Project 2 ETL Technical Report by Team 3

Anthony Stevens, Cecilia Zhang, Neda Mehdizadeh, and Sahar Jamal

Project Theme : Covid-19 in the USA

Step 1. Extract: Data Sources

- Covid-19 in USA-Number of Novel CoronaVirus 2019 cases in USA from Kaggle.com 3 csv files (primary)
  Reference Link: https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_counties_covid19_daily.csv
- State, County, CBSA, Country API Data from Covid Act Now (primary)
  Reference Link: https://apidocs.covidactnow.org/data-definitions/

Step 2. Transform: Data Wrangling

- Covid daily data is retrieved from “us_counties_covid19_daily”csv file, then encoded the fips (Federal Information Processing System) column to an int instead of a float. After that the null values were dropped to keep only usable data.

- The original data is scraped from json file “counties. timeseries” and is cleaned by selecting columns with useful information. Then dropped columns with NaN values such as “lat” or contained dictionaries which were not useful.. Leaving us with a handful of useful data. We ran into an issue with two of the columns throwing errors for “undefined columns”. The solution was to rename and lowercase the names of the columns.

Step 3. Loading:

We chose PostgreSQL as our database to load our data. The reason for this was due to this database being a common sql database, so we knew the expertise in this area would definitely challenge us to perfect our skills. After tackling the portion of data we decided to keep in pandas, we then lined up a table within postgres. To do so we used the <CREATE TABLE> statement, and then we identified each column we wanted in the table. The next step was to run the table and check back for the table to upload. We then entered into pandas and connected the engine so that the data would load into the database and align nicely. To confirm our work was done correctly, we utilized the <SELECT *> of the specific table to print what had loaded.

Project 2- A Case Study of ETL

Covid-19 Datasets

Team 3
Anthony Stevens, Cecilia Zhang, Neda Mehdizadeh, and Sahar Jamal

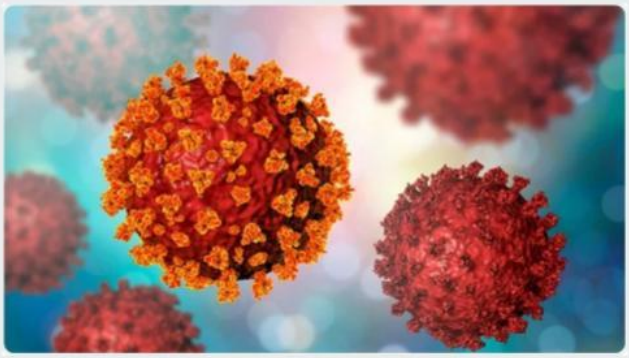


Illustration Credit: Kateryna Kon / Shutterstock

Extract: Data Sources

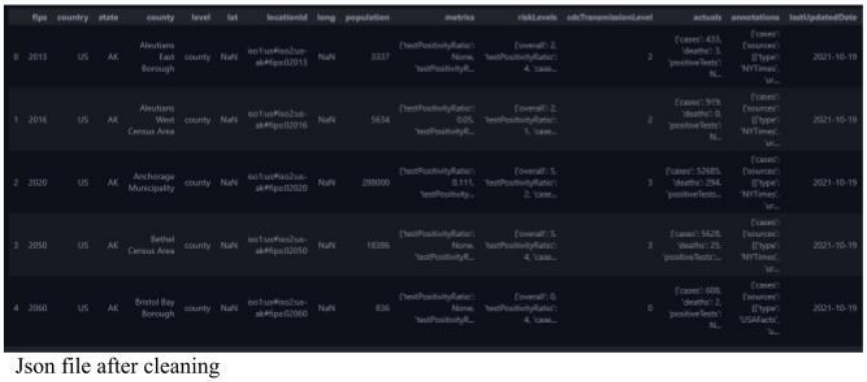
1. **Covid-19 in USA-Number of Novel Corona Virus 2019 cases in USA from Kaggle.com 3 csv files (primary)**
   Reference Link:
   https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_counties_covid19_daily.csv

2. **State, County, CBSA, Country API Data from Covid Act Now (primary)**
   Reference Link: https://apidocs.covidactnow.org/data-definitions/

API Value	CDC Community Transmission Level
0	Low
1	Moderate
2	Substantial
3	High
4	“Unknown”

