

Project 3: Data Transformation and Automation

Project Brief


Scenario

In the highly competitive movie streaming services market, your client has asked for help with enriching their data with publicly available data. Using a dataset from your client containing different data about movies, you are tasked with scraping online publicly available data from the Internet Movie Database (IMDb), one of the most popular websites that contains large amounts of data on movies. You'll need to transform the scraped data into a structured format and integrate it with your client's data to come up with an enriched dataset.


Key Tasks

This project has two parts with multiple tasks and separate deliverables for each part. Read each set of instructions carefully.


▼ PART A: Data Gathering, Transformation, and Enrichment.

Perform data gathering using web scraping to enrich your client's dataset ([Movies.csv](https://northeastern.instructure.com/courses/104383/files/14635965?wrap=1) ([https://northeastern.instructure.com/courses/104383/files/14635965?wrap=1](https://northeastern.instructure.com/courses/104383/files/14635965/download?download_frd=1))  https://northeastern.instructure.com/courses/104383/files/14635965/download?download_frd=1) containing top voted 500 movies released between 2018 and 2020. The dataset includes the following fields:

- movie_id: alphanumeric unique identifier of the title.
- originalTitle: original title, in the original language.
- titleType: the type/format of the title (e.g., movie, short, tvseries, tvepisode, video, etc.).
- isAdult: 0 for non-adult title; 1 for adult title.
- genres : includes up to three genres associated with the title.


Download the [Project 3 Part A.ipynb \(https://northeastern.instructure.com/courses/104383/files/14635964?wrap=1\)](https://northeastern.instructure.com/courses/104383/files/14635964?wrap=1)  (https://northeastern.instructure.com/courses/104383/files/14635964/download?download_frd=1) Jupyter Notebook adding your last name to the filename. Edit the code in the notebook to complete the following tasks:

1. Conduct Data Gathering:

- Scrape this [IMDb webpage](https://www.imdb.com/search/title/?at=0&sort=num_votes,desc&start=1&title_type=feature&year=2018,2020)  (https://www.imdb.com/search/title/?at=0&sort=num_votes,desc&start=1&title_type=feature&year=2018,2020) of movies released between 2018 and 2020, sorted by votes in descending order. Pull movie_id, rank, title, year, rating, runtime, and votes for the top 500 movies sorted by user number of votes in descending order.
- Transform the scraped data to a structured format and write it to a CSV file (name it IMDb_TopVoted.csv).

► Supporting Materials

2. Conduct Data Enrichment:

- Import the [Movies.csv \(https://northeastern.instructure.com/courses/104383/files/14635965?wrap=1\)](https://northeastern.instructure.com/courses/104383/files/14635965?wrap=1)  (https://northeastern.instructure.com/courses/104383/files/14635965/download?download_frd=1) file to a pandas DataFrame called df1.
- Import the scraped data from the IMDb_TopVoted.csv file to a pandas DataFrame called df2.
- Implement data cleansing and transformation for the df2.
- Enrich the given dataset (df1) by merging it with the scraped data (df2).
- Rearrange the dataset fields to be listed in the following order:
 - movie_id, rank, votes, title, originalTitle, year, rating, titleType, isAdult, runtime, genres
- Export the enriched dataset to a CSV file:
 - Use the following naming convention: Project_3_Part_A_Group#.csv

► Supporting Materials

▼ PART B: Automate Data Transformation and Integration.

Use Alteryx to automate the process that you applied in Part A to clean, transform, and integrate the data.

1. Create Alteryx workflow to:

- a. Import IMDb_TopVoted.csv dataset you created in Part A.

- b. Do the necessary data cleansing and transformation.
- c. Import **Movies.csv** (<https://northeastern.instructure.com/courses/104383/files/13597305?wrap=1>)_ ⬇ (https://northeastern.instructure.com/courses/104383/files/13597305/download?download_frd=1) dataset.
- d. Merge the two datasets to obtain the enriched dataset.
- e. Sort the enriched dataset by rank in ascending order, and rearrange the dataset fields to be listed as the following: movie_id, rank, votes, title, originalTitle, year, rating, titleType, isAdult, runtime, genres
- f. Export the enriched dataset to CSV file:
 - Use the following naming convention: Project_3_Part_B_Group#.csv

► Supporting Materials

2. Report in a Word document, a brief description of the following:

- What data was used to enrich the client's data?
- Describe the data cleaning and transformation that was implemented.

What to Submit:

PART A: Upload the following 4 files:

- The edited Jupyter notebook in .IPYNB format with annotations that explain and document your work.
- A copy of the Jupyter notebook in .HTML format.
- CSV file for the scraped data (IMDb_TopVoted_Group#.csv).
- CSV file for the enriched dataset (Project_3_Part_A_Group#.csv).

PART B: Upload the following 3 files:

- Alteryx file for the workflow (Project_3_Part_B_Group#.yxmd).
- CSV file for the output enriched dataset (Project_3_Part_B_Group#.csv).
- Word document with written description (Project_3_Part_B_Group#.doc).