Assignment 5

GLASSDOOR DATA SCIENCE JOBS 2024

Elaine Candido da Silva (W0445885) | DBAS3018 | November, 2024

Contents

[Introduction 3](#_Toc182594252)

[Activity 1 ERD and Data Dictionary 4](#_Toc182594253)

[Item 1 – ERD Image 4](#_Toc182594254)

[Item 2 – Data Dictionary 5](#_Toc182594255)

[Activity 2 Data Wrangling 6](#_Toc182594256)

[Item 1 Exploring data wrangling dealing with missing values 6](#_Toc182594257)

[Part A - Dataset information BEFORE data wrangling 6](#_Toc182594258)

[Part B - Replace rows with string value from ‘NaN' to 'Unknown' for the fields below 7](#_Toc182594259)

[Part C - Replace rows with string value from '--' to 'Unknown' for the fields below 7](#_Toc182594260)

[Part D – Convert the “avg\_salary\_estimate to INT for later calculations 8](#_Toc182594261)

[Part E – Convert the “avg\_salary\_estimate to NumPy for later calculations 8](#_Toc182594262)

[Part F - Dataset information AFTER data wrangling 8](#_Toc182594263)

[Activity 3 Exploring NumPy and Pandas 9](#_Toc182594264)

[Item 1 – Function, Loop, Conditionals, Read from & Write to CSV files 9](#_Toc182594265)

[PART A –Read from CSV files 9](#_Toc182594266)

[PART B –Write to CSV files 9](#_Toc182594267)

[PART C –Conditionals 9](#_Toc182594268)

[Item 2 – Exploring 2+ plots (visualization) 10](#_Toc182594269)

[PART A - Visualization 1 –World Cloud 10](#_Toc182594270)

[PART B - Visualization 2 – HEATMAP 12](#_Toc182594271)

[Item 3 – Exploring descriptive statistics 13](#_Toc182594272)

[PART A – DESCRIBE() method 13](#_Toc182594273)

[PART B – CORR() method 13](#_Toc182594274)

[Item 4 – Exploring Numpy Array and 2+ related functions 13](#_Toc182594275)

[PART A – np.MEAN() method 14](#_Toc182594276)

[PART B – np.MEDIAN() method 14](#_Toc182594277)

[PART C – np.STD() method 14](#_Toc182594278)

[PART D – np.VAR() method 14](#_Toc182594279)

[Item 5 – Exploring Pandas DataFrame and 2+ related functions 15](#_Toc182594280)

[PART A – FILLNA() method 15](#_Toc182594281)

[PART B – REPLACE() method 15](#_Toc182594282)

[PART C – GROUPBY() method 15](#_Toc182594283)

[PART D – APPLY() method 15](#_Toc182594284)

[Activity 4 Rating Analysis 16](#_Toc182594285)

[Item 1 – Best ranked companies by sector => mean(company\_rating) column. 16](#_Toc182594286)

[Item 2 – Best ranked companies by size => mean(company\_rating) column. 17](#_Toc182594287)

[Item 3 – Best ranked companies by revenue => mean(company\_rating) column. 17](#_Toc182594288)

[Item 4 – Best ranked companies by employment type => mean(company\_rating) column. 18](#_Toc182594289)

[Activity 5 Company Analysis 18](#_Toc182594290)

[Item 1 – Job vacancies by Profession (counter >= 10). 18](#_Toc182594291)

[Item 2 –Job vacancies by Location (counter >= 10). 19](#_Toc182594292)

[Item 3 – Job vacancies by sector (counter >= 10). 20](#_Toc182594293)

[Item 4 – Job vacancies by company size. 20](#_Toc182594294)

[Item 5 – Job vacancies by company revenue. 21](#_Toc182594295)

[Item 6 – Job vacancies by employment type. 21](#_Toc182594296)

[Activity 6 References 22](#_Toc182594297)

[Figure 1 5](#_Toc182594298)

[Figure 2 6](#_Toc182594299)

[Figure 3 7](#_Toc182594300)

[Figure 4 8](#_Toc182594301)

[Figure 5 8](#_Toc182594302)

[Figure 6 9](#_Toc182594303)

[Figure 7 9](#_Toc182594304)

[Figure 8 9](#_Toc182594305)

[Figure 9 10](#_Toc182594306)

[Figure 10 10](#_Toc182594307)

[Figure 11 10](#_Toc182594308)

[Figure 12 11](#_Toc182594309)

[Figure 13 12](#_Toc182594310)

[Figure 14 13](#_Toc182594311)

[Figure 15 13](#_Toc182594312)

[Figure 16 14](#_Toc182594313)

[Figure 17 14](#_Toc182594314)

[Figure 18 14](#_Toc182594315)

[Figure 19 15](#_Toc182594316)

[Figure 20 15](#_Toc182594317)

[Figure 21 15](#_Toc182594318)

[Figure 22 15](#_Toc182594319)

[Figure 23 16](#_Toc182594320)

[Figure 24 16](#_Toc182594321)

[Figure 25 16](#_Toc182594322)

[Figure 26 16](#_Toc182594323)

[Figure 27 17](#_Toc182594324)

[Figure 28 18](#_Toc182594325)

[Figure 29 18](#_Toc182594326)

[Figure 30 19](#_Toc182594327)

[Figure 31 19](#_Toc182594328)

[Figure 32 20](#_Toc182594329)

