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 Eletric 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`

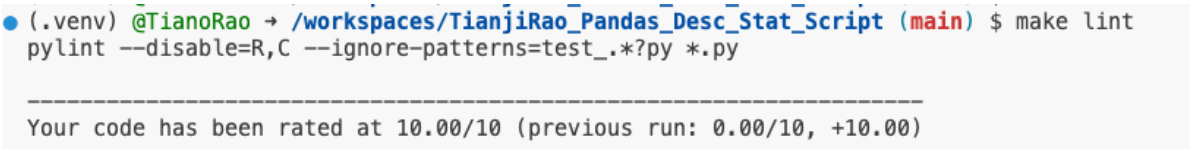

Figure 1: make lint

3. Test `make test`

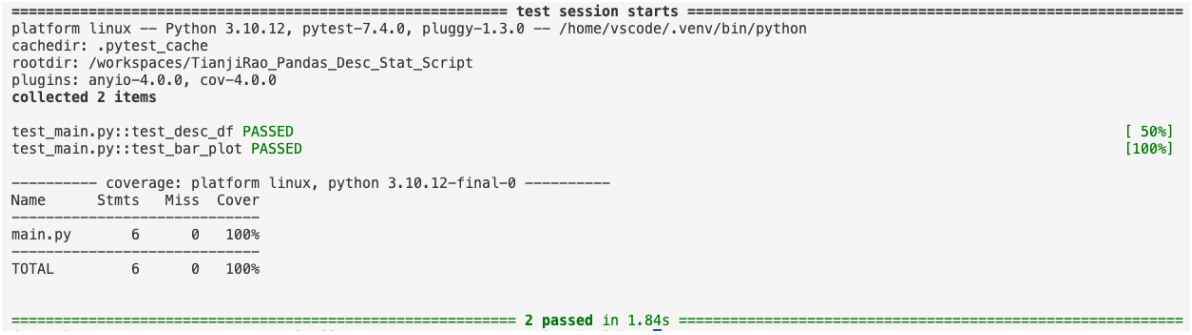

Figure 2: make test

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