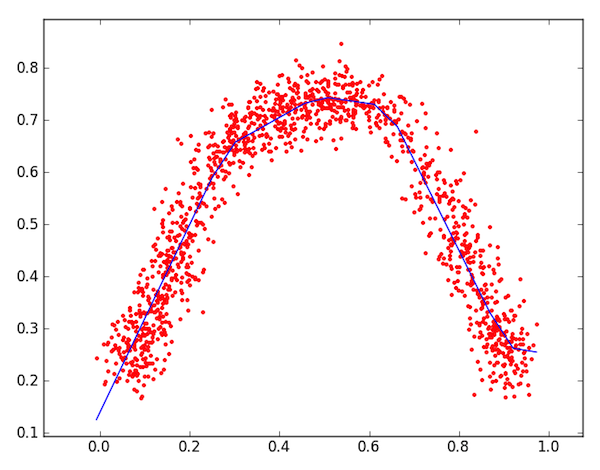
**What is Regression Analysis?**

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.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Regression_Line.png)

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.

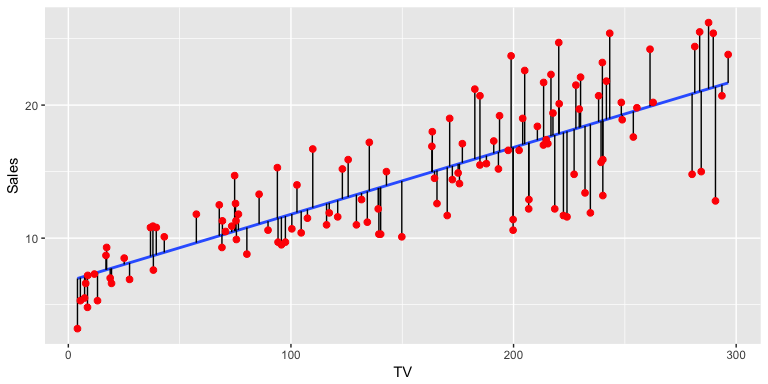
There are various kinds of regression techniques available to make predictions. In this study we have used linear regression and logistic regression.

**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).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression1.png)

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.

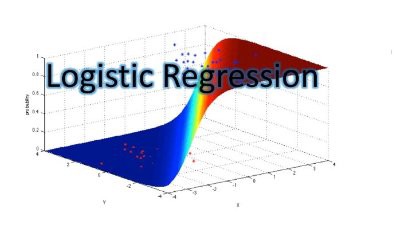
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

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" \t "_blank) 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).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Logistic_Regression.png)

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.

**CAR PRICE PREDICTION MODEL**

**Objective:**

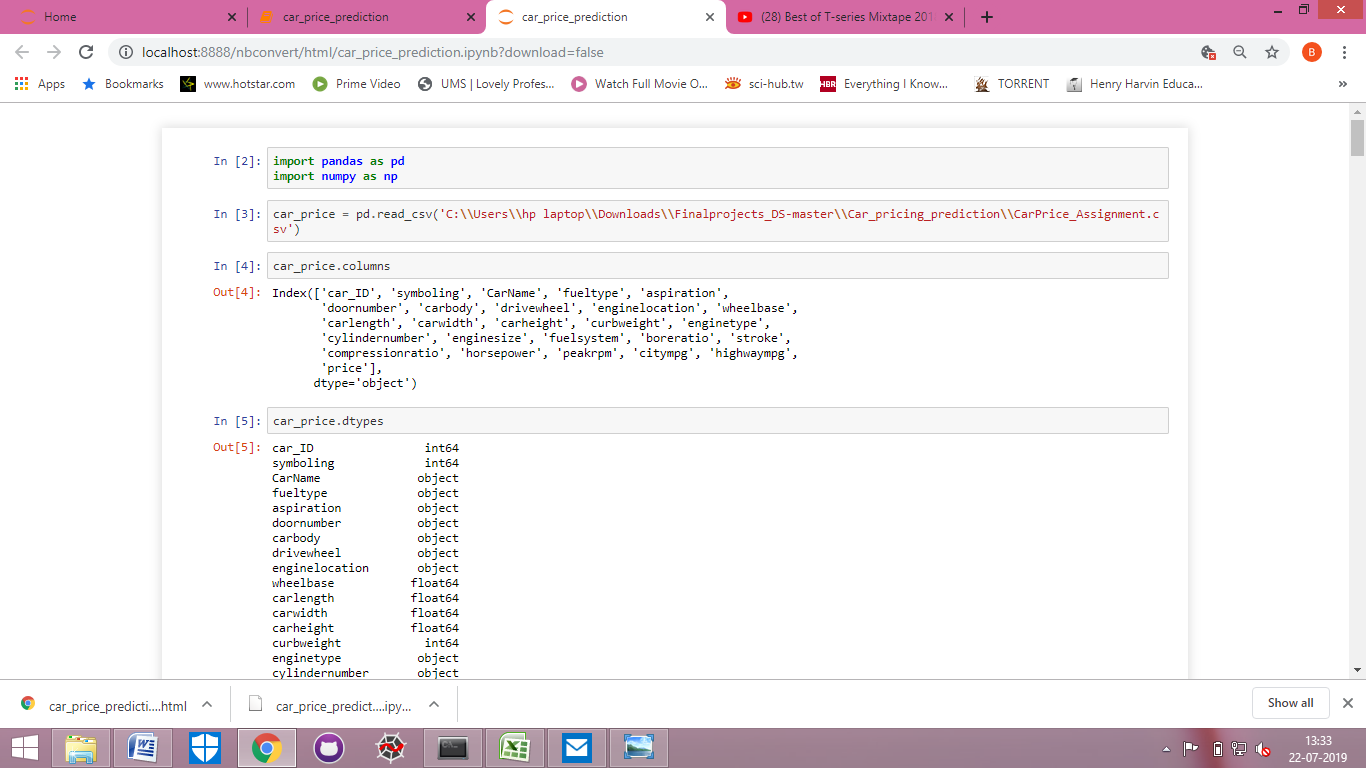
In this study, I have predicted the price of the cars based on the various factors which are given to me in the data. I have used various regression models on the data to find out the best possible model. And through various experiments also found out the factors which influence the price of the cars the most.

**Components of data available:**

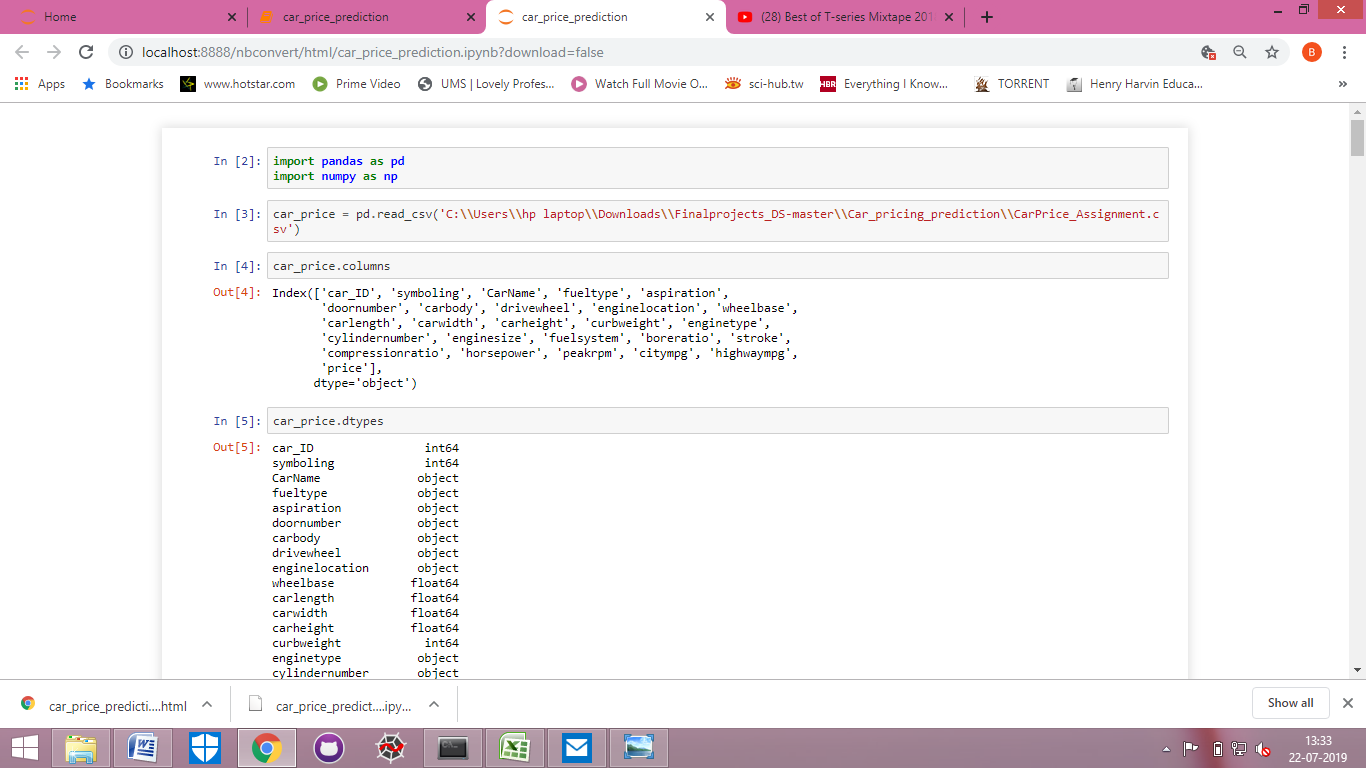
|  |  |  |
| --- | --- | --- |
| **DATA DICTONARY** | | |
|
| 1 | **Car ID** | Unique id of each observation (Integer) |
| 2 | **Symboling** | Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical) |
| 3 | **Car Company** | Name of car company (Categorical) |
| 4 | **Fuel Type** | Car fuel type i.e. gas or diesel (Categorical) |
| 5 | **Aspiration** | Aspiration used in a car (Categorical) |
| 6 | **Door Number** | Number of doors in a car (Categorical) |
| 7 | **Car Body** | body of car (Categorical) |
| 8 | **Drive Wheel** | type of drive wheel (Categorical) |
| 9 | **Engine Location** | Location of car engine (Categorical) |
| 10 | **Wheel Base** | Wheel base of car (Numeric) |
| 11 | **Car Length** | Length of car (Numeric) |
| 12 | **Car Width** | Width of car (Numeric) |
| 13 | **Car Height** | height of car (Numeric) |
| 14 | **Curb Weight** | The weight of a car without occupants or baggage. (Numeric) |
| 15 | **Engine Type** | Type of engine. (Categorical) |
| 16 | **Cylinder Number** | cylinder placed in the car (Categorical) |
| 17 | **Engine Size** | Size of car (Numeric) |
| 18 | **Fuel System** | Fuel system of car (Categorical) |
| 19 | **Bore Ratio** | Bore ratio of car (Numeric) |
| 20 | **Stroke** | Stroke or volume inside the engine (Numeric) |
| 21 | **Compression Ratio** | compression ratio of car (Numeric) |
| 22 | **Horse Power** | Horsepower (Numeric) |
| 23 | **Peak rpm** | car peak rpm (Numeric) |
| 24 | **City mpg** | Mileage in city (Numeric) |
| 25 | **Highway mpg** | Mileage on highway (Numeric) |
| 26 | **Price (Dependent variable)** | Price of car (Numeric) |

