```
!apt-get update
!apt-get install openjdk-8-jdk-headless -qq >/dev/null
!wget -q http://archive.apache.org/dist/spark/spark-2.3.1/spark-2.3.1-bin-hadoop2.7.tgz
!tar xf spark-2.3.1-bin-hadoop2.7.tgz
!pip install -q findspark

import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.3.1-bin-hadoop2.7"
!ls
import findspark
findspark.init()
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
spark
```

```
Get:1 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Hit:2 http://archive.ubuntu.com/ubuntu bionic InRelease
Get:3 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:4 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease [3,626 B]
Ign:5 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  InRelease
Get:6 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease [15.9 kB]
Ign:7 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  InRelease
Get:8 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  Release [696 B]
Get:9 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]
Hit:10 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  Release
Get:11 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  Release.gpg [836 B]
Hit:12 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Get:13 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease [15.9 kB]
Hit:14 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Get:15 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [2,695 kB]
Get:16 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [1,490 kB]
Get:17 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [3,134 kB]
Get:18 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [2,268 kB]
Get:20 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  Packages [953 kB]
Get:21 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic/main Sources [1,947 kB]
Get:22 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic/main amd64 Packages [996 kB]
Get:23 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic/main amd64 Packages [45.3 kB]
Fetched 13.8 MB in 8s (1,674 kB/s)
Reading package lists... Done
sample_data  spark-2.3.1-bin-hadoop2.7  spark-2.3.1-bin-hadoop2.7.tgz
```

SparkSession - in-memory

SparkContext

Spark UI

Version
    v2.3.1
Master
    local[*]
AppName
    pyspark-shell

```
from google.colab import files
 #uploading the files via google collab which is another
 #way to upload files directly from computer
 # Upload the file 1000 Sales_Records.csv in google colab

mydata = files.upload()
```

Choose Files covid19.csv

- **covid19.csv**(text/csv) - 1366592 bytes, last modified: 4/14/2022 - 100% done

```
Saving covid19.csv to covid19 (1).csv
```

```
# Upload the file in google colab
mydata=spark.read.format("csv").option("header","true").load("covid19 (1).csv")
# Upload the datafile in google colab ipython notebook and show the data file
mydata.show()
```

```
+-------------+-------------+-----------+----------+---------+----------------+----------+-----------------+
|Date_reported| Country_code|    Country|WHO_region|New_cases|Cumulative_cases|New_deaths|Cumulative_deaths|
+-------------+-------------+-----------+----------+---------+----------------+----------+-----------------+
|    2/24/2020|           AF|Afghanistan|      EMRO|        5|               5|         0|                0|
|    2/25/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|    2/26/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|    2/27/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|    2/28/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|    2/29/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/1/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/2/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/3/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/4/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/5/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/6/2020|           AF|Afghanistan|      EMRO|        0|               5|         0|                0|
|     3/7/2020|           AF|Afghanistan|      EMRO|        3|               8|         0|                0|
|     3/8/2020|           AF|Afghanistan|      EMRO|        0|               8|         0|                0|
|     3/9/2020|           AF|Afghanistan|      EMRO|        0|               8|         0|                0|
|    3/10/2020|           AF|Afghanistan|      EMRO|        0|               8|         0|                0|
|    3/11/2020|           AF|Afghanistan|      EMRO|        3|              11|         0|                0|
|    3/12/2020|           AF|Afghanistan|      EMRO|        0|              11|         0|                0|
|    3/13/2020|           AF|Afghanistan|      EMRO|        0|              11|         0|                0|
|    3/14/2020|           AF|Afghanistan|      EMRO|        3|              14|         0|                0|
+-------------+-------------+-----------+----------+---------+----------------+----------+-----------------+
only showing top 20 rows
```

```
#a) Import pandas and show 50 rows
import pandas as pd
mydata3 =pd.read_csv(r'covid19 (1).csv')
mydata3.head(50)
```

