CALIFORNIA HOUSING

Final Project

Nicole Aguilera

DSC 530 – T303

November 21, 2020

STATISTICAL QUESTIONS

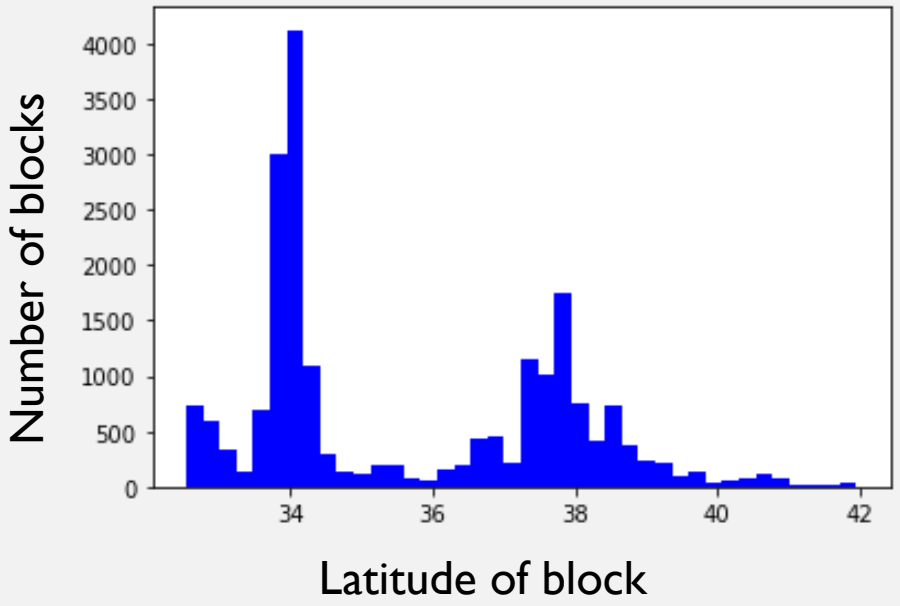
1. What variables drive the price of a house in the housing market?
2. What features of a house drive its price?
3. Do certain locations of houses attract a certain type of buyer? (i.e. income)
4. Do certain home features attract a certain type of buyer?

VARIABLES

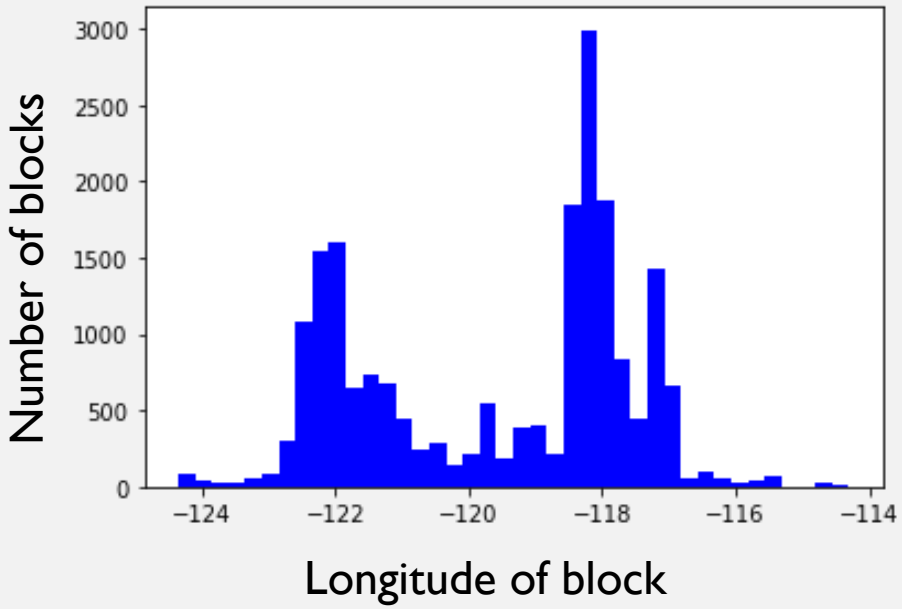
- Longitude - A measure of how far West a house is, geographically ; a larger negative number is West
- Latitude - A measure of how far North a house is, geographically; a higher value is North
- housing_median_age - Median age of a house within a block; a lower number is a newer building
- total_rooms - Total number of rooms within a block
- total_bedrooms - Total number of 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) within a block
- median_income - Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- median_house_value - Median house value for households within a block (measured in US Dollars)
- ocean_proximity – Location of house with relation to ocean

