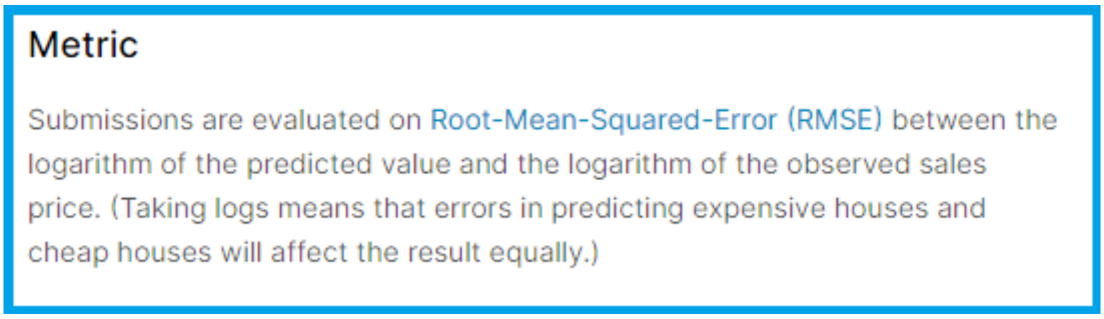
# 65020 Open Source Code for AI- Course Assignment

## Open -source descriptions:

This assignment is a Kaggle competition, the **House Prices: Advanced Regression Techniques** where you'll need to open a Kaggle account (see appendix 1).
In this competition you need to perform a Prediction task: To predict house prices in Iowa, USA
The data is described in the link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.
Read the description, understand the meaning of each column and think how each feature can impact the outcome = the house price.
For questions about the competition and data, write and read other participants' questions in the competition's forum: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion
You can also see other participants' notebooks and get inspiration from the way they solved the problem in: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/notebooks

Note that the score in the competition is given according to RMSE



Metric

Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

## Solution Specifications

The solution should be provided in one notebook, submitted in 2 forms: a ipynb file, downloadable from the File menu in the notebook, and html file that can be created from the ipynb file according to the instructions.

Upload your submission in a Zip (not RAR, please..) containing both files.

Write in the notebook your name, student number and a link to your user page in Kaggle.

## Solution Components and Structure

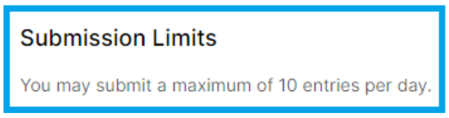
1. Data analysis.

Machine learning assignments can be looked at as riddle. The first step is solving a riddle is understanding all the data and deriving preliminary conclusions. Then, you can start looking for solutions.

Use visuals - graphs and tables - and explain and discuss your analysis in the Markdown cells. This is the place to investigate the features provided in the dataset and your additional or transformed features – how do they affect the prediction?

Practical recommendation: Divide the Training Dataset into two: Train and Validation. Use the Validation set as Test set for your experiments before the real submission.

Note that you can formally submit up to 10 submission per day:


Submission Limits

You may submit a maximum of 10 entries per day.

So, in order not to waste them on experiments, it is recommended to test your theories on the Validation Set, before a full training on the entire Training set and the submission of the Test for evaluation.

2. Machine Learning (ML) Choice

Try Linear Regression and additional 2 other prediction methods. Explain your choices, compare and discuss the results of your choices

3. ML Tuning

Try different values of hyper-parameters, try training on sub-sets of features and examine what are the best features for the prediction.

Try using Grid Search, Random Search for hyper-parameters search.

Provide graphs of the Training error (Loss) and Validation loss, for different hyper-parameters, and the training and validation accuracy. You can compare the training and validation performance.

Use different Cross validation (CV) methods, try feature selection algorithms, Ensembles (Bagging , Boosting) and more..

4. Submission File

Finally, create a submission file and submit to the competition. Print your submission page (if there are more than 10, print only the last 10 submissions) and highlight your best one. Upload the page print and your place in the Leaderboard to the notebook.

Briefly summarize your work. Explain what you did, what worked well and what did not so.

Put reference list of notebooks that inspired your work as well as the sites and tutorials that you learned from.

## Notebook Structure

1. Name, student number, link to your Kaggle user
2. Introduction: explaining the competition, what you're trying to do and what means you'll use
3. Data exploration
4. Experiments on different features, different models and different Hyper-parameters (reminder: evaluate on the validation set).
5. Graphs, tables and results analysis

6. Prints of your submissions and your Leadership place

7. Summary

8. References list

## Marking Structure

Each one of the speculation and notebook structure items above counts in your mark. The visibility of the notebook – organization and clarity - also counts, as well as the clear and simple code.

