Exploratory Analysis Real Estate Valuation Data Set Data

Carlos Hinrichsen

June 13th, 2019

# #

Real Estate Valuation Analysis

General Information

According to the data set official web page, https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set, the relevant information about the attributes is:

The inputs are: X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.), X2=the house age (unit: year), X3=the distance to the nearest MRT station (unit: meter), X4=the number of convenience stores in the living circle on foot (integer), X5=the geographic coordinate, latitude. (unit: degree), X6=the geographic coordinate, longitude. (unit: degree)

The output is: Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Exploratory Analysis

1. Loading the required libraries

```
library(ggplot2)
library(Hmisc)
library(GGally)
library(xlsx)
library(ggmap)
```

2. Loading the data

First we need to review the raw data set to understand if there are missing values, data type, etc.

```
# Select file
file <- file.choose()
# Read data as save as object
datao <- read.xlsx(file,1)
# Copy data ot manipulation purposes
data <- datao
```

```
# Dimension and Type
str(data)
```

```
## 'data.frame':    414 obs. of  8 variables:
##  $ No                                    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ X1.transaction.date                   : num  2013 2013 2014 2014 2013 ...
##  $ X2.house.age                          : num  32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
##  $ X3.distance.to.the.nearest.MRT.station: num  84.9 306.6 562 562 390.6 ...
##  $ X4.number.of.convenience.stores       : num  10 9 5 5 5 3 7 6 1 3 ...
##  $ X5.latitude                           : num  25 25 25 25 25 ...
##  $ X6.longitude                          : num  122 122 122 122 122 ...
##  $ Y.house.price.of.unit.area            : num  37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1 ...
```

```
# Descriptive
describe(data)
```

```
## data
##
##  8  Variables      414  Observations
## ---------------------------------------------------------------------------
## No
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0      414        1    207.5    138.3    21.65    42.30
##      .25      .50      .75      .90      .95
##   104.25   207.50   310.75   372.70   393.35
##
```

```
## lowest :   1   2   3   4   5, highest: 410 411 412 413 414
## ---------------------------------------------------------------------------
## X1.transaction.date
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0       12    0.991     2013   0.3239     2013     2013
##      .25      .50      .75      .90      .95
##     2013     2013     2013     2014     2014
##
## Value      2012.667 2012.750 2012.833 2012.917 2013.000 2013.083 2013.167
## Frequency        30       27       31       38       28       46       25
## Proportion    0.072    0.065    0.075    0.092    0.068    0.111    0.060
##
## Value      2013.250 2013.333 2013.417 2013.500 2013.583
## Frequency        32       29       58       47       23
## Proportion    0.077    0.070    0.140    0.114    0.056
## ---------------------------------------------------------------------------
## X2.house.age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0      236        1    17.71    12.93    1.100    3.500
##      .25      .50      .75      .90      .95
##    9.025   16.100   28.150   34.670   37.735
##
## lowest :  0.0  1.0  1.1  1.5  1.7, highest: 40.9 41.3 41.4 42.7 43.8
## ---------------------------------------------------------------------------
## X3.distance.to.the.nearest.MRT.station
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0      259        1     1084     1205    90.46   157.61
##      .25      .50      .75      .90      .95
##   289.32   492.23  1454.28  2697.66  4082.01
##
## lowest :   23.38284   49.66105   56.47425   57.58945   82.88643
## highest: 4605.74900 5512.03800 6306.15300 6396.28300 6488.02100
## ---------------------------------------------------------------------------
## X4.number.of.convenience.stores
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0       11    0.986    4.094    3.371        0        0
##      .25      .50      .75      .90      .95
##        1        4        6        8        9
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency     67    46    24    46    31    67    37    31    30    25
## Proportion 0.162 0.111 0.058 0.111 0.075 0.162 0.089 0.075 0.072 0.060
##
## Value         10
## Frequency     10
## Proportion 0.024
## ---------------------------------------------------------------------------
## X5.latitude
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0      234        1    24.97  0.01382    24.95    24.95
##      .25      .50      .75      .90      .95
##    24.96    24.97    24.98    24.98    24.99
##
## lowest : 24.93207 24.93293 24.93363 24.93885 24.94155
## highest: 24.99156 24.99176 24.99800 25.00115 25.01459
## ---------------------------------------------------------------------------
## X6.longitude
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0      232        1    121.5  0.01601    121.5    121.5
##      .25      .50      .75      .90      .95
##    121.5    121.5    121.5    121.5    121.5
##
## lowest : 121.4735 121.4752 121.4788 121.4846 121.4951
## highest: 121.5539 121.5548 121.5596 121.5617 121.5663
## ---------------------------------------------------------------------------
## Y.house.price.of.unit.area
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      414        0      270        1    37.98    15.13    16.49    21.02
##      .25      .50      .75      .90      .95
##    27.70    38.45    46.60    54.94    59.17
##
## lowest :   7.6  11.2  11.6  12.2  12.8, highest:  71.0  73.6  78.0  78.3 117.5
## ---------------------------------------------------------------------------
```

