

Anàlisi de dades i regressió

GRUP: GPA304-1030

Víctor Bosch Pueyo 1566583

Álvaro Caravaca Hernández 1566685

Juan Carlos Martínez Moreno 1566936

Contenido

Introducció	3
Explicació de la base de dades.....	3
Anàlisi numèric de cada atribut.....	5
Correlació entre dades.....	12
Regressió lineal.....	12
Error quadràtic.....	14
Atribut escollit	16
El descens de gradient	18
Conclusions.....	21

Introducció

En aquesta pràctica haurem d'aplicar els coneixements obtinguts a classe sobre aplicant-los a un problema real. Haurem d'analitzar els atributs d'una base de dades real a partir de diferents processos matemàtics i representacions gràfiques, per tal d'escollir els més representatius.

Un cop fet això, els normalitzarem per poder avaluar l'error del model i visualitzar les dades. Aplicarem també el descens de gradient als atributs més representatius, per ser capaços de fer prediccions del resultat a partir de valors nous.

Tot l'esmentat anteriorment l'haurem d'implementar aplicant models de regressió.

La nostra base de dades tracta de diferents mesures del producte interior brut en diversos sectors determinats de l'Índia. Aquestes dades estan organitzades per quadrimestres dels anys 2005 al 2016. Totes les dades que representen el producte interior brut estan expressats en bilions de rupies índies.

<https://www.kaggle.com/navoneel/fta-data>

Llibreries utilitzades

Per tal de realitzar la practica amb la màxima comoditat i eficiència possible, farem us d'algunes llibreries de Python especials per el aprenentatge computacional i la IA que ens donen moltes eines matemàtiques, d'anàlisi i algorismes.

Les principals llibreries utilitzades son:

- **Numpy**: llibreria que incorpora funcions matemàtiques de alt rendiment amb les que podem fer operacions de forma molt ràpida i eficient (computacionalment).
- **Pandas**: llibreria que incorpora estructures de dades que ens faciliten l'emmagatzematge de taules i la seva manipulació. A més, inclou algunes funcions d'anàlisi matemàtic també a considerar.
- **Matplotlib i seaborn**: dos llibreries especialitzades en gràfics, taules i anàlisi de variables.
- **Sklearn**: la llibreria d'aprenentatge computacional que incorpora tots els algorismes que hem utilitzat: regressions, normalitzacions, etc...
- **Scipy**: llibreria matemàtica que ens ha ajudat amb el test de distribució normal i algunes funcions matemàtiques, també.

Explicació de la base de dades

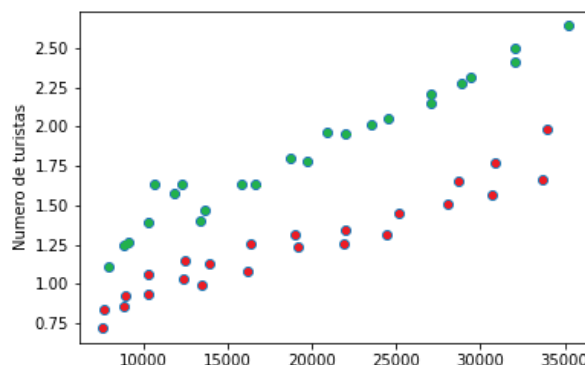
Aquest apartat el dedicarem a analitzar la base de dades que tenim entre mans per a entendre el problema al que ens enfrontem. És important saber que no treballem amb un conjunt de números sense sentit, sinó que el que tenim és una base de dades real que hem d'entendre per a no agafar com a atribut més important un que, per molt que hi hagi una correlació forta entre les dades, podria ser pura coincidència.

Com ja hem dit a l'apartat d'introducció, la nostra base de dades ens mostra el producte interior brut en diferents sectors de 2005 fins a 2016, dividit en quadrimestres. L'objectiu de treballar amb aquestes dades es poder aconseguir predir és el número de turistes que arriben a l'Índia. Els trobem en 4 sub-datasets de la següent manera:

- **q1.csv** : corresponent al 1r quadrimestre de 2005 fins a 2016
- **q2.csv** : corresponent al 2n quadrimestre de 2005 fins a 2016
- **q3.csv** : corresponent al 3r quadrimestre de 2005 fins a 2016
- **q4.csv** : corresponent al 4t quadrimestre de 2005 fins a 2016

Hem seguit les indicacions donades pel professor de treballar amb totes les dades juntes a l'hora de fer el descens de gradient i la predicció. Ens hem trobat amb el problema que les dades dels diferents quadrimestres varien dins un llinar, cosa totalment normal ja que en el primer i en l'últim, el número de turistes que arriben és molt més elevat que en els altres dos. Les dades d'aquests dos trimestres amb valors més alts comencen el primer d'Octubre fins el 31 de Març. Buscant informació per internet hem descobert que les dates més recomanades per a visitar l'Índia, per a raons climatològiques, són al nadal. Podem dir llavors que les nostres dades segurament siguin correctes. Vist això, hem decidit seguir amb la pràctica com ens va indicar el professor però, dedicarem un apartat a fer la regressió per a un sol quadrimestre.

En el següent gràfic de punts podem observar aquesta variància entre el primer i quart quadrimestre (marcats en verd) i el segon i el tercer (marcats en vermell).



Els sectors sobre els que treballem són els següents, expressats en bilions de rupies índies:

- PIB als preus de mercat (enfoc des de la producció)
- Valor afegit brut a preus bàsics (activitat total)
- Agricultura, silvicultura i pesca
- La indústria (inclosa l'energia)
- Fabricació
- Construcció
- Serveis
- Transport, allotjament i activitat de serveis alimentaris
- Activitats immobiliàries
- Administracions públiques (seguretat social, educació, etc.)

Tots aquests valors els veiem representats 4 cops segons el tipus de mesura en el que s'ha fet; aquestes són:

- **CQRSA:** Moneda nacional, preus corrents, nivells trimestrals, ajustos estacionals
- **CQR:** Moneda nacional, preus corrents, nivells trimestrals
- **VNBQRSA:** Moneda nacional, preus constants, any base nacional, nivells trimestrals, ajustos estacionals
- **VNBQR:** Moneda nacional, preus constants, any base nacional, nivells trimestrals

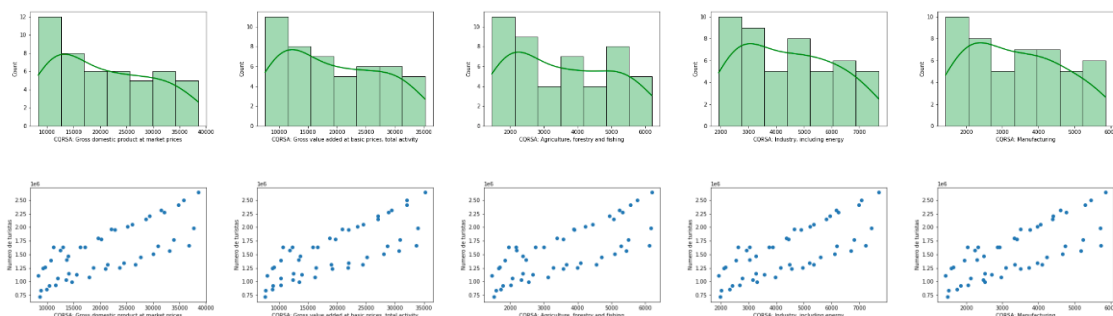
També trobem un atribut que són els guanys per canvi de divisa. Aquest atribut no el tindrem en compte, tot hi que es el que més relació té amb els turistes que arriben. Òbviament, com més turistes vinguin, més canvis de divisa hi haurà però en la vida real no ens ajudaria a predir el nombre de turistes. Podem dir que aquest atribut depèn dels turistes (no a la inversa) i no ens ajudarà a predir-ho. Com a exemple: si la nostra base de dades fos de predicció de les patates que recollirem d'un camp, es obvi que com més en plantem, més en sortiran. Per a fer una millor predicció escolliríem altres valors com: nivell d'acidesa de l'aigua, el nivell de substrat a la terra, etc.

