**PROJECT REPORT**

Of

**DATA MINING AND ASSOCIATION**

Practical File submitted in partial fulfilment of the requirements for the award of

Bachelor of Engineering IN

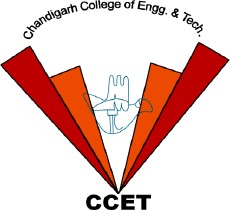
COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Agampreet kaur , Gursimar Singh , Gurdaman Singh

(Roll no: CO17307, CO17325,CO17323)

Under the Guidance of: Dr. Varun Gupta



**CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)**

Government Institute under Chandigarh (UT) Administration, Affiliated to Panjab University,

Chandigarh

Sector-26, Chandigarh. PIN-160019

**Jan – May, 2020**

****

**CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)**

Government Institute under Chandigarh (UT) Administration | Affiliated to Panjab University , Chandigarh Sector-26, Chandigarh. PIN-160019 | Tel. No. 0172-2750947, 2750943

Website: [www.ccet.ac.in |](http://www.ccet.ac.in/) Email: [principal@ccet.ac.in |](mailto:principal@ccet.ac.in) Fax. No**. :**0172-2750872

## Department of Computer Sc. & Engineering

**ACKNOWLEDGEMENT**

It is a great pleasure to present this Project report of data mining . We have taken efforts in this Lab. However, it would not have been possible without the kind support and help of our teacher Dr. Varun Gupta. We would like to extend my sincere thanks to him.

We are highly indebted to Chandigarh College of Engineering & Technology (Degree Wing) for their guidance and constant supervision as well as for providing necessary information regarding the practical & also for their support in completing the practical file.

They taught us all the basic concepts required for the practical and guided me through each step of building the programs whenever we were stuck.

We would like to express our special gratitude and thanks to institution (C.C.E.T.) persons for giving us such attention and time.

We would also like to thank the University for including this DMA Lab as a part of our curriculum.

**INDEX**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.**  **NO.** | **EXPERIMENT** | **PAGE NO.** | **DATE** | **REMARK** |
| **1** | INTRODUCTION | 1-2 |  |  |
| **2** | DATA COLLECTION | 3- 4 |  |  |
| **3** | DATA CLEANING | 5-6 |  |  |
| **4** | FEATURE ENGINEERING | 7-8 |  |  |
| **5** | OUTLIER REMOVAL | 9-10 |  |  |
| **6** | MODEL BUILDING | 11-12 |  |  |
| **7** | CONCLUSION | 13-14 |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

DATA SCIENCE PROJECT (HOUSING PRICE PREDICTION)

INTRODUCTION

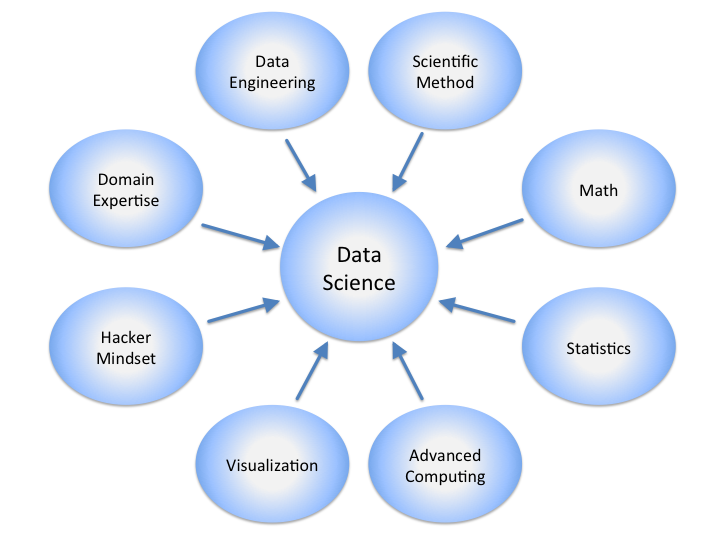
DATA SCIENCE is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data. Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution. Data science is the process of deriving knowledge and insights from a huge and diverse set of data through organizing, processing and analysing the data. It involves many different disciplines like mathematical and statistical modelling, extracting data from its source and applying data visualization techniques. Often it also involves handling big data technologies to gather both structured and unstructured data. Below we will see some example scenarios where Data science is used. Terms like artificial intelligence, [machine learning](https://www.digitalvidya.com/blog/introduction-to-machine-learning/), big data, and deep learning are often used interchangeably with data science in the general vocabulary. However, these are different areas that contribute to data science.

(i) Artificial intelligence focuses on creating machines that can think and behave as humans do.

(ii) Machine learning creates tools that extract useful information from data.

(iii) Big data deals with systems and tools that can handle tremendous amounts of data.

(iv) Deep learning focuses on creating multi-layered neural networks to work on more advanced algorithms than machine learning.



STEP 1- DATA COLLECTION

First step in any project is the collection of resources. Data collection is the process of gathering information on targeted variables.  A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information. In this project we have taken our dataset from Kaggle website (this website consists of various databases), we use Bengaluru house price data to train our data for this prediction. Our database is downloaded in csv format and then imported in the jupyter notebooks. We can perform various operations on our data. For our project we use pandas library.

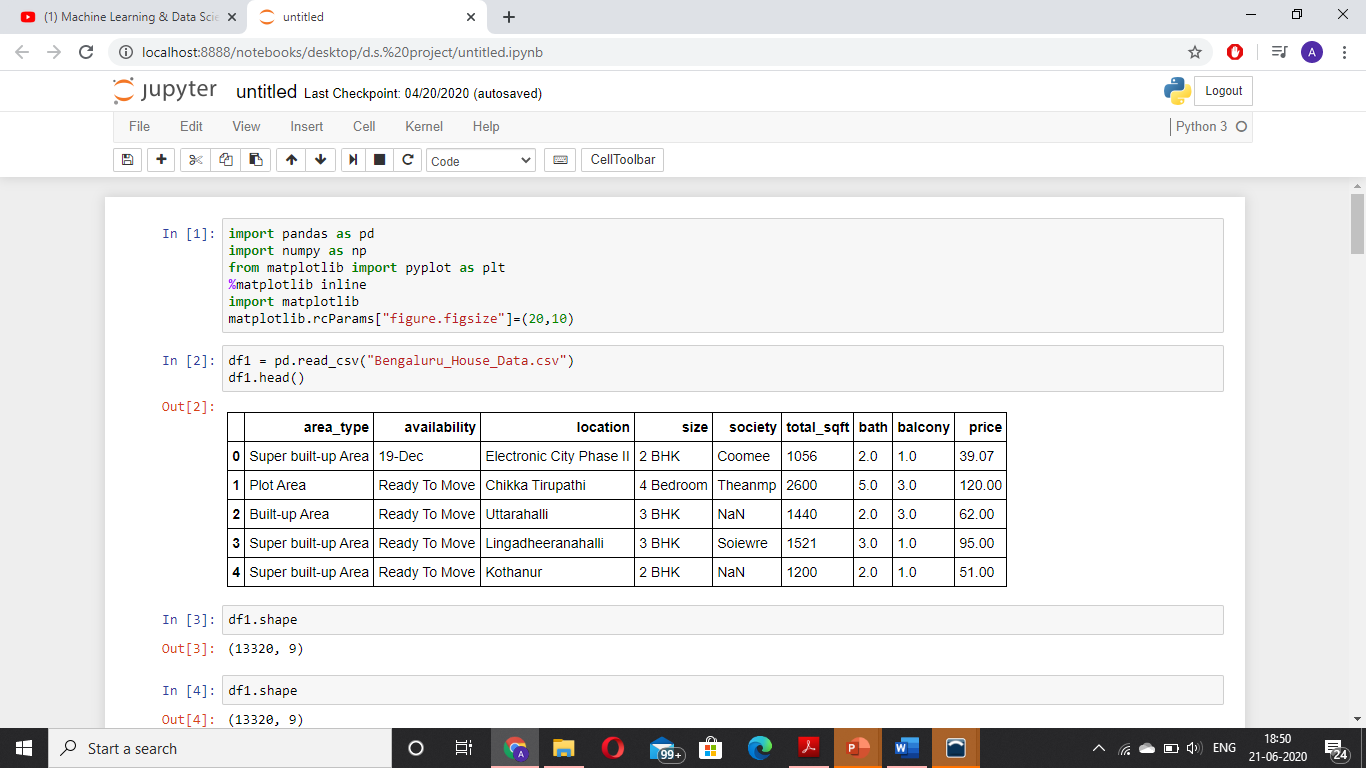


Fig 1.1 (showcasing few entries of database and size of database)

