|  |
| --- |
| University of Arkansas at Little Rock – Information Science |
| House Prices – Prediction using Advanced Regression Techniques |
| Data Science - Technologies Project |

|  |
| --- |
| Robinson Tamilselvan (T00693942)  5-3-2022 |

Table of Contents

[Introduction 2](#_Toc103027712)

[Data 2](#_Toc103027713)

[Technologies and Tools Used 2](#_Toc103027714)

[Analysis 3](#_Toc103027715)

[Data Cleansing 3](#_Toc103027716)

[Filtering and Dimensionality Reduction 3](#_Toc103027717)

[Normalization of Numerical variables 4](#_Toc103027718)

[Normalization of Categorical variables 5](#_Toc103027719)

[Feature selection 5](#_Toc103027720)

[Comparing various Models and Choosing the best 6](#_Toc103027721)

[Results 7](#_Toc103027722)

[Packages used 8](#_Toc103027723)

[Analysis in AWS Cloud 8](#_Toc103027724)

[Creating a MYSQL Instance in Cloud using RDS service 8](#_Toc103027725)

[Loading Data into the Cloud 10](#_Toc103027726)

[Modification of python code 10](#_Toc103027727)

[Using Shell script to check the modified code 12](#_Toc103027728)

[Creating Lambda function and S3 bucket in AWS 13](#_Toc103027729)

[Creating EC2(AWS) Instance-Ubuntu server to compute the data 14](#_Toc103027730)

[Packages used 16](#_Toc103027731)

[Conclusion 16](#_Toc103027732)

# Introduction

Predictive data analytics is nothing but applying Machine Learning technique to the data that we already have in-order to predict the future like what kind of movies, products that people are going to be interested in. In this Project I am predicting the house prices for the features it possesses, the location it is present, the year it’s been built etc. I have a historical data which has the SalePrice and the features and, I am going to analyse the data and teach the model to learn from it. And then using that model to find or predict the SalesPrice of houses which I do not know. Also, as an extension to this I also performed some ETL operations in the Cloud using the free-tier services offered by AWS cloud like Lambda function, S3 bucket, EC2 - Ubuntu server, RDS - MYSQL database and compute the result in the cloud.

# Data

The Data that I used to perform this project was taken from Kaggle. The data set contains details about the house features. And also, sale price is given for the house.

My Data set can be classified into two Training and Testing. My Training dataset contains 81 columns or features and 1460 rows which I used to train the model. My Testing dataset contains 80 columns or features except for Sales Price and 1459 rows which I used to test. Both of my dataset is in structured format in a csv file and its source is Kaggle. Attaching the data description, train.csv, test.csv below for your reference.

**https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques**

# Technologies and Tools Used

* + Python
  + Jupyter Notebook
  + AWS Cloud (RDS, Lambda­, S3)
  + MYSQL
  + Shell script
  + AWS EC2, Ubuntu, Putty, Filezilla

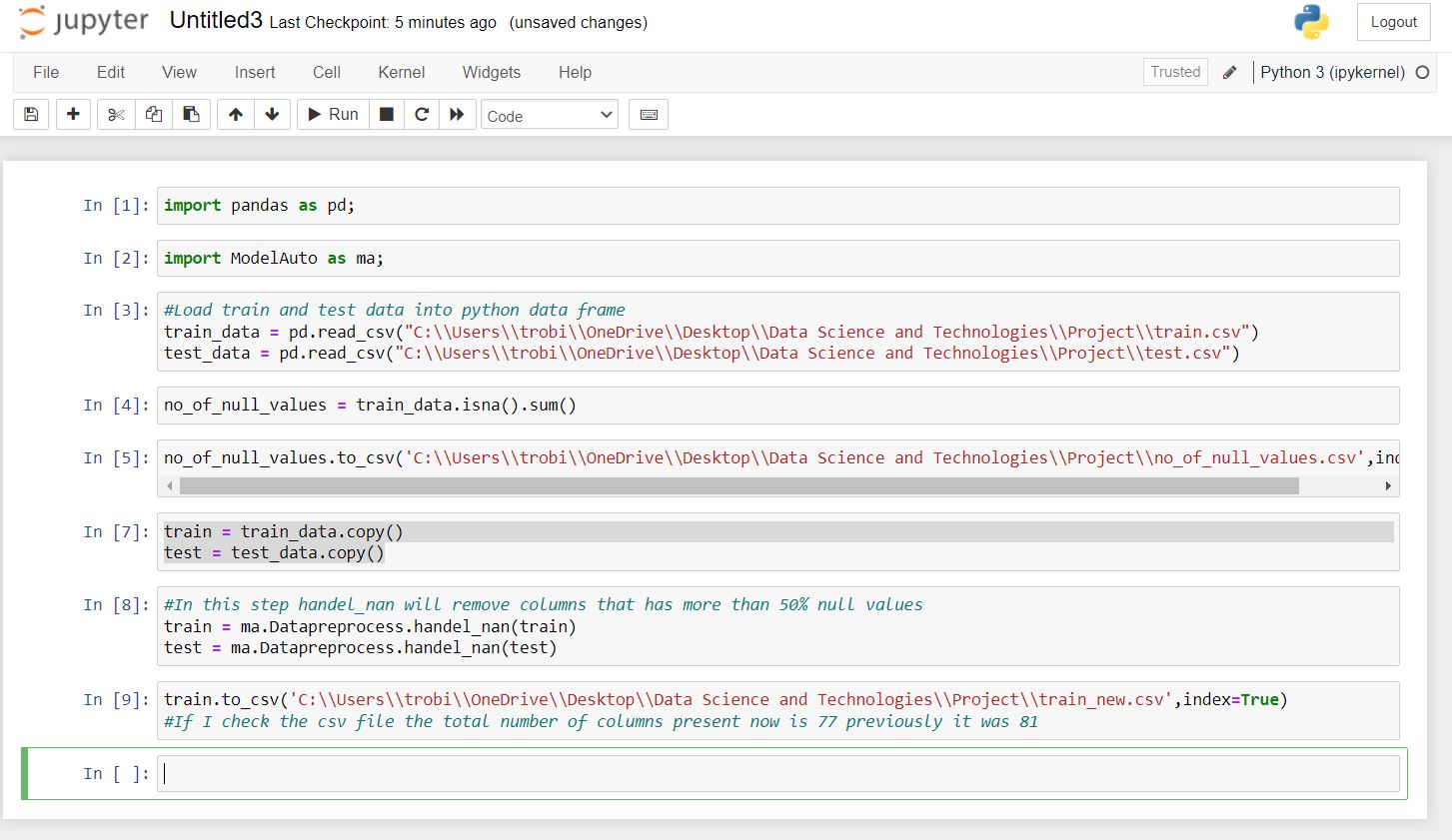
# Analysis

## Data Cleansing

Data Cleansing is the most important part in the prediction analytics. Even if the machine algorithms are well written if the data is not cleaned properly, it will not produce the desired result. An article about how amazon’s machine learning recruiting algorithm went horribly wrong because of faulty data is a good example for why Data cleansing is vital. In case of amazon since the data set used to train the model has not much women resume so the algorithm didn’t know how to rate a female applicant and hence showed bias against women.  
<https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e>

For my project I researched on various data cleansing techniques and methods and implemented those to clean my data properly. Below are the techniques and code which I used to clean my data.

### **Filtering and Dimensionality Reduction**

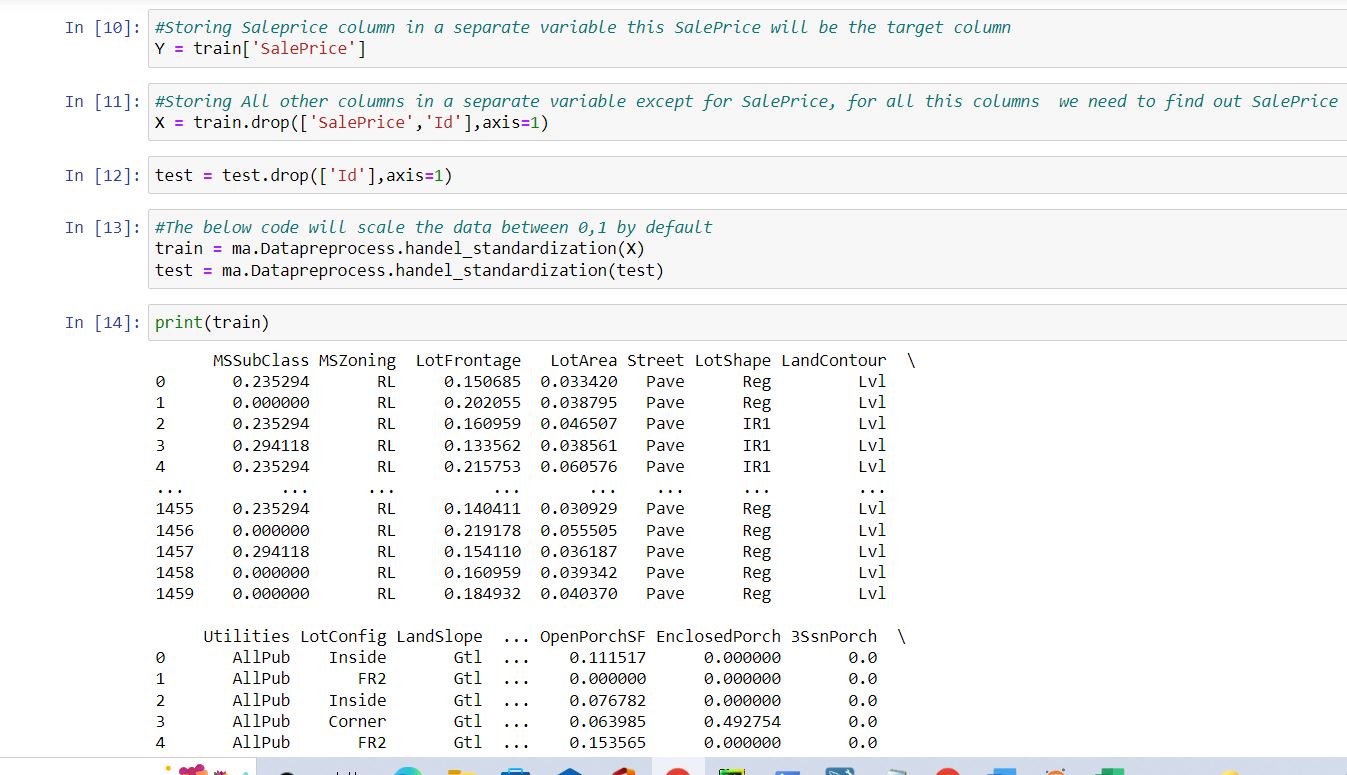


Filtering and Dimensionality Reduction are applied in the above code to filter the records to remove columns that has more than 50% null values.

