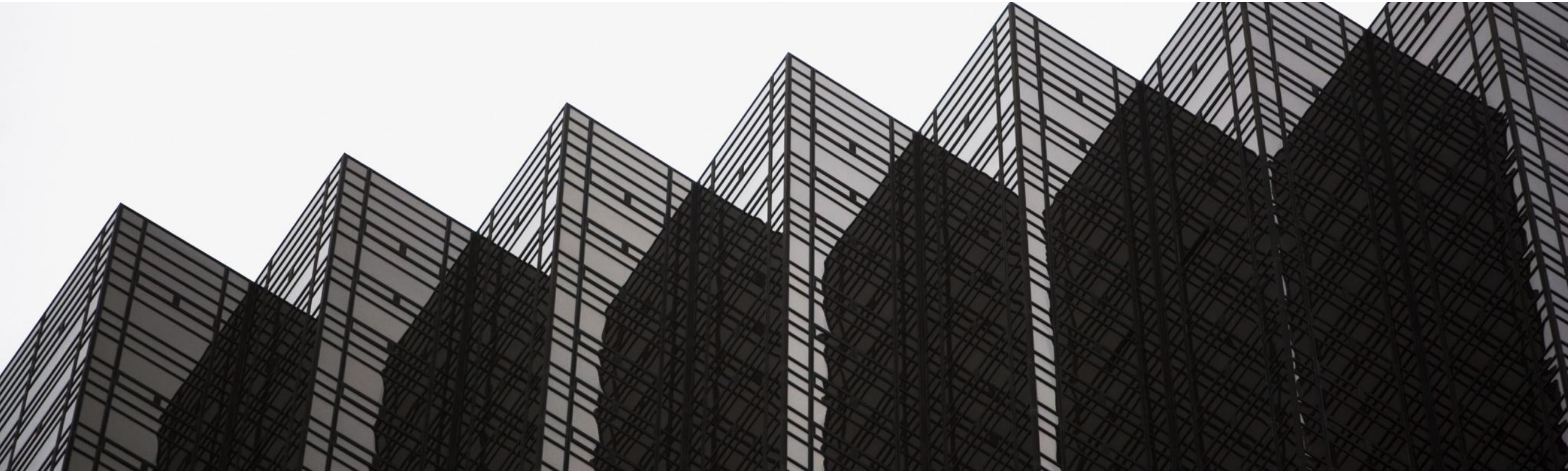
Data Engineering Platform
Final Project
Russell 3000 Companies Database



S&P CAPITAL IQ

Agenda

- Executive Summary
- Business Use Case and Data
- Technology and Process
- Data Preparation
- Data Modelling and Design
- Data Visualization and Findings
- Lesson Learned

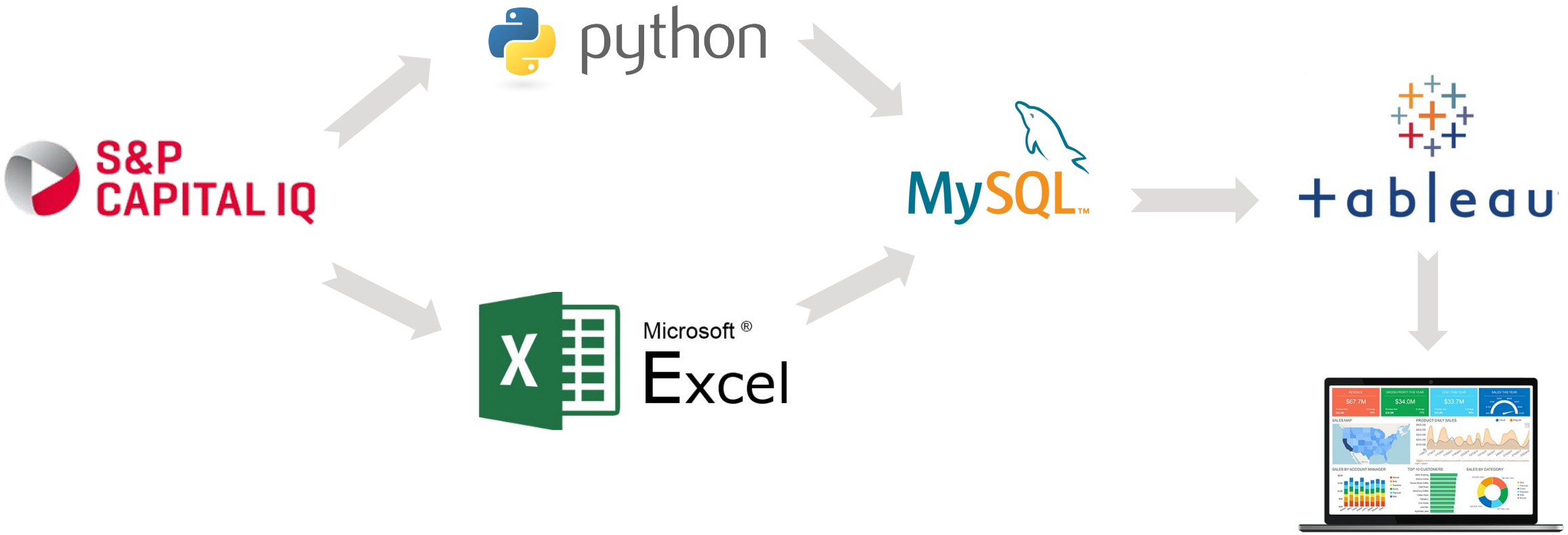
Business Use Case and Data

- Business Use Cases:
  - Analyze business events, characteristics, and personnel
    - Sample queries:
      - Who serves on the board, and are there patterns in board membership?
      - Who owns what? Who are a company's investors and where are investors concentrated?
      - Are there geographical patterns in number and size of companies?
      - Which industries have the highest P/E ratio?
  - Study relationships between entities
- Data:
  - Data – S&P Capital IQ

Design Considerations

- Analytical not Operational system
- Resembles but is not a dimensional model
  - but not modeling a business process, no fact table
  - means more complex queries
- Compromised ETL portion because limited access to underlying data (which was so good)
- Some information over time but most is snapshot

Technology and Process



Data Preparation

Raw data

- Common attributes: company ID,company name
- Distinguish attributes:

1. Company information:website, description, address,industry, country
2. Board: board member ID, board member name, board member title
3. Executive: executive ID, executive name, executive title
4. Owners: holder buyer ID, type,shares, values
5. Financial: stock price, year, market cap, P/E, EV/EBITDA,EBITA, sales,growth

Data Preparation

Excel

- Separate tables
- Create surrogate keys
- Check for different values

Company_ID	Company_Board_Person_ID
IQ24068100	1
IQ2878904	2
IQ13376023	3
IQ278649	4
IQ274166	5
IQ386991	6
IQ381013111	7
IQ342014	8
IQ33679	9
IQ44316034	10
IQ340047	11
IQ1797684	12
IQ317440	13
IQ254497048	14

Pandas

- Create new dataframe
- Delete data
- Fill data
- Replace
- Check for duplicates
- Check for missing values
- Separate non-atomic data

```
dataframe1 = dataframe1.drop(columns = ['Company Description'])
dataframe1 = dataframe1.drop(columns = ['Website'])
dataframe1.head()
```

	Company_ID	Company_name	Ticker	Industry_ID	Year_founded	Address_1	Address_2	City	State	Country
0	IQ24085	1-800-FLOWERS.COM, Inc.	NasdaqGS:FLWS	18	1976.0	One Old Country Road	Suite 500	Carle Place	NY	United States
1	IQ176404	1st Source Corporation	NasdaqGS:SRCE	2	1863.0	100 North Michigan Street	None	South Bend	IN	United States
2	IQ60414516	2U, Inc.	NasdaqGS:TWOU	20	2008.0	7900 Harkins Road	None	Lanham	MD	United States
3	IQ308402	3D Systems Corporation	NYSE:DDD	21	1986.0	333 Three D Systems Circle	None	Rock Hill	SC	United States
4	IQ289194	3M Company	NYSE:MMM	3	1902.0	3M Center	None	St. Paul	MN	United States

