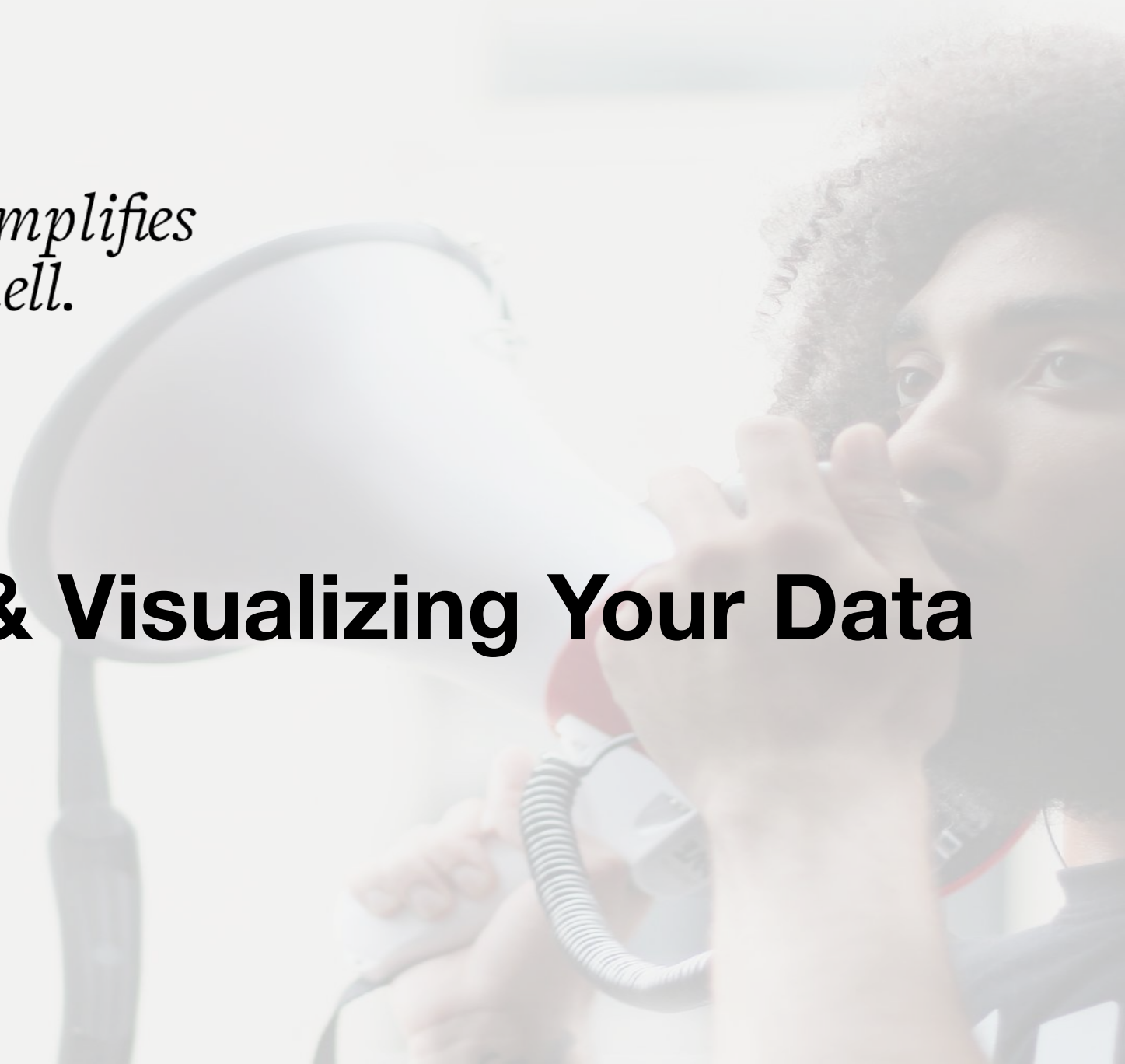
CORNELL CENTER for
SOCIAL SCIENCES

Accelerates, enhances, and amplifies social science research at Cornell.



Understanding & Visualizing Your Data

Remy Stewart
CCSS-ResearchSupport@cornell.edu

CCSS Research Support Code of Conduct

The Cornell Center for Social Sciences provides a welcoming environment for everyone embracing all backgrounds or identities. All instructors and attendees agree to abide by our community norms. We encourage the following behaviors in our workshops:

- Respect differing viewpoints and ideas
- Share your own perspectives and ask any questions
- Accept constructive criticism
- Use welcoming and inclusive language
- Show courtesy and respect for all instructors and attendees

If you believe that an instructor or attendee has violated the code of conduct, please report the violation to CCSS-ResearchSupport@cornell.edu. We take all reported incidents seriously.

Land Acknowledgement

Cornell University is located on the traditional homelands of the Gayogo̱hó:nǫ' (the Cayuga Nation). The Gayogo̱hó:nǫ' are members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic and contemporary presence on this land. The Confederacy precedes the establishment of Cornell University, New York state, and the United States of America. We acknowledge the painful history of Gayogo̱hó:nǫ' dispossession, and honor the ongoing connection of Gayogo̱hó:nǫ' people, past and present, to these lands and waters.

Here are additional links for more on the history of Cornell's violent, colonial formation, the movement to return native lands, and about the AIISP program at Cornell.

