# Tagging Doubt Analysis

## Unacademy Doubt Analysis

**Objective**

This project involves analyzing the tagged doubts dataset of Unacademy learners. The main goal is to gain detailed insight into doubt matching, with the ultimate aim of improving our instant match. This will help to modify the existing model to increase the IM rate and provide quick and accurate solutions to learners' doubts.

Overall, this project will help the team move forward with the detailed strategy.

In this project, we have conducted exploratory data analysis on two types of cases:

- No Match : These are cases where we did not provide any solution to the learners from our Question Bank (ES Top Match).
- Correct/Incorrect Matches : These are cases where we have offered a solution to learners from our Question Bank, but we are uncertain whether the solution meets their expectations.

**Data Source**

Here is the source for the tagged doubt dataset

https://docs.google.com/spreadsheets/d/1cK8zgtRjhQLnSP5ANlnNHmwgvQwmnKYYgWvjRCpT0A8/edit?usp=sharing

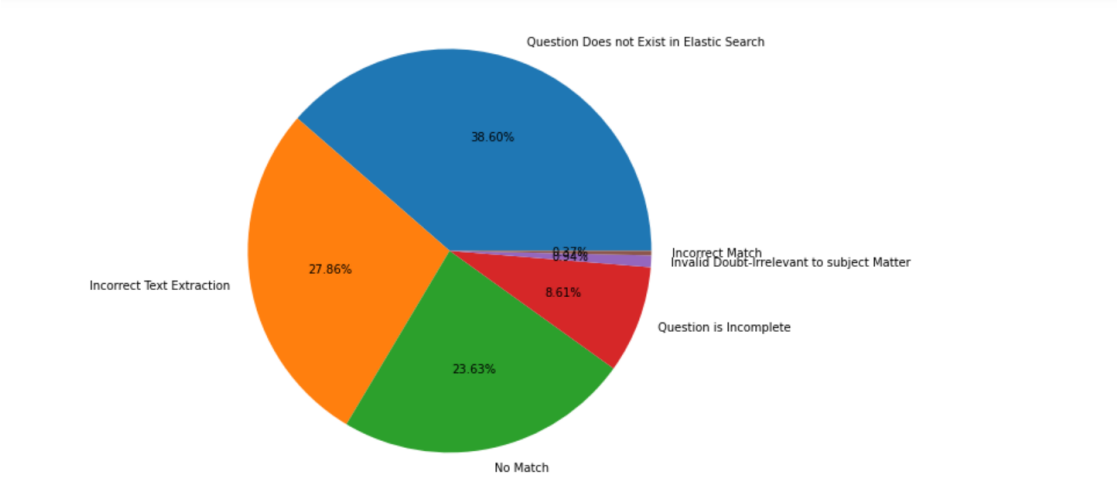
The dataset contains 3100 rows and 16 or 18 columns respectively. Out of these, 2000 rows are labeled as "No Match", and 1100 rows are labeled as "Correct/Incorrect Matches".

## “ No Match Cases “ Analysis

Reason based Distribution

```
Question Does not Exist in Elastic Search     740
Incorrect Text Extraction                     534
No Match                                      453
Question is Incomplete                        165
Invalid Doubt-Irrelevant to subject Matter     18
Incorrect Match                                 7
```

