

IDS706 Mini-Project 2 Report

Tianji Rao

Week 2 Pandas Descriptive Statistics Script

This repo contains:

- .devcontainer
- .github
- .gitignore
- Makefile
- README.md
- requirements.txt
- main.py
- test_main.py

Purpose

The purpose of this project is using the **Pandas** to show statistics descriptions. The author use a **pd.DataFrame** as a sample data and test its descriptions using the function **desc_df()**. The visualization focus on the bar plot, using **bar_plot()**. Both functions are tested in **test_main.py**.

Dataset

The experimental dataset is Electric Vehicle Population Data that provided by DATA.GOV. Here I downloaded the .csv file and made it the dataset for testing. The url address is <https://catalog.data.gov/dataset/electric-vehicle-population-data>. I used **pd.read_csv()** to read this dataset and save as a **pd.DataFrame**.

Functions

There are two main functions in the `main.py`: - `desc_df()`: this function can take a `DataFrame` as the input and return a statistical description summary based on the method `pd.DataFrame.describe()`. The default output of `describe()` can return a list of statistics including: `count`, `mean`, `std`, `min`, `25%`, `50%`, `75%`, and `max` (fullfill the requirements, which are mean, median, and standard deviation).

- `bar_plot()`: the `bar_plot()` function also take a `DataFrame` as the input and will plot of bar plot of the input data. This function is mainly based on the `pd.DataFrame.plot()`. Here, we set the `kind = bar` so we can get the desired bar plot.

Preparation

1. Setting up Codespaces
2. Check `make .` operations

Check format and test errors

1. Format `make format`
2. Lint `make lint`

```
(.venv) @TianoRao → /workspaces/TianjiRao_Pandas_Desc_Stat_Script (main) $ make lint
pylint --disable=R,C --ignore-patterns=test_.*?py *.py

-----
Your code has been rated at 10.00/10 (previous run: 0.00/10, +10.00)
```

Figure 1: make lint

3. Test `make test`

```
===== test session starts =====
platform linux -- Python 3.10.12, pytest-7.4.0, pluggy-1.3.0 -- /home/vscode/.venv/bin/python
cachedir: .pytest_cache
rootdir: /workspaces/TianjiRao_Pandas_Desc_Stat_Script
plugins: anyio-4.0.0, cov-4.0.0
collected 2 items

test_main.py::test_desc_df PASSED [ 50%]
test_main.py::test_bar_plot PASSED [100%]

----- coverage: platform linux, python 3.10.12-final-0 -----
Name      Stmts  Miss  Cover
-----
main.py         6     0   100%
TOTAL          6     0   100%

===== 2 passed in 1.84s =====
```

Figure 2: make test

There are three test cases for statistics: mean, median, and standard deviation.

Reference

<https://pandas.pydata.org/>

<https://catalog.data.gov/dataset/electric-vehicle-population-data>