I started by duplicating the Excel sheet so that the original data would be unaffected. Then in the copied file I deleted the rows containing the top 100 to top 1000 films (leaving 100 films remaining), I also deleted the “movie info” column since this would not be needed for the analysis. I then set column A to have numbers from 1 to 100 instead and gave it a heading off “Rank”.

A screenshot of a computer

AI-generated content may be incorrect.

I then turned on column filters and started using this to further clean the data. Under Distributor I noticed that the film “Hi, Mom” had the distributor of “February 12, 2021 (China)” which is clearly wrong.

A screenshot of a computer

AI-generated content may be incorrect.

I found the correct worldwide Distributor and inputted the data into the spreadsheet. I did this for many of the budgets on this document too by finding their budgets on IMDB. I also changed all the currency values to have the number type of “Currency” instead of “General”

A screenshot of a computer

AI-generated content may be incorrect.

I then copied and pasted the “Genre” column underneath the main data, I then used “Find and Replace” to remove the “[“, “]” and “ ‘ ” (without the spaces) characters then used split to split each genre into its own column.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

I then used the “Text to Columns” wizard to split all the genres into their own columns, using commas as delimiters.

A screenshot of a computer

AI-generated content may be incorrect.

I set the output destination to the end of the main sheet’s columns and deleted the combined “Genre” column.

A screenshot of a computer

AI-generated content may be incorrect.

To remove any spaces contained within the Genres columns I used the TRIM formula underneath the data, then I copied and pasted the values from the trimmed data on top of the original genres columns.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a phone

AI-generated content may be incorrect.A screenshot of a phone

AI-generated content may be incorrect.Then to make the “Running Time” column more usable I used “find and replace” to change every instance of “ hr ” (including the spaces) to instead be “:" and replaced “ min” with “” (meaning no value). I then changed the running times values to have the type of “Time”.

**Before**

**After**

To better make use of the running time in column T, named “Running Time No Formatting”, I used “=TEXT(K2,”[h]mm)” to convert the first “Running Time” number to a three digit number that showed the second hour digit and both minute digits. I then copied and pasted the values of these numbers to remove their formulas and set their types to be “Number”.

A screenshot of a computer

AI-generated content may be incorrect.

Once the numbers above were created I used the following formula in column U, applied this formula to all the rows, then hid column T:

=IFS(T2>=300, "3:00 hours or over", AND(T2<300, T2>= 230), "Between 2:30 and 3 hours", AND(T2<230, T2>=200), "Between 2 and 2:30 hours", AND(T2<200, T2>=130), "Between 1:30 and 2 hours", T2<130, "Under 1:30 hours").

cvzA screen shot of a computer

AI-generated content may be incorrect.

I then created a bar graph from this data showing what the most common length are for the top grossing films. This shows that the around half of the top grossing films are between 2 and 2:30 hours in length.

A graph of a number of hours

AI-generated content may be incorrect.

At this point I began to noticed that some of this data had been scraped incorrectly. Fixing this dataset would involve either rescraping it after determining what issues caused the data errors or spending a large amount of time manually changing incorrect data.

Since this project is for demonstrative purposes only I will continue with the existing data, however this data would not ordinarily be seen as clean data that could be analysed.

The next chart I made was a line chart to show the worldwide sales of the films, to see how much larger the sales of the highest grossing films where compared to the lower grossing films in the list.

After making the chart I formatted the y axis (with the currency values), set the display units to be in “billions”, set the symbol to be $ English (United States), and set the deminal places to 1. I also moved the “Billions” text to be vertically centre aligned. This helps makes the currency easier to read and makes it more immediate to see that the numbers are in billions of dollars.

A screenshot of a computer

AI-generated content may be incorrect.

I then added a horizontal title and labeled it as “Worldwide Box Office Revenue”, and set the interval unit for the X axis to 10.

A graph showing a line

AI-generated content may be incorrect.

I then created a pivot table for the Distributors and removed the Grand Total box and created a “Count” column next to it and used the “CountIF” to show the count of the number of rows that matched each label in the pivot table.

