Cappy Funding

Group Members:

Bryan Foo (foosuonchuang)
Gongzi Chen (gongzi)
Yujie Jiang (yujie0706)
Ziyang Chen (ziyangchen)

Table of Contents

Part 1: Project Requirements Compliance

- Two Data sources:
  - ✓ USA Spending API and USA Census
- Data analysis component
  - ✓ Three visualizations for the data analysis
- Specific components built by each member
  - ✓ Refer to the code responsibility part
- A visual or textual output
  - ✓ Refer to the following description of the three visualizations
- Each distinct component must be a subpackage
  - ✓ Refer to the relevant section
- Be able to run in a virtual environment
  - ✓ Refer to the code
- README.md
  - ✓ Refer to the Git repository

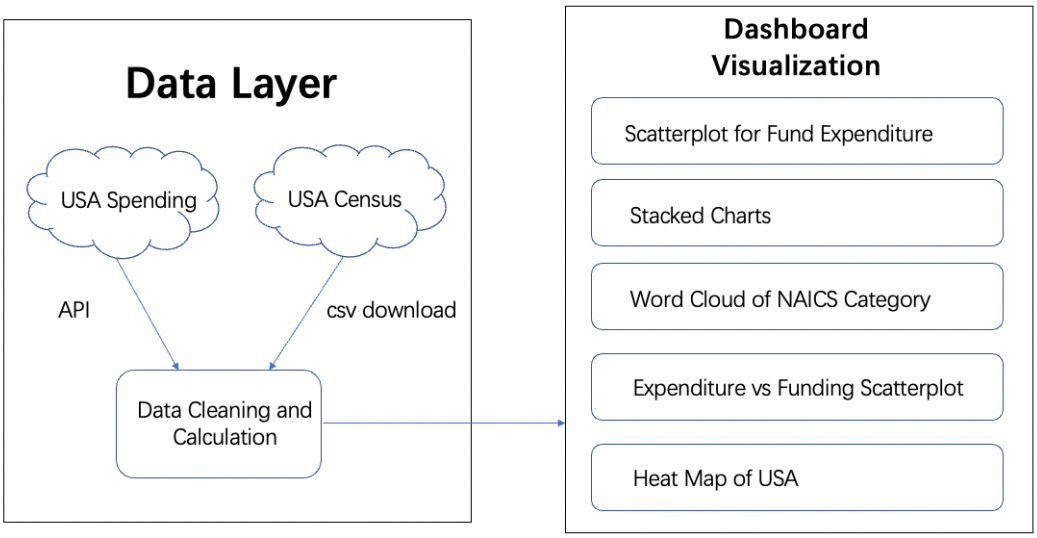
Part 2: Project Abstract

The main purpose of this project, Cappy Funding, is to visualize the allocation of federal fundings in the US to allow the audience to have a better sense of how funds at the federal level are spent on different industries and categories. Our expected target audience include federal foundation managers, funding seekers, and others who are interested in the way federal funds are spent. The complete data set contains a breakdown of federal funding by industry (using the NAICS categorisation) as well as by state.

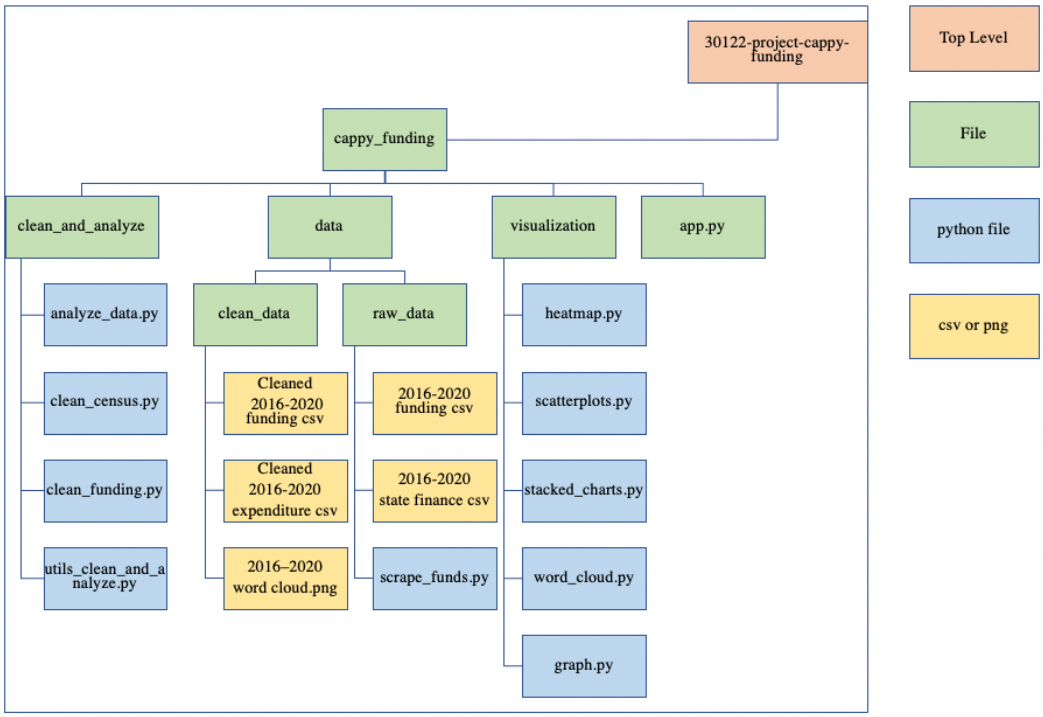
Our project performs data visualization and analysis for federal funding expenditure from 2016-2020 in the following five ways:

