Extract Transform Load Project

Team: John Chan, Emory Mansuetti, Khank-Linda Stark,

Jackson Freese, DeJuan Hall, Siddharth Das

UC Davis, Data Analytics Bootcamp

June 10, 2021

Abstract

Data literacy has become an essential skill for business managers and professionals in all fields. The ETL project is a case study that employs technical capabilities that we have learned in class. We wanted to compare how COVID-19 impacted to two coastal states – New York and California. In order to do this, we had to extract COVID-19 data from three disparate sources, transforming the data, e.g., performing data cleaning, concatenation, and aggregation, to suit business requirements, and finally load the data into a database that can be used for future analysis or business use.

Keywords: ETL, Extract, Transform, Load, Python, SQL, APIs, pgAdmin 4, iFrame

Extract Transform Load Project

# Proposal and Preparation

The team discussed the fact that we had a little over a week to accomplish our work. We agreed to start with modest goals that would allow us to complete the project earlier than the allotted time, which would allow us to work on more complex tasks if we had time.

## Extraction1

We initially agreed to use two different datasets from, “The COVID Tracking Project”: a CSV file and data extracted from an API. However, we found that we had extra time, so we took on an additional challenge of scraping dynamic chart URLs, embedded in iFrames, on the Syracuse website. We selected one of the URLs and loaded it into a browser. We then inspected the page, looking into the Network, then examined if there were XMLhttpRequests (XHR). We found that there was another URL, which led us to datawrapper.de, a provider of dynamic visual wrappers. We then downloaded the chart’s CSV data. – our third dataset.

* The CA CSV data was directly downloaded from the website
* JSON API was used to request services and retrieve data for the NY dataset
* For the Syracuse URL, urllib.request was used to fetch and read the website
* Beautifulsoup was used to parse the html page and find all src URLs that were embedded in iFrames.
* We selected one URL, “static.dwcdn.net/data/ijEiy.csv?v=1623222240000”, a dynamic chart, loaded it into a browser, used the inspect tool to peer into the Network activity after refreshing the page. We then filtered by XHR to find URL redirect.
* Python and Response were used to extract, then export the chart data to CSV.

## Selected Data Fields:

|  |
| --- |
| Data Field Name |
| Date (Primary Key) |
| state |
| deathIncrease |
| hospitalizedIncrease |
| inIcuCurrently |
| PositiveCasesViral |
| positiveIncrease |
| totalTestResults |
| totalTestResultsIncreas**e** |

## Data Sources:

* <https://api.covidtracking.com/v1/states/ca/current.csv>
* <https://api.covidtracking.com/v1/states/ny/daily.json>
* <https://www.syracuse.com/coronavirus-ny/>
* <https://datawrapper.dwcdn.net/ijEiy/2/>
* <https://static.dwcdn.net/data/ijEiy.csv?v=1623222240000>

## Transform

We performed transformation of the data to suit our needs, including:

* Pandas functions used:
* CA: Load CSV file, drop unnecessary columns, renamed NaN/missing values to zero, and exported to back to CSV
* NY: Dictionary of variables created to store data from the requested API calls
* Converted dictionary into DF, changed date format, exported to a CSV file
* Performed concatenation merge, inner of NY and CA datasets
* Removed “totalTestResultsIncrease” column, as it contained zeros
* Renamed columns to make it more readable
* Sort the merged DataFrame for each item listed below, and Paste image of dataframe.head().
* Deaths
* Daily Hospitalizations
* ICU Hospitalized

# Load

Load information goes here. **Jackson**, please write what you did to prepare the dataframe to load into SQL.

**Emory**, please write what you did to prepare the SQL database, e.g. create scheme, create database, fields, etc.

## Postgres Database:

* Paste screen grab of the Postgres Database after loading the python dataframes.
* Paste screen grab of the Postgress dataframe of

# Summary

DeJuan work with everyone to write the summary

Summary info goes here.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well. That’s at the end of the sample Heading 2 paragraph on the first page of body content in this template.)]

Tables

Table 1

[Table Title]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the APA Style Manual, 6th Edition.