University of Essex Online

MSc Artificial Intelligence

Numerical Analysis

End of Module Assessment 2: Individual Reflection

Due date: 27.01.2025

Contents

Reflection of Module

Data Analysis and statistical knowledge

At the start of this module, I was eager to deepen my understanding of numerical analysis and statistics, particularly because I enjoyed this topic during my bachelor's degree but had limited opportunities to work on statistical problems since then. This module, designed to provide a solid understanding of foundational mathematical and computing principles, has effectively enhanced my ability to interpret results from data science and AI tools, making it a perfect fit within our course program. The module's objectives align well with our MSc curriculum, strategically positioned after the Understanding Artificial Intelligence module and before the Machine Learning module, ensuring we build the necessary foundation to comprehend machine learning algorithms and their statistical underpinnings. One concept that has become much clearer to me is feature engineering, which focuses on preparing data for use with machine learning algorithms (Nargesian et al., 2017). Through hands-on experience with statistical measures such as mean, median, range, and standard deviation alongside data manipulation and visualization I have gained a better understanding of how feature engineering works and how machine learning algorithms, in general, function.

One of the challenges I encountered during this module was using R. Having not used the language for a while, I found it necessary to take extensive notes and engage in a lot of trial and error iterations to arrive at the correct solutions for the weekly data activities. Initially, this process was challenging and required additional time investment. However, it motivated me to document all data activities in my e-portfolio, which I believe will serve as a valuable cheat sheet for future reference.

Although documenting these activities was not a mandatory requirement, I consider it a worthwhile decision that will support my long-term learning. Given that AI is inherently data-driven, it inevitably involves statistical and mathematical foundations (Yu & Kumbier, 2018).

