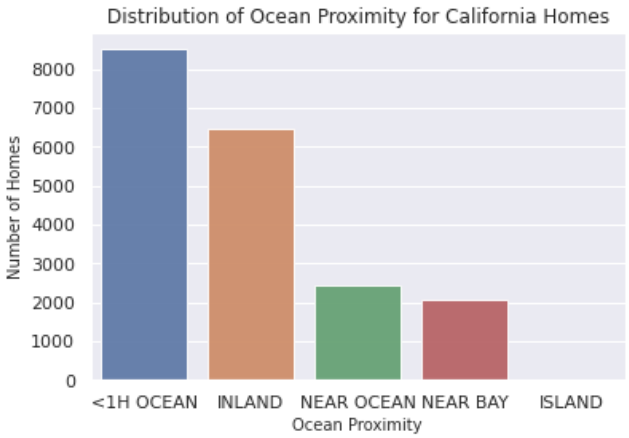
Assignment 1 Report

Matt Brown: meb180001

John Kenney: jfk150030

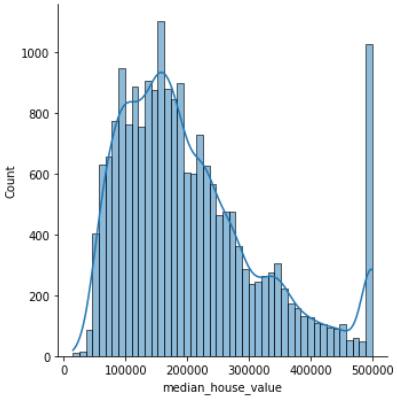
This report will supplement the text cells contained throughout the Google Colab notebook and comments supplied in the code cells throughout the notebook. The purpose of this report is to provide additional detail behind the decisions made in creating our model.

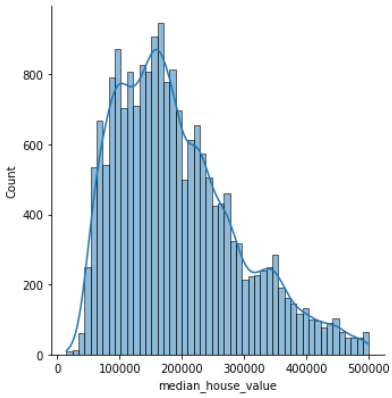
Exploratory Data Analysis



This graphic generated using seaborn gain us an idea of how many homes fit into each of the different buckets within the ocean proximity predictor. Out of ~20,000 roughly half of them are less than one hour from the ocean. Also, only 5 homes were on an island, this predictor was dropped due to it being an extreme outlier. After doing this we then transformed this categorical variable into numerical variables by using dummy variables.

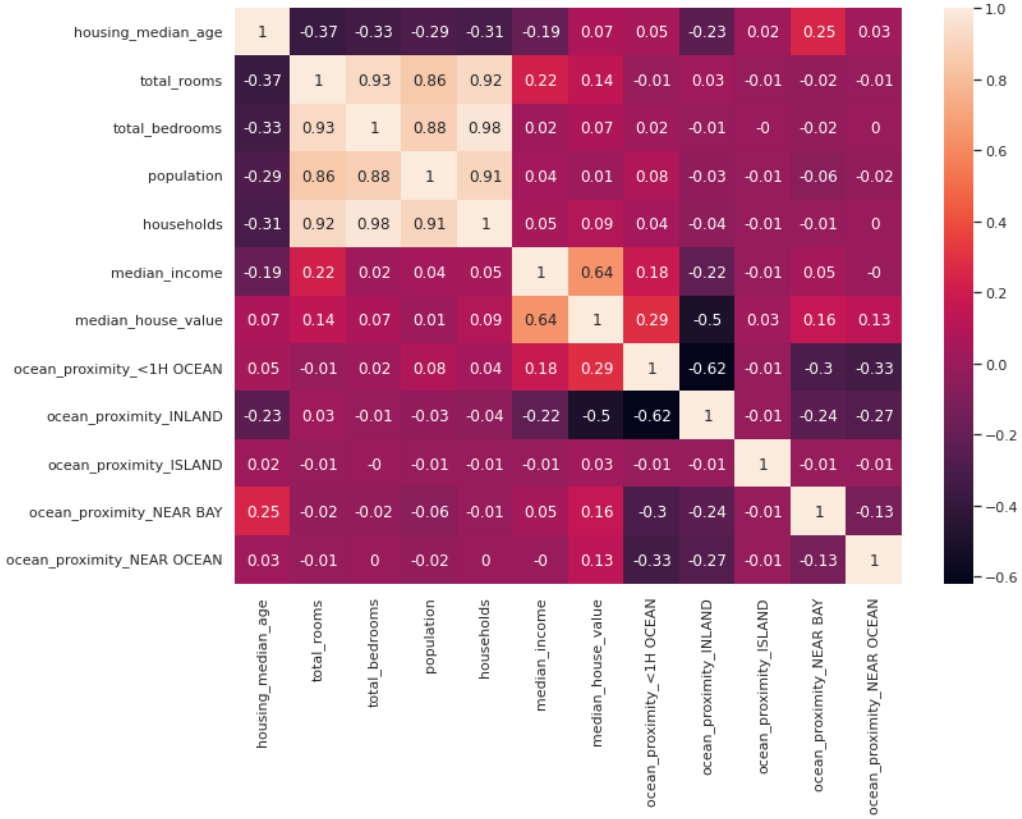
For our response variable we chose Median house value where we noticed that there seemed to be a case where there were a disporportionant amount of instances of median house values being max value of 500001 as seen in the graph below on the left. We then modeled our graph after removing all entries with median housing value of 500001. The after effects of or data after removing these outliers which we believe may be a possible cap imposed on the data or something can be seen on the bottom right where the the response varaible follows an approximately normal distribution.





