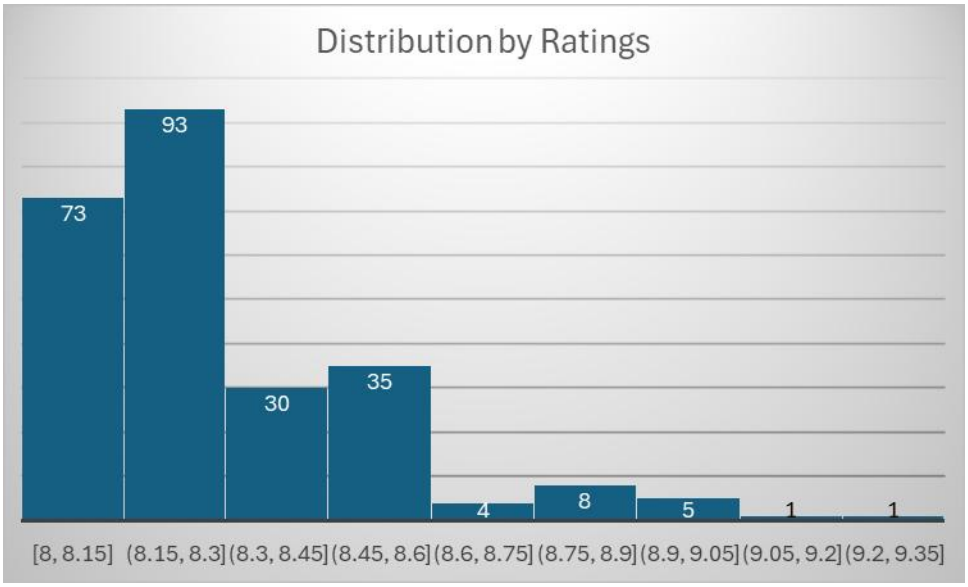
IMDb's Top 250 Movies of All Time and The Statistics Behind It by Miguel Reyes

The focus of this report, IMDb's Top 250 Movies up to the date February 10, 2023, covers numerous movies, specifically 250, and the statistics attached to each one. Whether it's regarding the type of movie or how much was made from each movie compared to what the budget was, there are numerous interesting tidbits concerning the movies we as a whole view to be the best. The dataset ascends or descends based on the rating of each movie, with the highest ones being at the top and descending the lower the rating is. Prior to drawing interesting conclusions regarding this dataset, we must take the time to comprehend what all this data means and how frequently each one shows up. Moving forward, we will be going over the foundation of this dataset and begin organizing the information in a coherent way.

To better understand the movie ratings, act as a movie researcher who is curious whether some decades produce significantly highly rated movies more than others. Using the IMDb Top 250 dataset, which includes ratings and release years, the goal is to investigate the following and answer these questions:

a) Construct a relative frequency histogram of IMDb ratings (e.g., use intervals like 8.0-8.2, 8.2-8.4, etc.) When making the histograms, make use of bins in a way that includes the outliers as well such as the highly rated movies.
b) One movie, *The Shawshank Redemption*, has a rating of 9.3. Does the movie's critical acclaim and legacy justify its exceptional rating, or is it just an outlier in this dataset?
c) The average ratings of movies from this dataset lie at about 8.3, after going through this dataset. What percentage of all movies in the dataset have ratings that are greater than the average?
d) Based on your findings, would you say the 1990s were an unusually strong decade for highly rated films?



Examining this histogram of the distribution of IMDb ratings for the Top 250 Movies of All Time, we can see there are two outliers (The Shawshank Redemption and The Godfather). One question asks if *The Shawshank Redemption* has a rating that fits its critical acclaim and legacy or if it is just an outlier in the dataset. The answer to that is the former since the outlier is not too far off from the dataset that makes it an anomaly. If the rating for *The Shawshank Redemption* were to be outside of the 9.2-9.35 bin, then it can be seen as an outlier due to how far it is from the typical ratings. Along with that, we can see that the average rating for movies is about 8.3 (specifically 8.3072), however we are curious to see how many movies go above the average. 82 movies go above the rating of 8.3 out of the 250, so doing simple math: $\frac{82}{250} = 0.328 = 32.8\%$

This shows most movies in the Top 250 lie below the average, specifically 67.2% of them. We can draw the conclusion that there is a great deal of difficulty in acquiring a rating above 8.3 when it comes to movies, specifically for IMDb's critics. Along with most movies lying below the average rating, we are curious as to see where most of the movies in this list can be found. It was said that the 1990s was a great time for film and media, so we shall use this time period to see if this is true. Considering 38 out of the 250 films were on this list, an argument can be made that this was the best decade for films. However, due to the 2000s having 42 films on this list, a stronger argument can be made that the 2000s was the best decade for having unusually high ratings. There is still an argument to be made that the 1990s was better since 8 of the 39 lies at the top 20 whereas 4 of the 42 films for the 2000s lies in the top 20. The question asks about being "unusually strong" which ties into the highest ratings, therefore the 1990s can be seen as an unusually strong decade for film considering it takes up 40% of the top 20 and has 38 films on the overall list.

With the focus currently lying on familiarizing ourselves with this dataset, the movie runtimes are an important factor due to a good chunk of time being dedicated to some movies, such as *The Godfather*, over other movies with ratings similar or close to it having way less time, such as *12 Angry Men*. The IMDb Top 250 movie runtimes (in minutes) are approximately normally distributed with a mean of 129.13 and a standard deviation of 30.13.

a) What proportion of movies are between 110 and 150 minutes?
b) What proportion of movies are longer than 170 minutes?
c) What fraction of movies are either shorter than 90 minutes or longer than 170 minutes?

In order to find these proportions, we use standard deviation, mean, and Z-scores for an accurate answer. First, we will be looking at what proportion of movies are between 110 and 150 minutes. We will use the following formula to find these Z-scores: $\boldsymbol{Z} = \frac{\boldsymbol{X-\mu}}{\boldsymbol{\sigma}}$. The X represents the number of focus (110 & 150 here), the lowercase mu symbol represents the mean, and lowercase sigma represents the standard deviation. After inserting the values we have for this equation, we got a Z-score of 110 equaling -0.63 and a Z-score of 150 equaling 0.69. We then look at a Z-table (negative and positive) to further proceed with getting the proportions:

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
+2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
+2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
+2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
+2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
+2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
+2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
+2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
+3	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
+3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
+3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
+3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
+3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
+3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
+3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
+3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
+3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
+3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
+4	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998

For -0.63 we get the value 0.2644 and for 0.69 we get the value 0.7549, both rounded to four decimal places. After getting both values, we subtract the negative table one from the positive table value and end up with 0.4905. This means that the proportion of movies that are between 110 and 150 minutes is 0.4905 of the movies in the Top 250 IMDb Movies list. Continuing onto the other two problems, we use the same formula since this problem set deals with proportions. After inserting 170 into the equation in order to find the proportion of movies longer than 170 minutes, we get 1.40 as a Z-score and end up with 0.9192, also rounded to 4 decimal places (we will be doing the same for the next problem). Due to only one value being considered and finding the proportion above this range, we take the equation 1 – 0.9192 in order to get the proportion 0.0808. This means that the proportion of movies longer than 170 minutes is 0.0808. The following problem is a two-tail one, meaning both sides end up getting added due to wanting to see the proportions from both ends. We already inserted 170 into the previous question so we have that value. However, inserting 90 into the Z equation, we get the value -1.34 which correlates to the value 0.0901 on the Z-table. Since we are getting the fraction of movies that are either shorter than 90 or longer than 170, we add both values up and get 0.1709. This means that the fraction of movies that are either shorter than 90 minutes or longer than 170 minutes is 0.1709 of the Top 250 Movies According to IMDb.

After gathering information on two important pieces of data (movie ratings and runtime), we can see the data varies in different areas, especially runtimes holding different values on a lot of occasions. With research done on those pieces of data, is there any curiosity toward the other pieces of data? I believe so, especially regarding the year movies were released and statistics on them by itself without any correlation to other data like ratings. I am curious to see how heavily the director plays into a movie being in the Top 250, especially since critically acclaimed directors have a good sum of their movies in this list. One specific director that comes to mind is Christoper Nolan, who produced pretty recent cult-classics such as *The Dark Knight (2008)*. Another interesting piece of data is the year the movies were released, with a lot of classics tied to before 2010.

Let C = All movies directed by Christoper Nolan and D = All movies released after 2010.

a) What does the set $C \cap D$ represent in this context?
b) What does $C \cup D$ represent in this context?
c) Which movies are in D | C?

