BSBINS401 - Analyse and PresentResearch Information

Session 5: Advanced Exploratory Data Analysis & Report Preparation

Lecturer: Jordan Hill

Learning Objectives

- 1. Explore advanced approaches to Exploratory Data Analysis (EDA) in Python.
- 2. Learn techniques to refine and interpret analysis results.
- 3. Begin structuring the Assessment 1 report, incorporating a clear data narrative.
- 4. Develop familiarity with correlation analysis and group-based insights.

Assessment 1 Reminder

Assessment 1 Due: Week 7

Components:

- 1. A report detailing your data cleaning, analysis steps, and findings
- 2. An in-class presentation summarizing your key insights

Session Overview

- 1. Common pitfalls in data analysis and how to avoid them
- 2. Advanced Pandas features (groupby, pivot tables)
- 3. Correlation and basic hypothesis testing concepts
- 4. Structuring the final report
- 5. Hands-on notebook walkthroughs and discussions

1. Common Pitfalls in Data Analysis

Overlooking Data Types:

- Ensure columns are correctly typed (numeric vs. categorical vs. datetime)
- Convert string-based numbers (e.g., "25") to int

Assuming Missing Data is Random:

Investigate patterns—missingness might be systematic

Failing to Document Changes:

Maintain a clear record of each cleaning, filtering, or transformation step

Relying on a Single Visualization:

Use multiple plot types to confirm or challenge your initial assumptions

2. Advanced Pandas Tools

Groupby and Pivot Tables

- groupby()
 - Perform split-apply-combine operations (e.g., average by department)
- pivot_table()

Reshape data and compute aggregated statistics in a spreadsheet-like pivot format

Example: Grouping Employee Data by Department

```
import pandas as pd

df = pd.read_csv("employee_attrition.csv")
dept_stats = df.groupby("Department")["MonthlyIncome"].mean()

print("Average Monthly Income by Department:")
print(dept stats)
```

Applying Pivot Tables for Deeper Insights

```
pivot_example = pd.pivot_table(
    df,
    values="MonthlyIncome",
    index="EducationField",
    columns="Gender",
    aggfunc="mean"
)

print("Pivot Table Example:\n", pivot example)
```

- Quick way to compare numeric outcomes across multiple dimensions (e.g., gender + education)
- Helps spot potential trends or disparities

3. Correlation Analysis

Why Correlation?

- Measures the linear relationship between pairs of variables
- Helps identify features that might be important in your research questions

Creating a Correlation Matrix

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

corr_matrix = df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap="Blues")
plt.title("Correlation Matrix of Dataset")
plt.show()
```

Interpretation:

- Values range from -1 to +1
- Closer to +1 indicates strong positive correlation, -1 indicates strong negative correlation
- 0 means no linear relationship

4. Structuring Your Assessment 1 Report

Section	Components	
1. Introduction	Research questions/objectivesOverview of dataset	
2. Methodology	Data cleaning steps and justificationTools and libraries used	
3. Analysis & Findings	EDA techniques, visualizations, correlation insightsObservations and potential explanations	
4. Conclusions & Recommendations	Summarize key insightsSuggest actions or further research	
5. References & Appendices	Citation of data sources and external references	

5. Hands-On Notebook Walkthrough

1. Data Wrangling

- Convert columns to correct data types
- Handle outliers or erroneous values

2. Exploratory Plots

- Bar plots of grouped means, standard deviations
- Correlation heatmap for feature relationships

3. Interpreting Results

- Identify top drivers or patterns
- Tie outcomes back to your research questions

Activity: Building Your Draft Analysis

1. Group Analysis:

• Using your dataset, apply a groupby or pivot table to examine at least one important metric by group (e.g., average exam score by major, average income by department).

2. Correlation Matrix:

• Generate a correlation heatmap to see how variables relate. Spot any strong correlations or interesting negative relationships.

3. Discussion:

Share any surprising findings or confirm previously known insights.

Key Takeaways

- Advanced EDA:
 - groupby, pivot tables, correlation
- Data Narrative:
 - Connect your plots and stats back to research questions
- Report Preparation:
 - Start drafting your methodology and initial findings now

Additional Resources

Resource	Description	Link
Pandas Documentation	Detailed guide on groupby, pivot tables	groupby
Kaggle Learn	Advanced data cleaning & pivot tables	Kaggle: Pandas
Seaborn Docs	Correlation heatmaps and advanced plots	Seaborn Examples

Questions & Next Steps

Any Questions?

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

Next Steps:

- Continue refining your EDA.
- Incorporate new findings into your Assessment 1 draft.
- Prepare to share partial analyses or rough visualizations for feedback in the next session.