BSBINS401 - Analyse and Present Research Information

Session 4: Exploring the Standard Dataset

Lecturer: Jordan Hill

Learning Objectives

1. Introduce the standard dataset provided for analysis.

2. Understand the dataset’s structure, key features, and context.

3. Practice data exploration techniques using Pandas.

4. Identify potential research questions from preliminary insights.

5. Reinforce best practices for data integrity and handling.

Session Overview

- Dataset Introduction & Context
- Research Objectives & Relevance
- Data Exploration Techniques
- In-Class Hands-On Activities
- Lab & Reading Resources
- Next Steps

The Standard Dataset

- **Source**: Provided from Kaggle or local CSV file
- **Domain Example**: HR, education, healthcare
- **Content Overview**:
  - Number of rows & columns
  - Key features (e.g., department, attendance, attrition)
- **Purpose**:
  - Practice data cleaning, exploration, and visualization

Research Alignment

- **Identify Metrics**:
  - Look for variables such as “employeeAttrition”, “department”.
- **Develop Questions**:
  - “What factors correlate most with [X]?”
  - “Which groups exhibit higher rates of [Y]?”
- **Brainstorm Ideas** during exploration.

Data Exploration Techniques

1. **Inspecting Data Structure**
   - Use: df.head(), df.tail(), df.info()
2. **Descriptive Statistics**
   - Use: df.describe() for means, std, quartiles
3. **Detecting Issues**
   - Use: df.dtypes, df.isna().sum()
4. **Visual Exploration**
   - Create histograms, bar plots, and value counts

Code Example: Basic Pandas Exploration

```python
import pandas as pd

# Load the dataset (update the file path as needed)
df = pd.read_csv("your_dataset.csv")

print("DataFrame Shape:", df.shape)
print("\nData Types:\n", df.dtypes)

# Display first few rows
display(df.head())

# Summary statistics
display(df.describe())

# Check for missing values
print("\nMissing Values:\n", df.isna().sum())
```

Lab 1

Lab Exercises:

- Complete the Hands-on Data Exploration with Pandas lessons.

In-Class Activity

1. **Load the Dataset**
    - Use your preferred environment (Jupyter, Kaggle, Colab).
2. **Inspect the Data**
    - Check dimensions, column names, and data types.
3. **Discussion Points**
    - Share your first impressions and any anomalies observed.
4. **Visualize**
    - Create a simple histogram or bar chart for a selected feature.

Lab 2

Load the Dataset > Inspect the Data > Discuss > Visualize

- Work using one of these datasets:
  - Employee Attrition
  - Student MBB Degree College Data
  - Student School Attendance

Share your first impressions and any anomalies observed.

Create a simple histogram or bar chart for a selected feature.

Reading Resources

Reading:

- Data Exploration Techniques
- Dataset Documentation on Kaggle
- “Python for Data Analysis” by Wes McKinney, Chapter 5

Quick Tips for Data Integrity

- **Backup Data**: Keep raw data unaltered.
- **Version Control**: Use Git for tracking changes.
- **Documentation**: Record every cleaning step and assumption.
- **Ethics & Compliance**: Ensure privacy and anonymize sensitive data.

Discussion & Next Steps

Discussion Points:

- Which dataset features caught your attention?
- What anomalies or missing data patterns did you observe?
- How could these insights inform research questions?

Next Steps:

1. Refine your research questions:

   submit 2-3 questions at the next session as a printed 1-page document .

2. Explore advanced Pandas functions like groupby and pivot tables.
3. Prepare initial visualizations for further analysis.

Questions or Comments?

Contact during office hours:

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