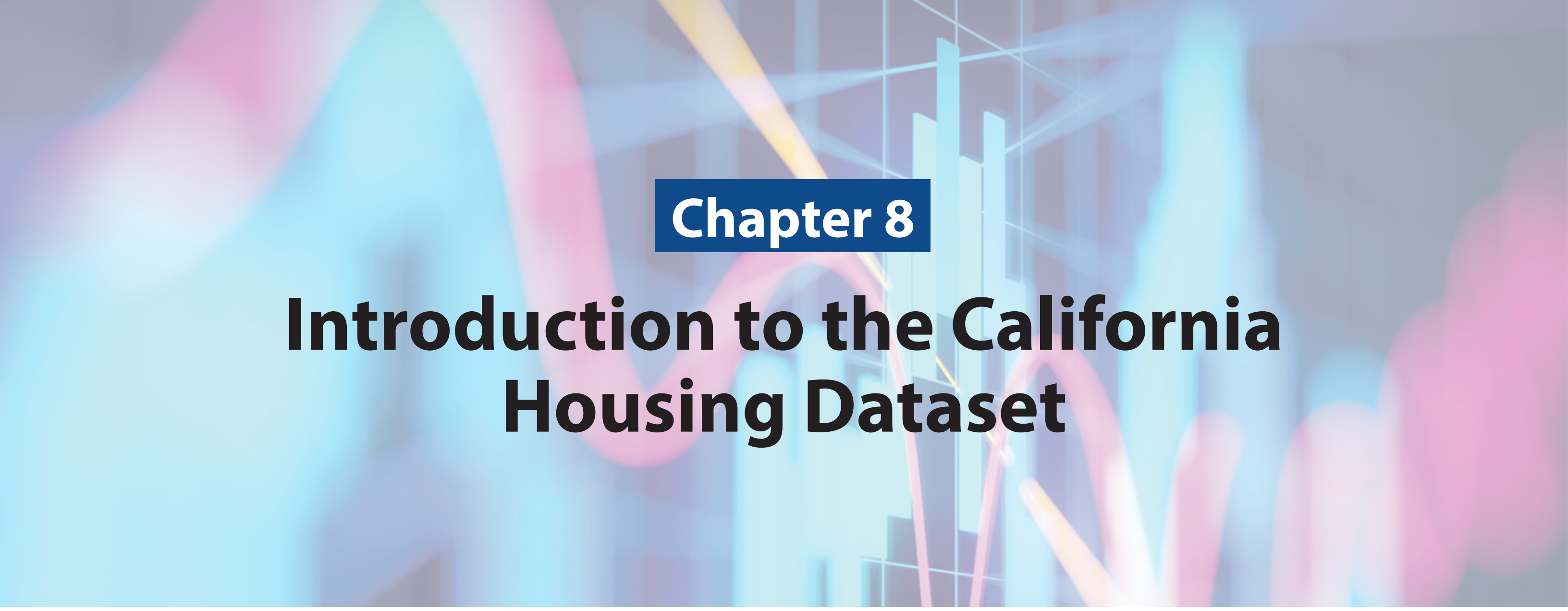


# Chapter 8

# Introduction to the California Housing Dataset

The California housing dataset is a widely used dataset in machine learning and statistics. It contains data collected from the 1990 California census and includes various features related to housing districts in California. The dataset is often used for regression analysis tasks, particularly for predicting housing prices.

## 8.1 Data Dictionary

Here is a brief description of the features typically included in the California housing dataset:

- Longitude and latitude: Geographic coordinates of the housing district.
- Median house value: The median value of owner-occupied homes in the district, which serves as the target variable for regression tasks.
- Median income: The median income of households in the district.
- Median house age: The median age of houses in the district.
- Average number of rooms: The average number of rooms per household in the district.
- Average number of bedrooms: The average number of bedrooms per household in the district.
- Population: The total population of the district.
- Average occupancy: The average household occupancy in the district.

These features provide insights into various aspects of housing districts in California, such as socioeconomic status, housing characteristics, and population density. Researchers and practitioners often use this dataset to explore patterns, relationships, and predictive models related to housing prices and demographic factors.

The above descriptions of the variables are what we call the data dictionary, which provides us with more insight into the way the data was collected. It is important to assess these things before moving forward with the analysis.

## 8.2 What Can Be Uncovered from the Dataset?

The California housing dataset offers a rich source of information that can be explored to uncover various insights and address a range of questions. Here are some examples of questions that could be investigated using this dataset:

- How does the median house value vary across different regions of California?
- What is the relationship between median income and median house value in California?
- Are there any patterns or trends in housing prices based on geographical location (longitude and latitude)?

