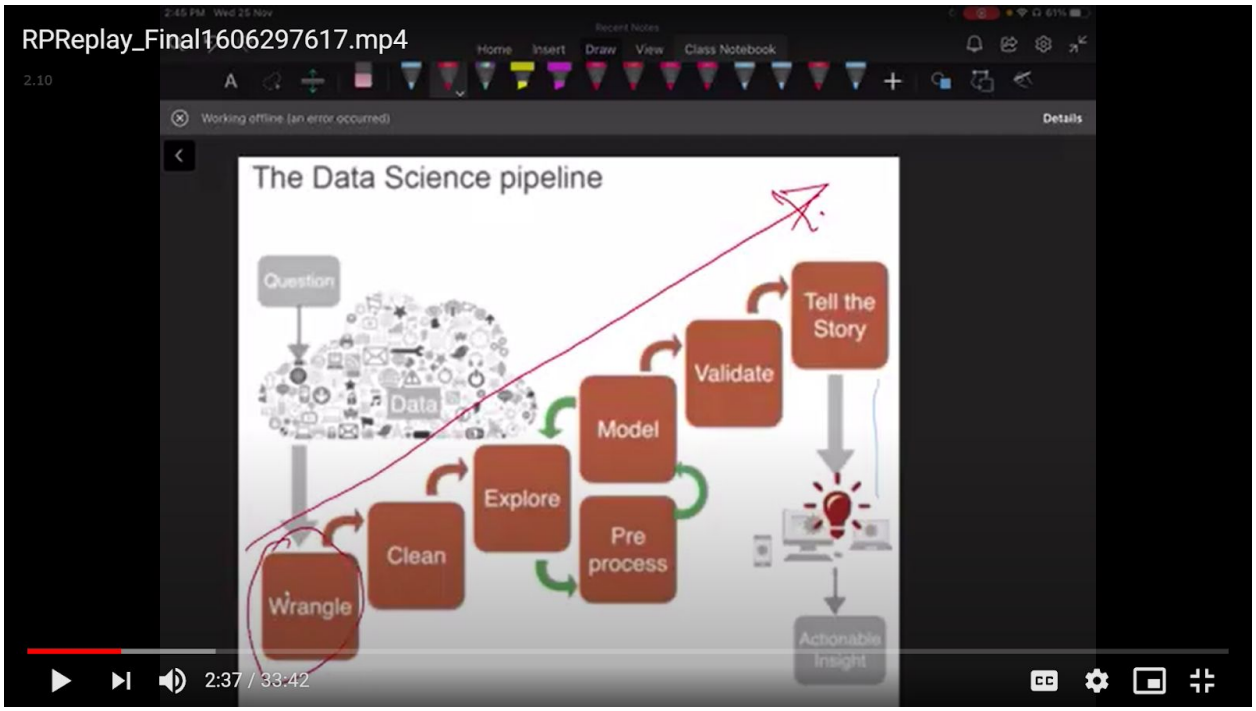
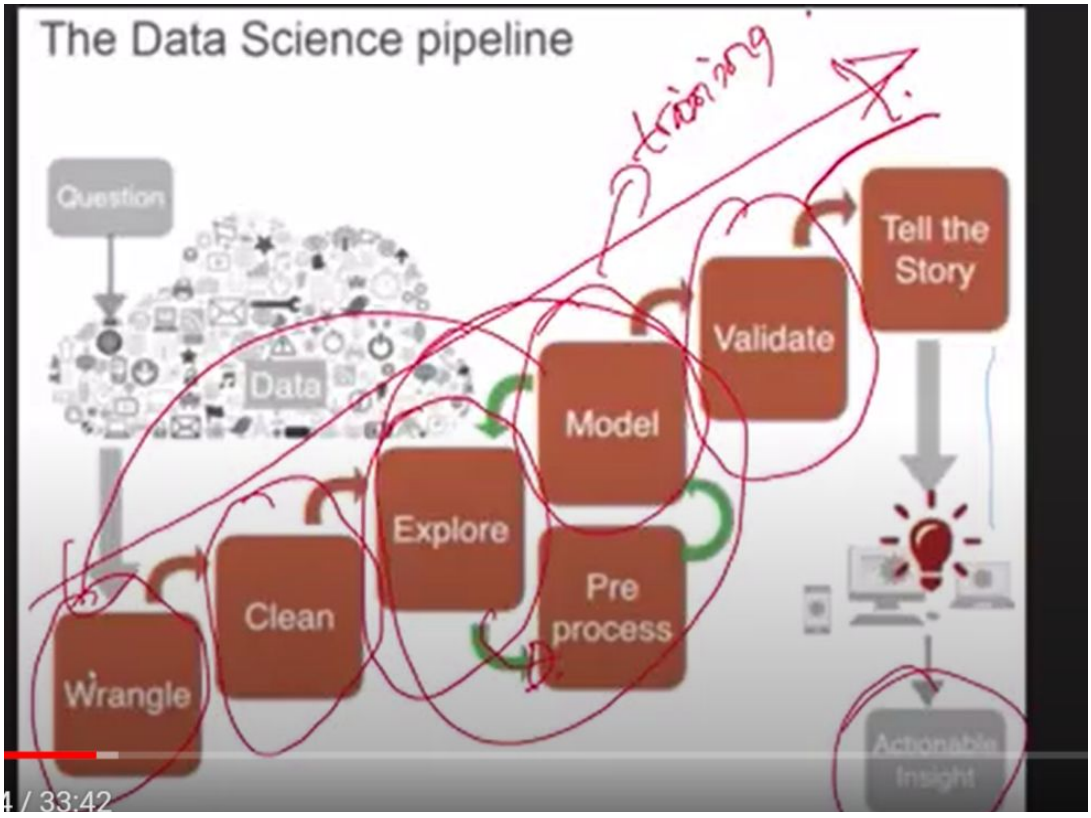
PIPELINING

