BSBINS401 - Analyse and Present Research Information

Session 3: Getting Started with Kaggle and Data Visualization

Lecturer: Jordan Hill

Learning Objectives

1. Understand how to create and navigate your Kaggle account and profile.
2. Explore Kaggle datasets for research and analysis.
3. Complete Kaggle Learn: Intro to Data Visualization, Lesson 1: Hello, Seaborn.
4. Apply fundamental data visualization principles in Python.
5. Review key NumPy concepts from “Python for Data Analysis” (Chapter 4).

Key Topics for Session 3

- Kaggle Account Creation & Profile Setup
- Touring the Kaggle Platform (Competitions, Datasets, Notebooks)
- Principles of Data Visualization:
  - Clarity, Comparisons, and Context
- Kaggle Learn: Intro to Data Visualization, Lesson 1 (Hello, Seaborn)
- NumPy Essentials and the Role of Arrays in Data Analysis

1. Kaggle Setup: Creating an Account (If you haven't already)

- Visit https://www.kaggle.com/account/signup
- Provide your preferred email and create a password (or link to Google/GitHub).
- Verify your account via email confirmation.
- Configure your Kaggle profile (optional but encouraged):
  - Add a profile picture and bio.
  - Indicate any areas of interest (e.g. machine learning, data visualization).

2. Navigating Kaggle

- **Kaggle Competitions**
  - Participate in real-world data science challenges.
  - Learn from other participants' public code notebooks.
- **Kaggle Datasets**
  - Kaggle Datasets: Repository of diverse public datasets.
  - Explore dataset previews, metadata, and discussions.
- **Kaggle Notebooks**
  - In-browser environment for Python and R.
  - Collaborative notebooks are shareable and reproducible.

3. Principles of Data Visualization

- **Clarity**: Ensure your plots tell a clear story—add labels, titles, legends.
- **Comparison**: Visuals should facilitate easy comparison across categories or over time.
- **Context**: Provide necessary background or supporting information so that viewers can interpret your graph.
- **Color and Aesthetics**: Use coherent color schemes; avoid visual clutter.

<iframe width="100%" height="580" src="https://www.youtube.com/embed/42RojZSB0Yg?si=4-tS5Lsnlh6bO1Xi" title="YouTube video player" frameborder="0" allow="accelerometer; autoplay; clipboard-write; encrypted-media; gyroscope; picture-in-picture; web-share" referrerpolicy="strict-origin-when-cross-origin" allowfullscreen></iframe>

4. Kaggle Learn: Intro to Data Visualization

- Go to Kaggle Learn and select “Lesson 1: Hello, Seaborn.”
- Follow the step-by-step notebook instructions.
- Practice:
  - Customize the example plots.
  - Try different Seaborn functions (e.g., barplot, scatterplot).
- Discuss:
  - Which charts best fit your data or prospective research questions?

