Homework 1

Chris Cirelli
09/08/2020

Contents

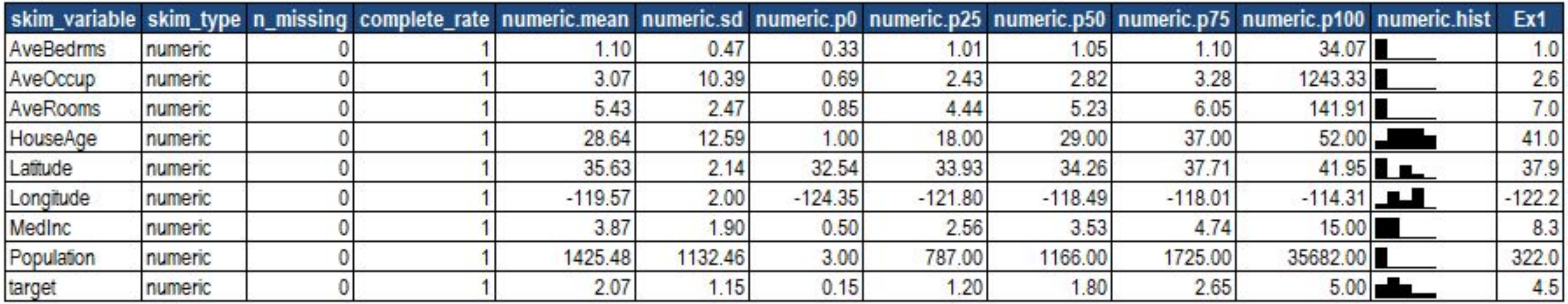
Dataset Description:

- California Housing Data
- Source : Scikit Learn
- “This dataset was derived from the 1990 U.S. census, using one row per census group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data”
- The target variable is the median house value for California districts

EDA

Observations:

- All data types are numeric | floats
- There are no null values.
- Avg Bedroom, Occup, Rooms, and Population are right skewed.
- Data is geocoded providing for the opportunity to compare the target and feature values spatially.



skim_variable	skim_type	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist	Ex1
AveBedrms	numeric	0	1	1.10	0.47	0.33	1.01	1.05	1.10	34.07		1.0
AveOccup	numeric	0	1	3.07	10.39	0.69	2.43	2.82	3.28	1243.33		2.6
AveRooms	numeric	0	1	5.43	2.47	0.85	4.44	5.23	6.05	141.91		7.0
HouseAge	numeric	0	1	28.64	12.59	1.00	18.00	29.00	37.00	52.00		41.0
Latitude	numeric	0	1	35.63	2.14	32.54	33.93	34.26	37.71	41.95		37.9
Longitude	numeric	0	1	-119.57	2.00	-124.35	-121.80	-118.49	-118.01	-114.31		-122.2
MedInc	numeric	0	1	3.87	1.90	0.50	2.56	3.53	4.74	15.00		8.3
Population	numeric	0	1	1425.48	1132.46	3.00	787.00	1166.00	1725.00	35682.00		322.0
target	numeric	0	1	2.07	1.15	0.15	1.20	1.80	2.65	5.00		4.5

