\_50\_STARTUPS POSSIBLE DATALOSS PREDICTION

1. INTRODUCTION

1.1.OVERVIEW

In this article, we will be predicting the profit from the startup’s dataset with the features available to us. We’re using the 50-startups dataset for this problem statement and we will be using the concept of Multiple linear regression to predict the profit of startups companies.

1.2.Purpose

Well, we can say that startups pipeline operates on the same principles which are similar to other MNCs the major difference between both of them is that on the one hand startups work to make products that are beneficial for the customers on a small scale while other established companies do that work on a large scale by re-doing something which is already being done.

As I mentioned above the startups are not such economical balanced company that has covered a path from an idea to a product so for the same reason no established investor will be going to come forward for those companies which don’t really have their market value hence, startups allow the**early investors** to start supporting in the format of [**seed funding**](https://www.feedough.com/seed-funding-explained/) which would help them to make a product out of their idea. In a nutshell, we can see that it’s hard to manage and analyze the investments and to make a profit out of them.

**We need a way by which we can analyze our expenditure on the startups and then know a profit put of them!**

How this model can help here?

This machine learning model will be quite helpful in such a situation where we need to find a profit based on how much we are spending in the market and for the market. In a nutshell, this machine learning model will help to find out the profit based on the amount which we spend from the 50 startups dataset.

2.LITERATURE SURVEY

2.1. Existing problem

This section describes the data that we used for testing our solution. We present the available features and the flow collection tool we used, as well as any modifications made in the original datasets. Our packet loss prediction model operates on flow data. We define a network flow as a five tuple identifier of Source IP, Destination IP, Source Port, Destination Port and Communication Protocol (TCP or UDP). Our data are collected from the 10 systems referred to as Data Transfer Nodes (DTNs) located at the National Energy Research Scientific Computing Centre (NERSC) [14]. DTNs are explicitly dedicated and fine tuned for performing large data transfers between the NERSC scientific facility and the external scientific community. They tend to have low-latency, high-bandwidth network interface cards (NICs) and I/O systems designed to limit disk-related bottlenecks. A variety of tools such as Globus online [15] and Grid FTP [16] are typically used in order to automate transfer of large datasets. We collected flow-like data from each DTN using the “tast” network monitoring tool [17]. Tstat is able to aggregate packet traces into flows and derive detailed statistics and performance metrics for each flow. Grouping packets into flows is particularly useful for optimizing efficiency in processing large amounts of network data. As opposed to other flow collection tools like NetFlow [18], tstat records non-sampled network data and also computes a wider set of performance features (the full list of the 53 metrics can be found here [17]). Furthermore, for ensuring anonymity of source and destination hosts, we drop the last octet of source and destination IPs. In our analysis we only use flow data where the percentage of packet retransmissions is greater than zero. We opt for flows that demonstrate packet loss and discard “perfect, loss-free data transfers.

2.2. Proposed solution

This particular dataset holds data from **50 startups in New York, California, and Florida.**The features in this dataset are**R&D spending, Administration Spending, Marketing Spending, and location features,**while the target variable is:**Profit. –**[Source](https://www.kaggle.com/farhanmd29/50-startups)**.**

**1. R&D spending:**The amount which startups are spending on Research and development.

**2. Administration spending:**The amount which startups are spending Admin panel.

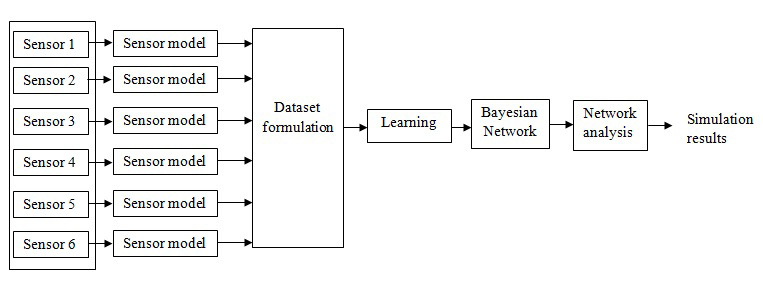
**3. Marketing spending:**The amount which startups are spending on marketing strategies.

**4. State:**To which state that particular startup belongs.

**5. Profit:**How much profit that particular startup is making.

3.THEORITICAL ANALYSIS

3.1.Block diagram



3.2. Hardware/Software designing

Hardware and software requirements of the project

Hardware: PC

Software: Anaconda (Jupyter)

Anaconda is package manager. Jupyter is a presentation layer. Anaconda tries to solve the dependency hell in python—where different projects have different dependency versions—so as to not make different project dependencies require different versions, which may interfere with each other.

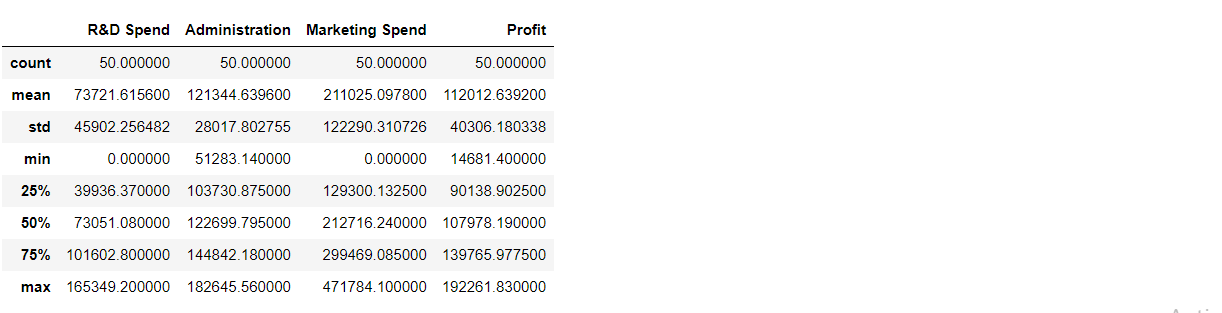
The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python or R language script. Once saved, you can share these files with others. NOTE: Python and R language are included by default, but with customization, Notebook can run several other kernel environments.

4.EXPERIMENTAL INVESTIGATIONS

**Numerical/Statistical analysis of the dataset**

dataset.describe()