Data Preparation

- Missing data: keep vs. drop

	Company_ID	Company Name	Ticker	Industry	Year founded	Address 1	Address 2	City	State	Country
0	IQ24937	Apple Inc.	NasdaqGS:AAPL	Technology Hardware and Equipment	1977	1 Infinite Loop	0	Cupertino	CA	United States
1	IQ29096	Alphabet Inc.	NasdaqGS:GOOGL	Software and Services	1998	1600 Amphitheatre Parkway	0	Mountain View	CA	United States
2	IQ18749	Amazon.com, Inc.	NasdaqGS:AMZN	Retailing	1994	410 Terry Avenue North	0	Seattle	WA	United States
3	IQ21835	Microsoft Corporation	NasdaqGS:MSFT	Software and Services	1975	One Microsoft Way	0	Redmond	WA	United States
4	IQ20765463	Facebook, Inc.	NasdaqGS:FB	Software and Services	2004	1601 Willow Road	0	Menlo Park	CA	United States

```
dataframe2=dataframe1[dataframe1.Company_ID!= '(Invalid Identifier)']
#Capital IQ did not recognize every company in the Russell 300 index, however this relatively negligible.
#We removed these unfound companies
```

```
dataframe2.isnull().any()
```

```
Company_ID      False
Company Name    False
Ticker          False
Industry        False
Year founded    False
Address 1       False
Address 2       False
City            False
State           False
Country         False
dtype: bool
```

- Duplicates: keep vs. drop

```
In [5]: dataframe2.drop_duplicates(subset=None, keep='first', inplace=True)
        #We got out company list from the Russell 3000 index,  Some companies are duplicated as a result of Russell's effort to provide
        #continuity in index pricing. However, for our purposes we do not need such duplicates.
```

37	IQ556799812	Abernathy	Robert	Independent Director
38	IQ556799812	Abernathy	Robert	Independent Director
39	IQ556799812	Abernathy	Robert	Independent Director

Data Preparation

- Non-atomic values

```
dataframe5[['Board_FirstName','Board_lastName']] = dataframe5['Board_Member_Name'].loc[dataframe5['Board_Member_Name'].str.split(
#We split the board member names into first and last name to keep the data atomic
```

```
dataframe5.head()
```

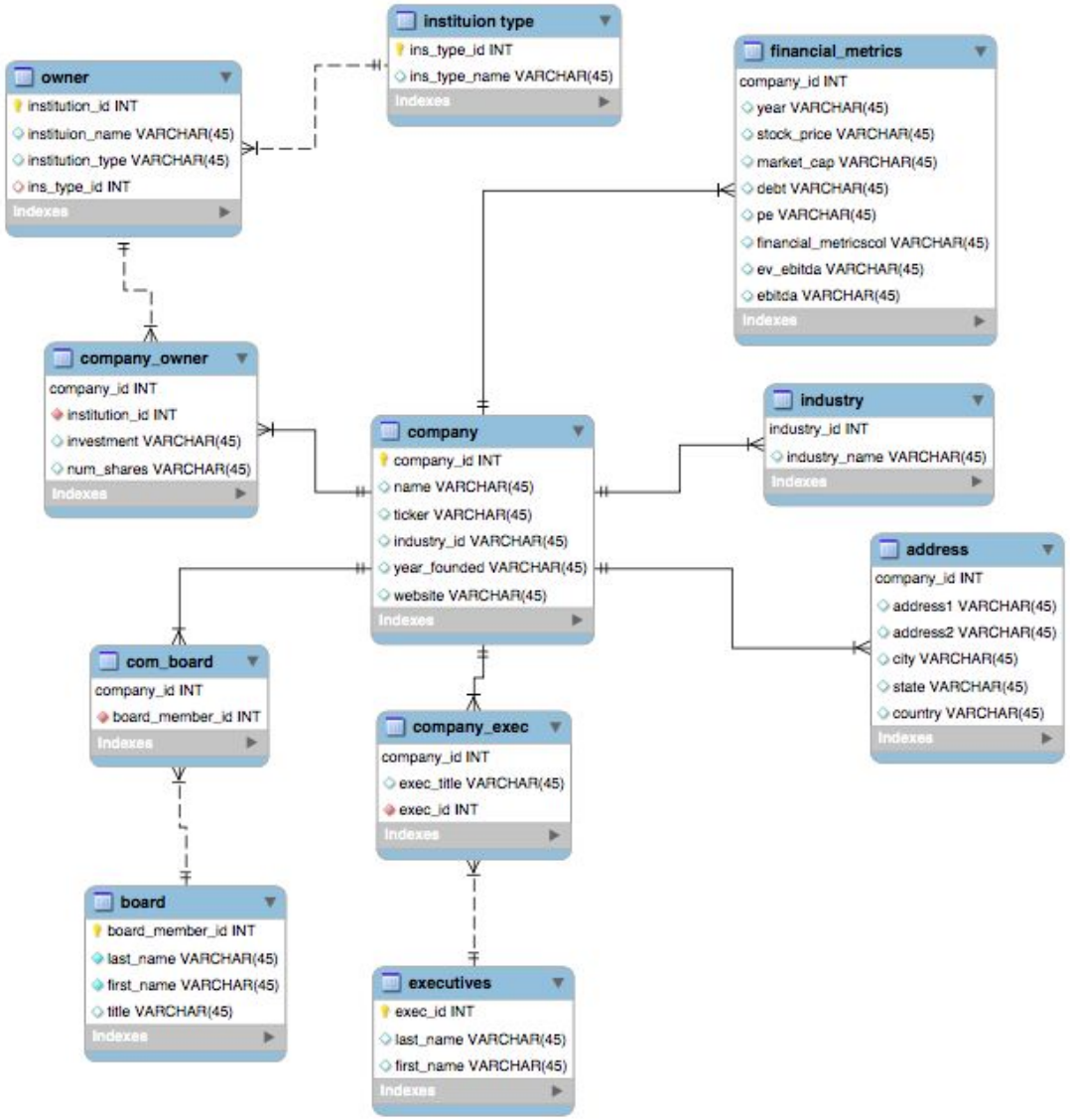
	Board Member ID	Board_Member_Name	Board Member Title	Board_FirstName	Board_lastName
0	IQ146193279	Geralyn Breig	Independent Director	Geralyn	Breig
1	IQ370363740	Celia Brown	Director	Celia	Brown
2	IQ35345098	James Cannavino	Independent Director	James	Cannavino
3	IQ146193192	Eugene DeMark	Director	Eugene	DeMark
4	IQ3721903	Leonard Elmore	Director	Leonard	Elmore

