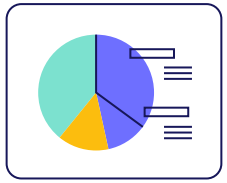
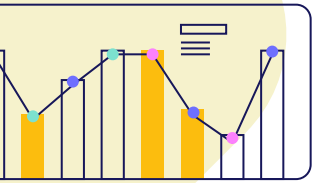
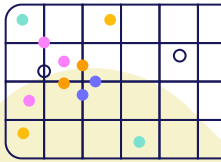


What determinants most and least significantly affect housing prices?

CS555 Term Project – Chen-Wei Hsu



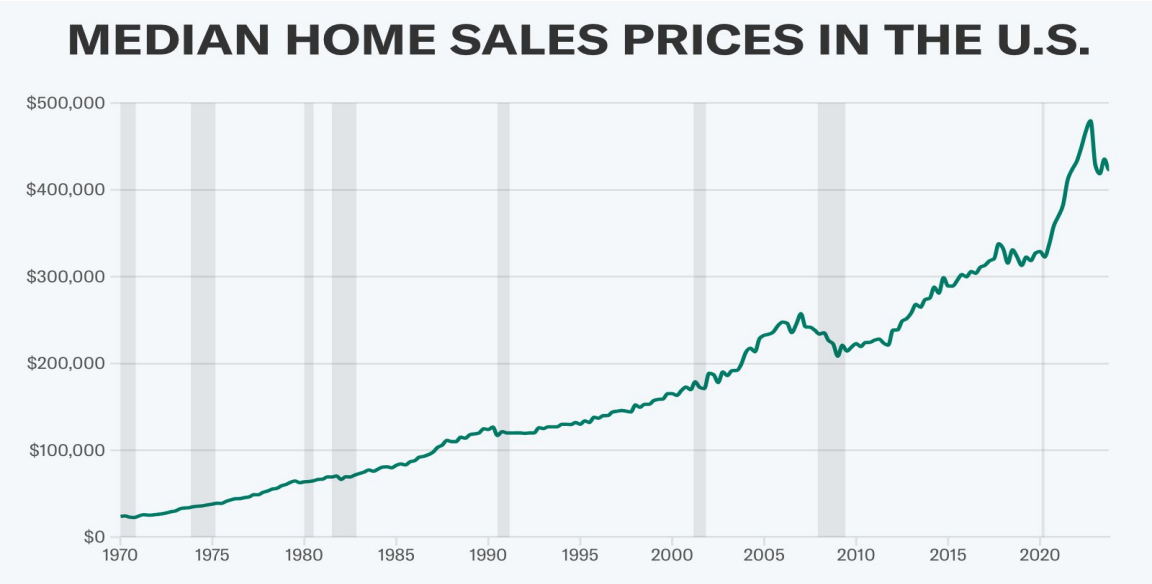


01.

Introduction and Research Scenario

Why this topic?

According to Yahoo Finance, median home prices have steadily risen over the years, and there is a dramatic jump in prices beginning in 2020. Prices peaked in the fourth quarter of 2022 and then fell sharply.



02.

Describe The Data Set and Statistical Methods

Data Introduction

Real_Estate

No	X1 house age	X2 distance to the nearest MRT station	X3 number of convenience stores	X4 latitude	X5 longitude	Y house price of unit area
1	32	84.87882	10	24.98298	121.54024	37.9
2	19.5	306.5947	9	24.98034	121.53951	42.2
3	13.3	561.9845	5	24.98746	121.54391	47.3
4	13.3	561.9845	5	24.98746	121.54391	54.8
5	5	390.5684	5	24.97937	121.54245	43.1
6	7.1	2175.03	3	24.96305	121.51254	32.1
7	34.5	623.4731	7	24.97933	121.53642	40.3
8	20.3	287.6025	6	24.98042	121.54228	46.7
9	31.7	5512.038	1	24.95095	121.48458	18.8
10	17.9	1783.18	3	24.96731	121.51486	22.1
11	34.8	405.2134	1	24.97349	121.53372	41.4
12	6.3	90.45606	9	24.97433	121.5431	58.1
13	13	492.2313	5	24.96515	121.53737	39.3
14	20.4	2469.645	4	24.96108	121.51046	23.8
15	13.2	1164.838	4	24.99156	121.53406	34.3
16	35.7	579.2083	2	24.9824	121.54619	50.5
17	0	292.9978	6	24.97744	121.54458	70.1
18	17.7	350.8515	1	24.97544	121.53119	37.4
19	16.9	368.1363	8	24.9675	121.54451	42.3
20	1.5	23.38284	7	24.96772	121.54102	47.7
21	4.5	2275.877	3	24.96314	121.51151	29.3
22	10.5	279.1726	7	24.97528	121.54541	51.6
23	14.7	1360.139	1	24.95204	121.54842	24.6
24	10.1	279.1726	7	24.97528	121.54541	47.9

