SIT731

Data Wrangling

Learning Summary Report

Prateek Singh
221218743

Self-Assessment Details

The following checklists provide an overview of my self-assessment for this unit.

	Pass (D)	Credit (C)	Distinction (B)	High Distinction (A)
Self-Assessment			✓	

Self-Assessment Statement

Declaration

I declare that this portfolio is my individual work. I have not copied from any other student's work or from any other source except where due acknowledgment is made explicitly in the text, nor has any part of this submission been written for me by another person.

Signature: **Prateek Singh**

Portfolio Overview

This portfolio contains all the work that demonstrates that I have achieved all the Unit Learning Outcomes for, SIT731 – Data Wrangling, to minimum Pass level and aiming for a Distinction Level.

I started learning this unit while having an Artificial Intelligence Engineer background. This unit taught theories that are directly applicable to the software I write. I learned about the issues specific to data handling and processing in the real world.

The initial task *1P: Introduction to Python and Jupyter Notebooks* was fundamental in understanding the development environment to be followed throughout the unit.

Task 2P: Working with NumPy Vectors (Unidimensional Data) taught me how NumPy library can be used to write functions to compute statistical measures from scratch and how it comes loaded with aggregation functions to help compute basic statistical measures out of the box.

Task 3P: Working with NumPy Matrices (Multidimensional Data) extended the learnings from Task 2P on multidimensional data. This task also taught how relationship between different variables can be exposed using correlation coefficients.

In *Task 4P: Working with pandas Data Frames (Heterogeneous Data)* I learned that how Pandas can be used to perform several statistical functions on heterogeneous data which help identifying underlying traits in a real-world dataset. This task also taught me how statistical measures can be inferred by both numerical as well as visual analysis of the dataset.

In *Task 5D: Working with pandas Data Frames Part 2* I learned how Pandas data frames can be utilized to mimic SQL queries and operations, thus offering convenient way of data wrangling within code and not worry about database connection and management.

In *Task 7C: Tableau or PowerBI Dashboard* I learned how 3rd party frameworks can be used for performing data visualization tasks, offering similar functionalities as python libraries, but with point and click capabilities. This becomes important in real world data wrangling as they offer a standardized/out of the box way of handling real world data, moreover, the development time and effort is also minimized.

During these tasks, I have displayed hands-on capability as well as made sure that the task submission was user friendly and properly documented. I have demonstrated the hands-on capacity by successfully implementing the given tasks along with their optional features. I believe this makes me a suitable candidate to achieve a Distinction grade.

Reflections

The most important things I learnt:

My expectation from this unit was awareness about new age data wrangling techniques and strategies, and practical applications of frameworks like Tableau public. This unit has provided me a comprehensive understanding of skills required to perform data wrangling.

I feel I learnt these topics, concepts, and/or tools really well:

After completing the *OnTrack* tasks I have confidence in my hands-on ability to quickly prototype and develop a baseline solution for real world data wrangling use cases.

I found the following topics particularly challenging:

The most challenging task for me was converting SQL queries into their respective Pandas code version. Implementing these taught me how to handle different types of data with Pandas and processing it to gain meaningful insights.

I found the following topics particularly interesting:

Interestingly, in addition with being the most challenging one, *Task 5D: Working with pandas Data Frames Part 2* was the most interesting topic of this unit for me as it allowed me to think about data processing from various perspective.

I still need to work on the following areas:

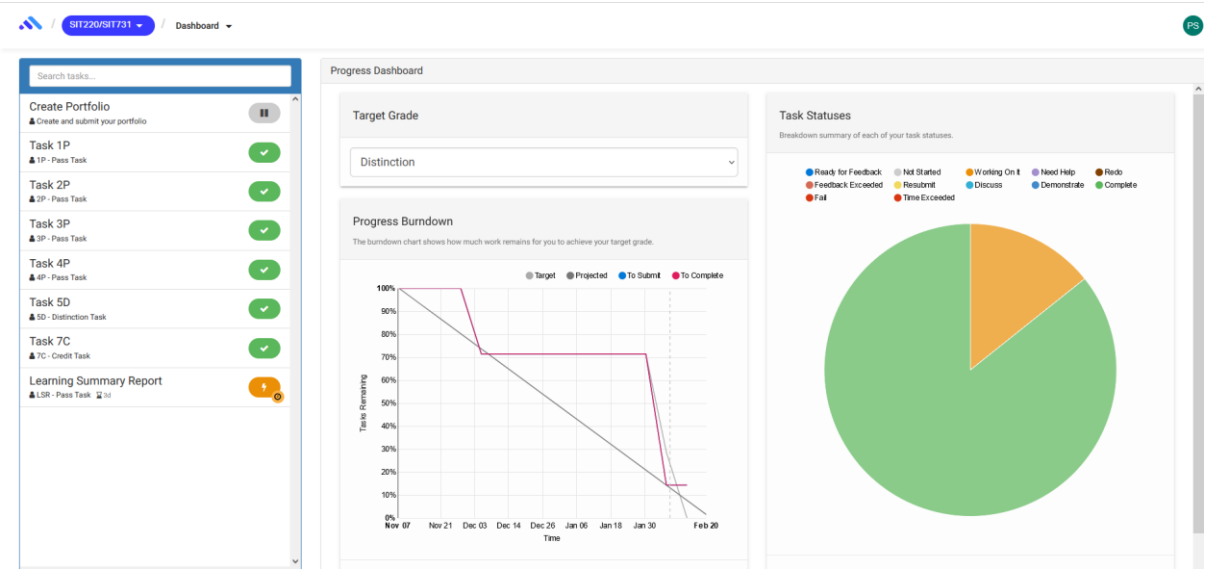
Along with the technical skills that I acquired during this unit, one of the most important skills which I picked up was effective time management. This taught me the importance of estimating the time required to come up with solutions for a problem statement. This is also a skill which I would like to improve in the future.

The things that helped me most were:

The material provided on the Cloud Deakin portal and the weekly lecture helped a lot in perfecting data wrangling. These were complemented by the weekly workshop sessions which helped me to put these skills into practice.

My progress in this unit was …:

The *OnTrack* screenshot denotes that I have taken up all distinction level tasks with best of my ability. It also shows that I could have finished some of the tasks sooner which would have allowed me time to revisit them and take up remaining tasks.



If I did this unit again, I would do the following things differently:

I have pursued this unit with satisfactory effort but if I did this unit again then I would aim for High Distinction grade. While I have completed all the distinction level tasks, I could not complete the last two HD tasks. I would also aim to improve my time management skills.

As an AI professional, this unit has made me aware of the importance of processing and visualizing real world data. I have also learned how these methods act as important tools to identify hidden patterns within a given dataset. Overall, this unit laid a strong foundation which is important for any professional working in AI and Data science domain.