# **Final Project Feedback: Malcolm**

## **Positives**

### **Overall**

An excellent example of end-to-end data analysis. Well delivered presentation of findings and data-driven insights into Stroke incidence in Scotland. Well done.

### **Data Preparation, including cleaning and wrangling**

Old Python Work

Totally fine to use Python and you used it well for cleaning and preparing your data. Obviously it’s quite atypical to complete the project using both R and Python, but as one of your goals was to practice using both, it’s fine in this case.

for column in community.select\_dtypes(include=['object']):

community[column] = community[column].astype('category')

This was a cool way to convert all the string columns. Might use this in the future. I’d be cautious about converting ALL of the columns to this category type. It may have been applicable in this case. Your code was nice, followed a logical process and adhered to modern standard practices.

Old R Work

Good development workflow and exploratory analysis.

### **Analysis**

Pretty sound analysis. I thought all the plots were good, informative, and weren’t misleading. Some spaghetti plots - lots of lines in one plot but I think they were informative nonetheless. Looking at crude rate was excellent - if you hadn’t I would have questioned if you had looked at population of health board/council area as a factor. Sound conclusions/recommendations.

### **Advanced analytics including modelling**

Great to see an attempt at modelling stroke incidence.

## **Potential Improvements**

### **Overall**

Overall this was a good project that exceeded MVP. The main focus for potential improvement I can see is cleaning up the project appearance. Everything you’ve got is totally fine during project development and is how I’d encourage you to proceed with your investigation. These are some suggestions to help turn your project into one you can showcase. At the top level, I’d expect to see different folder names. E.g. instead of “python\_work” - maybe: “cleaning\_using\_python”. This would show what you used the language for and still show at the surface level use of both languages. Flesh out your readme. I’d definitely highlight in your readme your choice for using both languages.

### **Data Preparation, including cleaning and wrangling**

Old Python Work

Fantastic to see you effective use of Python for data cleaning and preparation. Turn your Python notebook into a Python script. You can accomplish this similar to how you would take content from your Rmd files and place them into an R script file. For bonus points, whatever file location you place your Python script (.py) file also put a file titled: “\_\_init\_\_.py”. This is an industry standard for shared Python projects. Basically you need that \_\_init\_\_.py file to be able to source efficiently (running your script without “opening” it and hitting run.) More on this: <https://docs.python.org/3/reference/import.html#regular-packages>

<https://realpython.com/python-modules-packages/#package-initialization>

When you convert your notebook to a cleaning script, remember to take out the development checks (checking unique values, looking at the data etc.). If you want an extension extension extension you can look into assertive coding in Python and use these to verify the data has been cleaned properly.

I’m not sure if it’s okay practice to call os.makedirs() on a project that others may run. I’d remove this from your cleaning script and just include your clean\_data folder that has clean\_data inside so they don’t need to make the folder. In general, a few more comments will be useful for the future (both for your future self and others looking at your project.)

Old R Work

Don’t name your files things like: “last\_try”. Go for “exploratoration”, it’s more informative/ professional. More comments. If you were including this formally, I’d expect more section headings, and interpretation of results in the notebook.

R Work

This was more like it: section headings, interpretation. I’d maybe flesh out the interpretation a bit and include some knitted summary table outputs etc. And title it! Give your notebook a more informative name like: exploration.Rmd, rather than “working\_out”.

Final Project Documentation

This was good, comprehensive documentation. A lot of good information here, and data-driven insights. Call it something else: “final project documentation” isn’t particularly informative. “strokes\_analysis” would be more appropriate.

### **Analysis**

Not much I’d change about this.

### **Advanced analytics including modelling**

Great to see an attempt at a model. I think the logic was good, but was a bit confused at the decision to drop the oldest age group - who I think make up the bulk of the data points. Having a direct answer to: “can we predict/explain the incidence of stroke?” would be nice to have. The answer may well be: not with this data. But I’d include the “failed” modelling in your report/documentation. Just leave it out of the presentation and readme.