[Figure 33 21](#_Toc182594330)

[Figure 34 21](#_Toc182594331)

[Figure 35 22](#_Toc182594332)

[Figure 36 22](#_Toc182594333)

# Introduction

This project aims to explore the features of the Python programming language with an emphasis on Object-Oriented programming, alongside the Numpy and Pandas libraries. The analysis is conducted using the "Glassdoor Data Science Jobs - 2024" dataset (available on Kaggle), a comprehensive collection of job listings in data science sourced from Glassdoor. The dataset provides rich insights into job market trends, skill requirements, and employer characteristics. The analysis covers the following key areas:

* Python: functions, loops, and conditionals
* Pandas: reading and writing CSV files, handling and manipulating DataFrames and Series, and associated functions
* NumPy: handling and manipulating arrays, and relevant functions
* Data Wrangling, including managing missing data
* Visualization: creating and interpreting various plots

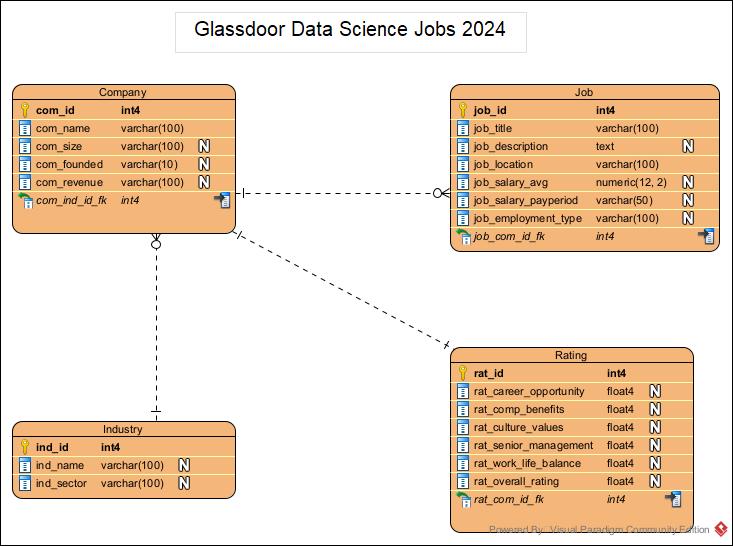
In addition, five data-related questions are analyzed, and answers are presented based on the findings.

Dataset overview

The "Glassdoor Job Postings Dataset" includes job titles, company details, job descriptions, qualifications, location, salary ranges, and rating information. It facilitates research in HR, labour market trends, and career development. The data is sourced from Glassdoor and anonymized for privacy, providing a valuable resource for job market analysis, salary trends, skills demand, and employer insights.

# Activity 1 ERD and Data Dictionary

## Item 1 – ERD Image



Figure

## Item 2 – Data Dictionary

A screenshot of a computer

Description automatically generated

Figure

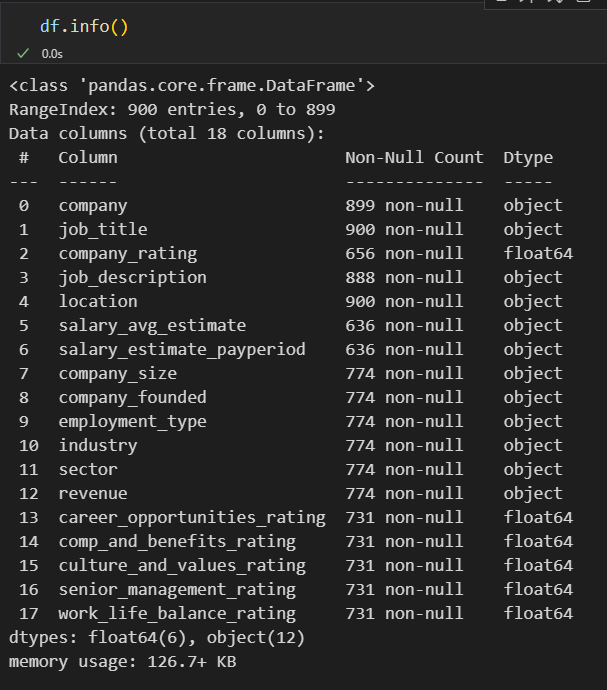
# Activity 2 Data Wrangling

As the dataset came with missing data, some transformation was needed to adjust the data before analyzing.

## Item 1 Exploring data wrangling dealing with missing values

## Part A - Dataset information BEFORE data wrangling

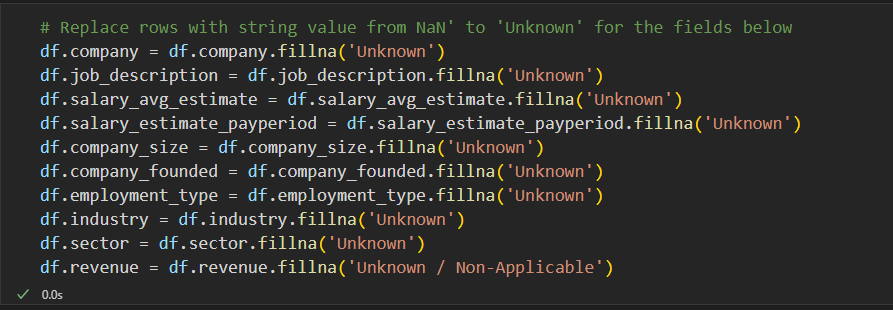
The image below shows the non-null values for each field.



