KNIME Clinical Trial Query Analysis

Sanjana Pukalay, Data Science Graduate Student, Indiana University

Introduction

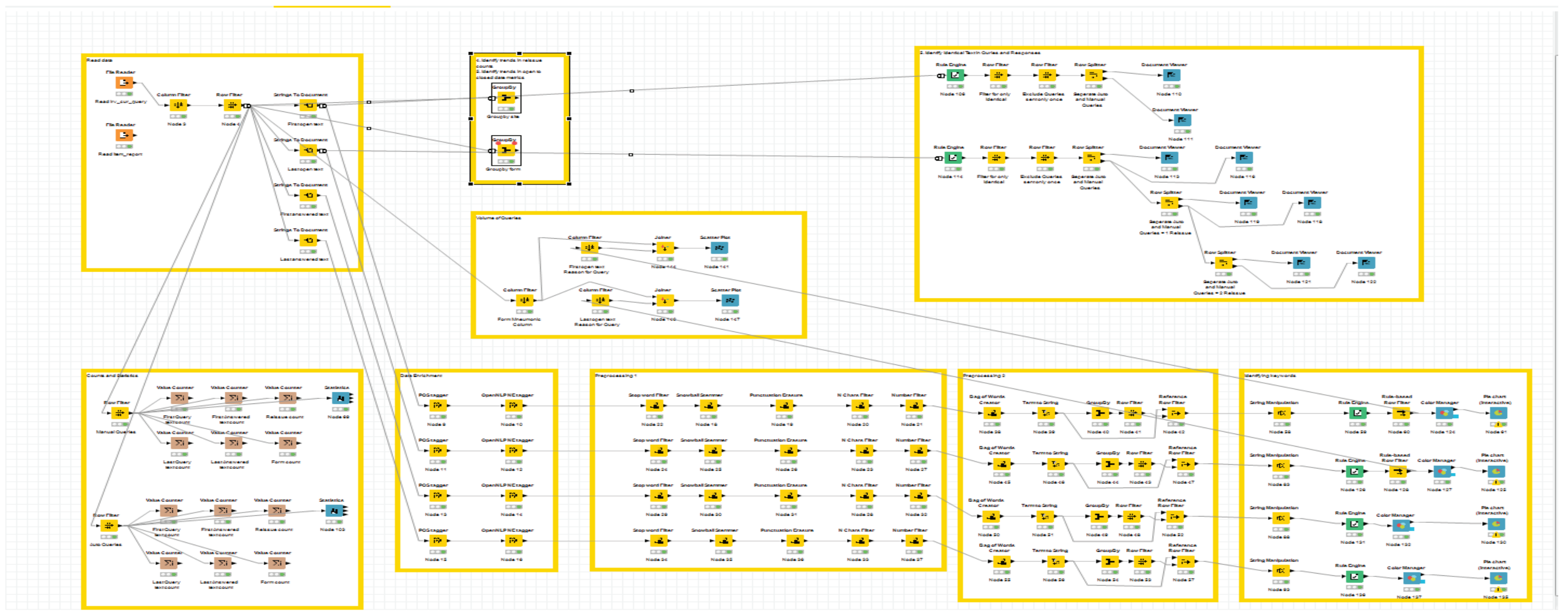
Queries can be generated during clinical trial regarding any missing or unusual values in the data. The queries may either be manual or automatically generated. Analyzing the various statistics like volume and other details of these queries like dates can provide significant insights from the data. Analysis of the textual queries is important because a lot of the future procedures and steps in the clinical trial can be altered based on existing results. This not only increases the accuracy and efficiency of the processes but also helps in fastening them. These aspects are extremely useful in the clinical trial life cycle. The clinical trial experiments usually aim at affecting a large group of the society and is non-trivial. The analysis will be done on the dataset provided by Eli Lilly and Company. This data will be analyzed using the KNIME tool accessed via the citrix virtual environment used for data analysis. The main idea is to to perform text analysis on queries along with volume related analysis on auto and manual queries.