STEP 2- DATA CLEANING

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analysing data because it may hinder the process or provide inaccurate results. There are several methods for cleaning data depending on how it is stored along with the answers being sought. Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set’s accuracy without necessarily deleting information.

For one, data cleaning includes more actions than removing data, such as fixing spelling and syntax errors, standardizing data sets, and correcting mistakes such as empty fields, missing codes, and identifying duplicate data points. Data cleaning is considered a foundational element of the [data science basics](https://www.sisense.com/glossary/data-science-basics/), as it plays an important role in the analytical process and uncovering reliable answers. Most importantly, the goal of data cleaning is to create data sets that are standardized and uniform to allow business intelligence and [data analytics tools](https://www.sisense.com/product/) to easily access and find the right data for each query.

There are many techniques for data cleaning. In our project for the simplicity of our program we drop few columns which we deem not useful for our analysis and prediction. Further we remove the null values, for doing so we can either eliminate null rows or fill the null values with median of the column values, removing null values is very important step in data cleaning. Next, we can explore our columns, here we take size column and see that there is discontinuity in the representation of the column values. We can resolve this by creating a new column by applying some function on the previous replacing column. It is a common problem of having un-structured data in the database, as we look in the total\_sqft column where some values are a range some in metre, so we apply functions to solve such inconstancies as well.

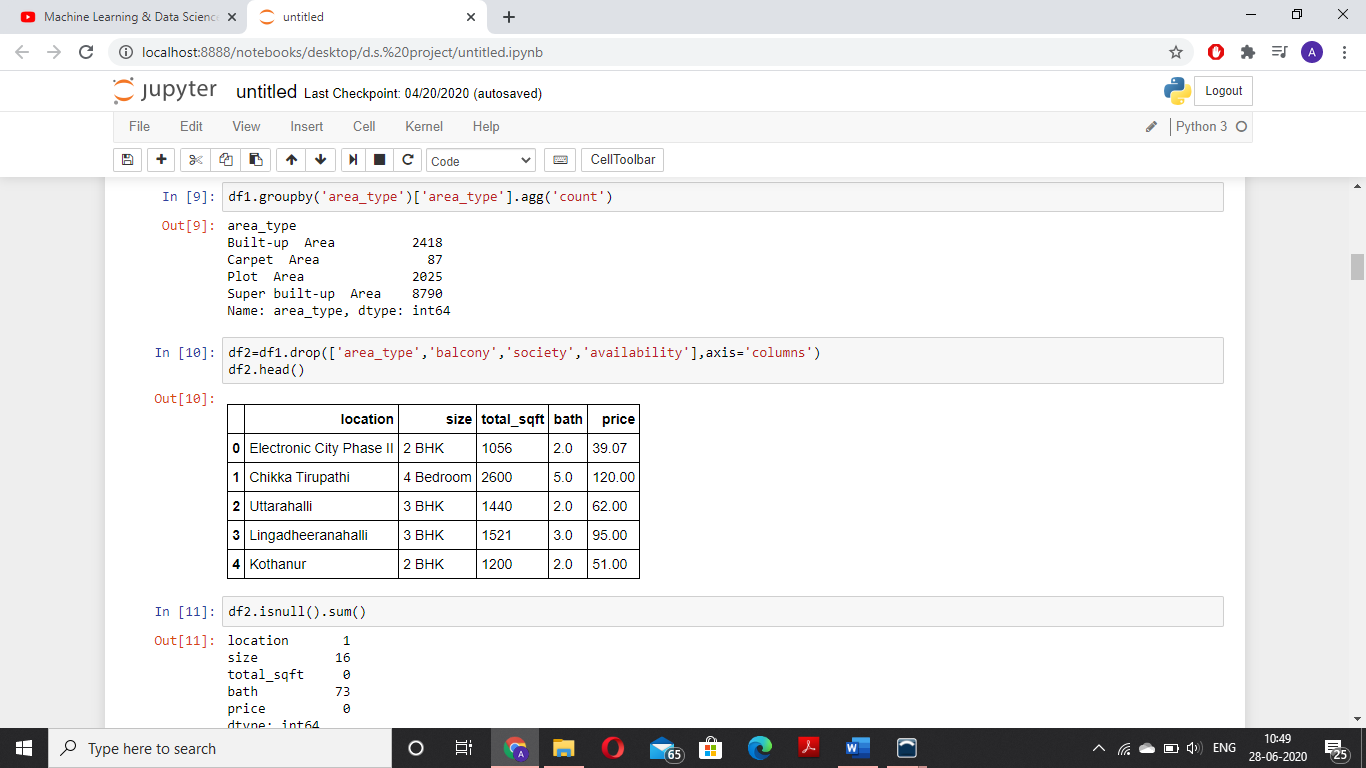


Fig 2.1 (Removing non useful columns)

