PRACTICE #02

1. Introduction

This paper details the implementation of data visualization functions using Python and Matplotlib. Functions include reading files, visualizing data, plotting defaults, plotting median hist, plotting top5 median, plotting three bars, plotting scatterplots, grouping histograms, pie charts, displaying histograms, and saving histograms. The goal is to demonstrate how to use Matplotlib to visualize data for easy analysis.

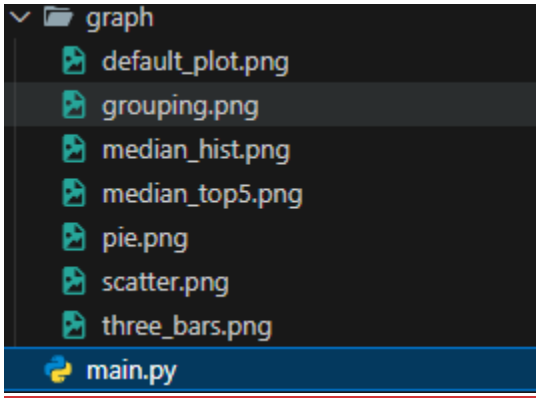
2. Implemented Functions

The following functions have been implemented:

- **read_file**: read file and load to data frame (using Pandas library).
- **visualize_data**: visualize the data (using Matplotlib).
- **plot_default**: create a default plot with some chosen info (Rank, P25th, Median, P75th).
- **plot_median_hist**: create a hist plot with median info (distributions and histograms).
- **plot_median_top5**: create a bar plot with median info (top 5).
- **plot_three_bars**: create a bar plot with median info (three bars per major).
- **plot_scatter**: create a scatter plot with median info.
- **plot_grouping**: create a plot with grouping.
- **plot_pie**: create a pie plot.
- **show_plot**: show the plot to screen.
- **save_plot**: save the plot figure to file.

3. Analytic Graphs

First, the structure of folders was organized as below:

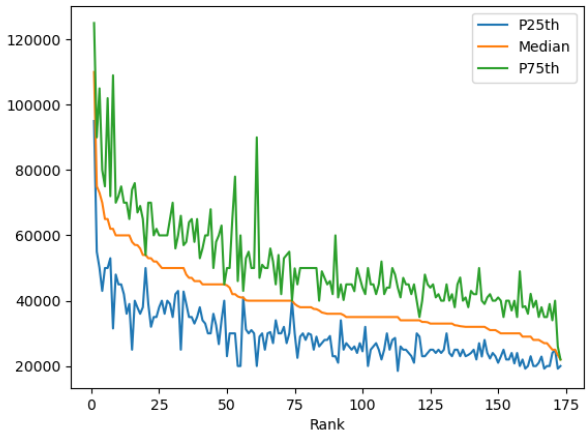


The .png files are the output after running main.py. Compiling main.py will display graphs according to the functions implemented inside the source code and these graphs visualize the data read from the file for analysis purposes and are then saved in the /graph directory.

3.1 plot_default function:

The **plot_default** is implemented by creating a default plot with some chosen info (Rank, P25th, Median, P75th):

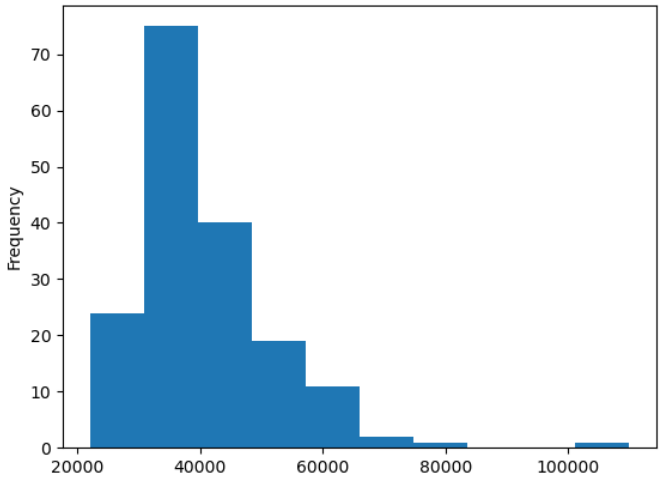
- "Median" is the median earnings of full-time, year-round workers
- "P25th" is the 25th percentile of earnings
- "P75th" is the 75th percentile of earnings
- "Rank" is the major's rank by median earnings



- This graph represents the distribution of salaries at the 25th, 50th (median), and 75th percentiles across different ranked groups.
- The decreasing trend suggests that higher-ranked groups tend to have higher salaries.
- The spread between the percentiles indicates salary variation within each rank.
- Fluctuations in the 75th percentile suggest some higher-earning outliers.

