XES Survey Results

IEEE TFPM XES Working Group: **Moe Thandar Wynn**, Julian Lebherz, Wil van der Aalst, Rafael Accorsi, Claudio Di Ciccio, Eric Verbeek

Lakmali Jayarathna (Queensland University of Technology)

ICPM 2021, Eindhoven, 2 Nov 2021

Survey Setup

Data Transformation Challenges for Process Mining (12 Qs)

Four perspectives: Academia, Professional Services, Vendors, End Users

Released to the international process mining community: June – July 2021

Received **290** responses

THANK YOU!

Q1. How much experience do you have with Process Mining?

Q2. Which area and role best describe how you have interacted with PM?

Q3. What share of effort is typically spent on data pre-processing?

Q4. Which process mining solutions have you used?

Q5. Which technologies have you used in data preprocessing for process mining?

Q6. In Which formats is your source data available in?

Q7.Which source systems have you analyzed with process mining?

Q8. How big was the largest data set you worked with in process mining?

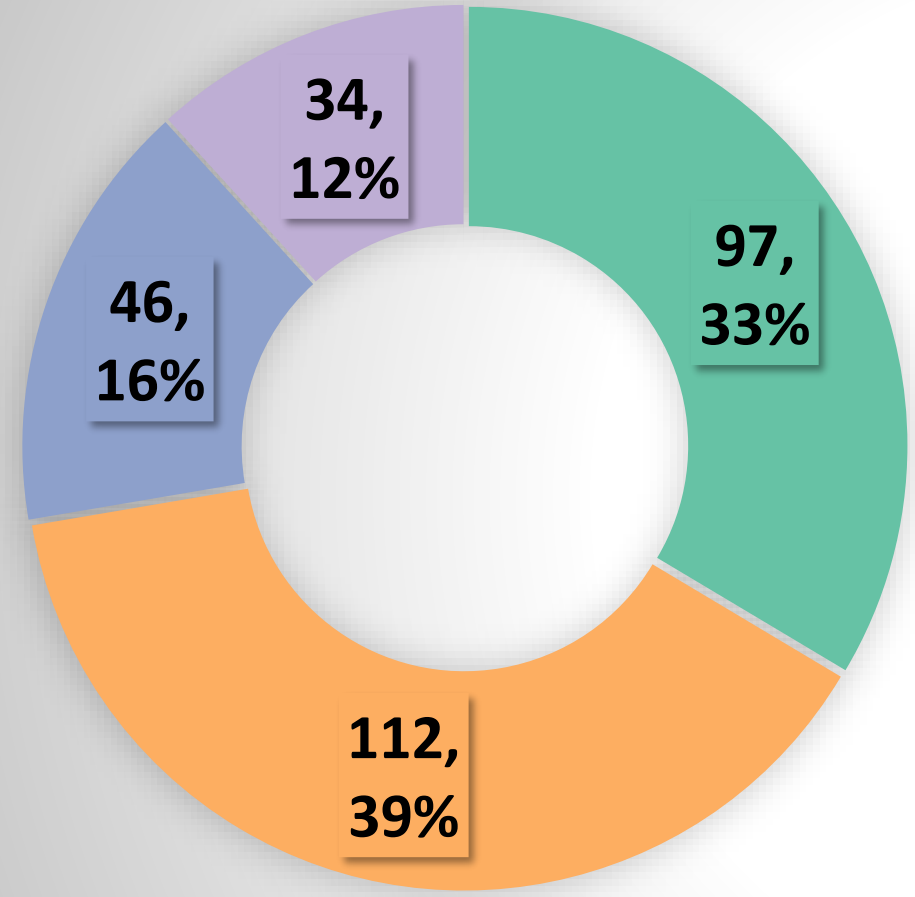
Q9. To what extent did you encounter the following data related challenges while undertaking PM projects?

Q10. Which data related challenges have you encountered beyond the ones listed in question #9?

Q11. There is general consensus amongst practitioners that data pre-processing tasks still consume most of the effort put into process mining initiatives. How could we speed up the data pre-processing to focus on analysis?

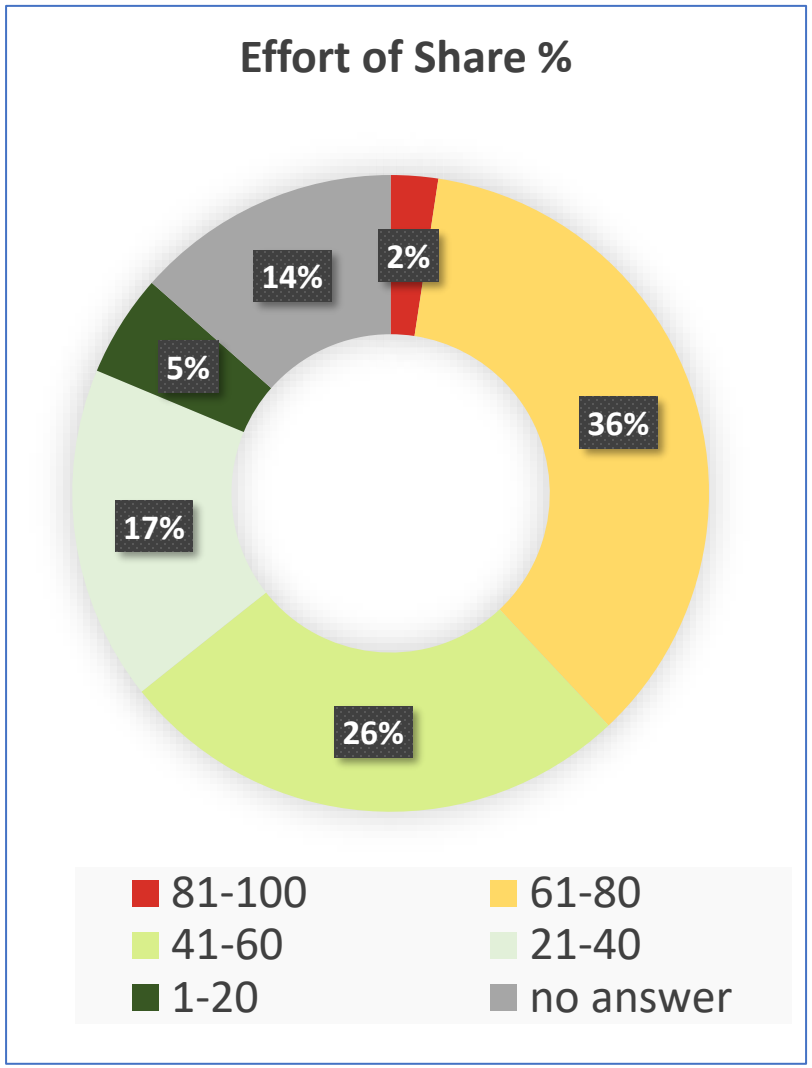
Q12. How could a reimagined industry-wide process mining data standard help you excel in your role?

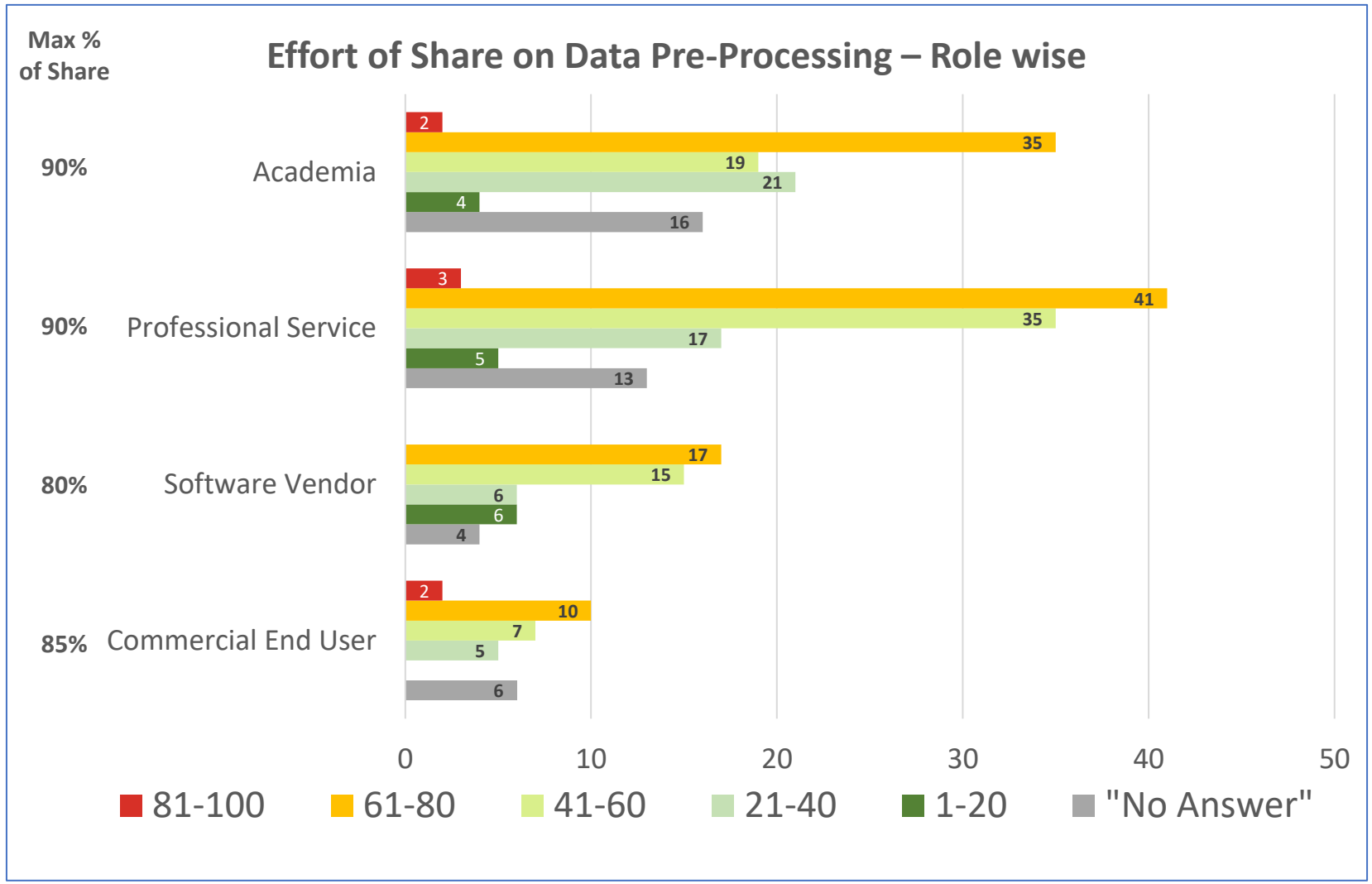
Survey Participation



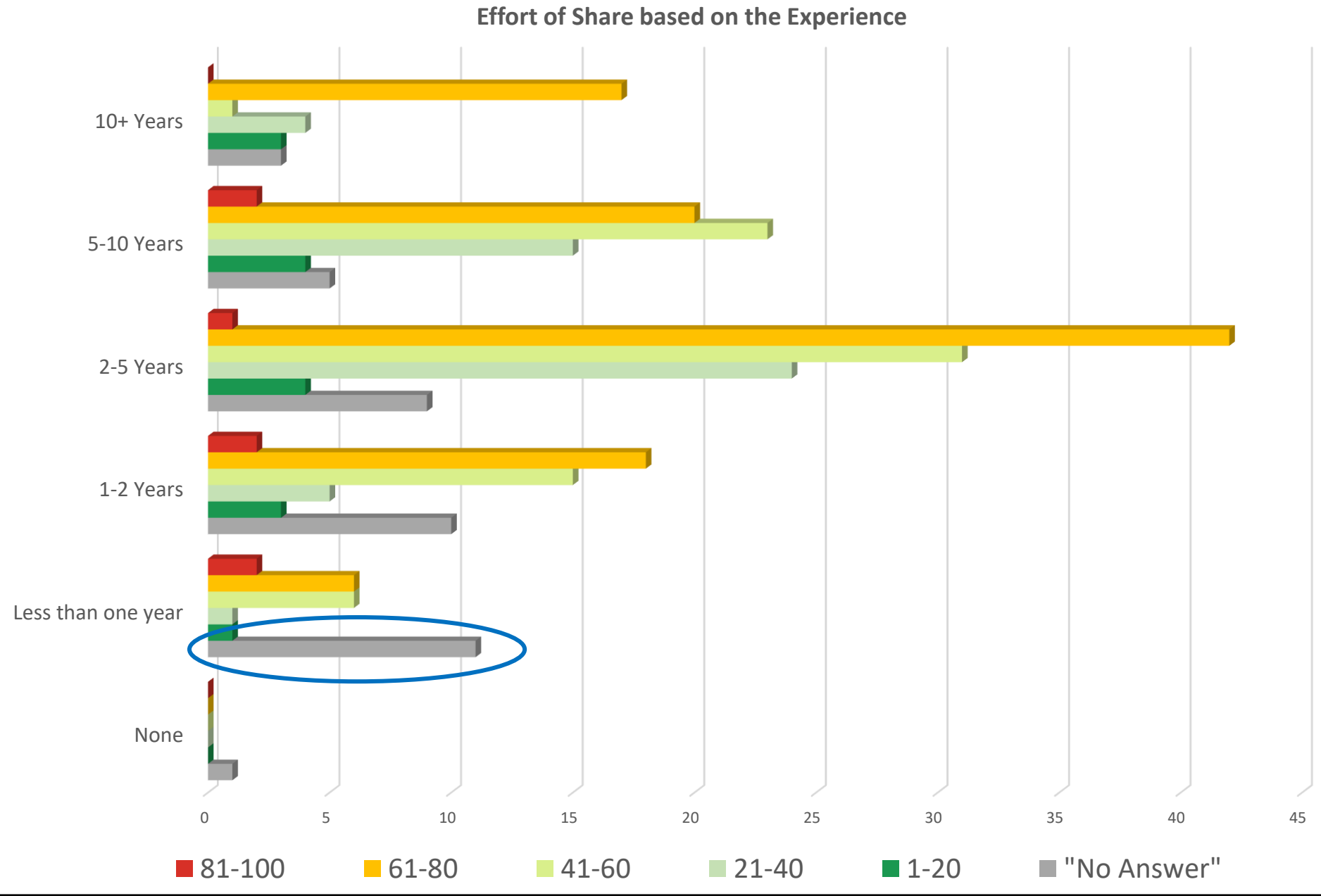
Total: 289 Survey Participants		
Academia	Researcher	59
	PhD Student	26
	MSC/BSc Student	12
Professional Service	Process Analyst	40
	Process Mining Engineer	37
	Project Manager	30
	Sales Representative	05
Software Vendor	Developer	13
	Process Analyst	07
	Process Mining Engineer	12
	Project Manager	08
	Sales Representative	06
Commercial End User	Process Analyst	16
	Process Mining Engineer	09
	Project Manager	07

