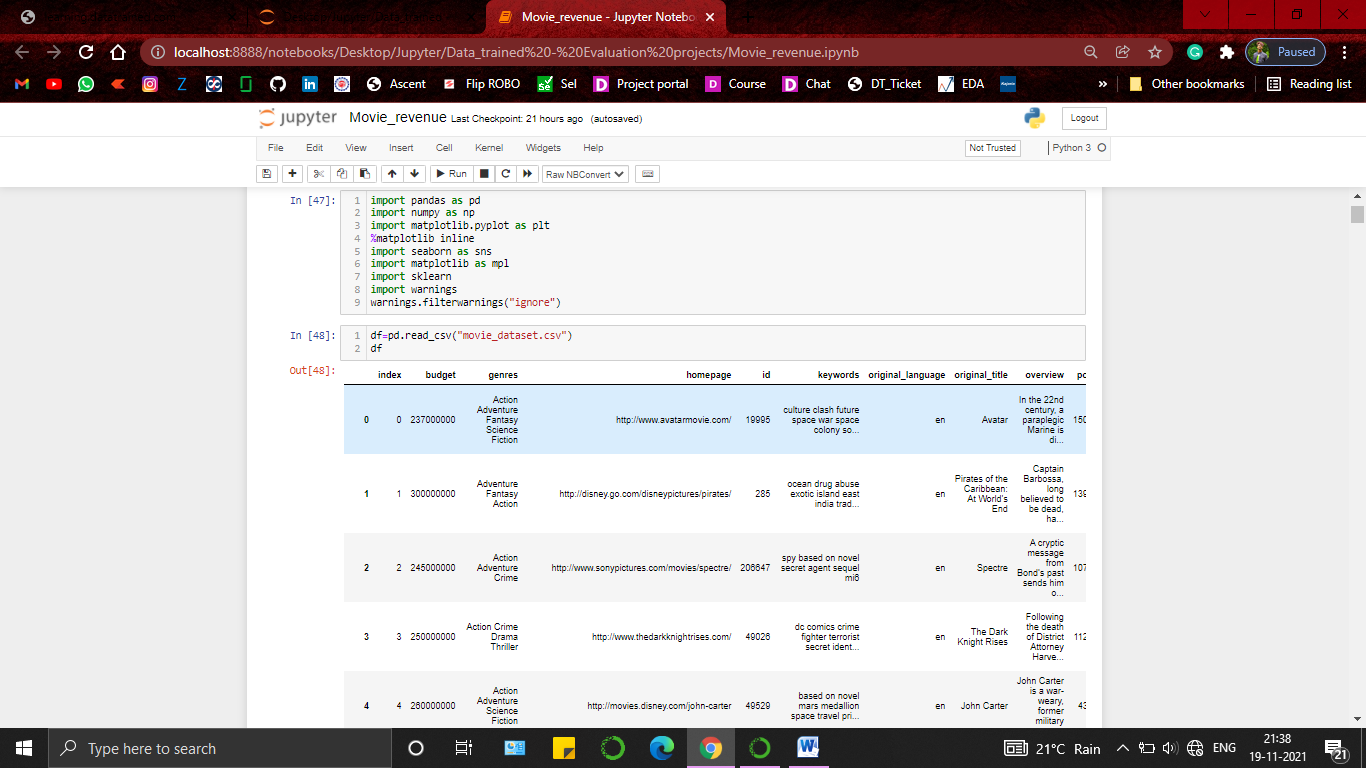
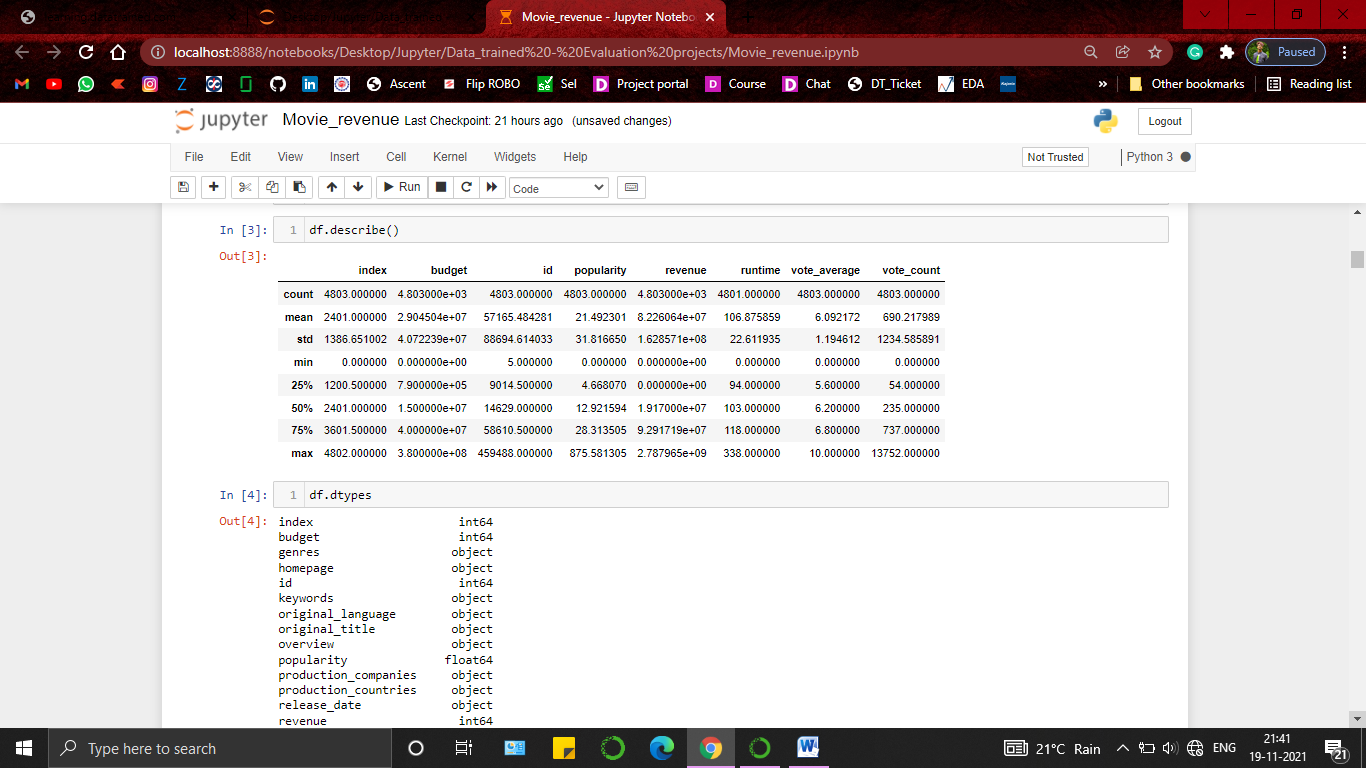
Prediction of Movie revenue collection

1. Problem Definition.

We have data set having movie details about most of the aspects of the movie like name , releasing year , genre , social media links as so on . Each variable affect the performance of a movie and its revenue collection using those variables we will try to evaluate the revenue a movie generates ..



1. Data Analysis.

Analysing the data gives us a better clarity about the available variables and also helps to figure out how to process it for training the model . like here We have a bunch of data in all int , float and object format. We will try to predict the revenue but not all variables are contributing towards the target variables . 

Some variables will not at all contribute a meaning full contribution towards the prediction . in this dataset precisely that is the index column as it’s a part of the dataset created during the data collection just to have a tracking of data .

variables like overviews , id , tagline will contribute minimally towards the target variable as they aren’t technically part of movie itself .

