



DS PORTFOLIO SESSION 9

**SKILLS
FOR LIFE**

SKILLS BOOTCAMPS



Department
for Education

Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
(FBV: Mutual Respect.)
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Open Classes.
You can submit these questions here: [Open Class Questions](#)

Data Science Session Housekeeping cont.

- For all **non-academic questions**, please submit a query:
www.hyperiondev.com/support
- Report a **safeguarding** incident:
www.hyperiondev.com/safeguardreporting
- We would love your **feedback** on lectures: [Feedback on Lectures](#)

Progression Criteria

✓ **Criterion 1: Initial Requirements**

- Complete 15 hours of Guided Learning Hours and the first four tasks within two weeks.

✓ **Criterion 2: Mid-Course Progress**

- Software Engineering: Finish 14 tasks by week 8.
- Data Science: Finish 13 tasks by week 8.

✓ **Criterion 3: Post-Course Progress**

- Complete all mandatory tasks by 24th March 2024.
- Record an Invitation to Interview within 4 weeks of course completion, or by 30th March 2024.
- Achieve 112 GLH by 24th March 2024.

✓ **Criterion 4: Employability**

- Record a Final Job Outcome within 12 weeks of graduation, or by 23rd September 2024.

Recap of Week 9: Data Visualisation

Approach to Visualisation

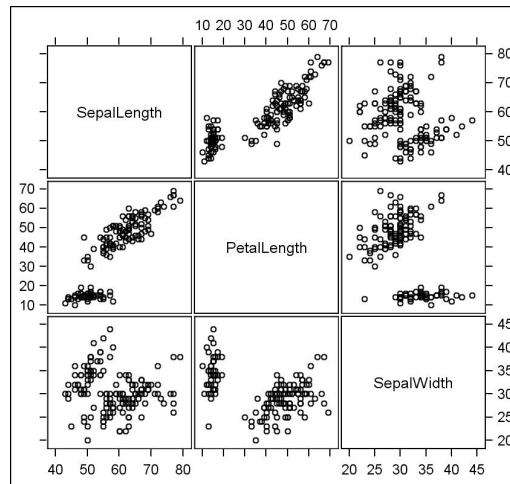
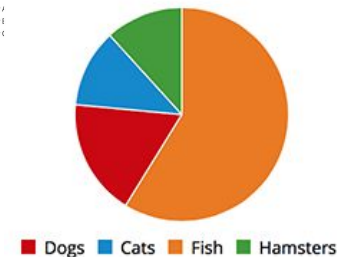
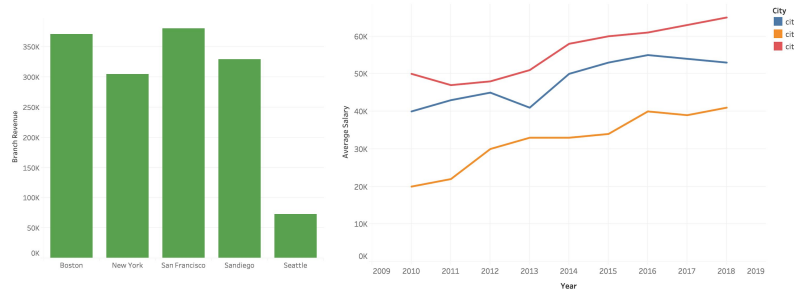
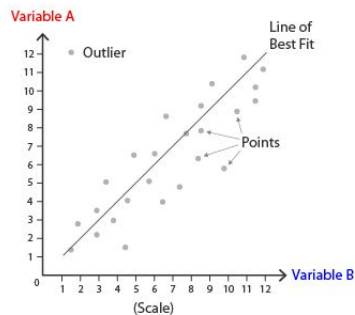
1. Start with a processed and clean dataset.
2. Know your dataset.
3. Determine what you want to find.
4. Create data visualisations.
5. Refine your visualisation.
6. Note down your findings.


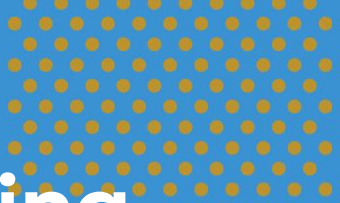
Types of Data:

- Discrete
- Categorical
- Continuous
- Time Series.

