Check-In Presentation Materials

Ethan Rodriguez-Shah

10/27/2017

Introduction

With the population changes and shifting demographics in the United States capital brought on in the aftermath of the 2008 economic crisis, real estate is an incredibly important business, coveted by real estate agencies and developers and peddled by the District government. Many different factors can affect property values, including number of bedrooms, number of bathrooms, style of kitchen, square footage, and style of flooring, among others. A University of Michigan study conducted by the Department of Horticulture found that even plant material, size, and even design can affect the value of a home (Barton et al). Through my research, I will be examining the effect that proximity to Metro stations has on property prices in the Washington, D.C., Metropolitan Area. In examining property data, I expect to find that Metro station proximity and property prices will have negative correlation; that is, property prices will increase the nearer a property is to a station.

Data Description

My research will be conducted using two datasets: the first, station_locations, is a list of the ninety-one Metro stations in the WMATA system, provided by Federal Reserve economist Neil Bhutta; the second, redfin_data, consists of all the properties in the Washington Metropolitan Area listed by Redfin, a Seattle-based residential real estate company, as sold in the three months prior to October 19, 2017. I compiled the panel data frame redfin_data manually from the Redfin website. Redfin only allows downloads of data sets with the information of 350 or less properties at one time, so I compiled the redfin_data by reading in many datasets and joining them by the ADDRESS variable, so that no single property would be included in the data twice. The station_locations data frame will be useful for answering my research question because it contains location data for all WMATA stations, and the redfin_data data frame will be useful because it includes not only location data but also data on other property characteristics like number of bedrooms, number of bathrooms, and square footage. This information should be sufficient to allow me to test my hypothesis.

Key economic variables

This research centers around only one completely dependent variable, two completely independent variable, and many other variables that may shift between being dependent and independent. The dependent variable is property price. The independent variables are date sold and distance from nearest Metro station. The flexible variables are as follows:

```
* Property type
* Number of bedrooms
* Number of bathrooms
* Square footage
* Days on market
```

Data loading and cleaning

Because of the nature in which I had to collect the data, my loading and cleaning process was very long.

Loading necessary libraries

```
##-----
## Necessary libraries
#install.packages("RCurl")
library(RCurl)
library(dplyr)
library(geosphere)
library(stringr)
library(lubridate)
```

Reading in data

Reading in station_locations wasn't complicated, so I won't include that code. Compiling redfin_data, on the other hand, was a little more of a process, so I'll show how I did it. First I read in each data set in ~/prelim_data/ individually. Then I joined them into one data frame using the base function rbind():

```
##-----
## Combine real estate data into single dataset

redfin_data <- studio %>%
  rbind(bed1_not_condo) %>%
  rbind(bed1_condo_0_750) %>%
  rbind(bed1_condo_750_inf) %>%
  rbind(bed2_not_condo) %>%
  rbind(bed2_condo_0_1250) %>%
  rbind(bed2_condo_1250_inf) %>%
  rbind(bed3_all_else) %>%
  rbind(bed3_house_condo) %>%
  rbind(bed3_townhouse_0_1500) %>%
  rbind(bed3_townhouse_1500_inf) %>%
  rbind(bed4_not_townhouse) %>%
  rbind(bed4_townhouse) %>%
  rbind(bed5) %>%
  rbind(bed6plus)
```

Cleaning the data

To clean the data, I removed unnecessary variables like STATUS (all properties are sold) and NEXT.OPEN.HOUSE.START.TIME and NEXT.OPEN.HOUSE.END.TIME (both of which are 2680 iterations of NA), among others, from redfin_data. I also ensured that all variables in redfin_data were the correct data class. It was necessary to alter the class of redfin_data$DATE:

```
# station_locations has all appropriate data classes. redfin_data needs work
redfin_data <- redfin_data %>%
  mutate(SOLD.DATE = mdy(SOLD.DATE))
```

For station_locations, I simply ensured that there were no missing values, as complete station location data is essential to answering my research question:

```
# Any missing values in station_locations?
na_rows_stat <- filter(station_locations,
                       is.na(station_locations$station) |
                    is.na(station_locations$address1) |
                    is.na(station_locations$address2) |
                    is.na(station_locations$Latitude) |
                    is.na(station_locations$Longitude))
na_rows_stat
```

```
## [1] station   address1  address2  Latitude  Longitude
## <0 rows> (or 0-length row.names)
```

```
# A total of 0 rows have missing values.
```

