Data Memo

Overview of the dataset

```
housing <- read.csv("housing.csv")
head(housing,10)
```

```
##    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
## 7    -122.25    37.84                 52        2535            489       1094
## 8    -122.25    37.84                 52        3104            687       1157
## 9    -122.26    37.84                 42        2555            665       1206
## 10   -122.25    37.84                 52        3549            707       1551
##    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
## 7         514        3.6591             299200        NEAR BAY
## 8         647        3.1200             241400        NEAR BAY
## 9         595        2.0804             226700        NEAR BAY
## 10        714        3.6912             261100        NEAR BAY
```

What does it include? The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data

Where and how will you be obtaining it? Include the link and source. The data was downloaded from Kaggle Link: https://www.kaggle.com/datasets/camnugent/california-housing-prices Source:This dataset appeared in a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in the Statistics and Probability Letters journal. They built it using the 1990 California census data.

About how many observations? How many predictors? There are 20640 observation, and 8 predictors

What types of variables will you be working with? Most of variables are numeric and ocean_proximity is character variable

```r
sum(is.na(housing))
```

Is there any missing data? About how much? Do you have an idea for how to handle it?

```
## [1] 207
```

There is 207 missing data. I am planning to remove the missing value.

Overview of the research question

What variable(s) are you interested in predicting? What question(s) are you interested in answering? The variable I am interested in predicting is median house value. The question I am interested in answering are following: + what are some factors relate to the house pricing + Which model will predict the housing price most accurately

Name your response/outcome variable(s) and briefly describe it/them. Response variable: median_house_value(The house median value in one census block group which typically has a population of 600 to 3000 people)

Will these questions be best answered with a classification or regression approach? These questions be best answered with a regression approach

Which predictors do you think will be especially useful? I think median_income and house_median_age will be especially useful

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain. My model will focus on prediction. I am planning to use machine learning method taught in class to predict California housing price and compare the accuracy of different models.

Proposed project timeline

When do you plan on having your data set loaded, beginning your exploratory data analysis, etc? I am planing to begin my exploratory data analysis next week and utilize different model based on class progress

Provide a general timeline for the rest of the quarter. General timeline is based on class progress - Week1-3 load and clean data - Week4-5 Exploratory data analysis - Week6-8 Run models and get results - Week9 Final edition

Questions and Concerns

Are there any problems or difficult aspects of the project you anticipate? My current questions are 1. Are there many points in my data that can be analyzed 2. Is it possible to change the data after I submitting data Memo