MIE1624 – Introduction to Data Science & Analytics

Group 10 – Project Presentation

MUJTABA ALI KHAN

SAHAR CHOUKIR

MOHAMMAD BA SALEEM

EKELE NNOROM

Outline

- Market Outlook
- Data Collection
- MIE1624 Course Outline
- MSDA and MBAI curriculum
- EdTech effort Startup

Job Market Glance : Canada

DATA SCIENCE FACTS CANADA

$1.1 → **$1.8**
Billion Annual Revenue 2019 — Billion Annual Revenue 2020
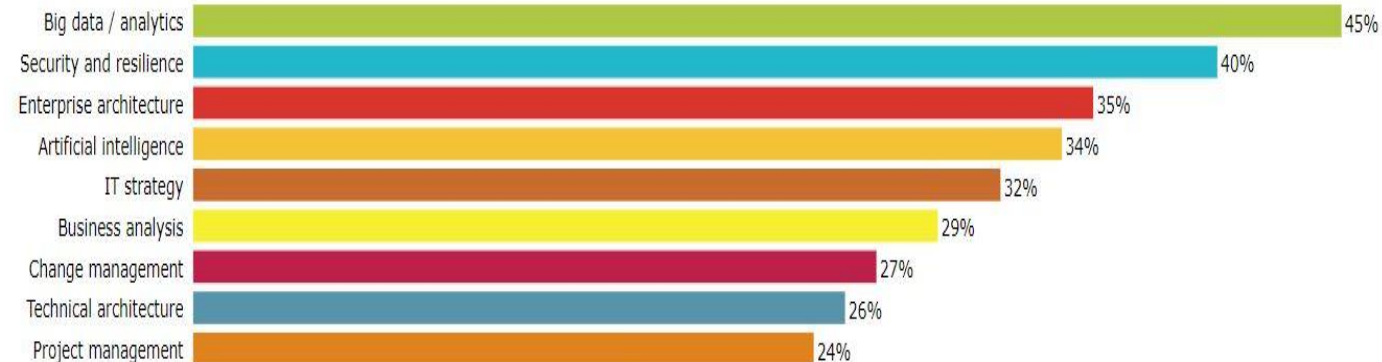
434%
Job Growth Rate

5th
Ranked job in Canada

82%
Job Satisfaction

Which functions do you feel suffer from a skills shortage?
Select all that apply

Function	Percentage
Big data / analytics	45%
Security and resilience	40%
Enterprise architecture	35%
Artificial intelligence	34%
IT strategy	32%
Business analysis	29%
Change management	27%
Technical architecture	26%
Project management	24%

Source: KPMG: Harvey Nash

Job Market Glance : Supply & Demand

Over 2/3 believe demand for talent will outpace the supply of data scientists

OVER THE NEXT FIVE YEARS, DEMAND FOR DATA SCIENTISTS WILL:

Be significantly less than the talent available **1%**

Be less than the talent available **5%**

Be met by the available talent **31%**

31% Significantly outpace the supply of talent

32% Somewhat outpace the supply of talent

Lack of training and resources are the biggest obstacle to data science in organizations

THE BIGGEST OBSTACLE TO DATA SCIENCE ADOPTION IN OUR ORGANIZATION IS:

| **32%** Employees don't have the right skills or training | **32%** Lack budget or resources | **14%** The wrong organizational structure | **10%** Lack of necessary tools or technology | **9%** Insufficient executive support | **3%** Other |

DATA SCIENTIST CAREER TRENDS

EDUCATION

92% OF DATA SCIENTISTS HAVE AN ADVANCED DEGREE

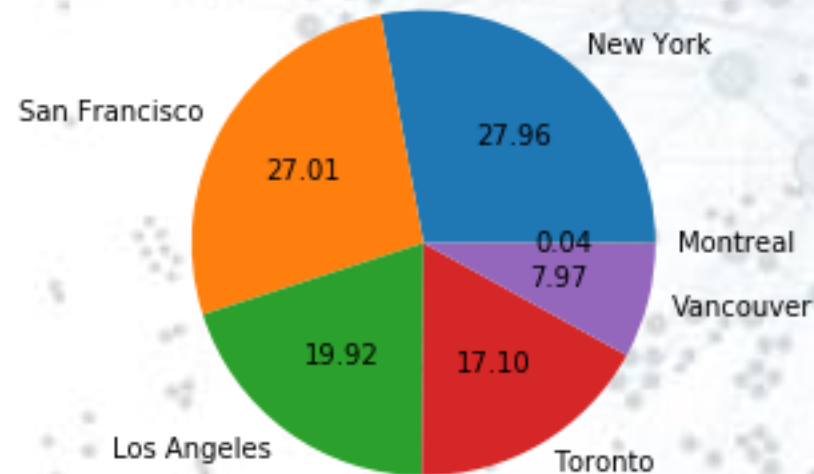**48%** have a Ph.D.

44% have a master's degree

8% have a bachelor's degree

What makes a data scientist ? Methodology

1. Data Science Survey-2018 : Kaggle

2. Web Scraping of more than 30000 Job Postings In Indeed

Distribution of the cities of web scrapped jobs

New York 27.96

Montreal 0.04

Vancouver 7.97

Toronto 17.10

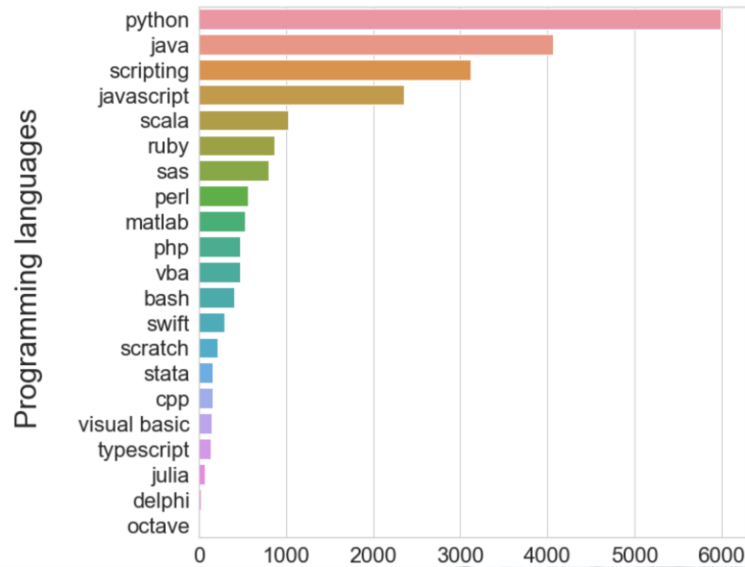Los Angeles 19.92

San Francisco 27.01

1. Exploratory analysis : frequency of skills

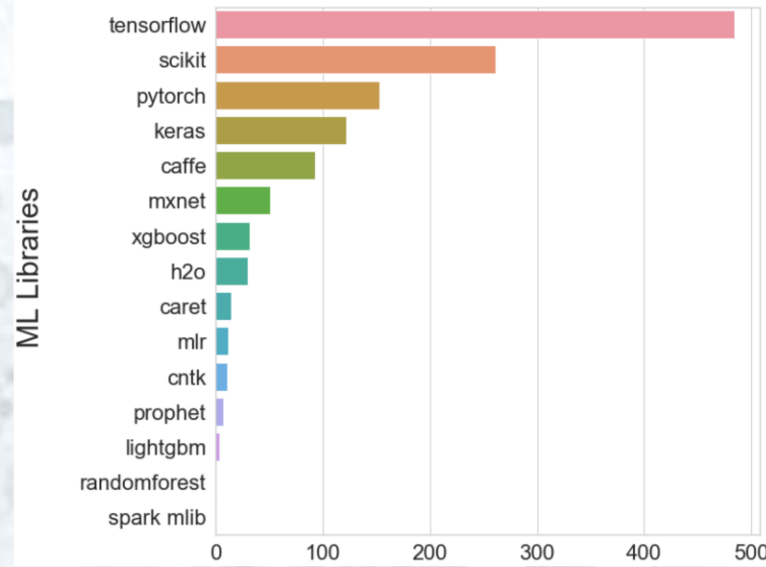2. Build a Machine Learning Model to identify features that lead to higher salaries