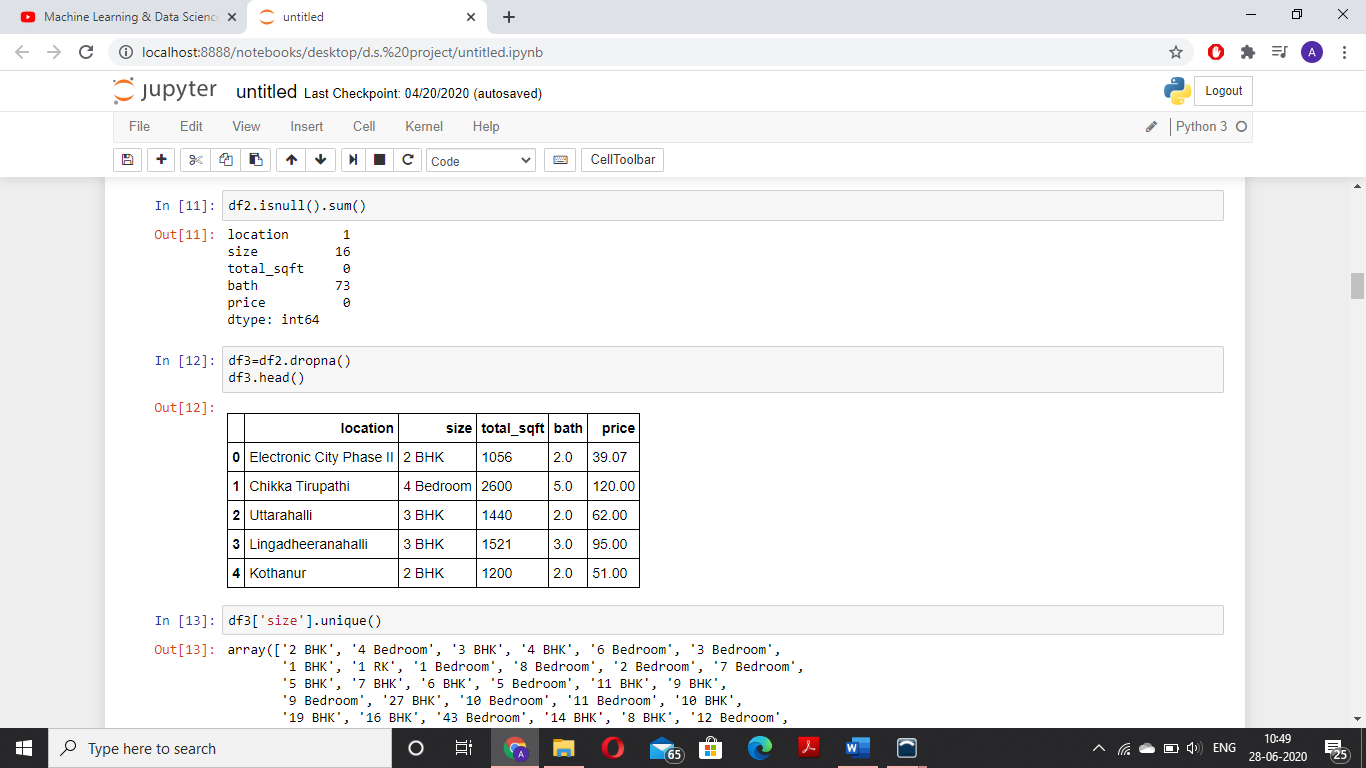


Fig 2.2 (Removing null rows)

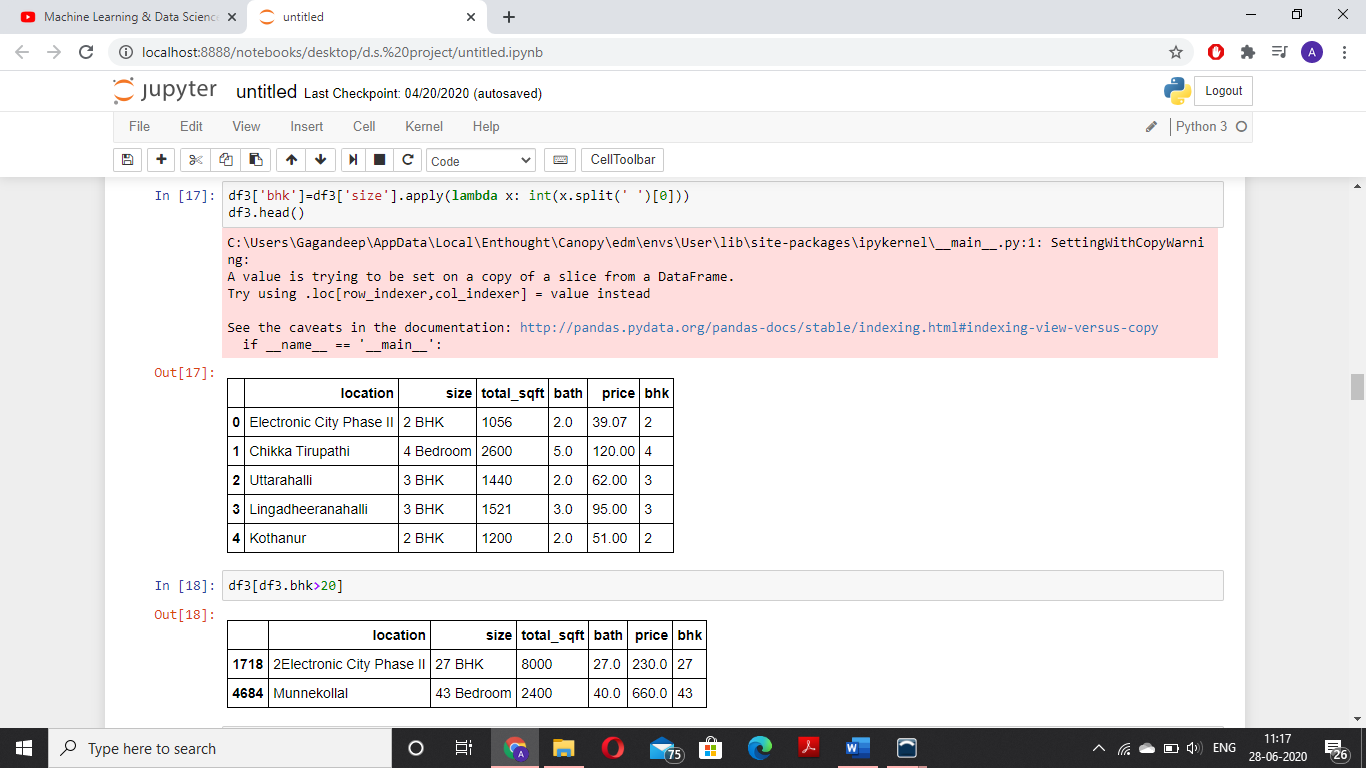


Fig 2.3 (Applying function on size column to remove inconsistency in its value)

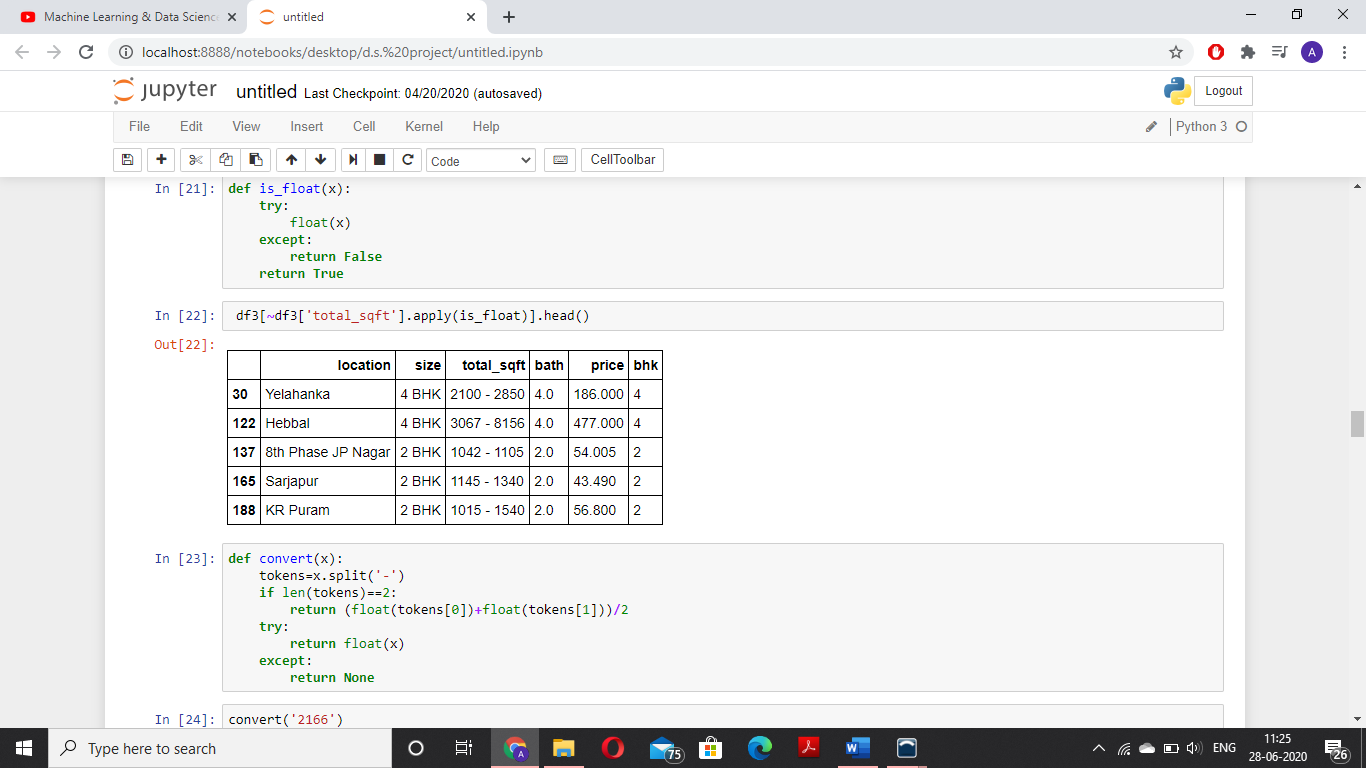


Fig 2.4 (Inconsistencies in total\_sqft column)

