Data Management Systems

Course Number: 610

Course Term: Summer, 2020

Assignment number: 2

Project title: Exploratory Data Analysis (EDA) using Cognos Analytics

Contact Information:

Melissa Hunfalvay

Email: [melissa@righteye.com](mailto:melissa@righteye.com)

Phone: 240-357-0026

Instructors Names:

Dr. Moretto

Dr. Goldberg

Introduction

This data set represents home value data for the Nashville, Tennessee market. The data set contains 56,636 number of cases or rows of data. There are 31 columns (variables) of data (see Figure 1).

The dataset provides the following variables that may be used to develop predictive models; information regarding the type of property e.g. condo (25%), single family home (61%) and other (15%). The address of the property shows 40,280 properties in Nashville (71%) and surrounding cities such as Antioch (11%), Brentwood and others (18%) (see Figure 2). Pricing includes sale date, price, and legal reference. It also includes whether the property was vacant when sold. The dataset includes additional information about the owner and property including name, taxation district and neighborhood. It also includes value of the land, building and total value. As well as other housing information such as foundation type and year built, types and number of rooms.

Data is missing in some address fields such as the address, suite/condo#, city and state. Data is also missing in the home features, some examples include land value, year built and number of full and half baths. Nevertheless, there are 45,069 unique values present in the property addresses and 56,636 sale prices listed with an average sale of $327,211.13. Therefore, although some data is missing there is a significant amount of unique data to make the dataset robust for exploration.

An interesting insight when initially examining this dataset was the Suite/Condo# had many “null” columns. Null columns reveal missing data. However, in this case, it is important to recognize that there could be two reasons why the data was missing; a) the property was not a suite/condo; or, b) the data was not known. This is an important distinction, as 75% of this column shows a null result.

Initial insights into this dataset reveal that it contains Personally Identifiable Information (PII) such as the Owner Name variable. It is important to be aware of sensitive information and how it should be properly stored and secured as data analysts.

Data Preparation

*Identify and remove irrelevant variables:* One method for identification and removal of irrelevant variables was to view the grid and determine if the data from a variable was the same and therefore added no value. For example, the “State” variable included two conditions, TN (Tennessee) or null, indicating that all the properties were within the state of Tennessee or the state was not reported, in the case of a null response. Furthermore, there is a replication of the variables Address, City, State in the dataset. The replicated data was deleted. This reduced the size of the overall dataset.

At this stage of the data preparation process, as an initial set of questions were not decided upon, and as the writer is not an expert in property data, no other variables were deleted. In other words, a decision was made to air on the side of leaving as much data in place as possible as “overcleaning” and removal of what may be seen as initially irrelevant data, may become relevant after the data is further explored. For instance, Parcel ID and Legal Reference are two variables that fit this reasoning.

*Missing data:* To identify and solve for missing data filters were added. For instance, if sale price was zero it was replaced with null. Then all null values within the column were filtered so that the remaining data set included only data with a valid sales price. An embedded filter was used to remove all null data from tax district, neighborhood, land value, building value, total value, finished area and foundation type.

After exploration of the data and determination of the initial questions more filters were added. For instance, the land use variable was reduced to single family only. As the suite/condo # was now no longer relevant this column was also removed.

An imputation was used to replace missing values in the acreage variable by calculating the average value. A selectable filter was used to remove all single-family homes that were not built between 2000 to 2016.

Data Exploration Process

The approach to data exploration was to use the visual relationship map to guide further questions and explorations using the Assistant. Ultimately this led to the creation of meaningful questions.

The purpose of the relationship map was to determine if a variable was connected to other variables and what the strength of those connections were. The line width and connections between variables was able to assist in visually exploring if the variable was contributing and by how much. If the relationship was meaningful the Assistant was employed to explore the data further.

Outliers were identified in the Exterior Wall (see Figure 3). The categories of Frame, Frame/Stone, Metal, Stone, and Stucco were all low in number and in value. They were therefore filtered and removed from the dataset. Brick and Frame were the most important categories of Exterior Wall with a total value of 810.96 (83.7 %).

Upon reviewing the relationship map (see Figure 4) and a follow-up visualization (bar chart; Figure 5) one half-bath variable was very highly related to sale price. However, at no half-baths and 3 half-baths the relationship to price was negligible. Therefore, another filter was added to the half-bath removing the no half baths and 3 half baths as they were outliers.

