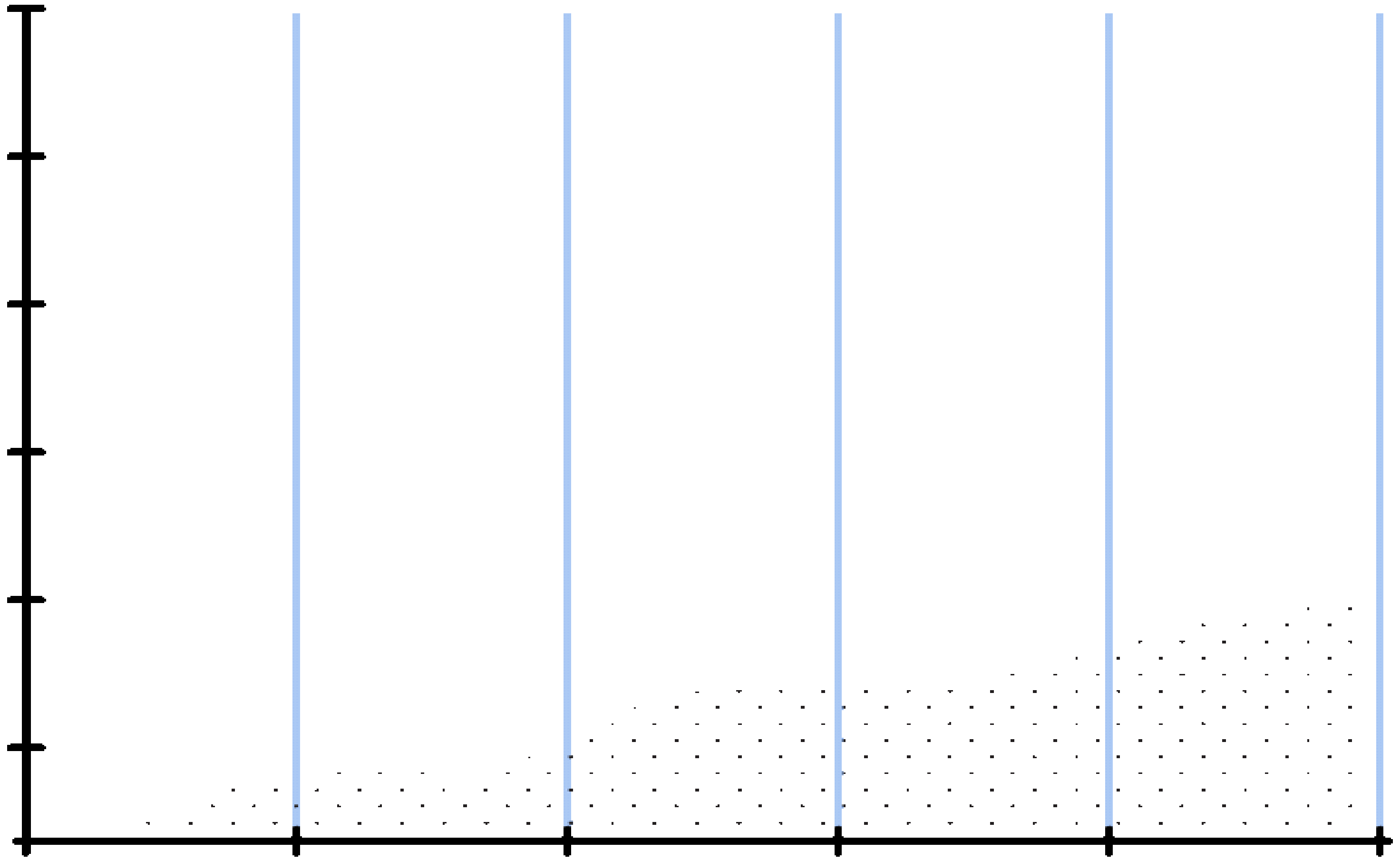
DSU

Data Science University



Analyzing the trends in Data Science Industry

Institutional Goal :

"Empowering the next generation of Analytics leaders through education and guidance."

BY THYNK ANALYTICS

OUR TEAM

Anagh Shrivastava

Khyati Desai

Kritika Dhingra

Parth Aloni

Sara Dhoot

Data Science University

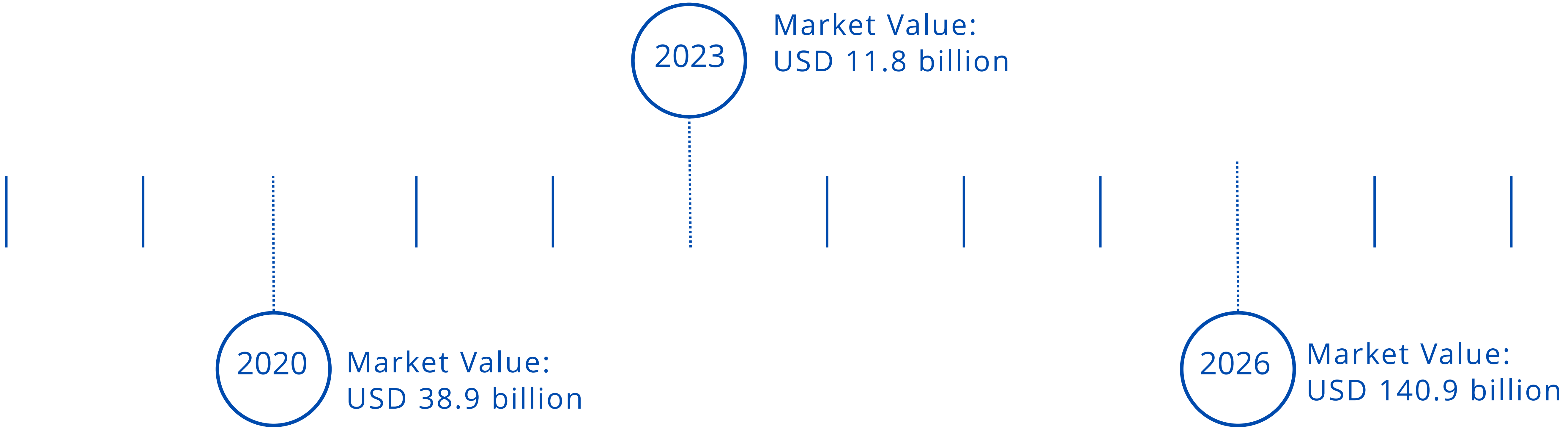
EXECUTIVE SUMMARY

THYNK ANALYTICS

- Examining trends in the analytics profession to provide insights/recommendations to an educational institution on survey responses from Kaggle Survey 2022 .
- It recorded over 25,000 data science practitioner responses which includes students and professionals.
- Conducted data cleaning/preparation and statistical analysis
- Used the kNN-supervised machine learning algorithm to build a tool to predict salaries based on user input.
- The Analysis provided insights into the significant factors that drive the salaries of data science professionals, and what kind of technologies and tools are prevalent and emerging in the field.
- Based on predictive modeling and insights the team is providing the following recommendations to the institute -
  - Focus on Higher Education
  - Build industry-specific knowledge
  - Keep up with the most used and emerging technologies and tools in Analytics, ML, and AI fields.

DATA SCIENCE

The global data science market size was valued at USD 38.9 billion in 2020 and is expected to reach USD 140.9 billion by 2026, growing at a compound annual growth rate (CAGR) of 24.2% from 2021 to 2026*



