

01 Machine Learning Fundamentals

Shahrokh Vahedi

2023-06-05

Contents

Challenge Summary	1
Objectives	2
Libraries	2
Data	2
Question	21
Step 1 - Convert stock prices to a standardized format (daily returns)	22
Step 2 - Convert to User-Item Format	23
Step 3 - Perform K-Means Clustering	24
Step 4 - Find the optimal value of K	24
Step 5 - Apply UMAP	25
Step 6 - Combine K-Means and UMAP	27

Challenge Summary

Your organization wants to know which companies are similar to each other to help in identifying potential customers of a SAAS software solution (e.g. Salesforce CRM or equivalent) in various segments of the market. The Sales Department is very interested in this analysis, which will help them more easily penetrate various market segments.

We will be using stock prices in this analysis. We come up with a method to classify companies based on how their stocks trade using their daily stock returns (percentage movement from one day to the next). This analysis will help our organization determine which companies are related to each other (competitors and have similar attributes).

We can analyze the stock prices using what we've learned in the unsupervised learning tools including K-Means and UMAP. We will use a combination of `kmeans()` to find groups and `umap()` to visualize similarity of daily stock returns.

Objectives

Apply our knowledge on K-Means and UMAP along with `dplyr`, `ggplot2`, and `purrr` to create a visualization that identifies subgroups in the S&P 500 Index. We will specifically apply:

- Modeling: `kmeans()` and `umap()`
- Iteration: `purrr`
- Data Manipulation: `dplyr`, `tidyr`, and `tibble`
- Visualization: `ggplot2` (bonus `plotly`)

Libraries

Load the following libraries.

```
library(tidyverse)
library(tidyquant)
library(broom)
library(umap)
library(plotly)
library(dplyr)
library(magrittr) # needs to be run every time you start R and want to use %>%
library(lubridate)
```

Data

We will be using stock prices in this analysis. We can read in the stock prices. The data is 1.2M observations. The most important columns for our analysis are:

- **symbol**: The stock ticker symbol that corresponds to a company's stock price
- **date**: The timestamp relating the symbol to the share price at that point in time
- **adjusted**: The stock price, adjusted for any splits and dividends (we use this when analyzing stock data over long periods of time)

```
sp_500_index_tbl <- readRDS("data01/sp_500_index_tbl.rds")
sp_500_index_tbl
```

##	symbol	company	weight
## 1	MSFT	Microsoft Corporation	3.589659e-02
## 2	AAPL	Apple Inc.	3.299844e-02
## 3	AMZN	Amazon.com Inc.	2.834845e-02
## 4	BRK.B	Berkshire Hathaway Inc. Class B	1.714493e-02
## 5	FB	Facebook Inc. Class A	1.676060e-02
## 6	JNJ	Johnson & Johnson	1.570168e-02
## 7	JPM	JPMorgan Chase & Co.	1.507235e-02
## 8	GOOG	Alphabet Inc. Class C	1.470747e-02
## 9	GOOGL	Alphabet Inc. Class A	1.436854e-02
## 10	XOM	Exxon Mobil Corporation	1.412361e-02
## 11	BAC	Bank of America Corp	1.141184e-02
## 12	UNH	UnitedHealth Group Incorporated	1.119696e-02

## 13	V	Visa Inc. Class A	1.093462e-02
## 14	PG	Procter & Gamble Company	1.053192e-02
## 15	PFE	Pfizer Inc.	1.052585e-02
## 16	INTC	Intel Corporation	1.012464e-02
## 17	CVX	Chevron Corporation	9.790428e-03
## 18	VZ	Verizon Communications Inc.	9.784105e-03
## 19	CSCO	Cisco Systems Inc.	9.544976e-03
## 20	T	AT&T Inc.	9.518933e-03
## 21	BA	Boeing Company	9.479343e-03
## 22	HD	Home Depot Inc.	9.332370e-03
## 23	WFC	Wells Fargo & Company	8.957703e-03
## 24	MRK	Merck & Co. Inc.	8.915412e-03
## 25	MA	Mastercard Incorporated Class A	8.635942e-03
## 26	KO	Coca-Cola Company	7.439485e-03
## 27	CMCSA	Comcast Corporation Class A	7.368779e-03
## 28	DIS	Walt Disney Company	7.194712e-03
## 29	PEP	PepsiCo Inc.	7.023937e-03
## 30	C	Citigroup Inc.	6.744127e-03
## 31	NFLX	Netflix Inc.	6.680205e-03
## 32	WMT	Walmart Inc.	6.127690e-03
## 33	MCD	McDonald's Corporation	5.961227e-03
## 34	ABT	Abbott Laboratories	5.673213e-03
## 35	ORCL	Oracle Corporation	5.649621e-03
## 36	PM	Philip Morris International Inc.	5.569381e-03
## 37	ADBE	Adobe Inc.	5.423148e-03
## 38	IBM	International Business Machines Corporation	5.378546e-03
## 39	UNP	Union Pacific Corporation	5.373873e-03
## 40	DWDP	DowDuPont Inc.	5.343558e-03
## 41	MDT	Medtronic plc	5.320276e-03
## 42	ABBV	AbbVie Inc.	5.227510e-03
## 43	MMM	3M Company	5.225519e-03
## 44	CRM	salesforce.com inc.	5.221497e-03
## 45	AMGN	Amgen Inc.	5.154364e-03
## 46	AVGO	Broadcom Inc.	5.010262e-03
## 47	LLY	Eli Lilly and Company	4.959256e-03
## 48	HON	Honeywell International Inc.	4.860377e-03
## 49	PYPL	PayPal Holdings Inc	4.796215e-03
## 50	NKE	NIKE Inc. Class B	4.662578e-03
## 51	UTX	United Technologies Corporation	4.450672e-03
## 52	TXN	Texas Instruments Incorporated	4.437206e-03
## 53	TMO	Thermo Fisher Scientific Inc.	4.363659e-03
## 54	ACN	Accenture Plc Class A	4.355615e-03
## 55	NVDA	NVIDIA Corporation	4.119677e-03
## 56	COST	Costco Wholesale Corporation	4.072393e-03
## 57	LIN	Linde plc	3.986931e-03
## 58	MO	Altria Group Inc	3.917806e-03
## 59	CVS	CVS Health Corporation	3.853464e-03
## 60	BKNG	Booking Holdings Inc.	3.844550e-03
## 61	NEE	NextEra Energy Inc.	3.774455e-03
## 62	GE	General Electric Company	3.767361e-03
## 63	SBUX	Starbucks Corporation	3.765740e-03
## 64	GILD	Gilead Sciences Inc.	3.740188e-03
## 65	BMJ	Bristol-Myers Squibb Company	3.620328e-03
## 66	LOW	Lowe's Companies Inc.	3.611353e-03

## 67	COP	ConocoPhillips	3.482849e-03
## 68	ANTM	Anthem Inc.	3.459557e-03
## 69	CAT	Caterpillar Inc.	3.452684e-03
## 70	AMT	American Tower Corporation	3.362129e-03
## 71	USB	U.S. Bancorp	3.355776e-03
## 72	UPS	United Parcel Service Inc. Class B	3.308592e-03
## 73	LMT	Lockheed Martin Corporation	3.248902e-03
## 74	AXP	American Express Company	3.238637e-03
## 75	CI	Cigna Corporation	3.172444e-03
## 76	MDLZ	Mondelez International Inc. Class A	3.018507e-03
## 77	GS	Goldman Sachs Group Inc.	2.949052e-03
## 78	DHR	Danaher Corporation	2.945331e-03
## 79	BIIB	Biogen Inc.	2.878567e-03
## 80	BDX	Becton Dickinson and Company	2.859908e-03
## 81	ADP	Automatic Data Processing Inc.	2.800078e-03
## 82	CELG	Celgene Corporation	2.724370e-03
## 83	QCOM	QUALCOMM Incorporated	2.704791e-03
## 84	CME	CME Group Inc. Class A	2.693005e-03
## 85	TJX	TJX Companies Inc	2.688162e-03
## 86	ISRG	Intuitive Surgical Inc.	2.687962e-03
## 87	DUK	Duke Energy Corporation	2.677987e-03
## 88	CHTR	Charter Communications Inc. Class A	2.661359e-03
## 89	SLB	Schlumberger NV	2.656396e-03
## 90	CB	Chubb Limited	2.635526e-03
## 91	INTU	Intuit Inc.	2.599428e-03
## 92	WBA	Walgreens Boots Alliance Inc	2.537987e-03
## 93	CSX	CSX Corporation	2.499398e-03
## 94	EOG	EOG Resources Inc.	2.479198e-03
## 95	SYK	Stryker Corporation	2.475356e-03
## 96	PNC	PNC Financial Services Group Inc.	2.461979e-03
## 97	CL	Colgate-Palmolive Company	2.459598e-03
## 98	SPG	Simon Property Group Inc.	2.420539e-03
## 99	SCHW	Charles Schwab Corporation	2.380268e-03
## 100	BSX	Boston Scientific Corporation	2.362500e-03
## 101	D	Dominion Energy Inc	2.362490e-03
## 102	MS	Morgan Stanley	2.357517e-03
## 103	FOXA	Twenty-First Century Fox Inc. Class A	2.286632e-03
## 104	BLK	BlackRock Inc.	2.263240e-03
## 105	RTN	Raytheon Company	2.245281e-03
## 106	OXY	Occidental Petroleum Corporation	2.194796e-03
## 107	GM	General Motors Company	2.190574e-03
## 108	DE	Deere & Company	2.180539e-03
## 109	SO	Southern Company	2.143840e-03
## 110	NSC	Norfolk Southern Corporation	2.139458e-03
## 111	NOC	Northrop Grumman Corporation	2.133255e-03
## 112	SPGI	S&P Global Inc.	2.128033e-03
## 113	CCI	Crown Castle International Corp	2.127052e-03
## 114	GD	General Dynamics Corporation	2.084551e-03
## 115	BK	Bank of New York Mellon Corporation	2.081129e-03
## 116	VRTX	Vertex Pharmaceuticals Incorporated	2.052085e-03
## 117	MU	Micron Technology Inc.	2.021239e-03
## 118	EXC	Exelon Corporation	1.989433e-03
## 119	MMC	Marsh & McLennan Companies Inc.	1.961229e-03
## 120	ZTS	Zoetis Inc. Class A	1.957868e-03

