Final Project Proposal

STAT 420, Summer 2022, D. Unger

2022/07/15

Project members

- Bryan Settles (brsettl2@illinois.edu)
- Yixing Zheng (yixingz3@illinois.edu)
- Yunfei Ouyang (yunfeio2@illinois.edu)

Tentative title for the project:

- Beijing Real Estate Price Prediction using Statistical Modeling

Description of the data file:

- The data file includes the Housing price of Beijing from 2011 to 2017, fetching from Lianjia.com (similar to Zillow or Redfin). It includes URL, ID, Lng, Lat, CommunityID, TradeTime, DOM (days on market), Followers, Total price, Price, Square, Living Room, Number of Drawing room, Kitchen and Bathroom, Building Type, Construction time, Renovation Condition, Building Structure, Ladder ratio (which is the proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident have on average), Elevator, Property Rights For Five Fears (It's related to China restricted purchase of houses policy), Subway, District, Community Average Price. Most data is collected from year 2011 - 2017, some of it is from Jan, 2018, and some is from earlier(2010, 2009). All the data was fetching from https://bj.lianjia.com/chengjiao.

Background information and source File

- Background information:
  - After some quick cleaning to remove the invalid values, we are left with 159376 obs. of 26 variables, from which we will select around 10 variables to build our model and test the performance with two splitted data sets - train and test, each might contain 80000 observations (depends on the calculation resources needed, we might reduce the number of observations used in building and testing the model).
  - One variable we would like to include but requires further data cleaning is `floor` that is composed by a Chinese character indicating the catalog of the height of the building and its total number of floors. For example, `\xb8\xdf 26` should be translated to `tall 26`, which means the building has 26 floors and is categorized as a tall building. Similarly, `\xd6\xd0 4` means a low building with 4 floors in total, and `\xb8\xdf 10` means a medium height building of in total 10 floors.
- Source file link: Housing price in Beijing

Statement of interest

- Real estate price prediction is attractive for both holders and traders. It is an interesting topic since many factors can inflate the house price in Beijing. For example, we want to investigate how housing prices in Beijing are related to the growth of its economy. We will construct several statistical models to predict the data on Beijing's house prices. Specifically, we will utilize multiple linear regression, categorical predictors, transformations and model building using AIC, and BIC. We then will use model selection tools and model diagnostic methods to decide which model is the best model for predicting house prices. Finally, we will do a deep analysis of the best model to see its performance.

