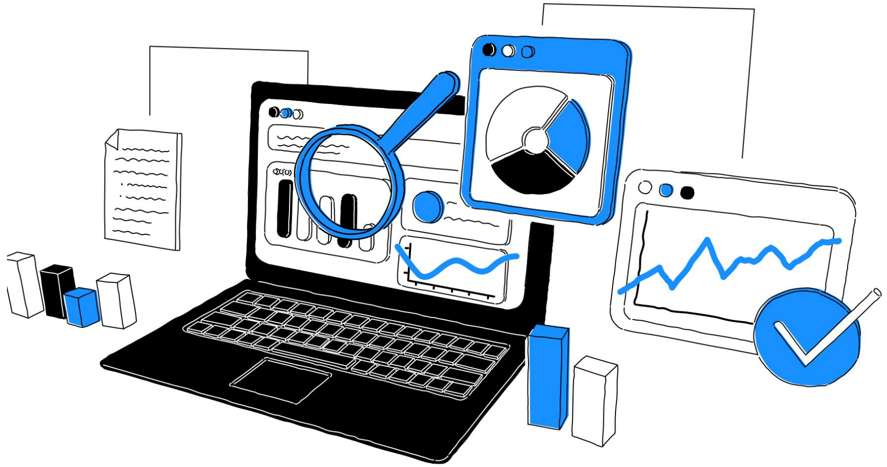
Exploring Your Data: An Introduction to EDA

Definition:

EDA (Exploratory Data Analysis) is the process of examining and summarizing data to uncover patterns, spot problems, and prepare for modeling.

Why It Matters:

- Understand your dataset.
- Detect issues (e.g., missing data, outliers).
- Discover relationships and trends.



SCALABLE PATH

Key Tools:

- Descriptive **statistics** (mean, median, count)
- **Data visualizations** (bar plot, histogram, scatterplot, boxplot)

Key Statistics in Exploratory Data Analysis (EDA):

- **Mean & Median**: Measure central tendency to understand typical values.
- **Standard Deviation & Variance**: Assess data spread and variability.
- **Skewness**: Indicates data asymmetry (left or right tail dominance).
- **Missing Values**: Analyze gaps in the data to ensure accurate modeling.

Correlations

- **Definition:** A measure of how strongly two variables are related.
- **Range:** -1 (perfect negative) to +1 (perfect positive).
- **Why It's Useful:** Helps identify predictive variables and uncover relationships.

Correlation & Functional Connectivity in fMRI Research (ADHD)

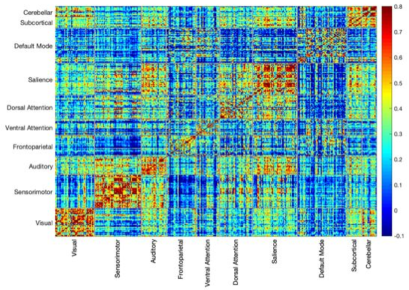
Women in Data Science Worldwide

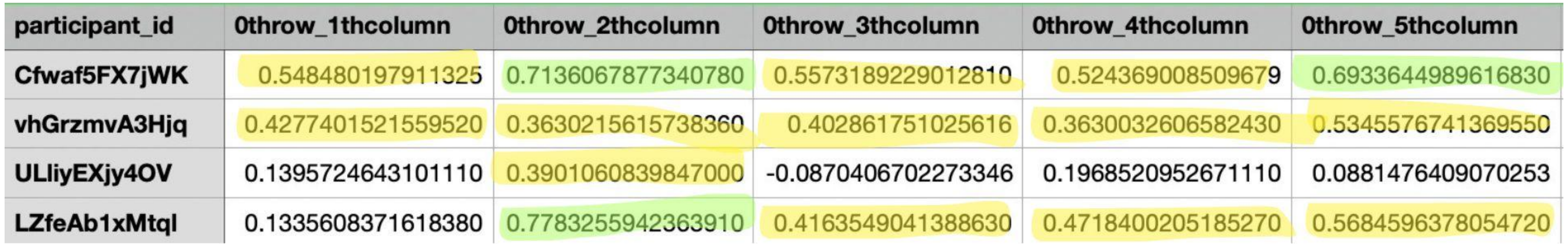
fMRI: Measuring Brain Activity

- **Resting-state fMRI:** Measures brain activity using **Blood Oxygen Level-Dependent (BOLD)** contrast.
- **How It Works:**
  - Active brain regions consume more oxygen. -> Blood flow increases to these regions. -> fMRI detects changes in oxygen levels to infer neural activity.