### Visual representation



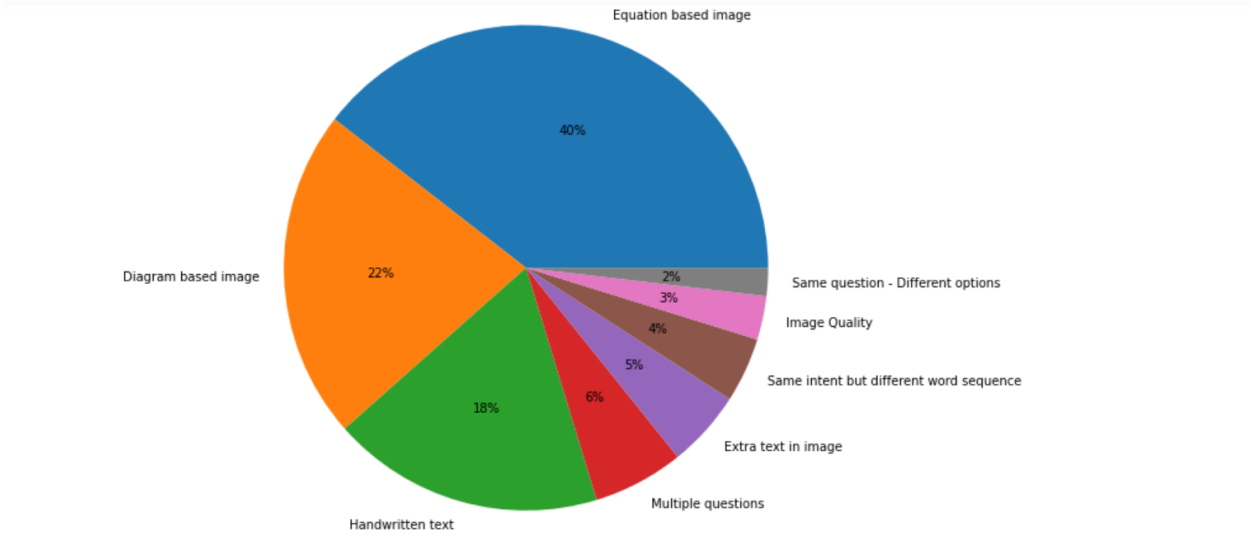

### Conclusion

Based on reason, the distribution indicates that 38.60% of data is questionable as we don't have similar questions to provide solutions to the learner's doubts. We  don’t have question bank in Elastic Search. Additionally, due to equation-based images or poor image quality, we couldn't extract the OCR of the doubts accurately, Incorrect text extraction contain 27.86%.

## Sub-Reason based Distribution

```
Equation based image                                339
Diagram based image                                 189
Handwritten text                                    156
Multiple questions                                   52
Extra text in image                                  44
Same intent but different word sequence              37
Image Quality                                        25
Same question - Different options                    16
Others - Instead get a new sub reason added          11
Different Numeric Values                              9
Object images                                         7
Same question multiple languages                      6
Same questions - Same options - Different Order       6
Random handwritten messages                           5
Selfies                                               4
Contains generic text - Data Sufficiency questions    2
Less text (small sentences)                           1
```

**Visual representation( Top 8)**




**Conclusion**

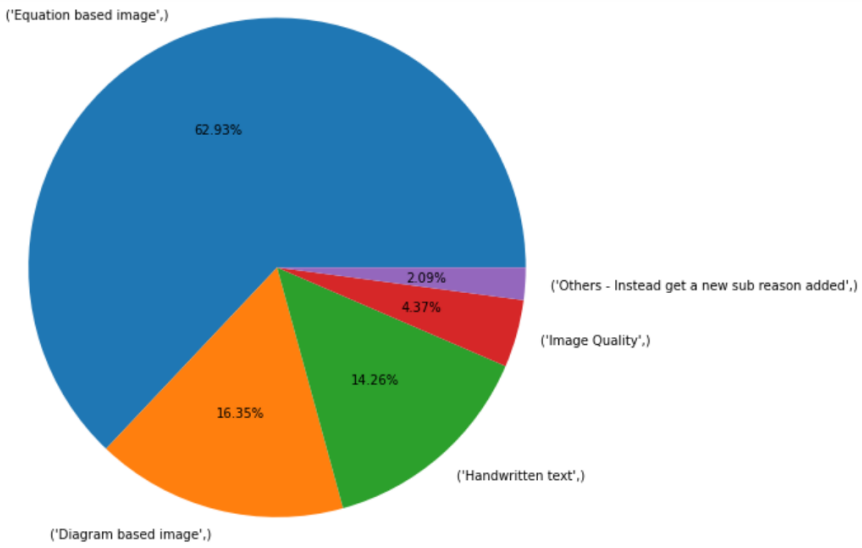
Based on the sub-reason, the distribution indicates that most of the doubts (40%) which we are receiving from learners contain equation or formula-based questions. Doubt_OCR didn't extract well. Diagram-based images (22%) also affect the top match, and the other sub-reasons are affecting respectively.