Per cada atribut hem dibuixat el seu histograma i la seva gràfica de punts per a veure quina distribució segueixen. També hem aplicat el test de Shapiro per a determinar quines variables no segueixen una distribució normal.

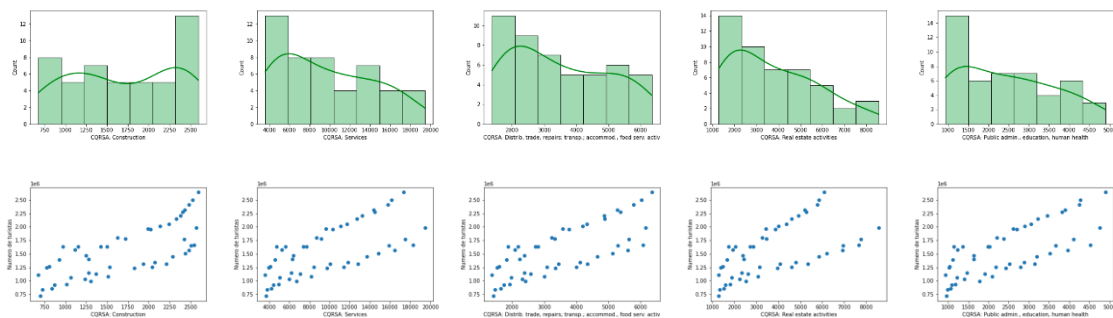
Anàlisi numèric de cada atribut

Aquests han estat tots els histogrames i diagrames de punts per a tots els atributs de la base de dades.

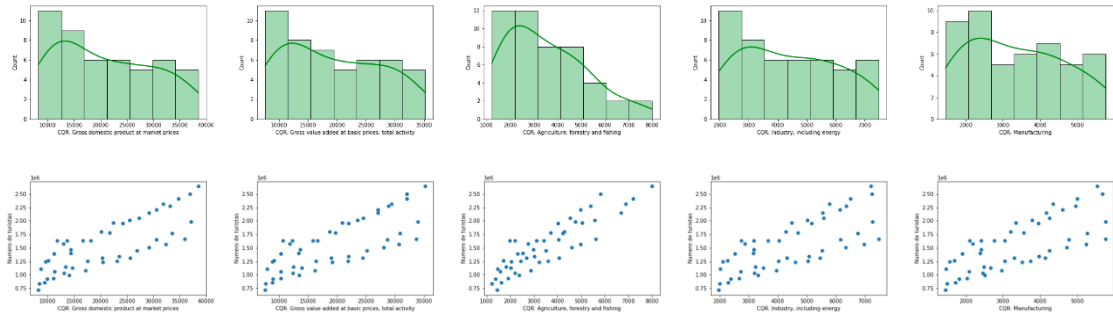
Atributs 1-5:



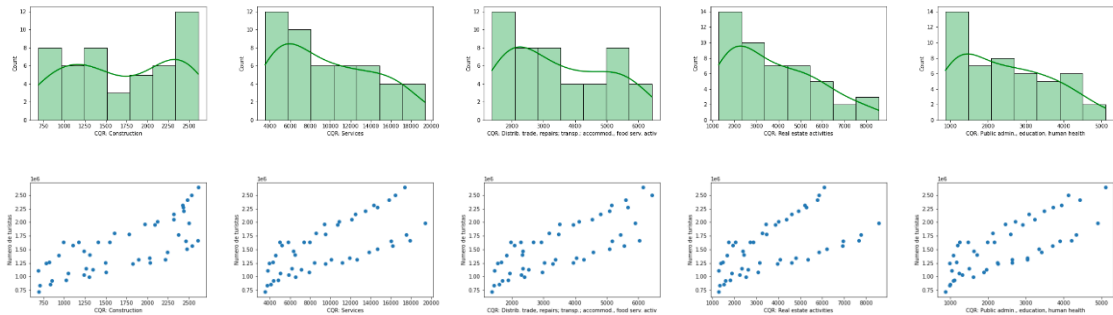
Atributs 6-10:



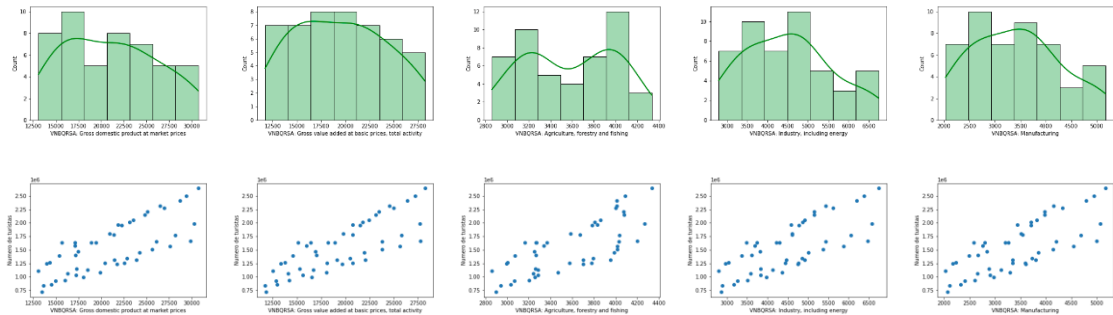
Atributs 11-15:



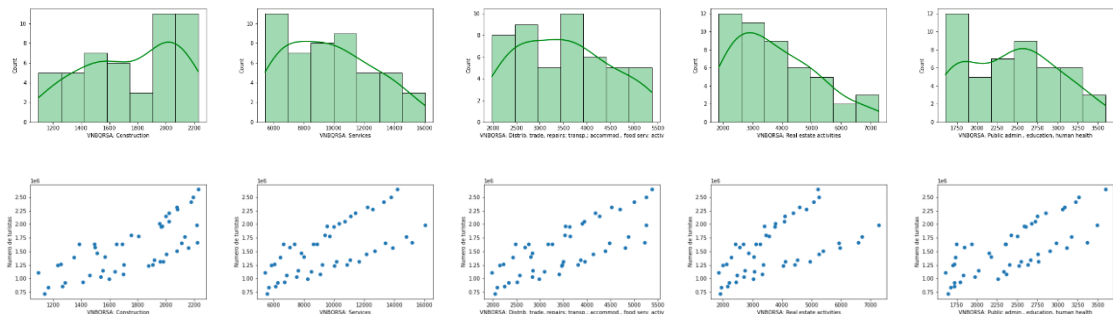
Atributs 16-20:



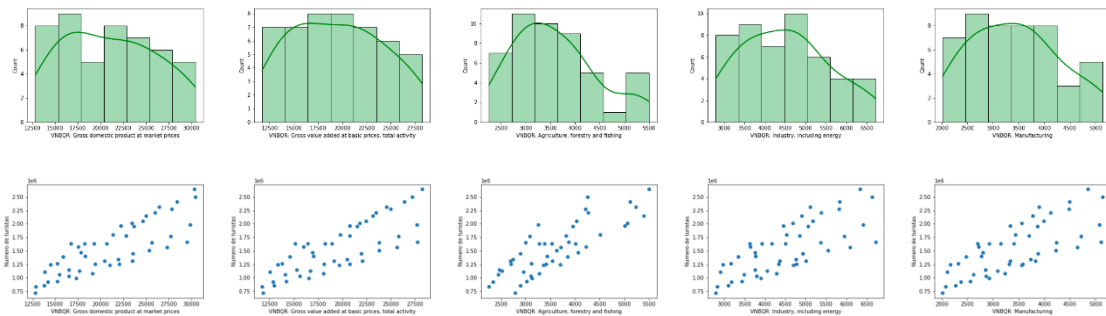
Atributs 21-25:



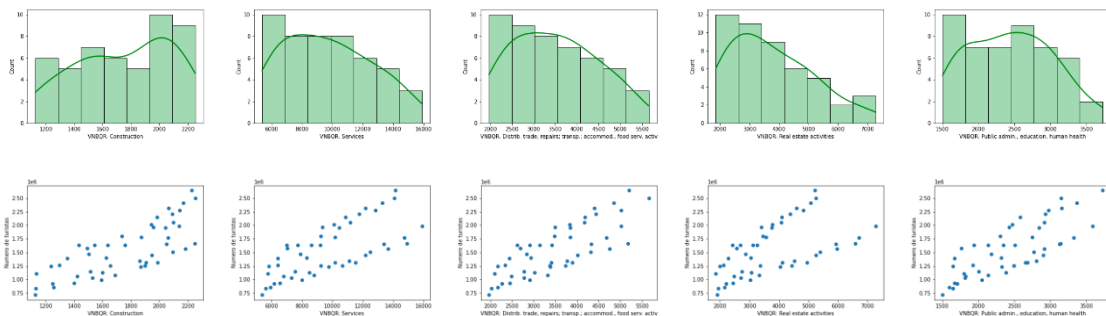
Atributs 26-30:



Atributs 31-35:



Atributs 36-40:



Quant als diagrames de punts, podem observar que tots els atributs tendeixen a augmentar el seu valor a mesura que augmenta la y . És a dir, que a mesura que passen els anys el valor del producte interior brut augmenta.

La regressió lineal assumeix tres qüestions importants: que la relació és de tipus lineal, que els residus segueixen una distribució normal i la variància d'aquests residus és constant.

El regressor funciona millor quan les dades estan disperses. Volem una dispersió elevada però tenint en compte que han de seguir una distribució normal.

Per tant, haurem de rebutjar tots aquells atributs que no segueixen una distribució normal. Per fer això, hem aplicat el test de Shapiro.

Aquest test planteja la hipòtesi nul·la que una mostra prové d'una distribució normal. Escollim un nivell de confiança (0.05) i tenim la hipòtesi alternativa que sosté que la distribució no és normal. El test de Shapiro intenta rebutjar la hipòtesi nul·la al nostre nivell de confiança. Per tant, rebutja aquells atributs que no segueixen una distribució normal.

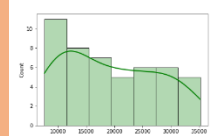
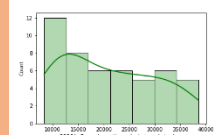
RESULTATS DEL TEST DE SHAPIRO


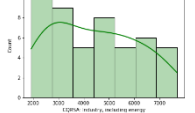

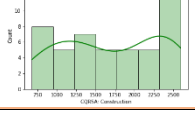
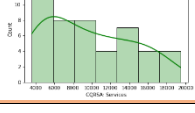
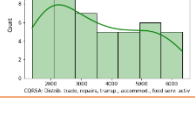
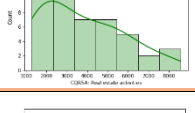
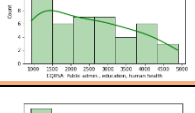
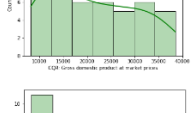
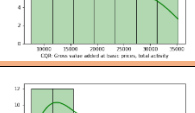
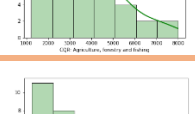
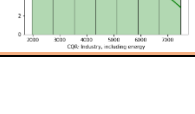
Atributo 1 : Estadístico: 0.931 | P-Valor: 0.007

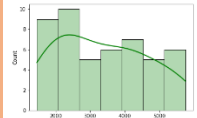
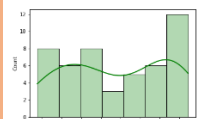
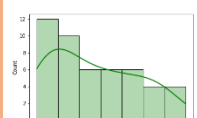
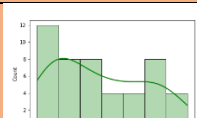
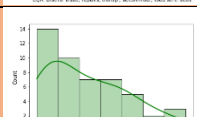
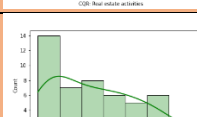
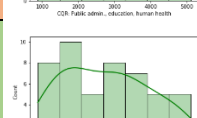
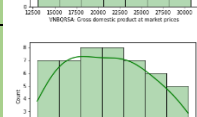
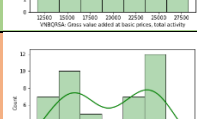
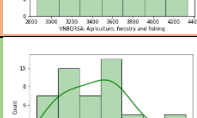
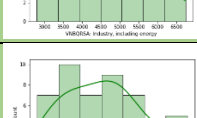
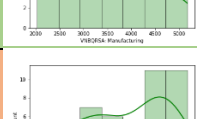
Se puede rechazar la hipotesis de que los datos de distribuyen de forma normal


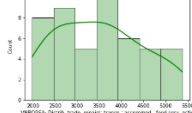
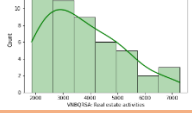
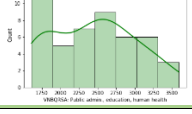
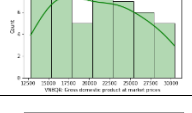
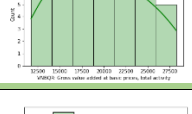
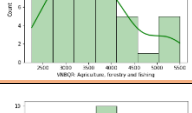
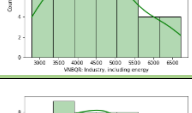
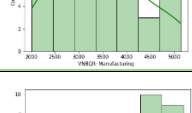
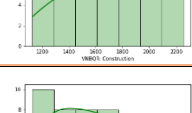
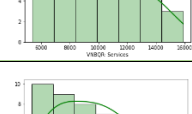
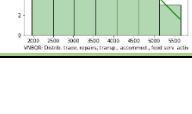
Atributo 2 : Estadístico: 0.932 | P-Valor: 0.008

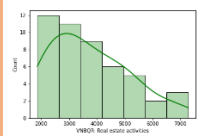
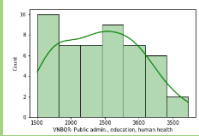
Se puede rechazar la hipotesis de que los datos de distribuyen de forma normal



