UFMG

lab-soft
software engineering laboratory

Unveiling Experts in Data Science: A Mining Software Repository Perspective

José Antônio
Eduardo Figueiredo (Advisor)
Johnatan Oliveira (Co Advisor)

Summary

- Introduction
- Study Design
- Dataset and Results
- Thread Validity
- Conclusion

Introduction

- Skills of data scientists as a pillar in software engineering projects
- Difficult to locate experts with strong technical skills in data science
- Addresses this problem by exploring the activity of software repositories

Study Design - Goal

- Discovery of fundamental skills of data science professional
- Provide an outline and description of experts in data-based fields
- Identify relevant repositories with the selected metrics

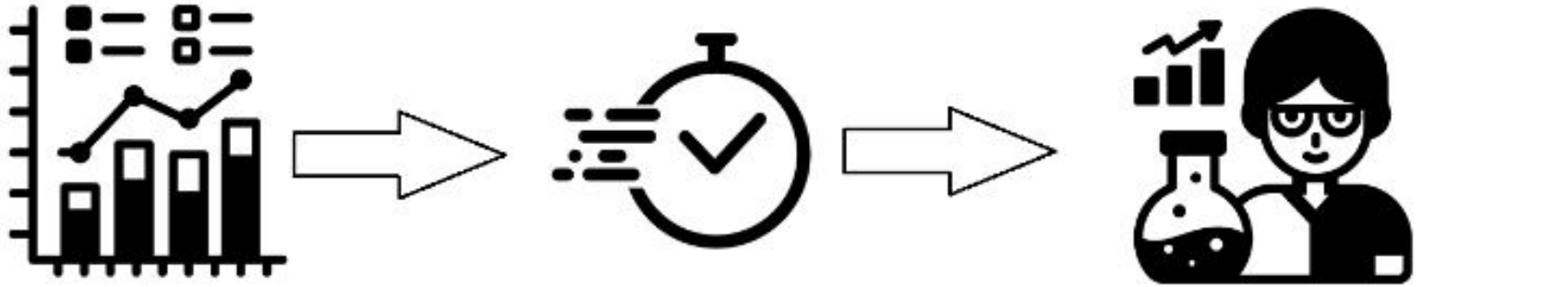
Research Questions

- RQ1 -How the selected metrics provide information on the averages of data scientists individual contributions?
- RQ2 - What are the characteristics of individuals identified as experts in data science projects?

Dataset

- Starting pool of 629 projects that were automatically searched on GitHub
  - Explorer tab using the keyword “Data Science stars:>100”
- Repository with at least two years
- More than 10 active contributors
- After all metrics, we chose a subset of 18

Repositories Select



Used GitHub Repositories

Repository	Stars	Contributors
OpenMined/PySyft	9.2k	423
kedro-org/	9.3k	211
goplus/gop	8.8k	39
Netflix/metaflow	7.5k	88
google/deepvariant	3.1k	24
quadratichq/quadratic	2.7k	22
colour-science/colour	1.9k	45
NannyML/nannyml	1.7k	29
apache/systemds	1k	180
visualpython/visualpython	799	6
LineaLabs/lineapy	653	21
googleapis/python-aiplatform	520	93
IBM/lale	320	25
nebari-dev/nebari	254	63
EpistasisLab/Aliro	219	20
mithril-security/bastionlab	165	12
vertica/VerticaPy	214	16
microsoft/MLOS	123	18

Used Technology and Metrics

- Python framework “Pydriller” to extract and analyze data
- Metrics used to evaluate each data scientist
  - Number of Commits
  - Lines of Code (LOC)
  - Number of Changed Files