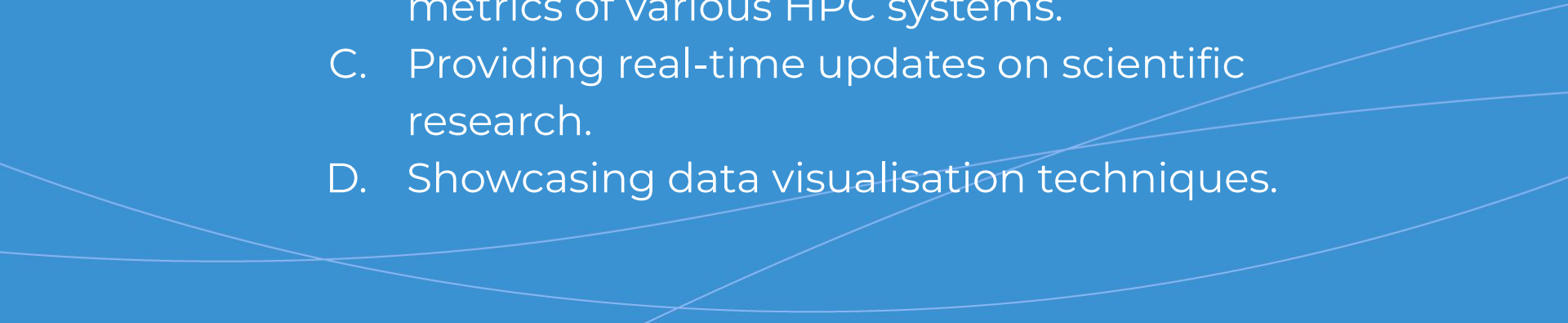
Types of Data:

- Bar Chart
- Line Graphs
- Pie Chart
- Scatterplot
- Scatterplot Matrix
- Double Axis Chart





What is the purpose of using visualisations like bar graphs and line graphs?

- A. Enhancing graphics.
 - B. Comparing and contrasting the performance metrics of various HPC systems.
 - C. Providing real-time updates on scientific research.
 - D. Showcasing data visualisation techniques.
- 

High-Performance Computing Dashboard

- **Background:** Building on the foundation laid in Week 9, we now delve deeper into the world of High Performance Computing (HPC) systems, shifting our focus from visualisation to in-depth analysis.
- **Challenge:** The HPC Insight Dashboard, developed in the previous week, has provided a comprehensive visual representation of the performance metrics of various HPC systems. Now, you're tasked with extracting deeper insights from this data. Using the power of Pandas in Jupyter Notebook, you will manipulate the benchmark data to uncover hidden trends, correlations, and anomalies.

- **Objective:**

- Computing and comparing statistical data to derive nuanced insights about HPC systems.

Demo: In-depth analysis

```
# Simple example to demonstrate the need for in-depth analysis in HPC
benchmark data
import pandas as pd

# Assume hpc_data is a Pandas DataFrame with benchmark data
hpc_data = pd.DataFrame({
    'HPC': ['HPC1', 'HPC2', 'HPC3'],
    'Iteration_Speed': [90, 120, 100],
    'Calculation_Speed': [120, 80, 100],
    'Memory_Capacity': [16, 32, 24]
})

# Extract deeper insights
average_iteration_speed = hpc_data['Iteration_Speed'].mean()
max_memory_capacity = hpc_data['Memory_Capacity'].max()

print(f"Average Iteration Speed: {average_iteration_speed}")
print(f"Maximum Memory Capacity: {max_memory_capacity}")
```

Demo: In-depth analysis Continued

```
# Extended example to demonstrate advanced data manipulation in Pandas for
HPC analysis
# (Demonstration may include additional features not covered in the example)

# Load the benchmark data into a Pandas DataFrame (assuming it's available
as a CSV file)
hpc_data = pd.read_csv('hpc_benchmark_data.csv')

# Selecting specific data segments
selected_data = hpc_data[['HPC', 'Iteration_Speed', 'Calculation_Speed']]

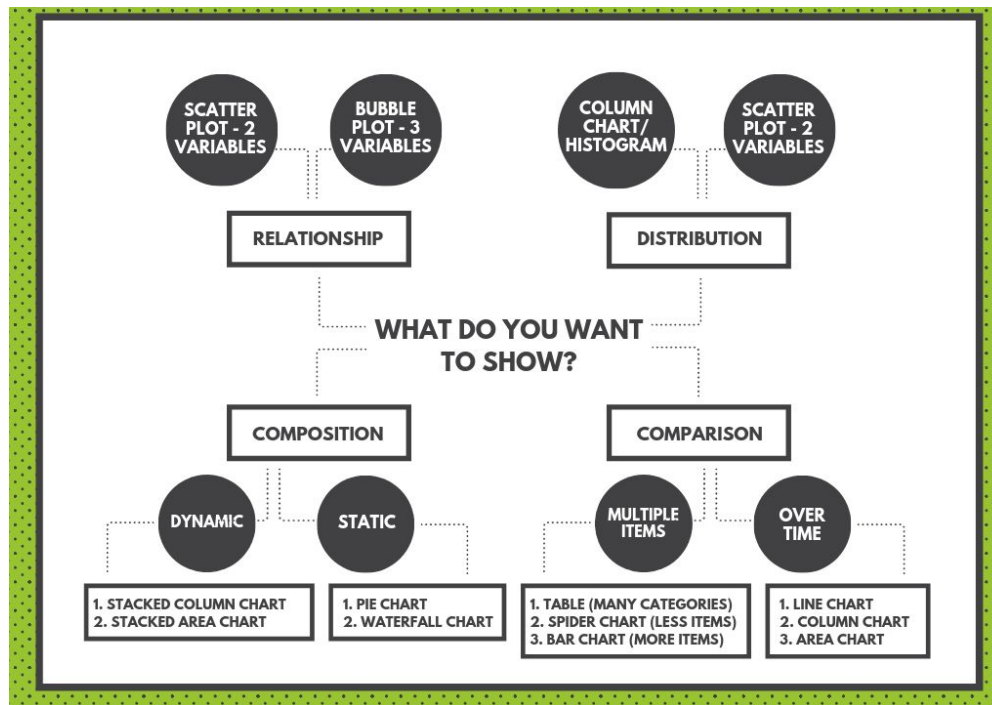
# Sorting data based on Iteration Speed
sorted_data = hpc_data.sort_values(by='Iteration_Speed', ascending=False)

# Analysing specific data using iloc
specific_data = hpc_data.iloc[0:3, 1:4] # Selecting rows 0 to 2 and columns
1 to 3

# Computing statistical data
average_calculation_speed = hpc_data['Calculation_Speed'].mean()
total_memory_capacity = hpc_data['Memory_Capacity'].sum()

print(f"Average Calculation Speed: {average_calculation_speed}")
print(f"Total Memory Capacity: {total_memory_capacity}")
```

