Practical Data Analytics
-- A Quick Guide to Generating Insights from Data

-- Mrityunjay Kumar & Gunjan Gupta

There is a lot of mystique around data analytics with buzzwords like Big Data, Predictive Analytics, Exploratory Analytics, Data wrangling, Data warehouse etc. doing the rounds. For beginners some of these can be intimidating because they seem to give the impression that a lot of theory needs to be understood and covered before any practical application of analytics can be made.

While it is indeed true that Data Analytics is a science with deep theory and requires significant preparation, what is also true is that a lot of analytics can be done using simple tools, basic maths and some analytical sense. In this paper we present a use case which showcases some simple techniques which can be used to generate powerful insights.

Use Case

The use case that we will cover in this paper is about assessments data. A publisher has created a huge question bank, which it gives out to schools along with a platform to administer tests. Schools conduct tests using questions from the question bank. All student attempt data is recorded in the platform.

Given the above use case we will go about finding meaningful insights from the data set. We are following the techniques of exploratory data analysis as opposed to hypothesis driven analytics. Here we are not starting with any preconceived hypothesis in mind, rather the approach is to start an exploration into data and discover something about it, without knowing before hand what it is we are looking for. Typically such an exercise throws surprises and insights which are least expected.

Step 1. Identify the data points that are present

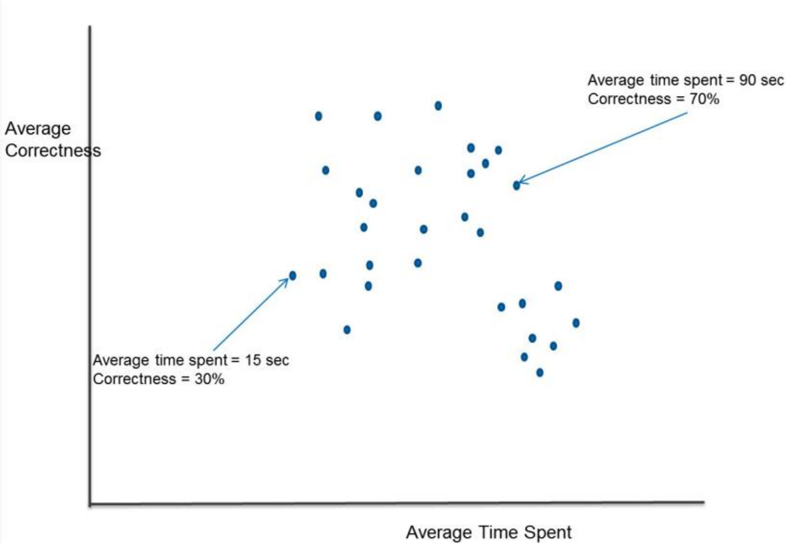
The assessment data set has recorded details of each and every attempt by the students. Following data points are present :-

1. If the question was attempted in the test
2. Was the question answered correctly?
3. How much time the student took in answering the question?
4. What was the question type (single choice, multiple choice, fill in the blank etc.)?
5. Did the question have a hint associated with it? Did the student take the hint?
6. Details about the student who attempted the question -- class, grade etc.

7. Special tags associated with the question

Step 2. Identify two metrics which seem important and start analysis with them

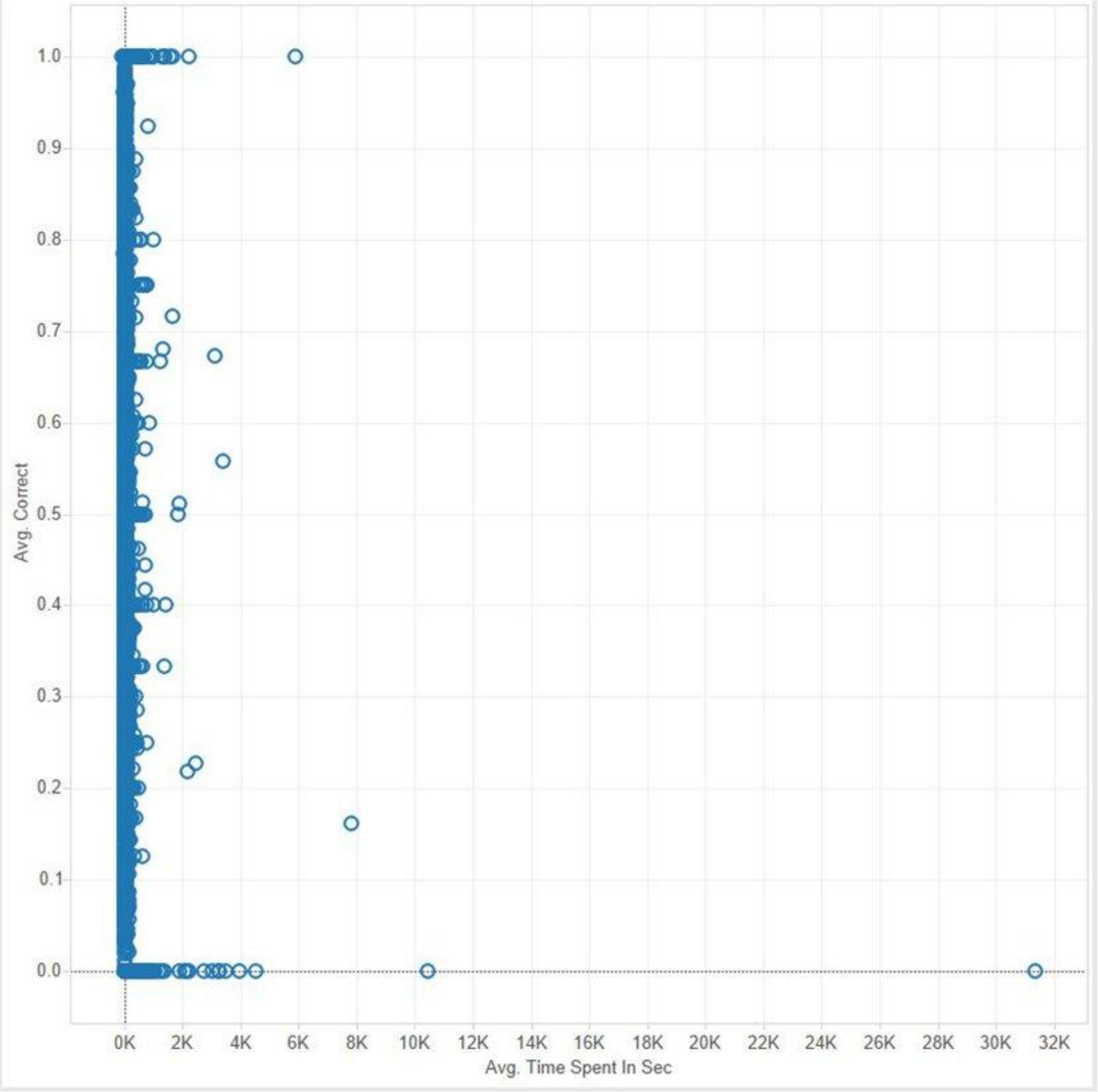
The idea of starting with two metrics as opposed to a single metric is that with two metrics, their interplay throws open many possibilities of analysis. For example in the above use case the two metrics we selected were average correctness of the question (of all attempts by all users, the average times the question was answered correctly) and the average time spent in answering the question. Now given a question with correctness 0.7, the only thing we know about it is that it is a relatively easy question because most of the time students answered it correctly. However, if we also include the property average time in answering it then two questions with same correctness but different answering times can indicate different things. Below is what a two dimensional plot of questions might look like.



Another advantage of choosing two metrics is that it is possible to easily visualize them on a two dimensional graph. If more than two metrics are chosen visualization can become very difficult and hence inferring any meaning out of that graph becomes impossible. For example with 3 metrics a 3D graph needs to be drawn which is not possible to visualize on a screen or paper.

Step 3. Plot all data points

Once the data points are plotted on a graph then it might turn out to be different from what was expected. One primary reason for this could be presence of outliers or data points which are so extreme that they distort the visualization of remaining data points. For example in the below plot of questions there are some questions with time spent so high that all the other questions have collapsed into a line thus hiding information. These outliers could be because they were test data added to check the robustness of the system, or they could occur because some extreme conditions occurred like student leaving the assessment midway to grab a coffee.



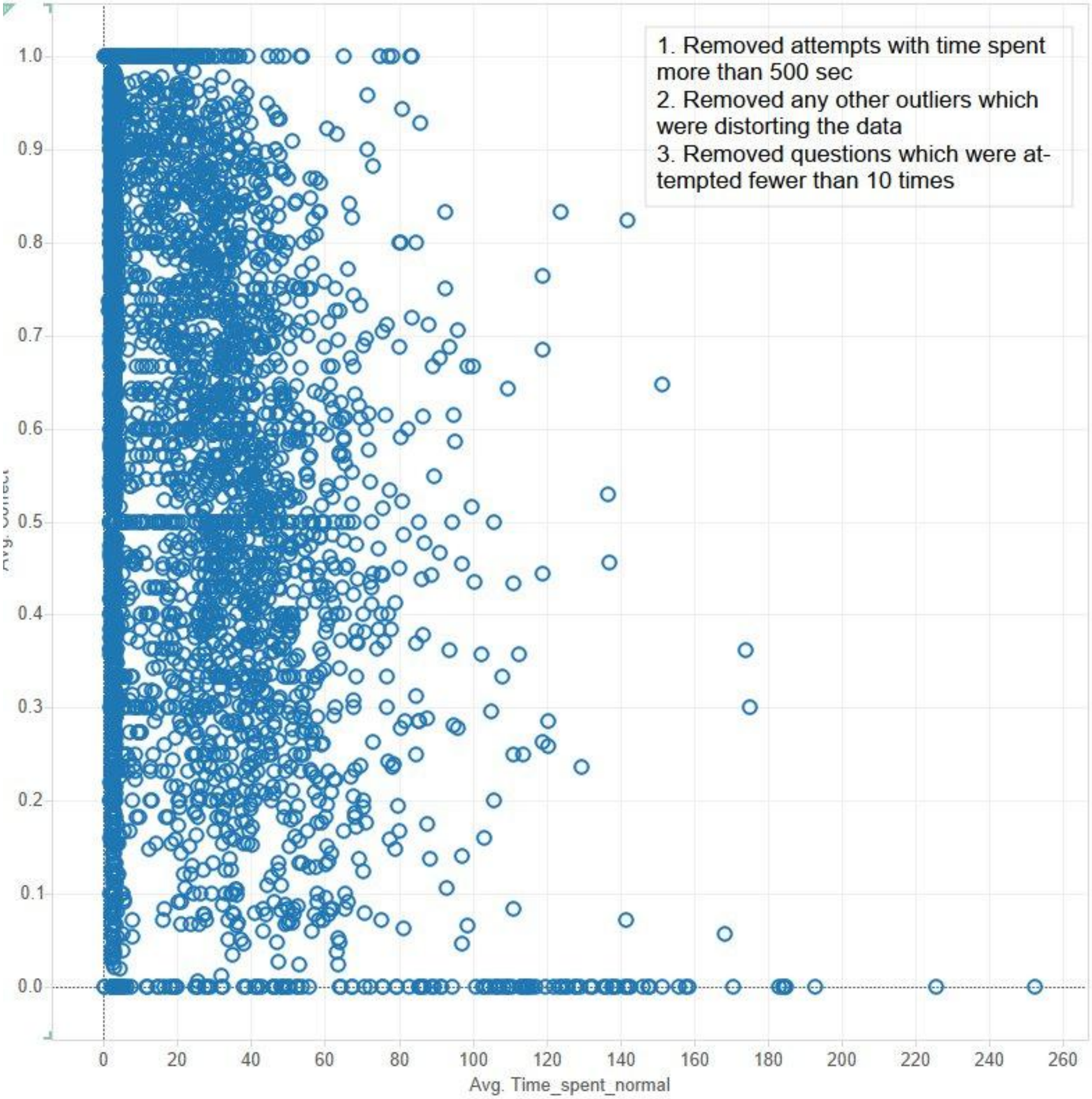
Step 4. Clean data and apply appropriate filters

This is the process which does not have any well defined rules. One has to go with instinct and apply reasoning to see if the visualization has reached a stage where it can provide meaningful results. Typically the steps undertaken are

1. Remove outliers which look improbable -- in our use case we excluded all data points with time spent on a question being more than 500 sec.
2. Only include the data points which provide meaningful information -- if averages are being taken then there should be sufficient events to make the average meaningful. For

example if a question is attempted only twice then the average is not indicative. In our use case we removed all questions which were attempted less than 10 times.

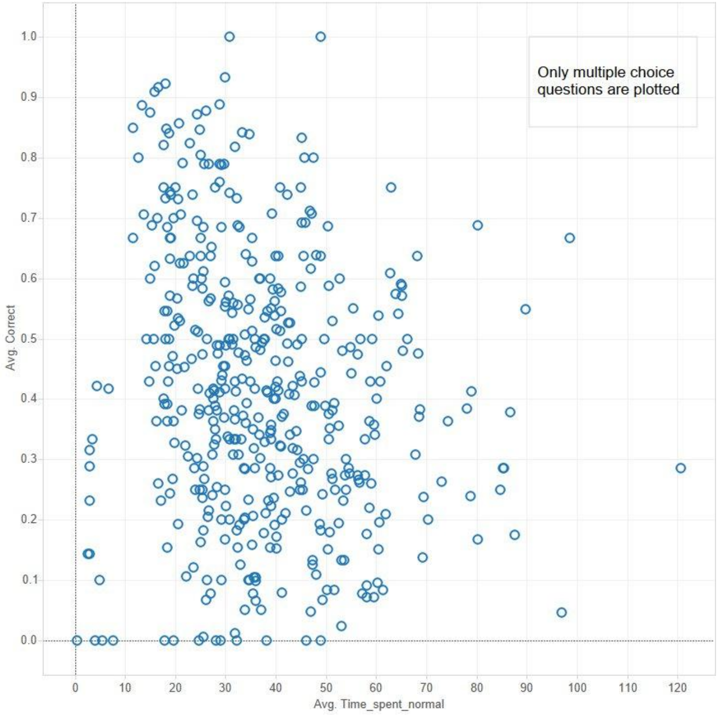
The graph after performing a series of filters looks like this.



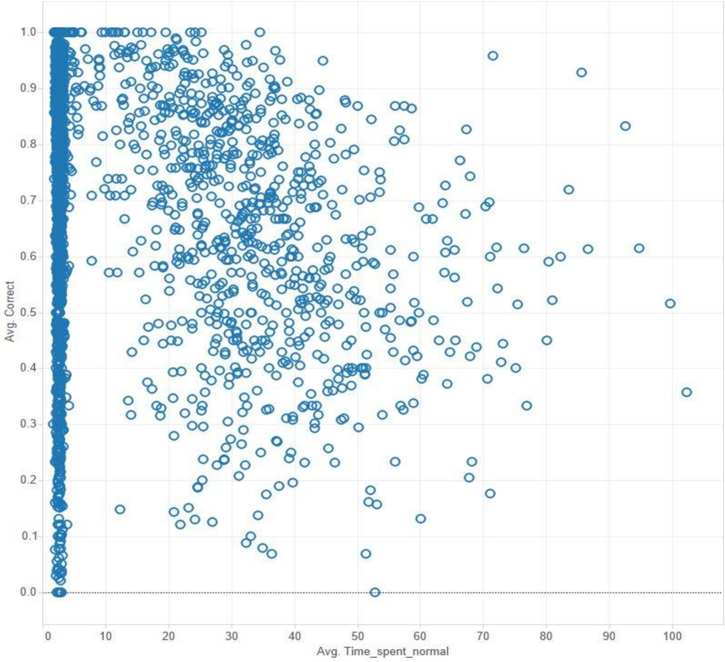
Step 5. Correlate with other metrics

Even though we started off by saying that we should concentrate on only two metrics, at this point it makes sense to see if there is some other metric which could be influencing the data. In our use case one interesting metric is the question type. So a question can be of type single choice, multiple choice, descriptive, fill in the blanks etc. It turns out that the relationship graph between correctness and time spent is vastly different for different question types. This means that the relationship graph needs to be analyzed separately for each of the question types.

