1. DATASET AND DATA CLEANING

California Housing Dataset (https://www.kaggle.com/camnugent/california-housing-prices)
The dataset was obtained from the StatLib repository. It was then modified and used in Aurélien Géron's text, 'Hands-On Machine Learning with Scikit-Learn and TensorFlow'. It is generally good for practice experimentation of machine learning algorithms.

The data itself pertains to a census that was done in California in 1990. The features of the dataset include:
longitude : A measure of how far west a house is; a higher value is farther west,
latitude : A measure of how far north a house is; a higher value is farther west,
housing_median_age : Median age of a house within a block; a lower number is a newer building,
total_rooms : Total number of rooms within a block,
total_bedrooms : Total number bedrooms within a block,
population : Total number of people residing within a block,
households : Total number of households, a group of people residing within a home unit, for a block,
median_income : Median income for households within a block of houses,
median_house_value : Median house value for households within a block, (this is the target variable)
ocean_proximity : Location of house w.r.t ocean/sea

The data was provided as a CSV file. We used pandas, a library in Python, to read it into a dataframe.

Data Cleaning:

- The input feature *total_bedrooms* had 207 null values. We replaced these null values with the mean of the *total_bedrooms* variable, which was 537.87
- There was one categorical variable - *ocean_proximity*. To handle this, we divided our experiments into 2 main types:
    - Experiments where categorical data was converted into continuous data using pd.get_dummies,
    - Experiments where the categorical data was dropped from the dataframe

2. MODEL

We used the Linear Regression module imported from Scikit-Learn to build our model.
sklearn.model_selection.train_test_split was used to split the dataset into training and testing data. For this, we chose an 80-20 train-test split, and a random_state of 5 for shuffling the data.

3. PLOTS

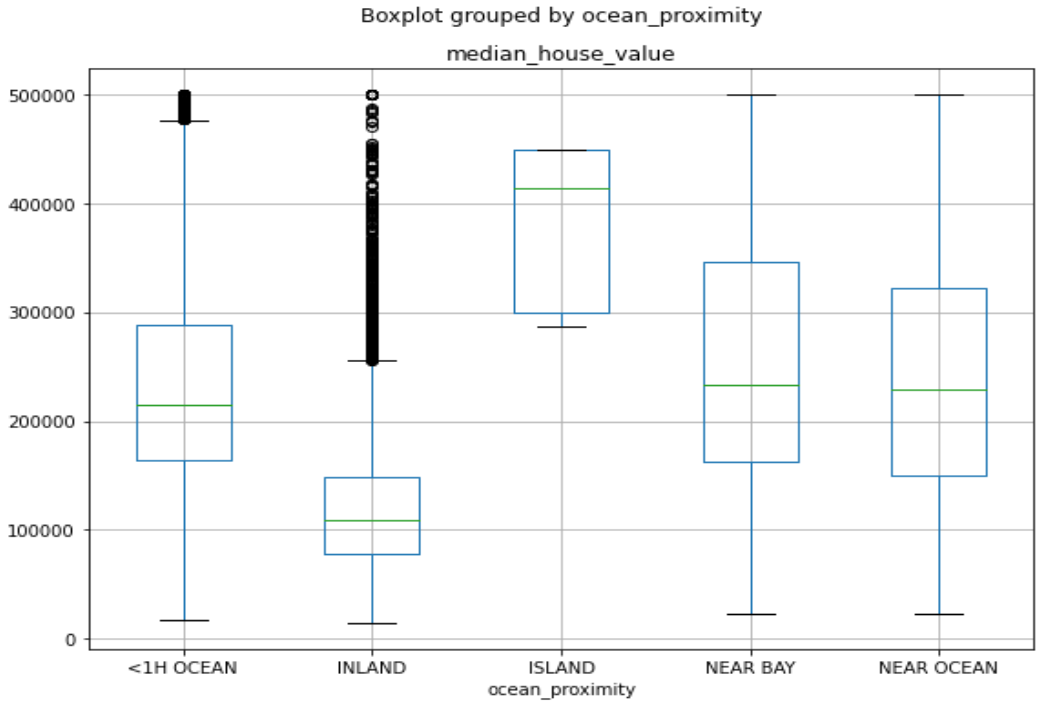

Figure 1: boxplot of ocean_proximity vs. median_house_value

Observations:

- o*cean_proximity_<1H OCEAN* and *ocean_proximity_INLAND* are the only *ocean_proximity* values that have outliers
- Houses that are inland have the lowest *median_house_value*