Q3 – What share of effort is typically spent on data pre-processing? (i.e., all tasks between data extraction and analysis design)



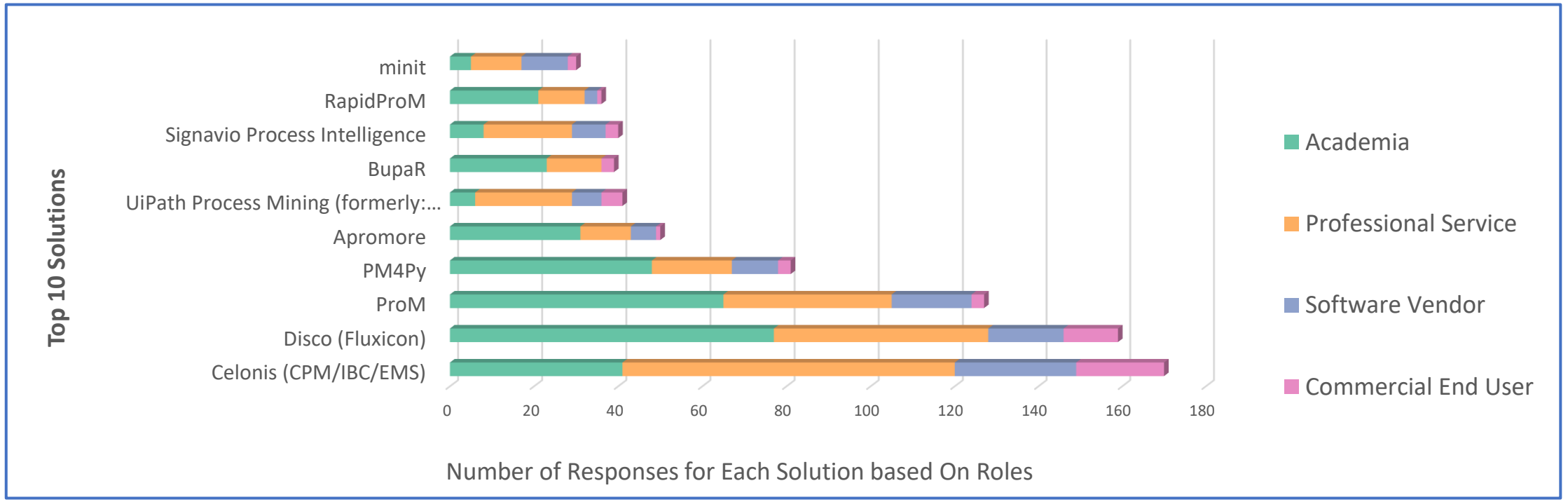


Q3 – What share of effort is typically spent on data pre-processing? (i.e., all tasks between data extraction and analysis design)



Q4 - Which process mining solutions have you used?

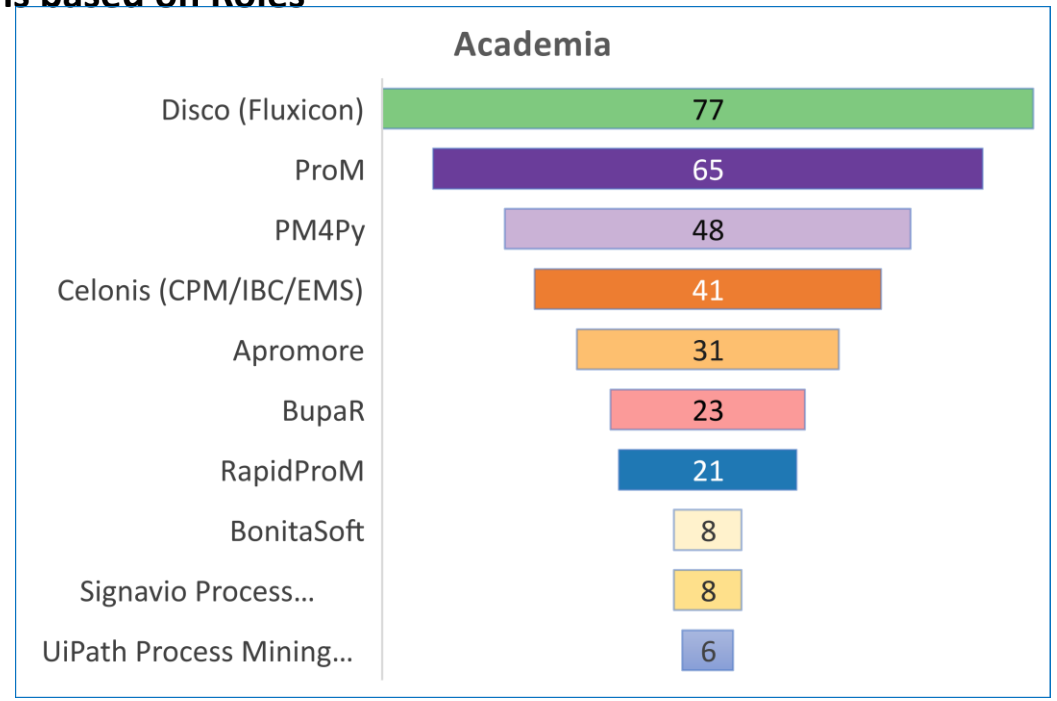
Top 10 Solutions

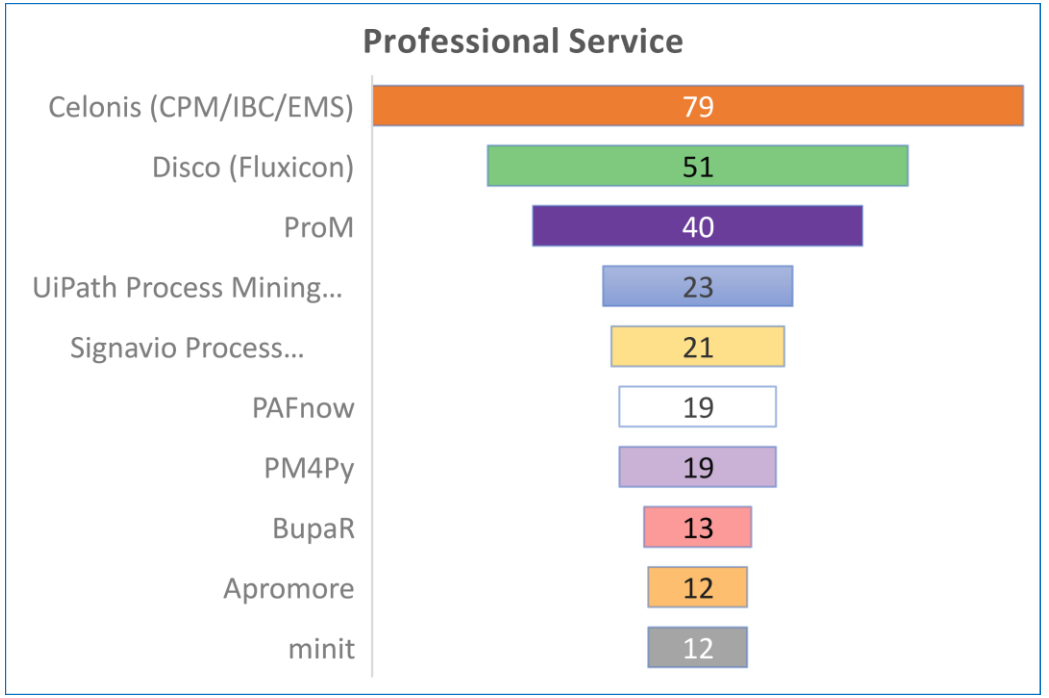


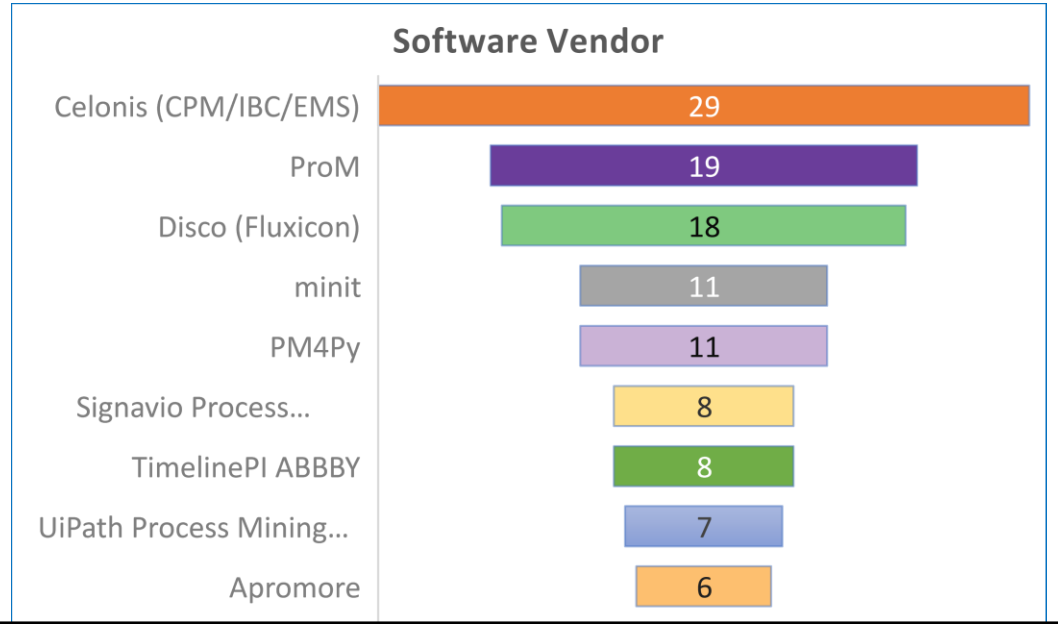
	Academia	Professional Service	Software Vendor	Commercial End User	Total
Celonis (CPM/IBC/EMS)	41	79	29	21	**170**
Disco (Fluxicon)	77	51	18	13	**159**
ProM	65	40	19	3	**127**
PM4Py	48	19	11	3	**81**
Apromore	31	12	6	1	**50**
UiPath Process Mining (formerly: ProcessGold)	6	23	7	5	**41**
BupaR	23	13	0	3	**39**
Signavio Process Intelligence	8	21	8	3	**40**
RapidProM	21	11	3	1	**36**
minit	5	12	11	2	**30**

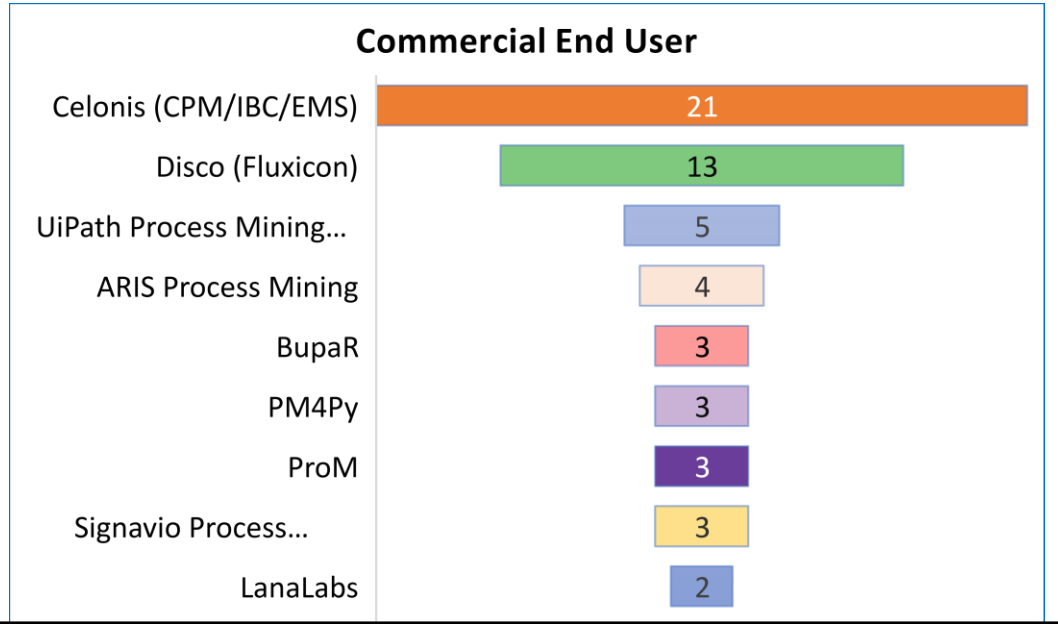
Q4 - Which process mining solutions have you used?