Ticker
NasdaqGS:AAPL
NasdaqGS:GOOGL
NasdaqGS:AMZN
NasdaqGS:MSFT
NasdaqGS:FB

ticker
AAPL
GOOGL
AMZN
MSFT
FB

- Data types: string vs. date

EER Diagram. v1



First version of EER Diagram

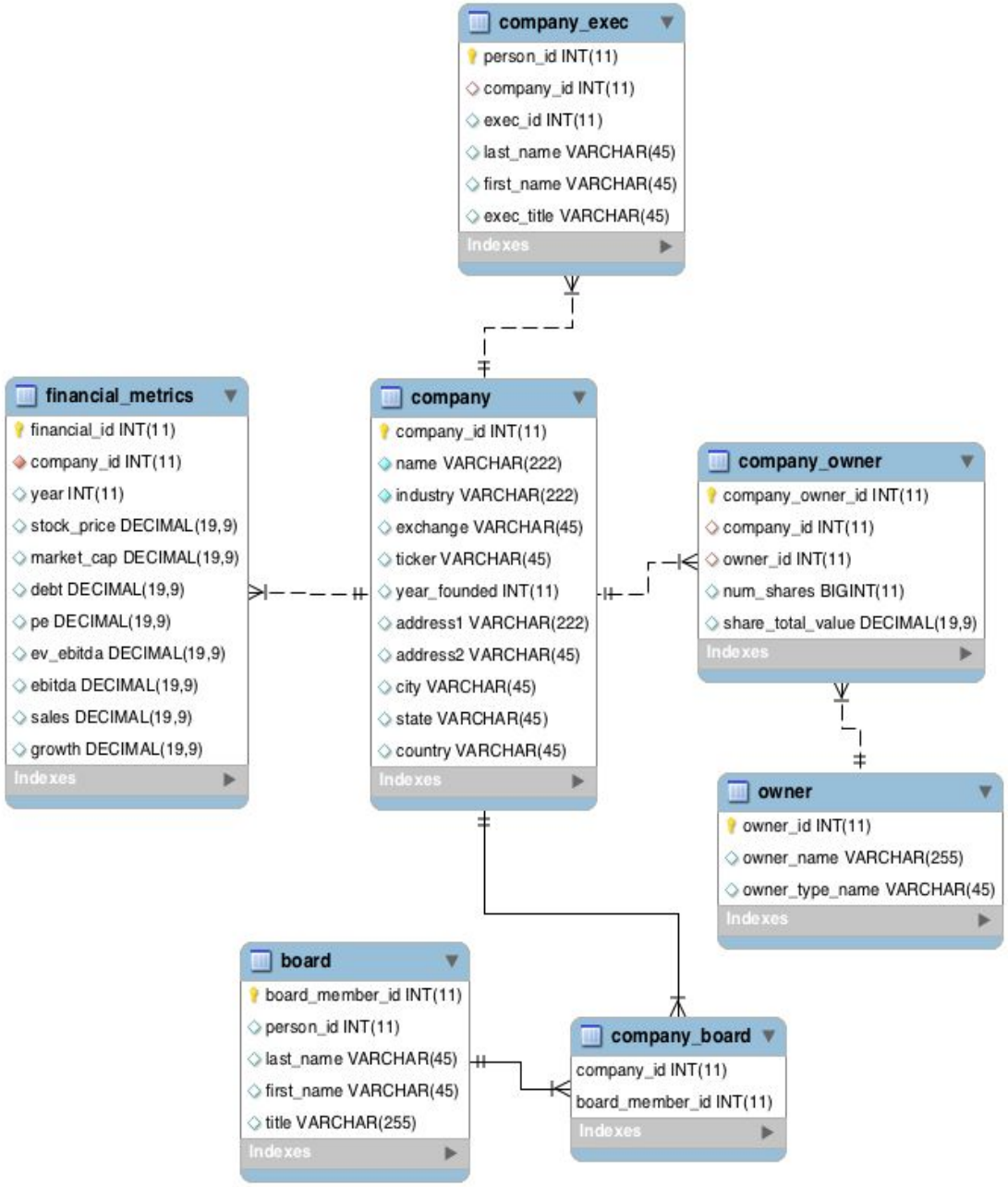
Company business characteristics, people, owners, some financial performance

“Cannot add foreign key constraint”

Problems?

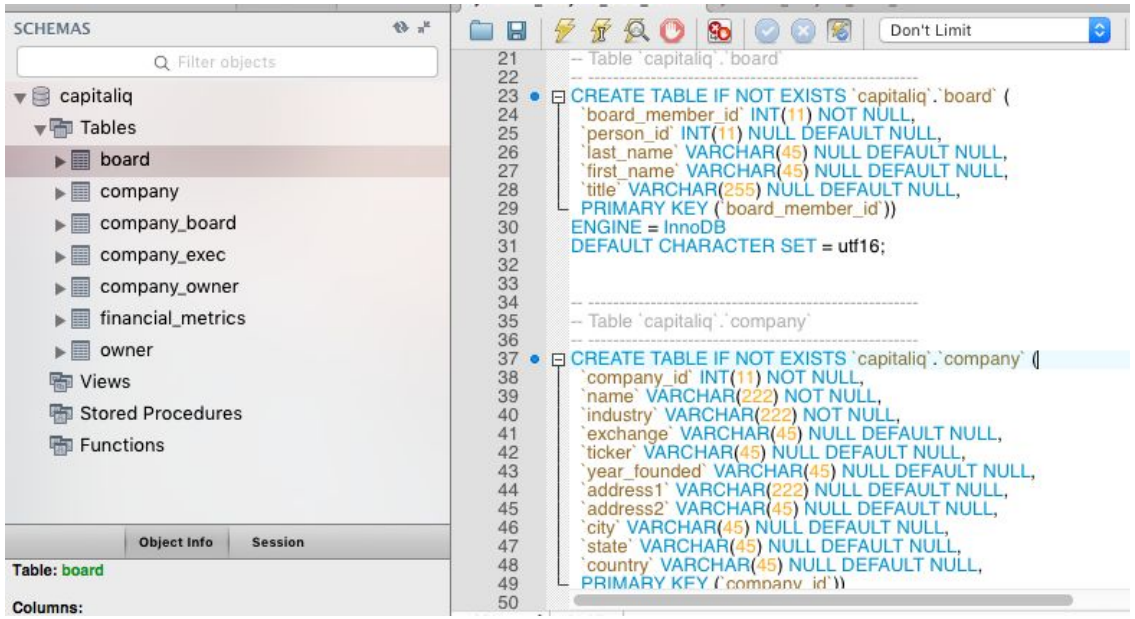
- Too many child tables (unnecessary table)
- Created duplicate primary key and foreign key

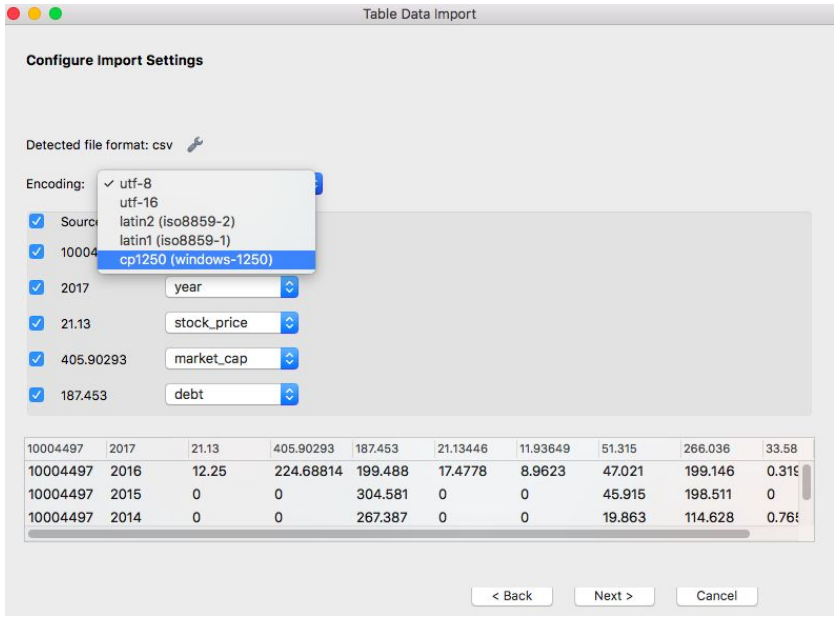
EER Diagram. v2



- Consolidated some tables together
  - industry, address -> company
  - institution type -> owner
  - exec -> company_exec
- Created composite primary keys, surrogate keys (compnay_owner_id, financial_id,board_member_id...)
- Redefined relationships between the entities
  - Some cases one-to-one (CEO)
  - Some cases one-to-many (board members, owners)
  - Many-to-many (owner positions in companies)