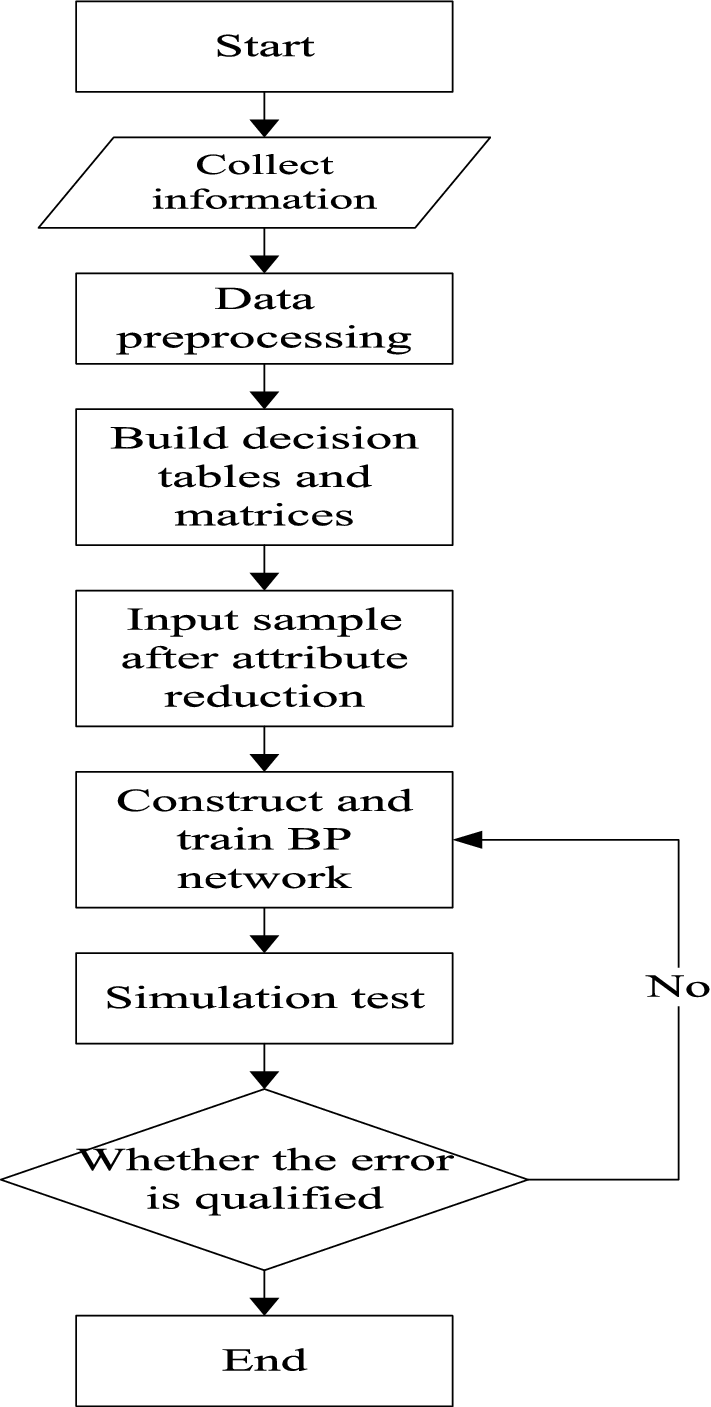
**Output:**



As a field, machine learning is closely related to computational statistics, so having a background knowledge in statistics is useful for understanding and leveraging machine learning algorithms. For those who may not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables. Correlation is a measure of association between two variables that are not designated as either dependent or independent. Regression at a basic level is used to examine the relationship between one dependent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities. Approaches to machine learning are continuously being developed. For our purposes, we’ll go through a few of the popular approaches that are being used in machine learning at the time of writing. The training process involves feeding the training dataset through the graph and optimizing the loss function. Every time the network iterates through a batch of more training images, it updates the parameters to reduce the loss in order to more accurately predict the digits shown. The testing process involves running our testing dataset through the trained graph, and keeping track of the number of images that are correctly predicted, so that we can calculate the accuracy.

