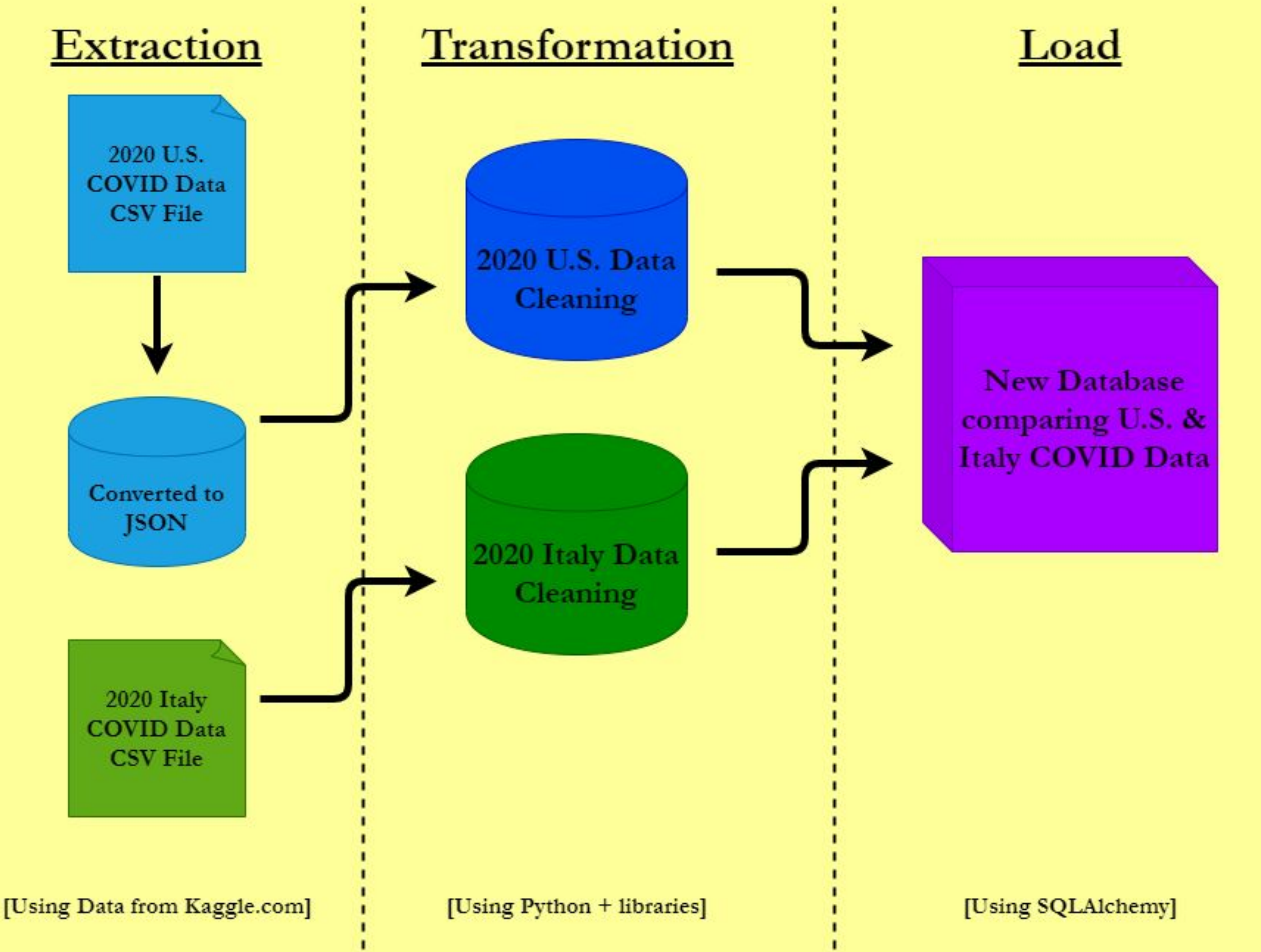
Covid 19: The United States VS. Italy

Group 11
Group members: Asim Syed, Connor Scherer, Digiant Patel, Stephany Obakpolor, Jessica Doanes, John Timmers