Figure

## Part B - Replace rows with string value from ‘NaN' to 'Unknown' for the fields below

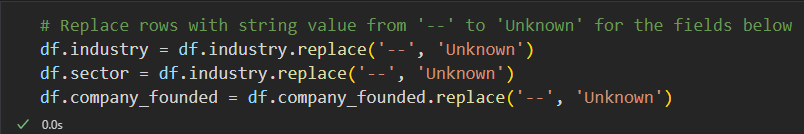
Replaced rows with string value from ‘NaN’ to “Unknown” to add a categorical value for calculations purposes.



Figure

## Part C - Replace rows with string value from '--' to 'Unknown' for the fields below

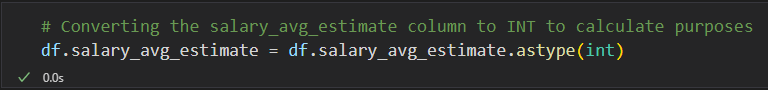
Replaced rows with string value from '--' to 'Unknown' to add a categorical value for calculations purposes.



Figure

## Part D – Convert the “avg\_salary\_estimate to INT for later calculations

Transformed this column to calculate the descriptive statistics on Activity 3.



Figure

## Part E – Convert the “avg\_salary\_estimate to NumPy for later calculations

Transformed this column to calculate the descriptive statistics on Activity 3.

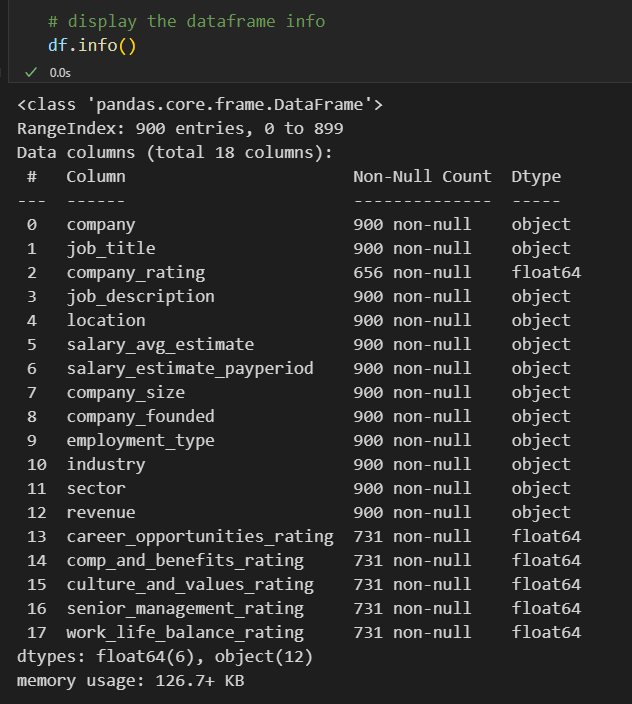
A black screen with white text

Description automatically generated

Figure

## Part F - Dataset information AFTER data wrangling

After the replacement, the data information became as shown in the image below. Just the float data remains with the same amount of ‘NaN’ values.