**Steps followed in making of Regression Model:**

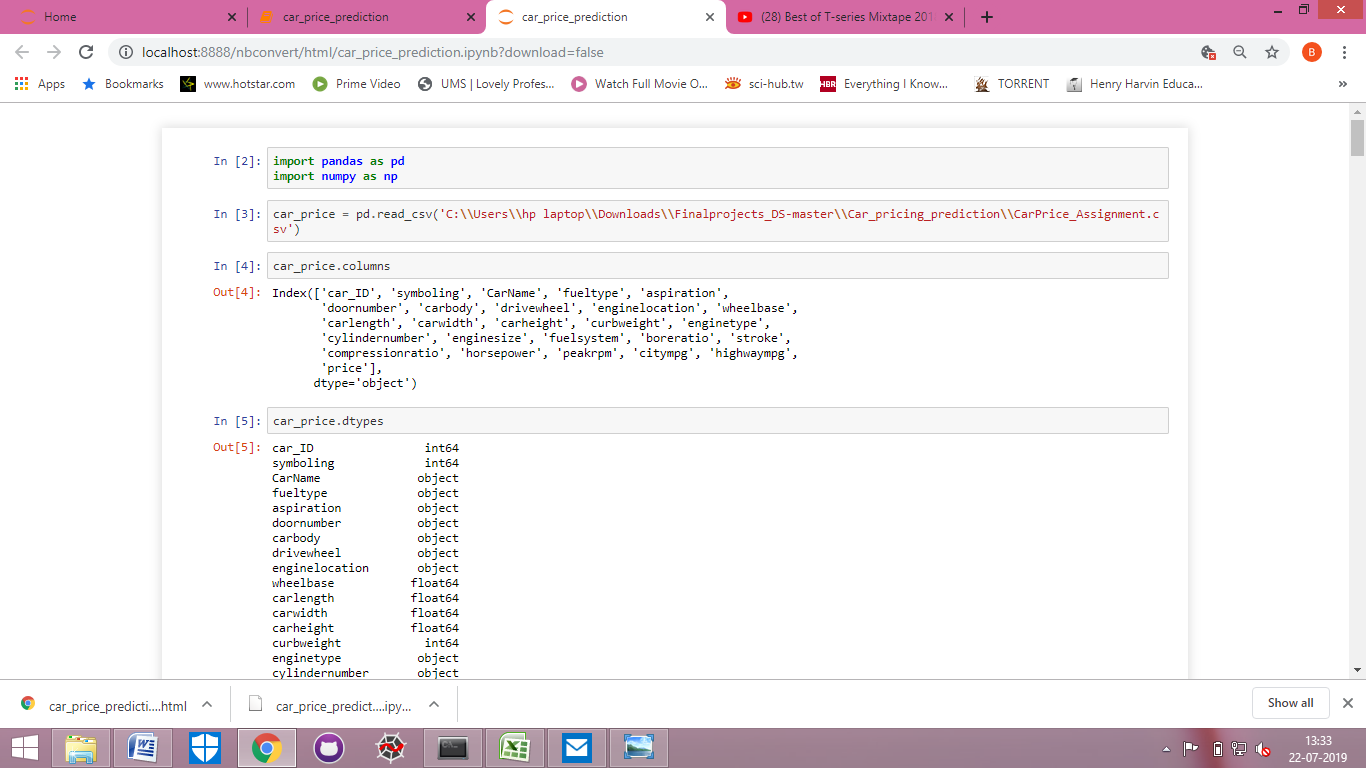
1. Import of various libraries like pandas, matplotlib and numpy.



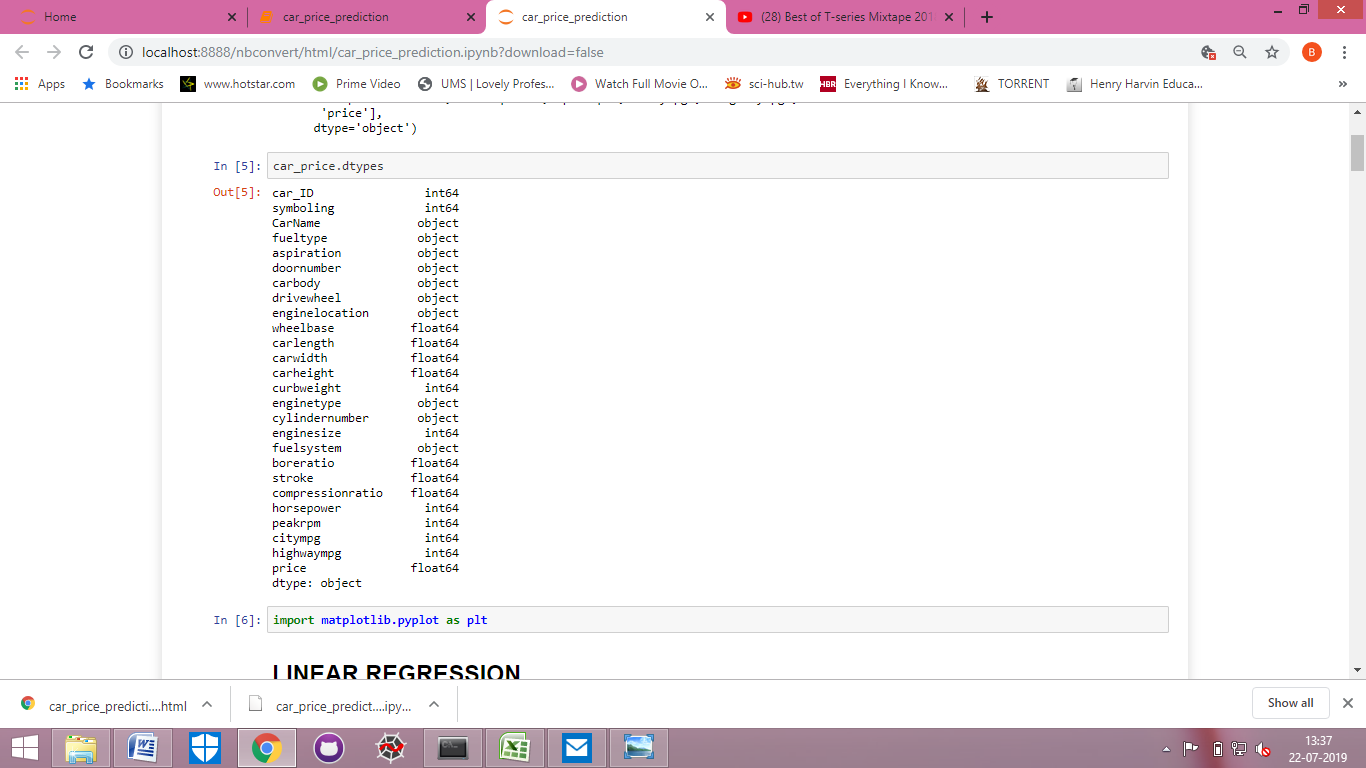
1. After the import of libraries, next step is to import data which is available to me in CSV form .



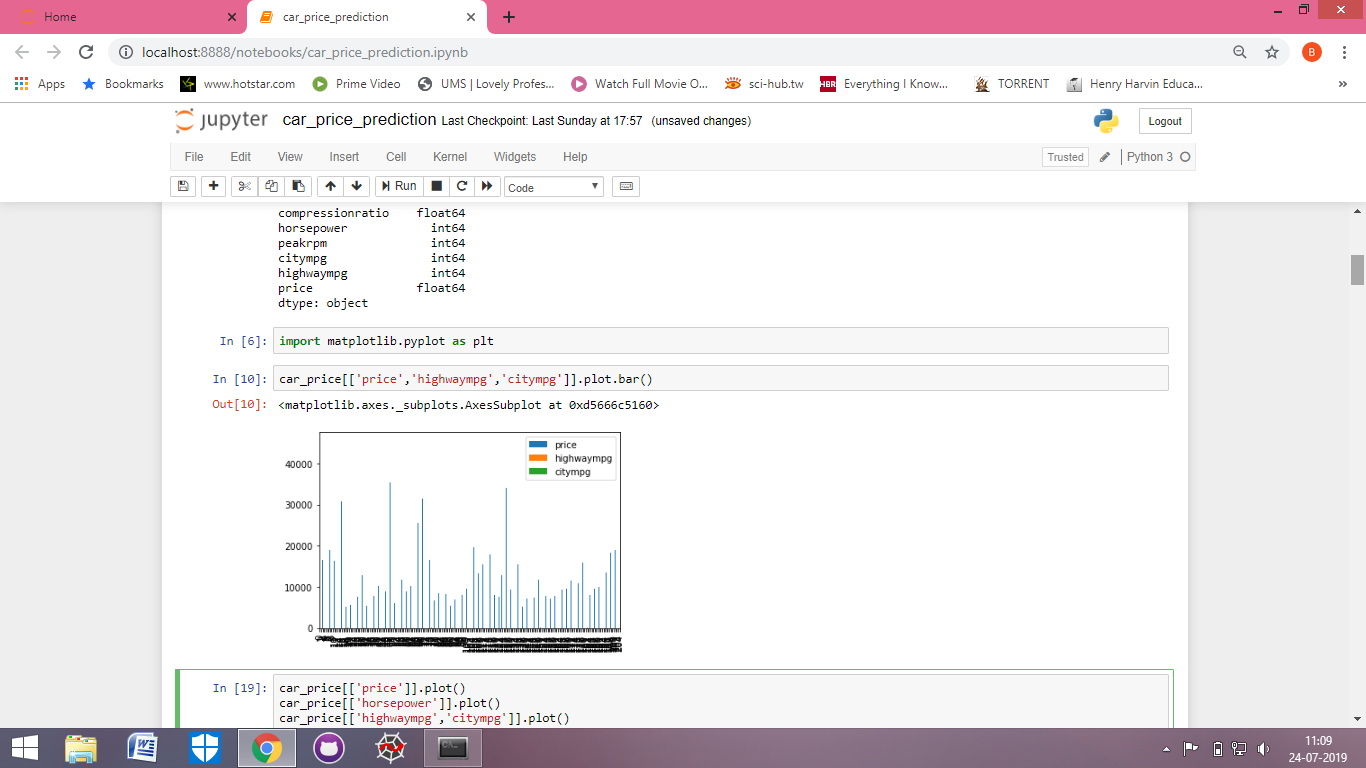
1. Once the data in imported, then i checked the names of the columns given to me in the data.



