



Machine Learning & Data Science

Survey provided by **Kaggle**

Data Analysts:

Sebastian Iacoboni

Bogdan Panduru

Vlad Ianoși

Mihai Giuroiu



CUPRINS

1. Context
2. Rezultate
3. Vizualizarea datelor
4. Concluzii

1. Context

Studiul se adresează cercetătorilor sau pasionaților de ML & Data Science, indiferent dacă aceștia au experiență sau se află la început de drum. Scopul analizei a fost de a observa și cartografia parcursul respondenților în funcție de gen, țară, parcurs academic, cursuri studiate, abilități dobândite sau nivel ierarhic.

Pornind de datele colectate de Kaggle începând cu anul 2017, am încercat să evidențiem anumite particularități pe care prospectorii în ML & Data Science le au, în funcție de diverși factori. Analizele rezultate oferă idei sub formă de date ce pot servi atât în scop de dezvoltare profesională, cât și la nivel de afacere.

1.2 Ipoteze

După vizualizarea datelor, am presupus că există o direcție ce poate fi urmată. Am pornit de la datele demografice și am făcut legături cu platformele de studii alese, sectorul de muncă, dimensiunea companiei și venituri. În final, deși am concluzionat că nu există o hartă care te poate duce din punctul A în punctul B, am reușit să identificăm rezultate care pot contura idei.

Analizele făcute oferă informații specifice, ce pot servi pe post de indicații și modele pentru un cercetător aflat la început de drum și pe post de răspunsuri pentru un cercetător experimentat.

1.3 Probleme

În colectarea datelor nu au fost incluse persoane care nu au optat să se aboneze la newsletter-ul Kaggle. Din acest motiv, identificarea altor platforme de același tip care sunt folosite frecvent a fost imposibilă.

Pe lângă problemele referitoare la eșantionul de cercetare, au mai apărut dileme legate de verificarea acurateții răspunsurilor, însă verificările au făcut ca analizele referitoare la calitatea datelor să aibă rezultate pozitive.