2023
Market Value:
USD 11.8 billion

2020
Market Value:
USD 38.9 billion

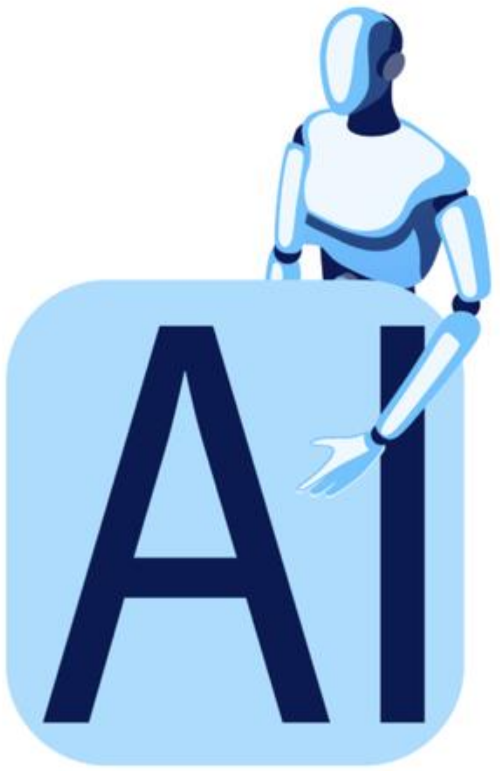
2026
Market Value:
USD 140.9 billion

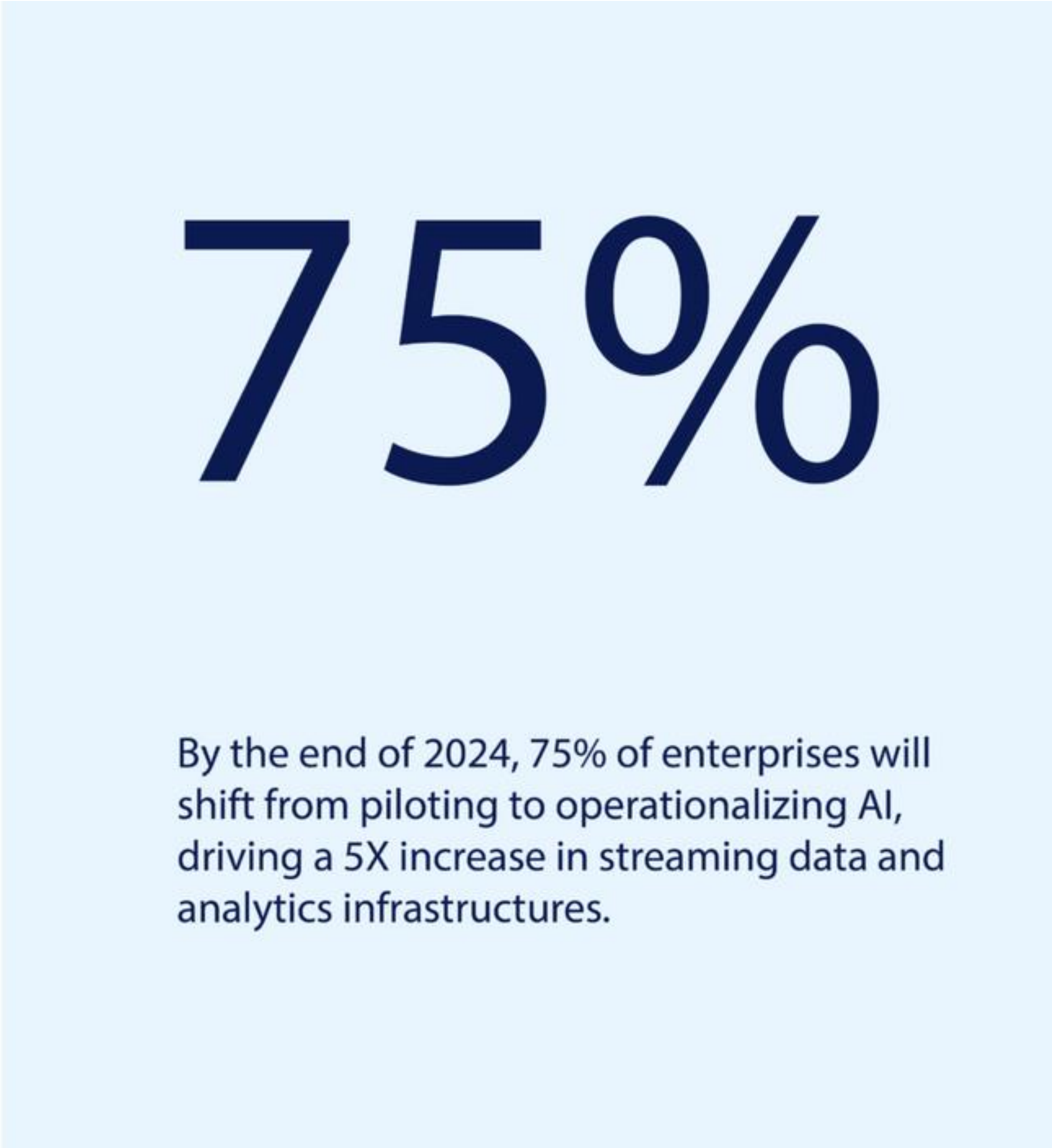
* : Source: MarketsandMarkets

PROBLEM STATEMENT

To effectively prepare and counsel students interested in analytic professions, but they lack understanding of the latest trends.
Also, they need a tool to provide students with a projected salary based on their qualifications.

Goal:
To solve this and to examine the trends and create a tool that will help students prepare for the fast-growing field of data analytics.



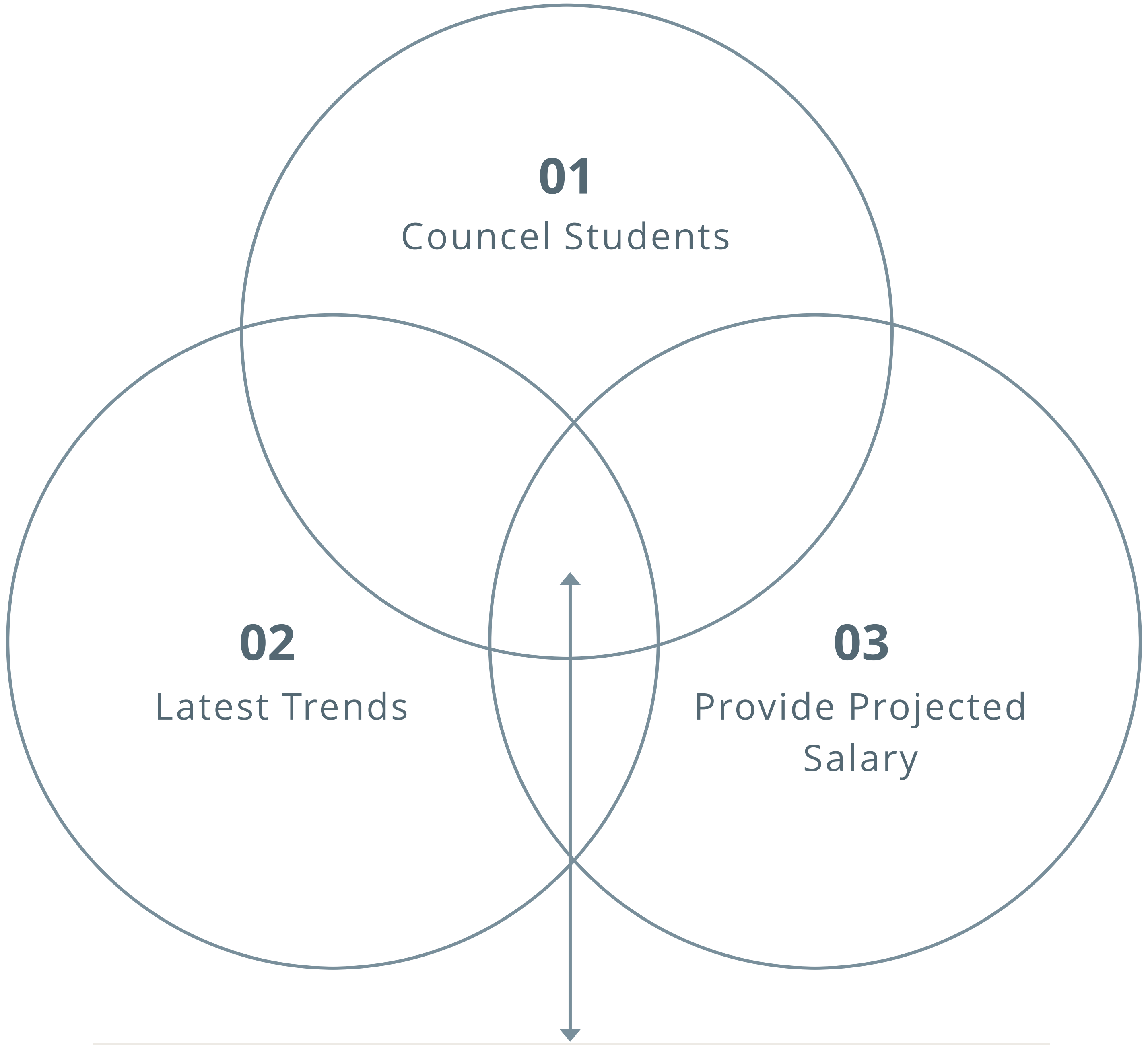


75%

By the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI, driving a 5X increase in streaming data and analytics infrastructures.

Data Science University

PROBLEM ANALYSIS



01
Councel Students

02
Latest Trends

03
Provide Projected Salary

Developing the perfect curriculum

THYNK ANALYTICS

DATA ANALYSIS

To ensure that the dataset is representative for the analysis, the following criteria will be applied to include only responses from employed professionals:

- The respondents must **not currently be students**.
- The respondents must **currently be employed**.
- The respondents must have provided information on the **industry in which they are currently employed**, or their most recent employer if retired.
- The respondents must have answered the question regarding the **industry of their current employer or contract.**

The resulting dataset will include **approximately 37.9%** of the total responses and will be used for the analysis.

