IFN650 BUSINESS PROCESS ANALYTICS

Assignment 2: Group 4

GROUP MEMBERS:

DEBORAH MEI HAR CHAN (N9781846)
DWISHANTH PANDURANGAN (N10840371)
KAVYA KORE (N10840371)
MEET CHIRAGKUMAR SOLANKI (N11569751)

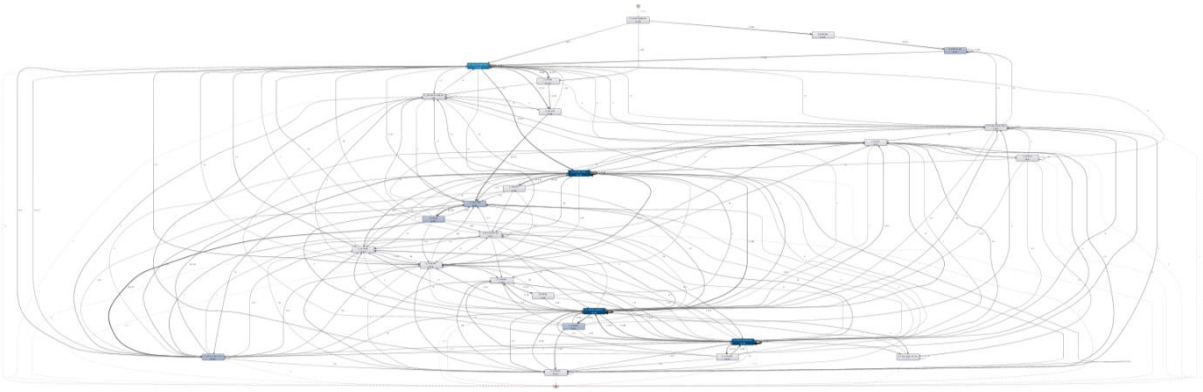
Contents

Part A: Disco Analysis

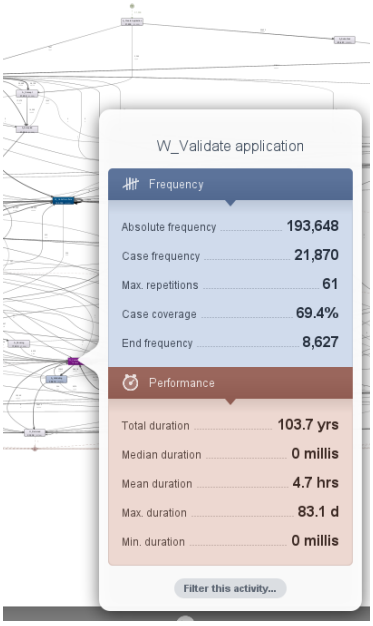
1.0 Analysis and Interpretation of Process Models

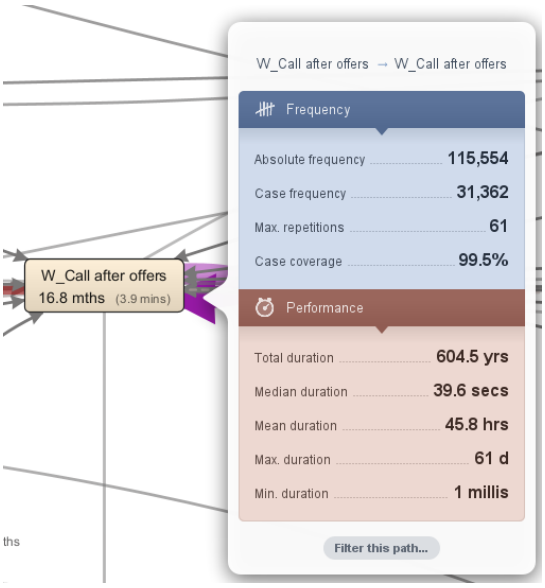
1.1 Compare and Contrast of 2 Process Models (100% activity/paths VS 50% activity/paths)

Comparing and contrasting process models with different percentages of activity and paths allows the user to obtain insights into the completeness and accuracy of the models. The figure below represents the process model with 100% activities and paths, covering all 26 activities and observed paths in the data.



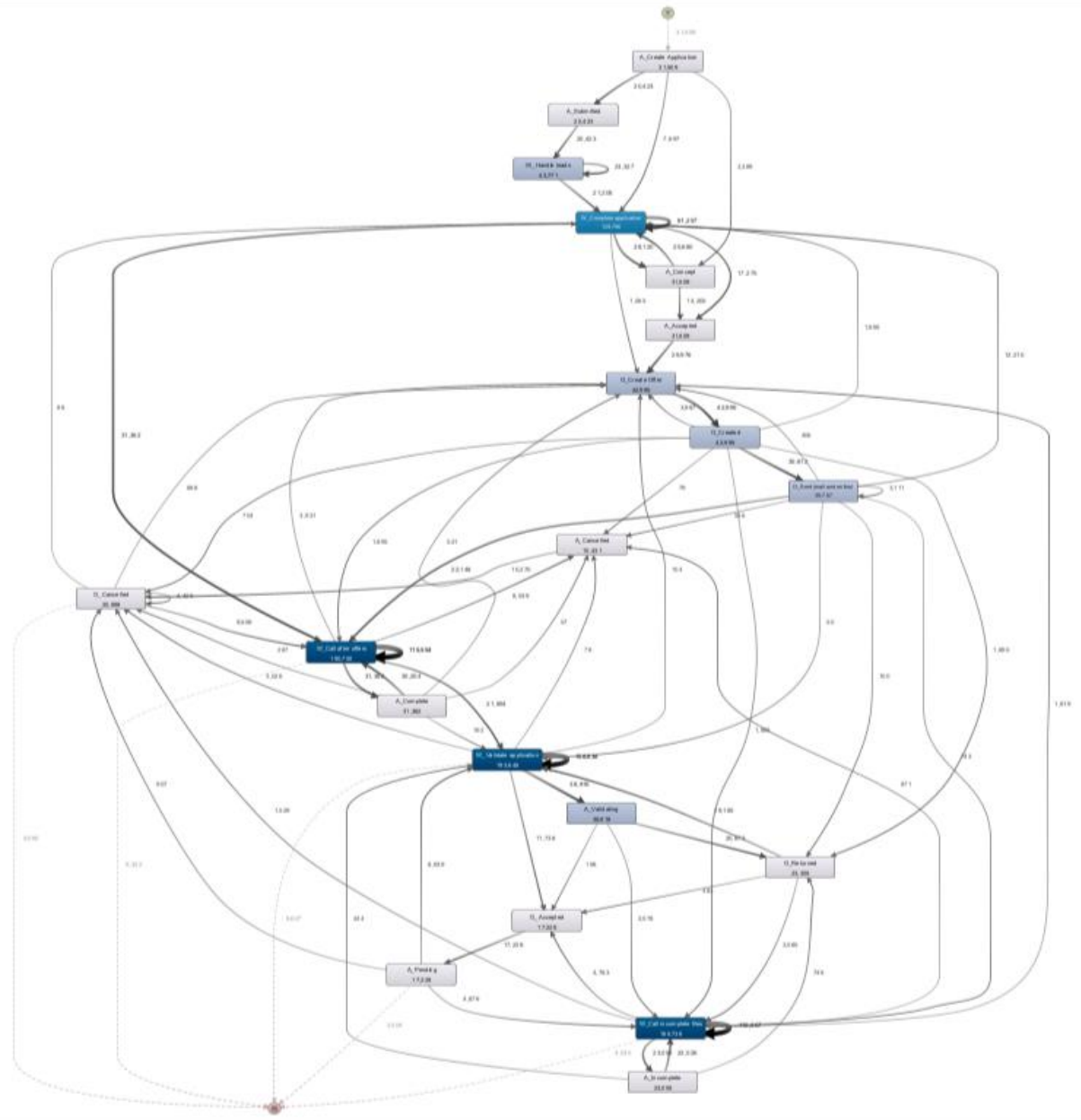
Using the “frequency” and “performance” metrics, it was identified that the highest absolute frequency activity was “W_Validate application”, being performed 193,648 times with a cumulative total duration of 103.7 years. In terms of paths, the highest absolute frequency occurred within the “W_Call after offers” being performed 115,554 times, cumulating to a total duration of 604.5 years.





Combining these metrics allows for an easy analysis of the activity(s) and path(s) that cause bottlenecks and the biggest overall delay. Through this process model, a more accurate and comprehensive representation of the processes that includes all possible process variants are provided. This allows for a better understanding of the process flow, the identification of uncommon behaviors, and more accurate predictions and/or recommendations being made for new cases.

The figure below represents the process model with 50% activities and 50% paths, displaying a total of 20 activities.



A significant clear difference between the number of paths displayed between the 100% and 50% process model can be seen. There are also 6 activities absent in the process model, which include:

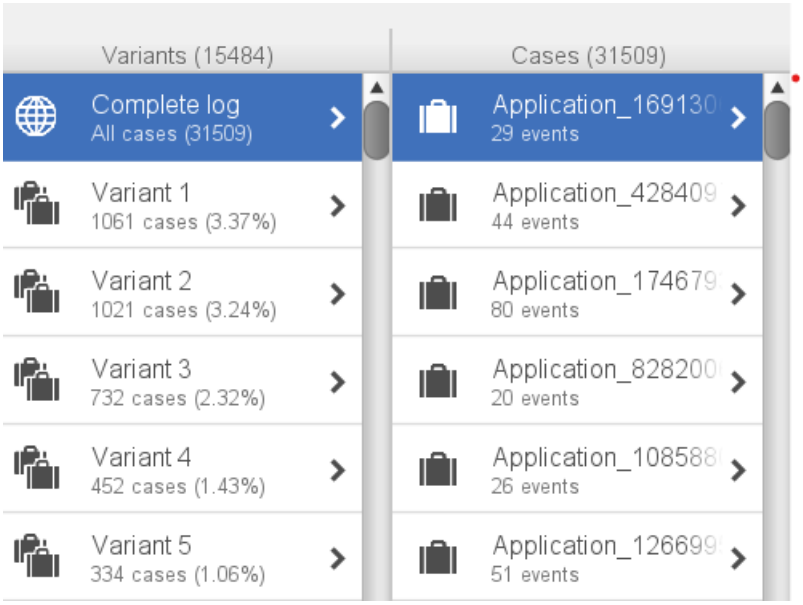
- W_Assess potential fraud
- W_Personal Loan collection
- W_Shortened completion
- O_Sent online only
- O_Refused
- A_Denied

Although there was no difference in the highest absolute frequency of activity and path respectively, the simplified process model does not allow the investigation of the missing activities and paths which, while still useful, would lead to potential inaccuracies in the representation of actual process behaviors, limited insights to unusual process patterns, and affects the reliability of recommendations.

Model with 0% paths should not be used to understand process behaviours due to several factors including, but not limited to:

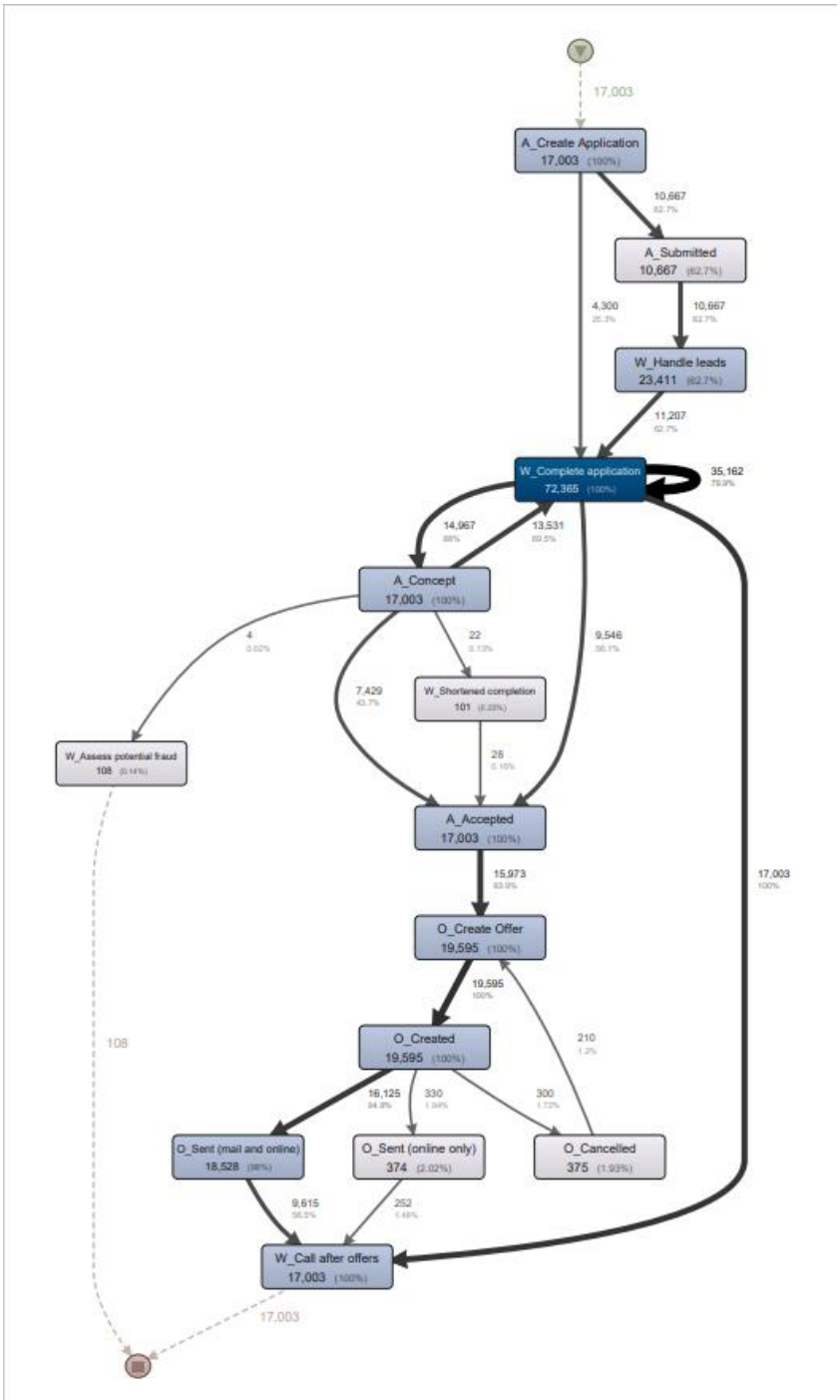
- **Incomplete Representation**: A model with 0% paths fails to capture the full complexity and diversity of the process.
- **Missed Insights**: Critical process behaviors, exceptions, or rare scenarios that may have a significant impact on the overall process performance may be overlooked and prevents identification of potential bottlenecks, deviations, or inefficiencies in the process.
- **Hidden Risks**: Having no paths in the model may lead to unforeseen risks or compliance issues that were not considered or addressed.
- **Limited Generalisation**: The ability to make accurate predictions and/or recommendations for decision making or process improvements will be hindered as unseen data fail to provide meaningful insights beyond the specific cases covered by the available data.

1.2 Investigation of Case Variants Detected from the Log

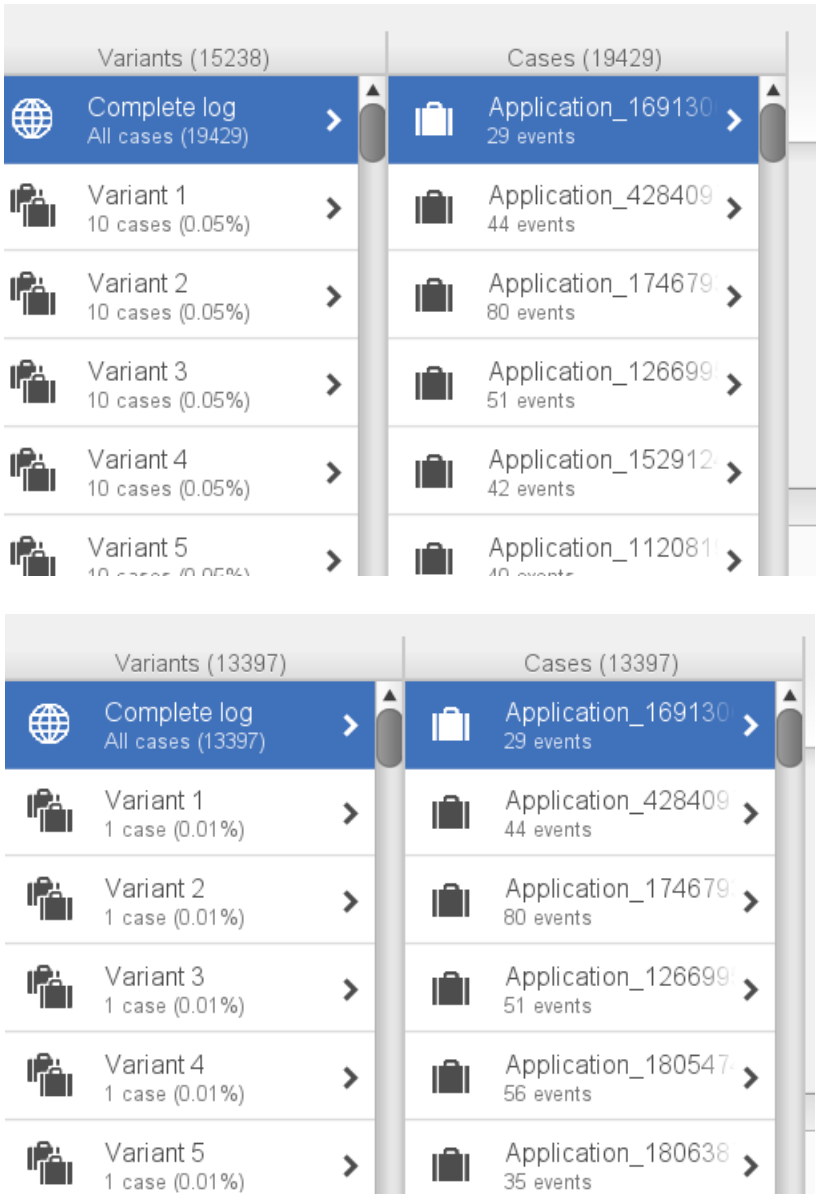


Overall, there are 15,484 case variants present in the log. The top five most frequent case variants and their respective frequencies are presented in the table below.

Top 5 Variants	**Frequencies**
Application_828200680	1061
Application_857715587	1021
Application_1710223761	732
Application_1413308979	452
Application_1091101077	334



These top 5 variants covered the log up to the activity **W_Call after offers.**



There are 13397 case variants with low frequency and 15238 variants only have less than 10 cases.

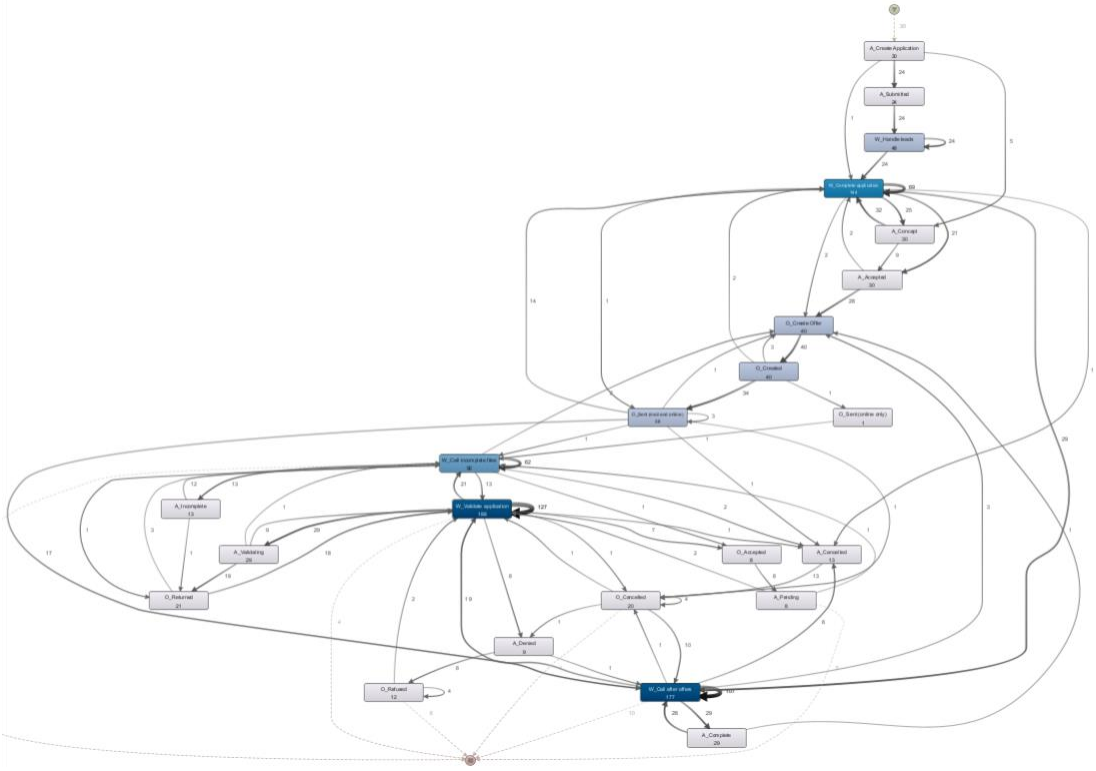
1.3 Implications of a High Number of Case Variants

Case variants are a part of the business, and it reflects the realities and being aware of what they are is crucial in achieving optimal process performance. Various implications can arise when dealing with a high number of case variants while generating a representative process model. A positive outcome would be in the focus of process improvement as the high number of case variants helps identify opportunities through identifying common patterns that require attention such as bottlenecks and inefficiencies. However, it is imperative to note that many challenges are associated with having too many case variants. For instance, it increases the complexity of the process and causes a challenge in accurately capturing the underlying process flow as well as the overall process behavior. This could also lead to difficulties in

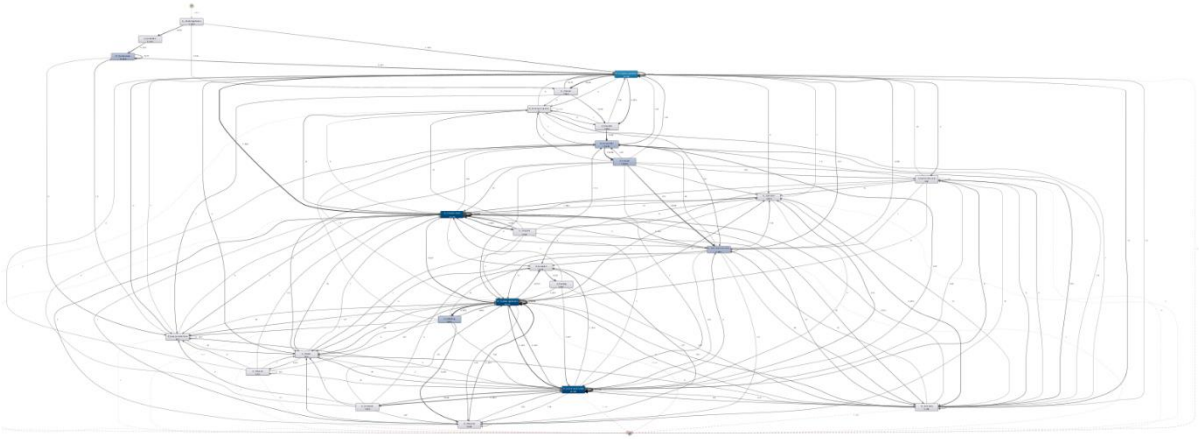
understanding and interpreting the model to gain meaningful insights. Different resources may be needed for case variants and a significantly high number of them would cause a challenge in allocating resources effectively and optimizing resource utilization. Additionally, having numerous case variants in a process may result in a decrease in the effectiveness of control mechanisms and maintaining compliance as each variant may have specific requirements which could lead to a more demanding task for the business. Designing and implementing automated workflows would be a challenging process when there are numerous case variants due to the possibility of each variant requiring specific rules and decision points, amongst other factors. Not to mention, it would also affect the complexity in handling process documentation and standardization, process modifications, change requests and change management as it is essential to involve all stakeholders to understand the process and effectively handle the change and ensure a smooth transition. Considering these implications is imperative as it enables grasping the complexities of generating a representative process model involving a high number of case variants, and thus, facilitates a more comprehensive decision-making and strategy implementation approach to manage processes effectively.

2.0 Comparison and Observation of Process Behaviours

The home improvement and business goal process models are somewhat similar to each other with regards to the number of activities, whereby the business goal consists of 23 activities while home improvement consists of 25. However, they differ greatly in terms of the path and number of cases as the home improvement process model is seen to be much more complex, as seen in the figures below.



