**Analiza Exploratorie**

Petcu Luminita Andreea, Nicula Ruth, Toth Akos

Setul nostru de date conține rezultatele analizei de date efectuate pe un set de date care conține informații despre procesul de fabricare a articolelor de îmbrăcăminte și productivitatea angajaților din industria confecțiilor pentru anumite grupe.

Producerea se poate categoriza dupa cele 5 parti ale lunii, 2 departamente, 6 zile, 12 echipe.

Sursa setului de date <https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees>

Primul tabel este obținut din depozitul original al UCI, care conține descrierea fiecărui atribut.

\* #Partile lunii (quarters) sunt repartizate in felul urmator

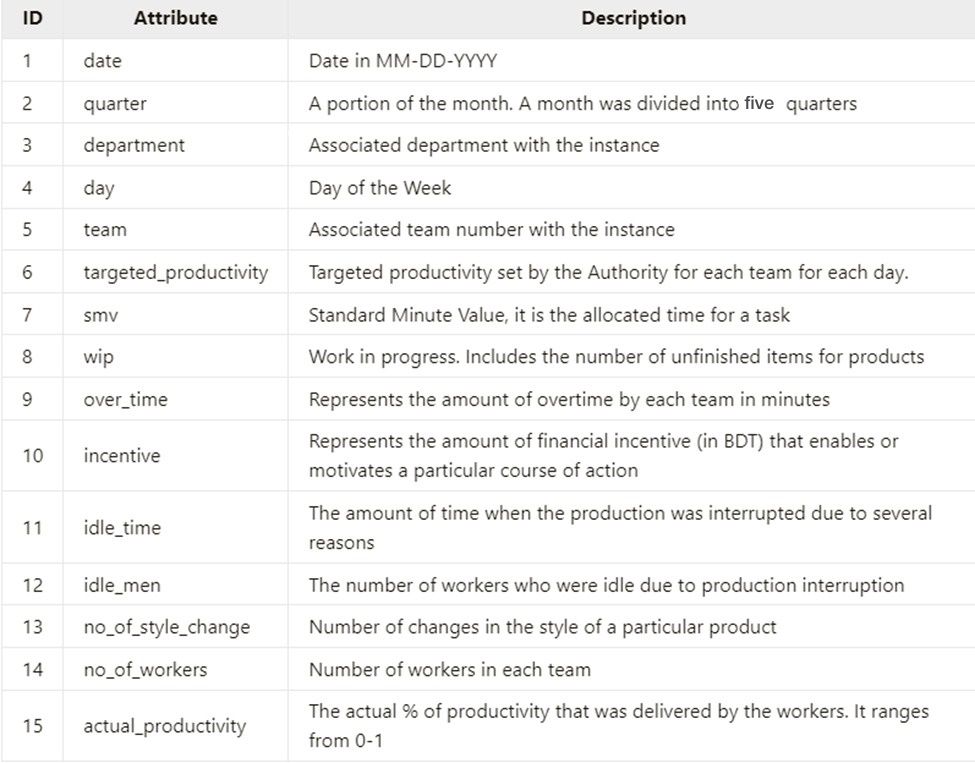
# -Quarter1: zilele 1-7 ale lunii

# -Quarter2: zilele 8-14 ale lunii

# -Quarter3: zilele 15-21 ale lunii

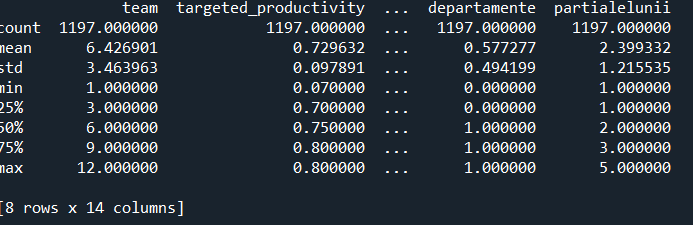
# -Quarter4: zilele 22-28 ale lunii

# -Quarter5: zilele 29-31 ale lunii



 Prelucrarea in python – citirea setului de date

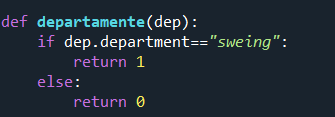
Folosind comanda df.describe(), am aflat că setul de date conține 1197 de înregistrări.



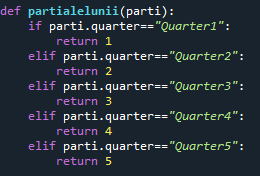
Datele provin din trei luni ale anului 2015 mai exact intre (01.ian si 03.mar) includ informații precum department, echipa care produce produsul, procentul de target al productivitatii si actualulul procent al productivitatii, numărul de lucrători ai echipelor, orele suplimentare.

Am analizat următoarele date:

1. Analiza totala de apariții departamente, zile si parti ale lunii între ian.2015 si mar.2015(transformare variabile categoriale in variabile dummy)







Nu am putut elimina celulele null ale coloanei „wip”, dar am inlocuit celulele Nan cu valoarea 0.

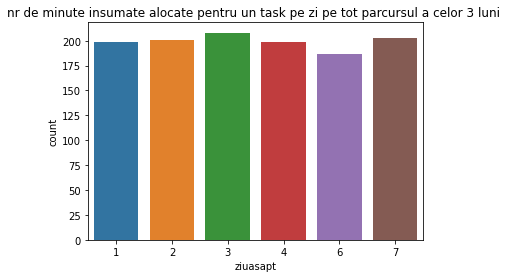


Am analizat:

- numarul de minute alocate pentru taskuri pe zile.

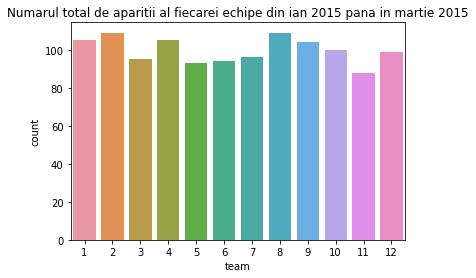
\*observam ca ziua 5 nu apare, deci putem considera ca vinerea este probabil zi libera

\*activitățile de producție sunt repartizate destul de uniform în timpul săptămânii și în weekend, cu excepția zilei de vineri, care pare a fi o zi liberă.



-numarul total de aparitii al fiecarei echipe(ian-mar 2015)

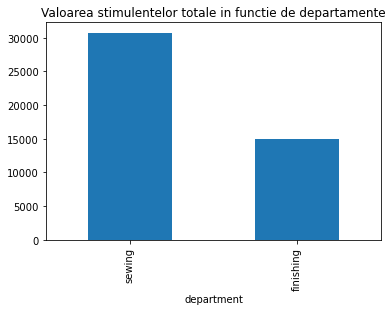
\*observam ca echipele sunt repartizate destul de uniform, in medie cam 90 de aparitii/echipa. De cele mai multe ori echipa 8 este prezenta cu 110 aparitii, iar de cele mai putine ori echipa 11 cu 90 de aparitii





-valoarea stimulentelor totale in functie de departamente

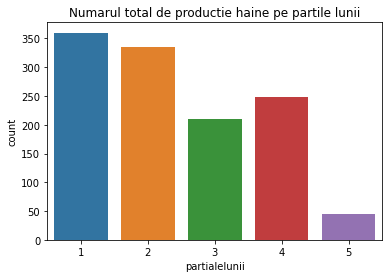
\* se observa ca cele mai multe stimulente sunt acordate depatamentului de cusut, fiind duble fata de cele de la finisare



-numarul total de productie haine repartizat pe cele 5 parti ale lunii

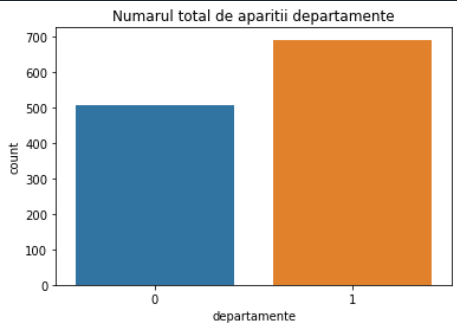
\* Mai mult de jumătate din activitățile de producție au loc în primele două trimestre ale lunii.

\* In ultima parte a lunii insemnand intervalul de zile 29-31, s-a inregistrat cea mai slaba productie.



-analiza repartizarii celor 2 departamente;

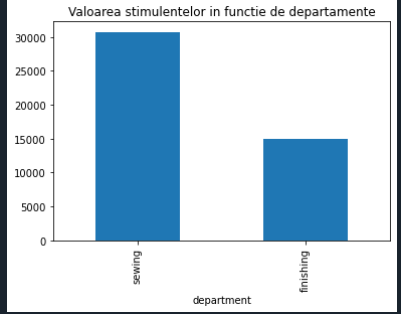
\*În total, există cu aproape 200 de sarcini de cusut în plus față de cele de finisare.



Am analizat:

-reprezentarea sumei de stimulente pe categorii de departamente;

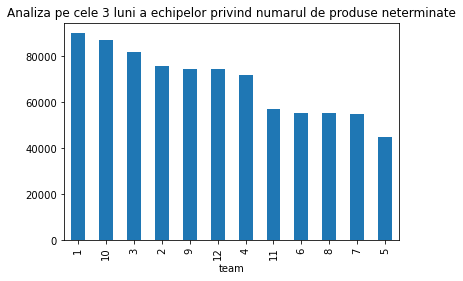
\*Prima linie de cod modifică valorile din coloana "department" prin aplicarea unei funcții lambda la fiecare valoare. Funcția lambda verifică dacă valoarea, după eliminarea spațiilor, este egală cu "finishing". În caz afirmativ, înlocuiește valoarea cu "finishing".





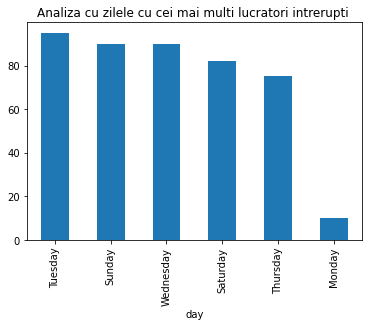
-echipele cu numarul de produse neterminate

\* observam ca echipa 1 dispune de o productie in curs de executie cu un nr. de produse de 9000, aceasta fiind cea mai mare, pe cand, la polul opus, se afla echipa nr. 5 cu 45000 de produse neterminate



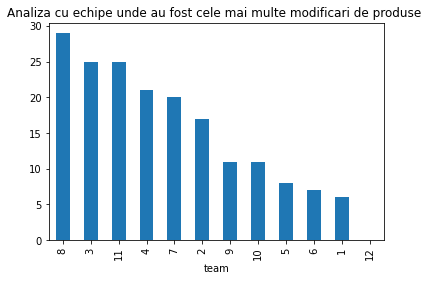
-top 3 zile cu cei mai multi lucratori intrerupti

\* In majoritatea timpului, lucratorii intrerupti sunt impartiti destul de uniform, cu exceptia zilei de luni unde observam ca sunt cei mai putini lucratori neintrerupti, in numar de 10.



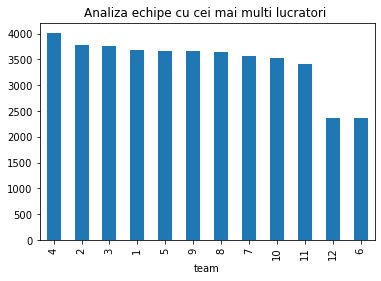
-top 5 echipe cu cele mai multe modificari de produse

\* In graficul rezultat, se evidentiaza echipa 8 cu cele mai multe modificari de produse (28), la capatul celelalt aflandu-se echipa 12 cu 0 modificari de produse pe parcursul celor 3 luni



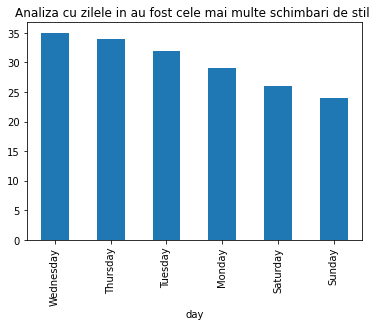
-echipe cu cei mai multi lucratori

\* Observam ca pe parcursul tuturor echipelor avem uniform impartiti muncitorii. Observam ca, in capatul listei se afla echipa 12 si 6 cu 2300 de muncitori



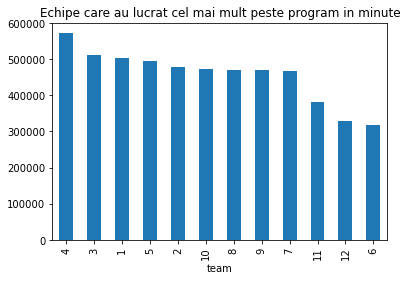
-zilele in care au fost cele mai multe schimbari de stil

\* In zilele de miercuri au existat cele mai multe schimbari de stil(35), iar in zilele de duminica au fost cele mai putine schimbari de stil(25)



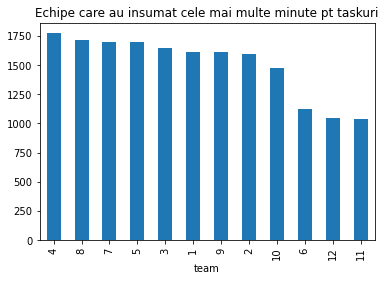
- echipe care au lucrat cel mai mult peste program/minute

\* Observam ca echipa 4 a avut cea mai mare productivitate peste program(580000 minute), in capatul graficului se afla echipa 6 cu cele mai putine minute petrecute peste program(350000). In general, majoritatea echipelor au avut o uniforma repartizare privind munca peste program



-echipe care au insumat cele mai multe minute pentru taskuri.

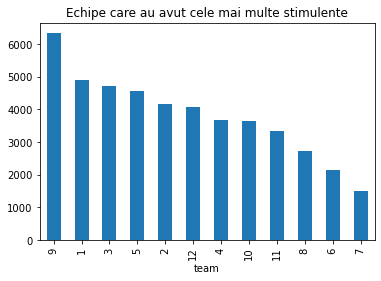
\* La prima vedere, graficul arata minutele taskurilor a fiecarei echipe in cele 3 luni. Se vede ca echipa 4 a insumat cele mai multe minute(1800), pe cand echipele 11 si 12, cele mai putine (1100)



-echipe care au avut cele mai multe stimulente

\* Din graficul cu bare de mai jos, putem vedea clar că echipa 9 a primit cea mai mare sumă de stimulente, iar majoritatea celorlalte echipe au primit 3/4 din această sumă.

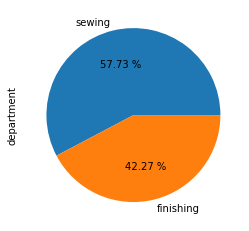
\* Mai mult, observam ca echipa 7 a beneficiat de cele mai putine stimulente (1800)



Am folosit pieplot pentru a vedea cat inseamna cele 2 departamente proportional din 100%.

\*Prima linie de cod modifică valorile din coloana "department" prin aplicarea unei funcții lambda la fiecare valoare. Funcția lambda verifică dacă valoarea, după eliminarea spațiilor, este egală cu "finishing". În caz afirmativ, înlocuiește valoarea cu "finishing".

\* Se remarca impartirea departamentelor aproximativ uniforma. Departamentul de cusut are o proportie mai mare, de 57.73 %, pe cand departamentul de finisare de 42.27%





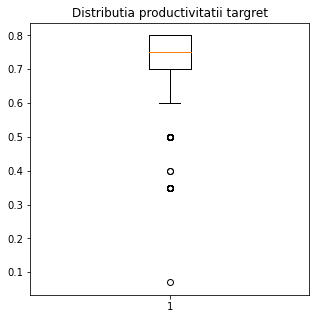
Am analizat:

-distributia productivitatii target

\* boxplot care ne ajuta la analiza distributiei, permitand sa vedem valorile minime,prima valoare, mediana, a treia valoare si valoarea maxima

\* Se poate vedea ca numarul general de target de productivitate este de 0,75 adica 75%.

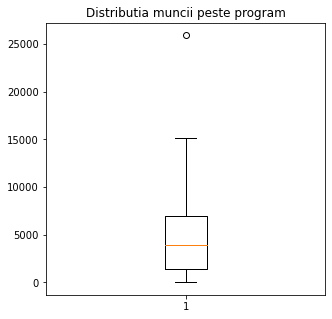
\* Tinand cont de linia medianei se afla la mijlocul casetei, se poate spune ca valorile impuse sunt simetrice





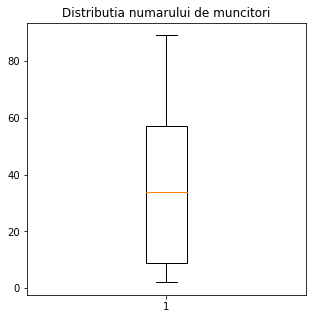
-distributia muncii peste program

\* Se observa ca minumul de lucru peste program este 0, maximul este 25920(punctul de sus), iar mediana 4567.460317



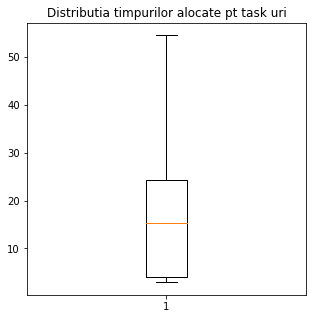
-distributia numarului de muncitori

\* Se observa ca minimul de muncitori este 2, maximul de muncitori este 89, iar media este 35



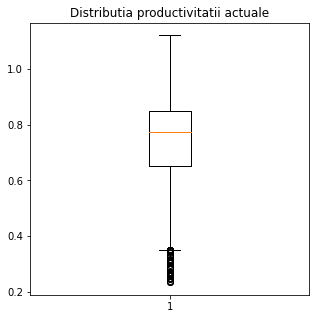
-distributia timpului alocat pentru task-uri

\* #Minimul de timp alocat pentru task-uri este 3 minute, maximul de timp alocat este de 55, iar media de 15 minute



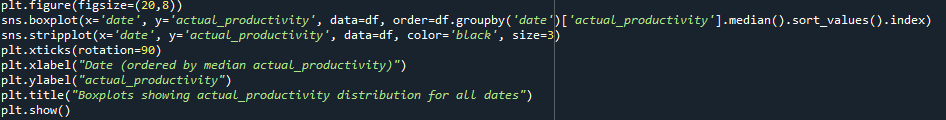
-distributia productivitatii actuale

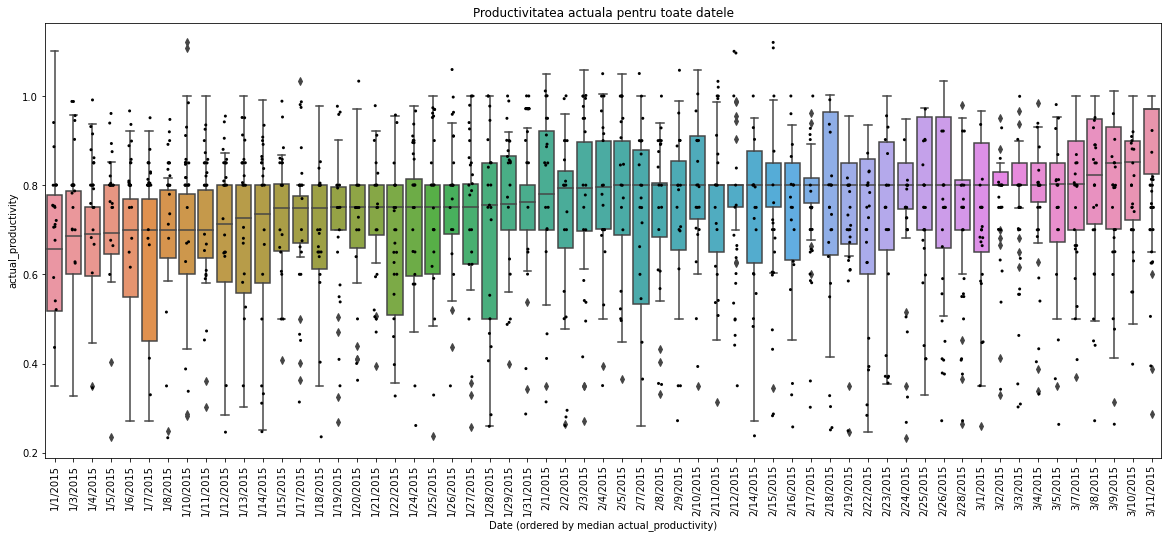
\* Minimul de productivitate este de 0,24(primul cerc), maximul de productivitate este de 1.12, media 0.73



Relatia distributiei productivitatii actuale pentru toate datele am aflat-o prin urmatorul cod:

\*Acest lucru confirmă ca trimestrul 5 este în medie mult mai productiv decât celelalte trimestre.

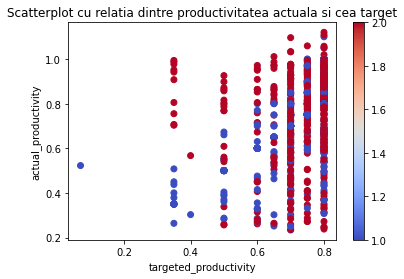




-Relatia dintre productivitatea actuala si cea target

\* Punctele sunt în mare parte grupate de-a lungul unei linii drepte cu pantă ascendentă, acest lucru indică o corelație pozitivă între productivitatea actuala si cea target, ceea ce înseamnă că atunci cand productivitatea target a fost mare si cea actuala a fost la fel.

\* există câteva puncte care se îndepărtează în mod semnificativ de grupul principal, acestea ar putea fi valori aberante și ar merita o investigație suplimentară. De exemplu, o productivitate target scăzuta, dar o productivitate reală ridicată, ar putea indica faptul că obiectivul a fost stabilit prea jos sau că au existat evenimente sau factori neașteptați care au contribuit la creșterea productivității. O productivitate țintă ridicată, dar o productivitate reală scăzută ar putea indica faptul că obiectivul a fost stabilit prea sus sau că au existat probleme care au împiedicat productivitatea.

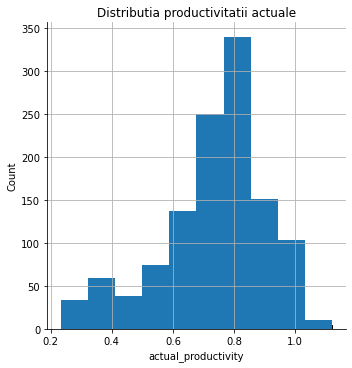


**Am analizat:**

-distributia productivitatii actuale

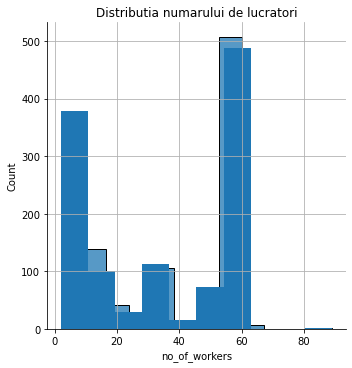
\*actual\_productivity are mai multe valori inferioare, ceea ce poate duce la o distorsiune negativă, deoarece se poate observa o coadă mai lungă în stânga.

\* Coada mai lunga in stanga a unei distributii inseamna ca exista o concentratie mai mare de valori mai mici decat media



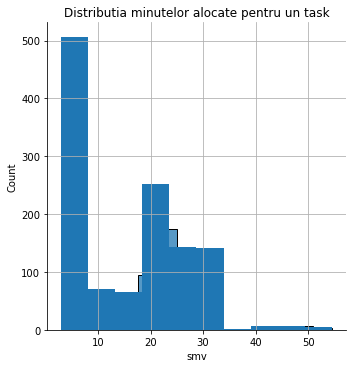
-distributia numarului de lucratori

\* Numărul de lucrători dintr-o echipă poate varia foarte mult, majoritatea având în jur de 8 sau 58 de lucrători.



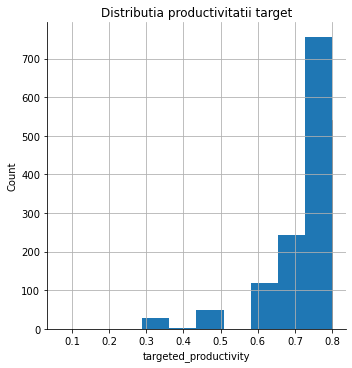
-distributia minutelor alocate pentru un task

\* Timpul alocat pentru o activitate este de obicei în limita a 20 de minute, majoritatea fiind de aproximativ 3 minute.



-distributia productivitatii target

\* Productivitatea țintită este stabilită la 0,8 în cea mai mare parte a timpului.



\*Cele mai multe dintre atributele noastre sunt puternic înclinate, atât pozitiv, cât și negativ, cu excepția probabil a productivității\_efective și a echipei, care par a fi aproximativ normale și, respectiv, uniforme.

\*12 echipe care lucrează cu o distribuție destul de uniformă a activităților. Echipele 1 și 2 au cel mai mare număr de activități.

**MEDIANE**

\* Mediana coloanei "smv" este 15,26;iar mediana coloanei "wip" este 586.0

\*Aceasta înseamnă că jumătate din valorile din coloana "smv" sunt mai mari sau egale cu 15,26. (smv) si cealalta jumatate mai mari decat 15.26

\*asemenea la coloana “wip” cu 580.6.

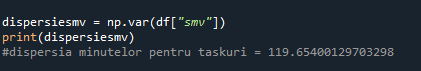
**DISPERSII**

\* Rezultatele oferă o indicație despre cât de răspândite sunt valorile din coloanele "smv","wip,"over\_time".

\* O valoare ridicată a varianței înseamnă că valorile sunt repartizate pe un interval mare, în timp ce o valoare scăzută a varianței indică faptul că valorile sunt grupate în apropierea mediei.

\* In acest caz, varianța coloanei "smv" este de 119.65400129703298,varianța coloanei "wip" este de 3371354.8595525227 si varianța coloanei "over\_time" este de 11205250.31692982 ceea ce sugerează că valorile din coloană sunt răspândite pe un interval relativ mare.

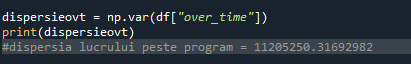
Comentariile cu # sunt ceea ce reprezinta fiecare dispersie si rezultatul acestora .



#dispersia minutelor pentru taskuri = 119.65400129703298



#dispersia lucrului in progres = 3371354.8595525227



#dispersia lucrului peste program = 11205250.31692982

Deviația standard este o măsură a dispersiei valorilor unui set de date în jurul mediei. Aceasta indică cât de mult se departează valorile de la valoarea medie. Cu cât deviația standard este mai mare, cu atât valorile sunt mai dispersate în jurul mediei. Dacă deviația standard este mică, aceasta înseamnă că valorile sunt aproape de media lor. Deviația standard poate fi utilizată pentru a evalua caracteristicile unui set de date, cum ar fi distribuția, simetria sau concentrarea valorilor. De exemplu, dacă deviația standard este mare, aceasta poate indica o distribuție neuniformă a valorilor. În resum, deviația standard este o măsură importantă a dispersiei valorilor unui set de date, care poate fi utilizată pentru a înțelege mai bine caracteristicile acestora.

Comentariul reprezinta rezultatul deviatiilor din output.



#10.938647141993062



#1836.1249574994952

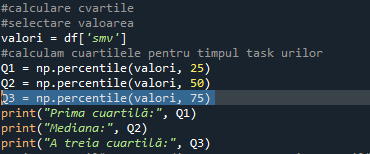


#11205250.31692982

**CVARTILE**

Cvartilele sunt valori utilizate pentru a descrie distribuția unui set de date, fiind un fel de medie mai avansată. Acestea impart setul de date in 4 parti(25% fiecare).

Am calculat cvartilele pentru timpul alocat task-urilor, respective pentru munca in progres implementand urmatoarele coduri:

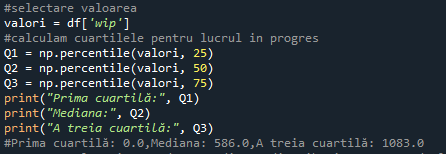


Observam ca:

\*Prima cuartilă: 3.94,Mediana: 15.26,A treia cuartilă: 24.26

\*Q3 este mult mai mare decat mediana, distributia are o coada lunga spre dreapta, înseamnă că majoritatea valorilor ce sunt timpul alocat pt task-uri sunt mai mari decât valoarea Q3. În alte cuvinte, această situație indică faptul că există o concentrație mai mare de valori ridicate în setul de date, în comparație cu valorile mai mici.În acest caz, putem să presupunem că există o valoare sau un număr mai mare de valori care se situează în partea superioară a intervalului, în comparație cu partea inferioară a intervalului.

\* 75% din date sunt mai mici decat 24.26, 25% mai mari decat acestea.



Observam ca:

\*Q3 este mult mai mare decat mediana, distributia are o coada lunga spre dreapta

\* Prima cuartilă:0.650307143(25% din date se afla sub valoarea q1), Mediana:0.773333333(50% din date se afla sub valoarea asta), A treia cuartilă: 0.850252525.

\* 75% din date sunt mai mici decat 0.850252525, 25% mai mari decat acestea.

**CORELATIE TOTALA**

\* Majoritatea corelațiilor dintre atribute sunt slabe și nesemnificative. Există doar câteva corelații pozitive moderate și puternice.

\* orele suplimentare și no\_of\_workers au o corelație pozitivă moderată, la fel ca și orele suplimentare și smv.

\* idle\_time și idle\_men au, de asemenea, o corelație pozitivă moderată, cel mai probabil pentru că sunt în mare parte 0-uri.

\*Gasim urmatoarele legaturi:- **lipsa legatura r<= 0.2**

#smv si no\_of\_style\_change 0.11

#wip actual\_productivity 0.13

#idle\_men si no\_of\_workers 0.05

- **legaturi directe,slabe,regasim intre 0.2-0.5**

#no\_of\_workers si over\_time 0.35

# over\_time si smv 0.26

#targeted\_productivity si incentive 0.49

**#legaturi puternice 0.7 si 0.95**

#actual-productivity si incentive 0.8

#targeted-productivity si actual\_productivity 0.7

**- In general, exista legaturi inverse lipsa intre parametrii nostri dpdv al corelatiilor**

#-0.06 intre idle\_men si targeted\_productivity

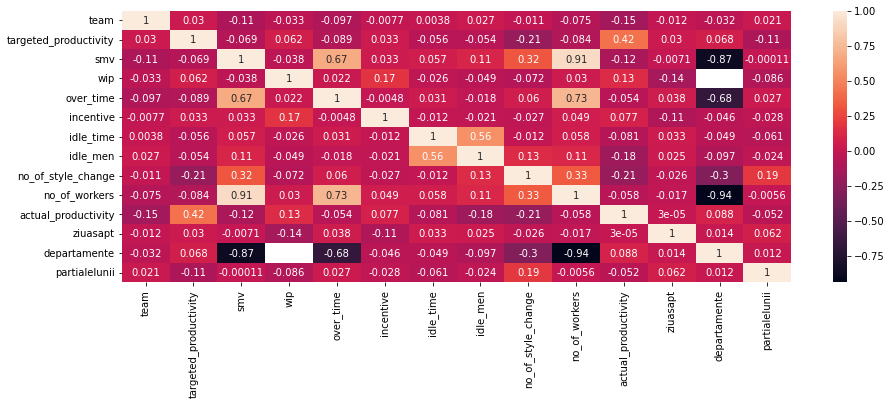
#-0.038 intre wip si smv

**#legaturi inverse ,slabe**

#smv si team -0.37

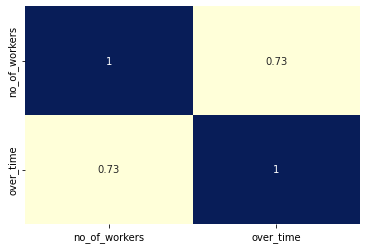
#no\_of\_style\_change si incentive -0.32



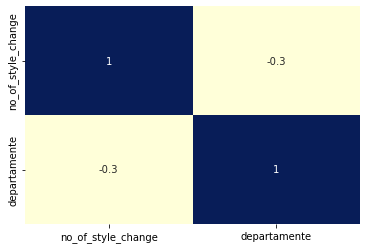


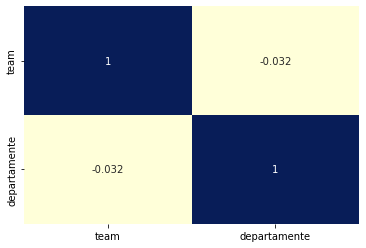
\*Observam ca o legatura stransa este intre intre numarul de lucratori si munca peste program (0.73).

\* Deci smv și no\_of\_workers au o corelație pozitivă puternică. Acest lucru are sens, deoarece ne-am aștepta ca, cu cât sarcina este mai lungă, cu atât mai mulți lucrători sunt alocați acesteia.



\*Observam ca o legatura inexistenta se afla intre departamente si team(-0.032), de asemenea, tot o legatura slaba se afla intre departamente si numarul de schimbari de stiluri(-0.3).





\* Majoritatea corelațiilor dintre atribute sunt slabe și nesemnificative. Există doar câteva corelații pozitive moderate și puternice.

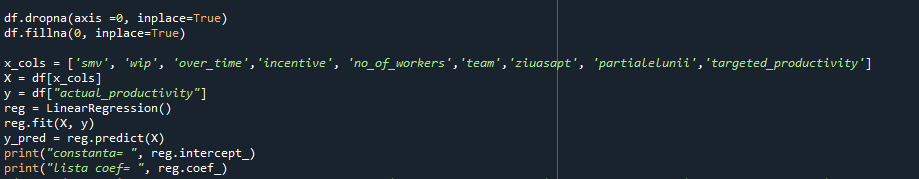
**REGRESII**

\*Acest cod ajustează un model de regresie liniară pentru a prezice productivitatea\_efectivă dintr-un set de variabile independente.

\*În primul rând, sunt tratate valorile lipsă din cadrul de date df. Rândurile care conțin valori lipsă sunt eliminate folosind df.dropna(axis=0, inplace=True), iar valorile lipsă rămase sunt completate cu 0 folosind df.fillna(0, inplace=True).

\*Variabilele independente, care sunt stocate în lista x\_cols, includ smv, wip, over\_time, incentive, no\_of\_workers, team, ziuasapt, partialelunii și targeted\_productivity. Aceste variabile sunt extrase din cadrul de date df și stocate în variabila X. Variabila dependentă productivitatea\_efectivă este stocată în variabila y.

\*Predicțiile modelului sunt apoi stocate în variabila y\_pred cu ajutorul metodei predict. Intersecția liniei de regresie (adică valoarea variabilei dependente atunci când toate variabilele independente sunt 0) este afișată cu ajutorul reg.intercept\_, iar coeficienții pentru fiecare variabilă independentă sunt afișați cu ajutorul reg.coef\_.



#constanta= 0.19675906117179487

#lista coef= [-3.13922032e-03 -5.00232255e-07 -2.01998747e-06 3.25520875e-03 9.85037953e-04 -3.04264366e-03 -1.70879370e-03 -4.06158086e-03 6.23650080e-01]

#index = 8 - prezicere = 0.74, val tabel = 0.75

#actprod= 0.19675906117179487+-3.13922032e-03\*19.87+-5.00232255e-07\*733+-2.01998747e-06\*6000+3.25520875e-03\*34+9.85037953e-04\*55+-3.04264366e-03\*2+(-1.70879370e-03)\*4+--4.06158086e-03\*1+6.23650080e-01\*0.75

#index = 26 -prezicere = 0.81, val tabel = 0.80

#actprod= 0.19675906117179487+-3.13922032e-03\*26.16+-5.00232255e-07\*1261+-2.01998747e-06\*7080+3.25520875e-03\*50+9.85037953e-04\*59+-3.04264366e-03\*2+(-1.70879370e-03)\*4+--4.06158086e-03\*1+6.23650080e-01\*0.8

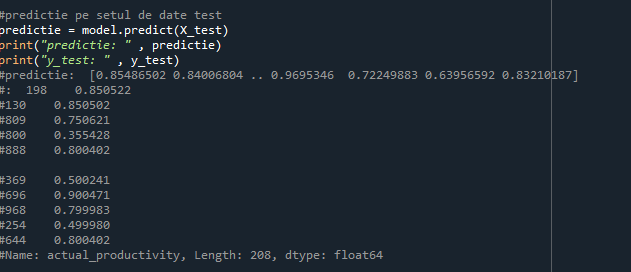
Constanta este interceptarea dreptei de regresie. Din dreapta de regresie intersecția este 0.19675906117179487. În regresie omitem unele variabile independente care nu au mare impact asupra variabilei dependente, interceptul indică valoarea medie a acestor variabile omise.

Dreapta de regresie este alcatuita si din coeficienti

Astfel, ea devine:

#dreaptaderegresie= 0.19675906117179487+-3.13922032e-03\*19.87+-5.00232255e-07\*733+-2.01998747e-06\*6000+3.25520875e-03\*34+9.85037953e-04\*55+-3.04264366e-03\*2+(-1.70879370e-03)\*4+--4.06158086e-03\*1+6.23650080e-01\*0.75

Predictie pe setul de date



Acesta este rezultatul predicțiilor făcute de modelul de învățare automată pe setul de date de test (X\_test) și valorile reale ale variabilei dependente (y\_test).

Rezultatul arată că predicțiile sunt stocate în variabila predictie, care este o matrice de 208 valori float. Valorile reale ale variabilei dependente sunt stocate în variabila y\_test, care este o serie de 208 valori float.

Valorile variabilelor predictie și y\_test pot fi utilizate pentru a evalua performanța modelului. De exemplu, am calculat eroarea medie pătratică (RMSE) pentru a obține o măsură cantitativă a acurateței modelului.

\*Eroarea pătratică medie (RMSE) măsoară diferența dintre valorile reale și valorile prezise. Cu cât valoarea RMSE este mai mică, cu atât modelul se potrivește mai bine datelor.

RMSE se calculează prin luarea rădăcinii pătrate a mediei pătratelor diferențelor dintre valorile reale și cele prezise. Diferența dintre fiecare valoare reală și cea prezisă se ridică la pătrat pentru a se asigura că toate diferențele sunt pozitive, iar apoi se ia media diferențelor la pătrat. În cele din urmă, se ia rădăcina pătrată a mediei pentru a obține valoarea RMSE în unitățile originale ale datelor.

O valoare RMSE mai mică indică faptul că modelul este mai corect în predicțiile sale.



In acest caz, valoarea RMSE de 0.09012262602437218 este relativ mică, ceea ce sugerează că modelul face predicții relativ corecte

\*Pentru R^2, prima data am impartit setul de date in set de antrenament si set de testare



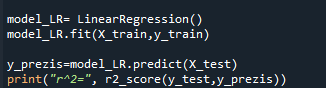
30% pentru testare, 70% pentru antrenare.

R2 reprezintă procentul de variație a variabilei care poate fi explicat prin caracteristici.Am facut o prezicere.



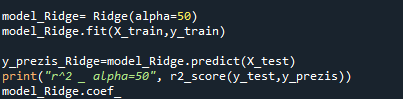
Modelul Linear Regression

\*Indicatorul ne arata ca nr de minute alocate pt un task, produsele neterminate ,timpul peste program a fiecarei echipe, valoarea stimulentului financiar care motiveaza o actiune ,nr asociat fiecarei echipe si comparatia cu ziua saptamanii influenteaza productivitatea actuala in proprotie de 78.5%,ceilalti factori influentand in proportie de 21.5%



coeficientul de determinare (R^2) este o măsură a modului în care modelul se potrivește datelor. O valoare de 1 indică o potrivire perfectă, iar o valoare de 0 indică faptul că modelul nu este mai bun decât o linie orizontală la media valorilor țintă. Funcția r2\_score ia ca intrări valorile țintă reale (y\_test) și valorile prezise (y\_prezis).

Modelul Ridge



Aceasta înseamnă că modelul Ridge cu alpha=50 este capabil să prezică variabila dependentă cu o acuratețe de 78,51%.

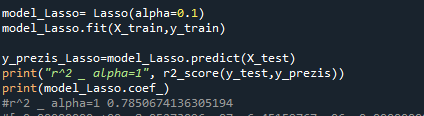
R pătrat, cunoscut și sub numele de coeficient de determinare, este o măsură statistică care reprezintă proporția de variație a variabilei dependente care poate fi prezisă de variabilele independente. În contextul regresiei ridge, R pătrat poate fi utilizat pentru a evalua gradul de adaptare a modelului.

În acest caz., valoarea lui R pătrat este 0,7850674136305194, ceea ce înseamnă că 78,5% din variația variabilei dependente poate fi explicată de variabilele independente din modelul de regresie. Aceasta este o valoare relativ mare, ceea ce sugerează că modelul se potrivește bine datelor.

Alpha este un parametru în regresia ridge care controlează puterea termenului de penalizare. O valoare mai mare a lui alfa înseamnă o contracție mai mare și un termen de penalizare mai puternic, în timp ce o valoare mai mică a lui alfa înseamnă o contracție mai mică și un termen de penalizare mai slab.

In cazul nostru valoarea lui alfa este 50, ceea ce înseamnă că termenul de penalizare este relativ puternic, iar coeficienții variabilelor independente sunt micșorați spre zero.

Pe model Lasso avem urmatoarele date:

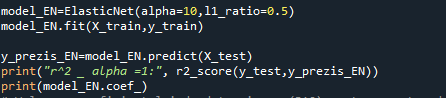


\*Scorul r2\_score variază de la -1 la 1, iar o valoare apropiată de 1 indică o potrivire bună, în timp ce o valoare apropiată de -1 indică o potrivire slabă. Coeficienții reprezintă ponderea sau importanța acordată fiecărei caracteristici din date atunci când se fac predicții. Un coeficient zero înseamnă că o caracteristică nu are niciun impact asupra variabilei țintă.

\* Acest lucru are ca efect micșorarea coeficienților caracteristicilor mai puțin importante la exact zero. Astfel, regresia Lasso realizează atât selecția variabilelor, cât și micșorarea pentru a spori precizia predicției.

Pe modelul ElasticNet, r patrat ne-a dat:





\*Valoarea coeficientului de determinare (R^2) pentru acest model este de 0,02117909686508379, ceea ce este relativ scăzut. Acest lucru sugerează că modelul explică doar 2,11% din variabilitatea variabilei țintă. Coeficienții modelului sunt, de asemenea, în mare parte 0, doar două caracteristici având coeficienți care nu sunt zero.

\*Multicoliniaritate

Acest cod calculează factorul de inflație a varianței (VIF) pentru un set de variabile independente. VIF este o măsură a gradului de creștere a varianței unui coeficient de regresie estimat din cauza prezenței altor variabile în model.

#Datele sunt mai întâi extrase din cadrul de date df și stocate într-un nou cadru de date dfVIF care include variabilele smv, wip, over\_time, incentive și no\_of\_workers.

# Variabila dependentă actual\_productivity și variabilele independente sunt specificate cu ajutorul unui șir de formule, iar datele rezultate sunt stocate în variabilele y și X.

#Se creează un nou cadru de date vif\_df pentru a stoca rezultatele calculului VIF. Funcția variance\_inflation\_factor din modulul statsmodels.stats.outliers\_influence este utilizată pentru a calcula VIF pentru fiecare variabilă din model.

#În cele din urmă, rezultatele sunt tipărite sub forma unui cadru de date în care variabilele sunt afișate în prima coloană, iar VIF-urile corespunzătoare acestora în a doua coloană.

#Multicoliniaritatea apare atunci când două sau mai multe variabile predictoare sunt foarte corelate între ele, astfel încât să nu furnizeze informații unice sau independente în modelul de regresie.

VIF măsoară puterea corelației dintre variabilele predictoare dintr-un model. Acesta capătă o valoare între 1 și infinitul pozitiv.

Dupa prelucrarea datelor, am obtinut datelede mai jos.

Intrucat toate valorile sunt foarte apropiate de 1, multicoliniaritatea nu este o problema.

