Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 01
Lecture - 05

Lecture 5: Data Description

Hello everyone. Welcome back. We are now going to start the 5th lecture of this series. And as you remember in the last lecture we tried talking about the data aspect of recommender system. In this context we saw what are various rating scales and we continue after that. Specifically, in this lecture we are going to talk about the data description part.

When we talk about the data, data can be structured or unstructured. Structured data is organized and it is very easy to work with. And it can be conveniently brought to numeric form which can ultimately be used by the algorithm. Whereas you have to little work extra on unstructured data and bring it to a structured form so that it can be used by the algorithm that you are targeting to.

Consider this case. This is the data about movies specifically from movie lens data set. This is the data about movie movies. This is the data about movie. This is the data about the user.

And this is joins movie data with user data. Let us look at the nature of this data. And in which kind of rating scale it is in. Look movie id is of course, we studied about 4 types of rating scales nominal, ordinal, interval and ratio. So obviously, this is a nominal.

This is also nominal because this is name. But if you look at this in this data there are 2 parts. One is the name of the movies and the year of release. So therefore, when we have to process this data we have to probably separate these 2. Similarly, look at the genre of the movie.

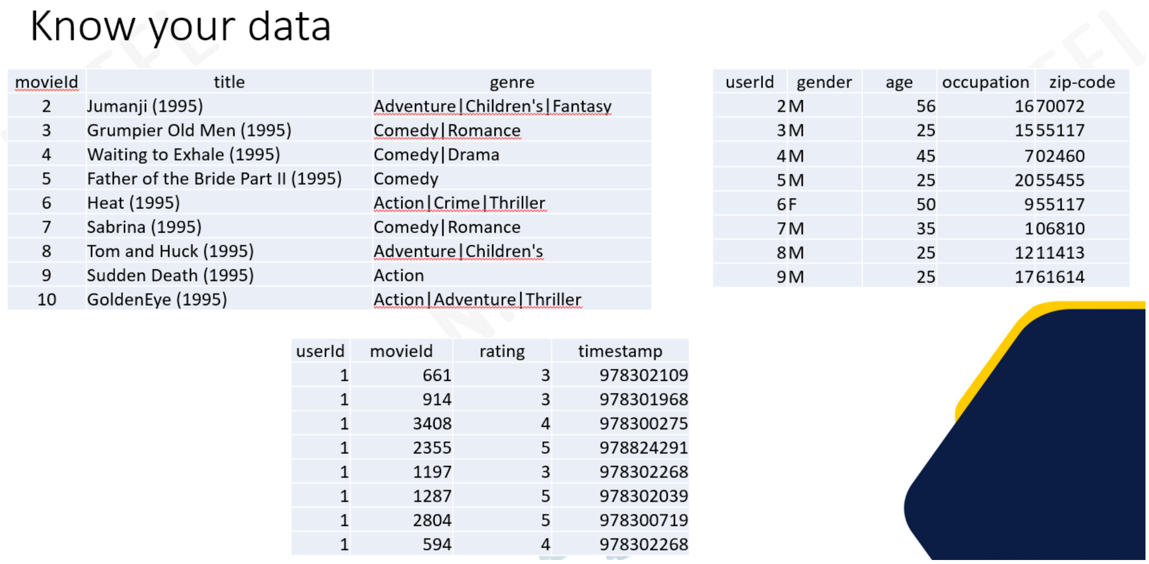
These are the 3 genres for this movie. Now if you have to work upon probably we have to represent it in a different manner because these are basically text data. So now come to the user. Similarly user has user id it is which is again nominal. Gender is also nominal.

Age is in ratio scale. And occupation which is again a nominal and zip code. Zip code is also a nominal value. Coming to this user id combination we have rating data as well where for a specific movie id how a user i user 1 has given the rating. Now to bring it to some processing convenience what we are supposed to do? Let us say this is our model.

We will be giving some data as input and we will be getting some output let us say some class as our output. So this input data which goes into the model. What is a model after all?

Model is basically certain computational function. So it is a kind of function very complex function that you will be building. Nature of the function can be anything.

Now when it you give input this input has to come in numeric form. And how to convert this data to various numeric form to let us say adventure, children, fantasy. Suppose all the genre you will be let us say there are total n number of genre 1 to n number of genre. So wherever it is 1 let us say this belongs to adventure, this one belongs to children's movie, this one belongs to fantasy. So probably we will be making this 1, this is as 1, this is as 1 and rest will be 0's.



Know your data

movieId	title	genre
2	Jumanji (1995)	Adventure|Children's|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action|Crime|Thriller
7	Sabrina (1995)	Comedy|Romance
8	Tom and Huck (1995)	Adventure|Children's
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action|Adventure|Thriller

userId	gender	age	occupation	zip-code
2	M	56	16	70072
3	M	25	15	55117
4	M	45	7	02460
5	M	25	20	55455
6	F	50	9	55117
7	M	35	1	06810
8	M	25	12	11413
9	M	25	17	61614