3. Data Transformation

Considering the descriptive statistics, we can see that there are no missing values. The other thing we can notice is that the name is the attributes is too long to manipulate, the house price is not in a well known units and the first column is just an ID that has to be removed. For this, first we will change the name of the attributes and the do tu related unit transformation considering that 1 New Taiwan Dollar is around 0.042 CAD (the new unit will be CAD/mt^2. Additionally, we will change the output from numerical to categorical (10 categories, to be discussed) thinking that a good approach for the problem is to be treated as a classification problem.

Once all the transformation are done, the variables of the data set will be:

TD=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.), AGE=the house age (unit: year), DIST=the distance to the nearest MRT station (unit: meter), STR=the number of convenience stores in the living circle on foot (integer), LAT=the geographic coordinate, latitude. (unit: degree), LONG=the geographic coordinate, longitude. (unit: degree)

NPRICE= house price of unit area (CAD/mt^2), CPRICE= house price according to 10 categories (A,. . . ,J)

```r
# Currency rate NTW/CAD
cr <- 0.042
# PING/meter squared
mt <- 1/3.3
# Dropping the first column
data <- data[,-c(1)]
# Rename variables
colnames(data) <- c("TD","AGE","DIST","STR","LAT","LONG","PRICE")
# Changing the unit of house price
data$NPRICE <- data$PRICE*cr*mt*1000
# Crreating 10 categories
CPRICE <- cut(data$NPRICE,breaks = 10,labels = c("A","B","C","D","E","F","G","H","I","J"))
describe(CPRICE)
```

```
## CPRICE
##        n  missing distinct
##      414        0        8
##
## Value           A     B     C     D     E     F     G     J
## Frequency      28    95   111   115    52     9     3     1
## Proportion 0.068 0.229 0.268 0.278 0.126 0.022 0.007 0.002
```

