HOUSE PRICE PREDICTION

Data 200 – Applied Statistical Analysis

INTRODUCTION

Objective:

- Predict median house values based on location, neighborhood characteristics, and economic factors.

Dataset Overview:

- Source: California Housing Dataset (housing.csv)
- Features:
  - Geographic (latitude, longitude, ocean proximity)
  - Neighborhood (rooms, population, households)
  - Economic (median income)
- Target Variable: median_house_value

A1 | longitude

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity									
2	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY									
3	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY									
4	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY									
5	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY									
6	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY									
7	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY									
8	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY									
9	-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY									
10	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY									
11	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY									
12	-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY									
13	-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY									
14	-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY									
15	-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY									
16	-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY									
17	-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY									
18	-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY									
19	-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY									
20	-122.26	37.84	50	2239	455	990	419	1.9911	158700	NEAR BAY									
21	-122.27	37.84	52	1503	298	690	275	2.6033	162900	NEAR BAY									
22	-122.27	37.85	40	751	184	409	166	1.3578	147500	NEAR BAY									
23	-122.27	37.85	42	1639	367	929	366	1.7135	159800	NEAR BAY									
24	-122.27	37.84	52	2436	541	1015	478	1.725	113900	NEAR BAY									
25	-122.27	37.84	52	1688	337	853	325	2.1806	99700	NEAR BAY									
26	-122.27	37.84	52	2224	437	1006	422	2.6	132600	NEAR BAY									
27	-122.28	37.85	41	535	123	317	119	2.4038	107500	NEAR BAY									
28	-122.28	37.85	49	1130	244	607	239	2.4597	93800	NEAR BAY									
29	-122.28	37.85	52	1898	421	1102	397	1.808	105500	NEAR BAY									
30	-122.28	37.84	50	2082	492	1131	473	1.6424	108900	NEAR BAY									
31	-122.28	37.84	52	729	160	395	155	1.6875	132000	NEAR BAY									
32	-122.28	37.84	49	1916	447	863	378	1.9274	122300	NEAR BAY									
33	-122.28	37.84	52	2153	481	1168	441	1.9615	115200	NEAR BAY									
34	-122.27	37.84	48	1922	409	1026	335	1.7969	110400	NEAR BAY									
35	-122.27	37.83	49	1655	366	754	329	1.375	104900	NEAR BAY									
36	-122.27	37.83	51	2665	574	1258	536	2.7303	109700	NEAR BAY									
37	-122.27	37.83	49	1215	282	570	264	1.4861	97200	NEAR BAY									
38	-122.27	37.83	48	1798	432	987	374	1.0972	104500	NEAR BAY									
39	-122.28	37.83	52	1511	390	901	403	1.4103	103900	NEAR BAY									
40	-122.26	37.83	52	1470	330	689	309	3.48	191400	NEAR BAY									

LITERATURE REVIEW

1. **California Housing Market Trends**
   - Median income strongly correlates with housing prices.
2. **Geographic Influence on Pricing**
   - Coastal properties are significantly more expensive than inland homes.
3. **Feature Importance in Real Estate**
   - Studies confirm that location, income, and housing density are top predictors.

DATA CLEANING & PREPROCESSING

1. **Handling Missing Values**
   - Dropped rows with missing total_bedrooms (~1% of data).
2. **Outlier Removal**
   - Used IQR method to remove extreme values in:
     - median_house_value, total_rooms, total_bedrooms, median_income.
3. **Categorical Encoding**
   - Converted ocean_proximity into dummy variables.

EXPLORATORY DATA ANALYSIS (EDA)

1. **Median Income vs. House Value**
    - Strong positive correlation ($r \approx 0.69$).
2. **Ocean Proximity Impact**
    - Homes near the ocean are significantly more expensive.
3. **Room Count vs. Value**
    - More rooms do not necessarily mean higher prices (weak correlation).

FEATURE SELECTION

Recursive Feature Elimination (RFE) Results:

- Selected top 6 features:
  - median_income
  - housing_median_age
  - population
  - households
  - <1H OCEAN
  - INLAND

Why RFE?

- Helps avoid overfitting.
- Focuses on the most predictive features.

MODEL SELECTION & HYPOTHESIS TESTING

Chosen Model: Linear Regression (OLS)

- Best for continuous target variables (median_house_value).
- Provides interpretable coefficients.

Hypotheses Tested:

1. **H1:** Higher median income → Higher house values (**Supported**).
2. **H2:** Coastal homes > Inland homes (**Supported**).
3. **H3:** Older homes have lower values (**Needs further testing**).

THANK YOU