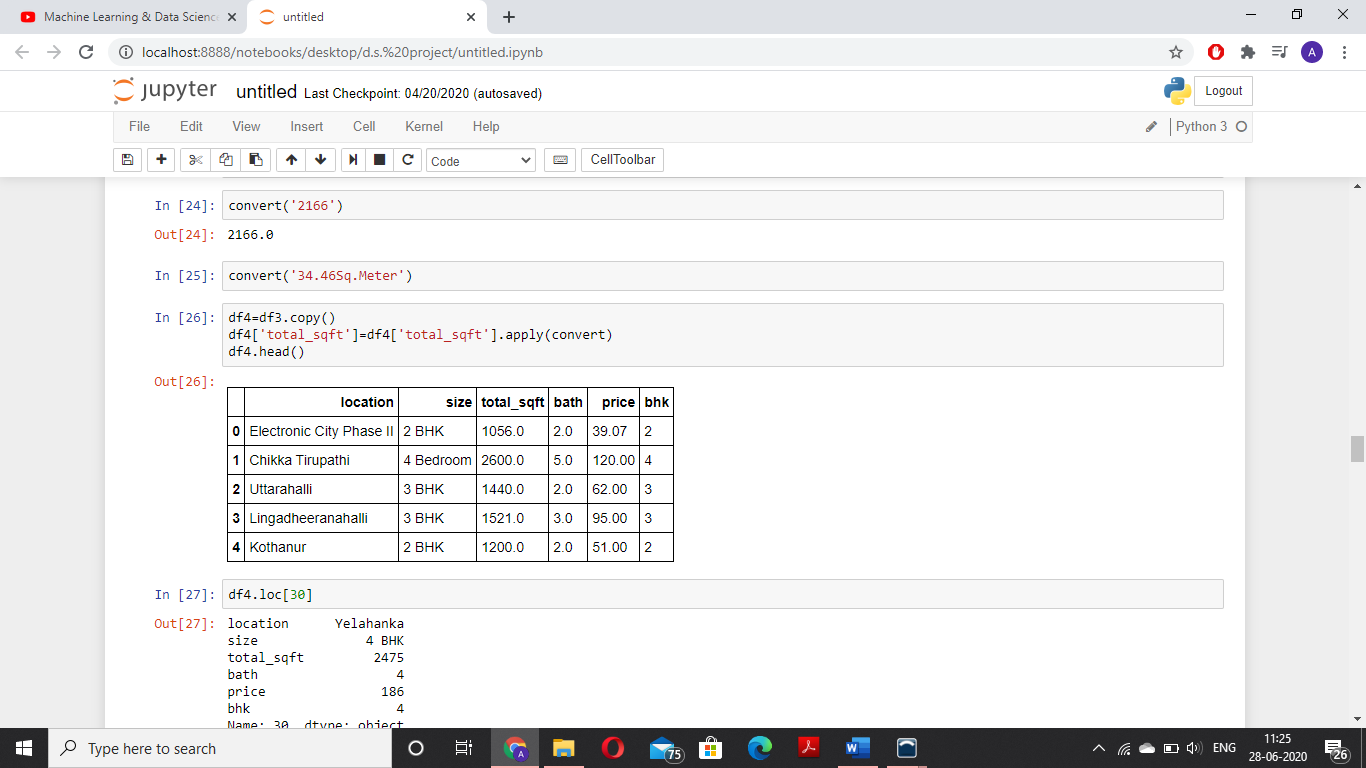


Fig 2.5 (After removing inconsistencies in total\_sqft column)

STEP 3 - FEATURE ENGINEERING

Feature engineering is the process of using [domain knowledge](https://en.wikipedia.org/wiki/Domain_knowledge) to extract [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) from raw [data](https://en.wikipedia.org/wiki/Data) via [data mining](https://en.wikipedia.org/wiki/Data_mining) techniques. These features can be used to improve the performance of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms. Feature engineering can be considered as applied machine learning itself. A [feature](https://en.wikipedia.org/wiki/Feature_(machine_learning)) is an attribute or property shared by all of the independent units on which analysis or prediction is to be done. Any attribute could be a feature, as long as it is useful to the model.

The purpose of a feature, other than being an attribute, would be much easier to understand in the context of a problem. A feature is a characteristic that might help when solving the problem main uses of feature engineering are preparing the proper input dataset, compatible with the machine learning algorithm requirements and improving the performance of machine learning models.

In our project we make a new column price per square feet, which is a very important feature in real estate market. Also, we can check our location column and further reduce our rows by applying function on it.