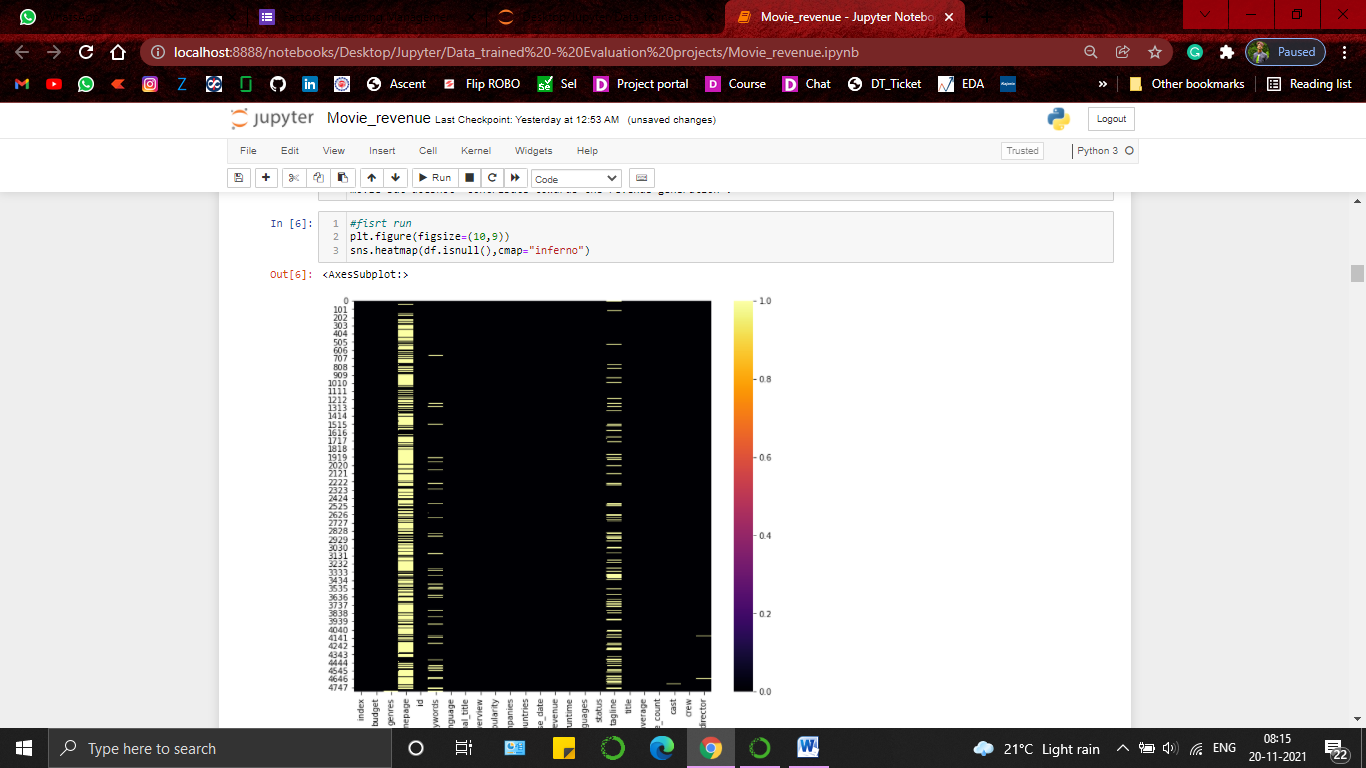
variables like Budget , cast , director , language spoken ,popularity , vote count and vote average etc. will impact the revenue highly because they are the variable that decides the quality and reach of the movie which is generally the key to earn higher revenue .

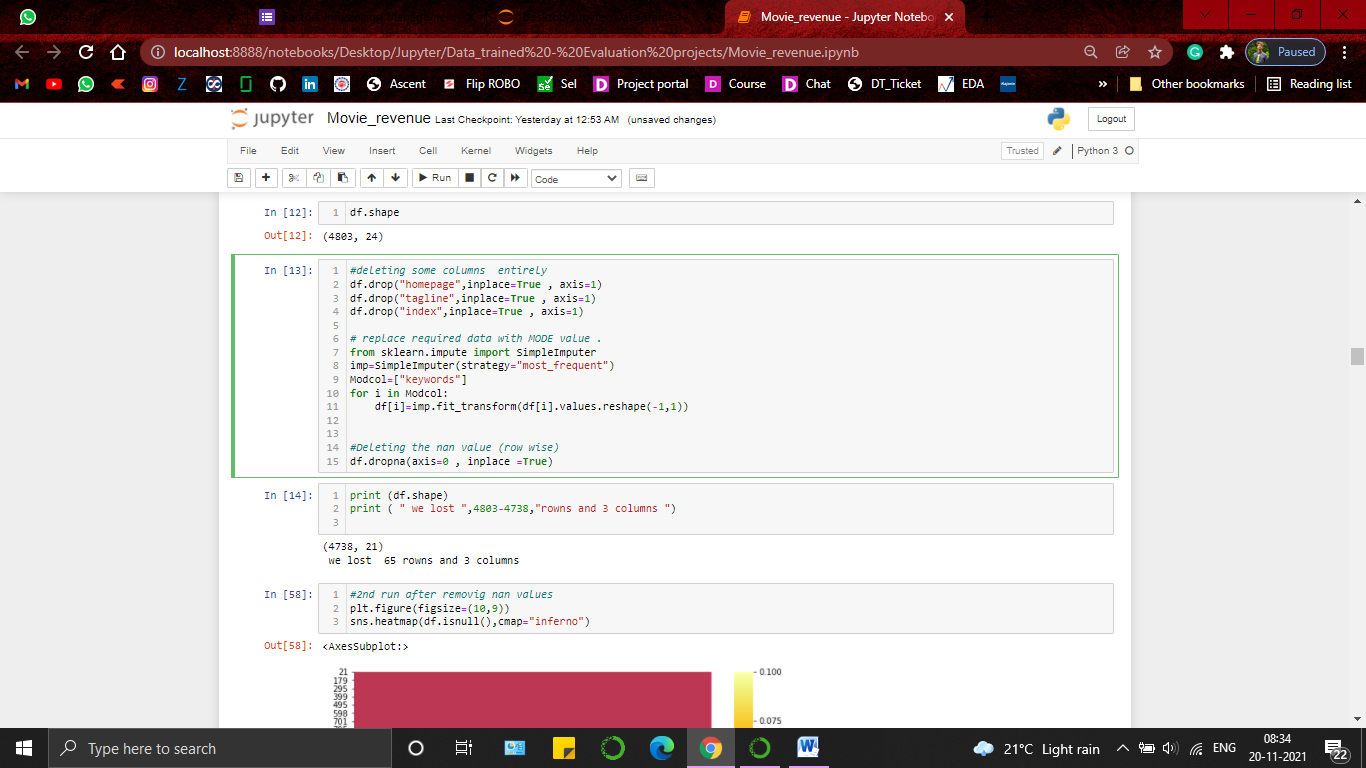
To predict the outcome we will use linear regression from sklearn and for the sake of diversity we will be using Ridge ,Lasso and Elastic net to evaluate which model will perform better to predict the outcome from this dataset

