APRENENTATGE COMPUTACIONAL

Pràctica 1

2022/23

GRUP 404-1130
Jonathan Rojas Granda - 1533448
Alex Fernández Ocón - 1571251
Pau Rovira Quiles - 1498591

ÍNDEX

Introducció

La Índia és el segon major productor de cebes del món i un dels centres del món de venta d'aquest vegetal. Durant l'any 2020, els preus de les cebes al país es van disparar a un nou nivell provocant que hi hagi problemes de subministrament a la població. Com a conseqüència, el govern govern prohibís la seva exportació en els mercats internacionals i va recollir tota la informació possible per fer-hi un anàlisi exhaustiu del mercat de la ceba.

En aquesta práctica ens demanen analitzar aquestes dades recollides sobre la venta de les cebes i poder extreure algun model sobre quina és la zona de la India on el preu de mercat serà més competitiu. Per això, ens proporcionen una col·lecció de dades amb una sèrie de factors com la data de venta, varietat de la ceba, el mercat, entre d'altres.

Objectius

- Aplicar models de regressió, ficant èmfasi en:
    1. Analitzar els atributs per seleccionar els més representatius i normalitzar-los.
    2. Avaluar correctament l'error del model.
    3. Visualitzar les dades i el model resultant.
    4. Saber aplicar el procès de descens del gradient.

- Ser capaç d'aplicar tècniques de regressió en casos reals.
- Validar els resultats de la col·lecció de dades sobre el preus de mercat de les cebes a l'Índia l'any 2020.
- Fomentar la capacitat per presentar resultats tècnics d'aprenentatge computacional de forma adequada davant altres persones.
- Trobar models que descriuen dades i permeten generar noves conclusions.

Raonament

La nostra col·lecció pertany a les dades recollides pel govern de la Índia sobre la venta de les cebes durant l'any 2020. Aquesta col·lecció conté fins a 107105 dades.

En la base de dades obtinguda trobem fins a 9 atributs, els quals representen :

- *State* : estat on s'ha venut la ceba
- *District* : districte on s'ha venut la ceba
- *Market*: mercat on s'ha venut la ceba
- *Variety*: varietat de la ceba
- *Commodity*: tipus de producte, en el nostre cas sempre serà la ceba (Onion)
- *Arrival_Date*: data d'arribada, venta de la ceba
- *Min_price*: preu mínim de venta de la ceba
- *Max_price*: preu màxim de venta de la ceba
- *Modal_price*: preu model de venta de la ceba

Variables de localització de la venta

Variable state

Les variables *state*, *district* i *market* ens aporten un valor referencial de les diferents zones de la India on s'han produit les ventes de cebes. D'aquesta forma podem saber de forma més precisa o més globalitzada, en funció del que ens interessi, en quin rang es mouen els diferents preus en funció de la seva geolocalització.

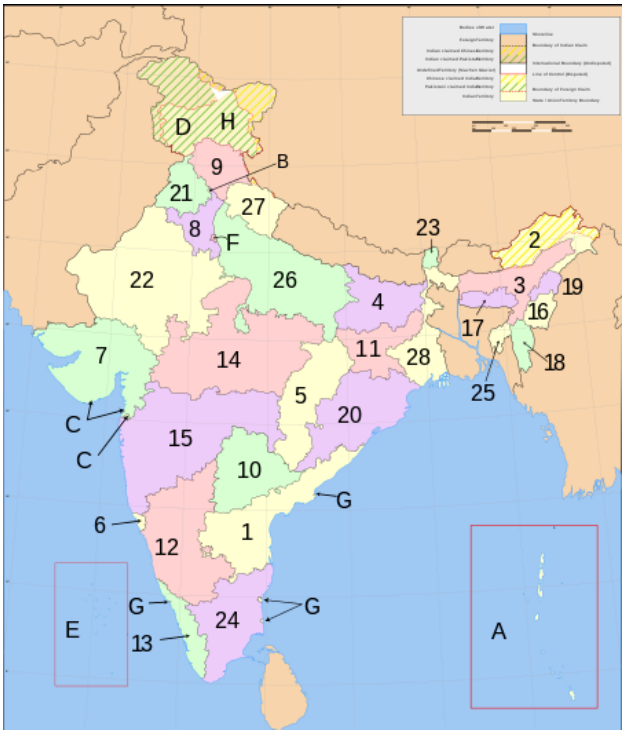
La variable ***state*** defineix en quin estat federat del país s'ha produit la venta. Tenim dades de fins a 22 estats diferents dels 28 totals que composen la Índia.