Data Modelling and Design

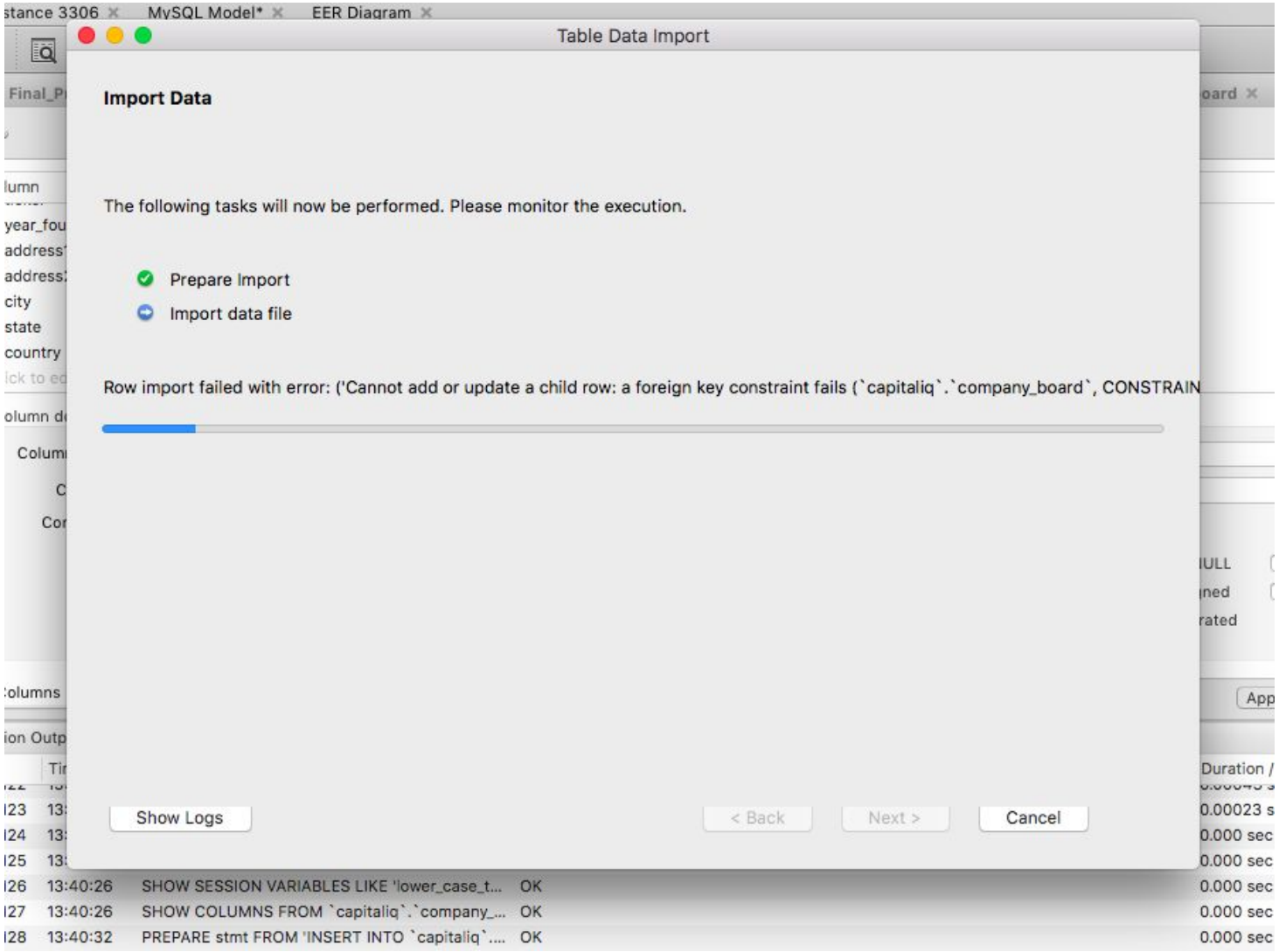




- Created schema and tables in the ER diagram and used reverse engineer to generate DDL script
- Imported .csv data by Table Data Import Wizard
- Used cp1250 encoding engine when importing large dataset (avoid losing information)

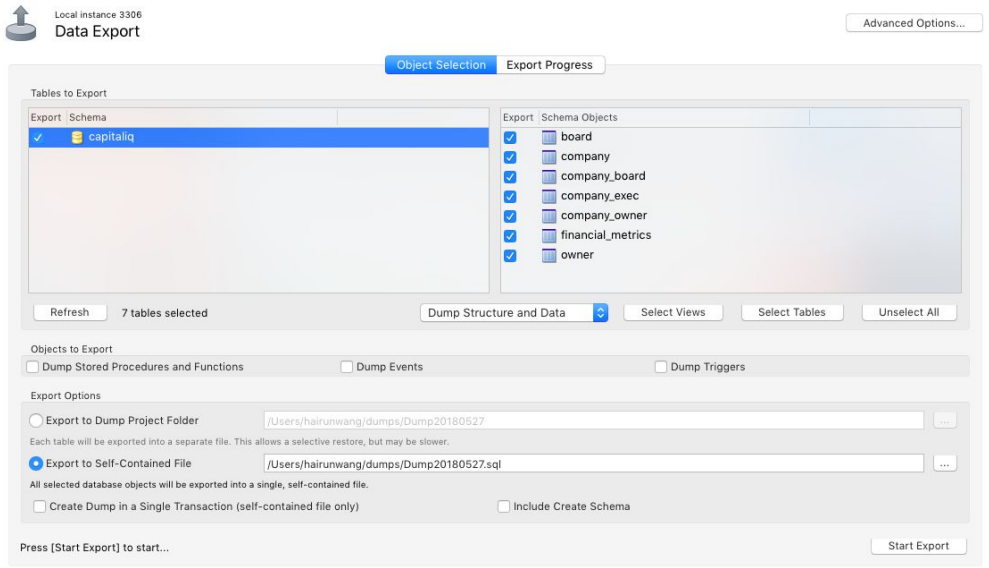
Data Modelling and Design

- Order matters! (Import data into parent table first)
- FK in the child table must match with parent table
- Data validation using Excel (EXACT function)
- Make sure data type is consistent (No string value in INT() column)
- Re-clean the data and truncate table and import data again until no error message popup