Multiple choice questions

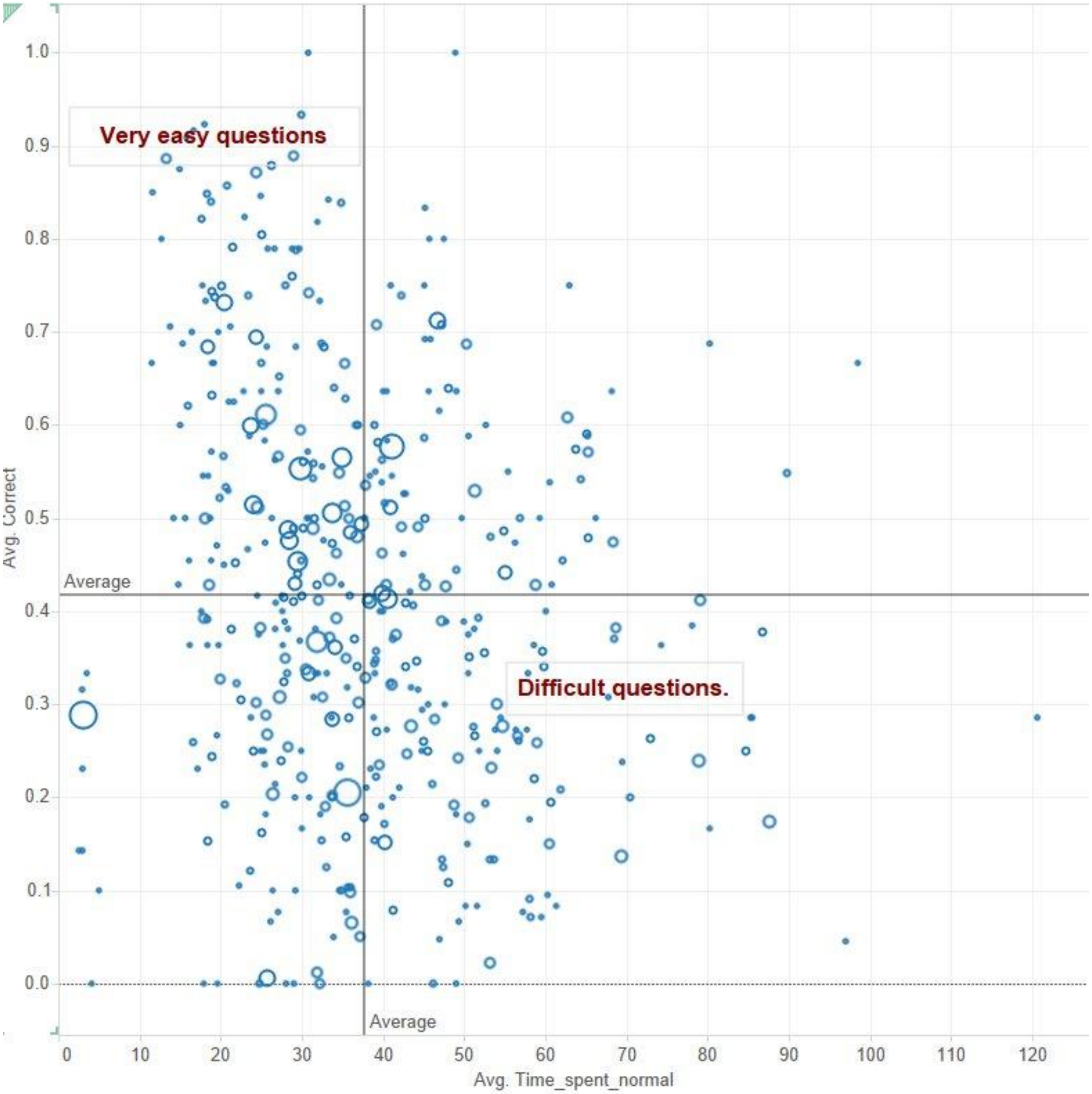


Single choice questions



Step 6. Find averages and infer results

Finally draw inferences from the graph. Creating averages, percentile lines in the graph can help. Another visualization trick that we did was to vary the bubble size depending on the number of times a question was attempted. Questions which were attempted much more stand out and we can now visually see which are the important questions which need extra analysis. Below is the final graph from our use case.