Discussion on 2 case studies:

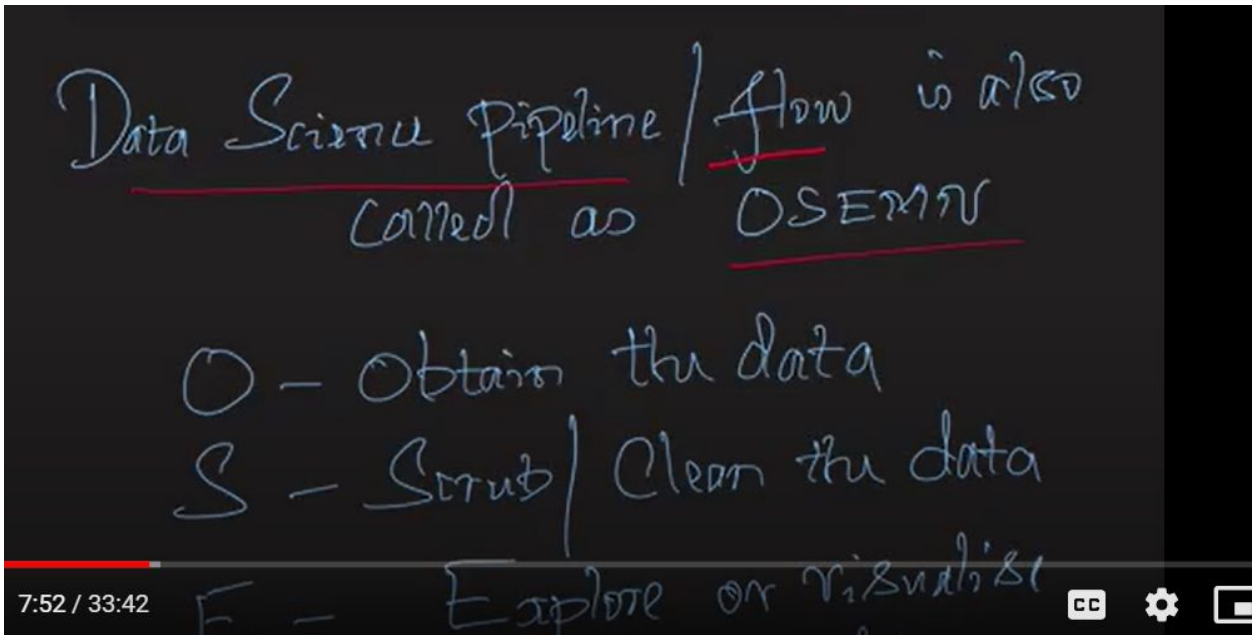
1. **Covid-19**
2. **Indian News Websites**

