Data Analysis Project Proposal

Team : Illini IN Stat - Anant Ashutosh Sharma (anantas2), Harpreet Siddhu (hsiddhu2)

Contents

Title

Can Statistics Help in Predicting House Prices?

Proposal Introduction

Over the years, transactions under the real estate industry have increased many folds. The pandemic has further driven the middle income populations mindset towards acquiring personal residential property. Housing prices, which depend on the different parameters associated with the house, are an important metric for closing and finalizing any such residential real estate transaction.

In this project, **we attempt to study the effect of different variables associated with a house on the price of the house.** While the data being used for this study is old, we present this project as an academic study on the utility of statistics and regression to model and predict the housing prices.

We plan to implement the concept learned from the STAT 420 course at UIUC to do preliminary statistical analysis on the 1990 California House Prices and plan to narrow down on an accurate mode to predict the house prices based on certain relevant parameters.

Background Information on Dataset

The dataset which is being used in this study is publicly available and has been taken from Kaggle. The dataset is named as **"California Housing Prices"** and can be accessed as a csv file at https://www.kaggle.com/camnugent/california-housing-prices.

Originally, the dataset was made available at the Statistics Library in CMU on the StatLib platform as `housing.zip`. The original dataset can be access at http://lib.stat.cmu.edu/datasets/ under the names `housing`.

The contains information from the 1990 California census and pertains to the houses found in a given California district. It contains some of the summary statistics about the houses based on the 1990 California census. As stated above, although the dataset is old, we wish to present this project as an academic study on the utility of statistics and regression to model and predict housing prices.

This data was initially featured in the following paper: Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297.

Description of Dataset

The California Housing Prices dataset provides the median house prices for California districts derived from the 1990 census data. 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).

It contains a total of **10 columns** which include the median house prices and other variables which might be associated with the price.

The data contains almost 20640 rows. We will build a model to predict is "Median-House-Value" and the rest are columns will be used as predictor variables. The dataset is not cleaned so we will be pre-processing it before using it for modeling.

The following are data points from the dataset:

- **1.longitude:** A measure of how far west a house is. The higher value of this parameter means farther west. The California longitude value ranges from: 114° 8' W to 124°.
- **2.latitude:** A measure of how far north a house is. The higher value of this parameter means farther north. The California Latitude value ranges from: 32° 30' N to 42° N
- **3. housingMedianAge:** Median age of a house within a block (a block has population of around 600 to 3000 people). The lower number for this parameter means the building is newer.
- **4. totalRooms:** Total number of rooms among all houses within a block.
- **5. totalBedrooms:** Total number of bedrooms among all houses 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. medianIncome:** Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- **9. medianHouseValue:** Median house value for households within a block (measured in US Dollars)
- **10. oceanProximity:** Location of the house w.r.t ocean/sea. `Ocean_proximity` indicating (very roughly) whether each block group is near the ocean, near the Bay area, inland or on an island. This parameter will allows us include and interpret the categorical variable while regressing the dataset.

Why we chose this dataset to perform the analysis?

By studying the effect of different variables associated with a house on the price of the house, one is able to obtain a better overview of which parameters drive the housing prices. Since a house transaction depends strongly on the quoted price, a predictive model would allow people to get an confidence interval of the expected price for their house which might help them close a deal and complete a transaction successfully.

We choose this model for our analysis because we want -

- Hands on expedience with real life dataset. We chose this dataset because it is a medium sized yet not small or too large. This dataset satisfy all the project requirements.
  - Minimum 2,000 observations : Our chosen dataset has 20640 observations
  - A Numeric Response Variable : `median_house_value`
  - At least one catergorical variable : `ocean_proximity`
  - Atleast two continuous numeric : `total_bedrooms`, `total_rooms`, `population`, `households` etc.
- The motivation of choosing this dataset is to implement certain techniques learned in STAT 420 course to perform statistical analysis on the dataset.
- Discover how applied statistics can help us building the most accurate model and predicting the subject of our study (House Prices).

Required Evidence for Dataset

Load into R

Here is a snippet of data being displayed

```
library(readr)
housing <- read.csv("housing.csv")
head(housing)
```

```
##   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
```

Structure of Dataset:

```
str(housing)
```

```
## 'data.frame':    20640 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
```

```
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

Categorical Variable `ocean_proximity` Levels

```
levels(as.factor(housing$ocean_proximity))
```

```
## [1] "<1H OCEAN"  "INLAND"     "ISLAND"     "NEAR BAY"   "NEAR OCEAN"
```

Credits

Pace, R. Kelley, and Ronald Barry. "Sparse spatial auto-regressions." Statistics & Probability Letters 33.3 (1997): 291-297.