CSCI183 HW2 Report

Codes for all other steps are in the ipynb file.

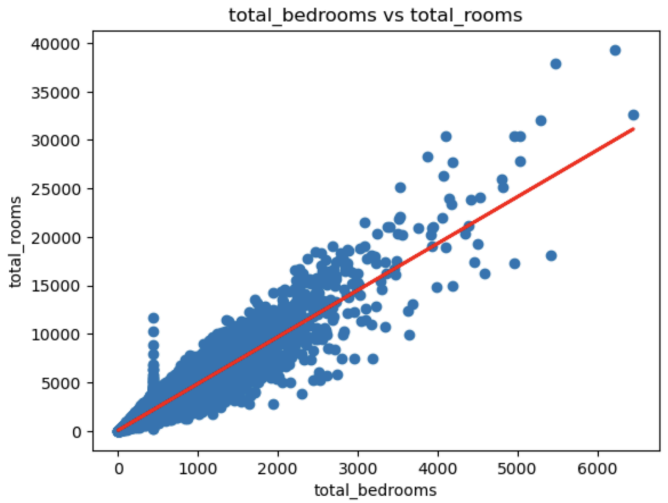
3.
there are 9 numerical features, and 1 categorical feature (ocean_proximity), assuming no target specified

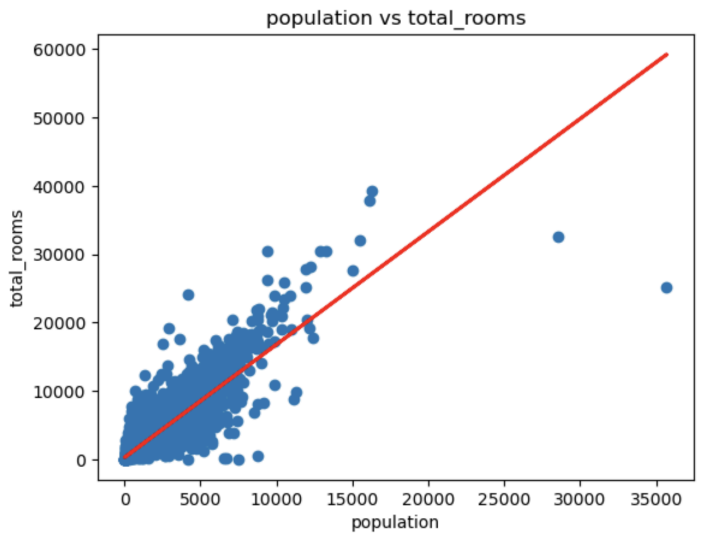
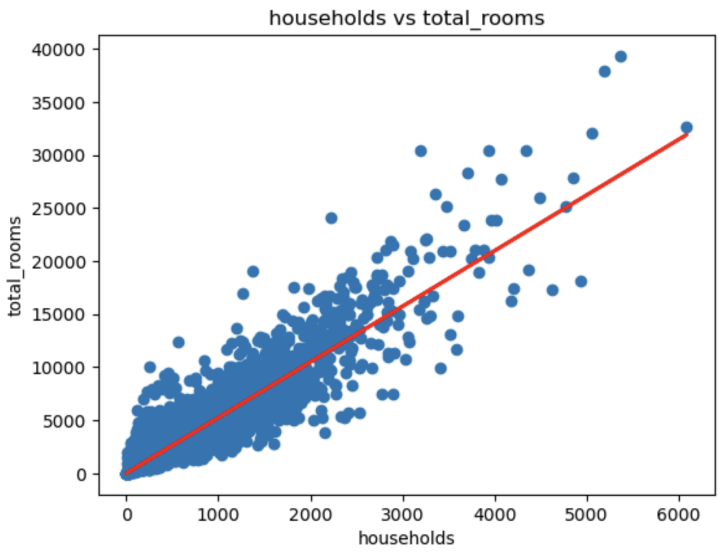
4.
I find high corr between the following pairs:
corr(longitude, latitude) = -0.924664 but I think this is because of the geographic location of california
corr(total_rooms, total_bedrooms) = 0.930380
corr(total_rooms, population) = 0.857126
corr(total_rooms, households) = 0.918484
corr(total_bedrooms, population) = 0.877747
corr(total_bedrooms, households) = 0.979728
corr(population, households) = 0.907222