Tenim informació dels 20 estats de :

1. Andhra Pradesh, 5. Chattisgarh, 6. Goa, 7. Gujarat, 8. Haryana, 9. Himachal Pradesh, 10. Telangana,, 11. Jharkhand, 12.Karnataka, 13. Kerala, 14. Madhya Pradesh, 15. Maharashtra, 19. Nagaland, 20. Odisha, 21. Punjab, 22. Rajasthan , 25. Tripura, 26. Uttar Pradesh, 27. Uttrakhand, 28. West Bengal

Els 2 estats que provenen de territoris de la unió de la India (divisió administrativa amb govern propi):

- D. Jammu and Kashmir
- F. NCT of Delhi

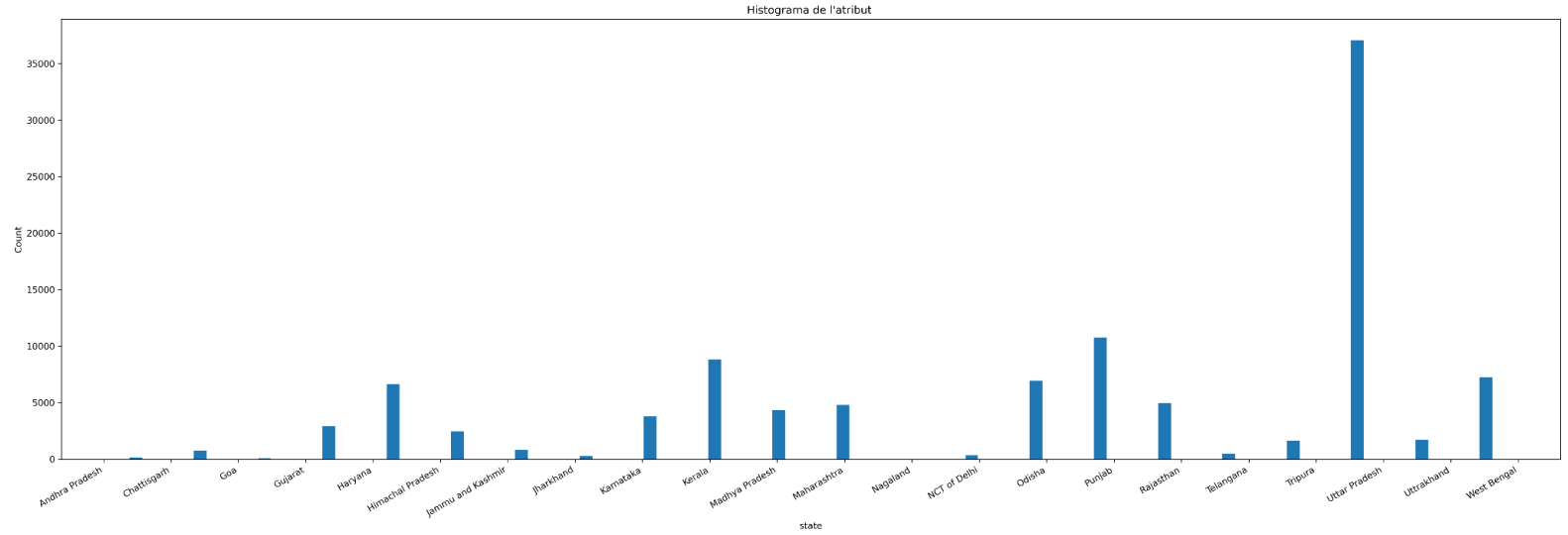


Només ens resten informació els 8 estats de :

2. Arunachal Pradesh
3. Assam
4. Bihar
16. Manipur
17. Megalaya
18. Mizoram
23. Sikkim
24. Tamil Nadu

Els estat amb més dades són:

Uttar Pradesh	34,6%
Punjab	10.0%
Kerala	8.2%
Remainder	47.2%



[Histograma sobre l'estat de la nostre col·lecció.]

Variable District

La variable ***district*** defineix en quin districte federat del país s'ha produit la venta. Tenim dades de fins a 315 districtes diferents dels 771 totals que composen la Índia.

Els districtes que obtenim més dades amb el seu percentatge són:

Kottayam	1,56%
Alappuzha	1,45%
Nashik	1,37%
Bulandshahar	1.30%
Hamirpur	1,11%
Remainder	93,20%

Variable Market

La variable ***market*** defineix en quin market federat del país s'ha produit la venta. Tenim dades de fins a 905 mercats diferents que composen la Índia.

És una variable que ens aporta molta precisió geogràfica sobre els punts de venta, i podrem determinar en exactitud quin té un mercat més competitiu de tot el país.

Al haver-hi una gran varietat de mercats, les dades per mercat son poques.

Els mercats que contenen més dades són:

Kayamkulam	0.430%	461 dades
Hubli (Amaragol)	0.388%	416 dades
Palakkad	0.345%	370 dades
Remainder	98,83%	105858 dades

Variables de informació del producte

Les variables *variety, commodity* són les variables que ens defineixen quin tipus de producte s'ha realitzat la venta.

Variable Commodity

La variable commodity representa el tipus de producte que s'ha venut.
En el nostre cas, només ens fixem en la venta de les seves cebes. Per tant, és una dada que no ens aporta cap mena de informació. Només ens ajuda a confirmar que les dades son sobre les cebes.