## Condition based distribution(Reason-SubReason)

what if reason = “Incorrect text Extraction” ?

```
Sub Reason
Equation based image                          331
Diagram based image                            86
Handwritten text                               75
Image Quality                                  23
Others - Instead get a new sub reason added    11
```

**Visual representation**



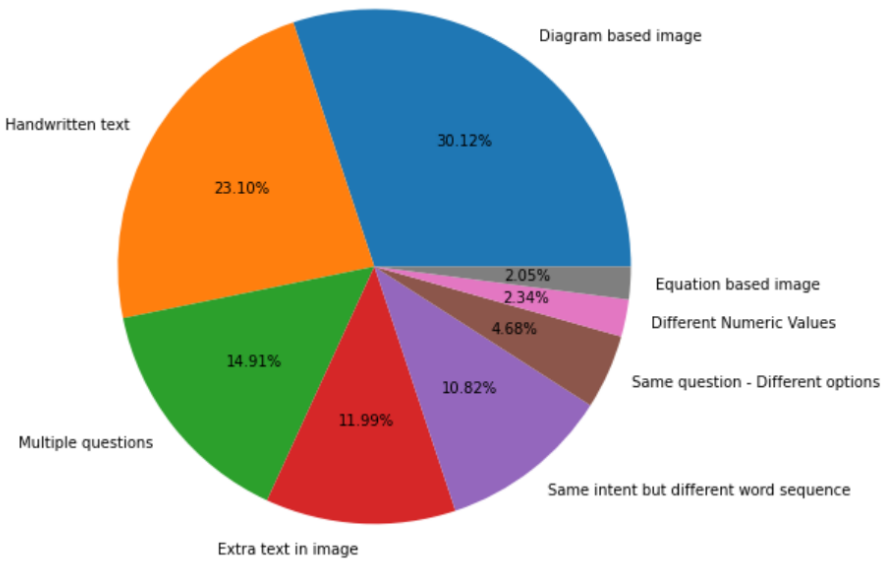

Equation-based image (62%) is a major sub-reason due to this doubt_OCR didn't extract well and we didn't get a match.

## What if reason = “No Match “?

```
Sub Reason

Diagram based image                                  103
Handwritten text                                      79
Multiple questions                                    51
Extra text in image                                   41
Same intent but different word sequence               37
Same question - Different options                     16
Different Numeric Values                               8
Equation based image                                   7
Same questions - Same options - Different Order        6
Same question multiple languages                       6
Image Quality                                          2
Contains generic text - Data Sufficiency questions     1
```