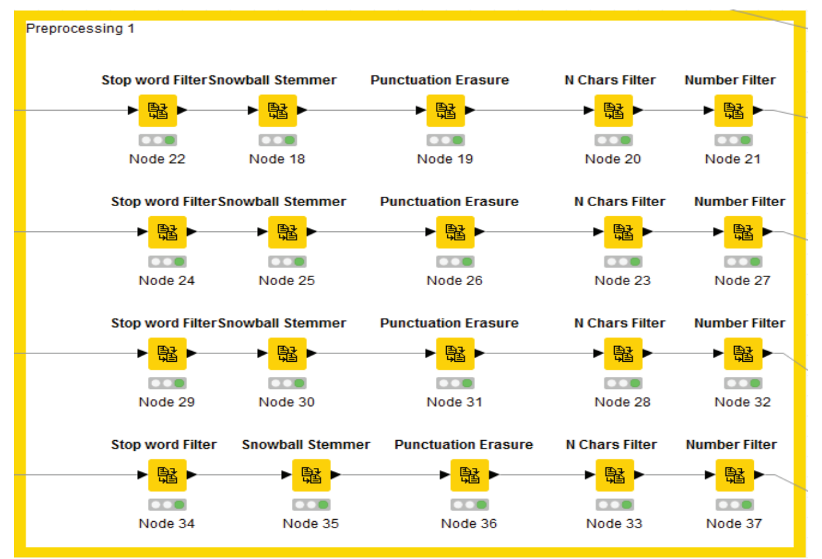
Background

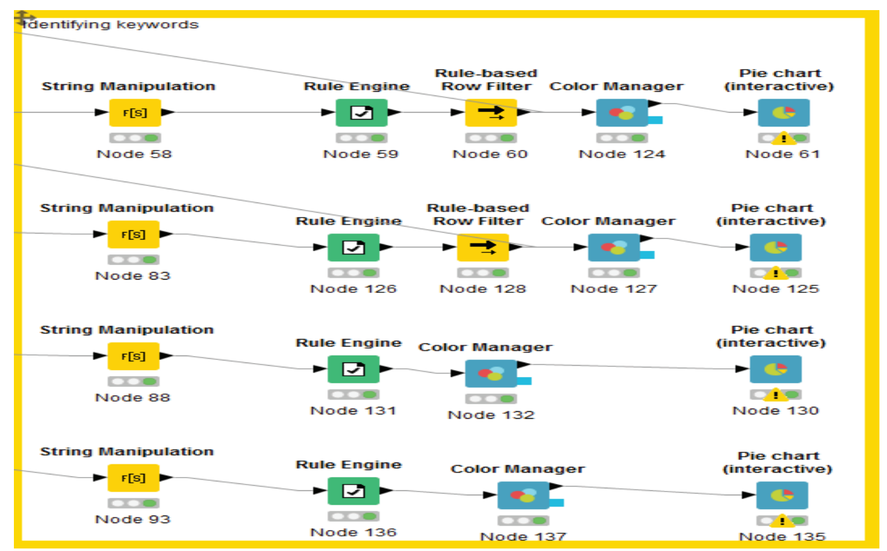
The dataset used for this project is the one directly provided by the Eli Lilly and Company as part of the Real World Data Science course at Indiana University. The data is provided in form of two de-identified CSV files with multiple rows and columns. There are many columns like QueryTime, FormID, QTypeManual, QTypeAuto, QDaysOpenToClose, QCountReIssued, FirstOpenedText, LastOpenedText, FirstAnsweredText, LastAnsweredText, Form_Mnemonic, Visit_Mnemonic, FormName, ItemCount and ItemQuestion giving us the details about type and attributes associated with queries. There are around 8000 and 1,45,000 odd rows in the two files which is a significantly good amount of data that can be used for analysis. The main idea is to find trends and unusuality associated with the data particularly the volume, time and similarity. This not only helps in standardizing the process, but also to rectify most of the existing problems in the clinical trial so it doesn't affect the advanced phases of what is called the clinical trial lifecycle. In a general sense, the development lifecycle has a protocol set. First, data is collected, enriched, pre-processed, analysis and statistical inferences are drawn and finally the data is visualized in order to gain information from the data. On similar lines, we can propose a 5 step clinical trial lifecycle and associate it with the proposed project. The first step is "Trial Design and Registration" which is essentially the Statistical Analysis Plan which includes finding count and volume analysis. The next step is "Participant Enrollment" which is basically cleaning the data and getting it into a form which can used for further sophisticated analysis. The next step that is the third step is pre-processing the cleaned data and performing statistical analysis on it using various popularly employed parameters. The fourth step is "Publication" which involves obtaining preliminary results in various forms and visualizing the data. The last step is "Regulatory Application" which involves evaluation and validation. The clinical trial lifecycle fits or rather acts as a blueprint which guides in carrying out project. It gives direction and a good approach to perform all the required steps. In the Pharmaceutical Industry, a lot of data that is generated is either numeric data which can analyzed by applying various statistical methods or textual data which can analyzed using various text analytics techniques. Text analysis can be used for finding structure and similarity in the data which can be used for predictive analytics and extracting information automatically.

