Brayden & Caleb Go Into the Real Estate Business

STAT 420 Project Proposal

Brayden Turner - brturne2
Caleb Cimmarrusti - Calebtc2

Description of the data file

Number of Columns: 14
Significant Columns:

1. Median House Value: Median house value for households within a block (measured in US Dollars) [$]
2. Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]
3. Median Age: Median age of a house within a block; a lower number is a newer building [years]
4. Total Rooms: Total number of rooms within a block
5. Total Bedrooms: Total number of bedrooms within a block
6. Population: Total number of people residing within a block
7. Households: Total number of households, a group of people residing within a home unit, for a block
8. Latitude: A measure of how far north a house is; a higher value is farther north [°]
9. Longitude: A measure of how far west a house is; a higher value is farther west [°]
10. Distance to coast: Distance to the nearest coast point [m]
11. Distance to Los Angeles: Distance to the centre of Los Angeles [m]
12. Distance to San Diego: Distance to the centre of San Diego [m]
13. Distance to San Jose: Distance to the centre of San Jose [m]
14. Distance to San Francisco: Distance to the centre of San Francisco [m]

Response Variable: Median household value

Number of Entries: 20,640

Background information on the data sets, including specific citation of their source (so that I can also access it).

The selected dataset is an analysis of real estate prices in California which include reference to the population density, income and statistics which describe the nature of the house in question. Some of these variable will be modified in the data cleaning process however it appears we have enough raw material to meet the project specifications.

Link:

https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features

A brief statement of the business, science, research, or personal interest you have in the data set which you hope to explore.

As renters with the hope of one day owning a home (with one of us currently living in one of the most expensive cities in the country in the state of California), understanding how certain variables go in to predicting a home price can help inform us of what to look out for once we get to that point.

Evidence that the data can be loaded into R. Load the data, and print the first few values of the response variable as evidence.

```
> head(California_Houses$Median_House_Value)
[1] 452600 358500 352100 341300 342200 269700
>
```