Loading source data into R

```
library(readr)
housing = read_csv("Housing_price_in_Beijing.csv")
housing = as.data.frame(housing)
housing = na.omit(housing)

str(housing)
```

```
## 'data.frame':    159376 obs. of  26 variables:
##  $ url                : chr  "https://bj.lianjia.com/chengjiao/101084782030.html" "https://bj.lianji
##  $ id                 : chr  "101084782030" "101086012217" "101086041636" "101086406841" ...
##  $ Lng                : num  116 116 117 116 116 ...
##  $ Lat                : num  40 39.9 39.9 40.1 39.9 ...
##  $ Cid                : num  1.11e+12 1.11e+12 1.11e+12 1.11e+12 1.11e+12 ...
##  $ tradeTime          : Date, format: "2016-08-09" "2016-07-28" ...
##  $ DOM                : num  1464 903 1271 965 927 ...
##  $ followers          : num  106 126 48 138 286 57 167 138 218 134 ...
##  $ totalPrice         : num  415 575 1030 298 392 ...
##  $ price              : num  31680 43436 52021 22202 48396 ...
##  $ square             : num  131 132 198 134 81 ...
##  $ livingRoom         : num  2 2 3 3 2 1 2 3 1 1 ...
##  $ drawingRoom        : num  1 2 2 1 1 0 1 2 0 0 ...
##  $ kitchen            : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ bathRoom           : num  1 2 3 1 1 1 1 2 1 0 ...
##  $ floor              : chr  "\xb8\xdf 26" "\xb8\xdf 22" "\xd6\xd0 4" "\xb5\xd7 21" ...
##  $ buildingType       : num  1 1 4 1 4 4 4 1 3 1 ...
##  $ constructionTime   : chr  "2005" "2004" "2005" "2008" ...
##  $ renovationCondition: num  3 4 3 1 2 3 4 4 1 4 ...
##  $ buildingStructure  : num  6 6 6 6 2 6 2 6 2 6 ...
##  $ ladderRatio        : num  0.217 0.667 0.5 0.273 0.333 0.333 0.5 0.667 0.333 0.308 ...
##  $ elevator           : num  1 1 1 1 0 1 0 1 0 1 ...
##  $ fiveYearsProperty  : num  0 1 0 0 1 1 0 1 0 1 ...
##  $ subway             : num  1 0 0 0 1 0 0 0 0 1 ...
##  $ district           : num  7 7 7 6 1 7 7 7 13 1 ...
##  $ communityAverage   : num  56021 71539 48160 51238 62588 ...
##  - attr(*, "na.action")= 'omit' Named int [1:159475] 11 13 80 94 133 146 190 209 223 232 ...
##    ..- attr(*, "names")= chr [1:159475] "11" "13" "80" "94" ...
```

```
head(housing, 10)
```

```
##                                                 url           id      Lng
## 1  https://bj.lianjia.com/chengjiao/101084782030.html 101084782030 116.4755
## 2  https://bj.lianjia.com/chengjiao/101086012217.html 101086012217 116.4539
## 3  https://bj.lianjia.com/chengjiao/101086041636.html 101086041636 116.5620
## 4  https://bj.lianjia.com/chengjiao/101086406841.html 101086406841 116.4380
## 5  https://bj.lianjia.com/chengjiao/101086920653.html 101086920653 116.4284
## 6  https://bj.lianjia.com/chengjiao/101087277815.html 101087277815 116.4663
## 7  https://bj.lianjia.com/chengjiao/101087292623.html 101087292623 116.4826
## 8  https://bj.lianjia.com/chengjiao/101087303800.html 101087303800 116.4539
## 9  https://bj.lianjia.com/chengjiao/101087463212.html 101087463212 116.5557
## 10 https://bj.lianjia.com/chengjiao/101087508625.html 101087508625 116.4531
##        Lat          Cid  tradeTime  DOM followers totalPrice price square
## 1  40.01952 1.111027e+12 2016-08-09 1464       106      415.0 31680 131.00
## 2  39.88153 1.111027e+12 2016-07-28  903       126      575.0 43436 132.38
## 3  39.87714 1.111041e+12 2016-12-11 1271        48     1030.0 52021 198.00
## 4  40.07611 1.111043e+12 2016-09-30  965       138      297.5 22202 134.00
## 5  39.88623 1.111027e+12 2016-08-28  927       286      392.0 48396  81.00
## 6  39.99136 1.111027e+12 2016-07-22  861        57      275.6 52000  53.00
## 7  39.89199 1.111027e+12 2016-07-14  851       167      275.0 37672  73.00
## 8  39.88153 1.111027e+12 2016-09-07  904       138      800.0 49521 161.55
## 9  40.16206 1.111027e+12 2016-09-04  873       218      134.0 27917  48.00
## 10 39.89840 1.111027e+12 2016-09-05  865       134      380.0 55883  68.00
##    livingRoom drawingRoom kitchen bathRoom       floor buildingType
## 1           2           1       1        1 \xb8\xdf 26            1
## 2           2           2       1        2 \xb8\xdf 22            1
## 3           3           2       1        3  \xd6\xd0 4            4
## 4           3           1       1        1 \xb5\xd7 21            1
## 5           2           1       1        1  \xd6\xd0 6            4
## 6           1           0       1        1  \xd6\xd0 8            4
## 7           2           1       1        1  \xb8\xdf 6            4
## 8           3           2       1        2 \xb8\xdf 22            1
## 9           1           0       1        1 \xb8\xdf 10            3
## 10          1           0       0        0 \xd6\xd0 23            1
##    constructionTime renovationCondition buildingStructure ladderRatio elevator
## 1              2005                   3                 6       0.217        1
## 2              2004                   4                 6       0.667        1
## 3              2005                   3                 6       0.500        1
## 4              2008                   1                 6       0.273        1
## 5              1960                   2                 2       0.333        0
## 6              2005                   3                 6       0.333        1
## 7              1997                   4                 2       0.500        0
## 8              2004                   4                 6       0.667        1
## 9              2009                   1                 2       0.333        0
## 10             2009                   4                 6       0.308        1
##    fiveYearsProperty subway district communityAverage
## 1                  0      1        7            56021
## 2                  1      0        7            71539
## 3                  0      0        7            48160
## 4                  0      0        6            51238
## 5                  1      1        1            62588
## 6                  1      0        7            67738
## 7                  0      0        7            50112
## 8                  1      0        7            71539
## 9                  0      0       13            44235
```

```
## 10                  1       1         1             78590
```