**About Dataset**

Link to repository: <https://github.com/csmfduro/salaries>

The salaries are from ai-jobs. Ai-jobs collects salary information anonymously from professionals all over the world in the AI/ML and Big Data space and makes it publicly available for anyone to use, share and play around with. The data is being updated regularly with new data coming in, usually on a weekly basis.

The primary goal is to have data that can provide better guidance regarding what's being paid globally. So, newbies, experienced pros, hiring managers, recruiters and startup founders or people wanting to make a career switch can make better informed decisions.

The dataset contains one table structured as follow:

1. **work\_year**: The year the salary was paid.

2. **experience\_level**: The experience level in the job during the year with the following possible values:

EN: Entry-level / Junior

MI: Mid-level / Intermediate

SE: Senior-level / Expert

EX: Executive-level / Director

3. **employment\_type**: The type of employment for the role:

PT: Part-time

FT: Full-time

CT: Contract

FL: Freelance

4. **job\_title**: The role worked in during the year.

5. **salary**: The total gross salary amount paid.

6. **salary\_currency**: The currency of the salary paid as an ISO 4217 currency code.

7. **salary\_in\_usd**: The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).

8. **employee\_residence**: Employee's primary country of residence in during the work year as an ISO 3166 country code.

9. **remote\_ratio**: The overall amount of work done remotely, possible values are as follows:

0: No remote work (less than 20%)

50: Partially remote

100: Fully remote (more than 80%)

10. **company\_location**: The country of the employer's main office or contracting branch as an ISO 3166 country code.

11. **company\_size**: The average number of people that worked for the company during the year:

S: less than 50 employees (small)

M: 50 to 250 employees (medium)

L: more than 250 employees (large)

The reason we chose this dataset is that it had the most relevance with all the members of the group being about AI job salaries. When compared to our other options being cars and cancer probabilities it also was the most appropriate and containing the most data to be able to extract assumptions from.

**Key Questions to Explore**

The below questions are the info each member of the group most wanted to find out involving the data included in the dataset.

**Fiyin**

1. Which countries offer the highest salaries for AI professionals?

* Group salaries by country and find the top countries with the highest average salaries.

1. What are the highest-paying AI job titles?

* Group salaries by job title and identify the highest-paying roles.

**Prajesh**

1. How much would an Entry Level/Junior of a high paying job make compared to a low paying job but at senior level or Executive level?
2. Do freelancers make more money or less money compared to employees who work fulltime and if so what job/ task allows a freelancer to get a good pay and if less what jobs are freelancers taking up?

**Jason**

1. what is the difference in average salary between companies with 100% remote work and 0% remote work
2. what is the difference between the average salary in lockdown (2020) compared to this year (2025)

**Jake**

1. What are the job titles for the top and bottom 10 for gross salary?
2. What is the main trend between experience level and salary?

**Eisha**

1. Do gender or diversity factors influence salaries in AI roles?
2. Are AI professionals in smaller companies paid less than those in large  
   enterprises?

**Faris**

1. How do salaries differ between employees with and without a degree?
2. Is there a correlation between years of experience and salary?

**Andrei**

1. What is the average salary difference between the small, medium and large companies?
2. What location of companies has the highest salary?

**Algorithms used**

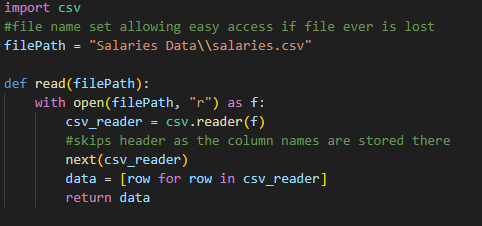
**Application Design**

Our application’s functionality will include a main menu as we will not display the results for every question at once allowing the user to choose which questions they would like the data from. The menu options included are the ability to view the dataset, view the column names, view the dataset info and choose the questions to view the info from. The program will interact with the user through the user inputting the menu option they want to view. The data will be processed through csv and pandas to extract the dataset from the csv file.

**Implementation**

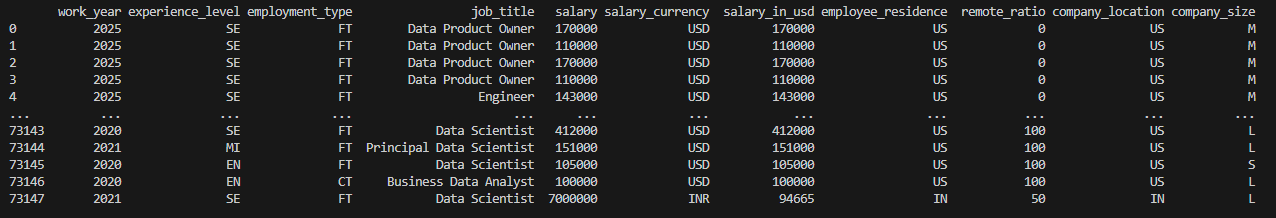
**File reading function – Jason**

To begin with I created a reusable function to read the csv file and extract the dataset as this would be used for all the questions and main menu functions. This went through a few iterations as originally it used csv but I found using pandas was more streamlines and easy to understand



This is the original function to be able to read the dataset file which uses a variable “filepath” that can be easily changed at the top of the code in case the file ever gets lost or renamed. Using this the code opens the file and inputs the contents into a list and returns it to a variable. This ended up looking messy when viewed so I switched the code to use the pandas module instead.



