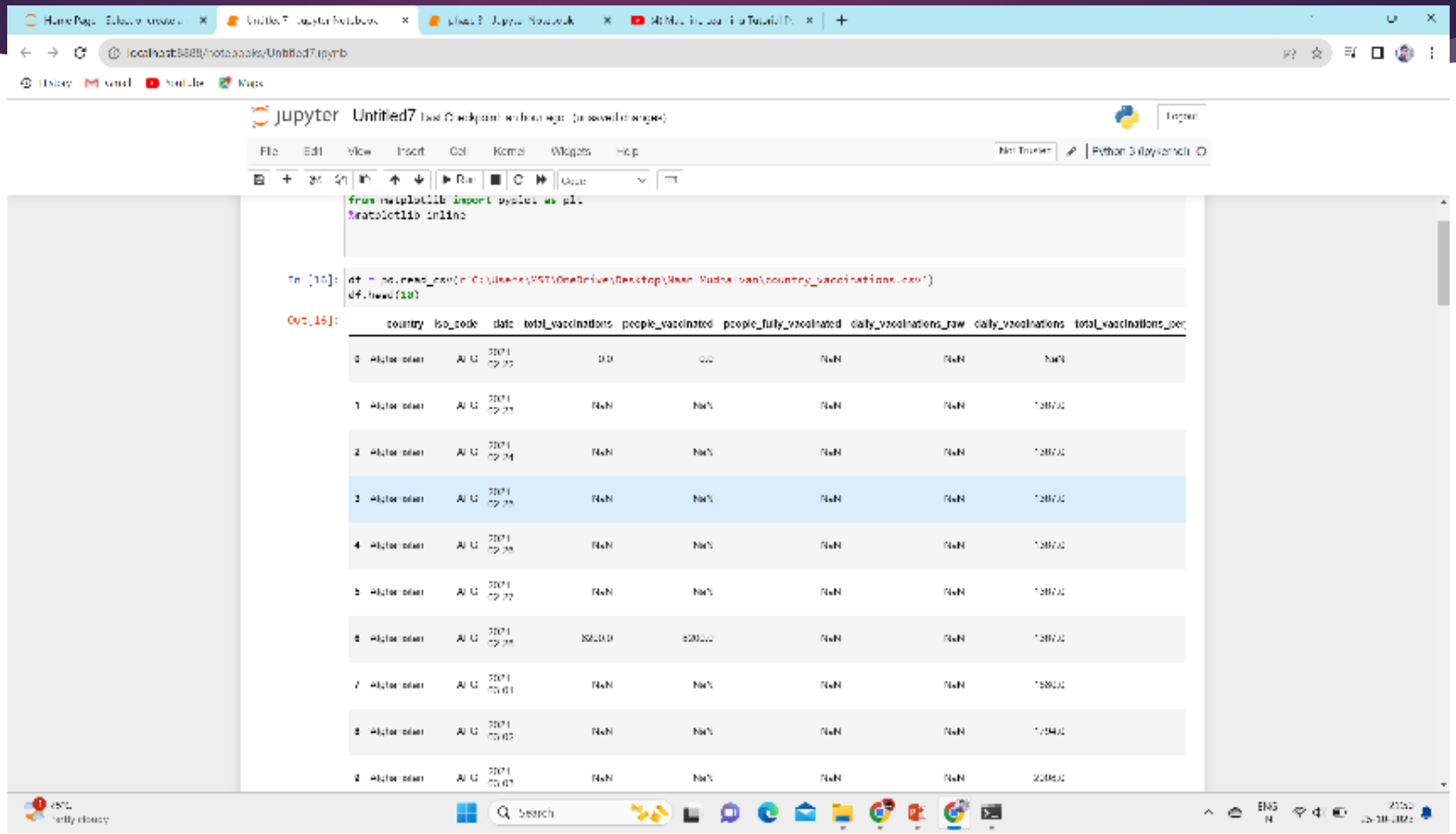
COVID VACCINE Analysis

PHASE -4

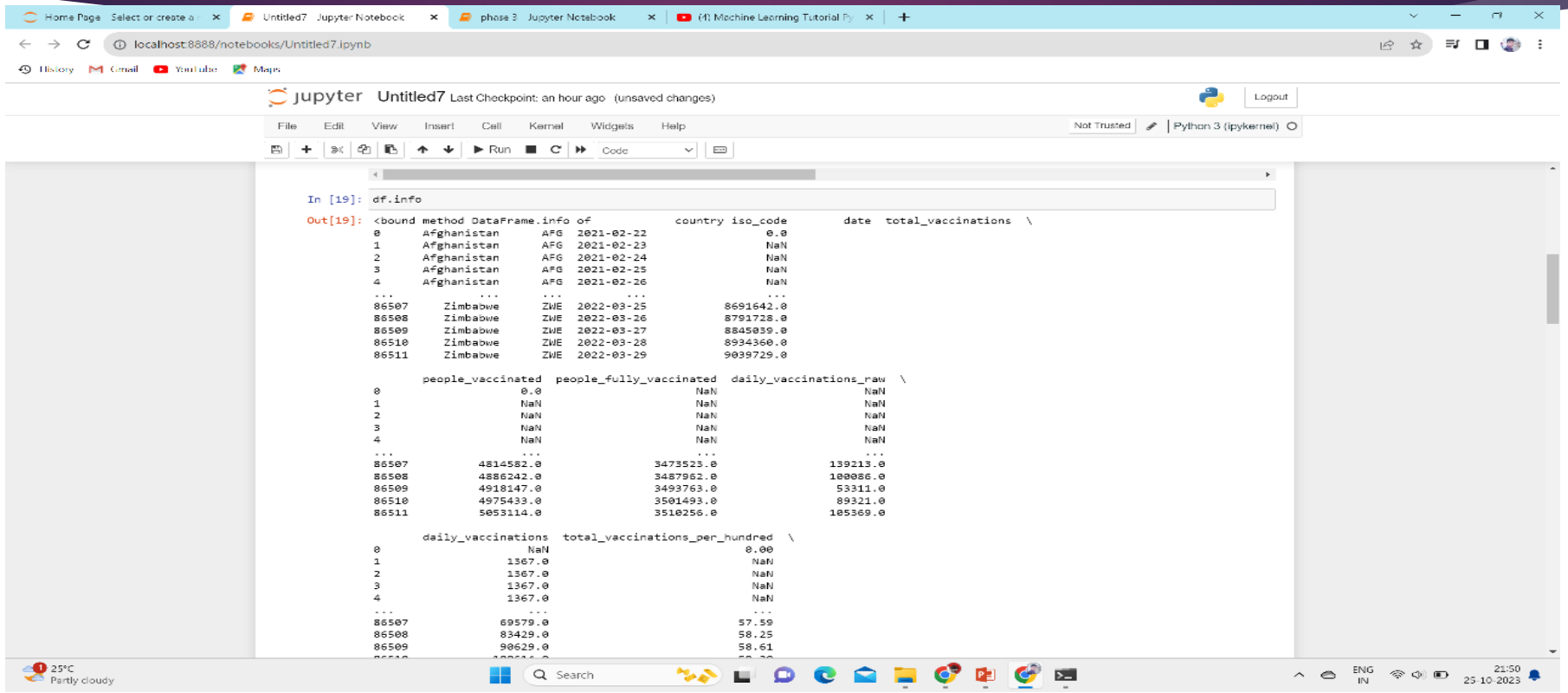
1: Performing exploratory data analysis

- **What Is Exploratory Data Analysis?**
- Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows you to get a better feel of your data and find useful patterns in it.
- You need to know the patterns in your data and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. You also need to recognize errors in your data.

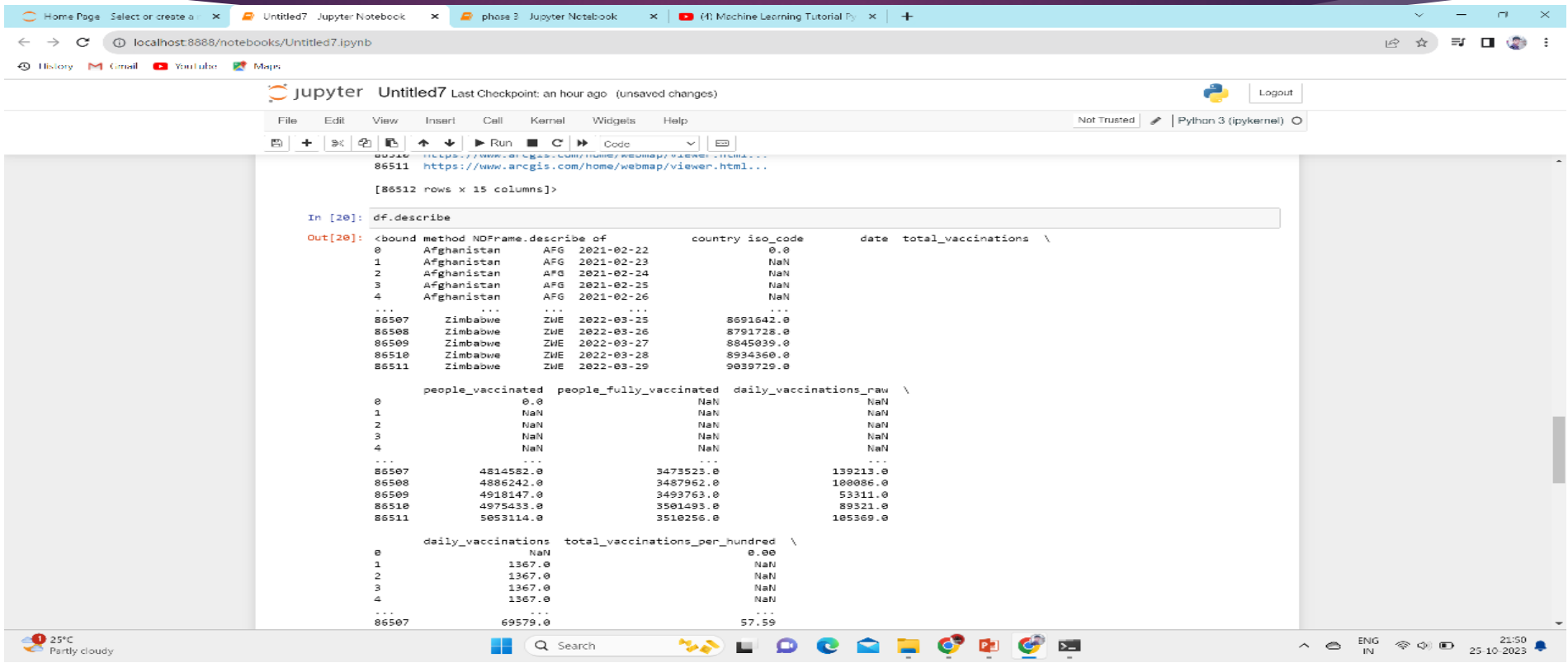
Importing a Dataset



Use info


Home Page Select or create a
Untitled7 Jupyter Notebook
phase 3 Jupyter Notebook
(4) Machine Learning Tutorial Py
localhost:8888/notebooks/Untitled7.ipynb
History
Gmail
YouTube
Maps
jupyter Untitled7 Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [19]: df.info
Out[19]: <bound method DataFrame.info of country iso_code date total_vaccinations \