Top 10 solutions based on Roles

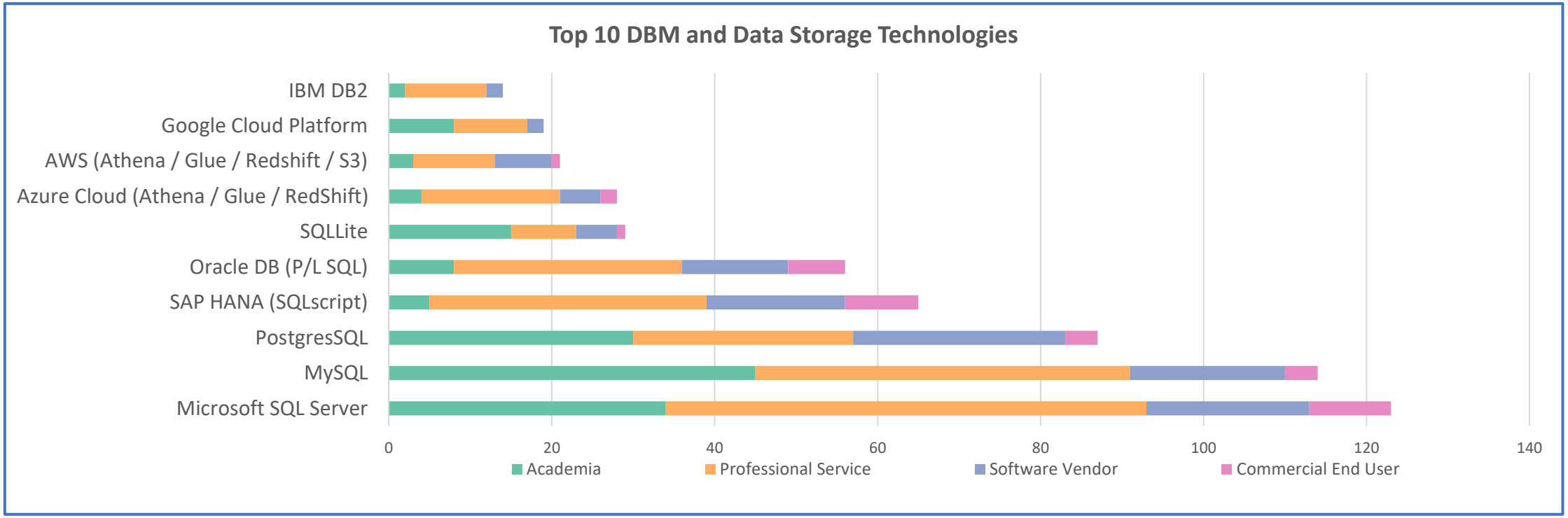






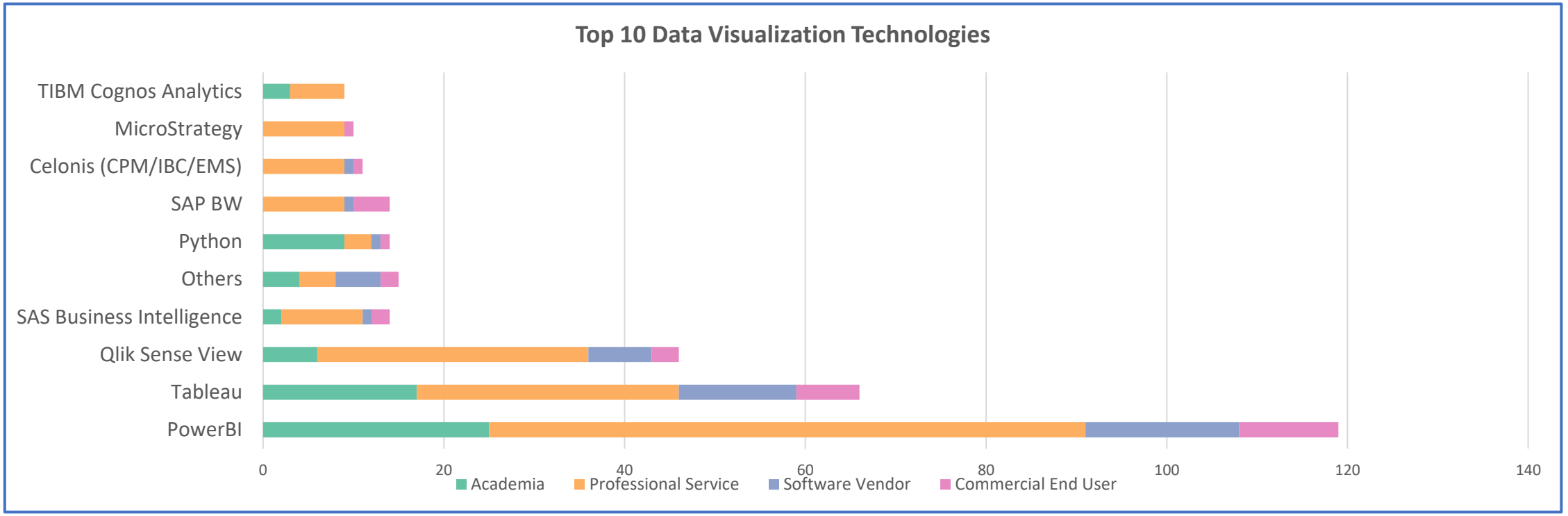


Q5 - Which technologies have you used in data prepossessing for process mining?



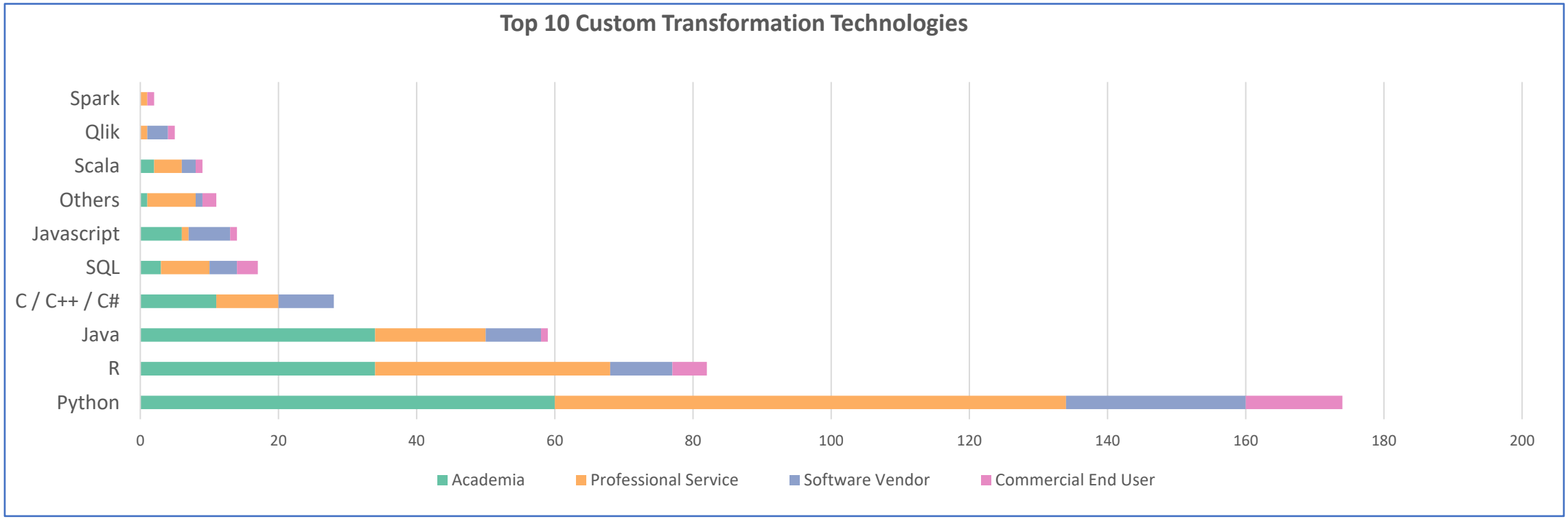
	Academia	Professional Service	Software Vendor	Commercial End User	Total
Microsoft SQL Server	34	59	20	10	**125**
MySQL	45	46	19	4	**116**
PostgresSQL	30	27	26	4	**89**
SAP HANA (SQLscript)	5	34	17	9	**68**
Oracle DB (P/L SQL)	8	28	13	7	**57**
SQLLite	15	8	5	1	**30**
Azure Cloud (Athena / Glue / RedShift)	4	17	5	2	**29**
AWS (Athena / Glue / Redshift / S3)	3	10	7	1	**21**
Google Cloud Platform	8	9	2	0	**19**
IBM DB2	2	10	2	0	**14**

Q5 - Which technologies have you used in data prepossessing for process mining?



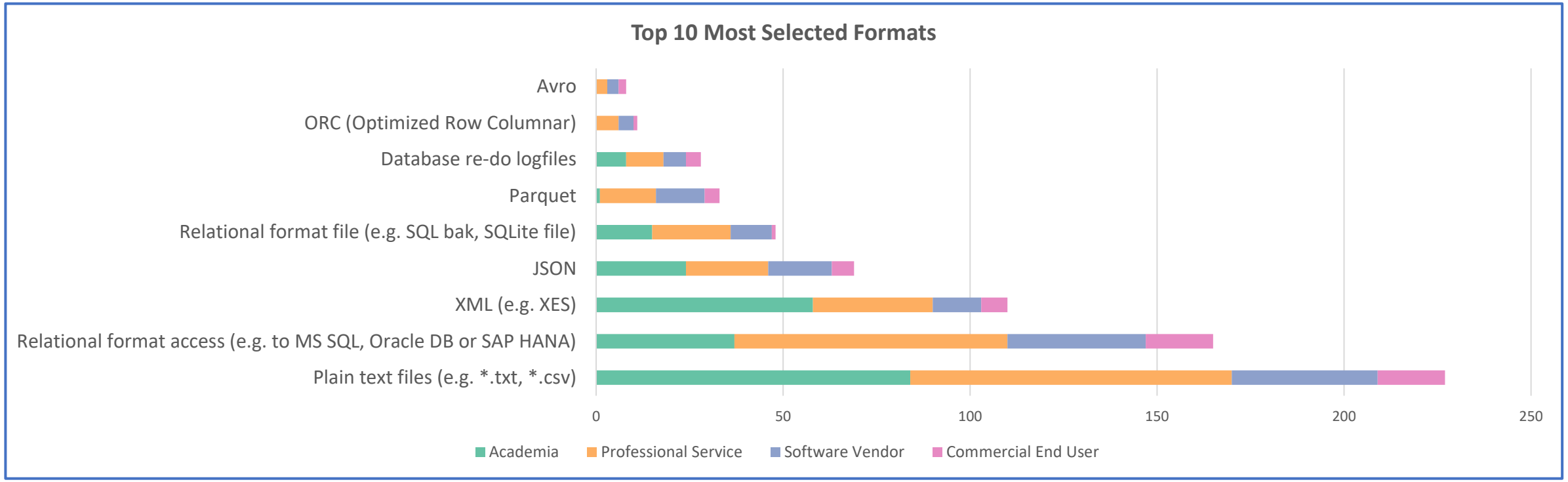
	Academia	Professional Service	Software Vendor	Commercial End User	Total
PowerBI	25	66	17	11	**122**
Tableau	17	29	13	7	**68**
Qlik Sense View	6	30	7	3	**47**
SAS Business Intelligence	2	9	1	2	**15**
Others	4	4	5	2	**15**
Python	9	3	1	1	**14**
SAP BW	0	9	1	4	**14**
Celonis (CPM/IBC/EMS)	0	9	1	1	**11**
MicroStrategy	0	9	0	1	**10**
TIBM Cognos Analytics	3	6	0	0	**9**

Q5 - Which technologies have you used in data prepossessing for process mining?



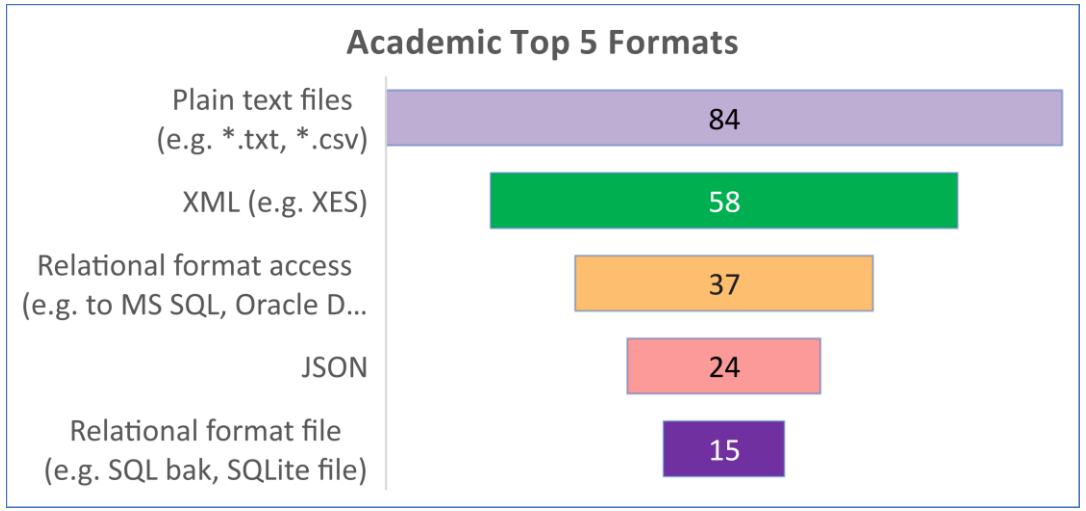
	Academia	Professional Service	Software Vendor	Commercial End User	Total
Python	60	74	26	14	**177**
R	34	34	9	5	**83**
Java	34	16	8	1	**60**
C / C++ / C#	11	9	8	0	**28**
SQL	3	7	4	3	**17**
Javascript	6	1	6	1	**14**
Others	1	7	1	2	**11**
Scala	2	4	2	1	**9**
Qlik	0	1	3	1	**5**
Spark	0	1	0	1	**2**

