Boston Housing Prediction Analysis  
Machain Learning[A]  
Department: CSE

Name: Mohammed Afridul Haque   
ID: 18-38794-3Computer Science and Engineering.  
American International University-Bangladesh.   
Dhaka, Bangladesh.   
18-38794-3@student.aiub.edu

Name: Jarin,Tasnim  
ID:19-39394-1Computer Science and Engineering.  
American International University-Bangladesh.   
Dhaka, Bangladesh.   
19-39394-1@student.aiub.edu

Name: Al-Amin Rasel   
ID:18-38622-2  
Computer Science and Engineering.  
American International University-Bangladesh.   
Dhaka, Bangladesh.   
18-38622-2@student.aiub.edu

*Abstract*— Every year, house prices rise. we tried to help people predict house prices by using machine learning. Our built models assist consumers in finding a home that meets their needs. The proposed work makes use of the house's properties or features, such as the number of bedrooms available, longitude, total bedrooms, and so on, as well as the house's location. House price prediction model is built for the city of Boston. The work uses Scikit-Learn and Machine Learning Others Tool to implement linear regression, DecisionTreeResgrassion , SVM.

*Keywords:*Machine learning for regression, housing dataset, Property stores, house price prediction, housing property value.

*Introduction*

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. We applied Linear Regression model. Regression in machine learning consists of mathematical methods that allow data scientists to predict a continuous outcome (y) based on the value of one or more predictor variables (x). Linear regression is probably the most popular form of regression analysis because of its ease-of-use in predicting and forecasting. The dataset for this project originates from the UCI Machine Learning Repository. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. Using Linear Regression model and dataset of housing price of new Boston, build a machine leaning model to predict median housing price. The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts.

***Related work***

The value of a property is determined by the infrastructure amenities that surround it. The scopes for obtaining clients for a few writers came with diverse technologies. Raghunandhan [1] mentioned the fundamentals of how data mining works, as well as supporting algorithms for prediction The easiest way to anticipate the price of a house is to employ a machine learning system. Often the location's environmental conditions decide what kind of price we can expect for different types of houses, Manjula [2] presents various important features to use when forecasting property prices with good precision using a regression model. A. Varma [3] designed a system that is a Google maps-based method that utilised real-time neighborhood data to obtain precise real-world valuations. Researchers discovered that there are links between physical appearance and non-visual characteristics such as crime statistics, housing prices, population density, etc. of a city. For instance, “City Forensics: Using Visual Elements to Predict Non-Visual City Attributes” [4], uses visual attributes to predict the sale price of the property. Hujia Yu, Jiafu Wu (2014) [5] used classification and regression algorithms. To the findings, the living space square feet, roof content, and neighborhood have the most statistical significance in determining a home's selling price. And the PCA technique can also help with prediction analysis. Li Li and kai-Hsuan Chu (2017) [6] studied various algorithms such as Backpropagation neural network (BPN) and Radial basis functional (RBF) neural networks. To detect the macroeconomic analysis, RBF and BPN models are used to identify the difference between house price indexes such as Cathy and Sinny price indexes and intricate correlation functions. Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) [7] studied linear regression algorithms for prediction of the houses. The purpose of this paper is to forecast the most cost-effective real estate price for clients based on their budgets and goals. Future house pricing will be predicted by analyzing past market trends and price ranges.

***Proposed Dataset***

## We collected our dataset from https://www.kaggle.com/puxama/bostoncsv

## This dataset is Boston House Prices Dataset. It was collected in 1978 and has 506 entries with 14 attributes or features for homes from various suburbs in Boston.

## The machine will smoothen the data if its nursery. It will also remove duplicate data and get data that is missing in the dataset by resampling the data.