Functional Connectivity: Understanding Brain Interactions

- **Definition:** Describes interactions between different brain regions.
- **How It's Measured:** Functional Connectivity Matrix: A matrix where rows and columns represent brain regions, and each cell shows the correlation between activity in paired regions.
- **Correlation** helps measure relationships, and **functional connectivity** uses these correlations to understand brain activity and interactions, especially in disorders like ADHD.





participant_id	0throw_1thcolumn	0throw_2thcolumn	0throw_3thcolumn	0throw_4thcolumn	0throw_5thcolumn
Cfwaf5FX7jWK	0.548480197911325	0.7136067877340780	0.5573189229012810	0.524369008509679	0.6933644989616830
vhGrzmvA3Hjq	0.4277401521559520	0.3630215615738360	0.402861751025616	0.3630032606582430	0.5345576741369550
ULliyEXjy4OV	0.1395724643101110	0.3901060839847000	-0.0870406702273346	0.1968520952671110	0.0881476409070253
LZfeAb1xMtql	0.1335608371618380	0.7783255942363910	0.4163549041388630	0.4718400205185270	0.5684596378054720

Women in Data Science Worldwide

What Makes a Variable Predictive?

A variable is predictive if its values are associated with the target variable (e.g., ADHD or gender).

How to Test Predictiveness:

1. **Quantitative Data**: Use histograms or boxplots to visualize distributions of quantitative (e.g., Color vision test score).
2. **Categorical Data**: Use bar plots to compare groups (e.g., ADHD rates by Parent 1 occupation).

Key Questions to Address:

- Which variables are most predictive of ADHD or gender?
- Are there strong correlations worth exploring further?
- What insights can be drawn for modeling?

NaN Values

Women in
Data Science
Worldwide

What are NaN values?

- Not a number values or missing data
- May originate from issues in data collection and curation

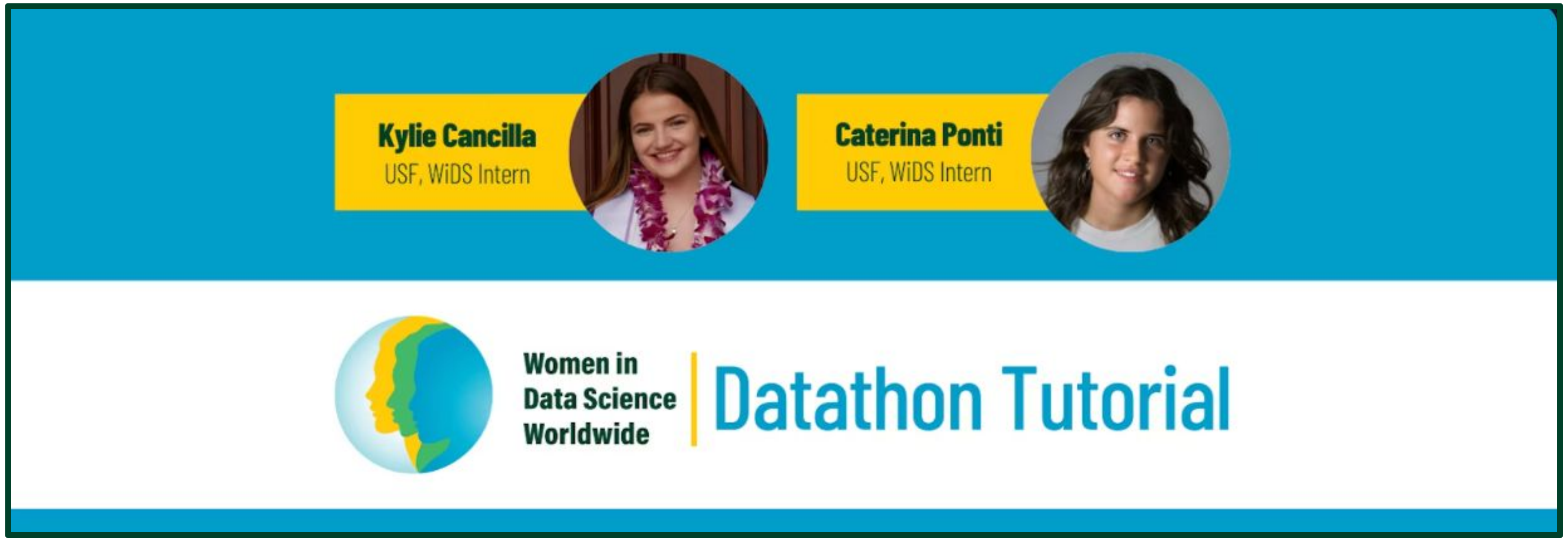
How do you deal with missing data?

- There are multiple methods
- Test method often depends on your dataset and chosen machine learning model

Some methods include the following:

- **Substitution of mean,** mode, or median
- Dropping rows with null values
- Replacing with a constant and arbitrary value

Save the Date: February 5, 2025



Workshop Topic: Building and Evaluating a Machine Learning Model

A. Building the Multi-Output Model

B. Accuracy score

C. Explaining F1-score as the metric for model evaluation

THANK YOU!

Questions?