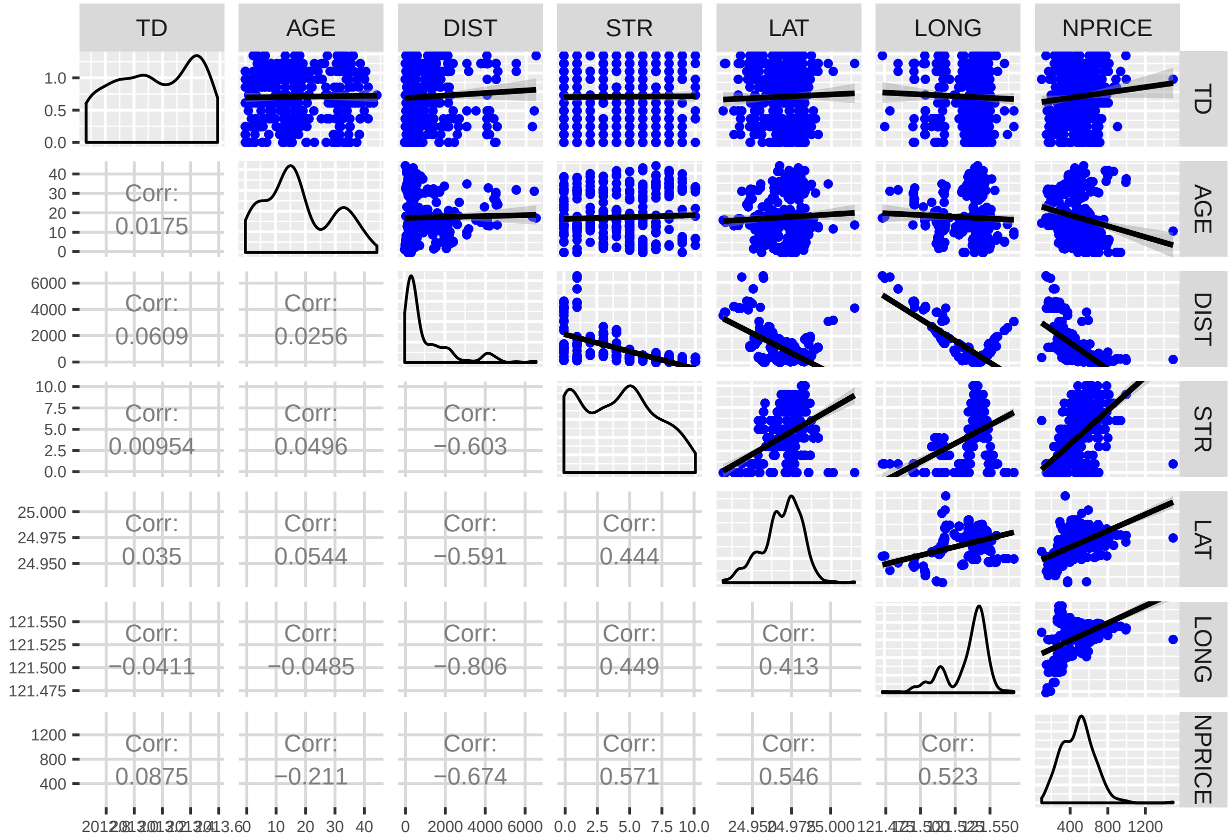
```r
data$CPRICE <- CPRICE
data <- data[,-c(7)]
head(data)
```

```
##         TD  AGE       DIST STR      LAT     LONG   NPRICE CPRICE
## 1 2012.917 32.0   84.87882  10 24.98298 121.5402 482.3636      C
## 2 2012.917 19.5  306.59470   9 24.98034 121.5395 537.0909      D
## 3 2013.583 13.3  561.98450   5 24.98746 121.5439 602.0000      D
## 4 2013.500 13.3  561.98450   5 24.98746 121.5439 697.4545      E
## 5 2012.833  5.0  390.56840   5 24.97937 121.5425 548.5455      D
## 6 2012.667  7.1 2175.03000   3 24.96305 121.5125 408.5455      C
```

4. Data Plotting

