Data-Driven Real Estate Investing

Team Members:

Wanyi (Evelyn) Dong
Shihhung (Angela) Ma
Jenny Shang
Kaiping (Rory) Wang
Yiran (Jenny) Wang

Agenda

01 Data Cleaning

02 Data Analysis

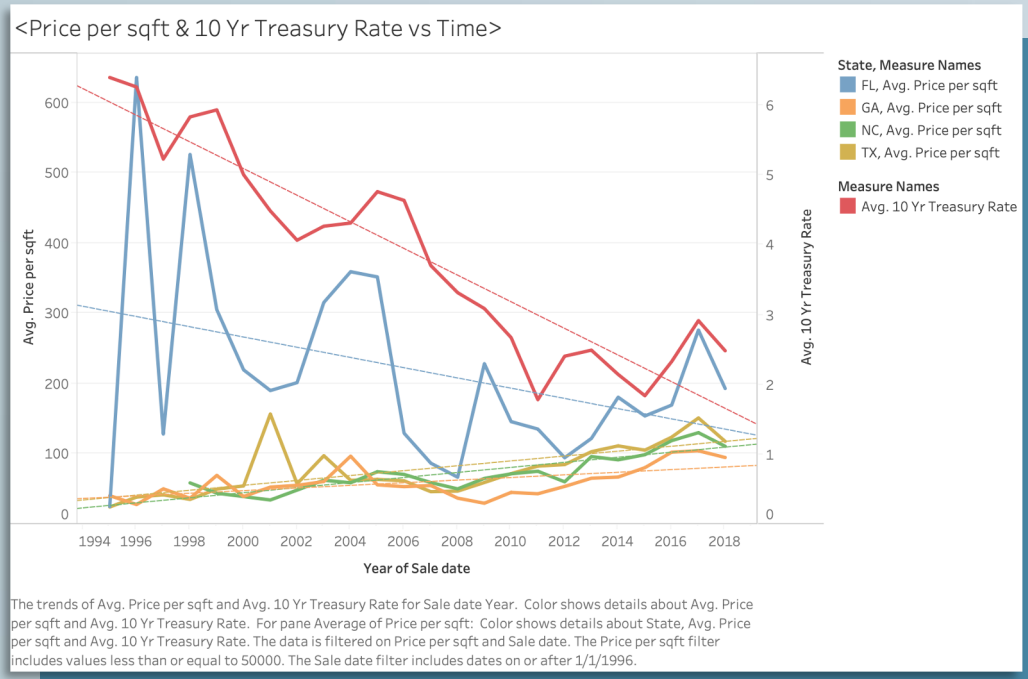
03 Regression Model

04 Recommendations & Future Works

Data Cleaning

Demographic Analysis

Allocate budget to different locations according to 10-year interest rate, geographic and demographic information



