Details:

I used k-means and hierarchical (with a euclidean distance linkage, aka ‘Ward’) algorithms to build cluster models. They’re similar enough in their outputs, so much of the results reported below are just k-means results.   
There were some 3000 unique names in the factor table Scott build. Being ultra-conservative and allowing no non-null elements on a given quarter, the number of securities considered fell to 1595 (unacceptable!). Assuming this truncation wouldn’t bias the results (probably a bad assumption), I used it as a starting point. The only other manipulation performed was standardization of factors, X-> (X – mu(X) )/sigma(X). the data matrix trained on was Nsymbol\*Nfactor-dimensional *for* *q0*. It might be useful to build in the remaining fifteen quarters via some pairwise distance between factor time series, but that might require building a Nsecurity\*Nsecurity-dimensional matrix, one for each factor, though could svd or pca the space to get rid of linear dependencies and reduce the rank.

Things to look at in the future:  
-Please add to this list,e.g., if there’s a subset of factors you think might give better resolution, let me know.   
-checking how much factor-based results change assuming 0s for all nulls (ultra-aggressive – can we find some middle-ground?).   
-concatenating text and factor features and repeating analysis  
-for homogeneity, run against basic, independent (to factor construction) quantites other than daily/hourly return. (I have price data, but that’s probably not useful.)  
-consider ways of including other quarter data—should come after the NULL problem is sorted out, e..g, by truncating the factor space to include (nearly-) complete data   
-backtest on labels?   
-consider different linkage kernels in HC models – could smooth volumes of labels

Classification models-

For each estimator, clusters with top-10 market-cap (mc) member names and cluster center (in the factor basis) ordered by absolute value of the projection onto that dimension are given below. E.g., for cluster ‘0’, “rect\_\_act\_m\_che” and “ppent\_ppegt” are the most important. This Do these centroids make any sense whatsoever? The blue counter reads as 822 names in cluster ‘8’, 508 in ‘0’, etc. The left-most column is just an index. I chose n\_clusters=11 to compare against gics sectors.

*factor-based-*

Counter({8: 822, 0: 508, 1: 189, 4: 38, 7: 26, 6: 6, 10: 2, 2: 1, 3: 1, 5: 1, 9: 1})

cluster 0:

name mc gics8

502 EXXON MOBIL CORP 359893.717184 10102010

285 MICROSOFT CORP 222801.603195 45103020

215 INTL BUSINESS MACHINES CORP 205308.369433 45102010

119 CHEVRON CORP 197974.835570 10102010

233 JOHNSON & JOHNSON 180330.177504 35202010

186 GOOGLE INC 174550.487974 45101010

178 GENERAL ELECTRIC CO 172891.489407 20105010

242 COCA-COLA CO 161541.843002 30201030

333 PFIZER INC 148084.366888 35202010

406 SCHLUMBERGER LTD 105412.577930 10101020

*Cluster 0 Centroid-*[['rect\_\_act\_m\_che, ppent\_ppegt, capxq\_\_act\_m\_che\_ppent, dpq\_revtq2, dpq\_gpq, cfroq\_revtq2, rect\_revtq2, oiadpq\_revtq2, gpq\_revtq2']]

cluster 1:

name mc gics8

509 APPLE INC 356700.380215 45202010

637 QUALCOMM INC 86298.418465 45201020

676 UNITEDHEALTH GROUP INC 50798.880489 35102030

683 VMWARE INC -CL A 39781.045135 45103020

570 EXPRESS SCRIPTS INC 22916.341399 35102015

613 MEDCO HEALTH SOLUTIONS INC 21198.012172 35102015

537 BROADCOM CORP 19072.750816 45301020

584 HUMANA INC 12952.643330 35102030

669 SYMANTEC CORP 12858.137894 45103020

554 CITRIX SYSTEMS INC 11362.108851 45103010

[['gpq\_\_act\_m\_che\_ppent, lct\_m\_dlc\_\_act\_m\_che\_ppent, cfroq\_\_act\_m\_che\_ppent, xrdq2\_\_act\_m\_che\_ppent, rect\_\_act\_m\_che, revtq2\_ppent, ap\_\_act\_m\_che\_ppent, lct\_m\_dlc\_\_act\_m\_che, capxq\_\_act\_m\_che\_ppent']]