Based on the exploration of data the following set of questions were explored. The overarching question was: *What are the most important features that determine the sale price of a home in Nashville* (see Figure 6).Specific questions included:

1. *What was the average sale price of a house in Nashville between 2000-2016?* From 417 records, the average price was $902,487.05.
2. *What variables impacted the sales price?* Results show a moderate relationship between sales price and total value (47%), building value (44%) grade (41%). A weaker relationship between sale price and acreage (21%), half bath (36%), full bath (31%), finished area (35%) and land value (34%).
3. *Did the year of sale impact the price? If so, how?* To do this the Sale Date needed to be split so that Year was parsed out into a separate column (titled year) for the analysis. Once done a visualization (line graph) was created (see Figure 7). The graph shows the Sales Price remained stable with only slight increases for 2013 ($823,663) to 2014 ($826,596) to 2015 ($853,511). However, in 2016 there was a significant increase in the housing price consistent with the economic boom at that time ([https://www.fortunebuilders.com/nashville-tn-real-estate-market-trends 2016/](https://www.fortunebuilders.com/nashville-tn-real-estate-market-trends%202016/#:~:text=The%20second%20quarter%20saw%20one,national%20average%20of%204.9%20percent.&text=Homes%20purchased%20in%20the%20Nashville%2C%20TN%20housing%20market%20three%20years,%2446%2C878%20over%20the%20same%20period.)). Housing prices in 2016 rose to an average of $1,162,270. Therefore, year of sale impacted the sale price by an average of 27%.
4. *Did the grade given to the house impact the sale price? If so, how?* Results reveal that the grade of the house drives the sales price by 41%. An important discovery however is that the Sale Price is unusually low when the property is given a “bad grade” of C (see Figure 8). The Grade had a moderate relationship with both Building Value (56%) and Total Value (61%). Building value and total value had a very strong relationship with one another (81%).
5. *Did the combined bathroom features within the house impact the sales price? If so, how?* The combined house features of interest (based on the relationship map; Figure 6) are full bath and half bath. A new variable was created called Bathrooms. A calculation was made to add the full and half bath variables together. Then a Data Group called Bathrooms (group) was created with 11 total categories. Results revealed that when combined the bathrooms group variable slightly drove the sales price at 24%. What is interesting is that 24% is a lower than the variables of full bath and half bath separately. Nevertheless, the number of bathrooms to sale price has a very linear upward trend (see Figure 9, left side). When combining the bathrooms with Grade of property 3 bathrooms occurs as an important factor related to grade.

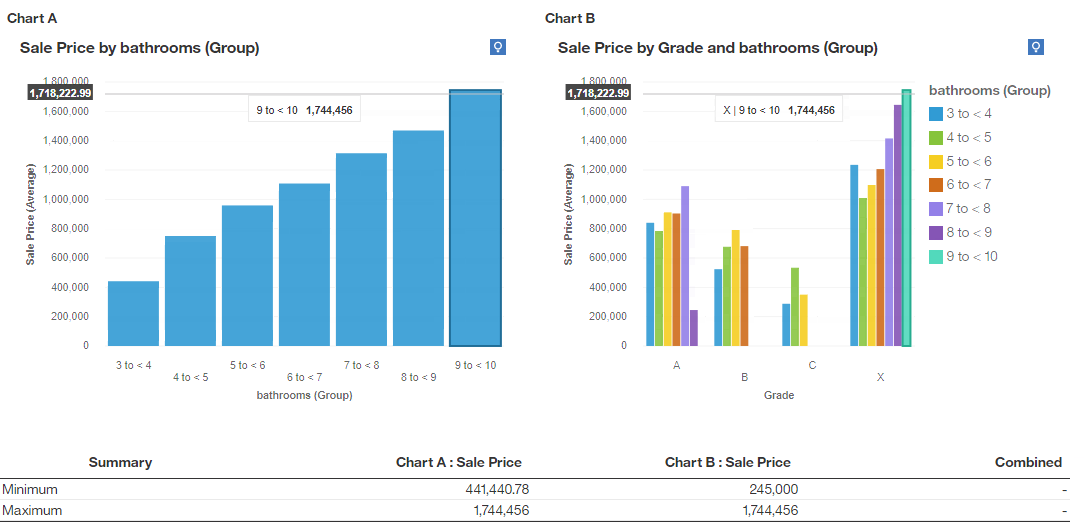


Figure 9: Number of bathrooms to sales price (left side) versus number of bathrooms to sales price and grade (right side)

1. *Was the sales price impacted if the property was vacant during the sale?* If occupants were in the house when it was sold the sale price was and average of $919,174. However, when the house was vacant the average sales price was $484,251. This is a staggering 52% difference and begs further investigation. When adding a third variable of Tax District the results reveal further insights (see Figure 10). Results show that the City of Oak Hill has the least discrepancy in sale price (22%) if the house was vacant or not during the sale. The City of Forest Hill has the largest discrepancy at 31%. The data reveals that when the houses are sold in all but the General Services District, they sell for a higher value when people are in the house. Interesting insight would have been if the house was “staged” or not and if that impacted the price (<https://www.forbes.com/sites/vanessamcgrady/2015/11/04/staging/#7c59881e50c9>). According to Betsy Wilber, who professionally stages homes for sale, “A staged home will sell for 17% more on average than a non-staged home, and 95% of staged homes sell in 11 days or less. That is statistically 87% faster than non-staged homes.” However, this information was not in the dataset. Another interesting insight would be adding average income level per household to the different Nashville burrows. Income plus the variable of house occupancy at the time of sale may provide further insight into the neighborhood (<https://www.businesswire.com/news/home/20200224005188/en/Regional-Variances-Household-Income-Influence-Affordability-American>).

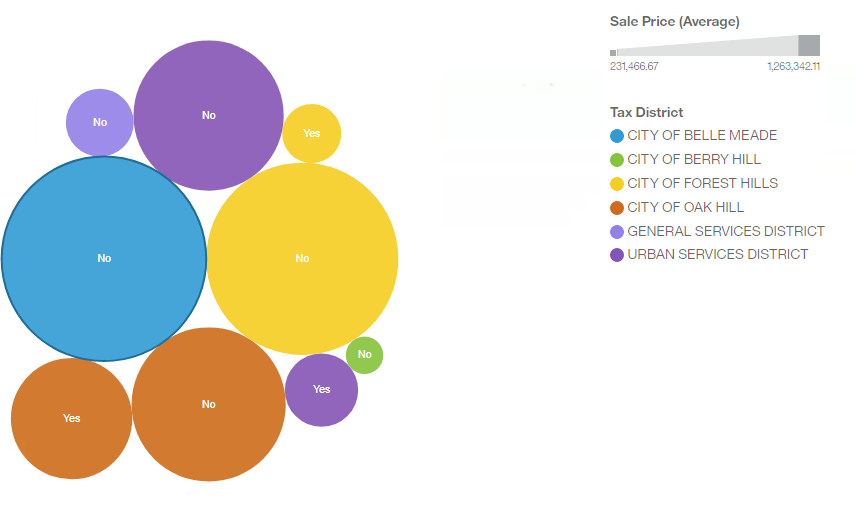


Figure 10: No = people were in the house when it was sold, i.e. it was not vacant. Yes = people were not in the house when it was sold i.e. it was vacant.

In conclusion, the most important features that determine the sale price of a home in Nashville include the year the home was sold, if the home was vacant or occupied when sold, the total value and building value of the home at the time of the sale.

References

Fleming, M. (2020). Regional Variances in Household Income Influence Affordability, According to First American Real House Price Index. Retrieved from <https://www.businesswire.com/news/home/20200224005188/en/Regional-Variances-Household-Income-Influence-Affordability-American>

McGrady, V. (2015). 7 Pro Tips to Help you Home Sell Faster, for More Money. Retrieved from: <https://www.forbes.com/sites/vanessamcgrady/2015/11/04/staging/#7c59881e50c9>

Merrill, T. (2016). Nashville, TN: Real Estate Market & Trends 2016. Retrieved from <https://www.fortunebuilders.com/nashville-tn-real-estate-market-trends-2016/#:~:text=The%20second%20quarter%20saw%20one,national%20average%20of%204.9%20percent.&text=Homes%20purchased%20in%20the%20Nashville%2C%20TN%20housing%20market%20three%20years,%2446%2C878%20over%20the%20same%20period.>