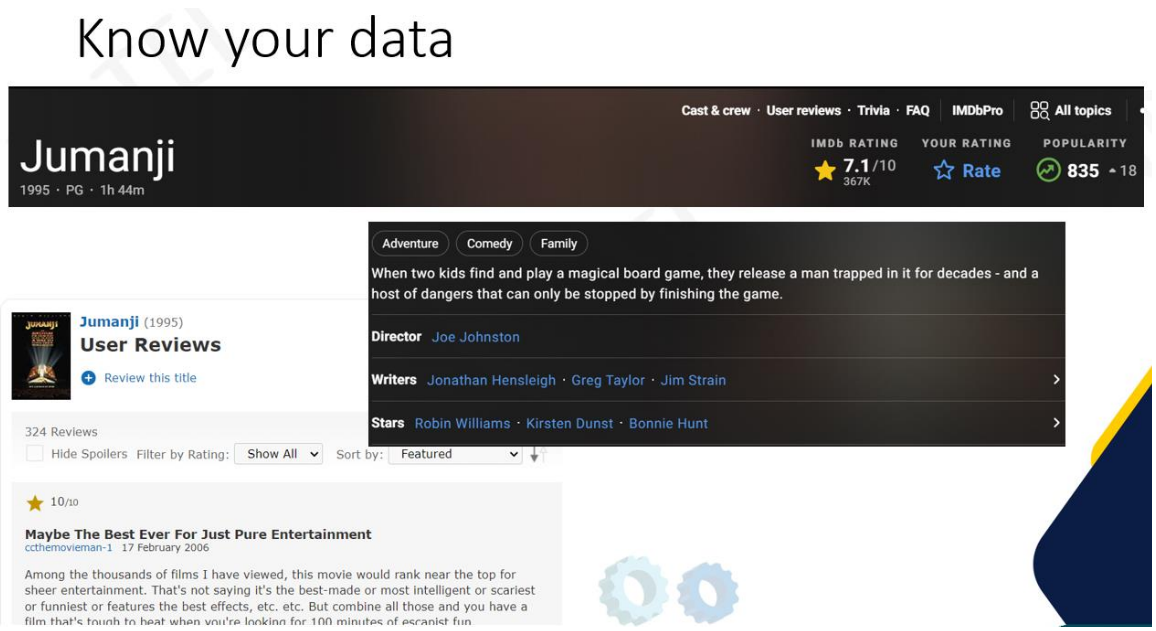
userId	movieId	rating	timestamp
1	661	3	978302109
1	914	3	978301968
1	3408	4	978300275
1	2355	5	978824291
1	1197	3	978302268
1	1287	5	978302039
1	2804	5	978300719
1	594	4	978302268

So that is how we will be making a vector to represent genre. So now let us go this data was pretty structured. So we can even think of that representing this name as instead of name we can use movie id and later on whenever required we can connect it. User id is of course, this is also nominal. Rating is anyway in a ordinal scale.

And now comes to this adventure comedy etc. as I told you we will be representing it in the form of this kind of vector. Let us go little bit more deep. Let us say some movie Jumanji and we have got this data that was from movie lens and we have got this data from IMDB. Of course, I have given certain snapshots of this the IMDB website.

So it has certain rating, then certain popularity score, cast and cues you can go into detail. Besides that it will also give you what kind of genre it is, who is the director, writer, leading stars and this is a short description which is in text. How do I convert this data into numeric? Similarly, look at this it has many user reviews and user reviews are detailed here. So there are total 324 reviews and this is just one of the reviews. This writes like among the thousands of films blah blah blah blah.

So which means this whole thing now I have to bring to certain numeric form. So what is that numeric form in which I will be using this and where I will be using this? Suppose if you remember in the basic form the matrix 3 there are 3 matrices. One is the user matrix which provides the user demographics, then we have item matrix which provides item features. And of course, we have the rating matrix. Now this is one item Jumanji as a movie is one item.



Know your data

Cast & crew · User reviews · Trivia · FAQ | IMDbPro | All topics

Jumanji
1995 · PG · 1h 44m

IMDb RATING 7.1/10 367K | YOUR RATING Rate | POPULARITY 835 ▲18

Adventure | Comedy | Family

When two kids find and play a magical board game, they release a man trapped in it for decades - and a host of dangers that can only be stopped by finishing the game.

Director Joe Johnston

Writers Jonathan Hensleigh · Greg Taylor · Jim Strain

Stars Robin Williams · Kirsten Dunst · Bonnie Hunt

Jumanji (1995)
User Reviews
Review this title

324 Reviews
□ Hide Spoilers Filter by Rating: Show All Sort by: Featured

★ 10/10

Maybe The Best Ever For Just Pure Entertainment
ccthemovieman-1 17 February 2006

Among the thousands of films I have viewed, this movie would rank near the top for sheer entertainment. That's not saying it's the best-made or most intelligent or scariest or funniest or features the best effects, etc. etc. But combine all those and you have a film that's tough to beat when you're looking for 100 minutes of escapist fun

So there are many features which you can directly consider like who is the director, the genres and so on. But there are certain features which are supposed to be extracted from this text data. This is one, this is one. So we have to have adequate procedures to convert it from numeric to text data. So one of the form could be you can have some kind of matrix in which the total if you have this is movie along with the explicit features like that of genre etc.

You also have few more values corresponding to this text data. What could be the value? Maybe there are certain important terms which are used frequently. So maybe that those terms and corresponding tf-idf rating which we are going to discuss shortly we can use. Similarly in case of reviews, we can find out the what is the total positive sentiment and what is total negative sentiment. So sentiment can be another important criteria to define the feature.

