CS 638 - Fall 2016
Team 20 - Project Stage 5

Group Members:
Daniel Kaczmarek <dkaczmarek@wisc.edu>,
Sahit Mandala, <mandala@wisc.edu>,
Zhilin Jiang <zjiang62@wisc.edu>

Analysis Questions/Topics

How did you combine the two tables A and B to obtain E? Did you add any other table? When you did the combination, did you run into any issues? Discuss the combination process in detail, e.g., when you merge tuples, what are the merging functions (such as to merge two age values, always select the age value from the tuple from Table A, unless this value is missing in which case we select the value from the tuple in Table B).

- We did a left-inner-join on table A and B based on the matching results from Stage 4. Specifically, we first ran a left-inner-join between the finaling matching tuple data set C and A, joining on the table A's id with C's ltable_id. Then, on this new joined table CA, we run another left-inner-join between CA and B, using table B's id with CA's rtable_id.

- After the join, we created the final scheme via merging attribute columns from A and B in the final joined table. When merging attributes shared by both table A and B (e.g. min_num_players, name) in our final joined table, if only one of them had data for a value we used that value, if both of them had a value we took the value from tableA which we got from BoardGameGeek. We do this because BoardGameGeek is the main hub for board game information and is also more likely (in our opinion) to have the correct data as opposed to the smaller BoardGamePrices.com

Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.

- Table E schema: id,ltable_id,rtable_id,id_C,name,year,rating,rank,num_players,min_num_players,max_num_players,gameplay_time,min_gameplay_time,max_gameplay_time,min_age,complexity_weight,category,mechanisms,type,BGG_link,store_names,store_prices,links_to_buy,availability,international_store

- Table E contains 563 tuples.
  Four sample tuples:
  https://github.com/malnoxon/board-game-data-science/blob/master/stage5/4_tuples.csv

What was the data analysis task that you wanted to do? (Example: we wanted to know if we can use the rest of the attributes to accurately predict the value of the attribute loan_repaid.) For that task, describe in detail the data analysis process that you went through.

- Correlation detection among the following attribute pairs:

  Rating vs. Year
  Complexity Weight vs. Year
  Price (mean) vs. Rating
  Num_players vs. Complexity Weight
  Gameplay_time vs. Complexity Weight

- Initially we plotted the variables against each other to visualize the relative relationships and whether we could see any semblance of a relationship. However, this we merely visual inspection before moving to more rigorous methods for correlation analysis.

- In order to determine the nature of the correlation, we use a Pearson correlation coefficient, which indicates the strength of the linear correlation of 2 variables. In order to verify the significance of this correlation, we generated P values from a 2-tailed p test that is run while generating the Pearson coefficient. Given our minimal dataset, we opted for $p<0.02$ for our threshold of significant results from this analysis

Give any accuracy numbers that you have obtained (such as precision and recall for your classification scheme).

- We opted for correlation analysis for this section, so we don't have any precision/recall values to report. However, we can report the Pearson correlation values we obtained and the corresponding p-value.

Correlation Variables	Pearson coefficient	2-tailed p-value
Rating vs Year	0.2546870576858738	1.0366175204534563e-09
Complexity weight vs year	-0.018067907504713075	0.67987091649355857
Price (mean) vs. Rating	0.20436585163223553	2.0503303236073994e-06
Min_Num_players vs. Complexity Weight	-0.11511794376748428	0.0083482417340441876
Gameplay_time vs. Complexity Weight	0.16917911927689597	9.9616299640073566e-05

What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?

- We had to make choices among different ways of handling missing values, especially those involving attributes we are doing analysis on. Possible alternatives include removing the tuple, imputing with mean value, imputing with zero, etc.

- Our usage of the Pearson coefficient implies several strong assumptions on how these variables work. For instance, the Pearson correlation assumes a purely linear correlation, though there may be more complicated behaviors between various variables, such as exponential or logarithmic. It also assumes the data is on a interval or ratio scale, which we can strongly assume for attributes like num_players and year but not complexity weight or price. Finally, it assumes a bivariate normal distribution among both of these variables, which seems reasonable for cases like ratings vs year, where games in a given year are related on a “curve” of sorts, but this again is a strong assumption that is hard to justify for cases like price vs year, where prices between games of a year are more complicated than just a normal around some “average price”.

- During our analysis, we also found missing values from among the non-matching attributes (e.g. rating, price, etc). In order to not skew our correlation analysis by using imputation via means or random values, we opted to prune these values from the corresponding analysis, hence only using analysis on variables, say x and y, only if the tuple were not missing both x and y.

If you have more time, what would you propose you can do next?

- Analysis on international prices, eg try to compare costs of games across currencies and use exchange rates to determine meaningful differences between prices of games in different countries.

- Analyze popularity of games vs how many stores sell them, popularity vs percentage of stores that are out of stock of the game

- We could try to do some analysis to try to predict the success of a game based on it’s price, mechanics, category, player count