Philadelphia Housing Analysis

Project 1
May 3, 2021

Group 2:
Jeremy Bar
Anthony Carannante
Dominique DeMoe
Anjanette Velazco
DeAngelo Williams

Task
Create a write-up summarizing your major findings. This should include a heading for each "question" you asked of your data, and under each heading, a short description of what you found and any relevant plots.

Background
Philadelphia is a city that is rich in history and diversity. With a population over 1.5 million and neighborhoods undergoing constant changes, our group wanted to dive into the housing market. Through several different explorations, our group endeavored to find the answer to following questions:

1. How have home sales in Philadelphia evolved over the past 5 years?
2. How does location (zip code) affect pricing and number of sales?
3. How has age of home affected pricing and number of sales?
4. How does size of home affect pricing and number of sales?

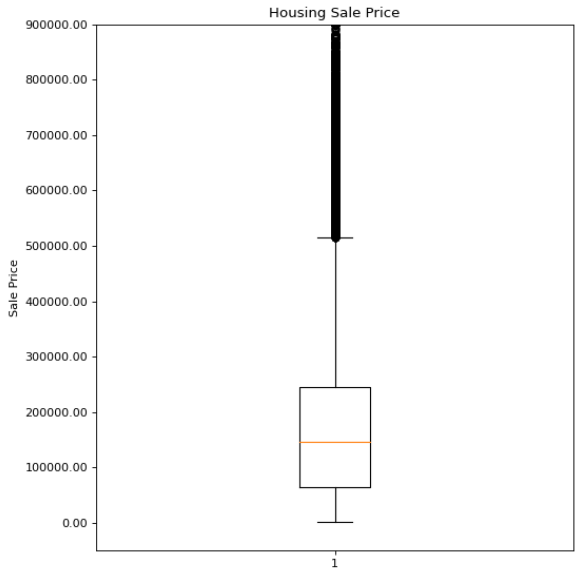
How we cleaned our data:
Before we began working, we had to clean our data.
1. Remove rows out of scope
2. Remove columns out of scope
3. Re-format columns

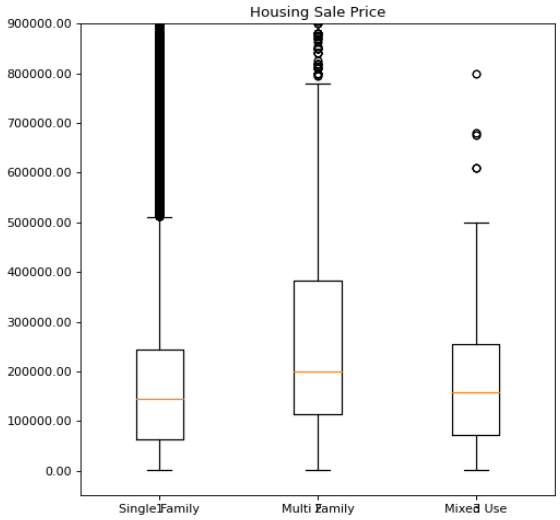
Summary Analysis
Question: How have home sales in Philadelphia evolved over the past 5 years?
Description: The total number of sales gradually increased from 2016-2018 and decreased from 2018-2020. The average house sale price generally increased except for 2017-2018 where it dropped about $13,000. Median sale price generally increased year-over-year. Standard deviation varied over the years which is likely due to outlier sales where houses sold for extremely large numbers.

Year	Total Number of Sales	Mean Sale Price	Median Sale Price	Standard Deviation of Sale Price
2016	14,060	$177,132	$125,000	$240,260
2017	15,619	$219,403	$140,000	$408,161
2018	16,283	$206,449	$155,000	$272,475
2019	15,787	$221,614	$153,000	$433,686
2020	6,736	$257,225	$170,000	$429,846

Box Plot Analysis
Question: How do outliers affect the data?
Description:
After looking at the summary table, we noticed that the mean housing price is larger than the median housing price anywhere from $50k-$100k. An outlier analysis was required to see why this was.