*train\_data.isna().sum()* is used to check the number of null values in each columns and then it is exported into a csv file. Attaching the csv file below.

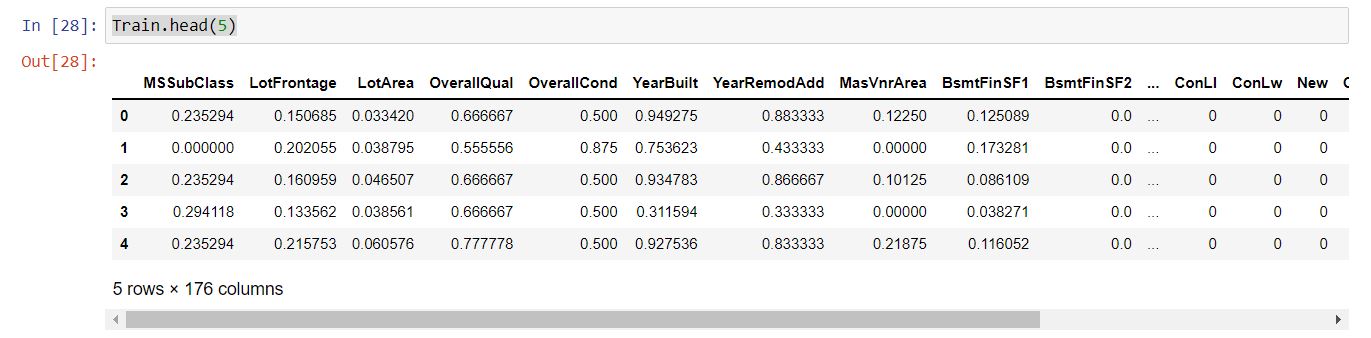
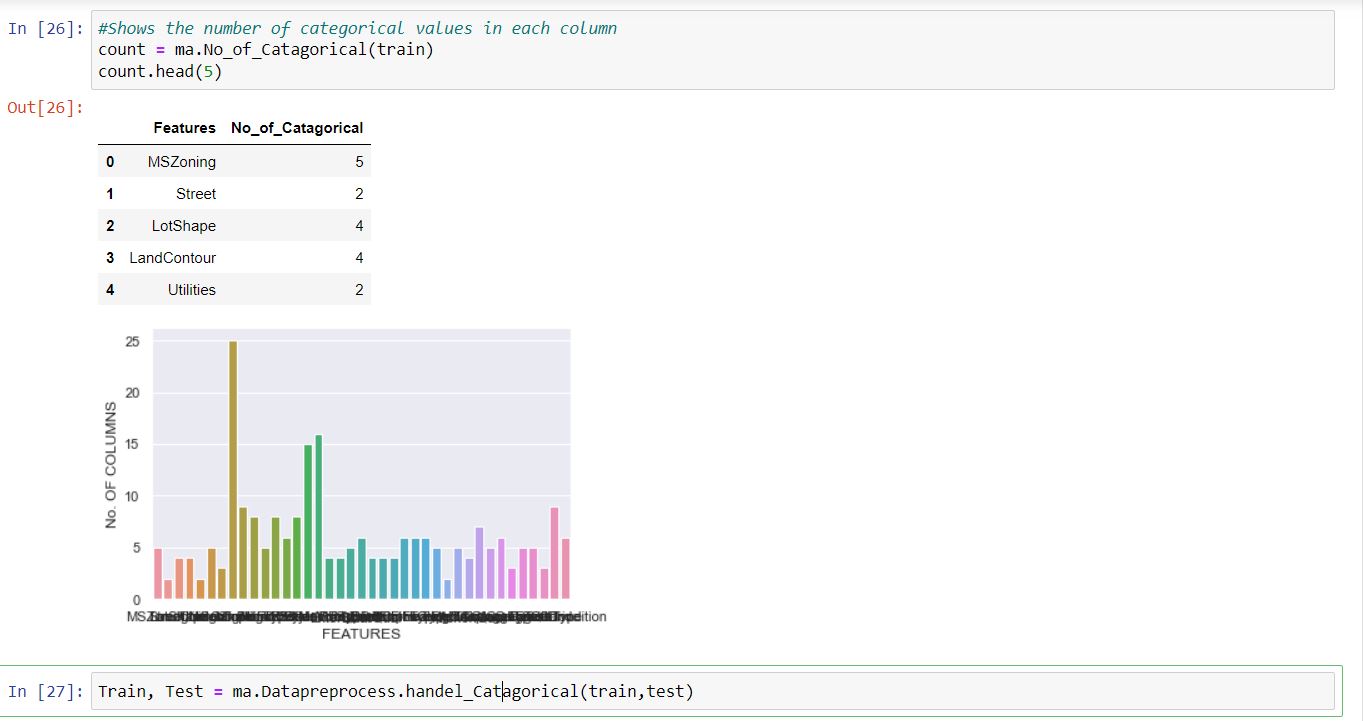
*train = ma.Datapreprocess.handel\_nan(train)* is used to remove columns that has more than 50% null values

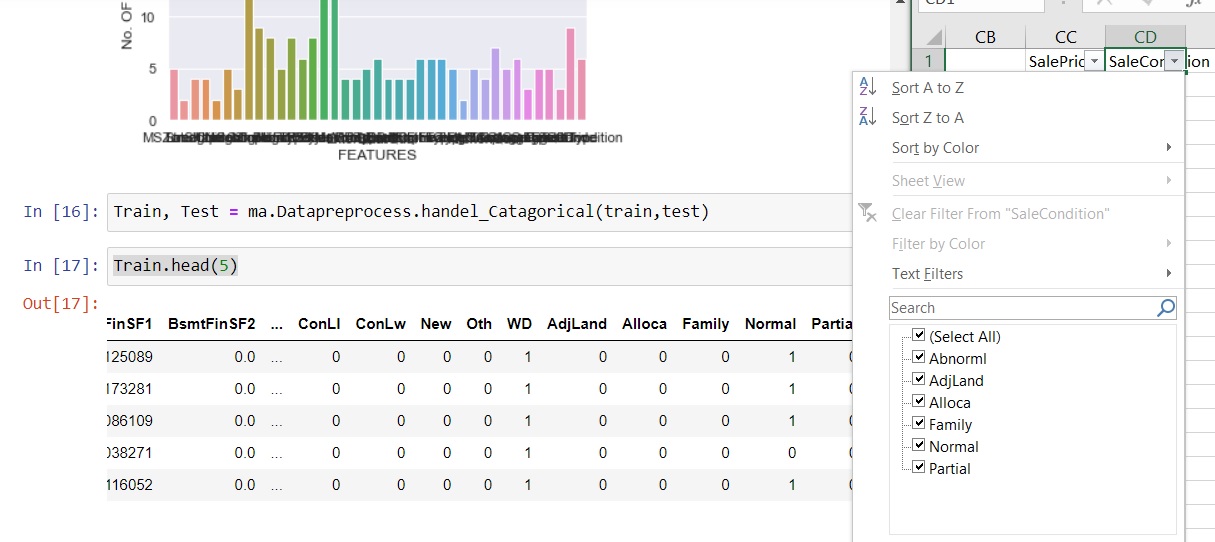
### Normalization of Numerical variables



In the above code [13] *train = ma.Datapreprocess.handel\_standardization(X)* Normalisation of data has been done which is to scale the numerical values between 0 and 1 after the columns are separated in a suitable way according to the task.

