Stat 420 Final Project Proposal

STAT 420, Summer 2022, D. Unger

Stat 420 Final Project Proposal

1. Team Member information

Team Member's Names:

- Kathryn DeWitt
- Shashank Thakur
- Avinash Tiwari

NetIDs:

- kdewitt3
- sthakur5
- tiwari6

2. Proposed Title:

We propose calling our project **Modeling California Housing Prices**.

3. Description of Data File

'California Housing Prices' dataset pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. There are 10 columns and 20.6K rows in the data. The columns are as mentioned below:

Column Name	Description
longitude	Longitudinal location of block group of houses.
latitude	Latitudinal location of block group of houses
housing_median_age	Median age of the houses in the block group
total_rooms	Total rooms in the group of houses
total_bedrooms	Total bedrooms in the block group
population	Total number of people in the block group
households	Households comprised in the block group
median_income	Median income calculated from the individual income of house.
median_house_value	Median house value of a block group.
ocean_proximity	Indicating whether each block group is near the ocean, near the Bay Area, inland or on an island.

4. Background on datasets

For our project we will be using the data set from kaggle California Housing Prices

The data contains information from the 1990 California census and pertains to houses found in a given California district. It contains one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

5. Research Question/Interest in Data Set

We would like to model median house value of a block group in CA in the 1990's based on administrative (Census) data. Given the recent turbulence in the real estate market and associated bidding wars, there's a renewed interest in what the value of a house is. The data set we propose using is interesting for the following reasons:

- Unlike scraped real estate data, the data set is gather by an administrative body, the Census. Since the stakeholder involved in collecting the data is not interested in selling a house, we see different features represented that relate more to geography (Latitude, longitude), population density (population, number of households), and age of the building (median_housing_age).
- The dataset is aggregated at the block level (rather than at the house level). This means our analysis is not directly modelling housing values, but the median housing value of a neighborhood. We believe this will help us better to understand the value of the neighborhood as opposed to the value of the amenities of the individual houses.
- Additionally, this data set meets the requirements set out by the project description:
  - The dataset has more than 20,000 observations
  - The dataset has a good mix of categorical and numerical variables to enable us to learn from the course
  - The dataset has 10 columns
  - The dataset coming from Kaggle means there has been extensive work done by many researchers / learners around the world providing our team with a baseline to compare our model & methodology with others. We also learn from their work and explore how much improvement could be achieved by applying modeling techniques other than regression to achieve better model performance

6. Evidence of Loading into R

We have included below evidence that the data can be loaded into R.

```
ca_housing_data = read.csv('../000_Data/california-housing-prices/housing.csv')
head(ca_housing_data)
```

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1    -122.2    37.88                41         880            129        322
## 2    -122.2    37.86                21        7099           1106       2401
## 3    -122.2    37.85                52        1467            190        496
## 4    -122.2    37.85                52        1274            235        558
## 5    -122.2    37.85                52        1627            280        565
## 6    -122.2    37.85                52         919            213        413
##   households median_income median_house_value ocean_proximity
## 1        126         8.325             452600        NEAR BAY
## 2       1138         8.301             358500        NEAR BAY
```

```
## 3        177         7.257           352100       NEAR BAY
## 4        219         5.643           341300       NEAR BAY
## 5        259         3.846           342200       NEAR BAY
## 6        193         4.037           269700       NEAR BAY
```