BSBINS401 - Analyse and Present Research Information

Week 8: Introducing More Advanced Data Analysis Techniques

Lecturer: Jordan Hill

Learning Objectives

1. Understand why we might need more detailed analysis methods beyond basic summaries.
2. Explore basic approaches to identifying patterns and trends when multiple variables are involved (multivariate thinking).
3. Gain familiarity with simple ways to compare groups (basic statistical tests).
4. Learn how to visualize more complex relationships (Kaggle Learn Lesson 4: Scatter Plots, Lesson 5: Distributions).
5. Reinforce best practices in data visualization and interpretation.

Session Overview

- Why Look Beyond Simple Charts?
- Comparing Groups: Basic Statistical Ideas
- Introduction to “Multivariate Thinking”
- Handling Complex Data (reminder: cleaning, missing values)
- Hands-on with Kaggle Learn: Scatter Plots & Distributions
- Activities and Next Steps

1. Why Look Beyond Simple Charts?

- When we have multiple variables (e.g., age, salary, department), a single average or bar chart might miss important information.
- Organizations often need deeper insights:
  - “Which factors might relate to higher sales?”
  - “Do certain groups consistently score higher or show different behaviors?”
- Moving beyond simple summaries can highlight hidden relationships or patterns.

2. Comparing Groups: Basic Statistical Ideas

Example Scenario

• Suppose you have two groups (Group A and Group B) and you want to see if there's a meaningful difference in their average test score or average salary.

Approach	What It Does
Visual Comparison	Simple bar chart or box plot side-by-side to see if one group is higher than the other.
Basic Statistical Test (e.g., T-test)	Checks (mathematically) if the difference you see might be due to chance or is likely real. *(We typically get a "p-value" – a smaller p-value often suggests a real difference.)*

3. Introduction to “Multivariate Thinking”

- **What does “multivariate” mean?**
  It simply means two or more variables in play (e.g., salary, age, department, and performance score).

- **Why is it helpful?**
  - Real-world situations rarely hinge on just one variable.
  - Sometimes, a combination of factors influences an outcome (e.g., department + age + education might all affect salary).

- **Example**:
  - Could an employee’s “salary” be related both to “age” and “years of education”?
  - Exploring more than one variable at a time gives richer insights.

4. Handling Complex Data (A Reminder)

1. Data Cleaning - Check for missing values or inconsistent data. - Decide how to handle them (drop, fill, or investigate further).	**3. Visualization** - Summaries first (like describe(), histograms). - Then consider more advanced plots (scatter plots across multiple variables).
2. Documentation - Keep a record of any cleaning steps or assumptions.	**4. Interpret Carefully** - A pattern doesn't always mean "cause and effect." - We look for associations or relationships, not instant proof of causation.

5. Hands-On: Kaggle Learn (Scatter Plots & Distributions)

Lesson	Key Focus	Link
Lesson 4: Scatter Plots	Visualizing relationships between two numeric variables.	Kaggle Scatter Plots
Lesson 5: Distributions	Understanding how observations spread out (histograms, box plots, etc.).	Kaggle Distributions

Activity: Visualizing Multiple Variables

1. **Scatter Plot**: Try plotting “salary” vs. “age,” and add color or markers for “department.”
2. **Distribution Plot**: Examine how “salary” (or another continuous variable) is distributed. Is it skewed? Are there outliers?
3. **Discussion**:
   - Which variables might you compare in your own project?
   - Did you notice any interesting clusters or big differences between groups?

Additional Notes on Basic Analysis

- **Correlation**:
  - A quick measure to see if two numeric variables move together (positively or negatively).
  - Ranges from -1 (perfect negative) to +1 (perfect positive). A correlation near 0 means little linear relationship.
  - Remember: correlation ≠ causation.
- **Regression**
  - A method to see how one or more variables predict another.

6. In-Class Exercises

1. **Scatter Plot & Correlation Check**
   - Choose any two numeric variables from a dataset (e.g., Kaggle's "Employee Attrition" or your own).
   - Create a scatter plot; calculate correlation with a single line of code (e.g., df.corr()).
2. **Compare Two Groups**
   - If you have a categorical variable with two groups (like "North" vs "South"), compare a numeric outcome (like average monthly income). A bar chart is fine!
   - (Optional) For the more adventurous, attempt a straightforward t-test with your lecturer's guidance.

Next Steps

1. **Practice** the scatter plot, correlation, and group comparison methods on your own dataset.
2. **Document** any new findings or interesting relationships.
3. **Consider** how these methods might inform the next stage of your research project or final report.

Questions?

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