Calculating the distance column

The final step in preparing my data was calculating the distance column for redfin_data.I defined a function Metro_dist() that takes one argument, locations, which is either a vector or a matrix of longitude and latitude coordinates (in that order). The function will return distance from the nearest DC Metro station in miles:

```
##-----
## Defining a function for distance to closest Metro station.
Metro_dist <- function(locations) {
  dist_var <- Inf
  test_dist <- 0
  for(n in 1:nrow(station_locations)) {

    test_dist <- locations %>%
      distHaversine(as.matrix(station_locations[n, c("Longitude", "Latitude")]))

    dist_var <- ifelse(test_dist < dist_var, test_dist, dist_var)
  }
  dist_var <- dist_var / 1609.344    # 1 mi = 1609.344 m
  return(dist_var)
}
```

I calculated redfin_data$dist using this function:

```
redfin_data <- redfin_data %>%
  mutate(dist = Metro_dist(redfin_data[,c("LONGITUDE", "LATITUDE")]))
```

At the end of my process, I added two new variables: dist_bin, which sorts distances from nearest Metro stations into bins based on redfin_data$dist rounded to the nearest tenth of a mile, and new_price, which shows property prices in millions:

```
##-----
## Cutting distance into bins and transforming price into multiples of one million
redfin_data <- redfin_data %>%
  mutate(dist_bin = round(dist, digits = 1),
         new_price = PRICE / 1000000)
```

Now that redfin_data is compiled, let's take a look at it:

```
head(redfin_data, n = 3)
```

```
##    SOLD.DATE PROPERTY.TYPE                       ADDRESS       CITY STATE
## 1 2017-10-13   Condo/Co-op          1725 17th St NW #503 Washington    DC
## 2 2017-08-31   Condo/Co-op         300 M St SW Unit N303 Washington    DC
## 3 2017-09-15   Condo/Co-op 1441 Rhode Island Ave NW #915 Washington    DC
##      ZIP   PRICE BEDS BATHS    LOCATION SQUARE.FEET LOT.SIZE YEAR.BUILT
## 1 20009 219500    0     1 Old City #2         440       NA       1917
## 2 20024 224900    0     1    Rla (sw)         463       NA       1967
## 3 20005 366500    0     1 Old City #2         552       NA       2003
##   DAYS.ON.MARKET X..SQUARE.FEET HOA.MONTH LATITUDE LONGITUDE      dist
## 1              5            499       536 38.91354 -77.03793 0.4131710
## 2             48            486       421 38.87613 -77.01649 0.0485816
## 3             33            664       299 38.90860 -77.03376 0.4902665
##   dist_bin new_price
## 1      0.4    0.2195
## 2      0.0    0.2249
## 3      0.5    0.3665
```

Summary Statistics

In working on initial summary statistics, I produced four charts providing some insight on relationships between variables in my data set.

Loading necessary libraries

In my original work, I sourced my loading/cleaning/validation script in order to have redfin_data on hand. Since it's all compiled here in the same document, the sourcing won't be necessary.