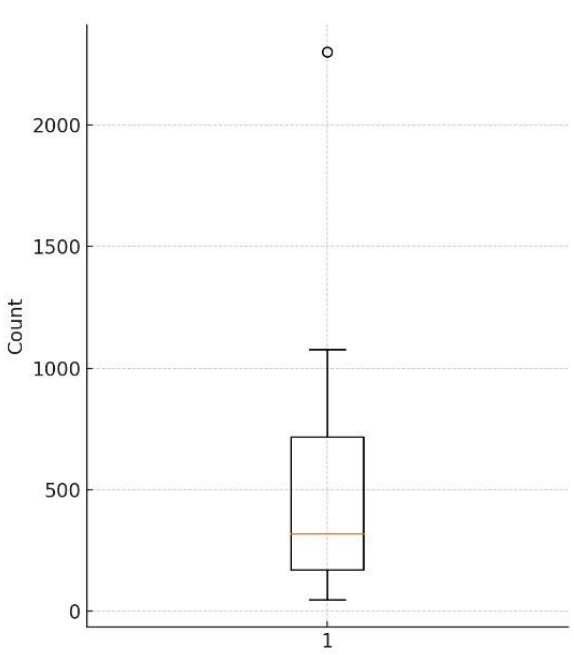
Results - RQ1

- Most developers have a median of more than 400 commits.
  - Few outliers with higher numbers with over than 2,000 commits
- Median number of files changed is around 5000.
  - Extreme outliers with over than 20.000 files
- Median number of LOC added is less than 100K
  - Extreme outliers with over than 5 million LOC added

