BUSINESS ANALYST COURSE

**REGRESSION ANALYSIS**

CAR PRICE PREDICTION

AND

HR ANALYTICS

horizontal line

# Introduction

## What is Regression Analysis?

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the [causal effect relationship](https://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/) between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.



Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

## Why do we use Regression Analysis?

As mentioned above, regression analysis estimates the relationship between two or more variables. Let’s understand this with an easy example:

Let’s say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

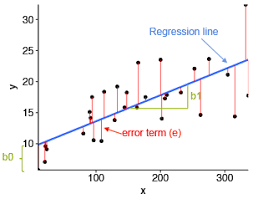
There are various kinds of regression techniques available to make predictions. In this study we have used linear regression and logistic regression.

### 1. Linear Regression

It is one of the most widely known modeling techniques. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be [continuous or discrete](https://en.wikipedia.org/wiki/Continuous_and_discrete_variables), and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation Y=a+b\*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).



The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

This task can be easily accomplished by Least Square Method. It is the most common method

We can evaluate the model performance using the metric R-square. To know more details about these metrics, you can read: Model Performance metrics [Part 1](https://www.analyticsvidhya.com/blog/2015/01/model-performance-metrics-classification/), [Part 2](https://www.analyticsvidhya.com/blog/2015/01/model-perform-part-2/) .

#### Important Points:

* There must be linear relationship between independent and dependent variables
* Multiple regression suffers from multicollinearity, autocorrelation, heteroskedasticity.
* Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.
* Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
* In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.

### 2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

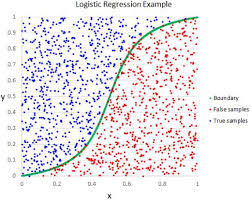
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is [logit](https://en.wikipedia.org/wiki/Logistic_function) function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).



#### Important Points:

* It is widely used for classification problems
* Logistic regression doesn’t require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
* To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
* It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
* The independent variables should not be correlated with each other i.e. no multi collinearity. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
* If the values of dependent variable is ordinal, then it is called as Ordinal logistic regression
* If dependent variable is multi class then it is known as Multinomial Logistic regression.

**How to select the right regression model?**

Life is usually simple, when you know only one or two techniques. One of the training institutes I know of tells their students – if the outcome is continuous – apply linear regression. If it is binary – use logistic regression! However, higher the number of options available at our disposal, more difficult it becomes to choose the right one. A similar case happens with regression models.

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:

1. Data exploration is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables
2. To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the [Mallow’s Cp](http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/goodness-of-fit-statistics/what-is-mallows-cp/) criterion. This essentially checks for possible bias in your model, by comparing the model with all possible submodels (or a careful selection of them).
3. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.
4. If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.
5. It’ll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.
6. Regression regularization methods (Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

# HR Analytics

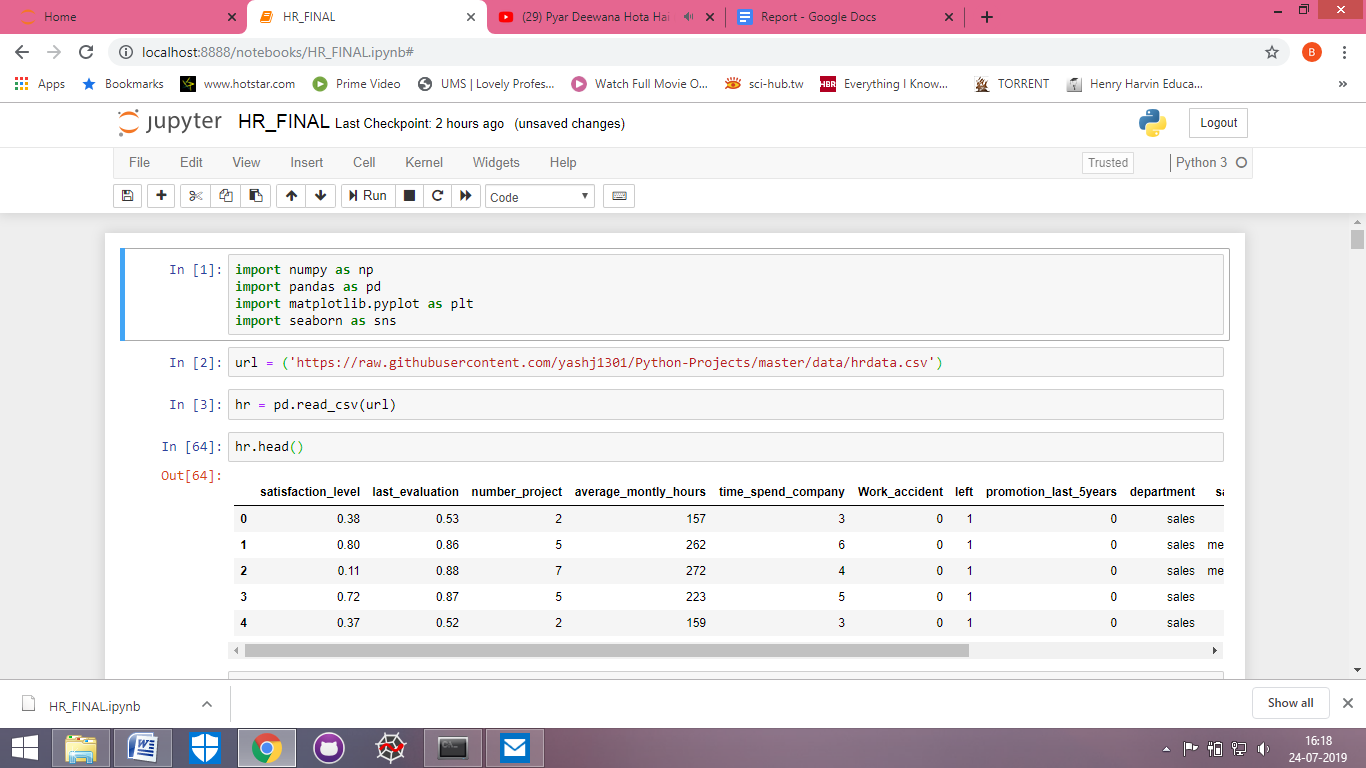
In this project I have given the role of HR manager in which i have to predict whether an employee will leave the company or not using various independent input factors provided to me. Data is of nearly 15000 employees. In this project I will be making various regression models and will see which model fits the best in the given data and will try to predict the outcome with the best possible accuracy.

## Components of Data Available:

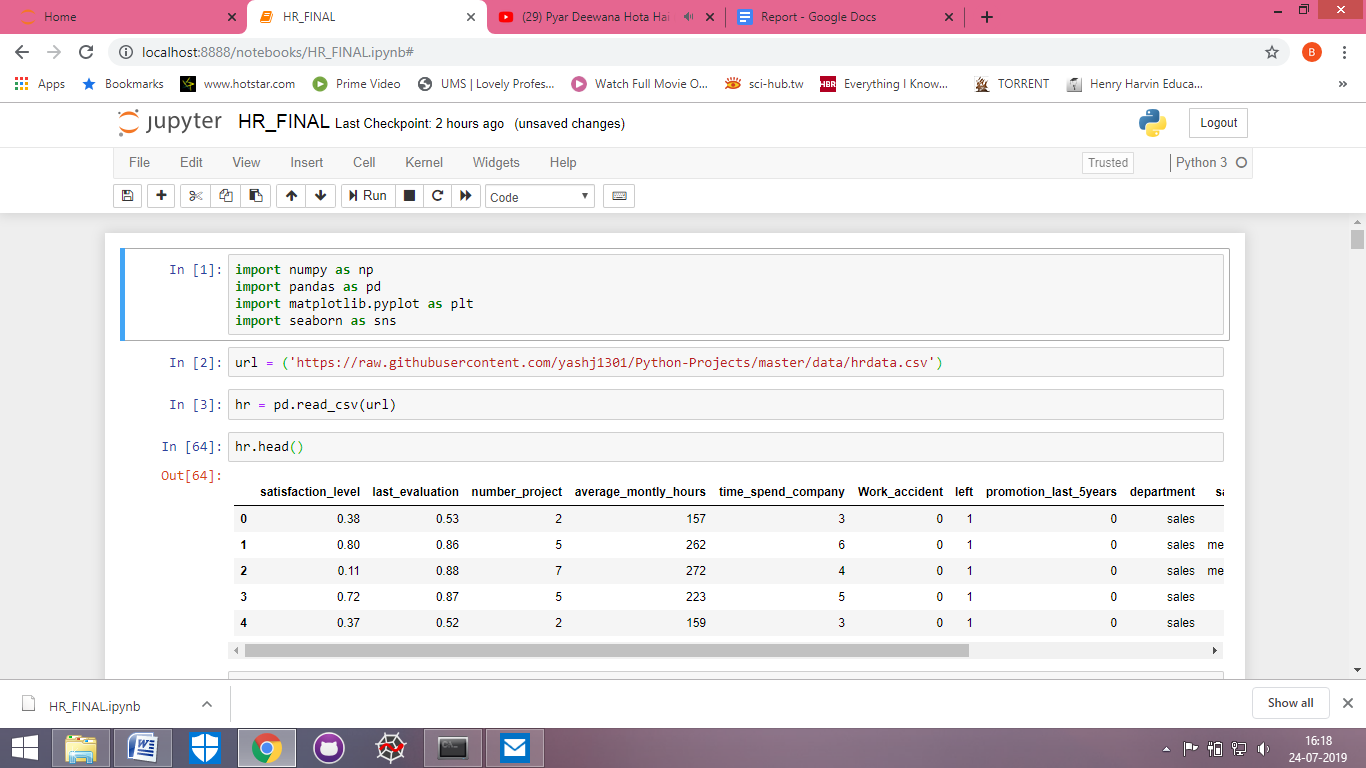
|  |  |
| --- | --- |
| sr.no. | Code Name |
| 1 | Satisfaction level |
| 2 | Last evaluation |
| 3 | Number of projects |
| 4 | Average monthly hours |
| 5 | Time spent in company |
| 6 | Work accident |
| 7 | Left |
| 8 | Promotion in last 5 years |
| 9 | Department |
| 10 | Salary |

## Steps Followed:

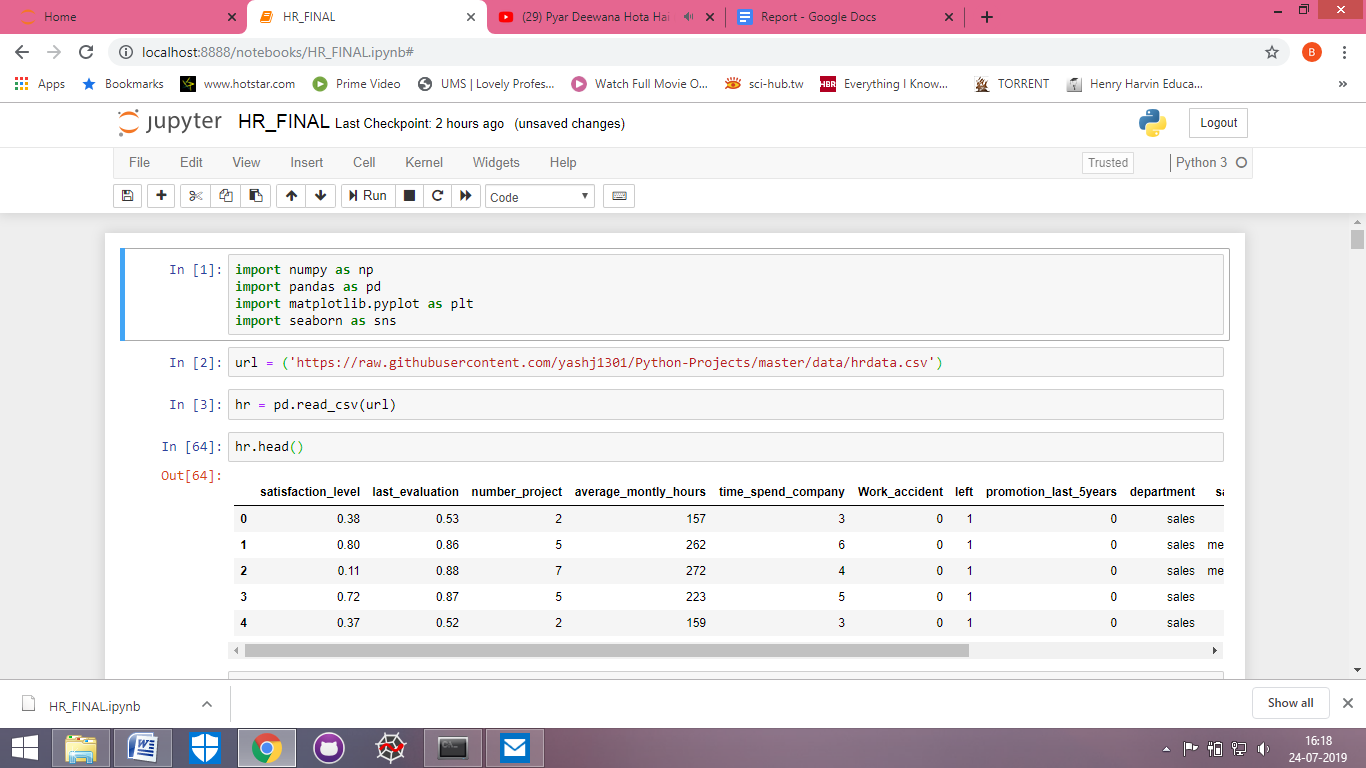
1. Firstly I have imported the required libraries which are pandas, numpy, matplotlib and seaborn.



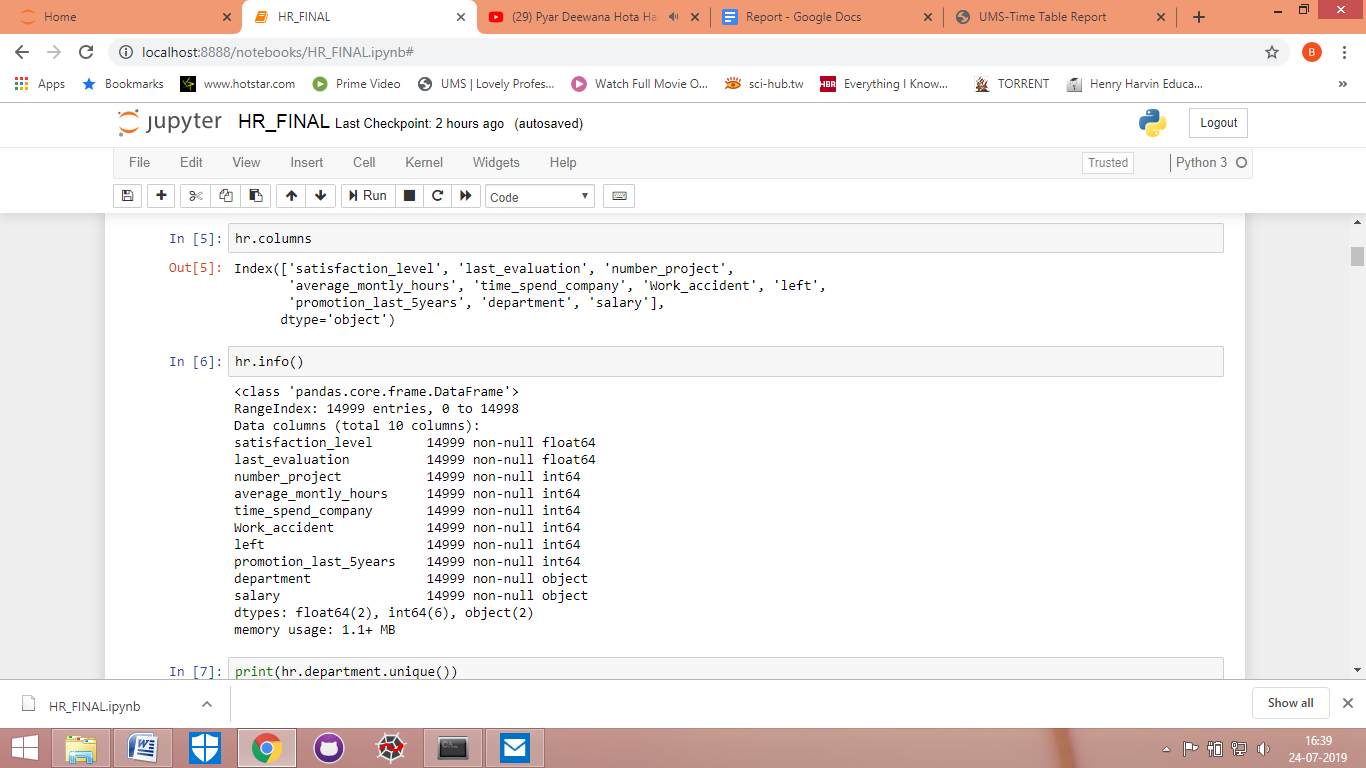
1. After importing the libraries, I have now imported the data set of HR analytics from a url and then named it as ‘hr’.