2. Rezultate

Date Demografice

Vârstă, gen, locație

01

Programare

Limbaj folosit, experiență

03

ML and Cloud

ML/Cloud Computing acasă
sau la job în ultimi 5 ani

04

02

Cursuri

Platforme pentru cursuri

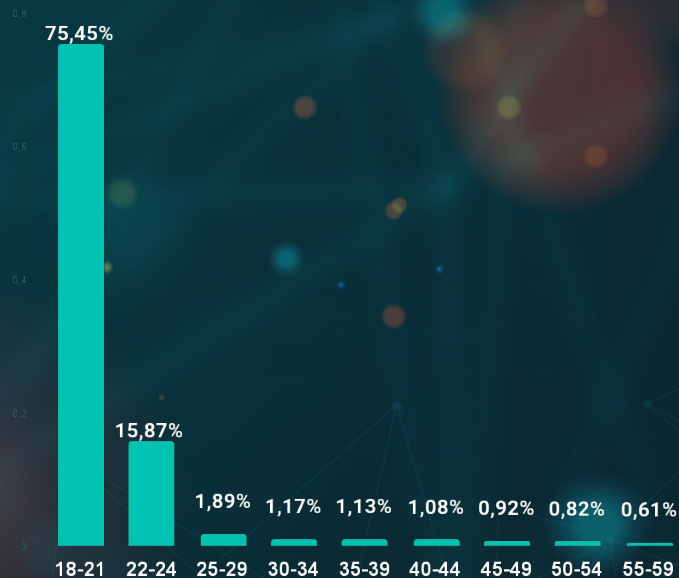
05

Job în industrie

Mărimea companiei,
salariul și numărul de
angajați în același domeniu

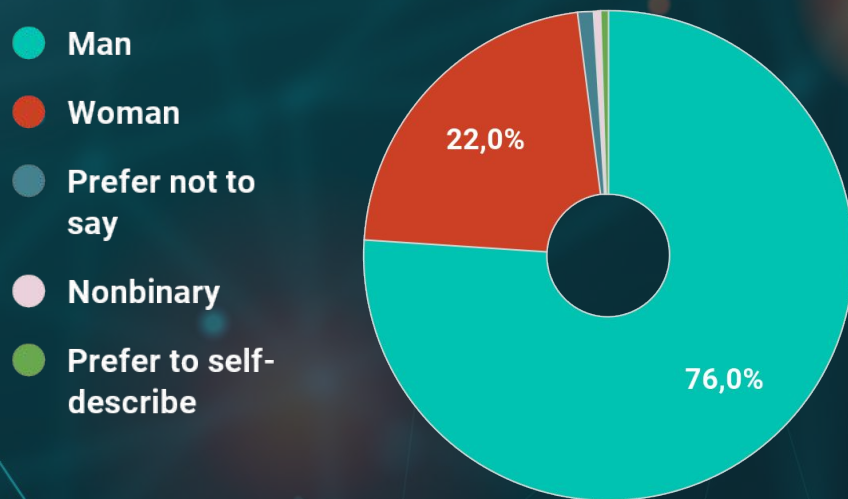
Care este distribuția pe vârste?

- 75% din respondenți au între **18 și 21 de ani**, în timp ce aproximativ 16% au între **22 și 24**.



Genul

Numărul **bărbaților** este de 3.45 ori mai mare decât numărul de **femei**.



Geographic



TOP 3

35% India

12% SUA

3% Brazilia

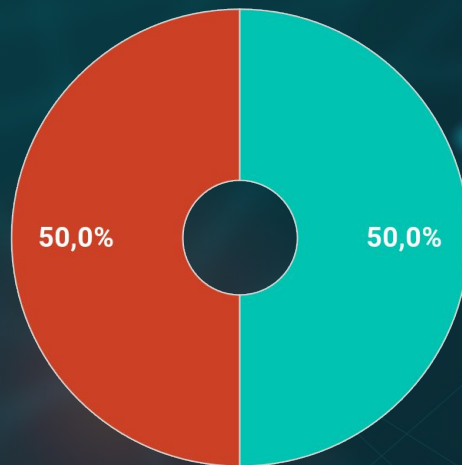
Cei mai mulți respondenți provin din **India** și **Statele Unite ale Americii**.

Mediul academic

Când vine vorba de mediul academic, jumătate din participanți sunt **studenți**.

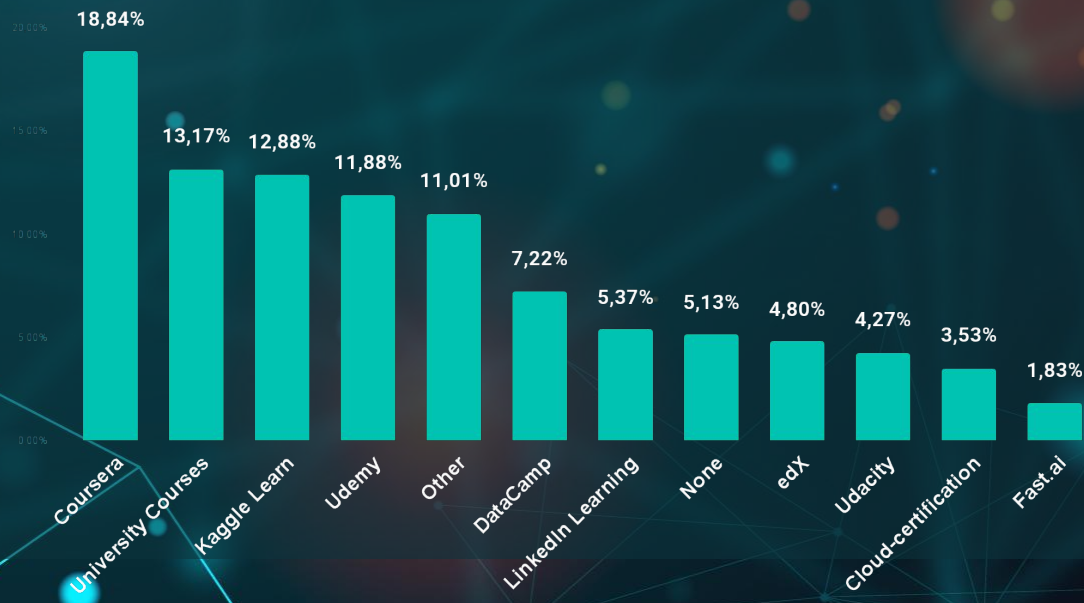
Are you currently a student?

