**University of Houston-Clear Lake**

2700 Bay Area Blvd

Houston, TX, 77058

(281)-283-7600

California Housing Price Prediction.

**[ Linear Regression with Gradient Descent ]**

**April 21, 2025**

# Introduction

This project implemented a Linear Regression model with Gradient Descent from scratch to predict house prices using a subsampled California Housing Prices dataset (4,999 rows, 9 numeric features). The algorithm is coded in python programming language from scratch , the model was trained, evaluated, and compared with scikit-learn’s LinearRegression. Visualizations illustrate model performance. The dataset, sourced from Subset\_Housing.csv, was cleaned with feature scaling . Subsampled from 20,640 rows, I use to be precise 5,000-rows of data. I have done some data pre-processing on dataset and focusing on numeric features. This report details the dataset, cleaning, methodology, implementation, results, and visualizations.

# Dataset Description

# The **California Housing Prices Dataset** (Subset\_Housing.csv, Kaggle: <https://www.kaggle.com/datasets/camnugent/california-housing-prices>) was subsampled to 4,999 rows from 20,640.

# **Features** (9, pre-cleaning):

# Longitude: -124.35 to -114.55

# Latitude: ~32.67 to 41.95

# Housing Median Age: ~1–52

# Total Rooms: ~2–28,258

# Total Bedrooms: ~ 2–4,457

# Population: ~ 6–12,203

# Households: ~2–4,204

# Median Income: 0.5–15.0001

* + Ocean\_proximity : - NEAR BAY,INLAND,<1H OCEAN

# Rooms per Household: 1–14

# **Target**: Median House Value (~$14,999–$500,001)

### Why This Dataset?

* The housing price dataset has nearly 20,000 rows of data specially designed for regression type of problem.
* One of the main reasons for choosing this dataset is , the dataset has only one column with string Datatype and with 9 feature having numeric values , which is ideal for Linear Regression problem.
* Dataset contains values that has high difference which requires data pre-processing.

**Data Cleaning**

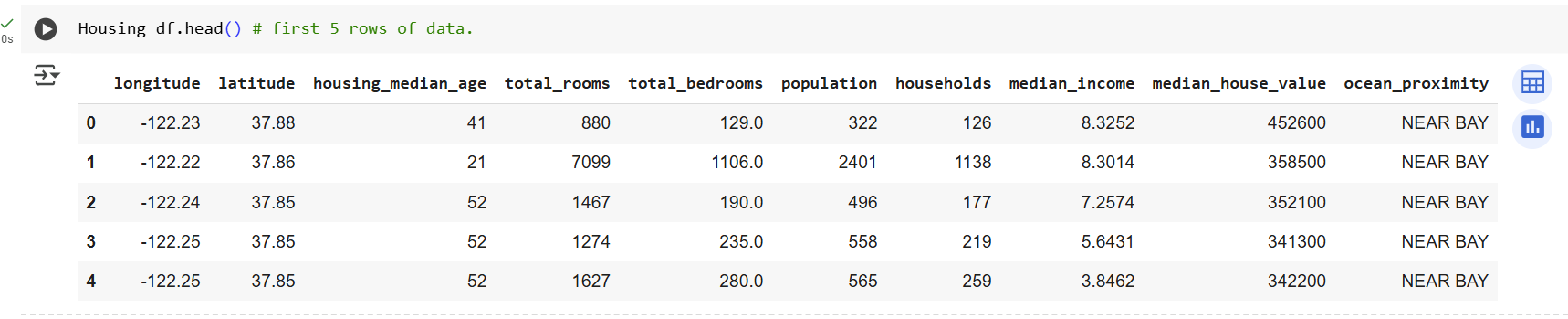
Here are the snapshots of the data cleaning part on the dataset.