<p>Atributo 3 : Estadístico: 0.928 P-Valor: 0.006</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 4 : Estadístico: 0.949 P-Valor: 0.035</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 5 : Estadístico: 0.950 P-Valor: 0.039</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 6 : Estadístico: 0.910 P-Valor: 0.001</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 7 : Estadístico: 0.931 P-Valor: 0.007</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 8 : Estadístico: 0.926 P-Valor: 0.005</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 9 : Estadístico: 0.929 P-Valor: 0.007</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 10 : Estadístico: 0.930 P-Valor: 0.007</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 11 : Estadístico: 0.930 P-Valor: 0.007</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 12 : Estadístico: 0.932 P-Valor: 0.008</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 13 : Estadístico: 0.939 P-Valor: 0.015</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 14 : Estadístico: 0.943 P-Valor: 0.021</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	

<p>Atributo 15 : Estadístico: 0.947 P-Valor: 0.029</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 16 : Estadístico: 0.911 P-Valor: 0.002</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 17 : Estadístico: 0.930 P-Valor: 0.007</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 18 : Estadístico: 0.928 P-Valor: 0.006</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 19 : Estadístico: 0.930 P-Valor: 0.007</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 20 : Estadístico: 0.934 P-Valor: 0.010</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 21 : Estadístico: 0.957 P-Valor: 0.073</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 22 : Estadístico: 0.962 P-Valor: 0.117</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 23 : Estadístico: 0.934 P-Valor: 0.010</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 24 : Estadístico: 0.967 P-Valor: 0.188</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 25 : Estadístico: 0.966 P-Valor: 0.182</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 26 : Estadístico: 0.939 P-Valor: 0.015</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	

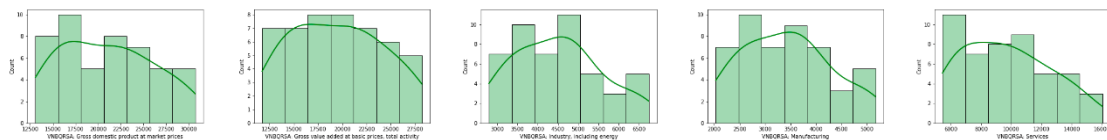
<p>Atributo 27 : Estadístico: 0.959 P-Valor: 0.089</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 28 : Estadístico: 0.957 P-Valor: 0.076</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 29 : Estadístico: 0.945 P-Valor: 0.025</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 30 : Estadístico: 0.953 P-Valor: 0.051</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 31 : Estadístico: 0.958 P-Valor: 0.082</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 32 : Estadístico: 0.962 P-Valor: 0.117</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 33 : Estadístico: 0.951 P-Valor: 0.044</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 34 : Estadístico: 0.969 P-Valor: 0.221</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 35 : Estadístico: 0.967 P-Valor: 0.188</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 36 : Estadístico: 0.942 P-Valor: 0.019</p> <p>Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 37 : Estadístico: 0.958 P-Valor: 0.086</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	
<p>Atributo 38 : Estadístico: 0.965 P-Valor: 0.164</p> <p>NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal</p>	

Atributo 39 : Estadístico: 0.945 P-Valor: 0.025 Se puede rechazar la hipótesis de que los datos de distribuyen de forma normal	
Atributo 40 : Estadístico: 0.967 P-Valor: 0.185 NO se puede rechazar la hipótesis de que los datos de distribuyen de forma normal	

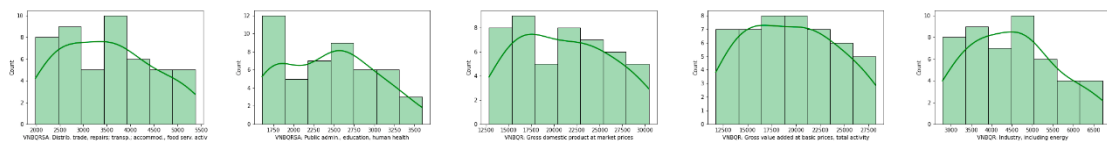
Després de veure els resultats del test de Shapiro, els atributs que no podem rebutjar, és a dir, que no es pot assegurar que no segueixen una distribució normal són els atributs 21, 22, 24, 25, 27, 28, 30, 31, 32, 34, 35, 37, 38 i 40.

Aquests són els histogrames dels atributs que poden seguir una distribució normal i que, per tant, els hem de considerar com possibles candidats a més representatius.

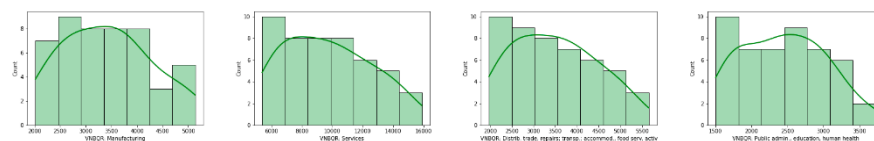
Atributs 21, 22, 24, 25 i 27:



Atributs 28, 30, 31, 32 i 34:



Atributs 35, 37, 38 i 40:



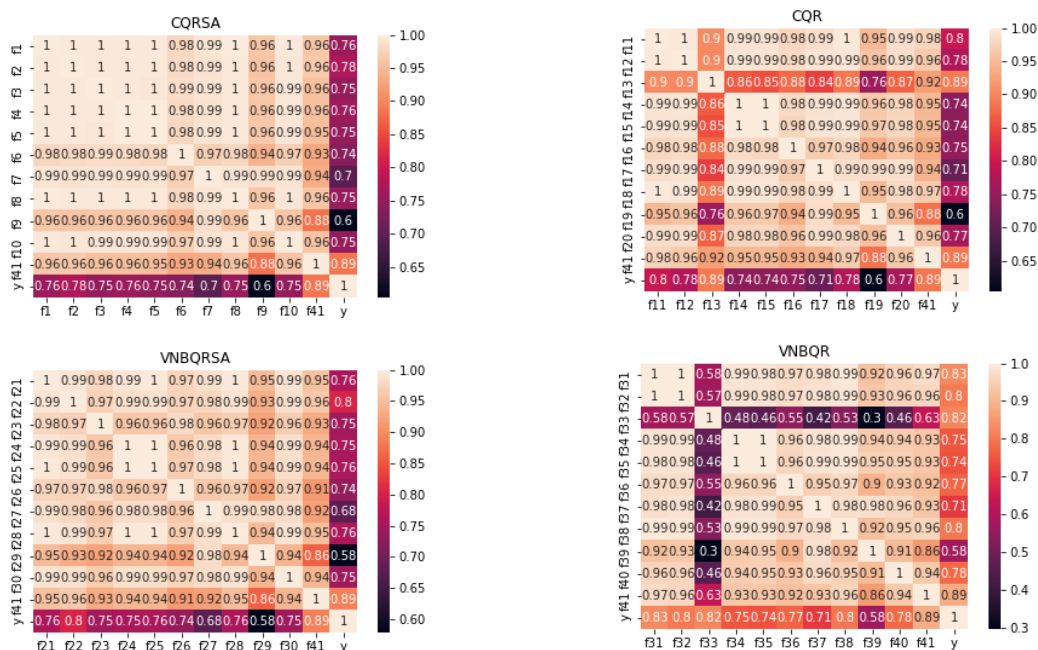
Com que ens interessen els atributs amb molta dispersió, hem de mirar també la dispersió de cada atribut, i rebutjar els que tinguin molt poca. Després de calcular la dispersió de cada atribut, aquest ha estat el resultat (la primera fila representa els atributs de l'1 al 10, la segona de l'11 al 20, i així successivament):

9188	8392	1476	1680	1297	635	4621	1531	1963	1172
9204	8392	1688	1674	1293	637	4610	1521	1963	1187
5060	4715	417	1058	859	334	2862	971	1382	552
5101	4715	830	1056	859	336	2852	969	1382	574

Per poder escollir els atributs amb més dispersió hem decidit seleccionar aquells que tinguin una dispersió de més de 2000. Aquests atributs són l'1, 2, 7, 11, 12, 17, 21, 22, 27, 31, 32 i 37.