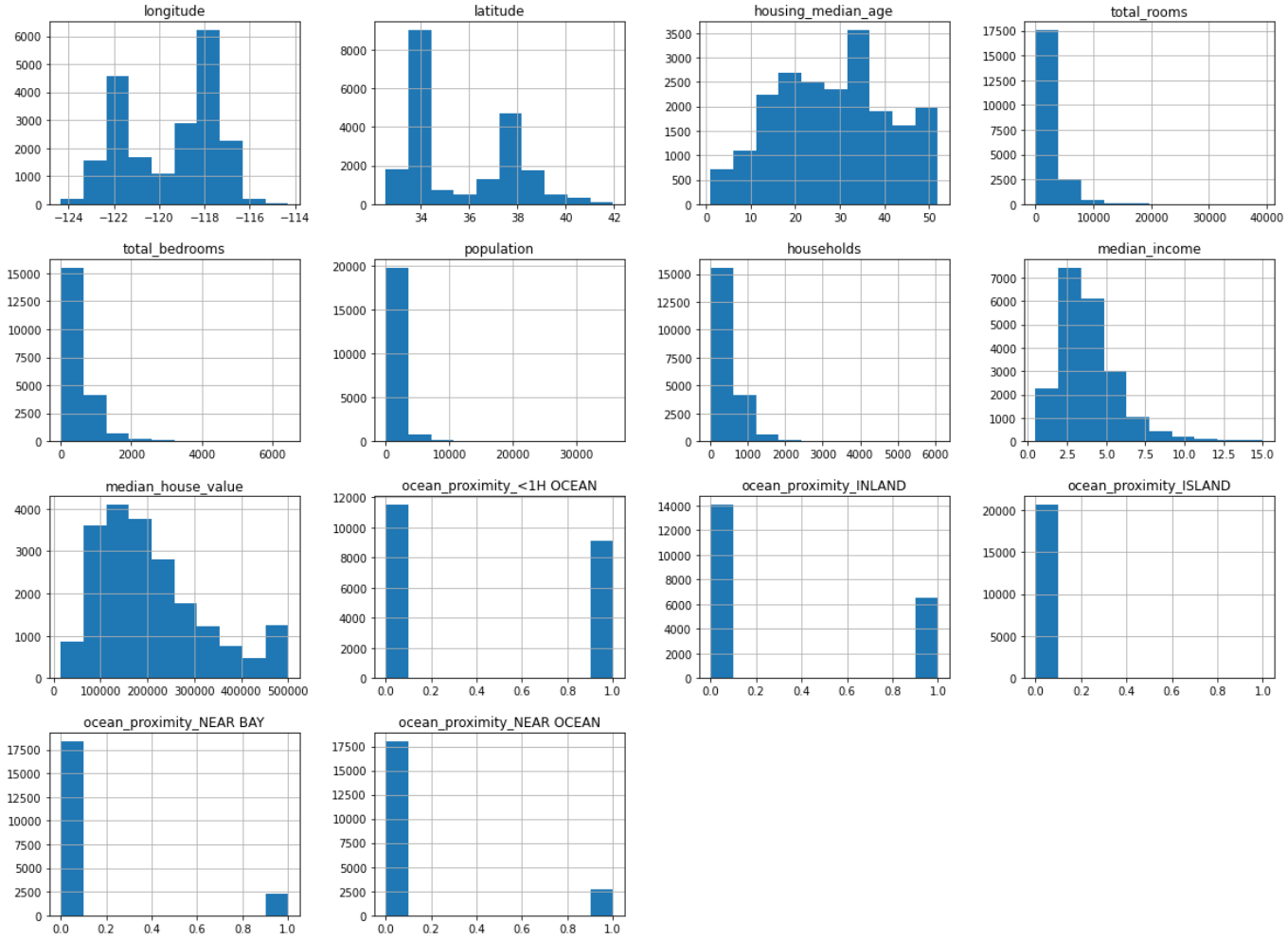

Figure 2: histograms for all the input features and the target variable

Observations:

- *median_house_value* is normally distributed, for the most part
- *total_bedrooms*, *total_rooms*, *population*, and *households*, are right-skewed

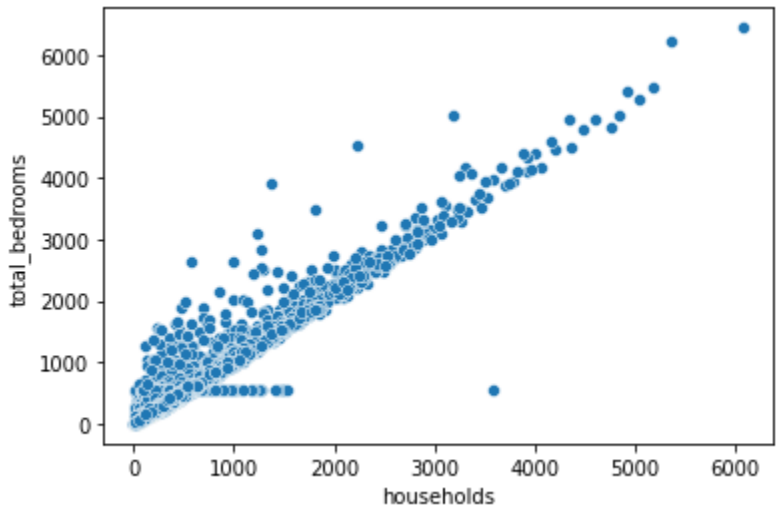

Figure 3: scatterplot of households vs. total_bedrooms

Observations:

- *households* and *total_bedrooms* have a linear relationship

---

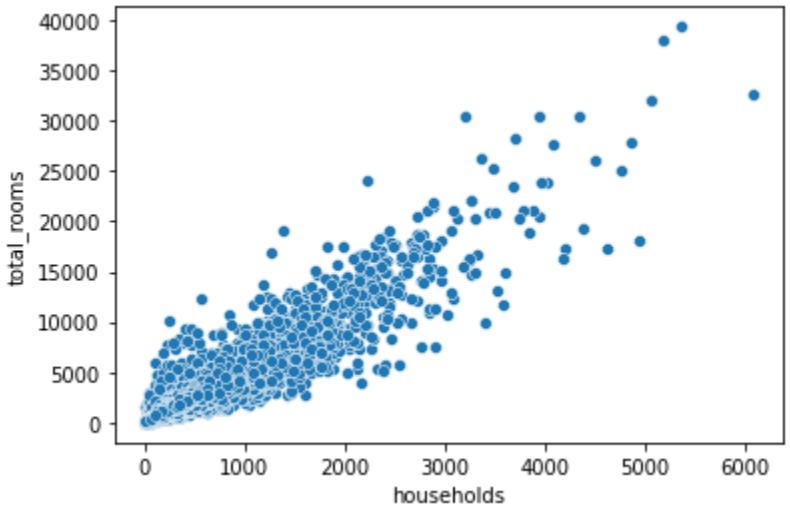

Figure 4: scatterplot of households vs. total_rooms

Observations:

- *households* and *total_rooms* have a linear relationship

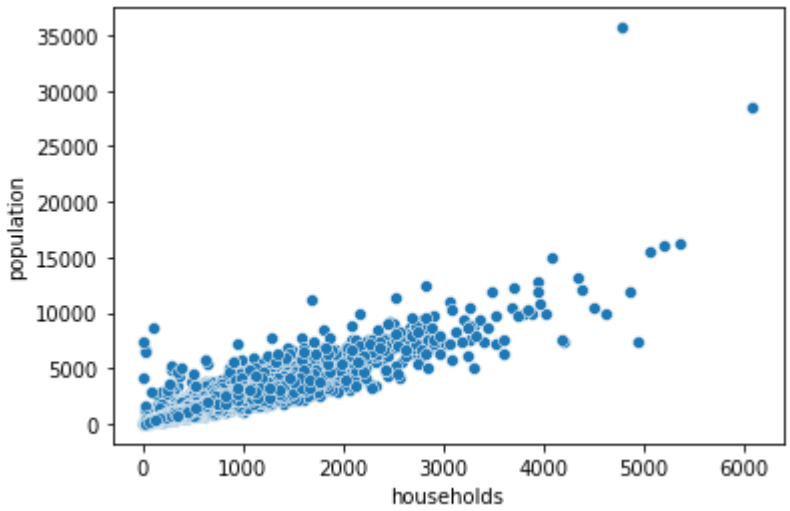

Figure 5: scatterplot of households vs. population

Observations:

- *households* and *population* have a linear relationship

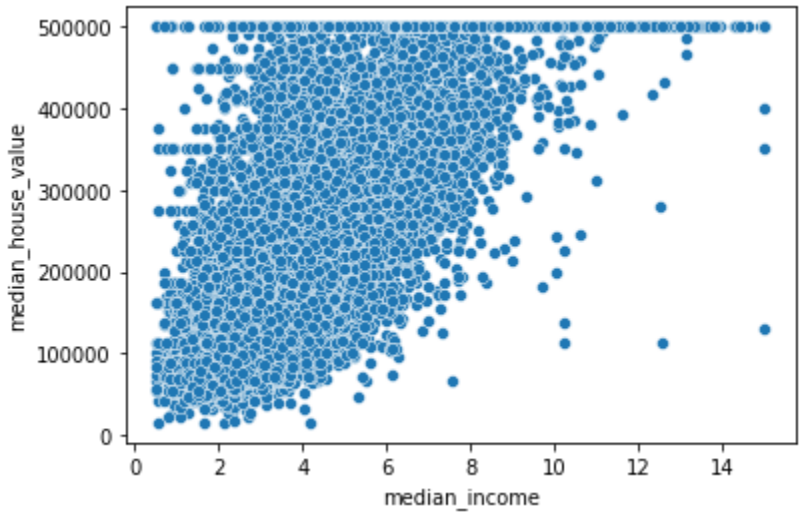

Figure 6: scatterplot of median_income vs. median_house_value

Observations:

- *median_income* is the only input feature that has a roughly linear relationship with the target variable

---

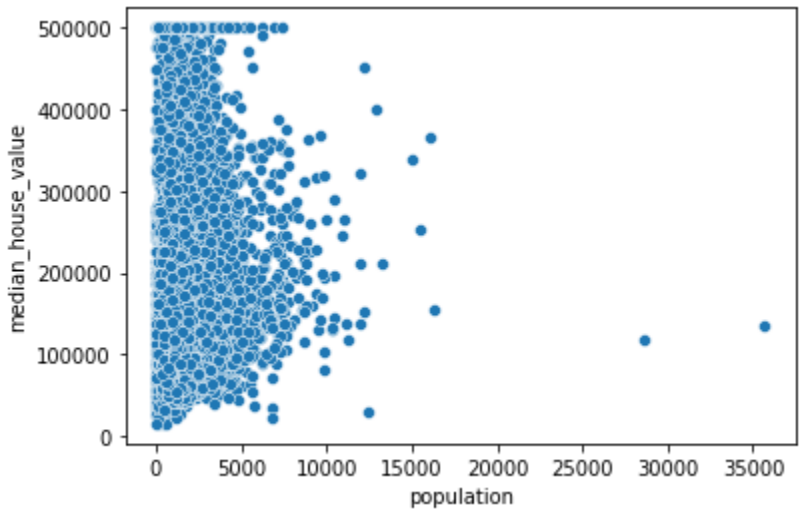

Figure 7: scatterplot of population vs. median_house_value

Observations:

- *population* does not have a linear relationship with the target variable

---

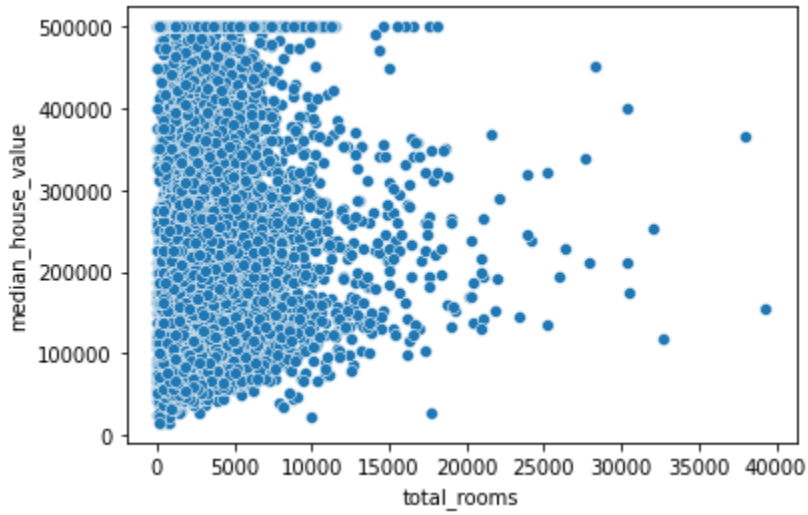

Figure 8: scatterplot of total_rooms vs. median_house_value

Observations:

- *total_rooms* does not have a linear relationship with the target variable

---

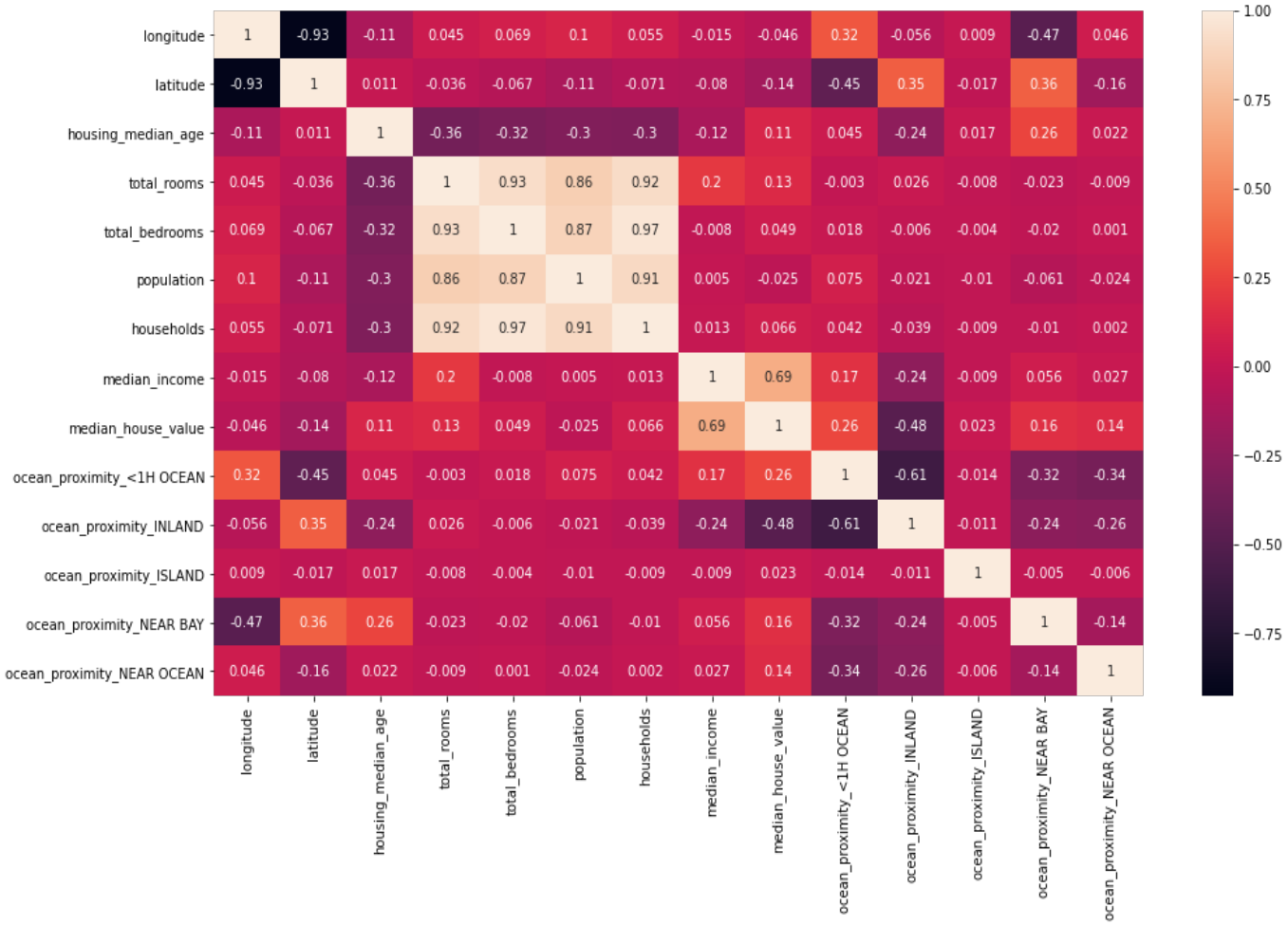

Figure 9: correlation matrix after encoding categorical data

Observations:

- *median_income* has the strongest correlation with the target variable
- Most other input features have a very weak correlation with *median_house_value*