**Visual representation**



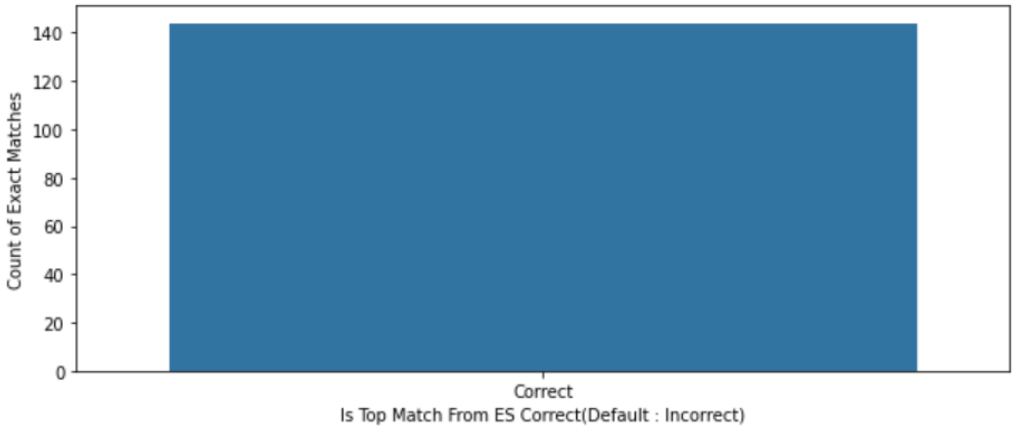

**Conclusion**

Based on the results, it appears that some of the learners' doubts have matches in ES, but not exact matches. Approximately 30.12% of the doubts contain image-based diagrams, making it difficult for us to provide an exact match. The handwritten text in the learners' doubts has also been a challenge. Although we were able to provide matches for one or two questions, due to the set of multiple questions asked by the learners, providing a match for all questions is not always possible. In cases where the doubt has the same objective but a different word sequence, we have assigned a weightage of 10.82%. Additionally, we have found that the weightage of the same question but with different options and no options is quite similar. Finally, about 2.34% of the learners' doubts have the same question and objective, but with different numeric values.

## How many cases have a top match even though we didn't give a match?

What if reason = “No Match “ and top match from ES =”Correct” ?

`144`



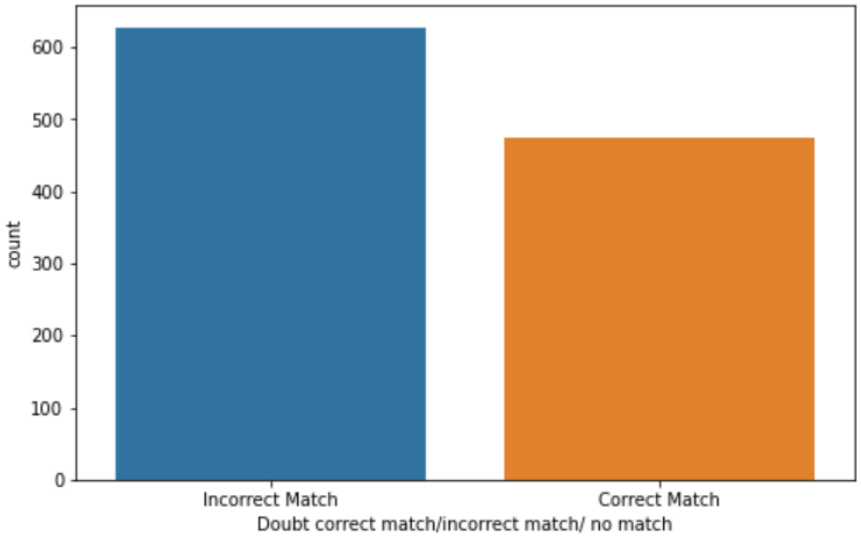

We have observed that total 144 cases among 2000 in which we get a exact match from top ES.

## "Incorrect Match / Correct Match" Cases Analysis

These are cases where we have provided a match, but are unsure whether it is correct or not.

Count for Correct/Incorrect match:-

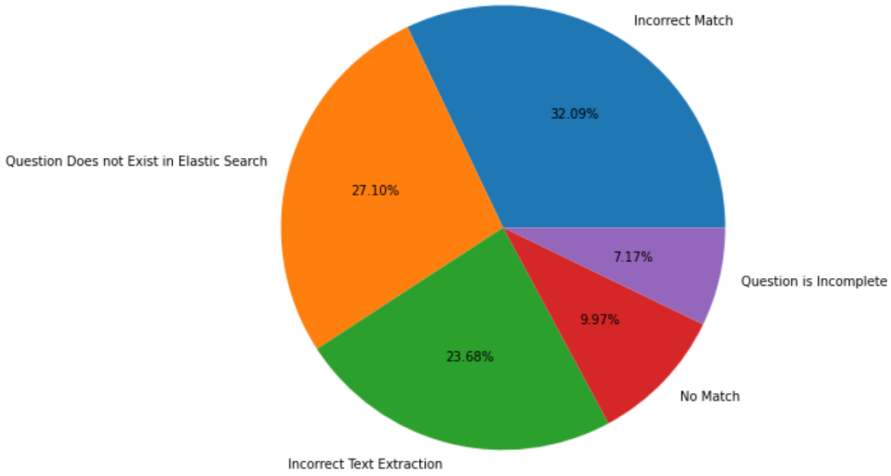
```
Incorrect Match    625
Correct Match      475
```



