CS 6375

ASSIGNMENT 1

Linear regression Analysis

Names of students in your group:

Ankita Gonnade (SMC200017)

Laxmi Niharika Epuri(LXE210008)

Number of free late days used: ____0____

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

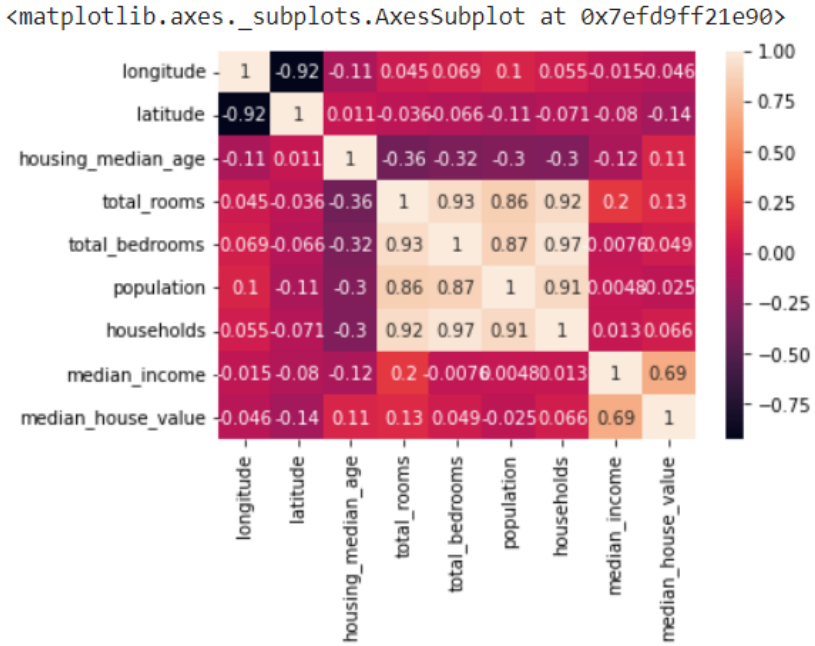
Please list clearly all the sources/references that you have used in this assignment.

The following are the references used in the assignment:

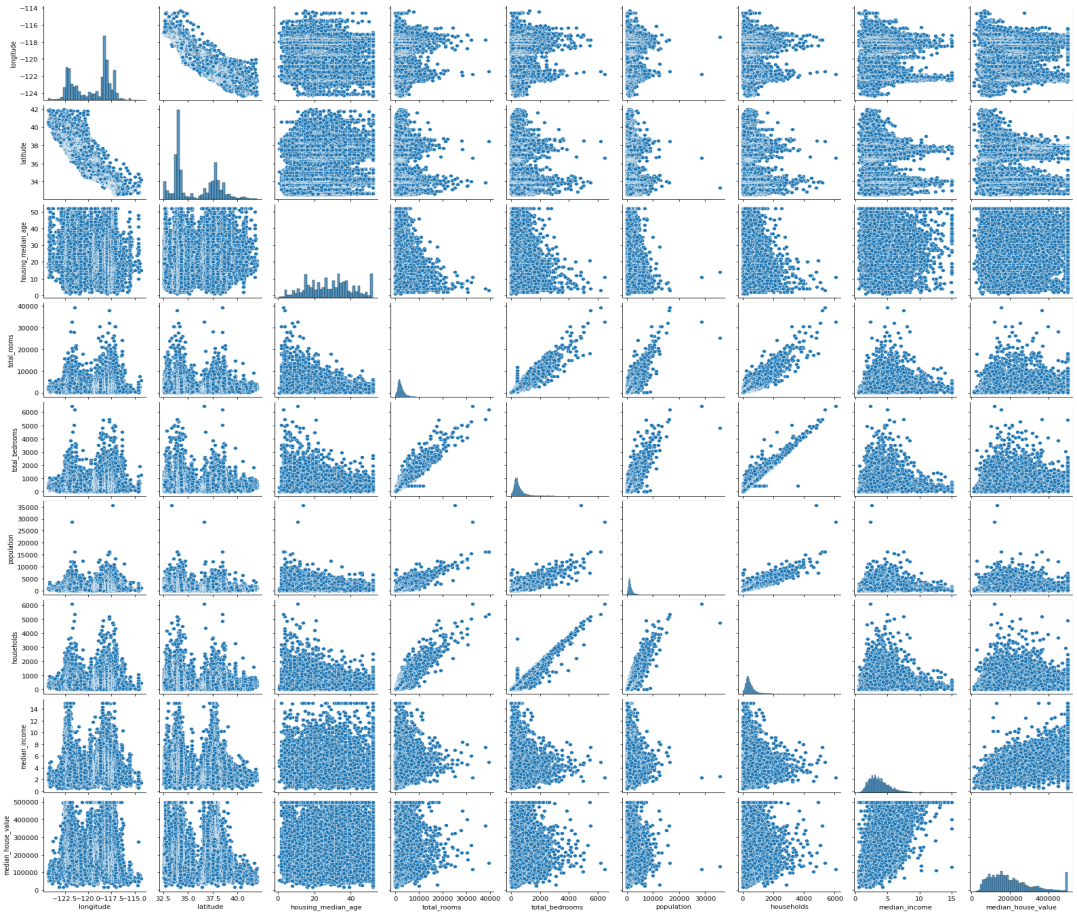
1. https://medium.com/@nikhilparmar9/simple-sgd-implementation-in-python-for-linear-regression-on-boston-housing-data-f63fcaaecfb1
2. https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc
3. https://www.statology.org/rmse-vs-r-squared/
4. https://www.kaggle.com/code/mennatallahnasr/housing-prices/notebook

