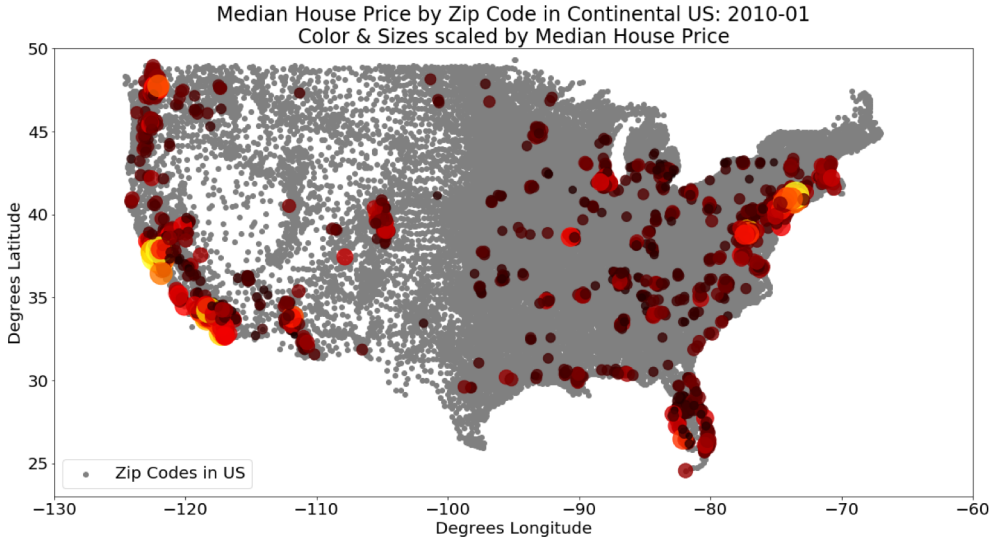
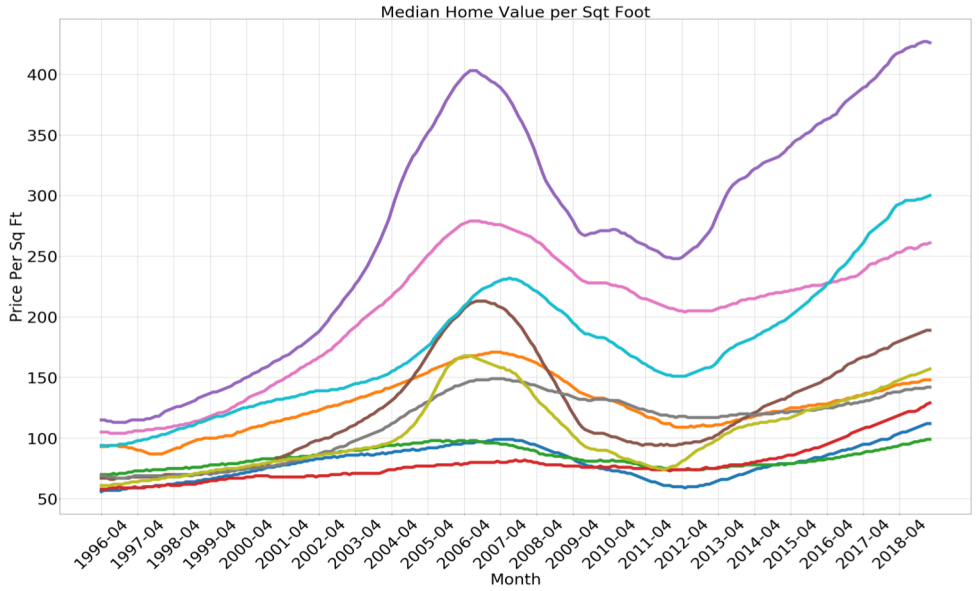
An Analysis of the Factors that Impact Housing Costs in the US

Kim Harrison
Kent Hazen
Tim Tang
Laura De Morneau

9 April 2019





Our approach was to utilize datasets from different sources to determine what factors impact housing prices

Rent or Buy?

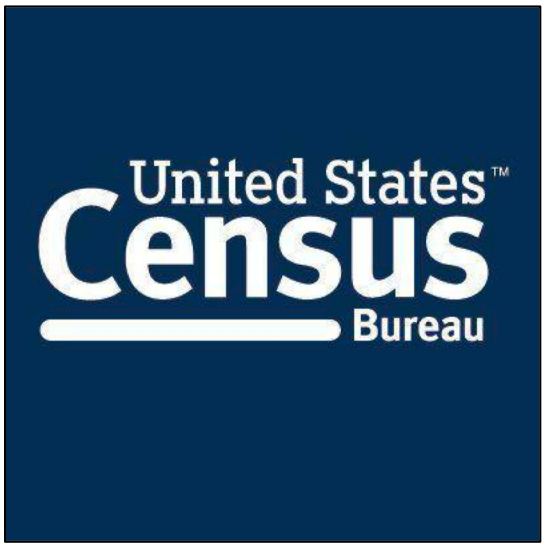
Demographics

Severe Weather Events

House Layout

Our approach was to utilize datasets from different sources to determine what factors impact housing prices

House Prices, Amenities

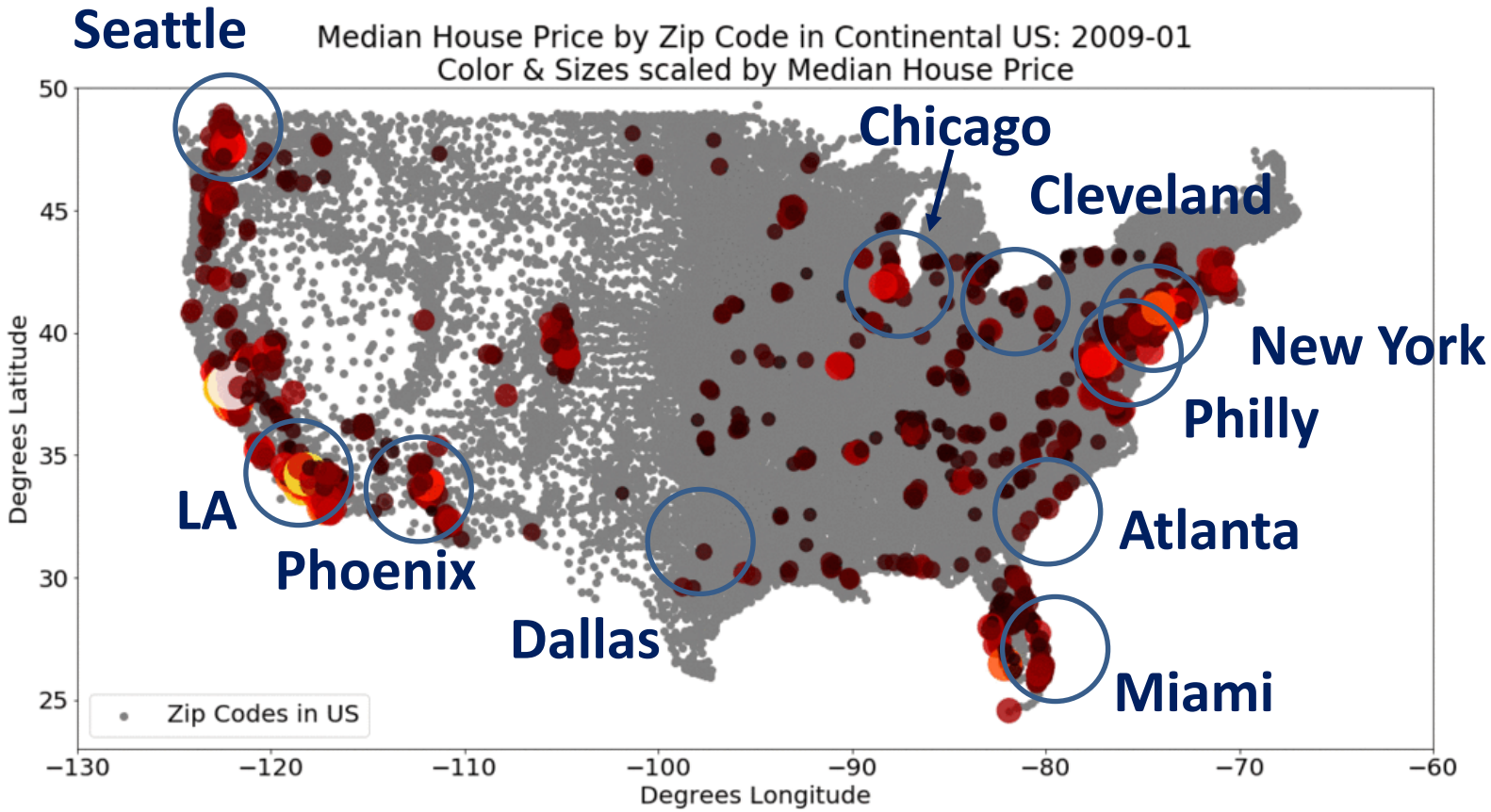


Demographic, City Info

FEMA

Severe Weather Events

While cleaning our data, we decided to select 10 target cities to analyze housing prices to answer our questions



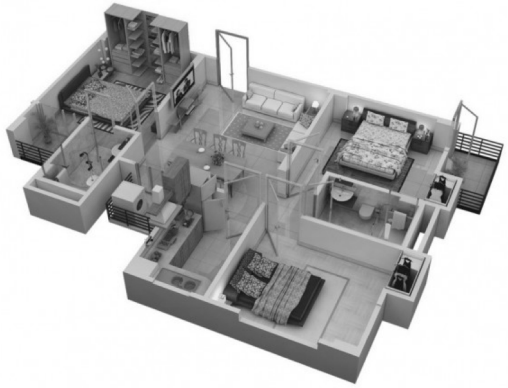
Housing Price Snap Shot every 6 months over the last 10 years

Our approach was to utilize datasets from different sources to determine what factors impact housing prices

Rent or Buy?