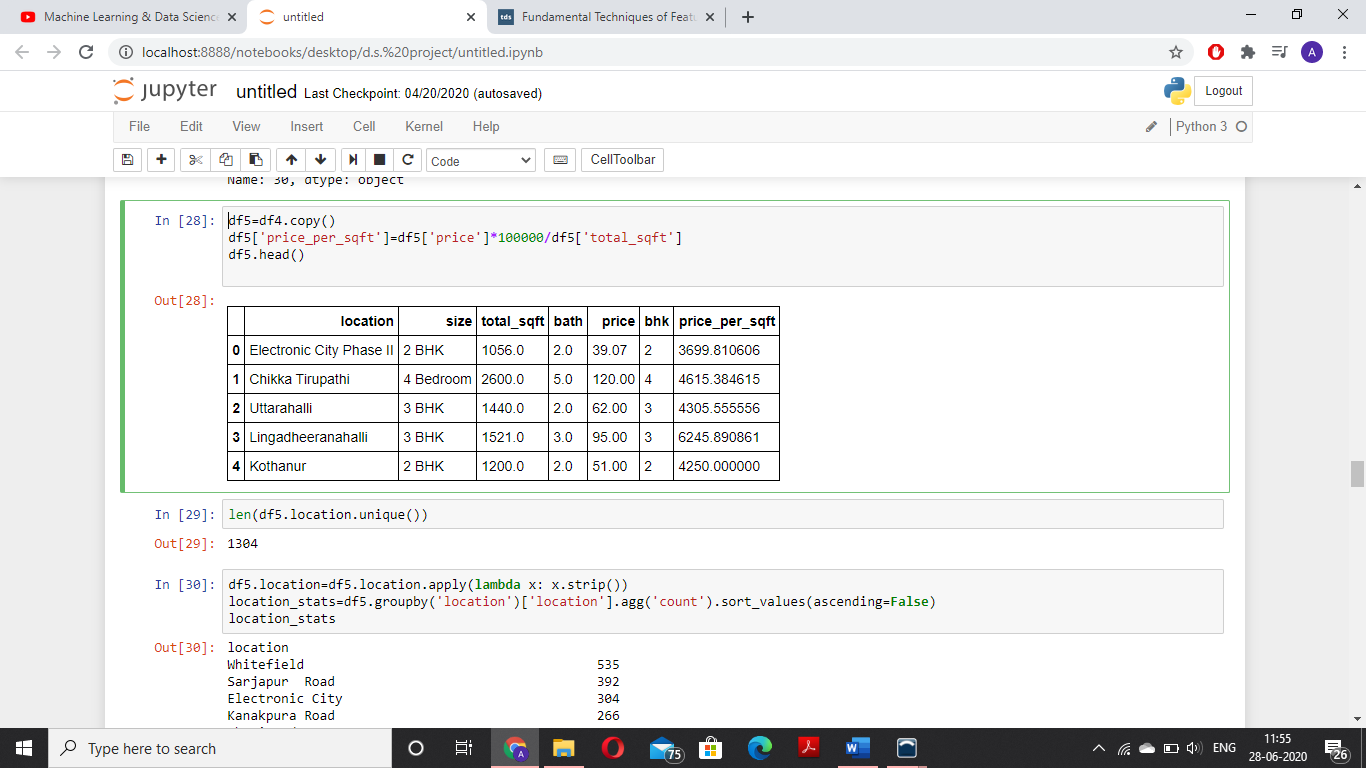


Fig 3.1 (New feature price\_per\_sqft)

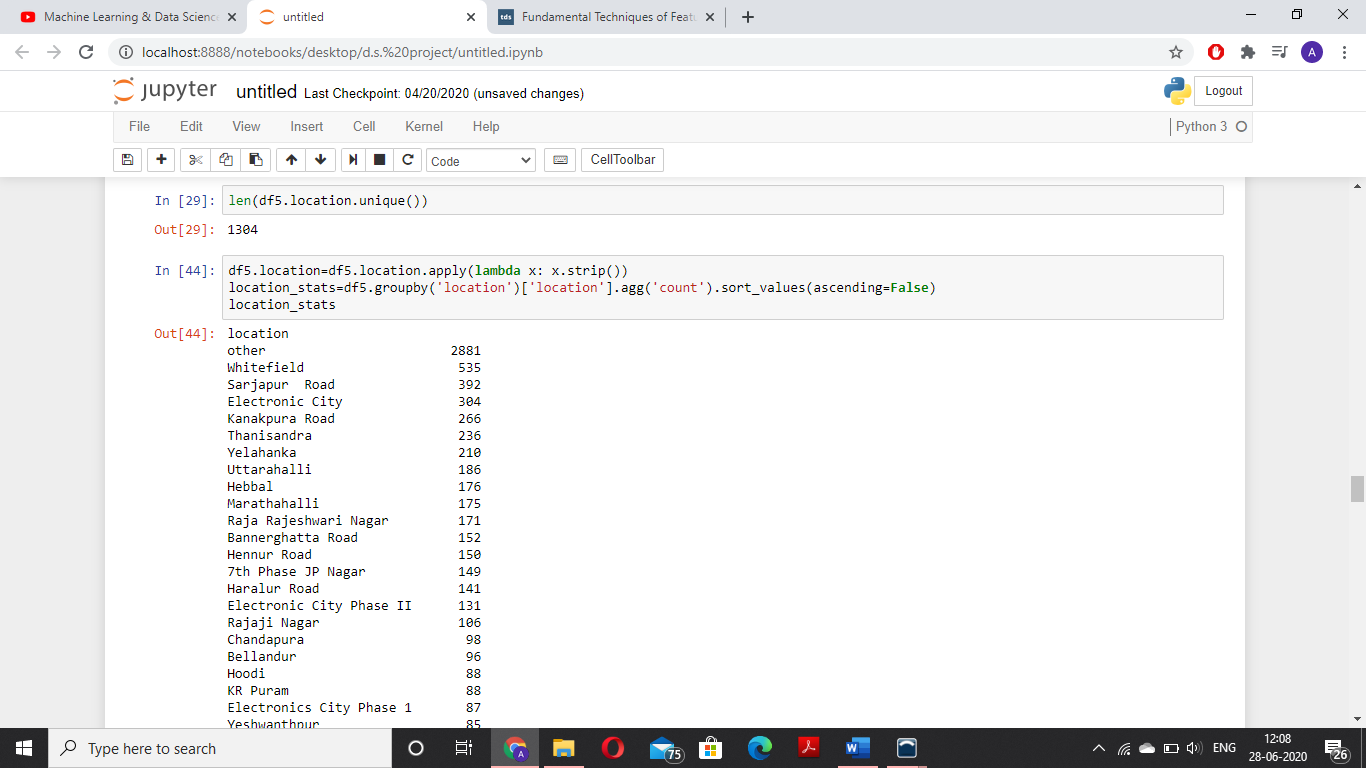


Fig 3.2 (Examining location column)

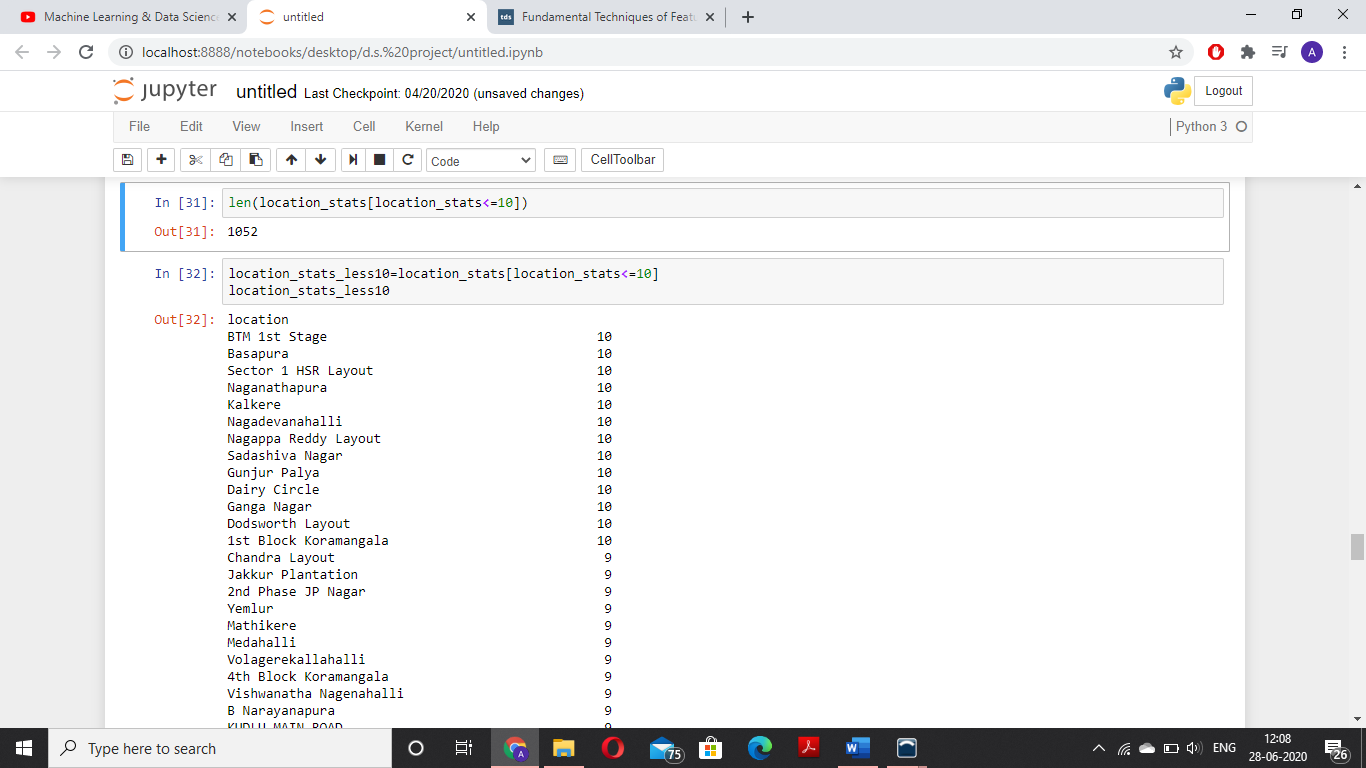


Fig 3.3 (Considering all locations with less than one count as other)

STEP 4 – OUTLIER REMOVAL

When modelling, it is important to clean the data sample to ensure that the observations best represent the problem. Sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data. These are called outliers and often machine learning modelling and model skill in general can be improved by understanding and even removing these outlier values.