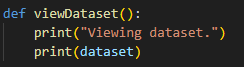
This is the final file reading function using the pandas module as the data outputted is automatically formatted making it easy to read as shown below.

**Main menu (structure) – Andrei & Prajesh**

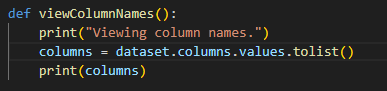
**Main menu (function) – Jason**

Once the menu structure was finished including the shells of each function that would be used as menu options, I added the code to make them work. The options that were given to me are: load dataset, view dataset, view column names, view dataset info and choose questions of which I added the code for the first 4.

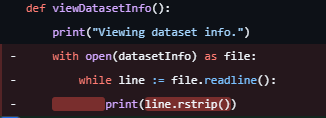
Starting with the load dataset function I thought that it made more sense to load it upon running the application so this option has been removed in favour for doing it upon program start.



The code to view the dataset is very simple as the bulk of the work was already done when creating the function to read the file as this function just prints out the info that was extracted from the file. This code went through no changes



The view column names function takes the dataset extracted using the file reading function and displays the top row of values being the column names. This code went through no changes.



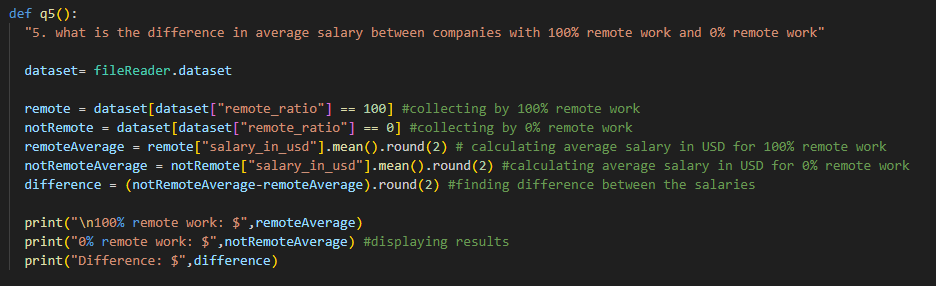
The view dataset function went through a couple changes as this is the original. It read the file every time the user called the function which isn’t ideal, this was later changed by another team member.

**Questions – Jason**

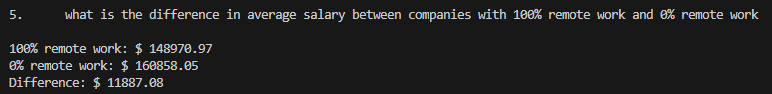
My questions that I had to represent were

5. what is the difference in average salary between companies with 100% remote work and 0% remote work

Both questions required similar code to calculate so there will be clear similarities when showing both pieces of code used for each.



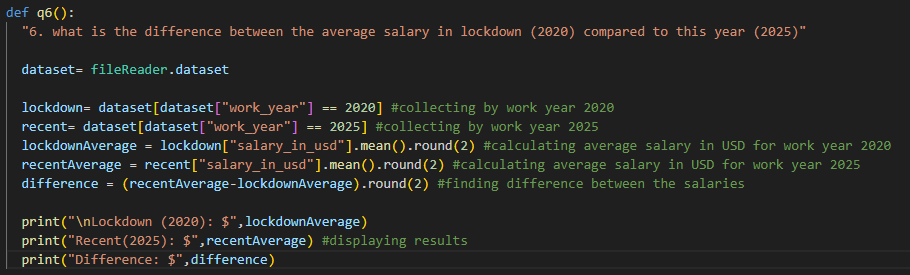
The code for the first question being the difference between full remote work and no remote work first retrieves the file from the dataset then extracts both the positions with 100% remote work and 0% remote work and calculates the mean of each respective salary. Finally, the code calculates the difference between the 2 averages before printing the results out in a clear way.



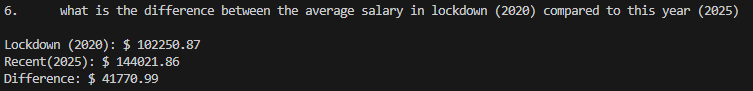
The result of the question is that on average jobs with no remote work pay slightly more than jobs that have 100% remote work.

The difference is very minimal.

6. what is the difference between the average salary in lockdown (2020) compared to this year (2025)



This code for this question works in a very similar manner to the other question substituting the remote work percentage for the work year comparing specifically 2020 to 2025.



The result shows that jobs in the current year pay more than jobs in lockdown (2020) although this can be contributed to many factors such as inflation and the inability for people to leave their houses during that year.

