An Investigation Into what Data Correlates to Revenue

Theodore Proctor

December 2022

1 Introduction

The Fortune500 are recognised due to their great success. By analysing available data, we can understand which elements correlate to their success. Although profit may equate to monetary gain for a stakeholder, revenue can be used to judge growth and company potential [2], determining the financial strength of a company. This can benefit the individual due to the potential financial gain. Alternatively, studying larger corporations can allow smaller businesses to prosper, since "forecasting sales quantity and sales revenue is very vital for a company to take action for the next period" [8]. This is useful when they don't have complete access to their own historical records.

This paper will discuss methodologies for studying correlation and which characteristics available in the data-set are most valuable for determining a company's revenue.

1.1 Objectives

1. Discover which aspects of the Fortune500 correlate to revenue.
2. Create clustering, regression and classification models to predict revenues.
3. Optimise the hyperparameters used for the aforementioned techniques.
4. Produce figures to display the data.

2 Data-set

Used columns

Column	Data type
Rank	Quantitative
Title	Qualitative
Employees	Quantitative
Sector	Qualitative
Industry	Qualitative
Hqcity	Qualitative
Hqstate	Qualitative
Revenues	Quantitative
Revchange	Quantitative
Profits	Quantitative
Prftchange	Quantitative
Assets	Quantitative
Totshequity	Quantitative

The data-set under consideration contains data regarding the Fortune500 from 2017. Although slightly deprecated, it can serve as a sample to train a models with. The data may lose some of its usefulness when trying to create a general model since it only represents the most successful companies.

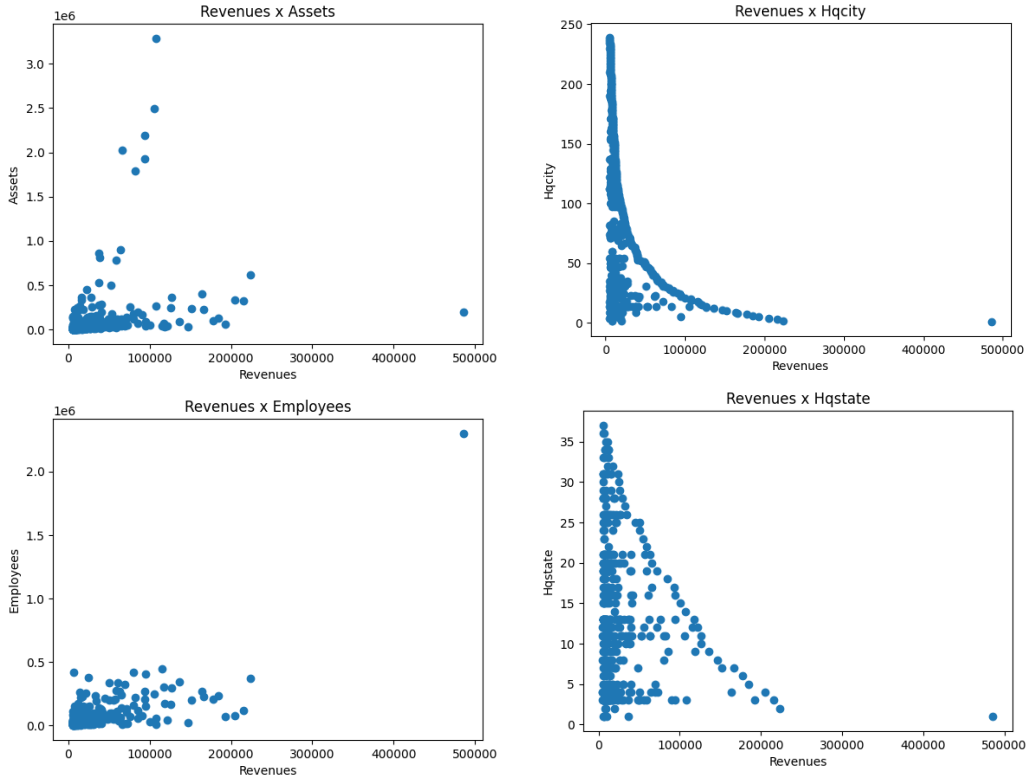
The set provided only provides data for 500 entries. Although sufficient for a simple model, a more accurate model will need more data. Upon initial inspection the dataset is partially incomplete, with elements displaying "NaN" instead of a numerical value. Unless a surrogate value [12] can be provided, these entries might have to be removed. In the instance a surrogate value is provided, "nonlinear algorithms can mistake linear correlations, in particular those of the power law type, for determinism" [11]; this can result in increased implementation complexity.

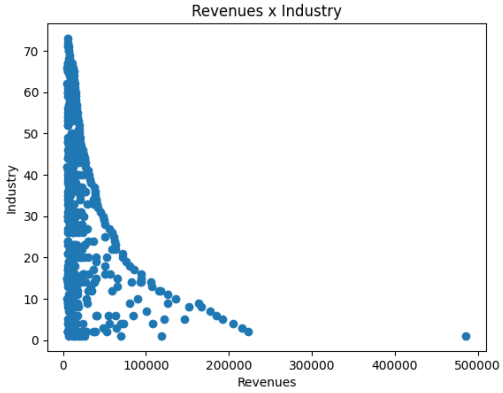
Furthermore, the data must be pre-processed before becoming usable. The data is stored as a ".csv" file, which is incompatible with standard Python. A library such as "Pandas" will be needed to convert the file into a Python-readable format. Normalisation of the data is necessary, ensuring that it is all in numerical form. For qualitative data such as "Sector",

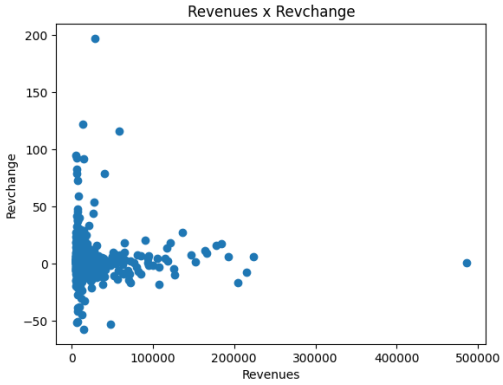
"Industry" and "Location", dictionaries will be defined relating the actual value with a numerical substitute. Since some data included is obsolete, columns such as "Ceo", "Rank" and "Hqaddress" will be removed. Finally, we must remove anomalous data. Although anomalies can represent "natural variations in the population" [14], these outliers are not conducive to the development of a model.

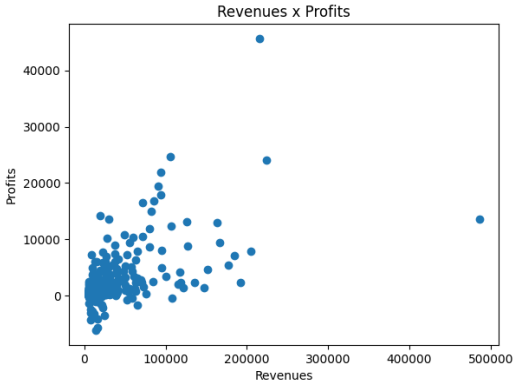
"Data splitting is the act of partitioning available data into two portions, usually for cross-validatory purposes" [10]. We need to apply a data split so that we can test the classification model. One segment of data will be used to train the model, and the other to test it.

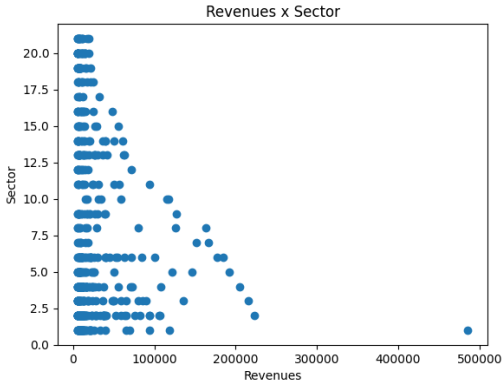
Displayed below are scatter graphs relating revenues to another subject, before anomalous values have been removed.

