Norah Masrour, Sam O'Brien, Masa Murray *DS 2002*

Step 6: Reflection Paper

The final data science project marked a significant step in our journey towards applying data science in the real world. As a team, we worked to begin an exploration into the complex world of energy storage data. Using datasets from the U.S. Department of Energy, we worked to pinpoint trends and insights that reflect energy practices throughout the United States. The course of this project left us with a deeper understanding of the impact of data science tools and techniques we learned throughout the semester.

During the project, we ran into several challenges. Data selection was one of the initial obstacles as we struggled to understand the purpose and relevance of each dataset. Differentiating between the geographic distribution of energy projects and cumulative storage capacity data was surprisingly challenging. Additionally, the complexity of the datasets, with technical terms like rated power and storage capacity, as long as the ETL setup and implementation added to the learning curve. Although the datasets were relatively clean, automating their upload to the cloud required us to backtrack and think back to what we learned throughout the course that could help us complete this task in the most efficient way possible. Setting up Google Cloud Storage and linking it to our analysis environment took longer than expected, and ensuring proper handling of data types during integration was challenging. The data analysis phase was also demanding as handling large datasets required more computational resources than anticipated. Finally, interpreting correlations between metrics such as storage capacity and discharge duration required us to quickly familiarize ourselves with statistical concepts.

Despite these challenges, we learned several valuable lessons throughout the project. One of the key takeaways was the importance of visualization in understanding complex datasets. Mapping energy storage projects geographically provided clear insights into regional trends and allowed us to better contextualize the data. We also realized the importance of data cleansing, as even pre-cleaned datasets required inspection to ensure accuracy. Additionally, analyzing correlations between variables such as storage capacity, rated power, and discharge duration gave us a deeper understanding of the data. On the teamwork front, dividing tasks among team members helped us tackle the workload more efficiently, though it also highlighted the importance of clear communication. Tools like GitHub were instrumental in facilitating collaboration and ensuring that everyone stayed on the same page. Looking ahead, we identified areas for improvement, such as prioritizing technical setup early in the project and allocating more time for data exploration to fully grasp the meaning of the data before diving into visualizations or analysis.

Through this project, we gained a range of new skills. We learned how to clean and prepare datasets for analysis, even when sourced from reliable fronts. Visualization skills improved as we worked with different graph types to create plots that effectively communicated insights. We also developed a basic understanding of cloud computing, particularly in storing and accessing data on platforms like Google Cloud Storage. Additionally, the project introduced us to statistical analysis concepts, such as correlations, which we applied to real-world datasets.

While we made significant progress, there are areas where we still need development. For instance, we aim to improve our proficiency with SQL to enable direct querying of datasets from

cloud storage. We also hope to deepen our understanding of advanced statistical techniques, such as regression modeling, to make more predictive analyses.

Despite the challenges we faced, the lessons learned and skills gained have laid a strong foundation for further studies in different topics using similar techniques. This project served as an excellent introduction to the field of data science on a zoomed-out scale, allowing us to explore real-world datasets and derive meaningful insights. As we reflect on our experience, we are excited to delve deeper into more advanced topics and improve our efficiency and effectiveness.