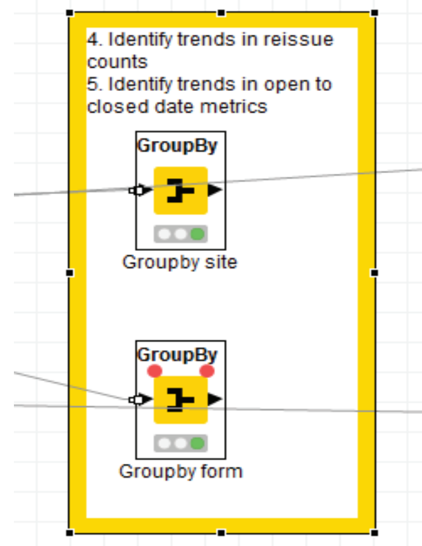
Methodology

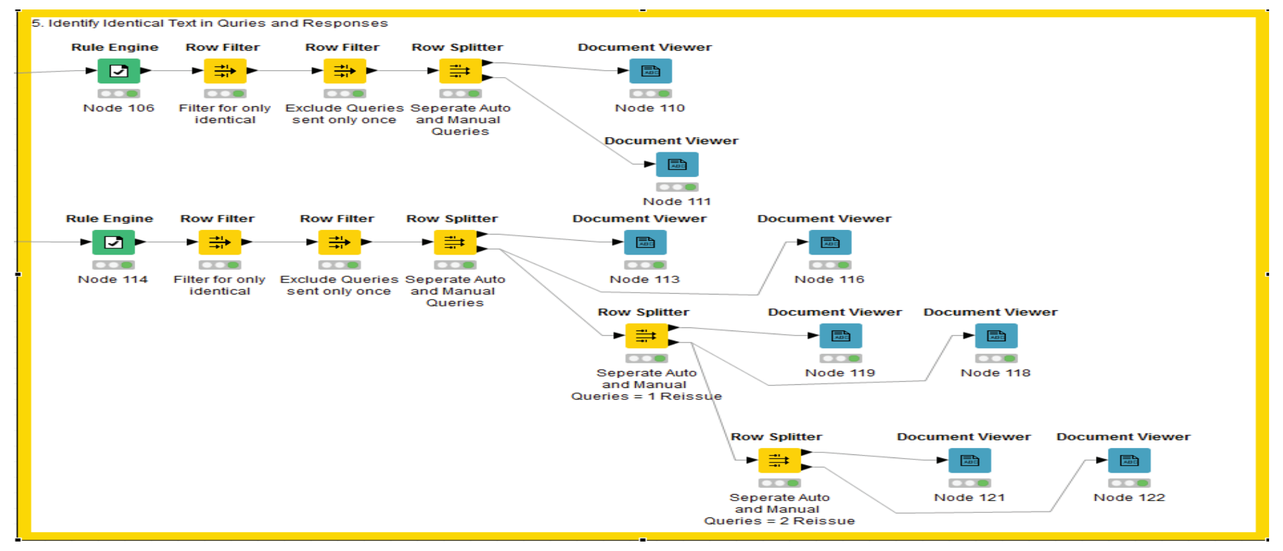
Two CSV files are used in this project - irv_cur_query.csv and item_report.csv. The data is de-identified to maintain the privacy and integrity of the data. The complete analysis and all the steps in the life cycle are performed using the data analytics tool - KNIME which is accessed through the citrix virtual environment. KNIME is a very helpful and can be adopted very easily. It provides most of the functionality for data analysis and can be used through its easy to use GUI. The major steps followed in the project is reading the data and applying basic column and row filters to it. The columns such as ids and rows with missing data is filtered and finally string to document is applied to it. Then, data enrichment is done where the data is passed through the POS tagger and Open NLP NE Tagger which essentially tags the date terms. Stop words, numbers, punctuation marks and low frequency words are removed. Then, the keywords are identified, rules are generated and the results are visualized to be able to get a good understanding. Finally, statistical analysis is performed to get an idea of the volume of auto and manually generated queries and the same is presented in a visual manner. For the most part, the workflow suggested in the course video is adopted and implemented.