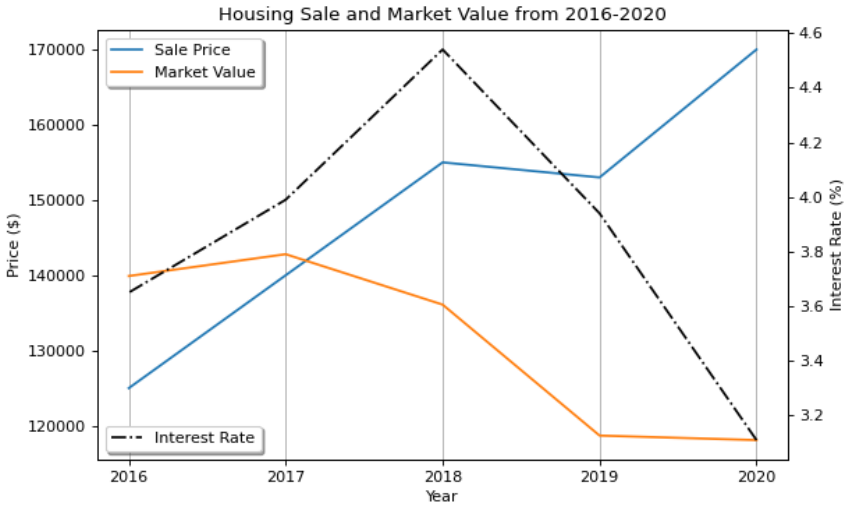
- Prices of all housing types had a median of about ~$150k with a significant number of outliers above the upper quartile. We then split this into types of houses (single family, multi family, mixed use)
- Single family house prices had the majority of the outliers. Multi-family houses had a wider range of values within the center quartile but still had some outliers that affected the mean price.
- Mixed use houses had few outliers, could be indicative of a small amount of data for these types of houses.





5 Year Trend Analysis
Question: How has the market value and sale price of homes changed over the last five years?
Description:

- We wanted to compare how the market value and sale prices changed over the last five years solely from a numbers standpoint.
    - It appears that the median value of sale prices continued to increase over this five-year period, while the median market value proceeded to decline over that time period.
- The annual interest rate for a 30-year mortgage increased and peaked in 2018 and then decreased until 2020.
    - From 2016-2018 the sale price and interest rates increased together, while from 2018-2020 the interest rate decreased as the sale price still increased.
    - The opposite holds true for the interest rate and market values.
    - In summary, sale prices increased while market values decreased, while interest rates appear have little direct effect on sale prices and market values.
- Similar to how stocks follow social and business trends, the interest rates appear to follow the housing prices. We decided to look at social factors that could affect housing prices.

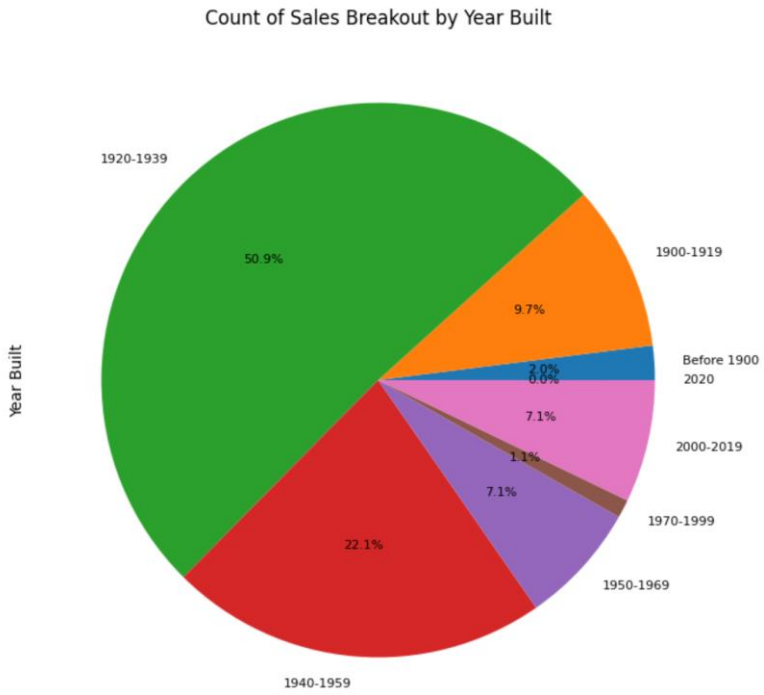


Pie Chart Analysis

Question: How has age of home affected pricing and number of sales?

Description:

Another one of the factors that our team looked at to determine the key drivers of home sales in Philadelphia was the year built. After breaking out the year built into 20-year bins, our time created a pie chart to compare the breakout of home sales between them.