cluster 2:

name mc gics8

697 WORLD HEART CORP 11.740225 35101010

[['invt\_revtq2, xrdq2\_revtq2, cogsq\_revtq2, rect\_revtq2, ap\_revtq2, dpq\_revtq2, invt\_\_act\_m\_che, xrdq2\_\_act\_m\_che\_ppent, capxq\_revtq2']]

cluster 3:

name mc gics8

698 DAXOR CORP 44.544089 35101010

[['lct\_m\_dlc\_\_act\_m\_che, lct\_m\_dlc\_\_act\_m\_che\_ppent, ap\_revtq2, ppent\_ppegt, dpq\_revtq2, xrdq2\_\_act\_m\_che\_ppent, invt\_\_act\_m\_che, xrdq2\_revtq2, ap\_\_act\_m\_che\_ppent']]

cluster 4:

name mc gics8

708 EOG RESOURCES INC 24865.567952 10102020

703 CONTINENTAL RESOURCES INC 10089.653920 10102020

724 PIONEER NATURAL RESOURCES CO 9086.761320 10102020

705 CABOT OIL & GAS CORP 7909.532500 10102020

719 NEWFIELD EXPLORATION CO 6871.516381 10102020

735 CIMAREX ENERGY CO 6083.070083 10102020

732 ULTRA PETROLEUM CORP 5122.904282 10102020

716 HERTZ GLOBAL HOLDINGS INC 4663.858612 20304020

725 PLAINS EXPLORATION & PROD CO 4146.809978 10102020

702 BRIGHAM EXPLORATION CO 3388.479379 10102020

[['capxq\_revtq2, capxq\_\_act\_m\_che\_ppent, dpq\_revtq2, dpq\_gpq, ppent\_ppegt, ap\_revtq2, rect\_\_act\_m\_che, lct\_m\_dlc\_\_act\_m\_che, revtq2\_pyoy']]

cluster 5:

name mc gics8

737 CADENCE PHARMACEUTICALS INC 415.833268 35202010

[['revtltm\_pqoq, revtq2\_pqoq, invt\_\_act\_m\_che, ppent\_ppegt, capxq\_\_act\_m\_che\_ppent, ap\_\_act\_m\_che\_ppent, xrdq2\_\_act\_m\_che\_ppent, lct\_m\_dlc\_\_act\_m\_che\_ppent, ap\_revtq2']]

cluster 6:

name mc gics8

742 SAVIENT PHARMACEUTICALS INC 306.800709 35201010

738 ADVANCED CELL TECHNOLOGY INC 263.719339 35201010

740 AMICUS THERAPEUTICS INC 158.799360 35201010

741 NABI BIOPHARMACEUTICALS 81.052293 35201010

743 VERMILLION INC 41.324363 35101010

739 ORAGENICS INC 13.355238 35201010

[['xrdq2\_\_act\_m\_che\_ppent, ap\_\_act\_m\_che\_ppent, lct\_m\_dlc\_\_act\_m\_che\_ppent, ap\_revtq2, lct\_m\_dlc\_\_act\_m\_che, xrdq2\_revtq2, dpq\_capxq, act\_m\_che\_\_ppent, revtltm\_pqoq']]

cluster 7:

name mc gics8

758 MCKESSON CORP 19687.102037 35102010

762 NVR INC 3551.027124 25201030

748 D R HORTON INC 3323.962822 25201030

752 INGRAM MICRO INC 2808.750628 45203030

753 WORLD FUEL SERVICES CORP 2641.953936 10102030

768 WATSCO INC 1965.187258 20107010

755 JAZZ PHARMACEUTICALS INC 1790.197724 35202010

759 MERITAGE HOMES CORP 607.031780 25201030

766 SAGENT PHARMACEUTICALS INC 600.522229 35202010

764 PRESTIGE BRANDS HOLDINGS 540.912492 30302010

[['act\_m\_che\_\_ppent, revtq2\_ppent, invt\_\_act\_m\_che, ap\_\_act\_m\_che\_ppent, rect\_revtq2, dpq*\_*capxq, revtq2\_pqoq, cogsq\_revtq2, invt\_revtq2']]

