Problem 1

I have written a short bash script problem_1/renameFiles.sh that will rename the files.

To explain to the researcher how to do this, I would first check how familiar they are with the use of terminal in a linux environment. Assuming this is fine (given the context in the introduction to the problems!) I would then explain to them the 'mv' command that allows them to rename files in this environment. Based on this knowledge, I can then explain to them the contents of the bash script, which applies that command to each of the files in the directory using a for loop (further explanation can be given here if necessary).

It is also possible to do the mv command all in the command line with a simple bash for loop written in one line. For example;

```
for i in {1..100}; do mv datafile"\$i".csv datafile"\$i".dat; done
```

could be used instead, if this feels easier for the researcher.

Problem 2

One very quick method for identifying problematic files in this scenario is to use the command;

```
ls -l | sort -r
```

This lists every file in the directory in order of size from largest to smallest. Since the files only contain a small amount of data, it is reasonable to assume that the smallest files are the ones missing data points. Using this method we see that values44 is the smallest output file, and therefore warrants further inspection. Indeed, when investigating this file we find that entry 37 in the expected value array is empty.

Of course, this method is hardly scalable when the number of samples grows larger, and is not perfect - values58 is only 2 bytes larger but contains all 100 entries correctly. So I have written a short Python script problem_2/checkOutputValues.py that reads each 'value' file in the directory in which it is run, and checks that there is a valid entry for each of the expected data values.

The script notifies the user if a file contains less than 100 entries in its tab separated list, and reports which entry is missing if there found to be a non-valid entry in the list.

To explain how this works to the researcher, I would highlight the checkOutputs method, which identifies missing data points in a single file. From there we can extrapolate to the section of code that checks each file in the directory.

Problem 3

I would suggest to the researcher the use of a Python profiler to establish where the bottlenecks in their code and any possible slow performance is coming from. Using the -m cProfile command line argument, for example, can give an idea of how long the method is taking to run. For this example only one method is called and this is, therefore, less immediately helpful, but it can still be useful to see how long things take to run when making adjustments to the code. She could, for example, remove parts of the code to see how much performance is improved.

Similarly, I would advise using the timeit command to compare how efficienctly different versions of code snippets run. This is an easy method to determine exactly what is better for running, for example, the long loop commands that will be running when the prime function is with a larger value of n.

Without pointing to anything specific in the the code, I would also highlight things that can sometimes be potential bottlenecks in Python code and advise her to start her investigations there. The use of long lists, for example, can be slow in Python, and nested loops are often quite innefficient. List comprehensions or numpy arrays can provide much faster solutions to while loops in particular, and could be considered here to improve the code's performance.

Problem 4

In order to carry out the Requirement Analysis, I would determine to have a meeting with the researcher in question, along with any other team members working on the project if applicable. I would consider the following questions important to beginning the development, and each question could spark further discussion on the scope and details of the study.

- **Who is the group of individuals we are collecting data from?**

  In particular, are they a pre-determined list of people or will they be selected based on certain criteria? This is an important question as it defines exactly how we go about data capture, and the sort of implementation that will be required of us, which would change dramatically depending on the answer.

  I would also posit that this could potentially raise ethical and GDPR concerns, but I suppose that this is the job of the ethics review committee rather than the RSEs.

- **What is the time window that counts as the run-up to the Indian general election?**

  As I am less familiar with Indian political cycles than, say, the UK or US, I would need clarifying information about the events around the election.

This would likely start off some more background questions on the election cycle itself, to familiarise myself with the material.

- **What is the hypothesis that the research is working to, and/or what is it that the researcher is hoping the data will show?**

  It is important that we know what the goal of the research is in order to capture comments and posts that are relevant and useful to the study. Potentially this could be multiple answers, in which case it would be useful to have a ranked hierarchy so we can prioritise the most relevant/useful information.

- **What tools is the researcher planning to use in the analysis of the collected data?**

  This will help determine the format in which we store and transfer the collected data.

- **What is the timeframe for the study?**

  Knowing how long we have to develop is critical in being to feedback our own limitations on what will be possible to implement.

I would expect further questions and more detailed specifications to arise organically during the discussion, and would therefore be taking detailed notes throughout.

After the discussion, I would make sure that our outcomes are recorded into a requirement analysis brief and shared around the researcher and team. This way we have a record of the desired outcomes and approaches as discussed by all, with plenty of opportunity for feedback and further discussion as necessary.

Problem 5

The output file name will be;

```
${Home}Ozone_2001-01-01_XXXX.png
```

where XXXX is a four digit number taken from simhour, padding zeroes to the left.

simhour and Home are missing initial definitions in this snippet. I can deduce that Home is likely the directory for output, and simhour is a part of the time stamp (simulated hour?), which corresponds to the frame number of the output png.

The TimeStamp variable does not match the date found in the received nc files. This could be a problem for reading in the files, which would require an adjustment either in the script or input file names. If the reading in is done elsewhere in the script, however, this would not be a problem here. Similarly the VarName Ozone and ozone are not case matched.