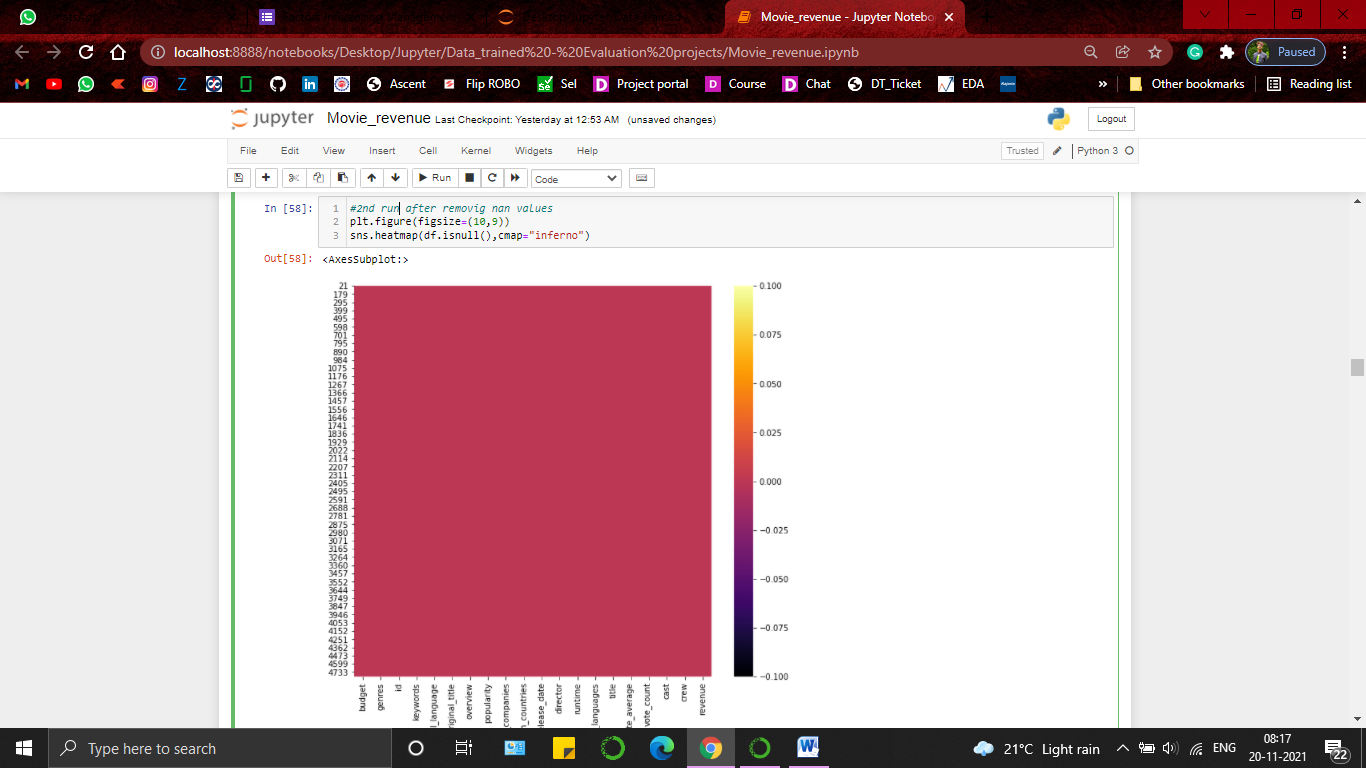
1. EDA and preprocessing

It was observed that data set is having missing / NaN values . I treated those missing values with considering the amount of its presence and the type of variable in which it is present . small amount of NaN values are deleted to keep the data pure , some NaN values were replace by Mode (most frequent) values which was done by keeping the randomness factor of variable in account . variables with higher (greater than 50%) number of NaN values are dropped .

I have reordered the dataset to keep the revenue at the end of the dataset , typically its done to have a better clarity about the predictor and target variable .

