House Sales Price Prediction in King County, USA

# Problem Statement

Real estate pricing is influenced by various factors such as location, size, age, and additional features of the property. King County, which includes Seattle, has a diverse housing market. This project aims to:  
  
Build machine learning models to predict the sale price of houses in King County based on their features.

# Dataset Description

Dataset Name: kc\_house\_data.csv

Total Records: 21,613

Total Features: 21 (including target variable 'price')

## Code: Importing Libraries

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
from sklearn.linear\_model import LinearRegression, Ridge  
from sklearn.preprocessing import PolynomialFeatures, StandardScaler  
from sklearn.pipeline import Pipeline  
from sklearn.model\_selection import train\_test\_split  
from sklearn.metrics import r2\_score

## Code: Load Dataset

df = pd.read\_csv('kc\_house\_data.csv')  
df.head()

# Data Cleaning

1. Dropped columns: 'id', 'Unnamed: 0'.  
2. Filled missing values in 'bedrooms' and 'bathrooms'.

df.drop(['id', 'Unnamed: 0'], axis=1, inplace=True)  
df['bedrooms'].fillna(df['bedrooms'].mean(), inplace=True)  
df['bathrooms'].fillna(df['bathrooms'].mean(), inplace=True)

# Exploratory Data Analysis Summary

- 'sqft\_living' and 'grade' are highly correlated with price.  
- Waterfront properties have higher outlier prices.

# Model Development

## 1. Simple Linear Regression

X = df[['sqft\_living']]  
y = df['price']  
slr = LinearRegression()  
slr.fit(X, y)  
print(slr.score(X, y)) # R² ≈ 0.49

## 2. Multiple Linear Regression

features = ['sqft\_living', 'bedrooms', 'bathrooms', 'floors', 'grade']  
X = df[features]  
y = df['price']  
mlr = LinearRegression()  
mlr.fit(X, y)  
print(mlr.score(X, y)) # R² ≈ 0.65

## 3. Polynomial Regression with Pipeline

Input = [('scale', StandardScaler()), ('poly', PolynomialFeatures(degree=2)), ('model', LinearRegression())]  
pipe = Pipeline(Input)  
pipe.fit(X, y)  
print(pipe.score(X, y)) # R² ≈ 0.77

## Model Evaluation - Ridge Regression

rr = Ridge(alpha=0.1)  
rr.fit(X, y)  
print(rr.score(X, y)) # R² ≈ 0.67

## Polynomial Ridge Regression (deg=2)

pipe\_ridge = Pipeline([('scale', StandardScaler()), ('poly', PolynomialFeatures(degree=2)), ('model', Ridge(alpha=0.1))])  
pipe\_ridge.fit(X, y)  
print(pipe\_ridge.score(X, y)) # R² ≈ 0.73

# Conclusion

This project successfully explored and modeled housing data from King County. Key takeaways:  
  
- Strongest predictors of price: 'sqft\_living', 'grade', 'bathrooms'.  
- Polynomial regression gives the best performance (R² ≈ 0.77).  
- Ridge regularization helps avoid overfitting.  
- Final recommendation: Use Polynomial Regression with StandardScaler for better accuracy.