

# CYNCLY TECHNICAL INTERVIEW

Dragana Trninic  
12 February 2024

# FIRST STEPS

Three datasets: xlsx, csv, json

csv & json – convert all to xlsx

json – missing quotation marks

Excel Power Query

Create Copies of all xlsx

```
C:\Users\Dragana\Downloads\DQ Analyst Interview\Dataset3_Companies_Updated_v2.json - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
change.log Dataset3_Companies_Updated_v2.json
["Technology", "Healthcare", "Finance", "Manufacturing", "Healthcare", "Retail", "N/A", "Retail", "Healthcare", "Retail", "N/A", "Tehnology", "Finance", "Technology", "Finance",
"Healthcare", "Retail", "Technology", "Healthcare", "Healthcare", "Finance", "Finance", "Finance", "Technology", "Healthcare", "Finance", "Manufacturing", "Tehnology", "Technology",
, "Healthcare", "Retail", "Finance", "Manufacturin", "N/A", "Manufacturing", "Manufacturing", "Healthcare", "Finance", "Manufacturing", "Manufacturing", "Retial", "Healthcare",
"Manufacturing", "Finance", "Retail", "Healthcare", "Technology", "Finance", "Manufacturin", "Manufacturing", "Manufacturing", "Retail", "Manufacturing", "Healthcare", "Healthcare",
, "Technology", "Finance", "Healthcare", "Retail", "Technology", "Healthcare", "Manufacturing", "Technology", "Finance", "Manufacturing", "Retail", "Retail", "Manufacturing",
"Technology", "Retail", "Healthcare", "Finance", "Retail", "Finance", "Finance", "Manufacturing", "Retail", "Technology", "Finance", "N/A", "Technology", "Finance", "Retail",
"Manufacturing", "Healthcare", "Manufacturing", "Healthcare", "Finance", "Healthcare", "Retail", "Retail", "Finance", "Manufacturing", "Technology", "N/A", "Technology", "Finance",
"Technology", "Retail", "Healthcare", "Retail", "Finance", "Technology", "Technology", "Finance", "Finance", "Finance", "Retail", "Retail", "Manufacturing", "Healthcare", "Finance",
"Finance", "Finance", "Technology", "Manufacturing", "Healthcare", "Technology", "Finance", "Healthcare", "Technology", "Manufacturing", "Technology", "Manufacturing",
"Manufacturing", "Healthcare", "Retail", "Technology", "Finance", "Healthcare", "Healthcare", "Manufacturing", "Retail", "Finance", "Healthcare", "Technology",
"Technology", "Technology", "Manufacturing", "Technology", "Technology", "Finance", "Healthcare", "Technology", "Manufacturing", "Retail", "Healthcare", "N/A", "Technology",
"Technology", "Retail", "Manufacturing", "Finance", "Finance", "Finance", "Technology", "Technology", "Manufacturing", "Healthcare", "Finance", "Retail", "Finance", "Healthcare",
"Technology", "Technology", "Healthcare", "Technology", "Finance", "Technology", "Healthcare", "Finance", "Healthcare", "Manufacturing", "Healthcare", "Healthcare", "Finance",
"Technology", "Retail", "Finance", "Technology", "Technology", "Healthcare", "Retail", "Healthcare", "Finance", "Finance", "Retail", "Manufacturing", "Technology", "Finance", "Technology",
"Retail", "Healthcare", "Manufacturing", "Finance", "Healthcare", "Healthcare", "Retail", "Healthcare", "Finance", "Healthcare", "Manufacturing", "Manufacturing",
"Healthcare", "Retail", "Healthcare", "Manufacturing", "Retail", "Manufacturing", "Manufacturing", "Technology", "Healthcar", "Healthcare", "Retail", "Finance", "Manufacturing",
"Retail", "Finance", "Manufacturing", "Technology", "Technology", "Technology", "Retail", "Retail", "Finance", "Finance", "Healthcare", "Retail", "Finance", "Healthcare", "Technology",
, 1552, 3076, 2015, 3637, 169, 3987, 1351, 1285, 4388, 767, 3671, 2825, 899, 4982, 3154, 108, 4314, 4314, 3834, 3680, 846, 437, 2624, 4794, 2658, 3915, 2113, 3194, 1268, 3032, 4404
, 771, 3935, 1740, 3140, 1195, 4428, 2075, 1878, 148, 3845, 2187, 3653, 158, 3378, 218, 3170, 3585, 1873, 1868, 2157, 889, 4432, 1116, 945, 2532, 1987, 2519, 4783, 1890, 519, 4124,
4738, 1861, 4039, 1217, 100, 3564, 2905, 4028, "NaN", 3688, 174, 1197, 4475, 3048, 3326, 1534, 1407, 4363, 3413, 1776, 1421, 2090, 528, 3000, 1202, 4270, 4125, 3537, 2482, 3259,
4436, 2643, 3127, 2826, 908, 1672, 4165, 4854, 1986, 497, 4571, 2719, 1222, 770, "NaN", 4863, 1043, 2546, 3532, 1568, 3956, 3366, 4603, 3939, 4934, 2827, 2598, 1373, 4050, 2731,
798, 1544, 3119, 1301, 3050, 4320, 2704, 4026, 1147, 4063, 4819, 2051, 4325, 668, 2784, 1846, 3214, 256, 2097, 2875, 4106, 2905, 2428, 2452, 3511, 1958, 2511, 1237, 3192, 469, 2258
, 3044, 506, 476, 2224, "NaN", 4675, 2406, 2133, 3087, 2911, 2851, 2736, 810, 3345, 4293, 3002, 3413, 1601, 1706, 3600, 1671, 2516, 666, 3711, 2838, 1404, 3887, 3147, 3325, 3190,
79, 4673, 2097, 3182, 4983, 76, 4648, 454, 3467, 4724, 407, 4569, 4038, 396, 362, 1845, 635, 2998, 2111, 4644, 590, 1164, 272, 1453, 3302, 2333, 630, 106, 1120, 3279, 2697, 1009,
4804, 2885, 92, 3787, 3500, 166, 3464, 332, 1390, 736, 3844, 4222, 2640, 2051, 258, 498, 1888, 3353, 1061, 3511, 1008, 1535, 3839, 1300, 548, 1663, 3866, "NaN", 4712, 93, 4256,
4190, 4081, 2482, 4667, 3680, 2232, 3453, 240, 4983, 4103, "NaN", 3595, 1545, 2778, 2545, 683, 1001, 2072, 3985, 4653, 3976, 4993, 3711, 4300, 1006, 2234, 3511, 1182, 1304, 1882,
209, 1546, "NaN", 2738, 851, 4642, 3067, 375, 4039, 3665, 4842, 2126, 3149, 1267, 2162, 3706, 2782, 2746, 448, 4976, 364, 3971, 3105, 181, 3503, "NaN", 429, 1369, 1722, 1873, 931,
4958, 241, 1865, 3412, 4445, 1278, 1258, 1212, 153, 138, 3311, 3473, 2889, 2366, 238, 2127, 1522, 291, 4439, 4527, 1231, 2854, 4972, 4936, 3109, 4943, 1263, 155, "NaN", 4960, 1231,
459, 4327, 4703, 427, 1931, 2356, 571, 378, 258, 155, 832, 4863, 1368, 1364, 1841, 558, 1902, 2837, 760, 90, 3498, 1783, 308, 786, 4719, 173, 589, 2714, 4609, 1313, 2722, 658,
2330, 3089, 2903, 1171, "NaN", 1196, 4093, 3826, 1543, 4163, 2428, 287, 1615, 211, 1794, 4998, 577, 910, 3316, 2953, 1178, 3947, 1377, 2881, 1454, 354, 3178, 582, 4368, 4569, 919,
990, 1968, 2735, 4297, 4136, 4148, 441, 4728, 2857, 3305, 3613, 1568, 134, 396, 3046, 94, 1846, 3374, 1190, 2980, "NaN", 3730, 4696, 1063, 2130, 2318, 4317, 1907, 4360, 915, 2168,
"NaN", 3995, 4027, 481, 4235, 1204, 3143, 1731, 3377, 1815, 3449, 3594, 4063, 4876, 3730, 917, 1159, 1755, 3606, 4313, 610, 1406, 4789, 4963, 615, 1412, 2424, 3409, 2180, 2507,
4138, 3212, 1458, 3068, 2250, 2534, 362, 4576, 1217, 4281, 1430, 4400, 3379, 2224, 1819, 4625, 3757, 282, 3778, 2762, 2335, 2303, 1154, 2715, 3914, 2612, 2408, 3285, 3399, 3385,
1232, 3386, 3737, 1397, 2697, 90, 2184, 1866, 2441, 4133, 4526, 4538, 4798, 4871, 2546, 645, 1422, 1854, 3872, 3170, 570, 3260, 3556, 4963, 3748, 4832, 188, 2191, 2595, 2795, 2556,
1655, 3751, 780, 1626, 2623, 4742, 4131, 245, 1062, 1659, 3909, 2826, 2179, 3308, 702, 733, 4025, 4419, 1391, 3143, 1171, 3500, 3267, 2957, 2814, 3681, 1896, 1131, 1436, 385, 1432
, 3662, 4955, 2496, 687, 674, 276, 2136, 1422, 3706, 272, 522, 1987, 917, 3849, 2250, 365, 1588, 1893, 4815, 2592, 1211, 3273, 3450, 1536, 667, 4801, 3806, 3100, 1890, 4133, 1195,
4330, 777, 663, 464, 2139, 4652, "NaN", 1394, 3744, 4583, 2337, "NaN", 2864, 1496, 183, 3981, 4557, 3515, 4896, 2406, 1493, 4931, 4184, 2340, 153, 711, 3387, 2706, 89, 267, 2368,
2660, 1121, 1465, 497, 2098, 4564, 1876, 1402, 4136, 2550, 4176, 4532, 2353, 982, 3042, 3272, 2777, 4026, 1623, 4165, 84, 3797, 696, 644, 118, 895, 1411, 4126, 2521, 4487, 2934,
3666, 1129, 2193, 1783, 341, 1807, 4099, 3778, 4146, 1114, 1227, 3375, 2975, 4458, 3004, 2925, 2974, 2627, 2880, 3226, 4229, 120, 4844, 3000, 3635, 2624, "NaN", 260, 104, 2343,
3705, 1678, 1112, 4796, 3870, 4628, 822, 4875, 1676, 1710, 1339, 2166, 3016, 1740, 517, 1422, 1212, 1036, 3021, 666, 4185, 2247, 2291, 2945, 3193, 1989, 2769, 2287, 3957, 175, 2310
, 2528, 1011, 1169, 1876, 3764, 97, 4828, 1335, 1694, 2239, 3507, 2678, 2987, 2045, 414, 100, 1192, 3527, 3924, 4831, 839, 1808, 1605, 3352, 329, 4521, 1287, 4200, 3036, 2520
, 2360, 2234, 2816, 3639, 1325, 2471, 3922, 4620, 2633, 3552, 243, 4790, 4040, 1597, 298, 377, 4717, 530, 738, 2185, 4620, 464, 845, 1081, 1950, 3936, 3582, 1454, 2210, 4255, 3709,
826, 3056, 4387, 1794, 4178, 2341, 1923, 4874, 1034, 156, 3357, 1209, 1809, 3251, 1141, 2488, 1016, 937, 3401, 1525, 3533, 2572, 4242, 3948, 1449, 909, 3925, 4401, 2090, 2954,
```