Below it is a matrix that shows all the numeric variables (categorical price is not included). The diagonal of the matrix shows the probability distribution of each variables. The upper triangular part of the matrix shows the relation of all variables (by pairs), with a linear trend. Finally, the lower triangular of the matrix shows the correlation coefficient of all the pairs of variables within the data set.

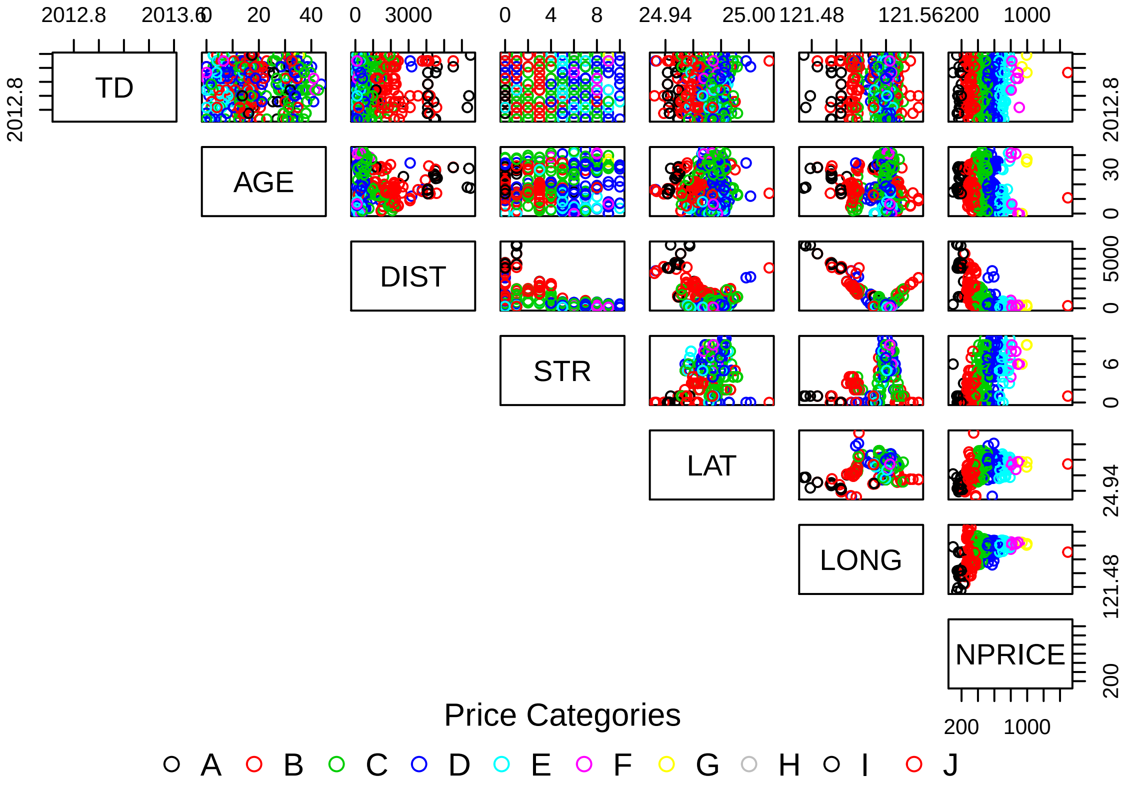
```r
ggpairs(data[,-c(8)],lower = list(continuous = wrap("cor", alpha = 1,size=3), combo = "box"),upper = list(continuous = wrap("smooth", alpha = 1, s
```



