NASHVILLE HOUSING DATA EDA

Ted Fitch
DATA610: Fall 2020
Assignment 2
Tedfitch4@gmail.com
Dr. Laila Moretto

Introduction:

Housing markets can be incredibly complicated being driven by a plethora of differing factors. These vary from economic factors (supply and demand, property value, average salary) to psychological factors (reservation of buyers, commitment of sellers) and beyond (Albrecht et al., 2016; Landvoigt et al., 2015). The Nashville Housing Data is an open source data repository used to practice basic data analysis. There are 56,636 rows of data with 31 columns. The data ranges from Jan 2013 until Oct 2016. There are a number of categories of data ranging from financial information, logistic information about the property, and identifying information about the sale and property. There are many rows with missing data. It's the goal of this Exploratory Data Analysis to take some complicated data and discover some basic trends.

The approach I would take towards this data is to take on the mindset of a first-time homeowner preparing to buy their property. Thus, the initial questions I want to explore would only pertain to single family residences: What drives sale price and total value? When are the highest volumes of sales and when are the highest average sales price? What factors into getting a good deal? These are the kinds of questions that would be useful both to sellers and buyers. Thus, this information may also be beneficial when it comes time to sell the property (providing the patterns hold). For example, sellers would want to know when the highest volumes and highest average sales price occur so that they can command the highest sales price for their property. Buyers would behoove themselves to know where the lowest sale prices are occurring, and especially to know where the highest ratio of total value to sales price is occurring.

Data Preparation:

There were several steps taken to prepare the data for any analysis. First, the data was observed to see how many rows there were, what kind of data this was, how much missing data was there, and what were the patterns of the missing data. The repository began with around 56,000 rows worth of data. The columns included many pertinent factors like: sale price, building value, acreage, finished area, and number of bedrooms. It was also noticed there were some irrelevant columns which would likely not contribute helpful information. These included: name, address, state, parcel ID, and legal reference. The name of the individual owner of the property likely wouldn't elicit any interesting patterns. There were two columns for address: owner address and property address. In every instance of a subset of the data which was checked, these addresses were the same. It was decided to keep property address and omit owner address. There was no reason to keep the state column since this data is inherently from Nashville; all data will be from Tennessee. Lastly, both the legal reference and parcel ID were thought to be arbitrary identifiers. The legal reference likely speaks to a government identification pointing to the specific instance when the property was sold. Likewise, the parcel ID likely is a government identifier for the land which would be arbitrary towards the questions of purpose. Once these columns were hidden, it was time to tackle the missing entries.

When examining the pattern of missing data, two things were noticed: 1. A majority of rows with missing values were missing data from the column "land value" through to the last column. 2. The other missing data were ad hoc entries. There were several columns chosen to serve as proxies to cull a majority of complete data rows. These were: land value, total value, acreage, and bedrooms. This was decided based on the patterning of missing data since most rows missing one of these values were also missing many other values. Thus, a filter was added demanding land value, total value, acreage, *and* bedrooms would be "not missing". After this, the data was scrutinized to see what else was missing. A set of 12 rows has 7 columns empty (which were not the ones mentioned above). Therefore, finished area was added to the above filter which excluded the 12 rows. All remaining rows contained all included information with the exception of one row where full bath and half bath was empty. This was kept since the questions I'm more interested in revolve around size of the house or land, and price instead of focusing on the number of rooms. This brings the number of rows to around 21,000 which is an acceptable number to work with when determining patterns over a 4-year period.

Data Exploration Process:

As previously mentioned, the questions a buyer would be primarily interested in center on: When is the best time to buy? What drives value? In order to determine this, a new variable was created: TV/SP ratio (*r* will be used as shorthand). This is the total value divided by the sales price. In this manner, a single variable shows if the sale was a better deal for the buyer or the seller. This ratio assumes that total value demonstrates the true value of the property and there aren't any confounding factors. Total value is comprised of building value (which is assumed to factor in finished area, bathroom number, bedroom number, foundation, etc.) and land value (which is assumed to factor in acreage and location). It is possible that total value doesn't represent the true value of the property (for example if this datapoint doesn't consider location, school zones, traffic, house condition, or other factors which would drive prices down). But, we'll move forward with the assumption that total value does represent the true value of the property. With that caveat stated, a high ratio ($r > 1$) demonstrates a better deal for the buyer since they are paying a lower amount compared to what the property is worth (and conversely $r < 1$ is a better deal for the seller). There were 5 outliers for this category: 1193, 571, 315.5, 166.9, and 53.767. No pattern was observed for these except that the sale price was an unusually round number (100, 100, 800, 1,000, and 3000 respectively). It's possible this data was mis-entered into the database. Alternatively, it's possible these sales were actually transfers of property between family or friends where they had to apply a price to the property for tax purposes (so they sold the property for an unreasonably cheap price). These 5 entries were also removed via filter so as to not skew the results dramatically. While there were a few other properties with low sales prices compared to the house total value, these 5 are significantly larger than all other values (for example, there are 2 values in the 30-40 range and 17 values in the 10-20 range).

With the outliers taken care of, the first exploration was to find the average TV/SP ratio across all time (r = 0.87). This means that the general housing market in Nashville has recently been a better deal for the seller than for the buyer on average. However, when examined over time, the ratio appears to decrease significantly (this will be explored further in a later visualization). Average price is around $290,000 for a single-family residence while average total value of a residence is $253,000. The range observed in *r* between cities was 0.16 however the range observed between tax districts was 0.34. There were 3 tax districts found to have an *r* > 1 (City of Belle Meade, City of Oak Hill, and City of Forest Hills). As someone searching for a good deal on a house, I would keep my eye on these 3 districts.

Visualizations Created:

The first visualization is a spiral diagram to determine drivers of sale price (**Figure 1**). As one might expect, grade (56%), total value (50%), land value (46%), building value (40%), and finished area (37%) were the largest single contributing factors to sale price. It’s interesting that land value has a 6% higher driver factor than building value. This makes sense in a region like Nashville which is the capital city of Tennessee because land tends to be scarce around city centers. Thus, having nicer land normally would be a higher contributor to sales price than having a nicer building. To no surprise, the other single drivers were bedroom count (25%), half-bath count (14%), bath count (11%), and acreage (10%). It’s important to see visually that there’s not a single outlier driving sales price more than anything else. For example, it would be noteworthy to find foundation, exterior wall, location/district, or year built were a critical, single driver. Year built and exterior wall both appear as combined factors; however, they don’t stand out on their own. Thus, it’s safe to say Nashville is a generally normal housing market driven by

Next, the average *r* was examined for residences being sold as vacant (**Figure 2**). Average *r* was over 3 times higher for properties sold as vacant versus properties occupied at the time of sale. It’s important to caveat that 99.4% of all entries were not sold as vacant. Thus, the takeaway is it’s rare to find vacant properties on the market; however, if a buyer can find one, they will likely have strong leverage to command a high *r*. In all likelihood, it’s because the seller is motivated to sell it quickly since the house is empty and not generating profit.

The third visualization demonstrates raw count of residences sold over time (**Figure 3**). Sale dates were binned into a single month (for each year). This is because visualizing the average sales per day would not be clear and precise. Showing the average per day would show too many fluctuations (because a day is too small of a timescale), so a greater timescale was chosen. There’s a clear trend of high volumes during the Summer with June being the highest month every year (with the exception of July in 2014). This is a key aspect to know when watching the market and is a well-characterized trend (Ngai & Tenreyro, 2019). If highest volumes are occurring through the Summer, the best time to be looking is through the entire Summer and was the season ends. As the “buying season” closes (towards Aug, Sept, Oct), sellers are likely to lower their prices if their houses haven’t sold yet. Unfortunately, this database doesn’t contain the data whether prices were lowered. However, an alternative time graph was made to observe what the average *r* is per each month. Unfortunately for buyers, this graph shows that the market is becoming worse over time. The average *r* begins at 1.35 in Jan 2013 and ends at 0.71 in Oct 2016. This dramatic 0.61 drop suggests that the Nashville market is becoming saturated. Either supply of housing has decreased or demand for housing has increased over the 2013 to 2016 period. While this graph doesn't support the above hypothesis that sales will have a high *r* during Aug, Sept, and Oct, it does show and important downward trend. If a potential buyer is deadest on the Nashville area, they should buy quickly before the trend gets worse (*r* = 0.80: average of 2016 data). Alternatively, they should look in another market where *r* is higher or wait until the Nashville trend reverses in favor of the buyer. Intuitively, it is the nature of cities to have a decreasing *r* trend over time because their populations increase causing demand to increase. An increased demand will always drive prices upwards (without an alternative supply…which is not likely in the case of single-family residential housing).