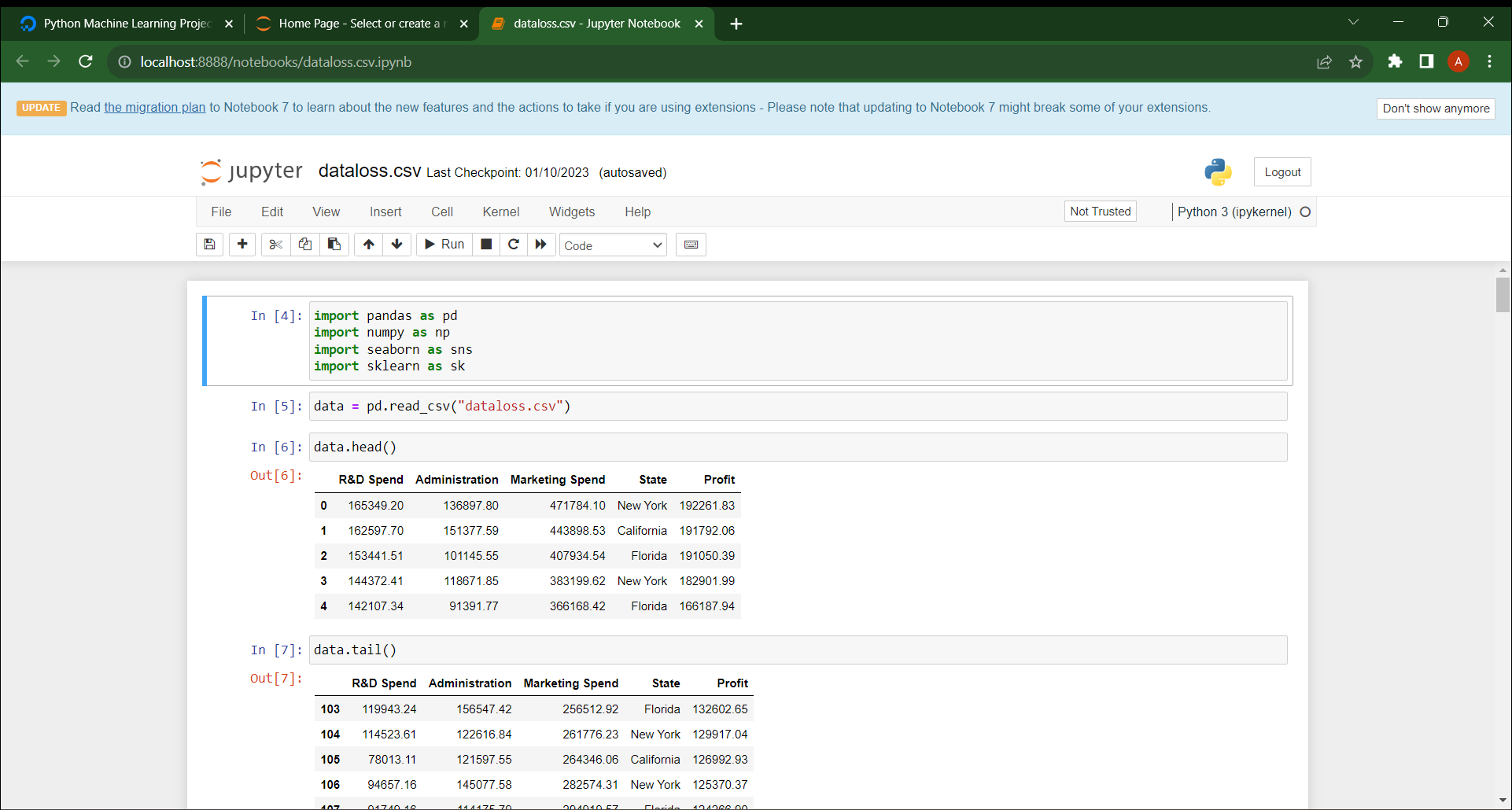
5.FLOWCHART

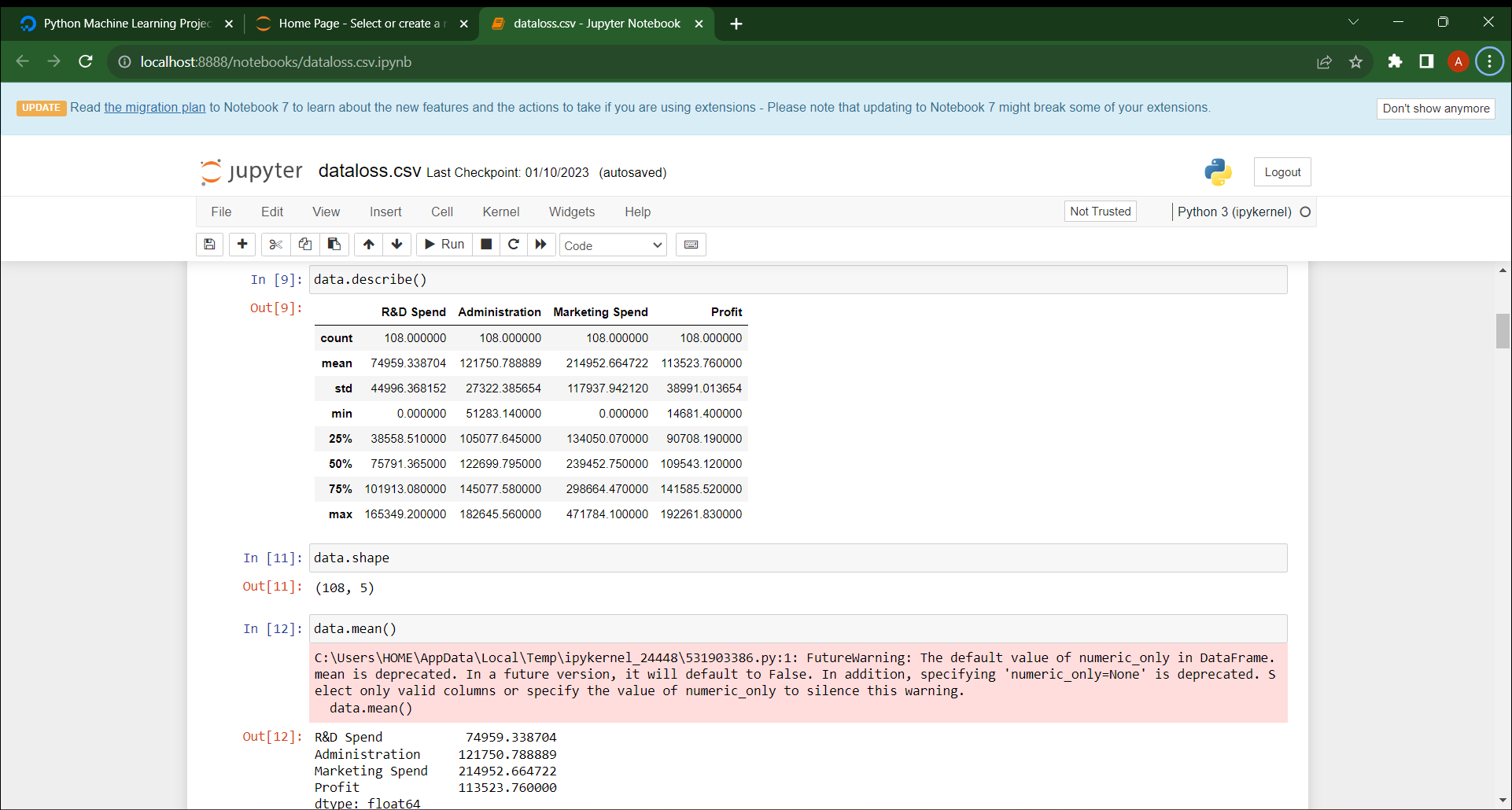
5.1.Diagram showing the control flow of the solution

