

Academia de Studii Economice

Facultatea de Cibernetică, Statistică și Informatică Economică

Sistem suport de decizie pentru fundamentarea deciziilor pentru o companie de taxi

Sora Cornel

# Introducere

Obiectivul acestui proiect este de a construi un model de machine learning care va prezice durata călătoriei pentru o cursă cu o mașină de taxi în orașul New York City, bazându-se pe date din anul 2018 culese pentru firma NYC Yellow Cab.

Pentru a avea o prediciție cât mai corectă și cu o marjă de eroare cât mai mică se vor urma patru faze principale:

* Preprocesare: curățarea și transformarea datelor pentru a putea fi utilizate în analiză.
* Analiza exploratorie: se vor exploata datele pentru a identifica parametri necesari pentru a fi folosiți pentru modelare.
* Implementarea algoritmului: se va utiliza arborele de decizie pentru a genera regresia liniară pe baza căreia se vor putea face previziuni
* Utilizarea modelului pentru a efectua previziuni pentru un set de date de test.

**Informații despre arhitectura proiectului**

În scopul realizării proiectului a fost utilizat limbajul de progrmare pytho, prin intermediul căruia au fost citite și prelucrate datele și a fost construit modelul în urma căruia putem face predicția dorită, iar mai târziu decizii pe baza acesteia.

**Informații despre date**

Datele au fost culese din baza de date oferită de Google Cloud, iar acestea au fost salvate ]ntr-un fișier de tip excel, având următoarele coloane:

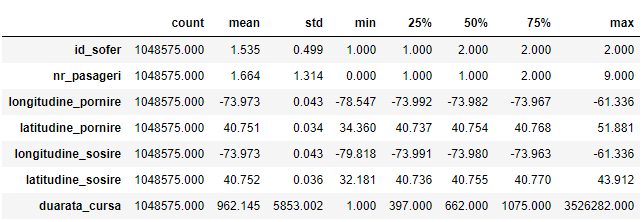
|  |  |
| --- | --- |
| **Nume variabilă** | **Descriere** |
| id | Identificator unic pentru fiecare cursa |
| id\_sofer | Identificator asociat soferului care a preluat cursa |
| data\_pornire | Data la care a fost solicitată cursa |
| data\_sosire | Data la care s-a încheiat cursa |
| nr\_pasageri | Numărul de pasageri din interiorul mașinii |
| longitudine\_pornire | Longitudinea de unde a fost solicitat taxi-ul |
| latitudine\_pornire | Latitudinea de unde a fost solicitat taxi-ul |
| longitudine\_sosire | Longitudinea unde cursa a fost încheiată |
| latitudine\_sosire | Latitudinea unde cursa a fost încheiată |
| cursa\_marcata | Y=cursa a fost înregistrată; N=cursa nu a fost înregistrată |
| durata\_cursa | Durata cursei măsurată în secunde |

# Componenta sistemului decizional

## Subsistemul de date

Mai întai datele vor fi încărcate și curătate într-o formă cât mai utila analizei. Pentru prelucrarea datelor este utilzat limbajul de programare Python. Datele sunt citite dintr-un fișier de tip CSV, excel. În urma citirii datelor se observă că fișierul conține 1048575 de linii.

Pentru a analiza dacă datele conțin anomalii sau dacă necesită modificări au fost calculate: media, dispersia, valoarea minimă și valoarea maximă pentru fiecare coloană:



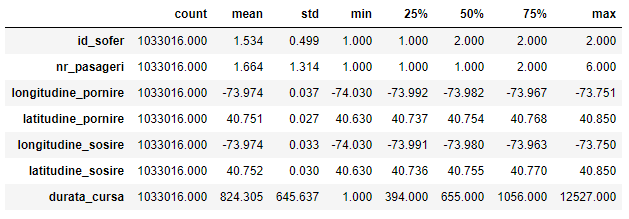
Conform tabelului de mai sus, se poate observa că analiza a fost facută pentru 2 șoferi, având id-ul 1 și respectiv id-ul 2. Numărul mediu de pasageri într-o mașină pentru o singură cursă este de 1.664, ceea ce ne arată că în mare parte din cazuri fost un singur pasager într-o cursă, în mai bine de 50% din curse. Numărul maxim de pasageri este de 9 pasageri într-o cursă, iar numărul minim este de 0, ambele arătând 2 valori anormale. De asemenea, anomalii se observă și în analiza coloanei ce arată durata cursei, fiind valori foarte mari.

