Analysis on Analytics Professions

Aaryan Sharma, Ben Poovey, Dan Casella, Gabe Walker

Executive Summary

Based on research done by Kaggle, the world's largest data science community, we've devised that current trends in data science include immersing oneself in **various open-source languages,** creating/having an **'always-be-ready-to-learn' mindset,** completing relevant **certifications** and **online courses,** diving deep into **machine learning methods** and deployment, and being **up-to-date on current events** lead to successful data scientists. Kaggle's data is from an in-house survey that asked over 20,000 opt-in participants to "capture the state of data science in 2021" via several multiple choice questions and responses. This report seeks to identify which of these trends correlate to/result in higher salaries in the field using various data modeling and analytical methods to best predict income for young professionals and students.

Problem Statement

“To Examine Trends in the Analytics Profession and Predict Salaries”

We can do this by considering the following questions:

1. What are the most significant factors driving Data Science salaries?
2. What are the most prevalent tools and techniques being applied by Data Scientists today?
3. What tools and techniques are currently emerging in the field?
4. How and where should aspiring data scientists invest their time and energy for preparedness?
5. Is formal education important to Data Science success?
6. How does the return on formal education compare to other types of learning?

Data and Methodology

Kaggle Survey 2022 Answer Choices

- Full list of questions asked
- Describe exactly which questions were asked to which respondents
- Survey live from 8/16/2022 to 9/16/2022
- Opt-in survey for Kaggle community
    - Anyone was able to respond via the email sent out or the Kaggle website promotion via “nudges”

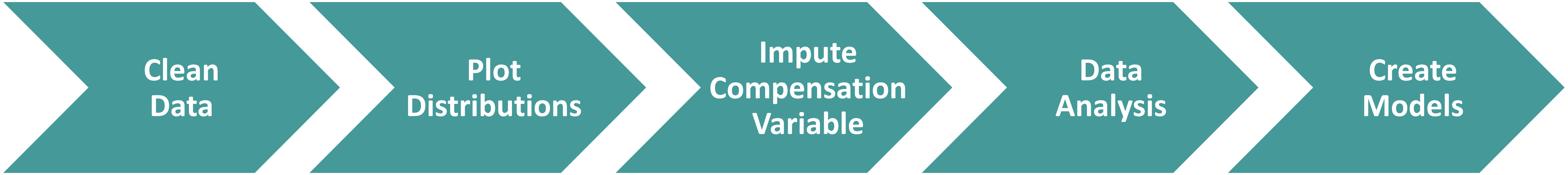
Kaggle Survey 2022 Responses

- Responses to the Kaggle survey that is described in more detail to the left
- Responses to multiple choice questions split into multiple columns
- Excluded responses flagged as “spam” or “duplicate”
- Excluded free-form responses
- Countries or territories with < 50 responses were grouped into the “Other” group

Data collected from the survey includes over 20,000 respondents

Analysis on Analytics Professions

Data and Methodology



Clean Data

- *Got rid of unnecessary variables*
- *Removed outliers*
- *Changed data typing*

Plot Distributions

- Found Nulls
- Created bar charts

Impute Compensation Variable

- *Turned our binned predictor into continuous*
- *Filled in missing values using regression → did not sacrifice any observations*

Data Analysis

- Correlation Matrix
- Summary Statistics
- New Distributions
- Variance Inflation Factors

Create Models

- *Partition Data (60-20-20)*
- *Create Multiple Linear Regression Models*
- *Obtain Metrics*

Technical Summary

Dataset	Adjusted R-Squared
Training	.47
Validation	.44
Test	.44

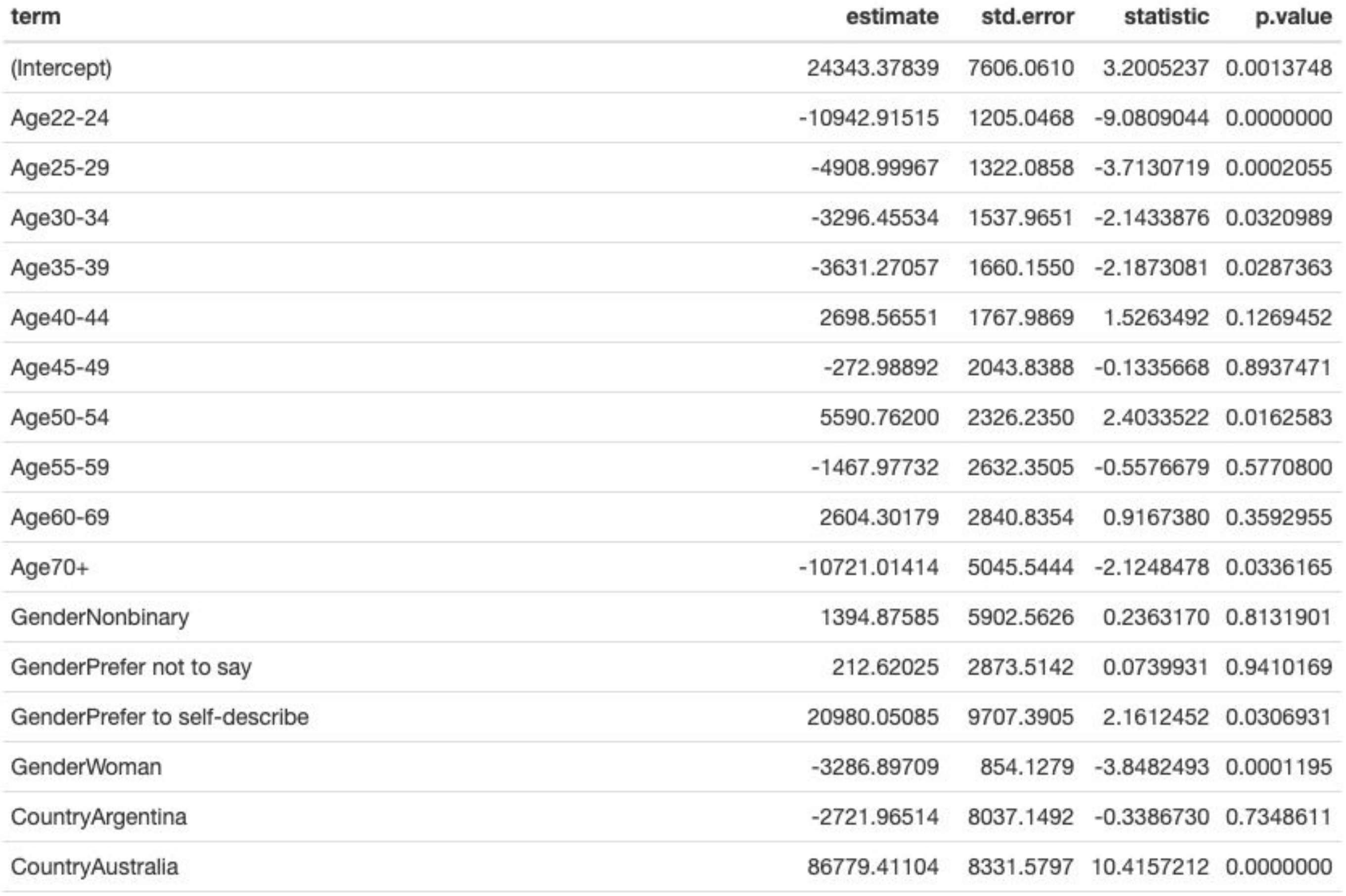
"Around 44% of the variability within our model is explained by the current variables"