Extraction

2020 U.S. COVID Data CSV File

Converted to JSON

2020 Italy COVID Data CSV File

[Using Data from Kaggle.com]

Transformation

2020 U.S. Data Cleaning

2020 Italy Data Cleaning

[Using Python + libraries]

Load

New Database comparing U.S. & Italy COVID Data

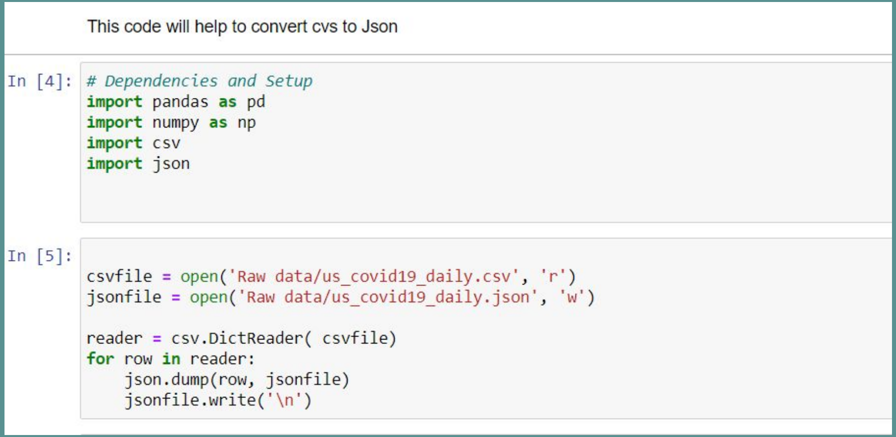
[Using SQLAlchemy]

Extraction

We retrieved data from COVID-19 in USA and COVID-19 in Italy from Kaggle.

- The assignment required that we use diverse data sources. To meet this requirement, we converted the USA data set from a CSV to JSON file.

- Then, we used Python to extract data from JSON and CSV.



This code will help to convert cvs to Json

In [4]:

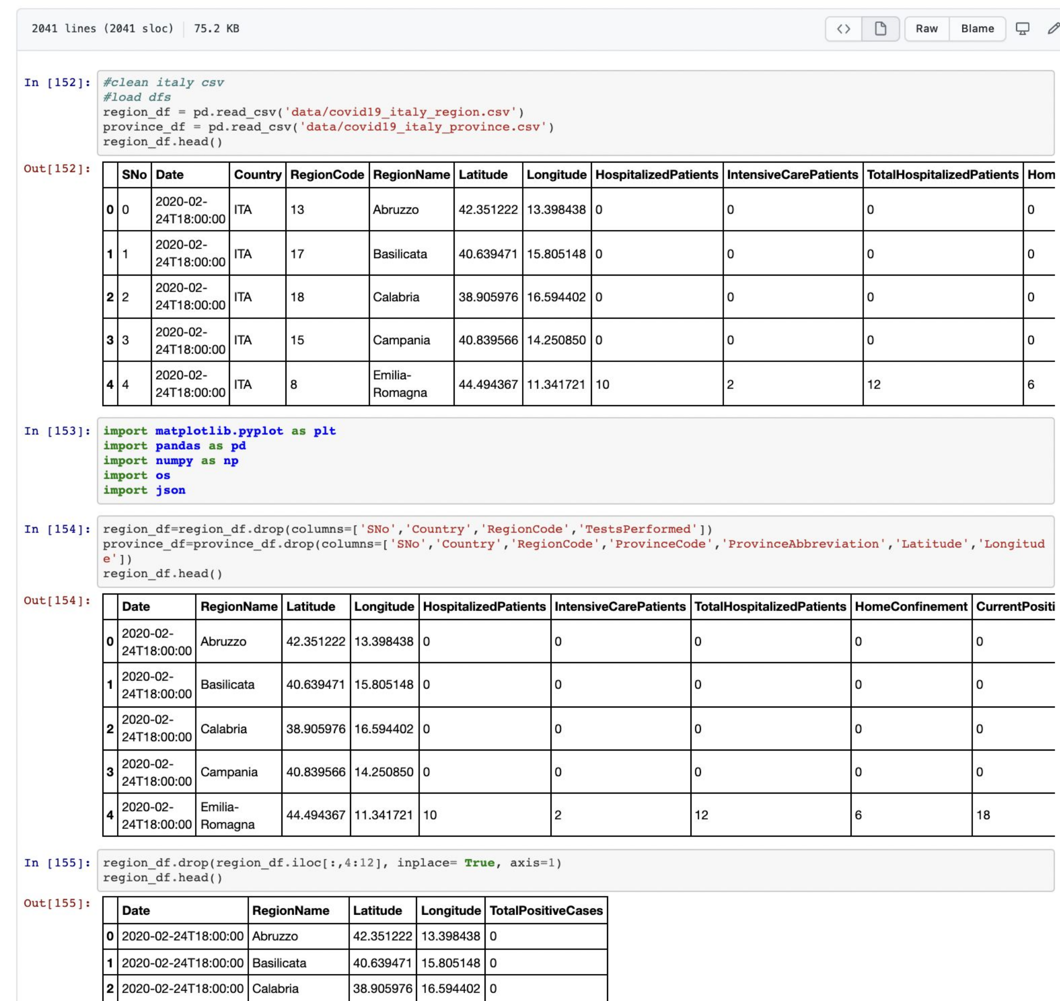
```
# Dependencies and Setup
import pandas as pd
import numpy as np
import csv
import json
```

In [5]:

```
csvfile = open('Raw data/us_covid19_daily.csv', 'r')
jsonfile = open('Raw data/us_covid19_daily.json', 'w')

reader = csv.DictReader( csvfile)
for row in reader:
    json.dump(row, jsonfile)
    jsonfile.write('\n')
```

Transformation

- Following successful conversion of the USA data set, we cleaned our data for USA and Italy.



2041 lines (2041 sloc) | 75.2 KB

In [152]:
```
#clean italy csv
#load dfs
region_df = pd.read_csv('data/covid19_italy_region.csv')
province_df = pd.read_csv('data/covid19_italy_province.csv')
region_df.head()
```

Out[152]:

	SNo	Date	Country	RegionCode	RegionName	Latitude	Longitude	HospitalizedPatients	IntensiveCarePatients	TotalHospitalizedPatients	Hom
0	0	2020-02-24T18:00:00	ITA	13	Abruzzo	42.351222	13.398438	0	0	0	0
1	1	2020-02-24T18:00:00	ITA	17	Basilicata	40.639471	15.805148	0	0	0	0
2	2	2020-02-24T18:00:00	ITA	18	Calabria	38.905976	16.594402	0	0	0	0
3	3	2020-02-24T18:00:00	ITA	15	Campania	40.839566	14.250850	0	0	0	0
4	4	2020-02-24T18:00:00	ITA	8	Emilia-Romagna	44.494367	11.341721	10	2	12	6