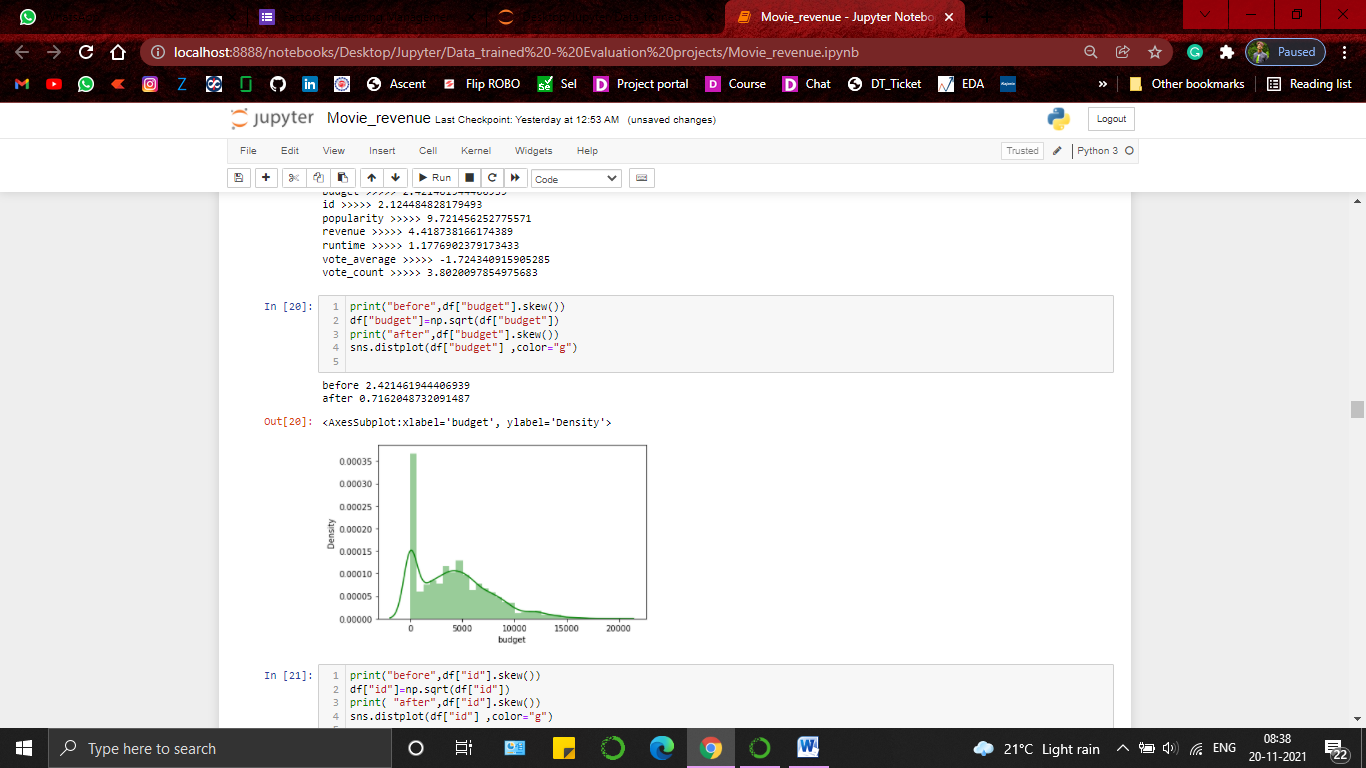




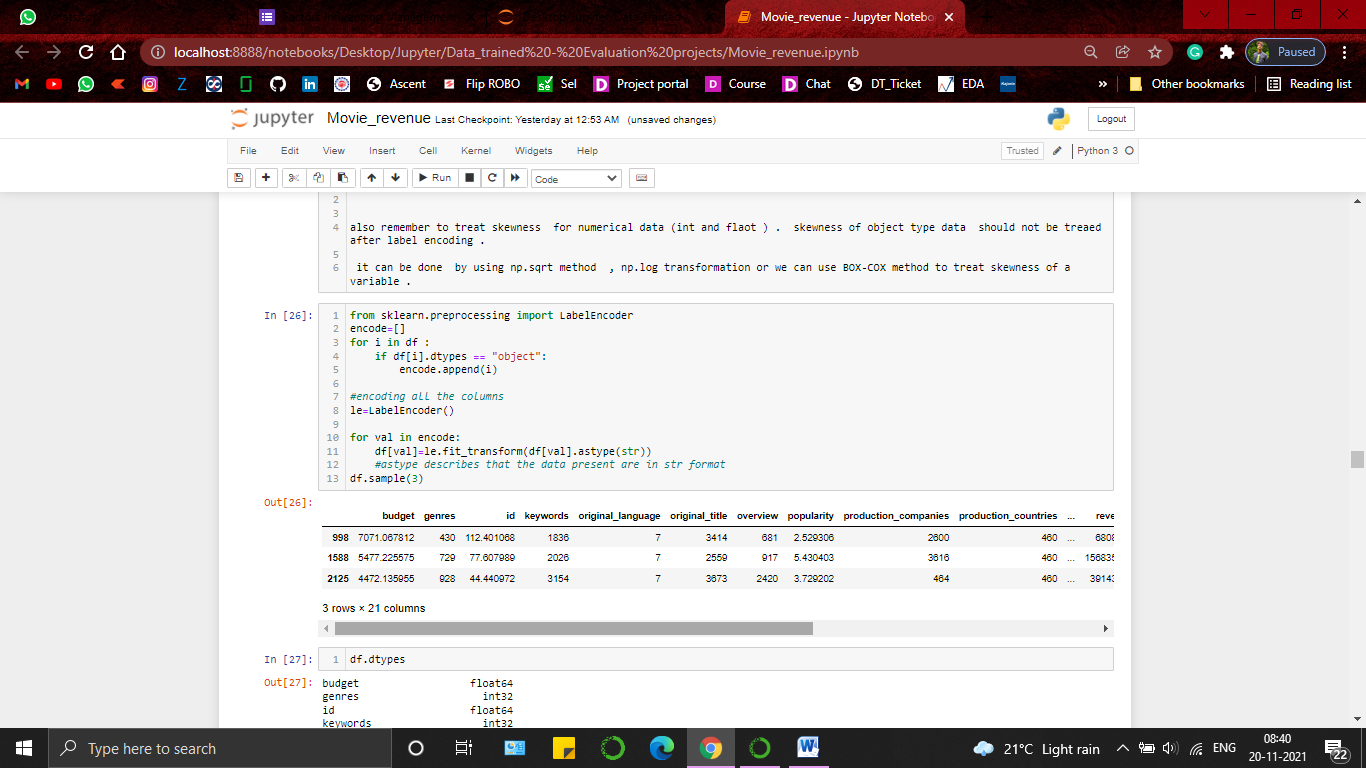


Checked for the skewness in the variables having int or float value . All the variables among them having more than 0.5 or less than 0.5 skewness are treated with numpy.squareroot method .

