**DS3: You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?**

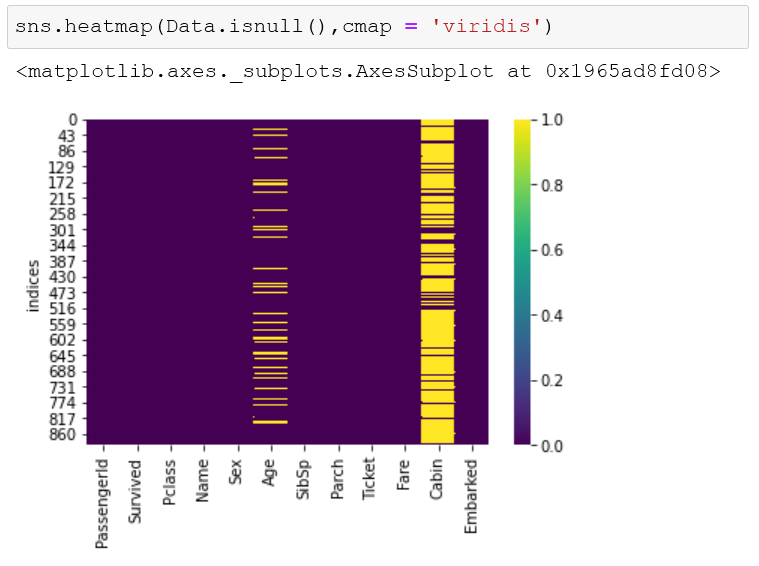
Ans:

Handling missing data depends on :

* the type of missing data
* the information that the dataset conveys
* what your aim/work with that data is (naïve bayes)
* the missing data might also be indicating some hidden information

1. **The type of missing data**

Usually I’d plot the various parameters to first understand how important those parameters are. A heat map or a bar graph will do the job just fine:



E.g.: In the titanic survivor dataset, cabin number would not be an important feature, but the age would be. So, we can drop the variable ‘cabin’ straight away.

Or we could use any the command:

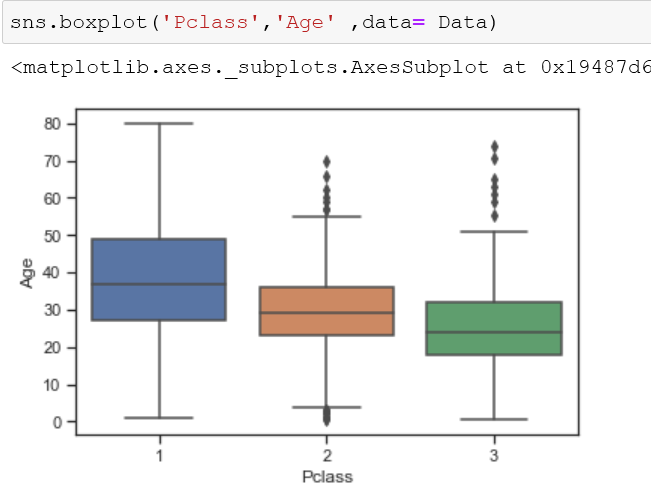


For parameter that are useful, it is fine if we drop then values given the dataset is sufficiently big. However, if otherwise, I might want to fill/replace the missing values with data.

1. **The information that the data conveys**

Filling of data depends upon the pattern of variable of which the data is missing. Which would include its own distribution or its relationship with other variables.

I would first look at the distribution of the ‘feature values’ using a box graph:



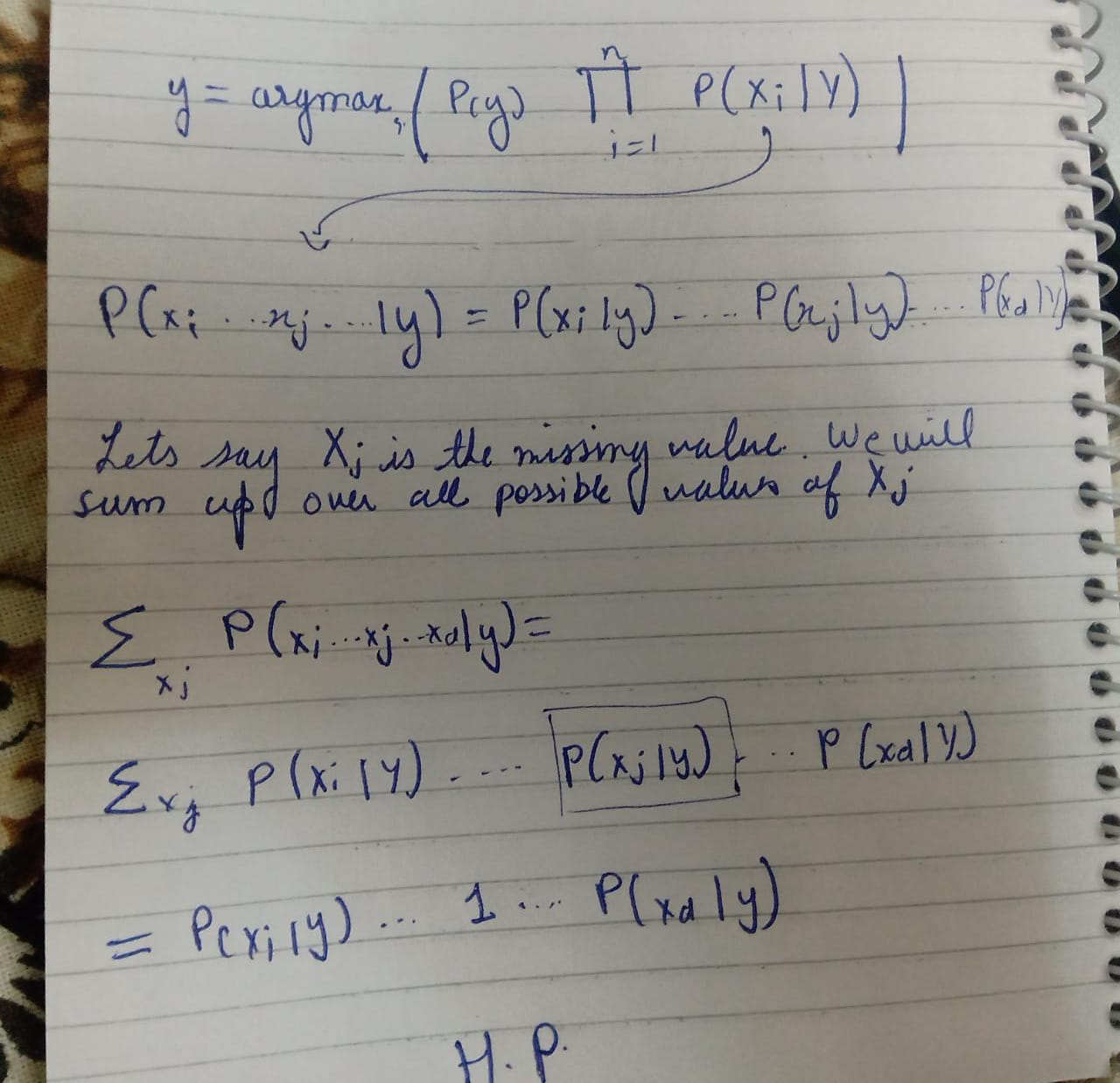
Then I would use the ‘fillna()’ function to fill the data based on trends. Either the mean of the values, or we could map the values based on the trends we see in the graphs. We could also use ffill and bfill methods



We could also use *interpolation* based on the trends. E.g. : if the data is increasing linearly, we could use forward linear interpolation. We could also use the ‘method’ parameter to a more precise interpolation of the values.

1. **What your aim/work with that data is (naïve bayes):**

Different algorithms have different properties with it comes to missing values. For e.g.: for a classification problem with a lot of missing data, I would use naïve bayes algorithms since missing values do not have any effect on the prediction. Proof:



1. **The missing data might also be indicating some hidden information:**

Missing data might also indicate some facts. For e.g:

* A particular test may not be conducted for a set of patients. (which might have some relationship with patient gender age etc)
* Or it is might indicate a socio-economic divide since a lot of time, a lot of parameters may be absent for poor people.
* It might also suggest that when data was being collected, there was an epidemic or riot etc during that time period in that area

***Therefore, data must be stored before deleting it forever.***