A Statistical Analysis of Saratoga's Property Features and Prices Infographic

Nicole Rodriguez

4/09/2025

Contents

Introduction

Analyzing the Characteristics that Increase House Prices in Saratoga, NY

In this analysis, I examined a dataset containing information on 1,057 homes in Saratoga, New York, with the goal of identifying the key factors that influence house prices in the area. My primary focus was to explore how variables like living area, number of bedrooms and bathrooms, lot size, age, and the presence of a fireplace related to the market price of these homes. I used descriptive statistics, including mean and median, to summarize the data and uncover trends in house pricing.

For visualization, I created several plots to better understand the data, including a violin plot to show the distribution of house prices with and without a fireplace. This revealed that homes with fireplaces generally sold for higher prices. Additionally, I developed two scatterplots and a bar plot to illustrate how various property features correlate with house prices. These visualizations helped me gain valuable insights into how different factors influence property values in Saratoga, NY.

Data

Load Libraries and Packages

```
# Load required packages
if (!require("mosaic"))
  install.packages("mosaic")
if (!require("moderndive"))
  install.packages("moderndive")
if (!require("tidyverse"))
  install.packages("tidyverse")
if (!require("ggplot2"))
  install.packages("ggplot2")

library(mosaic) # Stats analysis
library(moderndive) # Datasets
library(tidyverse) # Data packages
library(ggplot2) # Data visualization
```

Description of data

The Saratoga Houses dataset, available in the moderndive package, includes 1,057 records detailing house prices and distinct property features such as price, square footage of living area, number of bathrooms and bedrooms, fireplaces, lot size, and the age of the houses. I selected this dataset because the variables it contains are well-suited for examining the relationship between house prices and various property attributes. These features provide valuable insights for studying the key factors that drive real estate values in Saratoga, NY.

> R Package: `moderndive`
>
> Data set: `saratoga_houses`

Load and Clean Data

```
# Load Data
data("saratoga_houses")
# Dataset dimensions
dim(saratoga_houses)
```

```
## [1] 1057    8
```

```r
# Check and remove any missing values
sum(is.na(saratoga_houses))
```

```
## [1] 9
```

```r
saratoga_houses <- saratoga_houses %>%
  na.omit()
# Updated dataset dimensions
dim(saratoga_houses)
```

```
## [1] 1048    8
```

The dataset initially contained 1,057 observations, but 9 missing values were identified and removed using the **na.omit()** function. After cleaning, the dataset was reduced to 1,048 observations, which will be used for the analysis.

Preview Data

```r
# Top ten rows of data
head(saratoga_houses, 10)
```

```
## # A tibble: 10 x 8
##     price living_area bathrooms bedrooms fireplaces lot_size   age fireplace
##     <dbl>       <dbl>     <dbl>    <dbl>      <dbl>    <dbl> <dbl> <lgl>
##  1 142212        1982       1          3          0     2      133 FALSE
##  2 134865        1676       1.5        3          1     0.38    14 TRUE
##  3 118007        1694       2          3          1     0.96    15 TRUE
##  4 138297        1800       1          2          2     0.48    49 TRUE
##  5 129470        2088       1          3          1     1.84    29 TRUE
##  6 206512        1456       2          3          0     0.98    10 FALSE
##  7 108794        1464       1          2          0     0.11    87 FALSE
##  8  68353        1216       1          2          0     0.61   101 FALSE
##  9 123266        1632       1.5        3          0     0.23    14 FALSE
## 10 309808        2270       2.5        3          2     4.05     9 TRUE
```

Variables

```r
# Variables
names(saratoga_houses)
```

```
## [1] "price"       "living_area" "bathrooms"   "bedrooms"    "fireplaces"
## [6] "lot_size"    "age"         "fireplace"
```

The following variables are featured in this data visualization project:

1. `price`
   A **quantitative** variable representing the house's sale price in US dollars.

2. `living_area`
   A **quantitative** variable that measures the total living area of the house in square feet.

3. `bathrooms`
   A **quantitative** variable indicating the number of bathrooms in the house. Note that half bathrooms do not have a shower or tub.

4. `bedrooms`
   A **quantitative** variable representing the total number of bedrooms in the house.

5. `fireplaces`
   A **quantitative** variable representing the number of fireplaces in the house.

6. `lot_size`
   A **quantitative** variable that shows the house's lot size in acres.