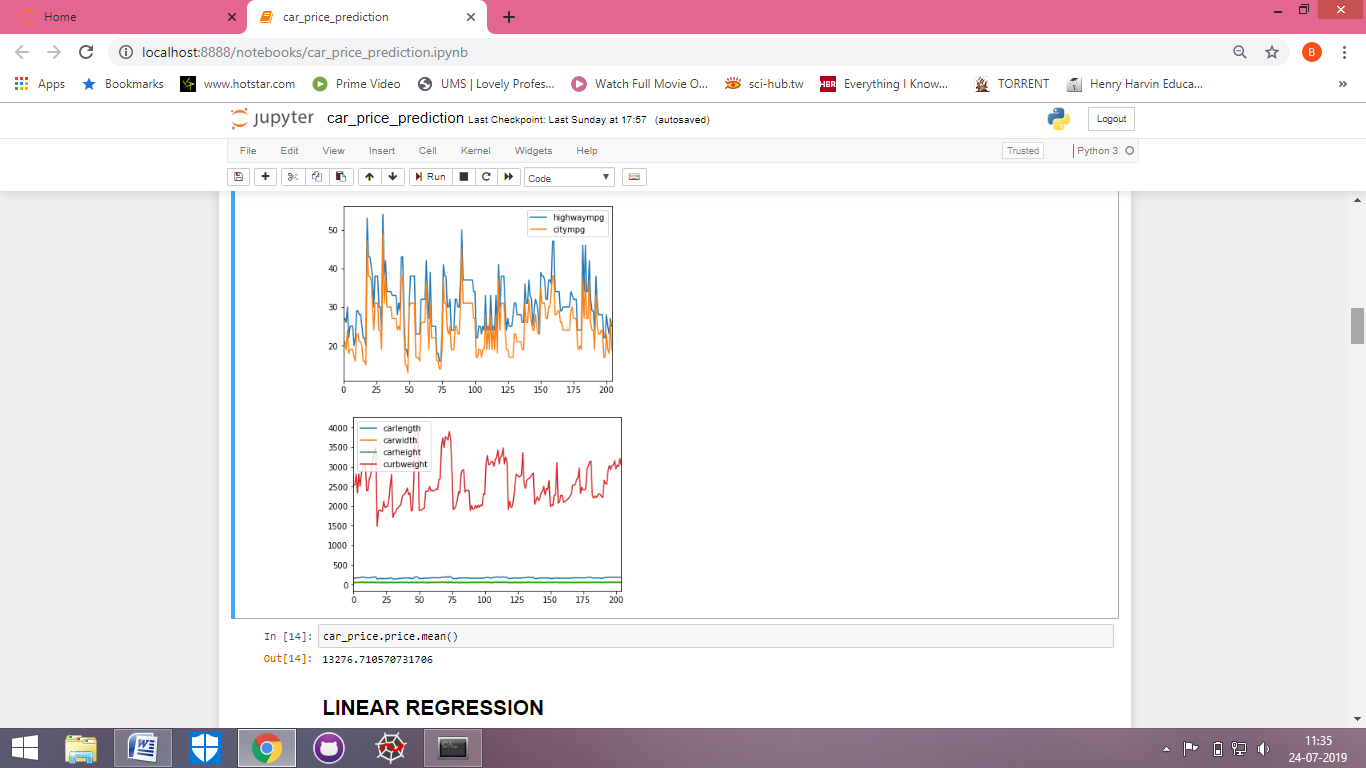
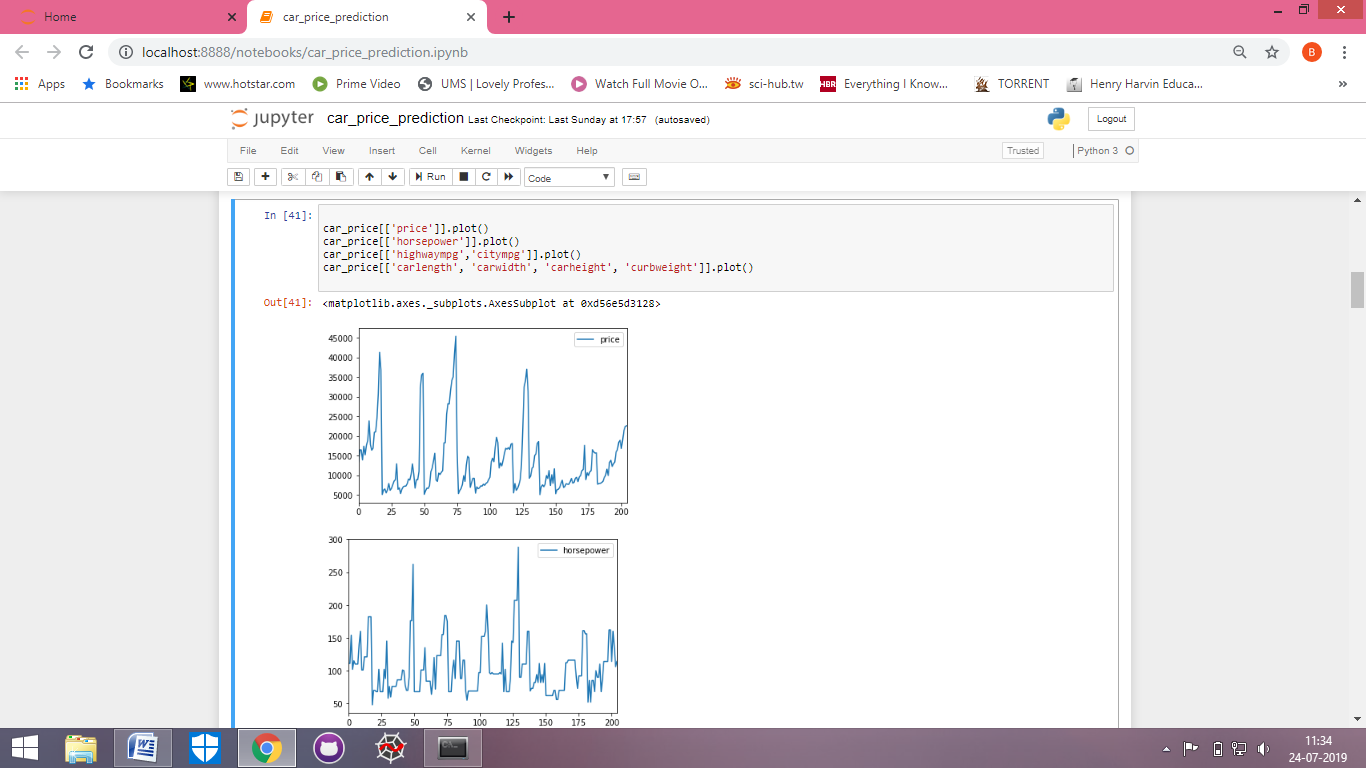
1. After checking out the names of the columns, I checked the type of data given in the each column.



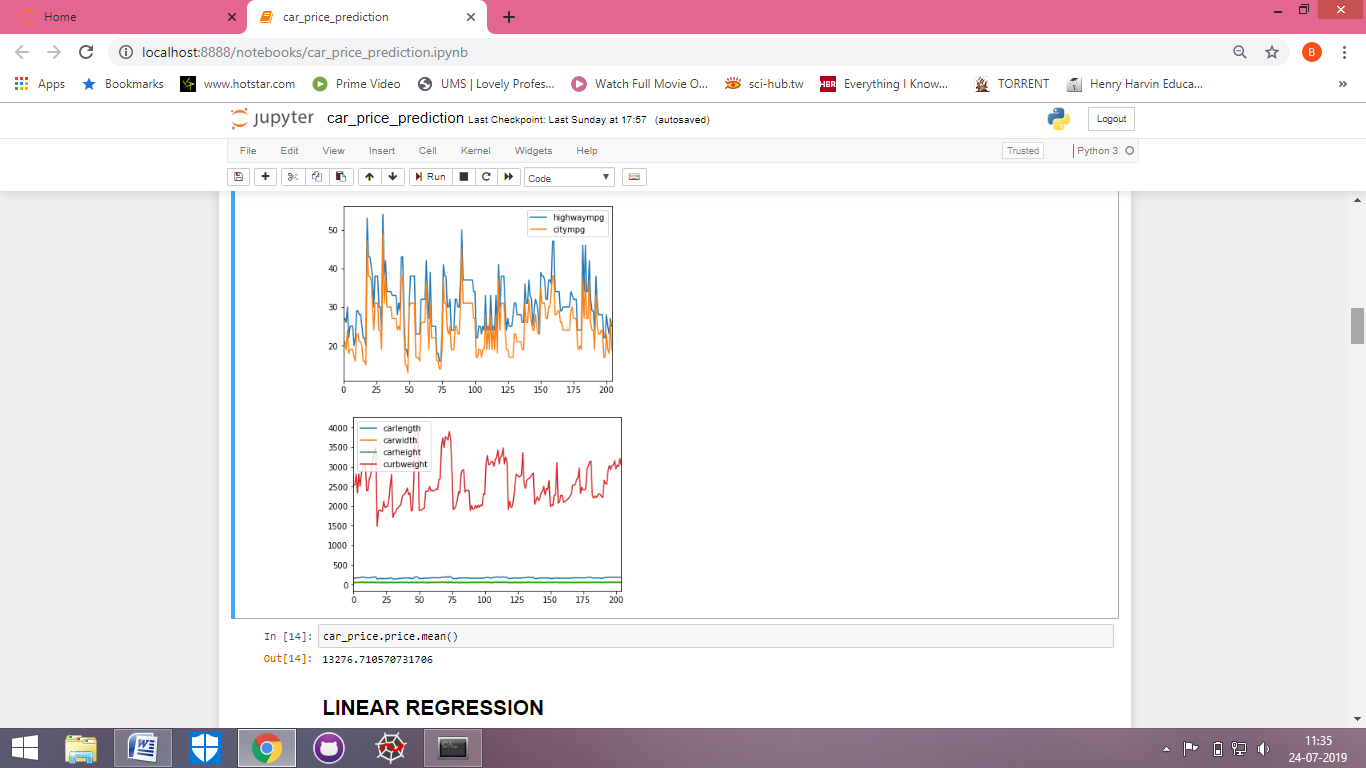
1. Here I have plotted a graph showing the relation between price, citympg and highwaympg.



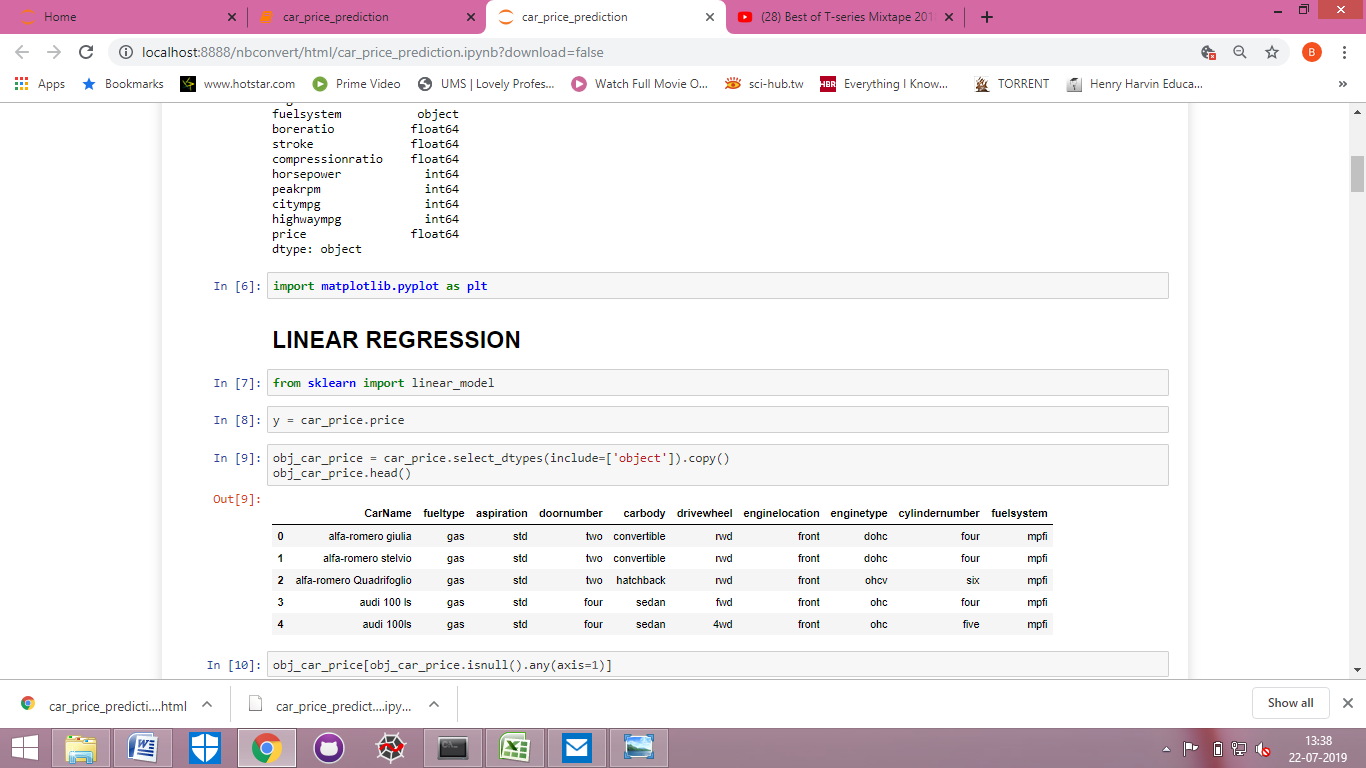
1. Here I have plotted various graphs which individually shows price, horsepower, highwaympg, citympg, and various other features of the cars.