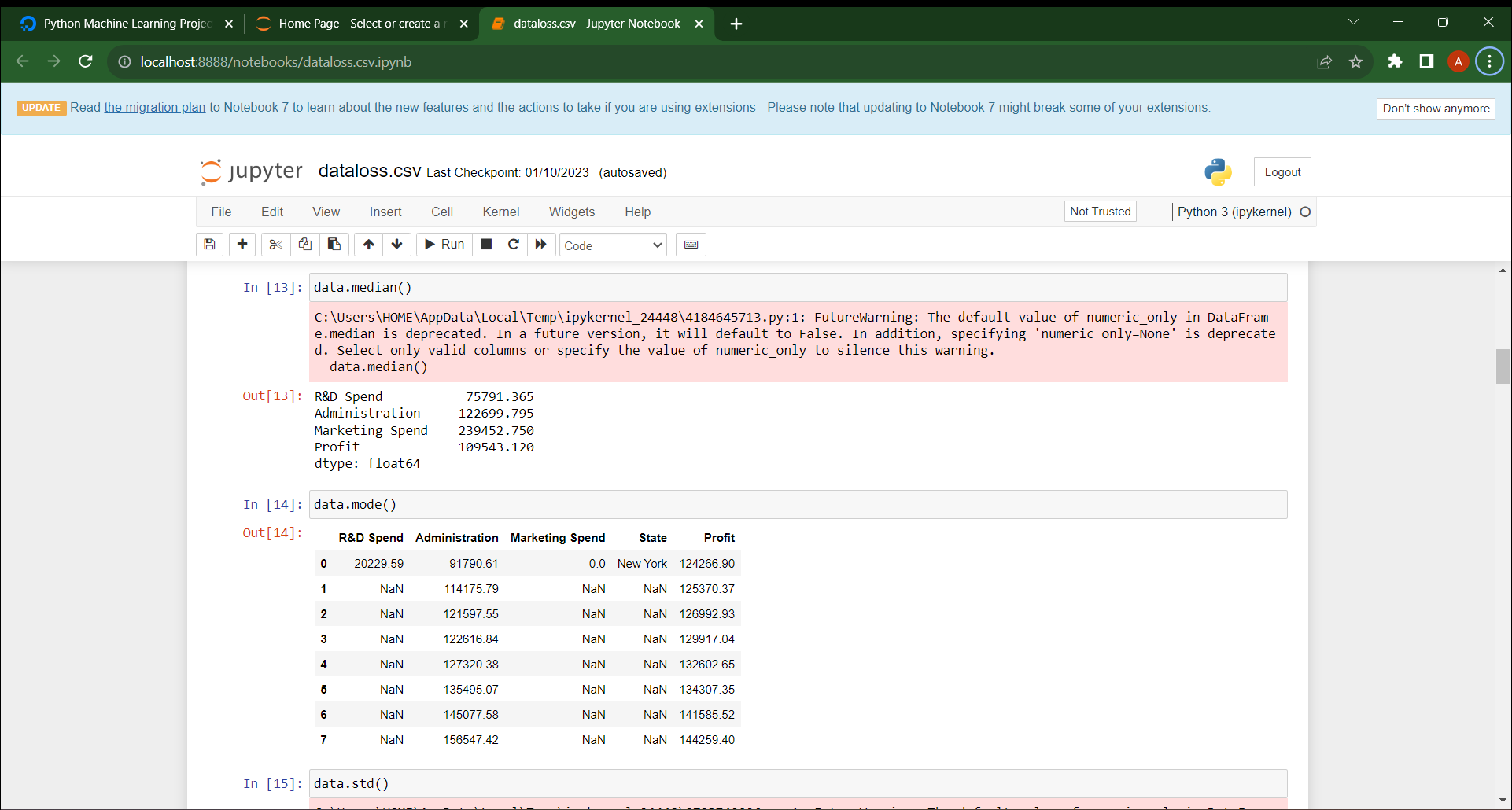


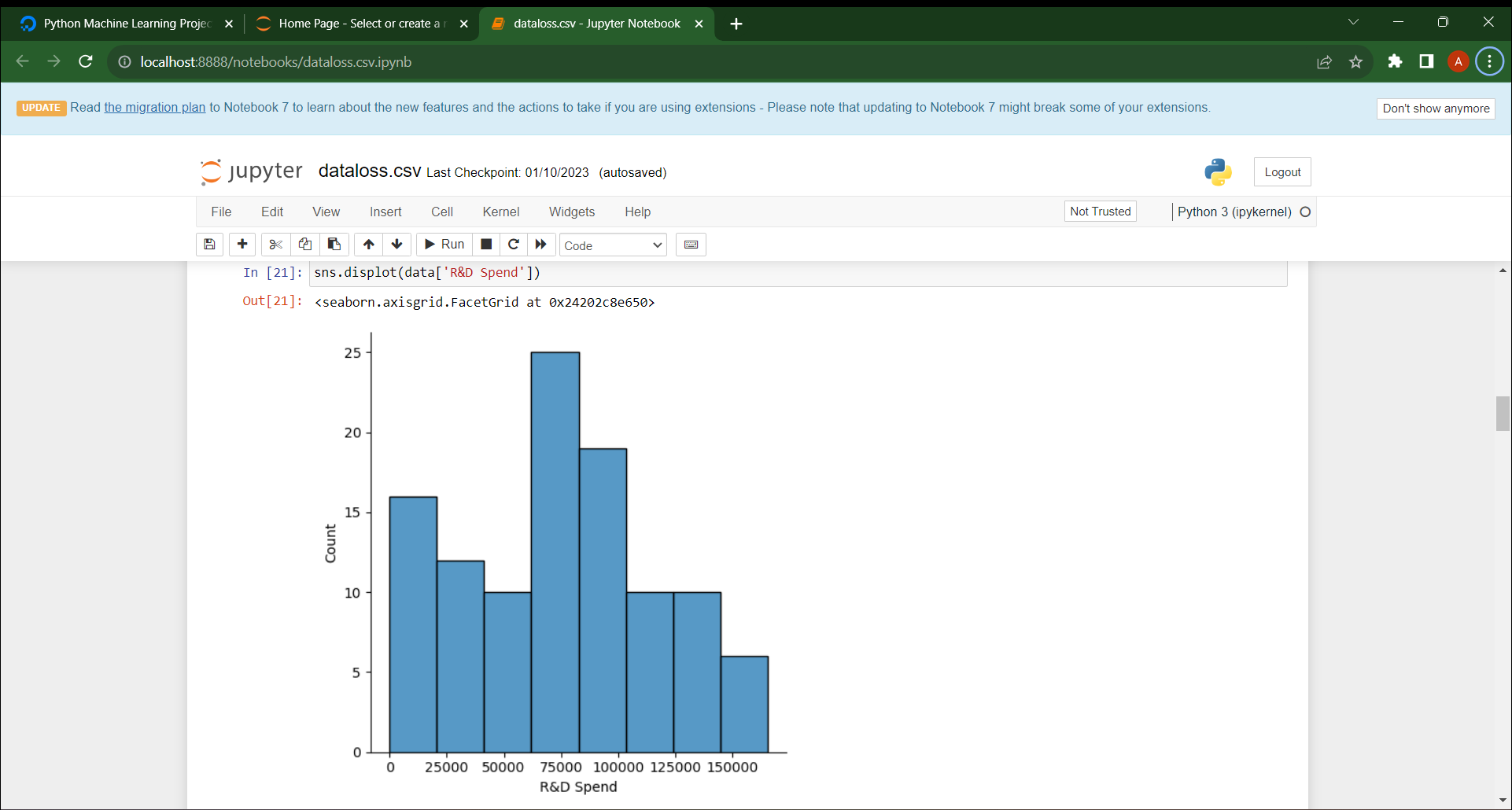
6.RESULT

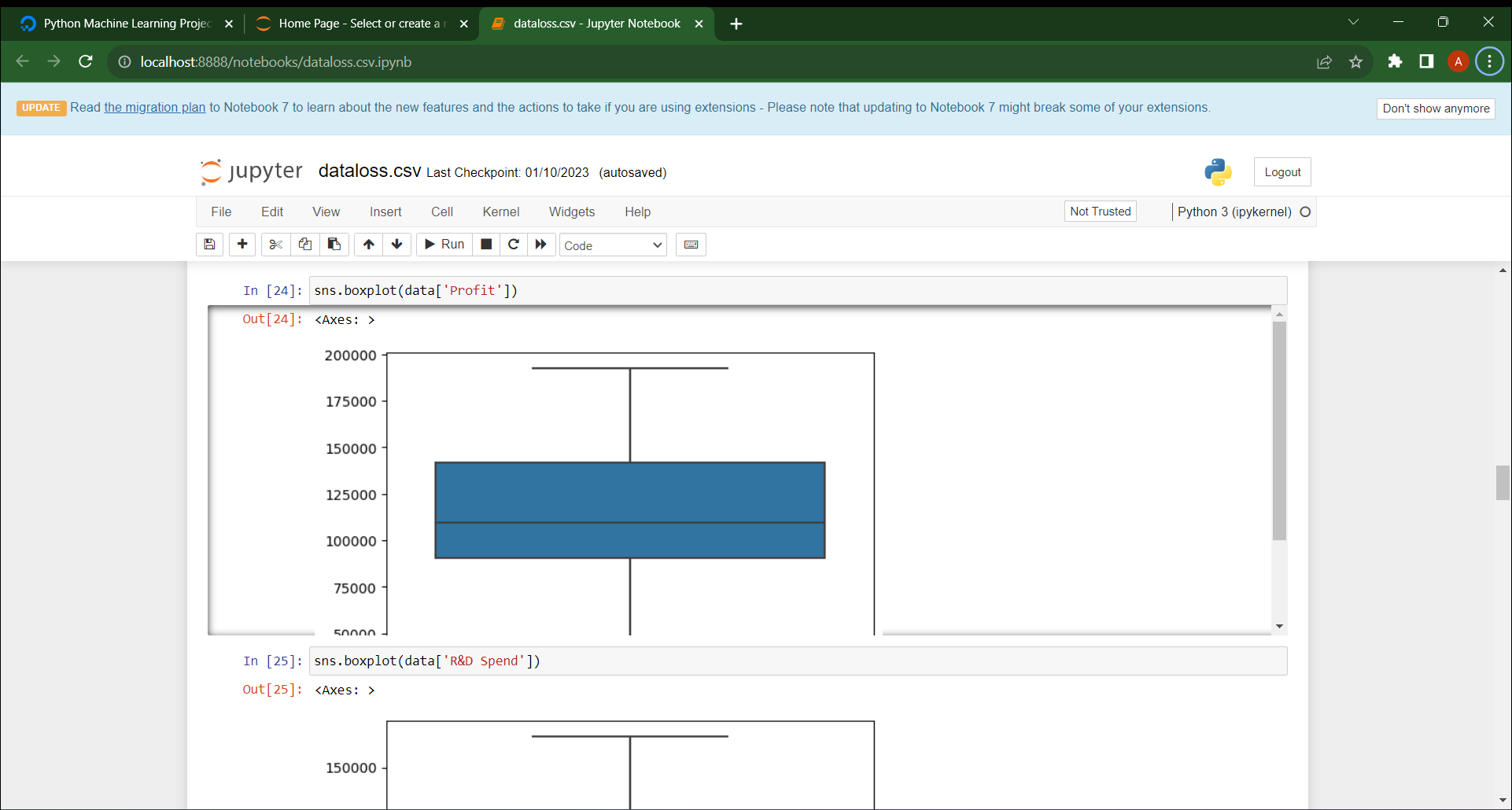
6.1.Final findings of the project along with screenshots