3.2 plot_median_hist function:

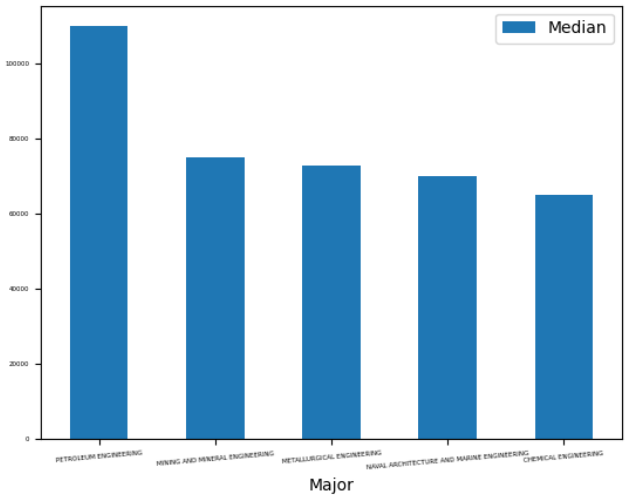
The **plot_median_hist** is implemented by creating a hist plot with median info (distributions and histograms):



- A histogram showing the frequency of median salaries.
- Most median salaries cluster between $30,000 and $50,000, with a sharp decline for higher salaries.
- A small number of majors have exceptionally high median salaries.

3.3 plot_median_top5 function:

The **plot_median_top5** is implemented by creating a bar plot with median info (top 5):

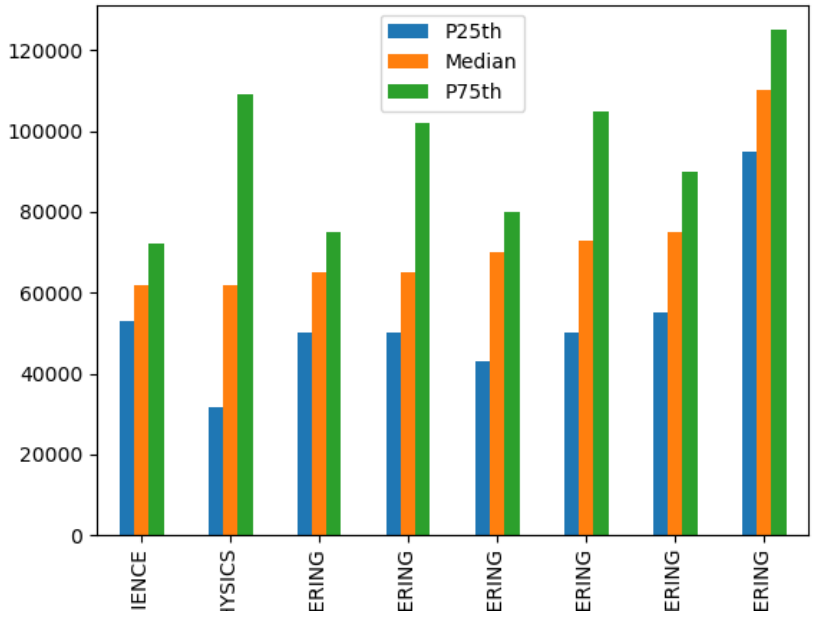


- A bar chart displaying the top 5 highest median salaries by major.
- Petroleum Engineering has the highest median salary, followed by Mining & Mineral Engineering.

- Engineering-related fields dominate the top 5, indicating high earning potential in these disciplines.

3.4 plot_three_bars function:

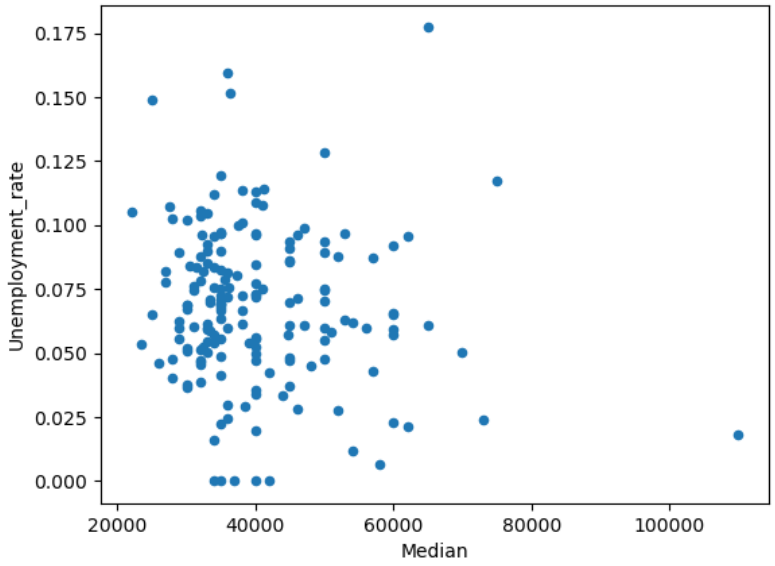
The **plot_three_bars** is implemented by creating a bar plot with median info (three bars per major).