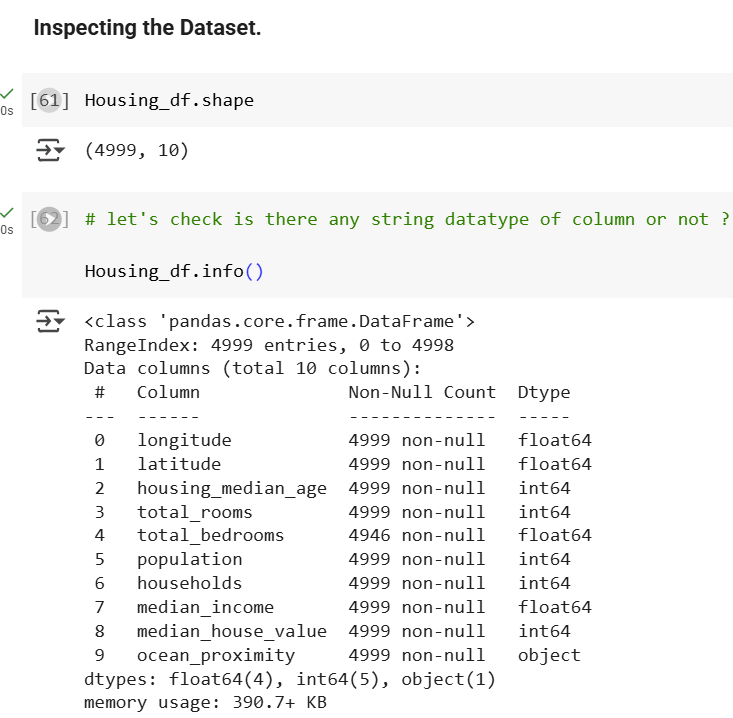
* After carefully inspection of data it seems that , the dataset has missing value on ~ 53 rows in column “total\_bedroom”. In order to clean this column we need to fill missing value with the median value of that column.
* We also have column called “ocean\_prximity” which has string datatype ,which can’t be process by Linear Regression Algorithm, so we need to get rid o that column for proceeding.

**SnapShots of Cleaning Part**

A screenshot of a computer program

AI-generated content may be incorrect.

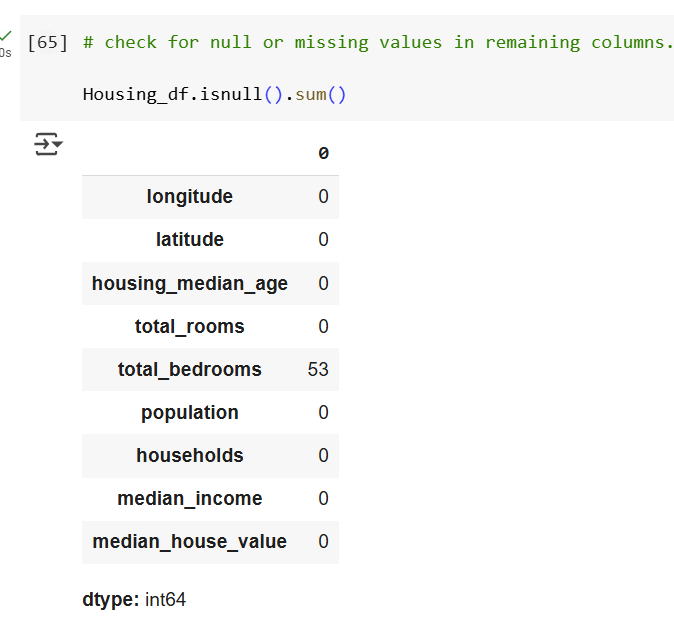




A screenshot of a computer

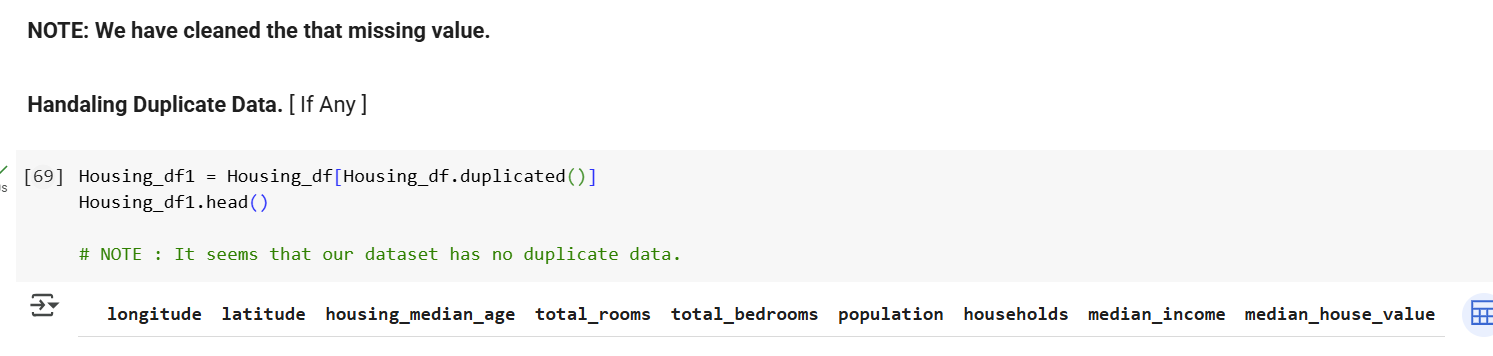
AI-generated content may be incorrect.