Deciding on Type of Visualisation



HPC

Building on the foundation laid in Week 9, we now delve deeper into the world of High Performance Computing (HPC) systems, shifting our focus from visualisation to in-depth analysis.

Vital libraries and concepts for this task:

Numpy Arrays
Matplotlib (PyPlot)
Pandas
Tableau


Important Concepts:

1. **Advanced data manipulation:** Using Pandas in Jupyter Notebook, focusing on selecting, sorting, and analysing HPC benchmark data.
2. **Visualisations tell a story:** Interpreting multi-dimensional data representations to uncover deeper insights about HPC performance metrics.
3. **Documentation:** Document and present data analysis findings in a structured, clear, and replicable manner within Jupyter Notebook.

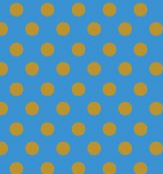
Advanced

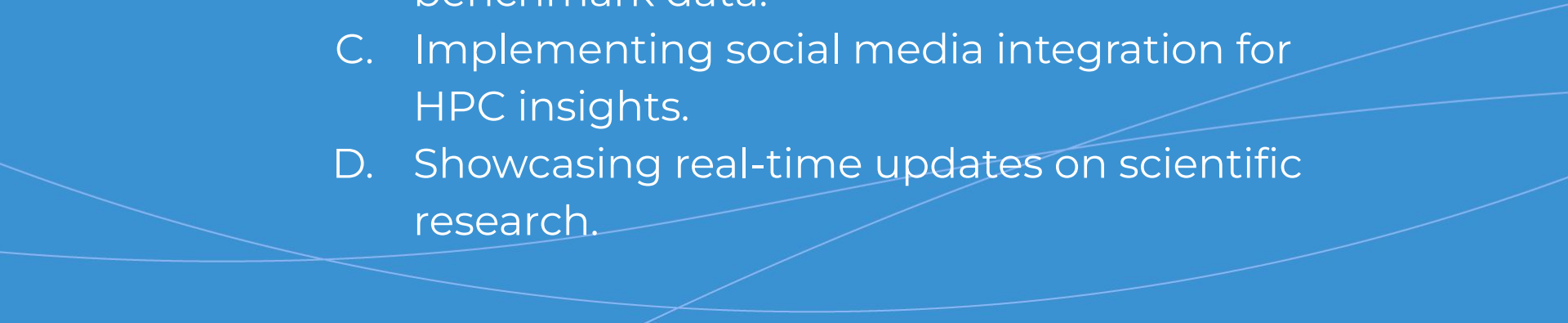
Challenge:

- If applicable, allow users to choose multiple visualisation types to view data.



In the HPC analysis project, what is the primary focus of using Pandas in Jupyter Notebook?



- A. Enhancing graphics for the HPC Insight Dashboard.
 - B. Conducting in-depth analysis of HPC benchmark data.
 - C. Implementing social media integration for HPC insights.
 - D. Showcasing real-time updates on scientific research.
- 

Summary

Data Visualisations

- ★ Data visualisation is the practice of translating information into a visual context. Remember to use the approach mentioned earlier.

Pandas

- ★ Library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data



Questions and Answers

Questions around the Case Study