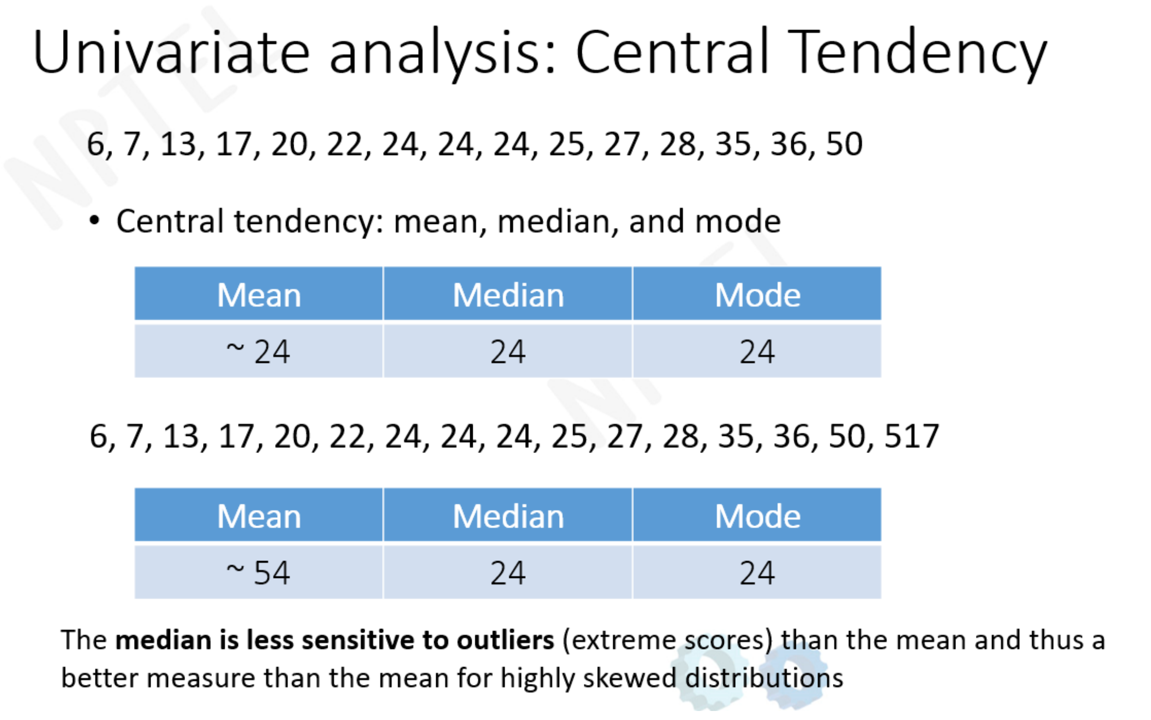
So whatever may be the case, we have to ultimately bring the data into numeric form. Now when we bring the data into numeric form, the data let us say the data that we consider let us say this item detail which has items maybe here i1, i2 etc. Let us say there are m items and there will be many features. So what are these features? Features are individual attributes which will be describing the data. So we must know what is the characteristics of each of these features.

Now when we talk about the characteristics of a particular feature as an individual feature, we can find out certain descriptive statistics which in a univariate way describes the data. And because there are multiple features comparing them also we can derive certain statistics. So basically descriptive statistics are brief descriptive coefficients that summarize a given data set. So individually if we look at each column, it involves describing the distribution of individual variables. So when we talk about this particular variable, as we know this can be in different scales and what is the corresponding descriptive statistics that also we saw.

But for the time being we let us make it a little bit generic and let us see this data is numeric only quantitative data. And in fact, all the data that will be fed to your algorithm has to be brought into a quantitative form. Quantitative in the sense even if it is a data which represents nominal represents a nominal attribute, it also has to be represented in the form of a numeric form. Just like we saw in case of genre, genre is a nominal value, but we represent the genre in terms of a 0, 1 vector. But anyway for the time being let us look at what are various univariate analysis that we must perform to characterize the data.

Let me tell you this particular course is on recommender systems. Therefore, while discussing about this basic foundational aspects on statistics, maybe something on matrices and all as well as machine learning, I will be limiting myself to only basic concepts. To go into in-depth details of those topics, we have to be referring to books or other online resources. So, let us go ahead with univariate analysis. So, measurement of central tendency as you know can be done in 3 different using 3 different statistics one is mean, median and mode.

These are the measurements of measures of central tendency. So, this you already know. However, I have to make little bit aware, I have to make you aware of little bit more details. Consider this particular data. And this is mean, this is median, this is mode.



Univariate analysis: Central Tendency

6, 7, 13, 17, 20, 22, 24, 24, 24, 25, 27, 28, 35, 36, 50

- Central tendency: mean, median, and mode

Mean	Median	Mode
~ 24	24	24

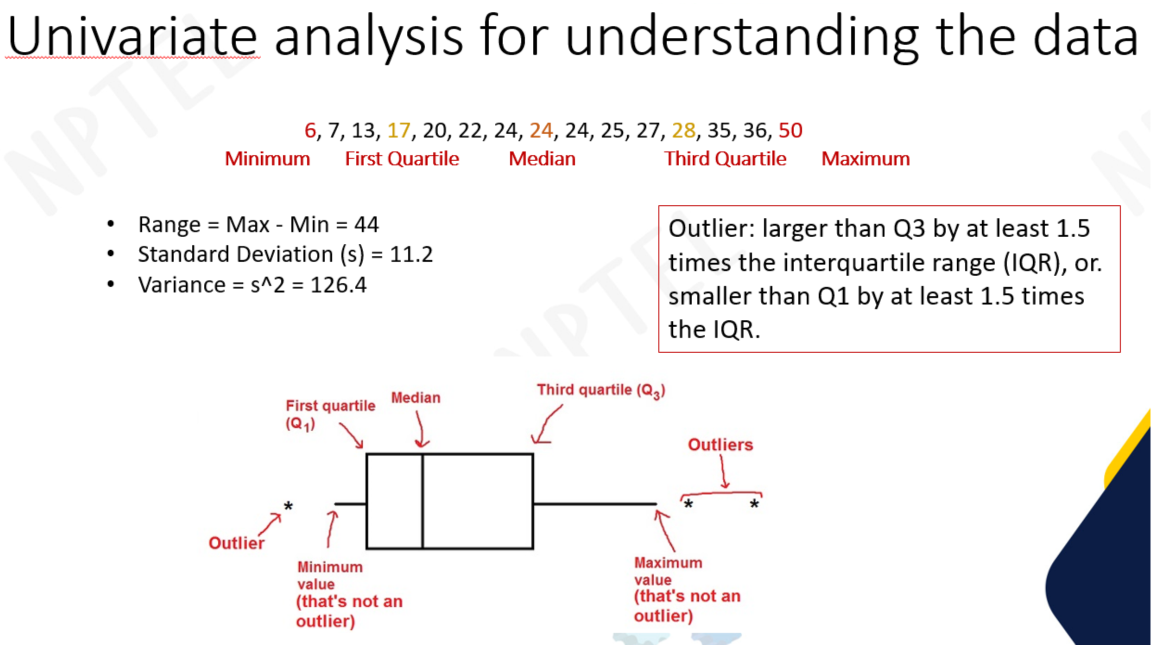
6, 7, 13, 17, 20, 22, 24, 24, 24, 25, 27, 28, 35, 36, 50, 517