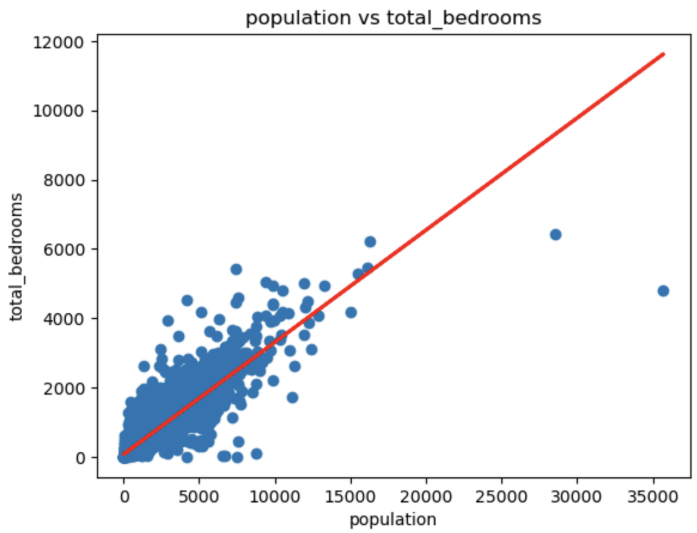
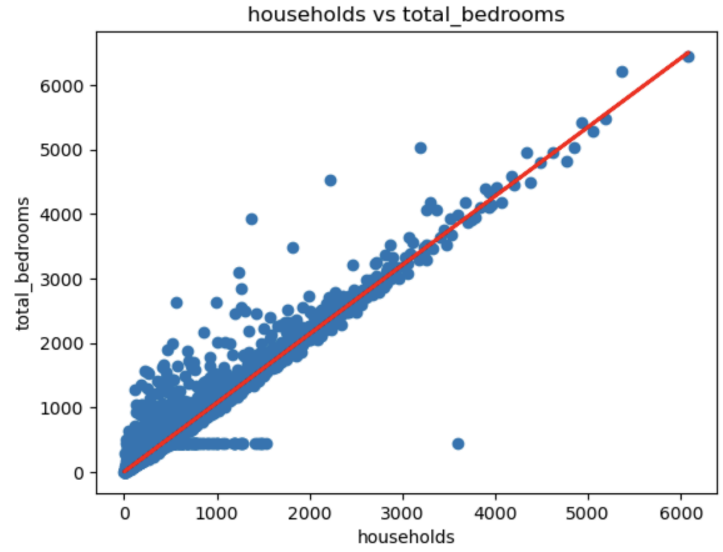
6. Based on the graphs in step 5, identify features that have a linear relationship with the target variable.

(total_rooms, total_bedrooms)
(total_rooms, population)
(total_rooms, households)
(total_bedrooms, population)
(total_bedrooms, households)
(population, households)
(total_bedrooms, total_rooms)
(population, total_rooms)
(households, total_rooms)
(population, total_bedrooms)
(households, total_bedrooms)
(households, population)

7-11.

Feature	Target	Slope/ Coefficient	Intercept	Graph
total_bedrooms	total_rooms	4.82242124	46.89997601810228	

population	total_rooms	1.65119792	282.01885268768046	
households	total_rooms	5.24097265	17.689280170347047	

population	total_bedrooms	0.3235017	75.69470115245974	
households	total_bedrooms	1.06881901	2.9213477383798363	

households	population	2.68719566	83.11588362697034