Demographics

Severe Weather Events



House Layout

Average house price was converted to mortgage payments using a function to compare buying versus renting

	RegionID	RegionName	SizeRank	2016-10	2016-11	2016-12	2017-01	2017-02	2017-03
0	394347	Atlanta	9	179500.0	182500.0	183500.0	185200.0	187100.0	191200.0
1	394463	Chicago	3	193900.0	195400.0	199600.0	202400.0	204400.0	206200.0
2	394475	Cleveland	28	122400.0	123100.0	123700.0	124900.0	126900.0	127000.0
3	394514	Dallas	4	220000.0	220200.0	220100.0	221300.0	226100.0	230900.0
4	753899	Los Angeles	2	531500.0	530200.0	527700.0	530500.0	525300.0	529100.0
5	394856	Miami	8	228200.0	228800.0	229600.0	228700.0	235200.0	237200.0
6	394913	New York	1	362100.0	361000.0	360700.0	356700.0	360200.0	363600.0

```
def sale_to_rent(sale):
    end = 0
    months = 360
    interest = .04
    Principal = sale - end
    pay_a = (interest / 12) / (1 - (1+interest/12)**(-months)) * Principal
    pay_b = interest / 12 * end
    monthly_payment = (interest / 12) * (1 / (1 - (1+interest/12) ** (-months))) * Principal + end


    return (monthly_payment)
```

```
range2.loc[:, '2016-10':'2019-01'] = range2.loc[:, '2016-10':'2019-01'].applymap(sale_to_rent)
range2.head()
```

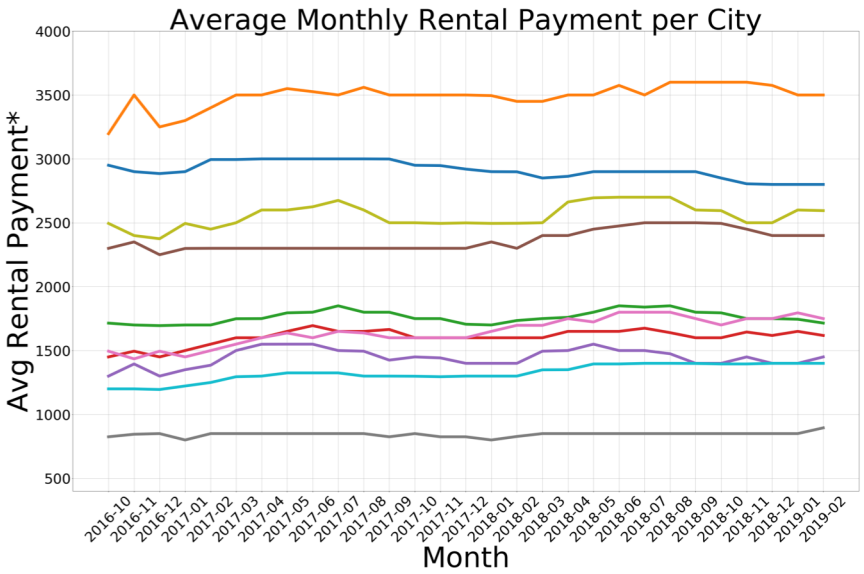
	2016-10	2016-11	2016-12	2017-01	2017-02	2017-03	2017-04	2017-05	2017-06	2017-07	...	2018-04
0	856.960455	871.282914	876.057067	884.173127	893.244018	912.818045	923.798597	930.959826	926.185673	919.979274	...	994.456060
1	925.708258	932.869487	952.920930	966.288558	975.836864	984.430339	988.727077	985.385170	972.972372	959.604744	...	1090.893950
2	584.356322	587.698229	590.562720	596.291704	605.840010	606.317425	607.749671	605.362595	600.588442	594.859458	...	646.420310
3	1050.313650	1051.268481	1050.791065	1056.520049	1079.435983	1102.351917	1110.467977	1107.603485	1099.487425	1102.351917	...	1181.602856

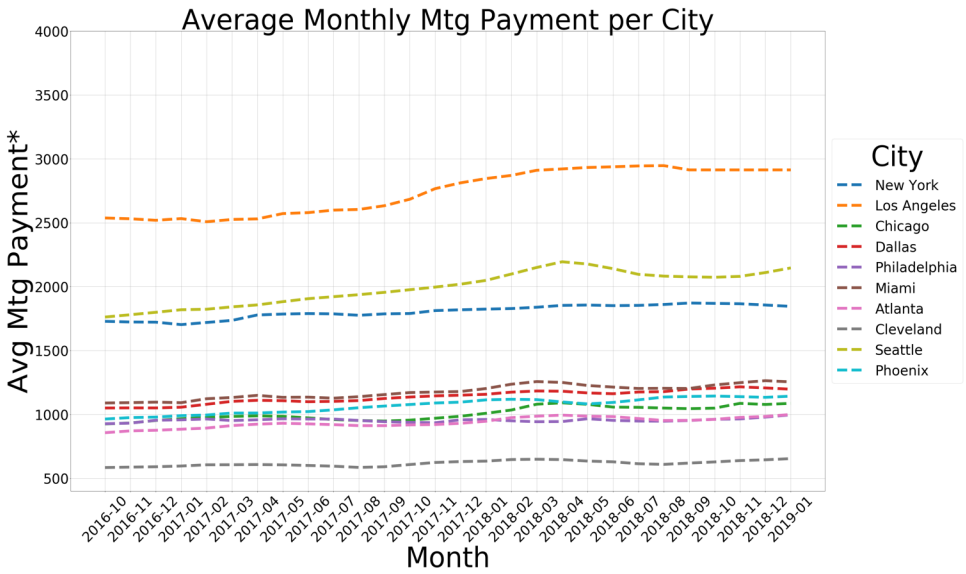
Sales Data Not in Monthly Payment Format

Created a function to convert to monthly payment

Applied the function to each cell

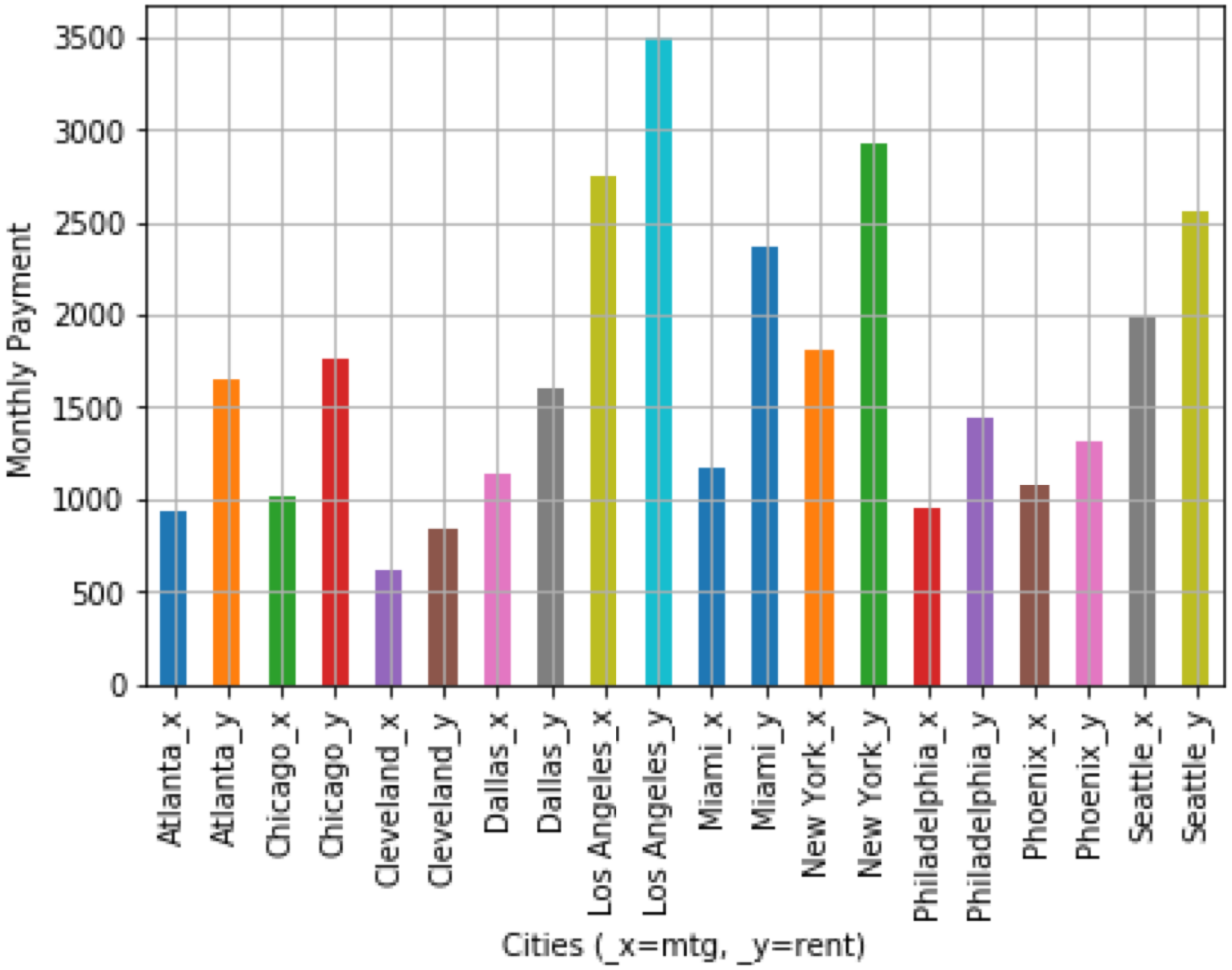
Mortgage payments versus rental payments show if you should buy or rent in a given city