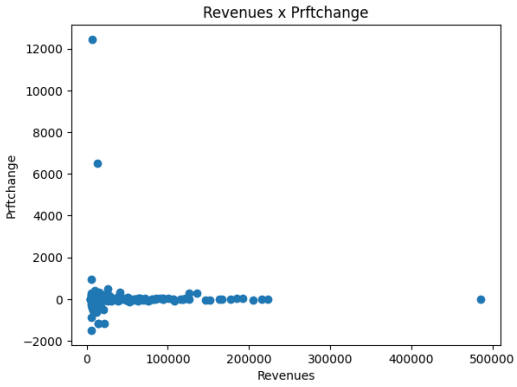


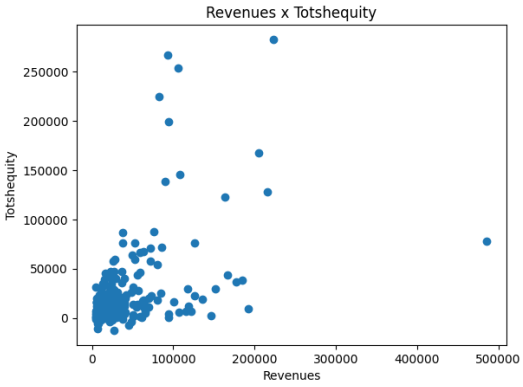

Revenues x Industry
Industry
Revenues
0
10
20
30
40
50
60
70
0
100000
200000
300000
400000
500000


Revenues x Revchange
Revchange
Revenues
−50
0
50
100
150
200
0
100000
200000
300000
400000
500000

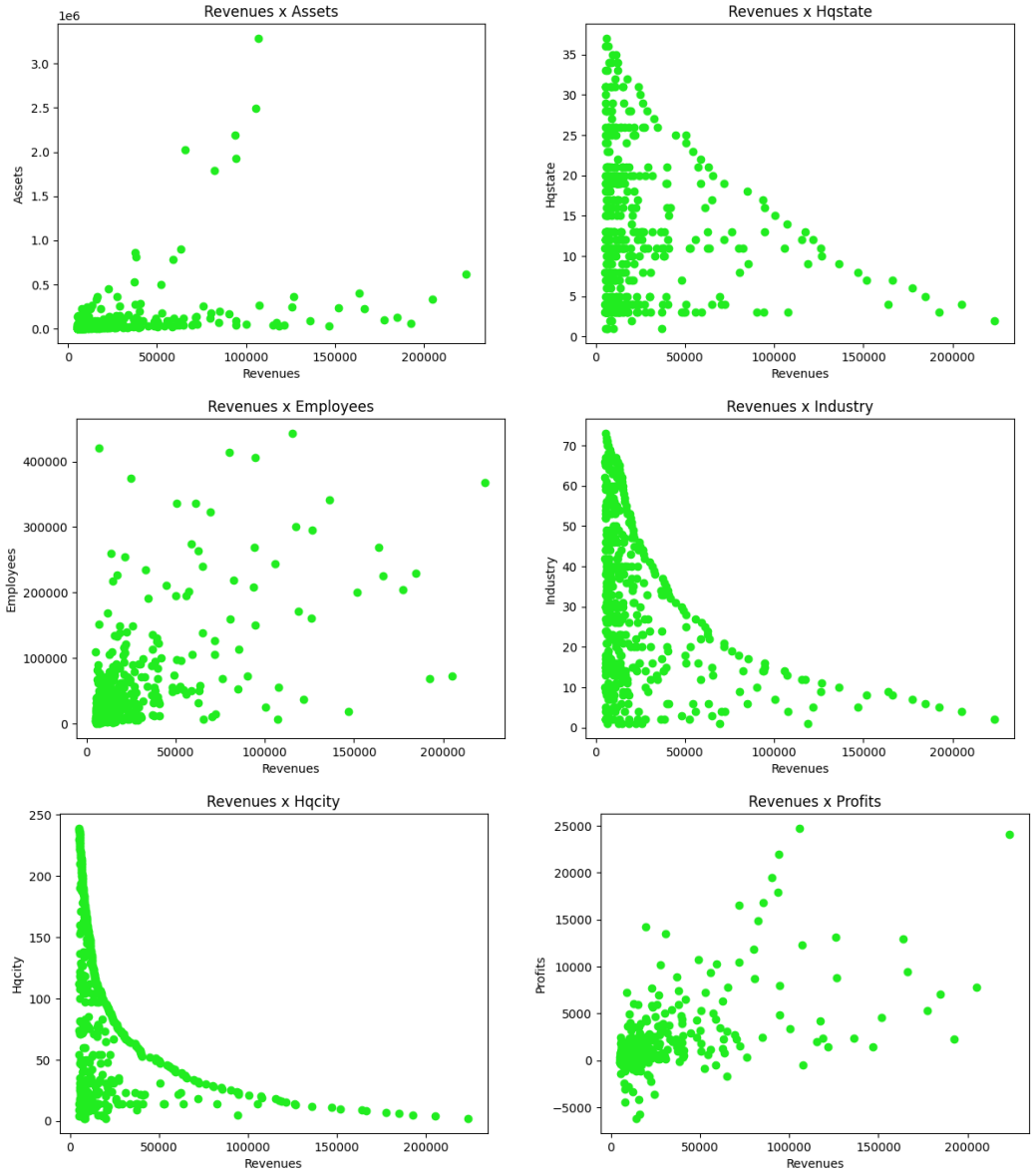

Revenues x Profits
Profits
Revenues
0
10000
20000
30000
40000
0
100000
200000
300000
400000
500000

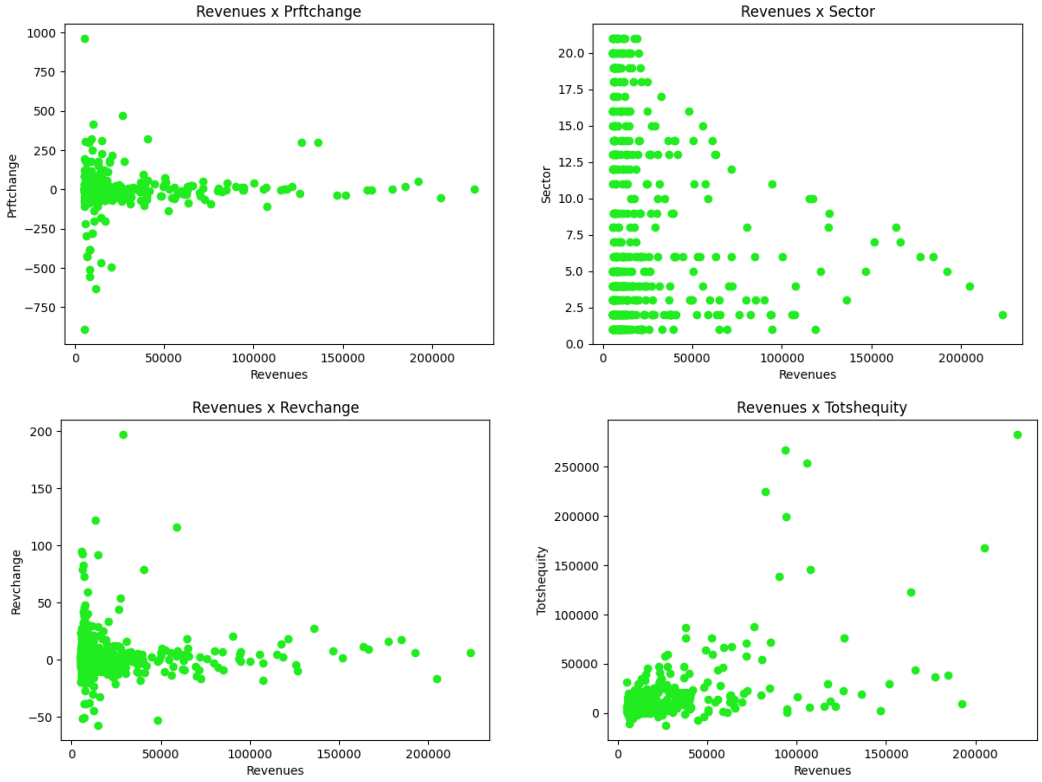

Revenues x Sector
Sector
Revenues
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
100000
200000
300000
400000
500000


Revenues x Prftchange
Prftchange
Revenues
−2000
0
2000
4000
6000
8000
10000
12000
0
100000
200000
300000
400000
500000


Revenues x Totshequity
Totshequity
Revenues
0
50000
100000
150000
200000
250000
0
100000
200000
300000
400000
500000

Post anomaly removal.