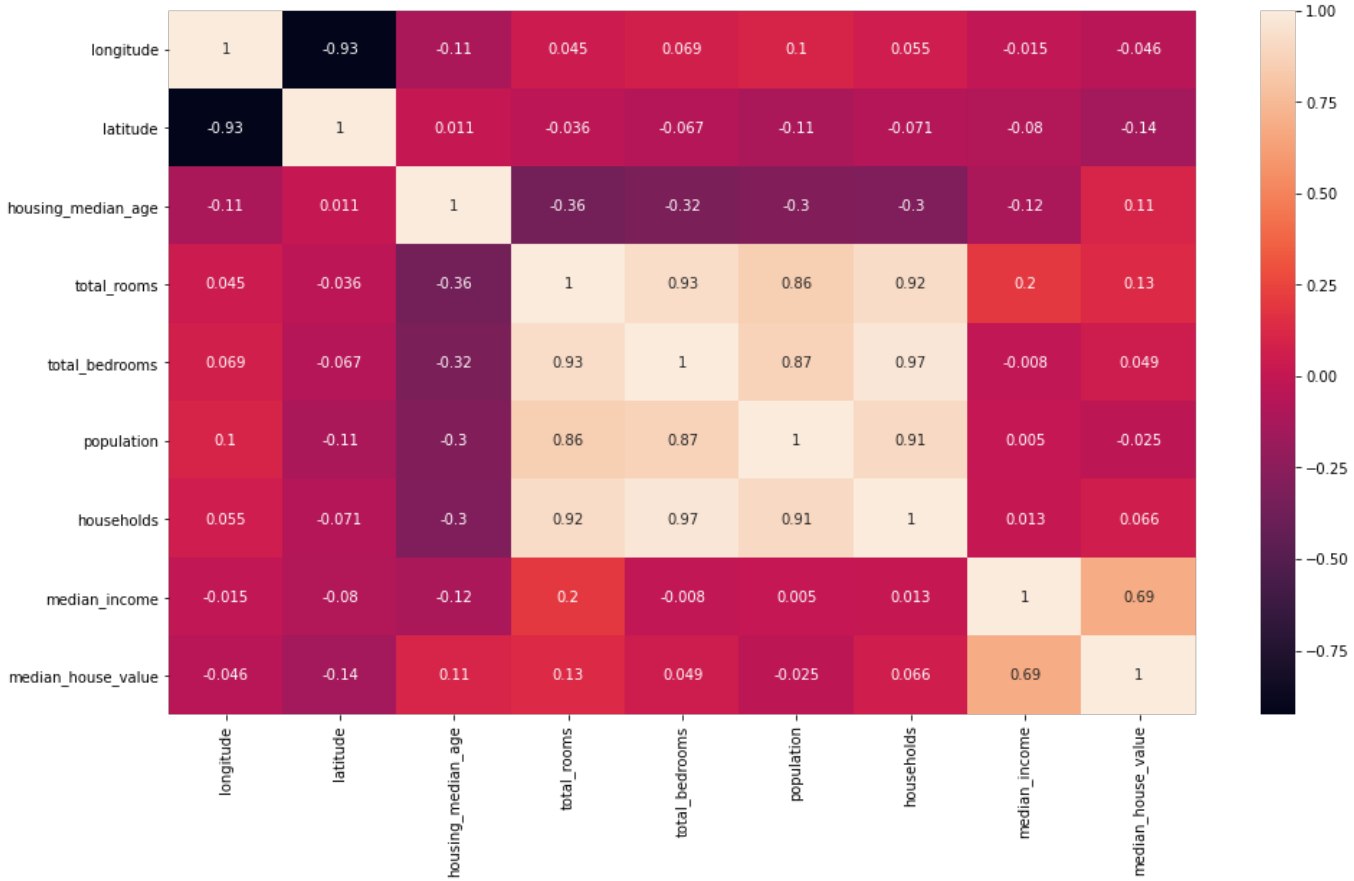

Figure 10: correlation matrix after dropping categorical data

Observations:

- *median_income* has the strongest correlation with target variable
- Most input features still have a very low correlation with *median_house_value*

4. EXPERIMENT LOG

- Using One-Hot Encoding for the Categorical Data - Input features were considered in non-increasing order of correlation with the target variable

#	Input Features	R^2 Score	RMSE	Conclusion
1.	median_income	0.48490837623606464	84300.87153316838	We include the median income feature in our model since it has the strongest correlation with the target variable and the highest impact on the r2 score
2.	median_income, ocean_proximity_INLAND	0.5925004537583891	74981.37245917703	We see that including Inland ocean proximity significantly increases our r2 score so we will include it in our model
3.	median_income, ocean_proximity_INLAND, ocean proximity_<1HOCEAN	0.595539886846771	74701.21580359849	We will not include the <1H ocean proximity feature in our model since it does not significantly increase the r2 score
4.	median_income, ocean_proximity_INLAND, ocean_proximity_NEAR BAY	0.5946131539657431	74786.74767777376	We will not include the Near Bay ocean proximity feature in our model since it does not significantly increase the r2 score
5.	median_income, ocean_proximity_INLAND, ocean_proximity_NEAR OCEAN	0.5929226813317592	74942.51670050713	We will not include the Near Ocean ocean proximity feature in our model since it does not significantly increase the r2 score

6.	median_income, ocean_proximity_INLAND, latitude	0.593065868461362	74929.33523945515	We will not include the latitude feature in our model since it does not significantly increase the r2 score
7.	median_income, ocean_proximity_INLAND, total_rooms	0.5930985794882815	74926.32361598866	We will not include the total rooms feature in our model since it does not significantly increase the r2 score
8.	median_income, ocean_proximity_INLAND, housing_median_age	0.604551142202245	73864.3675279136	The housing median age feature increases the r2 score significantly when compared to the impact on the score seen in rows 3 – 7. We will include it in our model
9.	median_income, ocean_proximity_INLAND, housing_median_age, households	0.6109294282230083	73266.25823553838	The households variable increases the r2 score significantly when compared to the impact on the score seen in rows 3 – 7. We will include it in our model
10.	median_income, ocean_proximity_INLAND, housing_median_age, households total_bedrooms	0.613833100120343	72992.34958542146	Total bedrooms does not significantly increase the r2 score, so we will not consider it in our model
11.	median_income, ocean_proximity_INLAND, housing_median_age, households, longitude	0.6124369997718668	73124.17433199659	The feature longitude does not significantly increase the r2 score, so we will not include it in our model

12.	median_income, ocean_proximity_INLAND, housing_median_age, households, population	0.637829171323185	70688.13997131982	Including population results in a significant increase to the r2 score. Here, the impact is significant relative to the increase seen in rows 9, 10, and 11
13.	median_income, ocean_proximity_INLAND, housing_median_age, households, population, ocean_proximity_ISLAND	0.6378031549443386	73268.66900443453	Island ocean proximity has a smaller impact on the r2 score than population, so we will not include it. We can conclude that the previous experiment represents the combination of features that produces the highest r2 score and the least root mean square error, with the exceptional case of considering all features in the input data
14.	median_income, ocean_proximity_INLAND, ocean_proximity_ISLAND, ocean_proximity_NEAR BAY, ocean_proximity_NEAR OCEAN, housing_median_age, households, population, ocean_proximity_<1H OCEAN, latitude, longitude, total_bedrooms, total_rooms	0.6508315183818004	69407.64961434438	Including all features in the model results in the highest r2 score and the lowest root mean square error

