

PRACTICA 1: REGRESSIÓ

Pol Espinasa Vilarrasa - 1566792

Marc Gonzalez Amores - 1564995

ÍNDEX

Introducció	3
Llibreries utilitzades	3
Explicació de la base de dades	4
Anàlisi numèric de cada atribut	8
Resultats test de Shapiro.....	14
Correlació entre dades	15
Regressió lineal.....	16
Error quadràtic mitjà (MSE)	22
Atribut escollit.....	23
Principal Component Analysis (PCA)	24
Conclusions	24

Introducció

En aquesta pràctica, tenint en compte els coneixements que hem anat adquirint en la assignatura, haurem d'aplicar aquests coneixements a un problema real. S'haurà d'analitzar una base de dades real utilitzant gràfiques i procediments matemàtics.

La nostra base de dades tracta sobre la esperança de vida dels països de tot el món separades per anys. La nostra base de dades està separada per tots els països i dels anys 1800 al 2016.

<https://www.kaggle.com/amarpandey/world-life-expectancy-18002016>

Llibreries utilitzades

Per portar a terme aquesta pràctica utilitzarem llibreries de Python per l'aprenentatge computacional i la IA, que ens facilita eines d'anàlisi i algorismes.

Les llibreries que hem utilitzat són:

- **NumPy**: és una biblioteca per al llenguatge de programació Python que dona suport per crear vectors i matrius grans multidimensionals, juntament amb una gran col·lecció de funcions matemàtiques d'alt nivell per operar-hi.
- **Pandas**: a Computació i Ciència de dades, pandas és una biblioteca de programari escrita com a extensió de NumPy per a manipulació i anàlisi de dades per al llenguatge de programació Python. En particular, ofereix estructures de dades i operacions per manipular taules numèriques i sèries temporals.
- **Matplotlib**: és una biblioteca per a la generació de gràfics a partir de dades contingudes en llistes o arrays en el llenguatge de programació Python i la seva extensió matemàtica NumPy. Proporciona una API, pylab, dissenyada per recordar la de MATLAB.
- **Sklearn**: és una biblioteca per a aprenentatge automàtic de programari lliure per al llenguatge de programació Python. Inclou diversos algorismes de classificació, regressió i anàlisi de grups
- **Scipy**: proporciona algorismes per a l'optimització, integració, interpolació, problemes de valors propis, equacions algebraïques, equacions diferencials, estadística i moltes altres classes de problemes.

Explicació de la base de dades

En aquesta secció analitzarem la nostra base de dades per tal d'entendre el problema. No treballarem amb un conjunt de dades sense sentit, sinó que darrera hi ha una base de dades real que hem d'analitzar per tal d'agafar els atributs més importants sense que siguin escollits de manera aleatòria.

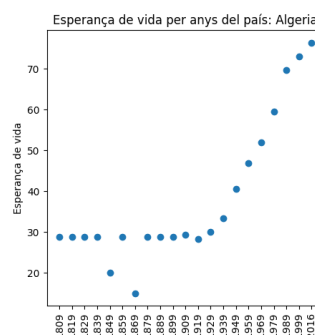
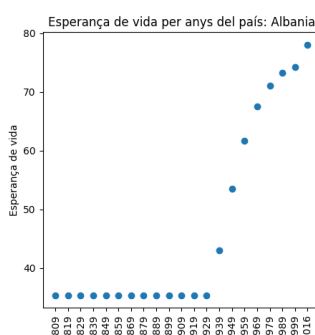
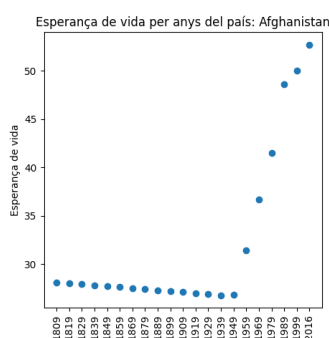
Com ja hem comentat anteriorment, la nostra base de dades tracta sobre la esperança de vida dels països de tot el món desde l'any 1800 al 2016. L'objectiu de treballar amb aquestes dades es intentar predir la esperança de vida al llarg d'aquest interval de temps. Tenim les dades organitzades en un dataset:

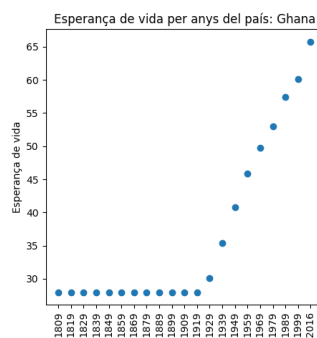
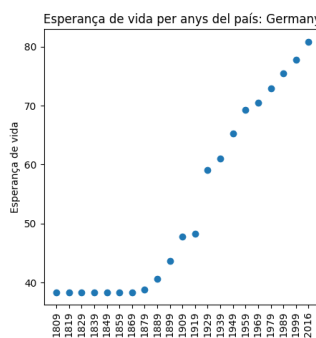
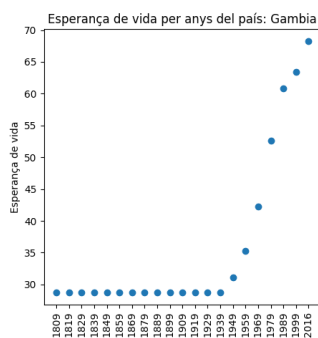
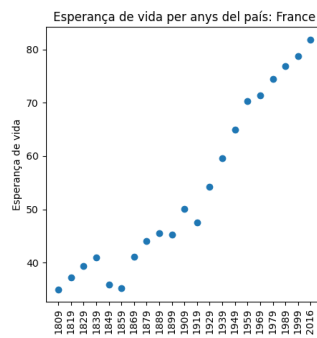
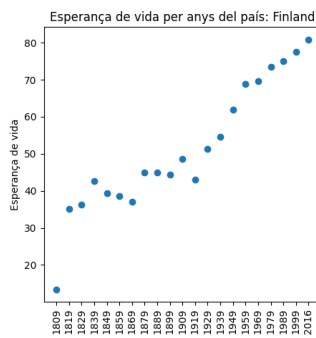
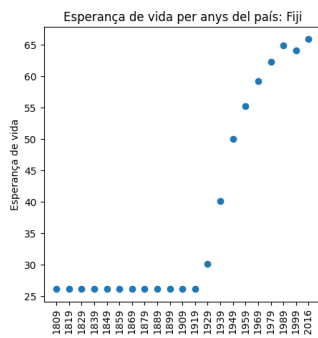
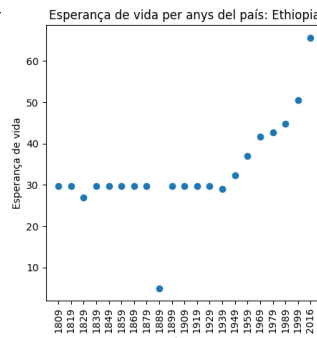
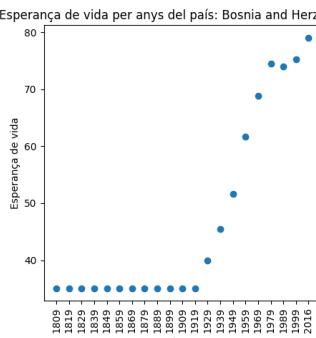
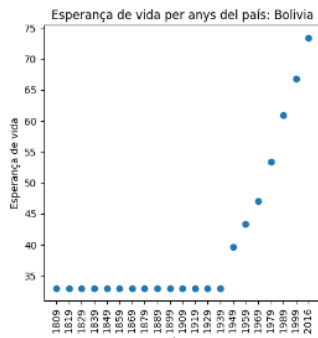
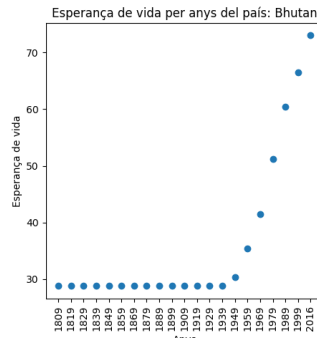
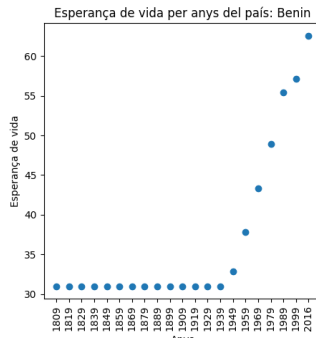
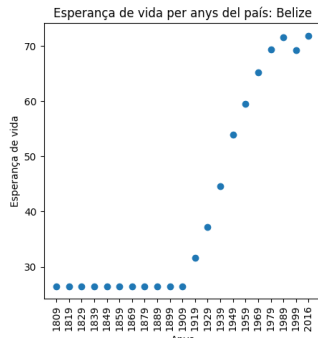
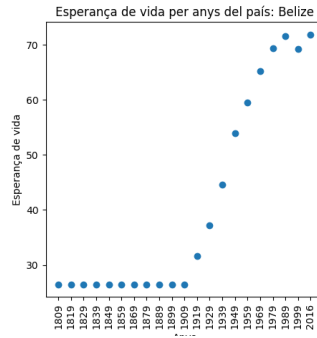
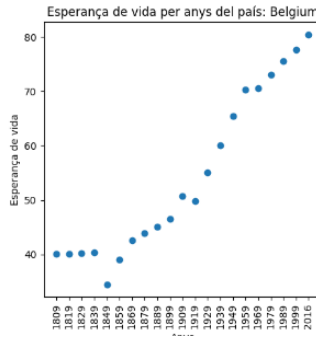
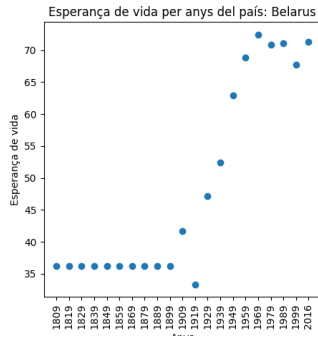
- **indicator-life_expectancy_at_birth.csv**