A screen shot of a list

AI-generated content may be incorrect.

I used $ signs before the row numbers in the range to ensure the data selections didn’t change when the formula was dragged down to the display information for the other distributors. As example of one of the formulas is this: =COUNTIF(D$2:D$101, X6)

This graph shows which film distributors have the highest film count from this data. I then copied and pasted the data to the “Graphs” sheet, sorted it from smallest to largest number, and created a bar chart based on this information. I sorted it to create a graph where the company with the highest film count is at the top and lowest is at the bottom. I then added data labels so that it would be easy to see exactly how many films each company released.

A graph of a number of films

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

To get a list of all of the genres in this data I used the “UNIQUE” function to run through each genre column individually and pull out distinct genres from those columns, I also encased the array in the TRIM formula to remove any spaces. An example of this formula was: =TRIM(UNIQUE(M2:M101)).

A screen shot of a computer

AI-generated content may be incorrect.

I then took all the results and copied and pasted their values in a column one under the other and then ran “UNIQUE” again to get a final list of unique genres. I then copied and pasted the results to remove the formula, and sorted them in alphabetical order.

A screenshot of a computer

AI-generated content may be incorrect.

Above the top genre I added the text “Genres” as a heading and then typed in “Count” next to them. I then used “Countif” to count each occurrence of the genres across all the genre columns. I used $ signs before the row numbers and column letters in the range to ensure the data selections didn’t change when the formula was dragged down to the display information for the other distributors. An example of the formula is: =COUNTIF(M$2:S$101,M161)

A screenshot of a computer

AI-generated content may be incorrect.

I then created a third column for the percentages and divided each individual genre count by the sum of all of the counts, then changed the number type to percentage to get the percentages. I then used $ signs next to the row numbers to ensure that the data rows used stayed the same.

A screenshot of a graph

AI-generated content may be incorrect.

I then took these three columns displaying the genre counts and percentages, cut and pasted it into the “Graphs” sheet, and sorted the rows from smallest to largest based on the count.

A table with numbers and percentages

AI-generated content may be incorrect.

I then created a bar chart based on the count and a tree map for the percentage. Once the bar chart was created, I gave the bars a gap width of 10% to make them easier to distinguish.

A graph of a movie genre

AI-generated content may be incorrect.

I was initially going to use a pie chart but decided that this would be hard to understand since there wasn’t space for the text on the pie and there were too many genres to have a legible legend.

For the treemap I then removed the legend and included the data labels for the percentages as well as the genre title. I then used =SUM to calculate the sum of the percentages from crime to the genres that have a smaller percentage, then I manually drew on a grey box and gave it the data label of Others and added the percentage of 10%. I did this since the data labels are not easily legible for the “other” genres. I also gave the chart the title of “Percentage Of Genre Listings For Top 100 Films”.

A chart of a variety of colors

AI-generated content may be incorrect.

In preparation for the final graph, I created a new column on the main sheet and gave it the heading of “Decade”. I then used the following “IFS” statement to work out the decades of the films: =IFS(C2<2000,"1990s",C2<2010,"2000s", C2<2020, "2010s", C2<2030, "2020s"), the cell number changed per row to match the film year on the same row. By adding a filter to the year column, I could easily see that the oldest film was the from 1990s and newest from the 2020s so I created decades for values within this range.

A screenshot of a computer

AI-generated content may be incorrect.

I then used the UNIQUE formula underneath the years to pick out the individual decade names. I then copied and pasted the decades as values to remove the formula and then moved “1990s” to the top of the 4 decades. Above then I wrote the heading “Individual Decades” and then gave the column next to them the heading of “Count”. I then used COUNTIF to count all the instances of each decade under the main “Decades” column.

A screenshot of a computer

AI-generated content may be incorrect.

I then turned this small table into a pie chart.

A pie chart with numbers and text

AI-generated content may be incorrect.

I chose pie chart style that contained the percentages within the pie slices and changed the pie chart colour to make the colours easier to distinguish from one-another.