## 121	MPC	Marathon Petroleum Corporation	1.930374e-03
## 122	PLD	Prologis Inc.	1.912765e-03
## 123	MET	MetLife Inc.	1.906152e-03
## 124	ITW	Illinois Tool Works Inc.	1.901359e-03
## 125	AGN	Allergan plc	1.894716e-03
## 126	ILMN	Illumina Inc.	1.881459e-03
## 127	FDX	FedEx Corporation	1.858768e-03
## 128	ICE	Intercontinental Exchange Inc.	1.849193e-03
## 129	EMR	Emerson Electric Co.	1.828073e-03
## 130	HUM	Humana Inc.	1.825011e-03
## 131	CTSH	Cognizant Technology Solutions Corporation Class A	1.818338e-03
## 132	AON	Aon plc	1.779839e-03
## 133	ECL	Ecolab Inc.	1.774246e-03
## 134	PGR	Progressive Corporation	1.773456e-03
## 135	KMB	Kimberly-Clark Corporation	1.754196e-03
## 136	PSX	Phillips 66	1.751575e-03
## 137	PRU	Prudential Financial Inc.	1.680439e-03
## 138	ADI	Analog Devices Inc.	1.677838e-03
## 139	BBT	BB&T Corporation	1.672945e-03
## 140	AEP	American Electric Power Company Inc.	1.667633e-03
## 141	COF	Capital One Financial Corporation	1.667192e-03
## 142	WM	Waste Management Inc.	1.664881e-03
## 143	HCA	HCA Healthcare Inc	1.662640e-03
## 144	AMAT	Applied Materials Inc.	1.640689e-03
## 145	TGT	Target Corporation	1.633986e-03
## 146	APD	Air Products and Chemicals Inc.	1.629554e-03
## 147	AFL	Aflac Incorporated	1.614086e-03
## 148	AIG	American International Group Inc.	1.602040e-03
## 149	EW	Edwards Lifesciences Corporation	1.591745e-03
## 150	HPQ	HP Inc.	1.572385e-03
## 151	BAX	Baxter International Inc.	1.557538e-03
## 152	SHW	Sherwin-Williams Company	1.553596e-03
## 153	VLO	Valero Energy Corporation	1.538518e-03
## 154	FIS	Fidelity National Information Services Inc.	1.527503e-03
## 155	KMI	Kinder Morgan Inc Class P	1.522860e-03
## 156	ROST	Ross Stores Inc.	1.512285e-03
## 157	ADSK	Autodesk Inc.	1.491685e-03
## 158	MAR	Marriott International Inc. Class A	1.489754e-03
## 159	EL	Estee Lauder Companies Inc. Class A	1.484361e-03
## 160	FISV	Fiserv Inc.	1.480919e-03
## 161	TRV	Travelers Companies Inc.	1.469054e-03
## 162	ETN	Eaton Corp. Plc	1.454436e-03
## 163	EQIX	Equinix Inc.	1.453206e-03
## 164	ATVI	Activision Blizzard Inc.	1.451855e-03
## 165	F	Ford Motor Company	1.427123e-03
## 166	EBAY	eBay Inc.	1.422420e-03
## 167	WMB	Williams Companies Inc.	1.408363e-03
## 168	REGN	Regeneron Pharmaceuticals Inc.	1.402270e-03
## 169	EA	Electronic Arts Inc.	1.393026e-03
## 170	ALL	Allstate Corporation	1.392405e-03
## 171	JCI	Johnson Controls International plc	1.385652e-03
## 172	SYN	Sysco Corporation	1.378898e-03
## 173	ROP	Roper Technologies Inc.	1.376557e-03
## 174	DAL	Delta Air Lines Inc.	1.372395e-03

## 175	RHT	Red Hat Inc.	1.369004e-03
## 176	DG	Dollar General Corporation	1.350724e-03
## 177	SRE	Sempra Energy	1.345972e-03
## 178	ORLY	O'Reilly Automotive Inc.	1.332325e-03
## 179	PSA	Public Storage	1.328894e-03
## 180	XLNX	Xilinx Inc.	1.285252e-03
## 181	STI	SunTrust Banks Inc.	1.279929e-03
## 182	YUM	Yum! Brands Inc.	1.273546e-03
## 183	LUV	Southwest Airlines Co.	1.271535e-03
## 184	KHC	Kraft Heinz Company	1.264782e-03
## 185	WELL	Welltower Inc.	1.251195e-03
## 186	STZ	Constellation Brands Inc. Class A	1.248293e-03
## 187	ALXN	Alexion Pharmaceuticals Inc.	1.231145e-03
## 188	TEL	TE Connectivity Ltd.	1.220970e-03
## 189	LRCX	Lam Research Corporation	1.219289e-03
## 190	PEG	Public Service Enterprise Group Inc	1.214677e-03
## 191	VFC	V.F. Corporation	1.208463e-03
## 192	MCO	Moody's Corporation	1.206382e-03
## 193	HAL	Halliburton Company	1.196988e-03
## 194	LYB	LyondellBasell Industries NV	1.196237e-03
## 195	GLW	Corning Inc	1.194216e-03
## 196	OKE	ONEOK Inc.	1.188553e-03
## 197	APH	Amphenol Corporation Class A	1.185542e-03
## 198	XEL	Xcel Energy Inc.	1.179399e-03
## 199	AVB	AvalonBay Communities Inc.	1.156808e-03
## 200	MCK	McKesson Corporation	1.156177e-03
## 201	EQR	Equity Residential	1.150735e-03
## 202	GIS	General Mills Inc.	1.145642e-03
## 203	STT	State Street Corporation	1.145102e-03
## 204	CNC	Centene Corporation	1.138238e-03
## 205	PPG	PPG Industries Inc.	1.127853e-03
## 206	IR	Ingersoll-Rand Plc	1.096718e-03
## 207	ZBH	Zimmer Biomet Holdings Inc.	1.073786e-03
## 208	A	Agilent Technologies Inc.	1.061240e-03
## 209	MTB	M&T Bank Corporation	1.058008e-03
## 210	PXD	Pioneer Natural Resources Company	1.051955e-03
## 211	CXO	Concho Resources Inc.	1.051925e-03
## 212	DFS	Discover Financial Services	1.043961e-03
## 213	ED	Consolidated Edison Inc.	1.042370e-03
## 214	HLT	Hilton Worldwide Holdings Inc	1.041860e-03
## 215	FOX	Twenty-First Century Fox Inc. Class B	1.041340e-03
## 216	PAYX	Paychex Inc.	1.035437e-03
## 217	FTV	Fortive Corp.	1.033396e-03
## 218	PCAR	PACCAR Inc	1.017388e-03
## 219	DLR	Digital Realty Trust Inc.	1.011595e-03
## 220	TROW	T. Rowe Price Group	1.010384e-03
## 221	ADM	Archer-Daniels-Midland Company	1.006973e-03
## 222	KR	Kroger Co.	1.001020e-03
## 223	CMI	Cummins Inc.	9.979983e-04
## 224	DLTR	Dollar Tree Inc.	9.979783e-04
## 225	AZO	AutoZone Inc.	9.977982e-04
## 226	WEC	WEC Energy Group Inc	9.976581e-04
## 227	MNST	Monster Beverage Corporation	9.970779e-04
## 228	APC	Anadarko Petroleum Corporation	9.903545e-04

