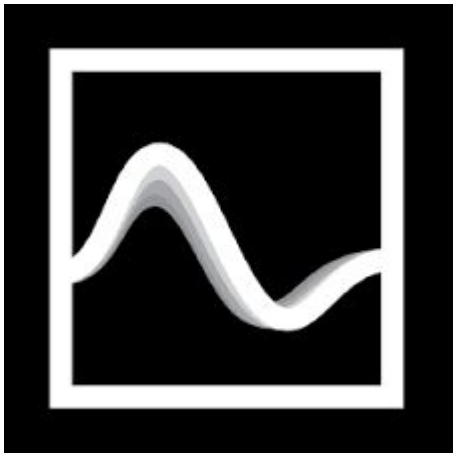
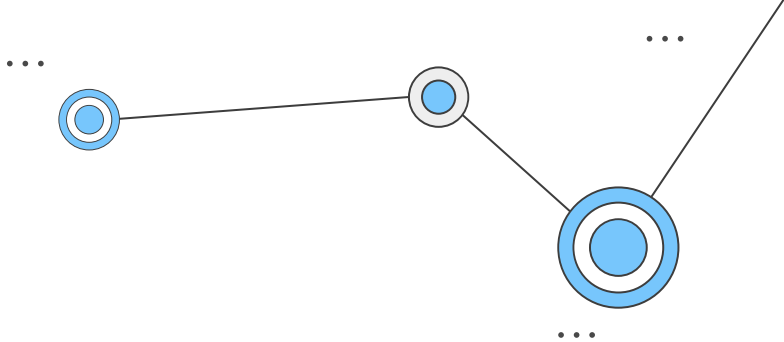




CRA Week 1: Intro to EDA

MAISI

Michigan Data Science Team

Fall 2025

MEET YOUR LEADS! - WILL MCKANNA

Hometown: Rockford, MI
Majors: DS and Statistics
Year: Sophomore
Ask me about: Studying abroad in Iceland, crocheting, trombone, Michigan and Detroit football

MEET YOUR LEADS! - RYAN ZIMMEL

Hometown: Fargo, North Dakota
Major: Information Analysis - UMSI
Minors: Data Science, Business
Year: Junior
Ask me about: A2 Coffee shops, Marching Band, School of Information, Music + Concerts

Session 1 Agenda

01

Fun Icebreaker :)

Get to know your projectmates (and maybe win a prize ?!?!)

02

Expectations

What you stand to gain and what we expect in return

03

Intro to EDA

Learning the first step in the data science process

04

Python Library Overview

What exactly are these libraries we're using?

05

Practice Time!

Work on a dataset utilizing homegrown data :0 (cool stuff)

Icebreaker Bingo!

	A	B	C	D	E
1	I'm a fan of the Detroit Lions	Slept overnight at a UofM non dorm building	I can whistle	I'm part of another CS/DS club	I get the supreme slice @ Joe's
2	I'm a member of MAISI	I have season tickets to Michigan Football	Touched grass this summer (3+ outdoor activities)	I'm a Data Science major	I know the capital of Mongolia
3	I play a sport	I'm a non CS/DS major	I'm a part of MDST	I pay for guac at chipotle	I live on North
4	I'm a Computer Science major	Took Math 215 at Michigan (WCC >>)	I play an instrument	I've taken a formal statistics class (HS/college)	I've visited the Upper Peninsula
5	I'm from the state of Michigan	Skipped < half of my lectures last week	I live on Central	I've customized my VSCode	Read 3+ books this year

Expectations:

Be responsible and show up!

