The basis of this analysis was to understand commercial tenant behavior through the lens of lease activity, using the transaction type variable as the response variable. According to the classification provided by the data donor, Savills, lease transactions can be broadly categorized into “Go” or “Stay” decisions. “Go” transactions—such as New and Relocation—represent a tenant moving into a new space, either by entering a new market or relocating within an existing one. In contrast, “Stay” transactions—such as Renewals, Expansions, Restructures, Extensions, and Renewal and Expansion—indicate a continued presence in the current location, often with renegotiated terms or added space. Transactions labeled as TBD (to be determined) were excluded from our analysis due to the lack of definitive classification. By applying this “Go” vs. “Stay” framework, our goal was to uncover patterns in commercial tenant mobility, renewal behavior, and commitment levels in the office leasing market.

The dataset includes five files covering U.S. commercial real estate leasing trends from 2018 to 2024. We began our analysis with the Leases.csv file, which contains 194,685 lease transactions, each representing market-level details of office rentals. To narrow our focus, we filtered for leases in the Technology sector with square footage of 10,000 SF or more. After removing duplicate records based on the CoStar ID and dropping variables deemed irrelevant through judgment, we were left with a cleaned dataset of 1,554 observations and 20 variables.

Preliminary exploration and data visualization guided our selection of predictor variables. We identified Region, Year, Internal Class, Space Type, Availability Proportion, Leased SF, and Overall Rent as the most relevant features for modeling and interpreting transaction behavior. After creating a new data-frame with only the variables we were to use, we ensured that they were all properly formatted for the models we decided to create. The numerical variables were scaled before model running due to the large differences in size of the variables.

We split our data set into two new data sets, one for training our models and one for testing them. 80% of the data was put into the training data frame, and the remaining 20% went to the test set. We then tested the proportions of the data set to see if there was an imbalance, which would later tell us which error metric to use in comparing the models.

We chose to run a General Logistic Regression Model, A Classification Tree model, and a Random Forest model due to the categorical nature of our response variable. After running the models, we predicted the values of our response variable for the test data set. We then compared those predicted values to the actual values of the test data set and created a confusion matrix. After the confusion matrix was created, we were able to calculate our error metrics. The error metrics we chose to find are the Accuracy, Recall, Precision, and F1-Score for each model.

The three models that we ran had similar error metrics, but the Classification Tree had superior metrics. The F1-Score of the Classification Tree model was the highest, with the Random Forest coming in second, and the General Logistic Regression model having the lowest score.

After performing our analysis, we found that the Year, Region, Internal Class, Space Type, Availability Proportion, and the Available Space variables were the most important in finding trends in the Transaction Type variable. The Space Type, Year, and Leased Square Feet variables were found to be the most important predictors from the Classification tree and Random Forest models.

In our search for the most appropriate model, we found that the Classifications tree had the best results in predicting our response variable, with Random Forest coming next.