Mean	Median	Mode
~ 54	24	24

The **median is less sensitive to outliers** (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions

What is median? The mid value. Mode is the highest number of occurrences of something in this data stream, some element in this data stream. Now look at this. Suddenly, we have added one more very large value in this list of in this data. Now, if we look at mean, median and mode again, we look at this mean value has increased.

But this one may be one outlier. What is an outlier? Some value which is unexpectedly high or low and that cannot follow this particular distribution, distribution from which these values are drawn. Let us say this value represents age of certain person. So, age of certain person being 517 is almost impossible. But consequently because of presence of this, this has increased. So therefore, in many situations, you are supposed to keep track of I mean you should not be blindly following the mean value if you are trying to understand the nature of your data.



Univariate analysis for understanding the data

6, 7, 13, 17, 20, 22, 24, 24, 24, 25, 27, 28, 35, 36, 50

Minimum First Quartile Median Third Quartile Maximum

- Range = Max - Min = 44
- Standard Deviation (s) = 11.2
- Variance = s^2 = 126.4

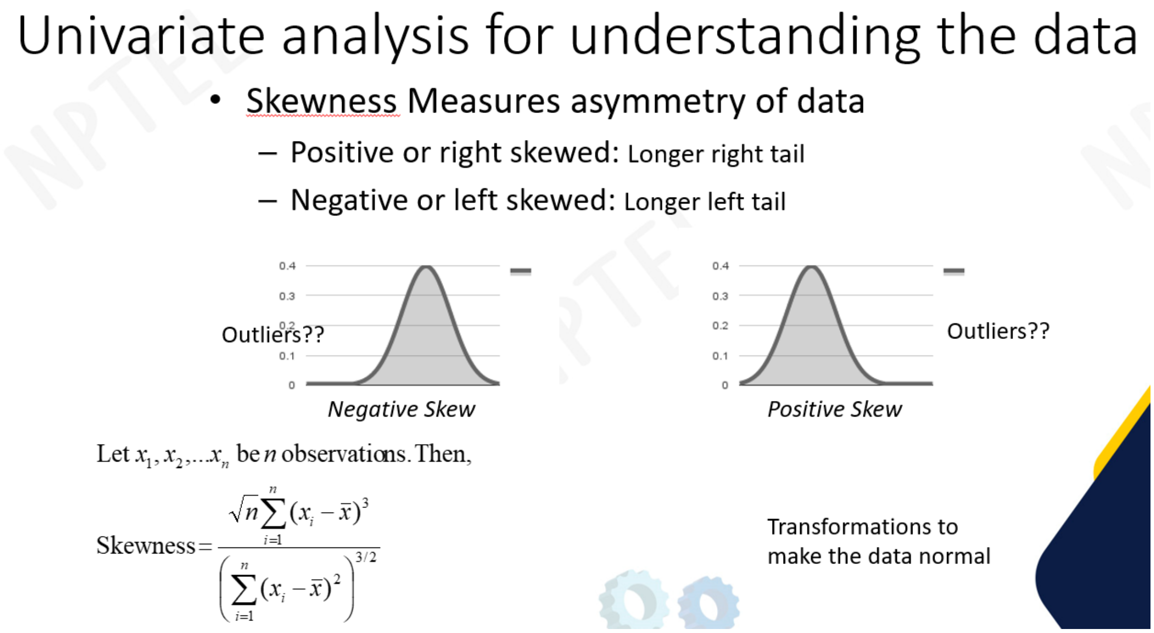
Outlier: larger than Q3 by at least 1.5 times the interquartile range (IQR), or. smaller than Q1 by at least 1.5 times the IQR.

First quartile (Q_1)
Median
Third quartile (Q_3)
Outliers
Outlier
Minimum value (that's not an outlier)
Maximum value (that's not an outlier)

Here in this context, median is a better representative. Similarly, talk about mode. If the data is nominal type, mode makes sense. Suppose this data is in nominal scale. So in that case, median also is also not so important.

In the sense, it does not make any sense and mode is the important thing. Suppose this data is in ordinal scale, then median takes important thing, important place. But in case the data is numeric, then also you have to be taking care of this mean, median and mode and observing how the mean changes with respect to this median and mode. So, the point that is being made here, the median is less sensitive to outliers, ok, to more than mean. So thus, if you would like to understand whether your data is being skewed, skewed means tilted to a particular either to the left or to the right, then you have to be comparing mean and median.