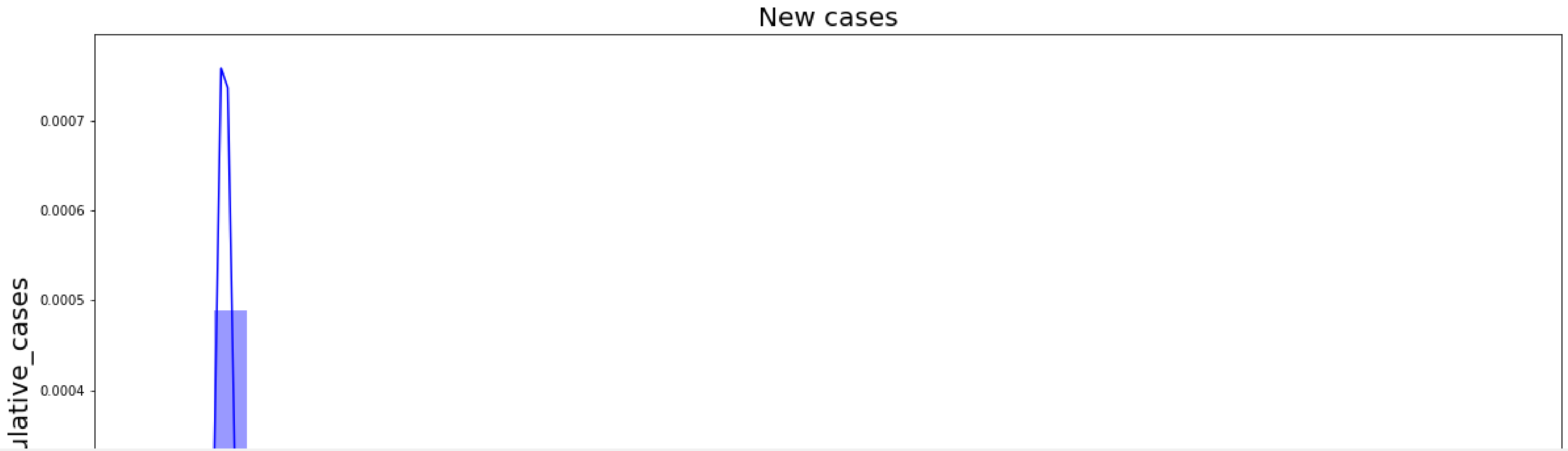
	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
0	2/24/2020	AF	Afghanistan	EMRO	5	5	0	0
1	2/25/2020	AF	Afghanistan	EMRO	0	5	0	0
2	2/26/2020	AF	Afghanistan	EMRO	0	5	0	0
3	2/27/2020	AF	Afghanistan	EMRO	0	5	0	0
4	2/28/2020	AF	Afghanistan	EMRO	0	5	0	0
5	2/29/2020	AF	Afghanistan	EMRO	0	5	0	0
6	3/1/2020	AF	Afghanistan	EMRO	0	5	0	0
7	3/2/2020	AF	Afghanistan	EMRO	0	5	0	0
8	3/3/2020	AF	Afghanistan	EMRO	0	5	0	0
9	3/4/2020	AF	Afghanistan	EMRO	0	5	0	0
10	3/5/2020	AF	Afghanistan	EMRO	0	5	0	0
11	3/6/2020	AF	Afghanistan	EMRO	0	5	0	0
12	3/7/2020	AF	Afghanistan	EMRO	3	8	0	0
13	3/8/2020	AF	Afghanistan	EMRO	0	8	0	0
14	3/9/2020	AF	Afghanistan	EMRO	0	8	0	0
15	3/10/2020	AF	Afghanistan	EMRO	0	8	0	0
16	3/11/2020	AF	Afghanistan	EMRO	3	11	0	0
17	3/12/2020	AF	Afghanistan	EMRO	0	11	0	0
18	3/13/2020	AF	Afghanistan	EMRO	0	11	0	0
19	3/14/2020	AF	Afghanistan	EMRO	3	14	0	0
20	3/15/2020	AF	Afghanistan	EMRO	6	20	0	0
21	3/16/2020	AF	Afghanistan	EMRO	5	25	0	0
22	3/17/2020	AF	Afghanistan	EMRO	1	26	0	0
23	3/18/2020	AF	Afghanistan	EMRO	0	26	0	0
24	3/19/2020	AF	Afghanistan	EMRO	0	26	0	0
25	3/20/2020	AF	Afghanistan	EMRO	-2	24	0	0
26	3/21/2020	AF	Afghanistan	EMRO	0	24	0	0
27	3/22/2020	AF	Afghanistan	EMRO	10	34	0	0
28	3/23/2020	AF	Afghanistan	EMRO	6	40	1	1
29	3/24/2020	AF	Afghanistan	EMRO	2	42	0	1
30	3/25/2020	AF	Afghanistan	EMRO	32	74	0	1
31	3/26/2020	AF	Afghanistan	EMRO	6	80	1	2
32	3/27/2020	AF	Afghanistan	EMRO	11	91	0	2
33	3/28/2020	AF	Afghanistan	EMRO	15	106	1	3
34	3/29/2020	AF	Afghanistan	EMRO	8	114	1	4
35	3/30/2020	AF	Afghanistan	EMRO	0	114	0	4

```
#b)By using matplotlib, numpy and
#seaborn libraries you will plot histogram of new cases, your chart
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

f,(axl)=plt.subplots(1,1,figsize=(20,10),sharey=True)
#changed the csv file title to remove the space in the front of the columns
sns.distplot(mydata3['New_cases'], kde=True,color='blue',hist=True,bins=40)

#histogram
plt.title('New cases',fontsize = 20)
plt.xlabel('New_cases',fontsize = 20)
plt.ylabel('Cumulative_cases',fontsize =20)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar
  warnings.warn(msg, FutureWarning)
Text(0, 0.5, 'Cumulative_cases')
```



```
#c) You should use scientific python and sckit
#learn packages for importing linear regression and
#statistics models-with these packages you need to create
#a prediction model for the new cases and the new deaths

from scipy import stats
from sklearn.linear_model import LinearRegression
linear_regression = stats.linregress(x=mydata3.New_cases,y=mydata3.New_deaths)
slope=linear_regression.slope
print(format(slope,'.6f'))
intercept = linear_regression.intercept
print(format(intercept,'.4f'))
```

```
0.026971
6.3321
```

```
printing=linear_regression.slope *(1000)+linear_regression.intercept
print(format(printing,'.4f'))

printing2=linear_regression.slope *(-1000)+linear_regression.intercept
print(format(printing,'.4f'))
```