Appendix A

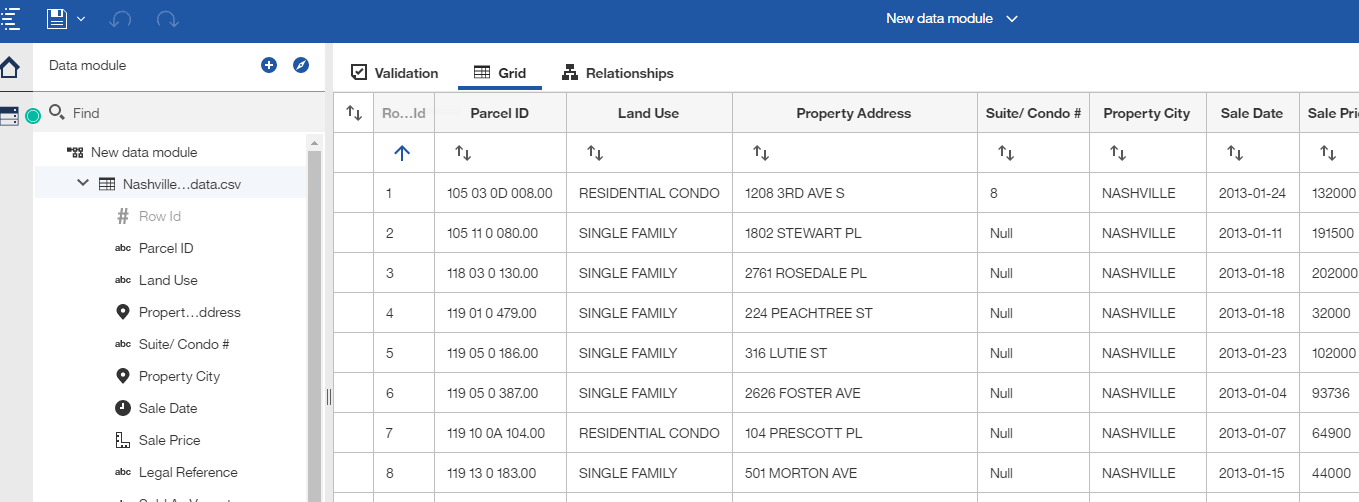


Figure 1: View of a portion of the Nashville data in My Content Cognos Analytics

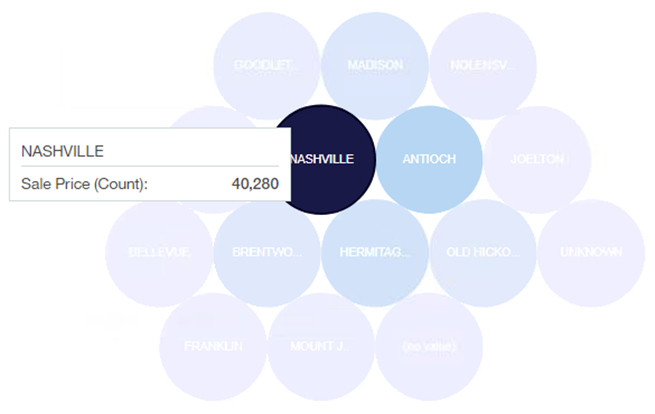


Figure 2: Property city and sales price by count

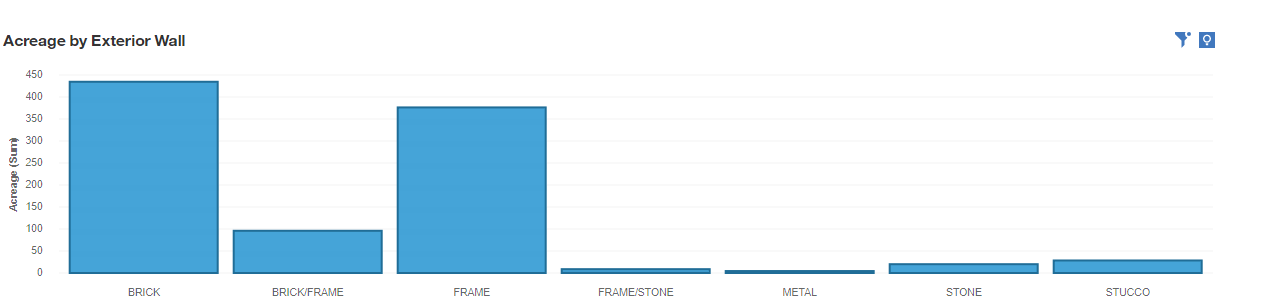


Figure 3: Identification of outliers in the exterior wall variable.

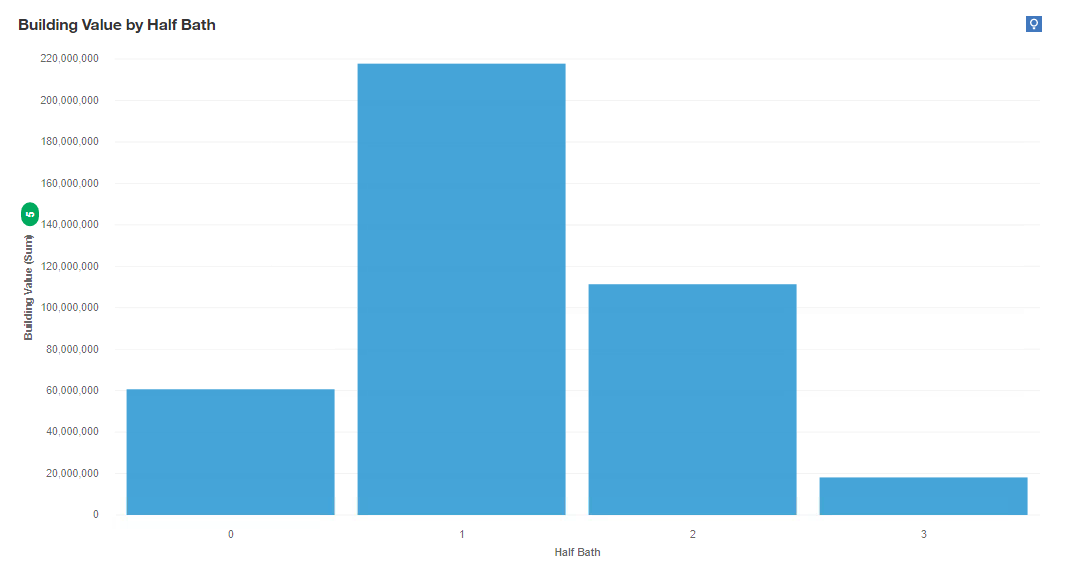


Figure 4: Half-bath bar chart to identify outliers.

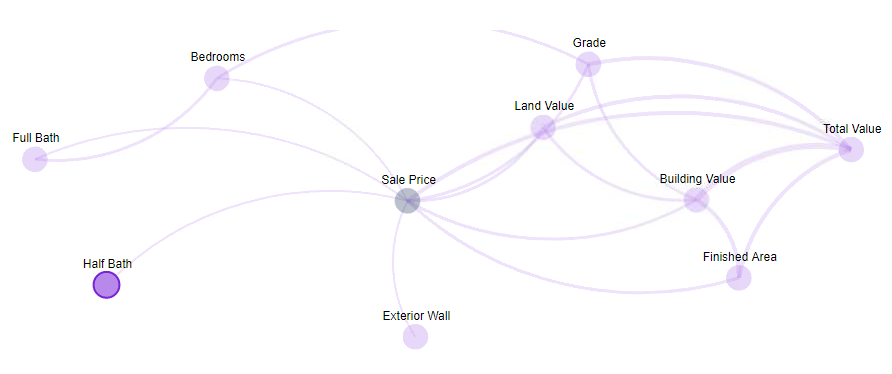


Figure 5: Half bath relationship map to identify outliers

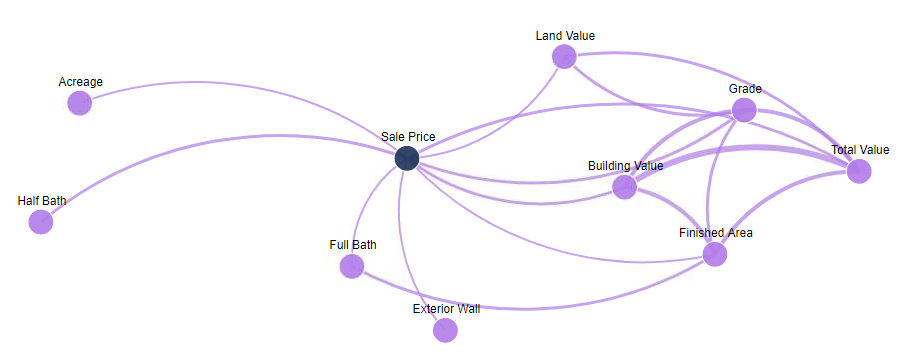


Figure 6: Relationship map used to identify the most important features that determine the sale price of a home in Nashville

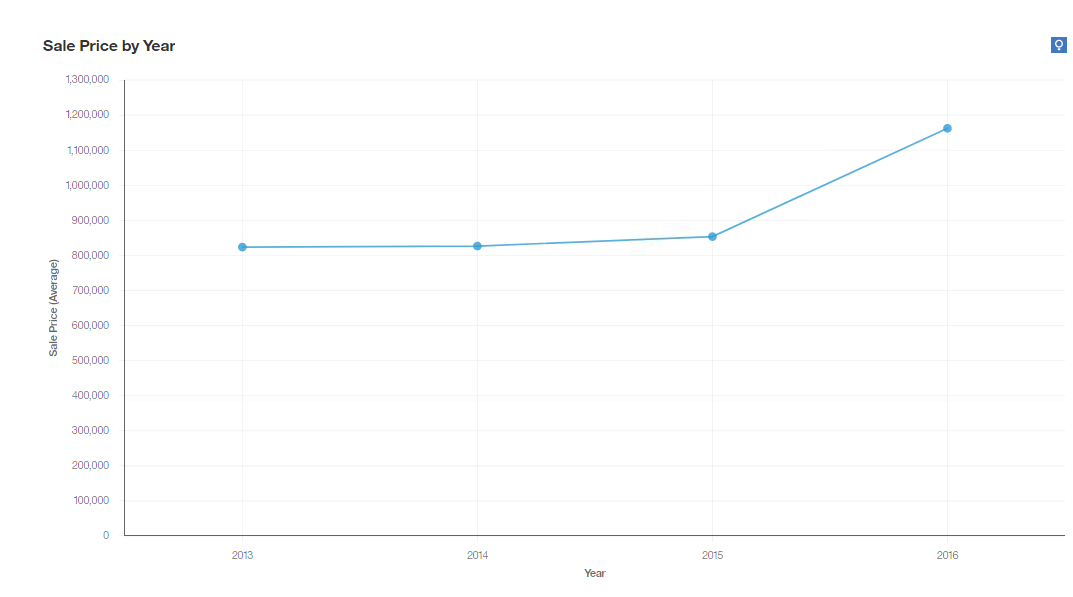


Figure 7: Sales price by year

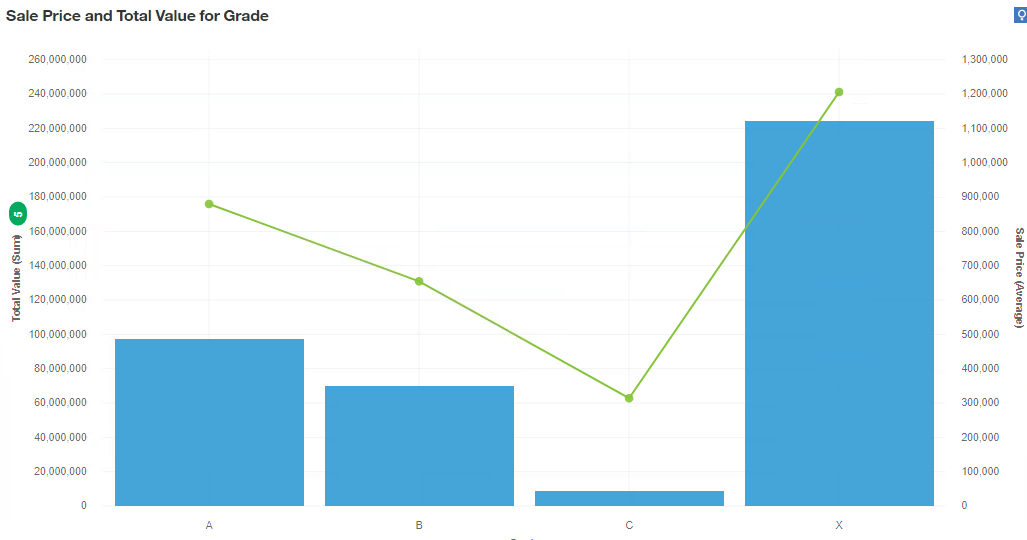


Figure 8: House grade (A, B, C, X) and sales price