Remember that hi-tech today is working with people, within and across company groups – other people will need to understand your report in a simple and convenient way.

So:

Code = 20% (simple, organized, explained and clean)

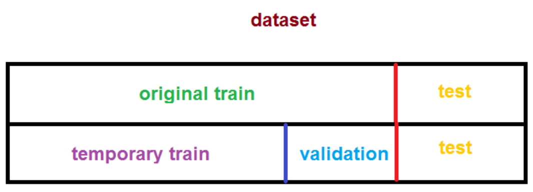
Notebook = 20% (organized, clean and easy to read, containing explanations, conclusions and discussion)

Correct Implementation of the solution specification and notebook structure = 50%

Extra = 10% (self-learning, exceptional aspects in solution, code or writing).

## Notes

For experimenting before the submission, you should divide your original Training set into Temporary Train and Validation, as illustrated below:



Where the Temporary Train serves for training and the validation serves for testing. The errors in this testing will guide us if our modeling is good.

When you settle with the final feature-set, model and hyper parameters , you can go ahead and re-train on the original Training set and provide the submission on the Test set.

You can experiment with different divisions of Temporary Train and Validation sets. You are welcome to source online for explanations on these division strategies.

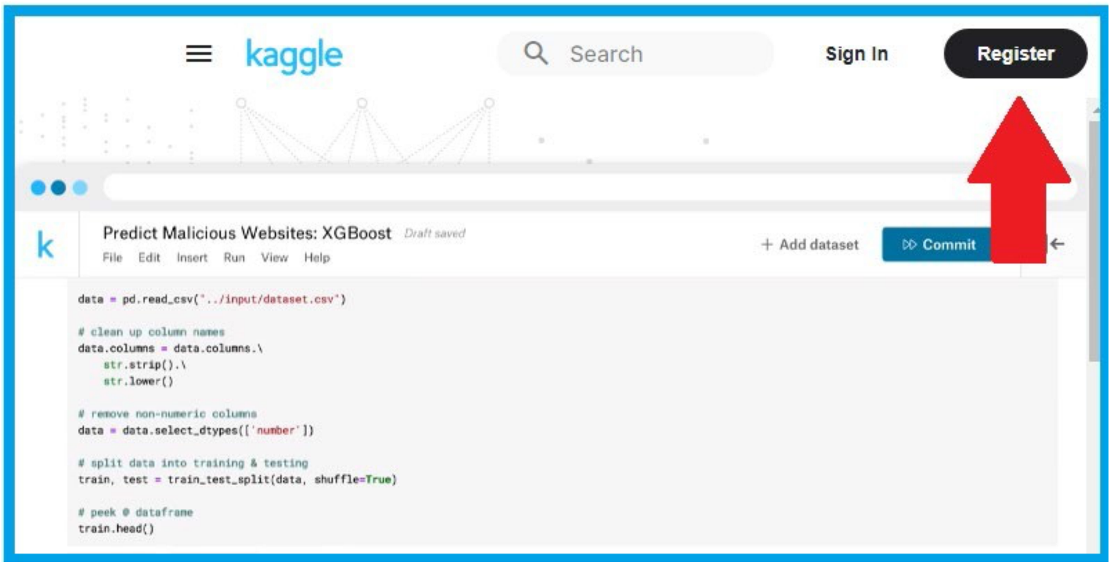
Basics: Try to use targeted, short functions and to avoid double-codes. Give your functions and variables meaningful names and document the operation of each function.

Important ML note: In many real-life cases there is no “right” and “wrong” solutions. ML is about trials and errors and interpretation/understanding of yours (and the machine’s) moves. The only “wrong” is incorrect methodology, i.e. training and testing on the same set.