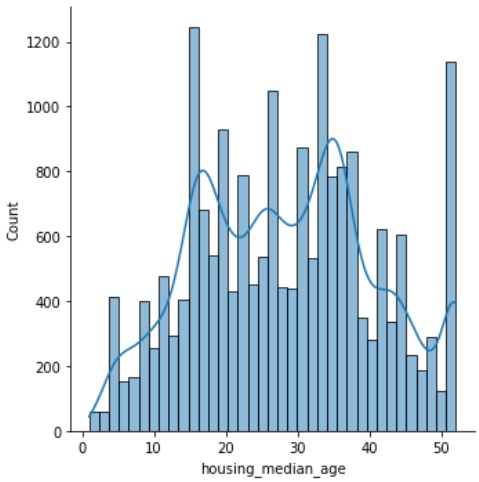
Variable Selection

For variable selection we looked at the heat map of our correlation matrix for all our variable except longitude and latitude that we dropped right away as location data would not be helpful for regression.

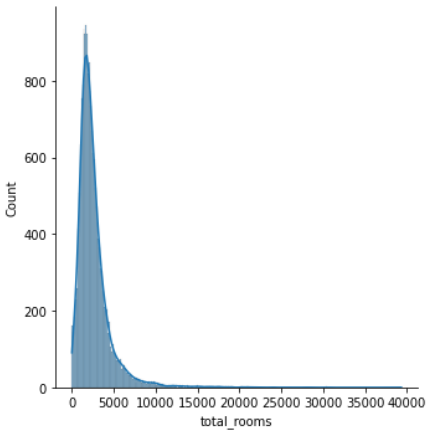


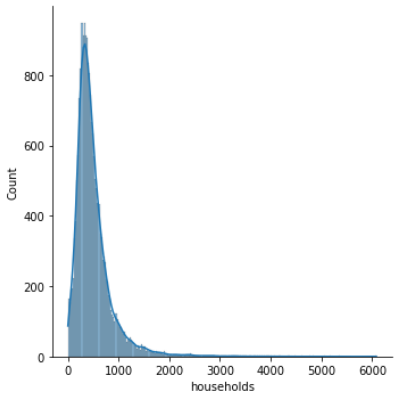
From the heat map we picked as our predictor variables as housing median age, total rooms, households, median income, and ocean proximity Inland against our response varaible median house value. We picked these varaibles from looking at our correlation matrix and picking the 4 most highly correlated variables to our response and picking just one variable that we got from our categorical varaible that had the highest correlation to median house value.

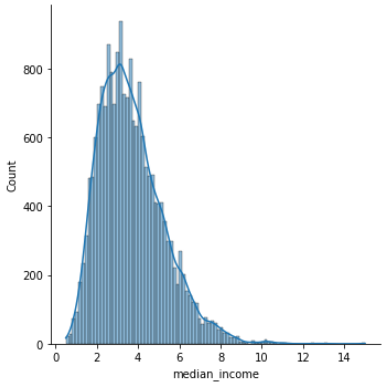
For housing median age the ditribution is roughly normal but could use a transformation or closer investigation for finding possible outliers.



For total rooms, households, and median income they all appear roughly normal but seem to be skewed to the left especially in the case of total rooms and households.

