# Tipologia i cicle de dades · PRACTICA 2 · 2017-2018

EEES · Màster de data Science

Relació i patrons entre moviments de índexs de borsa i cryptomoneda

**Nom i Cognoms:** Albert Costas Gutiérrez

## UOC practica 2. Tipologia i cicle de vida.

## Descripció

Datasets per la comparació de moviments i patrons entre els principals indexs borsatils espanyols i les crypto-monedes

| **Links** | **Fitxers** |
|---|---|
| Repositori github: https://github.com/acostasg/scraping<br><br>Repositori kaggle Open data:<br>• https://www.kaggle.com/acostasg/stock-index<br>• https://www.kaggle.com/acostasg/crypto-currencies<br>• **https://www.kaggle.com/acostasg/cryptocurrenciesvsstockindex** | • Document PDF amb les respostes de les preguntes i els noms dels components del grup.<br>• Fitxer amb el codi **Pyhton** per obtenir les dades<br>• Fitxer **R** amb correlació d'atributs<br>• Carpeta CSV amb les dades |

# Estructura

```
scraping
├── pdf
│   ├── acostasg-PRACTICA_1.pdf  # Document pdf de la practica 1, components grup
│   └── acostasg-PRACTICA_2.pdf  # Document pdf de la practica 2, components grup
│
├── csv  # datasets
│   ├── cypto_currencies
│   │   └── ... # fitxers csv
│   ├── stock_index
│   │   └── ... # directoris per data amb els csv
│   └── dataset
│       └── dataset.csv # dataset preparat per al script R unifica els anteriors
│
├── projects
│   ├── scraping_crypto_currencies.py # scraping url criptomoneda
│   ├── scraping_stock_indexs.py # scraping url el economista
│   └── cleanAndTransform.py # script para limpiar i unificar en un dataset
├── R
│   ├── matriu_de_correlacio_index_borsatils.xlsx #matriu de correlació dels index borsatils
│   └── script.r # script R  i amb correlació d'atributs i model
│
├── README.md
├── cleanData.py #fitxer python per netejar, unificar i transformar les dades
├── scraping.py # fitxer python incial
└── setup.py
```

# # Script R

* Anàlisis de les observacions i el domini de les dades
* Anàlisis en especial de Bitcoin i la IOTA.
* Test de Levene per veure la homogeneitat
* Kmeans per creació de cluster per veure la homegeneitat
* Freqüències de les distribucions
* Test de contrast d'hipòtesis de variables dependents (Wilcoxon)
* Test de Shapiro-Wilk per veure la normalitat de les dades, per normalitzar-les o no
* Correlació d'índexs borsatils, per eliminar complexitat dels índexs amb grau més alt de correlació
* Iteració de Regressions lineals per obtenir el model amb més qualitat, observa'n el p-valor i l'índex de correlació
* Validació de la qualitat del model
* Representació gràcia

# Autors

Albert Costas Gutierrez - acostasg@uoc.edu

# Llicència



# Fonts de dades

- http://www.eleconomista.es
- https://coinmarketcap.com

Les dades de borsa i crypto-moneda estan en última instància sota llicència de les webs respectivament.

## Respostes a les preguntes

### 1. Descripció del dataset

Perquè és important i quina pregunta/problema pretén respondre?

En aquest hi ha 2 datasets, els quals netejarem i transformarem en un únic dataset amb l'objectiu de poder **comparar** en el mateix període de temps si hi ha **relació o es podreixen patrons** comuns entre els **moviments borsatils** dels principals indexes espanyols i els **moviments de les crypto-monedes**.

En aquest cas **el «trading» en cryptomoneda** és relativament nou, força popular per la seva formulació com a mitja digital d'intercanvi, **utilitzant un protocol que garanteix la seguretat, integritat i equilibri** del seu estat de compte per mitjà d'un entramat d'agents.

En aquest cas el context és detectar o preveure els **diferents moviments que es produeixen per una sèrie factors**, tant de moviment interns (compra-venda), com externs (moviments polítics, econòmics, etc...), en els principals índexs borsatils espanyols i de les crypto-monedes.

Hem seleccionat diferents fonts de dades per generar fitxers «csv» i **guardar diferents valors en el mateix període de temps**. És important destacar que ens interessa més les tendències alcistes o baixes que el volum de les transaccions.

Cal destacar per altra banda, que en el nou dataset un cop netejades les dades passàrem **totes les monedes a Euros, ja que en la cryptomendes està en dolors, ho farem el script de neteja de python.**

La comunitat podrà respondre, entre altres preguntes, a:

Està afectant o hi ha **patrons comuns** en les cotitzacions de cryptomonedes i el mercat de valors principals del país d'Espanya?

Els efectes o agents **externs afecten per igual a les accions o cryptomonedes**

Hi ha **relacions cause efecte** entre les acciones i cryptomonedes?

## 2. Neteja de les dades

### 2.1. Selecció de les dades d'interès a analitzar. Quins són els camps més rellevants per tal de respondre al problema?

En aquest cas el contingut està format per diferents csv, especialment tenim els fitxers de moviments de **cryptomoneda**, els quals s'ha generat **un fitxer per dia del període de temps estudiat**.

Pel que fa als moviments dels principals **índexs borsatils s'ha generat una carpeta per dia del període, en cada directori un fitxer amb cadascun dels noms dels índexs.** Degut això s'han comprimit aquests últims abans de publicar-los en el directori de «*open data*» kaggle.com.

Pel que fa als camps, ens **interessà detectar els moviments alcistes i baixistes**, o almenys aquelles que tenen un patró similar en les cryptomonedes i els índexs. Els camps especialment són els camps comuns els quals netejarem i crearem el dataset parat per l'«script» d'R on generarem la correlació d'atributs i el model:

Camps comuns o nous (transformació i discretització):

- **Data:** Data de l'observació
- **Nom**: Nom empresa o cryptomoneda, per identificar de quina moneda o index estem representant.
- **Símbol**: Símbol de la moneda o del index borsatil, per realitzar gràfic posteriorment d'una forma més senzilla que el nom.
- **Preu**: Valor en euros d'una acció o una cryptomoneda (transformarem la moneda a euros en el cas que estigui en dòlars amb l'última cotització (un dollar a 0,8501 euro)

```
Min.    1st Qu.   Median     Mean     3rd Qu.       Max.
 0.00     0.00     0.09      51.33     2.17        82001.24
```

Clarament tenim valors extrem legítims a causa del bitcoin.

- **Volum:** En euros/volum 24 hores, acumulat de les transaccions diàries en milions d'euros.
- **Tipus_cotitzacio**: Valor nou que agregarem per discretitzar entre la cotització: baix (0 i 1), normal (1 i 100), alt (100 i 1000), molt_alt (>1000)

```
     alta     baixa     molt_alta    normal
      625     33980       242        13653
```

```
def get_type_value(value):
    if value <= 1:
        return 'baixa'
    if 1 < value <= 100:
        return 'normal'
    if 100 < value <= 1000:
        return 'alta'
    else:
        return 'molt_alta'

   return 'molt_alta'
```

- **Tipus**: Generarem aquest per agrupar identificar cryptomonedes d'índex de borsa, tindrà 2 valors: cryptomoneda/borsa.

```
crypto_moneda :38090
index_borsatil:10432
```

Crypto-currencies (no els utilitzarem):

- **Símbol**: Símbol o acrònim de la moneda
- **Cap de mercat**: Valor total de totes les monedes en el moment actual
- **Oferta circulant**: Valor en oportunitat de negoci
- **% 1h, % 2h i %7d**, tant per cent del valor la moneda en 1h, 2h o 7d sobre la resta de cyprtomonedes.

Stock Index (no els utilitzarem):

- **Estat**: Estat final en tancament en alta o baixa del dia.
- **Var. Per cent**: Variació en el moment del tancament amb tant per cent respecte al dia anterior
- **Var. En euros**: Variació en el moment del tancament amb euros respecte al dia anterior.
- **Capitalització**: Valor de l'empresa respecte a les seves accions.
- **PER**: La ràtio preu-benefici
- **Rent./Div:** Rendibilitat de l'acció respecte al valor inicial de l'acció.

> **2.2.** Les dades contenen zeros o elements buits? I valors extrems? Com gestionaries cadascun d'aquests casos?

Com es pot veure en script de python per netejar les dades: **clearData.py, el fitxer** netejar, elimina, transforma i unifica les dades procedents de les diferents fonts en el fitxer dataset.csv.

El dataset.csv és el conjunt de resultats que utilitzarem per a l'anàlisi amb els camps:

```
HEADERS = ['Data', 'Tipus', 'Nom', 'Simbol', 'Preu (Euros)', 'Tipus_cotitzacio']
```

Filter

| | Data | Tipus | Nom | Simbol | Preu..Euros. | Tipus_cotitzacio |
|---|---|---|---|---|---|---|
| 1 | 2017-11-19 | crypto_moneda | BTCBitcoin | BTC | 6.832132e+03 | molt_alta |
| 2 | 2017-11-19 | crypto_moneda | ETHEthereum | ETH | 2.985163e+02 | alta |
| 3 | 2017-11-19 | crypto_moneda | BCHBitcoin Cash | BCH | 1.003034e+03 | molt_alta |
| 4 | 2017-11-19 | crypto_moneda | XRPRipple | XRP | 1.952729e-01 | baixa |
| 5 | 2017-11-19 | crypto_moneda | LTCLitecoin | LTC | 6.010265e+01 | normal |
| 6 | 2017-11-19 | crypto_moneda | DASHDash | DASH | 3.720960e+02 | alta |

dataset 48522 obs. of 6 variables

Disposem de 48522 observacions amb 6 atributs:

Pel que fa a les estratègies que hem seguit per als zeros i elements buits són:

- Les monedes o índex borsatils que puntualment **no hi havia preu, està en blanc, eliminem l'observació**, en aquest no disposar d'un valor d'una cotització podria contaminar la tendència, en aquest sentit es mantindrà la tendència del dia anterior.

```
def get_value_euros(row):
    value = clear_currency_data(row)
    if value:
        return float(value) * EURO_VALUE_FROM_DOLAR
    else:
        return None
```

(de 48522 observacions em passat a 46859)

- Pel que fa als índexs borsatils, que no hi havia símbol o no es disposava, s'ha agregat els tres primers caràcters del nom del valor, per poder realitzar gràfics posteriorment amb un nom curt.

```
format(get_value_euros(row[4][1:])
```

Finalment sobre les monedes s'ha passat a euros, ja que les cryptomonedes el preu era amb dollars, i els índexs borsatils en euros, amb una precisió de decimals de 9 dígits, què es recomana per monedes.

A més s'ha unificat la data dels valors amb un format estàndard YYYY-m-d.

```
FORMAT_DATA = '%d_%m_%Y'
```

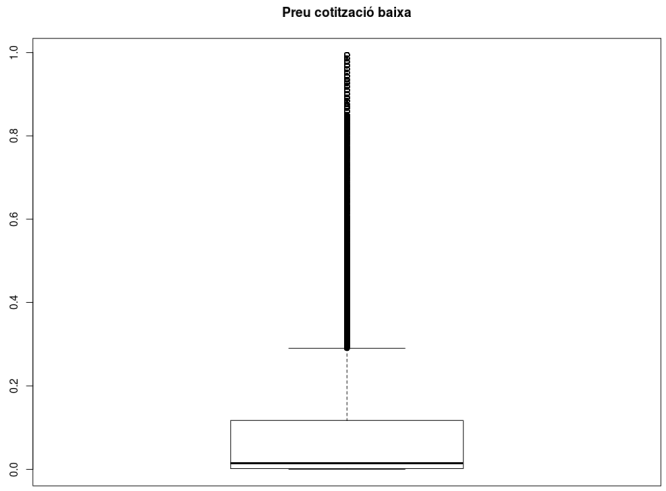
Pel que fa als valors extrems hem d'observar primer les freqüències, ho farem pel tipus de

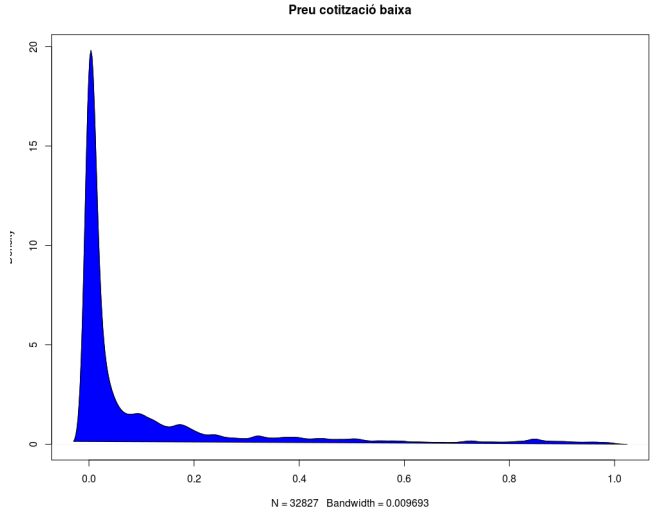
cotització que hem discretitzat, per tenir una primera aproximació a la similitud de la tendència.

Cal tenir present, que la diferència entre cryptomonedes és molt alta, especialment amb el màxim o valors extrems del bitcoin.

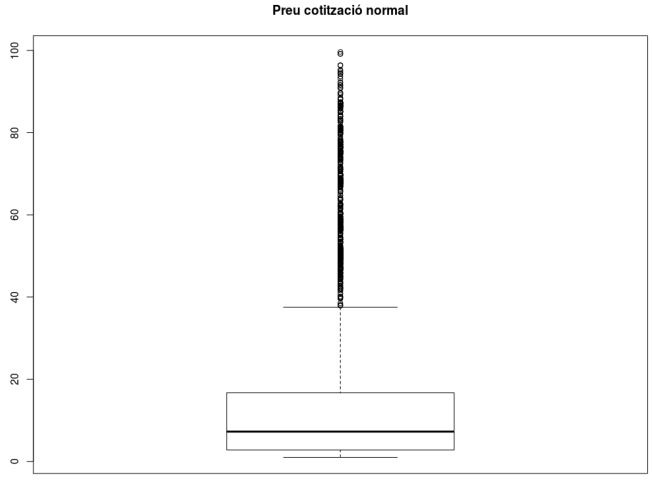
En definitiva els **valors extrems en el nostre cas es tracta de valors atípics legítims**, no son degut a cap error humà o de mostreig, i ens aporta una major font d'informació, degut a que pot haver-hi ha una tendència entres els «stock» index i les cryptomendes amb el cotització alta del bitcoin (es veuran «arrossegades» per la tendència), com valor extrem de la resta de cryptomonedes.

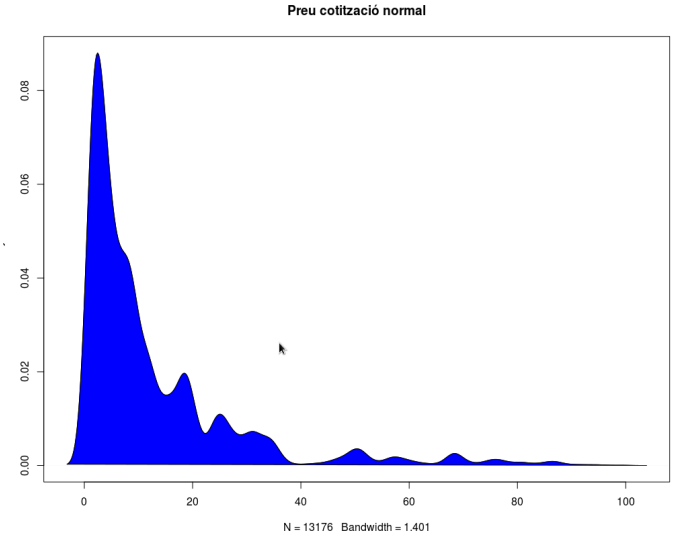
**Cotització baixa:**



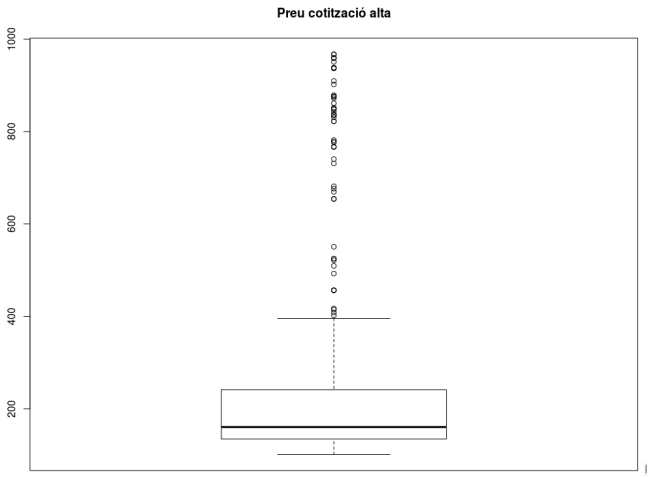



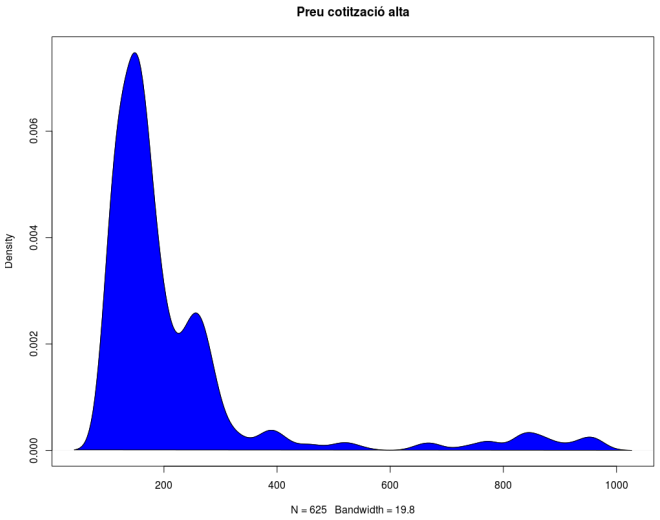

**Cotització normal:**



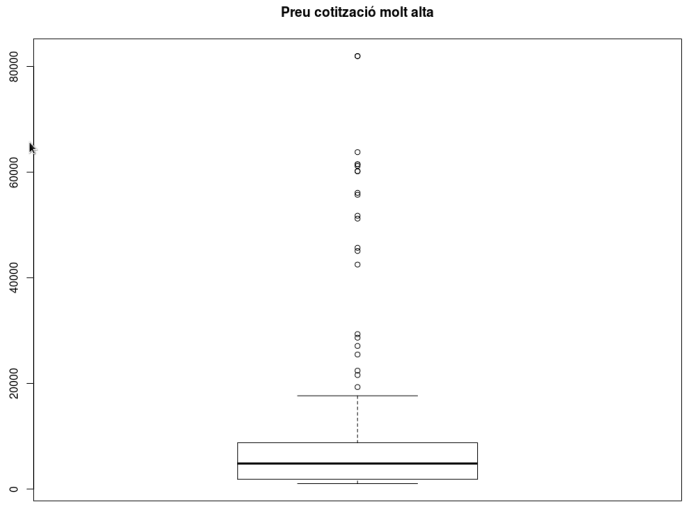



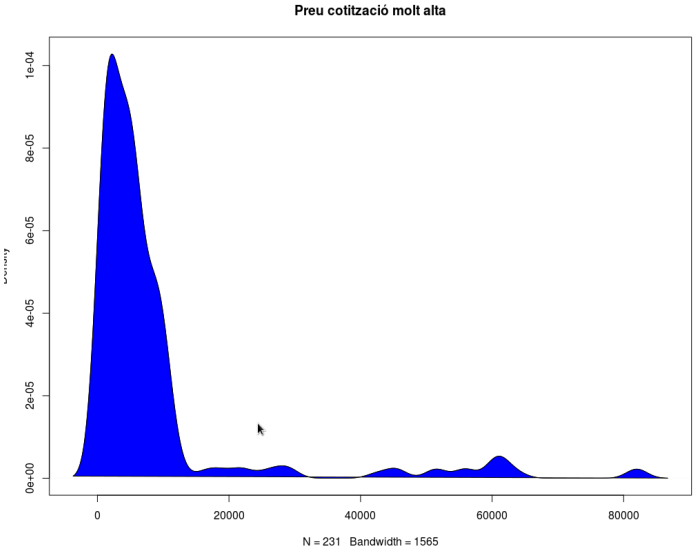
**Cotització alta:**




**Cotització molt alta:**

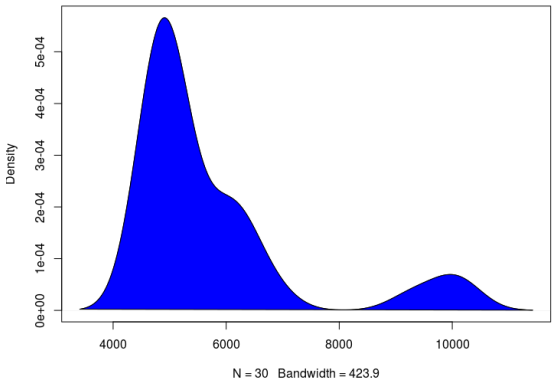


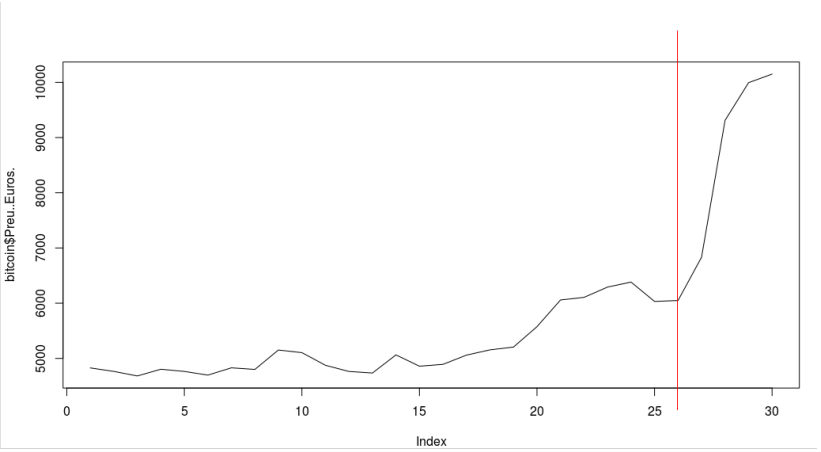




Com podem veure en cada grup del tipus de cotització la distribució es comporta de la mateixa forma, la gran concertació de valors està en el rang més vaig de preu de cada discretització, per tant tenim representat valors extrems legítims que ens indiquen que sempre hi ha cotització molt més altes que la majoria.

Una observació curiosa és un valor extrem el **bitcoin com a cryptomoneda** que està en

un moviment alcista molt clar. Podem fer una ampliació de la distribució del bitcoin per veure aquesta tendència:




## 3. Anàlisis de les dades:

### 3.1. Selecció dels grups de dades que es volen analitzar/comparar.

En aquest cas el grup de dades que volem comparar son el preu de les cotitzacions, i **discretitzar per grup de baixa, normal, alta i molt alta**, per veure relacions del valor qualitatiu de tipus (si es cryptomoneda o índex borsatil):

- **Data:** Data de l'observació
- **Nom**: Nom de l'empresa o cryptomoneda, per identificar de quina moneda o index estem representant.
- **Símbol**: Símbol de la moneda o del index borsatil, per realitzar gràfic posteriorment d'una forma més senzilla que el nom.
- **Preu**: Valor en euros d'una acció o una cryptomoneda (transformarem la moneda a euros en el cas que estigui en dòlars amb l'última cotització (un dollar a 0,8501 euro)
- **Tipus_cotitzacio**: Valor nou que agregarem per discretitzar entre la cotització: baix (0 i 1), normal (1 i 100), alt (100 i 1000), molt_alt (>1000)
- **Tipus**: Tipus de valor: «stock» índex o cryptomoneda.

Tenim 2 grups cyptomoneda i «stock index», a més de 4 grups per tipus de preu en la cotització, posteriorment utilitzarem algorismes d'agrupació per veure similitud amb aquest tipus de valor.

### 3.2. Comprovació de la normalitat i homogeneïtat de la variància. Si és necessari (i possible), aplicar transformacions que normalitzin les dades.

Com hem vist en la pregunta 2, no hi hagut més remei que discretitzar per tipus

de preu en la transacció, ja que els valors extrems d'algunes cryptomonedes eren molt alts (bitcoin), igual ha passat en les stock indexs.

Amb el test de Levene amb els grups de cryptomoneda i stock index, p-valor és menor de 0.05 per tant es **rebutja la hipotiposi nul·la**:

```
Rcmdr>  with(dataset, tapply(Preu..Euros., Tipus, var, na.rm=TRUE))
 crypto_moneda index_borsatil
   1758612.334        501.601

Rcmdr>  leveneTest(Preu..Euros. ~ Tipus, data=dataset, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
         Df F value    Pr(>F)
group     1  13.562 0.000231 ***
      47183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Per tant hem **aplicat transformació per obtenir grups homogenis i poder tenir un gràfic adequat amb la freqüència dels valors**. Tot i això aplicarem un algorisme d'agrupació en la pregunta 5, i ara realitzarem el test de Levene per **veure l'homogeneïtat dels grups** per tipus de valor:

```
Rcmdr>  with(dataset, tapply(Preu..Euros., list(Tipus, Tipus_cotitzacio), var, na.rm=TRUE))
                    alta      baixa molt_alta   normal
crypto_moneda  38146.8202 0.03796463 203807965 236.7990
index_borsatil   506.1773 0.06039860        NA 238.1323

Rcmdr>  leveneTest(Preu..Euros. ~ Tipus*Tipus_cotitzacio, data=dataset, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
         Df F value    Pr(>F)
group     6  1864.7 < 2.2e-16 ***
      47178
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**No hi ha grups homogenis**, degut als valors extrems que no hem descartat segurament, en tot cas, el que farem al punt 3.3 és usar algorisme de kmeans per trobar la agrupació més optima i poder veure realment si hi ha una relació entre als valors de la cryptomoneda i els index borsatil.

Pel que fa la normalitat de les dades podem veure amb el test de Shapiro-Wilk, hi ha **normalitat**:

Bitcoin:

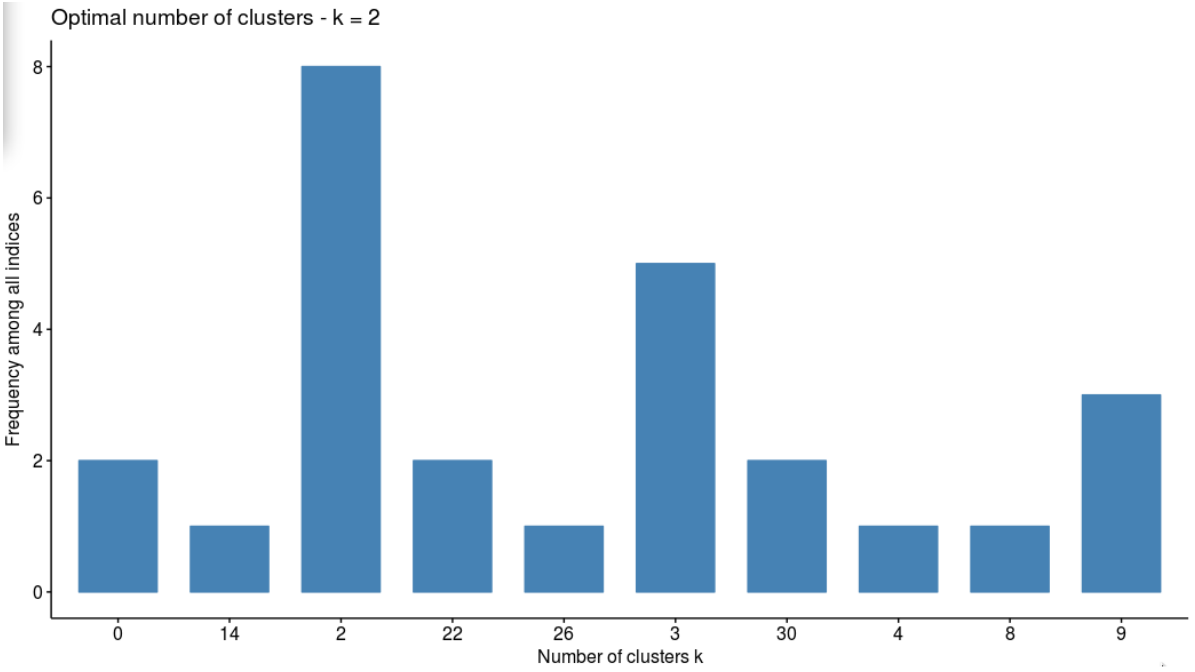
```
test <- wilcox.test(bitcoin_october$Preu..Euros., bitcoin_novembre$Preu..Euros.)
print(test)
```

```
# Wilcoxon rank sum test
#
# data:  bitcoin_october$Preu..Euros. and bitcoin_novembre$Preu..Euros.
# W = 0, p-value = 3.661e-08
# alternative hypothesis: true location shift is not equal to 0
#Com es compleix p < 0.05  refusem que sigui uns distribució normal
      Wilcoxon rank sum test

data:  x1 and x2
W = 136, p-value = 0.1853
alternative hypothesis: true location shift is not equal to 0

#Com no es complex p > 0.05 refusem que sigui uns distribució normal

### normalitzem ###
x1 = rnorm(bitcoin_october$Preu..Euros.)
x2 = rnorm(bitcoin_novembre$Preu..Euros.)
test <- wilcox.test(x1,x2)
print(test)
# W = 112, p-value = 0.767
#Com es compleix p > 0.05 no podem refusar que sigui uns distribució normal
```

IOTA:

```
test <- wilcox.test(miota_octuber$Preu..Euros., miota_octuber$Preu..Euros.)
print(test)
# Wilcoxon rank sum test with continuity correction
#
# data:  miota_octuber$Preu..Euros. and miota_octuber$Preu..Euros.
# W = 180.5, p-value = 1
# alternative hypothesis: true location shift is not equal to 0
#Com es compleix p > 0.05 no refusem que sigui uns distribució normal
```

En aquest cas utilitzarem un **model d'agregació que ens permetrà fer perdicions d'atributs observant els veïns més pròxims**, per majoria o mitjana. El que farem és normalitzar els valors i utilitzarem l'algorisme kmeans amb R per crear cluster o grups de dades òptims, a més del NbClust un paquet que iteraré amb algorisme kmeans per proposar el número de clusters o grups més òptims:




```
Conclusion

=========================
* According to the majority rule, the best number of clusters is  2 .
```

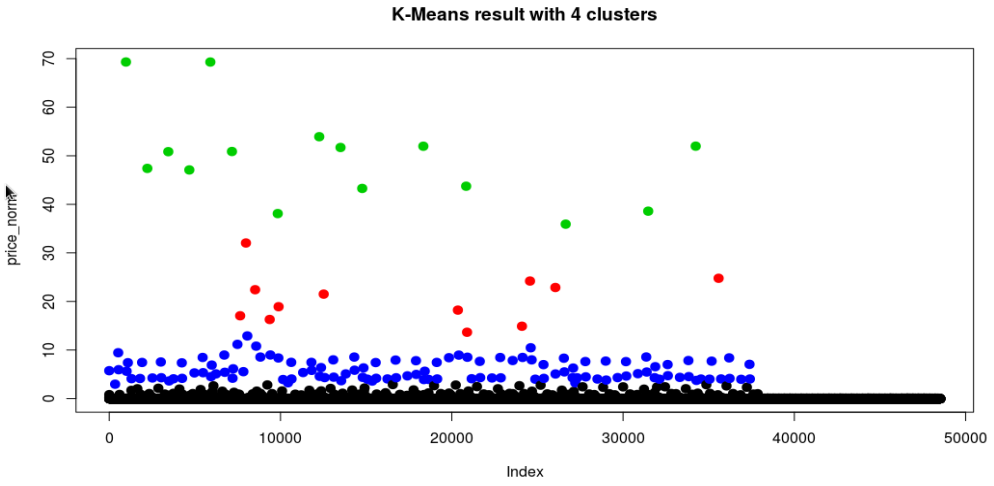
Podem veure que el cluster més optim està format per 2 grups, tot i això, amb 4 grups ens dóna un valor més alt d'homogeneïtat:

```
clusters_2 <- kmeans(price_norm,2, 15)
print(clusters_2)

K-means clustering with 2 clusters of sizes 47168, 17

Cluster means:
   Preu..Euros.
1  -0.01698065
2  47.11429870

Within cluster sum of squares by cluster:
[1] 6930.837 2503.691
 (between_SS / total_SS =  78.1 %)
```

```
clusters_4 <- kmeans(price_norm,4, 15)
print(clusters_4)

Within cluster sum of squares by cluster:
[1]  284.6638  291.0051 1341.0138  480.6168
 (between_SS / total_SS =  95.1 %)
```

Per alta banda, si realitzem **una agrupació per 4 grups, podem veure es molt similar a la discretització per tipus de preu** (baix, normal, alt i molt alt):

```
plot(price_norm, col =(clusters_4$cluster) , main="K-Means result with 4 clusters", pch=20, cex=2)
```

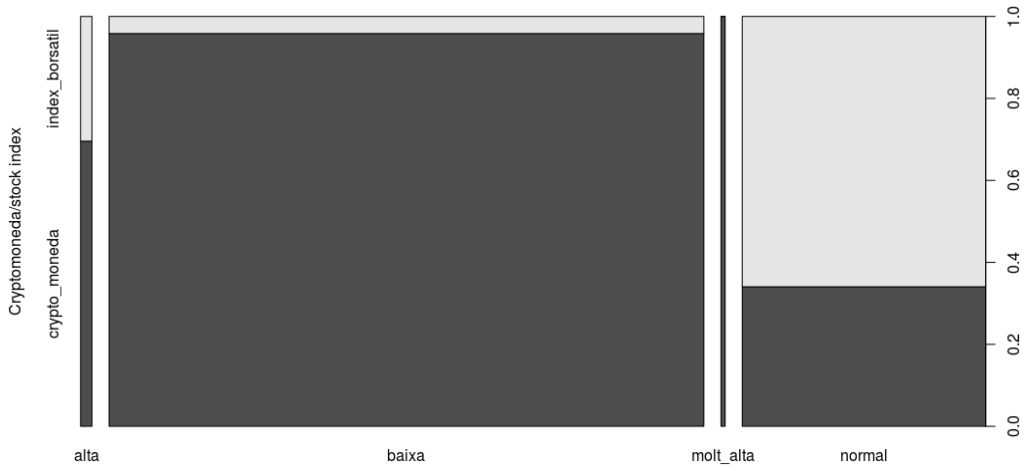



- Grup Verd:
  - Degut això, podem veure que en el rang de cotització (preu diari de la cotització del valor) alt (més de 1000 euros el valor) **són sòls valors de cryptomonedes**, i no hi ha valors d'índexs borsatils que tinguin aquest volum, **té una densitat baixa**. Segurament aquests valors **estan ocasionant tendències sobre la resta de grups**.
- Grup Vermell:
  - El tipus de cotització alta (100 i 1000 euros), **és un rang predominat pels index borsatils**, també hi ha cryptomonedes. Aquest grup t**é una densitat molt baixa,** és el grup que menys patrons podrem trobar.
- Grup Blau:
  - La cotització està en el rang d'1 i 100 euros, normal, **té una densitat alta de valors predominat pels index borsatils**, hi podrem trobar bastantes tendències.
- Grup Negre:
  - Cotització baixa, entre O i 1 euro, **és el grup més dens, a causa de la gran quantitat de cryptomondes noves**, també disposem d'índexs borsatils. Per altra banda, **és on podem trobar tendències entre els dos tipus de valors** (cryptomoneda i «stock» índexs).

### **3.3.** Aplicació de proves estadístiques (tantes com sigui possible) per comparar els grups de dades.

Inicialment podem veure la distribució de les discretitzacions o grups del tipus de preu:

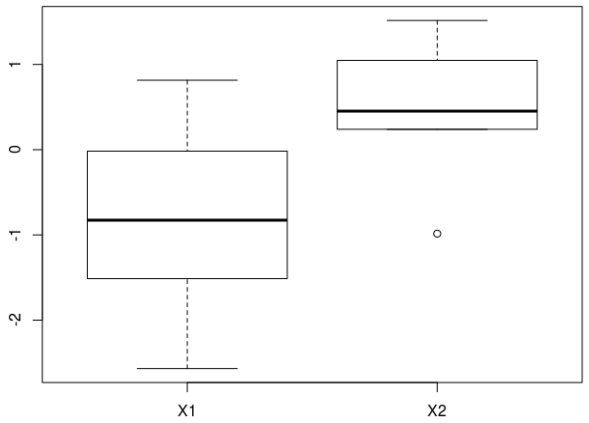
```
#comparació tipus de preus o rangs amb cryptomonedes/stock índex
plot(dataset$Tipus_cotitzacio, dataset$Tipus, xlab = "Tipus de cotització", ylab = "Cryptomoneda/stock index")
```




Finalment per respondre a la pregunta farem les següents proves estadístiques:

1. Un **test de contrast d'hipòtesis de variables dependents (Wilcoxon)** d'una mateixa moneda per comparar dos mesos, iterem sobre les monedes, **comprovem la normalitat per normalitzar o no (Shapiro-Wilk)**, creant subconjunts dels mesos per cada moneda i aplicant contrast, veiem que una les que hi ha més diferencia és:

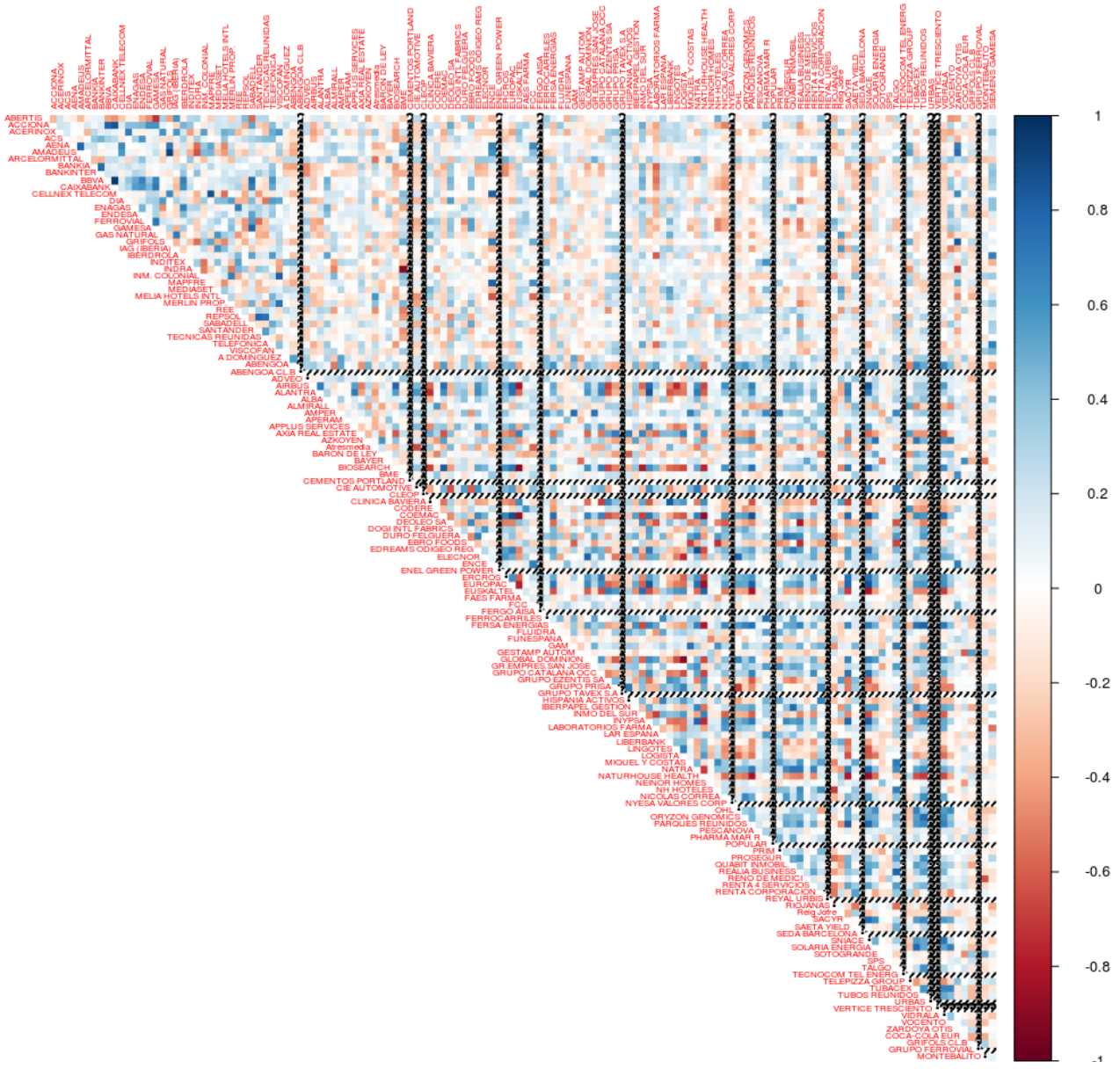
```
"Moneda més diferencia: HBT  ' p-valor: 0.00221343873517787"

Wilcoxon rank sum test

data:  x1 and x2
W = 16, p-value = 0.002213
alternative hypothesis: true location shift is not equal to 0
```




Algunes altres monedes que p-value < 0.05 per tant la tendència central de les mostres no es la mateixa entre els dos mesos, es el que es el mateix, hi ha mes diferencies significatives:

```
[1] "Moneda: TRICK , p-valor: 0.0254727880814837"
[1] "Moneda: SHELL , p-valor: 0.0351657038613560"
[1] "Moneda: NXX , p-valor: 0.0359640359640360"
[1] "Moneda: APP , p-valor: 0.0421103060061589"
[1] "Moneda: DUR , p-valor: 0.0421103060061589"
[1] "Moneda: INY , p-valor: 0.0344580427206545"
[1] "Moneda: RIO , p-valor: 0.0135642302361737"
```

2. Correlacionarem els diferents índexs borsàtils per veure si hi ha similars, deixarem sols un dels mes similars per poder fer una regressió múltiple amb menys variables i tenir un model, seleccionarem mes d'octubre com a dataset i les dades per fer test del model el mes de nombre.



Descartem de cada grup o index correlacionats tots menys, cada grup té un index de correlació superior a 0.70.

| | ABERTIS | ACCIONA | ACERINOX | ACS | AENA | AMADEUS | ARCELORMITTAL | BANKIA | BANKINTER | BBVA | CAIXABANK | CELLNEX TELECOM | DIA | ENAGAS | END |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ABERTIS | 1.000000000 | -0.326598632 | 0.197545919 | -0.062241200 | -0.100858941 | -0.324016785 | -0.084721511 | 0.085561189 | 0.101255245 | -0.390729513 | -0.448136638 | -0.254164532 | -0.071735705 | -0.065848640 | -0.3 |
| ACCIONA | -0.326598632 | 1.000000000 | 0.209684238 | 0.174819860 | -0.019764235 | -0.047620547 | 0.177878118 | 0.059880359 | 0.146830021 | 0.794799112 | 0.853771410 | 0.146255342 | 0.421718098 | 0.427433254 | 0.4 |
| ACERINOX | 0.197545919 | 0.209684238 | 1.000000000 | 0.163939934 | 0.003984854 | 0.024003073 | -0.338712629 | 0.261582654 | 0.116014851 | 0.155213031 | 0.081969967 | -0.067742526 | 0.708554791 | -0.073170732 | -0.0 |
| ACS | -0.062241200 | 0.174819860 | 0.163939934 | 1.000000000 | -0.044194174 | -0.116969722 | -0.004017652 | 0.389517245 | 0.467878887 | 0.321440000 | 0.256198347 | -0.100441304 | -0.004202274 | 0.319682870 | 0.5 |
| AENA | -0.100858941 | -0.019764235 | 0.003984854 | -0.044194174 | 1.000000000 | 0.784319756 | 0.195312500 | 0.209083146 | -0.007843198 | 0.082243962 | 0.060264782 | 0.648437500 | -0.004085753 | -0.003984854 | 0.0 |
| AMADEUS | -0.324016785 | -0.047620547 | 0.024003073 | -0.116969722 | 0.784319756 | 1.000000000 | -0.007843198 | 0.162379103 | -0.070866142 | 0.053668629 | 0.032267509 | 0.572553422 | 0.069730718 | -0.080010242 | -0.0 |
| ARCELORMITTAL | -0.084721511 | 0.177878118 | -0.338712629 | -0.004017652 | 0.195312500 | -0.007843198 | 1.000000000 | -0.303762306 | -0.345100693 | 0.222058697 | 0.245076782 | 0.125000000 | -0.192030382 | 0.227136704 | 0.4 |
| BANKIA | 0.085561189 | 0.059880359 | 0.261582654 | 0.389517245 | 0.209083146 | 0.162379103 | -0.303762306 | 1.000000000 | 0.772290855 | 0.141200445 | 0.085206897 | 0.161743566 | 0.222817332 | 0.124754804 | 0.1 |
| BANKINTER | 0.101255245 | 0.146830021 | 0.116014851 | 0.467878887 | -0.007843198 | -0.070866142 | -0.345100693 | 0.772290855 | 1.000000000 | 0.189904381 | 0.149237231 | -0.078431976 | 0.118952401 | 0.292037383 | 0.2 |
| BBVA | -0.390729513 | 0.794799112 | 0.155213031 | 0.321440000 | 0.082243962 | 0.053668629 | 0.222058697 | 0.141200445 | 0.189904381 | 1.000000000 | 0.960090526 | 0.131590339 | 0.395707160 | 0.515978455 | 0.6 |
| CAIXABANK | -0.448136638 | 0.853771410 | 0.081969967 | 0.256198347 | 0.060264782 | 0.032267509 | 0.245076782 | 0.085206897 | 0.149237231 | 0.960090526 | 1.000000000 | 0.172759043 | 0.306766018 | 0.557395774 | 0.5 |
| CELLNEX TELECOM | -0.254164532 | 0.146255342 | -0.067742526 | -0.100441304 | 0.648437500 | 0.572553422 | 0.125000000 | 0.161743566 | -0.078431976 | 0.131590339 | 0.172759043 | 1.000000000 | -0.053114787 | 0.067742526 | -0.0 |
| DIA | -0.071735705 | 0.421718098 | 0.708554791 | -0.004202274 | -0.004085753 | 0.069730718 | -0.192030382 | 0.222817332 | 0.118952401 | 0.395707160 | 0.306766018 | -0.053114787 | 1.000000000 | -0.070855479 | 0.1 |
| ENAGAS | -0.065848640 | 0.427433254 | -0.073170732 | 0.319682870 | -0.003984854 | -0.080010242 | 0.227136704 | 0.124754804 | 0.292037383 | 0.515978455 | 0.557395774 | 0.067742526 | -0.070855479 | 1.000000000 | 0.4 |
| ENDESA | -0.331269330 | 0.442235250 | -0.069530961 | 0.585568255 | 0.048112522 | -0.028175916 | 0.400937687 | 0.194357471 | 0.241507852 | 0.679542697 | 0.597939415 | -0.072168784 | 0.142583034 | 0.462176387 | 1.0 |
| FERROVIAL | -0.463777062 | 0.596409007 | 0.061350848 | -0.061855802 | 0.096225045 | 0.076477486 | 0.104243799 | -0.048589368 | -0.080502617 | 0.388310112 | 0.457732932 | 0.264618873 | 0.327102254 | 0.053170735 | 0.0 |

Exportant les dades a full de càlcul, i amb ajuda de filtre per veure els valors superior a 0.7 descartarem tots menys un de cada grup:

- ACCIONA, FERROVIAL, BBVA, SABADALL, GAMESA i CAIXABANK tenen un correlació del aprox. entre 0.75 i 0.8.
- ACERINOX i DIA del 0.7
- AENA, AMADEUS, ZARDOYA OTIS i GRIFOLS del aprox. 0.70
- BANKIA i BANKINTER en un 0.77
- CELLNEX TELECOM, MAFRE, ENCE i FCC aprox. 0.75
- GAS NATURAL i IAG (IBERIA) un 0.70
- INM. COLONIAL, MERLIN PROP. I VISCOFAN en un 0,73
- REPSOL, SANTANDER i TECNICAS REUNIDAS en un 0.75
- ALANTRA, APPLUS SERVICES, EUROPAC, INYPSA, NATRA i SNIACE en un aprox. 0.75
- AMPER, GAM i GR.EMPRES.SAN JOSE en un 0,75
- ALANTRA i APPLUS SERVICES en 0,70
- AZKOYEN, SOLARIA ENERGIA i TUBACEX en un 0,75
- AIRBUS, BIOSEARCH, COEMAC, EUROPAC, FERSA ENERGIAS, INYPSA, PARQUES REUNIDOS, SACYR, VIDRALA i SNIACE en un 0,80
- CIE AUTOMOTIVE i REALIA BUSINESS en un 0,77
- CLINICA BAVIERA i LIBERBANK en un 0.84
- DURO FELGUERA i PROSEGUR en un 0.75
- EDREAMS ODIGEO REG, INYPSA, NATRA, QUABIT INMOBIL. I SNIACE en un aprox. 0.75
- ELECNOR, GLOBAL DOMINION i OHL en un 0.80
- CELLNEX TELECOM, MAPFRE, i ENCE en un 0.75
- ERCROS, GRUPO CATALANA OCC, SACYR i SNIACE en un 0.80
- EUSKALTEL, INMO DEL SUR, NATURHOUSE HEALTH en un 0,70
- CELLNEX TELECOM i FCC en un 0.80
- AMPER, GAM. MIQUEL i COSTAS en un 0.75
- ELECNOR i GLOBAL DOMINION en un 0,80

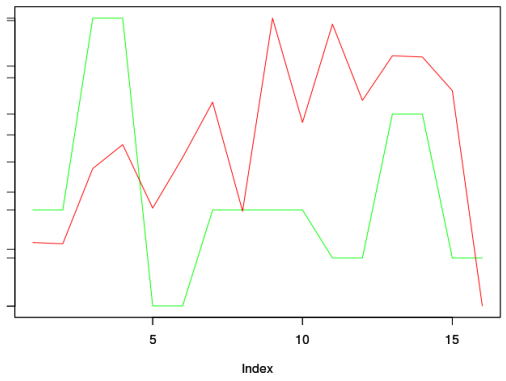
3. Finalment realitzarem els gràfics en la pregunta 4 i en la pregunta 5 respondrem a: han canviat els índex que afecten? Són els mateixos però tenen un pes diferent? Tenen el mateix pes però han crescut considerablement?

Finalment creem diferents regressions lineals, tenim per una banda les cotitzacions d'octubre de la moneda HBT, una de les que el test de **Wilcoxon**, ens indica que hi ha més diferencia entre els meses d'octubre i novembre, i com dataset de test les cotitzacions de HBT del mes de nombre. Normalitzam les dades o no depenen del test de normalitat de **Shapiro-Wilk**.

Després creem un bucle i realitzem n models en cadascun dels index borsatils amb les dades del mes de octubre d'aquest, veiem els index que tenen un p-valor inferior a 0.05 per tant on la hipòtesi n'indica que els índexs afecten, ja que tenen la mateixa diferencia (tendència) en el mes d'octubre:

#[1] "Hi ha relació en la moneda HBT amb index borsatil: FUNESPANA , p-valor: 0.0380989010567089"
#[1] "Hi ha relació en la moneda HBT amb index borsatil: GRUPO PRISA , p-valor: 0.00627180926584772"
#[1] "Hi ha relació en la moneda HBT amb index borsatil: ORYZON GENOMICS , p-valor: 0.0269072098667691"
#[1] "Hi ha relació en la moneda HBT amb index borsatil: Reig Jofre , p-valor: 0.00217669767591883"
#[1] "Hi ha relació en la moneda HBT amb index borsatil: URBAS , p-valor: 0.0122987380270188"
#[1] "Hi ha relació en la moneda HBT amb index borsatil: MONTEBALITO , p-valor: 0.0423629291066837"

La resposta a la nostra pregunta, es que si hi ha relació entre les monedes i els index borsatils del IBEX-35, en aquest entre la cryptomoneda HBT en el mes d'octubre i novembre.




Realitzem un predicció amb les dades del index borsatils amb p-valor mes baix, **i el coeficient de determinació (coeficient de correlació al cuadrat) que ens indica la bondat de la recta de les dades**. Despres fem comprovació amb les dades de test del mes veiem que, tot i tenir molt poques dades, la perdició observem que la tendència si es pot preveure.

Amb vermes està les dades reals del més novembre i amb ver la perdició del model.

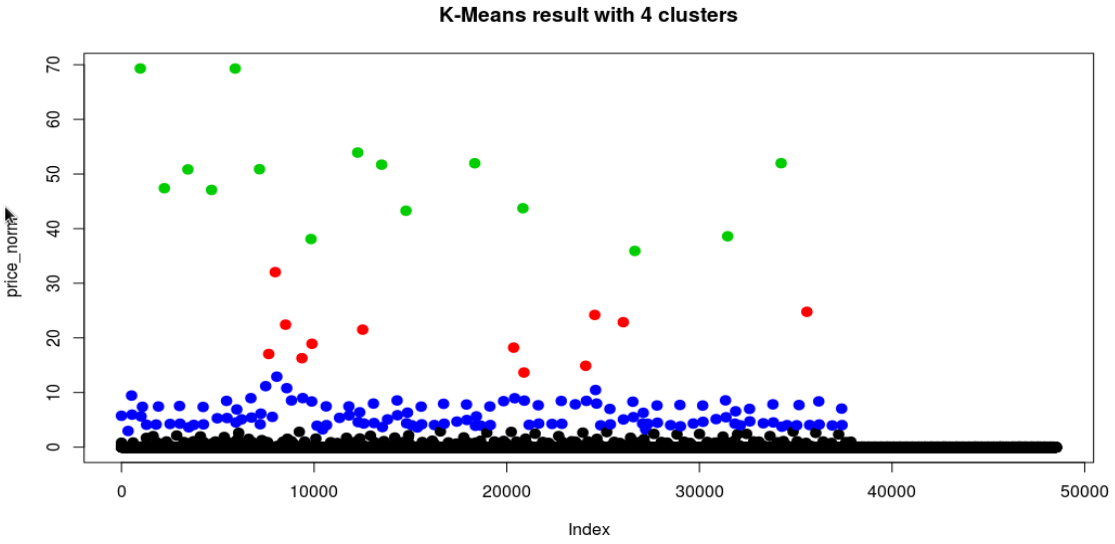
```
# Call:
#   lm(formula = octrubre_norm_HBT[1:16] ~ index_octuber_norma[1:16])
#
# Residuals:
#   Min       1Q   Median       3Q      Max
# -1.36465 -0.58417 -0.03334  0.56504  1.80891
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)                -0.5667     0.2433   -2.33  0.03530 *
#   index_octuber_norma[1:16]  -0.9315     0.2511   -3.71  0.00233 **
```

```
#   ---
#   Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.8479 on 14 degrees of freedom
# Multiple R-squared:  0.4958,      Adjusted R-squared:  0.4598
# F-statistic: 13.77 on 1 and 14 DF,  p-value: 0.00233
```

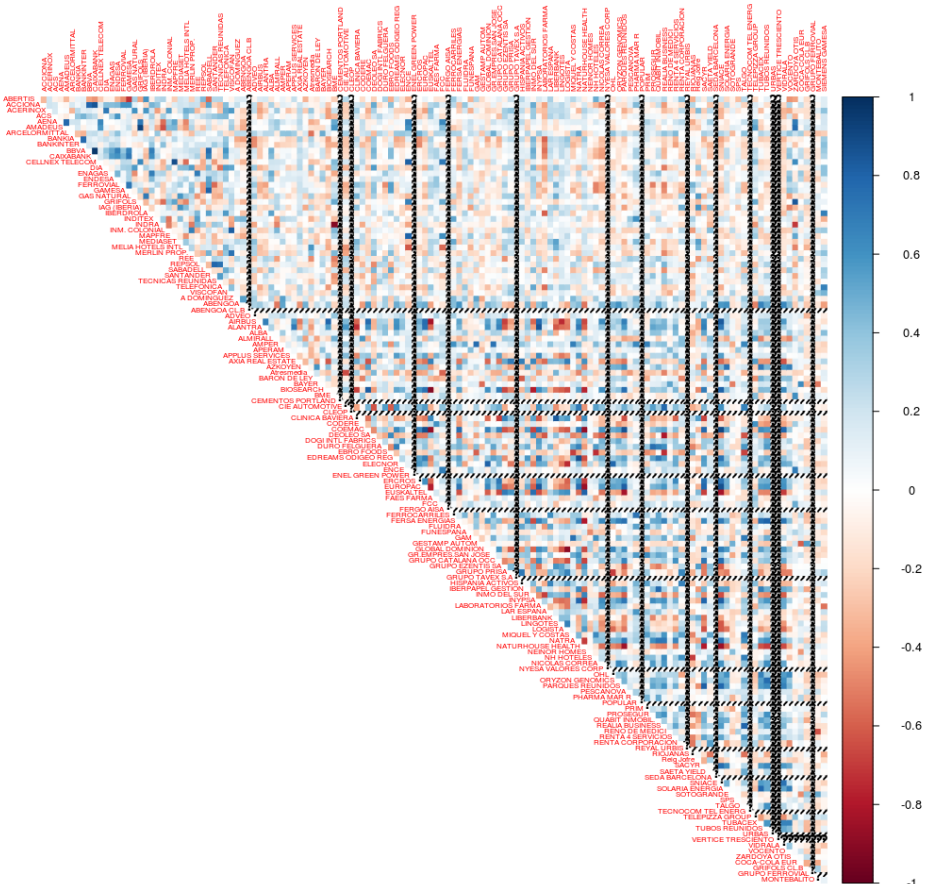
## 4. Representació dels resultats a partir de taules i gràfiques.

Durant tota la practica hem anat representat els resultats, tant les freqüències per veure els valors extrems, així com la matriu de correlació per eliminar atributs per realitzar la regressió i el model predictiu, i els resultats amb un gràfic per veure la perdició envers les dades reals del test en el mes de novembre.

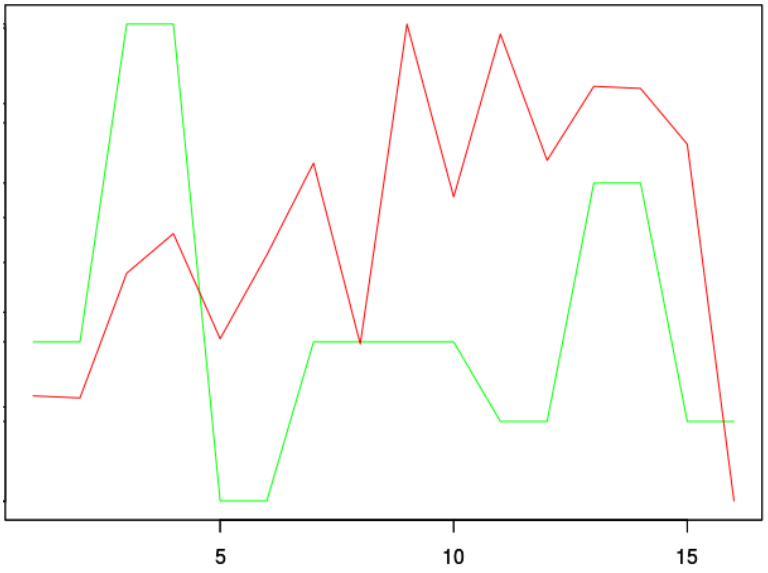
- Els 4 grups de index i monedes pel seu valor de cotització:




- Podem veure el gràfic de la matriu de correlació dels index borsatils per tenir menys variables per als model de regressió:




I finalment, visualment podem veure la diferència entre perdició de les dades de la moneda HBT (la qual te més diferencia entre els dos mesos), i les dades de test reals del mes de novembre de HBT:



En vermell és la cotització real i amb verda la predicció, si hem pogut predir la tendència, clarament, en la poca informació de la distribucional normal que tenim no podem preveure el preu de la cotització, a més com podem veure també influeixen altres factors.

He agregat el script d'R com a kernel en el kaggle.com:

- https://www.kaggle.com/acostasg/grafic-distribucion-and-kmeans-algorim-group/output

## 5. Resolució del problema.

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Finalment els resultats han permès de respondre al problema, les diferents regressions lineals dels index borsatils (s'han eliminat índexs borsatis del més correlacionats), envers una de les criptomoneda HBT amb més diferencia entre els mesos d'octubre i nombres (test de contrast d'hipòtesis de variables dependents Wilcoxon), ha permès primer veure que la hipòtesi que la tendència d'algunes index borsatils ha sigut igual que la moneda en qüestió, i que posterior el model de regressió ha permès predir la tendència, amb les poques dades de la distribució normal que disposem.

Per tant hem pogut veure que si els canvis produïts en algunes cryptomonedes també s'ha **replicat i s'ha vist afectat per la tendència d'alguns index borsatils de l'Ibex35** en el mes de novembre i octubre, tot i que en pesos diferents. Amb les diferents models de cada index borsatil sobre la moenda HBT hem vist que ni tots els indexs borsatils ni en el mateix pes.

En conclusió, la tendència de creixement o decreixement d'algunes **criptomònades estan relacionades amb la mateixa tendència dels index borsatils** duran el temps de la mostra, clarament també afecten molts altres factors.

## 6. Codi:

Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi amb python i R està en la wiki https://github.com/acostasg/scraping, exactament hem de destacar:

- El fitxer de codi de python per la netejat, eliminació i transformació de dades
  - https://github.com/acostasg/scraping/blob/master/clearData.py
- I el fitxer de codi R amb l'analisi i els gràfics del conjunt de dades (s'ha discretitzat un atribut):
  - https://github.com/acostasg/scraping/blob/master/R/script.r

# Referencies


- https://stackoverflow.com/questions/6159900/correct-way-to-write-line-to-file-in-python
- Llibre manual: Richard Lawson. Web Scraping with Python. Packt Publishing Ltd, October 2015. 174 p. ISBN 9781782164371
- Regressió lineal simple amb R. https://www.uam.es/personal_pdi/ciencias/joser/paginaR/regresion.html
- The R Graph Gallery. https://www.r-graph-gallery.com/
- Exemple de com fer un gràfic de Kmeans usan ACP. http://apuntes-r.blogspot.com.es/2014/10/ejemplo-graficar-kmeans-usando-acp.html
- https://docs.python.org/2/tutorial/inputoutput.html
- Correlations - https://www.statmethods.net/stats/correlations.html
- Test de Shapiro-Wilk - https://rpro.wikispaces.com/Test+de+Shapiro-Wilk