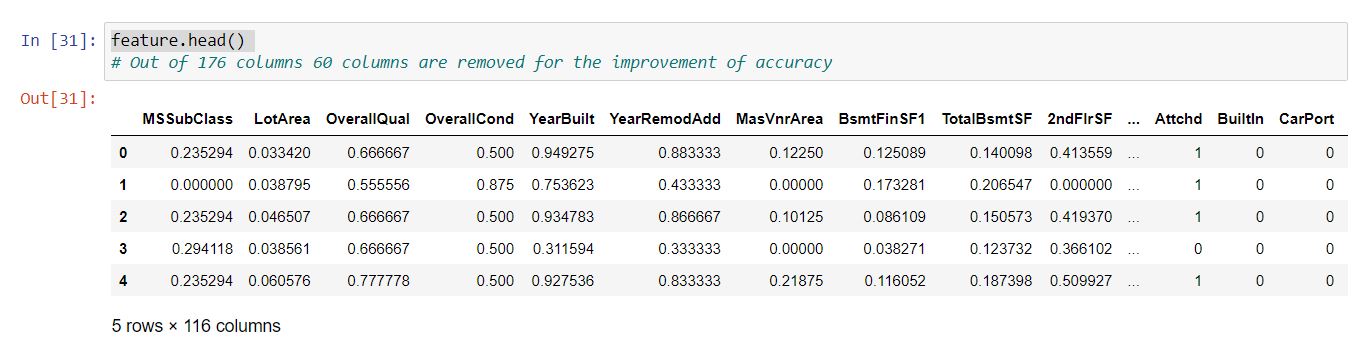
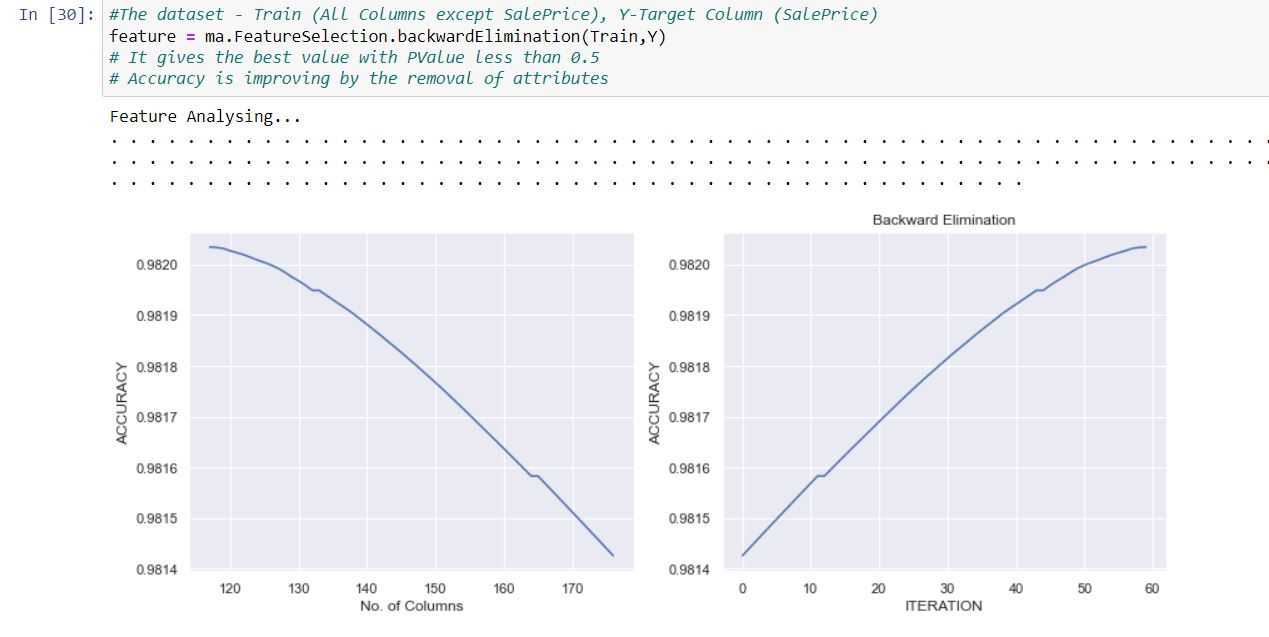
### Normalization of Categorical variables



*count = ma.No\_of\_Catagorical(train)* is used to count the number of categories present in the categorical variables. *Train, Test = ma.Datapreprocess.handel\_Catagorical(train,test)* assigns the value between 0 and 1 for the categorical variables. The reason behind the increase in the number of columns is that all the categorical values are converted as columns and value between 0 and 1 is assigned for them. Kindly refer the below screenshot.  


### Feature selection

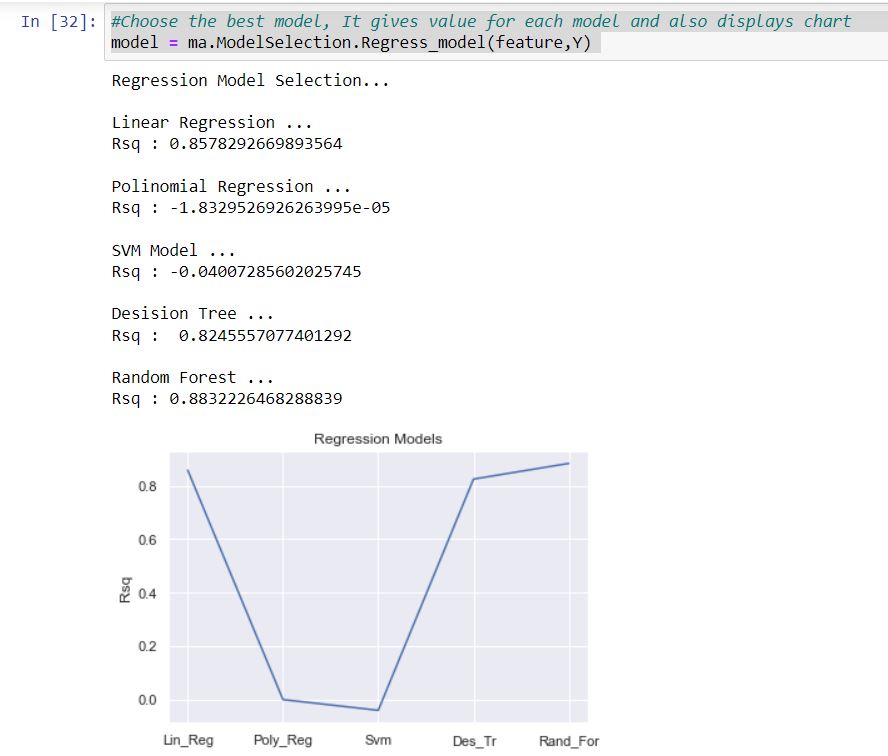
Feature selection is the process of reducing the number of input variables when developing a predictive model.



The total number of columns present now is 116 whereas previously the number of columns was 176. And, 60 columns have been removed in the process of Feature selection to improve accuracy.

## Comparing various Models and Choosing the best

For my Model selection I tried five models and chose the best one which Random Forest based on the results.



Random forest is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. The random forest model is made up of multiple decision trees.

## Results

Using the Random Forest model to produce results which is to calculate the SalePrice of the house for the Testdata and exporting these Results to a csv file.



## Packages used

pandas, ModelAuto  
However in ModelAuto below packages were used

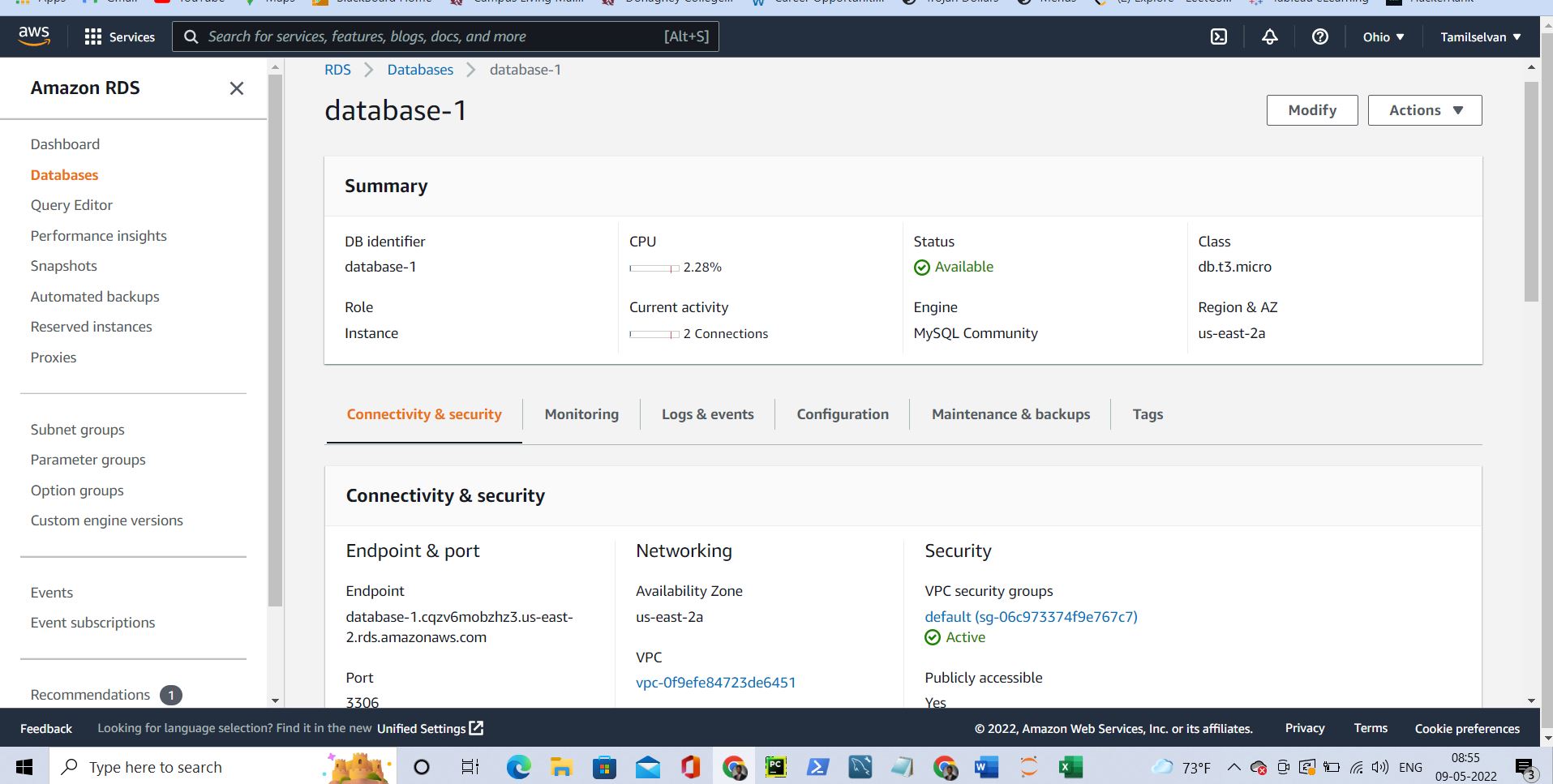
pytz, threadpoolctl, six, pyparsing, pillow, numpy, kiwisolver, joblib, fonttools, cycler, scipy, python-dateutil, patsy, packaging, scikit-learn, pandas, matplotlib, statsmodels, sklearn, seaborn,

