Progress Update – 2024-07-13

Summary of Work Completed

1. Project Structure Setup

- Created the main project folder (`DataMining_GroupProject_GroupMembersName`) and all required subfolders:
    - `data/raw`, `data/transformed`, `data/final`
    - `notebooks` (with 4 placeholder .ipynb files)
    - `report` (with placeholder PDF and PPTX)
    - Added `requirements.txt`, `.gitignore`, and a comprehensive `README.md`

2. Repository Management

- Set up the GitHub repository: https://github.com/Testertesting-create/DSA2040A_DataMining_group1
- Configured the remote to the correct team repo and pushed all changes to the `Nicholas_Work` branch
- Ensured all files and folders are tracked and up to date on GitHub

3. README and Documentation

- Drafted a detailed `README.md` including:
    - Group/project info (placeholders for names/IDs)
    - Project summary (OSMI Mental Health in Tech Survey 2014)
    - ETL summary and planned techniques
    - Tools used and instructions to run notebooks
    - Team Members & Contributions table
- Updated the `todo.md` to reflect all completed steps and added detailed ETL/data cleaning plans

4. Data Preparation Planning

- Inspected the OSMI dataset and identified formatting and consistency issues
- Added a step-by-step plan for preprocessing and cleaning the raw CSV (removing extra quotes/semicolons, standardizing missing values, etc.)
- Outlined improvements/expansions for the ETL pipeline, including:
    - Standardizing categorical variables
    - Handling advanced missing values
    - Feature engineering (age groups, region, binary flags)
    - Outlier detection
    - Modular, well-documented code

5. Next Steps

- Begin implementing the ETL notebook (`notebooks/1_extract_transform.ipynb`) to clean and prepare the dataset for analysis
- Continue updating documentation and todo as progress is made
- Assign and track contributions as the project advances

This document summarizes all progress made to date and is intended to keep all group members informed and aligned. Please review the README and todo list for your next steps and assigned roles.