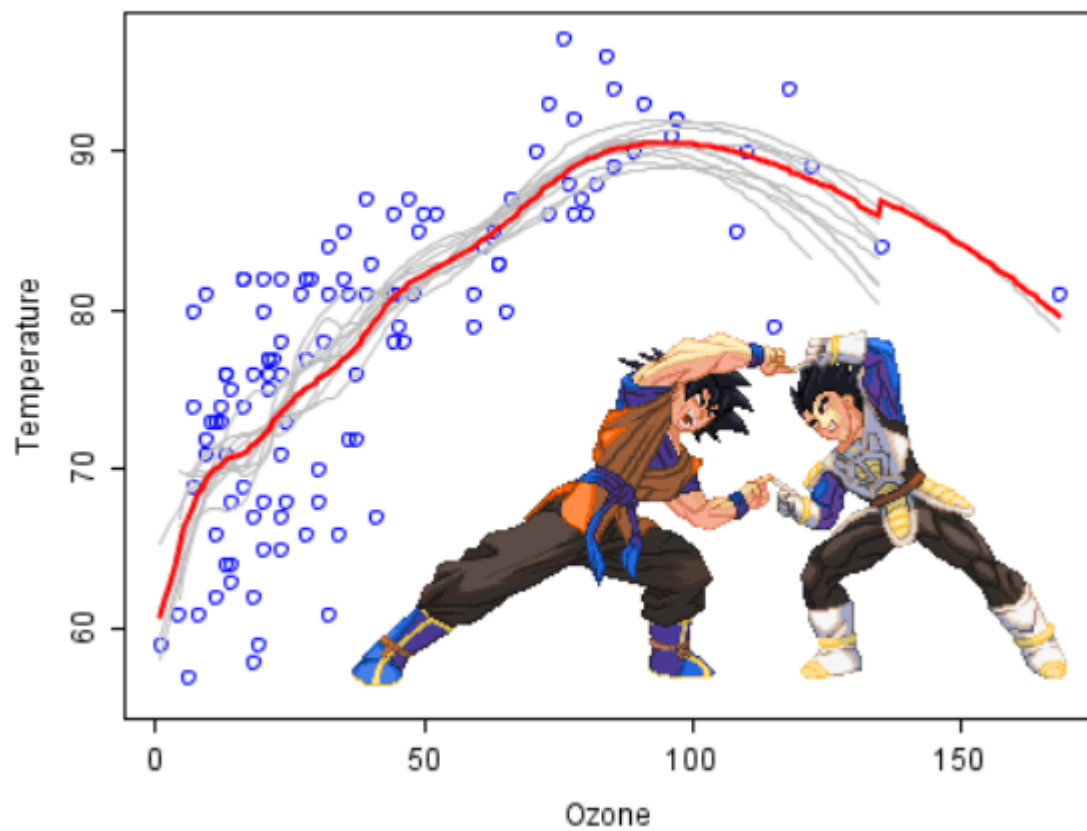


Aprentatge Computacional

Practica 1: Regressió



Grup GPA603-1530

Mario González 1566235

Ferran Bernabé 1564845

Daniel Gutiérrez 1563389

Introducció

En aquesta pràctica aplicarem els coneixements de classe (teòrics) de manera pràctica en un problema d'una base de dades real. Analitzarem els seus atributs amb mètodes i fórmules matemàtiques les quals podrem visualitzar per després treure conclusions. L'objectiu principal serà predir el número de visites a través d'altres atributs, i veure amb quins dels atributs es pot predir el nostre atribut objectiu i amb quins no.

Concretament, un cop haguem normalitzat els valors, aplicarem el **descens de gradient** a aquells atributs que puguin ser més representatius i per conseqüència ens permetin fer millors prediccions i/o anàlisi.

Per tal de poder realitzar aquesta tasca, utilitzarem les llibreries que disposa Python per fer aprenentatge computacional / machine learning, les quals ens permetran fer tot el que necessitem per la pràctica.

Les llibreries que utilitzarem seran:

- **Numpy:** llibreria de càlculs i funcions matemàtiques amb l'objectiu de obtenir un alt rendiment en totes les seves operacions.
- **Scikit-learn:** llibreria d'aprenentatge computacional per poder aplicar tots els algorismes que necessitem, com per exemple el **descens de gradient**.
- **Matplotlib / seaborn:** llibreries de visualització de dades, fer gràfics, taules, anàlisi d'aquests, ...
- **Scipy:** llibreria de càlculs matemàtics.
- **Pandas:** llibreries per manipulació d'estructures de dades per poder emmagatzemar les dades en taules, i també inclou mètodes d'anàlisi.

Explicació de la base de dades a analitzar

<https://www.kaggle.com/code/daan4k/yellow-stone-visits-eda/notebook>

Durant aquesta secció s'analitzarà la base de dades que ens ha estat assignada per tal d'entendre la relació entre l'activitat humana al parc de Yellowstone amb factors climàtics o macroeconomics. En especial ens centrem en com varia el nombre de visites en funció de la temperatura, clima i altres valors.

Aquestes dades no han estat escollides al atzar ni res per l'estil, es tracta d'una base de dades generada a partir de dades reals. A l'hora de realitzar l'estudi sobre aquesta cal tenir en compte la diferència entre coincidències i correlacions. Per molt que pugui semblar que hi hagi una relació directe entre dos valors, cal ser conscients de si es tracta d'una coincidència o no.