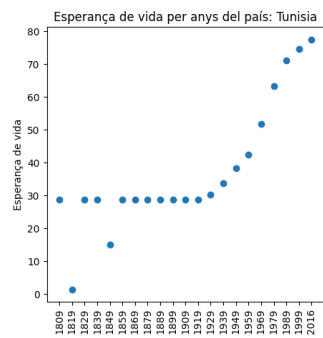
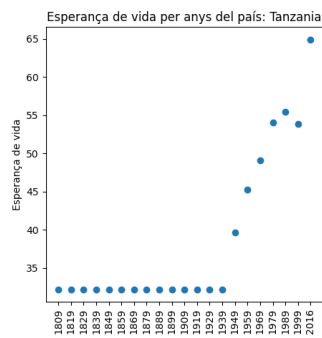
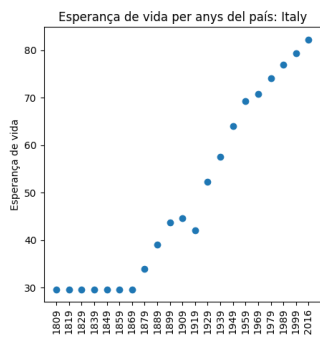
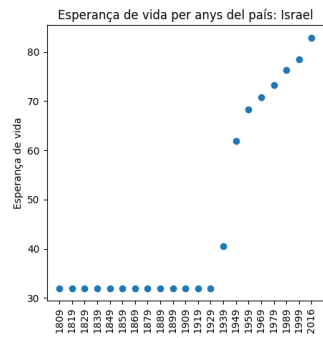
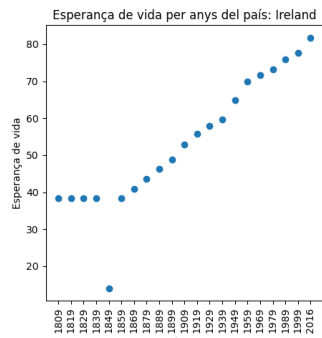
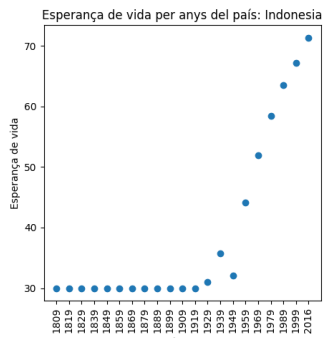
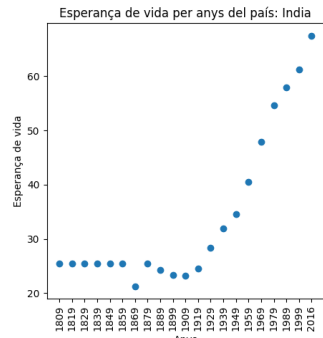
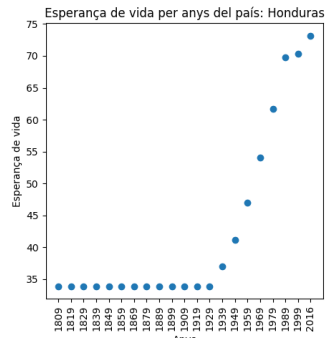
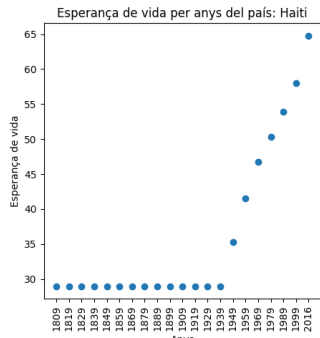
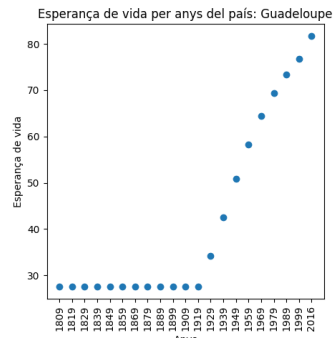
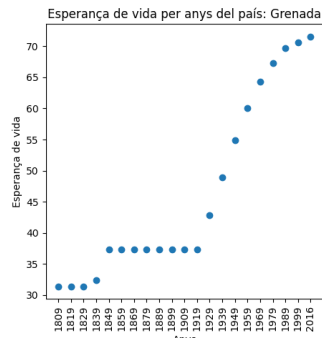
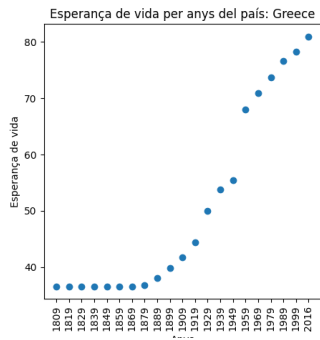
El primer problema que hem hagut d'afrontar era el fet que alguns països tenien dades buides, és a dir que hi ha països dels quals no tenim informació sobre les dades de la esperança de vida o altres que no tenim dades seves fins a algun any en concret. Per tal de poder tractar aquestes dades, hem eliminat aquells països que no tenien dades.

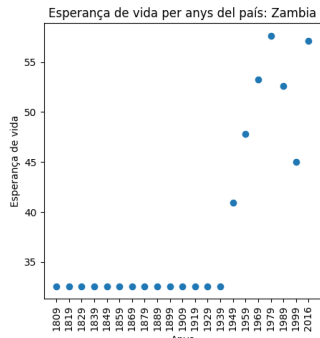
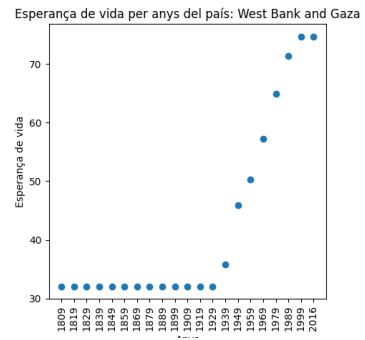
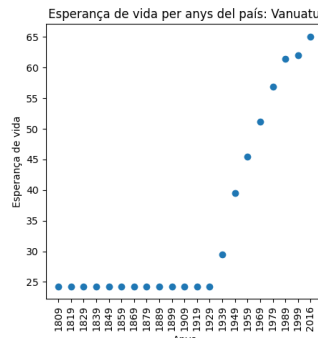
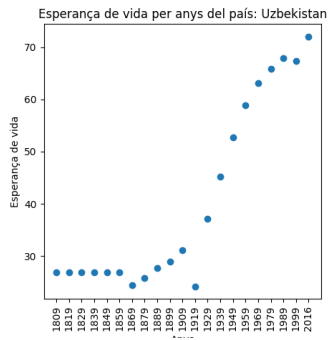
Hem analitzat per cada país les dades, i hem pogut veure com la esperança de vida en la gran majoria de països la esperança de vida a anat augmentant. Hem buscat informació a internet i hem pogut veure que el nostre anàlisi té sentit, ja que actualment hem evolucionat molt sobre la manera de tractar enfermetats cosa que fa en gran majoria que puguem viure més.