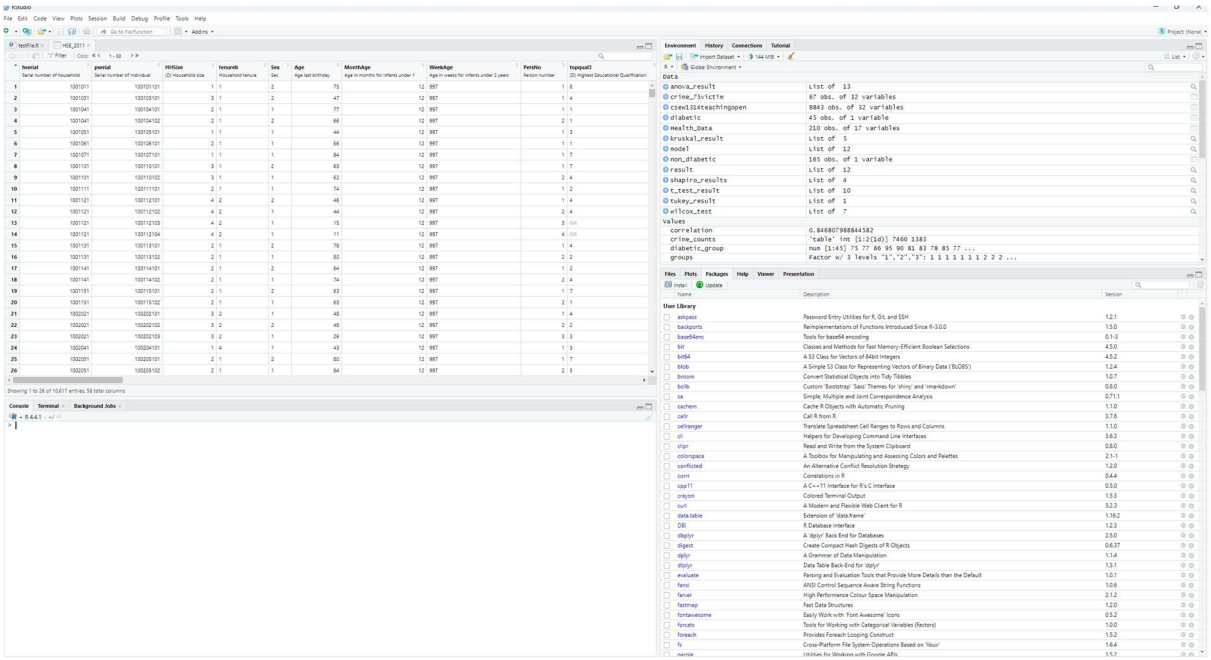
Interpreting statistical findings

During the module, we had the opportunity to work with p-values and confidence intervals, both of which are essential for interpreting statistical findings. P-values help determine whether a null hypothesis holds statistical significance and whether it can be rejected or not (Vidgen & Yasseri, 2016). In contrast, confidence intervals provide a range within which the true value is likely to fall, offering insights into the precision of the results (Simundic, 2008). Understanding these concepts has equipped me with valuable tools to derive insights from data by assessing statistical significance. Applying these techniques beyond the module exercises, I analyzed my personal data and identified a statistically significant correlation between the sleep scores provided by my smartwatch and the number of hours I slept during the night. While the relationship between sleep duration and sleep quality was expected, it was an excellent opportunity to apply the skills acquired during the module in a real-world context.

I plan to continue utilizing these statistical methods to effectively identify patterns within data. Applying these techniques to analyze raw datasets prior to their use in machine learning could provide valuable insights. Gaining a deeper understanding of the data before applying machine learning algorithms may enhance model performance and facilitate more informed decision-making.

Practical Application and Visualization

Reflecting further on the practical applications and visualization aspects, it is important to provide additional insights into R and R Studio. R is an open-source programming language designed for statistical computing, data analysis, and visualization. It is widely used in data science and statistical research due to its extensive libraries and built-in statistical functions. R is commonly utilized alongside R Studio, an integrated development environment (IDE) that offers a structured and user-friendly interface to enhance productivity and streamline the data analysis process (Verzani, 2011).

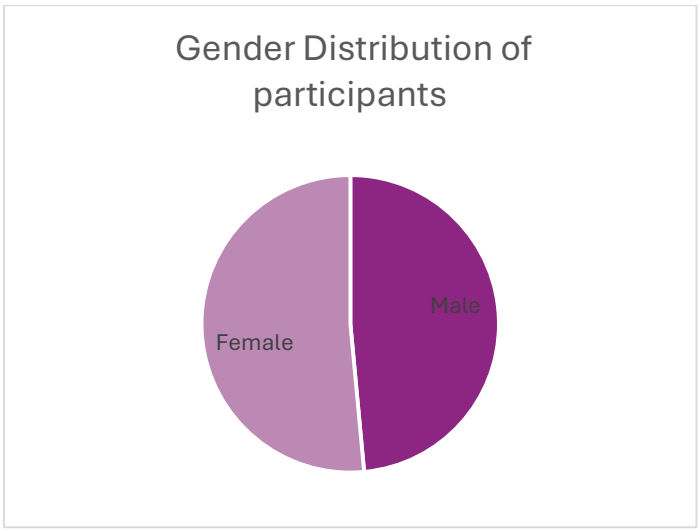


R Studio Screenshot

R Studio provided an excellent platform for visualizing data, creating summary tables, generating contingency tables, and analyzing data from various perspectives. Its user-friendly interface greatly facilitated data exploration and manipulation. However, despite these advantages, I encountered several challenges that required additional research to successfully complete tasks. Overcoming these challenges not