## Appendix 1: creating a user in Kaggle

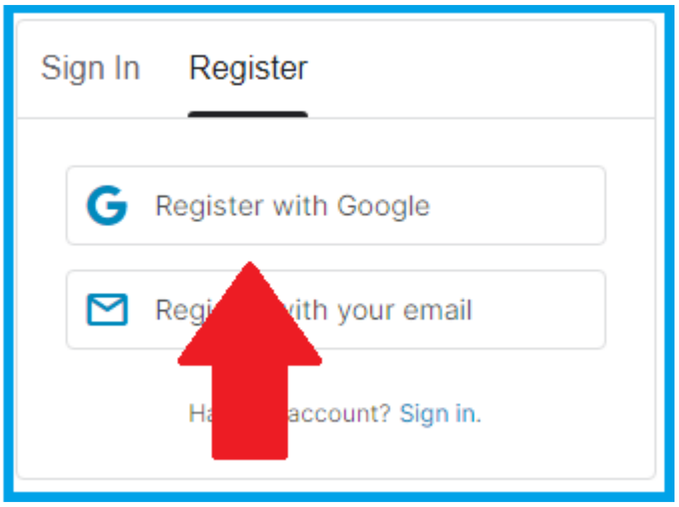
/https://www.kaggle.com

Register


kaggle
Search
Sign In
Register
Predict Malicious Websites: XGBoost Draft saved
File Edit Insert Run View Help
+ Add dataset
Commit

```
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()
```

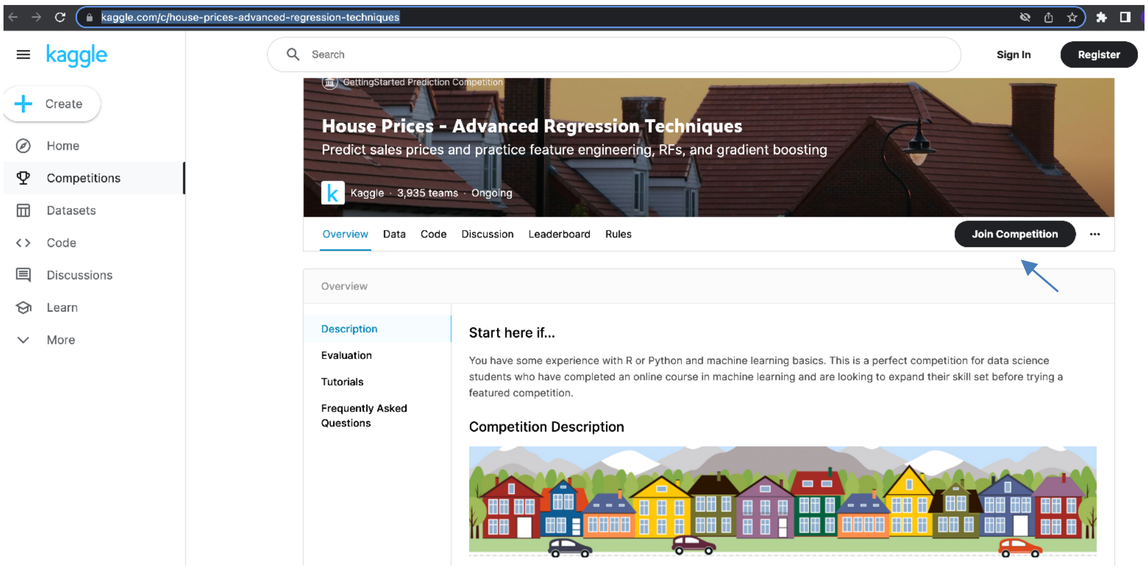
Register with Google



## Joining the competition

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

Join Competition


kaggle.com/c/house-prices-advanced-regression-techniques

GettingStarted Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 3,935 teams · Ongoing

Overview

Start here if...

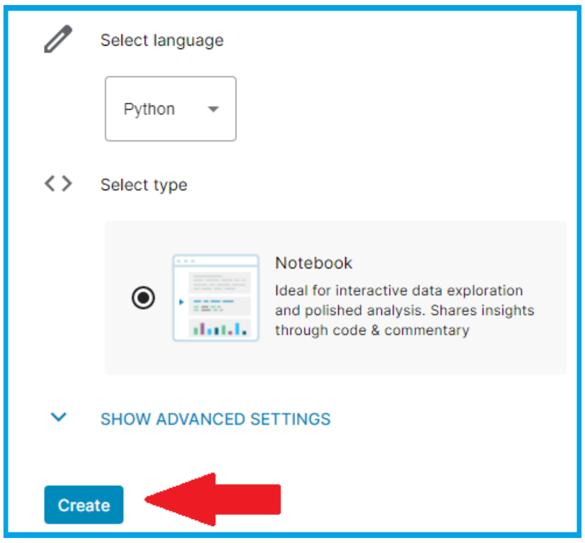
You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