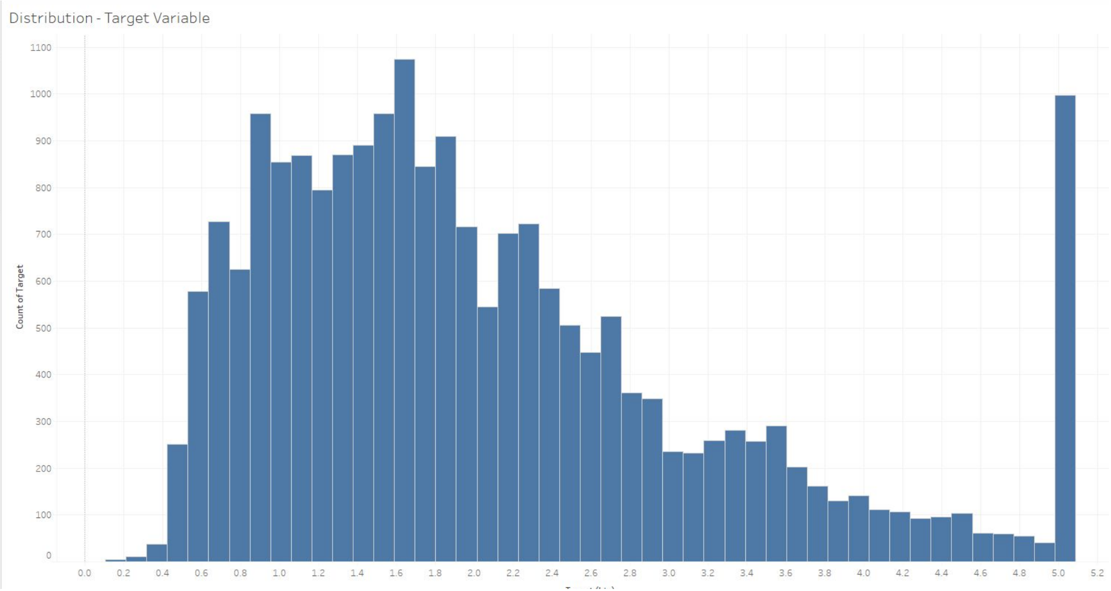
Target Variable
Distribution

Plot

- Distribution of median property value

Observations

- Right skewed distribution
- There appears to be some anomalies around bin 5 where the count of observations spikes