cluster 8:

name mc gics8

1339 PROCTER & GAMBLE CO 176122.136860 30301010

1349 PHILIP MORRIS INTERNATIONAL 122100.113764 30203010

1140 INTEL CORP 106387.045565 45301020

1255 MERCK & CO 102032.300835 35202010

1526 UNITED TECHNOLOGIES CORP 67471.936661 20101010

1175 KRAFT FOODS INC 61217.290622 30202030

1241 3M CO 58877.275230 20105010

1247 ALTRIA GROUP INC 56317.275330 30203010

809 AMGEN INC 51194.218872 35201010

876 BRISTOL-MYERS SQUIBB CO 50575.000000 35202010

[['invt\_\_act\_m\_che, oiadpq\_revtq2, cfroq\_revtq2, cogsq\_revtq2, invt\_revtq2, revtq2\_pqoq, gpq\_revtq2, ap\_revtq2, xrdq2\_revtq2']]

cluster 9:

name mc gics8

1592 TRANSWITCH CORP 86.67639 45301020

[['dpq\_capxq, xrdq2\_\_act\_m\_che\_ppent, lct\_m\_dlc\_\_act\_m\_che\_ppent, rect\_\_act\_m\_che, gpq\_\_act\_m\_che\_ppent, act\_m\_che\_\_ppent, rect\_revtq2, revtq2\_ppent, gpq\_revtq2']]

cluster 10:

name mc gics8

1594 MOLYCORP INC 4741.765084 15104020

1593 GEVO INC 276.963744 10102030

[['revtq2\_pyoy, capxq\_\_act\_m\_che\_ppent, ppent\_ppegt, revtltm\_pqoq, revtq2\_pqoq, invt\_\_act\_m\_che, capxq\_revtq2, xrdq2\_\_act\_m\_che\_ppent, dpq\_gpq']]

*text-based-*

Counter({7: 580, 0: 511, 2: 466, 4: 314, 10: 306, 5: 258, 3: 237, 6: 232, 9: 218, 1: 99, 8: 43})

cluster 0:

name mc gics8

0 APPLE INC 794094.976211 45202030

156 ALPHABET INC 657848.626617 45101010

250 MICROSOFT CORP 522924.322929 45103020

364 AT&T INC 233297.240554 50101020

393 VISA INC 211559.697815 45102020

410 VERIZON COMMUNICATIONS INC 183734.783495 50101020

280 ORACLE CORP 181759.542465 45103020

192 INTEL CORP 165850.985748 45301020

86 CISCO SYSTEMS INC 157119.655798 45201020

179 INTL BUSINESS MACHINES CORP 141657.339506 45102010

[[‘services, solutions, software, data, management, company, cloud, technology, provides, security, digital, products, segment, communications, information, platform, video, enterprise, mobile']]

cluster 1:

name mc gics8

708 WAL-MART STORES INC 236341.922791 30101040

547 HOME DEPOT INC 187307.096329 25504030

606 PEPSICO INC 163062.898254 30201030

586 MCDONALD'S CORP 119813.145523 25301040

700 WALGREENS BOOTS ALLIANCE INC 88022.629905 30101010

638 STARBUCKS CORP 86601.410637 25301040

597 NIKE INC 85375.360504 25203020

484 COSTCO WHOLESALE CORP 74936.009681 30101040

580 LOWE'S COMPANIES INC 72787.302643 25504030

587 MONDELEZ INTERNATIONAL INC 68441.160278 30202030

[[‘stores, products, company, restaurants, brands, retail, accessories, apparel, food, brand, operates, segment, home, footwear, foods, states, merchandise, com, operated']]