7. `age`
   A **quantitative** variable indicating the age of the house in years.

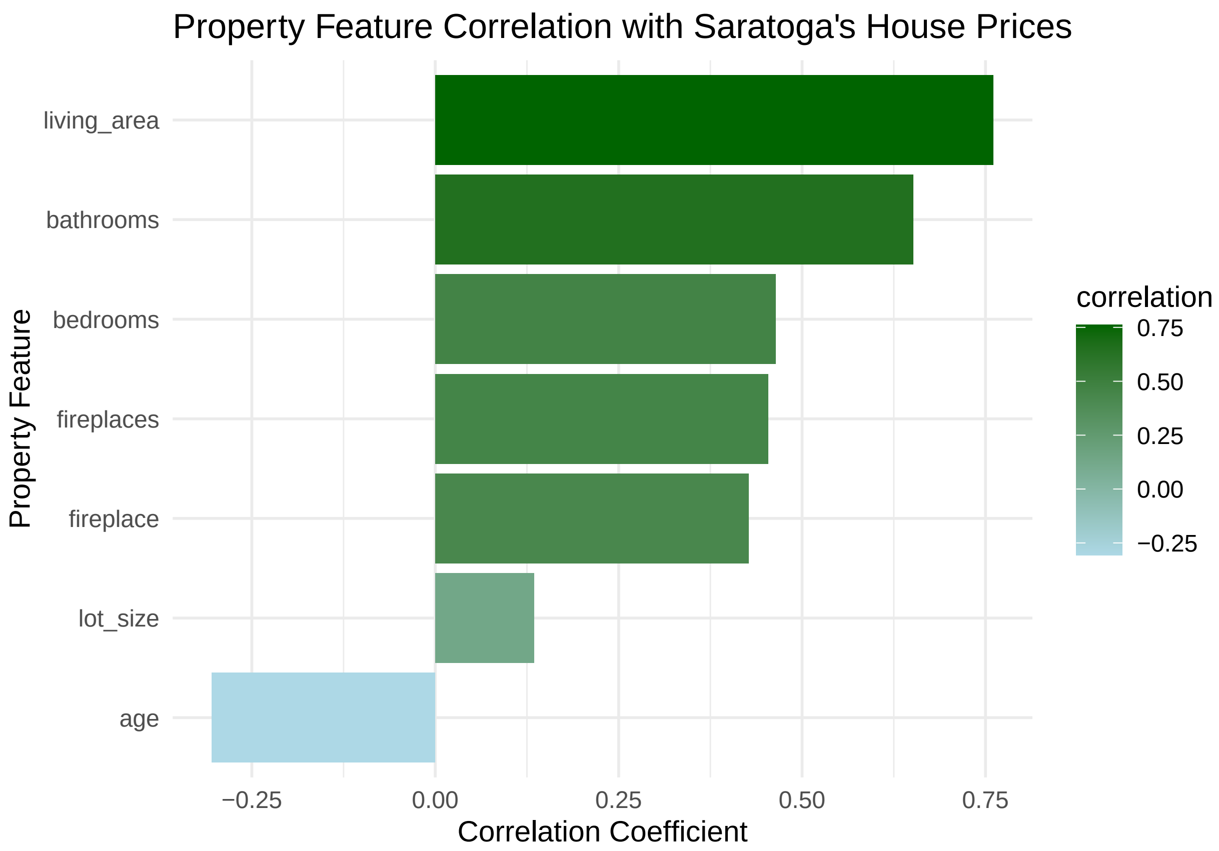
8. `fireplace`
   A **qualitative** variable that indicates whether the house has a fireplace or not, with values "TRUE" or "FALSE".

Data Analysis

Chart/Graph

```r
# Create a data frame of the correlation values
cor_values <- data.frame(
  variable = c("living_area", "bathrooms", "bedrooms", "fireplaces", "lot_size",
               "age", "fireplace"),
  correlation = c(0.7607, 0.6513, 0.4639, 0.4538, 0.1349, -0.3046, 0.4270)
)

# Correlation barplot
ggplot(cor_values, aes(x = reorder(variable, correlation), y = correlation,
                       fill = correlation)) +
  # Bar plot
  geom_bar(stat = "identity") +
  # Adjust color gradient
  scale_fill_gradient(low = "lightblue", high = "darkgreen") +
  # Horizontal bar plot
  coord_flip() +
  # Add title and labels
  labs(title = "Property Feature Correlation with Saratoga's House Prices",
       x = "Property Feature",
       y = "Correlation Coefficient") +
  # Minimal theme
  theme_minimal()
```



Upon calculating the correlations between house prices and the various variables in the dataset, it is evident that house prices tend to rise with an increase in living area size. Notably, the number of bathrooms shows a stronger correlation with price compared to the number of bedrooms. The barplot further highlights that age has the weakest correlation with price, which makes sense as newer homes are often perceived as more modern and, therefore, more expensive.