3 Methodology

3.1 Clustering - K-means

K-means is a clustering algorithm, which is an unsupervised machine learning technique. It aims to find naturally occurring groups of data in an unlabelled set. "The aim of the K-means algorithm is to divide M points in N dimensions into K clusters so that the within-cluster sum of squares is minimised" [4]. The clusters are based around centroids (artificial values used to represent the middle of the cluster), from which all sum of squares calculations are made. Evaluation will occur through visual inspection.

3.2 Regression - Polynomial regression

Polynomial regression attempts to fit a polynomial equation to the data points provided. "Regression analysis involves identifying the relationship

between a dependent variable and one or more independent variables" [7]. R-squared error is used to evaluate the model.

3.3 Classification - Neural network (NN)

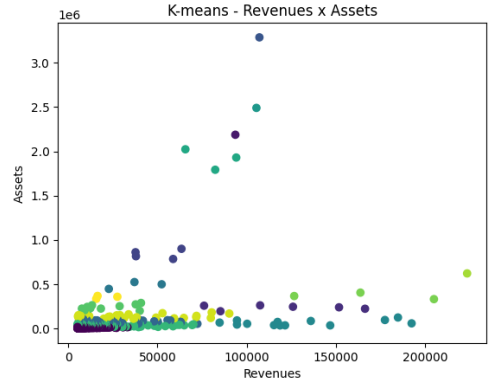
NN's are composed of input nodes (neurons), hidden layers and output nodes [13]. Each connection is associated with some weight, which is calculated by some activation function. "The purpose of the activation function is, besides introducing non-linearity into the neural network, to bound the value of the neuron so that the neural network is not paralysed by divergent neurons" [13]. Nodes may also possess biases, which translate the data to another value. NN's have many hyperparameters, all of which need to be fine-tuned. To evaluate, some denomination of the mean squared error loss function [9] will be employed.

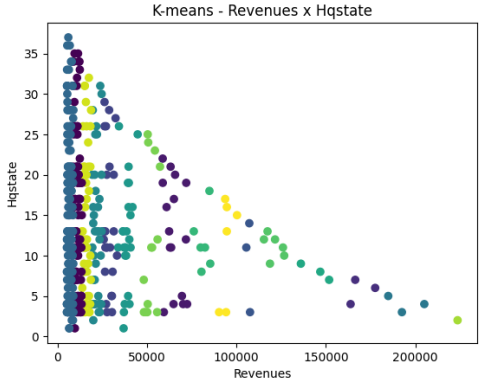
4 Results

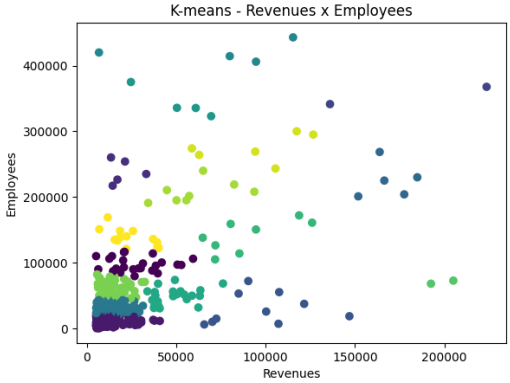
4.1 K-means

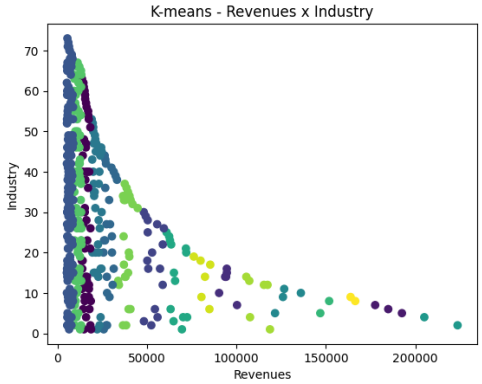
The hyper-parameters for K-means are the value K and the initialisation methodology. Two primary initialisation methods are available. Since K-means++ "improves both the speed and the accuracy of k-means, often quite dramatically" [1], it was used as the initialisation method. K denotes the number of clusters that is optimal for grouping. This value can be calculated with: the Elbow method [6], RSQRT heuristic [3] and the silhouette method [15]. The RSQRT method provides a quick, inaccurate method for calculating K. It is not case specific either.

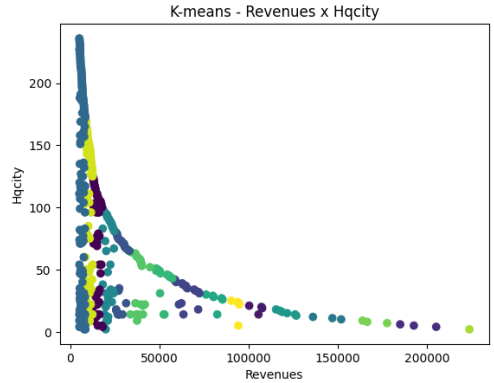
RSQRT graphs:

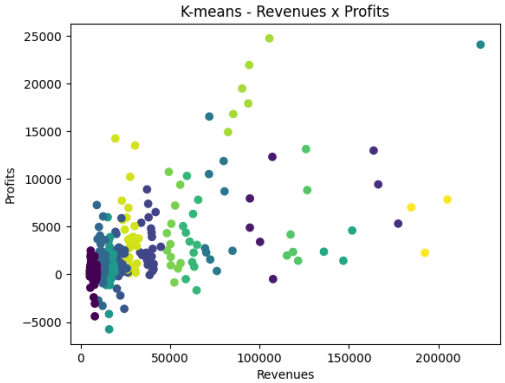

K-means - Revenues x Assets
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Assets
0
50000
100000
150000
200000
Revenues

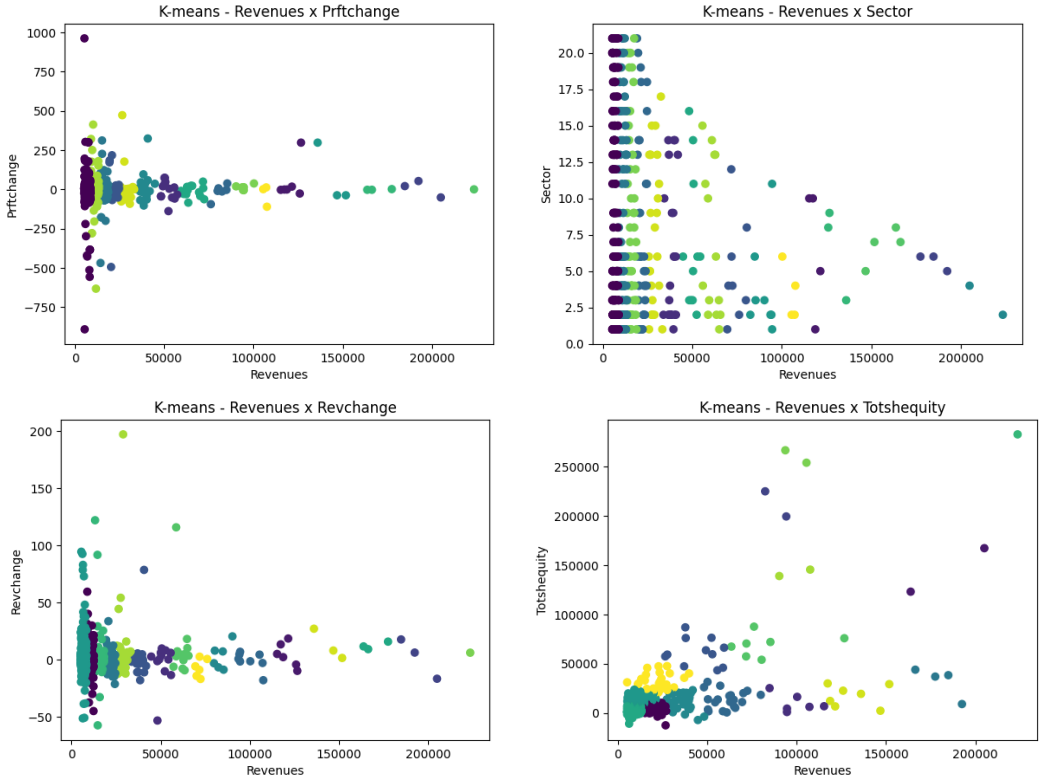

K-means - Revenues x Hqstate
35
30
25
20
15
10
5
0
Hqstate
0
50000
100000
150000
200000
Revenues