Now next thing is the spread of the data starting from which point as the minimum value, which point as the maximum value and how it spreads. So, this is again there are many terms, many to characterize this, characterize this dispersion or spread. First one is range. So, range is the, I mean the range which is the difference between the largest and smallest value in the data set. So, range is something which is the difference between the smallest and largest value in the data set.



Univariate analysis for understanding the data

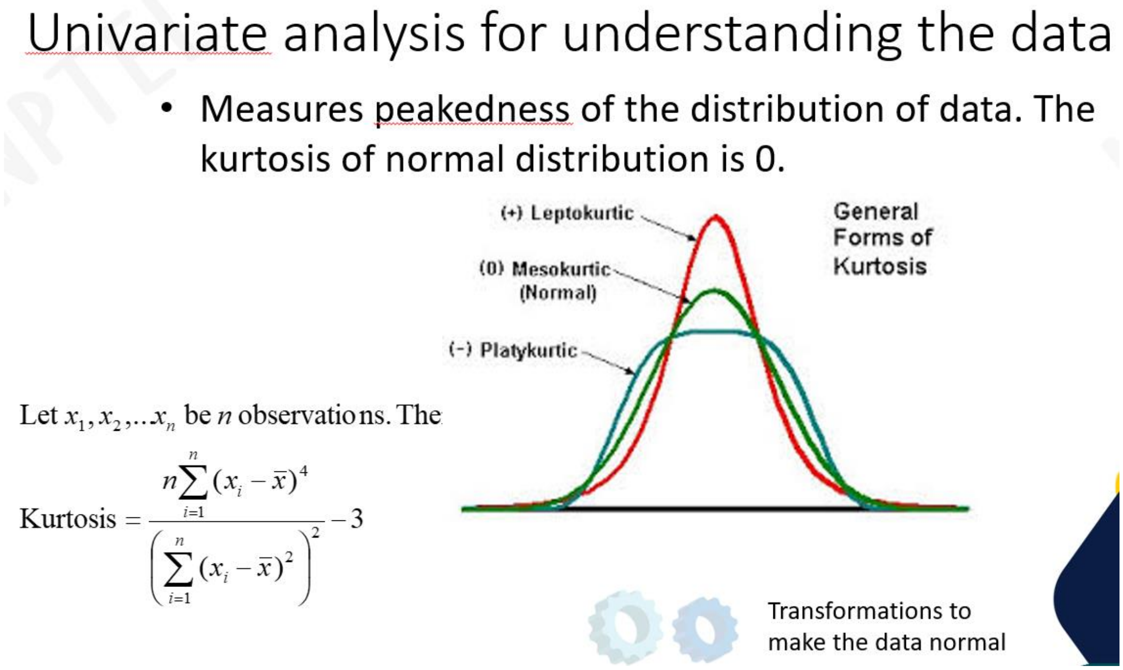
- Skewness Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let $x_1, x_2, \ldots x_n$ be n observations. Then,

$$\text{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Transformations to make the data normal

Variance. So, variance measures the fluctuation of the observation around the mean. The larger the value, the greater is the fluctuation. And standard deviation as you know it is the root over of variance. So, in this particular context, we have considered the sample variance.



Univariate analysis for understanding the data

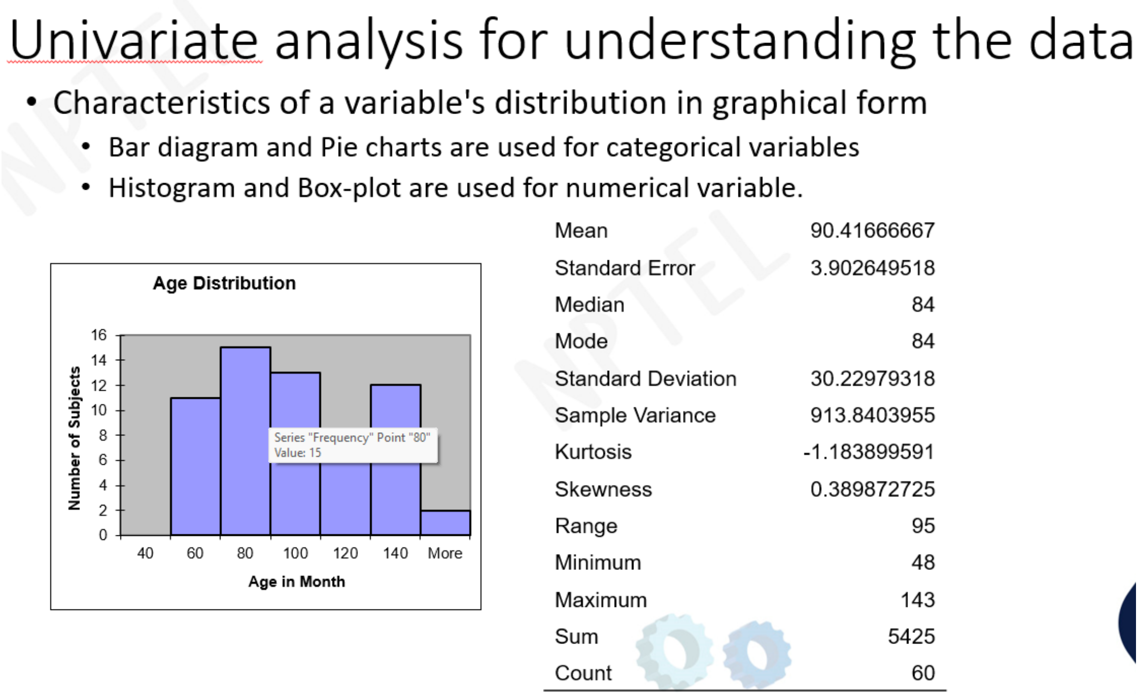
- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

Let $x_1, x_2, \ldots x_n$ be n observations. The

$$\text{Kurtosis} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}} - 3$$

Transformations to make the data normal

As you know there are two concepts. One is the population which represents the entire data about something and sample is a very small part. So, whenever we are using sample instead of using n, we will be using n minus 1. So, with this now this range that we consider that can be broken into 4 quartiles. So, this lower quartile, this is connected to range. So, this lower quartile Q 1 is the value such that one-fourth of the observation falls below it and three-fourth falls above it.



