UCLA Extension Introduction to Data Science: Class Project

Assignment Summary:

Produce a data science project report for a management-level audience.

Demonstrate your ability to work through the Data Science Process of:

1. Ask an interesting question.
2. Get the data.
3. Explore the data.
4. Model the data.
5. Communicate and visualize the results.

Project Objective

To build a model that predicts median house values, using 1990's California Census Data

Data Source & Newly Created Metrics

Data Source: This data set appeared in a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in the Statistics and Probability Letters journal. The researchers built it using the 1990 California census data.

The original data file, housing.csv, contains 20,640 observations of 10 variables.

Features in the original data set:

- longitude (Numeric)
- latitude (Numeric)
- housing_median_age (Numeric)
- total_rooms (Numeric)
- total_bedrooms (Numeric)
  - This variable had 207 missing data points.
  - These missing data points were filled in with the median value of the overall variable: 435.0
- population (Numeric)
- households (Numeric)
- median_income (Numeric)

- median_house_value (Numeric)
- ocean_proximity (Factor with 5 levels)

New metrics created in the process of this analysis:

- **New Binary Variables from Ocean_Proximity**: In order to simplify the analysis process for the algorithm, this factor variable “ocean_proximity”, which contained geographical categories each census block group belongs to, became separate categorical variables. This transformed one categorical variable with 5 different possible values into 5 different binary variables with 2 possible values, 0 or 1.
  - <H OCEAN
  - INLAND
  - ISLAND
  - NEAR BAY
  - NEAR OCEAN

- **Mean Bedrooms and Mean Rooms**: The variables that contained total number of bedrooms and rooms were not very useful to be directly tied to the response variable, the median house value. Since each data point could represent different number of households, feeding the total number of bedrooms or rooms directly to the predictive algorithm would make it difficult to identify any meaningful correlation to help predict median house value. For example, here are two random data appoints that could misrepresent the relationship between an individual house’s number of bedrooms and the median house value. Data Point A has a higher total bedrooms, yet it shows a lower median home value:

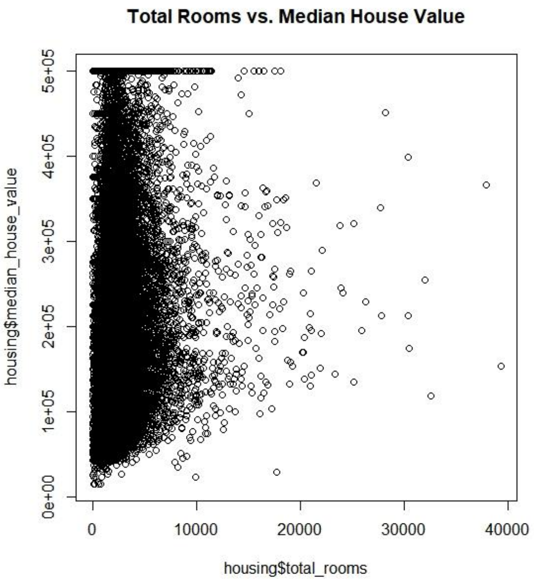
	Total Bedrooms	Households	Median House Value
Data Point A	1,106	1,138	$358,500
Data Point B	129	126	$452,600

Providing the mean bedroom metric in this case, make the relationship much more clear as Mean Bedrooms values and Median House Value data increases together:

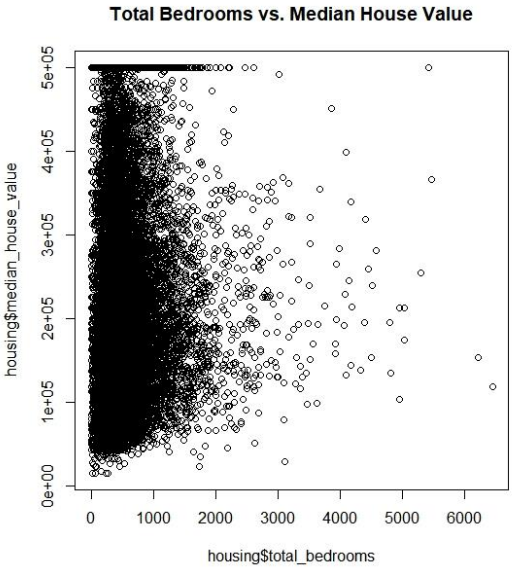
	Total Bedrooms	Households	Mean Bedrooms	Median House Value
Data Point A	1,106	1,138	0.97	$358,500
Data Point B	129	126	1.02	$452,600

