Table of Contents

[INTRODUCTION 2](#_Toc184742248)

[1.2: Histogram 3](#_Toc184742249)

[1.3: Variance, mean, standard deviation 3](#_Toc184742250)

[2.3: Set notation 5](#_Toc184742251)

[2.4: A Probabilistic Model for an Experiment: The Discrete Case 6](#_Toc184742252)

[2.5: Calculating the Probability of an Event: The Sample-Point Method 7](#_Toc184742253)

[2.6: Tools for Counting Sample Points 8](#_Toc184742254)

[2.7: Conditional Probability and the Independence of Events 9](#_Toc184742255)

[2.8: two laws of probability 10](#_Toc184742256)

[2.9: Calculating the Probability of an Event: The Event-Composition Method 11](#_Toc184742257)

[2.10: The Law of Total Probability and Bayes’ Rule 11](#_Toc184742258)

[3.2: The Probability Distribution for a Discrete Random Variable 12](#_Toc184742259)

[3.3: The Expected Value of a Random Variable or a Function of a Random Variable 14](#_Toc184742260)

[3.4: Binominal Distribution 15](#_Toc184742261)

[3.5: Geometric Distribution 16](#_Toc184742262)

[3.6: Negative Binominal Distribution 17](#_Toc184742263)

[3.7: Hypergeometric Distribution 18](#_Toc184742264)

[3.8: Poisson Distribution 19](#_Toc184742265)

[3.11: Tchebysheff’s Theorem 20](#_Toc184742266)

[4.2: The probability Distribution for a Continuous Random Variable 21](#_Toc184742267)

[4.3: Expected Values for Continuous Random Variables 22](#_Toc184742268)

[4.4: Uniform Probability Distribution 23](#_Toc184742269)

[4.6: Gamma Probability distribution (exponential) 24](#_Toc184742270)

[5.2: Bivariate and Multivariate probability distribution 24](#_Toc184742271)

[5.3: Marginal and Conditional Probability Distributions 26](#_Toc184742272)

[CONCLUSION + FINDINGS 27](#_Toc184742273)

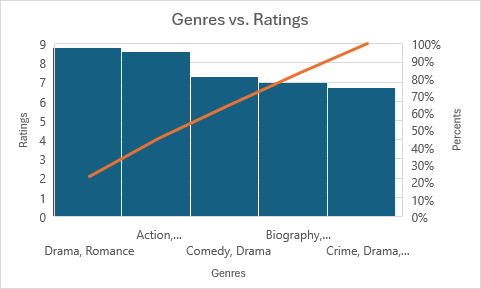
# INTRODUCTION

Netflix is a popular streaming service, released in 2007. A large amount of content has been released on Netflix. From older content to brand new content, Netflix has everything. It also splits it library into Tv shows and movies. Historically, the platform has been populated by a large degree of movies more so than tv shows. Over half of Netflix’s tv show catalog are Netflix originials, however, that can’t be said for movies. Movies are added to the catalog, off and on, from many locations. Netflix is available all around the world and the content offered changes based on the location. Movies and Tv shows can be analyzed by their ratings on Imdb. They are rated accordingly from 1-10. Throughout Netflix’s large catalog overtime, the ratings can viewed and compared accordingly.

Researchers are looking to find many qualities of this large dataset. Some of them being the randomness of different genres content amounts and the variance, mean and standard deviation of a group of movies. This research study works to find all of the specific statistical information based off of genre, movie and tv show count, rankings of catalog items based off of their imdb score, randomness of the content, release order and the alphabetical apperance of content titles.

From this information, researchers want to better understand the entire Netflix catalog, reporting it back to Netflix in order to enhance what is on Netflix. By analyzing the content release dates and their ratings, researchers can come to a conclusion of what is most loved on Netflix. This varies from content type (movie or tv show) and genre, which helps researchers provide even more insight to specifics of the data. Below the study begins, navigating different statistical approaches.

# 1.2: Histogram



From a survey, there were a couple genres picked and compared. A histogram was formed. It shows the distribution of ratings for each genre.

The drama, romance group was rated the highest.

The Crime drama group was rated the lowest.

Researchers analyze the skew of data. The highest group (drama, romance) is left-skewed but the ratings rise above that. The lowest group (Crime, drama) is right-skewed and the ratings are low under.