The property feature that has the *highest correlation* level to `price` is **living_area**

The property feature that has the *lowest correlation* level to `price` is **age**

Summary Statistics

```
# Calculate correlation matrix
corr_matrix <- cor(saratoga_houses[,c("price", "living_area", "bathrooms",
                                      "bedrooms", "fireplaces", "lot_size",
                                      "age", "fireplace")])

# Correlation of each variable with house prices
corr_matrix_prices <- corr_matrix["price", ]
corr_matrix_prices
```

```
##       price living_area   bathrooms    bedrooms  fireplaces    lot_size
##   1.0000000   0.7607178   0.6512717   0.4638828   0.4538355   0.1349124
##         age   fireplace
##  -0.3045985   0.4269994
```

```
# Summary statistics
favstats(~ price, data = saratoga_houses) # price
```

```
##    min     Q1 median     Q3    max     mean       sd    n missing
##  16858 112614 152258 206512 599701 167918.8 77152.67 1048       0
```

```
favstats(~ living_area, data = saratoga_houses) # living_area
```

```
##  min   Q1 median      Q3  max     mean       sd    n missing
##  672 1344   1679 2223.75 5228 1822.385 662.6726 1048       0
```

```
favstats(~ bathrooms, data = saratoga_houses) # bathrooms
```

```
##  min  Q1 median  Q3 max     mean        sd    n missing
##    1 1.5      2 2.5 4.5 1.928435 0.6516694 1048       0
```

```
favstats(~ bedrooms, data = saratoga_houses) # bedrooms
```

```
##  min Q1 median Q3 max    mean        sd    n missing
##    1  3      3  4   5 3.19084 0.7377345 1048       0
```

```
favstats(~ fireplaces, data = saratoga_houses) # fireplaces
```

```
##  min Q1 median Q3 max      mean        sd    n missing
##    0  0      1  1   4 0.6259542 0.5505698 1048       0
```

```
favstats(~ lot_size, data = saratoga_houses) # lot_size
```

```
##  min   Q1 median   Q3 max      mean        sd    n missing
##    0 0.21   0.39 0.61   9 0.5700477 0.7662969 1048       0
```

```
favstats(~ age, data = saratoga_houses) # age
```

```
##  min Q1 median Q3 max     mean      sd    n missing
##    0  6     18 34 247 28.10592 35.0353 1048       0
```

```
favstats(~ fireplace, data = saratoga_houses) # fireplace
```