# GETTING TO KNOW THE DATA

Dataset 1 Columns	Dataset 2 Columns	Dataset 3 Columns
Company Name	Name of Company	Corporation
Headquarters Address	Country of Headquarters	Main Office Location
Industry Sector	Industry	Industry Category
Website URL	Website	Website
Number of Employees	Employee Count	Number of Staff
Year Founded	Founding Year	Establishment Year
Revenue (USD)	Market Cap (USD)	Annual Sales (USD)
	CEO Name	Operating Income (USD)
		Number of Offices Worldwide

# QUESTIONS I WOULD ASK

1. Are the companies in this list leads, companies we work with, companies we have worked with in the past?

2. The Company Name and Website URL are missing in some cases. I would prefer to delete as they lack basic crucial information, however the following gives me pause: *“Describe a strategy to deal with missing information in critical fields like ‘Number of Employees’ or ‘Revenue (USD)’.”*

Would deleting these rows be acceptable?

3. “Given the datasets contain information about company revenues and the number of employees, can you calculate and analyze the average revenue per employee across different industries?”

Dataset 1 contains information about company revenues.

Dataset 2 contains Market Cap data which cannot be used to calculate revenue.

Dataset 3 can we assume that Annual Sales can be interchangeable with Revenue?

# CLEANING DATASETS INDIVIDUALLY

12 February 2024

# CHECKING FOR DUPLICATES WITHIN DATASETS

Excel interface showing the 'Remove Duplicates' dialog box. The dialog box is open, displaying the columns selected for duplicate removal: Company Name, Headquarters Address, Industry Sector, Website URL, Number of Employees, and Revenue. The 'My data has headers' checkbox is checked.

Columns selected for duplicate removal:

- Company Name
- Headquarters Address
- Industry Sector
- Website URL
- Number of Employees
- Revenue

Buttons: Select All, Unselect All, My data has headers, OK, Cancel.

18	Green-Palmer	Juarezborough, Finland				
19	Walsh-Joseph	Lake Scottchester, Botswana				
20	Boone, Henry and Callahan	North Annette, Kazakhstan				
21	Cunningham-Soto	Fergusonberg, Tokelau				
22	Morales, Gonzalez and Collins	Jenniferfort, Faroe Islands				
23	Mueller Groux	Laurafort, Guernsey				
24	Dyer, Wolfe and Stevens	Paulview, Luxembourg				
25	Moody-Harris	Lake Douglas, Venezuela				
26	Jones, Vaughn and Brewer	North Catherine, Heard Island and McDonald Islands				
27	Moore-Cooper	Heatherville, Israel				
28	Smith-Nelson	Ashleybury, Bulgaria				
29	Harper LLC	Hernandezberg, Anguilla				

Excel interface showing the 'Remove Duplicates' dialog box. The dialog box is open, displaying the columns selected for duplicate removal: Industry Sector, Website URL, Number of Employees, and Revenue. The 'My data has headers' checkbox is checked.

Columns selected for duplicate removal:

- Industry Sector
- Website URL
- Number of Employees
- Revenue

Buttons: Select All, Unselect All, My data has headers, OK, Cancel.

Microsoft Excel dialog box: No duplicate values found. OK

# UNIQUE FIELD — URLS

COPY\_Dataset1\_Companies\_Updated\_v2.xlsx - Excel

Company Name	Headquarters Address	Industry Sector	Website URL	Number of Employees	Year Founded	Revenue
Booker-Price	New Deanna, British Indian Ocean Territory (Chagos Archipelago)	Manufacturing	www.bp.com	4520	1943	
Martin, James and Conley	South Dianestad, Antigua and Barbuda	Manufacturing	www.martin-conley.com	2103	1979	
Martin, Moon and Jones	Wilsonfurt, Saudi Arabia	Retail	www.martin-jones.com	1259	1939	
White PLC	New Jennifer, Somalia	Healthcare	www.wp.com	889	2013	
Lee-Gjoss	New Brandon, Panama	Retail	www.lee-gjoss.com	598	1959	
Atkins-Ramirez	West Rileyshire, Anguilla	Technology	www.ar.com	2921	1963	
Kennedy LLC	Matthewfurt, Aruba	Retail	www.kennedy-llc.com	125		
Patterson-Thompson	Knappmouth, Nicaragua	Manufacturing	www.pt.com	3357	1931	
Glenn LLC	New Melissafort, Argentina	Finance	www.glenn-llc.com	4888	2006	
Powell-Vaughn	New Rachel, Israel	Technology	www.powell-vaughn.com	1630	1949	
Kim and Sons	Danielsfurt, Guernsey	Healthcare	www.kas.com	3552	1936	
Hayes, Parker and Todd	N/A	Finance	www.hayes-parker-and-todd.com	3046	1957	
Ward-Thomas	Mayshire, Lao People's Democratic Republic	Healthcare	www.ward-thomas.com	721	1959	
Campbell-Mullen	Robinside, Saint Lucia	Healthcare	www.campbell-mullen.com	3022	1992	
Fischer-Lee	N/A	Finance	www.fischer-lee.com	3124	1965	
Murphy and Sons	Sandersland, Jersey	Manufacturing	www.murphy-sons.com	3185	1907	
Green-Palmer	Juarezborough, Finland	Manufacturing	www.gp.com	3265	1955	
Walsh-Joseph	Lake Scottchester, Botswana	Healthcare	www.walsh-joseph.com	2114	1919	
Boone, Henry and Callahan	North Annette, Kazakhstan	Retail	www.boone-henry-and-callahan.com	4788		
Cunningham-Soto	Fergusonberg, Tokelau	Technology	www.cs.com	3179	1938	
Morales, Gonzalez and Collins	Jenniferfort, Faroe Islands	Retail	www.mgac.com	3652	1999	
Mueller Grou	Laurafort, Guernsey	Retail	www.mueller-grou.com	4852	1998	
Dyer, Wolfe and Stevens	Paulview, Luxembourg	Technology	www.dyer-wolfe-and-stevens.com	4023	1979	
Moody-Harris	Lake Douglas, Venezuela	Technology	www.moody-harris.com	3420	1949	
Jones, Vaughn and Brewer	North Catherine, Heard Island and McDonald Islands	Manufacturing	www.jones-vaughn-and-brewer.com	4154	2004	
Moore-Cooper	Heatherville, Israel	Manufacturing	www.mc.com	605	1901	
Smith-Nelson	Ashleybury, Bulgaria	Healthcare	www.smith-nelson.com	639	1992	
Harper LLC	Hernandezberg, Anguilla	Healthcare	www.harper-llc.com	890	1935	