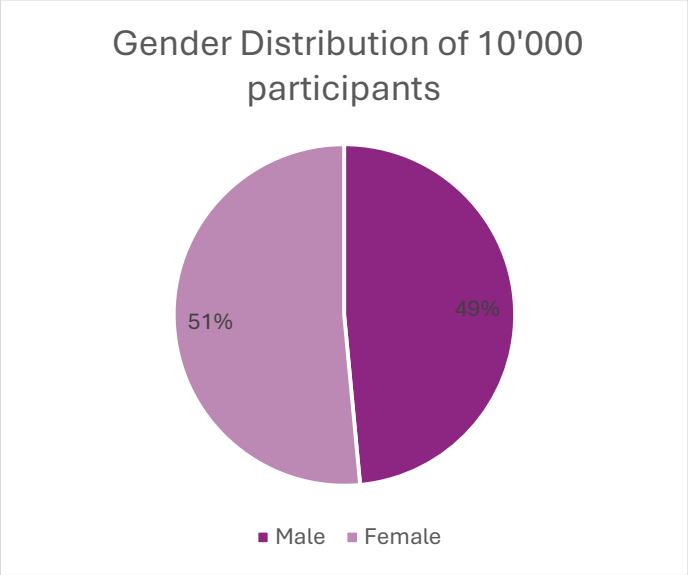
only deepened my understanding but also made the completion of each task more rewarding. One of my key takeaways from this module is the critical role of proper data visualization in the analytical process. Visualizing data using the ggplot2 package in R often provided me with valuable insights, enabling a deeper understanding of the data before proceeding with further analysis.

Throughout the module, we had the opportunity to collaborate with fellow students on the principles of effective data visualization, focusing on how to present data in a way that successfully conveys the intended message to the audience. One of the key learning experiences involved analyzing a poorly designed visualization and working together to enhance it. During our discussions, we agreed that a critical aspect of data presentation is ensuring transparency by always including the full source of the data. Additionally, selecting the most appropriate visualization technique to support the specific use case and context of the data is essential. For example, when the goal is to highlight the distribution of men and women within a sample, it is crucial to choose a visualization that effectively communicates this distribution rather than one that obscures or misrepresents the data like this one:



For a good visualization in this example, it is crucial to include the total sample size and present the data in percentages. By doing so, readers can easily interpret the

distribution of male and female participants while also gaining insights into the dataset's overall size. This approach enhances the reader's ability to assess the significance of the findings, providing essential context that aids in drawing meaningful conclusions from the data.



Data visualization is an important task that aims at simplifying complex data by presenting it in a clear and concise format, making it easier to understand and communicate key insights effectively to stakeholders (Embarak & Embarak, 2018).

Conclusion

PDP

My short-term goal after completing this module is to further enhance my proficiency in R, focusing on improving my practical skills and deepening my understanding of its applications. In the mid-term, I aim to apply the knowledge gained in this module to upcoming coursework, integrating statistical concepts into future learning. From a long-term perspective, my objective is to incorporate the statistical knowledge

acquired into AI projects, particularly in the field of machine learning, to develop more robust and data-driven solutions.

General Conclusion

The Numerical Analysis module has significantly enhanced my understanding of data analysis and statistics. It has equipped me with a valuable toolset that I can apply in future projects, while also clarifying concepts introduced in previous modules. This has not only boosted my confidence in statistical methodologies but has also strengthened my overall understanding of artificial intelligence.
I am eager to further develop my data analytics skills and apply the knowledge acquired in this module to upcoming coursework and practical applications.

References

Yu, B. and Kumbier, K., 2018. Artificial intelligence and statistics. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), pp.6-9.

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E.B. and Turaga, D.S., 2017, August. Learning Feature Engineering for Classification. In *Ijcai* (Vol. 17, pp. 2529-2535).

Vidgen, B. and Yasseri, T., 2016. P-values: misunderstood and misused. *Frontiers in Physics*, *4*, p.6.

Simundic, A.M., 2008. Confidence interval. *Biochemia Medica*, *18*(2), pp.154-161.

Verzani, J., 2011. *Getting started with RStudio*. " O'Reilly Media, Inc.".

Embarak, D.O. and Embarak, O., 2018. The importance of data visualization in business intelligence. *Data analysis and visualization using python: analyze data to create visualizations for BI systems*, pp.85-124.