HOUSE PRICE PREDICTION

By Grace Gan

The Problem:

House prices fluctuate frequently and form significant bubbles, misleading both investors and buyers.

The Data: (1460,81)

Id	MSSubClas	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContou	Utilities	LotConfig	LandSlope	Neighborhc	Condition1	Condition2	BldgType	HouseStyle	OverallQua	OverallCon	YearBui
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5	5	1
7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	2
8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7	6	1
9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7	5	1
10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1.5Unf	5	6	1
11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	5	1
12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5	2
13	20	RL	NA	12968	Pave	NA	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1
14	20	RL	91	10652	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	2
15	20	RL	NA	10920	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	1
16	45	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Unf	7	8	1
17	20	RL	NA	11241	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	7	1
18	90	RL	72	10791	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5	1
19	20	RL	66	13695	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	RRAe	Norm	1Fam	1Story	5	5	2
20	20	RL	70	7560	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1
21	60	RL	101	14215	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	2
22	45	RM	57	7449	Pave	Grvl	Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Unf	7	7	1
23	20	RL	75	9742	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	2
24	120	RM	44	4224	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1
25	20	RL	NA	8246	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	8	1
26	20	RL	110	14230	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	2
27	20	RL	60	7200	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	1
28	20	RL	98	11478	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	2
29	20	RL	47	16321	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1
30	30	RM	60	6324	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr	RRNn	1Fam	1Story	4	6	1
31	70	C (all)	50	8500	Pave	Pave	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Feedr	Norm	1Fam	2Story	4	4	1
32	20	RL	NA	8544	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1
33	20	RL	85	11049	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	2
34	20	RL	70	10552	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	5	1

Data Wrangling

- Summary Statistics: Generated summary statistics for both numerical and categorical features. This helped us understand the distribution and central tendencies of the data.

- Identifying Missing Values: Identified missing values by calculating the percentage of missing data for each feature. This allowed us to determine which features had significant amounts of missing data.

- Removing High Null Value Features: Removed the top 5 features with the highest percentage of missing values to improve model accuracy. These features were: 'PoolQC' 、'MiscFeature'、'Alley' 、'Fence' 、'FireplaceQu'

Exploratory Data Analysis

Distribution of dwelling types and their relation to sale prices?

Does zoning impact sale price?

Does street type effect on sale price?

What is the Average sale price by property shape?

Is there a Correlation between Property Age and Sale Price

Is there a Correlation between Living Area and Sale Price

Does price change year to year

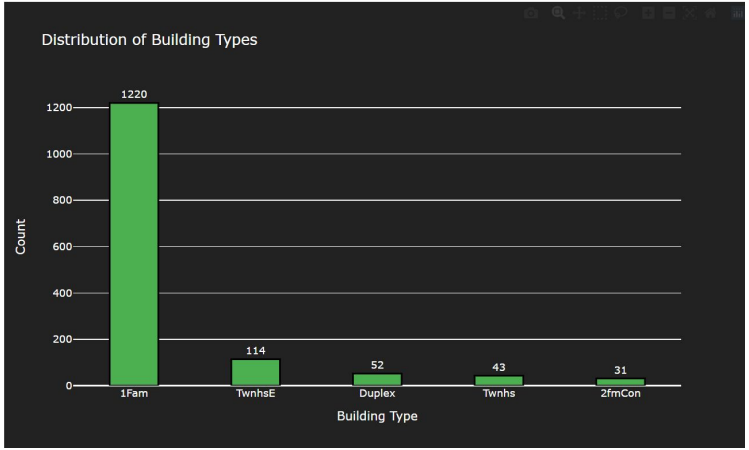

Figure1: Distribution of dwelling types and their relation to sale prices?

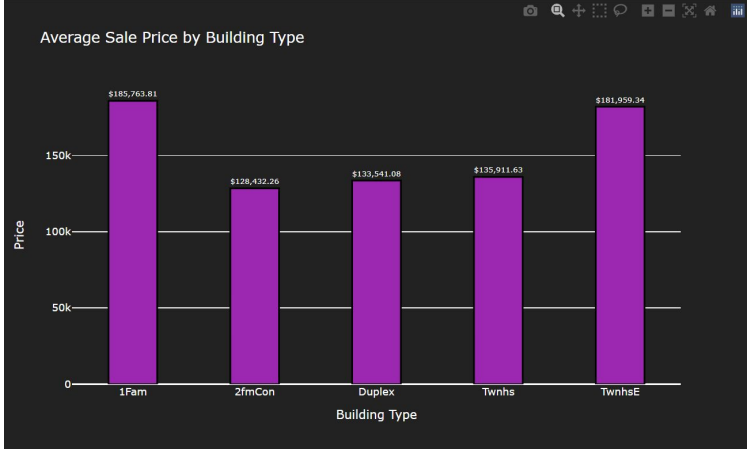

Figure 2 : Distribution of dwelling types and their relation to sale prices?

Figure3: Is there correlation between Property Age and Sale Price?

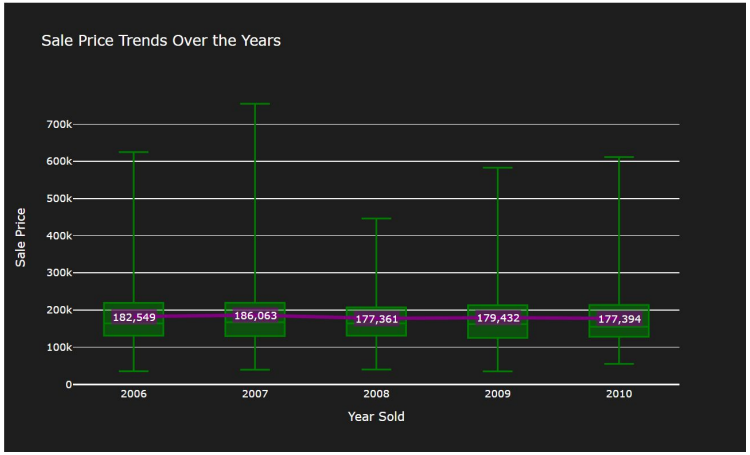

Figure4: Does the price change year to year?

Model Selection

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor

Hyperparameter Tuning (Grid Search Cross Validation)

- Linear Regression : 0.134
- Random Forest Regressor : 0.150
- XGBoost Regressor : 0.135

Conclusions:

- Linear regression model achieved an average error of 13.39% relative to actual house prices.
- By incorporating various housing features, this model can effectively estimate property values, thereby mitigating market bubbles.
- This predictive capability supports real estate agents, homebuyers, and financial institutions in making data-driven decisions.

Future Work

- Model Evaluation Metrics: Using percentage metric
- Feature Engineering: Transforming the 'year' feature into a categorical variable
- Data Volume: Expanding the dataset beyond the current 1460 rows

Thank You!

Questions?