(F23) PSTAT 126: Project Step 1

Anthony Cu and William Mahnke

2023-10-22

Our Dataset

We plan on using the California Housing dataset curated from Kaggle. The data is compiled from a 1990 California survey, providing information on a specified district in the state. The data describes homes in a district of California in 1990 in order to represent the larger population of this district during the 1990's. Some limitations of this data include the lack of data about individual houses and the individuals in each house. Each independent observation corresponds to a different block within the district.

We view our dataset and explored the information provided in columns.

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1   -122.23    37.88                41         880            129        322
## 2   -122.22    37.86                21        7099           1106       2401
## 3   -122.24    37.85                52        1467            190        496
## 4   -122.25    37.85                52        1274            235        558
## 5   -122.25    37.85                52        1627            280        565
## 6   -122.25    37.85                52         919            213        413
##   households median_income median_house_value ocean_proximity
## 1        126        8.3252            452600        NEAR BAY
## 2       1138        8.3014            358500        NEAR BAY
## 3        177        7.2574            352100        NEAR BAY
## 4        219        5.6431            341300        NEAR BAY
## 5        259        3.8462            342200        NEAR BAY
## 6        193        4.0368            269700        NEAR BAY
```

```
##  [1] "longitude"          "latitude"           "housing_median_age"
##  [4] "total_rooms"        "total_bedrooms"     "population"
##  [7] "households"         "median_income"      "median_house_value"
## [10] "ocean_proximity"
```

The independent quantitative variable is ocean_proximity. This indicates the location of the house in regards to its distance from the ocean, categorized by less than an 1 hour, inland, near bay, near ocean, or an island.

The independent qualitative variables are longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and median_house_value. Longitude and latitude respectively measure how far west and north a home is. Housing median age indicates the median age of a home within a block. Thus, older homes would have greater values and newer homes would have smaller values. Total rooms and total bedrooms represent the number of these rooms in a home within a block. Population indicates the total number of people living within a block. Household indicates the total number of groups living within a home, in a block. Median income represents the median income (measured in $10,000) for households within a block. Median house value measures the median house value (measured in $) for households within a block.

Manipulating Dataset

In order to create another independent categorical variable, we convert the housing_median_age into age buckets. We use the dplyr package to add an additional column categorizing the median ages into their respective age buckets.

```
houseData <- houseData %>% mutate(ageRange = ifelse(housing_median_age < 10, "<10",
                                    ifelse(housing_median_age >= 10 & housing_median_age < 20, "10-20",
                                           ifelse(housing_median_age >= 20 & housing_median_age < 30, "20-30",
                                                  ifelse(housing_median_age >= 30 & housing_median_age < 40,
"30-40",
                                                         ifelse(housing_median_age >= 40 & housing_median_age
< 50, "40-50", "50+")))))) %>%
   filter(!is.na(total_bedrooms))
```

We randomly select 300 observations from the dataset of over 500 rows, using the sample() function. This sample was curated randomly, so it is a representative sample of the population. The total proportion of homes based on categorical variables is similar to the proportion breakdown in our random sample.

We will then store this random sample into a new dataframe named houseData, one that we can collaboratively access for future analyses.

```
houseData <- sample_n(houseData, 300, replace = F)
```

Summarizing Statistics