And open a new notebook:

press “code” and then “new notebook”

Ensure the language is Python and the file type is Notebook, then press **Create**


Select language: Python

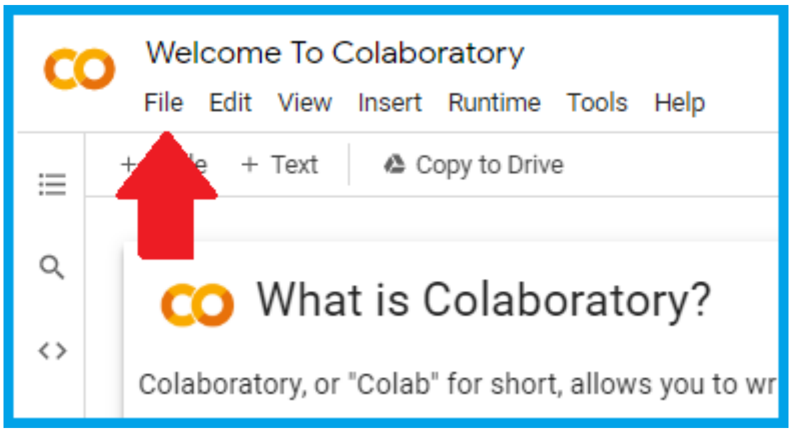
Select type: [✓ Notebook]

Notebook

Ideal for interactive data exploration and polished analysis. Shares insights through code & commentary