Totodată, trebuie analizate și latitudinea și longitudinea punctului de plecare pentru a ne asigura că datele analizate sunt din interiorul orașului New York. Cunoscând că datele geografice despre New York City: latitudinea este între 40.7128 și 40.748817, iar longitudinea este între: -74.0059 și -73.968285, a fost analizată harta de mai jos pentru punctele care sunt în afara orașului New York.



Prin urmare, aceste date au fost eliminate, analiza fiind efectuata exclusiv pentru cursele din interiorul orașului.

În urma eliminării anomaliilor s-a obținut tabelul următor:



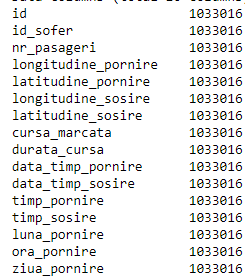
De asemenea, se poate observa că nu mai sunt locații în afara orașului New York.



Astfel, putem afirma în final că numărul de pasageri maxim a fost de 6, iar numărul minim: 1 singur pasager. Cea mai lungă cursă a fost de 2115 secunde, ceea ce înseamnă 35.25 minute. Valoarea medie pentru durata unei curse este aproximativ 728 de secunde, aproximativ 12 minute.

În urma eliminării anomaliilor, datele sunt în număr de 1033016, ce vor fi utilizate în pașii următori pentru realizarea modelului.

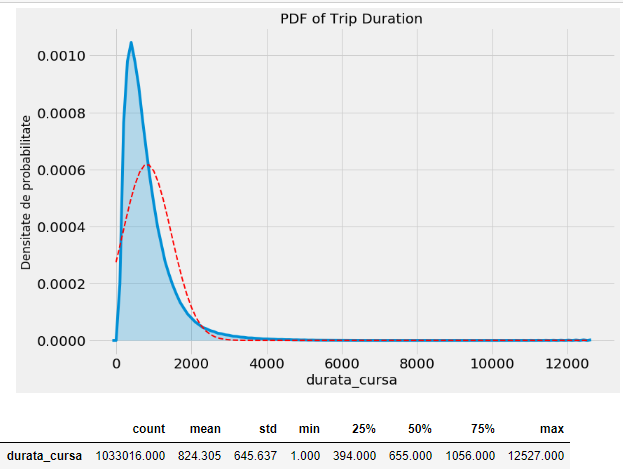
Data plecării, respectiv data sosirii, conțin atât informații despre dată cât și despre ora plecării/sosirii. Mai departe, vom separa această informație în: lună, ziua săptămânii, oră pentru data plecării pentru o analiză mai ușoară. Avem nevoie de aceste informații detalitate deoarece pot influența durata cursei. Nu este necesar acest proces și pentru data plecării deoarece, ziua și ora la care pasagerul este lăsat nu va influenta durata cursei, deoarce durata este înregistrată la final. Astfel, datele cu care vom continua analiza vor avea următoarele coloane:



### Analiza exploratorie a datelor

* Variabila dependenta: durata cursei

Variabila ce dorim să fie previzionată este durata cursei. Avem nevoie sa analizăm distribuția pentru a vedea dacă este nevoie de transformări asupra valorilor.



