Testele statistice și rolul lor în proiect

Testele statistice presupun procesul de verificare a unei ipoteze statistice! Pe baza datelor colectate și pe baza testelor rulate, pot fi analizate două ipoteze:

- ipoteza nulă: presupunem că nu există diferențe semnificative și/sau eventualele diferențe au un caracter pur întâmplător;
- ipoteza alternativă: neagă ipoteza nulă și presupune că există diferențe semnificative.

De asemenea, este important să precizăm faptul că testele statistice sunt de două feluri:

- parametrice: datele sunt distribuite normal și/sau variațiile "populațiilor" sunt cunoscute;
- neparametrice: datele nu sunt distribuite normal și/sau variațiile "populațiilor" nu sunt cunoscute.

Înainte de a explica prima parte a lucrării științifice, este esențial să prezentăm și cele două tipuri de teste care reprezintă baza testelor statistice: T test și Z test.

Testul T se folosește în cazul eșantioanelor mici și deviația standard a eșantionului este cunoscută. Pentru a efectua un test t, se calculează statisticile t prin împărțirea diferenței dintre media eșantionului și valoarea nulă (valoarea de referință) cu deviația standard estimată a mediei eșantionului.

Testul z, pe de altă parte, este folosit atunci când dimensiunea eșantionului este mare și deviația standard a populației este cunoscută sau poate fi estimată. Statistica z este calculată prin împărțirea diferenței dintre două medii ale eșantioanelor cu eroarea standard a diferenței. Ambele teste generează o statistică de testare și un p-value, care indică probabilitatea ca diferența observată dintre medii să fie pur întâmplătoare. Dacă valoarea p este mai mică decât nivelul de semnificație ales, se poate respinge ipoteza nulă și se acceptă ipoteza alternativă că diferența dintre medii este semnificativă.

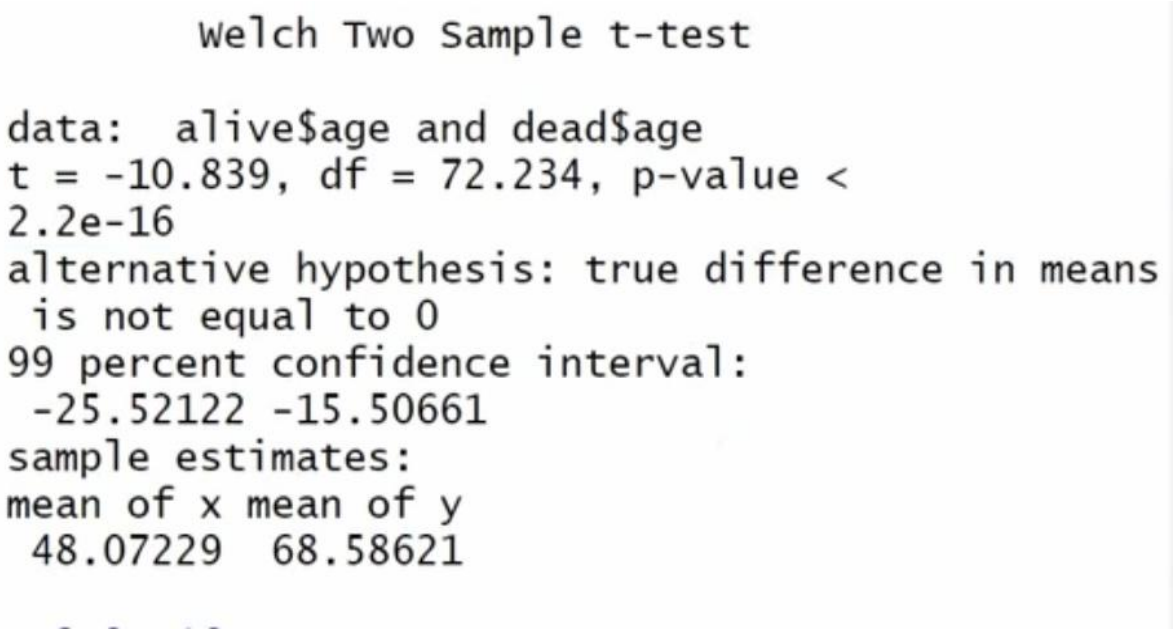
În cadrul lucrării noastre, am rulat testul Z în limbajul de programare R. Pentru acest obiectiv am folosit funcția z.test() integrată prin pachetul stats.

Rezultatul testelor statistice

În cazul primului test, ne-am propus să testăm una dintre cele mai întâlnite ipoteze din perioada pandemiei. Aceasta presupune că vârstnicii au un risc crescut de deces în cazul îmbolnăvirii cu virusul Covid-19. Cazurile analizate cuprind anii 2020 și 2021.
H0: “Vârstnicii au un risc crescut de deces în cazul îmbolnăvirii cu virusul Covid-19”

```r
# Varsta
# H0: oamenii care au murit sunt mai batrani
dead = subset(data, death_dummy == 1)
alive = subset(data, death_dummy == 0)
mean(dead$age, na.rm = TRUE)
mean(alive$age, na.rm = TRUE)
# Statistic significat?
z.test(alive$age, dead$age, alternative="two.sided", conf.level = 0.99)
# diferenta dintre o persoana care moare si una care traieste este in intervalul -25.52122 si -15.50661
# p-value is 2.2e-16 ~ 0 < 0.05, deci oamenii care au murit de Covid 19 sunt mai batrani
```

Output-ul codului R:



```
        Welch Two Sample t-test

data:  alive$age and dead$age
t = -10.839, df = 72.234, p-value <
2.2e-16
alternative hypothesis: true difference in means
 is not equal to 0
99 percent confidence interval:
 -25.52122 -15.50661
sample estimates:
mean of x mean of y
 48.07229  68.58621
```

Ce înseamnă aceste rezultate? În medie, diferenta dintre o persoana care se vindeca și o persoana care moare în urma infectarii cu virusul Covid-19 este în intervalul [-25,5; -15,5]. Ce înseamnă acest lucru? În medie, persoanele care mor în urma contactului cu virusul sunt cu cu 15 pana la 25 de ani mai în varsta decat cele care s-au vindecat. Probabilitatea acestui eveniment a fost considerat 0.99, adică 99%. Deşi intervalul este unul de dimensiune de 10 ani, una semnificativ de mare, capetele intervalului demonstrează/confirma aceeaşi teorie. P-value este una care tinde la 0 (2.23 - 16). Acest fapt arată ca rezultatul nostru este statistic semnificativ.

În cadrul celui de al doilea test, am considerat H0: “Genul nu afectează mortalitatea”.

```r
# Gen
# H0: genul nu are efect
men = subset(data, gender == "male")
women = subset(data, gender == "female")
mean(men$death_dummy, na.rm = TRUE) #8.5%!
mean(women$death_dummy, na.rm = TRUE) #3.7%
# Statistic significat?
z.test(men$death_dummy, women$death_dummy, alternative="two.sided", conf.level = 0.99)
# Barbatii au cu 0,8% pana la 8.8% sanse mai mari de mortalitate in cazul infectarii cu Covid19
# p-value = 0.002 < 0.05, semnificativ statistic
```

Rezultatul acestui test oferă următoarele date:

- mortalitatea barbatilor: 8,5%
- mortalitatea femeilor: 3,7%

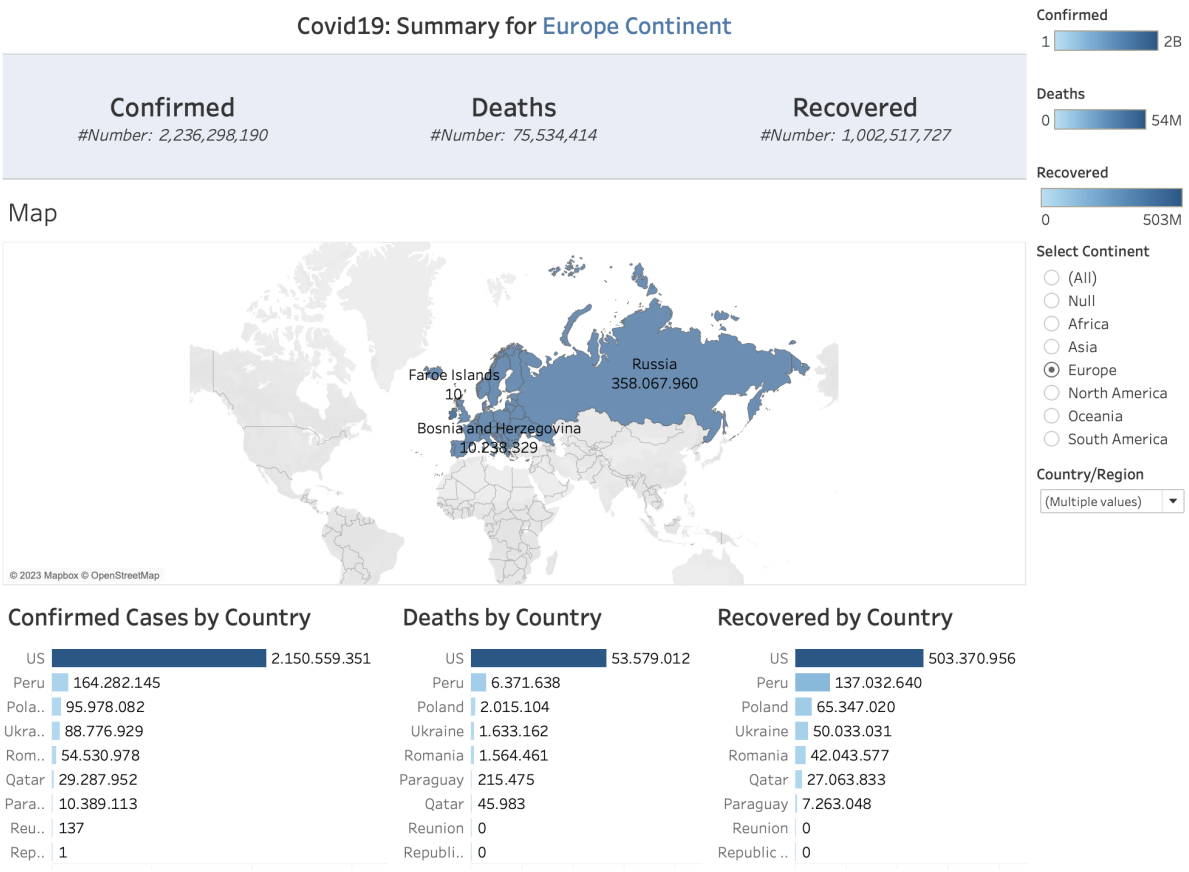
Rulând testul Z, înțelegem că riscul bărbaților de a muri în urma infectării este cu 0,8% până la 8,8%. Din cauza probabilității crescute (de 99%) afectează rezultatul din punct de vedere al dimensiunii de 8 unități de măsură. Totuși, p-value este de 0,002 care arată faptul că testul este semnificativ statistic. Așadar, este confirmată H0: "Genul nu afectează mortalitatea".

Tableau & Data Analytics

Domeniul Data Analytics se referă la procesul de analiză a datelor pentru a descoperi modele, tendințe și informații utile în vederea luării deciziilor. Acesta se bazează pe tehnologii avansate de calcul, precum inteligența artificială și analiza datelor în timp real, și poate fi aplicat în diferite domenii, cum ar fi afaceri, sănătate, sport sau finanțe.

Una dintre cele mai puternice unelte pentru data analytics este Tableau, o platformă de analiză și vizualizare a datelor care oferă o gamă largă de funcții pentru a extrage informații relevante din date și a le prezentă sub formă de grafice, tabele și rapoarte interactive. Rolul acestei tehnologii a jucat un rol important pe parcursul pandemiei de Covid-19 pentru analiza evolutiei acestei pandemii.

În cele ce urmează, atasez un dashboard creat cu ajutorul Tableau. Obiectivul acestuia este de a sintetiza un dataset amplu în care sunt raportate zilnic, în funcție de locație, toate cazurile noi, decesele și numărul persoanelor vindecate. Prin intermediul hărții interactive, un user poate vedea numărul de cazuri de pe un anumit continent sau chiar țara. Rolul acesteia este de a fi o modalitate de input folosind un simplu click pentru o prima analiza. Uitandu-ne la bar-chart-urile din josul paginii, putem vedea clasamentului țărilor care au avut cel mai mare număr de cazuri de Covid-19. Acest feature se bazează pe filtrele din dreapta paginii unde putem alege țările pe care dorim sa le urmarim.



Acest dashboard oferă o imagine de ansamblu asupra situației globale. O astfel de aplicație poate fi folosită pentru luarea deciziilor importante, precum restricțiile de călătorie sau pentru a identifica trenduri sau modele de raspandire.

Tableau este o unealtă de analiza a datelor foarte importanta, mai ales datorita capacității sale de a rula query-uri pe baza de date mari (Big Data). Aceasta utilitate se poate extinde in domenii precum IoT, unde cantitatea de date este una semnificativ mai mare.

Ce probleme am intampinat?

Principala problema pe care am descoperit-o în cazul cercetării noastre este cea legată de data set. Acesta cuprindea erori și date redundante, fapt care ne-a afecta precizia datelor expuse. Din acest motiv, procesul de data cleaning este foarte important, dar efectul acestuia este redus în cazul unui set de date alcătuit greșit (din punct de vedere atât al designului bazei de date, cat si a acurateței informației).

JupyterNotebook, Python & Machine Learning

JupyterNotebook este o platforma foarte des intalnita in domeniul data science și AI (cu subdomeniile conexe).

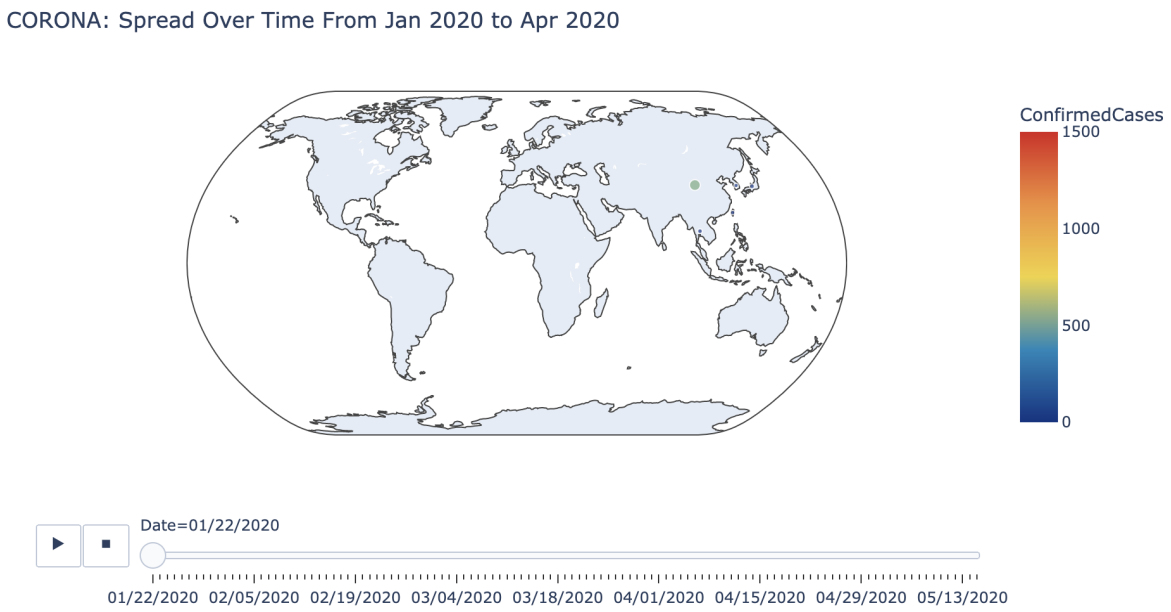
Va propun sa analizam o secventa din cadrul cercetării noastre:

```
formated_gdf = train.groupby(['Date', 'Country_Region'])['ConfirmedCases'].sum()
formated_gdf = formated_gdf.reset_index()
formated_gdf['Date'] = pd.to_datetime(formated_gdf['Date'])
formated_gdf['Date'] = formated_gdf['Date'].dt.strftime('%m/%d/%Y')
formated_gdf['size'] = formated_gdf['ConfirmedCases'].pow(0.3)

fig = px.scatter_geo(formated_gdf, locations="Country_Region", locationmode='country names',
                     color="ConfirmedCases", size='size', hover_name="Country_Region",
                     range_color= [0, 1500],
                     projection="natural earth", animation_frame="Date",
                     title='CORONA: Spread Over Time From Jan 2020 to Apr 2020', color_continuous_scale="portland")
fig.show()
```

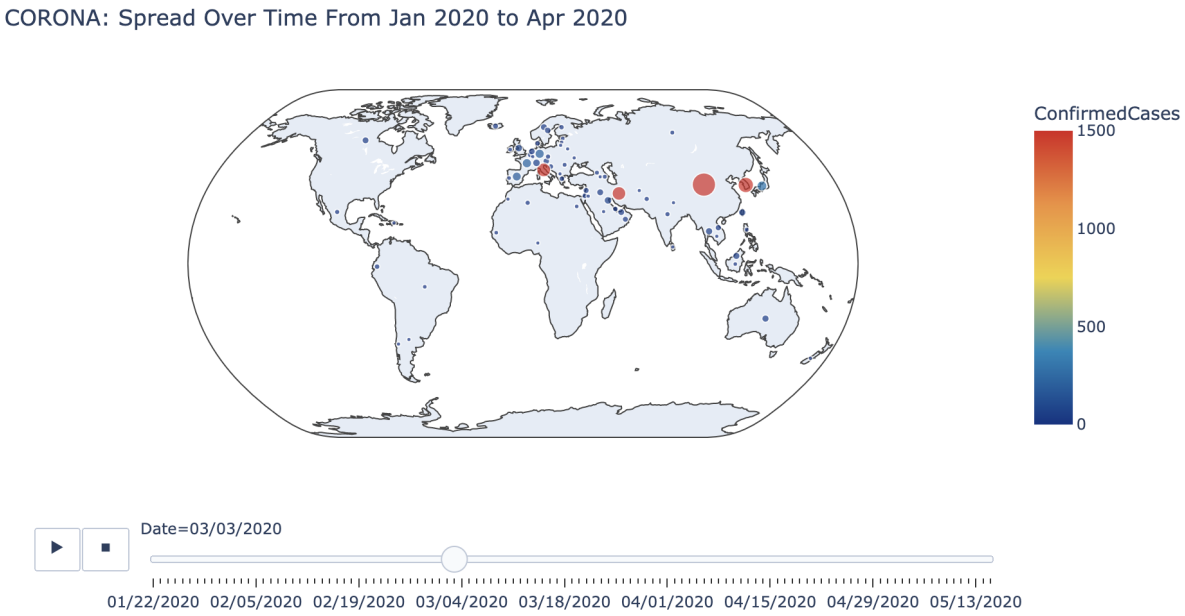
Obiectivul nostru este de a crea o harta globală prin care sa fie evidentiata împrăștierea virusului Covid-19 pe glob. Pentru acest lucru, am folosit libraria plotly.express. Data evenimentelor reprezinta axa de evolutie, iar numarul de cazuri confirmate este reprezentat prin cercuri de diferite culori pe harta.

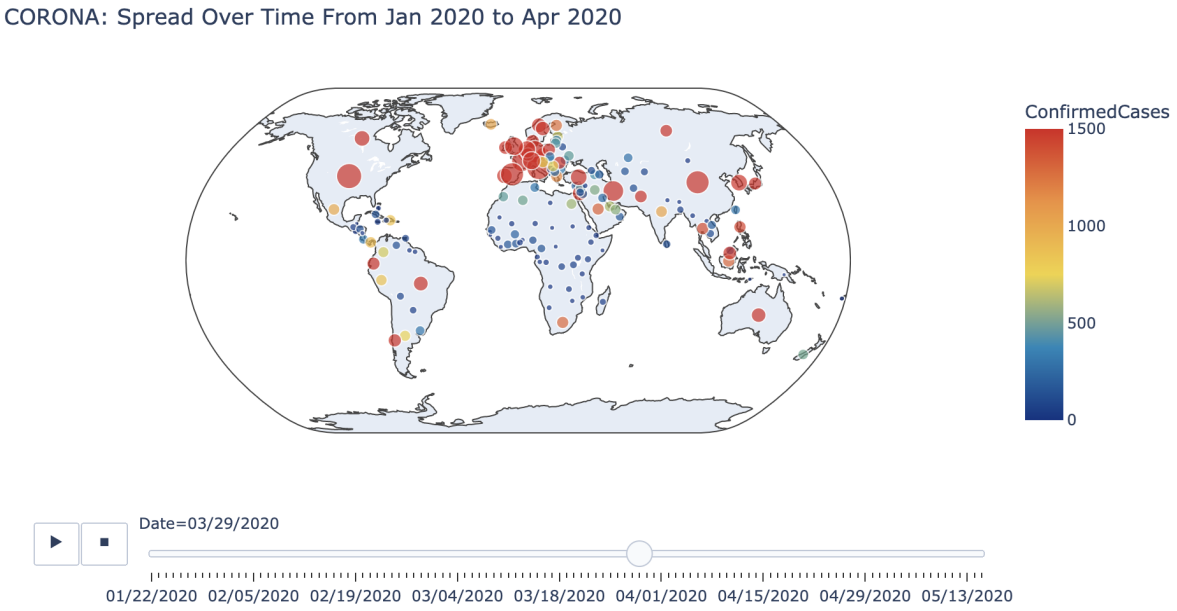
Prima forma a hartii la momentul T0 (data de 01/22/2020) este aceasta:



Ne propunem sa identificăm un trend de raspandire. Dar, ce este un trend din punct de vedere statistic? Trendul reprezinta o directie generala de evolutie a unui eveniment/fenomen, obicei. Uitandu-ne la contextul actual, un trend este o evolutie a variabilei noastre principale (numărul de cazuri confirmate). Intensitatea de creștere este un element crucial în analiza sau descoperirea unui trend. Pe harta noastra, aceasta este prezentata prin cercuri în dreptul țărilor de diferite culori si marimi, conform legendei din dreapta. Utilizăm analiza trendurilor, pentru

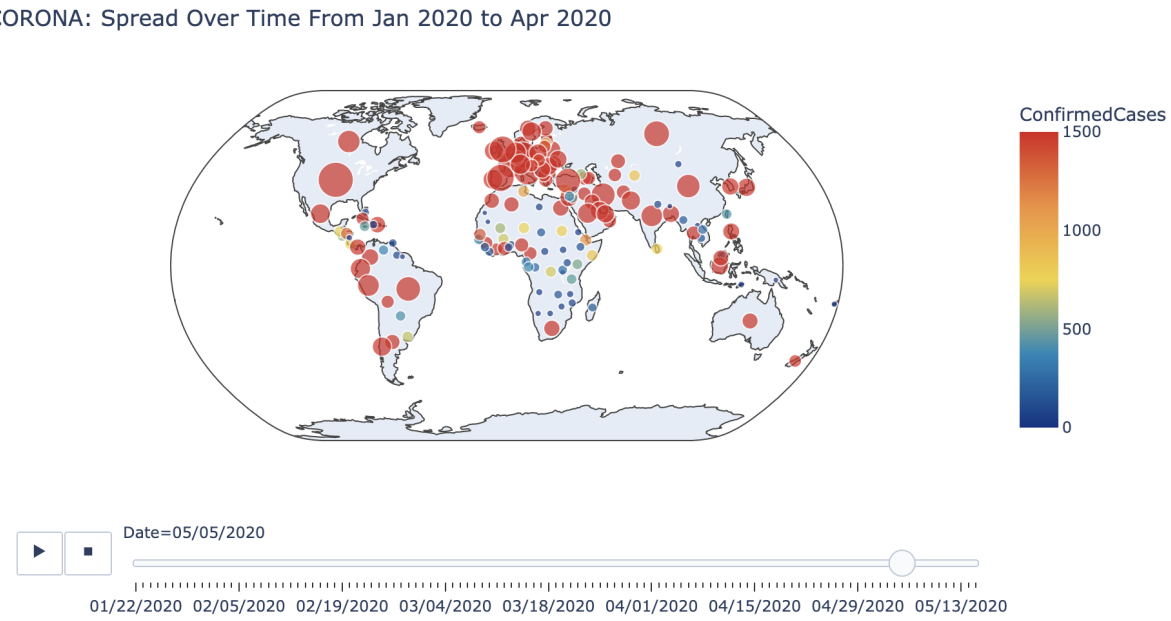
a observa modele și comportamente care pot sa se repete intr-o anumita perioada de timp. Scopul acestora este de a planifica deciziile în funcție de evolutia unui eveniment.





Din aceste doua poze, observam evolutia exponențială a numărului de cazuri. Trendul este unul clar cand ne uitam la aceste două poze. Dar dacă ne uităm la prima și la a doua poza? Observam faptul ca apar noi tari pe traiectoria E - V (Europa) unde numărul de cazuri este foarte ridicat într-o perioada scurta de timp. Așadar, observăm un potențial trend.

Cand analizam si cea de a treia data, trendul este deja confirmat. Raspandirea pe traiectoria E - V a cuprins toata harta. Toate continentele au cel puțin o țară cu intensitate mare.



Aceasta ultima poza arată clar efectul pandemiei. Virusul a ajuns în toate țările. Trendul confirmat este de extindere din E spre V, iar cele mai multe cazuri sunt în zonele aglomerate. Europa este o zona geografica foarte afectata, mai ales în zona de Veste.

Concluzie si o intrebarea de analizat in continuare

Viteza de imprastiere a pandemiei este strans legată de viteza si posibilitatea de deplasare a oamenilor în jurul lumii?

Machine Learning & Predicția numărului de cazuri

Machine Learning reprezinta un domeniu a inteligentei artificiale care permite calculatoarelor sa invete si sa se adapteze automat asupra datelor fără a fi programate explicit. În cadrul lucrării noastre, am utilizat tehnologia Machine Learning pentru a rula predicții asupra evoluției numărului de infectii. Pe baza analizei lunilor februarie și martie 2020, ne-am propus sa vedem cum putem anticipa numărul de cazuri pentru primele 5 zile din luna aprilie.

Pentru acest obiectiv, am ales modelul SARIMA (Seasonal Autoregressive Integrated Moving Average). Acesta utilizează informațiile din seriile temporale precum tendința, sezonul și intensitatea. Tendinta reflecta o crestere sau scadere generala a valorii (numarul cazurilor confirmate) in cadrul seriei temporale. Sezonul, un element important în cadrul analizei noastre, reflecta variații aleatoare dintr-un cadru temporal.

Un element care ar trebui sa fie luat în considerare în cazul unei analize actuale ar trebui sa fie rata de vaccinare. De asemenea, existenta unei raportări reale este foarte importantă.

5 Days Cases Prediction

Spain - Prediction	Spain - Reality	Deta	Error
188699	170537	18162	-10,64988829
189501	174621	14880	-8,52131187
190303	179143	11160	-6,229660104
191104	185870	5234	-2,815946629
191906	191444	462	-0,241323834

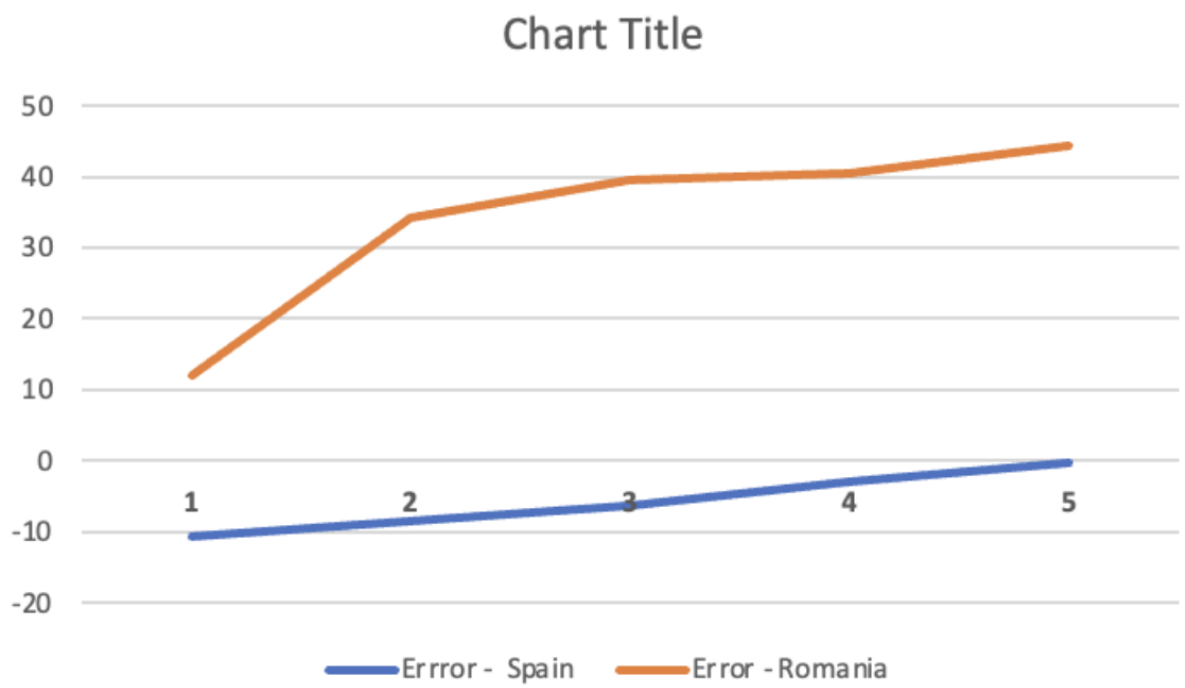
Ro - Prediction	Ro - Reality	Delta	Error
2076	2360	-284	12,03389831
2097	3183	-1086	34,11875589
2121	3502	-1381	39,43460879
2148	3613	-1465	40,54802104
2230	4010	-1780	44,38902743

În aceasta imagine, vedem rezultatele predicției in cazul a doua tari din Europa: România și Spania. Este foarte interesant faptul ca în cazul primei simulari (data: 01.04), valoarea erorii în modul este foarte apropiata. De asemenea, trendul este unul identificat corect. Creșterea numărului de cazuri a fost anticipată. Intensitatea este elementul greșit estimat în cazul României. Deși am început cu o eroare în modul de aproximativ 12%, în cea de a cincea zi, eroarea în modul ajunge la 44%. În cazul Spaniei, începem de la valoarea de 10% și ajungem până la 0,2%.

Concluzia 1

Transparența si onestitatea sunt elemente foarte importante. Raportarea greșită a numărului de cazuri de la începutul pandemiei este un factor decisiv. Spania, unde Guvernul a pus la punct un sistem de raportare foarte bine organizat, oferă o credibilitate mult mai mare a predicției. Exemplul Chinei (unde eroarea este constant de 98%) este un alt exemplu relevant pentru concluzia noastra.

Graficul atașat arată evoluția erorii de predicție (fără a fi considerată valoarea in modul). În cazul României, creșterea exponențială vine după prima zi de estimare. De ce? Din cauza datelor insuficiente și greșite. Eroare initiala este considerata normala, deoarece toate simularile relevante au avut o eroare inițială (în modul) în intervalul [9,43; 15,1].



Regresia si efectul vaccinului

Regresia este o metoda statistică care este utilizata si in domeniul data analytics pentru a evolua relația dintre două sau mai multe variabile.

Din punct de vedere medical, vaccinul începe sa isi faca efectul in aproximativ doua saptamani după administrare. Din acest motiv, am pregatit doua teste.

Primul caz analizat presupune identificarea corelației dintre numărul de persoane vaccinate lunar în perioada Mai - Iulie 2021 cu numărul de cazuri confirmate în perioada Iulie - Septembrie 2021. Aceasta analiza, care se bazează pe un algoritm de regresie, arată o corelație puternică, negativă (r = 0.6123). De asemenea, studiul scoate în evidenta ca o modificare cu o unitate în randul persoanelor vaccinate duce la scăderea cu o valoare de 1,11 în randul persoanelor confirmate.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,61233071
R Square	0,3749489
Adjusted R S	0,06242335
Standard Err	334734,123
Observations	4

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1,3443E+11	1,3443E+11	1,1997384	0,38766929
Residual	2	2,2409E+11	1,1205E+11		
Total	3	3,5852E+11			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	633524,324	222605,326	2,84595312	0,10447256	-324269,09	1591317,74	-324269,09	1591317,74
X Variable 1	-1,1198923	1,0224286	-1,0953257	0,38766929	-5,5190475	3,27926288	-5,5190475	3,27926288

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	630704,435	352482,565
2	614635,1	-226124,1
3	471871,226	-205893,23
4	173843,24	79534,7604

Cel de-al doilea caz analizat presupune aceleași perioade de timp, dar analizăm corelatia dintre numărul de vaccinuri și numărul de decese datorate Covid-19. De aceasta data, corelația este de 46,9% (r = 0.469), adică moderata. Impactul a unei unități în cazul numărul persoanelor vaccinate duce la scăderea cu 33,46 a numărului de decese.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,46900799
R Square	0,2199685
Adjusted R S	-0,1700473
Standard Err	373936,877
Observations	4

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7,8863E+10	7,8863E+10	0,563999	0,53099201
Residual	2	2,7966E+11	1,3983E+11		
Total	3	3,5852E+11			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	597826,2	250377,608	2,38769835	0,13959489	-479461,7	1675114,1	-479461,7	1675114,1
X Variable 1	-33,461592	44,5561269	-0,7509987	0,53099201	-225,17113	158,247949	-225,17113	158,247949

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	558910,369	424276,631
2	587620,415	-199109,41
3	510190,291	-244212,29
4	234332,926	19045,074