DATA ANALYST INTERVIEW

City of Edmonton (Assessment & Taxation)

Questions

What are the top residential neighborhoods that retain their value over time?
What factors could affect market value?

AvielChow88@gmail.com

CONTENTS

TOP RESIDENTIAL NEIGHBORHOODS THAT RETAIN THEIR VALUE OVER TIME

Purpose

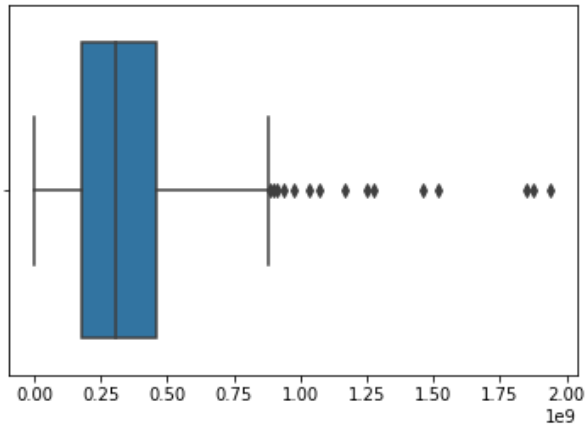
For this question, I assumed that the purpose of the question was **to forecast revenue** and to determine which neighborhoods generated the **steadiest stream of revenue**. Based on that assumption, I looked at the total market value of all properties within each neighborhood from 2012 – 2018 to determine which neighborhoods had the least year to year change. I used regression analysis to confirm my results.

Data

I used the Assessment Data from 2012 – 2018 on the City of Edmonton Open Data portal to conduct my analysis. Assessments are legislated to be at market value and I made the assumption that the assessments are correct. Market sales data was not available on the portal. I started by cleaning the data.

I started by investigating why certain neighborhoods had missing data for entire assessment years. I made **data corrections**. For example, neighborhoods such as “Terwilligar South” were renamed “South Terwilligar” in 2012 and the data needed to be corrected. Then I checked the average of the total assessed values of each neighborhood to check for outliers. The Interquartile range was $180,000,000 to $460,000,000; however, since I assumed the purpose of the study was forecasting, I didn’t want to eliminate outliers, especially outliers in the upper quartile.

Box Plot of Average of Total Assessed Values per Neighborhood From 2012-2018



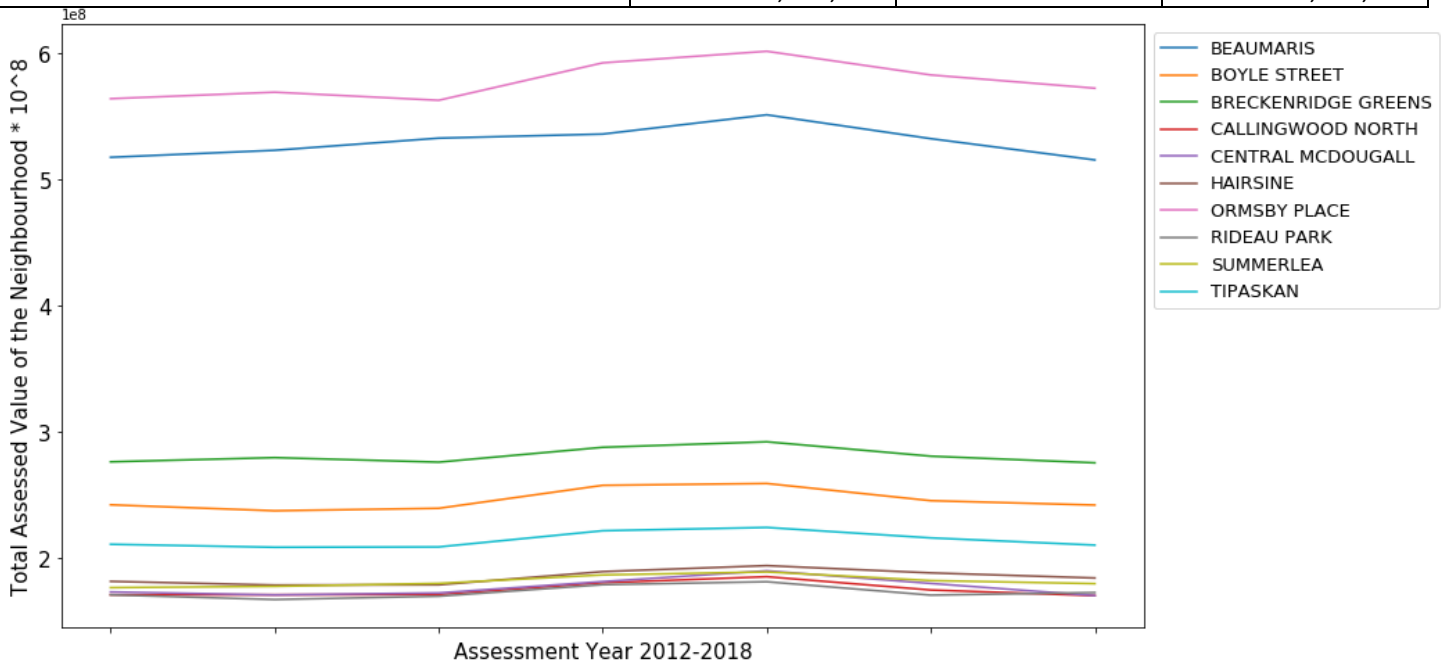
From the data, it was clear that the **high value neighborhoods tended to have higher percent change**. In order to ensure data on high revenue generating, upper quartile neighborhoods was not lost, I segmented my research and conducted one study for neighborhoods between $150,000,000 to $600,000,000 in total market value. I conducted another study for all neighborhoods with over $600,000,000 in total market value.