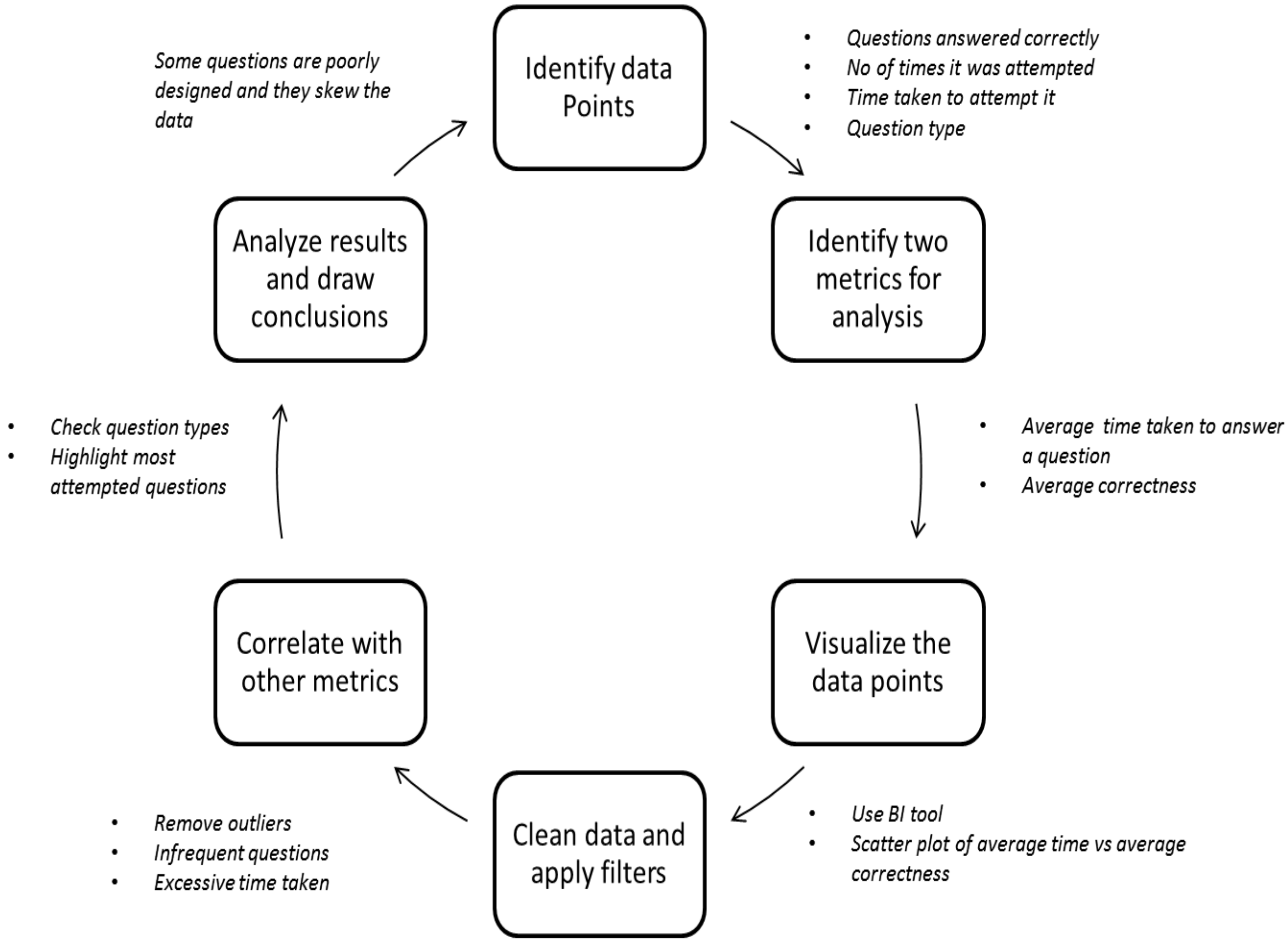


The inference that we were able to draw was that questions fell in different learning zones. For example an easy zone included questions which could be answered very easily and very quickly. On the other hand there was a difficult zone, with questions which most of the time students got wrong and also took a lot of time in answering.

We also found some interesting results, which is typically the case when doing exploratory analytics. We found there were questions which were almost never answered correctly. When we looked into the content of those questions we realized that their answer choices were ambiguous or incorrect. We also found questions which were answered very quickly but their correctness percentage was low. This indicates that the questions could be misleading because the student answered them immediately on seeing them (the student thought they knew the answer to the question) but got them wrong (the answer is not what it appears to be).

Conclusion

The purpose of this paper is to show that simple exploratory data analysis can lead to interesting and sometimes unexpected results. Below is a summary of the steps involved.


Identify data Points
• Questions answered correctly
• No of times it was attempted
• Time taken to attempt it
• Question type
Identify two metrics for analysis
• Average time taken to answer a question
• Average correctness
Visualize the data points
• Use BI tool
• Scatter plot of average time vs average correctness
Clean data and apply filters
• Remove outliers
• Infrequent questions
• Excessive time taken
Correlate with other metrics
• Check question types
• Highlight most attempted questions
Analyze results and draw conclusions
Some questions are poorly designed and they skew the data

Publishers and content creators interested in knowing the use of data analytics to improve their content and its learning efficacy should tune into the recording of our webinar on “Discover the power of Content Analytics” In the webinar we cover different topics on analytics and techniques for solving use cases similar to the one mentioned in this paper.

<<Link to webinar recording>>