Data Storytelling Submission

Questions Asked:

1. Are the “mega-mansions” skewing the results of my data?
2. Is there a relationship between the area of the lot and the value of the home?
3. Is there a relationship between bedroom count and value of the home?
4. Is there a relationship between the city in which the property is located and the value of the home?
5. Are 3 story houses more expensive than 1 story houses?
6. Do mostly more expensive homes have hot tubs/spas?
7. Is there a relationship between the amount of bathrooms the home has and the value of the home?

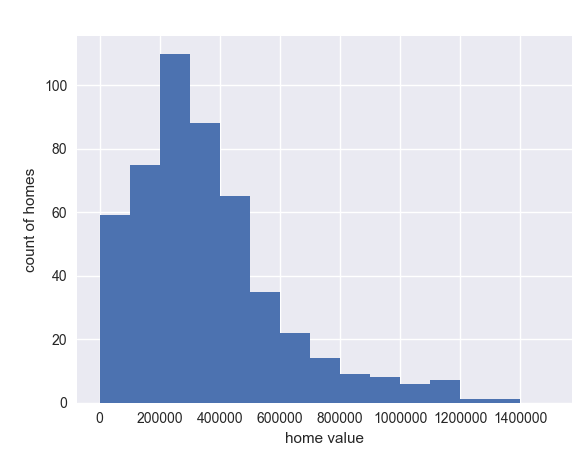
Trends Investigated/Conclusions Made:

1. Yes they are. The mean house value is $443.5k, yet I had values reaching as high as $90 million. I removed all outliers with an absolute zscore > 1.5. I saw similar situations with lot size, so I removed all outliers on that column as well.
2. It doesn’t seem like there is as strong of one as I would have imagined. Most homes are bunched on the smaller lot size and have a wide range of home values.
3. Most homes seem to have between 2-4 bedrooms. 3 bedrooms has the most amount of higher priced homes.
4. Can’t figure out how to make this plot. Trying to make a groupby with a median and plot that. I can get the code. Just can’t plot it:  
   df\_explore['taxvaluedollarcnt'].groupby(df\_explore['regionidcity']).median()
5. It doesn’t seem like it. While most houses in general are only 1 story, there are also more pricier homes that are 1 story than there are that are 3 storied.
6. Yes they do. Most homes that have a hot tub/spa are above the mean house price.
7. Yes, if the home has over 3.5 bathrooms, there is a very high chance the home is above the mean house price. Meanwhile, if the home only has 1 bathroom, odds are it is below the mean house price.

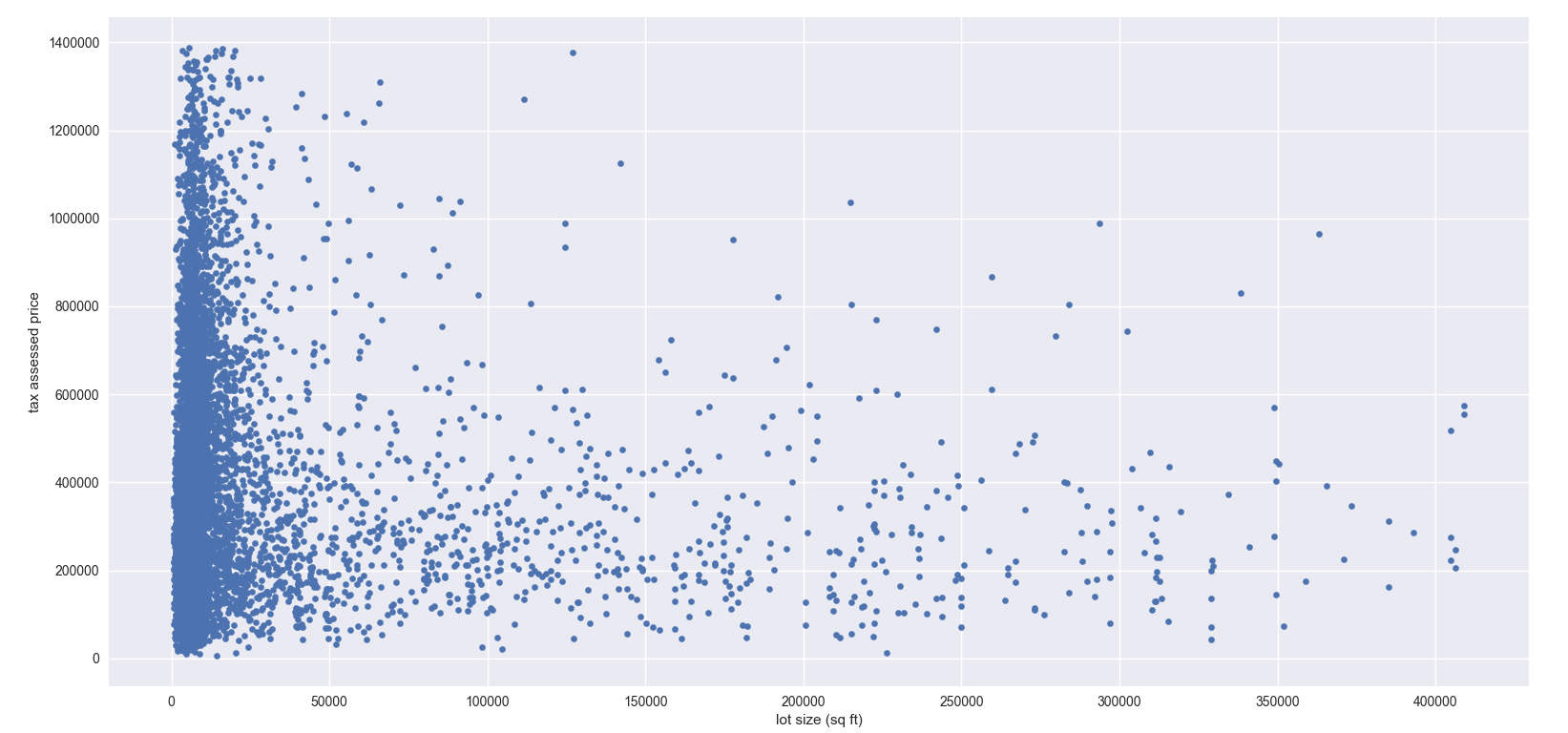
Resulting Visualizations:

1. (taking up too much memory to create/display)

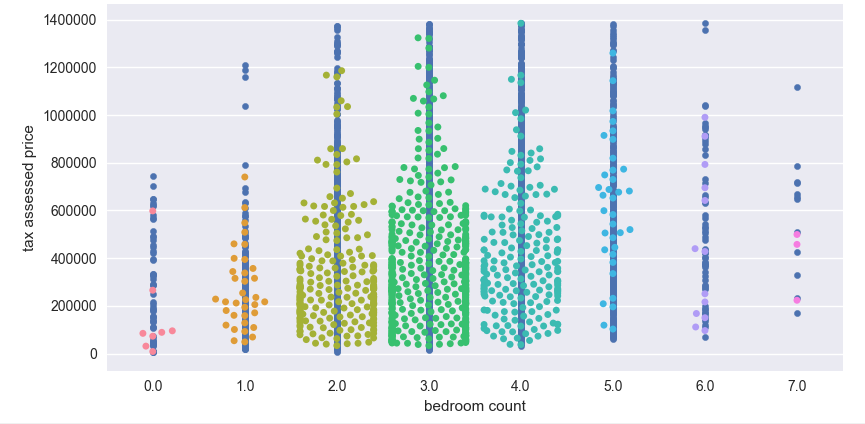
I created the below histogram showing the effect. However, even this histogram has already removed outliers and you can still see how heavily skewed towards the left it is.



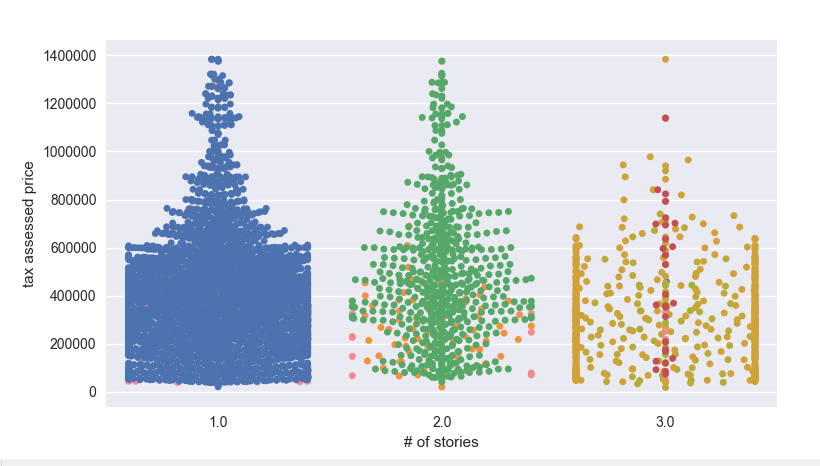
1. Relationship between area of lot and tax assessed value of home



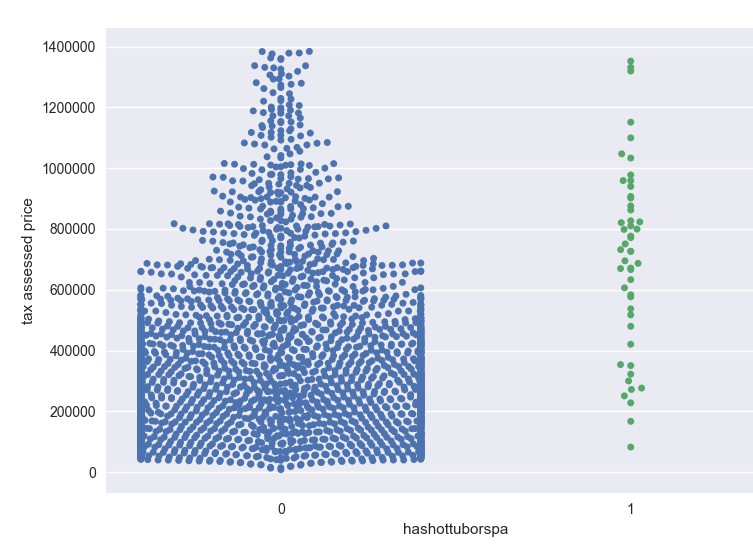
1. Relationship between bedroom count and value of the home