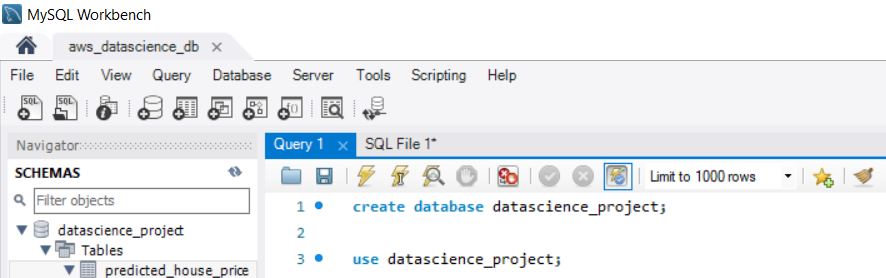
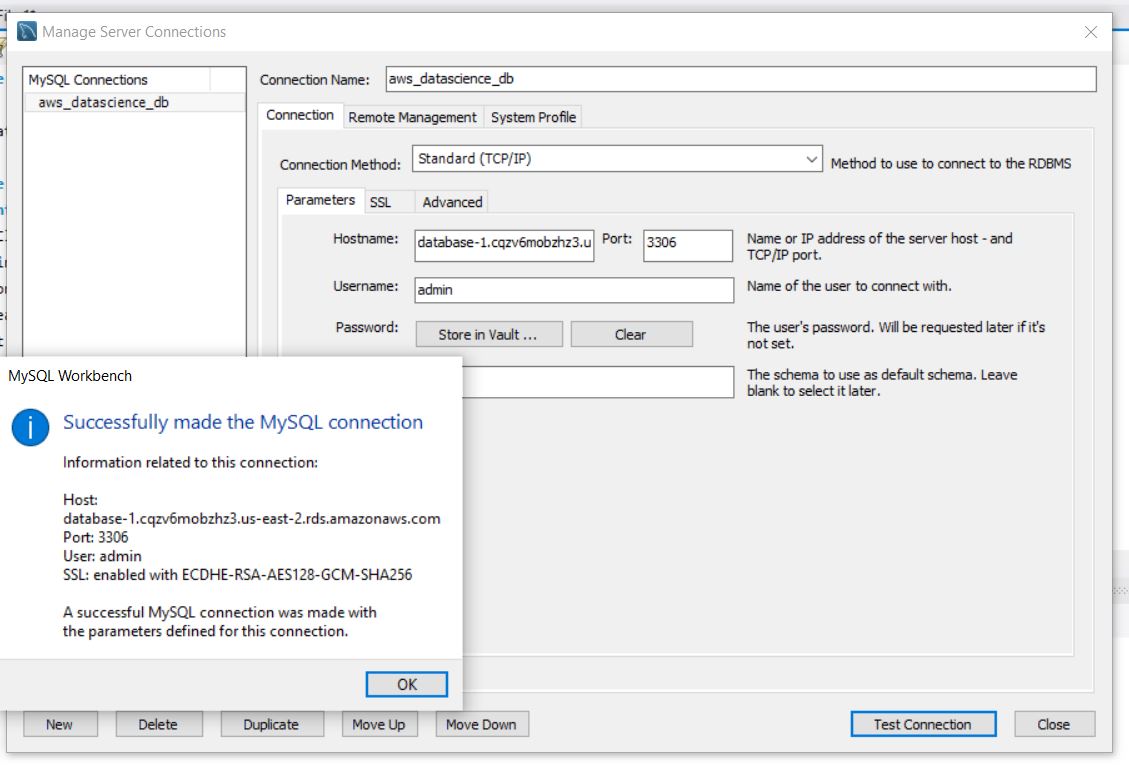
# Analysis in AWS Cloud

I am performing the same operation in the AWS Cloud in order to understand how to work in the cloud environment and also how to Extract data from a csv file into the cloud and Load it into a cloud Database and perform some Data Cleansing and Analysis in the cloud and Storing the results (Transformed data) into the cloud.  
For the scale of the data which I used for this project cloud in not required However I have done this in order to understand how should work on the cloud technologies and what are the difficulties and difference between cloud and the normal environment.

## Creating a MYSQL Instance in Cloud using RDS service

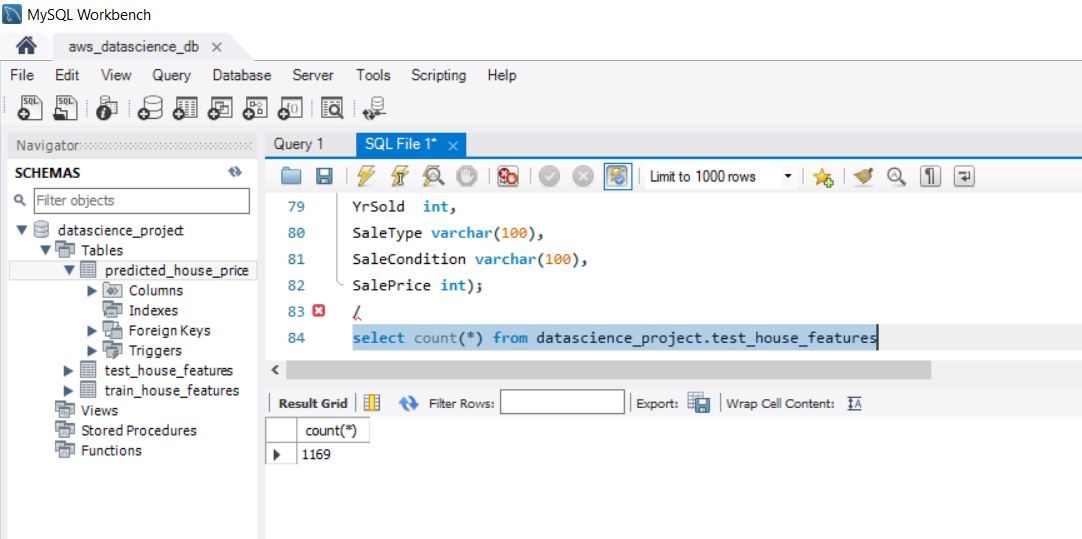
I have created a MYSQL Instance in the AWS Cloud using the RDS Service.



Connected the AWS Cloud instance and created a database “datascience\_project” in the cloud using MYSQL workbench. 

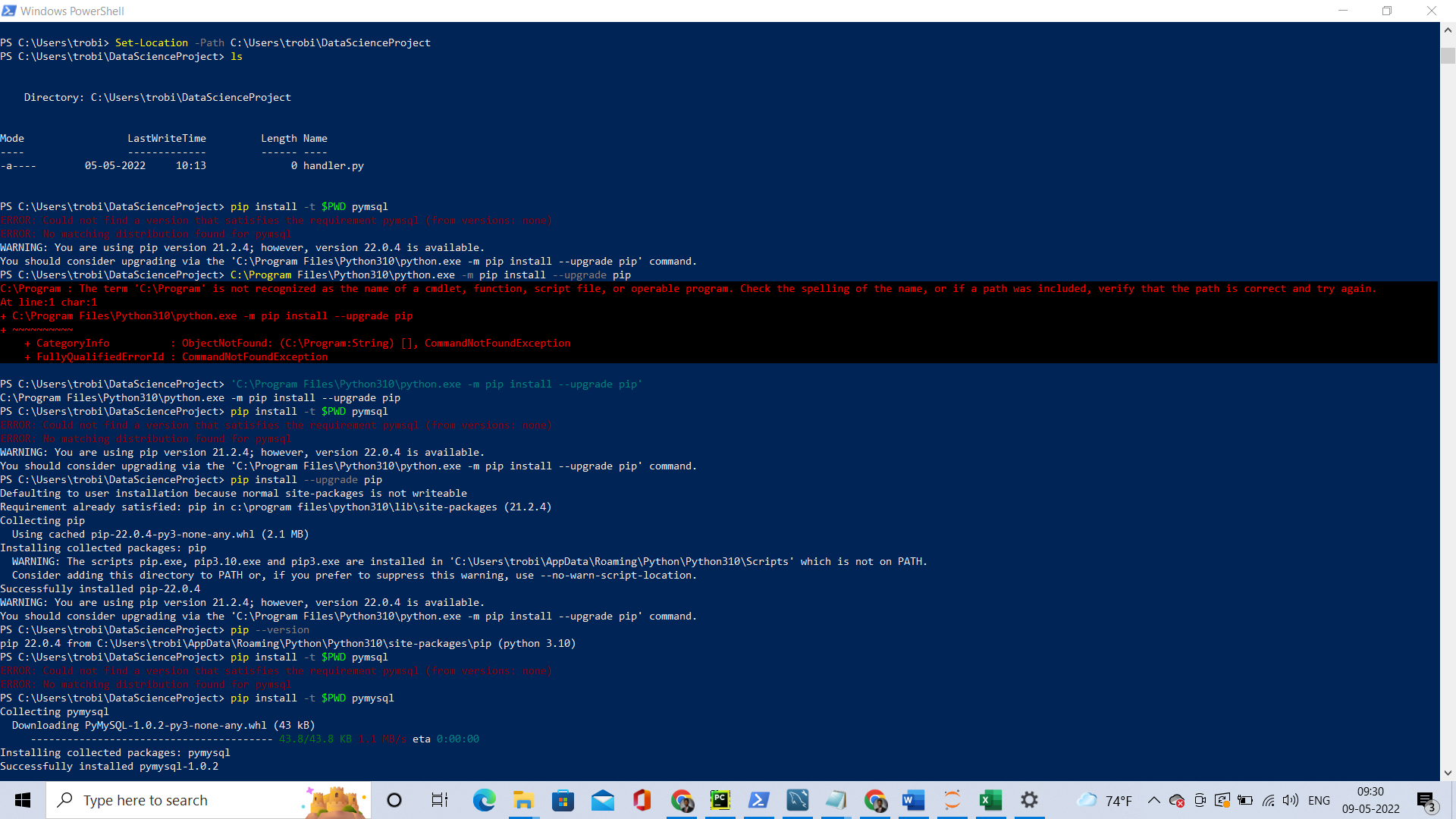
## Loading Data into the Cloud

Created tables for both the train and test data and imported all the records into the cloud.