Correlació entre dades

També hem calculat la correlació entre els diferents atributs per tal de saber si estan relacionats entre ells. El que més ens interessa es si estan relacionats amb l'atribut que volem predir. Els mapes de calor obtinguts són els següents:



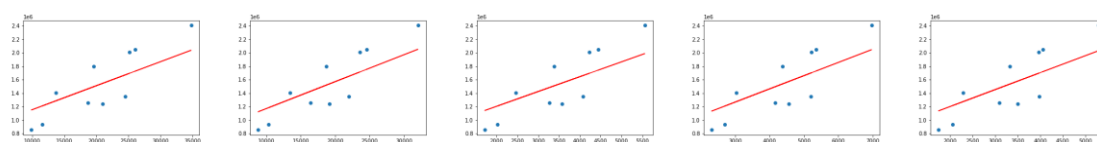
Aquests tipus de mapa son molt útils ja que, gràficament podem observar molt fàcilment quins atributs tenen una correlació més alta i més baixa segons el color que presenten. Com hem dit abans, el que més ens interessa és la seva relació amb l'atribut a predir. Hem descartat tots aquells valors que, comparats amb **y** no superen el 0.75, tots els que si que ho fan els hem guardat com a possibles candidats.

Regressió lineal

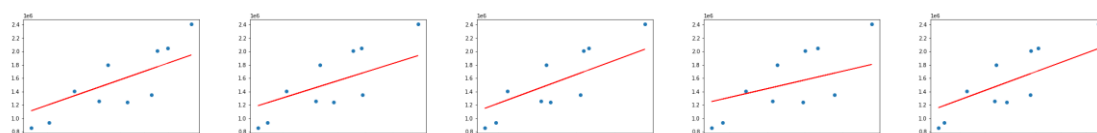
En aquest apartat farem la regressió lineal de cada atribut. Primer, la farem sense normalitzar les dades, i després ho tornarem a fer normalitzant-les. Un cop fet això calcularem l'error quadràtic mitjà del regressor tant normalitzat com sense normalitzar.

Regressió sense normalitzar dades:

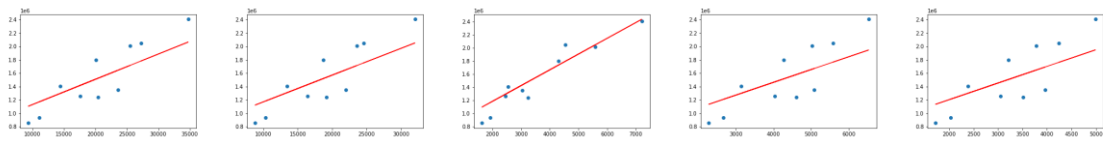
Atributs 1-5:



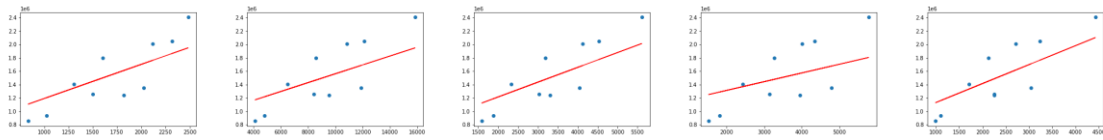
Atributs 6-10:



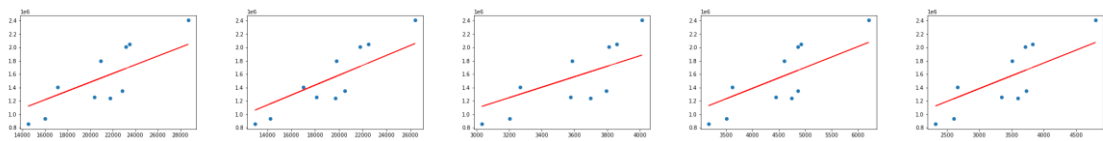
Atributs 11-15:



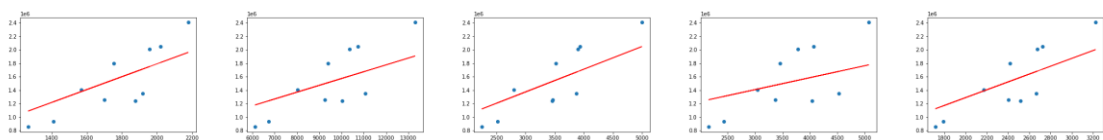
Atributs 16-20:



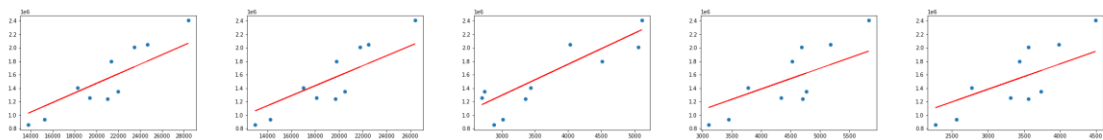
Atributs 21-25:



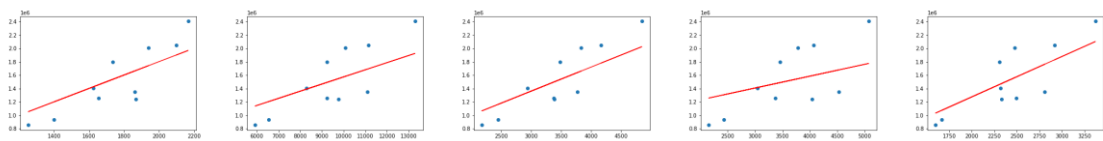
Atributs 26-30:



Atributs 31-35:

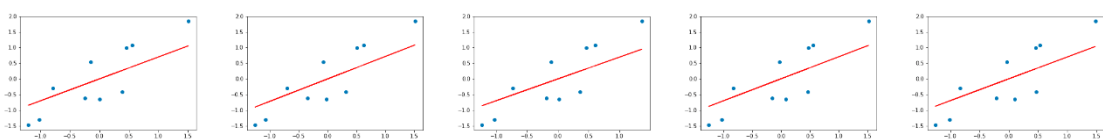


Atributs 36-40:

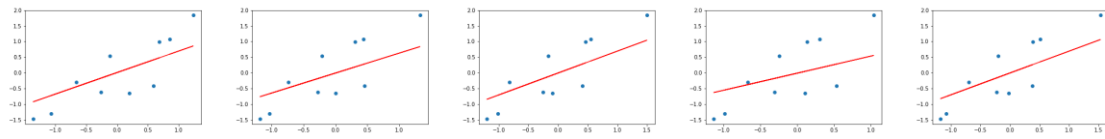


Regressió normalitzant dades:

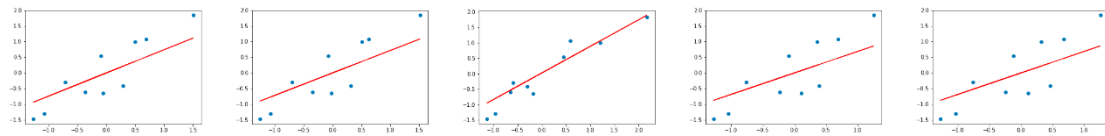
Atributs 1-5:



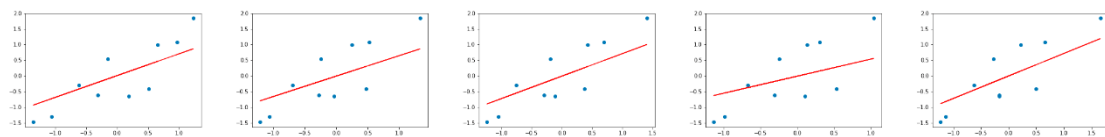
Atributs 6-10:



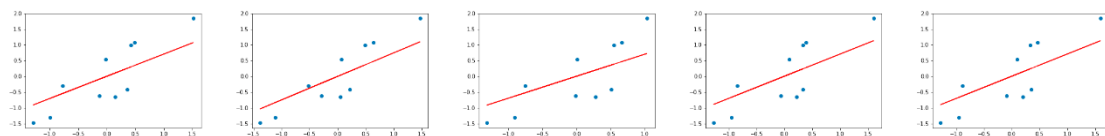
Atributs 11-15:



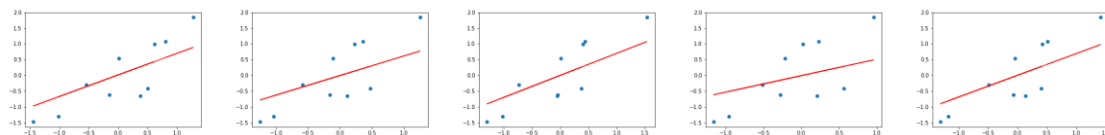
Atributs 16-20:



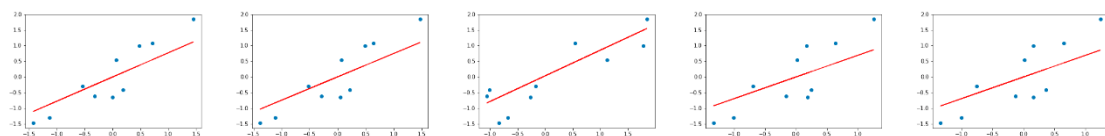
Atributs 21-25:



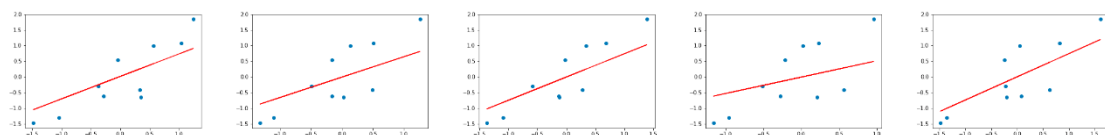
Atributs 26-30:



Atributs 31-35:



Atributs 36-40:



En totes les gràfiques es pot observar com, sense diferenciar entre atributs normalitzats o no, la recta de regressió no acaba de cobrir molt bé tots els punts. Per tant, per algunes mostres que li arribin al regressor, l'error serà molt gran, però per altres mostres que estiguin més a prop de la recta de regressió, l'error serà més petit.

Aquest problema ve de la distribució que tenen les dades; a les gràfiques de dispersió s'observen dues línies de punts ben separades. Això passa perquè els quatre trimestres que tenim a la base de dades tenen valors diferents, i justament el primer trimestre i l'últim tenen valors molt semblants, i els 2 trimestres del mig, tenen valors més baixos. Per tant, el regressor calcula la w_0 i w_1 per deixar al recta enmig d'aquestes dues línies de punts.

Error quadràtic mitjà (MSE)

En aquest apartat calcularem l'error quadràtic mitjà del regressor per a cada atribut de la base de dades. Amb aquest pas podrem observar quin atribut mostra un error quadràtic menor, que voldrà dir que difereix menys entre la predicció i la realitat.

Hem calculat aquest error amb les dades sense normalitzar i amb les dades normalitzades per a poder observar la diferència entre les dues maneres i per a veure la importància de normalitzar els atributs per a poder tractar els rangs de dades per igual.

Hem estandarditzat el resultat per a que l'error doni entre 0 i 1.

Error quadràtic	Sense normalitzar	Normalitzat
<i>Atribut 1</i>	0.2589	0.3838
<i>Atribut 2</i>	0.1985	0.3315
<i>Atribut 3</i>	0.2461	0.3927
<i>Atribut 4</i>	0.2820	0.3867
<i>Atribut 5</i>	0.2892	0.3984
<i>Atribut 6</i>	0.2965	0.3991
<i>Atribut 7</i>	0.3159	0.4778
<i>Atribut 8</i>	0.2552	0.3813
<i>Atribut 9</i>	0.4365	0.6252
<i>Atribut 10</i>	0.2629	0.3921
<i>Atribut 11</i>	0.1850	0.3076
<i>Atribut 12</i>	0.1985	0.3315
<i>Atribut 13</i>	0.0822	0.1156
<i>Atribut 14</i>	0.2799	0.4174
<i>Atribut 15</i>	0.3071	0.4319
<i>Atribut 16</i>	0.2738	0.3808
<i>Atribut 17</i>	0.3104	0.4608
<i>Atribut 18</i>	0.2294	0.3591
<i>Atribut 19</i>	0.4365	0.6252
<i>Atribut 20</i>	0.2898	0.3729
<i>Atribut 21</i>	0.2809	0.3937
<i>Atribut 22</i>	0.1601	0.2920
<i>Atribut 23</i>	0.3738	0.4757
<i>Atribut 24</i>	0.3238	0.4158
<i>Atribut 25</i>	0.3155	0.4059
<i>Atribut 26</i>	0.2998	0.4105
<i>Atribut 27</i>	0.3475	0.5141

<i>Atribut 28</i>	0.2565	0.3837
<i>Atribut 29</i>	0.5137	0.6774
<i>Atribut 30</i>	0.2675	0.4001
<i>Atribut 31</i>	0.1462	0.2624
<i>Atribut 32</i>	0.1601	0.2920
<i>Atribut 33</i>	0.2261	0.2580
<i>Atribut 34</i>	0.2849	0.4283
<i>Atribut 35</i>	0.3198	0.4416
<i>Atribut 36</i>	0.2494	0.3561
<i>Atribut 37</i>	0.3305	0.4765
<i>Atribut 38</i>	0.1915	0.3255
<i>Atribut 39</i>	0.5137	0.6774
<i>Atribut 40</i>	0.3466	0.3780
<i>Atribut 41</i>	0.1117	0.1571

Podem observar clarament que l'atribut amb un error quadràtic menor es el 41 però, com ja hem explicat al primer apartat, aquest atribut el descartem.