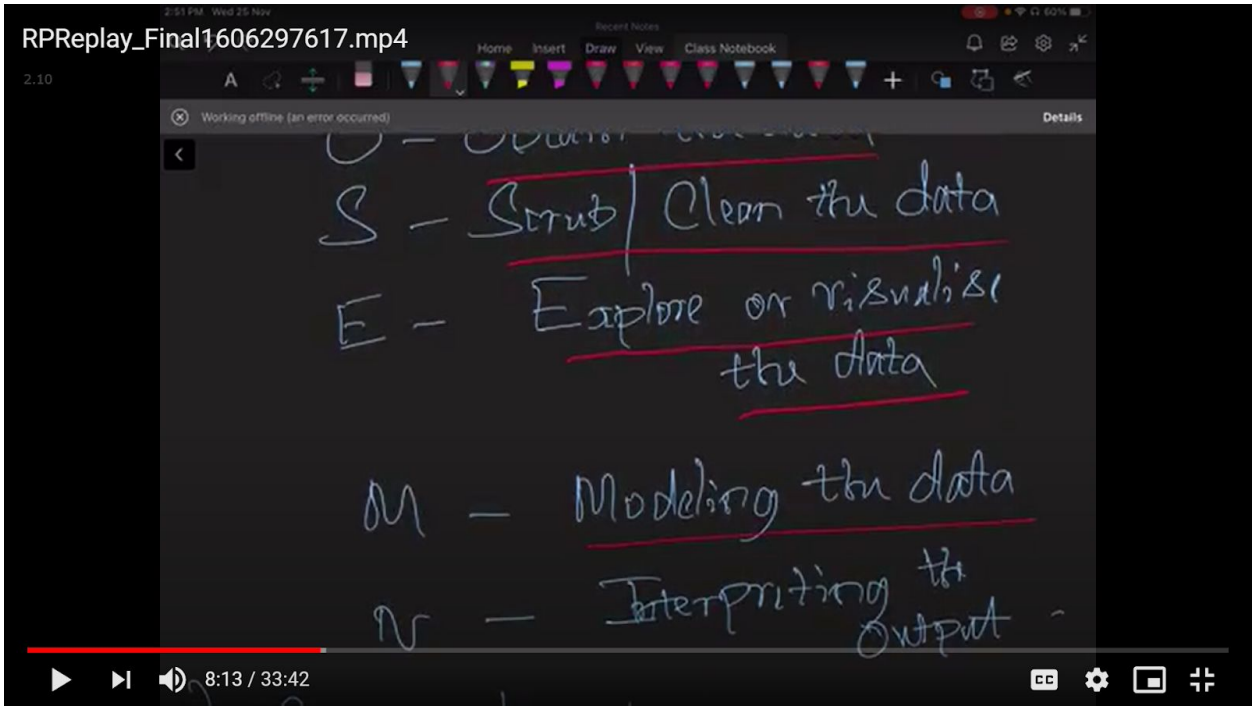


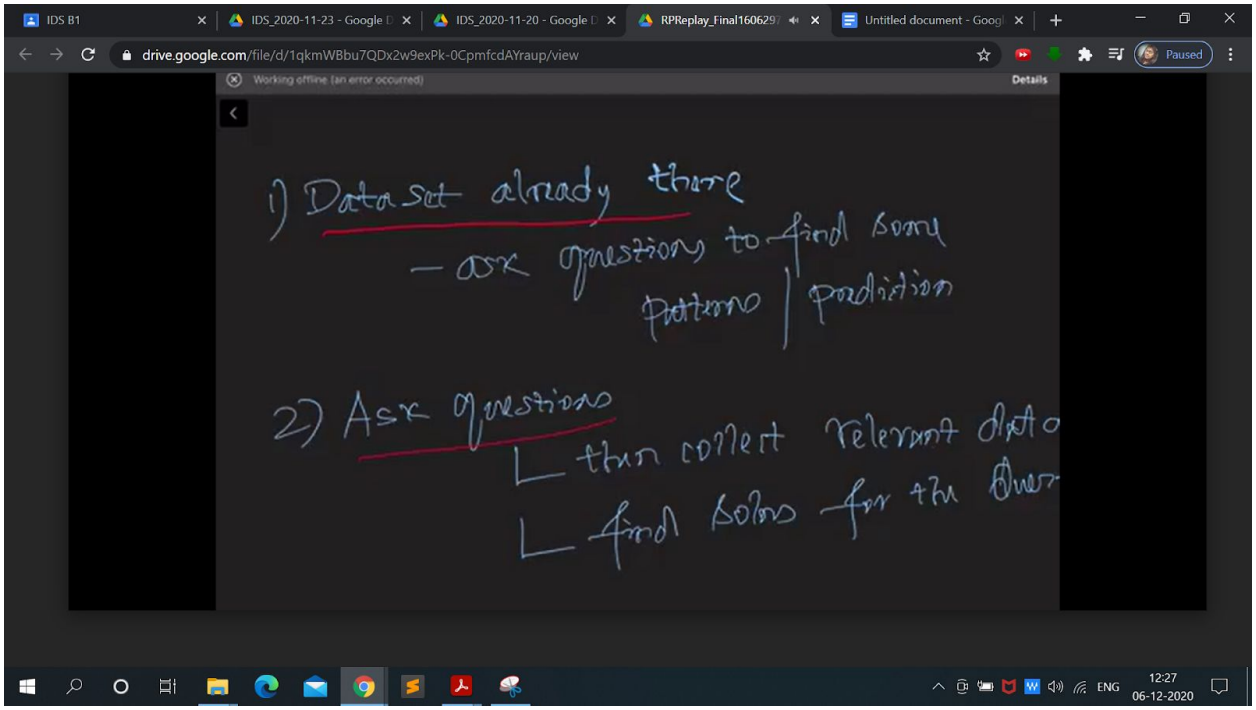
1. Wrangle: data collection
2. Data cleaning: You would be leaving some of the data or You would be modifying some of the data. More of the noise removal.
3. Explore: Descriptive analysis
   Exploration is studying about different kinds of variables, what is the range, what is the relationship between other attributes, whether there is a correlation between attributes and among which attributes, after all this you look into what kind of model you to need to build.
   Preprocessing is different from data cleaning there will be some outliers or exception cases that need to be removed.
4. Modelling(and training): Supervised learning or unsupervised learning: Classification methods: training data, testing data. Or deciding upon the clustering algorithms you will be using.
5. Validation: that your algorithm is working correctly or not. By checking the accuracy, precision and recall.
6. Tell the Story: Making out inferences of what you have done.
7. Actionable insight: Decision making kind of a thing



Though the steps are sequential it does not mean that you cannot come back. Suppose, your model is wrong and you need some more data to train it than from the modelling phase you will come back to the wrangle phase again.