At a first glance, we can see that TD and AGE don't have clear relation with NPRICE and with the rest of the variables. They are two good candidates to eliminate of the modelling part of this project.

Anyway, the above plot doesn't show how the different house prices are distributed within the rest of the variables. In the following plot it's possible to see this effect

```
pairs(data[,-c(8)],lower.panel = NULL,col=data$CPRICE)
par(oma = c(4, 1, 1, 1))
par(fig = c(0, 1, 0, 1), oma = c(0, 0, 0, 0), mar = c(0, 0, 0, 0), new = TRUE)
plot(0, 0, type = "n", bty = "n", xaxt = "n", yaxt = "n")
legend("bottom", legend=levels(data$CPRICE), xpd = TRUE, horiz = TRUE, inset = c(0,
    0), bty = "n", col = 1:10, pch=1, cex = 1,title="Price Categories")
```

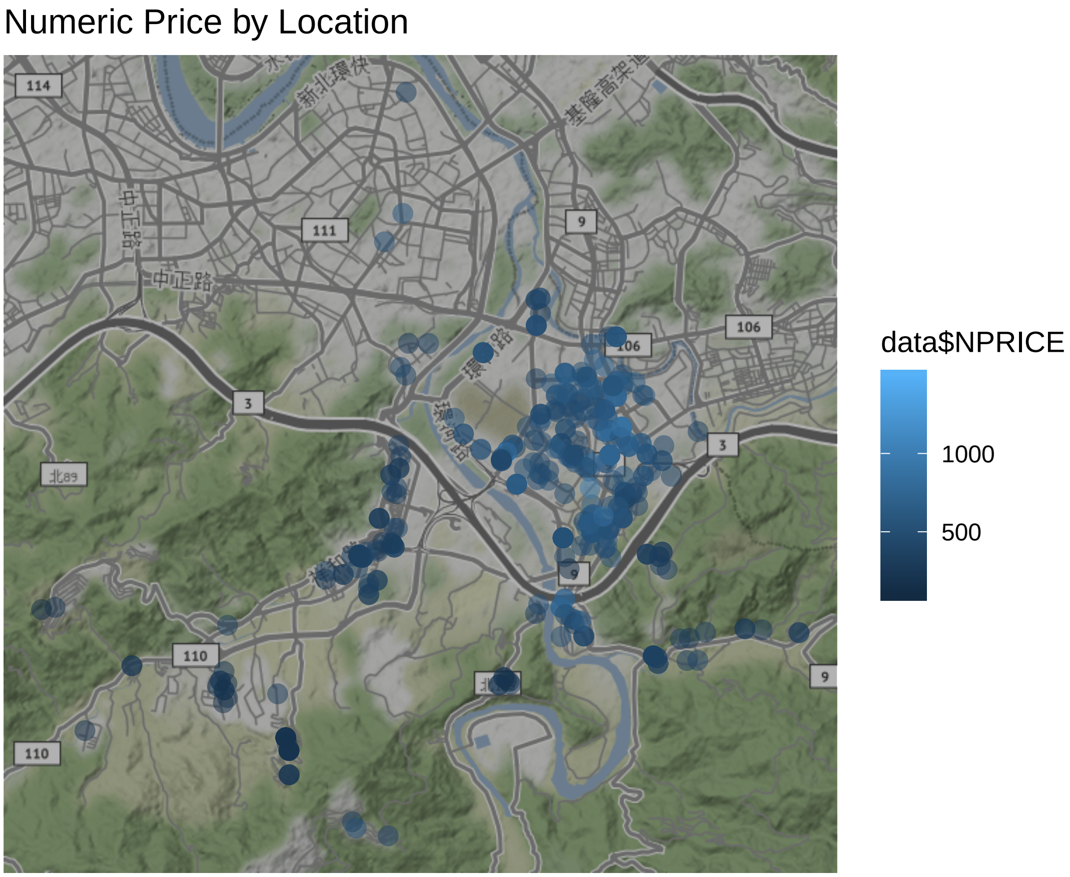


Overall is difficult to see the differences between the 10 categories. Finally, we will use a map application to show the different variables by location.

5. Geographical Data

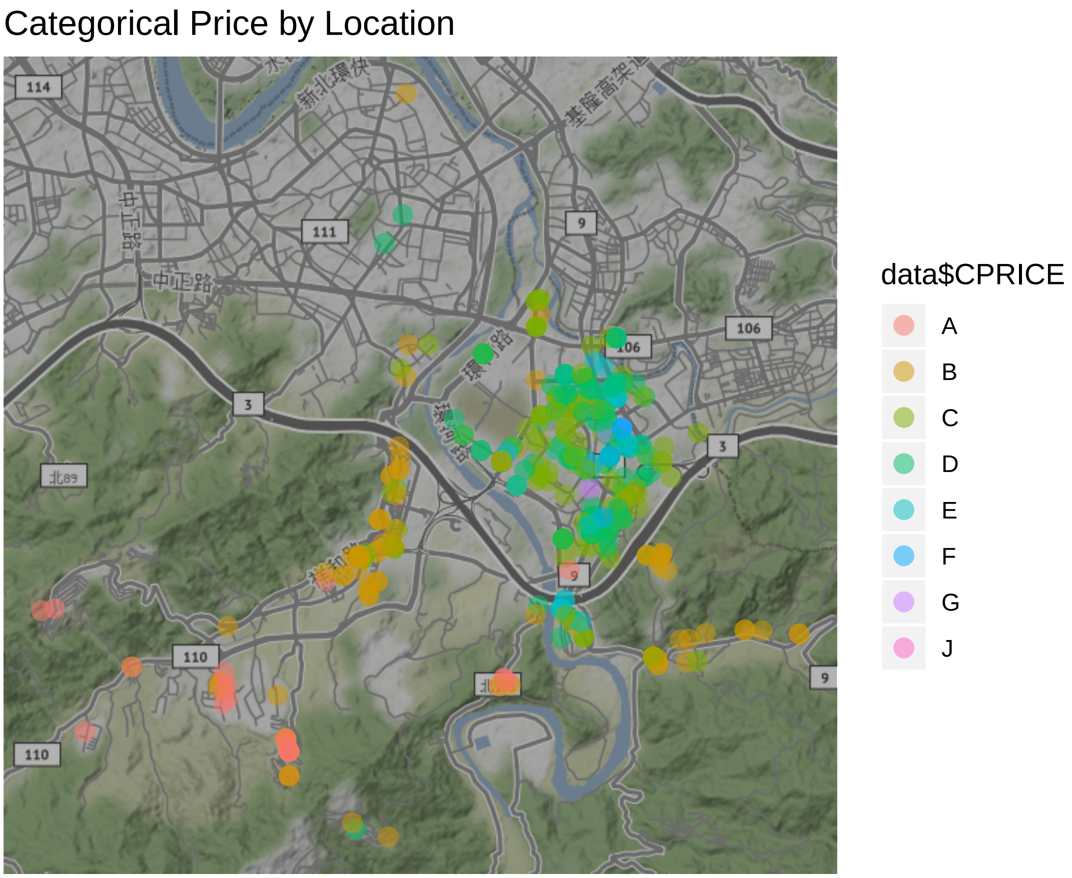
Below is the map that shows the prices of the houses (NPRICE).

```
qmplot(LONG, LAT, data = data, colour = data$NPRICE,size = I(3), darken = .3,  alpha = I(0.5),main ="Numeric Price by Location")
```