## Appendix 2: Converting ipynb to html

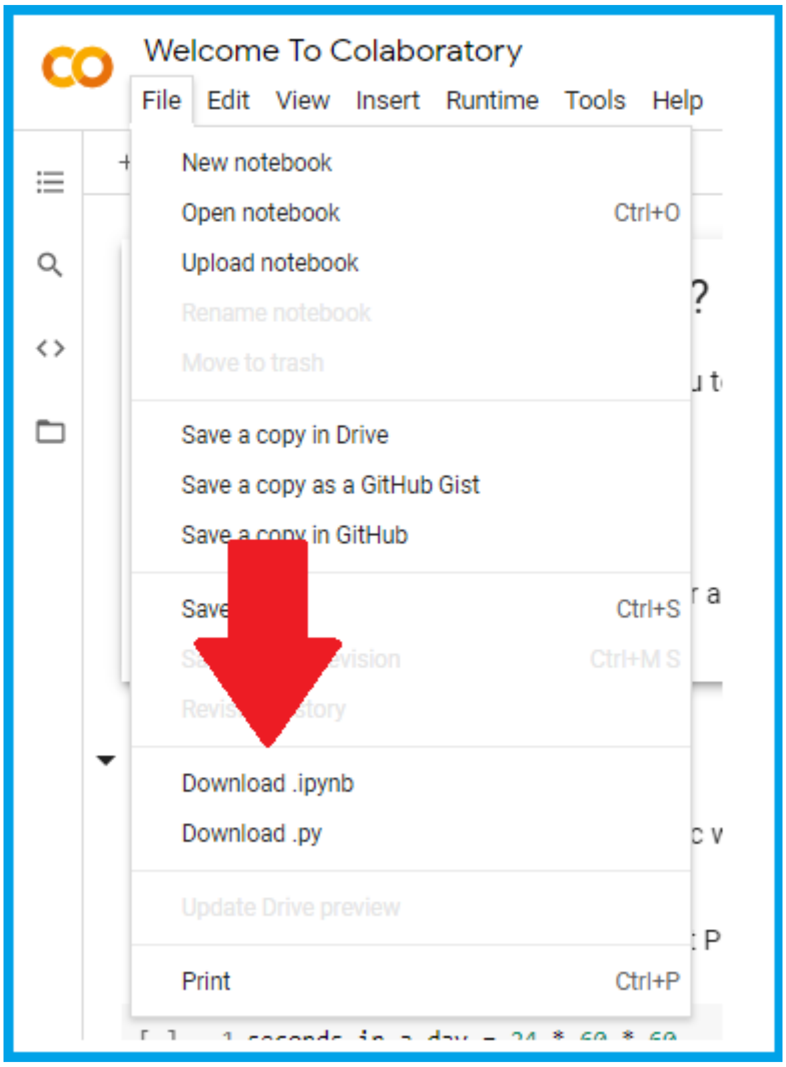
When opening the notebook, press File

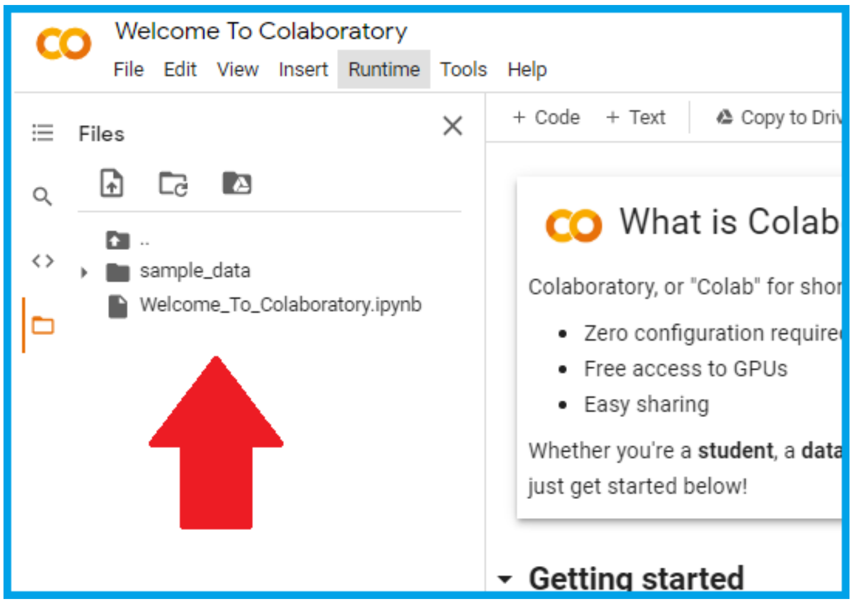


What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to wr

**In the file menu press Download.ipynb**





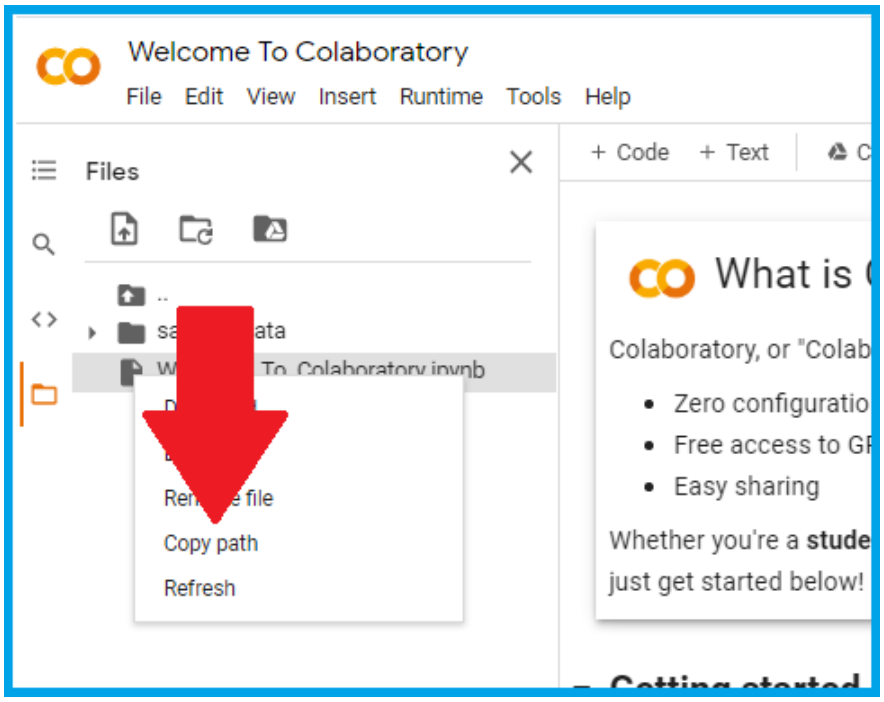
Colaboratory, or "Colab" for shor

- Zero configuration require
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data**
just get started below!

When the notebook is downloaded, upload it to the files area in Colab

Press right-mouse on the notebook, and in the menu choose Copy path

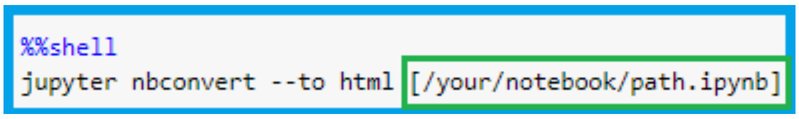


Colaboratory, or "Colab

- Zero configuratio
- Free access to GI
- Easy sharing

Whether you're a **stude**
just get started below!