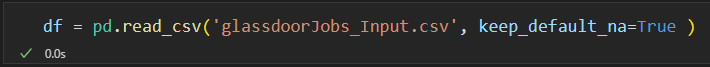


Figure

# Activity 3 Exploring NumPy and Pandas

## Item 1 – Function, Loop, Conditionals, Read from & Write to CSV files

## PART A –Read from CSV files



Figure

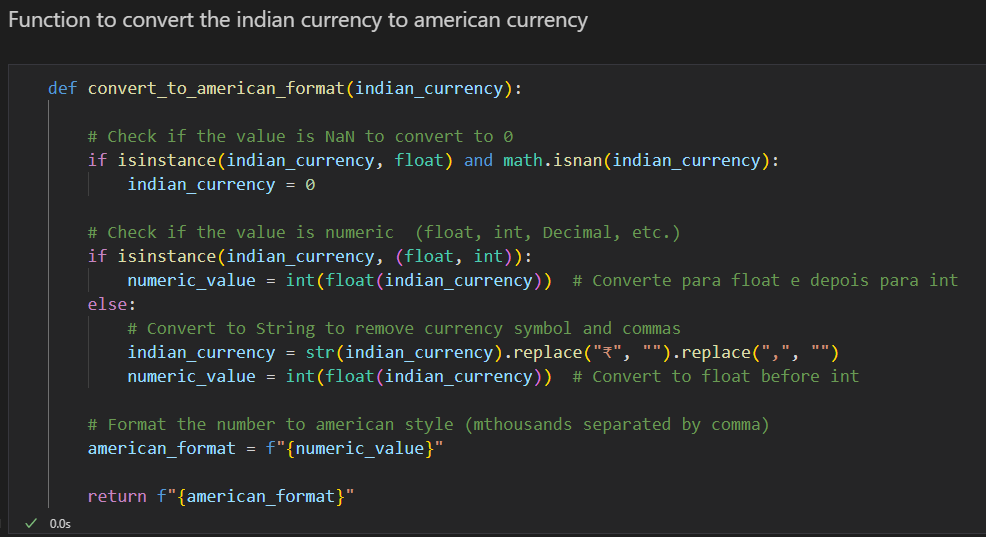
## PART B –Write to CSV files

A black screen with text

Description automatically generated

Figure

## PART C –Conditionals



Figure

## Item 2 – Exploring 2+ plots (visualization)

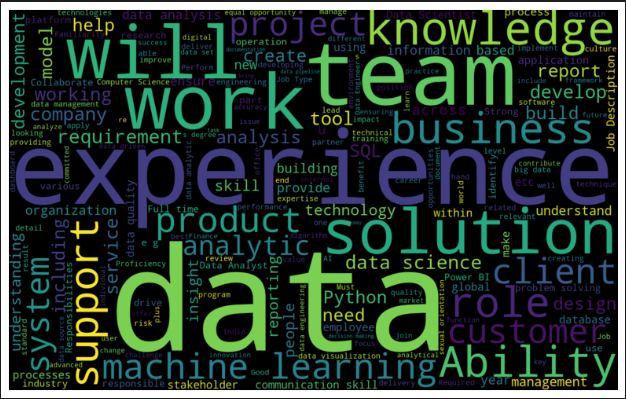
## PART A - Visualization 1 –World Cloud

In this visualization a word cloud was used to show the most used words on the job\_description column. The steps are on the image below.

A screenshot of a computer program

Description automatically generated

Figure



Figure

## PART B - Visualization 2 – HEATMAP

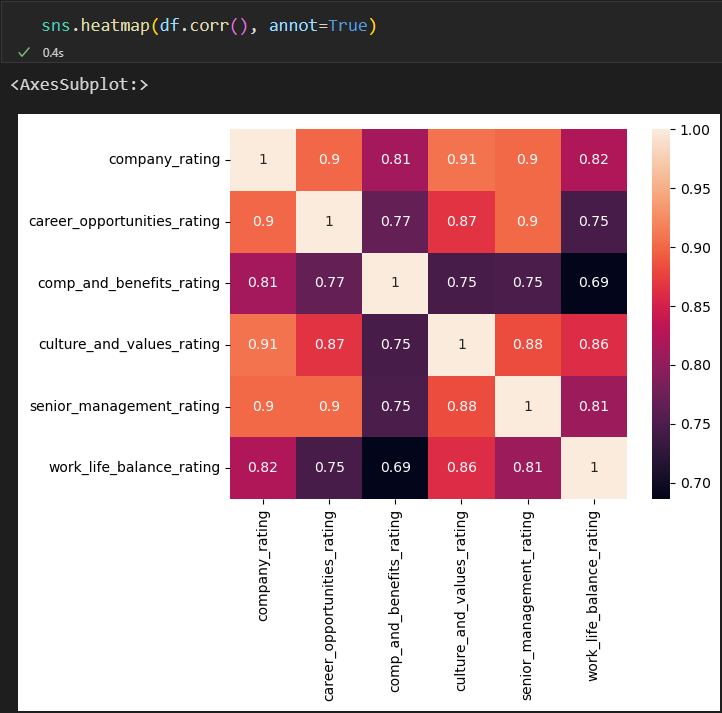
In this visualization I used the heatmap to show the correlation between the rating columns.

sns.heatmap(df.corr(), annot=True)

A black screen with white text

Description automatically generated

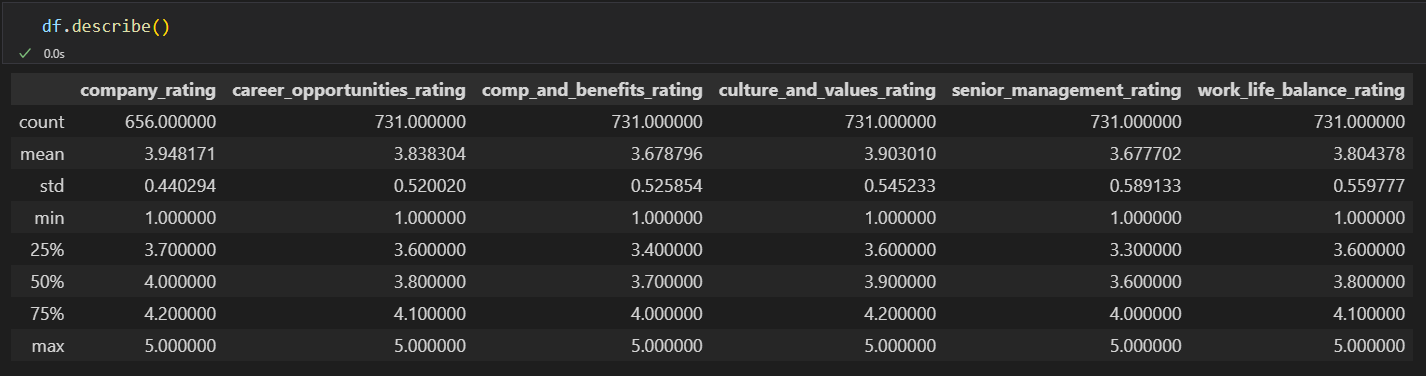
Figure



Figure

## Item 3 – Exploring descriptive statistics

## PART A – DESCRIBE() method



Figure

## PART B – CORR() method

A screenshot of a computer

Description automatically generated

Figure

## Item 4 – Exploring Numpy Array and 2+ related functions

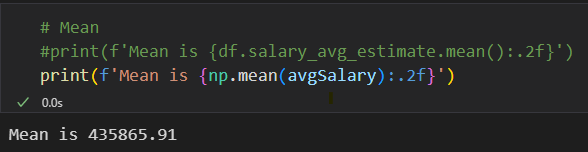
Initially I calculated descriptive statistics using Pandas. However, to explore NumPy, I converted the “df.salary\_avg\_estimate” to NumPy array to use it to calculate the mean, median