# 1.3: Variance, mean, standard deviation

Netflix data researchers have chosen 6 movies to be denoted as (1, 2, 3, 2, 5, 3) which are some ratings they took from a survey sent out randomly to Netflix users. They want to find the mean, variance and standard deviation out of these six movies’ ratings. The total of movies in their library is 15,334, but these four are from the fantasy genre.

Researchers find the **mean** of these 6 movies is

They find that the mean is slightly below the mid-point, showing that users are rating fantasy movies from the survey moderately.

The **variance** was found by researchers and is below:

= approximately 1.56.

This shows that the values are moderately spread around the mean, not super close or clustered but not too far either.

The **standard deviation** was found and is as follows:

=1.25

There is a moderate consistency in the data.

With this survey, it was found that out of all 6 movies, the users rated them moderately the same. 5 was the large outlier, however, the other values showed a similarity and moderate consistency.

# 2.3: Set notation

A diagram of comedy and comedy

Description automatically generated

Suppose that A (Movies) and B(comedy) are two independent events. Researchers want to find the union, intersection, and complement of these two events.

There are two movies being picked from the Netflix catalog.

A movie can be a different genre or the comedy genre.

A non-comedy movie is denoted as “N” and comedy is denoted as “C”.

The universal set is based below:

S = {CC, NC, CN, NN}

Both movies can be comedy (CC), the first movie picked can be not comedy and the second may be comedy (NC), the first movie picked may be comedy and the second not (CN), both movies picked may not be comedy.

Let A denote the subset where there are no comedy movies, B, the subset containing two comedy movies, and C the subset containing at least one comedy movie.

Our subsets are:

Now we must find A, B, C, C B, A B, and C

A = {NN}

B = {CC}

C = {NC, CN}

C B = empty set.

A B = {NN, CC}

C = {NC, CN}

Researchers find that audiences are more interested in diverse content. They are more likely to watch multiple genres and not just stick to comedy.

# 2.4: A Probabilistic Model for an Experiment: The Discrete Case

Netflix has a wide array of genres on its platform. However, a lot of that content has been added in recent years. The amount of content added in 2018, 2019, 2022, and 2023 are 0.073, 0.0740, 0.0956, and 0.0857, respectively. A single year is chosen at random based on this information.

Researchers find that the sample space is as below:

S = {2018, 2019, 2022, 2023}

According to the data provided, researchers match the years with their probabilities, finding that 2022 has the highest probability and 2019 the lowest probability—however, these four years consist of the largest quantity of all genre releases that have been on Netflix.

2019 = 0.073. 2019 = 0.0740. 2022 = 0.0956. 2023 = 0.0857.

Researchers decided to analyze the probability of 2018 or 2022, finding the probability.

P(2018 or 2022) = 0.0723 + 0.0956 = 16.79%

They find that looking into either 2018 or 2022, gives a probability of 16.79%. So, these two years add up to 16.79% which is a high value of content released in those two years on Netflix, compared to other years.

gave the values in the problem. 😊

# 2.5: Calculating the Probability of an Event: The Sample-Point Method

Four Netflix movies are picked, being ranked from lowest to highest by IMDB users. That is one movie is ranked best, one second best, one third best and one movie is ranked the worst.

Researchers discovered one sample point.

Movie 1: best

Movie 2: second best

Movie 3: third best

Movie 4: worst

Researchers developed a sample space consisting of all four movies in different combinations to portray how the movies can be ordered.

(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), (1, 3, 4, 2), (1, 4, 2, 3), (1, 4, 3, 2), (2, 1, 3, 4), (2, 1, 4, 3), (2, 3, 1, 4), (2, 3, 4, 1), (2, 4, 1, 3), (2, 4, 3, 1), (3, 1, 2, 4), (3, 1, 4, 2), (3, 2, 1, 4), (3, 2, 4, 1), (3, 4, 1, 2), (3, 4, 2, 1), (4, 1, 2, 3), (4, 1, 3, 2), (4, 2, 1, 3), (4, 2, 3, 1), (4, 3, 1, 2), (4, 3, 2, 1)

Researchers want to look at the case where the people rating didn’t really watch the movie and give the movies random ratings. One of the movies is of much better rating than the others. They want to find the probability that people will rate the best movie no worse than second best.

12/24 = 0.5.