Com ens aporta molt poca informació, per millorar el càlcul i anàlisi de les dades vam decidir eliminar-la. Ja que com farem un anàlisi exhaustiu dels preus en funció de la seva ubicació, no afectarà al seu resultat.

Variable Variety

La variable *variety* defineix quina varietat del producte, és a dir, de la ceba, s'ha produit la venta.

Tenim recollides fins a 21 tipus diferents de cebes:

Local, **Other**, **Onion**, Nasik, Red, White, Beelary-Red, **1stSort**, Bangalore-Samall, Puna, Pusa-Red, Bombay (U.P.), Telagi, Hybrid, **Big, Small**, **2nd Sort**, **Pole**, Dry F.A.Q., **Medium**, Bellary.

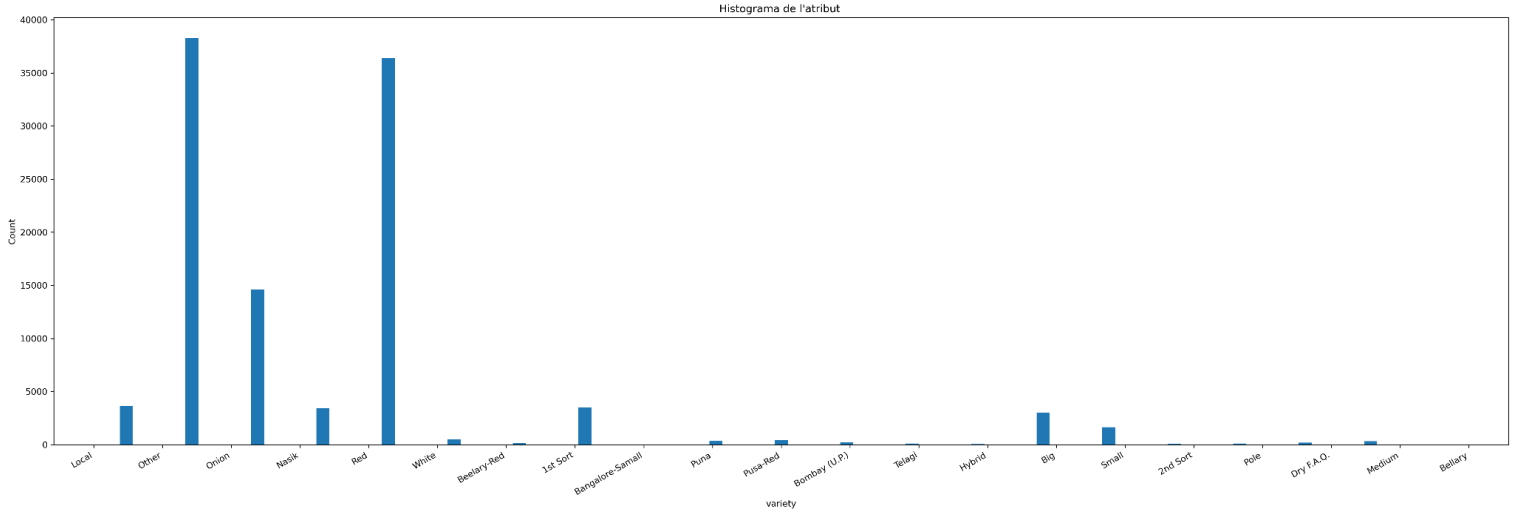
Hem pogut observar que algunes de les que hem analitzat no es corresponen realment amb varietats reals de cebes, o simplement son característiques o aproximacions del tipus de ceba. Per exemple, les varietats Big, Small, Medium, que representen simplement la mida.
Entre aquestes destaquem com que no son variants les següents:

- Other
- Onion
- 1st Sort
- 2n Sort
- Big
- Small
- Pole
- Medium

Hem contrastat la informació i no hem trobat ningun indici que puguin referir-se algun tipus concret de ceba.
Amb la informació que sen's dona no tenim informació prou contrastada per relacionar-les i és per això que hem decidit no prendre ninguna mesura al respecte i deixar-les com a tipus de varietat. Potser, amb un anàlisi més exhaustiu dels preus podrem determinar-ho.

Les varietats que obtenim més dades amb el seu percentatge són:

Other	35,76%
Red	33,98%
Onion	13,63%
Local	3,42%
Remainder	13,21%



[Histograma sobre la varietat de la nostra col·lecció.]

Variable Arrival_date

La variable *arrival_date* defineix en quina data s'ha produit la venta. El format de la variable és de tipus *object*. Com ens interessava poder-ho comprovar segons el mes de l'any 2020, l'hem convertit en tipus *data* per tal de desglosar-la en dos nous atributs, el mes i el dia.

Aquest canvi ens ha facilitat poder plasmar les dades als gràfics i poder fer un anàlisi més general, perque amb diferents parelles de valors no s'apreciava res.

Variables de informació dels preus

Les variables *min price, max price i modal price* són les variables que ens defineixen quin preus de producte quan s'ha realitzat la venta.

El min_price representa el valor mínim que s'ha venut la ceba en rupies.
El max_price representa el valor màxim que s'ha venut la ceba en rupies.
El modal_price representa el preu esperat de venut de la ceba en rupies.

