**Intro** - The real estate sector is a significant sector with such a wide range of participants, including financiers, private corporations, and governing agencies. There is a significant need among these participants for a deeper comprehension of the driving forces and operational mechanisms of the sector. Since the market for real estate is among the most competing in terms of pricing and that price can differ significantly due to numerous factors, predicting real estate prices is a crucial device in decision making for both buyers and sellers in continuing to support funds allocated, attempting to find real estate finding strategies, and deciding appropriate policies, trying to make it one of the top fields to pertain the machine learning techniques to optimise and forecast the prices with high precision

Following were the primary steps in our research.

·      Explanatory data analysis (EDA)

·      Feature Selection

·      Modelling

**Research Questions -** In order to help buyers and sellers make better choices, the research study's goal is to pinpoint the variables that affect the price of real estate properties. The following research questions have been created to achieve the study's goals:

1. Which machine-learning algorithm works better and produces the most accurate results for predicting home prices? And why?
2. What variables have influenced property values over time?

**Results -** The data is unquestionably the most important component of any machine learning activity, one that requires special attention. The data's origin, formatting, consistency, presence of outliers, and other factors will all have a significant impact on the outcomes. Many concerns must be addressed at this stage in order to ensure the effectiveness and accuracy of the learning algorithm.

We will use visuals to explore the data in this part. This will enable us to better comprehend the information and the connections among variables, enabling us to create a more accurate model.

There are many variables in our data, but the goal parameter is the one we should focus on(Target variable is salesprices). We must comprehend how it is distributed. Actively planning the violin plot again for target attribute comes first. The frequency is represented by the violin's breadth. As a result, the region around 300 and 400 includes fewer pieces of data than other regions if a violin is widest between these numbers, The plot demonstrates that the majority of home values decrease between 100,000 and 25,000,000. The 3 quartiles, Q1, Q2 (the median), and Q3, are indicated by the dashed lines

**Heat map -** We are interested in the relationships between the variables in the dataset and the relationships between the predictors and the goal variable. We want to examine the relationship between Lot Area and Sale Price, for instance: Do they have a positive association, rising and falling together? Do they have a negative association, where one increases when the other decreases? Or do they not have a connection?

A number between -1 and +1 is used to express correlation, with +1 signifying the strongest positive correlation, -1 the greatest negative relationship, and 0 signifying no connection.

We can observe that our dataset has a large number of associated factors. We observe a strong positive association between garage cars and garage area, this makes sense given that as garage area grows, so does its car capacity. We also observe a strong positive correlation between Gross Liveable Area and Total Rooms Above Ground, which is consistent with the idea as such living space above ground rises, so should the number of rooms above ground. In terms of negative correlation, we can observe that Bsmt Unf SF and BsmtFin SF 1 are inversely associated. This relationship stands to reason because having more incomplete space entails having less finished space. We also observe the negative relationship between Bsmt Unf SF and Bsmt Full Bath, which is reasonable as well.

**Overall quality Vs sale price -** The majority of homes get an overall performance around 5 and 7, according to Overall Qual, which accepts integer values from 1 and 10. To examine their association, we now present the scatter graph of SalePrice and Overall Qual: It is clear that the two are indeed positively associated; typically, as overall quality rises, so does the sale price. This supports the results of the heatmap above.

**Engineering Features –**

We have designed the attributes of our data in this part using the knowledge gained out from sections on exploratory data analysis.

* **New Derived Features Creation**

First, we found that the target variable SalePrice had a strong positive association with both Overall Qual and Gr Liv Area. This suggests that the last two characteristics are crucial in determining the sale price. Therefore, we shall transform these features into polynomials values: We will create a characteristic where values are the squared of the actual numbers for these attributes, as well as a characteristic where values are indeed the cubes of the original values. Additionally, we will develop a characteristic where values are the summation of our two features.

Additionally, we discovered that some predictive traits had a strong correlation with one another. We will eliminate one characteristic out of each pair of strongly correlated predictors in order to prevent the problem of multicollinearity. The very first pair is made up of Garage Cars and Garage Area, and the second is made up of Gr Liv Area and TotRms AbvGrd. We will delete the Garage Cars function from the initial pair, and the TotRms AbvGrd element from the second pair.

**Managing Ordinary Variables**

Our data contains some ordinal features. The Bsmt Cond feature, for instance, supports various unique combinations.

According to the dataset description, "Gd" stands for "Good," "TA" for "Typical," "Po" for "Poor," "Fa" for "Fair," and "Ex" for "Excellent." However, the issue is now that machine learning algorithms would consider this characteristic as though it were another categorical feature, not realising that it reflects a rating. Therefore, in order to resolve this problem, we will convert each potential value for this feature to a numeric value. " No Basement" will be mapped to 0," Po" to

**Comparison and analysis -** The graph show that the Support Vector Regression model has the largest MAE, 12974.93, while the XGBoost model has the shortest MAE, 12556.68. Following that, the errors for the Random Forest and Elastic Net models are comparable, coming in at 14506.46 and 14767.91, respectively. Next with similar errors are the Ridge and Neural Network models (15270.46 and 15656.38, respectively). The Decision Tree model follows with an MAE of 20873.95, followed by the K-Nearest Neighbours model with a 22780.14 error.

Therefore, in our study, XGBoost is the most effective model while K-Nearest Neighbours is the least effective model. We could see that the best model has approximately half the inaccuracy of worst model, indicating a substantial difference in MAE between two models

**Interpretation of Performance**

To evaluate and contrast algorithms, we picked the mean absolute error (MAE) as our quality indicator. MAE displays a simple number; it displays the average value of model uncertainties. For instance, the MAE of our XGBoost algorithm is 12556.68, which implies that on average, XGBoost will forecast a number that is 12556.68 more or lower than the real value. To appreciate how excellent this MAE is, we have to examine the data's range and spread. In our example, the numbers of the target attribute SalePrice, which includes the real property prices, must be seen. Let's look at SalePrice's violin plot, box plot, and histogram in our dataset.