Assumptions:

- **For Sales Price to Monthly Payment conversion, payments are based on 4% interest rate and 30 year term**
- **Monthly payment comparison does not factor down payment, property taxes, home maintenance**
- **Population used does not reflect rural or suburban area**
- **Used historical average to fill in blank months**
- **Limited Date Range, used from 2016 thru 2019 data only**



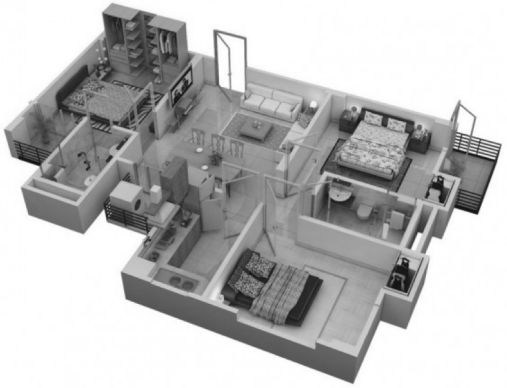
- **Highest Rental Payments**
  - #1 Los Angeles
  - #2 New York City
- **Lowest Rental Payments**
  - #1 Cleveland
  - #2 Phoenix
- **Highest Mortgage Payments**
  - #1 Los Angeles
  - #2 Seattle
- **Lowest Mortgage Payments**
  - #1 Cleveland
  - #2 Atlanta
- **Biggest Differences in Payments**
  - #1 Miami
  - #2 New York

Our approach was to utilize datasets from different sources to determine what factors impact housing prices

Rent or Buy?

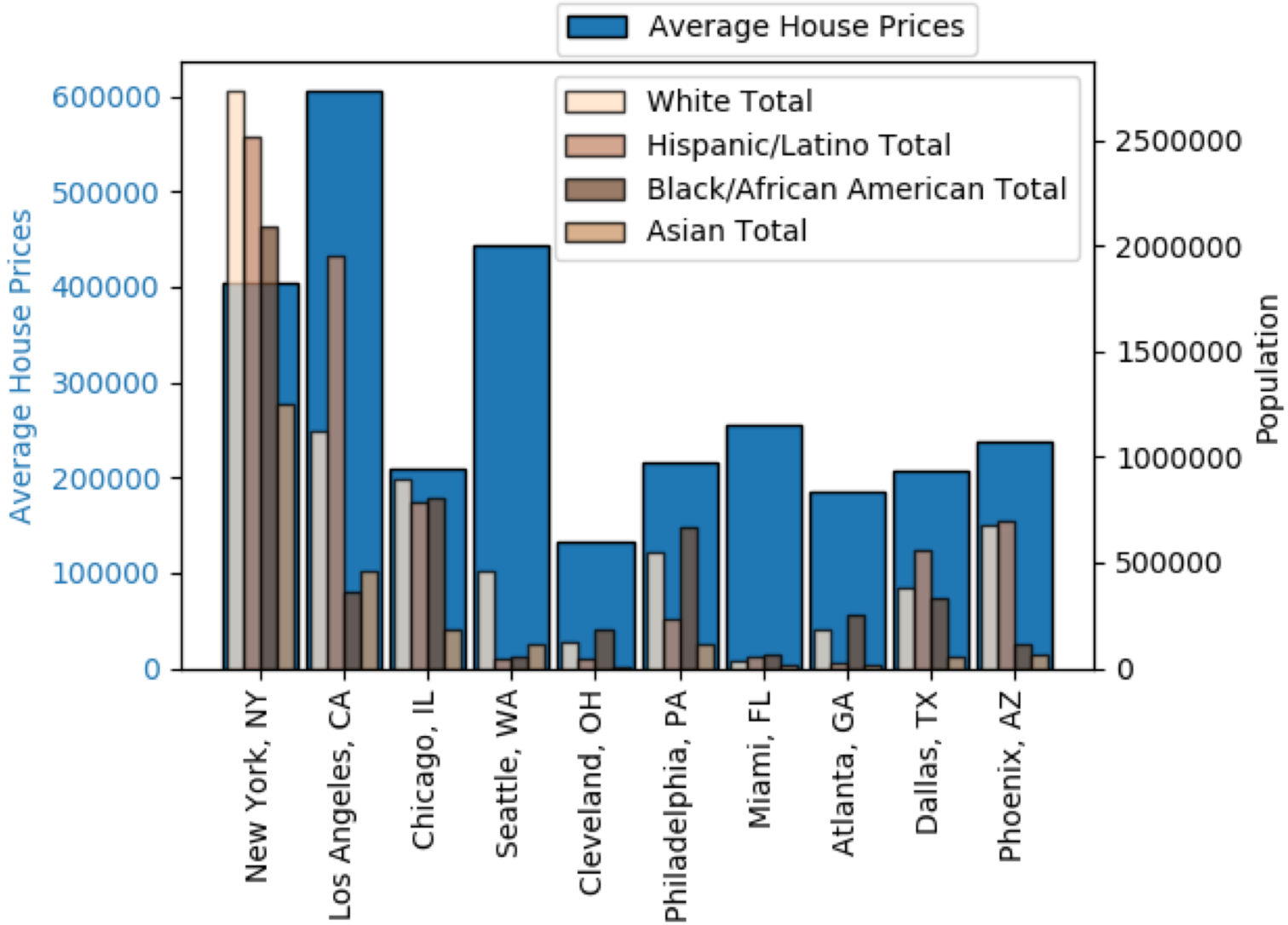
Demographics

Severe Weather Events



House Layout

Using a t-test, racial makeup of cities showed no statistical significance on housing prices

