Homework 2 (Linear Regression)

Name: Nasir Khan

Student Number:0075244

To perform the regression analysis, we utilize the California housing data set which is available on Kaggle (https://www.kaggle.com/camnugent/california-housing-prices). The original data set has 10 columns(9 features can be utilized). Moreover, for the purpose of this homework, we retrieve only 1000 samples out of 17000 samples in the original data set and Write the resulting data set to a csv file.

We choose the total number of rooms as an independent variable and the outcome of interest in our case (dependent variable) is the median house price.

First, we define the linear regression function which reads the input dataset. The original data The defined function ***linear_regress_california*** checks for presence and then deletion of NaN values.

The first step for performing the regression is the data preprocessing so we standardize our data to make the regression analysis convenient and to ensure that the transformation does not affect the results and no overflow occurs during the regression analysis. We normalize and scale our inputs and outcomes to perform the standardization of the data.

Next, we utilize the *scipy.stats* library and utilize the *t-statistics* module to find the 95% confidence interval for our data.

We then perform the regression analysis ad estimate the residual error (actual- predicted) and the standard error for our outcome. Our function returns the regression estimates, their standard error, and 0.95 credible intervals along with the residual errors.

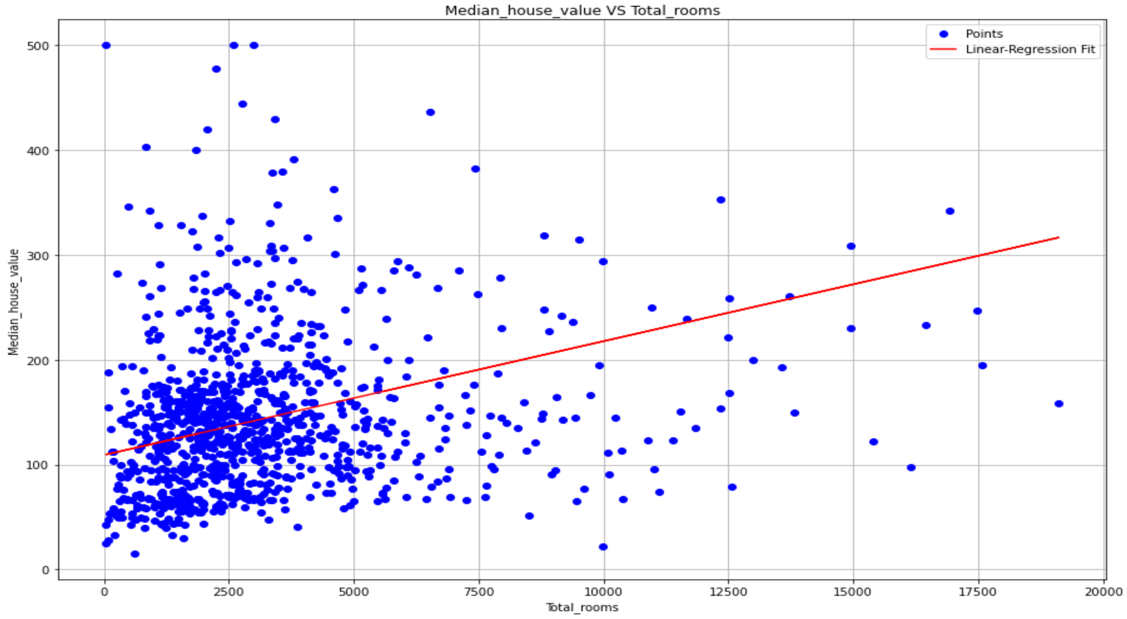
Before defining the null and alternate hypothesis for our analysis we pos ethe problem statement:

Is total room count a significant linear predictor of the house prices in California?