De les dades obtingudes s'han eliminat les files el cual el mínim preu o el máxim preu són zero, pel fet que no ens aporten cap valor en la nostre col·lecció apart d'això no tindria cap sentit que un valor máxim sigui zero o valor mínim sigui zero quan el preu model és més gran que zero.

Aquests valors són numerics, representat en rupies. Això significa que el valor que tractarem és bastant gran, respect el l'euro o el dollar per a preus com la ceba.

Explorant els atributs veiem que el preu més baix que conté és de 20, i el preu més alt es 25000. La mitja del preu model és de 2109 rupies.

Una altre estadística són els percentils de cada atribut, els mostrem a continuació:

	min price	max price	modal price
25%	1000	1225	1150
50%	1400	1800	1600
75%	2400	2800	2550

Aquesta taula ens dona informació per comparar els resultats obtinguts de cada atribut, i apart ens serveix per saber en quin percentatge estan ubicats. Com veiem el 25% de cada atribut ronda entre 1000 a 1225, el 50% entre 1400 a 1600 y el 75% entre 2400 a 2800.

Atribut objectiu

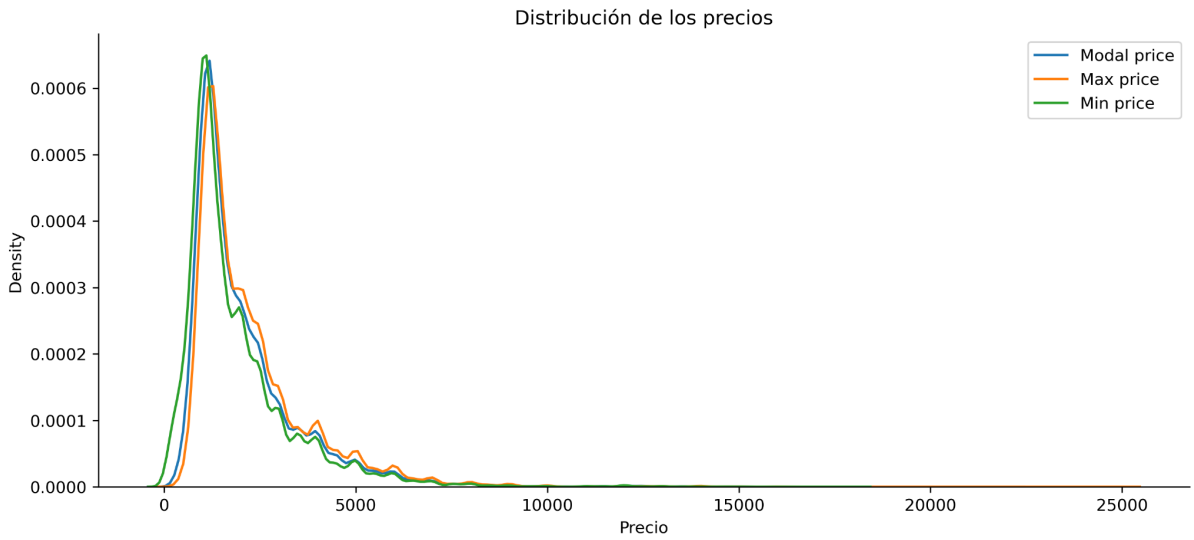
Amb tota la informació de la col·lecció, pensem que l'atribut que hem de prendre com a objectiu és el modal_price, per diverses raons :

- És l'únic que ens dona una mitja entre el valor del preu mínim i el preu maxim.
- És el més representatiu per analitzar la col·lecció de dades.
- És una aproximació del preu al que es vendrà la verdura.

Experiments realitzats

Distribució

S'han avaluat els atributs números **min price, max price i modal price**.
En el gràfica podem observar que els tres valors es tenen una gran densitat entre els preus de 0 i 5000 rupies. Per tant, el gran nombre de preus es corresponen a aquesta franja.
Podem observar que a partir de 10000 rupies no s'aprecia gairebé ninguna densitat, i s'acosta fins el valor de 25000 rupies.
Aleshores entenen que hi ha valors excepcionals molt superlatius respecte la distribució normal.
Aleshores, amb aquesta informació i amb la forma de la gràfica, arribem a la conclusió que la distribució és logarítmica-normal. Conté una distribució de valors molt concentrada en una certa regió i es va diluint degut a valors molt grans.



[Gráfic de densitat sobre els preus de la nostra col·lecció.]

Correlació

S'ha calculat la correlació que hi ha entre els valors numerics de la nostre col·lecció, els atributs que s'han analitzat son: *min price, max price, modal price, month i day* (desglosats de *arrival_date*) .

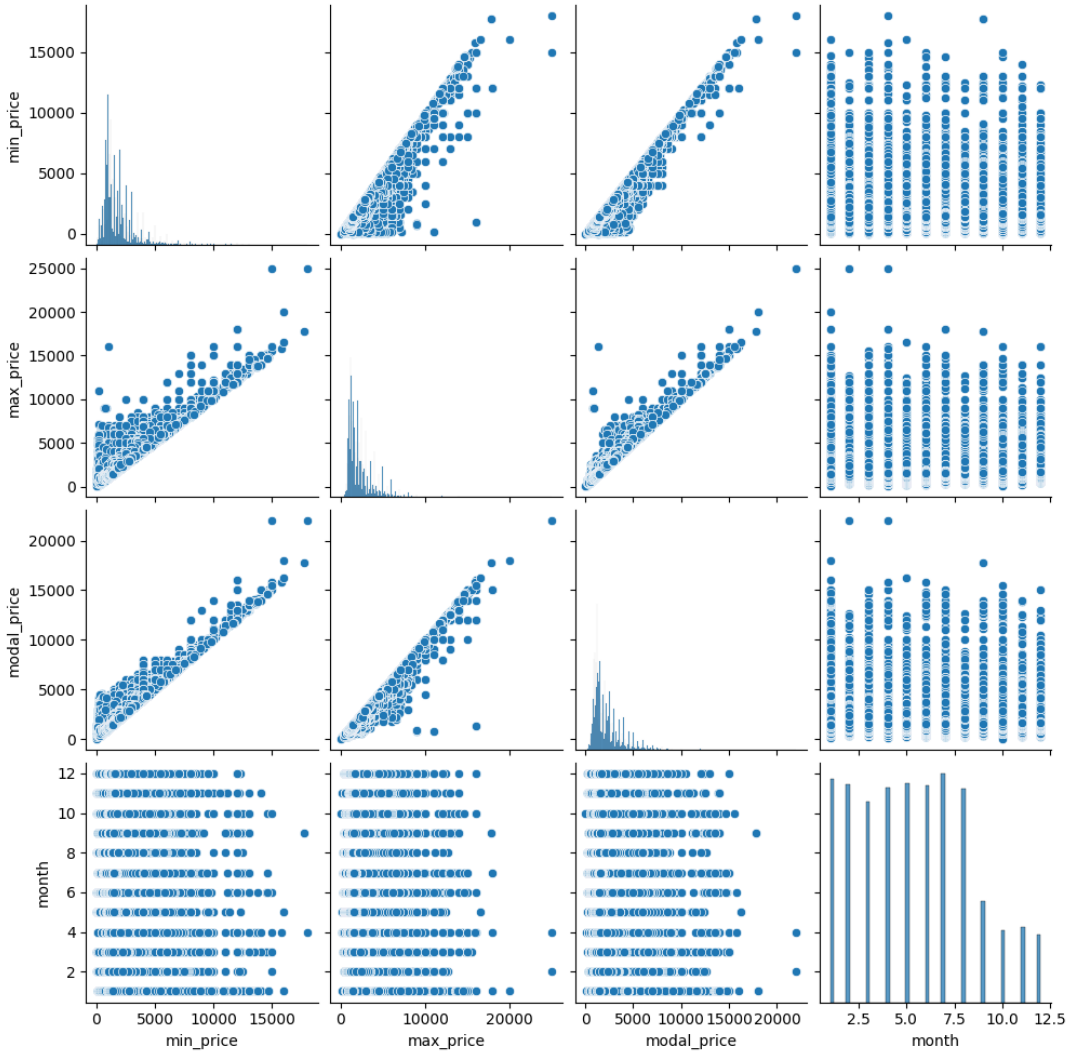
[Mapa de calor sobre la nostra col·lecció.]

S'ha arribat a la conclusió que el atribut modal price té una alta relació lineal amb els atributs max price i min price són els més alts i per tant suposa que hi ha una gran relació lineal entre aquests atributs.

Relacions lineal obtingudes:

- Entre el modal price i min price hi ha una relació de **0.98**
- Entre el modal price i max price hi ha una relació de **0.99**
- Entre el max price i min price hi ha una relació de **0.94**
- Les relacions de mínim preu, máxim preu i el preu model entre el mes i el dia dona valors negatiu. Per tant, no té cap mena de relació.

S'observa que entre els treus preus hi ha una aproximació molta alta. Com podem veure entre modal_price i max_price hi ha gairabé la mateixa relació que amb modal_price i max_price: 0.99 a 0.98 respectivament.
La relació no és tan alta entre max_price i min_price, de 0.94 només. això ens dona més motius per pensar que l'atribut objectiu ha de ser el modal_price, ja que és qui té més relació entre les dades, i podem pensar que els valors frontera son el màx_price i min_price.



[Correlació de cada atribut número de la nostra col·lecció.]

Categorització

Abans de categoritzar les dades, s'han eliminat el atribut *district* i *market*. En el nostre cas s'ha fet per tal de poder treballar millor amb les dades. Pensem que no és important tenir una presició territoral, i podem resumir-ho per l'atribut *state*, ja que formen part de l'estat.

Per poder utilitzar les variables categóricas el que s'ha fet és utilitzar *dummies*, per convertirles en numériques. La hem fet sobre l'atribut variety.

	min_price	max_price	month	day	state_Andhra Pradesh	state_Chattisgarh	state_Goa	state_Gujarat	state_Haryana	state_Himachal Pradesh	...	variety_Nasik	variety
0	1350	4390	3	1	1	0	0	0	0	0	...	0	
1	1390	4400	4	1	1	0	0	0	0	0	...	0	
2	1460	5150	6	1	1	0	0	0	0	0	...	0	

3 rows × 47 columns

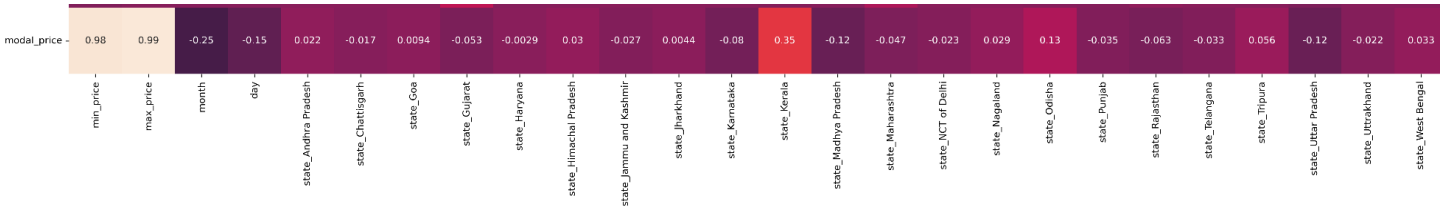
[Dataset amb la categorització]

Correlació amb varietat i estat

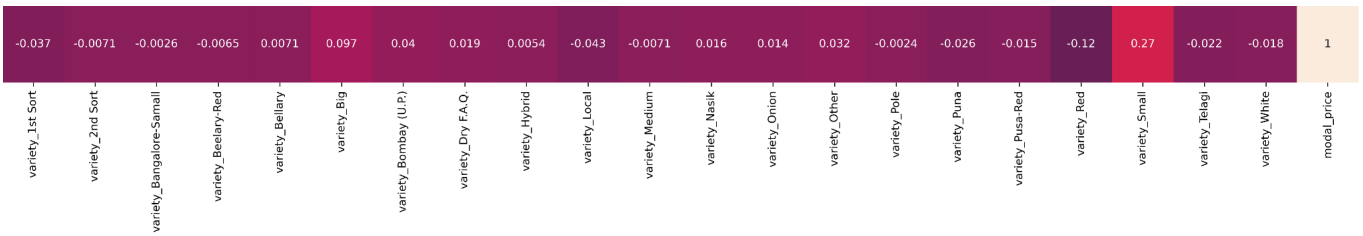
En aquest experiment en optat per fer la correlació del *state* i *variety* sobre el *modal price* ja que és el nostre atribut objectiu.
S'ha comprovat que no hi ha molta relació entre els estat ni les varietat amb el preu model, però hi ha un parell que tenen una relació bastant positiva respecte a les altres.

- La relació lineal a destacar en aquest gràfic és *state kerala* amb un **0.35** de 1 respecte el modal_price.
- Una altra relació a destacar és la *variety small* amb un **0.27** de 1 respecte el modal_price.

Son relacions petites però prou significatives del conjunt de dades perquè ens inidiquen que pot ser el *state kerala* on hi hagi un *modal_price* més pròxim al esperat. Al igual que també ho pot voler dir per la varietat de ceba Small.



[Mapa de calor sobre el preu model]