Enjoy working collaboratively

Bring excitement about data science and analysis

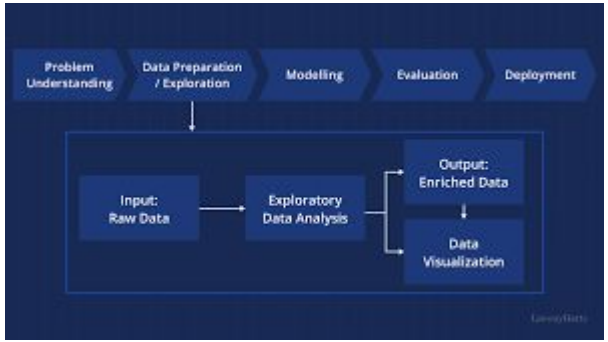
Hands-on practice (lectures 👎)

Gain hard skills each week

Final slides deliverable!

What is EDA?

- **Definition:**
  - EDA stands for Exploratory Data Analysis.
  - It is a process used to analyze datasets to summarize their main characteristics, often using visualization tools.



Main Goals of EDA:

- Understand Data Structure: Inspect the dataset to understand shape, types, and features.
- Identify Patterns and Relationships: Use statistics and plots to find correlations and trends.
- Detect Data Quality Issues: Locate missing values, outliers, and inconsistencies that may impact analysis.
- Generate Hypotheses: Develop questions and insights that can be further investigated or modeled.

Significance of EDA

- **EDA** is a philosophy of exploring data **without assumptions** to gain a thorough understanding of its **context**, **patterns**, and **limitations**.
- It reveals insights and helps in identifying **biases** and **relationships** that guide further analysis.
- **EDA** ensures data is **well-understood** and **clean**.
    - Many datasets have missing values such as NaNs or null values. This can mess up your code!
- It informs **stronger hypotheses** and leads to better **data-driven decisions**.

How We Will Use It Today

Team Member Data!

- Begin by answering a survey based on your information (Name, year, major, number of pets, etc.)
- Practice importing libraries and reading .csv files
- Understand basic data cleaning steps, such as removing unnecessary columns and imputing missing data.
- Make some discoveries about people on our project team!!

Python Libraries

Some of the most useful Python libraries known to man!! (and panda)

What is Pandas?

- pandas is a Python library for data manipulation and analysis.
- It provides two main data structures: **Series (1D)** and DataFrame **(2D, similar to a table in a database).**
- The name "pandas" is derived from "Panel Data" and "Python Data Analysis" - helps us **understand data structures**
- Identify **missing values** and discover **key patterns**
- Enables **data cleaning and preprocessing**—crucial steps to prepare the data for further analysis or modeling.

What is NumPy?

What is NumPy, and Why Are We Using It?

- NumPy stands for “Numerical Python” library
- Provides fast, efficient arrays and a wide range of mathematical functions.
- Key tool for scientific computing and handling large datasets effectively.

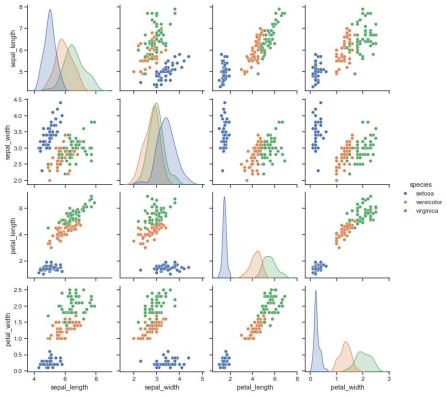
NumPy's Relationship with pandas and Its Role in EDA

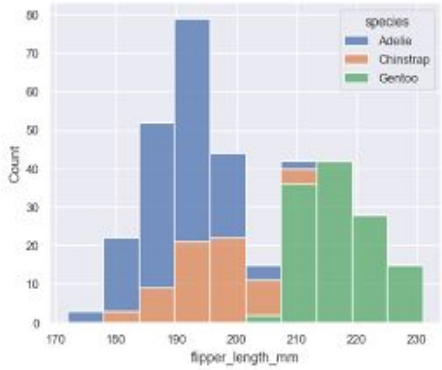
- Foundation for pandas: pandas is built on top of NumPy; pandas DataFrames internally use NumPy arrays.
- Exploratory Data Analysis (EDA): Enables data manipulation, statistical analysis, and complex calculations required for understanding data before visualization or modeling.

What is Seaborn?

What is Seaborn, and Why Are We Using It?

- Seaborn is a data visualization library built on top of Matplotlib, designed for creating attractive and informative statistical graphics.
- Provides high-level functions for visualizing data patterns, such as distributions, relationships, and trends.
- Easier and more aesthetic compared to Matplotlib for generating complex plots with fewer lines of code.
- Functions like **sns.countplot()**, **sns.boxplot()**, **sns.histplot()**, and **sns.pairplot()** provide valuable insights into dataset structure.





Important Functions to Know (pt. 1)

Syntax	Description
import pandas as pd	Import the **pandas** library, using the alias **pd** for convenience
import numpy as np	Import the **NumPy** library, using alias **np** for numerical operations
import seaborn as sns	Import **Seaborn** for statistical data visualization
df = pd.read_csv('file.csv')	Load a **CSV** file into a **pandas DataFrame** for data manipulation
df.head()	Returns the **first 5 rows** of the DataFrame, useful for quickly inspecting data
df.info()	Provides a concise **summary** of the DataFrame, including **data types** and **non-null values**

Important Functions to Know (pt. 2)

Syntax	Description
df.describe()	Generates **descriptive statistics** of numerical columns (mean, median, quartiles, etc.)
df[‘column_name’]	Access a **specific column** in the DataFrame, works like a key in a dictionary
df.drop(columns=[‘col’...])	**Drops specified columns** from the DataFrame,use inplace=True to modify the original DataFrame
df.index	Returns the **index labels** of the DataFrame
df.isnull()	returns a DataFrame of **boolean values**, where each entry indicates whether the corresponding value in df is NaN (missing).

Important Functions to Know (pt. 3)

Syntax	Description
df[‘column’].value_counts()	Returns the **count** of **unique values** in a specific column
df[‘column’].mean()	Returns the **mean value** of a numerical column
df.corr()	Computes **correlation** for numerical columns to understand relationships
df.columns	Lists all **column names** in the **DataFrame**, useful for renaming or viewing dataset structure
df.shape	Returns **# of rows and columns**

Important Functions to Know (pt. 4)

Syntax	Description
df.rename(columns={'old' : 'new'}, inplace=True)	Renames **specific columns** to new names
df.groupby(‘category’)[‘value’]	Groups the DataFrame using a specified column to perform **aggregate functions** (e.g., .sum(), .mean()).
df[‘new_column’] = df[‘column’].apply(function)	**Applies** a function to each element or column/row
df.shape	Returns a **tuple** representing the **dimensions** (rows, columns) of the DataFrame.
df.dropna()	Removes rows or columns containing **missing values**

Timeline

Week 1: Icebreaker/EDA intro (Programming/Python basics)

Week 2: EDA/Data Cleaning

Week 3: Error/Bias Analysis

Fall Break!! (No meeting 10/12)

Week 4: Logistic Regression

Week 5: Cox Proportional Hazards

Week 6: Kaplan Meier Curves

Week 7: Work Session (Form teams and brainstorm ideas)

Week 8: Work Session (Create slides for final presentation)




Survey!

Practice Time!

Let's learn more about each other while practicing Exploratory Data Analysis!

Hands-On Data Science!! :O

Next Steps:

1. Find the F25 CRA repo and download week1_pandas_practice.ipynb and F25_survey_data.csv in the MDST GitHub
   a. You can just Google “https://github.com/MichiganDataScienceTeam”
2. Split into teams of 3-5 and introduce yourselves!
   a. Name, hometown, year, intended major, favorite UMich memory, hobbies
3. Work on the exercises in the notebook!
   a. You are free to go as soon as you’re finished, but we encourage you to stick around and help your teammates!

Pandas Cheat Sheet Seaborn Cheat Sheet