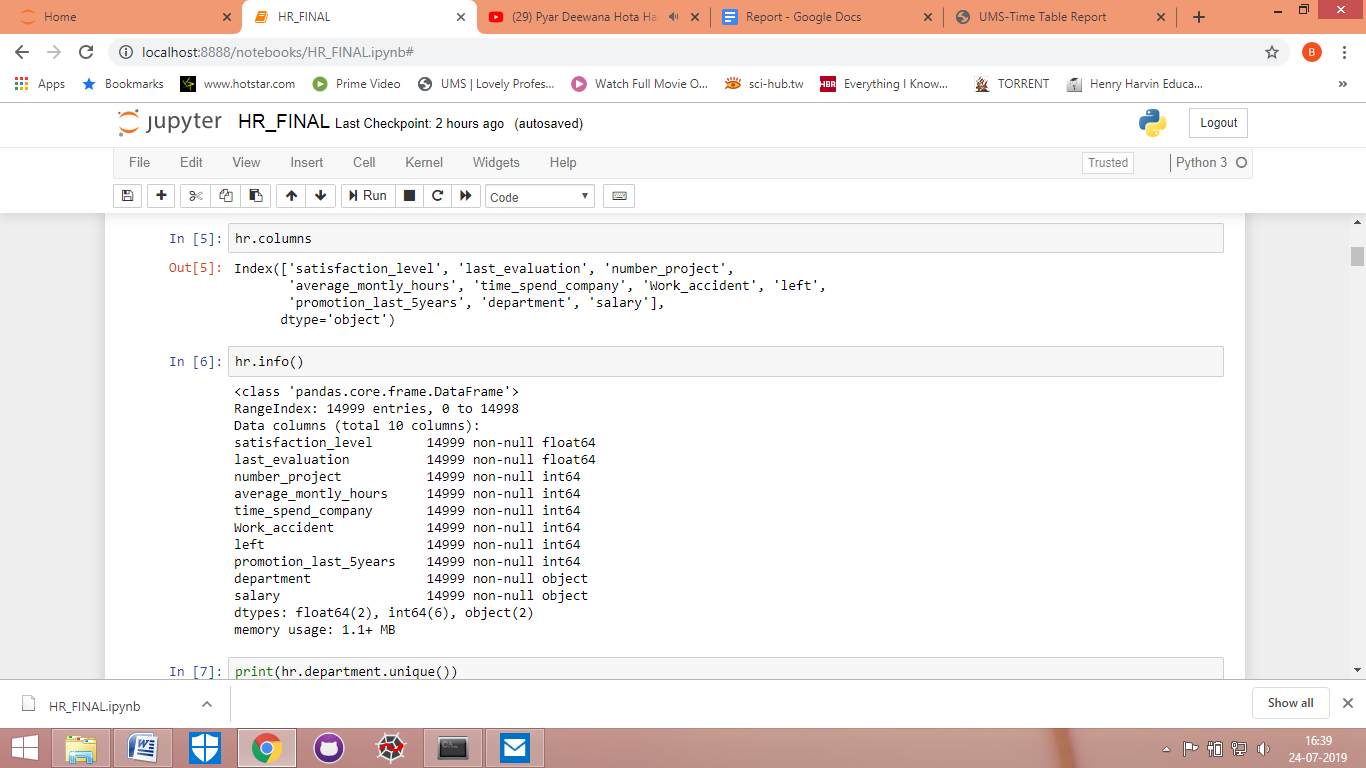
1. After importing and naming the data set, now I checked the first five rows od the data and all the column heads.



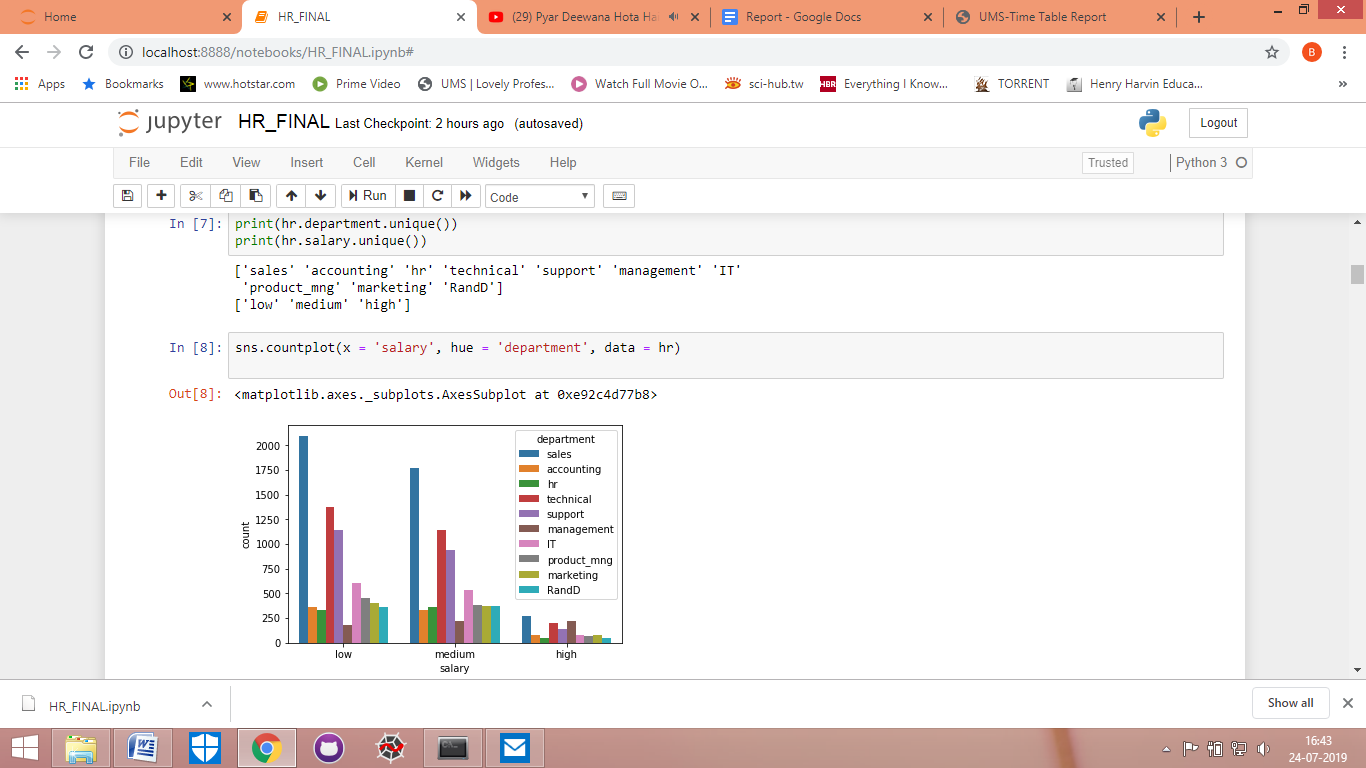
1. Now in this step I checked the name of all the columns.



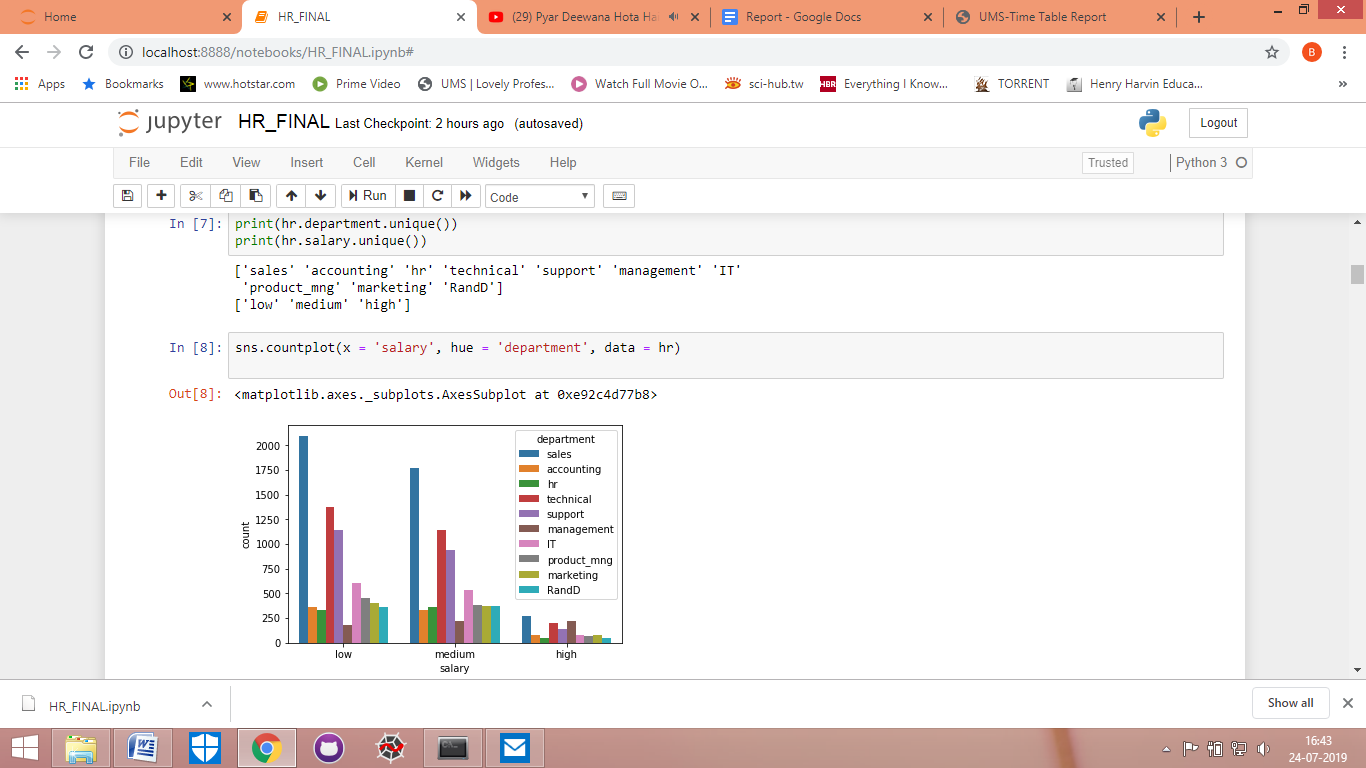
1. In this step I saw the detailed information of all the columns.



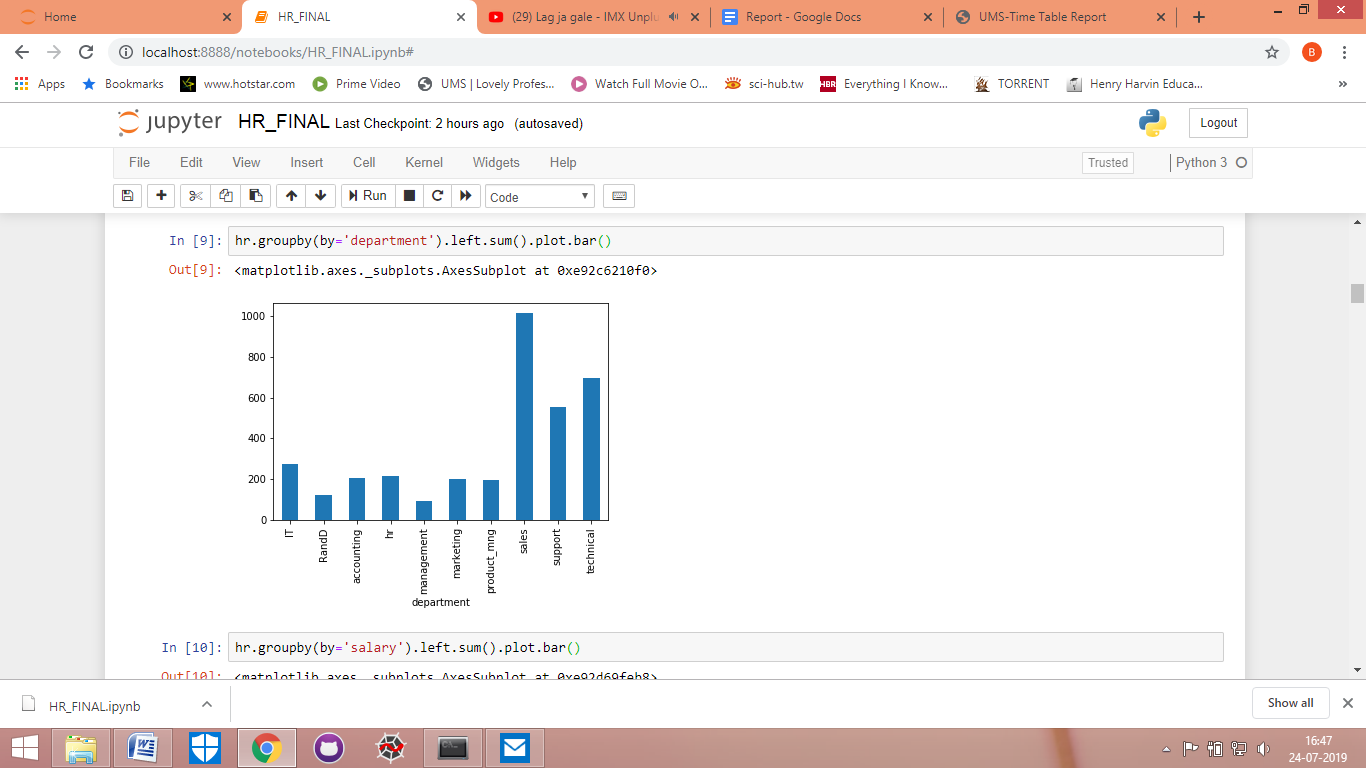
1. In this step I found out the unique categories of department and salary.



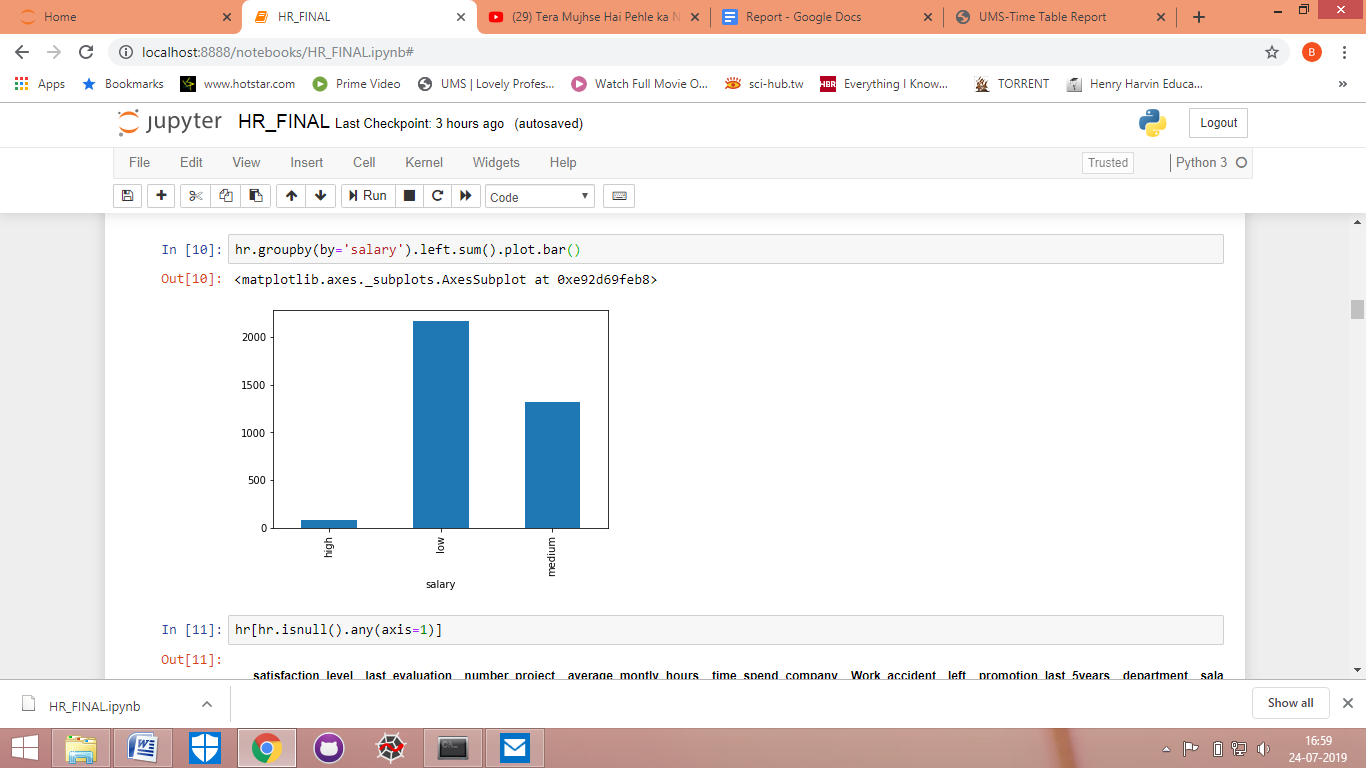
1. Here I have plotted the number of employees in each department on the basis of salary.