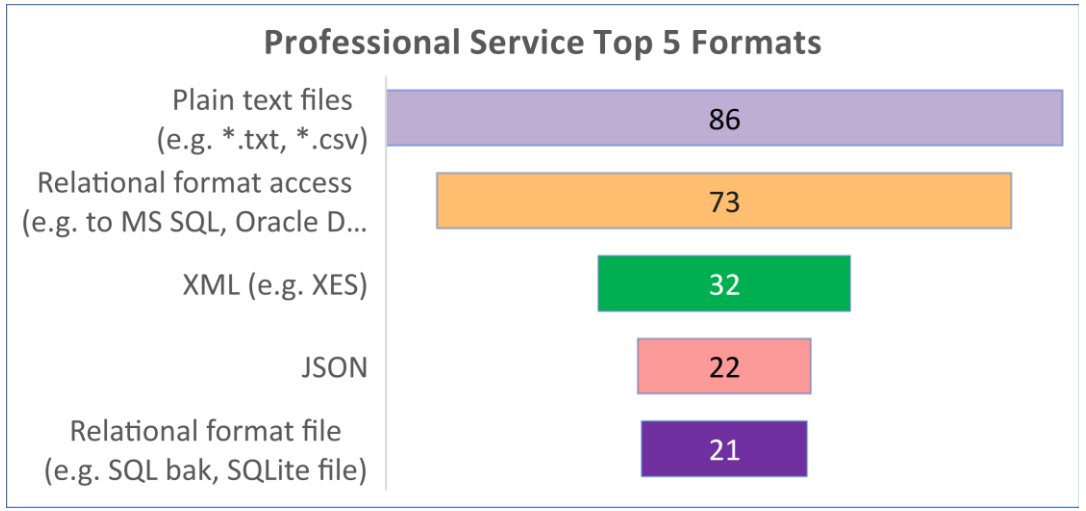
Q6 - In which formats is your source data available? – Top 10 responses

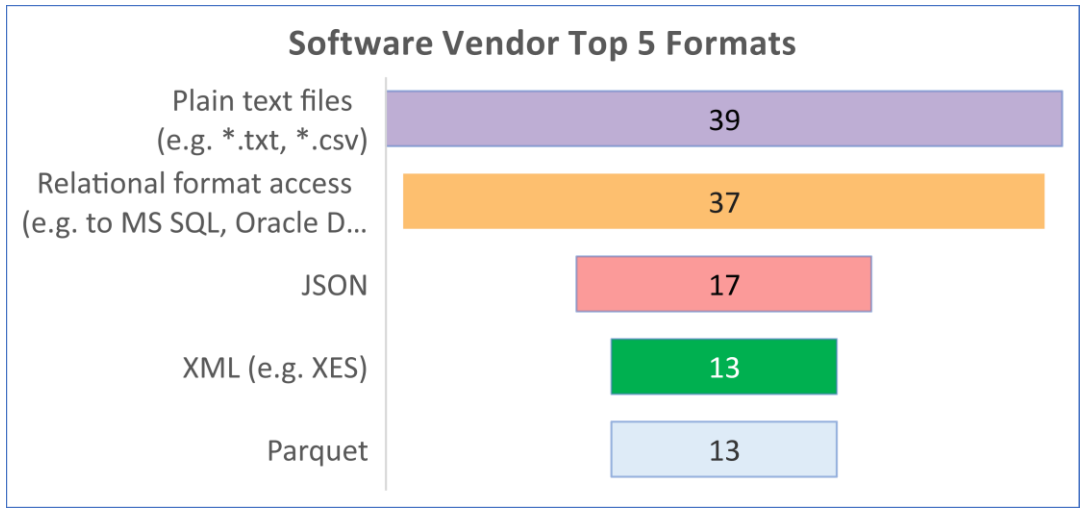


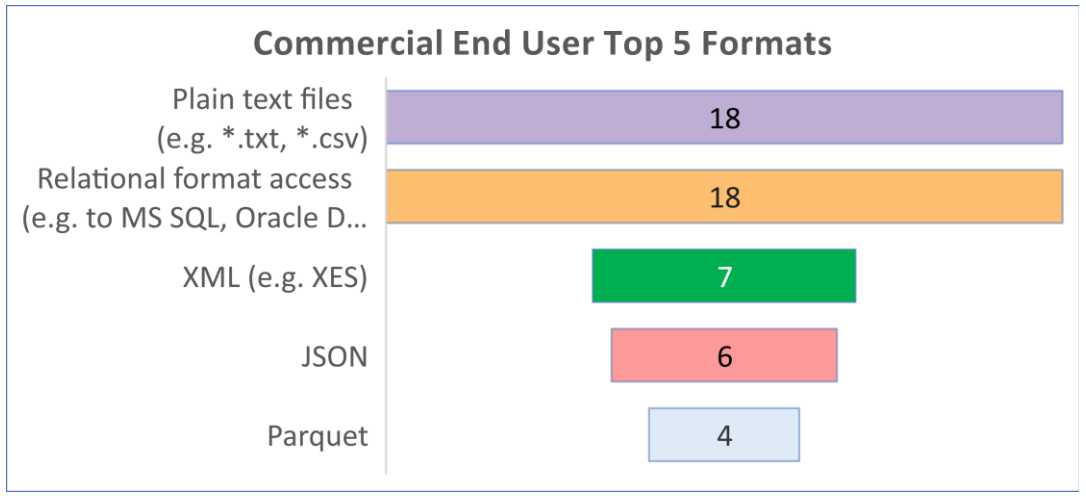
	Academia	Professional Service	Software Vendor	Commercial End User	Total
Plain text files (e.g. *.txt, *.csv)	84	86	39	18	**229**
Relational format access (e.g., to MS SQL, Oracle DB or SAP HANA)	37	73	37	18	**168**
XML (e.g. XES)	58	32	13	7	**112**
JSON	24	22	17	6	**70**
Relational format file (e.g. SQL bak, SQLite file)	15	21	11	1	**48**
Parquet	1	15	13	4	**34**
Database re-do logfiles	8	10	6	4	**29**
ORC (Optimized Row Columnar)	0	6	4	1	**11**
Avro	0	3	3	2	**9**

Q6 - In which formats is your source data available? (check all that apply).

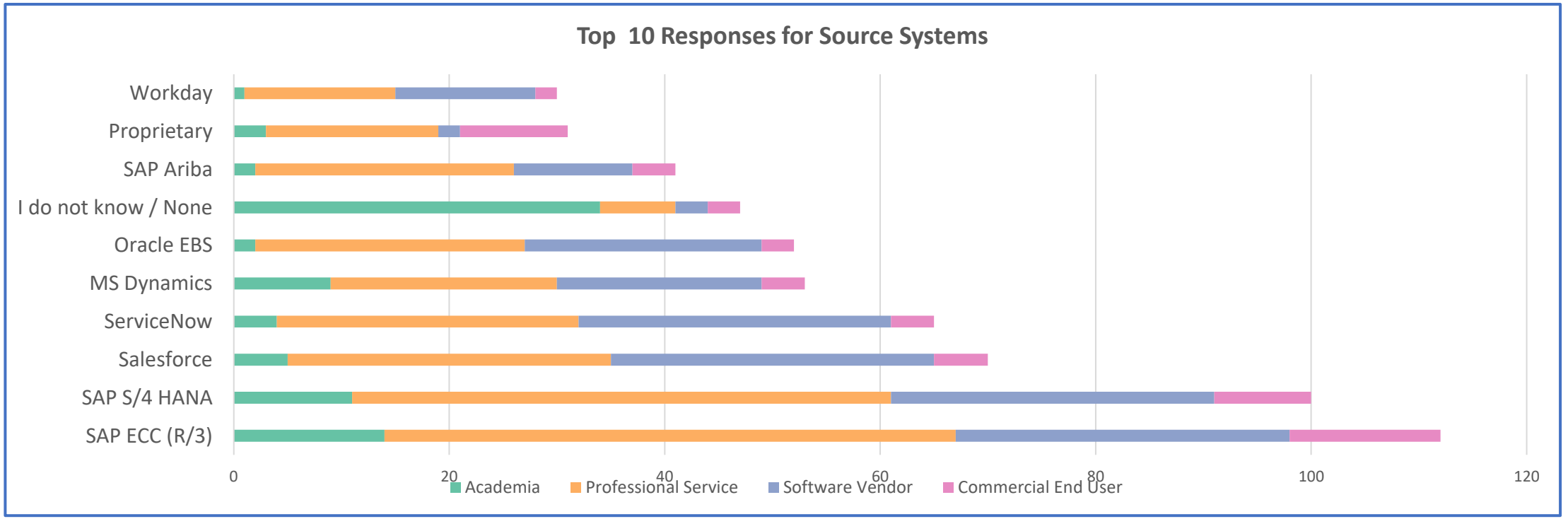








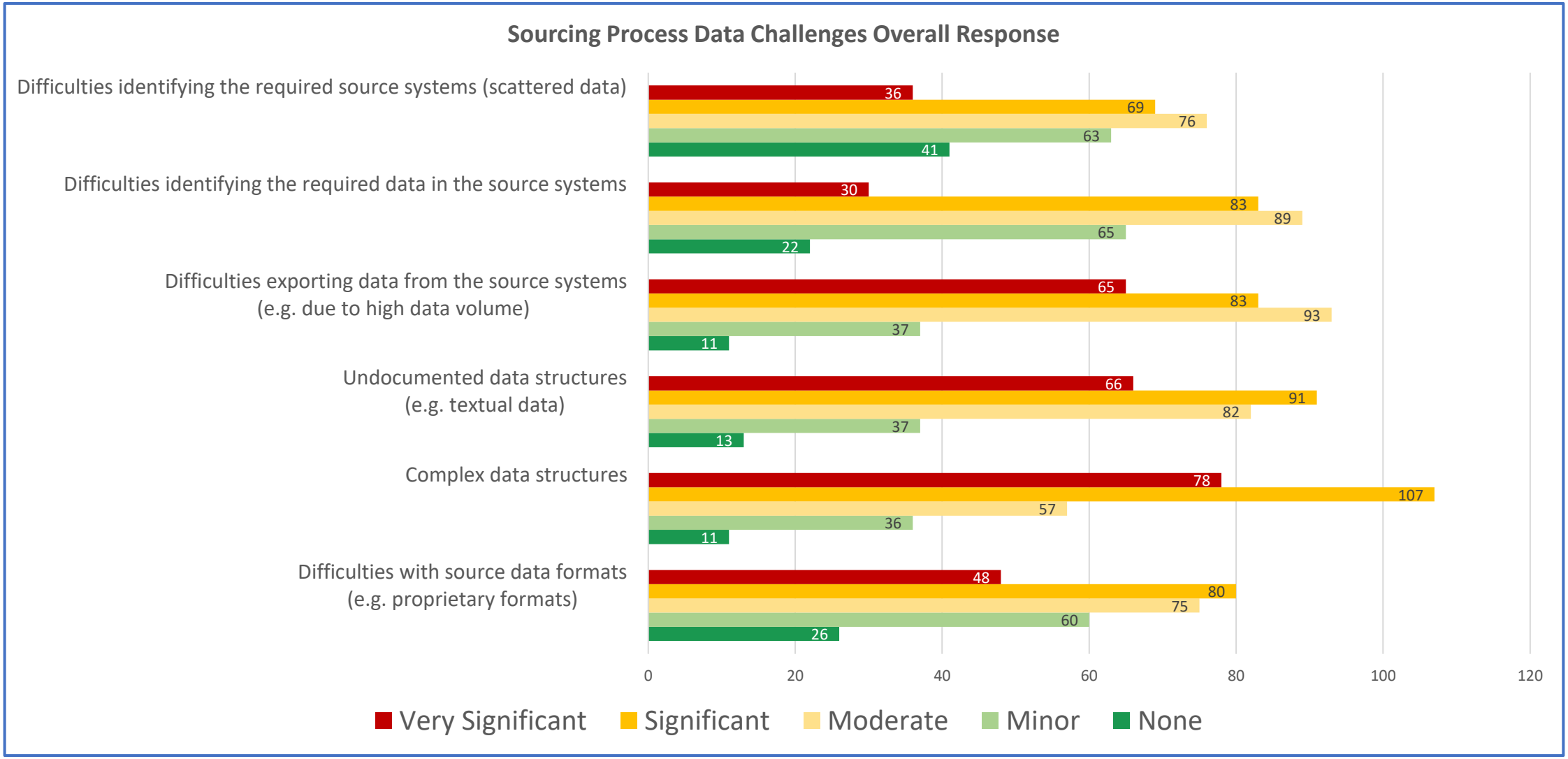
Q8 - Which source systems have you analyzed with process mining?