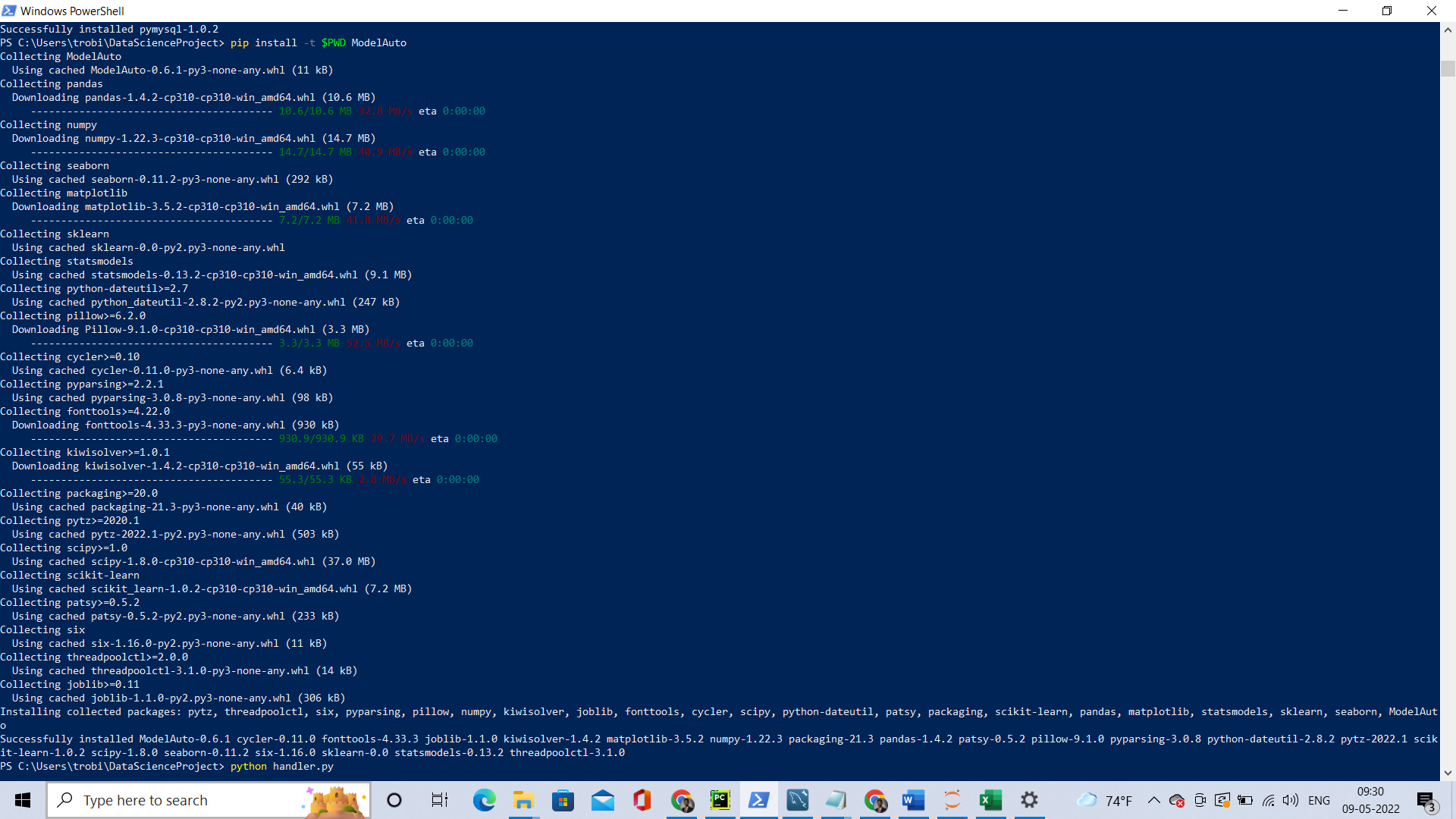


## Modification of python code

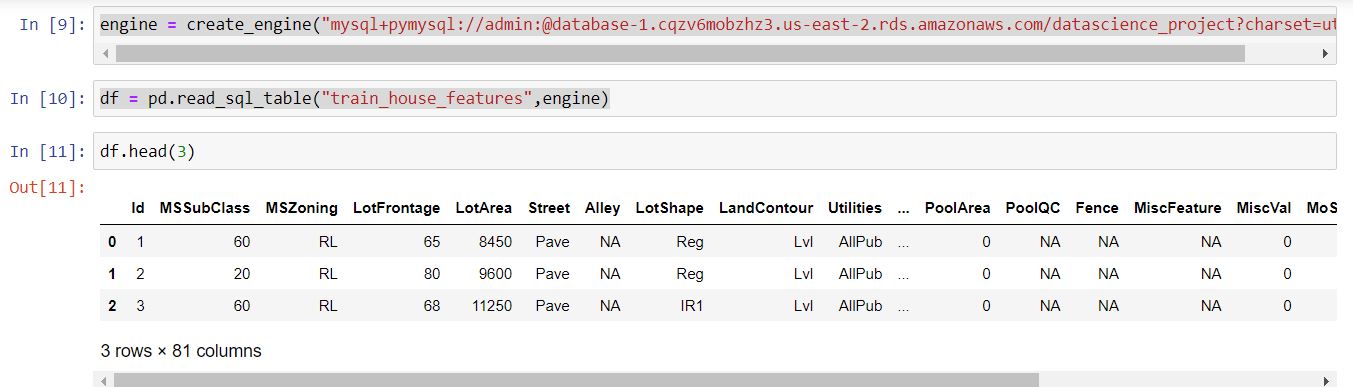
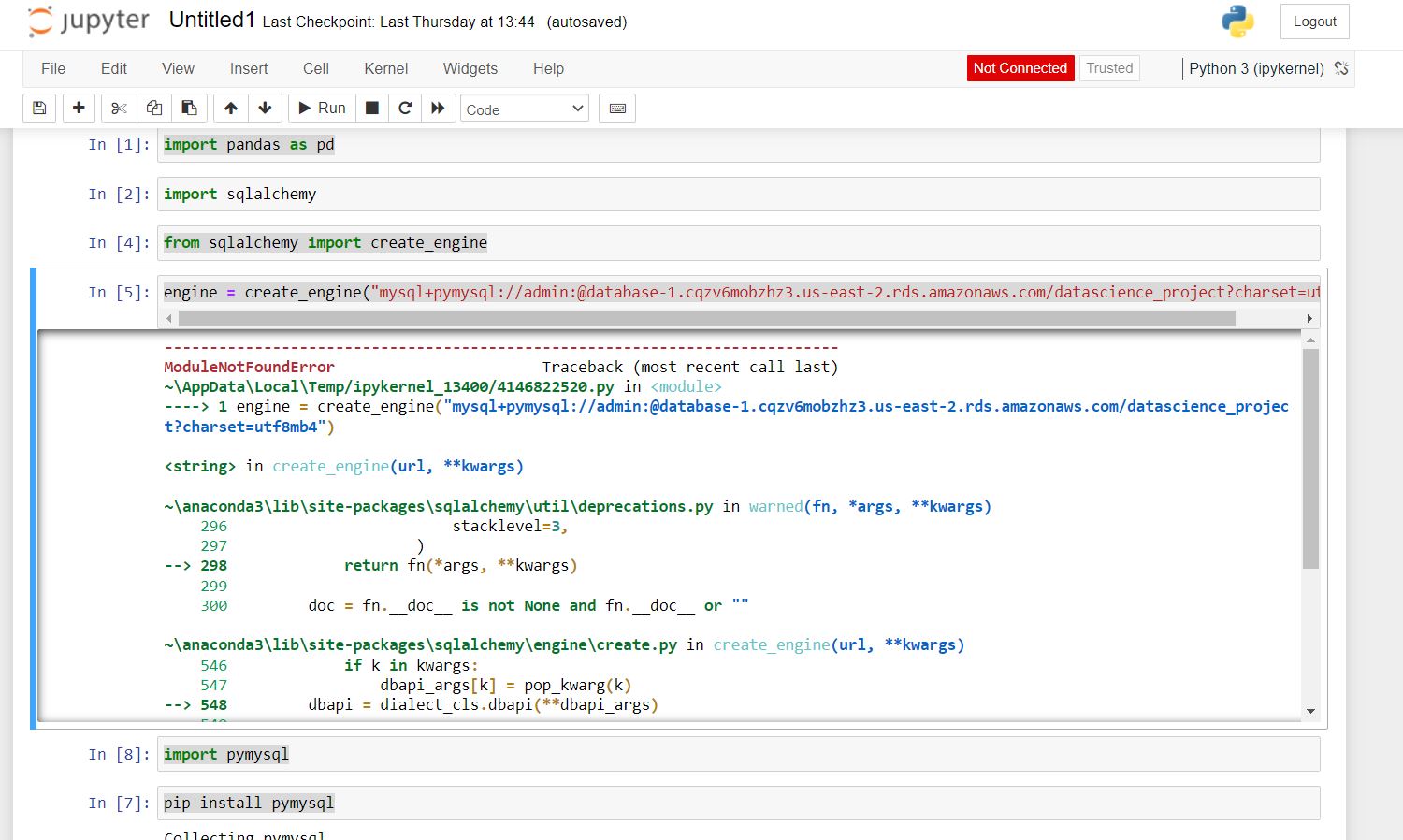
In order to analysis on the cloud