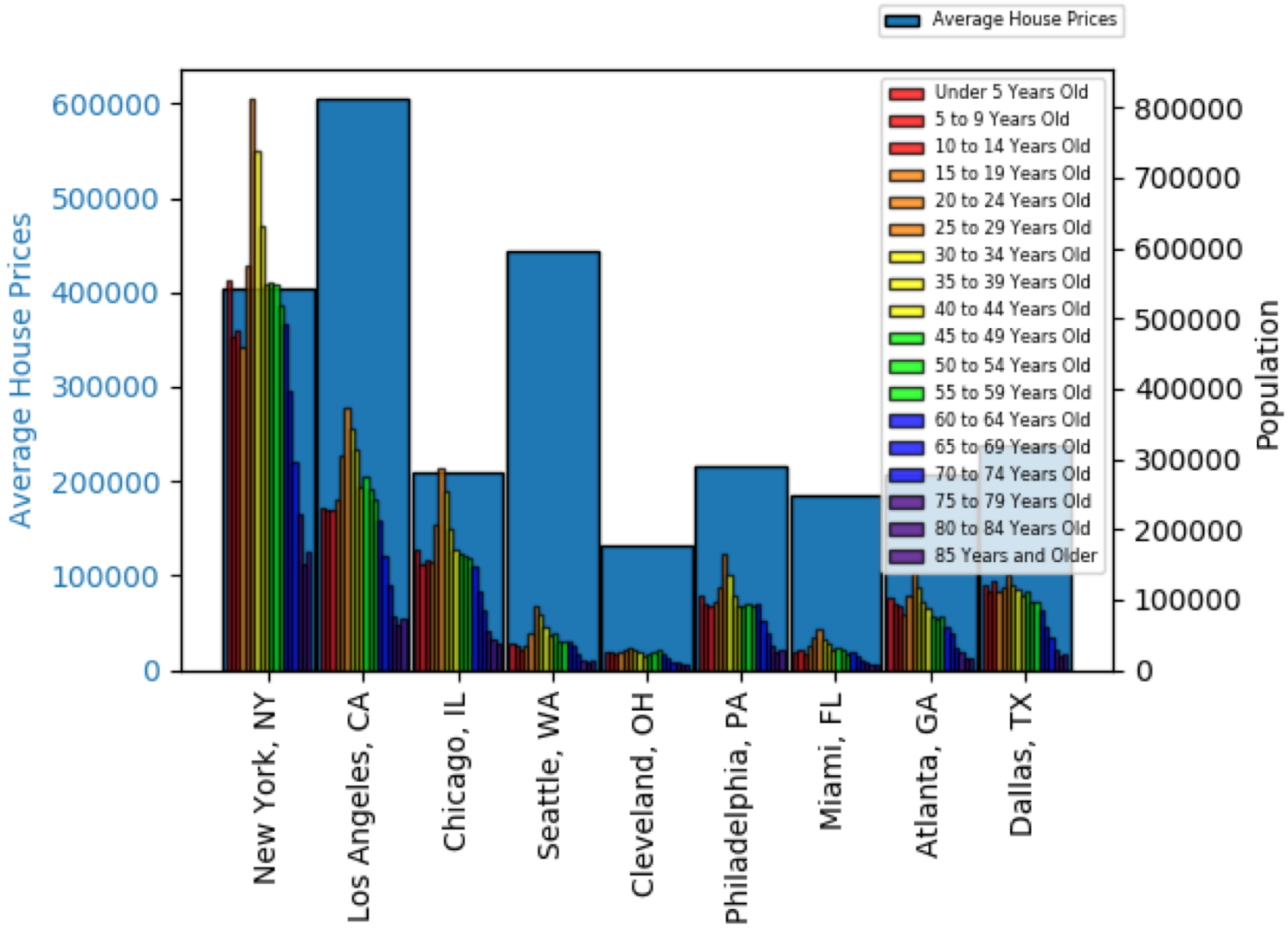


Using a t-test, racial makeup of cities showed no statistical significance on housing prices

In [49]:
```
print(f"White P-Value: {stats.ttest_ind(race_demographics_final_df['White Total'
print(f"Hispanic/Latino P-Value: {stats.ttest_ind(race_demographics_final_df['Hi
print(f"Black/African American P-Value: {stats.ttest_ind(race_demographics_final
print(f"Asian P-Value: {stats.ttest_ind(race_demographics_final_df['Asian Total'
```

```
White P-Value: 0.12366086648631457
Hispanic/Latino P-Value: 0.18216677315602117
Black/African American P-Value: 0.33307482309733916
Asian P-Value: 0.6462703001067476
```

Using a t-test, age was shown to have a statistical significance on housing prices



Using a t-test, age was shown to have a statistical significance on housing prices

In [56]:
```
print(f"Under 5 Years Old P-Value:  {stats.ttest_ind(age_demographics_final_df['Under 5 Years Old'], age_demographics_final_df['A
print(f"10 to 14 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['10 to 14 Years Old'], age_demographics_final_df[
print(f"15 to 19 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['15 to 19 Years Old'], age_demographics_final_df[
print(f"20 to 24 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['20 to 24 Years Old'], age_demographics_final_df[
print(f"25 to 29 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['25 to 29 Years Old'], age_demographics_final_df[
print(f"30 to 34 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['30 to 34 Years Old'], age_demographics_final_df[
print(f"35 to 39 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['35 to 39 Years Old'], age_demographics_final_df[
print(f"40 to 44 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['40 to 44 Years Old'], age_demographics_final_df[
print(f"45 to 49 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['45 to 49 Years Old'], age_demographics_final_df[
print(f"50 to 54 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['50 to 54 Years Old'], age_demographics_final_df[
print(f"55 to 59 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['55 to 59 Years Old'], age_demographics_final_df[
print(f"60 to 64 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['60 to 64 Years Old'], age_demographics_final_df[
print(f"65 to 69 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['65 to 69 Years Old'], age_demographics_final_df[
print(f"70 to 74 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['70 to 74 Years Old'], age_demographics_final_df[
print(f"75 to 79 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['75 to 79 Years Old'], age_demographics_final_df[
print(f"80 to 84 Years Old P-Value: {stats.ttest_ind(age_demographics_final_df['80 to 84 Years Old'], age_demographics_final_df[
print(f"85 Years and Older P-Value: {stats.ttest_ind(age_demographics_final_df['85 Years and Older'], age_demographics_final_df[
```

```
Under 5 Years Old P-Value:  0.08071595809872698
10 to 14 Years Old P-Value: 0.0455818478053208
15 to 19 Years Old P-Value: 0.04016827640582783
20 to 24 Years Old P-Value: 0.14016317765728406
25 to 29 Years Old P-Value: 0.5367427185502082
30 to 34 Years Old P-Value: 0.35511350101204975
35 to 39 Years Old P-Value: 0.17751944694765193
40 to 44 Years Old P-Value: 0.07990826361242098
45 to 49 Years Old P-Value: 0.08477679442958697
50 to 54 Years Old P-Value: 0.07335182492806476
55 to 59 Years Old P-Value: 0.05691960682299336
60 to 64 Years Old P-Value: 0.037347181660099656
65 to 69 Years Old P-Value: 0.010707767643484365
70 to 74 Years Old P-Value: 0.0031071298232364774
75 to 79 Years Old P-Value: 0.0013878118771841097
80 to 84 Years Old P-Value: 0.0009729498204263203
85 Years and Older P-Value: 0.0010374386907740856
```

Using a t-test, age was shown to have a statistical significance on housing prices

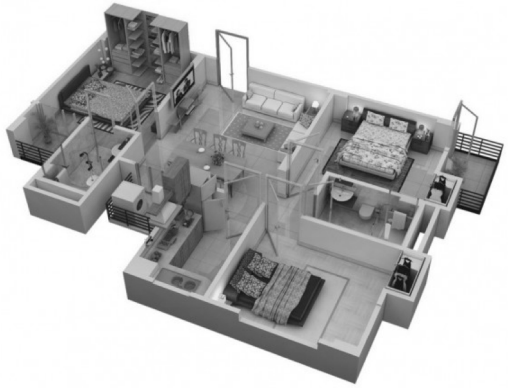
```
Under 5 Years Old P-Value:  0.080715958098726 98
10 to 14 Years Old P-Value: 0.0455818478053208
15 to 19 Years Old P-Value: 0.04016827640582783
20 to 24 Years Old P-Value: 0.14016317765728406
25 to 29 Years Old P-Value: 0.5367427185502082
30 to 34 Years Old P-Value: 0.35511350101204975
35 to 39 Years Old P-Value: 0.17751944694765193
40 to 44 Years Old P-Value: 0.07990826361242098
45 to 49 Years Old P-Value: 0.08477679442958697
50 to 54 Years Old P-Value: 0.07335182492806476
55 to 59 Years Old P-Value: 0.05691960682299336
60 to 64 Years Old P-Value: 0.037347181660099656
65 to 69 Years Old P-Value: 0.010707767643484365
70 to 74 Years Old P-Value: 0.0031071298232364774
75 to 79 Years Old P-Value: 0.0013878118771841097
80 to 84 Years Old P-Value: 0.0009729498204263203
85 Years and Older P-Value: 0.0010374386907740856
```

Our approach was to utilize datasets from different sources to determine what factors impact housing prices

Rent or Buy?

