SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Job Recommendation

Team 1

Meet Our Team

Abdelrahman Walid

Yassin Osama

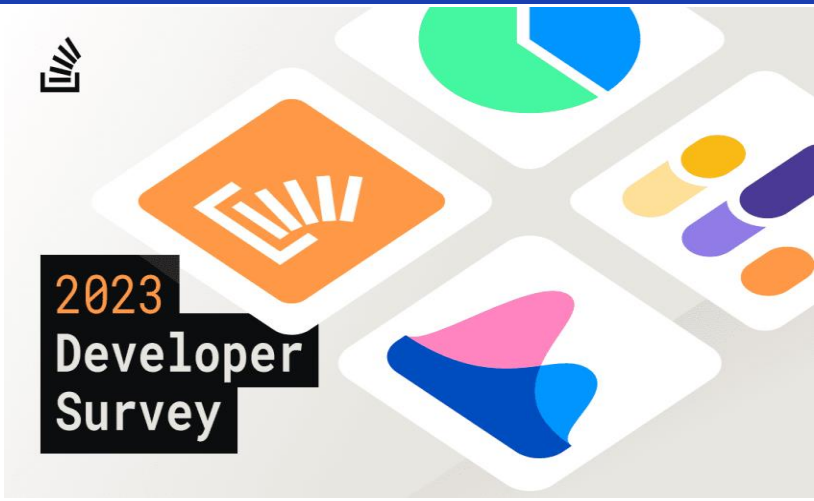
Yara Hisham

Under supervision of : Zead Omar

Agenda

- Project Overview
- Data Source
- Exploratory Data Analysis
- Data Preprocessing & Feature Engineering
- Modeling & Deployment

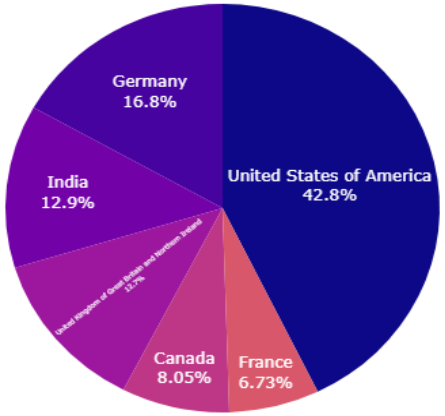
Project Overview
Job Recommendation System

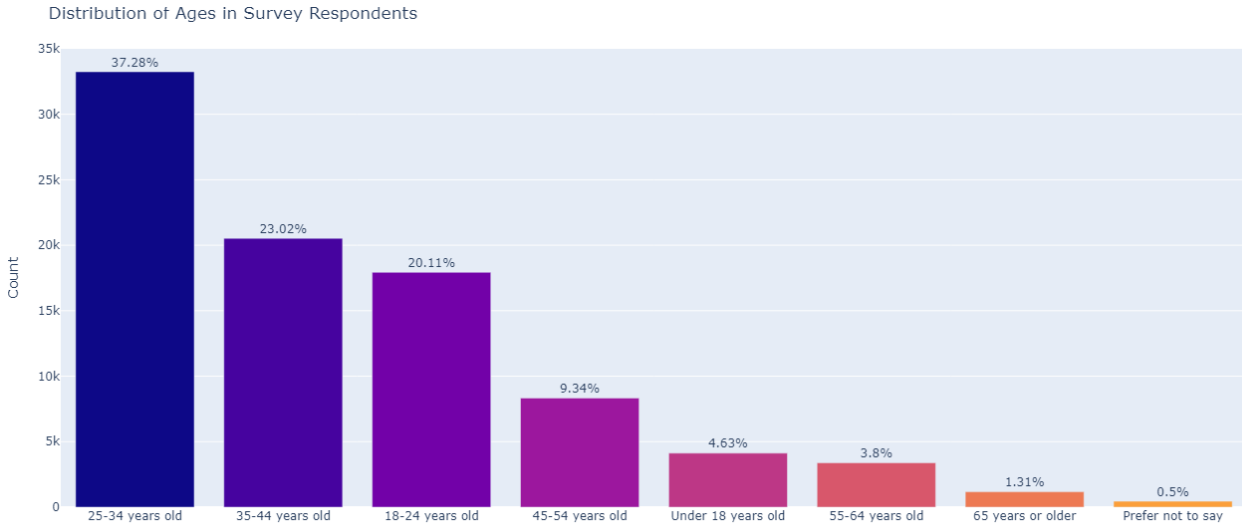


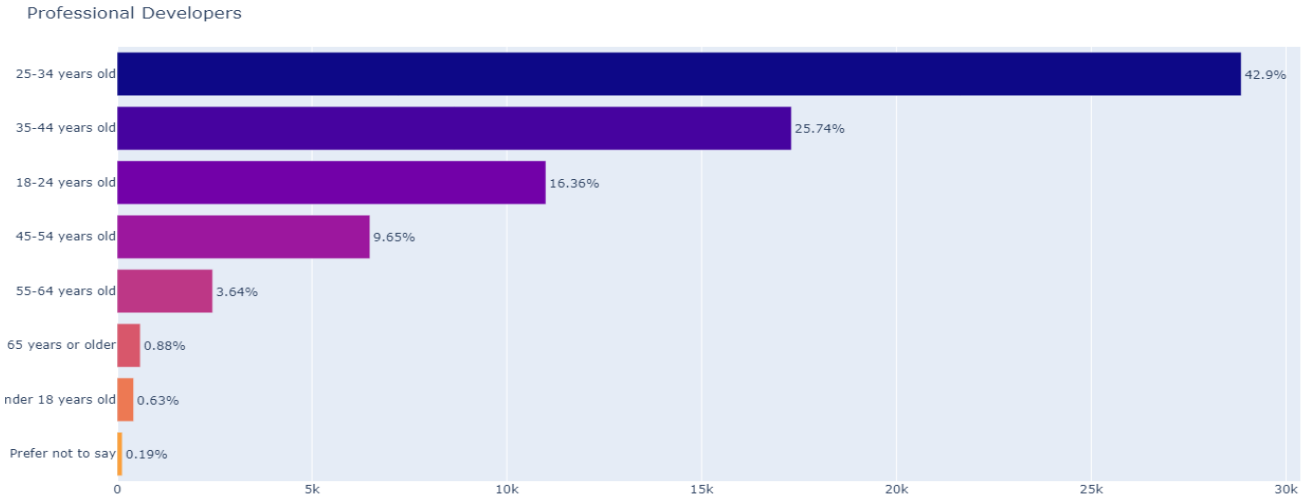
- Stack Overflow Annual Developer Survey
- Importance of Job Recommendation System
- **Dataset**: Stack Overflow runs a developer survey about technology, work, community, and more.