Geographic

- Trend from 1996 to 2018

Florida ↓

Texas, North Carolina, Georgia ↑

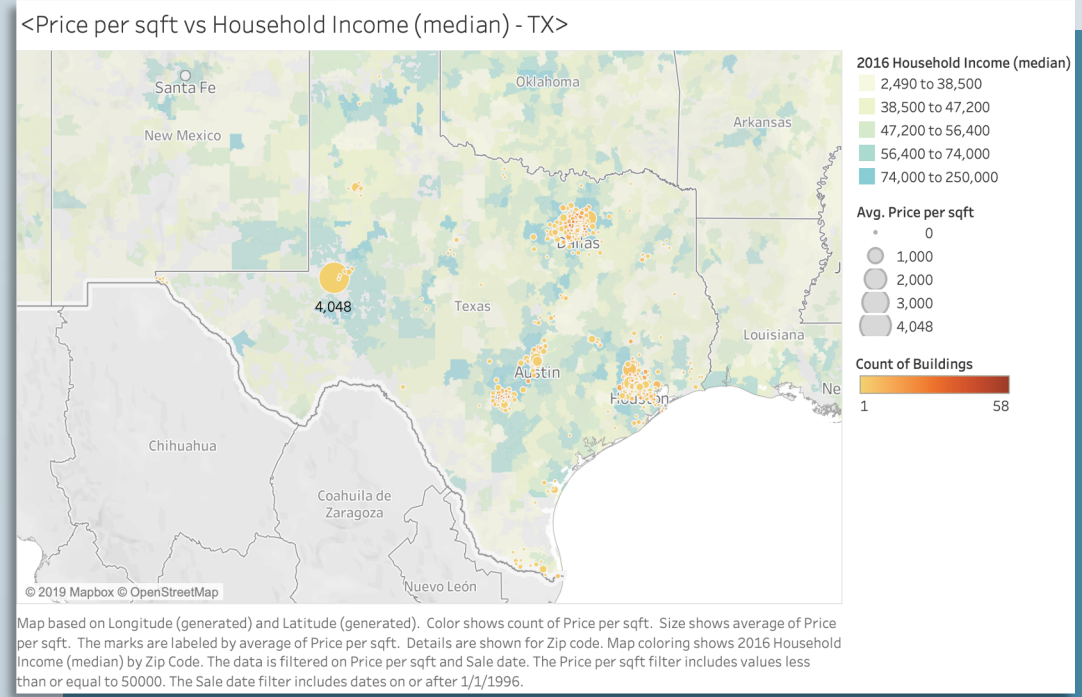
- **10-year treasury rate -** highly correlated with **Florida**'s building price

Demographic Analysis

Allocate budget to different locations according to 10-year interest rate, geographic and demographic information

Demographic

- Eight heatmaps
  - Four states
  - Household income
  - Per capita income
- Positive relationship between building's price per sqft and median household income
- Example: Ector at Texas