A continuació veurem diferents gràfics dels països des de el any 1969-2016, on podem veure aquesta evolució de la que estem parlant:







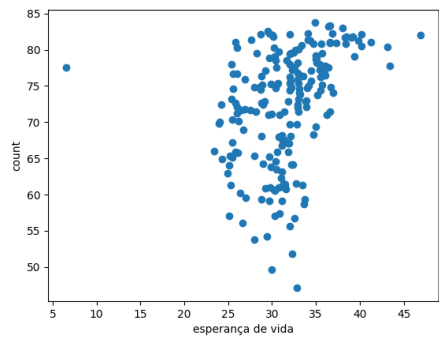
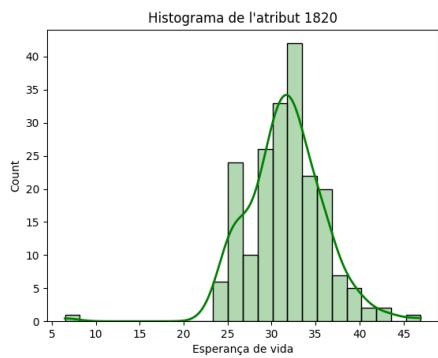
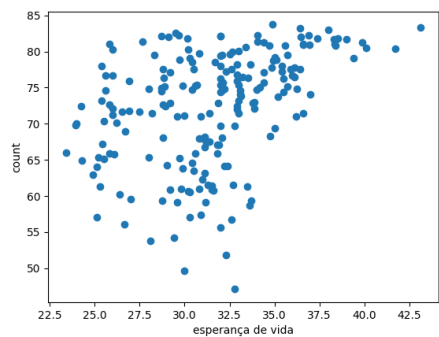
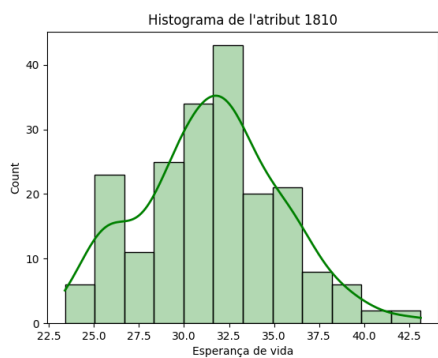
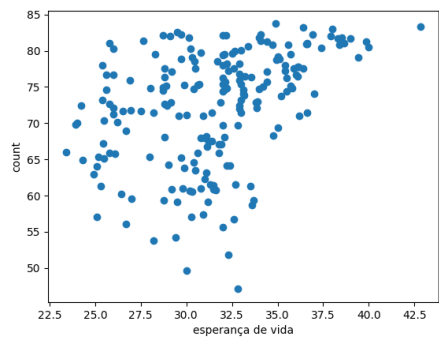
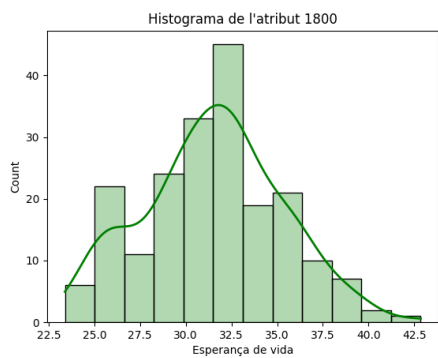


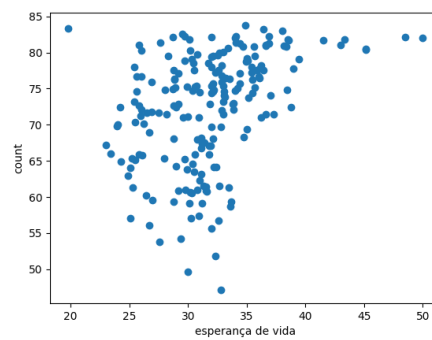
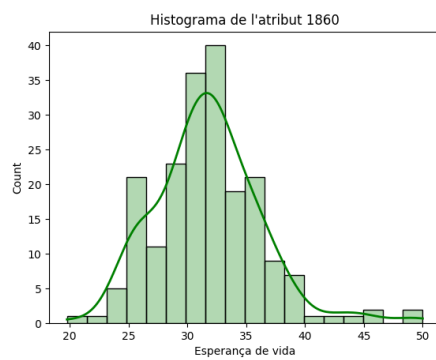
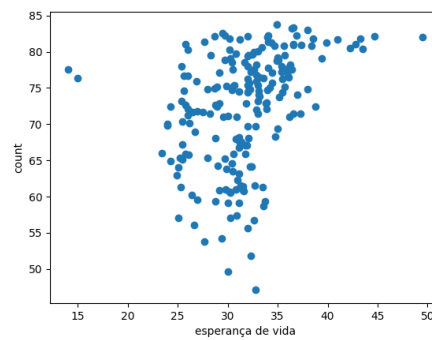
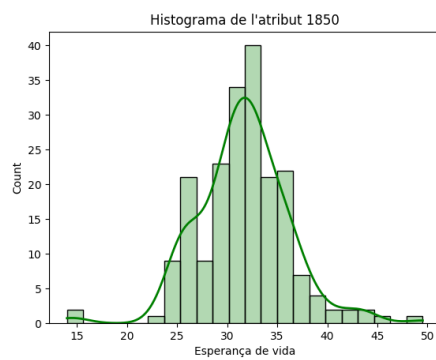
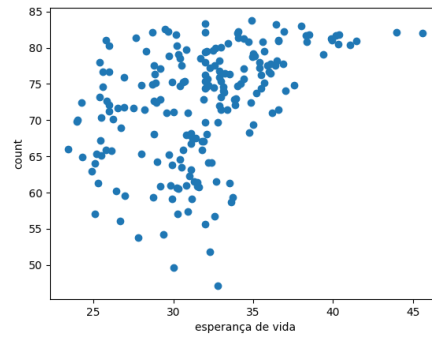
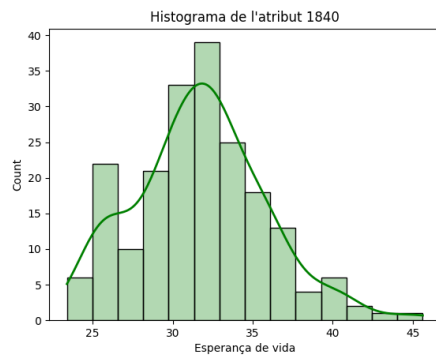
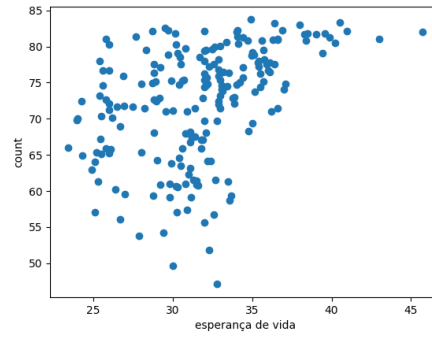
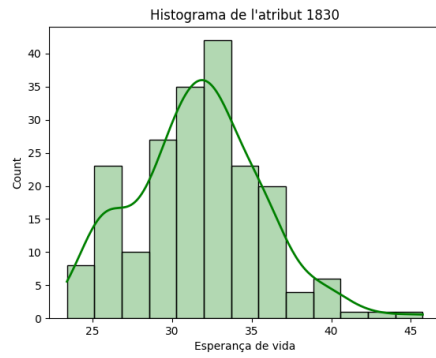
Com podem observar als diagrames de punts, podem veure la evolució de la esperança de vida tal i com hem comentat amb anterioritat, tot i que alguns països tenen més esperança de vida que d'altres. Els diagrames de punts dels altres països estan penjats al GitHub.

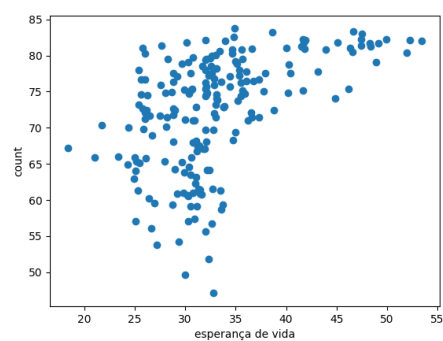
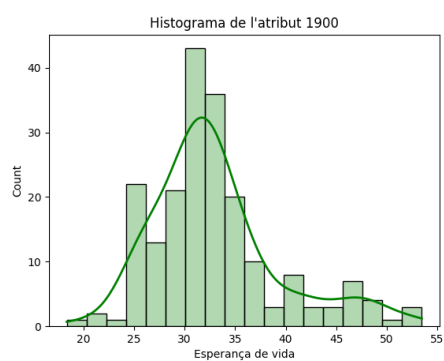
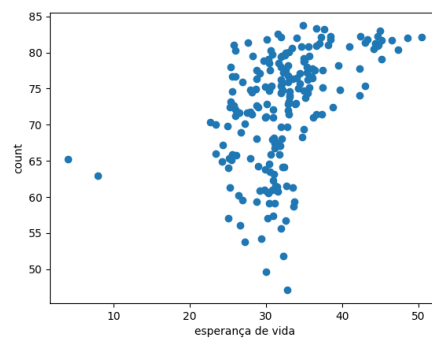
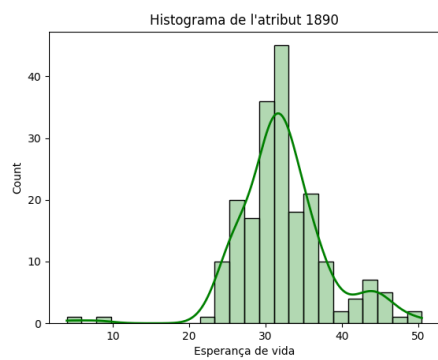
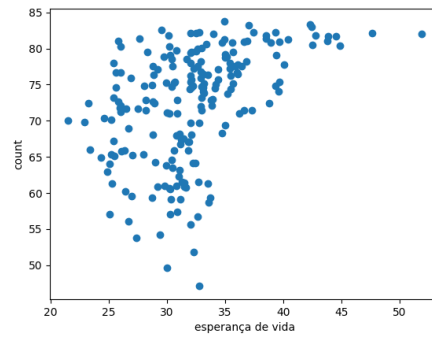
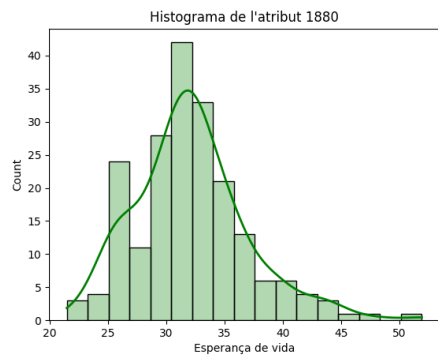
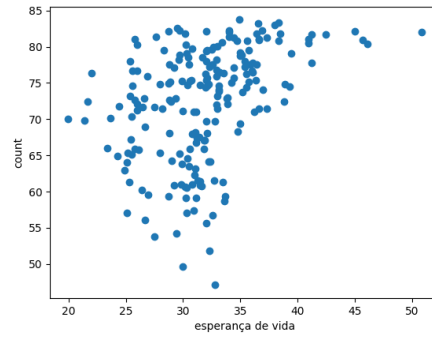
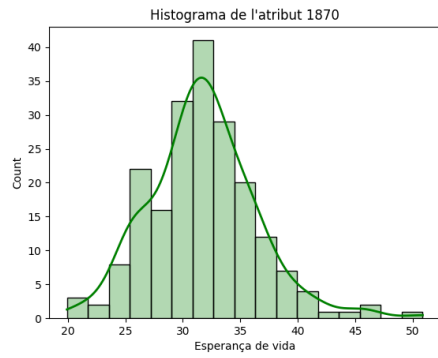
Per a cada any de la mostra hem dibuixat el seu histograma i la seva gràfica de punts per a veure la distribució que segueix la esperança de vida de cada any. També hem aplicat el test de Shapiro per a determinar quins anys de la mostra no segueixen una distribució normal.