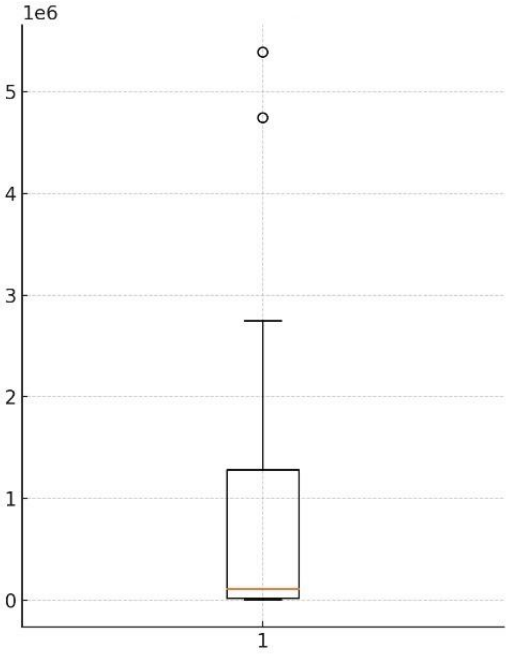
Data Scientists metrics

All Data Scientists Boxplot

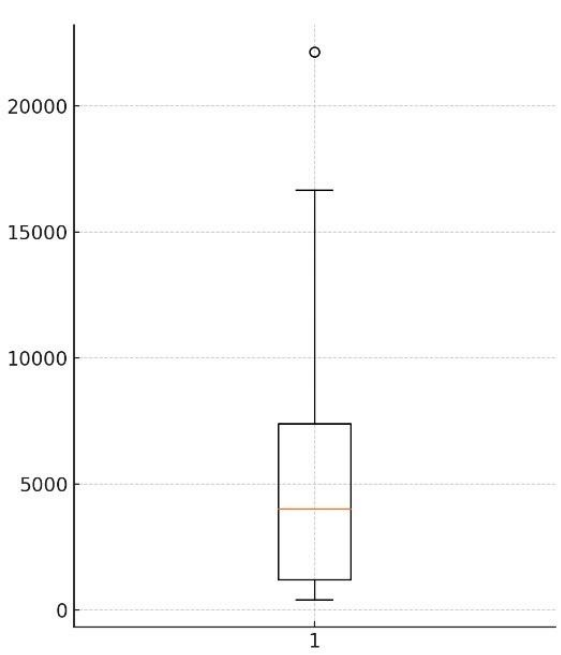
Commits



LOC Added



Files Changed



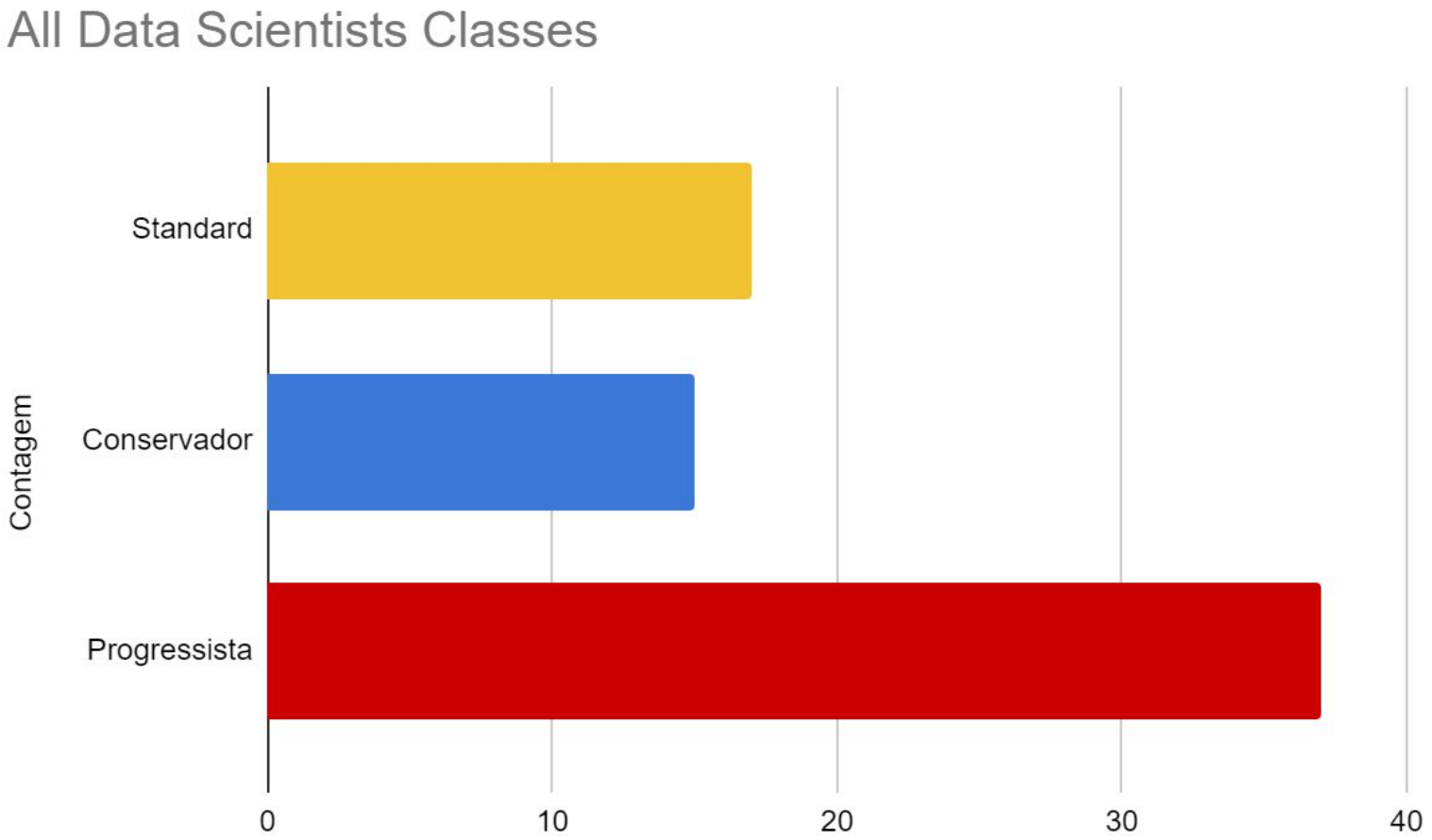
Results - RQ2

- Classes to separate the data scientists
- Progressive - at least 60% of their activity spent on additions and no more than 40% on deletions.
- Conservative - at least 60% of their activity focused on deletions and no more than 40% on additions
- Standard - Neither of metrics exceeding 60%

