STAT 155 Final Data Analysis Project

Draft

Xi Feng, He Lian, Yunyang Zhong, Chenxin Zhu

12/5/2019

Title

A study of house age and house price in Taiwan

Scientific background

The overall scientific goal is to assess the association between house age and house prices in Taiwan.

Specifically, we will address the following questions: 1. Is house age associated with decreased house prices in Taiwan? 2. Is house age associated with decreased house prices in Taiwan with latitude as a confounder? 3. If there is an association between house age and decreased house-price, is this effect different when the distance to the nearest MRT station varies with latitude as a confounder? 4. (Potential/if time) If there is an association between house age and decreased house-price, is this effect different when the number of convenience stores in the living circle on foot varies with latitude as a confounder?

All the columns in this data set are important factors that people tend to consider when comparing different houses. Theoretically, people are willing to pay more to get the better in each variable and thus maximize their utility. By exploring and analyzing these data using regression models, we can determine if these factors can truly influence house prices and the extent to which each factor influences house prices.

Because houses are relatively expensive but essential goods, when people want to buy houses, they tend to consider many factors related to house conditions and living environment. Theoretically, if a house has good conditions and living environment, people are willing to pay more to get it.

Description of study/data

This data set was collected from Taiwan in 2012 and 2013 by Prof. I-Cheng Yeh and was used to build real estate valuation models in his paper(Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271).

The data include transaction date, house age(year), distance to the nearest metro station(meter), number of convenience stores in the living circle on foot, latitude(degree), longitude(degree), and house price of unit area(10000 New Taiwan Dollar per Ping, where Ping is a local unit, 1 Ping = 3.3 meters squared).

The link of the data set is provided below:
https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set (https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set)

This study was observational because subjects included in this study were not randomly assigned to buy different types of houses.

House age is our predictor of interest while house price is our outcome. The underlying hypothesis is that people tend to be more willing to purchase new houses. We consider latitude as a confounder as it can be associate with house age (more advanced development in the North of Taiwan). House age cannot cause its latitude and higher latitudes may lead to higher prices. The distance from metro stations as well as number of convenience stores around are two potential effect modifieres in our model. They are not confounders since they are not associated with house age. However, they are important factors consumers consider before making a final choice. We believe that the relationship between house age and house price changes as the distance from metro stations and the number of convenience stores change.

Statistical Methods and Results

Describing any modifications to your data

```
library("tidyverse")
data <- read_csv("RealEstate.csv")
```

This step is to create a categorical variable from a quantitative variable. The quantative variable distance is divided into three categories: a walking distance from 0 to 1500 meters, a biking distance from 1500 to 3000 meters, and a driving distance farther than 3000 meters.

```
dis <- data %>%
  select(distance)
distanceCat <- 0
distanceCat[dis <= 1500] <- "Walk"
distanceCat[dis > 1500 & dis <= 3000] <- "Bike"
distanceCat[dis > 3000] <- "Drive"
data <- data %>%
  mutate(distanceCat)
```

This step is to create a categorical variable from a quantitative variable. The quantative variable store is divided into four categories: None for 0 stores, Few for 1-3 stores, Some for 4-6 stores, and Many for more than 6 stores.

```
store <- data %>%
  select(store)
storeCat <- 0
storeCat[store == 0] <- "None"
storeCat[store > 0 & store <= 3] <- "Few"
storeCat[store > 3 & store <= 6] <- "Some"
storeCat[store > 6] <- "Many"
data <- data %>%
  mutate(storeCat)
```

This step is to create a categorical variable from a quantitative variable. The quantative variable latitude is divided into three categories: low for latitudes smaller or equal to 24.96, medium for latitudes from 24.96 to 24.98, and high for latitudes larger than 24.98.

```
lat <- data %>%
  select(latitude)
latitudeCat <- 0
latitudeCat[lat<=24.96] <- "Low"
latitudeCat[24.96<lat & lat<=24.98] <- "Medium"
latitudeCat[24.98<lat] <- "High"
data <- data %>%
  mutate(latitudeCat)
```

Describing your outcome variable(s)

```
anyNA(data)
```

```
## [1] FALSE
```

```
data %>%
  select(price) %>%
  summary()
```