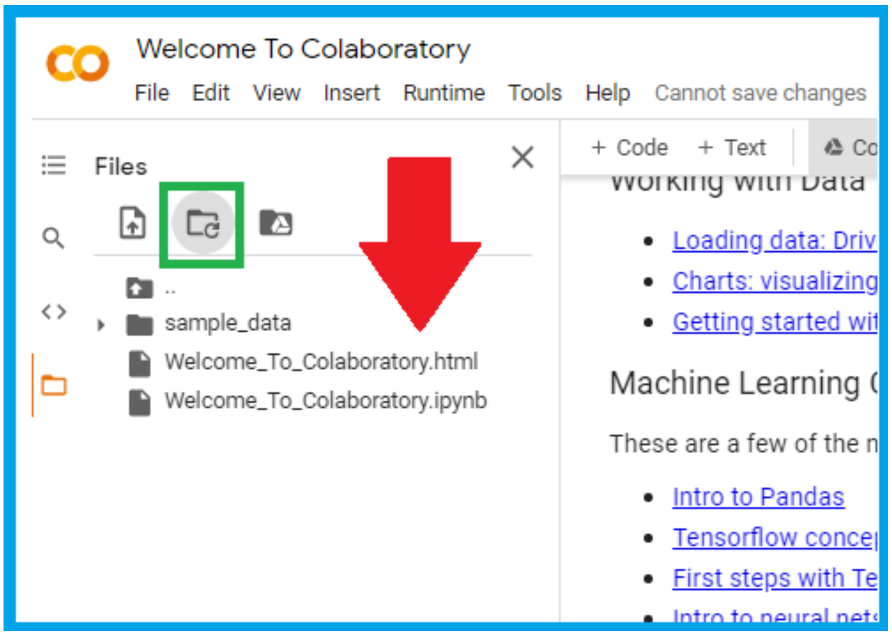
Add a code cell in the notebook and write:



```
%%shell
jupyter nbconvert --to html [/your/notebook/path.ipynb]
```

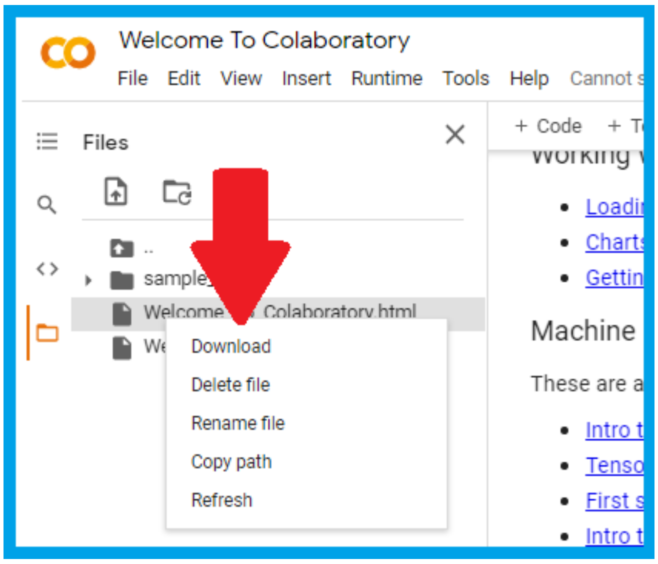
Where instead of the squared brackets (in the green rectangle) paste the path you copied

Run the cell  - an HTML file is created next to the ipynb that you uploaded:


Welcome To Colaboratory
File Edit View Insert Runtime Tools Help Cannot save changes
Files
+ Code + Text
sample_data
Welcome_To_Colaboratory.html
Welcome_To_Colaboratory.ipynb
Machine Learning
Intro to Pandas

Note: if this file is not created, press “refresh directory” icon (green square) and it will appear