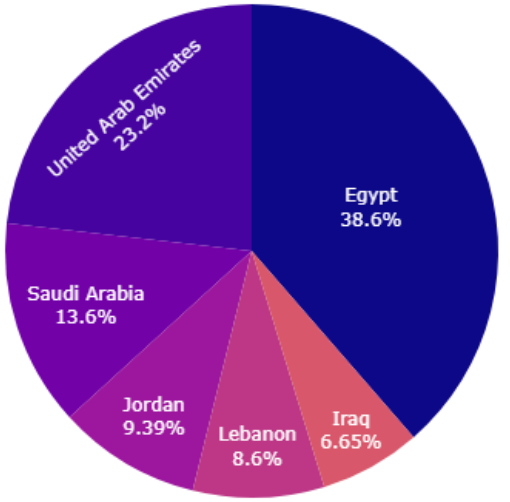
Exploratory Data Analysis

- **Demographics**





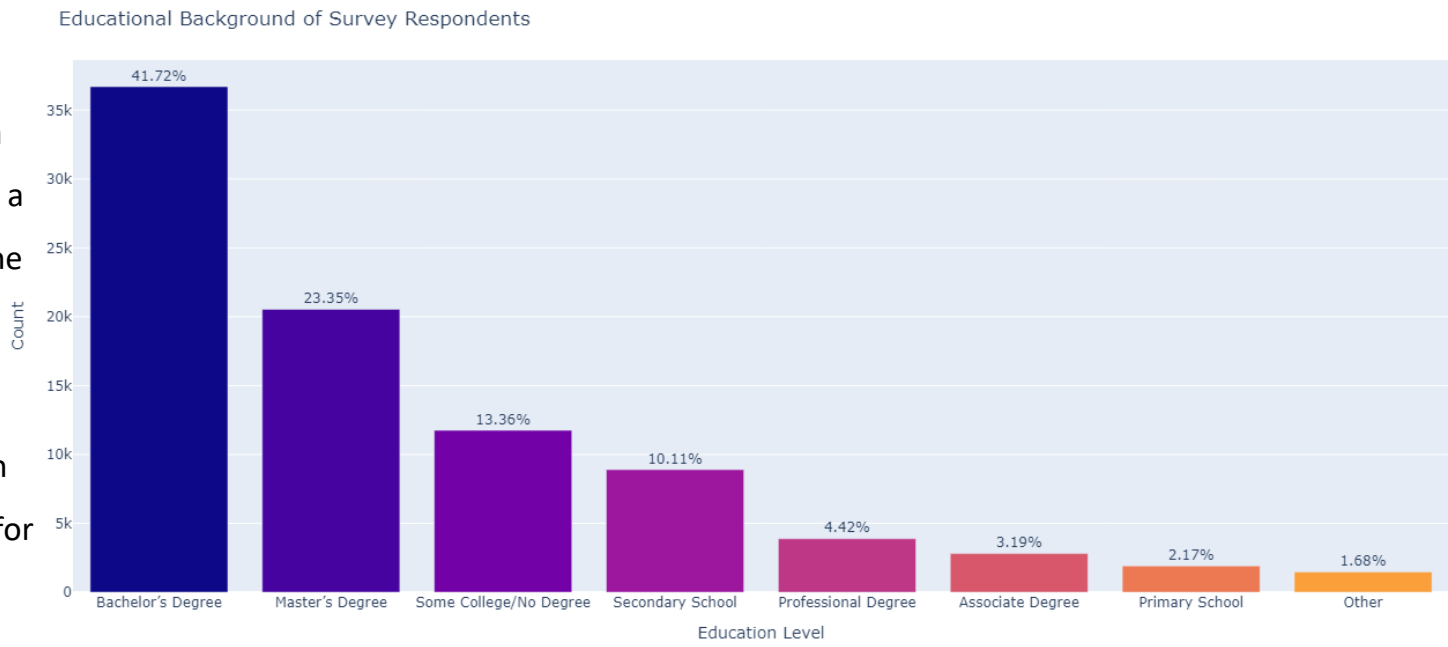




Exploratory Data Analysis

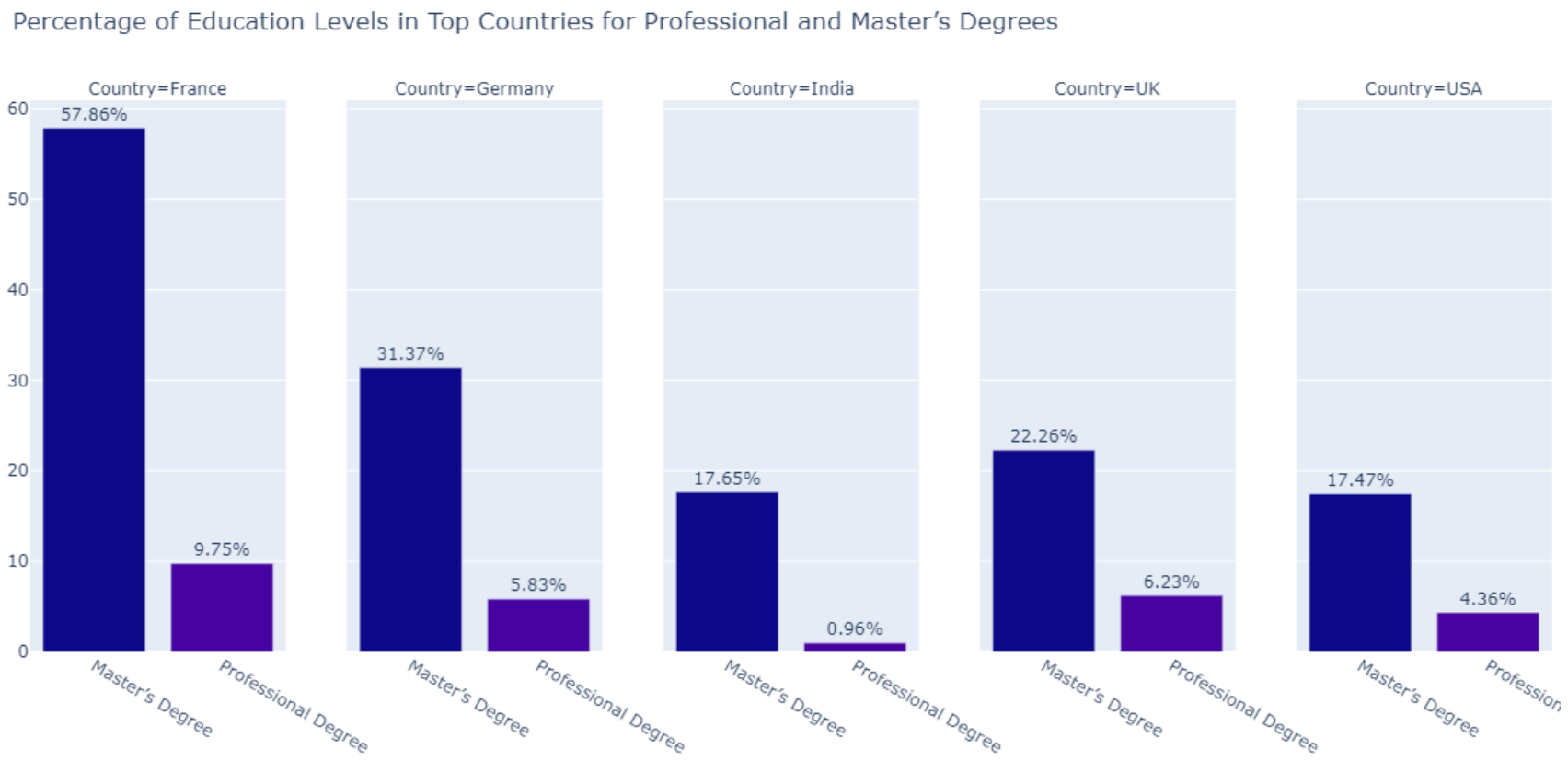
Education

- The majority of respondents have a bachelor's degree, highlighting it as a key educational level for hiring in the tech industry
- Professional degrees often indicate highly specialized knowledge, which may translate into a reduced need for resources like StackOverflow.



Exploratory Data Analysis

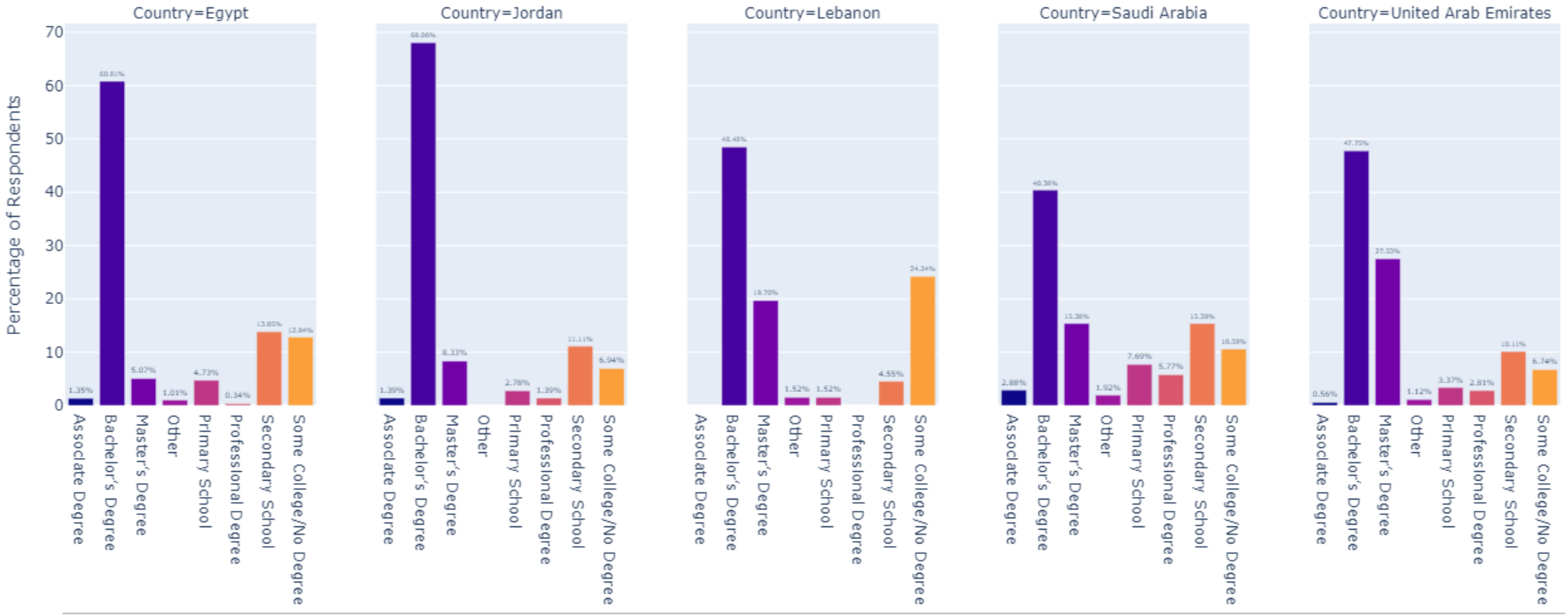
- **Education**



Exploratory Data Analysis

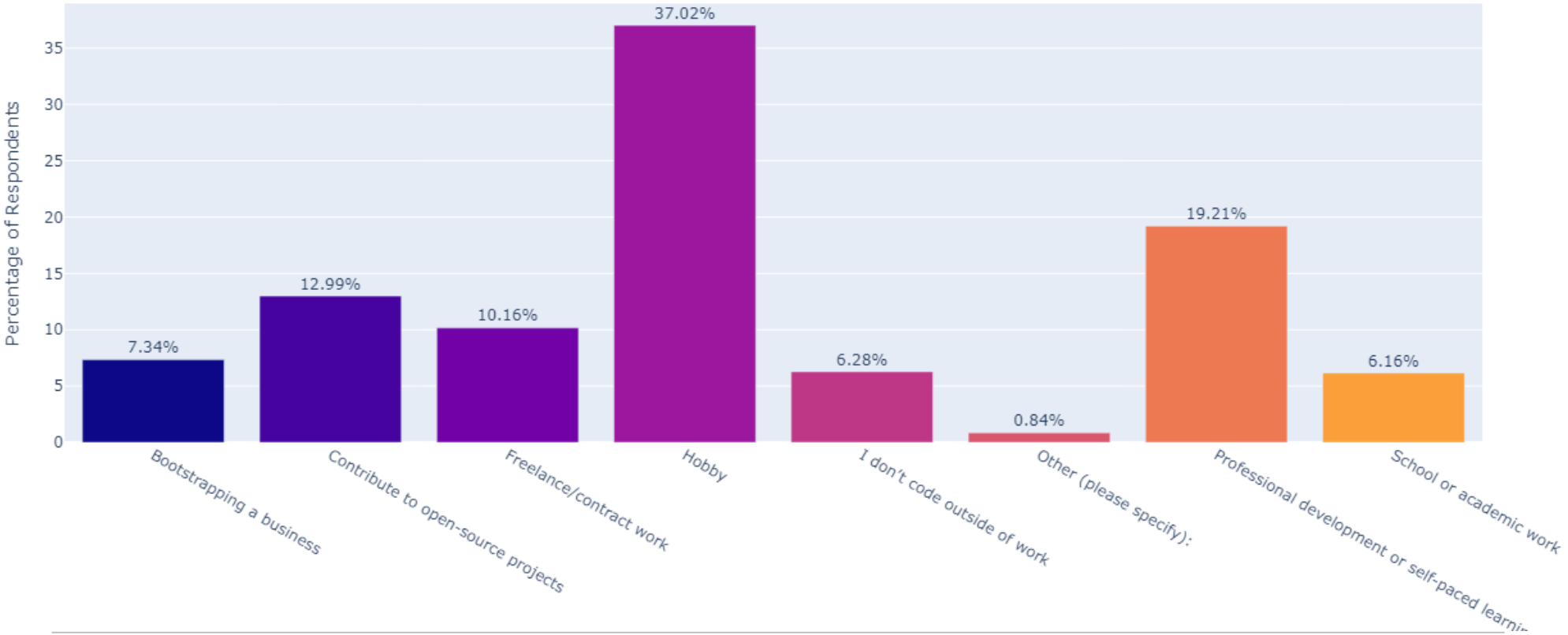
Education

- Egypt and Jordan have the highest proportion of Bachelor’s Degree holders, while the United Arab Emirates leads in Master’s Degree holders