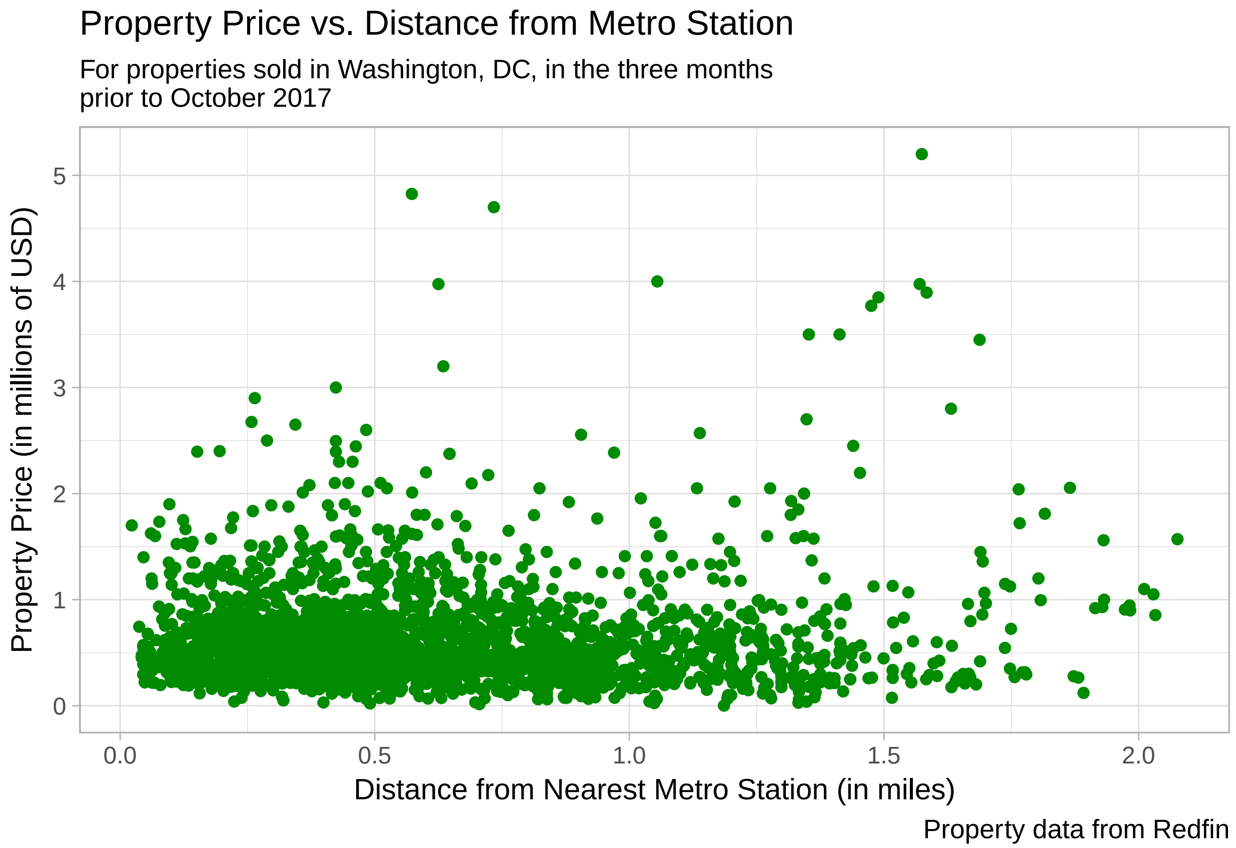
```r
##-----
## Necessary libraries
library(ggplot2)
library(dplyr)
#source("ETHAN_RSHAH_loading_cleaning_validation.R")
```

Property price vs. distance from nearest Metro station

This plot shows the relationship between property prices and the distance from the nearest Metro station for properties sold in the Washington Metropolitan Area in the three months prior to October 2017.

```r
##-----
## Plotting property price vs. distance from nearest Metro station
price_dist <- redfin_data %>%
  ggplot(aes(x = dist,
             y = new_price)) +
  geom_point(color = "green4") +
  labs(title = "Property Price vs. Distance from Metro Station",
       subtitle = "For properties sold in Washington, DC, in the three months\nprior to October 2017",
       x = "Distance from Nearest Metro Station (in miles)",
       y = "Property Price (in millions of USD)",
       caption = "Property data from Redfin") +
  theme_light()

price_dist
```