Word Cloud

Florida

North Carolina

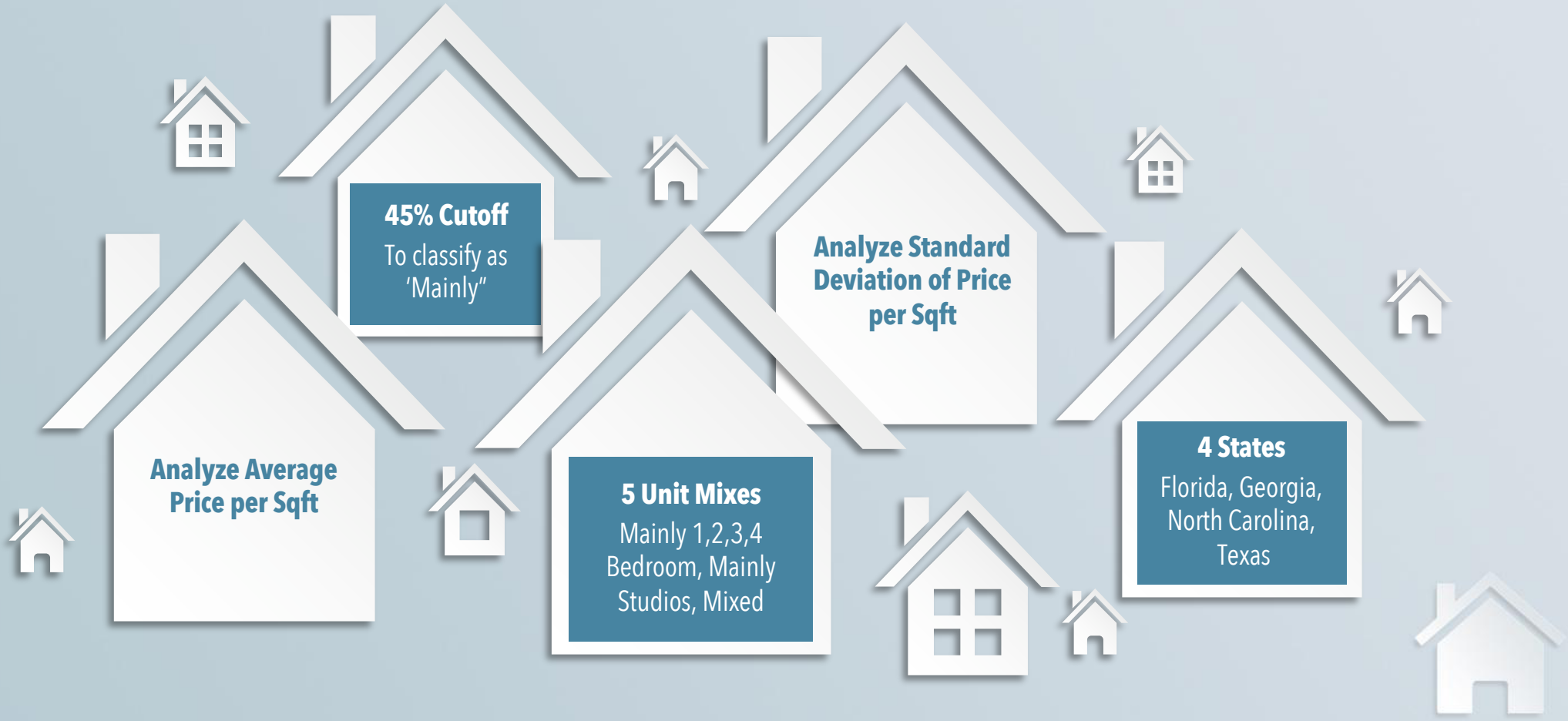
Texas

Georgia

Most common amenities

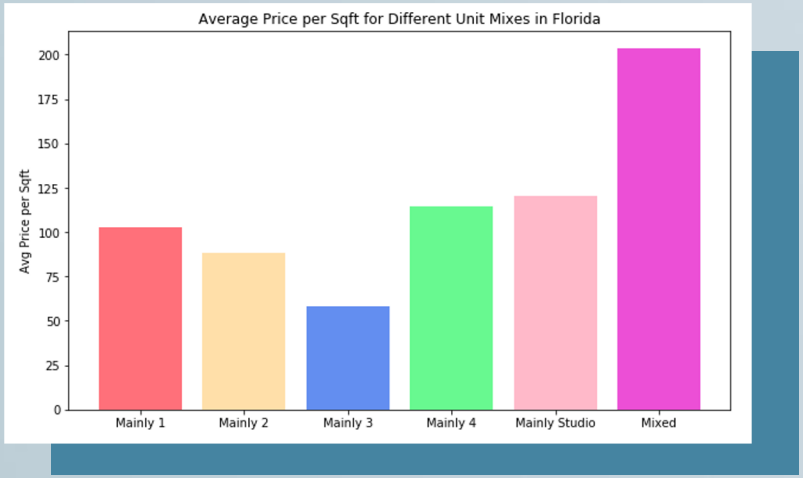
- Fitness center
- Business center
- Laundry Facilities
- Property Manager

