Progress Report

Which tasks have been completed?

- Selected a dataset which contains sensitive information (2 hours)
- Completed the user flow diagram on how the dataset is created (2 hours)
- Finished designing the ERD/UML of the system that needs to built to create this dataset (3 hours)
- Implemented the evaluation metric pmse to show that the distribution between the private and synthetic dataset are the same:: https://arxiv.org/pdf/1805.09392.pdf (15 hours)

Which tasks are pending?

- Implement the synthetic data generation model found in this paper: https://arxiv.org/pdf/1802.06739.pdf (10 hours)
- Implement the evaluation metrics to show privacy guarantees (10 hours)

Are you facing any challenges?

There are many different ways to prove that two datasets have the same distribution but there is not as much research done on the topic of whether there has been any private information that has been leaked from the real dataset A to the fake dataset B. I am currently reading papers regarding different re-identification attacks as well as theoretical guarantees of **differential privacy**.