In [153]:
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import json
```

In [154]:
```
region_df=region_df.drop(columns=['SNo','Country','RegionCode','TestsPerformed'])
province_df=province_df.drop(columns=['SNo','Country','RegionCode','ProvinceCode','ProvinceAbbreviation','Latitude','Longitud
e'])
region_df.head()
```

Out[154]:

	Date	RegionName	Latitude	Longitude	HospitalizedPatients	IntensiveCarePatients	TotalHospitalizedPatients	HomeConfinement	CurrentPositi
0	2020-02-24T18:00:00	Abruzzo	42.351222	13.398438	0	0	0	0	0
1	2020-02-24T18:00:00	Basilicata	40.639471	15.805148	0	0	0	0	0
2	2020-02-24T18:00:00	Calabria	38.905976	16.594402	0	0	0	0	0
3	2020-02-24T18:00:00	Campania	40.839566	14.250850	0	0	0	0	0
4	2020-02-24T18:00:00	Emilia-Romagna	44.494367	11.341721	10	2	12	6	18

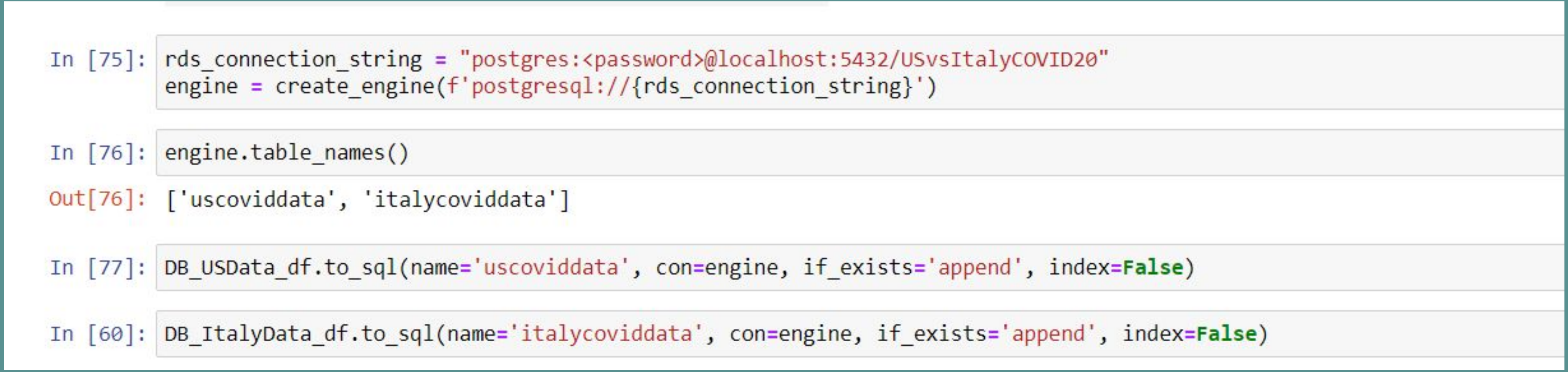
In [155]:
```
region_df.drop(region_df.iloc[:,4:12], inplace= True, axis=1)
region_df.head()
```

Out[155]:

	Date	RegionName	Latitude	Longitude	TotalPositiveCases
0	2020-02-24T18:00:00	Abruzzo	42.351222	13.398438	0
1	2020-02-24T18:00:00	Basilicata	40.639471	15.805148	0
2	2020-02-24T18:00:00	Calabria	38.905976	16.594402	0

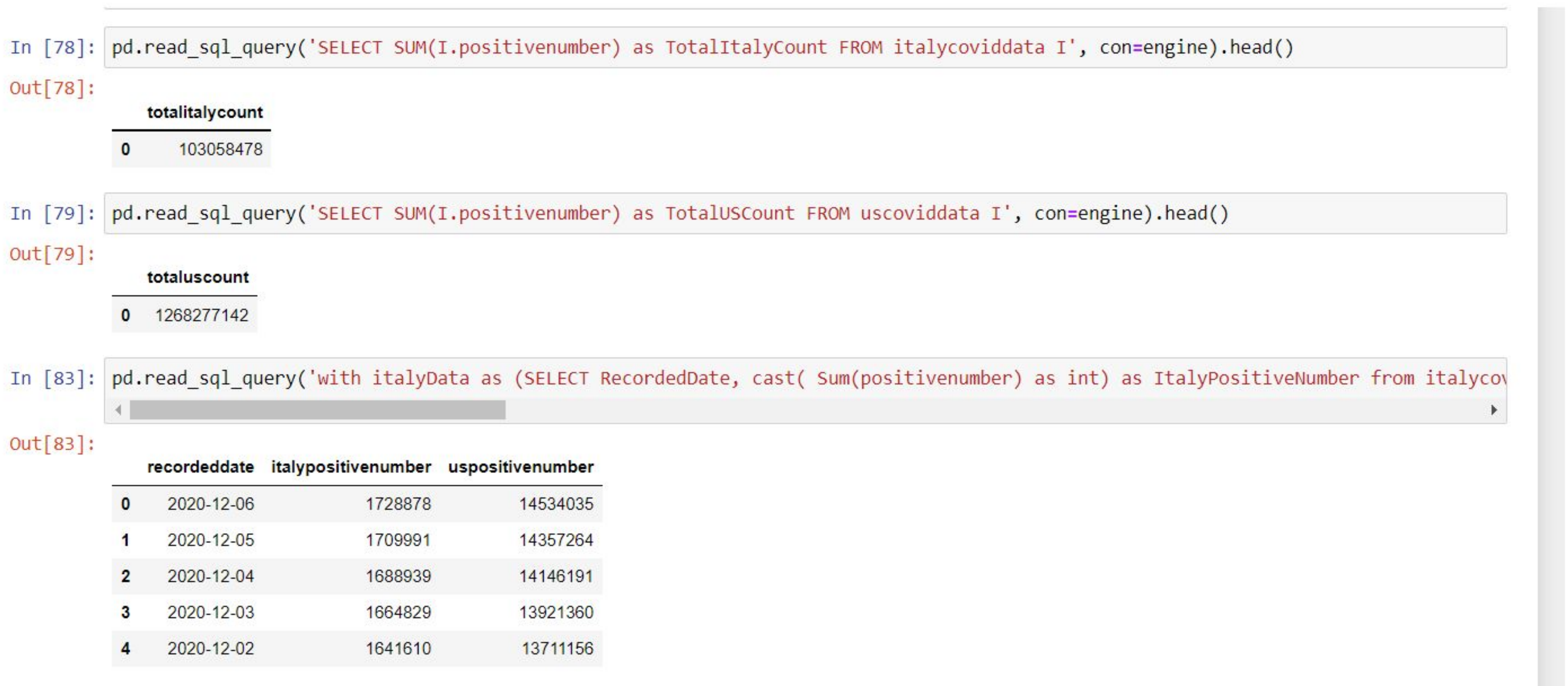
Load

Finally loaded data into Database