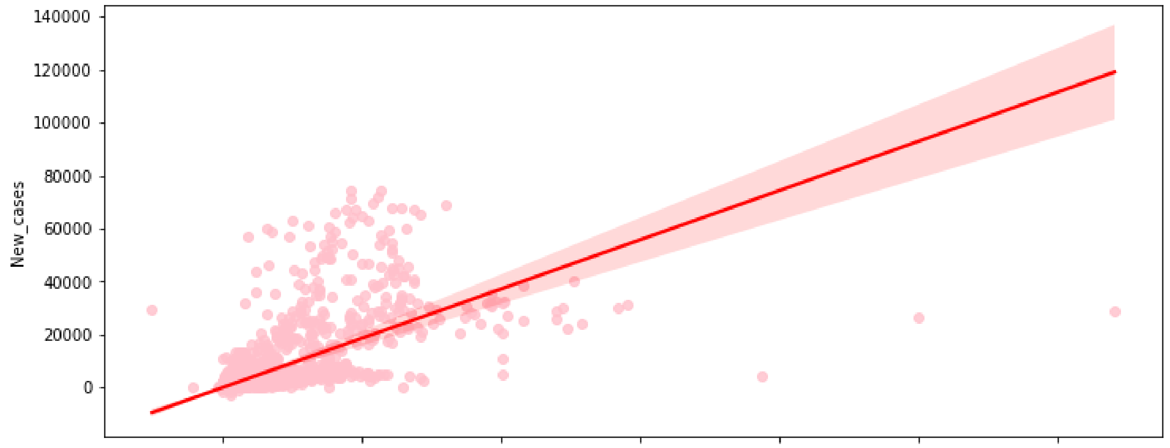
```
33.3028
33.3028
```

```
#d) Create a linear regression line with the above data,
#new cases versus new deaths.

f,(axl)=plt.subplots(1,1,figsize=(12,5),sharey=True)
```

```
f,(ax1)=plt.subplots(1,1,figsize=(12,5),sharey=True)
#newcases =y cause its over new deaths
sns.regplot(x="New_deaths",y="New_cases",data=mydata3,scatter_kws={"color":"pink"},
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff68f721490>
```



✓ 5s completed at 11:12 AM