Apartat (C): Analitzant Dades

1. Quin és el tipus de cada atribut?

La nostra base de dades disposa de 20 atributs, els tipus dels quals son:

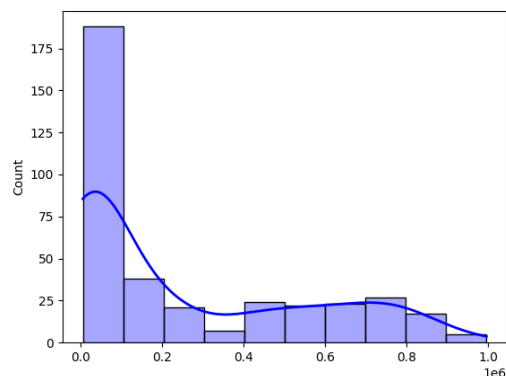
Recreation Visits	int64
LowestTemperature(F)	float64
HighestTemperature(F)	float64
WarmestMinimumTemperature(F)	float64
ColdestMaximumTemperature(F)	float64
AverageMinimumTemperature(F)	float64
AverageMaximumTemperature(F)	float64
MeanTemperature(F)	float64
TotalPrecipitation(In)	float64
TotalSnowfall(In)	float64
Max 24hrPrecipitation(In)	float64
Max 24hrSnowfall(In)	float64
Year/Month/Day	object
3month Percent Change Airfare Costs	float64
3month Percent Change Food Away From Home Costs	float64
3month Percent Change Gasoline Costs	float64
3month Percent Change Jet Fuel Costs	float64
Consumer Price Index	float64
Consumer Sentiment Index	float64
Unemployment Rate	float64

Com podem observar, disposem d'un atribut de tipus enter què és el de Recreation Visits, un de tipus objecte (date) que és l'atribut referent a la data (Year/Month/Day) i la resta d'atributs és de tipus flotant.

2. Quins atributs tenen una distribució Gaussiana?

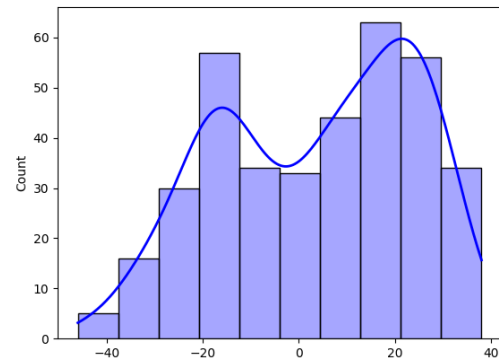
Recreation Visits

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



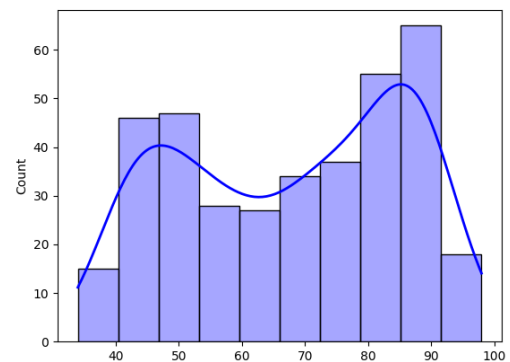
1 LowestTempreature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



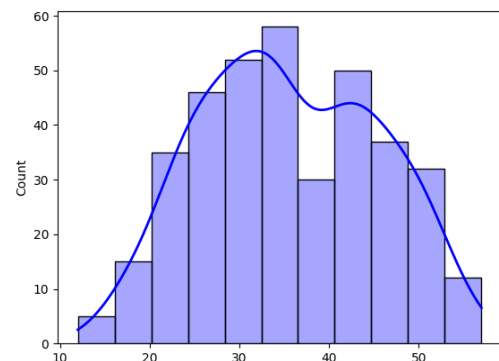
2 HighestTemperature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



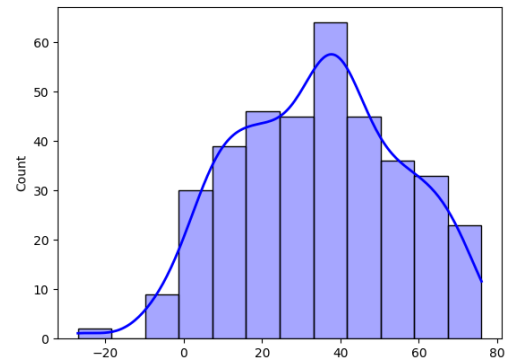
3 WarmestMinimumTemperature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



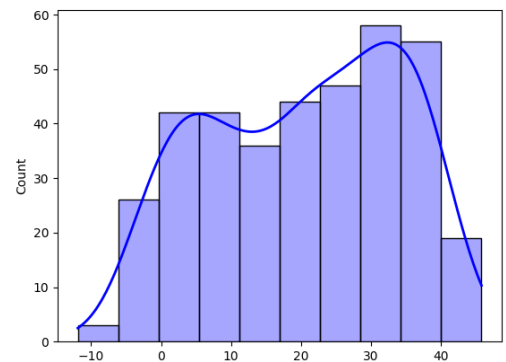
4 ColdestMaximumTemperature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



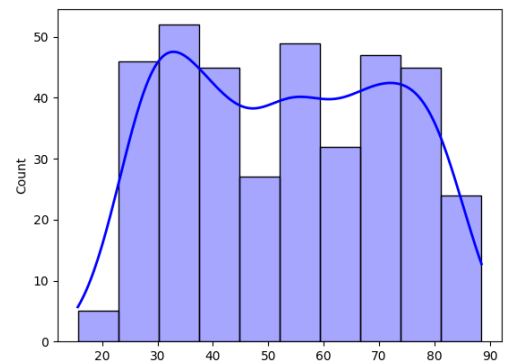
5 AverageMinimumTemperature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



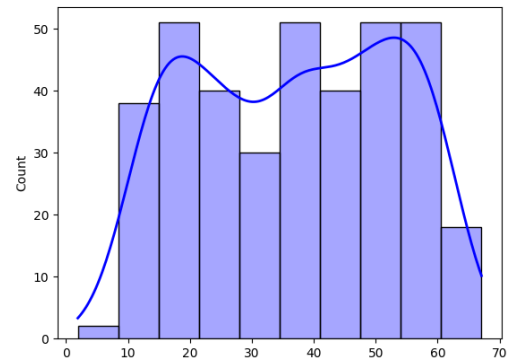
6 AverageMaximumTemperature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



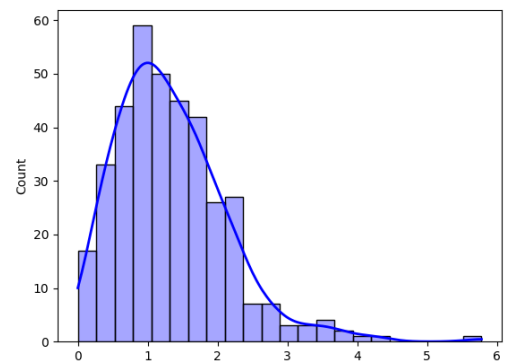
7 MeanTemperature(F):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