0 Afghanistan AFG 2021-02-22 0.0
1 Afghanistan AFG 2021-02-23 NaN
2 Afghanistan AFG 2021-02-24 NaN
3 Afghanistan AFG 2021-02-25 NaN
4 Afghanistan AFG 2021-02-26 NaN
... ...
86507 Zimbabwe ZWE 2022-03-25 8691642.0
86508 Zimbabwe ZWE 2022-03-26 8791728.0
86509 Zimbabwe ZWE 2022-03-27 8845039.0
86510 Zimbabwe ZWE 2022-03-28 8934360.0
86511 Zimbabwe ZWE 2022-03-29 9039729.0
people_vaccinated people_fully_vaccinated daily_vaccinations_raw \
0 0.0 NaN NaN
1 NaN NaN NaN
2 NaN NaN NaN
3 NaN NaN NaN
4 NaN NaN NaN
... ...
86507 4814582.0 3473523.0 139213.0
86508 4886242.0 3487962.0 100886.0
86509 4918147.0 3493763.0 53311.0
86510 4975433.0 3501493.0 89321.0
86511 5053114.0 3510256.0 105569.0
daily_vaccinations total_vaccinations_per_hundred \
0 NaN 0.00
1 1367.0 NaN
2 1367.0 NaN
3 1367.0 NaN
4 1367.0 NaN
... ...
86507 69579.0 57.59
86508 83429.0 58.25
86509 90629.0 58.61
25°C
Partly cloudy
Search
ENG
IN
21:50
25-10-2023