	Academia	Professional Service	Software Vendor	Commercial End User	Total
SAP ECC (R/3)	14	53	31	14	**114**
SAP S/4 HANA	11	50	30	9	**101**
Salesforce	5	30	30	5	**71**
ServiceNow	4	28	29	4	**66**
MS Dynamics	9	21	19	4	**54**
Oracle EBS	2	25	22	3	**53**
I do not know / None	34	7	3	3	**48**
SAP Ariba	2	24	11	4	**42**
Proprietary	3	16	2	10	**31**
Workday	1	14	13	2	**30**

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

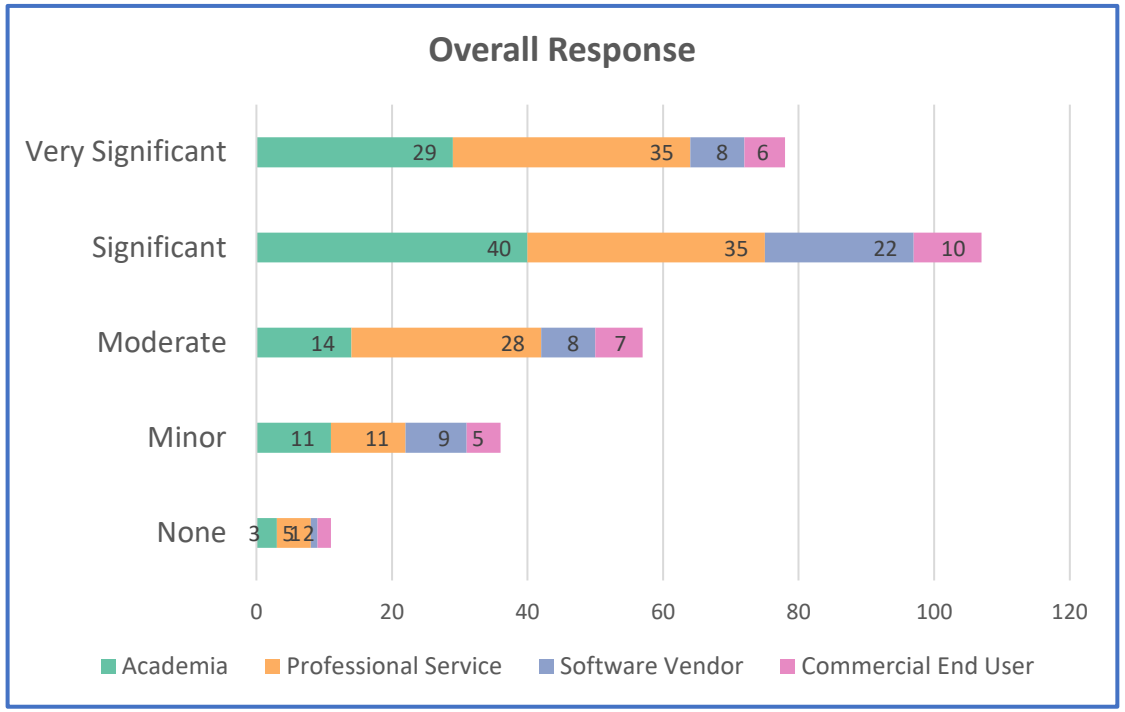
Sourcing Process Data Overall

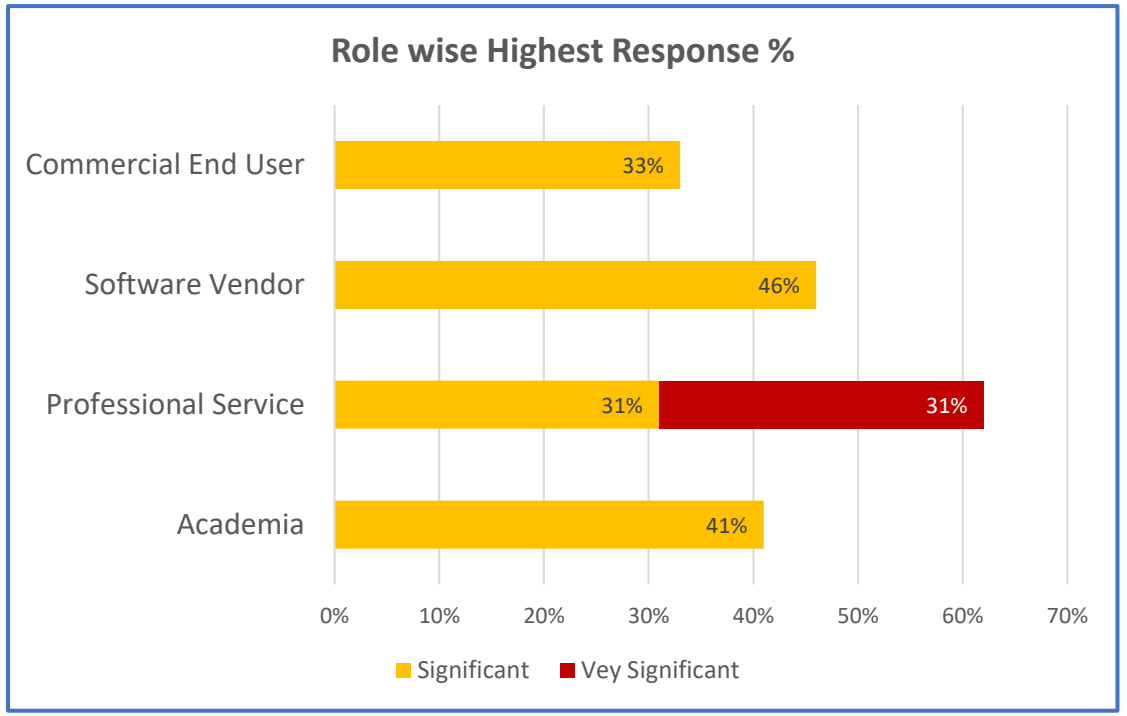


Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

Sourcing Process Data

5. Complex data structures

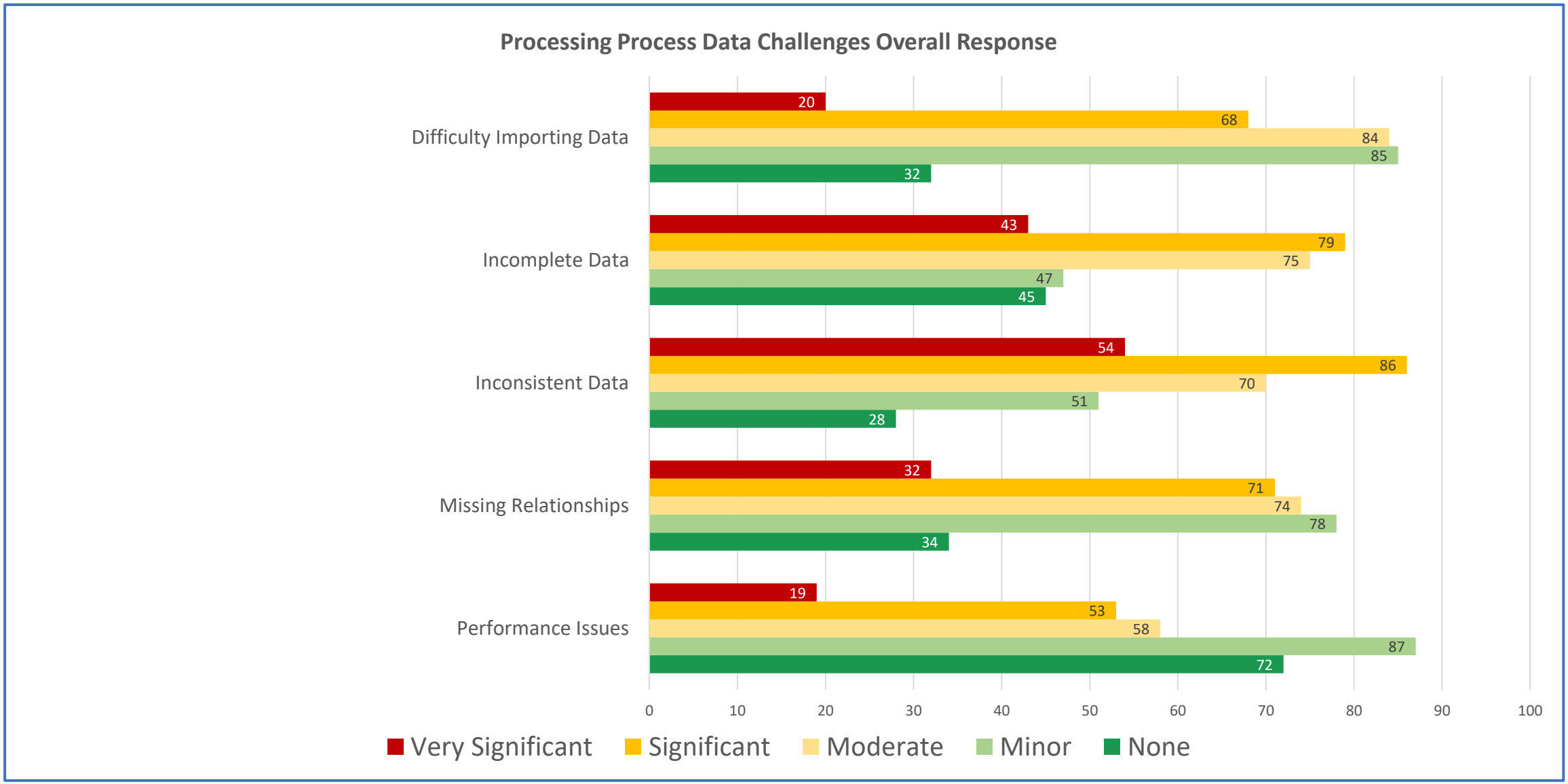




	None	Minor	Moderate	Significant	Very Significant
Academia	3 (3%)	11 (11%)	14 (14%)	40 (41%)	29 (30%)
Professional Service	5 (4%)	11 (10%)	28 (25%)	35 (31%)	35 (31%)
Software Vendor	1 (2%)	9 (19%)	8 (17%)	22 (46%)	8 (17%)
Commercial End User	2 (7%)	5 (17%)	7 (23%)	10 (33%)	6 (20%)
Total	**11**	**36**	**57**	**107**	**78**