Univariate analysis for understanding the data

- Characteristics of a variable's distribution in graphical form
  - Bar diagram and Pie charts are used for categorical variables
  - Histogram and Box-plot are used for numerical variable.

Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

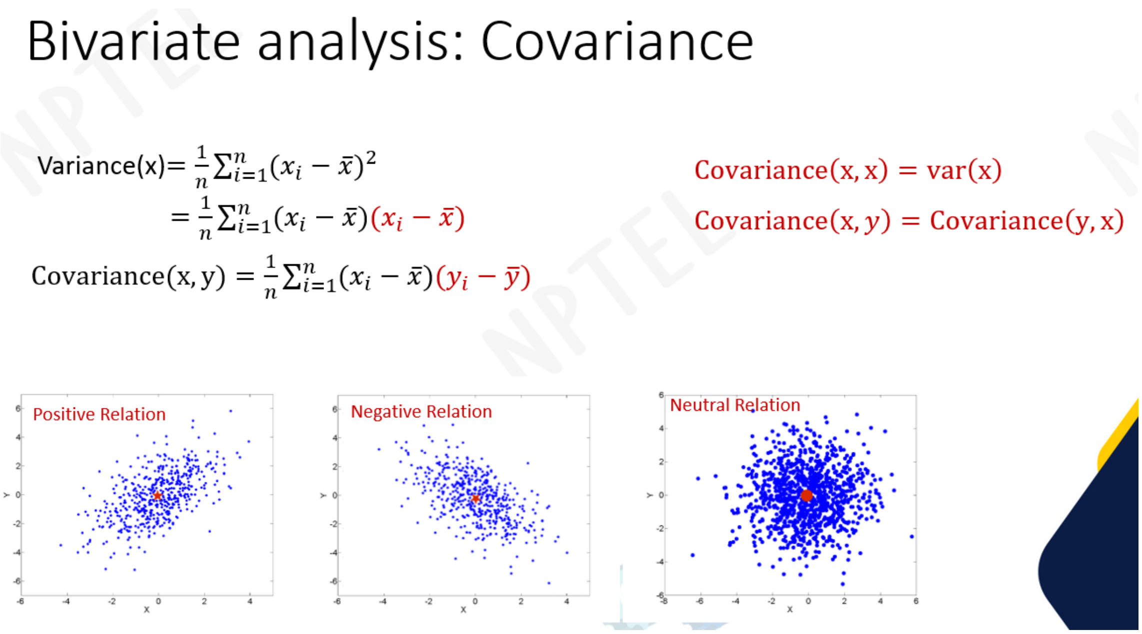
And similarly median is the middle value, Q 3 is the other part, this side Q 1, Q 3 is the other part. So, this inter quartile range is the difference between the third quartile and first quartile. So, if you look at this, this is the data, this is the minimum value, this is the median, this is the maximum value, median is the mid value. Then if you take this part of the data and make it again into two parts, this part makes your first quartile, the middle of this is first quartile and middle of this is the third quartile. And we have actually made the data into 4 different parts.

So, here the range is this, standard deviation as following that formula and various variants etcetera. Now, this is something called a box plot. When you look at the data feature wise, one of the approach in which you can have a visual representation to comprehend better is a box plot. It is a box and whisker plot.

These two sides are called whiskers. And there is a some value in the middle, there is some line, this is the median line. As you saw median is away from the mean. In this particular diagram of course, this is not corresponding to this data.

So, median is little bit away. So, to this side. So, what does this indicate? The data is been little bit skewed. So now, this is the minimum value, this is the maximum value, but there are some values which are beyond minimum and beyond maximum. What are they? They are called outliers. So, what are the outliers? Any value which is larger than Q 3 by at least 1.5 times the inter quartile range or smaller than Q 1 by at least 1.

5 times of IQR are termed as outlier. Now look at this. In case of understanding univariate data, we have considered its central tendency, its spread and while talking out the spread, we also understood that we have to figure out whether the data there are certain values where higher importance is given or for example, let us say an user's rating you are trying to find out. Let us say this is rating 1, this is this many times rating 2, this is these are the ratings and this is the frequency. How frequently he has rated? This is rating 2, then this is rating 3, 4 and 5, 3, 4 and 5. So, which means this person mostly gives a very low rating to whatever he rates. So, such kind of observation can be figured out if we study the skewness of the data.