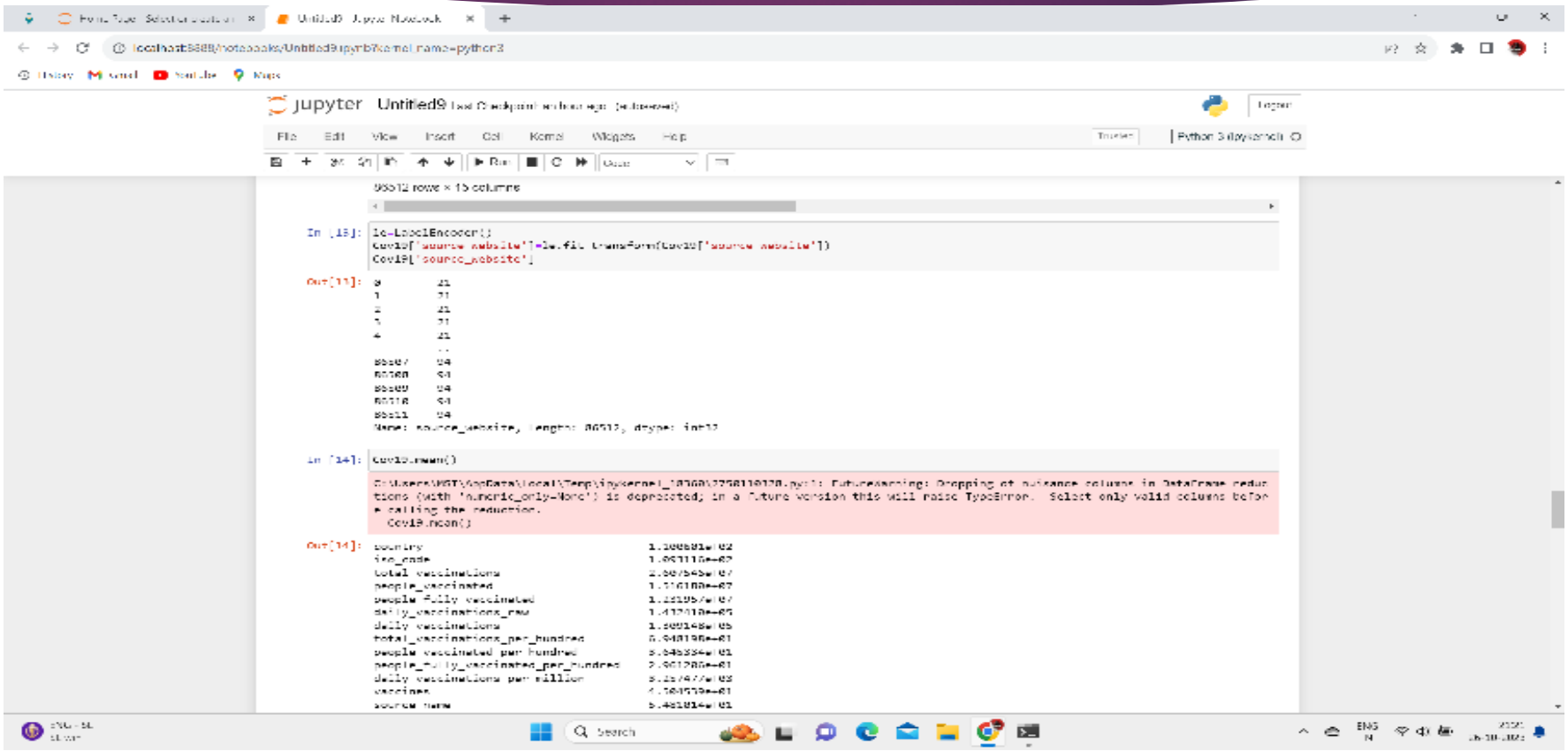
Use describe



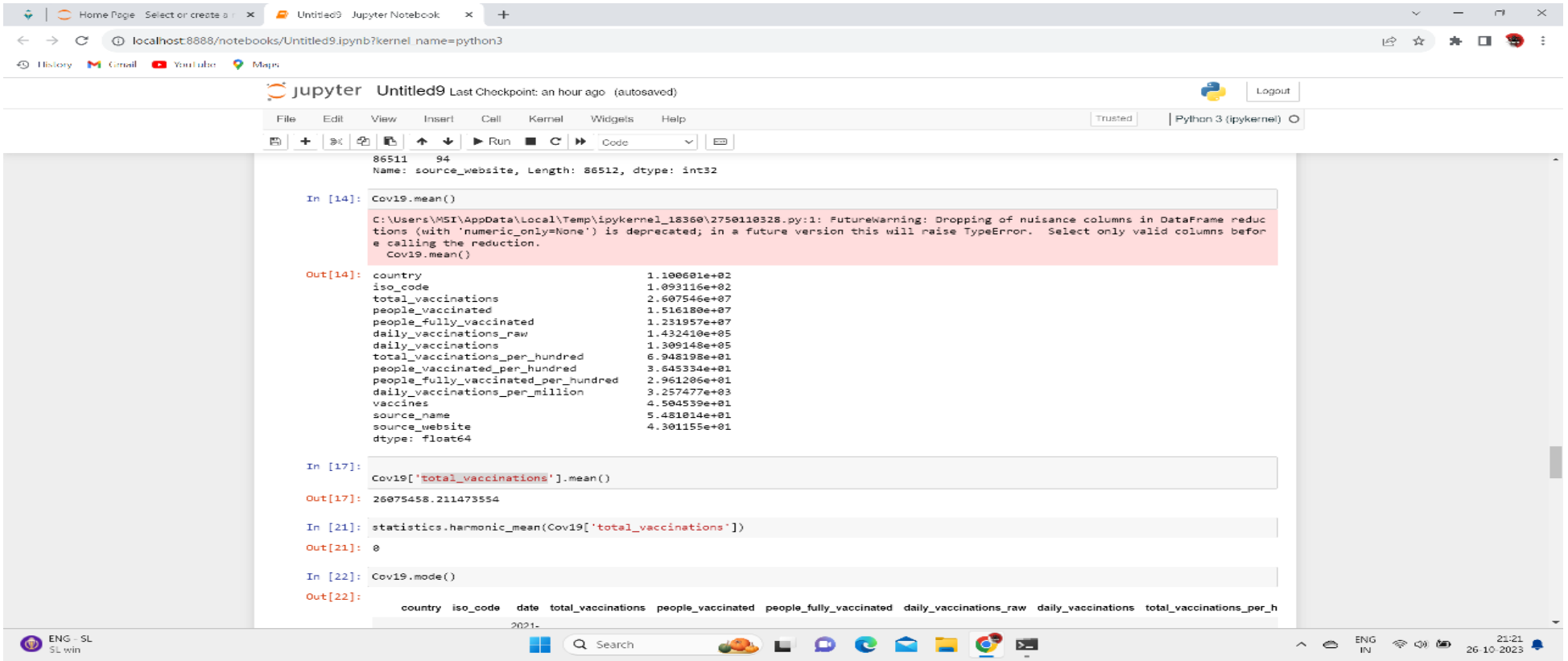
2: Statistical analysis

- Statistical analysis is the collection and interpretation of data in order to uncover patterns and trends. It is a component of data analytics. Statistical analysis can be used in situations like gathering research interpretations, statistical modeling or designing surveys and studies. It can also be useful for business intelligence organizations that have to work with large data volumes.
- Describe the nature of the data to be analyzed.