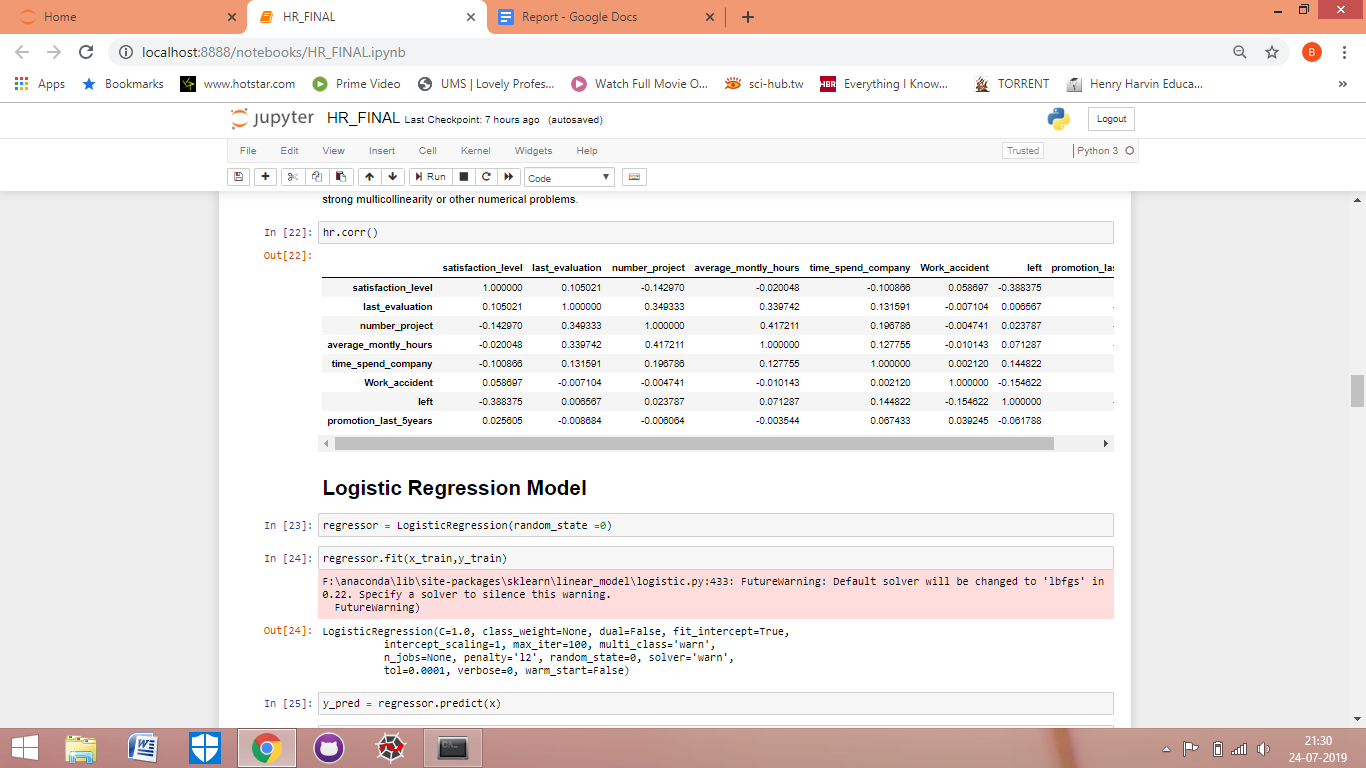
1. In this graph, I have shown the number of employees that have left in each department. And this graph shows that maximum employees left were from sales department.



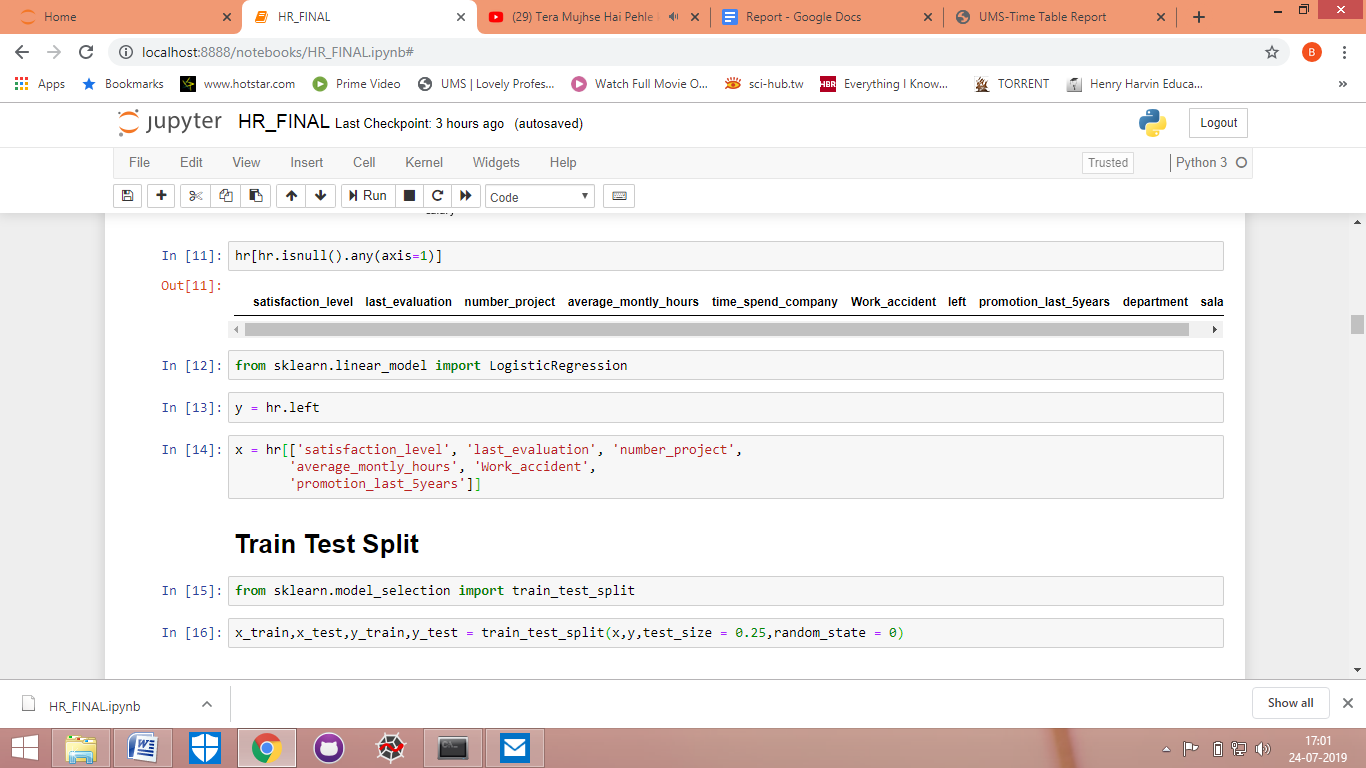
1. In this graph, I have plotted the number of employees on the basis of their salary, who have left the company. Employees with the low salary left the most.



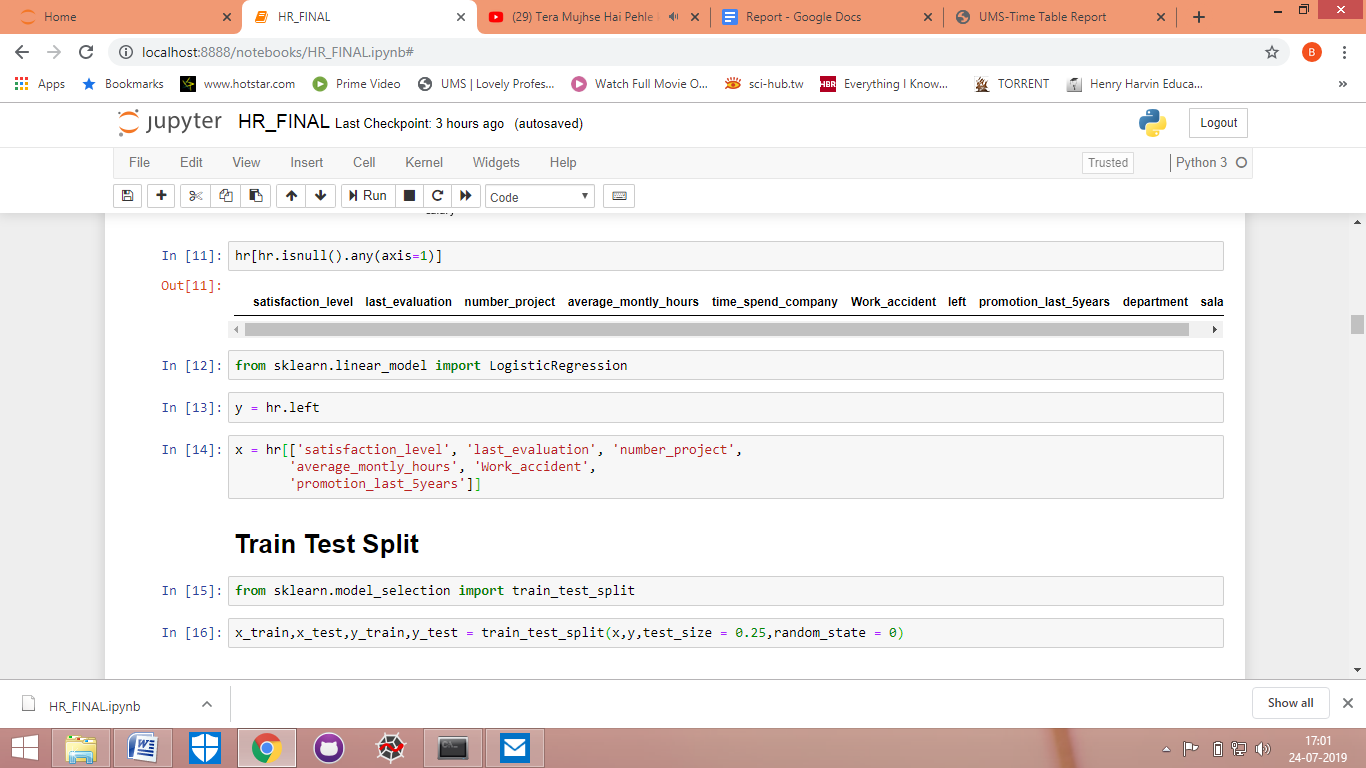
1. Now here I am checking the correlation of various variables with left variable.



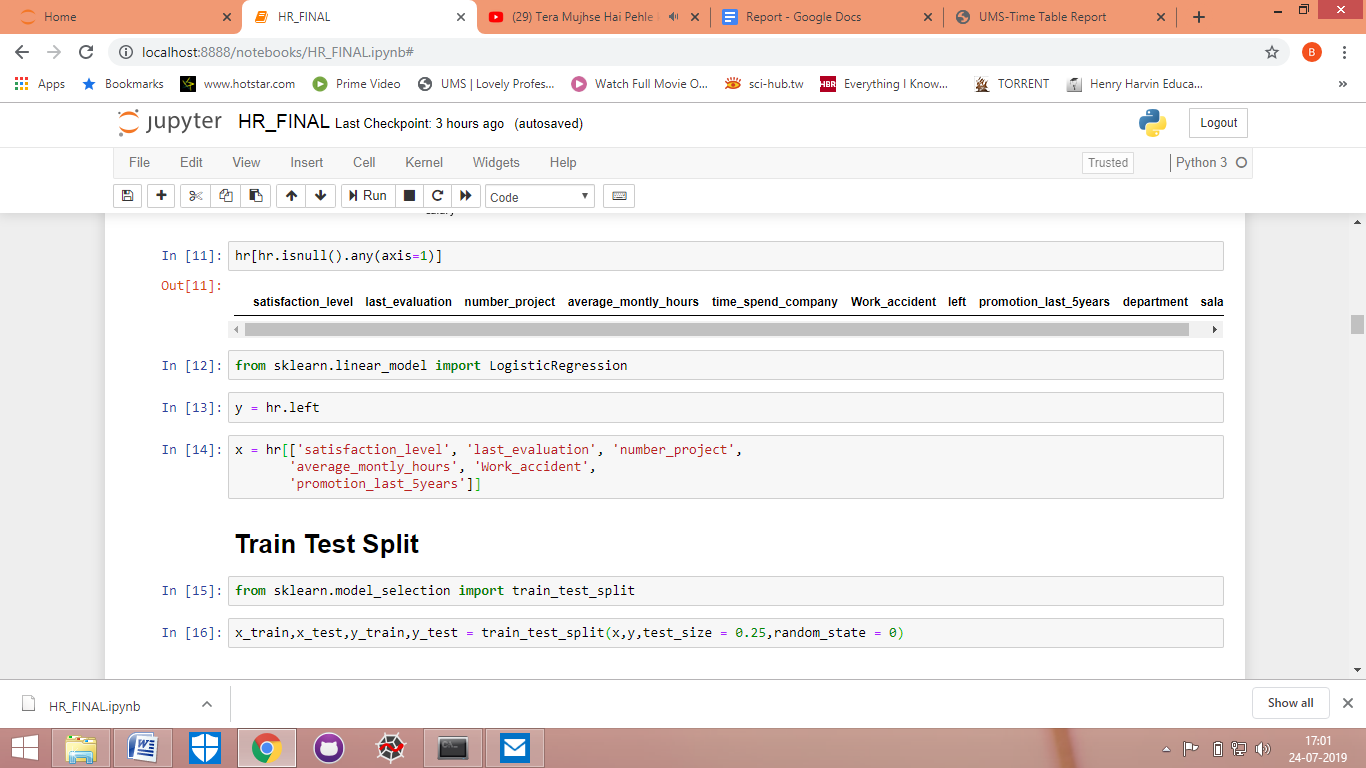
1. Now I have checked the data for any NAN values.



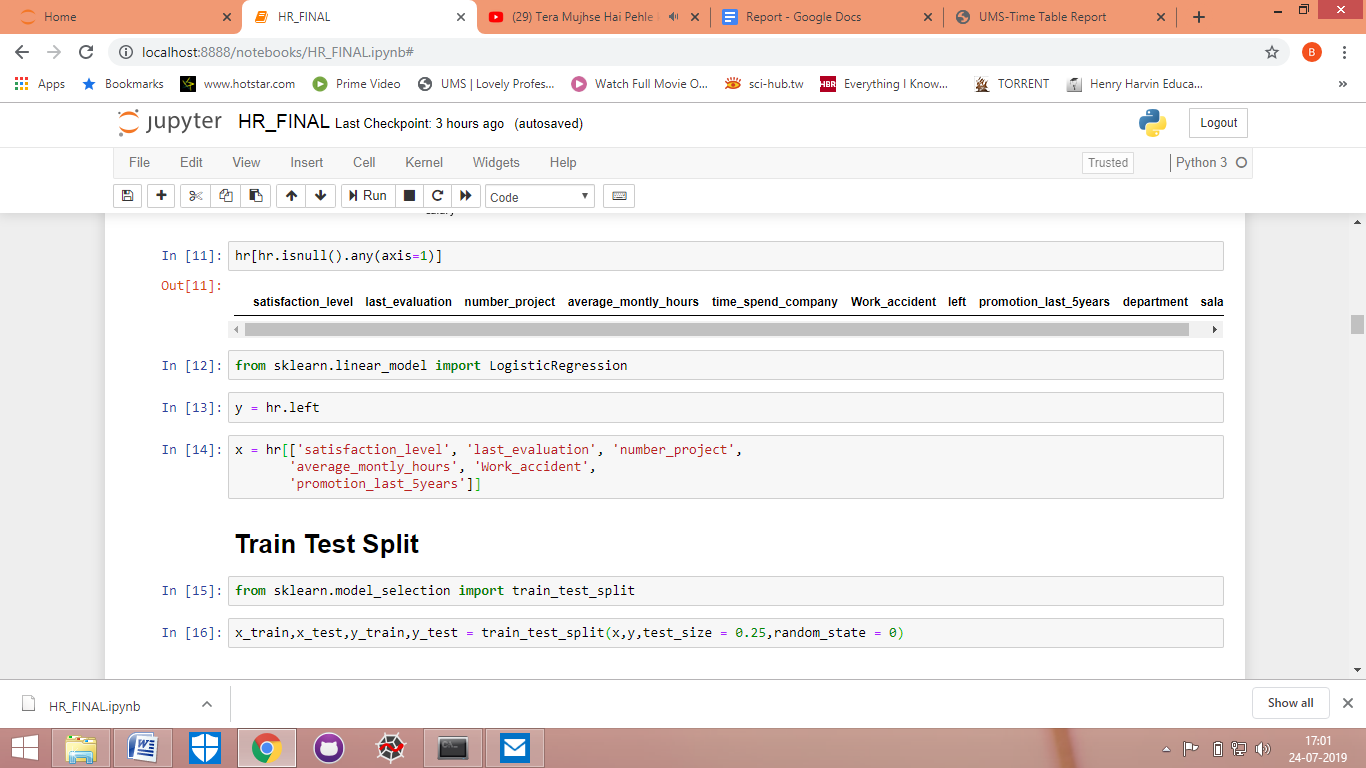
1. Now I have imported the logistic regression model from sklearn.linear\_model library.