```
skim(houseData)
```

Data summary

Name	houseData
Number of rows	300
Number of columns	12
_____	
Column type frequency:	
character	2
numeric	10
_____	
Group variables	None

Variable type: character

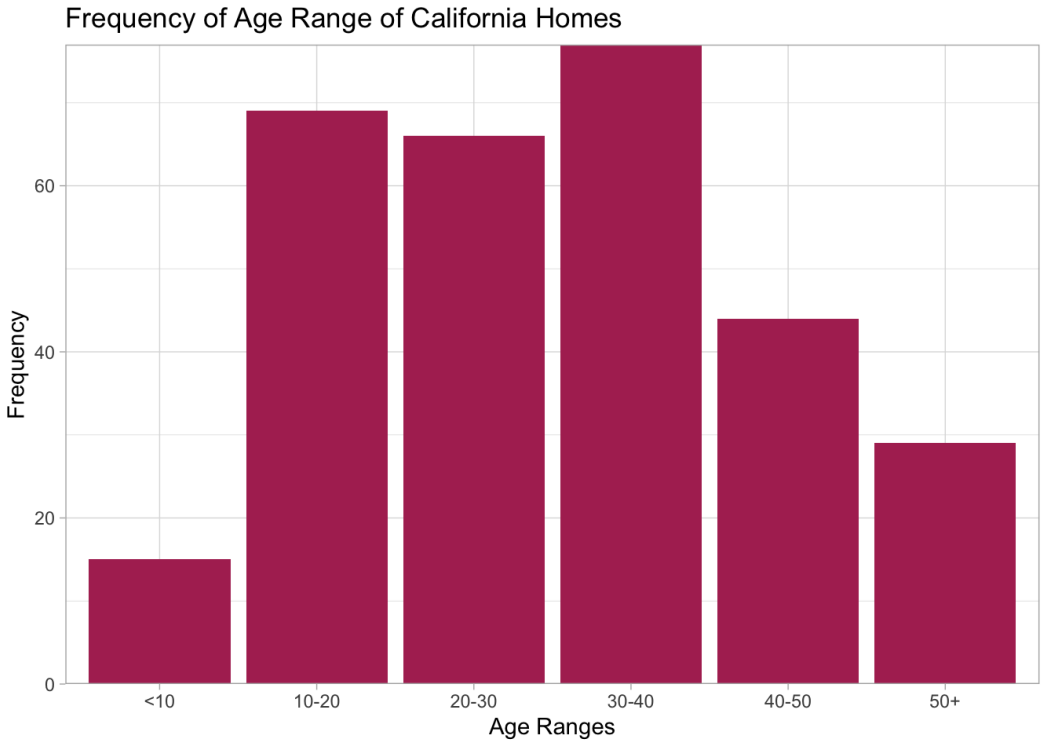
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ocean_proximity	0	1	6	10	0	4	0
ageRange	0	1	3	5	0	6	0

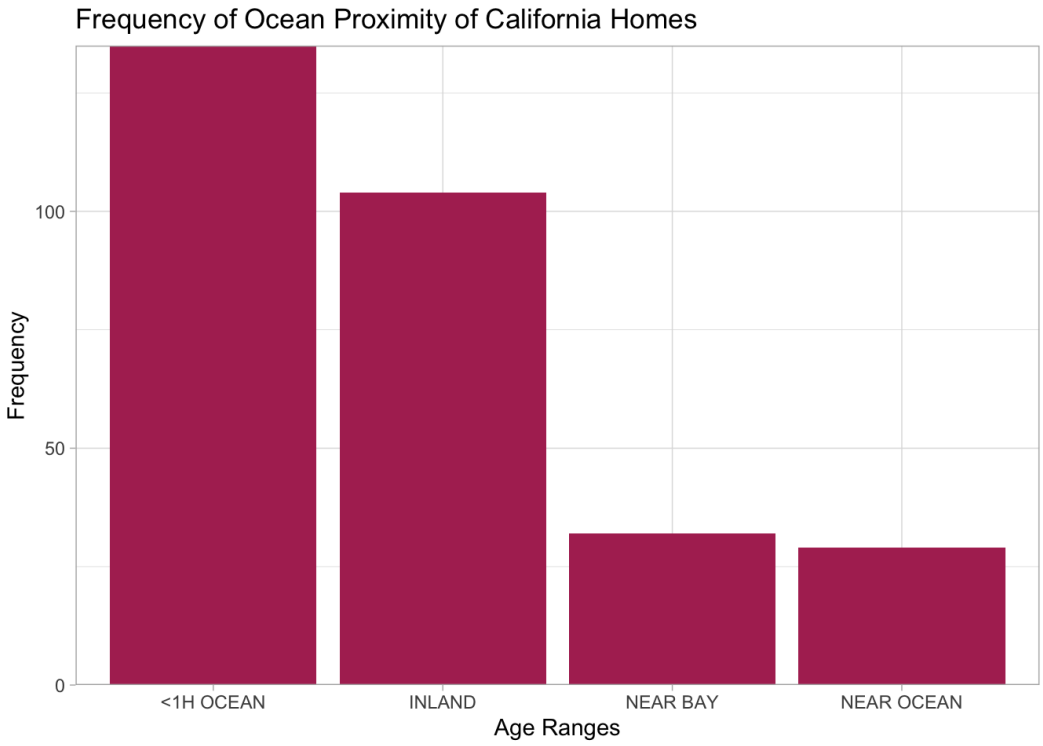
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
X	0	1	150.50	86.75	1.00	75.75	150.50	225.25	300.00	
longitude	0	1	-119.64	2.01	-124.23	-121.77	-118.71	-118.09	-115.38	
latitude	0	1	35.76	2.17	32.57	33.95	34.40	37.75	41.78	
housing_median_age	0	1	29.56	12.93	3.00	18.00	29.50	39.00	52.00	
total_rooms	0	1	2473.66	2037.71	16.00	1331.50	1984.50	3069.00	20377.00	
