

EDA for 2020-2021 Kaggle Machine Learning Data Science Surveys

Presented by: Amal ALThaqafi.





Outline

- Introduction.
- Dataset Description.
- Processing.
- Results.
- Conclusion.


Introduction

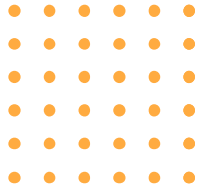
Analyzing data for a set of data related to Kaggle surveys to see the difference and the change between 2020 and 2021.



Dataset Description.

The datasets come from Kaggle "2020 Kaggle Data Science & Machine Learning Survey and 2021 Kaggle Machine Learning & Data Science Survey". The survey questions range from demographic questions, such as gender and level of higher education, to questions about programming languages, tools, and machine learning algorithms used. Most of the columns are categorical. The dataset has 25974 rows and 369 columns.





Processing



1

Filtering

To get columns that related to the questions



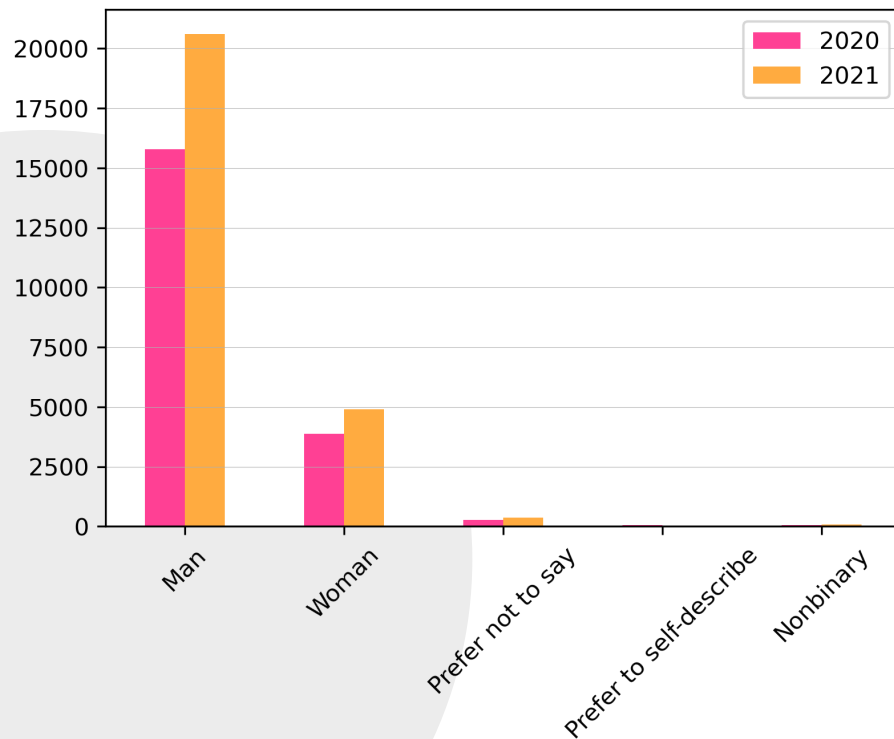
2

Visualization

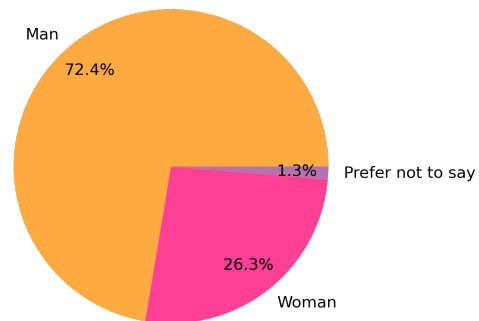
Using the Matplotlib and Seaborn

Results

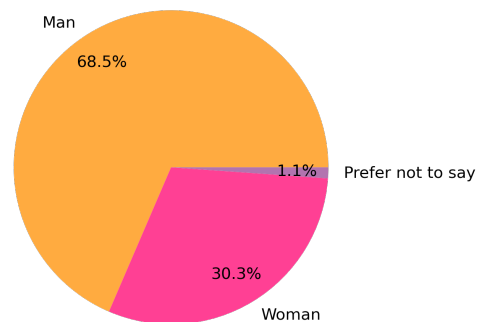
Gender Worldwide



Gender in Saudi Arabia 2020



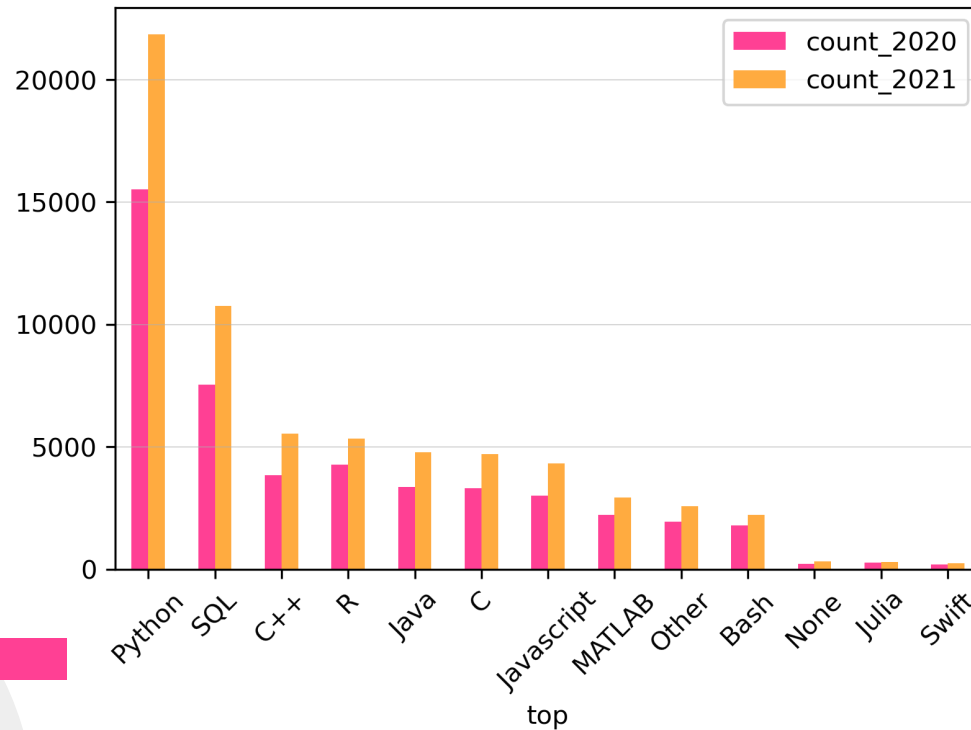
Gender in Saudi Arabia 2021





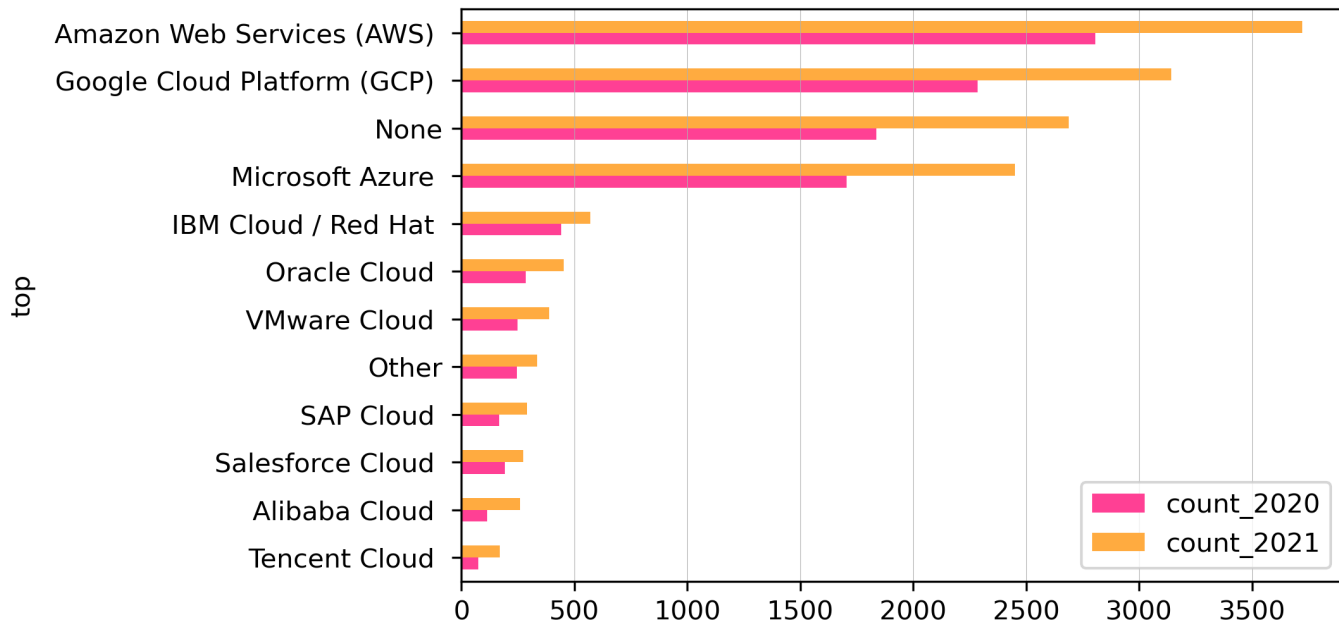