Group A: Business goal - 1,044 total frequency of cases



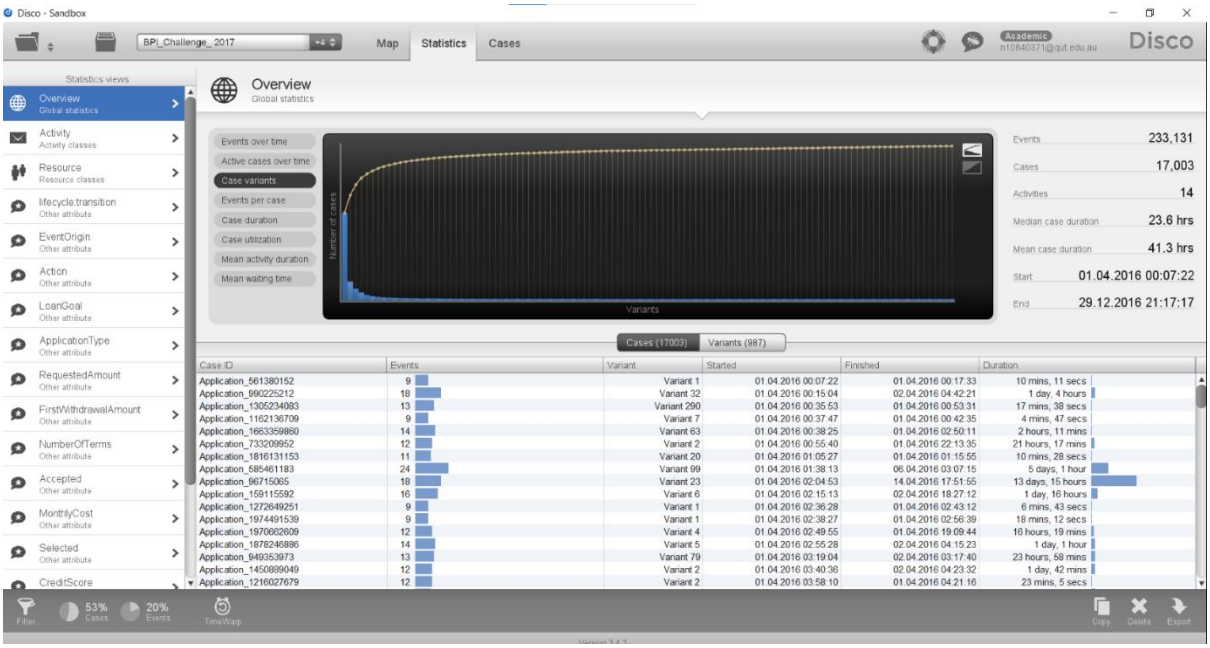
Group B: Home improvement - 284,745 total frequency of cases

Using the combination of “frequency” and “performance” metrics, the following information of the two process models was identified.

Process Model	Frequency Metrics		Performance Metrics	
	Primary: Absolute frequency Secondary: Maximum repetitions		Primary: Total duration Secondary: Mean duration	
	Activity	Path	Activity	Path
Business Goal	W_Validate application. Absolute frequency: 188. Throughput time: 70 days. Case frequency: 19. Max repetitions: 20	W_Validate application -> W_Validate application. Absolute frequency: 127. Throughput time: 61.4 days Case frequency: 19 Max repetitions: 16	W_Validate application. Absolute frequency: 188. Throughput time: 70 days. Case frequency: 19. Max repetitions: 20	W_Call after offers -> W_Call after offers. Absolute frequency: 107. Throughput time: 29 weeks. Case frequency: 29. Max repetitions: 7
Home Improvement	W_Call after offers. Absolute frequency: 47,059. Throughput time: 29.3 weeks. Case frequency: 7,648. Max repetitions: 64.	W_Call after offers -> W_Call after offers. Absolute frequency: 28,880. Throughput time: 159.9 years. Case frequency: 7,648. Max repetitions: 61.	W_Validate application. Absolute frequency: 46,481. Throughput time: 23.8 years Case frequency: 5,542 Max repetitions: 42.	W_Call after offers -> W_Call after offers. Absolute frequency: 28,880. Throughput time: 159.9 years. Case frequency: 7,648. Max repetitions: 61.

From the observation in the table above, the bottlenecks for both processes lie within the “W_Validate application” and “W_Call after offers” activities and paths. Hence, it can be proposed that the two process models are quite like each other although they differ a lot from each other in terms of their respective total frequency of cases and paths.

3.0 Applying Filters to Log



Q: **How many cases are there in the log?**

A: There are 17003 cases in the log.

Q: **What is their mean duration?**

A: 41.3 hours (about 1 day 17 and a half hours)

Q: **How many variants are there in the log?**

A: There are 987 variants present in the log.

Q: **How significant is the problem of rework for this process?**

A: The large number of cases and variants implies that significant challenges will occur for rework due to the fact that it would be a substantial cumulative effect on the overall process efficiency and extra time, effort and resources will be needed to rectify these inefficiencies.

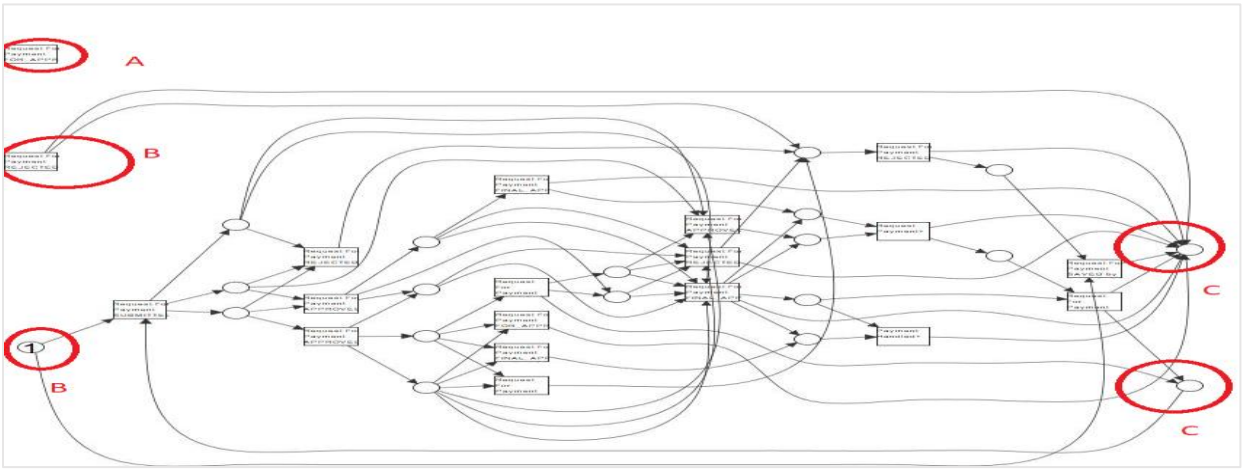
Q: **Are there any bottlenecks detected? If so, which activities/paths are involved?**

A: The bottleneck falls in the “W_Complete application” activity, with a total frequency of 3,682, a maximum 8 repetitions within the case. The total duration for this activity accumulates to 21.4 weeks, meaning this area is a high-impact area. When using the performance metrics, the bottleneck is detected in the “A_Concept -> W_Complete application” path that takes a total duration of 31.9 months; proving that this path causes a delay in the processes.

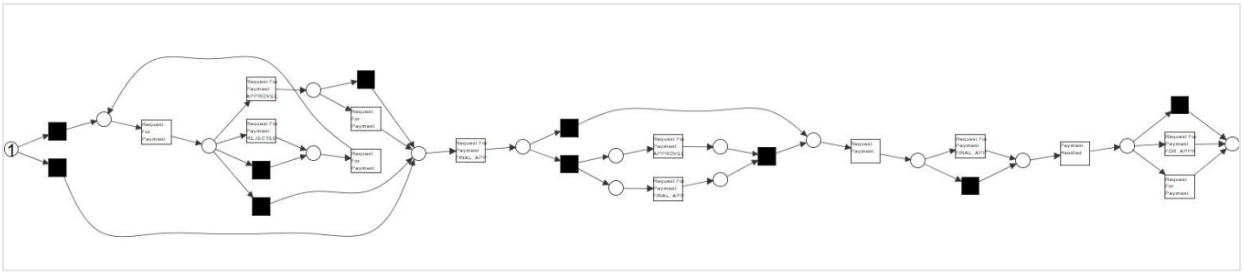
Part B: ProM

1.0 Alpha Miner and Inductive Miner Petri Nets

Below figures give the output from Alpha miner and Inductive miner, exported as .jpg and copied. The inductive miner algorithm has been used with 90% path or with at-most 10% noise.



Output from Alpha Miner add-on



Output from Inductive Miner Add-on

Start events

Total number of classes: **2**

Class	Occurrences (absolute)	Occurrences (relative)
Request For Payment SUBMITTED by EMPLOYEE+	6814	98.954%
Request For Payment SAVED by EMPLOYEE+	72	1.046%

End events

Total number of classes: **10**

Class	Occurrences (absolute)	Occurrences (relative)
Payment Handled+	6305	91.563%
Request For Payment REJECTED by EMPLOYEE+	448	6.506%
Request For Payment SAVED by EMPLOYEE+	74	1.075%
Request For Payment REJECTED by MISSING+	41	0.595%
Request For Payment FINAL_APPROVED by SUPERVISOR+	11	0.16%
Request Payment+	2	0.029%
Request For Payment REJECTED by ADMINISTRATION+	2	0.029%
Request For Payment REJECTED by SUPERVISOR+	1	0.015%
Request For Payment FOR_APPROVAL by SUPERVISOR+	1	0.015%
Request For Payment FINAL_APPROVED by DIRECTOR+	1	0.015%

Log Summary View (Exploratory Analysis)

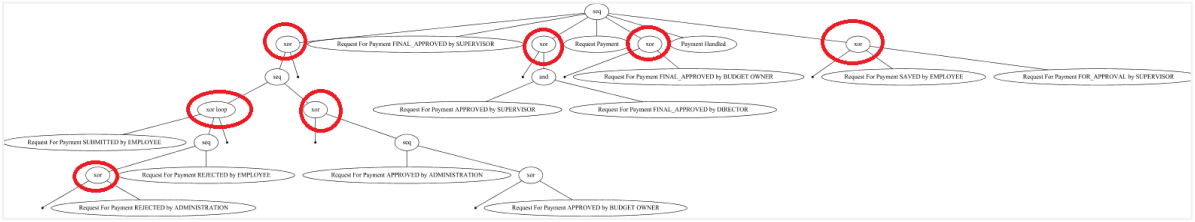
The models from Alpha miner and Inductive miner are clearly different from the above output.

- The noise level from 'log summary view' indicates that less noise i.e., traces start with events other than "start" and end with events other than "end". [Refer 'A', ''B' and 'C' circled in red in figure-1 & Figure-3 for noise log analysis]
- Because the Alpha miner algorithm cannot cope very well with noise, its discovered model presents a lower fitness, as compared to the Inductive Miner.
- There are also no loops shown in the model, all forward arrows

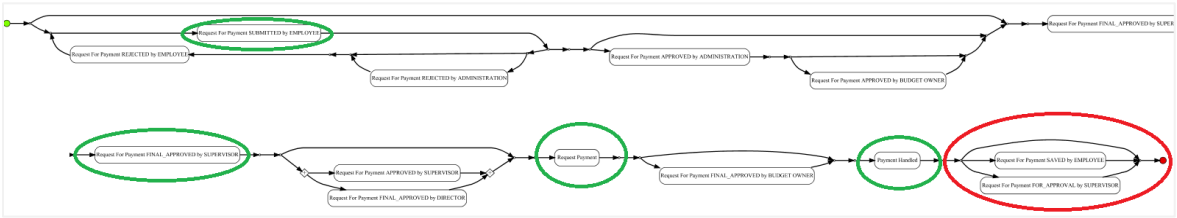
The model given by the Inductive Miner has a good fitness value but can see several number of silent transactions in the model that it is not that precise.

2.0 Inductive Miner algorithm (Process tree)

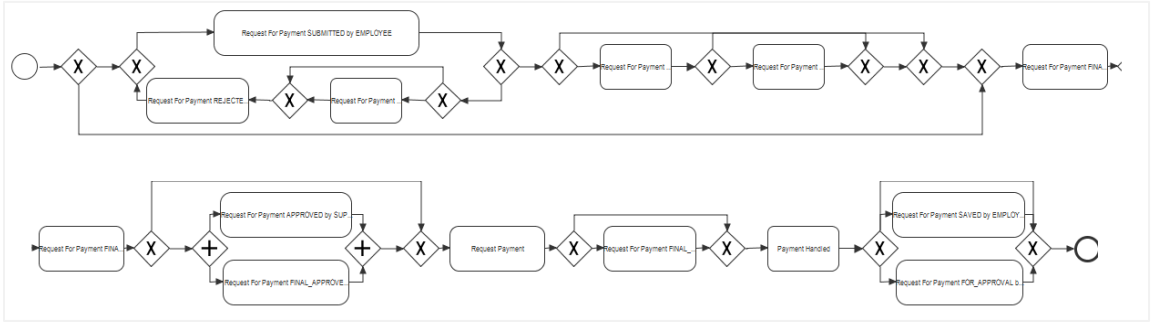
Below are the figures from Inductive miner with process tree plug-in.



The Graphviz Process Tree Visualisation above shows several silent transactions especially with XOR, the process is not precise.



