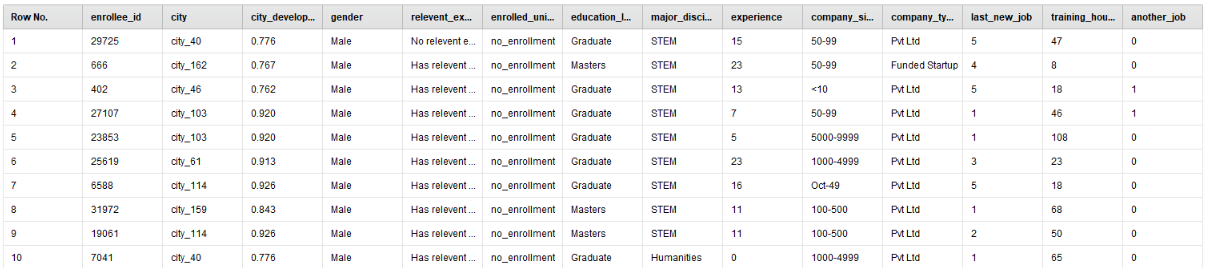
405801 Anna Garoufali
392326 Na Young Ahn
414382 Jahnavi Jaiswal

Assignment Part 1

Business Process Intelligence

Question 1 - Preprocessing the data

Please refer to file: **dataset1.csv**



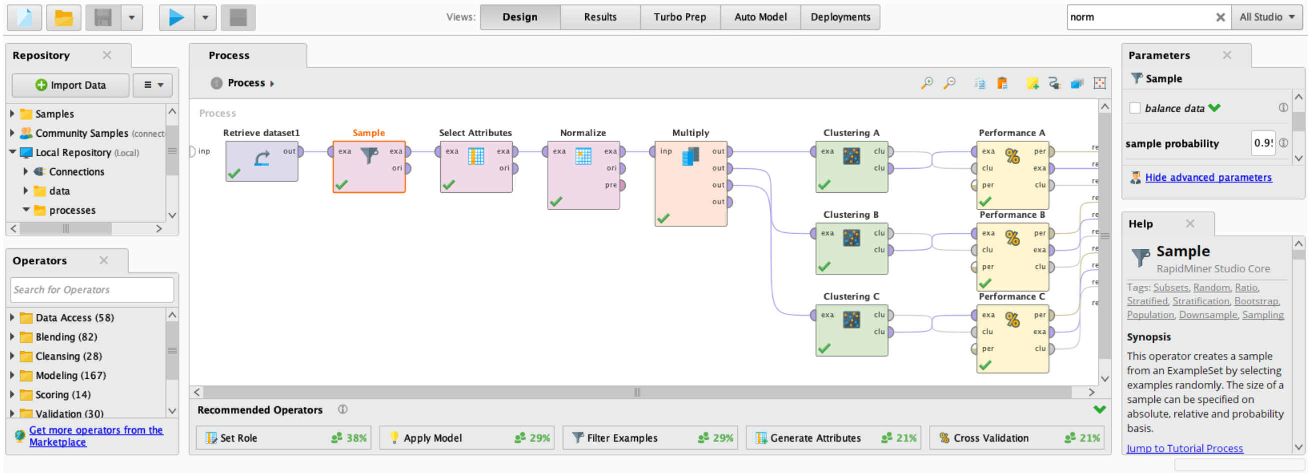
Row No.	enrollee_id	city	city_develop...	gender	relevent_ex...	enrolled_uni...	education_l...	major_disci...	experience	company_si...	company_ty...	last_new_job	training_hou...	another_job
1	29725	city_40	0.776	Male	No relevent e...	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	5	47	0
2	666	city_162	0.767	Male	Has relevent ...	no_enrollment	Masters	STEM	23	50-99	Funded Startup	4	8	0
3	402	city_46	0.762	Male	Has relevent ...	no_enrollment	Graduate	STEM	13	<10	Pvt Ltd	5	18	1
4	27107	city_103	0.920	Male	Has relevent ...	no_enrollment	Graduate	STEM	7	50-99	Pvt Ltd	1	46	1
5	23853	city_103	0.920	Male	Has relevent ...	no_enrollment	Graduate	STEM	5	5000-9999	Pvt Ltd	1	108	0
6	25619	city_61	0.913	Male	Has relevent ...	no_enrollment	Graduate	STEM	23	1000-4999	Pvt Ltd	3	23	0
7	6588	city_114	0.926	Male	Has relevent ...	no_enrollment	Graduate	STEM	16	Oct-49	Pvt Ltd	5	18	0
8	31972	city_159	0.843	Male	Has relevent ...	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	1	68	0
9	19061	city_114	0.926	Male	Has relevent ...	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	2	50	0
10	7041	city_40	0.776	Male	Has relevent ...	no_enrollment	Graduate	Humanities	0	1000-4999	Pvt Ltd	1	65	0

Figure 1: Snapshot of the first 10 rows of dataset1

Question 2 - Clustering

Student ID used as seed: **405801**

1.



□ balance data ✓

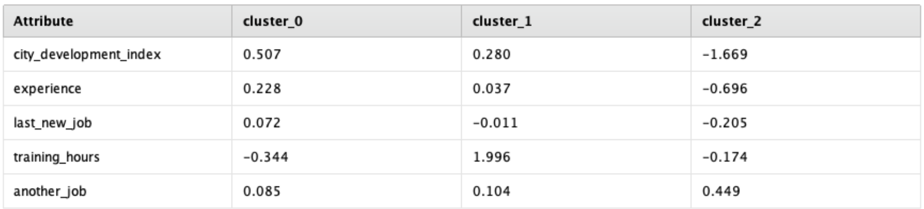
sample probability 0.9!