Results

The most popular programming language in 2020 and 2021



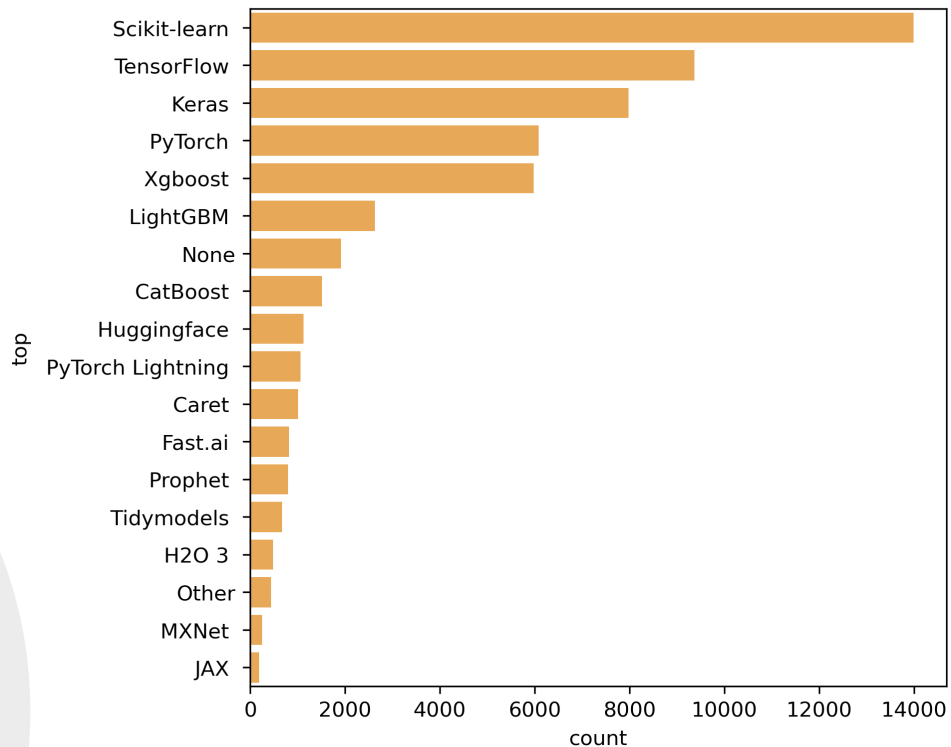
Results

The most used cloud platforms in 2020 and 2021



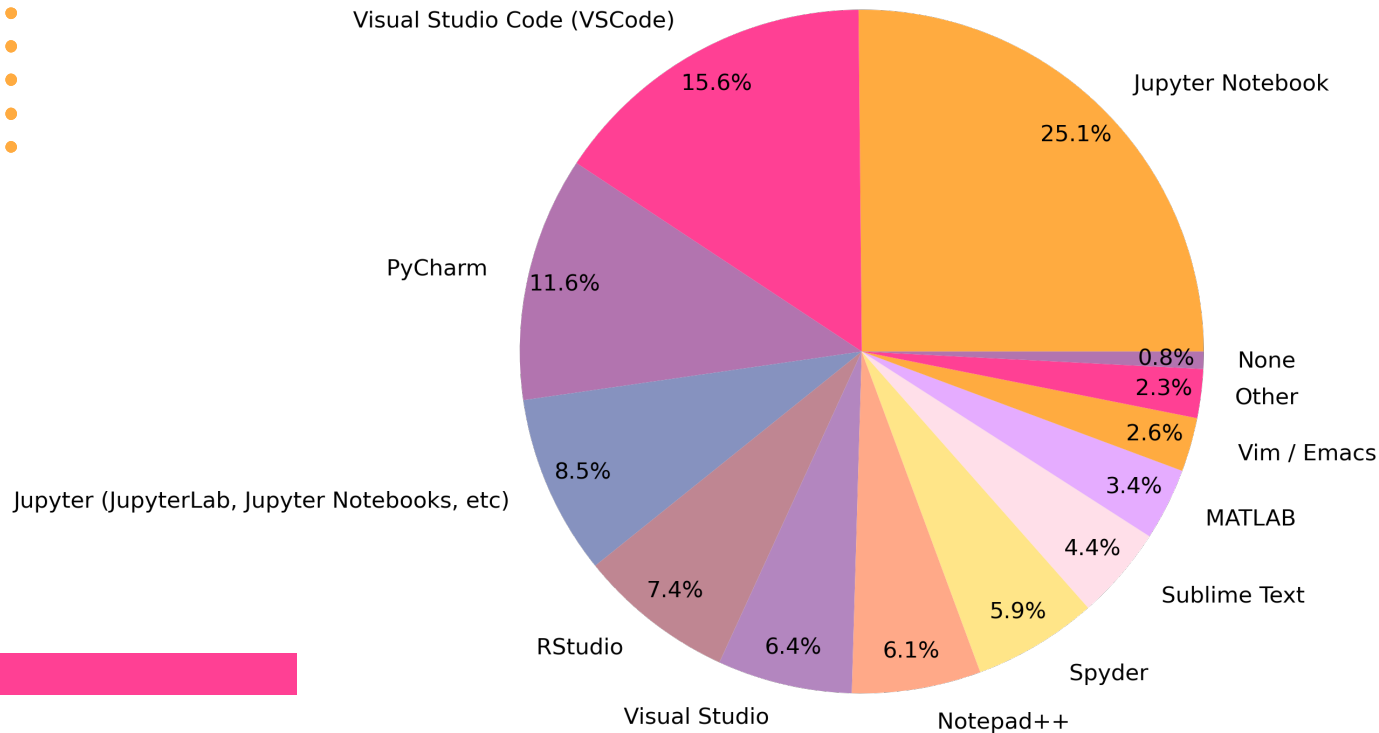
Results

The most common framework in 2021



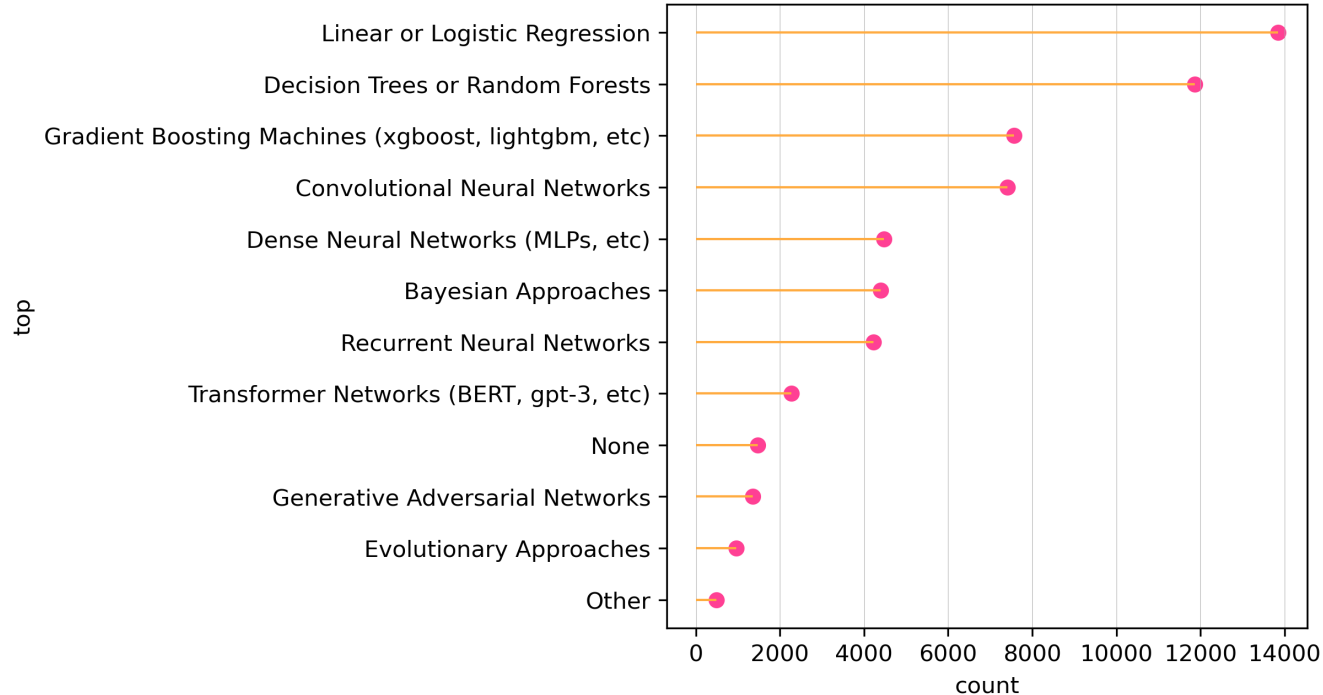
Results

The most popular Integrated Development Environment (IDE) in 2021



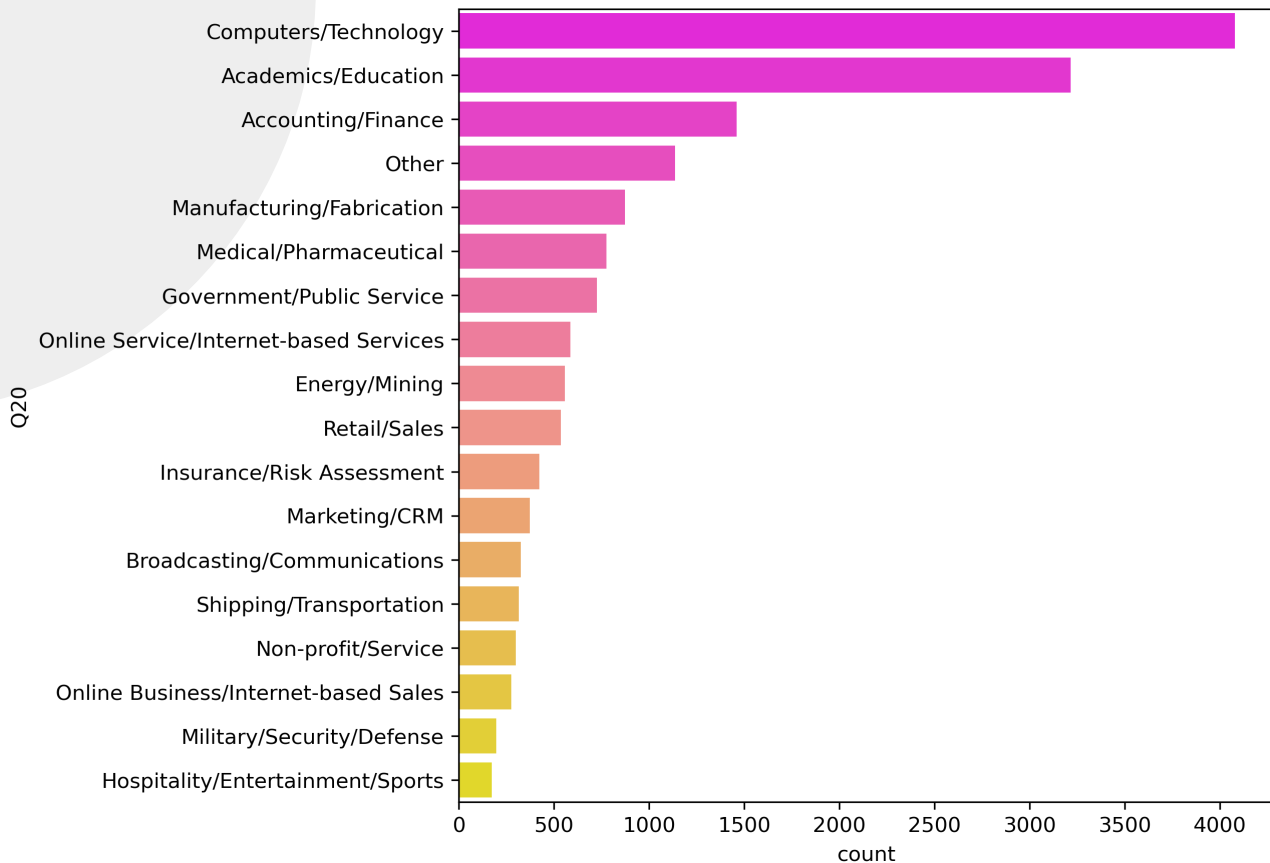
