Introducere

In cadrul studiului realizat, s-a urmarit extragerea de informatii relevante pentru persoanele/companiile care realizeaza online advertising.

Marile engine-uri de advertising (Facebook ads, Google ads, TikTok ads) permit persoanelor care administreaza campaniile de advertising sa selecteze anumite audiente care sa fie tintite in cadrul acestora, in functie de anumiti factori, cum ar fi: localizare, varsta, interese.

Campaniile platite de advertising, de cele mai multe ori se bazeaza pe sistemul PPC (pay per click) sau pe unul similar. Profitabilitatea acestor campanii se poate masura cu indicatorul universal ROAS(Return On Ad Spend), care se calculeaza raportand veniturile generate din advertising la costurile campaniilor aferente.

Se doreste ca aceste campanii sa dea un randament cat mai mare, prin urmare este necesar ca persoanele targetate pentru a vedea o anumita reclama sa fie cat mai predispusi in achizitionarea produsului/serviciului promovat, acest lucru fiind posibil prin selectarea audientelor.

Pentru acest studiu, am asumat ca mai multe industrii ar s-ar putea folosi de insight-uri legate de fericire/tristete, pentru a targeta in mod eficient persoane. De exemplu, o companie care comercializeaza suplimente alimentare pe baza de plante, care pot creste in mod natural nivelul de endorfine (chimicale produse natural in sistemul nervos pentru a combate stresul si implicit tristetea), poate beneficia direct de informatiile rezultate in urma cercetarii.

Asadar, ne propunem sa raspundem urmatoarelor intrebari:

- Care sunt cele mai favorabile zone pentru a targeta o astfel de audienta?
- Care sunt factorii cheie care pot influenta fericirea?
- Cum se poate specula fericirea pentru o audienta anume?

Setul de date

Pentru a putea analiza aceste aspecte, am gasit setul de date “World Happiness Report” de pe Kaggle. Datele sunt extrase Gallup World Poll, o serie globala de sondaje care care urmareste o serie de indici care reflecta bunastarea tarilor (cum ar fi accesul la hrana, rata somajului sau performanta conducerii). Aceste sondaje se desfasoara din 2005, in peste 160 de tari. Pe baza

acestor sondaje, s-au creat 6 indici pe baza carora se calculeaza scorul de fericire: indicele de productie economica, indicele de suport social, indicele de speranta la viata, indicele de libertate, indicele de absenta a coruptiei si indicele de generozitate. Prin insumarea acestor factori, se optine indicele de fericire. Intrebarile din sondajul Gallup pe baza carora s-au realizat acesti indici, au fost adresate sub forma unei scari Cantril. Participantii la sondaj au fost nevoiti sa se gandeasca la o scara cu 10 trepte, unde treapta 0 reprezinta cea mai neplacuta viata posibila pentru ei, in timp ce treapta 10 reprezinta cea mai buna viata posibila. Pe baza acestor raspunsuri ,s-a creat o tara fictiva numita Dystopia, care incorporeaza cei mai josi indici inregistrati pentru fiecare categorie selectata in cadrul tarilor evaluate. Modul de calculare al indicilor enumerati anterior, este raportul indicelui mediu al fiecarei tari (la categoria respectiva) la indicele Dystopiei -1, astfel el fiind intotdeauna mai mare de 0.

Acest set de date prezinta mult potential in ceea ce priveste obiectivul studiului nostru, intrucat pe inregistrarile acestuia se pot efectua diverse analize care sa rezulte in noi insight-uri pentru targetarea audientelor.

Setul de date este alcatuit din 5 fisiere CSV, acestea reprezentand datele inregistrate din anii 2015, 2016, 2017, 2018 si 2019. Acest lucru ne permite nu doar sa ne folosim de informatiile cele mai recente pentru a obtine rezultate, dar si sa demaram analize comparative utilizand datele din anii precedenti. Acest lucru ne permite sa observam evolutia si trend-urile.

Datele sunt atat de tip float cu 3 zecimale in cazul indicilor, cat si de tip string in cazul denumirilor proprii.

Primul pas este extragerea CSV-urilor , fiecare in cate un data frame. Astfel, vom avea 5 data frame-uri initiale: data_2015, data_2016, data_2017, data_2018, data_2019. Pe baza acestora urmeaza sa creem variabile auxiliare pe baza carora sa creem analize.

Observam ca nu toate datele au aceeasi denumire pentru coloane, dar totodata, in functie de an, difera si coloanele in sine. Astfel, ne vom raporta la cele mai recente inregistrari pentru a decide care indici sa fie luati in considerare. Aceste inregistrari se regasesc in dataframe-ul data_2019, intrucat acesta este cel mai relevant pentru obiectivele noastre.

Primul pas pentru aplatizarea datelor este eliminarea coloanelor pe care nu le folosim. Metodologia chestionarelor difera de la an la an, asadar este posibil si apara coloane in plus sau in minus. Nu ne dorim elemente care sa nu se poata asocia cu elementele aferente din celelalte tabele, asadar s-au identificat pentru stergere urmatoarele coloane:

- data_2015: Standard.Error, Dystopia.Residual

- data_2016: Lower.Confidence.Interval, Upper.Confidence.Interval, Dystopia.Residual
- data_2017: Whisker.high,Whisker.low, Dystopia.Residual

Acest lucru este realizat prin atribuirtribuirea la data frame-ul tinta a valorii unui subset al acestuia care nu include coloanele enumerate anterior.