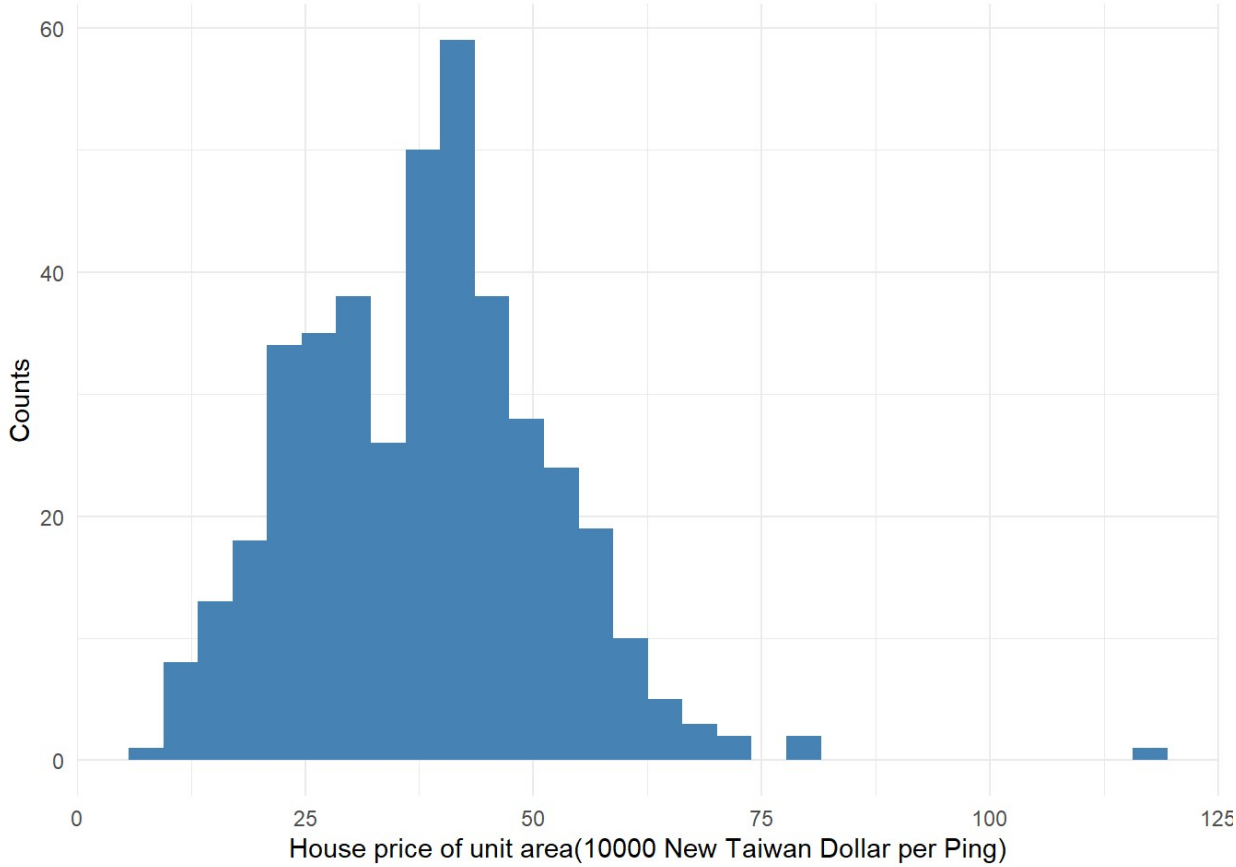
```
##      price
##  Min.   :  7.60
##  1st Qu.: 27.70
##  Median : 38.45
##  Mean   : 37.98
##  3rd Qu.: 46.60
##  Max.   :117.50
```

```
data %>%
  summarize(sd(price))
```

```
## # A tibble: 1 x 1
##   `sd(price)`
##         <dbl>
## 1        13.6
```

```
data %>%
    ggplot(aes(x = price)) +
    geom_histogram(fill = "steelblue") +
    xlab('House price of unit area(10000 New Taiwan Dollar per Ping)') +
    ylab('Counts') +
    theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There is no missing values in our dataset. For our outcome variable “price”, the range is from 7.60 to 117.50 (10000 New Taiwan Dollar). The mean of “price” is 37.98 (10000 New Taiwan Dollar). The standard deviation is 13.6 (10000 New Taiwan Dollar).

Describing your predictor(s) of interest