## Reason Based Distribution

```
Incorrect Match                              206
Question Does not Exist in Elastic Search    174
Incorrect Text Extraction                    152
No Match                                      64
Question is Incomplete                        46
```

### Visual representation




### Conclusion

Based on the above visualization, we can see that the match provided to learners is not an exact match, accounting for 32.09% of the whole tagged data. In addition, 27.10% of the learners' doubts do not exist in Elastic Search.

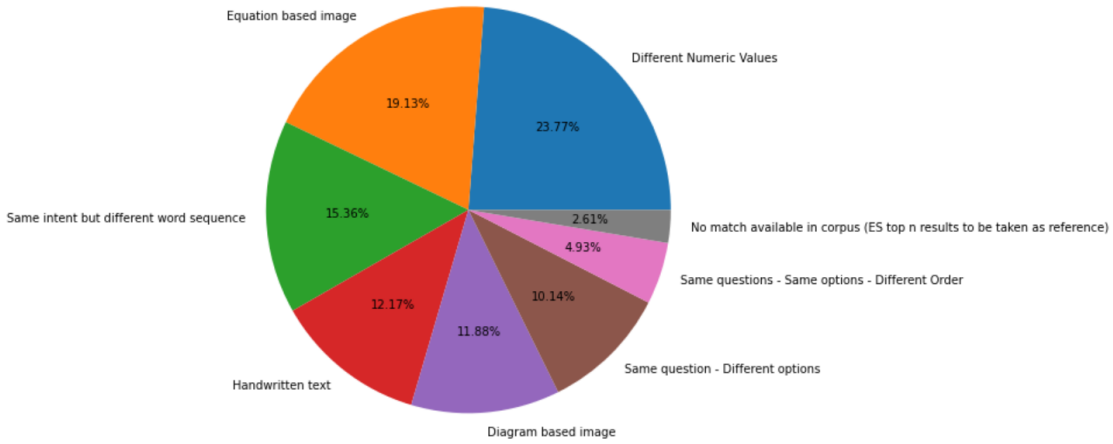
Due to poor text extraction, we have incorrectly given matches (23.68%). Here, "No Match" indicates that when the parent question did not have a match, ES provided matches that were close enough but not an exact match. No match represents 9.97% of the whole tagged data.

## Sub-Reason Based Distribution

```
Different Numeric Values                                    82
Equation based image                                        66
```

```
Same intent but different word sequence                                  53
Handwritten text                                                         42
Diagram based image                                                      41
Same question - Different options                                        35
Same questions - Same options - Different Order                          17
No match available in corpus (ES top n results to be taken as reference)  9
Multiple questions                                                        8
Extra text in image                                                       7
Same question multiple languages                                          4
Image Quality                                                             3
Contains generic text - Data Sufficiency questions                        3
Others - Instead get a new sub reason added                               2
```

### Visual representation




### Conclusion

From the above visual,we found that we almost have similar weightage for different numeric values and equation based image which means we had given the similar kind of match with same objective but numeric values are different. Most of the doubts contain equation based image. But these are only based on sub-reasons. Therefore, we cannot conclude a solution, and we need to consider more conditions to draw insights.