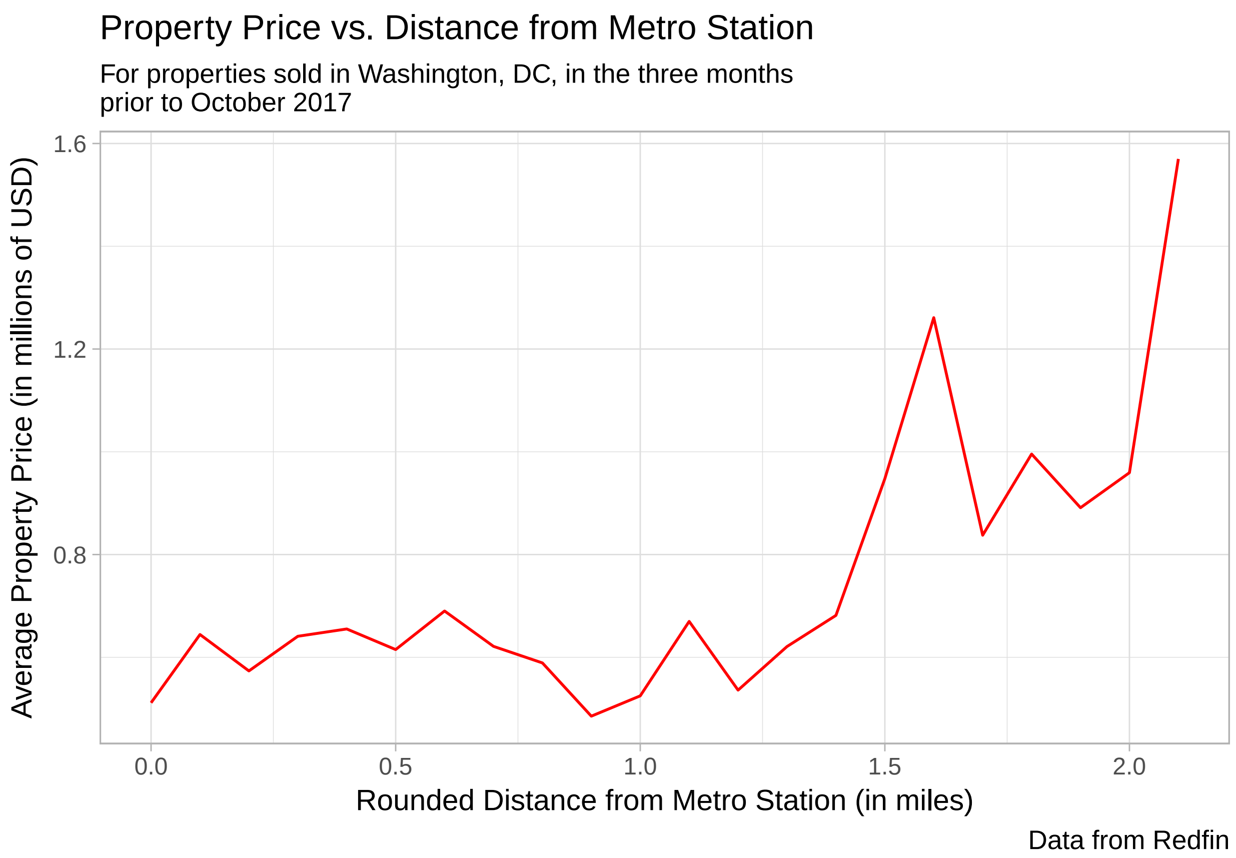
This isn’t very clear, so I approached it a different way.

Property price vs. distance bins

This plot cuts data into bins rounded to the nearest tenth of a mile. The price values are the average prices by bin.

```
##-----
## Plotting property price vs. distance bins
price_distbin <- redfin_data %>%
  group_by(dist_bin) %>%
  summarise(avg_price = mean(new_price)) %>%
  ggplot(aes(x = dist_bin,
             y = avg_price)) +
  geom_line(color = "red") +
  labs(title = "Property Price vs. Distance from Metro Station",
       subtitle = "For properties sold in Washington, DC, in the three months\nprior to October 2017",
       x = "Rounded Distance from Metro Station (in miles)",
       y = "Average Property Price (in millions of USD)",
       caption = "Data from Redfin") +
  theme_light()

price_distbin
```