# 2.6: Tools for Counting Sample Points

A Netflix content manager is analyzing the release order for four released movies. The release order, and hence the impact on subscriber engagement, will depend on the order in which the movies are released.

The Netflix content manager works to discover what different release schedules (and engagement outcomes) are possible.

(4!) = 24 release schedules are possible based on this information.

The content manager wants to find out if a release sequence is chosen randomly, and two of the movies released are "Marlene" and "Saw IV" what is the probability that "Marlene" was released before "Saw IV"?

There are 24 release schedules. There are two movies mentioned here.

24 / 2 = 12.

12/24 = 0.5.

The probability that Marlene was released before Saw IV is 50% or 1/2.

# 2.7: Conditional Probability and the Independence of Events

In a certain population of movies and TV shows on Netflix, the percentage of family movies is given based on the total number of movies and TV shows. 0.0604% of tv shows are in the family genre, 99.93% are not in the family genre. 0.252% of movies are in the movie genre and 99.7% are not in the family genre. Based on this information, Netflix wants to find the independent events between the data to find out how the family genre quantity compares against every other genre.

Let F be the event that a movie or TV show is in the family genre, and let M be the event that the selected item is a movie. Let T be the event where the selected item is a TV show. Let be the event that a movie or TV show is not of the family genre.

While analyzing the data, researchers find that the total number of Netflix movies is 19,861. The total number of movies is 15,334. Finding P(F ∩ M) comes out to be 50. Family content of both Tv shows and movies comes out to 62.

Number of family movies / total number of genres =

Number of family tv shows / total number of genres =

|  |  |  |  |
| --- | --- | --- | --- |
| Outcome | Tv show (T) | Movie (M) | Total |
| Family (F) | 0.0604% | 0.0252% | 0.312% |
| Non-family movie ( | 99.9396%. | 99.748% | |  | | --- | | 99.688% | |
| Total | 100 | 100 | |  | | --- | |  | |

Researchers are looking to find if T and F are independent.

P(tv show = 0.0604

100 x 0.025804 = 2.5804

Researchers find that these values are not independent.

# 2.8: two laws of probability

From our previous data, researchers have analyzed that 0.0604% of TV shows are in the family genre and approximately 99.9% are not. Researchers now want to compare two distinct values 99.9% and 99.7% as well. They want to see if any of these values are mutually exclusive to analyze data trends.

They look to find if P(A) being .0604 and P(B) being 99.9 makes A and B mutually exclusive as well as the case where P(A) is 99.9 and P(B) is 99.7.

**Can A and B be mutually exclusive if P(A)=** 0.0604 **and P(B)=** 0.0252**? If P(A)=99.9 and P(B)=99.7? Why?**

In the first case, they are mutually exclusive due to being less than one. That means that a piece of content cannot be both a movie and a TV show. In the second case, researchers find that they are also not mutually exclusive due to being more than one. This information tells us that both events can occur on their own. For example, a piece of content can be a movie without being of the family genre.

# 2.9: Calculating the Probability of an Event: The Event-Composition Method

A Netflix user is found to have a 75% chance of watching a movie starting with a specific letter. (Ex: F, T, K, J, U). The movies are independent from each other. Researchers want to find the probability that the user will watch all five movies.

P(M) = probability of movie. There are 5 movies, so they are denoted by the number.

P(Watch all five movies) = P(M1) \* P(M2) \* P(M3) \* P(M4) \* P(M5)

= 0.2373

Researchers find that the probability of the user watching all five movies is 23.73%.

# 2.10: The Law of Total Probability and Bayes’ Rule

Researchers find that there is a population of movies with ratings over 8 which is 40% and movies that are under 6 which is 60%. It is found that 30% of the movies over 8 are in for review 70% of the movies under 6 are in for review.

(0.70 \* 0.60) + (0.30 \* 0.40) = 0.54

Researchers find that 0.54 represents the total probability that a movie is in for review across the entire population of movies.

0.42 / 0.54 = .78

This means that the conditional probability that a movie in for review has a rating over 8 is 78%. In other words, movies with higher ratings (over 8) are more likely to be selected for review compared to movies with lower ratings (under 6). The movies for review are likely to be added to the Netflix catalog.

# 3.2: The Probability Distribution for a Discrete Random Variable

A study is being done to determine randomness of how many movies/tv shows were released from 1992, 1999 and 2023.

