To clean the data, we first developed a thorough understanding of the data. It was clear that we had to deal with the missingness in the csv files. We were mainly looking at the barts_to_hotspots file, and trying to clean this file initially. We were considering ways to impute the missing values, bringing into consideration methods like probabilistic imputation, backfill, frontfill, and single imputation. This discussion took a lot of time, and because of the limited time constraint, we thought it would be fairly best for us to impute by using the mean times based on other times of the same routes. If we had more time, we would have considered imputing based on looking at the same routes, on the same day of the week, at around the same time because we do recognize that there is a lot of room for error when just imputing based on the same origin and destination, as it can vary drastically during rush hour on a Monday compared to midday during Saturday. This is one thing we could improve on in the future. After filling in the missingness, we were able to find out which of the movement IDs corresponded to which location. We found out that the ID 3396 represented the Palace of Fine Arts, 3792 to Oracle Park, and 3394 Fisherman's Wharf, which represents the hotspots. For the BART stations, 3603 represents the Embarcadero, 3760 for Powell, and 3692 to be Montgomery Station.

After that, we were able to answer a couple questions for the first part of the task, where we were able to explore different variations between BART stations and hotspots, as well as vice versa, making inferences and analysis on the features that were given, including the different mean times throughout the week, correlations between the different features, and comparing the commuting times throughout the day as the time changed. For future continuation of this project, we would work on better visualizations based on the same time at the same day of the week, so we would compare seven days of commuting times based on early morning, midday, and evening to see how it differs on different days of the week. We can also look at the AM and PM times to see if there is also a difference there.