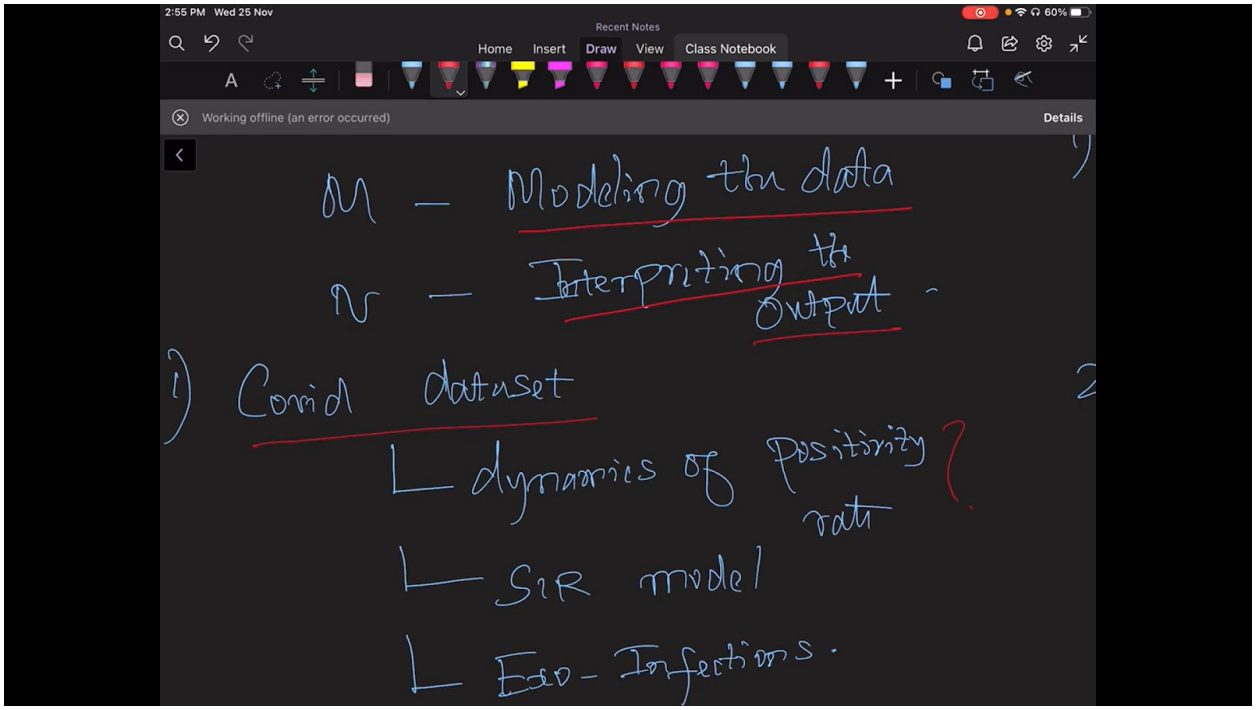


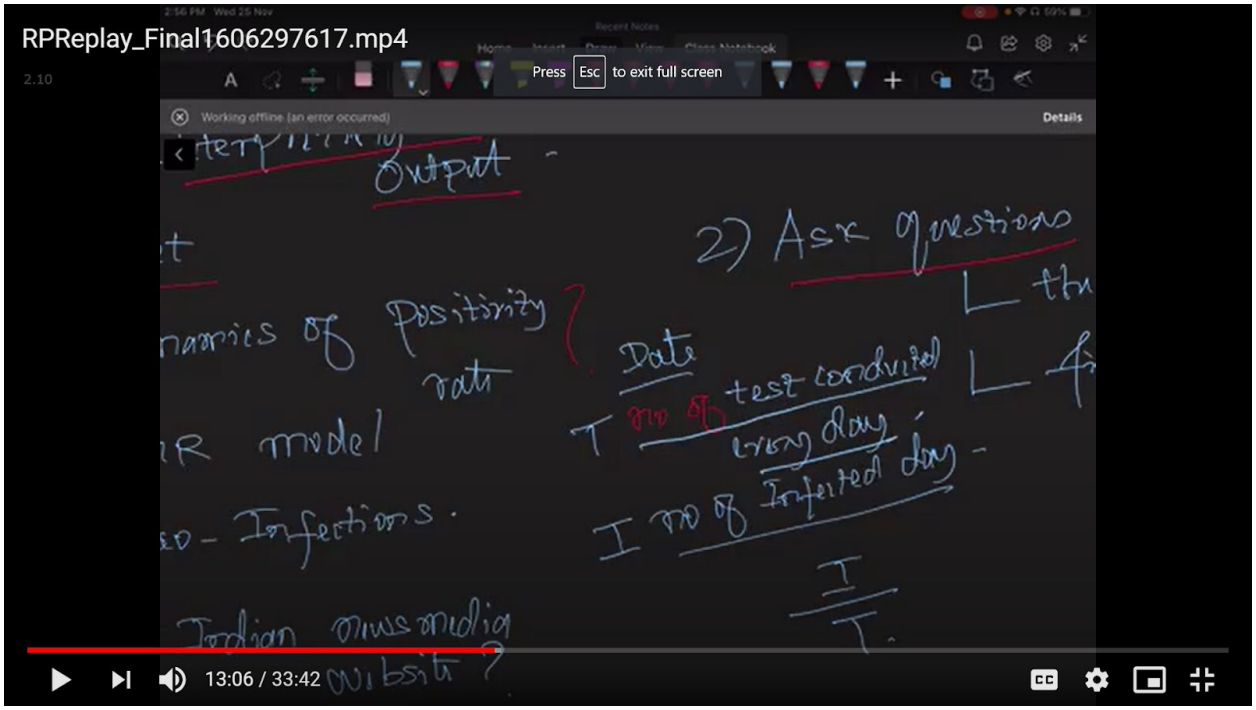


1) Dataset is already present: Just like covid data is already present.
You need to ask questions:
What are the patterns you need to look at
What are the predictions you want to look at
How many people have been infected? When will the peak will come?