## 229	CCL.U	Carnival Corporation	9.873830e-04
## 230	VTR	Ventas Inc.	9.864626e-04
## 231	HPE	Hewlett Packard Enterprise Co.	9.839113e-04
## 232	PH	Parker-Hannifin Corporation	9.801494e-04
## 233	IQV	IQVIA Holdings Inc	9.759873e-04
## 234	MSI	Motorola Solutions Inc.	9.730559e-04
## 235	TWTR	Twitter Inc.	9.606297e-04
## 236	WLTW	Willis Towers Watson Public Limited Company	9.449018e-04
## 237	ES	Eversource Energy	9.445016e-04
## 238	PPL	PPL Corporation	9.408298e-04
## 239	ROK	Rockwell Automation Inc.	9.317452e-04
## 240	DTE	DTE Energy Company	9.294641e-04
## 241	SYF	Synchrony Financial	9.250619e-04
## 242	APTIV	Aptiv PLC	9.162075e-04
## 243	MCHP	Microchip Technology Incorporated	9.087137e-04
## 244	TDG	TransDigm Group Incorporated	9.086637e-04
## 245	SBAC	SBA Communications Corp. Class A	9.083636e-04
## 246	SWK	Stanley Black & Decker Inc.	8.942865e-04
## 247	AMD	Advanced Micro Devices Inc.	8.933361e-04
## 248	NTRS	Northern Trust Corporation	8.885036e-04
## 249	O	Realty Income Corporation	8.874931e-04
## 250	VRSK	Verisk Analytics Inc	8.821104e-04
## 251	BXP	Boston Properties Inc.	8.754071e-04
## 252	UAL	United Continental Holdings Inc.	8.631610e-04
## 253	RCL	Royal Caribbean Cruises Ltd.	8.626407e-04
## 254	CLX	Clorox Company	8.558573e-04
## 255	FLT	FleetCor Technologies Inc.	8.436713e-04
## 256	EIX	Edison International	8.369679e-04
## 257	GPN	Global Payments Inc.	8.353671e-04
## 258	HRS	Harris Corporation	8.295442e-04
## 259	IP	International Paper Company	8.223906e-04
## 260	FE	FirstEnergy Corp.	8.221105e-04
## 261	WY	Weyerhaeuser Company	8.213501e-04
## 262	INFO	IHS Markit Ltd.	8.174982e-04
## 263	CERN	Cerner Corporation	8.136563e-04
## 264	VRSN	VeriSign Inc.	8.048018e-04
## 265	NUE	Nucor Corporation	8.041515e-04
## 266	BLL	Ball Corporation	7.976183e-04
## 267	ALGN	Align Technology Inc.	7.962776e-04
## 268	TSN	Tyson Foods Inc. Class A	7.915152e-04
## 269	AMP	Ameriprise Financial Inc.	7.904947e-04
## 270	KEY	KeyCorp	7.861025e-04
## 271	DXC	DXC Technology Co.	7.849519e-04
## 272	ESS	Essex Property Trust Inc.	7.820805e-04
## 273	IDXX	IDEXX Laboratories Inc.	7.770380e-04
## 274	AME	AMETEK Inc.	7.766478e-04
## 275	FITB	Fifth Third Bancorp	7.736063e-04
## 276	WAT	Waters Corporation	7.714552e-04
## 277	FAST	Fastenal Company	7.690040e-04
## 278	FCX	Freeport-McMoRan Inc.	7.602396e-04
## 279	AWK	American Water Works Company Inc.	7.555272e-04
## 280	NEM	Newmont Mining Corporation	7.529760e-04
## 281	CTAS	Cintas Corporation	7.481436e-04
## 282	CFG	Citizens Financial Group Inc.	7.464627e-04

## 283	ULTA	Ulta Beauty Inc	7.426108e-04
## 284	HIG	Hartford Financial Services Group Inc.	7.408799e-04
## 285	FRC	First Republic Bank	7.325258e-04
## 286	CBS	CBS Corporation Class B	7.311351e-04
## 287	RSG	Republic Services Inc.	7.226008e-04
## 288	KLAC	KLA-Tencor Corporation	7.217504e-04
## 289	AEE	Ameren Corporation	7.206898e-04
## 290	MTD	Mettler-Toledo International Inc.	7.184887e-04
## 291	OMC	Omnicom Group Inc	7.175283e-04
## 292	CAH	Cardinal Health Inc.	7.103347e-04
## 293	RF	Regions Financial Corporation	7.101146e-04
## 294	NTAP	NetApp Inc.	7.097444e-04
## 295	LLL	L3 Technologies Inc	7.066829e-04
## 296	MYL	Mylan N.V.	7.006999e-04
## 297	ABMD	ABIOMED Inc.	6.935163e-04
## 298	FANG	Diamondback Energy Inc.	6.920756e-04
## 299	ETR	Entergy Corporation	6.896343e-04
## 300	EVRG	Evergy Inc.	6.780386e-04
## 301	CBRE	CBRE Group Inc. Class A	6.753972e-04
## 302	MXIM	Maxim Integrated Products Inc.	6.719555e-04
## 303	CHD	Church & Dwight Co. Inc.	6.718755e-04
## 304	GPC	Genuine Parts Company	6.684037e-04
## 305	MKC	McCormick & Company Incorporated	6.650521e-04
## 306	TSS	Total System Services Inc.	6.647519e-04
## 307	MSCI	MSCI Inc. Class A	6.631811e-04
## 308	LH	Laboratory Corporation of America Holdings	6.526559e-04
## 309	HSY	Hershey Company	6.524658e-04
## 310	CNP	CenterPoint Energy Inc.	6.480336e-04
## 311	EXPE	Expedia Group Inc.	6.438815e-04
## 312	KEYS	Keysight Technologies Inc	6.404798e-04
## 313	HBAN	Huntington Bancshares Incorporated	6.392192e-04
## 314	SNPS	Synopsys Inc.	6.389991e-04
## 315	CMS	CMS Energy Corporation	6.380286e-04
## 316	CMG	Chipotle Mexican Grill Inc.	6.367779e-04
## 317	SWKS	Skyworks Solutions Inc.	6.361676e-04
## 318	VMC	Vulcan Materials Company	6.352072e-04
## 319	SYMC	Symantec Corporation	6.343768e-04
## 320	HCP	HCP Inc.	6.318055e-04
## 321	CDNS	Cadence Design Systems Inc.	6.312652e-04
## 322	INCY	Incyte Corporation	6.284338e-04
## 323	ANSS	ANSYS Inc.	6.232512e-04
## 324	MRO	Marathon Oil Corporation	6.196694e-04
## 325	HES	Hess Corporation	6.190591e-04
## 326	AJG	Arthur J. Gallagher & Co.	6.175784e-04
## 327	MGM	MGM Resorts International	6.163778e-04
## 328	GWW	W.W. Grainger Inc.	6.140466e-04
## 329	K	Kellogg Company	6.137565e-04
## 330	AAL	American Airlines Group Inc.	6.124658e-04
## 331	LEN	Lennar Corporation Class A	6.092742e-04
## 332	ARE	Alexandria Real Estate Equities Inc.	6.082837e-04
## 333	RMD	ResMed Inc.	6.071331e-04
## 334	BBY	Best Buy Co. Inc.	6.056024e-04
## 335	CMA	Comerica Incorporated	5.996494e-04
## 336	WCG	WellCare Health Plans Inc.	5.953673e-04

## 337	DRI	Darden Restaurants Inc.	5.935464e-04
## 338	ABC	AmerisourceBergen Corporation	5.918655e-04
## 339	WDC	Western Digital Corporation	5.916054e-04
## 340	HST	Host Hotels & Resorts Inc.	5.866429e-04
## 341	DHI	D.R. Horton Inc.	5.829311e-04
## 342	CTXS	Citrix Systems Inc.	5.824008e-04
## 343	TXT	Textron Inc.	5.822107e-04
## 344	ANET	Arista Networks Inc.	5.821006e-04
## 345	COO	Cooper Companies Inc.	5.809801e-04
## 346	DOV	Dover Corporation	5.751772e-04
## 347	PFG	Principal Financial Group Inc.	5.747170e-04
## 348	TFX	Teleflex Incorporated	5.732262e-04
## 349	DVN	Devon Energy Corporation	5.712452e-04
## 350	IFF	International Flavors & Fragrances Inc.	5.703548e-04
## 351	LNC	Lincoln National Corporation	5.692042e-04
## 352	XYL	Xylem Inc.	5.646419e-04
## 353	BHGE	Baker Hughes a GE Company Class A	5.601297e-04
## 354	EFX	Equifax Inc.	5.579286e-04
## 355	CE	Celanese Corporation	5.579186e-04
## 356	CTL	CenturyLink Inc.	5.574683e-04
## 357	L	Loews Corporation	5.549771e-04
## 358	SIVB	SVB Financial Group	5.546369e-04
## 359	IT	Gartner Inc.	5.531062e-04
## 360	CINF	Cincinnati Financial Corporation	5.520556e-04
## 361	EXPD	Expeditors International of Washington Inc.	5.477835e-04
## 362	CHRW	C.H. Robinson Worldwide Inc.	5.420206e-04
## 363	APA	Apache Corporation	5.405199e-04
## 364	EXR	Extra Space Storage Inc.	5.384689e-04
## 365	HOLX	Hologic Inc.	5.374083e-04
## 366	AAP	Advance Auto Parts Inc.	5.350071e-04
## 367	NRG	NRG Energy Inc.	5.292443e-04
## 368	UDR	UDR Inc.	5.268831e-04
## 369	ETFC	E*TRADE Financial Corporation	5.260527e-04
## 370	VAR	Varian Medical Systems Inc.	5.234614e-04
## 371	WYNN	Wynn Resorts Limited	5.205199e-04
## 372	CAG	Conagra Brands Inc.	5.101947e-04
## 373	FTNT	Fortinet Inc.	5.097445e-04
## 374	STX	Seagate Technology PLC	5.096245e-04
## 375	DGX	Quest Diagnostics Incorporated	5.065029e-04
## 376	MLM	Martin Marietta Materials Inc.	5.062028e-04
## 377	TSCO	Tractor Supply Company	5.057225e-04
## 378	VNO	Vornado Realty Trust	5.054024e-04
## 379	MAA	Mid-America Apartment Communities Inc.	5.031012e-04
## 380	HRL	Hormel Foods Corporation	5.012603e-04
## 381	BR	Broadridge Financial Solutions Inc.	4.989791e-04
## 382	SJM	J.M. Smucker Company	4.983588e-04
## 383	EMN	Eastman Chemical Company	4.979786e-04
## 384	UHS	Universal Health Services Inc. Class B	4.975784e-04
## 385	NCLH	Norwegian Cruise Line Holdings Ltd.	4.973683e-04
## 386	MAS	Masco Corporation	4.912953e-04
## 387	AKAM	Akamai Technologies Inc.	4.905549e-04
## 388	ATO	Atmos Energy Corporation	4.900247e-04
## 389	TAP	Molson Coors Brewing Company Class B	4.867230e-04
## 390	MOS	Mosaic Company	4.859326e-04