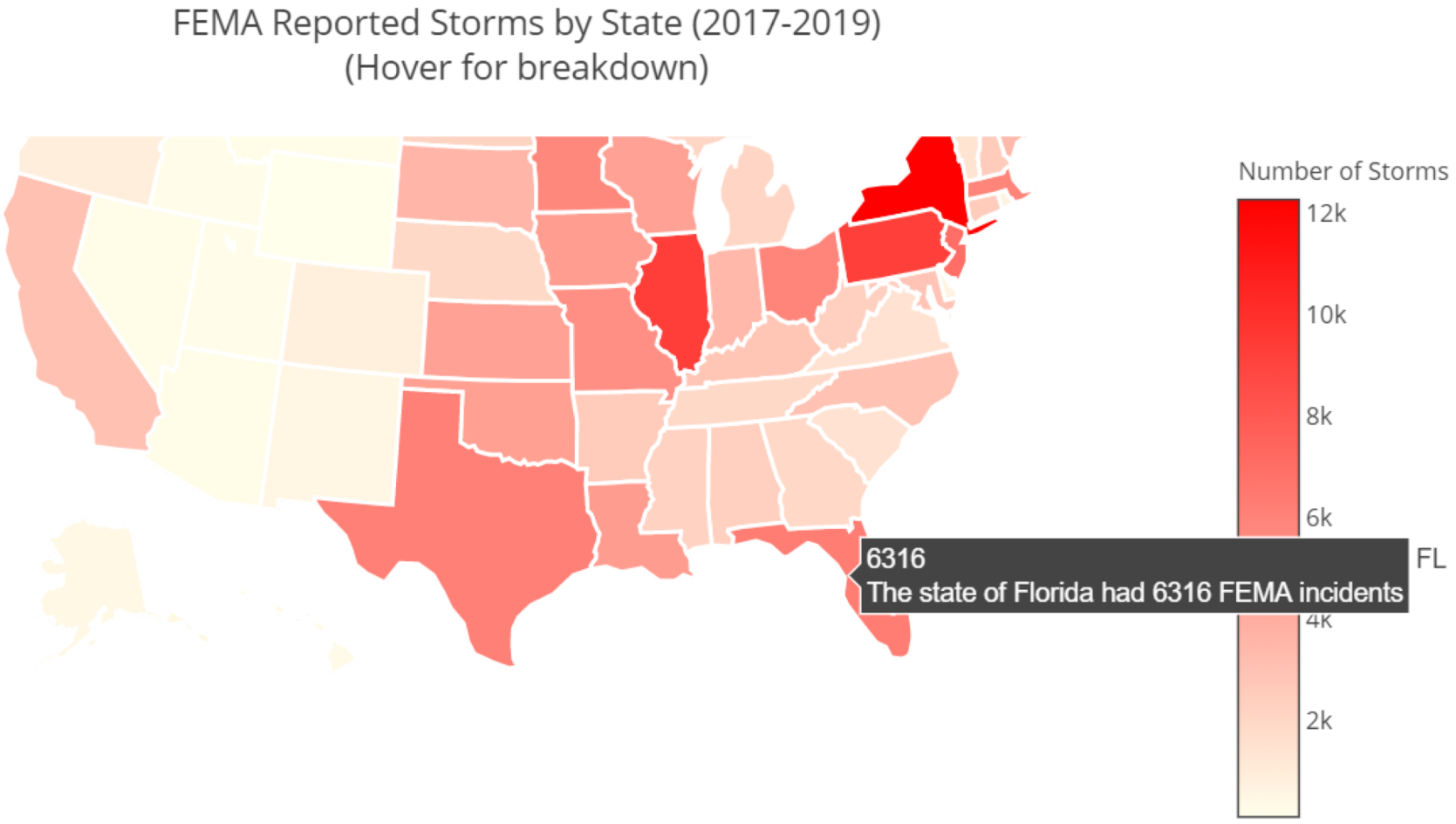
Demographics

Severe Weather Events



House Layout

FEMA data was analyzed to calculate the number of severe weather events by location



FEMA data was analyzed to calculate the number of severe weather events by location

In [34]:

```
example = fema_data_clean.groupby('State')['County'] \
                         .count() \
                         .reset_index() \
                         .rename(columns={'County': 'Number of Storms'})

state_codes = {
    'District of Columbia' : 'dc','Mississippi': 'MS', 'Oklahoma': 'OK',
    'Delaware': 'DE', 'Minnesota': 'MN', 'Illinois': 'IL', 'Arkansas': 'AR',
    'New Mexico': 'NM', 'Indiana': 'IN', 'Maryland': 'MD', 'Louisiana': 'LA',
    'Idaho': 'ID', 'Wyoming': 'WY', 'Tennessee': 'TN', 'Arizona': 'AZ',
    'Iowa': 'IA', 'Michigan': 'MI', 'Kansas': 'KS', 'Utah': 'UT',
    'Virginia': 'VA', 'Oregon': 'OR', 'Connecticut': 'CT', 'Montana': 'MT',
    'California': 'CA', 'Massachusetts': 'MA', 'West Virginia': 'WV',
    'South Carolina': 'SC', 'New Hampshire': 'NH', 'Wisconsin': 'WI',
    'Vermont': 'VT', 'Georgia': 'GA', 'North Dakota': 'ND',
    'Pennsylvania': 'PA', 'Florida': 'FL', 'Alaska': 'AK', 'Kentucky': 'KY',
    'Hawaii': 'HI', 'Nebraska': 'NE', 'Missouri': 'MO', 'Ohio': 'OH',
    'Alabama': 'AL', 'Rhode Island': 'RI', 'South Dakota': 'SD',
    'Colorado': 'CO', 'New Jersey': 'NJ', 'Washington': 'WA',
    'North Carolina': 'NC', 'New York': 'NY', 'Texas': 'TX',
    'Nevada': 'NV', 'Maine': 'ME'}

example['Code'] = example['State'].map(state_codes)
example = example.dropna()
example.head()
```

Out[34]:

	State	Number of Storms	Code
0	Alabama	2360	AL
1	Alaska	438	AK

FEMA data was analyzed to calculate the number of severe weather events by location

In [15]:

```
#Creating a data frame for the cities to analyze.
stateprices = {"State":["IL","MI","NY","PA","FL","GA","TX","AZ","CA","CA","CA","WA"], \
            "City":["Chicago","Detroit","New York","Philadelphia","Miami","Atlanta","Dallas","Phoenix","Los Angeles",\
                   "Long Beach", "Anaheim","Seattle"],
            "FIPS":['17031', '26163', '36063', '42103', '12086', '13013', '48113', '85003', '06037','06430','06020','536
            "Housing Price Rate":["231500","135000","388800","205100","264800","206200","252900","237500","610350", \
                                  "490000","521000","442000"]
                }
df_stateprices=pd.DataFrame(stateprices)
df_stateprices.head()
```

Out[15]:

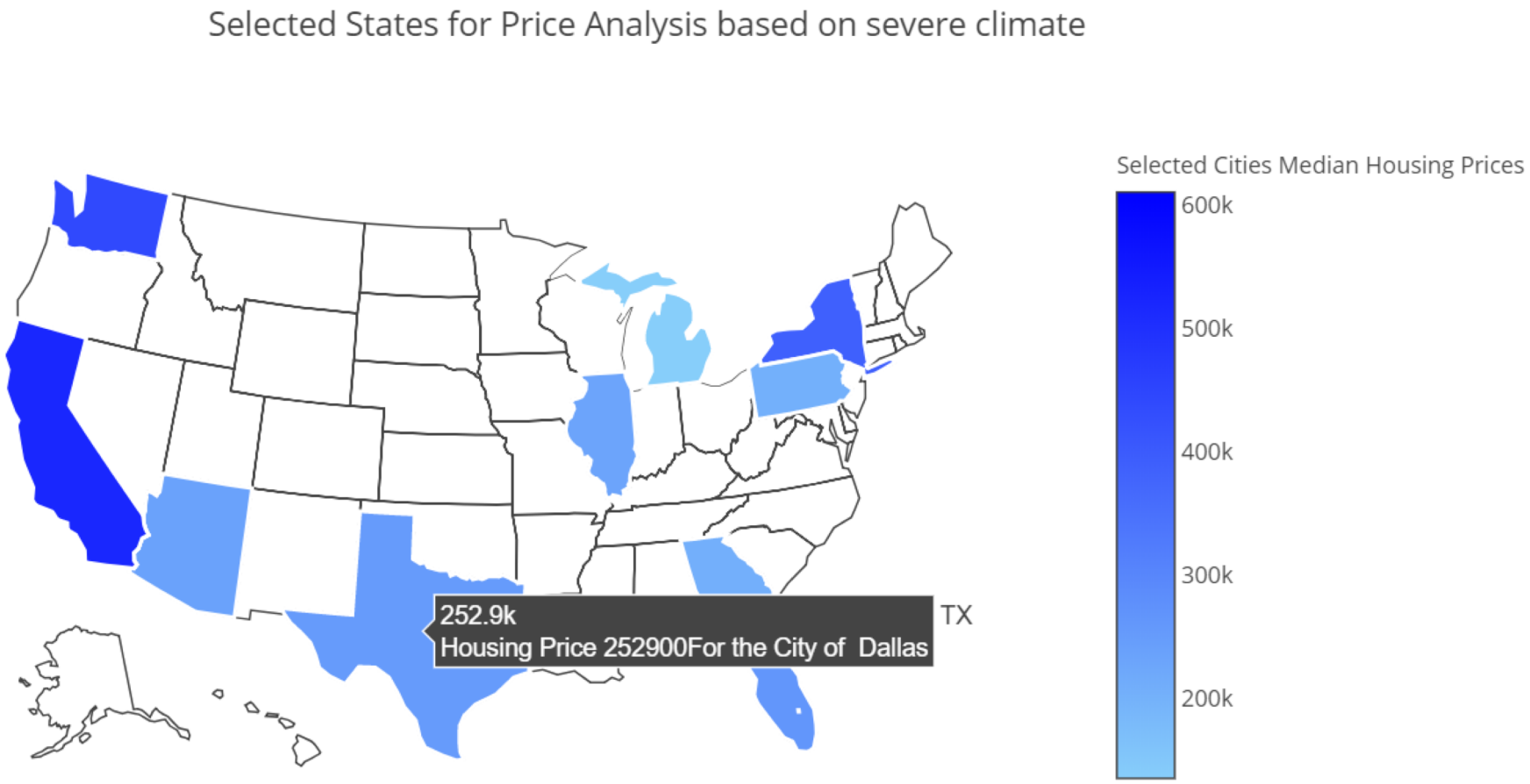
	State	City	FIPS	Housing Price Rate
0	IL	Chicago	17031	231500
1	MI	Detroit	26163	135000
2	NY	New York	36063	388800
3	PA	Philadelphia	42103	205100
4	FL	Miami	12086	264800

In [21]:

```
#Graph of the house prices selected for analysis
import plotly.plotly as py
import plotly.graph_objs as go
import plotly.figure_factory as ff

import pandas as pd
```

FEMA data was analyzed to calculate the number of severe weather events by location



Our approach was to utilize datasets from different sources to determine what factors impact housing prices

Rent or Buy?

