Statistics For Dummies Tool

Hriday Baghar
Ernesto Cediel
Keegan Freeman
Andy Tsai

UNIVERSITY *of* WASHINGTON

Background and Use Cases

- Python tool that allows users to import data sets to perform data visualization and statistical hypothesis testing in a web application without writing any code.
- Built and Deployed with Streamlit:

  Streamlit is an open-source app framework in Python built specifically for Machine Learning and Data Science

Streamlit

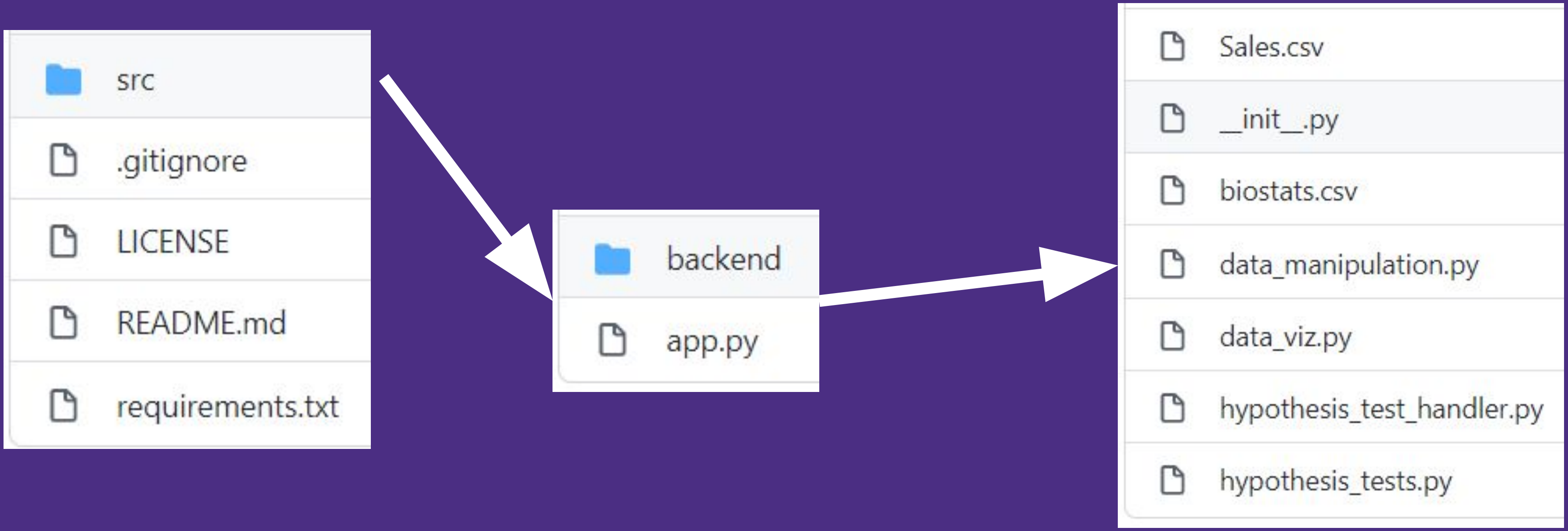
User Workflow

1. Upload a data file (CSV, XLSX, JSON)
2. Select Data Action
   - Data Visualization
   - Statistical Testing
3. Select
   - Visualization Type (Histogram, Scatterplot, Bar Graph, Line Graph, Box Plot, Correlation Heatmap)

   OR

   - Hypothesis Test (Z-Test, T-Test, ANOVA)
4. Select Dataframe Columns

GitHub Structure



src
.gitignore
LICENSE
README.md
requirements.txt

backend
app.py

Sales.csv
__init__.py
biostats.csv
data_manipulation.py
data_viz.py
hypothesis_test_handler.py
hypothesis_tests.py

GitHub Link:
https://github.com/hbaghar/statistics-for-dummies

Code Structure

```
statistics-for-dummies
├─ .gitignore
├─ LICENSE
├─ README.md
├─ datasets
│  ├─ Iris.csv
│  ├─ Sales.csv
│  └─ biostats.csv
├─ requirements.txt
└─ src
   ├─ app.py
   └─ backend
      ├─ __init__.py
      ├─ data_manipulation.py
      ├─ data_viz.py
      ├─ hypothesis_test_handler.py
      └─ hypothesis_tests.py
```

GitHub Repo Link:
https://github.com/hbaghar/statistics-for-dummies

Design

Separation of Concerns

- App script for UI
- Separate modules/classes for
  - Data Manipulation
    - Functionality to manipulate uploaded file
    - Used in UI and backend modules
  - Data Visualization
    - Returns plotly objects based on inputs from UI
  - Hypothesis Tests
    - Module for various hypothesis tests
    - Handler class to simplify module calls from UI

Demo

https://share.streamlit.io/hbaghar/statistics-for-dummies/main/src/app.py

Lessons Learned

Current Limitations:

- Streamlit
  - File upload limits
  - Scripts run from top to bottom on every input change (some workarounds are available)
  - Releases have some bugs
- Tool Design
  - Assumptions about input data shape in hypothesis tests

Future Work

- Revisit UI and incorporate UI tests
- Improve representation of results (add graphical explanations etc.)
- Support more types of statistical analysis (regression etc.)
- Reduce Streamlit's top to bottom re-running of code
- Support unpivoted datasets (time series data etc.)
- Add more error handling capabilities