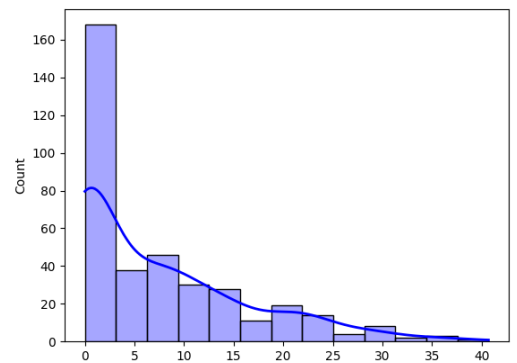
8 TotalPrecipitation(In):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



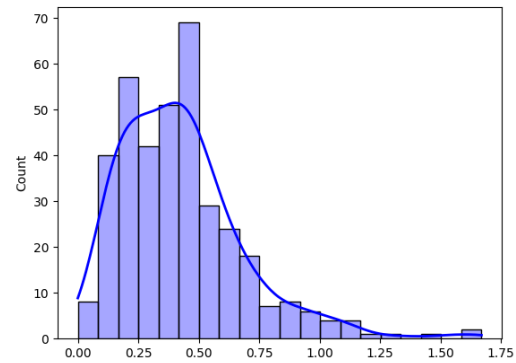
9 TotalSnowfall(In):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



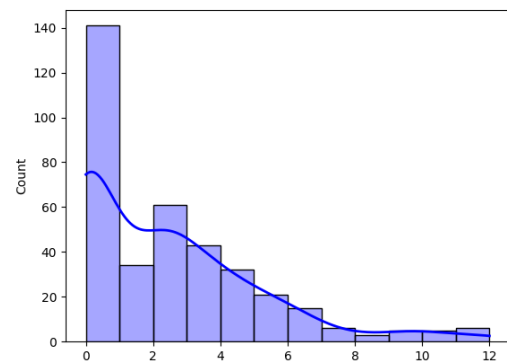
10 Max 24hrPrecipitation(In):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



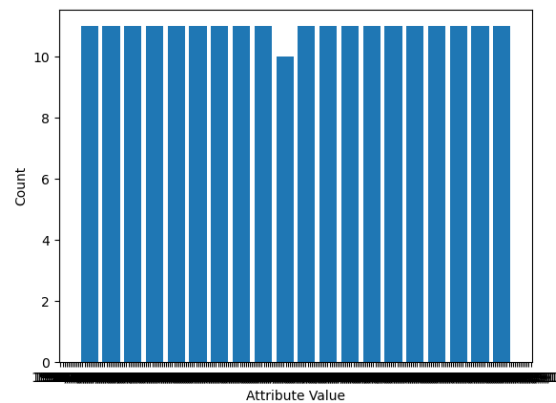
11 Max 24hrSnowfall(In):

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



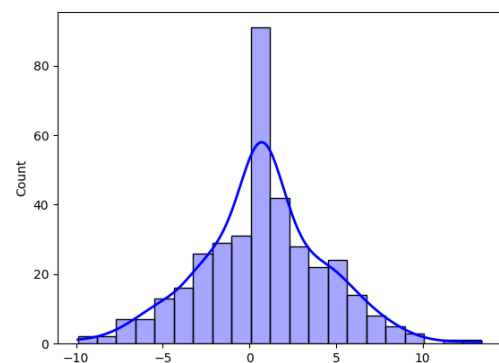
12 Year/Month/Day:

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



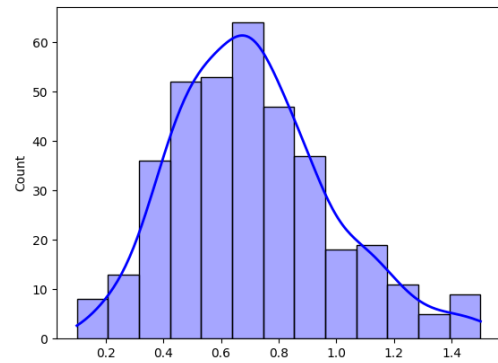
13 3month Percent Change Airfare Costs:

No es pot descartar que aquest atribut tingui una distribució Gaussiana.



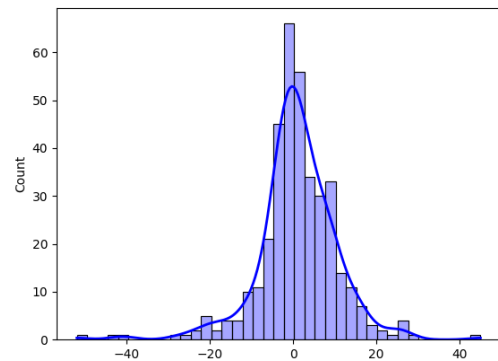
14 3month Percent Change Food Away From Home Costs:

No es pot descartar que aquest atribut tingui una distribució Gaussiana.



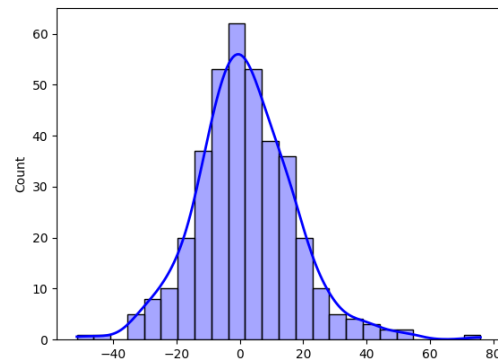
15 3month Percent Change Gasoline Costs:

No es pot descartar que aquest atribut tingui una distribució Gaussiana.



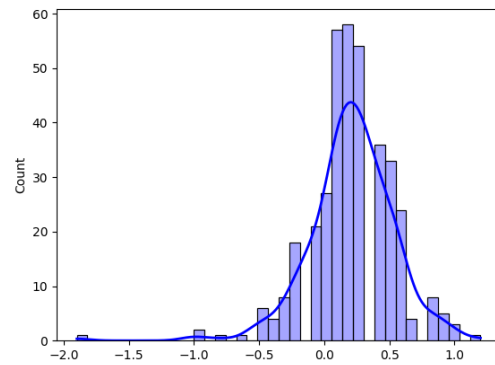
16 3month Percent Change Jet Fuel Costs:

No es pot descartar que aquest atribut tingui una distribució Gaussiana.