In your report, you should mention details of the dataset, what you are trying to predict, plots and figures that show data distribution, and correlation. You should also output your results and analysis of the results. Do not put any code in your report

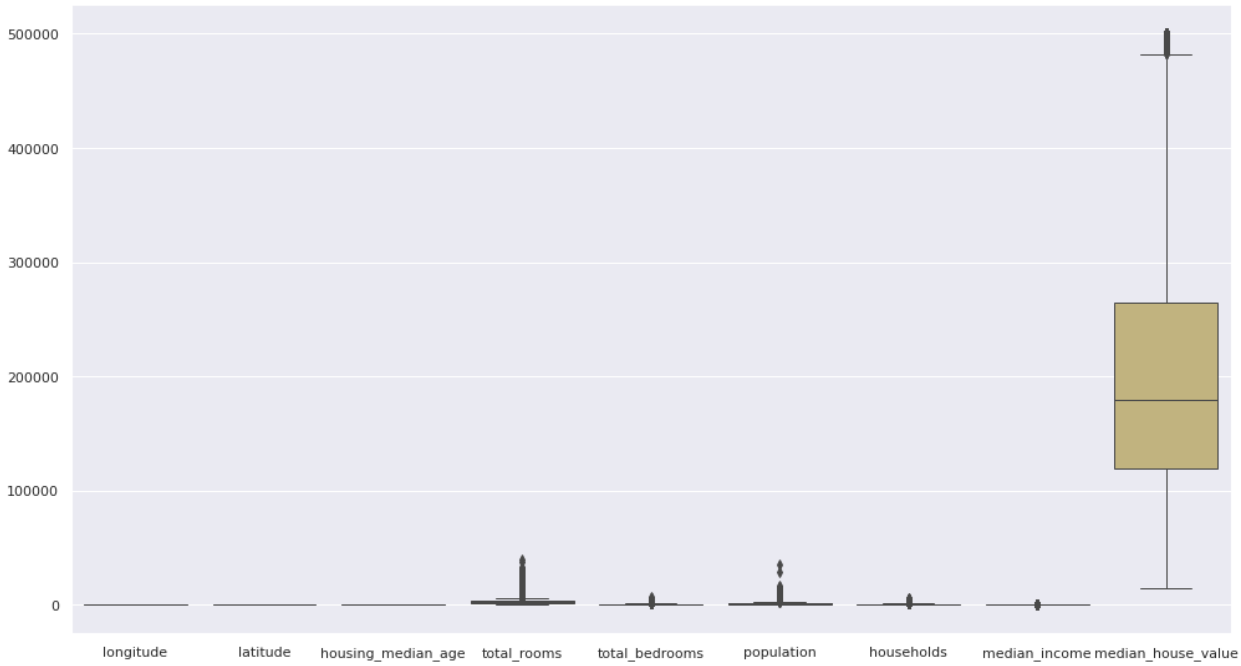
1. Dataset used:
   • California Housing Dataset
   https://www.kaggle.com/camnugent/california-housing-prices
   a. We have hosted the dataset on github and accessed it in google colab notebook
2. We are trying to predict the ‘median house value’ with respect to housing_median_age', 'total_rooms', 'total_bedrooms','median_income','population' and 'households’.
3. Plots/Figures
   a. HeatMap to get the correlation between variables:



   b. Pairplot:


longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value

c. Boxplot


500000
400000
300000
200000
100000
0
longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value

4. Observations/Conclusions:

We have split the data to training and testing dataset in 80:20 ratio

Ordinary Least Squares Observations:

```
The model performance for training set
--------------------------------------
RMSE is 65134.50509249206
R2 score is 0.6426309045026166


The model performance for testing set
--------------------------------------
RMSE is 69415.46977283146
R2 score is 0.6141381195464177
```

We found that Root Mean Square error for testing dataset is more than the training dataset
Also, R-Squared score for training dataset is better than testing dataset

SGD Regression Observations:

#	Alpha	Max iteration	Training		Testing	
			RMSE	R2	RMSE	R2
1.	0.1	2000	69774.58680475531	0.5899004390025446	71739.6968252492	0.5878660265381388
2.		3000	69775.03190070912	0.5898952068902295	71737.97304438193	0.5878858320336859
3.		4000	69775.72841437193	0.5898870192904679	71739.08300974098	0.5878730790670265
4.	0.15	2000	70365.08711479438	0.5829297454543911	72289.93084247624	0.5815197554933556
5.		3000	70363.41273646957	0.582949594078592	72288.56934370565	0.5815355185454536
6.		4000	70361.11378787321	0.5829768457920947	72286.07906845697	0.5815643494878977
7.	0.2	2000	70825.3000345753	0.5774563264793379	72736.54472361977	0.5763329632549491
8.		3000	70828.15341152326	0.5774222793051271	72739.06578323735	0.5763035940191192
9.		4000	70831.9374181971	0.577377125519444	72744.53143604832	0.5762399180719018

As the max iteration value and alpha value increases, Root Mean Square error also increases and R-Squared score decreases.

We found that Root Mean Square error for testing dataset is more than the training dataset
Also, R-Squared score for training dataset is better than testing dataset

Based on the RMSE and R^2 values of SGD, from the table its identified that alpha = 0.1, Max_iterations = 2000 show better results.

RMSE Vs R2
In order to evaluate how well the model will perform in making predictions for future observations—that is, calculating accuracy on an unknown observation—we should calculate the RMSE on a test set.

R-squared measures how much of the variance in your training set is explained by how well your model fits.

SGD Regression Observations Vs Ordinary Least Squares Observations

Keeping the below point into consideration:

1. The higher the R2 value, the better a model fits a dataset.
   The lower the RMSE, the better a model fits a dataset.

Observations:

Root mean square error:
Training:
SGD Regression values for training dataset is higher than the Ordinary Least Squares
Testing:
SGD Regression values for testing dataset is higher than the Ordinary Least Squares

R-squared:
Training:
SGD Regression values for training dataset is lower than the Ordinary Least Squares
Testing:
SGD Regression values for testing dataset is lower than the Ordinary Least Squares

Conclusion:

SGD is dependent on parameter tuning, and based on our observations over a few hyperparameters, we found OLS is better than SGD regression with the dataset we used.

However, in general SGD is better as we can tune a lot of hyperparameters and it is used in gradient descent.