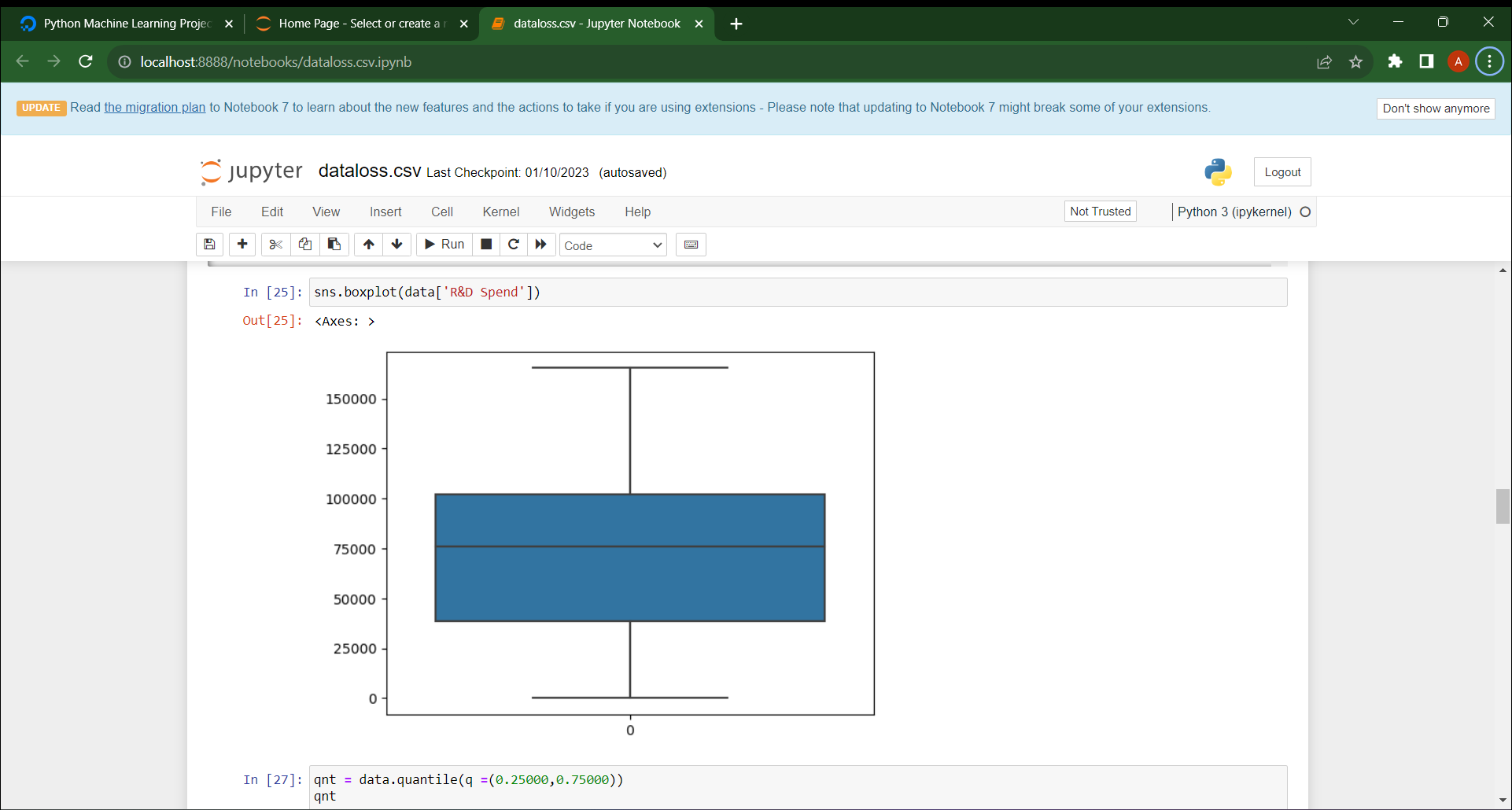


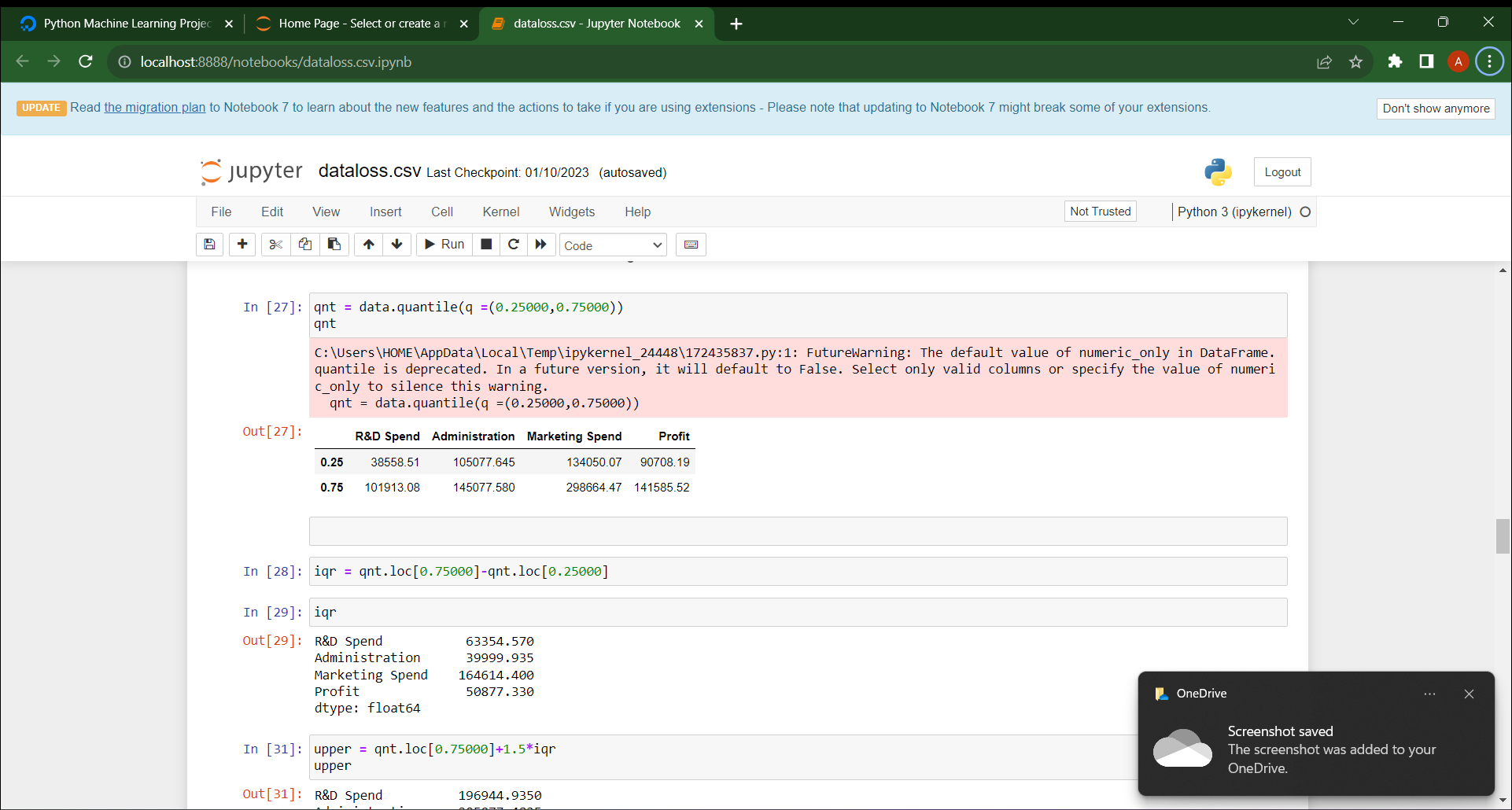


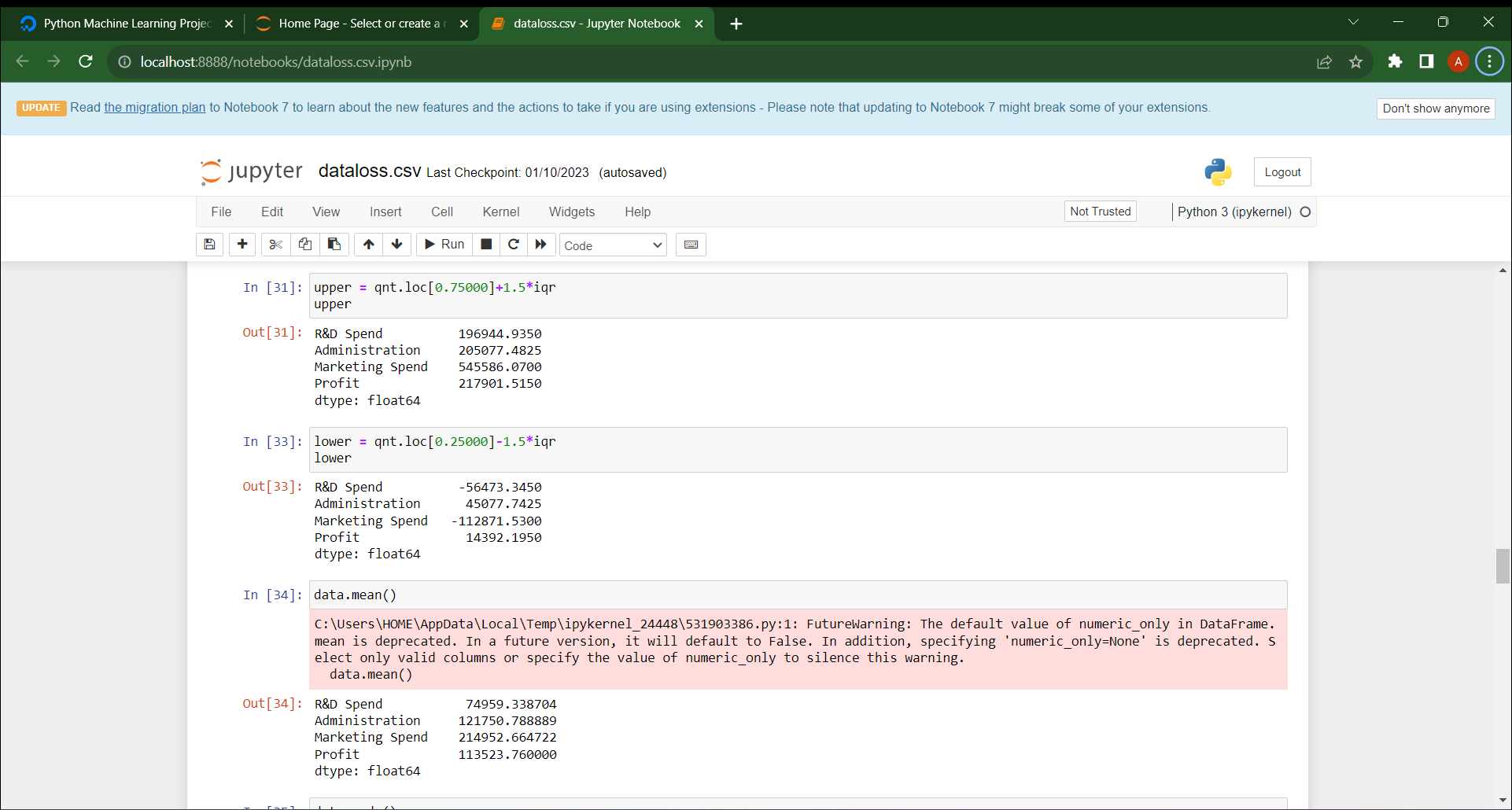


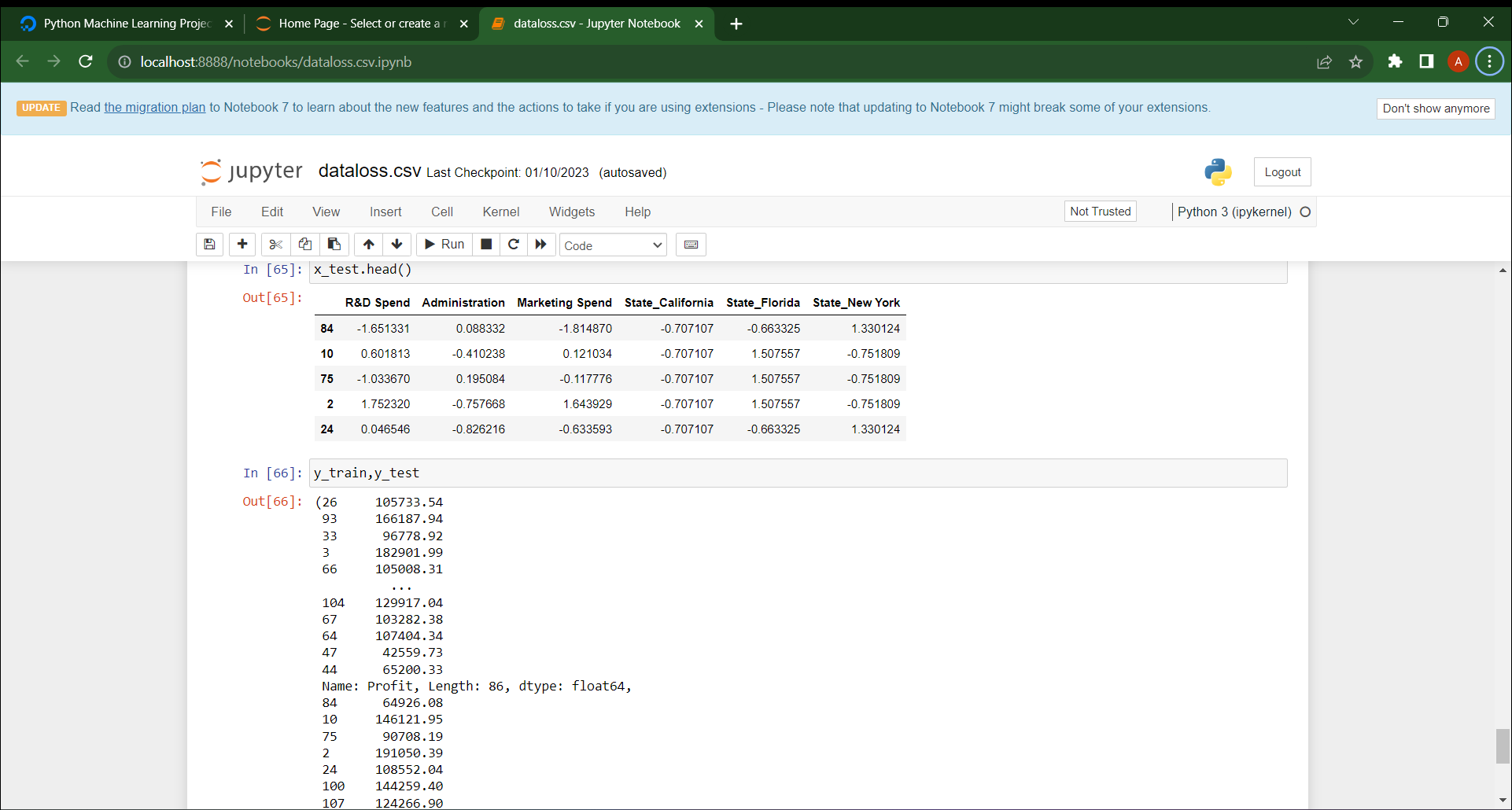


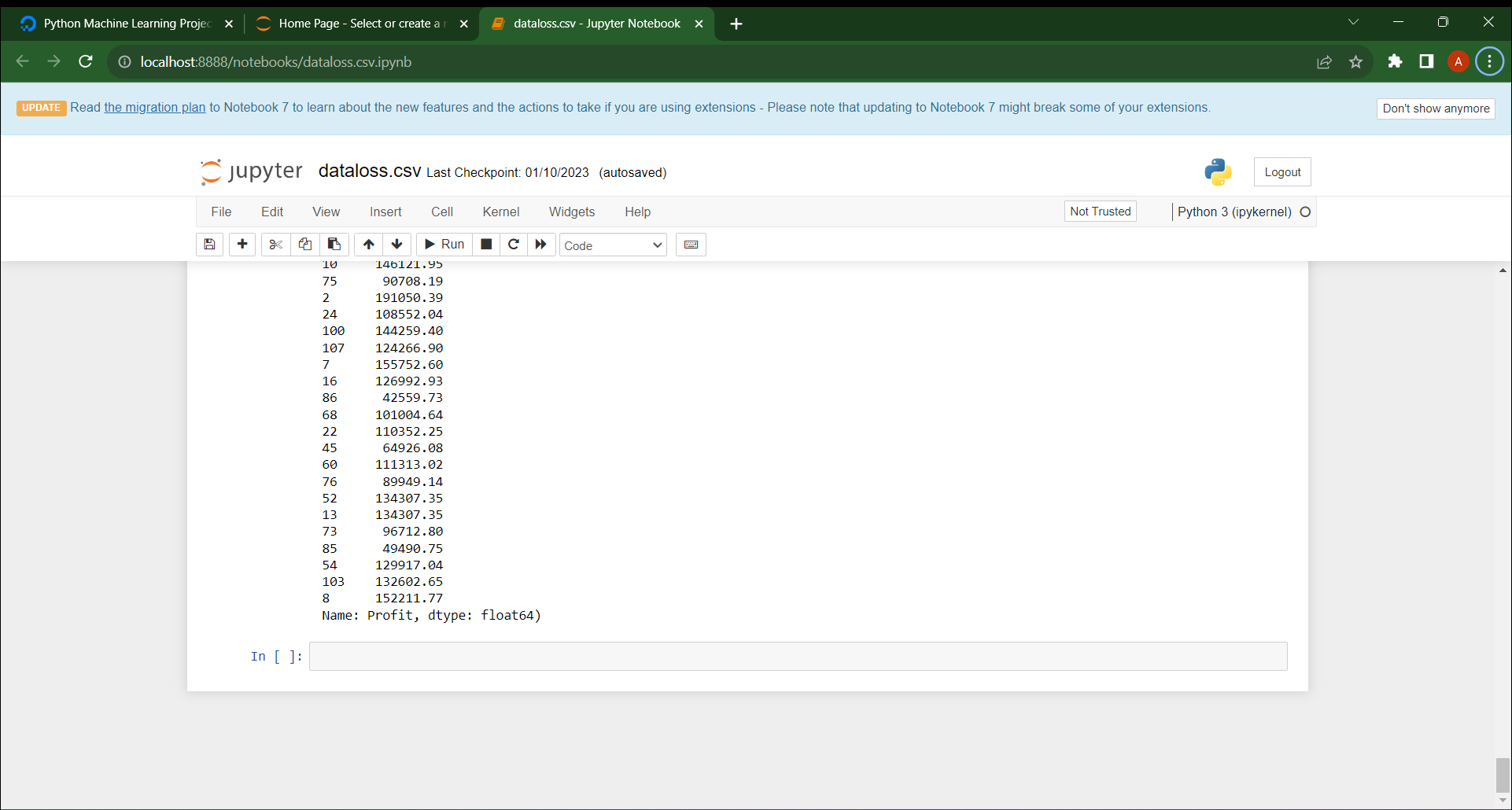












ADVANTAGES

**Freedom and Flexibility**

When you own a business, you are your own boss. This means that you decide your schedule, call the shots for your business, and get to create a business that is unique to your vision. Forget the typical nine-to-five day job — as a business owner, your time isn’t dictated by punching a clock. Instead, you get to dedicate your time to a project you’re passionate about that is centered around your business goals and ambitions.