```
In [75]: rds_connection_string = "postgres:<password>@localhost:5432/USvsItalyCOVID20"
         engine = create_engine(f'postgresql://{rds_connection_string}')

In [76]: engine.table_names()

Out[76]: ['uscoviddata', 'italycoviddata']

In [77]: DB_USData_df.to_sql(name='uscoviddata', con=engine, if_exists='append', index=False)

In [60]: DB_ItalyData_df.to_sql(name='italycoviddata', con=engine, if_exists='append', index=False)
```

Final result: Comparing US vs Italy



```
In [78]: pd.read_sql_query('SELECT SUM(I.positivenumber) as TotalItalyCount FROM italycoviddata I', con=engine).head()
```

Out[78]:

	totalitalycount
0	103058478

```
In [79]: pd.read_sql_query('SELECT SUM(I.positivenumber) as TotalUSCount FROM uscoviddata I', con=engine).head()
```

Out[79]:

	totaluscount
0	1268277142

```
In [83]: pd.read_sql_query('with italyData as (SELECT RecordedDate, cast( Sum(positivenumber) as int) as ItalyPositiveNumber from italycov
```

Out[83]:

	recordeddate	italypositivenumber	uspositivenumber
0	2020-12-06	1728878	14534035
1	2020-12-05	1709991	14357264
2	2020-12-04	1688939	14146191
3	2020-12-03	1664829	13921360
4	2020-12-02	1641610	13711156