Now even in our data many extremities lie and it is important to get rid of such data as it affects our prediction poorly. One outlier we can fine is in the square feet column, we find that below or above a certain threshold the area is not possible so we remove such inconsistent data. Also we can remove outliers by basing them on the threshold for price.

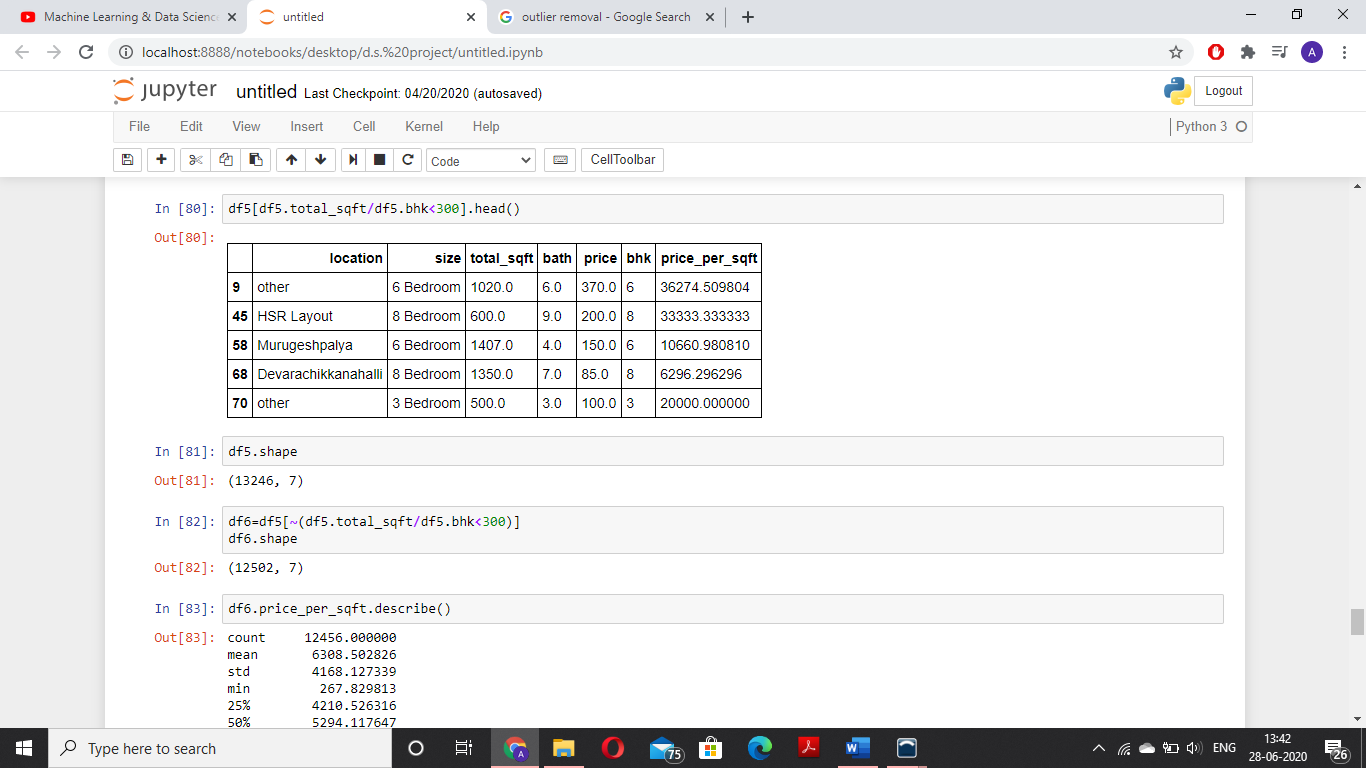


Fig 4.1 (Outliers)

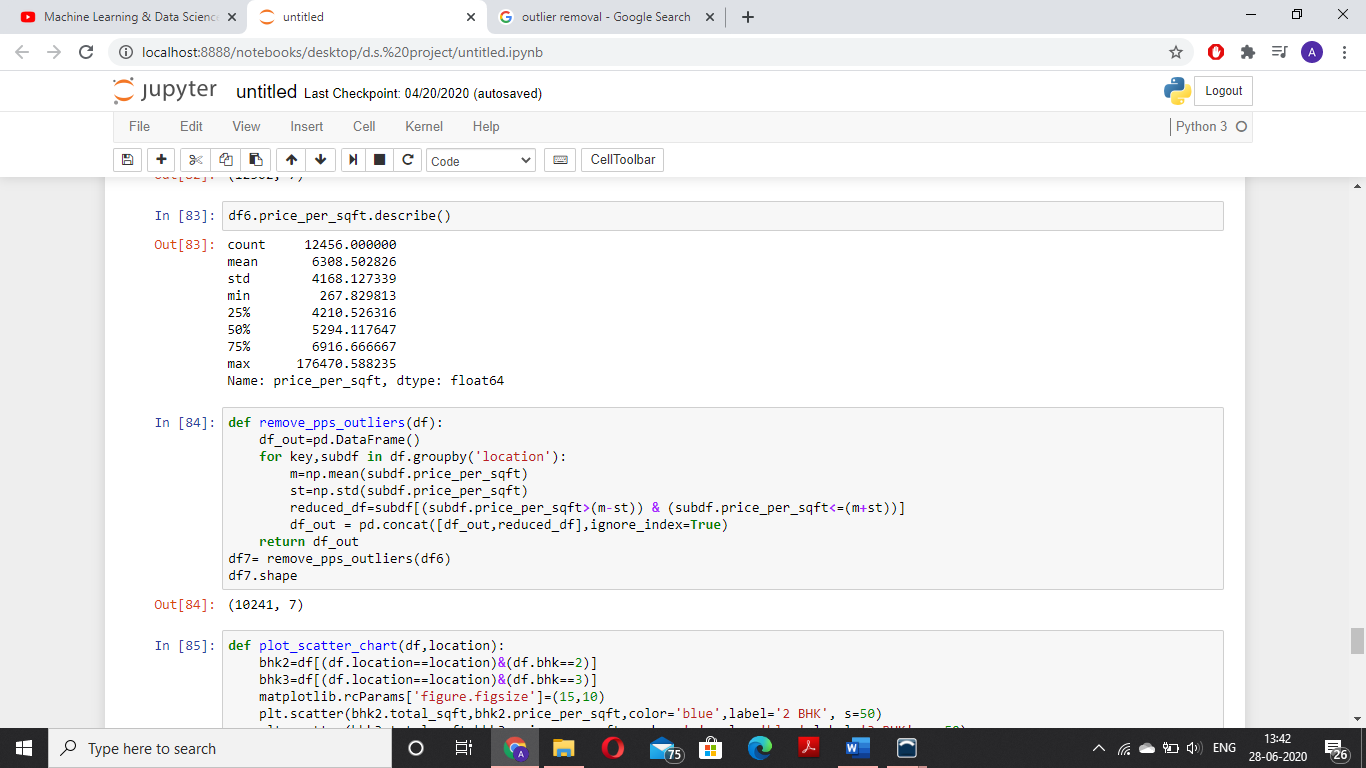


Fig 4.2 (Outlier removal)