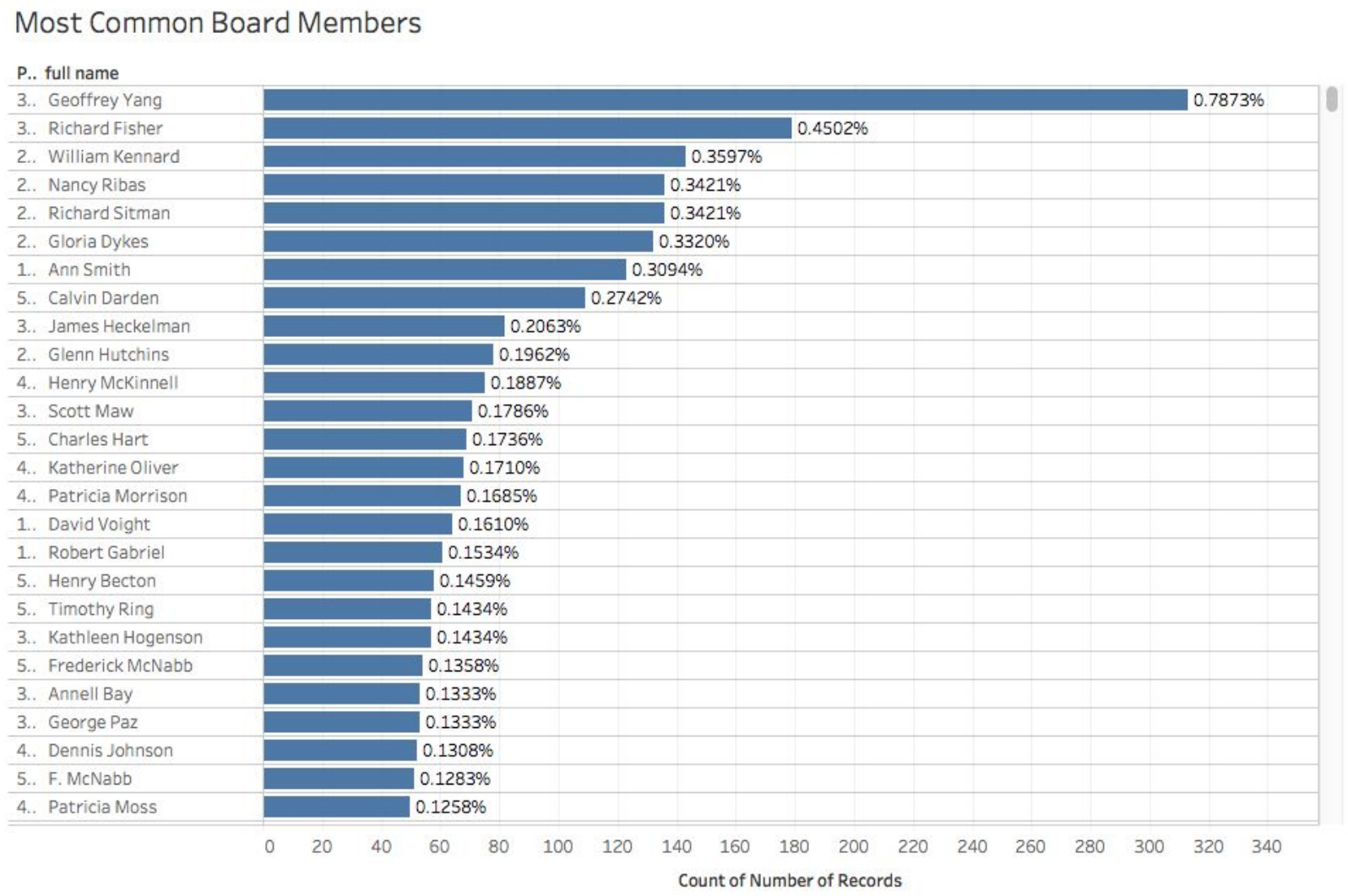
Data Modelling and Design

```
  `board_member_id` int(11) NOT NULL,
  `person_id` int(11) DEFAULT NULL,
  `last_name` varchar(45) DEFAULT NULL,
  `first_name` varchar(45) DEFAULT NULL,
  `title` varchar(255) DEFAULT NULL,
  PRIMARY KEY (`board_member_id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf16;
/*!40101 SET character_set_client = @saved_cs_client */;

--
-- Dumping data for table `board`
--

LOCK TABLES `board` WRITE;
/*!40000 ALTER TABLE `board` DISABLE KEYS */;
INSERT INTO `board` VALUES (1,53556254,'Aalaei',' Faraj','Chairman, President & CEO'),(2,248879511,'Aamir',' Mir','President, CEO & Director'
INSERT INTO `board` VALUES (18327,276369338,'Kennard',' William','Director'),(18328,276369338,'Kennard',' William','Director'),(18329,276369
INSERT INTO `board` VALUES (36417,545407528,'Veihmeyer',' John','Independent Director'),(36418,545407528,'Veihmeyer',' John','Independen
/*!40000 ALTER TABLE `board` ENABLE KEYS */;
UNLOCK TABLES;

--
-- Table structure for table `company`
--

DROP TABLE IF EXISTS `company`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `company` (
```



- All SQL scripts were saved in local machine, so we exported data as dump file in order to share among team members
- Drawback: you have to redo the entire process every time when you update data

Data Visualization and Findings



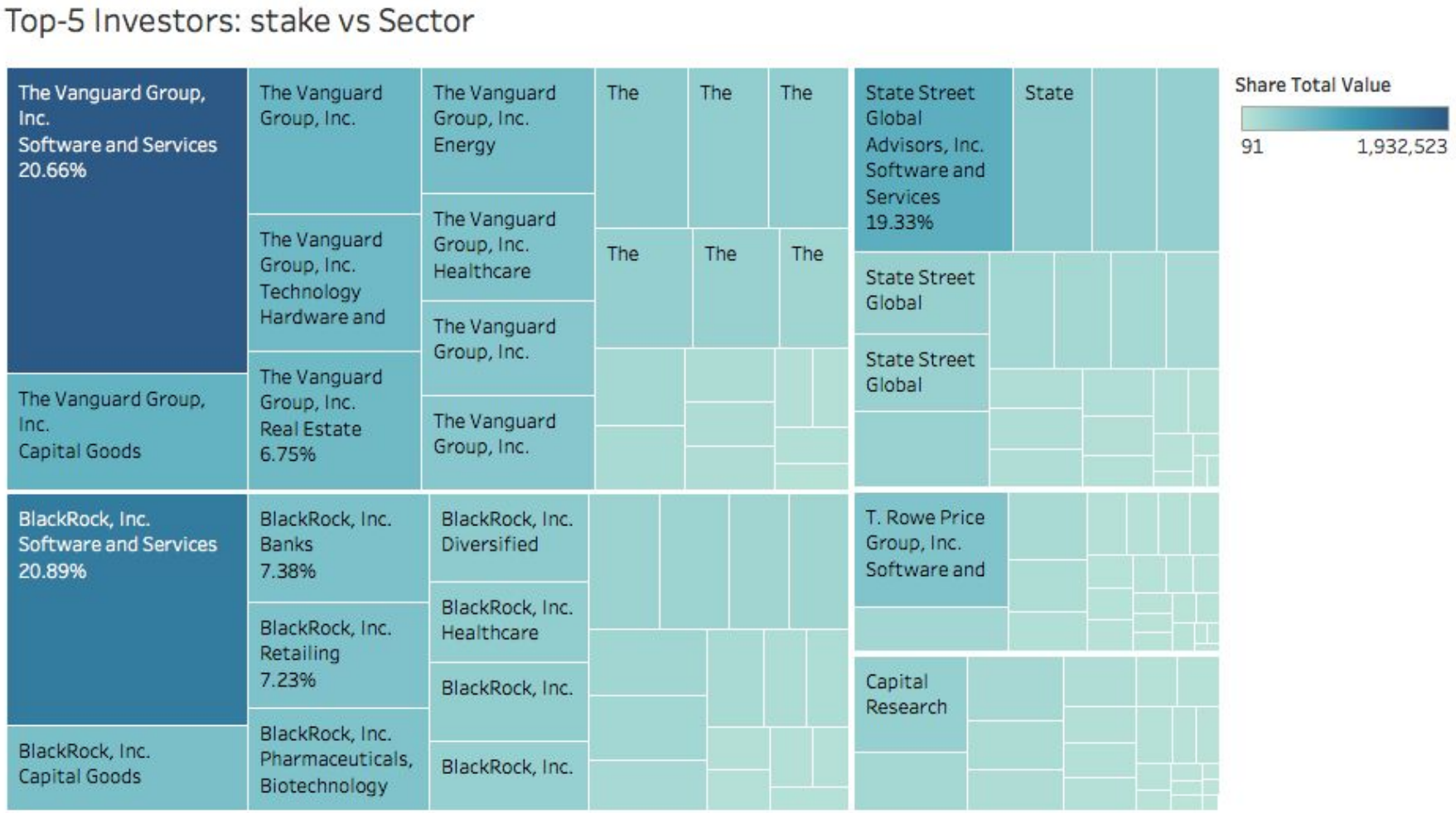
- Gathered info on 25 board members for each of the Russell 3000 companies
- Aggregated to rank board members with the highest seats
- A fund manager, Geoffrey Yang, holds the most at 313
- He holds ~1% of all public board seats in the Russell 3000 companies