Looking at these notations and variables, we can see that C and D, which is what $C \cap D$ mean in English form, means the movies that were directed by Christopher Nolan and were released after 2010 in the Top 250 list. Along with that, $C \cup D$, which means C or D, represents the movies that were either made by Christoper Nolan or were released after 2010 instead of looking at both as one value. After looking at and using the COUNTIF function in Excel, we can see that there are 43 movies that were released after 2010, and we can also see that 7 of the movies on this list were directed by Christopher Nolan. For A, we can see that this equals 2, meaning that 2 movies were released after 2010 and were directed by Christoper Nolan, and it's interesting to note that 3 out of the 7 movies were apart of the Batman series. For B, we can use the equation 43 + 7 – 2 since 2 of his movies were made after 2010 and already directed by him. Therefore, we get 48 movies that were directed by Christopher Nolan and were made after the year 2010 in the Top 250 IMDb Movies list. Now, if we look at more specific instances, like only movies produced after 2010, we get 2 out of the 43 movies that were directed by Christopher Nolan, given that they were released after 2010. This means that Christopher Nolan has made 4.65% of the movies in the Top 250 IMDb Movies list that were released after 2010, making him not as successful as a director compared to his earlier stuff.

Moving on from specific instance regarding release dates and directors, a conclusion that can be made on what makes a great movie is the rating it has regarding what audience it is intended for. These ratings are classified as "certificate" in the dataset, however the actual name tied to these numbers are MPAA ratings. Suppose we randomly select one movie from the IMDb Top 250 list. Let the experiment be defined as observing the MPAA rating of the selected movie. From the dataset, the most common certifications and their approximate frequencies are as follows:

PG: 37

PG-13: 35

R: 97

G: 19

Approved: 14

Not Rated: 24

a) What is the probability that a randomly selected movie is Rated R or PG-13 out of the most common certifications?
b) What is the probability that the selected movie is not rated R out of the most common certifications?
c) Define the sample space and assign the appropriate probabilities to each outcome when it comes to most common certifications
d) Are the sample events equally likely? Why or why not?

The MPAA ratings may have a correlation to what makes a great movie, after all there are some movies that have a high quantity over others. According to Stephen Follows, "41% of releases (of the top 100 films of the past 20 years beginning from 2014) are PG-13 but they make up 47% of the box office". This shows a rating that is clearly being used a lot more than the others and appears more frequently when it comes to big draws. We can draw more conclusions on this after answering the questions presented to us. 226 movies hold the most common certifications, and we are wondering what the odds of a random one being Rated R or PG-13 out of them. There are 97 Rated R movies and 35 Rated PG-13 movies, so that means there are 132 movies that are either Rated R or PG-13. We would have to take the following fraction in order to draw our conclusion: $\frac{132}{226}$. This fraction equals about 0.5841, meaning that there is a 58.41% chance that a randomly selected movie is Rated R or PG-13 out of the most common ratings. If we were to look at movies that are specifically not Rated R, we can take the equation 226 – 97 to find how many movies are not Rated R and we end up with the fraction $\frac{129}{226}$. This fraction equals about 0.5708, meaning that there is a 57.08% chance that the randomly selected movie is not Rated R out of the most common ratings. In order to keep track of this information, we have to define a sample space to understand what our option is for when looking at the most common certifications. We will be using the notation R_n with "R" equaling all the possible ratings and "n" equaling the specific rating.

$\{\boldsymbol{R_1, R_2, R_3, R_4, R_5, R_6}\}$ ***is*** the sample space for the problem with the following probabilities being attached to each one: R_1 (Rated PG) = 16.37%, R_2 (Rated PG-13) = 15.49%, R_3 (Rated R) = 42.92%, R_4 (Rated G) = 8.41%, R_5 (Rated Approved) = 6.19%, & R_6 (Rated Not Rated) = 10.62%. Reviewing the probabilities and the sample space, the sample events are not likely since each sample event holds different weight to it, with Rated R and Rated PG being the most common ones out of the common certifications. In contrast to the site that reviewed the top 100 films list on their MPAA ratings, PG-13 have a higher quantity of movies does not lead to it having more films in the Top 250 IMDb Movie list, making an argument against the summary Stephen Follows typed up.

The following research questions done on the dataset so far have covered the ratings of the movies, the MPAA ratings of the movies, the years it was released, the director of the films, and the runtime (in minutes) of the films. We covered a good portion of the dataset in different hypothetical situations but what if we were to do a real-life example and how the data would be reflected.

Let's say that three movies are randomly selected from the IMDb Top 250 list to be featured in a weekend film showcase. The selected films will be shown in a randomly assigned order: Friday, Saturday, and Sunday. Let's suppose the three movies chosen are:

- Blade Runner
- Pulp Fiction
- Taxi Driver

a) Define the experiment and describe a single sample point.
b) List the full sample space of all possible ordering in which these three movies can be shown.
c) Assuming all orderings are equally likely, what is the probability that *Taxi Driver* will be shown on Sunday?
d) What is the probability that *Blade Runner* is shown before *Pulp Fiction*?

The experiment presented to use is taking three randomly selected specific movies (Blade Runner, Pulp Fiction, Taxi Driver) and assigning each one to a unique day in a weekend showcase (Friday, Saturday, Sunday). We will be using the following notations to signify each possible movie: **Blade Runner = M_1, Pulp Fiction = M_2, & Taxi Driver = M_3.** A sample point example for the following sample set would be $M_2M_1M_3$. The sample points will always follow the order of the weekend, which is Friday, Saturday, and Sunday. To list all sample points, permutation is used to find how many total sample points there are, which is the formula $(n!)$ with n = 3. That means we are looking for six sample points. The following is the full sample space S:

S = $\{M_1M_2M_3, M_1M_3M_2, M_2M_1M_3, M_2M_3M_1, M_3M_1M_2, M_3M_2M_1\}$.

With the sample space written up, we draw conclusions on how the movies will be presented regarding the weekend order. An example problem would be what is the probability that *Taxi Driver* will be shown on Sunday. Looking at the sample set, we can see that there is a 33.3% chance of this movie being shown on Sunday due to looking at the notation that was assigned to it. We can also see that the probability that *Blade Runner* is shown before *Pulp Fiction* is a 50% chance due to the way the notations are set up in the sample set.

Focusing on more real-life situations, we can look at this from different perspectives such as a streaming service exploring something similar. Instead of a weekend showcase, a streaming service has decided to plan a feature of a five-movie spotlight on their homepage selected from the IMDb Top 250 list. The movies will be displayed in a specific left-to-right order in a row of 5 tiles.

a) If the service selects 5 distinct movies to display from the list, how many different ordered arrangements are possible?
b) Suppose *The Shawshank Redemption* and *The Godfather* are both selected and must be placed next to each other (in any order). How many such arrangements of the 5 featured movies being shown are possible?
c) If *The Dark Knight* must appear first, how many spotlight arrangements are possible?

For this problem, there are movies being selected out of the Top 250 List with 5 distinct ones being selected where order matters. Since we are choosing and arranging 5 out of 250 movies, we use the following formula (P represents the permutation formula):

$\mathbf{P(250,5)} = \frac{\mathbf{250!}}{\mathbf{(250-5)!}} = \mathbf{250 \times 249 \times 248 \times 247 \times 246}$. The answer for this is 938,043,756,000, meaning that there are 938,043,756,000 possible 5-movie lineups where order matters when picking from the Top 250 IMDb Movie list. If we were to look at two specific movies and making sure they are next to each other, specifically *The Shawshank Redemption* and *The Godfather*, we would have to treat these two as a block with two different appearances allowed being: [Shawshank, Godfather] & [Godfather, Shawshank]. This would leave us to choose 3 more movies from the remainder of the list and arrange them in an ordered manner. This would lead us to use the following formula to pick the last three movies:

$$\mathbf{P(248,3)} = \frac{\mathbf{248!}}{\mathbf{(248-3)!}} = \mathbf{248 \times 247 \times 246 = 15,068,976.}$$

With the block representing the two movies being put together, we have to order these "4 movies" using the formula (**4!)**. Also, since there are only two ways to organize the block, we just use the number 2 in the following formula: 15,068,976 * 24 * 2 (4! = 24). This will equate to 723,310,848 different arrangements of movies using the Top 250 if we have to put two specific movies next to each other. If we wanted a specific movie to be first (*The Dark Knight)*, we would employ the formula **P(249,4) * 4!** due to one movie already being accounted for and there are 24 separate ways to organize the 4 distinct movies. This would equate to 90,052,200,576 different possible combinations of showing the movies if they were to select them from the Top 250 list and have *The Dark Knight* as the headliner. With values as large as this, this shows how it is near impossible to run out of combinations when it comes to watching movies on this list in specific orders. It's not near impossible but I don't see anyone pulling off a stunt like this ever in their lifetime.