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

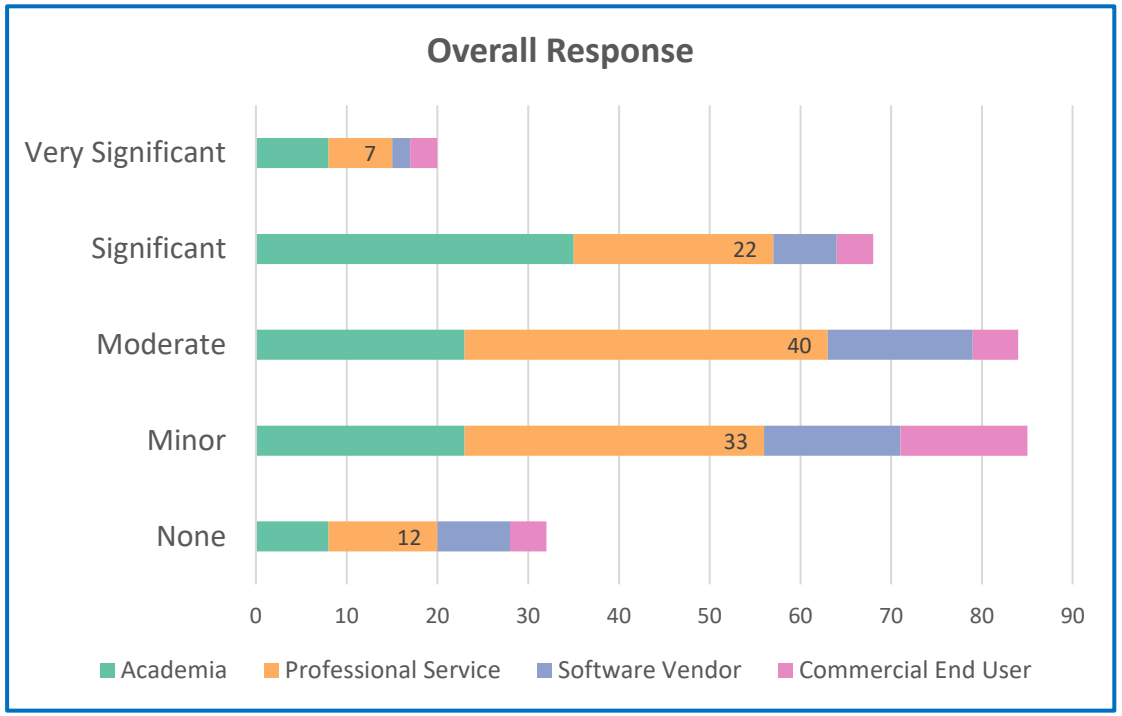
Processing Process Data Overall

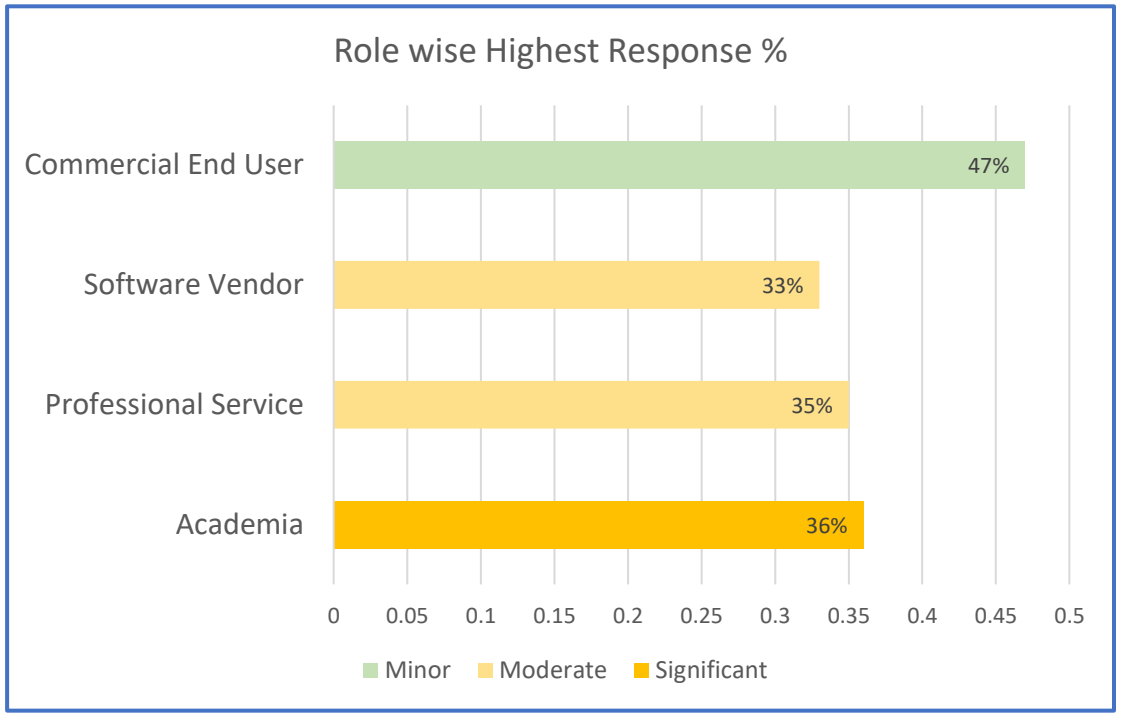


Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

Processing Process Data

1. Difficulties importing data into the data processing environment



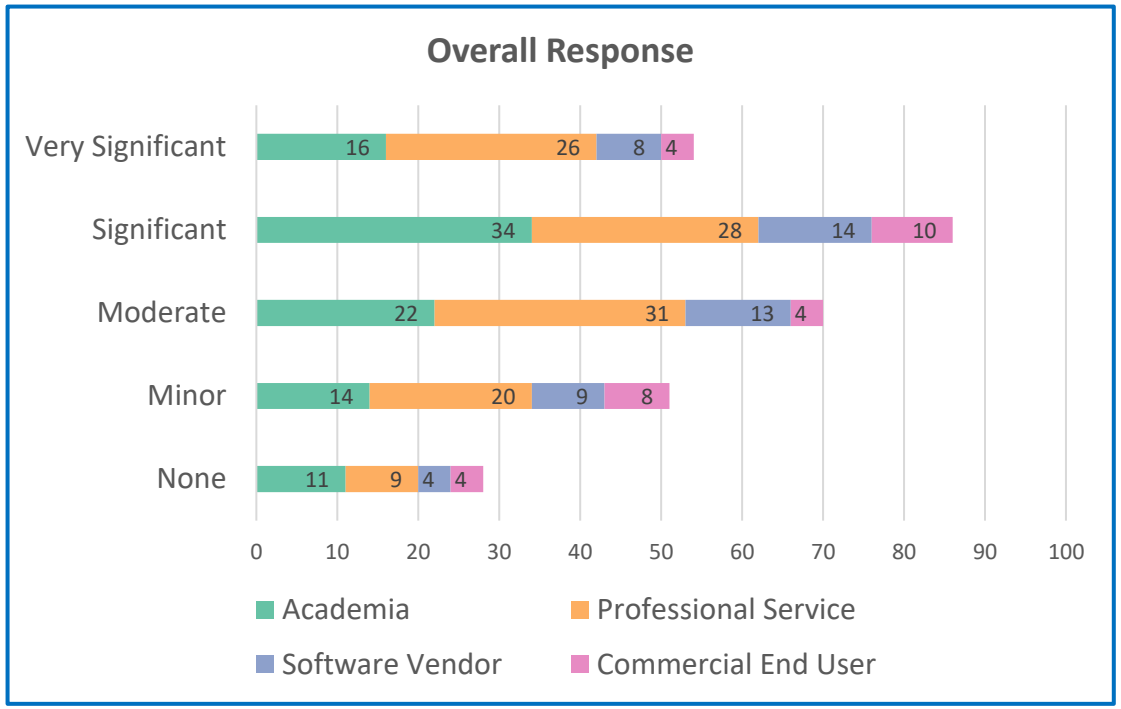


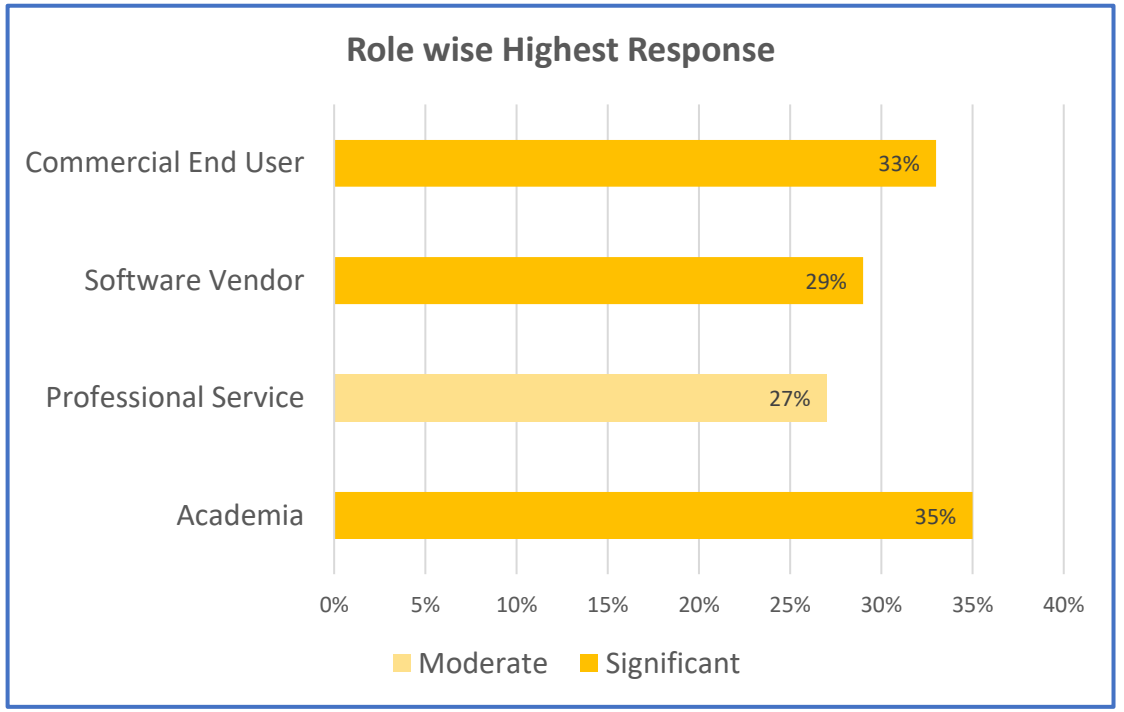
	None	Minor	Moderate	Significant	Very Significant
Academia	8 (8%)	23 (24%)	23 (24%)	35 (36%)	8 (8%)
Professional Service	12 (11%)	33 (29%)	40 (35%)	22 (19%)	7 (6%)
Software Vendor	8 (17%)	15 (31%)	16 (33%)	7 (15%)	2 (4%)
Commercial End User	4 (13%)	14 (47%)	5 (17%)	4 (13%)	3 (10%)
Total	**32**	**85**	**84**	**68**	**20**

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

Processing Process Data

3. Inconsistent Source Data



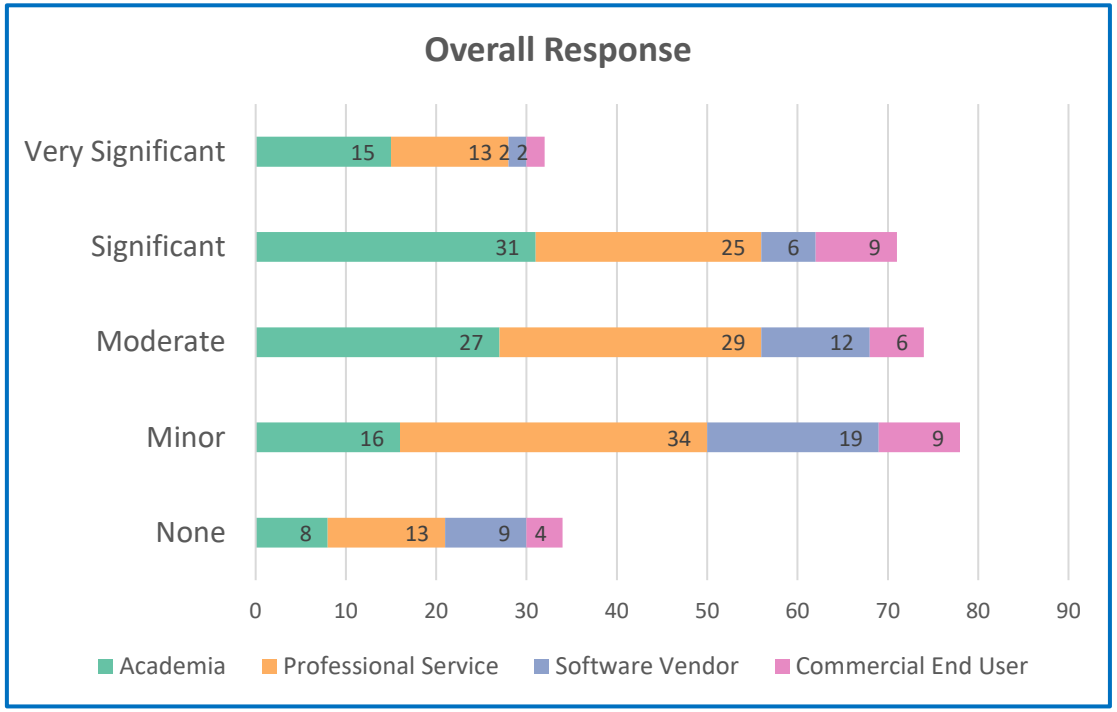


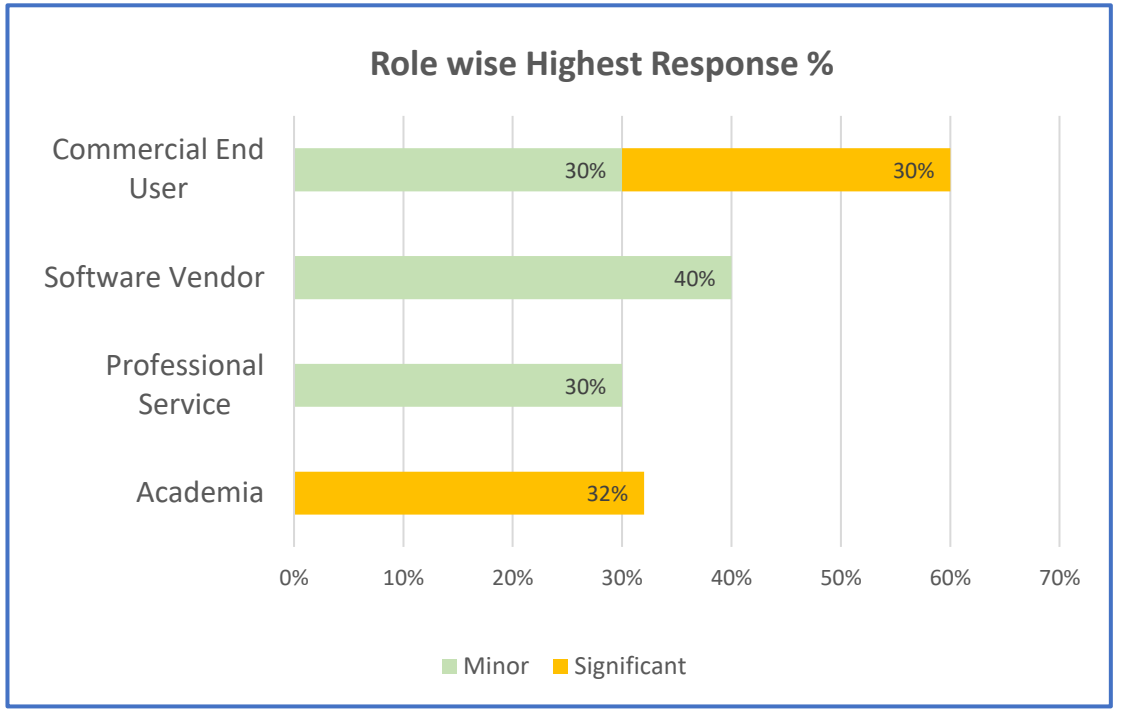
	None	Minor	Moderate	Significant	Very Significant
Academia	11 (11%)	14 (14%)	22 (23%)	34 (35%)	16 (16%)
Professional Service	9 (8%)	20 (18%)	31 (27%)	28 (25%)	26 (23%)
Software Vendor	4 (8%)	9 (19%)	13 (27%)	14 (29%)	8 (17%)
Commercial End User	4 (13%)	8 (27%)	4 (13%)	10 (33%)	4 (13%)
Total	**28**	**51**	**70**	**86**	**54**

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

Processing Process Data

4. Missing information about relationships in the data

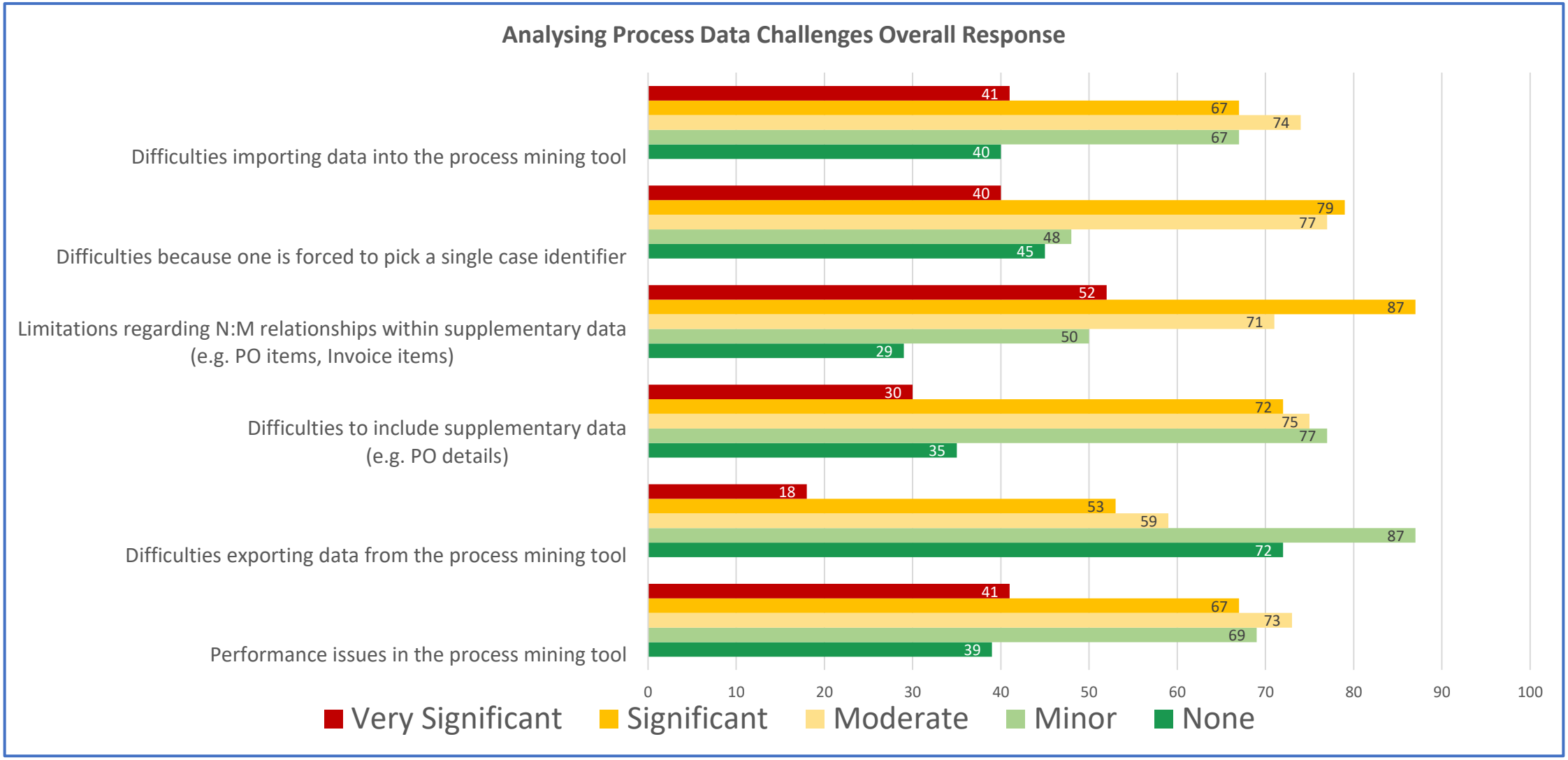




	None	Minor	Moderate	Significant	Very Significant
Academia	8 (8%)	16 (16%)	27 (28%)	31 (32%)	15 (15%)
Professional Service	13 (11%)	34 (30%)	29 (25%)	25 (22%)	13 (11%)
Software Vendor	9 (19%)	19 (40%)	12 (25%)	6 (13%)	2 (4%)
Commercial End User	4 (13%)	9 (30%)	6 (20%)	9 (30%)	2 (7%)
Total	**34**	**78**	**74**	**71**	**32**