- Explore the relation of the data to the underlying population.
- Create a model to summarize an understanding of how the data relates to the underlying population.
- Prove (or disprove) the validity of the model.
- Employ predictive analytics to run scenarios that will help guide future actions.

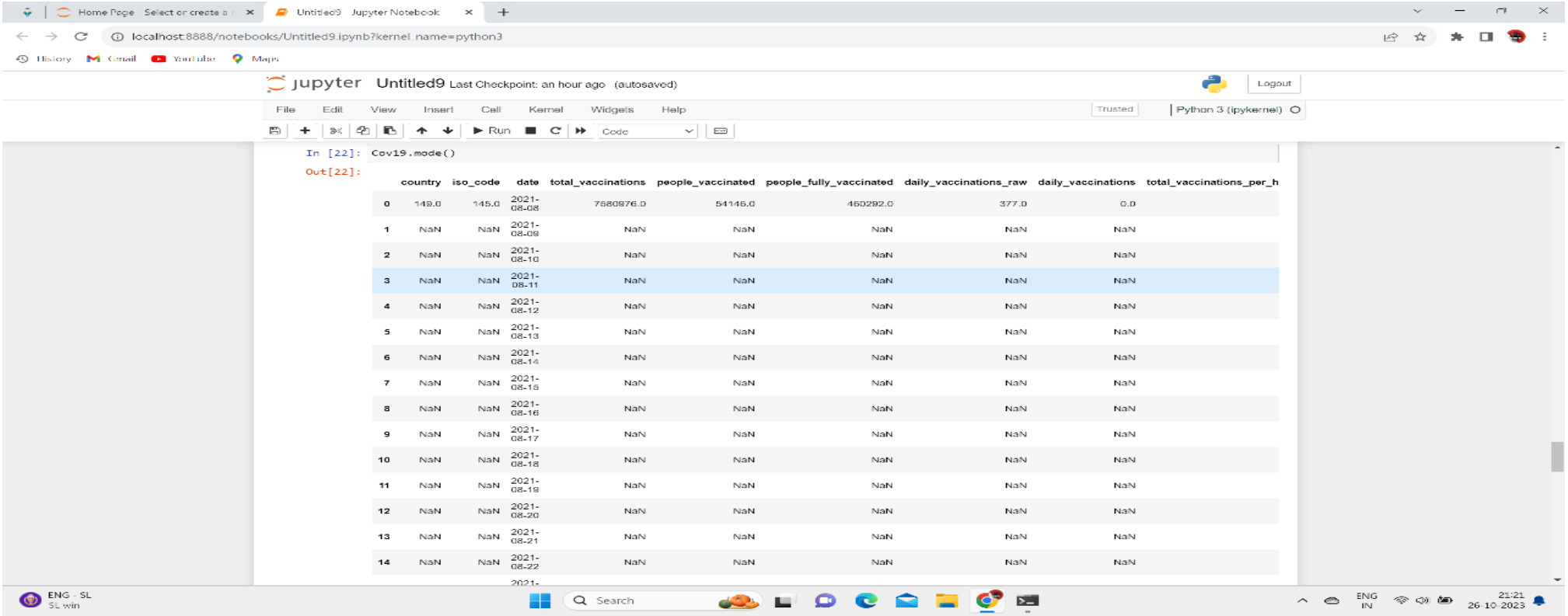
- The five basic methods are mean, standard deviation, regression, hypothesis testing, and sample size determination. It is widely used by governments, businesses, banking entities, insurance companies, etc



MEAN()