1. **Interactive Funding Heat Map of the US:** illustrates how funding grants are distributed geographically to different states for each NAICS sector.
2. **Interactive Scatter Plot on expenditure per capita of every state:** illustrates how state expenditure per capita is correlated with population size.
3. **Time-series Stacked Charts:** illustrates how federal funding for each sector changes over time.
4. **Interactive Scatter Plot for Expenditure per Capita and Funding per Capita:** illustrates how state expenditure per capita is correlated with funding per capita in a given year and how the relationship changes over time.
5. **Top 10 Categories Word Cloud:** illustrates the top ten NAICS industries that have received the greatest amount of federal funds over the last five years.

Part 3: Project Structure



Top-level directory structure



Part 4: Code Responsibility

Modules	Tasks	Name
Data Scraping	Used API to scrape data.	Yujie Jiang
Data Cleaning and Environment construction	(1) Cleaned and manipulated data to make it usable for data visualization. (2) Made the application workable in virtual environment.	Bryan Foo
Data Visualization -1	(1) Interactive Scatter Graph (2) Time-series Stacked Charts (3) Interactive Time-series Graph (4) Word Cloud	Ziyang Chen
Data Visualization -2	Interactive Heat Map	Gongzi Chen

Part 5: Interacting with the Software

1. Interacting with the application

To interact with the application, clone the git repository and run the command line (explained in the README.md file). Our application allows the user to do one of the following four actions:

1. Open the visualization dashboard: Selecting this option starts the process of data visualization using the pre-stored datasets. In the process, two static data visualizations (time-series stacked chart and word cloud graphics) are saved into the visualization/ directory. Following which the user is provided with a URL to use to access the data visualization dashboard on a browser of choice.
2. Run API and download files: Selecting this option starts the back-end process of API scraping from the USA Spending database. The user is provided with the option of selecting a range of years to download the raw files.
3. Data cleaning and analysis: Selecting this option starts the back-end process of using data already scraped from the API and cleans and processes the data sets. The user is provided with the option of selecting a range of years of raw datasets to clean. For the data visualization to work properly, the user should use the system default starting year (2016) and ending year (2020) by pressing the Enter button when prompted. Five main datasets are cleaned in the process for each year selected: (a) federal funding by state, (b) federal funding by category by state (percentages), (c) federal funding by category by state (absolute values), (d) state expenditure per capita, (e) federal funding per capita. In addition, two more datasets from the USA census are cleaned: (a) 2020 population census, and (b) 2018-2020 poverty rate by state.
4. Exit application: Selecting this option allows the user to exit the poetry virtual environment.

2. Interacting with API

To interact with the API, run the command line 'poetry run python3 -m cappy_funding' at the top-level directory and select option 2 for API interaction. It will ask for the input of a year range: a start year and an end year. This program will run a query on the USASpending API of the entered range. The input year should be in the string format (e.g. "2016", "2017"). Note that the default start year is 2016 and the default end year is 2017.
The output will be stored in csv-files at "/cappy_funding/data/raw_data". For each year, there will be a csv file containing the fundings of every category in every state.
The average run time for one year is around 10 minutes. As suggestion, a shorter year range query is recommended for API testing.

Part 6: Goals and Accomplishments of the Project

The project successfully delivered visualization on the distribution of federal fundings in the states of every category of industries (defined by NAICS) and also visualized some information on federal expenditure. However, there are three objectives we had to reevaluate.

The outcomes generated are not as significant as expected. Initially, we would like to see the correlation between federal expenditure and federal fundings and were expecting to see some relations between the two variables. However, after visualizing the correlation scatter plot, we found weak relation between the two. Same for the top-10 categories that receive most fundings: we were expecting to see some changes in the fundings of certain categories to change over time, however, it appears that from 2016 to 2020, there is no significant change.

Additionally, we were trying to do some analysis on recipients in every state to give more insights for funding seekers. However, the database containing recipient information has limited access, we can neither use API or web scraping, so we had to give up this idea.