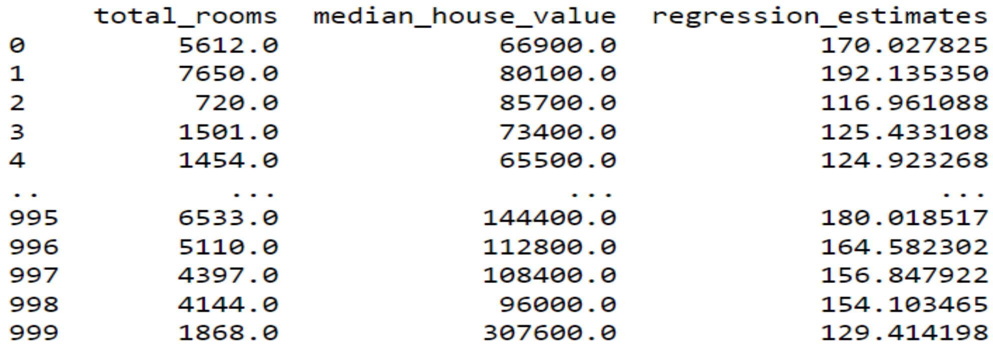
- Null hypothesis: There is no association between total room count and house prices in California
- Alternative hypothesis: There is an association between total room count and house prices in California.

We perform the linear regression by utilizing the gradient descent based approach we set the learning rate to 0.0001 and we choose 1000 iteration (epochs). This is enough as we have choosen a subset of data frame which converges easilty. The gradient descent method essentialy calculates the equations described in the homework pdf.

The outputs of the funtion include the regression estimates (cofficient values * the input) which are then used to plot the following figure:

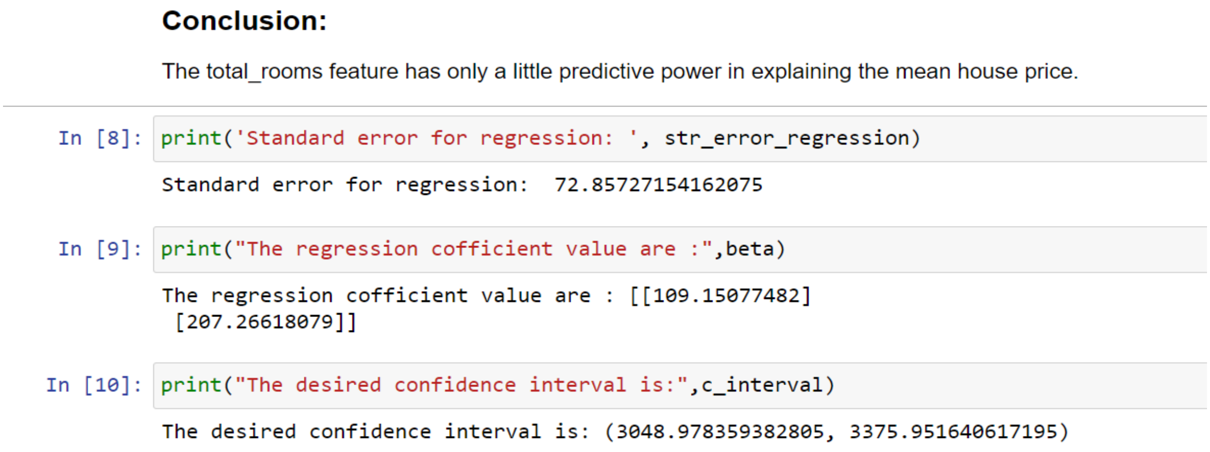


The desired output in tabular form is shown below:



```
     total_rooms  median_house_value  regression_estimates
0         5612.0             66900.0            170.027825
1         7650.0             80100.0            192.135350
2          720.0             85700.0            116.961088
3         1501.0             73400.0            125.433108
4         1454.0             65500.0            124.923268
..           ...                 ...                   ...
995       6533.0            144400.0            180.018517
996       5110.0            112800.0            164.582302
997       4397.0            108400.0            156.847922
998       4144.0             96000.0            154.103465
999       1868.0            307600.0            129.414198
```

We have a strong reason to accept the null hypothesis which is evident from the std error values. We could also use the p-value to validate this claim.



Conclusion:

The total_rooms feature has only a little predictive power in explaining the mean house price.

```
In [8]: print('Standard error for regression: ', str_error_regression)
```

```
Standard error for regression:  72.85727154162075
```

```
In [9]: print("The regression cofficient value are :",beta)
```

```
The regression cofficient value are : [[109.15077482]
 [207.26618079]]
```

```
In [10]: print("The desired confidence interval is:",c_interval)
```

```
The desired confidence interval is: (3048.978359382805, 3375.951640617195)
```