Data Appendix File

Section 1: bechdel_movies.csv

Unit of Observation:

Each row in this dataset contains detailed information on a movie that is assessed by the Bechdel test after a sentiment analysis was performed.

Variables:

- Title: This categorical variable is the title of the movie, contained as a string. There are no missing observations out of 10,357 entries.
- Year: This numerical variable is the year the movie was released, contained as an integer. There are no missing observations out of 10,357 entries.
- Rating: This numerical variable indicates whether the movie passed or failed the Bechdel test using a 0 or 1, contained as an integer. There are no missing observations out of 10,357 entries.
- Dubious: This numerical variable indicated whether the review could be viewed as unreliable or dubious using 0.0 or 1.0, contained as a float. There are no missing observations out of 10,357 entries.
- Imdbid: The numerical variable gives the IMDb movie ID for each movie tested, contained as an integer. There are no missing observations out of 10,357 entries.
- Id: This numerical variable is an unique identifier for each entry, contained as an integer. There are no missing observations out of 10,357 entries.
- Submitterid: This numerical variable is another unique identifier for the person who submitted the Bechdel test rating, contained as an integer. There are no missing observations out of 10,357 entries.
- Date: This numerical variable details the dates and times that the reviews were added to the dataset, contained as a datetime variable. There are no missing observations out of 10,357 entries.
- Visible: This numerical variable indicated whether the movie has been approved with a 0 or 1, contained as an integer. For this dataset, all movies have been approved, meaning the return value will always be 1. There are no missing observations out of 10,357 entries.
- Sentiment: This numerical variable indicates the sentiment score for the reviews after they have been scraped from the IMDb website from -1 to 1, contained as a float. There are no missing observations out of 10,357 entries.

	year	rating	dubious	imdbid	id	submitterid	visible	sentiment
count	10357.000000	10357.000000	10058.000000	1.035700e+04	10357.000000	10357.000000	10357.0	10216.000000
mean	1996.927489	2.132471	0.090873	2.606133e+06	5538.922661	10318.339094	1.0	0.521243
std	25.079985	1.099714	0.287443	6.031473e+07	3247.689426	6763.568483	0.0	0.326541
min	1874.000000	0.000000	0.000000	1.000000e+00	1.000000	1.000000	1.0	-0.990000
25%	1989.000000	1.000000	0.000000	1.005190e+05	2700.000000	4063.000000	1.0	0.340000
50%	2006.000000	3.000000	0.000000	4.525980e+05	5497.000000	10187.000000	1.0	0.590000
75%	2014.000000	3.000000	0.000000	2.199543e+06	8363.000000	16373.000000	1.0	0.770000
max	2024.000000	3.000000	1.000000	6.129999e+09	11429.000000	22119.000000	1.0	1.000000

Section 2: bechdel_movies_2023_FEB.csv

Unit of Observation:

Each row in this dataset contains detailed information on a movie that is assessed by the Bechdel test.

Variables:

- Title: This categorical variable is the title of the movie, contained as a string. There are no missing observations out of 9,902 entries.
- Year: This numerical variable is the year the movie was released, contained as an integer. There are no missing observations out of 9,902 entries.
- Rating: This numerical variable indicates whether the movie passed or failed the Bechdel test using a 0 or 1, contained as an integer. There are no missing observations out of 9,902 entries.
- Dubious: This numerical variable indicated whether the review could be viewed as unreliable or dubious using 0.0 or 1.0, contained as a float. There are no missing observations out of 9,902 entries.
- Imdbid: The numerical variable gives the IMDb movie ID for each movie tested, contained as an integer. There are no missing observations out of 9,902 entries.
- Id: This numerical variable is an unique identifier for each entry, contained as an integer. There are no missing observations out of 9,902 entries.
- Submitterid: This numerical variable is another unique identifier for the person who submitted the Bechdel test rating, contained as an integer. There are no missing observations out of 9,902 entries.
- Date: This numerical variable details the dates and times that the reviews were added to the dataset, contained as a datetime variable. There are no missing observations out of 9,902 entries.
- Visible: This numerical variable indicated whether the movie has been approved with a 0 or 1, contained as an integer. For this dataset, all movies have been approved, meaning the return value will always be 1. There are no missing observations out of 9,902 entries.