59 pieces of content from Netflix are released in 1992.

96 pieces of content are released in 1999.

1701 pieces of content are released in 2023.

Netflix researchers have a group of people randomly arrange the content to each year.

The number of correct matches is:

**Y = 0** (no matches)

**Y = 1** (one correct match)

**Y = 2** (two correct matches)

**Y = 3** (three correct matches)

What do they mean?

**Y = 0**: No correct matches mean all years are assigned incorrectly.

**Y = 1**: One correct match means one year is assigned correctly, and the other two are mismatched.

**Y = 2**: Two correct matches mean two years are assigned correctly, and one is incorrect.

**Y = 3**: All three years are assigned correctly.

3! = 6 ways.

There are three options, so the outcomes come out in 6 different ways.

Below shows the 6 unique cases:

59 – 1992, 96 – 1999, 1701 – 2023 (Y=3)

59 – 1999, 96 – 2023, 1701 – 1992 (Y=0)

59 – 1999, 96 – 1992, 1701 – 2023 (Y=1)

59 – 1992, 96 – 2023, 1701 – 1999 (Y=1)

59 – 2023, 96 – 1992, 1701 – 1999 (Y=0)

59 – 2023, 96 – 1999, 1701 – 1992 (Y=1)

Six arrangements are possible here. All the content numbers can all be arranged with the right year. They can all be arranged wrong twice. Three cases show only one of the years being correlated with the correct content amount.

Y(1) = 3 = 3/6 = 2/3

Y(0) = 2 = 2/6 = 1/3

Y(3) = 1 = 1/6

Researchers find that 67% of the cases show only one year being correlated correctly. 33% of the time shows no outcome being correct in the listing. 17% of the time shows all three choices being arranged correctly.

They find there is an in-the-middle case where a person may match nothing correct.

They found that it is rare to get all three choices correct in a case where a person does not know the correct arrangement.

Most of the time a person can match at least one correctly.

# 3.3: The Expected Value of a Random Variable or a Function of a Random Variable

Researchers spin a wheel with all 5 western movies on it at *once.* Y is the number that is picked. They want to find the expected and variance of Y.

The PMF is 1/5 (they have 5 western movies to land on, which is the total of them in the Netflix catalog).

1(1/5) + 2(1/5) + 3(1/5) + 4(1/5) + 5(1/5) = = 15/3 = 3 is the expected value.

11.

11 - = 2 is the variance.

From this data, researchers find that from the random data in the wheel, they are most likely to land on the third western movie.

The pieces of data they find that will be landed on other than this (the variance) suggest that it will be two away from the mean (3).

# 3.4: Binominal Distribution

Researchers pick 10 movies from the dataset. The probability of landing on a western movie (5 of them) is 20%. They want to find the probability that exactly 3 of the movies are of the western genre.

N = 10 (the number of trials).

K = 3 (number of successes)

P = 1/5 (probability of success)

= 20.13%.

Researchers find that the probability that exactly of 3 movies picked being western is 20.13%.

# 3.5: Geometric Distribution

A study proposes 20 movies. The probability that a movie starts with the letter A is 10% out of the group of movies picked (two started with A). Movies are randomly presented to users to review in no order. Researchers want to find the probability that the first movie starting with the letter A will be picked for review by Netflix users on the second try.

Probability of success (p) = 10%

The first success will occur (k) = 2

The probability of failure (1-p or q) = 90%

= the probability of failure.

0.90 \* 0.10 = 0.09% the probability of success.

Researchers find the probability of a movie starting with A popping up on the review page is 0.09%. Since the percentage is very low (10%), it is unlikely that a user will see the movies starting with A on the second try on the review page. That’s what the 90% failure rate suggests as well.

# 3.6: Negative Binominal Distribution

Movies are being analyzed for their ratings. 42% of the movies in the grouping are above rating 8, and researchers want to find the probability that that when the 10th movie is picked by a user, exactly three movies are over the rating 8.

The probability of success (p) is 42%.

The number of successes (r) is 3 (over rating of 8).

The total number of trials (n) is 10 movies.

The probability of failure is 1 – p (q) = 0.58%.

Researchers find that there is an 8.03% success rate of exactly three movies being over the rating 8 by the 10th movie.

