BM23BAM: Data Science Project Description

The data science project marks the capstone of BM23BAM. In groups of four to five students, you will apply everything you have learned about Python, data engineering, and analytics in one single project. The project's deliverables consist of a written report (in Jupyter notebooks) and a presentation. Please carefully read the instructions below.

1 Questions

You receive data on start-ups founded between 1990 and 2015. Using Jupyter notebooks, analyse the data and create a model that can predict the amount of funding start-ups receive. Specifically, answer the following two questions in separate Jupyter notebooks:

1. (Descriptives) How have start-up markets developed between 1995 and 2015? Can you spot major trends? What are the dominant markets? Use Pandas' built-in graphing tools (or any other Python packages that you are familiar with, such as Plotly) and text to visualise and describe the trends you can find in the data (all in the same notebook). You can use secondary data (e.g., GDP) to enhance your analyses (you can re-use any secondary data for Question 2).
2. (Model) Build a predictive model that can estimate the amount of funding start-ups receive. Use `funding_total_usd` as the target variable. You can use feature engineering and secondary data to improve the model's predictive performance. Use 5-fold cross-validation and test data to create a model that generalises well and does not overfit.

2 Report (75%)

You can use as many Jupyter notebooks as you need while you work on the assignment, but for your final answers please submit only one Jupyter notebook for each question. Upload the notebooks to your group's GitHub repository in the folder `final_submissions`. The notebook for the first question should be called `question1.ipynb` and the notebook for the second question should be called `question2.ipynb`. Both notebooks need to contain all code necessary to execute the analyses. Do not use external scripts.

You can use any other folder in your group's GitHub repository as you see fit. Only content in `final_submissions` will be evaluated.

The report is due 5 April, 12am.

Questions 1 is evaluated based on the extent of your analyses and the quality your insights. It is worth 33% of the report grade. There is no limit to your creativity for this question. The only requirement is that all analyses, graphing, and description are Python-based and in the same Jupyter notebook, and that the insights are broadly related to trends in entrepreneurship.

Question 2 is evaluated based on the rigour and correctness of your modelling process. For instance, how did you train the model and how did you evaluate its performance? Have you tried other models and how did they perform? It is worth 67% of the report grade. Consult the lecture material for how to structure your modelling process.

For both questions, make sure that every part of your data engineering and modelling is clearly described and can be replicated within the Jupyter notebooks. Do not use external code; all code

must be contained in the Jupyter notebooks. If a notebook doesn't run it will result in the deduction of points. You can use any external package that is available through conda. If you use secondary data, make sure it is included in your repository.

3 Presentation (25%)

Every group prepares a 7-minute presentation based on its report, highlighting the key insights from Questions 1 and 2. All presentations are scheduled 1-4pm on 5/4. The presentation slides need to be uploaded to Canvas (as a PDF file) before 5/4, 12am.

Group	Time Slot
1	1-1.15pm
2	1.15-1.30pm
3	1.30-1.45pm
4	1.45-2pm
5	2-2.15pm
6	2.30-2.45pm
7	2.45-3pm
8	3-3.15pm
9	3.15-3.30pm
10	3.30-3.45pm