The problem with that map is that it's difficult to see cluster. Therefore, another map is created below but with the categorical variable CPRICE

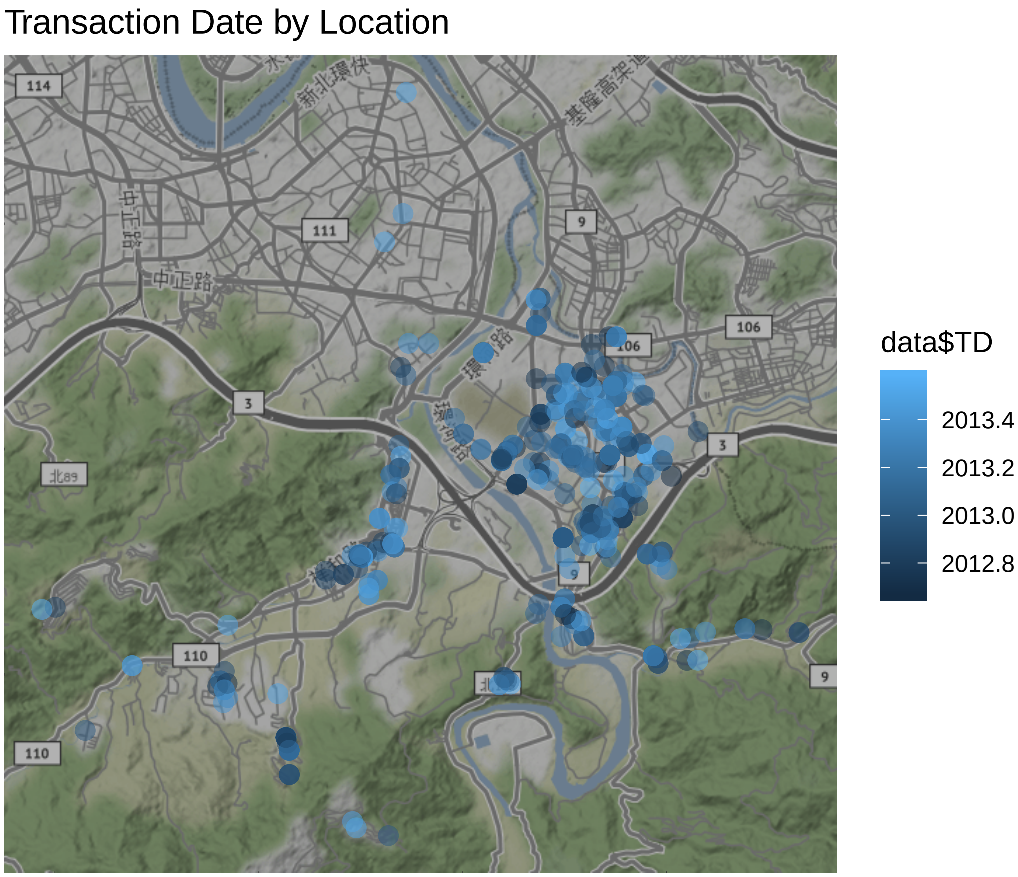
```
qmplot(LONG, LAT, data = data, colour = data$CPRICE,size = I(3), darken = .3,  alpha = I(0.5),main ="Categorical Price by Location")
```



Developing the map by categories, it's clearer that there are some cluster (or neighborhoods), with houses with similar price range.