The difference is significant

**Questions - Name here**

**Questions - Name here**

**Questions - Name here**

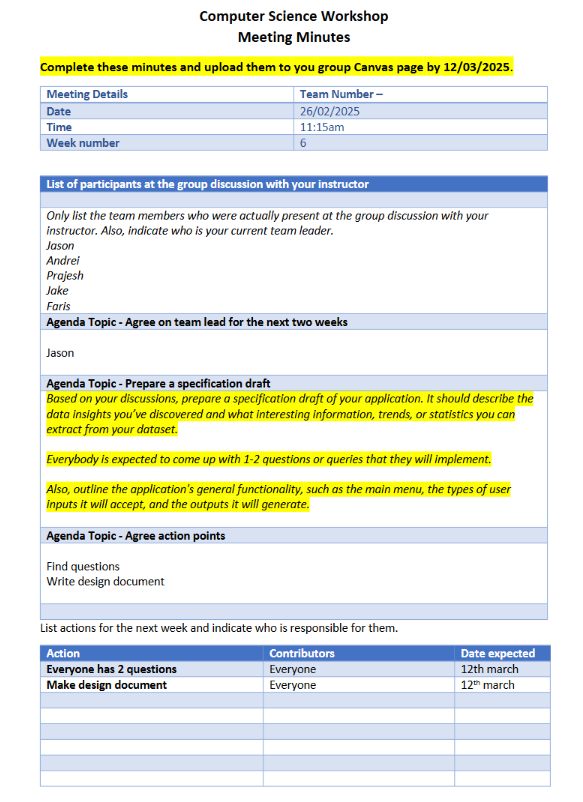
**Questions - Name here**

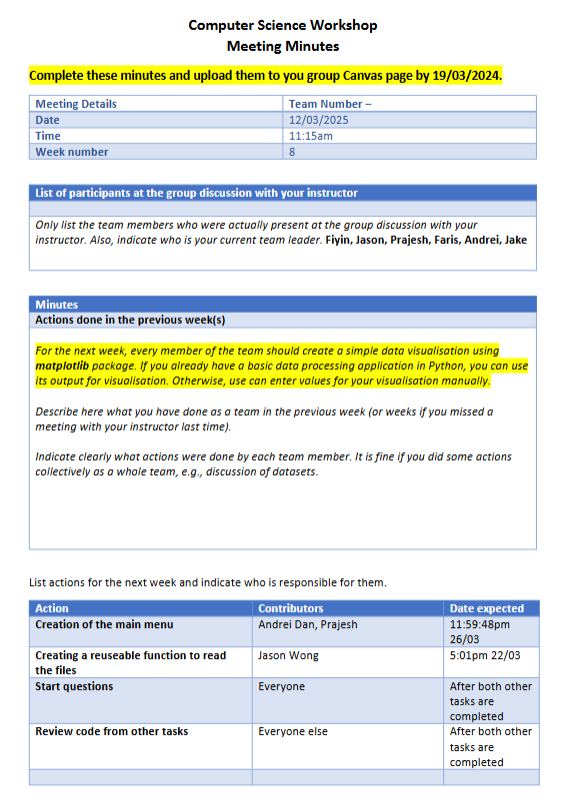
**Questions - Name here**

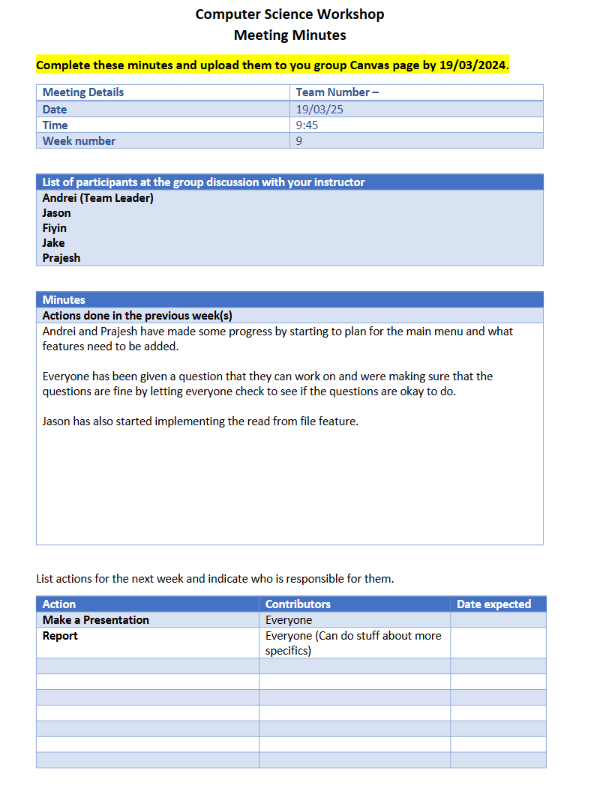
**Questions - Name here**

**Testing**

**Meeting Minutes**







**Group Evaluation**