K-means - Revenues x Employees
400000
300000
200000
100000
0
Employees
0
50000
100000
150000
200000
Revenues

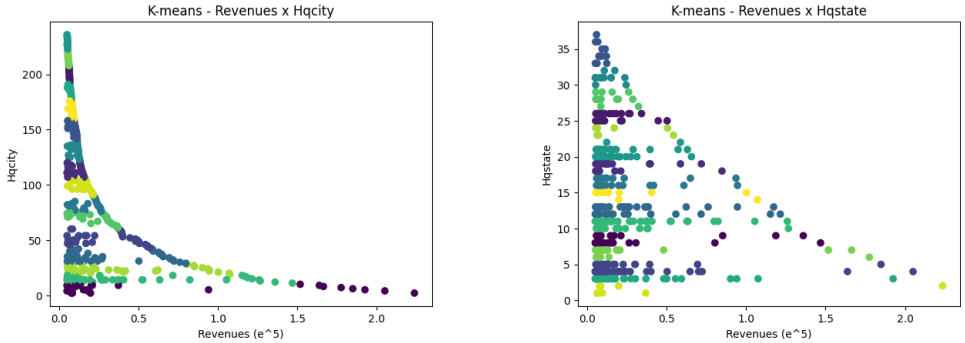

K-means - Revenues x Industry
70
60
50
40
30
20
10
0
Industry
0
50000
100000
150000
200000
Revenues

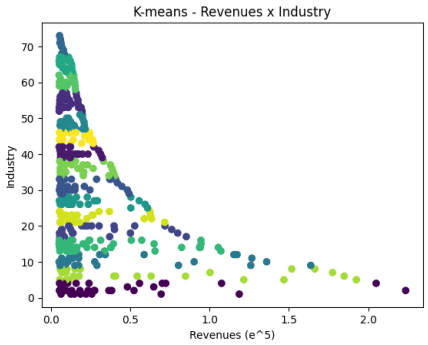

K-means - Revenues x Hqcity
200
150
100
50
0
Hqcity
0
50000
100000
150000
200000
Revenues

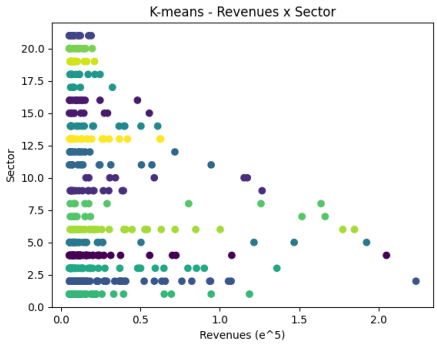

K-means - Revenues x Profits
25000
20000
15000
10000
5000
0
−5000
Profits
0
50000
100000
150000
200000
Revenues



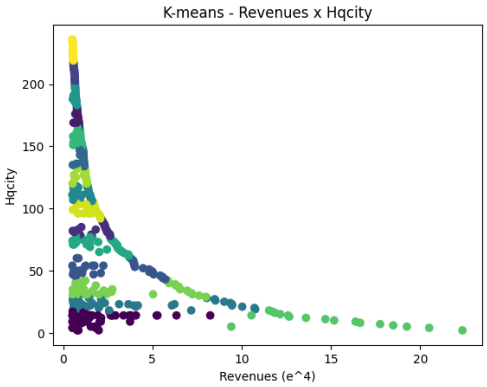
We can see that some of the data was incorrectly modelled due to the fact that the axis scales are orders of magnitude apart. This results in K-means sets tending towards a linear form. Since this occurs for qualitative data types, we need to transform the data so that K-means can be applied properly. This can be done by applying a scalar so that the axis operate on a similar range. Graphs based on quantitative data show little to no trends when clustered, and won't be explored further.

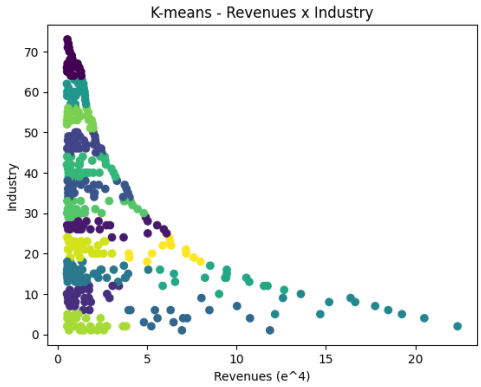


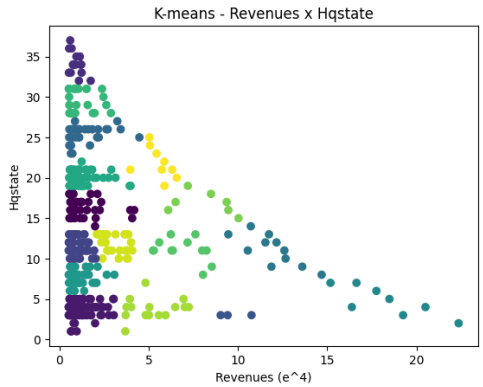


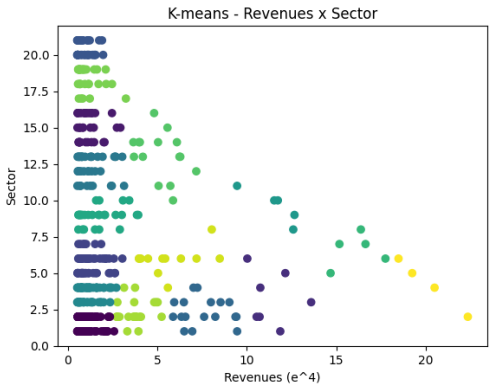


This set of graphs is defunct as there is too strong a correlation between elements of the same subject, resulting in more linear clustering.