1

Data Cleaning

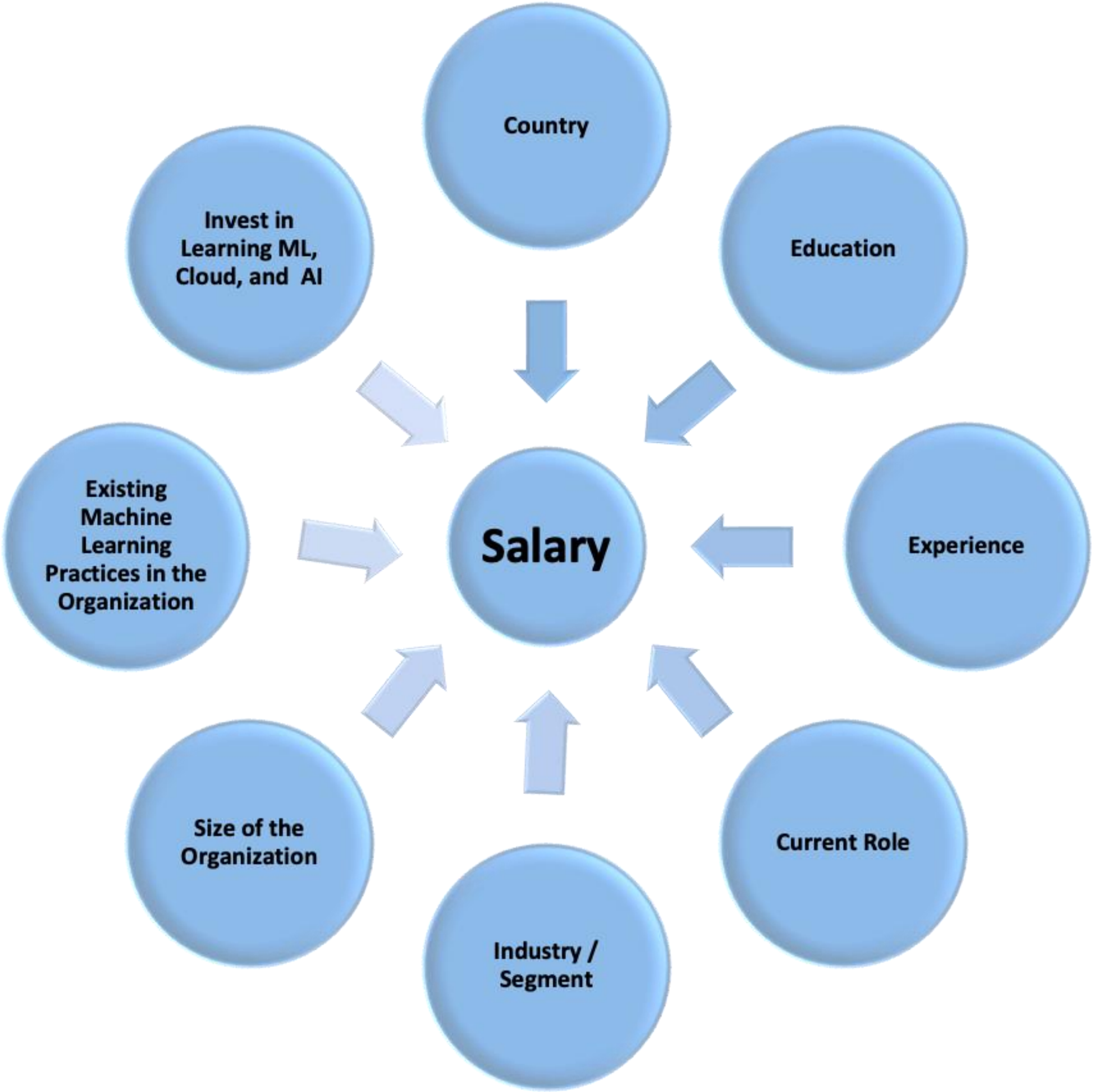
2

Data Manipulation

3

Data Analysis

SIGNIFICANT FACTORS DRIVING DATA SCIENCE PROFESSIONALS' SALARIES?


Country
Invest in Learning ML, Cloud, and AI
Education
Existing Machine Learning Practices in the Organization
Salary
Experience
Size of the Organization
Current Role
Industry / Segment

- Country
- Education
- Experience
- Current Role
- Industry/ Segment
- Size of Organization
- Existing ML practices in Org.
- Investment in Learning ML, Cloud, AI

THE ROLES AND RESPONSIBILITES OF A DATA SCIENCE PROFESSIONAL:

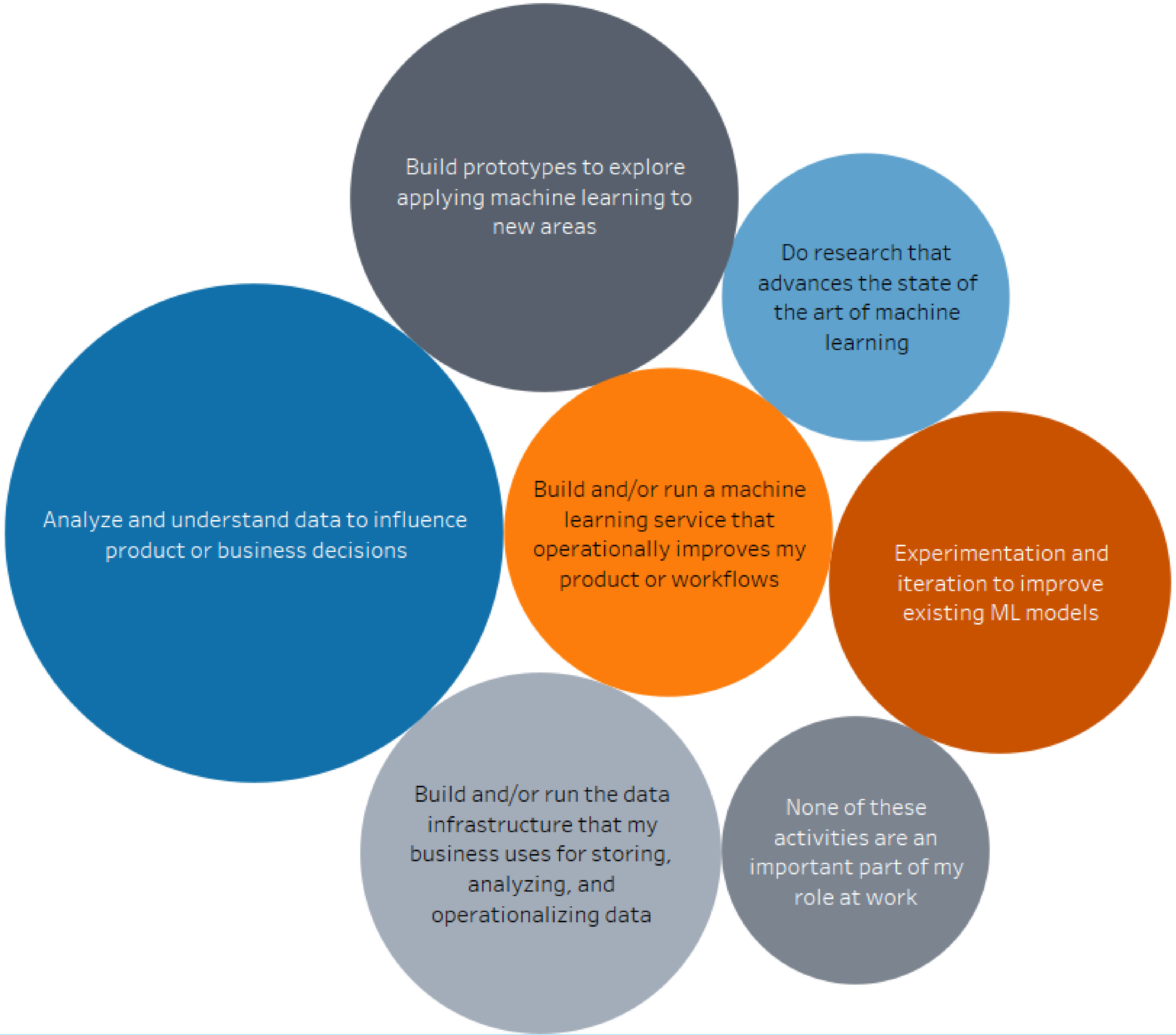
Important parts of the role:

1 Analyze and understand data to influence product or business decisions

2 Build prototypes to explore applying machine learning to new areas

3 Build and/or run a machine learning service that operationally improves my product or workflows

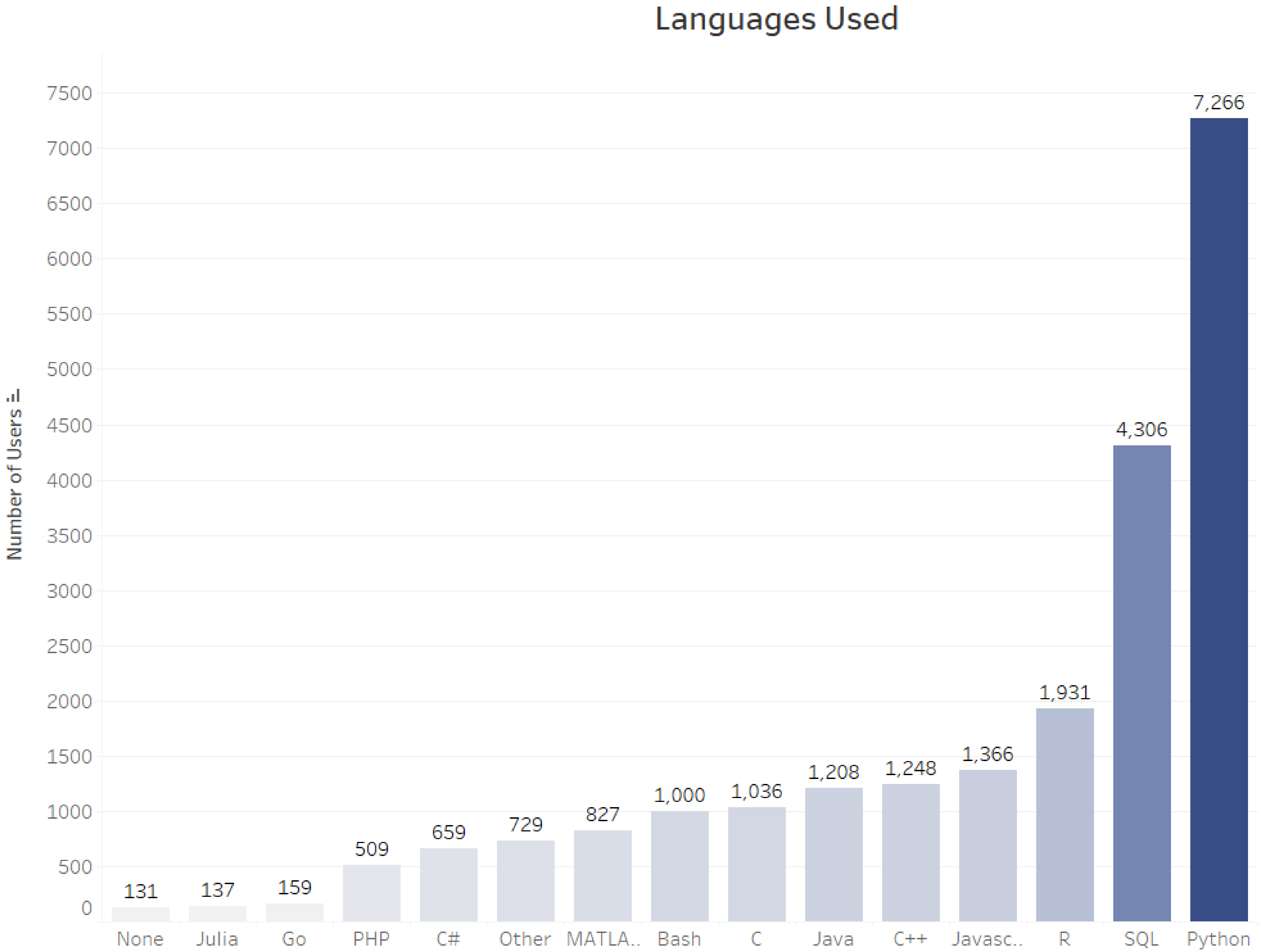
4 Experimentation and iteration to improve existing ML models



