SUMMARY

- Several variables have significant correlation with the Sale Price which enables us to predict the pricing trend.
- The variables that are used for this analyses comprises of numerical and categorical. The numerical variables are 'YearBuilt','SalePrice','FullBath','Fireplaces' while the categorical variables are 'MSZoning','SaleType','Neighborhood','LotShape'.
- These variables are chosen because they show a correlation with the Sale Price which indicates that there are relationships between Sales Price and each of the variables
- The summary correlation of the numerical variables:
    - Fireplaces: The more the fireplaces, the higher the Sale Price
    - Full Bath: The more the number of full bathrooms above grade, the higher the Sale Price
    - Year Built: Although weakly correlated, the tendency of the data shows that the newer houses have a rising sale price trend. However, year built alone is not a accurate enough tool to predict the sale price of houses.

- The summary correlation of the categorical variables:
    - MS Zoning: In general sales price are higher in Floating Village. However, there are many outliers in the Residential Low Density (RL) zones with a much higher sales price. As for the outliers, it is very possible that the high numbers are due to errors of classifying into the wrong zone. It is observed that several zones have zero values in the dataset.
      Thus, for investments, it may be more profitable for investors to invest in Floating Village (FV) zones due to the higher average sale price and the data spread.

    - Lot Shape: The average sale price is highest in the IR 2 lot shape but with the least data dispersion. However, it is observed that if outliers are included IR 1 would have the highest sale price that far exceeds the other lot shapes. Thus, if it were assumed that it is a possibility of high bid buyer, then IR 1 may pose as a more profitable option to invest in.

    - Sale Type: The most favoured sale type is New sale type where home is just constructed and sold. It is closely followed by Con where contract is 15% down payment on regular terms. It is also observed that there are many outliers in the WD sale type which can either be an error or indicates that some buyers prefer warranty deed for house sales above 300,000.
      For investments, it is better and less risky to opt for New sale type as it has a large data disperse with a large median value.

    - Neighborhood: Certain neighborhoods like Northridge Heights (NridgHt) and StoneBr (Stone Brook) have a large data dispersion with relatively high median sale price. The highest average sale price is located at Northridge. However, there are outliers with sale price that far exceeds the rest. For instance, the neighborhood North Ridge has two outliers above the level of 700,000. Considering the data dispersion and the median of the sale price, it is better for investors to not take the risk and invest in neighborhoods such as Northridge Heights.

- For better visualization, refer to the plots located in the Results folder within the repository.