**Creativity**

As the founder, your startup is based on your vision. This means you have the opportunity to get creative in your approach and really see your ideas come to life. Whether you’re in a creative industry or not, entrepreneurship by its nature encourages a creative perspective towards everything from the startup idea to daily problem-solving. Not to mention, you have the final say over branding and marketing materials – giving you the opportunity to put your artistic side to work.

### **Valuable Relationships**

### Networking is a huge part of startup ownership. While this may sound a little daunting, networking with other entrepreneurs, founders, and professionals can result in a community of motivated individuals who inspire you. Not only does this benefit you professionally, but surrounding yourself with a network of passionate, motivated people can also benefit your personal life and help you grow as an individual.

### **Personal Growth**

As an entrepreneur, you will almost certainly be required to overcome challenges that will test you. These hurdles, in many ways, are blessings in disguise. The challenges you face as an entrepreneur will contribute to your professional growth and help you evolve personally. From fine-tuning your communication skills to better manage staff and strengthening your problem-solving abilities, many of the lessons you will learn as an entrepreneur will manifest in a positive way through your personal life as well.

**Positive Impact**

As a startup founder, you have the opportunity to make a positive impact on your community and the world around you regardless of the type of startup you’re founding. Whether you’re a [**social entrepreneur**](https://startupsavant.com/what-is-a-social-entrepreneurship-startup) or simply prioritize incorporating equitable company practices that contribute to the betterment of society, you have the authority to use your position to do good.

DISADVANTAGES

**Long Hours**

Being your own boss means you get to make your own schedule. However, as a business owner, you should anticipate working longer hours than you would at a typical nine-to-five job. A huge part of [**being an entrepreneur**](https://startupsavant.com/entrepreneur-anatomy) is being willing to dedicate yourself and your time to maintaining and growing your business. This means less time spent on your hobbies and with friends and family, which is a dealbreaker for some entrepreneurs that cannot commit this much time.

**Stress and Uncertainty**

Launching your own startup poses inherent risks, including a degree of uncertainty even for the most well-planned businesses. Therefore, adaptability and resilience are critical traits for an entrepreneur to be able to combat the stress and uncertainty of founding a startup. As a cornerstone of entrepreneurship, it's a good idea to consider the level of stress and uncertainty you’re confident handling before launching your own startup.

**Self-Doubt**

There is a level of confidence that you need to have in order to start and maintain a successful startup. Most entrepreneurs will experience self-doubt to some capacity when they start their own business. Understandably, there is a great deal riding on their vision, leadership, and business savvy. As one of the cons of launching a startup, an entrepreneur's ability to persevere through bouts of self-doubt and potential failures is crucial.

**Financial Risk**

Whether you’re [**bootstrapping a startup**](https://startupsavant.com/tips-for-bootstrapping-your-startup) using your own personal savings or partnering with an investment firm that expects a high return on their investment, founding a startup requires unavoidable financial risk. While financial risk can be mitigated somewhat by business planning and financial forecasting, there is no way to eliminate financial risk completely. Therefore, before you start a startup of your own, you should be absolutely sure that you are comfortable taking on the financial risk.

8.APPLICATIONS

Beyond the asymptotic oracle performance of the super learner, our evaluation of the practical performance of the super learner shows that the super learner is also an adaptive and robust estimator selection procedure for small samples. Combining estimators with the weights (i.e. positive and summing up till 1) based on minimizing cross-validated risk appears to control for over-fitting of the final ensemble fit generated by the super learning algorithm, even when using a large collection of candidate estimators. The above examples demonstrate that the super learner framework allows a researcher to try many prediction algorithms, and many a priori guessed models about the true regression model for a given problem, knowing that the final combined super learner fit will either be the best fit or near the best fit. Combining estimators with the convex combination algorithm proposed here appears to also improve on the usual cross-validation selector (i.e. discrete super learner). Selection of a single algorithm based on V-fold cross-validated risk minimization may be unstable with the small sample sizes of the data sets presented here, while the super learner can average a few of the best algorithms in the library to give a more stable estimator compared to the discrete super learner.

9.CONCLUSION

With the development and deployment of 5G networks, network planning puts forward higher

requirements on the accuracy, complexity, and versatility of path loss prediction. Machine learning

methods, especially supervised learning, can model hidden non-linear relationships and thus can

be used for path loss prediction. Based on historical data, machine-learning-based models can build

relationship between path loss and input features. It has been shown that machine-learning-based

models, including ANN, SVR, and RF, are in good agreement with measured data. In order to satisfy the

demand for training data, two data expansion schemes have been proposed to make full use of existing

data and classical models. Through the measured data, the feasibility of the proposed schemes has also

been veriﬁed. Finally, we have summarized the problems still faced by the machine-learning-based.

10.FUTURE SCOPE

* **Real-time information:** The application gives us the current details of the market trends. Also, it uses techniques of Machine Learning to process the information and find the hidden trends in the data to provide us with the proper market information.
* **Stock prediction**: Upstox visualizes the data of traders and predicts the ups and downs of the market. For smart prediction, it uses Machine Learning algorithms. This helps us properly invest money in stocks with lesser chances of losing them.

### **Security**: The app uses built-in Machine Learning systems to predict fraudulent activities that make it secure for users.

### **Automotive Industry:** The automotive industry is one of the areas where Machine Learning is excelling by changing the definition of ‘safe’ driving. There are a few major companies such as Google, Tesla, Mercedes Benz, Nissan, etc. that have invested hugely in Machine Learning to come up with novel innovations. However, Tesla’s self-driving car is the best in the industry. These self-driving cars are built using Machine Learning, IoT sensors, high-definition cameras, voice recognition systems, etc.

### **Robotics:** Robotics is one of the fields that always gain the interest of researchers as well as the common. In 1954, George Devol invented the first robot that was programmable and it was named **Unimate**. After that, in the 21st century, Hanson Robotics created the first AI-robot, **Sophia**. These inventions were possible with the help of Machine Learning and Artificial Intelligence.

### 10.BIBILOGRAPHY

There were several facilitation techniques used by the trainer which included question and answer, brainstorming, group discussions, case study discussions and practical implementation of some of the topics by trainees on flip charts and paper sheets. The multitude of training methodologies was utilized in order to make sure all the participants get the whole concepts and they practice what they learn, because only listening to the trainers can be forgotten, but what the trainees do by themselves they will never forget. After the post-tests were administered and the final course evaluation forms were filled in by the participants, the trainer expressed his closing remarks and reiterated the importance of the training for the trainees in their daily activities and their readiness for applying the learnt concepts in their assigned tasks. Certificates of completion were distributed among the participants at the end.

11. APPENDIX

# Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score

# Load your dataset

data = pd.read\_csv('data\_loss\_dataset.csv')

# Define features and target variable

X = data.drop('data\_loss\_event', axis=1)

y = data['data\_loss\_event']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Create a decision tree classifier

clf = DecisionTreeClassifier()

# Train the classifier

clf.fit(X\_train, y\_train)

# Make predictions on the test set

y\_pred = clf.predict(X\_test)

# Evaluate the model's performance

accuracy = accuracy\_score(y\_test, y\_pred)

print(f"Accuracy: {accuracy}")