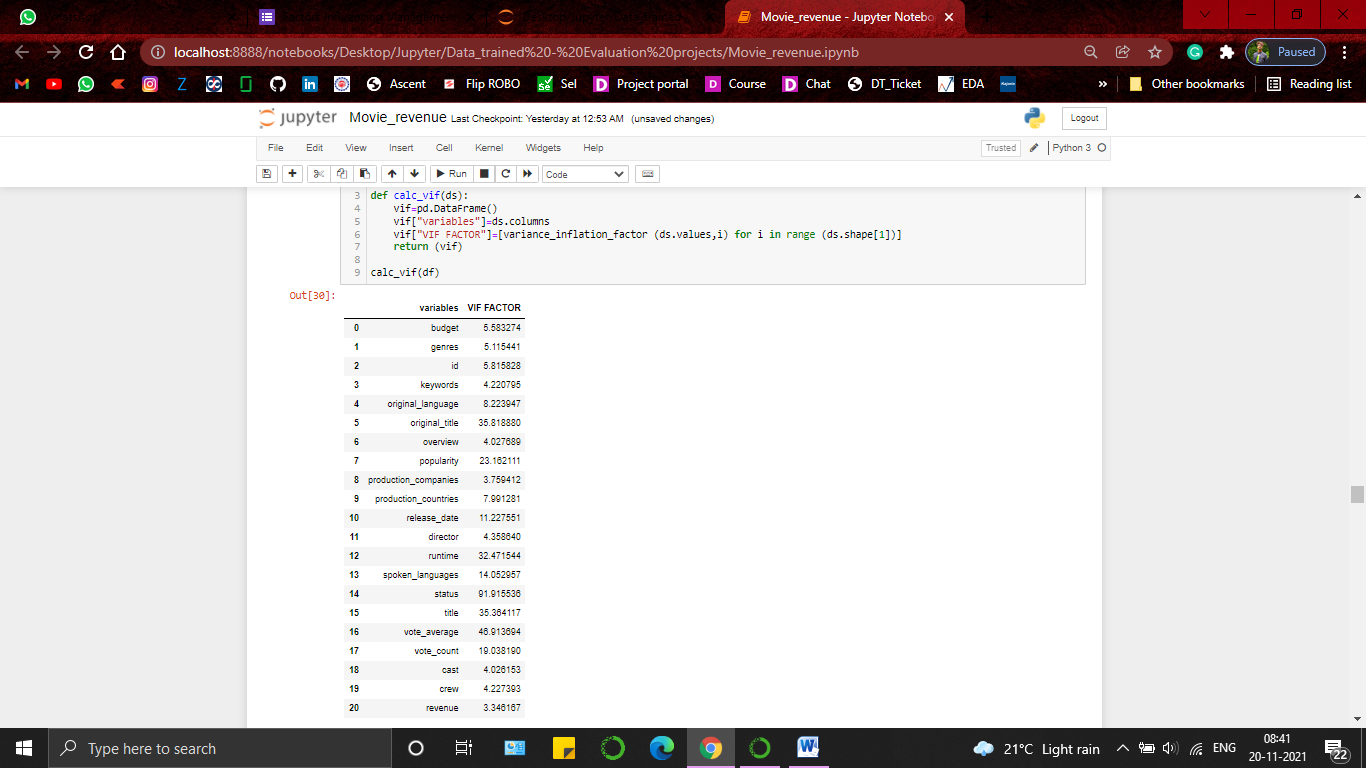




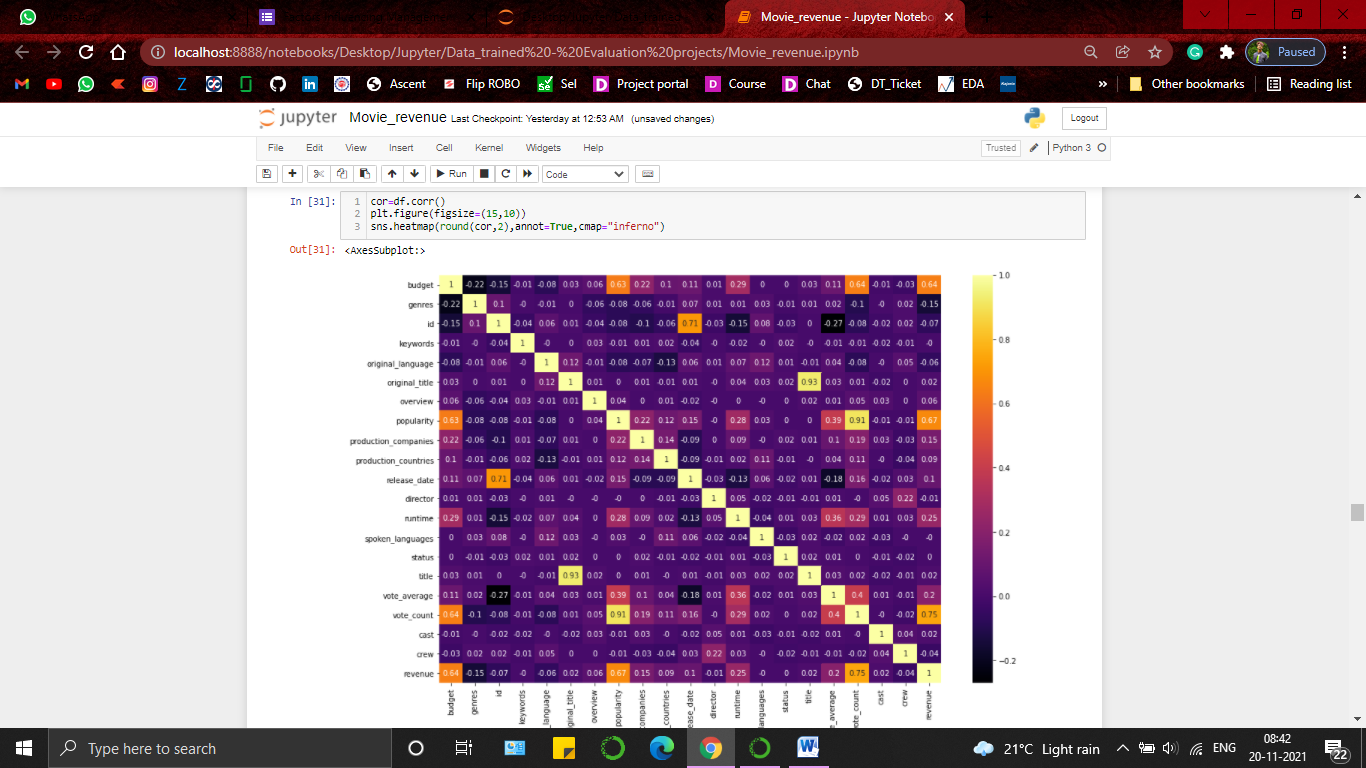
Categorical values were encoded . To encode , Label encoder is used as the variables are having more than 2 values .



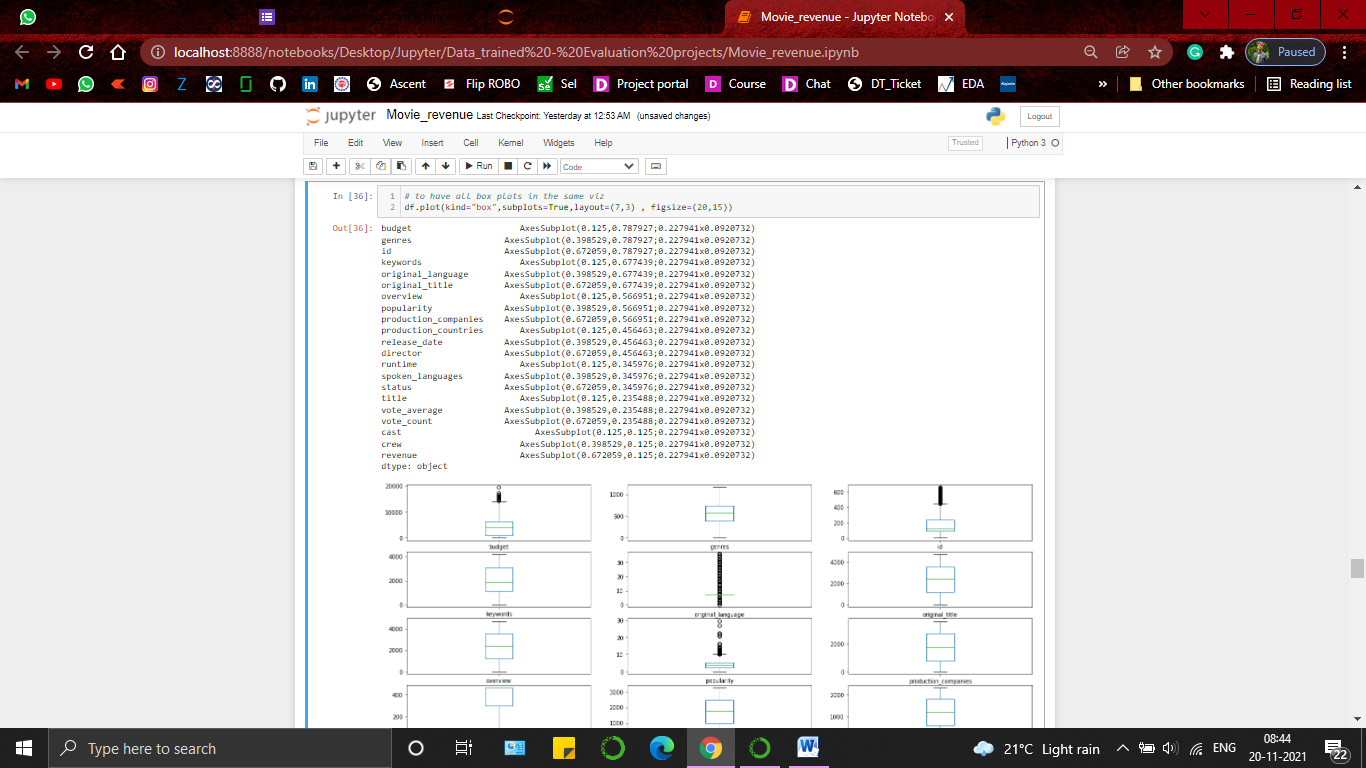
Variable influence factor ( VIF factor) is also check based on which we had option to drop some variables but to avoid data loss and keeping the legitimacy of data in account no variable is dropped after checking the VIF factor .