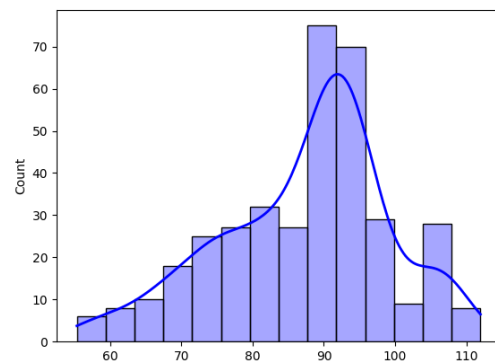
17 Consumer Price Index:

No es pot descartar que aquest atribut tingui una distribució Gaussiana.



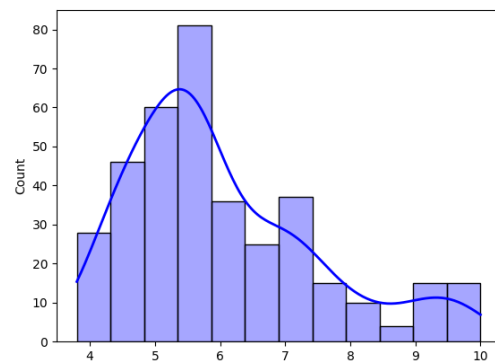
18 Consumer Sentiment Index:

Es pot descartar que aquest atribut tingui una distribució Gaussiana.



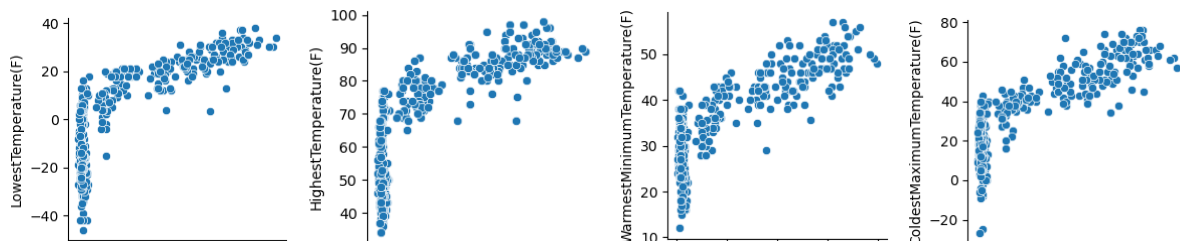
19 Unemployment Rate:

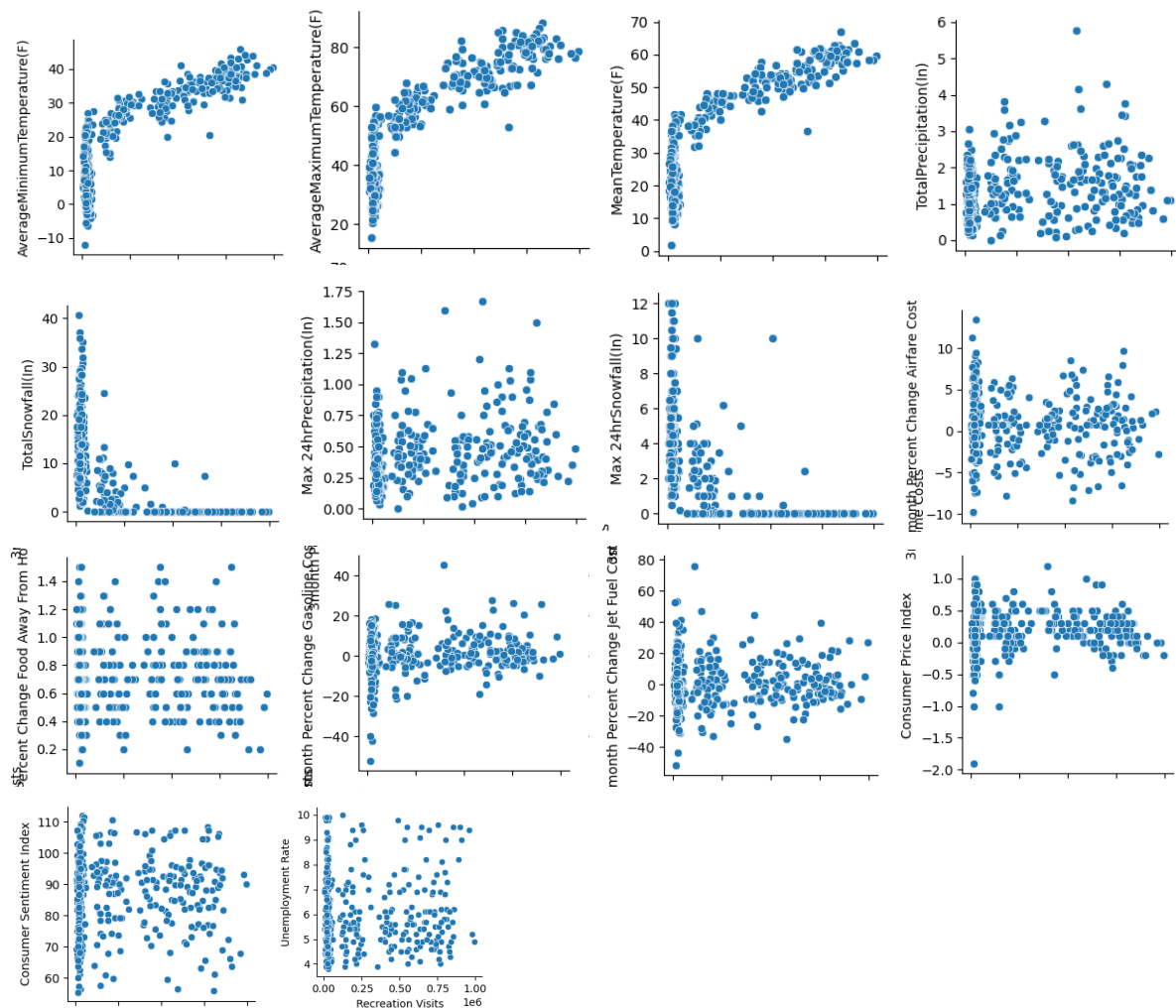
Es pot descartar que aquest atribut tingui una distribució Gaussiana.



