COMP2200/COMP6200 Assignment 3

Due date: 11:55pm on Sunday June 8th, 2025

Scenario

You are still working for the venture capital fund from assignment 1. They decided to invest in the company you were analysing.

Now they are regretting it, so they want that company to get bought as soon as possible by some company in the USA, preferably a big one with lots of money that is listed on a USA-based stock exchange. The venture capital fund wants to speed this up by identifying influential board members who might be able to make connections to wealthy companies.

(Company directors represent the interests of shareholders, ensuring that executives act responsibly to make business decisions that increase share value. Alongside the CEO, they form the company's board.)

The VC fund's data scientist just quit (or maybe they were fired, nobody knows for sure), and you have been asked to look at what the data scientist was working on during their last day at work. You found a Jupyter notebook, along with two CSV files. These contain real data extracted from DEF 14A filings from US companies, detailing company directors, their ages, and compensation.

(Note: the data you will be working with in this assignment is not synthetic. It genuinely is extracted from DEF 14A filings.)

About this assignment

The relevant Unit Learning Outcomes (ULO) are:

- Identify the appropriate Data Science analysis for a problem and apply that method to the problem.
- Interpret Data Science analyses and summarise and identify the most important aspects of a Data Science analysis.
- Present the results of their Data Science analyses both verbally and in written form.
- Discuss the broader implications of Data Science analyses.

Total: 20 marks

Submission

Use `https://classroom.github.com/a/LtSmqM9i` to create a repository.
For submission:

- Include your modified Jupyter notebook.
- Optionally include any additional data files you used.
- Provide a separate document clearly explaining your modifications, rationale, and analytical decisions.
- If applicable, include presentation files (especially for option 4a).

Data

This zip file should contain:

assignment3.pdf The file you are reading now

company_directorships.csv Company directors and the boards that they are serving on (with a small amount of background information)

director-details.csv Compensation and demographic details for directors

directors-network.ipynb The Jupyter notebook that the previous data scientist left

Use of Generative AI

Feel free to use Generative AI tools (ChatGPT, Claude, DeepSeek, etc.) to assist your analysis and documentation. However, clearly indicate in your documentation where and how these tools were used.

Sample Gen-AI prompts

Programming prompt

```
Don't give me the answer directly, but help guide me to answering this
question.  I will be programming in Python using the pandas and scikit-learn
libraries.  If it appears that I don't know something that I would
need to answer a question, suggest functions or classes that I should
learn about.  Always add little bits of information that will guide
my journey.
   Here is my code so far, and the question I am trying to answer (....)
```

Proof-reading

If you are concerned about your grammar or language, try a prompt like this:

`Find any grammatical mistakes, typos or other language errors in this text. Don't make the corrections, just list for me what was wrong and explain the problem.`

Late Assessment Submission Penalty

Unless a Special Consideration request has been submitted and approved, a 5% penalty (of the total possible mark of the task) will be applied for each day a written report or presentation assessment is not submitted, up until the 7th day (including weekends). After the 7th day, a grade of '0' will be awarded even if the assessment is submitted. The submission time for all uploaded assessments is 11:55 pm. A 1-hour grace period will be provided to students who experience a technical concern.

For example, if the assignment is worth 8 marks (of the entire unit) and your submission is late by 19 hours (or 23 hours 59 minutes 59 seconds), 0.4 marks (5% of 8 marks) will be deducted. If your submission is late by 24 hours (or 47 hours 59 minutes 59 seconds), 0.8 marks (10% of 8 marks) will be deducted, and so on.

For details about Special Considerations, see `https://students.mq.edu.au/study/assessment-exams/special-consideration`

Tasks

1. Centrality extension (1.5 marks)

Find where the eigenvector and degree centrality measures are used. Add another centrality measure (be aware that it might be a lot slower than the ones already there) and explain why you chose that measure.

Explain what each of these measures means in the context of this data. That is, don't give us a definition of what the centrality measure means, tell us what it means for a company or individual to have a high or low measure and how that fits into the goals of the project.

(0.5 marks for the new measure, 0.5 marks for the existing measures) = 1.5 marks

2. Code repair (8 marks)

The code is not well-documented, and is also quite dense. (In your career you will often find this when you take over someone else's programs, particularly someone who was still working on it when you took over. There's often not a lot of time to fix these problems, so you need to choose what's the highest priority.)

Pick four places where you find the code to be inadequate (is hard to understand, needlessly discards some data, is poorly documented or has some other problem) and correct it accordingly.

Explain why you chose those four.

(If you choose to document the code better, that only counts as one fix, regardless of how many places you document the code.)

(1 mark for each explanation, 1 marks for each successful fix) = 8 marks

3. Explore something in the existing dataset (0.5 marks)

Identify one interesting feature in the dataset that is not currently being used in the analysis.

(0.5 marks)

4. Complementary dataset (2 marks)

Find a complementary dataset. The secret to a good data science project is to bring together data from different sources. If you could choose anything at all, what would you add to this project?

It does not have to be related to the VC scenario — but it does need to be related to the existing data.

It's OK if this is a proprietary (paid-for) dataset (e.g. from Stata or other commercial data provider); it's OK if it is a dataset created by hand by someone that you found on the internet: just make sure the URLs that you give for this dataset are something that we can find and consider.

Explain what you would do with this extra data.

(1 mark for a valid dataset, 1 mark for an explanation) = 2 marks

5. Choose two refinement options (8 marks)

Take two options from (a) - (c) below. *Each is worth 4 marks.*

(a) Improve the data visualisations and turn this into a presentation that you could give to a non-technical audience.

(b) Take the dataset from (4) and implement whatever it was you wanted to do with that data.

(c) Write up your take on the ethics of this project.