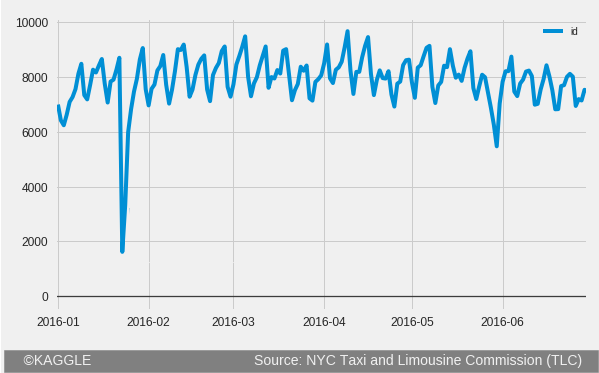
Conform graficului prezentat se observă clar că durata\_cursa este foarte înclinată spre dreapta, având valoarea asimetriei este mai mare ca 1.0, ceea ce ne arată că distribuția este înclinată spre dreapta, având foare multe vlori extreme spre dreapta.

Mediana pentru durata\_cursei este de 655 secunde, aproximativ 11 minute. Totuși, se observa că sunt și curse lungi de 12527 secunde, aproximativ 208 minute, aprox. 3 ore jumătate. Aceste valori mari nu au fost șterse în pasul anterior deoarece acestea sunt totuși între cele 2 valori ale deviației standar ale mediei.

* Variabilele independete

Coloanele rămase de analizat sunt incluse în variabilele independete. Aceste variabile for fi folosite pentru a antrena modelul de machine learning pentru a prezice variabila dependenta, *durata\_cursa*.

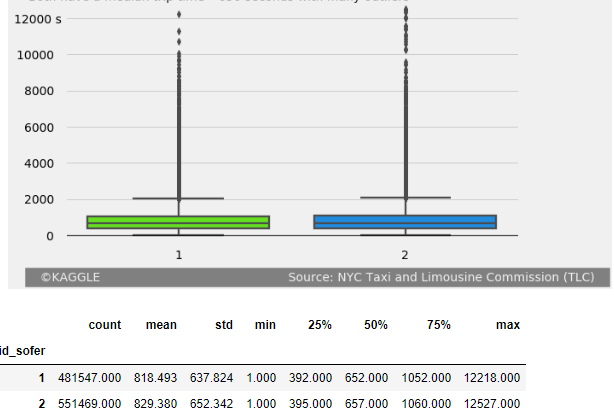
1. **Id**: variabila id este un identificator unic al fiecărei curse. Se va analiza cum variază aceata caractistică în timp. În mod clar, această variabilă nu influentează durata cursei, dar ne poate ajuta în analiza comenzilor pe diferite luni.



Se poate observa conform graficului că este un declin al valorii la finalul lunii Ianuarie. Ca primă vedere se poate spune că este vorba de sezonalitate, dar declinul este mult prea mare față de celelalte zile de iarnă înainte și după această scădere bruscă, ceea ce ne arată că în acea zi circulația taxi-urilor a fost aproape oprită.

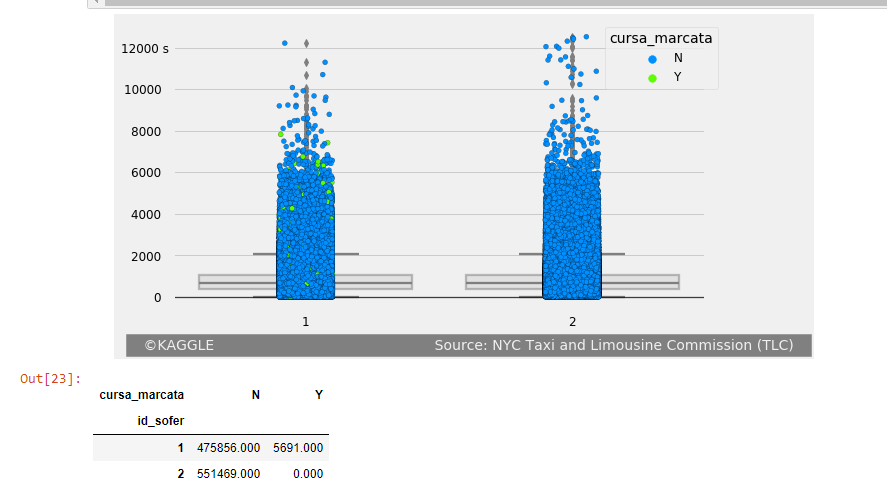
Deși variabila id, ne arată informații interesante în legătură cu traseele efectuate de firma de taxi, forma actuală a id-ului cursei nu va fi folositoare pentru a prezice durata cursei în algoritm.