3. Topic Modeling: clustering of skills

Indeed Web Scraping
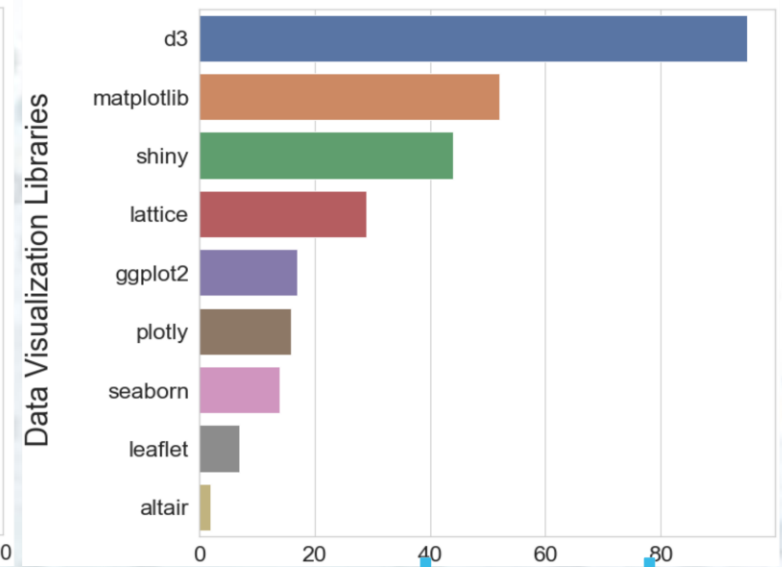

Indeed Web Scraping: Top Programming Languages

Indeed Web Scraping : ML Libraries

Indeed Web Scraping : Data Visualization Libraries

Kaggle Survey 2018

Kaggle Survey: Top Programming Languages
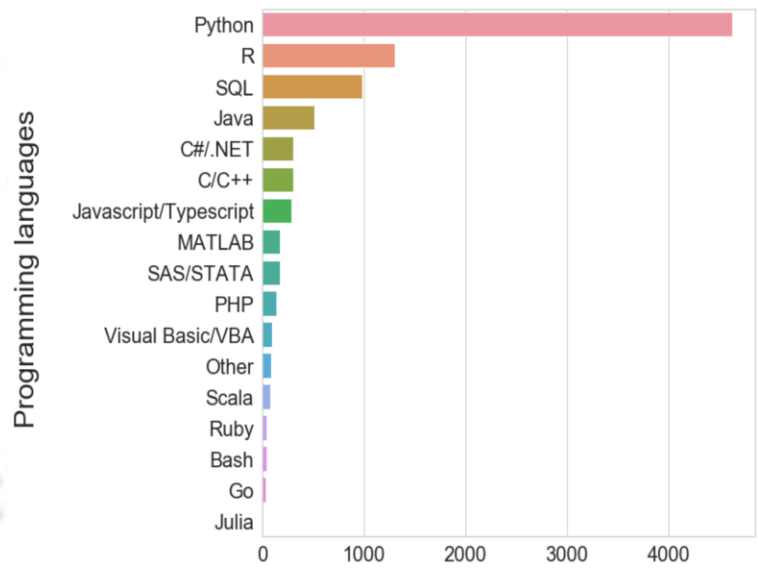
Kaggle Survey: Top ML Libraries
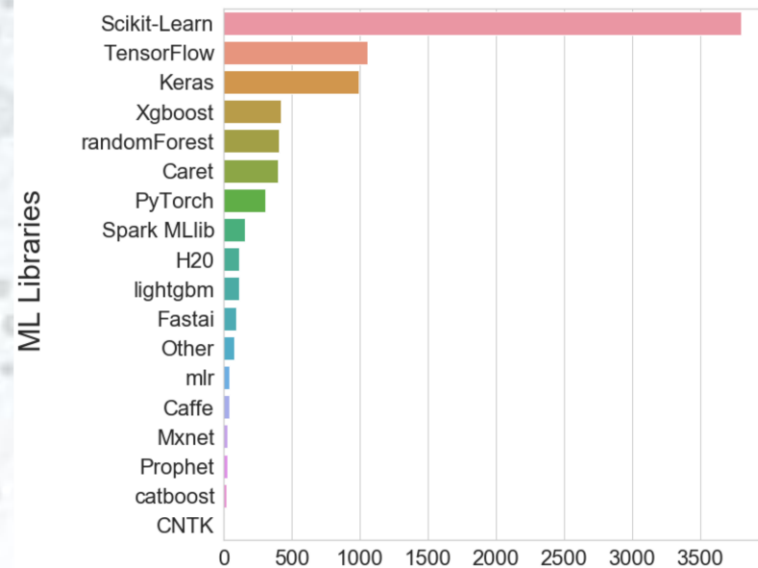
Kaggle Survey: Data Visualization Libraries
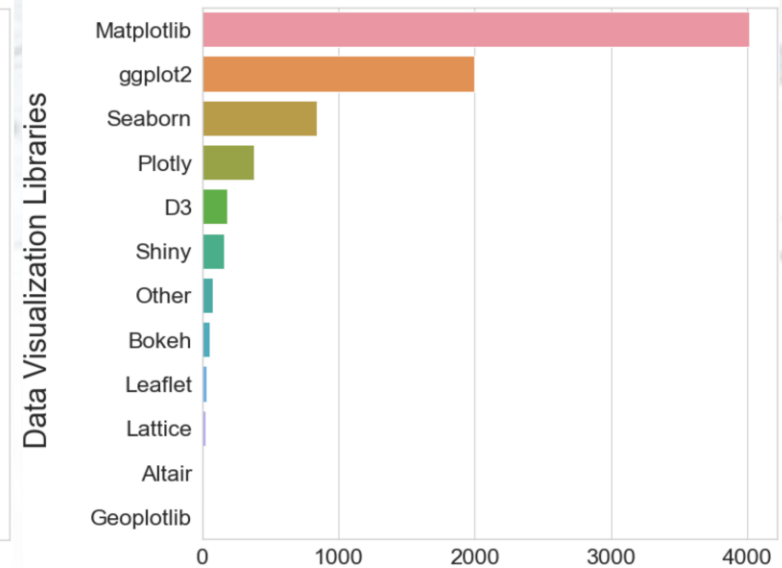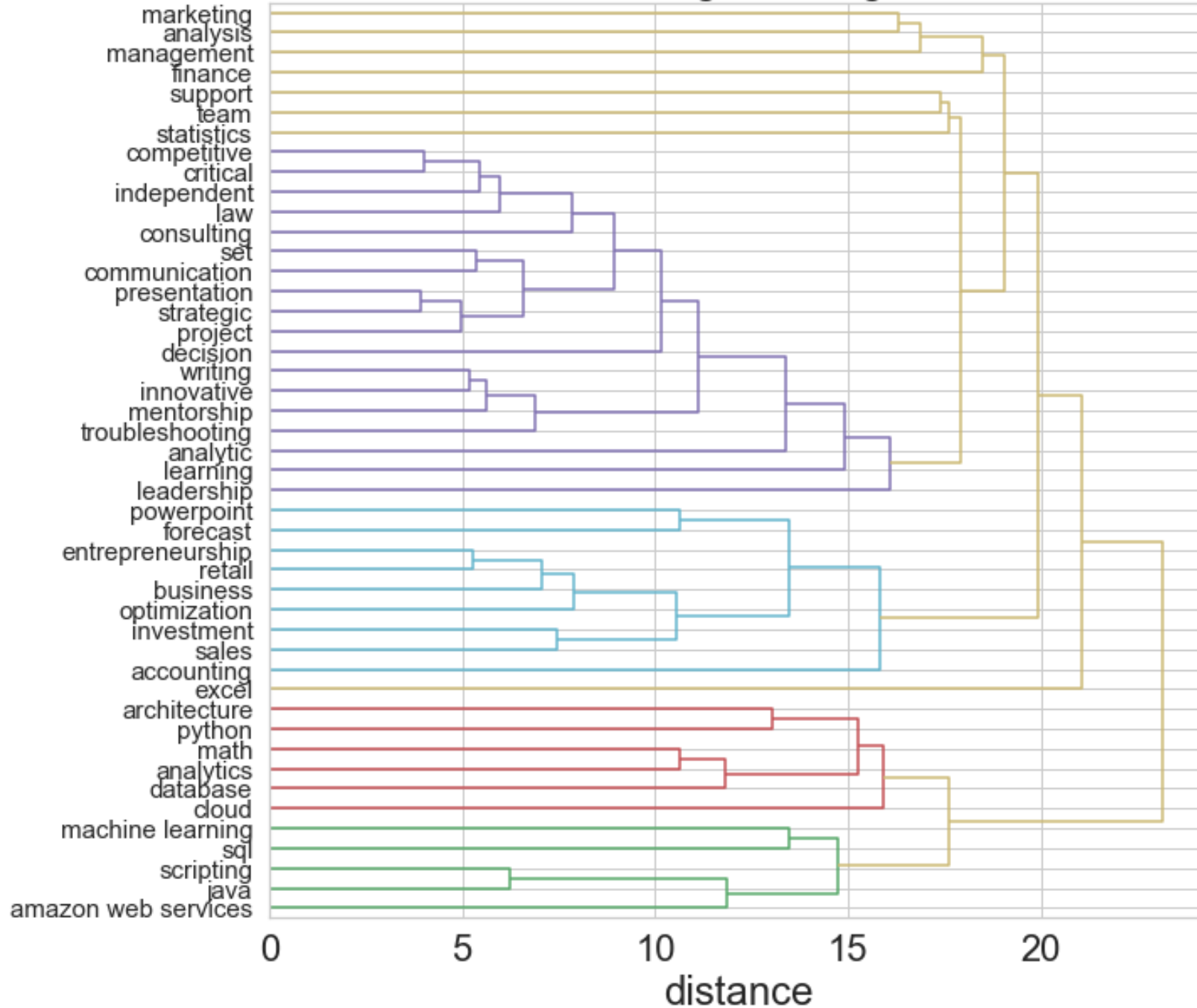
Hierarchical Clustering of Web Scraped Indeed Jobs as a mean to develop modules and courses
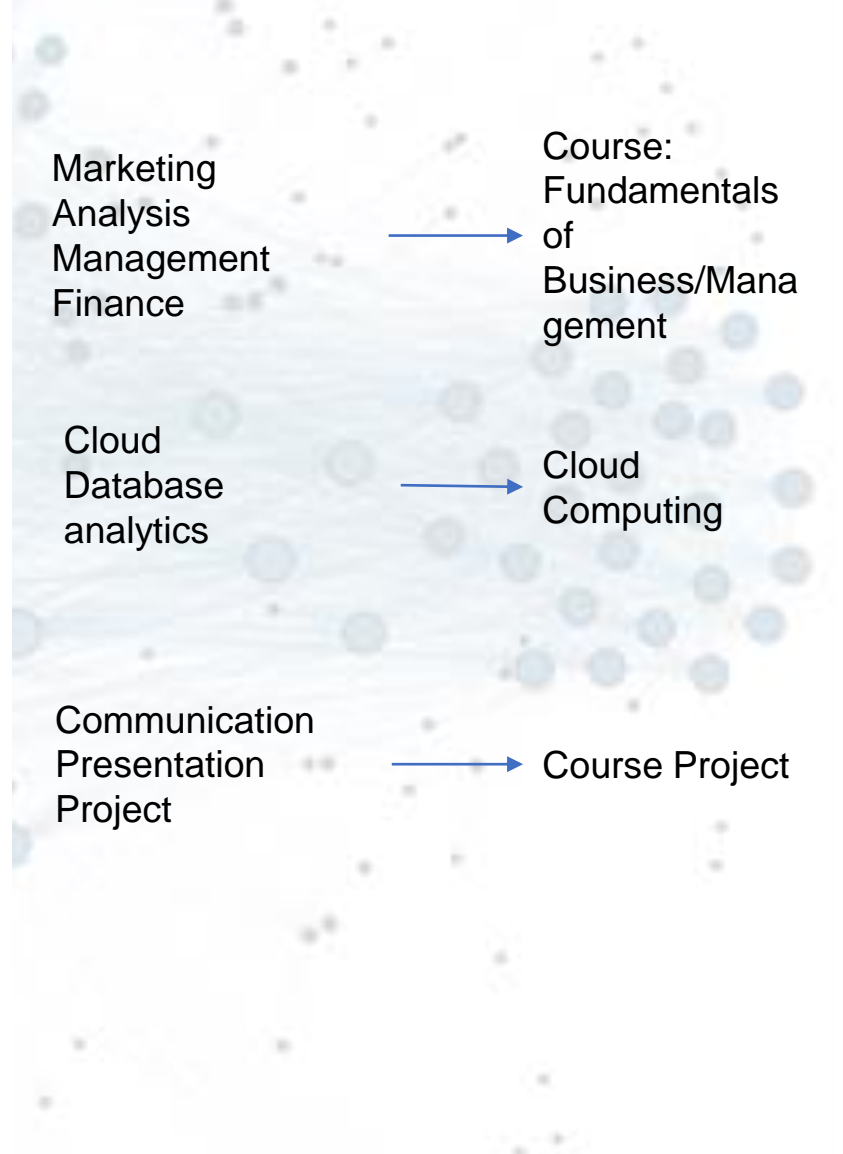
Data Inferences