Synopsis

This operator creates a sample from an ExampleSet by selecting examples randomly. The size of a sample can be specified on absolute, relative and probability basis.

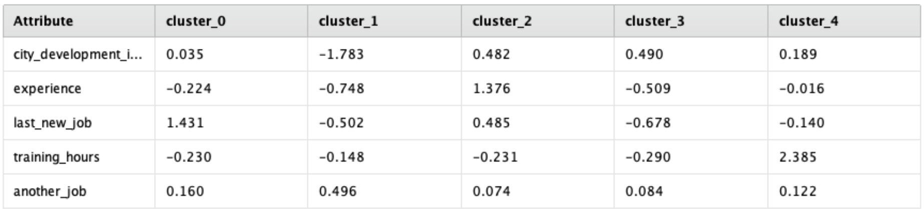
Figure 2: Design tab of processQ21

a.



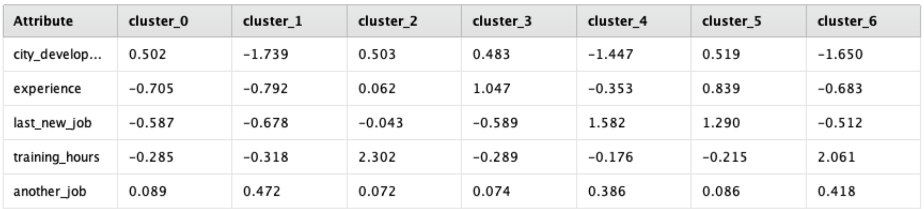
Attribute	cluster_0	cluster_1	cluster_2
city_development_index	0.507	0.280	-1.669
experience	0.228	0.037	-0.696
last_new_job	0.072	-0.011	-0.205
training_hours	-0.344	1.996	-0.174
another_job	0.085	0.104	0.449

Figure 3: Centroids of clustering A



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
city_development_i...	0.035	−1.783	0.482	0.490	0.189
experience	−0.224	−0.748	1.376	−0.509	−0.016
last_new_job	1.431	−0.502	0.485	−0.678	−0.140
training_hours	−0.230	−0.148	−0.231	−0.290	2.385
another_job	0.160	0.496	0.074	0.084	0.122

Figure 4: Centroids of clustering B



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6
city_develop...	0.502	−1.739	0.503	0.483	−1.447	0.519	−1.650
experience	−0.705	−0.792	0.062	1.047	−0.353	0.839	−0.683
last_new_job	−0.587	−0.678	−0.043	−0.589	1.582	1.290	−0.512
training_hours	−0.285	−0.318	2.302	−0.289	−0.176	−0.215	2.061
another_job	0.089	0.472	0.072	0.074	0.386	0.086	0.418

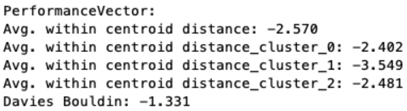
Figure 5: Centroids of clustering C

b. Coherence measures how close the objects are within the same cluster. Using the Cluster Distance Performance operator, we find cluster with the smallest average within centroid distance. Checking values from figures 6 ,7 and 8 we get the following answer:

Clustering	most coherent cluster
A	cluster 0
B	cluster 3
C	cluster 0

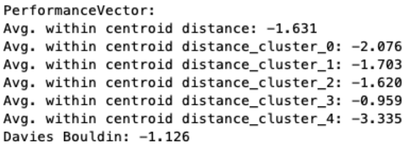
Table 1: Most coherent cluster for each clustering

c. **Clustering C is the best clustering** as it has the smallest absolute Davies-Bouldin index -1.122, which indicates that the clusters have low intra-cluster distances and high inter-cluster distances.



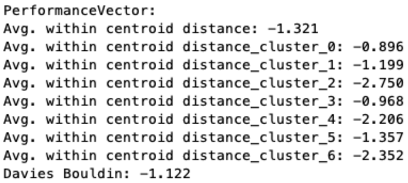
```
PerformanceVector:
Avg. within centroid distance: −2.570
Avg. within centroid distance_cluster_0: −2.402
Avg. within centroid distance_cluster_1: −3.549
Avg. within centroid distance_cluster_2: −2.481
Davies Bouldin: −1.331
```

Figure 6: Cluster distance performance of clustering A



```
PerformanceVector:
Avg. within centroid distance: −1.631
Avg. within centroid distance_cluster_0: −2.076
Avg. within centroid distance_cluster_1: −1.703
Avg. within centroid distance_cluster_2: −1.620
Avg. within centroid distance_cluster_3: −0.959
Avg. within centroid distance_cluster_4: −3.335
Davies Bouldin: −1.126
```

Figure 7: Cluster distance performance of clustering B

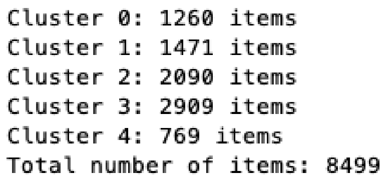


```
PerformanceVector:
Avg. within centroid distance: −1.321
Avg. within centroid distance_cluster_0: −0.896
Avg. within centroid distance_cluster_1: −1.199
Avg. within centroid distance_cluster_2: −2.750
Avg. within centroid distance_cluster_3: −0.968
Avg. within centroid distance_cluster_4: −2.206
Avg. within centroid distance_cluster_5: −1.357
Avg. within centroid distance_cluster_6: −2.352
Davies Bouldin: −1.122
```

Figure 8: Cluster distance performance of clustering C

2.