1. **Id șofer**: această variabilă este un cod ce indică șoferul asociat fiecărei curse. Vom examina dacă un șofer are cursele mai lungi decât ceilalți.



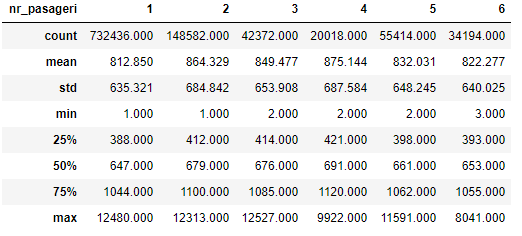
Conform figurii de mai sus se observă că nu sunt diferențe mari între timpul cursei efectuat de cei doi șoferi, observându-se valorile medianelor aproximativ egale: 652 și 657. Această varabilă va fi folosită în modelul pentru prezicerea duratei cursei, având o influență mică asupra duratei cursei. Totuși, dacă ne dorim ca un client final să fie un pasager, este dificil ca acesta să cunoască id-ul șoferului.

1. **cursa\_marcata**: această variabilă indică dacă ultima cursă efectuată este înregistrată în memoria mașinii înainte de a o trimite către producptor deoarece mașina nu a avut conexiune directă către server.



Se poate observa că doar șoferul cu id-ul 1 are curse care nu au fost marcate direct pe server. Astfel, această variabilă nu influentează cu mult predicția pentru durata cursei pentru șoferul cu id-ul 2, dar puțin pentru șoferul cu id-ul 1.

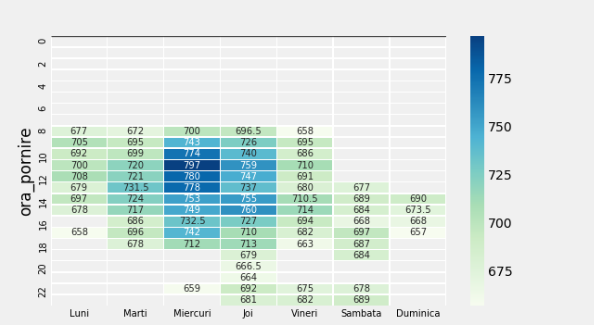
1. **Numărul de pasageri:**această variabilă este numărul de pasageri din mașină introduse de șofer. Prezumția este că acele curse cu mai mulți pasageri durează mai mult din cauza opririlor. Vom analiza dacă această variabilă are un impact și asupra datelor curente.



Conform tabelului rezultat, se observă că nuumărul de pasageri nu are o implicație foarte mare asupra duratei cursei, mediana variind foarte puțin. Totuși, această variabilă va fi folosită în model, ținând cont de ipoteză: pot apărea mai multe opriri ceea ce duce la creșterea timpului.

1. **Durata cursei în funcție de dată și oră.** Această variabilă influențează în mod clar durata cursei, traficul variind în funcție de zi și oră. Mai departe vom analiza dacă durata cursei durează mai mult cu adevărat în anumite zile față de celelalte.