total_bedrooms	0	1	514.34	408.98	4.00	277.75	415.00	636.00	4335.00	
population	0	1	1371.69	1083.81	8.00	779.00	1154.50	1712.50	11973.00	
households	0	1	481.18	379.72	3.00	266.75	389.50	600.75	3933.00	
median_income	0	1	3.67	1.82	1.09	2.41	3.32	4.37	15.00	
median_house_value	0	1	193931.71	114825.14	40000.00	109650.00	165800.00	250375.00	500001.00	

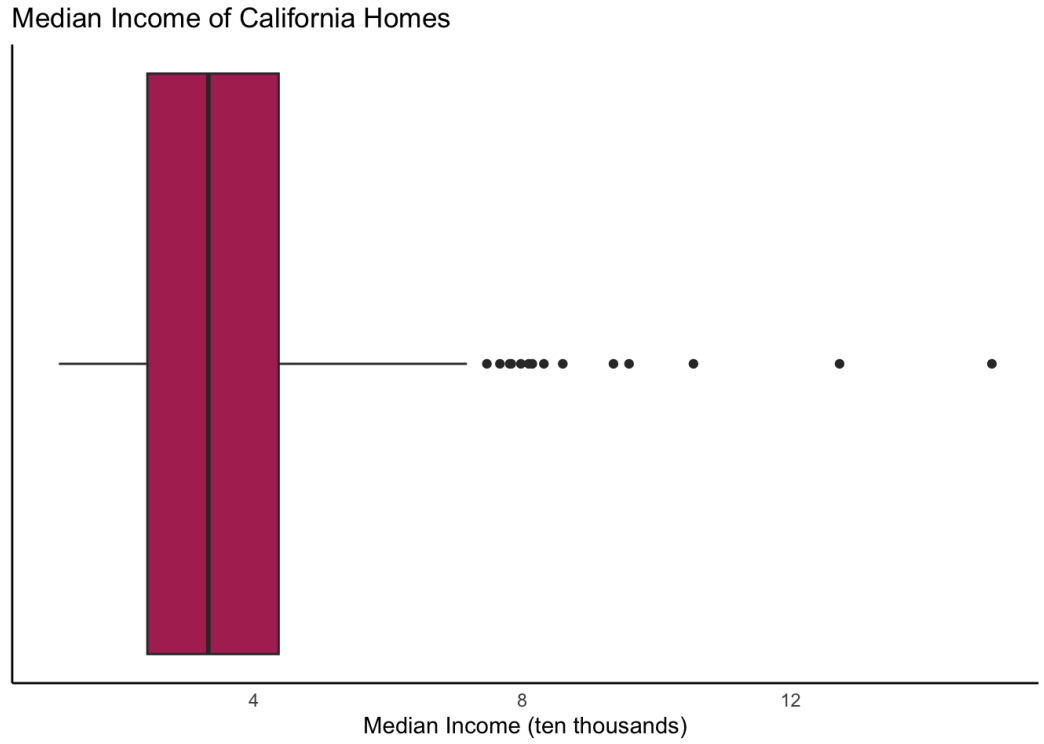
We observe that the median house value and the median income are left skewed in their distribution and the housing median age is symmetrically distributed. This leads us to infer that the age of a house doesn't strongly impact the value of the house.

Visualizations

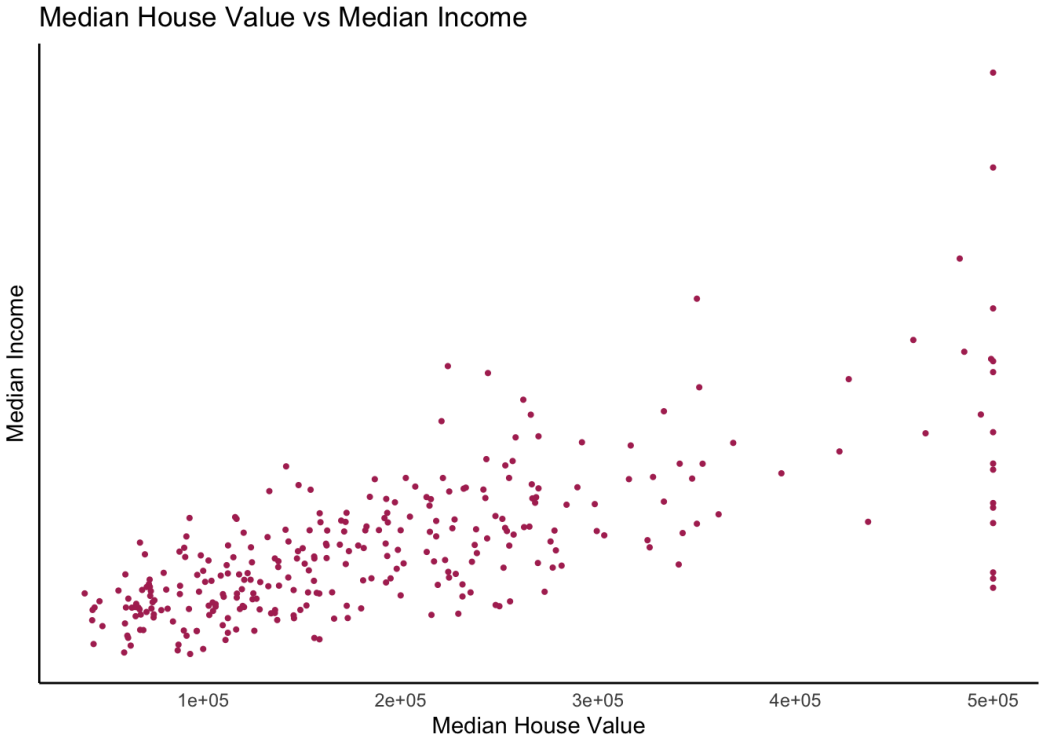




The histograms of the categorical variables suggest that most of the houses are between 10-40 years old and most of the houses are less than an hour from the ocean.



The box plot indicates that the median median income was about $35000 with a few outliers to the right. This motivates our study to research how the median income of the block affects the median value of the houses on the block.



The figure suggests that median house value and median income are linearly related, i.e. as house values increases in price, median incomes of those homeowners rises.

Next Steps

We are interested in exploring the correlation between median income of homes with their location, measured through longitude and latitude. Furthermore, we hope to analyze the income and house value with other house variables, such as its age range.