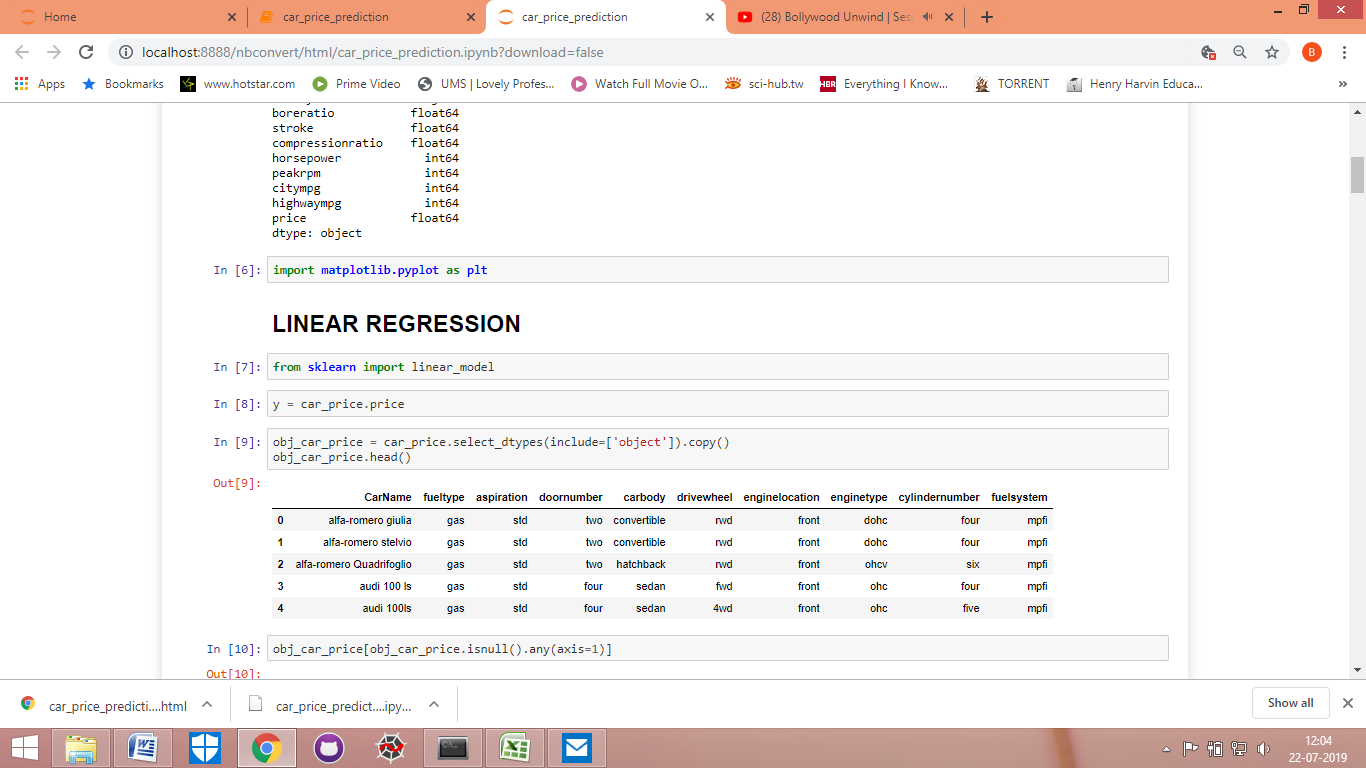
1. Now I have also checked the average price of cars.



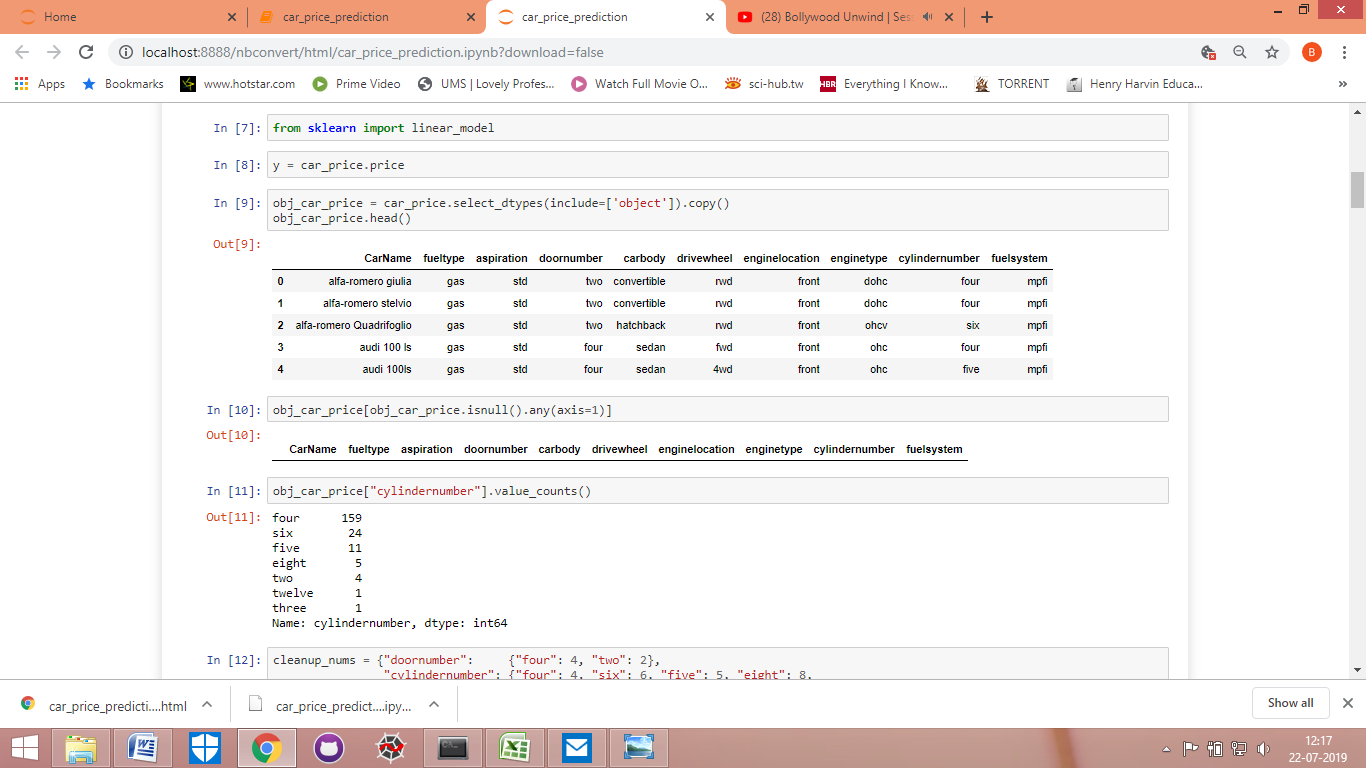
1. After checking the data types, I have imported linear regression model from sklearn library.



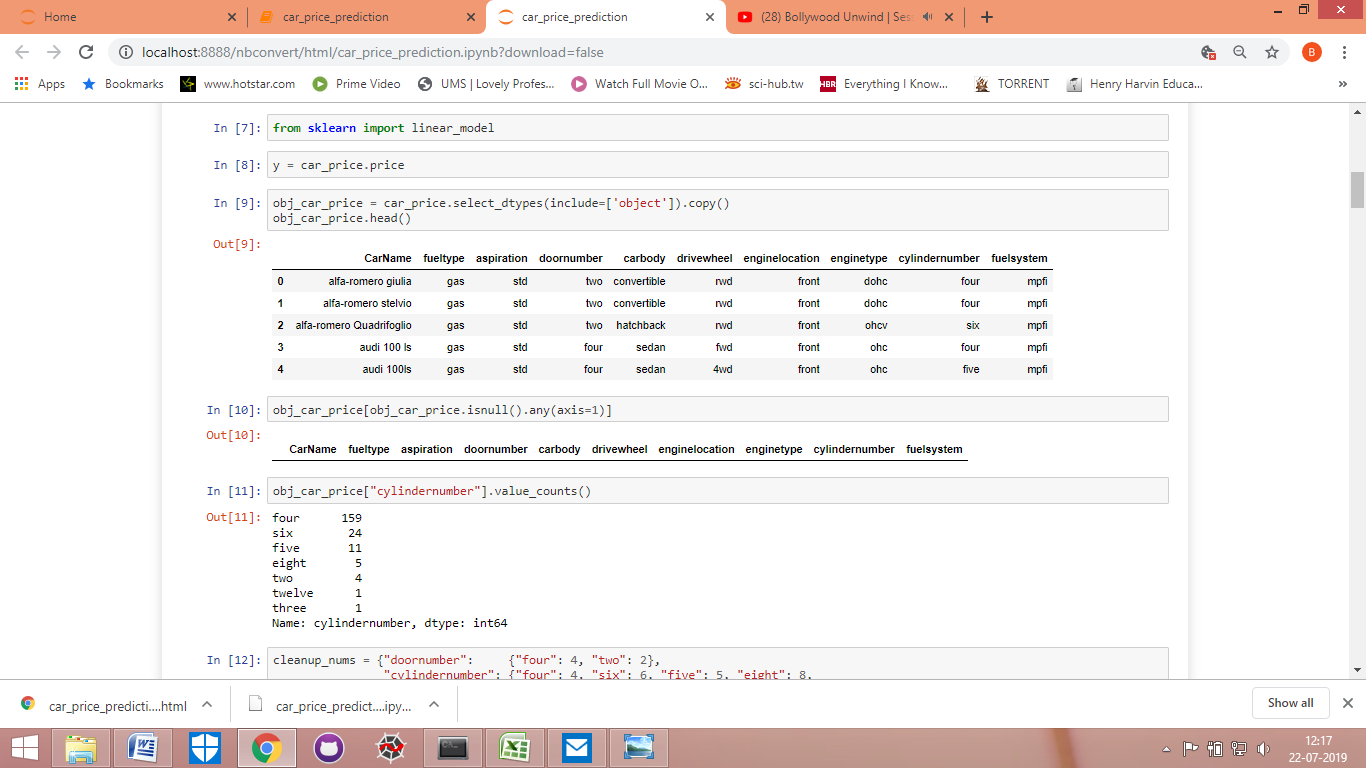
1. Now I have created a separate data set with columns having objects data type. To that data set I have given the name obj\_car\_price.