RESULTS

In the $150,000,000 to $600,000,000 range, the Top 10 Neighborhoods with low % year to year change were **Breckenridge Greens, Callingwood North, Rideau Park, Central McDougall, Summerlea, Tipaskan, Beaumaris, Boyle Street, Hairsine, and Ormsby Place**. The Sum of the Year to Year % Change and the regression slope for the Total Assessed Value of the Neighborhood from 2012-2018 are in the chart below.

Top Ten Neighborhoods That Retain Value Over Time ($150,000,000 - $600,000,000)

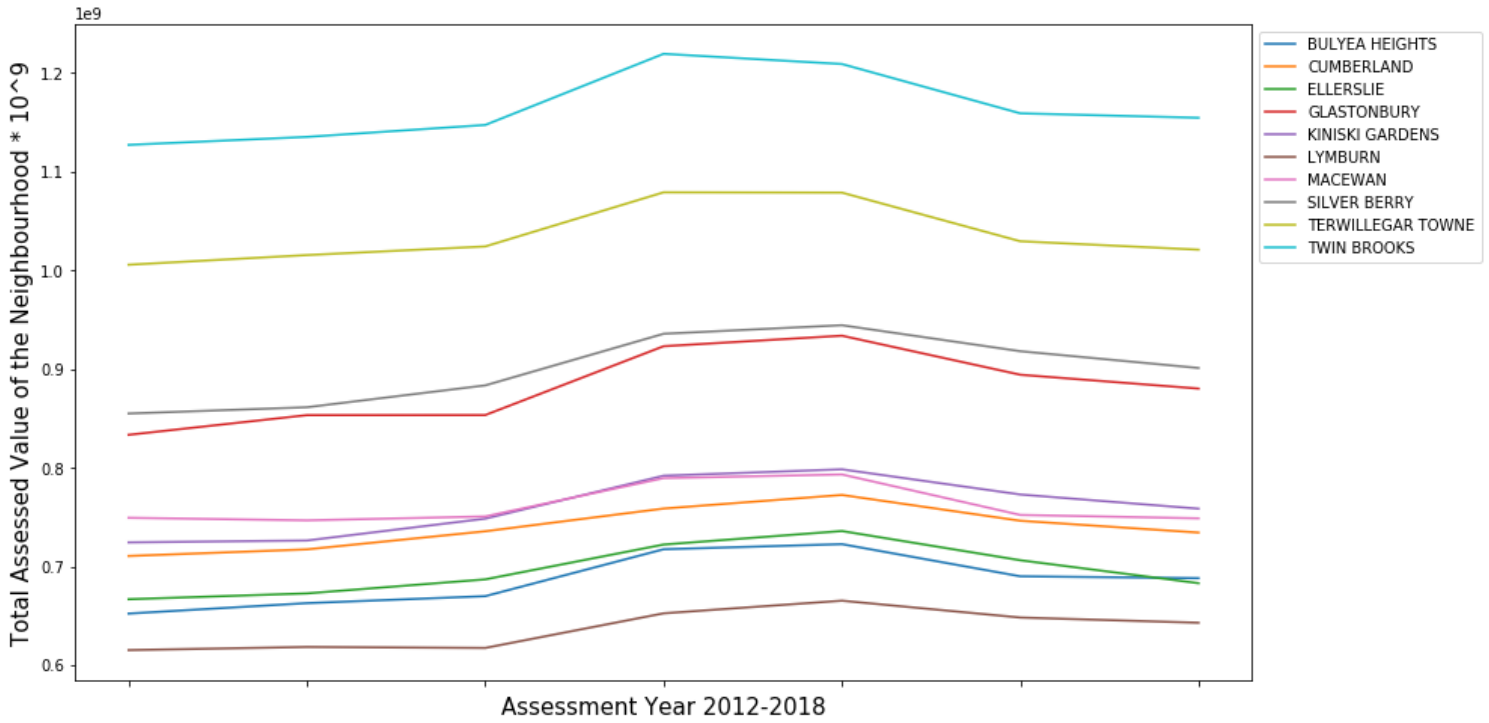
	Average Total Assessed Value	Sum of Year to Year % Change	Regression Slope
BRECKENRIDGE GREENS	281,467,357	-0.08%	572,375
CALLINGWOOD NORTH	175,186,429	-0.01%	702,821
RIDEAU PARK	173,437,143	1.43%	854,464
CENTRAL MCDOUGALL	177,222,214	-1.13%	966,446
SUMMERLEA	182,064,929	1.92%	987,321
TIPASKAN	214,731,643	-0.01%	1,008,768
BEAUMARIS	529,940,786	-0.23%	1,094,554
BOYLE STREET	246,541,786	0.37%	1,251,500
HAIRSINE	185,354,857	1.74%	1,516,571
ORMSBY PLACE	577,990,286	1.68%	3,251,750



In the $600,000,000 and up range, the Top 10 Neighborhoods with low % year to year change were **Macewan, Terwillegar Towne, Ellerslie, Cumberland, Lymburn, Twin Brooks, Bulyea Heights, Kiniski Gardens, Glastonbury, Silver Berry**. The Sum of the Year to Year % Change and the regression slope for the Total Assessed Value of the Neighborhood from 2012-2018 are in the chart below. The regression slope and % change in many of these neighborhoods is fairly high and **many would not be in the Top 10 if I did not segment the data**.

Top Ten Neighborhoods That Retain Value Over Time ($600,000,000+)

	Average Total Assessed Value	Sum of Year to Year % Change	Regression Slope
MACEWAN	761,691,929	0.20%	1,855,750
TERWILLEGAR TOWNE	1,036,289,143	1.76%	4,581,179
ELLERSLIE	696,521,214	2.71%	5,886,929
CUMBERLAND	739,561,000	3.45%	5,925,089
LYMBURN	637,459,000	4.62%	6,806,054
TWIN BROOKS	1,164,520,857	2.69%	6,849,304
BULYEA HEIGHTS	686,435,643	5.71%	7,663,964
KINISKI GARDENS	760,354,714	4.88%	8,771,000
GLASTONBURY	881,787,929	5.92%	10,810,143
SILVER BERRY	899,998,143	5.52%	11,169,911



IMPROVEMENTS

One improvement would be to utilize more data from previous years. It would also be valuable to be able to validate whether the assessment values are representative of market value using market sales. Less obviously, it would be valuable to **measure change in each individual property within our top neighborhoods to ensure that the steadiness in market value is not due to a temporary offset**. For example, a neighborhood where individual property values are steadily decreasing but new properties are being built and sold, may appear on our Top 10 list.

Focus

This is a big question and I decided to limit the scope partially because of the availability of data and partially due to complexity. After exploring the data, I limited the scope by **excluding Condos and Apartments**. The **Assessment Data online does not include livable square footage.** Condos don't have accurate lot sizes, and without livable square footage, a regression analysis of condos would be quite tough. It would be difficult to determine whether the impact of something like Neighborhood was due to size.

The raw data did not give data on the Building type, so I excluded all data with "Unit" in the legal description to filter out condos. I eliminated remaining outliers from the remaining dataset. (For the most part, Condo's and Apartment's Legal Description includes Unit, while other residential properties will use Block and Lot). With the remaining data, I only focused on **lot size, neighborhood, and actual year built.**

Data Cleaning

I used the 2018 Assessment Data available on the City of Edmonton open data portal. It included lot size, neighborhood, legal description, actual year built, garage, and zoning. I did not analyze garage and zoning to reduce complexity and to have a deeper analysis on lot size, neighborhood, and actual year built. As stated earlier, I **filtered out properties with legal description unit to eliminate condos**. I recognize this could skew the data by eliminating a specific subclass of properties that are not condo's or apartments but have unit in their legal description.

Next, I **eliminated properties with assessed values under $50,000 and over $1,500,000**. From listening to Assessment Review Board hearings, properties under $50,000 are usually parking spots or properties with such significant deficiencies that are not captured in the available data, that they would skew the results. Similarly, properties over $1,500,000 have such unusual special features that they would also skew the results.

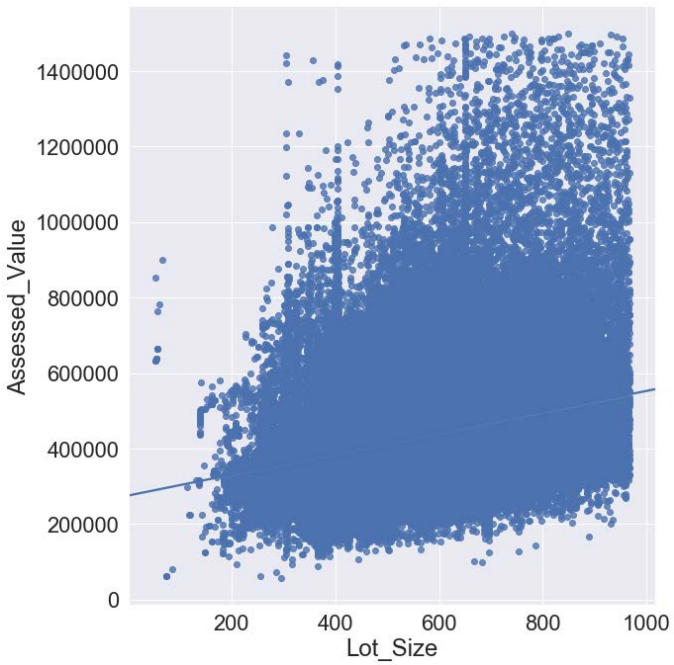
With the remaining data**, I eliminated all statistical outliers (interquartile range * 1.5) based on lot size, and removed all properties with an Actual Year Built of under 1600**. I used the remaining data for my initial data exploration

LOT SIZE, NEIGHBORHOOD, ACTUAL YEAR BUILT - EXPLORATION

I started by performing some simple linear regressions with the cleaned data to get some sense of the information. I tested **Lot Size vs. Assessed Value, Lot Size vs Price per meter squared of Lot Size, Year Built vs. Assessed Value using regression**. In all three cases, the p-value against the null hypothesis of a slope of 0 was below 0.001. The R^2 values ranged from 0.15 to 0.25. This tells us that we can be quite certain that our **predictors have an effect on assessed value but that none of the models would be good at making predictions** because there is too much variance unaccounted for.

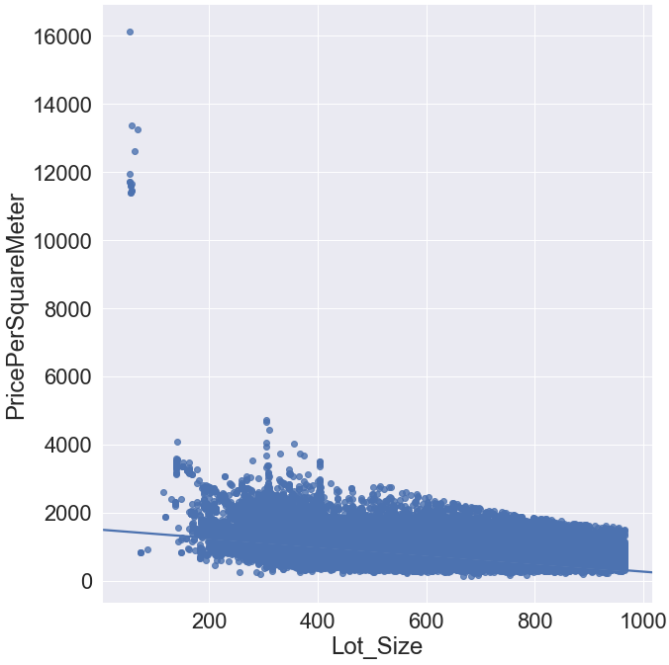
Lot Size

Lot Size in Square Meters vs Assessed Value



Slope = 153, R Value = 0.29, P Value < 0.0001

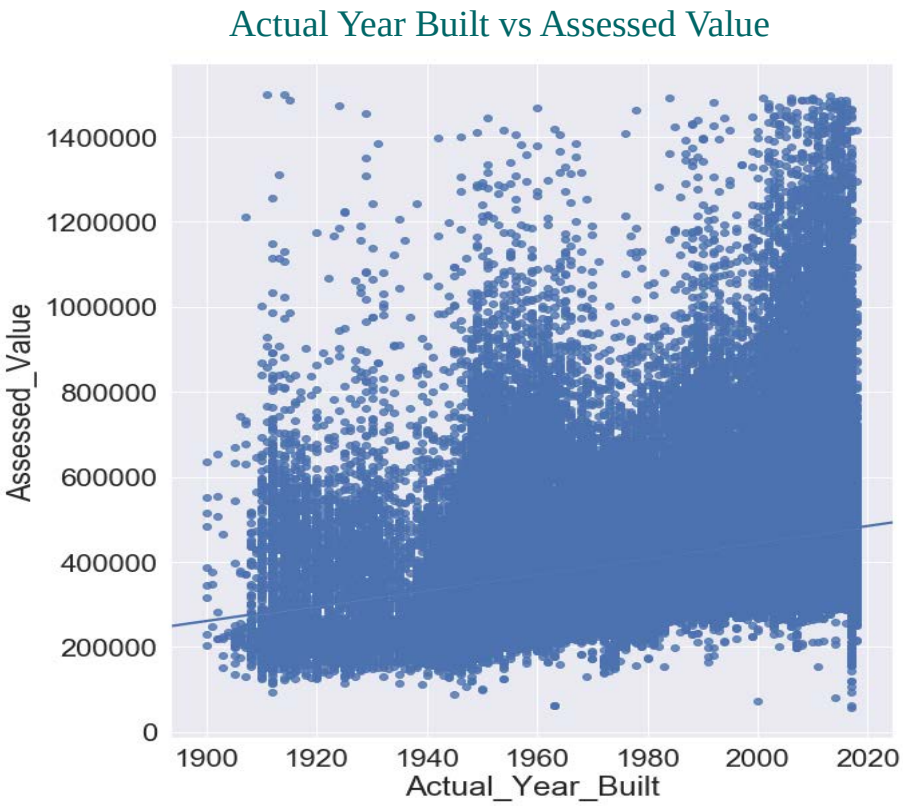
Lot Size vs Price per Square Meter



Slope = -1.23, R Value = -0.56, P Value < 0.0001

From both the P values and the graphs, we can tell that Lot Size has an effect on Assessed value. We can also see that Lot Size does not explain a lot of the variance and lot size alone would not give us much predictability. The second chart demonstrates **"economies of scale".** Often, in hearings, assessors will explain that price per square foot decreases as lot size increases. This shows that the effect of lot size on market value decreases as lot size increases.

Actual Year Built

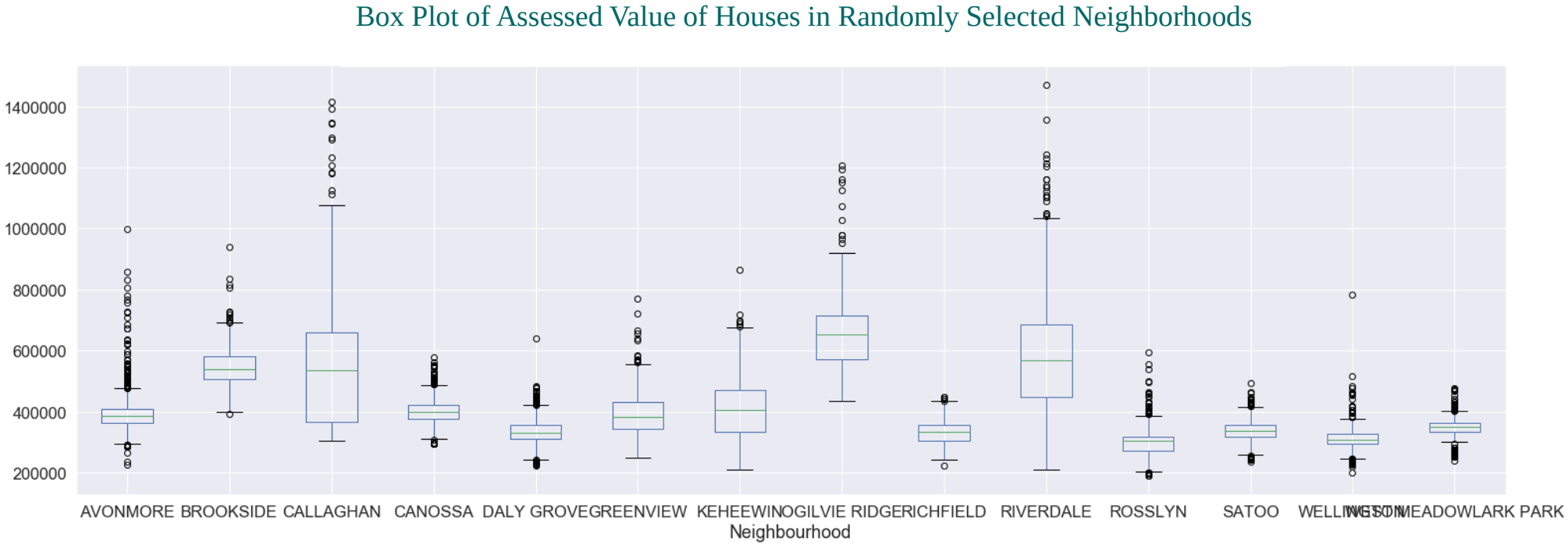


Slope = 2122, R Value = 0.6, P value <0.0001

Actual Year Built vs. Assessed Value has a higher R value than Lot Size indicating there is a greater correlation between Year Built and Assessed Value than Lot Size and Assessed Value. In either case, the R^2 value is quite low so, as we would expect, there is a lot of variance unaccounted for.

To ensure that the Actual Year Built and Lot Size are asserting different impact on Assessed Value I performed a linear regression for Lot Size v. Actual Year Built. As expected, there is a correlation (R value = -0.38, P value < 0.001) where Lot Size decreases as Year Built Increases. Since this relationship is the inverse of the relationship between Lot Size and Assessed Value, and Actual Year Built and Assessed Value, we can assume that **the two assert different impacts on Assessed Value.**

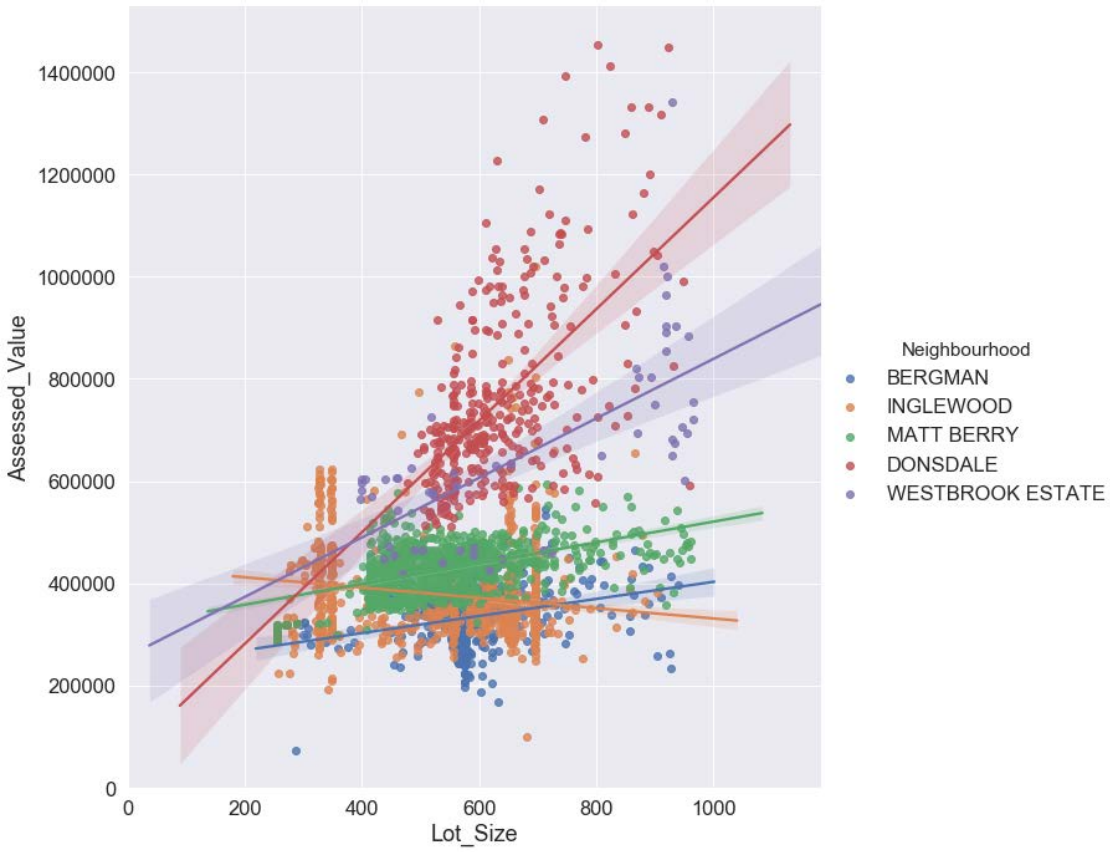
Neighborhood



When considering how to view Neighborhoods, I considered creating dummy variables for the Neighborhoods and then performing regression; however, since I was working entirely in Python, I was able to create dummy variables but managing the regression on the data set required more time. Instead, **I used a random sampling and box plot visualizations to demonstrate the relationship between Neighborhood and Assessed Value**. At a quick glance, we can clearly see that Neighborhood has an impact on assessed value.