```
anyNA(data)
```

```
## [1] FALSE
```

```
data %>%
  select(age) %>%
  summary()
```

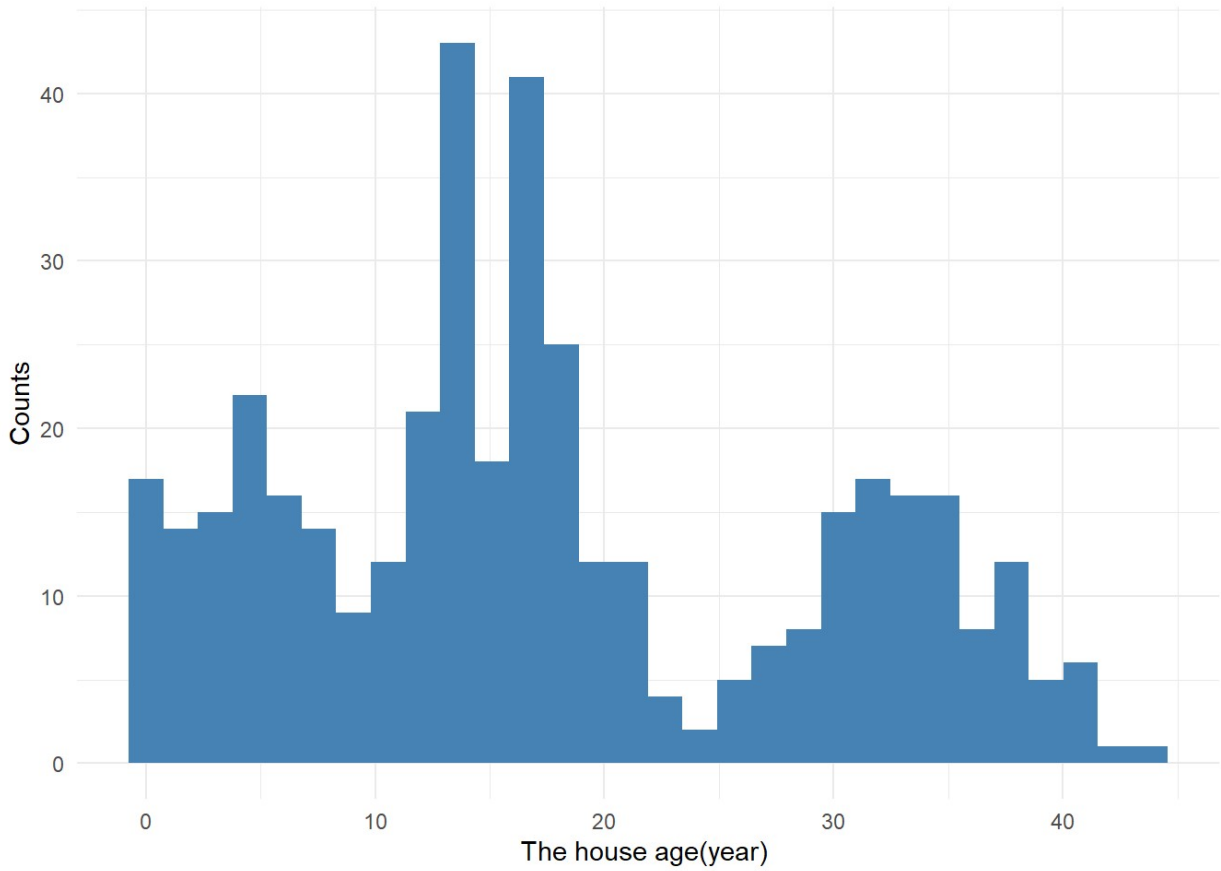
```
##       age
##  Min.   : 0.000
##  1st Qu.: 9.025
##  Median :16.100
##  Mean   :17.713
##  3rd Qu.:28.150
##  Max.   :43.800
```

```
data %>%
  summarize(sd(age))
```

```
## # A tibble: 1 x 1
##   `sd(age)`
##       <dbl>
## 1      11.4
```

```
data %>%
    ggplot(aes(x = age)) +
    geom_histogram(fill = "steelblue") +
    xlab('The house age(year) ') +
    ylab('Counts') +
    theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

