# 2020 Data Science Survey Highlights

Karl Rexer, PhD
President, Rexer Analytics

www.RexerAnalytics.com

# Highlights

- This was the 9th survey in an ongoing research program that began in 2007. The survey was conducted May-July 2020, and 579 analytic professionals from 71 countries participated.

- SUBSTANTIAL GROWTH IN PYTHON USE: In the last 3 years an additional 15-20% of data scientists have begun using python. Python is now the most commonly used tool in Corporate and Consulting environments. In Academica and NGO/Government environments python use is now surpassed only by R.

- DATA SCIENTISTS ARE CAUTIOUS ABOUT AUTO-ML: 40% feel that Auto-ML tools are very or somewhat problematic. Many respondents provided extreme examples of misuse and bad decisions resulting from Auto-ML use.

- IMPACT OF ENHANCED DATA SECURITY & PRIVACY: Almost half of respondents report that enhanced data security and data privacy (e.g., GDPR and HIPAA) have had a real impact on the analytics at their organization.

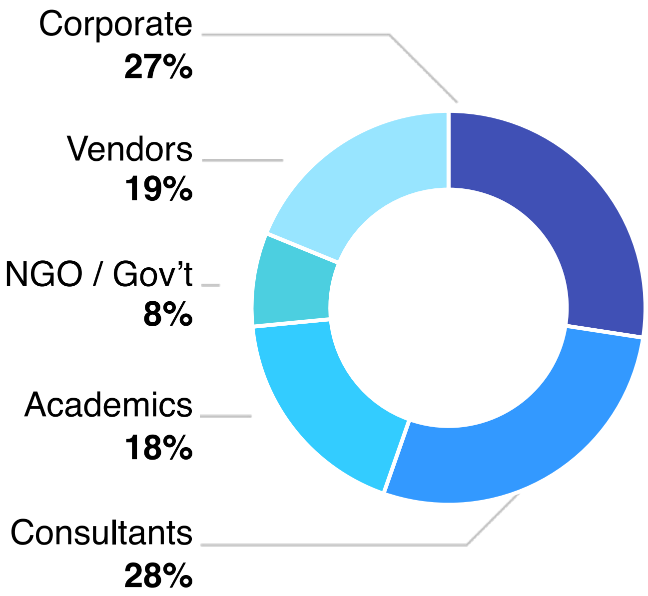
- EARLY COVID IMPACT: In May-July 2020 very few analytic professionals reported losing a job due to the pandemic. Consultants were more likely to report a reduction in projects and income. Over 20% of analytic people working in corporate settings saw an increase in their workload – twice the rate of consultants.

- JOB SATISFACTION: Even in the middle of a global pandemic, data science professionals reported very high job satisfaction.

- EXPANSION OF ANALYTICS: 63% of people report their companies conducted more analytic projects in 2019 than in previous years. Only 8% said their companies were doing fewer analytic projects.

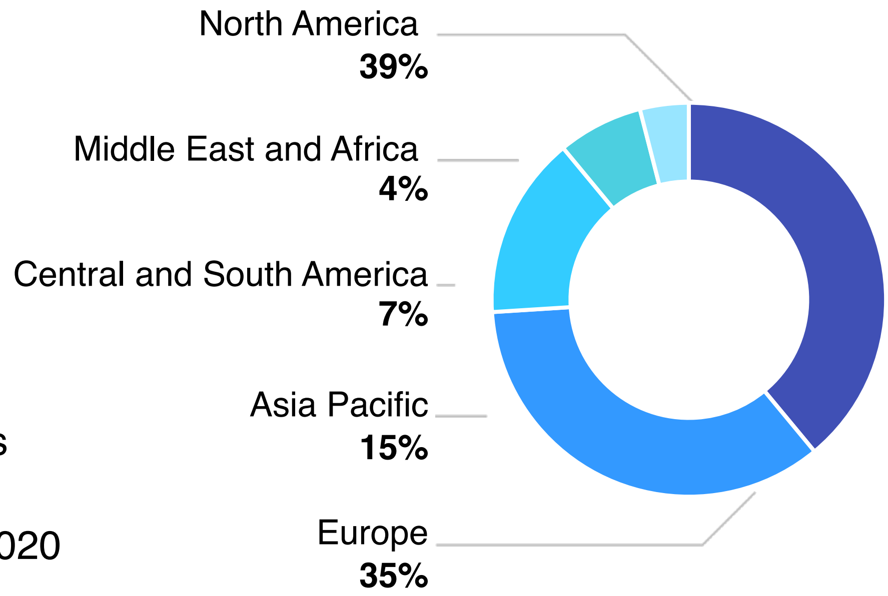
# 2020 Data Science Survey: Overview

- 9th survey since 2007
- 49 questions
- 10,000+ invitations emailed & promoted by newsgroups, vendors and bloggers
- Respondents: 579 analytic professionals from 71 countries
- Data collected in May to July 2020

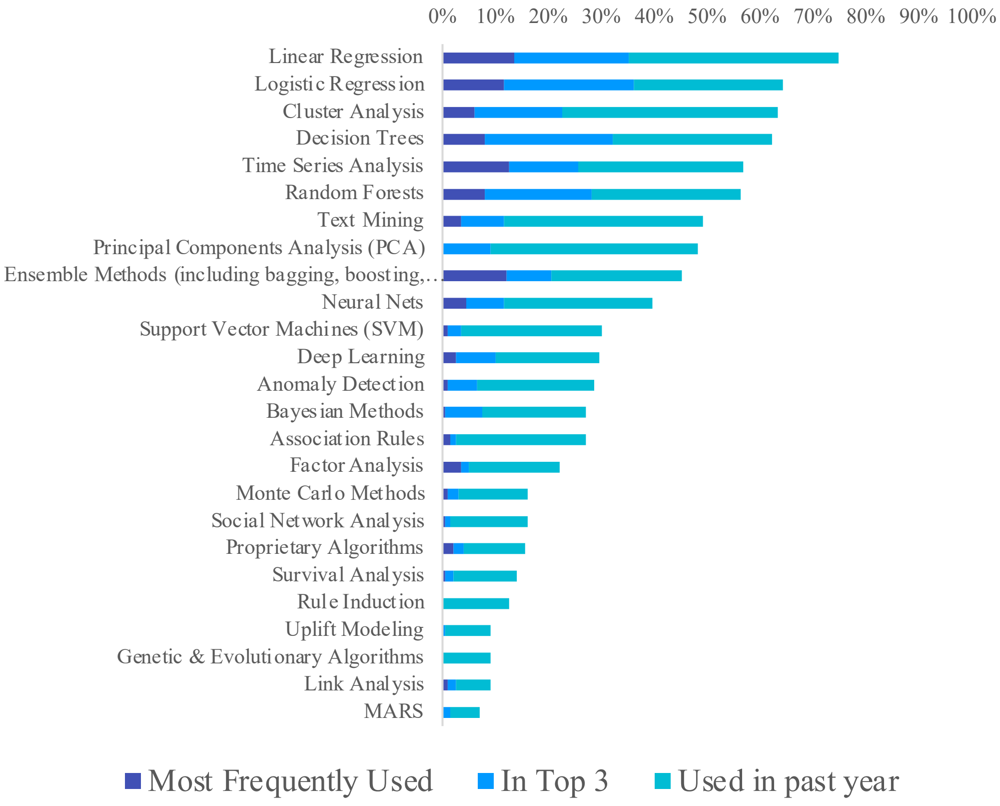


# 2020 Data Science Survey: Overview - Geography

- 9th survey since 2007
- 49 questions
- 10,000+ invitations emailed & promoted by newsgroups, vendors and bloggers
- Respondents: 579 analytic professionals from 71 countries
- Data collected in May to July 2020



# What Algorithms are being used?




The average 2020 respondent reports using 9 algorithms

In every survey (since 2007), regression, decision trees and cluster analysis consistently form the top triad of algorithms

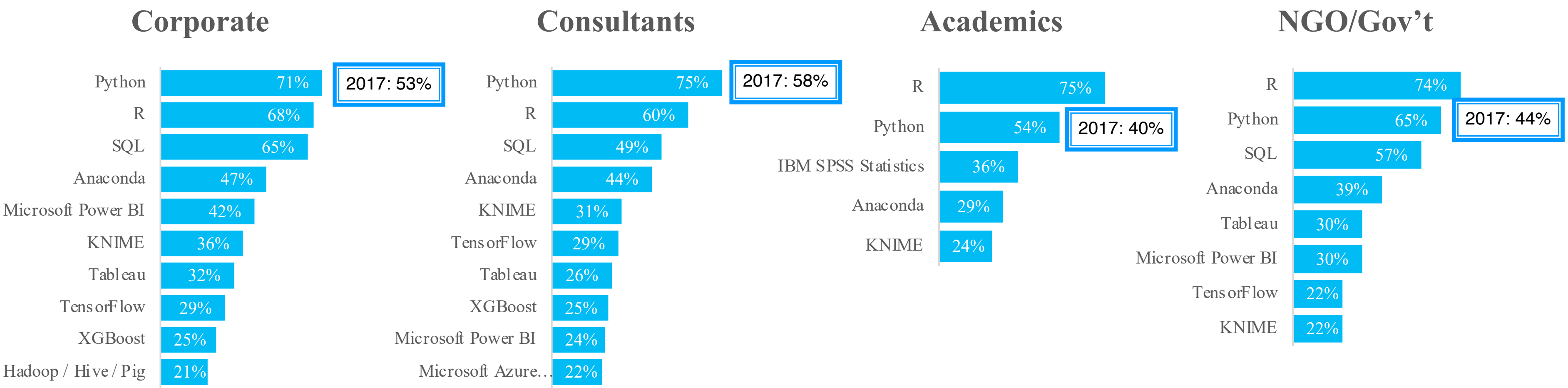
This year we separately asked about linear and logistic regression: Both are used very often

This year we also asked about Principal Components Analysis: Close to 50% of people report using it

Several algorithms, such as Association Rules and SVMs are not likely to be the top analytic techniques people use, but are important in certain situations

# Most Data Scientists use Multiple Tools

What data science / analytic tools, technologies, and languages did you use in the past year?

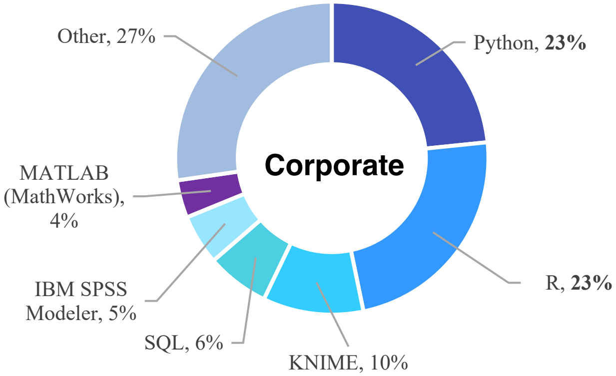


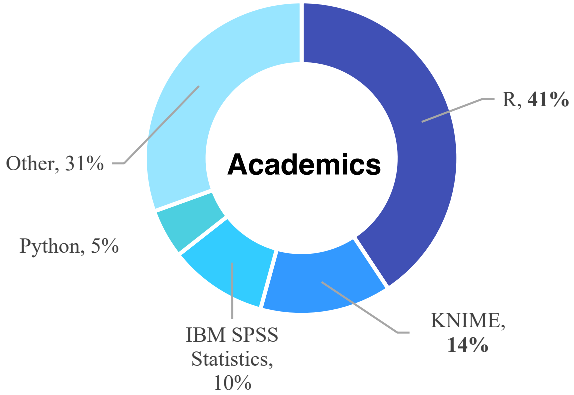

Across our four primary professional categories, Python and R are unsurprisingly the dominant analytics tools used. Python, in particular, has shown substantial growth in adoption in the past 3-5 years,

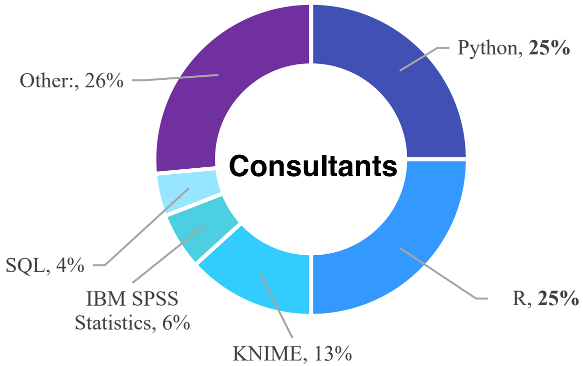
All tools used by more than 20% of a group are shown

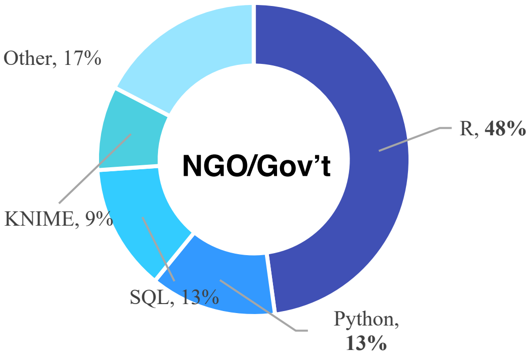
# Primary Tools

What one data science tool, technology or language do you use most frequently?



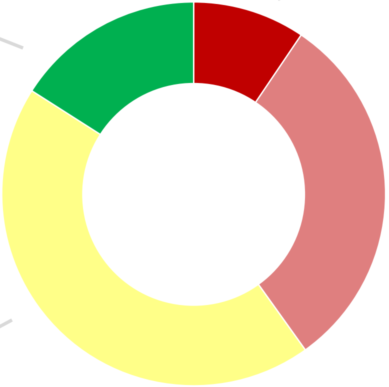



# Citizen Data Scientists and Auto-ML Tools

Feelings about Automated Machine Learning Tools are mixed

**These tools are great:** – **16%** Citizen Data Scientists using these tools are accomplishing great things.

**9%** – **These tools are very problematic:** Most of the analyses and models Citizen Data Scientists create using automated analytic tools have serious flaws and are problematic for the businesses using them.



**31%** – **These tools are somewhat problematic:** Occasionally Citizen Data Scientists using these tools accomplish good things, but mostly the results are problematic.

**44%** – **These tools are OK:** Citizen Data Scientists using these tools are accomplishing some good things, but there are also substantial problems with some of the analyses and models.

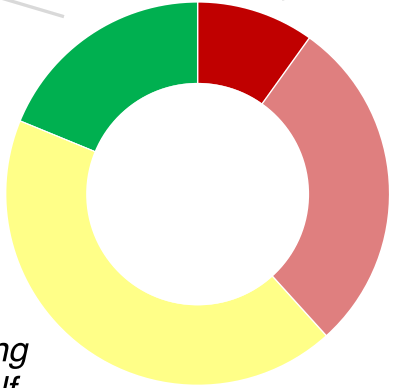
# Citizen Data Scientists and Auto-ML Tools: Quotes

Feelings about Automated Machine Learning Tools are mixed

**These tools are great**

*"It makes [Citizen Data Scientists] feel empowered and encourages them to experiment with modelling. It also teaches them that [there's] no such thing as THE model."*

*"I'm a citizen data scientist and my work is effective because the results, when used, provide the right framework for making decisions. Automatic tools (not AutoML itself but other visual tools that automate parts of the data science process) are excellent at removing the coding burden and the need to spend a lot of time coding that which can be done via a button (or two). The problem is when Citizen Data Scientists can't read the analysis. It's not the tool."*



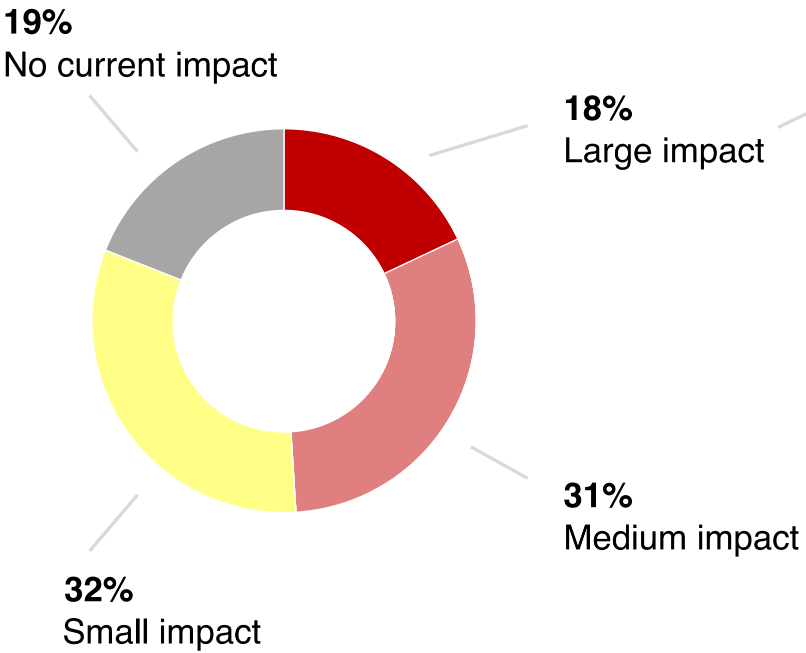
**These tools are very problematic**

*"Overfitting. Confusion about correlation and causality. p-hacking. Meaningless analyses. Misinterpretation of inputs, intermediates and outputs."*

*"Most people do no understand problem solving basic (but hard core) concepts. Such as cause-effect, generalization, cognitive biases, overfitting, GIGO, etc. Data Science is not about using algorithms. That's the easy part. Data Science is about defining your domain problem in terms of data."*

# The Impact of Enhanced Data Security and Privacy

Almost half of people report that enhanced data security and data privacy (e.g., GDPR and HIPAA) have had a real impact on the analytics at their organization



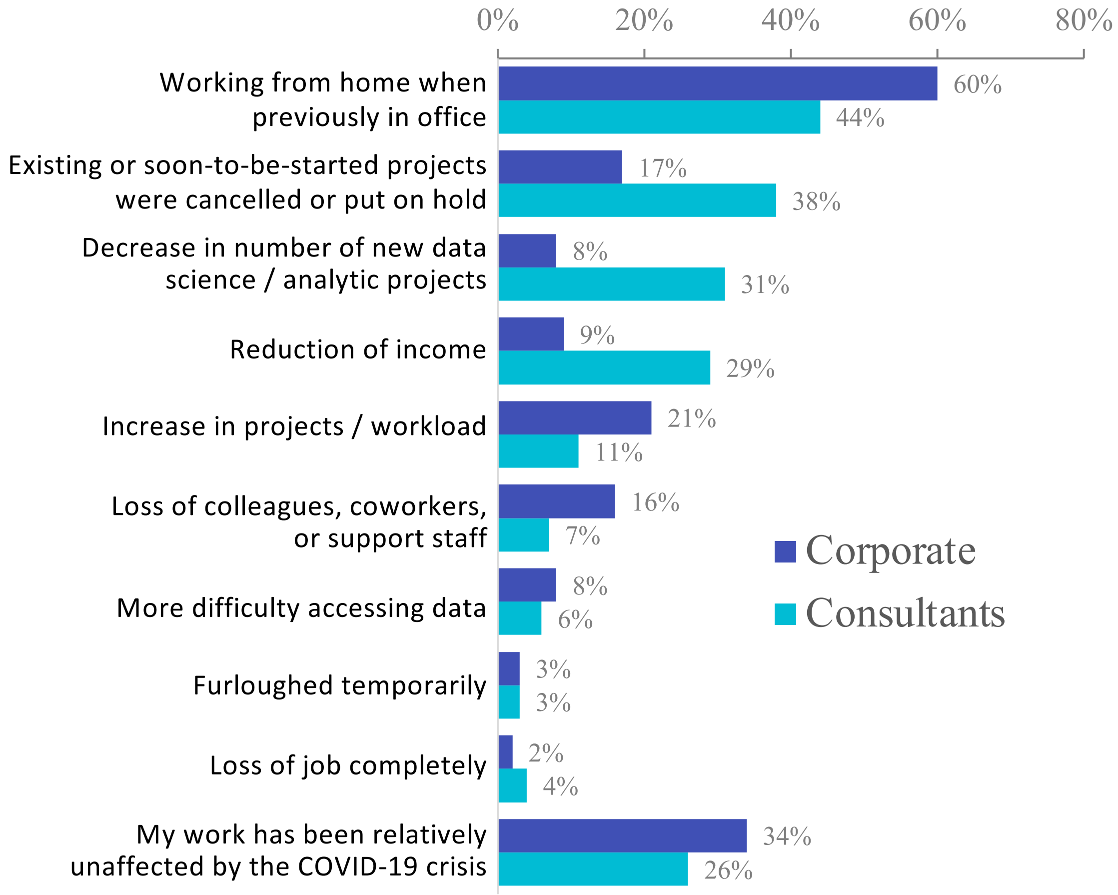

*"In many cases, the data exists, is of good quality, but cannot be used due to debatable security or business confidentiality issues."*

*"GDPR means that data collected by third parties (such as when looking at eCommerce data) may not be shared at the individual level so doing anything that links purchase behavior to attitude to journey becomes impossible at the granular level."*

*"The levels of administrative oversight of human subjects protection issues; long delays between submitting protocols and receiving approval; insufficient IT support related to creating HIPPA approved databases."*

*"This year we completed a project for an organization where nearly half the project time was used working through security issues."*

# The Impact of COVID-19




Very few data analytic professionals report losing their jobs due to COVID-19

The most common COVID impact is the transition to working at home. Data analytic professionals who work directly for companies are experiencing this more than consultants.

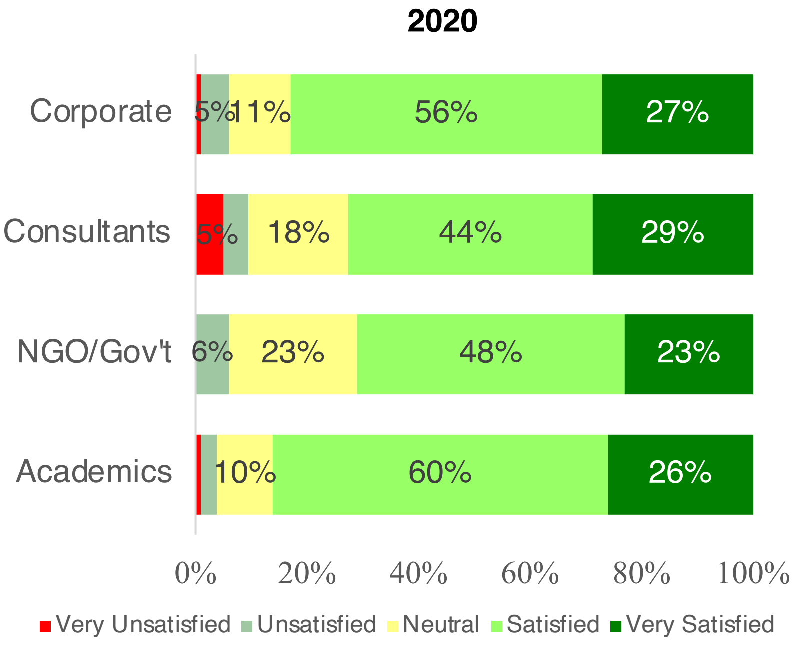
Consultants are more likely to have a reduction in projects and income

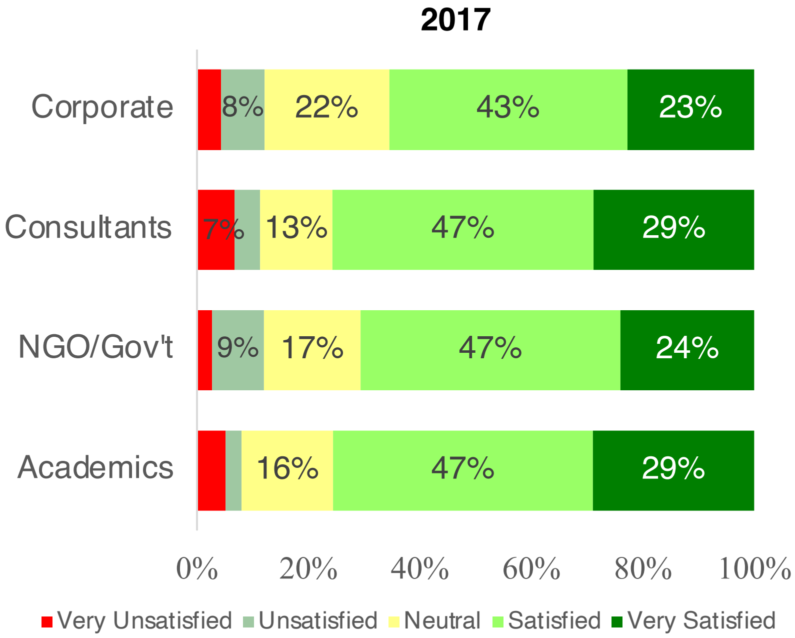
Over 20% of analytic people working in corporate settings are seeing an increase in their workload – twice the rate of consultants

# Employment:  Job Satisfaction Remains High

Even in the middle of a global pandemic, data science professionals reported very high job satisfaction.

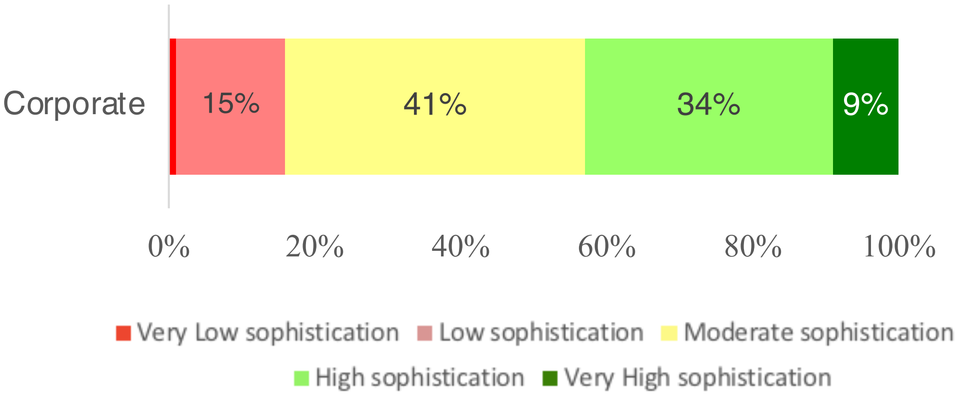
What is your current level of job satisfaction?



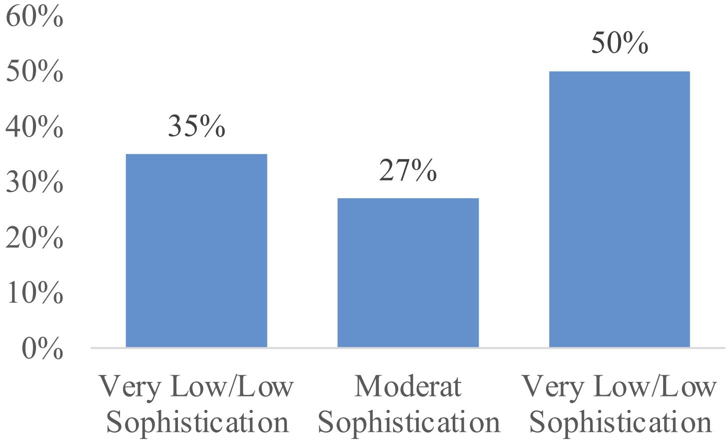
# Analytic Sophistication: Opportunity for Improvement

In general, with what degree of sophistication does your company / organization approach analytic problems?




Only a minority of data scientists feel that their firms have high analytic sophistication

Percent Highly Satisfied with Job




Interpretation: When a company is judged to have high or very high analytic sophistication, 50% of analytic staff report being very satisfied with their jobs.

# Good News

- Very few data analytic professionals report losing their jobs due to COVID-19

- Even in the middle of a global pandemic, data science professionals are reporting very high job satisfaction levels (no drop from previous years)

- 63% of people report that their companies conducted more analytic projects in 2019 than in previous years. Only 8% said their companies were doing fewer analytic projects.

- A slightly increasing number of people report the analytic sophistication of their organizations is high or very high (43% in 2020, 39% in 2017)

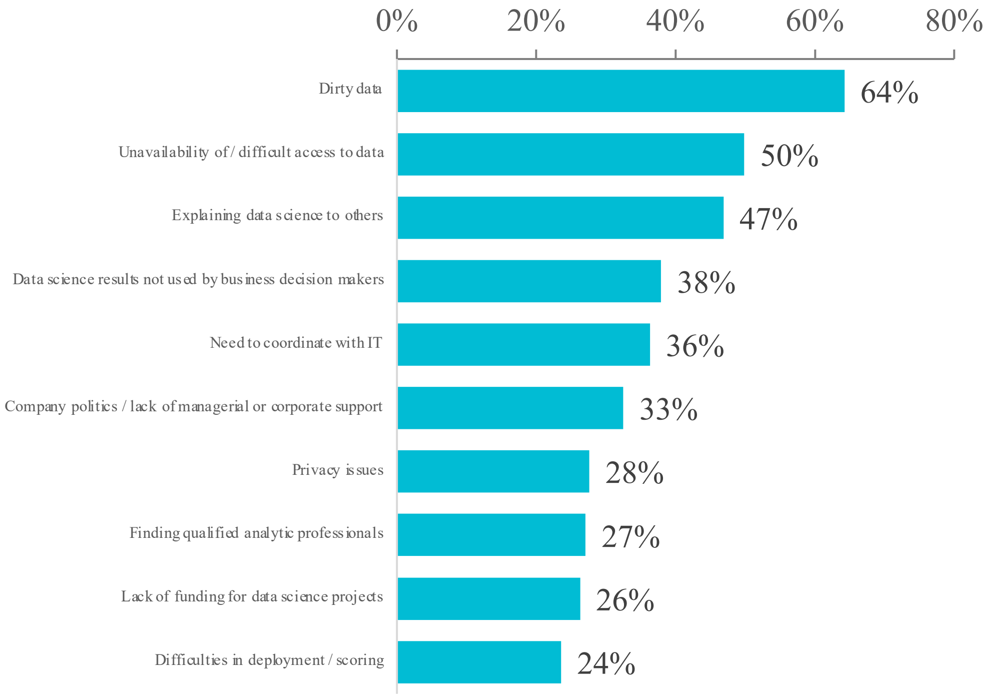
REXER ANALYTICS

# Appendix: Additional Results from the 2020 Survey

www.RexerAnalytics.com

# Barriers Faced by Analytic Professionals

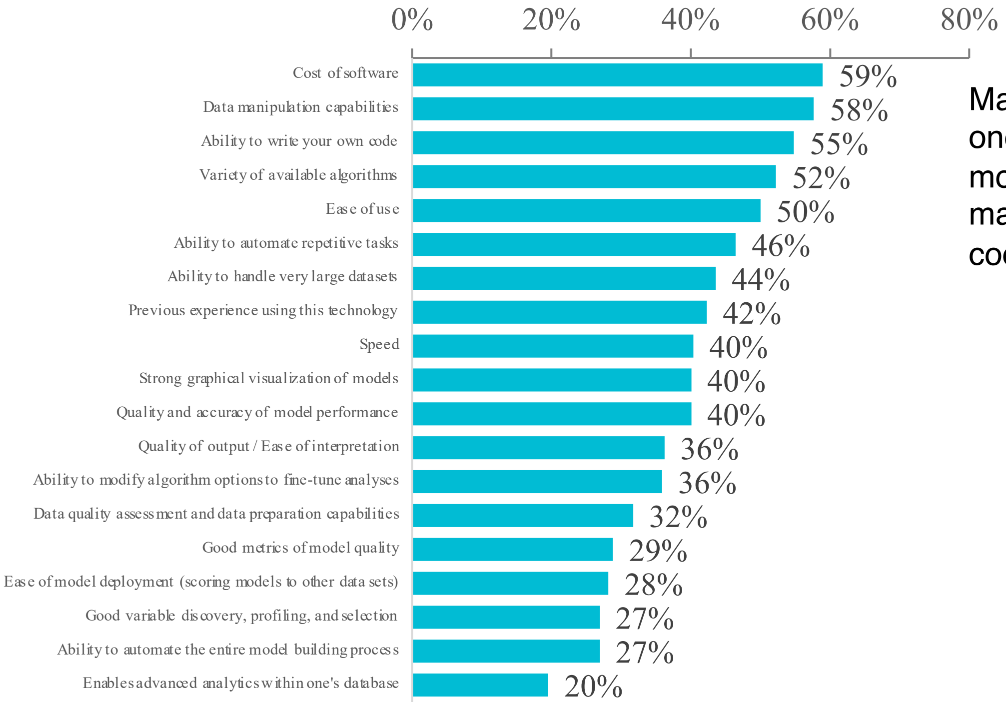
What barriers or challenges do analytic professionals in your organization TYPICALLY face? (Select all that apply)




Dirty data, lack of access to data, and trying to explain data science to others are the most common barriers faced by analytic professionals.

# Primary Factors in Selecting Primary Data Science Technology

Which of the following factors did you consider when selecting your primary data science technology? (Select all that apply)

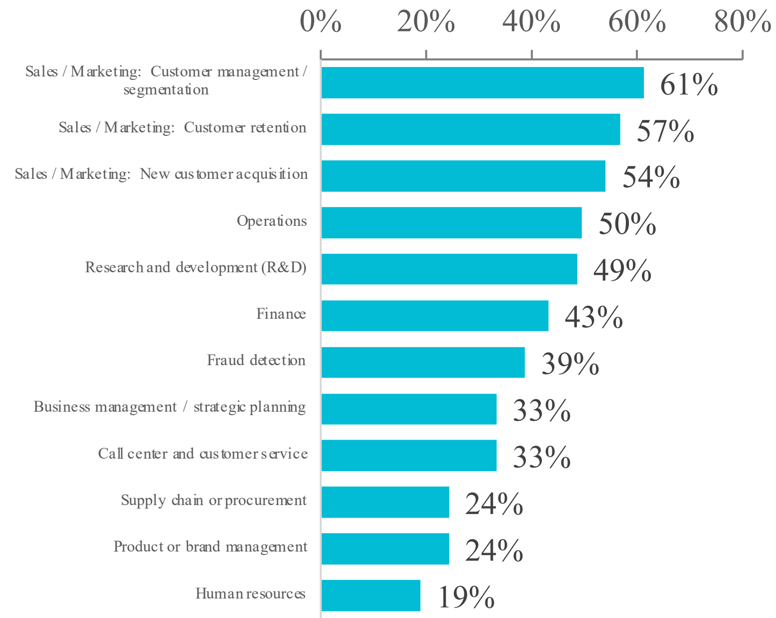



Many factors are considered when selecting one's primary data science software. Among the most frequent factors are cost of software, data manipulation capabilities, ability to write own code, and variety of algorithms.

# Where is Data Science being Applied?

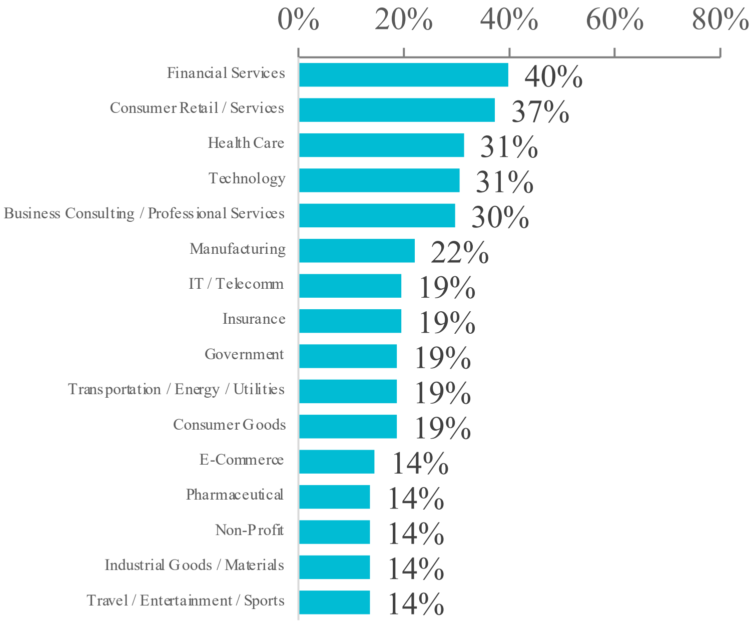
## Corporate

In what field do you typically apply data science?




## Consultants

For what industries did you provide consulting in the past year? (Select all that apply)



For additional information, please contact

Karl Rexer, PhD
President
krexer@RexerAnalytics
617-233-8185

www.RexerAnalytics.com