Una altra observació, una mica curiosa, es que l'error quadràtic mitjà és més gran quan normalitzem les dades que quan no ho fem. A la teoria, això seria una contradicció, ja que amb una normalització de les dades, el regressor hauria d'aprendre millor, però si les dades ja estan en un interval més o menys normal, no fa falta normalitzar. El que podem observar es que els nostres atributs d'entrada (del 1 al 40), més o menys estan en l'interval del 1.000 al 10.000, i la sortida es del ordre de 10^6 , per tant, aquesta diferencia hauria de interferir molt en la capacitat del regressor per aprendre, però no ho esta fent. Per tant, en aquest cas concret, no ens fa falta normalitzar les dades per aconseguir bons resultats.

Atribut escollit

Mesura	Possibles atributs
Histogrames i test de Shapiro	21, 22, 24, 25, 27, 28, 30, 31, 32, 34, 35, 37, 38 i 40
Desviacions	1, 2, 7, 11, 12, 17, 21, 22, 27, 31, 32 i 37
Mapes de calor	1, 2, 4, 11, 12, 13, 18, 20, 21, 22, 25, 28, 31, 32, 33, 36, 38 i 40

Els atributs que assoleixen aquests tres requisits són: 21, 22, 31 i 32.

Els que segueixen una distribució normal però solament assoleixen un dels altres dos requisits són: 25, 27, 28, 37, 38 i 40.

Agrupant tots aquests possibles atributs ens queden com a més possibles: 21, 22, 31 i 32, i també: 25, 27, 28, 37, 38 i 40 com a altres possibilitats.

Entre aquests atributs haurem d'escollir aquells que tinguin un error quadràtic mitjà menor.

<i>Error quadràtic</i>	<i>Sense normalitzar</i>	<i>Normalitzat</i>
<i>Atribut 21</i>	0.2809	0.3937
<i>Atribut 22</i>	0.1601	0.2920
<i>Atribut 25</i>	0.3155	0.4059
<i>Atribut 27</i>	0.3475	0.5141
<i>Atribut 28</i>	0.2565	0.3837
<i>Atribut 31</i>	0.1462	0.2624
<i>Atribut 32</i>	0.1601	0.2920
<i>Atribut 37</i>	0.3305	0.4765
<i>Atribut 38</i>	0.1915	0.3255
<i>Atribut 40</i>	0.3466	0.3780

Els atributs que menys error quadràtic tenen són el 22, 31, 32 i 38.

Per tant, aquests seran els atributs que seleccionarem com a més representatius.

Per aquests atributs més representatius, el regressor millora significativament les seves prediccions, però encara dona errors quadràtics una mica grans degut a aquesta separació entre els punts, ja comentada anteriorment.

PCA (Principal Component Analysis)

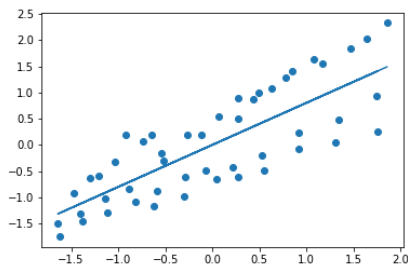
Per poder visualitzar els 40 atributs que té la nostra base de dades en un espai visualitzable, es podria aplicar un PCA per reduir la dimensió del espai a una observable (com 2 o 3). Si ho apliquéssim a la nostra base de dades, ens quedaria un espai de 2 dimensions (2D perquè es visualitza millor) i podríem veure de forma més directa la relació que hi ha entre els 40 atributs d'entrada de la nostra base de dades.

El descens de gradient

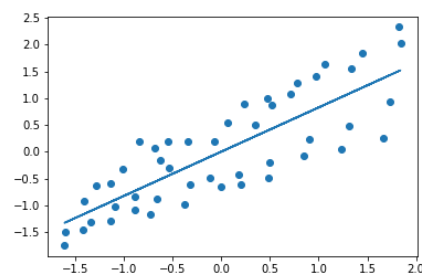
En aquest apartat hem aplicat el descens de gradient per a predir el número de turistes que arriben a l'Índia cada quadrimestre a cadascun dels atributs escollits en l'apartat anterior i hem calculat el seu error quadràtic.

Aquests han estat els resultats:

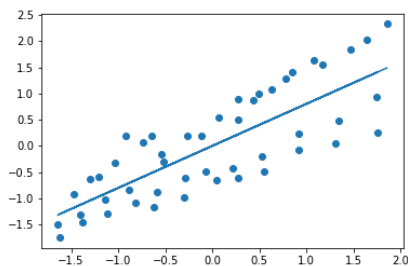
Atribut 22:



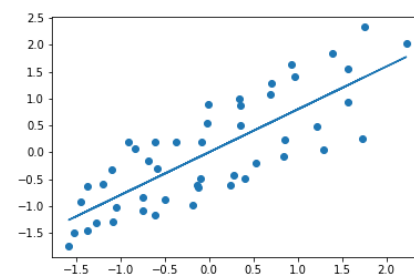
Atribut 31:



Atribut 32:



Atribut 38:



L'error quadràtic mitjà de cada atribut ha estat el següent:

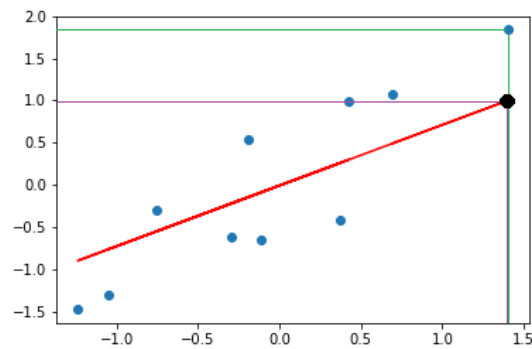
Atribut	Error quadràtic
<i>Atribut 22</i>	0.038
<i>Atribut 31</i>	0.030
<i>Atribut 32</i>	0.038
<i>Atribut 38</i>	0.040

Com es pot observar a la taula anterior, l'atribut que té un valor d'error menor és el 31. Per altra banda, el que en té un major és l'atribut 38, que té sentit, ja que era dels quatre, el que menys assolía els requisits. Els atributs 22 i 32 tenen aproximadament el mateix error.

Per tant, el millor atribut per predir noves dades és el que expressa la "moneda nacional, preus constants, any base nacional, nivells trimestrals del PIB als preus de mercat".

Aplicar el procés per un quadrimestre

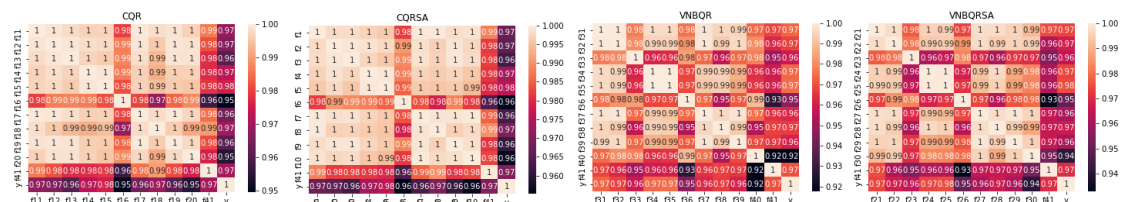
Les dades estan molt correlacionades amb el quadrimestre, per això la predicció que hem fet al treballar amb tots junts no es molt ajustada. Al treballar amb totes les dades alhora, fem la mitja i ens retorna un valor que no reflecteix la realitat.