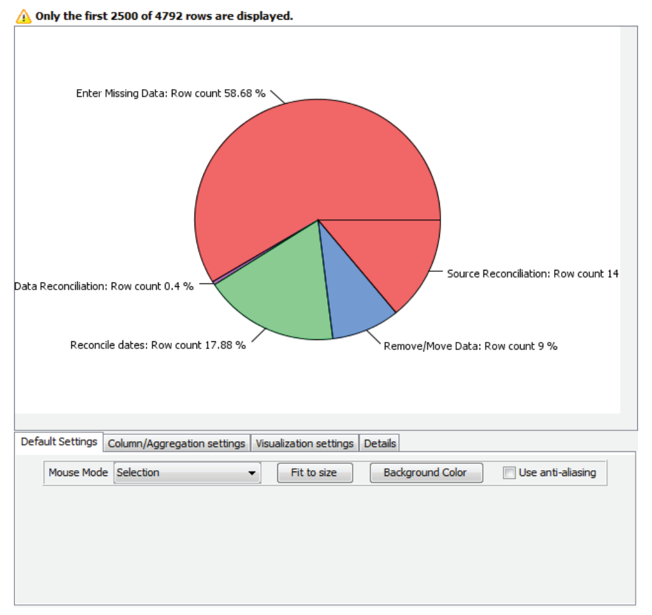


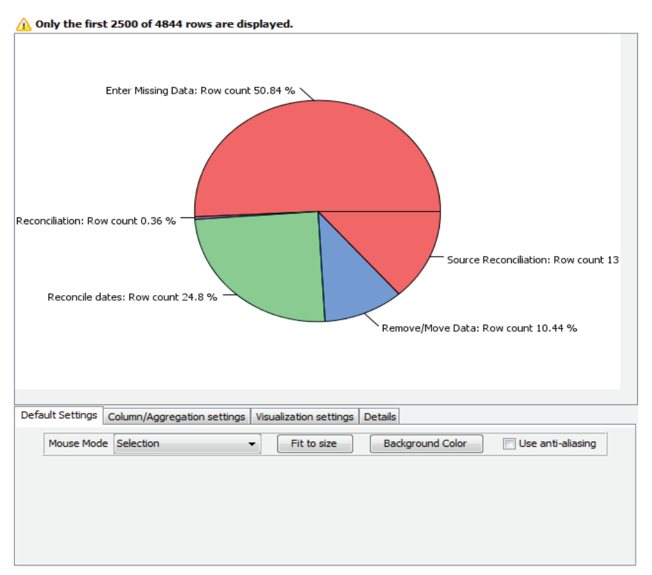

4. Identify trends in reissue counts
5. Identify trends in open to closed date metrics
GroupBy
Groupby site
GroupBy
Groupby form

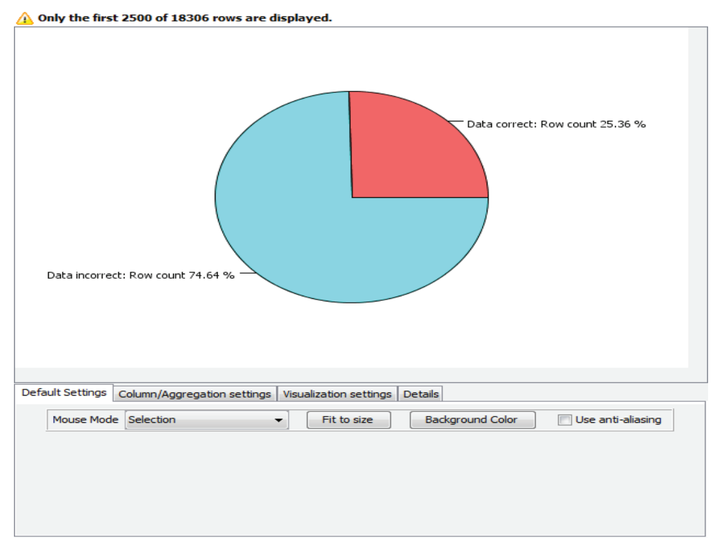

5. Identify Identical Text in Quries and Responses
Rule Engine
Row Filter
Row Filter
Row Splitter
Document Viewer
Node 106
Filter for only identical
Exclude Queries sent only once
Seperate Auto and Manual Queries
Node 110
Document Viewer
Node 111
Node 114
Node 113
Node 116
Row Splitter
Seperate Auto and Manual Queries = 1 Reissue
Node 119
Node 118
Seperate Auto and Manual Queries = 2 Reissue
Node 121
Node 122