***Data set description:***

As discourse in the introduction, we used the data set from kugel named Boston housing price. This is a csv formatted file. The data set is distributed into 10 Attributes like longitude, latitude, housing\_median\_age,total\_rooms total\_bedrooms ,populationhouseholds,median\_income,median\_house\_value ,ocean\_proximity where total number of row or instance is 20640 , where , all the attribute without ocean\_preximity are numeric and ocean\_preximity is text

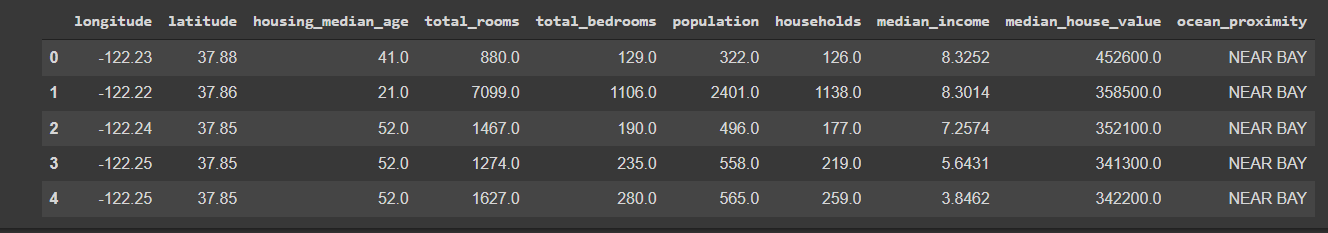
****

Fig 1: preview of dataset

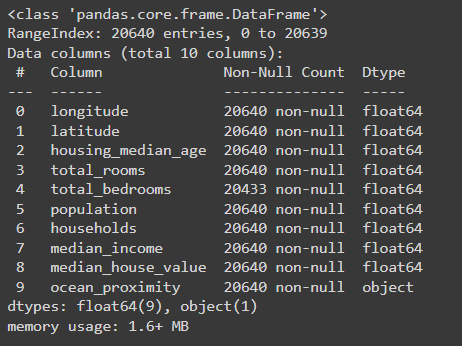
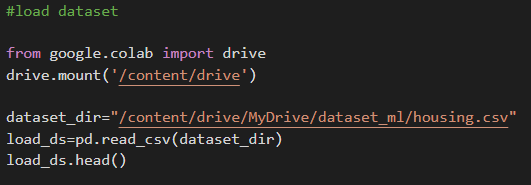
****

Fig 2: number of elements and datatype

***Load dataset***

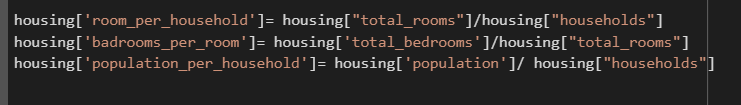
We discourse in the description we use google colab as platform to write the program because it provide free GPU . to train the dataset we need to load. to load the date, we use panda read\_csv function

****

***Dataset improvement:***

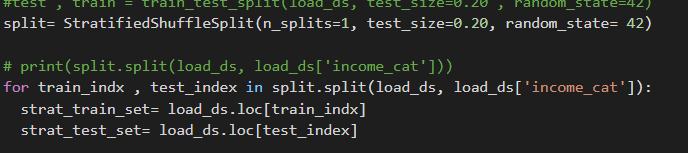
To improve the regression accuracy, we added few new attributes like

Room\_per\_household which is derive from total rooms / households, bedroom\_per\_room is derive from total\_bedrooms/ total\_rooms and population\_per\_household =population / households

****

***Split the dataset for train and test set***

In this section we divided the dataset into 80 % in train and 20 % for testing. The main target of this split is to prevent our model from overfitting. to split the date, we also used the StratifiedShuffleSplit from sklearn which not only split the data and also pick the date from randomly from shuffled data which makes the data more generalized

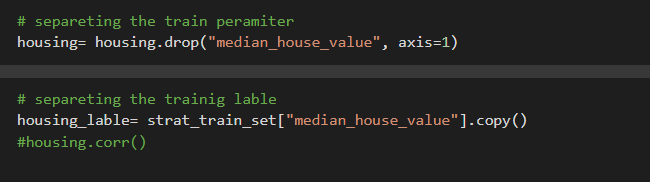
****

***Separating training and labels attributes***

As discourse earlier our model is regression model which will predict the medien\_house\_value

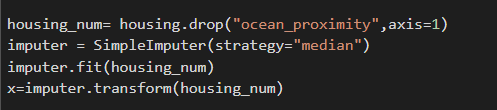
So in out model maiden\_housing\_price is the label attributes and other attributes are its training attribute

To do that I used pandas drop function to a copy of the actual training dataset(strat\_train\_set).housing\_lable hold only median\_house\_value