Now that we have seen different ways we can view the movies on this list, we can circle back to the ratings in order to ensure we are dedicating our time to the highest quality of film instead of randomizing it with tons of combinations. However, what if we want to compare that to the age of film that was before the beginning of a new millennium? Out of the 250 IMDb Top movies:

Using the COUNTIF function in Excel, we see 15 movies have a rating of 8.8 or higher, 154 movies were made before 2000, and 9 of those movies have a rating of 8.8 or higher as well as being made before 2000. Let A = movie with a rating of 8.8 or higher and Let B = movie was released before 2000.

a) What is the probability that a randomly selected movie has both a rating of 8.8 or higher and was released before 2000, $P(A \cap B)$?
b) What is the conditional probability that a movie was released before 2000 given that it has a rating of 8.8 or higher, $P(B|A)$?
c) Are events A and B independent? Justify your answer using probabilities.

In order to find out the probability of a randomly selected movie having both a rating of 8.8 or higher and being released before 2000, we use the formula $P(A \cap B)$ which translates to taking the fraction $\frac{9}{250}$ due to already having the value for this combination of events. This equates to 0.036, meaning there is a 3.6% chance that a randomly selected movie contains both of these qualities. As for finding the conditional property, we use the formula and equations $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{9/250}{15/250} = \frac{9}{15} = 0.6$. This means that there is 60% chance that a movie was released before 2000 give that it has a rating of 8.8 or higher. When looking at the events we have, we tend to wonder if they affect one another or chart their own paths. We can check if they are independent of each other by using the following problem: $P(A \cap B) = P(A) \times P(B)$. Multiplying these two events leads us to having the value of 0.03696. The value for $P(A \cap B)$ = 0.036 and even though that values are remarkably close, rounding up shows us that they are not equal, so A and B are not independent.

A majority of this report focuses on what we can do with ratings and the correlation it has to the other pieces of data in this dataset. Without ratings, it is difficult to separate the "filler" media from the ones with meaningful and deep stories. Otherwise, we would all constantly watch remakes of the Disney animated classics. With the continued focus on ratings, we also take a look back at the release years for movies. Let A = the event that a movie is from the 1990s (There are 42 such movies that fall in the 1990-1999 range). Let B = the event that a movie has a rating above 8.8 (There are 15 such movies). Bear in mind that there are also 250 movies in total on this list.

a) What is the smallest possible value of $P(A \cap B)$?
b) What is the largest possible value of $P(A \cap B)$?
c) Based on these answers, explain whether it is possible that *all* highly rated movies (higher than 8.8) are from the 1990s.

Before we move onto answering these questions, we have to find the probability of each event. Since we worked with ratings higher than 8.8 in the previous question, we know that P(B) = 0.06. For P(A), we take the fraction $\frac{42}{250}$ and get the probability 0.168. Moving on, to find the answers and statistics tied to the main questions, we use the Additive Law of Probability: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. However, since $P(A \cup B) \leq 1$, we get: $P(A \cap B) \geq P(A) + P(B) - 1 = 0.168 + 0.06 - 1$ = -0.826. However, since we can't have negative probabilities, we get the answer for the smallest possible value of $\boldsymbol{P(A \cap B) = 0}$. This is after employing the max formula which is max(0, -0.826). For finding the largest possible value, we take the problem $P(A \cap B)_{max} = min(P(A), P(B)) = min(0.168, 0.06) = 0.06$. This indicates that the largest possible value for $\boldsymbol{P(A \cap B) = 0.06}$. An interesting hypothetical was posed in this question regarding *all* highly rated movies being from the 1990s. Even though we have the list at our disposal, we will be going about this in a different way. If all 15 of the highly rated movies are from the 1990s, then $P(B) = P(A \cap B)$. Then, this would lead to $P(A \cap B) = P(B) = 0.06$. After going through the possible values for $P(A \cap B)$, we know that the maximum possible value for $P(A \cap B)$ is 0.06. So yes, it is possible that all highly rated movies are from

the 1990s. However, it would also require all 15 movies to be part of the 42 movies from the 1990s. By that, we mean $15 \leq 42$ which is True. Without checking the list, it is highly possible.

The nature of this report must have you wanting to become a movie expert at this point. Let's put ourselves in another hypothetical situation regarding these movies and let's use five randomly selected movies again but this time you play the role of a movie critic. As a movie critic, you plan to watch five randomly selected movies from the IMDb Top 250 list over the weekend. From prior experiences, the critic knows that there is a 90% chance they will enjoy any given movie, independent of the others.

a) What is the probability that the critic enjoys all five movies?
b) What is the probability that the critic enjoys at least one of the five movies?
c) What is the probability that the critic enjoys exactly three of the five movies?

The following problems are at times easy to follow and only make use of one value presented in the problem format. However, there are problems that make use of the Binomial Probability Distribution to understand single instances and the probability of their success. Before getting to the problem and their answers, the formula for Binomial Probability Distribution (for future reference use): $\boldsymbol{p(y) = \binom{n}{y} p^y q^{n-y}}$**.** The "p" represents a success rate, which is mainly what we focus on when using this formula. The "q" represents the complement, which would be a failure rate. The "n" represents the total for the problem of focus, here it would be 5 since we're focused on 5 films. The "y" represents the specific instance, within the total, that we are looking to get a probability for. Also, for the purpose of the problem, note that the complement to the 90% chance the film is enjoyed by you is the odds of you disliking the film, which is 10% (0.10).

To find the probability that you as the critic enjoy all five movies, you take the odds of liking any given movie and get the answer via the following problem (answer shown afterward): $\boldsymbol{0.9^5 = 0.59049}$. Therefore, there is a 59.05% chance that you will be having yourself a grand time when watching all five of these films from the Top 250 list. However, in case you are expecting the worst and want to enjoy at least one film, the probability of this happening for you is found using the following formula (the solution is presented at the end):

$$\boldsymbol{P(enjoy\ at\ least\ one) = 1 - P(enjoy\ none) = 1 - (0.10)^5 = 1 - 0.00001 = 0.99999}.$$

Therefore, there is a 99.999% chance that your time will be bearable 20% of the time (well actually that is dependent on runtime, but that's a problem for another time). If we were to look at specific number of films you would enjoy as the critic, we would have to use binomial probability distribution to find this out. The following formula will be used:): $\boldsymbol{P(X = 3) = \binom{5}{3}(0.90)^3(0.10)^2}$**.** The Combination(5,3) equates to 10 whereas the other values for the problem equate to 0.729 and 0.01, respectively. When multiplying these numbers, we get 0.0729. Therefore, there is a 7.29% chance that you will enjoy 3 out of 5 films. Good thing that there is a higher chance of you enjoying all the films!

We have reached the point of the report where we can begin to tell what pieces of info have been important to understanding this dataset to its full extent. Do you know what pieces I am referring to? In case you don't, I am talking about the movie rating and the year it was released.

Elaborating on that, there are people who prefer highly rated movies and would like to only watch them, but year tends to get in the way since a film's format ties into the enjoyability an audience derives from it. By focusing on these two pieces, we see if that is actually true by putting these two pieces through different equations/problems to complete our understanding of the correlation between the two. Refocusing ourselves, we can see that among the Top 250 IMDb movies, 91 out of 250 were released after the year 2000, which would be 36.4% of the list. Suppose that in a hypothetical scenario where we don't refer to the list other than finding how many movies were released after 2000 that 70% of the post-2000 movies have an IMDb rating above 8.5, while only 50% of pre-2000 movies have a rating above 8.5. Let event A be "a randomly selected Top 250 movie with a rating above 8.5", and let B be the event "the movie was released after 2000." (The purpose of the problem is to train ourselves to not focus on the data in front of us and gain the ability to work with numbers and values we are asked to experiment with.)

a) What is the probability that a randomly selected Top 250 movie has a rating above 8.5?

b) Given that a movie has a rating above 8.5, what is the probability that it was released after 2000?

For the following problems, we are making use of the Law of Total Probability, where the formula for that is as follows: $\boldsymbol{P(A) = \sum P(A \cup B_n)}$**.** The total number of movies that we have are 250, we have 91 post-2000 movies as well as 159 pre-2000 movies. The hypothetical values used here are 70% of post-2000 movies having a rating of 8.5 or higher whereas 50% of pre-2000 movies have a rating of 8.5 or higher. When making use of the Law of Total Probability, we get the following**:**

$$\boldsymbol{P(R > 8.5) = P(R > 8.5 \mid post - 2000) \times P(post - 2000) + P(R > 8.5 \mid pre - 2000) \times P(pre - 2000)}$$
$$\boldsymbol{= (0.70)(0.364) + (0.50)(0.636) = 0.2548 + 0.318 = 0.5728.}$$

This means that using the hypothetical values mixed with an actual count of the movies before or after 2000, we have a 57.28% of a randomly selected movie having a rating above an 8.5. We also make use of Bayes' Rule in this problem with the equation for that being: $\boldsymbol{P(B_j \mid A) = \frac{P(A \cap B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}}$**.** Using this formula, we get the following:

$$\boldsymbol{P(post - 2000 \mid R > 8.5) = \frac{P(R>8.5 \mid post-2000) \times P(post-2000)}{P(R>8.5)} = \frac{(0.70)(0.364)}{0.5728} = \frac{0.2548}{0.5728} \approx 0.4446}$$

This shows that there is a 44.46% chance that the movie was released post-2000 given a movie rating higher than 8.5. Using different values from what the list would say, we remain sharp in dealing with this data, especially since the purpose of the report is to comprehend the data fully. This is successfully done through practice and changing up the actual values attached to certain pieces of data.

