Assignment1

Import the required libraries

Installing the Spark Dependancies

Creating a Spark Session

```
import findspark

findspark.init()

from pyspark import SparkContext, SparkConf

from pyspark.sql import SQLContext, SparkSession
```

Configuring a Spark Session

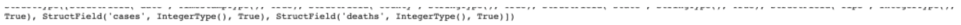
```
conf = SparkConf().set('spark.ui.port', '4050')

sc = SparkContext(conf=conf)

spark = SparkSession.builder.master('local[*]').getOrCreate()
```

Load dateset into a dataframe :

By using the spark.read(), we can load the data

Schema

We can check the structure of the data by using data.schema()



```
True), StructField('cases', IntegerType(), True), StructField('deaths', IntegerType(), True)])
```

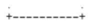
Task 1: Find the total number of new cases added in the entire US in the month of March 2020.

Basically, we can use Pyspark dataframe and can query the results, but in real time scenarios, the complex queries is difficult to write by using dataframe syntax.

Hence, We connected to spark SQL and then by using SQL queries we generate the results.

Create temp table where in Pyspark we can use spark SQL

Created CTE in which we will be returning date and sum of cases by grouping date column and then with condition having March 2020



```
+----------+
```

Task 2: Calculate the total new cases added in three consecutive months of June, July, and August of 2020 in Jackson county, Missouri (fips code 29095).

By using Common Table Expressions, we generated the results.

First CTE: extract3_months - where it returns records with condition having three consecutive months of June, July, and August of 2020

Second CTE: sum3_months - returns data with 3 months having total new cases added

Finally, by using case when statement, will be extracting month from the date and returning records.

```
|August 2020 cases|  [illegible]|
+-----------------+--------+
```

Task 3: Find the daily new cases per month per 1000 population in Missouri state (MO) since the beginning of the pandemic (assume MO's population is 6,154,913). [Plot the data]

Rounding sum of new cases per month per 1000 population and by using where condition we will be filtering state = 'Missouri' and Year(date) >= 2020 grouping the year and date

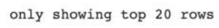
```
|   7159.25|     3|2022|
|   6973.02|     4|2022|
```

For plotting, Convert Spark dataframe to a Panda dataframe to plot

Find news cases from cumulative cases and then finally plot as follows:

Task 4: On which date all 50 US states have at least 100 cases? At least one death?

CTE- Firstly filtered records with cases>=100 AND deaths>=1and then by using CTE, filtering data by applying group by condition having all the states for particular date

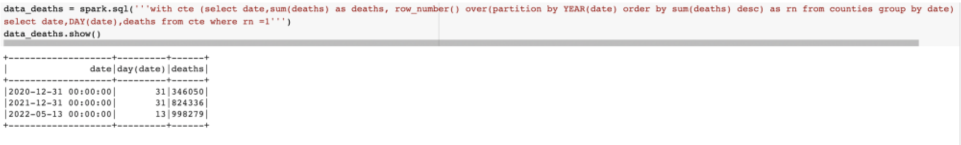


```
only showing top 20 rows
```

Task 5: Which single day in the year 2020 and 2021 had the largest number of deaths in the entire US (if there are multiple such dates, choose the earliest one)?

CTE - retuns data with all the dates having sum(deaths) and row_number with partition by YEAR(date)

Finally filtering 3 records in each year with rownumber=1



```
data_deaths = spark.sql('''with cte (select date,sum(deaths) as deaths, row_number() over(partition by YEAR(date) order by sum(deaths) desc) as rn from counties group by date)
select date,DAY(date),deaths from cte where rn =1''')
data_deaths.show()
```

```
+-------------------+---------+------+
|               date|day(date)|deaths|
+-------------------+---------+------+
|2020-12-31 00:00:00|       31|346050|
|2021-12-31 00:00:00|       31|824336|
|2022-05-13 00:00:00|       13|998279|
+-------------------+---------+------+
```