Práctica 1: Regressió

Aprenentatge Computacional (102787) - Grau en Enginyeria Informàtica [MO52745]

Enrique Gómez Becerra - 1566725

Borja Arcos Comas - 1568307

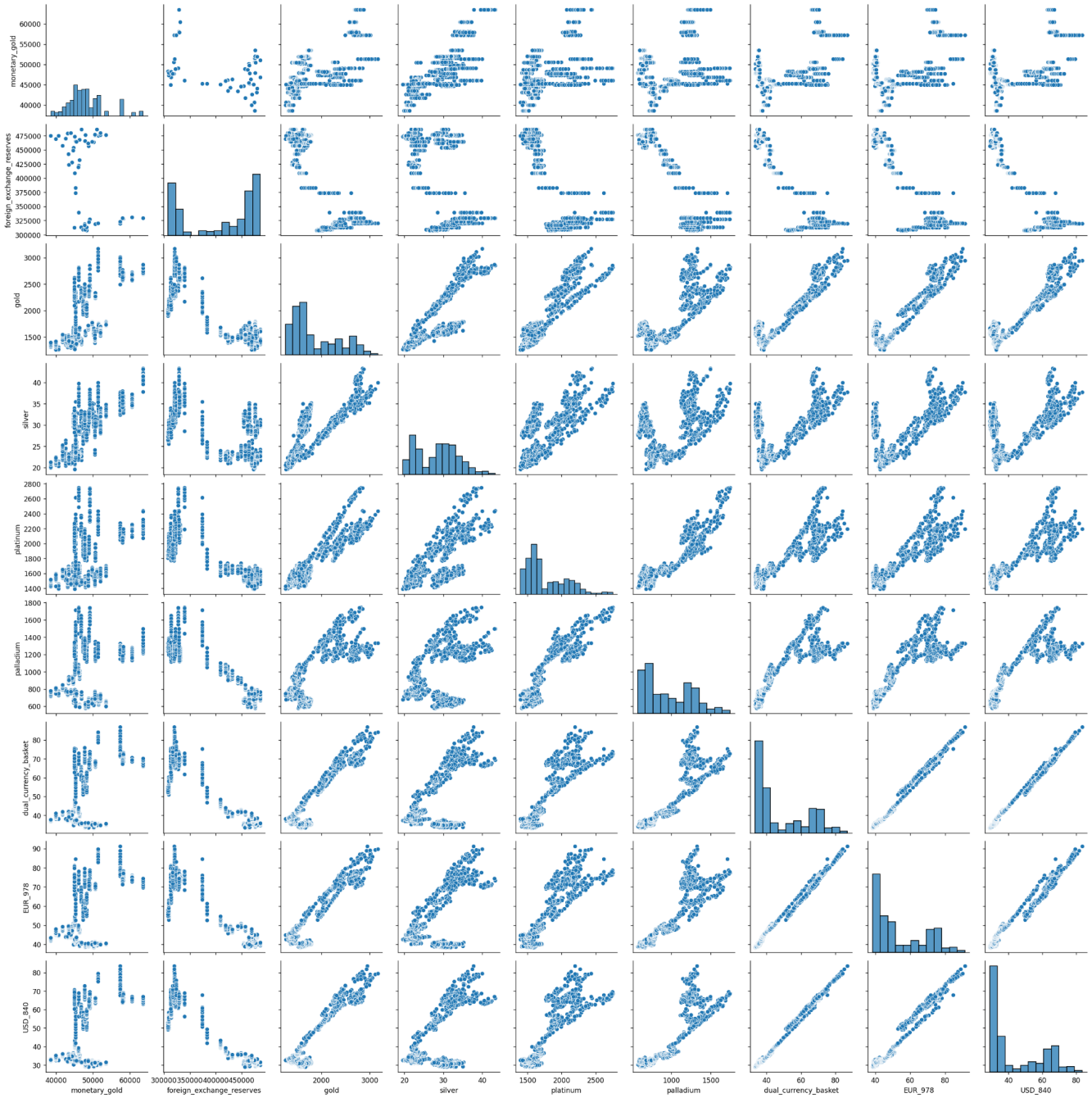
Javier Méndez Leiva - 1496052

Index

Introducció

En aquest document es detallen els anàlisis de dades fets en base als indicadors financers rusos. El nostre análisis es centrará en el valor del ruble front a altres monedes com l'Euro i el USD en base als valors financers de les exportacions minerals d'or, plata, pal·ladi i platí, també prenen en consideració les reserves d'intercanvi estranger i les reserves d'or del país.

És per aquest motiu que altres dades com poden ser els valors monetaris d'altres divises han sigut omesos i deprecats per tal de no fer d'algunes prediccions una trivialitat.



Objectius

Amb aquest projecte el que es busca és posar en pràctica els conceptes estudiats a classe i aprendre a aplicar models de regressió sobre dades reals. A més, serà necessari assolir el coneixement necessari per aplicar les tècniques de regressió i anàlisi necessàries per poder treure els resultats desitjats, tot amb la búsqueda d'informació per internet i el nostre propi criteri. Per acabar, respondre les preguntes que es plantegen dins del notebook que es dóna com a guia per realitzar els gràfics, a partir de les conclusions que es poden extreure.

Descripció de la base de dades

Dintre de la base de dades, es poden diferenciar 3 tipus de paràmetres els paràmetres de la tresoreria actual del país com poden ser el valor monetari en or de les reserves i les reserves per al comerç exterior.

Per altra banda tenim el valors dels minerals exportables com poden ser l'or, la plata, el pal·ladi i el platí.

Finalment, tenim el valors de divises sobre el ruble, com poden ser l'euro, el dòlar nord-americà, el i en japonés, el yuan xinès, la rupia índia i la brazil real coin.

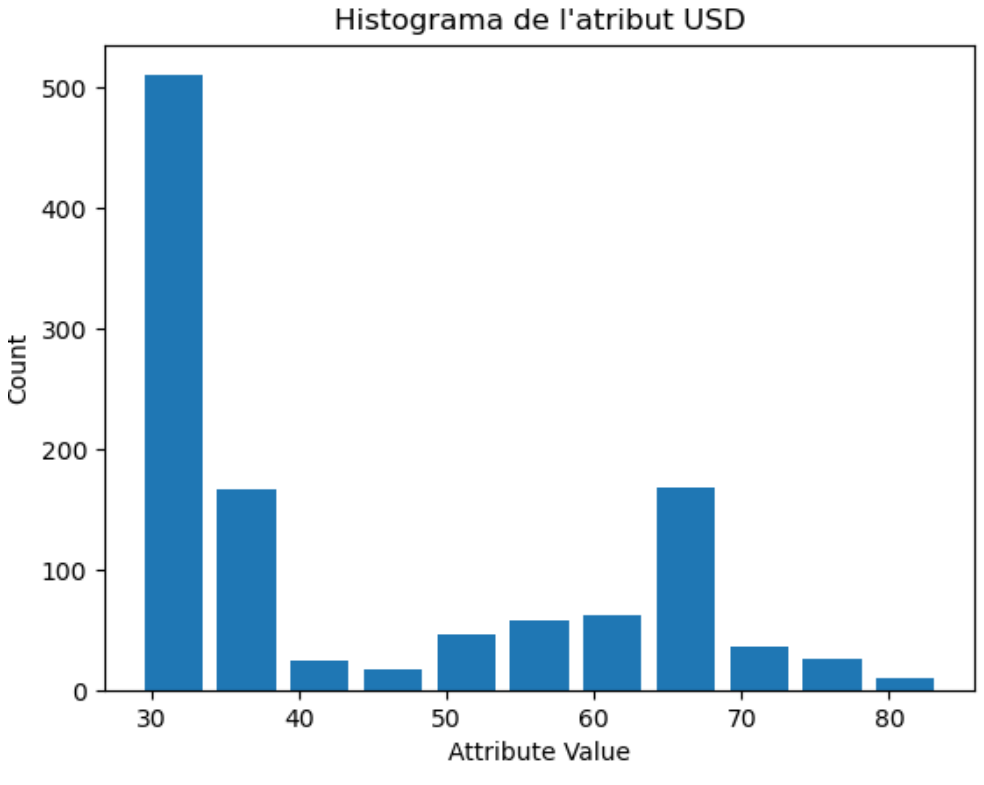
D'aquests valors i per tal de fer les prediccions més significatives i els anàlisis mes centralitzats sobre l'impacte dels minerals en l'economia s'ha centralitzat l'anàlisi en l'euro i el dòlar nord-americà, netejant el nostre dataset de dades com les altres divises.

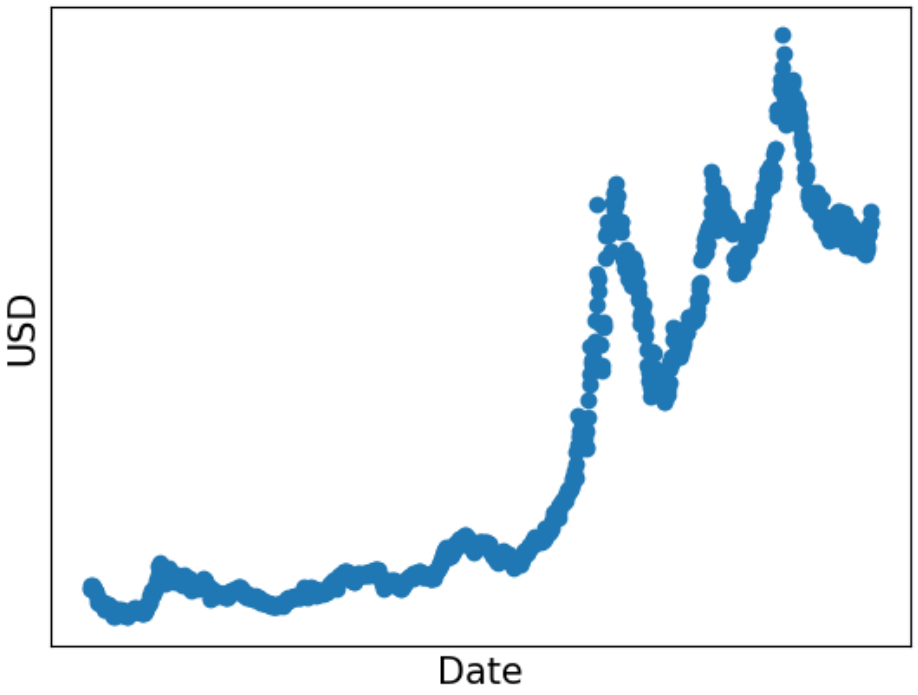
A l'hora de suprimir altres dades vam trobar que, a la nostre base de dades, no hi havia cap columna que pogués ser una molestia de cara a la realització de la pràctica. Això es deu a que existeixen valors per tots els atributs i, ademés aquests valors son correctes. És a dir, no hi ha cap columna/valor 0 o NaN. Per tant, s'ha concluit que aquests valors i atributs son els indicats per començar el nostre anàlisi i no cal eliminar, fer "neteja" o fer una replica de valors mitjans. Això es pot veure en la següent imatge.

```
print("Per comptar el nombre de valors no existents:")
print(dataset.isnull().sum())
```

```
Per comptar el nombre de valors no existents:
date                          0
monetary_gold                 0
foreign_exchange_reserves     0
gold                          0
silver                        0
platinum                      0
palladium                     0
dual_currency_basket          0
EUR_978                       0
USD_840                       0
JPY_392                       0
CNY_156                       0
INR_356                       0
BRL_986                       0
```

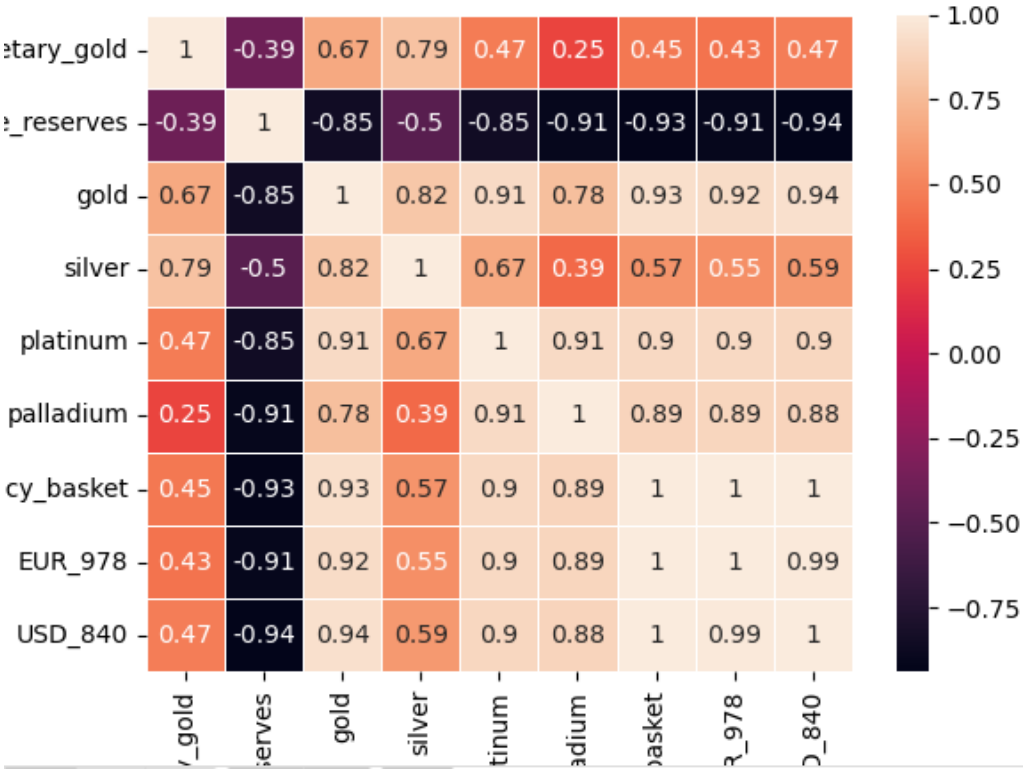
Per tal de comprendre la nostra dada principal, el dòlar nord-americà, es va realitzar un histograma analitzant la frecuencia d'aparició d'un valor conjuntament amb un histograma de l'aparició del valor durant el temps, el que ens mostra els creixements econòmics dels valors de rusia durant períodes de temps concret i ens deixa veure com durant el temps en que el ruble va experimentar el creixement que observem de cara al final del mostreig, el valor que sembla una excepció en el primer histograma es mostra com un valor més que comprensible en el segon.





Gràfic de correlacions

El gràfic mostrat a continuació ens indica les correlacions entre tots els atributs de la nostra base de dades.

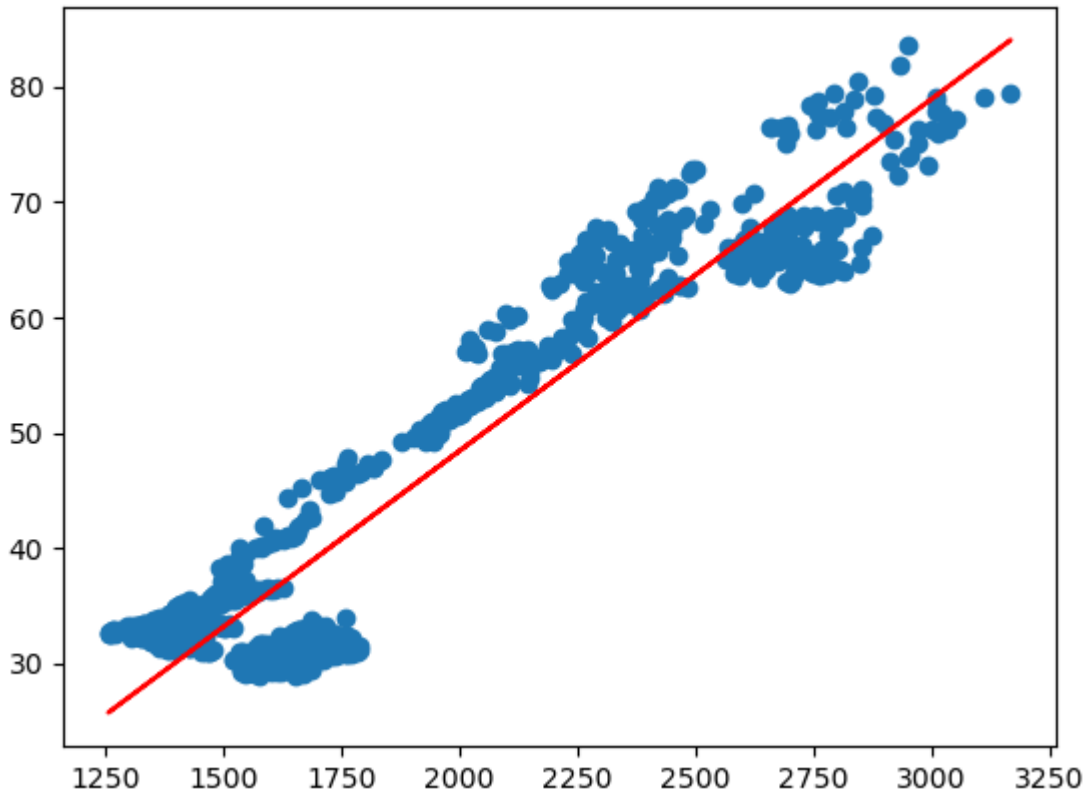


S'observa que en general existeix una correlació alta entre gran part dels atributs, a excepció de '*foreign_exchange_reserves*' que té una correlació negativa amb tota la base de dades. També és necessari comentar que els atributs '*monetary_gold*' i '*silver*' obtenen valors de correlació bastant baixos en comparació amb la resta.

Anàlisi de les dades

A continuació es veurà una sèrie de diagrames de dispersió que relacionen els atribut '*USD_840*' i '*EUR_978*' amb la resta d'atributs utilitzats al nostre model de dades, és a dir, els pertanyents als valors dels minerals (or, plata, platí i pal·ladi). A partir d'aquests diagrames es pot deduir el nivell de correlació entre els atributs i fer prediccions sobre quin d'ells aporta un menor error quadràtic mitjà a la regressió.

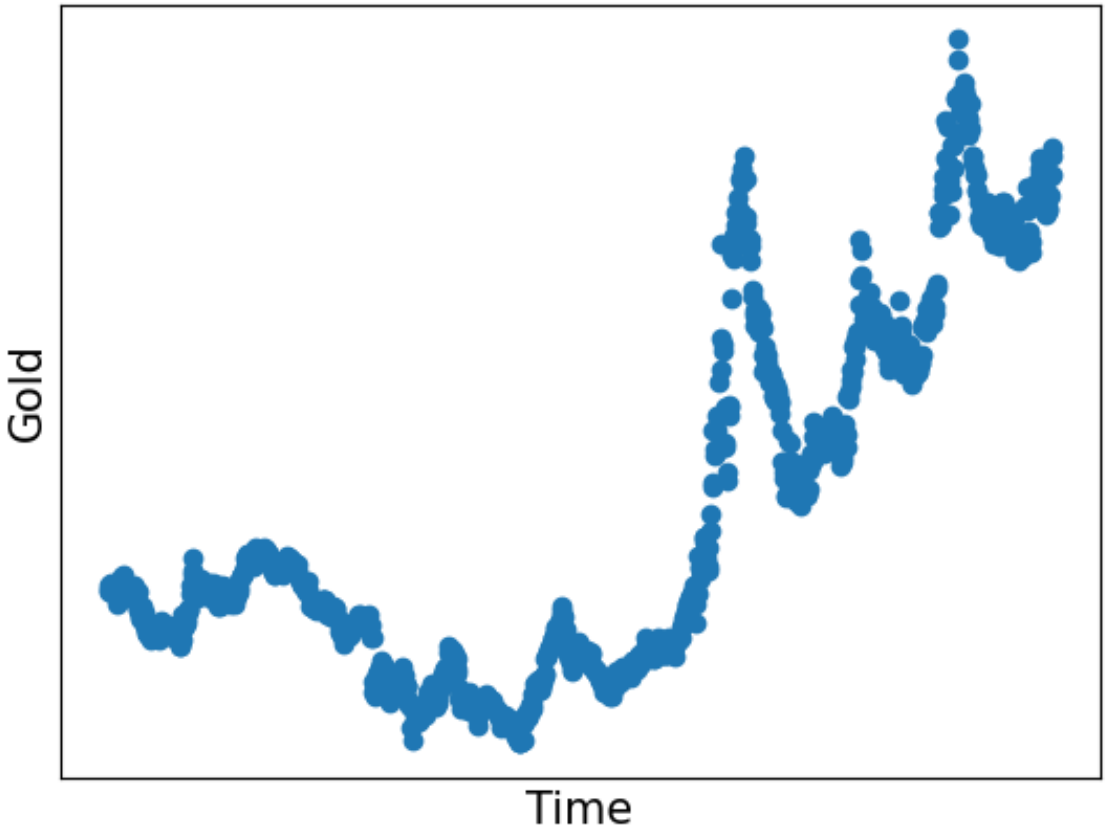
USD - GOLD



Mean squeared error: 28.295223274898426

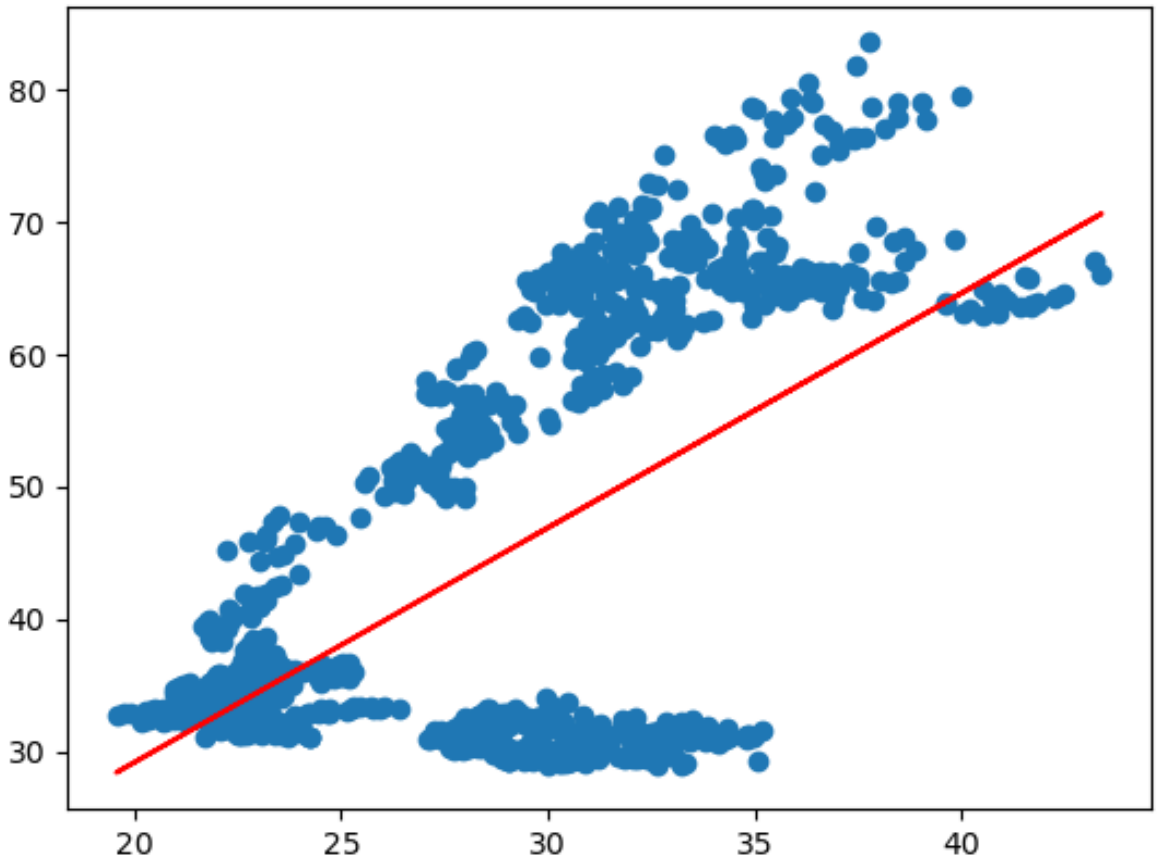
R2 score: 0.8825792585706665

En aquest cas s'observa una correlació positiva bastant evident. Molts dels valors s'allunyen lleugerament de la recta de regressió i aquesta petita distància es reflexa en un MSE de 28,3 i un R2 de 0,88. El coeficient de determinació (R2) és alt i, tal com es pot veure a continuació, és el major de tots els resultats.



Aquesta correlació es fa evident en l'evolució temporal de l'or sobre la del dolar, on es pot observar com presenten estructures molt similars, denotant una forta connexió entre els dos valors i indicant-nos l'impacte de l'or dintre de l'economia rusa.

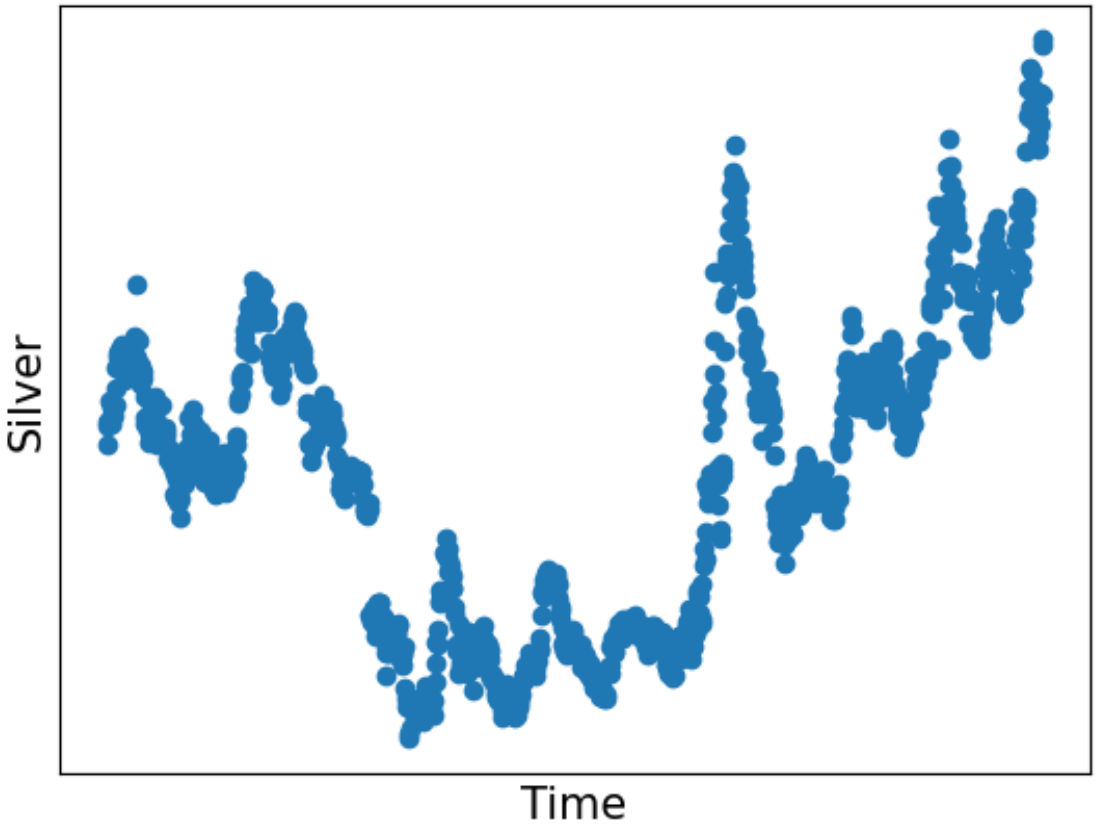
USD - SILVER



Mean squeared error: 156.1390525860088

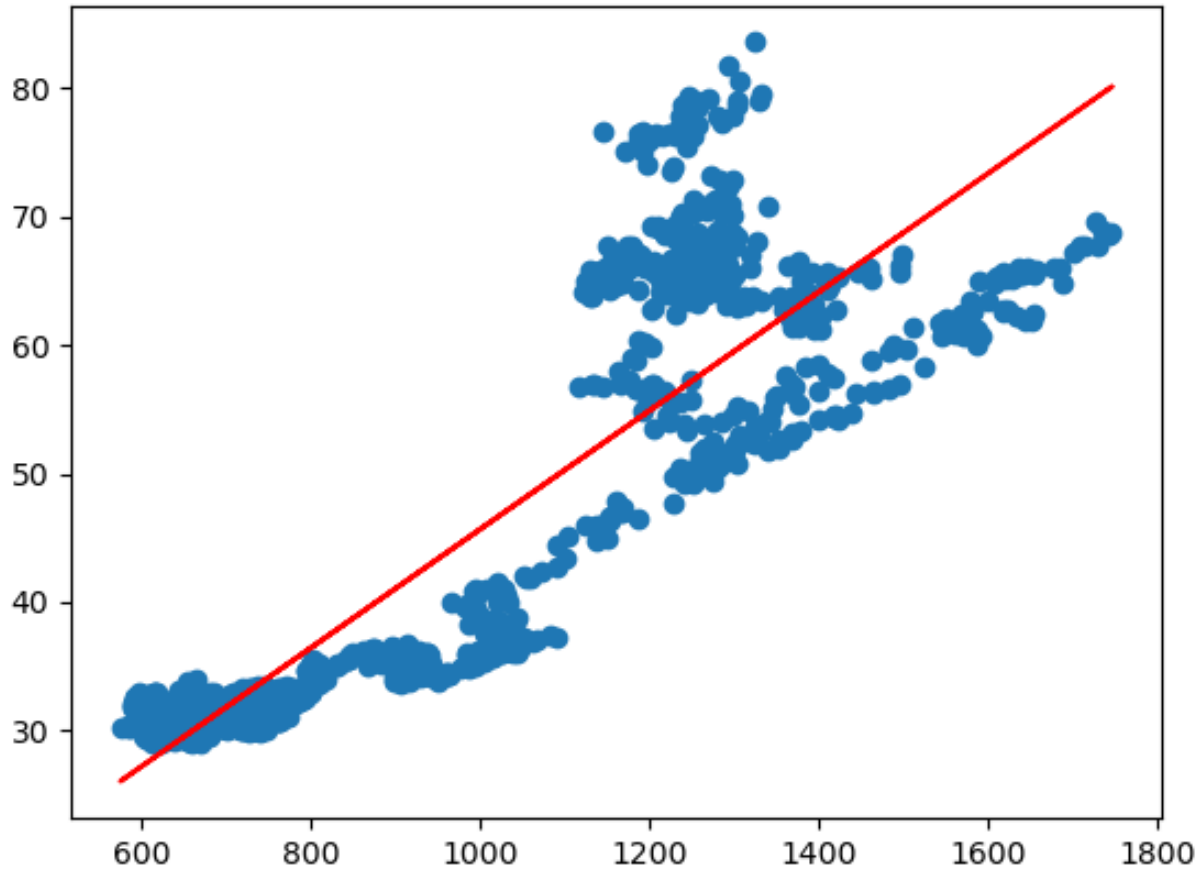
R2 score: 0.35204740593132355

En aquest cas la correlació positiva no és tan evident. La majoria dels valors s'allunyen significativament de la recta de regressió, el que porta a un MSE de 156,14 i un R2 de 0,35. El coeficient de determinació en aquest cas és bastant baix. Ja es pot deduir que aquest no será el valor idoni per fer una predicció.



A nivell d'evolució temporal es pot observar que en les etapes primeres del histograma temporal la plata va tenir uns valors molts majors que els del valoren USD donant origen als valors més allunyats de la recta de regressió verticalment, també en la part final es pot observar que els creixements del valor del USD no acaben de correspondre amb els de la plata, donant lloc als valors més inferiors de la gráfica de regressió mostrant així la justificació als valors allunyats de la regressió i implicants sobretot amb la primera meitat del graf que la correlació i l'impacte de la plata sobre la economía no es gaire significativa front a altres minerals.

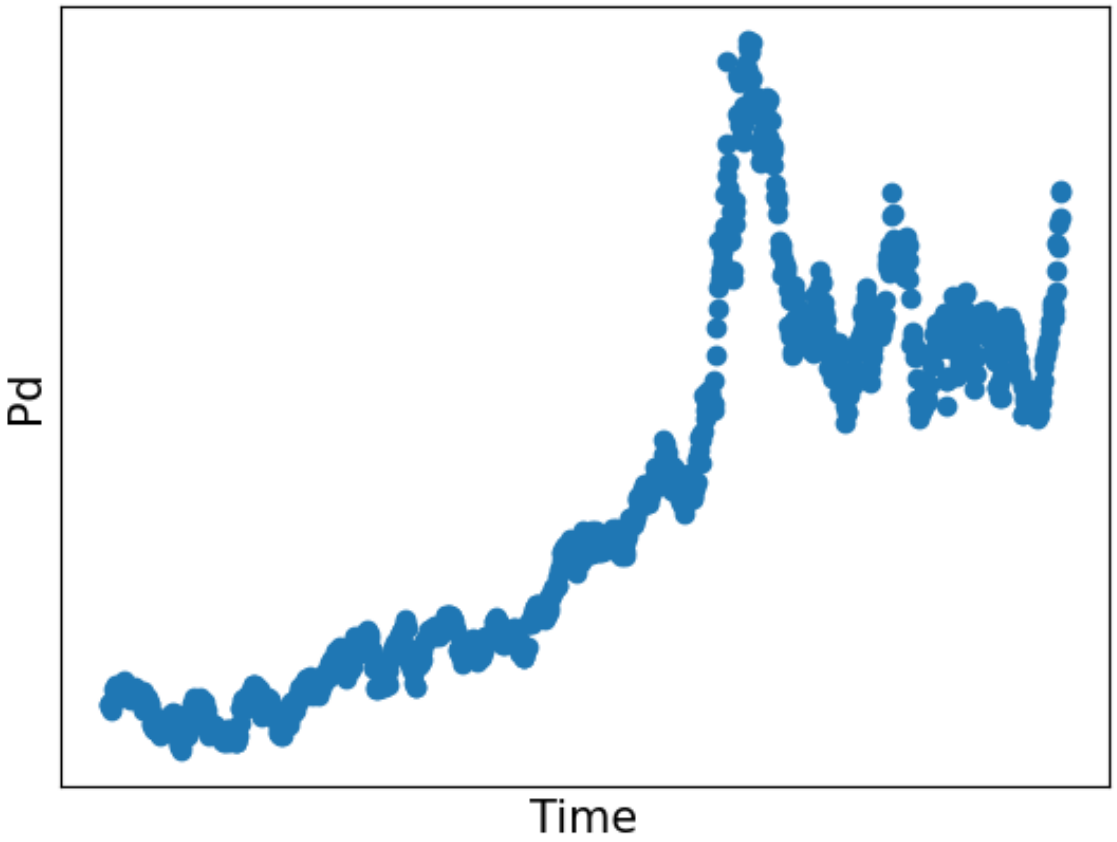
USD - PAL·LADI



Mean squeared error: 47.26102811897045

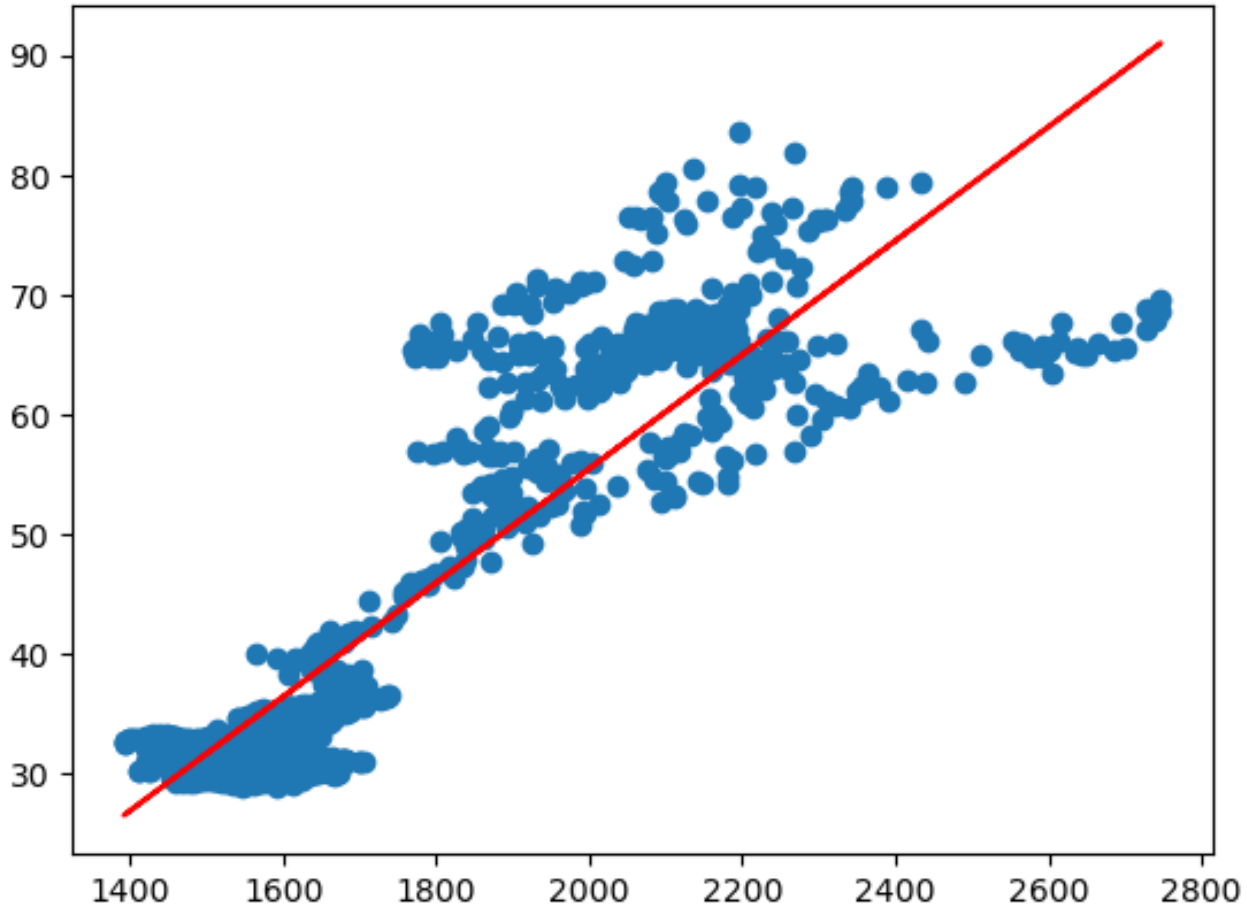
R2 score: 0.8038741412807597

De forma similar al que s'ha vist anteriorment, la correlació positiva no és tan evident. La majoria de valors s'allunyen de forma significativa de la recta de regressió, tot i que en menor mesura en comparació amb el cas anterior, el que porta a un MSE de 47,26 i un R2 de 0,8. El coeficient de determinació en aquest cas és elevat, però menor que al primer cas. Un fet molt evident és com en els valors més alts es tornen, menys correlació mostren, fet significatiu que pot indicar que no marca tant el flux de l'economia.



Un cop més es pot trobar una gráfica de evolució molt similar a la del valor del ruble en USD, en aquest cas, tenim un clúster de punts en la gráfica de regressió que despunta molt sobre la resta, i en aquest cas en l'histograma es pot observar com el primer pic que es troba en el valor del pal·ladi es molt mes pronunciat, donant potencialment origen a aquest clúster de valors, cal destacar que mostra una gran correlació en la gran majoria de valors i que podría significar que l'economia en veritat es troba molt impactada pel valor del pal·ladi però per altra banda els valors alts no semblen reflexar-se tant en l'histograma del valor del ruble en USD.

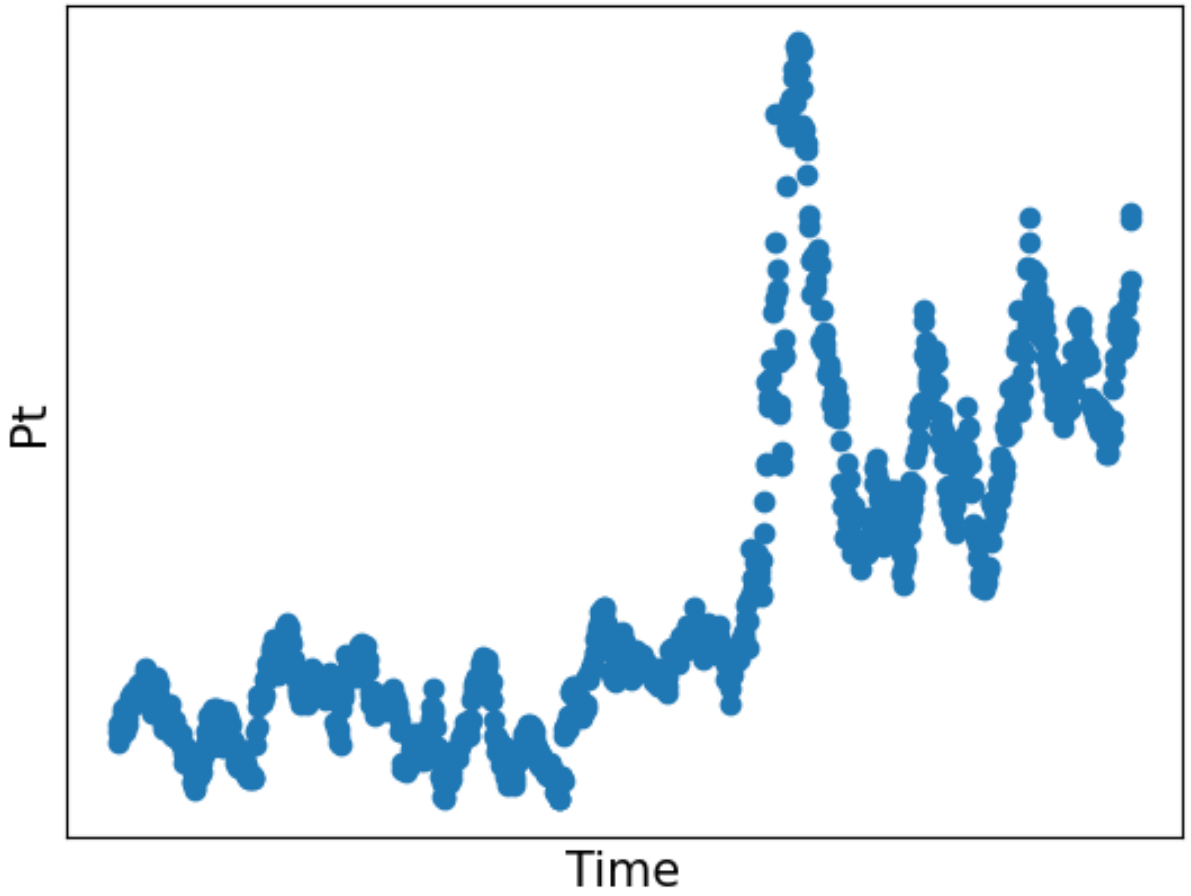
USD - PLATÍ



Mean squeared error: 54.50392937098879

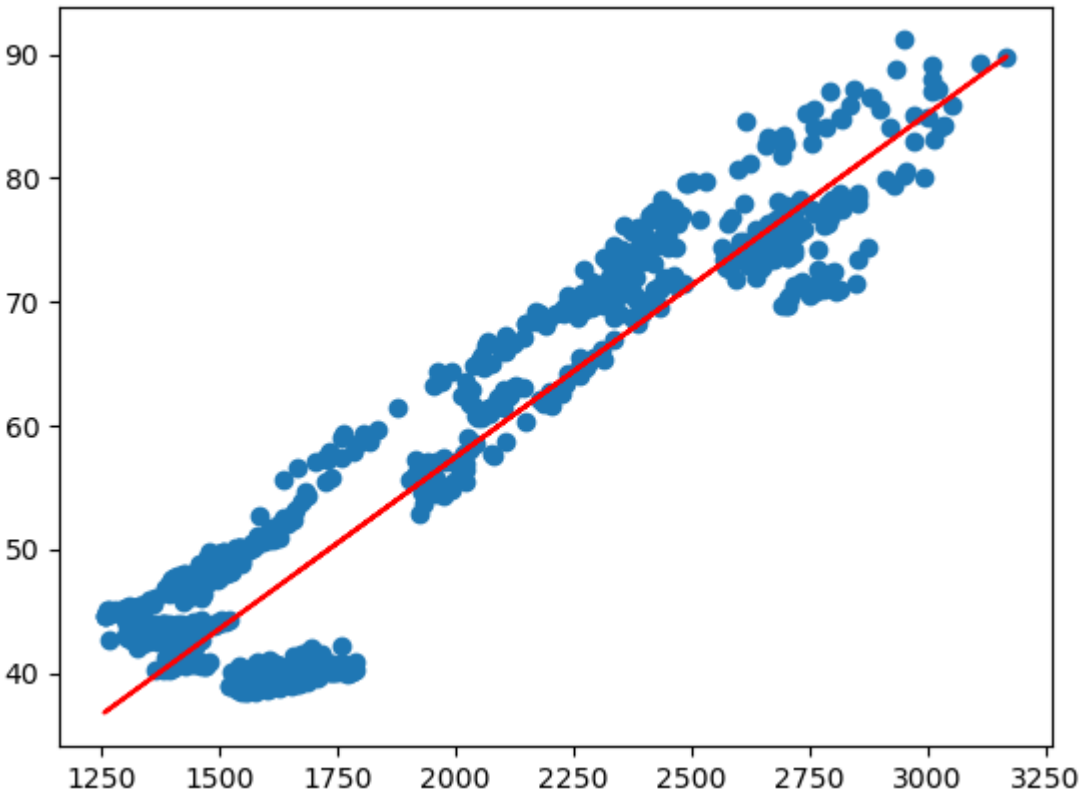
R2 score: 0.7738172363802808

La correlació observada en aquest cas és positiva. Gran part dels valors s'allunyen de la recta de regressió, i això provoca l'obtenció d'un MSE de 54,5 i un R2 de 077. El coeficient de determinació en aquest cas és elevat, però tot i així no supera els resultats obtinguts al primer cas. Un cop més encara que de forma menys pronunciada, cap tenir en compte que els valors a més elevats menys correlació mostren, cas repetit amb la gráfica del pal·ladi.



Finalment, es troba l'evolució del platí en el temps, en aquest gràfic es pot observar una correlació molt més propera a la de l'or i la del valor del ruble en USD coincidint en semblança amb el pal·ladi, on el pic és molt més pronunciat i no es reflexa tampoc en l'economía rusa, indicant que es posible que aquests minerals tinguin un impacte dintre de l'economía pero no que marquen el flux de la mateixa, per altra banda cal identificar que en les primeres parts de la gráfica es poden trobar valors generalment més positius que els del ruble en USD, un altra factor a tenir en compte i que es pot veure reflexat en la gràfica de regressió és com el cluster inferior de punts de la part dreta de la recta de regressió potencialmente correspon històricament amb l'espai entre el segón i tercer pic de l'histograma, degut a la significant caiguda que aquest ens mostra.

EUR - GOLD

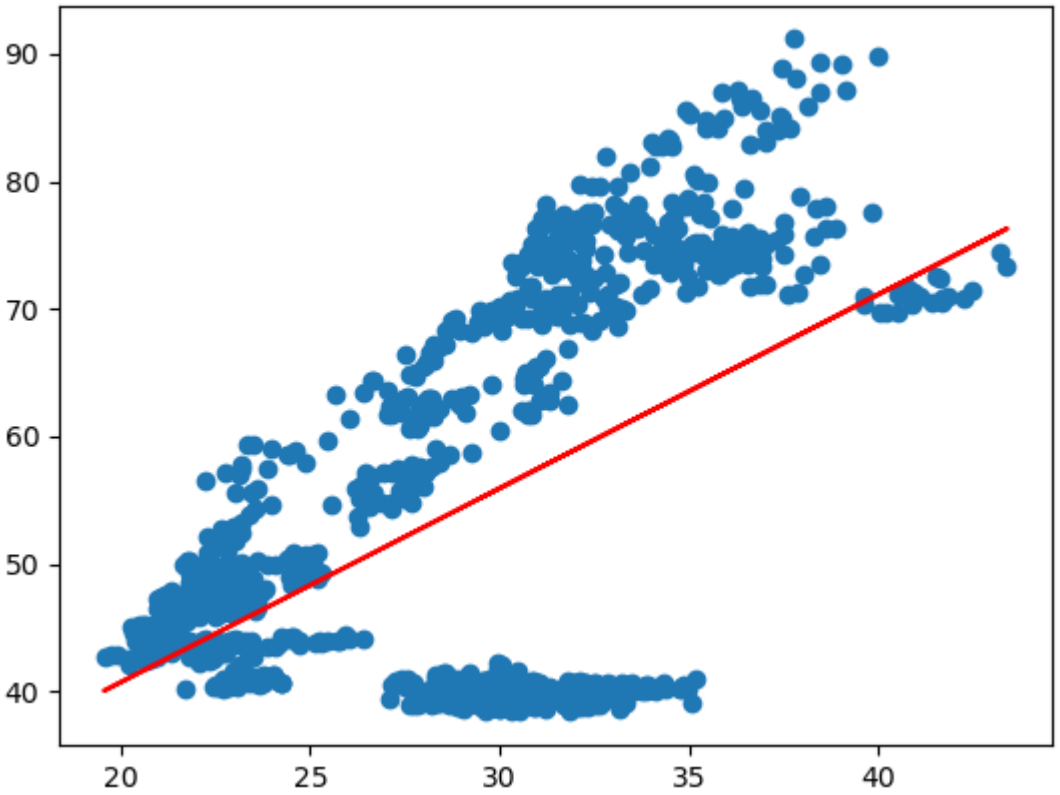


Mean squeared error: 33.33118423866212

R2 score: 0.8407692814141479

En aquest cas s'obté una correlació significativament positiva molt evident. El valors no s'allunyen significativament de la recta de regressió i es pot veure això representat amb un MSE de 33.33, també es pot observar que l'error no es tampoc un mal valor amb un 0.84. Obtenint d'aquest mode un dels millors coeficients de determinació (R2).

EUR - SILVER

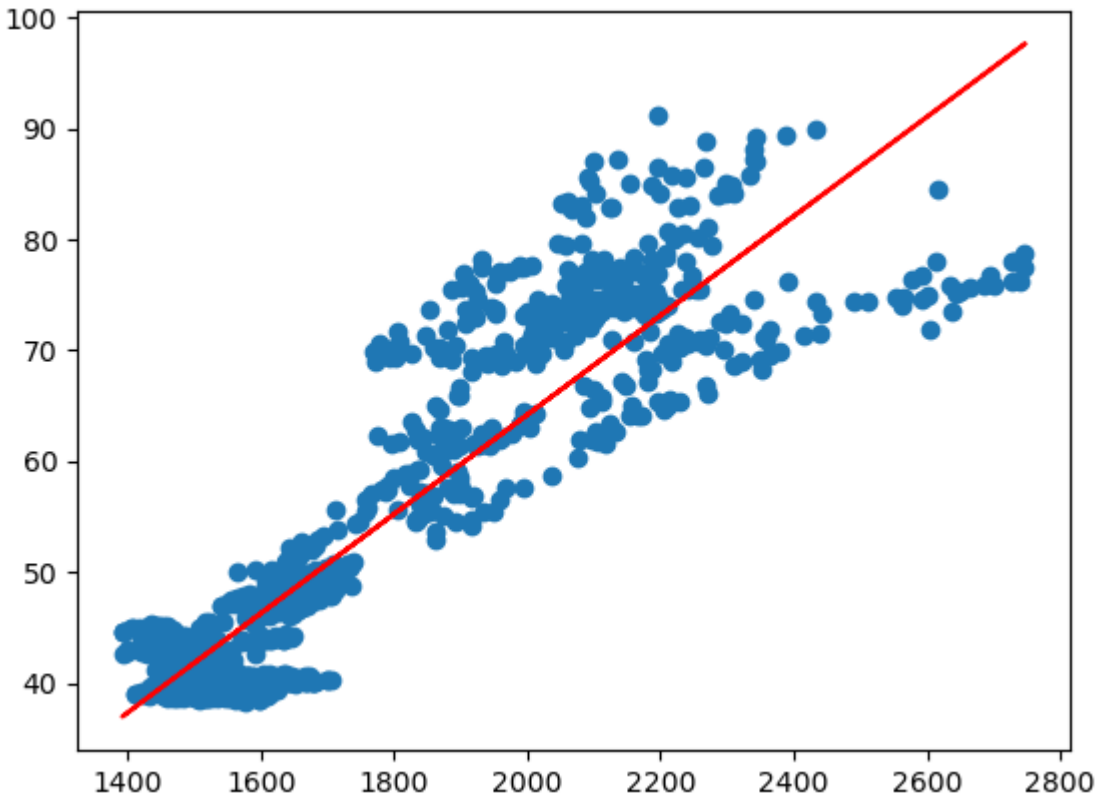


Mean squeared error: 146.87812901171648

R2 score: 0.29832946049504616

Un cop més la plata aporta una correlació més cuestionable amb poc punts propers a la recta i alguns molt allunyats de la mateixa, el MSE es de 146,88, un valor que mostra lo desproporcionadament allunyats que estan la major part dels valors a la recta de regressió, per altra banda, el coeficient de determinació, és quasi 0.3 un valor molt baix que ens mostra la poca correlació entre les dues dades i ens mostra un cop més l'extrany comportament de la plata front a l'economia rusa.

EUR - PLATÍ

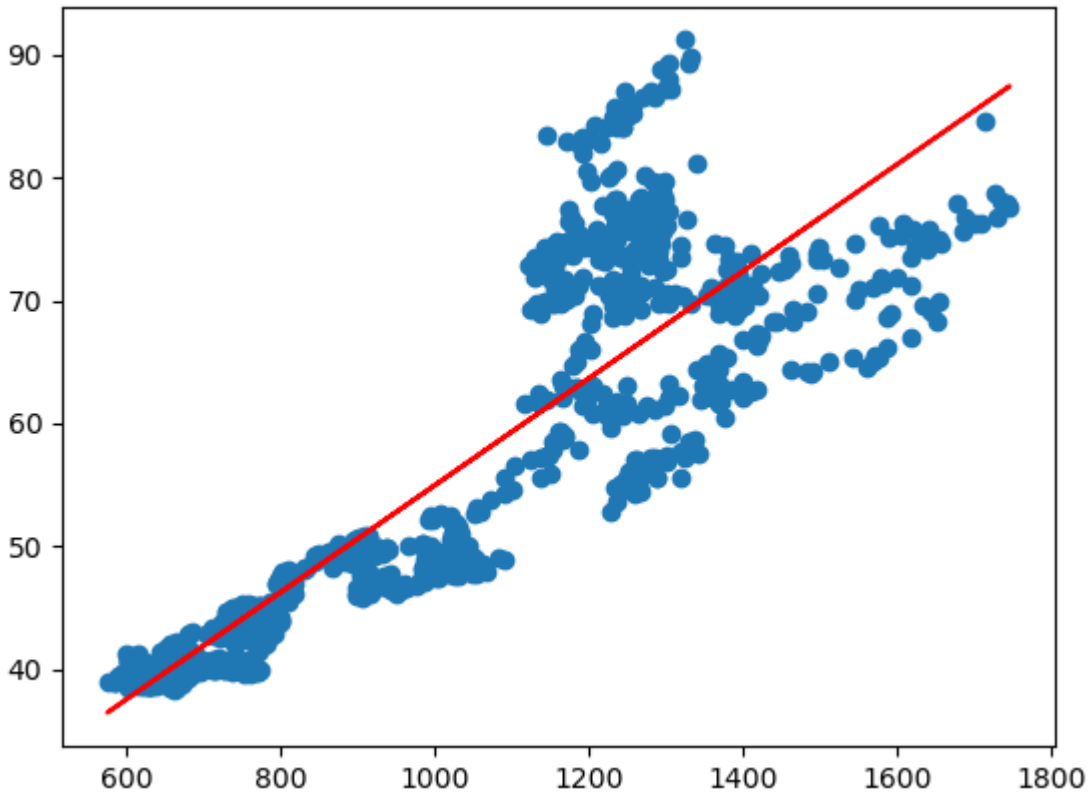


Mean squeared error: 38.47189863749971

R2 score: 0.8162109086329592

En aquest diagrama es pot observar una regressió majorment positiva, un altre factor a veure que tenim una gran quantitat de valors desviats de la recta de regressió, cosa que dóna lloc a un pitjor rendiment en el MSE (38,47) de cara al de l'or, i es pot veure que en els valors baixos tenim un molt bon clúster amb moltes dades properes a la regressió, el que compensa les dades altes patint una major desviació, això es reflexa amb el coeficient de correlació (R2) que es manté en 0.82, un molt bon valor.

EUR - PAL·LADI



Mean squeared error: 43.824792753348135

R2 score: 0.7906389046358181

En aquest diagrama es pot apreciar l'existència d'una correlació positiva. La correlació és poc clara, ja que molts dels punts es troben a gran distància de la recta de regressió. És observable que alguns dels valors per sobre de la recta presenten valors molt diferents a la resta del conjunt i són aquests els que obtenen distàncies més elevades com a resultat. Aquesta distància amb la recta té una influència als resultats, el que porta a un MSE de 43,83 i un coeficient de determinació (R2) del 0,79. El coeficient R2 arriba a un valor prou elevat, però no millora els resultats obtinguts amb l'or.

Respostes a les preguntes

Amb els resultats obtinguts anteriorment, es respondran les preguntes plantejades en els diferents apartats de la pràctica:

Apartat C:

1. **Quin és el tipus de cada atribut?**
   Tots son atributs numèrics a excepció de l'atribut date que és alfanumèric. Per la resta dels atributs; monetary_gold i foreign_exchange_reserves són valors enters, mentre que la resta són valors reals.

2. **Quins atributs tenen una distribució Gaussiana?**
   D'acord amb els histogrames dels diferents atributs, observem que no existeix cap atribut amb una distribució Gaussiana.

3. **Quin és l'atribut objectiu? Per què?**
   Hem escollit com a atribut objectiu USD_840. Aquest atribut fa referència als valors de conversió del dòlar al ruble i l'hem escollit perquè pren una gran varietat de valors reals que ens poden ser més útils que no un valor enter o binari. Per acabar, cal dir que el dòlar pertany a la 1a potència mundial EEUU, per tant és la moneda més important i amb la que cal comparar-se.

Apartat B:

1. **Quin són els atributs més importants per fer una bona predicció?**
   Els valors que són més propers a la recta de regressió i que tenen un valor de error menor.

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

3. **Quina correlació hi ha entre els atributs de la vostra base de dades?**
   Hi ha una correlació positiva a excepció del segon atribut. Aquest té una correlació negativa amb tots els atributs de la base de dades. Per tant aquest atribut no té un impacte en el mercat del pais, contràriament a l'or que és el

major que es veu com es podia preveure. Per altre banda, tenim els altres materials minerals. El pal·ladi i el platí tenen una correlació molt alta amb el ruble, molt semblant a la de l'or tot i que essent una mica inferior. Per contra, la plata no té un impacte tant significatiu.

4. **Com influeix la normalització en la regressió?**
   Evita que els valors anòmals tinguin una gran influència a la predicció.

5. **Com millora la regressió quan es filtren aquells atributs de les mostres que no contenen informació?**
   La predicció experimenta un creixement en la seva precisió.

6. **Si s'aplica un PCA, a quants components es redueix l'espai? Per què?**
   Per predir el valor del USD (que es l'atribut escollit per fer la regressió) aplicant un PCA, es redueix a dos components: or i platí. Hem pensat que és lo apropiat per aquesta situació, ja que són els dos valors de correlació més alts que hi ha a la taula sense tenir en compte les altres divises(no veiem apropiat utilitzar-les ja que quan una divisa creix les altres també degut al comerç).

Conclusions

Arrel dels anàlisis prèviament esmentats, s'han extret múltiples conclusions. És cert que l'impacte dels minerals esmentats sobre l'economia rusa es significatiu fins a cert extent, i s'ha concluit que fins a cert punt es podria justificar un model predictiu amb dos valors com són el or i el platí, però també tenim la idea de que fins a cert punt es posible que la economía sigui qui ha dictaminat els rendiments dels minerals que no son or, degut a que quan l'economía més s'accentúa, és quan més pronunciats son els pics d'aquests i a causa d'això obtenim majors desviacions en les nostres regressions, causant repressions majors també quan l'economia es resigna degut a possiblement retalls en l'extracció dels mateixos per millorar l'impacte i el creixement econòmic, encara amb això, per reforçar aquesta teoria faltarien més dades i seria només un anàlisi especulatiu de les dades en base al seu comportament.