The final comparison made was using grade (**Figure 4**, **Figure 5**). Little information was found online for how grades were defined. Thus, it's assumed this grade is information the realtor keeps or is found in particular databases but can be found by buyers when investigating deeper into a property. Three analyses were made, one demonstrating average TV/SP ratio per grade, another demonstrating average sales price per grade, and lastly a tree map showing the proportionality of each grade. The first thing noticed is there are 5 grades where *r* > 1: A, B, E, SSC, and X. This is necessary to know when attempting to purchase a house which grades are likely to be the best deal. A, B, OFB, and X are all above average sales price ($290,000) while C, D, E, and SSC are all below average sales price. Grade X appears to be a category for mansions based on most entries having unusually high acreage, finished area, bedroom count, bathroom count, and price (which sits at an average $1,400,000 at 409 entries). In contrast, OFB and SSC each have 1 entry which appears to be an average house. Because there's only 1 entry each, there's nothing that can be said about a trend for these categories. However, a clear trend is observed for grades A and B where they both have above average sales price and have a ratio which generally favors the buyer ($r > 1$). Importantly, these both carry enough entries to call it a trend (452 and 2802 respectively). Lastly, it's important to point out the vast majority (70%) of entries are grade C and yet grade C has a $r < 1$ meaning it's not a good deal for the buyer. A buyer should automatically filter out all grade C entries (and D) while looking mostly towards categories A, B, E, or X depending on what fits their budget.

Calculation:

The primary calculation which has driven the analysis in this report is the TV/SP ratio (total value / sales price). As previously discussed, this factor demonstrates when a sale is a good deal ($r > 1$) or a bad deal ($r < 1$) for the buyer. This is mission critical in evaluating an individual property and in determining trends over the long-term. Another equation would also be useful for first-time buyers:

$$D = R(t) * P(t)$$

This describes a tradeoff between saving up a down payment (P) over time versus the likelihood that market will become worse over time. R is the average r for any given month and is likely to decrease over time. There are certainly fluctuations between buyer's markets and seller's markets; however, housing in and near cities tends to become more expensive over time because demand increases while supply remains static. Both R and P are functions of time and would need to be modeled more thoroughly in order for this calculation to work effectively. As time increases, R will decrease and P will increase; the point at which they intersect would be the ideal time for a person to buy (D). Buyers need to be aware of the how the R in their region is decreasing. This will help them calculate how much money they need to save up before reaching a critical mass (where the tradeoff with *R* is no longer acceptable). It's important to also note that in a buyer's market *R* will be > 1 and thus will increase the value of the down payment.

Summary:

Based on the data and attached analyses, the Nashville area housing market has recently been more of a seller's market. There are still plenty of opportunities for first-time homeowners to find a great home at a good deal; however, they'll have to apply the principles discussed here. Buyers should:

1. **Know** that finished area, acreage, bedroom count, and half-bath count are the largest single drivers of sale price. Buyers need to determine what's most important to them.
2. **Look** for homes being sold as vacant.
3. **Understand** the market is currently getting worse for buyers. They may want to consider metro areas outside of Nashville or consider moving quickly before the market gets worse.
4. **Keep** watch during the Summer when the highest volumes of houses are sold. While no months were found to have the best deals, most houses are sold during the Summer months. Buyers need to keep their eyes peeled to have the highest likelihood to find what they are looking for.
5. **Filter** out grade C houses and look towards A, B, E, or X (depending on their budget).

With persistence and by following the above advice, first-time buyers are likely to find themselves a great deal.