STEP 5 - MODEL BUILDING

Our final step is model building, now that we have cleaned our database and got rid of all inconsistencies, we can apply algorithm to train our model and then expect predictions. For that we have to do few things such as change text to numeric format then apply a suitable algorithm, in this project we apply linear regression. **Linear Regression** is a machine learning algorithm based on**supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Now we have tried two models here test and cross validation and gridcv is used to decide the better amongst the two models

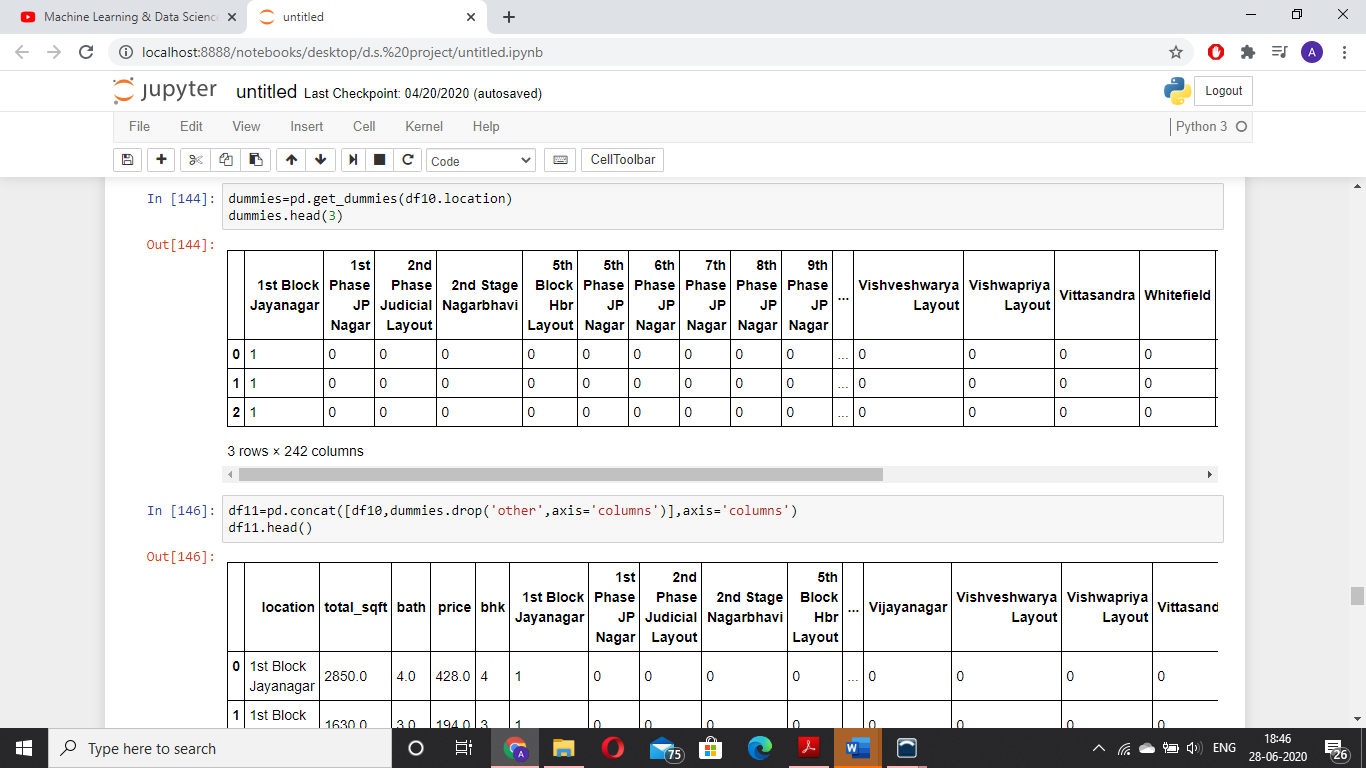


Fig 5.1 (Using hot encoding to convert text to numeric format)

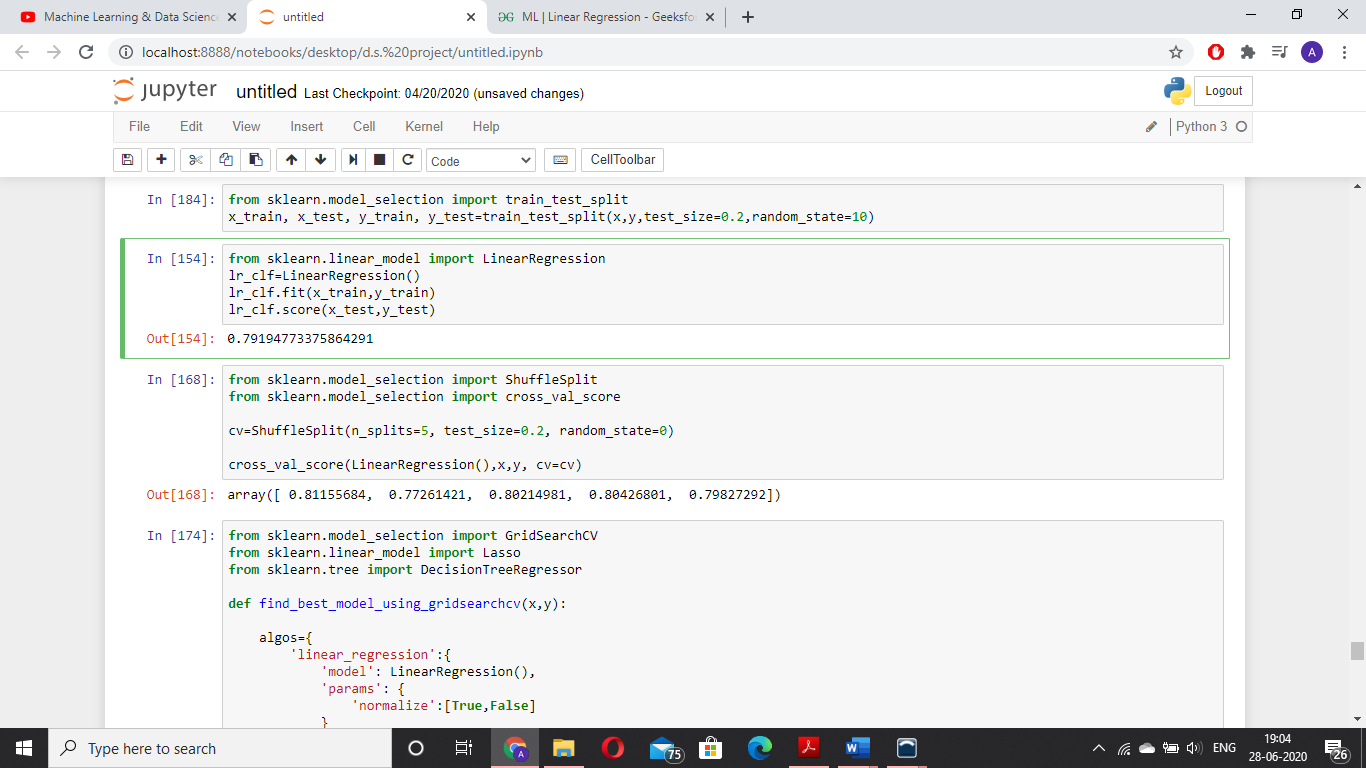


Fig 5.2(Importing sklearn, creation of train and test sets and calculating score of models)