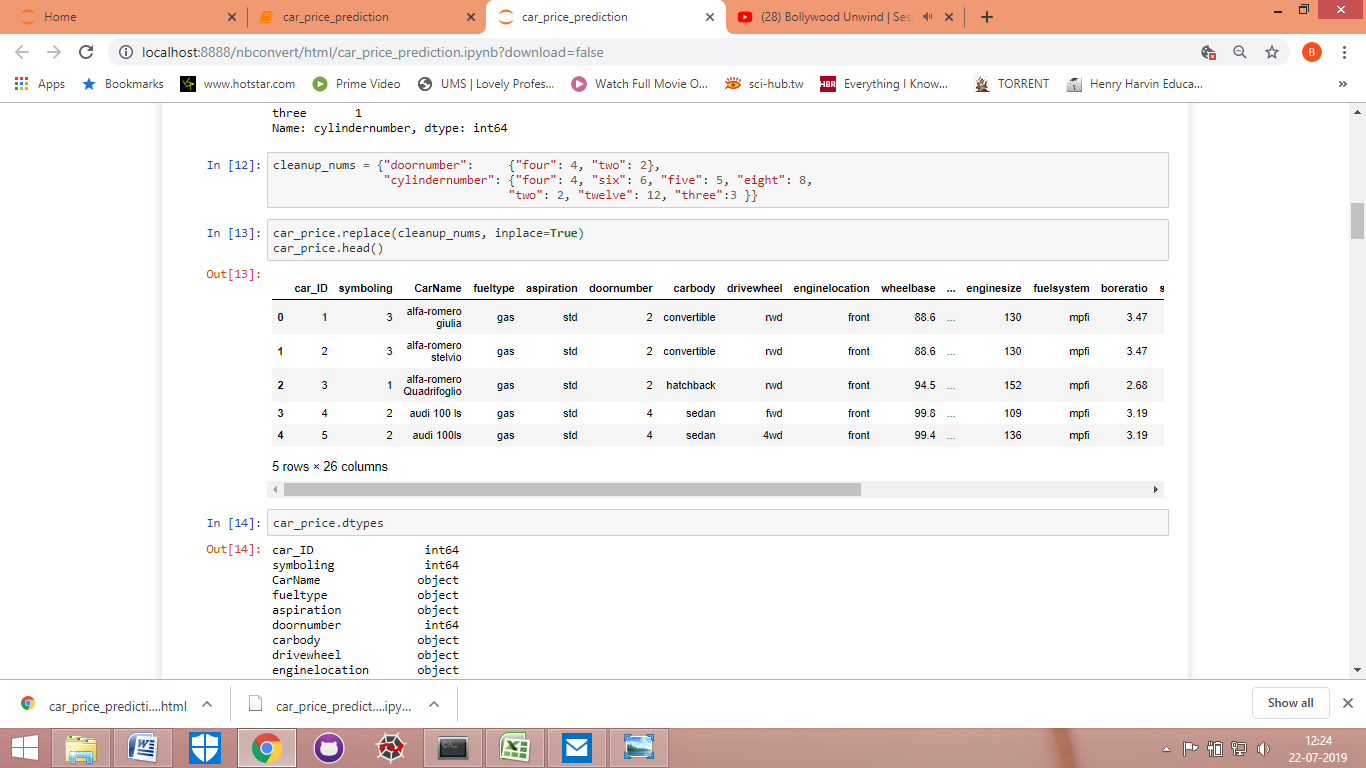
1. After getting the obj\_car\_price data set now I checked if there any NAN values in the data set.



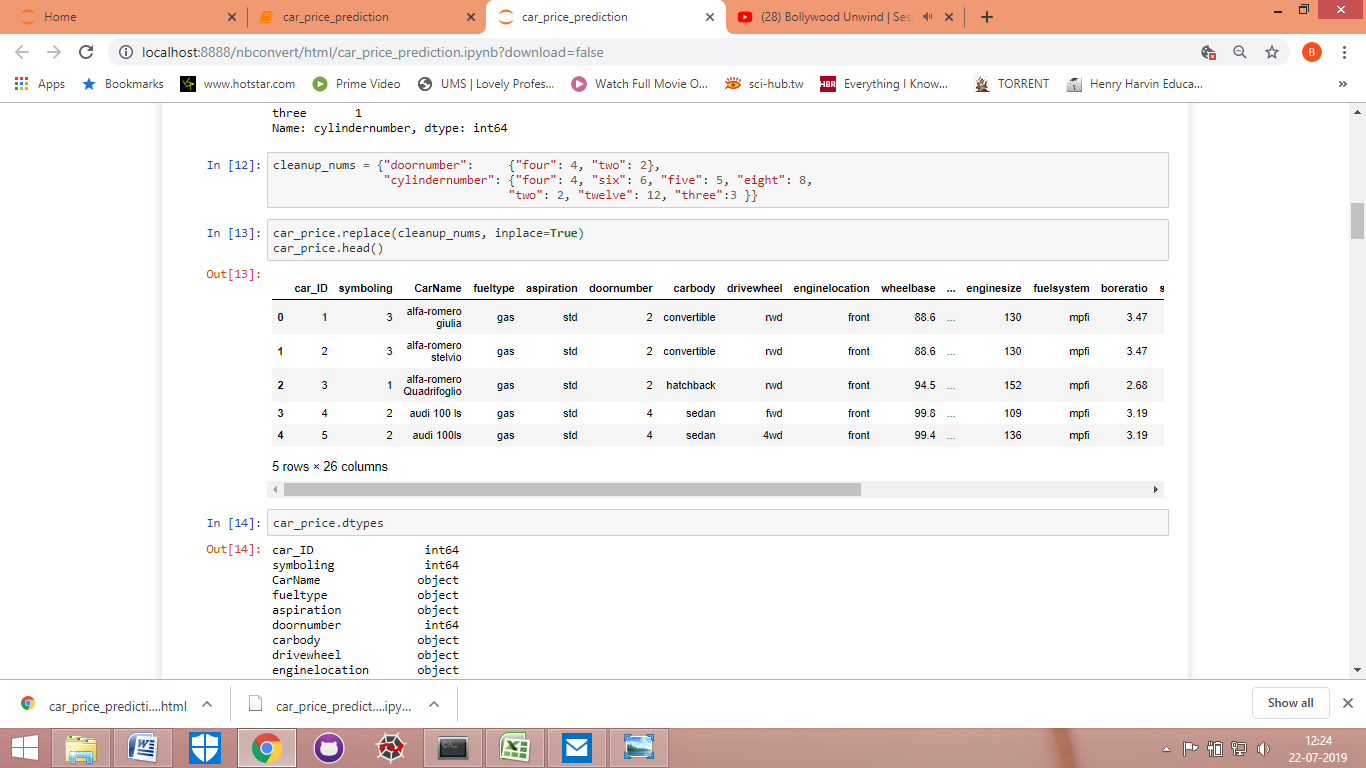
1. As there are no NAN values in the data set, now I checked the value counts in the cylinder number column.



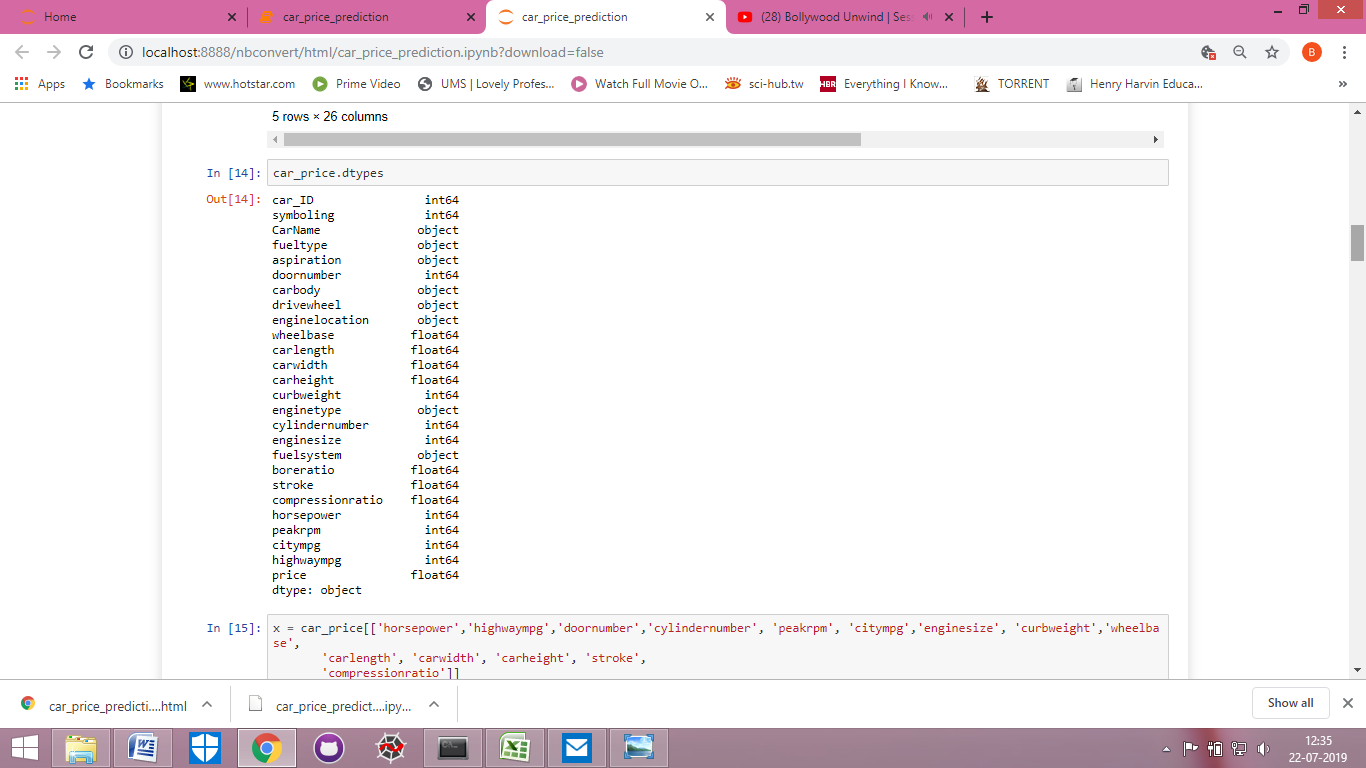
1. Now in this step I have created a dictionary for door number and cylinder numbers. So that to convert both these data types from object to integer.



