Code ▾

Missingness Imputation

Yiran Qin

Hide

```
library(quantmod)
library(magrittr)
library(VIM)
library(DMwR)
library(FNN)
```

Hide

```
start <- as.Date("2020-03-01")
end <- as.Date("2020-06-02")
```

According to the assignment requirements on the slides, I need to get Dow Jones index and 10 more companies data. Since they all used abbreviations on Yahoo finance, I will write their names down here. They are Apple, Goldman Sachs, Microsoft, Snapchat, Boeing, Google, Amazon, JP Morgan Chase, Alibaba and Nike.

Hide

```
getSymbols(c("^DJI", "AAPL", "GS", "MSFT", "SNAP", "BA", "GOOG", "AMZN", "JPM", "BABA", "NKE"),
 src = "yahoo", from = start, to = end)
```

Hide

```
stocks <- as.xts(data.frame(DJI = DJI[, "DJI.Close"], AAPL = AAPL[, "AAPL.Close"], GS = GS[, "G
S.Close"], MSFT = MSFT[, "MSFT.Close"], SNAP = SNAP[, "SNAP.Close"], BA = BA[, "BA.Close"], GOOG
= GOOG[, "GOOG.Close"], AMZN = AMZN[, "AMZN.Close"], JPM = JPM[, "JPM.Close"], BABA = BABA[, "BA
BA.Close"], NKE = NKE[, "NKE.Close"]))
```

The reason why the Date begins with ‘2020-03-02’ because ‘2020-03-01’ is a Saturday, stock market closes on weekends, holidays and meltdown.I think this also the part of the reason we need to find out missing data.

Hide

```
head(stocks)
```

```
           DJI.Close AAPL.Close GS.Close MSFT.Close SNAP.Close BA.Close GOOG.Close AMZN.Close JP
M.Close
2020-03-02  26703.32     298.81   209.47     172.79      14.39   289.27    1389.11    1953.95
121.52
2020-03-03  25917.41     289.32   203.43     164.51      13.55   280.62    1341.39    1908.99
116.96
2020-03-04  27090.86     302.74   208.74     170.55      13.63   283.12    1386.52    1975.83
119.85
2020-03-05  26121.28     292.92   198.79     166.27      13.85   260.37    1319.04    1924.03
113.97
2020-03-06  25864.78     289.03   192.85     161.57      13.00   262.33    1298.41    1901.09
108.08
2020-03-09  23851.02     266.17   172.81     150.62      11.45   227.17    1215.56    1800.61
93.44
           BABA.Close NKE.Close
2020-03-02     210.98     92.68
2020-03-03     207.41     90.93
2020-03-04     211.96     93.79
2020-03-05     211.46     90.58
2020-03-06     204.64     88.36
2020-03-09     197.66     84.11
```

Hide

```
summary(stocks)
```

```
     Index                         DJI.Close       AAPL.Close      GS.Close       MSFT.Close   
 Min.   :2020-03-02 00:00:00   Min.   :18592   Min.   :224.4   Min.   :135.0   Min.   :135.4  
 1st Qu.:2020-03-23 18:00:00   1st Qu.:22628   1st Qu.:259.2   1st Qu.:164.2   1st Qu.:157.3  
 Median :2020-04-15 12:00:00   Median :23702   Median :283.8   Median :177.1   Median :171.7  
 Mean   :2020-04-15 04:34:41   Mean   :23408   Mean   :281.5   Mean   :175.0   Mean   :167.7  
 3rd Qu.:2020-05-07 06:00:00   3rd Qu.:24376   3rd Qu.:304.7   3rd Qu.:183.5   3rd Qu.:180.9  
 Max.   :2020-06-01 00:00:00   Max.   :27091   Max.   :321.9   Max.   :209.7   Max.   :186.7  
   SNAP.Close       BA.Close        GOOG.Close     AMZN.Close     JPM.Close         BABA.Close   
 Min.   : 8.37   Min.   : 95.01   Min.   :1057   Min.   :1677   Min.   : 79.03   Min.   :176.3  
 1st Qu.:11.88   1st Qu.:128.85   1st Qu.:1181   1st Qu.:1909   1st Qu.: 88.95   1st Qu.:194.4  
 Median :13.62   Median :138.37   Median :1276   Median :2297   Median : 91.52   Median :200.0  
 Mean   :14.32   Mean   :150.84   Mean   :1267   Mean   :2162   Mean   : 93.63   Mean   :199.3  
 3rd Qu.:17.34   3rd Qu.:151.50   3rd Qu.:1373   3rd Qu.:2395   3rd Qu.: 95.86   3rd Qu.:206.6  
 Max.   :19.55   Max.   :289.27   Max.   :1432   Max.   :2498   Max.   :121.52   Max.   :217.2  
   NKE.Close    
 Min.   :62.80  
 1st Qu.:84.08  
 Median :87.08  
 Mean   :85.61  
 3rd Qu.:90.05  
 Max.   :99.87  
```