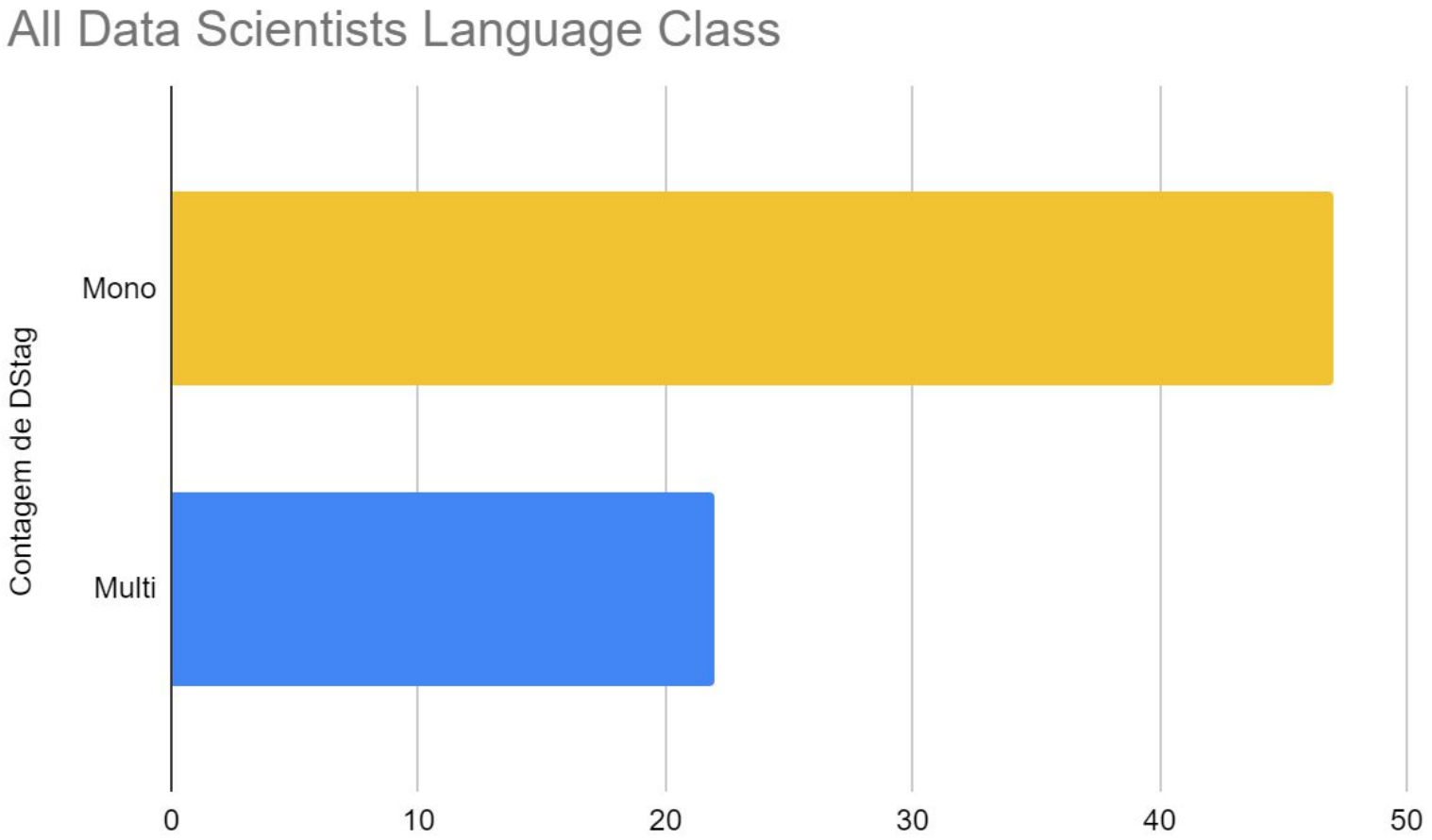
Results - RQ2

- Data scientists progressive coding (53%)
- Data scientists standard approach (25%)
- Data scientists conservative in their coding practices (22%).
- Significant majority of data scientists (68%) are Mono-language. 32% of the data scientists were classified as Multi-language

Results - RQ2


All Data Scientists Classes
Contagem
Standard
Conservador
Progressista
0
10
20
30
40

Results - RQ2


All Data Scientists Language Class
Contagem de DStag
Mono
Multi
0
10
20
30
40
50

Threats of Validity

- Small sample size, future research may expand to a more larger number of projects
- Quantifying data science proficiency through GitHub activities might not adequately represent its complexity,

Related Work

- Oliveira et al. presented a study on the efficacy of two source code analysis models (Changed Files and Changed Lines of Code) in detecting programming talents.
- Saltz et al. investigated software engineers transitioning to data engineering roles using a case study at a big data consulting firm.
- Kim et al. examined the changing role of data scientists in software development using Microsoft as a case study.