Com podem veure en la següent gràfica d'exemple, la predicció que ens dona per a $x = 1.4$ es troba al voltant de l'1.0 però, el valor real és de gairebé 2.0.

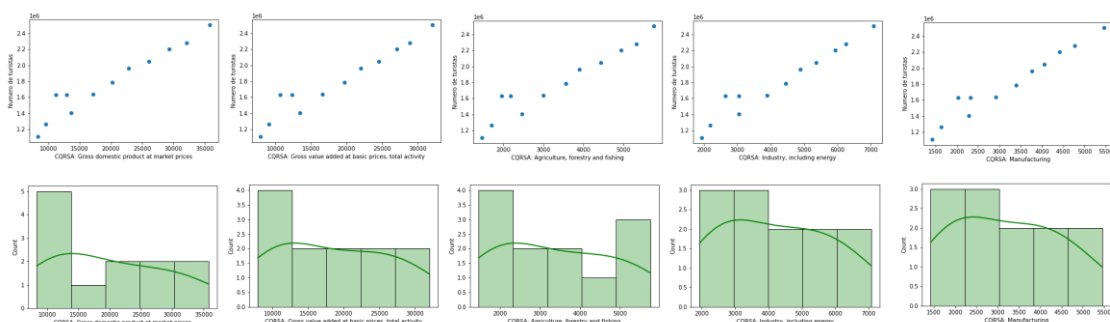
En aquest apartat plantejarem el problema des d'un altre enfoc, buscarem fer una predicció però només per a un quadrimestre. Les dades estaran més regularitzades i treballarem amb un marge d'error menor. Un cop plantejada aquesta hipòtesi, ho posarem en pràctica seguint el mateix procés que hem fet en tota la pràctica per a només el primer quadrimestre.

Aplicant els mateixos procediments d'anàlisi numèric dels diferents atributs. El primer que observem clarament és que les dades estan molt més correlacionades entre si, ja que la dispersió de les dades es menor.



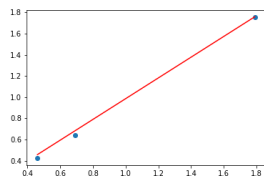
Com podem observar, cap correlació amb l'atribut objectiu baixa de 0.92. Podem veure que les gràfiques de punts ja no presenten el que abans semblaven dues fileres de punts, sinó que mostren una sola recta. En els histogrames veiem que s'ha suavitzat molt la seva corba.

Atributs 1-5:

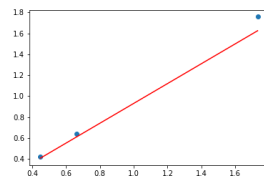


És important destacar que el test de Shapiro ja no ens és útil per a descartar atributs a l'hora d'escollir el que farem servir per a fer la predicció així que haurem d'escollir-ho per altres vies.

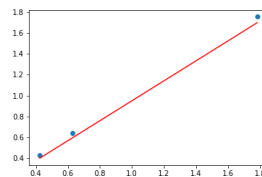
Per a triar l'atribut ens basarem en la regressió lineal i en l'error quadràtic. Al contrari que en la part anterior de la pràctica, observem que la recta s'ajusta molt a les nostres dades ja que no es troben tant distribuïdes en l'espai. De la mateixa manera, tots els errors quadràtics oscil·len entre $2.13 \cdot 10^{-5}$ i **0.011**. Els atribut amb els que farem el descens de gradient donat que tenen el menor error quadràtic i segueixen una distribució normal són: **15, 22, 31 i 32**. Les seves gràfiques de punts i els seus errors quadràtics son els següents respectivament (un cop normalitzats):



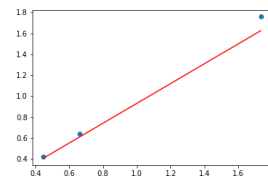
15: 0.0009



22: 0.0062



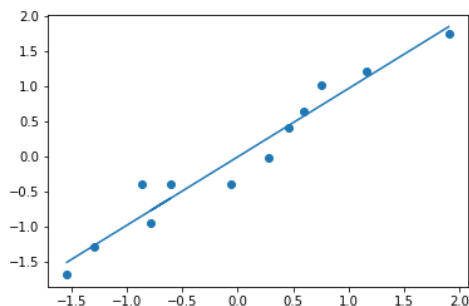
31: 0.0022



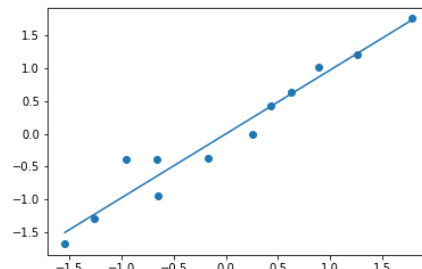
32: 0.0062

Per acabar, només farà falta aplicar el descens de gradient amb els atributs escollits. Seguirem el mateix procediment que el que hem aplicat anteriorment així que només mostrem i expliquem els resultats.

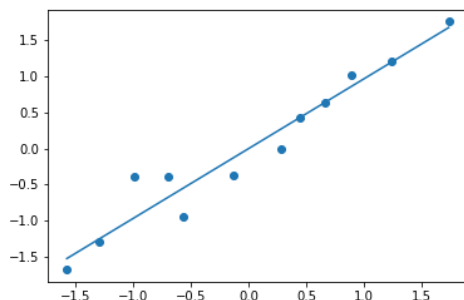
Atribut 15.- mostra error del 0.00064



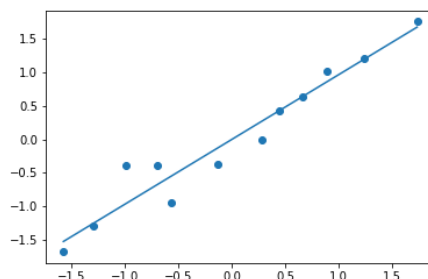
Atribut 31.- mostra error del 0.00072



Atribut 22.- mostra error del 0.00104



Atribut 32.- mostra error del 0.00104



Aplicant el mateix raonament d'abans, el millor atribut per a predir el número de turistes que arriben de l'Índia el primer quadrimestre és l'atribut 15, la Fabricació segons la moneda nacional, preus constants, any base nacional, ajustats estacionalment.

Amb aquest resultat tant satisfactori corroborem la nostra hipòtesi que, al fer l'estudi per quadrimestres podem predir amb molta més precisió.

Conclusions

Aquesta pràctica ens ha servit d'introducció al Machine Learning, aplicant de manera pràctica tota la teoria que hem anat veient a les classes.

Ara sabem com plantejar el problema, els mètodes que hem d'aplicar i com interpretar els resultats obtinguts. Tot això aplicant-lo a un problema real, afrontant els diferents dubtes que hem tingut durant la realització de la pràctica.

Hem entès la importància d'entendre els atributs de la base de dades, de visualitzar aquestes dades, de normalitzar-les i de fer una bona regressió per poder arribar a un bon resultat en el descens de gradient.

Un cop acabada la pràctica som capaços d'afrontar un problema d'aquest estil amb les eines i coneixements adquirits a través d'aquesta assignatura.