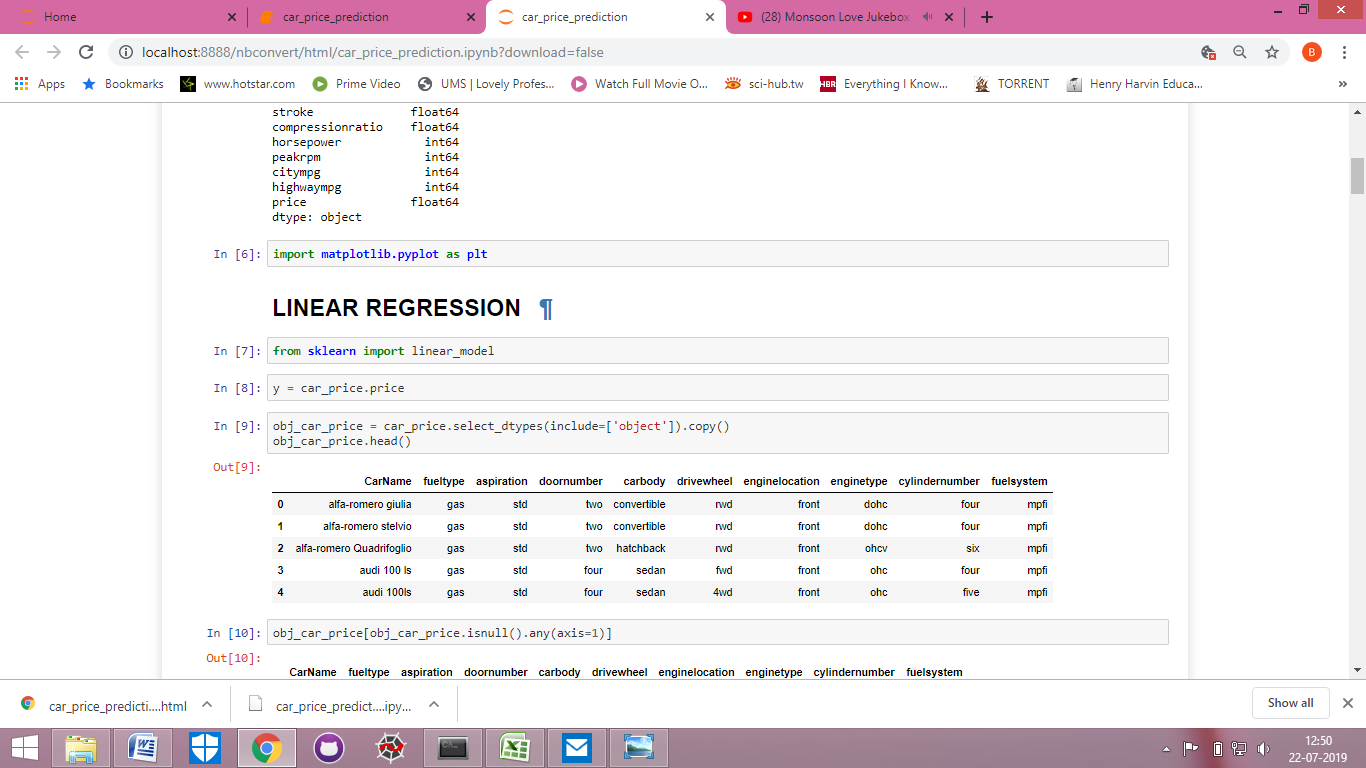
1. After creating the dictionary, I have replaced it in the original data set which is by the name of car\_price.



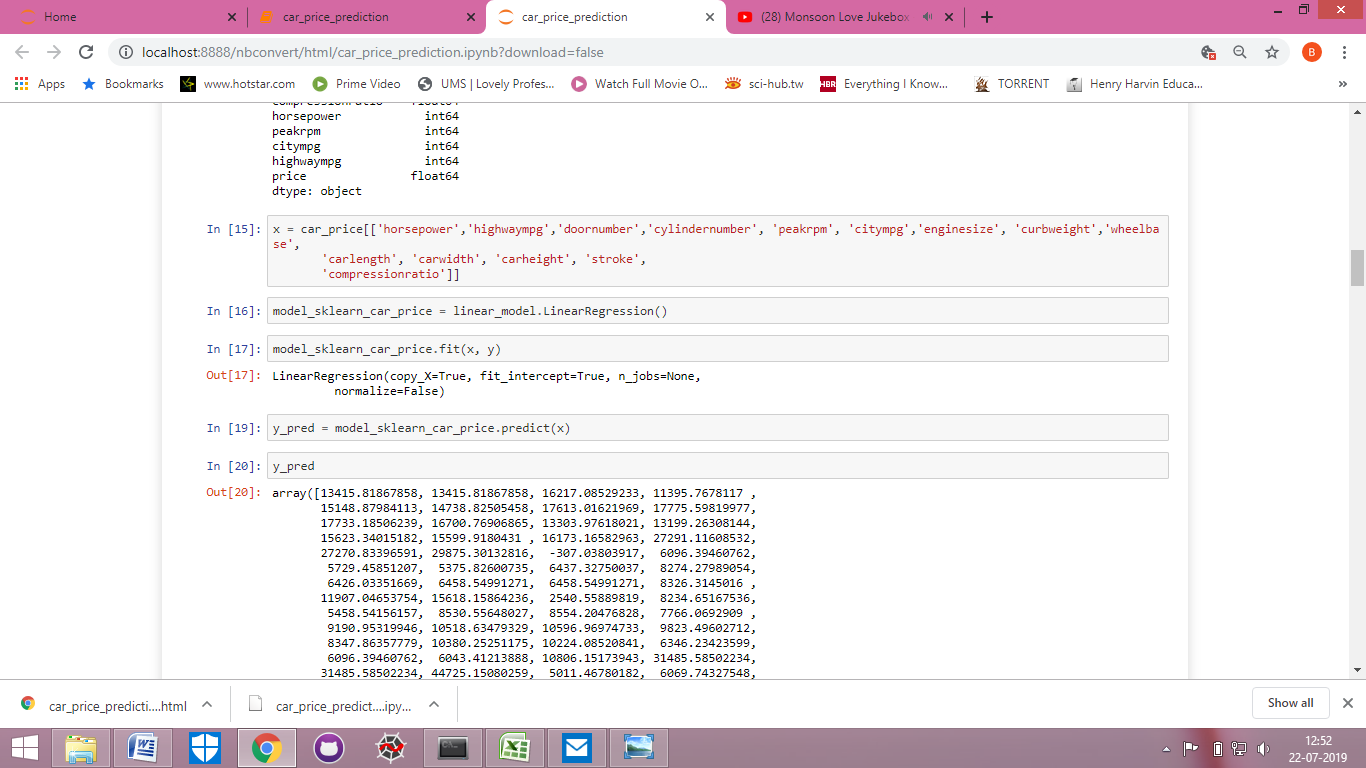
1. Now after changing the data type, I check it again whether the data type has changed or not. In this output we can see that the data type of door number and cylinder number has been changed.



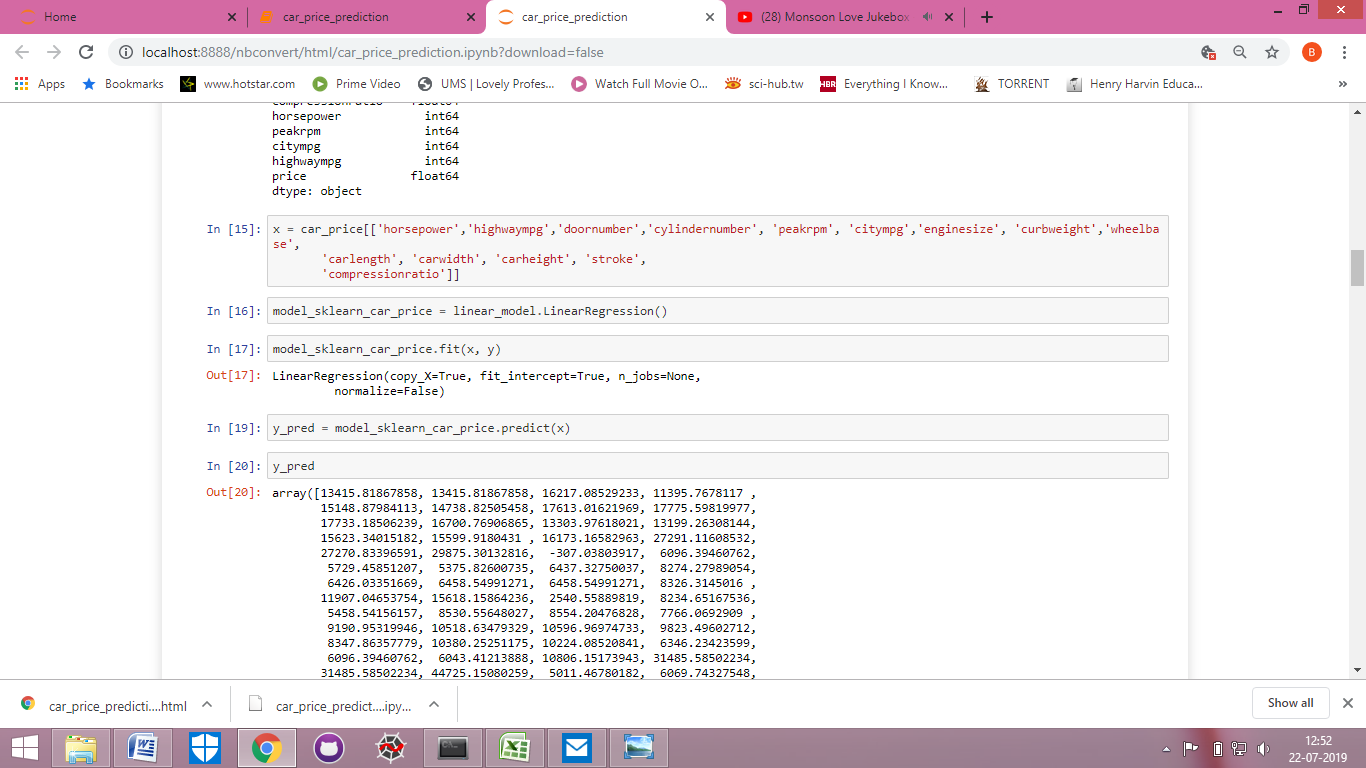
1. Now I assign the price column to y. As price is according to me the target or dependent variable which I have to predict.