With ratings being prevalent in most of the problems pertaining to the material in Chapter 2, we will be examining runtime to close off the questions pertaining to this chapter. Specifically, we will be looking at random sampling and probability mass function (PMF), which has

usefulness regarding finding the probability of discreet random variables that were assigned specific values. We would need something like this in cases we wanted to see the spread of certain pieces of data or get specific probabilities for possible values, improving the efficiency of possible further experimentation that may be done with these pieces of data outside of this report. For future reference, the notation we will be using to specify a specific random variable will be $P = (Y = N)$ where N equals the specific instance of variable occurring.

In the IMDb Top 250 Movies list, the runtimes of the films are distributed as follows:

- Short (under 100 minutes): 43
- Medium (100-150 minutes): 155
- Long (more than 150 minutes): 52

Let the random variable *Y* represent the category of a randomly selected movie's runtime:

- Y = 0 if the movie is Short
- Y = 1 if the movie is Medium
- Y = 2 if the movie is Long

a) Determine the PMF of Y
b) Calculate the expected value E[Y] of the random variable Y.

When conducting the work for PMF, we look at the number of movies that are classified as either short, medium, or long. First, we take the different instances by doing the following: $\frac{43}{250}, \frac{155}{250}$, & $\frac{52}{250}$, with each number tied to the respective Y variable starting from 0-2. The decimals for these values are 0.172, 0.62, and 0.208. All these decimals represent their own respective PMF, with medium length movies having the highest value and being more than the combination of Y = 0 & Y = 2. This shows how much more likely one of the Top 250 movies is to be of medium length. Calculating the expected value of E[Y], we do the following**:**

$$\boldsymbol{E[Y] = 0 \times P(Y = 0) + 1 \times P(Y = 1) + 2 \times P(Y = 2) = 0 \times 0.172 + 1 \times 0.62 + 2 \times 0.208 = 0 + 0.62 + 0.416 = 1.036.}$$

The interpretation of this is the expected value suggests that, on average, a randomly selected movie from the IMDb Top 250 list falls between the Medium and Long, more toward the medium side. We can conclude that if we are going for highly rated movies using IMDb, it will most likely be a medium length film.

As we can see, there are many things to learn about our dataset and what each variable can do as well as what values they hold. These different statistics help improve the efficiency and comprehensiveness of the dataset for future uses, when it comes to placing a value to variables, we are aware of. Then there are those variables that are harder to detect. What if we want to know more about the hard to see variables, or discrete variables, and their probability distributions? Why are probability distributions important to use again? It's basically a method to the randomness presented to us since, according to Geeks for Geeks, "It provides a way of

modeling the likelihood of each outcome in a random experiment." Continuing with the report, we will be looking at different ways of finding these likelihoods when it comes to random experimentation through different means. Let's focus on the overall probability distribution for discrete random variables before getting into specifics.

Building off our previous question, let the random variable Y represent the number of movies in the IMDb Top 250 list that fall into the following runtime categories:

- Y = 0: Short (less than 100 minutes)
- Y = 1: Medium (100-150 minutes)
- Y = 2: Long (greater than 150 minutes)

Using the IMDb Top 250 dataset:

a) Construct the probability distribution for Y
b) What is the probability that a randomly selected movie is medium length?
c) What is the probability that a randomly selected movie is either short or long?

Probability Distribution Table for Y

y	0	1	2
P(Y = y)	$\frac{43}{250}$	$\frac{155}{250}$	$\frac{52}{250}$

With the table above, we gain a clear picture of the previous numbers we worked with and the value it holds tied to each segment of runtime. Using the table, we can also easily answer any questions we have concerning a movie's runtime, such as the probability of getting a randomly selected movie of medium length which is 155 out of 250, or 62%. As for getting the likelihood of two events if it's randomly selected based on either or, such as randomly selecting a short or long movie, which equals 95 out of 250, or 38%.

In the prior problem, we see an example of the usefulness probability distribution holds and the means of displaying it. There are different ways to go about and display this information. Let's continue with our usage of runtimes and let the random variable **Y** represent the runtime category of a movie from the IMDb Top 250 list, defined as:

- Y = 0: Short (less than 100 minutes)
- Y = 1: Medium (100-150 minutes)
- Y = 2: Long (greater than 150 minutes)

Where we have the values of each one pertaining to the movies: 43, 155, and 52, respectively. Suppose that each movie had a certain profit to make off it such as:

- Short movies = $1
- Medium movies = $2
- Long movies = $4

Let P(Y=y) represent her proportion of movies in each category.

a) Construct the probability distribution table for Y.
b) Compute the expected value E(Y), interpreted as the average "payoff" from randomly selecting a movie.
c) Suppose it costs $2.30 to stream a randomly selected movie. What is your expected profit or loss per selection

Probability Distribution Table for Y

Category (Y)	Description	Count	Probability P(Y=y)	Payoff
0	Short (< 100 min)	43	0.172	$1
1	Medium (100-150)	155	0.62	$2
2	Long(> 150 min)	52	0.208	$4

With this problem, we take a real-life scenario based around the cost of operating a movie streaming service and measuring the best costs in order to make the most money. There is money to be made off these movies, but it depends on how the service goes about streaming these movies. If the service were to base their streaming around the Top 250 IMDb Movie list, there would be an expected "payoff" to calculate to see if the business is worth going through with. In order to get the expected "payoff," we use the formula:

$$\boldsymbol{E(Y) = \sum[Payoff(y) \times P(Y = y)].}$$

Using the values we have set aside for this table, we get the problem: $E(Y) = (1 \times 0.172) + (2 \times 0.62) + (4 \times 0.208) = 0.172 + 1.24 + 0.832 = 2.244$. This means that we have an expected "payoff" of $2.24 per movie, which you would expect to add up over time and be good profit, especially since there 250 movies to cycle through. However, what if the cost to stream these movies was $2.30? The expected profit would then be -0.056, meaning that 5.6 cents would be lost per movie if this were the cost of operating. A problem such as this shows the importance of probability distribution through a real-life business model since the majority of the report has dealt with the nitty gritty of statistics and not how this could be beneficial to us. Maybe if you decide to open your own movie streaming service, you have a formula to use to make cost calculation more efficient.

Prior to this section of the report, we briefly discussed Binomial Probability Distribution and now we shall go more into detail with it. The purpose of getting binomial probability distribution is to get the probability of success on a single trial by using the two outcomes possible: success or failure. All the trials are independent of each other which makes it easier to

find a value for each instance. We base our successes using the random variable of interest, or Y, which is the number of successes found during the trials, n. To summarize, if we are interested in a specific instance of a piece of data in the Top 250 list, we use this distribution if there are success and failure rates attached to them. Since those aren't present in this dataset, we will be going through hypothetical scenarios for not only this distribution, but for the others in this chapter as well.

A film critic claims, not you this time, that a well-respected director typically has an extremely high success rate when it comes to getting their movies into the IMDb Top 250 list. Suppose historical data shows that 63.63% of Christopher Nolan's movies make it into the Top 250. Assume Nolan has directed 11 movies to date (using the time of the dataset's creation; excluding his film *Oppenheimer* due to its release data and *Man of Steel* which he was only a producer on). Treat each movie as an independent trial with a probability of success p = 0.6363, where "success" is defined as the movie making it into this dataset at some point.

a) What is the probability that exactly 6 of his movies are in the Top 250?
b) What is the probability that at least 8 of his movies are in the Top 250?
c) What is the probability that fewer than 5 of his movies will make it into the Top 250?