Exploratory Data Analysis & Visualization:

- **Total Rooms and Bedrooms**: You can also see that these variables’ raw data does not provide much value to the overall analysis by reviewing the plots of these variables against the response variable, Median House Value:



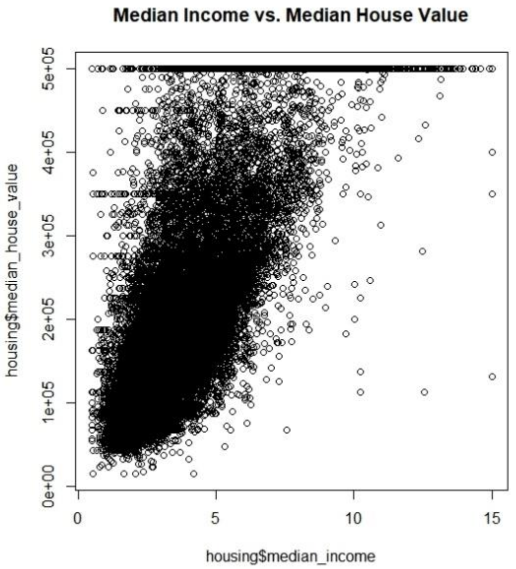
Total Rooms vs. Median House Value

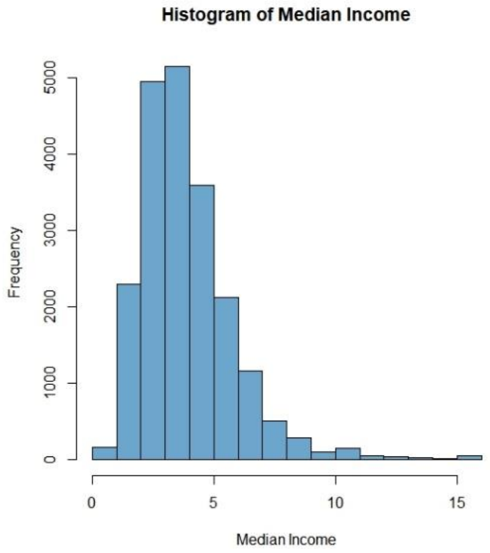


Total Bedrooms vs. Median House Value

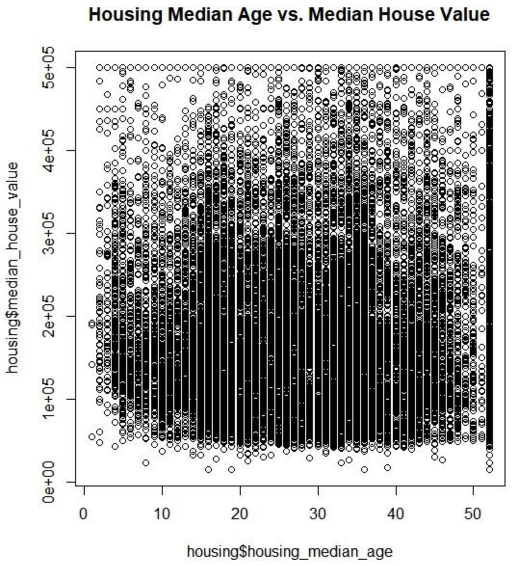
However, the total rooms and bedrooms can still be relevant if we divide them by the number of households in each Census block group to make the direct connection between the average rooms or average bedrooms per house and the median house value.

