2020-2021 Kaggle Machine Learning Data Science Surveys

Amal AlThaqafi

Abstract

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

Design

This project originates from the Data Science Bootcamp (T5) to find answers to some questions which is:

- What is the most gender answering the surveys in Saudi Arabia and worldwide?
- What's the most popular programming language in 2020 and 2021?
- What's the most used cloud platforms in 2020 and 2021?
- What's the most common framework in 2021?
- What's the most popular Integrated Development Environment (IDE) in 2021?
- What's the most popular ML algorithm used for Data Science 2021?
- What's the most common industry rely on Machine Learning products in 2021?

by using the Kaggle surveys datasets through exploratory data analysis.

Data

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. These datasets are available as a comma-separated values (.csv) file.

Algorithms

Feature Engineering:

Divide the data set to two different dataframes and filter it to work on the columns that related to the questions and applied function to get the rate.

Visualization:

Using the Matplotlib and Seaborn to show the Top gender who answering the surveys in Saudi Arabia in 2020 and 2021 and compare it, show the gender worldwide in 2020 and 2021 and compare it, show the top popular programming language in 2020 and 2021 and compare it and the rate of increasing between the 2 years, show the most used cloud platforms in 2020 and 2021 and compare it, show the most common framework in 2021, show the most popular Integrated Development Environment (IDE) in 2021, show the most popular ML algorithm used for Data Science 2021 and the most common industry rely on Machine Learning products in 2021.

Tools

- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting

Communication

Presentation.