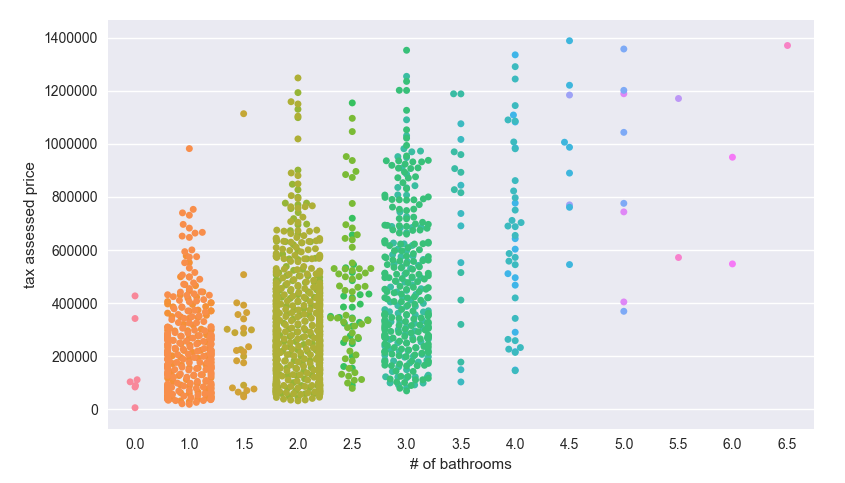
1. N/A
2. Number of Stories the house house and how it relates to home value



1. Relationship b/t hot tub/spa and home value



1. Relationship between number of bathrooms and home value



Plot Code:

1. df\_sample = df\_explore.sample(500)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

bin\_edges = [0,100000,200000,300000,400000,500000,600000,700000,800000,900000,1000000,1100000,1200000,1300000,1400000,1500000]

df\_sample['taxvaluedollarcnt'].plot('hist', bins=bin\_edges)

plt.ylabel('count of homes')

plt.xlabel('home value')

plt.show()

2.

df\_explore = df\_outlier\_removal2

df\_sample = df\_explore.sample(10000)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

var\_x = 'lotsizesquarefeet'

var\_y = 'taxvaluedollarcnt'

\_ = df\_sample.plot(kind='scatter', x=var\_x,y=var\_y)

\_ = plt.xlabel('lot size (sq ft)')

\_ = plt.ylabel('tax assessed price')

plt.show()

3. df\_explore = df\_outlier\_removal2

df\_sample = df\_explore.sample(1000)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

var\_x = 'bedroomcnt'

var\_y = 'taxvaluedollarcnt'

\_ = sns.swarmplot(x=var\_x,y=var\_y, data =df\_sample)

\_ = plt.xlabel('bedroom count')

\_ = plt.ylabel('tax assessed price')

plt.show()

4. df\_explore = df\_outlier\_removal2

df\_sample = df\_explore.sample(1000)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

var\_x = 'numberofstories'

var\_y = 'taxvaluedollarcnt'

\_ = sns.swarmplot(x=var\_x,y=var\_y, data =df\_sample)

\_ = plt.xlabel('# of stories')

\_ = plt.ylabel('tax assessed price')

plt.show()

5. df\_sample = df\_explore.sample(10000)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

var\_x = 'numberofstories'

var\_y = 'taxvaluedollarcnt'

\_ = sns.swarmplot(x=var\_x,y=var\_y, data =df\_sample)

\_ = plt.xlabel('# of stories')

\_ = plt.ylabel('tax assessed price')

plt.show()

6.

df\_sample = df\_explore.sample(2000)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

var\_x = 'hashottuborspa'

var\_y = 'taxvaluedollarcnt'

\_ = sns.swarmplot(x=var\_x,y=var\_y, data =df\_sample)

\_ = plt.xlabel('hashottuborspa')

\_ = plt.ylabel('tax assessed price')

plt.show()

7. df\_sample = df\_explore.sample(1000)

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

var\_x = 'bathroomcnt'

var\_y = 'taxvaluedollarcnt'

\_ = sns.swarmplot(x=var\_x,y=var\_y, data =df\_sample)

\_ = plt.xlabel('# of bathrooms')

\_ = plt.ylabel('tax assessed price')

plt.show()