**PRÀCTICA 2: Tipologia i cicle de vida de les dades**

Gerard Rosés Terrón i Gisela Claret Tortajada

08/06/2021

# 1. Descripció del dataset

Cal incloure:

* Descripció del dataset principal (hapiness)(variables + explicació)
* Perquè és important? Motivació per l’elecció del conjunt de dades.
* Què volem respondre?
  + Factors (o variables) que determinen nivell de happiness, especialment pel que fa al continent. (hi ha continents on la felicitat és major que en la resta, cas concret: Africa)
* Llicència de les dades

# 2. Integració i selecció de les dades d’interès a analitzar

* Incloure el tractament previ?
* Presentació breu del segon csv?
* Llicència de les dades
* Merge?
* Selecció de columnes importants?
* Canvi de noms de les variables
* Arrodoniment de Hapiness\_rate
* Creació Hapiness i Years.

# 3. Neteja de les dades

## Zeros o elements buits

* Descripció del tractament dels valors nuls:
* Correcció de les dades de l’atribut continent
* AVG i KNN
* Adjuntar captures de les taules
* Motius de l’elecció de KNN

## Identificació i tractament dels valors extrems

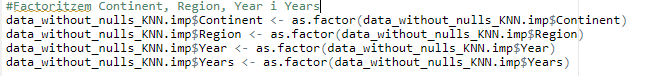
- explicació com s’ha detectat valors extrems

- motius per no eliminar-los

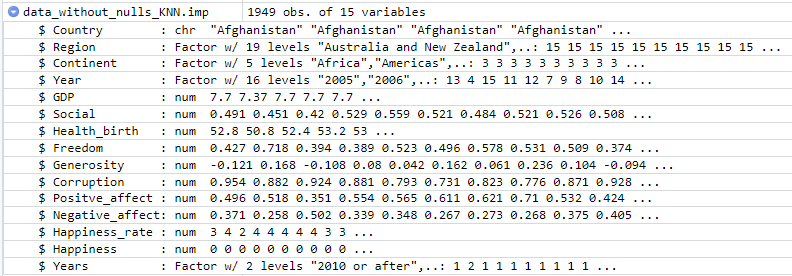
# 4. Anàlisi de dades

En aquest apartat s’explicarà la realització de l’anàlisi del conjunt de dades. En primer lloc, s’explicaran els diferents subconjunts de dades seleccionats per dur a terme les anàlisis estadístiques. També s’explicarà com s’ha dut a terme la comprovació de la normalitat de les dades i l’homogeneïtat de la variància. Per últim, es presentaran les proves estadístiques realitzades per comparar els grups de dades; correlació entre les variables, contrast d’hipòtesis i una regressió lineal.

Prèviament a la selecció dels grups de dades, però, s’ha realitzat la factorització d’alguns dels atributs qualitatius. Aquests atributs són: Continent, Region, Year i Years, fent ús de la funció *as.factor()*, tal com es mostra en la següent figura:



Observem que finalment obtenim un *dataframe* amb atributs nous tal com es mostra a continuació.



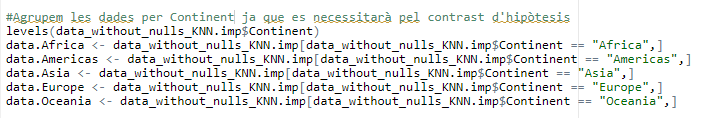
El motiu pel qual s’ha decidit no factoritzar l’atribut “Country” és, en primer lloc, un atribut amb un gran nombre de valors únics, i en segon lloc, no serà necessari per a l’estudi de la felicitat.

## Selecció de grups de dades

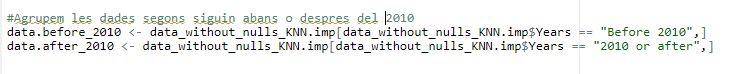
En primer lloc se separen les dades segons els continents. En la següent imatge es mostra quins són els diferents valors que pot prendre la variable “Continent”:



Així, s’han creat cinc grups diferents en els quals s’inclouen tots els valors pertanyents al continent corresponent, utilitzant el codi mostrat a continuació.



En segon lloc, també s’han creat dos grups de dades que poden resultar interessants en la realització de les proves estadístiques. El primer grup conté les dades pertanyents a abans del 2010 i el segon grup conté les dades pertanyents a després del 2010. Aquesta separació de les dades en dos grups s’ha creat fent ús de la columna que s’ha creat anteriorment anomenada “Years”. Per tal de crear aquests dos grups s’ha utilitzat el codi següent:

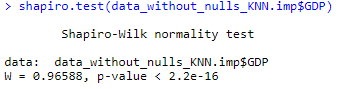


## Comprovació de la normalitat i homogeneïtat de la variància

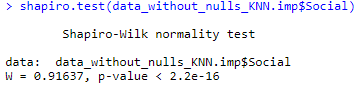
Per tal de comprovar si les dades estan o no normalitzades, s’aplica el test de *Shapiro-Wilk.* Aquest test planteja com a hipòtesi nul·la, que una mostra té una població amb una distribució normal. Així, en aplicar-lo, es rebutja la hipòtesi nul·la si el valor p és inferior al nivell de significança (alfa), i per tant es pot afirmar que la distribució no és normal.

En el cas que ens ocupa es té un valor α = 0.05, per tant, si en l’aplicar el test s’obtenen valors majors que 0.05, es pot afirmar que la variable segueix una distribució normal. Si, al contrari aquest valor és menor que 0.05, la distribució no serà normal. S’observarà que en tots els casos, s’obté un valor p menor que el valor alfa, i, per tant, cap de les variables serà distribuïda normalment.

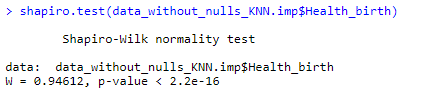
* **GDP per capita:**



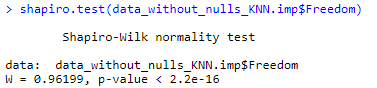
* **Social support:**



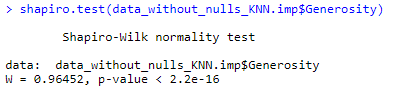
* **Healthy life expectancy at birth:**



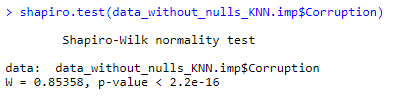
* **Freedom to make life choices:**



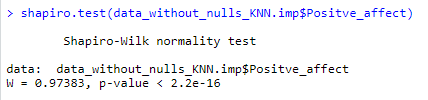
* **Generosity:**



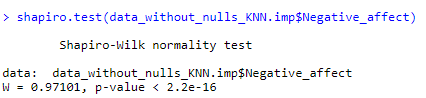
* **Perceptions of corruption:**



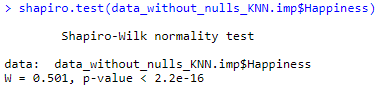
* **Positive affect**



* **Negative affect**



* **Happiness**:



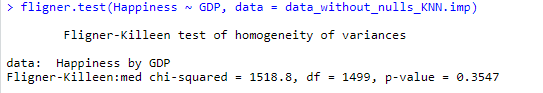
Tal com s’ha dit abans, podem observar que cap de les variables segueix una distribució normal.

Homogeneïtat de la variància:

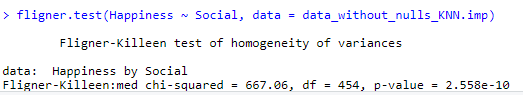
Per tal d’estudiar l’homogeneïtat de la variància també s’utilitza una tècnica basada en el contrast d’hipòtesis, anomenat el test de Flinger-Killeen. En aquest cas, la hipòtesi nul·la és que les variàncies dels dos grups són la mateixa.

A continuació es mostra els resultats obtinguts:

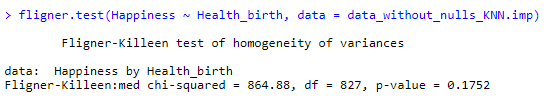
* **GDP per capita:**



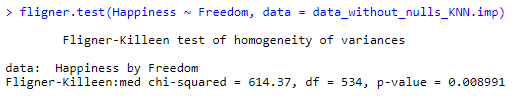
* **Social Support:**



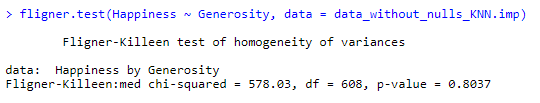
* **Healthy life expectancy at birth:**



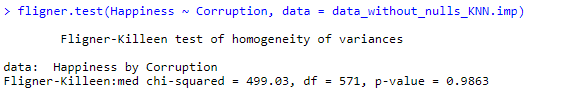
* **Freedom to make life choices:**



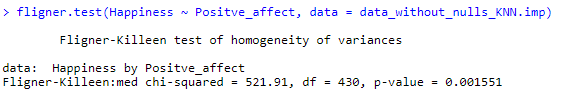
* **Generosity:**



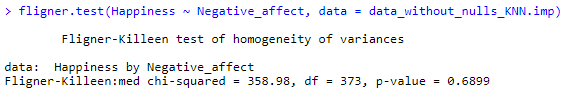
* **Perceptions of corruption:**



* **Positive affect:**



* **Negative affect:**



En aquest cas observem que les variables homogènies són:

* GDP per capita
* Healthy life expectancy at birth
* Generosity
* Perceptions of corruption
* Negative affect

Les variables no homogènies són:

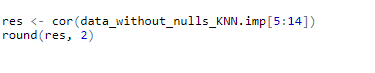
* Social Support
* Freedom to make life choices:
* Positive affect

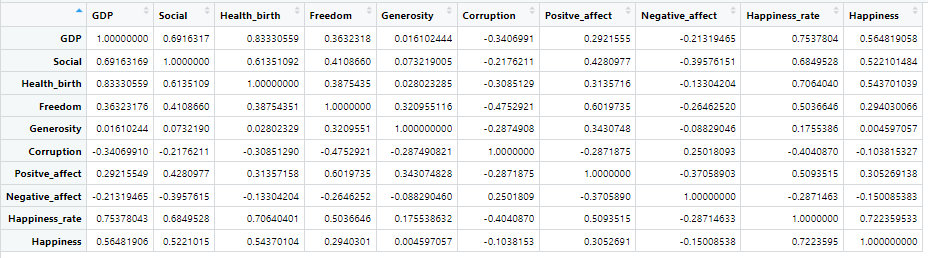
## Aplicació de les proves estadístiques per comparar els grups de dades

* **Matriu de correlació**

La matriu de correlació ens permet observar quines són les variables quantitatives que més correlació tenen entre elles. Conté nombres entre -1 i 1, essent les variables amb un valor més proper a 1 i -1 les que més correlació tenen. Els valors amb correlació negativa, afecten negativament i els que tenen una correlació positiva afecten positivament.

Per tal d’obtenir-la s’ha utilitzat el següent codi:



En la següent imatge es mostra una taula que representa la matriu de correlació obtinguda:

Pel que fa a la variable objectiu Hapiness, s’observa que les variables que més correlació tenen amb ella són: GPD per capita, Social support, Healthy life expectancy at birth i Hapiness\_rate. Aquesta última era d’esperar, ja que els valors de Happiness han estat obtinguts a través de Hapiness rate.

**- Contrast d’Hipòtesi**

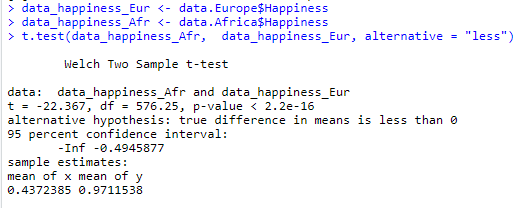
En aquest apartat s’estudiarà mitjançant contrast d’hipòtesis quins són els continents en què la felicitat és menor. Cal tenir en compte que els conjunts de dades que utilitzarem tindran entre 319 i 684 observacions, i per tant, es pot utilitzar el contrast d’hipòtesi tot i no tenir una distribució normal de les dades.

Així, es plantejaran diferents contrasts paramètrics d’hipòtesis de dues mostres sobre la diferència de les mitges, per tal d’observar la diferència entre la felicitat entre continents.

En el primer s’assumirà

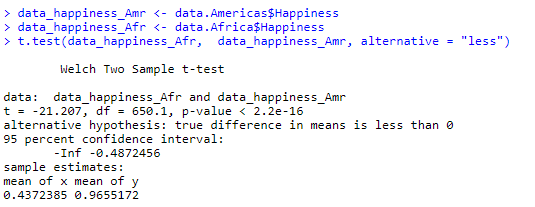
On s’assumirà α = 0.05

Obtenim:



S’obté un *valor p* menor que 0.05, per tant es rebutja la hipòtesi nul·la i per tant podem concloure que la felicitat és major en Europa que en Africa.

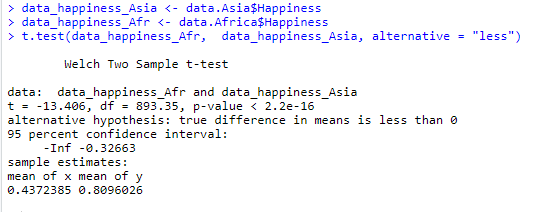
Pel segon cas, s’estudiarà Africa i Amèrica. De nou es plantegen les hipòtesis següents:



En aquest cas també és menor que 0.05, per tant, es pot tornar a afirmar que la felicitat és major en Amèrica que en Africa.

El tercer cas que s’estudia és el de Africa i Asia. Com en els dos casos anteriors es plantegen les següents hipòtesis:

Els resultats que s’obtenen són els següents:

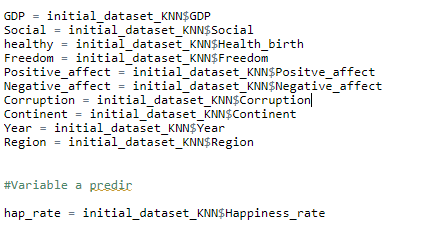


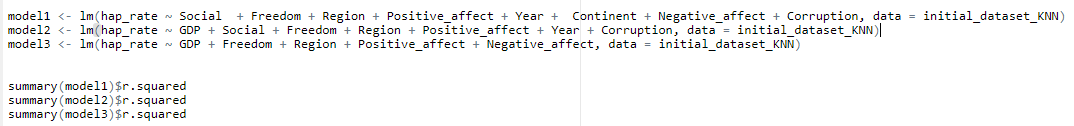
De nou, rebutgem la hipòtesi nul·la i obtenim que a Africa la felicitat és menor que a Asia.

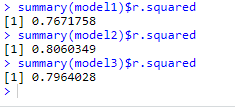
* **Finalment s’ha realitzat un model de regressió lineal**

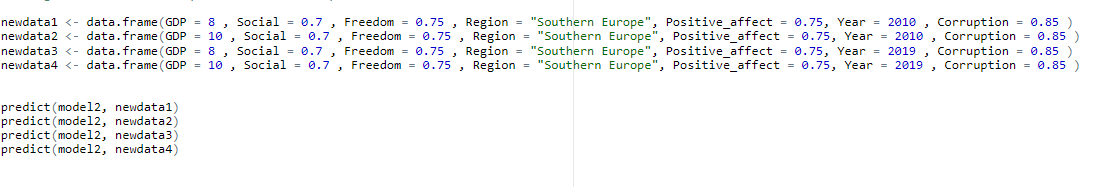
S’han estudiat tres models de regressió lineals diferents per tal de realitzar prediccions sobre quin serà el *Happiness rate* segons unes condicions concretes. D’aquests tres models estudiats s’ha escollit el de més precisió i s’han realitzat un seguit de prediccions que es presentaran a continuació. Es vol mencionar que el conjunt de dades que s’ha utilitzat és el que s’ha obtingut després de tractar els valors nuls però abans d’afegir les columnes de *Happiness* i *Years.*

Es defineixen els regressors i la variable a predir:

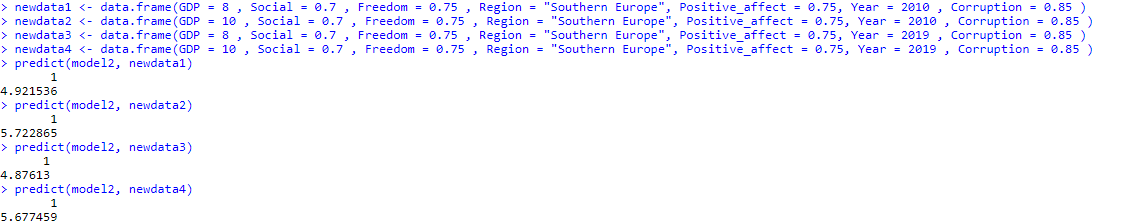


Seguidament es defineixen els models i s’avaluen per tal d’obtenir quin és amb el que s’obtindran millors resultats.



Observem que s’obtenen millors resultats pel segon model, així que s’utilitzarà aquest per realitzar les prediccions.

El valor de *Happiness* *rate* predit és el següent:

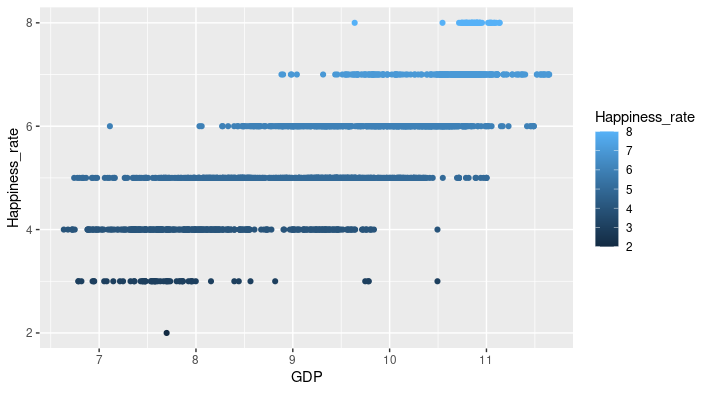


S’observa que si es manté el valor de *social support, freedom to make life choices, Region, Positive affect i corruption* fixes obtenim que:

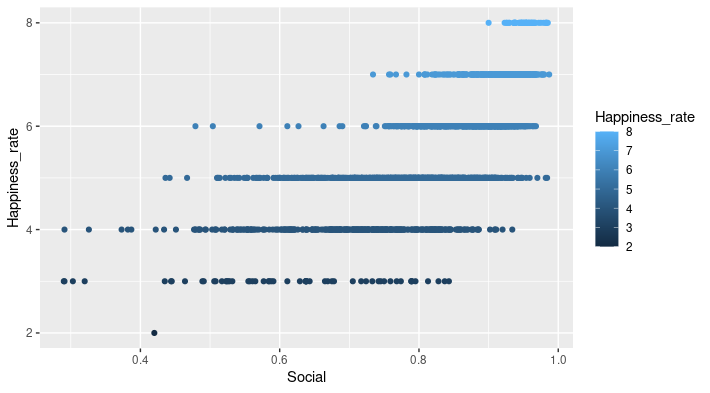
* En augmentar GDP per càpita de 8 a 10 augmenta *hapiness rate* de 4.92 a 5.72
* En l’any 2010 el *happiness rate* era lleugerament major que en el 2019.

# 5. Representació dels resultats a partir de taules i gràfiques

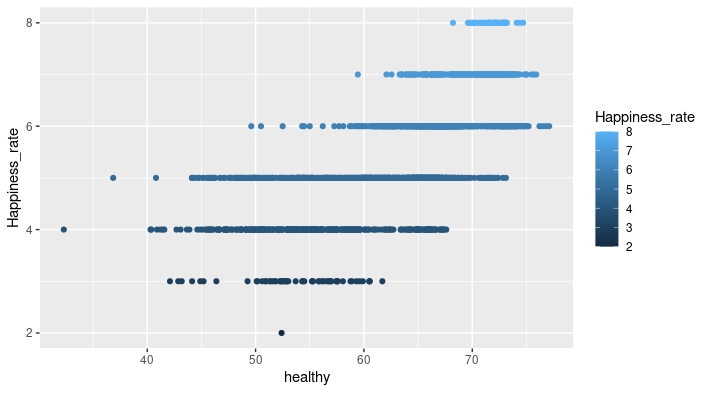
En primer lloc s’ha estudiat matriu de correlació i s’ha observat que les variables que més correlació tenen amb l’atribut *happiness\_rate* són *GPD per capita, social support i healthy life expectancy at birth.* En canvi, l’atribut que menys correlació té amb la variable *happiness\_rate* és *Generosity.* Per tal de comprovar-ho, s’han representat les gràfiques *GDP – Happiness\_rate*, *Social\_support – happiness\_rate, healthy life expectancy at birth – Happiness\_rate* i *generosity – happiness\_rate*. Es pot comprovar com efectivament, pels tres primer es mostra una tendència creixent de la felicitat a mesura que augmenten els valors de *GDP*, *Social support* i *healthy life expectancy at birht.*  En canvi, pel que fa a *generosity* no es mostra cap tendència.



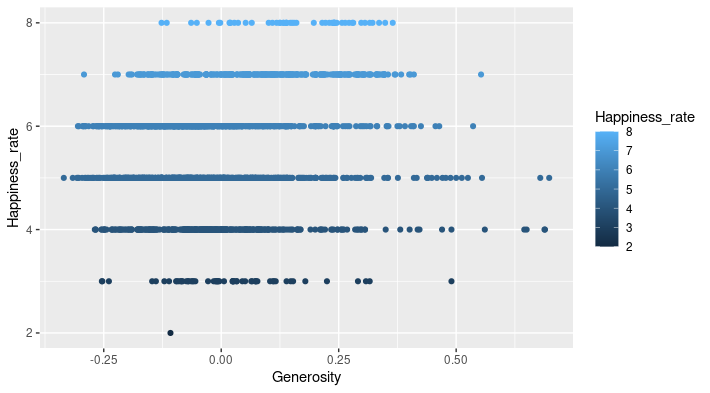
En aquesta primera gràfica s’observa que a mesurar que augmenta el valor de GDP augmenten també els valors de happiness.



En aquest cas també obtenim que a mesura que augmenten els valors de *social support* també augmenten els valors de la felicitat.

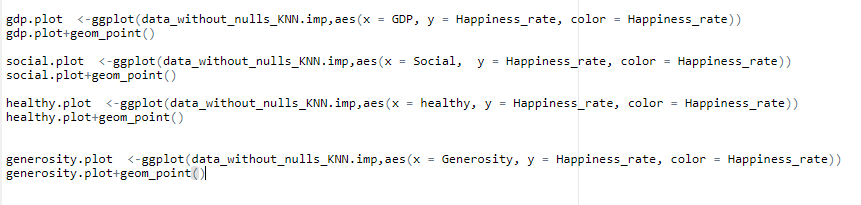


El valor de *happiness\_rate* també augmenta a mesura que augmenten els valors de *healthy life expectancy at birth.*

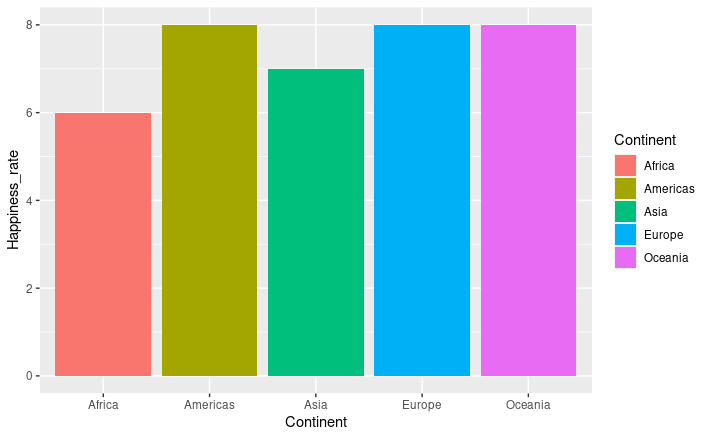


En canvi, aquí es demostra que per valors alts i baixos de generosity obtenim *happiness\_rate* de tots els rangs, així que aquestes dos variables no estan altament correlacionades.

El codi utilitzat per la representació d’aquestes gràfiques és el següent:



En segon lloc, mitjançant contrast d’hipòtesis s’ha arribat a la conclusió que les persones que viuen a Africa tenen un *happiness\_rate* menor. En el següent gràfic de barres es demostra que efectivament això és així:



Pel que fa a la regressió lineal, s’han obtingut els resultats següents:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **GDP** | **SOCIAL** | **FREEDOM** | **REGION** | **POSITIVE AFFECT** | **YEAR** | **CORRUPTION** | **HAPPINESS**  **RATE** |
| 8 | 0.7 | 0.75 | Sourthern Europe | 0.75 | 2010 | 0.85 | 4.92 |
| 10 | 0.7 | 0.75 | Sourthern Europe | 0.75 | 2010 | 0.85 | 5.72 |
| 8 | 0.7 | 0.75 | Sourthern Europe | 0.75 | 2019 | 0.85 | 4.87 |
| 10 | 0.7 | 0.75 | Sourthern Europe | 0.75 | 2019 | 0.85 | 5.68 |

# 6. Resolució dels resultats

Pel que fa als objectius que ens hem proposat, hem aconseguit respondre totes les preguntes plantejades. S’ha plantejat comprovar quins factors afavorien la felicitat, quins és el continent amb uns menors valors de *happiness rate* i també es volia realitzar prediccions del valor de felicitat que es tindria segons uns

Així, una vegada s’ha realitzat la neteja de dades i realitzant un seguit d’anàlisis estadístiques s’han arribat a les conclusions següents:

* Les persones que viuen a Africa tenen més probabilitats de ser infelices.
* Tenir una GPD per capita major afecta positivament la felicitat.
* El social support també té molt a veure amb un augment de la felictiat.
* Tenir un *healthy life expectancy at birth* alt afecta positivament la felicitat.
* La generositat no és un factor que cal tenir gaire en compte a l’hora de predit la felicitat.
* En la regió del sud d’Europa la població era lleugerament més infeliç al 2019 que en el 2010

# 7. Codi

L’enllaç al Github creat, el qual conté el codi utilitzat per la realització d’aquesta pràctica és el següent:

\*\*\*\*\*\*\*Enllaç github\*\*\*\*\*\*\*\*\*

# 8.Taula de contribucions

|  |  |
| --- | --- |
| **CONTRIBUCIONS** | **SIGNA** |
| Recerca prèvia | G. R. T., G. C. T. |
| Redacció de les respostes | G. R. T., G. C. T. |
| Desenvolupament codi | G. R. T., G. C. T. |