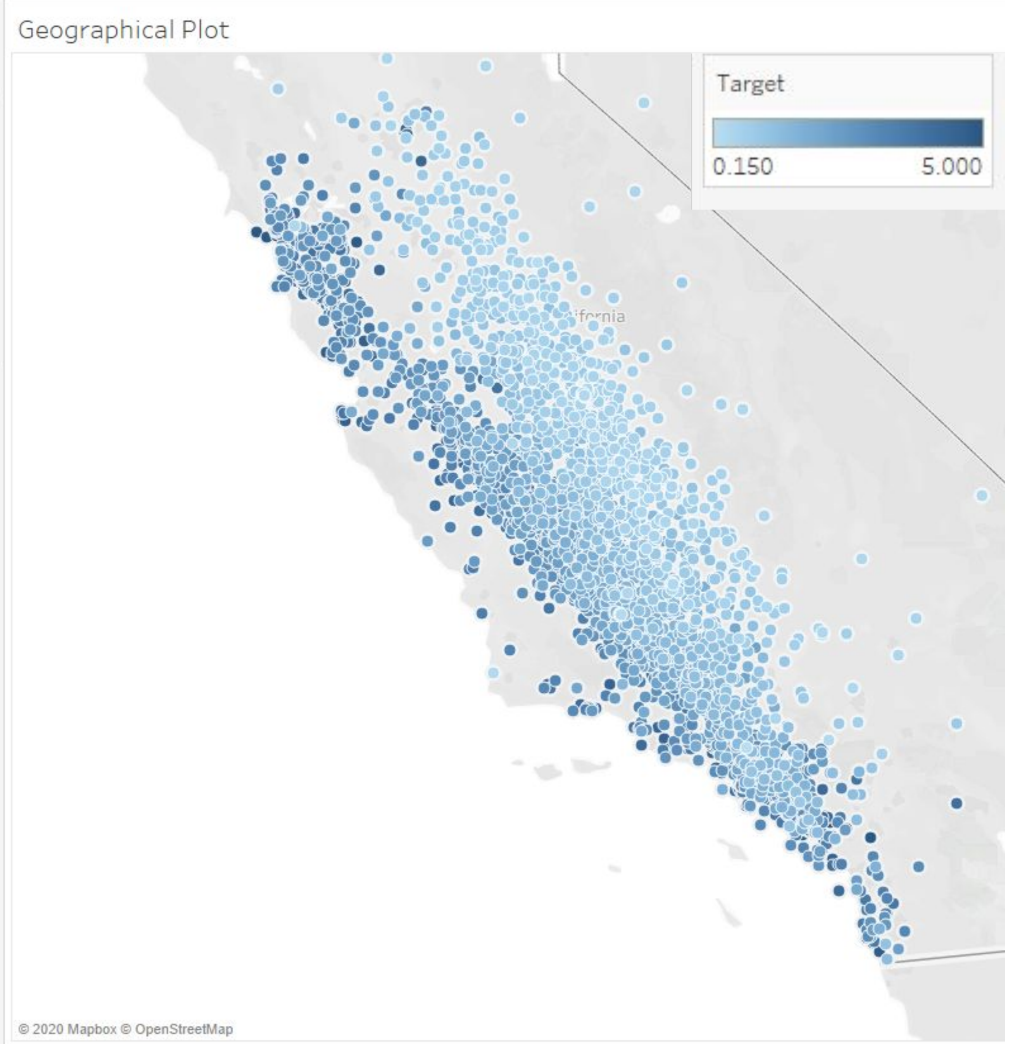
Target Variable
Geographic Plot

Plot

- Geographical plot of target value using lat lon coordinates.

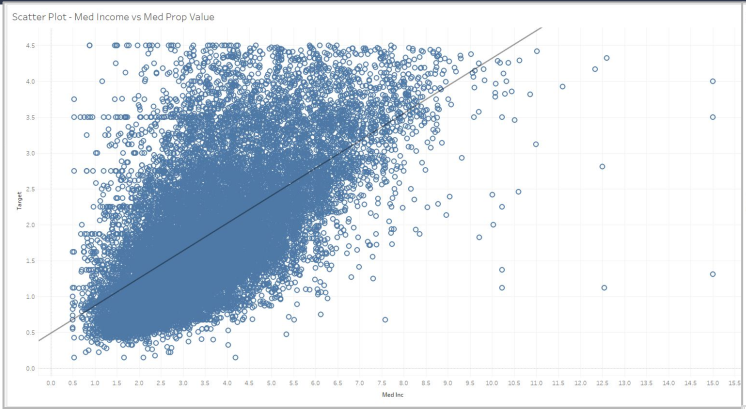
Observations

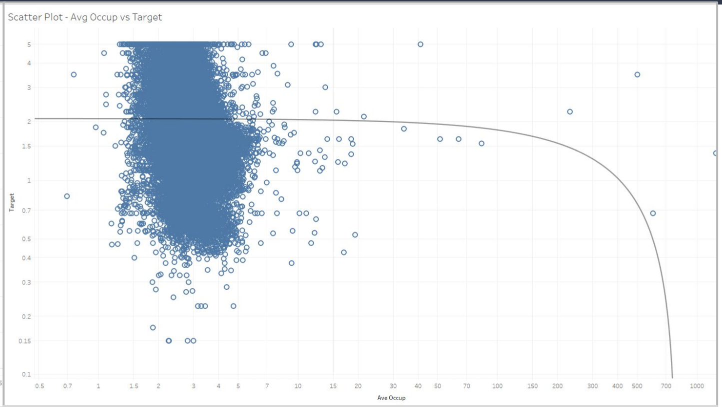
- Clearly there is a difference in median income as one gets closer to the coast, which is to be expected.