- Soft Skills are extremely vital
- Many Data Scientist Jobs require business knowledge in particular in project management, consulting and accounting
- Machine learning and cloud computing skills form the backbone of high in –demand technical toolkits
- Many programs are not fully compliant with current market needs
- Project-based learning is the key to develop both required technical and soft skills

MIE1624 Course Curriculum Design

What is Data Science & Analytics
- Overview
- Kaggle
- Github
- Application of DS
- Stack Overflow.com

Ass. 1 Exploratory Analysis

Programming Languages
- Python
- R
- Jupyter Notebook
- Databases & SQL
- Installation
- TA Data Vizs + Python libraries

Introduction to Statistics
- Probability Distributions
- Confidence Intervals
- Tests & Test Scores RMSE, R^2
- Ass. 2 Sentiment Analysis

Linear Algebra
- Gaussian Elimination
- Regression Techniques
- TA Linear Regression Solver

Machine Learning – Supervised ML
- Logistic & Linear Regression
- Classification & Regression
- Train & Test, Cross Validation
- Hyper-parameter Tuning
- Ass. 3 Anomaly Detection

Machine Learning – Unsupervised ML
- Clustering – K Means, DB Scan, Hierarchical Clustering
- Dimensionality Reduction – PCA, RFE
- TA Text Analytics

Neural Networks & Deep Learning
- Gradient Descent
- Back Propagation
- Tensorflow, Keras
- Ass. 4 Face Recognition

Convolutional & Recurrent Neural Networks
- GPU CUDA, ILSVRC
- LSTM, GRU
- Tensorflow, Keras
- TA Parallel Programming

Reinforcement Learning
- Markov Decision Processes
- Value Function
- Q Function, Q table
- DQN

Optimization & Simulation
- Optimization & Simulation – Linear & Nonlinear
- Monte Carlo Simulation – Risk Analysis

+ Course Project

Kaggle Survey : Most used programming languages

Kaggle Survey: Top Programming Languages

Programming languages: Python, R, SQL — Frequency Count (0, 1000, 2000, 3000, 4000)
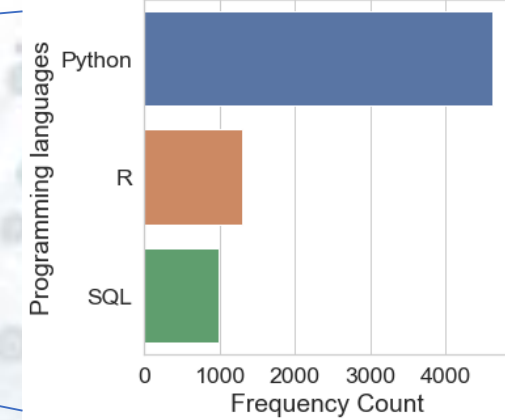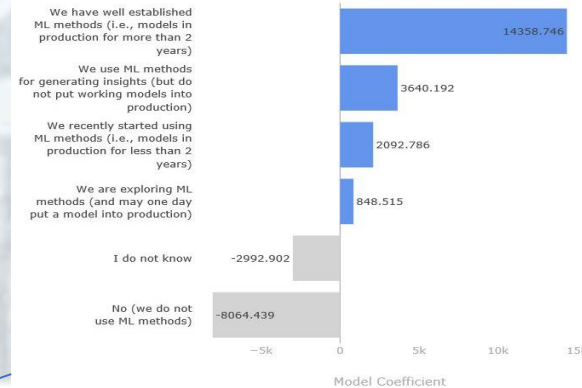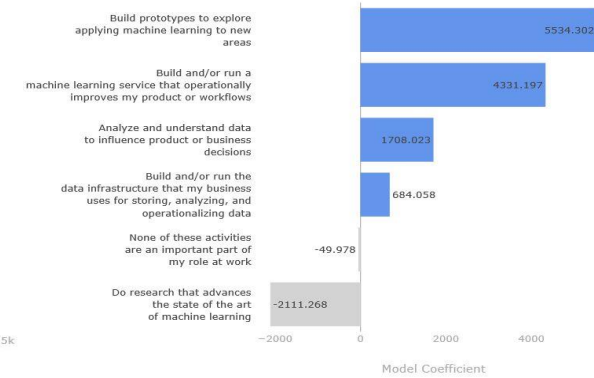
Kaggle Survey : ML correlates with high compensation

Q10. Does your current employer incorporate machine learning methods into their business?
- We have well established ML methods (i.e., models in production for more than 2 years): 14358.746
- We use ML methods for generating insights (but do not put working models into production): 3640.192
- We recently started using ML methods (i.e., models in production for less than 2 years): 2092.786
- We are exploring ML methods (and may one day put a model into production): 848.515
- I do not know: -2992.902
- No (we do not use ML methods): -8064.439

Model Coefficient

Q11. Select any activities that make up an important part of your role at work
- Build prototypes to explore applying machine learning to new areas: 5534.302
- Build and/or run a machine learning service that operationally improves my product or workflows: 4331.197
- Analyze and understand data to influence product or business decisions: 1708.023
- Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data: 684.058
- None of these activities are an important part of my role at work: -49.978
- Do research that advances the state of the art of machine learning: -2111.268

Model Coefficient

Indeed Web Scrapping : Most frequent skills reveals a high demand for soft skills : Communication, Team and Presentation

(word cloud: communication, support, analytic, management, cloud, competitive, critical, innovative, marketing, law, accounting, sql, analysis, learning, decision, project, investment, python, excel, strategic, machine, statistics, writing, leadership, sales, consulting, team, finance, architecture, presentation, database, set)

Masters Curricula

Core Courses | **Professional Experience** | **Elective Courses**

MDSA

- Intro to Data Science → Python | R
- Mathematics for Data Science → Statistics | Linear Algebra | Linear & Logistic Regression
- Data Science & Databases → SQL, NoSQL, PostgreSQL, Oracle
- Machine Learning → Supervised ML | Unsupervised ML | ML Frameworks Tensorflow etc
- Cloud Computing → Amazon Web Services | Microsoft Azure | IBM
- Big Data Analytics → Hadoop, Map Reduce, Hive, Pig | Spark and PySpark
- Ethics& Communication in Data Science → Topic Presentation | Best Ethical Practice of DS | Interpretation Techniques & Skills
- Professional Experience → Summer Internship or Research Project
- Bayesian Statistics & Machine → Uncertainty Modelling | Bayesian Inferences
- NLP → Text Processing Fundamentals | Advanced Techniques
- Reinforcement Learning → Markov Decision Processes | DQN | Q Function and Dynamic Programming
- Advanced Data Visualization → Tableau | Visualization for Multivariate, Temporal, Text-Based, Geospatial, Hierarchical Data