- Yes
- No



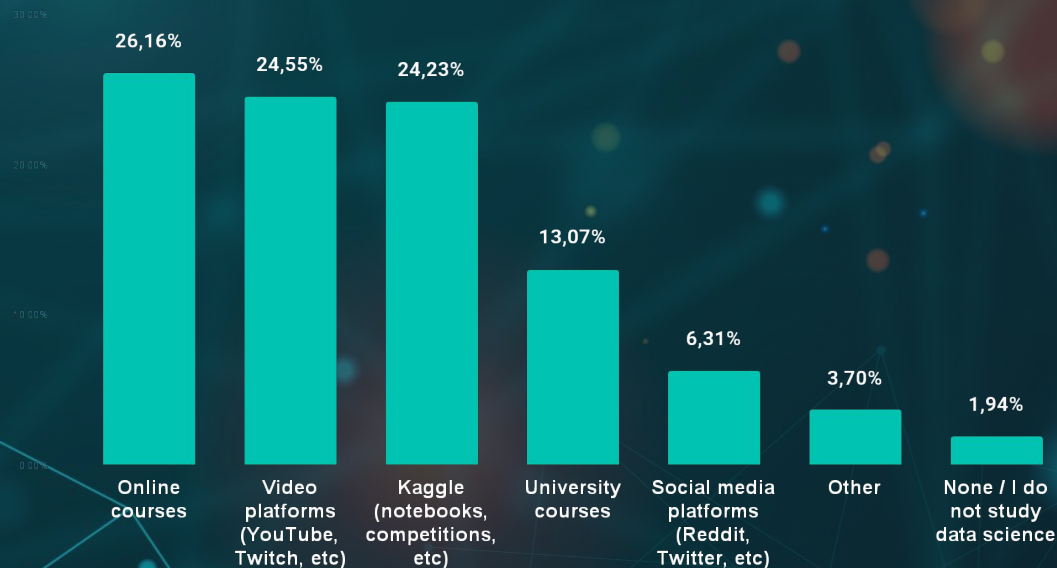
Care sunt cele mai folosite platforme pentru cursuri?

Coursera este platforma cea mai folosită. Un aspect interesant este însă distanța dintre următoarele 3 locuri. Putem observa varietatea cursurilor și numărul mare de oportunități.



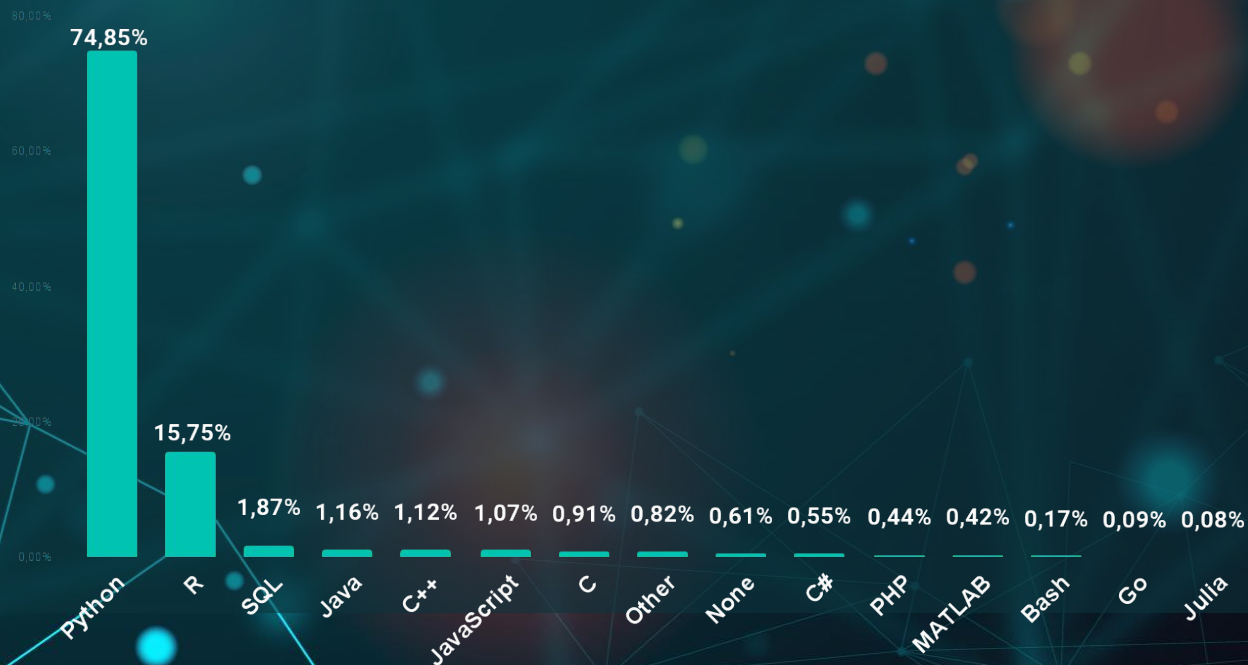
Care sunt principalele forme de învățare?