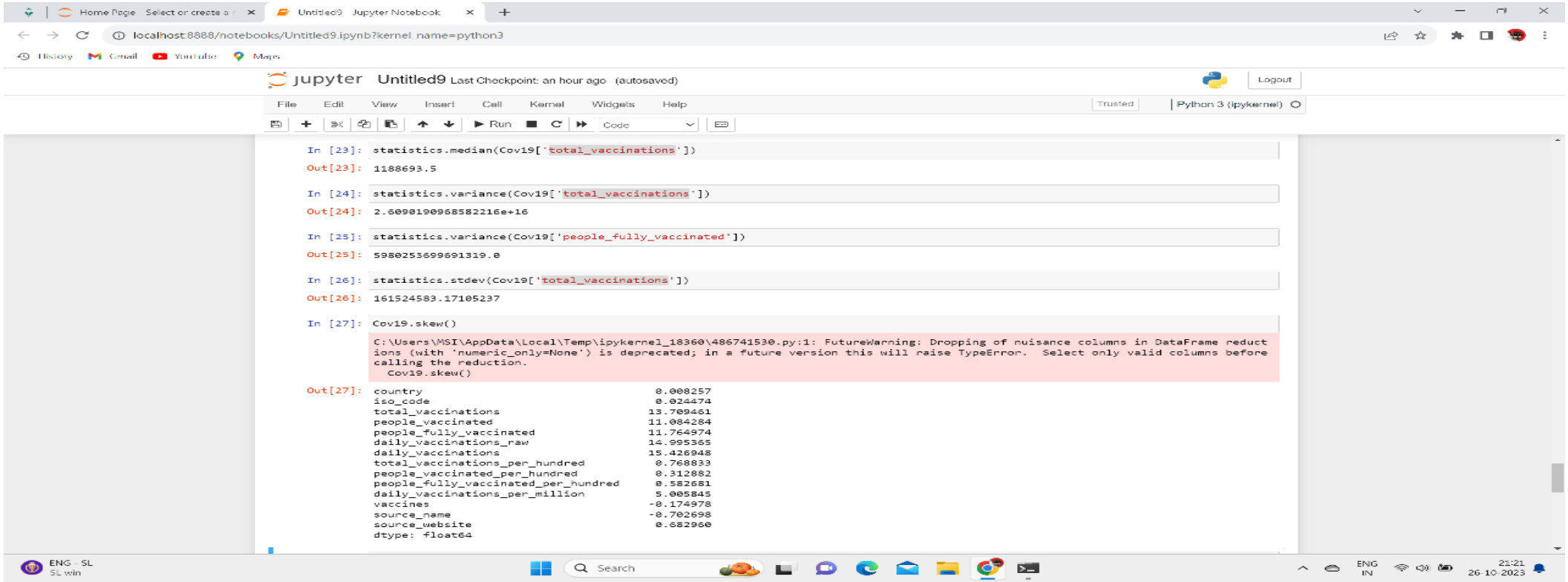

Home Page
Untitled9 Jupyter Notebook
localhost:8888/notebooks/Untitled9.ipynb?kernel_name=python3
jupyter Untitled9 Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
86511 94
Name: source_website, Length: 86512, dtype: int32
In [14]:
Cov19.mean()
C:\Users\MSI\AppData\Local\Temp\ipykernel_18360\2750110328.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
Cov19.mean()
Out[14]:
country 1.100601e+02
iso_code 1.093116e+02
total_vaccinations 2.607546e+07
people_vaccinated 1.516180e+07
people_fully_vaccinated 1.231957e+07
daily_vaccinations_raw 1.432410e+05
daily_vaccinations 1.309148e+05
total_vaccinations_per_hundred 6.948198e+01
people_vaccinated_per_hundred 3.645334e+01
people_fully_vaccinated_per_hundred 2.961206e+01
daily_vaccinations_per_million 3.257477e+03
vaccines 4.504539e+01
source_name 5.481014e+01
source_website 4.301155e+01
dtype: float64
In [17]:
Cov19['total_vaccinations'].mean()
Out[17]:
26075458.211473554
In [21]:
statistics.harmonic_mean(Cov19['total_vaccinations'])
Out[21]:
0
In [22]:
Cov19.mode()
Out[22]:
country iso_code date total_vaccinations people_vaccinated people_fully_vaccinated daily_vaccinations_raw daily_vaccinations total_vaccinations_per_h
2021-
Search
ENG IN
21:21
26-10-2023

MODE()



	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_h
0	149.0	145.0	2021-08-08	7580976.0	54145.0	460292.0	377.0	0.0	
1	NaN	NaN	2021-08-09	NaN	NaN	NaN	NaN	NaN	
2	NaN	NaN	2021-08-10	NaN	NaN	NaN	NaN	NaN	
3	NaN	NaN	2021-08-11	NaN	NaN	NaN	NaN	NaN	
4	NaN	NaN	2021-08-12	NaN	NaN	NaN	NaN	NaN	
5	NaN	NaN	2021-08-13	NaN	NaN	NaN	NaN	NaN	
6	NaN	NaN	2021-08-14	NaN	NaN	NaN	NaN	NaN	
7	NaN	NaN	2021-08-15	NaN	NaN	NaN	NaN	NaN	
8	NaN	NaN	2021-08-16	NaN	NaN	NaN	NaN	NaN	
9	NaN	NaN	2021-08-17	NaN	NaN	NaN	NaN	NaN	
10	NaN	NaN	2021-08-18	NaN	NaN	NaN	NaN	NaN	
11	NaN	NaN	2021-08-19	NaN	NaN	NaN	NaN	NaN	
12	NaN	NaN	2021-08-20	NaN	NaN	NaN	NaN	NaN	
13	NaN	NaN	2021-08-21	NaN	NaN	NaN	NaN	NaN	
14	NaN	NaN	2021-08-22	NaN	NaN	NaN	NaN	NaN	

Median , Variance



Home Page Select or create a | Untitled9 Jupyter Notebook

localhost:8888/notebooks/Untitled9.ipynb?kernel_name=python3

History Gmail YouTube Maps

jupyter Untitled9 Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [26]:
```
statistics.stdev(Cov19['total_vaccinations'])
```

Out[26]:
```
161524583.17105237
```

In [27]:
```
Cov19.skew()
```

```
C:\Users\MSI\AppData\Local\Temp\ipykernel_18360\486741530.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError.  Select only valid columns before calling the reduction.
  Cov19.skew()
```

Out[27]:
```
country                               0.008257
iso_code                              0.024474
total_vaccinations                   13.709461
people_vaccinated                    11.084284
people_fully_vaccinated              11.764974
daily_vaccinations_raw               14.995365
daily_vaccinations                   15.426948
total_vaccinations_per_hundred        0.768833
people_vaccinated_per_hundred         0.312882
people_fully_vaccinated_per_hundred   0.582681
daily_vaccinations_per_million        5.005845
vaccines                             -0.174978
source_name                          -0.702698
source_website                        0.682960
dtype: float64
```