- A bar chart comparing the 25th percentile, median, and 75th percentile salaries across different fields.
- The significant gap between percentiles suggests salary dispersion within fields.
- Some engineering and science majors show high variability between the 25th and 75th percentiles.

3.5 plot_scatter function:

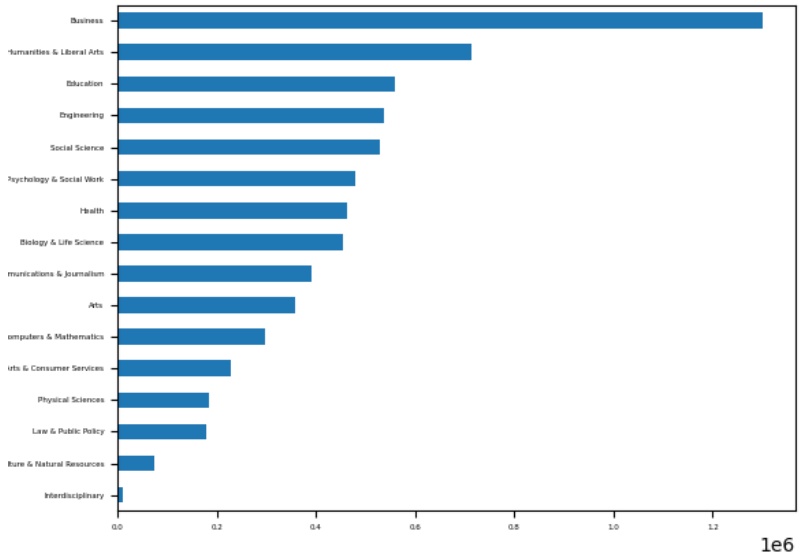
The **plot_scatter** is implemented by creating a scatter plot with median info:



- A scatter plot showing the relationship between median salaries and unemployment rates.
- A general trend suggests that lower median salaries tend to have higher unemployment rates.
- Some majors with high median salaries still exhibit a range of unemployment rates.

3.6 plot_grouping function:

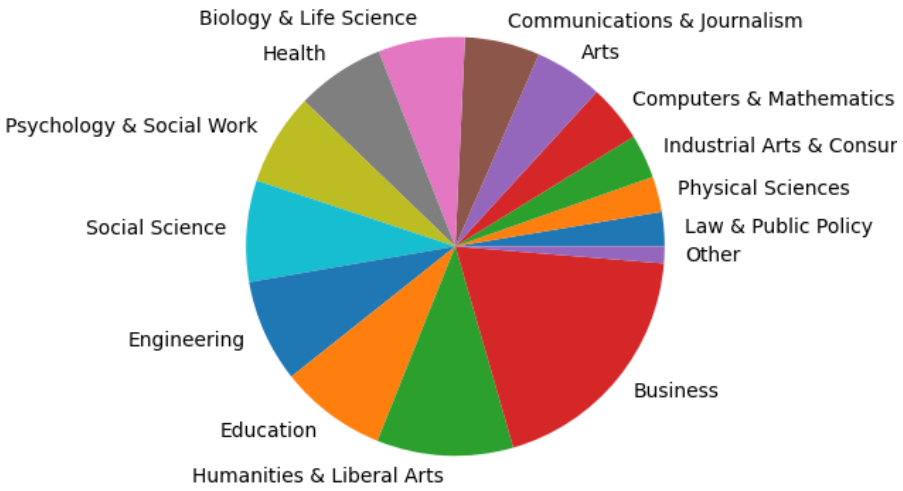
The **plot_scatter** is implemented by creating a plot with grouping.



- A horizontal bar chart showing the number of people in different fields.
- Business has the highest number of people, followed by Humanities & Liberal Arts.
- STEM-related fields like Engineering and Computers & Mathematics have lower numbers but might have higher median salaries.
- Fields such as Natural Resources and Interdisciplinary Studies have the lowest numbers.

3.7 plot_pie function:

The **plot_pie** is implemented by creating a pie plot.



- A pie chart displaying the distribution of different fields.
- Business, Humanities & Liberal Arts, and Education take up a large portion of the chart.
- STEM fields have a smaller share, which might correlate with higher salaries but fewer graduates.

All functions have been successfully implemented and tested for accuracy.