Hide

```
stocks1 <- knnImputation(stocks, k = 3, scale = T, meth = "median", distData = NULL)
```

```
No case has missing values. Stopping as there is nothing to do.
```

I checked ‘median’ column as a example by using knnImputation function. It shows there is no missing values, so we need to check other columns for example ‘Date’. First oF all, we need to add index as a variable in the dataset.

Hide

```
stocks <- as.data.frame(stocks)
stocks$Date <- row.names(stocks)
cln <- ncol(stocks)
stocks <- stocks[, c(cln, 1:(cln-1))]
row.names(stocks) <- NULL
```

Hide

```
head(stocks)
```

	Date <chr>	DJI.Close <dbl>	AAPL.Clo… <dbl>	GS.Clo… <dbl>	MSFT.Close <dbl>	SNAP.Close <dbl>	BA.Clo… <dbl>	GOOG.Cl… <dbl>
1	2020-03-02	26703.32	298.81	209.47	172.79	14.39	289.27	1389.11
2	2020-03-03	25917.41	289.32	203.43	164.51	13.55	280.62	1341.39
3	2020-03-04	27090.86	302.74	208.74	170.55	13.63	283.12	1386.52
4	2020-03-05	26121.28	292.92	198.79	166.27	13.85	260.37	1319.04
5	2020-03-06	25864.78	289.03	192.85	161.57	13.00	262.33	1298.41
6	2020-03-09	23851.02	266.17	172.81	150.62	11.45	227.17	1215.56

6 rows | 1-10 of 12 columns

Hide

```
summary(stocks)
```

```
     Date             DJI.Close       AAPL.Close        GS.Close        MSFT.Close       SNAP.Close
 Length:64          Min.   :18592   Min.   :224.4   Min.   :135.0   Min.   :135.4   Min.   : 8.3
7
 Class :character   1st Qu.:22628   1st Qu.:259.2   1st Qu.:164.2   1st Qu.:157.3   1st Qu.:11.8
8
 Mode  :character   Median :23702   Median :283.8   Median :177.1   Median :171.7   Median :13.6
2
                    Mean   :23408   Mean   :281.5   Mean   :175.0   Mean   :167.7   Mean   :14.3
2
                    3rd Qu.:24376   3rd Qu.:304.7   3rd Qu.:183.5   3rd Qu.:180.9   3rd Qu.:17.3
4
                    Max.   :27091   Max.   :321.9   Max.   :209.7   Max.   :186.7   Max.   :19.5
5
    BA.Close        GOOG.Close      AMZN.Close       JPM.Close         BABA.Close      NKE.Close
 Min.   : 95.01   Min.   :1057   Min.   :1677   Min.   : 79.03   Min.   :176.3   Min.   :62.80
 1st Qu.:128.85   1st Qu.:1181   1st Qu.:1909   1st Qu.: 88.95   1st Qu.:194.4   1st Qu.:84.08
 Median :138.37   Median :1276   Median :2297   Median : 91.52   Median :200.0   Median :87.08
 Mean   :150.84   Mean   :1267   Mean   :2162   Mean   : 93.63   Mean   :199.3   Mean   :85.61
 3rd Qu.:151.50   3rd Qu.:1373   3rd Qu.:2395   3rd Qu.: 95.86   3rd Qu.:206.6   3rd Qu.:90.05
 Max.   :289.27   Max.   :1432   Max.   :2498   Max.   :121.52   Max.   :217.2   Max.   :99.87
```

I realized that the data type of “Date” is not numeric and different than other variables, that’s why when I was using knnImputation function gave me 0 missing value. I removed the “Date” column, and use get.knn function on the new dataset. Finally, I got the knn index below.