Demographics

Severe Weather Events

House Layout

Layout data was analyzed for 2.9 million housing properties in 2016 for LA, Ventura, and Orange County

In [2]:
```
#read in dataset.  Dataset from Zillow.
file = 'data/properties_2016.csv'
df_raw = pd.read_csv(file)
```

In [40]:
```
df_raw.shape
```

Out[40]: (2985217, 58)

2.98M data points, 58 columns

In [3]:
```
#Clean data.  Data contains 2.9M data points
#remove properties with no value
df = df_raw.dropna(subset = ['structuretaxvaluedollarcnt'])

#remove listings with 0 bedrooms or bathrooms
df_box = df[['bedroomcnt', 'bathroomcnt']]
df_box1 = df_box[df_box['bedroomcnt']  > 0]
df_box2 = df_box1[df_box1['bathroomcnt']  > 0]
```

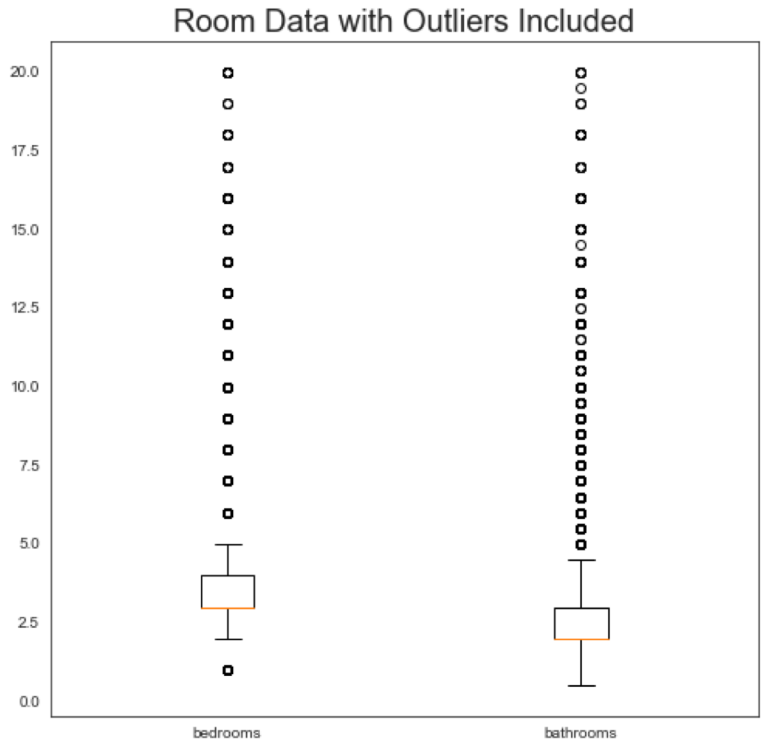
Removed datasets with 0 values in bed and bath

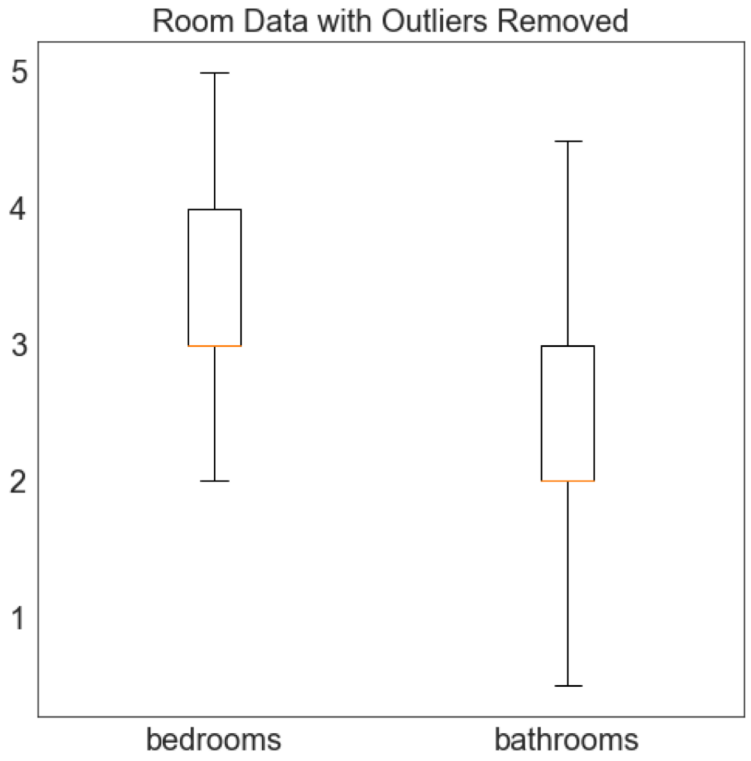
In [32]:
```
#create boxplot to determine the outliers

data = (df_box2['bedroomcnt'], df_box2['bathroomcnt'])
labels = ['bedrooms', 'bathrooms']
fig1, (ax1, ax6) = plt.subplots(nrows=1, ncols=2, figsize=(18, 8), sharey=False)
ax1.set_title('Room Data with Outliers Included', fontsize=20)
ax1.boxplot(data, showfliers=True, labels=labels)
plt.tick_params(axis='both', which='major', labelsize=20)
ax6.set_title('Room Data with Outliers Removed', fontsize=20)
ax6.boxplot(data, showfliers=False, labels=labels)
plt.tick_params(axis='both', which='major', labelsize=20)

plt.savefig('figures/BoxPlotofBedBath.png')
plt.show()
```

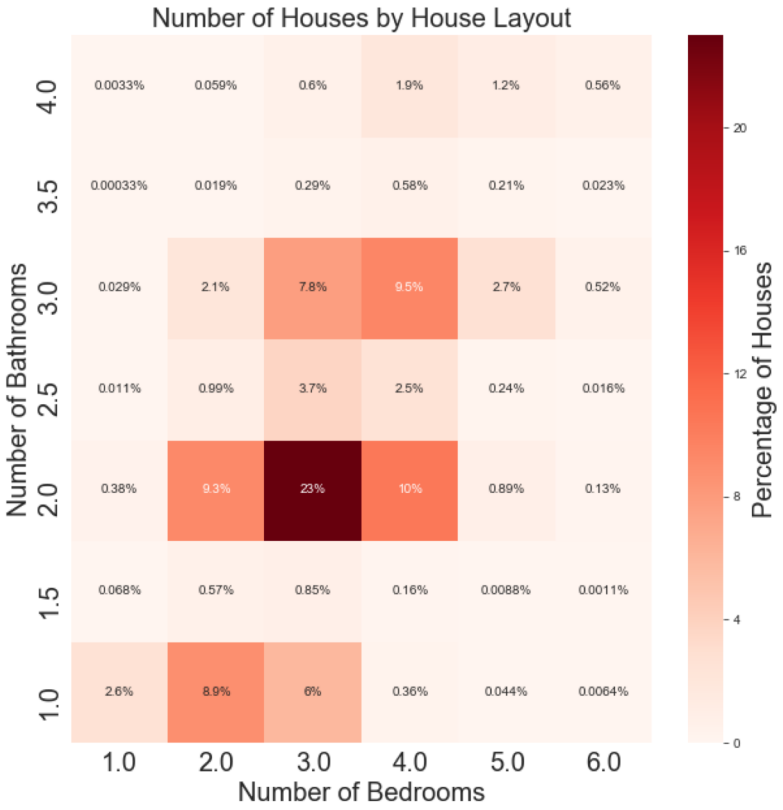
Layout data was analyzed for 2.9 million housing properties in 2016 for LA, Ventura, and Orange County

