MCIS-6273 Data Mining (Prof. Maull) / Fall 2022 / HW3a

Points Possible	Due Date	Time Commitment (estimated)
10	Sunday, Dec 4 @ Midnight	*up to* 2 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Learn about data exploration tools in the commercial space.

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw3`. Put all of your files in that directory. Then zip that directory, rename it with your name as the first part of the filename (e.g. `maull_hw3_files.zip`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(100%) Learn about data exploration tools in the commercial space.

Most of the work we have done in this class has been centered around manually building data analyses in Jupyter notebooks. This is a wonderful way to do work when you are given freedom and the analysis may not follow a specific target output – when your focus is on getting very preliminary exploration underway.

As we learned in this course data cleaning is a significant task and requires care to get right.

There are a multitude of commercial tools that provide more structure and flexibility to pursue data cleaning and analyses which may be more appropriate, especially when the data is stored in Excel spreadsheets, the grandfather of tabular data formats (let's for now say CSV is the "great"-grandfather of tabular data).

You will be listening to a podcast about a product that makes cleaning and exploring data stroed in Excel sheets more efficient, especially for non-programmers, but also can be placed in a pipeline leading up to more complex analyses which might happen in the Jupyter (or other) space."

This podcast is an episode from the IBMs Make Data Simple Podcast:

> Aug. 3, 2022. Make Data Simple Podcast: *"Of course you're tired of spreadsheets. Data exploration is better. Ryan Buick, Co-Founder of Canvas."*

You can listen to the podcast from these links:

- Spotify
- Player.FM
- direct from player.fm download of MP3
- Apple Podcast (DRM required)

§ Task: Listen to the podcast and answer the following 4 questions:

1. List 3 things you learned from this podcast and relate them to things you have experienced in this course.

2. What motivated the core Canvas company business, and what product space did it emerge from? Your answer should include mentioning data and management team concerns.
3. What problem does Canvas solve? Please include details from what Ryan mentions in his interview.
4. What or who are the target market for the Canvas project?
5. In two or three sentences relate what you learned about data analysis among non-data scientists with what you learned in the podcast. Be complete, but brief (no more than 3 sentences, please).