A black screen with white text

Description automatically generated

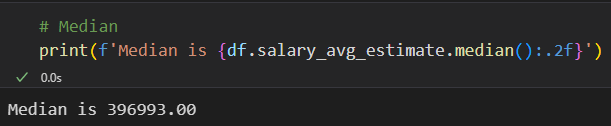
Figure

## PART A – np.MEAN() method



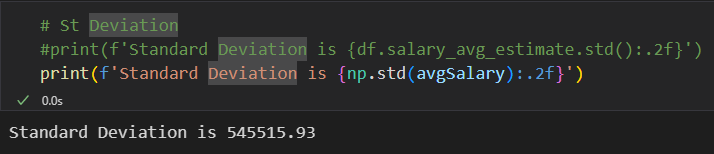
Figure

## PART B – np.MEDIAN() method



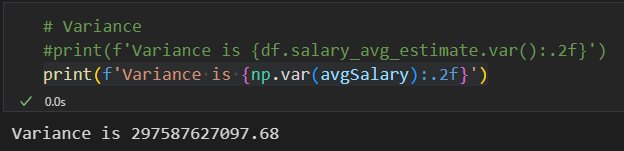
Figure

## PART C – np.STD() method



Figure

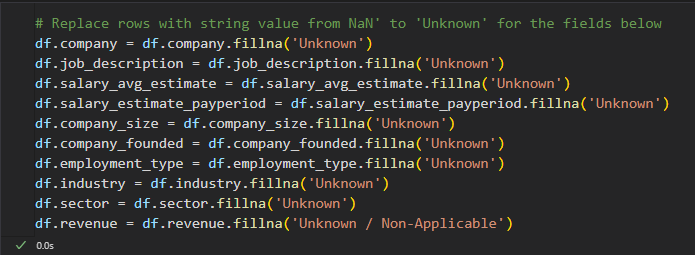
## PART D – np.VAR() method



Figure

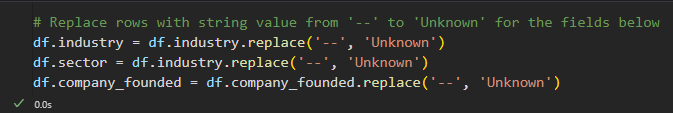
## Item 5 – Exploring Pandas DataFrame and 2+ related functions