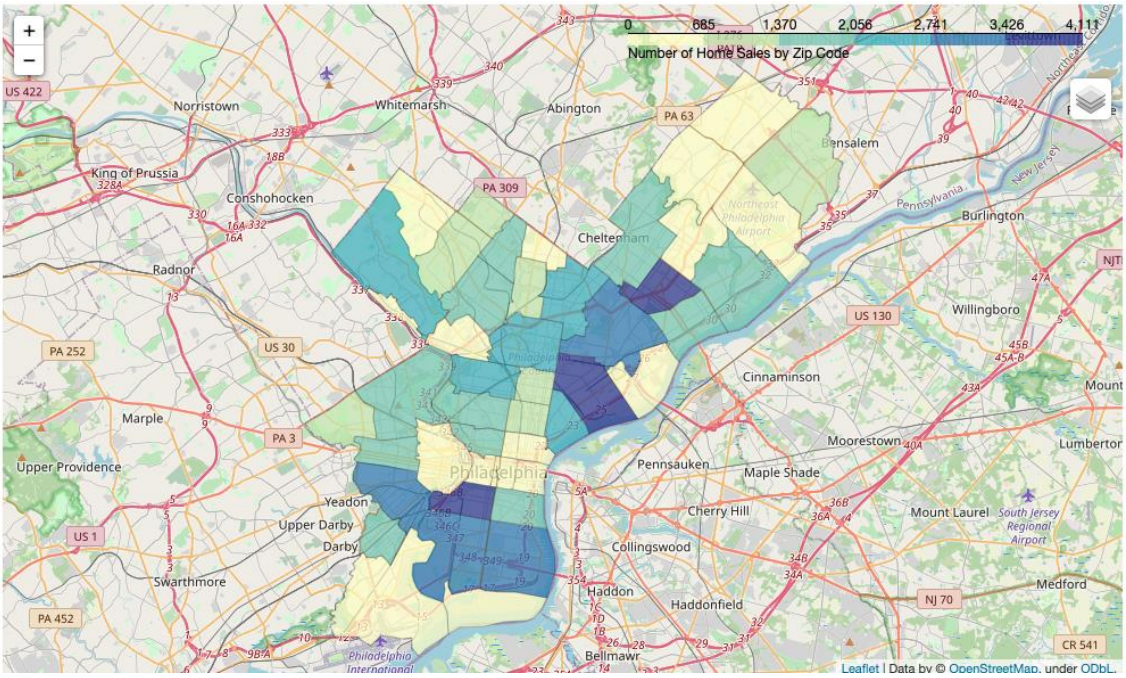
The vast majority of home sales were for homes that were built between 1920 and 1939, as can be seen in the pie chart above. About half of all home sales between 2016 and 2020 were for homes that were 80 – 100 years old. This could indicate that there has been a higher demand for these types of homes as opposed to homes built during different time periods. If there is higher demand, the most basic economic principles tell us that prices should be higher as well. When looking at this chart in a vacuum that would be a reasonable conclusion, but it is clear from other analyses in this dataset that there are many factors at play in determining house prices. Likely an interaction amongst several of the characteristics analyzed as part of this dataset would allow for a more accurate prediction of what house prices should be.

Zip Code Analysis

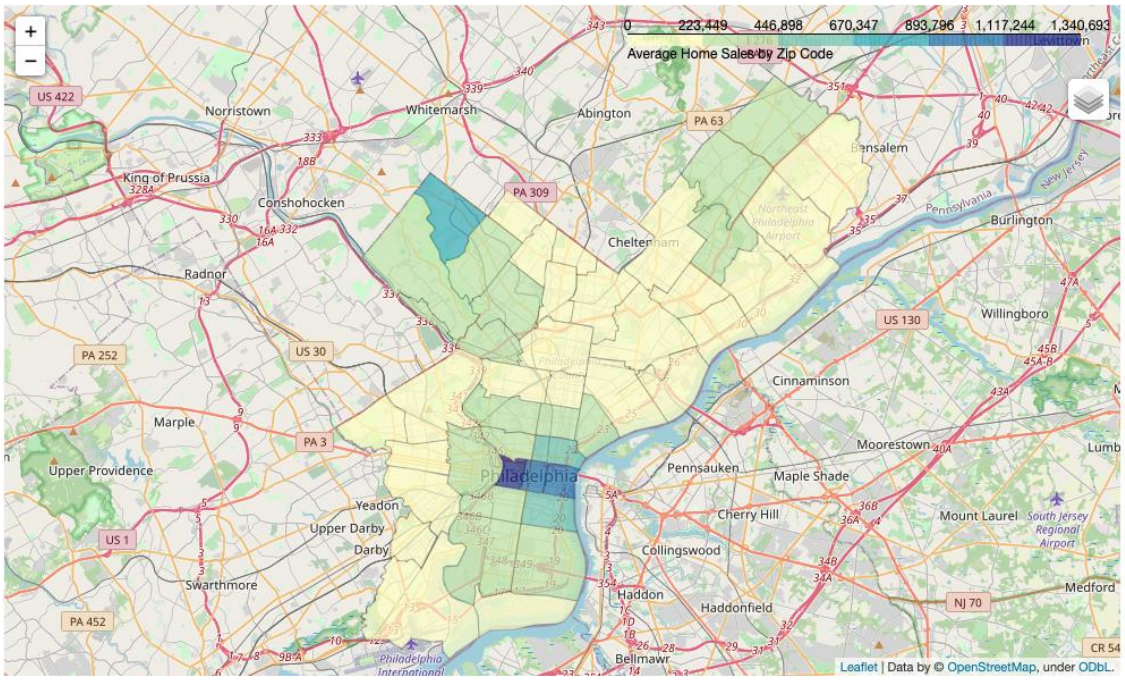
Question: How does location (zip code) affect pricing and number of sales?