Before moving onto math, we can understand that p = 0.6363, so this would end up with q = 0.3637. Along with that, n = 11 since that is the total number of movies we are basing this on, and the "y" is dependent on what the question is asking. If we want to know the exact probability that 6 of his movies are in the top 250, we use the following equation:

$$\boldsymbol{p(Y=6) = \binom{11}{6}(0.6363)^6(0.3637)^5}$$

After calculating this, we end up with the value 0.195132, which means that there is a 19.5132% chance that exactly 6 of Christopher Nolan's movies make it onto the Top 250. If we didn't want to make exact calculations and look at general questions such as that probability of at least 8 of his movies making it, we would then use the following equations:

$$\boldsymbol{P(Y \geq 8) = P(Y=8) + P(Y=9) + P(Y=10) + P(Y=11)}$$

$$\boldsymbol{\left[p(Y=8) = \binom{11}{8}(0.6363)^8(0.3637)^3\right] + \left[p(Y=9) = \binom{11}{9}(0.6363)^9(0.3637)^2\right] +}$$

$$\boldsymbol{[p(Y=10) = \binom{11}{10}(0.6363)^{10}(0.3637)^1] + [p(Y=11) = \binom{11}{11}(0.6363)^{11}(0.3637)^0]}$$

This equation would equate to 0.213309 + 0.124396 + 0.043527 + 0.006923 = 0.388155. This means that there is a 38.8155% chance that Christopher Nolan ends up having 8 of his movies in the Top 250 list. If we wanted to look at the complementary operations, which would be the opposite of "at least", we could do so by checking the probability that fewer than 5 of his movies will make it into the Top 250. We would have to use the following equation that's fitted to our binomial distribution property:

$$\boldsymbol{P(Y<5) = P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) + P(Y=4)}$$

$$\left[p(Y=0)=\binom{11}{0}(0.6363)^0(0.3637)^{11}\right]+\left[p(Y=1)=\binom{11}{1}(0.6363)^1(0.3637)^{10}\right]+$$

$$\left[p(Y=2)=\binom{11}{2}(0.6363)^2(0.3637)^9\right]+\left[p(Y=3)=\binom{11}{3}(0.6363)^3(0.3637)^8\right]+$$

$$\left[p(Y=4)=\binom{11}{4}(0.6363)^4(0.3637)^7\right]=$$

$$0.000015+0.000283+0.002480+0.013014+0.045537=0.061329$$

After going through the steps for this problem, we reach the value 0.061329, which means that there is a 6.1329% chance that less than 5 of Nolan's movies make it into the Top 250 IMDb Movie List, showing how much of a powerhouse every time he produces a film. With equations like this, we can also have another means of incorporating this into real-life scenarios, especially if we want to tie it back to the payoff for streaming movies with certain runtimes. If the quality of Christopher Nolan's films is that high and the expectation almost all the time is for it to succeed with excellence, the streaming service owner can make a good profit since Nolan's films are practically either medium or long length, with his shortest film being *Dunkirk* (not on the list) having a runtime of 107 minutes.

Other than binomial probability, which focuses on single trials regarding a piece of data's success rate, there are other probabilities to gather in order to apply them to real-life scenarios regarding this list and they are also used to better understand the content before us. Moving on, we also have geometric probability distribution, which makes use of the formula:

$$p(y)=q^{y-1}p, \qquad y=1,2,3,\ldots,0\leq p\leq 1$$

The purpose of this distribution is to gather the chances of achieving success in a series, rather than a singular focus, of independent trials. To summarize, it helps measure what success will be after going pass a given number of trials. The following problem will make use of this concept in regard to the runtime again, this time focusing on one specific length of film.

Suppose a streaming platform, potentially your hypothetical one, is randomly selecting movies from the IMDb Top 250 list to feature in a "Short Film Spotlight." From the dataset, 43 of the 250 movies are classified as short (less than 100 minutes). Assume the (or your) platform randomly selects and reviews one movie at a time without prior knowledge of runtime until they find a short movie. Let the random variable Y represent the number of movies the platform must review until it finds the first short movie.

a) What is the probability that the first short movie appears on the 5th movie selected?
b) What is the expected number of movies the platform will need to review before selecting the first short movie?
c) What is the probability that the platform needs to review more than 6 movies to find a short one?

Using the formula we had shown above, we will be determining the following with the values of the movie runtimes and their classifications based off the distribution tables we have set up for it. If you notice, we have "p" and "q" in the formula, but we don't have a clear rate set up yet. Due to short films being the focus, **p = 0.172** and **q = 0.828**. If we want to find the probability that the first short movie is the 5th movie selected, we do the following equation:

$$\boldsymbol{p(Y = 5) = (0.828)^{5-1} \times (0.172)}$$

After going through the steps for this equation, we end up with P(Y = 5) = 0.0808. The conclusion is that there is a 8.08% of the first short movie appearing on the 5th movie selected. As for the other questions, we are looking to get the expected number of movies the platform will need to go through before getting the first short movie. Using the expected value μ for a geometric distribution, which is:

$$\boldsymbol{E(Y) = \frac{1}{p} = \frac{1}{0.172} \approx 5.814}$$

This means that on average, about 5.814 movies need to be reviewed before finding a short one. Since you can't watch movies in that manner, 6 movies is the safest number to go with if you want to be certain about how many movies need to be run before the first short one. Continuing with predicting the movie count before finding our desired length, we wonder what is the probability that more than 6 movies are reviewed before finding that short one. We use:

$$\boldsymbol{P(Y > 6) = q^6 = (0.828)^6 \approx 0.3222}$$

This means that there is a 32.22% chance that more than 6 movies are reviewed before a short one is found. The benefit of doing this form of probability analysis is seeing what time has to be spent in order to reach a certain goal or intended goal, in this case it would be reaching a short film when going through the Top 250 list. This can also be useful when it comes to calculating the cost of your time if the business revolved around this were to be done. As you can see, a good amount of time is spent since medium-length movies, which is the closest to the short ones, is 100 minutes minimum. That means 500 minutes are spent before being close to reaching a short movie, an important statistic when it comes to factoring in your time.

Having gone through binomial and geometric probability distribution, we can see the usefulness each one holds from the real-life perspective. Moving on, we begin to work with negative binomial probability distribution, with the formula for that being:

$$\boldsymbol{p(y) = \binom{y-1}{r-1} p^r q^{y-r}, \quad y = r, r+1, r+2, \ldots, \quad 0 \leq p \leq 1}$$

The following probability distribution is primarily used when focusing on single instances in a certain number of trials. Primarily, the focus is on finding the specific success event while also looking at and taking into consideration the previous successes. The next problem will do so in a meaningful way that holds up in statistics and in real-life scenarios, perhaps relating to business activities. With this next problem, we see begin to gain understanding of another tool that can be used to comprehend the dataset in its entirety.

According to IMDb Top 250 movie list, approximately 2.8% of the movies are directed by Martin Scorsese. Suppose you randomly go through the Top 250 list in hopes of reaching his films, one movie at a time, and you specifically want to find 3 movies directed by him. Let the random variable **Y** represent the number of movies you must check to find 3 Scorsese-directed films.

a) What is the probability that the 3rd Scorsese movie appears on the 10th movie you check?
b) What is the probability that you will find the 3rd Scorsese-directed movie on or before the 10th movie?
c) Find the mean and variance of the number of movies that must be checked to find 3 Scorsese movies.

To start this off, we have to familiarize ourselves with the values being used. The following are the values of the variables used throughout the problem: p = 0.028, q = 0.972. The Y concerns the trial on which the 3rd success occurs. We have r = 3 since our focus in on getting 3 Scorsese-directed films. For the first problem, we use the following equation:

$$\boldsymbol{p(Y = 10) = \binom{9}{2}(0.028)^3(0.972)^7 \approx 0.000648}$$

From this equation, we can conclude that there is a 0.0648% chance that the 3rd Scorsese movie is found on the 10th movie check. If we wanted to see if time can be saved and not have to wait while taking this probability into consideration, you can check the chance of getting the 3rd film on and prior to the 10th film. We would have to add the values we get from this process as such:

$$\boldsymbol{p(Y \leq 10) = \sum_{y=3}^{10} P(Y = y)}$$

For this formula, you start from three since the equation concerns the three movies we have as the theme from this problem and ends at 10 since we that is the last trial we are considering when hoping for getting all three Scorsese-directed films. The values that we get when going through the values 3-10 are:

$$\boldsymbol{0.000022 + 0.000064 + 0.000124 + 0.000202 + 0.000294 + 0.000400 + 0.000518 + 0.000648 = 0.002272}$$

This would mean that there is a 0.2272% chance of finding the 3rd Scorsese-directed movie on or before the 10th movie. If we wanted to gather more statistics around the 3 movies other than negative binomial distribution-based ones, we could gather what the mean and variance of the number of movies that must be checked to find the three movies. The formulas for that are:

$$\textbf{Mean: } \mu = \boldsymbol{E(Y) = \frac{r}{p}}$$

$$\textbf{Variance: } \boldsymbol{\sigma^2 = V(Y) = \frac{r(1-p)}{p^2}}$$

We use these formulas so that we can better understand what values to expect when going through negative binomial distribution and improve our comprehension of the data being put through it. When putting the values we have set aside for these formulas, we end up with the

following: μ = 107.142857 and σ^2 = 3719.387755. The values appear to be a lot larger than usual since the odds of getting a Martin Scorsese film, let alone 3, is like finding a needle in the haystack. However, using him, we are able to see the usefulness negative binomial distribution holds and how we can go about employing this to real-life scenarios by posing real-to-life questions.