## 391	FMC	FMC Corporation	4.848721e-04
## 392	NOV	National Oilwell Varco Inc.	4.833113e-04
## 393	REG	Regency Centers Corporation	4.802398e-04
## 394	URI	United Rentals Inc.	4.765079e-04
## 395	COG	Cabot Oil & Gas Corporation	4.752273e-04
## 396	AES	AES Corporation	4.750672e-04
## 397	KSS	Kohl's Corporation	4.714254e-04
## 398	KMX	CarMax Inc.	4.696045e-04
## 399	KSU	Kansas City Southern	4.682138e-04
## 400	CPRT	Copart Inc.	4.660327e-04
## 401	NBL	Noble Energy Inc.	4.633514e-04
## 402	RJF	Raymond James Financial Inc.	4.619907e-04
## 403	TTWO	Take-Two Interactive Software Inc.	4.529462e-04
## 404	DRE	Duke Realty Corporation	4.512753e-04
## 405	LNT	Alliant Energy Corp	4.485540e-04
## 406	CBOE	Cboe Global Markets Inc	4.448721e-04
## 407	FFIV	F5 Networks Inc.	4.446620e-04
## 408	JKHY	Jack Henry & Associates Inc.	4.428011e-04
## 409	LW	Lamb Weston Holdings Inc.	4.403699e-04
## 410	PKI	PerkinElmer Inc.	4.390992e-04
## 411	TPR	Tapestry Inc.	4.385890e-04
## 412	VIAB	Viacom Inc. Class B	4.367881e-04
## 413	FTI	TechnipFMC Plc	4.341868e-04
## 414	HAS	Hasbro Inc.	4.310552e-04
## 415	NDAQ	Nasdaq Inc.	4.308651e-04
## 416	IRM	Iron Mountain Inc.	4.289042e-04
## 417	PNW	Pinnacle West Capital Corporation	4.272633e-04
## 418	WRK	WestRock Company	4.253424e-04
## 419	DISCK	Discovery Inc. Class C	4.248021e-04
## 420	BEN	Franklin Resources Inc.	4.240517e-04
## 421	TIF	Tiffany & Co.	4.230212e-04
## 422	FRT	Federal Realty Investment Trust	4.207801e-04
## 423	NI	NiSource Inc	4.200297e-04
## 424	JBHT	J.B. Hunt Transport Services Inc.	4.181888e-04
## 425	ZION	Zions Bancorporation N.A.	4.156776e-04
## 426	CF	CF Industries Holdings Inc.	4.141368e-04
## 427	XRAY	DENTSPLY SIRONA Inc.	4.130963e-04
## 428	HSIC	Henry Schein Inc.	4.068632e-04
## 429	HII	Huntington Ingalls Industries Inc.	4.054825e-04
## 430	JNPR	Juniper Networks Inc.	4.021908e-04
## 431	HFC	HollyFrontier Corporation	4.006400e-04
## 432	NLSN	Nielsen Holdings Plc	4.002899e-04
## 433	PKG	Packaging Corporation of America	3.944870e-04
## 434	AVY	Avery Dennison Corporation	3.883139e-04
## 435	IPG	Interpublic Group of Companies Inc.	3.825310e-04
## 436	SNA	Snap-on Incorporated	3.805400e-04
## 437	WHR	Whirlpool Corporation	3.793094e-04
## 438	RE	Everest Re Group Ltd.	3.785190e-04
## 439	ALB	Albemarle Corporation	3.767281e-04
## 440	MHK	Mohawk Industries Inc.	3.750773e-04
## 441	BWA	BorgWarner Inc.	3.740268e-04
## 442	ALLE	Allegion PLC	3.714054e-04
## 443	GRMN	Garmin Ltd.	3.696846e-04
## 444	PVH	PVH Corp.	3.696646e-04

## 445	TMK	Torchmark Corporation	3.691343e-04
## 446	LKQ	LKQ Corporation	3.688242e-04
## 447	ADS	Alliance Data Systems Corporation	3.505650e-04
## 448	BF.B	Brown-Forman Corporation Class B	3.505650e-04
## 449	RHI	Robert Half International Inc.	3.495245e-04
## 450	JEC	Jacobs Engineering Group Inc.	3.482939e-04
## 451	ALK	Alaska Air Group Inc.	3.471433e-04
## 452	QRVO	Qorvo Inc.	3.466931e-04
## 453	WU	Western Union Company	3.396196e-04
## 454	UNM	Unum Group	3.383189e-04
## 455	SLG	SL Green Realty Corp.	3.379187e-04
## 456	AIV	Apartment Investment and Management Company Class A	3.317456e-04
## 457	IVZ	Invesco Ltd.	3.289042e-04
## 458	M	Macy's Inc	3.234915e-04
## 459	ARNC	Arconic Inc.	3.221308e-04
## 460	KIM	Kimco Realty Corporation	3.205200e-04
## 461	DVA	DaVita Inc.	3.198897e-04
## 462	NWL	Newell Brands Inc	3.183990e-04
## 463	AOS	A. O. Smith Corporation	3.176186e-04
## 464	XEC	Cimarex Energy Co.	3.077136e-04
## 465	NKTR	Nektar Therapeutics	3.070933e-04
## 466	FLIR	FLIR Systems Inc.	3.052424e-04
## 467	PHM	PulteGroup Inc.	3.023210e-04
## 468	RL	Ralph Lauren Corporation Class A	3.008902e-04
## 469	DISH	DISH Network Corporation Class A	3.003400e-04
## 470	PNR	Pentair plc	2.993395e-04
## 471	FL	Foot Locker Inc.	2.956276e-04
## 472	HBI	Hanesbrands Inc.	2.951674e-04
## 473	XRX	Xerox Corporation	2.891744e-04
## 474	FBHS	Fortune Brands Home & Security Inc.	2.876636e-04
## 475	SEE	Sealed Air Corporation	2.856726e-04
## 476	CPRI	Capri Holdings Limited	2.830813e-04
## 477	CPB	Campbell Soup Company	2.824010e-04
## 478	PBCT	People's United Financial Inc.	2.756476e-04
## 479	LB	L Brands Inc.	2.663130e-04
## 480	FLS	Flowserve Corporation	2.653125e-04
## 481	HP	Helmerich & Payne Inc.	2.652424e-04
## 482	HOG	Harley-Davidson Inc.	2.649123e-04
## 483	PRGO	Perrigo Co. Plc	2.626111e-04
## 484	JEF	Jefferies Financial Group Inc.	2.601699e-04
## 485	ROL	Rollins Inc.	2.528763e-04
## 486	LEG	Leggett & Platt Incorporated	2.523660e-04
## 487	AMG	Affiliated Managers Group Inc.	2.485041e-04
## 488	TRIP	TripAdvisor Inc.	2.479238e-04
## 489	IPGP	IPG Photonics Corporation	2.379788e-04
## 490	GPS	Gap Inc.	2.343470e-04
## 491	PWR	Quanta Services Inc.	2.245821e-04
## 492	AIZ	Assurant Inc.	2.211204e-04
## 493	JWN	Nordstrom Inc.	2.172385e-04
## 494	FLR	Fluor Corporation	2.166282e-04
## 495	BHF	Brighthouse Financial Inc.	2.135266e-04
## 496	COTY	Coty Inc. Class A	2.120159e-04
## 497	HRB	H&R Block Inc.	2.095646e-04
## 498	NWSA	News Corporation Class A	2.074036e-04

## 499	MAT	Mattel Inc.	2.017908e-04
## 500	MAC	Macerich Company	1.996897e-04
## 501	DISCA	Discovery Inc. Class A	1.940669e-04
## 502	GT	Goodyear Tire & Rubber Company	1.899448e-04
## 503	UAA	Under Armour Inc. Class A	1.670134e-04
## 504	UA	Under Armour Inc. Class C	1.521260e-04
## 505	NWS	News Corporation Class B	6.290141e-05
## 506	ECA-CA	Encana Corporation	7.893942e-06
##		sector	shares_held
## 1	Information Technology		84853600
## 2	Information Technology		49533308
## 3	Consumer Discretionary		4510051
## 4	Financials		21364490
## 5	Communication Services		26385216
## 6	Health Care		29452358
## 7	Financials		36529800
## 8	Communication Services		3378423
## 9	Communication Services		3282939
## 10	Energy		46493644
## 11	Financials		100285460
## 12	Health Care		10564454
## 13	Information Technology		19303236
## 14	Consumer Staples		27357982
## 15	Health Care		63506110
## 16	Information Technology		50135996
## 17	Energy		20984748
## 18	Communication Services		45375500
## 19	Information Technology		49397930
## 20	Communication Services		79917256
## 21	Industrials		5801735
## 22	Consumer Discretionary		12408924
## 23	Financials		46556430
## 24	Health Care		28576474
## 25	Information Technology		9946411
## 26	Consumer Staples		42067310
## 27	Communication Services		49908344
## 28	Communication Services		16347008
## 29	Consumer Staples		15501865
## 30	Financials		26843670
## 31	Communication Services		4788556
## 32	Consumer Staples		15677056
## 33	Consumer Discretionary		8473443
## 34	Health Care		19286212
## 35	Information Technology		27943780
## 36	Consumer Staples		17070768
## 37	Information Technology		5347148
## 38	Information Technology		9968187
## 39	Industrials		8075142
## 40	Materials		25202004
## 41	Health Care		14750220
## 42	Health Care		16540164
## 43	Industrials		6400274
## 44	Information Technology		8398198
## 45	Health Care		7000953

