COMP5721M: Programming for Data Science

Group project (Coursework 2): Data Analysis Project

Project Title

Give names and emails of group members here:

- *NameA*, *usernameA*@leeds.ac.uk
- *NameB*, *usernameB*@leeds.ac.uk
- *NameC*, *usernameC*@leeds.ac.uk
- *NameD*, *usernameD*@leeds.ac.uk

Project Requirements

PLEASE DELETE THIS WHOLE CELL BEFORE SUBMITTING YOUR PROJECT

The purpose of this assignment is to develop your skills in organising and presenting a data analysis project.

Since most of the marks will be awarded for organisation and presentation, it is suggested that you do not initially attempt anything too complicated. However, once you have managed to get a basic pipeline working that fits the guidelines, you are encouraged to extend and elaborate your analysis.

Your project should entirely be contained within this template file. You should keep the basic structure indicated below. To facilitate grading according to the marking scheme.

You *may* import any module that is provided with Anaconda3 Python.

Marking Scheme

The marking scheme is as follows:

- Project Plan:
  - Description of data to be used (10)
  - Overview of Project Aims (5)
  - Design (5)
- Program Code: (15)
  Code should be laid out in steps with explanations and intermediate output with comments. You should ensure that the steps do not require a large amount of processing time.

- Project Outcome:
  - Explanation of Results (10)
    This should include a qualitative description of the results as well as key figures and tables of results.
  - Results visualisation (10)
    This should be graphical representations of the results with brief explanations (ordinary tables will be graded as part of the explanation of results)
  - Conclusion (5)

Data Resources

You can use any data you like. Many useful resources are available.

The Data Resources section of the module (Unit 4.3 on Minerva) has links to several example data sets.

As a starting point you could browse the following:

- Kaggle
- Our World in Data
- scikit-learn datasets
- scikit-learn tutorial

Using this Notebook Template

Please use this notebook as a template for your project file. In the following cells of the notebook, *italic text* giving explanations and examples should be either deleted, or, in most cases, replaced by appropriate text describing your project. Text that is not in italic (which is mostly headings) should be left as it is. **Your project report notebook should the same overall structure as this template notebook.** An exception to this is the current markup cell describing the project requiements. You should delete this before submitting your notebook.

Project Plan

The Data (10 marks)

Here you should describe the data indluding details of: where it comes from, what data it contains, is it accurate. (Potentially you could create data from a simulation but you should still explain why and how you intend to generate this data.)

It can be just one dataset or several that can be combined somehow.

There are 10 marks for this, so a fairly detailed description of the data is expected (around 300-400 words)

Project Aim and Objectives (5 marks)

Here you should describe the general aim of your project in around 200-300 words.

This can can be anything from classifying items according to their characteristic features (which mushrooms are poisonous?) to simulating an evolving process (will the rabbits eat all the carrots or get eaten by the foxes?)

Here some ideas of general types of processing functionality that you could implement:

- *Classification: separate data items into classes according to their charactersitics (can be either a definite or a statistical kind of classification)*
- *Corellation: find correspondences between different attributes within a dataset*
- *Search: find solutions matching some criteria*
- *Visualisation: find informative ways to display the structure of a large and/or complex dataset*
- *Query Answering: create a system that enables one to retrieve information by evaluating some form of query representation*
- *Simulation: model the evolution of a complex process*

Specific Objective(s)

You should chose and list ***up to 4*** *specific objectives suited to the data you will be working with and the type of project you wish to carry out. There should be* ***at least one per person doing the project****. There is no need for the objectives them to be completely different. They could be different stages of the processing requirements, or different processing functions that the system provides. Or just different aspects of data analysis that will be conducted. Typically, it is expected that there would be one objective per person. Replace the following examples with your own objectives:*

- **Objective 1:** *create a Python datastructure that will ensure world peace*
- **Objective 2:** *stop global warming using a pandas dataframe*
- **Objective 3:** *end poverty using matplotlib*
- **Objective 4:** *end suffering with sklearn*

System Design (5 marks)

Describe your code in terms of the following two sections.

Architecture

Typically this would be a pipeline in which data goes through several stages of transformation and analysis, but other architectures are possible. This does not need to be particularly complicated. A simple diagram with 100-150 words of explanation would be a good way to present your architecture.

Processing Modules and Algorithms

Briefly list and describe the most significant computational components of your system and the algorithms you will use to implement them. This could include things like:

- *cleaning the data by removing outliers*
- *combining different datasets*
- *converting samples to a special representaion (such as feature vectors)*
- *constructing a special data-structure (such as a decision tree)*
- *running some kind of analysis*

Your list can be presented in similar form to the one just given, but should include a brief but more specific description of the components and/or algorithms. Probably three or four components is sufficient for most projects, but you may want to have more.

Program Code (15 marks)

Your code should be divided into relatively short cells, with brief explanation in markup cells between.

As noted in the assigment overview, it is not necessary that your coding be super complex in order to get a good mark. Although there is a mark for the coding achievement, it is only a quarter of the total.

The suggested length of the code is about 150 lines for 1 person or 500 for a 4 peope project. You should not use more than 500 lines of code.

You should divide the code in accordance with the specification of modules and/or algorithms you gave in the previous section. Complex modules should be further divided into several code cells.