Original 1 | Copy 1 | Modified for Duplicates 1

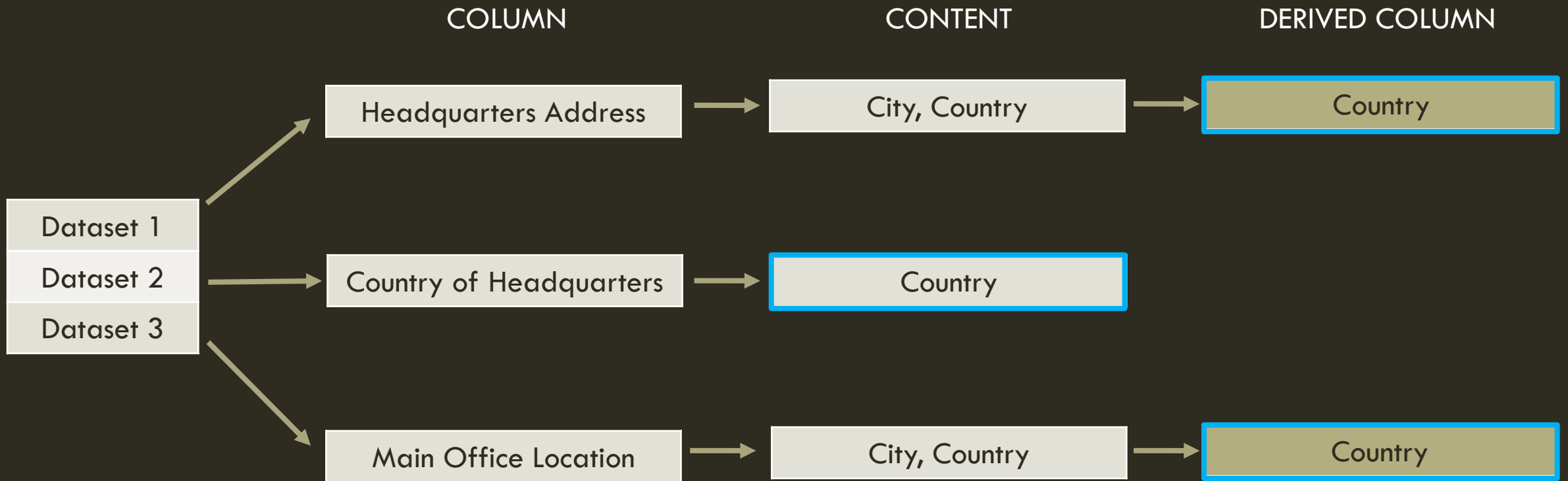
# UNIQUE FIELD — URLS

	A	B	C	D	E	F	G
1	Company Name	Headquarters Address	Industry Sector	Website URL	Number of Employees	Year Founded	Revenue (USD)
2	Booker-Price	New Deanna, British Indian Ocean Territory (Chagos Archipelago)	Manufacturing	www.bp.com	4520	1943	99623962
3	Bentley PLC	Hoffmanmouth, Congo	Finance	www.bp.com	4154	1924	21764093
4							
5							
6	Company Name	Headquarters Address	Industry Sector	Website URL	Number of Employees	Year Founded	Revenue (USD)
7	White PLC	New Jennifer, Somalia	Healthcare	www.wp.com	889	2013	7685196
8	Webster PLC	New Kevin, Norfolk Island	Retail	www.wp.com	536	1950	48921065
9							
10							
11	Company Name	Headquarters Address	Industry Sector	Website URL	Number of Employees	Year Founded	Revenue (USD)
12	N/A	Jenkinsport, Cayman Islands	Technology	N/A	1922	2012	87076344
13	N/A	West Patriciashire, Ethiopia	Healthcare	N/A	4962	2008	15153068
14	N/A	West Joshua, Panama	Healthcare	N/A	1870	1902	14419292
15							
16							
17							



[illegible]

# DERIVED COUNTRY COLUMN: SETS 1 & 3



# DERIVED COUNTRY COLUMN: SETS 1 & 3

C2				British Indian Ocean Territory (Chagos Archipelago)
	A	B	C	D
1	Company Name	Headquarters Address	Country of Headquarters	Industry Sector
2	Booker-Price	New Deanna, British Indian Ocean Territory (Chagos Archipelago)	British Indian Ocean Territory (Chagos Archipelago)	Manufacturing
3	Martin, James and Conley	South Dianestad, Antigua and Barbuda	Antigua and Barbuda	Manufacturing
4	Martin, Moon and Jones	Wilsonfurt, Saudi Arabia	Saudi Arabia	Retail
5	White PLC	New Jennifer, Somalia	Somalia	Healthcare
6	Lee-Gjoss	New Brandon, Panama	Panama	Retail
7	Atkins-Ramirez	West Rileyshire, Anguilla	Anguilla	Technology
8	Kennedy LLC	Matthewfurt, Aruba	Aruba	Retail
9	Patterson-Thompson	Knappmouth, Nicaragua	Nicaragua	Manufacturing
10	Glenn LLC	New Melissaft, Argentina	Argentina	Finance
11	Powell-Vaughn	New Rachel, Israel	Israel	Technology
12	Kim and Sons	Danielsfurt, Guernsey	Guernsey	Healthcare
13	Hayes, Parker and Todd	N/A	N/A	Finance
14	Ward-Thomas	Mayshire, Lao People's Democratic Republic	Lao People's Democratic Republic	Healthcare
15	Campbell-Mullen	Robinside, Saint Lucia	Saint Lucia	Healthcare
16	Fischer-Lee	N/A	N/A	Finance
17	Murphy and Sons	Sandersland, Jersey	Jersey	Manufacturing
18	Green-Palmer	Juarezborough, Finland	Finland	Manufacturing
19	Walsh-Joseph	Lake Scottchester, Botswana	Botswana	Healthcare
20	Boone, Henry and Callahan	North Annette, Kazakhstan	Kazakhstan	Retail
21	Cunningham-Soto	Fergusonberg, Tokelau	Tokelau	Technology
22	Morales, Gonzalez and Collins	Jenniferfurt, Faroe Islands	Faroe Islands	Retail

	C
	Country of Headquarters
	Zambia
	Eritrea
	Barbados
	Mauritius
	Swaziland
	Israel
	Gabon
	Turkmenistan
	Czech Republic
	Cook Islands
	Jordan
	Cote d'Ivoire
	Saint Martin
	French Polynesia
	China
	Israel
	N/A
	Taiwan

# MISSING DATA

Identifying "N/A" and "NaN" entries and missing data:

E	F	G
Number	Establishment Year	Number of Offices
4794	Sort Smallest to Largest	
2658	Sort Largest to Smallest	
3032	Sort by Color	
2333		
1522	Clear Filter From "Number of Offices..."	
155	Filter by Color	
2224	Number Filters	
3778		
2546		
188	Search	
3308	<input type="checkbox"/> 92	
2957	<input type="checkbox"/> 93	
3273	<input type="checkbox"/> 94	
4133	<input type="checkbox"/> 95	
3981	<input type="checkbox"/> 96	
1402	<input type="checkbox"/> 97	
2934	<input type="checkbox"/> 98	
3226	<input type="checkbox"/> 99	
666	<input type="checkbox"/> 100	
4011	<input checked="" type="checkbox"/> NaN	
3764		
3900	1988 NaN	
1094	1980 NaN	
3236	1910 NaN	
2900	1984 NaN	
NaN	1930 NaN	
4419	NaN NaN	
4586	1983 NaN	
202	1925 NaN	
3588	2021 NaN	

Founding Year	Market Cap (USD)	CEO Name
4068	Sort Smallest to Largest	Barbara Allen
4680	Sort Largest to Smallest	Jordan Ramos
322	Sort by Color	Michael Brooks
743		Diane Hoffman
3136	Clear Filter From "Market Cap (USD)"	Jordan Wallace
4454	Filter by Color	Luke Edwards
3710	Number Filters	Barbara Davis
516		Matthew Russell
4667	Search	James Savage
3982	<input type="checkbox"/> \$499,244,384	Robert Stevens
2613	<input type="checkbox"/> \$499,393,815	Amanda Thomas
2040	<input type="checkbox"/> \$499,409,706	Meghan Jimenez
1173	<input type="checkbox"/> \$499,448,085	Andrew Juarez
3659	<input type="checkbox"/> \$499,495,892	Rachel Carroll
1608	<input type="checkbox"/> \$499,538,362	Haley Camacho
1893	<input type="checkbox"/> \$499,548,746	Krystal Shaw
4825	<input type="checkbox"/> \$499,573,449	Michelle Rivas
763	<input type="checkbox"/> \$499,942,578	Trevor Knight
2191	<input checked="" type="checkbox"/> (Blanks)	Chris Kennedy
2191		Patrick Leblanc
1768		Shannon Hernandez
246	1978	Gwendolyn Martinez
235	1948	Mr. Edward Burns
3083	1973	Michael Stewart
2350	1937	Briana Joseph
2063	1953	Erika Cox
2471	2013	N/A
1976	1905	Glen Burns
1431	1974	Dana Edwards
3761	1968	Robert Thomas
2273	1934	Shawn Dominguez
4546	2002	Joshua Brown
3231	1996	Charles Page
3825	1989	Michael Garcia
2963	2021	Antonio Johnson
1918	2008	Elizabeth Cline
4422	1971	Joshua Hall
4278	1958	Michael Rich
3475	1902	Karen Gibson