.4% of our data removed

Removed observations with compensation values $500,000 and greater ⟶ Increased R^2 by 20 points

Removed 6 Variables:

Published.Academic.Research.Papers, How.many.individuals.are.responsible, Company.Size, Years.Used.Machine.Learning, Similar.Title,and Industry.of.Work ⟶ Variable Coefficients are more true



term	estimate	std.error	statistic	p.value
(Intercept)	24343.37839	7606.0610	3.2005237	0.0013748
Age22-24	-10942.91515	1205.0468	-9.0809044	0.0000000
Age25-29	-4908.99967	1322.0858	-3.7130719	0.0002055
Age30-34	-3296.45534	1537.9651	-2.1433876	0.0320989
Age35-39	-3631.27057	1660.1550	-2.1873081	0.0287363
Age40-44	2698.56551	1767.9869	1.5263492	0.1269452
Age45-49	-272.98892	2043.8388	-0.1335668	0.8937471
Age50-54	5590.76200	2326.2350	2.4033522	0.0162583
Age55-59	-1467.97732	2632.3505	-0.5576679	0.5770800
Age60-69	2604.30179	2840.8354	0.9167380	0.3592955
Age70+	-10721.01414	5045.5444	-2.1248478	0.0336165
GenderNonbinary	1394.87585	5902.5626	0.2363170	0.8131901
GenderPrefer not to say	212.62025	2873.5142	0.0739931	0.9410169
GenderPrefer to self-describe	20980.05085	9707.3905	2.1612452	0.0306931
GenderWoman	-3286.89709	854.1279	-3.8482493	0.0001195
CountryArgentina	-2721.96514	8037.1492	-0.3386730	0.7348611
CountryAustralia	86779.41104	8331.5797	10.4157212	0.0000000

Glimpse of our Model

Analysis on Data Scientist Salaries

What are the most significant factors driving Data Scientist salaries?

Taking data science courses at your university

Consuming podcast media

Consuming journal publication media

Programming for 10-20+ Years

Working in the United Kingdom

Using Hugging Face

Programming in Bash

Programming in PHP

Working in Australia

Ages 25-29

Utilizing Kaggle dataset machine learning repositories

Taking data science courses on Coursera

Found video platforms (YouTube) helpful

Programming for 5-10 Years

Programming in C

Taking cloud certification programs

Working in the United States

Working in Canada

Ages 40-54

Programming in Julia

Working in Hong Kong

Using TensorFlow Hub

Company you work for has incorporated Machine Learning Methods

Analysis on Current Tools and Techniques

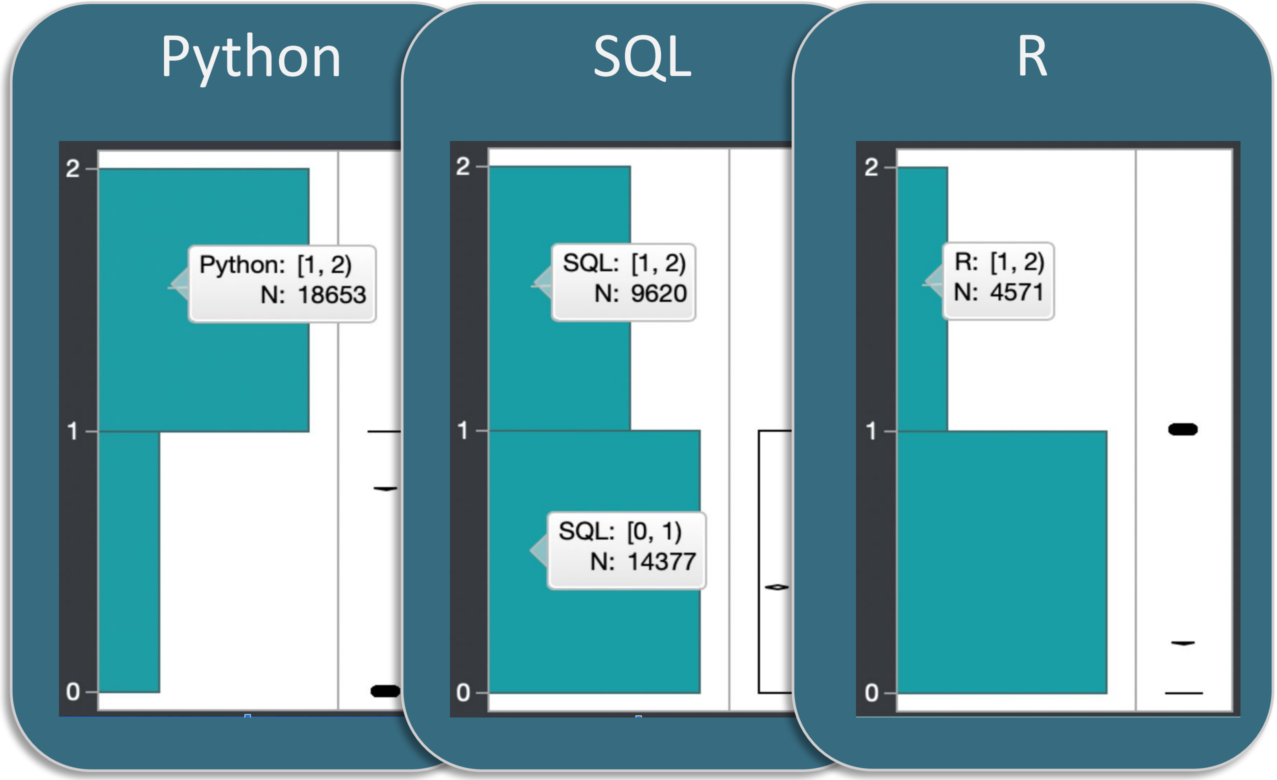
What are the most prevalent tools (software) being applied by Data Scientists Today?

The most prevalent languages applied in data science:

- **78%** of survey respondents know Python
- **40%** of survey respondents know SQL
- **19%** of survey respondents know R

Although most successful data scientists use Python, several utilize 2+ languages in their line of work:

- **36%** of survey respondents know Python and SQL → **$47000** average salary
- **17%** of survey respondents know Python and R → **$55000** average salary
- **10%** of survey respondents know Python, SQL, and R → **$57000** average salary



Analysis on Current Tools and Techniques

What are the most prevalent techniques (methods) being applied by Data Scientists Today?

