Bechdel test - report

D2
Anette Habanen, Triin Schaffrik

Task 1:
Repository: https://github.com/TriinSchaffrik/Bechdel-test

Task 2:

- Identifying your business goals
    - Background
        - We wanted to make this project about movies. After searching for a dataset for some time, we found a dataset about Bechdel test. This topic seemed interesting so we decided to use that data for this project.
    - Business goals
        - This is not relevant in our project. But this information could be used by the entertainment industry to consider before making their own movie. For instance, if movies, which pass the Bechdel test tend to make more or less profit than movies, which does not pass this test.
    - Business success criteria
        - This is not relevant in our project.
- Assessing your situation
    - Inventory of resources
        - This project is conducted by Anette Habanen and Triin Schaffrik
        - For this project we use two datasets. One dataset has a long list of movies with a lot of information (https://data.world/emmahyams/do-the-best-shows-on-netflix-pass-the-bechdel-test/workspace/file?filename=updated_movie_data.csv). The second dataset has less movies but also consists the Bechdel test result for that film (https://data.world/emmahyams/do-the-best-shows-on-netflix-pass-the-bechdel-test/workspace/file?filename=bechdel_clean.csv).
        - To analyze the data we use Google Colab.
    - Requirements, assumptions, and constraints
        - To complete the project, we make video calls twice per week. If this is not enough, we will work on the project outside of the video calls.
        - The video for this project will be submitted by the noon (12:00) of Monday, Dec 13.
        - We will present the project in the poster session on Thursday, December 16, 2021 at 14:00-17:00.
    - Risks and contingencies
        - One risk could be that we underestimate how much time a task takes and it messes up the schedule.
        - Another risk could be that we do not find interesting relations to choose the features for our model.
    - Terminology

        - Bechdel test - measure of the representation of women in fiction. It asks whether a work features at least two women who talk to each other about something other than a man.
    - Costs and benefits
        - This is not relevant in your project.
- Defining your data-mining goals
    - Data-mining goals
        - During this project we intent to uncover interesting relations that are not obvious from the original data.
        - We want to make a model that would predict if a movie passes the Bechdel test or not. To test that model we can input information about the movies that are not in the dataset (for example Estonian movies) and see if the prediction is correct.
        - We also want to visualize our findings so that they are easily understandable to other people.
        - Present our findings and trained model to the other students.
    - Data-mining success criteria
        - Firstly, we will analyze how different features affect the Bechdel test result. We will find at least 3 relations that influence the outcome.
        - Based on that we will train a model to predict if a movie passes the Bechdel test. It takes some attributes, which we have find that affects significantly test result, as input, so the movies about which we want to predict does not have to have unimportant information.

Task 3:

- Gathering data
    - Outline data requirements
        - The datasets must have a selection of different properties about the movies. It is also important to have the Bechdel test results for some of the movies in our dataset.
    - Verify data availability
        - To get access to the data, it is necessary to have a data.world account. After that it is easy to download the dataset and share it with other teammembers. Both datasets are also available in our Github repository.
    - Define selection criteria
        - Our projects dataset is from the webpage: https://data.world/emmahyams/do-the-best-shows-on-netflix-pass-the-bechdel-test/workspace/file?filename=bechdel_clean.csv .
- Describing data
    - updated_movie_data.csv
        - This dataset has 29 features and 4974 rows. Each row describes a movie.
        - Some of the important features are: movie_title, gross, facenumber_in_poster, imdb_score.

    - bechdel_clean.csv
        - This dataset has 14 features and 1756 rows. Each row describes a movie. Most of the movies that are represented here are also in the ‘updated_movie_data.csv’ dataset.
        - Some of the important features are: binary - shows if movie passed the bechdel test or not, budget, year.
- Exploring data
    - updated_movie_data.csv
        - imdb_id (int) - imdb code
        - color (string) - black and white or color movie
        - director_name (string)
        - num_critic_for_reviews (int)
        - duration (int) - in minutes
        - director_facebook_likes (int)
        - actor_3_facebook_likes (int)
        - actor_2_name (string)
        - actor_1_facebook_likes (int)
        - gross (int)
        - genres (string) - multiple genres as a string, separated with ‘|’
        - actor_1_name (string)
        - movie_title (string)
        - num_voted_users (string)
        - cast_total_facebook_likes (string)
        - actor_3_name (string)
        - facenumber_in_poster (string)
        - plot_keywords (string) - multiple keywords as a string, separated with ‘|’
        - movie_imdb_link (url)
        - num_user_for_reviews (string)
        - language (string)
        - country (string)
        - content_rating (string)
        - budget (string)
        - title_year (string)
        - actor_2_facebook_likes (string)
        - imdb_score (decimal) - from 0 to 10
        - aspect_ratio (decimal)
        - movie_facebook_likes (int)
    - bechdel_clean.csv
        - imdb_id (int) - min 29356 and max 2425486
        - year (year) - min 1970 and max 2013
        - imdb (string) - max length 10
        - title (string) - title of the movie
        - test (string) - string that shows what the Bechdel test result is. Possible answers: notalk, ok, disagree, men, nowomen, dubious. The answers might consist of two possibilities.
        - clean_test (string) - if there were two answers for the previous feature, then one is chosen.

        - binary (string) - shows if the movie passed the test or not (pass or fail)
        - domgross (string)
        - intgross (int) - min 828 and max 2,783,918,982
        - code (string) - max length 9
        - budget_2013 (string)
        - domgross_2013 (string)
        - intgross_2013 (int) - min 899 and max 3,171,930,973
        - period_code (int)
        - decade_code (int)
- Verifying data quality
    - The data is good for our project. It needed little bit of cleaning and luckily it did not eliminate a lot of instances. There are also some NaN values but we are probably not going to use features that have a lot of NaN values.

Task 4:

- Tasks:
    - Project plan - done
        - 2 x 1.5 hours (both members contributed at the same time)
    - Cleaning the data - done
        - updated_movie_data.csv - Anette. 1.5 hours
        - bechdel_clean.csv - Triin, 1.5 hours
    - Report - done
        - 1.5 hours (both members contributed at the same time)
        - Anette 1 hour
        - Triin 1 hour
    - Finding interesting relations - in progress
        - 2 x 2 hours (both members contributed at the same time)
        - Estimation: 2 x 5 hours more
    - Training the model
        - Estimation: 2 x 14 hours
    - Making the video
        - Anette (estimation: 6 hours)
    - Making poster
        - Triin (estimation: 4 hours)

- Methods and tools:
    - Google Colab
    - Google Docs
    - Video calls + in person meetings
    - GitHub