Data Visualization and Findings

By Industry

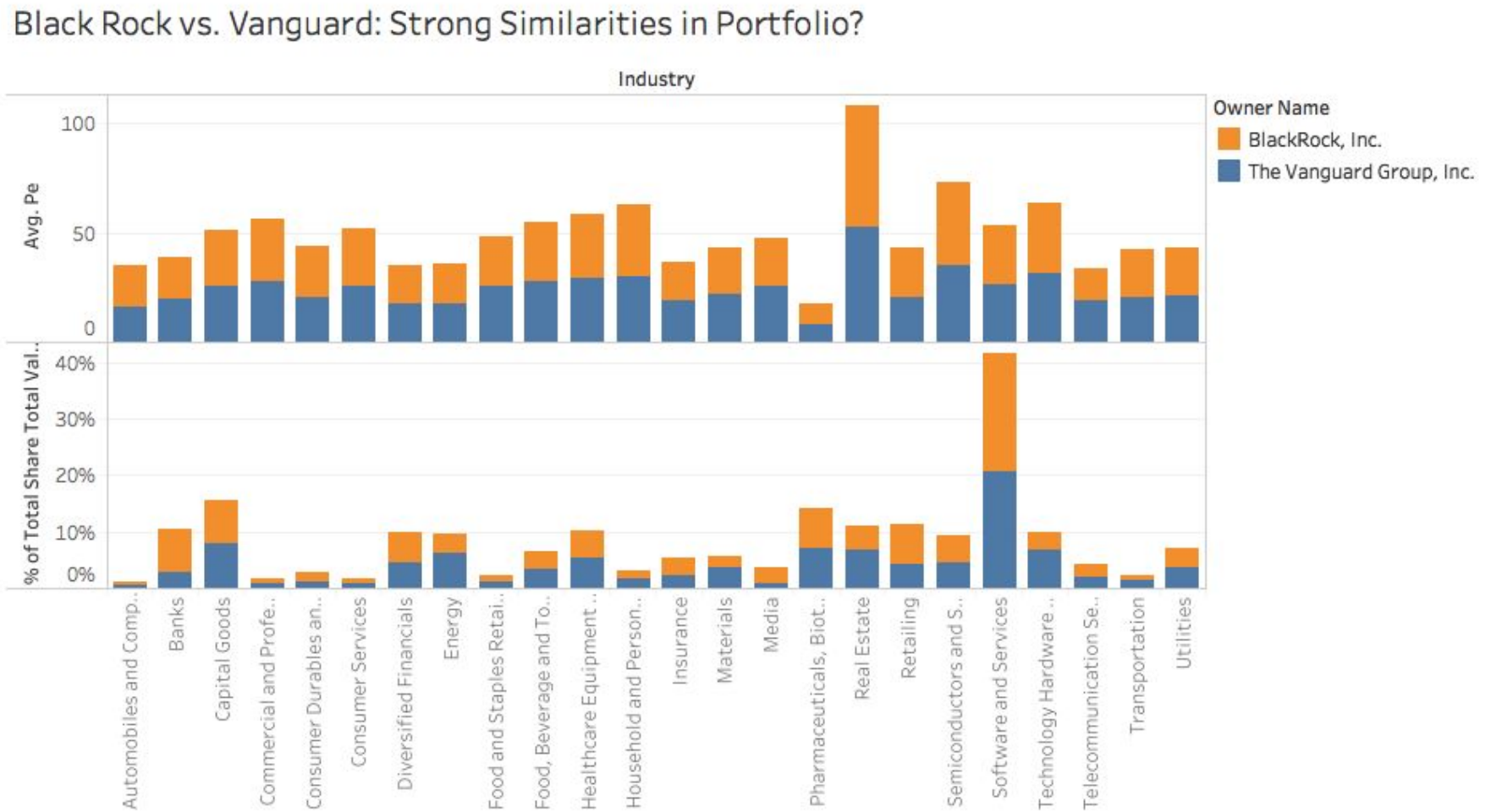
Person Id	full name	Industry												
		Automobiles and Components	Banks	Capital Goods	Commercial and Professional ..	Consumer Durables and Apparel	Consumer Services	Diversified Financials	Energy	Food and Staples Retailing	Food, Beverage and Tobacco	Healthcare Equipment and Ser..	Household and Personal Produ..	Insuran..	Mat
3..	Geoffrey Yang	1%	5%	10%	3%	4%	2%	4%	9%	0%	3%	6%	1%	3%	
3..	Richard Fisher	1%	4%	8%	2%	2%	2%	5%	8%	1%	2%	8%	1%	3%	
2..	William Kennard	1%	3%	9%	2%	2%	1%	6%	8%	1%	1%	8%	1%	3%	
2..	Nancy Ribas	1%	10%	13%	3%	6%	4%	2%	7%	1%	2%	3%	1%	3%	
2..	Richard Sitman	1%	10%	13%	3%	6%	4%	2%	7%	1%	2%	3%	1%	3%	
2..	Gloria Dykes	2%	9%	13%	3%	6%	5%	2%	7%	1%	2%	3%	1%	3%	
1..	Ann Smith	1%	7%	13%	2%	6%	5%	2%	7%	1%	2%	3%	1%	3%	
5..	Calvin Darden	2%	5%	15%	6%	2%	6%	5%	8%		2%	8%	1%	3%	
3..	James Heckelman		5%	6%	4%	1%	7%	7%	11%		1%	4%	1%	4%	
2..	Glenn Hutchins			10%	3%	3%	3%	5%	9%	1%	3%	8%		1%	
4..	Henry McKinnell	3%	5%	15%	1%	8%	4%	1%	7%	1%	1%	11%		3%	
3..	Scott Maw		6%	8%	1%	7%		10%	7%			6%	1%		
5..	Charles Hart	3%	14%	12%		3%	7%	1%	6%	3%	3%	6%		3%	
4..	Katherine Oliver		3%	10%	3%	4%	3%	1%	7%		3%	9%		3%	
4..	Patricia Morrison	3%	4%	10%	7%	1%	9%	6%	10%		1%	1%	1%	3%	
1..	David Voight		3%	8%	2%		8%	8%	13%		2%	5%	2%	5%	
1..	Robert Gabriel		11%	15%		8%	5%	3%	7%		2%	2%			
5..	Henry Becton		3%	14%	3%			5%	7%			10%		2%	
5..	Timothy Ring		4%	14%	4%			5%	7%			11%		2%	
3..	Kathleen Hogens..	2%	5%	14%	5%	2%	2%	5%	7%			5%			
5..	Frederick McNabb	2%	2%	11%	6%	4%	6%	2%	9%	2%	4%	6%		4%	
3..	Annell Bay		4%	13%	6%	2%	2%	6%	8%			6%			
3..	George Paz		4%	13%	4%	6%	4%	4%	11%		4%	4%		4%	
4..	Dennis Johnson	4%	17%	12%		4%	8%	2%	6%	4%	2%	6%		2%	

