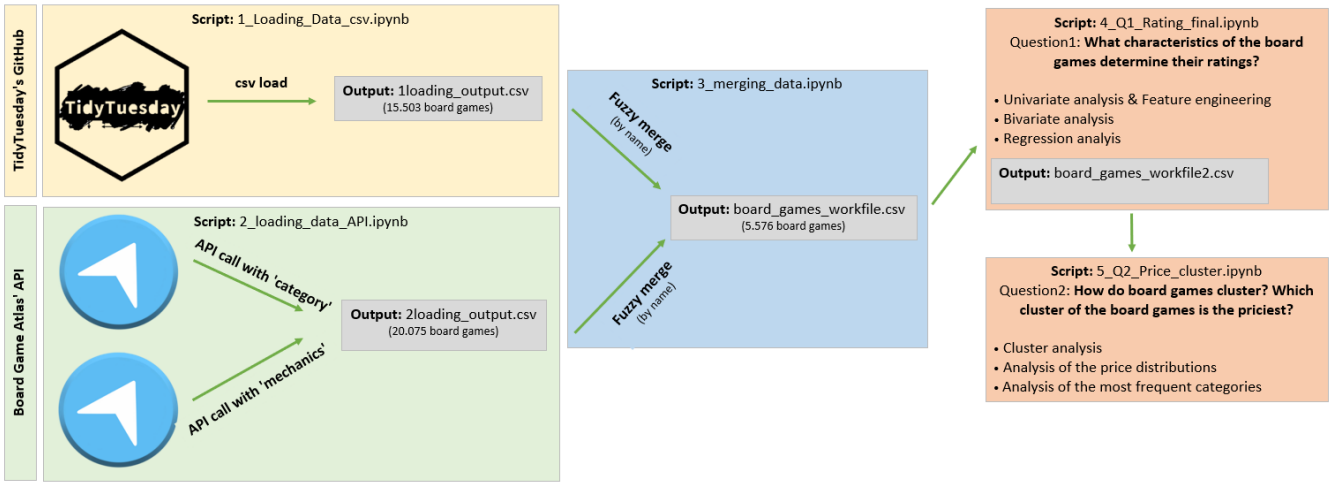
Analysis of board games

András Bognár, Viktória Kónya

Description of the project

For our term project we chose to analyze characteristics of board games. The main source of our data is the GitHub repo of TidyTuesday from where we downloaded the dataset in csv format. It contains rich information about board games but lacks details about their prices. Therefore, we requested this data using the Board Game Atlas' API along with information about the learning and strategy complexity of the games.

The chart below summarizes our data pipeline with the corresponding notebook references.



The **1_Loading_Data_csv.ipynb** contains the loading of the TidyTuesday's data with initial cleaning (removing duplicates, dealing with high missing rate and dropping unnecessary fields). The dataset consists of 15.503 unique board games.

The **2_loading_data_API.ipynb** contains the API requests along with the same preliminary data cleaning as described above. We created our dataset using two different API calls (one using the category and another call using the mechanics as keys in the API call). The dataset consists of 20.075 unique board games.

The **3_merging_data.ipynb** contains the joining of the two sources. We joined them on the names of board games via fuzzy merge. The final workfile consists of 5.576 unique board games. Note that we suffixed the fields from the TidyTuesday's dataset with '_base' and the ones from the API with '_api'.

We chose to examine two analytical questions about the board games:

- **What characteristics of the board games determine their ratings?**
- **How do board games cluster? Which cluster of the board games is the priciest?**

The **4_Q1_Rating_final.ipynb** contains the analysis of the first question, including a detailed analysis of each variable of interest, additional feature engineering steps, elimination of extreme values and recoding of the variables. The analysis starts from the univariate analysis, then explores the relationship of each predictor with the outcome and finally examines the pattern of association in regression framework.

The **5_Q2_Price_cluster.ipynb** contains our second analysis where we used cluster analysis to create separate groups based on different characteristics. Once we created separate clusters we examined how the distribution of the prices differ across the clusters. We also examined which categories are the most frequent in each cluster.

31 March 2022

Appendix

List of fields in the TidyTuesday's dataset:

variable	description
game_id	Unique game identifier
description	A paragraph of text describing the game
max_players	Maximum recommended players
max_playtime	Maximum recommended playtime (min)
min_age	Minimum recommended age
min_players	Minimum recommended players
min_playtime	Minimum recommended playtime (min)
playing_time	Average playtime
year_published	Year game was published
artist	Artist for game art
category	Categories for the game (separated by commas)
compilation	If part of a multi-compilation - name of compilation
family	Family of game - equivalent to a publisher
mechanic	Game mechanic - how game is played, separated by comma
publisher	Comoany/person who published the game, separated by comma
publisher	Average rating on Board Games Geek (1-10)
users_rated	Number of users that rated the game
name	Name of the game (PK)

List of used fields from the API:

variable	description
average_learning_complexity	Average learning complexity score
average_strategy_complexity	Average strategy complexity score
name	Name of the game (PK)
price	Price in USD
price_au	Price in AUD
price_ca	Price in CAD
price_uk	Price GBP