This week, we spent a significant portion of time sourcing and processing data for the second and primary model of the project. Understanding that some gaps in our prior datasets existed which would ultimately limit user experience, we worked to process data related to national public education and leisure activities. In order to quantify education, the allocated annual spending in the field per country was used. An important acknowledgement here is that these values are the gross spending, and thus will need to be converted to a per capita rate to be more easily compared. We hope that the additions of these data sources will provide the user with a higher degree of aggregation capability and improve the overall performance of the final model.

Currently, the primary model used to match user inputs to a target nation will likely be completed in the form of K-Means Clustering or K-Nearest Neighbors. Both of these models have a similar fundamental design and implementation, with some potential benefits and hardships with either one. The features to be used in this model include the spending on public education per country, the rate of crime (by type) per country, the happiness of populations per country, leisure rates per country, cost of living per country, and train coverage per country. All of these features were selected to provide the most accurate image of a given nation to a user, either by direct mapping or by proxy (for example, train coverage serving as a proxy for overall transit availability).

One shortcoming found within the first linear regression model during the residuals analysis was the model's apparent goodness of fit. When plotting Residuals vs. Predicted Values and Residuals vs. Index, a strong skew was seen. While no clear patterns proved a passing indication for many assumptions, this skew shows an extremely poor overall fit of the model.