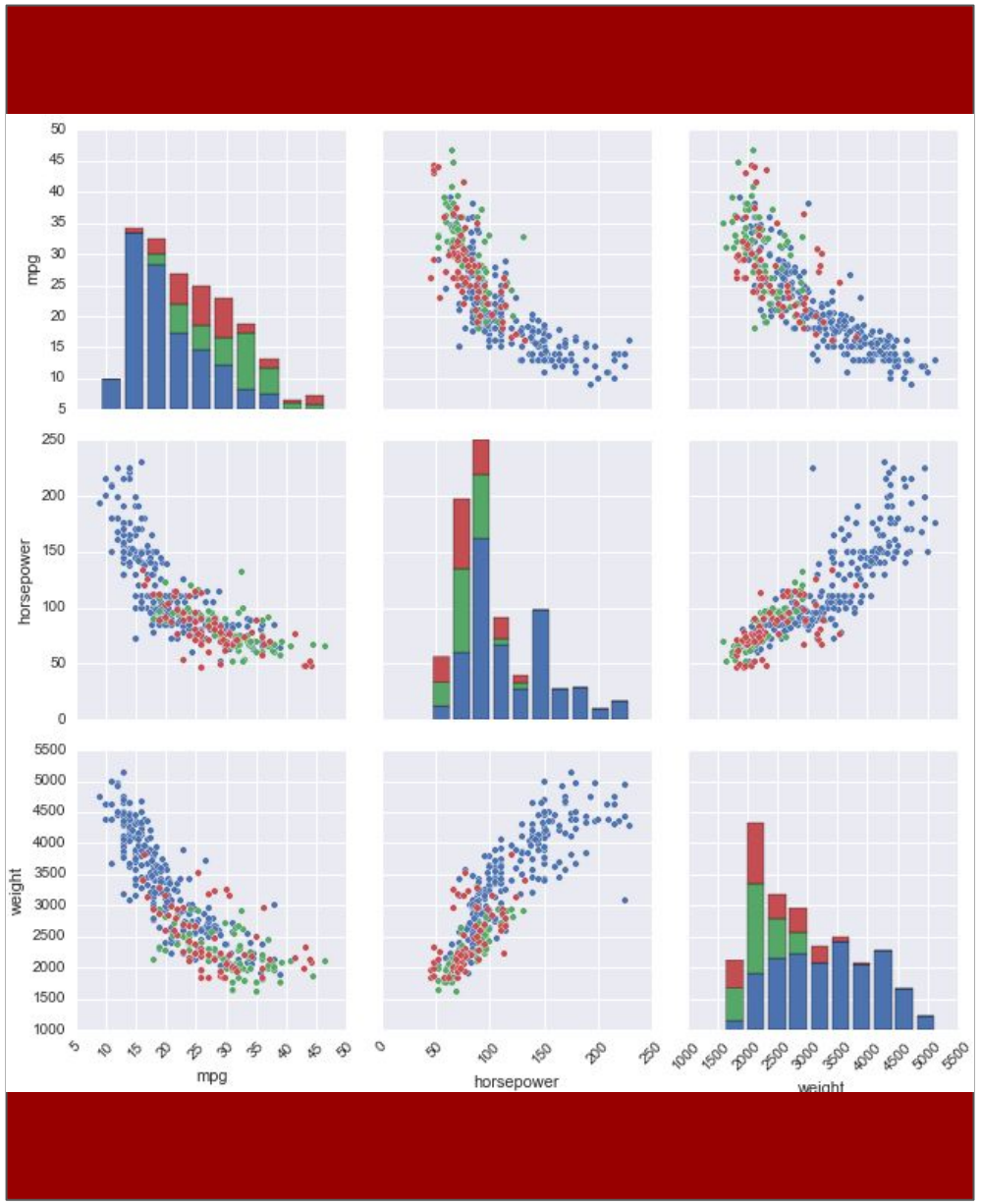
Consider donating to the Gayogo̱hó:nǫ' sovereignty initiative here.

Understanding Data through Exploratory Data Analysis (EDA)

- EDA refers to the examination, transformation, and preparation of data before initializing any ML modeling
- Core preliminary step to build robust ML workflows
- Different approaches for data types such as numeric, categorical, text, etc.
- While methods themselves vary, EDA's importance is widespread across ML applications

The Power of Data Visualization

- Guides interpretation beyond numbers & summary statistics
- Unveils trends & unexpected patterns
- Alerts us to potential issues within data
- Facilitates comparisons between groups & variables
- Translates key takeaways for diverse audiences



Questions regarding EDA for Machine Learning

How does data need to be uniquely prepared for ML?

How can we prioritize certain variables over others to build robust ML models?

What makes EDA different for structured vs. unstructured data?

What are the pros and cons of different EDA decisions?

Coding Demo Outline

Structured Data

Descriptive Statistics

- Averages
- Distributions
- Co-Occurances

→

Data Visualization

- Histograms
- Count Plots
- Box Plots

→

Feature Selection & Creation

- Heatmaps & Correlations
- Outliers
- Missing Data

→

Feature Engineering

- Binarization
- One-Hot Encoding
- Min-Max Scaling
- Standardization

Unstructured (Text) Data

String Manipulation

- Deletion
-Concatenation
-Lowercasing

→

Data Distributions

- Label Frequencies
- Word Counts

→

Tokenization & Vectorization

- "Bag of Words"
- TF-IDF

→

Vocabulary Manipulation

- Token Frequencies
- Stopwords
- Stemming

Let's head to the Github to start our first coding demo!

Alternatively, you can access this session's Colab file directly via the following link:

https://colab.research.google.com/drive/1Tj2dRofpZ2Nj5zojOWsKVKJ9tr2GxbgR?usp=sharing