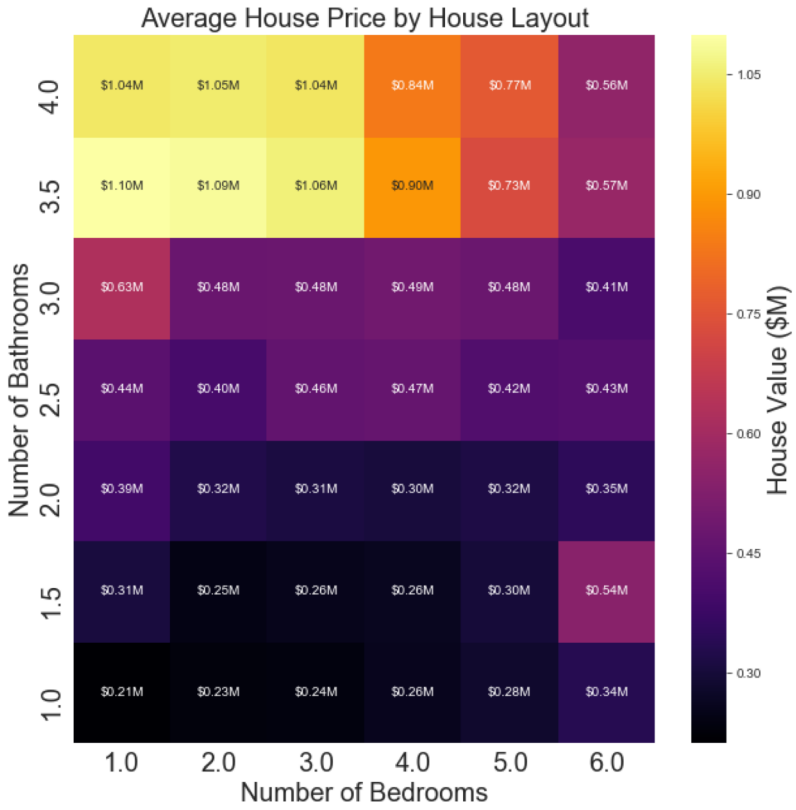




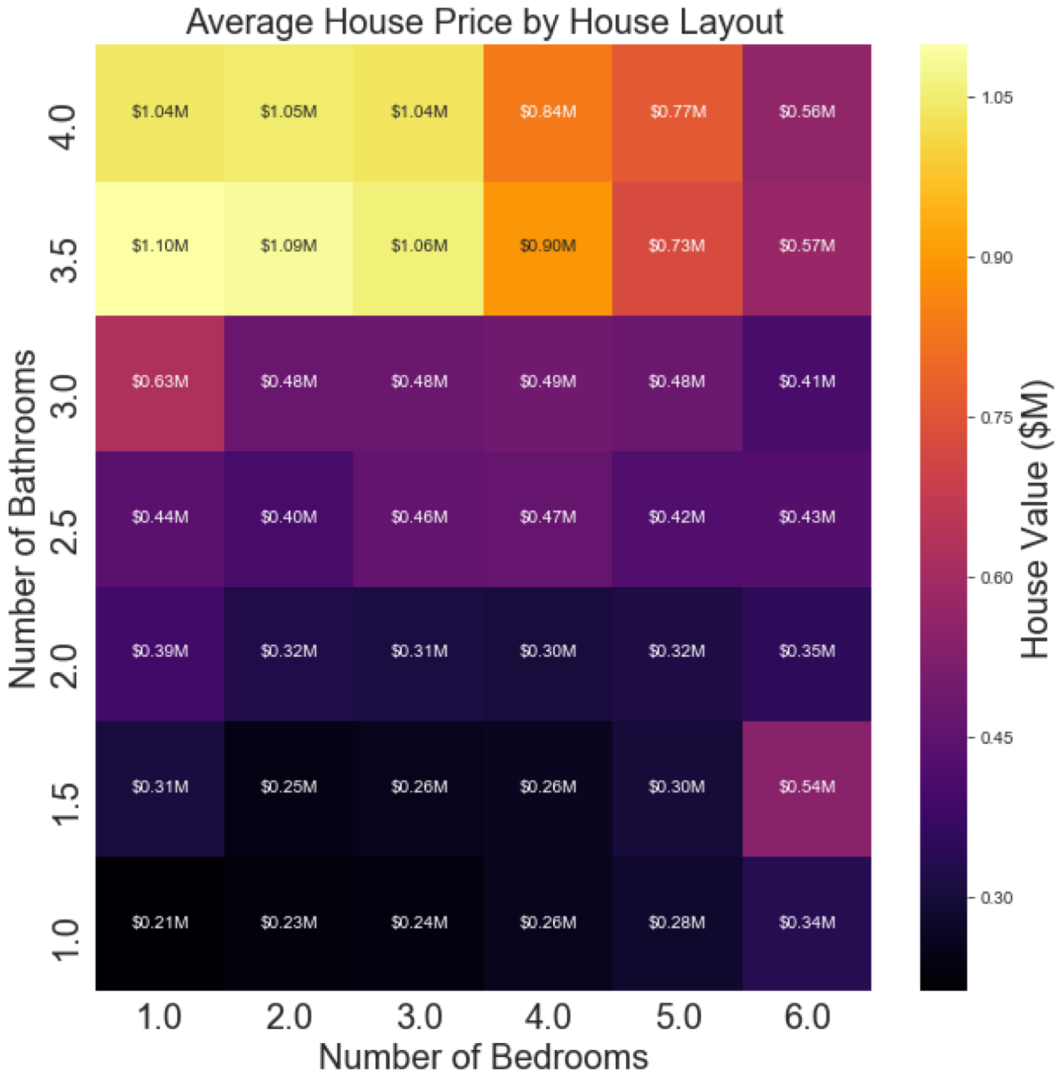
Data cleanup showed many outliers that needed to be filtered

Heatmaps were used to show the most common layouts, as well as the average price per layout

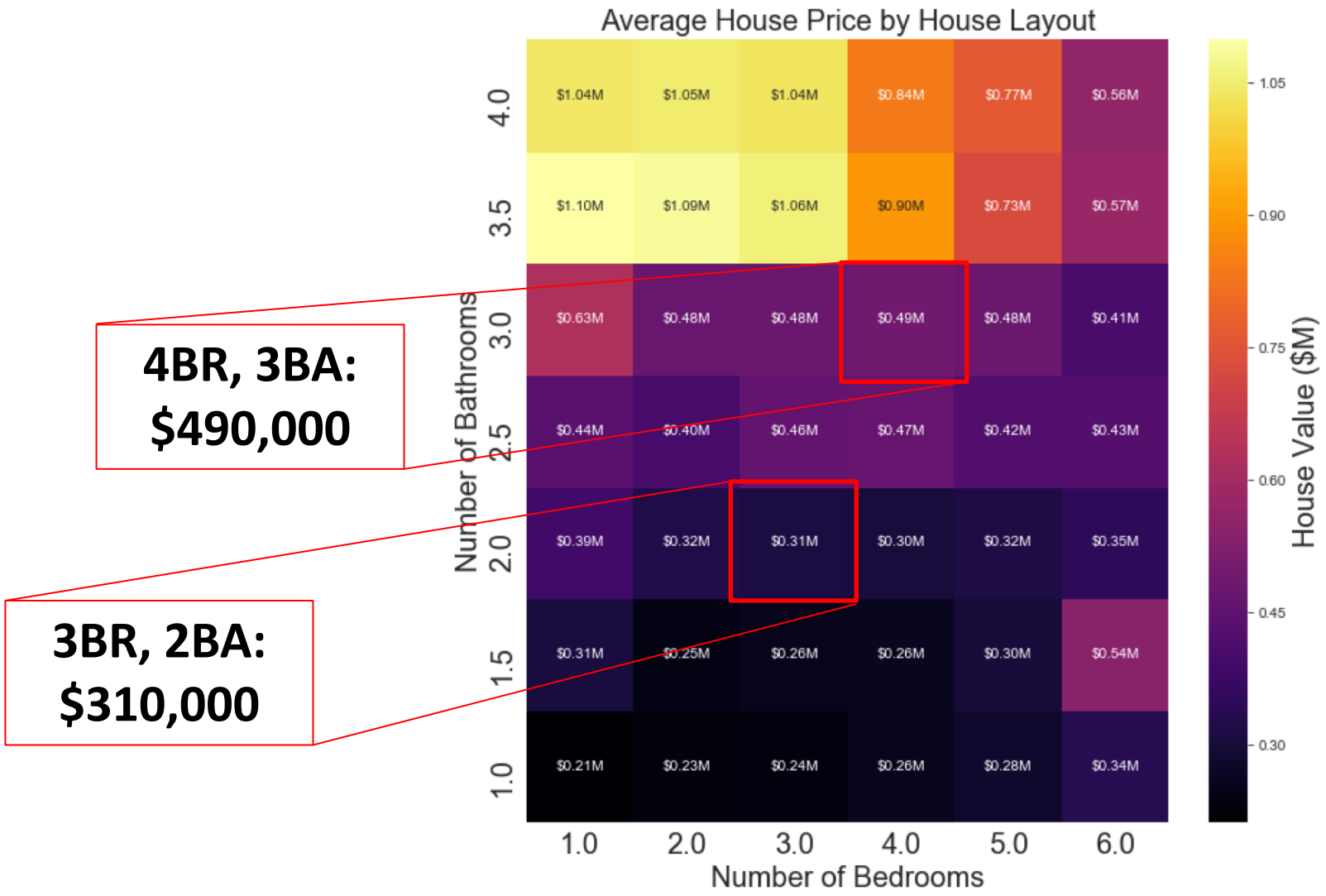




Heatmaps were used to show the most common layouts, as well as the average price per layout



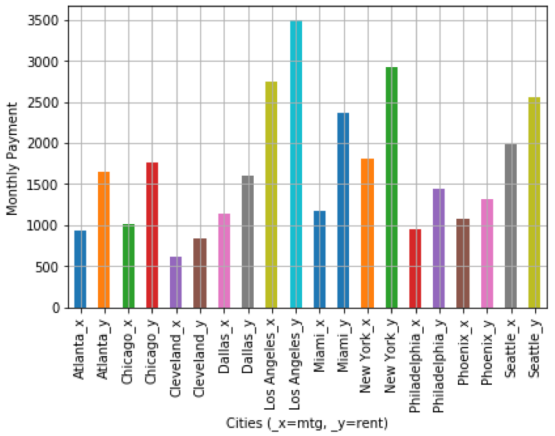
Heatmaps were used to show the most common layouts, as well as the average price per layout



