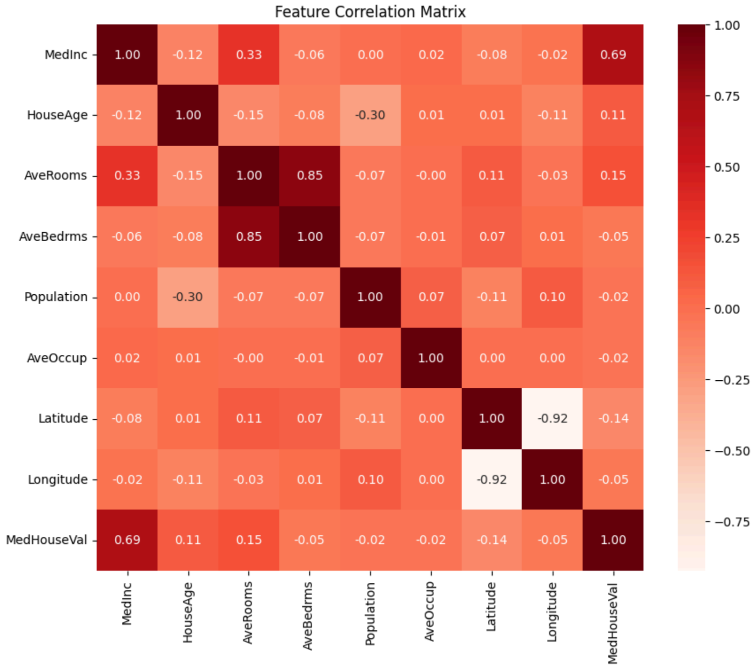
The objective for the predictive model we made for our final project was to be able to input features into it, and have them predict their effect on one specific other variable. In the case of our project, we decided to choose median horse value as our main variable. We chose this as our topic to further understand the drivers of housing value, and what features of a house can be used to predict whether or not it will be valued more or less than any other house. This was particularly important to us because in a time such as this, where the housing market is so crazy, and we are all beginning to look at housing options for our future, it was very useful to create this model which told us what to look out for. In the prediction, we found that the High Value homes are not randomly disturbed, instead, they are tightly clustered around specific areas. This means that if houses around a property are expensive, that would raise the value of the property itself. We also learned that the most important features for a home, in predicting its value, are its location, the owners median income, and the number of rooms in the house. This led us to believe that purchasing power of the buyer is a far greater predictor of home value, as compared to the physical size of the home.

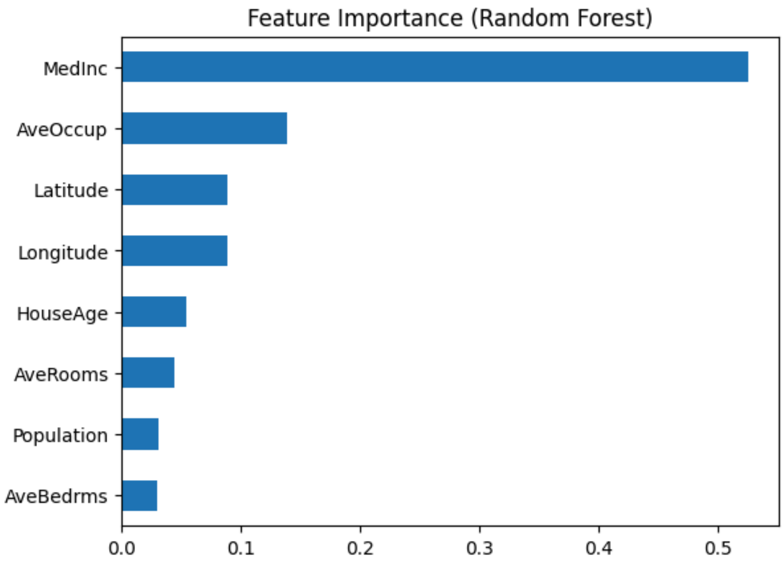
Our model is predictive because it uses algorithms to learn and understand different relationships in our data. It uses a Linear Regression model as well as a Random Forest model.

The data set we chose is the California Housing Dataset. It contained 20,640 records of homes with 9 different noted attributes, with their price. Those attributes were divided into three subsections pertaining to Demographics (Medium Income, population, and occupancy), Structural information (The houses age, number of rooms, and number of bedrooms), as well as Geographic information (the latitude and longitude). We thought that this data set was particularly informative, because it mixed the properties physical attributes, with its socioeconomic context, providing a fuller picture of the properties as compared to other data sets that we looked at. This also allowed us to create a far better predictive model, because if the data is better, then the model is going to be better.



We used the predictive model to create two different visualizations of our findings. The first, as seen to the right, is a heatmap which uses linear regression. It is a model which establishes direct relationships between features stated within the dataset, and value of home. This was very useful to us, because it showed us very simply what parts of the model were best suited to be used as predictors of house value. As you can see by the heatmap, it is clear that the most valuable homes were all distributed closely together. We thought this made sense, as it was probably referring to coastal/urban centers where the average property value is higher, due to demand and lifestyle. You can also clearly see that the best predictor of cost for a house is the medium income of the buyer, with the value closest to one, of .69. The second best predictor seems to be the house age with a value of .11. We decided that this also made sense, as newer homes will have newer appliances, and therefore will be furnished with nicer things, and will therefore be more expensive.

The second model that we used utilized a Random Forest Regressor. This estimator is designed to fit multiple decision trees to improve accuracy and control the fitting of the model. That model, as seen on the right, was a bit confusing when taken in consideration with the first model which we created. This model said that the most important features in this prediction was the medium income of the home's buyer, as well as the average number of occupants of the home. We decided that these both made sense because in American culture, a high proportion of your income should go towards buying your home, and therefore if you make more money, more money can be put towards buying your home. The average occupants also made sense because of things such as roommates, but also kids. The more kids someone has, the higher the number of occupants, also the older they probably are, and therefore the more time they've been able to devote to their career. It implies that someone would have a larger income.



When we looked at both of these models, and how they performed, we saw that for the Linear Regression model, there was a Mean Squared Error of 0.5559 and a R2 value of 0.5758. The Random Forest model had a Mean Squared Error of 0.2552 and a R2 value of 0.8053. These values were okay, but led us to believe that the housing market is, for the most part, non-linear and therefore, since the interaction of a property's age, location, and buyer's income cannot be captured by a single line, that the Random Forest model was the better choice.

While not perfect, our models were able to fairly accurately draw connections between a property's value and its many different features. While not perfect, the reliability of the Random Forest model is .80, it is fairly high. In summary, both models were usable, but the Random Forest model was better, with a higher reliability value. In the future if we wanted to refine this, we should have chosen a larger dataset which accounted for property outside of California, as restricting ourselves to just those properties gave a different view of property values as compared to all of America. We should have also chosen a more recent dataset, as the housing market has changed significantly in even just the past couple of years, such that we do not know how those predictors may have shifted. What may have been significant in the past, may not be now, or vice versa.