HISTOGRAMS

LATITUDE

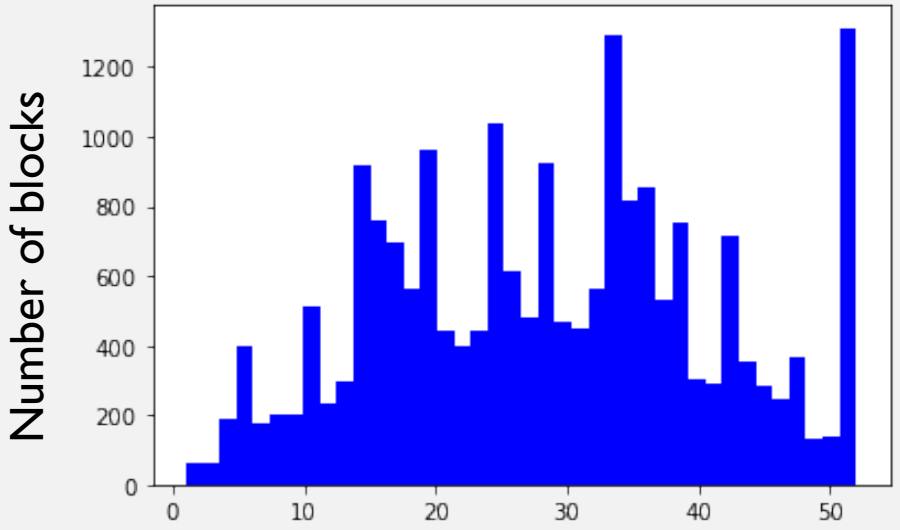


LONGITUDE



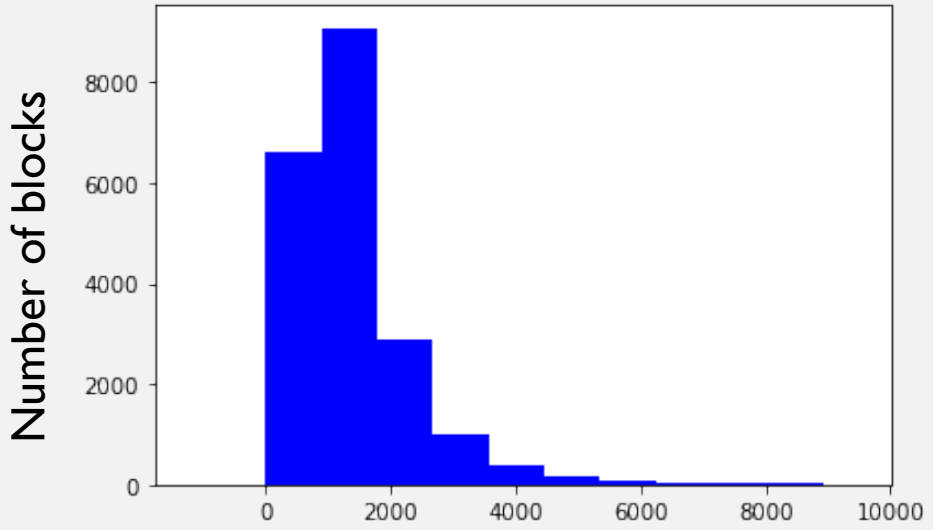
HISTOGRAMS

HOUSING MEDIAN AGE



Median age of house within a block
Lower number is a newer house

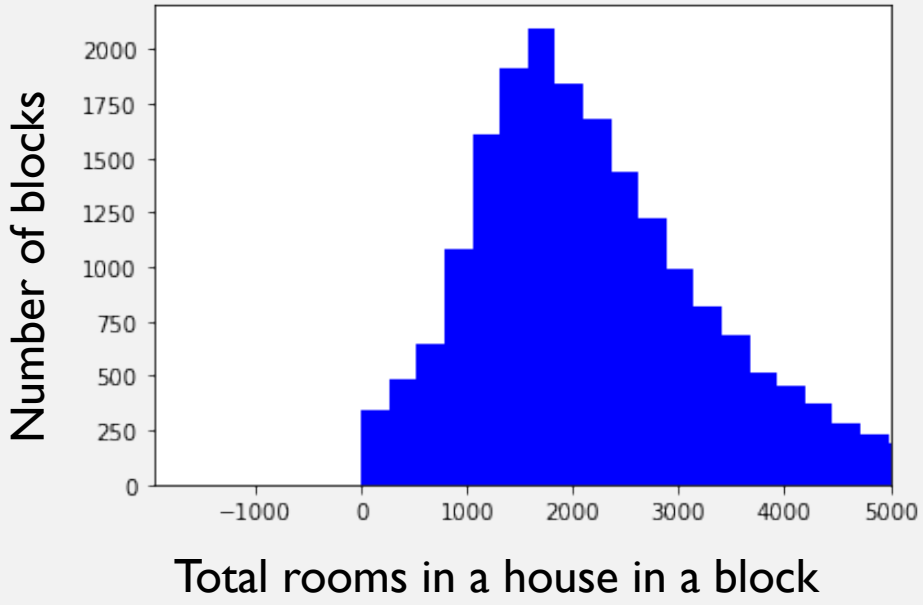
POPULATION



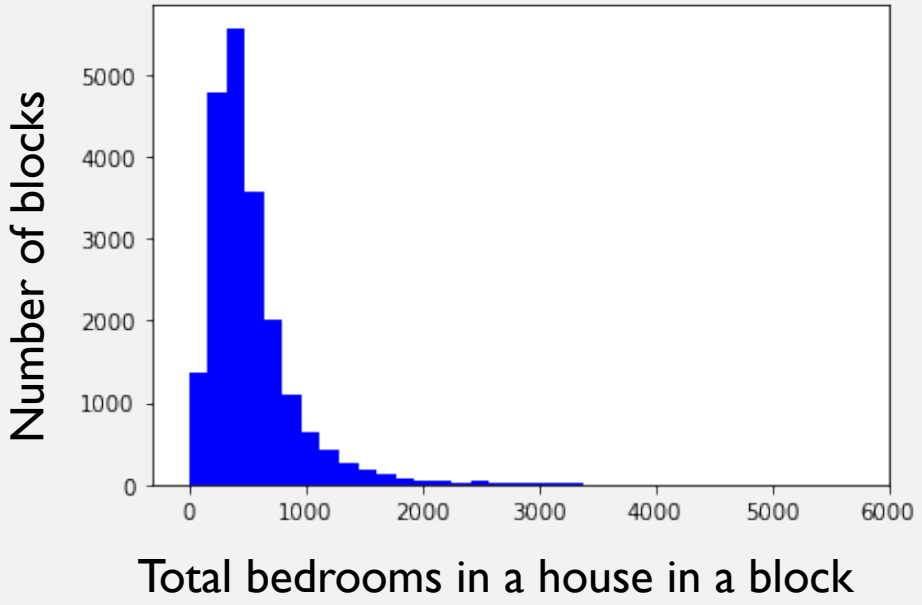
Number of people residing within a block