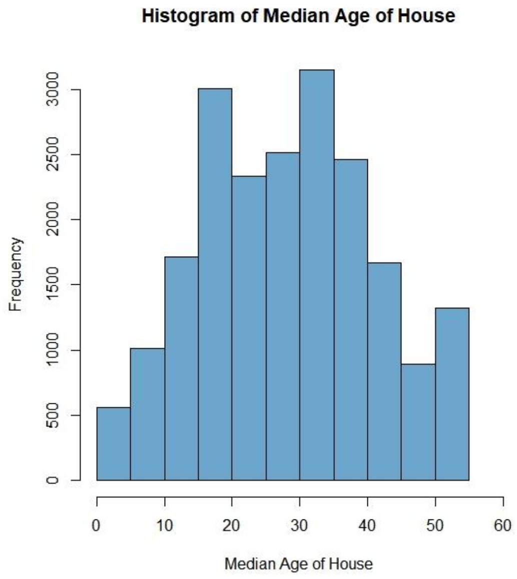
- Variables that suggest potential relationships with Median House Value:
  - **Median Income**: Median income seems to suggest some linear relationship with Median House Value. The density on the top line suggests that there may be some outliners that will affect the accuracy of the predictive model.



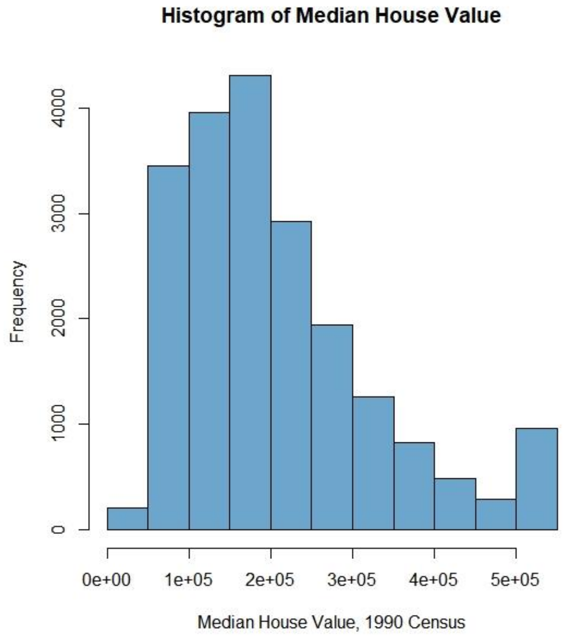


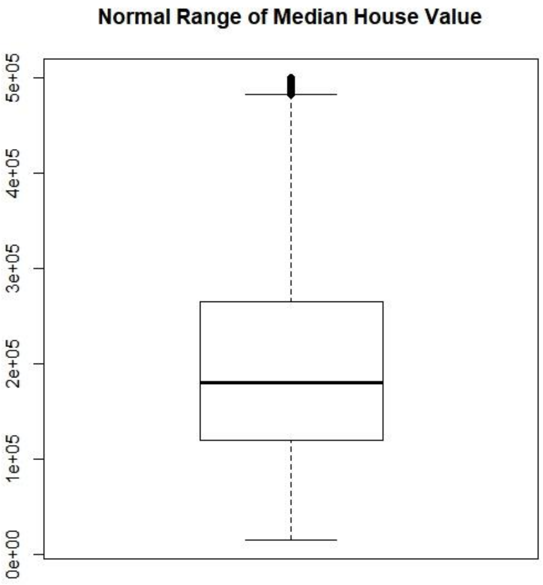
  - **Housing Median Age**: The plot does not suggest a linear relationship. However, the density of the data points does seem to suggest a potential non-linear relationship.