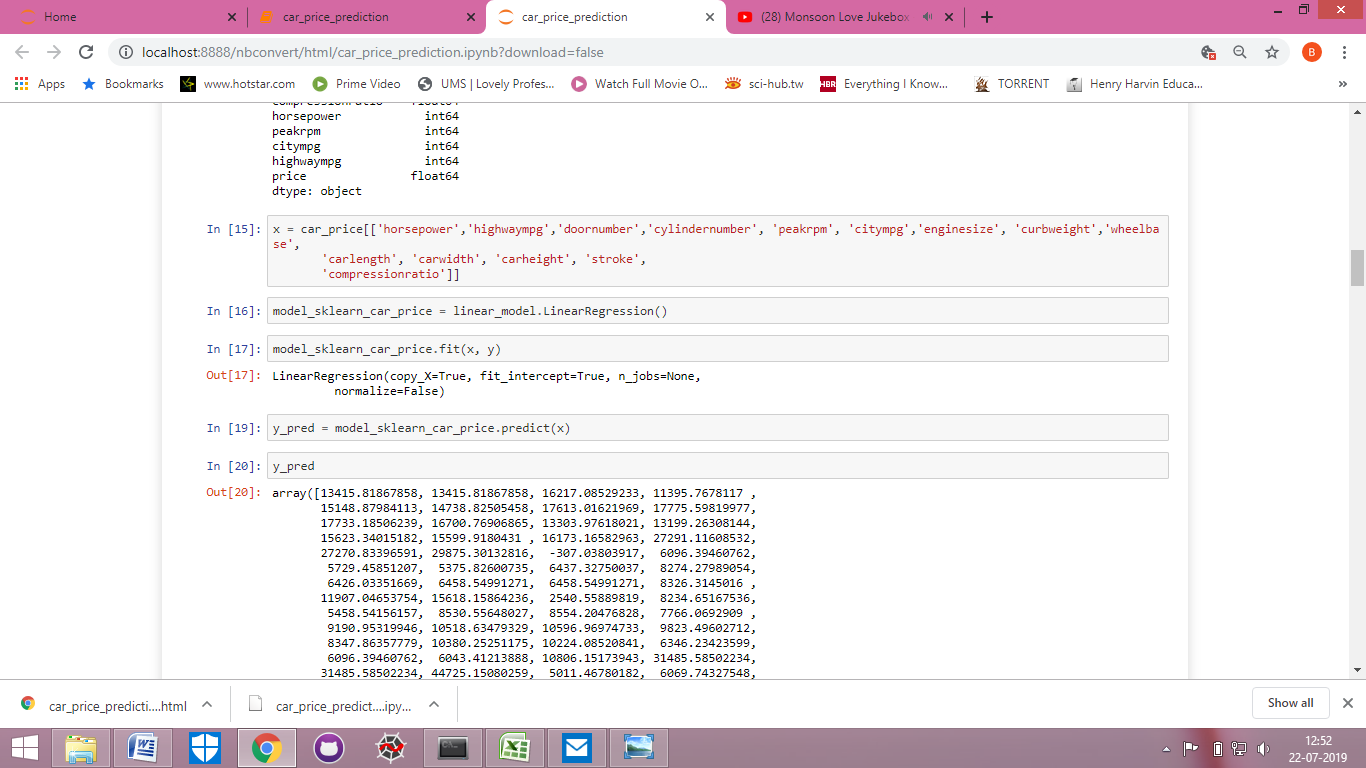
1. And here in this step I assign independent variables to x. I have selected those variables which according to me are having the maximum effect on the price.



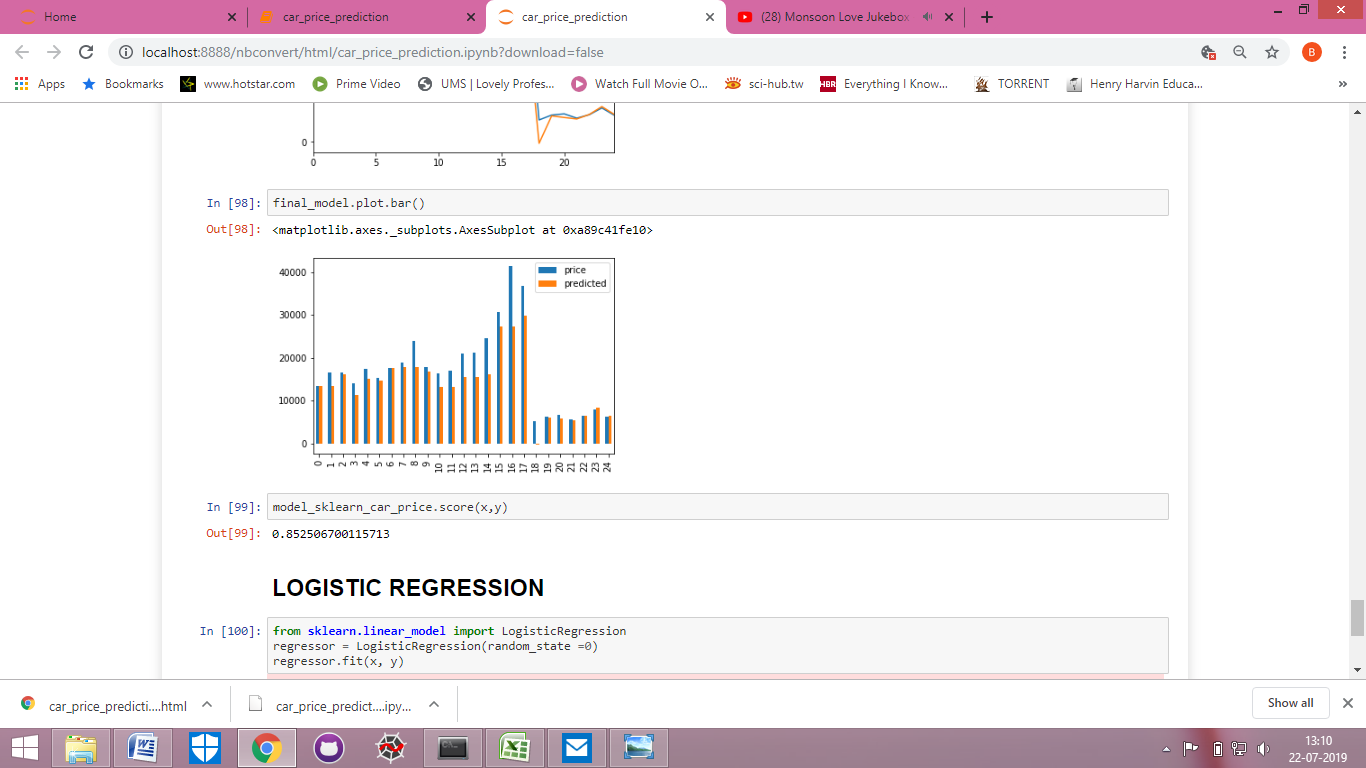
1. Now I prepare a linear regression model by the name of model\_sklearn\_car\_price.



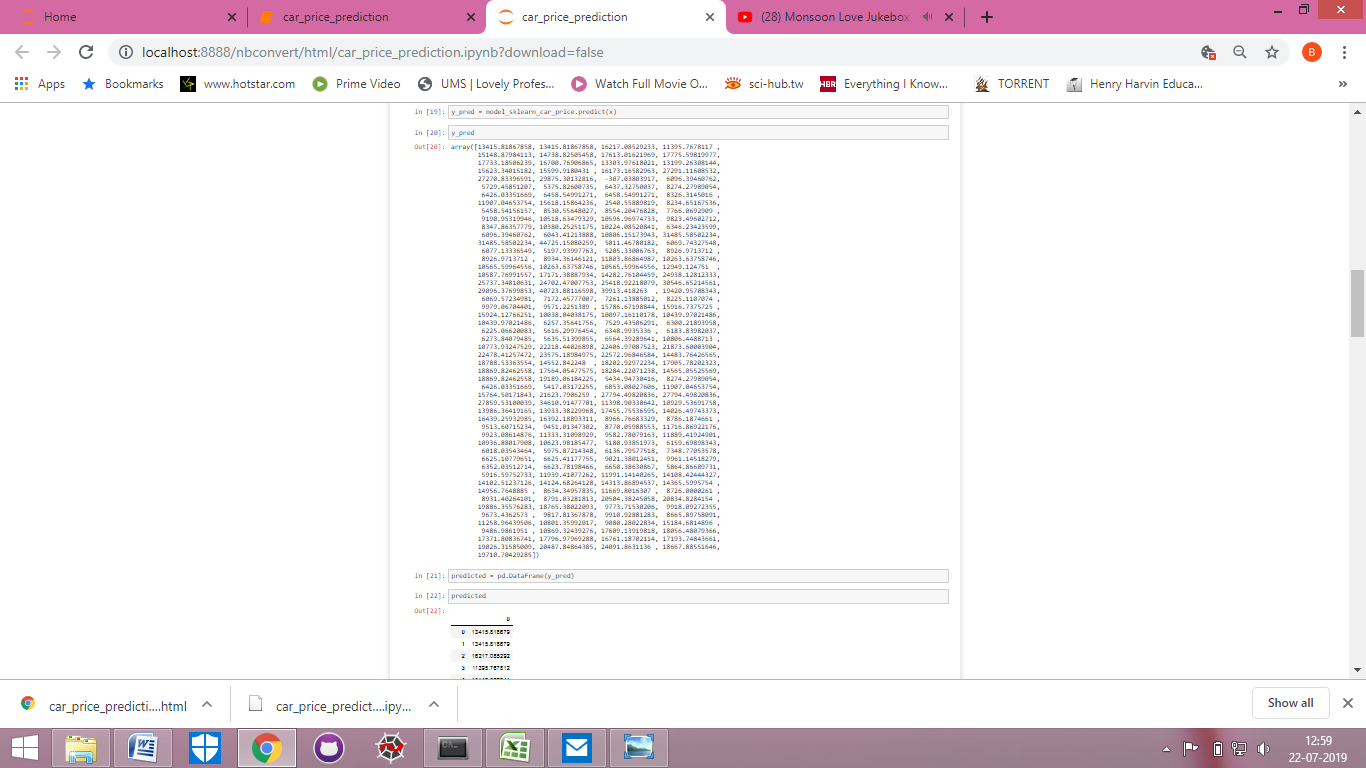
1. After preparing a model, now I have fitted x, y in the model.



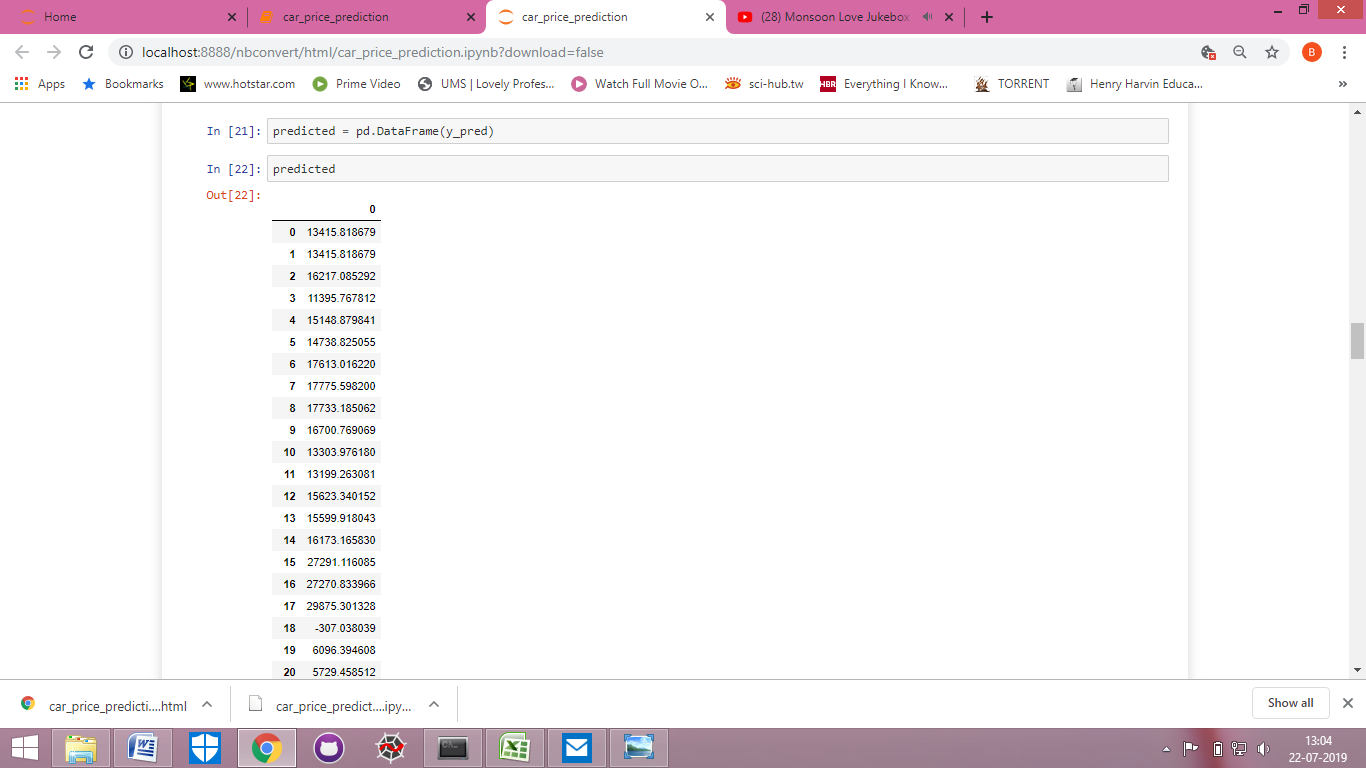
1. Now I check the score of my regression model.