In [29]:
```
Cov19['total_vaccinations'].plot(kind = 'density')
```

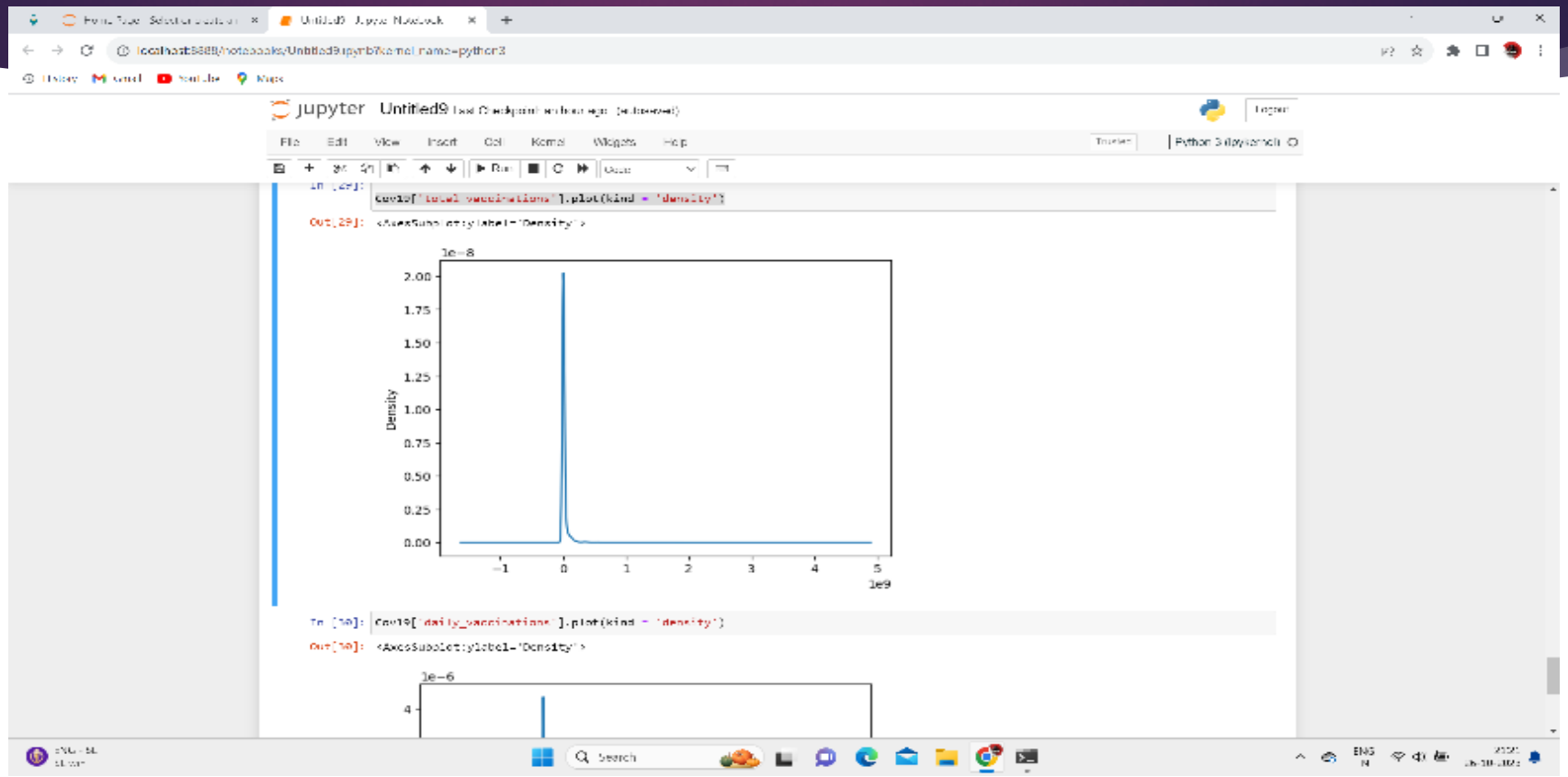
Out[29]:
```
<AxesSubplot:ylabel='Density'>
```