Looking at this report, it's noticeable that the genre of films has not been discussed that much. The next problem looks to shed light on the genres of these films, but mainly the primary genre it is identified by. Specifically, some movies can be identified with numerous genres but we will be focusing on the primary genre (which can be found in the 1st genre column when referring to the work done on the dataset) to get our information and base it off a real-life scenario to further improve our understanding of the material. Before continuing, we will be using hypergeometric distribution for this section, which is useful for getting calculations from finite populations without replacement where we are interested in the number of successes that can occur in the problem. The formula used for hypergeometric distribution is (the problem will be presented after the formula):

$$\boldsymbol{p(y)} = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$$

Out of the IMDb Top 250 movies, we have 68 that are primarily classified as Drama, while the remaining 188 are not primarily classified as Drama or is not classified as Drama at all. You decide to host a weekend of movies with your friend group and decide to randomly select 15 movies from the Top 250 list without replacement, with the focus on having this weekend filled with variety. Let the random variable **Y** represent the number of primarily drama films selected.

a) What is the probability that exactly 5 of the selected movies are primarily defined as drama?
b) What is the probability that no more than 3 of the selected movies are primarily drama?

In order to start this problem off, we need to understand what values we will be using for this problem. The variables and values attached are as follows: N = 250 (total movies), r = 68 (number of drama movies; the main focus of the problem), N-r = 182, n = 15 (sample size we set for the problem). "y" is meant to change when going through these problems so it's not consistent like the others. In order to find the probability that exactly 5 of the selected movies are primarily drama, we do the following equation:

$$\boldsymbol{p(Y = 5)} = \frac{\binom{68}{5}\binom{182}{10}}{\binom{250}{15}}$$

After going through the steps for this problem, we end up with the value 0.191923. This means that there is a 19.1923% chance that exactly 5 of the selected movies are primarily defined as a drama. If we want to focus on more than one instance for this problem, we can move on to finding the probability that no more than 3 of the selected movies are primarily drama. We would achieve this answer by using $P(Y \leq 3)$ when getting the answers, which means using hypergeometric distribution and adding up all the instances for the answer. This would mean doing:

$$p(Y \geq 3) = \frac{\binom{68}{0}\binom{182}{15}}{\binom{250}{15}} + \frac{\binom{68}{1}\binom{182}{14}}{\binom{250}{15}} + \frac{\binom{68}{2}\binom{182}{13}}{\binom{250}{15}} + \frac{\binom{68}{3}\binom{182}{12}}{\binom{250}{15}} \approx$$

$$\mathbf{0.007253 + 0.044036 + 0.122207 + 0.205596 = 0.379092}$$

This means that there is 37.9092% chance that no more than 3 of the selected films are mainly drama-based. These simple questions show use the usefulness of hypergeometric distribution, especially since this can be used to calculate time cost when doing funny activities regarding movies and not always using them from a business perspective. Exploring binomial, geometric, negative binomial, and now hypergeometric distribution, we understand and comprehend the usefulness these tools pose underneath all the notations. Moving on, we will next be looking at the usefulness of Poisson distribution.

Poisson distribution's purpose is to model the likelihood of certain events occurring at a specific rate, operating in a fixed area (time or space). This is all under the assumption of independence and constant occurrence. We use the following formula for Poisson:

$$p(y) = \frac{\lambda^y}{y!}e^{-\lambda}, \quad y = 0, 1, 2, \ldots, \lambda > 0$$

This will be employed in the question we will have regarding this. However, instead of sticking to the data we have already set up, we will be posing another fun hypothetical to improve the flexibility we have when going through this dataset. Like I mentioned before, being able to adapt to the concepts and not what's physically written down helps to improve our comprehension skills of the dataset, especially when doing so in real-life situations.

Suppose IMDb users submit trivia entries about movies on the Top 250 list at an average rate of 3 trivia submissions per day across all 250 movies. Assume this follows a Poisson process and employs Poisson concepts. Let the random variable **Y** represent the number of trivia submissions received in a given day.

a) What is the probability that exactly 4 trivia submissions are made in a day?
b) What is the probability that at least 2 submissions are made in a day
c) What is the probability that no more than 1 submission is made in a day?
d) What is the probability that exactly 6 trivia submissions are made across two days?

Before beginning, like we have done before, we will be getting the variable and the values to the forefront for better comprehension of the problem: $\lambda = 3$ (the average rate of trivia submissions; the specific rate of focus). "y" is interchangeable depending on the problem instructions since Poisson focuses on certain events. As for "e", it's just representative of Euler's number and isn't a placeholder for a certain value in these equations. Diving in, to get the probability of exactly 4 trivia submissions being made in a day, we do:

$$p(Y = 4) = \frac{3^4}{4!}e^{-3} \approx 0.168031$$

We can conclude that there is a 16.8031% chance of exactly 4 trivia submissions being made in a day. If we wanted to not look at specific instances and look toward lower bound problems, we work on finding the probability that at least 2 submissions are made in a day. This is done with:

$$\boldsymbol{P(Y \geq 2) = 1 - (P(Y = 0) + P(Y = 1))}$$

If we were to gather those specific instances, we would end up with the values 0.049787 and 0.149361 being added, each belonging to the respective variable in the equation above. We end up getting the value 0.199148, which we subtract from 1, ending us with 0.800852. This means that there is a 80.0852% chance of at least 2 submissions being made in a day. If we wanted to understand what the complement of that is, we would use the value we got from the previous problem, which helps in solving the next problem. By using previous values, we can see that the probability that no more than 1 submission is made in a day is 19.9148%. Sometimes sticking to the block of time (1 day here) isn't always an option and want to see more extensive data. For the sake of not getting too extensive, we want to find out what the probability that exactly 6 submissions are made in two days. This means changing the specific rate value we have ($\lambda = 3$) to $\lambda = 3 \times 2 = 6$ in order to account for the new time block. With this, we do the following:

$$\boldsymbol{p(Y = 6) = \frac{6^6}{6!} e^{-6} \approx 0.160623}$$

We can conclude from this that there is a 16.0623% chance of getting exactly 6 submissions in two days. After going through all these different scenarios, we see that Poisson distribution is another means to an end when calculating the probability of a specific instance occurring when interacting with specific rates, making this another of cost analysis. Not focusing specifically on time, this focuses on the likelihood of an instance occurring in a given period, a particularly important tool for more real-life scenarios such as the one posted above. There is only one more probability distribution tool to go over after this, which is Tchebysheff's Theorem. This tool is meant to better understand bell-shaped graphs from pieces of data and is mainly used to determine a lower bound for the probability of a random variable **Y** of interest fall in an interval.

Tchebysheff's Theorem employs the following method to reach its conclusions:

$$\boldsymbol{P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \ \ or \ \ P(|Y - \mu| \geq k\sigma) \leq 1 - \frac{1}{k^2}}$$

As mentioned before, the purpose of this tool is to see the limit a value falls within, specifically when using this to get lower and upper bounds. The following problem will be working with runtimes for convenience's sake, also due to the fact we have gathered more data on that piece over the others. However, do not fear, we will continue to cover the different pieces of data this data set holds.

The runtimes (in minutes) of the Top 250 IMDb movies are approximately distributed with a mean of 129.13 minutes and a standard deviation of 30.13 minutes. Let **Y** be the runtime of a randomly selected movie from the Top 250 list.

a) Use Tchebysheff's Theorem to find the minimum proportion of movies that have runtimes between 70.87 minutes and 187.39 minutes.
b) What is the maximum proportion of movies that fall outside this interval?