Y = house price of a unit area

X1 = house age

X2 = distance to the nearest MRT station

X3 = number of convenience stores

X4 = latitude

X5 = longitude

Statistical Methods

Independent variables: house age, distance to the nearest MRT station, number of convenience stores, latitude, and longitude.

Dependent variable being the house price of a unit area.

I will conduct further analysis using sequential sum-of-squares, which reveals insightful findings.

Furthermore, through additional experiments, the original linear regression model was refined by excluding a variable.

Example (1/3):

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-34.546  -5.267  -1.600   4.247  76.372

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.946e+03  6.211e+03  -0.796    0.426
X1          -2.689e-01  3.900e-02  -6.896 2.04e-11 ***
X2          -4.259e-03  7.233e-04  -5.888 8.17e-09 ***
X3           1.163e+00  1.902e-01   6.114 2.27e-09 ***
X4           2.378e+02  4.495e+01   5.290 2.00e-07 ***
X5          -7.805e+00  4.915e+01  -0.159    0.874
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.965 on 408 degrees of freedom
Multiple R-squared:  0.5712,    Adjusted R-squared:  0.5659
F-statistic: 108.7 on 5 and 408 DF,  p-value: < 2.2e-16
```

The regression equation derived from this model is Y = -4946 - 0.2689 (X1) -4.259 (10^-3) (X2) - 1.163 (X3) + 237.8 (X4) - 7.805 (X5)

Example (2/3):

Check whether there is evidence of a linear relationship between the dependent variable and the independent variable.

At $\alpha = 0.05$:

$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

From the summary table: the P-value is 0.00, which is small, and smaller than $\alpha = 0.05$. Thus, H_0 is rejected. There is sufficient evidence of a linear relationship between house age and house price of a unit area.

Example (3/3)

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq  F value    Pr(>F)
X1          1   3390    3390  42.1840 2.420e-10 ***
X2          1  34164   34164 425.0967 < 2.2e-16 ***
X3          1   3817    3817  47.4910 2.106e-11 ***
X4          1   2299    2299  28.6106 1.478e-07 ***
X5          1      2       2   0.0252    0.8739
Residuals 408  32790      80
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Check collinearity:

```
      X1       X2       X3       X4       X5
1.014249 4.282985 1.613339 1.599017 2.923881
```

03. Report the Results

Results

Including all independent variable	Most significant: house age, Least significant: longitude
Excluding house age	Most significant: distance to the nearest MRT station, Least significant: longitude
Excluding distance to the nearest MRT station	Most significant: number of convenience stores, Least significant: longitude
Excluding number of convenience stores	Most significant: distance to the nearest MRT station, Least significant: longitude
Excluding latitude	Most significant: distance to the nearest MRT station, Least significant: longitude
Excluding longitude	Most significant: distance to the nearest MRT station, Least significant: longitude

Conclusion

Based on the analysis, there are four independent variables significant to the linear model, which are house age, distance to the nearest MRT station, number of convenience stores, and latitude. We can say they are all tied to the most significant factors affecting housing prices. House age negatively impacts price, as newer houses typically cost more. Distance to the nearest MRT station negatively impacts price, as houses close to MRT stations cost more. The number of convenience stores positively impacts price, as more convenience stores make it a more appealing location to live in, making the house cost more. Latitude positively impacts price because it's colder, so the house needs to be built with better quality materials, which increases the price. On the other hand, factors like longitude showed weaker or no consistent impact on housing prices.

Reference

Bundrick, H. (2024, November 19). Why are home prices so high?. Yahoo! Finance. https://finance.yahoo.com/personal-finance/why-are-house-prices-so-high-184935574.html

Thank you!