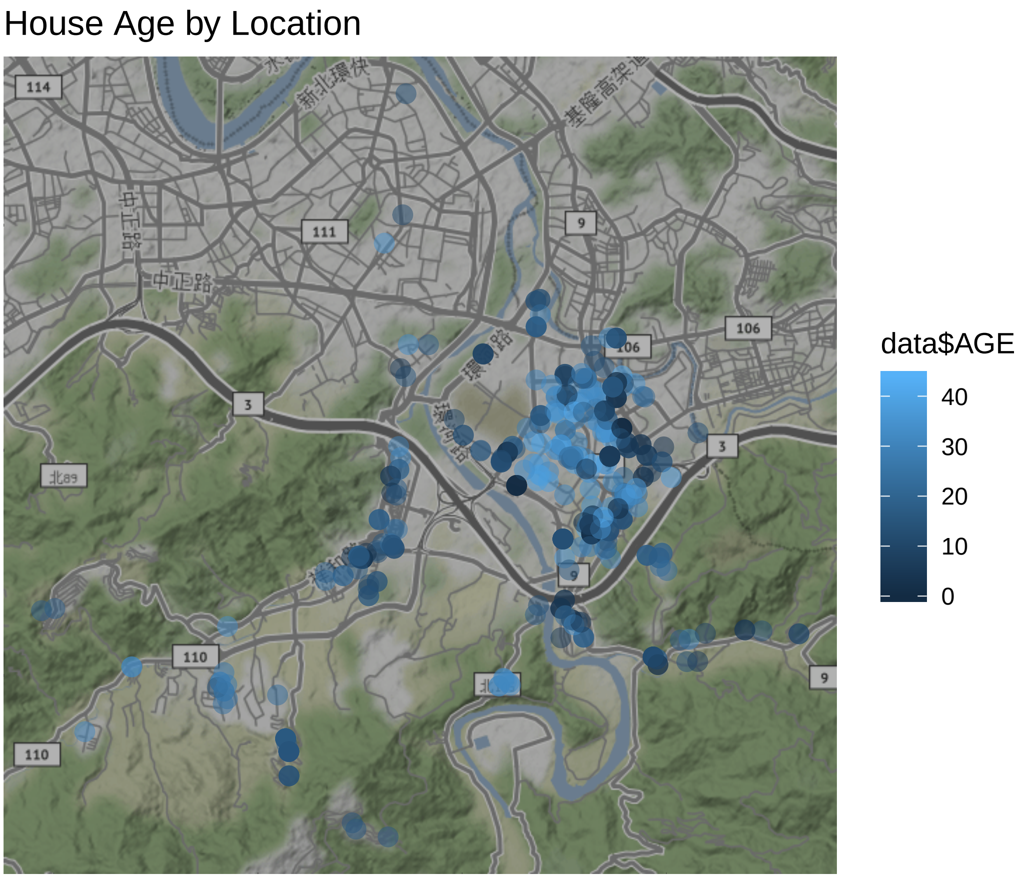
The following maps correspond to the rest of the variables by location

```
qmplot(LONG, LAT, data = data, colour = data$TD,size = I(3), darken = .3,  alpha = I(0.5),main ="Transaction Date by Location")
```



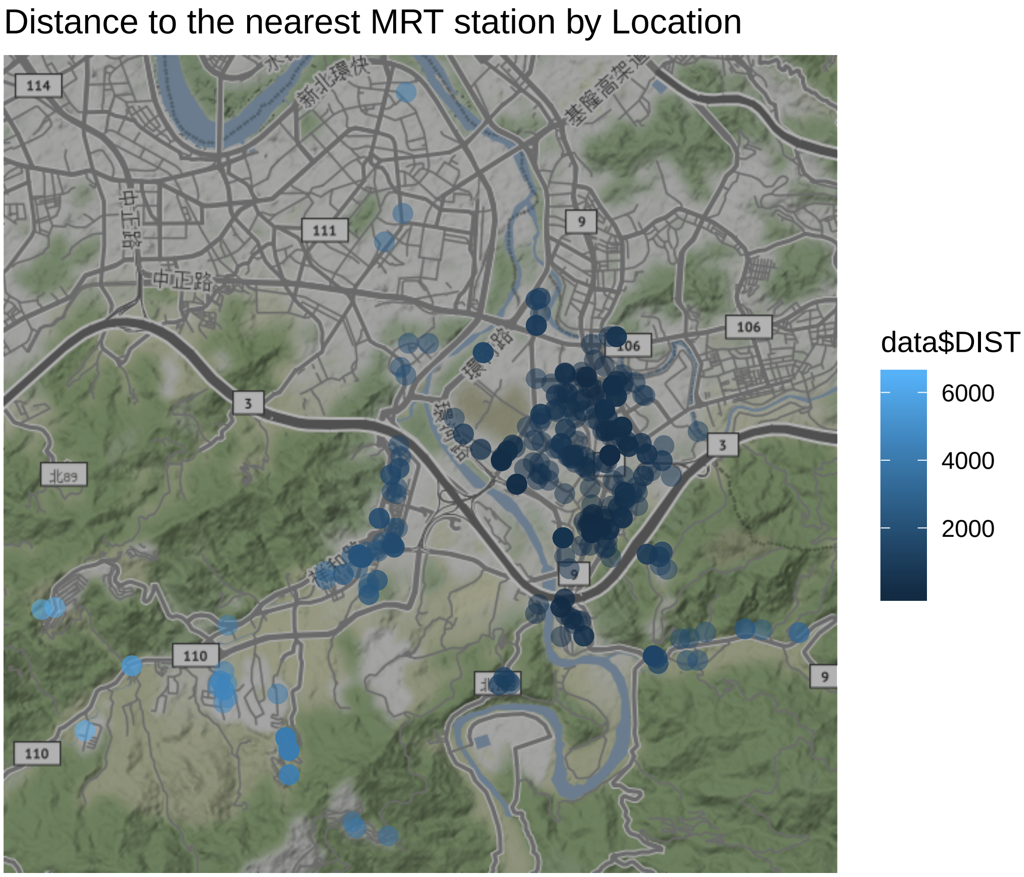
As expected, the transaction date has nothing to do with the location.

```
qmplot(LONG, LAT, data = data, colour = data$AGE,size = I(3), darken = .3,  alpha = I(0.5),main ="House Age by Location")
```