References:

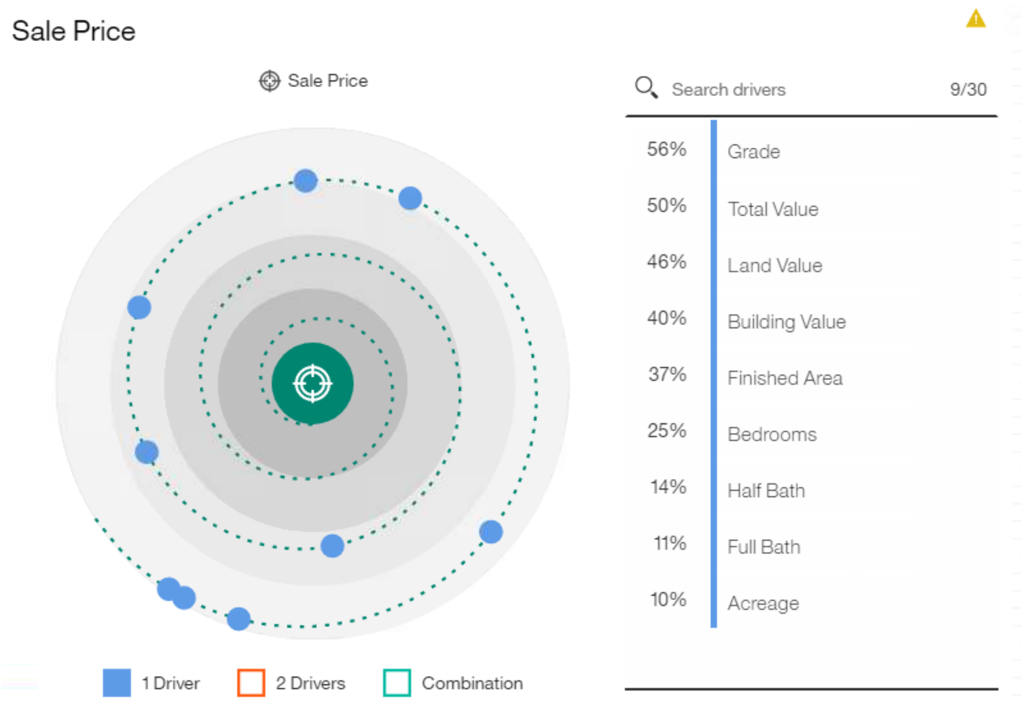
Albrecht, J., Gautier, P. A., & Vroman, S. (2016). Directed search in the housing market. Review of Economic Dynamics, 19, 218-231.

Landvoigt, T., Piazzesi, M., & Schneider, M. (2015). The housing market (s) of San Diego. American Economic Review, 105(4), 1371-1407.

Ngai, L. R., & Tenreyro, S. (2019). Replication data for: Hot and Cold Seasons in the Housing Market.

Appendix:

Figure 1. Spiral graph showing the calculated drives of TV/SP Ratio (Total Value divided by Sales Price). Note: only single drivers are shown in this visualization.



56%	Grade
50%	Total Value
46%	Land Value
40%	Building Value
37%	Finished Area
25%	Bedrooms
14%	Half Bath
11%	Full Bath
10%	Acreage

Figure 2. Graph showing the average TV/SP Ratio (Total Value divided by Sales Price) whether a house is vacant when sold or not.