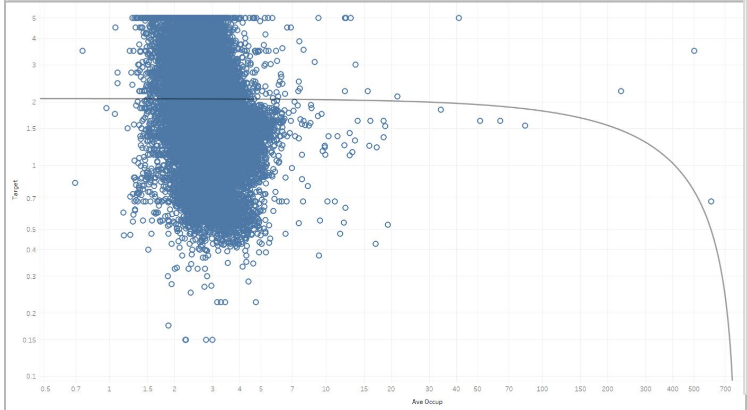


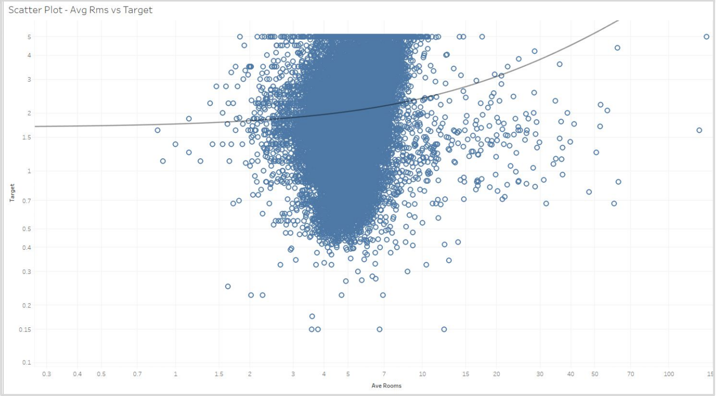
Scatter Plots

Dependent vs Independent Variables


Scatter Plot - Med Income vs Med Prop Value

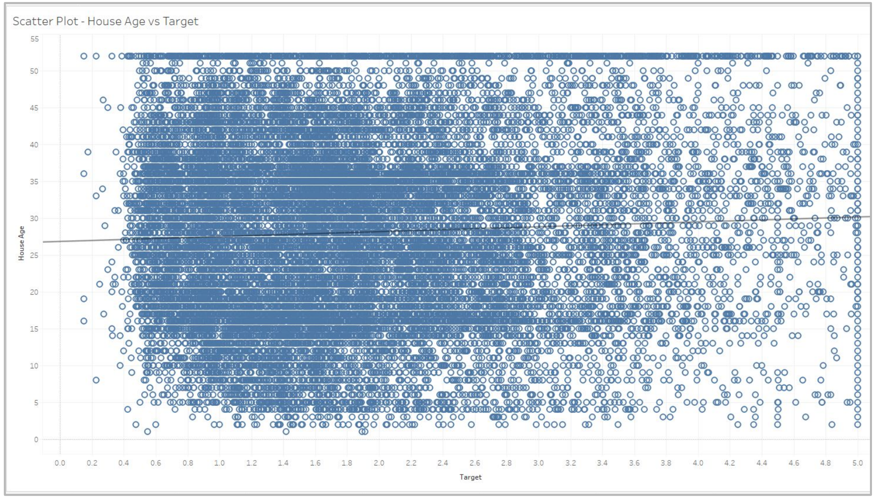

Scatter Plot - Avg Occup vs Target

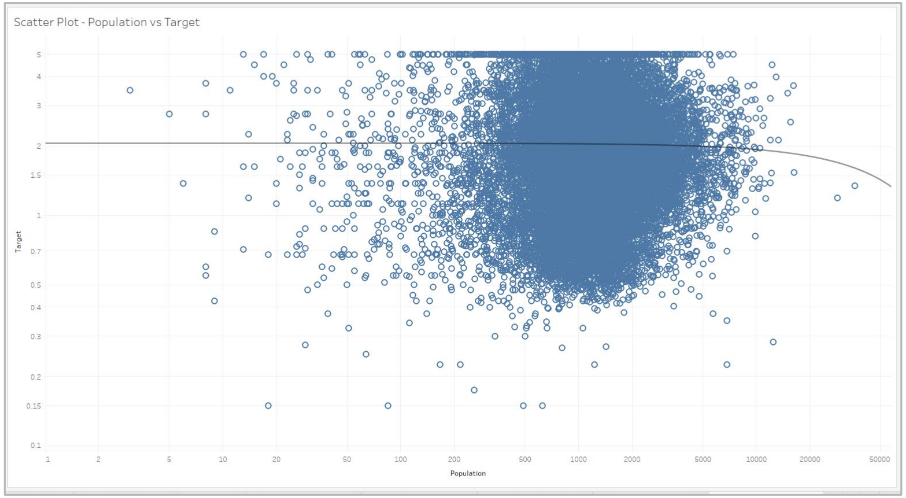



Scatter Plot - Avg Rms vs Target

Scatter Plots

Dependent vs Independent Variables


Scatter Plot - House Age vs Target