You need to precisely frame your questions

2) Collect the relevant data: You are doing a kind of research.

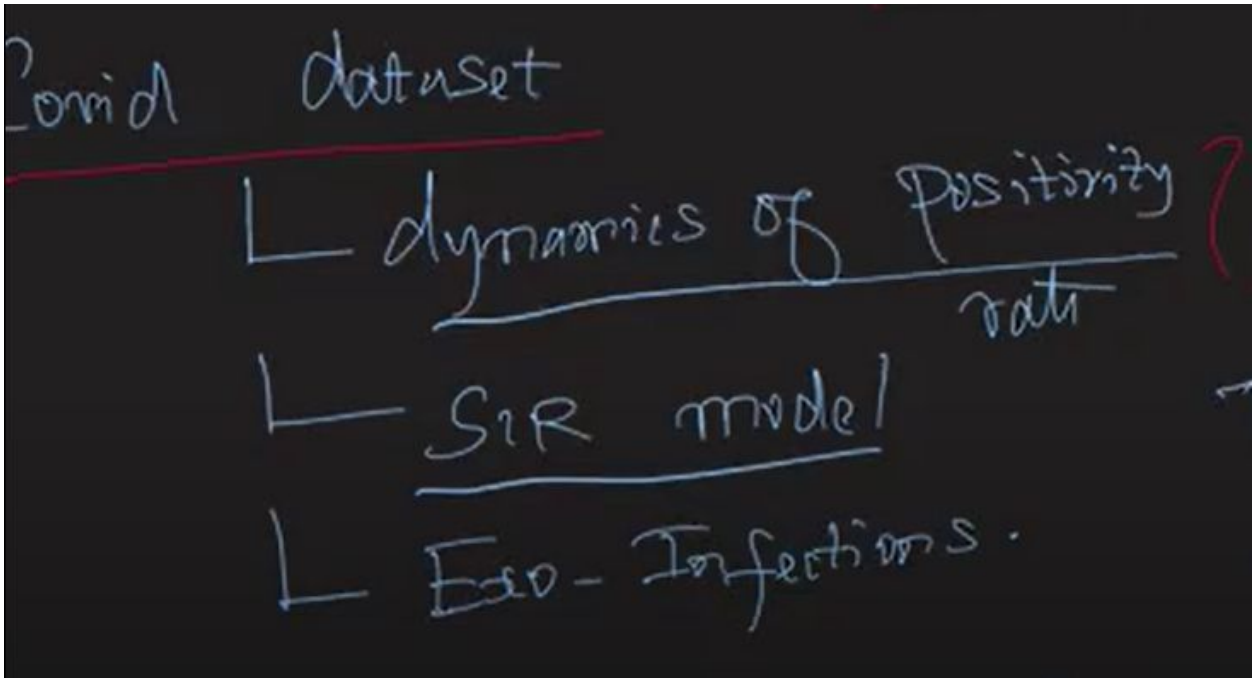
Analysing Covid data:
Questions: 1) Dynamics of positivity rate? From march to december hoe has the positivity rate changed?