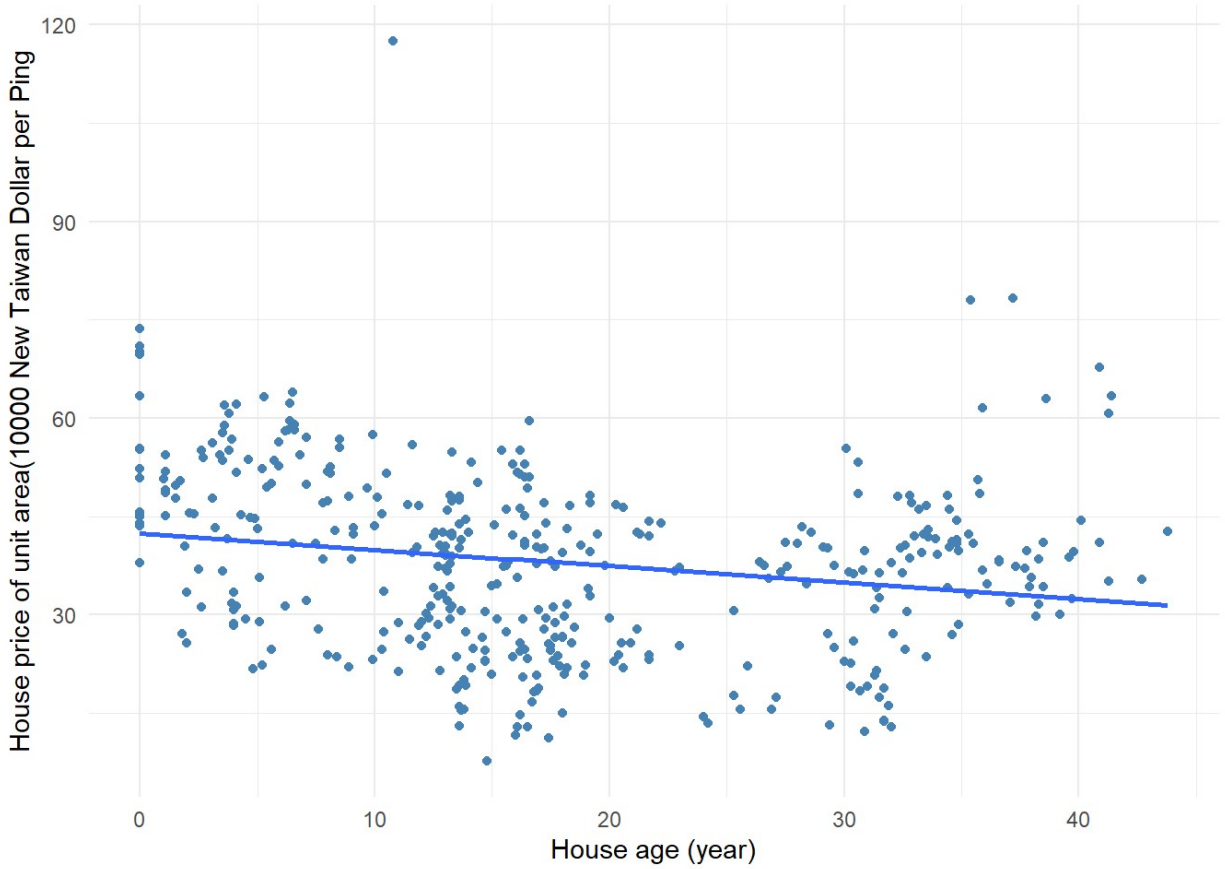


Missing observations: no missing values Range: 0.000 to 43.800 (year)
Mean: 17.713(year) Standard deviation: 11.392(year)

Answering your first scientific question

Descriptive

```
data %>%
  ggplot(aes(x = age, y = price)) +
  geom_point(color = 'steelblue') +
  geom_smooth(method = 'lm', se = FALSE) +
  xlab('House age (year)') +
  ylab('House price of unit area(10000 New Taiwan Dollar per Ping') +
  theme_minimal()
```




```
data %>%
  summarize(cor(age,price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.211
```

> The correlation coefficient is -0.211, so the direction of relationship seems to be slightly negative, and the form is generally linear. The compactness around the average relationship is relatively weak. For unusual features, there is one outlier on the top-left of the graph.

Inferential

```
lm(price ~ age, data = data)  %>%
  summary()
```

```
##
## Call:
## lm(formula = price ~ age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.113 -10.738   1.626   8.199  77.781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.43470    1.21098  35.042  < 2e-16 ***
## age         -0.25149    0.05752  -4.372 1.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.32 on 412 degrees of freedom
## Multiple R-squared:  0.04434,    Adjusted R-squared:  0.04202
## F-statistic: 19.11 on 1 and 412 DF,  p-value: 1.56e-05
```

```
lm(price ~ age, data = data) %>%
  confint()
```

```
##                 2.5 %    97.5 %
## (Intercept) 40.0542333 44.815161
## age         -0.3645609 -0.138416
```

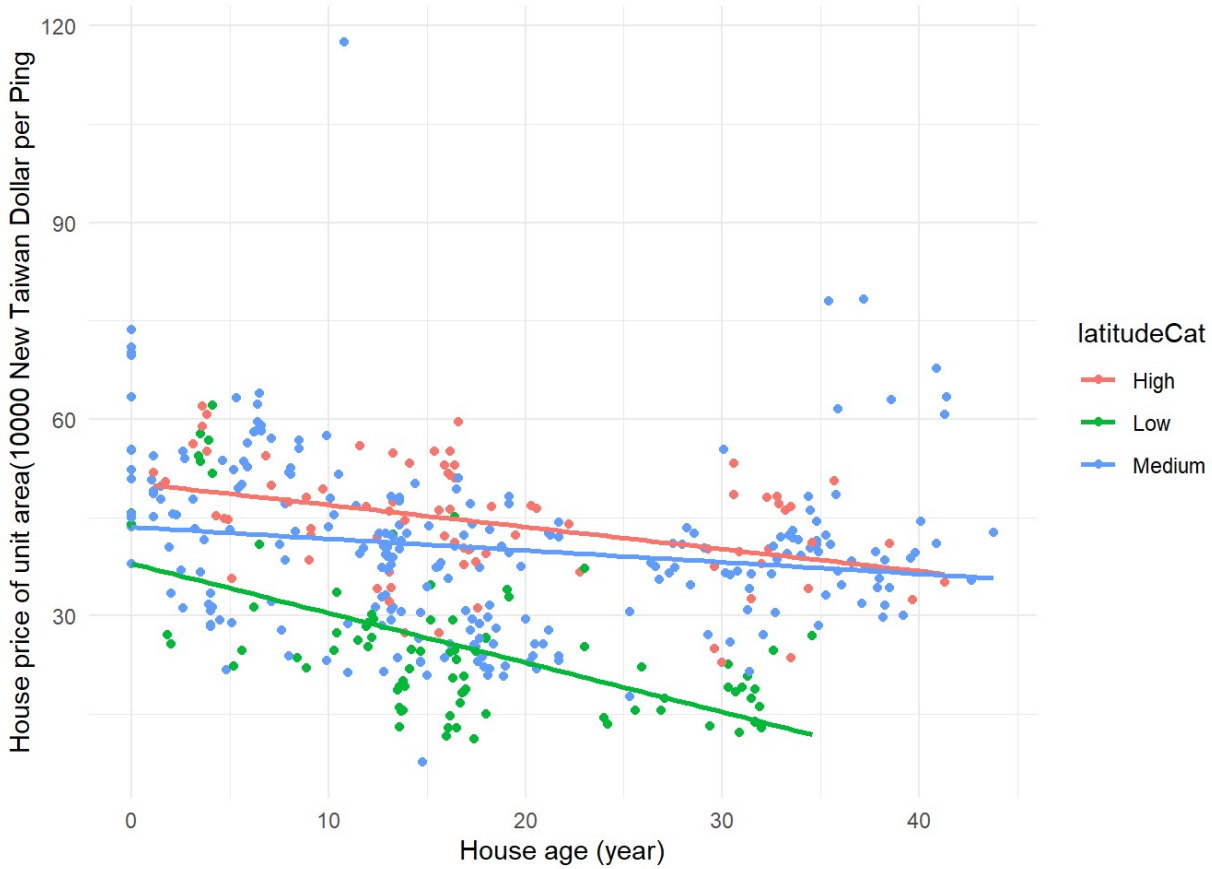
For all of our regression analyses, we use linear regression models because our outcomes are quantitative. We do not need to worry about probabilities out of the range of 0 to 1 as we talk about differences only. Also, logistic regression models involve odds ratios which are hard to interpret.

p-value for the age coefficient is 1.56e-05. Assuming H_0 is ture, the probability of observing a test statistic which is " as or even more extreme" than -4.372 is 1.56e-05. 95% confidence interval: Assuming the sampling distribution model is accurate, we are 95% confident that the true difference of average price(10000 New Taiwan Dollar) comparing two groups of houses that differ in 1 year is between -0.365 and -0.138.

Answering your second scientific question

Descriptive

```
data %>%
  ggplot(aes(x = age, y = price, color = latitudeCat)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  xlab('House age (year)') +
  ylab('House price of unit area(10000 New Taiwan Dollar per Ping') +
  theme_minimal()
```




```
data %>%
    filter(latitudeCat == "Low") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.585
```

```
data %>%
    filter(latitudeCat == "Medium") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.172
```

```
data %>%
    filter(latitudeCat == "High") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.405
```

In the graph we can observe that houses with low latitude (<24.96 degree) have a lower average house price of unit area. The correlation coefficient is -0.585, which means a moderate negative linear relationship. Houses with medium latitude (24.96, 24.98 degree) have a generally higher average unit area price. The correlation coefficient is -0.172, a weak negative linear relationship. And houses with high latitude (>24.98 degree) have the highest average unit area price, with a correlation coefficient of -0.405, a moderate negative lienar relationship.

Inferential

```
lm(price ~ age+latitude,  data = data)  %>%
  summary()
```

```
##
## Call:
## lm(formula = price ~ age + latitude, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.722  -6.731  -1.195   5.598  74.114
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.527e+04  1.085e+03 -14.076  < 2e-16 ***
## age         -2.878e-01  4.733e-02  -6.081 2.73e-09 ***
## latitude     6.133e+02  4.345e+01  14.116  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.94 on 411 degrees of freedom
## Multiple R-squared:  0.3564, Adjusted R-squared:  0.3532
## F-statistic: 113.8 on 2 and 411 DF,  p-value: < 2.2e-16
```

```
lm(price ~ age+latitude,  data = data) %>%
  confint()
```

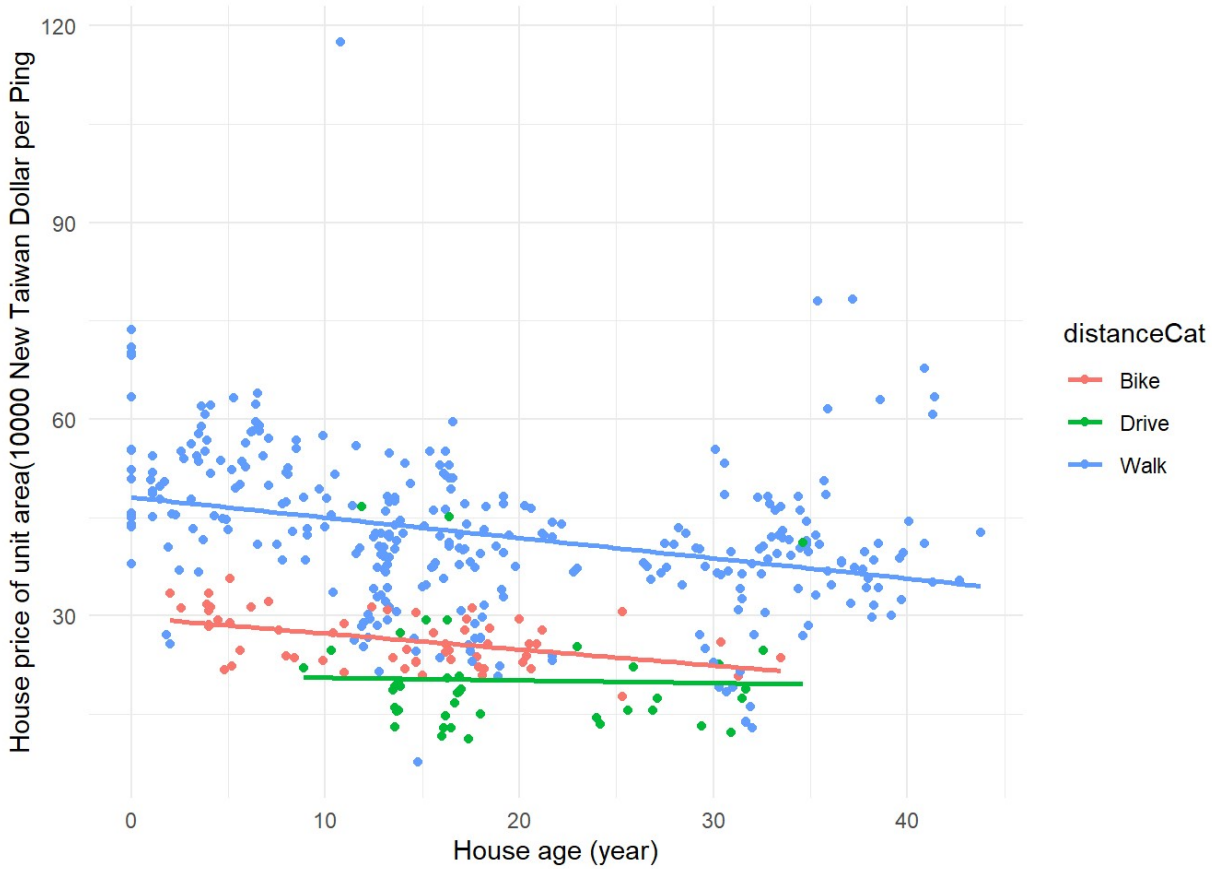
```
##                     2.5 %        97.5 %
## (Intercept) -1.740430e+04 -1.313897e+04
## age         -3.808949e-01 -1.948022e-01
## latitude     5.279324e+02  6.987644e+02
```

p-value for the age coefficient is 2.73e-09. Assuming H_0 is ture, the probability of observing a test statistic which is " as or even more extreme" than -6.081 is 2.73e-09. 95% confidence interval: Assuming the sampling distribution model is accurate, we are 95% confident that holding the latitude of the houses constant, the true difference of average price (10000 New Taiwan Dollar) comparing two groups of houses that differ in 1 year is between -0.381 and -0.195.