MBAI

- Intro to Data Science → Python | R
- Mathematics for Data Science → Statistics | Linear Algebra | Non-Linear Algebra
- Data Science & Databases → SQL, NoSQL, PostgreSQL, Oracle
- Fundamentals of Management → Management Theory | Strategic & Global Management | Influence & Leadership
- Finance & Accounting → Financial Reporting & Analysis | Audit Analytics
- Optimization → Unconstrained Nonlinear Optimization | Quadratic Programs (e.g.: SVM)
- Ethics& Communication in Data Science → Topic Presentation | Interpretation Techniques & Skills
- Stochastic Modelling & Time Series Analysis → Stochastic Processes | ARIMA Models
- Data Visualization → D3 | Marks and Channels | Interaction Techniques
- Professional Experience → Summer Internship OR Research Project
- Elective 1 (e.g.: Big Data in Finance) → Classical Econometrics, Portfolio Management | Data Mining & ML in Finance
- Elective 2 (e.g.: Supply Chain Analytics) → Exponential Smoothening Models | Data Driven Inventory Optimization
- Elective 3 (e.g.: Marketing Analytics) → Text Analytics | Segmentation

DataTO

Market Analysis

Canada

- Scotiabank — 13.12
- Hudson's — 17.77
- DICK'S Sporting Goods — 5.92
- Providence Health & Services — 5.92
- Deloitte — 6.77
- Oregon Health & Science University
- Loblaw — 8.60
- Intel — 9.87
- BMO — 12.13
- NIKE — 13.12

USA

- Amazon — 14.20
- Uber — 15.43
- Accenture — 6.56
- Kaiser Permanente — 6.92
- Citi — 6.92
- Salesforce — 8.44
- Northrop Grumman — 9.59
- Goldman Sachs — 9.59
- Deloitte — 10.67
- Google — 11.68

Jobs & Skills

Boot Camp

- ❑ Intro Data Science
- ❑ Mathematics for DS
- ❑ Intermediate Algorithms
- ❑ Data Pre-Processing
- ❑ Machine Learning
- ❑ Visualization Techniques
- ❑ Web & Cloud Computing
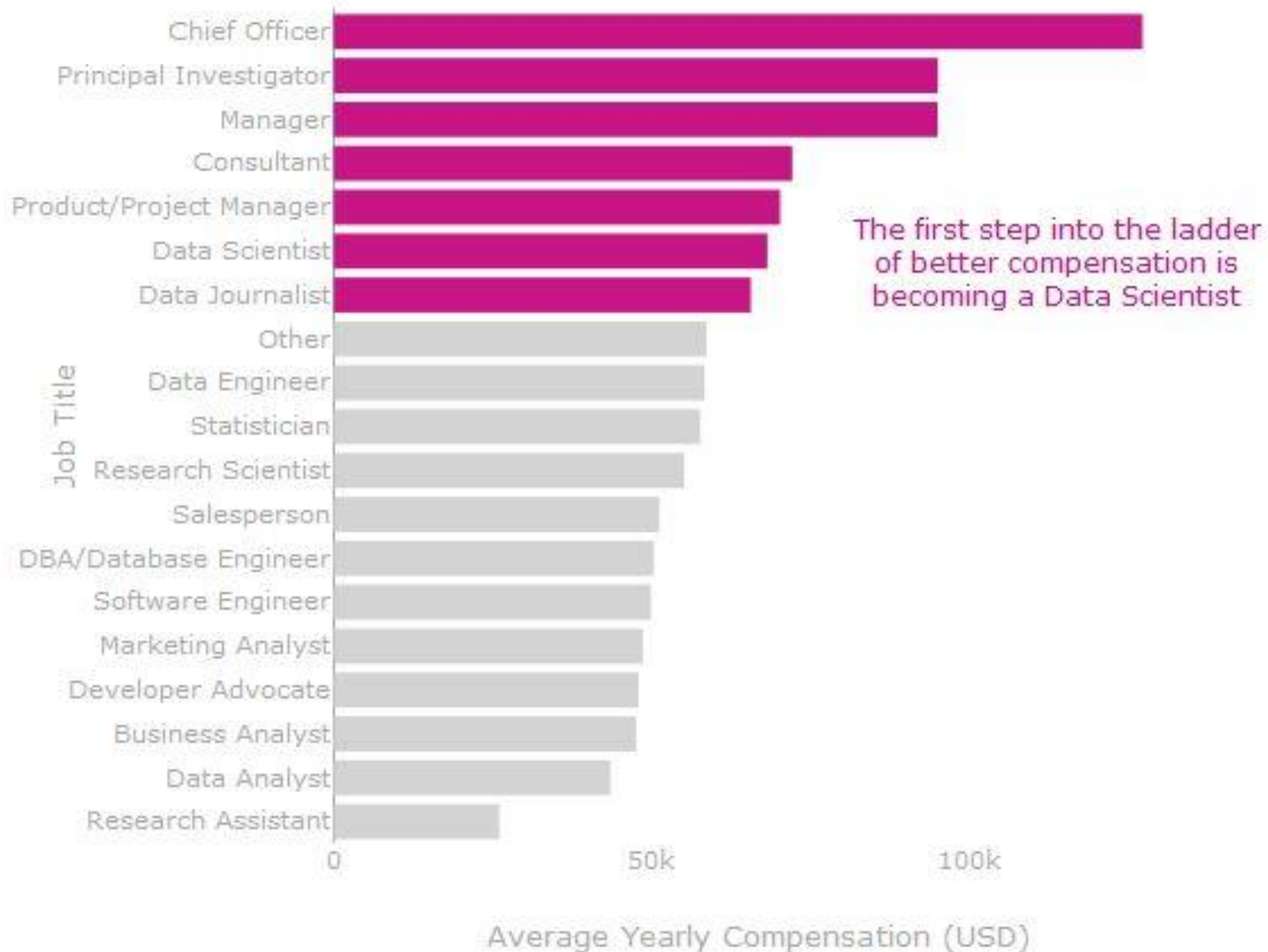- ❑ Neural Networks
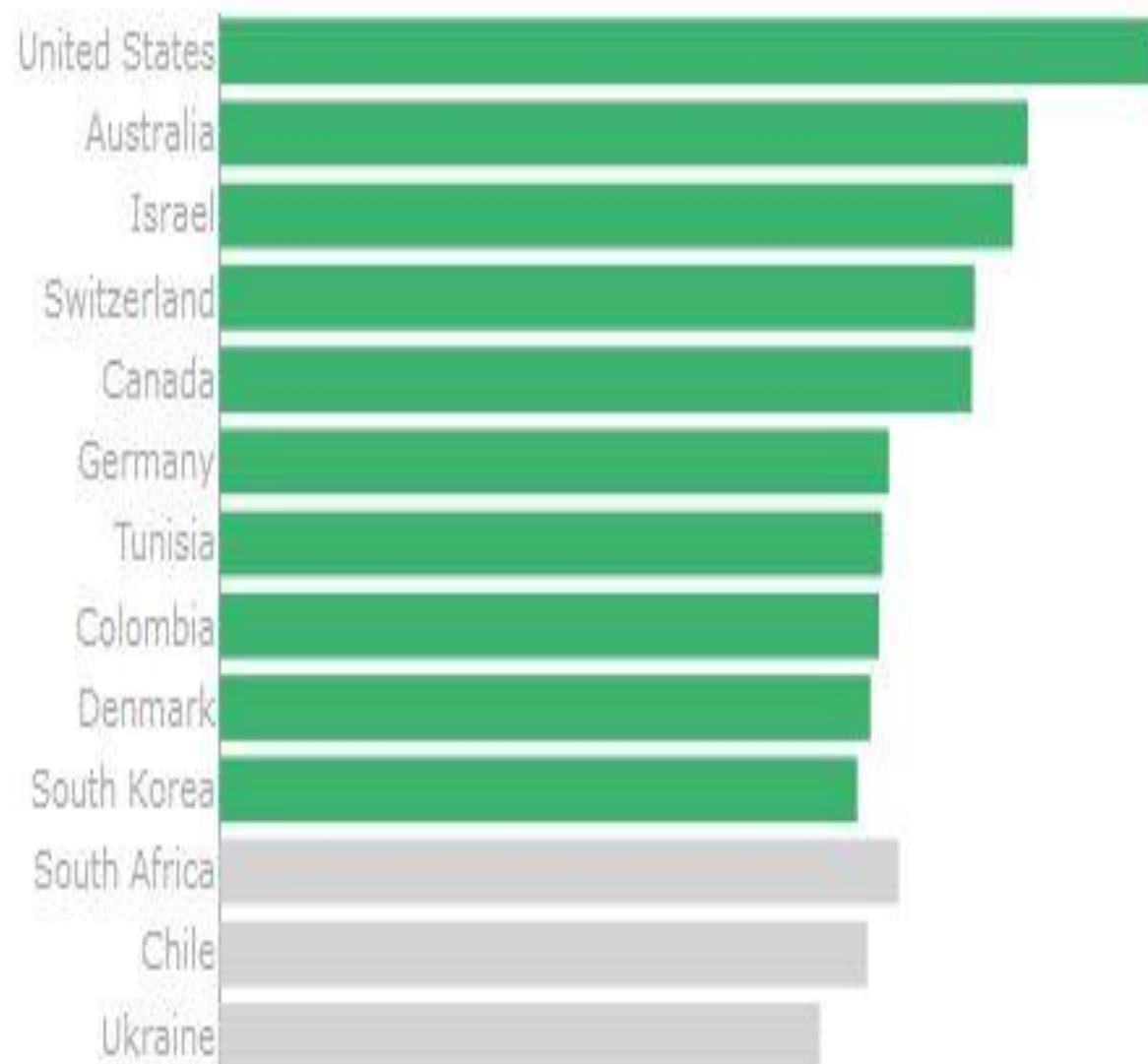- ❑ Ethics & Communication

RESUME

Thank You

Any Questions?