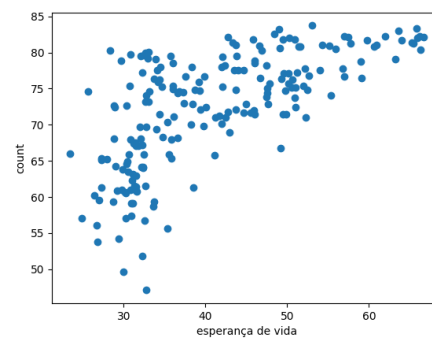
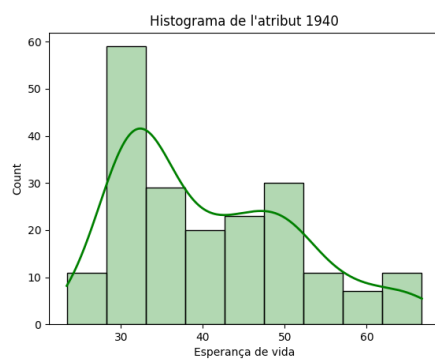
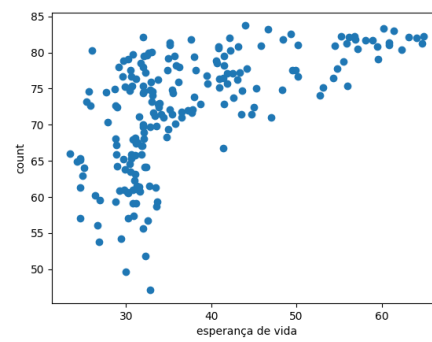
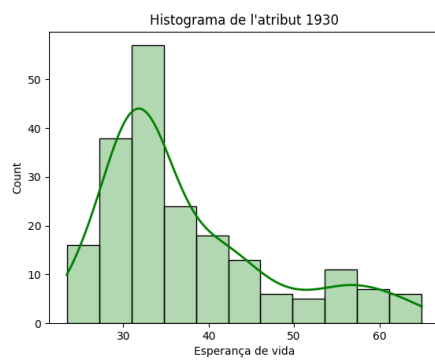
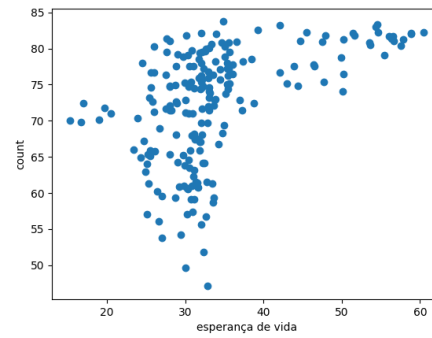
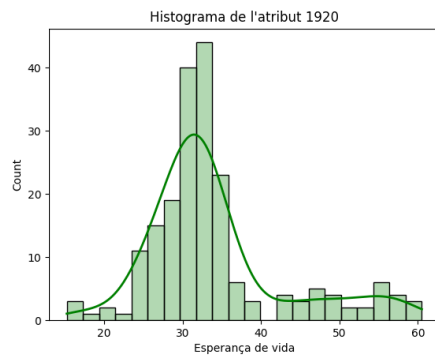
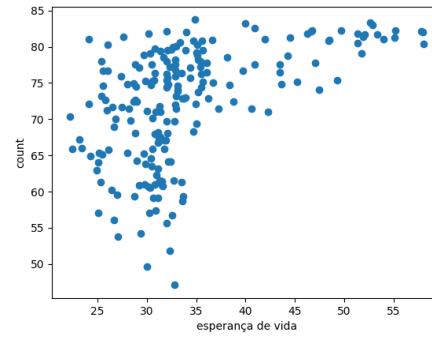
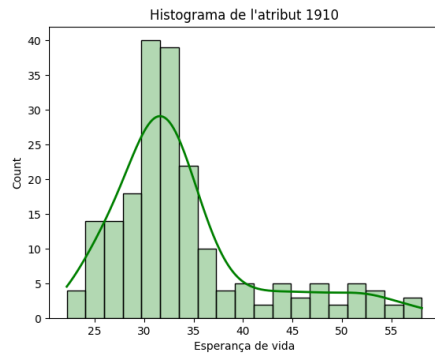
Anàlisi numèric de cada atribut

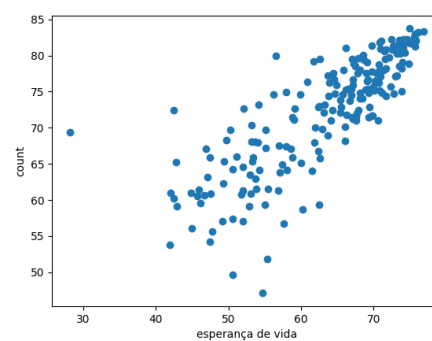
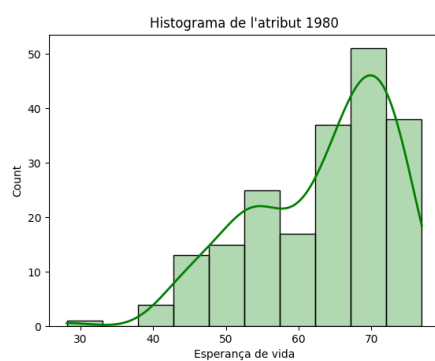
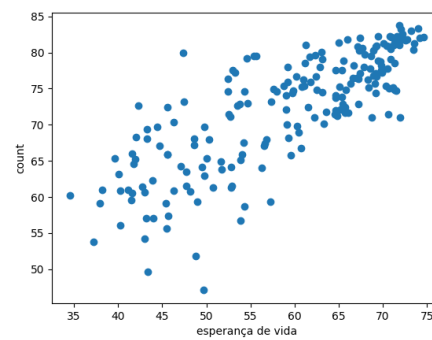
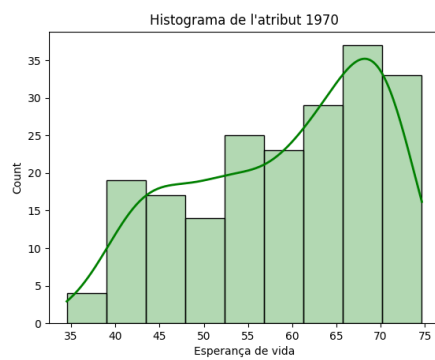
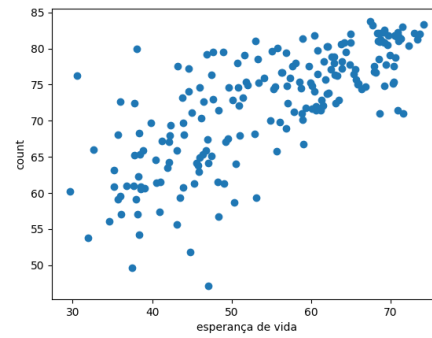
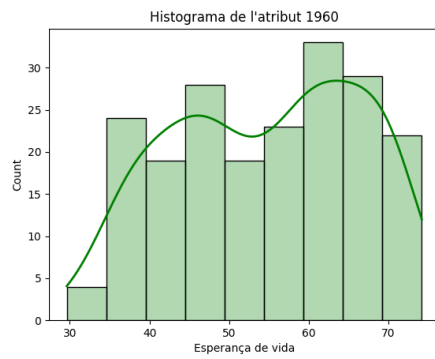
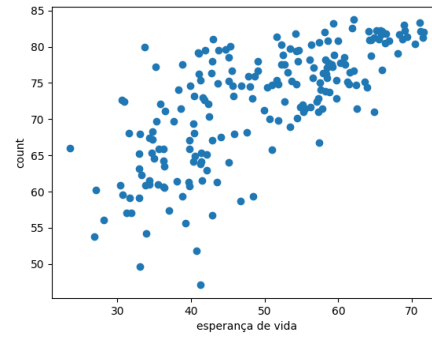
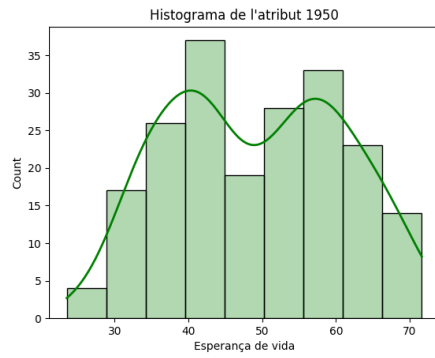
En el informe només mostrarem els anys de 10 en 10, ja que sinó l'informe seria molt extens:

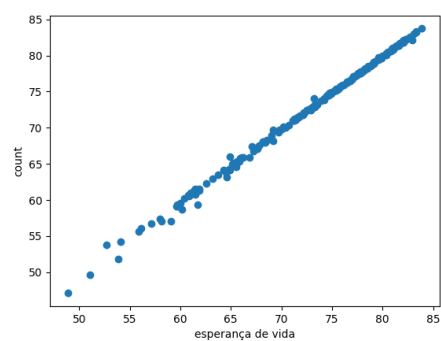
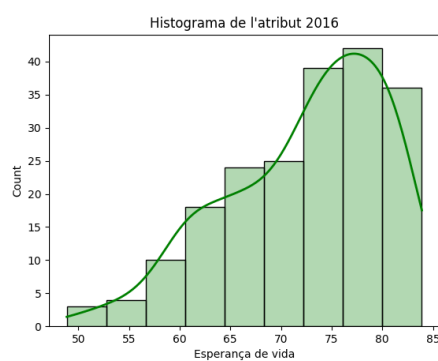
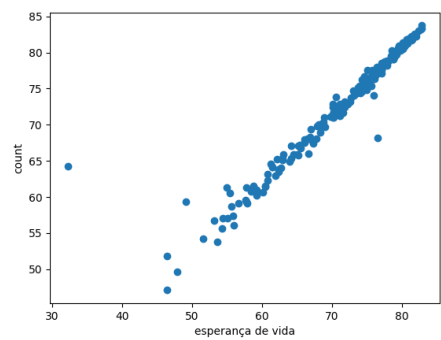
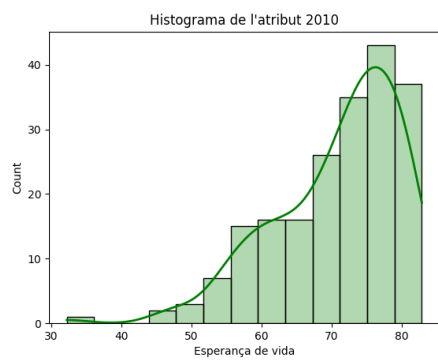
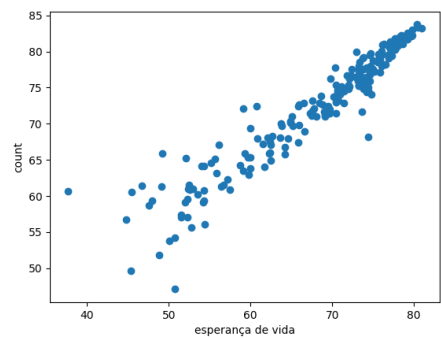
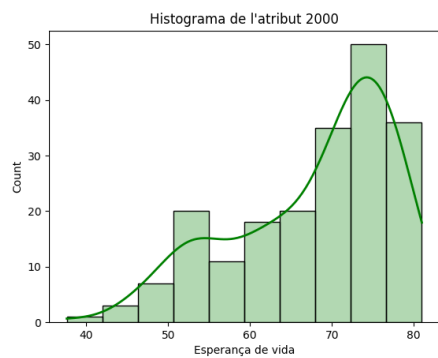
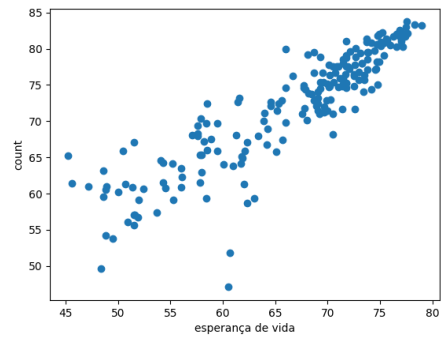
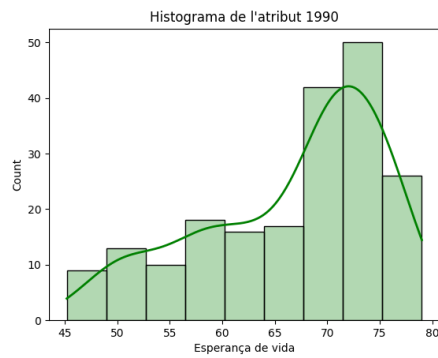












Pel que fa als diagrames de punts, podem observar que tots els països a mesura que van passant els anys augmenten la esperança de vida, tot i que hi ha alguns països que es queden al darrere en aquesta evolució, i augmenten en menor mesura la seva esperança de vida, tot i que la majoria com podem observar al diagrama del 2016 es troben en una esperança de vida major.

La regressió lineal es basa en tres bases; que la relació sigui de tipus lineal, que els residus segueixin una distribució normal i que la variància d'aquests residus sigui constant.

Quan les dades estan disperses, el regressor funciona millor. Tenint en compte que han de seguir una distribució normal, volem una dispersió elevada. Per tant hem de rebutjar tots aquells valors que no segueixin una distribució normal, aplicant el test de Shapiro.

El test de Shapiro rebutja aquells atributs que no segueixen una distribució normal.

Resultats test de Shapiro

Atribut 1 Any: 1800 --> Estadistic: 0.9887773394584656 P-Valor: 0.11583148688077927 No es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 2 Any: 1801 --> Estadistic: 0.9861128926277161 P-Valor: 0.04586431756615639 Es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 3 Any: 1802 --> Estadistic: 0.9872831106185913 P-Valor: 0.06893709301948547 No es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 4 Any: 1803 --> Estadistic: 0.9921090006828308 P-Valor: 0.34993278980255127 No es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 5 Any: 1804 --> Estadistic: 0.9880924820899963 P-Valor: 0.0913628488779068 No es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 6 Any: 1805 --> Estadistic: 0.983538806438446 P-Valor: 0.018868504092097282 Es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 7 Any: 1806 --> Estadistic: 0.9772395491600037 P-Valor: 0.0023758108727633953 Es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 8 Any: 1807 --> Estadistic: 0.9863826632499695 P-Valor: 0.050377920269966125 No es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 9 Any: 1808 --> Estadistic: 0.9765198230743408 P-Valor: 0.00189547601621598

Es pot descartar la hipòtesi de que les dades es distribueixen de forma normal
Atribut 10 Any: 1809 --> Estadistic: 0.97993004322052 P-Valor: 0.005640340968966484
Es pot descartar la hipòtesi de que les dades es distribueixen de forma normal

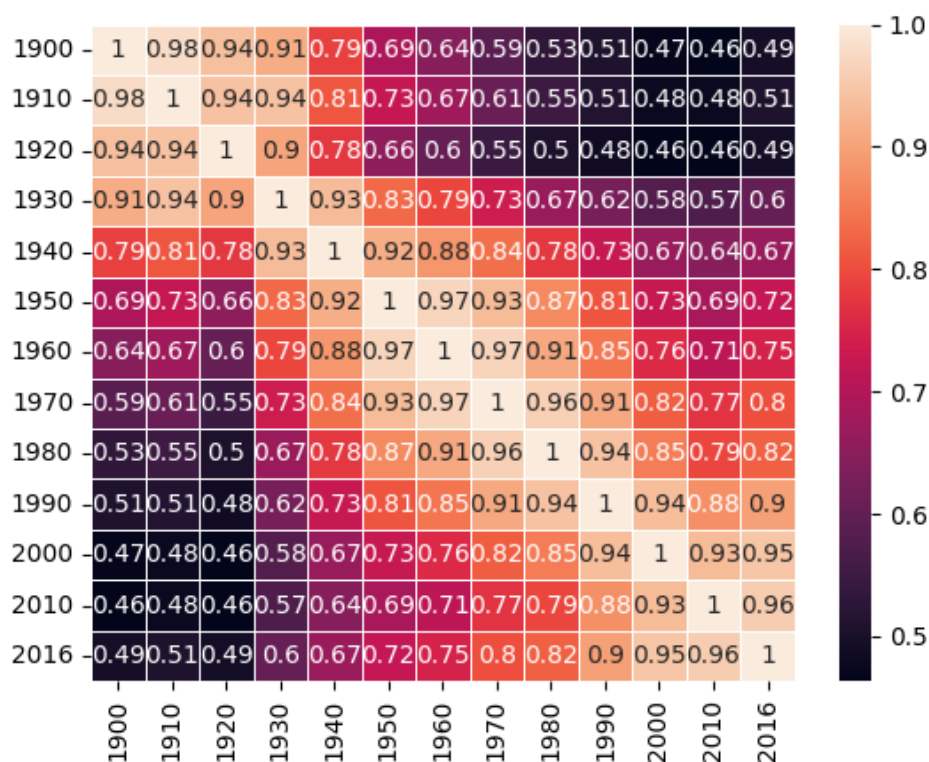
Hem posat d'exemple els resultats dels primers 10 anys, per tal de no allargar el informe, els altres resultats estan al GitHub.

