EDA of Kaggle Machine Learning & Data Science Survey (2021)

Abstract

This year, Kaggle set out to conduct an industry-wide survey that presents a truly comprehensive view of the state of data science and machine learning and the best ways for new data scientists to break into the field. Survey data provides an overview of the sector on an aggregate scale. The main objective of this project is to identify the most popular programming languages in 2021 and compare the results with previous years, find the relationship between salaries and years of experience, and identify the five countries that are most aware of data science.

Design

This project originates from the Kaggle survey competition. Kaggle a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Data

The dataset is provided in .csv format. It contains 369 columns and 25974 rows. . The columns are all the questions asked and they are detailed into many columns because the multiple answers were separated into more columns, but the basis of the questionnaire questions were 38 questions in the questionnaire and rows are responses from the Kaggle community. (https://www.kaggle.com/c/kaggle-survey-2021/overview)

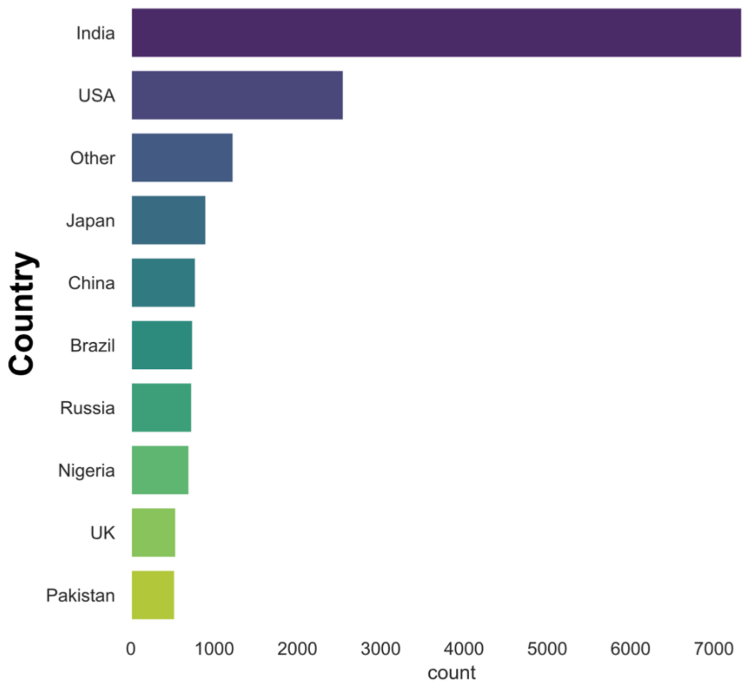
Algorithms

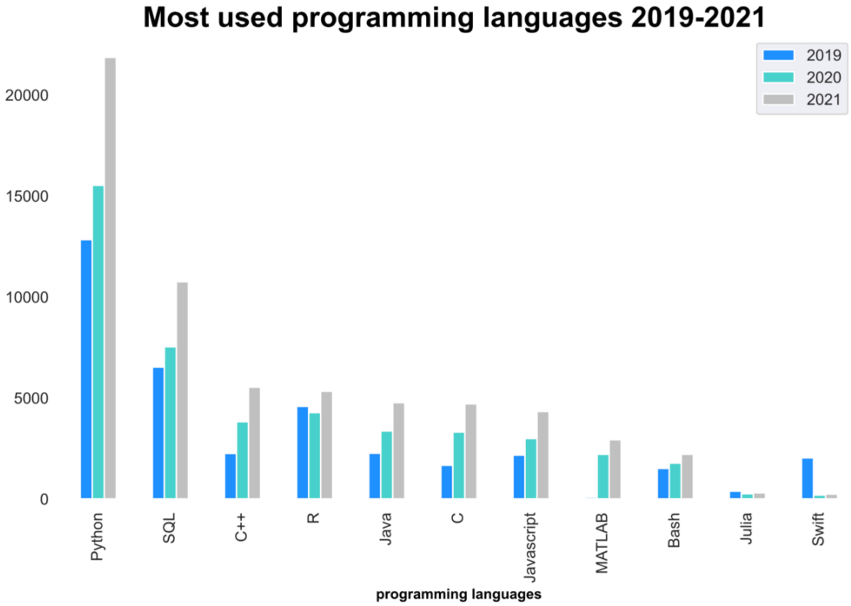
1- Checking the nulls
2- Drop heterosexuals from the data
3- Seeing the questions answered by 50% of the Kaggle community
4- Abbreviation of long country names
5- Adding Year column needed for exploring the data.
6- Comparison with other years
7- Convert variables to categorical so we can calculate correlation

Tools

- Python and Jupyter Notebook
- Numpy and Pandas for data manipulation
- Matplotlib ,Seaborn , for plotting visuialization

Communication


Country
India
USA
Other
Japan
China
Brazil
Russia
Nigeria
UK
Pakistan
count


Most used programming languages 2019-2021
Python
SQL
C++
R
Java
C
Javascript
MATLAB
Bash
Julia
Swift
programming languages