Back-up Slides

# Average Compensation per Role | Top 7 in Color
without students

The first step into the ladder of better compensation is becoming a Data Scientist

Job Title (top to bottom):
- Chief Officer
- Principal Investigator
- Manager
- Consultant
- Product/Project Manager
- Data Scientist
- Data Journalist
- Other
- Data Engineer
- Statistician
- Research Scientist
- Salesperson
- DBA/Database Engineer
- Software Engineer
- Marketing Analyst
- Developer Advocate
- Business Analyst
- Data Analyst
- Research Assistant

X-axis: 0, 50k, 100k

Average Yearly Compensation (USD)

Normalized Average Compensation per Country - Top 10 in Color
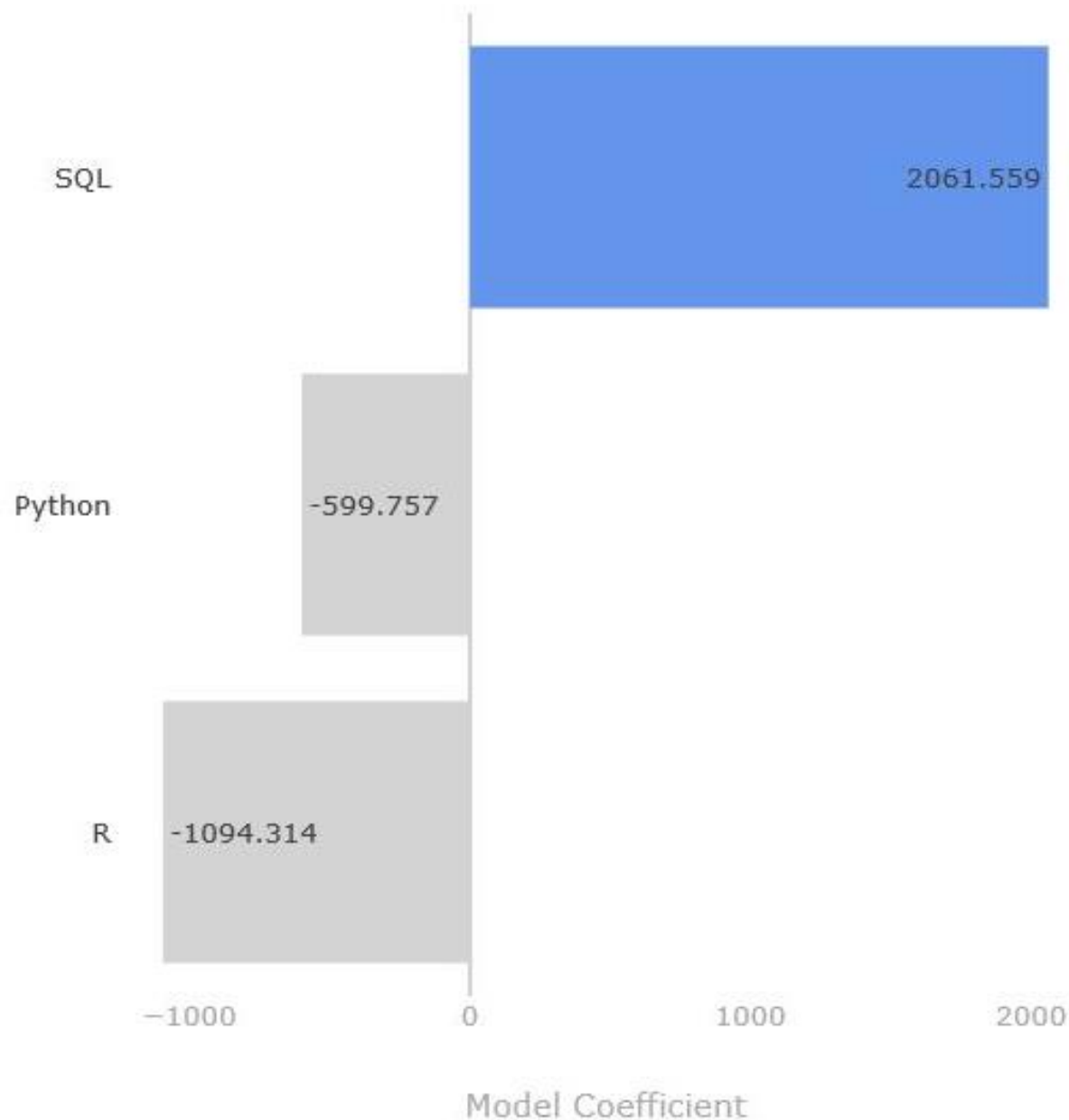## *without students*

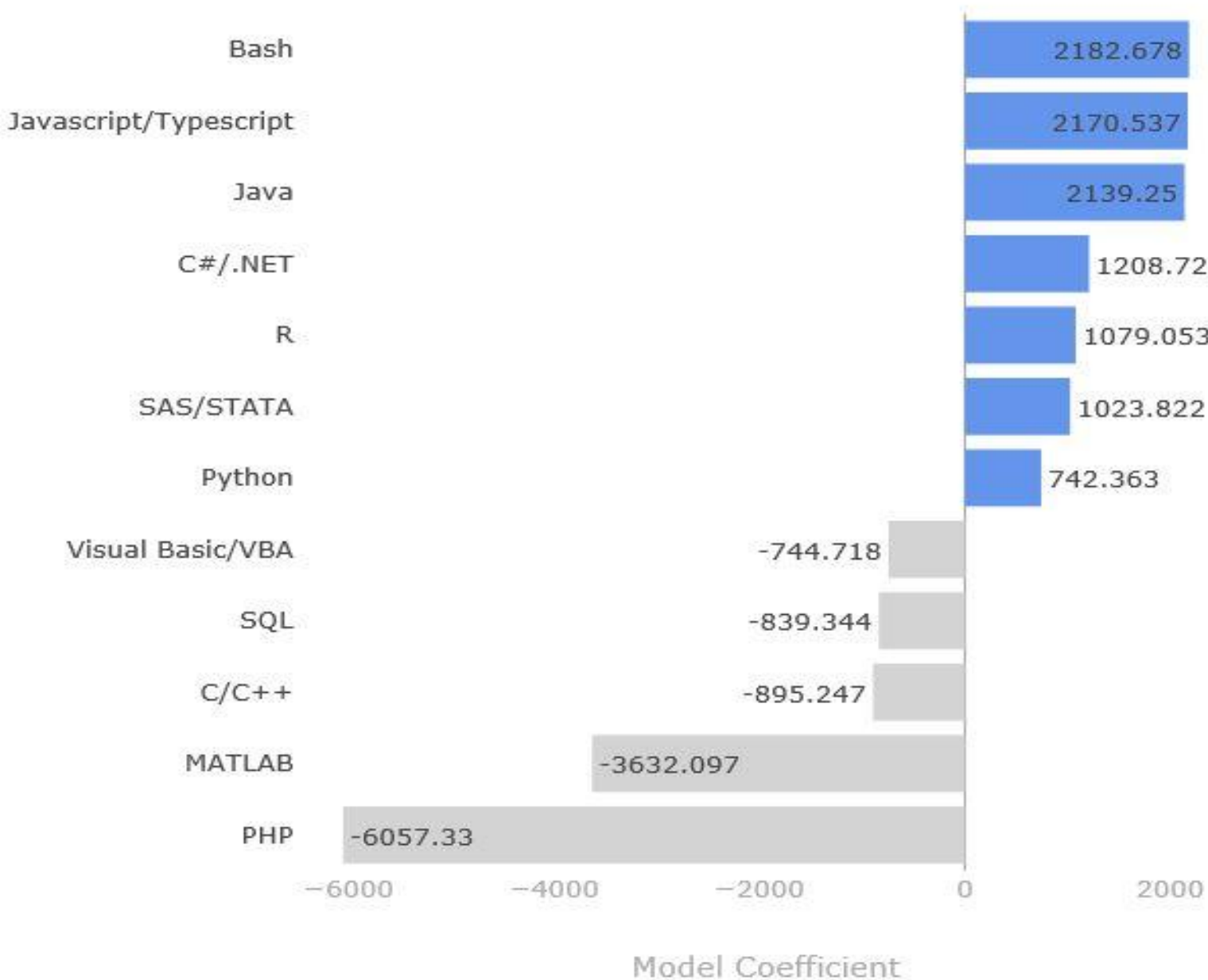# Impact of Formal Education on Compenstaion
without students
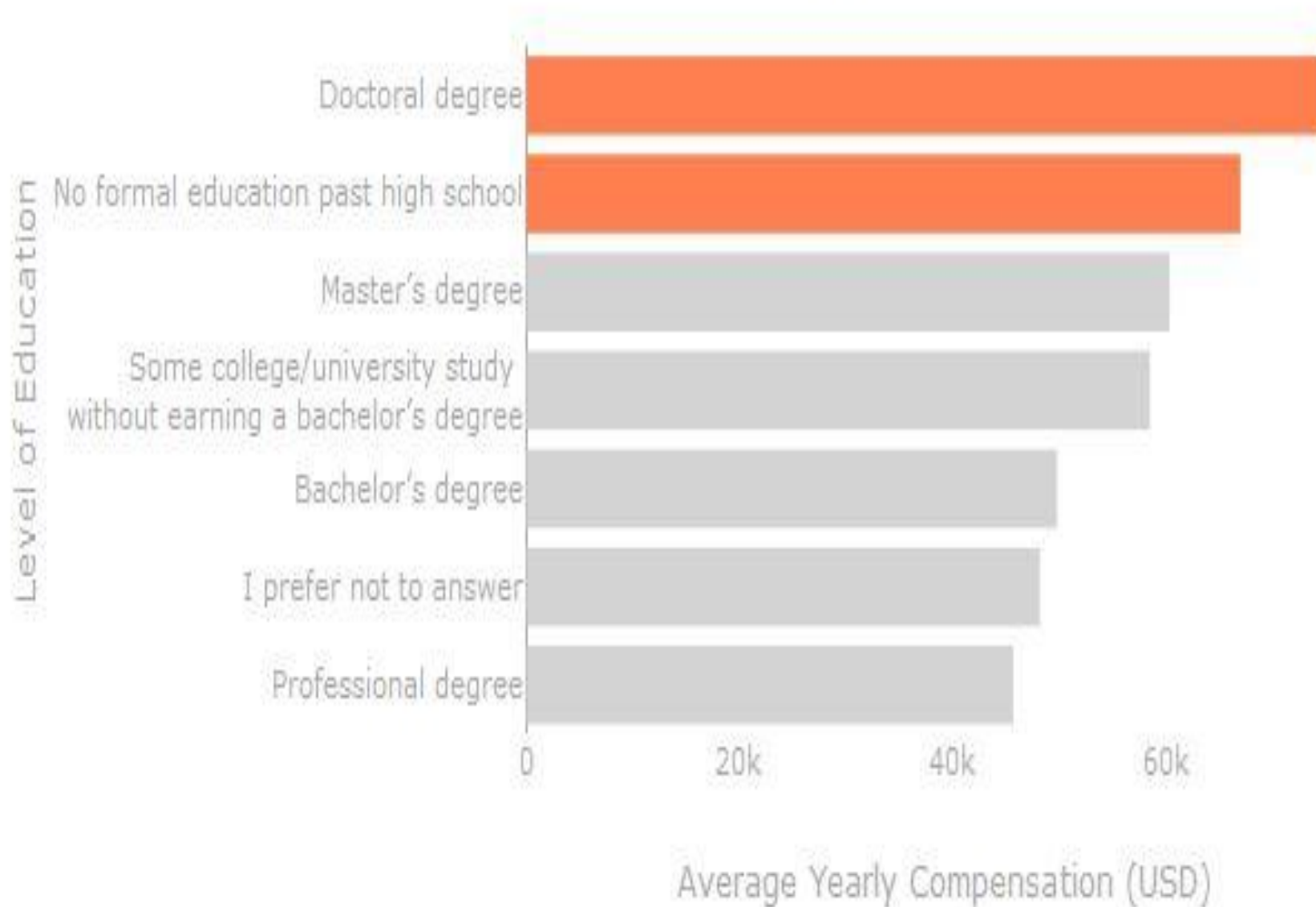
Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q17. What specific programming language do you use most often?