cluster 2:

name mc gics8

739 AMGEN INC 116274.357086 35201010

716 ABBVIE INC 104441.344030 35201010

779 CELGENE CORP 91689.401192 35201010

765 BRISTOL-MYERS SQUIBB CO 88575.657990 35202010

833 GILEAD SCIENCES INC 84210.010399 35201010

759 BIOGEN INC 53624.844169 35201010

919 REGENERON PHARMACEUTICALS 48598.100961 35201010

958 VERTEX PHARMACEUTICALS INC 29351.716469 35201010

850 INCYTE CORP 26347.805886 35201010

737 ALEXION PHARMACEUTICALS INC 26134.079590 35201010

[[‘treatment, clinical, phase, company, cancer, therapeutics, development, diseases, patients, biopharmaceutical, product, pharmaceuticals, developing, stage, candidates, candidate, trial, drug, cell']]

cluster 3:

name mc gics8

1029 GOLDMAN SACHS GROUP INC 84943.354559 40203020

1060 MORGAN STANLEY 76503.752463 40203020

993 BLACKROCK INC 62855.652022 40203010

1084 PNC FINANCIAL SVCS GROUP INC 56919.600296 40101015

1094 SCHWAB (CHARLES) CORP 51241.724034 40203020

991 BANK OF NEW YORK MELLON CORP 48042.318193 40203010

1099 STATE STREET CORP 30067.400170 40203010

990 FRANKLIN RESOURCES INC 23534.445603 40203010

1071 NORTHERN TRUST CORP 19914.289022 40203010

983 TD AMERITRADE HOLDING CORP 19298.399597 40203020

[[‘investment, investments, mortgage, debt, loans, equity, securities, funds, services, capital, companies, management, invests, million, company, fund, asset, income, financial']]

cluster 4:

name mc gics8

1274 SIMON PROPERTY GROUP INC 48082.260566 60101070

1250 PUBLIC STORAGE 37506.181407 60101080

1249 PROLOGIS INC 29039.766948 60101020

1197 WELLTOWER INC 26587.627811 60101050

1138 AVALONBAY COMMUNITIES INC 26435.738280 60101060

1179 EQUITY RESIDENTIAL 24014.481987 60101060

1292 VENTAS INC 23928.412643 60101050

1191 GGP INC 19761.541373 60101070

1146 BOSTON PROPERTIES INC 18503.446124 60101040

1170 DIGITAL REALTY TRUST INC 18292.843967 60101080

[[‘estate, properties, real, trust, investment, reit, company, office, realty, ownership, invests, property, hotels, communities, shopping, self, centers, income, development']]

cluster 5:

name mc gics8

1436 JOHNSON & JOHNSON 342172.590836 35202010

1489 PROCTER & GAMBLE CO 229101.167328 30301010

1488 PFIZER INC 192283.940965 35202010

1462 MERCK & CO 175086.441015 35202010

1533 UNITEDHEALTH GROUP INC 165449.247055 35102030

1445 LILLY (ELI) & CO 86254.191826 35202010

1377 CVS HEALTH CORP 78688.742821 30101010

1302 ABBOTT LABORATORIES 74581.993837 35101010

1524 THERMO FISHER SCIENTIFIC INC 67188.109435 35203010

1360 COLGATE-PALMOLIVE CO 65301.840425 30301010

[[‘care, products, health, medical, services, company, healthcare, surgical, diagnostic, segment, hospitals, systems, pharmaceutical, blood, therapy, patient, provides, devices, treatment']]

cluster 6:

name mc gics8

1783 EXXON MOBIL CORP 346374.750000 10102010

1595 CHEVRON CORP 199250.531264 10102010

1677 NEXTERA ENERGY INC 64087.921143 55101010

1605 DUKE ENERGY CORP 58275.000000 55101010

1585 CONOCOPHILLIPS 57426.365596 10102020

1622 ENTERPRISE PRODS PRTNRS -LP 56635.612403 10102040

1621 EOG RESOURCES INC 52928.803815 10102020

1740 SOUTHERN CO 49501.192685 55101010

1597 DOMINION RESOURCES INC 49263.279808 55103010

1697 OCCIDENTAL PETROLEUM CORP 46096.632965 10102010

[[‘gas, natural, oil, energy, crude, company, exploration, basin, texas, production, gathering, properties, midstream, segment, storage, petroleum, wells, liquids, interests']]