As seen from the pictures on the next page, **cluster 3** contains **2909 instances** and **cluster 4** contains **769** instances, biggest and smallest number of items respectively.



```
Cluster 0: 1260 items
Cluster 1: 1471 items
Cluster 2: 2090 items
Cluster 3: 2909 items
Cluster 4: 769 items
Total number of items: 8499
```

Figure 9: Smallest and biggest cluster of clustering B

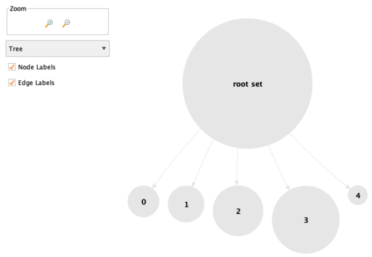


Figure 10: Smallest and biggest cluster of clustering B

3.

Redundant attributes can not differentiate clusters or instances very well. For attribute **another job** we observe small distances between centroid mean values for all clusters. Hence, it does not contribute significantly in differentiating between clusters. Additionally, **another job** does not provide much information to clusterization with its boolean values. Similarly, we observe centroids for all clusters except cluster 4 are mustered for the attribute **training hours**. However, as cluster 4 is distinguished, we do not consider it as redundant as **another job**.

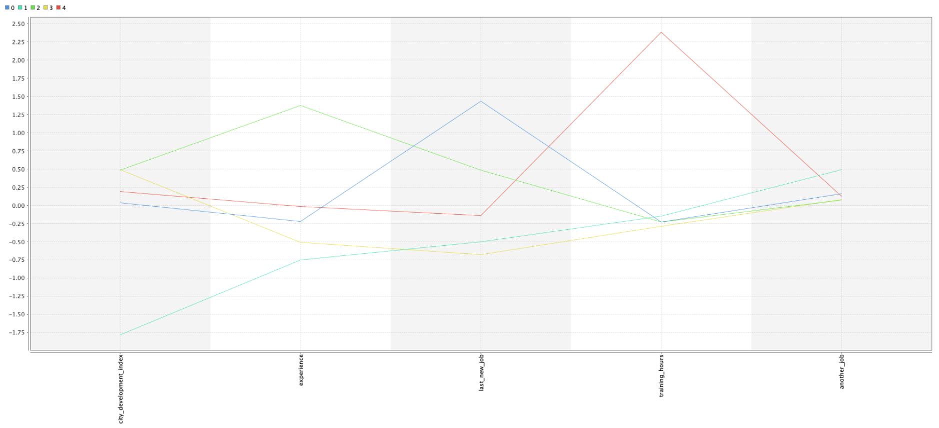


Figure 11: Clusters means for clustering B

4.

Sample 95 percent of instances using the student ID: **405801**. Afterwards, divide it into Subset 1: **Candidates that are looking for a job change** and Subset 2: **Candidates that are not looking for a job change**.

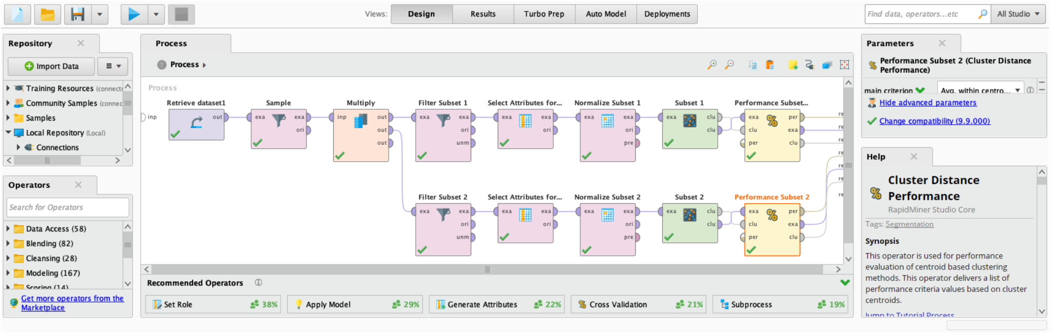


Figure 12: Design tab of processQ24

5.

For each subset, and for each centroid attribute, we group the clusters with similar centroid mean values. Then we define "similar" for values within the group and "dissimilar" if the compared cluster centroid value is outside of the group.

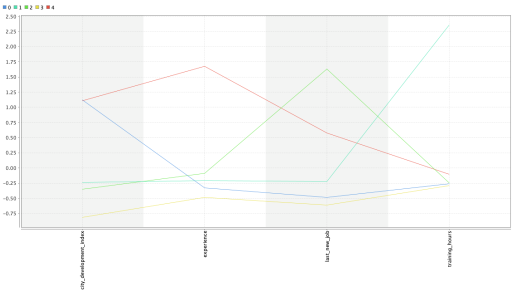


Figure 13: Candidates that are looking for a job change

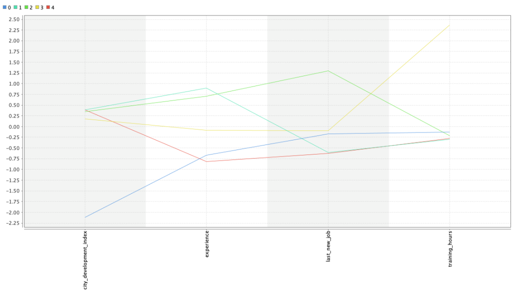
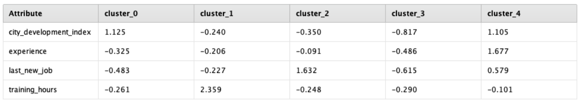
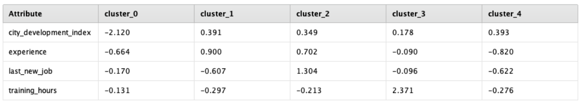


Figure 14: Candidates that are not looking for a job change



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
city_development_index	1.125	−0.240	−0.350	−0.817	1.105
experience	−0.325	−0.206	−0.091	−0.486	1.677
last_new_job	−0.483	−0.227	1.632	−0.615	0.579
training_hours	−0.261	2.359	−0.248	−0.290	−0.101

