# Report

**Data preprocessing:** A very important part of data analysis, this involves

1. importing useful libraries such as NumPy, pandas matplotlib, seaborn, and sklearn,

2. importing the dataset under consideration.

3. Data Cleaning which involves removing of duplicates, replacement of missing values or removal of affected rows if any.

4. Data conversion in case of categorical values

5. Scaling in case where values are either too large or far apart. StandardScaler in scikit library will be where necessary to readjust the distribution to have a mean value 0 and standard deviation of 1.

6. Splitting dataset into train and test sets. Selecting at random, the dataset can be divided in ratio usually with test set taking 0.2-0.33 of the datasets while the rest, train set is used in training the model. If the training data is usually much as more data tends to improve modelling.

## TASK 1:

Multiple Linear regression (MLR): a statistical modelling approach that could be used type for predictive analysis, like simple linear regression (SLR) is used to show relationship between values of a dataset. Unlike SLR, which shows between two variables, MLR tends to use more than one independent variable hence showing relationship between multiple independent variables against a targeted point or column of interest.

**Fitting MLR to the Training set and Predicting the Test set results:**

- The dataset is divided between test and train in the ratio 1 to 5 to afford more data point for train consideration the data size.
- varying the random state has little effect on the r2-score, best at around 10, hence it can be said that they are closely related
- Using scikit package, linear least-squares regression line is being fitted to the training data then used to predict target values for the test set.

**Data Analysis, Result Interpretation, Suggestion/Conclusion:**

- Random state does not really affect the outcome which shows features and targets are closely related.
- the coefficient shows the relationship of individual independent variable to the target. Only 7 of the independent variables are negatively related to the target i.e., price.
- Visuals shows inclusion of zipcode, longitude and latitude improves performance of the model. It can then be said that Increasing features bring about increase in performance of the model.
- An accuracy of 71% was observed using all the features repeating d same analyses without zipcode, longitude and latitude columns gives a result with regression score of 0.65.

# Task2:

Cluster analysis, or clustering, is an unsupervised machine learning approach that tends to group input data based on observed properties or features into natural group. It only interprets the input data and find natural groups or clusters in feature space.

Clustering has a large no. of applications spread across various domains. Some of the most popular applications of clustering are: Recommendation engines, Market segmentation, social network analysis, Search result grouping, medical imaging, Image segmentation, Anomaly detection etc.

**Cluster:**

A cluster may then be referred to an area with density of a particular feature or group of properties with data points closer to a central point than those in another cluster

**Clustering Algorithms:** clustering algorithms use similarity or distance measures between data points in the feature space in an attempt to show regions with dense observations; data scaling is often used to ensure data points are spread around a central point with mean of 0 and standard deviation of 1.

**K-means:** For the purpose of this analysis K-means, a popular centroid-based algorithm will be used. it is best used on smaller data sets because it iterates over all of the data points to classify them.

**Analysis:**

1. Specify the desired number of clusters K: Let us choose k=11
2. Scree plot is a plot which can be used to get the elbow point which is used to break the dataset for optimal classification. The K-means is determined from scree plotx, **NB:** K-means is an integer and is selected as the point where further changes down the graph is almost linear.
3. Compute cluster centroids(center of cluster): The centroid of data points are the yellow cluster as shown in the visuals.
4. Increasing the features does not really improve the output as seen by varying the selected features. **NB:** a couple of test variations of the features already commented out in the code.

**Conclusions:**

It is better to use the optimal number as suggested or deduced by the scree plot.

# Task3:

**Logistic regression (LR):** a statistical method similar to linear regression that tends to find an equation that predicts an outcome for a binary variable.

**Naïve Bayes algorithm:** is a supervised learning algorithm, which is based on Gaussian normal distribution and supports continuous data and used for solving classification problems. It tends to predict the probability of different class based on various attributes.

**Neural networks:** a deep learning approach which tends to simulating the way human brain works. it can be used for classification or regression.

A typical Neural network comprises of:

**input layer:** where data is being feed into the model

**hidden layer(s):** Any layer(s) which can be found between the input and output layers.

**output layer:** which is the result of our modelling or algorithm computations.

**Deep learning:** is a technique in which the neural network is left to figure out important features instead of applying feature engineering techniques.

**Synapses:** take a value from their input, multiply it by a specific weight, then output the value.

**Neurons:** being more complicated, they add together the outputs of all their synapses, and apply an activation function.

**activation functions:** they allow neural nets to model complex non-linear patterns, that simpler models may miss.

**Gradient descent:** It involves the use of the derivative or slope to find the direction and the rate to update the parameters.

**Confusion matrix:** is a matrix (table), usually 2x2 that can be used to measure the performance of a machine learning algorithm. it shows false positive, true positive, true negative and false negative. It helps to see how well a model predicts the values of the test dataset.

In neural net visuals, circles connote neurons while lines synapses. the process is very similar: initializing with random weights and bias vectors, the model makes a prediction, compare it to the desired output, and adjust the vectors, i.e., weights and bias to predict more accurately the next time. The process continues until the difference between the prediction and the correct targets is minimal. In the process of training the neural network, you first assess the error and then adjust the weights accordingly. To adjust the weights, gradient descent and backpropagation algorithms could be put to use.

# Comparison and Conclusion:

- It can be observed that increasing features increases the accuracy of the **Naïve Bayes** model as seen in the test case 2 till an upper limit is reached as observed with the case 3.
- Neural network gives best result compare to the rest followed by Logistic Regression using its random state property gives a better result than Naive Bayes.
- It can be observed that the at some point increasing hidden layers in a neural network hardly improves the system but takes time to complete
- The neural network is a really cool approach having to improve itself based on corrections while iterating make the system less rigid on decision.