HISTOGRAMS

TOTAL ROOMS

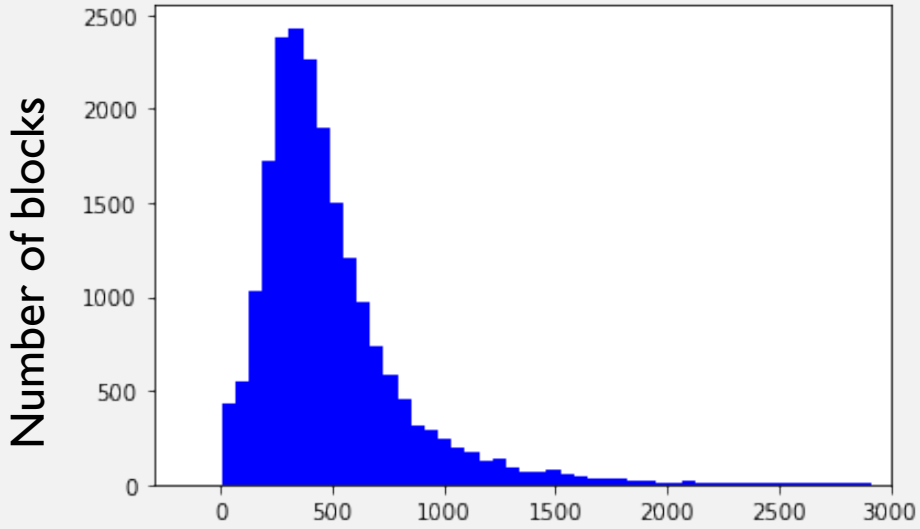


TOTAL BEDROOMS



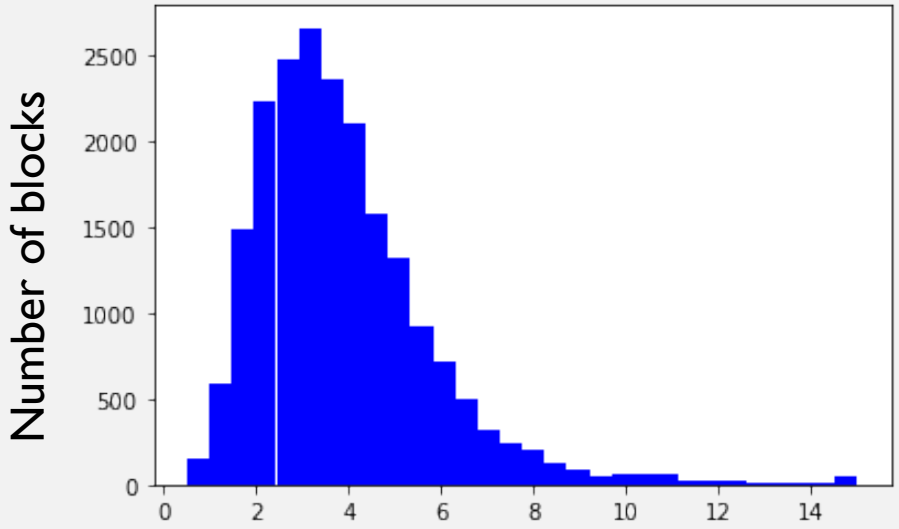
HISTOGRAMS

HOUSEHOLDS



Total number of households
within a block

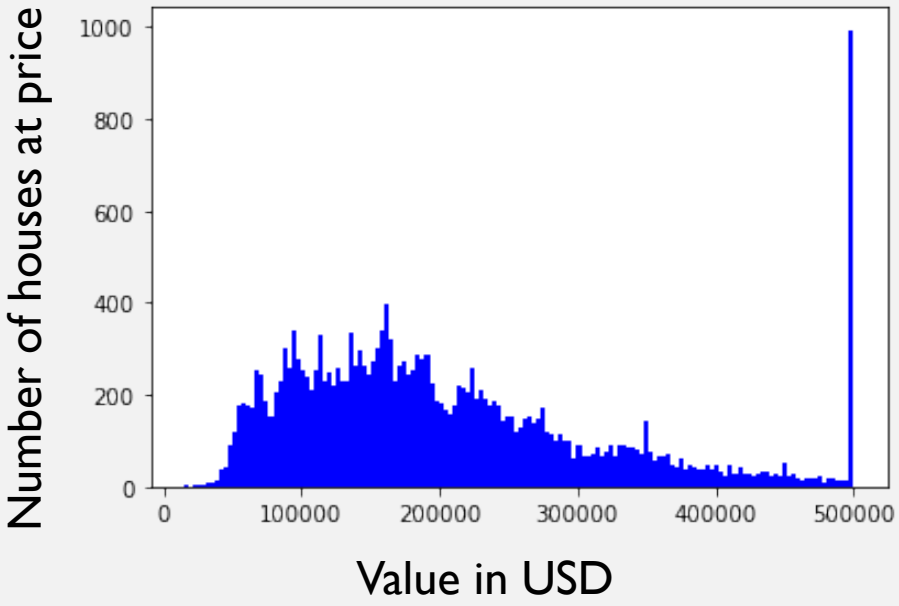
MEDIAN INCOME



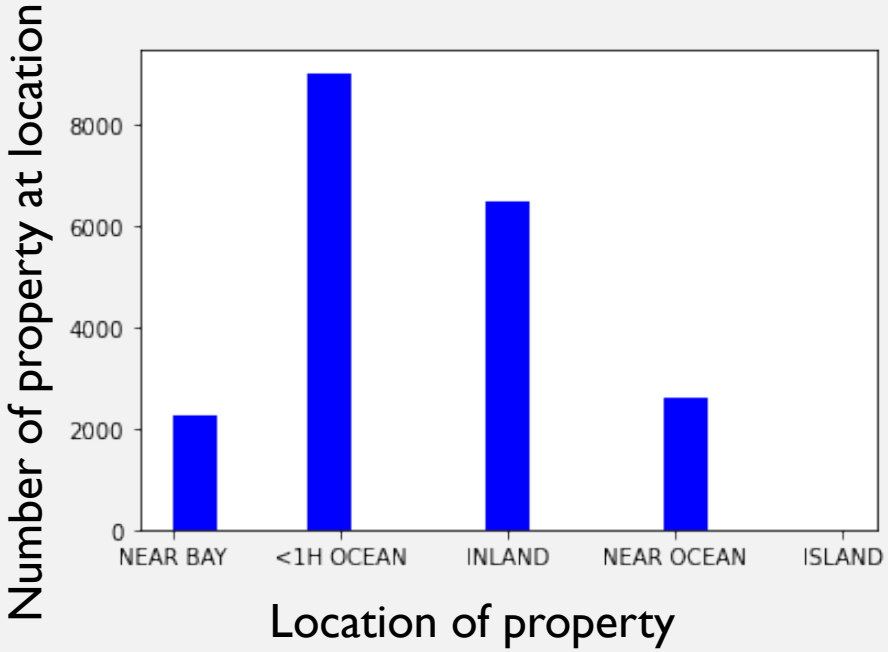
Median income for households
Within a block
(Tens of thousands of USD)