5. Review of NumPy Basics (“Python for Data Analysis,” Chapter 4)

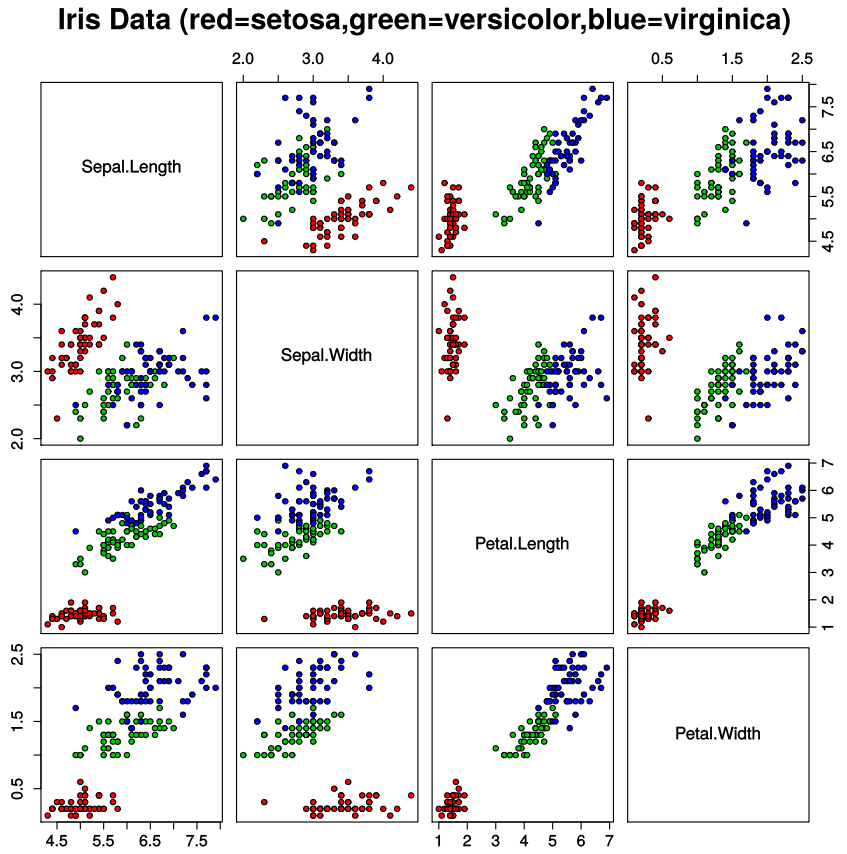
- **What is NumPy?**
  - A Python package for scientific computing, focusing on arrays and fast operations.
- **Arrays**
  - Homogeneous collection of items (same data type) in a continuous block of memory.
- **Key Functions**
  - `np.array()`, `np.arange()`, `np.ones()`, `np.zeros()`, shape manipulations, indexing, and slicing.
- **Why it Matters**
  - Nearly every data manipulation or ML pipeline in Python starts with NumPy arrays or uses libraries that rely on them.

Code Snippet: NumPy Demo

```
import numpy as np

# Create a 1D array
arr_1d = np.array([1, 2, 3, 4, 5])

# Create a 2D array (matrix)
arr_2d = np.array([[1, 2, 3],
                   [4, 5, 6]])

print("1D Array:", arr_1d)
print("2D Array Shape:", arr_2d.shape)

# Perform arithmetic
arr_sum = arr_1d + 10
print("Add 10 to each element in 1D Array:", arr_sum)
```

Sample Seaborn Visualization

```
import seaborn as sns
import matplotlib.pyplot as plt

# Load built-in example dataset
data = sns.load_dataset("iris")

# Create a pairplot
sns.pairplot(data, hue="species")
plt.suptitle("Pairplot of Iris Dataset", y=1.02)
plt.show()
```



In-Class Activities

1. **Hands-On Kaggle Platform Exploration**

   - Explore Datasets and Competitions.
   - Create a new Notebook on Kaggle.
   - Experiment with simple Python commands.

2. **Find a dataset on Kaggle, Can you locate the original source of the data?**

- What columns are in the dataset?
- What kinds of datatypes are in the dataset?
- Try to find the original source of the data.

Discussion Prompt

1. What types of datasets on Kaggle spark your interest and why?
2. How do you see data visualization principles influencing the way you present your research questions?
3. Any barriers or issues encountered while signing up/in to Kaggle or running notebooks?

Discuss these questions with the person next to you.

In-Class Activities

3. **Practice: Kaggle Learn Module**
    - Open the “Hello, Seaborn” tutorial and read through the lesson.
    - Using the tutorial write a quick snippet of Seaborn code to visualize a sample dataset found previously.
    - Complete the Exercise: “Hello, Seaborn”

Lab Resources

1. Kaggle Interface Exploration
   - Create a simple Python notebook; run basic commands.

2. Kaggle Learn Module - Lesson 1
   - Follow the “Hello, Seaborn” steps to practice data visualization.

3. NumPy Basics Reference
   - Reinforce array manipulation skills for future data exploration.

Next Steps

- **Continue** working through Kaggle Learn: Data Visualization lessons.
- **Review** Chapter 4 of “Python for Data Analysis” to strengthen NumPy and array manipulation skills.
- **Brainstorm**: Think about future datasets and research questions you might explore for assessments.

Questions?

- Office Hours: Mon–Fri, 9 AM – 5 PM
- Email: jordan.hill@nmtafe.wa.edu.au