Datafest Workshop

Learning Objectives

1. Clean your data
2. Visualize Data
3. Understand differences between observations, values and variables
4. Gain knowledge on Pandas

Workshop Overview

Dataset Basics

0-10

Google Colab

10–12

What is Pandas?

12-20

Understanding Dataset

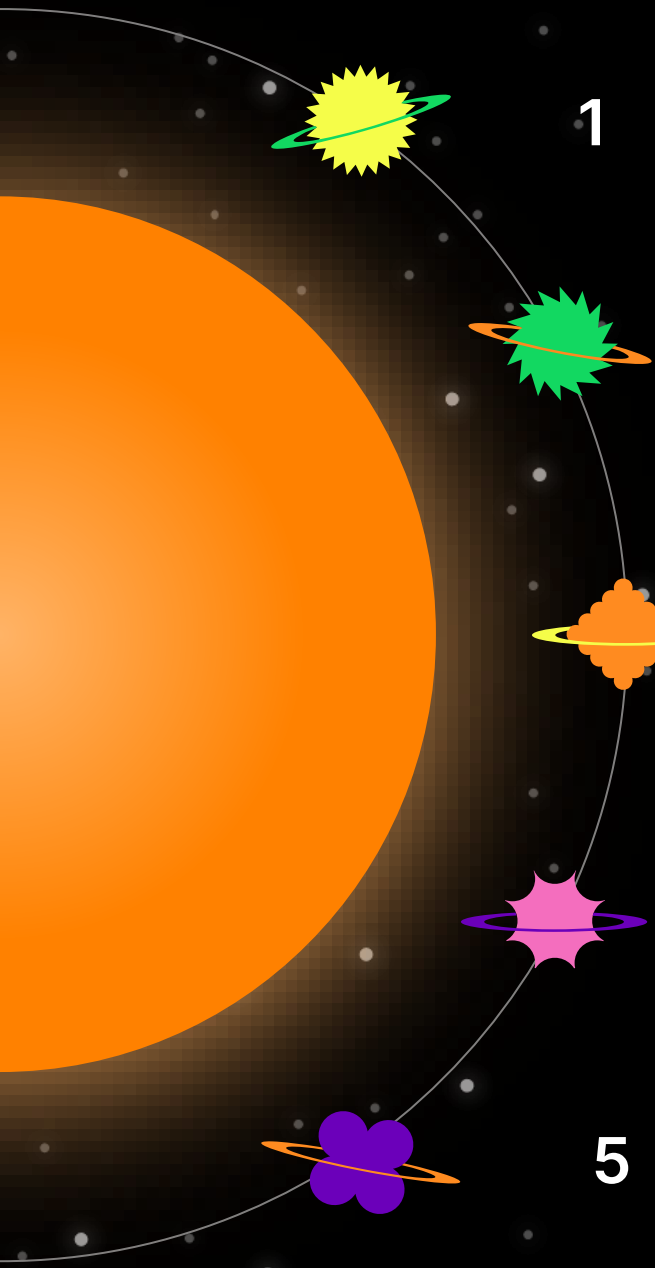
20-26

Transform Data

26-38

Visualize Data

38-50



1 What is a dataset?

2 Rows vs Columns

3 Data Types

4 Structured vs Unstructured Data

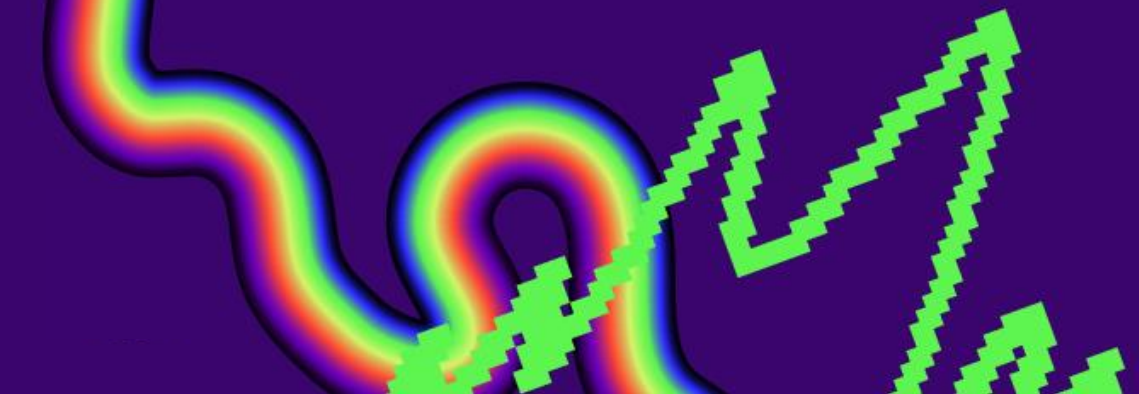
5 Common Data Issues

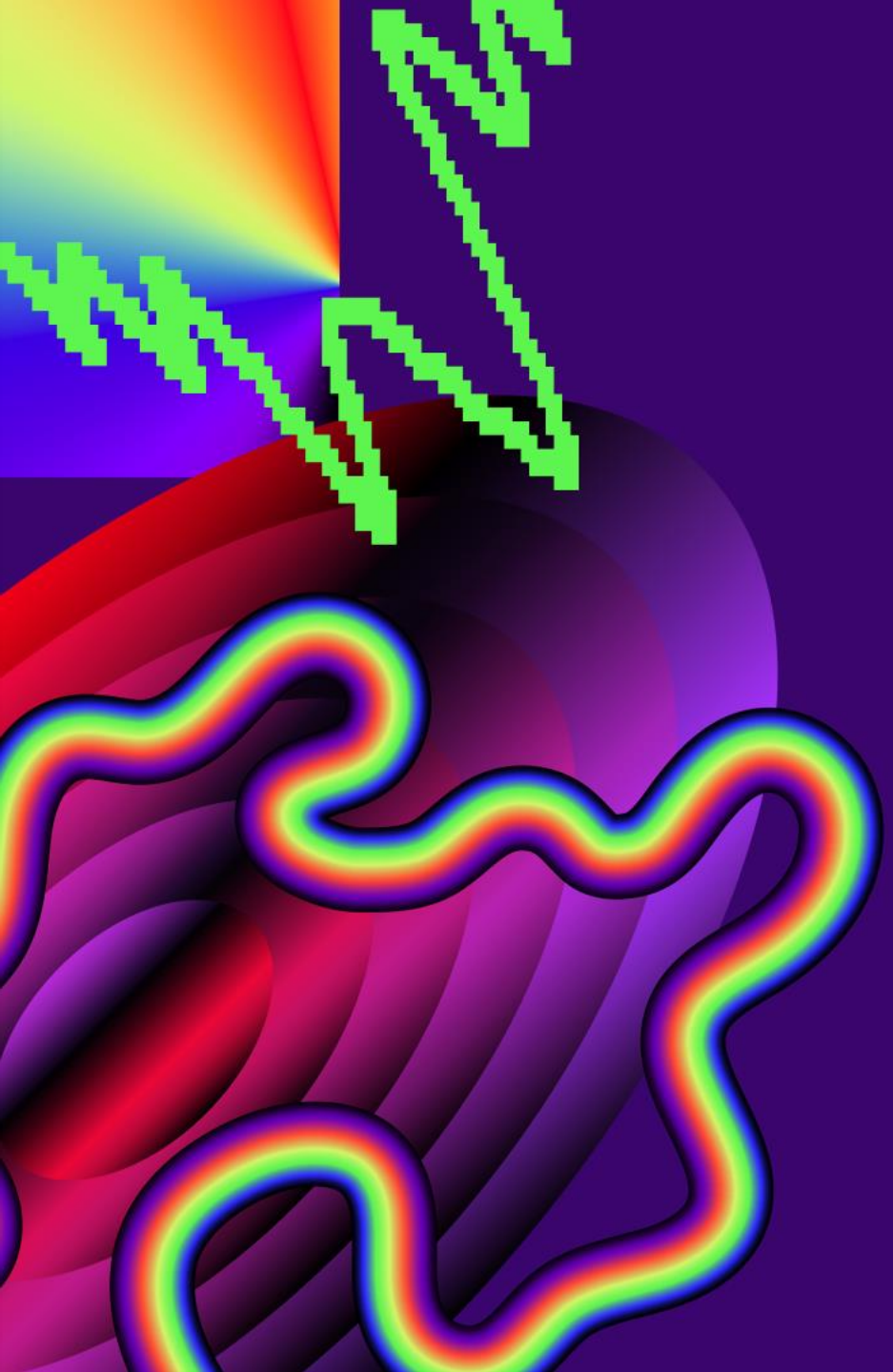
Dataset Basics


What is a
Dataset?

A dataset is a structured collection of data

It is typically organized into rows and columns for analysis





Rows

Records

Columns

Features

Data Types

4 main

Missing Values

NaN

Structured

Tables

Real-World Data

Messy

Rows vs Columns

Example Dataset: Student info

Age	Major	GPA
18	Biology	3.2
19	Psychology	3.7
22	History	3.1
20	Computer Science	3.8

Data Types

Understanding different kinds of data in a Pandas dataset

Main Data Types

1. 

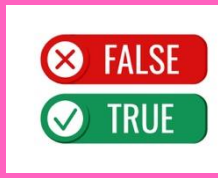
Numbers

Integers, float

2. 

Objects (String)

Text data

3. 

Boolean

True/False values

4. 

Datetime

Dates and Timestamps

5. 