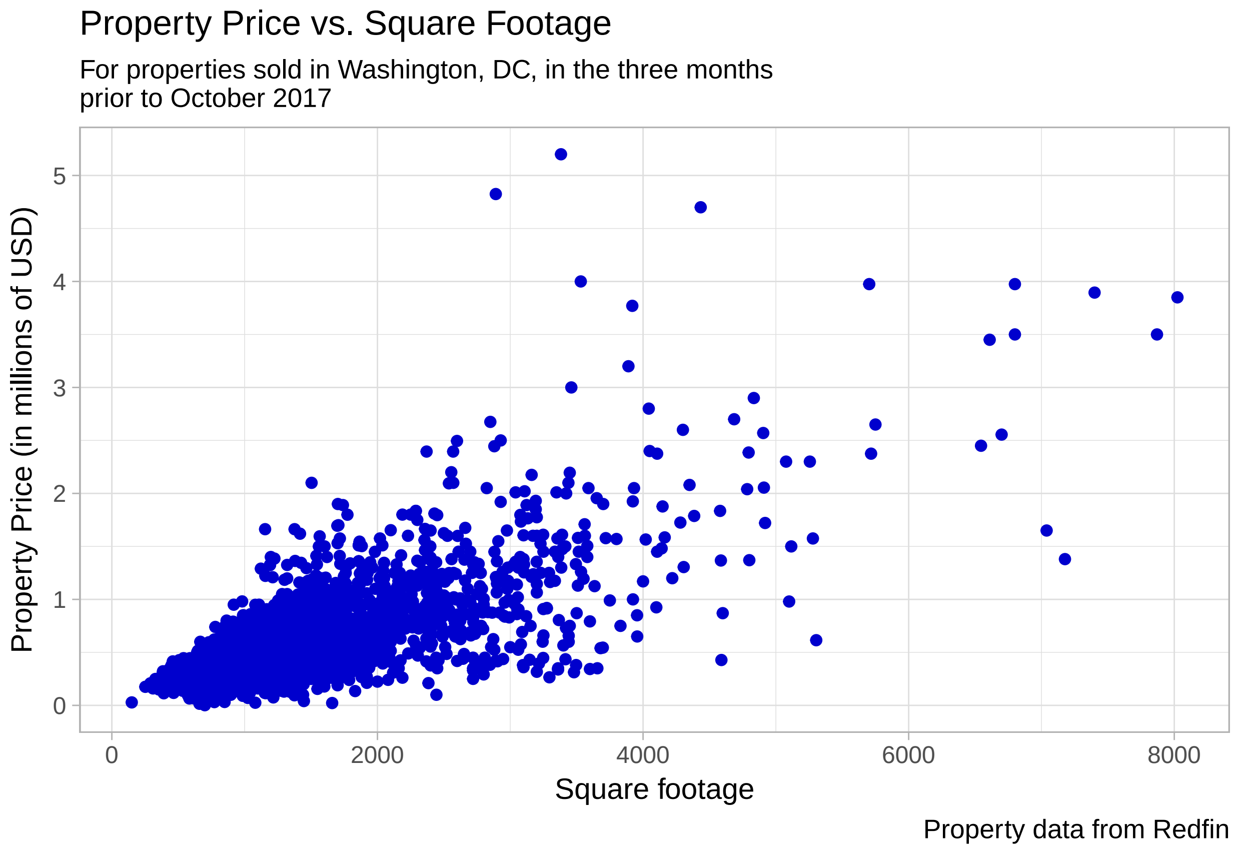
This chart is a lot clearer. It seems as though property prices generally increase the further the property is from a Metro station; that is, property prices and proximity to the nearest Metro station have a positive correlation, seemingly disproving my hypothesis. However, other factors may be at play here, so this plot is not definitive.

Property price vs. square footage

This plot shows the relation between property prices and square footage.

```
##-----
## Plotting property price vs. square footage
price_sqft <- redfin_data %>%
  ggplot(aes(x = SQUARE.FEET,
             y = new_price)) +
  geom_point(color = "blue3") +
  labs(title = "Property Price vs. Square Footage",
       subtitle = "For properties sold in Washington, DC, in the three months\nprior to October 2017",
       x = "Square footage",
       y = "Property Price (in millions of USD)",
       caption = "Property data from Redfin") +
  theme_light()

price_sqft
```



