2018 Kaggle ML & DS Survey

- **Source of data:**
  - **https://www.kaggle.com/kaggle/kaggle-survey-2018**

Data Science Flex - Unit 3
Supervised Learning Capstone

David J.R. Gay
February 2019

2018 Kaggle ML & DS Survey

- **Introduction:**
  - Why this dataset?
    - Dataset captures many details associated with the field
    - “Meta” factor
    - I can relate to some questions & responses
  - Overview
    - Nature of the raw data
    - Some recurring themes in the data
    - Visuals / Bar charts for respondents’ Q&A totals (16 slides)
    - Supervised Learning models - Predicting income tiers (10 slides)
    - The notebook contains more charts & source code

2018 Kaggle ML & DS Survey

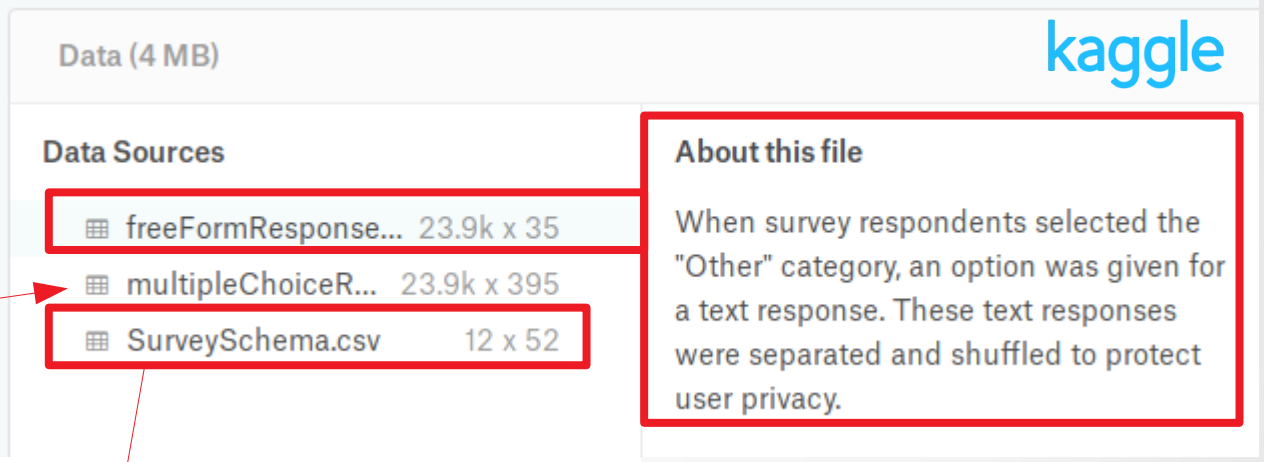
Survey Methodology

kaggle

- This survey received 23,859 usable respondents from 147 countries and territories. If a country or territory received less than 50 respondents, we grouped them into a group named "Other" for anonymity.
- We excluded respondents who were flagged by our survey system as "Spam".
- Most of our respondents were found primarily through Kaggle channels, like our email list, discussion forums and social media channels.
- The survey was live from October 22nd to October 29th. We allowed respondents to complete the survey at any time during that window. The median response time for those who participated in the survey was 15-20 minutes.
- Not every question was shown to every respondent. You can learn more about the different segments we used in the schema.csv file.
- To protect the respondents' identity, the answers to multiple choice questions have been separated into a separate data file from the open-ended responses. We do not provide a key to match up the multiple choice and free form responses. Further, the free form responses have been randomized column-wise such that the responses that appear on the same row did not necessarily come from the same survey-taker.

2018 Kaggle ML & DS Survey

- **Raw Data:**
  - (3) CSV files
  - Main set of data contained here



2.1 Dataset Schema

```
In [6]: schema.shape
Out[6]: (12, 52)
```

```
In [7]: schema.head()
```

Out[7]:

	2018 Kaggle Machine Learning and Data Science Survey	Q1	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	...	Q47	Q48	Q49
0	Question:	What is your gender? - Selected Choice	Does your current employer incorporate machine...	Select any activities that make up an importan...	What is the primary tool that you use at work ...	Which of the following integrated development ...	Which of the following hosted notebooks have y...	Which of the following cloud computing service...	What programming languages do you use on a reg...	What specific programming language do you use ...	...	What methods do you prefer for explaining and/...	Do you consider ML models to be "black boxes" ...	What tools and methods do you use to make your...
1	# of Respondents:	23860	20670	19518	19199	19117	18971	18864	18828	15223	...	13418	13369	12891
2	Who was excluded? (0 = not excluded; 1 = exclu...	0	0	0	0	0	0	0	0	0	...	0	0	0
3	If What is your age (# years)? 0-17 Is Selecte...	0	1	1	1	1	1	1	1	1	...	1	1	1
	If What is the													

- **freeFormResponses.csv**
  - Sparse
  - Decoupled from main set of data
- **SurveySchema.csv**
  - “Not every question was shown to every respondent.”
  - Some questions were prerequisites for others

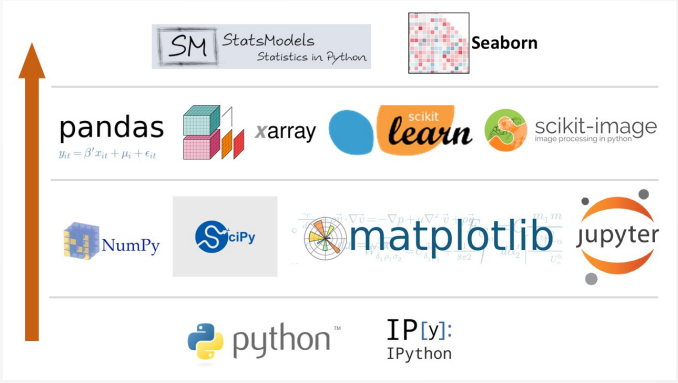
2018 Kaggle ML & DS Survey

- **Main Set of Data:**
  - multipleChoiceResponses.csv
  - Any columns referencing free-form text responses were ignored
  - Answers to:
    - Single-choice questions - in one (1) column
    - Questions with multiple parts or choices - spread across multiple columns
  - Mostly categorical data
    - Nominal
    - Interval / Ordinal
    - Bar charts are sorted by total count of respondents, regardless of any logical order in the categories / labels
  - Data for Questions 34 & 35 was both categorical & numerical

Some Recurring Themes in the Data:

- *Youth*
  - Respondents
  - The field itself

- *Python ML & DS Ecosystem*



Keras

PYTORCH

- *Possible Disruption in Education*

THINKFUL

GENERAL ASSEMBLY

coursera

UDACITY

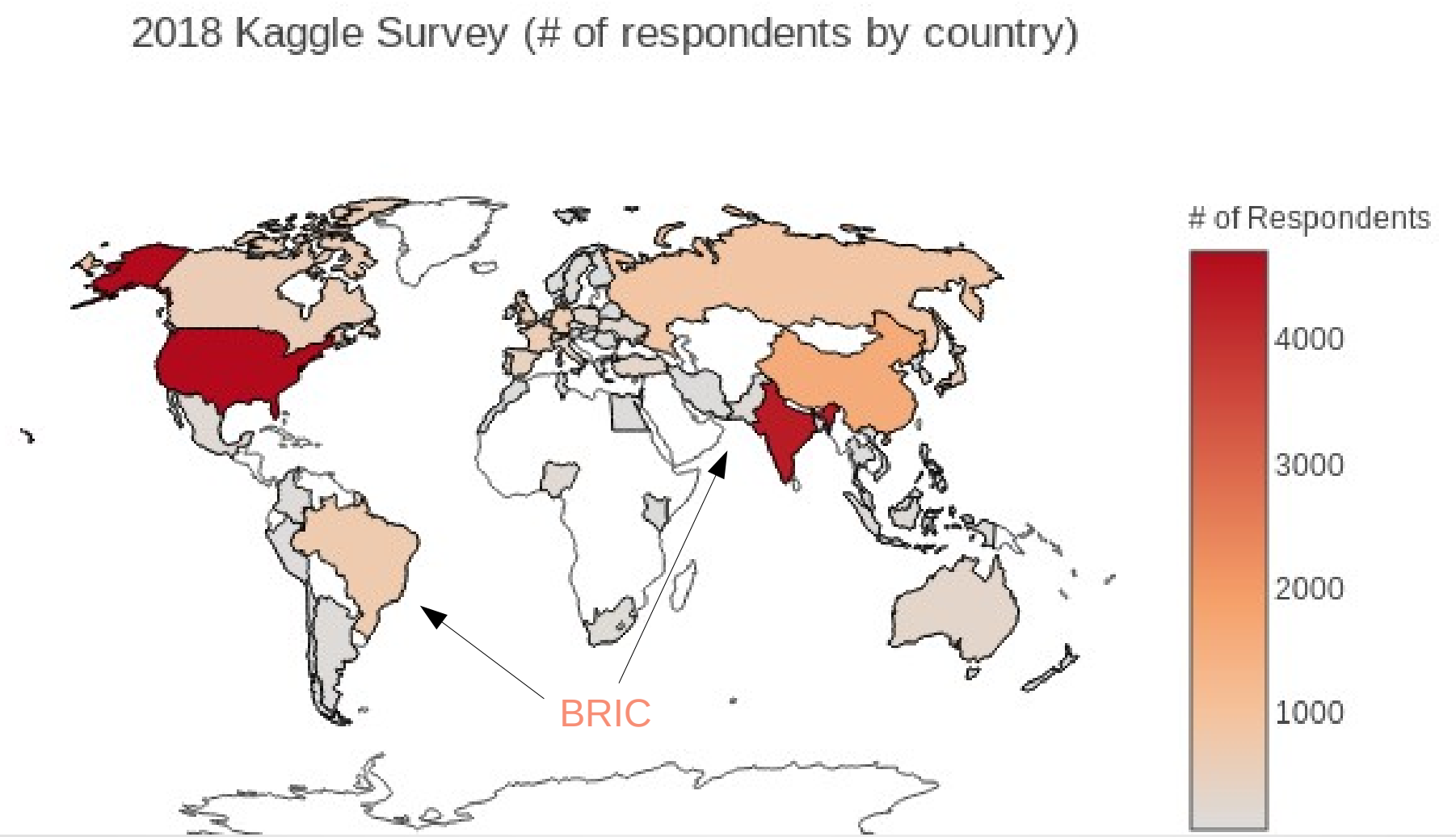
edX

kaggle

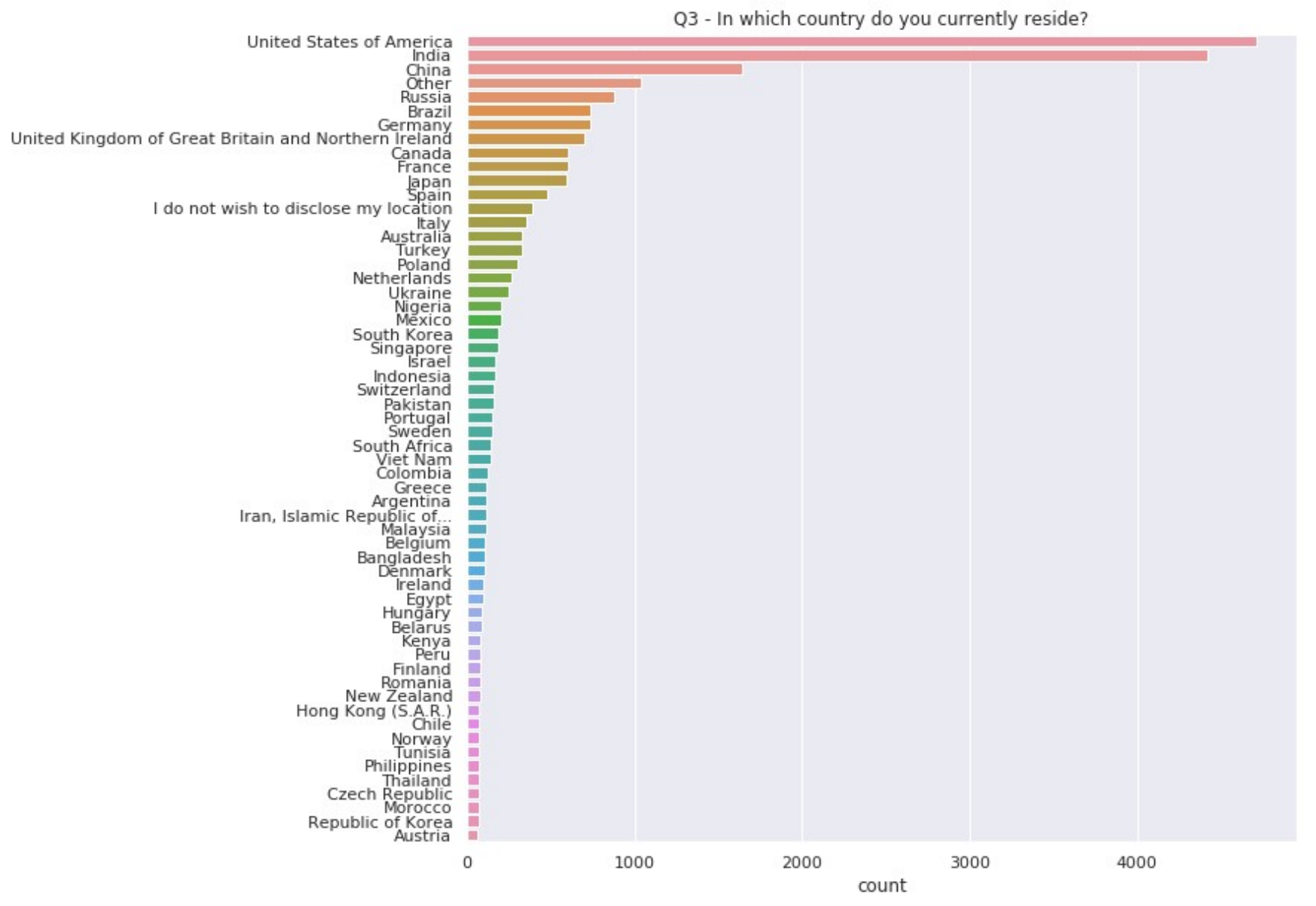
DataCamp

Udemy

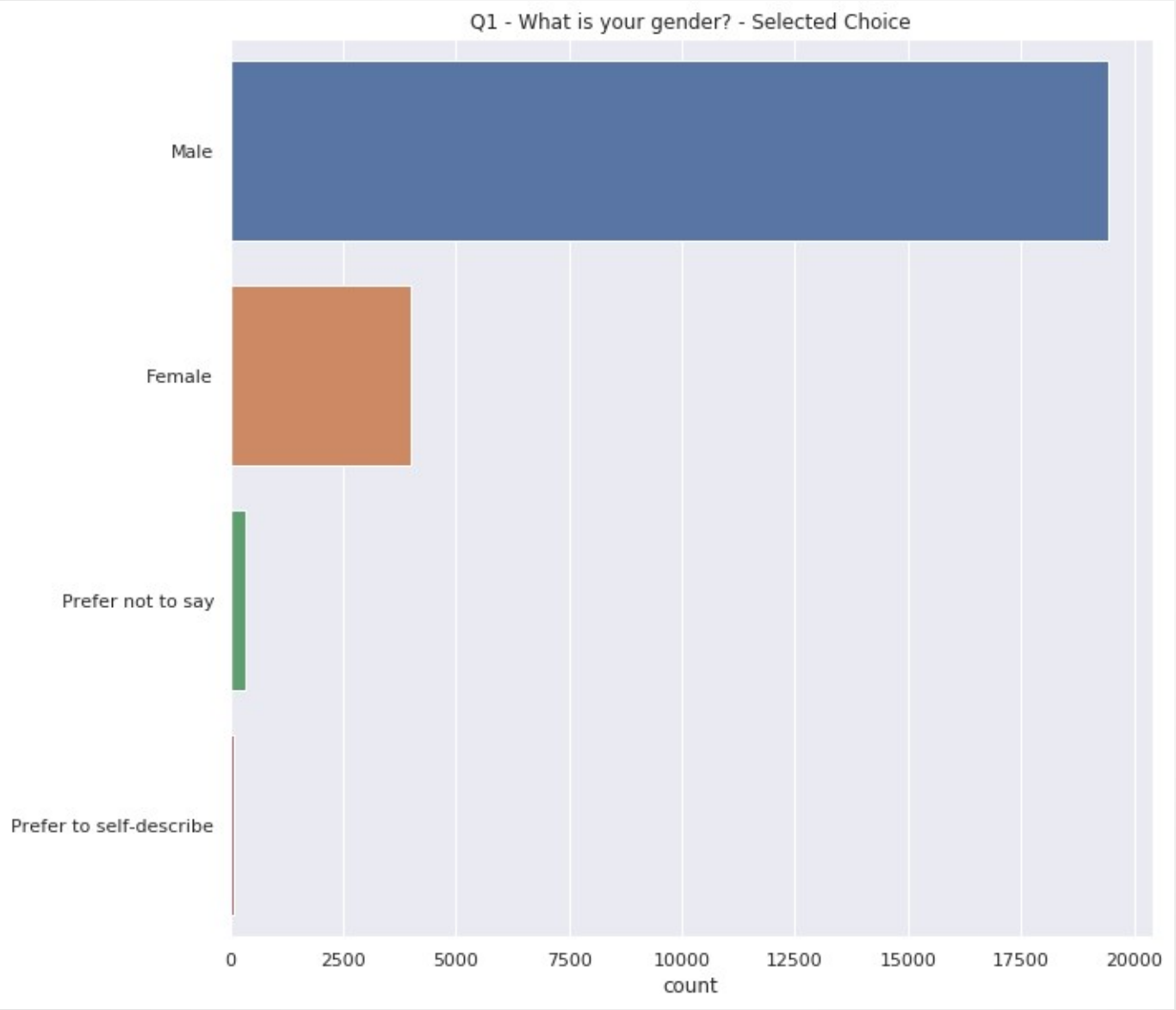
Locations of Survey Respondents (Map):



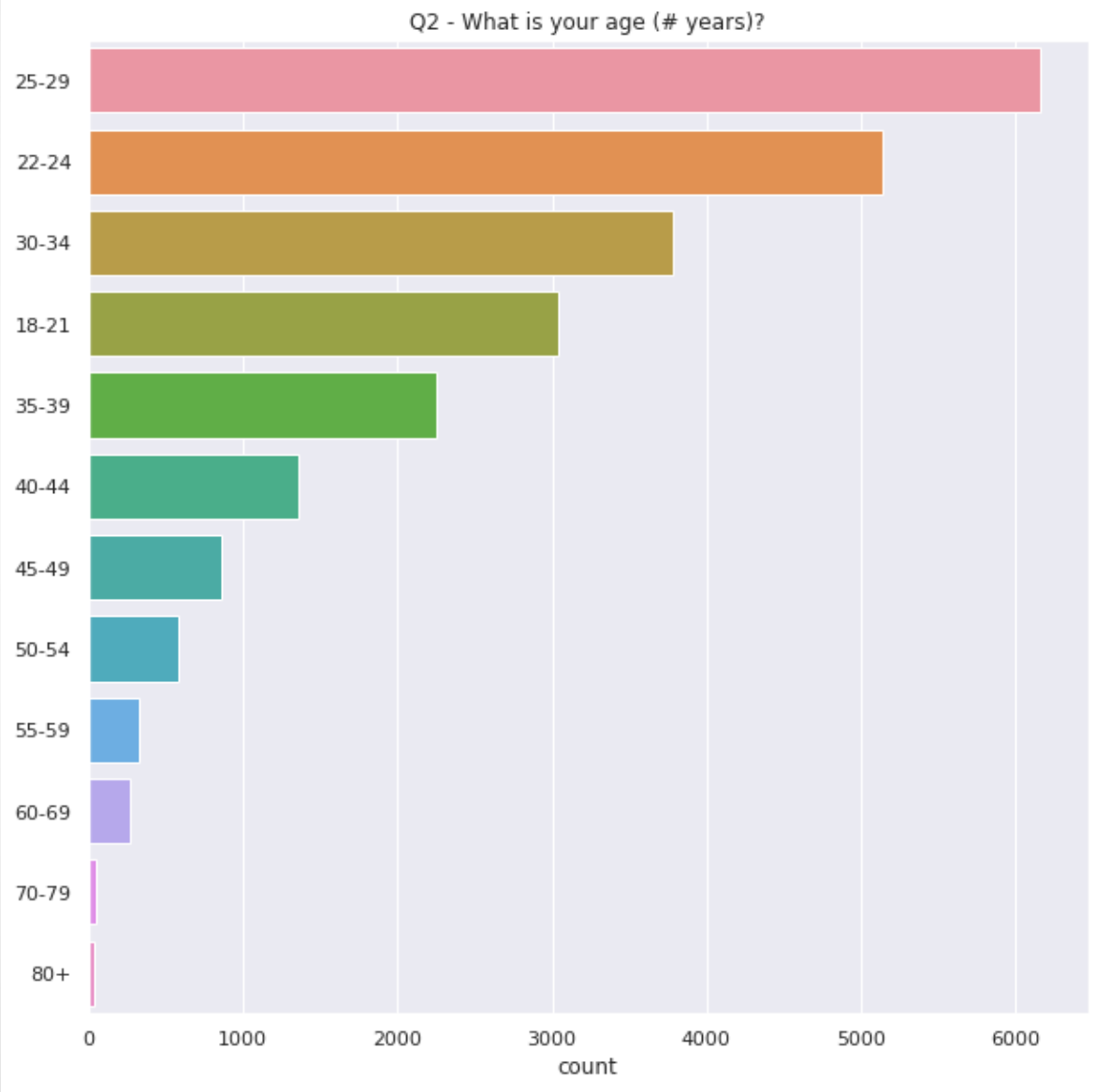
Locations of Survey Respondents (Chart):



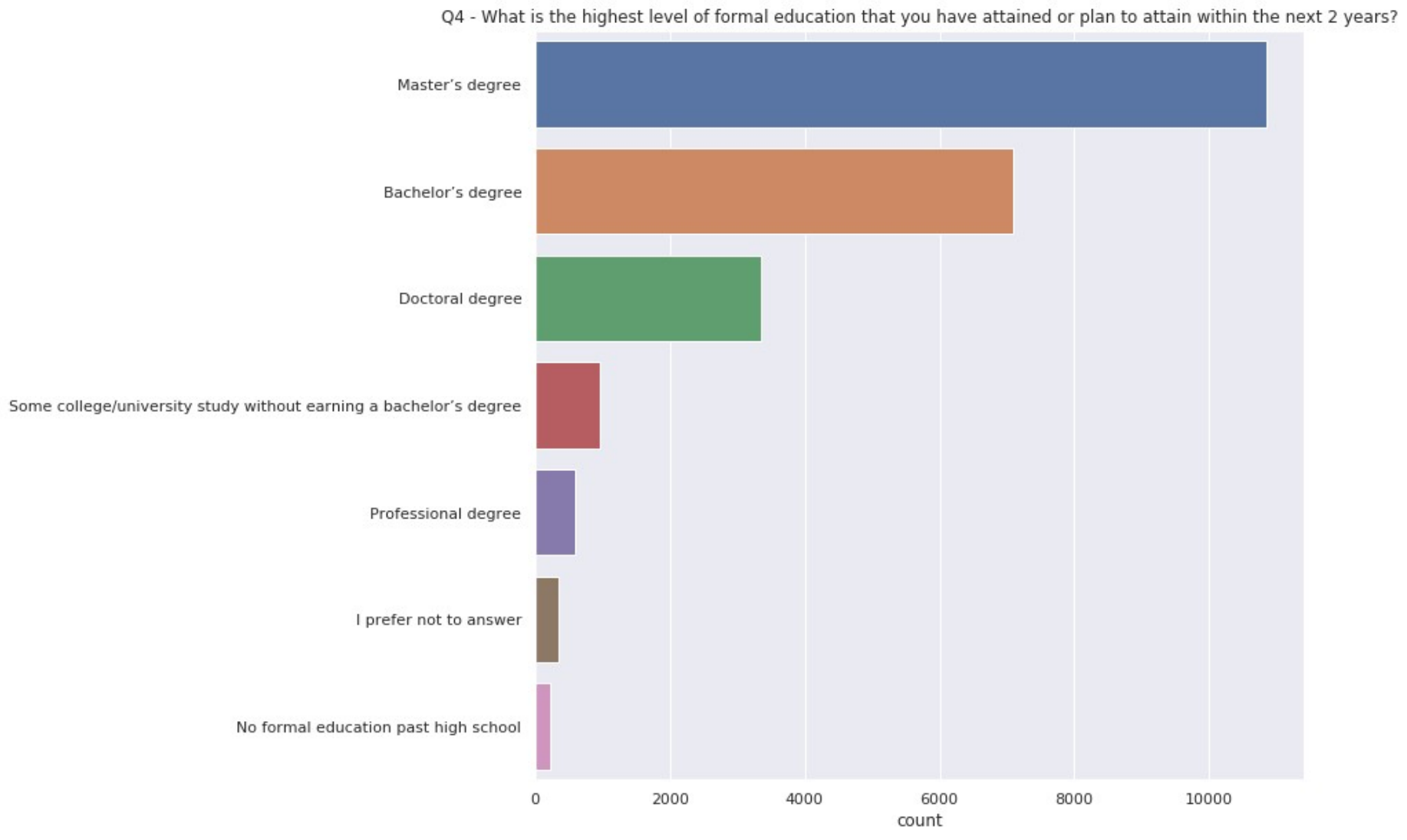
Gender:


Q1 - What is your gender? - Selected Choice
Male
Female
Prefer not to say
Prefer to self-describe
0
2500
5000
7500
10000
12500
15000
17500
20000
count

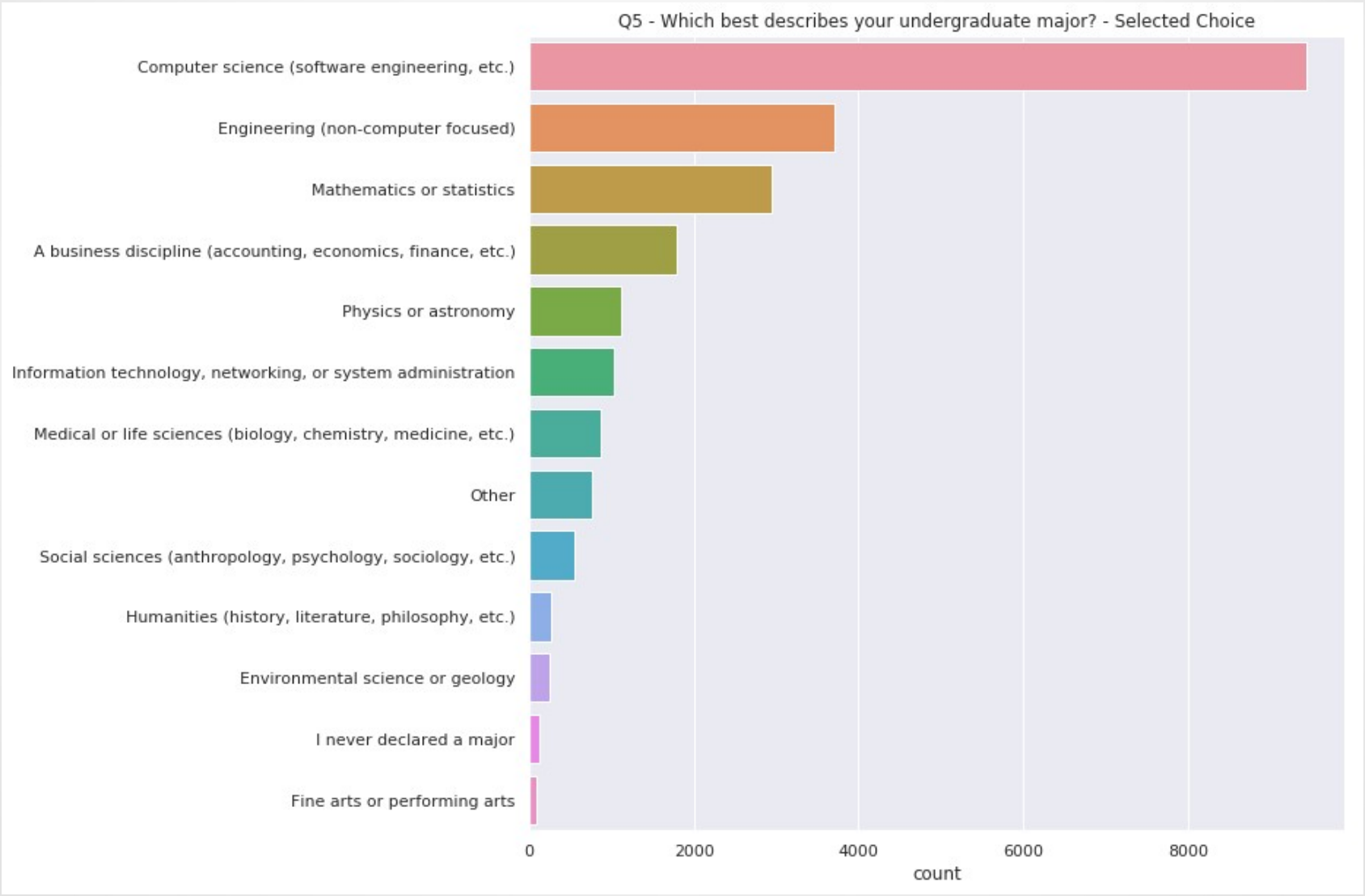
Age:


Q2 - What is your age (# years)?
25-29
22-24
30-34
18-21
35-39
40-44
45-49
50-54
55-59
60-69
70-79
80+
0
1000
2000
3000
4000
5000
6000
count

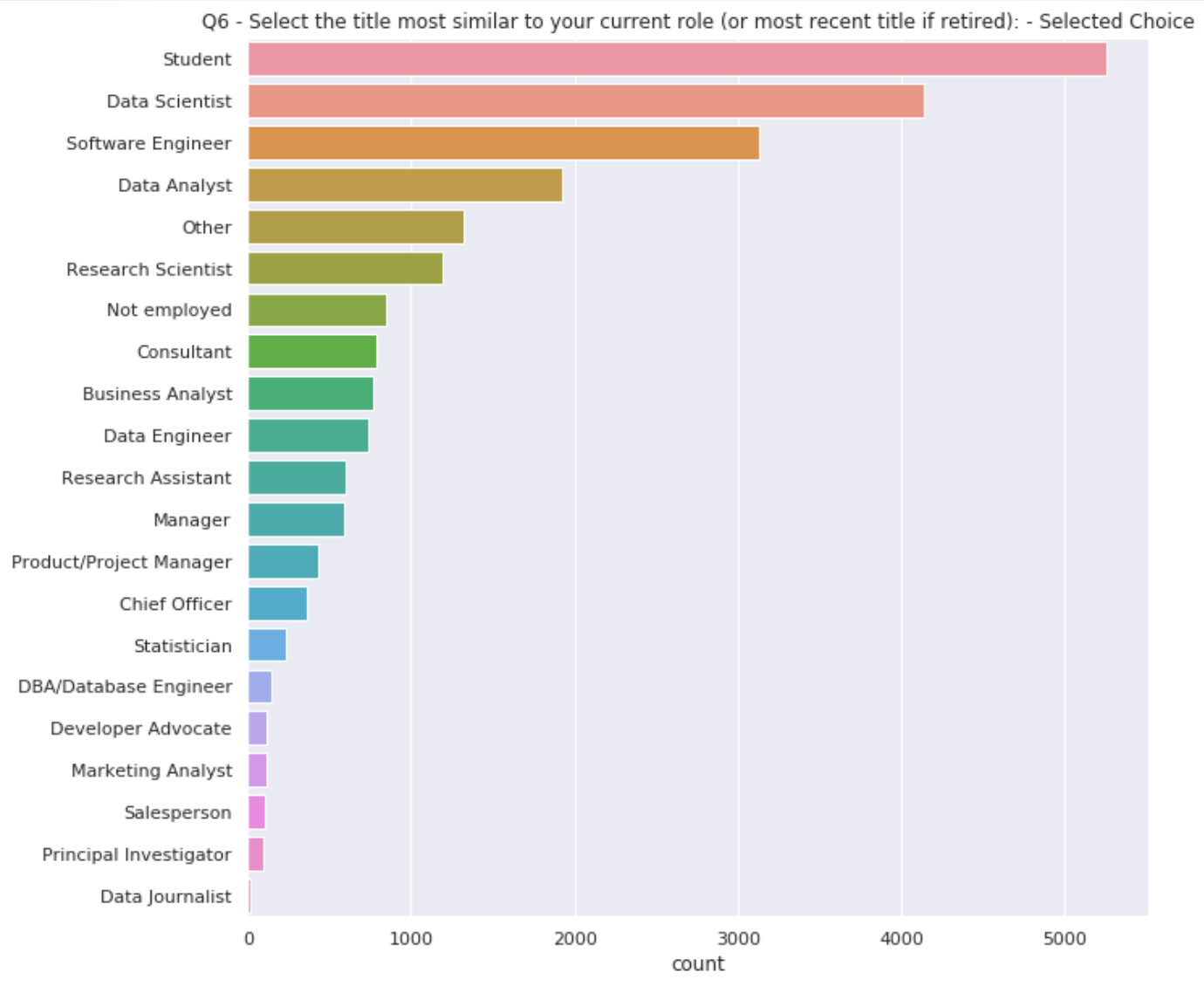
Formal Education (degree):



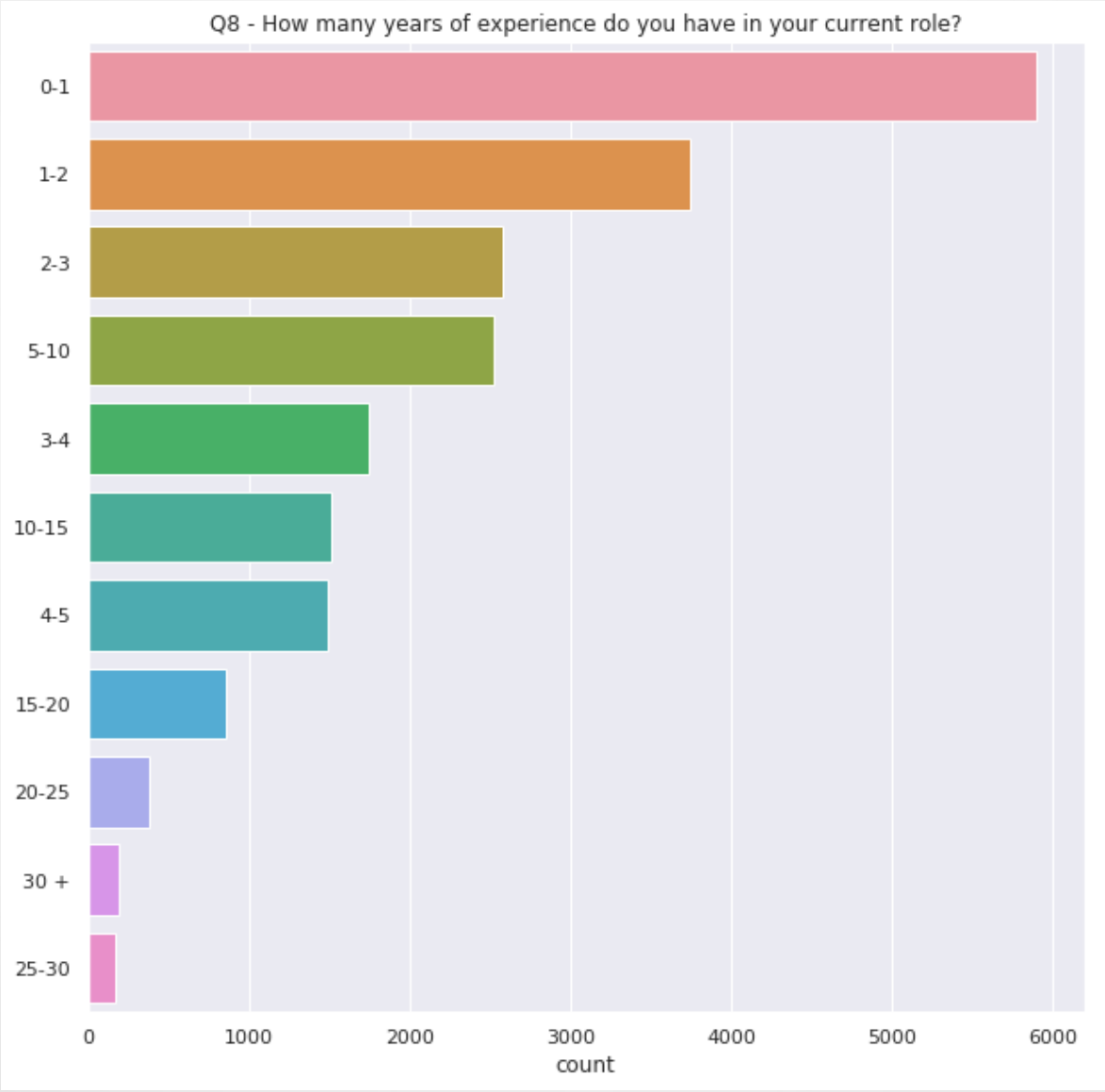
Formal Education (undergrad major):



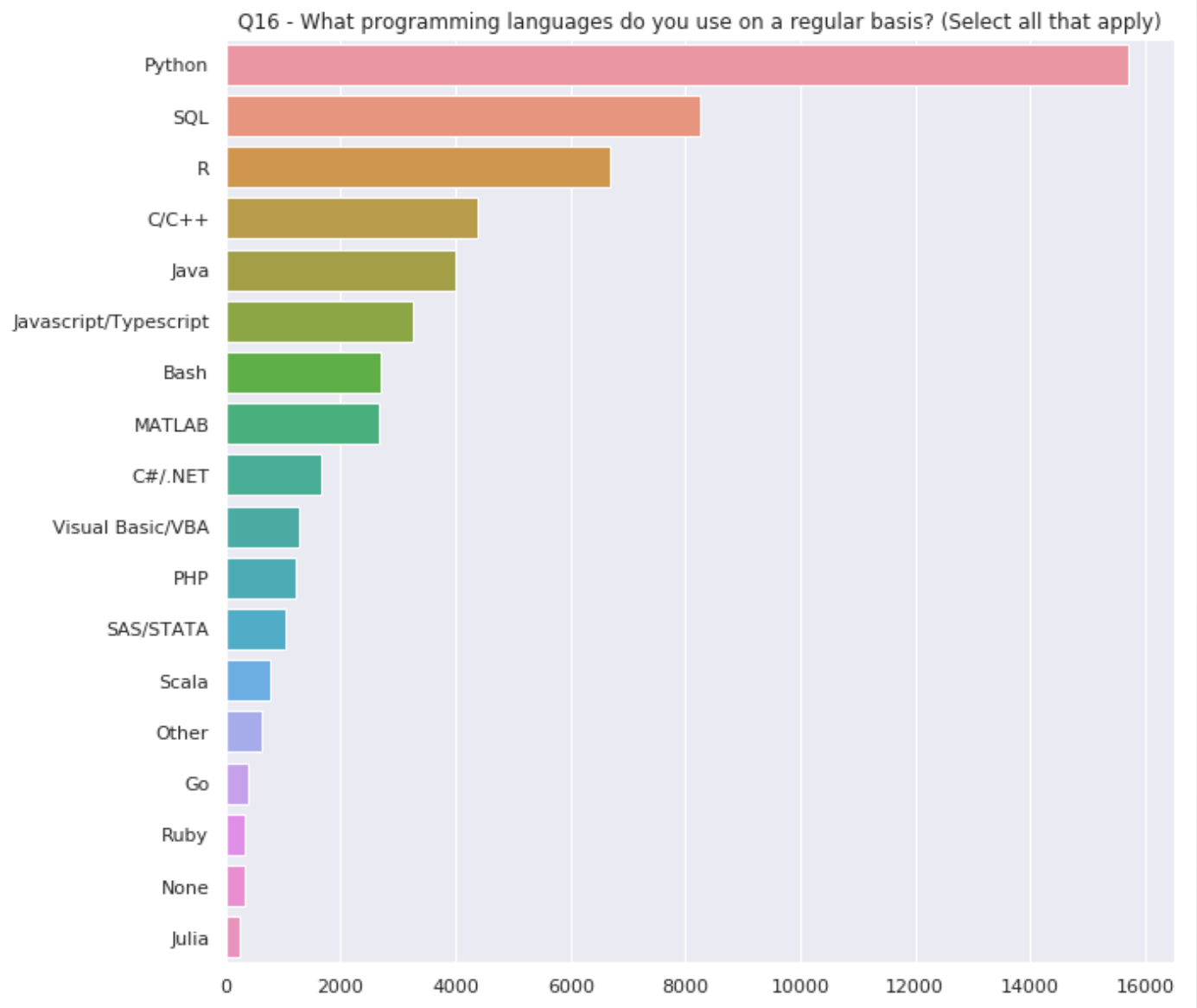
Job Roles:


Q6 - Select the title most similar to your current role (or most recent title if retired): - Selected Choice
Student
Data Scientist
Software Engineer
Data Analyst
Other
Research Scientist
Not employed
Consultant
Business Analyst
Data Engineer
Research Assistant
Manager
Product/Project Manager
Chief Officer
Statistician
DBA/Database Engineer
Developer Advocate
Marketing Analyst
Salesperson
Principal Investigator
Data Journalist
0
1000
2000
3000
4000
5000
count

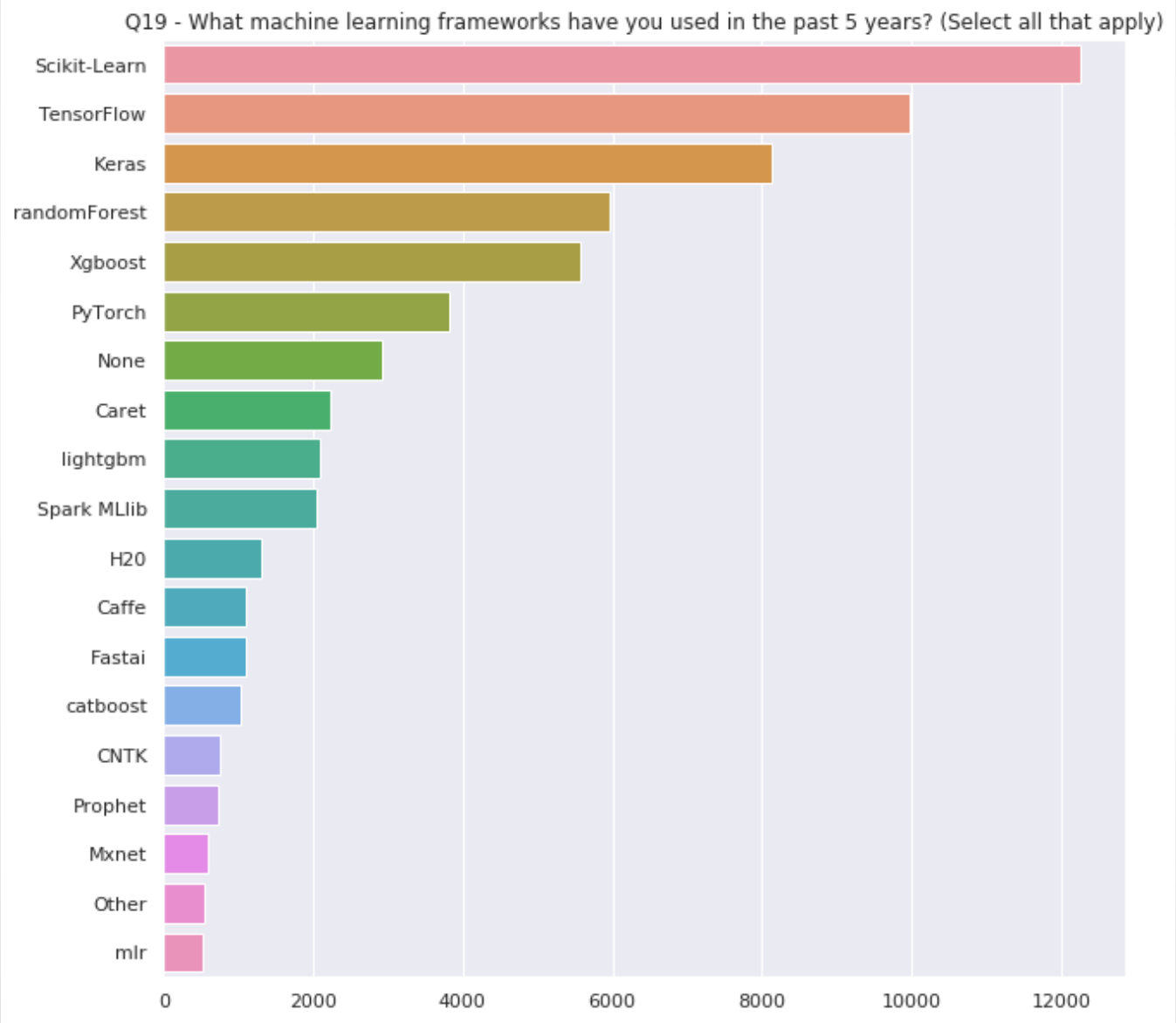
Experience in Job Role (years):



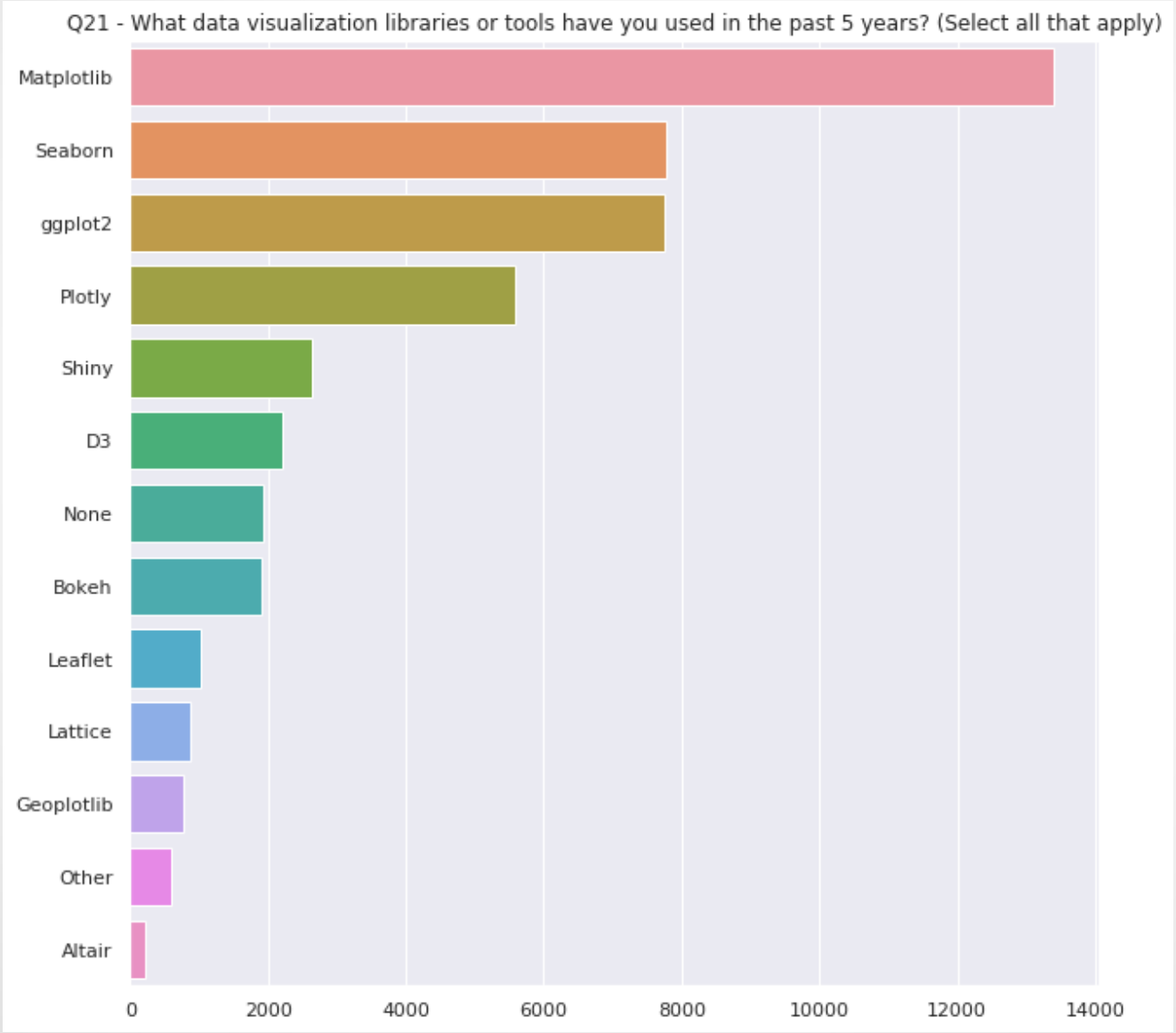
Experience w/ Programming Languages:



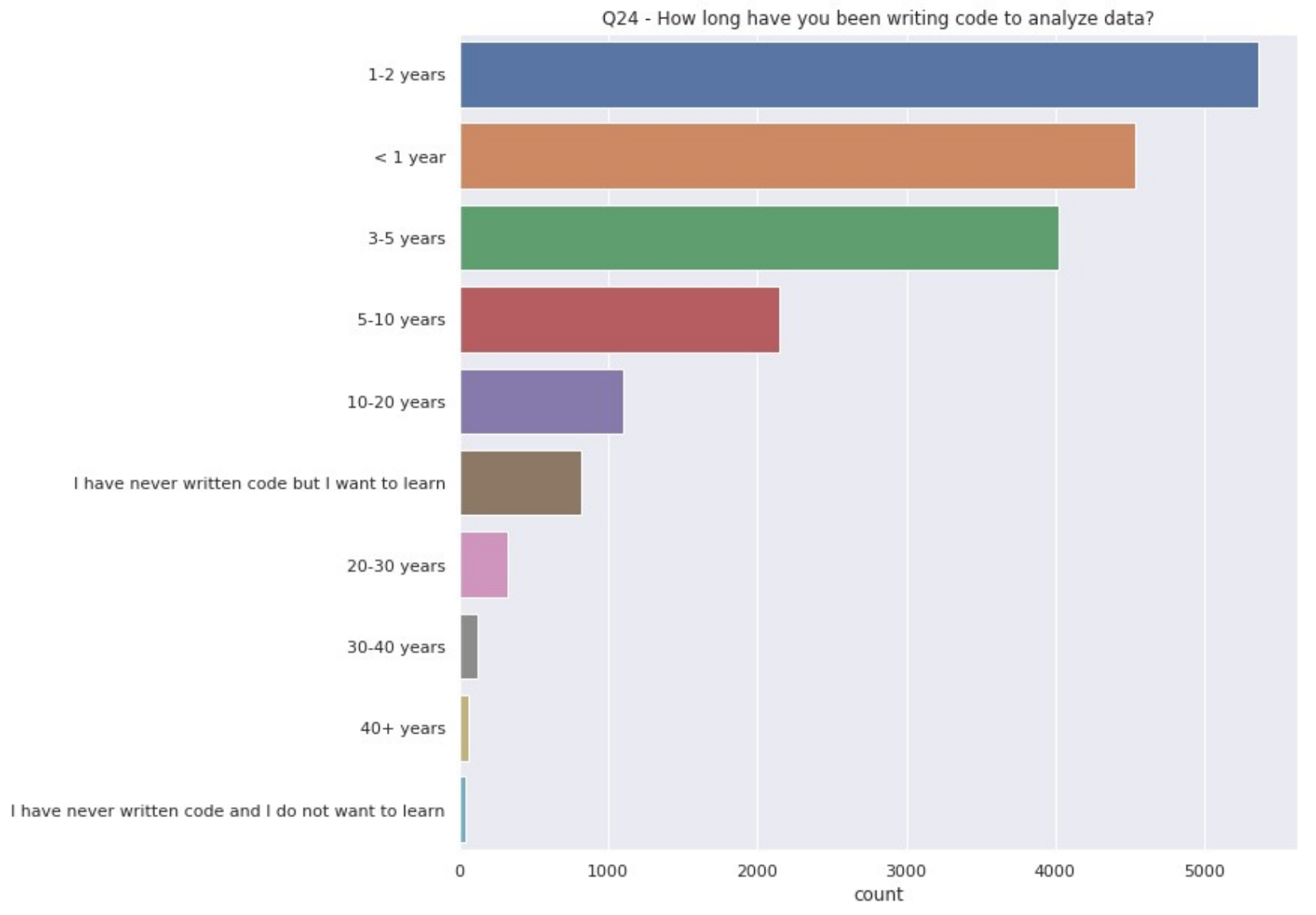
Experience w/ ML Frameworks:



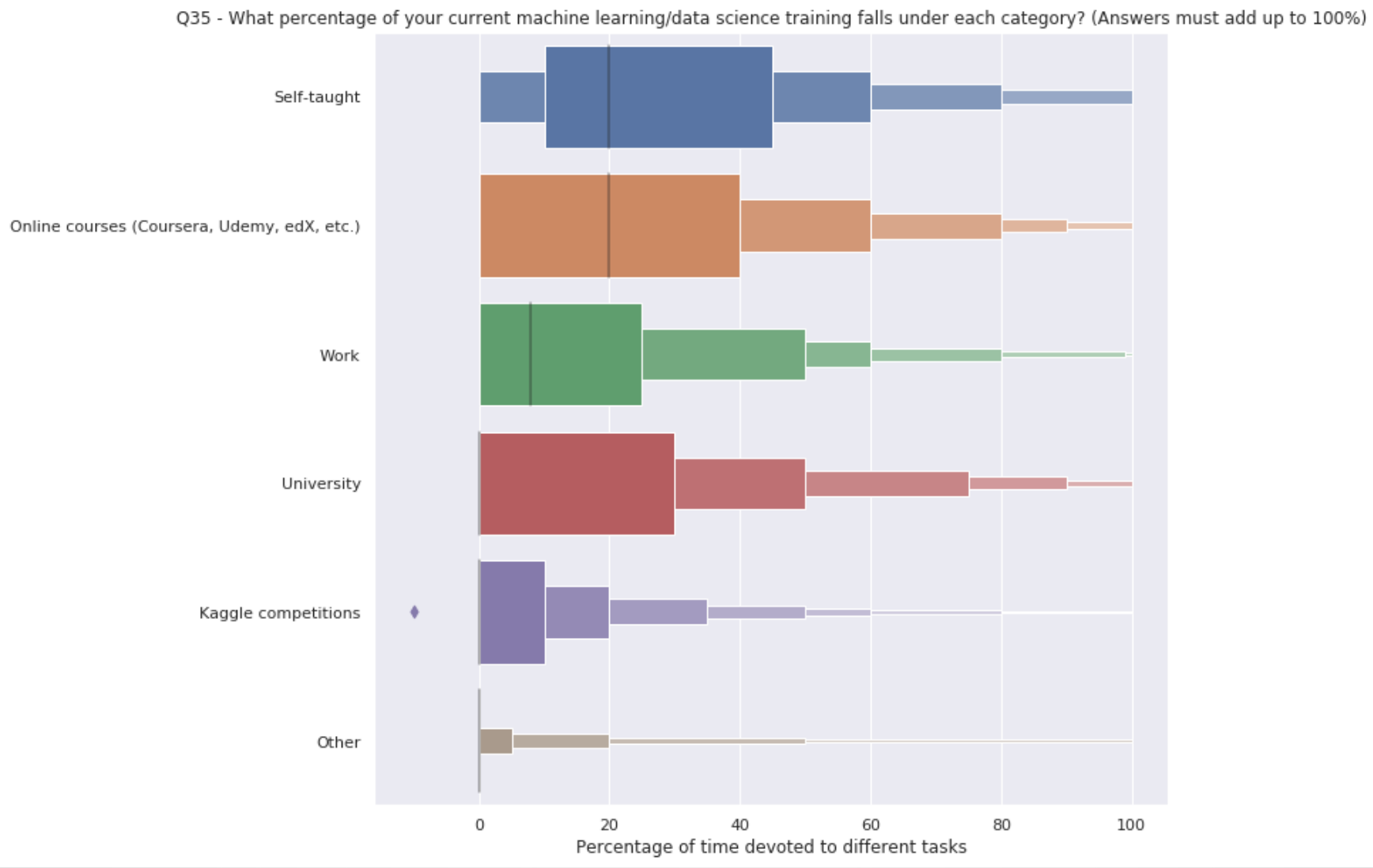
Experience w/ Data Visualization Tools:



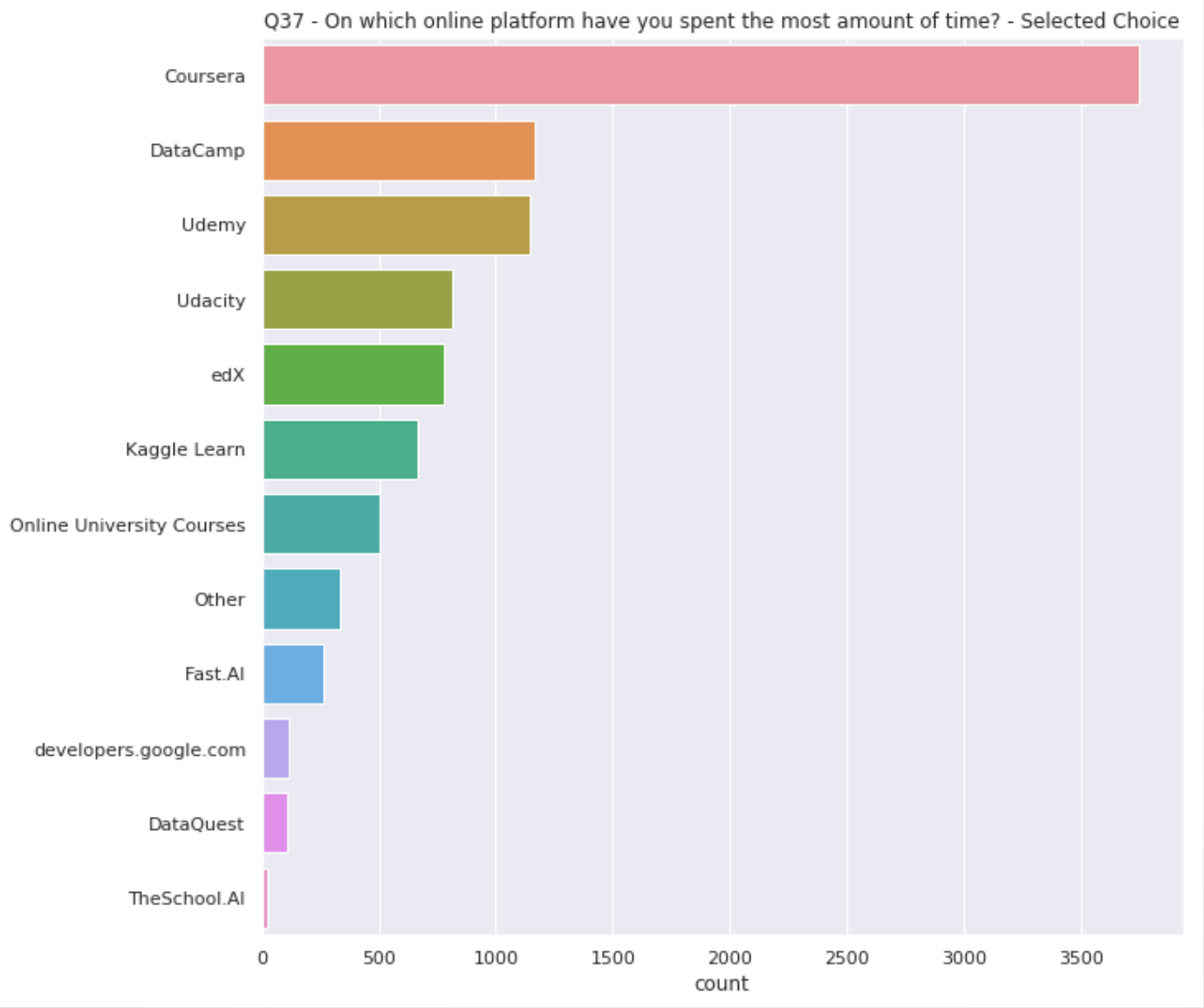
Experience w/ Programming for Data Analysis:



Training Ingredients for ML & DS

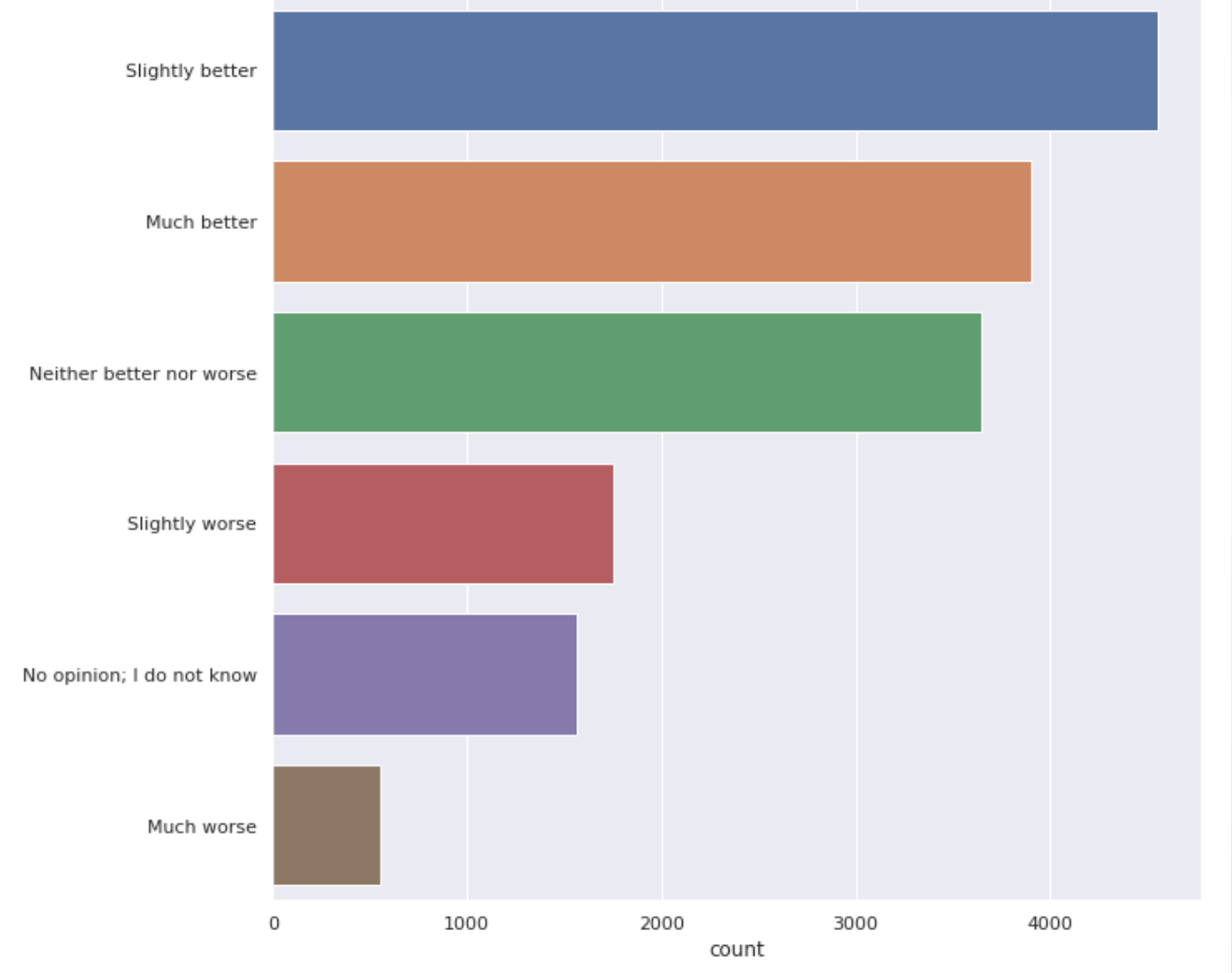


Online Training Platforms:



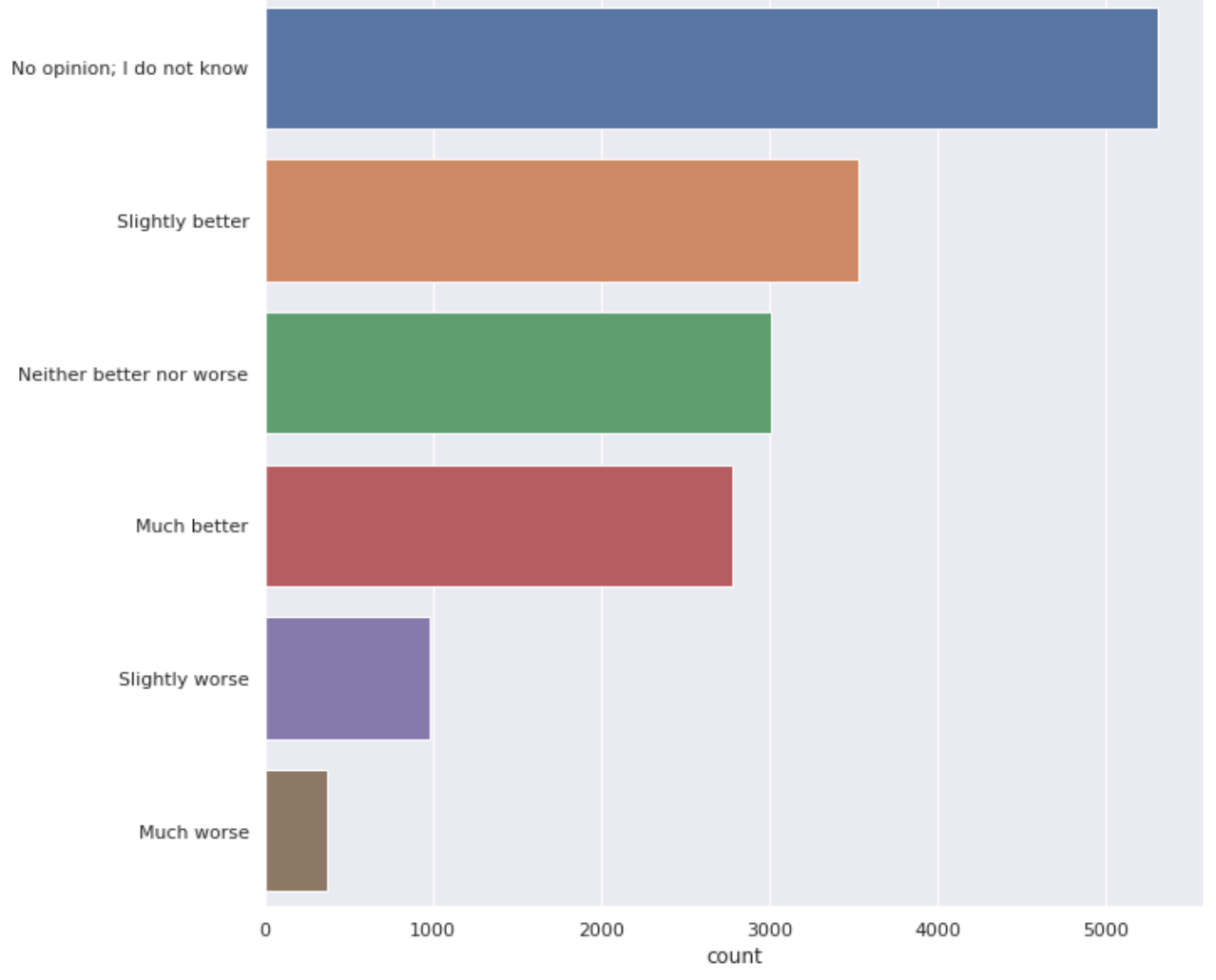
Perceived Quality of Non-Traditional Education (Online):

Online Learning / MOOCs Vs. Traditional Brick & Mortar Institutions



Perceived Quality of Non-Traditional Education (Offline):

In-person Bootcamps Vs. Traditional Brick & Mortar Institutions

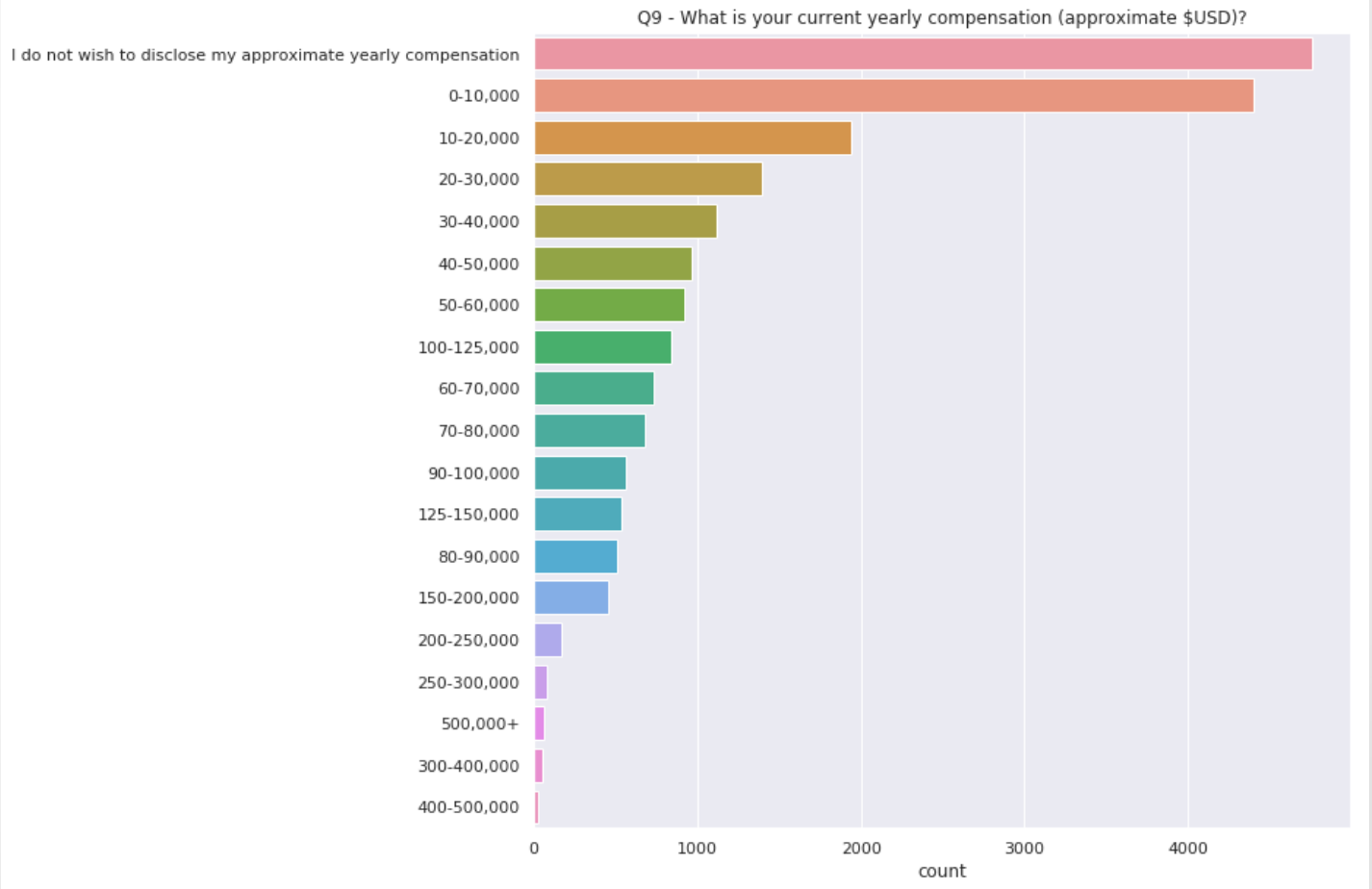


Supervised Learning:

- **<u>Outcome Variable for Supervised Learning Models (Income Tiers – Multi-Class Classification)</u>**

- *Initial # of rows in data – 23,859*
- *After dropping nulls – 14,576*
- *After dropping respondents that did not disclose income – 11,644*
- *Income tier groupings*
  - *Inside US – 2,645 rows*
  - *Outside US – 8,999 rows*
- *Income variable is categorical*
- *Chi-square test to test null hypothesis*
  - Critical value = 27.587
  - Chi-square statistic exceeds critical value
  - Reject null hypothesis
- *SL models tested on:*
  - Initial 18 categories
  - Regrouping by 7 categories

All Respondents



Supervised Learning Models:

- **(4) Models:**
  - Logistic Regression (Multinomial)
  - Random Forest
  - Gradient Boosting
  - Neural Network
    - fastai.tabular

- **Location:**
  - Inside US – 2,645 rows
  - Outside US – 8,999 rows

- **# of Income Tiers:**
  - Initial 18 Groupings
  - Adjusted 7 Groupings

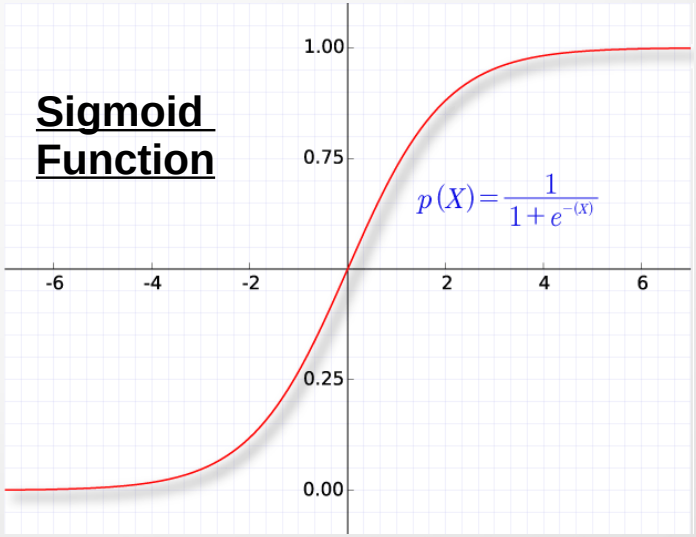
Multinomial Logistic Regression:

- **Income tier (dependent variable)**
  - More than two (2) possible outcomes
    - 18 income tiers
    - 7 income tiers
  - Model converts from *multinomial* to multiple *binomial* logistic regression problems; generates probabilities for:
    - Income Tier 1 (1) vs. Rest (0)
    - Income Tier 2 (1) vs. Rest (0)
    - Income Tier 3 (1) vs. Rest (0)
    - Etc.
  - For each respondent, the model predicts an income tier with the highest probability from those generated

Income Tiers

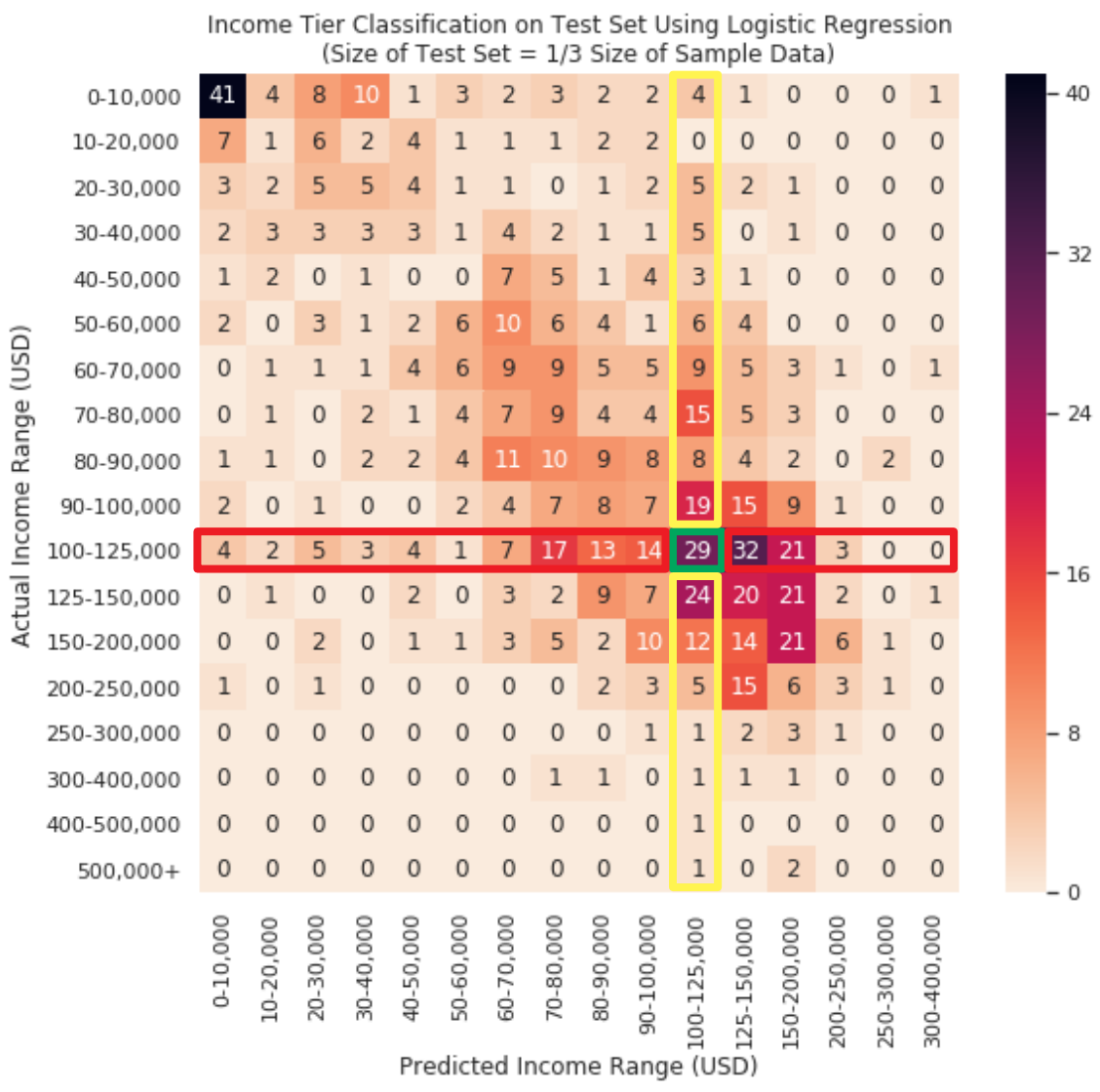
```
In [60]: salary_tiers = {
             '0-10,000': '0-49,999',  # Tier 1
             '10-20,000': '0-49,999',
             '20-30,000': '0-49,999',
             '30-40,000': '0-49,999',
             '40-50,000': '0-49,999',
             '50-60,000': '50,000-99,999',  # Tier 2
             '60-70,000': '50,000-99,999',
             '70-80,000': '50,000-99,999',
             '80-90,000': '50,000-99,999',
             '90-100,000': '50,000-99,999',
             '100-125,000': '100,000-149,999',  # Tier 3
             '125-150,000': '100,000-149,999',
             '150-200,000': '150,000-199,999',  # Tier 4
             '200-250,000': '200,000-249,999',  # Tier 5
             '250-300,000': '250,000-299,999',  # Tier 6
             '300-400,000': '300,000+',  # Tier 7
             '400-500,000': '300,000+',
             '500,000+': '300,000+'
         }
```

Sigmoid Function

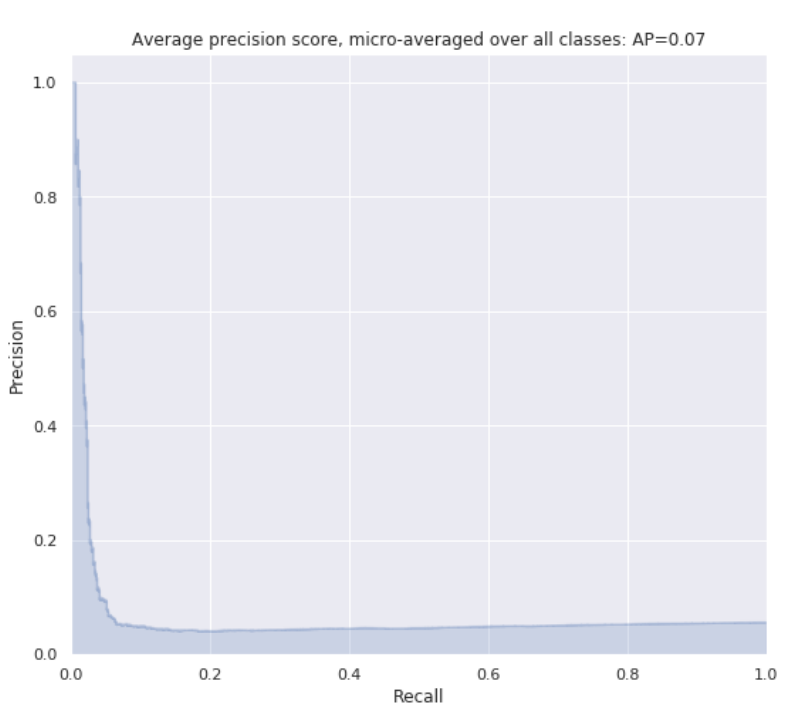


Multinomial Logistic Regression:

- **Predicted Incomes (Inside US) – Worst Model**



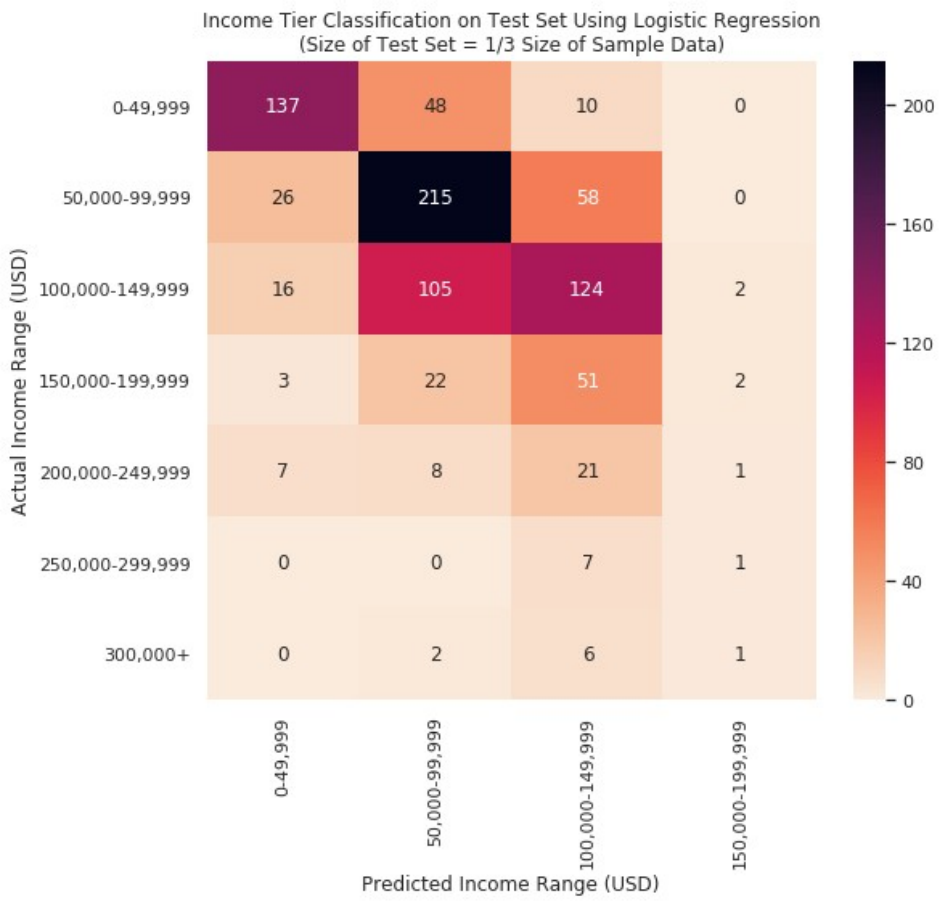
- **18 Income Tiers (Without Regularization)**
- **Example Heatmap Annotations:**
  - **True Positives**
  - **False Positives**
  - **False Negatives**

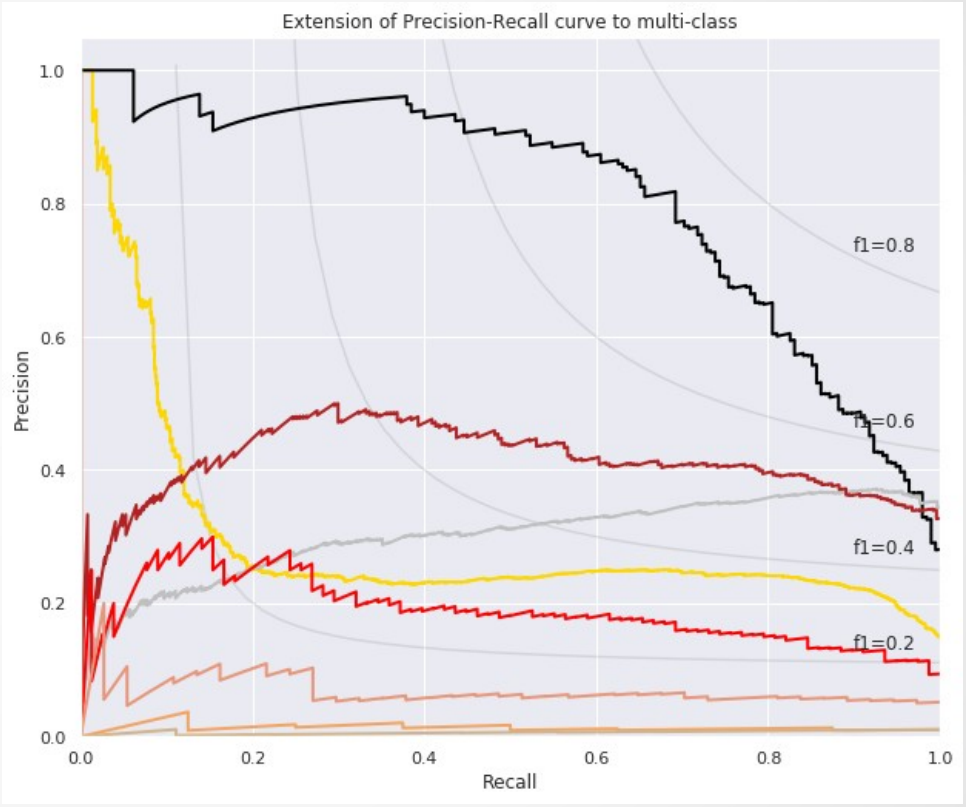


Multinomial Logistic Regression:

- **Predicted Incomes (Inside US) – Best Model**

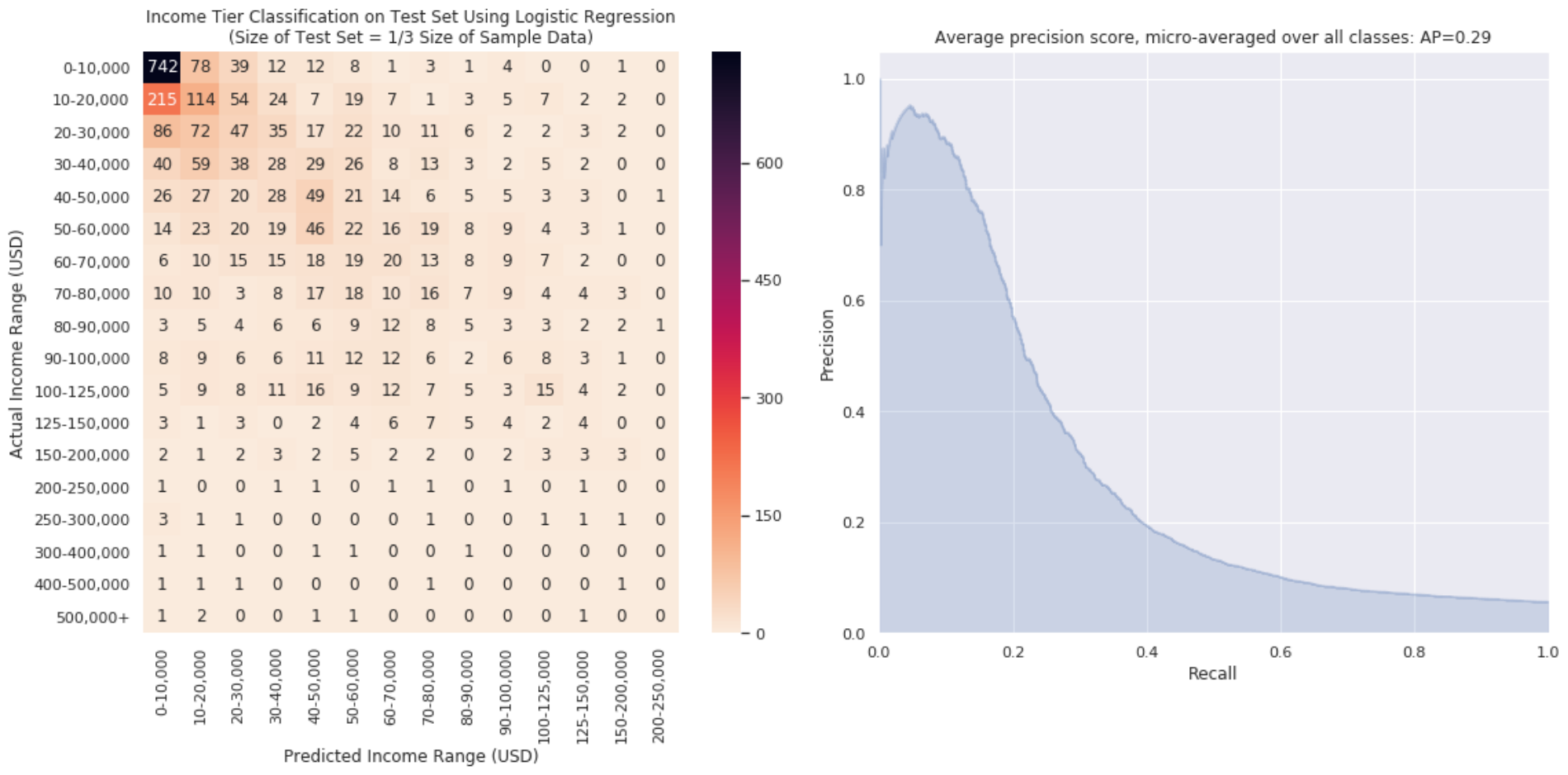
  - **7 Income Tiers (With L2 Regularization)**





Multinomial Logistic Regression:

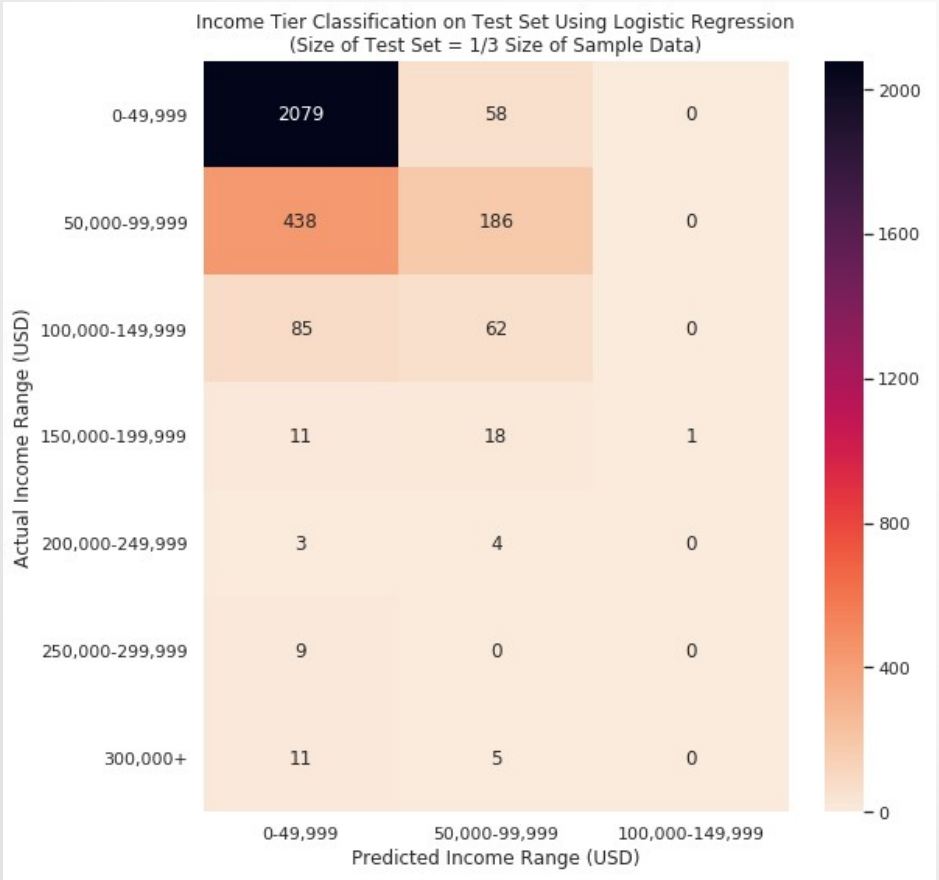
- **Predicted Incomes (Outside US) – Worst Model**
  - **18 Income Tiers (Without Regularization)**

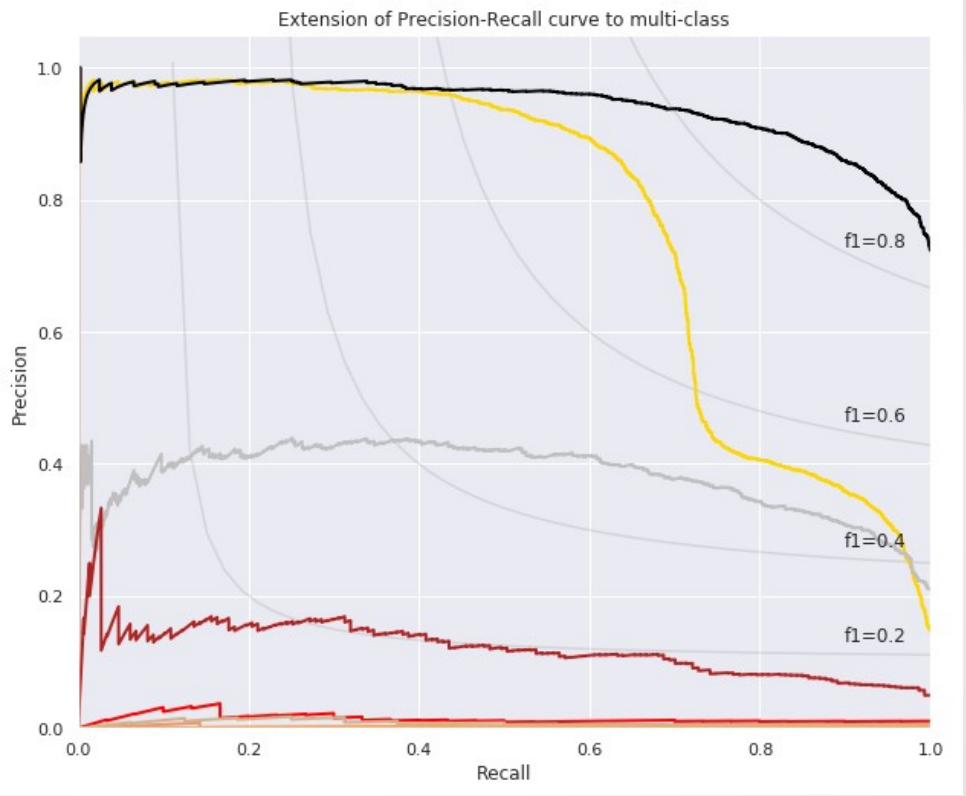


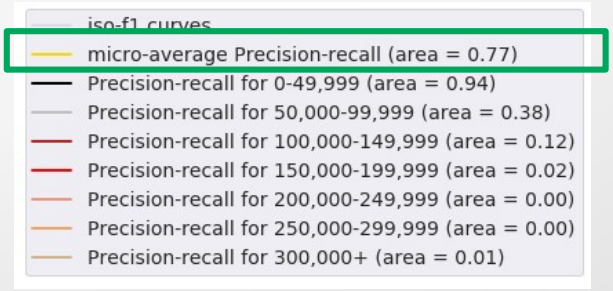
Multinomial Logistic Regression:

- **Predicted Incomes (Outside US) – Best Model**

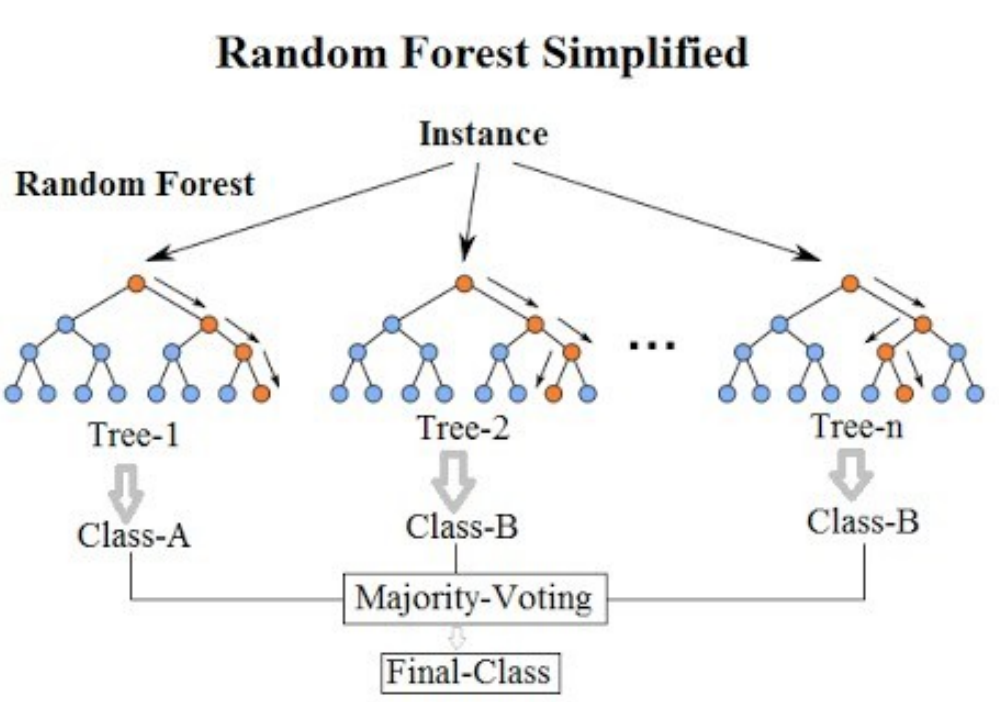
  - **7 Income Tiers (With L2 Regularization)**







Random Forest:

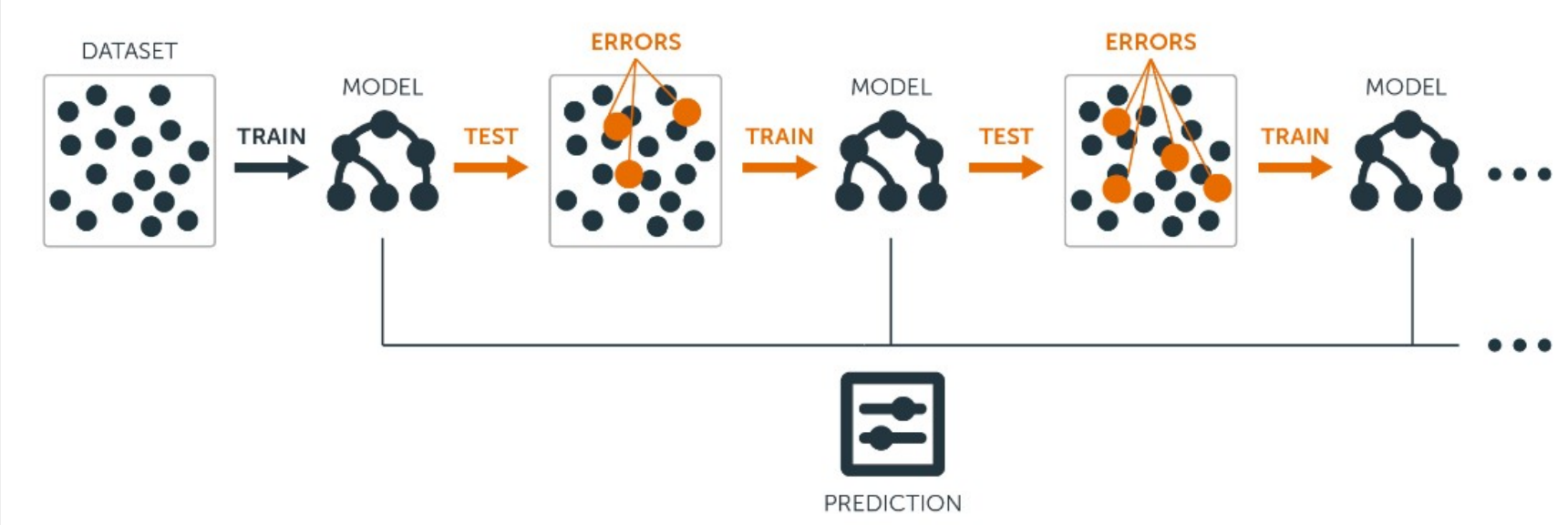


- A r*andom forest* model is an *ensemble method* comprised of the simultaneous, independent results from multiple d*ecision trees*
- Individual decision trees are likely to overfit on part of the data
- Overfitting can be reduced by averaging or taking the majority of the results from multiple decision trees

RF - Best averages of cross-validation scores (5 folds)	18 Income Tiers n_estimators=300 max_depth=4	7 Income Tiers n_estimators=300 max_depth=4
Inside US	~0.2171	~0.5149
Outside US	~0.3039	~0.7213

Gradient Boosting:

Gradient Boosting Simplified

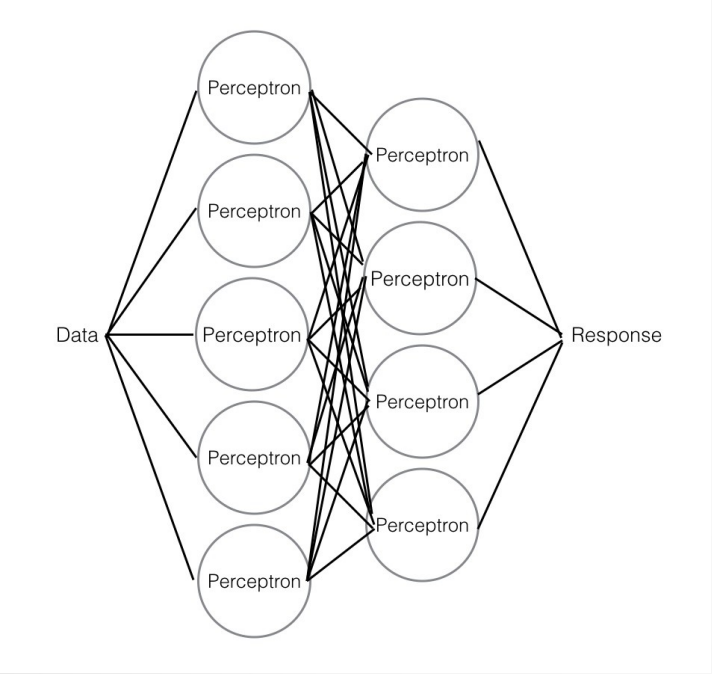


- A *gradient boosting* model is an *ensemble method* where each model in a sequence of models learns from the mistakes of previous models in the sequence
- Observations for successive model training are based on errors instead of bootstrap sampling

GBDT - Best averages of cross-validation scores (5 folds)	18 Income Tiers n_estimators=300 learning_rate=0.1 max_depth=2	7 Income Tiers n_estimators=300 learning_rate=0.1 max_depth=2
Inside US	~0.2280	~0.5395
Outside US	~0.3672	~0.7883

Neural Network:

Neural Network Simplified



- Input layer of features
- Hidden layers (1 or more)
  - Multiple perceptrons in each hidden layer
  - Activation function for a perceptron allows for a binary or continuous output (relu, sigmoid, tanh)
- Each feature from input layer connected to each perceptron in the 1st hidden layer with different weightings
- Each perceptron in each hidden layer is connected to other perceptrons in the next hidden layer OR the output layer with different weightings
- Output layer

Using fastai.tabular - Best accuracy after 10 epochs	18 Income Tiers epochs=10 learning_rate=0.05 layers=[200,100]	7 Income Tiers epochs=10 learning_rate=0.05 layers=[200,100]
Inside US	~0.9783	~0.9447
Outside US	~0.9913	~0.9747

Summary:

- Some themes in the data:
  - Youth
  - Python ecosystem
  - Possible disruption in education
- Supervised Learning – Predicting Income Tiers Inside / Outside US
  - Logistic Regression – Better than RF (7 Tiers – Outside US)
  - Random Forest – Better than LR (7 Tiers - Inside US)
  - Gradient Boosting – Better than LR & RF
  - Neural Net – Best
- More detailed results in the notebook