## 46	Information Technology	4546147
## 47	Health Care	10358008
## 48	Industrials	8109534
## 49	Information Technology	12927395
## 50	Consumer Discretionary	13969941
## 51	Industrials	8911586
## 52	Information Technology	10552200
## 53	Health Care	4421107
## 54	Information Technology	6997603
## 55	Information Technology	6698059
## 56	Consumer Staples	4812566
## 57	Materials	6051462
## 58	Consumer Staples	20562772
## 59	Health Care	14177615
## 60	Consumer Discretionary	508639
## 61	Utilities	5246468
## 62	Industrials	95514760
## 63	Consumer Discretionary	13623676
## 64	Health Care	14155833
## 65	Health Care	17923948
## 66	Consumer Discretionary	8862585
## 67	Energy	12695394
## 68	Health Care	2840374
## 69	Industrials	6484917
## 70	Real Estate	4836797
## 71	Financials	16730737
## 72	Industrials	7634037
## 73	Industrials	2705191
## 74	Financials	7715496
## 75	Health Care	4101458
## 76	Consumer Staples	16033591
## 77	Financials	3800548
## 78	Health Care	6773251
## 79	Health Care	2212211
## 80	Health Care	2945983
## 81	Information Technology	4785852
## 82	Health Care	7684789
## 83	Information Technology	13311328
## 84	Financials	3927643
## 85	Consumer Discretionary	13690399
## 86	Health Care	1253634
## 87	Utilities	7828438
## 88	Communication Services	1950416
## 89	Energy	15117770
## 90	Financials	5061222
## 91	Information Technology	2850040
## 92	Consumer Staples	8841804
## 93	Industrials	8814182
## 94	Energy	6326332
## 95	Health Care	3389350
## 96	Financials	5072966
## 97	Consumer Staples	9464497
## 98	Real Estate	3377012
## 99	Financials	13134280

## 100	Health Care	15105250
## 101	Utilities	8249501
## 102	Financials	14362621
## 103	Communication Services	11606211
## 104	Financials	1340362
## 105	Industrials	3111283
## 106	Energy	8356286
## 107	Consumer Discretionary	14335656
## 108	Industrials	3508474
## 109	Utilities	11293749
## 110	Industrials	2992855
## 111	Industrials	1901582
## 112	Financials	2744956
## 113	Real Estate	4529516
## 114	Industrials	3043009
## 115	Financials	10046857
## 116	Health Care	2791257
## 117	Information Technology	12313897
## 118	Utilities	10545466
## 119	Financials	5503610
## 120	Health Care	5251650
## 121	Energy	7579710
## 122	Real Estate	6873344
## 123	Financials	10869825
## 124	Industrials	3372816
## 125	Health Care	3485269
## 126	Health Care	1605308
## 127	Industrials	2651972
## 128	Financials	6265545
## 129	Industrials	6845957
## 130	Health Care	1504560
## 131	Information Technology	6323489
## 132	Financials	2647597
## 133	Materials	2776346
## 134	Financials	6366478
## 135	Consumer Staples	3788826
## 136	Energy	4667958
## 137	Financials	4565907
## 138	Information Technology	4059161
## 139	Financials	8441104
## 140	Utilities	5381653
## 141	Financials	5228546
## 142	Industrials	4299405
## 143	Health Care	2948341
## 144	Information Technology	10734333
## 145	Consumer Discretionary	5738562
## 146	Materials	2394666
## 147	Financials	8378780
## 148	Financials	9683932
## 149	Health Care	2290413
## 150	Information Technology	17278248
## 151	Health Care	5426107
## 152	Materials	904940
## 153	Energy	4663221

## 154	Information Technology	3587225
## 155	Energy	20721726
## 156	Consumer Discretionary	4115142
## 157	Information Technology	2379755
## 158	Consumer Discretionary	3148294
## 159	Consumer Staples	2449646
## 160	Information Technology	4416413
## 161	Financials	2919089
## 162	Industrials	4725856
## 163	Real Estate	882199
## 164	Communication Services	8327456
## 165	Consumer Discretionary	42749332
## 166	Consumer Discretionary	9947385
## 167	Energy	13211548
## 168	Health Care	846096
## 169	Communication Services	3335426
## 170	Financials	3774086
## 171	Industrials	10098861
## 172	Consumer Staples	5222715
## 173	Industrials	1128806
## 174	Industrials	6873085
## 175	Information Technology	1935077
## 176	Consumer Discretionary	2903658
## 177	Utilities	2987326
## 178	Consumer Discretionary	878425
## 179	Real Estate	1636418
## 180	Information Technology	2755723
## 181	Financials	5028027
## 182	Consumer Discretionary	3461448
## 183	Industrials	5637370
## 184	Consumer Staples	6794417
## 185	Real Estate	4124124
## 186	Consumer Staples	1825683
## 187	Health Care	2433510
## 188	Information Technology	3802557
## 189	Information Technology	1722416
## 190	Utilities	5517268
## 191	Consumer Discretionary	3557434
## 192	Financials	1823290
## 193	Energy	9613979
## 194	Materials	3491712
## 195	Information Technology	8852467
## 196	Energy	4490440
## 197	Information Technology	3271969
## 198	Utilities	5643593
## 199	Real Estate	1509296
## 200	Health Care	2181810
## 201	Real Estate	4020945
## 202	Consumer Staples	6506921
## 203	Financials	4145673
## 204	Health Care	4481716
## 205	Materials	2644321
## 206	Industrials	2678846
## 207	Health Care	2221510

## 208	Health Care	3467173
## 209	Financials	1571095
## 210	Energy	1851432
## 211	Energy	2188314
## 212	Financials	3746094
## 213	Utilities	3396411
## 214	Consumer Discretionary	3260214
## 215	Communication Services	5325635
## 216	Information Technology	3495366
## 217	Industrials	3231765
## 218	Industrials	3821761
## 219	Real Estate	2249676
## 220	Financials	2648802
## 221	Consumer Staples	6136183
## 222	Consumer Staples	8680506
## 223	Industrials	1642124
## 224	Consumer Discretionary	2582660
## 225	Consumer Discretionary	277521
## 226	Utilities	3427101
## 227	Consumer Staples	4343785
## 228	Energy	5601322
## 229	Consumer Discretionary	4412013
## 230	Real Estate	3869811
## 231	Information Technology	15643081
## 232	Industrials	1443835
## 233	Health Care	1760100
## 234	Information Technology	1795575
## 235	Communication Services	7868838
## 236	Financials	1425562
## 237	Utilities	3441836
## 238	Utilities	7901413
## 239	Industrials	1343665
## 240	Utilities	1984554
## 241	Financials	7453370
## 242	Consumer Discretionary	2895350
## 243	Information Technology	2551726
## 244	Industrials	528184
## 245	Real Estate	1252681
## 246	Industrials	1677327
## 247	Information Technology	9650679
## 248	Financials	2441622
## 249	Real Estate	3239184
## 250	Industrials	1799774
## 251	Real Estate	1675955
## 252	Industrials	2500381
## 253	Consumer Discretionary	1870616
## 254	Consumer Staples	1400955
## 255	Information Technology	963283
## 256	Utilities	3538982
## 257	Information Technology	1733167
## 258	Industrials	1287679
## 259	Materials	4461923
## 260	Utilities	5312109
## 261	Real Estate	8226775

## 262	Industrials	3914760
## 263	Health Care	3594290
## 264	Information Technology	1173766
## 265	Materials	3457348
## 266	Materials	3769475
## 267	Health Care	798947
## 268	Consumer Staples	3234016
## 269	Financials	1547442
## 270	Financials	11484608
## 271	Information Technology	3067559
## 272	Real Estate	717552
## 273	Health Care	945392
## 274	Industrials	2514572
## 275	Financials	7283773
## 276	Health Care	840888
## 277	Industrials	3124053
## 278	Materials	15837152
## 279	Utilities	1971184
## 280	Materials	5794860
## 281	Industrials	939128
## 282	Financials	5203126
## 283	Consumer Discretionary	620090
## 284	Financials	3889356
## 285	Financials	1843857
## 286	Communication Services	3693436
## 287	Industrials	2381179
## 288	Information Technology	1706564
## 289	Utilities	2645846
## 290	Health Care	274795
## 291	Communication Services	2447402
## 292	Health Care	3275280
## 293	Financials	11353641
## 294	Information Technology	2835577
## 295	Industrials	851775
## 296	Health Care	5629673
## 297	Health Care	490005
## 298	Energy	1678142
## 299	Utilities	1963267
## 300	Utilities	2951350
## 301	Real Estate	3452740
## 302	Information Technology	3077792
## 303	Consumer Staples	2656922
## 304	Consumer Discretionary	1589238
## 305	Consumer Staples	1324279
## 306	Information Technology	1832858
## 307	Financials	968497
## 308	Health Care	1111794
## 309	Consumer Staples	1526384
## 310	Utilities	5366678
## 311	Consumer Discretionary	1296049
## 312	Information Technology	2038376
## 313	Financials	11671771
## 314	Information Technology	1608117
## 315	Utilities	3092984