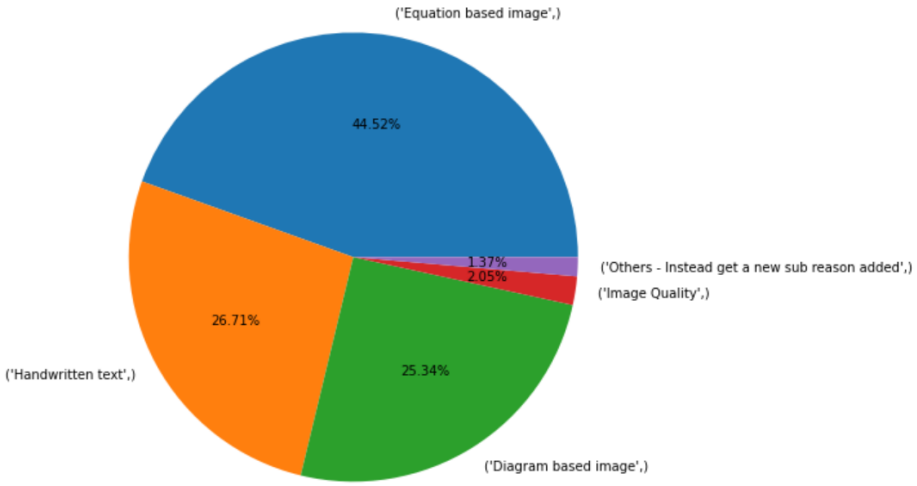
## Condition Based Distribution(Reason-Subreason)

what if reason = “Incorrect text Extraction” ?

```
Sub Reason
Equation based image                        65
```

```
Handwritten text                                39
Diagram based image                             37
Image Quality                                    3
Others - Instead get a new sub reason added      2
Extra text in image                              1
Multiple questions                               1
```

### Visual representation




### Conclusion

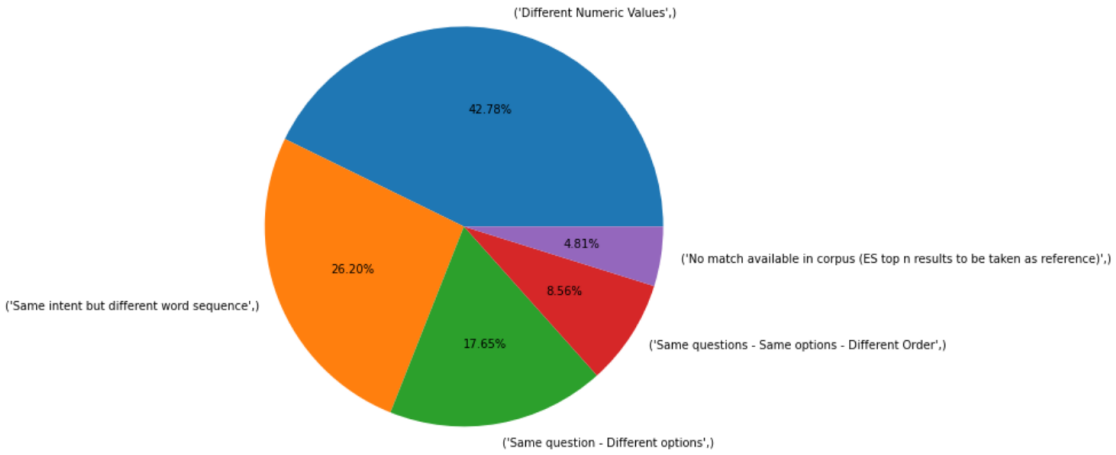
From the above visual, we can see that the major reason for poor extraction is equation based image and then handwritten text and so on. Majority has involve Equation based-Handwritten).

We need to focus on improving the model that extract equation question well.

## What if reason = “Incorrect Match” ?

```
Sub Reason
Different Numeric Values                                                  80
Same intent but different word sequence                                   49
Same question - Different options                                         33
Same questions - Same options - Different Order                           16
No match available in corpus (ES top n results to be taken as reference)   9
Multiple questions                                                         6
Contains generic text - Data Sufficiency questions                         3
Extra text in image                                                        2
Same question multiple languages                                           1
```