Com podem observar els atributs que presenten una distribució gaussiana a l'histograma de barres són el 13, 14, 15, 16 i 17.

Relació de l'atribut Recreation Visits amb els altres atributs:





Es pot observar que els 7 primers atributs, tenen una relació logarítmica, de manera que es dibuixa una curva elíptica on s'observa que les visites no augmenten fins un punt determinat i després d'aquest punt, augmenten les visites de forma logarítmica.

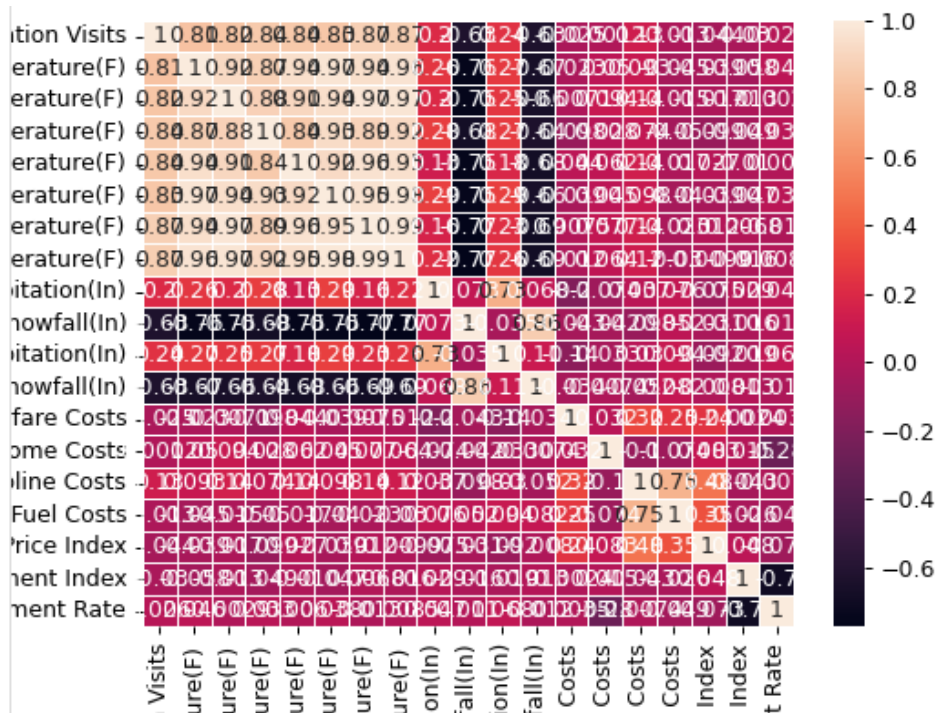
En quant a les gràfiques 9 i 11, relacionades amb l'efecte de la neu, podem observar que hi ha una relació directe entre si hi ha neu, pel qual la gent no visitarà el parc, i si no hi ha neu, les visites augmenten.

Per la resta de gràfiques, no hi ha una relació clara definida.

3. Quin és l'atribut objectiu? Per què?

L'atribut objectiu serà el **Recreation Visits**, ja que com bé diu l'autor el focus de la base de dades es analitzar el nombre de visites al parc respecte a altres factors (Temperatura, precipitació, preus de béns bàsics, taxa de desocupació, etc.) i com tots aquests l'afecten.

Per un primer anàlisi podem observar la matriu de correlació, que ens indica gràficament la relació i proporcionalitat entre dos atributs concrets. La matriu de correlació general és la següent:



D'aquesta manera, podem observar a primer cop d'ull que les visites tenen una alta correlació amb els atributs referents a la temperatura, de manera que una hipòtesi podria ser que les visites tenen una relació de causalitat amb la temperatura.

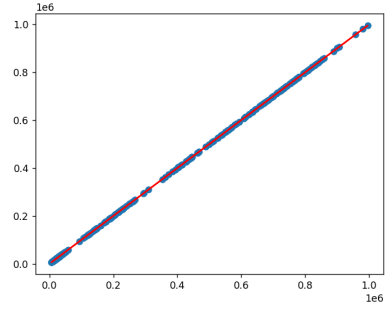
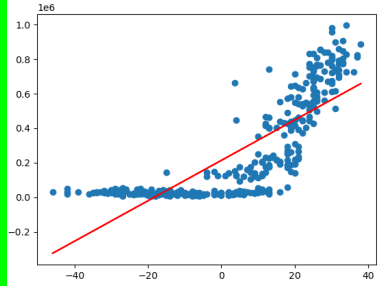
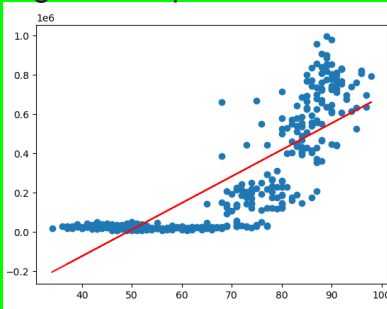
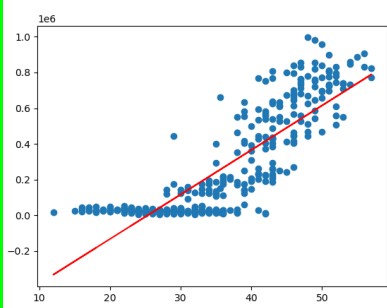
Per poder analitzar els efectes de causalitat i casualitat, utilitzarem models de regressió, específicament el **descens de gradient**.

