First, a webpage request was made to the IMDB website   
(URL: <https://www.imdb.com/search/title/?title_type=feature>).

As the request was confirmed, and the web page was successfully fetched, then HTML content had to be parsed. For parsing the HTML content from the webpage and scraping out necessary data, beautifulsoup python library was used. Using the beautifulsoup library, relevant movie details: movie title, release date, movie duration, critics rating, audience rating, total audience ratings, metascore rating, and movie description was extracted from consequent HTML tag sections.

Representative code:

soup = BeautifulSoup(pgrequest.text, 'html.parser')

movies = soup.find\_all("li", class\_="ipc-metadata-list-summary-item")

The scrapped movies data were stored in a pandas dataframe format for further computational analysis.

For the data cleaning part, first the dataframe was checked for NaN or None type values. It was done using the df.isna() function which consequently did also give null values which could have been checked using a separate df.isnull() function. There was only one NaN value in the entre table.

Then it was subjected to check for duplicate values out of which there were none. Then the NaN values were assigned value 0 since it fell in Metascore field and conincidentally not given any score in the original website. Several handlings of data had to be sorted out, notably changing the Metascore field from None typle attribute (got no idea why it was so!) to a numeric attribute. Also, brackets were removed. The string representation of units were converted to their respective appropriate numeric values.

**Optional part:** A customtokenizer custom function was defined to tokenize the movie description texts. Without utilizing more optimized and advanced techniques such as ‘nltk’ or ‘textblob’, punctuation were removed and individual words were split out from sentences to represent as tokens.

The data visualization consisted of box plots for movie duration and audience rating which showed average measures as well as variance from mean. To correlate different variables, scatter plot between numerical representative variables were plotted as well as correlation matrix heatmap were done.  
  
Finally, the final customized data were saved as a csv file.