F	G	H
Founding Year	Market Cap (USD)	CEO Name
1973	\$	Sort A to Z
1940	\$	Sort Z to A
2016	\$	Sort by Color
1901	\$	Clear Filter From "CEO Name"
2019	\$	Filter by Color
1980	\$	Text Filters
1927	\$	N/A
2015	\$	(Select All Search Results)
1941	\$	Add current selection to filter
2009	\$	N/A
1903	\$	
2012	\$	
1953	\$	
1947	\$	
1952	\$	
1922	\$	
1982	\$	
1938	\$	
1947	\$	
2000	\$	
1926	\$	
1989	\$	308,761,045 N/A
1913	\$	208,805,422 N/A
1978	\$	356,191,258 N/A
1919	\$	273,649,512 N/A
1992	\$	409,272,774 N/A
2002	\$	452,338,857 N/A
1961	\$	472,963,171 N/A
1933	\$	142,993,998 N/A
2013		N/A
2007	\$	245,876,656 N/A
2015	\$	110,923,071 N/A
1904	\$	226,884,145 N/A
2010	\$	8,673,500 N/A
1911	\$	243,269,253 N/A
1951	\$	454,201,726 N/A
1956	\$	185,093,079 N/A
	\$	12,641,556 N/A
1906	\$	484,252,347 N/A

# HANDLING MISSING DATA

Numerical: “NaN” (Set 3) to  
NULL i.e. Blank Values.

(Ctrl H.

Find NaN.

Replace with:

Replace All)

Text: “N/A” NULL i.e. Blank  
Values for all DataSets.

Industry Category	Number of Staff	Establishment Year	Annual Sales (USD)	Operating Income (USD)	Number of Offices Worldwide
Finance	332	1902	\$ 352,813,468	\$ 36,453,886	4
Finance	1390	1954	\$ 383,612,685	\$ 93,758,369	54
Healthcare	736	1947	\$ 316,349,719	\$ 72,783,186	45
Technology	3844	1917	\$ 167,002,341	\$ 48,203,051	20
Retail	4222	1914	\$ 392,926,029	\$ 9,067,364	3
Finance	2640	1943	\$ 235,664,603	\$ 92,990,841	7
Healthcare	2051	1912	\$ 411,117,643	\$ 33,858,246	45
Manufacturing	258	1973	\$ 167,348,747	\$ 35,985,930	54
Healthcare	498	1999	\$ 435,802,420	\$ 21,409,284	19
Technology	1888	2009	\$ 391,796,435	\$ 705,428	44
Manufacturing	3353	2012	\$ 1,034,877	\$ 90,021,101	32
Finance	1061	1972	\$ 21,126,846	\$ 28,856,853	77
Retail	3511	1920	\$ 436,398,486	\$ 56,116,597	1
Technology	1008	1908	\$ 76,559,441	\$ 18,011,632	88
Technology	1535	1922	\$ 17,618,744	\$ 74,289,922	57
Technology	3839	1936	\$ 5,146,794	\$ 48,674,662	83
Technology	1300	1955	\$ 349,310,925	\$ 21,838,355	62
Retail	548	1917	\$ 321,054,011	\$ 63,081,881	74
Healthcare	1663	1951	\$ 312,579,205	\$ 35,710,567	52
Retail	3866	1983	\$ 219,269,045	\$ 12,900,847	96
Manufacturing		2015			
Healthcare	4712	1921			
Healthcare	93	NaN	\$		
Technology	4256	1993	\$		
Finance	4190	1991	\$		
Finance	4081	1955	\$		
Healthcare	2482	1964	\$		
Manufacturing	4667	1951	\$		
Technology	3680	1950	\$		
Finance	2232	1944	\$		
Finance	3453	2014	\$ 353,348,828	\$ 34,831,710	28
Finance	240	1942	\$ 321,157,828	\$ 10,589,335	45
Retail	4983	1936	\$ 350,379,968	\$ 23,897,713	67
Technology	4103	1906	\$ 70,527,375	\$ 86,150,222	26
Healthcare	NaN	1980	\$ 43,079,431	\$ 73,937,789	38
Technology	3595	2011	\$ 71,873,651	\$ 33,369,396	72
Manufacturing	1545	1950	\$ 260,708,152	\$ 69,675,478	84
Healthcare	2778	1955	\$ 478,374,012	\$ 16,187,110	60
Technology	2545	1900	\$ 27,348,743	\$ 52,719,180	28
Retail	683	1938	\$ 118,412,469	\$ 93,758,582	100

Find and Replace

Find

Replace

Find what: NaN

Replace with:

Options >>

Replace All

Replace

Find All

Find Next

Close

# PATTERNS IN MISSING DATA?

All Datasets:

- When Company Name N/A, Website URL N/A (vice-versa).

- No obvious pattern:

Country

Industry

Founding Year

etc.

	A	B	C	D	E	F	G	H
1	Name of Company	Industry	Website	Country of Headquarters	Employee Count	Founding Year	Market Cap (USD)	CEO Name
2	N/A				3884	2021	\$ 239,962,681	Debra Marshall
134	N/A				217	2020	\$ 252,242,144	Cassandra Moore
139	N/A				456	2020	\$ 432,866,078	William Greene
200	N/A				2333	2020	\$ 421,606,048	Anthony Grant
201	N/A				2384	2020	\$ 499,448,085	Jason Davis
257	N/A	Healthcare	N/A	Slovenia	3662	2017	\$ 233,186,587	Joshua Rich
318	N/A	Retail	N/A	Saint Barthelemy	3094	2016	\$ 337,072,259	David Holland
359	N/A	Manufacturing	N/A	Anguilla	3325	2015	\$ 318,690,173	Jordan Garcia
367	N/A	Manufacturing	N/A	Equatorial Guinea	2014	2014	\$ 353,896,894	Thomas Morgan
379	N/A	Retail	N/A	Afghanistan	317	2012	\$ 261,308,568	Ryan Berger
394	N/A			Solomon Islands	450	2012	\$ 209,832,228	Brady Nichols
522	N/A			Serbia	3156	2012	\$ 313,748,347	Pamela Wade
593	N/A			Guam	3203	2011	\$ 1,414,674	Heidi Williamson
657	N/A	Manufacturing	N/A	Mozambique	267	2010	\$ 88,297,137	Steven Petersen
741	N/A						\$ 63,041,817	Carmen Reid MD
815	N/A						\$ 147,514,905	Michael Williams
822	N/A	Manufacturing	N/A	Reunion	1153	2008	\$ 22,551,468	Bryan Holmes
963	N/A			Lebanon	1357	2006	\$ 107,163,759	Douglas Webster
995	N/A			Guinea-Bissau	2020	2003	\$ 297,427,526	Mary Dunn
1012	N/A			Solomon Islands	1323	2001	\$ 370,806,590	Brett Valdez
1069	N/A			Austria	4251	2001	\$ 232,686,769	Gabrielle Farrell
1102	N/A			Sao Tome and Principe	4172	1998	\$ 257,108,380	Cynthia Hoover
1132	N/A	Healthcare	N/A	Uganda	4880	1997	\$ 381,558,788	Chad Davila
1195	N/A			Sierra Leone	494	1995	\$ 195,847,177	Carla Salas
1220	N/A			Ghana	4118	1995	\$ 254,405,432	Robert Glenn
1225	N/A	Healthcare	N/A	Lithuania	128	1994	\$ 146,878,751	Danielle Rodriguez
1239	N/A			Cook Islands	2548	1994	\$ 312,768,514	Brooke Richard
1287	N/A			Solomon Islands	4547	1993	\$ 355,141,364	Andre Scott
1379	N/A			Saint Barthelemy	2697	1992	\$ 426,202,920	Kimberly Morris
1448	N/A			United Kingdom	1181	1990	\$ 348,290,614	Bradley Yoder
1497	N/A	Finance	N/A	Solomon Islands	3994	1990	\$ 115,712,730	William Sanders
1582	N/A	Manufacturing	N/A	Albania	4436	1989	\$ 428,782,632	Rachel Dominguez
1631	N/A	Retail	N/A	Lesotho	1854	1988	\$ 53,212,328	Robert Tyler
1693	N/A	Technology	N/A	Namibia	4632	1988	\$ 302,089,425	Patrick Anderson
1711	N/A	Manufacturing	N/A	Ghana	1244	1987	\$ 43,273,981	Joshua Vaughan
1774	N/A	Retail	N/A	Jamaica	691	1984	\$ 129,094,328	John Hays
1784	N/A	Healthcare	N/A	Sierra Leone	772	1984	\$ 247,820,269	Kristi Hartman
1967	N/A	Finance	N/A	Romania	2629	1984	\$ 236,514,200	Kimberly Morse
1987	N/A	Retail	N/A	Iceland	3677	1984	\$ 164,985,258	Michael Haney

# MISSING DATA IN CRUCIAL FIELDS

Strategy to deal with missing information in critical fields like 'Number of Employees' or 'Revenue (USD)'.