Data Science University

PREVELANT TOOLS AND TECHNIQUES

THYNK ANALYTICS



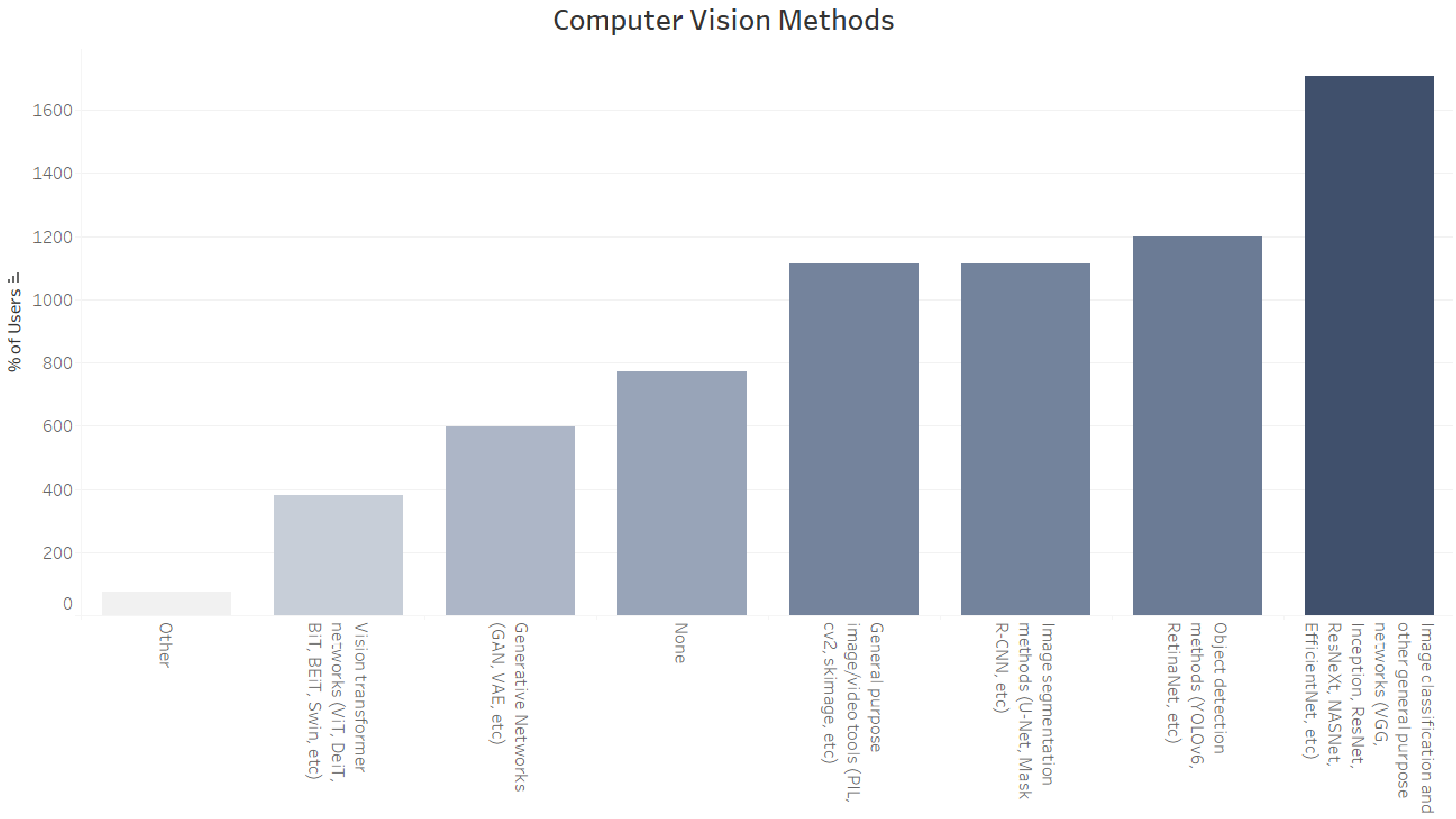
Most Used Languages in the field of Data Science:

- Python
- SQL
- R

Why to use these languages:

Python, R, and SQL are popular choices in the field of data science due to their versatility, functionality, and ability to work together to analyze and manipulate large datasets.

MOST USED COMPUTER VISION METHODS



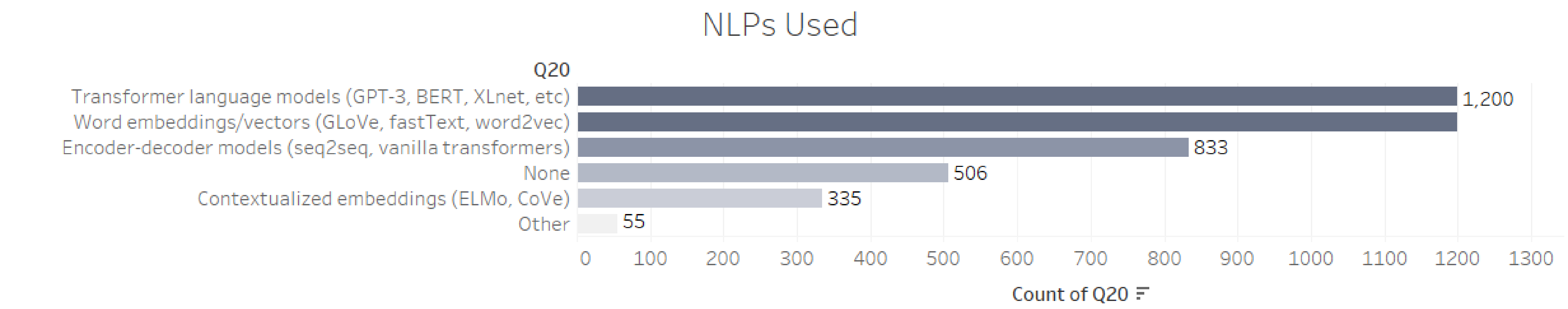
Why to use computer vision methods?

Computer vision techniques can generate new data from existing datasets, improving accuracy of machine learning models with limited data.

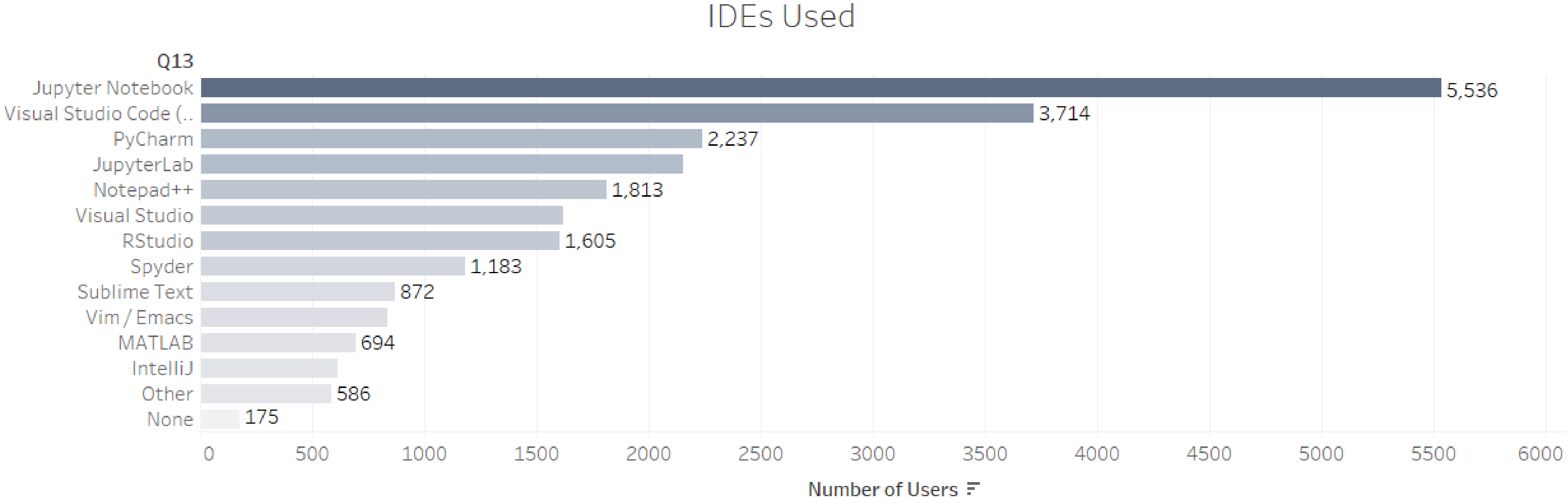
Most used:

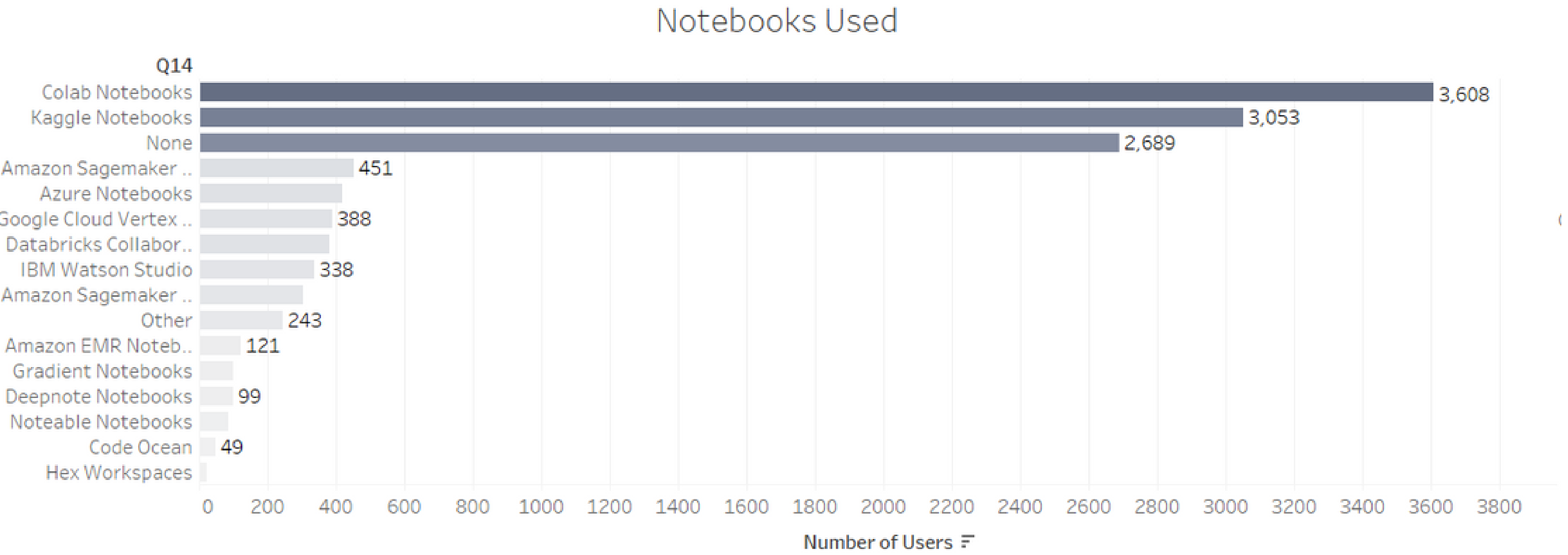
Image classification and other general purpose networks (VGG, Inception, ResNet, ResNeXt, NASNet, EfficientNet, etc)

THE MOST POPULAR USED TOOLS BY DATA SCIENCE PROFESSIONALS:



Count of Q20 for each Q20. Color shows count of Q20. The marks are labeled by count of Q20. The view is filtered on Q20, which excludes NA.

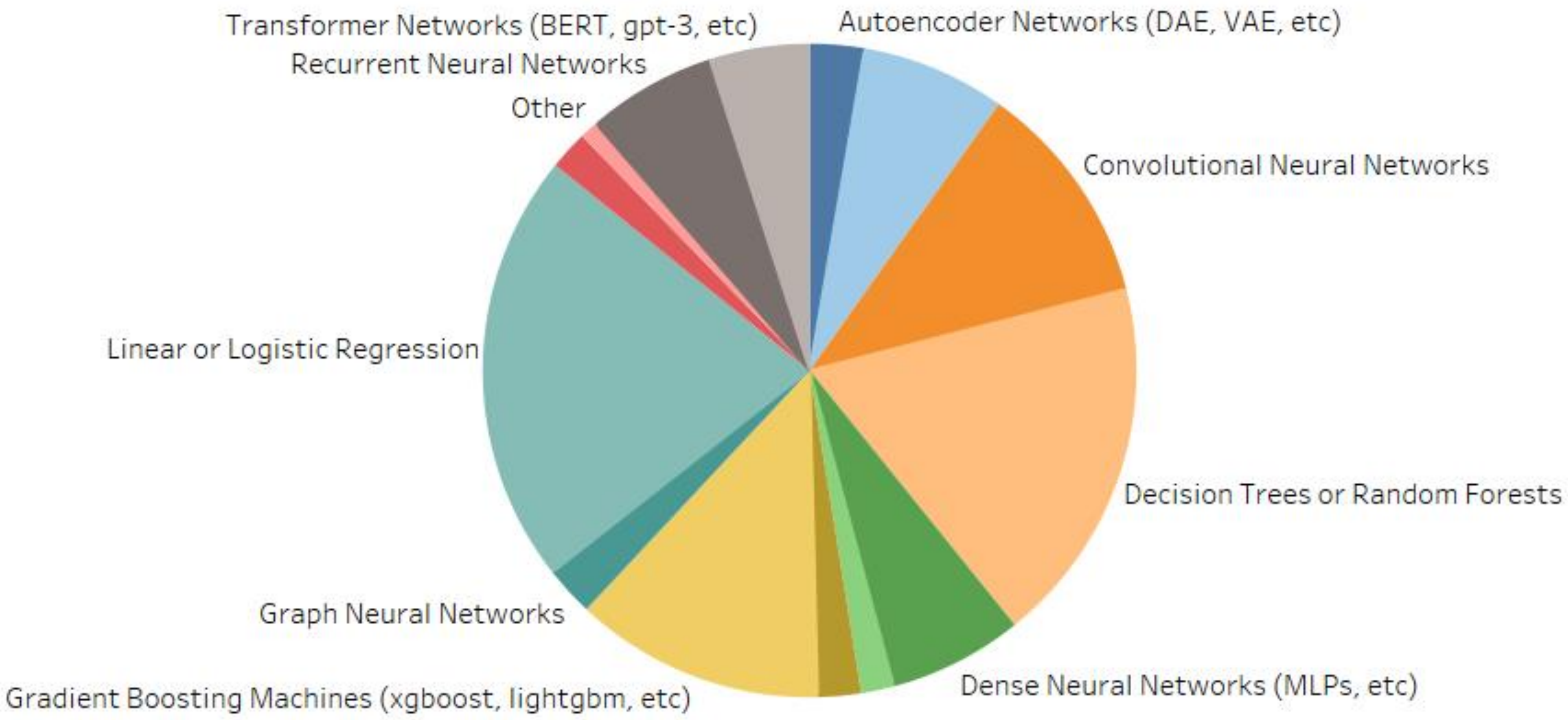




- NLP: Transformer language models (GPT-3, BERT, XLnet, etc)
- IDE: Jupyter Notebook
- Notebook: Colab Notebooks

THYNK ANALYTICS

MACHINE LEARNING ALGORITHMS



Most used:

- Linear or Logistic Regression
- Decision Trees or Random Forests
- Gradient Boosting Machines

Why to use ML Algorithms?

Identify:

- Patterns and trends in large datasets
- Automate processes
- accurate predictions

Advantages:

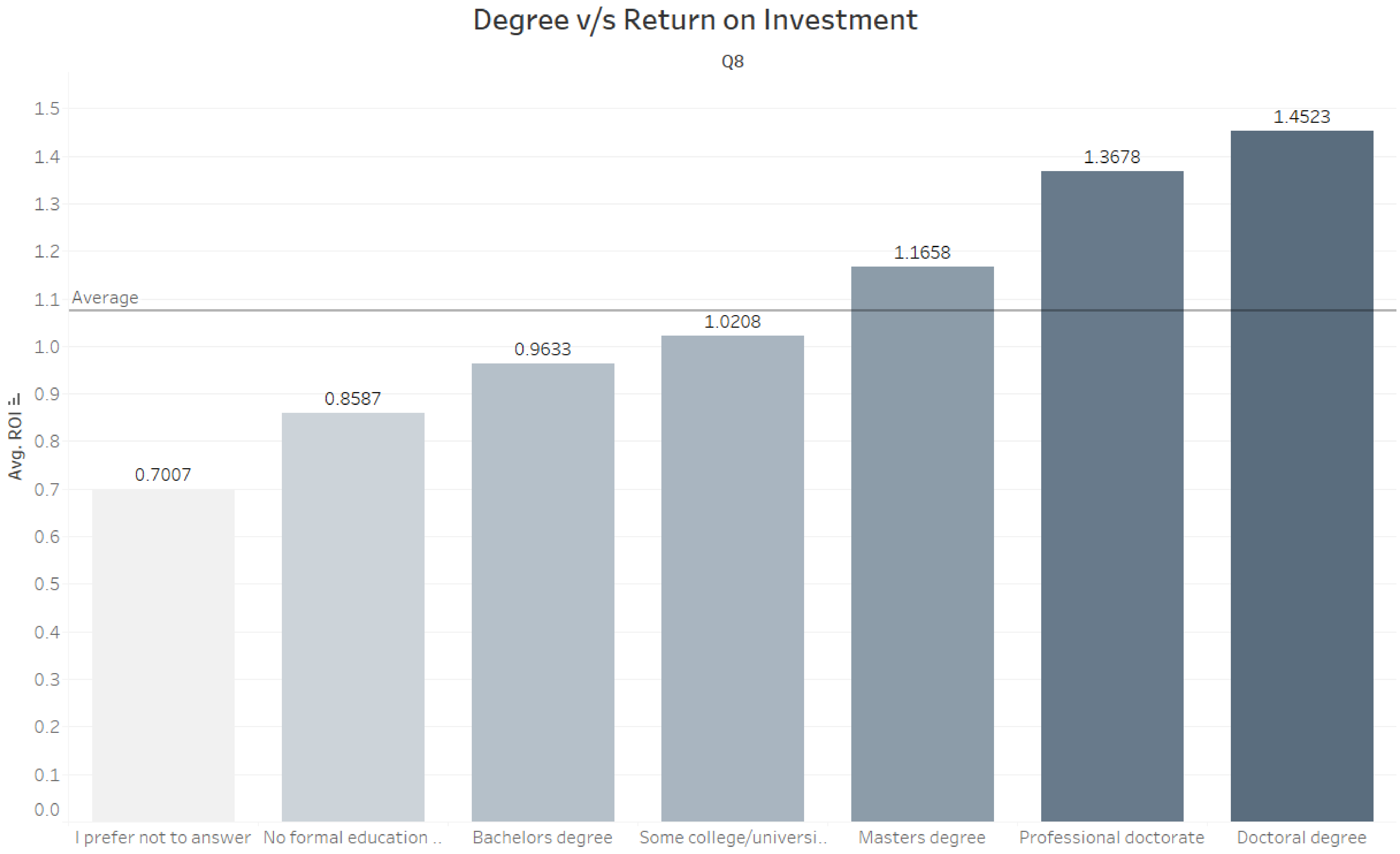
- Businesses make data-driven decisions
- Improve efficiency
- Competitive Edge