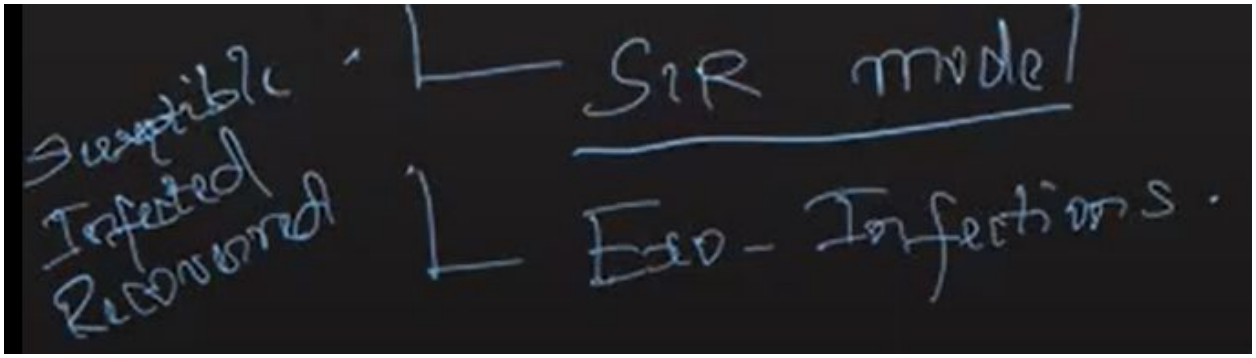


Positivity rate = I/T

You can plot this and can get various inferences out of it.
When it will be less than 5.?
WHO says that when positivity rate comes less than 5 then only your country is at safer side



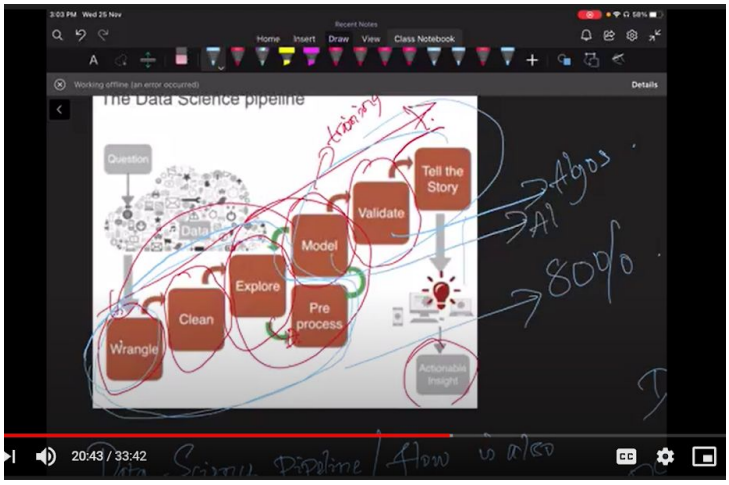
SIR Model: (a very simplistic model)
Dataset related to no. of susceptible, infected and recovered people will be required.



