edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

zhushun0008 (/dashboard) ▾

Courseware (/courses/MITx/15.071x/1T2014/courseware) Course Info (/courses/MITx/15.071x/1T2014/info)

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum) Progress (/courses/MITx/15.071x/1T2014/progress)

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

Help

CLUSTERING STOCK RETURNS

When building portfolios of stocks, investors seek to obtain good returns while limiting the variability in those returns over time. This can be achieved by selecting stocks that show different patterns of returns. In this problem, we will use clustering to identify clusters of stocks that have similar returns over time; an investor might select a diverse portfolio by selecting stocks from different clusters.

For this problem, we'll use nasdaq_returns.csv (/c4x/MITx/15.071x/asset/nasdaq_returns.csv), which contains monthly stock returns from the NASDAQ stock exchange from 2000-2009, limiting to tickers that were listed on the exchange that entire period and whose stock price never fell below $1. The NASDAQ is the second-largest stock exchange in the world, and it lists many technology companies. The stock price data used in this problem was obtained from infochimps (http://www.infochimps.com/datasets/nasdaq-exchange-daily-1970-2010-open-close-high-low-and-volume), a website providing access to many datasets, and the industry information was obtained from Yahoo! Finance (http://biz.yahoo.com/). This dataset contains the following variables:

- **stock_symbol**: The symbol identifying the company for the stock
- **industry**: The industry the stock is classified under
- **subindustry**: The sub-industry the stock is classified under
- **ret2000.01-ret2009.12**: The return for the stock during the variable's indicated month. The variable names have format "retYYYY.MM", where YYYY is the year and MM is the month. For instance, variable ret2005.02 refers to February 2005. The value stored is a proportional change in stock value during that month. For instance, a value of 0.05 means the stock increased in value 5% during the month, while a value of -0.02 means the stock decreased in value 2% during the month. There are 120 of these variables, for the 120 months in our dataset.

PROBLEM 1 - LOADING THE DATA (1 point possible)

Load nasdaq_returns.csv into a data frame called "stocks". How many companies are in the dataset?

Check Save *You have used 0 of 2 submissions*

PROBLEM 2 - SUMMARIZING THE DATA (1 point possible)

For which industries are there 40 or more companies in our dataset?

☐ Basic Materials
☐ Conglomerates
☐ Consumer Cyclical
☐ Consumer Goods
☐ Financial

- □ Healthcare
- □ Industrial Goods
- □ Services
- □ Technology
- □ Utilities

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 3 - STOCK TRENDS IN THE DATA (2 points possible)

In the aftermath of the dot-com bubble bursting in the early 2000s, the NASDAQ was quite tumultuous. In December 2000, how many stocks in our dataset saw their value increase by 10% or more?

In December 2000, how many stocks in our dataset saw their value decrease by 10% or more?

Check Save *You have used 0 of 2 submissions*

PROBLEM 4 - STOCK TRENDS IN THE DATA (2 points possible)

Entering the Great Recession most stocks lost significant value, but some sectors were hit harder than others. In October 2008, which of the following industries had the worst average return across the stocks in that industry?

- ○ Basic Materials
- ○ Consumer Goods
- ○ Financial
- ○ Healthcare
- ○ Industrial Goods
- ○ Services
- ○ Technology

February 2000 was the third strongest month in the dataset in terms of average returns. However, which of the following industries actually had a negative average return during that month?

- ○ Basic Materials
- ○ Consumer Goods
- ○ Financial
- ○ Healthcare
- ○ Industrial Goods
- ○ Services
- ○ Technology

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 5 - PREPARING THE DATASET (2 points possible)

Copy the stocks data frame into a new data frame called "limited", and remove the first three variables of limited: stock_symbol, industry, and subindustry.

Now, identify the month with the largest average return across all stocks in the dataset. What is the variable name associated with this month (for instance, if your answer were February 2004, you would answer ret2004.02)?

Identify the month with the lowest average return across all the stocks in the dataset. What is the variable name associated with this month?

Check Save *You have used 0 of 2 submissions*

PROBLEM 6 - PREPARING FOR CLUSTERING (1 point possible)

We are about to cluster our data. Why did we remove the stock_symbol, industry, and subindustry variables prior to clustering our data?

○ No reason -- we could have clustered the dataset with these variables still in it.

○ If we had included these variables in our clustering analysis, they would have caused some of the pairwise distance calculations to fail.