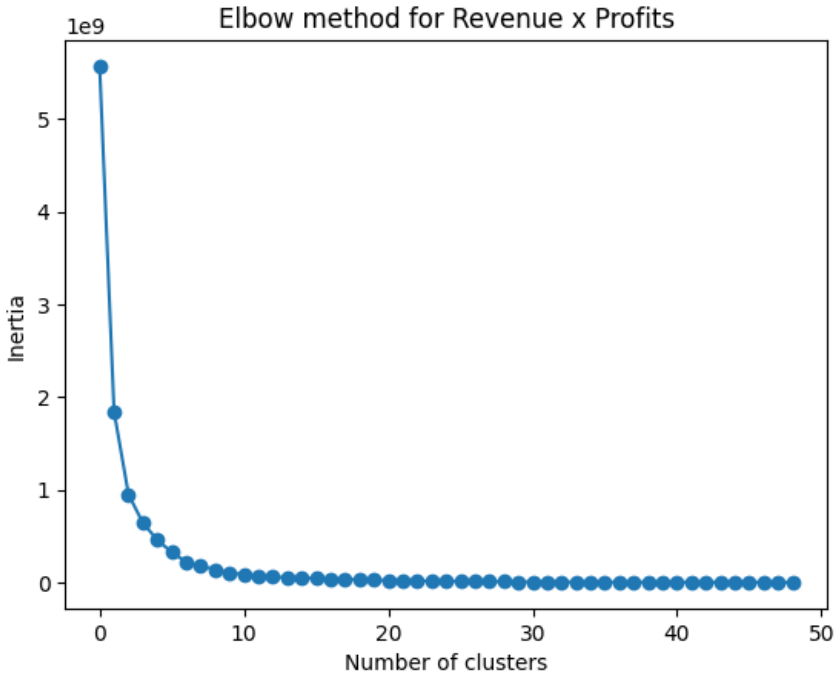




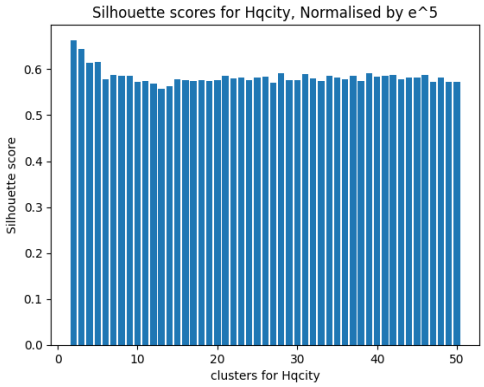


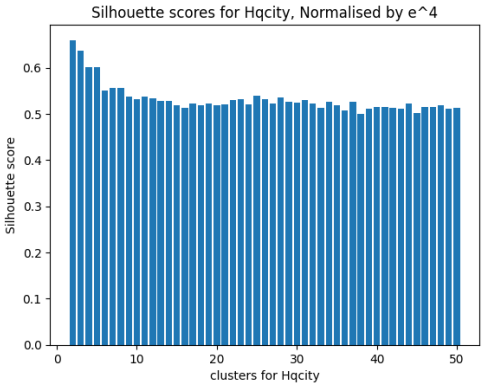
These clusters are more circular, meaning groups are more representative of similarity. Therefore the optimal scalar is one of e^4 magnitude. The Elbow method was not used since it stated that K=2 is optimal. The Elbow method states the greatest change in gradient (where the Elbow forms) is the K value

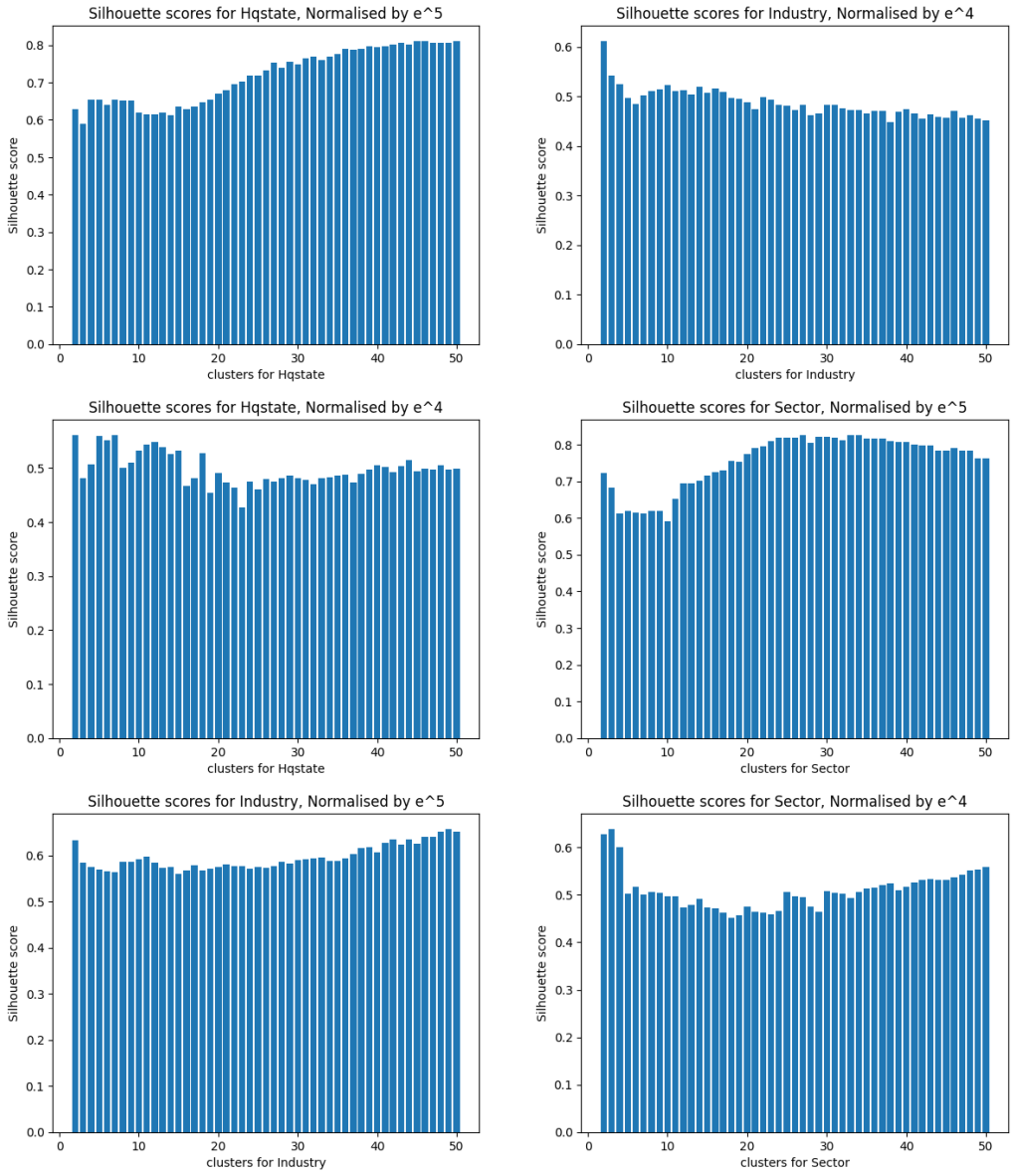
that should be used. This can be found by calculating $\frac{\partial^2 f}{\partial x^2}$ for each data point. Due to the very high inertia [5] of each cluster, this method was ineffective.



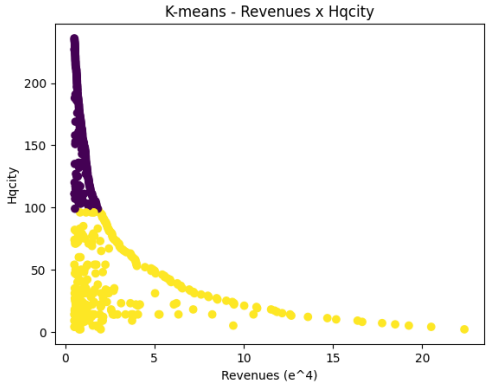
The Silhouette method suggested high values for K when "Revenues" was normalised by a factor of e^5, and low values when the factor was e^4. The maximum number of clusters was limited to 50 to prevent over-fitting.