Create a directory on my local machine

Installed the packages in the directory created using Shell Command line

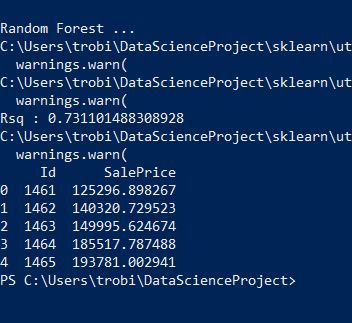
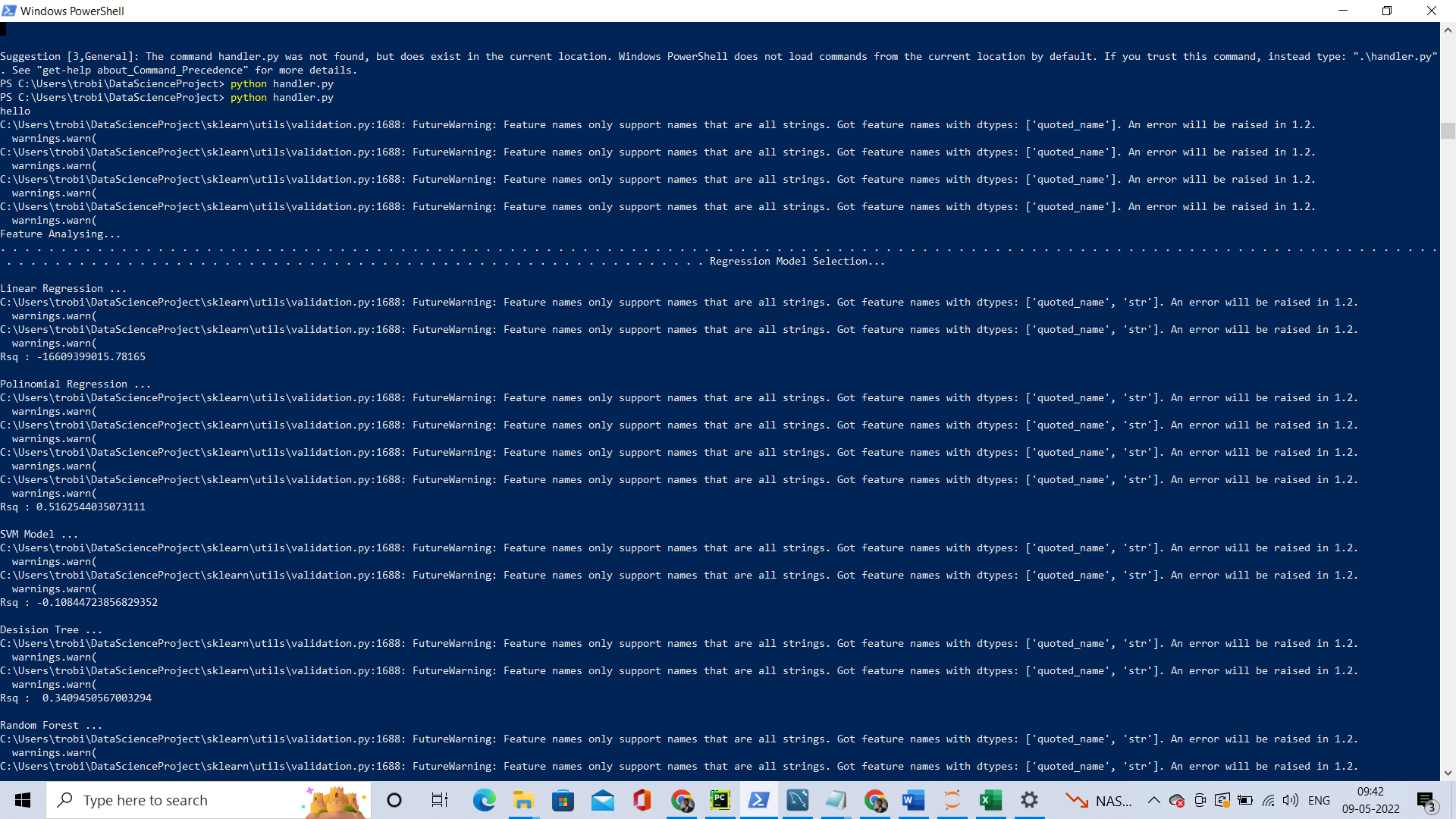


Modified the python code according to the cloud to connect to MYSQL instance in cloud and store it in a data frame and then do the process.

