REBECA PRACTICE: DATA SCIENTIST SOLUTIONS

SOLUTION TO TASK 2

In [ ]:
```
# Before cleaning, I want to replace the space in columns names with underscores
movies_data.columns = movies_data.columns.str.replace(' ','_')
```

In [ ]:
```
# First I try to drop all observation with missing data
movies_data.dropna()
```

In [ ]:
```
# Then I try to drop all columns with missing data
movies_data.dropna(axis=1)
```

If I drop all observations with missing data, I am left with <10% of the data set, which I find is too small of a fraction.

If I drop all columns of the data set which contain missing data, I am only left with the release date. Which is useless.

I need to find a smarter way to filter the missing data out.

In [ ]:
```
# First I select only the numerical data type columns
# Which could be included in a model
movies_data_filtered = movies_data.drop(columns=movies_data.dtypes[
    movies_data.dtypes !=float].index)
movies_data_filtered
```

In [ ]:
```
!conda env list
```

In [ ]:
```
# Let's check where there are nans:
for coln in movies_data_filtered.columns:
    print(coln,'\n',movies_data_filtered.isna()[coln].value_counts(),'\n\n')
```

In [ ]:
```
# I decide to also drop the US_DVD_Sales, Running_Time_min, Rotten_Tomatoes_Rating
movies_data_filtered.drop(columns = ['US_DVD_Sales',
                                     'Running_Time_min',
                                     'Rotten_Tomatoes_Rating'
                                     ],
                          inplace = True)
```

In [ ]:
```
# Remove NaNs from this subset
movies_data_filtered.dropna(axis = 0, inplace=True)

# Check how many rows are left
movies_data_filtered
```