Bivariate analysis: Covariance

$$\text{Variance(x)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

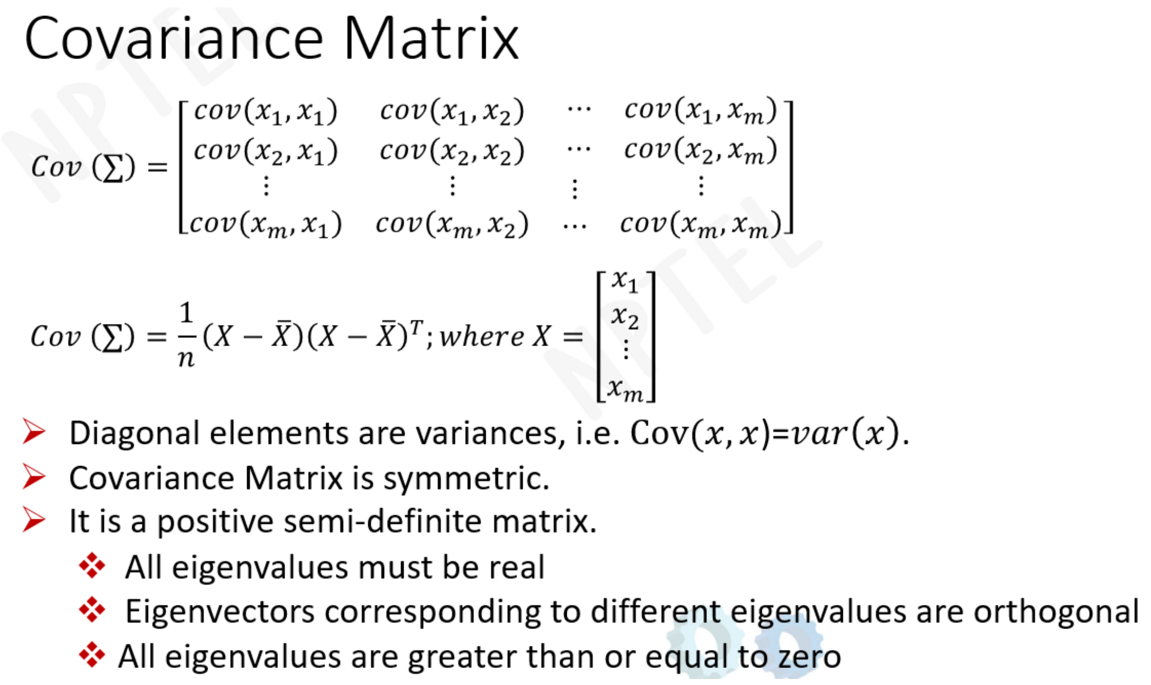
$$\text{Covariance(x, y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Covariance(x, x)} = \text{var(x)}$$

$$\text{Covariance(x, }y) = \text{Covariance(y, x)}$$

 So, skewness measures the asymmetry of the data. Of course, here the figures that you see they basically are for quantitative values like in the ratio scale or in the interval scale. But as I told you through a histogram also you can study for ordinal scale as well as for ordinal scale where the order also matters. So, now here this particular thing is called negatively skewed and this is positively skewed. So, if it is positive or right skewed you have a longer right tail and if it is negative or left skewed you have a longer left tail. So, it longer you have a longer left tail over here and if you have a longer right tail over here.

 This is the formula. Formula of course, is not important. All the things that you do probably you will be using certain software or you will be certain language where all these codes will be available to you. So, the formula anyway you will be able to. So, why exactly do you need to study? Because you will be understanding the practical aspect of how the what is the nature of the data. So, therefore, understanding all this is very important. So, how picked is your data? So, if look at this data here the spread might be same, but look at this.



Covariance Matrix

$$Cov\ (\Sigma) = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_m) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_m) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_m, x_1) & cov(x_m, x_2) & \cdots & cov(x_m, x_m) \end{bmatrix}$$

$$Cov\ (\Sigma) = \frac{1}{n}(X - \bar{X})(X - \bar{X})^T; where\ X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

- Diagonal elements are variances, i.e. Cov(x, x)=$var(x)$.
- Covariance Matrix is symmetric.
- It is a positive semi-definite matrix.
  - All eigenvalues must be real
  - Eigenvectors corresponding to different eigenvalues are orthogonal
  - All eigenvalues are greater than or equal to zero

This is very picked and this is pretty flat. So, these things are also important like let us say one person, the age let us say age of the people. You have people from almost every age. So, this is the let us say histogram representing age and these are various bins age groups and let us say this is young adult, this is young, this is let us say middle aged and this is say middle aged old and this is old. So, if we look at this the data looks quite flat. So, which means the people who are viewing your movies this is the frequency.

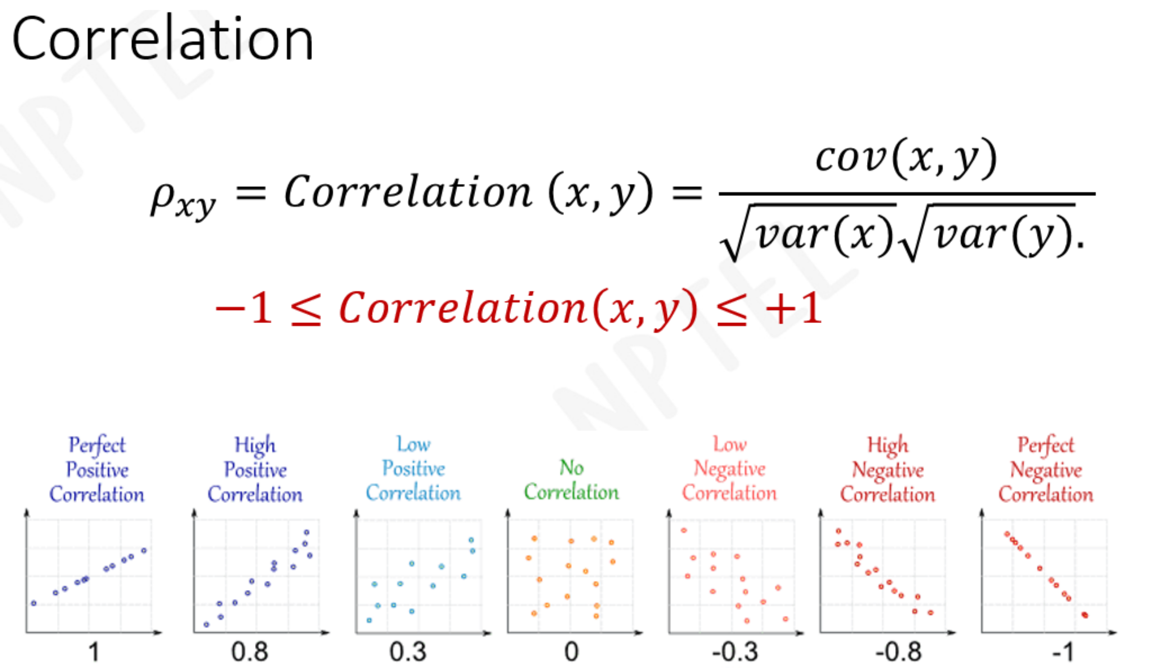
So, people who are viewing your movies belong to all age groups. Now, suppose the data would have been in a different form. Then you would have said your viewership is more for the middle aged people. So, therefore, understanding the picketness of the distribution is also important. So, what we studied so far is we can study the central tendency, then median, mode of course, standard error we did not do standard deviation, but these are the typical descriptive statistics you can find out from some software.