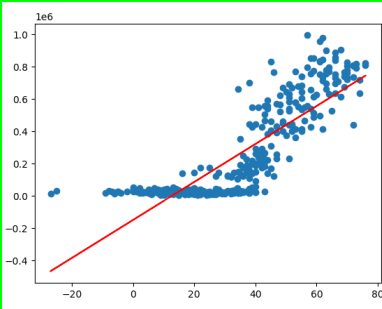
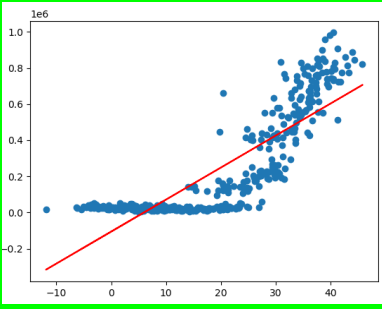
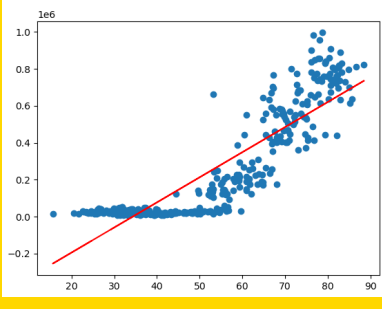
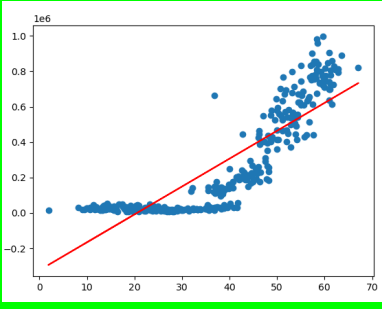
Apartat (B): Primeres regressions

1. Quin són els atributs més importants per fer una bona predicció?

Els atributs més importants per fer una bona predicció seran els que tenen un baix MSE al comparar-se amb el resultat de l'atribut objectiu. Aquests atributs seran els que en la matriu de correlació tenien un valor alt en la fila de l'atribut objectiu. En la nostra base de dades aquests atributs son els que estan relacionats amb la temperatura. Per comprovar aquest fet, utilitzarem mètriques com el MSE, R2...

2. Amb quin atribut s'assoleix un MSE menor?

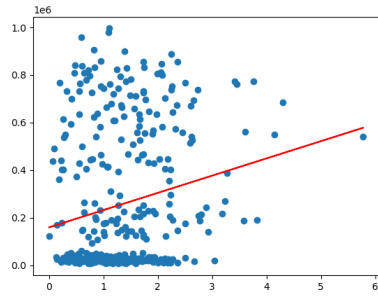
Atribut	MSE (no normalitzat)	MSE (normalitzat)
Recreation Visits 	0.000000	0.000000
LowestTemperature 	28940377249.70259	0.3409128671277541
HighestTemperature 	27753677157.985798	0.32693373592302527
WarmestMinimumTemperature 	24623723241.10622	0.2900633954096815

<div>ColdestMaximumTemperature</div> <div>A scatter plot showing the relationship between ColdestMaximumTemperature and an unlabeled variable. The x-axis ranges from -20 to 80, and the y-axis ranges from -0.4 to 1.0 (scaled by 1e6). A red regression line shows a positive correlation, starting near (-20, -0.4) and ending near (80, 0.8).</div>	25478142991.34341	0.30012831903768755
<div>AverageMinimumTemperature</div> <div>A scatter plot showing the relationship between AverageMinimumTemperature and an unlabeled variable. The x-axis ranges from -10 to 40, and the y-axis ranges from -0.2 to 1.0 (scaled by 1e6). A red regression line shows a positive correlation, starting near (-10, -0.2) and ending near (40, 0.7).</div>	26259642721.500072	0.30933425686526067
<div>AverageMaximumTemperature</div> <div>A scatter plot showing the relationship between AverageMaximumTemperature and an unlabeled variable. The x-axis ranges from 20 to 90, and the y-axis ranges from -0.2 to 1.0 (scaled by 1e6). A red regression line shows a positive correlation, starting near (20, -0.2) and ending near (90, 0.8).</div>	20030670686.00058	0.2359579944358358
<div>MeanTemperature</div> <div>A scatter plot showing the relationship between MeanTemperature and an unlabeled variable. The x-axis ranges from 0 to 70, and the y-axis ranges from -0.2 to 1.0 (scaled by 1e6). A red regression line shows a positive correlation, starting near (0, -0.2) and ending near (70, 0.8).</div>	21234794994.757767	0.25014238004127426

TotalPrecipitation

81586862197.06857

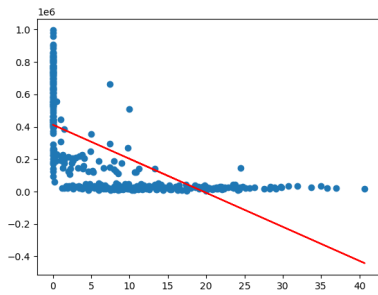
0.9610797700242644



TotalSnowfall

51063852220.06794

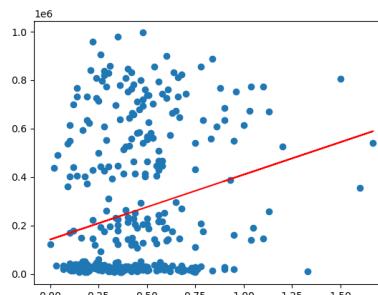
0.6015237505969343



Max 24hrPrecipitation

80133020916.02104

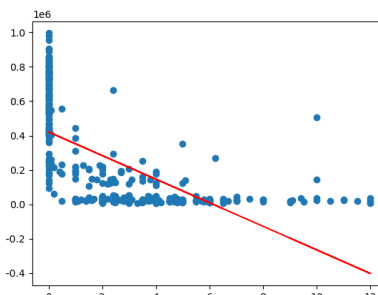
0.9439537597033142



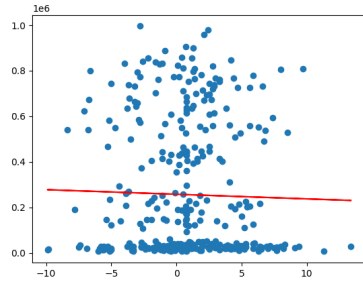
Max 24hrSnowfall

51321432040.92007

0.6045579983706689



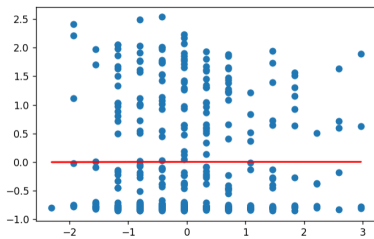
3month Percent Change
Airfare Costs



84838850298.31592

0.9993876530866032

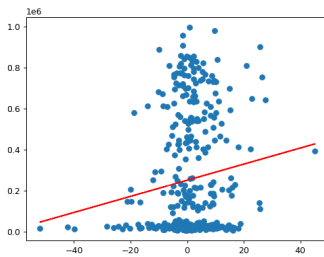
3month Percent Change
Food Away From Home
Costs