# 3.7: Hypergeometric Distribution

There are 20 movies that have been chosen which include 3 that are horror films. Researchers want to find out the minimum number of movies that must be selected if we require that P(at least one horror movie) .80

Total movies = 20

Horror movies = 3

Non-horror movies = 20−3=17

Therefore, the hypergeometric formula is as follows:

(20-n)(19-n)(18-n)0.2

(20-n)(19-n)(18-n)1368

Seven movies are chosen.

n=7

(20-n)(19-n)(18-n)1368

This doesn’t meet the criteria as it’s larger than the threshold.

Eight movies are chosen.

n=8

(20-n)(19-n)(18-n)1368

This is less than the threshold—becoming the minimum and meeting the criteria.

Therefore, the minimum number of movies will be 8. When 8 is hit, the threshold passes 80% likely to have at least one horror movie. Researchers find that this study applies to certain algorithmic procedures in the Netflix database, leading to people who watch horror movies and other genres being more likely to see a horror movie recommended by the 8th movie shown.

# 3.8: Poisson Distribution

Researchers analyze how Netflix users pick movies from the database based off genre. There are 4 crime documentaries. The average number of crime, musical documentaries to click on per hour is λ = 3.

What is the probability that exactly 4 crime musical documentaries are chosen?

What is the probability at least 2 crime musical documentaries are chosen?

1 – P(0) – P(1)

1 - = 0.800852

What is the probability no more than 3 crime musical documentaries are chosen?

= 0.6472

From this, researchers find there is approximately a 16.8% probability that exactly 4 crime musical documentaries are chosen. There is an 80% chance that at least 2 crime musical documentaries are chosen. There is a 65% chance that no more than 3 crime musical documentaries are chosen. From this, they find that it is a low probability that exactly 4 crime musical documentaries will be chosen. However, at least two being chosen is the highest with no more than 3 being chosen being second highest, which makes sense.

# 3.11: Tchebysheff’s Theorem

Let’s go back to our standard deviation and mean section.

Netflix data researchers have chosen 6 movies to be denoted as (1, 2, 3, 2, 5, 3) which are some ratings they took from a survey sent out randomly to Netflix users.

Researchers find the **mean** of these 6 movies is

The **standard deviation** was found and is as follows:

=1.25

The average of these movies is 2.6. The standard deviation is 1.25. Researchers want to find a lower bound for the number of movies which there being 6 of them that are expected to have ratings from 0.02 to 0.05.

P(0.02 < X < 0.05) > 1 -

The upper and lower bound are found:

|2.6 – 0.02| / 1.25= 2.064

|0.05 – 2.6| / 1.25= 2.04

Then the formula is done:

1 – (= 0.7596

The final expected answer is found:

6 \* 0.7596 = 4.56

Researchers find that the lower bound of the movies with ratings 0.02 to 0.05 is 4.56. So approximately, 5 movies (if rounded) will be in the lower bound of that. The number could be higher since this is a lower bound, but at least 5 movies will be in that range.

# 4.2: The probability Distribution for a Continuous Random Variable

Netflix researchers select five TV shows, with only one beginning with the letter S. Only one of them starts with S. Movies are selected and tried, one at a time, until the movie is chosen (movies that do not start with an s are discarded before another is picked). Let Y be the number of the trial on which the movie starting with S is chosen. This part of the study’s purpose is to find out how long it would take to choose one letter out of 26 while choosing 5.

We find that all cases have a probability of 1/5 or 0.2.

P(Y=1) = 1/5

P(Y=2) = 1/5

P(y=3) – 1/5

P(Y=4) – 1/5

P(Y=5) – 1/5

The distribution function was shown below:

P(y < 0 ) = 0

P(0 (1/5)

P(1 < 2) = 0.2 + 1/5 = 0.4

P(2

P(3

P(P

They want to find the probability of it showing up on the first, second and third trial.

P(Y=1) + P(Y=2) + P(Y=3)

= (1/5) + (1/5) + (1/5)

= 3/5

So, the probability of Y being less than or equal to three is 0.6.

So, there is a probability of 60% that the movie starting with S will appear from the first trial to the third trial.

# 4.3: Expected Values for Continuous Random Variables

Ratings from Netflix (Y) tend to range from 5 to 8 statistically which produces the probability function below:

Researchers want to find the mean and variance of this function they have set up.