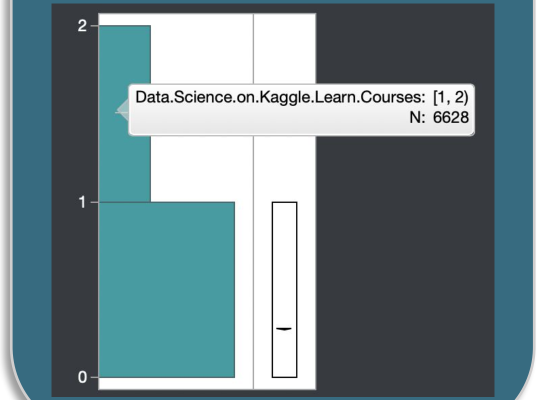
The most prevalent methods used to learn data science:

- YouTube Media - 50% utilize YouTube
- Kaggle Media - 47% utilize Kaggle Media
- 41% of survey respondents utilize data science courses on Coursera

Less prevalent methods include:

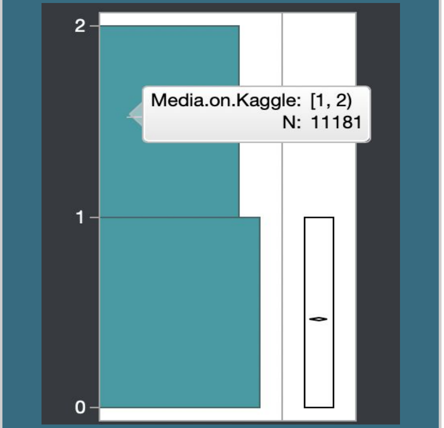
- Fast.ai - 5% of survey respondents have learned from Fast.ai
- Kaggle Data Science courses - 29%
- Udemy - 26%

Kaggle Courses

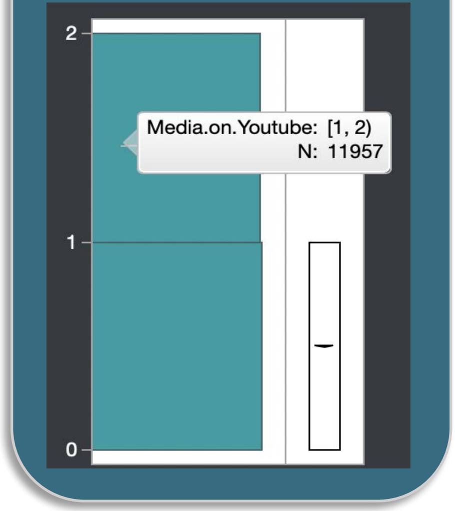


Kaggle is a hub for data. Individuals visit Kaggle to obtain data for their own practice, and consult YouTube and other media for aid. This explains why our Kaggle distributions differ

Kaggle Media



Youtube Media



Analysis on Current Tools and Techniques

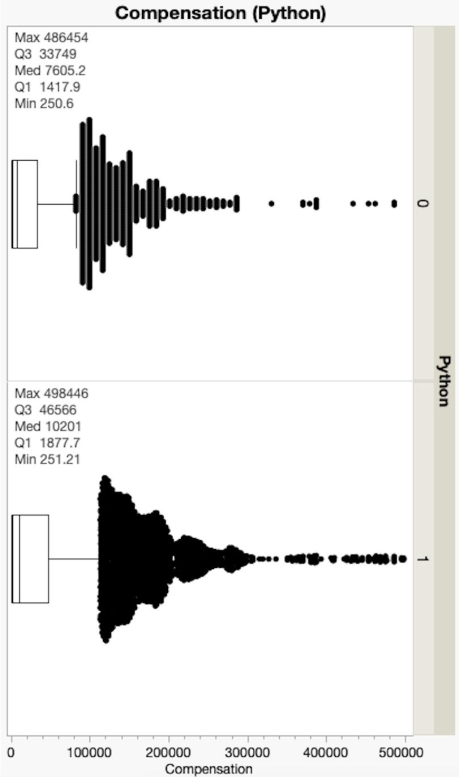
What tools and techniques have the highest ROI?

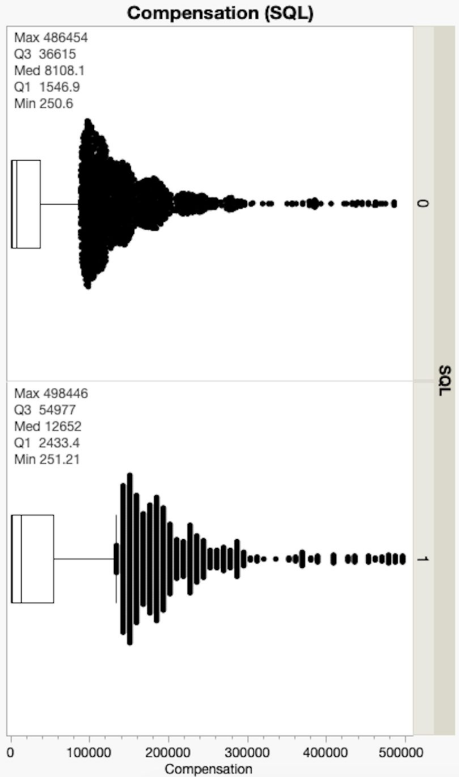
Summary Statistics:

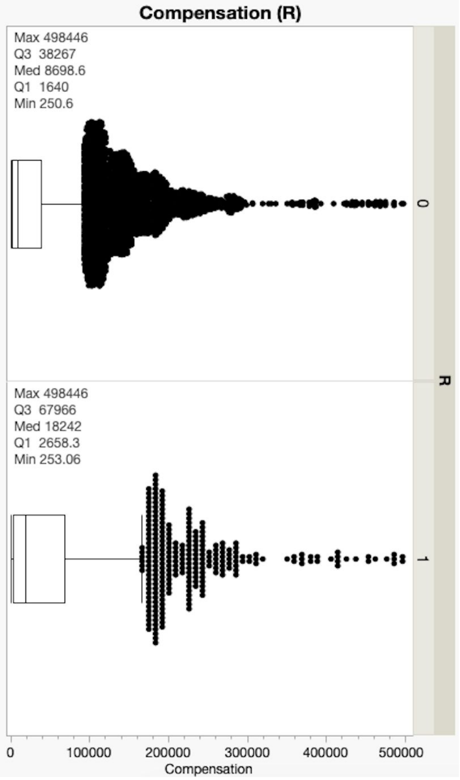
- The maximum salary for a data scientist is nearly identical for each language
  - This is likely due to an individual knowing more than one language.
- The distribution of each language is nearly identical, with each having a heavy right skew.
- Knowing any of the following languages increases median compensation (as opposed to not knowing the language)

Conclusions:

- Salaries do not appear to differ among the top three programming languages for data science.
- Although SQL has a higher number of outliers, users of the language also have a slightly higher number of users making $0-$20,000/yr