RETURN ON INVESTMENT

How much a person has invested in their education and are they getting return on their investment:

Return on Investment (ROI)

ROI = Salary/Investment



The ROI

The professionals who have pursued :

- **Doctoral Degree**
- **Professional Degree**
- **Master's Degree**

Have the return on investment higher than the average return on investment in the market.

THYNK ANALYTICS

PREDICTION MODEL PERFORMANCE

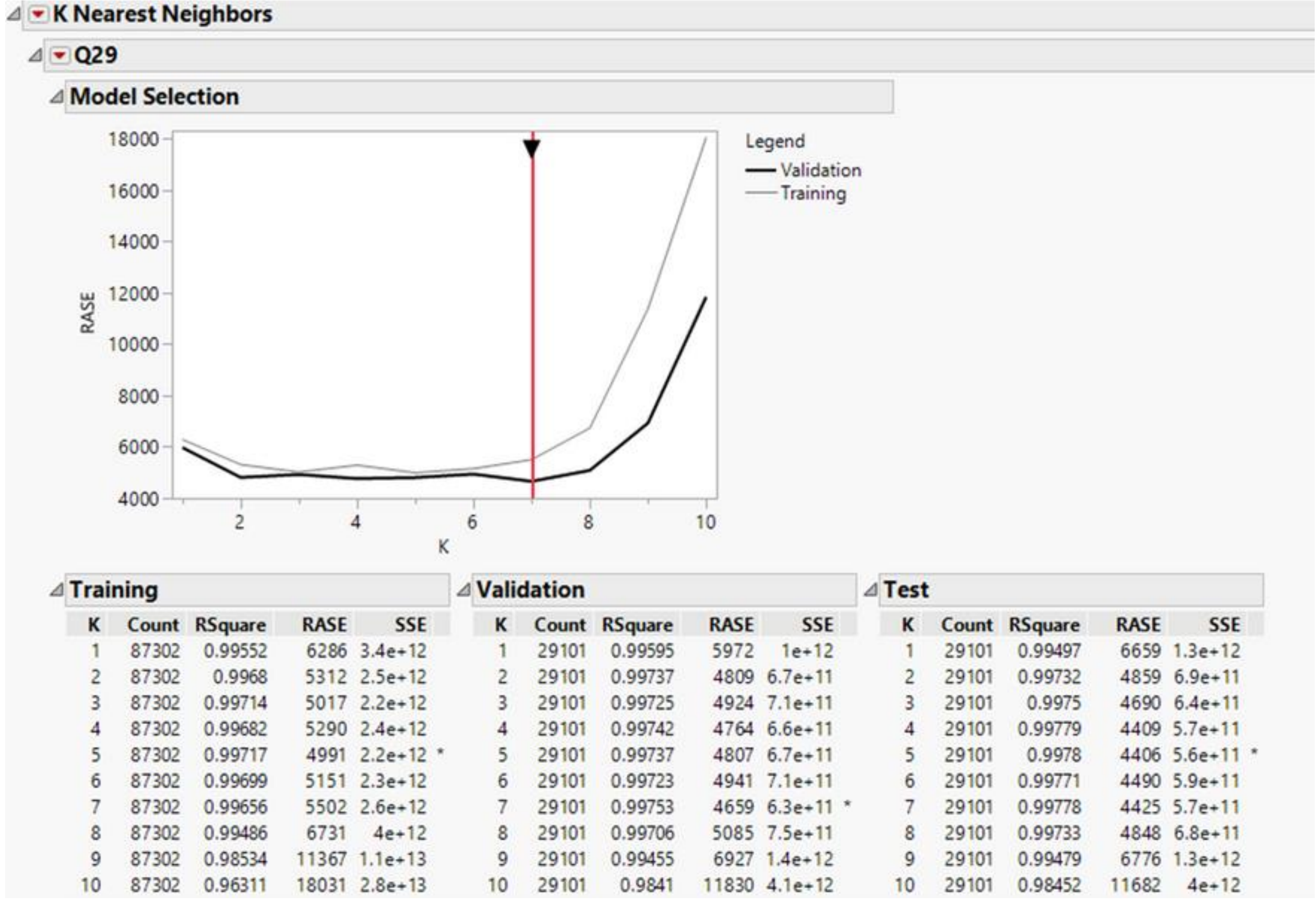
The model is based on kNN (K-Nearest Neighbor) Algorithm of supervised Learning

Best Model fit with tuning parameter: **k = 7**

Model Performance is analyzed with statistical measures :

r-square = 0.99 (on the test set)

RASE = 4425



K Nearest Neighbors

Q29

Model Selection

Training

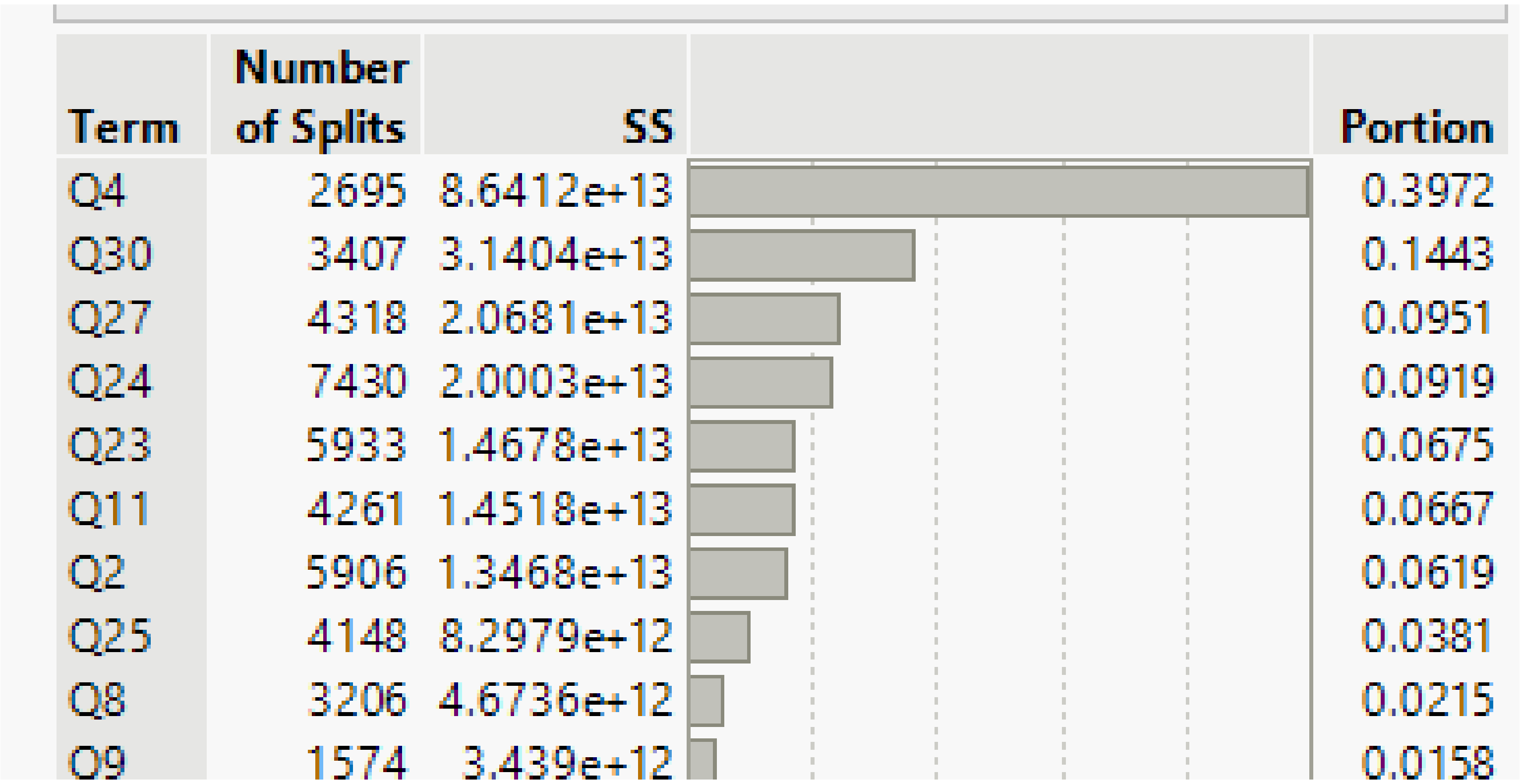
K	Count	RSquare	RASE	SSE
1	87302	0.99552	6286	3.4e+12
2	87302	0.9968	5312	2.5e+12
3	87302	0.99714	5017	2.2e+12
4	87302	0.99682	5290	2.4e+12
5	87302	0.99717	4991	2.2e+12 *
6	87302	0.99699	5151	2.3e+12
7	87302	0.99656	5502	2.6e+12
8	87302	0.99486	6731	4e+12
9	87302	0.98534	11367	1.1e+13
10	87302	0.96311	18031	2.8e+13