Figure 15: Centroids of subset 1



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
city_development_index	−2.120	0.391	0.349	0.178	0.393
experience	−0.664	0.900	0.702	−0.090	−0.820
last_new_job	−0.170	−0.607	1.304	−0.096	−0.622
training_hours	−0.131	−0.297	−0.213	2.371	−0.276

Figure 16: Centroids of subset 2

Subset 1	city development index	experience	last new job	training hours
groups with similar centroid values	Clusters 1,2,3 Clusters 0,4	Clusters 0,1,2,3	Clusters 0,1,3	Clusters 0,2,3,4
dissimilar cluster(s)		Cluster 4	Cluster 2 Cluster 4	Cluster 1

Table 2: Similarities and dissimilarities between the clusters on subset 1

Subset 2	city development index	experience	last new job	training hours
groups with similar centroid values	Clusters 1,2,3,4	Clusters 1,2 Cluster 0,4	Clusters 1,4 Cluster 0,3	Clusters 0,1,2,4
dissimilar cluster(s)	Cluster 0	Cluster 3	Cluster 2	Cluster 3

Table 3: Similarities and dissimilarities between the clusters on subset 2

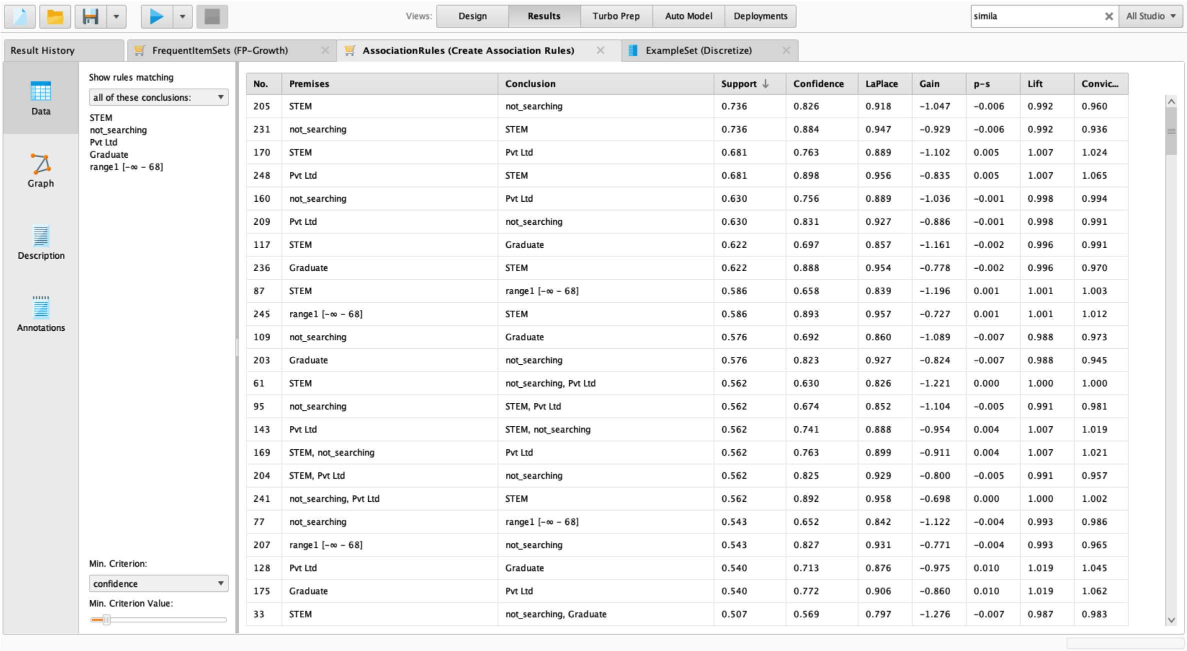
Question 3 - Association Rule

1.

Please refer to file: **dataset2.csv**

2.

The same student ID **405801** is used as a seed and 95% of the instances are sampled. And according to the setting as per the question, the list of rules with their support and confidence in decreasing order w.r.t. their support are as follows:

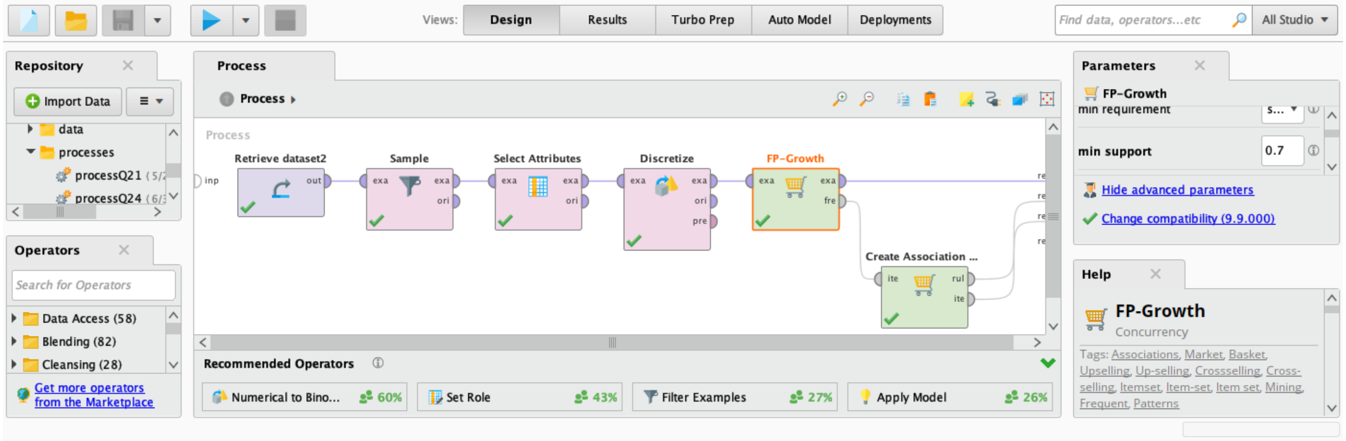

Views: Design Results Turbo Prep Auto Model Deployments
simila
All Studio
Result History
FrequentItemSets (FP-Growth)
AssociationRules (Create Association Rules)
ExampleSet (Discretize)
Data
Graph
Description
Annotations