Results

The most popular ML algorithm used for Data Science 2021



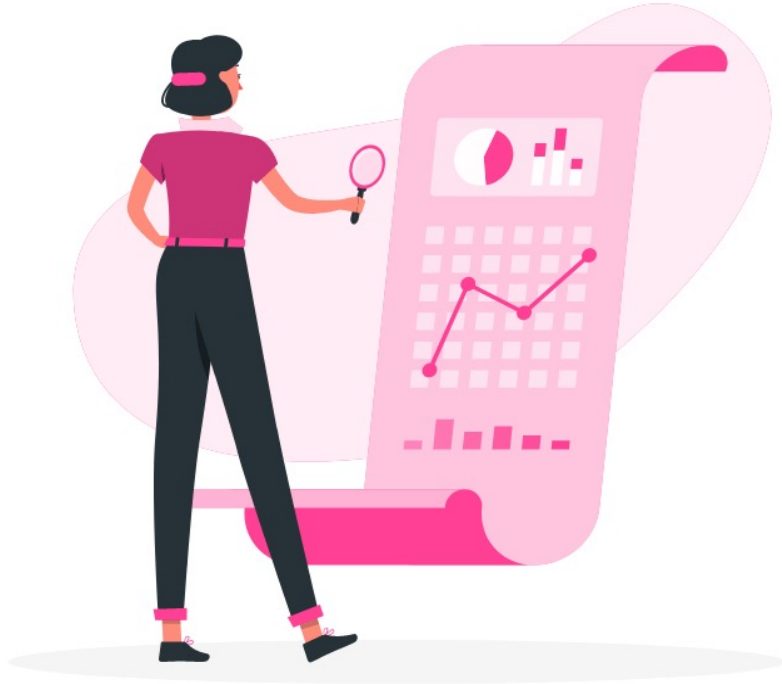
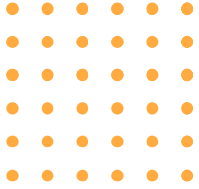
Results

The most common industry rely on Machine Learning products in 2021



Conclusion

- Increasing in the number of men using ML worldwide and the number of women worldwide and in SA.
- The most popular programming language in 2020 and 2021 is Python.
- The most used cloud platforms in 2020 and 2021 is AWS.
- The most common framework in 2021 is Scikit-learn.
- The most popular Integrated Development Environment (IDE) in 2021 is Jupyter Notebook.
- The most popular ML algorithm used for Data Science 2021 is Linear or Logistic Regression.
- The most common industry rely on Machine Learning products in 2021 is Computers and Technology.



THANK
YOU!