Answering your third scientific question

Descriptive

```
data %>%
  ggplot(aes(x = age, y = price, color = distanceCat)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  xlab('House age (year)') +
  ylab('House price of unit area(10000 New Taiwan Dollar per Ping') +
  theme_minimal()
```




```
data %>%
    filter(distanceCat == "Walk") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.315
```

```
data %>%
    filter(distanceCat == "Bike") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.461
```

```
data %>%
    filter(distanceCat == "Drive") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1           -0.0343
```

> In this graph, houses with biking distance to the nearest MRT station have a median average house price, and the correlation coefficient is -0.461 which means a moderate negative linear relationship. Houses with driving distance to the nearest MRT station have a low average house price, and the correlation coefficient is -0.0342, a weak negative linear relationship. Houses with walking distance to nearest MRT station have a high average house price, with a correlation coefficient of -0.315, a weak negative linear relationship.

Inferential

```
lm(price ~ age+latitude+distance+age:distance,  data = data) %>%
  summary()
```

```
##
## Call:
## lm(formula = price ~ age + latitude + distance + age:distance,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.312  -4.877  -0.691   4.254  70.969
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.634e+03  1.136e+03  -5.839 1.07e-08 ***
## age          -3.581e-01  4.970e-02  -7.205 2.81e-12 ***
## latitude      2.677e+02  4.550e+01   5.885 8.31e-09 ***
## distance     -8.647e-03  9.556e-04  -9.049  < 2e-16 ***
## age:distance  1.546e-04  4.301e-05   3.596 0.000363 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.213 on 409 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.5415
## F-statistic: 122.9 on 4 and 409 DF,  p-value: < 2.2e-16
```

```
lm(price ~ age+latitude+distance+age:distance,  data = data) %>%
  confint()
```

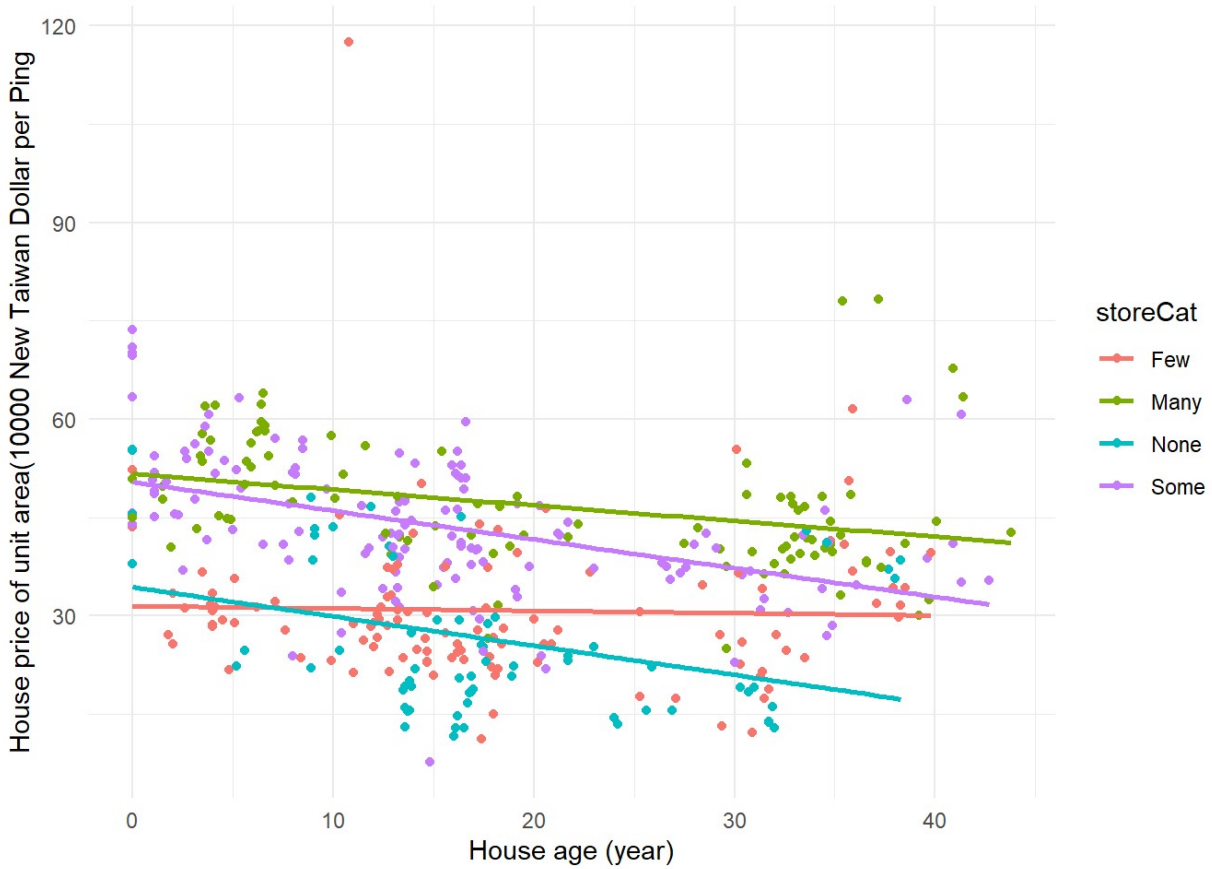
```
##                      2.5 %        97.5 %
## (Intercept)  -8.867862e+03 -4.400676e+03
## age          -4.557865e-01 -2.603964e-01
## latitude      1.782944e+02  3.571640e+02
## distance     -1.052572e-02 -6.768594e-03
## age:distance  7.009198e-05  2.391746e-04
```