Validation

K	Count	RSquare	RASE	SSE
1	29101	0.99595	5972	1e+12
2	29101	0.99737	4809	6.7e+11
3	29101	0.99725	4924	7.1e+11
4	29101	0.99742	4764	6.6e+11
5	29101	0.99737	4807	6.7e+11
6	29101	0.99723	4941	7.1e+11
7	29101	0.99753	4659	6.3e+11 *
8	29101	0.99706	5085	7.5e+11
9	29101	0.99455	6927	1.4e+12
10	29101	0.9841	11830	4.1e+12

Test

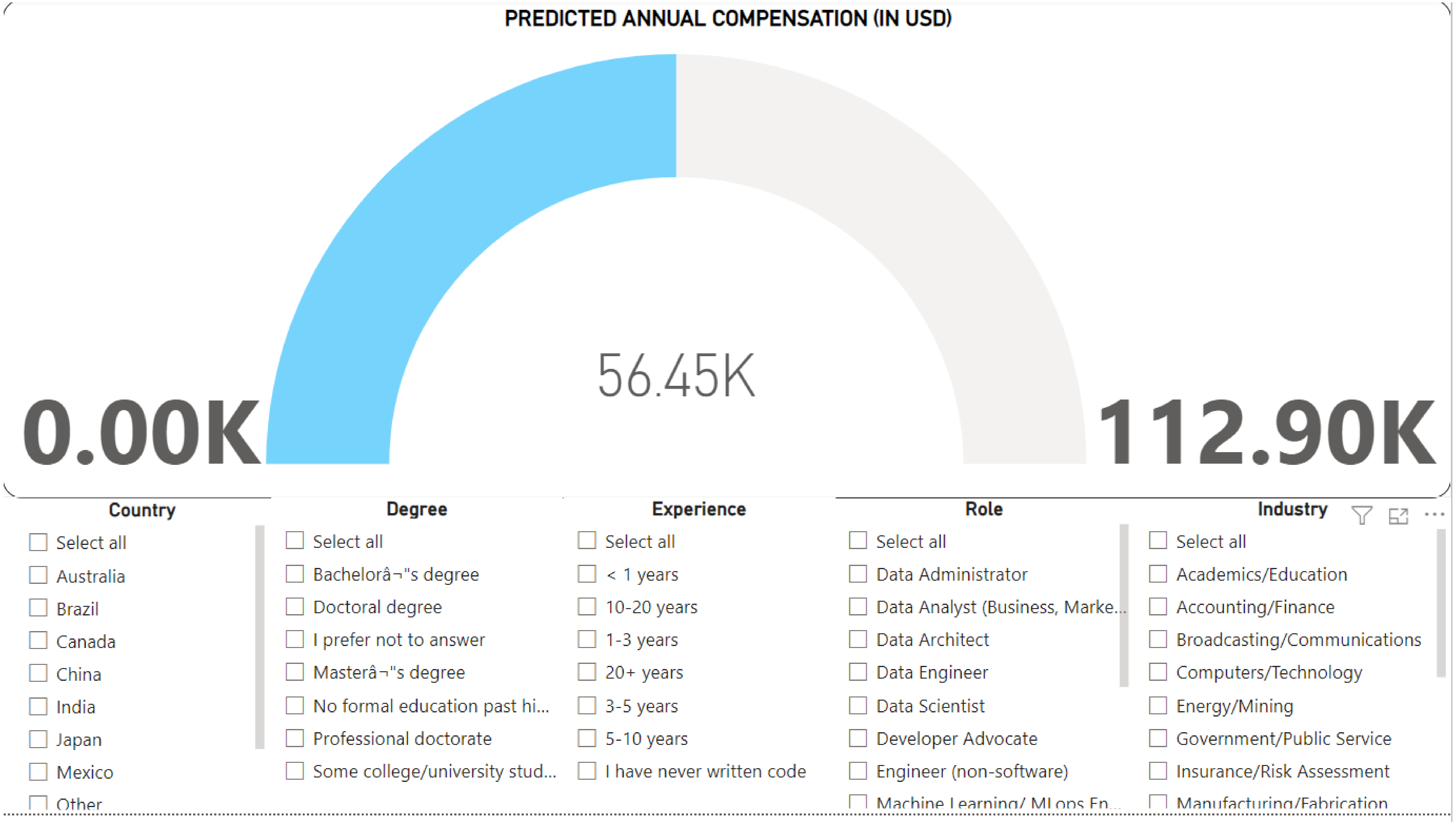
K	Count	RSquare	RASE	SSE
1	29101	0.99497	6659	1.3e+12
2	29101	0.99732	4859	6.9e+11
3	29101	0.9975	4690	6.4e+11
4	29101	0.99779	4409	5.7e+11
5	29101	0.9978	4406	5.6e+11 *
6	29101	0.99771	4490	5.9e+11
7	29101	0.99778	4425	5.7e+11
8	29101	0.99733	4848	6.8e+11
9	29101	0.99479	6776	1.3e+12
10	29101	0.98452	11682	4e+12



Term	Number of Splits	SS	Portion
Q4	2695	8.6412e+13	0.3972
Q30	3407	3.1404e+13	0.1443
Q27	4318	2.0681e+13	0.0951
Q24	7430	2.0003e+13	0.0919
Q23	5933	1.4678e+13	0.0675
Q11	4261	1.4518e+13	0.0667
Q2	5906	1.3468e+13	0.0619
Q25	4148	8.2979e+12	0.0381
Q8	3206	4.6736e+12	0.0215
Q9	1574	3.439e+12	0.0158

** Data only for current professionals

Income Prediction Tool



Country

- □ Select all
- □ Australia
- □ Brazil
- □ Canada
- □ China
- □ India
- □ Japan
- □ Mexico
- □ Other

Degree

- □ Select all
- □ Bachelorâ¬"s degree
- □ Doctoral degree
- □ I prefer not to answer
- □ Masterâ¬"s degree
- □ No formal education past hi...
- □ Professional doctorate
- □ Some college/university stud...

Experience

- □ Select all
- □ < 1 years
- □ 10-20 years
- □ 1-3 years
- □ 20+ years
- □ 3-5 years
- □ 5-10 years
- □ I have never written code

Role

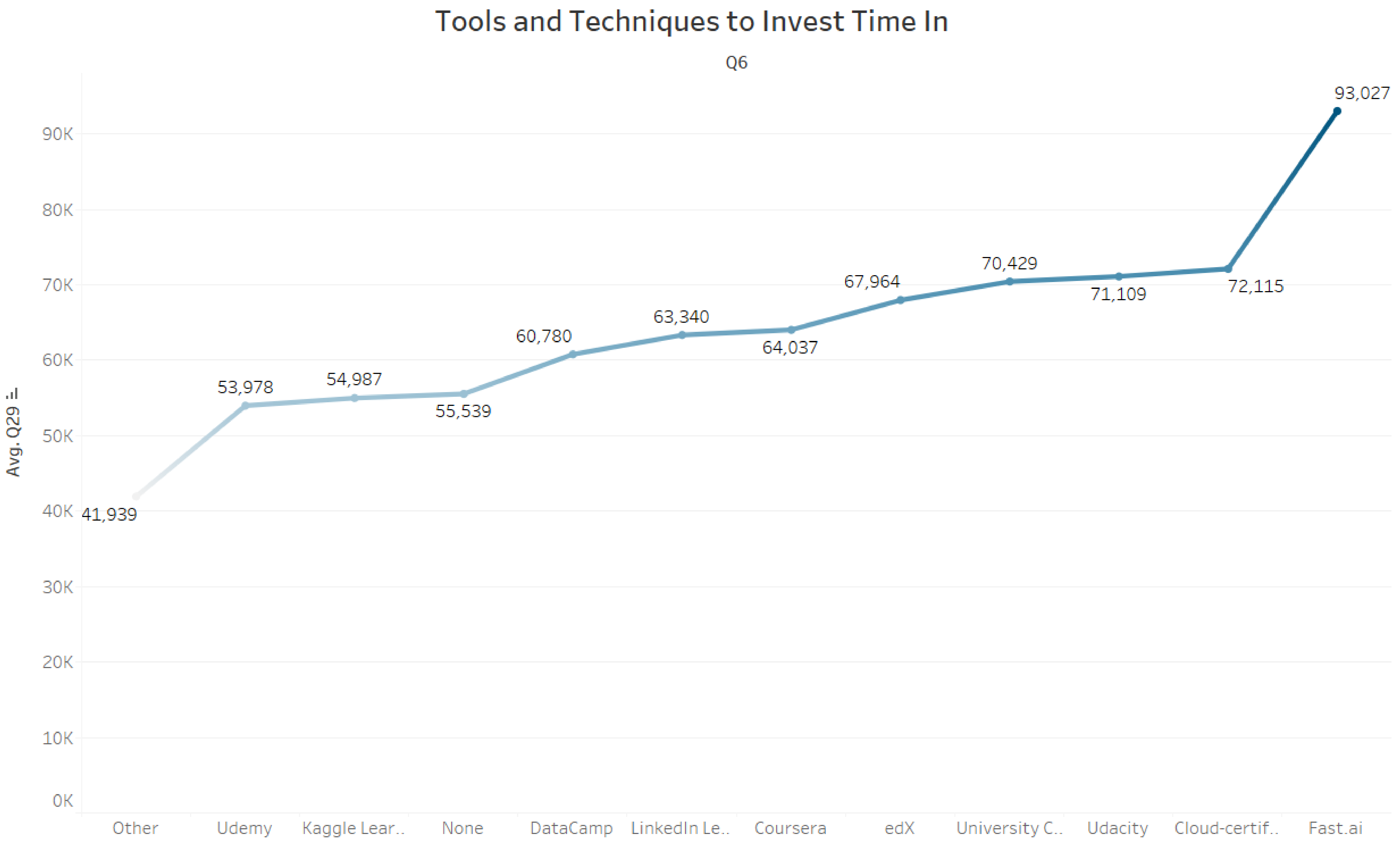
- □ Select all
- □ Data Administrator
- □ Data Analyst (Business, Marke...
- □ Data Architect
- □ Data Engineer
- □ Data Scientist
- □ Developer Advocate
- □ Engineer (non-software)
- □ Machine Learning/ MLops En...