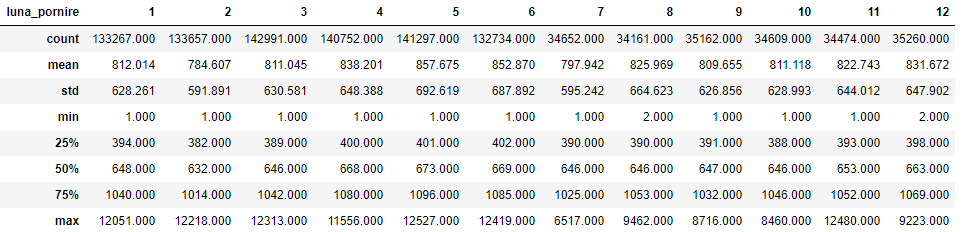
**Durata cursei în funcție de zi și oră**



Așa cum a fost precizat mai sus mediana pentru durata cursei este de 655 secunde, iar zilele momentele din săptămână care tind să depășească această valoare sunt:

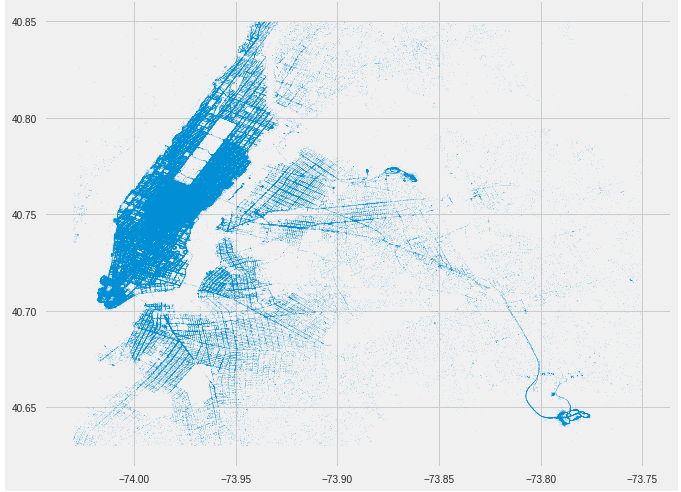
* luni – joi, între orele 8-18, reprezentând orele de muncă;
* joi – sâmbătă, între orele 18-24
* sâmbătă – duminică, între orele 24-1
* duminică, între orele 14-16

**Durata cursei în funcție de lună**

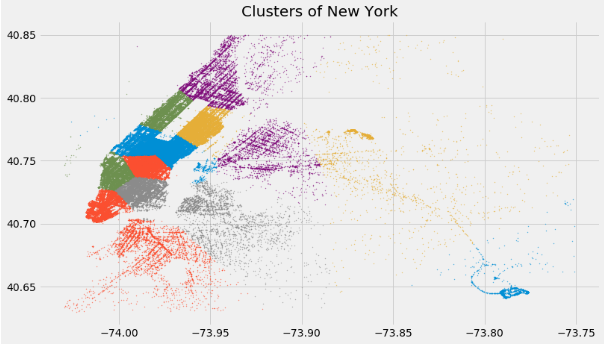


Se observă că mediana cu valoarea cea mai mare este în luna Mai. Totuși, se observă că nu variază foarte mult de la o lună la alta.Ceea ce ne arată că luna nu va fi neapărat un ajutor în predicția pentru variabila: durata\_cursa. Se poate afirma astfel, că luna are un impact minimal.