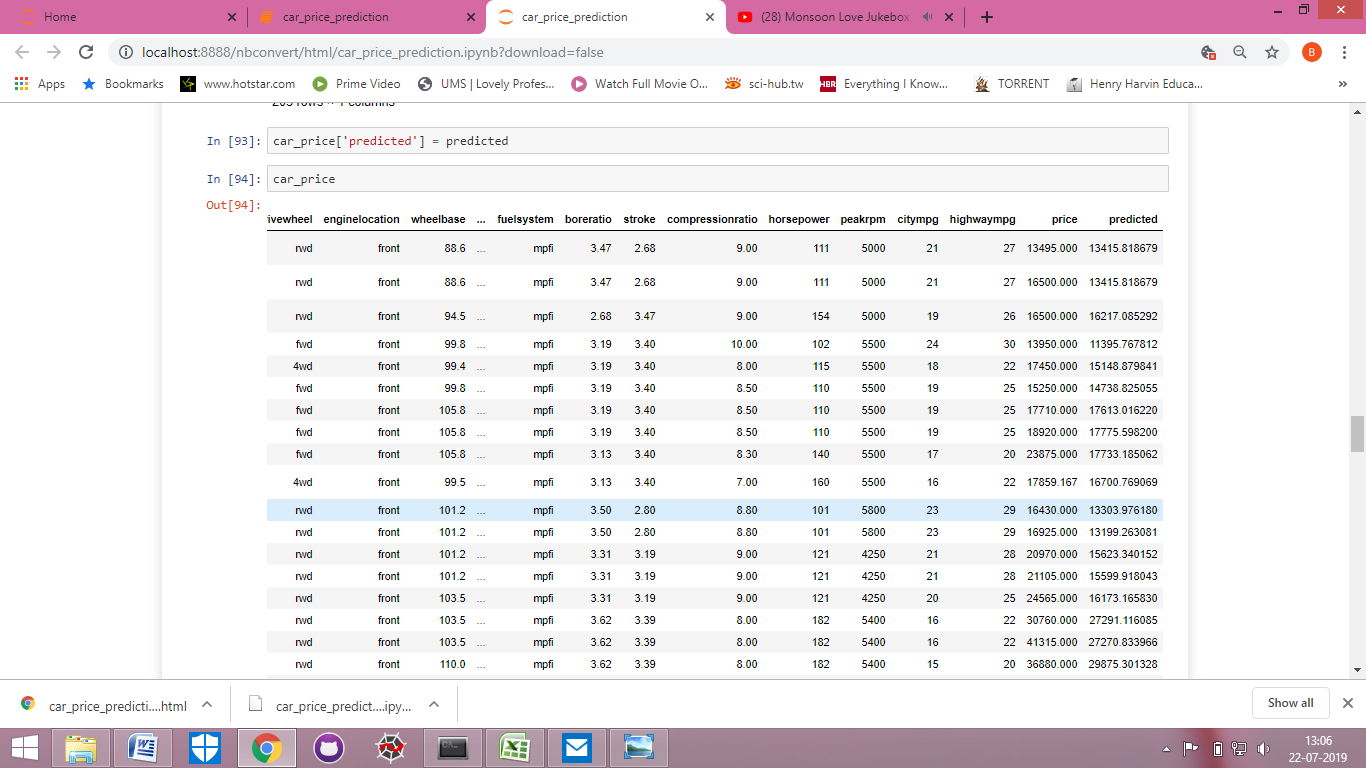
1. After fitting the x, y, now I predict the price columns and see the results through two commands. The result is shown in the form of array.



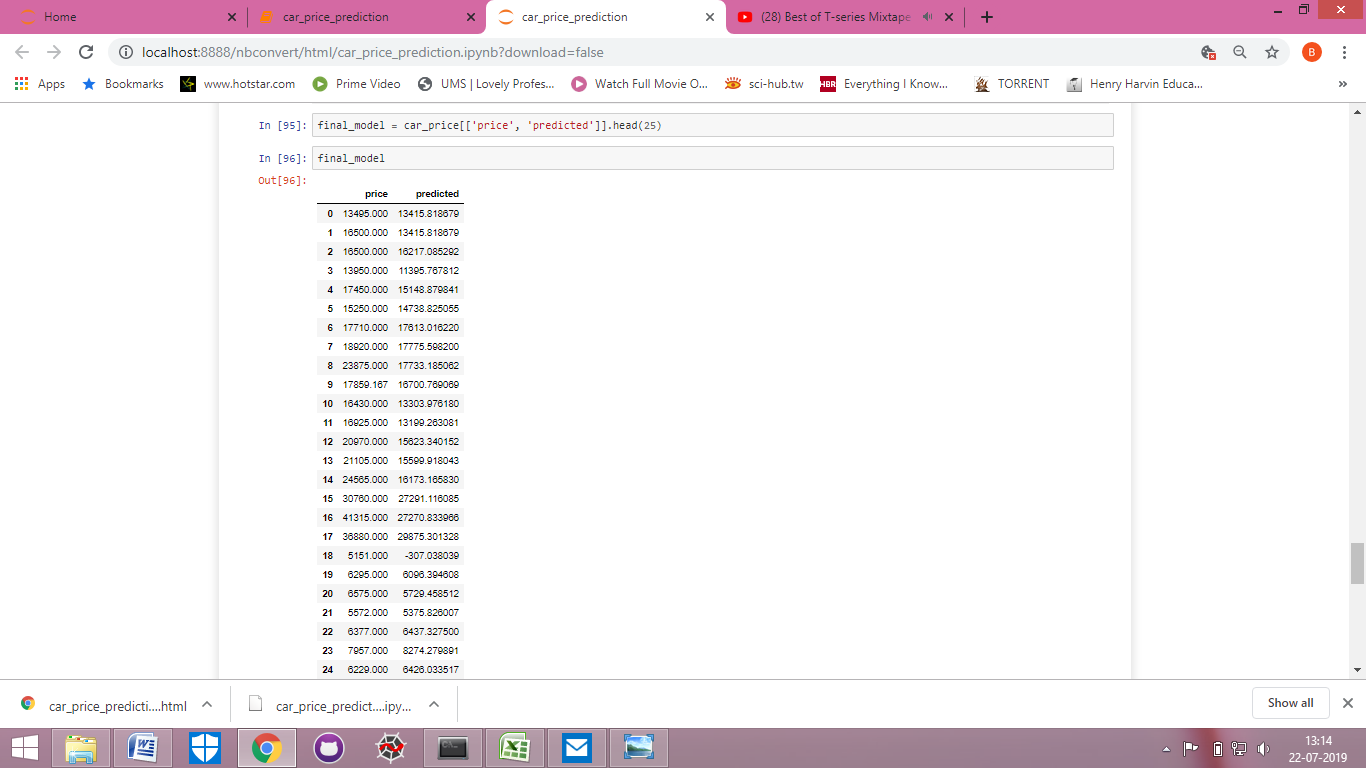
1. Now for better understanding, I have converted the predicted results from array to data frame form. And named it as predicted.



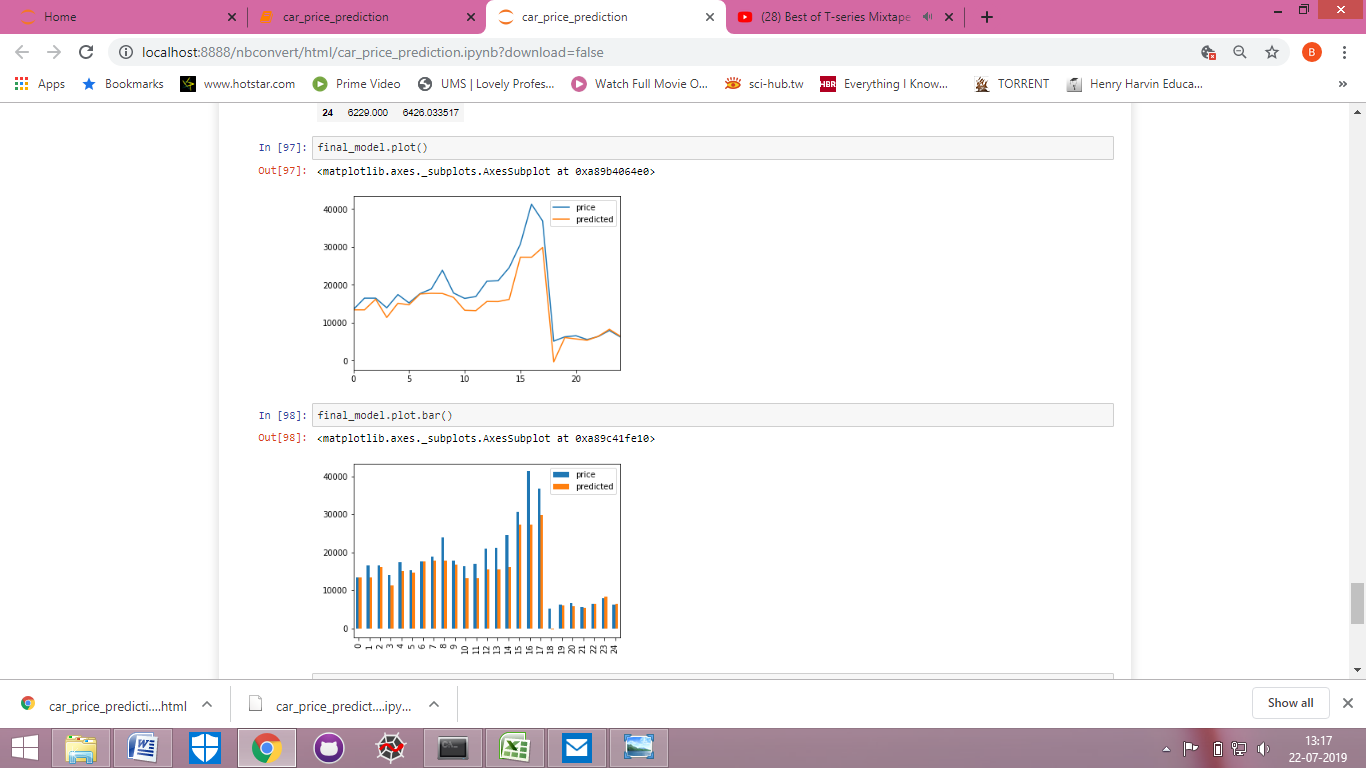
1. Now after converted the predicted results in the data frame form, I have added that predicted column in the original data set named car\_price.



1. Now I have separated just the actual price and predicted price columns from the data set. And just seeing the first 25 rows right now.



1. Now I have plotted the data in the form of line and bar graphs.



1. After completing my linear regression model, I checked once if logistic regression model can also be used but as the data given in the continuous form, therefore logistic regression model cannot be followed.