Please note the following about your code:

- ***You may import any packages/modules you wish to use for your project.***

- ***You should submit your notebook in a state where all cells have been executed and are displaying the output that you wish to present for grading.*** (We will not normally run your code, unless for some reason we wish to check that the code functionality matches the given description.)
- Although you may be programming in a "pipeline" style, it is strongly suggested you use function definitions to organise your code. As well as being generally good programming practice, it will be helpful for easy and flexibily presentation of results visualisations later in the document. In the visualisation sections you will be able to produce images in a concise and flexible way by calling functions defined in the program code secion.

Brief Explanation of following code cell

Below is a silly example of some trivial data. Replace this markup cell and the one below with something more interesting. And go on adding more until you have achieved your objectives (at least to some extent).

In [1]:

```
## Code Cell
## This will typically consist of:

## (a) Code doing some data manipulation:

fm_data = { "souvenir"     :  9,
            "cute animal"  :  5,
            "meme"         : 36,
            "smiley"       :  3,
            "random image" : 13
          }
total = sum([fm_data[f] for f in fm_data])

## (b) Code for displaying some output:

print("The total number of fridge magnets is:", total)
```

```
The total number of fridge magnets is: 66
```

Comment on previous cell output (optional)

As well as describing code, it will in many cases be informative to describe the output that has been generated by a cell.

The previous output cell shows a key number in our fridge magnet analysis.

Brief Explanation of following code cell

Since fridge magnets often take the form of cute animals, we use `pandas` to convert the raw data into a `DataFrame`.

In [30]:

```
## Code Cell
import pandas
df = pandas.DataFrame.from_dict(fm_data, orient='index')
df
```

Out[30]:

	0
souvenir	9
cute animal	5
meme	36
smiley	3
random image	13

Comment on previous cell output (optional)

The output from the previous cell is very interesting.

The following cell defines a visualisation function for the data.

In [31]:

```
def fridge_sorted_bar(color='blue'):
    df.sort_values(0).plot.bar( color=color)
```

More code cells

You can add as many code cells as you require, but it is recommended that you break code into relatively small chunks and do not exceed the maximum number of lines stated above.

Project Outcome (10 + 10 marks)

This section should describe the outcome of the project by means of both explanation of the results and by graphical visualisation in the form of graphs, charts or or other kinds of diagram

The section should begin with a general overview of the results and then have a section for each of the project objectives. For each of these objectives an explanation of more specific results relating to that objective shoud be given, followed by a section presenting some visualisation of the results obtained. (In the case where the project had just one objective, you should still have a section describing the results from a general perspective followed by a section that focuses on the particular objective.)

The marks for this section will be divided into 10 marks for Explanation and 10 marks for Visualisation. These marks will be awarded for the Project Outcome section as a whole, not for each objective individually. Hence, you do not have to pay equal attention to each. However, you are expected to have a some explanation and visualisation for each. It is suggested you have 200-400 words explanation for each objective.

Overview of Results

Give a general overview of the results (around 200 words).

Objective 1

Explanation of Results

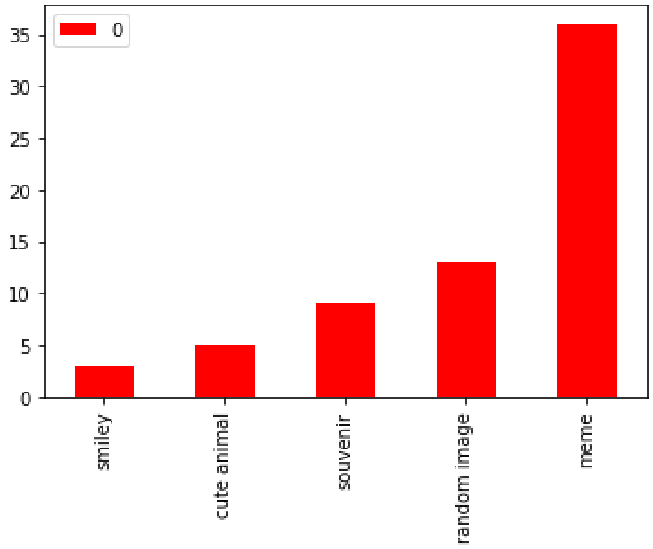
200-400 words

Visualisation

The following bar chart gives a vivid representation of the distribution of fridge magnet types, in which the dominance of 'meme' type magnets is dramatically illustrated.

In [32]:

```
fridge_sorted_bar(color='red')
```



Objective 2

Explanation of Results

200-400 Words

Visualisation

Objective 3

Explanation of Results

200-400 Words

Visualisation

Objective 4

Explanation of Results

200-400 Words

Visualisation

Conclusion (5 marks)

Your concluding section should be around 200-400 words. It is recommended that you divide it into the following sections.

Achievements

As we had expected, the most popular fridge magnets were of the 'meme' kind. We were surprised that 'smiley' fridge magnets were less common than expected. We conjecture that this is because, although they are apparently very popular, few fridges display more than one smiley. However, 'meme' based magnets can be found in large numbers, even on quite small fridges.

Limitations

The project was limited to a small number of fridge magents, which may not be typical of fridges found in the global fridge magnet ecosystem.

Future Work

In future work we would like to obtain more diverse data and study fridge magnets beyond the limited confines of student accomodation. We hypothesise that there could be a link between fridge magnet types and social class and/or educational achievement.