- Dropping Categorical Data - Input features were considered in non-increasing order of correlation with the target variable

#	Input Features	R^2 Score	RMSE	Conclusion
15.	median_income	0.48490837623606464	84300.87153316838	We include the median income feature in our model since it has the strongest correlation with the target variable and the highest impact on the r2 score
16.	median_income, latitude	0.4923887092703556	83686.51014347991	Including latitude does not significantly increase the r2 score. We will not consider it in our model
17.	median_income, total_rooms	0.48491789379158734	84300.09269898168	Total rooms has a negligible impact on the r2 score; we will not include it in our model
18.	median_income, housing_median_age	0.5244593831562463	80999.74080097071	Housing median age increases the r2 score significantly when compared to the increase in rows 2 and 3. We will include it in our model
19.	median_income, housing_median_age, households	0.5362814644543781	79986.56537552178	We will include households in our model since it increases the r2 score
20.	median_income, housing_median_age, households, total_bedrooms	0.537222925385503	79905.3280722402	Total bedrooms does not have a notable impact on the r2 score. We will not consider it in our model

21.	median_income, housing_median_age, households, longitude	0.5365637359988217	79962.21723486198	Longitude has a negligible impact on the r2 score; we will not include it in our model
22.	median_income, housing_median_age, households, population	0.563974653098303	77561.40880477776	Adding population results in a notable increase in the r2 score. This is the combination of features that gives us the highest r2 score, excluding the case when all features are considered
23.	median_income, latitude, housing_median_age, households, longitude, population, total_bedrooms, total_rooms	0.642320753449664	70248.441197765	Including all features in the model results in the highest r2 score, and a significantly lower root mean square error

5. RESULTS

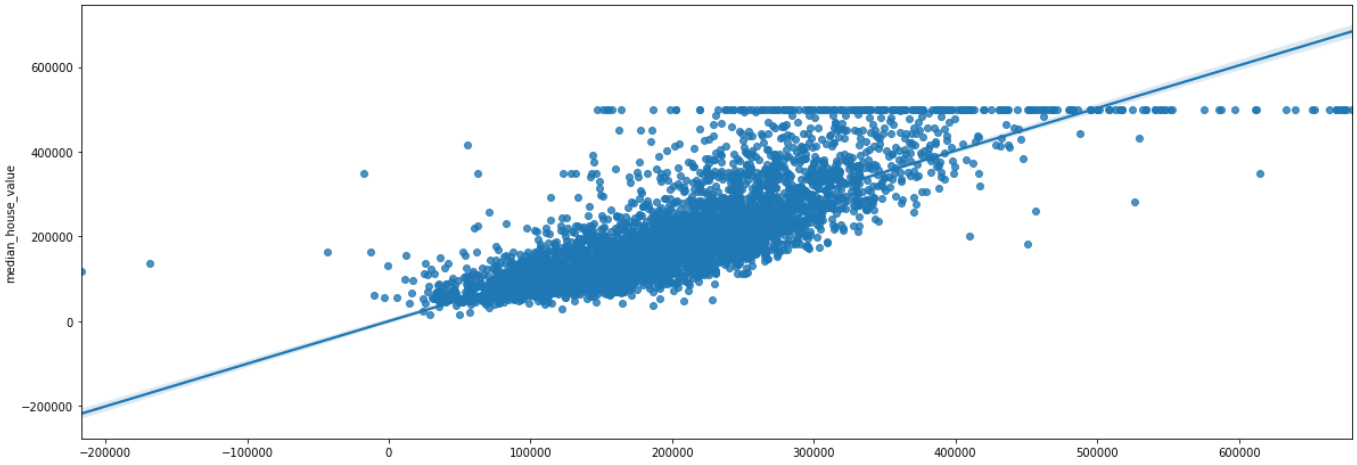

Figure 11: regression plot of predicted data vs. test data for experiment 12

Features considered :

- median_income
- ocean_proximity_INLAND
- housing_median_age
- households
- population

R^2 Score : 0.637829171323185
Coefficients : [3.81744453e+04, -7.31616500e+04, 1.14727411e+03, 1.39048180e+02, -4.21140366e+01]
Intercept : 39729.87804063852
RMSE : 70688.13997131982