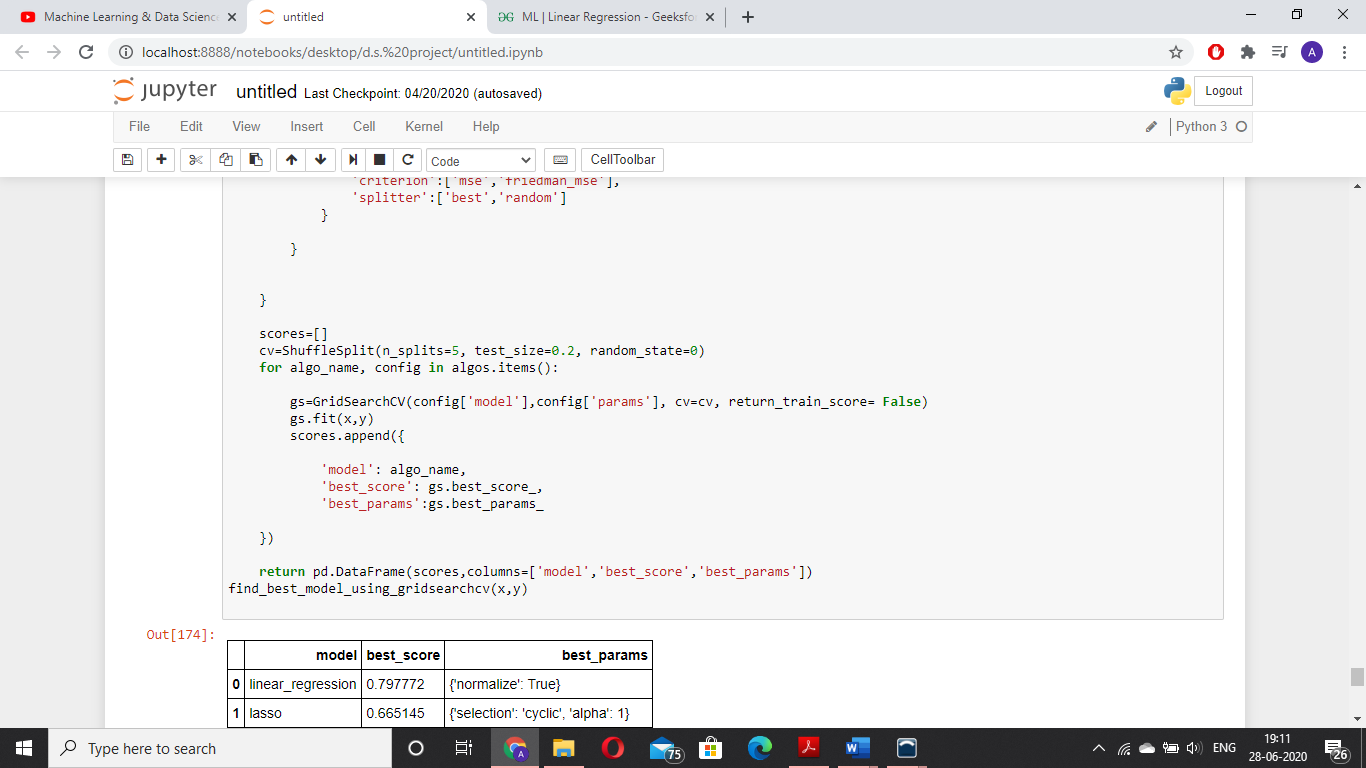


Fig 5.3(Using gridcv to decide the better model)

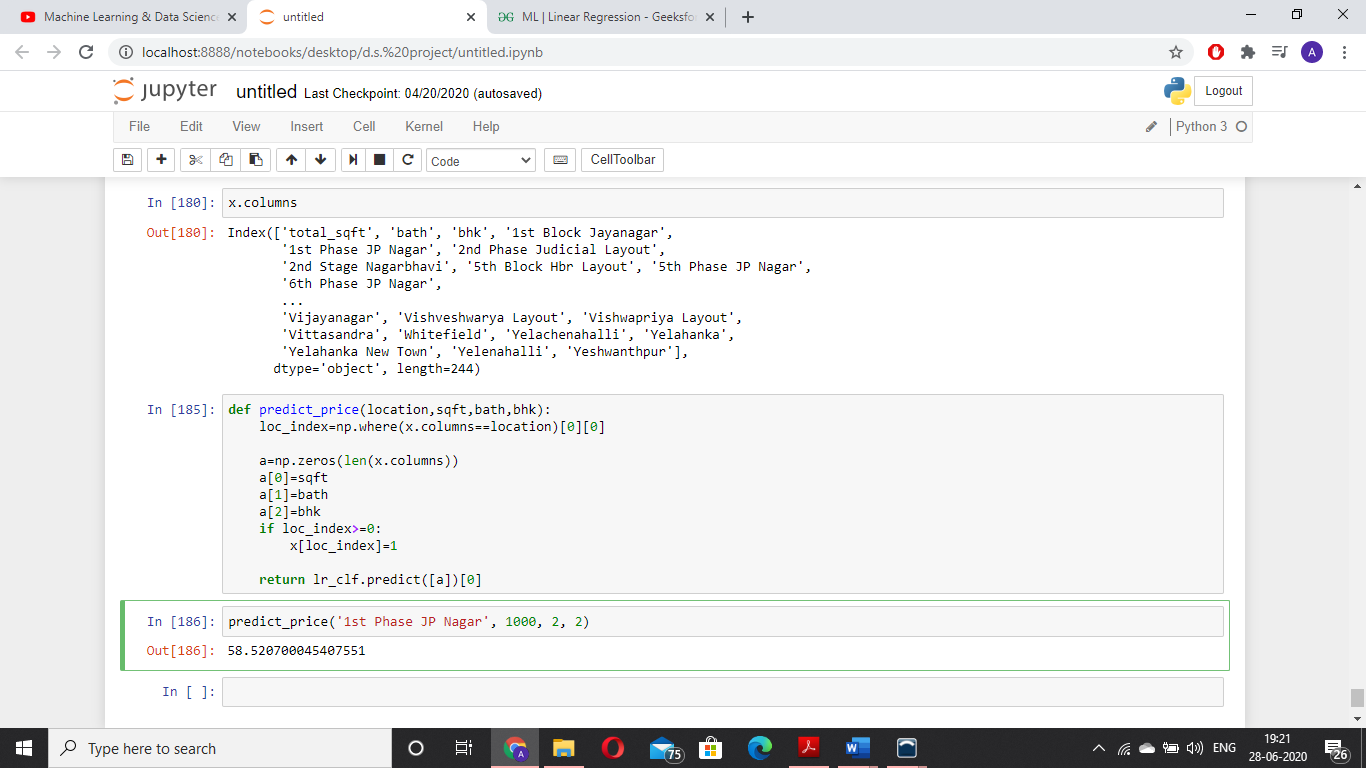


Fig 5.4 (Price prediction)

**CONCLUSION**

Usually, here are the way of addressing overfitting/underfitting problem:

● Overfitting

○ Reduce number of features.

○ Increase λ.

○ Add training examples.

● Underfitting ○ Increase number of features.

○ Polynomial feature.

○ Decrease λ.