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

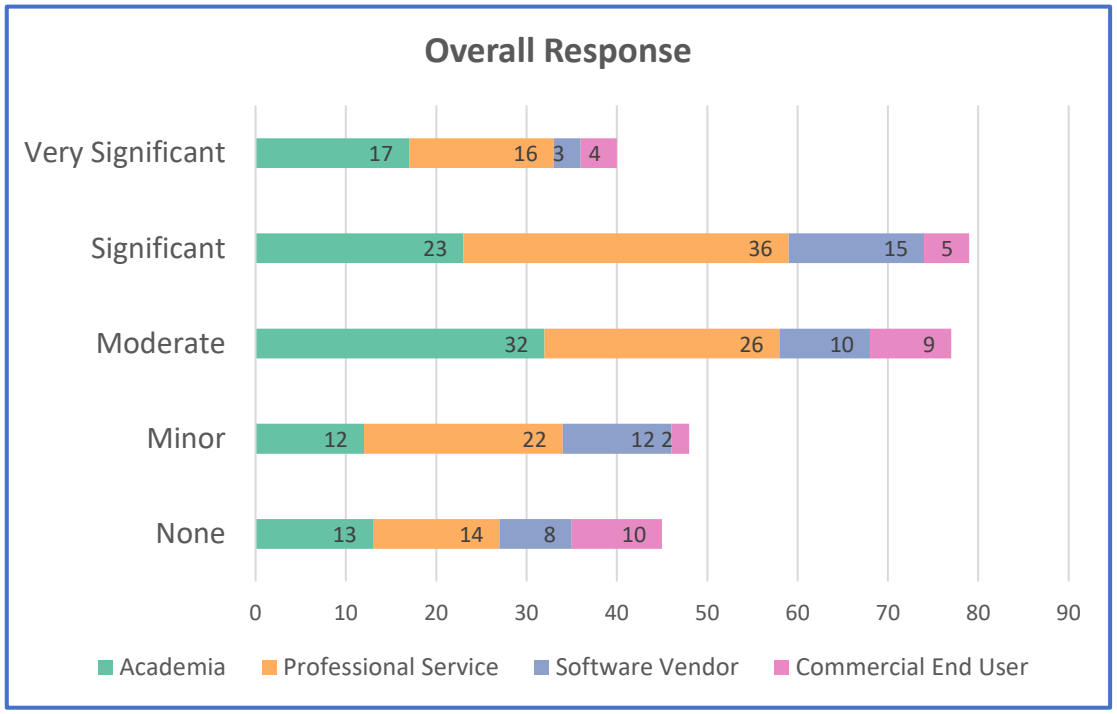
Analysing Process Data Overall

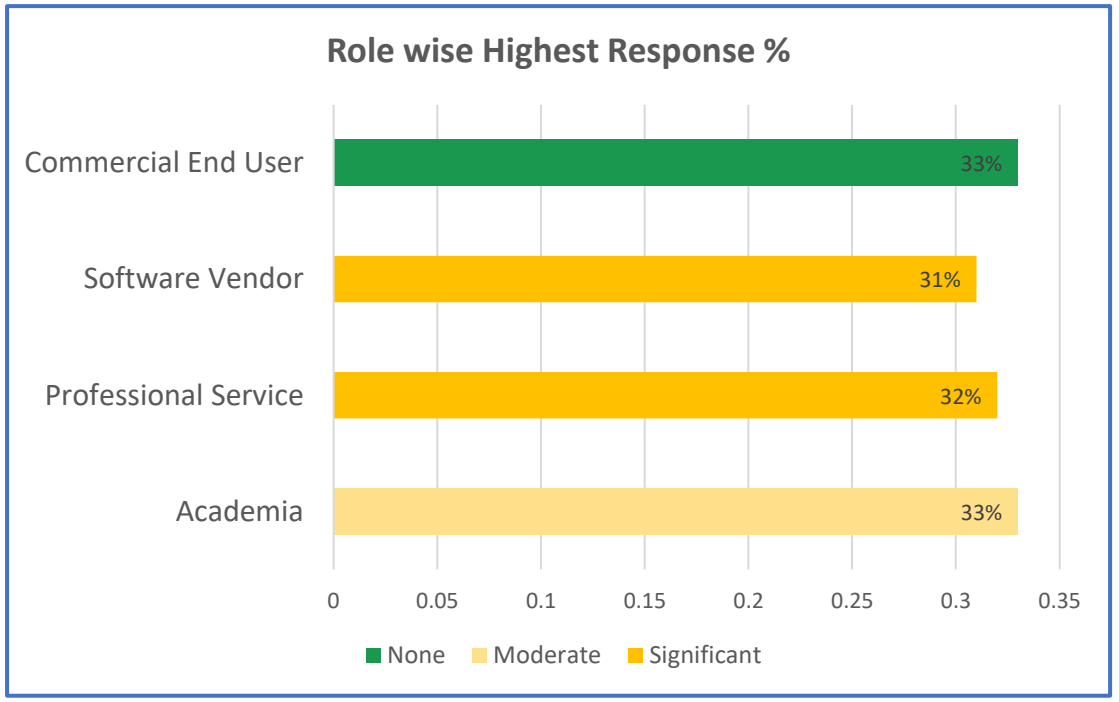


Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

Analyzing Process Data

2. Difficulties because one is forced to pick a single case identifier



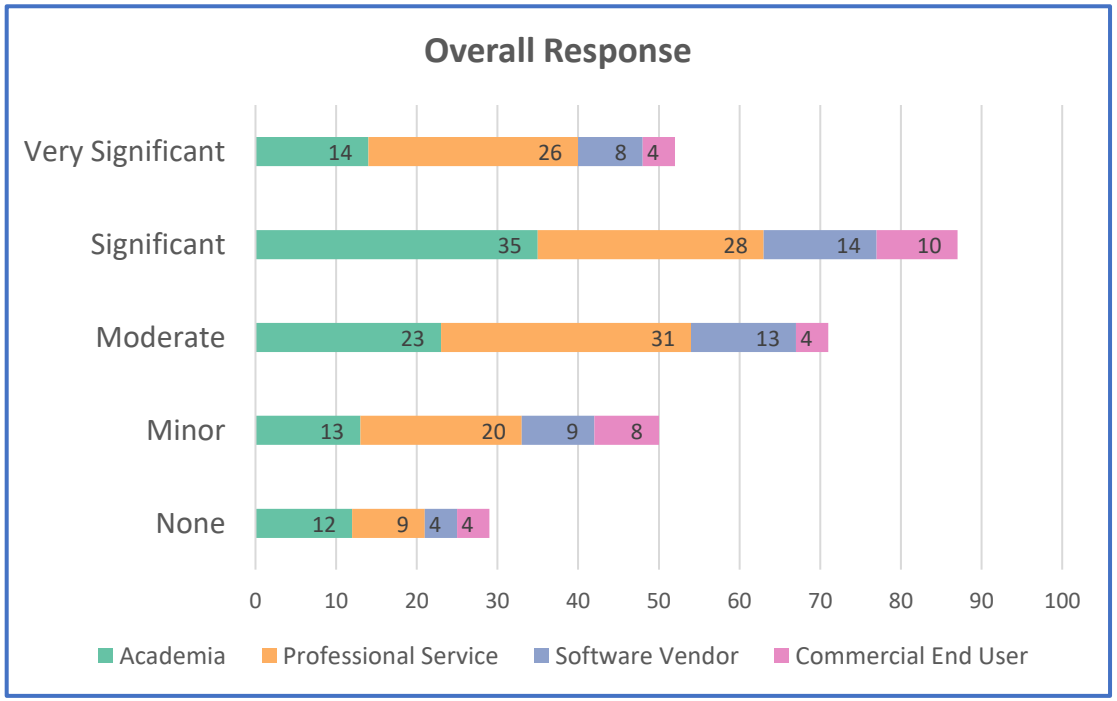


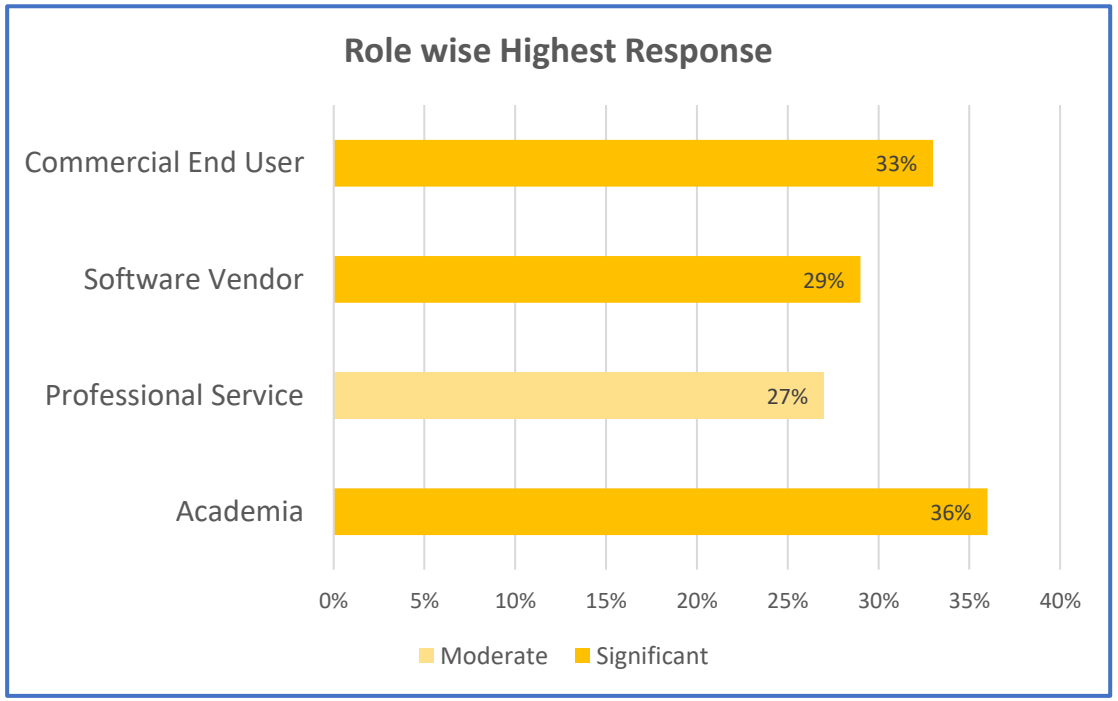
	None	Minor	Moderate	Significant	Very Significant
Academia	13 (13%)	12 (12%)	32 (33%)	23 (24%)	17 (18%)
Professional Service	14 (12%)	22 (19%)	26 (23%)	36 (32%)	16 (14%)
Software Vendor	8 (17%)	12 (25%)	10 (21%)	15 (31%)	3 (6%)
Commercial End User	10 (33%)	2 (7%)	9 (30%)	5 (17%)	4 (13%)
Total	**45**	**48**	**77**	**79**	**40**

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?

Analyzing Process Data

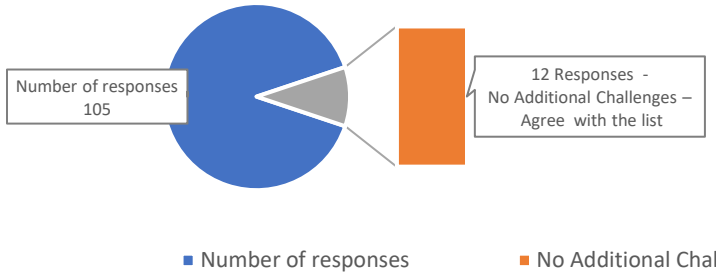
3. Limitations Regarding N:M Relationships within Supplementary Data (E.G. PO Items, Invoice Items)





	None	Minor	Moderate	Significant	Very Significant
Academia	12 (12%)	13 (13%)	23 (24%)	35 (36%)	14 (14%)
Professional Service	9 (8%)	20 (18%)	31 (27%)	28 (25%)	26 (23%)
Software Vendor	4 (8%)	9 (19%)	13 (27%)	14 (29%)	8 (17%)
Commercial End User	4 (13%)	8 (27%)	4 (13%)	10 (33%)	4 (13%)
Total	**29**	**50**	**71**	**87**	**52**

Q10 - Which data related challenges have you encountered beyond the ones listed inquestion #9?