Unit Mixes Analysis

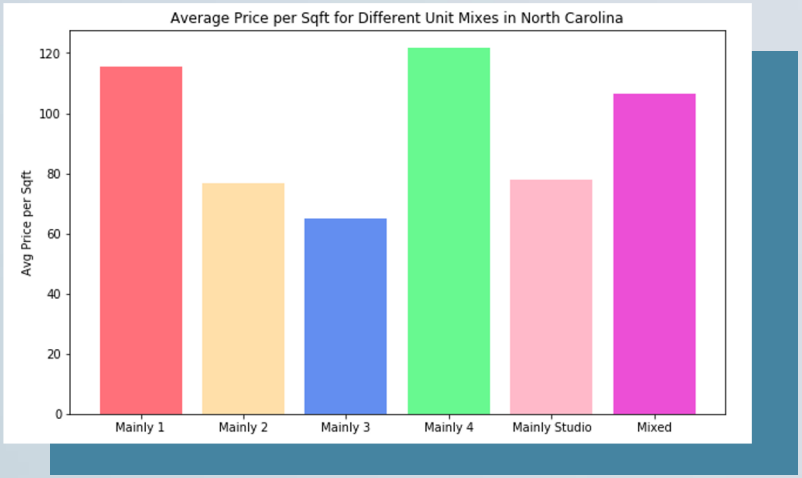


Unit Mixes Analysis

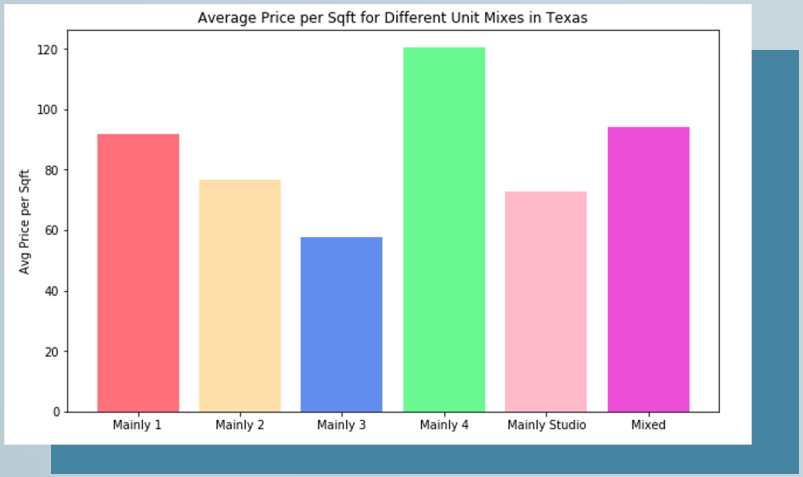
Florida



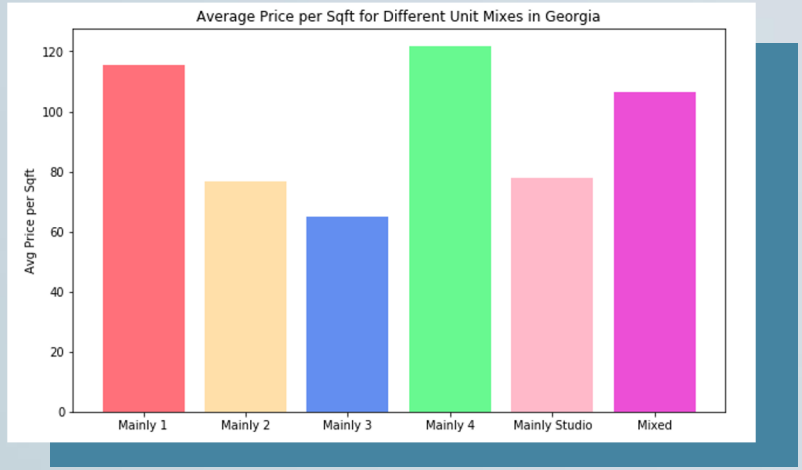
North Carolina



Texas



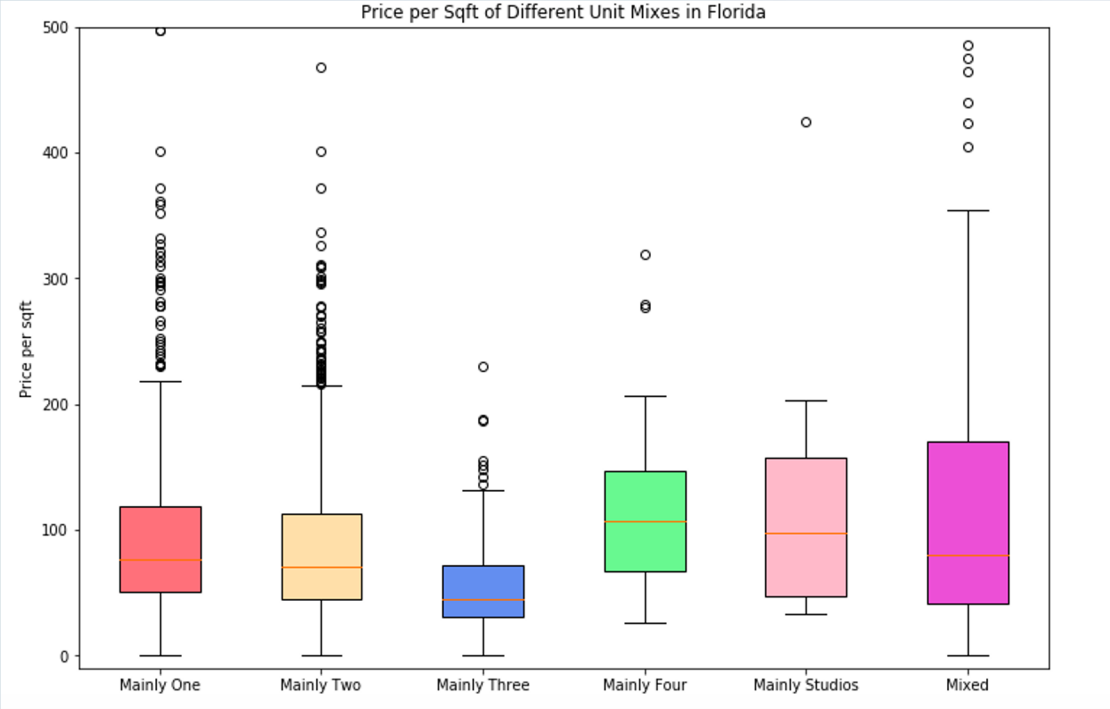
Georgia