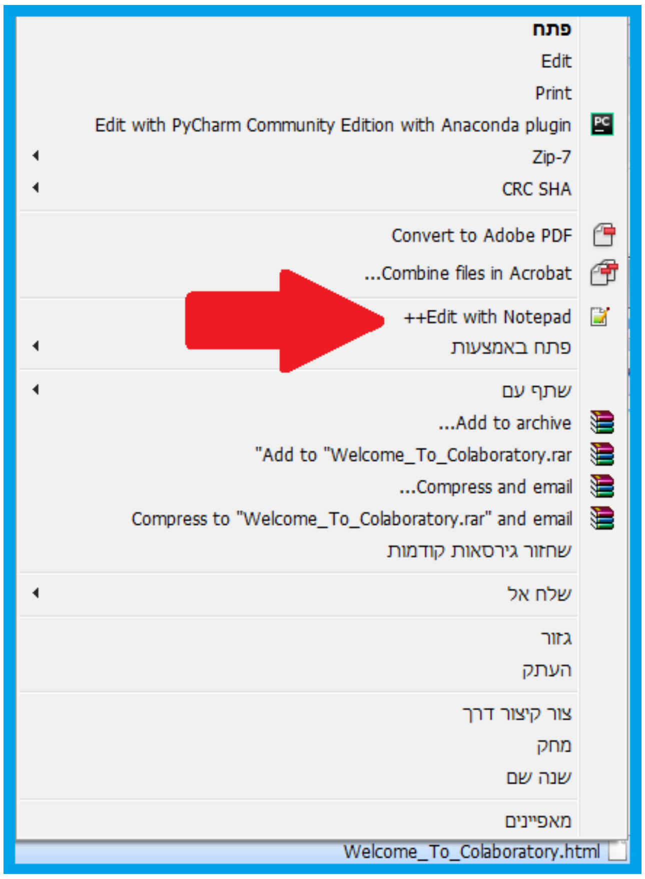
Download the HTML file by clicking the right mouse and choosing Download



## Appendix 3: Plotly graphs in HTML file

When an HTML file created from ipynb that contains Plotly graphs, they can work within the HTML by adding a script prior to all other scripts in the file

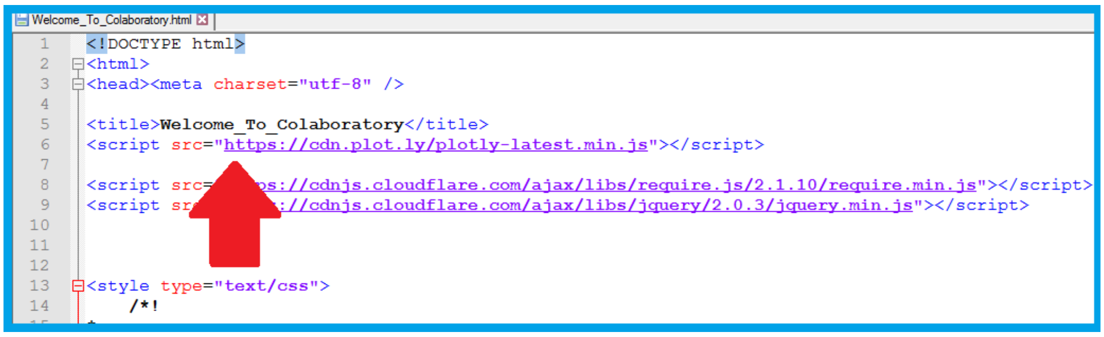
Install Nogtepad++ (open source) , then click on your HTML file and in the menu choose “Edit with Notepad”


פתח
Edit
Print
Edit with PyCharm Community Edition with Anaconda plugin
Zip-7
CRC SHA
Convert to Adobe PDF
...Combine files in Acrobat
++Edit with Notepad
פתח באמצעות
שתף עם
...Add to archive
"Add to "Welcome_To_Colaboratory.rar
...Compress and email
Compress to "Welcome_To_Colaboratory.rar" and email
שחזר גירסאות קודמות
שלח אל
גזור
העתק
צור קיצור דרך
מחק
שנה שם
מאפיינים
Welcome_To_Colaboratory.html

Copy the line:

>script  src="https://cdn.plot.ly/plotly-latest.min.js"></script<

At the head of the file, before all other scripts



```
<!DOCTYPE html>
<html>
<head><meta charset="utf-8" />

<title>Welcome_To_Colaboratory</title>
<script src="https://cdn.plot.ly/plotly-latest.min.js"></script>

<script src=      ps://cdnjs.cloudflare.com/ajax/libs/require.js/2.1.10/require.min.js"></script>
<script sr        ://cdnjs.cloudflare.com/ajax/libs/jquery/2.0.3/jquery.min.js"></script>



<style type="text/css">
    /*!
```

And save the file