84890706659.5314

0.9999985124623553

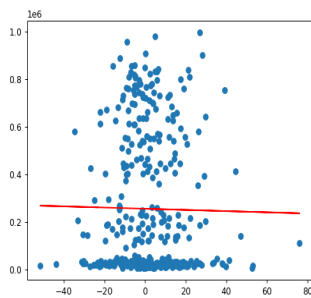
3month Percent Change
Gasoline Costs



83471220190.70598

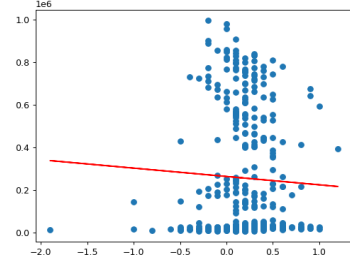
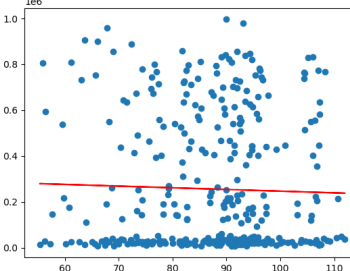
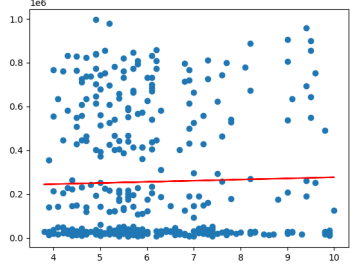
0.9832771961588057

3month Percent Change
Jet Fuel Costs



84875601572.70674

0.9998205770327933

<p>Consumer Price Index</p> 	84727417784.11115	0.9980749964622257
<p>Consumer Sentiment Index</p> 	84814425923.95772	0.9990999379881297
<p>Unemployment Rate</p> 	84832147714.52428	0.9993086977558603

Com podem veure, els atributs que assoleixen un menor MSE (Mean Squared Error), són els atributs 2,3,4,5,6,7, que corresponen a:

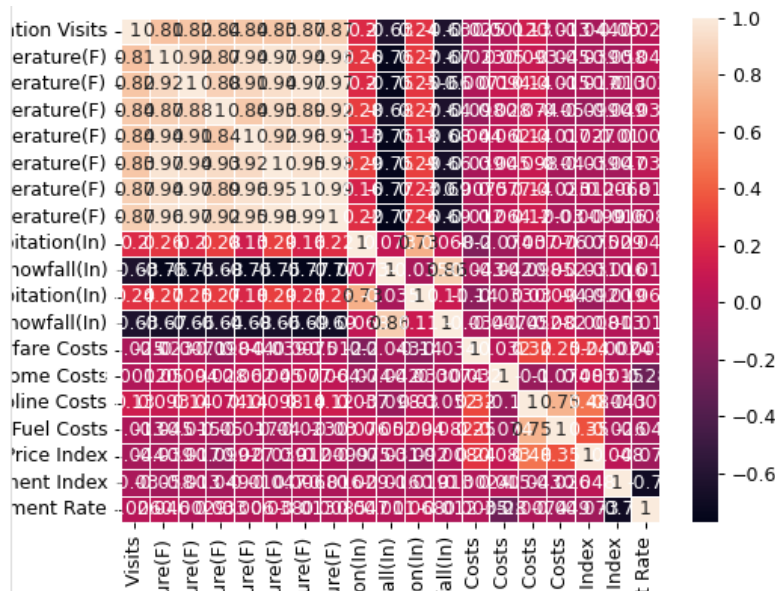
- LowestTemperature
- HighestTemperature
- WarmestMinimumTemperature
- ColdestMaximumTemperature
- AverageMinimumTemperature
- **AverageMaximumTemperature**
- MeanTemperature

On **AverageMaximumTemperature** és l'atribut que minimitza el MSE normalitzat, amb un valor de 0.23595.

També observem que aquests són els atributs que més correlació tenen i per tant és més fàcil fer una regressió a partir d'aquests.

3. Quina correlació hi ha entre els atributs de la vostra base de dades?

Com podem observar en la taula de l'apartat anterior, els MSE més baixos són els dels atributs que tenen una gran correlació amb l'atribut objectiu, en canvi, els atributs amb baixa correlació amb el Recreation Visits tenen un MSE més gran. Aquests resultats són lògics ja que, si tenim una baixa correlació entre dos atributs serà difícil predir-ne un a través de l'altre. Podem observar les correlacions en la matriu de correlació:



4. Com influeix la normalització en la regressió?

En el nostre cas influeix molt, ja que com podem observar a la nostra base de dades, els valors de l'atribut objectiu (Visites), són de l'ordre de 10^4 , mentre que la resta d'atributs en la majoria dels casos són d'un parell o tres d'ordres menys de magnitud, com podem veure:

	Recreation Visits	LowestTemperature(F)	HighestTemperature(F)	WarmestMinimumTemperature(F)	ColdestMaximumTemperature(F)
count	372.000	372.000	372.000	372.000	372.000
mean	255659.575	3.595	67.986	35.590	34.326
std	291752.719	20.216	17.671	9.863	20.749
min	6261.000	-46.000	34.000	12.000	-27.000
25%	24485.500	-15.000	51.000	28.000	18.000
50%	55627.500	6.500	71.000	35.000	35.000
75%	509026.500	21.000	84.000	43.000	50.000
max	995917.000	38.000	98.000	57.000	76.000