○ While we could have run the clustering analysis with these variables in our dataset, we removed them so they don't bias our analysis.

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 7 - NORMALIZING (1 point possible)

In this analysis, we will not normalize our data prior to clustering. Why is this a valid approach?

○ All the variables have the same scale, so no normalization is necessary

○ Because this dataset is so large, normalization would be prohibitively slow

○ Normalization would have caused an error for this dataset

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 8 - HIERARCHICAL CLUSTERING (1 point possible)

Using Euclidean distances (the default) and the Ward method, perform hierarchical clustering on the "limited" data frame, and plot the resulting dendrogram.

Which of the following number of clusters is least appropriate, based on the dendrogram?

○ 2

○ 3

○ 4

○ 5

○ 6

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 9 - THE HIERARCHICAL CLUSTERS (1 point possible)

Extract cluster assignments from your hierarchical clustering object, using 5 clusters in total. Which cluster has the largest number of stocks?

○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 10 - UNDERSTANDING THE CLUSTERS (2 points possible)

Which cluster best fits the description "healthcare and technology stocks"?

○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

Which of the following industries have more than half of their stocks assigned to a single cluster?

□ Basic Materials

□ Consumer Goods

□ Financial

□ Healthcare

□ Industrial Goods

□ Services

□ Technology

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 11 - SUB-INDUSTRIES (2 points possible)

We can get a finer-grained understanding of the composition of the clusters by looking at subindustry information. Which cluster contains nearly all companies categorized in the subindustry "Apparel Stores" (part of the services industry)?

○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

Which cluster contains all stocks categorized in sub-industry "Electronics Wholesale" (another part of the services industry)?

○ Cluster 1
○ Cluster 2
○ Cluster 3
○ Cluster 4
○ Cluster 5

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 12 - STOCK TRENDS IN THE CLUSTERS (2 points possible)

For some months, we expect there to be significant differences between the returns of stocks in different clusters. In February 2000, the average return of stocks in Cluster 3 was negative, while the average return of stocks in one of the other clusters was more than 100%. What cluster had the average return exceeding 100%?

○ Cluster 1
○ Cluster 3
○ Cluster 4
○ Cluster 5

For which of the following months did one cluster have an average return exceeding 30% and another cluster have a negative average return?

□ March 2000
□ May 2005
□ October 2009
□ December 2009

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 13 - USING A VISUALIZATION (1 point possible)

Which of the following visualizations could be used to observe the distribution of stock returns in February 2000, broken down by cluster? Select all that apply.

□ A box plot of the variable ret2000.02, subdivided by cluster
□ A box plot of the clusters, subdivided by ret2000.02 values
□ ggplot with the cluster number on the x-axis and ret2000.02 on the y-axis, plotting with geom_line()
□ ggplot with ret2000.02 on the x-axis and the cluster number on the y-axis, plotting with geom_point()

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 14 - K-MEANS CLUSTERING (1 point possible)

Now set the seed to 144 and immediately afterward run k-means clustering on the "limited" data frame, using 5 clusters. How many stocks are in the smallest cluster?

Check Save *You have used 0 of 2 submissions*

PROBLEM 15 - COMPARING CLUSTERING ALGORITHMS (1 point possible)

k-means cluster number 4 contains more than half of its members from which hierarchical cluster?

- ○ Hierarchical Cluster 1
- ○ Hierarchical Cluster 2
- ○ Hierarchical Cluster 3
- ○ Hierarchical Cluster 4
- ○ Hierarchical Cluster 5
- ○ It contains fewer than half of its members from any one hierarchical cluster

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 16 - RANDOM BEHAVIOR (2 points possible)

If we re-ran hierarchical clustering a second time without making any additional calls to set.seed(), would we expect:

- ○ Different results from the first hierarchical clustering
- ○ Identical results to the first hierarchical clustering

If we re-ran k-means clustering a second time without making any additional calls to set.seed(), would we expect:

- ○ Different results from the first k-means clustering
- ○ Identical results to the first k-means clustering

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 17 - CREATING A DIVERSE PORTFOLIO (1 point possible)

In the introduction to the problem, we discussed the value of a diverse portfolio and how we might achieve this objective by selecting stocks from different clusters. Consider an investor with a large holding of stock from the company with stock_symbol AAPL. Which of the following stock symbols is neither in the same hierarchical cluster nor in the same k-means cluster as AAPL?

- □ AMZN
- □ MSFT
- □ TROW

Final Check Save *You have used 0 of 1 submissions*

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.