final-proj-proposal

April 24, 2022

1 Princeton House Prices and Ethnicity

By Brandon Feder and Atticus Wang

2 Introduction

Princeton has many diverse houses ranging in size, location, age, and many other factors. With this comes a huge range of house prices. But how do the characteristics of a house determine it's price? This our research question: **"How do factors such as the number of bathrooms, location, year built, and number of bathrooms effect the price of the house?"**

Beatrice Bloom, a Princeton Residential Specialist, provides many great resources about the Princeton housing market including a table of houses sold in Princeton since 2011. This data can be found here. We indented to use this data to answer our question.

We also hope to use US census data about the income, ethnicity, and age of people in different areas of Princeton to help in our prediction.

Based off our own experience, we predict that neighborhood/address and year built have a significant impact on the price of the house. In addition, we believe that the ethnicity and income of the residents of the neighborhood the houses reside will be good predictors of the house's price.

3 Analysis Plan

3.0.1 Variables

The response variable in our analysis will be the price of a house in dollars.

The predictors will include location, number of bedrooms, number of full bathrooms, number of half bathrooms, style, year built, parking-lot size, and previous selling price.

Some other relevant variables include the number of days on the market and data about the human population in the area of town the houses resides (such as age, race/ethnicity, income, etc).

3.0.2 Analysis Plan

First, we will tidy up the Princeton real estate market dataset, and extract US census data pertaining to age, race/ethnicity, income, and other variables in the Princeton area on the level of census blocks. We will then analyze the correlation between house prices and resident race and ethnicity, and potentially other correlations.

Next, we will build a linear model using a part of the data set with house prices as the response variable. Because there are many potential predictors, we will use the step function to select the best model. Then, for the rest of the dataset, we will predict house prices using the chosen predictors and compare our predictions with actual house prices. Finally, we will study why the model differs from real data, and whether there are temporal trends to house prices.

3.0.3 Preliminary Analysis

Load required libraries

[ ]:
```
install.packages("tidyverse")
library(tidyverse)
```

```
Installing package into '/mnt/MainStorage/bfeder/R/x86_64-pc-linux-gnu-
library/4.0'
(as 'lib' is unspecified)

  Attaching packages                                   tidyverse
1.3.1

 ggplot2 3.3.5      purrr   0.3.4
 tibble  3.1.6      dplyr   1.0.7
 tidyr   1.1.4      stringr 1.4.0
 readr   2.1.1      forcats 0.5.1

  Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()
```

Load raw housing data

[ ]:
```
house <- read.csv("./data/pton-market-data.csv")
```

Format price column

[ ]:
```
house <- house %>%
mutate(Price = strtoi(str_replace_all(str_sub(Sold.Price, 2, -4), ",", "")))
```

Calculate top 10 most expensive neighborhoods (on average)

[ ]:
```
house %>%
group_by(Neighborhood) %>%
summarise(meanPrice = mean(Price, na.rm = TRUE)) %>%
arrange(desc(meanPrice)) %>%
head(10)
```

A tibble: 10 × 2

Neighborhood <chr>	meanPrice <dbl>
The Preserve	2150000.0
Carnegie Lake	1645562.5
Battelfield Area	1350950.0
Institute	1338176.6
Pretty Brook Area	1312639.5
princeton Ridge	970000.0
Western Section	897657.4
The Glen	883113.5
Hun Area	862042.5
Institute Area	831696.7

Calculate top 10 most expensive neighborhoods (on average)

[ ]:
```
house %>%
group_by(Address) %>%
summarise(meanPrice = mean(Price, na.rm = TRUE)) %>%
arrange(desc(meanPrice)) %>%
head(10)
```

A tibble: 10 × 2

Address <chr>	meanPrice <dbl>
Garrett Ln	2695000
Pheasant Hil Rd	2610000
Libary Pl	2476938
Fredrick Ct	2290000
Bogart Ct	2213125
Cradle Rock Rd	2138000
Morven Pl	2042000
Grasmere Way	2031250
Running Cedar Rd	1967458
Province Line	1950000

4 Data

Here is a summary of the house price data. We were not yet able to obtain the US census data do to an issue with our API key.

[ ]:
```
dplyr::glimpse(house)
```

```
Rows: 3,331
Columns: 16
$ X                          <chr> "49-F", "44-H", "218",
"12", "93-95", "58", …
$ Address                    <chr> "Palmer Sq", "Nassau St",
"Birch Ave", "Birc…
$ Neighborhood               <chr> "Princeton Center",
"Princeton Center", "Pri…
```

```
$ Bed.Rooms                 <int> 0, 0, 3, 3, 3, 3, 3, 3, 2,
3, 3, 3, 3, 2, 3,…
$ Full.Baths                <int> 1, 1, 1, 1, 2, 2, 1, 2, 1,
2, 2, 2, 2, 2, 2,…
$ Half.Baths                <int> 0, 0, 0, 0, 0, 0, 1, 0, 1,
0, 0, 0, 0, 1, 1,…
$ Style                     <chr> "Flat", "Flat", "Twin",
"Twin", "Bungalow", …
$ Year.Built                <chr> "1932", "1932", "1929", NA,
"1940", "1900", …
$ Lot.Size                  <chr> NA, NA, NA, "0.04", "0.07",
"0.08", "0.11", …
$ Original.Price            <chr> "$320,000.00",
"$369,000.00", "$395,000.00",…
$ Last.Price                <chr> "$320,000.00",
"$329,000.00", "$395,000.00",…
$ Sold.Price                <chr> "$320,000.00",
"$320,000.00", "$395,000.00",…
$ Sold.Date                 <chr> "3/14/22", "1/4/22",
"2/28/22", "3/3/22", "1…
$ Days.on.Market            <chr> "7", "13", "6", "85", "17",
"11", "146", "11…
$ Property.Marketing.Period <chr> "7", "175", "6", "85",
"17", "11", "146", "1…
$ Price                     <int> 320000, 320000, 395000,
475000, 590000, 6400…
```

5 Reference

Beatrice Bloom website: https://www.realestate-princeton.com/market-analysis/princeton-pending/

CRAN `tidycensus` package: https://walker-data.com/tidycensus/

US census data website: https://www.census.gov/data.html