Unit Mixes Analysis

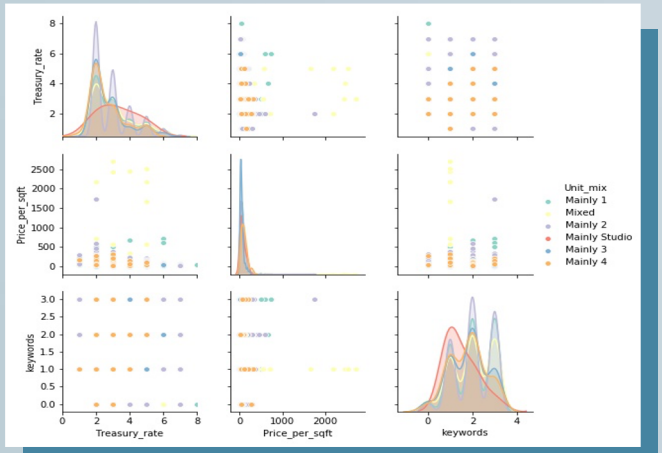
Recommendation

- Different states expand different unit mixes
    - **Florida**: Mainly Studios
    - **Texas**: Mainly 4 Bedrooms
    - **Georgia**: Mainly 3 Bedrooms
    - **North Carolina**: Mainly 4 Bedrooms