p-value for the age coefficient is 2.81e-12. Assuming H_0 is ture, the probability of observing a test statistic which is " as or even more extreme" than -7.205 is 2.81e-12. 95% confidence interval: Assuming the sampling distribution model is accurate, we are 95% confident that holding the latitude of the houses and the distance to the nearest MRT station constant, the true difference of average price (10000 New Taiwan Dollar) comparing two groups of houses that differ in 1 year is between -0.456 and -0.260.

Answering your fourth scientific question

Descriptive

```
data %>%
  ggplot(aes(x = age, y = price, color = storeCat)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  xlab('House age (year)') +
  ylab('House price of unit area(10000 New Taiwan Dollar per Ping') +
  theme_minimal()
```




```
data %>%
    filter(storeCat == "None") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.355
```

```
data %>%
    filter(storeCat == "Few") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1           -0.0316
```

```
data %>%
    filter(storeCat == "Some") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.448
```

```
data %>%
    filter(storeCat == "Many") %>%
    summarize(cor(age, price))
```

```
## # A tibble: 1 x 1
##   `cor(age, price)`
##               <dbl>
## 1            -0.331
```

> In this graph houses with zero convenience stores in the living circle on foot have a lower average house price, and the correlation coefficient is -0.355, a weak negative linear relationship. Houses with few convenience stores have a median average house price, with a correlation coefficient of -0.0316, a weak negative linear relationship. Houses with some convenience stores have a second highest average house price, and the correlation coefficient is -0.448, a moderate negative linear relationship. And houses with many convenience stores have a highest average house price, with a correlation coefficient of -0.331, a weak negative linear relationship.

Inferential

```
lm(price ~ age+latitude+store+age:store,  data = data)  %>%
  summary()
```

```
##
## Call:
## lm(formula = price ~ age + latitude + store + age:store, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.971  -5.582  -1.209   4.182  81.363
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.026e+04  1.074e+03  -9.556  < 2e-16 ***
## age         -2.540e-01  8.037e-02  -3.161  0.00169 **
## latitude     4.125e+02  4.303e+01   9.586  < 2e-16 ***
## store        2.103e+00  3.163e-01   6.647 9.58e-11 ***
## age:store   -9.606e-03  1.415e-02  -0.679  0.49752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.703 on 409 degrees of freedom
## Multiple R-squared:  0.4964, Adjusted R-squared:  0.4915
## F-statistic: 100.8 on 4 and 409 DF,  p-value: < 2.2e-16
```

```
lm(price ~ age+latitude+store+age:store,  data = data) %>%
  confint()
```

```
##                     2.5 %        97.5 %
## (Intercept) -1.237546e+04 -8.152693e+03
## age         -4.120018e-01 -9.603451e-02
## latitude     3.278746e+02  4.970403e+02
## store        1.480734e+00  2.724333e+00
## age:store   -3.741803e-02  1.820506e-02
```

p-value for the age coefficient is 0.00169. Assuming H_0 is ture, the probability of observing a test statistic which is " as or even more extreme" than -3.161 is 0.00169. 95% confidence interval: Assuming the sampling distribution model is accurate, we are 95% confident that holding the latitude of the houses and the number of convenience stores in the living circle on foot constant, the true difference of average price(10000 New Taiwan Dollar) comparing two groups of houses that differ in 1 year is between -0.412 and -0.0960.