## 316	Consumer Discretionary	268854
## 317	Information Technology	1951800
## 318	Materials	1436228
## 319	Information Technology	7010013
## 320	Real Estate	5240775
## 321	Information Technology	3071238
## 322	Health Care	1910436
## 323	Information Technology	913731
## 324	Energy	9281057
## 325	Energy	2750316
## 326	Financials	1978764
## 327	Consumer Discretionary	5589458
## 328	Industrials	496123
## 329	Consumer Staples	2764699
## 330	Industrials	4470119
## 331	Consumer Discretionary	3187318
## 332	Real Estate	1153852
## 333	Health Care	1546184
## 334	Consumer Discretionary	2574321
## 335	Financials	1776056
## 336	Health Care	545599
## 337	Consumer Discretionary	1354233
## 338	Health Care	1747146
## 339	Information Technology	3176075
## 340	Real Estate	8051085
## 341	Consumer Discretionary	3730843
## 342	Information Technology	1399057
## 343	Industrials	2716830
## 344	Information Technology	564156
## 345	Health Care	532562
## 346	Industrials	1613347
## 347	Financials	2894667
## 348	Health Care	512924
## 349	Energy	5141867
## 350	Materials	1105247
## 351	Financials	2363409
## 352	Industrials	1953788
## 353	Energy	5547419
## 354	Industrials	1305700
## 355	Materials	1423524
## 356	Communication Services	10379065
## 357	Financials	3040086
## 358	Financials	574548
## 359	Information Technology	987044
## 360	Financials	1652133
## 361	Industrials	1902907
## 362	Industrials	1516193
## 363	Energy	4152349
## 364	Real Estate	1381347
## 365	Health Care	2969890
## 366	Consumer Discretionary	805121
## 367	Utilities	3188619
## 368	Real Estate	3022753
## 369	Financials	2835503

## 370	Health Care	1008799
## 371	Consumer Discretionary	1069451
## 372	Consumer Staples	5327094
## 373	Information Technology	1569001
## 374	Information Technology	2852071
## 375	Health Care	1475566
## 376	Materials	681336
## 377	Consumer Discretionary	1325791
## 378	Real Estate	1879215
## 379	Real Estate	1235370
## 380	Consumer Staples	2968258
## 381	Information Technology	1272004
## 382	Consumer Staples	1233183
## 383	Materials	1540210
## 384	Health Care	939188
## 385	Consumer Discretionary	2399286
## 386	Industrials	3355027
## 387	Information Technology	1790425
## 388	Utilities	1291661
## 389	Consumer Staples	2044159
## 390	Materials	3871360
## 391	Materials	1453268
## 392	Energy	4147562
## 393	Real Estate	1852404
## 394	Industrials	903877
## 395	Energy	4818499
## 396	Utilities	7182509
## 397	Consumer Discretionary	1829183
## 398	Consumer Discretionary	1935735
## 399	Industrials	1113767
## 400	Industrials	2206502
## 401	Energy	5235531
## 402	Financials	1435488
## 403	Communication Services	1241900
## 404	Real Estate	3876499
## 405	Utilities	2551075
## 406	Financials	1216629
## 407	Information Technology	664116
## 408	Information Technology	846668
## 409	Consumer Staples	1620658
## 410	Health Care	1195097
## 411	Consumer Discretionary	3124383
## 412	Communication Services	3833174
## 413	Energy	4718422
## 414	Consumer Discretionary	1276288
## 415	Financials	1255942
## 416	Real Estate	3124109
## 417	Utilities	1214980
## 418	Materials	2786398
## 419	Communication Services	3930185
## 420	Financials	3342977
## 421	Consumer Discretionary	1191398
## 422	Real Estate	794526
## 423	Utilities	3967883

## 424	Industrials	941785
## 425	Financials	2122059
## 426	Materials	2536327
## 427	Health Care	2424846
## 428	Health Care	1668249
## 429	Industrials	472130
## 430	Information Technology	3793694
## 431	Energy	1770424
## 432	Industrials	3893577
## 433	Materials	1024844
## 434	Materials	946952
## 435	Communication Services	4185522
## 436	Industrials	616986
## 437	Consumer Discretionary	696269
## 438	Financials	446368
## 439	Materials	1184371
## 440	Consumer Discretionary	689241
## 441	Consumer Discretionary	2282471
## 442	Industrials	1037700
## 443	Consumer Discretionary	1320464
## 444	Consumer Discretionary	837593
## 445	Financials	1132802
## 446	Consumer Discretionary	3475186
## 447	Information Technology	517747
## 448	Consumer Staples	1832793
## 449	Industrials	1334917
## 450	Industrials	1296545
## 451	Industrials	1339031
## 452	Information Technology	1361089
## 453	Information Technology	4880875
## 454	Financials	2382921
## 455	Real Estate	943705
## 456	Real Estate	1710661
## 457	Financials	4461157
## 458	Consumer Discretionary	3326163
## 459	Industrials	4695504
## 460	Real Estate	4616746
## 461	Health Care	1384655
## 462	Consumer Discretionary	4746547
## 463	Industrials	1573427
## 464	Energy	1035895
## 465	Health Care	1884332
## 466	Information Technology	1489880
## 467	Consumer Discretionary	2851670
## 468	Consumer Discretionary	615328
## 469	Communication Services	2486454
## 470	Industrials	1792505
## 471	Consumer Discretionary	1275513
## 472	Consumer Discretionary	4022810
## 473	Information Technology	2428599
## 474	Industrials	1552478
## 475	Materials	1733776
## 476	Consumer Discretionary	1627708
## 477	Consumer Staples	2076504

```
## 478 Financials 4066548
## 479 Consumer Discretionary 2491820
## 480 Industrials 1420509
## 481 Energy 1182082
## 482 Consumer Discretionary 1816851
## 483 Health Care 1372418
## 484 Financials 3166258
## 485 Industrials 1612804
## 486 Consumer Discretionary 1436217
## 487 Financials 583435
## 488 Communication Services 1116383
## 489 Information Technology 393061
## 490 Consumer Discretionary 2391279
## 491 Industrials 1606564
## 492 Financials 575956
## 493 Consumer Discretionary 1250523
## 494 Industrials 1526223
## 495 Financials 1302087
## 496 Consumer Staples 4912631
## 497 Consumer Discretionary 2224465
## 498 Communication Services 4161396
## 499 Consumer Discretionary 3735261
## 500 Real Estate 1169252
## 501 Communication Services 1694339
## 502 Consumer Discretionary 2585962
## 503 Consumer Discretionary 2019164
## 504 Consumer Discretionary 2043966
## 505 Communication Services 1244468
## 506 Energy 292529
```

```
sp_500_prices_tbl <- readRDS("data01/sp_500_prices_tbl.rds")
sp_500_prices_tbl
```

```
## # A tibble: 1,225,765 x 8
##   symbol date      open high low close volume adjusted
##   <chr> <date>    <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 MSFT 2009-01-02 19.5 20.4 19.4 20.3 50084000 15.9
## 2 MSFT 2009-01-05 20.2 20.7 20.1 20.5 61475200 16.0
## 3 MSFT 2009-01-06 20.8 21 20.6 20.8 58083400 16.2
## 4 MSFT 2009-01-07 20.2 20.3 19.5 19.5 72709900 15.2
## 5 MSFT 2009-01-08 19.6 20.2 19.5 20.1 70255400 15.7
## 6 MSFT 2009-01-09 20.2 20.3 19.4 19.5 49815300 15.2
## 7 MSFT 2009-01-12 19.7 19.8 19.3 19.5 52163500 15.2
## 8 MSFT 2009-01-13 19.5 20.0 19.5 19.8 65843500 15.5
## 9 MSFT 2009-01-14 19.5 19.7 19.0 19.1 80257500 14.9
## 10 MSFT 2009-01-15 19.1 19.3 18.5 19.2 96169800 15.0
## # i 1,225,755 more rows
```

Question

Which stock prices behave similarly?

Answering this question helps us **understand which companies are related**, and we can use clustering to help us answer it!

Even if we're not interested in finance, this is still a great analysis because it will tell us which companies are competitors and which are likely in the same space (often called sectors) and can be categorized together. Bottom line - This analysis can help us better understand the dynamics of the market and competition, which is useful for all types of analyses from finance to sales to marketing.

Let's get started.

Step 1 - Convert stock prices to a standardized format (daily returns)

What we first need to do is get the data in a format that can be converted to a "user-item" style matrix. The challenge here is to connect the dots between what we have and what we need to do to format it properly.