Hide

```
new <- within(stocks, rm("Date"))
new
```

DJI.Close	AAPL.Clo…	GS.Clo…	MSFT.Close	SNAP.Close	BA.Clo…	GOOG.Cl…	AMZN.Cl…	JPI
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
26703.32	298.81	209.47	172.79	14.39	289.27	1389.110	1953.95	
25917.41	289.32	203.43	164.51	13.55	280.62	1341.390	1908.99	
27090.86	302.74	208.74	170.55	13.63	283.12	1386.520	1975.83	
26121.28	292.92	198.79	166.27	13.85	260.37	1319.040	1924.03	
25864.78	289.03	192.85	161.57	13.00	262.33	1298.410	1901.09	
23851.02	266.17	172.81	150.62	11.45	227.17	1215.560	1800.61	
25018.16	285.34	184.35	160.92	11.99	231.01	1280.390	1891.82	
23553.22	275.43	171.89	153.63	10.81	189.08	1215.410	1820.86	
21200.62	248.23	150.68	139.06	10.42	154.84	1114.910	1676.61	
23185.62	277.97	177.17	158.83	11.35	170.20	1219.730	1785.00	

1-10 of 64 rows | 1-9 of 11 columns Previous **1** 2 3 4 5 6 7 Next

Hide

```
get.knn(new, k=5)
```

```
$nn.index
      [,1] [,2] [,3] [,4] [,5]
 [1,]    3    4    2    5   61
 [2,]    5    4   61   62   64
 [3,]    1    4    2    5   61
 [4,]    2    5    1   61   64
 [5,]    2    4   61   62   64
 [6,]   29    8   28   31   44
 [7,]   60   42   62   63   55
 [8,]   28   29    6   10   30
 [9,]   12   18   25   24   23
[10,]   28    8   30   26   36
[11,]   14   13   17   23   25
[12,]   18    9   24   25   23
[13,]   14   11   15   17   23
[14,]   13   11   17   23   15
[15,]   16   13   14   11   17
[16,]   15   13   14   11   17
[17,]   23   25   18   12    9
[18,]   12   25    9   24   23
[19,]   27   26   21   36   22
[20,]   24   22   12   18    9
[21,]   19   27   26   22   20
[22,]   20   21   24   19   12
[23,]   25   17   18   12    9
[24,]   12   18   20    9   25
[25,]   23   18   12    9   17
[26,]   27   19   21   36   10
[27,]   26   19   21   36   10
[28,]   30    8   32   29   37
[29,]   44    8   28    6   45
[30,]   28   32   37   38   52
[31,]   46   48   41   40   45
[32,]   37   38   33   35   53
[33,]   38   37   32   35   53
[34,]   40   56   50   49   43
[35,]   47   53   33   54   39
[36,]   52   30   37   26   27
[37,]   38   32   33   35   53
[38,]   33   37   32   35   53
[39,]   45   51   35   54   44
[40,]   41   34   56   50   31
[41,]   40   31   34   56   50
[42,]   55   57   58   59   43
[43,]   49   59   58   56   50
[44,]   45   47   51   39   54
[45,]   44   51   47   39   54
[46,]   48   31   51   45   39
[47,]   53   54   44   45   35
[48,]   46   51   45   39   31
[49,]   43   50   34   56   59
[50,]   56   49   34   43   40
[51,]   45   54   44   47   48
```

```
[52,]   37   36   32   38   30
[53,]   47   54   35   33   38
[54,]   53   47   51   35   45
[55,]   57   42   58   59   43
[56,]   50   34   49   43   40
[57,]   55   58   59   42   43
[58,]   59   57   55   43   49
[59,]   58   57   55   43   49
[60,]   42   63   55   62   57
[61,]   64   62   63   60    5
[62,]   63   64   61   60    7
[63,]   62   64   61   60    7
[64,]   63   61   62   60    5

$nn.dist
           [,1]       [,2]       [,3]       [,4]       [,5]
 [1,] 388.24779  587.93295  788.83496  845.95310 1250.53975
 [2,]  72.30712  206.78867  642.05850  730.45799  733.71772
 [3,] 388.24779  973.70365 1176.33993 1231.98909 1608.76523
 [4,] 206.78867  258.66445  587.93295  767.00072  861.94720
 [5,]  72.30712  258.66445  623.13546  702.59071  713.23376
 [6,] 286.54072  301.07983  489.62090  504.11713  522.34695
 [7,] 556.12859  625.69401  657.88738  683.72660  695.86514
 [8,] 256.38800  280.45782  301.07983  370.15109  386.60195
 [9,] 140.48057  209.96383  276.31795  324.52960  347.09563
[10,] 359.23943  370.15109  436.10503  550.75279  569.75132
[11,] 221.66695  323.90263  576.56208  786.66897  891.42489
[12,]  95.45022  140.48057  209.04421  211.91859  311.63801
[13,] 196.23071  323.90263  725.78479  814.95531 1048.09435
[14,] 196.23071  221.66695  621.64174  857.57303  915.10838
[15,] 585.20556  725.78479  915.10838 1027.48634 1535.65743
[16,] 585.20556 1309.87329 1496.85570 1611.49610 2115.15853
[17,] 242.91316  351.39479  500.65919  548.95324  562.47232
[18,]  95.45022  154.03089  209.96383  219.36318  259.89190
[19,] 125.45703  141.37774  227.34591  601.77706  635.98131
[20,] 227.96325  290.02193  412.06109  436.59931  490.43939
[21,] 227.34591  332.58301  356.57390  410.96476  694.78297
[22,] 290.02193  410.96476  507.22721  635.98131  696.51288
[23,] 109.50318  242.91316  259.89190  311.63801  347.09563
[24,] 209.04421  219.36318  227.96325  324.52960  361.90465
[25,] 109.50318  154.03089  211.91859  276.31795  351.39479
[26,]  31.75028  141.37774  356.57390  474.89002  550.75279
[27,]  31.75028  125.45703  332.58301  484.34842  580.34017
[28,] 133.46513  256.38800  279.88561  286.07371  328.42120
[29,] 268.60964  280.45782  286.07371  286.54072  300.56854
[30,] 133.46513  185.70040  218.28730  269.05268  282.30622
[31,] 113.41625  154.56971  159.78472  206.65794  211.54736
[32,]  64.49779   94.46088  106.86412  169.83531  177.03982
[33,]  31.58009   77.35814  106.86412  114.70613  132.16286
[34,] 112.18349  128.81878  133.35337  142.39221  158.43527
[35,]  99.59096  101.94486  114.70613  120.87621  127.87942
[36,] 271.60347  404.95670  460.84799  474.89002  484.34842
[37,]  55.31270   64.49779   77.35814  177.58289  181.77272
[38,]  31.58009   55.31270   94.46088  136.17892  141.87164
```

```
[39,] 109.99678  114.73757  127.87942  133.06558  141.26583
[40,]  81.97476  112.18349  146.60572  162.17036  206.65794
[41,]  81.97476  159.78472  163.52010  225.19102  232.30205
[42,]  83.76140  156.94746  189.38706  196.45280  305.68676
[43,] 105.23464  142.20372  144.55470  146.02830  152.39403
[44,]  40.83127   93.81094  102.29802  141.26583  142.50965
[45,]  40.83127   69.29416   95.38806  109.99678  125.15865
[46,]  55.41662  113.41625  128.01042  135.84310  160.25099
[47,]  56.34155   68.55511   93.81094   95.38806   99.59096
[48,]  55.41662  112.25026  144.52696  145.75431  154.56971
[49,] 105.23464  114.53172  142.39221  144.64959  147.74157
[50,]  54.10700  114.53172  133.35337  152.39403  162.17036
[51,]  69.29416   95.81040  102.29802  105.04044  112.25026
[52,] 246.43644  271.60347  279.78901  281.49894  282.30622
[53,]  56.34155   66.10978  101.94486  132.16286  141.87164
[54,]  66.10978   68.55511   95.81040  120.87621  125.15865
[55,]  78.41290   83.76140  126.52069  136.27121  259.93572
[56,]  54.10700  128.81878  144.64959  146.02830  146.60572
[57,]  78.41290  114.30608  127.77576  156.94746  240.68847
[58,]  19.94802  114.30608  126.52069  144.55470  159.22756
[59,]  19.94802  127.77576  136.27121  142.20372  147.74157
[60,] 373.70198  388.78342  399.82054  406.16119  426.91770
[61,]  96.93285  148.19590  169.33067  553.50708  623.13546
[62,]  47.64674  103.56268  148.19590  406.16119  657.88738
[63,]  47.64674   96.64094  169.33067  388.78342  683.72660
[64,]  96.64094   96.93285  103.56268  482.78665  713.23376
```