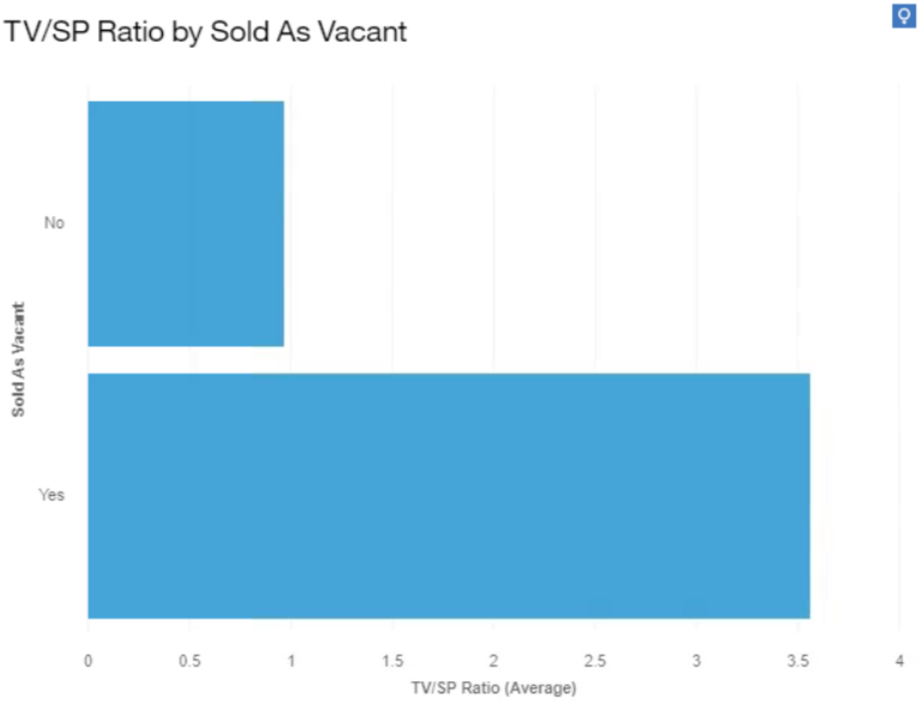


Figure 3. Graph showing the average TV/SP Ratio (Total Value divided by Sales Price) for each month (left). All sales for each month were binned together and averaged. Graph showing the count of sales per month (right).

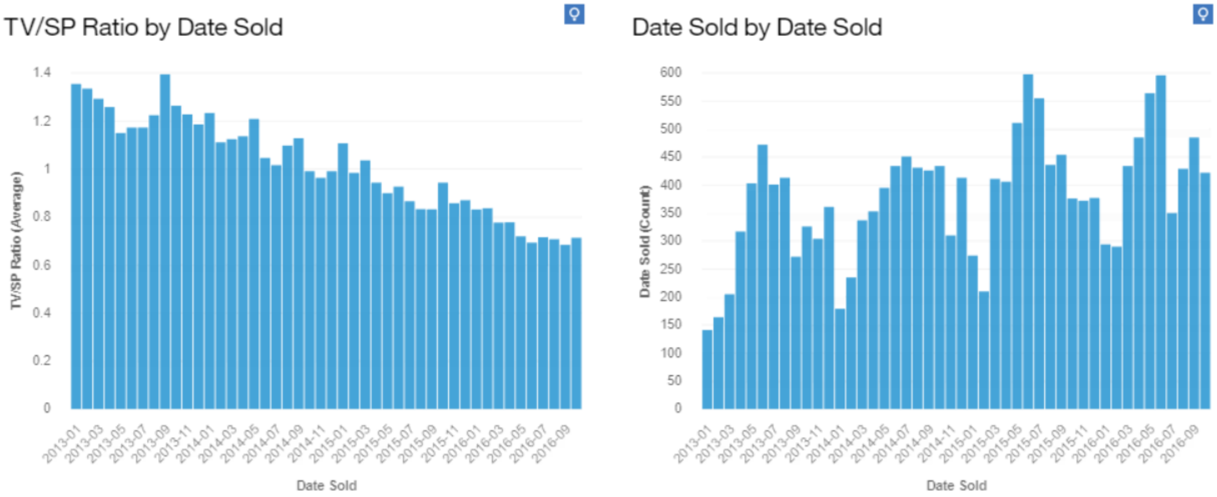


Figure 4. Graph showing the average TV/SP Ratio (Total Value divided by Sales Price) for each grade (left). Graph showing the average Sales Price for each grade (right).