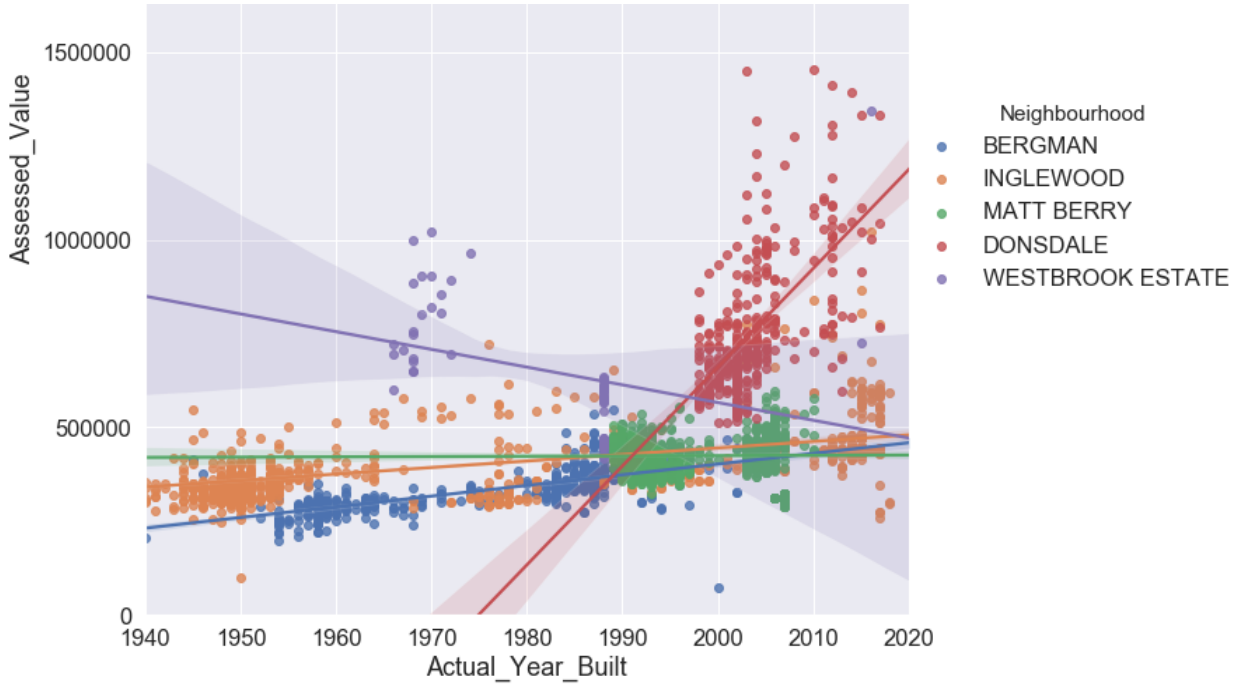
Clearly, there is a relationship between Neighborhood and Assessed Value; however, **Neighborhood must also have correlations with a lot of the variables** such as lot size and actual year built. In order to investigate further, I turned Neighborhood into a constant value, and tried to determine whether the regression analysis between Lot Size and Assessed Value, and actual year built and assessed value would improve.

LOT SIZE, ACTUAL YEAR BUILT – NEIGHBORHOOD CONSTANT



I performed regression analysis with various constant Neighborhoods but for visual ease, I only displayed 5 Neighborhoods. These 5 neighborhoods are randomly generated by the computer from a list of Neighborhoods with a total assessed value of $150,000,000 to $600,000,000. Even without the statistics, we can see that the model is significantly better after controlling Neighborhood. We can see that Lot Size generally increases as Assessed Value increases. Controlling the Neighborhood also gives us the opportunity to dive in and investigate anomalies such as Inglewood where assessed value decreases as lot size increases.

Actual Year Built vs Assessed Value by Neighborhood



The same can be said for the graph above. The regression is clearly better, and it gives us the opportunity to investigate further into anomalies such as Westbrook Estate. It also appears that Actual Year Built has a steady but limited effect on value in older neighborhoods such as Bergman and Inglewood. We could dive deeper into such hypotheses.

CONCLUSION

Actual Year Built, Lot Size, and Neighborhood clearly affect value. In all cases of regression, the P value is far below 0.0001. That said, no factor alone is a strong predictors of Assessed Value as no factor can explain a significant amount of variance. Based on a random analysis, Neighborhood also clearly affects value. The Box Plot shows that assessed value changes as neighborhood changes.

When we keep the neighborhood the same, the regression analysis for Actual Year Built and Lot Size improve significantly. Given more time, it could be valuable to conduct a multiple linear regression analysis on various factors in each neighborhood.

Other factors to consider would have been livable square footage, property type (duplex, single family detached, condo, etc.), quality, effective year built, various influences (lake abutting, major roadways etc.), and garage. If I was able to obtain square footage, investigating condos and apartments would also have been more meaningful.

In conclusion, many factors affect assessed value, and under the assumption that assessed value is close to market value, this report shows that Actual Year Built, Lot Size, and Neighborhood certainly have an effect on market value.

TECHNICAL DOCUMENTATION

All of the analysis was done exclusively using Python. Most of the data cleaning and manipulation was done using Python Libraries Numpy and Pandas. The Data Visualization was done using Seaborn, and the statistics was done using Scipy. I used Anaconda

The file “CleanData.py” was used to clean and reduce the csv data from 255mb to about 10mb. It created various files that were used to conduct the analysis on the neighborhoods with the least change in total assessed value. The file “StableMarketValue.py” was used to analyze and visualize the data.

The file “CleanData1.py” was used to clean and reduce the csv data for the second problem. Analysis.py was used for the filtering based on distribution, regression analysis, and visualization.

The files are included for your review.