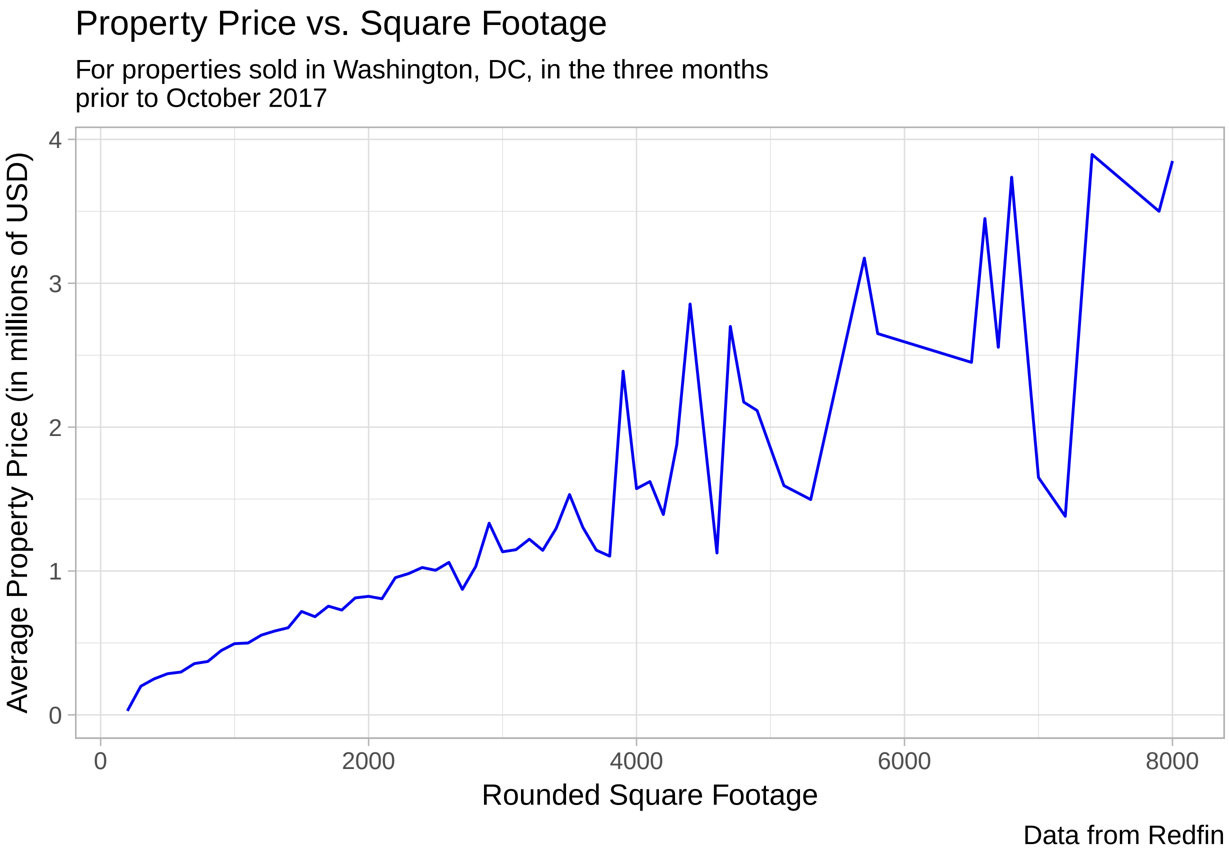
This isn't too clear either, so I approached it as I did with the property price vs. distance chart.

Property price vs. square footage bins

This plot cuts square footage into bins rounded to the nearest 100 square feet. The price values are the average prices by bin.

```r
##-----
## Plotting property price vs. square footage bins
price_sqftbin <- redfin_data %>%
  mutate(sqft_bin = round(SQUARE.FEET, digits = -2)) %>%
  group_by(sqft_bin) %>%
  summarise(avg_price = mean(new_price)) %>%
  ggplot(aes(x = sqft_bin,
             y = avg_price)) +
  geom_line(color = "blue2") +
  labs(title = "Property Price vs. Square Footage",
       subtitle = "For properties sold in Washington, DC, in the three months\nprior to October 2017",
       x = "Rounded Square Footage",
       y = "Average Property Price (in millions of USD)",
       caption = "Data from Redfin") +
  theme_light()

price_sqftbin
```



Property Price vs. Square Footage

For properties sold in Washington, DC, in the three months prior to October 2017

Data from Redfin

This chart shows that property prices and square footage have a positive relationship, which is believable. Again, however, regressions are necessary to determine true correlations between property price and several independent variables.

Current Conclusions and Next Steps

So far, I have been able to compile the necessary real estate data from Redfin, calculate the distances from each of more than 2,500 properties to the nearest Metro stations, and begin to produce some meaningful charts. The next steps in answering my research question is to determine the relationships between the rest of the independent, dependent, and flexible variables, and then run regressions to determine the true relationship between proximity to Metro stations and property prices.