	year	rating	dubious	imdbid	id	submitterid	visible
count	9902.000000	9902.000000	9603.000000	9.902000e+03	9902.000000	9902.000000	9902.0
mean	1996.405171	2.133205	0.090909	2.342054e+06	5283.208645	9801.553828	1.0
std	25.110394	1.100234	0.287495	6.164546e+07	3088.966247	6462.398855	0.0
min	1874.000000	0.000000	0.000000	1.000000e+00	1.000000	1.000000	1.0
25%	1988.000000	1.000000	0.000000	9.977825e+04	2583.250000	3865.500000	1.0
50%	2006.000000	3.000000	0.000000	4.382925e+05	5265.500000	9626.500000	1.0
75%	2013.000000	3.000000	0.000000	2.035422e+06	7891.500000	15388.500000	1.0
max	2023.000000	3.000000	1.000000	6.129999e+09	10785.000000	20969.000000	1.0

Section 3: bechdel_movies_combined.csv

Unit of Observation:

Each row in this dataset contains information on a movie that is assessed by the Bechdel test.

Variables:

- Title: This categorical variable is the title of the movie, contained as a string. There are no missing observations out of 10,357 entries.
- Year: This numerical variable is the year the movie was released, contained as an integer. There are no missing observations out of 10,357 entries.
- Rating: This numerical variable indicates whether the movie passed or failed the Bechdel test using a 0 or 1, contained as an integer. There are no missing observations out of 10,357 entries.
- Dubious: This numerical variable indicated whether the review could be viewed as unreliable or dubious using 0.0 or 1.0, contained as a float. There are no missing observations out of 10,357 entries.
- Imdbid: The numerical variable gives the IMDb movie ID for each movie tested, contained as an integer. There are no missing observations out of 10,357 entries.
- Id: This numerical variable is an unique identifier for each entry, contained as an integer. There are no missing observations out of 10,357 entries.
- Submitterid: This numerical variable is another unique identifier for the person who submitted the Bechdel test rating, contained as an integer. There are no missing observations out of 10,357 entries.
- Date: This numerical variable details the dates and times that the reviews were added to the dataset, contained as a datetime variable. There are no missing observations out of 10,357 entries.
- Visible: This numerical variable indicated whether the movie has been approved with a 0 or 1, contained as an integer. For this dataset, all movies have been approved, meaning the return value will always be 1. There are no missing observations out of 10,357 entries.

	year	rating	dubious	imdbid	id	submitterid	visible
count	10357.000000	10357.000000	10058.000000	1.035700e+04	10357.000000	10357.000000	10357.0
mean	1996.927489	2.132471	0.090873	2.606133e+06	5538.922661	10318.339094	1.0
std	25.079985	1.099714	0.287443	6.031473e+07	3247.689426	6763.568483	0.0
min	1874.000000	0.000000	0.000000	1.000000e+00	1.000000	1.000000	1.0
25%	1989.000000	1.000000	0.000000	1.005190e+05	2700.000000	4063.000000	1.0
50%	2006.000000	3.000000	0.000000	4.525980e+05	5497.000000	10187.000000	1.0
75%	2014.000000	3.000000	0.000000	2.199543e+06	8363.000000	16373.000000	1.0
max	2024.000000	3.000000	1.000000	6.129999e+09	11429.000000	22119.000000	1.0

Section 4: bechdel_movies_with_sentiment.csv

Unit of Observation:

Each row in this dataset contains detailed information on a movie that is assessed by the

Bechdel test after a sentiment analysis was performed.

Variables:

- Title: This categorical variable is the title of the movie, contained as a string. There are no missing observations out of 10,357 entries.
- Year: This numerical variable is the year the movie was released, contained as an integer. There are no missing observations out of 10,357 entries.
- Rating: This numerical variable indicates whether the movie passed or failed the Bechdel test using a 0 or 1, contained as an integer. There are no missing observations out of 10,357 entries.
- Dubious: This numerical variable indicated whether the review could be viewed as unreliable or dubious using 0.0 or 1.0, contained as a float. There are no missing observations out of 10,357 entries.
- Imdbid: The numerical variable gives the IMDb movie ID for each movie tested, contained as an integer. There are no missing observations out of 10,357 entries.
- Id: This numerical variable is an unique identifier for each entry, contained as an integer. There are no missing observations out of 10,357 entries.
- Submitterid: This numerical variable is another unique identifier for the person who submitted the Bechdel test rating, contained as an integer. There are no missing observations out of 10,357 entries.
- Date: This numerical variable details the dates and times that the reviews were added to the dataset, contained as a datetime variable. There are no missing observations out of 10,357 entries.
- Visible: This numerical variable indicated whether the movie has been approved with a 0 or 1, contained as an integer. For this dataset, all movies have been approved, meaning the return value will always be 1. There are no missing observations out of 10,357 entries.
- Sentiment: This numerical variable indicates the sentiment score for the reviews after they have been scraped from the IMDb website from -1 to 1, contained as a float. There are no missing observations out of 10,357 entries.

	year	rating	dubious	imdbid	id	submitterid	visible	sentiment
count	10357.000000	10357.000000	10058.000000	1.035700e+04	10357.000000	10357.000000	10357.0	10216.000000
mean	1996.927489	2.132471	0.090873	2.606133e+06	5538.922661	10318.339094	1.0	0.521269
std	25.079985	1.099714	0.287443	6.031473e+07	3247.689426	6763.568483	0.0	0.326516
min	1874.000000	0.000000	0.000000	1.000000e+00	1.000000	1.000000	1.0	-0.991800
25%	1989.000000	1.000000	0.000000	1.005190e+05	2700.000000	4063.000000	1.0	0.335320
50%	2006.000000	3.000000	0.000000	4.525980e+05	5497.000000	10187.000000	1.0	0.594566
75%	2014.000000	3.000000	0.000000	2.199543e+06	8363.000000	16373.000000	1.0	0.771213
max	2024.000000	3.000000	1.000000	6.129999e+09	11429.000000	22119.000000	1.0	0.999300

Section 5: new_movies.csv

Unit of Observation:

Each row in this dataset contains information on a movie that is assessed by the Bechdel test.

Variables:

- Title: This categorical variable is the title of the movie, contained as a string. There are no missing observations out of 455 entries.
- Year: This numerical variable is the year the movie was released, contained as an integer. There are no missing observations out of 455 entries.
- Rating: This numerical variable indicates whether the movie passed or failed the Bechdel test using a 0 or 1, contained as an integer. There are no missing observations out of 455 entries.
- Dubious: This numerical variable indicated whether the review could be viewed as unreliable or dubious using 0.0 or 1.0, contained as a float. There are no missing observations out of 455 entries.
- Imdbid: The numerical variable gives the IMDb movie ID for each movie tested, contained as an integer. There are no missing observations out of 455 entries.
- Id: This numerical variable is an unique identifier for each entry, contained as an integer. There are no missing observations out of 455 entries.
- Submitterid: This numerical variable is another unique identifier for the person who submitted the Bechdel test rating, contained as an integer. There are no missing observations out of 455 entries.
- Date: This numerical variable details the dates and times that the reviews were added to the dataset, contained as a datetime variable. There are no missing observations out of 455 entries.
- Visible: This numerical variable indicated whether the movie has been approved with a 0 or 1, contained as an integer. For this dataset, all movies have been approved, meaning the return value will always be 1. There are no missing observations out of 455 entries.

	year	imdbid	rating	id	dubious	submitterid	visible
count	455.000000	4.550000e+02	455.000000	455.000000	455.000000	455.000000	455.0
mean	2008.294505	8.353181e+06	2.116484	11103.934066	0.090110	21564.949451	1.0
std	21.486321	8.475479e+06	1.089423	203.128245	0.286654	353.960767	0.0
min	1915.000000	4.134000e+03	0.000000	10786.000000	0.000000	20972.000000	1.0
25%	1999.000000	2.210895e+05	1.000000	10917.500000	0.000000	21221.500000	1.0
50%	2020.000000	6.263850e+06	3.000000	11143.000000	0.000000	21640.000000	1.0
75%	2023.000000	1.423042e+07	3.000000	11289.500000	0.000000	21880.500000	1.0
max	2024.000000	3.253778e+07	3.000000	11429.000000	1.000000	22119.000000	1.0