Hands-On Activity #4:

Comparing T-Test Results from Python and R

Objective:

Compare the results of a t-test calculated using Python's `scipy.stats.ttest` and R's `t.test` to identify and visualize differences.

Data:

T-test results comparing drugs that preferentially kill pancreatic vs. lung cancer cells.

1. Derived using `tidyverse` and `t.test()`
    a. https://github.com/WayScience/CPBS7601/blob/main/lectures/4.data_wrangling/results/ttest_results_tidyverse.csv
2. Derived using `pandas` and `scipy.stats.ttest_ind()`
    a. https://github.com/WayScience/CPBS7601/blob/main/lectures/4.data_wrangling/results/ttest_results_pandas.csv

In *a jupyter notebook* follow the instructions:

Instructions:

1. **Prepare your workspace:**
    - Start by organizing your workspace and loading the two files: one containing t-test results calculated with Python's `scipy` and the other with R's `t.test`.
    - Ensure both files are in a tidy data format, with each row representing a single observation and each column representing a variable.
2. **Load and align the data:**
    - Load both files into dataframes (e.g., using `pandas` in Python).
    - Ensure that the dataframes are properly aligned for comparison. Check that the same observations (rows) and variables (columns) are matched between the two dataframes.
3. **Identify differences:**
    - Calculate the differences between the corresponding values in the two dataframes. Consider differences in the t-statistic, p-values, or any other relevant metrics.
    - Quantify the differences and store the results in a new dataframe.
4. **Visualize the differences:**
    - Choose a visualization method (e.g., histogram, density plot, boxplot, or heatmap) to represent the differences between the two sets of t-test results.
    - Create the chosen visualization to effectively illustrate the magnitude and distribution of the differences.
5. **Analyze the biggest differences:**
    - Identify the largest differences in your comparison.

    - Provide an analysis of why these differences might exist. Consider factors such as numerical precision, algorithmic implementation differences, or other potential causes.

Deliverables:

- A jupyter notebook html file that includes:
    - A brief description of the two datasets and how they were aligned.
    - The method used to quantify differences between the t-test results.
    - The visualization(s) of the differences.
    - An analysis of the biggest differences and possible reasons for their occurrence.

Tips:

- Ensure your data is properly aligned before calculating differences.
- Experiment with different types of visualizations to best represent the data.
- Consider both statistical and computational factors that could lead to differences in the results.

Tools:

- `jupyter notebook`
- Use Python libraries such as `pandas`, `matplotlib`, `seaborn`, or `plotly` for data manipulation and visualization.
- Consider using `R` if you prefer to work with R-based tools for any part of the analysis.