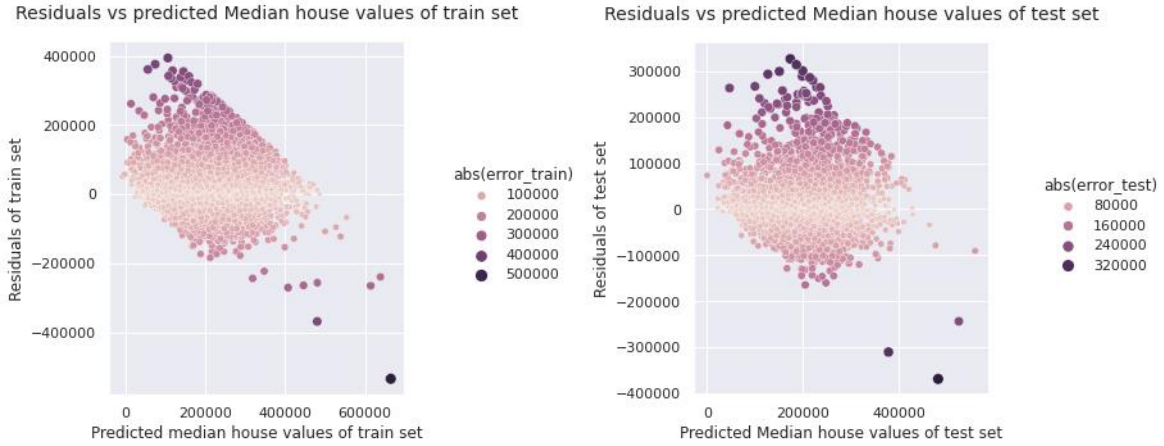






Model Diagnostics and Interpretation of the Results

At the bottom of this report includes a list of the parameters used and there corresponding MSE and R^2 values on our training and test data sets.
For this assignment we choose the SGDRegressor linear model from Sklearn to model our multivariate linear model. After running through variable parameter as can be seen below, we found that the best model gave us a RMSE of 64745.550282224074 and a R^2 of 0.5581069535263399 on our test data set. The exact parameters used on the model were: Loss Function "squared_loss", Penalty function: "l1", alpha: "0.0001", Max Iterations: "150", Learning rate: "adaptive", Initial learning rate: "0.1".
Finally, the best linear model fit found was:
y_hat = (192352.9732205975)*intercept + (12431.441005985747)*housing_median_age + (13644.026413561383)*total_rooms + (21082.164466225353)*households + (60785.87168355857)*median_income + (-31071.8637134103)*ocean_proximity_INLAND
Residual Analysis



From the diagnostic plots shown above we can see that while the residuals may be large, they are approximately even distributed across the fitted values. From looking at the range of the residuals we can see that our linear model may not be good for accurately predicting median house value, but because the residuals are randomly distributed across y = 0 that our model is a good linear model for our data. This tells us that we can get a good understanding of how our predicters are related to median house value. For further residual analysis our python code shows our residuals vs each predictor.

Also, we found that a simple linear model of median house value vs median income gave us a better R^2 and RMSE, but we stuck with the multivariate model so we could have a better understanding of how the different predictors influenced our model.

Log of parameters used, and the equation modeled in addition to its training and test RMSE and R^2 output