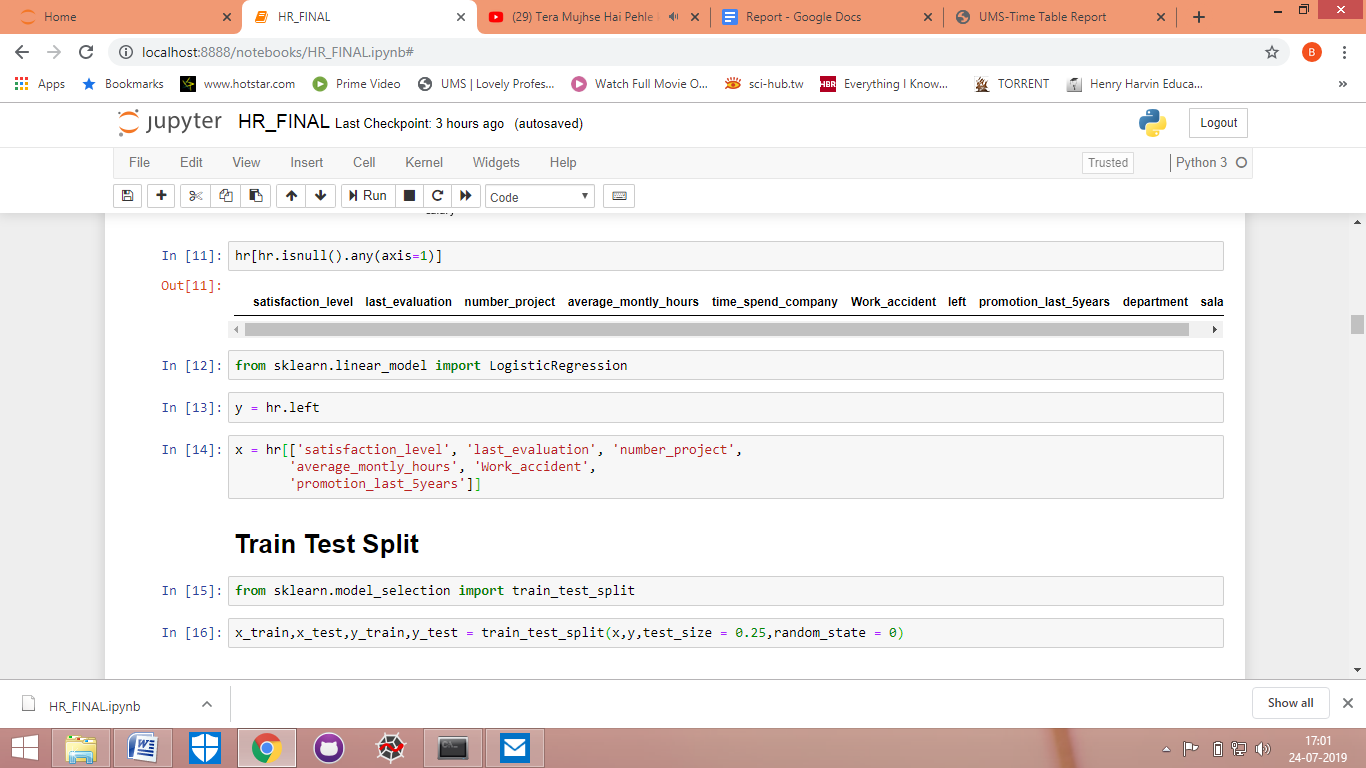
1. In this step i have assigned dependent variable i.e. left column to y.



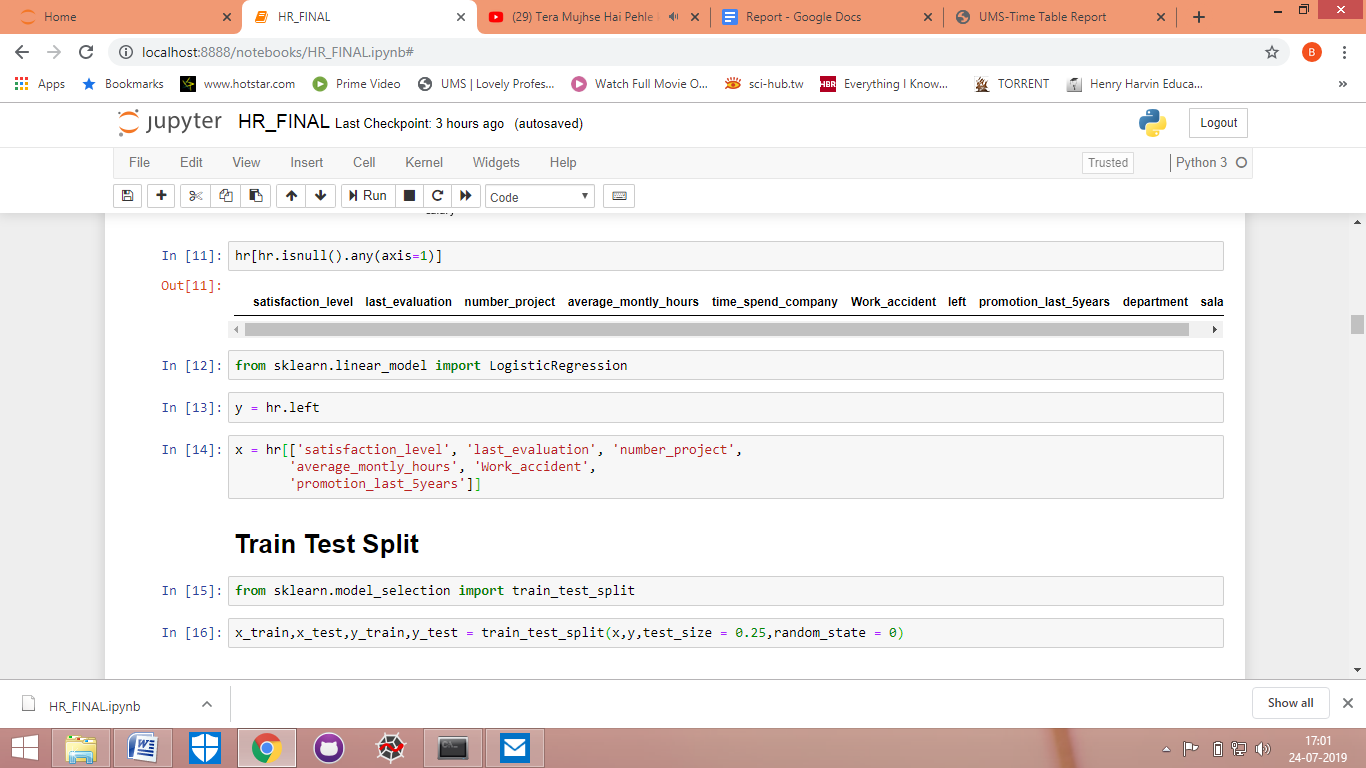
1. In this step, I have assigned independent variables to x. I have taken satisfaction level, last evaluation, number project, average monthly hours, Work accident, promotion last 5 years.



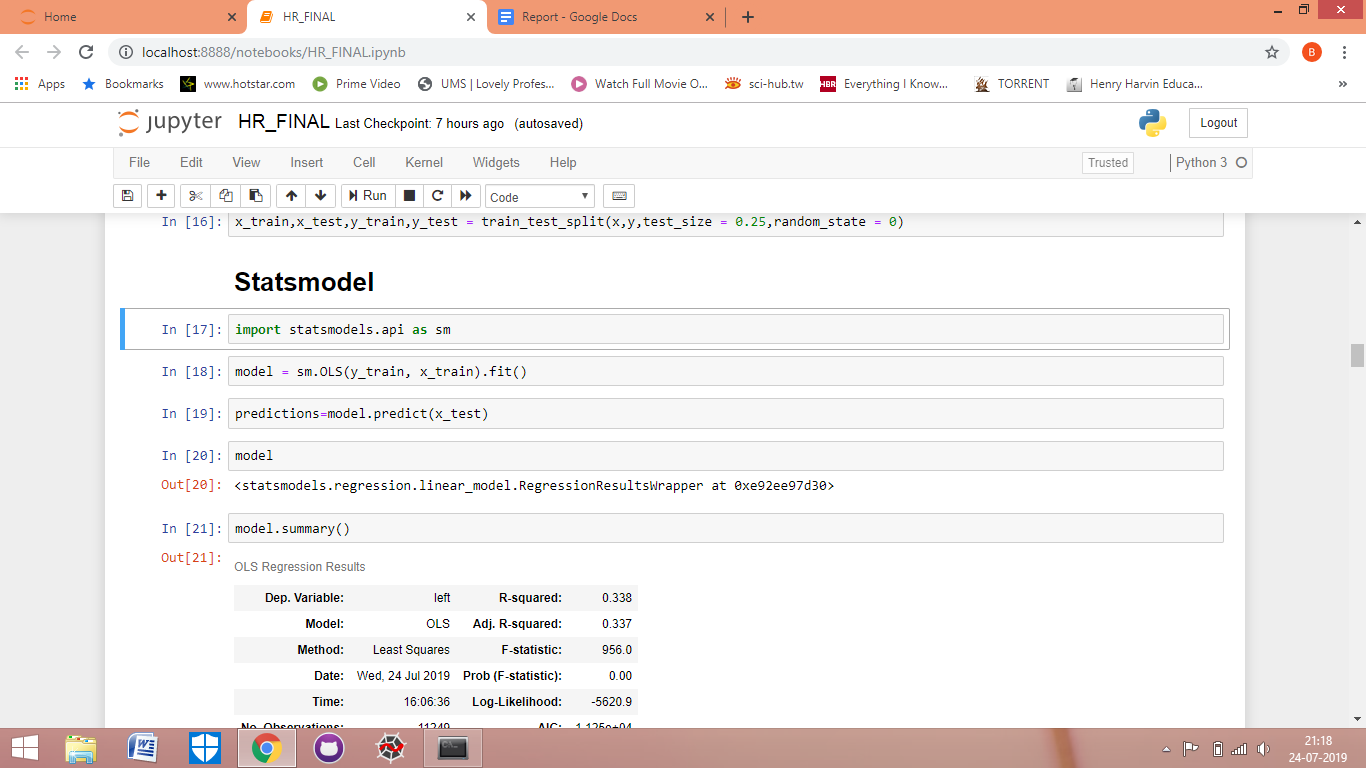
1. In this step I have imported train test split from sklearn.model selection library.



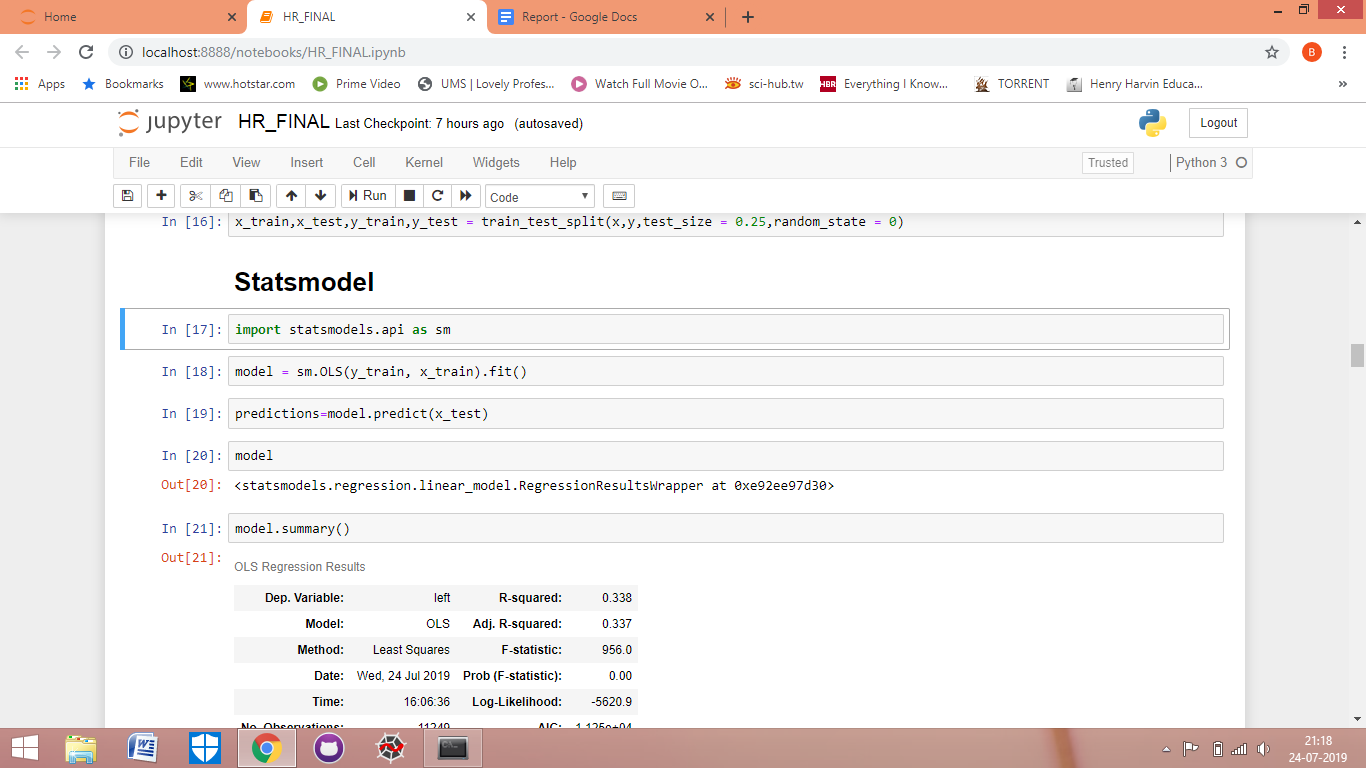
1. Here I have split the data set in test and train in the ratio of 25:75.



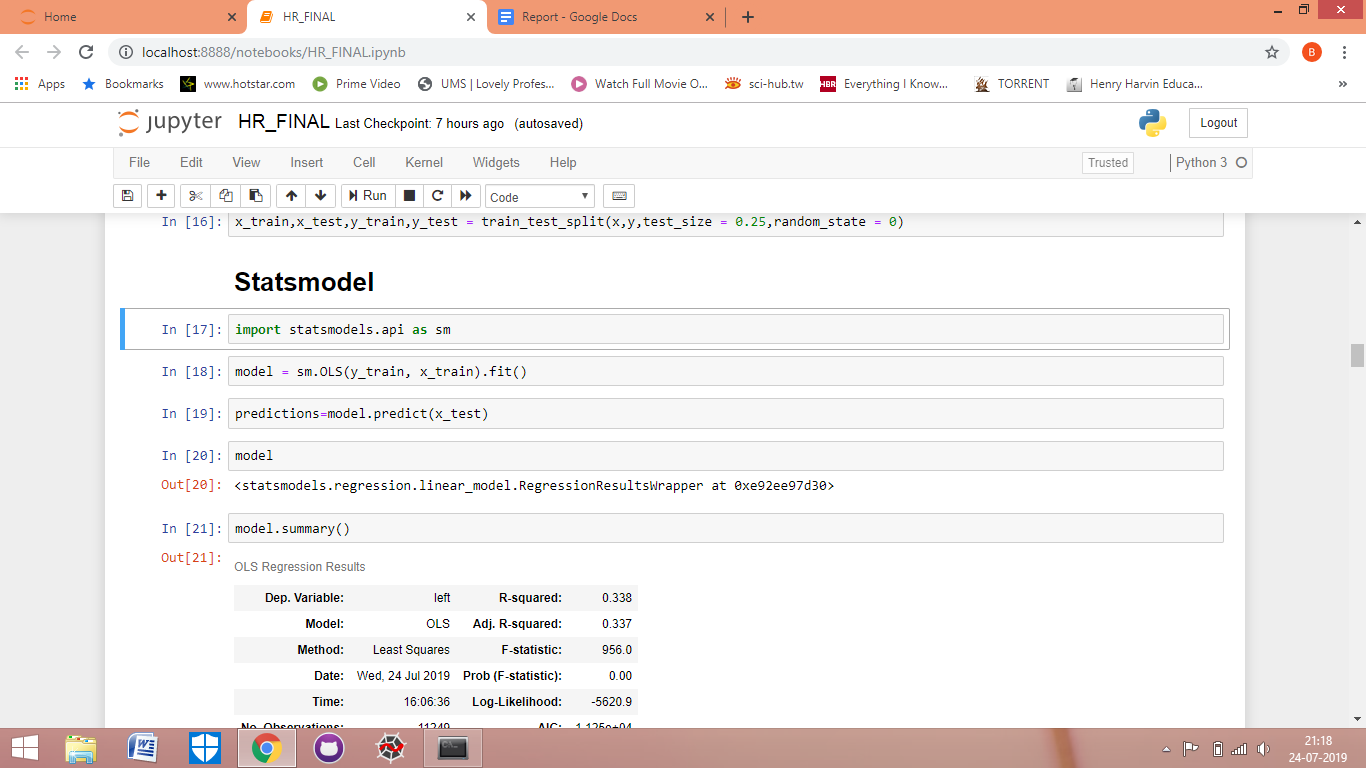
1. In this step I have imported statsmodel.api library as sm.