## PART A – FILLNA() method



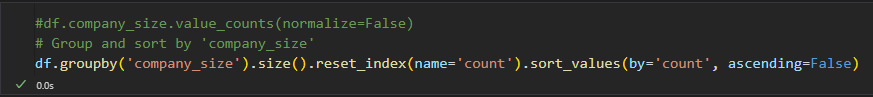
Figure

## PART B – REPLACE() method



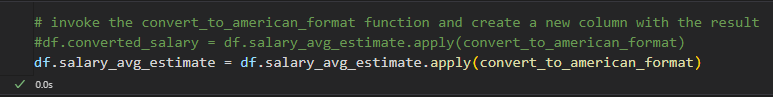
Figure

## PART C – GROUPBY() method



Figure

## PART D – APPLY() method



Figure

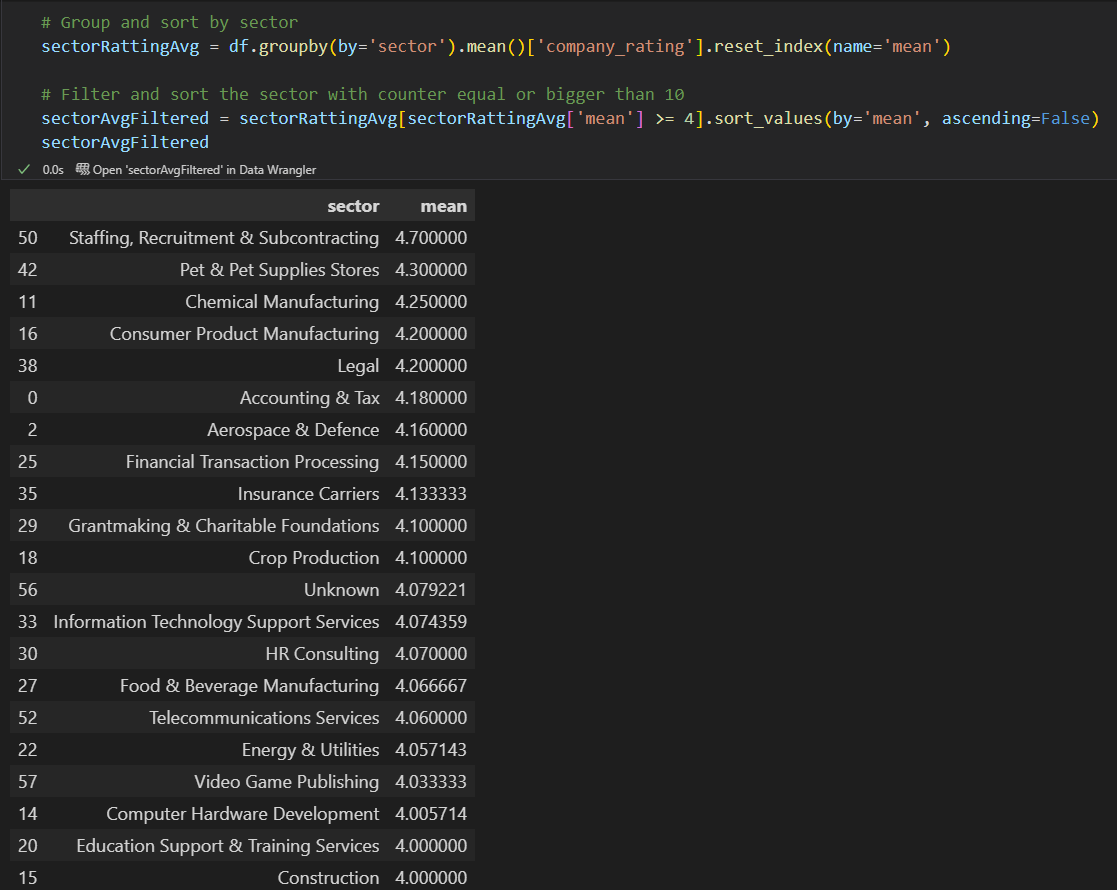
# Activity 4 Rating Analysis

After exploring the dataset, it is important to note:

* A significant amount of missing information, evident in the 'Unknown' category.
* A substantial discrepancy in ”salary\_avg\_estimate” values, resulting in numerous outliers that complicate the evaluation process.
* To analyze the rating values, I grouped them by calculating the mean of the “company\_rating” column

## Item 1 – Best ranked companies by sector => mean(company\_rating) column.

This grouping highlights the top-ranked sectors. To streamline the data displayed, a condition was applied to filter for a mean score of 4 or higher. Interestingly, the Recruitment sector emerges as the highest-ranked sector. This may be due to the sector's focus on the human aspect of labor relations and its concern for the overall well-being of employees.



Figure

## Item 2 – Best ranked companies by size => mean(company\_rating) column.

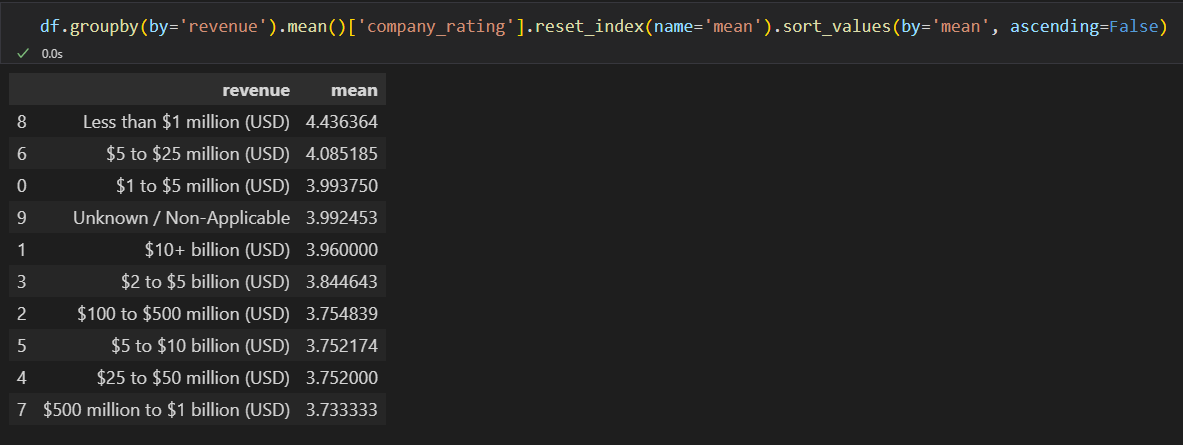
This grouping demonstrates that small companies are typically well-rated by their employees. Specifically, companies with up to 500 employees are highly ranked. This may be due to the closer proximity between employees and leadership, as well as the more horizontal management structure typically found in smaller organizations.



Figure

## Item 3 – Best ranked companies by revenue => mean(company\_rating) column.

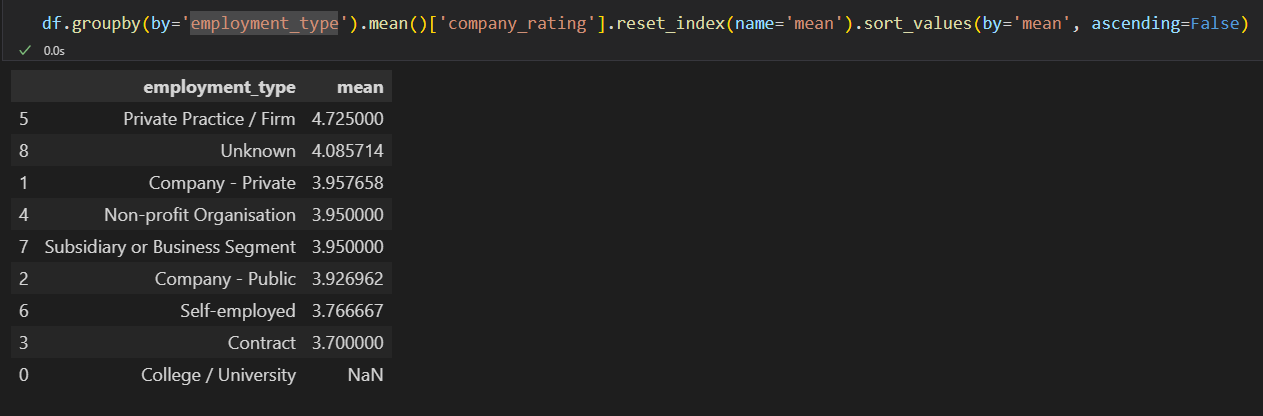
This grouping once again illustrates that small companies are typically well-rated by their employees. Specifically, companies with revenues below 1 million are highly ranked.