HISTOGRAMS

MEDIAN HOUSE VALUE



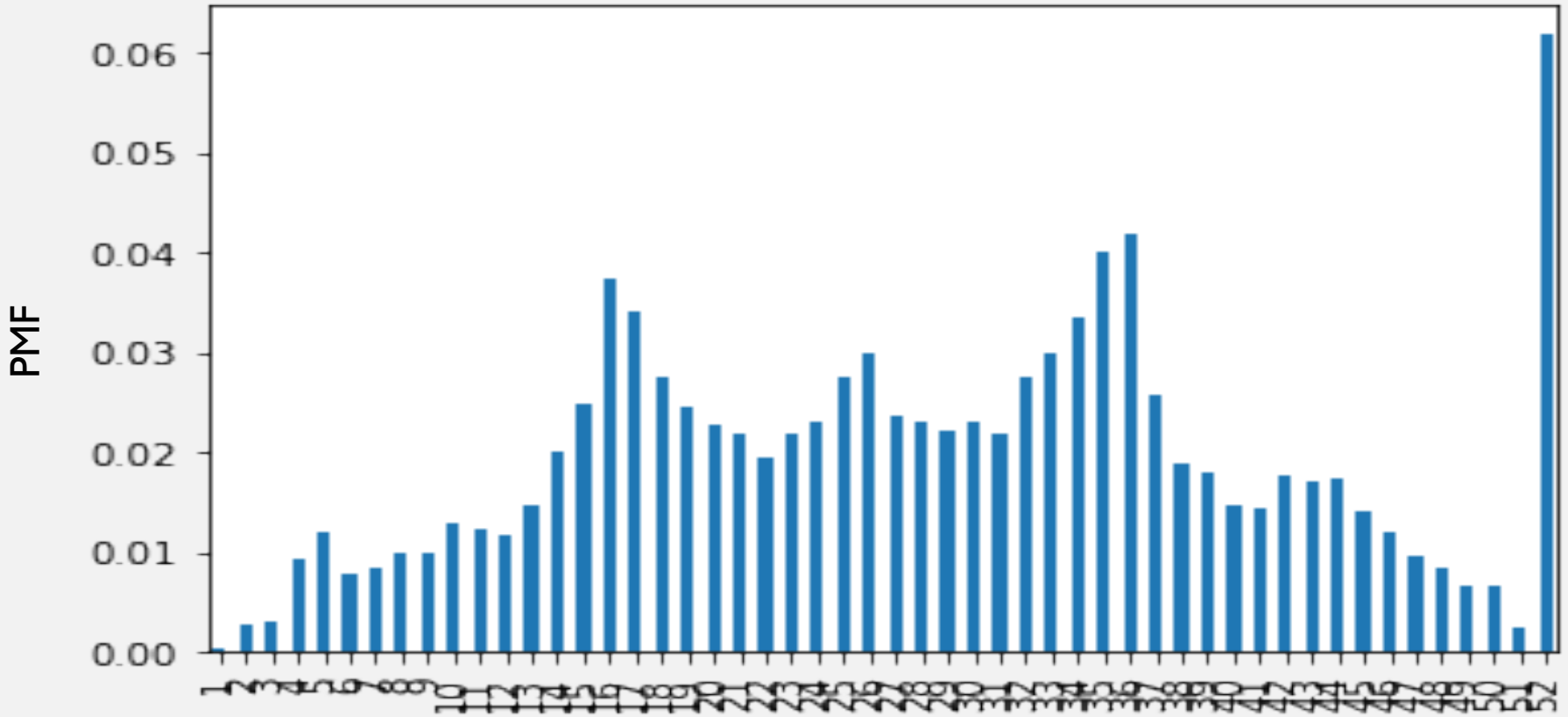
OCEAN PROXIMITY

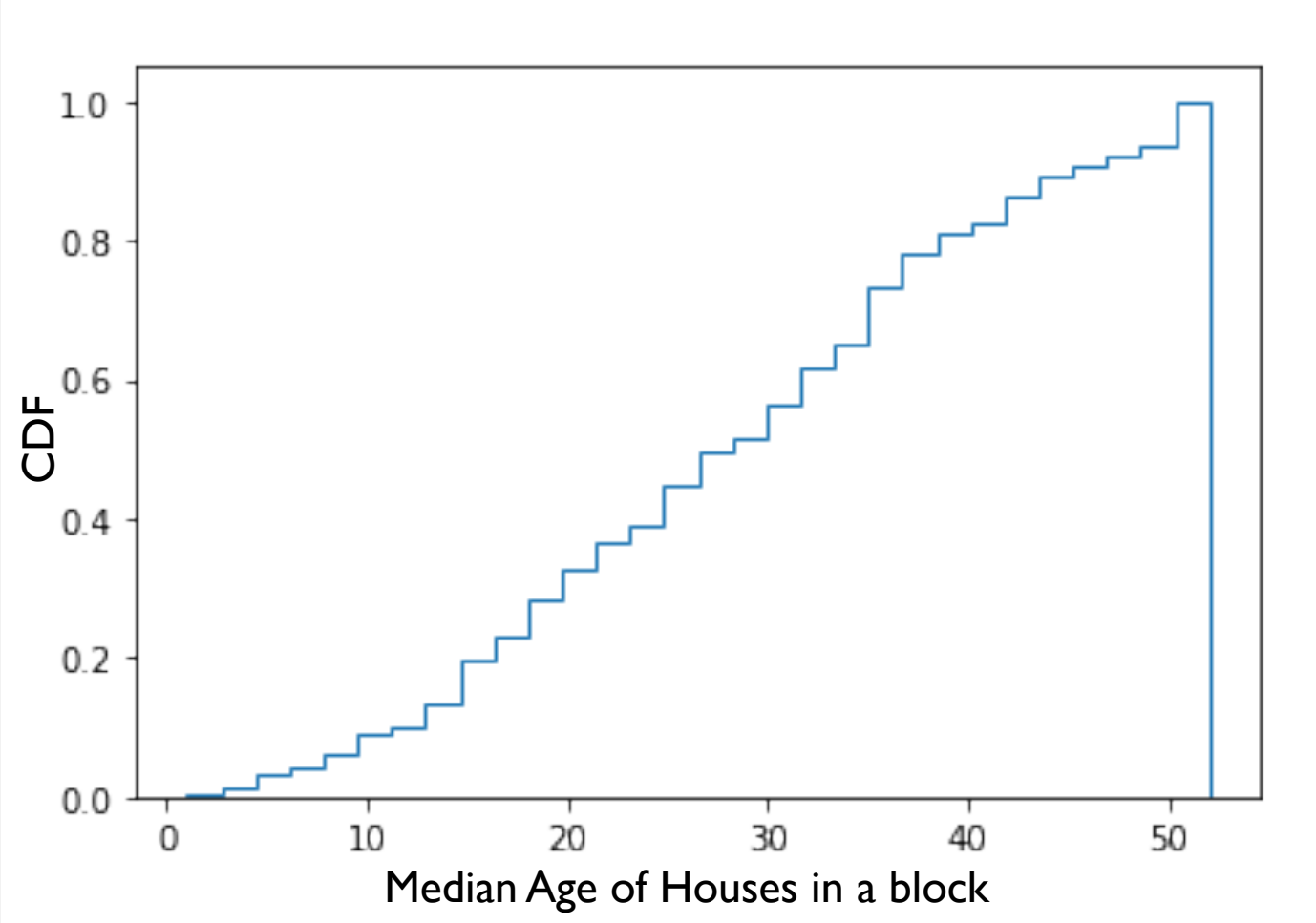


DESCRIPTIVE STATISTICS LISTED AS "MEAN, MODE, VARIANCE"

- Housing Median Age: 28.63, 52, 158.55
- Total Rooms: 2636.5, 1527, 4775403.1
- Total Bedrooms: 537.9, 280, 177565.4
- Population: 1425, 891, 1284161.5
- Households: 499.4, 306, 146152.7
- Median Income: 3.87, 3.125, 3.61
- Median House Value: 206864.41, 500001, 13325393238.50