Învățatul **online** sau **de la distanță** sunt principalele forme de învățare.

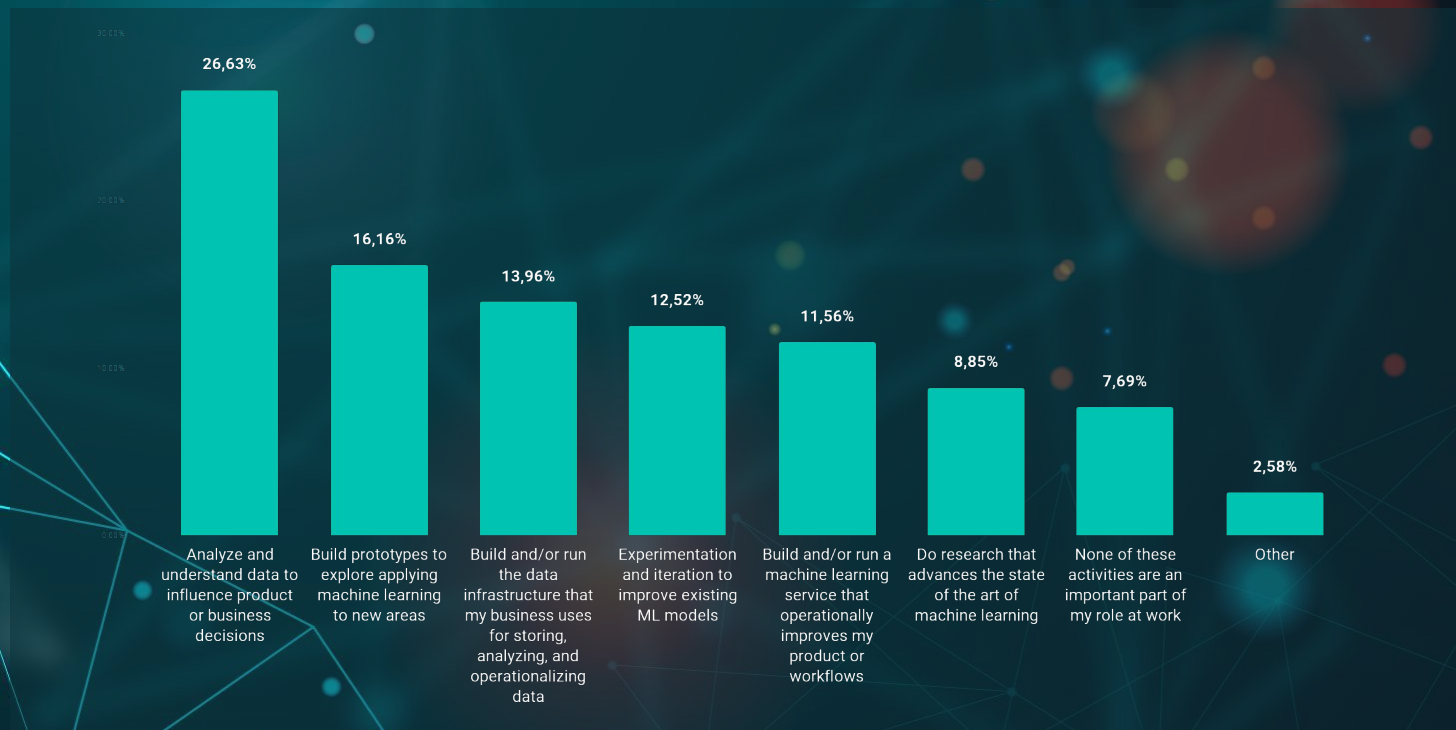


Care sunt principalele limbaje de programare folosite?

Python este principala alegere când vine vorba de limbaje de programare folosite.