Industry

- □ Select all
- □ Academics/Education
- □ Accounting/Finance
- □ Broadcasting/Communications
- □ Computers/Technology
- □ Energy/Mining
- □ Government/Public Service
- □ Insurance/Risk Assessment
- □ Manufacturing/Fabrication

Link:
https://app.powerbi.com/links/IJ67FqhCAP?ctid=80f23f4a-91a4-4566-8db1-3bcabb21d1cb&pbi_source=linkShare

EMERGING TOOLS TO INVEST TIME IN



The trend of average of Q29 for Q6. Color shows average of Q29. The marks are labeled by average of Q29. Details are shown for Q6. The view is filtered on Q6, which excludes NA and Null.

The comparsion between the emerging online learning platforms using the average salaries in the data science industry:

Fast.ai:

Eases building & training advanced ML models with pre-built models & tools for data preprocessing, visualization & interpretation. Built on PyTorch & TensorFlow.

RECOMMENDATIONS:

To design the curriculum holistically to integrate the use of the emerging technologies and tools.

Higher Education

Data science higher education offers valuable skills and knowledge with a high return on investment, making it a must for students seeking better job opportunities.

Technologies

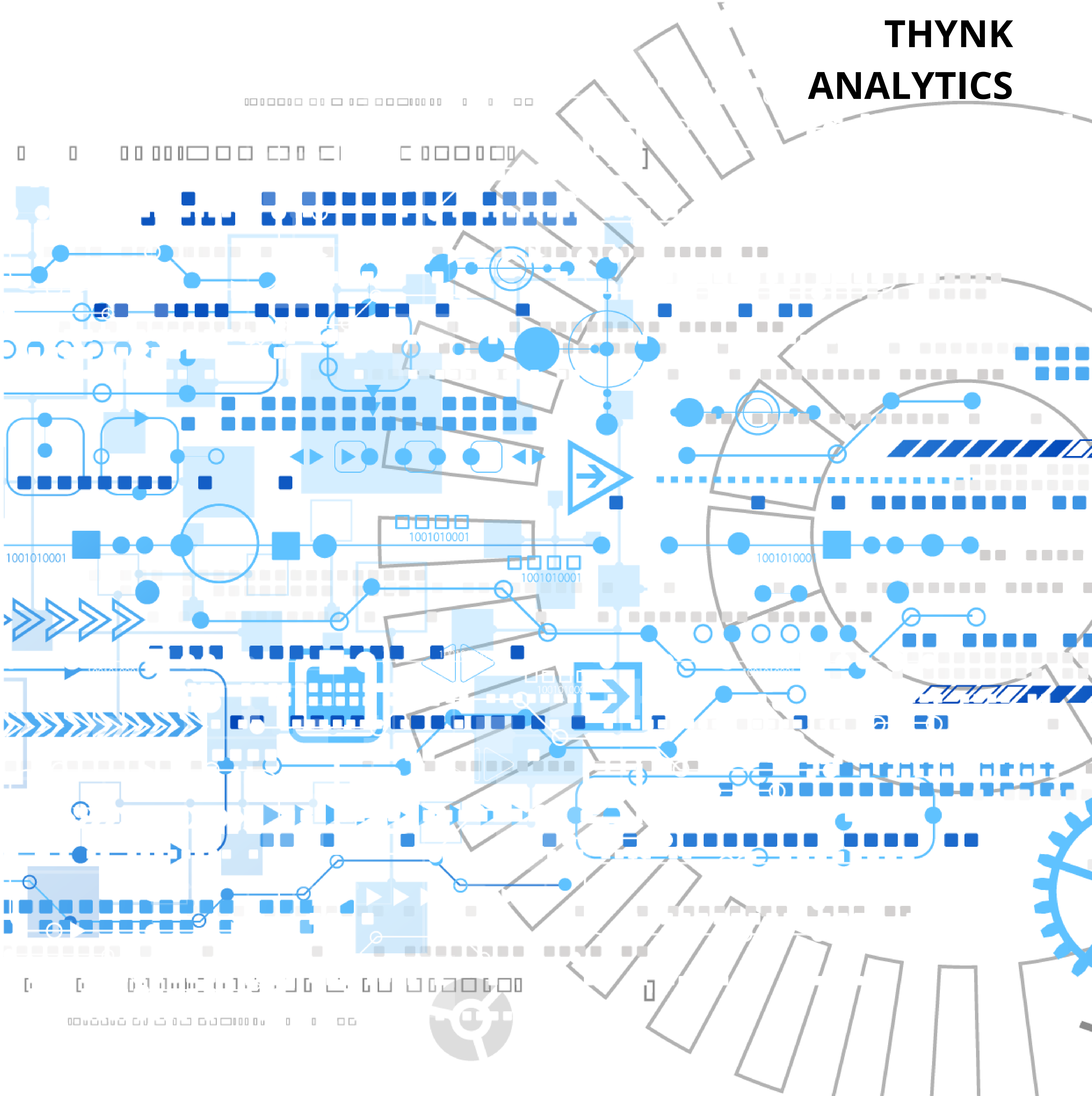
Focusing on the most used technologies in the industry, i.e Python, SQL, Jupyter, Collab, etc.

Industry Specific Knowledge

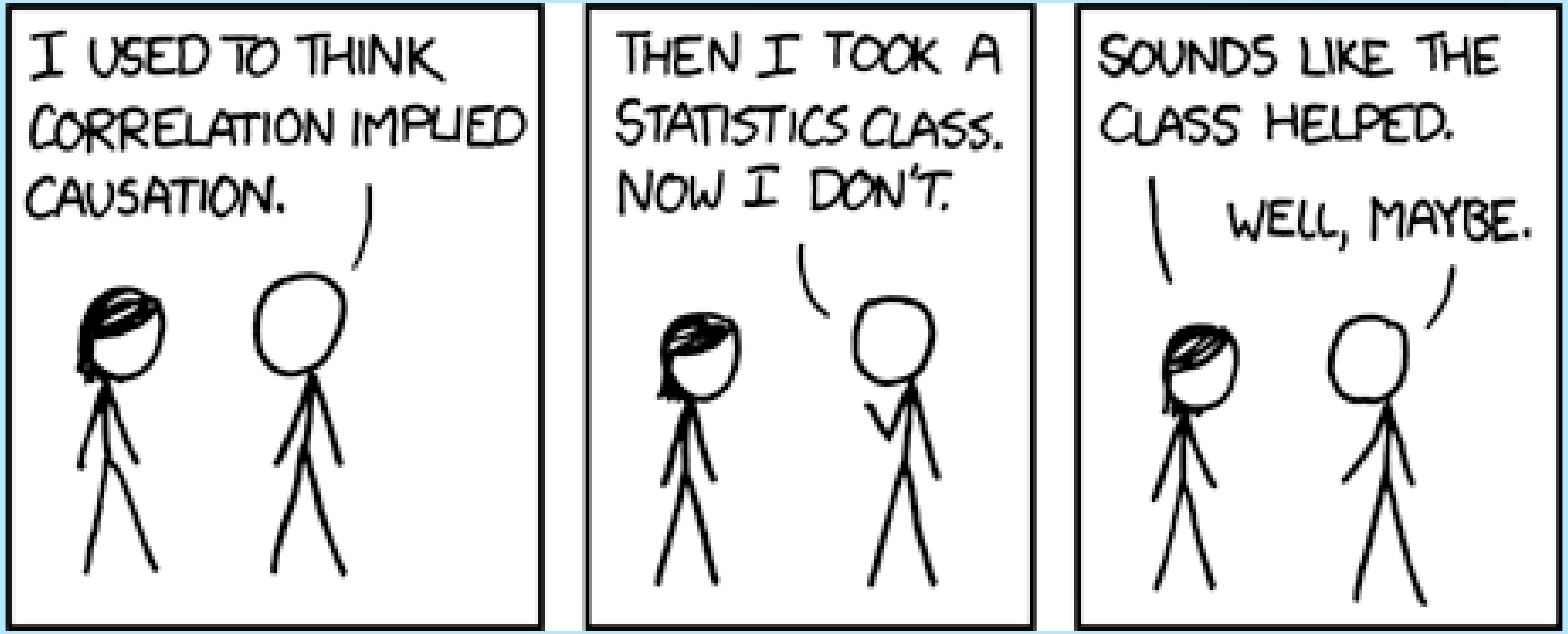
. Students should be encouraged to focus on a major area of interest like education, computers/IT, Banking, etc. and build industry-specific knowledge.

LIMITATIONS

- The dataset is biased as most of the responses are from India.
- The model has overfitted the data.



Thank You


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.