ENG - SL
SL win

Search

ENG
IN

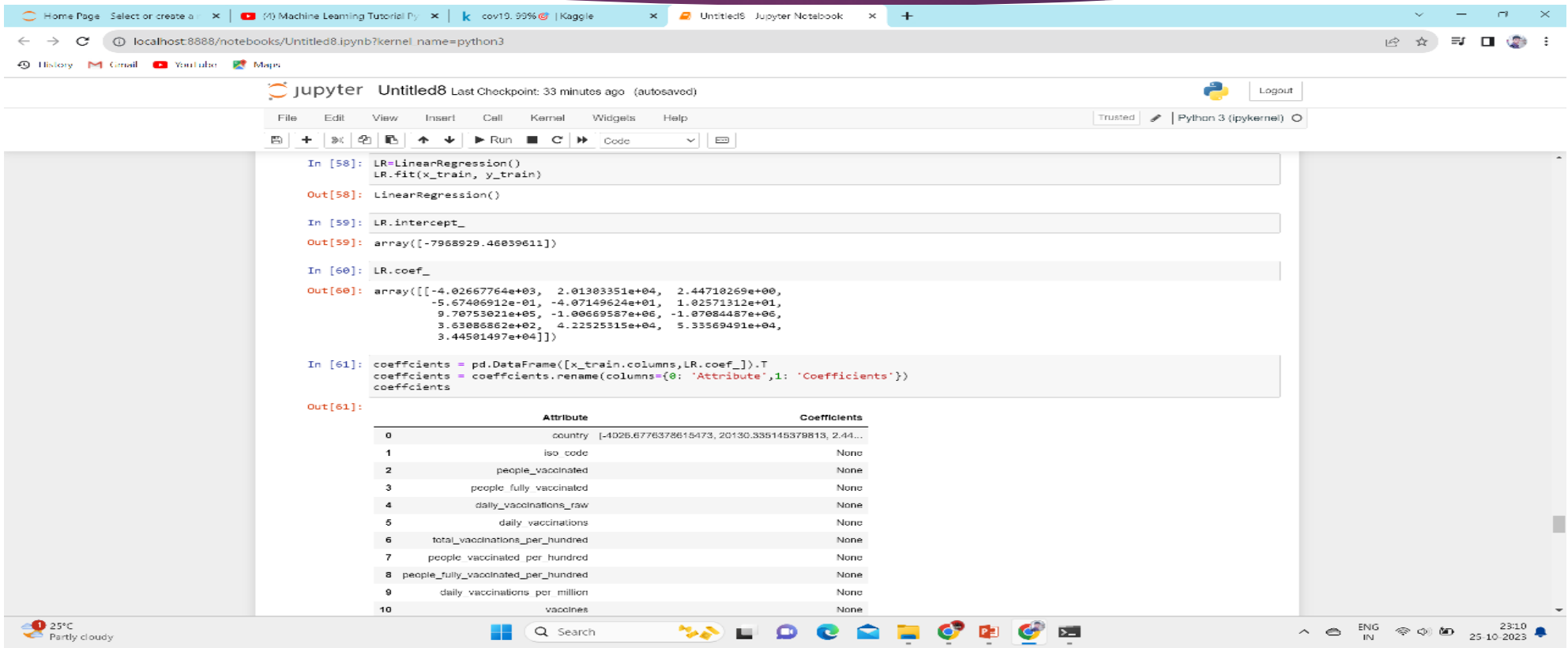
21:21
26-10-2023

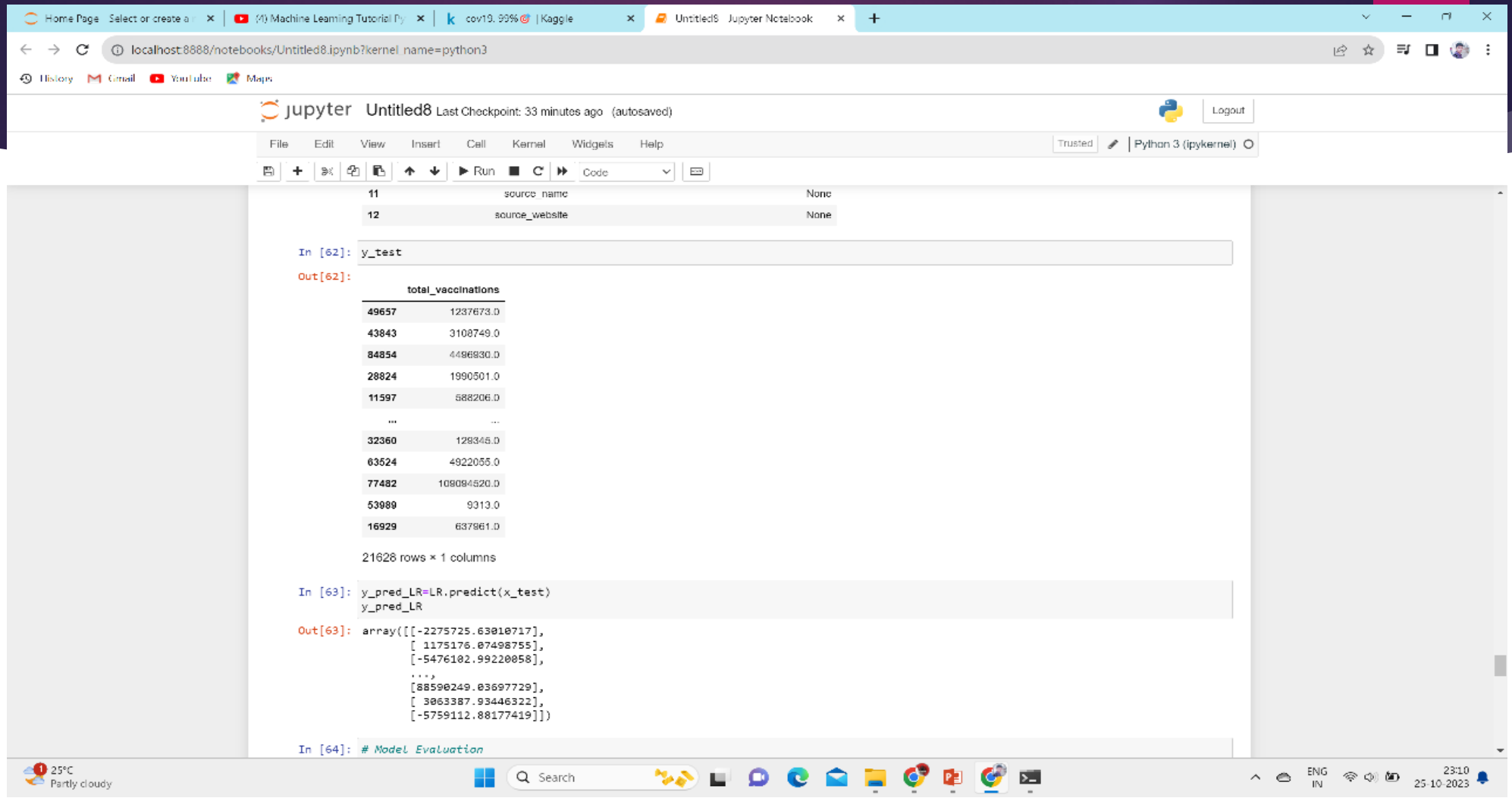

jupyter Untitled9
File Edit View Insert Cell Kernel Widgets Help
Python 3 (ipykernel)
In [29]: covid['total_vaccinations'].plot(kind = 'density')
Out[29]: <AxesSubplot:ylabel="Density">
1e-8
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Density
−1
0
1
2
3
4
5
1e9
In [30]: covid['daily_vaccinations'].plot(kind = 'density')
Out[30]: <AxesSubplot:ylabel="Density">
1e−6
4
Search

3: Visualization

- Visualization (graphics), the physical or imagining creation of images, diagrams, or animations to communicate a message. Data and information visualization, the practice of creating visual representations of complex data and information.
- Like:
  - Linear Regression
  - Kmeans
  - Decision Tree
  - KNN,etc..

Using Linear Regression




Home Page - Select or create a
(4) Machine Learning Tutorial Py
cov19: 99%
| Kaggle
Untitled8 - Jupyter Notebook
localhost:8888/notebooks/Untitled8.ipynb?kernel_name=python3
History
Gmail
YouTube
Maps
jupyter Untitled8 Last Checkpoint: 33 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
11 source_name None
12 source_website None
In [62]: y_test
Out[62]:
total_vaccinations
49657 1237673.0
43843 3108749.0
84854 4496930.0
28824 1990501.0
11597 588206.0
... ...
32360 128346.0
63524 4922055.0
77482 109094520.0
53989 9313.0
16929 637961.0
21628 rows × 1 columns
In [63]: y_pred_LR=LR.predict(x_test)
y_pred_LR
Out[63]: array([[-2275725.63010717],
[1175176.07498755],
[-5476102.99220058],
...,
[88590249.03697729],
[3063387.93446322],
[-5759112.88177419]])
In [64]: # Model Evaluation
25°C
Partly cloudy
Search
ENG
IN
23:10
25-10-2023

Home Page Select or create a | (4) Machine Learning Tutorial Py | cov19: 99%🎯 | Kaggle | Untitled8 Jupyter Notebook

localhost:8888/notebooks/Untitled8.ipynb?kernel_name=python3

History Gmail YouTube Maps

jupyter Untitled8 Last Checkpoint: 33 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