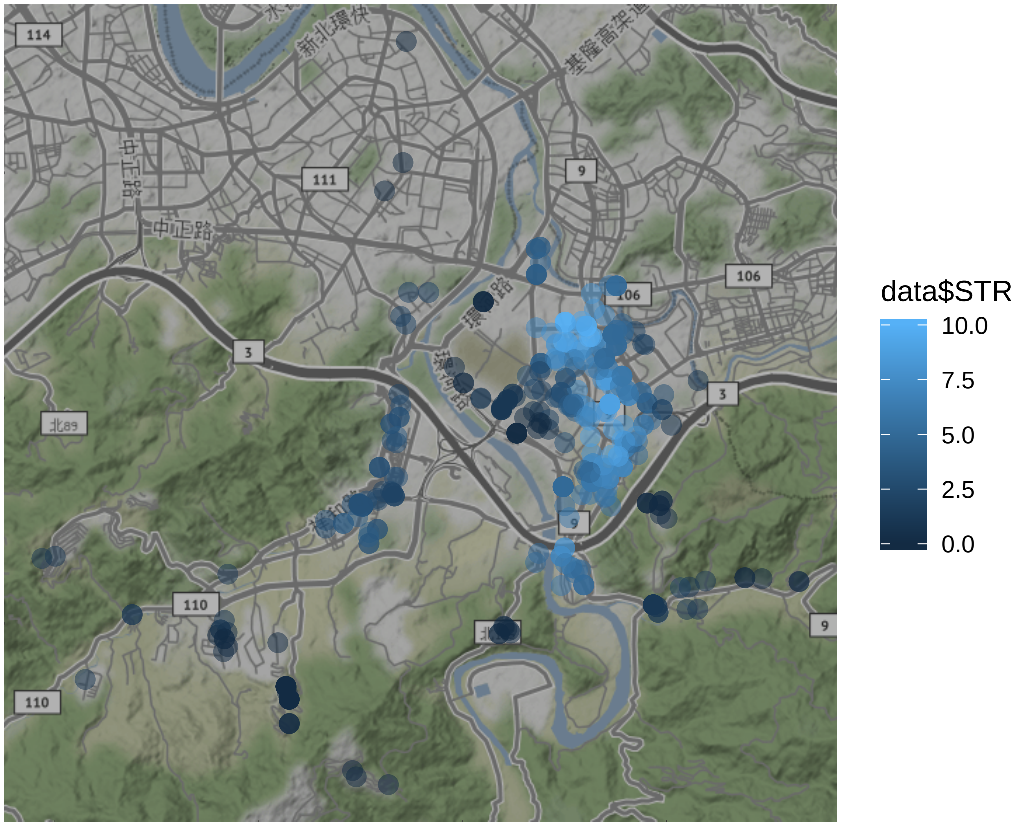
Additionally, there are some spots with houses with similar ages.

```
qmplot(LONG, LAT, data = data, colour = data$DIST,size = I(3), darken = .3,  alpha = I(0.5),main ="Distance to the nearest MRT station by Location
```



As expected, the distance to the nearest MRT station is related with the location.

```
qmplot(LONG, LAT, data = data, colour = data$STR,size = I(3), darken = .3,  alpha = I(0.5),main ="Number of convenience stores by Location")
```


Number of convenience stores by Location

```
qmplot(LONG, LAT, data = data,alpha=I(0))
```

Similar as previous one, the number of convenice stores is related with the location.

6. Prediction Analysis

We are working for you. . .