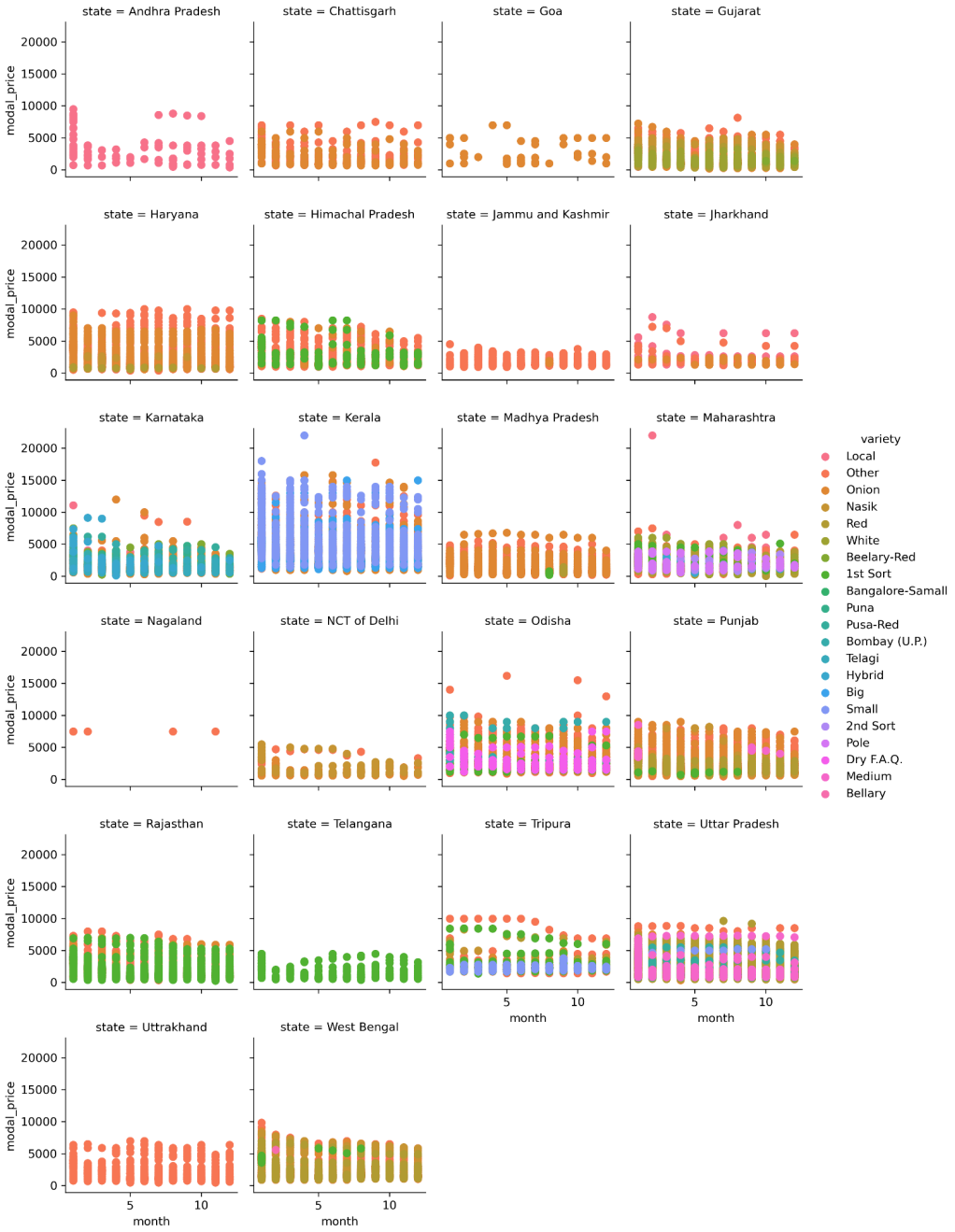


[Mapa de calor sobre el preu model]

Avaluació de paràmetres

En la següent imatge sobre els diferents grafics que obtenim ens dona una idea de com eran les ventes de cada varietat en cada mes sobre cada estat.
Els resultats obtinguts son els següents:

- El gràfic de l'estat **Nagaland**, veiem que no té casi ventes o s'han recollit molt poques dades. Per exemple, a mitjans d'any no hi han vendes de cebes.
- El gràfic l'estat **Kerala** el qual veiem que té moltes vendes i són d'un preu bastant gran. La varietat **Small** es la que ven bastant ja que predomina sobre altres varietat. També es comprova que cada mes de l'any l'estat **Kerala** ven bastantes cebes.



[*Gràfiques sobre cada estat respecte al mes i el preu model de cada varietat]*

Anàlisis

Mean Squared Error (MSE) i Squared Correlation Coefficient (R2)

Un cop normalitzat les dades sobre les qual volem treballar. En el nostre cas, els atributs escollits per saber el mean squared error i el r2 són els següents

- Mínim preu
- Màxim preu
- Mes de l'any
- Dies
- Totes les varietats
- Tots els estats

El MSE amb un error molt petit significa que el model és millor. S'aproxima més al valor esperat, és el preu maxim amb un MSE de **0.03**.
També, el preu mínim també obte un error molt petit, de **0.05**.

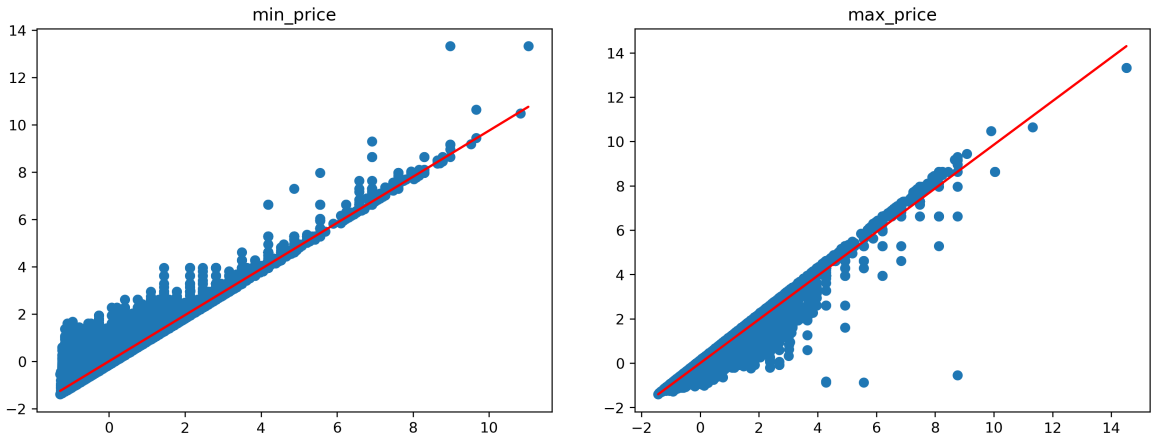
El MSE amb un valor molt gran són els estat de **Jharkhand** i **Haryana** amb un **0.99** ambos.
Respecte a la varietat **Bangalore-Samall** i **Pole** també tenen un valor molt gran de MSE, és del **0.99** ambos

A continuació comprovem Squared Correlation Coefficient (R2), el cual el valor més gran signifca que la predicció será més bona. En aquest cas, els atributs que son millor per predir són el **máxim preu** i el **mínim preu** amb un **0.95** i **0.97** respectivament.

Model de predicció

En la primera imatge ens mostra el model segons les dades normalitzades del preu mínim (x) amb el preu model (y). La línia en vermell ens mostra la predicció que s'ha fet sobre el model. Com la predicció és optima ens demostra que el nostre model està ben escollit.

Hem fet el càlcul sobre el máxim preu. També es comprova que la predicció és correcte.



[*Gràfiques sobre la regressió lineal de min price i max price segons el model price]*

Regressions lineals

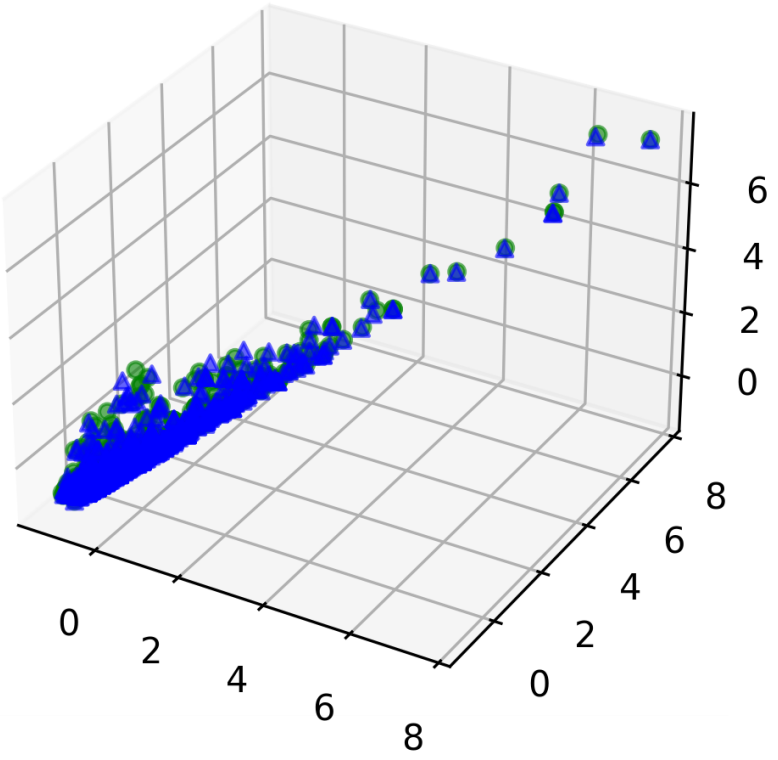
Hem desenvolupat 4 regressions lineals per predir el modal_price. Per valorar els resultats de les regressions hem dividit les dades en train i test. Amb el train hem creat el model i amb el test l'hem provat visualitzant el MSE.

Hem provat 4 combinacions. Primer sol utilitzem el min_price i el max_price, en aquesta prova dona un més elevat, aproximadament 18000 que eliminant l'elevat dona una mitja de 134 d'error a les prediccions.

Considerant que els valors del preu normalment oscil·len entre 500 i 4000 (sense considerar valors extrems) és un error assequible.

Les següents proves van donar resultats similars, les proves van ser amb min_price, max_price i els valors de variety. Després amb min_price, max_price i els valors de states. Per últim amb min_price, max_price, valors de states i els valors de variety.
En les 3 últimes proves els resultats variaven molt segons l'execució del model, a vegades donava un MSE de 13000-17000 i altres vegades donaven de 0,5-10.

En aquests casos milloraven els resultats respecte a la primera combinació, sobretot en el cas del MSE menor que 10, que suposa un error de 3,16. Que considerant el nostre cas és un error pràcticament nul.



Problemes trobats

Al principi vam tenir dificultats per veure quines variables es podien utilitzar en la regressió, ja que la majoria de les dades eren categòriques i no sabíem com tractar-les.

Un altre problema va ser tenir 2 variables amb un score R2 molt elevat (min_price i max_price) comparat amb les altres, ja que en el millor dels casos aconseguirem un score de 0,4 i no sabíem si sol utilitzar les dues variables o utilitzar més. Al final vam decidir fer models pels dos casos.