**Observations and insights through basic EDA:**

1. There in one instance in the dataset where an entire row is duplicated. The dataset needs to be deduped for processing
2. There is an instance in the dataset where the same house has been sold twice. In the housing market, renovations etc. can be accounted for in the latest sale and hence the latest sale entry is only retained. The older entry is excluded from the analysis
3. Additionally, another variable indicating the number of times a house has been sold historically was also created to understand the impact of resale as a phenomenon
4. The year in which the house was built is available. Using the year as such in the analytical dataset might not be the best way since interpretation of the variable is not straightforward. Instead, it indicated the age of the building. Hence, a more direct relationship needs to be created in the form of the age of the house
5. The variable SALE\_PRICE\_LOG10 is the log of the actual sale price. It will not make sense in the real world to look at logarithmic values to understand the actual value of a property. Hence the modelling should be done against the base sale price which can be calculated by taking an antilog of the mentioned variable
6. A few of the variables have a higher standard deviation as elicited by the univariate analysis. Eg. FIRST\_STORY, SECOND\_STORY, UNIT\_PRICE. Some of these variables also have a high correlation with the dependent variable indicating a good predictive capability and the ability to explain the variability in the dependent variable as well
7. The dataset provides geo-coordinates which are extremely valuable. Plotting them on a map by categorizing the houses as high priced, medium priced and low priced can help understand where the premium localities are vs where are some of the affordable neighbourhoods in the market
8. The sale date in the dataset plays a much bigger role than just recordkeeping. Plotting the house prices in a temporal sequence by using the sale date will indicate a trend towards increasing or decreasing prices. This can be broken by region using the NBHD code
9. Additional cuts can be prepared by square footage, number of bedrooms etc. to understand how are different categories performing over time in terms of sale price
10. The variable ‘buildings\_300m’ is an excellent indicator of how densely packed is the neighbourhood and concomitant average prices for the houses within the same NBHD code will provide a look at how densely populated region are priced compared to how sparsely populated regions are priced in the city of Saturn
11. The construction grades from Grade A to Grade E reflect the quality of construction and can turn out to be a good indicator of the sale price
12. Additionally, the distribution of zeroes in these binary variables is such that houses in the best and worst categories are rare, followed by the next two categories and the maximum houses have an average construction quality. It might also be due to the balance of construction quality and cost
    1. **Average price for grade A construction is USD 556K vs for construction that is not grade A is only USD 97K**
    2. **Grade B average: 215K USD vs Not grade B average: 91K USD**
    3. **Grade C average: 96K USD vs Not grade C average: 99K USD**
    4. **Grade D average: 62K USD vs Not grade D average: 102K USD**
    5. **Grade E average: 57K vs not grade E average: 96K USD**
13. It is worth investigating the differences in property rates across the different types of intended usage – residential, commercial, agricultural etc.
14. Sale price difference across:
    1. **No sewage facilities: 62K USD**
    2. **Private: 297K USD**
    3. **Public: 97K USD**
15. Average sale price for:
    1. **Houses with fireplace: 258K USD**
    2. **Houses without a fireplace: 77K USD**
16. Average sale capacity of the basement garage:
    1. **3 cars: 324K USD**
    2. **Other capacities: 108K USD**
17. Interestingly, houses built in the first three quarters of the 20th century see a lower price fluctuation compared to houses built before and after
18. As the number of floors increase, price increases gradually
19. The **top three NBHD codes** with the highest average sale price are **15780, 15540 and 15310**
20. The number of buildings in the 300 m radius and the average sale price have an **inverse and strong correlation with a correlation coefficient of 0.87**. This indicates that the fewer houses around the property, the more likely it is to be priced higher. It may be asserted that fewer houses in the neighbourhood elevates the sale price for the property