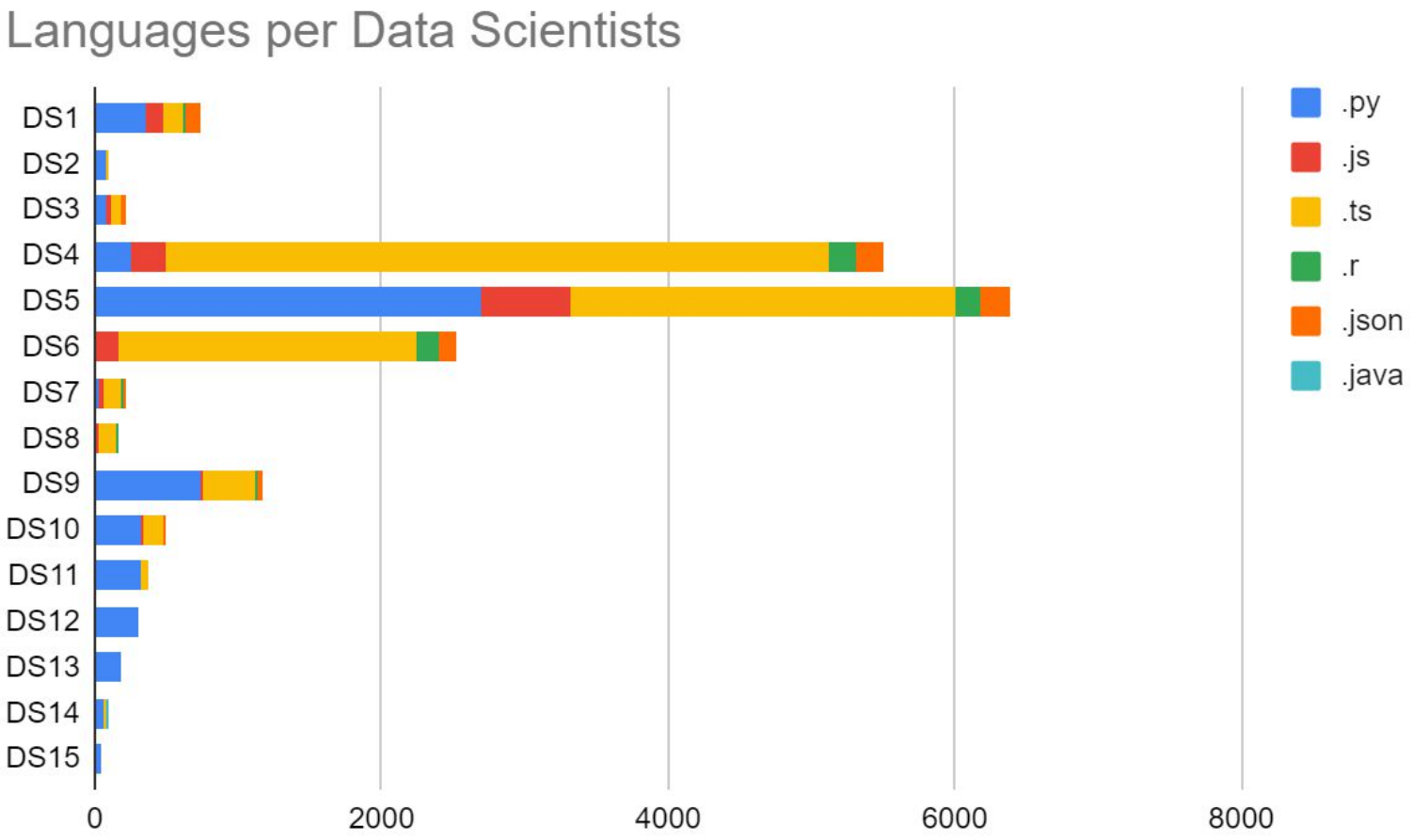
Conclusion

- We analyzed specific metrics and focused on 69 developers for deeper analysis, categorizing the main contributors based on these metrics.
- We created profiles based on selected metrics (LOC, Number of Commits, and Changed Files) and the programming languages used

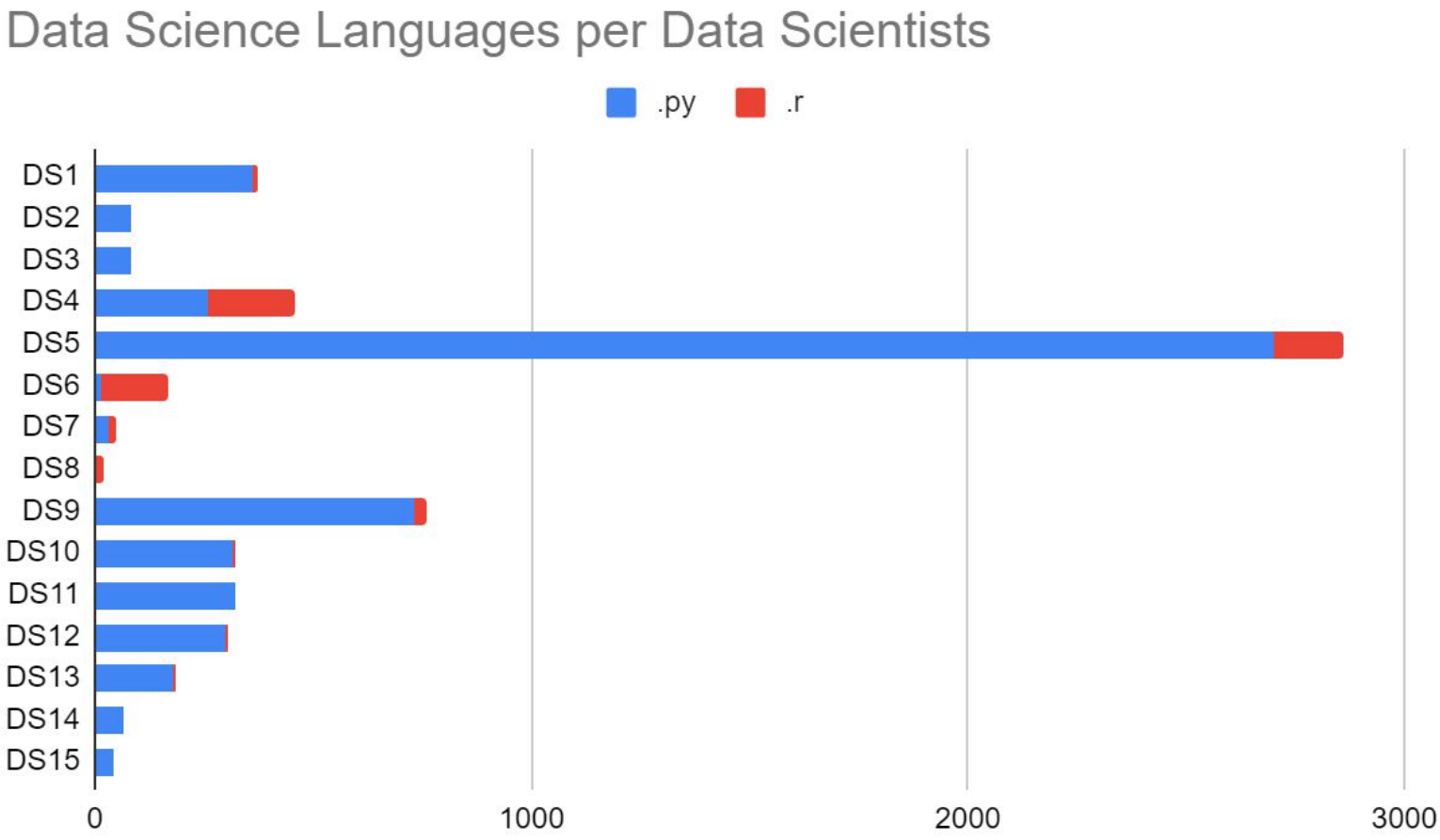
Ongoing Work

- Analyze multi-language data scientists and projects
- Study the team formation in these projects
- How will the team manage if one of the data scientists leaves the team?
- 12 of 15 data scientists are multi-language (Inside the multi-language projects)
- TS is more used among multi-language data scientists

Ongoing Work



Ongoing Work



UFMG

lab-soft
software engineering laboratory

Any Questions?