The mean E(Y) is found as follows:

The variance V(Y) is found as follows:

Ratings are equally likely to occur in the interval [5, 8].

Researchers observing ratings between 5 and 8 can confidently expect most values to hover near the mean, with equal chances of being higher or lower within the range.

# 4.4: Uniform Probability Distribution

Netflix researchers choose five different movies in which they want to test the algorithm. They test their algorithm’s power by seeing which movie it lands. The algorithm lands on a random movie between markers A and B. They want to find the probability that the distance of the movie in the algorithm is more than four times the distance to B.

P(X > 4(L-X))

P(X > 4L -4X)

P(5X > 4L)

P(X > 0.8L)

Researchers find that 20% of the time, the algorithm is expected to select a movie in the last 20% between A and B. Based off user preference and genre, they find that the movies in the last 20% are picked by the algorithm to show to these users.

# 4.6: Gamma Probability distribution (exponential)

The magnitude of Netflix movies within the data set is modeled with a mean rating score of 2.6.

Researchers want to find the probability that a movie will exceed a rating of 3 and the probability that it will fall between 2.0 and 3.0 on the rating scale.

P(Y > 3) =

P(Y > 3) =

P(Y > 3) =

P(Y > 3) = 0.2865

So, the probability that a movie will exceed a rating of three is 28.7% approximately.

So, the probability that a movie will have between a 2 and 3 rating is 14.8%.

Researchers analyze these findings and can see that movies that have more than a three rating are more likely to exist than movies that have one between a 2 and 3 rating.

# 5.2: Bivariate and Multivariate probability distribution

Netflix wants to randomly assign countries for two horror movies to one or more of three groups. A, B and C. Researchers let ‘’ denote how many countries are assigned to group A and the number of countries assigned to group B.

They want to find the joint probability function.

They find out that the sample space of this information is:

{(0,0)(0,1)(0,2)(1,0)(1,1)(1,2)(2,0)(2,1)(2,2)}

This leads them to the conclusion that P(= 0) =

So, this repeats for all of the sample space.

P(0, 1) =

P(0,2) =

P(1, 0) =

P(1, 1) =

P(1, 2) = 0

P(2, 0) =

P(2, 1) = 0

P(2, 2) = 0

**JOINT PDF TABLE**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
| 0 | 1/9 | 2/9 | 1/9 |
| 1 | 2/9 | 2/9 | 0 |
| 2 | 1/9 | 0 | 0 |

They find F(1, 0) next:

P(0,0) + p(1,0) = 1/9 + 2/9 = 1/3.

Researchers find out that assigning at most one country to group A and none to group B is 33%.

# 5.3: Marginal and Conditional Probability Distributions

In section 5.2, researchers determined the joint distribution of To add onto their findings, they want to discover the marginal probability distribution of .

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
|  | 4/9 | 4/9 | 1/9 |
|  | 0 | 4/9 | 2/9 |
|  | 0 | 4/9 | 4/9 |

There is a 44.4% chance no countries are assigned to Group A.

There is a 44.4% chance one country is assigned to Group A.

There is an 11.1% chance two countries are assigned to Group A.

0+4/9​+2/9​=6/9=2/3 (mean)

0+4/9​+4/9=8/9

8/9 - =4/9 (variance)

On average, less than one country is assigned to Group A.

# CONCLUSION + FINDINGS

To sum things up, there were multiple fascinating findings that researchers found out regarding content ratings, content schedules, content countries, content genres and much more. They used all these statistical approaches to see what they could find in terms of the algorithm and ratings. This helped them determine which movies should be pushed out to users based off the rating and genre mostly. The report aims to provide customer satisfaction as the algorithm suggests content, they’ll like based off what they have already watched. It dives into the rating system to see how many low and high rating pieces of content have been on Netflix. It investigates genres to see the amount of each which has been on Netflix even versus user rating.

By completing this report, using data analysis and statistical mathematical computations, researchers were able to provide Netflix with the ability to make data-driven decisions to improve their content library and overall customer satisfaction. In the future, researchers plan to use AI machine learning tasks to force more insight out of the Netflix algorithm and platform.

**Researchers have prepared an important findings PowerPoint called:**

Netflix Interesting Findings.ppt

Found in the same folder as this report.