Després de veure els resultats, els atributs que no podem rebutjar són: 1, 3, 4, 5, 8, 11, 12, 13, 14, 15, 34, 35, 38, 47, 48 que són els corresponents als anys 1800, 1802, 1803, 1804, 1807, 1810, 1811, 1812, 1813, 1814, 1833, 1834, 1837, 1846, 1847.

Ens interessen aquells atributs amb molta dispersió.

Correlació entre dades

Hem analitzat la correlació dels atributs, per poder detectar si estan relacionats amb l'any objectiu, el 2016. Podem observar-ho al mapa de calor:



En el mapa de calor podem observar que com més van avançant els anys, la correlació de dades augmenta, és a dir els valors s'assemblen més al atribut objectiu. El atribut objectiu que hem agafat és la esperança de vida dels països al 2016.

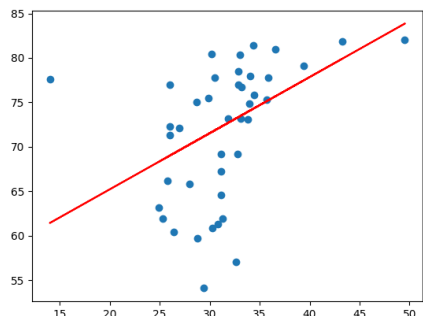
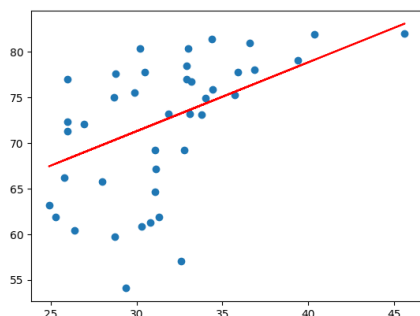
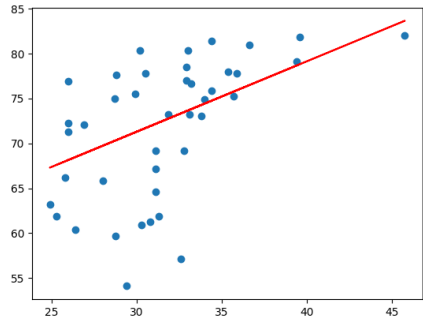
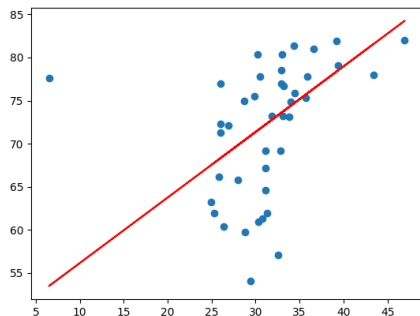
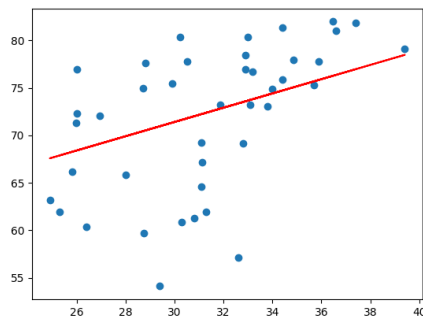
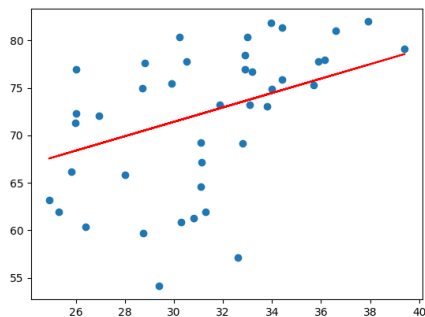
Regressió lineal

En aquest apartat farem la regressió lineal de cada atribut, en primer lloc sense normalitzar i després normalitzant-les. Tot i seguit calcularem l'error quadràtic mitjà del regressor, normalitzat i sense normalitzar.

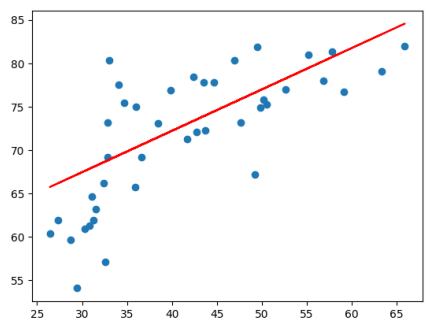
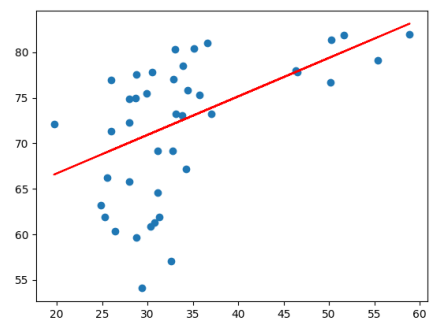
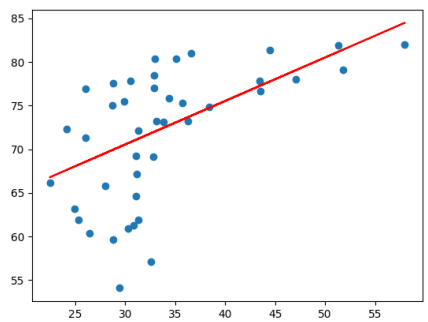
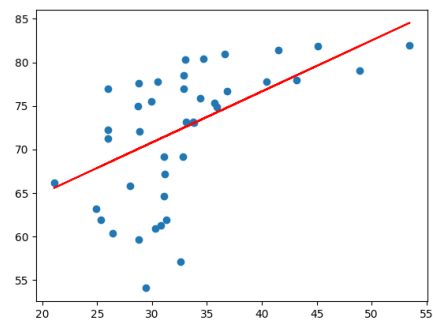
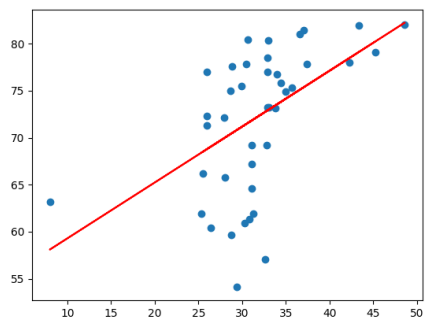
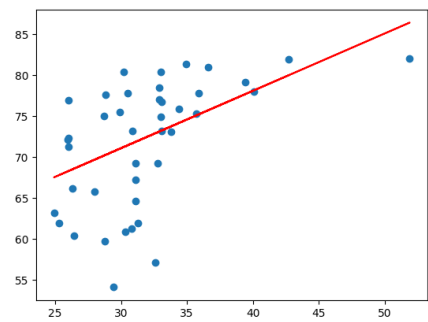
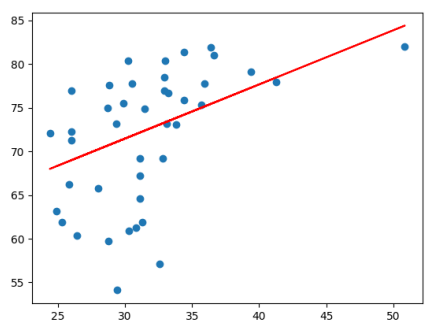
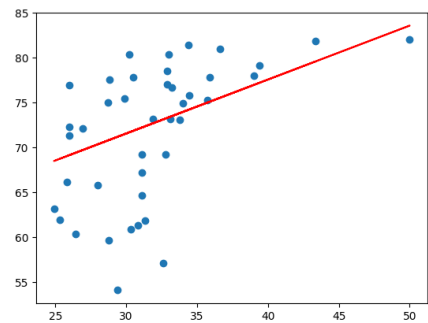
Regressió sense normalitzar

Per tal de no allargar el informe, mostrarem els anys de 10 en 10, la resta de gràfiques estan penjades al GitHub