Process Tree Visualisation



BPMN Process Tree Visualisation

The Process Tree Visualisation and BPMN Process Tree Visualisation figures above show the following 3 steps in the process (the flow is lengthy so shown each view in two rows in the display),

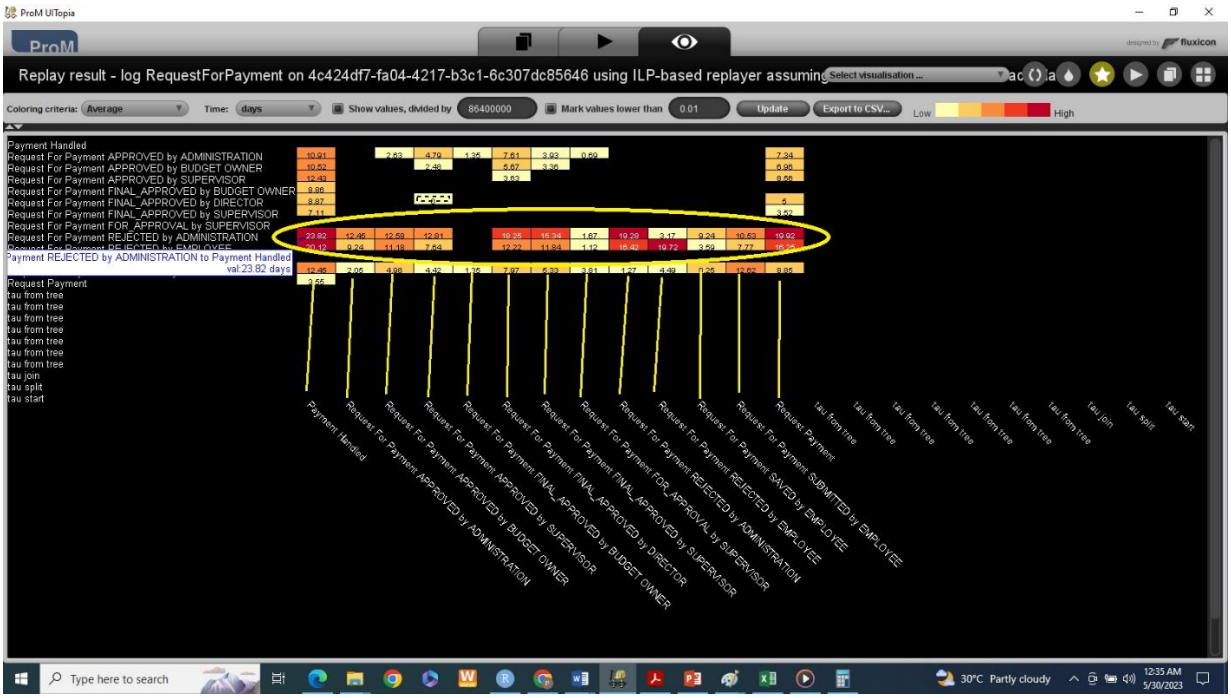
1. Request for payment is submitted by employee
2. It gets approved or rejected by different roles
3. Payment is handled

<Payment handled> should be the end process but in this log/model highlighted in red circles in the Process Tree Visualisation figure shows below two tasks after payment is handled.

1. Request for Payment SAVE by EMPLOYEE
2. Request for Payment FOR_APPROVAL by SUPERVISOR

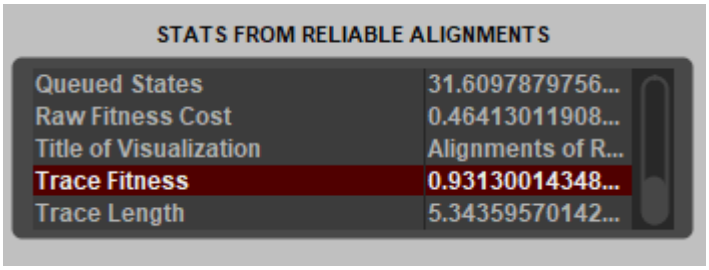
3.0 Replay a Log on Petri Net for Conformance Analysis

From the "time between transactions view" below, "Request for payment REJECTED by ADMINISTRATION" taking more than 28 days average (changed time from minutes to days) in heat map below, so the elapsed times might be high in this workflow.



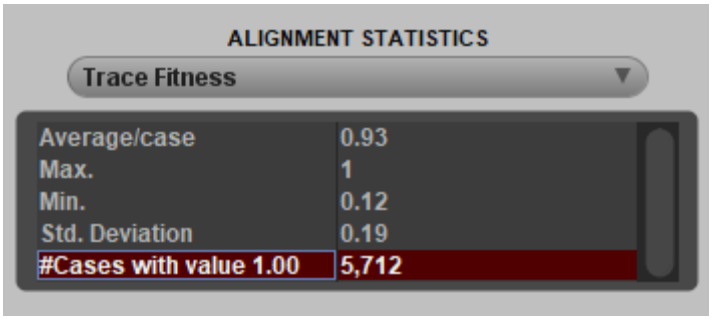
3.1 <u>Does the model completely fit the log?</u>

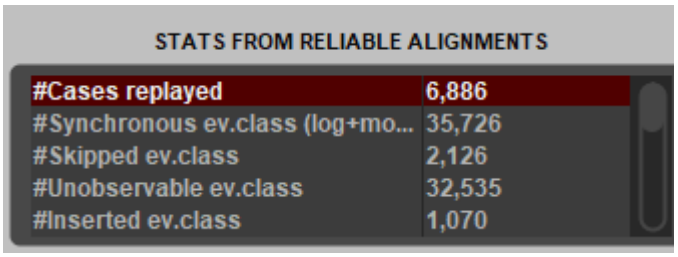
No, the model does not completely fit the log, but it fits to 93.13% as shown in the figure below.



3.2 <u>If not, how many cases fit the models and how many do not?</u>

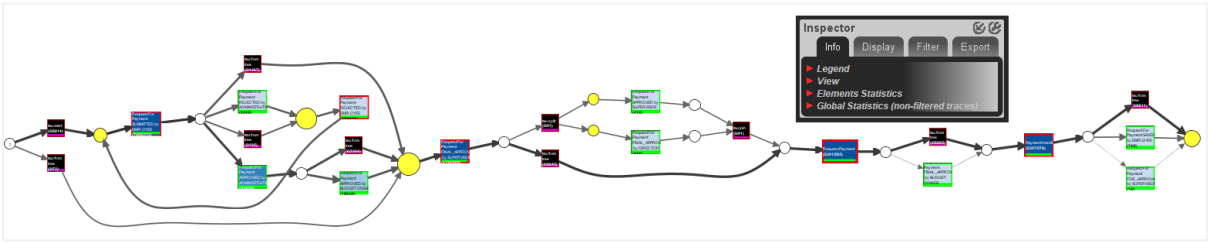
Based on alignment statistics, **5712 cases fit the model from "Trace fitness"** view below and out of **6886**. The cases doesn't fit the model are then 6886 – 5712 **= 1174**





3.3 <u>Where are the problems for the non-fitting process cases?</u>

Based on the figure-8 below, the cases where the model is deviating are listed in the table



Replay of inductive mining for conformance

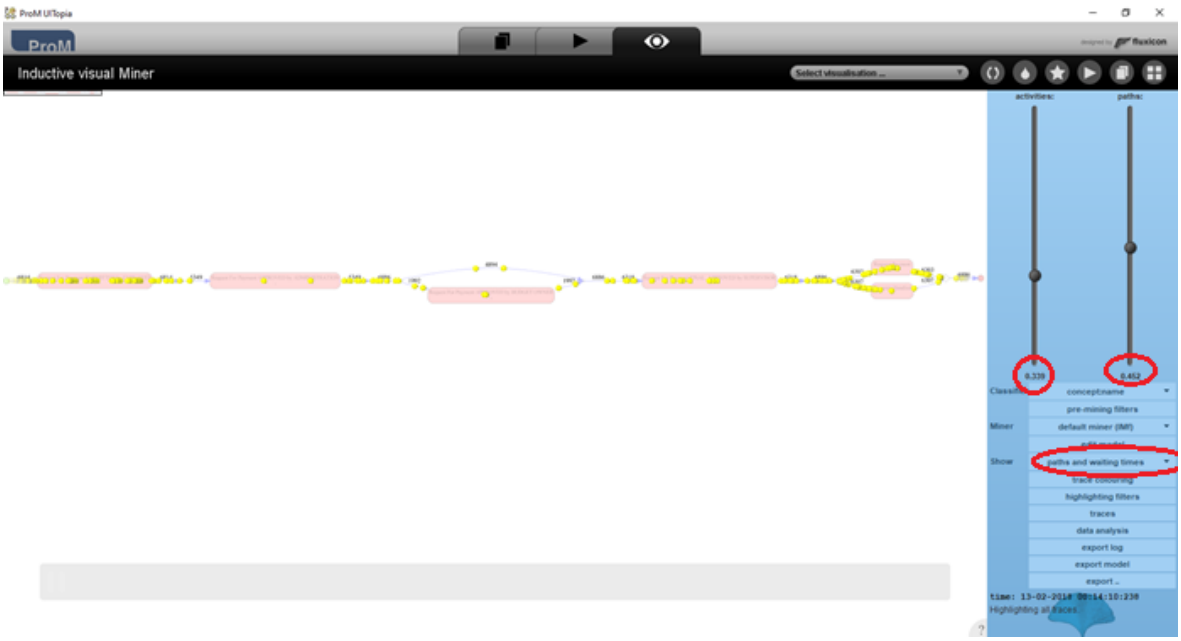
Element	Total Moves in log + model (A)	Move in model only (B)	% C = (B/A) *100
Request for payment "SUBMITED BY EMPLOYEE"	7469	374	5%
Request for payment "REJECTED by EMPLOYEE"	1009	20	1.9%
Request for payment "FINAL_APPROVED by SUPERVISOR"	6318	568	8.99%
Request Payment	6301	585	9.3%
Payment Handled	6307	579	9.18%

About 9% of the non-fitting cases are from each activity, Request for payment "FINAL_APPROVED by SUPERVISOR", "Request Payment" & "Payment Handled"

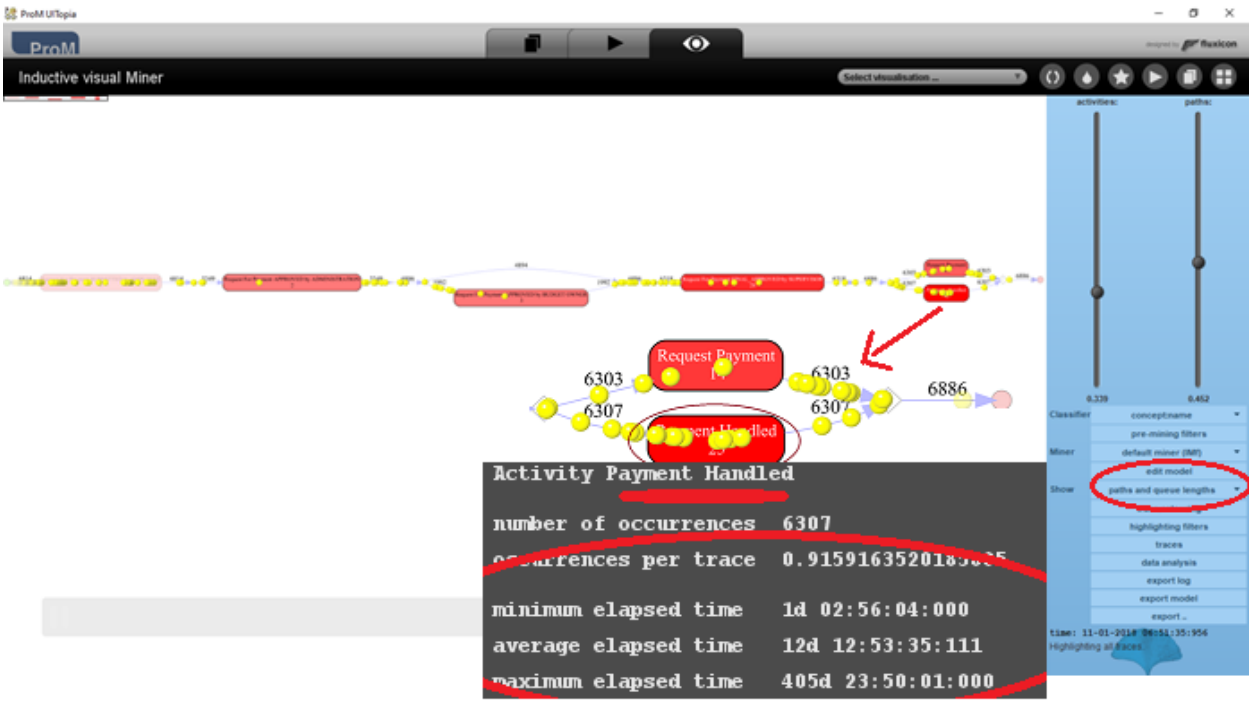
4.0 Process Analysis using Inductive Visual Miner

4.1 Bottlenecks

✓ Paths and Waiting Times view, there are no issues with waiting time in process.



✓ Paths and Queue lengths view, there are long waiting of cases (tokens) in queue, especially in handling payment i.e., elapsed time is long though service time is less. This may be ok as the company might process the payments monthly once.

