DSCI 560 Lab 6 Meeting Notes

2/18
- Created a rough project timeline
- Assigned roles and responsibilities to each team member
- Outlined the initial steps for implementation

2/19
- Setup and connected Azure MySQL database
- Downloads PDFs from the provided Google Drive folder
- Started writing Python scripts or PDF text extraction
- Discovered the PDF extraction takes 25 minutes per file, which is too slow

2/20
- Began researching optimization strategies about PDF extraction
- Verified text extraction accuracy and handled scanned images PDFs
- Started implementing the web scraper for drillingedge.com

2/21
- Debugged web scraping script and ensured correct retrieval of production data
- Conducted initial tests on data preprocessing
- Implemented and tested optimizations to reduce PDF processing time
- Completed final implementation of scraper, database storage and data preprocessing

2/22
- Wrote and reviewed the lab report
- Ensured error handling and logging mechanisms were in place
- Tested data visualization setup for displaying oil well locations and attributes
- Discussed potential improvements and prepared for submission