Data Visualization and Findings



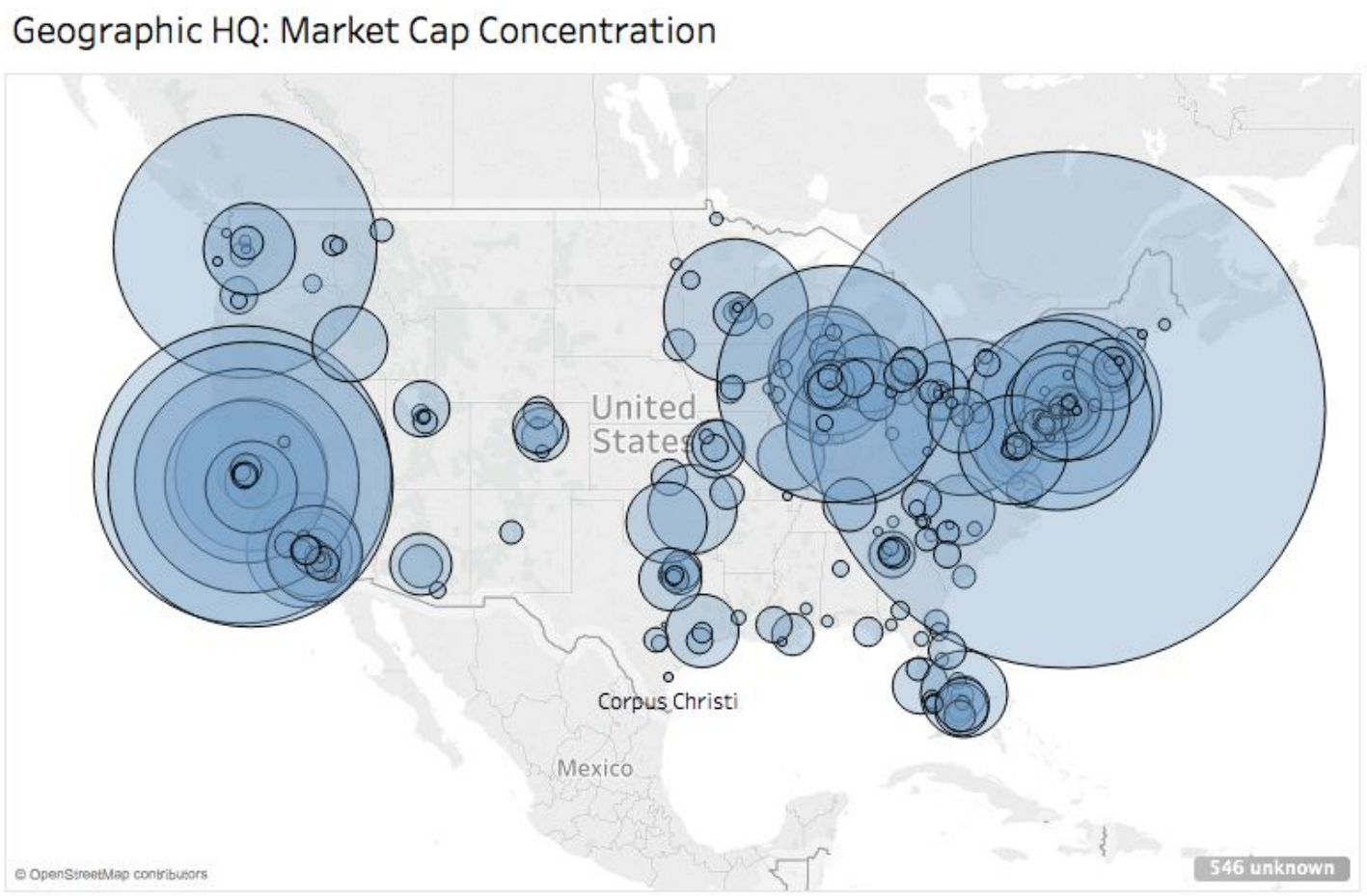
- We observed ownership stakes of each public company
- Not surprisingly, Vanguard and BlackRock held frequent stakes and large positions
- What was peculiar is Vanguard and BlackRock holdings similar market stakes in size and industry

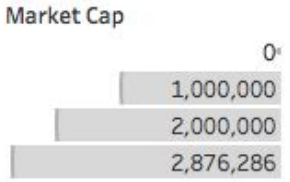
Data Visualization and Findings



- Observing more closely, the top-2 funds held near-identical stakes
- This reveals similar portfolio strategies and market assumptions
- These allocations also match the broader market index