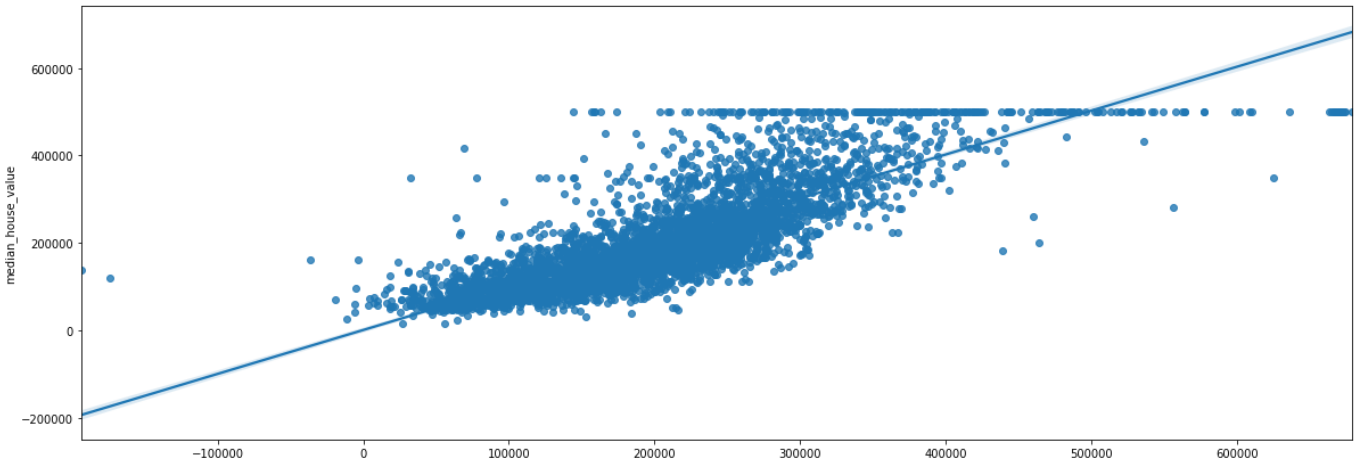

Figure 12: regression plot of predicted data vs. test data for experiment 14

Features considered :

- median_income
- ocean_proximity_INLAND
- housing_median_age
- households
- population
- ocean_proximity_ISLAND
- ocean_proximity_NEAR BAY
- ocean_proximity_NEAR OCEAN
- ocean_proximity_<1H OCEAN
- latitude
- longitude
- total_bedrooms
- total_rooms

R^2 Score : 0.6508315183818004
Coefficients : [3.89138790e+04, -6.36654267e+04, 1.33701996e+05, -2.78015484e+04, -1.80818126e+04, 1.03800988e+03, 8.20780051e+01, -3.91192124e+01, -2.41532079e+04, -2.48330321e+04, -2.61550697e+04, 7.13095793e+01, -5.47147673e+00]
Intercept : -2189180.7784567396
RMSE : 69407.64961434438

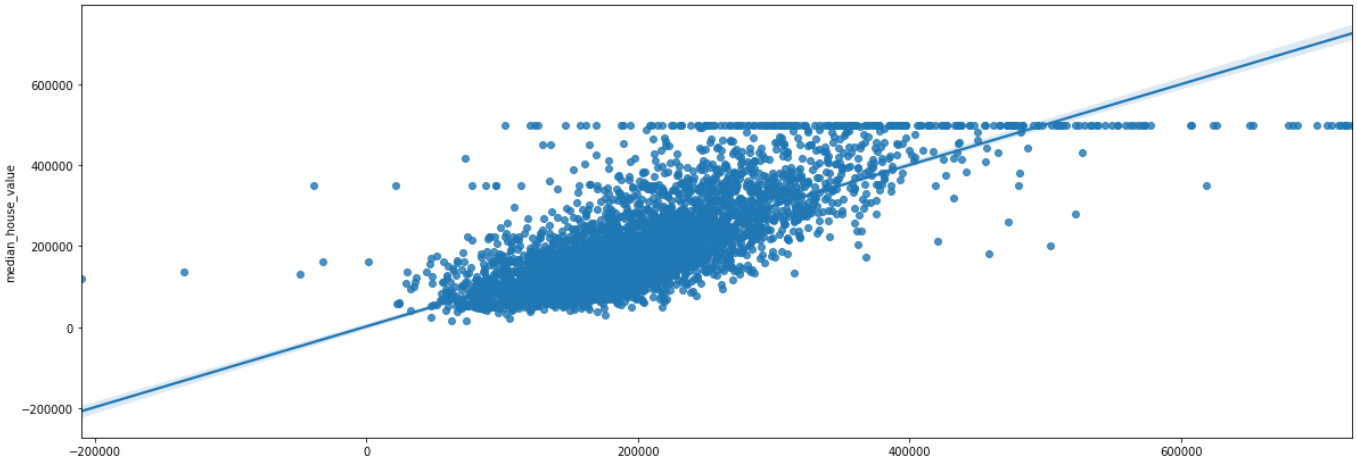

Figure 13: regression plot of predicted data vs. test data for experiment 22

Features considered :
- median_income
- housing_median_age
- households
- population

R^2 Score : 0.563974653098303
Coefficients : [43095.58522718, 1977.47012148, 154.44441591, -43.3416565]
Intercept : -32203.65413126463

RMSE : 77561.40880477776

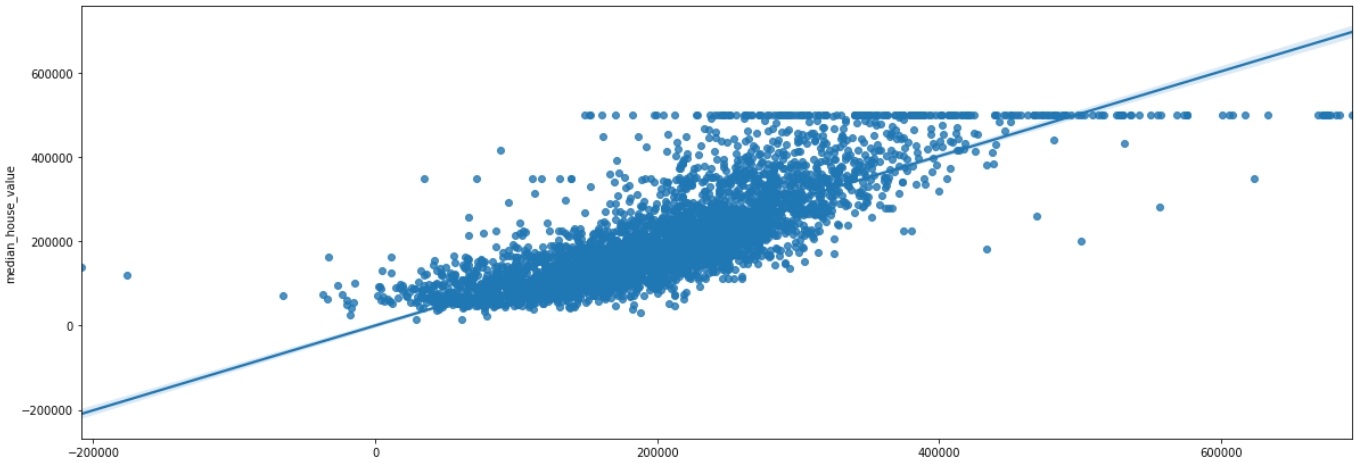

Figure 14: regression plot of predicted data vs. test data for experiment 23

Features considered :