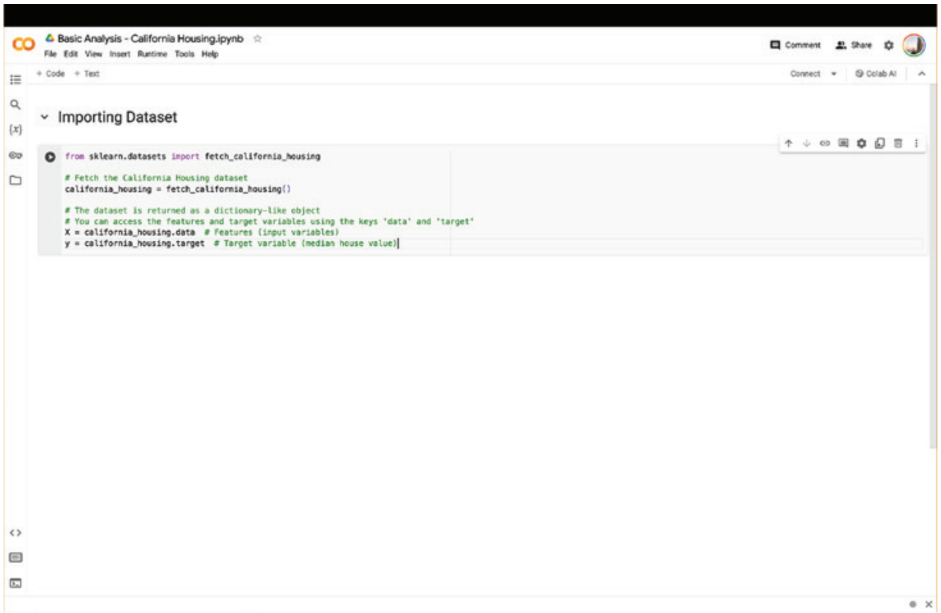
- How does the age of houses in a district affect the median house value?
- Is there a correlation between the average number of rooms or bedrooms in a household and the median house value?
- How does population density impact housing prices in California?
- Are there any significant differences in housing prices between urban, suburban, and rural areas?
- Can demographic factors such as population size and median income predict variations in housing prices?
- How has the median house value changed over time since the 1990 census?
- Are there any outliers or anomalies in the dataset that may indicate unique factors influencing housing prices in certain districts?

Exploring these questions can provide valuable insights into the factors influencing housing prices in California and contribute to a better understanding of housing markets and socioeconomic dynamics in the region.

## 8.3 Importing the Dataset to Colab

*Step 1: Create a new Colab notebook and title it "Basic Analysis—California Housing."*

*Step 2: Import the necessary libraries and fetch the data.*


Basic Analysis - California Housing.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
Connect
Colab AI
Importing Dataset

```
from sklearn.datasets import fetch_california_housing

# Fetch the California Housing dataset
california_housing = fetch_california_housing()

# The dataset is returned as a dictionary-like object
# You can access the features and target variables using the keys 'data' and 'target'
X = california_housing.data  # Features (input variables)
y = california_housing.target  # Target variable (median house value)
```

© Anusha Vissapragada

## 8.4 Python Libraries

In Python, libraries (also known as modules or packages) are collections of pre-written code that provide functionality to perform specific tasks. These libraries contain functions, classes, and constants that programmers can use to simplify their code and avoid reinventing the wheel.

Libraries are an essential part of the Python ecosystem and are created and maintained by developers to address various needs across different domains, such as data science, web development, machine learning, and more. They allow programmers to leverage existing code and tools, saving time and effort in the development process.

## 8.5 Common Python Libraries

Python offers a rich ecosystem of libraries for data science that are widely used for tasks such as data manipulation, visualization, machine learning, and statistical analysis. Some common Python libraries for data science include

- NumPy
  - NumPy is a fundamental package for scientific computing in Python. It provides support for large, multidimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
- Pandas
  - Pandas is a powerful library for data manipulation and analysis. It provides data structures like DataFrames and Series, which are highly efficient for handling structured data. Pandas facilitates tasks such as data cleaning, exploration, transformation, and visualization.
- Matplotlib
  - Matplotlib is a versatile library for creating static, interactive, and animated visualizations in Python. It offers a wide range of plotting functions to generate various types of plots, including line plots, scatter plots, bar plots, histograms, and more.
- Seaborn
  - Seaborn is built on top of Matplotlib and provides a higher-level interface for creating attractive and informative statistical graphics. It simplifies the process of creating complex visualizations and offers built-in support for tasks like data aggregation and visualization of categorical data.
- Scikit-learn
  - Scikit-learn is a comprehensive library for machine learning in Python. It includes a wide range of algorithms for classification, regression, clustering, dimensionality reduction, and more. Scikit-learn also provides tools for model evaluation, hyperparameter tuning, and feature selection. It is also commonly known as sklearn.
- SciPy
  - SciPy is a library for scientific computing and technical computing in Python. It builds on top of NumPy and provides additional functionality for optimization, integration, interpolation, linear algebra, signal processing, and more.
- Statsmodels
  - Statsmodels is a library for statistical modeling and hypothesis testing in Python. It offers a wide range of statistical models, including linear regression, logistic regression, time series analysis, and generalized linear models. Statsmodels also provides tools for conducting statistical tests and exploring relationships in data.

In this textbook, we will mostly be using the first five libraries.

## 8.6 Importing Data from an Existing Library

In the above code, we used sklearn to fetch the California housing dataset. Sklearn has a few in-built datasets, the California housing dataset is one of them. In the first few lines of code, we created variables:

1. california_housing is storing the data.
2. X, is storing all the input or independent variables.
3. y, is storing the output or dependent variables.

## 8.7 Input Variables and Output Variables

Going back to the definition of a function, this is a supervised learning model that we are aiming to build. The *y* variable that we want to predict is the price of the house. The *x* variables are the remaining variables that will serve as the input to predict the output.

*Input Variables—Predictor Variable—Independent Variable*

- Longitude and latitude: Geographic coordinates of the housing district.
- Median income: The median income of households in the district.
- Median house age: The median age of houses in the district.
- Average number of rooms: The average number of rooms per household in the district.
- Average number of bedrooms: The average number of bedrooms per household in the district.
- Population: The total population of the district.
- Average occupancy: The average household occupancy in the district.

*Output Variable—Target Variable—Dependent Variable*

- Median house value: The median value of owner-occupied homes in the district, which serves as the target variable for regression tasks.

After this setup, we are now ready to begin the analysis.

## 8.8 Colab Notebook

https://colab.research.google.com/drive/1GtSe8ePdsWs0yqu1fiX4rDzE0WWY3s8z?usp=sharing