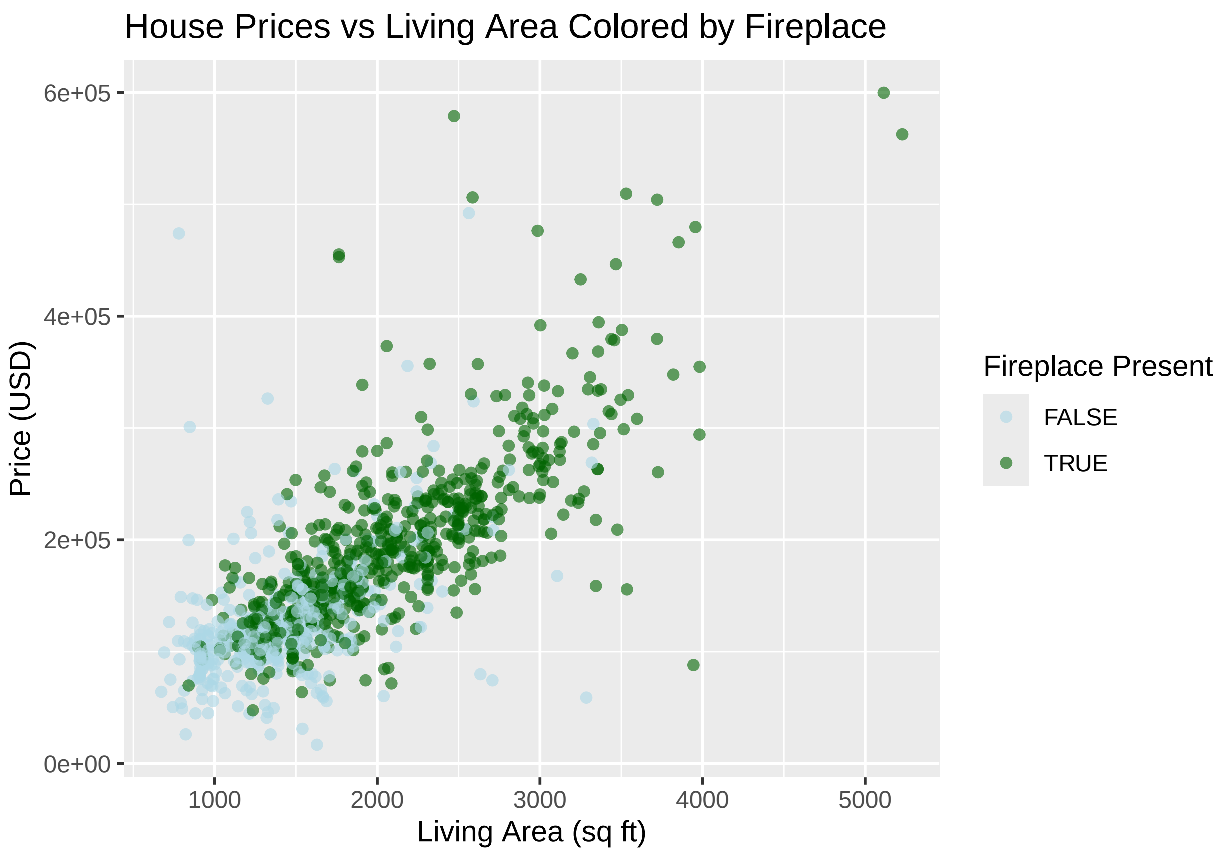
```
##  min Q1 median Q3 max     mean        sd    n missing
##    0  0      1  1   1 0.596374 0.4908584 1048       0
```

From the summary statistics above we can gather that the following average statistics for each property feature:

- `Price`: $167,918.80
- `Living Area`: 1822.39 sq ft
- `Bathrooms`: 1.93
- `Bedrooms`: 3.19
- `Fireplaces`: 0.63
- `Lot Size`: 0.57 acres
- `Age`: 28.11
- `Fireplace`: 0.60

Additional Analyses

```r
# Scatterplot house prices vs living area by presence of fireplace
ggplot(saratoga_houses, aes(x = living_area, y = price, color = fireplace)) +
  # Scatterplot
  geom_point(alpha = 0.6) +
  # Title and labels
  labs(title = "House Prices vs Living Area Colored by Fireplace",
       x = "Living Area (sq ft)",
       y = "Price (USD)",
       color = "Fireplace Present") +
  # Adjusting colors
  scale_color_manual(values = c("TRUE" = "darkgreen", "FALSE" = "lightblue"))
```



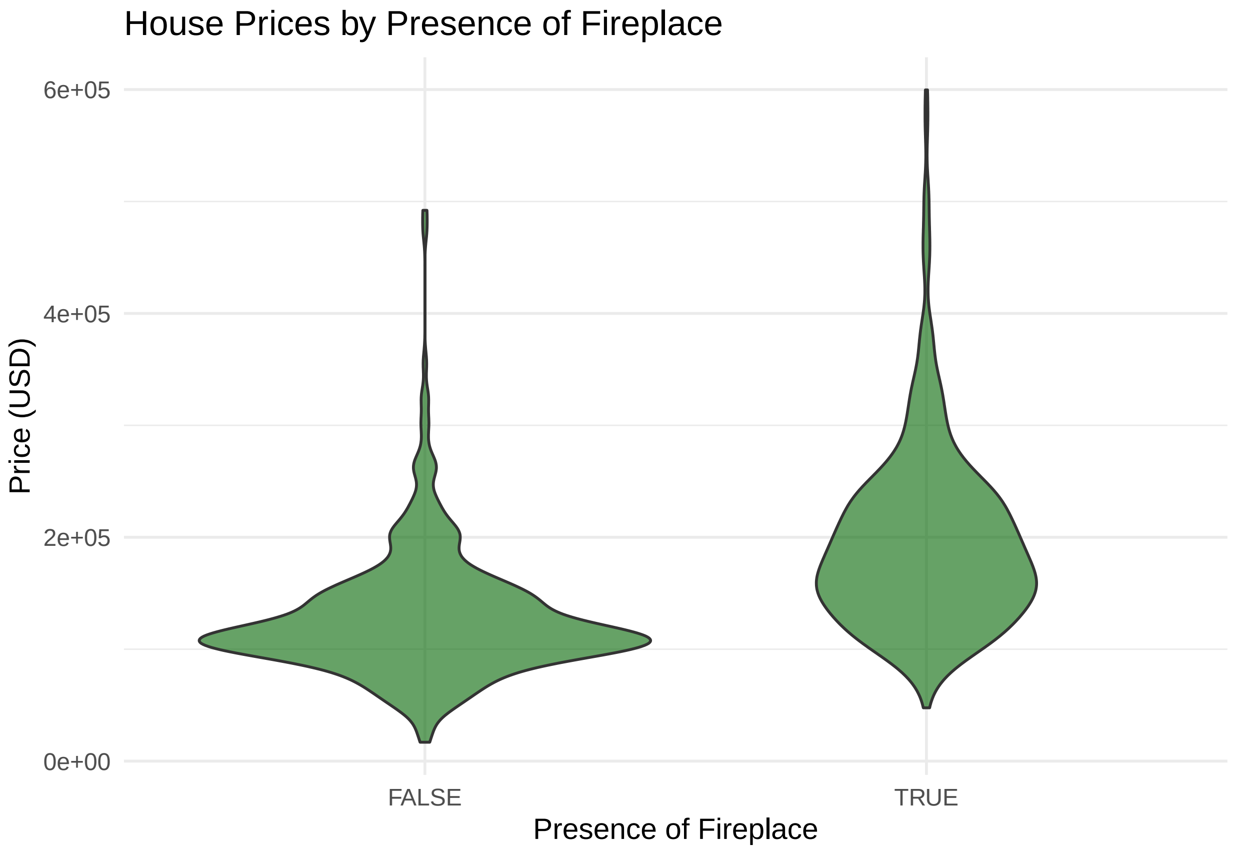
```r
  # Minimal theme
  theme_minimal()

# Total number of homes with/without a fireplace
y_fp <- sum(saratoga_houses$fireplace == "TRUE")
n_fp <- sum(saratoga_houses$fireplace == "FALSE")
```

This scatterplot illustrates the relationship between living area and price, while also indicating whether the homes have a fireplace. The plot provides valuable insights, showing that more affordable homes with smaller living spaces tend to lack a fireplace. This suggests that fireplaces are typically a luxury feature found in higher-priced homes. From this visualization, the following observations can be made:

- The size of a home tends to grow with its price.
- Homes with higher prices are more likely to feature a fireplace.
- Homes featuring a fireplace are more common than those without.
  - There are a total of 625 homes with a fireplace and 423 homes without one.