- median_income
- housing_median_age
- households
- population
- latitude
- longitude
- total_bedrooms
- total_rooms

R^2 Score : 0.642320753449664
Coefficients : [3.98585379e+04, -4.21450042e+04, 1.12523498e+03, 8.44778892e+01, -4.23129358e+04, -3.98200093e+01, 8.03007026e+01, -7.25820820e+00]
Intercept : -3547042.6243745657
RMSE : 70248.441197765

---

6. INTERPRETATION OF RESULTS

In experiment 12, we considered 5 input features. The resulting model achieved an R^2 Score of 0.63, and a Root Mean Squared Error of 70688.13

- This means that the model explained 63% of the variance in the *median_house_value*
- The RMSE indicates how close the actual data is to the model's predicted values. It is the standard deviation of the unexplained variance
- This model had an RMSE of 70688.13. This means that $70,688 was the square root of the average of squared differences between the predicted data and the observed data
- From the coefficients and the intercept, we can construct the regression equation :

Y = 39729.87804063852 + 3.81744453e+04 (median_income) - 7.31616500e+04 (ocean_proximity_INLAND) + 1.14727411e+03 (housing_median_age) + 1.39048180e+02 (households) - 4.21140366e+01 (population)

In experiment 14, we considered all 13 input features. The resulting model achieved an R^2 Score of 0.65, and a Root Mean Squared Error of 69407.64

- This means that the model explained 65% of the variance in the *median_house_value*
- The RMSE indicates how close the actual data is to the model's predicted values. It is the standard deviation of the unexplained variance
- This model had an RMSE of 69407.64. This means that $69,407 was the square root of the average of squared differences between the predicted data and the observed data
- From the coefficients and the intercept, we can construct the regression equation :

Y = -2189180.7784567396 + 3.89138790e+04 (median_income) - 6.36654267e+04 (ocean_proximity_INLAND) + 1.33701996e+05 (ocean_proximity_ISLAND) - 2.78015484e+04 (ocean_proximity_NEAR BAY) - 1.80818126e+04 (ocean_proximity_NEAR OCEAN) + 1.03800988e+03 (housing_median_age) + 8.20780051e+01 (households) - 3.91192124e+01 (population) - 2.41532079e+04 (ocean_proximity_<1H OCEAN)- 2.48330321e+04 (latitude) - 2.61550697e+04 (longitude) + 7.13095793e+01 (total_bedrooms) - 5.47147673e+00 (total_rooms)

In experiment 22, we considered all 4 input features. The resulting model achieved an R^2 Score of 0.56, and a Root Mean Squared Error of 77561.40

- This means that the model explained 56% of the variance in the *median_house_value*
- The RMSE indicates how close the actual data is to the model's predicted values. It is the standard deviation of the unexplained variance
- This model had an RMSE of 77561.4. This means that $77,561 was the square root of the average of squared differences between the predicted data and the observed data
- From the coefficients and the intercept, we can construct the regression equation :

Y = -32203.65413126463 + 43095.58522718 (median_income) + 1977.47012148 (housing_median_age) + 154.44441591 (households) - 43.3416565 (population)

In experiment 23, we considered all 8 input features after dropping *ocean_proximity*. The resulting model achieved an R^2 Score of 0.64, and a Root Mean Squared Error of 70248.44

- This means that the model explained 56% of the variance in the *median_house_value*
- The RMSE indicates how close the actual data is to the model's predicted values. It is the standard deviation of the unexplained variance
- This model had an RMSE of 70248.44. This means that $70,248 was the square root of the average of squared differences between the predicted data and the observed data
- From the coefficients and the intercept, we can construct the regression equation :

Y = -3547042.6243745657 + 3.98585379e+04 (median_income) - 4.21450042e+04 (latitude) + 1.12523498e+03 (housing_median_age) + 8.44778892e+01 (households) - 4.23129358e+04 (longitude) - 3.98585379e+04 (population) + 8.03007026e+01 (total_bedrooms) - 7.25820820e+001 (total_rooms)

7. CONCLUSION

Based on the results, we concluded that the model with the R^2 score of 0.56 from experiment 22 was not satisfactory. While the models created in experiment 12, experiment 14, and experiment 23 achieved similar results, we selected the linear regression model that was produced in experiment 14, when we performed one-hot encoding to convert *ocean_proximity* into continuous data. This model includes all 13 of the input features and obtains the best results. The model explains 65% of the variance in the target variable. $69,407 is the square root of the average of squared differences between the predicted data and the observed data in this experiment. Taken in context of the mean of *median_house_value* - $206,855.81 – this error is acceptable.

In addition, the following observations were made while performing linear regression :

- The only input feature that is roughly linear with the target variable is *median_income*
- median_income alone explains 48% of the variance in the target variable
- *ocean_proximity_INLAND*, when *ocean_proximity* is converted into continuous data, has a strong correlation with *median_house_value*, and increases the percentage variance explained from 48% to 59%
- When we drop *ocean_proximity* from the data frame however, we still obtain a satisfactory R^2 score of 0.64
- The model can explain at most 65% of the variance in the *median_house_value*,, even when all 13 input features are considered. Out of this 65%, the bulk of the variance is explained by the 2 input features - *median_income*, and *ocean_proximity_INLAND*