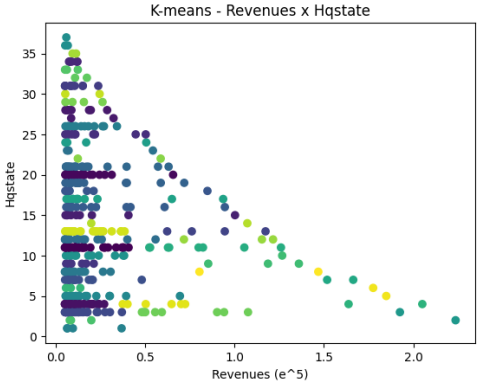


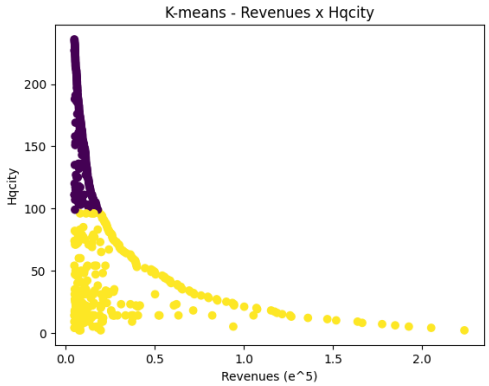


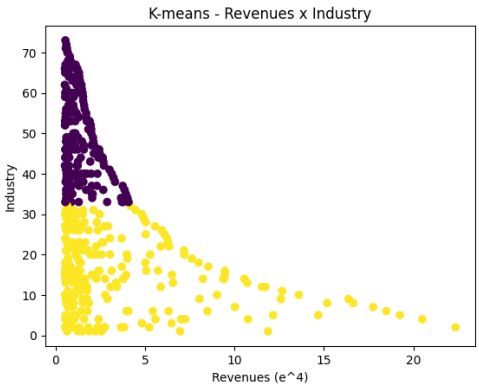


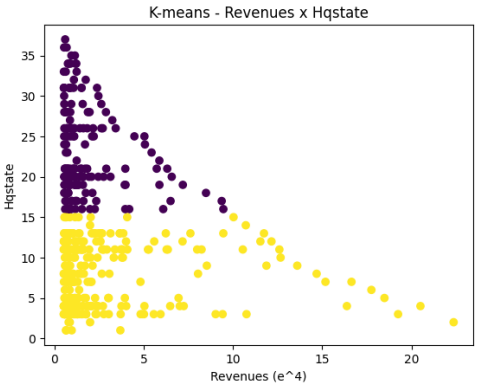
The following are graphs for K-means, with factor reductions of e^4 and e^5, optimised for each.

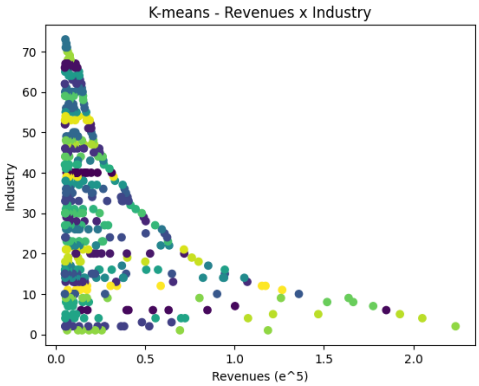

K-means - Revenues x Hqcity
Hqcity
Revenues (e^4)

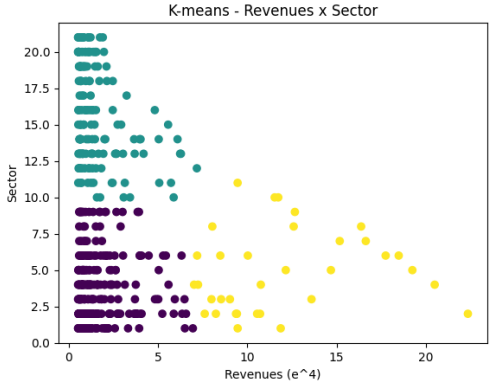

K-means - Revenues x Hqstate
Hqstate
Revenues (e^5)

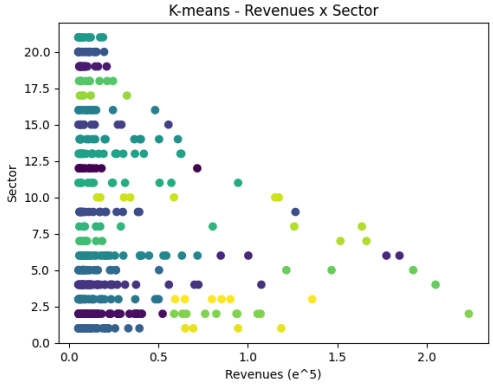

K-means - Revenues x Hqcity
Hqcity
Revenues (e^5)


K-means - Revenues x Industry
Industry
Revenues (e^4)


K-means - Revenues x Hqstate
Hqstate
Revenues (e^4)

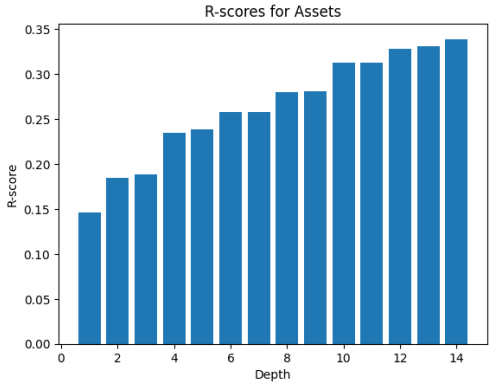

K-means - Revenues x Industry
Industry
Revenues (e^5)