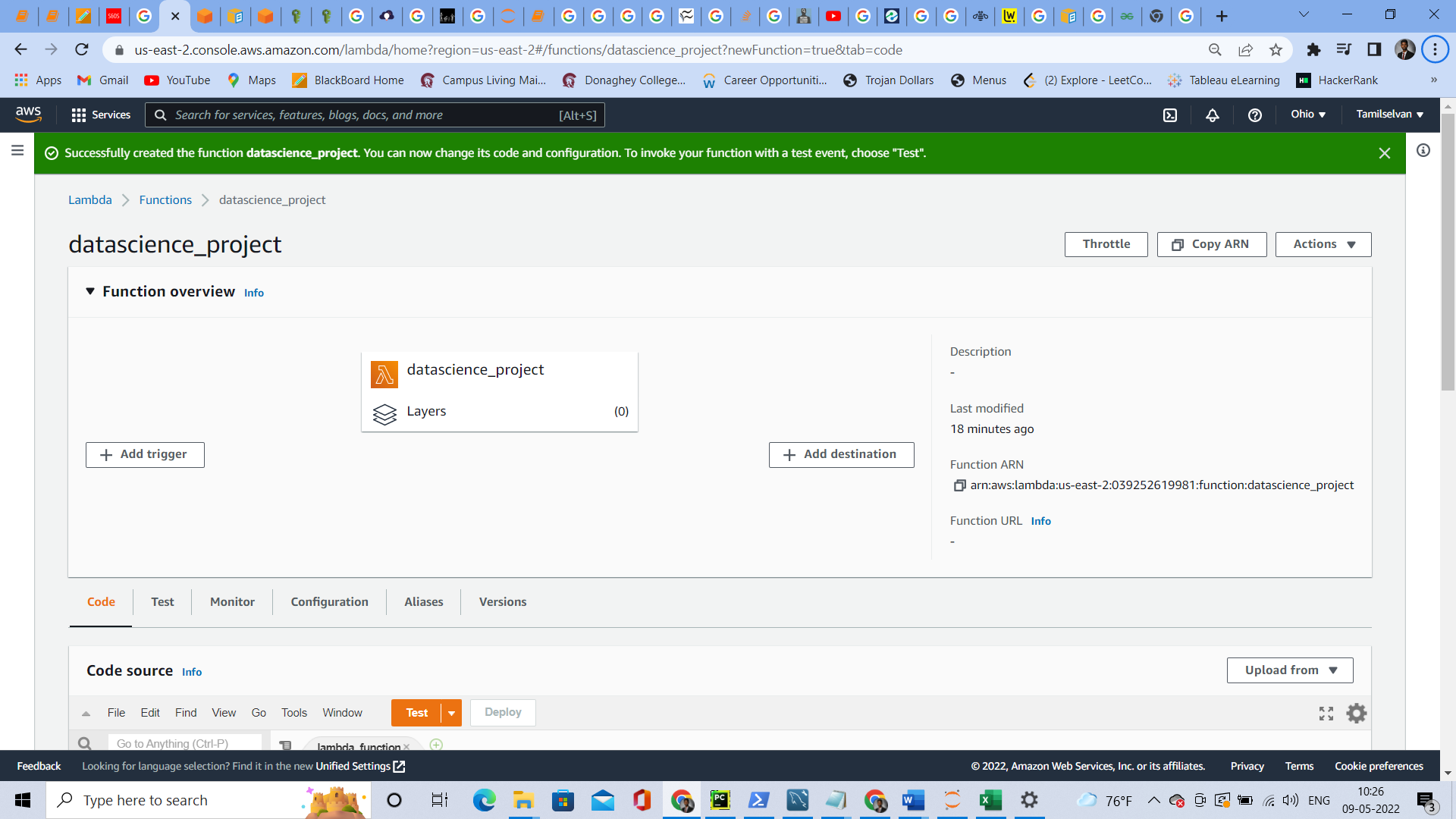


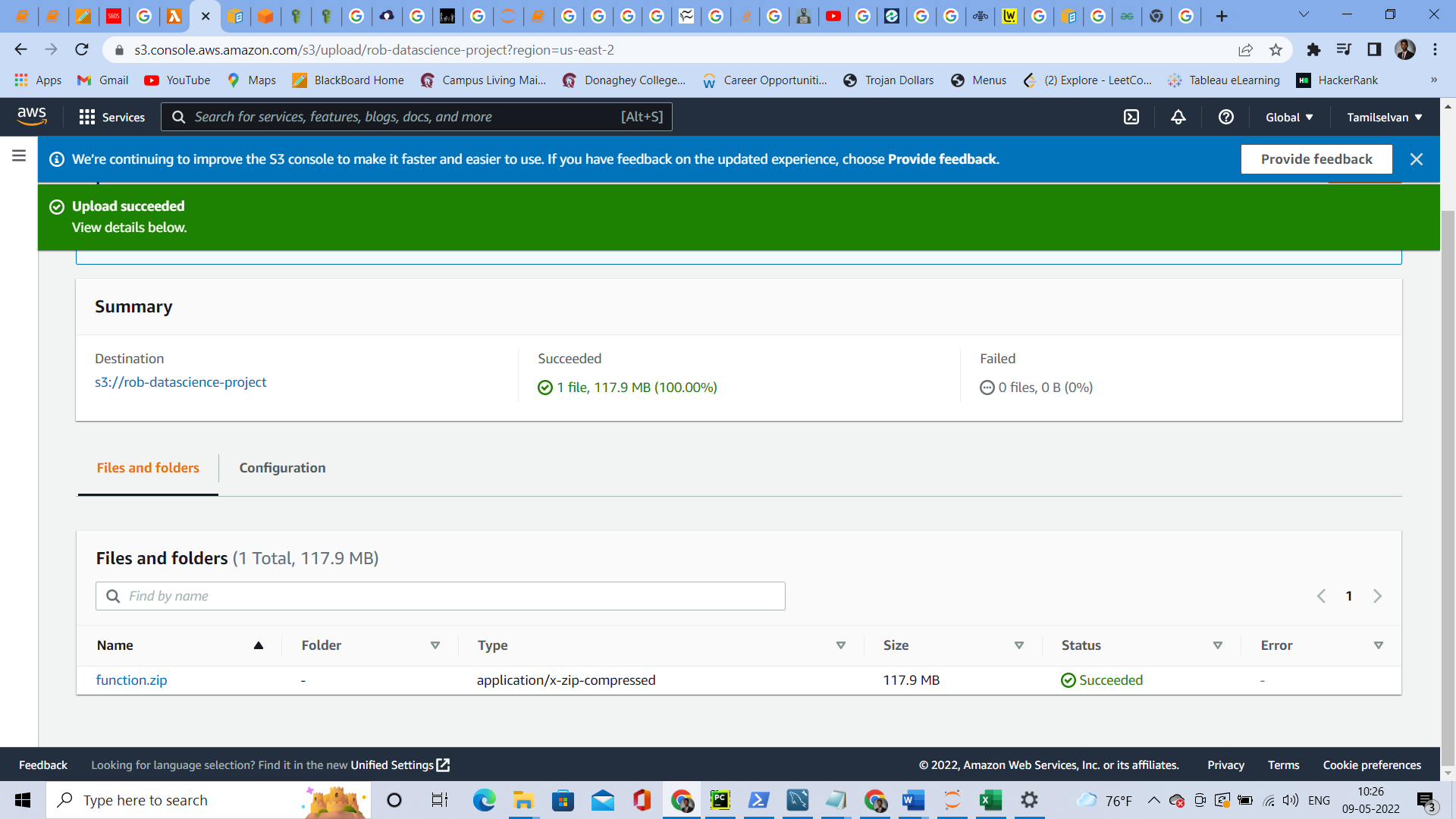
## Using Shell script to check the modified code

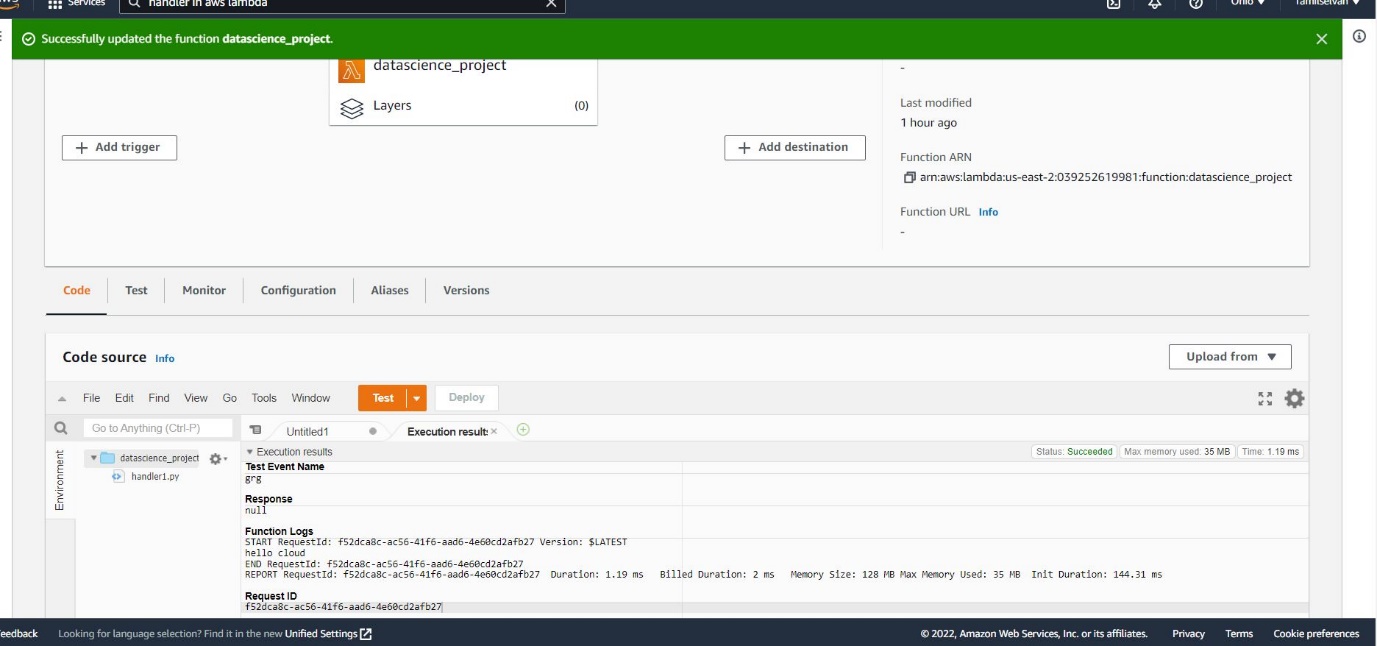
Store all these code in the .py file in the directory and executing the file using the Shell command line.