Figure

## Item 4 – Best ranked companies by employment type => mean(company\_rating) column.

This grouping shows that the 'Private Practice / Firm' employment type is the highest rated.

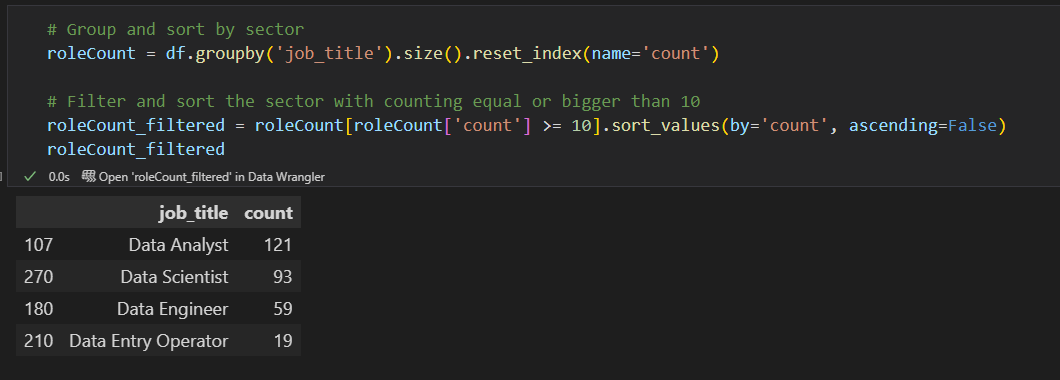


Figure

# Activity 5 Company Analysis

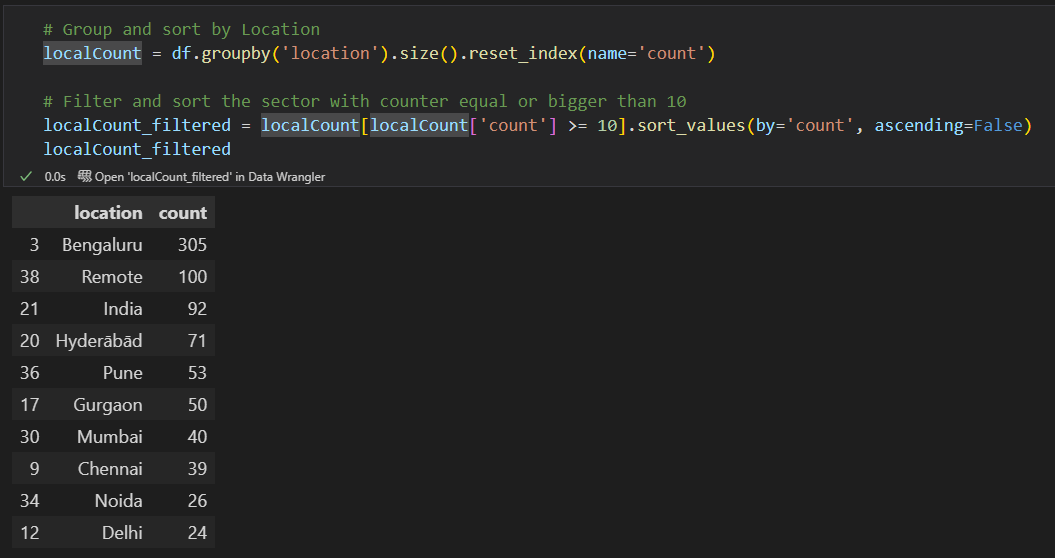
## Item 1 – Job vacancies by Profession (counter >= 10).

This grouping highlights the top-ranked roles: Data Analyst, Data Scientist, Data Engineer, and Data Entry Operators. To streamline the data displayed, a condition was applied to filter for a count of 10 or higher. This may be due to the increasing data generation driven by Big Data and the growing value that companies are placing on data.