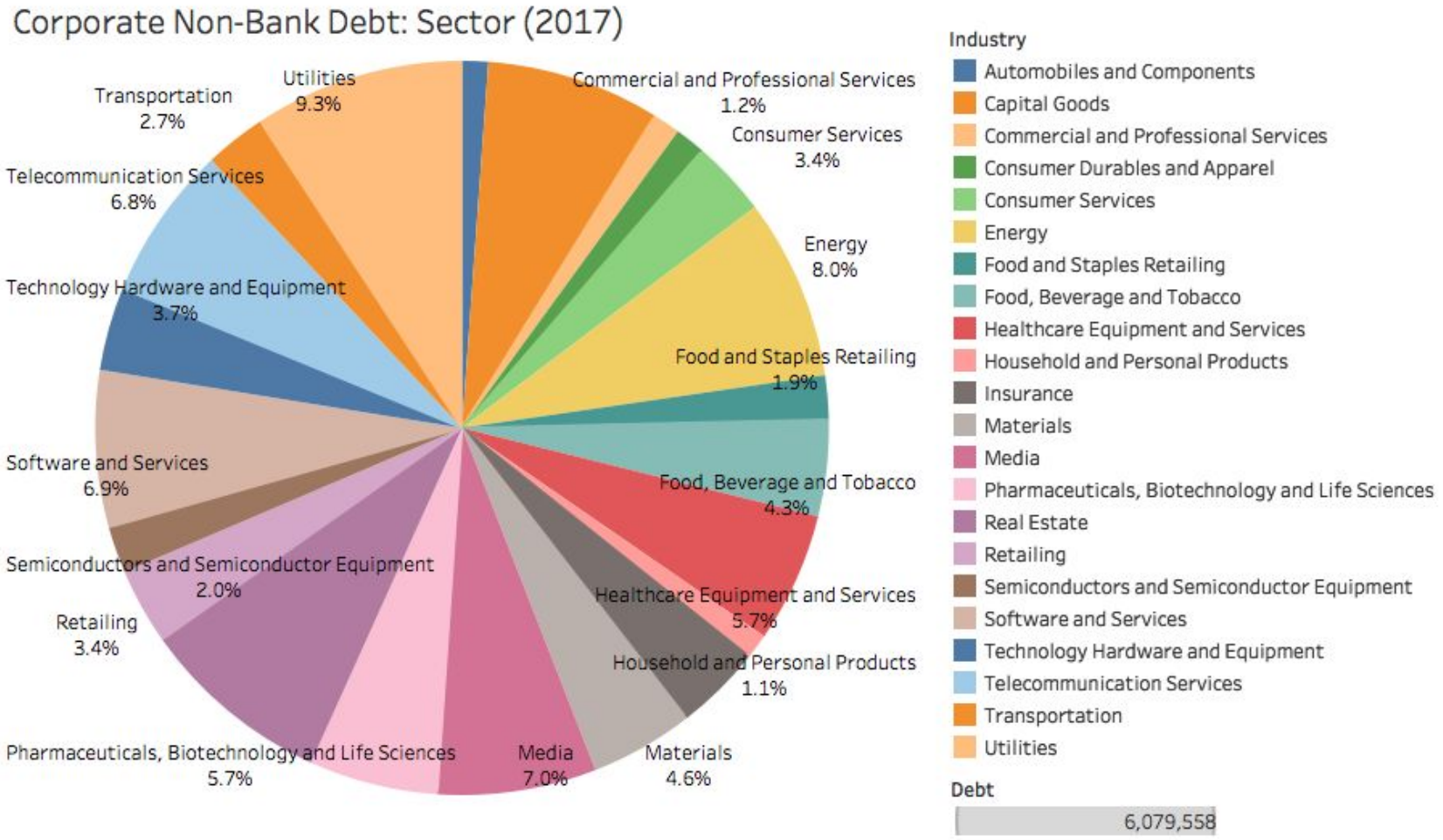
Data Visualization and Findings





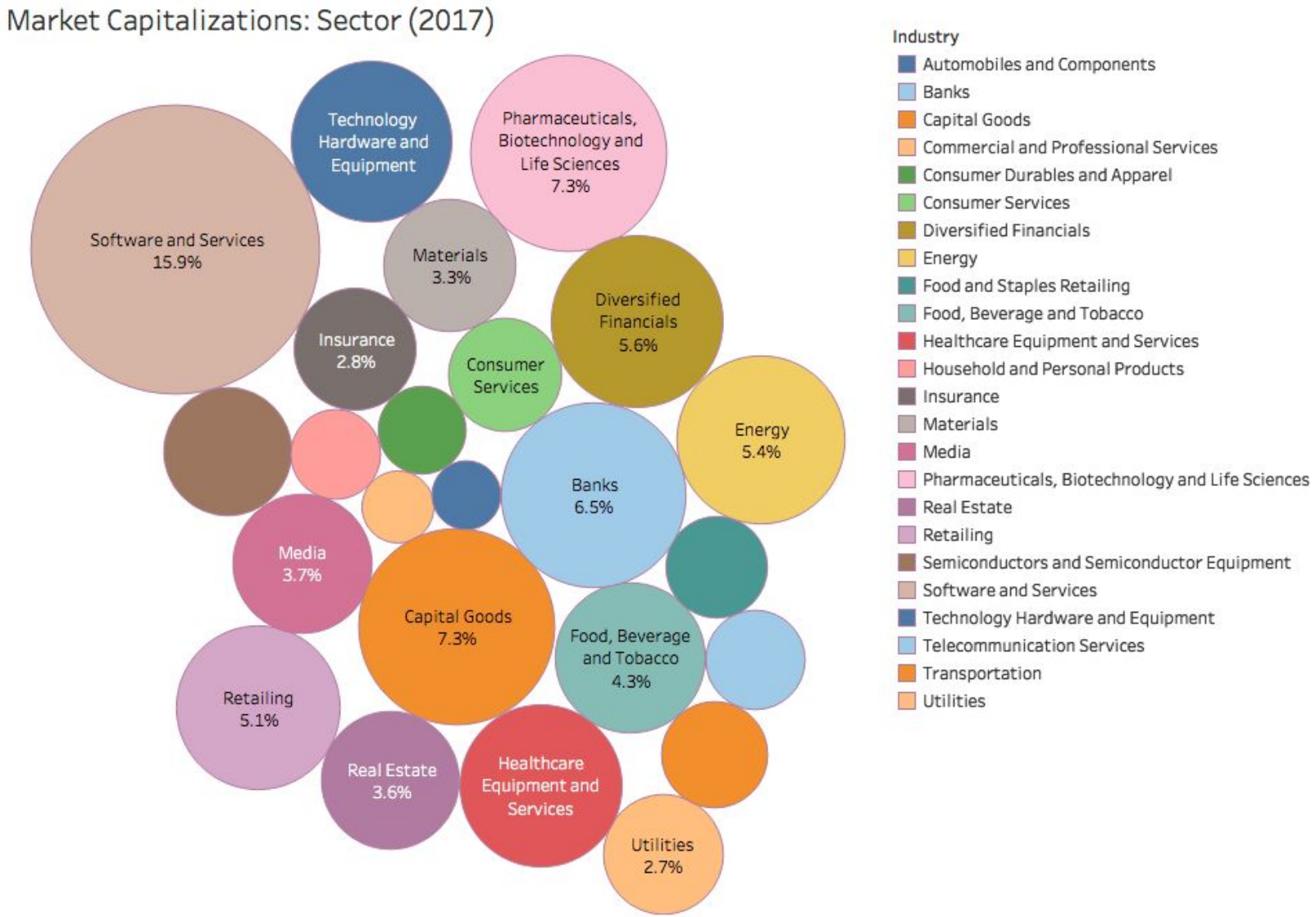
- Company headquarters location and market capitalization
- Not surprisingly, San Fran-bay area and NYC held largest spot
- Further research could find this useful over time to see if a leading indicator for real estate prices?

Data Visualization and Findings



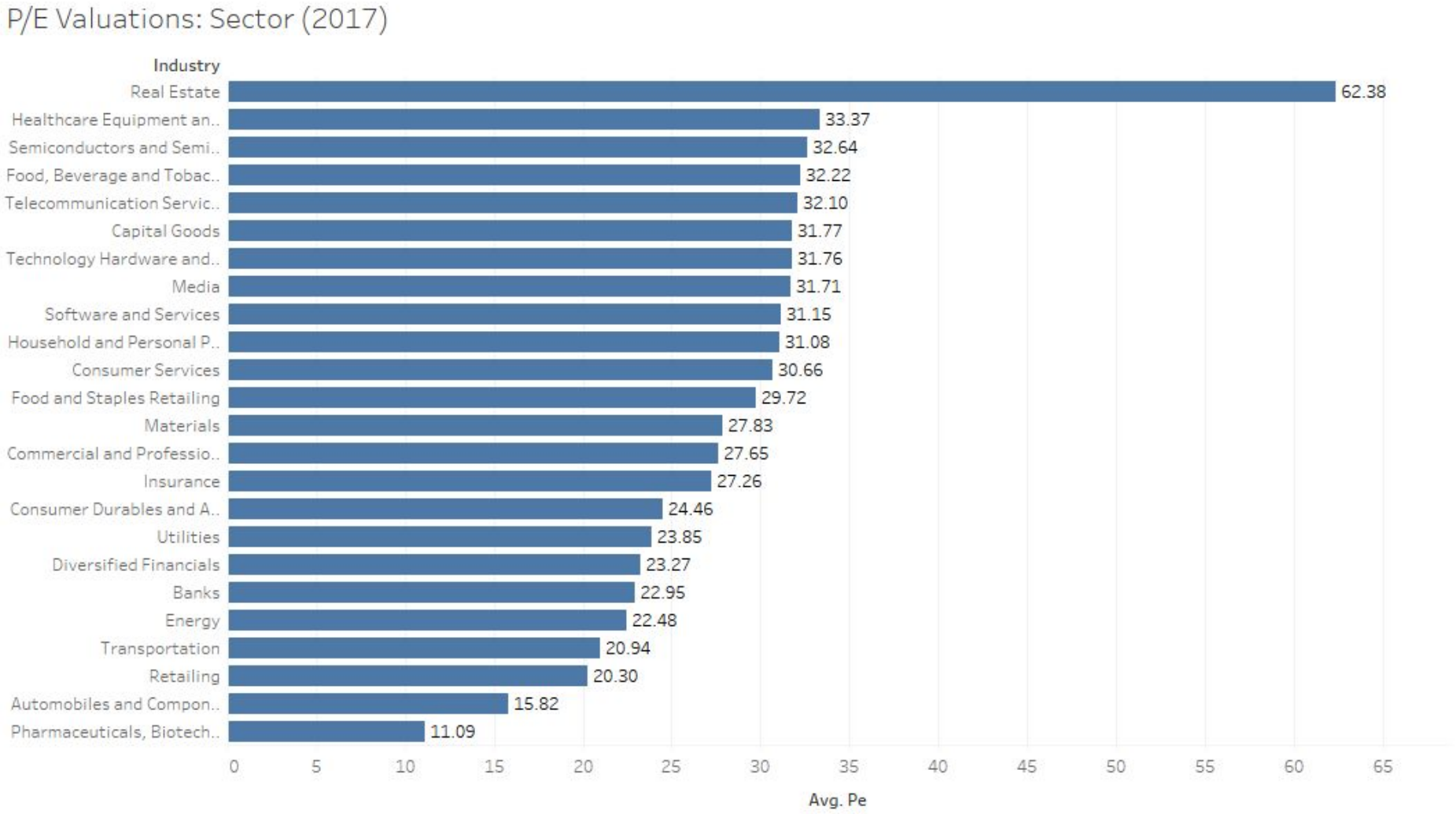
- Non-bank corporate debt listed by sector shows a relatively even dispersion
- This was fascinating since sector differences are usually so wide, depending on industry size and credit quality
- •Further research might find this interesting for debt pricing

Data Visualization and Findings



- Market capitalization tells a more familiar story: technology in the lead
- Outsizes proportions can be an interesting indicator for bubbles or overvaluation in general

Data Visualization and Findings



- However comparing price to earnings (i.e. PE ratio) shows valuations for tech are not over stretched
- It would seems real estate and capital goods could be potentially overvalued
- Real Estate might be overvalued given the secular decline in retail sector

Lesson Learned

- You might need to clean your data even if you don’t think you need to
- The nature of relationships really influences your design
- Surrogate keys are very useful
- The nature of the data influences what you can do with it
- Market capitalization is really concentrated!