Description:

We wanted to explore how location (defined by zip code) affects pricing and number of sales. Initially, we looked at the count of sales by zip code. This explores the total number of home sales by zip code over time where most homes are sold. With this analysis we found that areas with the top number of sales are the zip codes 19149 (Oxford Circle), 19134 (Port Richmond) and 19146 (Center City West). From this we can see that the greater Center City area generally has a high number of sales, but up and coming areas do as well such as the area around Fishtown and Kensington.



After looking into the count of sales by zip code, we still had more questions. We not only wanted know how many homes were sold per zip code but the average price per zip code. Our second map explores the average of sales by zip code, showing overtime the average price of a residence sold. From this we found that the zip code with the highest average price of a residence sold over time is 19103 which is the Logan area of Center City.



Other things that we explored are total number of sales by zip over the last 5 years (an analysis for each of the years 2016-2020) and from this we found similar results. Areas around Center City generally had a lot of home sales and it decreased as you moved away from the center, with certain up and coming areas like Kensington, Fishtown and some parts of the city south of Center City having an increase in sales on various years.

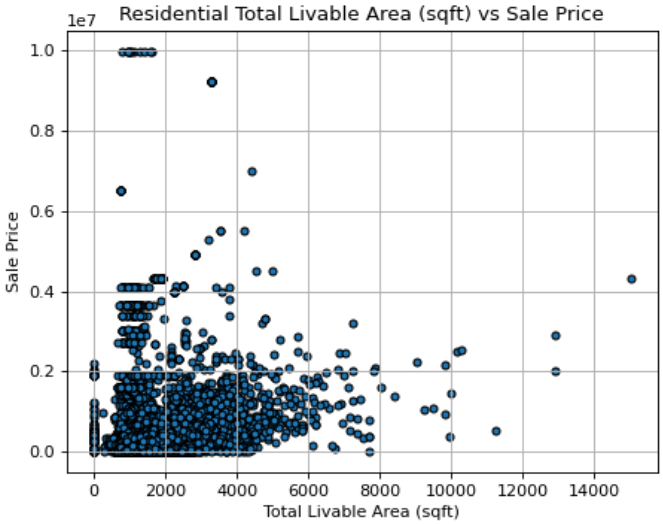
Scatter Plot Analysis

Question: Is bigger always better?

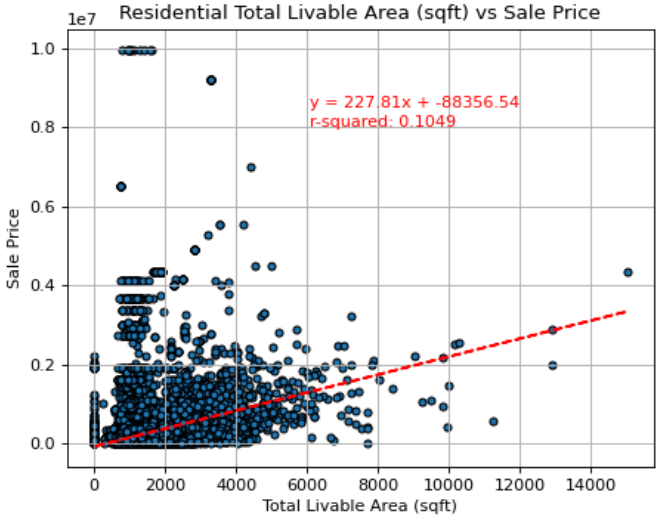
Description:

Throughout the course of history, humanity has pondered the age-old question: does size matter? Our group decided to attempt to answer this question as it relates to housing prices by utilizing scatter plots and simple linear regression. It may seem logical that the bigger the house, the higher the price. While this is what intuition may tell you, the data contradicts this intuition.

Initially, we plotted total livable area versus sale price across our entire dataset:



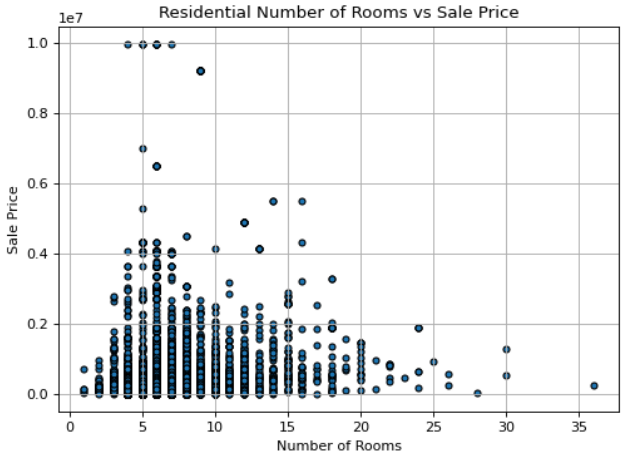
At first glance, it seems that there is may be a positive relationship between the two variables, meaning that as total livable area increases, sale price would also increase. However, when you layer in a simple linear regression, it becomes immediately clear that size plays a rather minimal role in driving sale prices.



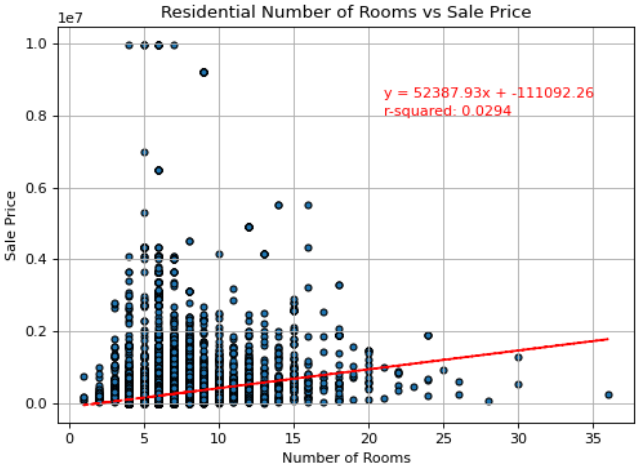
While the line of best fit is trending upward, the r-squared value of 0.1049 calculated from the linear regression tells us that only 10.49% of the variation in sale price can be explained by total livable area. Based on this

result, it can be determined that there are other variables at play when it comes to determining sale prices, and total livable area on its own is a relatively insignificant driver.

Analysis of total number of rooms versus sale price tells a similar story. When plotting total number of rooms against sale price, it seems as if there may be a slight positive trend.



Again, however, the calculated linear regression model shows us that the number of rooms on its own is an insignificant driver of sale price. The r-squared value of 0.0294 indicates that only approximately 3% of the variation in sale price is explained by the number of rooms.



Based on this result, it is clear there are other variables at play that are much stronger drivers of sale price than just size alone.

Conclusion

In conclusion, we found that Philadelphia has an ever-changing housing market.

From our summary chart, we learned median sale price gradually increased year over year, but the mean housing price is much larger than the median. From our five-year trend analysis we found, there is an inverse relationship between prices and interest rates. Our pie chart analysis showed us that many homes sold in Philadelphia are old construction. We also found that there is little correlation between sale price and total livable area as well as between sale price and number of rooms. Lastly our zip code analysis showed us that

Center City and the neighborhoods surrounding it are the most expensive and many of the homes in the city are sold in only a few zip codes. Certain areas of the city see a lot more activity than others. Which can explain a lot of outliers in our data as Philadelphia as a city has a lot of wealthy areas mixed with areas that are not as well off.

And overall, 2020 was an anomaly that we can't read too far into due to a pandemic that changed lives forever. If we were to continue this project, we would be interested in doing a forecasting model with housing to try to determine the future direction of the housing market.