**Grafic cu punctele de plecare și sosire în orașul New York**

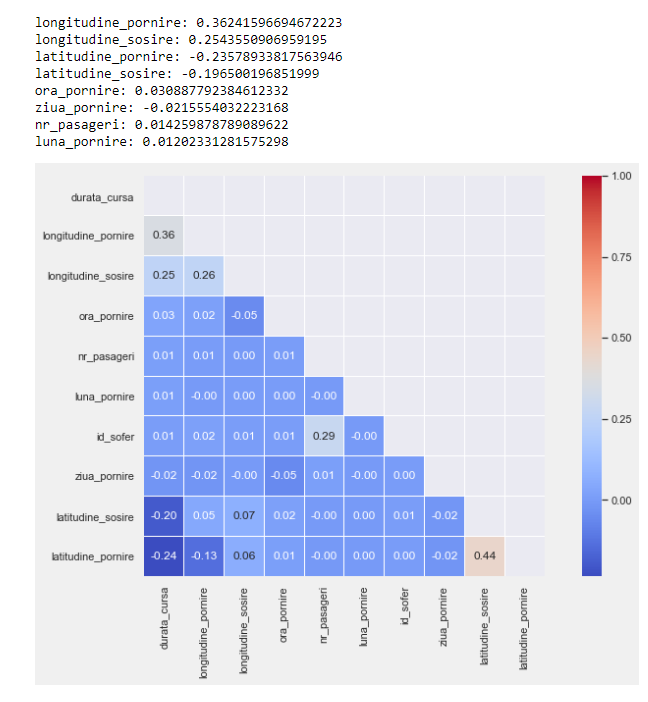


**Clusterizarea datelor în funcție de regiune în orașul New York**



**Relația dintre variabile**

În urma exploatării, vom analiza în continuare corelația dintre date. Dacă coeficientul de corelație dintre variabilele independente este mai mare de 0.7 putem renunța la acele variabile, prezentând informație redundantă.



Prin urmare, coeficientul de corelație ale coordantelor: latitudine\_plecare, longitudine\_plecare, latitudine\_sosire, longitudine\_sosire indică o relație liniară între ele. Totuși, coeficientul de corelație este mai mic de 0.5 și nu este suficient de mare încât să ștergem aceasta valoare.

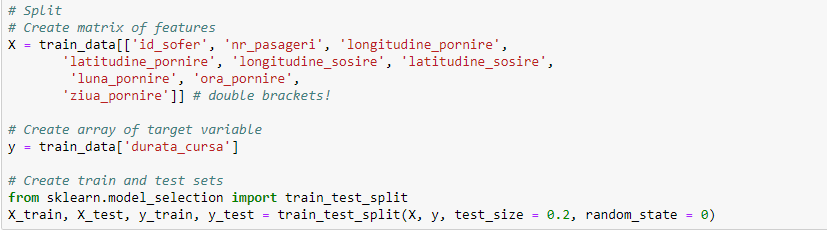
# Susbistemul de modele

Conform analizei efectuate mai sus, coloanele ce vor fi folosite pentru realizarea modelului din datele inițiale sunt: **id\_sofer**, **nr\_pasageri**, **longitudine\_pornire**, **latitudine\_pornire, longitudine\_sosire**, **latitudine\_sosire**, **luna\_pornire**, **ora\_pornire**, **ziua\_pornire** ca variabile independente, iar variabila dependeta este **durata\_cursa**.

Prin urmare, variabilele au fost alese din coloanele prezente in excel, anomaliile ce puteau duce la rezultate eronate au fost eliminate, astfel modelul fiind gata pentru implementarea modelul arborelui de regresie.

A fost ales algoritmul arobre de regresie deoarece conform analizei datelor modelul este unul non-liniar. Arborele de regresie este un model de estimare (în acest caz) folosit în situația în care se dorește o prognoză pentru o variabilă dependentă.

Pentru a antrena modelul, arborele de regresie, avem nevoie de un set de date de antrenament și un set de date de test. Pe baza datelor de antrenament se va construi arborele de decizie, iar pe baza acestuia se poate face prognoza datelor de test și astfel se pot determina metricele acestui algoritm. Astfel, vom extrage din datele inițiale 80% pentru a antrena modelul, iar 20% pentru a testa arborele de decizie rezultat. Pentru a selecta datele este folosit în continuare limbajul Python, selectând 80% din date aleator prin funcția *train\_test\_split*.



* train\_data reprezintă datele analizate și modificate
* X reprezintă variabilele independente
* y reprezintă variabila dependentă
* X\_train, y\_train reprezintă datele de antrenament
* X\_test, y\_test reprezintă datele pentru testarea rezultatelor

Mai departe, modelul de arobre de decizie este antrenat, iar în urma acestei etape putem face predicții pe baza lui. În acest scop a fost folosit DecisionTreeRegressor din biblioteca sklearn în limbajul Python. Acest pachet prezintă clasa modelul *tree* cu metodele: *fit* pentru antrenarea modelului și metoda *predit* pentru a realiza predicția pe baza arborelui de decizie.