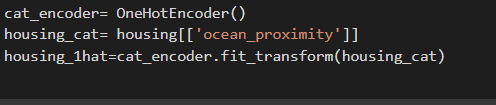
****

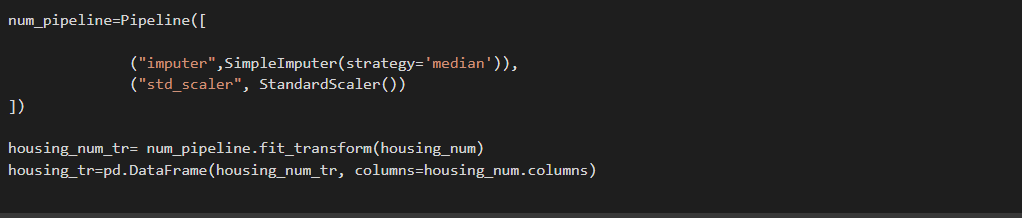
***Dataset transformation***

As we are using regression model missing value will play an important role in model accuracy, so we need to fill up those missing value on the other hand ocean\_preximity is a textual attribute but machine learning model don’t understand the textual value so, we need to convert the taxtual value into numeric value to handle those transformation we use SimpleImputer which will handle the missing vale replacing with a average value of corresponding attribute

****

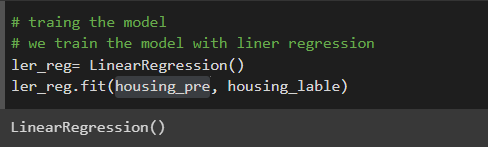
And OneHotEncoder transformer to conver the ocean\_proximity to one hot format

****

****

***Model selection and training the model***

Skein provide the prebuild liner regression and we can train our model with.fit() function by providing the trainable attribute and corresponding label