Transform: Data Wrangling pt.1

Json file before cleaning



	fips	country	state	county	level	lat	locationId	long	population	metrics	riskLevels	cdcTransmissionLevel	actuals	annotations	lastUpdatedDate
0	2013	US	AK	Aleutians East Borough	county	NaN	iso1us#iso2us-ak#fips02013	NaN	3337	{'testPositivityRatio': None, 'testPositivityR...	{'overall': 2, 'testPositivityRatio': 4, 'case...	2	{'cases': 433, 'deaths': 3, 'positiveTests': N...	{'cases': {'sources': [{'type': 'NYTimes', 'ur...	2021-10-19
1	2016	US	AK	Aleutians West Census Area	county	NaN	iso1us#iso2us-ak#fips02016	NaN	5634	{'testPositivityRatio': 0.05, 'testPositivityR...	{'overall': 2, 'testPositivityRatio': 1, 'case...	2	{'cases': 919, 'deaths': 0, 'positiveTests': N...	{'cases': {'sources': [{'type': 'NYTimes', 'ur...	2021-10-19
2	2020	US	AK	Anchorage Municipality	county	NaN	iso1us#iso2us-ak#fips02020	NaN	288000	{'testPositivityRatio': 0.111, 'testPositivity...	{'overall': 3, 'testPositivityRatio': 2, 'case...	3	{'cases': 52685, 'deaths': 294, 'positiveTests...	{'cases': {'sources': [{'type': 'NYTimes', 'ur...	2021-10-19
3	2050	US	AK	Bethel Census Area	county	NaN	iso1us#iso2us-ak#fips02050	NaN	18386	{'testPositivityRatio': None, 'testPositivityR...	{'overall': 3, 'testPositivityRatio': 4, 'case...	3	{'cases': 5620, 'deaths': 25, 'positiveTests':...	{'cases': {'sources': [{'type': 'NYTimes', 'ur...	2021-10-19
4	2060	US	AK	Bristol Bay Borough	county	NaN	iso1us#iso2us-ak#fips02060	NaN	836	{'testPositivityRatio': None, 'testPositivityR...	{'overall': 0, 'testPositivityRatio': 4, 'case...	0	{'cases': 608, 'deaths': 2, 'positiveTests': N...	{'cases': {'sources': [{'type': 'USAFacts', 'u...	2021-10-19

Json file after cleaning

	fips	country	state	county	population	transmission_level	last_updated
0	2013	US	AK	Aleutians East Borough	3337	2	2021-10-19
1	2016	US	AK	Aleutians West Census Area	5634	2	2021-10-19
2	2020	US	AK	Anchorage Municipality	288000	3	2021-10-19
3	2050	US	AK	Bethel Census Area	18386	3	2021-10-19
4	2060	US	AK	Bristol Bay Borough	836	0	2021-10-19

Data Wrangling pt. 2

Reading csv file

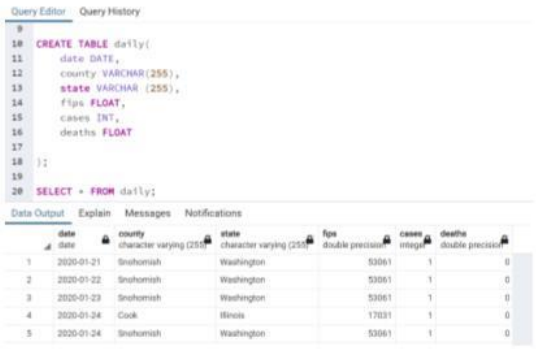
```
covid_daily_data_df.fillna(0, inplace=True)
covid_daily_data_df['fips'] = covid_daily_data_df['fips'].astype(int)
covid_daily_data_df = covid_daily_data_df[covid_daily_data_df['fips'] != 0]
```

	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061	1	0.0
1	2020-01-22	Snohomish	Washington	53061	1	0.0
2	2020-01-23	Snohomish	Washington	53061	1	0.0
3	2020-01-24	Cook	Illinois	17031	1	0.0
4	2020-01-24	Snohomish	Washington	53061	1	0.0

```
cleaned_df.to_excel("cleaned_df.xlsx")
```

Confirm data has been added by using pgAdmin

Reading table “daily”



```
CREATE TABLE daily(
    date DATE,
    county VARCHAR(255),
    state VARCHAR (255),
    fips FLOAT,
    cases INT,
    deaths FLOAT

);

SELECT * FROM daily;
```

	date	county	state	fips	cases	deaths
1	2020-01-21	Snohomish	Washington	53061	1	0
2	2020-01-22	Snohomish	Washington	53061	1	0
3	2020-01-23	Snohomish	Washington	53061	1	0
4	2020-01-24	Cook	Illinois	17031	1	0
5	2020-01-24	Snohomish	Washington	53061	1	0

Reading table “current_data”

```
CREATE TABLE current_data(
    fips FLOAT,
    country VARCHAR(255),
    state VARCHAR(255),
    county VARCHAR (255),
    population INT,
    transmission_level INT,
    last_updated DATE
);

SELECT * FROM current_data;
```

	fips	country	state	county	population	transmission_level	last_updated
1	2013	US	AK	Aleutians East Borough	3337	2	2021-10-19
2	2016	US	AK	Aleutians West Census Area	5634	2	2021-10-19
3	2020	US	AK	Anchorage Municipality	288000	3	2021-10-19
4	2050	US	AK	Bethel Census Area	18386	3	2021-10-19

Connect to local database and Use pandas to load csv/json converted DataFrame into database

```
covid_daily_data_df.to_sql(name="daily", con=engine, if_exists='append', index=False)
new_covid_json.to_sql(name='current_data', con=engine, if_exists='append', index=False)
```

Thank You!