When going through this problem, we make use of Tchebysheff's after understanding what each variable means and the value attached to it: $\boldsymbol{\mu = 129.13}$ ($\boldsymbol{the\ mean\ runtime\ of\ the\ data}$), $\boldsymbol{\sigma = 30.13}$ (the standard deviation of the data). We are also working with a specific interval for the first part of the problem so that interval is [70.87,187.39]. First step in the problem is getting "k", which is done with:

$$\boldsymbol{k = \frac{Distance\ from\ the\ mean}{\sigma} = \frac{129.13 - 70.87}{80.13} \approx 1.9 \approx 2}$$

Using the rounded values, we end up with the equation $\mu \pm 2\sigma$. We then apply the theorem of focus, and we get the following equation:

$$\boldsymbol{P(|Y - \mu| < 2\sigma) \geq 1 - \frac{1}{4} = 0.75}$$

The interpretation of this is that at least 75% of the movie runtimes lie between the [70.87,187.39] range. For the other part of this problem, this is simply the complement to this equation, which would be:

$$\boldsymbol{P(|Y - \mu| \geq 2\sigma) \leq \frac{1}{4} = 0.25}$$

The interpretation of this is that at most 25% of the movie's runtimes fall outside the interval of focus, [70.87,187.39] minutes. With this brief and simple equation, we gain a better understanding of Tchebysheff's and its usefulness. After reviewing all the probability distribution tools, we can see each one has their own purpose, whether it is in regard to specific instances or finding the limits of an interval, all forms of distribution have their place in statistics. With this chapter done, we will be moving on to more advanced concepts and how they can apply to the Top 250 IMDb Movie list.

This report has made an emphasis to focus on numerous pieces of data, with runtime being the main one so far in this report. Before shifting focus, let's clear up what the problems for this portion of the report will be about. To keep it short, the previous section covered discrete variables and this one covers continuous, which parallels calculus concepts and practices. The difference between the two will reveal itself with the problems we encounter, with the first two being big indicators on how this portion will. Instead of looking at runtimes, let's look at movie ratings again, the main way the list is organized, especially since the list travels in descending order according to the movie ratings.

Let **Y** be the random variable representing the IMDb rating of a randomly selected movie from the Top 250 list. Based on the provided data below:

RATING (y)	P(Y=y)
8.0	5/250 = 0.02
8.1	68/250 = 0.272
8.2	55/250 = 0.22
8.3	38/250 = 0.152
8.4	30/250 = 0.12
8.5	22/250 = 0.088
8.6	13/250 = 0.052
8.7	4/250 = 0.016
8.8	7/250 = 0.028
8.9	1/250 = 0.004
9.0	5/250 = 0.02
9.1	0/250 = 0
9.2	1/250 = 0.004
9.3	1/250 = 0.004

a) Construct the distribution function $F(y) = P(Y \leq y)$
b) What is $P(Y \leq 8.5)$?
c) What is $P(8.2 < Y \leq 8.8)$?
d) Sketch or describe the shape of the distribution function. Is it left-skewed, right-skewed, or symmetric?

Starting off with distribution function, we start off by converting the counts that we have into probabilities, which is already done in the table above. Therefore, we can move onto computing cumulative probabilities F(y). To make it easy, we start from the smallest and work our way up the ratings. The formula for this is:

$$F(y) = P(Y \leq y) = \sum_{ratings \leq y} P(Y = y)$$

This formula is our way of getting these cumulative probabilities, which we then put into the table down below:

RATING (y)	P(Y=y)	F(y) = Cumulative Sum
8.0	5/250 = 0.02	0.02
8.1	68/250 = 0.272	0.292
8.2	55/250 = 0.22	0.512
8.3	38/250 = 0.152	0.664
8.4	30/250 = 0.12	0.784
8.5	22/250 = 0.088	0.872
8.6	13/250 = 0.052	0.924
8.7	4/250 = 0.016	0.94
8.8	7/250 = 0.028	0.968
8.9	1/250 = 0.004	0.972
9.0	5/250 = 0.02	0.992
9.1	0/250 = 0	0.992
9.2	1/250 = 0.004	0.996
9.3	1/250 = 0.004	1.000

After getting this table, the answers come to use easily. When they are asking for $P(Y \leq 8.5)$, the simple way to solve it is to look at the cell that holds the cumulative sum pertaining to this number, which works out since cumulative sum works in favor of what the function is asking of you. By that, I mean looking for values that are less or equal to 8.5 can be easily found using a cumulative sum table like this. We get the answer **F(8.5) = 0.872**. Moving onto the next question, which asks about $P(8.2 < Y \leq 8.8)$, it works similar to finding the answer in a table except we are now working with a lower and upper bound. We get the value attached to F(8.8), which is **0.968**. We then look to find the next value which is **F(8.2) = 0.512**. Afterward, we are taking F(8.8) – F(8.2) to find our answer since we are looking for the value within these bounds, and we end up getting 0.456. Before moving onto the next part, the interpretation for the previous two questions is as follows: **87.2% of the ratings in the IMDb Top 250 Movie list are 8.5 or less. Also, 45.6% of the ratings fall in between an 8.2 and an 8.8 rating**. When it comes to this data, we would like to know what shape is made and how it is distributed. If we were to describe the shape of this distribution function, we would first make observations such as there being rapid growth from 8.0 to 8.4. Afterward, there is slower growth past the 8.6 rating and slow increase beyond 9.0. From this, we can conclude that the distribution is **right-skewed** due to the slow growth toward the higher ratings and most movies fall between 8.1 and 8.5, which forms the center bulk for the graph. As we can see, the pace for this problem is different from the others since we are looking at more pieces of data and not using specific statistics formulas to discover certain instances that we were unaware of at first. The next question should do the same as well when it comes to distinguishing the difference between discrete and continuous.

The next section deals with expected values for continuous random variables which follows the formula:

$$\boldsymbol{E(Y) = \int_{-\infty}^{\infty} yf(y)dy}$$

There is no specific way to describe the way the expected value works other than providing formulas revolving around it, however with work and equations we can begin to understand what its purpose is. Staying on the theme of movie ratings, we will use this piece of data to understand what the purpose of using expected values is when it comes to continuous random variables.

Based on the IMDb Top 250 movie list, assume the movie ratings are modeled by a continuous random variable **Y** with the following simplified probability density function:

$$\boldsymbol{f(y) = \begin{cases} 5(y-8), & 8 \leq y \leq 8.4 \\ 0, & otherwise \end{cases}}$$

The reason we used the bounds above is because of the previous question pointing out how clustered the ratings are around this range.

a) Verify that f(y) is a valid probability density function.
b) Find the **mean** (expected value) of **Y**, the rating of a randomly selected movie

When starting this problem off, we need to verify that f(y) is a valid probability density function by checking the bounds and the integral. We use the equation:

$$\int_{-\infty}^{\infty} \boldsymbol{f(y)dy = 1} \;\Rightarrow\; \int_{\mathbf{8}}^{\mathbf{8.4}} \mathbf{5(y-8)dy}$$

Once we get the values put where they need to be, specifically the upper and lower limit, along with the function placed in f(y), we work out the problem. The following is the work for the problem:

$$5\int_{8}^{8.4}(y-8)dy = 5[\frac{(y-8)^2}{2}]\begin{matrix}8.4\\8\end{matrix} = 5[\frac{(8.4-8)^2}{2} - \frac{(8-8)^2}{2}] = 5[\frac{0.4^2}{2}] = 5 \times \frac{0.16}{2} = 0.4$$

The integral does not end up equaling one like we asked above, meaning we have to normalize the function. Continuing on with the equation, we look to get a normalized version of the probability distribution function and define it as such:

$$f(y) = \frac{1}{0.4} \times 5(y-8) = 12.5(y-8)$$

This leads to the corrected PDF being:

$$\boldsymbol{f(y)} = \begin{cases} \mathbf{12.5(y-8)}, & \mathbf{8 \le y \le 8.4} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases}$$

Now that we have a corrected PDF after trying to verify f(y), we look to find other measurements and statistics attached to it, such as the mean of **Y.** We use the following formula to find the mean, or expected value:

$$\boldsymbol{E(Y) = \int_{z}^{b} y \times f(y)dy = \int_{8}^{8.4} y \times 12.5(y-8)dy}$$

After setting up the equation, we continue by distributing *y*:

$$\mathbf{12.5\int_{8}^{8.4} y(y-8)dy = 12.5\int_{8}^{8.4}(y^2-8y)dy = 12.5\left[\frac{y^3}{3} - 4y^2\right]\begin{matrix}8.4\\8\end{matrix} =}$$

$$\mathbf{12.5\left[\left(\frac{(8.4)^3}{3} - 4(8.4)^2\right) - \left(\frac{(8)^3}{3} - 4(8)^2\right)\right]}$$

$$\mathbf{= 12.5[(197.568 - 282.24) - (170.667 - 256)]}$$

$$\mathbf{= 12.5[-84.672 + 85.333] = 12.5(0.661) = 8.2625}$$

When all the work is done, we can see that we have an expected value, or mean, of

μ = E(Y) = 8.26. The usefulness of all this work is that we gain an understanding of how continuous variables work and comprehending it leads to being able to formula problems alongside datasets, or pieces of information tied to a dataset. Moving on, we will continue to look at continuous probability distribution in the form of a tool that parallels the names and formulas used in discrete but now for continuous instead: Gamma Probability Distribution

The Gamma probability distribution is one of the continuous probability distribution formulas that we learn about in this chapter, which is a specific distribution that some people would look to get over other functions that can be found when working with continuous functions. The formula for Gamma probability distribution is as follows:

$$f(y) = \begin{cases} \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \\ 0, \ elsewhere \end{cases}, \quad 0 \le y < \infty, \text{ where } \Gamma(\alpha) = \int_{-\infty}^{\infty} y^{\alpha-1} e^{-y} dy$$

We will be using this formula for the following problem concerning this probability distribution. Switching up from the serious topics regarding the hard and physical data, let's pose another hypothetical since we have yet to do one in this section, where we will circle back to runtimes.