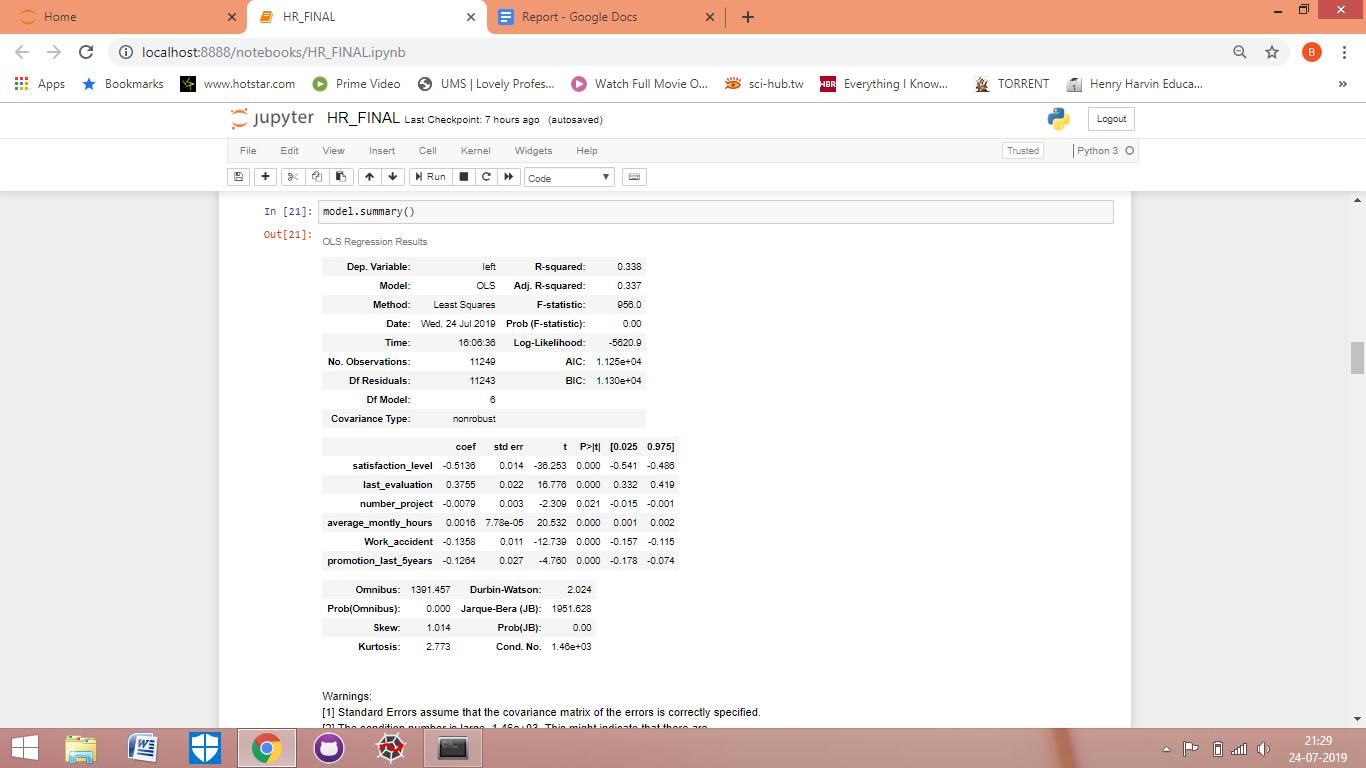
1. In this step i have created a statsmodel and in that i fitted y\_train and x\_train.



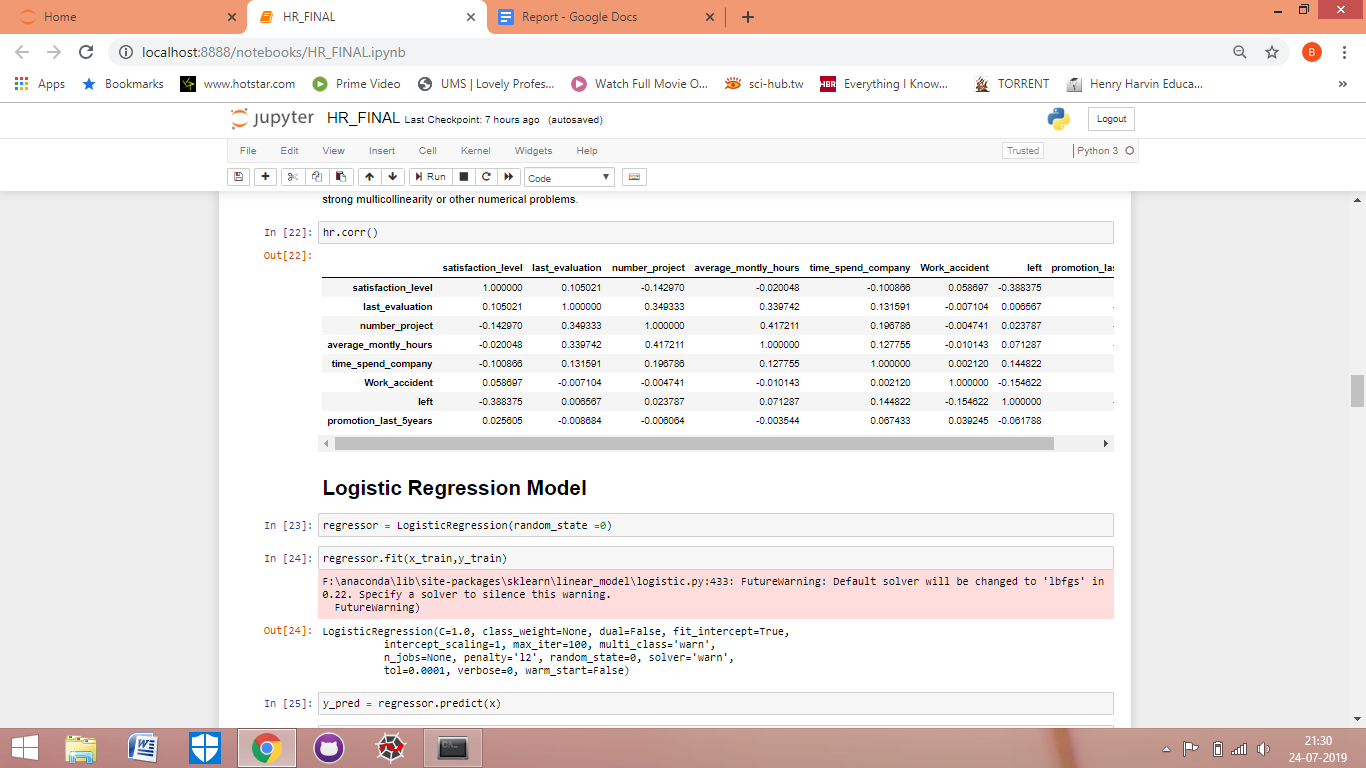
1. In this step i predicted the statsmodel which I had created in the last step.



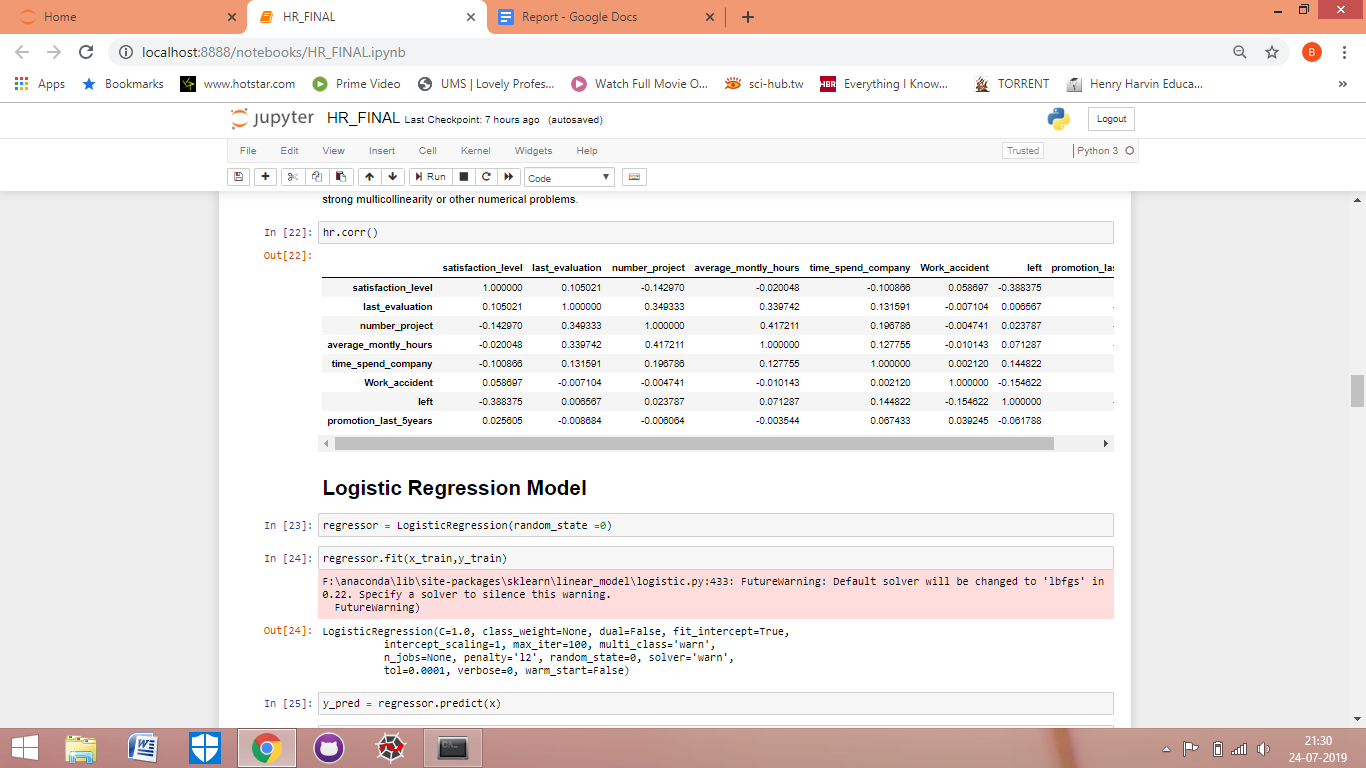
1. Now here I am seeing the summary of the statsmodel. And checking the various parameters like R square, adjusted R square and p values.



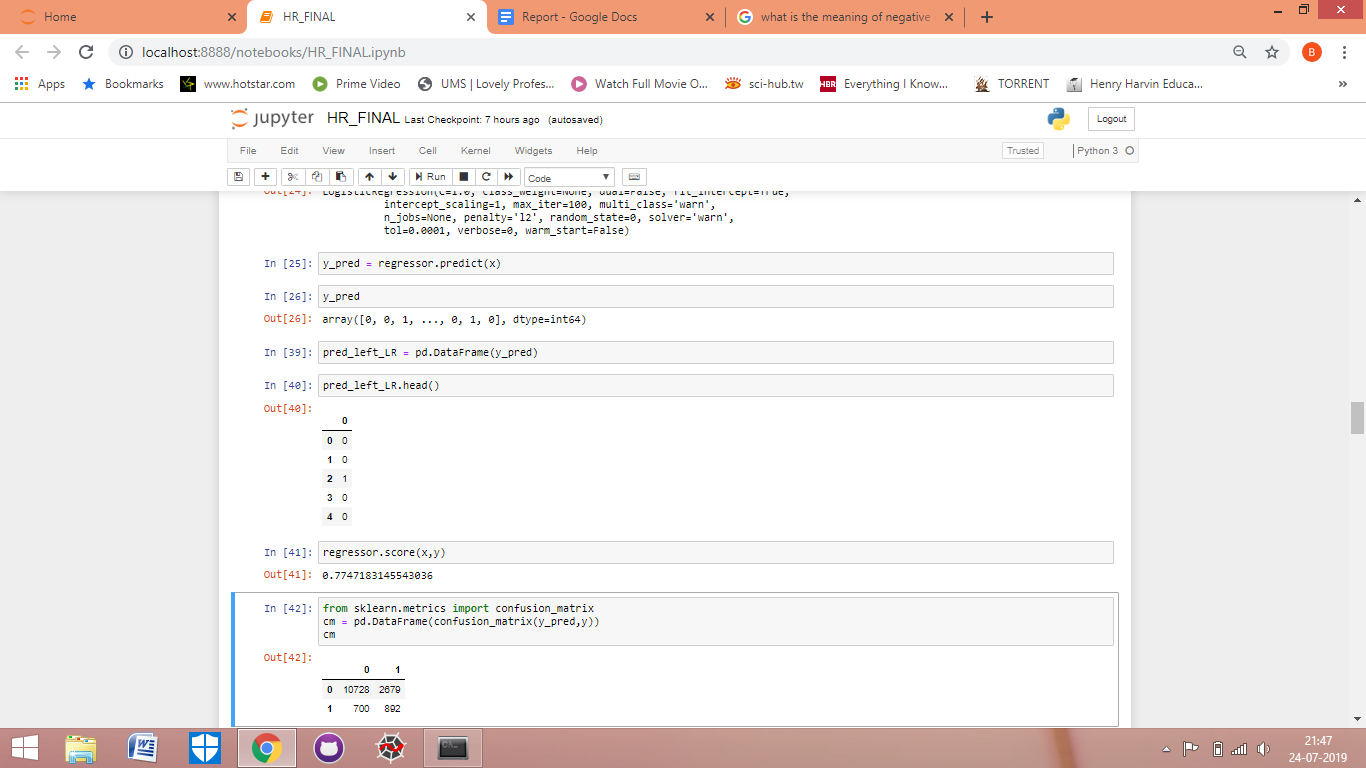
1. Now here I am creating a logistic regression model by the name of regressor.



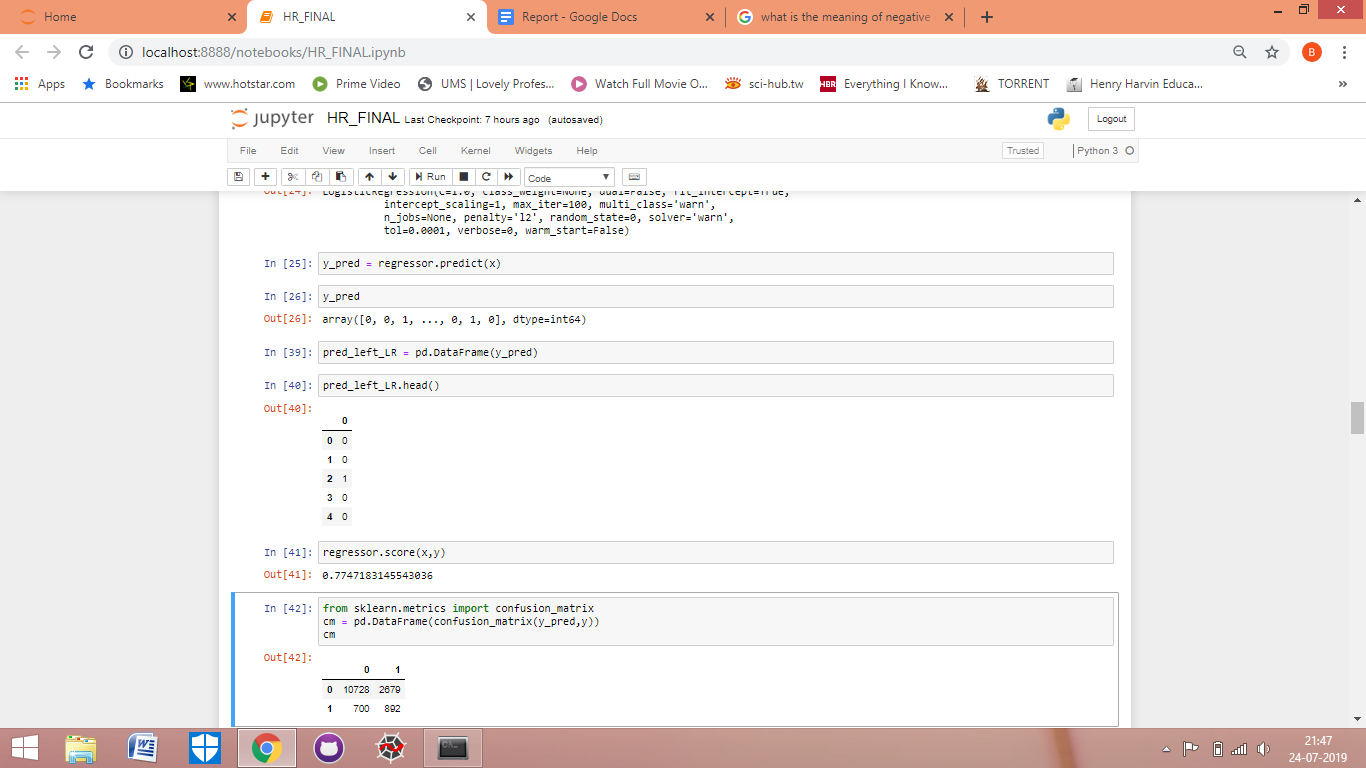
1. Here i have fitted the training data set in the model.