Scatter Plot - Population vs Target

Target Variable

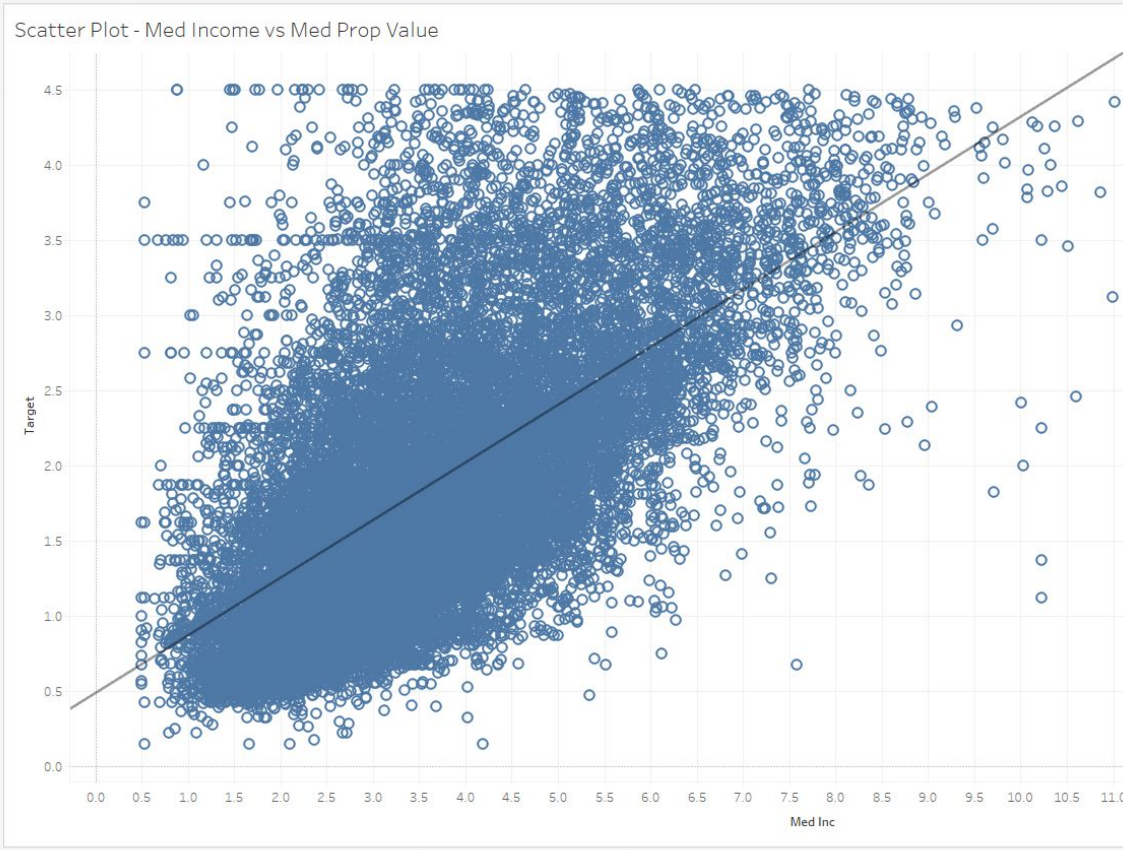
Scatter Plot

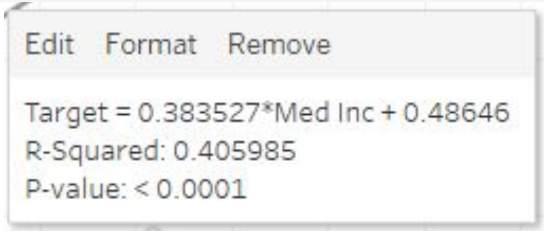
Plot

- Scatter plot comparing target variable versus median income.

Observations

- Based on R-squared and p-value there appears to be some level of linear relationships between median income and property value.





Target = 0.383527*Med Inc + 0.48646
R-Squared: 0.405985
P-value: < 0.0001