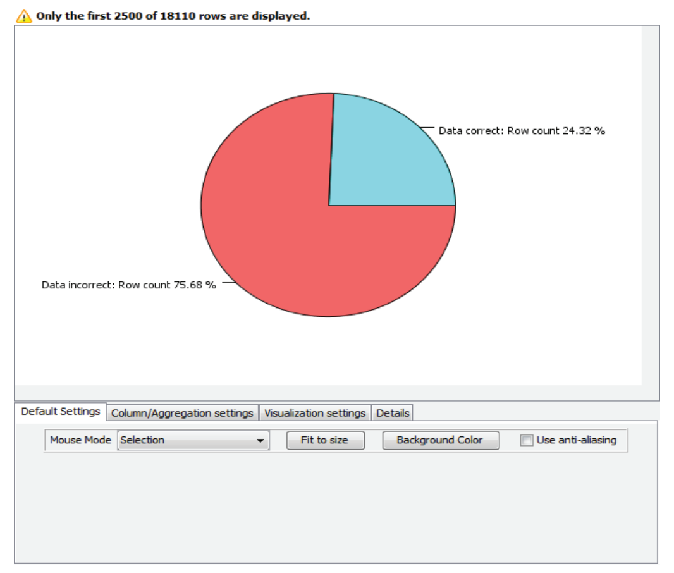
Results

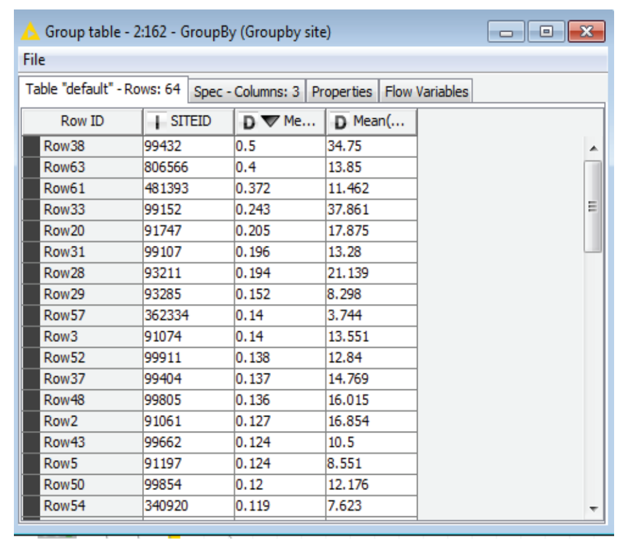
The results obtained in this analysis are rather interesting. Many queries had the same source and there were many queries which were essentially similar in nature. This analysis can be used in the future stages of the clinical trial lifecycle in such a way that similar queries can be grouped and dealt with together to save time produce faster results. Some of the problems that were seen is that the average time of attending to a query at the moment is a bit longer which is about a little more than 15 days. Efforts can be put in this direction to reduce this time. Also, just like how there is always issues associated with the analysis due to messy data, there was some unusuality in the filtering wherein there was a problem eliminating certain words present in the form of phrases instead of individual words. Visualizations have been created to enable better understanding of the obtained results and can be found below.