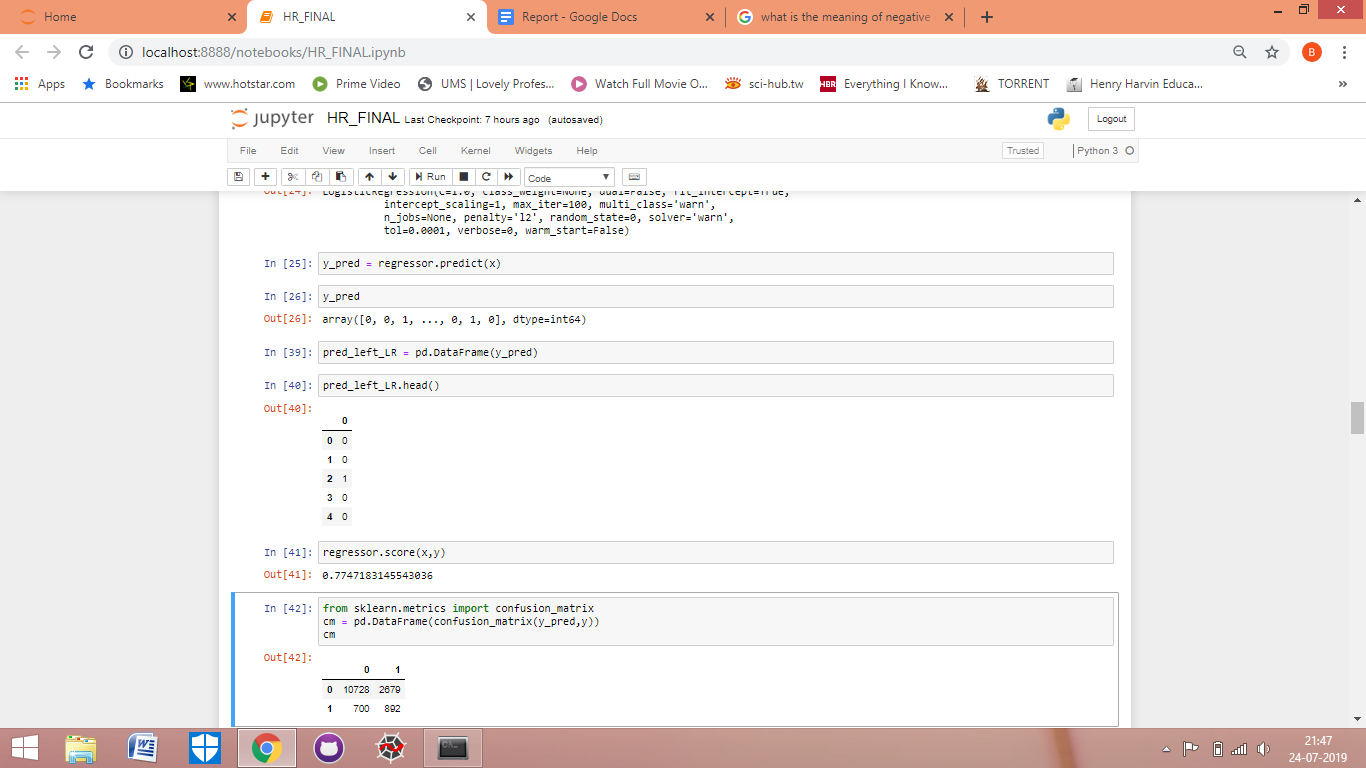
1. Now after creating the model, I am now predicting the y on the whole data set.



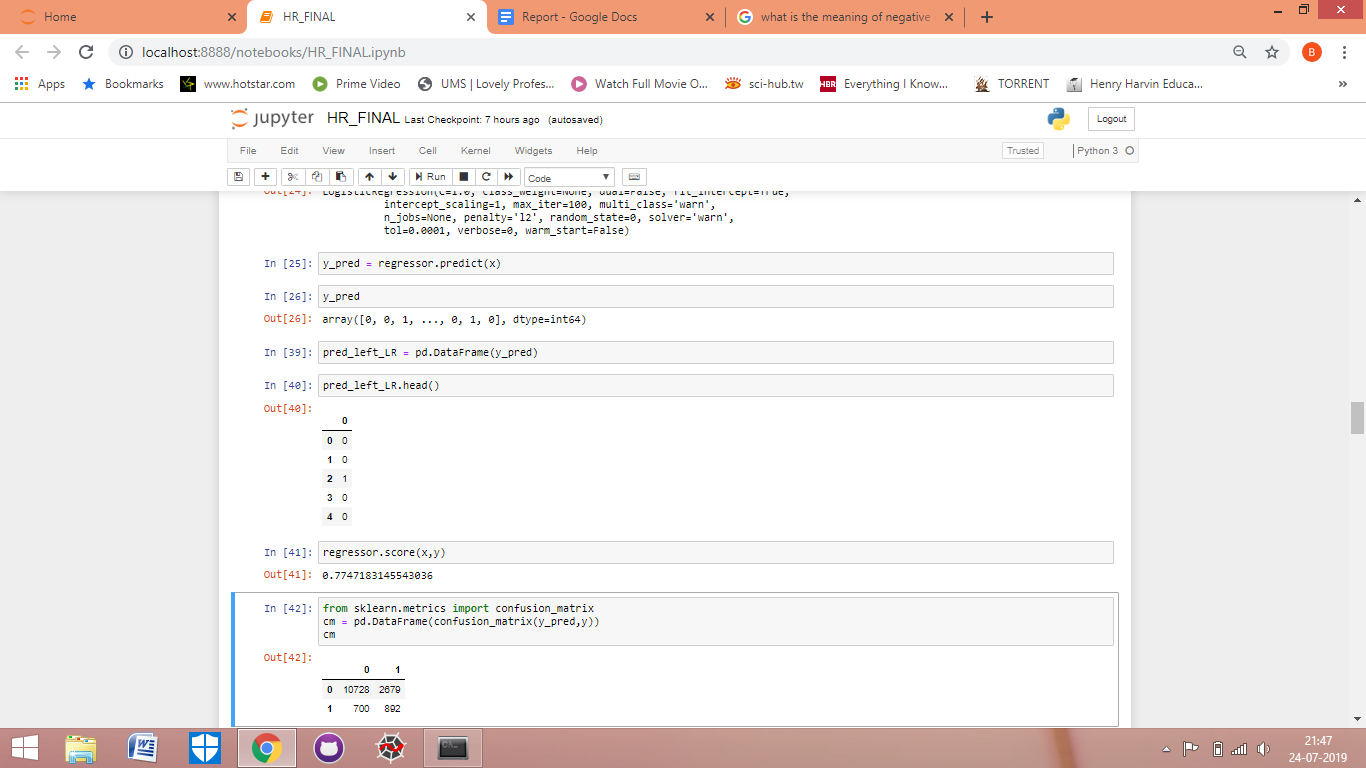
1. Now the predicted results are in array form. Using code, I will now convert it into dataframe, which is easier to understand.



1. Now I am checking the accuracy of my model by checking its score, which comes out to be more than 77.



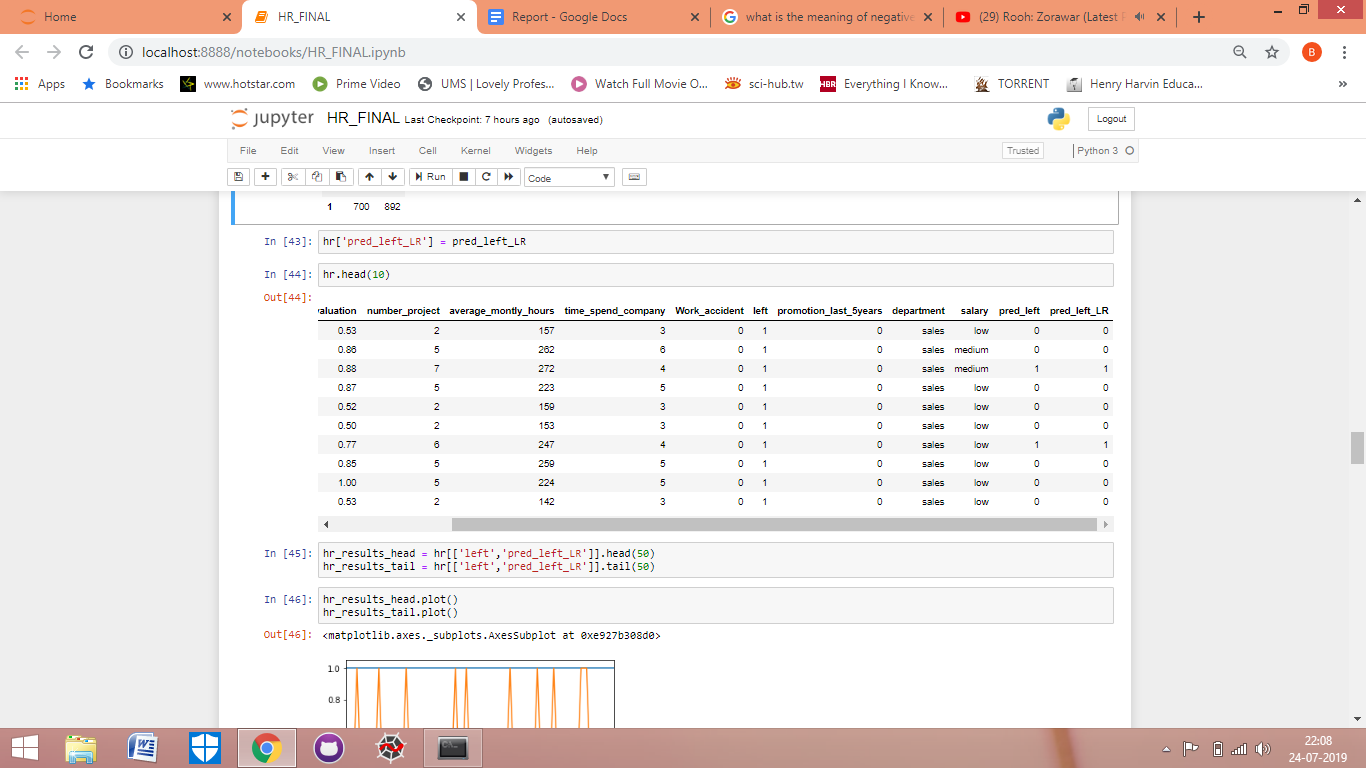
1. Now I am using the validation technique called confusion matrix to validate my model.



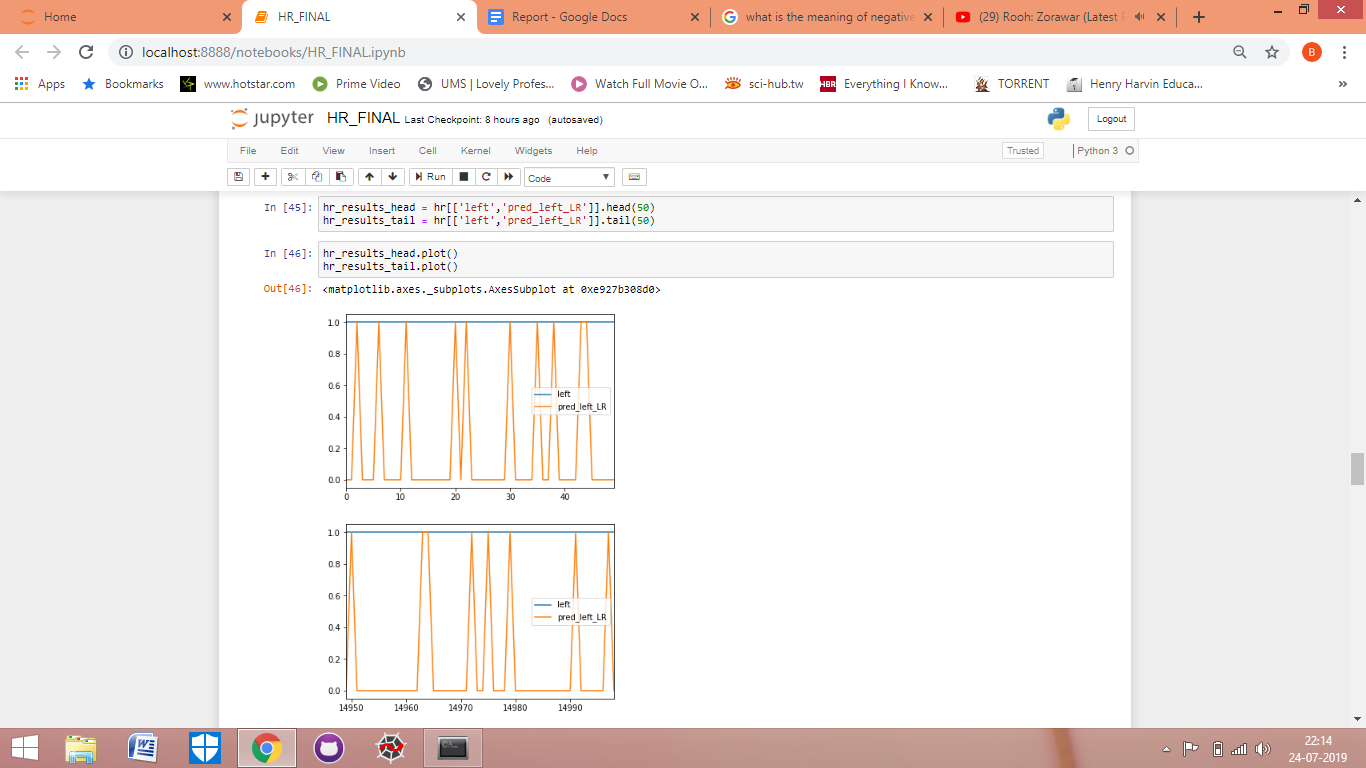
This matrix shows that how many times actual 0 is predicted 0 (true result) and actual 0 is predicted as 1 (false result).

And how many times actual 1 is predicted as 0 (false result) and actual 1 predicted as 1 (true result).

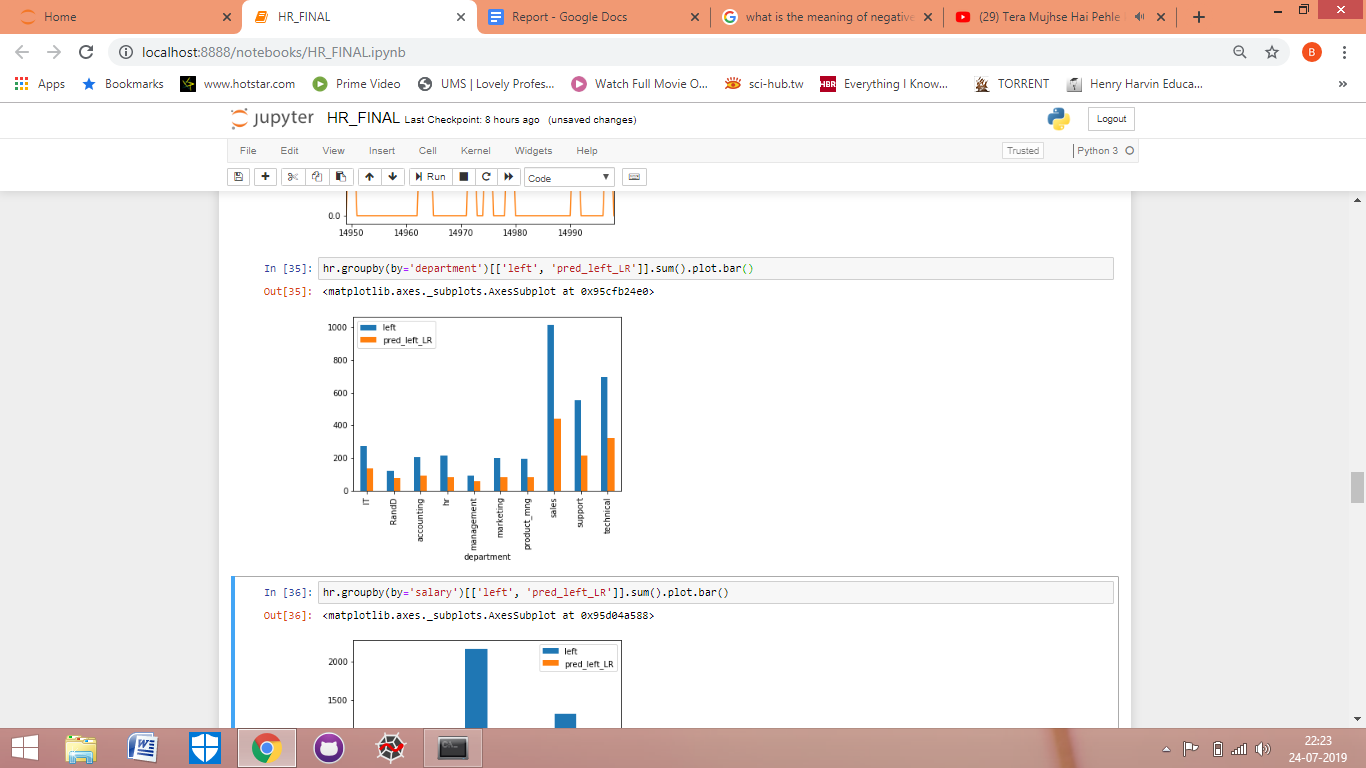
1. Now here I have added the predicted left column in the main data set of hr and then checked the complete data set.