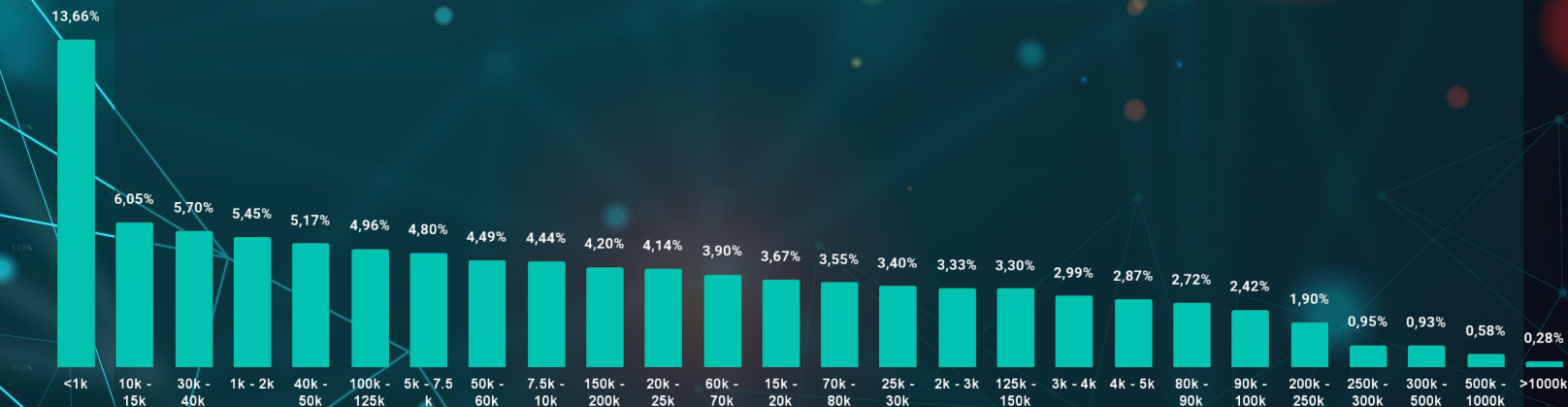


Ce activități se desfășoară cel mai des la locul de muncă?



Care este nivelul remunerației în acest domeniu?

Deși sumele nu sunt mari, cercetătorii în ML & Data Science nu investesc mult în cursuri și formare profesională. De asemenea, graficul reprezintă o justificare a faptului că domeniul rămâne unul foarte atractiv financiar.



Concluzii

În urma studiului putem observa că deși există diverse particularități comune, nu există un drum definit care asigură un rezultat.

Totuși, în urma analizelor, am reușit să cartografiem eficient trasee și metode de învățare. Analiza și rezultatele pot servi atât cercetătorilor aflați la început, cât și celor mai experimentați.

Cert este că, domeniile studiate rămân principale puncte de interes pentru cercetătorii din toată lumea. În acest sens, rămâne datoria fiecăruia să aleagă și să-și dezvolte propria carieră.

Metode de colectare și raportare a datelor

- Datele sunt raportate de cei de la **Kaggle**, iar pentru alegerea participanților a fost trimis **un mail tuturor persoanelor** care au optat pentru opțiunea de a primi mail-uri de la Kaggle.
- Pentru a asigura **calitatea datelor**, răspunsurile tuturor celor care au petrecut mai puțin de 2 minute pe chestionar sau au răspuns la **mai puțin de 15** întrebări au fost **șterse**.

Variable afectate de calitatea raportării	Cross-referencing	Estimarea
Fiecare întrebare cu răspuns multiplu. Datorită tipului de chestionar, fiecare răspuns neales, rămâne blank, aspect ce necesită mult data cleaning.	• Desk research • Familiarizarea cu domeniul și baza de date • Scopul și motivul survey-ului • Identificarea problemelor	• Rezultat general pe o singură platformă • Numărul mare de respondenți • Discrepanța dintre genuri

Metode de raportare a datelor

F

Puncte Forte

Răspunsuri prestabilite;
Date calculate anual;
Calitate superioară a datelor;

S

Puncte Slabe

Nu cuprind oameni care nu
Primesc mail-uri de la Kaggle;
Eșantionul nu e destul de divers;
Inabilitatea de a verifica acuratețea
răspunsurilor.



Mulțumim pentru atenție!