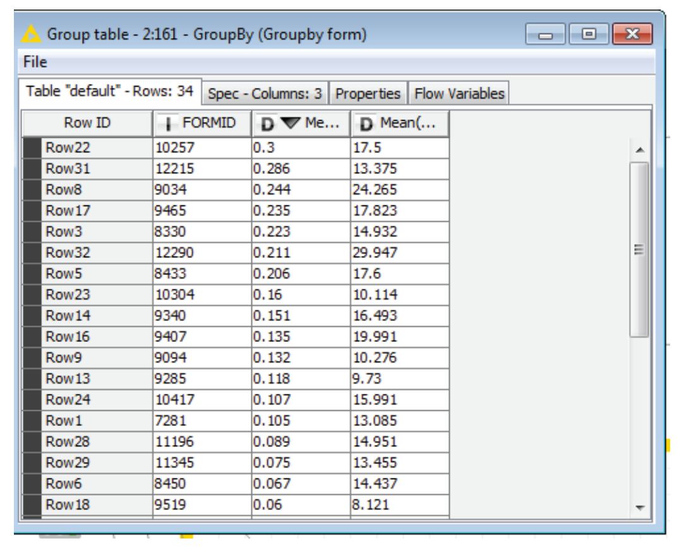







Group table - 2:162 - GroupBy (Groupby site)
File
Table "default" - Rows: 64
Spec - Columns: 3
Properties
Flow Variables

Row ID	SITEID	Me...	Mean(...
Row38	99432	0.5	34.75
Row63	806566	0.4	13.85
Row61	481393	0.372	11.462
Row33	99152	0.243	37.861
Row20	91747	0.205	17.875
Row31	99107	0.196	13.28
Row28	93211	0.194	21.139
Row29	93285	0.152	8.298
Row57	362334	0.14	3.744
Row3	91074	0.14	13.551
Row52	99911	0.138	12.84
Row37	99404	0.137	14.769
Row48	99805	0.136	16.015
Row2	91061	0.127	16.854
Row43	99662	0.124	10.5
Row5	91197	0.124	8.551
Row50	99854	0.12	12.176
Row54	340920	0.119	7.623


Group table - 2:161 - GroupBy (Groupby form)
File
Table "default" - Rows: 34
Spec - Columns: 3
Properties
Flow Variables

Row ID	FORMID	Me...	Mean(...
Row22	10257	0.3	17.5
Row31	12215	0.286	13.375
Row8	9034	0.244	24.265
Row17	9465	0.235	17.823
Row3	8330	0.223	14.932
Row32	12290	0.211	29.947
Row5	8433	0.206	17.6
Row23	10304	0.16	10.114
Row14	9340	0.151	16.493
Row16	9407	0.135	19.991
Row9	9094	0.132	10.276
Row13	9285	0.118	9.73
Row24	10417	0.107	15.991
Row1	7281	0.105	13.085
Row28	11196	0.089	14.951
Row29	11345	0.075	13.455
Row6	8450	0.067	14.437
Row18	9519	0.06	8.121

Discussion

Some of the key take aways from this analysis are that, the quality of data can play an important role in determining the accuracy of the analysis. If the data is too messy or complicated, thorough analysis cannot be performed with ease. Also, a significant amount of preprocessing needs to be done on the data to be able to make it ready for analysis. Preprocessing could be anything from filtering out missing values to removing stop words and other things which do not contribute positively to the analysis. As mentioned, this analysis can be used to determine the quality of the more advanced stages of clinical trial lifecycle. It is observed that, some queries are more prominent. They can be grouped together and dealt with together and in a similar manner. Also, we can reduce the average time taken to resolve a query. All these are important factors which helps improve the quality of the lifecycle and indirectly affect businesses. Better policies and business models can be developed. The action items that are accomplished in the analysis are finding the number of days required for queries to be resolved, the number of times the queries are reissued, finding the volume of auto and manual queries, spotting similarity in queries and responses for manual and auto queries, identifying frequency of keywords in manual and auto queries and identifying general trends. It has been observed that some queries are more frequently generated in comparison to the others. This can be used to train some particular sites and speeding up the query acknowledgement process. These queries are essentially an outcome of the design flaws in the clinical trial lifecycle which needs to be eliminated in order to get a positive outcome from the trials. These metrics can be used as a basis for evaluating and validating the progress in the clinical trial lifecycle. Measures can be taken for better management of data as we know that data is essentially generated at every stage of the lifecycle. They can be scaled up to handle big data and further analysis. This kind of analysis has a significant impact on the pharmaceutical industry and the clinical trial lifecycle and also on the drug development lifecycle. It can be used to identify trends and similarity in the data which can be used for further analysis.

References

Transcelerate White Paper
KNIME White Paper
Polyanalyst White Paper
Big Data in Pharma Papers

SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
Department of Information and Library Science
Bloomington