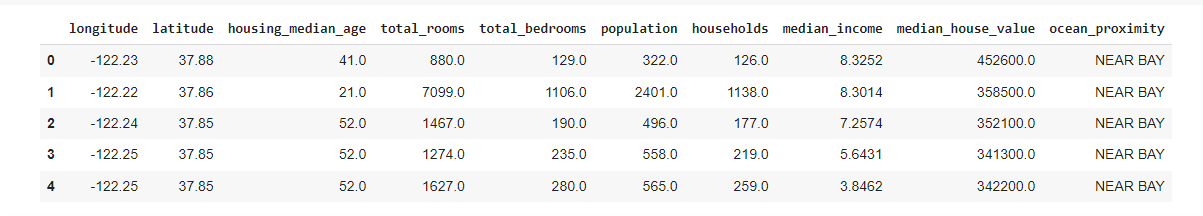
****

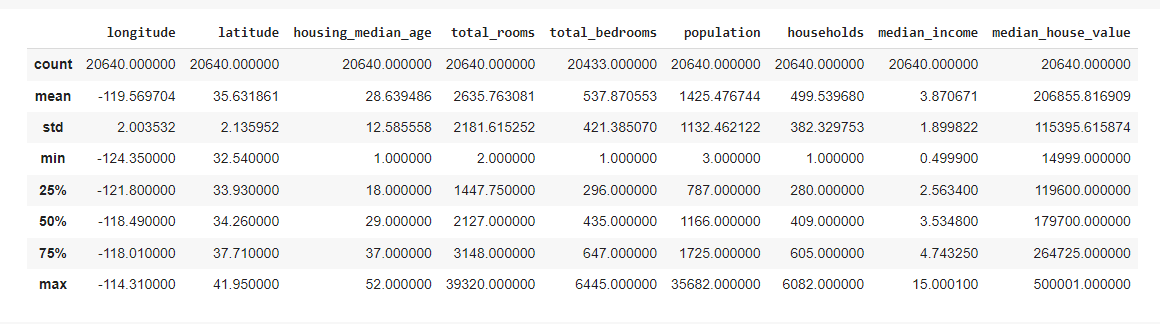
***Figures and Tables***

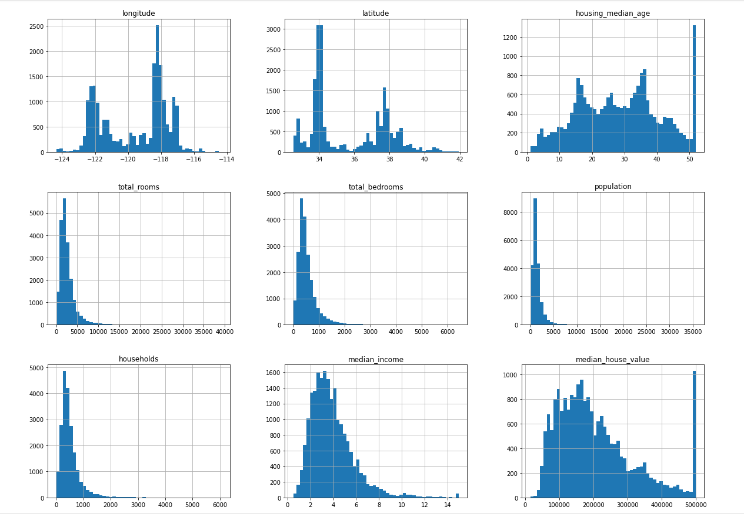
In the code cell below, you will need to implement the following:

•Calculate the minimum, maximum, mean, median, and standard deviation of 'MEDV', which is stored in prices.

•Store each calculation in their respective variable.







## **Sensitivity**

## An optimal model is not necessarily a robust model. Sometimes, a model is either too complex or too simple to sufficiently generalize to new data. Sometimes, a model could use a learning algorithm that is not appropriate for the structure of the data given. Other times, the data itself could be too noisy or contain too few samples to allow a model to adequately capture the target variable, the model is underfitted. Run the code cell below to run the fit\_model function ten times with different training and testing sets to see how the prediction for a specific client change with the data it's trained on.

As a reminder, we are using three features from the Boston housing dataset: 'RM', 'LSTAT', and 'PTRATIO'. For each data point (neighborhood):

•'RM' is the average number of rooms among homes in the neighborhood.

•'LSTAT' is the percentage of homeowners in the neighborhood considered "lower class" (working poor).

•'PTRATIO' is the ratio of students to teachers in primary and secondary schools in the neighborhood.