## ACCEPTING MISSING VALUES

Dataset is not overly large.\*

Data seems to be missing at random.

Conservative.

## WHY NOT DELETE?

May be the only information missing.

Need to answer which data do we consider most critical?

## WHY NOT IMPUTE?

Does not make sense for this use case.

Avoids potential confusion and error in the future.

# HANDLING INCONSISTENCIES

In case of different formats (not so much an issue here): e.g. `fx = PROPER(C2)`

In case of typos (Industries) – Slide 8

In case of typos (company names) – Harder to determine. Will explore more when merging.

(Ensuring data quality before merging)



# MERGING THE DATASETS

Moving on to SQL

# STANDARDIZING COLUMNS BETWEEN DATASETS BEFORE MERGING

Dataset 1 Columns	Dataset 2 Columns	Dataset 3 Columns
Company Name	Company Name	Company Name
Country of Headquarters	Country of Headquarters	Country of Headquarters
Industry	Industry	Industry
Website	Website	Website
Number of Employees	Number of Employees	Number of Employees
Year Founded	Year Founded	Year Founded
Revenue (USD)	Market Cap (USD)	Annual Sales (USD)
	CEO Name	Operating Income (USD)
		Number of Offices Worldwide

# MERGING USING POWER QUERY & FUZZY MATCHING

Merged in PowerQuery, unfortunately no Fuzzy Matching option.

Upload data into Microsoft SQL Server to do merging and analysis there.

Fuzzy Matching – Couldn't manage as no key column(s)

# UNION

All three datasets with NULL  
data = 18,975 rows.\*

All three datasets only with  
rows where Company Name IS  
NOT NULL = 18,597\*\*

(i.e. 378 rows without Company  
Name or Website or about 2%  
of the data.)

```
SELECT
    [Company Name]
    ,[Country of Headquarters]
    ,[Industry]
    ,[Website]
    ,[Number of Employees]
    ,[Year Founded]

FROM [dbo].['Modified for Merging 1A$']
WHERE [Company Name] IS NOT NULL

UNION

...

FROM [dbo].['Modified for Merging 3A$']
WHERE [Company Name] IS NOT NULL
ORDER BY [Company Name];
```

DATA ANALYSIS

# DATA ANALYSIS QUESTIONS

1. How many unique companies are present in the merged dataset?

18,597

# DATA ANALYSIS QUESTIONS

2. Which industry sector has the highest representation in the dataset?

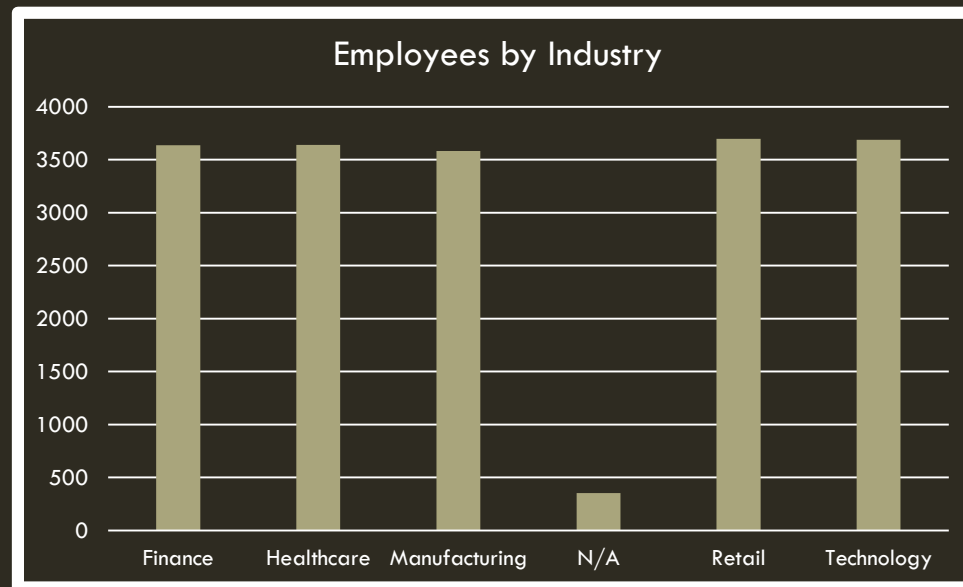
Retail - 3,697

Row Labels	Count of Industry
Finance	3637
Healthcare	3640
Manufacturing	3581
N/A	353
<b>Retail</b>	<b>3697</b>
Technology	3689

# DATA ANALYSIS QUESTIONS

3. Can you identify any trends in the number of employees relative to the industry sector?

Row Labels	Sum of Number of Employees
Finance	8890165
Healthcare	9151254
Manufacturing	8693896
N/A	905476
Retail	9057566
Technology	9067927
Grand Total	45766284





# LOOKING AT REVENUE

COLUMN



4. Given the datasets contain information about company revenues and the number of employees, can you calculate and analyze the average revenue per employee across different industries?

# ASSUMING: ALL THREE VALUES CORRESPOND TO REVENUE

Revenue (USD)

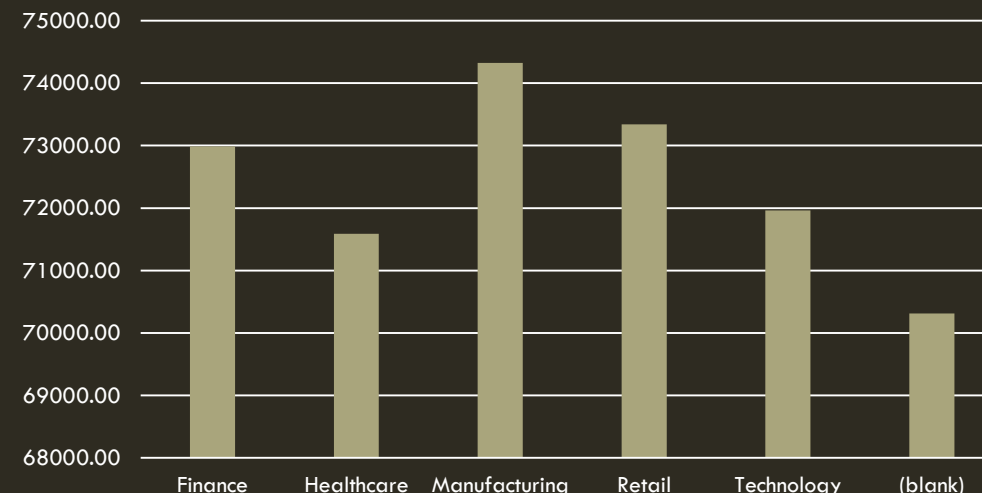
Market Cap (USD)

Annual Sales (USD)

Row Labels	Average Revenue per Employee (USD)
Finance	72988.34
Healthcare	71586.25
Manufacturing	74325.38
Retail	73338.64
Technology	71959.57
(blank)	70310.45

4. Given the datasets contain information about company revenues and the number of employees, can you calculate and analyze the average revenue per employee across different industries?

Average Revenue per Employee by Industry (USD)



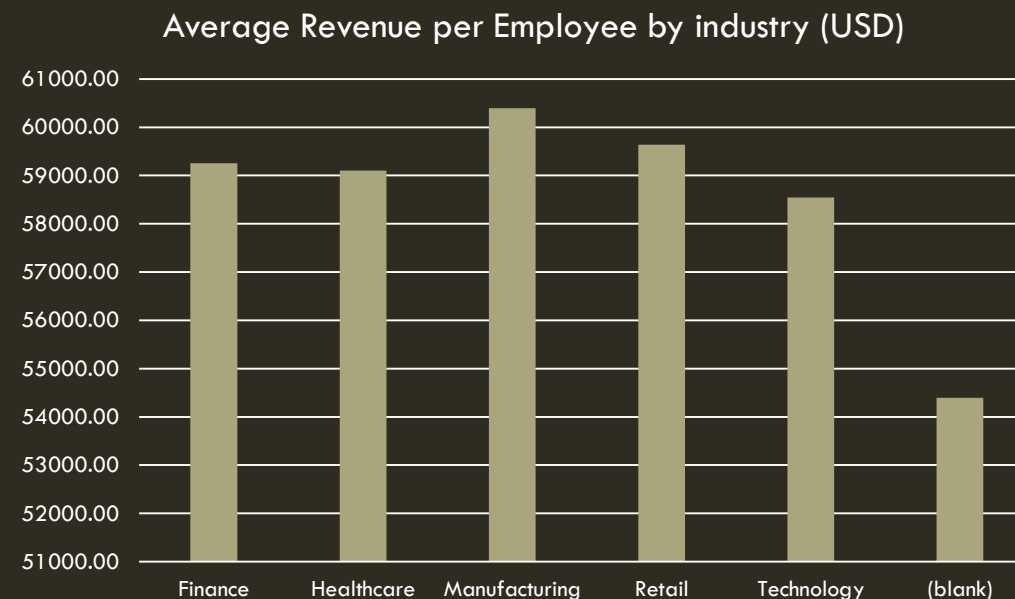
# ASSUMING: MARKET CAP CANNOT CORRESPOND TO REVENUE

Revenue (USD)

Annual Sales (USD)

Row Labels	Average Revenue per Employee (USD)
Finance	59252.06
Healthcare	59103.26
Manufacturing	60394.22
Retail	59639.25
Technology	58542.74
(blank)	54396.69

4. Given the datasets contain information about company revenues and the number of employees, can you calculate and analyze the average revenue per employee across different industries?