In the IMDb Top 250 movie list, let's say that the runtime of select epic films (movies over 160 minutes) can be modeled using a gamma distribution with shape parameter $\alpha = 3$ and scale parameter $\beta = 20$. Let the random variable **Y** represent the runtime in minutes of a randomly selected epic film from this list.

a) What is the mean and variance of the runtime distribution?
b) What is the probability that a randomly selected epic film has a runtime longer than 80 minutes?

Starting with the mean and variance, which use the following formulas:

$$\mu = E(Y) = \alpha \times \beta$$

$$\sigma^2 = V(Y) = \alpha \times \beta^2$$

We then plug in the values we have into these equations and end up with E(Y) = 60 minutes & V(Y) = 1200 minutes squared. That means that our mean is 60 minutes and our variance is 1200 $\boldsymbol{minutes^2}$**.** Now we want to continue with finding what epic films have a runtime longer than 80 minutes. We first start off with P(Y>80) since we are looking for the films greater than 80 minutes in runtime. After getting the cumulative for this, we then move to finding the gamma CDF. After taking the steps to go through that, we compute:

$$P(Y > 80) = 1 - F(80; \alpha = 3, \beta = 20) = 0.1987$$

This tells us that there is a 19.87% chance that a randomly selected epic film has a runtime longer than 80 minutes, when going through this hypothetical scenario. With that being done, we move onto the last chapter that still deals with calculus-related variables but in a longer and more intricate manner.

From the IMDb Top 250 Movie list, consider the following:

68 movies are primarily classified as Drama, 32 movies are directed by one of the following five directors: Christopher Nolan, Steven Spielberg, Martin Scorsese, Quentin Tarantino, or Alfred Hitchcock. 3 movies are both primarily classified as Drama and directed by one of the five directors listed above.

Let $Y_1 = 1$ if a randomly selected movie is primarily Drama, 0 otherwise

Let $Y_2 = 1$ if a randomly selected movie is directed by one of the five listed directors, 0 otherwise

a) Construct the joint probability function $p(y_1, y_2)$
b) Find the joint cumulative distribution function value $F(1,0) = P(Y_1 \leq 1, Y_2 \leq 0)$

We start this problem off by building a joint probability function as follows:

Joint Frequency Table

	$Y_2 = 1$(5 Directors)	$Y_2 = 0$(Other)	**Total**
Y_1 = 1 (Drama)	3 (given): **0.012**	68-3 = 65: **0.26**	68
Y_1 = 0 (Not Drama)	32-3 = 29: **0.116**	250 – (3 + 65 + 29) = 153: **0.612**	182
Total	**32**	**218**	**250**

We continue with the problem by continuing onto finding the joint distribution function value. Specifically, we are looking for $F(1,0) = P(Y_1 \leq 1, Y_2 \leq 0)$. Looking at the table, the values can be gathered to form a conclusion by looking at the column with none of the main five directors. Adding those values together gets us 0.872, which means that there is a 87.2% chance that you can find a movie not directed by one of the main five directors listed in this problem. We explore more intricate problems like this throughout this chapter, gathering information on possibilities that combine two pieces of data. Also, there might be more potential uses for these concepts as we continue throughout the report.

Marginal and conditional distributions is another means of acquiring statistics on this dataset when viewing it in a continuous way. We will be sticking with the variables we used before in order to stay consistent with what we are interested in regarding this dataset and to avoid doing more problems on runtimes or movie ratings, leaving more room for us to explore the other pieces of information of this dataset.

Out of the IMDb Top 250 movies:

68 are primarily classified as Drama

32 are directed by Christopher Nolan, Steven Spielberg, Martin Scorsese, Quentin Tarantino, or Alfred Hitchcock

3 movies are both primarily classified as Drama and directed by one of those five directors

Let Y_1 = 1 if a randomly selected movie is primarily Drama, 0 otherwise

Let Y_2 = 1 if a randomly selected movie is directed by one of the five listed directors, 0 otherwise

Assume one movie is randomly selected from the Top 250.

a) Construct a joint probability distribution table for Y_1 and Y_2
b) Find the marginal distributions of Y_1 and Y_2
c) What is the conditional probability that a selected movie is directed by one of the five directors given that it is a drama?

Since this problem isn't too far off from the previous one, we will be using the table we had before (no need to reprint it). In order to find the marginal distributions, we look at the values presented in each row (or column) and add them up based on the notation it holds. Concerning Y_1, we have marginal distribution set up as: $P(Y_1 = 1) = 0.012 + 0.26 =$ **0.272** & $P(Y_1 = 0) = 0.116 + 0.612 =$ **0.728**. As for the marginal distribution for the director-based notation: $P(Y_2 = 1) = 0.012 + 0.116 =$ **0.128** & $P(Y_2 = 0) = 0.26 + 0.612 =$ **0.872**. Now that we familiarize ourselves with marginal distribution a bit, let us look at conditional probability, specifically the conditional probability that a selected movie is directed by one of the five directors given that it is a drama. We find the answer by using the following formula:

$$\boldsymbol{P(Y_2 = 1 \mid Y_1 = 1) = \frac{P(Y_1 = 1, Y_2 = 1)}{P(Y_1 = 1)} = \frac{0.012}{0.272} \approx 0.0441}$$

This means that there is a **4.41%** chance that a primarily drama-based movie is directed by one of the five directors. With the understanding becoming clearer on the distribution techniques being employed here (joint cumulative distribution, conditional probability, and marginal probability), we become to see that this chapter specializes in taking pieces of information for the sake of calculating probabilities connecting the two in a meaningful way. We now move onto the last section of focus which deals with independent random variables.

The importance of independent random variables lies in making systems based around that where different factors can operate without having influence over each other, which provides a clear and understandable insight into complicated systems. Independent random variables make use of the following formula if the variables are said to be independent:

$$\boldsymbol{F(y_1, y_2) = F_1(y_1)F_2(y_2)} \textbf{ for every pair of real numbers } \boldsymbol{(y_1, y_2)}.$$

We will be using this formula to understand the importance independent random variables hold in regard to the Top 250 IMDb Movie list. This will continue to follow the theme of this chapter by ending off with a question regarding the five directors and genre.

Out of the IMDb Top 250 movies:

68 are primarily classified as Drama

32 are directed by Christopher Nolan, Steven Spielberg, Martin Scorsese, Quentin Tarantino, or Alfred Hitchcock

3 movies are both primarily classified as Drama and directed by one of those five directors

Let $Y_1 = 1$ if a randomly selected movie is primarily Drama, 0 otherwise

Let $Y_2 = 1$ if a randomly selected movie is directed by one of the five listed directors, 0 otherwise

Are Y_1 and Y_2 independent? Justify your answer

In order to answer the question, we need to understand that Y_1 and Y_2 are independent if and only if

$$\mathbf{P(}Y_1 \mathbf{= a,}\ Y_1\ Y_2 \mathbf{= b) = P(}Y_1 \mathbf{= a) * P(}Y_2 \mathbf{= b) \text{ for all values a, b}}$$

First, we have to identify the total number of movies which we know to be 250. Now we have to gather the total for our values of focus with $P(Y_1 = 1) = 68/250$ & $P(Y_2 = 1) = 32/250$. The combination of the two is found in the problem instructions, 3/250. We then check for the multiplication condition before seeing if both sides equal each other, which ends up being 2176/62500 = 0.0348 and when computing the actual joint value, 3/250, we get 0.012 (can be seen in the table for this chapter). As we can see, 0.0348 does not equal 0.012, showing that the two variables are not independent of each other, meaning there is some dependency between being a Drama and being directed by one of the five directors. This concludes the research we have done for this paper regarding the chapters in the *Mathematical Statistics* textbook, with the main focus being the Top 250 IMDb Movie dataset. The resources that were used outside of the textbook can be found on the next page.

Resources Used for the Research of this Report

Z-score table resource: ***https://www.ztable.net/***

MPAA rating information: ***https://stephenfollows.com/p/which-mpaa-rating-earns-the-most-money***

Geeks for Geeks Reference: ***https://www.geeksforgeeks.org/probability-distribution/***

Top 250 Movies Dataset (Kaggle): ***https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset***