

Mahalanobis_Distance_Stock

September 29, 2021

1 Mahalanobis Distance

1.1 Mahalanobis distance is the distance between two points in a multivariate space. It's used in statistical analyses to find outliers that involve several variables.

1.2 Formula: $d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2}$

```
[1]: import numpy as np
import scipy as stats
from scipy.stats import chi2

import warnings
warnings.filterwarnings("ignore")

# yfinance is used to fetch data
import yfinance as yf
yf.pdr_override()
```

```
[2]: symbol = 'AMD'

start = '2018-01-01'
end = '2019-01-01'

# Read data
dataset = yf.download(symbol,start,end)

# View Columns
dataset.head()
```

```
[*****100%*****] 1 of 1 completed
```

```
[2]:
```

	Open	High	Low	Close	Adj Close	Volume
Date						
2018-01-02	10.42	11.02	10.34	10.98	10.98	44146300
2018-01-03	11.61	12.14	11.36	11.55	11.55	154066700
2018-01-04	12.10	12.43	11.97	12.12	12.12	109503000
2018-01-05	12.19	12.22	11.66	11.88	11.88	63808900

```
2018-01-08 12.01 12.30 11.85 12.28 12.28 63346000
```

```
[3]: dataset.tail()
```

```
[3]:
```

	Open	High	Low	Close	Adj Close	Volume
Date						
2018-12-24	16.520000	17.219999	16.370001	16.650000	16.650000	62933100
2018-12-26	16.879999	17.910000	16.030001	17.900000	17.900000	108811800
2018-12-27	17.430000	17.740000	16.440001	17.490000	17.490000	111373000
2018-12-28	17.530001	18.309999	17.139999	17.820000	17.820000	109214400
2018-12-31	18.150000	18.510000	17.850000	18.459999	18.459999	84732200

```
[4]: dataset = dataset.drop(['Adj Close', 'Volume'], axis=1)
dataset.head()
```

```
[4]:
```

	Open	High	Low	Close
Date				
2018-01-02	10.42	11.02	10.34	10.98
2018-01-03	11.61	12.14	11.36	11.55
2018-01-04	12.10	12.43	11.97	12.12
2018-01-05	12.19	12.22	11.66	11.88
2018-01-08	12.01	12.30	11.85	12.28

```
[5]: def mahalanobis_distance(x=None, data=None, cov=None):

    x_mu = x - np.mean(data)
    if not cov:
        cov = np.cov(data.values.T)
    inv_covmat = np.linalg.inv(cov)
    left = np.dot(x_mu, inv_covmat)
    mahal = np.dot(left, x_mu.T)
    return mahal.diagonal()
```

```
[6]: df = mahalanobis_distance(x=dataset, data=dataset)
df
```

```
[6]: array([[ 2.34360202,  2.44314893,  1.00051049,  1.21842069,  1.0076011 ,
          1.1097397 ,  1.16944107,  1.06911884,  1.00390335,  0.85728349,
          0.9411238 ,  1.0507168 ,  1.47890511,  0.93506169,  1.04127015,
          0.88274656,  0.91767493,  0.90369209,  0.64917069,  0.61214883,
          0.90104305,  1.04678643,  1.24920747,  6.04067172,  2.26186731,
          1.07026566,  1.62663777,  5.99515177,  0.96906497,  0.9987026 ,
          1.41960866,  0.86803649,  1.14565132,  1.10774379,  1.18389276,
          0.83843433,  1.14792339,  0.9038267 ,  1.51593834,  1.29103672,
          1.43003373,  0.96934244,  0.87583202,  1.06060443,  8.04445156,
          1.44196559,  1.10803103,  1.08750665,  2.09671055,  1.15534847,
          1.14637643,  1.22763692,  1.47350503,  1.13439172,  1.15168505,
```

```

1.19876804, 1.26982093, 1.79159781, 2.17916227, 1.71617204,
1.51686826, 2.41694351, 1.80552464, 4.09364976, 1.3772735 ,
2.16947298, 1.67262181, 1.49514589, 1.60919369, 1.59990887,
1.52116268, 1.38034091, 1.37348925, 1.39492868, 1.33123752,
1.50803544, 1.436479 , 1.34931475, 1.9372756 , 1.37951499,
1.08405128, 1.2938159 , 1.18946552, 1.39442585, 1.22697137,
1.26384276, 1.0224159 , 1.06575806, 0.95675545, 1.07214764,
0.99820106, 1.09601038, 0.94396288, 0.91789148, 1.08170523,
0.90845572, 0.80435321, 0.75119906, 0.9454025 , 0.75931927,
0.89817409, 0.82743458, 0.76268299, 1.40870538, 0.7091193 ,
0.62541728, 1.38165634, 0.75782601, 2.25922605, 1.99888624,
0.70671984, 1.34816781, 0.80227047, 2.75211161, 0.26133333,
2.90741091, 0.15353398, 0.67282188, 2.98762093, 1.26364464,
3.03560477, 0.61901927, 1.2084012 , 0.56670838, 1.0579387 ,
0.73752369, 1.1850576 , 0.77833398, 1.5289542 , 1.21784498,
1.81804514, 0.80584398, 1.52090812, 1.43044018, 2.15565348,
1.28304516, 1.71467422, 1.66423832, 1.50595928, 2.83929483,
0.873116 , 0.82069335, 6.40107702, 1.71670569, 5.20948199,
2.28678886, 3.01120617, 0.99693978, 1.42082089, 2.21824395,
2.11009392, 2.05654467, 2.3358634 , 1.65092443, 1.34648951,
1.58783655, 1.05393364, 1.76195833, 1.95122771, 2.03076006,
2.31910308, 2.00207133, 2.866545 , 3.61963406, 39.23553794,
3.79515163, 3.01886825, 5.70062074, 7.30367551, 20.29354593,
20.6136468 , 4.45230036, 9.29035089, 10.9225366 , 7.53404552,
22.13293678, 38.07259698, 12.60837117, 12.93607504, 8.60834295,
9.79193128, 10.36880876, 21.23745326, 10.44550239, 11.56685523,
15.57855679, 12.04568533, 10.30210199, 6.86874433, 9.65658274,
30.97454035, 6.40053723, 4.54581346, 4.06106886, 6.02332255,
13.94292451, 7.85099643, 5.16014179, 4.68740178, 9.94387838,
6.0290411 , 6.61744159, 34.14657979, 2.948217 , 3.23277446,
12.49290731, 20.32505734, 3.81026758, 9.13902562, 1.91948225,
3.99761567, 16.47815701, 1.60130587, 8.27284557, 18.47464404,
0.95377367, 1.62160657, 2.64663925, 7.69234529, 0.49857021,
1.38041939, 1.71850861, 1.86860555, 4.10475717, 19.62015777,
4.9335346 , 3.24461049, 5.00677544, 7.05205896, 11.65790375,
1.82827698, 5.81347656, 3.9226461 , 12.76084232, 2.91101583,
9.75890136, 1.20231284, 1.79904888, 0.40866211, 1.92855546,
6.06748072, 2.90058778, 0.49860425, 4.76976202, 1.6716313 ,
3.94486996, 1.88578812, 14.41782414, 5.46589456, 0.57557327,
1.28460693])

```

```
[7]: dataset = dataset.reset_index(drop=True)
```

```
[8]: dataset.head()
```

```
[8]:
```

	Open	High	Low	Close
0	10.42	11.02	10.34	10.98

1	11.61	12.14	11.36	11.55
2	12.10	12.43	11.97	12.12
3	12.19	12.22	11.66	11.88
4	12.01	12.30	11.85	12.28

```
[9]: dataset['mahalanobis'] = mahalanobis_distance(x=dataset, data=dataset[['Open', 'High', 'Low', 'Close']])
dataset.head()
```

```
[9]:
```

	Open	High	Low	Close	mahalanobis
0	10.42	11.02	10.34	10.98	2.343602
1	11.61	12.14	11.36	11.55	2.443149
2	12.10	12.43	11.97	12.12	1.000510
3	12.19	12.22	11.66	11.88	1.218421
4	12.01	12.30	11.85	12.28	1.007601

```
[10]: dataset['p'] = 1 - chi2.cdf(dataset['mahalanobis'], 4)
dataset.head()
```

```
[10]:
```

	Open	High	Low	Close	mahalanobis	p
0	10.42	11.02	10.34	10.98	2.343602	0.672842
1	11.61	12.14	11.36	11.55	2.443149	0.654844
2	12.10	12.43	11.97	12.12	1.000510	0.909719
3	12.19	12.22	11.66	11.88	1.218421	0.875057
4	12.01	12.30	11.85	12.28	1.007601	0.908641