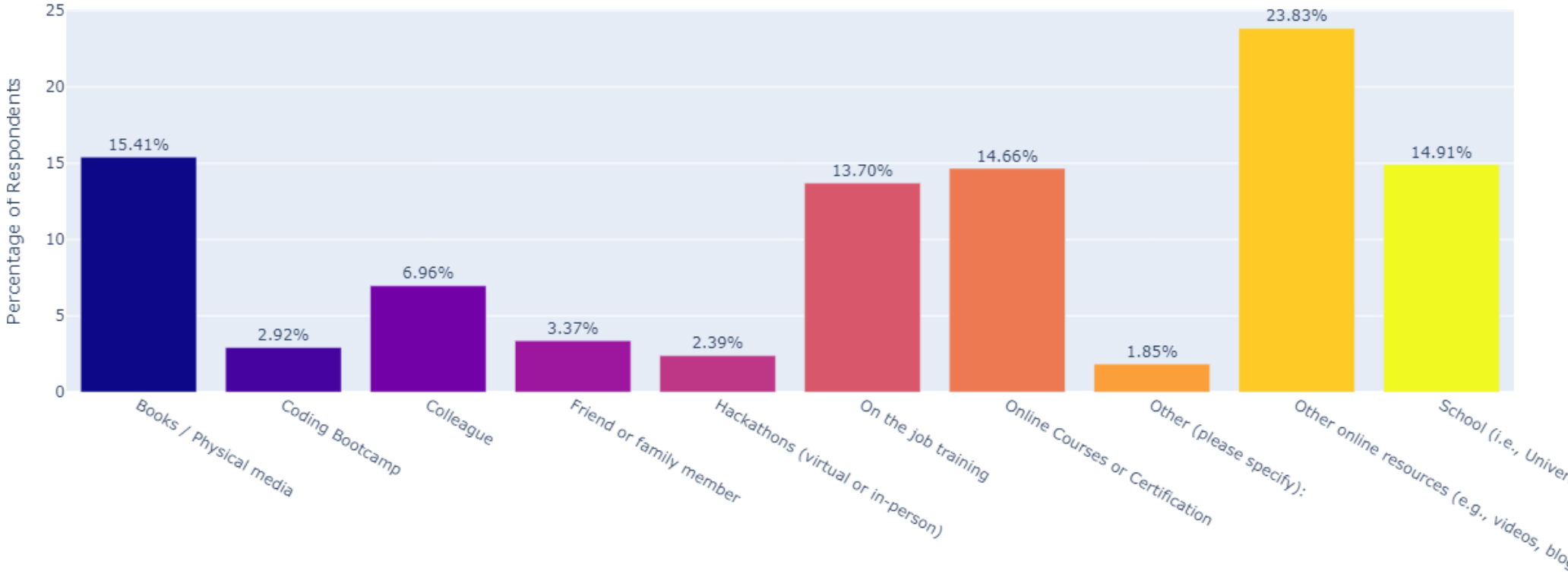
Exploratory Data Analysis

- **Learning to code**
  - Coding outside of work



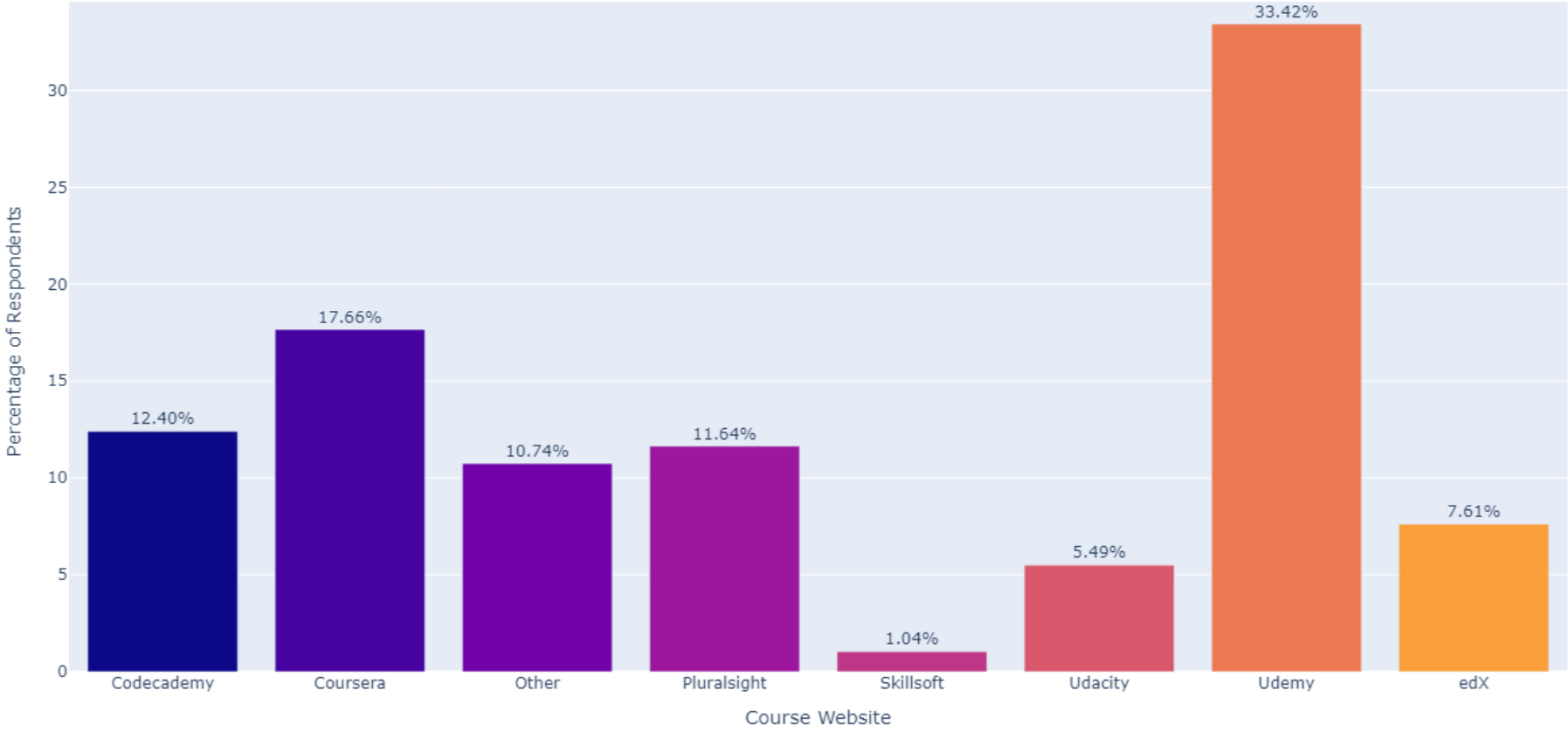
Exploratory Data Analysis

- **Learning to code**
  - Where Do People Learn to Code?



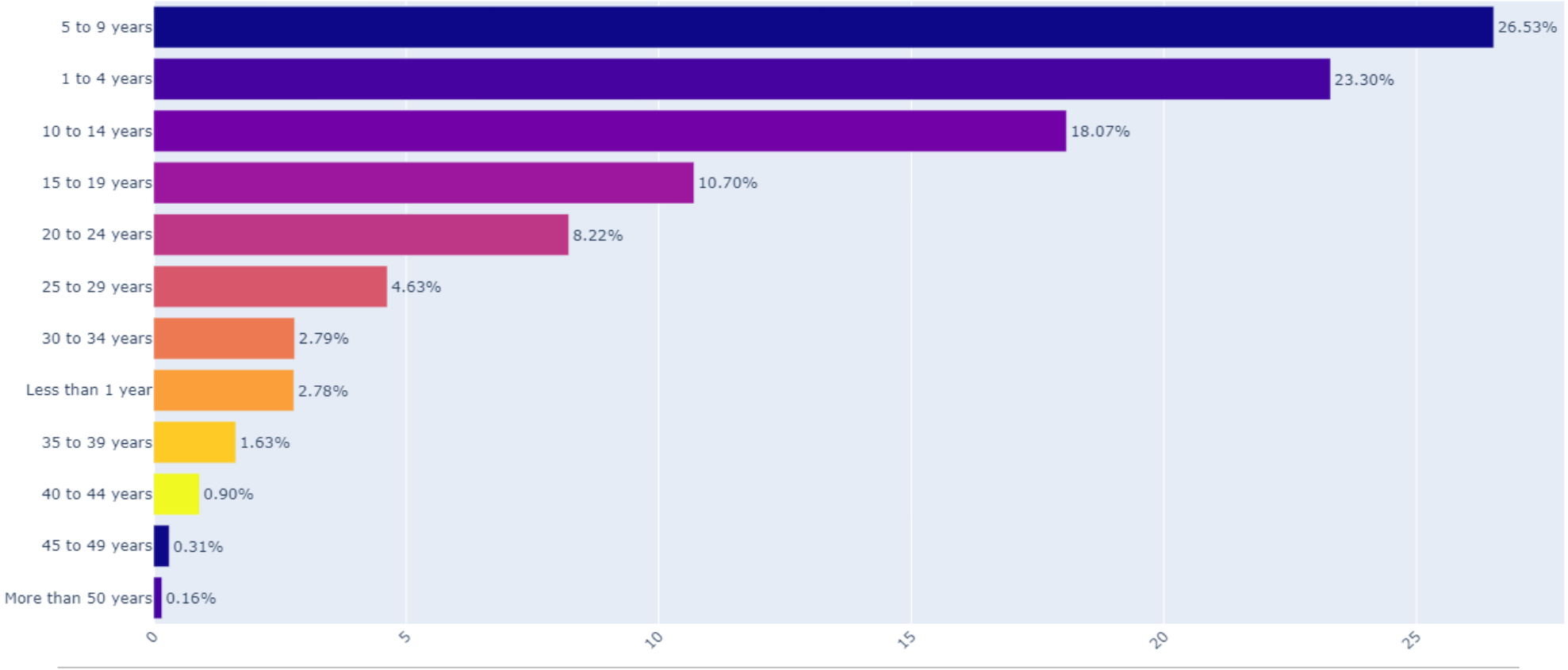
Exploratory Data Analysis

- **Learning to code**
  - Online course platforms to learn how to code



Exploratory Data Analysis

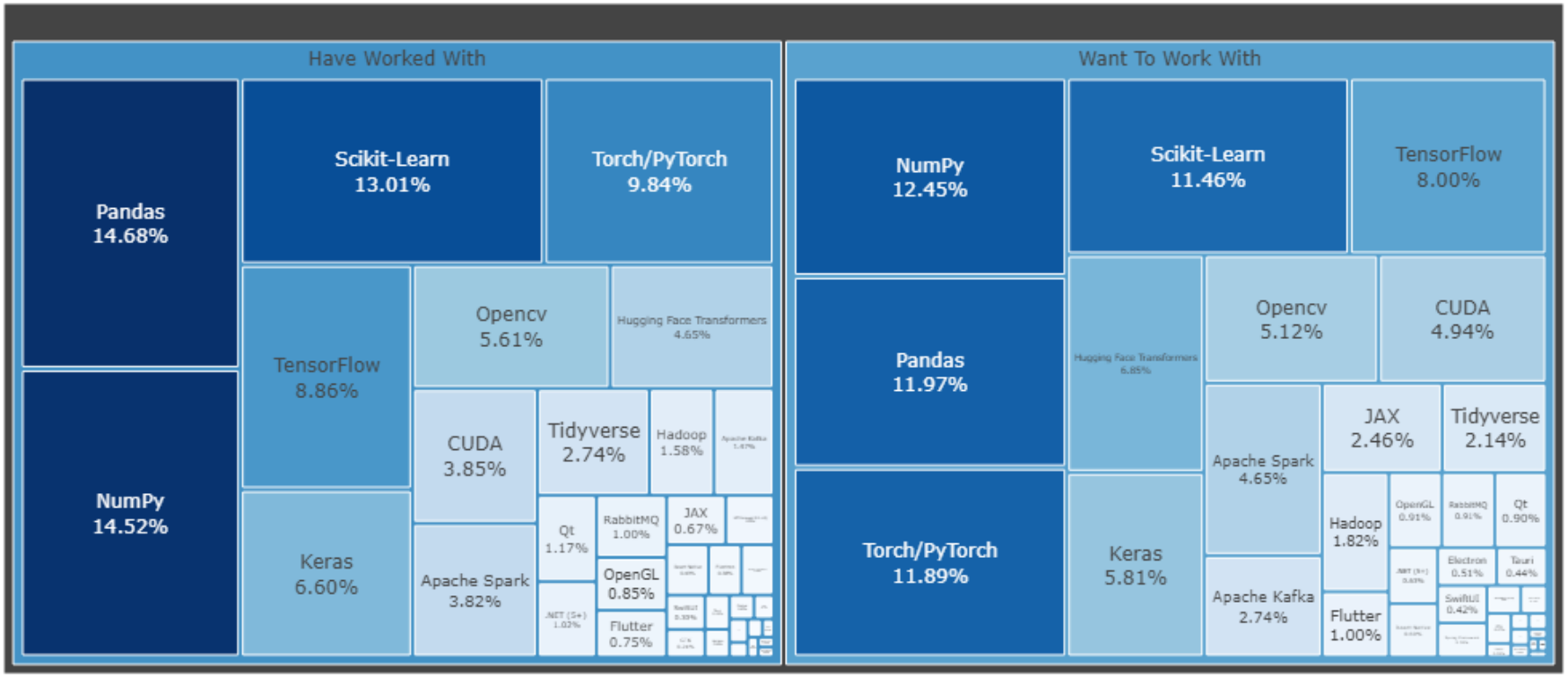
- **Experience**
  - Years coding professionally



