**Data Analysis Example: IMDB Top 1000 Movies**

Before learning how to write Python code for data analysis, let's see how we can analyze a dataset without Python:

Example data source: imdb\_top\_1000.csv (438KB)

Source: <https://www.kaggle.com/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>

Open the file in Excel. Save it as an Excel document.

Try to answer the following questions. For each question, type your answer or include a screenshot of your result:

1. Getting an overall idea about the data set:
   1. What is the total number of rows? 1000 rows
   2. What is the total number of columns? 16 columns
   3. What are the data types of columns? 5 integer, 11 character or text
   4. How can we statistically describe each column? Could look at measures of central tendency for integer rows, and or sort by ascending or descending to look at top movies or bottom movies.
2. Filtering by rows:
   1. Select movies ranked 11~20. A screenshot of a computer screen

      Description automatically generated
   2. Select all movies directed by Quentin Tarantino. A screenshot of a computer

      Description automatically generated
   3. Select all movies released after (>=) 2010  and with IMDB\_Rating>=8.0A screenshot of a computer

      Description automatically generated
   4. More challenging tasks:
      1. Movie genres contain "Comedy"- Used the contains filter term. A screenshot of a computer

         Description automatically generated
      2. Starring: Al PacinoA screenshot of a computer

         Description automatically generated
3. Filter by columns:
   1. Create a new sheet called "movie\_stars" including the following columns: Series\_Title, Released\_Year, Star1, Star2, Star3, Star4
   2. Create a new sheet called "movie\_genres" including the following columns: Series\_Title, Released\_Year, Genre.
4. Sorting:
   1. Sort all movies in ascending order of "Released Year"A screen shot of a computer

      Description automatically generated
   2. Sort all movies in descending order of "Gross"A screenshot of a computer

      Description automatically generated
5. Add a new column:
   1. Add a new column "Runtime\_min" by removing the string ' min" in "Runtime"A screen shot of a computer

      Description automatically generated
   2. Add a new column "Age\_Year" by the expression: [current year (e.g., 2022)] - Released\_YearA screenshot of a computer

      Description automatically generated
6. Aggregation & summary:
   1. Total "Gross" of all top 1000 moviesA screen shot of a computer

      Description automatically generated
   2. Average "No\_of\_Votes" of all moviesA screenshot of a computer

      Description automatically generated
   3. Count movies in each decade (e.g., ..., 1980, 1990, 2000, 2010, 2020). I learned that the Quotient function in excel returns just the integer part of division. So I was able to divide every year by 10, and extract the decade then multiply by 10 to make it look like a year again. A screenshot of a graph

      Description automatically generated
   4. Average "IMDB\_Rating" of movies by different directors, ranked in descending order. I could do this in SQL or Tableau, but not excel (and/or) I cannot think a good way to group things in excel before aggregating them. But that is the skill that is being asked.
   5. Count movies by "Certificate" in different decades- Just a few of them, but could use COUNTIFS to do any certificate. A screenshot of a table

      Description automatically generated
   6. Total "Gross" of movies by starts. Once again the lack of ability or knowledge to group and aggregate is making things difficult.
   7. Count movies by genre (e.g., drama, crime, comedy) A screenshot of a computer

      Description automatically generated
7. Exploratory data analysis
   1. What type of **variation** occurs within my variables?
      1. Distribution of "Released\_year"A screenshot of a graph

         Description automatically generated
      2. Distribution of "Runtime\_min"A screenshot of a computer

         Description automatically generated
      3. Piechart of "Certificate" A screenshot of a computer

         Description automatically generated
   2. What type of **covariation** occurs between my variables?
      1. Gross vs. Released\_YearA screen shot of a graph

         Description automatically generated
      2. Gross vs. DecadesA screen shot of a graph

         Description automatically generated
      3. IMDB\_Rating vs. Director- I would probably want to create an average rating for each director before then displaying it as a bar chart.
      4. IMDB\_Rating vs. GrossA screenshot of a computer

         Description automatically generated