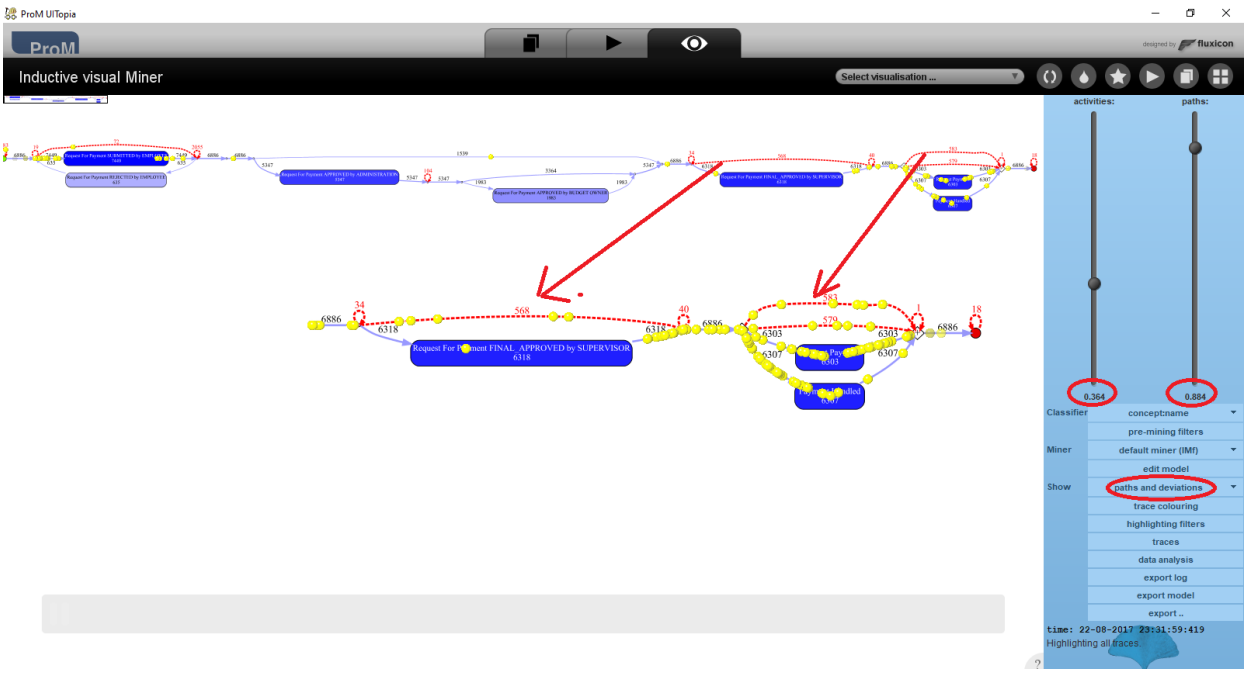


4.2 Deviations

✓ Paths and Deviations: There are 658 cases out of 6886 (9.5%) are skipping the activity called "**Request for payment FINAL_APPROVED by SUPERVISOR**"

✓ 579 out of 6307 cases (9.1%) are deviating from the model activity called "**Payment handled**"

✓ 583 out of 6303 cases (9.2%) are deviating from the model activity called "**Request Payment**"


ProM UITopia
Inductive visual Miner
Select visualisation ...
activities:
paths:
0.364
0.884
Classifier
conceptname
pre-mining filters
Miner
default miner (IMf)
edit model
Show
paths and deviations
trace colouring
highlighting filters
traces
data analysis
export log
export model
export ..
time: 22-08-2017 23:31:59:419
Highlighting all traces.

Part C: BPI Challenge 2011 Log Exploratory Analysis

1.0 Analysis by Group Member: Dwishanth Pandurangan (n10933816)

ProM BMP data mining tool has been used for exploratory analysis of Hospital log from Dutch Academic Hospital.

Selecting right attributes for Filtering and Analysis

Imported given *Hospital_log.xes* file in ProM and exported and saved as .csv file, to-do visual inspection of data. Here is the summary,

<table>
<tr><th>Attribute</th><th>Count</th><th>Remarks</th></tr>
<tr><td>Total number of attributes in log file</td><td>69</td><td>Large number of attributes, need to select important applicable fields for BPM analysis</td></tr>
<tr><td>Number of event attribute types</td><td>390</td><td>Has 390 different types written in medical terms and more over they are in ‘Dutch language’, so ‘event’ is not right choice for analysis.</td></tr>
<tr><td>Number of org:group</td><td>27</td><td>There are 27 departments performing various medical functionaries and I am familiar so this attribute <u>has been selected for analysis</u></td></tr>
<tr><td>Treatment codes</td><td>31</td><td>This is more of specific to doctors and treatment procedures, my domain knowledge is not sufficient, so not good field to take for analysis</td></tr>
<tr><td>Section</td><td>7</td><td>There are totally 7 sections of which ‘section 2’ performs the Gynecology related treatments shown below, so this field <u>has been selected</u> for <u>filtering</u> the cases.
<table>
<tr><th>Section 2</th><th>13922</th></tr>
<tr><td>IVF clinic</td><td>9</td></tr>
<tr><td>Maternity ward</td><td>12</td></tr>
<tr><td>Nursing ward</td><td>9018</td></tr>
<tr><td>Obstetrics & Gynaecology clinic</td><td>4883</td></tr>
</table>
</td></tr>
</table>

Process Mining Questions Formation

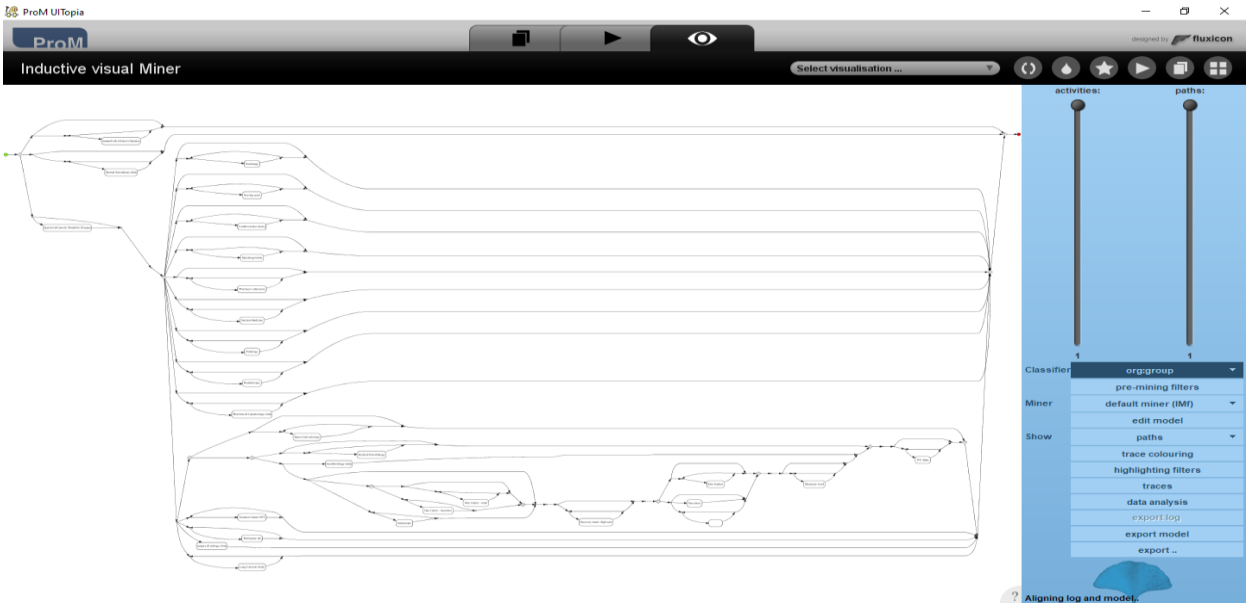
As the analysis is specific to gynecology and various groups involved in gynecology treatment the following questions are formulated for analysis.

Q1. Which org group has more queues?

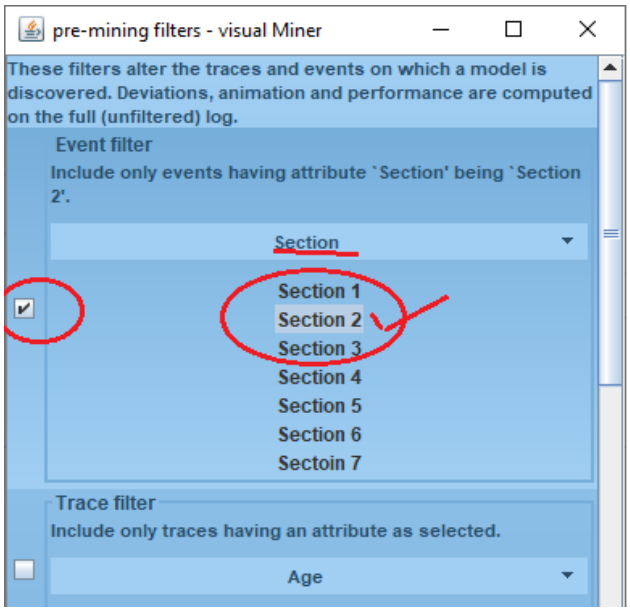
Q2. What is the recommendations for process improvement?

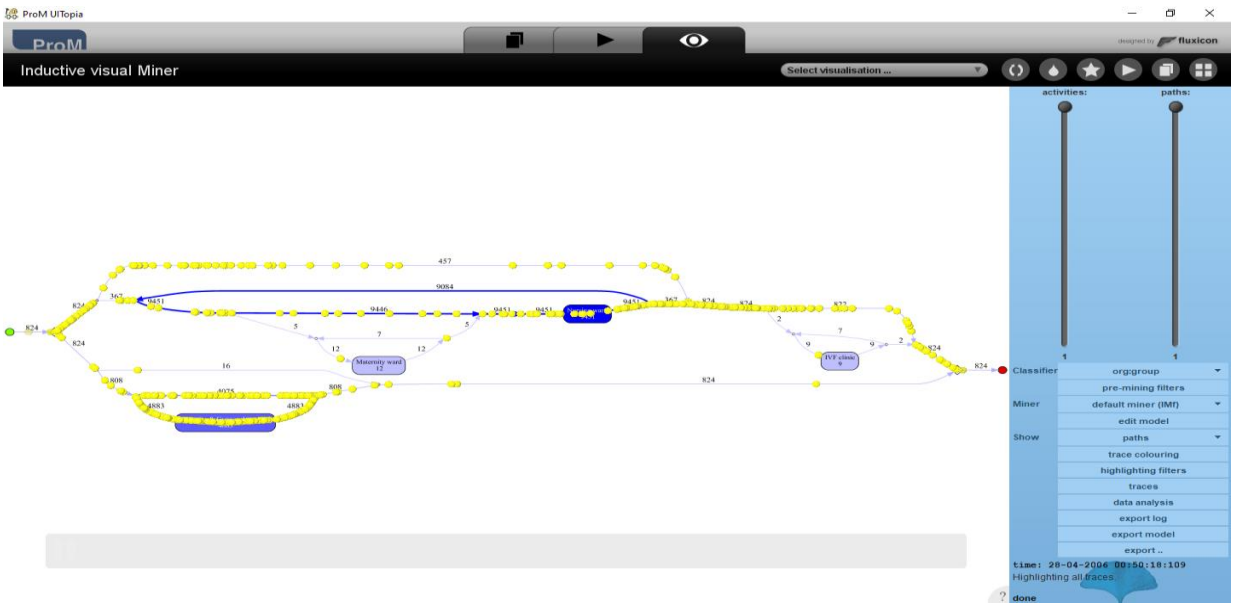
Analysis using 'Inductive Visual Miner'

a. Apply 'Mine with Inductive Visual Miner' for Hospital log data, change classifier as Org:group. Not much readable in screenshot below but will apply filter in next step