Tot odata, denumirile coloanelor care reprezinta acelasi indice, difera de la an la an, prin urmare este nevoie sa stabilim o serie de denumiri standard:

- indicele de productie economica - GDP
- indicele de suport social - SocialSupport
- indicele de speranta la viata - LifeExpectancy
- indicele de libertate - Freedom
- indicele de absenta a coruptiei - Corruption
- indicele de generozitate - Generosity
- indicele de fericire - Happiness
- pozitia in clasamentul de fericire - Rank
- tara - Country

Acest lucru se realizeaza cu ajutorul functiei “rename” din cadrul librariei “plyr”

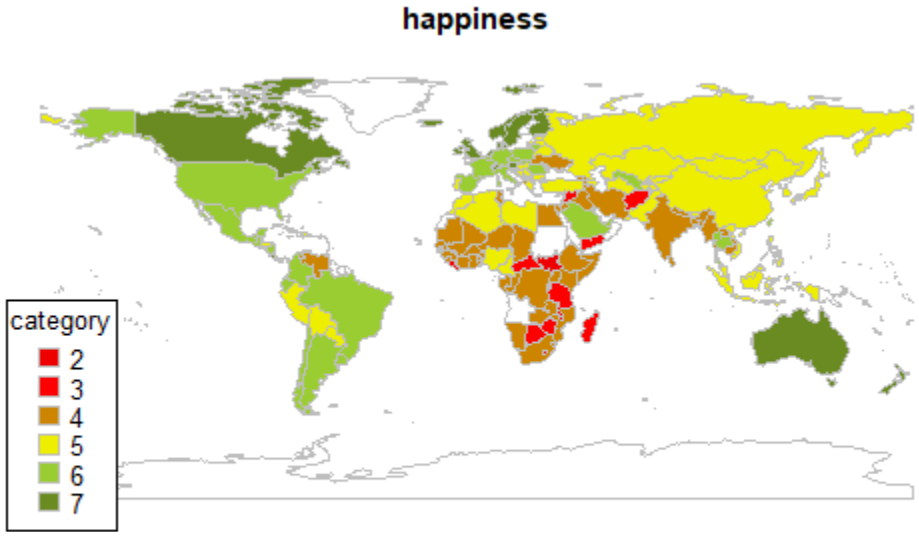
Observam dupa efectuarea modificarilor ca una din coloanele din data_2018 este de tip “char” in loc de “double”. Aceasta este o consecinta al unui simbol “N/A” intr-una din celule. Pentru a remedia aceasta inconvenienta, inlocuim valorile de tip string “N/A” cu valoarea nula, dupa care convertim coloana in double utilizand functia “as.double”.

Dorim sa introducem o coloana care va fi utila in analiza, si anume o coloana care sa indice regiunea din care tara face parte. Aceasta va purta denumirea “Region”. Pentru a executa acest lucru, vom folosi libraria countrycode. Aceasta va primi ca input coloana “Country” aferenta numelui tarii sub identificatorul “country.name” si va returna un sir de caractere reprezentand numele regiunii aferente.

Dupa procesul de curatare a datelor, datele arata in felul urmator:

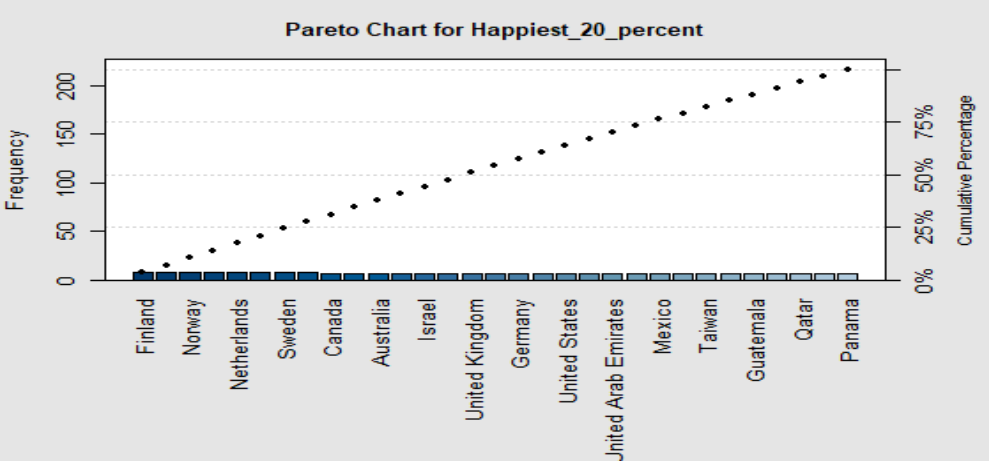
Rezultate si discutii

In primul rand, pentru a ne decide de unde sa incepem analiza, vom vizualiza datele intr-o harta a lumii, cu ajutorul librariei “rworldmap”. Pentru aceasta vom crea un dataframe de unde vom extrage codul tarii si indicele de fericire, pe care il vom asociahartii globale.



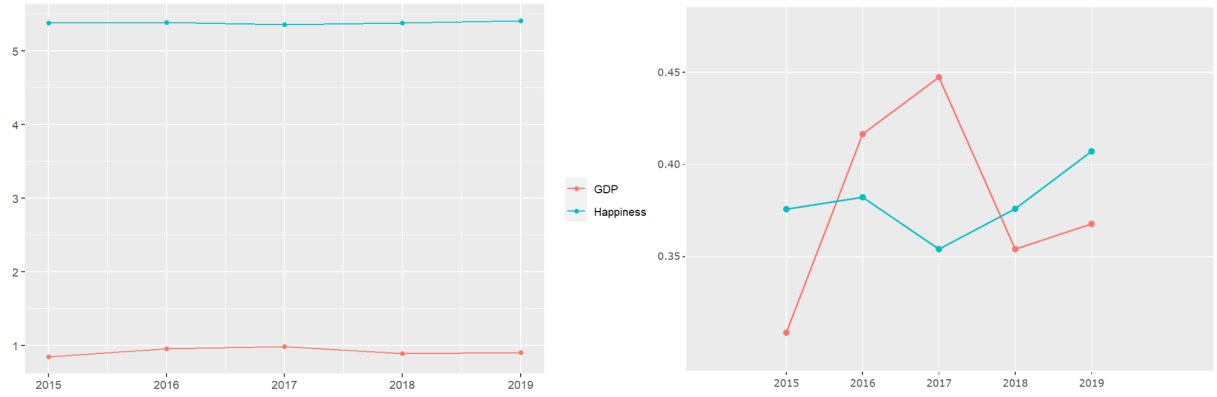
Ca o prima parte a analizei, dorim sa aflam care au fost cele mai fericite, respectiv cele mai triste tari in anul 2019. Ca o analogie la principiul lui Paretto, vom asuma ca 20% din tari acapareaza 80% din fericirea globala, respectiv 20% din tari acapareaza 80% din tristetea globala. In cazul advertising-ului online, dorim sa ne adresam acelor 20%. Asadar, calculam acest lucru raportat la

tarile evaluate in 2019 prin formula length(data_2019$rank)/100*20, iar, folosindu-ne de acest rezultat, returnam doua liste cu 31 cele mai triste respectiv cele mai fericite tari din anul 2019. Mai departe, dorim sa vedem daca acestea au diferente mari in procentaje, motiv care ar justifica eliminarea unora dintre ele, acestea contribuind prea putin la procentajul cumulativ. Pentru a verifica acest lucru, vom realiza o diagrama Pareto.



Dupa cum se poate observa in diagrama, procentajul cumulativ creste aproximativ constant, lucru care inseamna ca tarile din top 20% au ponderi aproximativ egale. Astfel, toate acestea pot constitui un grup de interes pentru advertiseri.

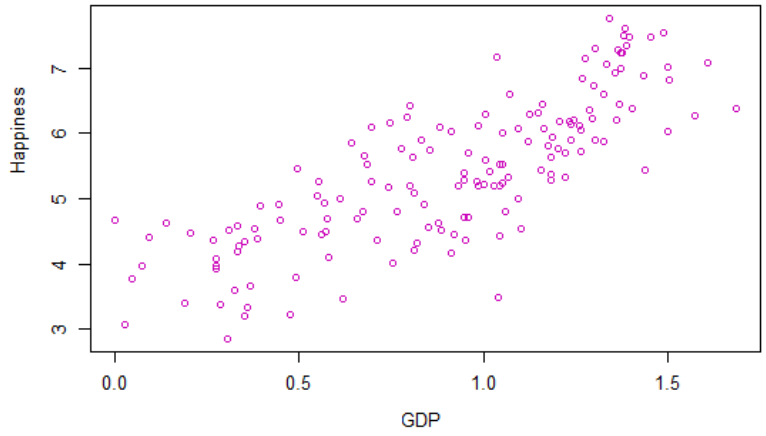
Una din premise ar fi ca produsul intern brut este unul din factorii principali care determina fericirea, asa ca dorim sa observam cum afecteaza evolutia acestuia in timp nivelurile de fericire. Pentru aceasta se va realiza un data frame care va cuprinde mediile celor doi indici in fiecare an inregistrat, dar si o coloana cu etichete pentru a diferentia categoriile. Acesti indici se vor urmarii pe un grafic cu linii.



Graficul generat (cel din partea stanga) indica o relatie liniara, variatiile fiind prea mici pentru a se putea influenta reciproc. Totusi, de dragul curiozitatii, am dorit sa apropiem aceste valori pentru a le suprapune. Pentru a realiza acest lucru , am adus valorile la aceeasi origine,

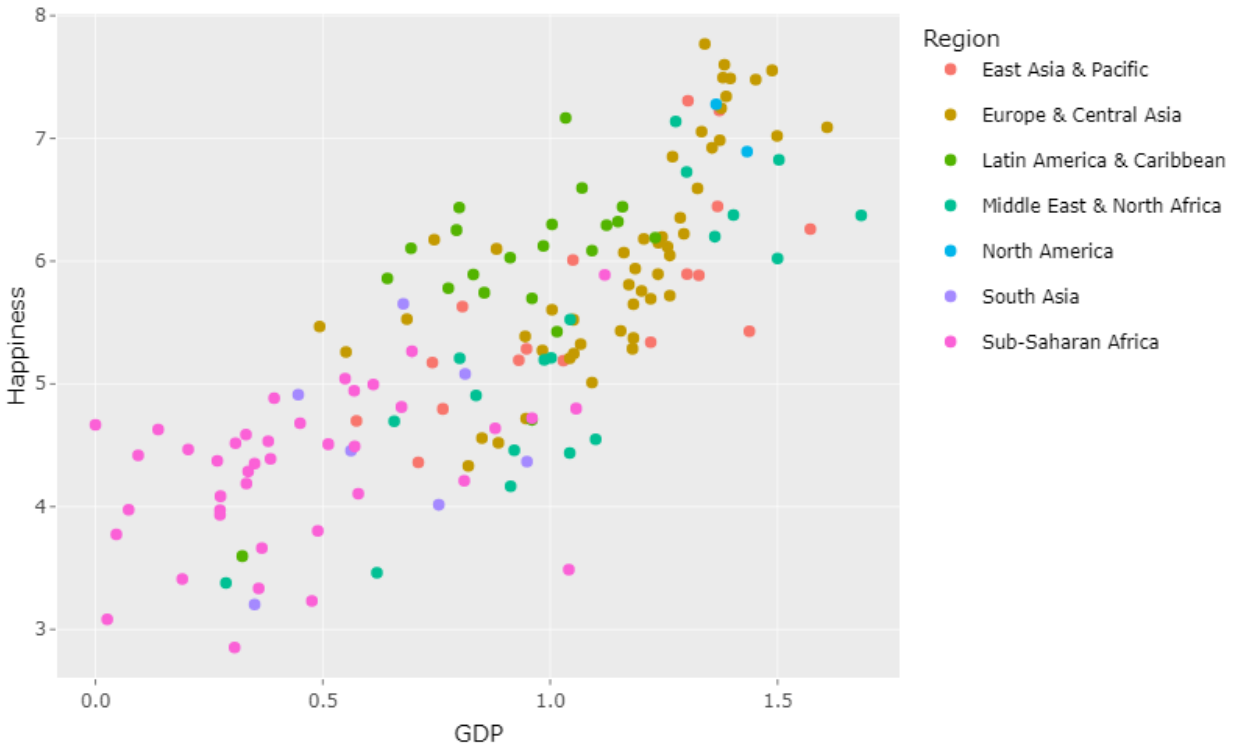
determinand valorile pentru aceasta. Astfel, scazand din media PIB-ului valoarea determinata 0.5374293, respectiv din indicele de fericire valuarea 5, s-a obtinut graficul din partea dreapta, unde se poate observa mai clar relatia dintre cele doua variabile.

In continuare, pentru a observa influenta PIB-ului in indicele de fericire, s-a realizar un grafic cu puncte pe doua axe, unde fiecare punct reprezinta o tara



Din grafic se poate observa ca produsul intern brut are o oarecare influenta asupra indicelui de fericire, in masura in care daca un anumit nivel economic nu este atins, tara respectiva nu poate trece la nivelul urmator pe scara fericirii (de exemplu, tarile care au indicele PIB mai mic de 0.5 nu depasesc indicele de fericire 5, in timp ce nici o tara care doar tarile care au indicele PIB mai mare de 1 depasesc indicele de fericire 7)

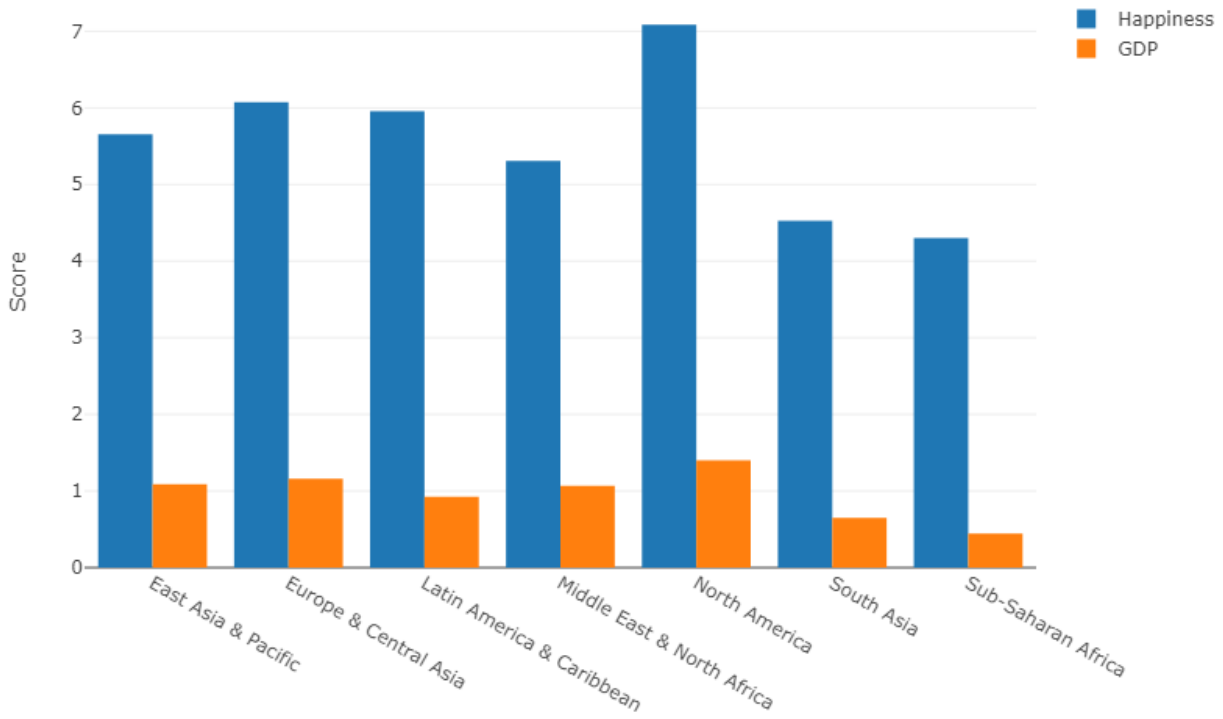
Din acest motiv, apare ipoteza unui cluster, iar daca ne uitam la harta generata la inceput, un posibil element pentru clustering este reprezentat de regiune. Vom genera graficul anterior, colorand punctele in functie de aceasta.



In figura alaturata se pot identifica anumite grupuri, de sus in jos. In partea de sus avem tari care fac parte din Europa si asia centrala, impreuna cu cele doua din nordul continentului American. In cadranul urmator, se pot diferentia alte doua grupuri, si anume in partea stanga tarile din America latinta, iar in partea dreapta tarile din Orientul Mijlociu si Africa de nord. Urmatoarea sectiune este populata de tari din Asia de est, Asia centrala, dar si Asia de sud acestea fiind amestecate. Ultimul grup care se poate identifica (avand tot odata si cea mai mare arie acoperita) este Africa sub-sahariana.

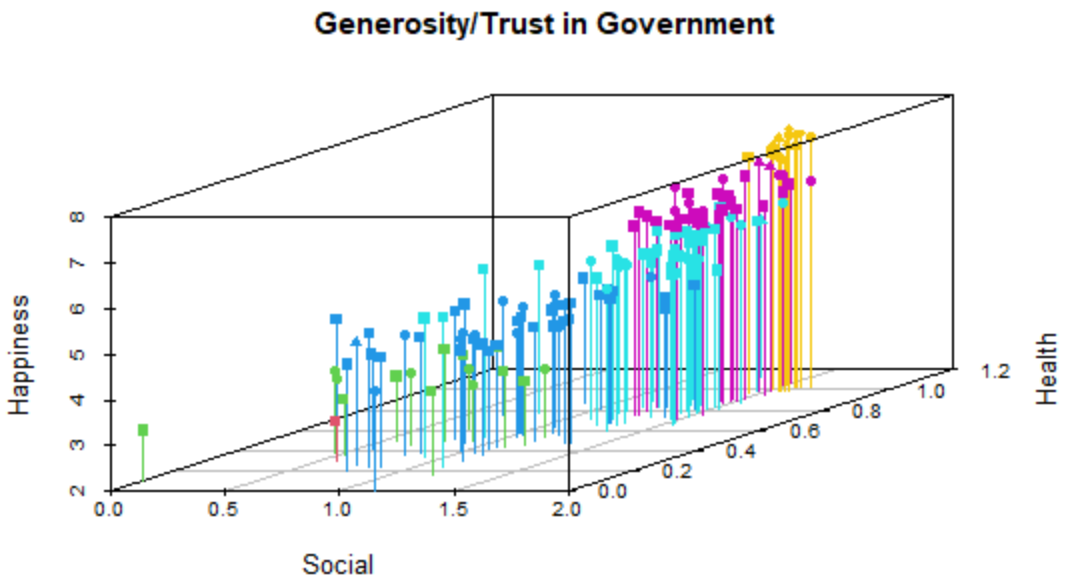
Asadar, de aici putem trage concluzia ca tarile care se afla in aceeasi regiune, impartasesc atat un nivel similar de fericire, cat si o economie similara. Pentru a verifica aceasta asumptie, vom vom calcula media acestor indici in functie de regiune. Ne vom folosi de functia “aggregate”, care ne permite sa grupam inregistrarile in functie de regiune, si tot odata sa aplicam functia “mean” asupra acestora.

Pe baza acestui nou dataframe, vom construi un bar plot pe baza caruia sa putem analiza vizual datele.



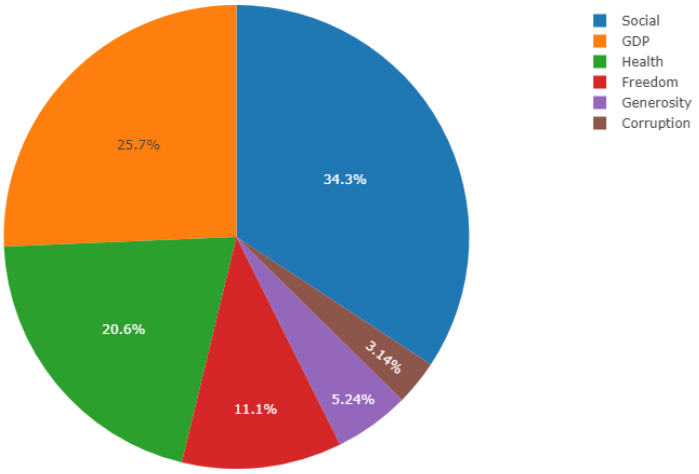
Din cate se poate observa, proportia PIB/Fericire se respecta, exceptie facand orientul mijlociu si tarile nord-africane, care in ciuda economiei mai ridicate, sunt in medie mai triste decat tarile din america latina. Cel mai probabil, acest lucru se datoreaza considerentilor socio-politici din aceasta zona.

Suportul social, perceptia coruptiei precum si sanatatea si speranta la viata sunt factori de natura administrativa care influenteaza indicele de fericire. Dorim sa vizualizam aceste elemente in plan 3D pentru a trage concluzii. Pentru aceasta, vom utiliza libraria “scatterplot3d”. Doua dintre axe vor fi reprezentate de sanatate, respectiv suport social, in timp ce inaltimea va fi data de catre fericire si forma de catre perceptia coruptiei.



Din figura se observa ca suportul social nu are o contributie atat de mare la nivelul fericirii, aceasta crescand mai degraba pe axa sanatatii. Chiar din contra, anumite tari care au in indice de suport social mic prezinta o indici mai mari de fericire decat unele tari cu sanatate similara si suport social mai ridicat. De asemenea, nici in cazul percepriei coruptiei nu se observa o mare influenta asupra fericirii, diverse forme regasindu-se amestecate pe fiecare nivel de inaltime.

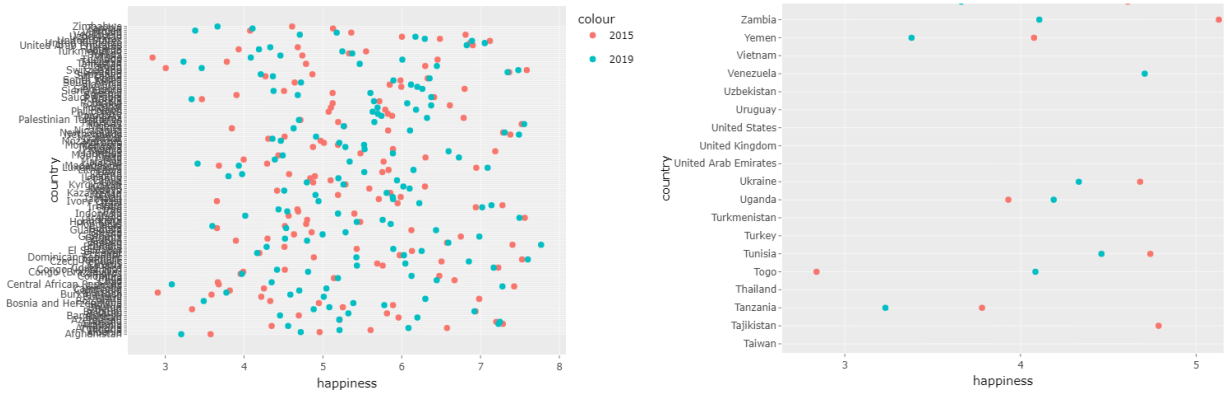
In ciuda faptului ca variatia indicelui social nu pare sa influenteze in vreun fel indicele de fericire, acesta totusi are o valuare destul de ridicata petru toate tarile in comparatie cu alti indici. Prin urmare, dorim sa vedem atat ponderea acestuia in indicele de fericire, cat si ponderile celorlalti indici contributori. Pentru acest lucru vom realiza un pie chart compus din mediile indicilor in anul 2019.



Din grafic se observa nu doar ca indicele social are cea mai mare pondere, ceea ce inseamna ca el are o valoarea apropiata in majoritatea tarilor, dar si faptul ca perceptia coruptiei are un procentaj infim, neavand posibilitatea sa schimbe foarte multe. Asadar, cel mai mare factor de influenta al fericirii ramane produsul intern brut, urmat de sanatate.

Industria de online advertising ar putea beneficia totodata, nu doar stiind indicii de fericire, dar si dinamica acestora in timp. In promovarea diverselor produse, adesea conteaza rata schimbarii indicelui mai mult decat indicele in sine. Asadar, dorim sa aflam care sunt tarile care au avut cea mai mare rata de schimbare (atat pozitiva cat si negativa) din anul 2015 pana in anul 2019.

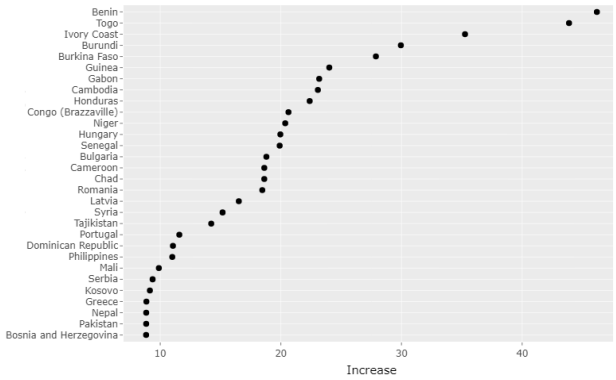
In primul rand, vom crea un nou data frame care va contine atat indicele din 2015 cat si cel din 2019. Functia “merge” va alatura aceste date in functie de coloana comuna (in acest caz, coloana “country”). Cu ajutorul librariei plotly vom genera un grafic aferent. Intrucat setul de date este foarte lung, graficul plotly ofera posibilitatea de a mari anumite zone din grafic si a naviga pe acestea intr-un mod interactiv, pentru o analiza amanuntita. Fiecare tara are propiul rand, pe care este indicat cu rosu indicele din anul 2015, iar cu albastru cel din anul 2019.

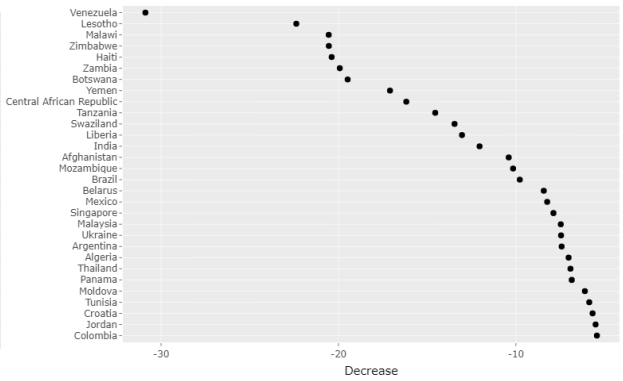


Se observa faptul ca pentru unele tari exista diferente in scor chiar si de 3 puncte intre cei 2 ani de referinta. Asadar, dorim sa vedem care sunt tarile cu cele mai mari rate de schimbare. Pentru acest lucru vom crea un nou data frame care va contine o coloana avand ca valoare raportul dintre indicele din 2019 si cel din 2015 minus unu pentru a obtine si valori negative, ulterior inmultit cu 100 pentru a obtine raportul procentual.

Astfel aflam ca cel mai mare procent de crestere a fericirii a fost inregistrat de Benin (46%) urmat de Togo (44%) si Coasta de Fildes (35%). La polul opus avem Venezuela (-31%) urmata de Lesotho (-22%) si Malawi (-20%).

Acest lucru se poate ilustra vizual, astfel, pentru primele 30 de tari cu cea mai mare rata de crestere, cat si pentru cele cu cea mai mica rata, s-au realizat doua grafice.





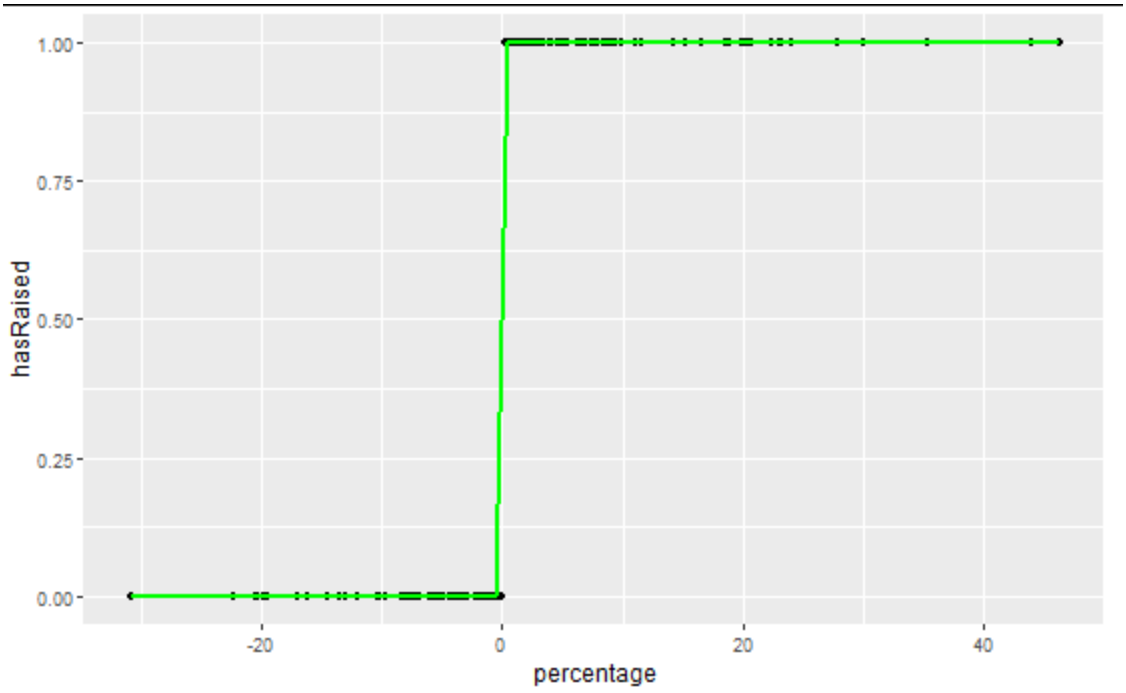
Pornind de la aceste data, se doreste realizarea unui model de regresie logistica care sa pezica daca economia va creste sau va scadea in functie de variatia fericirii. Pentru acest lucru se va crea un data frame nou care sa includa atat variatia fericirii 2015-2019, cat si valori de tip boolean care sa indice daca fericirea a crescut sau a scazut in respectivul an. Aceste valori se calculeaza pe baza variatiei economiei, comparand-o cu 0. Rezultatul returnat este de tip true/false care ulterior este convertit in 1/0.

Pe baza acestuia se creaza modelul de date stocat intr-o noua variabila. La actiunea comenzii summary(), ne sunt prezentate doura avertismente.

```
Warning: glm.fit: algorithm did not converge
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Call:
glm(formula = hasRaised ~ percentage, family = binomial, data = modelDF)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-4.739e-04  -2.000e-08   2.000e-08   2.000e-08   4.919e-04

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -10.24    1420.05  -0.007    0.994
percentage    101.20    7523.52   0.013    0.989

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.0639e+02  on 148  degrees of freedom
Residual deviance: 5.0883e-07  on 147  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

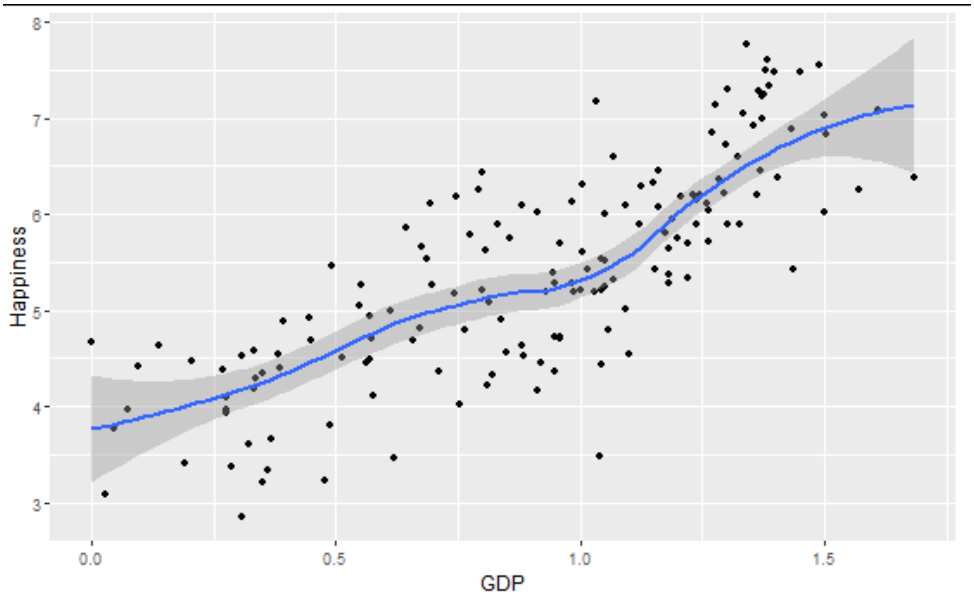
Acestea se datoreaza faptului ca variabila predictor x este capabila sa separe perfect variabilele raspuns y in 0 si 1. Cu alte cuvinte, pentru fiecare dintre tarile inregistrate, in momentul in care indicele fericirii a scazut, economia a scazut de asemenea, iar cand acesta a crescut, a crescut si economia.

Acest lucru se poate ilustra printr-un grafic reprezentand curba regresiei.



Din grafic se observa ca exista o linie verticala in dreptul indicatorului de procentaj 0. Acest lucru inseamna ca (cel putin in cazul tarilor evaluate) atunci cand fericirea este negativa, si economia este influentata negativ, acesta regula fiind valabila si in caz contrar. De aici se poate deduce ca produsul intern prut este probabil cel mai important factor in influenta indicelui de fericire.

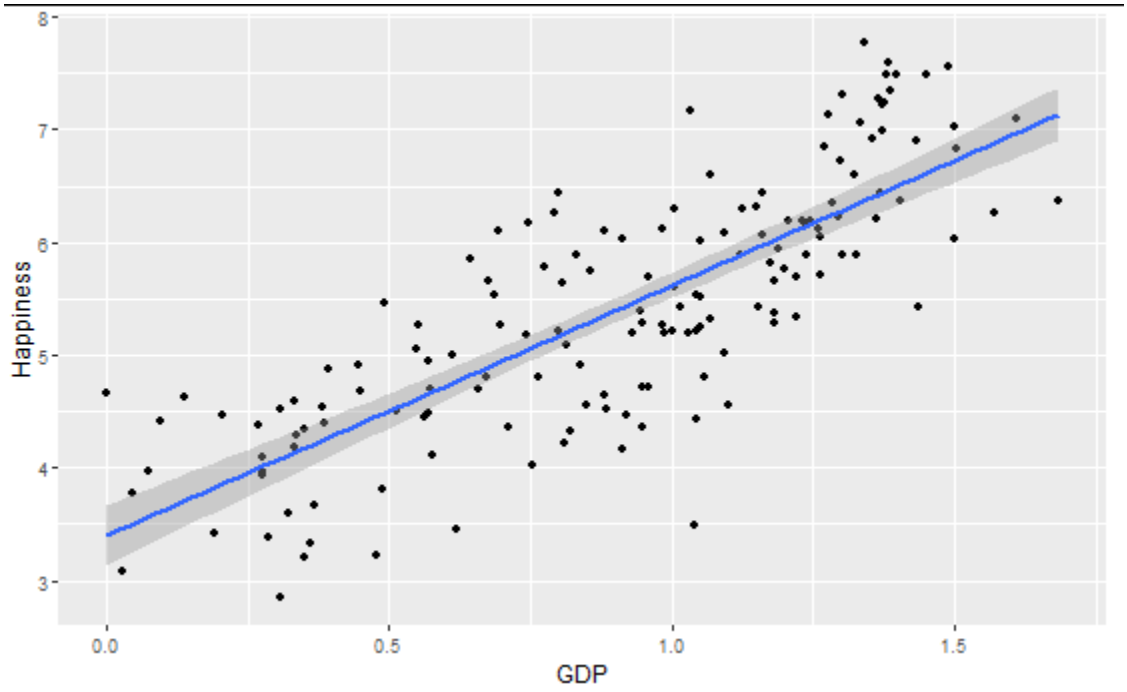
Asadar, regresia logistica s-a dovedit a fi nepotrivita pentru acest set de date. Vom incerca sa implementam o regresie liniara, pentru aceeasi coeficienti. In primul rand, dorim sa vizualizam cu ajutorul unei linii daca exista o relatie liniara intre cele doua variabile



Din grafic reiese ca fericirea creste liniar in raport cu produsul intern brut, lucru care este confirmat si de valoarea coeficientului de corelatie (0.7938829), acesta fiind apropiat de 1. Prin urmare putem continua analiza.

Vom crea un model liniar de forma: Happiness = b0 + b1 * GDP. Coeficientii beta pot fi determinati cu ajutorul functiei “lm”.

Aplicand functia “summary” aflam ca interceptorul(b0) are valoarea 3.39, ceea ce se traduce ca pentru un PIB egal cu 0, coeficientul de fericire mediu va fi egal cu 3.39. Coeficientul beta al PIB-ului este egal cu 2.21, ceea ce inseamna ca pentru fiecare unitate aditionala de PIB, indicele de fericire creste cu 2.21 unitati. Prin urmare, formula de predictie a fericirii in functie de PIB devine: Happiness = 3.39 + 2.21 * GDP. Acest lucru se poate vizualiza grafic daca adaugam o linie de regresie.



La o prima vedere, modelul pare ca se potriveste cu datele noastre. Pentru a verifica acest lucru, vom analiza indicii Residual standard error (RSE) si R-squared (R2) din sectiunea “summary”.

RSE-ul este egal cu 0.67, ceea ce inseamna ca valorile observate deviaza de la dreapta de regresie cu aproximatic 0.67 unitati. R patrat este egal cu 0.63, ceea ce inseamna ca mai mult de jumatate din variantele observate pot fi explicate de datele de intrare din model (in ce masura varianta variabilei dependente poate fi explicata de varianta variabilei independente).

Astfel, in cazul in care cei care practica advertising online doresc sa speculeze fericirea unei anumite aduiente, se pot folosi de formula determinata pentru regresie.

Setul de date are o serie de limitari in ceea ce priveste metodela care se pot aplica asupra acestuia, si implicit asupra analizelor care se pot realiza. O posibila solutionare a acestor limitari ar fi asocierea acestuia cu mai multe seturi de date. De exeplu, asocierea cu un set care contine populatia planetei in 2019 (anul de referinta), astfel putand sa ne raportam la date / cap de locuitor. O alta solutie ar fi aflarea valorilor numerice a indicilor din Dystopia, lucru care ne-ar permite sa lucram cu unitati, nu doar cu proportii.

Concluzia

In urma analizei demarate, se poate raspunde la fiecare din intrebarile adresate la inceput. Astfel, cele mai favorabile zone pentru a targeta o audienta pentru suplimente alimentare care favorizeaza fericirea si starea de bine sunt: Columbia, Iordan, Croatia, tunisia, Moldova, Panama, Tailanda, Algeria, Argentina si Ucraina, precum si regiunea Africa sub-sahariana. Factorii principali care influenteaza fericirea sunt economia, suportul social, sistemul de sanatate si libertatea. Fericirea se poate specula dupa formula fericire = 3.39 + 2.21 * (PIB / cel mai mic PIB inregistrat). Aceasta formula este accesibila advertiserilor in rice moment, ea putand fi adaptata si in concordanta cu alti factori economici.

In concluzie, pe fondul analizei au rezultat particularitati importante ale datelor care pot fi utilizate atat de catre beneficiarii consacrati (administratorii de campanii de online marketing), cat si de alte persoane.