Identified critical challenges out of the provided list under Q#9

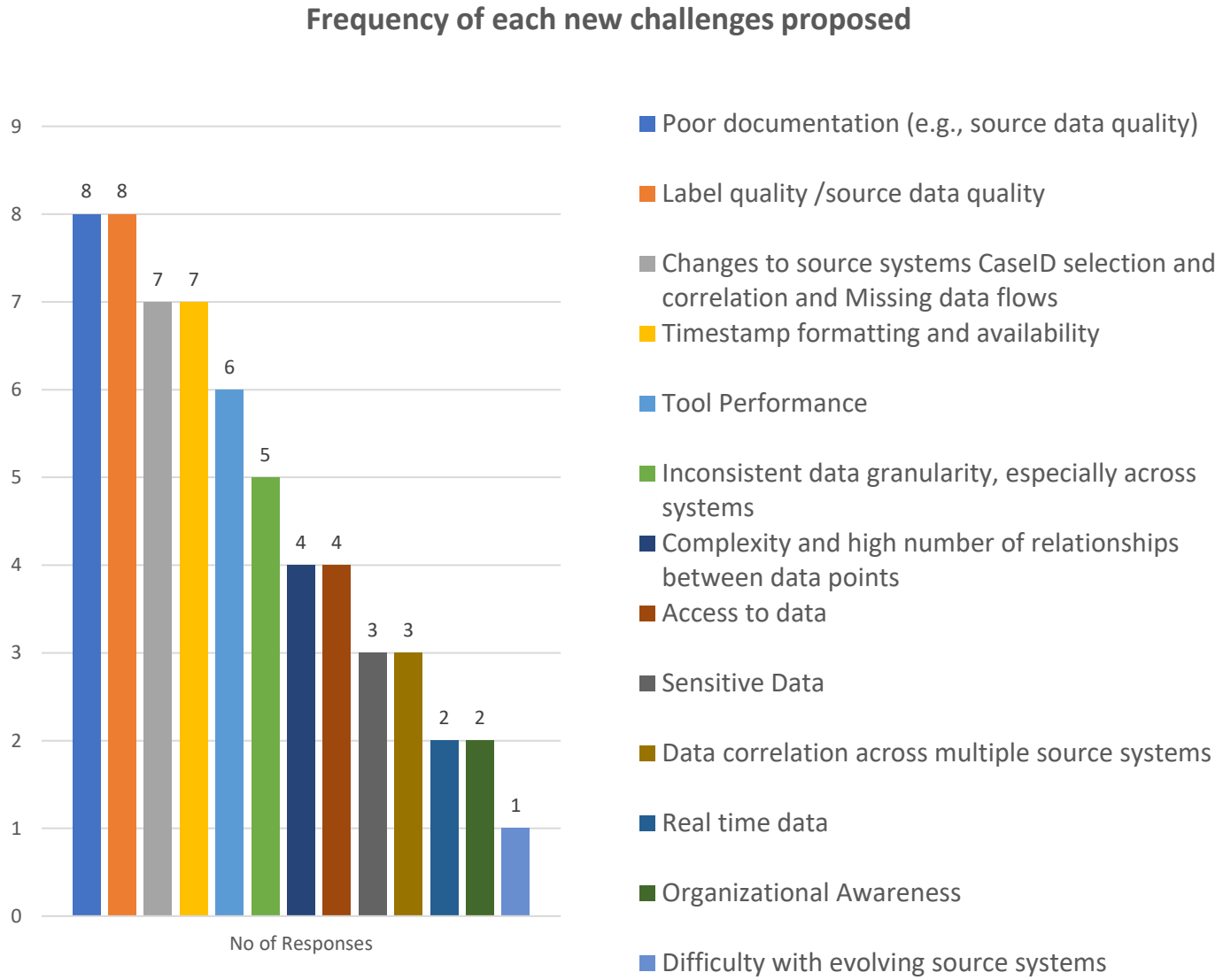
"I found that many of the above challenges are very relevant for the extraction task. The most critical according to my experience:
Difficulties identifying the required source systems (scattered data)
Difficulties identifying the required data in the source systems
Missing information about relationships in the data
Difficulties because one is forced to pick a single case identifier"

(Participant 159)

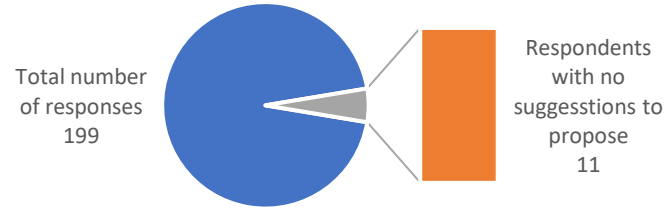
"Performance of tools" has been proposed as a critical challenge

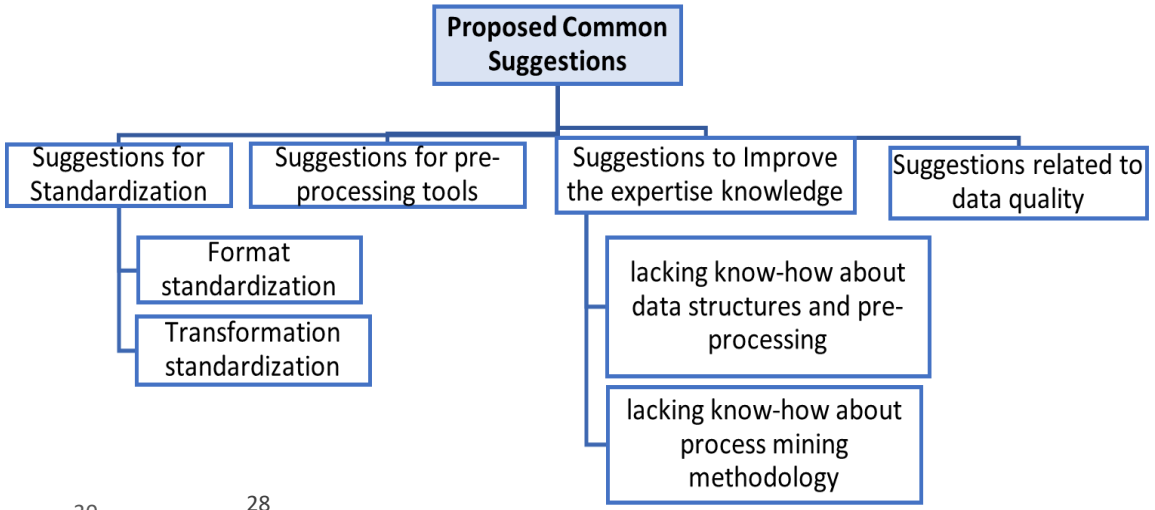
"A major challenge I've encountered regards the performance of tools whilst processing large portions of data. This is both true for "industrial" tools, such as Celonis (on-prem and IBC) and simpler tools such as Disco."

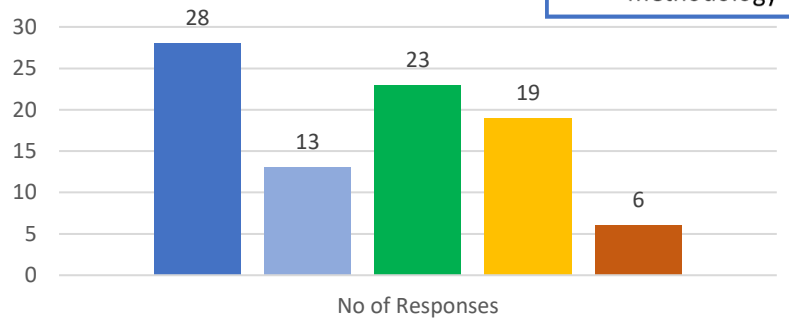
(Participant 33)



Q11 - How could we speed up the data pre-processing to focus on analysis?







Theme (Proposed suggestion)		Example Quote
Suggestions for Standardization *To standardize the data format both in source systems and mining tools. Based on that, to develop standardized ETL tools easy to use*. **(Participant 131)**	Format standardization	*"Standardized formats and public export adapter for common source systems"* **(Participant 1)** *"Standardising a data catalogue for those systems would most speed up the data pre-processing."* **(Participant 11)**
	Transformation standardization	*"Accelerators for Standard Source systems transformation, Standardised Applications will help in focusing on Analysis"* **(Participant 85)** *"Pre-processing is available for standard systems. Otherwise, availability for a CDM (common data model) so once information is mapped, the event log can be calculated"* **(Participant 130)**
Suggestions for pre-processing tools		*"Develop proper tooling to do common tasks; I have a framework to do that, but it's nowhere near user-friendly; however there's little research value in developing that."* **(Participant 31)** Build dedicated data-preprocessing tools (independent of process mining tool) **(participant 142)** A Graphic interface which address almost all the preprocessing issue and can be automated **(participant 200)**
Suggestions to Improve the expertise knowledge	lacking know-how about data structures and pre-processing	*Data pre-processing goes hand-in-hand with organisational context. That is, it is important to understand the domain, the organisation, external forces acting on process participants, system constraints on the way process activities are scheduled, carried out, recorded, etc. Such understanding allows the analyst to explain how/why every row of data in the source logs arose. This understanding also guides data pre-processing and informs transformation of source data to the final event log(s). This activity does take time, but I have found it time well spent.* **(Participant 61)**
	lacking know-how about process mining methodology	*Process focused mindset to be followed, to start understand the process, performance levels and then focus on what data is required to discover process boundaries, variants and performance* **(Participant 70)**
Suggestions related to data quality		*"Statistics to identify quickly the most common quality issues with data." (***Participant 166)**

- **Q12 - How could a reimagined industry-wide process mining data standard help you excel in your role?**

•Efficiency gains / Increased Impact

•accelerate pre-processing

•simplifying data exchange

•Ease of extraction

•Enhance the awareness Easy availability

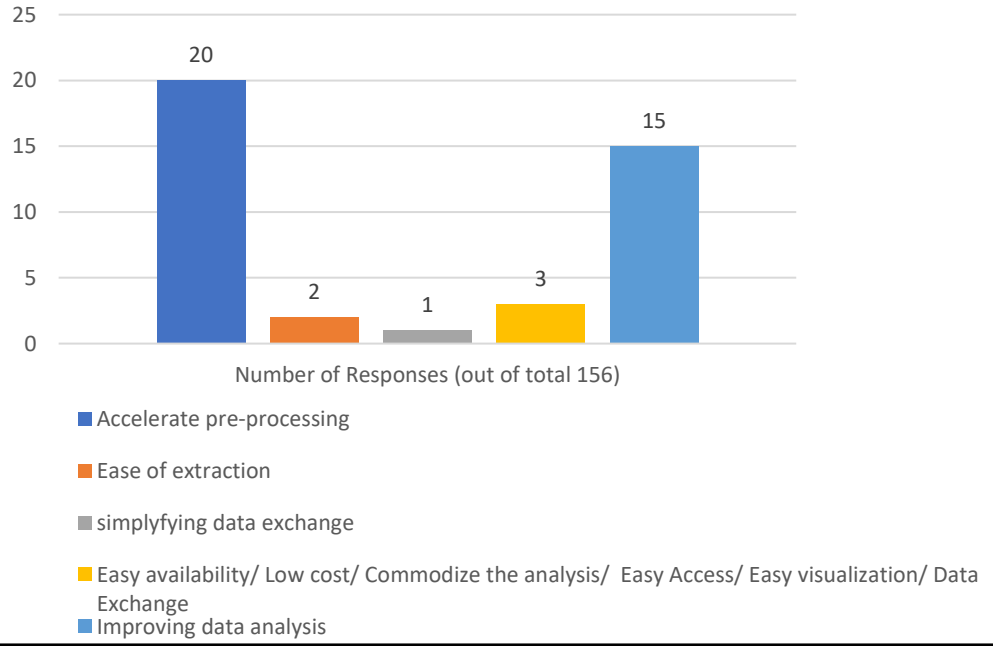
•Low cost

•Commodize the analysis

•Easy Access

•Easy visualization

•Emphasis of PM rather than Data Prep



Identified Benefits	Example Quote
Accelerate pre-processing	*"This would significantly reduce the project delivery time as the data pre-processing time will be reduced"* **(Participant 51)** *"Better reusability and reduction of time in data preparation; more effort of Process Mining providers on really crucial tasks in analytics; better possibilities for end customers to switch between Process Mining providers depending on use case."* **(Participant 112)**
Ease of extraction	*"It makes work of the data scientists easy to extract the data from the sources and build the data models for multi sourced datasets to marry different tables from the sources. "* **(Participant 13)**
Simplifying data exchange	*"The major issue today is that no source system will likely to support a process mining standard format. However, it would help to use different process mining tools since exchange between them would be much easier. A data standard like XES that supports multiple hierarchies and different granularity levels of events / tracked entities would probably also lead to more development regarding large data storages for faster and easier querying."* **(Participant 79)**
Improving data analysis	*"It would help to obtain and analyze correct and complete details. Data pre-processing is expected to be simpler as all industries would be using the same standard."* **(Participant 67)**
Easy availability/ Low cost/ Commodize the analysis/ Easy Access/ Easy visualization/ Data Exchange	*"It would help commoditize the analysis and make it more easily available at a lower price, thus driving operational excellence also in smaller corporations.."* **(Participant 30)** *"Easy way to export / import PM data, Easy way to exchange data with customers and partners, Broader acceptance of PM, Increased user experience for PM tool users"* **(Participant 81)**

Next steps …

- Share ideas on how we can collectively address some of these challenges
- Beyond XES 1.0: what does it mean for the next version of XES?