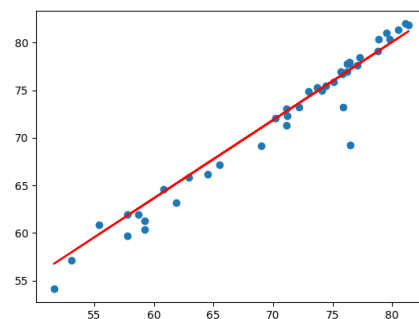
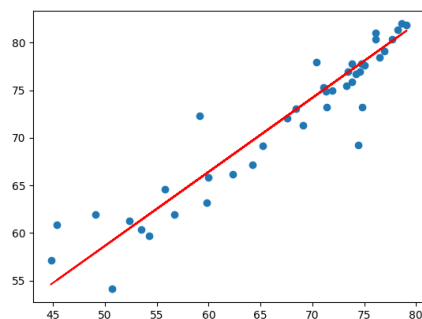
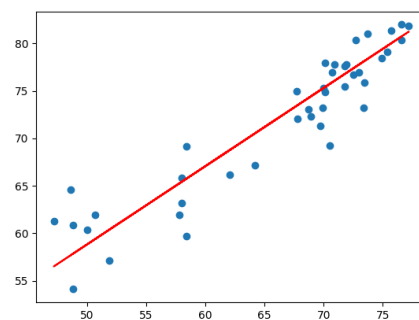
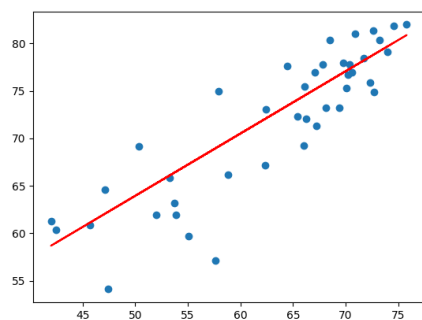
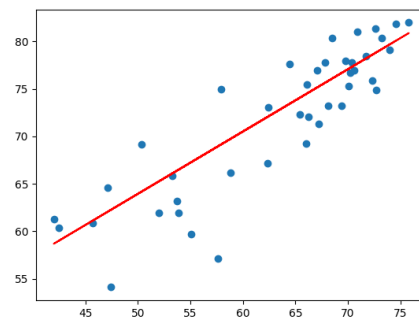
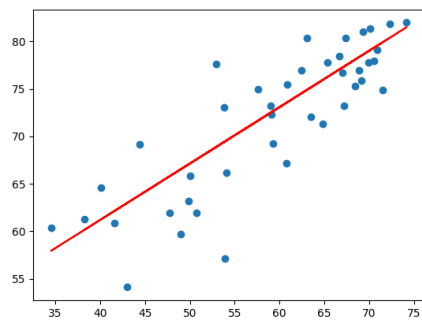
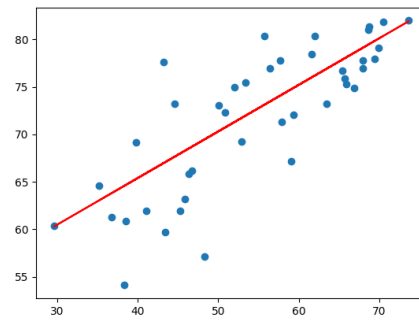
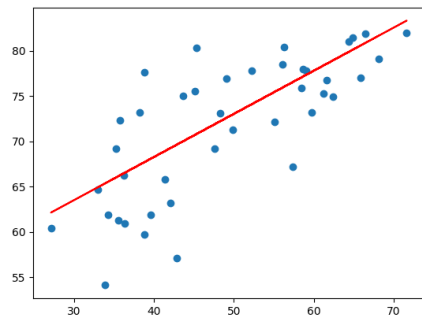
Anys 1800-1850



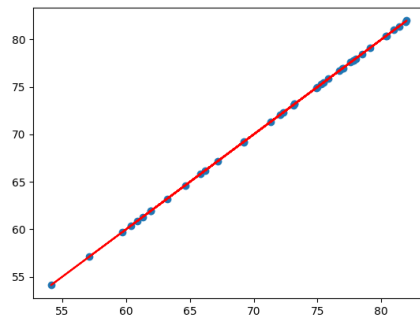
Anys 1860-1930



Anys 1940-2010



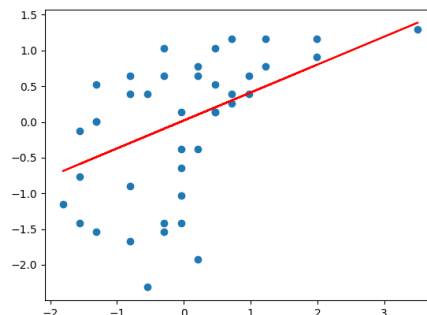
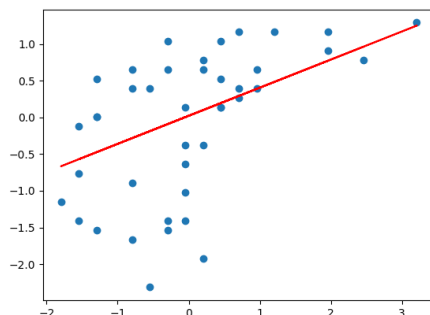
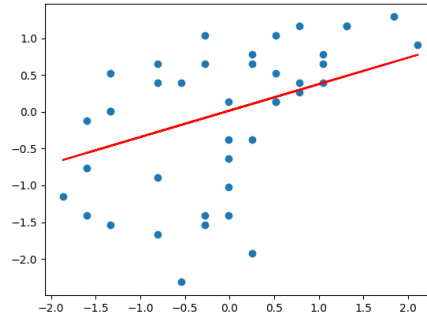
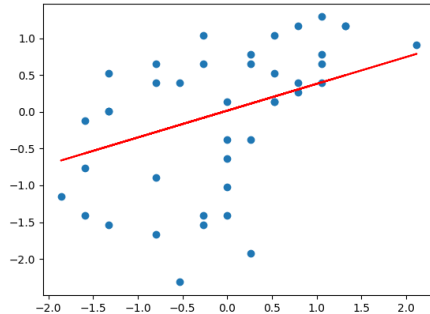
Any 2016



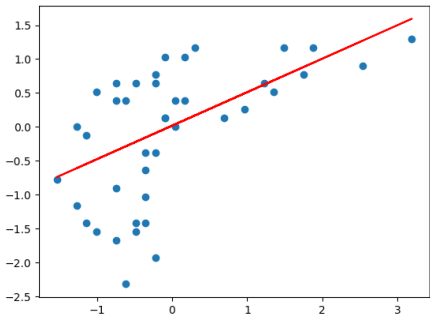
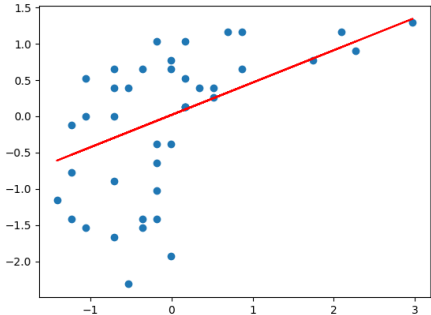
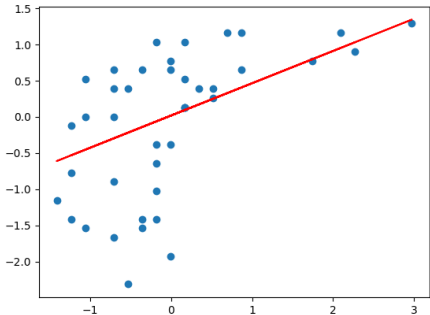
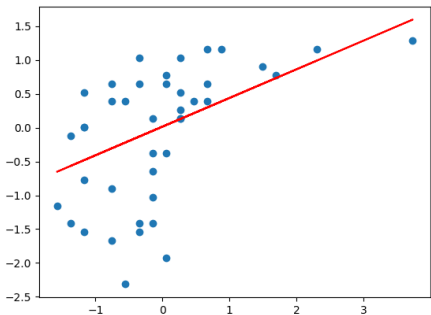
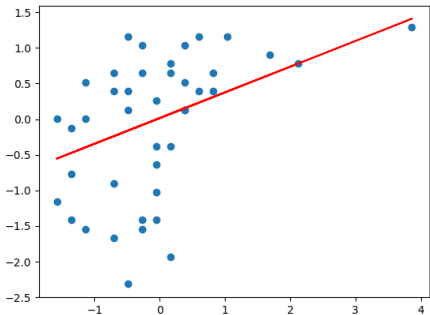
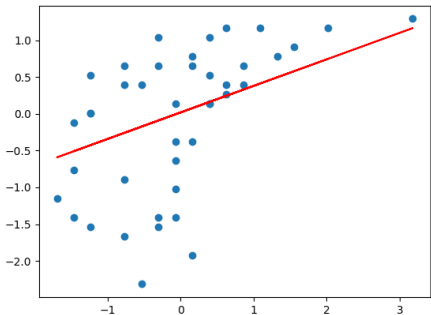
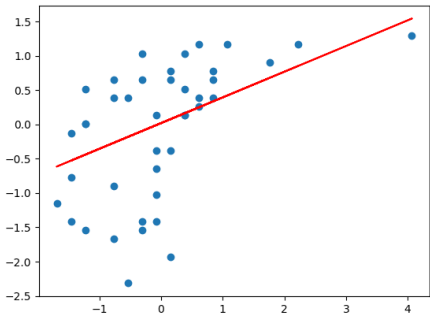
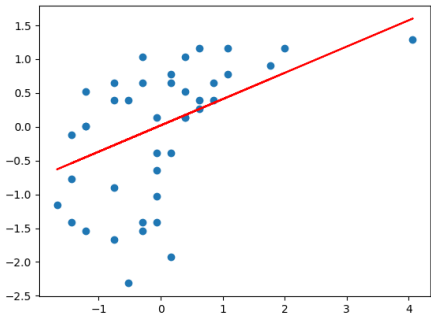
Amb els valors sense normalitzar podem veure com la recta de regressió cobreix en casi tota la majoria tots els punts. Això vol dir que per aquelles mostres que arribin al regressor, aquelles que estiguin més a prop de la recta de regressió el error serà més petit, mentre que aquelles que estiguin més lluny el error serà més gran.