# ***Methodology***

# We will be use five algorithms.

# Linear Regression

# Decision Tree

# Random Forest

# Extra Tress

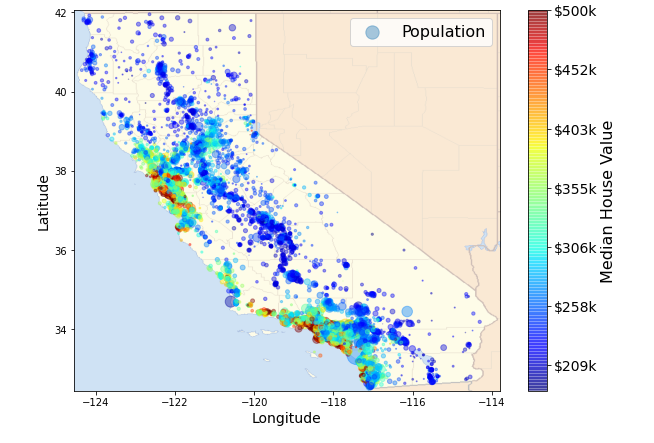
# XGBoost

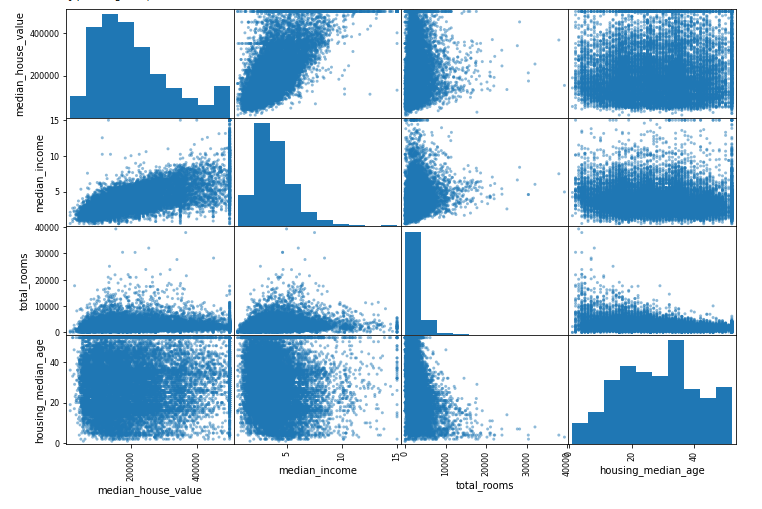
## **Objective**

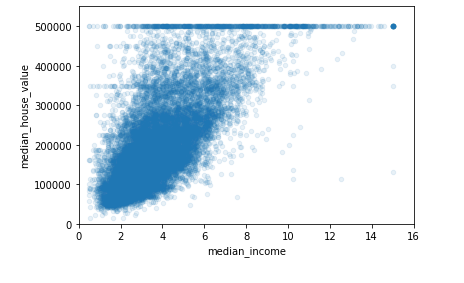
## In this decade, machine learning is growing at a faster rate. So, with this machine learning, we can fix many problems. House prices go up or down from time to time. Nowadays, most people face so many problems with renting a house. For this problem, people need a solution where they can get the predicted price of a house. So, we tried to help people predict house prices by using machine learning. This house price prediction can help people arrange a perfect time to purchase a house. Also, the people who tried to buy a house can understand whether they have received a fair price for a house. Financial technology applications that require a reasonable evaluation system for mortgage calculation and house auctions also use house price predictions. In current era of globalization, investment is a commercial activity that most people are interested in. There are a variety of items that are frequently utilized for investment. People can invest in real estate. Property investment has risen and fallen in response to demand and property sales. In here house price prediction will help most. We want to solve those problem by predicting the price of house. so, we tried to exploratory and analysis on the collected dataset. Then develop a machine learning that can determine the suitable rental price from training dataset using the algorithm. After that we tested the functionality and accuracy of the model using tested dataset.

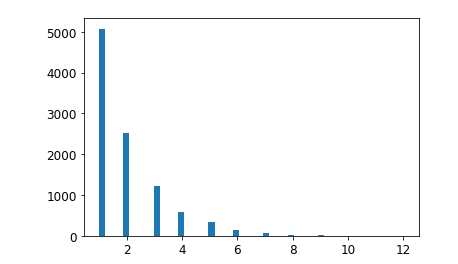
## **Data Exploration**

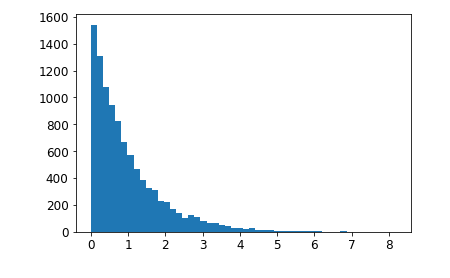
## Since the main goal of this project is to construct a working model which has the capability of predicting the value of houses, we will need to separate the dataset into features and the target variable. The features, 'RM', 'LSTAT', and 'PTRATIO', give us quantitative information about each data point. The target variable, 'MEDV', will be the variable we seek to predict. These are stored in features and prices, respectively.

