HackerRank Developer survey 22.05.2023

Aleksander Blok, Marcel Boxberger, Rafał Curyło, Wiktor Czetrybok, Marcel Dąbrowski

1. General description of the data set

The dataset consists of responses collected from a survey conducted in late 2016 among a global community of developers. With a total of 25,000 responses, the dataset provides valuable insights into various aspects of the developer community. It includes information about participants' skills, educational background, current roles, and more. The dataset encompasses both students and professionals, offering a comprehensive view of the developer landscape. This rich dataset serves as a valuable resource for understanding the characteristics and trends within the global developer community.

2. Discussion of data mining goals and success criteria

Data Mining goals:

- Identify trends and patterns: Explore the dataset to identify trends and patterns related to various aspects such as skills, educational background, roles, and more. This could involve analyzing the distribution of different variables, identifying correlations between variables, and uncovering hidden relationships within the data.
- Understand the developer community: Gain insights into the demographics and characteristics of the developer community. This could involve analyzing factors like age, gender, location, educational background, and employment status to understand the profile of developers

Success criteria:

- Accurate analysis: Ensure that the analysis of the dataset is performed accurately, taking into account relevant statistical techniques and methodologies.
- Validated findings: Verify the findings through appropriate statistical tests and validation techniques to ensure the reliability and robustness of the results.

3. Characteristics of the data set:

Origin: The dataset was created by HackerRank, a popular online platform for coding challenges and technical skill assessments. The data was collected through a survey conducted by HackerRank in 2018, targeting developers from various backgrounds.

Format: The dataset is available in CSV (Comma-Separated Values) format. CSV is a common file format used for tabular data, where each line represents a row, and the values within each line are separated by commas.

Number of samples: The exact number of samples in the dataset is 25091.

Single or multiple sets: The dataset consists of one set of data. It represents the responses collected from developers who participated in the HackerRank Developer Survey 2018. Each row in the dataset represents a separate response, and the columns represent different attributes or variables related to the survey questions.

4. Description of the attributes:

q1AgeBeginCoding:

Type: Categorical

Meaning: Age at which the respondent started coding

Unit of measurement: Not applicable (since it is a categorical variable)

q4Education:

Type: Nominal

Meaning: Highest level of education attained or planned to obtain

Special values: Only shown if the respondent chooses "Some College" or above as their education level

q23Frame:

Attribute Name: Frameworks

Type: Nominal (Categorical)

Meaning: The frameworks that the hiring manager looks for in potential candidates.

Special Values: None

Unit of Measurement: Not applicable (since it is a categorical variable)

q25Lang:

Attribute Name: Language

Type: Nominal (Categorical)

Meaning: Most used programming language

Special Values: None

Unit of Measurement: Not applicable (since it is a categorical variable)

q28LoveLang:

Attribute Name: Loved Language

Type: Nominal (Categorical)

Meaning: Most loved programming language

Special Values: None

Unit of Measurement: Not applicable (since it is a categorical variable)

q13EmpMeas:

Attribute Name: Measurement Methods

Type: Nominal (Categorical)

Meaning: The methods used by employers to assess the skills of the job seeker.

Special Values: None

Unit of Measurement: Not applicable (since it is a categorical variable)

q30LearnCode:

Attribute Name: Learning Resources

Type: Nominal (Categorical)

Meaning: The additional resources used by individuals to practice and learn coding.

Special Values: None

Unit of Measurement: Not applicable (since it is a categorical variable)

country

Attribute Name: CountryNumeric2

Type: Nominal (Categorical)

Meaning: The country of origin for the respondent.

Unit of Measurement: Not applicable (since it is a categorical variable)

Special Values: The values for this attribute correspond to country codes based on the "Country-Code-mapping.csv" file. Each value represents a specific country.

q12JobCrit

Attribute Names: Job criteria that people look for in a company.

Type: Nominal (Categorical)

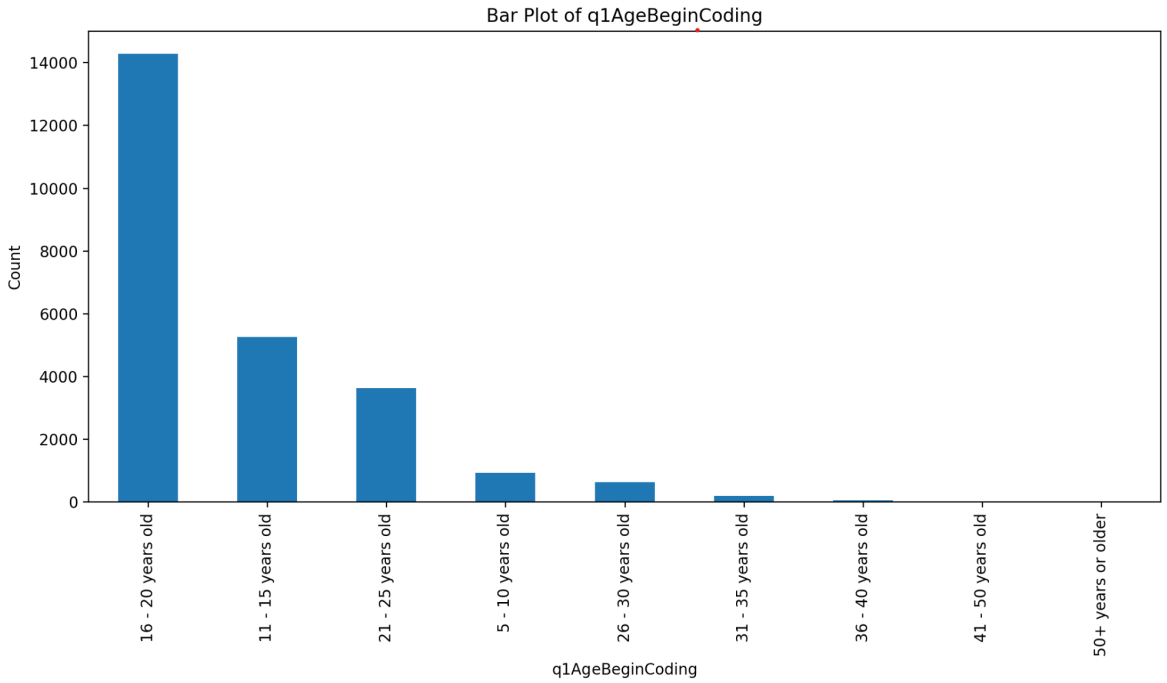
Meaning: The top 3 criteria that respondents consider important when evaluating job opportunities.

Unit of Measurement: Not applicable (since it is a categorical variable)

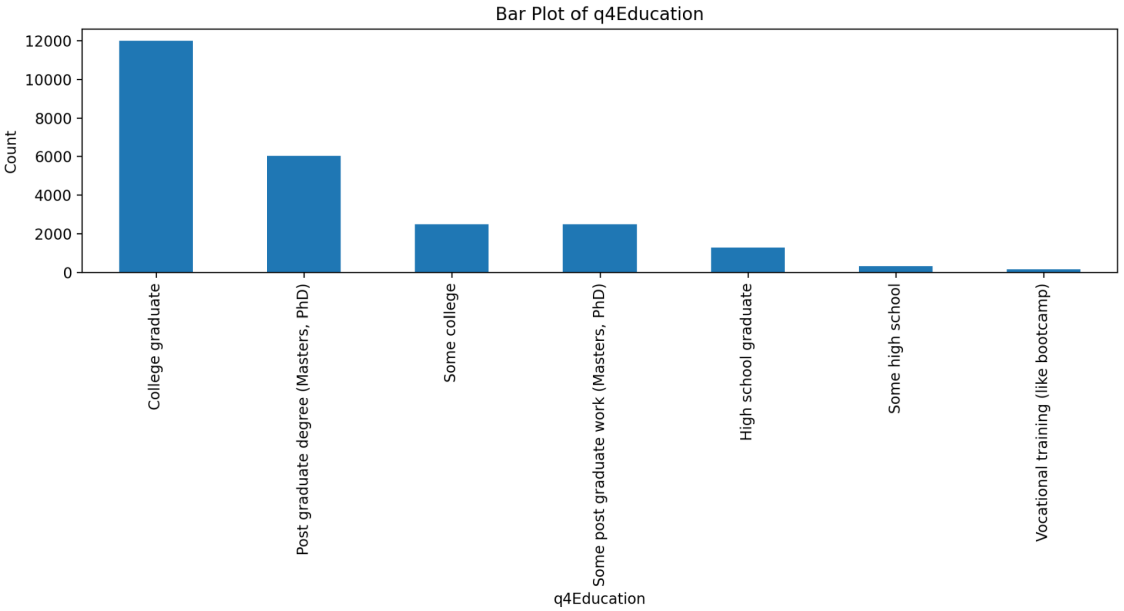
Special Values: Respondents are asked to choose 3 options from a predefined list of criteria. Each attribute represents a different aspect of a company that individuals prioritize when looking for job opportunities.

5. Exploratory data analysis:

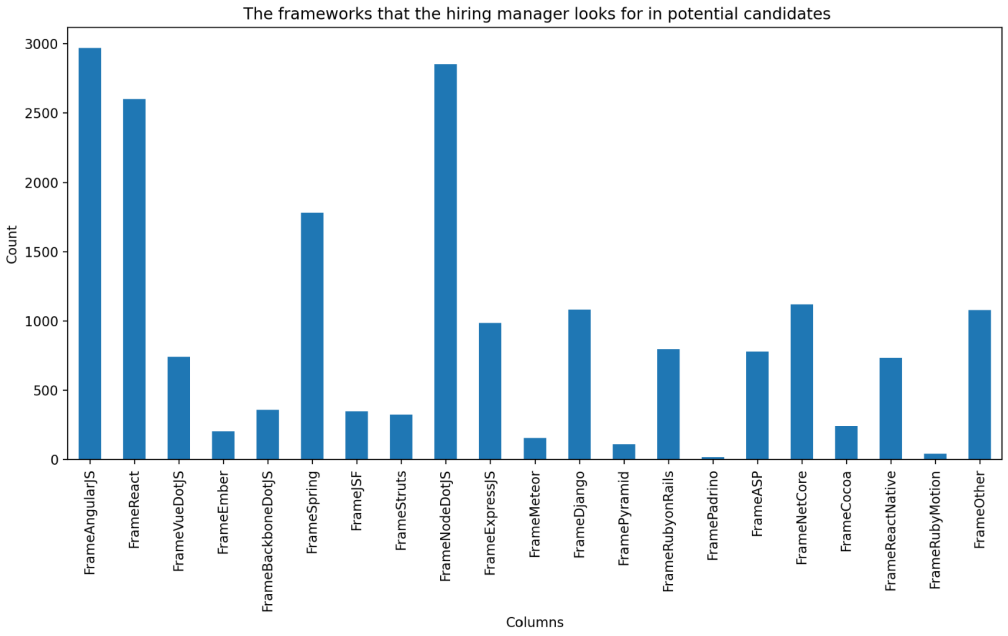
Age at which the respondent started coding



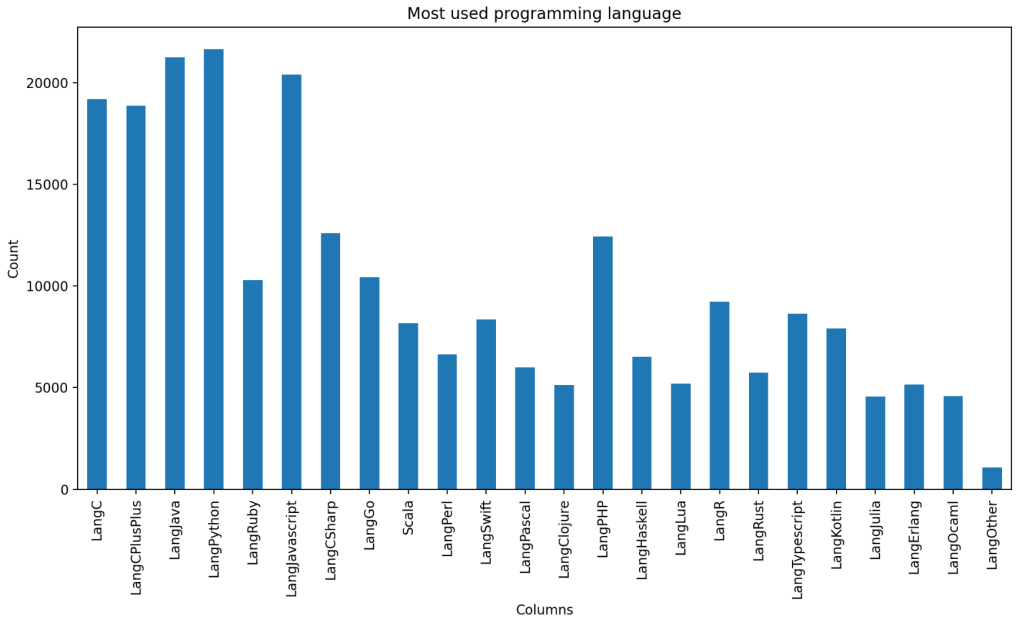
Highest level of education attained or planned to obtain



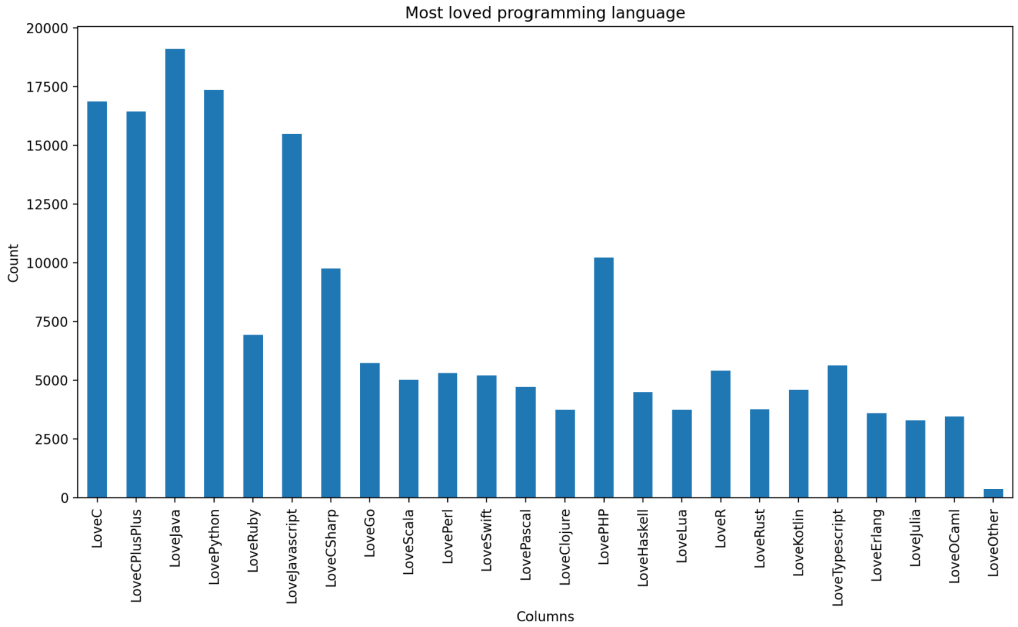
The frameworks that the hiring manager looks for in potential candidates



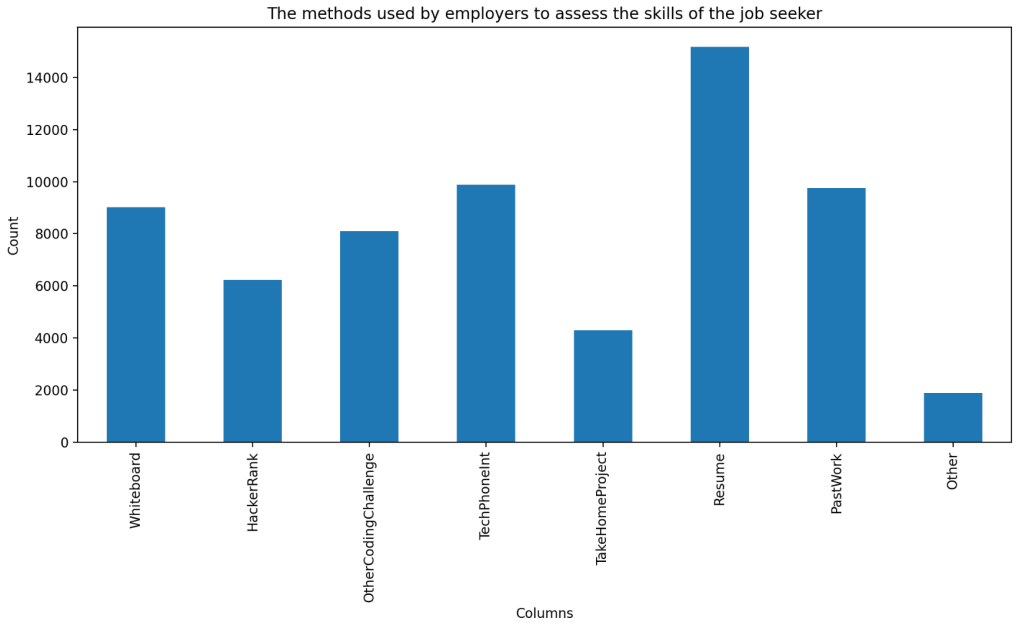
Most used programming language



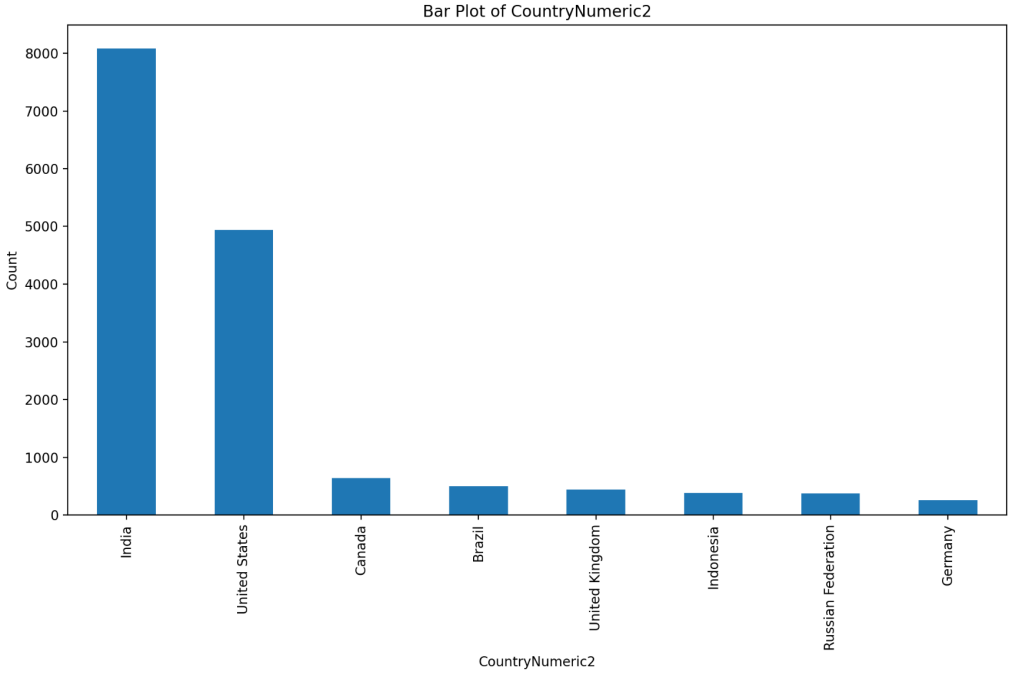
Most loved programming language



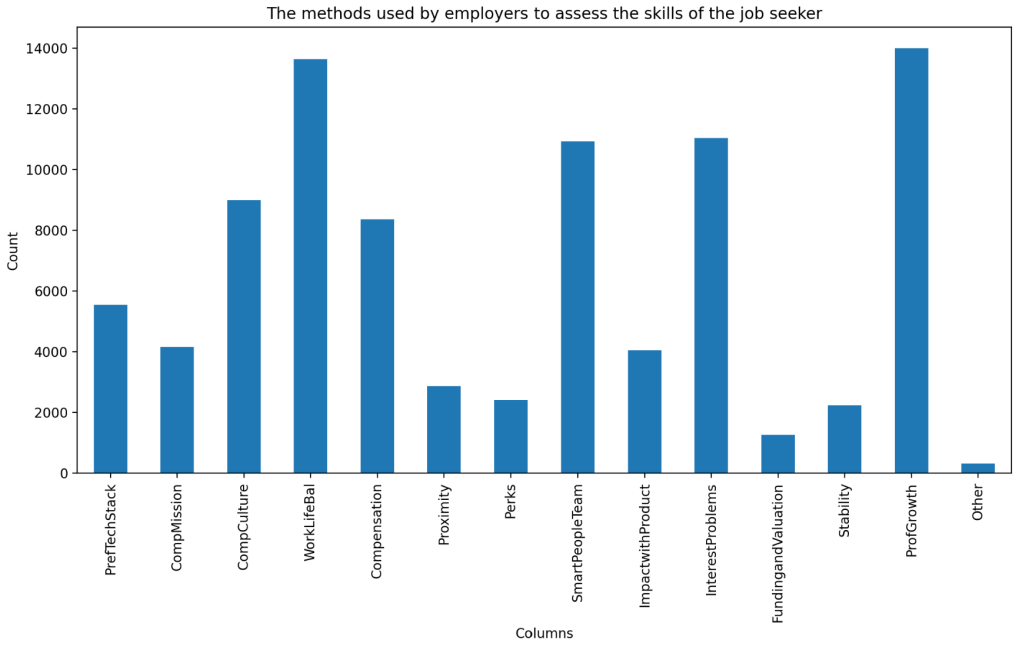
The methods used by employers to assess the skills of the job seeker



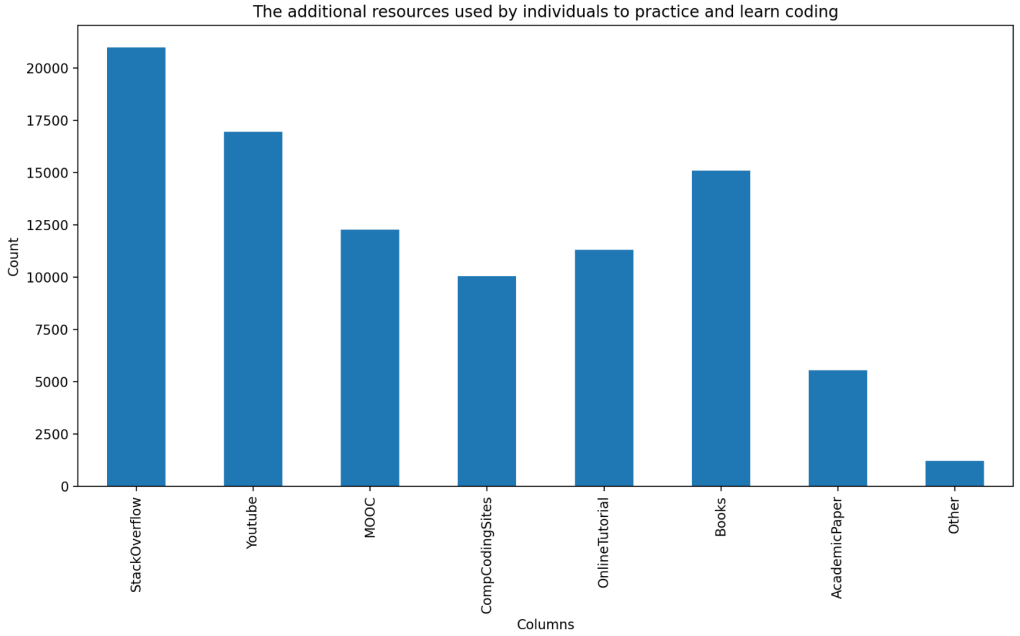
The country of origin for the respondent.



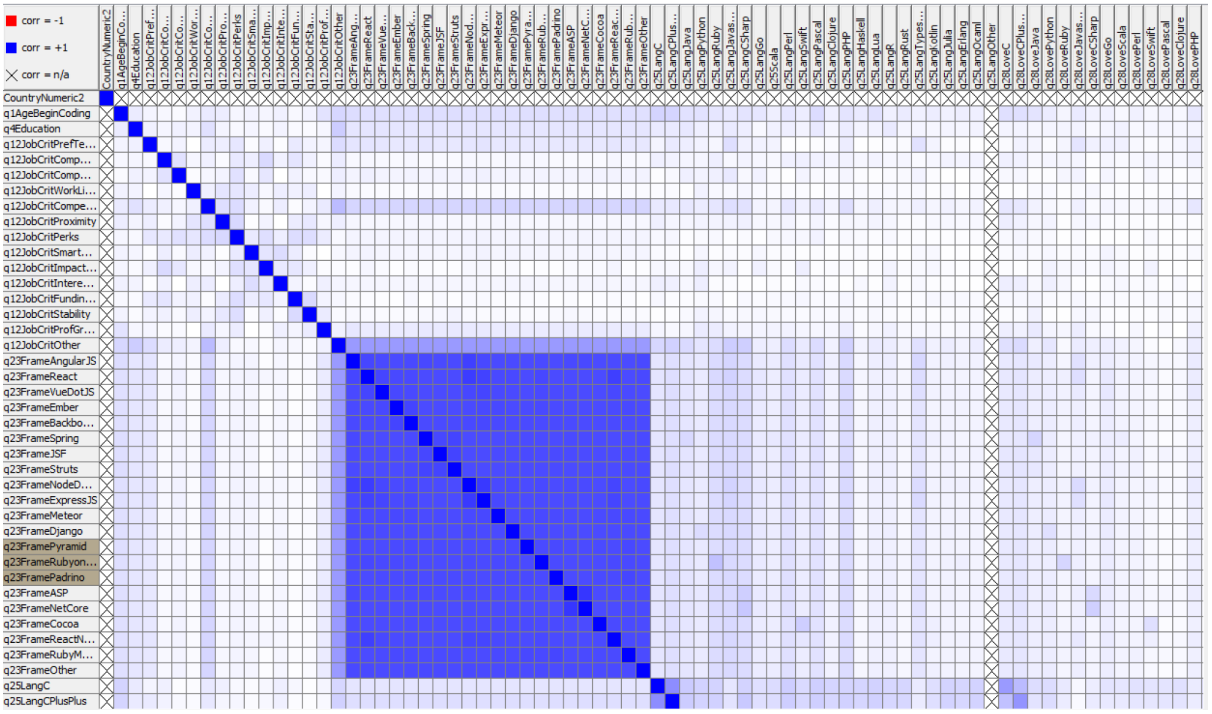
Criteria that respondents consider important when evaluating job opportunities



The additional resources used by individuals to practice and learn coding



Linear correlation:



There is no strong linear correlation observed among the following attributes: q1AgeBeginCoding, q4Education, q23Frame, q25Lang, q28LoveLang, q12EmpMeas, q30LearnCode, country, and q12JobCrit. **These attributes are categorical variables**, meaning they represent different categories or options rather than numerical values. As such, they do not have a numerical relationship that can be measured using linear correlation. Each attribute captures distinct aspects such as age at which coding began, education level, preferred frameworks, programming languages, measurement methods, learning resources, country of origin, and job criteria. Their relationships are qualitative rather than quantitative.

6. Discussion of data quality:

The overall quality of the dataset is considered to be on a very high level. The data comes from a trusted source, namely **HackerRank** which is *"the market-leading technical assessment and remote interview solution for hiring developers*". The dataset was cleaned from any incomplete answers as well as from any obvious spam submissions.

As most of the answers were **nominal**, with rest of them being **ordinal**, it is impossible to determine outliers as so.

As few of the survey's questions enabled respondents to provide their **custom answer** (option: *"other")* the dataset **includes some values of unknown meaning. Examples:**

- **question 4**. (regarding highest level of education attained or planned to obtain):
  2% of all submissions

- **question 12.** (regarding the most important things in a company that candidates are looking for):
  1% of all submissions

- **question 13.** (regarding the methods used by employers to assess the skills of the candidates):
  7% of all submissions

- **question 23.** (regarding the frameworks the hiring managers are looking for in candidates):
  4% of all submissions

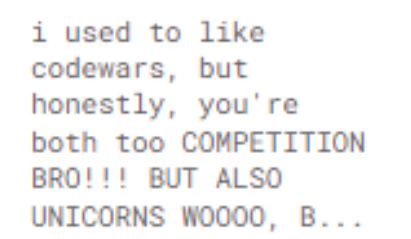
- **question 25.** (regarding most used programming languages):
  4% of all submissions

- **question 28.** (regarding most loved programming languages):
  1% of all submissions

- **question 30.** (regarding additional resources used by respondents to practice and learn coding):
  5% of all submissions

Image below shows an example of what was written in a "other" option for the question:

"Besides HackerRank, which other resources do you use to practice and learn coding? Check all that apply."



We can see that it is in fact **informative** but rather unique and not really useful in EDA ;)

Summary:

Accuracy	As the size of the dataset is large and its creators (respondents) are professionals in the field of IT, or people that want to become such, we can assume (and check using other sources) that the data given in the dataset is correct and reflects reality.
Completeness	In case of our research goals the dataset is complete and comprehensive.
Reliability	Given the respondents experience and HackeRank reputation we can freely assume that the dataset comes from a reliable source
Relevance	The dataset contains some values of unknown meaning.
Timeliness	The survey was conducted in 2016 and given the rapid development of the IT sector it might not be ideal for reflecting the current job market situation.