Regressió amb dades normalitzades

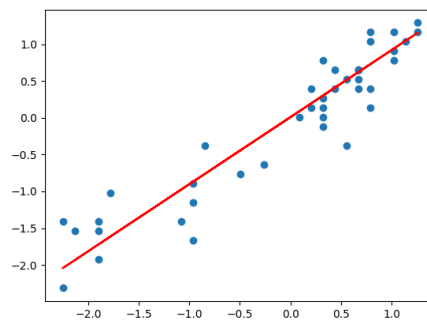
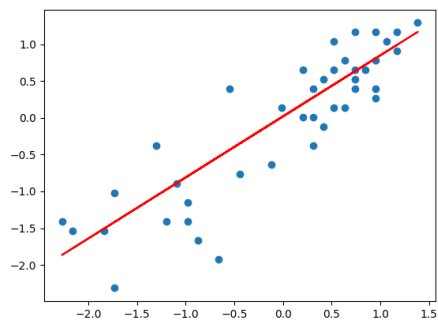
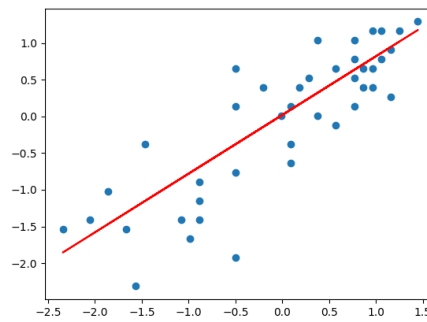
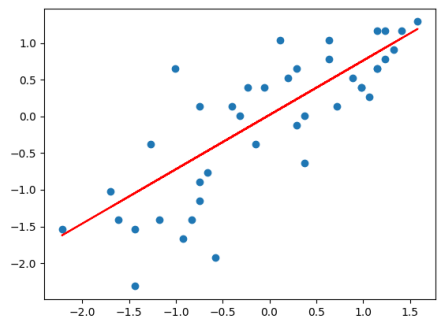
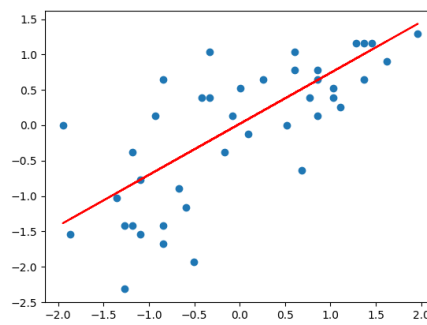
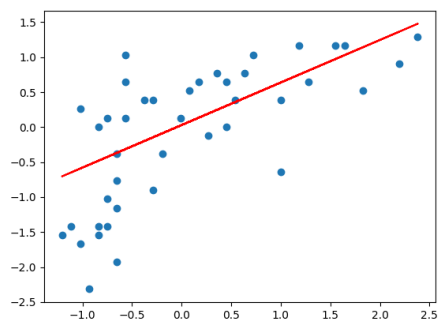
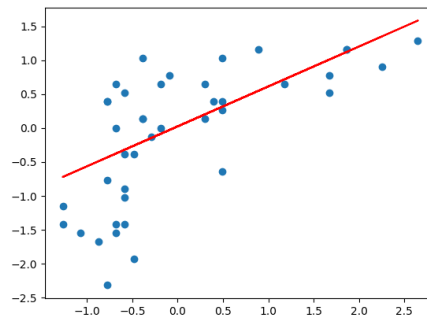
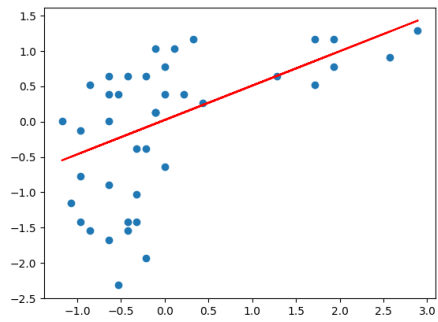
Anys 1800-1830



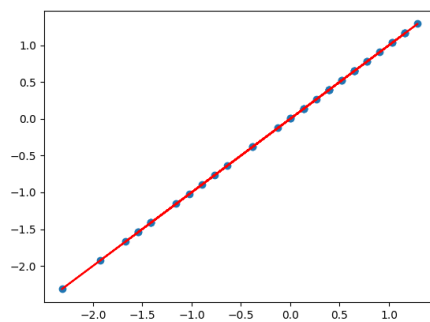
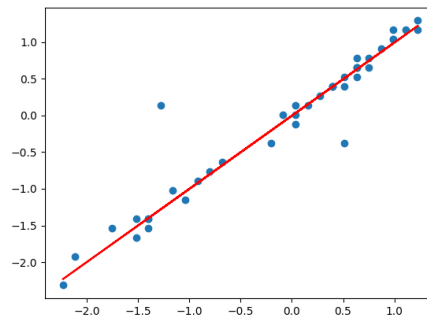
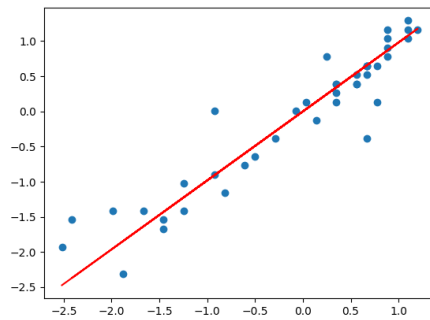
Anys 1840-1910



Anys 1920-1990



Anys 2000-2016



Observem que el fet de normalitzar les dades no fa canviar molt els gràfics.

Error quadràtic mitjà (MSE)

En aquesta secció es calcula l'error quadràtic mitjà del regressor per cada atribut, és a dir per cada any, per tal de detectar aquells atributs amb un MSE menor. Al informe només posarem resultats d'alguns atributs per no estendre l'informe.

Atributo 1 Any: 1800 :

MSE: 44.21246510698685 R2 score: -4.881399661218707

Atributo 11 Any: 1810 :

MSE: 43.84808616079501 R2 score: -4.922067014933038

Atributo 31 Any: 1830 :

MSE: 42.38370179335465 R2 score: -2.749826511153811

Atributo 51 Any: 1850 :

MSE: 48.862189561219125 R2 score: -3.065682480954413

Atributo 71 Any: 1870 :
MSE: 44.680822534197446 R2 score: -3.809068722300725
Atributo 91 Any: 1890 :
MSE: 41.91354868870076 R2 score: -1.8771824915583322
Atributo 111 Any: 1910 :
MSE: 39.44282456737262 R2 score: -1.6049298436703499
Atributo 131 Any: 1930 :
MSE: 33.478544124918194 R2 score: -0.5556762414901519
Atributo 151 Any: 1950 :
MSE: 25.946506573853956 R2 score: 0.18085976561084494
Atributo 171 Any: 1970 :
MSE: 17.467762443140433 R2 score: 0.5597816668939357
Atributo 191 Any: 1990 :
MSE: 8.129141058466812 R2 score: 0.8583021403391822
Atributo 201 Any: 2000 :
MSE: 7.062119172191595 R2 score: 0.8811598725572126
Atributo 217 Any: 2016 :
MSE: 3.07848158135029e-29 R2 score: 1.0

Hem observat que l'error quadràtic va disminuint quan més s'apropa al any objectiu, el 2016.

Atribut escollit

Després d'aplicar d'analitzar els mapes de calor i el MSE de cada atribut, els millors atributs que podem escollir són aquells que es troben més a prop de l'any que volem predir, el 2016. Quan més ens apropem a aquest any, les diferències son poc notòries.

Principal Component Analysis (PCA)

Ja que la nostra base de dades té molts atributs (anys des del 1800-2016), podríem aplicar un PCA per reduir la dimensió del espai observable. És a dir, si es redueix el nombre de variables a dues o tres de noves, es poden representar les dades originals en el pla o en un gràfic de 3-dimensions i, així, es visualitza de manera gràfica un resum de les nostres dades.

Conclusions

Gràcies a aquesta pràctica hem pogut entendre el funcionament del Machine Learning, utilitzant tècniques que havíem vist a classe, i com aplicar-les per solucionar un problema real. També hem après a analitzar aquests resultats.