SQL — 2061.559

Python — -599.757

R — -1094.314

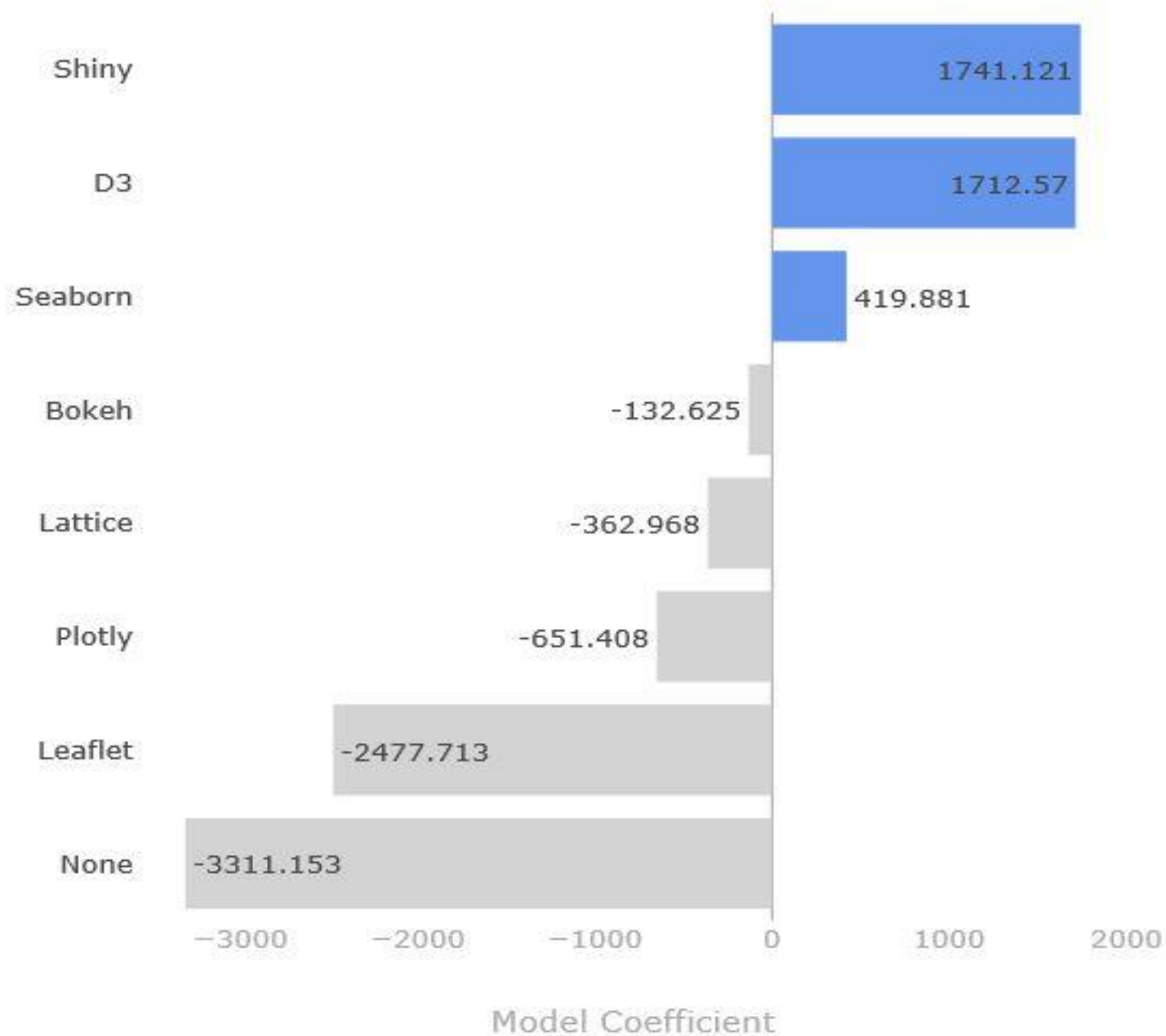-1000 · 0 · 1000 · 2000

Model Coefficient

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

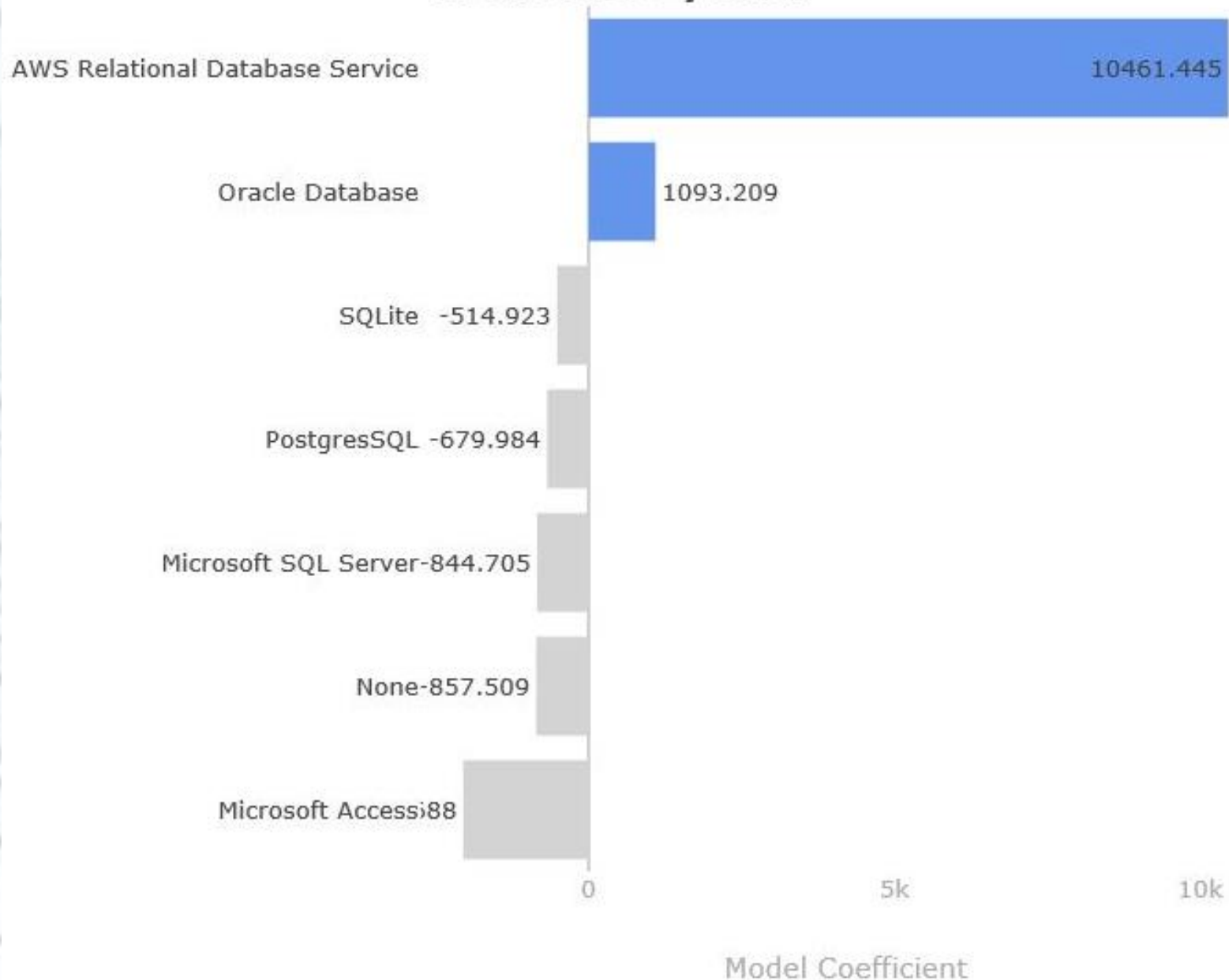Q16. What programming languages do you use on a regular basis?