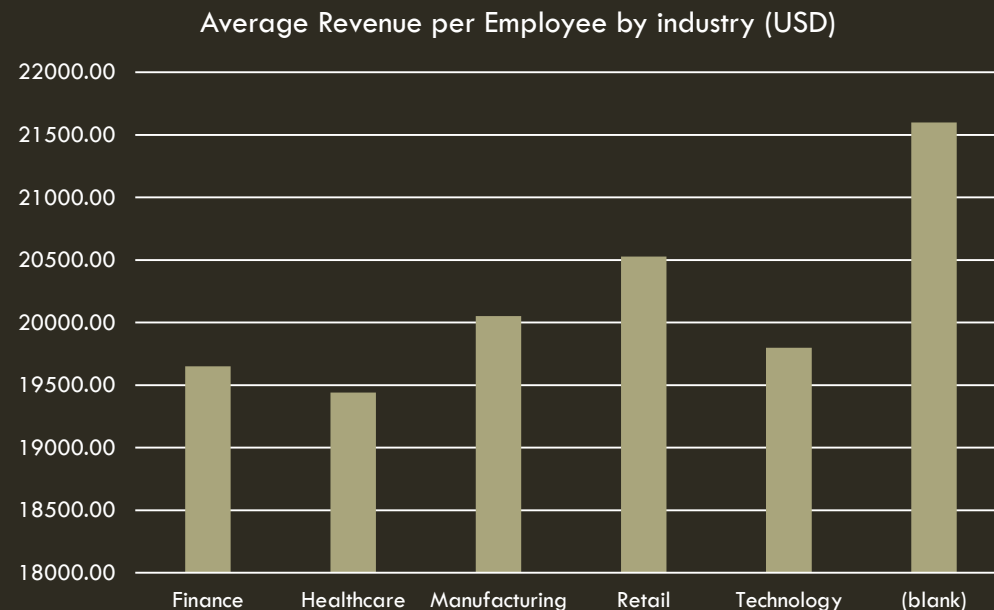


# ASSUMING: ONLY REVENUE (USD) IN DATASET 1 IS REVENUE

Revenue (USD)

Row Labels	Average Revenue per Employee (USD)
Finance	19651.01
Healthcare	19441.32
Manufacturing	20052.12
Retail	20527.81
Technology	19799.32
(blank)	21598.67

4. Given the datasets contain information about company revenues and the number of employees, can you calculate and analyze the average revenue per employee across different industries?



# DATA INSIGHTS

Almost there 😊

# DATA INSIGHTS

1. Identify the top 5 companies by number of employees.
2. Analyze the distribution of company foundation years. Are there any particular decades that saw a boom in company formations?
3. Based on the headquarters location, which country has the highest number of companies?
4. Considering the 'Year Founded' information, can you identify any correlation between a company's age (how long it has been in business) and its reported revenue or number of employees? - linear regression in Excel

# TOP FIVE COMPANIES BY NUMBER OF EMPLOYEES

1. Identify the top 5 companies by number of employees.

1	Company Name	Country of Headquarters	Industry	Website	Number of Employees	Year Founded
2	Blackburn-Diaz	Liechtenstein	Manufacturing	www.blackburn-diaz.com	5000	1925
3	Garner-Chambers	Bermuda	Retail	www.garner-chambers.com	5000	1952
4	Kim, Mitchell and Gonzalez	Pakistan	Healthcare	www.kim-gonzalez.com	5000	1979
5	Berry, Mcknight and Ferguson	N/A	N/A	www.berry-ferguson.com	4999	1953
6	Bradley, Daniels and Meadows	Bermuda	Technology	www.bradley-daniels-and-meadows.com	4999	1908

```
SELECT TOP (5) [Company Name]
, [Country of Headquarters]
, [Industry]
, [Website]
, [Number of Employees]
, [Year Founded]
FROM [Cyncly].[dbo].[All_Datasets]
ORDER BY [Number of Employees] DESC, [Company Name]
```

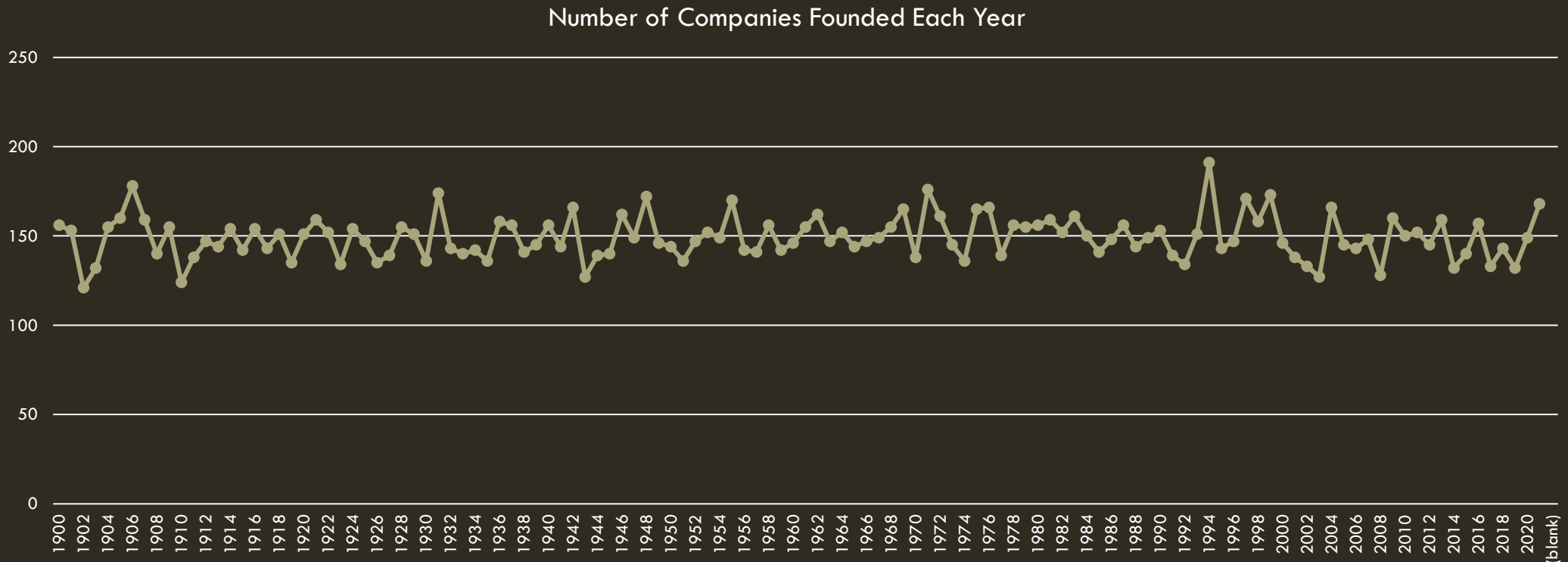
110 %

Results Messages

	Company Name	Country of Headquarters	Industry	Website	Number of Employees	Year Founded
1	Blackburn-Diaz	Liechtenstein	Manufacturing	www.blackburn-diaz.com	5000	1925
2	Garner-Chambers	Bermuda	Retail	www.garner-chambers.com	5000	1952
3	Kim, Mitchell and Gonzalez	Pakistan	Healthcare	www.kim-gonzalez.com	5000	1979
4	Berry, Mcknight and Ferguson	N/A	N/A	www.berry-ferguson.com	4999	1953
5	Bradley, Daniels and Meadows	Bermuda	Technology	www.bradley-daniels-and-meadows.com	4999	1908

# COMPANY FOUNDATIONS THROUGH THE YEARS

*1906 and 1994 saw a boom compared to the surrounding years (with 178 and 191 companies founded respectively).*