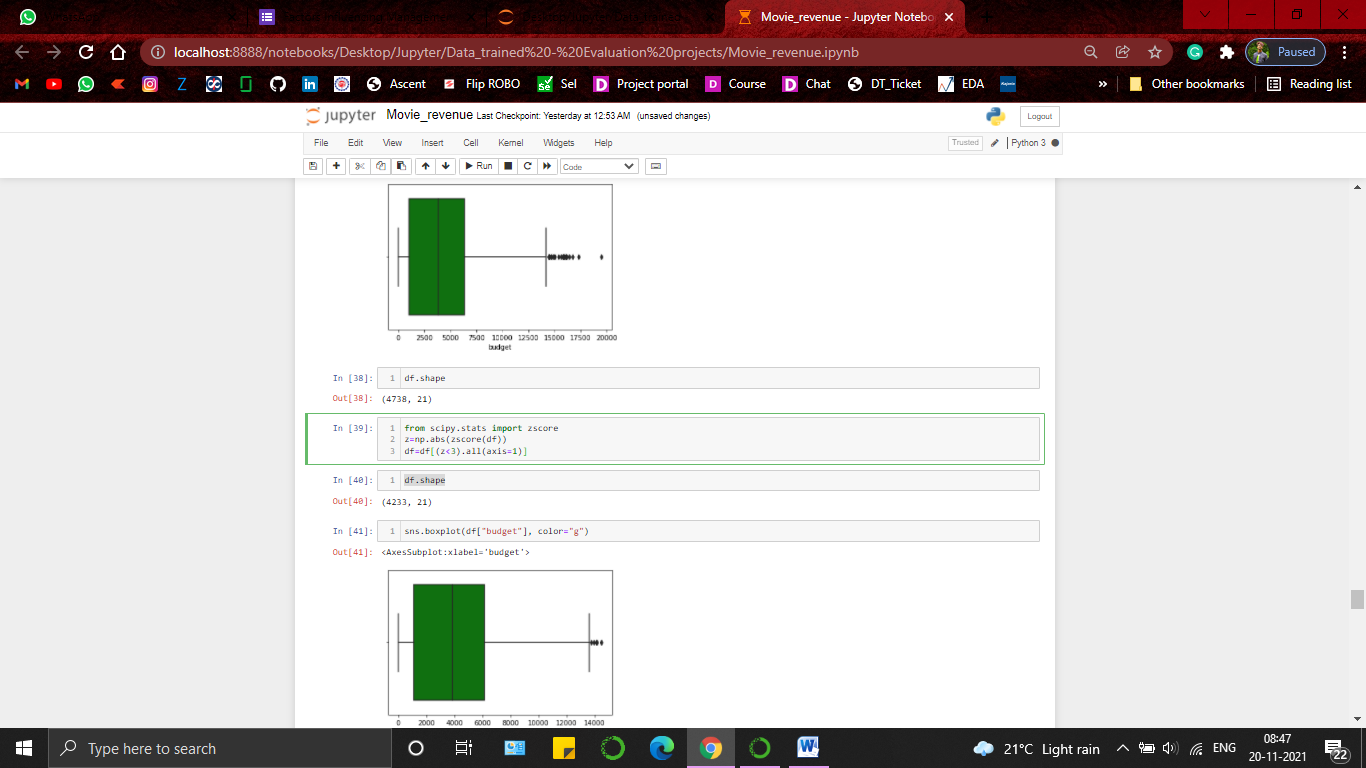


Correlation of variables were checked using .corr and heatmap(from seaborn) . found the most negative correlation towards the target variable is -0.15 so considering that as acceptable no change was done to the dataset .

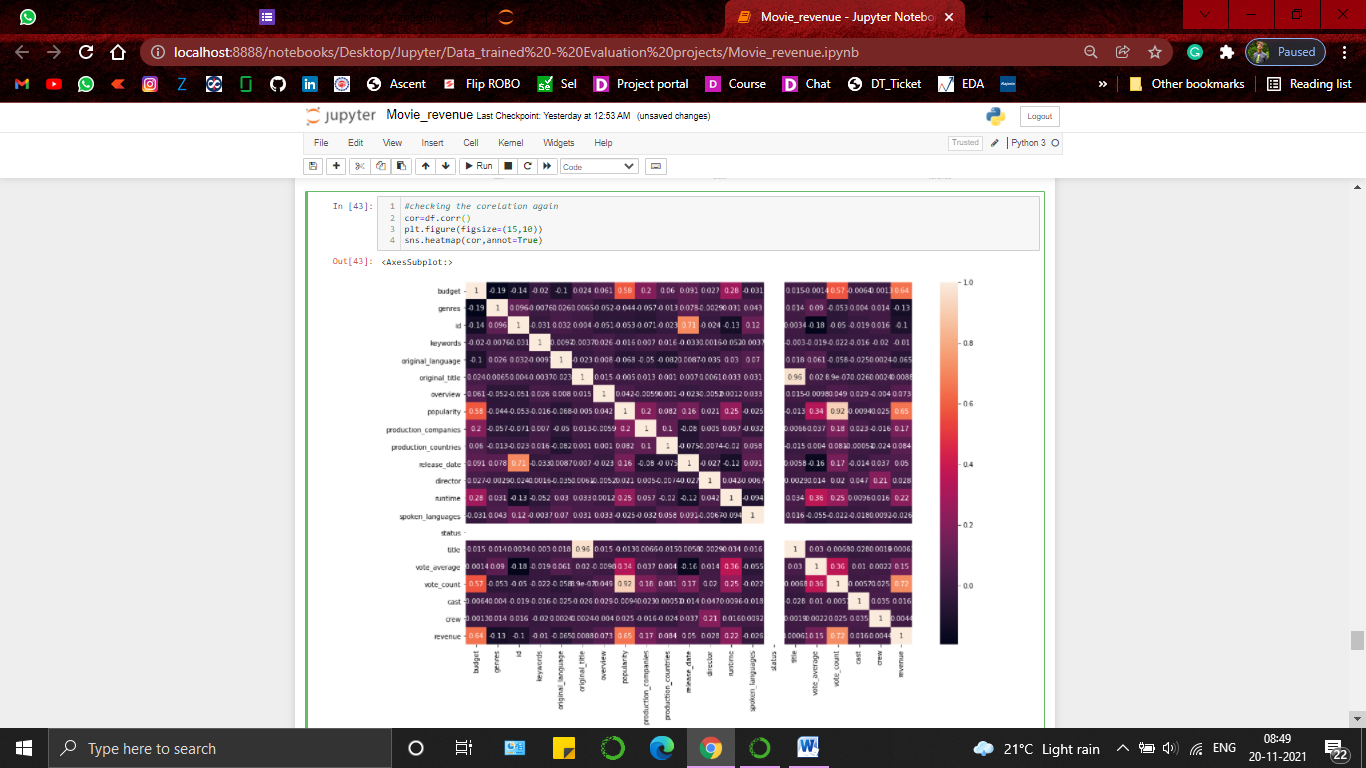


Outliers were spotted using boxplot and removed using ZSCORE method from Scipy.stats .



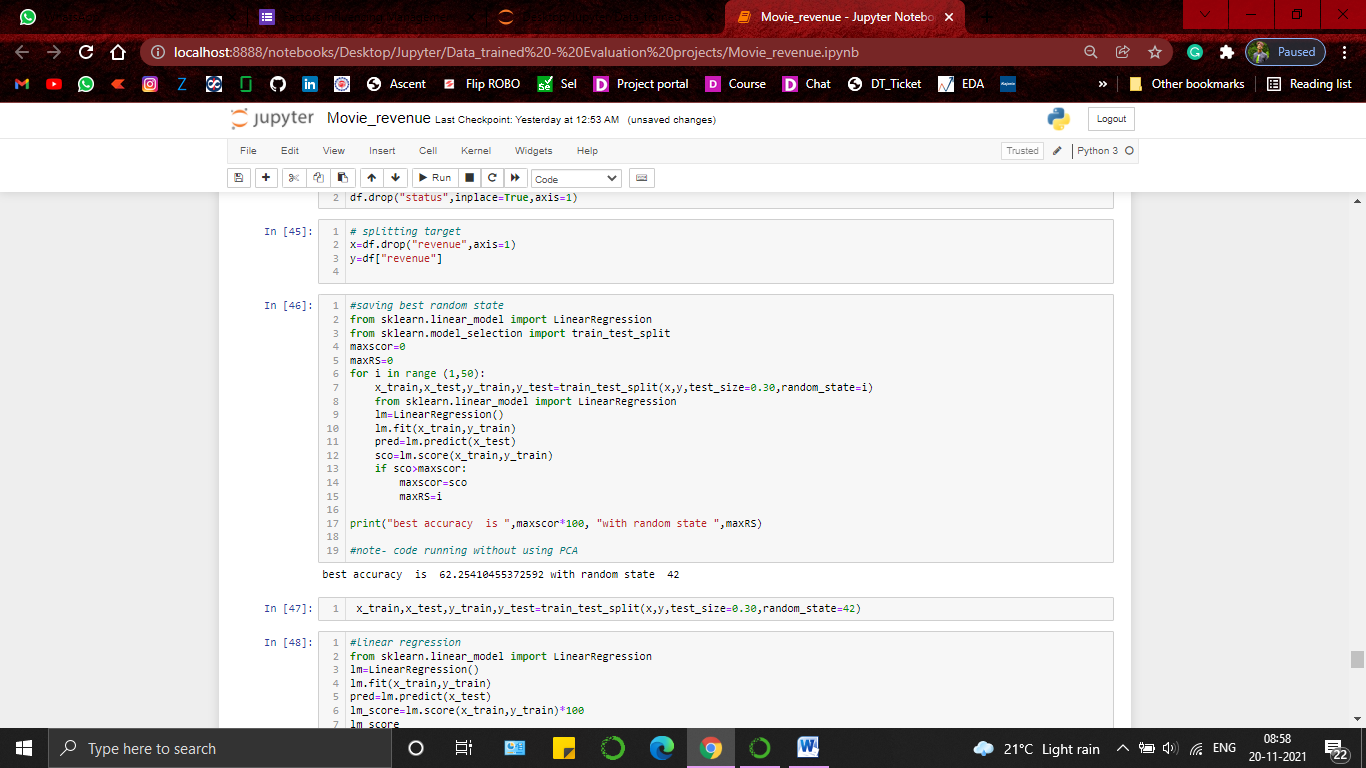


after removing the outliers the status variable is not contributing anything to any variable so dropping it