PMF – HOUSING MEDIAN AGE



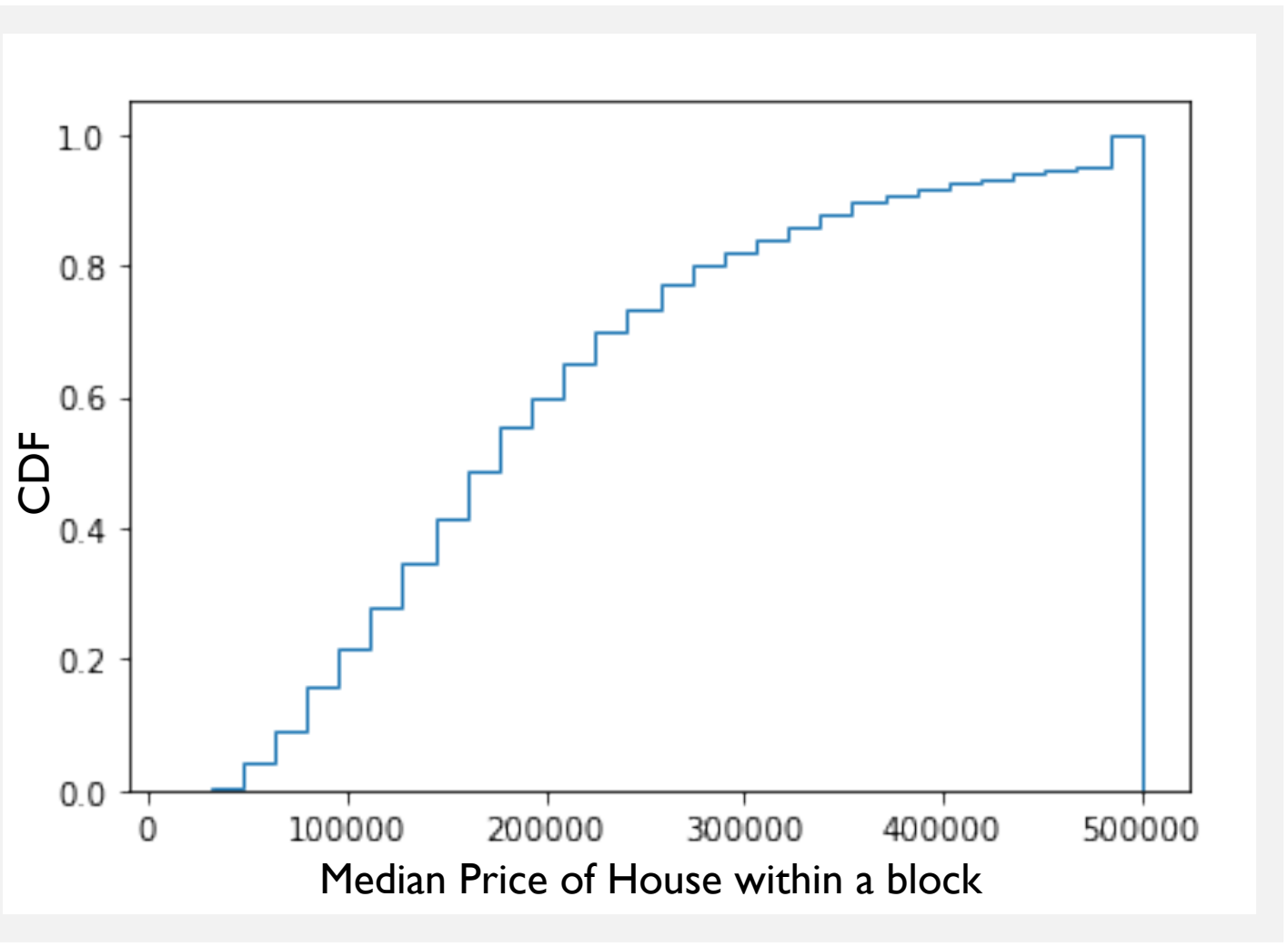


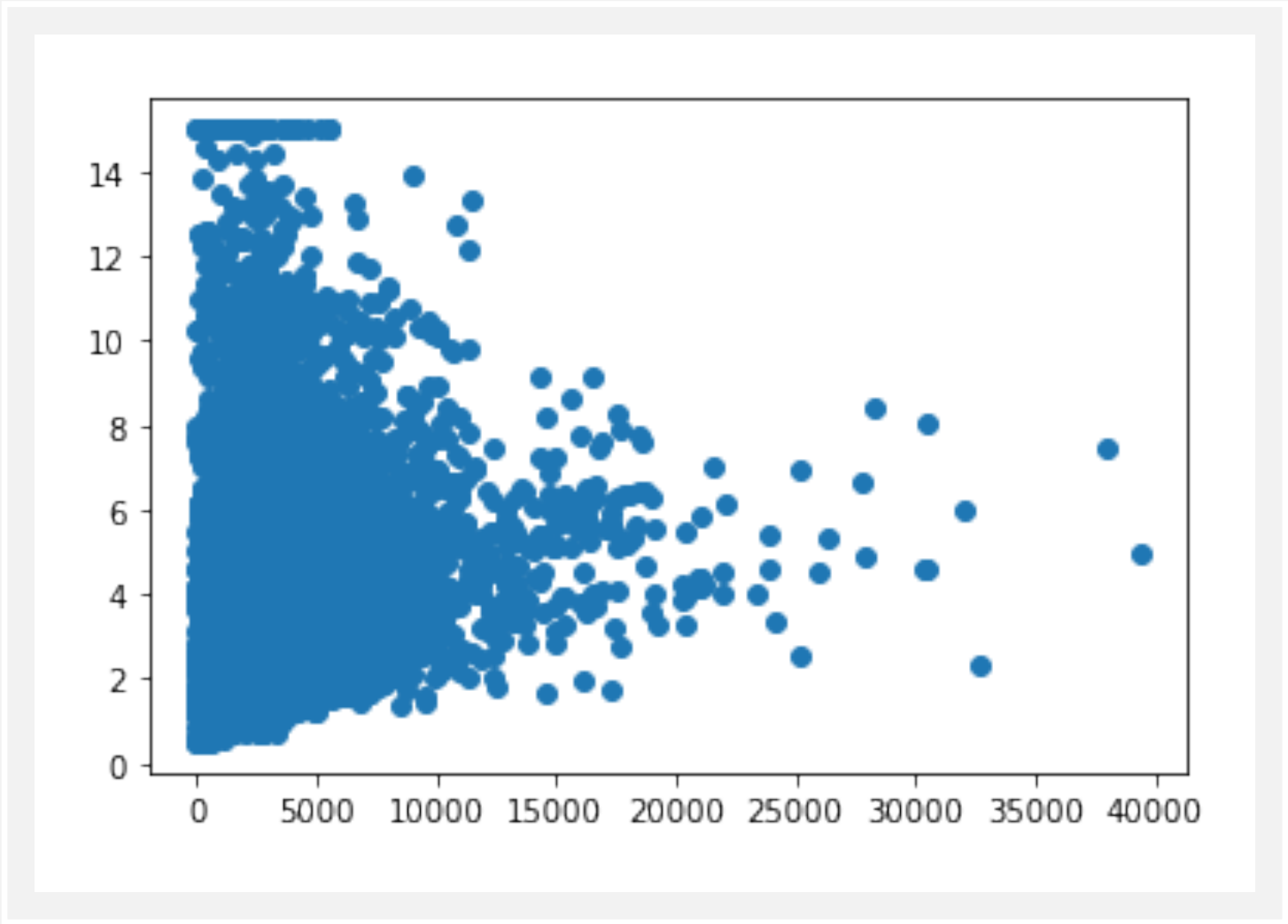
CDF – HOUSING MEDIAN AGE

- Normal distribution of CDF
- It seems that there are the same number of new houses on the market (new to 25 years old) as older houses on the market
- This may mean that it does not matter to Californians how old the house is when buying real estate