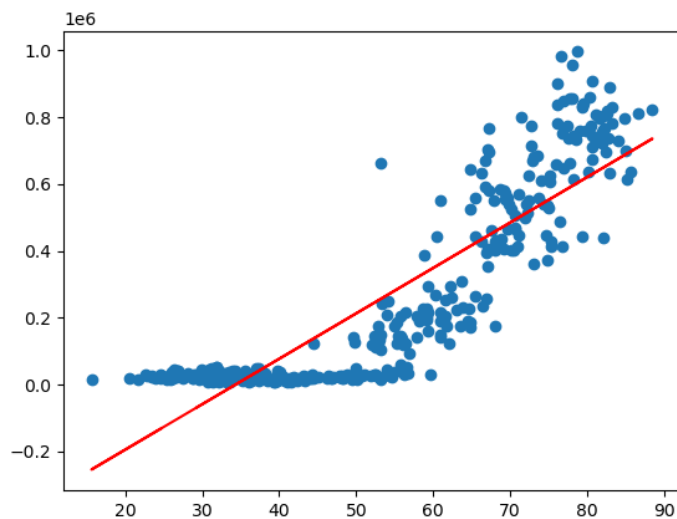
Per això, com hem pogut observar quan calculem el MSE, si no normalitzem les dades tenim valors de MSE d'ordre 10^{11} , en canvi al normalitzar-les el MSE es troba entre 0 i 1.

En quant als resultats de l'ordre de MSE no afecten, ja que els que tenen un MSE alt normalitzat el tenen també sense normalitzar, i els que tenen un MSE baix normalitzat també el tenen sense normalitzar.

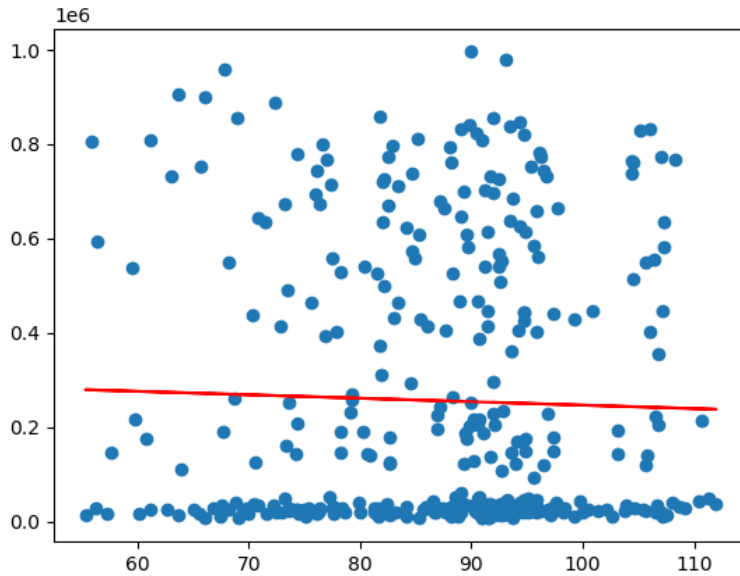
5. Com millora la regressió quan es filtren aquells atributs de les mostres que no contenen informació?

Hem observar que la regressió es molt més òptima en aquells atributs on la correlació amb l'atribut objectiu (Visites) es molt mes alta. Com hem comentat abans, aquests atributs son els que estan relacionats amb la **temperatura**.

En aquest objectius, podem observar de manera molt clara com hi ha una tendència que la regressió pot traçar una línea clara:



En canvi, en altres atributs com per exemple **Consumer Sentiment Index** , podem observar que al haver-hi una baixa correlació, en la gràfica no podem observar cap tipus de tendència.



Per això hem escollit com a atributs principals per a fer la regressió son tots els relacionats amb la **temperatura**.

6. Si s'aplica un PCA, a quants components es redueix l'espai? Per què?

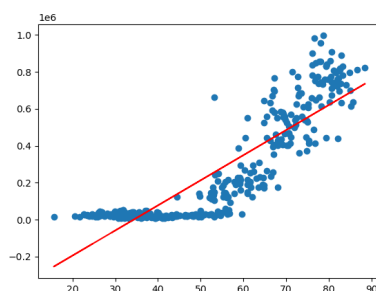
Si apliquem un PCA, podem reduir el espai de n-components, a un espai que podem visualitzar (2D, 3D), per tant podem reduir un conjunt molt gran de components a 2 o tres components principals.

Apartat (A): El descens de gradient

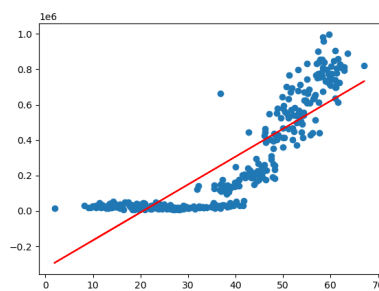
Com influeixen tots els paràmetres en el procés de descens? Quins valors de learning rate convergeixen més ràpid a la solució òptima? Com influeix la inicialització del model en el resultat final?

Amb les proves que hem fet hem trobat que un learning rate (alpha) de valor 0.05 ens minimitzava l'error resultant amb un temps d'execució no molt més gran que altres valors semblants.

Hem escollit els atributs 6, AverageMaximumTemperature i 7, MeanTemperature, que son els que menys MSE generaven:



Atribut 6, MSE = 0.235



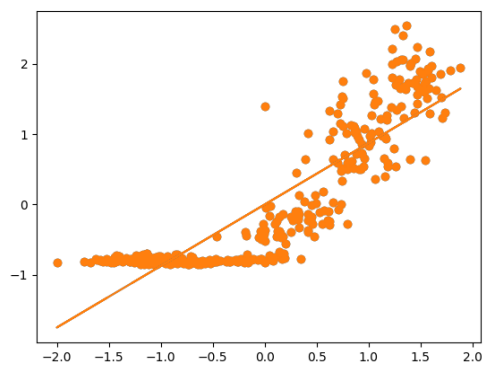
Atribut 7, MSE = 0.250

Hem aplicat el **descens de gradient**, amb els següents paràmetres:

- Coeficient d'aprenentatge (alpha) = 0.05
- N° d'iteracions = 500

Amb aquests valors i generant la recta pels atributs 6 i 7, hem generat les següents gràfiques de regressió:

Atribut 6 (AverageMaximumTemperature)



MeanTemperature