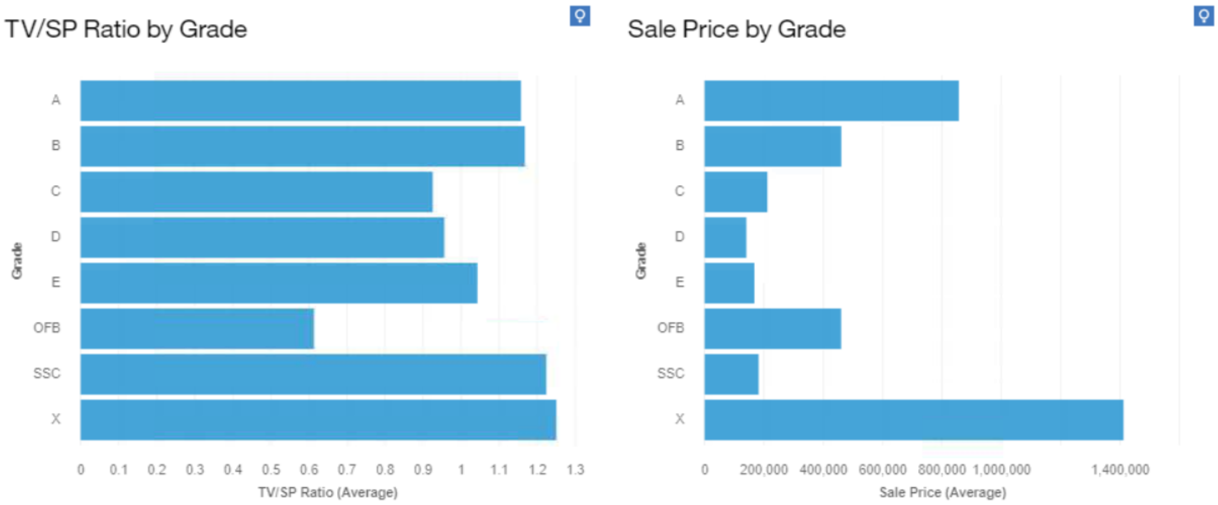
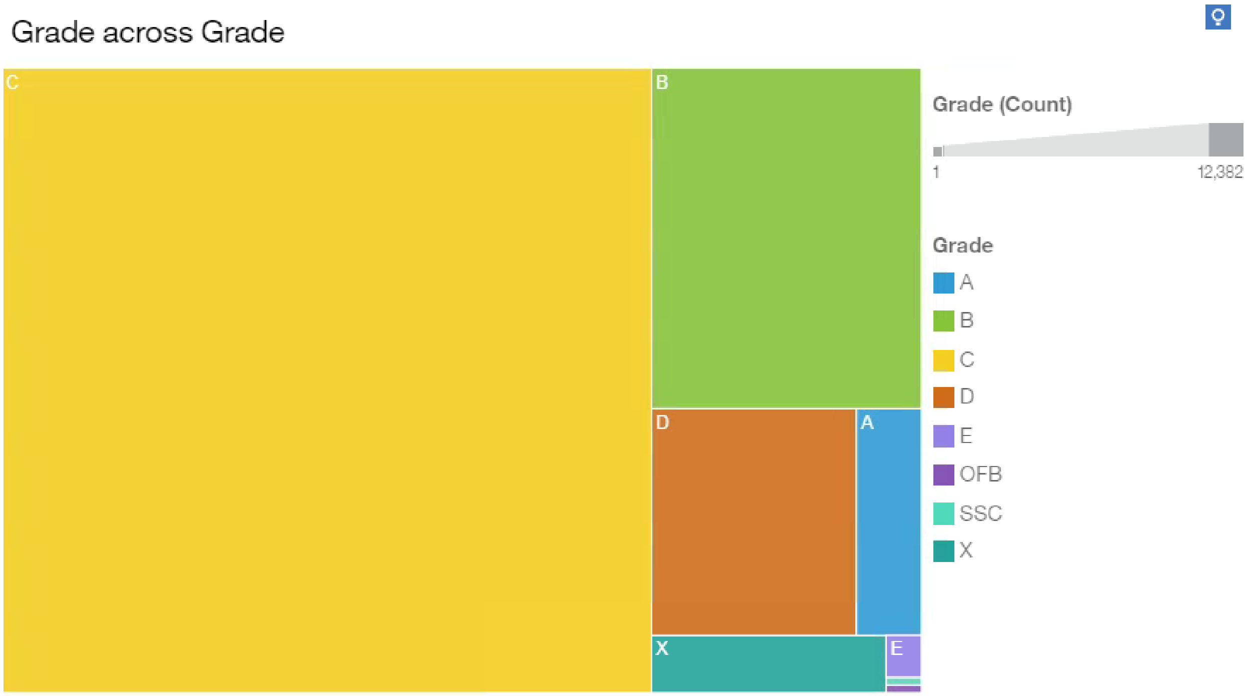


Figure 5. Graphic showing the raw count of each grade to demonstrate proportion of each grade.


Grade across Grade
C
B
D
A
X
E
Grade (Count)
Grade
A
B
C
D
E
OFB
SSC
X