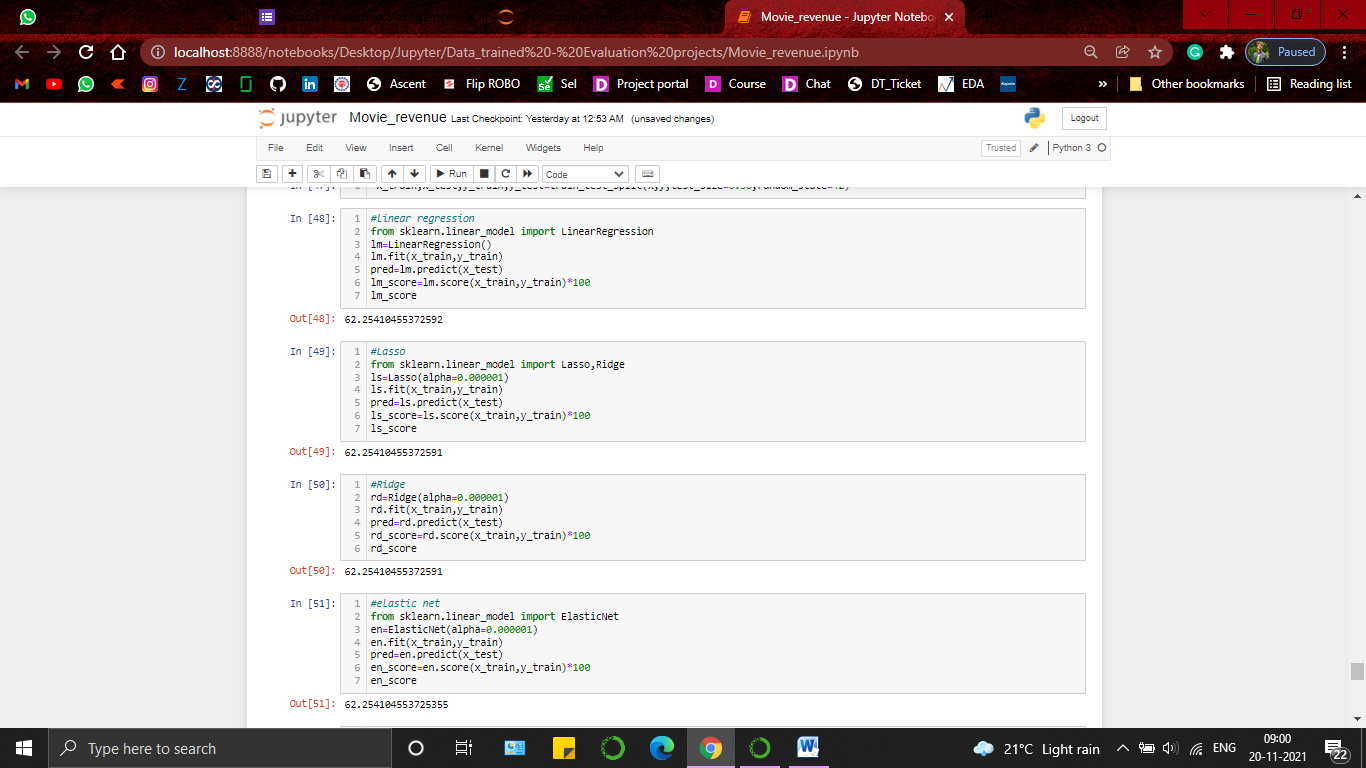


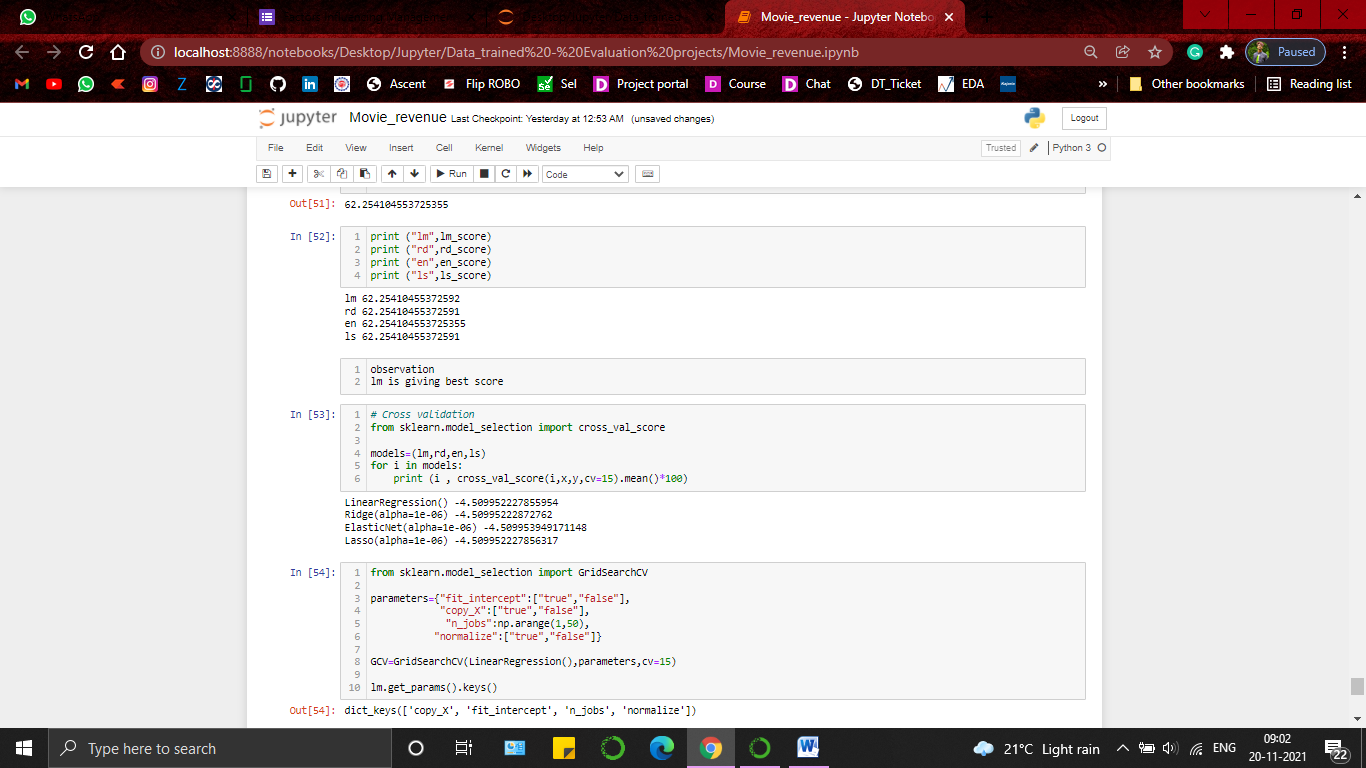
4.Building Machine Learning Models.