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q21. What data visualization libraries or tools have you used in the past 5 years?

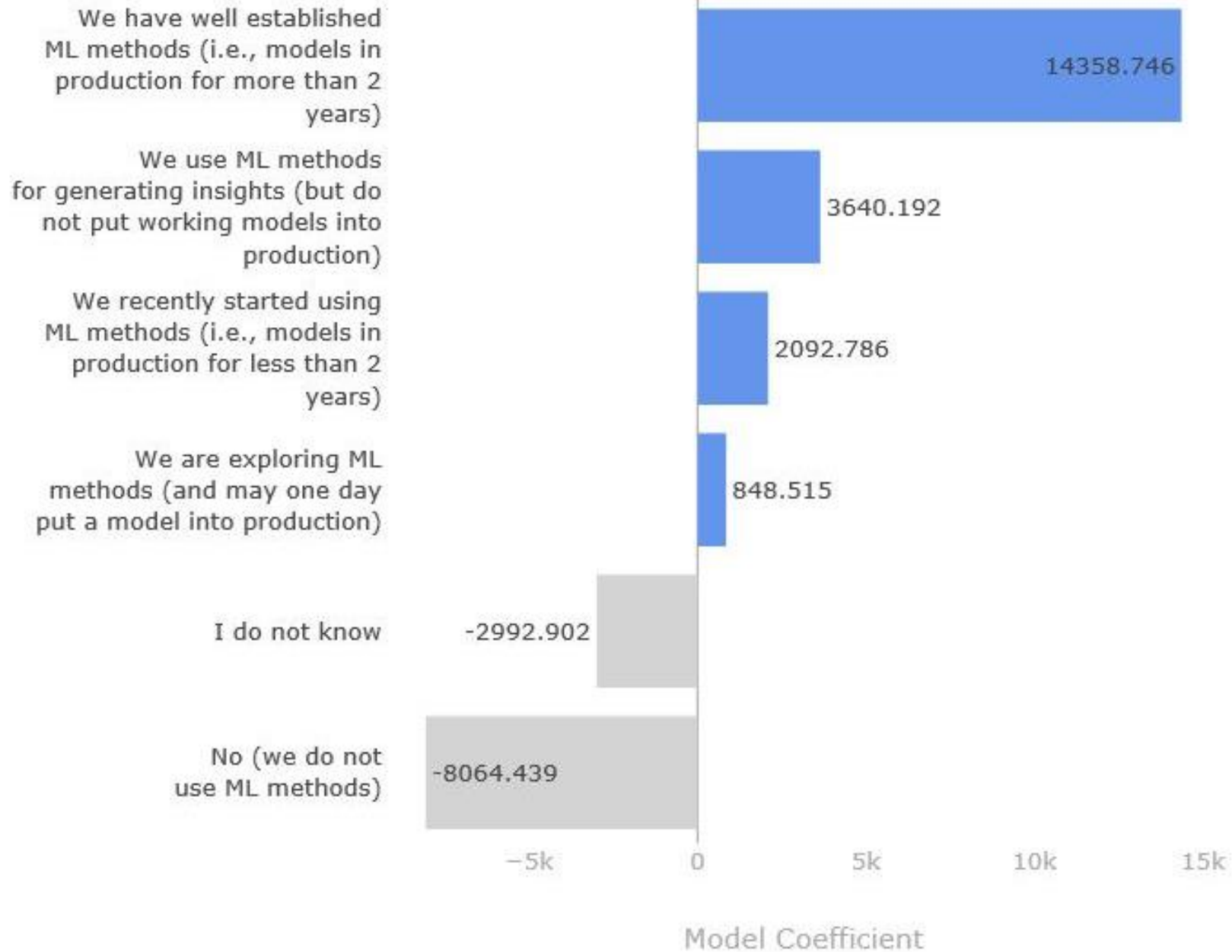| Library	Model Coefficient
Shiny	1741.121
D3	1712.57
Seaborn	419.881
Bokeh	-132.625
Lattice	-362.968
Plotly	-651.408
Leaflet	-2477.713
None	-3311.153

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q29. Which of the following relational database products have you used at work or school in the last 5 years?

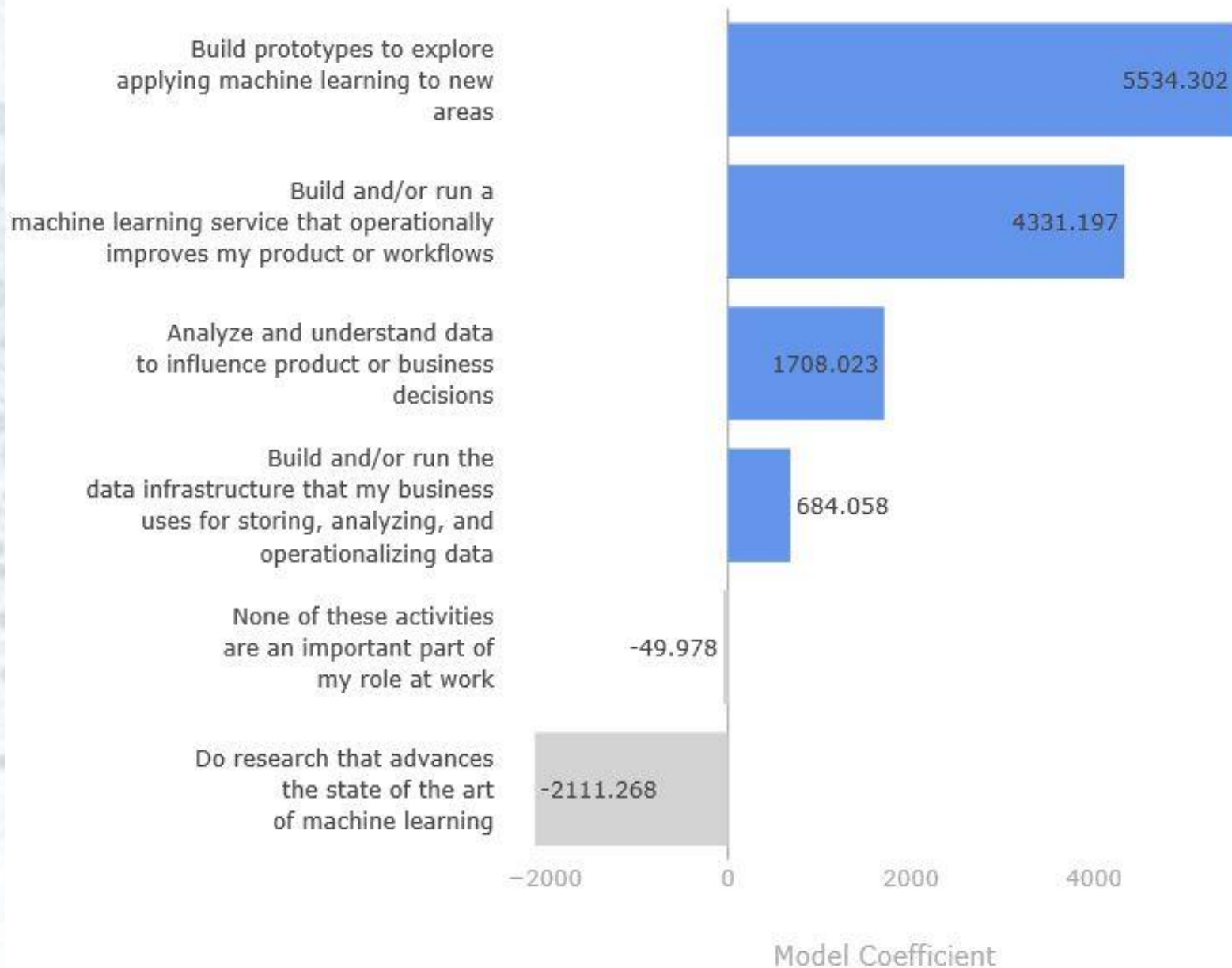|	
AWS Relational Database Service	10461.445
Oracle Database	1093.209
SQLite	-514.923
PostgresSQL	-679.984
Microsoft SQL Server	-844.705
None	-857.509
Microsoft Access	88

Model Coefficient

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q10. Does your current employer incorporate machine learning methods into their business?