But finding these values is not important. What I am trying to stress is try to relate it to the real problem. Let us say you have an outlier before you apply the algorithm you have to remove the outlier. Let us say you find that all your data from viewer data is equally distributed among all the age groups. So, why not we make separate models for different age groups. So, such kind of decisions can be understood if you look at univariate data.

Look at as I told you when we talk about a data matrix, if we have the entities here and if we have attributes here the features ok. So, these features and attributes together make your data matrix. Now suppose here you have how many features you have? You have F1, F2, F3, F4, F5. So, how these features relate to each other? Now why should you try to relate the features with each other? We try to relating the features with each other because sometimes the features are too correlated which means they basically give you the same data.

So, in that case you can you should be able to drop one feature. Besides this observation they can also lead to certain computational issues that computational issues called the multicolinearity issues. So, to know about all these detail in depth probably you have to take some course on statistics specifically on multivariate statistics. But anyway that is not important ah here because most of the things you will be actually doing using some available code which your specific software or language provides language library provides. But here the point that I am trying to make is the matrix the statistics which you are trying to use to represent this relationship.

One is your covariance. So, if your matrix this data matrix let us say this data matrix there are m number of users and n number of attributes. So, to show the relationship among the attributes you will be computing covariance considering any two attribute together. So, if remember the formula for variance what did you do while trying to find out the variance it is the spread with respect to a single variable. So, the formula was like this.



Correlation

$$\rho_{xy} = Correlation\ (x, y) = \frac{cov(x, y)}{\sqrt{var(x)}\sqrt{var(y)}.}$$

$$-1 \leq Correlation(x, y) \leq +1$$

Now you have another variable y x and along with that another variable y. So, ah here when we consider two variables let us say this is variable x this is variable y and these are the data points. Looking at the data points you see what does it indicate when the variable value of x increase y increase as well. So, most of the data follow this relationship. So, as value of x increase y increases too. So, this is a positive relationship this is a oh sorry this actually probably I made a mistake.

So, that this points would have been like this this again the positive one is shown as negative point. So, this is the negative relationship suppose the points would have been like this this is the negative relationship and here you cannot say anything about the relationship. So, this is a neutral relationship. Moving ahead if we have total m number of attributes m number of total attributes then we have connecting one variable with another we have a m cross m covariance matrix ok. And what is x here x is the data matrix x bar is the individual mean and this is a basically mean centered data matrix and we find out the covariance.

When we find out the covariance the diagonal elements are relating one variable with itself. So, they are variances this covariance matrix is symmetric and it is a positive semidefinite matrix and what is a positive semidefinite matrix which satisfies this criteria all the eigenvalues must be real eigenvectors correspond to different eigenvalues are orthonormal all the eigenvalues are greater than or equal to 0. So, this relationship as we move ahead to study about dimensionality reduction this we will be knowing more about it in detail. Now, from covariance the problem with covariance is covariance can only say whether the relationship is positive or negative, but it cannot say to what degree they are related how much positive and how much negative ok. So, in addition to covariance we use another measure called correlation which is basically the normalized value of covariance and how the covariances are normalized by dividing by root over of variances.

So, this as a result of this normalization value of correlation lies between minus 1 to plus 1. So, looking at how negative how it is away from the 0 you say how much negatively they are correlated minus 1 means completely uncorrelated and this is fully correlated. This is shows perfect positive correlation, this is a perfect negative correlation and these are the values where this is highly positive, this is low positive, no correlation, low negative and high negative correlation.

So, with this we end today's lecture. These are the references. So, this if you would like to know more this is quite a good book chapter 2 probably or chapter 4 of this fundamentals of quality control in a very lucid manner it gives you the fundamentals of statistics as well the statistical background, but there are many statistics good books, but basically I draw certain things from here and Han and Kamber not the recent one old version also you can get little detail and I also got various stuffs from different contribution slides from different other authors. So, to conclude data can be structured or unstructured. Unstructured data needs to be brought to a structure form for analysis purpose, data description helps us understand a data sets main features and characteristics and it helps us identifying the patterns trends etcetera so that we can gain further insight from the data. So, univariate analysis considers one feature at a time whereas, bivariate analysis finds the relationship among the variables. With this we end this lecture. Thank you everyone. Thank you.