**Data Pre-processing**

****

**A screenshot of a computer program

AI-generated content may be incorrect.**

****

**A black text on a white background

AI-generated content may be incorrect.**A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

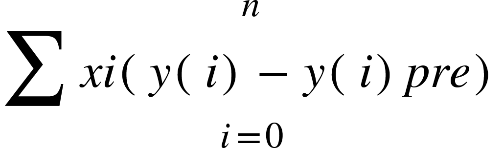
Note : Standardization is refer to scaled down your features to **“Standard Normal Distribution”**

Formula : z = ( x - µ ) / Standard Deviation

# Implemenation

1. Loading Cleaned Dataset : -
   1. Used pandas.read\_csv("Subset\_Housing.csv") to load data.
2. Cleaning : -
   1. Dropped ocean\_proximity (Housing\_df.drop), imputed total\_bedrooms median, log-transformed skewed features, removed outliers/duplicates.Objectives
3. Splitting: Applied train\_test\_split (80/20 split, random\_state=42
4. Gradient Descent : -
   1. w = w - α \* dw
   2. b = b - α \* db

Where , α = Learning Rate

dw = -2/n  
Start from here.

The main objectives of the project are:

* Determine the factors that influence the vaccination rates in the counties of a state.
* Understanding the unique characteristics of counties with high or low vaccination rates.
* Propose strategies for counties to reduce the infection rates.
* Explore the relation between vaccination rates and the level of infection and find out if the rates of COVID-19 vaccination reduces the rates of COVID-19 spread.
* Compare between the counties and determine the COVID-19 trends with the vaccination level.
* Analyze why some counties perform differently than others.

# Preliminary Literature Review (not final)

Research on COVID-19 vaccination efforts at the county level in Arizona highlights disparities influenced by demographics, healthcare access, and policy interventions.

1. **Urban-Rural Vaccination Disparities**
   * A study by the CDC found that, as of early 2022, COVID-19 vaccination rates in rural counties were significantly lower than in urban areas due to vaccine hesitancy, limited healthcare infrastructure, and geographic barriers (CDC, 2022).
   * In Arizona, counties like La Paz and Apache had some of the lowest vaccination rates, reflecting broader national trends in rural areas (CDC MMWR, 2022).  
     (<https://www.cdc.gov/mmwr/volumes/71/wr/mm7109a2.htm>)
2. **Equitable Access Initiatives in Arizona**
   * In response to disparities, health officials in counties such as Cochise and Pinal established mobile vaccination units to improve accessibility. These efforts increased uptake but faced challenges such as logistical limitations and staffing shortages (NIH, 2023).  
     (<https://pmc.ncbi.nlm.nih.gov/articles/PMC11425004/>)
3. **Data-Driven Approaches in Pima County**
   * Pima County leveraged real-time vaccination dashboards and partnerships with community organizations to ensure more targeted outreach. This approach effectively identified vaccine deserts and informed intervention strategies (Partners in Health, 2022).  
     (<https://www.pih.org/article/giving-county-health-leaders-right-data-boost-covid-19-vaccination-rates>)

## **Strengths and Weaknesses of Existing Research**

### Strengths

* **Comprehensive Overview of Urban-Rural Gaps:** Studies by the CDC and NIH provide a clear picture of urban versus rural disparities in COVID-19 vaccination rates across Arizona counties, showing the impact of geographic and socioeconomic factors.  
  ([CDC, 2022](https://www.cdc.gov/mmwr/volumes/71/wr/mm7109a2.htm?utm_source=chatgpt.com))
* **Evaluation of County-Level Health Interventions:** Research from NIH and public health agencies highlights county-specific strategies such as mobile vaccination units, targeted outreach in underserved areas, and collaborations with local healthcare providers.  
  ([NIH, 2023](https://pmc.ncbi.nlm.nih.gov/articles/PMC11425004/?utm_source=chatgpt.com))
* **Use of Real-Time Data Analytics:** Pima County’s success in using vaccination dashboards for targeted interventions serves as a model for other counties. Such data-driven approaches have helped pinpoint areas with lower vaccination rates and deploy resources effectively.  
  ([Partners in Health, 2022](https://www.pih.org/article/giving-county-health-leaders-right-data-boost-covid-19-vaccination-rates?utm_source=chatgpt.com))