Show rules matching
all of these conclusions:
STEM
not_searching
Pvt Ltd
Graduate
range1 [-∞ – 68]

Min. Criterion: confidence
Min. Criterion Value:

No.	Premises	Conclusion	Support ↓	Confidence	LaPlace	Gain	p-s	Lift	Convic...
205	STEM	not_searching	0.736	0.826	0.918	-1.047	-0.006	0.992	0.960
231	not_searching	STEM	0.736	0.884	0.947	-0.929	-0.006	0.992	0.936
170	STEM	Pvt Ltd	0.681	0.763	0.889	-1.102	0.005	1.007	1.024
248	Pvt Ltd	STEM	0.681	0.898	0.956	-0.835	0.005	1.007	1.065
160	not_searching	Pvt Ltd	0.630	0.756	0.889	-1.036	-0.001	0.998	0.994
209	Pvt Ltd	not_searching	0.630	0.831	0.927	-0.886	-0.001	0.998	0.991
117	STEM	Graduate	0.622	0.697	0.857	-1.161	-0.002	0.996	0.991
236	Graduate	STEM	0.622	0.888	0.954	-0.778	-0.002	0.996	0.970
87	STEM	range1 [-∞ – 68]	0.586	0.658	0.839	-1.196	0.001	1.001	1.003
245	range1 [-∞ – 68]	STEM	0.586	0.893	0.957	-0.727	0.001	1.001	1.012
109	not_searching	Graduate	0.576	0.692	0.860	-1.089	-0.007	0.988	0.973
203	Graduate	not_searching	0.576	0.823	0.927	-0.824	-0.007	0.988	0.945
61	STEM	not_searching, Pvt Ltd	0.562	0.630	0.826	-1.221	0.000	1.000	1.000
95	not_searching	STEM, Pvt Ltd	0.562	0.674	0.852	-1.104	-0.005	0.991	0.981
143	Pvt Ltd	STEM, not_searching	0.562	0.741	0.888	-0.954	0.004	1.007	1.019
169	STEM, not_searching	Pvt Ltd	0.562	0.763	0.899	-0.911	0.004	1.007	1.021
204	STEM, Pvt Ltd	not_searching	0.562	0.825	0.929	-0.800	-0.005	0.991	0.957
241	not_searching, Pvt Ltd	STEM	0.562	0.892	0.958	-0.698	0.000	1.000	1.002
77	not_searching	range1 [-∞ – 68]	0.543	0.652	0.842	-1.122	-0.004	0.993	0.986
207	range1 [-∞ – 68]	not_searching	0.543	0.827	0.931	-0.771	-0.004	0.993	0.965
128	Pvt Ltd	Graduate	0.540	0.713	0.876	-0.975	0.010	1.019	1.045
175	Graduate	Pvt Ltd	0.540	0.772	0.906	-0.860	0.010	1.019	1.062
33	STEM	not_searching, Graduate	0.507	0.569	0.797	-1.276	-0.007	0.987	0.983

Figure 17: Rules with support and confidence in decreasing order w.r.t. their support



FP-Growth
min requirement s...
min support 0.7

Figure 18: Design tab of processQ32

3.

When we look at association rules for support values greater than 0.5, we observe that most of the association rules have the relationship p → q and q → p from which we can infer a p $\iff$ q relationship. With this in mind, we chose the following three interesting rules that have a support value larger than 0.5:

1. **STEM** → **range1[−∞ - 68]** means that most of those who studied **STEM** tend to have less training hours in **range1[−∞ - 68]**.
   range1[−∞ - 68] → **STEM** means that most of those who have less training hours in **range1[−∞ - 68]** tend to have studied **STEM**.
   STEM $\iff$ **range1[−∞ - 68]** is interesting because it could be thought that one normally needs a lot of training hours before one gets used to the technically specific work setup and environment at work. Vice versa, it could be assumed that those with less training hours would not have come from such a technical background, however, with the association rules we find it is not the case.

2. **STEM** → **not_searching, Pvt Ltd** means that most of those who studied **STEM** work for a **Pvt**

Ltd and are **not_searching** for another job.
not_searching, Pvt Ltd → **STEM** means that most of those who are **not_searching** for another job and work for a **Pvt Ltd**, tend to have studied **STEM**.
STEM ⟺ **not_searching, Pvt Ltd** is interesting because STEM majors are often considered innovative and nonconforming to social conventions. However, the association rules suggest that if you are a STEM major you are likely to work for a private company and are likely to stay. Moreover, that if you are someone who is working for a private company, and not intending to seek another job, you are likely to come from a STEM background. This again disproves one's prejudice that STEM majors are likely to begin or work for a startup, whether it be after studies or as their next career step.

3. **range1[$-\infty$ - 68]** → **not_searching** means that most of those who had less training of **range1[$-\infty$ - 68]** are **not_searching** for another job.
not_searching → **range1[$-\infty$ - 68]** means that most of those who are **not_searching** for another job, tend to have had less training hours of **range1[$-\infty$ - 68]** .
range1[$-\infty$ - 68] ⟺ **not_searching** is interesting because it can be considered that those who spent a lot of hours for training are less likely to look for new opportunities and those who spent less hours for training are likely to have less opportunity cost in changing jobs. However, the association rule suggests that those who had low training hours tend not to look for another job. Moreover, those who are not looking for another job are likely to have had less training hours. This goes against the conventional idea that high training hours are correlated to long-term loyal employees.