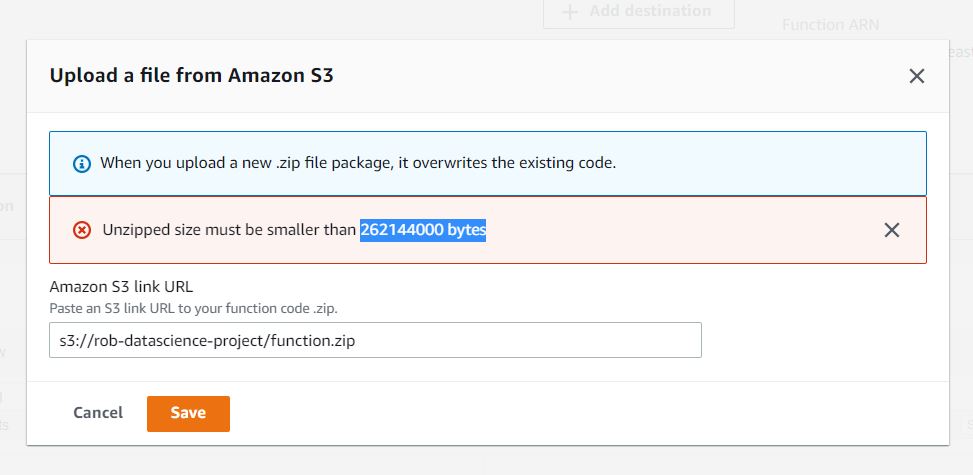


## Creating Lambda function and S3 bucket in AWS





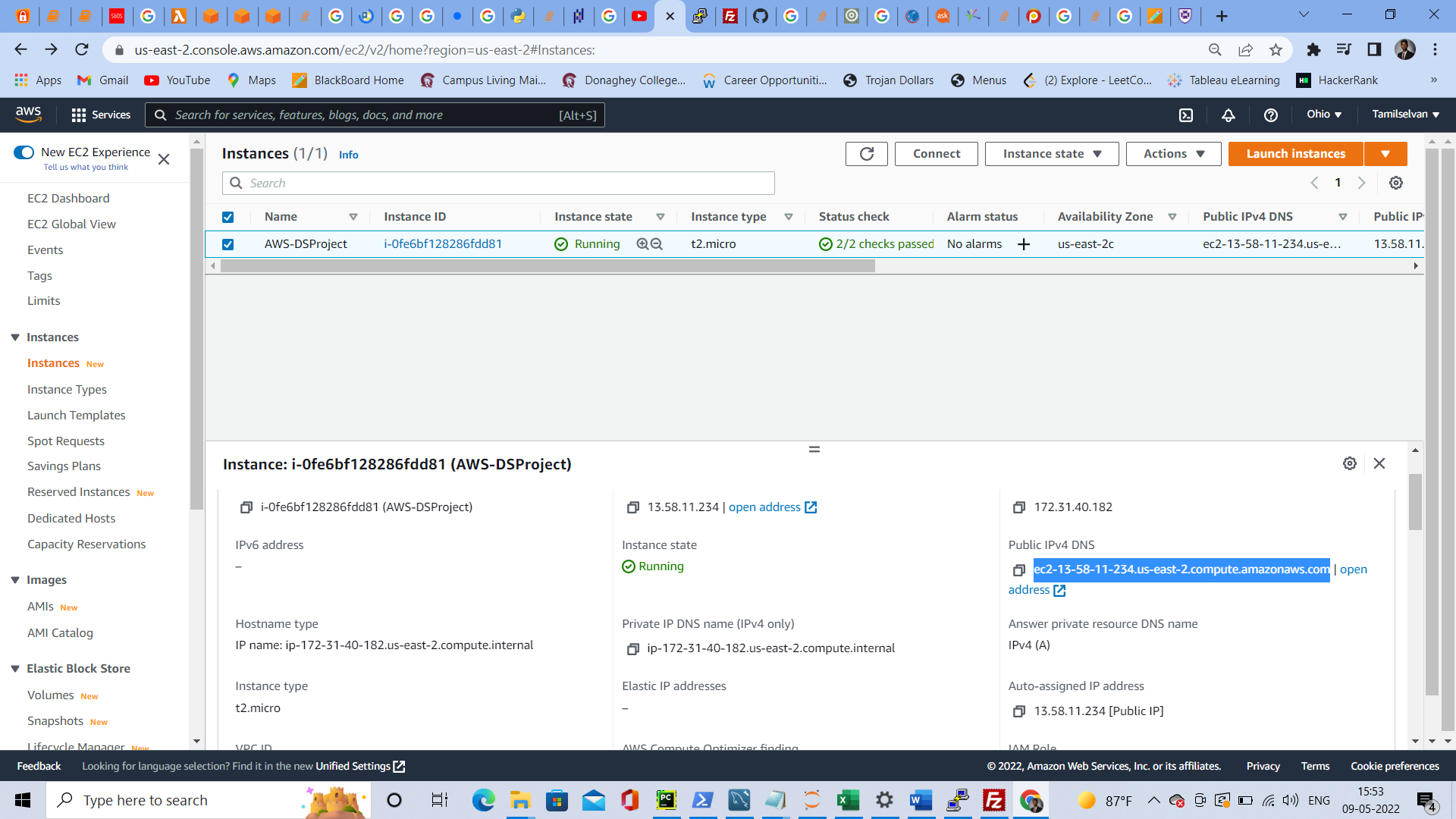


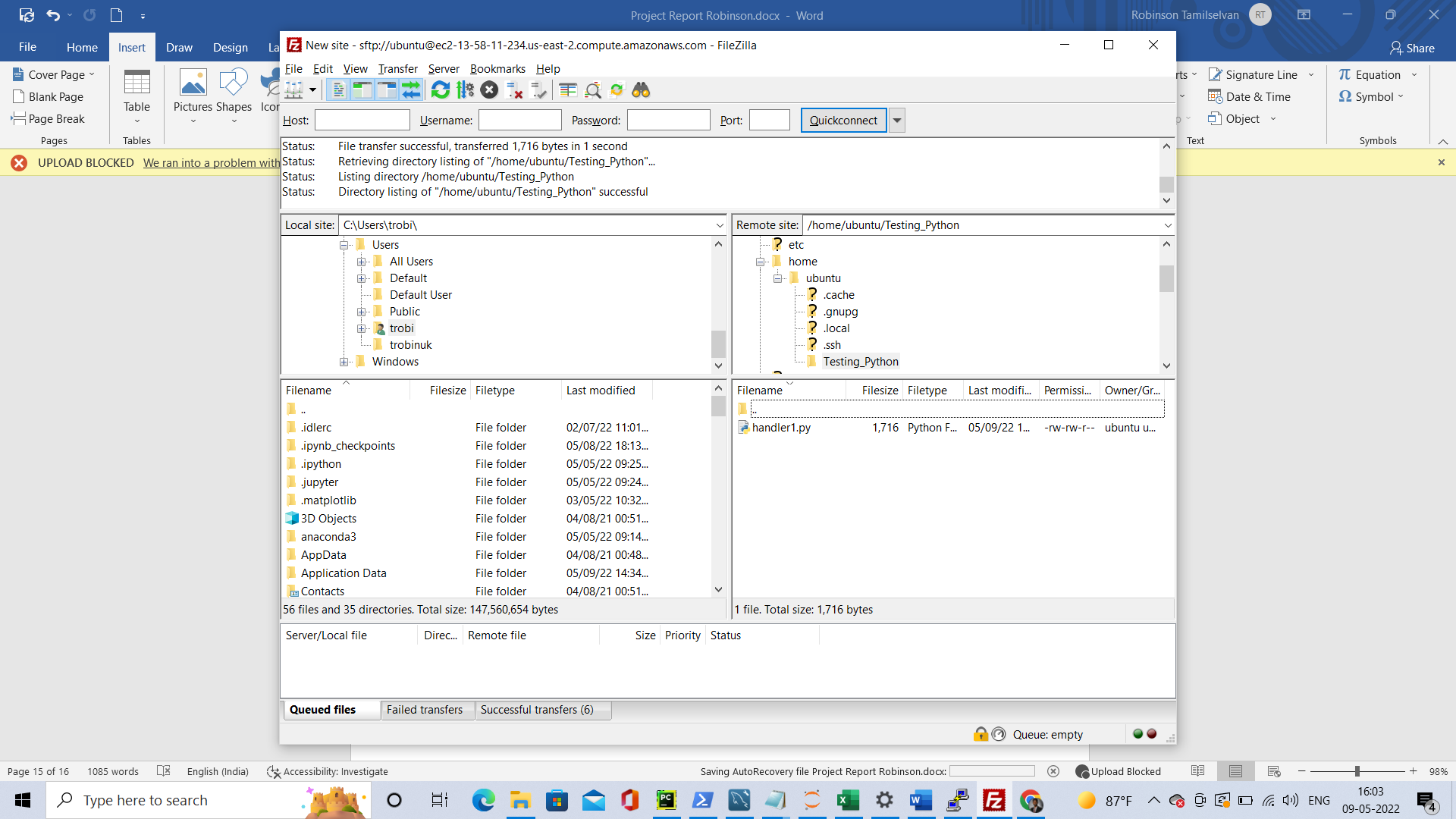


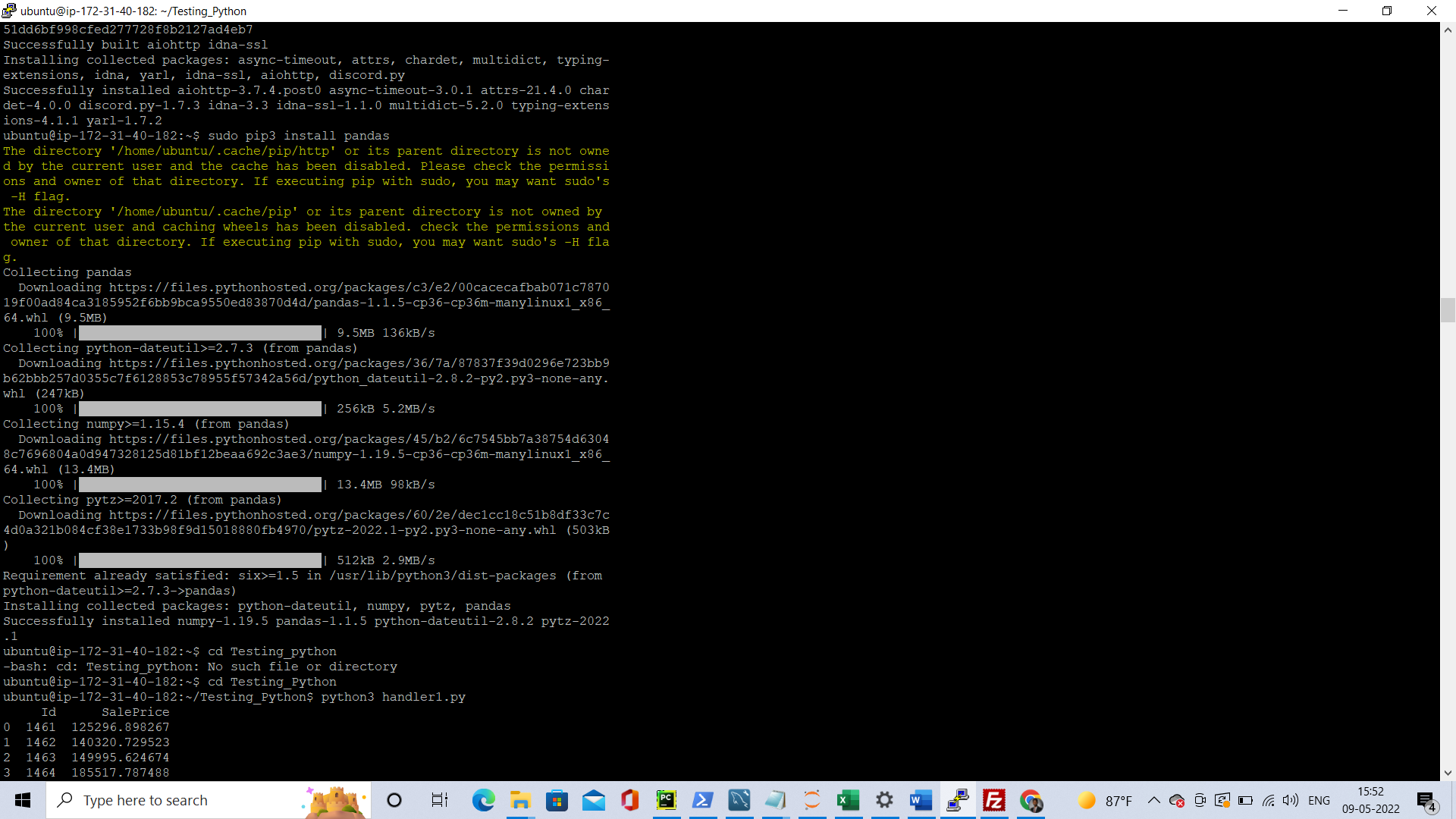
Converting the .py file and all the libraries into zip file and upload to AWS Lambda function (Serverless) but since for my Project involves lots of libraries the file size is large and I couldn’t do it using the AWS Lambda function. Hence, I am trying another AWS Service which is EC2.

## Creating EC2(AWS) Instance-Ubuntu server to compute the data

I installed AWS EC2 instance – Choose ubuntu server and executed the .py file in the Cloud (Ubuntu Server)







Here I have stored my code in the .py file Created a EC2 AWS Instance – Ubuntu Server and then used putty to connect with Ubuntu server and also used Filezilla to copy the file into the directory and executed the .py file in the Cloud and got the desired result.

## Packages used

In addition to ModelAuto and pandas which were used in normal environment. For cloud additionally I used **sqlalchemy** and **pymysql** to connect to MYSQL database.

# Conclusion

In summary, in this project I have predicted the house pricing by creating and training a machine learning model. I have also used various AWS cloud services like Lambda function, S3 bucket, EC2 - Ubuntu server, RDS - MYSQL database to perform Extract, Load, Transform operations and compute the data using the same Machine Learning model in the cloud and produced the results.