EXPONENTIAL CDF OF MEDIAN HOUSE VALUE

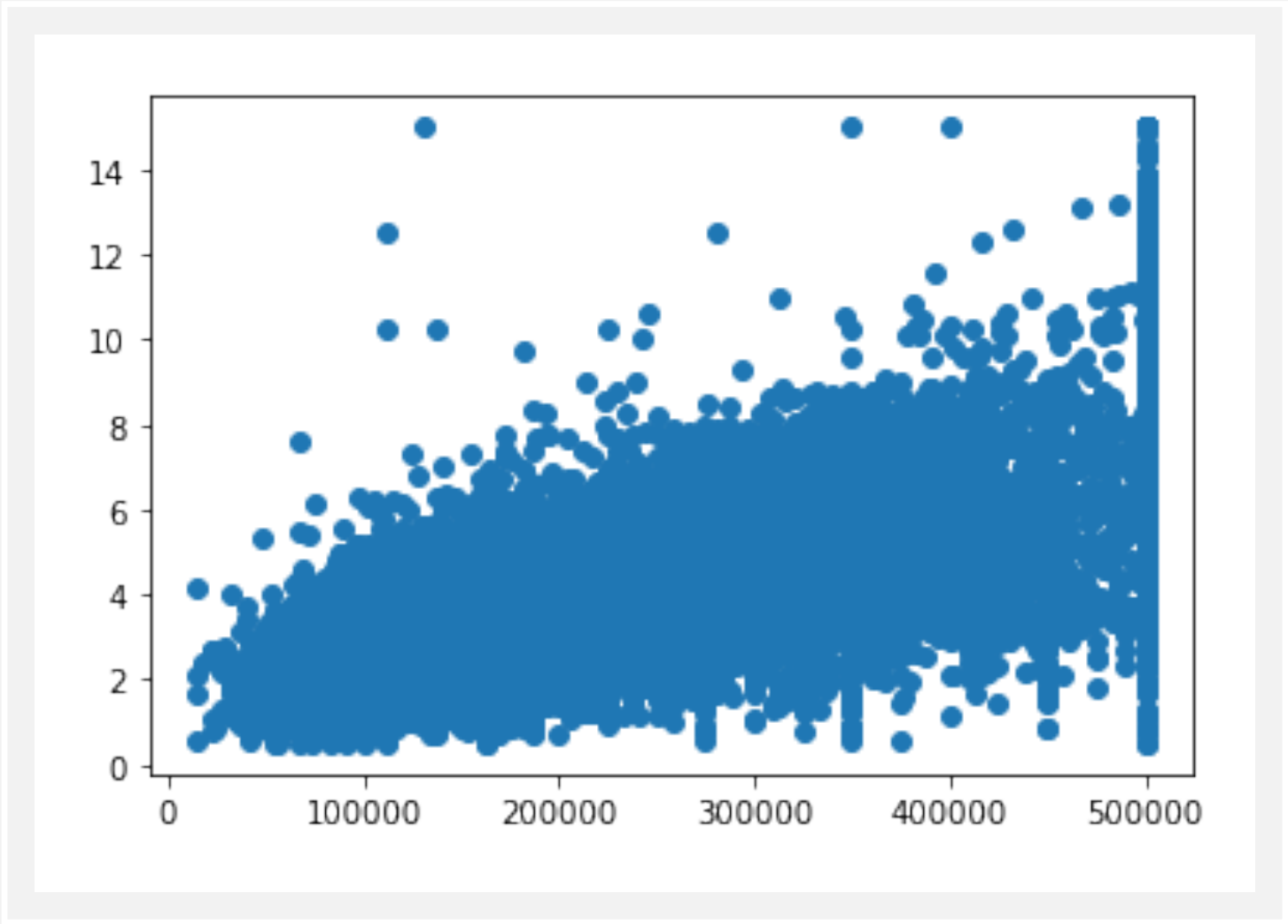
- Shows a relatively exponential distribution
- Many houses increase in price at lower end of the x-axis
- After about $350,000, not many houses vary in price
- $500,000 mark is an outlier; may be accounting for all the houses in the dataset more expensive





SCATTERPLOT

- I had originally thought that a higher income would mean a higher number of rooms (bigger house)
- Plot shows even with a large income, neighborhoods still like to have less than about 17,000 rooms
- Only a few neighborhoods have a lot of bedrooms
- Possible to have many houses within a block (outliers) and only a few houses within a block (rural areas)



SCATTERPLOT

- I was expecting there to be clumps on the left lower corner and right upper corner
- Somewhat uniform distribution, slight increase after $200,000
- Again, $500,000 mark is an outlier; may be accounting for all the houses in the dataset more expensive

MULTIPLE LINEAR REGRESSION

OLS Regression Results

Dep. Variable:	median_house_value	**R-squared:**	0.637
Model:	OLS	**Adj. R-squared:**	0.637
Method:	Least Squares	**F-statistic:**	4478.
Date:	Wed, 18 Nov 2020	**Prob (F-statistic):**	0.00
Time:	19:50:14	**Log-Likelihood:**	-2.5682e+05
No. Observations:	20433	**AIC:**	5.137e+05
Df Residuals:	20424	**BIC:**	5.137e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.585e+06	6.29e+04	-57.001	0.000	-3.71e+06	-3.46e+06
longitude	-4.273e+04	717.087	-59.588	0.000	-4.41e+04	-4.13e+04
latitude	-4.251e+04	676.952	-62.796	0.000	-4.38e+04	-4.12e+04
housing_median_age	1157.9003	43.389	26.687	0.000	1072.855	1242.945
total_rooms	-8.2497	0.794	-10.387	0.000	-9.807	-6.693
total_bedrooms	113.8207	6.931	16.423	0.000	100.236	127.405
population	-38.3856	1.084	-35.407	0.000	-40.511	-36.261
households	47.7014	7.547	6.321	0.000	32.909	62.493
median_income	4.03e+04	337.207	119.504	0.000	3.96e+04	4.1e+04

SOURCES

- Nugent, C. (2017). *California Housing Prices*. Kaggle.
- https://www.kaggle.com/camnugent/california-housing-prices