### Visual representation



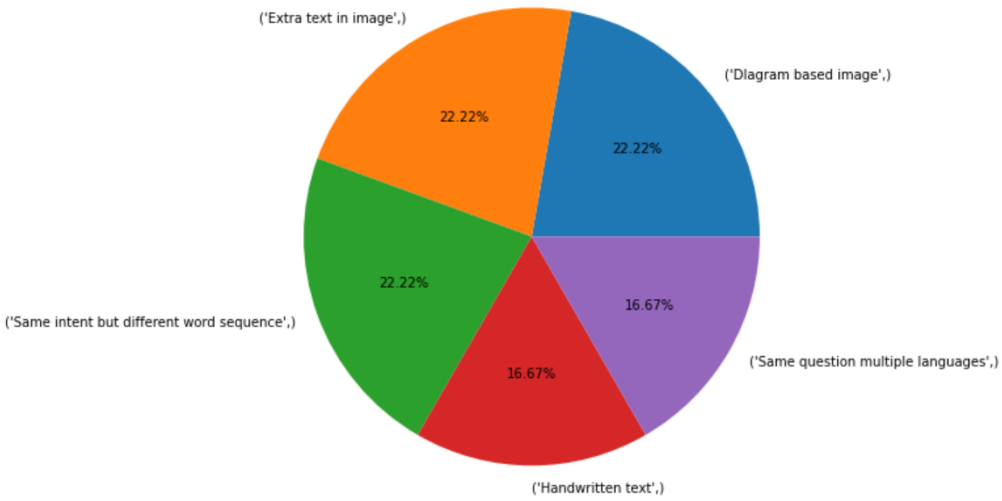

### Conclusion

As we already concluded the insight from sub-reason based distribution. We have majority of the weightage has Different numeric values and same intent but different word sequence(68.98 %) which means we have given the doubt match with same objective but text are jumbled and values are different.

## What if reason= “No Match” ?

```
Sub Reason
DIagram based image                              4
Extra text in image                              4
Same intent but different word sequence          4
Handwritten text                                 3
Same question multiple languages                 3
Different Numeric Values                         2
Same question - Different options                2
Multiple questions                               1
Same questions - Same options - Different Order  1
```

### Visual representation

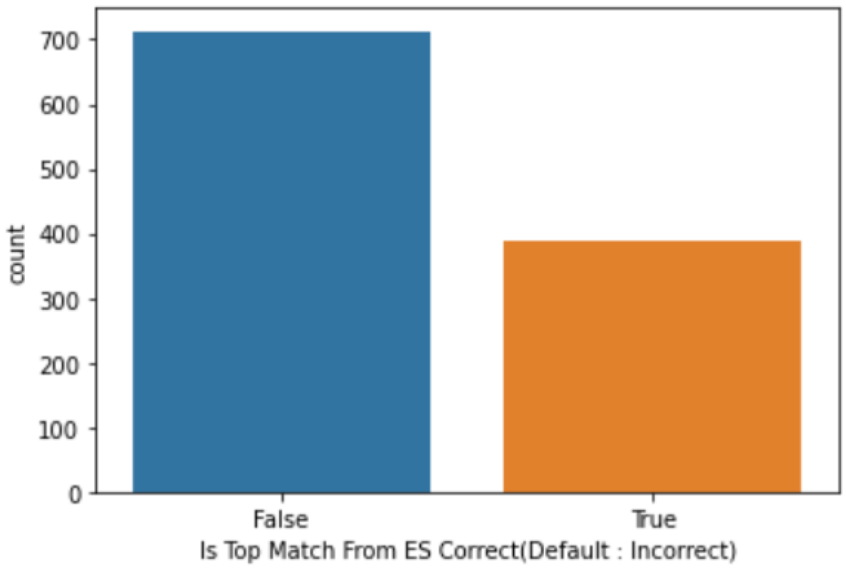



## How many cases have a top match from ES even though we have given the match(parent question)?

What if reason is “Incorrect Match”,”Incorrect text extraction”,”No match” and top match from ES =”Correct” ?

Total such cases found = **389**

**Visual**


700
600
500
400
300
200
100
0
count
False
True
Is Top Match From ES Correct(Default : Incorrect)