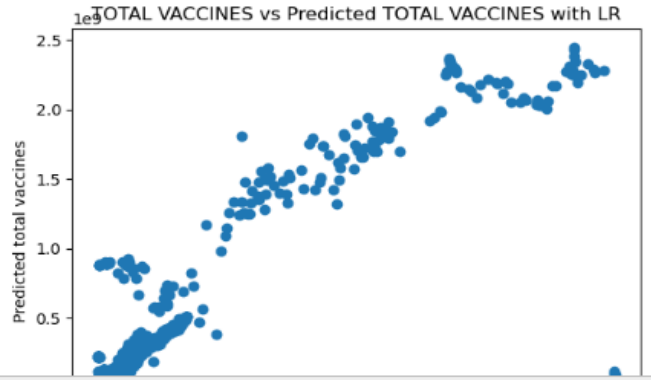
Run Code

```
[88590249.03697729],
[ 3063387.93446322],
[-5759112.88177419]])
```

In [64]:
```
print('R^2:',metrics.r2_score(y_test, y_pred_LR))
print('MAE:',metrics.mean_absolute_error(y_test, y_pred_LR))
print('MSE:',metrics.mean_squared_error(y_test, y_pred_LR))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test, y_pred_LR)))
```

```
R^2: 0.8719183276933511
MAE: 11884018.238719486
MSE: 3581668998225302.5
RMSE: 59847046.69593398
```

In [65]:
```
plt.scatter(y_test, y_pred_LR)
plt.xlabel("total vaccines")
plt.ylabel("Predicted total vaccines")
plt.title("TOTAL VACCINES vs Predicted TOTAL VACCINES with LR")
plt.show()
```

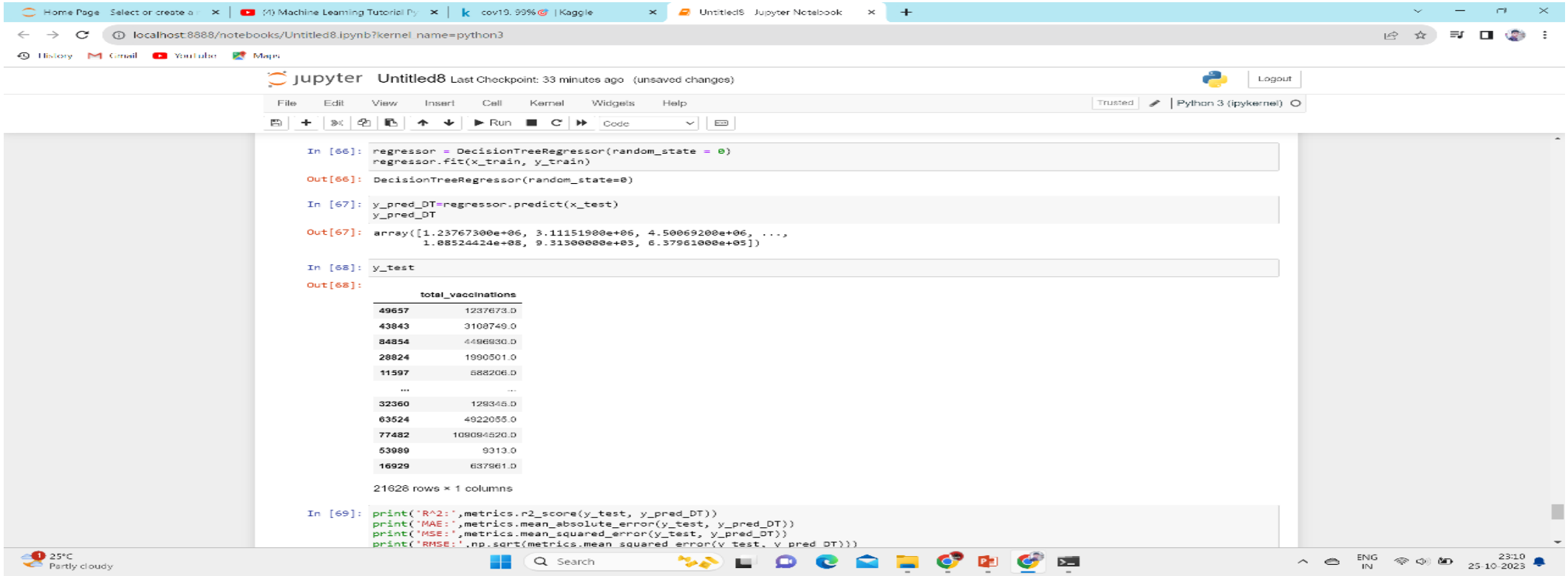


25°C Partly cloudy

Search

ENG IN 23:10 25-10-2023

2.2 Decision Tree Algorithm



Home Page Select or create a | (1) Machine Learning Tutorial Py | cov19: 99% | Kaggle | Untitled8 Jupyter Notebook

localhost:8888/notebooks/Untitled8.ipynb?kernel_name=python3

History Gmail YouTube Maps

jupyter Untitled8 Last Checkpoint: 33 minutes ago (unsaved changes) Logout

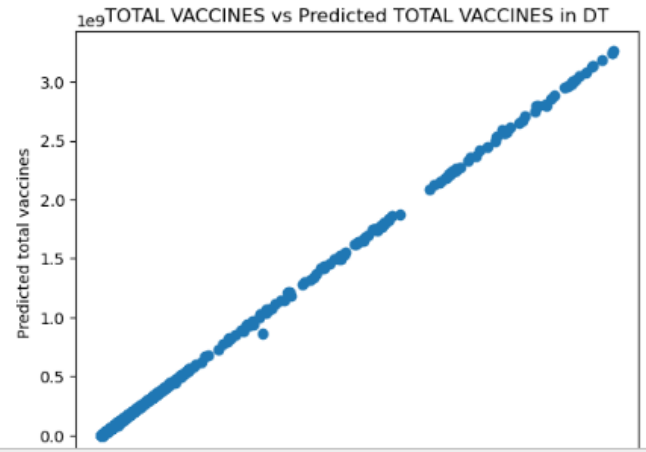
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [69]:
```
print('R^2:',metrics.r2_score(y_test, y_pred_DT))
print('MAE:',metrics.mean_absolute_error(y_test, y_pred_DT))
print('MSE:',metrics.mean_squared_error(y_test, y_pred_DT))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test, y_pred_DT)))
```

```
R^2: 0.9999132840454302
MAE: 156448.73247641945
MSE: 2424920291410.6694
RMSE: 1557215.557143798
```

In [70]:
```
plt.scatter(y_test, y_pred_DT)
plt.xlabel("total vaccines")
plt.ylabel("Predicted total vaccines")
plt.title("TOTAL VACCINES vs Predicted TOTAL VACCINES in DT")
plt.show()
```



25°C Partly cloudy Search ENG IN 23:10 25-10-2023


Thank you