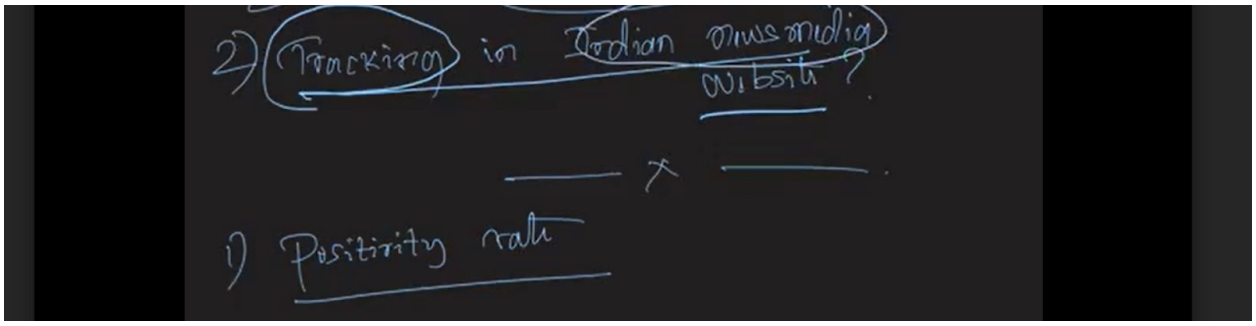
3) Exo-Infections: People in Rajasthan now, but have come from some other state or country and are infected.
For this you will have to collect data from outside also. Data from covid-19.com will not be sufficient.
Therefore, data collection is not from one place only. It can be from different sources.
Gather that data to one place. Some data maybe in image format some maybe in video format.
So, you need to convert that in one source and then do the analysis.

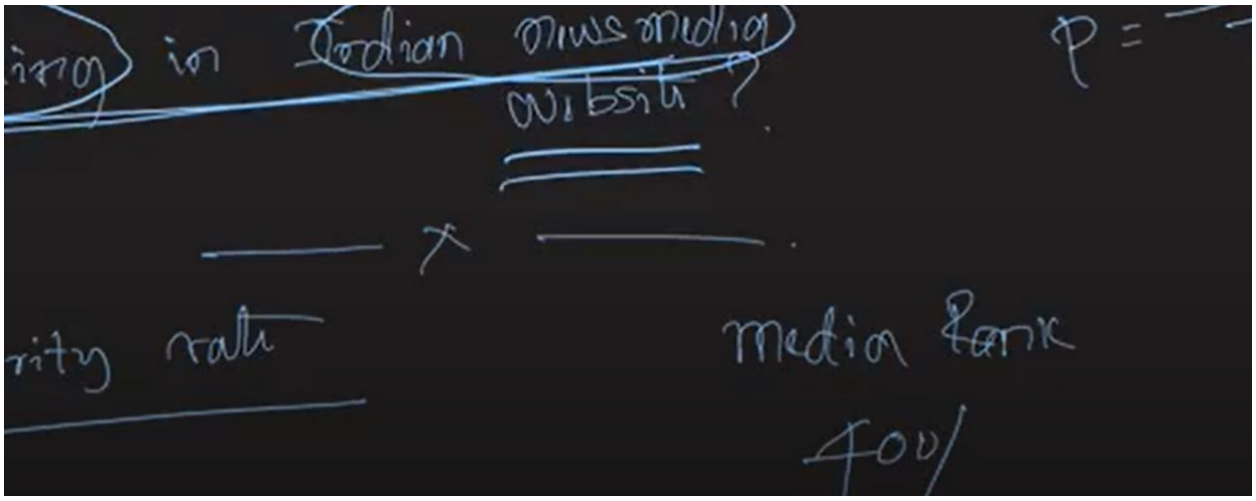
Sometimes, when data is not sufficient you will have to do data augmentation also.

80% of time takes the initial phases
Last 3 phases from modelling only take 20% of time because you have algorithms for that



Research Question:
You need to collect relevant data to answer these kind of questions





Suppose, you use alexa(amazon). It ranks the news channels. But has only 40% of the news channels only. So, you need to do data augmentation.
So, you collect the data. There are various sources: various government websites are also there which provide the ranks of news channels.

40% you got from alexa and another 60% from indian ranking websites.
Amount of tracking in Indian News websites: example
Good tracking and bad tracking of websites examples discussed!
India still does not have privacy policies in this.
Your data can be leaked because of cookies, etc.