My group project was finding out the variables that determine 3-year returns of small cap mutual funds. I, along with Meha Mehta, Charan Musunuru, Jacob Pammer and Chandler Wann, together worked on this project.

Upon deciding on our problem statement, each of us individually started hunting for relevant datasets. I and Chandler jointly discovered the small cap mutual funds dataset on the Fidelity Investments website. The data was part of 7 separate workbooks each with 7 worksheets, for a total of 49 worksheets. I was tasked with unifying the data into a single data frame. Upon creating a unified data frame, I further proceeded with data cleaning. This involved eliminating duplicate columns and dropping rows where the predictors had null values. After we had the relevant columns, each of the columns had to be further transformed to make the data ready for feeding into our statistical models. For this, prefixes such as the ‘$’ sign and suffixes such as the ‘%’ sign had to be removed and the strings had to be parsed as numbers. Further, new calculated columns were created based on existing columns in the data, such as creation of ‘Life of Fund’ column based on ‘Fund Start Date’.

Apart from data sourcing and data cleaning, I worked on fitting a Random Forest regression model on the data. The model was run with m=4 (total predictors were 12) and was able to achieve a test root mean squared error of 2.32%. Further, important predictors were identified, and results were corroborated with other team members who had worked on different regression models.

Lastly, I jointly worked on the presentation deck that was used for presenting our problem, work and the outcomes with the class. This involved selecting appropriate formatting and theme for the deck so as to make our presentation crisp and comprehensible.