Question 4 - Process Mining

1. Preprocessing

Student ID used for sampling: **405801**

2. Activities

a. There are 8 activities in the log.

b. Most frequently executed activity: **Validation**
Least frequently executed activity: **Personal loan collection**

3. Performance

a. Activity that takes on average the most time: **Fraud suspicion check**
The average time: 3 day 1 hour

b.

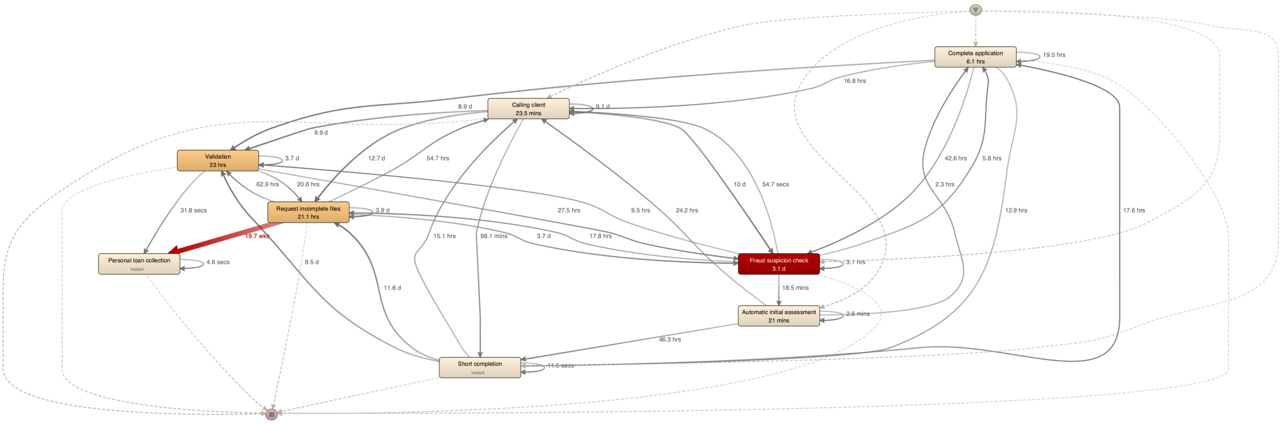
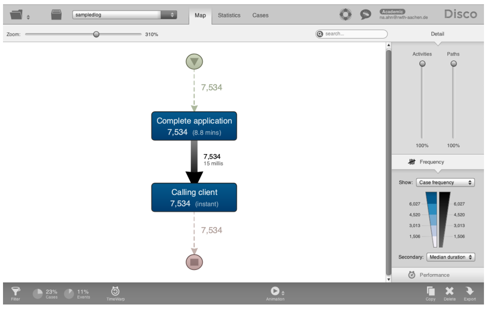


Figure 19: Process model discovered with disco with activity 100% path 100%

4. Special Behavior

a. Short completion did not appear before 30-06-2016.

b.



Zoom: 310%

Show: Case frequency

Secondary: Median duration

Figure 20: Activity 100% path 100% model of all cases that include activity **Short completion**

c. There are 2 variants of the cases that include activity **Short completion**

5. Resources

a. **User_138** takes on average the longest time of 2 days 22 hours.

b.

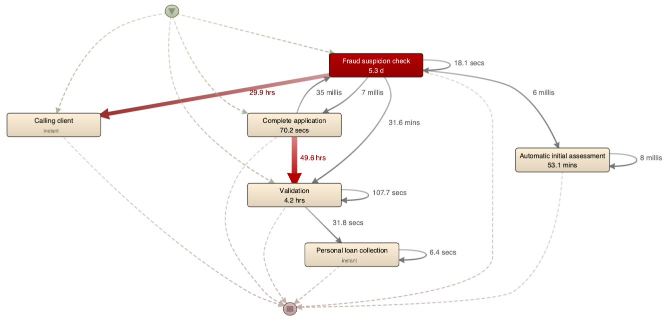


Figure 21: Activity 100% path 100% process model for resource **User_13**

c. Resource **User_138** is primarily doing **Fraud suspicion**, for, in total, 88 days and 4 hours.

6. Variants

a. The most frequent variant is **Variant 1**

b.

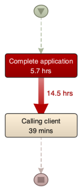


Figure 22: Activity 100% path 100% process model for variant **Variant 1**

c. The longest case of **Variant 1** is **Application_715212713**, which takes 29 days and 5 hours.

7. Other Data

a. **Car** is the most frequent reason for customers to apply for a loan.

b. **35586** cases follow this objective i.e. **Car**. Out of these cases, **3247** applied for just a limit raise of a previously approved loan.

c.

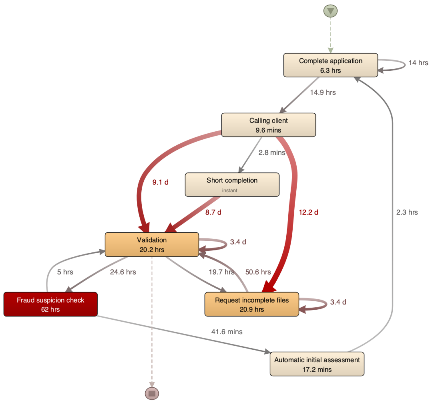
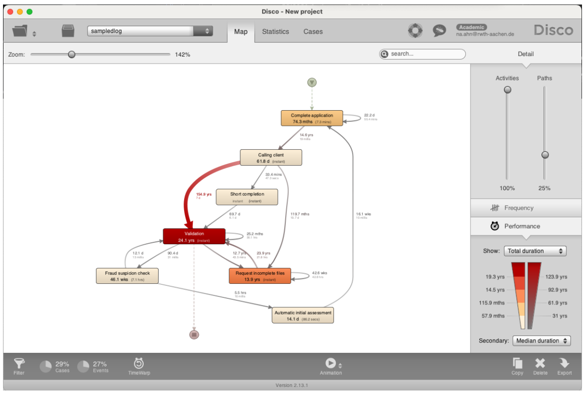