cluster 7:

name mc gics8

1972 GENERAL ELECTRIC CO 238635.296259 20105010

2102 ALTRIA GROUP INC 136505.157282 30203010

2100 3M CO 116605.055559 20105010

1845 BOEING CO 107467.497103 20101010

2003 HONEYWELL INTERNATIONAL INC 99223.533015 20105010

2300 UNITED TECHNOLOGIES CORP 95642.526253 20101010

2287 TEXAS INSTRUMENTS INC 79050.149702 45301020

2144 NVIDIA CORP 77845.954285 45301020

2071 LOCKHEED MARTIN CORP 77373.919368 20101010

1923 DOW CHEMICAL 74010.755364 15101020

[[‘products, systems, segment, equipment, company, industrial, power, manufactures, components, markets, solutions, materials, applications, used, services, manufacturing, technologies, control, designs']]

cluster 8:

name mc gics8

2351 AMERICAN INTERNATIONAL GROUP 57858.846625 40301030

2400 METLIFE INC 53860.600495 40301020

2346 AETNA INC 46514.289583 35102030

2414 PRUDENTIAL FINANCIAL INC 44410.669890 40301020

2360 CIGNA CORP 41020.343264 35102030

2402 MARSH & MCLENNAN COS 38369.296112 40301010

2428 TRAVELERS COS INC 33773.871233 40301040

2354 ALLSTATE CORP 30922.800446 40301040

2349 AFLAC INC 29340.707926 40301020

2410 PROGRESSIVE CORP-OHIO 24148.012823 40301040

[[‘insurance, life, property, casualty, products, segment, liability, company, services, reinsurance, group, commercial, health, personal, policies, coverage, financial, annuities, automobile']]

cluster 9:

name mc gics8

2468 AMAZON.COM INC 458158.215332 25502020

2593 FACEBOOK INC 427918.690613 45101010

2685 COCA-COLA CO 187157.396740 30201030

2528 COMCAST CORP 183410.139714 25401025

2785 PHILIP MORRIS INTERNATIONAL 176405.715031 30203010

2562 DISNEY (WALT) CO 170688.000488 25401030

2850 SCHLUMBERGER LTD 99216.270424 10101020

2798 REYNOLDS AMERICAN INC 93572.119539 30203010

2900 UNITED PARCEL SERVICE INC 89672.112122 20301010

2772 PRICELINE GROUP INC 88329.024590 25502020

[[‘services, company, segment, operates, water, transportation, energy, operations, provides, united, states, products, entertainment, segments, international, america, homes, management, offers']]

cluster 10:

name mc gics8

3118 JPMORGAN CHASE & CO 298293.420587 40101010

3256 WELLS FARGO & CO 261728.984637 40101010

2965 BANK OF AMERICA CORP 226813.078317 40101010

2997 CITIGROUP INC 165415.745681 40101010

3246 U S BANCORP 85840.615600 40101010

3028 CAPITAL ONE FINANCIAL CORP 38268.779929 40202010

2969 BB&T CORP 34183.018967 40101015

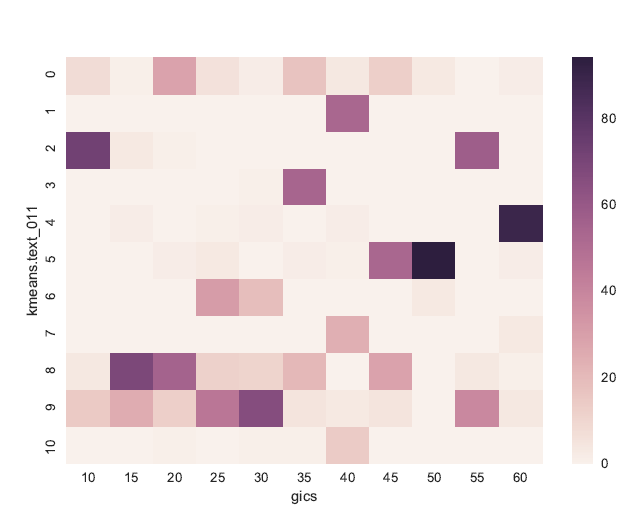
3216 SUNTRUST BANKS INC 26262.447692 40101015

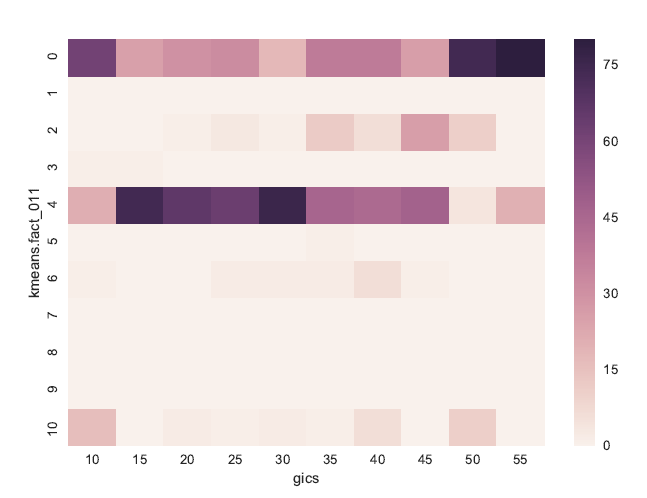
3138 M & T BANK CORP 24151.364442 40101015

3042 DISCOVER FINANCIAL SVCS 22665.826093 40202010

[[‘loans, bank, banking, services, commercial, deposit, accounts, financial, company, estate, mortgage, savings, holding, real, credit, deposits, consumer, bancorp, management']]

Below are heatmaps of cluster labels versus gics sector label. Columns sum to 100 and dark squares mean high overlap, e.g., gics 60 is a combination of ~80% kmeans.text cluster ‘4’ and small amplitude everywhere else.   
the model trained on the factor matrix smears gics across three labels mostly (highlighted in red above).





The below (top-daily, bottom-hourly) are meant to quantify homogeneity of clusters furnished by labeling schemes. They’re plotted as follows:   
For a given scheme, for each cluster label assigned by that scheme, compute the average group return over some time window (below are 1d and 1h-window returns since last September). then for each security (s) assigned to each group (G) at time t, collect the R2 on fitting its return as a function of its group mean at that time, e.g, return\_S\_t = alpha+beta\*<return\_G >\_t. For each group (a group must have more than 1 member!), take an average of member returns, then for each scheme, take a weighted average of group average R2s as the homogeneity measure (the y axis below is mislabeled). A measure of inter-group similarity is given the same way, but with <return\_G> replaced by <return not-G>. The O-marker series are intra-group homogeneities. The x-marker series are inter-group. Gics-,factor-,and text-based clustering all perform similarly across n\_clusters. (Clearly, though, the clustering due to factor data is on different footing because of NULL issues and the awkward distribution of cluster volumes for a given n\_cluster value – most of the ~1600 companies are in 3 clusters and 4 clusters are 1-member clusters in the n\_clusters=11 scheme. For n\_clusters=68, half of clusters have fewer than 7 members.)  
By inspection of companies and examination against gics bins, clustering on factors as it stands now, as I’ve done it, doesn’t look that good. If you accept the homogeneity measure, it’s indistinguishable from the other model and gics, but you have to admit that, for these results, n\_clusters=11 is really n\_clusters= <<11, e.g., ~3-4.

If possible, it might be useful to use rankings by one or more of these labelings in a back-test to measure differences produced in observables relative to those from the current method of grouping securities (by gics or alam-gics designation, right?). Relative values of observables might be the ultimate measure of the usefulness of any one grouping scheme.