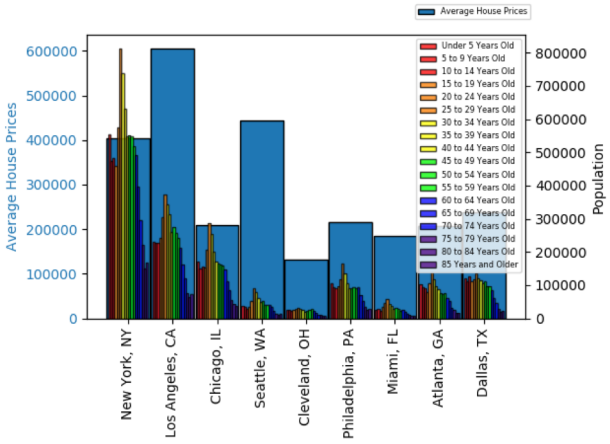
4BR, 3BA: $490,000

3BR, 2BA: $310,000

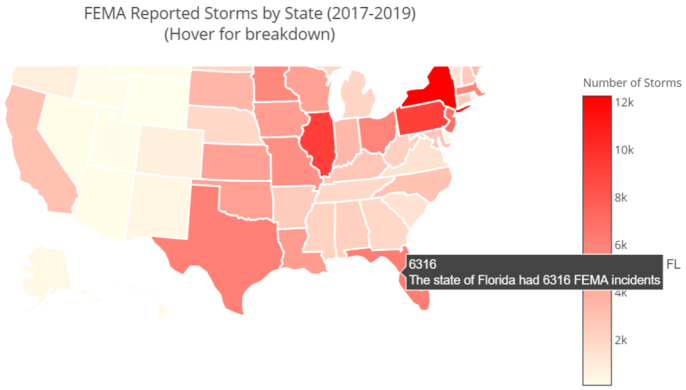
In summary, demographics, severe weather events, rental markets, and housing layouts can all impact housing costs



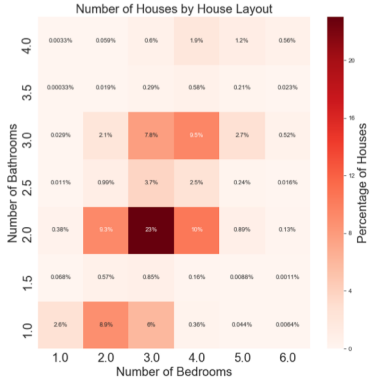
In most cities, it makes more financial sense to buy vs. rent

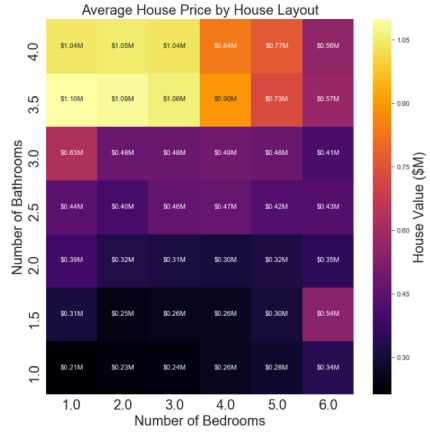


Racial make up did not have an impact on housing, but ages did



Severe weather events did not directly impact housing prices





Heatmaps can be used to show how much value a bedroom or half-bath adds to a house

If given more time, we would look at the following factors to further answer our questions

Go back and look at t-tests on entire list of cities, rather than subset

Demographics

Look at extreme temperatures; analyze potential development areas and where would be a good place to buy

Severe Weather

Look more into historical trends of rentals; get better data sets for mortgage rates

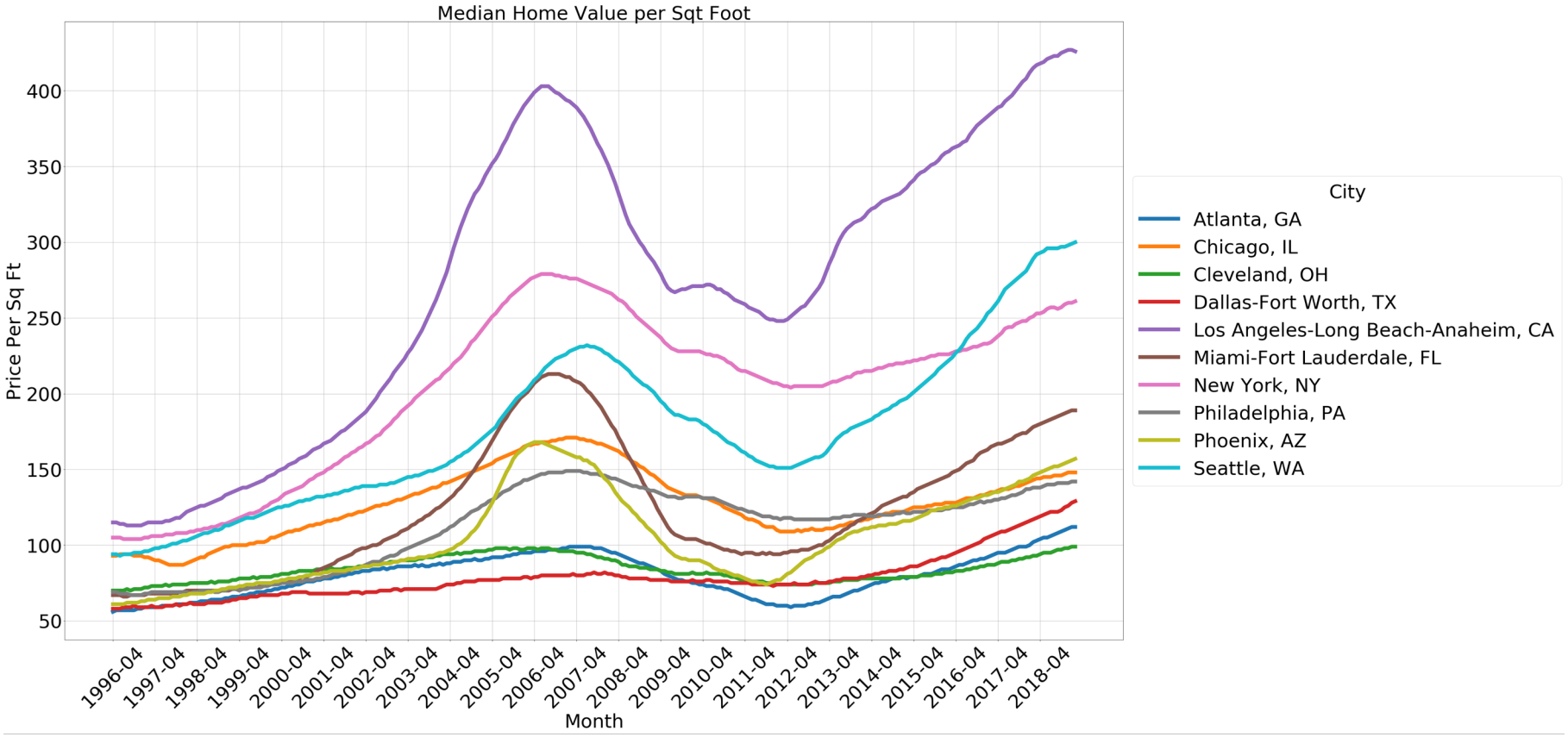
Rent vs. Buy?

Look at how square footage factors into layouts; examine other house amenities

House Layout

BACK UP SLIDES

The historical look at $/sqft shows the housing market crash and start of the recession in 2008



Biases for Rent or Buy

Msa	Rent Zestimate Accuracy	Homes on Zillow	Homes with Rent Zestimates	Within 5% of Rent Price	Within 10% of Rent Price	Within 20% of Rent Price	Median Error
Atlanta, GA	4	2,166,823	1,984,546	46.0%	69.2%	88.0%	5.6%
Chicago, IL	2	3,444,739	3,246,900	29.1%	51.4%	79.8%	9.5%
Cleveland, OH	2	852,407	785,229	37.6%	59.6%	82.6%	7.5%
Dallas-Fort Worth, TX	4	2,412,675	2,210,983	45.3%	69.0%	88.1%	5.7%
Miami-Fort Lauderdale, FL	2	2,309,476	2,247,312	35.9%	57.9%	80.0%	7.8%
New York, NY	2	5,988,420	5,647,225	33.9%	56.1%	80.9%	8.3%
Philadelphia, PA	2	2,230,251	2,119,442	34.7%	55.1%	78.1%	8.6%
Phoenix, AZ	3	1,749,655	1,596,643	41.2%	64.4%	84.5%	6.6%
Seattle, WA	3	1,375,305	1,282,147	39.4%	63.6%	86.0%	6.8%
Los Angeles-Long Beach-An	2	3,307,951	3,095,721	35.4%	57.4%	80.2%	8.0%

What is a Rent Zestimate?

A Rent Zestimate (pronounced ZEST-ti-met, rhymes with estimate) is Zillow's estimated monthly rent price, computed using a proprietary formula. It is a starting point in determining the monthly rental price for a specific property. The Rent Zestimate is pulled from public property data and similar local properties listed for rent; there may be special features, location, and market conditions our algorithms have not taken into account. Variations in rental price can also occur because of negotiating factors, special incentives, and length of lease.