### Weaknesses

* **Lack of Granular Demographic Data:** While studies provide a broad analysis of vaccination trends, there is a lack of in-depth research on the impact of race, income, education, and local healthcare infrastructure on vaccine uptake at the county level in Arizona.  
  ([NIH, 2023](https://pmc.ncbi.nlm.nih.gov/articles/PMC11425004/?utm_source=chatgpt.com))
* **Limited Focus on Vaccine Hesitancy Factors:** Studies acknowledge vaccine hesitancy as a challenge but do not explore the specific cultural, political, or misinformation-related factors that contribute to low vaccination rates in some Arizona counties.  
  ([CDC, 2022](https://www.cdc.gov/mmwr/volumes/71/wr/mm7109a2.htm?utm_source=chatgpt.com))
* **Insufficient Longitudinal Analysis:** Most research has focused on short-term vaccination efforts rather than long-term trends, booster uptake, and the sustainability of vaccination programs.  
  ([Partners in Health, 2022](https://www.pih.org/article/giving-county-health-leaders-right-data-boost-covid-19-vaccination-rates?utm_source=chatgpt.com))

# Methodology

This study analyzes COVID-19 vaccination trends at the county level in Arizona, focusing on disparities in vaccine distribution, accessibility, and intervention efforts. The research will utilize **quantitative data analysis**, **geospatial mapping**, and **correlation studies** to evaluate the impact of demographic and health care factors on vaccination rates. Data sources will primarily include the **Covid-19 Arizona County Vaccination Data** dataset and the **Covid-19 vaccination details** document, along with supplementary reports from health agencies.

**1. Data Collection**

We will collect and analyze the following datasets:

* **COVID-19 Vaccination Data**
  + Data on vaccine administration at the county level, including:
    - Total doses administered
    - Number of fully vaccinated individuals
    - Vaccination trends over time
* **Demographic and Socioeconomic Data**
  + Information on population characteristics such as:
    - Age distribution
    - Racial and ethnic composition
    - Median income and poverty rates
    - Education levels
* **Healthcare Infrastructure Data**
  + Availability of healthcare facilities, including:
    - Number of vaccination sites per county
    - Presence of mobile vaccination units
    - Number of hospitals, clinics, and pharmacies
* **Policy and Intervention Data**
  + Strategies implemented by Arizona counties, including:
    - County-level vaccine mandates and incentives
    - Targeted outreach programs
    - Mobile vaccination efforts

**2. Data Analysis**

A combination of statistical and geospatial methods will be used to examine vaccination trends and their correlation with demographic and health care factors.

* **Descriptive Statistics**
  + Calculate mean, median, and standard deviation for vaccination rates across counties.
  + Identify counties with exceptionally high or low vaccination rates.
* **Geospatial Analysis (GIS Mapping)**
  + Use GIS tools to visualize county-level vaccination rates.
  + Identify "vaccine desserts" with limited access to healthcare facilities.
* **Regression and Correlation Analysis**
  + Multiple regression models will analyze the relationship between vaccination rates (dependent variable) and:
    - Population density
    - Socioeconomic factors (income, education)
    - Healthcare infrastructure (vaccination sites per capita)
    - Policy interventions (mobile clinics, mandates)
  + Correlation tests will examine associations between vaccine hesitancy factors and vaccination rates.
* **Time-Series Analysis** *(if data permits)*
  + Evaluate trends in vaccine uptake before and after policy interventions.
  + Identify seasonal or event-driven fluctuations in vaccination rates.

**3. Limitations**

This study acknowledges the following limitations:

* **Data Completeness and Accuracy**
  + Inconsistent reporting across counties may result in missing or incomplete data.
  + Vaccine hesitancy data may not fully capture cultural and political influences.
* **Geographic Disparities**
  + Rural counties may have less reliable data due to lower reporting capabilities.

**4. Ethical Considerations**

* All data used in this study is sourced from publicly available records.
* No personally identifiable information will be collected.
* Findings will be reported objectively, ensuring transparency and accuracy.

# References