Tested best random state for train\_test\_split . Comes out to be 42 .so splitted the data into 70:30 ratio using 42 as the random state.

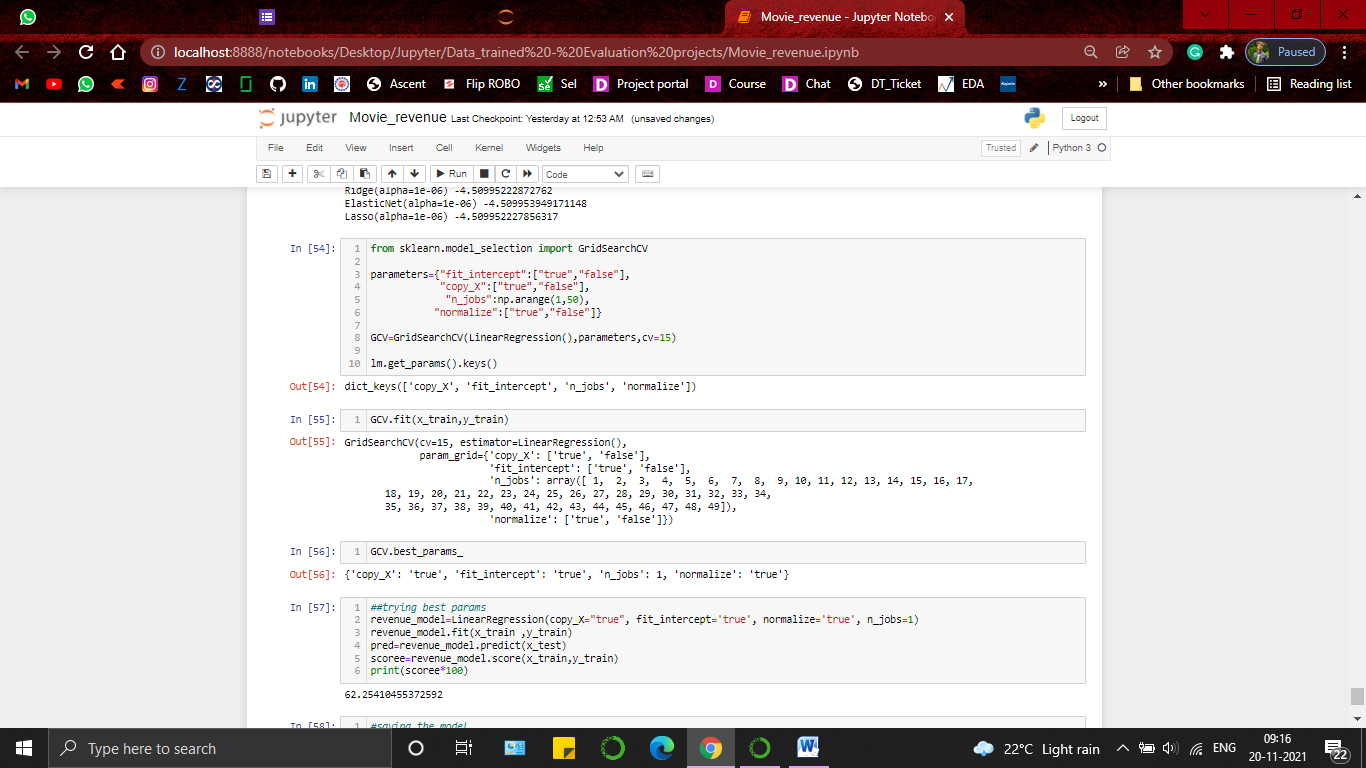


Trained Linear Regression module with the splitted data and as said for diversity , trained Lasso , Ridge and ElasticNet and measured the accuracy of the model prediction using accuracy score .

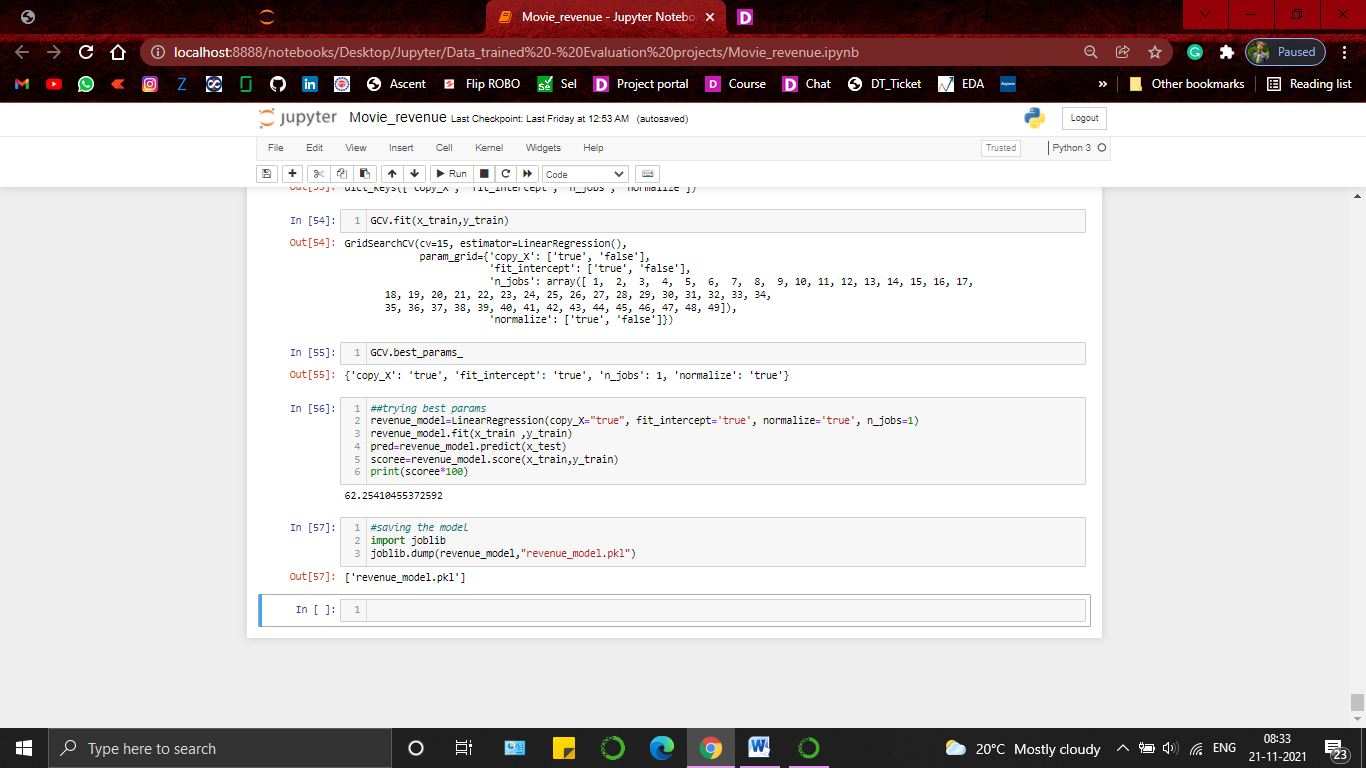


Used cross validation to evaluate the model performance 

Keeping the cross validation score and model accuracy score in consideration, we can say linear regression is giving the best result for this dataset so took linear regression as best model to go with and used grid search CV to get the best params to amplify the result .



5. conclusion.



Using the parameters that we got from grid search cv \_bestparams , I trained LinearRegression model which is predicting the outcomes with an accuracy of 62.25 percent .

Prepared by –Amit Patel .