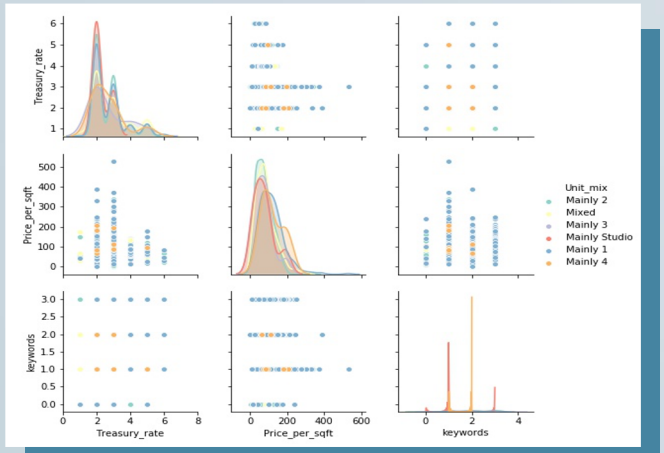


Pair Plot Analysis

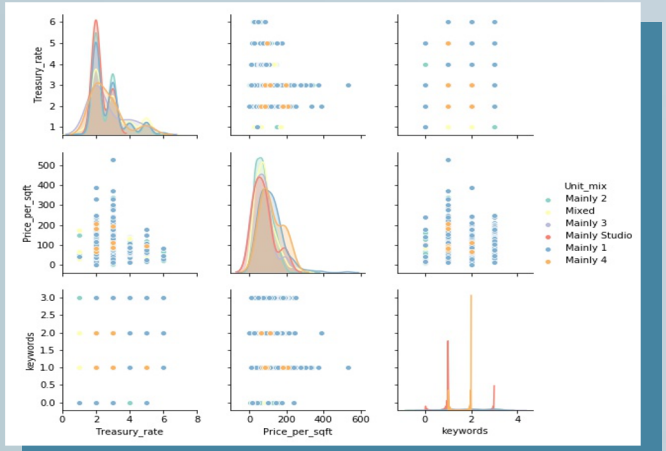
Florida



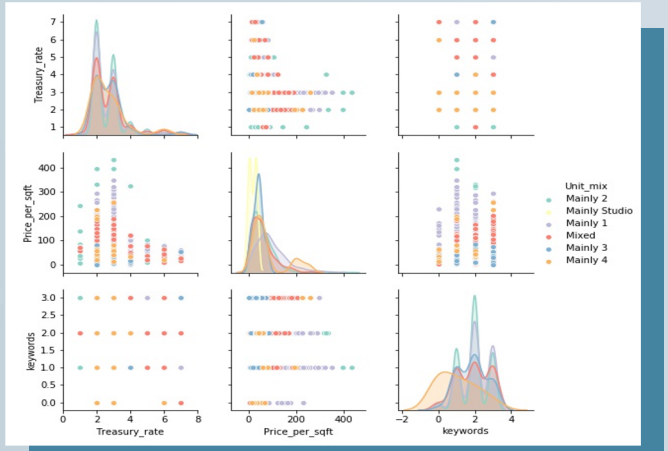
North Carolina



Texas



Georgia



Recommendation

- Highest price in each region happened when 10-year treasury rate is around 2-3
- Buildings with equal to or more than one amenities have relatively higher price

Regression Model

Florida

```
Call:
lm(formula = Price_per_sqft ~ Star_rating + Land_area + Num_units +
    Avg_unit_sqft + Parking_per_unit + Year_built + Floor_area_ratio +
    Num_floors, data = florida)

Residuals:
    Min      1Q  Median      3Q     Max
-1064.5   -51.0    -6.6    37.2 18517.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.803e+03  1.490e+03  -1.210  0.22654
Star_rating      -3.718e+01  1.654e+01  -2.248  0.02459 *
Land_area         3.395e+00  6.091e-01   5.573 2.77e-08 ***
Num_units         2.271e-01  7.263e-02   3.126  0.00179 **
Avg_unit_sqft     9.084e-02  4.360e-02   2.084  0.03730 *
Parking_per_unit -3.070e+01  1.190e+01  -2.580  0.00995 **
Year_built        9.287e-01  7.712e-01   1.204  0.22863
Floor_area_ratio  1.236e+01  1.563e+01   0.791  0.42918
Num_floors        3.042e+00  5.129e+00   0.593  0.55322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 388.7 on 2467 degrees of freedom
  (4103 observations deleted due to missingness)
Multiple R-squared:  0.03885,   Adjusted R-squared:  0.03574
F-statistic: 12.47 on 8 and 2467 DF,  p-value: < 2.2e-16
```

North Carolina

```
Call:
lm(formula = Price_per_sqft ~ Star_rating + Land_area + Num_units +
    Avg_unit_sqft + Parking_per_unit + Year_built + Floor_area_ratio +
    Num_floors, data = nc)

Residuals:
     Min       1Q   Median       3Q      Max
-134.730  -24.196   -5.476   21.764  234.487

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.958e+03  2.421e+02 -12.218  < 2e-16 ***
Star_rating       1.465e+01  2.975e+00   4.924 1.00e-06 ***
Land_area         1.488e-01  1.752e-01   0.849 0.396042
Num_units         1.602e-02  1.809e-02   0.886 0.376025
Avg_unit_sqft    -3.204e-02  9.078e-03  -3.529 0.000437 ***
Parking_per_unit -9.218e+00  2.132e+00  -4.324 1.70e-05 ***
Year_built        1.519e+00  1.253e-01  12.130  < 2e-16 ***
Floor_area_ratio  1.388e+01  2.690e+00   5.161 3.01e-07 ***
Num_floors        2.105e+00  7.952e-01   2.648 0.008245 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.47 on 926 degrees of freedom
  (1930 observations deleted due to missingness)
Multiple R-squared:  0.483,     Adjusted R-squared:  0.4785
F-statistic: 108.1 on 8 and 926 DF,  p-value: < 2.2e-16
```

Regression Model

Texas

```
Call:
lm(formula = Price_per_sqft ~ Star_rating + Land_area + Num_units +
    Avg_unit_sqft + Parking_per_unit + Year_built + Floor_area_ratio +
    Num_floors, data = texas)

Residuals:
    Min      1Q  Median      3Q     Max
-113.32  -23.85   -5.76   19.84  567.93

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.823e+03  2.227e+02 -17.167  < 2e-16 ***
Star_rating      3.763e+00  2.333e+00   1.613  0.10691
Land_area       -3.113e-01  1.622e-01  -1.920  0.05507 .
Num_units        2.731e-02  9.628e-03   2.836  0.00461 **
Avg_unit_sqft   -1.768e-02  7.775e-03  -2.274  0.02309 *
Parking_per_unit 8.537e-01  1.390e+00   0.614  0.53917
Year_built       1.956e+00  1.152e-01  16.977  < 2e-16 ***
Floor_area_ratio -2.331e-02  2.576e-02  -0.905  0.36561
Num_floors       2.897e+00  6.302e-01   4.597 4.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.54 on 1901 degrees of freedom
  (9718 observations deleted due to missingness)
Multiple R-squared:  0.3584,    Adjusted R-squared:  0.3557
F-statistic: 132.8 on 8 and 1901 DF,  p-value: < 2.2e-16
```

Georgia

```
Call:
lm(formula = Price_per_sqft ~ Star_rating + Land_area + Num_units +
    Avg_unit_sqft + Parking_per_unit + Year_built + Floor_area_ratio +
    Num_floors, data = georgia)

Residuals:
    Min      1Q  Median      3Q     Max
-159.37  -27.94   -7.22   16.43 2248.58

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.336e+03  3.661e+02  -6.381 2.30e-10 ***
Star_rating      1.589e+01  3.982e+00   3.991 6.87e-05 ***
Land_area        8.460e-01  2.268e-01   3.730 0.000198 ***
Num_units       -3.242e-02  2.445e-02  -1.326 0.184953
Avg_unit_sqft   -2.915e-02  1.153e-02  -2.527 0.011609 *
Parking_per_unit -9.720e+00  2.718e+00  -3.576 0.000360 ***
Year_built       1.201e+00  1.882e-01   6.382 2.28e-10 ***
Floor_area_ratio 9.600e+00  4.130e+00   2.324 0.020233 *
Num_floors       2.245e+00  2.170e+00   1.035 0.301051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.28 on 1585 degrees of freedom
  (2001 observations deleted due to missingness)
Multiple R-squared:  0.1512,    Adjusted R-squared:  0.1469
F-statistic:  35.3 on 8 and 1585 DF,  p-value: < 2.2e-16
```

Recommendations

01 **Allocate budget according to changes in 10-year US Treasury rate**

Invest in GA, NC, and TX when rate is expected to decrease or stay the same; invest in FL when the rate is expected to increase.

02 **Select properties that contain amenities that positively contribute to price per sqft**

Focus on buildings that include fitness centers and business centers as amenities.

03 **Invest more into buildings that are either mainly four bedrooms or a mix of different units**

Mainly 4-bedroom in Georgia and North Carolina, Texas, and studios in Florida.

Future Works

01 Continue to record amenities

Look further into the amenities that do not increase the price per sqft, such as laundry facilities and property manager on site.

02 Continue to record the percentage of unit type

Further analysis can be conducted to see if the change of unit mix of a building would change the profit.

03 Use external variables to select optimal locations

Variables such as crime rate, the schools nearby, the number of grocery stores or coffee chains in the area, or distance to the nearest bus stop could also be analyzed.

04 Use machine learning to select locations

By automating the building of analytical models, this would make Centro's goal of finding better locations more efficient.

Thank you!

Questions?