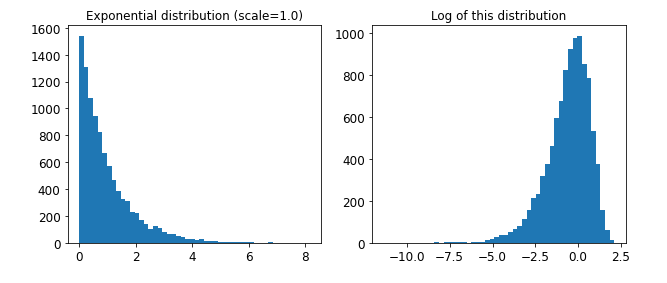




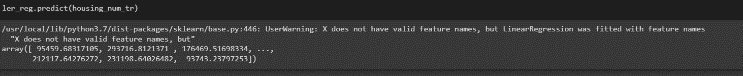




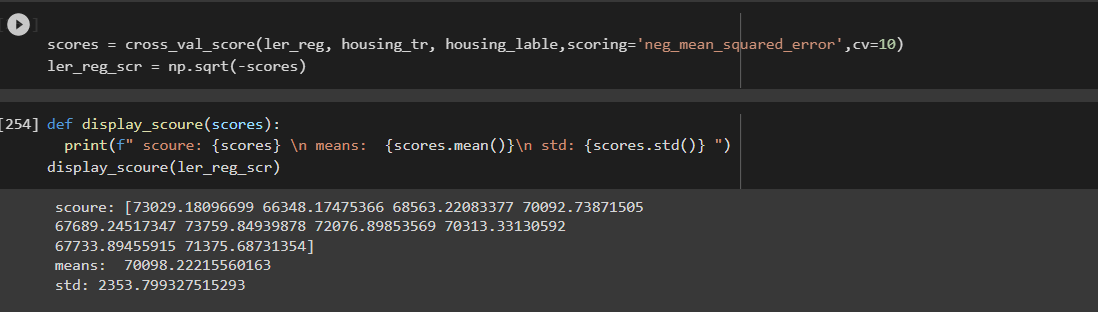




***Prediction***

****

***Model accuracy***

****

***Result and analysis***

In this project we tried various machine learning algorithms like linear regression, svm, decision tree with our prepare data set. At the fast stage of our project we tried to train the model with direct dataset and obtain batter accorece on (algorithm name) which is (scoure ) in below all the machine learning algorithm that we train are describe accordingly including its mean square error.

|  |  |  |
| --- | --- | --- |
| Algorithm | Mse | Mse |
| Linear regression | 68627.87 | 49438.66 |
| DecisionTreeResgrassion | 0.0 | 0.0 |
| SVM | 111095.066 | 1234113767.96 |
| RandomForestRegressor | 347848562.213 | 18650.699 |

***Cross validation***

We also trained out machine learning algorithm in cross validation to find out the better out come in the cross validation process we use cross validation fold =10. that meas we divided the the data in 10 part and and suffuld it in every new training and the scoring was based on neg\_mean\_squard\_erro

#### 

|  |  |
| --- | --- |
| *Algorithm* | Mse (best out of 10 training ) |
| DecisionTreeResgrassion | 79094.74 |
| Linear regression | 73997.08 |
| RandomForestRegressor | 2832232.213 |

***Fine-Tune***

Here, we used for find tuning for finding the best hyperparameter we used GridSearch cv and the list of parameters are astimetor 'bootstrap': [False], 'n\_estimators': [3, 10], 'max\_features': [2, 3, 4] [those parameters are used for randomforestregression] And train across 10 fold cross validation.After training we found best hyperparameters (max\_feature:8,n\_estimators:30)

Which is almost true for all the other machine learning algorithm

|  |  |
| --- | --- |
| Algorithm | Mse |
| RandomForestRegression | 63895.16 |
| DecisionTreeResgrassion | 60117.02 |
| Linear regression | 62448.32 |

After find tune RandomForestRegression outcome is worst and the DecisionTreeResgrassion perform well.Before using the find tune RandomForestRegression was giving bad mse and the DecisionTreeResgrassion gives best mse. DecisionTreeResgrassion mse 0 and the RandomForestRegression mse is maximum out of all the result.

#### **Conclusion**

##### Machine learning is a branch of artificial intelligence that is defined as a machine's ability to replicate intelligent human behavior. The proposed method will try to minimal human intervention. It is simple to find patterns in the housing field and how those patterns alter with price when utilizing a machine learning approach. We will used datasets and try to predict Housing price or estimation is a major trending project related to the field of Machine learning. We will research about the prices change of house. In this work we will compare previous data from there we can find the increase or decrease performance and accuracy.

##### ***References***

1. https://github.com/aswintechguy/Machine-Learning-Projects
2. https://github.com/aswintechguy/Machine-Learning-Projects/tree/master/Boston%20Housing%20Prediction%20Analysis%20-%20Regression
3. https://www.youtube.com/watch?v=WXXHPiX5LoQ

https://colab.research.google.com/drive/1YFSkEzoJXTm3GEn3CZ6MUcg2kJktAHHc?usp=sharing&fbclid=IwAR2gQPNpSERCd2UvGn7cmRY28FIegYSYyUSbPuUYJpNr7UFjCT4flglPaec#scrollTo=aYTq75K4RcVE