Analysis on Emerging Tools and Techniques

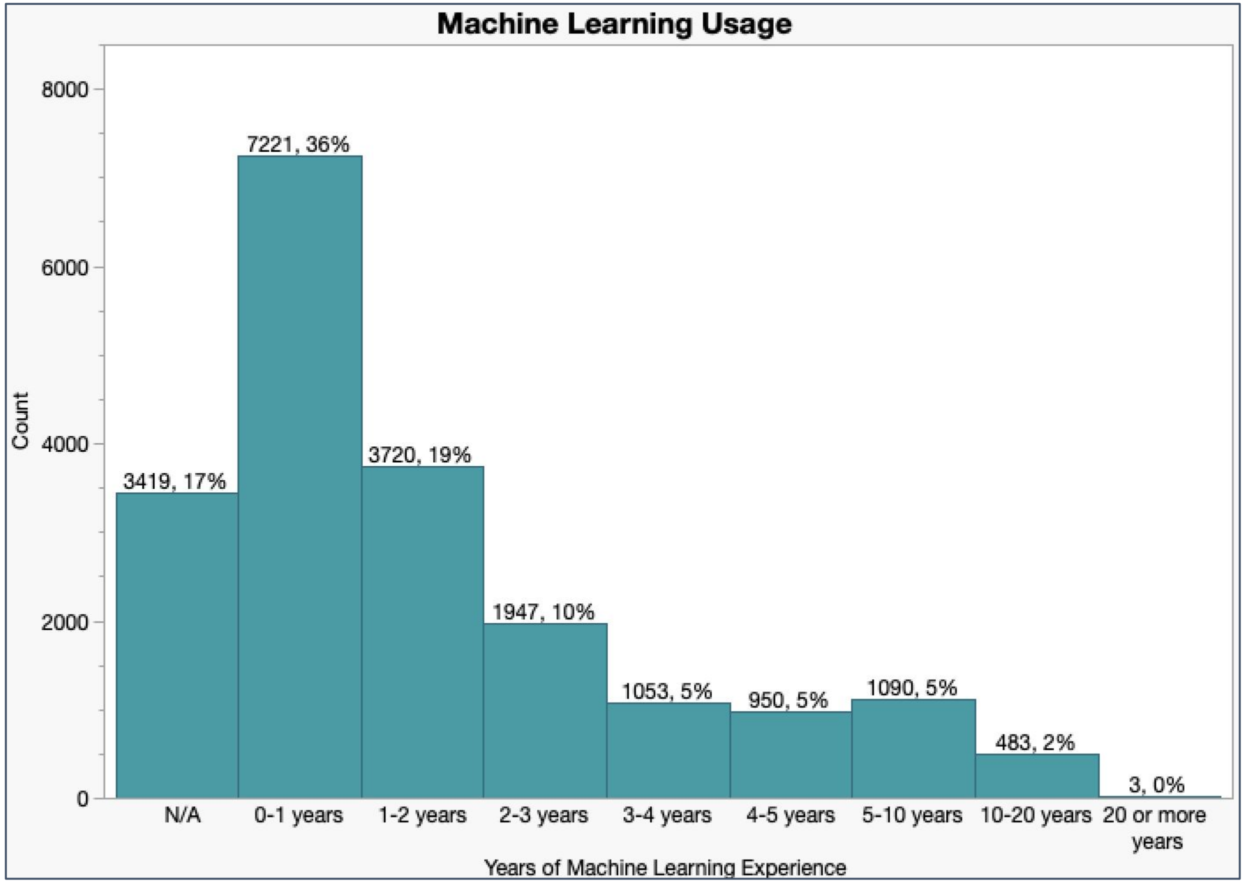
What tools and techniques are emerging in the field of Data Science

Summary Statistics:

- The machine learning usage histogram is heavily *right skewed*, with a disproportionate amount of users with little to no usage.
- Still a good margin of people who have not worked with ML (17%)
- Just over 70% of the respondents have used (or did not use) machine learning for < 2 years
- Under 10% of respondents have used ML for 5+ years

Conclusions:

- Many of machine learning practitioners are relatively new to the technique
- Importance of machine learning techniques are rising, with an increasing number of data science positions relying on machine learning techniques.



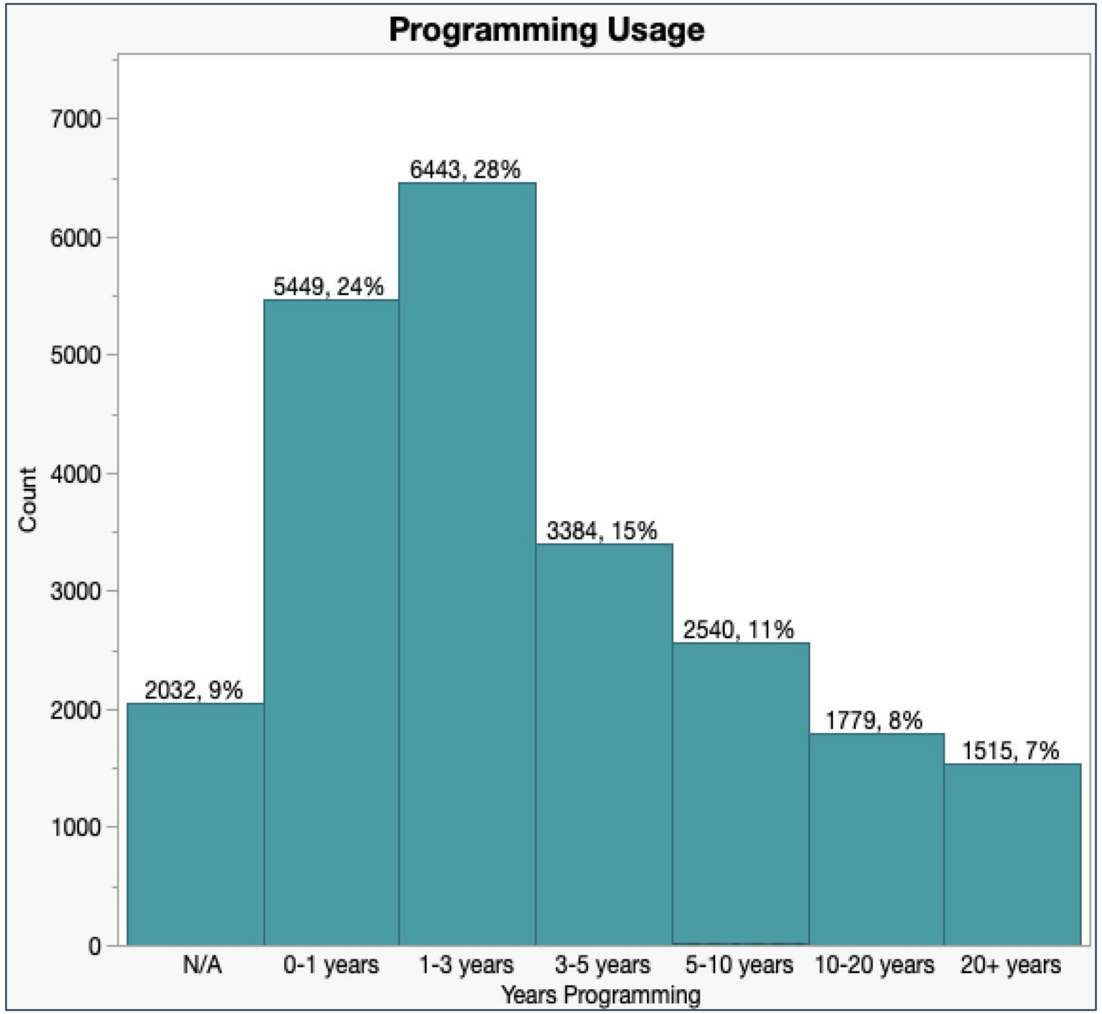
What tools and techniques are emerging in the field of Data Science

Summary Statistics:

- The programming usage histogram is slightly *right skewed*, with a over 60% of users with < 3 years of programming experience.
- There is only a small number of people with a lot of programming experience (15% with 10+ years)
- Less than 10% of survey participants have not used programming skills in their jobs (Only 9% fall into N/A)

Conclusions:

- Programming usage is slightly more integrated into the data science field than machine learning
  - A lower percentage of participants have been programming for 1 year or less (*53%* for ML, only *33%* for programming)
- Importance and value of programming in data science is increasing as more users practice



How and where should aspiring data scientists invest their time and energy to prepare for the current and future Data Science environment?

Prepare emerging skills

The application of machine learning is currently deployed by about half of data scientists with over 70% of these users having used it for less than 2 years.

Exposure to more languages

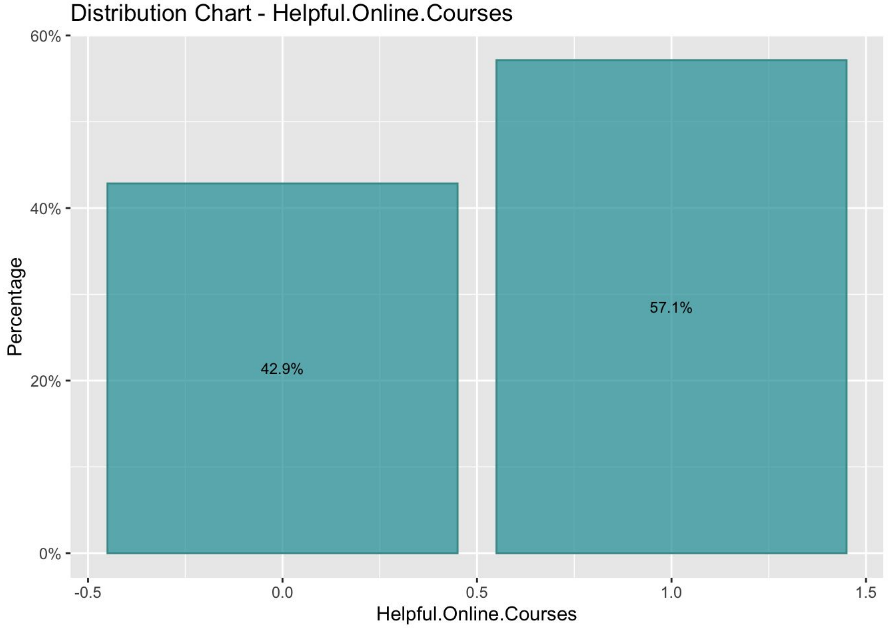
Although the salaries for the top three data science languages are roughly the same, exposure to multiple languages only solidifies a data scientist's position in their salary bracket and gives them potential to earn more

Length of knowledge

Data scientists who have been practicing one or more languages for over 5 years have a significant advantage over data scientists who have only practiced their language for 0-5 years.

Analysis on Formal Education and Success

Is formal education important to success as a Data Scientist?

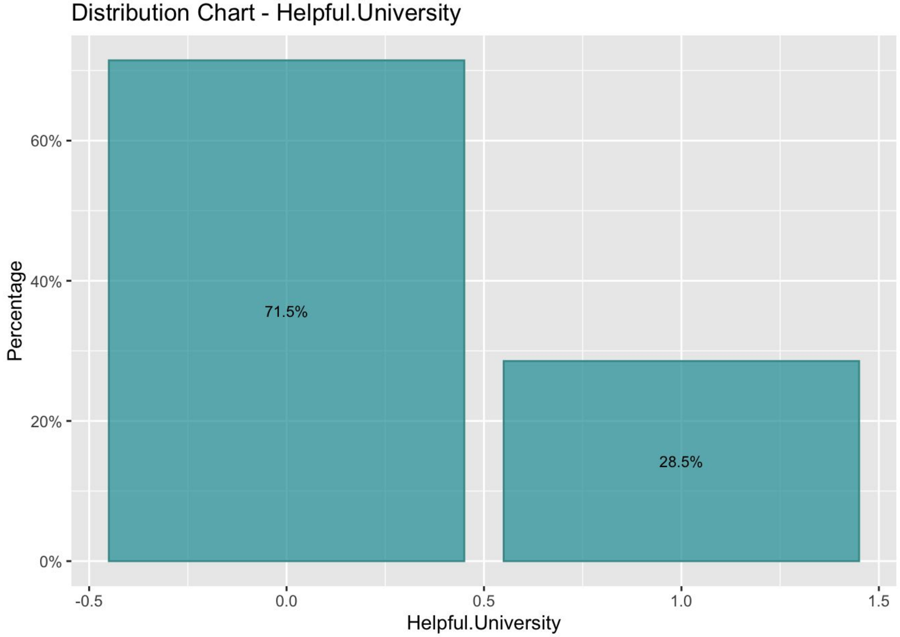


Many data science professionals found online courses to be more helpful than not

57% of professionals found online courses helpful

Many data science professionals surprisingly found university courses to not be helpful in their careers:

only 29% found it helpful



Analysis on Formal Education and Success

Is formal education important to success as a Data Scientist?

Compensation Data

No Formal Education

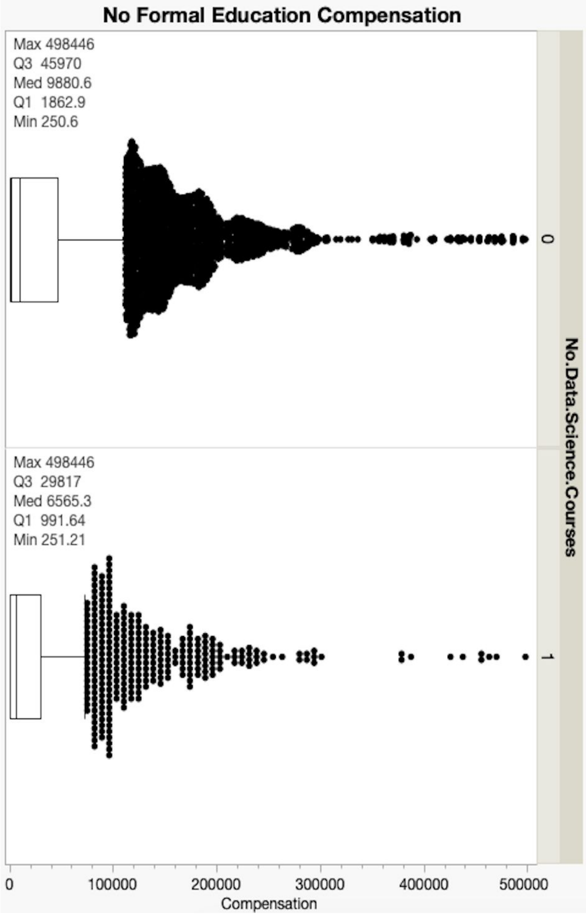
- Not having a formal education results in a median salary of $6,565
- Not taking data science courses leads to a $3,315 decrease in median compensation

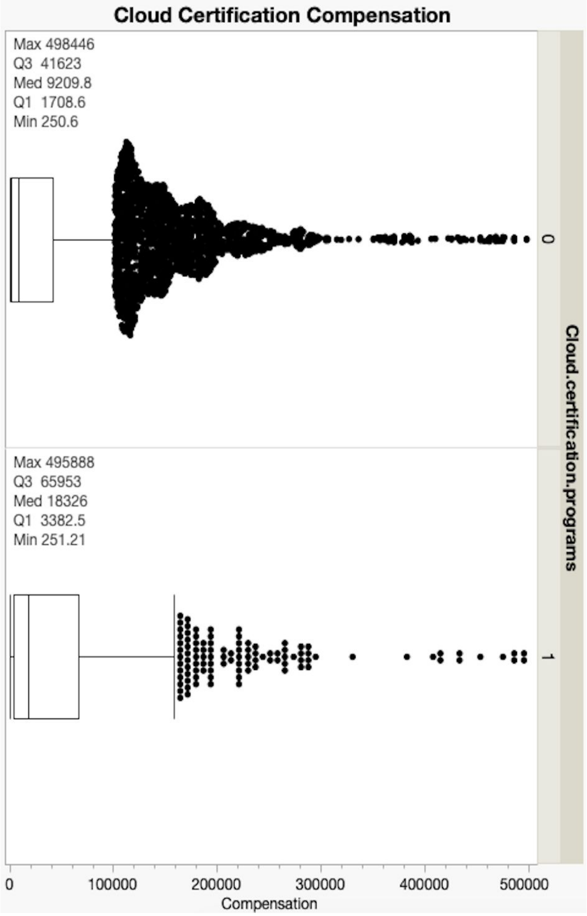
Cloud Certification

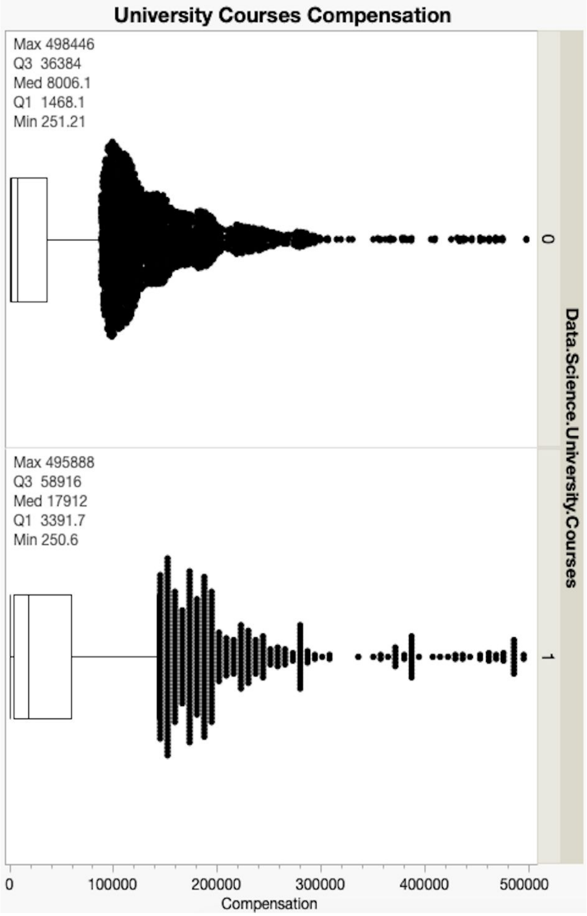
- Professionals with cloud certifications have a median salary of $18,326
- $9,116 increase in median compensation opposed to not having cloud certifications

University Courses

- Taking university courses is associated with a median salary of $17,912
- $9,904 increase from not taking courses

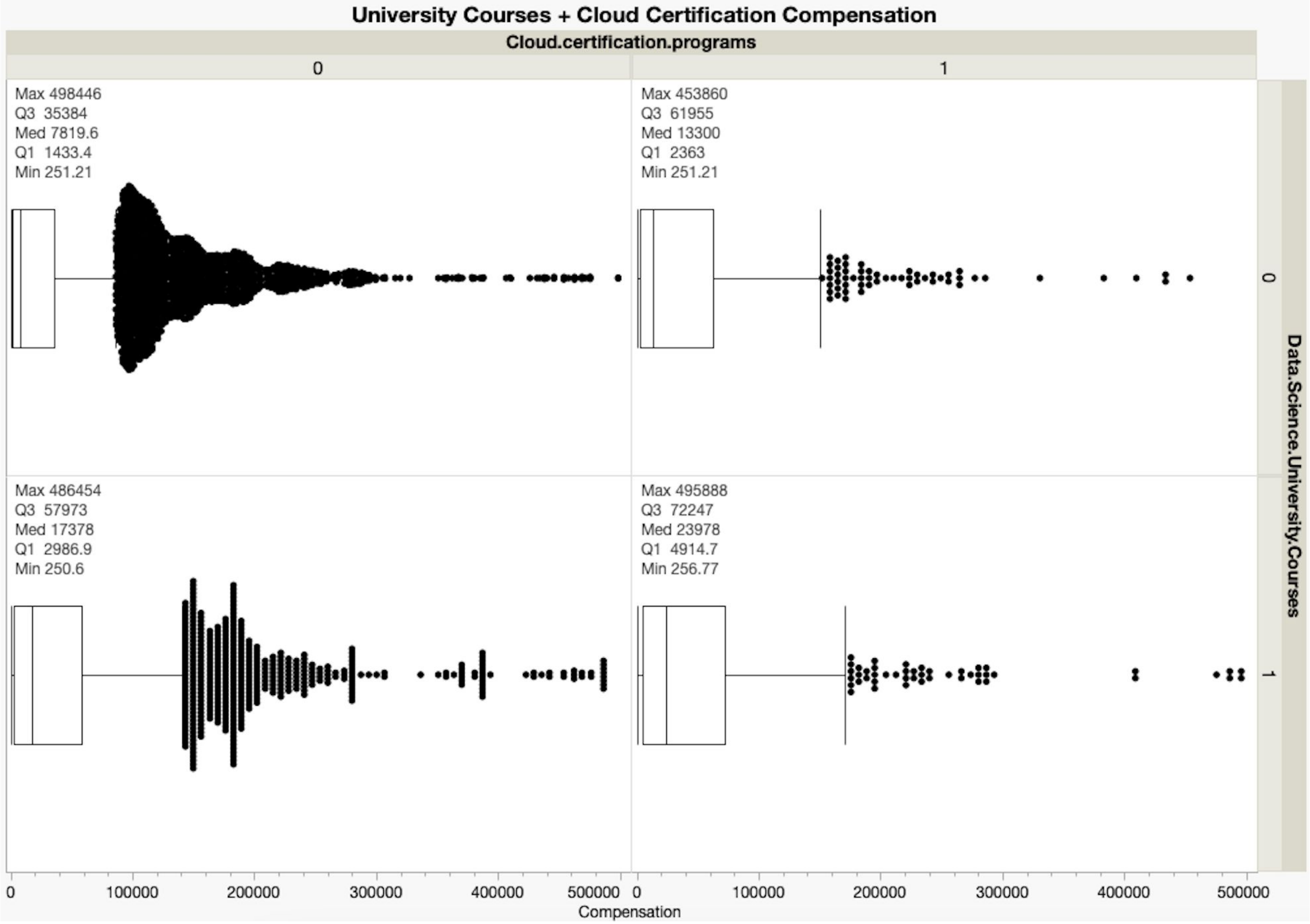






Analysis on Formal Education and Success

Is formal education important to success as a Data Scientist?



Given that formal education and cloud certifications increase median compensation for data science professionals…

Professionals who take **university courses *AND*** are **cloud certified** see a median salary of **$23,978**, $16,151 higher than if someone were neither formally educated nor cloud certified

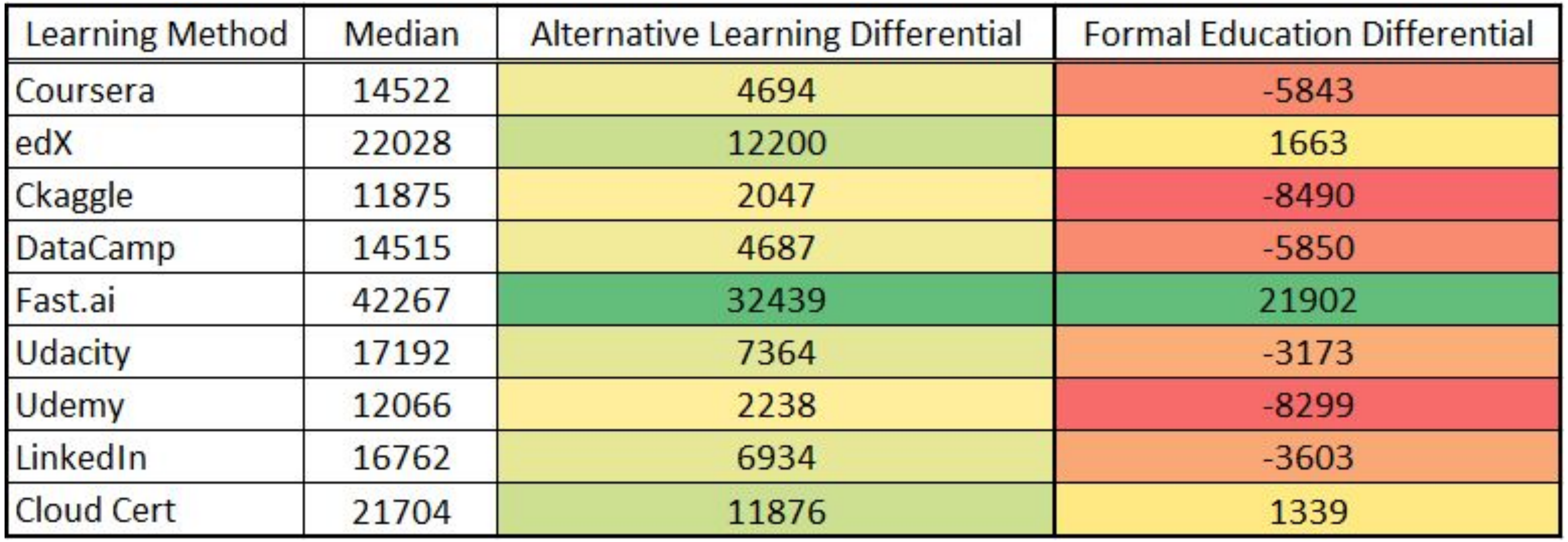
Professionals who are **cloud certified** but **NOT** formally educated have a median compensation of **$13,300**

Professionals who are **formally educated**, but **NOT** cloud certified earn a median compensation of **$17,378**

If a professional must choose 1, formal education has a higher return on investment than getting cloud certified

Analysis on Formal Education and Return Comparison

How does the return on formal education compare to other types of learning?



Learning Method	Median	Alternative Learning Differential	Formal Education Differential
Coursera	14522	4694	-5843
edX	22028	12200	1663
Ckaggle	11875	2047	-8490
DataCamp	14515	4687	-5850
Fast.ai	42267	32439	21902
Udacity	17192	7364	-3173
Udemy	12066	2238	-8299
LinkedIn	16762	6934	-3603
Cloud Cert	21704	11876	1339

Alternative Learning Differential
Difference in median income if the professional used the learning method

Formal Education Differential
Difference between median income from formal education and alternative learning

Conclusions:

- Utilizing any alternative learning measure listed above increases median compensation for data science professionals
- When compared with University courses, only **3** alternative learning methods see a higher return on investment; Fast.ai, edX, and Cloud Certification Programs
    - Those who learn with *Fast.ai* earn a **median salary of $42,267**, $32,439 higher than if one chose to not learn with Fast.ai
    - **Fast.ai** saw a median salary of **$21,902 higher** than learning via university courses

Analysis on Data Science from the Institution’s Point of View

How should educational institutions think about the role of formal education in the world of data science?

The Role of Formal Education

- While formal education is important for success in the data science field, there are mixed reviews about its importance
  - Taking university courses increases average compensation for data science professionals as per our model
  - That being said, many of the survey participants (**71%**) found that attending university was not helpful
  - Many professionals turn to alternative learning sources to supplement their formal education

Formal education sets budding professionals up for success by building a foundation upon which a student can succeed, but the execution and focus of formal education can use tweaking…

Understand Formal Education Shortcomings

Discover what Fast.ai (or any other alternative method) teaches that university courses don’t provide, and integrate that knowledge into the curriculum

Tailored Learning Opportunities

Improving students’ satisfaction with formal education by utilizing more “free-form” classes, allowing students to customize their learning experience to be more similar to other online learning options

Utilize Emerging Technologies to keep Students Ahead

Tools and techniques like machine learning are up and coming; It is critical that educational institutions realize their importance in the future job market

What specific methodologies should formal education institutions use to train data scientists?

Advocate for online aid for class

Knowing how to navigate social media, kaggle, and YouTube videos proves to be helpful in the field

Endorse outside learning

Making sure students are in the know and go above and beyond makes them more:

1) Outstanding to employers
2) Creates a habit of knowing what's going on in the current economy and state of the world, making them well-rounded data scientists

Teach machine learning methods and guiding through repositories

Knowing how to navigate PyTorch, HuggingFace, and TensorFlow are not only useful in the data science community, it’s also a huge salary booster

~$11,000 salary increase in knowing HuggingFace
~$4,200 salary increase in knowing PyTorch
~$2,000 salary increase in knowing TensorFlow

What specific methodologies should formal education institutions use to train data scientists?

Making sure students start programming in their first semester

Python, R, and SQL are important languages to start on

Offer courses that teach JavaScript, Bash, and Go

These languages are salary boosters

Bash: Suited for automating tasks

Go: good for scaling development at no additional computational cost

Offering Data Science courses are crucial

~$5,000 dollar increase in salary

Students learn how to create models, the ethics of data science, machine learning methods, and the real-world applications of data science

Implementing certifications and LinkedIn learning courses prove to be useful

Adding these in the curriculum as extra activities will be beneficial to the student

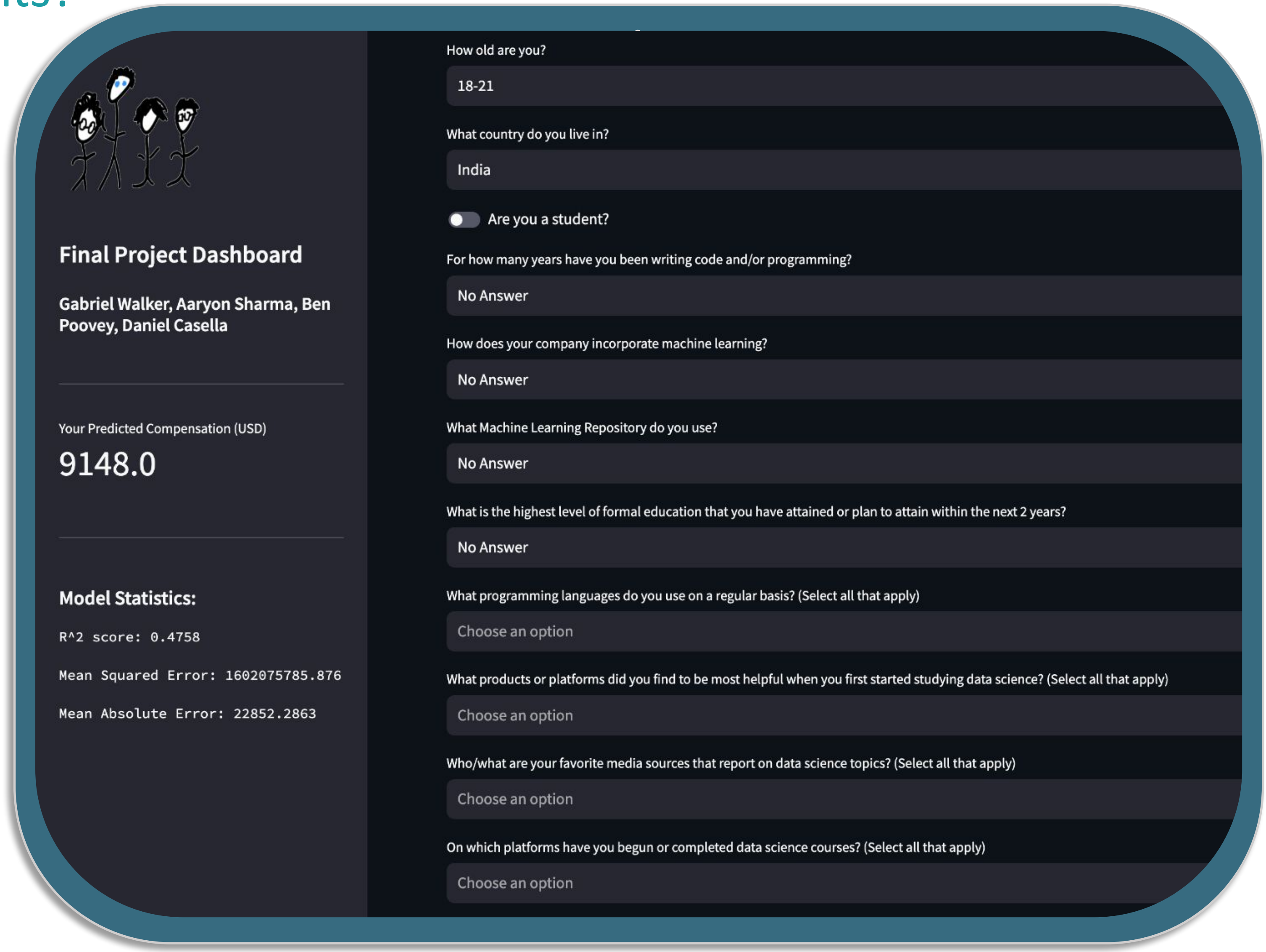
~$1,200 dollar increase in salary

Dashboard: Viewing the Model

How can we leverage our data insights?

Dashboard: https://analyticsdashboard.streamlit.app/

Allows students to **predict their compensation** based on our important features. They can also use this tool to experiment with what they should learn in the future



Final Project Dashboard

Gabriel Walker, Aaryon Sharma, Ben Poovey, Daniel Casella

Your Predicted Compensation (USD)

9148.0

Model Statistics:

```
R^2 score: 0.4758

Mean Squared Error: 1602075785.876

Mean Absolute Error: 22852.2863
```

How old are you?
18-21

What country do you live in?
India

○ Are you a student?

For how many years have you been writing code and/or programming?
No Answer

How does your company incorporate machine learning?
No Answer

What Machine Learning Repository do you use?
No Answer

What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
No Answer

What programming languages do you use on a regular basis? (Select all that apply)
Choose an option

What products or platforms did you find to be most helpful when you first started studying data science? (Select all that apply)
Choose an option

Who/what are your favorite media sources that report on data science topics? (Select all that apply)
Choose an option

On which platforms have you begun or completed data science courses? (Select all that apply)
Choose an option

Limitations and Next Steps

How should we go about implementation and next steps?

Sharing our findings with educational institutions and those interested in data science

- Institutions can rework curricula to reflect import findings
- Individuals can get a headstart on their career

Human Resource Divisions and current data scientists can use our model as a salary negotiation tool

- makes sure salaries are at a competitive standpoint
- makes sure everyone is getting what they deserve

Mutate our factor variables into multiple binary variables

- Ie. Turn our Age variable into 10 binary variables, one for each bin → this will result in a more accurate model

Adding in new variables related to the model and reworking current variables

- Adding in a mixed effect typing to compensation based on country will give more accurate salary results → removing demographic bias in total is the goal
- A variable that tracks their educational background (ie. major, minor) or one that tracks extracurriculars (personal projects, # of hackathons and datathons participated in)

Obtain data from other sources, not just a Kaggle survey

- We believe that some of the metrics used in this model may be biased because of Kaggle's involvement
- Data about usefulness of Kaggle as an alternative learning source may suffer from selection bias or the halo effect that may skew model results

Q&A