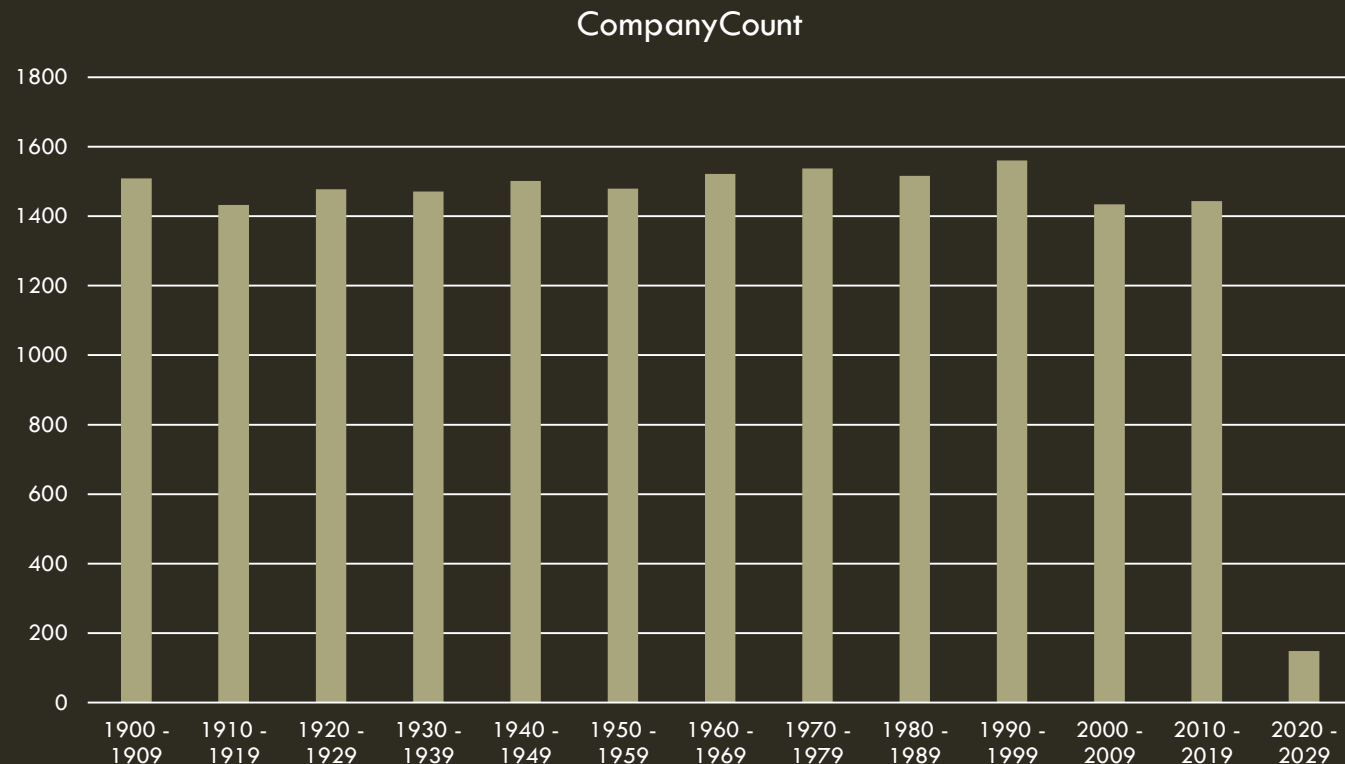




# COMPANY FOUNDATIONS THROUGH THE DECADES

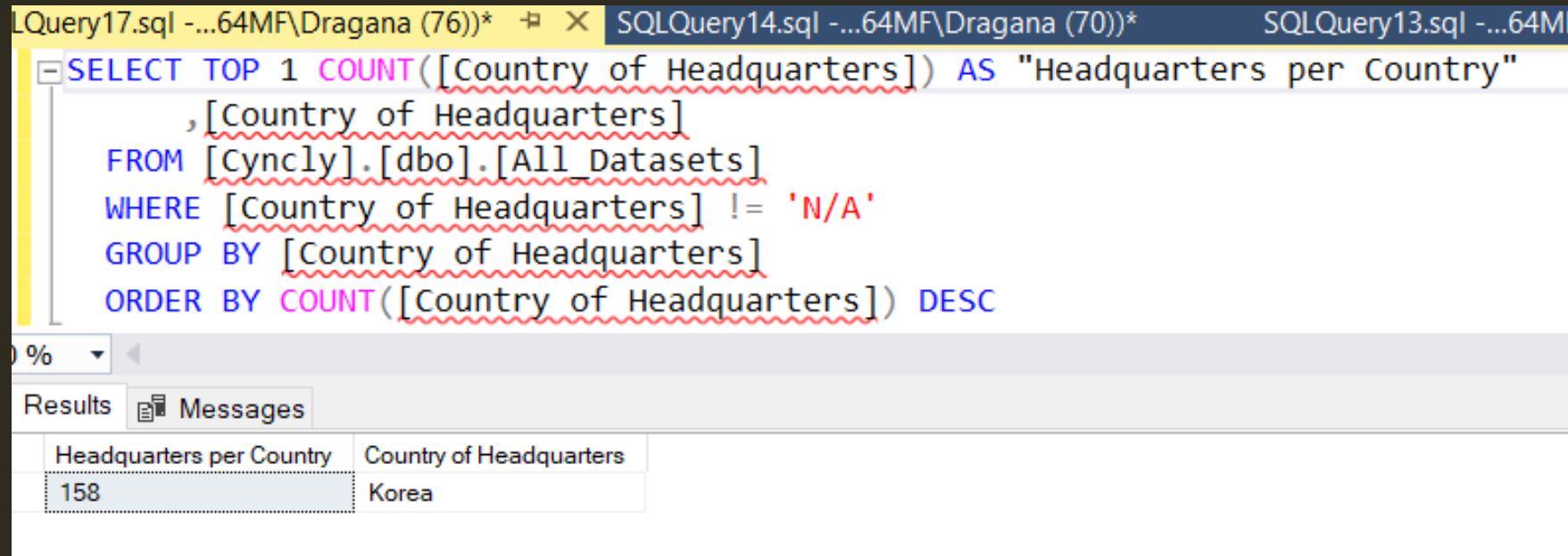
2. Analyze the distribution of company foundation years. Are there any particular decades that saw a boom in company formations?

Decade	Company Count
1900 - 1909	1509
1910 - 1919	1432
1920 - 1929	1477
1930 - 1939	1471
1940 - 1949	1501
1950 - 1959	1479
1960 - 1969	1522
1970 - 1979	1537
1980 - 1989	1516
1990 - 1999	1560
2000 - 2009	1434
2010 - 2019	1443
2020 - 2029	149



# DATA INSIGHTS

3. Based on the headquarters location, which country has the highest number of companies?



The screenshot displays a SQL Server Enterprise Manager window with three tabs: 'SQLQuery17.sql - ...64MF\Dragana (76))\*', 'SQLQuery14.sql - ...64MF\Dragana (70))\*', and 'SQLQuery13.sql - ...64M'. The active tab shows a SQL query designed to find the country with the most headquarters. The query is as follows:

```
SELECT TOP 1 COUNT([Country of Headquarters]) AS "Headquarters per Country",  
[Country of Headquarters]  
FROM [Cyncly].[dbo].[All_Datasets]  
WHERE [Country of Headquarters] != 'N/A'  
GROUP BY [Country of Headquarters]  
ORDER BY COUNT([Country of Headquarters]) DESC
```

Below the query editor, the 'Results' tab is selected, showing a single row of data:

Headquarters per Country	Country of Headquarters
158	Korea

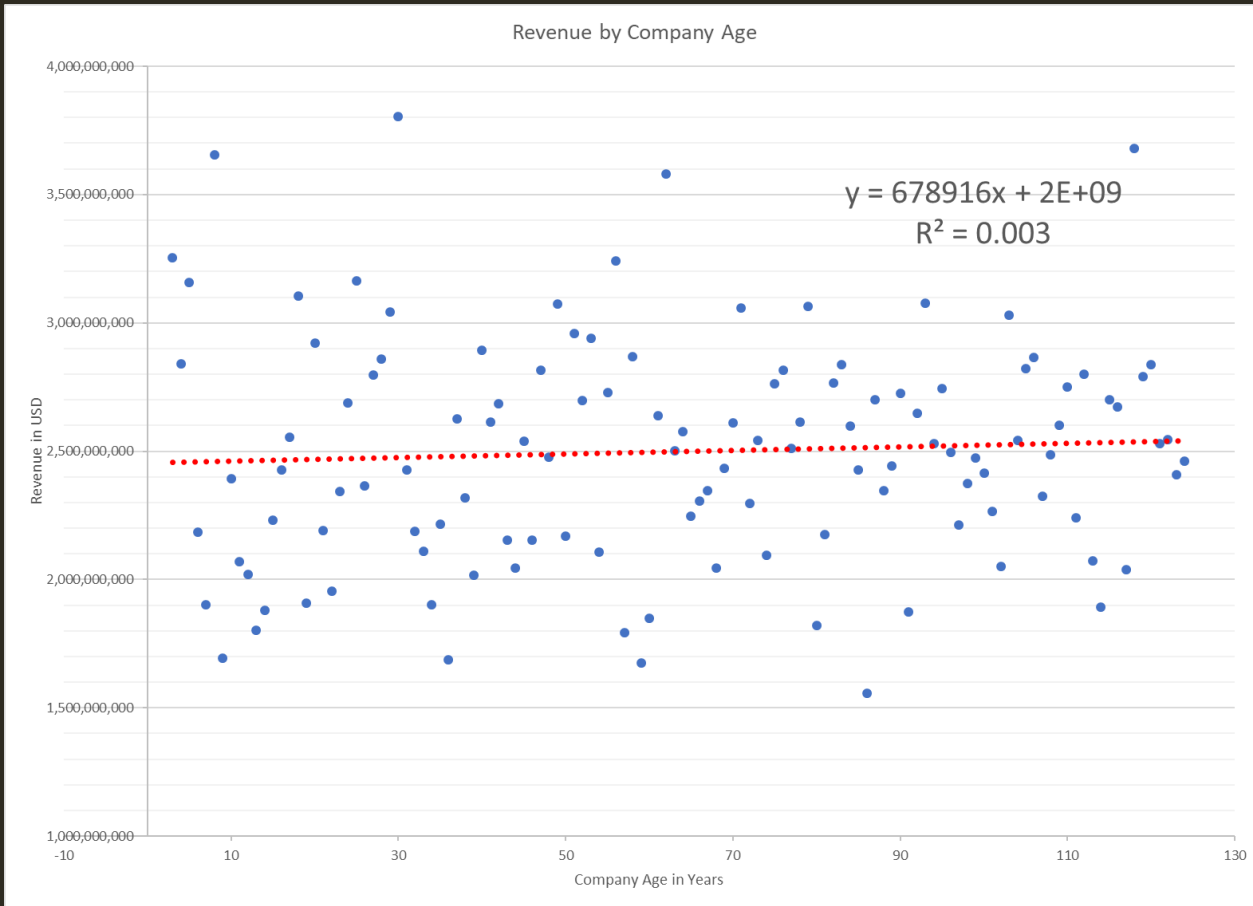
# ASSUMING: ONLY REVENUE (USD) IN DATASET 1 IS REVENUE

Revenue (USD)

4. Considering the 'Year Founded' information, can you identify any correlation between a company's age (how long it has been in business) and its reported revenue or number of employees?

- linear regression in Excel

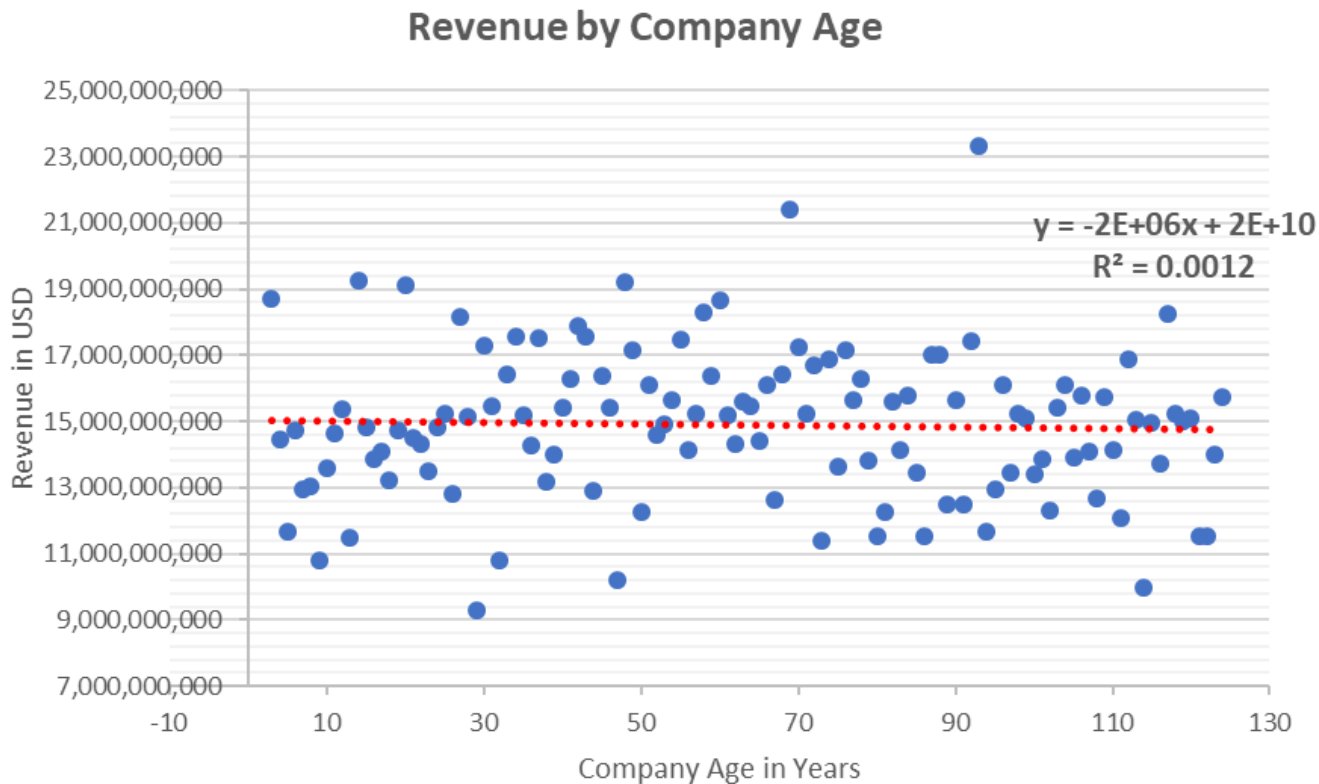
$R = .054549$



# ASSUMING: MARKET CAP CANNOT CORRESPOND TO REVENUE

Revenue (USD)

Annual Sales (USD)



4. Considering the 'Year Founded' information, can you identify any correlation between a company's age (how long it has been in business) and its reported revenue or number of employees?

- linear regression in Excel

$R = -0.0349276$

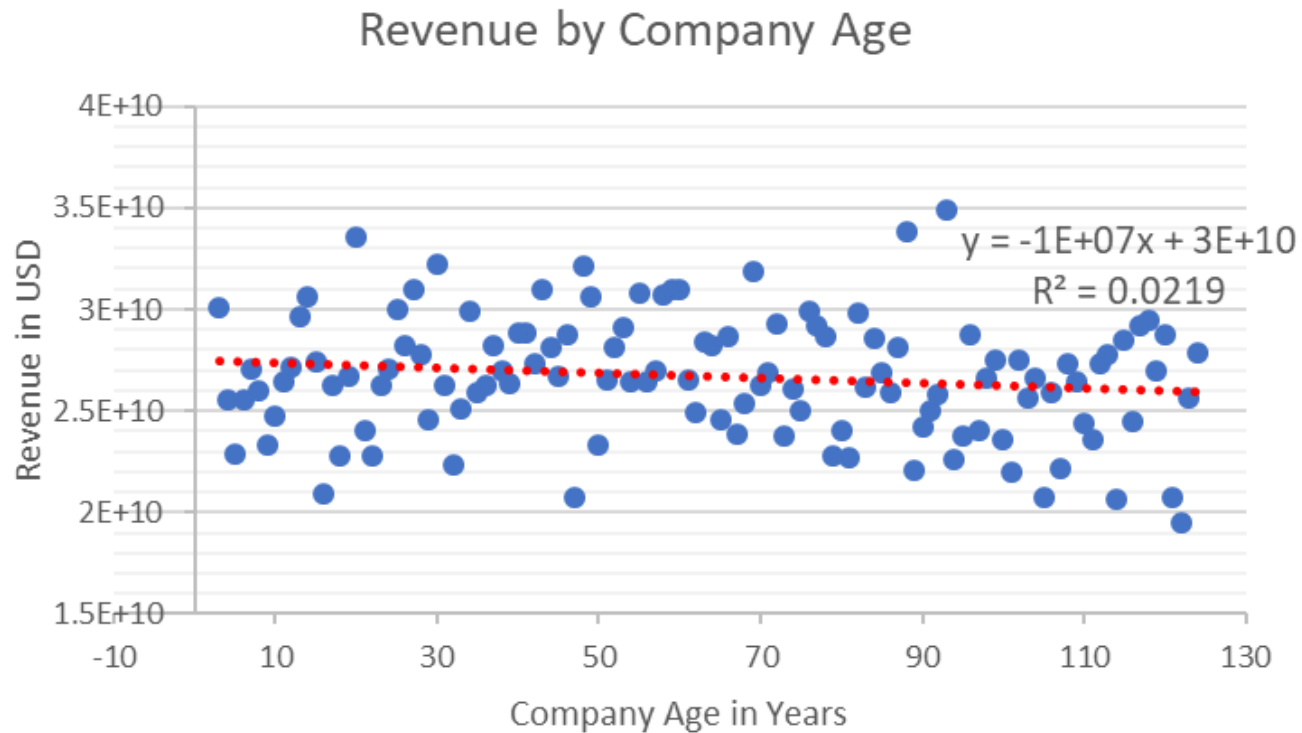
ASSUMING:

ALL THREE VALUES CORRESPOND TO REVENUE

Revenue (USD)

Market Cap (USD)

Annual Sales (USD)



4. Considering the 'Year Founded' information, can you identify any correlation between a company's age (how long it has been in business) and its reported revenue or number of employees?

- linear regression in Excel

$R = -0.14810133$

**THANK YOU!**