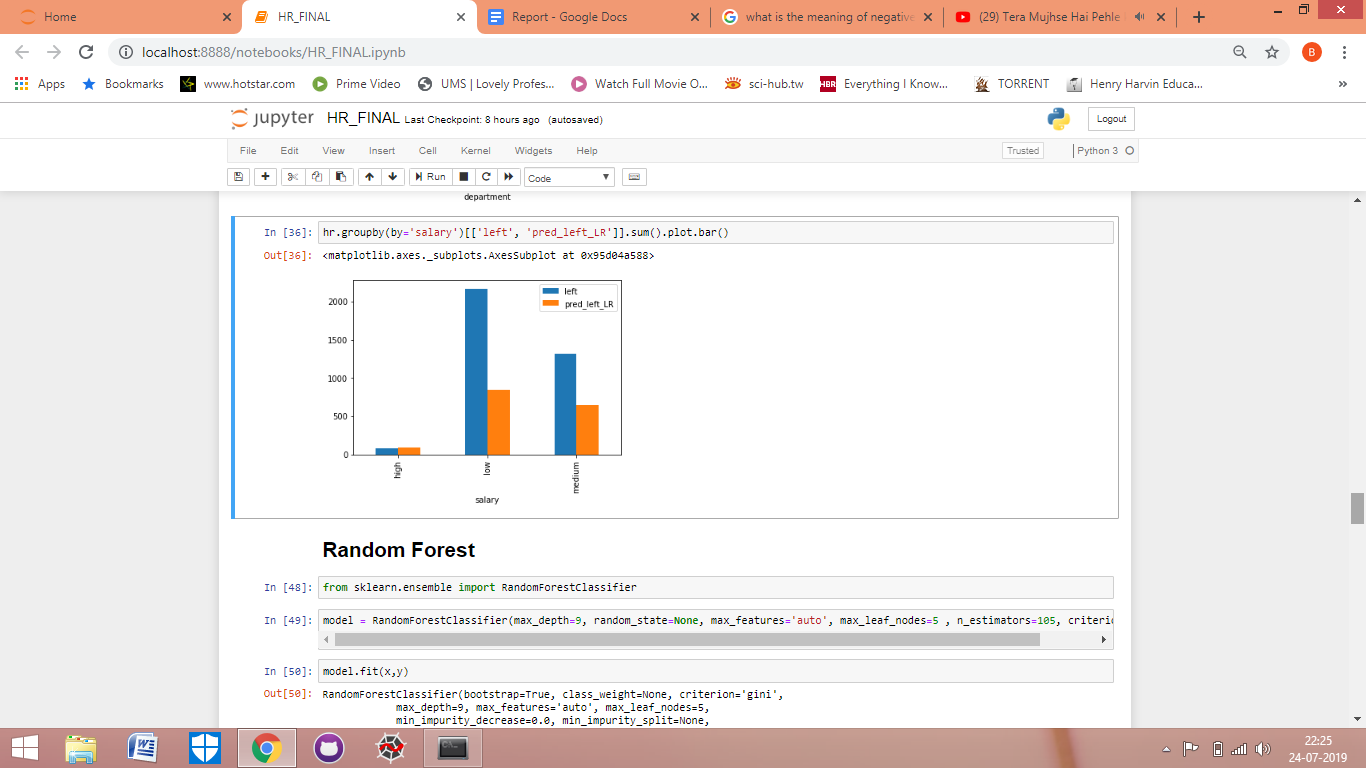


1. Here now I have created two different subsets of data from the main hr data set by taking columns of actual left and predicted left (by logistic regression) only. And also in the first data set I took top 50 rows while in the second data set i took the last 50 rows. Then after creating the new data sets I prepared their line graphs for comparison.

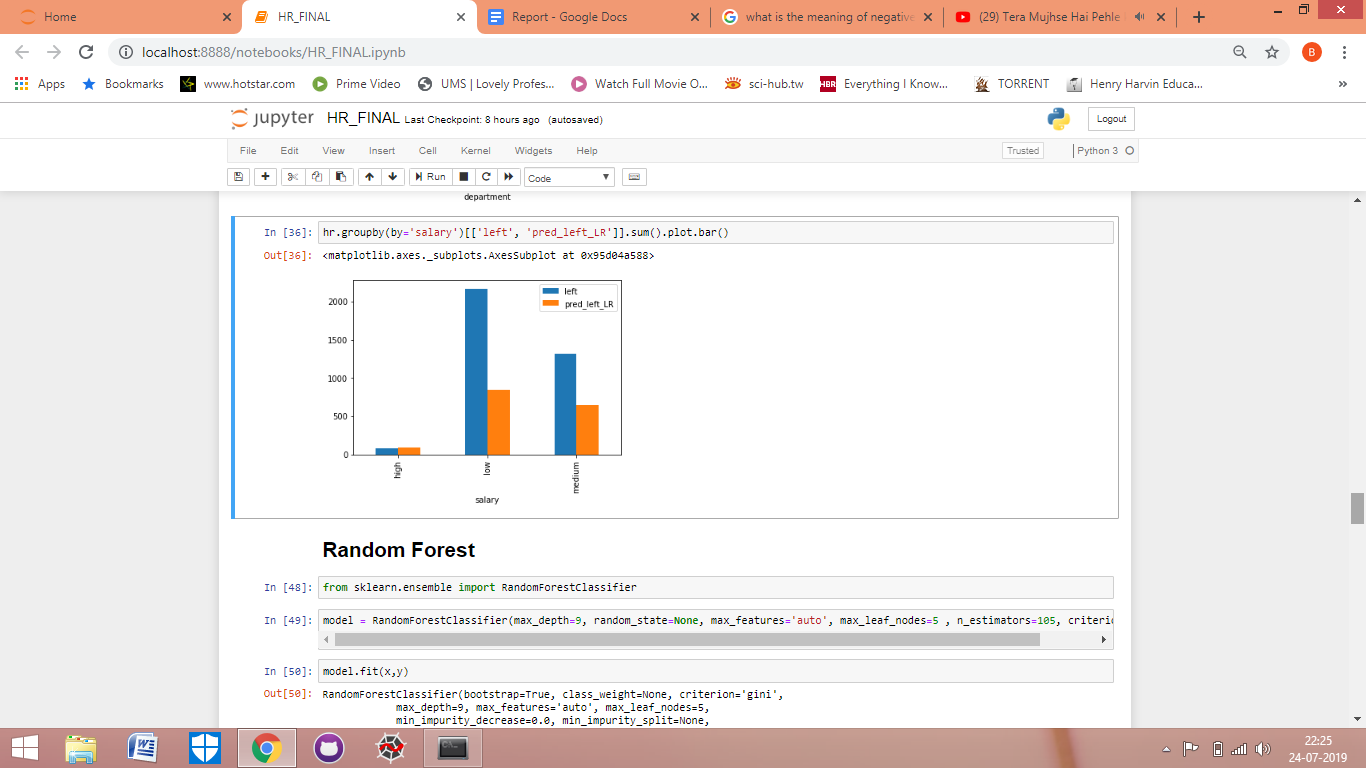


1. Here I plotted graphs of actual left and predicted left on the basis of departments and salary respectively using groupby function.

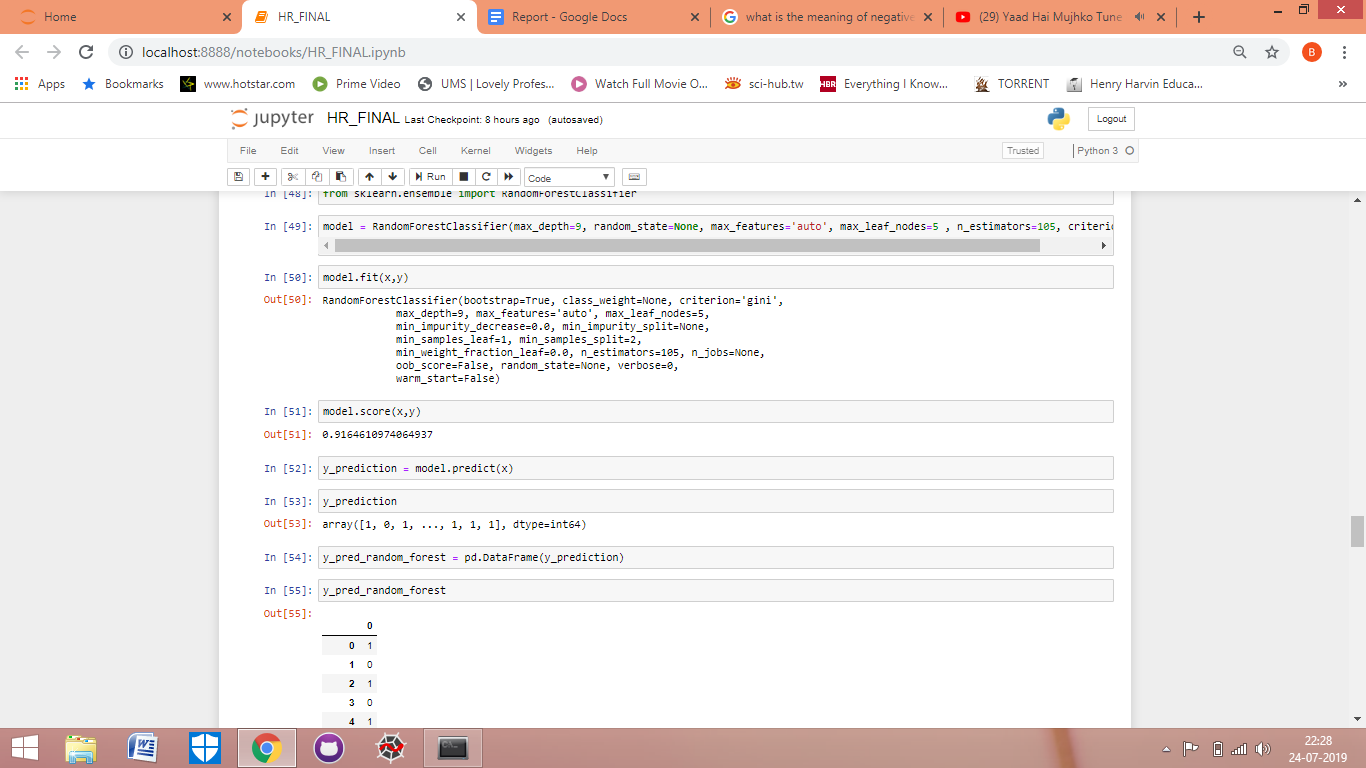




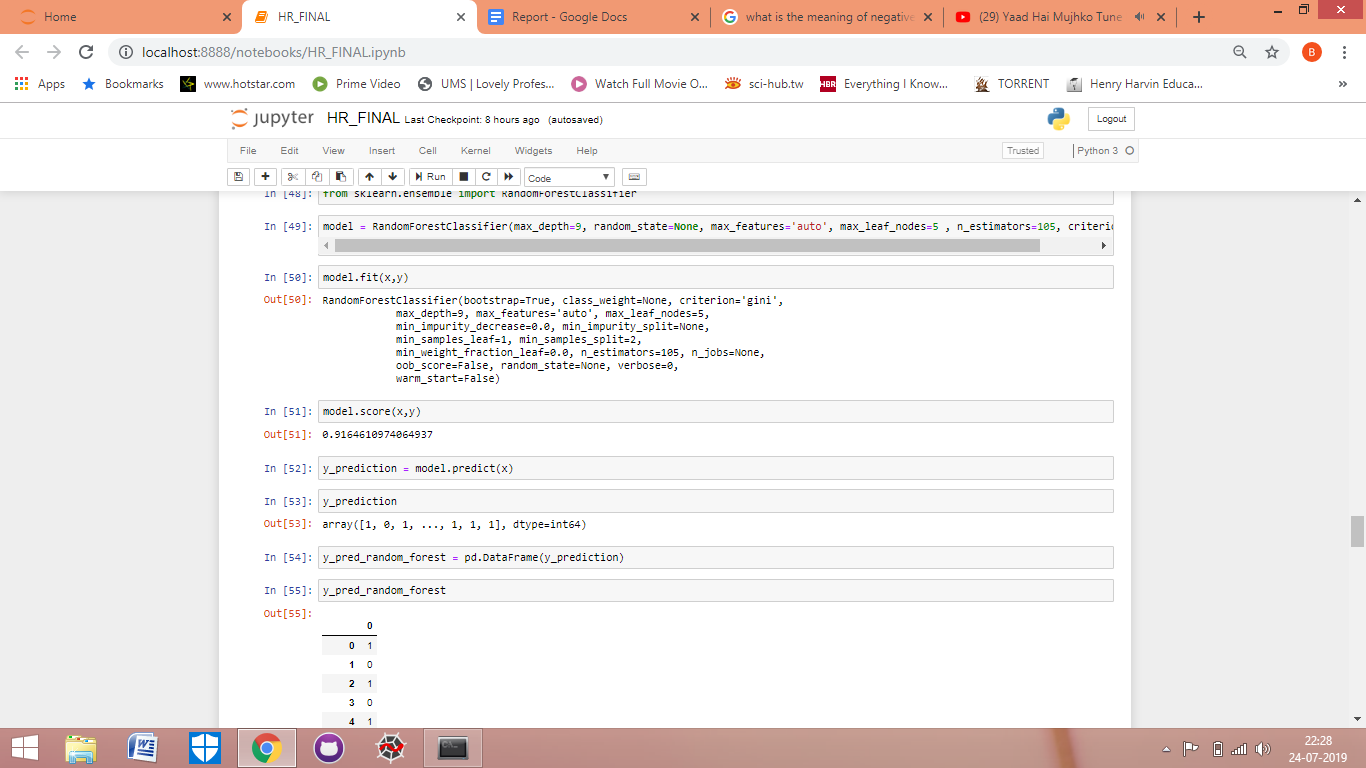
1. After completing the analysis by using logistic regression, now I imported randomforestclassifer to analyse and predict using random forest.



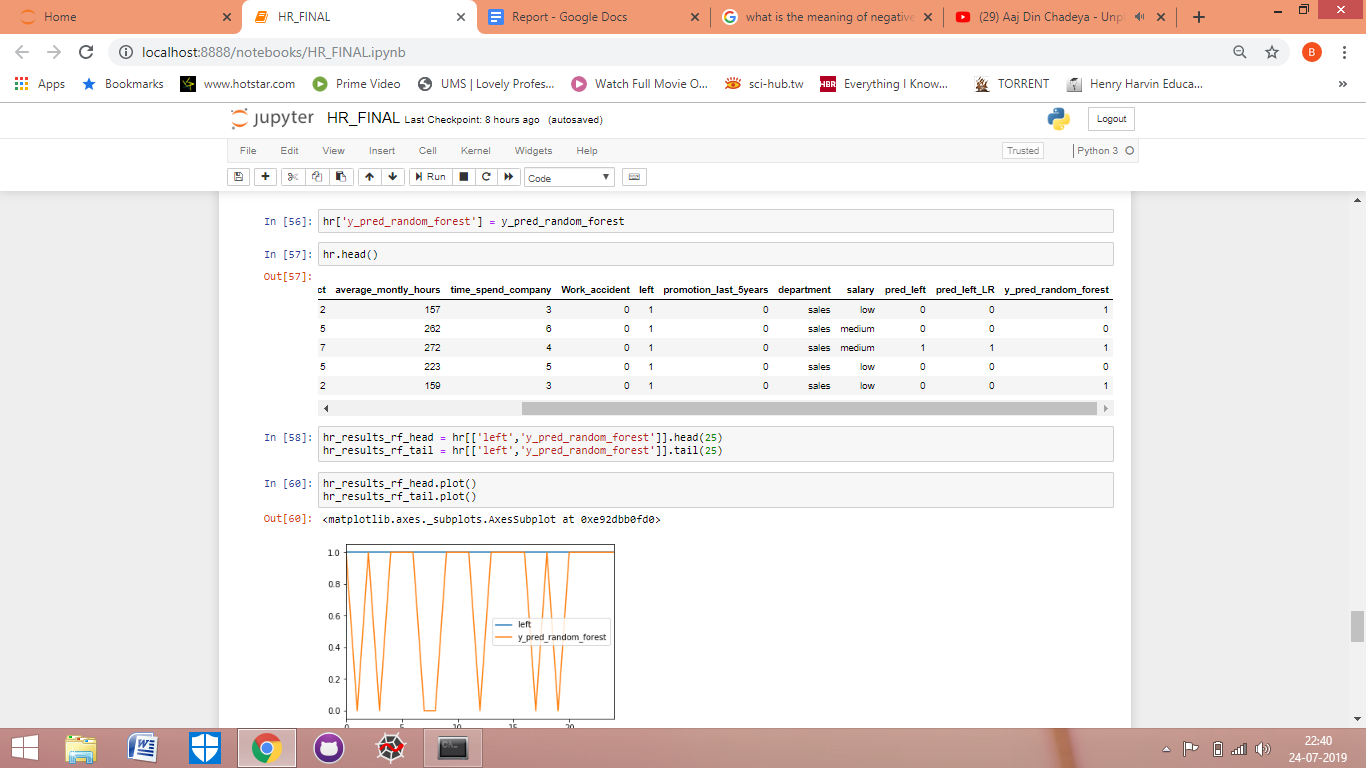
1. After importing the required model, I created the random forest model and fitted the model with x and y. And then checked out its score for accuracy, which comes out to be 0.9164.