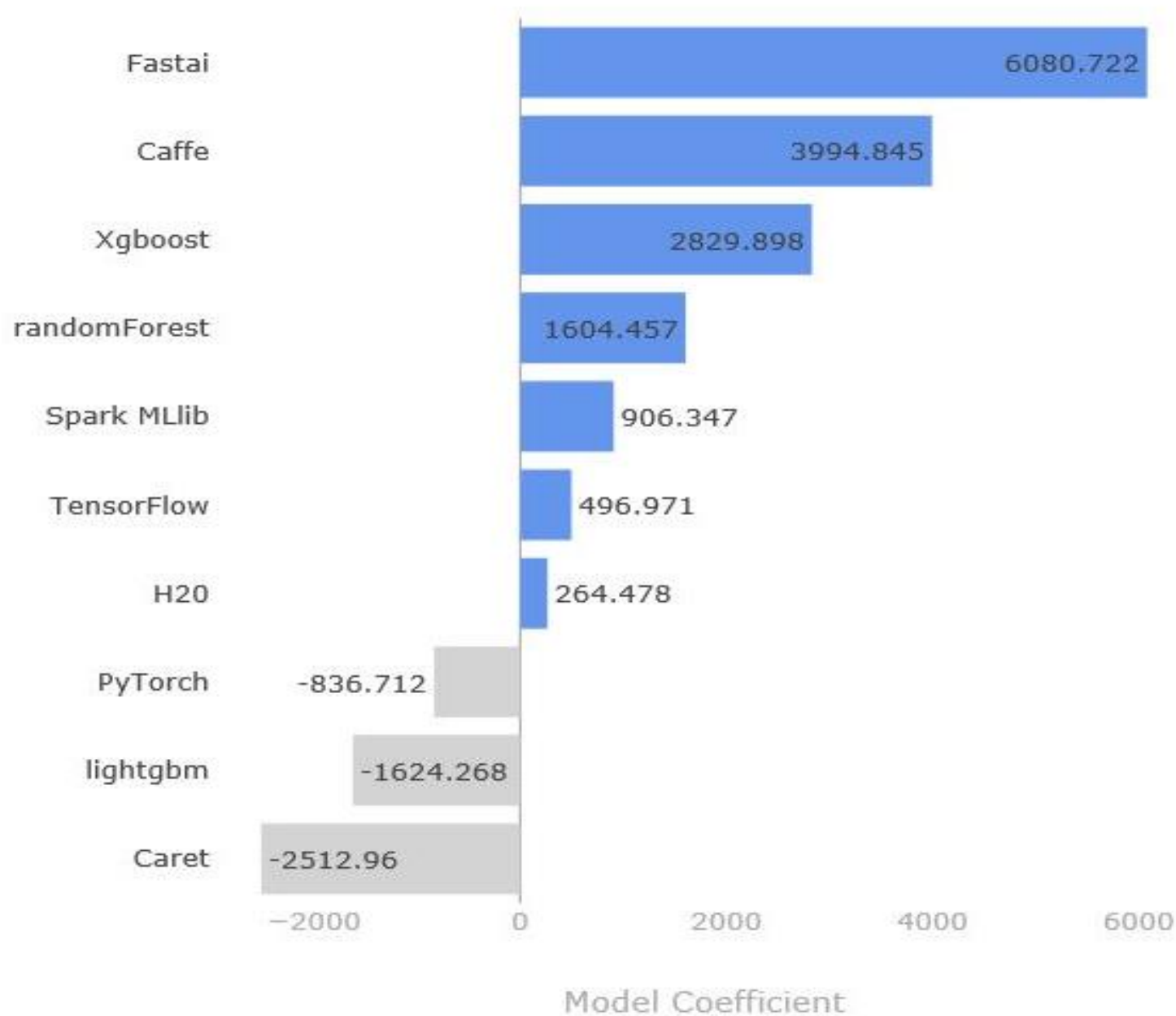| Response	Model Coefficient
We have well established ML methods (i.e., models in production for more than 2 years)	14358.746
We use ML methods for generating insights (but do not put working models into production)	3640.192
We recently started using ML methods (i.e., models in production for less than 2 years)	2092.786
We are exploring ML methods (and may one day put a model into production)	848.515
I do not know	-2992.902
No (we do not use ML methods)	-8064.439

Model Coefficient

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q11. Select any activities that make up an important part of your role at work

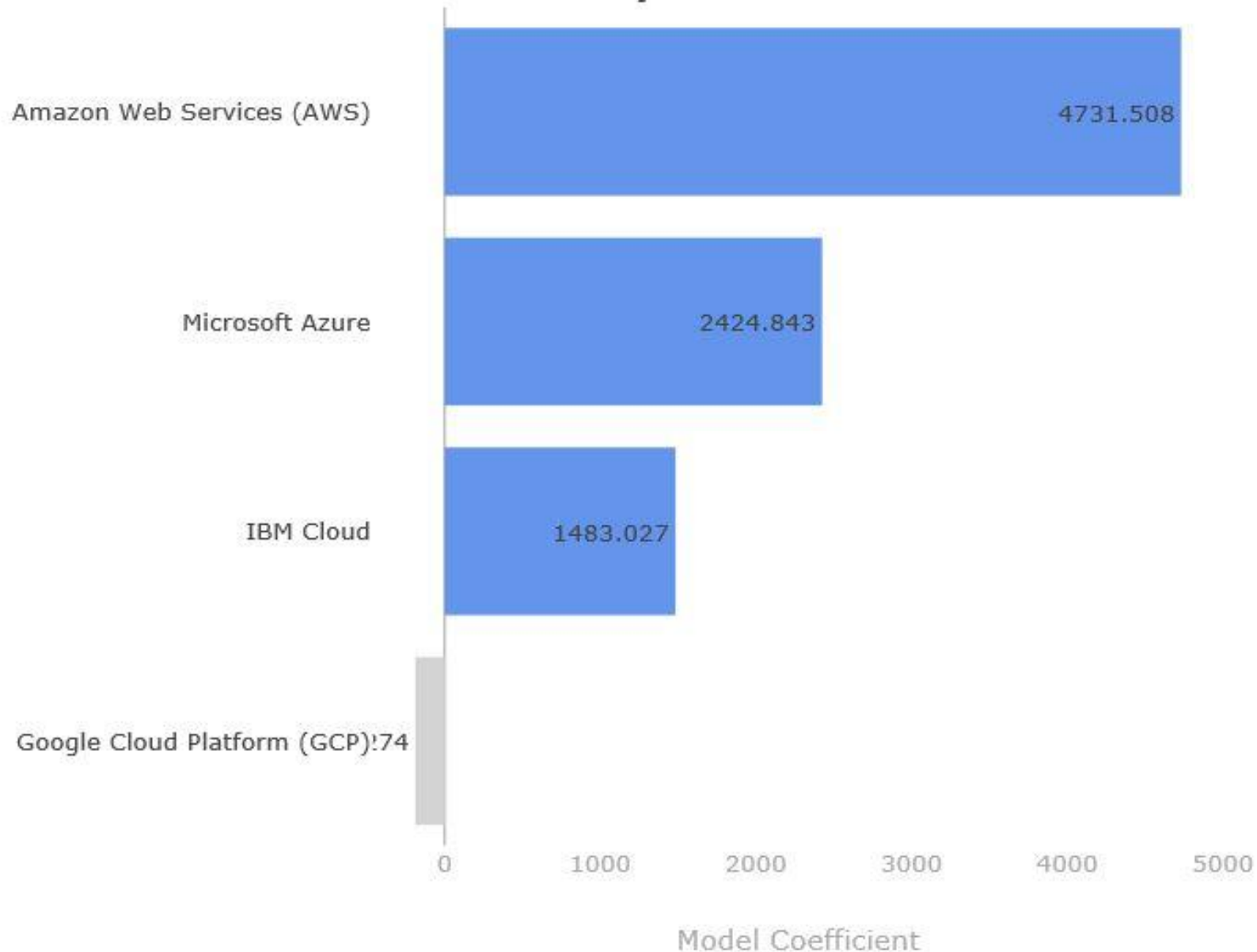| Activity	Model Coefficient
Build prototypes to explore applying machine learning to new areas	5534.302
Build and/or run a machine learning service that operationally improves my product or workflows	4331.197
Analyze and understand data to influence product or business decisions	1708.023
Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data	684.058
None of these activities are an important part of my role at work	-49.978
Do research that advances the state of the art of machine learning	-2111.268

Model Coefficient

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q19. What machine learning frameworks have you used in the past 5 years?

Framework	Model Coefficient
Fastai	6080.722
Caffe	3994.845
Xgboost	2829.898
randomForest	1604.457
Spark MLlib	906.347
TensorFlow	496.971
H20	264.478
PyTorch	-836.712
lightgbm	-1624.268
Caret	-2512.96

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q15. Which of the following cloud computing services have you used at work or school in the last 5 years?

Service	Model Coefficient
Amazon Web Services (AWS)	4731.508
Microsoft Azure	2424.843
IBM Cloud	1483.027
Google Cloud Platform (GCP)	?74

Model Coefficient

Features Importance for Higher compensation by applying supervised machine learning on the Kaggle survey dataset

Q40. Which better demonstrates expertise in data science: academic achievements or independent projects?

	Model Coefficient
Independent projects are much more important than academic achievements	2367.404
Independent projects are slightly more important than academic achievements	789.067
Independent projects are equally important as academic achievements	-526.72

Model Coefficient