Missing Values

NaN (not a number)

Structured vs Unstructured Data

Structured Data

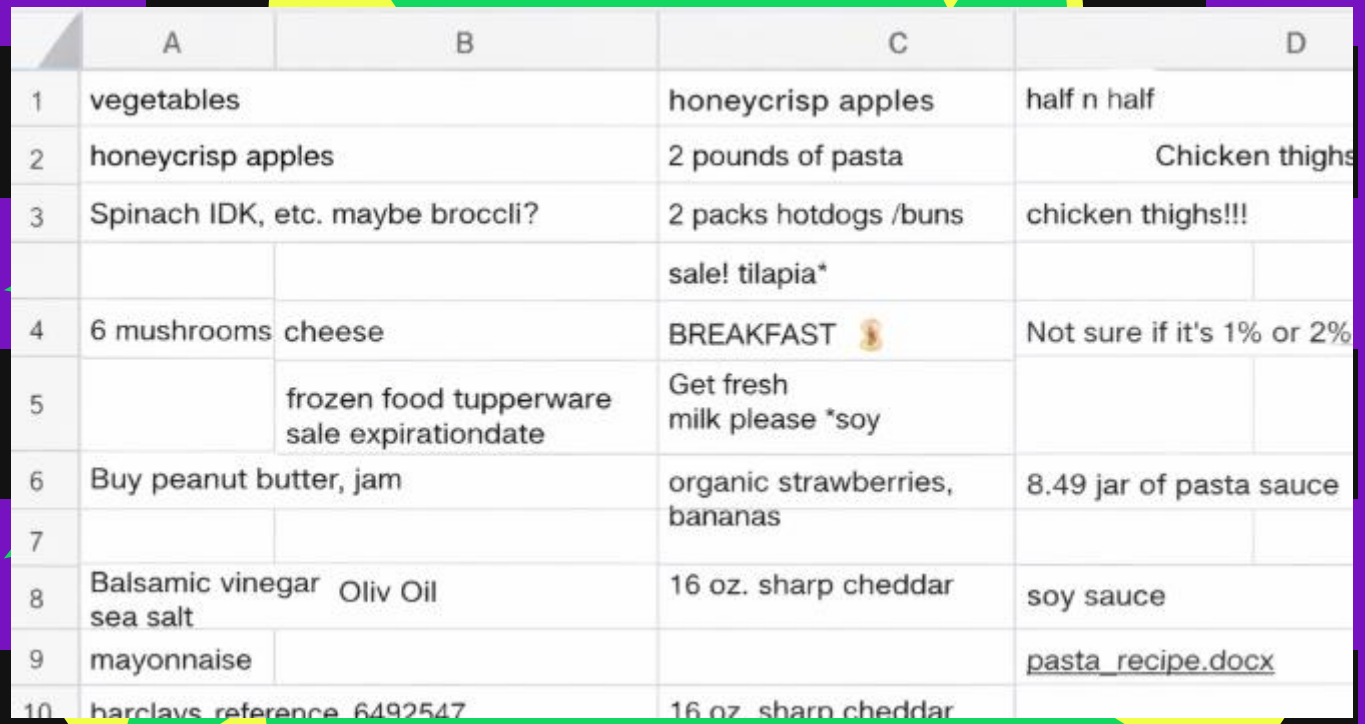
Name	Age	Purchase
Alice	29	Laptop
Bob	35	Shoes
Charlie	22	Book
Dana	40	Watch

Organized in rows and columns

Unstructured Data

Images, audio, video, and free text.

Not organized in rows and columns



	A	B	C	D
1	vegetables		honeycrisp apples	half n half
2	honeycrisp apples		2 pounds of pasta	Chicken thighs
3	Spinach IDK, etc. maybe broccli?		2 packs hotdogs /buns	chicken thighs!!!
			sale! tilapia*	
4	6 mushrooms	cheese	BREAKFAST	Not sure if it's 1% or 2%
5		frozen food tupperware sale expirationdate	Get fresh milk please *soy	
6	Buy peanut butter, jam		organic strawberries, bananas	8.49 jar of pasta sauce
7				
8	Balsamic vinegar sea salt	Oliv Oil	16 oz. sharp cheddar	soy sauce
9	mayonnaise			pasta_recipe.docx
10	barclays reference 6492547		16 oz. sharp cheddar	

This image is unstructured data. It is not organized in rows and columns

Common Data Issues

	A	B	C	D
1	Name	Age	Blood Pressure	Sales
2	Jane	45	130/85	$12,000
3	Matt	58	NaN	$25,500
4	Lisa	34	120/80	$9,800
5	Sam	50	145/90	$18,300

Example Issue

NaN

Real-world datasets often contain missing values.

In Pandas, missing data appears as NaN

Missing values can affect calculations and analysis.

We must detect and handle them properly

Google Colab

https://bit.ly/4sxr8qV