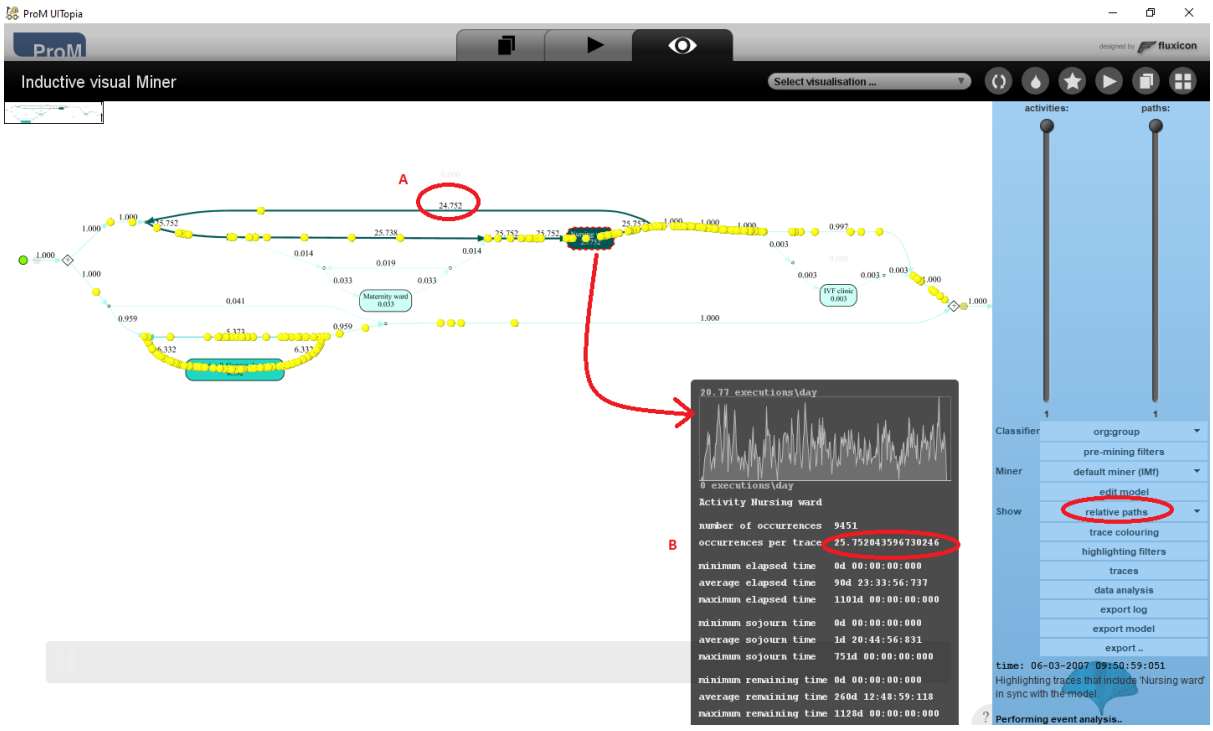
b. Selected 'Section 2' under 'section' attribute category in 'Pre-mining filters' and then apply. This filters gets only cases related to Gynecology org. Below screenshots shows the filter applied and the default output.





c. In ‘Show’ selected the ‘Paths and Queue lengths”, below screenshot shows ‘Nursing Ward” has high queue length and followed by “Obstetrics & Gynecology clinic”

d. In ‘Show’ selected the ‘Relative Paths”, below screenshot shows ‘Nursing Ward’ has high path impact.



Findings and Recommendations

- ✓ **"Nursing Ward' in Section 2** (Gynecology department) **has highest queue lengths** (11.46 occurrences/trace) based on figure #12 above
- ✓ Based on figure #13 above, the cases are repeatedly visiting 'Nursing Ward" i.e., A (24.752)*100/B (25.752) = **96.12% cases are repeatedly visiting the "Nursing Ward'**
- ✓ **Recommended to increase the number of "nursing wards"** or **avoid repeating workflows within "Nursing Ward"** to improve the performance

2.0 Analysis by Group Member: Kavya Kore (n10840371)

ProM Lite tool has been used for analysis of hospital log.

Process Mining Question

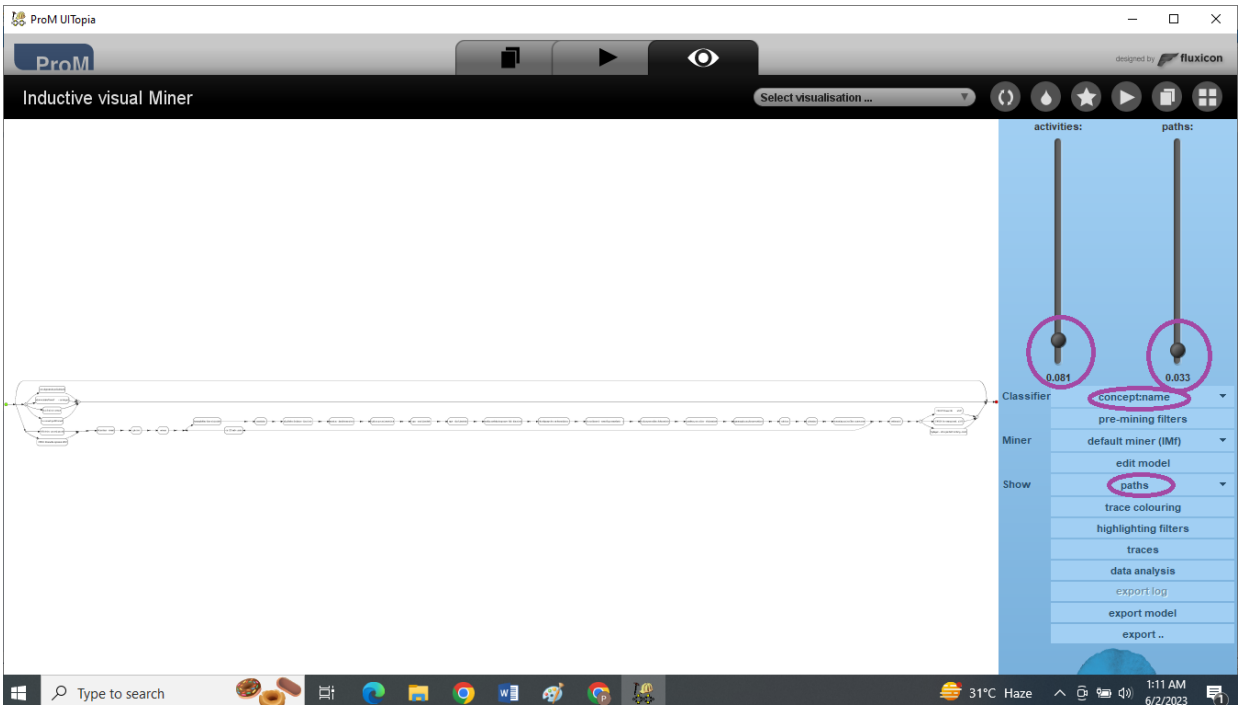
The activities in the log are in Dutch language and there are many activities which requires domain knowledge for any recommendations, so the current study is on group, the analysis various departments targeting the following questions.

Q1. Which group has more path deviations?

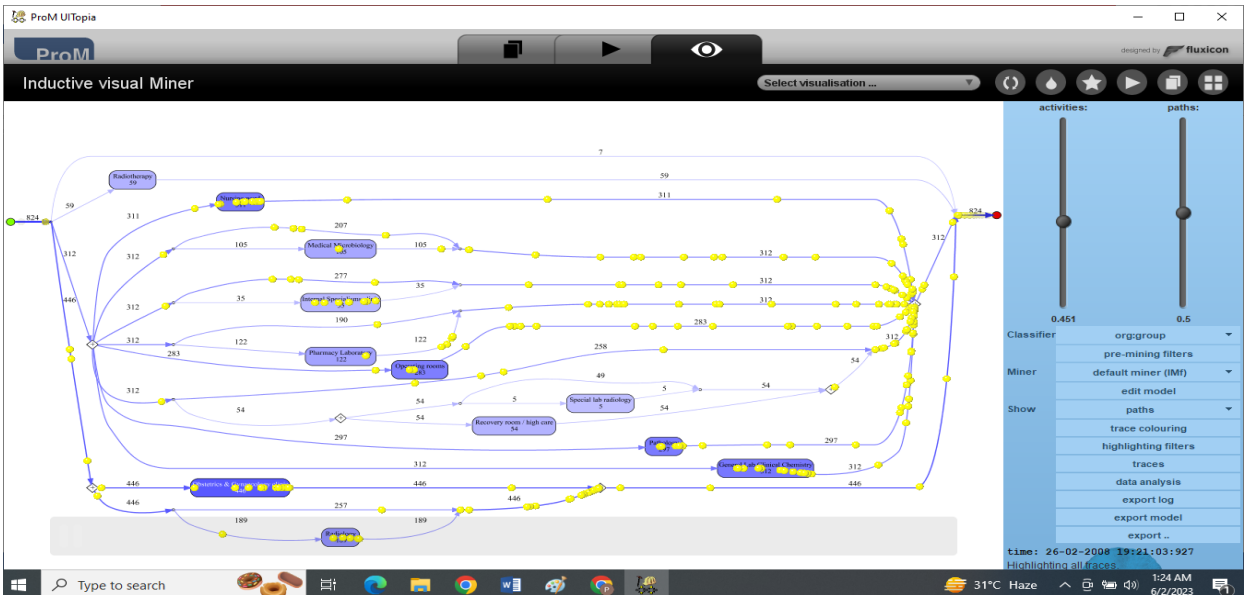
Q2. Which org group has more queues?

Process Analysis

I. Apply 'Mine with Inductive Visual Miner' for Hospital log data. The screenshot below shows the flow after applying filters for path and activities close to zero, to accommodate the whole view.

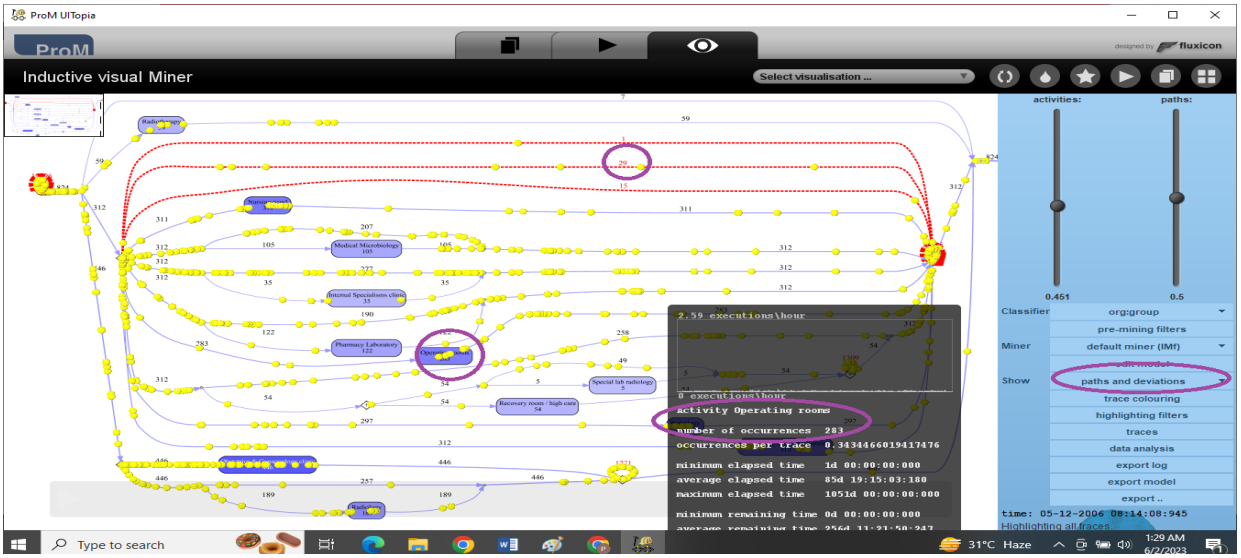


II. The view shows that the activities are too large and bringing them down to zero even to view. Moreover, analysis of activities in the medical field requires lot of domain knowledge. In next step will change the classifier to group for further analysis

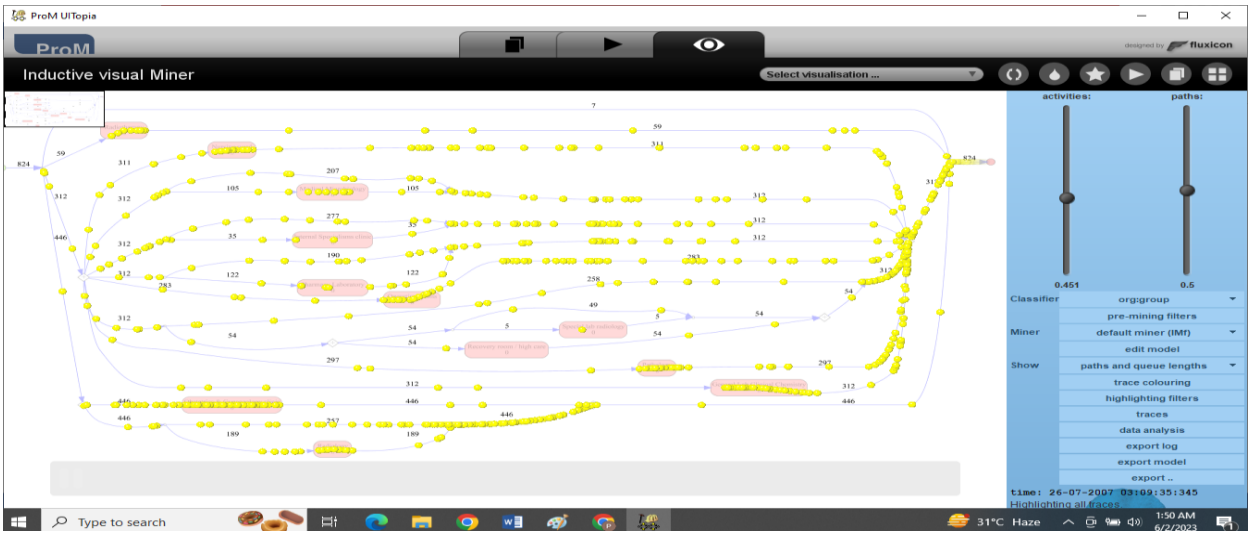


From the screenshot above the paths and activities are of reasonable for further analysis towards solving our mining question.

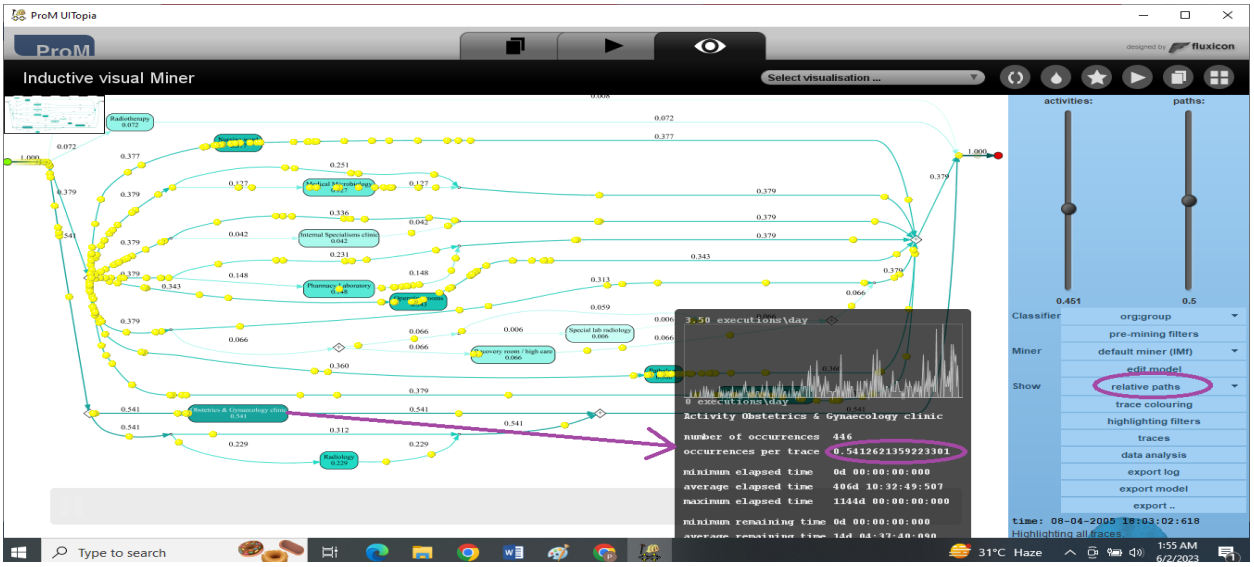
III. Apply Path and Deviations in "Show" filter and see the groups which are deviating the process largely with this model. From screenshot below the 'operating rooms' have most deviations compared to other groups. But overall, 29 deviations out of 283 i.e., **10% deviations**.



IV. Apply 'Paths and queue lengths' in "Show" filter and found not queue waiting within any group



V. Apply 'Relative paths' in "Show" filter and found that "Obstetrics & Gynecology Clinic" has highest traces compared to other groups.



Recommendations

The analysis has been performed and found that there are no waiting times though there are deviations in 'operating rooms' group which is recommended to fix, Refer figure#23 above. The relative paths shows that the "Obstetrics & Gynecology Clinic" department is frequently used at 54%, so increasing the capacity and maintenance should be critical for success.

3.0 Analysis by Group Member: Meet Chiragkumar Solanki (n11569751)

Introduction

The purpose of this report is to present the findings of the analysis conducted on the 'Hospital_log.xes' dataset, which represents the process of the Gynaecology department in a Dutch Academic Hospital. The objective of the analysis is to identify areas of improvement in the process and provide data-informed recommendations for process enhancement.

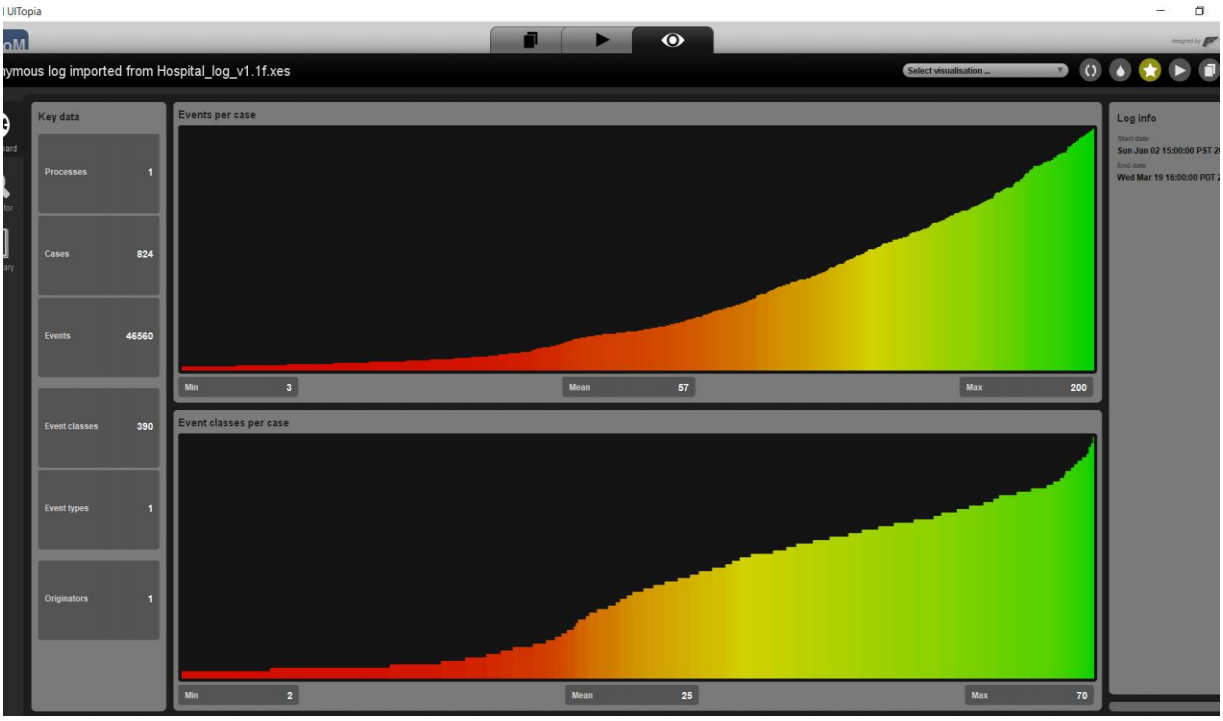
Analysis Methodology

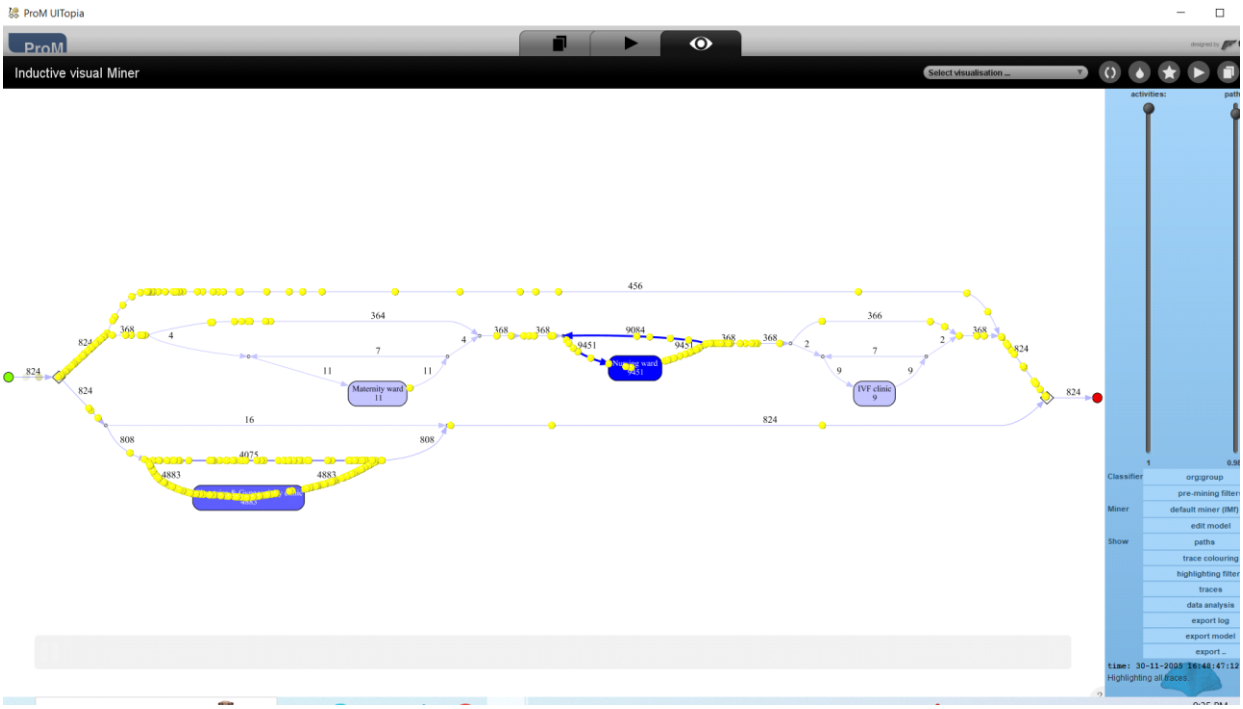
The analysis was carried out using ProM Lite, a process mining tool, to explore the log data and derive insights. The following steps were followed in the analysis:

Data Preparation and Filtering

To focus the analysis on the Gynaecology department cases, the log data was filtered based on the assumption that a specific attribute (e.g., "Department") distinguishes the Gynaecology department cases from other departments. By applying the filtering rule, only the cases relevant to the Gynaecology department were included in the analysis.

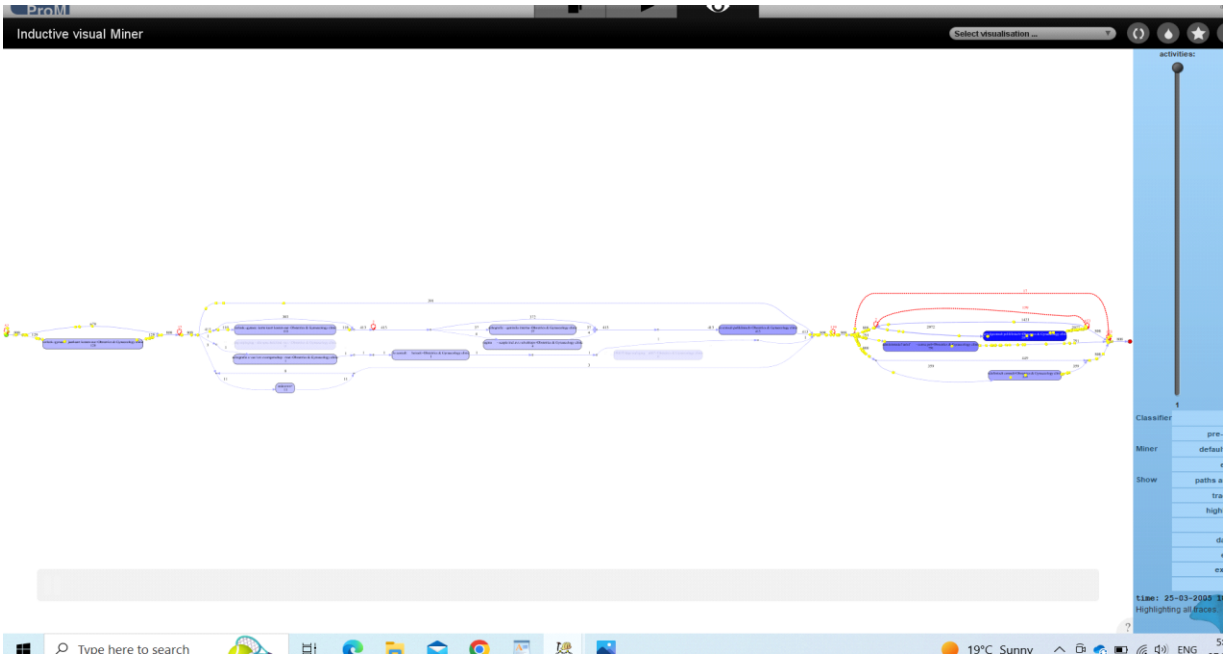
Below is the image of process for gync cases and events.





Process Discovery

The Inductive Miner algorithm was employed to discover the process model from the log data. The resulting process model provides a visual representation of the Gynaecology department process, showcasing the activities and their relationships. The process model helps in understanding the overall flow of the process and identifying bottlenecks or inefficiencies.



Process Analysis

The process analysis involved calculating various performance metrics to gain insights into the Gynaecology department process. The following metrics were considered:

Case Durations: The time taken for each Gynaecology case from start to completion.

Waiting Times: The duration patients spend waiting between activities or phases.

Visualizations, including process maps, performance metrics charts, and queueing simulation results, were generated to present the analysis findings. These visualizations aid in identifying areas of improvement and understanding the performance of the process.

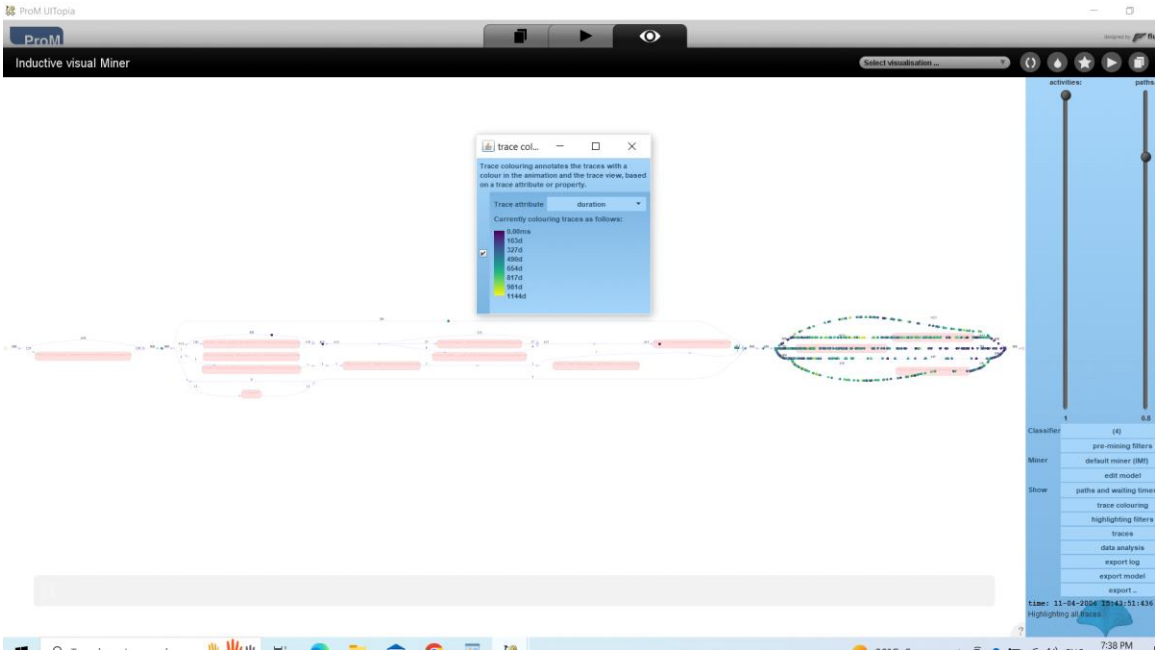
Identification of Improvement Opportunities

Based on the analysis results, several areas of improvement were identified in the Gynaecology department process. These include:

High waiting times between specific activities, indicating potential inefficiencies in resource allocation or scheduling.

Overutilization or underutilization of certain resources, leading to bottlenecks or idle capacity.

Variations or delays in specific phases of the process, indicating potential issues in diagnosis or treatment protocols.



It is seen that nursing ward has highest waiting time from gynecology department following maternity ward,ivf clinic.

Data-Informed Recommendation

Based on the identified improvement opportunities, the data-informed process improvement recommendation is as follows:

Implement a centralized scheduling system: Introduce a centralized scheduling system that optimizes the allocation of resources and minimizes waiting times between activities. This system should consider factors such as resource availability, patient priorities, and treatment requirements. By efficiently scheduling appointments and procedures, the waiting times can be reduced, resource utilization can be optimized, and the overall process flow can be improved.

This recommendation is supported by the analysis insights, which revealed significant waiting times and resource utilization issues. By implementing a centralized scheduling system, the Gynaecology department can enhance patient flow, improve resource allocation, and enhance the overall efficiency of the process.

Assumptions and Filtering Rules

This assumption ensures that the analysis focuses specifically on the Gynaecology department process and provides accurate insights for improvement

Conclusion

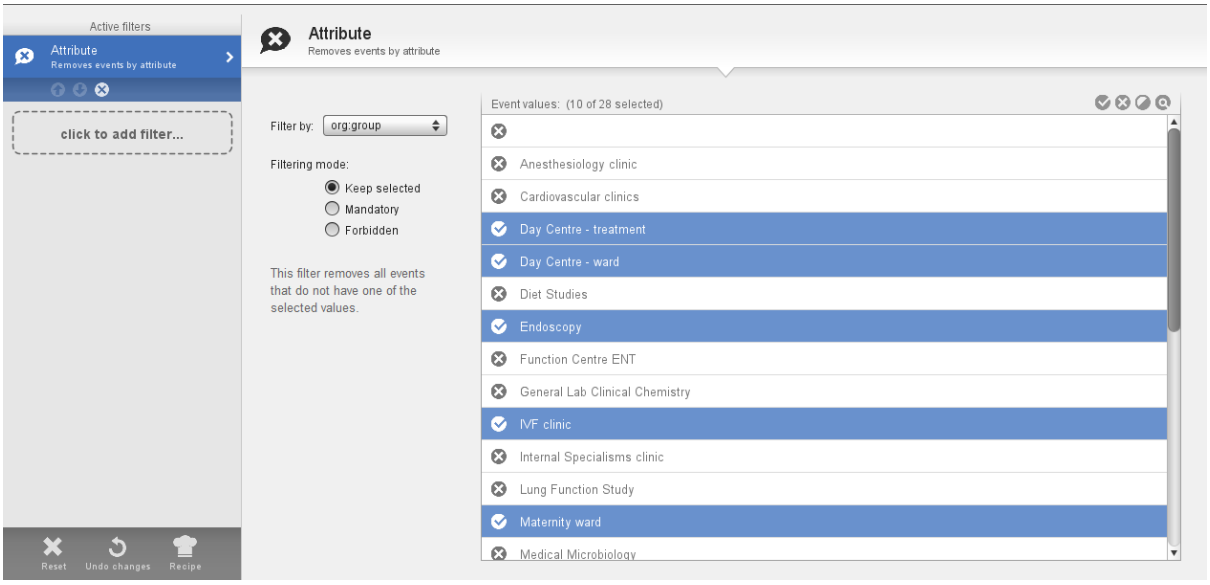
In conclusion, the analysis of the 'Hospital_log.xes' dataset provided valuable insights into the Gynaecology department process. By leveraging process mining techniques, performance metrics calculations, and visualizations, areas of improvement were identified. The data-informed recommendation of implementing a centralized scheduling system has the potential to enhance patient flow, reduce waiting times, and optimize resource utilization.

4.0 Analysis by Group Member: Deborah Mei Har Chan (n9781846)

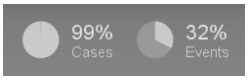
The hospital log taken from a Dutch Academic Hospital, was analyzed through the process mining tool, Disco. The "Hospital_log_v1.1f.xes" file consists of 46,560 events in 824 cases. Upon importing the log, the process model automatically generated was at 10% activities and 0% path, showing a very non-complex flow of processes whereby only the most frequent activities and most dominant paths are shown. There is also a high number of attributes present in the log file, therefore, it is imperative to carefully identify and select attributes which will be highly applicable for the analysis.

Using the filter settings, the list of filters for "attributes" were each looked into to identify which ones were able to be understood. The "org:group" attribute, consisting of 27 departments, was first selected, and then further filtered to include only the event values that are most directly related to gynaecology procedures, treatments, diagnostics, and care. These event values are:Day Centre - treatment

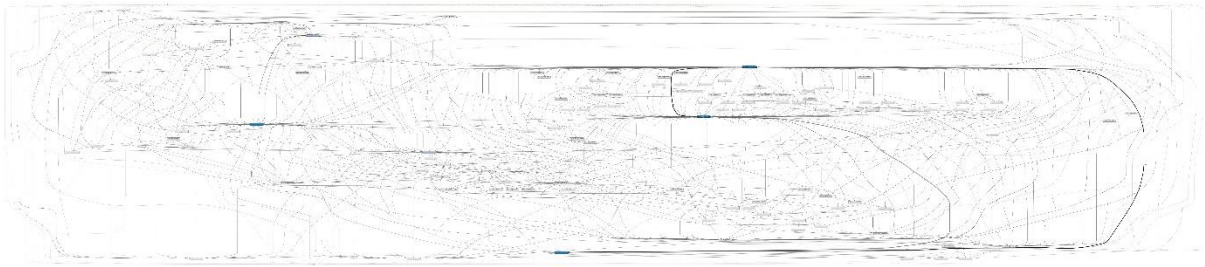
1. Day Centre - ward
2. Endoscopy
3. IVF clinic
4. Maternity ward
5. Nursing ward
6. Obstetrics & Gynaecology clinic
7. Operating Rooms
8. Recovery room/high care
9. Special lab radiology



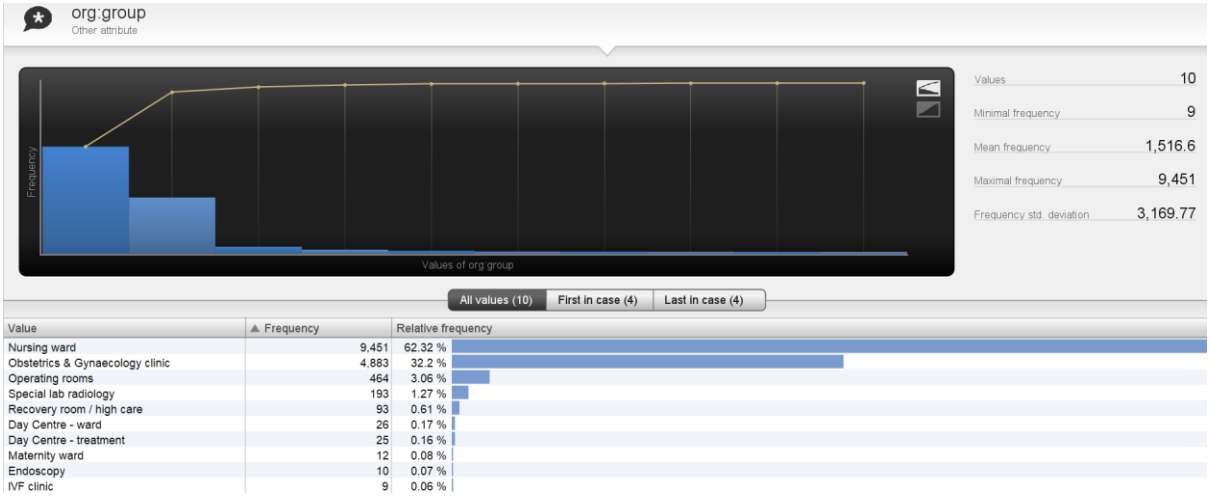
Applying these filters, the percentage indicators show that 99% of cases and 32% events of the original data set remains after applying this filter.



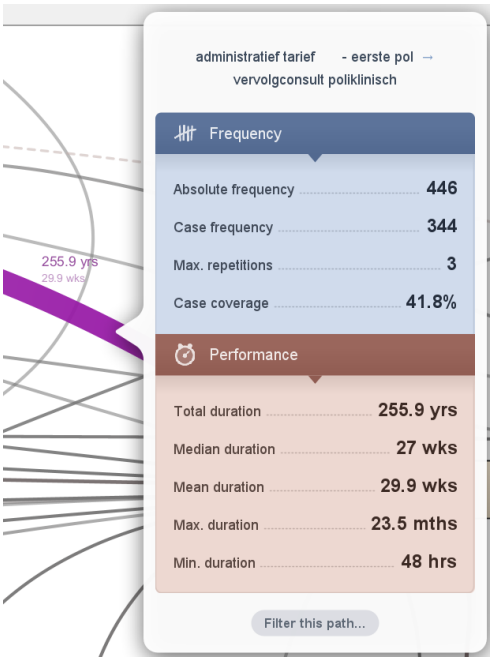
The process model is generated as shown below, with the sliders set to 100% to capture all activities and paths.



In the “Statistics” tab, the graph of the “org:group” attribute shows that the “Nursing ward” has the highest frequency of 9,451, followed by the “Obstetrics & Gynaecology clinic” with a frequency of 4,883.



In terms of path, the highest impact for delays falls within the “Nursing ward” department, taking a cumulative time of 255.9 years, and a mean duration of 29.9 weeks.



Through this analysis, it is clearly shown that the bottleneck occurs mainly in the “Nursing ward” and a recommended solution to tackle this issue would be to implement an obstetric triage system. The triage system is used in the healthcare industry to categorise patients based on the urgency of their condition. Considering the highest delay falls in the path from administration to consultation, utilising the obstetric triage system would help expedite the assessment, prioritisation, and management of patients, and thus, providing patients with a more effective service.