Figure 23: Activity 100% path 25% process model for all cases that follow the loan objective **Car**

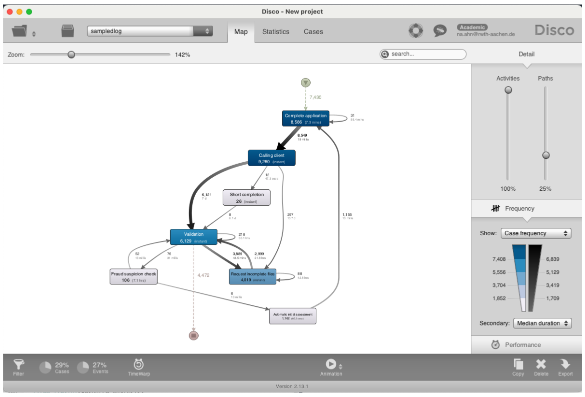
8. Analysis

a.



Show: Total duration
Secondary: Median duration

Figure 24: Activity 100% path 25% process model for all cases that follow the loan objective **Car** for bottleneck analysis by performance



Show: Case frequency
Secondary: Median duration

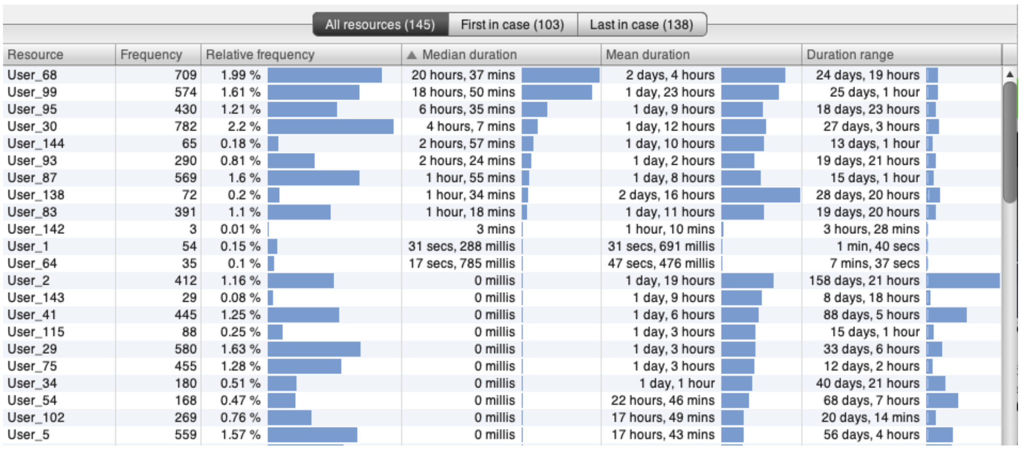
Figure 25: Activity 100% path 25% process model for all cases that follow the loan objective **Car** for bottleneck analysis by frequency

1. **Process Analysis** To improve the throughput of the process we can try to identify and resolve bottlenecks in the process. To check for bottleneck, we look at median duration and total duration, instead of mean duration, which is susceptible to extreme outliers. From figure 24 and 25 we see that activities: **Complete application**, **Validation**, and **Request incomplete files** have the most potential for improvement i.e. to reduce time duration.
   Suggestions:
   a) We observe that half of the entire cases end up under **Request incomplete application**. We can reduce the time wasted for half of the cases, by only accepting complete application.

b) Only 106 cases out of 8586 applications are sent for **Fraud suspicion check**, which takes in median 7.1 hrs to process. We can improve the situation by outsourcing validation to a third party which specialises on it to reduce time, or embed in the application process via software to reduce manual validation efforts and time wasted.

c) It seems the path from activity **Calling client** to **Validation** takes far too long, of 7 days in median, and 154.9 years in total. We can make further analysis of how and when to better reach the client and most efficient means to validate the information.

2. **Resource Analysis**

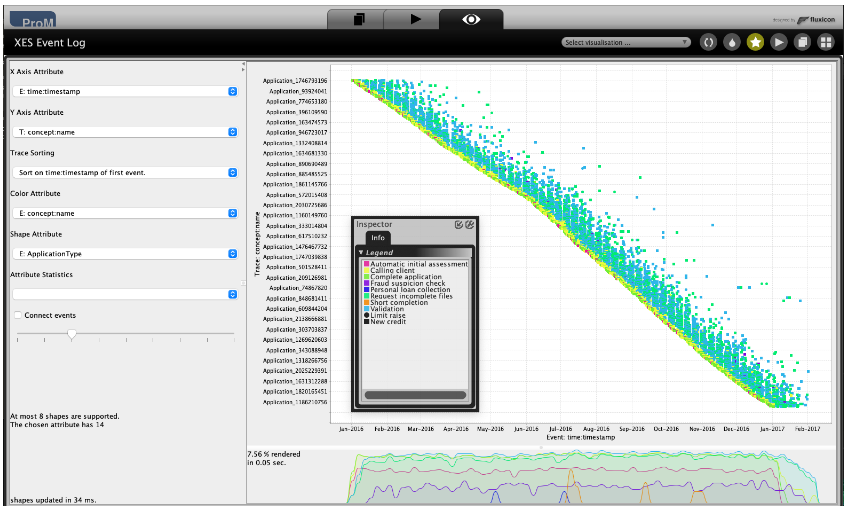

All resources (145)
First in case (103)
Last in case (138)