We know that in order to compare the data, it needs to be standardized or normalized. Why? Because we cannot compare values (stock prices) that are of completely different magnitudes. In order to standardize, we will convert from adjusted stock price (dollar value) to daily returns (percent change from previous day). Here is the formula.

$$return_{daily} = \frac{price_i - price_{i-1}}{price_{i-1}}$$

First, what do we have? We have stock prices for every stock in the SP 500 Index, which is the daily stock prices for over 500 stocks. The data set is over 1.2M observations.

```
sp_500_prices_tbl %>% glimpse()
```

```
## Rows: 1,225,765
## Columns: 8
## $ symbol   <chr> "MSFT", "MSFT", "MSFT", "MSFT", "MSFT", "MSFT", "MSFT", "MSFT~
## $ date     <date> 2009-01-02, 2009-01-05, 2009-01-06, 2009-01-07, 2009-01-08, ~
## $ open     <dbl> 19.53, 20.20, 20.75, 20.19, 19.63, 20.17, 19.71, 19.52, 19.53~
## $ high     <dbl> 20.40, 20.67, 21.00, 20.29, 20.19, 20.30, 19.79, 19.99, 19.68~
## $ low      <dbl> 19.37, 20.06, 20.61, 19.48, 19.55, 19.41, 19.30, 19.52, 19.01~
## $ close    <dbl> 20.33, 20.52, 20.76, 19.51, 20.12, 19.52, 19.47, 19.82, 19.09~
## $ volume   <dbl> 50084000, 61475200, 58083400, 72709900, 70255400, 49815300, 5~
## $ adjusted <dbl> 15.86624, 16.01451, 16.20183, 15.22628, 15.70234, 15.23408, 1~
```

```
sp_500_daily_returns_tbl <- sp_500_prices_tbl %>%
```

```
  select(symbol, date, adjusted) %>%
```

```
  filter(date >= ymd("2018-01-01")) %>%
```

```
  group_by(symbol) %>%
```

```
  mutate(lag_1 = lag(adjusted)) %>%
```

```
  ungroup() %>%
```

```
  filter(!is.na(lag_1)) %>%
```

```
  mutate(diff = adjusted - lag_1) %>%
```

```
  mutate(pct_return = diff / lag_1) %>%
```

```
select(symbol, date, pct_return)

# Apply your data transformation skills
sp_500_daily_returns_tbl
```

```
## # A tibble: 141,340 x 3
##   symbol date      pct_return
##   <chr> <date>      <dbl>
## 1 MSFT  2018-01-03    0.00465
## 2 MSFT  2018-01-04    0.00880
## 3 MSFT  2018-01-05    0.0124
## 4 MSFT  2018-01-08    0.00102
## 5 MSFT  2018-01-09   -0.000680
## 6 MSFT  2018-01-10   -0.00453
## 7 MSFT  2018-01-11    0.00296
## 8 MSFT  2018-01-12    0.0173
## 9 MSFT  2018-01-16   -0.0140
## 10 MSFT 2018-01-17    0.0203
## # i 141,330 more rows
```

```
saveRDS(sp_500_daily_returns_tbl, file = "data01/sp_500_daily_returns_tbl.RDS")
```

Step 2 - Convert to User-Item Format

```
stock_date_matrix_tbl <- readRDS("data01/sp_500_daily_returns_tbl.rds") %>%
  spread(key = date, value = pct_return, fill = 0)
```

```
stock_date_matrix_tbl
```

```
## # A tibble: 502 x 283
##   symbol '2018-01-03' '2018-01-04' '2018-01-05' '2018-01-08' '2018-01-09'
##   <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 A          0.0254      -0.00750    0.0160      0.00215    0.0246
## 2 AAL        -0.0123       0.00630   -0.000380   -0.00988   -0.000959
## 3 AAP         0.00905       0.0369    0.0106     -0.00704   -0.00808
## 4 AAPL       -0.000174      0.00465    0.0114     -0.00371  -0.000115
## 5 ABBV        0.0156      -0.00570    0.0174     -0.0160    0.00754
## 6 ABC         0.00372     -0.00222    0.0121      0.0166    0.00640
## 7 ABMD        0.0173       0.0175    0.0154      0.0271    0.00943
## 8 ABT         0.00221     -0.00170    0.00289    -0.00288    0.00170
## 9 ACN         0.00462       0.0118    0.00825     0.00799    0.00333
## 10 ADBE        0.0188       0.0120    0.0116     -0.00162    0.00897
## # i 492 more rows
## # i 277 more variables: '2018-01-10' <dbl>, '2018-01-11' <dbl>,
## #   '2018-01-12' <dbl>, '2018-01-16' <dbl>, '2018-01-17' <dbl>,
## #   '2018-01-18' <dbl>, '2018-01-19' <dbl>, '2018-01-22' <dbl>,
## #   '2018-01-23' <dbl>, '2018-01-24' <dbl>, '2018-01-25' <dbl>,
## #   '2018-01-26' <dbl>, '2018-01-29' <dbl>, '2018-01-30' <dbl>,
## #   '2018-01-31' <dbl>, '2018-02-01' <dbl>, '2018-02-02' <dbl>, ...
```

Step 3 - Perform K-Means Clustering

```
kmeans_obj <- stock_date_matrix_tbl %>%  
  select(-symbol) %>%  
  kmeans(centers = 4, nstart = 20)  
  
kmeans_obj %>% glance()
```

```
## # A tibble: 1 x 4  
##   totss tot.withinss betweenss iter  
##   <dbl>      <dbl>      <dbl> <int>  
## 1  33.6         29.2         4.40     3
```

Step 4 - Find the optimal value of K

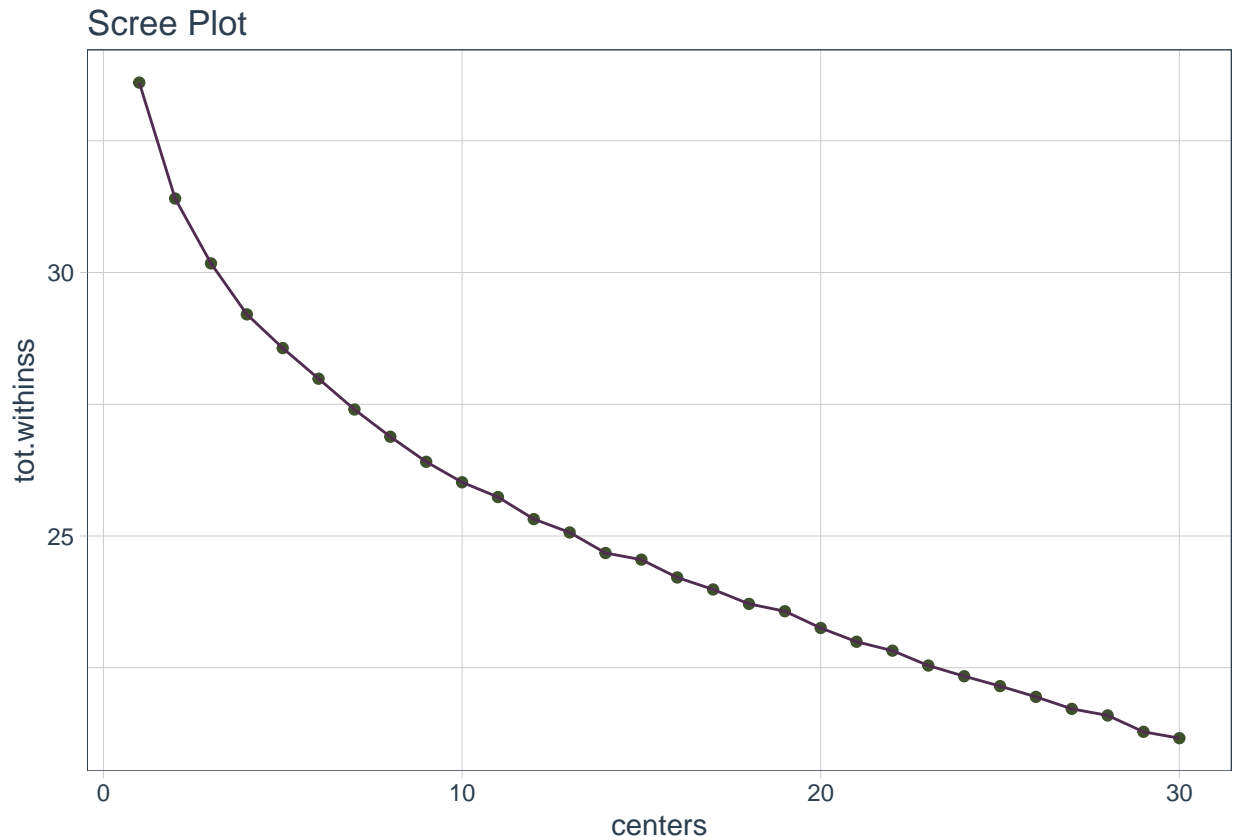
```
# Lets use `purrr` to iterate over many values of "k" using the `centers` argument.  
kmeans_mapper <- function(center = 3) {  
  stock_date_matrix_tbl %>%  
    select(-symbol) %>%  
    kmeans(centers = center, nstart = 20)  
}  
  
# Apply the `kmeans_mapper()` and `glance()` functions iteratively using `purrr`.  
  
k_means_mapped_tbl <- tibble(centers = 1:30) %>%  
  mutate(k_means = centers %>% map(kmeans_mapper)) %>%  
  mutate(glance = k_means %>% map(glance))  
  
#Output: k_means_mapped_tbl  
k_means_mapped_tbl
```

```
## # A tibble: 30 x 3  
##   centers k_means glance  
##   <int> <list> <list>  
## 1     1 1 <kmeans> <tibble [1 x 4]>  
## 2     2 2 <kmeans> <tibble [1 x 4]>  
## 3     3 3 <kmeans> <tibble [1 x 4]>  
## 4     4 4 <kmeans> <tibble [1 x 4]>  
## 5     5 5 <kmeans> <tibble [1 x 4]>  
## 6     6 6 <kmeans> <tibble [1 x 4]>  
## 7     7 7 <kmeans> <tibble [1 x 4]>  
## 8     8 8 <kmeans> <tibble [1 x 4]>  
## 9     9 9 <kmeans> <tibble [1 x 4]>  
## 10    10 10 <kmeans> <tibble [1 x 4]>  
## # i 20 more rows
```

Scree Plot

Next, let's visualize the "tot.withinss" from the glance output as a *Scree Plot*.


```
k_means_mapped_tbl %>%
  unnest(glance) %>%
  ggplot(aes(centers, tot.withinss)) +
  geom_point(color = "#3e502c") +
  geom_line(color = "#502c50") +
  labs(title = "Scree Plot") +
  theme_tq()
```



We can see that the Scree Plot becomes linear (constant rate of change) between 5 and 10 centers for K.

Step 5 - Apply UMAP

Next, let's plot the UMAP 2D visualization to help us investigate cluster assignments.

```
# let's plot the `UMAP` 2D visualization to help us investigate cluster assignments.
umap_results <- stock_date_matrix_tbl %>%
  select(-symbol) %>%
  umap()

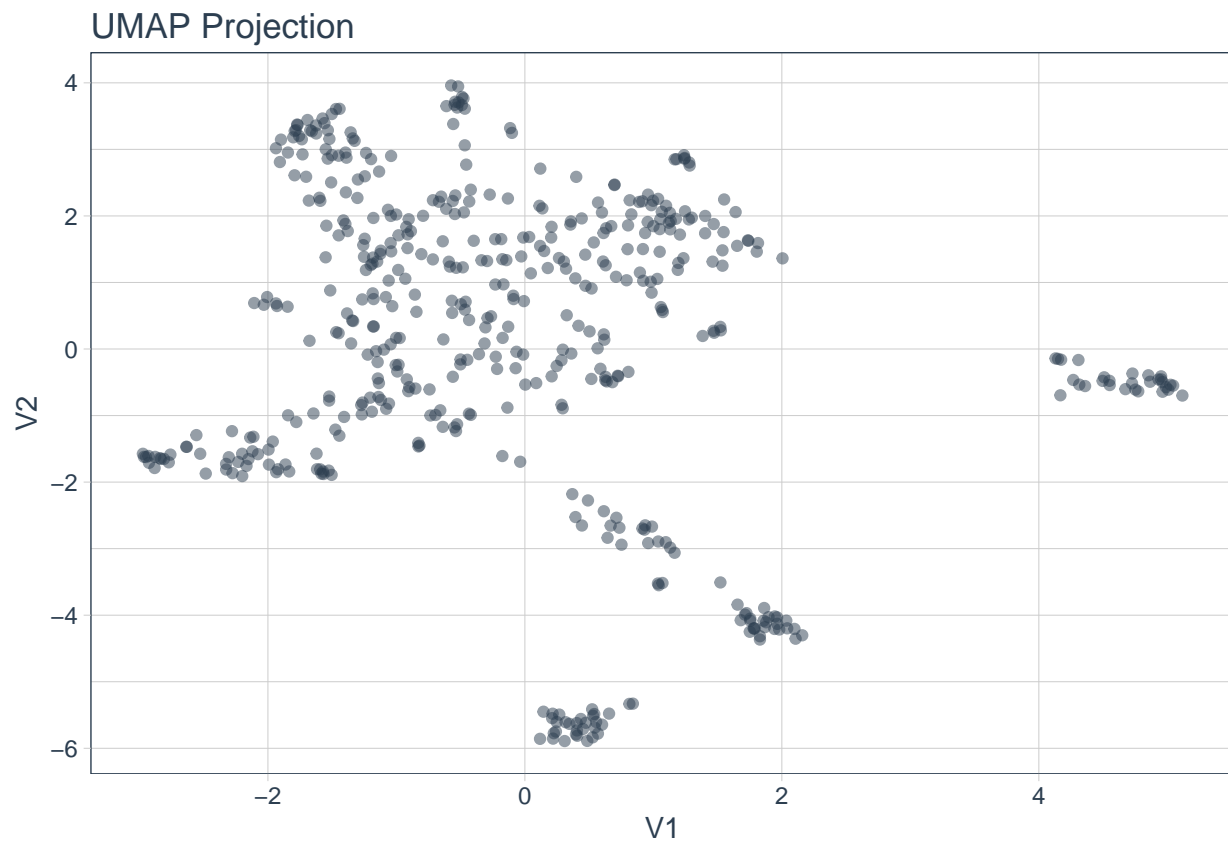
# Next, we want to combine the `layout` from the `umap_results` with the `symbol` column from the `stock_date_matrix_tbl`
umap_results_tbl <- umap_results$layout %>%
  as_tibble() %>%
  bind_cols(stock_date_matrix_tbl %>% select(symbol))

umap_results_tbl
```

```
## # A tibble: 502 x 3
##       V1      V2 symbol
##   <dbl> <dbl> <chr>
## 1 -0.612  2.11    A
## 2  1.52   0.333  AAL
## 3  0.361 -0.0678  AAP
## 4 -1.33   3.13   AAPL
## 5 -1.18   0.338  ABBV
## 6 -1.22  -0.0847  ABC
## 7 -1.79   3.28   ABMD
## 8 -1.12   1.48   ABT
## 9 -0.983  1.71   ACN
## 10 -1.74   3.15  ADBE
## # i 492 more rows
```

Finally, let's make a quick visualization of the `umap_results_tbl`.

```
umap_results_tbl %>%
  ggplot(aes(V1, V2)) +
  geom_point(alpha = 0.5, color = "#2c3e50") +
  theme_tq() +
  labs(title = "UMAP Projection")
```



We can now see that we have some clusters. However, we still need to combine the K-Means clusters and the UMAP 2D representation.

Step 6 - Combine K-Means and UMAP

```
# Next, we combine the K-Means clusters and the UMAP 2D representation
# First, pull out the K-Means for 10 Centers. Use this since beyond this value the Scree Plot flattens.

k_means_mapped_tbl <- read_rds("data01/k_means_mapped_tbl.rds")
umap_results_tbl   <- read_rds("data01/umap_results_tbl.rds")

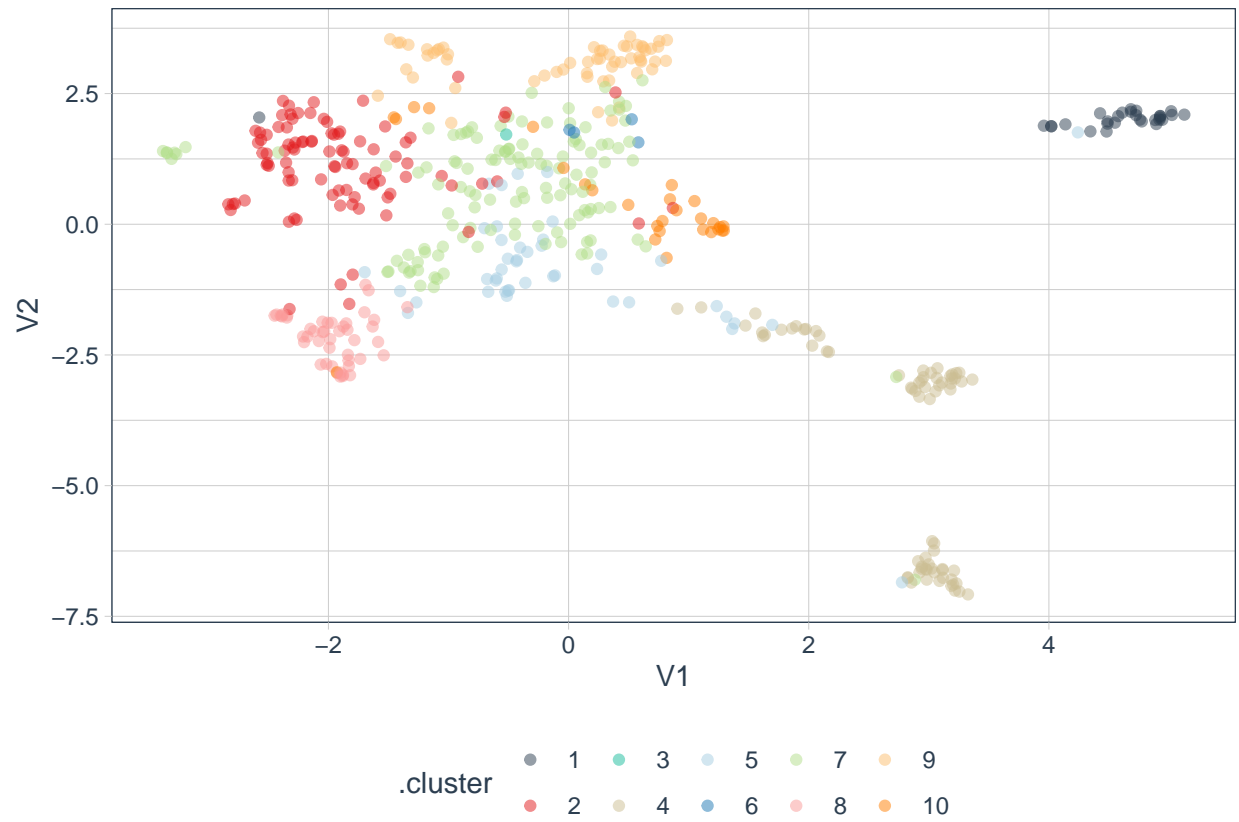
k_means_obj <- k_means_mapped_tbl %>%
  filter(centers == 10) %>%
  pull(k_means) %>%
  pluck(1)

# Next, we'll combine the clusters from the `k_means_obj` with the `umap_results_tbl`.

umap_kmeans_results_tbl <- k_means_obj %>%
  augment(stock_date_matrix_tbl) %>%
  select(symbol, .cluster) %>%
  left_join(umap_results_tbl, by = "symbol") %>%
  left_join(sp_500_index_tbl %>% select(symbol, company, sector),
            by = "symbol")

# Plot the K-Means and UMAP results.

umap_kmeans_results_tbl %>%
  ggplot(aes(V1, V2, color = .cluster)) +
  geom_point(alpha = 0.5) +
  theme_tq() +
  scale_color_tq()
```



Congratulations! We are done with the 1st challenge!