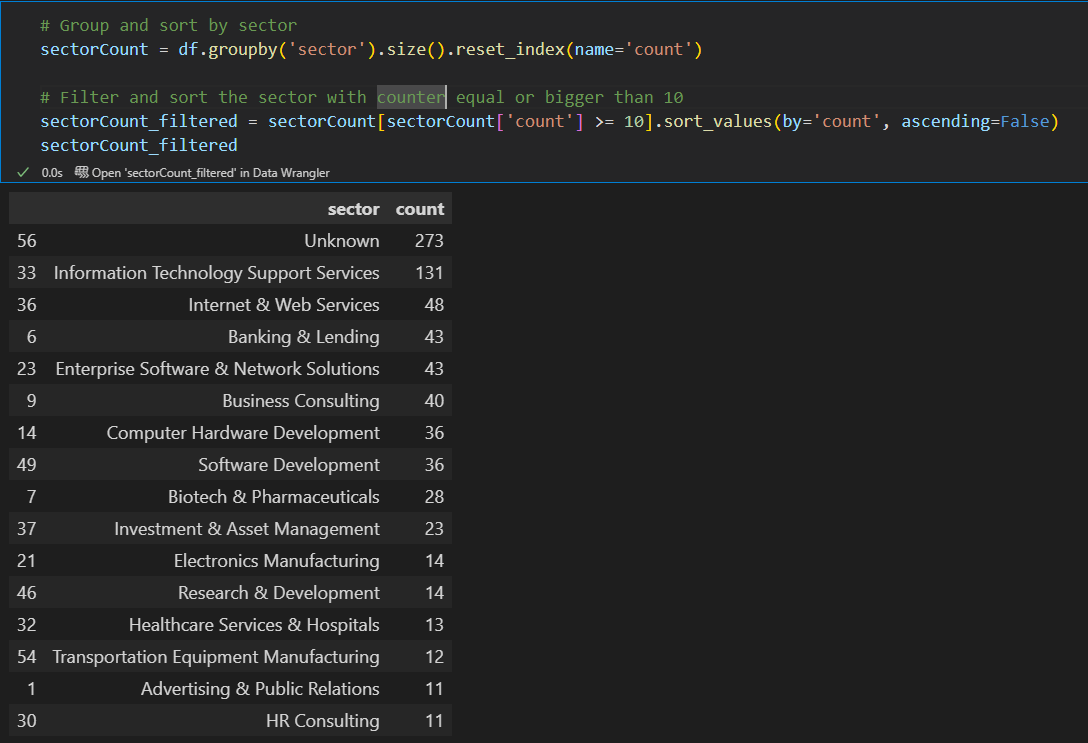
Figure

## Item 2 –Job vacancies by Location (counter >= 10).



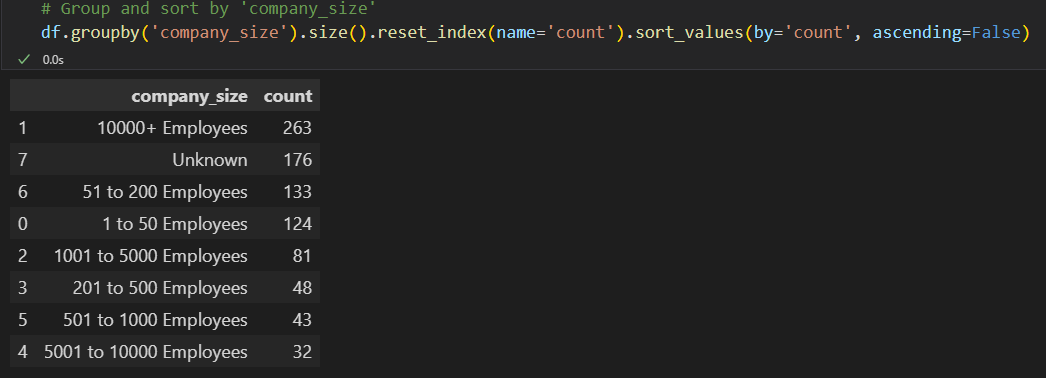
Figure

## Item 3 – Job vacancies by sector (counter >= 10).



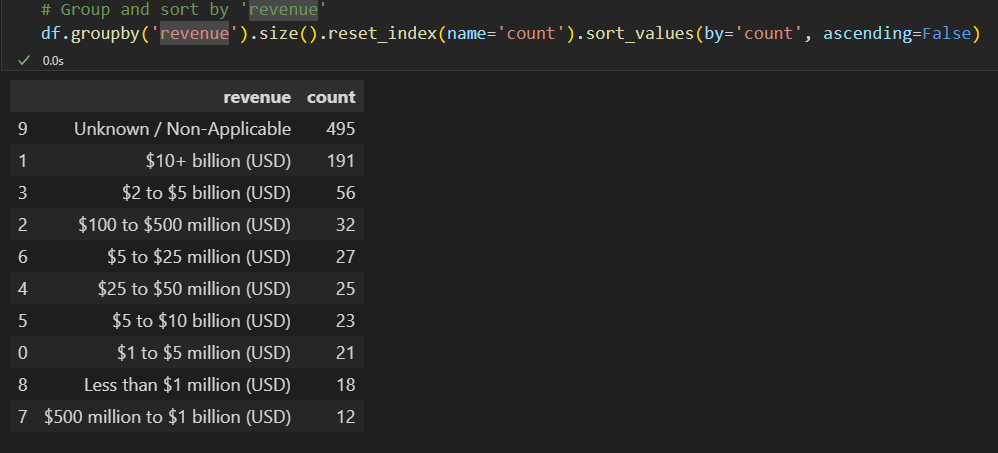
Figure

## Item 4 – Job vacancies by company size.



Figure

## Item 5 – Job vacancies by company revenue.



Figure

## Item 6 – Job vacancies by employment type.



Figure

# Activity 6 References

*ChatGPT*. (n.d.). https://chatgpt.com/c/6737ae89-ab84-800e-8747-f0b6b1d9f4de

DataFrame — pandas 2.2.3 documentation. Accessed November 15, 2024. https://pandas.pydata.org/docs/reference/frame.html

Array objects — NumPy v2.1 Manual. Accessed November 15, 2024. https://numpy.org/doc/stable/reference/arrays.html

Python Developer’s Guide. Python Developer’s Guide. Accessed November 15, 2024. https://devguide.python.org/