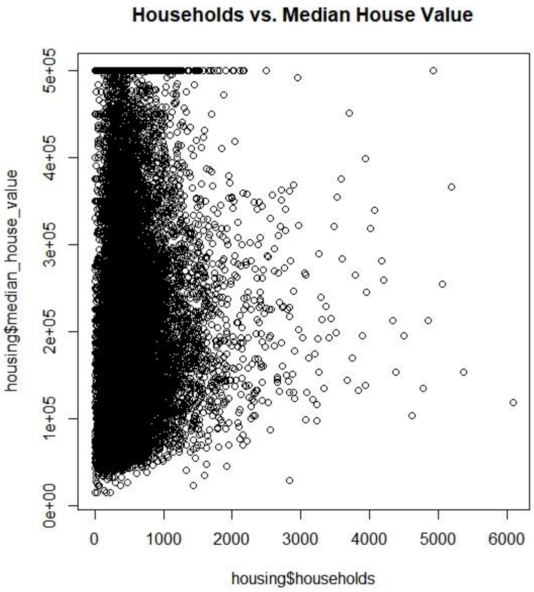


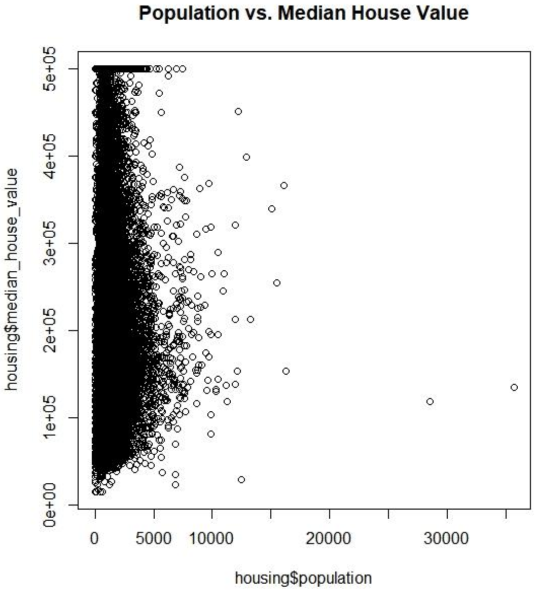
- Exploring the response variable: **Median House Value**
  - The boxplot shows that majority of the house median values range between $100,000 and $300,000 within the time period of 1990 Census.


Histogram of Median House Value


Normal Range of Median House Value

- Other numeric variables that don't seem to form any identifiable patterns:

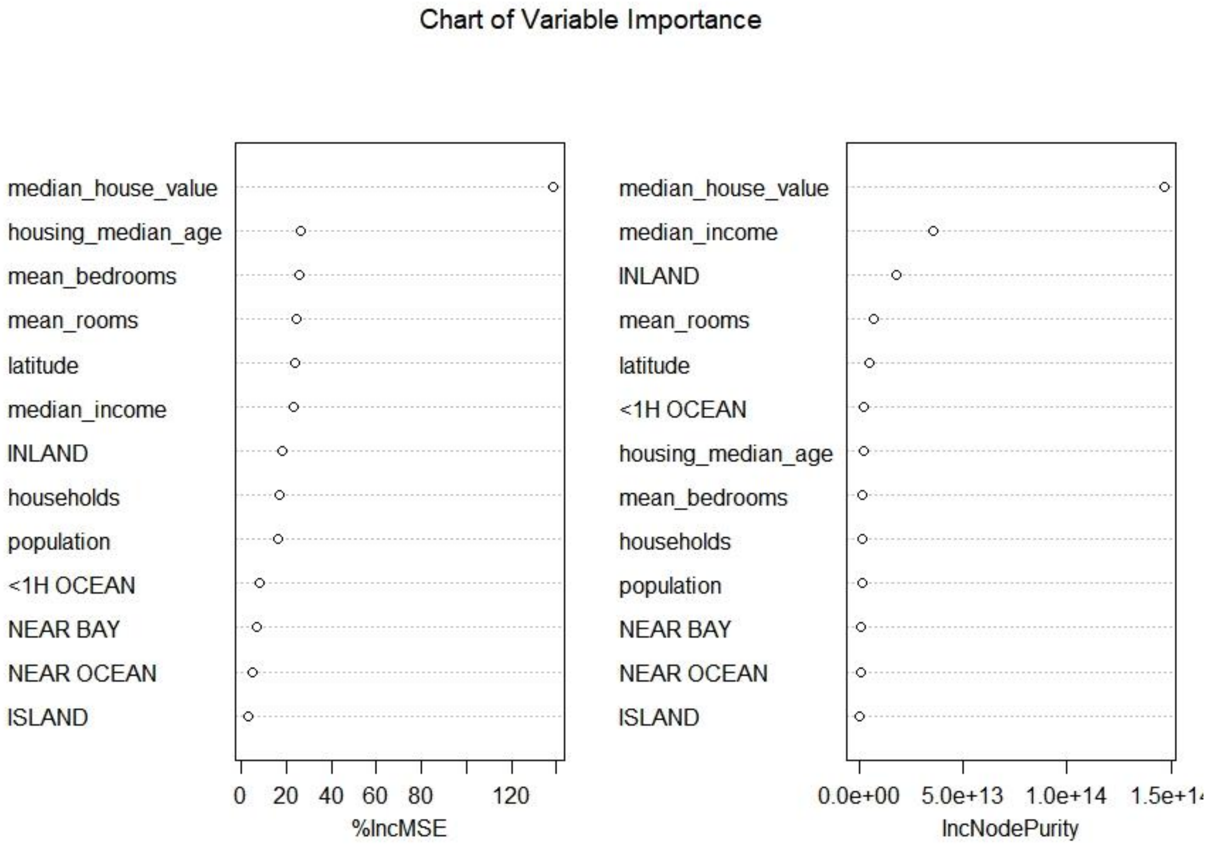

Households vs. Median House Value


Population vs. Median House Value

Overview of the Algorithm

Random Forest Approach:

- **Overview on Random Forest Algorithm Approach**:
    - As you’ve seen in the data exploration section, selecting relevant variables play an important role in developing a high-performing algorithm. Often times, selecting the right variables can be time-consuming and may not reflect the best combination of variables when done manually.
    - Selecting Random Forest method offer a simple solution to the issues mentioned above. The algorithm randomly tries many different versions of models using different random samples.
    - **Pros of this approach**: Random Forest will automatically generate a fairly good quality model without requiring a lot of time for developing the model manually. This approach promotes model variance and helps to avoid overfitting.
    - **Cons**: If you want to customize any aspects of how variables are selected or would like a high accuracy model, this may not be the best fit.
    -
- **Training Set:**
    - 80% of the data was used to train the model as a random sample.
        - 16,512 observations
    - This regression model accounts 99.84% of the data variance.
    - Reviewing the Chart of Variable Importance, the following variables appear to be good predictors:
        - Housing median age
        - Mean bedrooms
        - Mean rooms
        - Latitude
            - In terms of the variables that indicate locations, latitude appears to be the most effective variable since it may be sufficiently capturing California coastal properties that may be a significant predictor of housing prices.
            - Inland binary variable also seems to be a good predictor. However, having both Inland and Latitude may be redundant.
        - Median income



- **Test Set:**
  - 20% of the data was used to evaluate the quality of the model.
    - 9,272 observations

RMSE (The Square Root of Mean Squared Error)

- The Mean Squared Error (MSE) measures how close a fitted line is to data points. The fitted line here is the regression model we are using to predict the median house value. So, the bigger MSE or RMSE is the worse the model's performance is.
- The current Random Forest model shows RMSE of 4587.927 for the training model and 6029.655 for its performance against the test set. This means that when the model was tested on a new data, it didn't predict the median house value as well as it did with the data set the model was trained on.

Final Conclusion

Current Model Evaluation:

- **RMSE**: Although the model's performance was worse on the test set compared to its performance on the training set, it is an expected result for most cases. The model is developed using the training data set. So, the model will of course perform better on the training set, compared to a new data set.
- **Overall Observation**: It is hard to evaluate the model when it doesn't have a comparison. It is crucial that this model goes through iterations to further refine itself based on the insights we gather through each iteration.
- **Key Findings**: Here are key findings we uncovered in this iteration.
    - After exploring data with visualization, we learned that some variables are not meaningful in its original format. We created more meaningful metrics such as mean_bedrooms and mean_rooms.
    - Building a Random Forest regression model provided us with a smaller set of important variables that we can review further:
        - Housing median age
        - Mean bedrooms
        - Mean rooms
        - Latitude
        - Median income
    - Knowing that Latitude may be the best performing spatial predictor for median house value, we may be able to skip any efforts to transform the ocean_proximity categorical variable.

Next Steps:

- **Recommendations:** Here are some of the recommendations on what we may try for the next iteration:
    - Develop a regression model only with the important variables identified above and evaluate the model's performance against the current one.
    - Since mean rooms and bedrooms were good predictors, try a metric that gives population divided by households.
    - Since the missing data points for Total Bedrooms were a small portion of the overall data set, try removing them (instead of imputation) to mitigate any potential cases of creating unintended outliners when those imputed values are divided by a very low or a very high value of households variable.
    - Spend more time identifying and treating outliners. For this iteration, we saw some potential outliners, but we did not spend much time addressing them. Mitigating the impact of outliers can improve the model's quality.