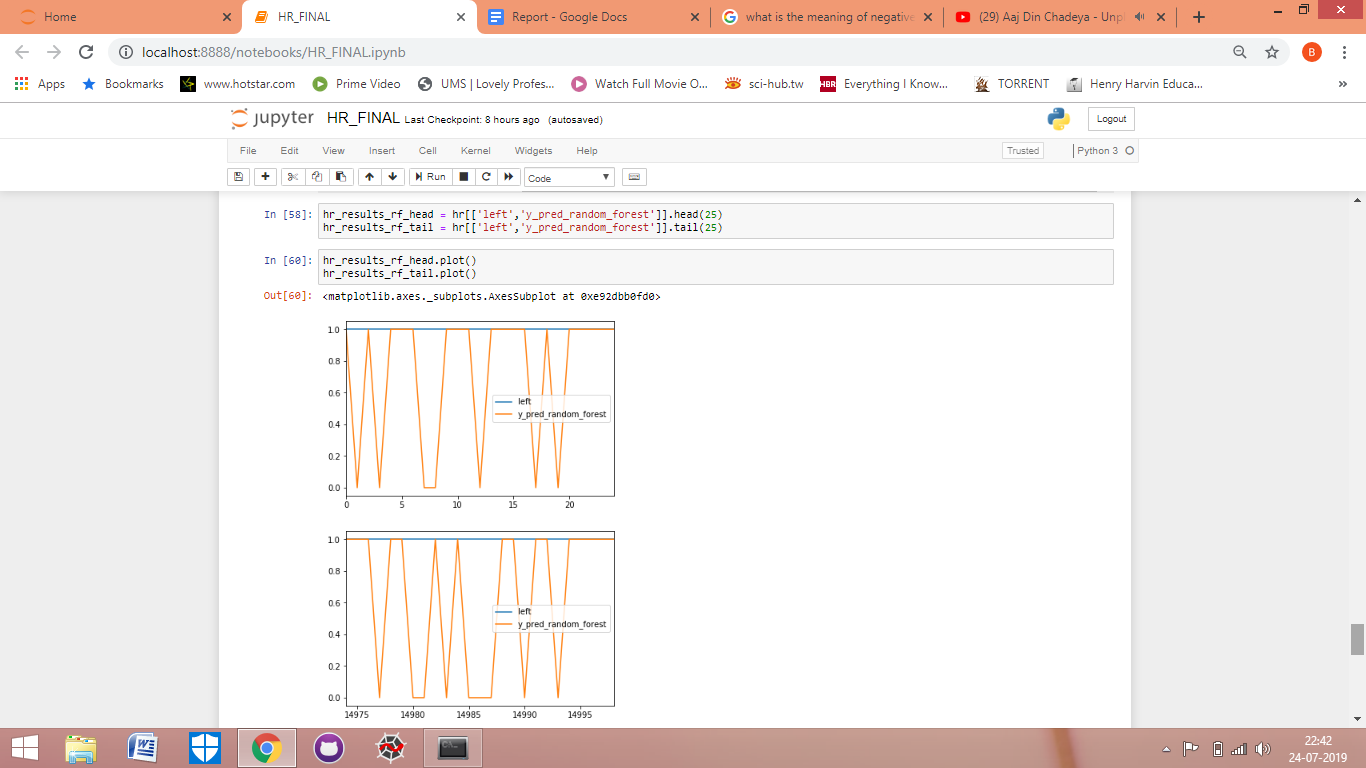
1. Now i predicted the value of y through model and then converted the results from array form to dataframe for better understanding.

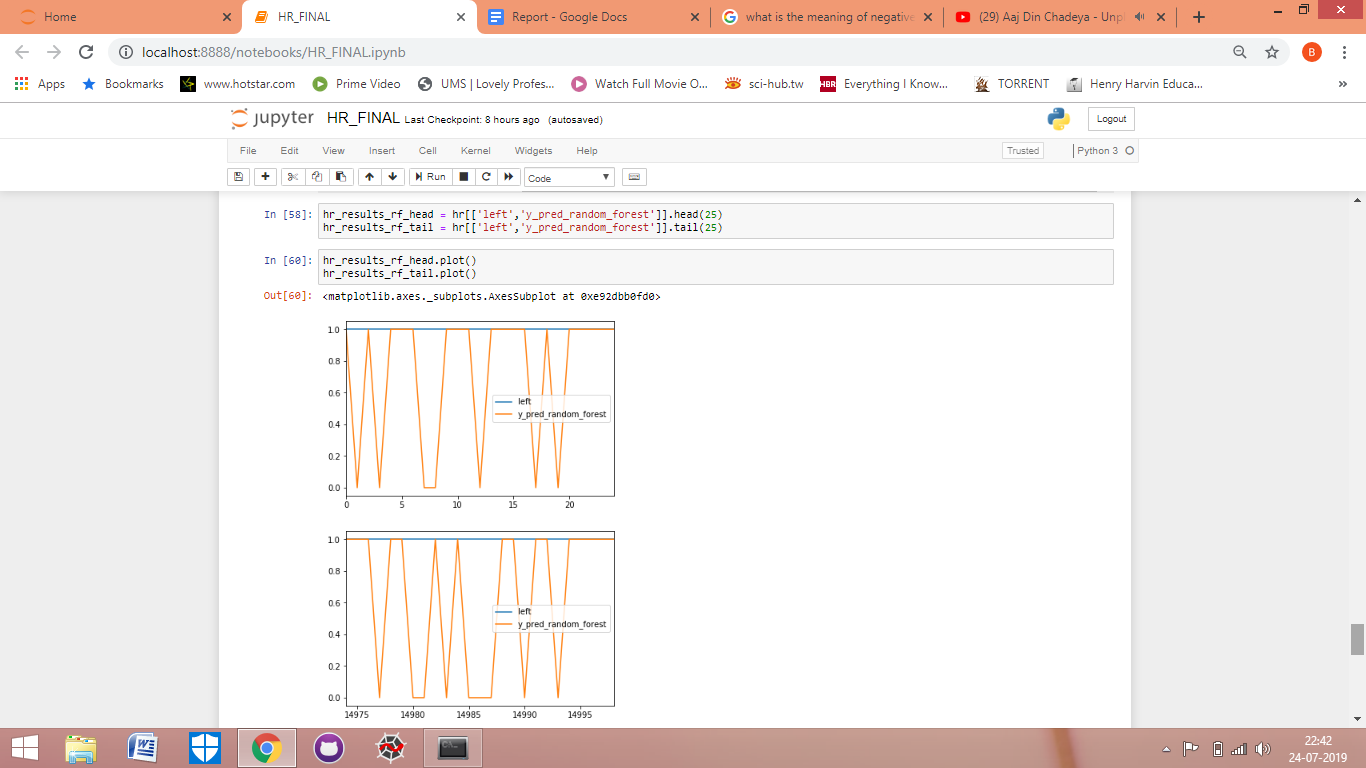


1. After converting the predicted results into data frame. I added those results in the main data set of hr by the name of column y\_pred\_random\_forest.

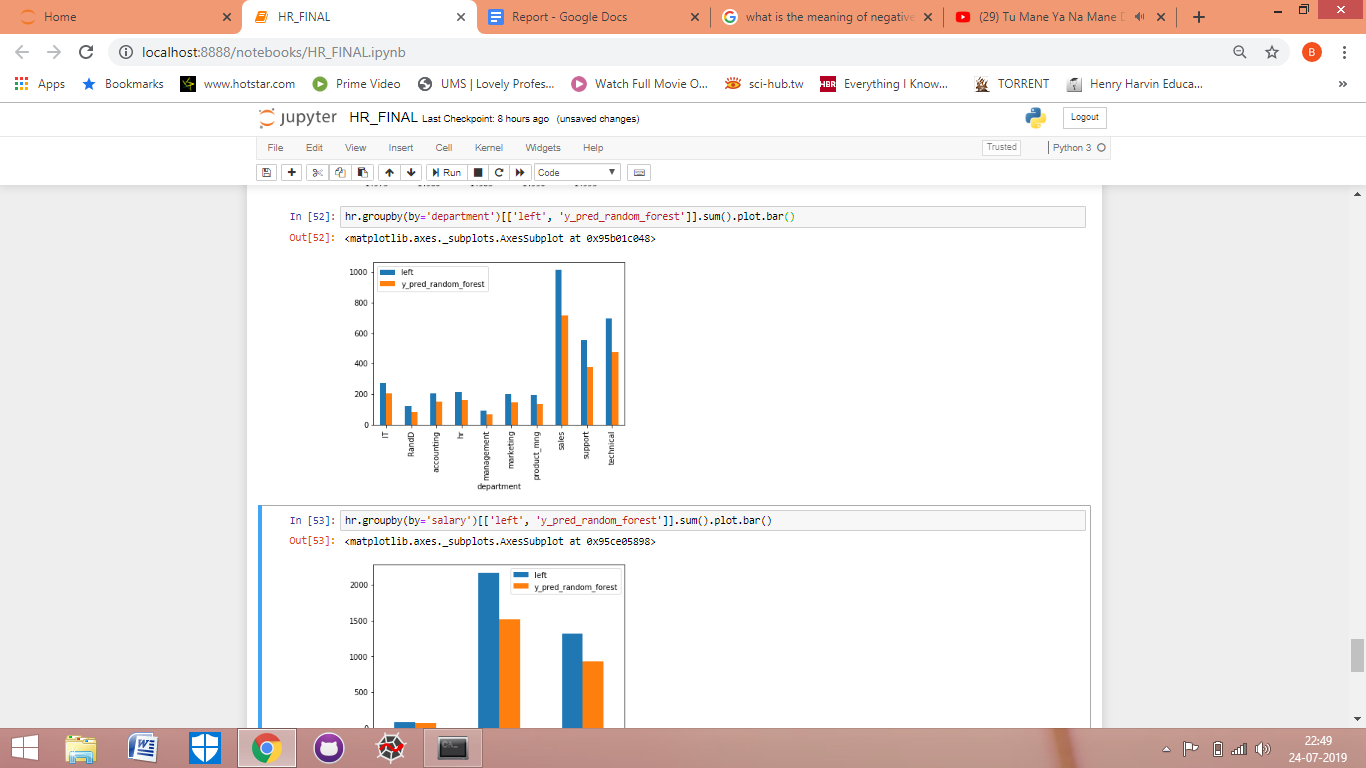


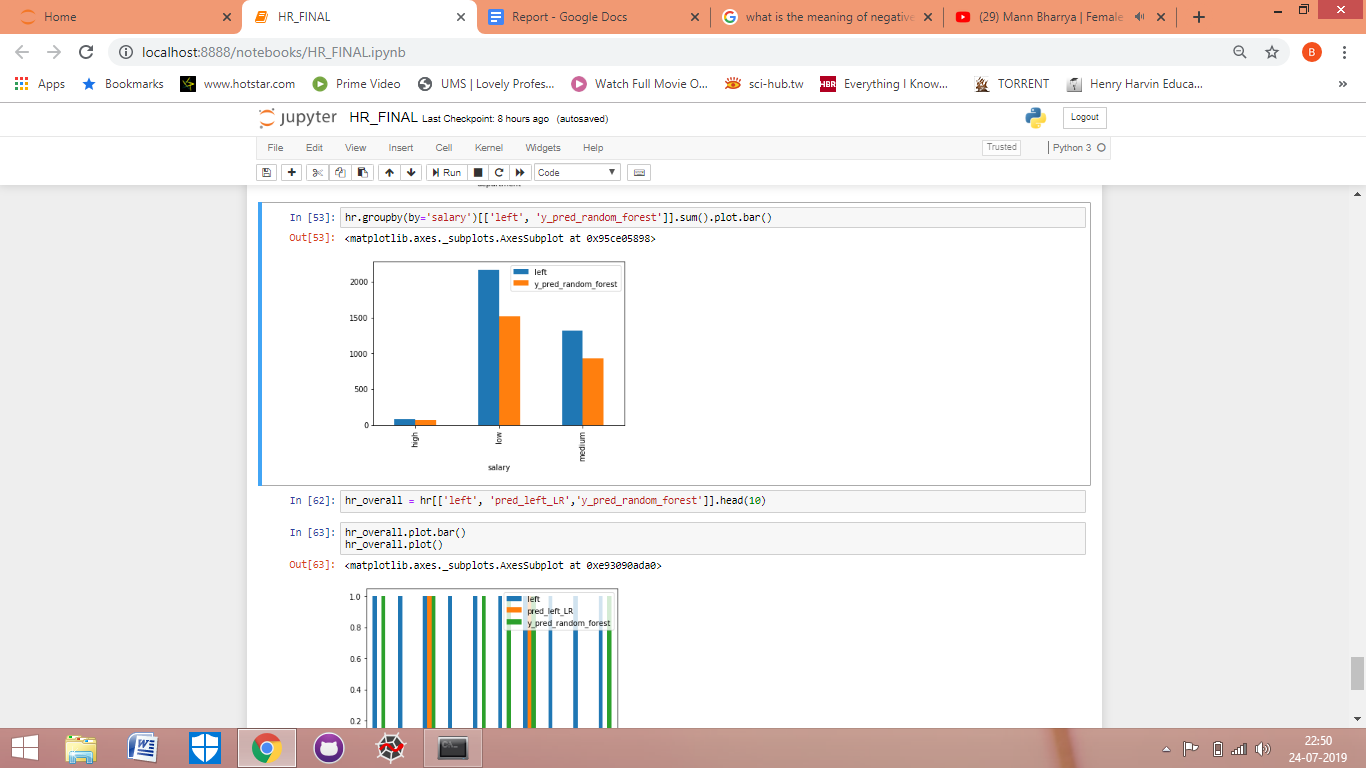
1. Here now I have created two different subsets of data from the main hr data set by taking columns of actual left and predicted left (by random forest) only. And also in the first data set I took top 25 rows while in the second data set i took the last 25 rows. Then after creating the new data sets I prepared their line graphs for comparison.



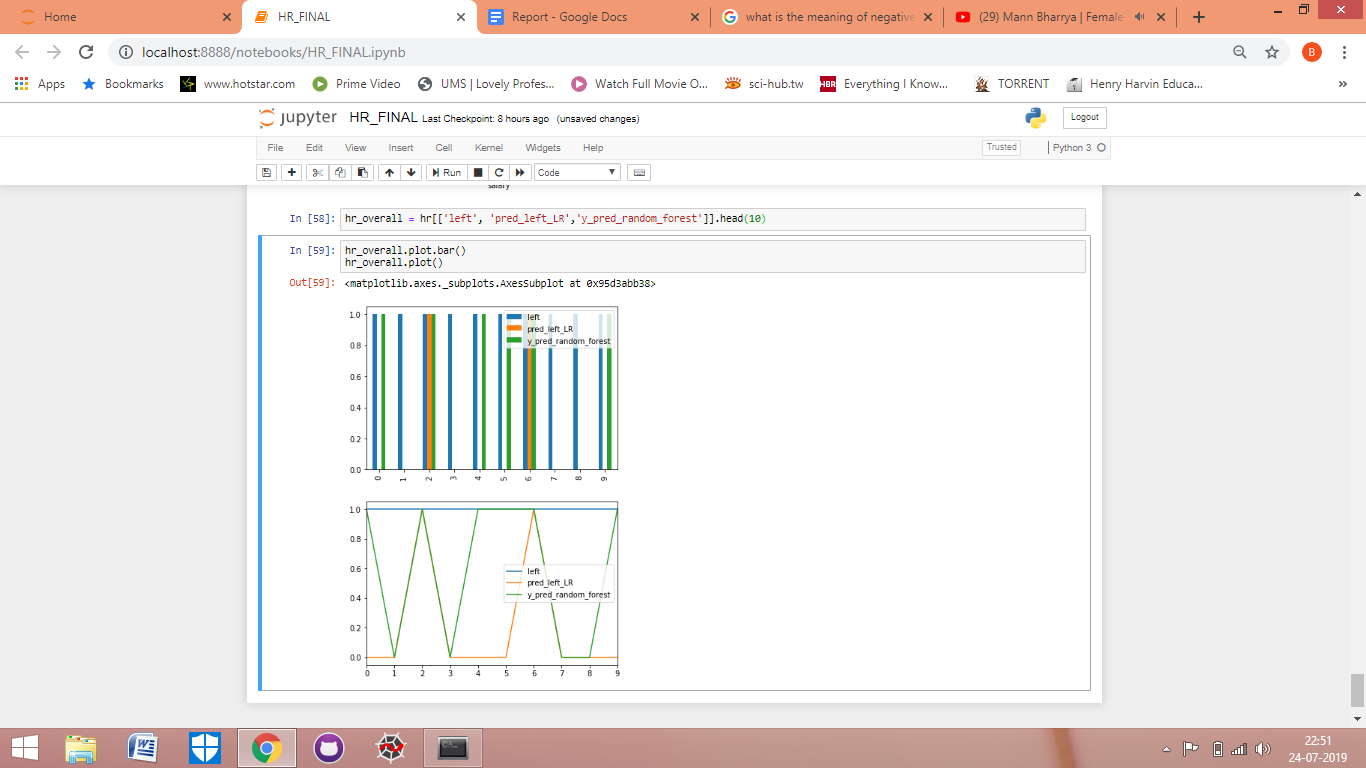


1. Here I plotted graphs of actual left and predicted left (by random forest) on the basis of departments and salary respectively using groupby function.





1. Now I created another subset by the name of hr\_overall using three columns which are actual left, predicted left ( logistic regression) and predicted left (random forest) for more comparisons.



## 

## 

## 