The highlighted run is the parameters used for our final model

```
-------------------------------------
-------------------------------------
New Run with New parameters:
```

```
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l2", alpha: "0.0001", Max Iterations: "100", Learning rate:
"invscaling", Initial learning rate: "0.01"
-------------------------------------
Model equation: y_hat = (192503.8759154889)*intercept +
(11101.121428303066)*housing_median_age + (-
14004.549645112364)*total_rooms + (20510.546550707622)*households +
(60390.734009468295)*median_income + (-
30856.881000610145)*ocean_proximity_INLAND
-------------------------------------
The model performance for training set
RMSE is 64151.813914807746
R2 score is 0.5695129281615786
-------------------------------------
The model performance for testing set
RMSE is 64805.92589419467
R2 score is 0.5572824334653953

-------------------------------------
-------------------------------------
New Run with New parameters:
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.0001", Max Iterations: "100", Learning rate:
"invscaling", Initial learning rate: "0.01"
-------------------------------------
Model equation: y_hat = (192504.15370707252)*intercept +
(11102.858839464492)*housing_median_age + (-14042.63279011226)*total_rooms
+ (20548.809567709326)*households + (60404.855684177084)*median_income +
(-30853.47972852039)*ocean_proximity_INLAND
-------------------------------------
The model performance for training set
RMSE is 64151.72427178158
R2 score is 0.5695141312495204
-------------------------------------
The model performance for testing set
RMSE is 64805.38779968956
R2 score is 0.5572897853517413

-------------------------------------
-------------------------------------
New Run with New parameters:
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.01", Max Iterations: "100", Learning rate:
"invscaling", Initial learning rate: "0.01"
-------------------------------------
Model equation: y_hat = (192504.15254617637)*intercept +
(11102.849433743346)*housing_median_age + (-
14042.452160986439)*total_rooms + (20548.626673655428)*households +
(60404.80714889313)*median_income + (-
30853.496842531444)*ocean_proximity_INLAND
-------------------------------------
The model performance for training set
RMSE is 64151.72462368961
R2 score is 0.5695141265266097
```

```
--------------------------------------
The model performance for testing set
RMSE is 64805.390780098984
R2 score is 0.5572897446311231


--------------------------------------
--------------------------------------
New Run with New parameters:
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.01", Max Iterations: "10000", Learning rate:
"invscaling", Initial learning rate: "0.1"
--------------------------------------
Model equation: y_hat = (194091.3121566784)*intercept +
(8650.07995924299)*housing_median_age + (-14875.481880845855)*total_rooms
+ (16079.378345831681)*households + (62620.96765643513)*median_income + (-
28730.663643108222)*ocean_proximity_INLAND
--------------------------------------
The model performance for training set
RMSE is 64544.79206711944
R2 score is 0.5642226594474902
--------------------------------------
The model performance for testing set
RMSE is 65469.88927722583
R2 score is 0.5481643147272001


--------------------------------------
--------------------------------------
New Run with New parameters:
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.0001", Max Iterations: "100", Learning rate:
"optimal", Initial learning rate: "0.01"
--------------------------------------
Model equation: y_hat = (194762.46495924247)*intercept +
(22151.59483650746)*housing_median_age + (-6392.436147005027)*total_rooms
+ (19849.038157135088)*households + (64402.11664023705)*median_income + (-
29347.0000421016)*ocean_proximity_INLAND
--------------------------------------
The model performance for training set
RMSE is 64887.84987772622
R2 score is 0.559578005527117
--------------------------------------
The model performance for testing set
RMSE is 65323.173303780844
R2 score is 0.5501871449024279


--------------------------------------
--------------------------------------
New Run with New parameters:
```

```
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.0001", Max Iterations: "100", Learning rate:
"adaptive", Initial learning rate: "0.01"
-------------------------------------
Model equation: y_hat = (192340.73526681922)*intercept +
(12426.907773000457)*housing_median_age + (-
13726.899772659546)*total_rooms + (21141.085080722085)*households +
(60821.14546710913)*median_income + (-
31044.047940538505)*ocean_proximity_INLAND
-------------------------------------
The model performance for training set
RMSE is 64135.61062779363
R2 score is 0.5697303631568998
-------------------------------------
The model performance for testing set
RMSE is 64745.97595258864
R2 score is 0.5581011430454124


-------------------------------------
-------------------------------------
New Run with New parameters:
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.0001", Max Iterations: "100", Learning rate:
"adaptive", Initial learning rate: "0.001"
-------------------------------------
Model equation: y_hat = (192342.81723833547)*intercept +
(12414.828879293245)*housing_median_age + (-
13638.703039181783)*total_rooms + (21063.84067946464)*households +
(60788.63900144657)*median_income + (-
31066.42709881929)*ocean_proximity_INLAND
-------------------------------------
The model performance for training set
RMSE is 64135.611058420334
R2 score is 0.5697303573789679
-------------------------------------
The model performance for testing set
RMSE is 64746.664385745
R2 score is 0.5580917457218643


-------------------------------------
-------------------------------------
New Run with New parameters:
Parameters used on the model: Loss Function "squared_loss", Penalty
function: "l1", alpha: "0.0001", Max Iterations: "150", Learning rate:
"adaptive", Initial learning rate: "0.1"
-------------------------------------
Model equation: y_hat = (192352.9732205975)*intercept +
(12431.441005985747)*housing_median_age + (-
13644.026413561383)*total_rooms + (21082.164466225353)*households +
(60785.87168355857)*median_income + (-
31071.8637134103)*ocean_proximity_INLAND
```

```
--------------------------------------
The model performance for training set
RMSE is 64135.60722897425
R2 score is 0.5697304087605428
--------------------------------------
The model performance for testing set
RMSE is 64745.550282224074
R2 score is 0.5581069535263399
```