Resource	Frequency	Relative frequency	▲ Median duration	Mean duration	Duration range
User_68	709	1.99 %	20 hours, 37 mins	2 days, 4 hours	24 days, 19 hours
User_99	574	1.61 %	18 hours, 50 mins	1 day, 23 hours	25 days, 1 hour
User_95	430	1.21 %	6 hours, 35 mins	1 day, 9 hours	18 days, 23 hours
User_30	782	2.2 %	4 hours, 7 mins	1 day, 12 hours	27 days, 3 hours
User_144	65	0.18 %	2 hours, 57 mins	1 day, 10 hours	13 days, 1 hour
User_93	290	0.81 %	2 hours, 24 mins	1 day, 2 hours	19 days, 21 hours
User_87	569	1.6 %	1 hour, 55 mins	1 day, 8 hours	15 days, 1 hour
User_138	72	0.2 %	1 hour, 34 mins	2 days, 16 hours	28 days, 20 hours
User_83	391	1.1 %	1 hour, 18 mins	1 day, 11 hours	19 days, 20 hours
User_142	3	0.01 %	3 mins	1 hour, 10 mins	3 hours, 28 mins
User_1	54	0.15 %	31 secs, 288 millis	31 secs, 691 millis	1 min, 40 secs
User_64	35	0.1 %	17 secs, 785 millis	47 secs, 476 millis	7 mins, 37 secs
User_2	412	1.16 %	0 millis	1 day, 19 hours	158 days, 21 hours
User_143	29	0.08 %	0 millis	1 day, 9 hours	8 days, 18 hours
User_41	445	1.25 %	0 millis	1 day, 6 hours	88 days, 5 hours
User_115	88	0.25 %	0 millis	1 day, 3 hours	15 days, 1 hour
User_29	580	1.63 %	0 millis	1 day, 3 hours	33 days, 6 hours
User_75	455	1.28 %	0 millis	1 day, 3 hours	12 days, 2 hours
User_34	180	0.51 %	0 millis	1 day, 1 hour	40 days, 21 hours
User_54	168	0.47 %	0 millis	22 hours, 46 mins	68 days, 7 hours
User_102	269	0.76 %	0 millis	17 hours, 49 mins	20 days, 14 mins
User_5	559	1.57 %	0 millis	17 hours, 43 mins	56 days, 4 hours

Figure 26: Resource event classes by median duration

To increase the throughput of the process, we can also consider better distribution and increasing efficiency of the resources e.g. employees or **User(s)**. As we consider the Resource event classes as shown on figure 26, we observe that median duration for certain users are significantly longer than others. e.g. user_68, user_99. This can be improved by providing more training e.g. time management, performance improving skills to such users to increase employee productivity and faster processing of loan applications.

b.



X Axis Attribute: E: time:timestamp

Y Axis Attribute: T: concept:name

Trace Sorting: Sort on time:timestamp of first event.

Color Attribute: E: concept:name

Shape Attribute: E: ApplicationType

Attribute Statistics:

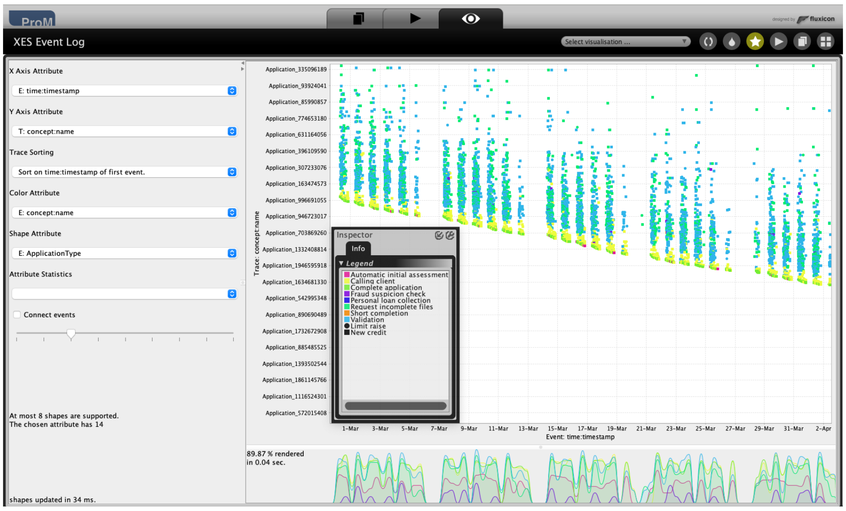
□ Connect events

At most 8 shapes are supported.
The chosen attribute has 14

shapes updated in 34 ms.

7.56 % rendered
in 0.05 sec.

Figure 27: Dotted chart via ProM for the global timeframe



X Axis Attribute: E: time:timestamp

Y Axis Attribute: T: concept:name

Trace Sorting: Sort on time:timestamp of first event.

Color Attribute: E: concept:name

Shape Attribute: E: ApplicationType

Attribute Statistics:

□ Connect events

At most 8 shapes are supported.
The chosen attribute has 14

shapes updated in 34 ms.

89.87 % rendered
in 0.04 sec.

Figure 28: Dotted chart via ProM for the local timeframe: March

We create the dotted plot for the unsampled XES event log with event timestamp as the X axis and instances or applications as out Y axis. Then we sort the trace according to the time stamp of the first event and we set various colors for different activities. As a result, we get a dotted plot that shows relatively linear behavior showing when each application was started and finished. When we zoom in the timestamp as in 28 we observe a regular gap on Sundays and reduced amount of instances i.e. work on Saturdays. We can also see a gap during holidays such as Christmas week, which infers the company where the data was gathered observed could be located in a region where they celebrate Christmas. Thus, we can deduce that no work or relatively low amount of work is handled and application is not processed during off work hours, Sundays, and holidays.