Big Data_Final Project_Taxi Fare Predictors_Shrumi's Individual Report

My Part:
In the NYC TLC Taxi Fare Prediction project, my primary responsibility centered around the crucial stages of data loading, preprocessing, and manipulation. My task was to handle the large dataset from the New York City Taxi and Limousine Commission, ensuring it was accurately and efficiently prepared for the subsequent machine learning processes. This role was vital as it laid the groundwork for the entire predictive modeling pipeline, and my expertise in data handling was instrumental in navigating the complexities of transportation data. Utilizing Amazon S3 for data storage, I ensured efficient handling and accessibility of data throughout the project.

Challenges Overcome and Solutions Implemented:
One of the most significant challenges I faced was managing the extensive and complex dataset. I developed a systematic approach for data loading, ensuring accuracy and efficiency in importing the data into our working environment. The preprocessing stage required meticulous attention to detail; I dealt with cleaning and organizing the data, handling missing values, normalizing formats, and extracting relevant features. In data manipulation, I transformed the raw data into a structured and analysis-ready format, optimizing it for the predictive models. My methodological approach during these phases was critical in enhancing the dataset's suitability for accurate fare prediction.

Project Outcomes and Personal Reflection:
My efforts in data preparation were reflected in the seamless integration of the dataset into the machine learning pipeline, contributing significantly to the robustness and reliability of the predictive models. The efficiency and effectiveness of my work facilitated a smooth workflow and set a high standard for data handling in the project. This experience underscored my skills in data analytics and adaptability in dealing with large-scale, real-world datasets, reinforcing my role as a valuable asset to the team.