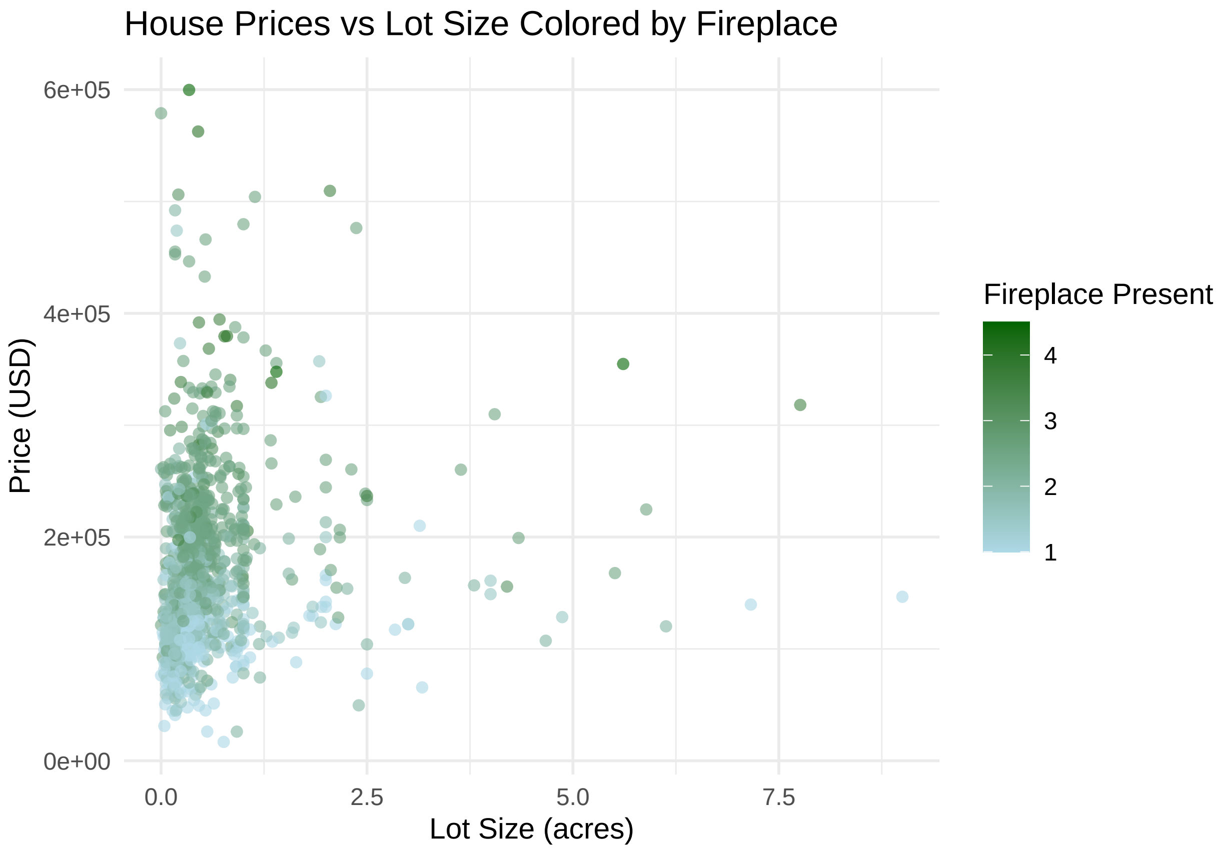
```r
# Violin Plot: House prices by presence of fireplace
ggplot(saratoga_houses, aes(x = as.factor(fireplace), y = price)) +
  # Violin plot
  geom_violin(fill = "darkgreen", alpha = 0.6) +
  # Title and labels
  labs(title = "House Prices by Presence of Fireplace",
       x = "Presence of Fireplace",
       y = "Price (USD)") +
  # Minimal theme
  theme_minimal()
```



The violin plot displays the distribution of house prices based on whether or not a fireplace is present. From this plot, we can observe the following:

- Homes that include a fireplace are priced higher than those that don't.
- A fireplace is often viewed as a luxury feature, signaling strong demand for this addition.
- The extended tail on the right suggests the presence of a few exceptional homes with fireplaces that are priced much higher than the rest.

```r
# Scatterplot house prices vs lot size by presence of fireplace
ggplot(saratoga_houses, aes(x = lot_size, y = price, color = bathrooms)) +
  # Scatterplot
  geom_point(alpha = 0.6) +
  # Title and labels
  labs(title = "House Prices vs Lot Size Colored by Fireplace",
       x = "Lot Size (acres)",
       y = "Price (USD)",
       color = "Fireplace Present") +
  # Adjusting colors
  scale_color_gradient(low = "lightblue", high = "darkgreen") +
  # Minimal theme
  theme_minimal()
```



Since living area size showed a strong correlation with price, I decided to create a visualization for lot size as well. Considering that fireplaces are often seen as a luxury feature in Saratoga homes, I included this variable in the scatterplot. From this visualization, the following observations can be made:

- As home prices rise, it's common to find homes with more than three fireplaces.
- The price of a home doesn't show a strong relationship with its lot size.
- Homes with 2-3 fireplaces are more common than those with just one or four fireplaces.

Infographic

A Statistical Analysis of Saratoga's Property Features and Prices by Nicole Rodriguez

Infographic Link

References

1. Violin Plot
2. The Best Visualization for You
3. R Markdown