Exploratory Data Analysis

Technology

MiscTech for Data scientist or machine learning specialist

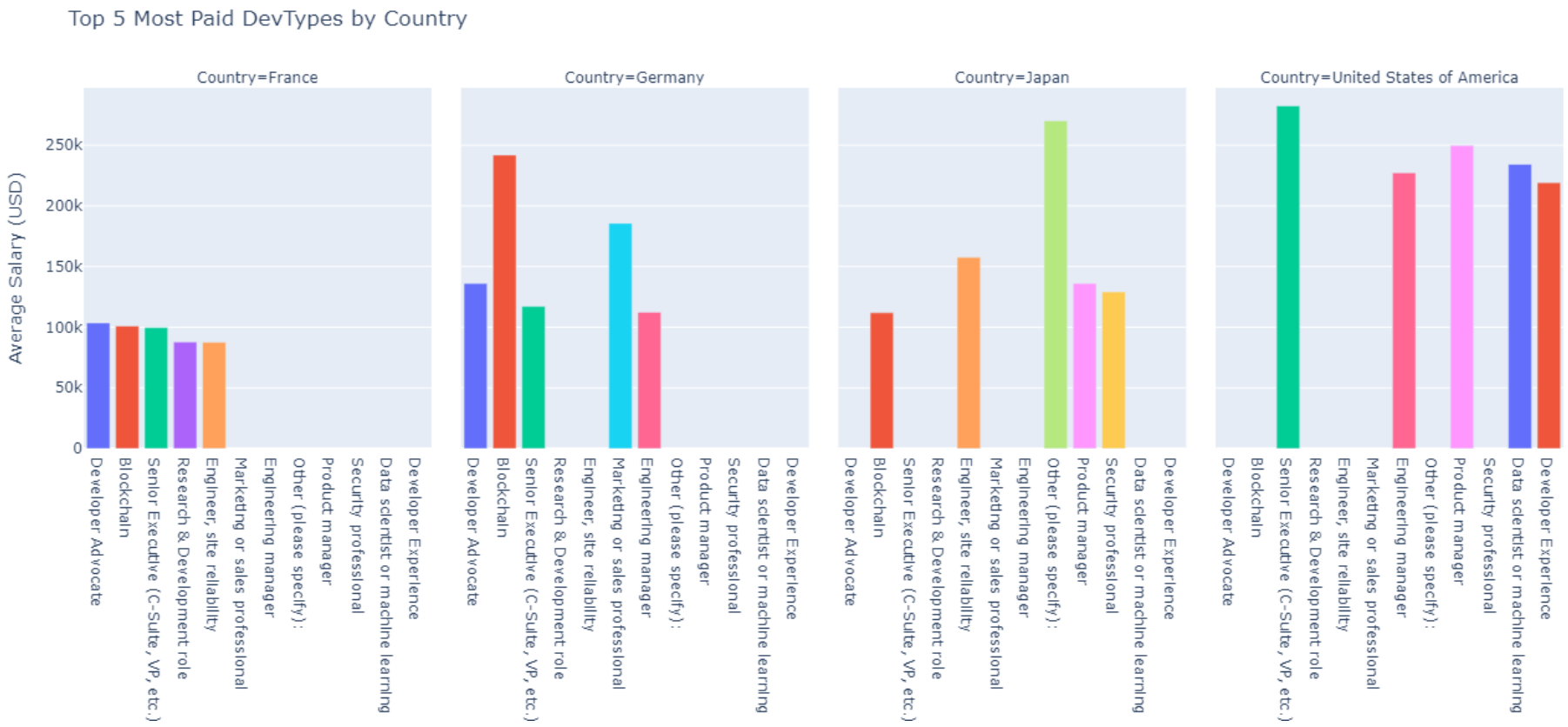


Exploratory Data Analysis

- **Work**
  - Average salary for junior developers by each development type

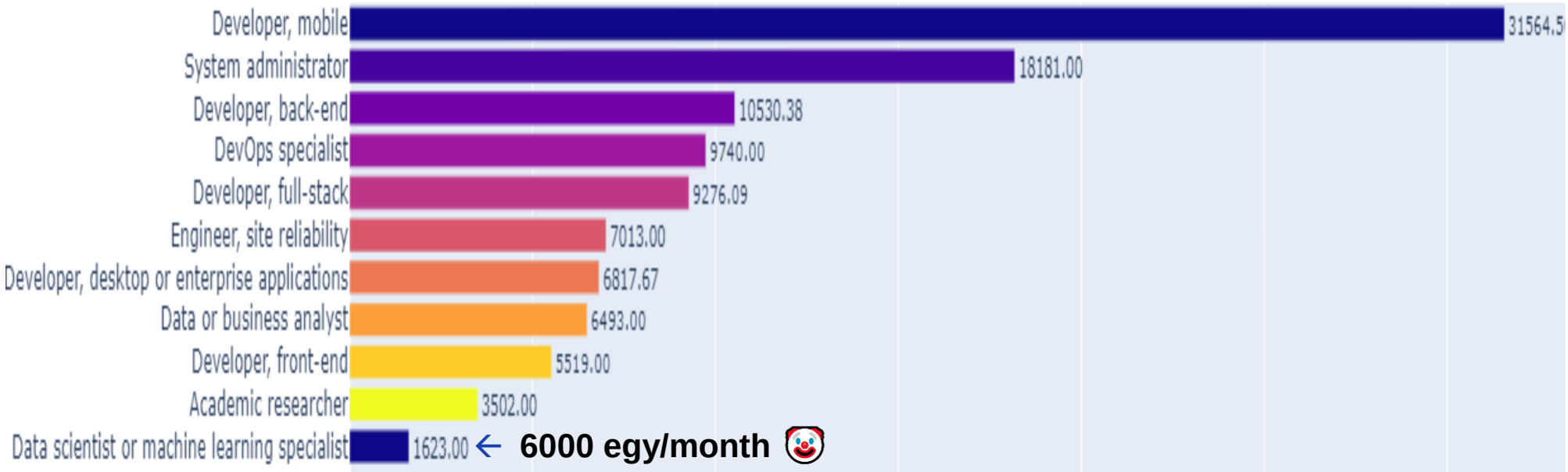
Exploratory Data Analysis

- **Work**



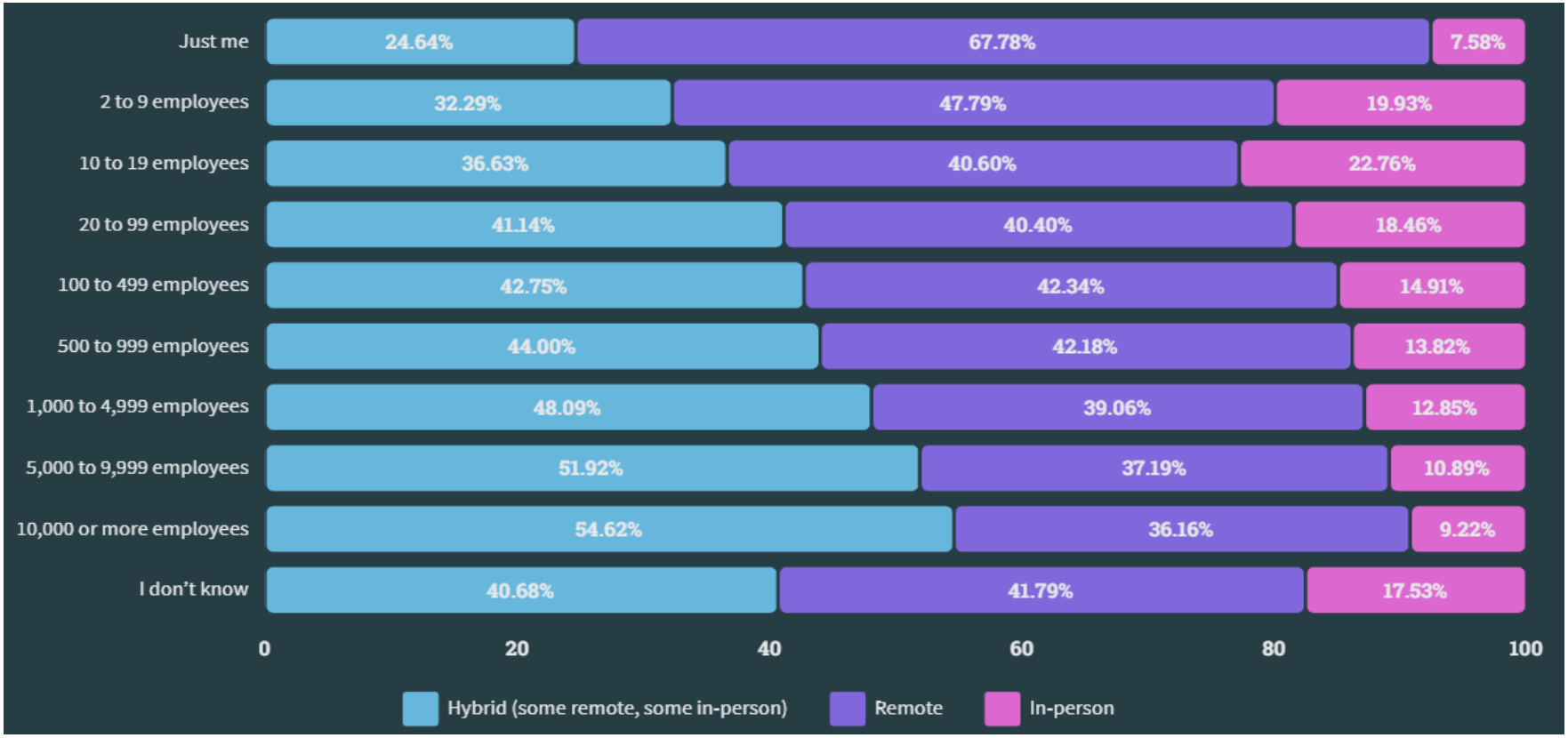
Exploratory Data Analysis

- **Work**
  - Average salary for junior developers in ***Egypt***



Exploratory Data Analysis

- **Work**
  - Work environment



Exploratory Data Analysis

In conclusion

- The 2023 Stack Overflow Developer Survey offers a detailed look at the developer landscape, highlighting emerging trends and the industry's continuous evolution. From programming languages and frameworks to cloud services and remote work preferences, developers are constantly adapting to the shifting demands of technology. With increasing focus on ethics and diversity, the developer community is driving efforts toward a more inclusive and responsible future. As we wrap up another insightful survey year, we look forward to the innovations and changes the tech world will bring in the future.

Preprocessing

Data Preprocessing

- The Data Shape is: (89184, 84).
- The Data has many nulls, but we will work on skills features only and dropped the nulls in target column.
- The skills features are separated with semi colon in the same sample
- So, we will split the answers in a list.
- The null samples will be converted to empty list.

LanguageHaveWorkedWith	LanguageWantToWorkWith	DatabaseHaveWorkedWith	DatabaseWantToWorkWith
Bash/Shell (all shells);HTML/CSS;JavaScript;PH...	Bash/Shell (all shells);Go;HTML/CSS;JavaScript...	PostgreSQL;Snowflake	PostgreSQL
HTML/CSS;Java;JavaScript;Python;SQL;TypeScript	HTML/CSS;JavaScript;Python;SQL	BigQuery	BigQuery
C#;HTML/CSS;Java;JavaScript;Kotlin;Objective-C...	Objective-C;Python;SQL;Swift;TypeScript	MongoDB;MySQL;PostgreSQL	BigQuery;MongoDB;MySQL

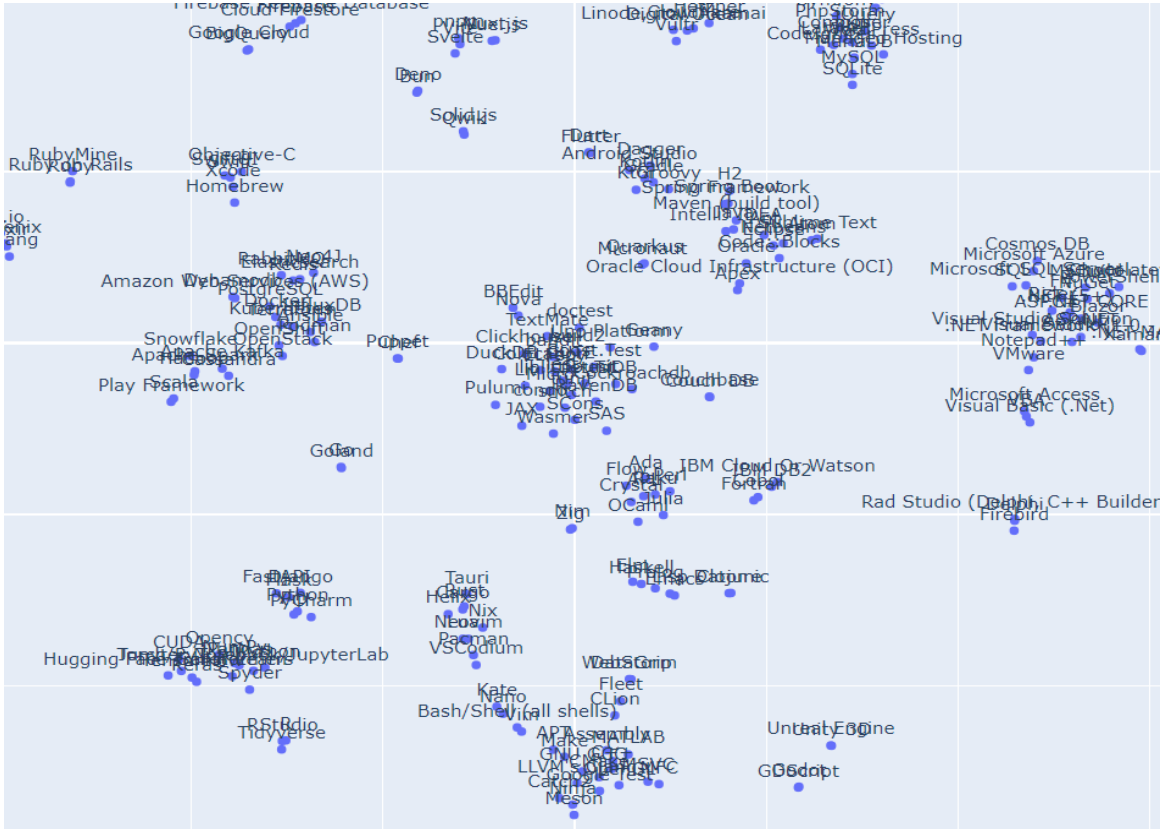
Data Preprocessing

- For better features handling, we encoded them using “MultiLabelBinarizer”.

APL	Ada	Apex	Assembly	Bash/Shell (all shells)	C	C#	C++	Clojure	Cobol	Crystal	Dart	Delphi	Elixir	Erlang	F#	Flow	Fortran	GDScript	Go	Groovy	HTML/CSS
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

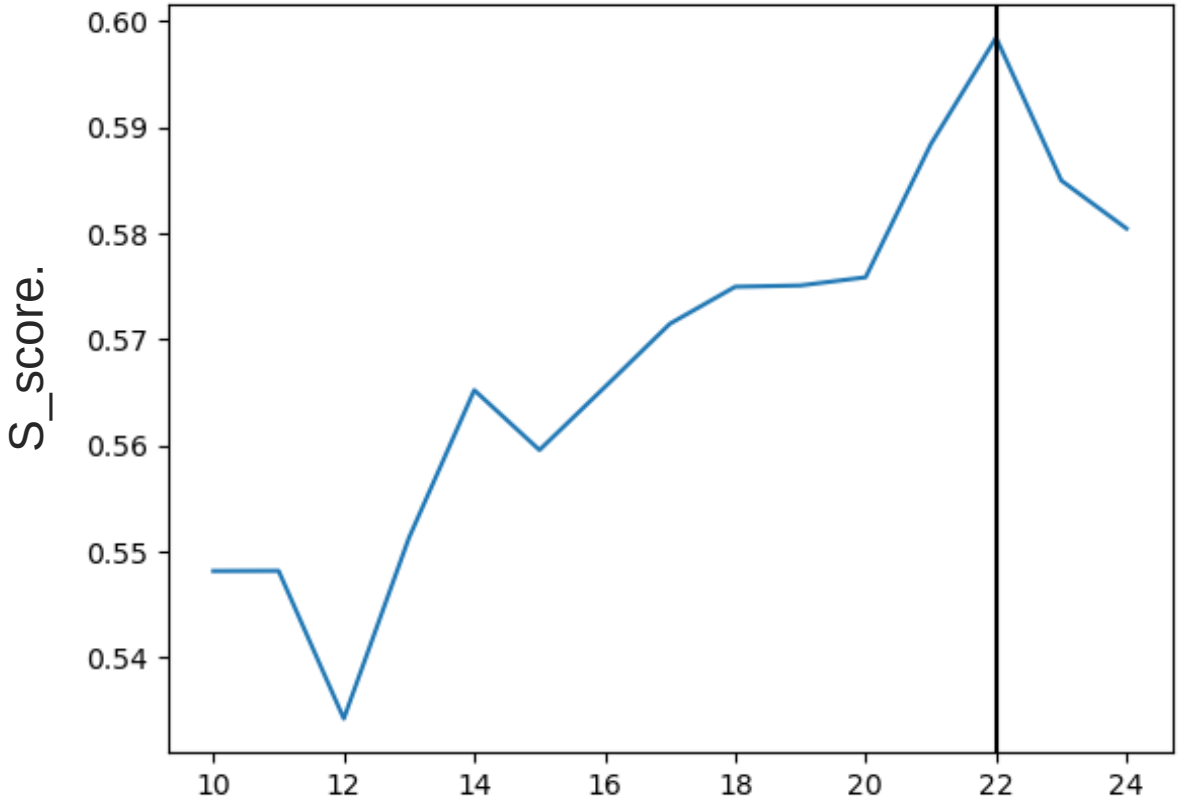
Data Preprocessing

- Now The number of features = 270, So we applied clustering to reduce the number of features.



Data Preprocessing

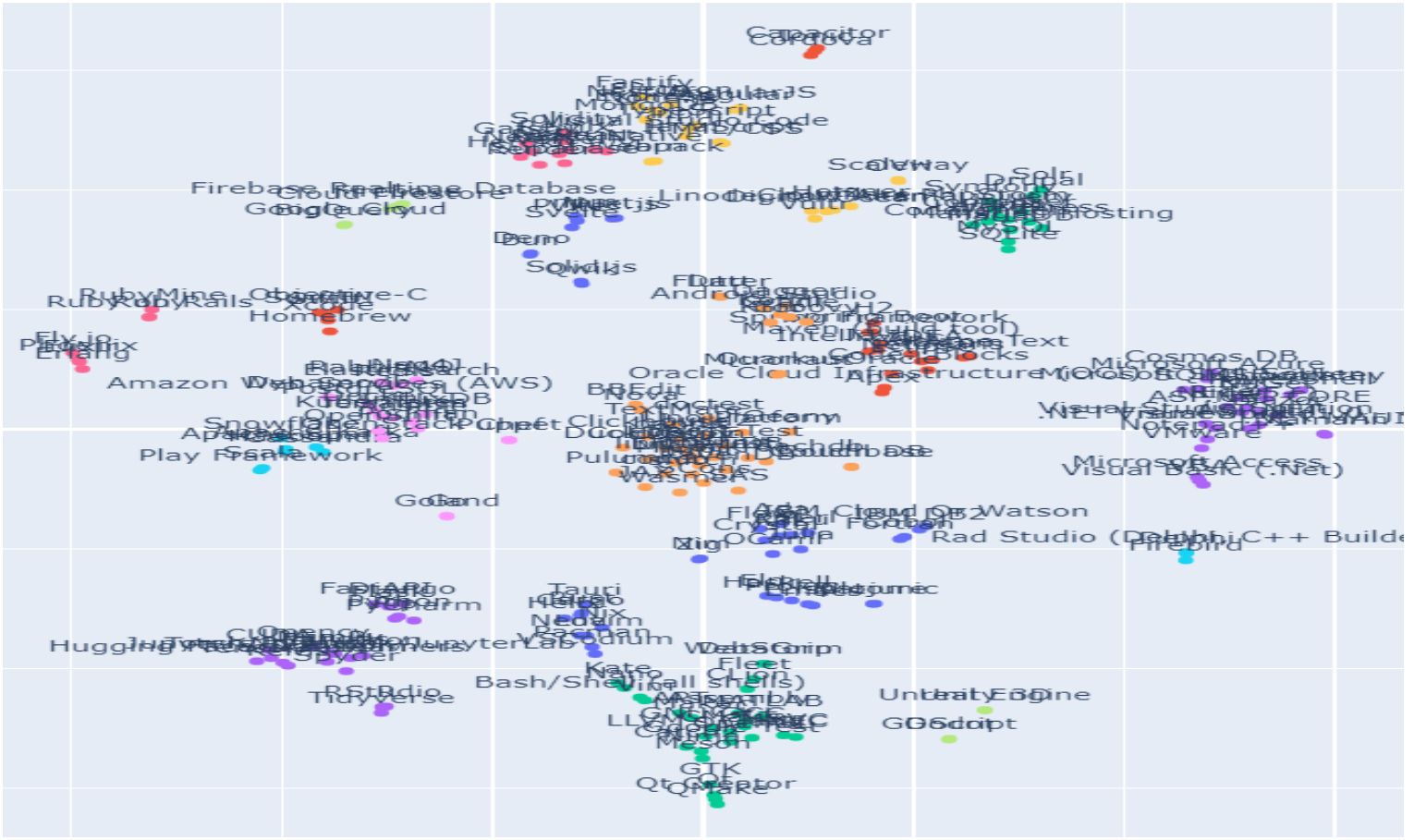
- Using Silhouette Score we found the best number of clusters is 22.



Data Preprocessing

- After Clustering:

- Now we can use it for feature engineering to make a new features then merge it with the original skills.



Data Preprocessing

- Another problem we have is the variance in answers, so we decided to take the answers that more than or equal 3 and less than or equal 30 only.
- The target column has 33 unique class some of them are not important, so we decided to exclude some classes like Student, Designer, Educator, Scientist, and others.

Data Preprocessing

- The final target classes are:

- To fix the unbalancing in data, We decided to take 1459 samples per class “Q1 of the distribution”.

Target Classes	Count
Full-stack dev	19335
Back-end dev	11604
Front-end dev	4484
Desktop dev	3377
Mobile dev	2290
Embedded dev	1632
Data Scientist	1286
DevOps	1018

Data Preprocessing

- Finally, we applied one hot encoding on target too

DevType								
Back-end dev	Data Scientist	Desktop dev	DevOps	Embedded dev	Front-end dev	Full-stack dev	Mobile dev	Researcher
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0

Modeling & Deployment

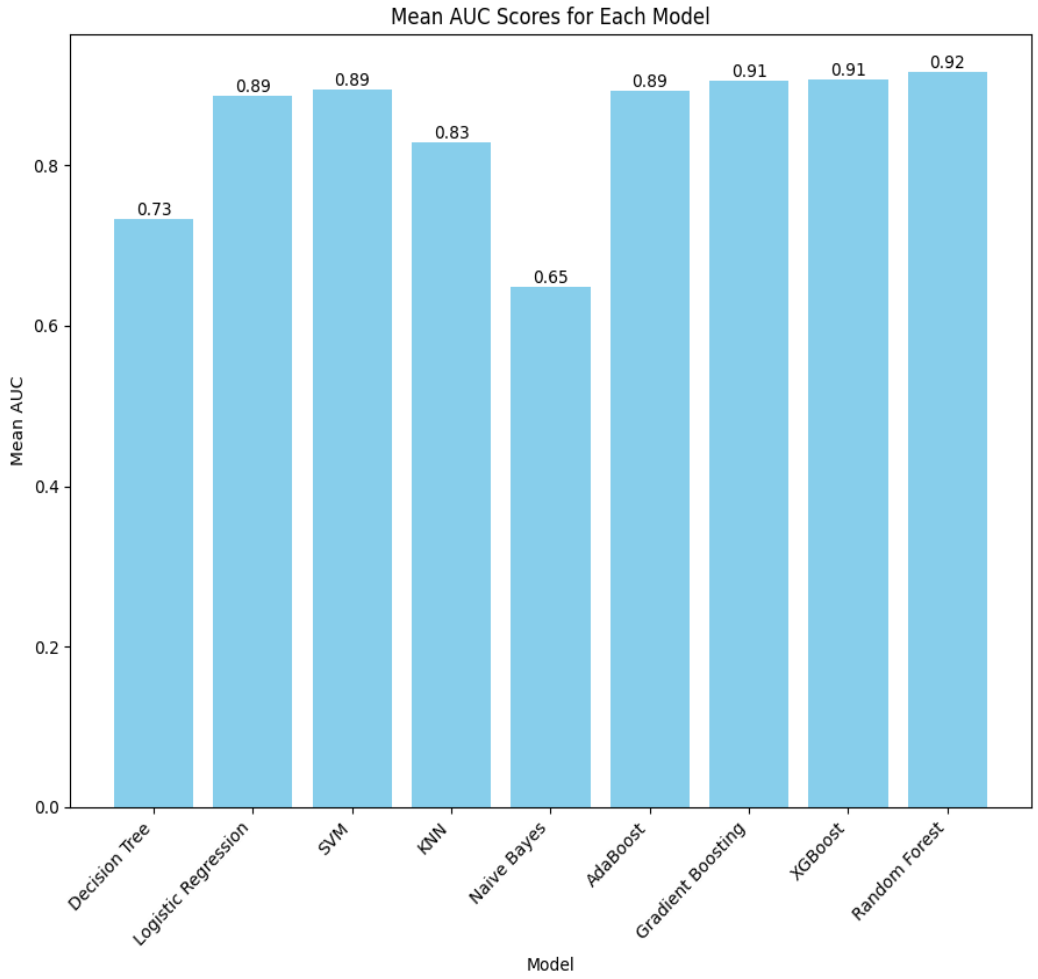
Modeling & Deployment

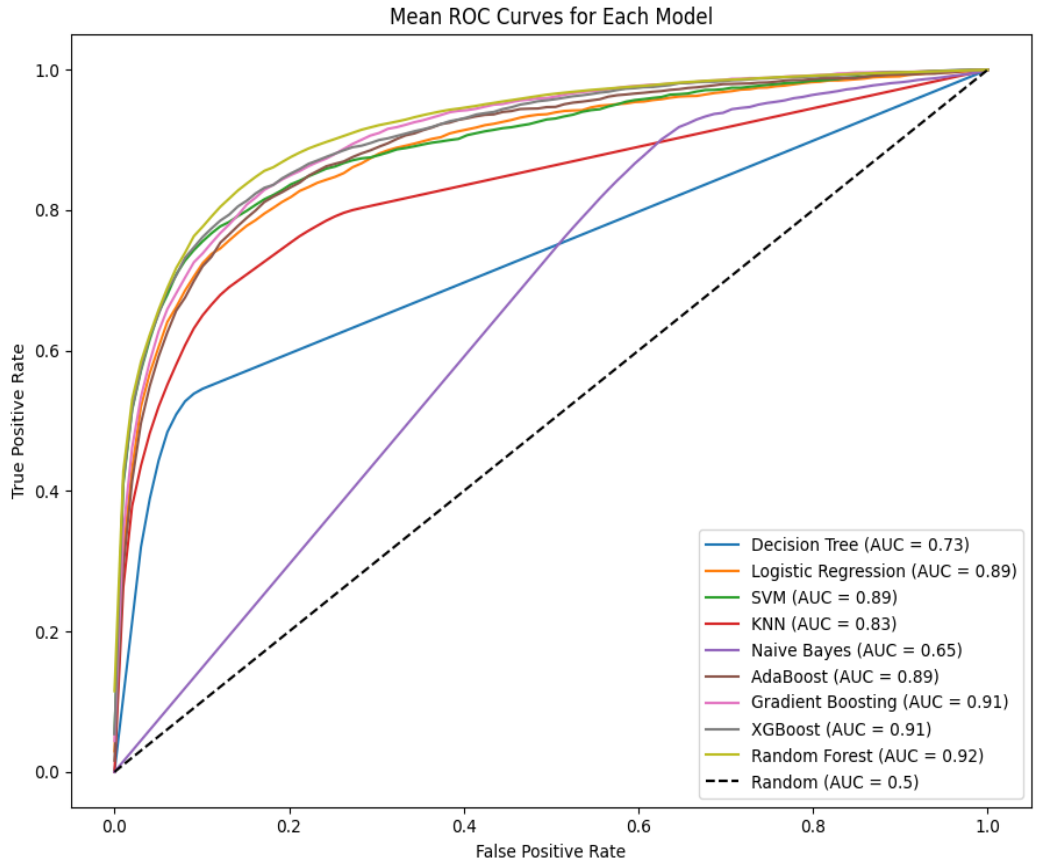
Base Models

Trying all the classification models with their default parameters, the XGBoost had the best f1 score so far

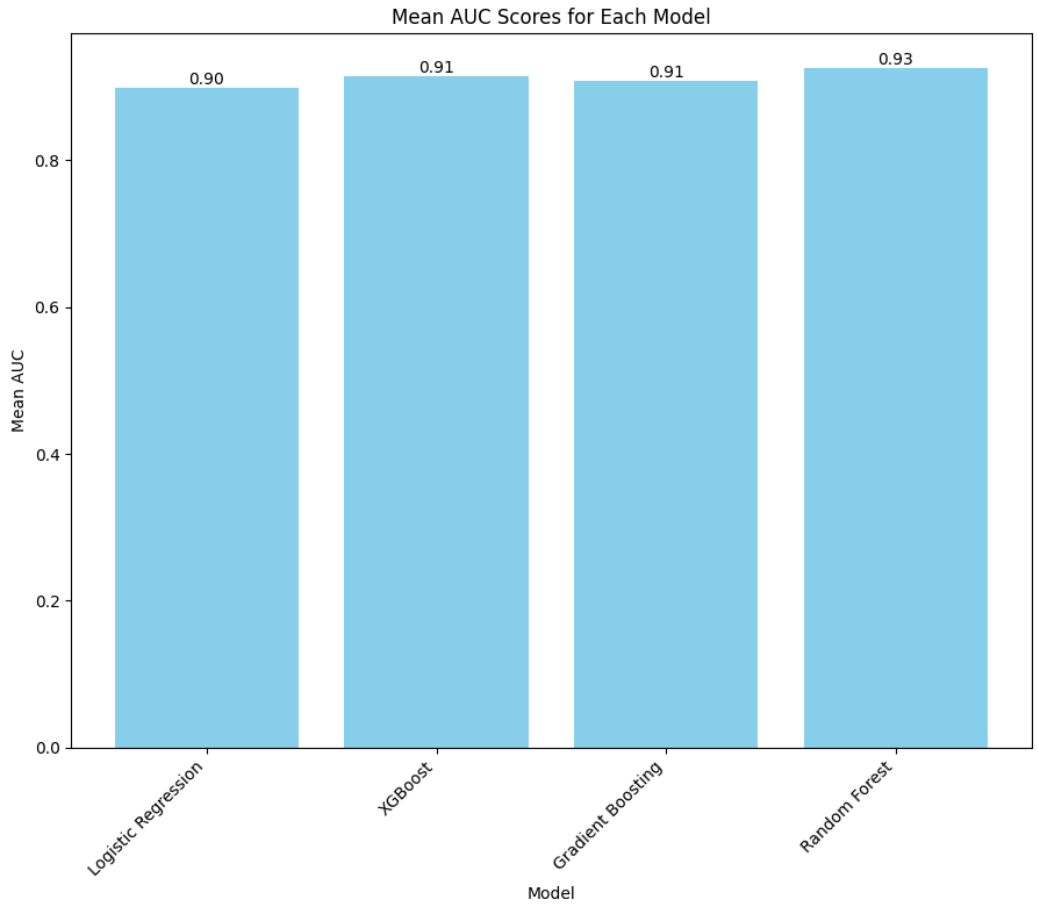
Test Scores	
Accuracy	92.802222
Precision	70.963333
Recall	53.715556
F1 score	59.877778

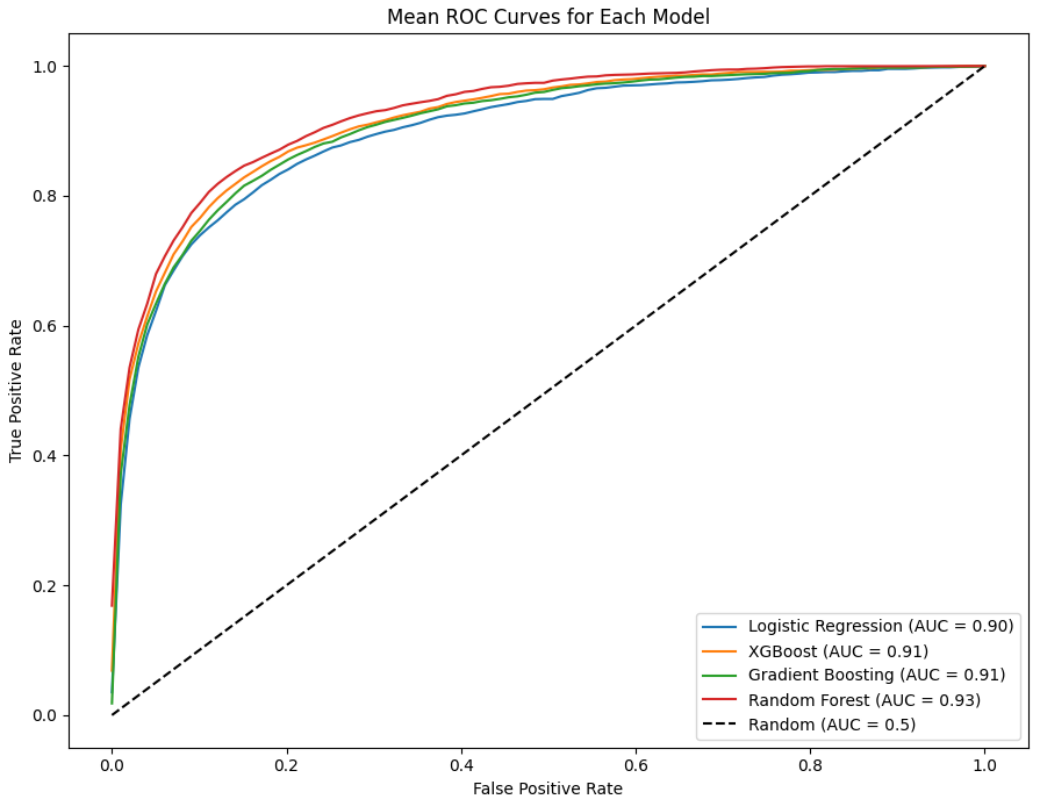
Modeling & Deployment
Base Models



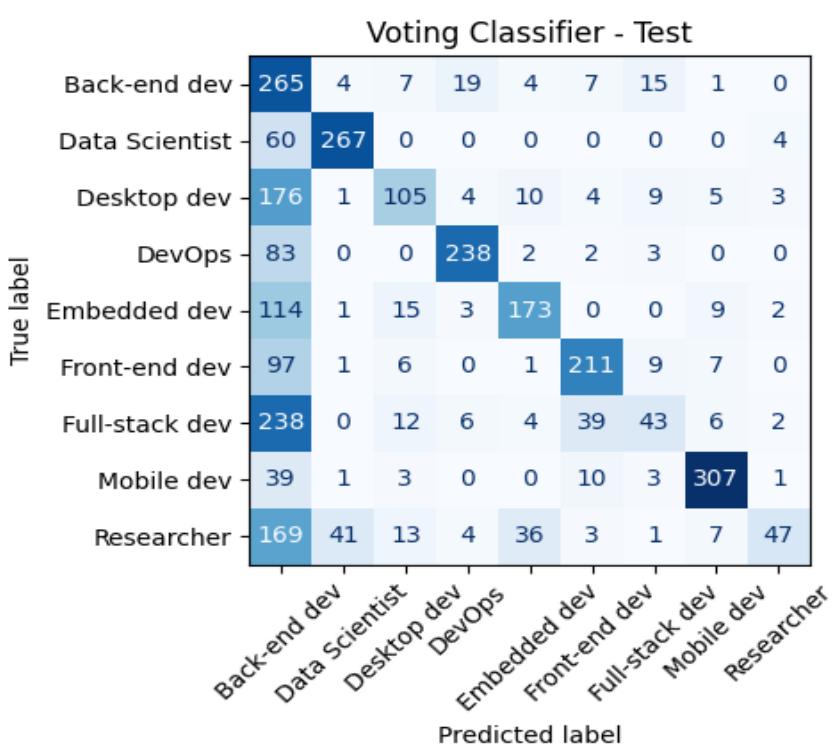


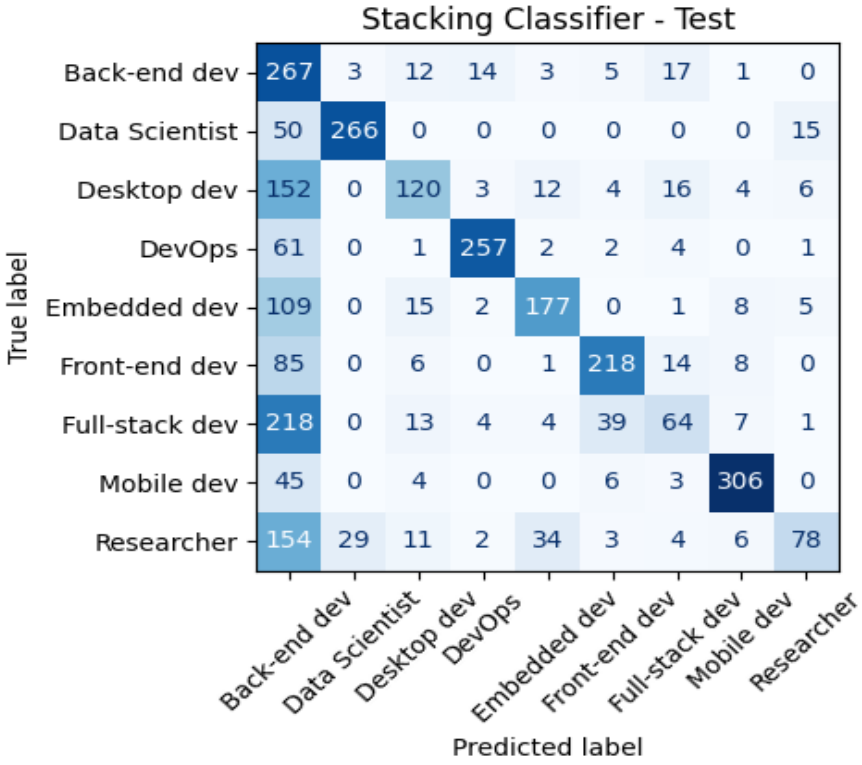
Modeling & Deployment
Hyperparameters Tuning





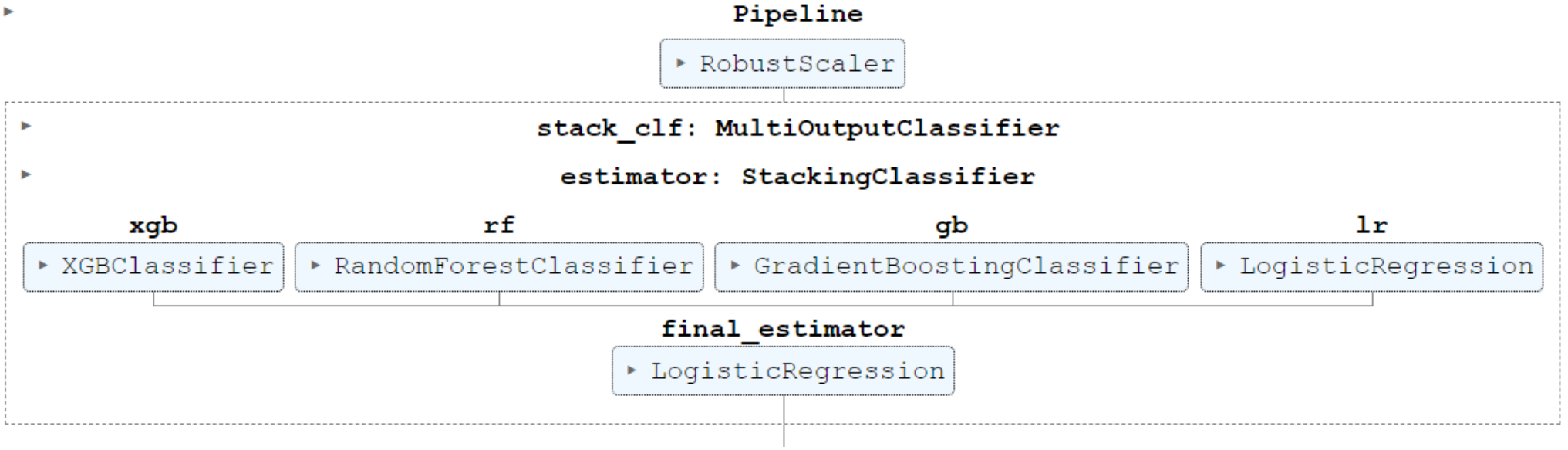
Modeling & Deployment
Voting & Stacking Models



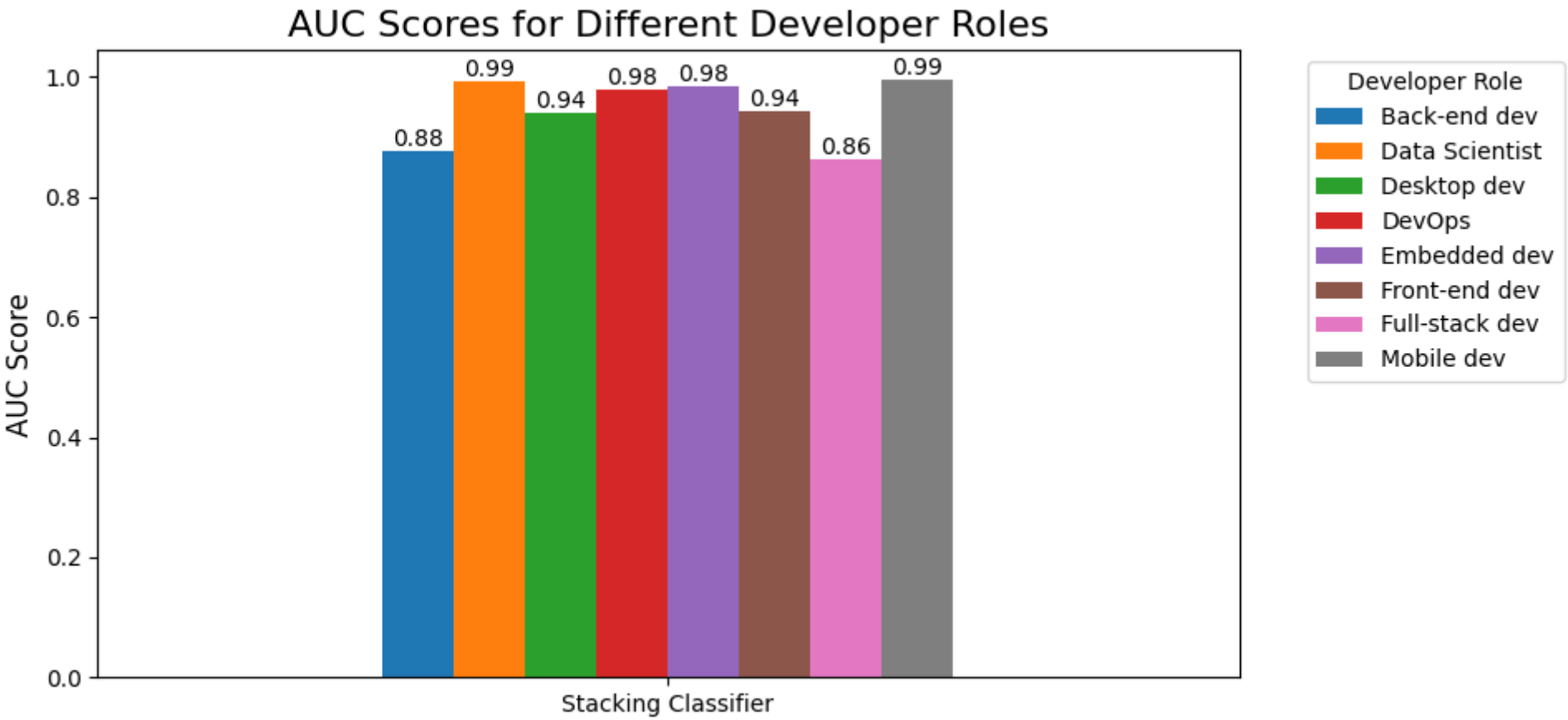


Modeling & Deployment
Best Model

Test Scores	
Accuracy	91.57875
Precision	72.79625
Recall	73.01875
F1 score	70.34500

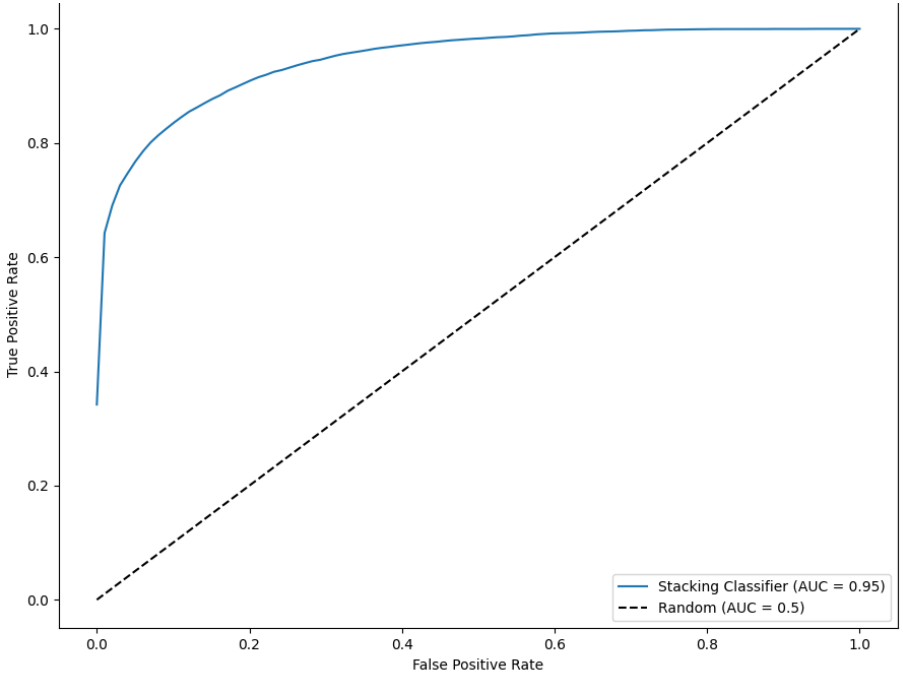


Modeling & Deployment
Best Model

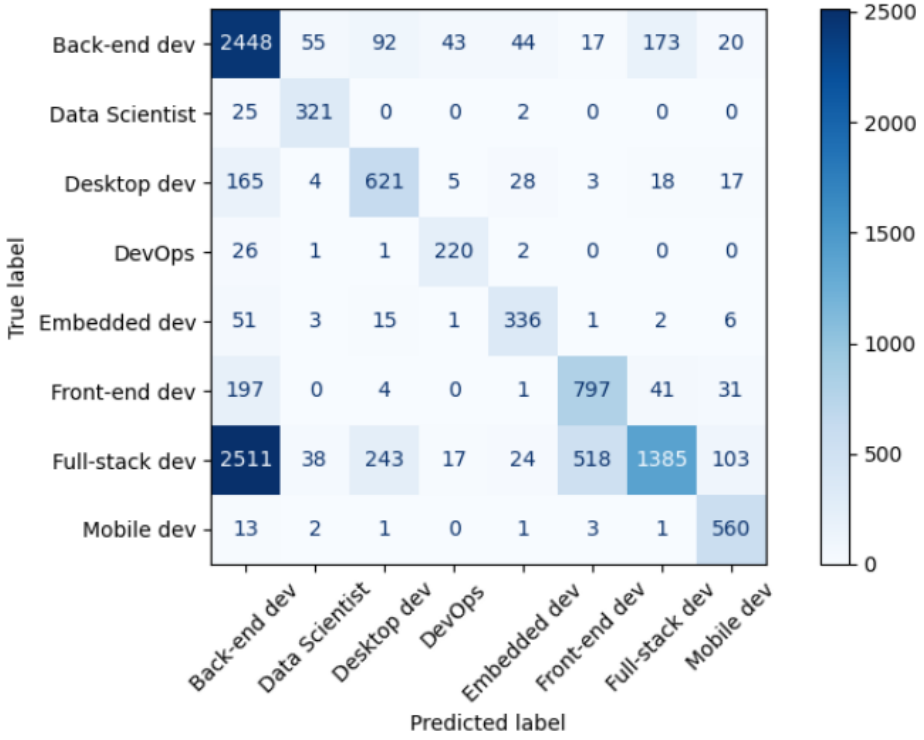


Modeling & Deployment
Best Model

Mean ROC curve for stacking model on test data



Confusion matrix of stacking model on test data



SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Thank You.
Any Questions ?

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations