

PRÀCTICA 2: Tipologia i cicle de vida de les dades

Gerard Rosés Terrón i Gisela Claret Tortajada

08/06/2021

1. Descripció dels jocs de dades

- El primer joc de dades és el *world-happiness-report-2021.csv* conté la informació de l'estat de la felicitat en els països del món en els últims anys.

La taxa de felicitat, que és la variable dependent, és el resultat dels factors següents:

- Producte brut nacional
- Suport social
- Esperança de vida
- Llibertat
- Absència de corrupció
- Generositat
- Sentiments positius (https://en.wikipedia.org/wiki/Positive_affectivity)
- Sentiments negatius (https://en.wikipedia.org/wiki/Negative_affectivity)

Aquests factors són les variables independents que determinen la taxa de felicitat d'un país en concret en un any concret.

- El segon joc de dades és el *countryContinent.csv* i consisteix en una definició dels països que ens ajuda a agrupar-los per regions o continents en aquest exercici.
- La llicència dels jocs de dades utilitzats per aquest exercici són de domini públic.
- La motivació principal és comprendre quins factors contribueixen més en la felicitat d'un país, especialment a nivell de continent. I fer comparacions entre continents.

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

- Em intentat utilitzar una API per integrar kaggle.com amb el nostre codi en R. Actualment existeix una API no oficial que no em pogut fer funcionar per extraure les dades de manera automàtica. Aparentment només existeix una API oficial i aquesta és per Python.

Finalment, de manera manual, em baixat les dades de la xarxa i les hem guardat en local.

- Per poder fer agrupacions entre continents i regions em afegit el joc de dades *countryContinent.csv*. I s'han creuat les dades dels dos conjunts de dades.

```
data <- merge(x = whr_df, y = countries, by.x = "Country.name", by.y = "country", all.x = TRUE)
```

- Degut a que la clau dels jocs de dades era el nom de cada país, hi ha hagut registres que no s'han creuat correctament perquè els noms eren diferents (per exemple, "United States of America" en *countryContinent.csv* i "United States" en *world-happiness-report.csv*). A conseqüència em hagut de assignar la informació manualment imputant el continent i regió d'un país veí.

```
data$continent[data$Country.name=="United States"] <- data$continent[data$Country.name=="Canada"][1]  
data$sub_region[data$Country.name=="United States"] <- data$sub_region[data$Country.name=="Canada"][1]
```

- Finalment em seleccionat les columnes d'interès per aquest exercici i les em ordenat d'una manera més entenedora.

```
data <- data[,c(1,13,12,2,4,5,6,7,8,9,10,11,3)]
```

- S'han assignat noms més entenedors per les columnes del joc de dades.

```
names(data)[names(data) == "Country.name"] <- "Country"  
names(data)[names(data) == "sub_region"] <- "Region"  
names(data)[names(data) == "continent"] <- "Continent"  
names(data)[names(data) == "year"] <- "Year"  
names(data)[names(data) == "Log.GDP.per.capita"] <- "GDP"  
names(data)[names(data) == "Social.support"] <- "Social"  
names(data)[names(data) == "Healthy.life.expectancy.at.birth"] <- "Health_birth"  
names(data)[names(data) == "Freedom.to.make.life.choices"] <- "Freedom"  
names(data)[names(data) == "Perceptions.of.corruption"] <- "Corruption"  
names(data)[names(data) == "Positive.affect"] <- "Positive_affect"  
names(data)[names(data) == "Negative.affect"] <- "Negative_affect"  
names(data)[names(data) == "Life.Ladder"] <- "Happiness_rate"
```

- Tractament de valors nulls amb 2 mètodes. Imputació de la mitjana agrupat per país i per predicció amb KNN.

Amb imputació de la mitjana el codi utilitzat és el següent:

```
data_without_nulls_AVG.imp <- data %>% group_by(`Country`) %>%
  mutate(`GDP` = ifelse(is.na(`GDP`), mean(`GDP`, na.rm=TRUE), `GDP`),
         `Social` = ifelse(is.na(`Social`), mean(`Social`, na.rm=TRUE), `Social`),
         `Health_birth` = ifelse(is.na(`Health_birth`), mean(`Health_birth`, na.rm=TRUE), `Health_birth`),
         `Freedom` = ifelse(is.na(`Freedom`), mean(`Freedom`, na.rm=TRUE), `Freedom`),
         `Generosity` = ifelse(is.na(`Generosity`), mean(`Generosity`, na.rm=TRUE), `Generosity`),
         `Corruption` = ifelse(is.na(`Corruption`), mean(`Corruption`, na.rm=TRUE), `Corruption`),
         `Positive_affect` = ifelse(is.na(`Positive_affect`), mean(`Positive_affect`, na.rm=TRUE), `Positive_affect`),
         `Negative_affect` = ifelse(is.na(`Negative_affect`), mean(`Negative_affect`, na.rm=TRUE), `Negative_affect`))
)
```

Com que quedaven valors nulls (per als grups que no tenien cap valor). s'han eliminat els registres que quedaven amb el codi següent:

```
data_without_nulls_AVG.imp <- data_without_nulls_AVG.imp[complete.cases(data_without_nulls_AVG.imp),]
```

- Amb la llibreria VIM i el codi següent s'han implementat la imputació de valors aplicant el mètode KNN:

```
library(VIM)
data_without_nulls_KNN.imp <- knn(data)
```

- Finalment s'han afegit les columnes "Hapiness" i "Years" per aplicar posteriorment contrastos d'hipòtesis.

3. Neteja de les dades

3.1 Zeros o elements buits

En el codi s'ha definit la funció `nulls_function()` per processar columna a columna aquest anàlisi.

En aquest apartat s'han empleat 2 mètodes:

- Imputació de la mitjana per país. S'han imputat als valors nuls el valor de la mitjana per al país en qüestió. Amb aquest mètode hi havia països sense dades i per tant no era possible calcular la mitjana. Per tant s'han eliminat certs registres i el joc de dades ha quedat reduït.
- Imputació segons KNN. S'han imputat als valors nuls el resultat d'aplicar l'algoritme KNN per predir el valor perdut d'un registre donat.

Finalment es tria el mètode de KNN per la imputació de valors perduts. D'aquesta manera no s'han d'eliminar cap registre.

3.2 Identificació i tractament dels valors extrems

S'ha considerat que un valor allunyat més de 3 desviacions típiques de la mitjana és un valor extrem.

En el codi s'ha definit la funció `ourliers_function()` per processar columna a columna aquest anàlisi.

Un cop identificats els valors nuls es pren la decisió de no eliminar-los perquè es considera que en el món hi ha molta diversitat i diferències entre països i eliminar els valors extrems faria que es perdés part de la variància per explicar aquest joc de dades.

A de més es considera que <https://worldhappiness.report/> és una font d'informació de qualitat.

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 qualitius. Aquests atributs són: Continent, Region, Year i Years, fent ús de la funció *as.factor()*, tal com es mostra en la següent figura:

```
#Factoritzem Continent, Region, Year i Years|
data_without_nulls_KNN.imp$Continent <- as.factor(data_without_nulls_KNN.imp$Continent)
data_without_nulls_KNN.imp$Region <- as.factor(data_without_nulls_KNN.imp$Region)
data_without_nulls_KNN.imp$Year <- as.factor(data_without_nulls_KNN.imp$Year)
data_without_nulls_KNN.imp$Years <- as.factor(data_without_nulls_KNN.imp$Years)
```

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

data_without_nulls_KNN.imp	1949 obs. of 15 variables													
\$ Country	:	chr	"Afghanistan"	"Afghanistan"	"Afghanistan"	"Afghanistan"	"Afghanistan"	...						
\$ Region	:	Factor w/ 19 levels	"Australia and New Zealand",...	15	15	15	15	15	15	15	15	15	15	...
\$ Continent	:	Factor w/ 5 levels	"Africa", "Americas",...	3	3	3	3	3	3	3	3	3	3	...
\$ Year	:	Factor w/ 16 levels	"2005", "2006",...	13	4	15	11	12	7	9	8	10	14	...
\$ GDP	:	num	7.7	7.37	7.7	7.7	7.7	...						
\$ Social	:	num	0.491	0.451	0.42	0.529	0.559	0.521	0.484	0.521	0.526	0.508	...	
\$ Health_birth	:	num	52.8	50.8	52.4	53.2	53	...						
\$ Freedom	:	num	0.427	0.718	0.394	0.389	0.523	0.496	0.578	0.531	0.509	0.374	...	
\$ Generosity	:	num	-0.121	0.168	-0.108	0.08	0.042	0.162	0.061	0.236	0.104	-0.094	...	
\$ Corruption	:	num	0.954	0.882	0.924	0.881	0.793	0.731	0.823	0.776	0.871	0.928	...	
\$ Positive_affect	:	num	0.496	0.518	0.351	0.554	0.565	0.611	0.621	0.71	0.532	0.424	...	
\$ Negative_affect	:	num	0.371	0.258	0.502	0.339	0.348	0.267	0.273	0.268	0.375	0.405	...	
\$ Happiness_rate	:	num	3	4	2	4	4	4	4	3	3	...		
\$ Happiness	:	num	0	0	0	0	0	0	0	0	0	...		
\$ Years	:	Factor w/ 2 levels	"2010 or after",...	1	2	1	1	1	1	1	1	1	...	

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.

4.1 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":

```
> levels(data_without_nulls_KNN.imp$Continent)
[1] "Africa" "Americas" "Asia" "Europe" "Oceania"
```

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.

```
#Agrupem les dades per Continent ja que es necessitarà pel contrast d'hipòtesis
levels(data_without_nulls_KNN.imp$Continent)
data.Africa <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Continent == "Africa",]
data.Americas <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Continent == "Americas",]
data.Asia <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Continent == "Asia",]
data.Europe <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Continent == "Europe",]
data.Oceania <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Continent == "Oceania",]
```

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:

```
#Agrupem les dades segons siguin abans o després del 2010
data.before_2010 <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Years == "Before 2010",]
data.after_2010 <- data_without_nulls_KNN.imp[data_without_nulls_KNN.imp$Years == "2010 or after",]
```

4.2 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 $\alpha = 0.05$, 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 està distribuïda normalment.

- GDP per capita:

```
> shapiro.test(data_without_nulls_KNN.imp$GDP)
```

Shapiro-Wilk normality test

```
data: data_without_nulls_KNN.imp$GDP
W = 0.96588, p-value < 2.2e-16
```

- Social support:

```
> shapiro.test(data_without_nulls_KNN.imp$Social)
```

Shapiro-Wilk normality test

```
data: data_without_nulls_KNN.imp$Social
W = 0.91637, p-value < 2.2e-16
```

- **Healthy life expectancy at birth:**

```
> shapiro.test(data_without_nulls_KNN.imp$Health_birth)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Health_birth  
W = 0.94612, p-value < 2.2e-16
```

- **Freedom to make life choices:**

```
> shapiro.test(data_without_nulls_KNN.imp$Freedom)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Freedom  
W = 0.96199, p-value < 2.2e-16
```

- **Generosity:**

```
> shapiro.test(data_without_nulls_KNN.imp$Generosity)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Generosity  
W = 0.96452, p-value < 2.2e-16
```

- **Perceptions of corruption:**

```
> shapiro.test(data_without_nulls_KNN.imp$Corruption)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Corruption  
W = 0.85358, p-value < 2.2e-16
```

- **Positive affect**

```
> shapiro.test(data_without_nulls_KNN.imp$Positive_affect)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Positive_affect  
W = 0.97383, p-value < 2.2e-16
```

- **Negative affect**

```
> shapiro.test(data_without_nulls_KNN.imp$Negative_affect)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Negative_affect
```

```
W = 0.97101, p-value < 2.2e-16
```

- **Happiness:**

```
> shapiro.test(data_without_nulls_KNN.imp$Happiness)
```

```
Shapiro-Wilk normality test
```

```
data: data_without_nulls_KNN.imp$Happiness
```

```
W = 0.501, p-value < 2.2e-16
```

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 *Fligner-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:**

```
> fligner.test(Happiness ~ GDP, data = data_without_nulls_KNN.imp)
```

```
Fligner-Killeen test of homogeneity of variances
```

```
data: Happiness by GDP
```

```
Fligner-Killeen:med chi-squared = 1518.8, df = 1499, p-value = 0.3547
```

- **Social Support:**

```
> fligner.test(Happiness ~ Social, data = data_without_nulls_KNN.imp)
```

```
Fligner-Killeen test of homogeneity of variances
```

```
data: Happiness by Social
```

```
Fligner-Killeen:med chi-squared = 667.06, df = 454, p-value = 2.558e-10
```


- **Healthy life expectancy at birth:**

```
> fligner.test(Happiness ~ Health_birth, data = data_without_nulls_KNN.imp)

    Fligner-Killeen test of homogeneity of variances

data:  Happiness by Health_birth
Fligner-Killeen:med chi-squared = 864.88, df = 827, p-value = 0.1752
```

- **Freedom to make life choices:**

```
> fligner.test(Happiness ~ Freedom, data = data_without_nulls_KNN.imp)

    Fligner-Killeen test of homogeneity of variances

data:  Happiness by Freedom
Fligner-Killeen:med chi-squared = 614.37, df = 534, p-value = 0.008991
```

- **Generosity:**

```
> fligner.test(Happiness ~ Generosity, data = data_without_nulls_KNN.imp)

    Fligner-Killeen test of homogeneity of variances

data:  Happiness by Generosity
Fligner-Killeen:med chi-squared = 578.03, df = 608, p-value = 0.8037
```

- **Perceptions of corruption:**

```
> fligner.test(Happiness ~ Corruption, data = data_without_nulls_KNN.imp)

    Fligner-Killeen test of homogeneity of variances

data:  Happiness by Corruption
Fligner-Killeen:med chi-squared = 499.03, df = 571, p-value = 0.9863
```

- **Positive affect:**

```
> fligner.test(Happiness ~ Positive_affect, data = data_without_nulls_KNN.imp)

    Fligner-Killeen test of homogeneity of variances

data:  Happiness by Positive_affect
Fligner-Killeen:med chi-squared = 521.91, df = 430, p-value = 0.001551
```

- **Negative affect:**

```
> fligner.test(Happiness ~ Negative_affect, data = data_without_nulls_KNN.imp)

    Fligner-Killeen test of homogeneity of variances

data:  Happiness by Negative_affect
Fligner-Killeen:med chi-squared = 358.98, df = 373, p-value = 0.6899
```

En aquest cas observem que les variables homogènies són:

- GDP per càpita
- 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

4.3 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:

```
res <- cor(data_without_nulls_KNN.imp[5:14])  
round(res, 2)
```

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

	GDP	Social	Health_birth	Freedom	Generosity	Corruption	Positive_affect	Negative_affect	Happiness_rate	Happiness
GDP	1.00000000	0.6916317	0.83330559	0.3632318	0.016102444	-0.3406991	0.2921555	-0.21319465	0.7537804	0.564819058
Social	0.69163169	1.0000000	0.61351092	0.4108660	0.073219005	-0.2176211	0.4280977	-0.39576151	0.6849528	0.522101484
Health_birth	0.83330559	0.6135109	1.00000000	0.3875435	0.028023285	-0.3085129	0.3135716	-0.13304204	0.7064040	0.543701039
Freedom	0.36323176	0.4108660	0.38754351	1.0000000	0.320955116	-0.4752921	0.6019735	-0.26462520	0.5036646	0.294030066
Generosity	0.01610244	0.0732190	0.02802329	0.3209551	1.000000000	-0.2874908	0.3430748	-0.08829046	0.1755386	0.004597057
Corruption	-0.34069910	-0.2176211	-0.30851290	-0.4752921	-0.287490821	1.0000000	-0.2871875	0.25018093	-0.4040870	-0.103815327
Positive_affect	0.29215549	0.4280977	0.31357158	0.6019735	0.343074828	-0.2871875	1.0000000	-0.37058903	0.5093515	0.305269138
Negative_affect	-0.21319465	-0.3957615	-0.13304204	-0.2646252	-0.088290460	0.2501809	-0.3705890	1.00000000	-0.2871463	-0.150085383
Happiness_rate	0.75378043	0.6849528	0.70640401	0.5036646	0.175538632	-0.4040870	0.5093515	-0.28714633	1.0000000	0.722359533
Happiness	0.56481906	0.5221015	0.54370104	0.2940301	0.004597057	-0.1038153	0.3052691	-0.15008538	0.7223595	1.000000000

Pel que fa a la variable objectiu Happiness, s'observa que les variables que més correlació tenen amb ella són: GDP per càpita, Social support, Healthy life expectancy at birth i *Happiness_rate*. Aquesta última era d'esperar, ja que els valors de Happiness han estat obtinguts a través de Happiness 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 contrastos 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

$$H_0 = \mu_{Africa} - \mu_{Europa} = 0$$

$$H_1: \mu_{Africa} - \mu_{Europa} < 0$$

On s'assumirà $\alpha = 0.05$

Obtenim:

```
> data_happiness_Eur <- data.Europe$Happiness
> data_happiness_Afr <- data.Africa$Happiness
> t.test(data_happiness_Afr, data_happiness_Eur, alternative = "less")

Welch Two Sample t-test

data: data_happiness_Afr and data_happiness_Eur
t = -22.367, df = 576.25, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.4945877
sample estimates:
mean of x mean of y
0.4372385 0.9711538
```

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 Àfrica.

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

$$H_0 = \mu_{Africa} - \mu_{Americas} = 0$$

$$H_1: \mu_{Africa} - \mu_{Americas} < 0$$

```
> data_happiness_Amr <- data.Americas$Happiness
> data_happiness_Afr <- data.Africa$Happiness
> t.test(data_happiness_Afr, data_happiness_Amr, alternative = "less")

Welch Two Sample t-test

data: data_happiness_Afr and data_happiness_Amr
t = -21.207, df = 650.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.4872456
sample estimates:
mean of x mean of y
0.4372385 0.9655172
```

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 Àfrica.

El tercer cas que s'estudia és el de Àfrica i Àsia. Com en els dos casos anteriors es plantegen les següents hipòtesis:

$$H_0 = \mu_{Africa} - \mu_{Asia} = 0$$

$$H_1: \mu_{Africa} - \mu_{Asia} < 0$$

Els resultats que s'obtenen són els següents:

```
> data_happiness_Asia <- data.Asia$Happiness
> data_happiness_Afr <- data.Africa$Happiness
> t.test(data_happiness_Afr, data_happiness_Asia, alternative = "less")

Welch Two Sample t-test

data: data_happiness_Afr and data_happiness_Asia
t = -13.406, df = 893.35, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.32663
sample estimates:
mean of x mean of y
0.4372385 0.8096026
```

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

- 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:

```
GDP = initial_dataset_KNN$GDP
Social = initial_dataset_KNN$Social
healthy = initial_dataset_KNN$Health_birth
Freedom = initial_dataset_KNN$Freedom
Positive_affect = initial_dataset_KNN$Positive_affect
Negative_affect = initial_dataset_KNN$Negative_affect
Corruption = initial_dataset_KNN$Corruption
Continent = initial_dataset_KNN$Continent
Year = initial_dataset_KNN$Year
Region = initial_dataset_KNN$Region

#Variable a predir

hap_rate = initial_dataset_KNN$Happiness_rate
```

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

```
model1 <- lm(hap_rate ~ Social + Freedom + Region + Positive_affect + Year + Continent + Negative_affect + Corruption, data = initial_dataset_KNN)
model2 <- lm(hap_rate ~ GDP + Social + Freedom + Region + Positive_affect + Year + Corruption, data = initial_dataset_KNN)
model3 <- lm(hap_rate ~ GDP + Freedom + Region + Positive_affect + Negative_affect, data = initial_dataset_KNN)
```

```
summary(model1)$r.squared
summary(model2)$r.squared
summary(model3)$r.squared
```

```
> summary(model1)$r.squared
[1] 0.7671758
> summary(model2)$r.squared
[1] 0.8060349
> summary(model3)$r.squared
[1] 0.7964028
>
```

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

```
newdata1 <- data.frame(GDP = 8 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2010 , Corruption = 0.85 )
newdata2 <- data.frame(GDP = 10 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2010 , Corruption = 0.85 )
newdata3 <- data.frame(GDP = 8 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2019 , Corruption = 0.85 )
newdata4 <- data.frame(GDP = 10 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2019 , Corruption = 0.85 )
```

```
predict(model2, newdata1)
predict(model2, newdata2)
predict(model2, newdata3)
predict(model2, newdata4)
```

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

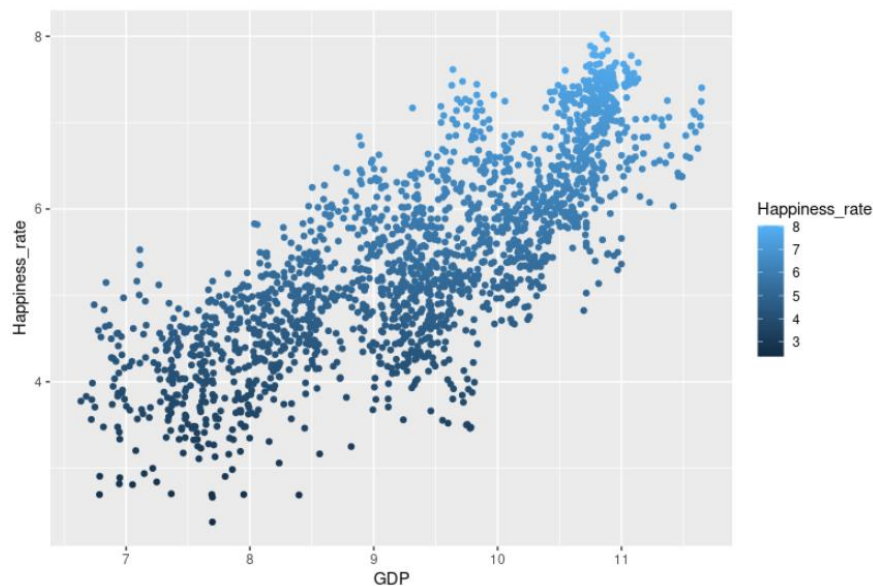
```
> predict(model2, newdata1)
1
4.921536
> predict(model2, newdata2)
1
5.722865
> predict(model2, newdata3)
1
4.87613
> predict(model2, newdata4)
1
5.677459
```

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

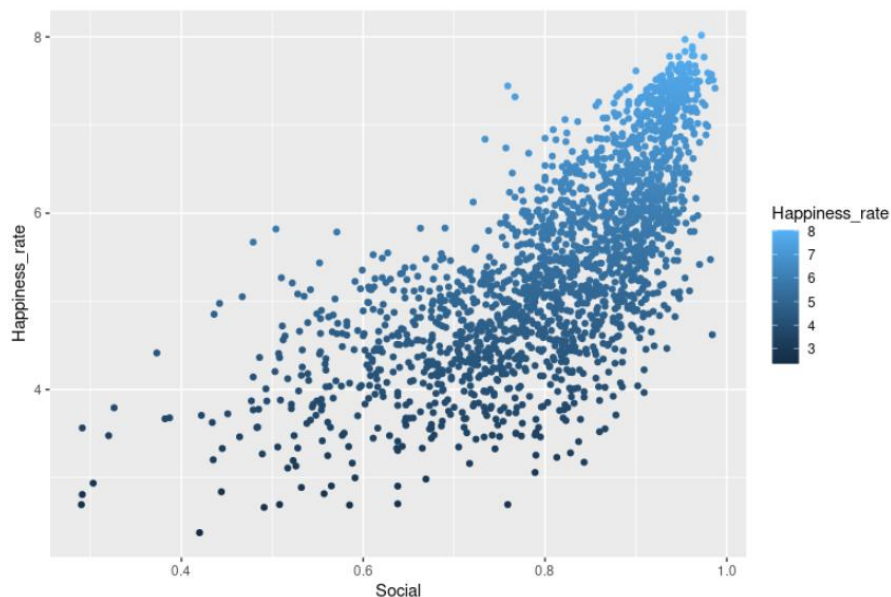
- En augmentar GDP per càpita de 8 a 10 augmenta *happiness rate* de 4.92 a 5.72
- En l'any 2010 el *happiness rate* era lleugerament major que en el 2019.

5. Representació dels resultats a partir de taules i gràfiques

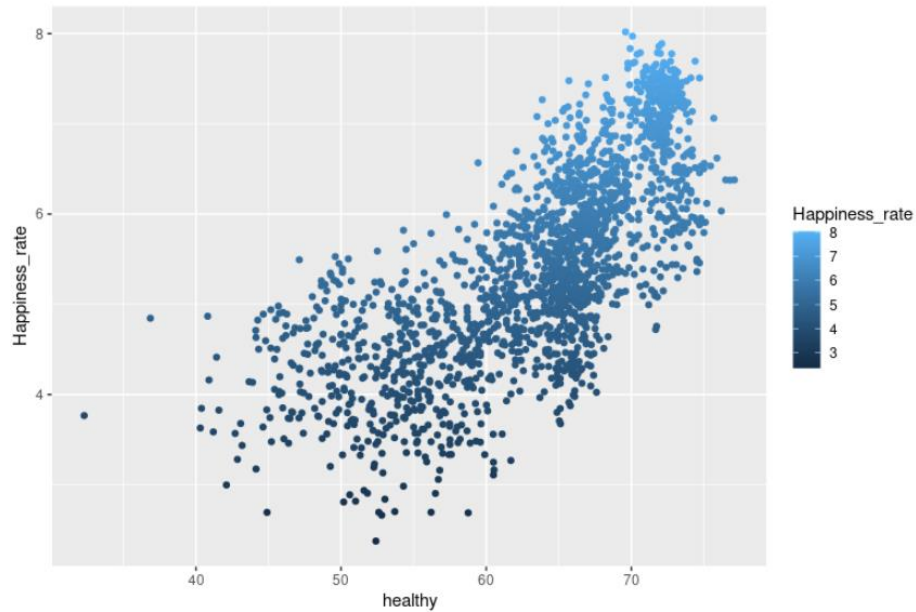
En primer lloc s'ha estudiat matriu de correlació i s'ha observat que les variables que més correlació tenen amb l'atribut *happiness_rate* són *GDP per càpita*, *social support* i *healthy life expectancy at birth*. En canvi, l'atribut que menys correlació té amb la variable *happiness_rate* és *Generosity*. 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 birth*. En canvi, pel que fa a *generosity* no es mostra cap tendència.



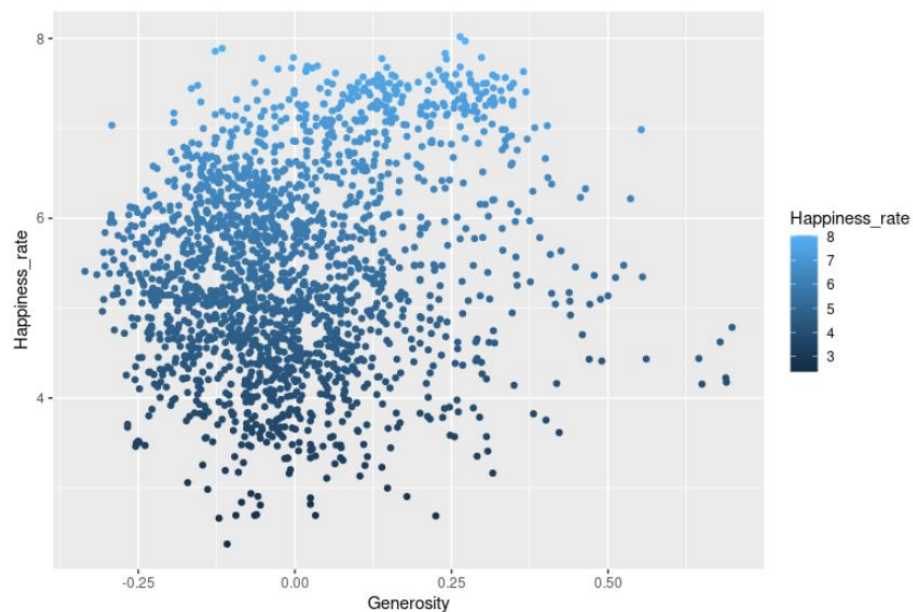
En aquesta primera gràfica s'observa que a mesura que augmenta el valor de GDP augmenten també els valors de happiness.



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:

```
gdp.plot <-ggplot(data_without_nulls_KNN.imp,aes(x = GDP, y = Happiness_rate, color = Happiness_rate))
gdp.plot+geom_point()

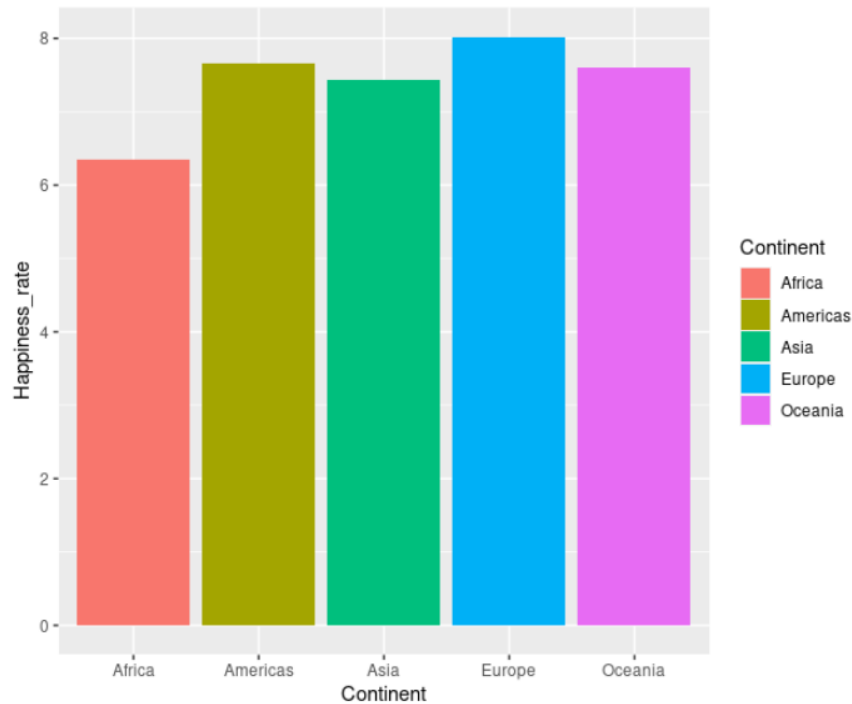
social.plot <-ggplot(data_without_nulls_KNN.imp,aes(x = Social, y = Happiness_rate, color = Happiness_rate))
social.plot+geom_point()

healthy.plot <-ggplot(data_without_nulls_KNN.imp,aes(x = healthy, y = Happiness_rate, color = Happiness_rate))
healthy.plot+geom_point()

generosity.plot <-ggplot(data_without_nulls_KNN.imp,aes(x = Generosity, y = Happiness_rate, color = Happiness_rate))
generosity.plot+geom_point()
```


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

```
happiness_continent.plot <- ggplot(data_without_nulls_KNN.imp, aes(x = Continent, y = Happiness_rate, fill = Continent))
happiness_continent.plot + geom_bar(stat="identity", position="identity")
```



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

```
newdata1 <- data.frame(GDP = 8 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2010 , Corruption = 0.85 )
newdata2 <- data.frame(GDP = 10 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2010 , Corruption = 0.85 )
newdata3 <- data.frame(GDP = 8 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2019 , Corruption = 0.85 )
newdata4 <- data.frame(GDP = 10 , Social = 0.7 , Freedom = 0.75 , Region = "Southern Europe", Positive_affect = 0.75, Year = 2019 , Corruption = 0.85 )

predict(model2, newdata1)
predict(model2, newdata2)
predict(model2, newdata3)
predict(model2, newdata4)
```

GDP	SOCIAL	FREEDOM	REGION	POSITIVE AFFECT	YEAR	CORRUPTION	HAPPINESS RATE
8	0.7	0.75	Southern Europe	0.75	2010	0.85	4.92
10	0.7	0.75	Southern Europe	0.75	2010	0.85	5.72
8	0.7	0.75	Southern Europe	0.75	2019	0.85	4.87
10	0.7	0.75	Southern 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 registres donats.

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 Àfrica tenen més probabilitats de ser infelices.
- Tenir una GPD per càpita major afecta positivament la felicitat.
- El *social support* també té molt a veure amb un augment de la felicitat.
- 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 predir 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:

<https://github.com/gerardroseterron/PRAC2/blob/main/codi/PRACTICA2.R>

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.