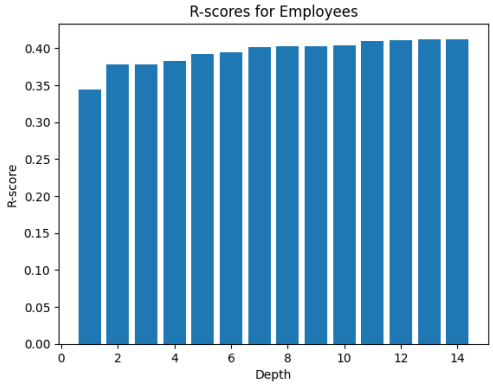


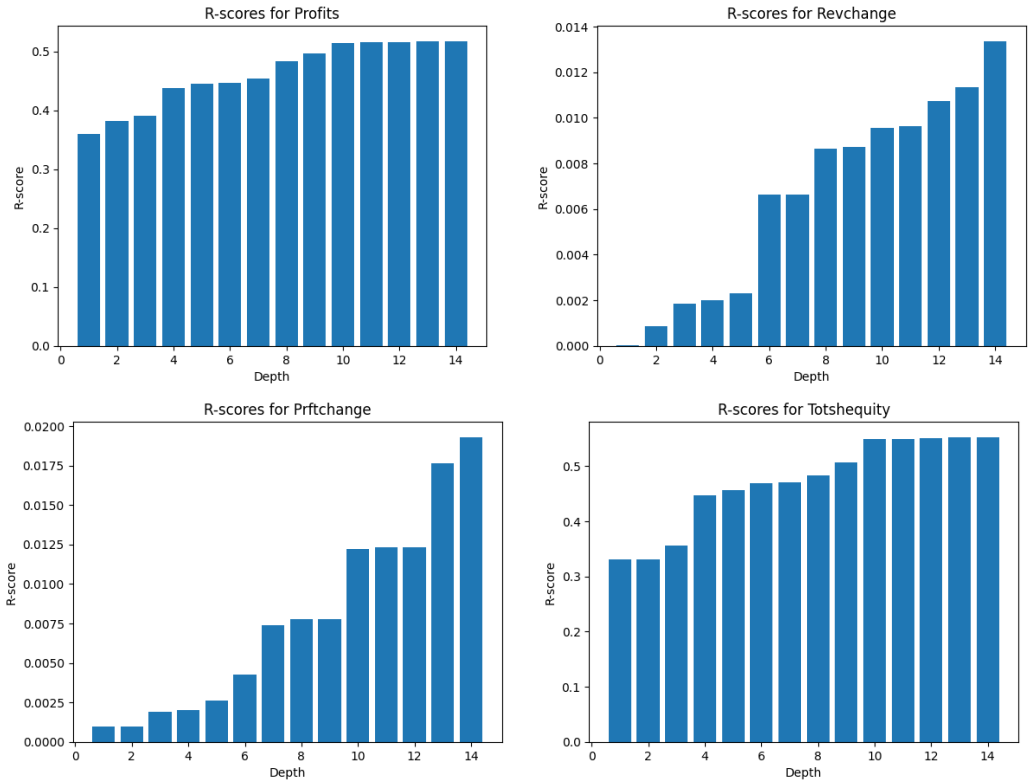


4.2 Polynomial Regression

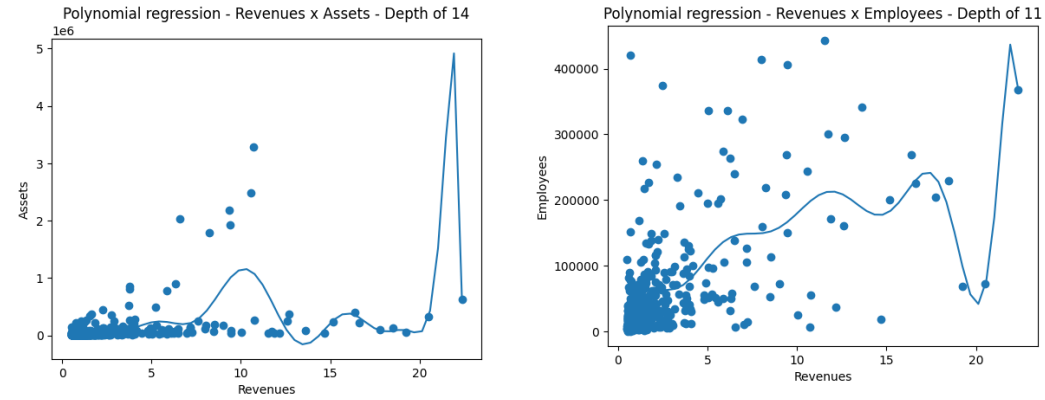
The exponent depth is the only hyperparameter for polynomial regression. We can calculate the R-squared score for each depth, and use the best depth for each subject. This algorithm will not be applied to quantitative data. Firstly, we need to find the R-squared scores for each depth (limiting the maximum depth to prevent excessive overfitting).

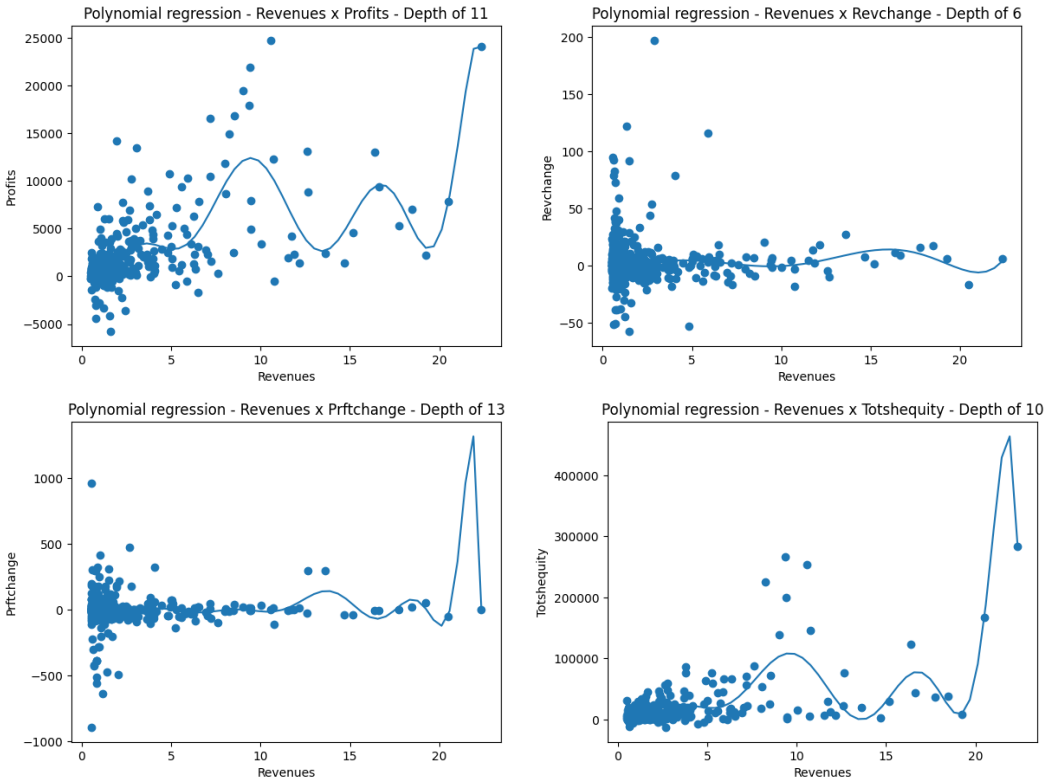






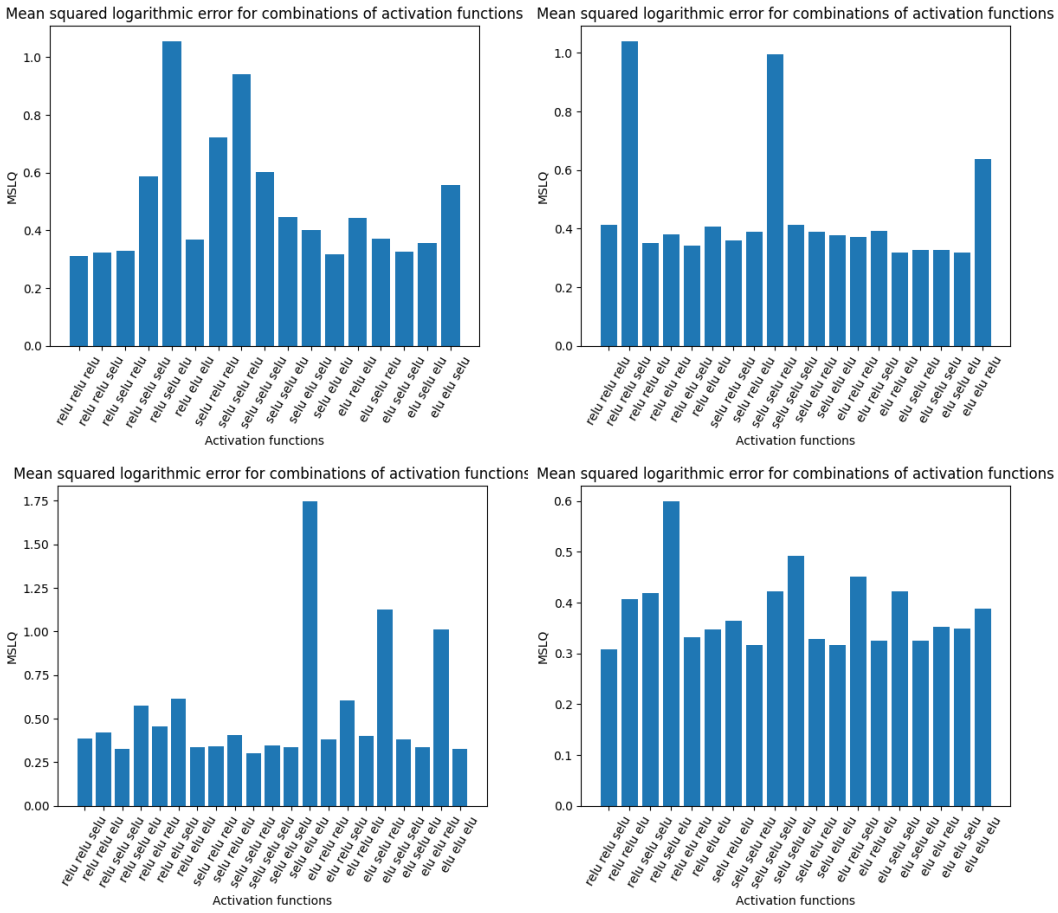
All the charts suggest that the highest exponent value. However this would create byzantine models, due to overfitting. A model with a similar R-squared value but lower complexity must be found ("GoodEnough" depth), finding a value within two decimal places of accuracy will suffice. GoodEnough depth graphs:



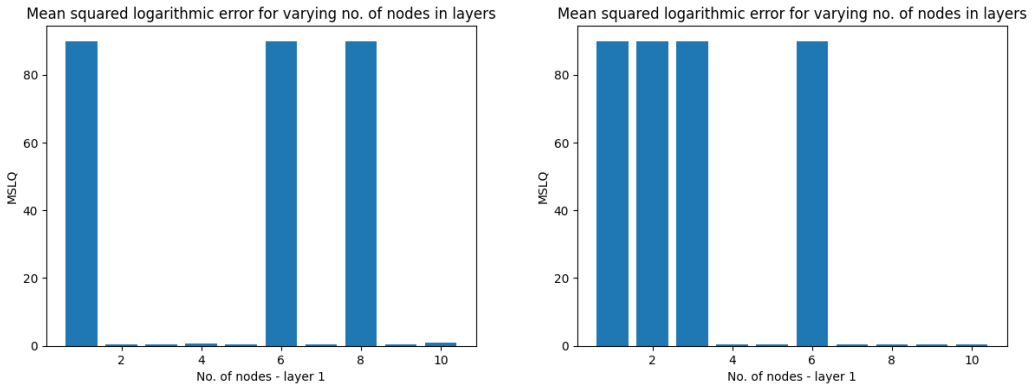


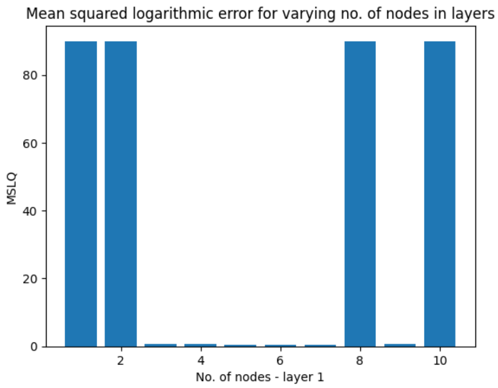
4.3 Neural Network

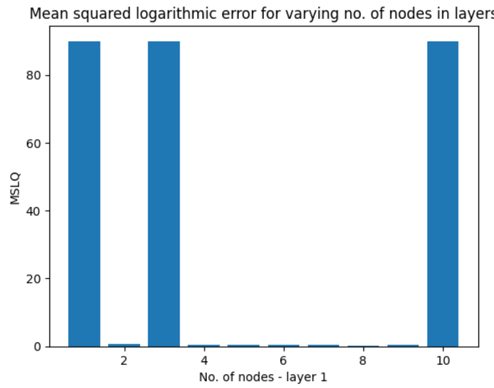
The dataset has been renewed for the NN, with anomalous values being removed and the data being normalised. It was decided to have one input layer, two hidden layers and one output layer. Neural networks have many hyperparameters, one of which is the loss function. Since we are dealing with large values, the "MeanSquaredLogarithmicError" will suffice. Another hyperparameter is the activation functions. Some combinations are obsolete or can produce a deficient model, and have been removed from the displayed graphs. To combat the chance of deficiency in a model, these tests are performed four times. Activation functions tested below:



It is clear that no one group performs better than others, however, the best result was garnered with ”elu, relu, elu”. Tuning the number of nodes per layer is the next hyperparameter to optimise. Layer 1:

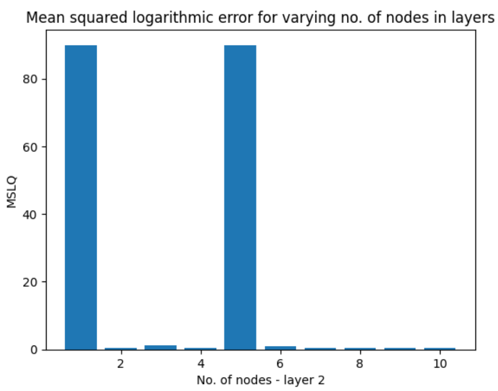


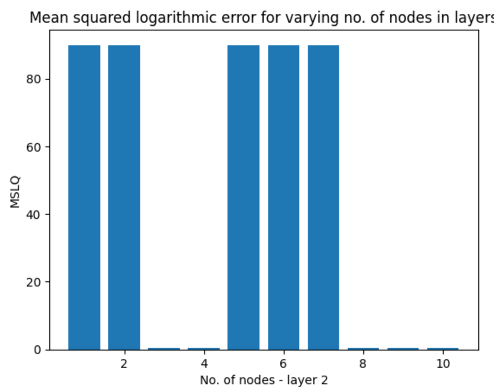


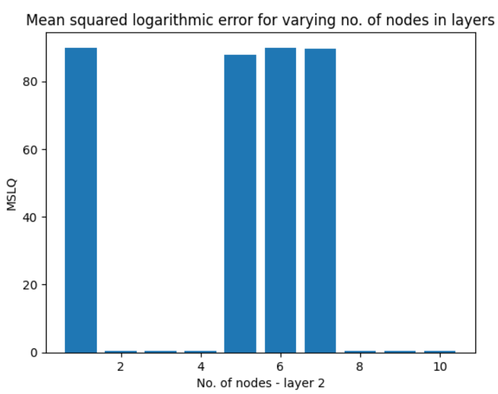


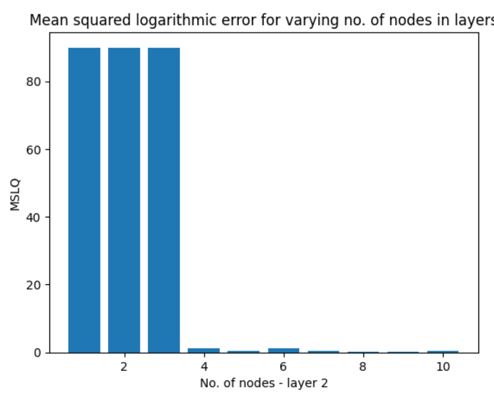
Values of four, five and nine perform well for the first layer.

Second layer

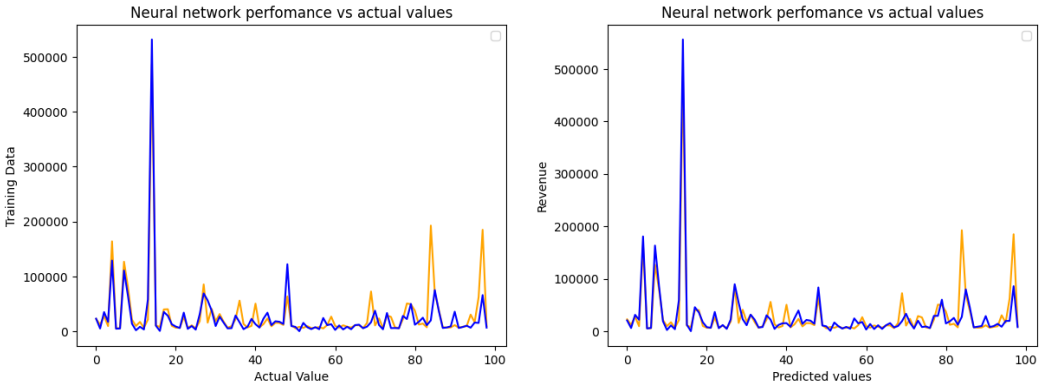








Values of four, eight, nine and ten perform well. Additional study showed that values greater than 50 epochs produced good results. With fully optimised parameters, The NN can produce models as such.



5 Discussion

5.1 K-means

Our exploration of the subject shows that some columns do correlate to revenue. Industry (with large K values) and sector both show evidence of strong clustering once fully optimised. This is in accorddance with real-world knowledge, with certain sectors (such as retailing) being known for high revenue. Clustering on "Hqcity" and "Hqstate" is less conclusive, since clusters are formed across numerous cities or states. An improvement on these subject correlations might be to increase the weight of the subject, so horizontal "clusters" can form. By measuring the "tightness" of these new clusters, we could determine whether "Hqcity" or "Hqstate" do correlate to revenue. By using these methods, another business could examine its industry and sector to try and predict its profit (relative to similarly sized companies).

5.2 Polynomial Regression

This machine-learning methodology seems ineffective for this dataset. Although a R-squared score of 0.5 can suggest a relatively strong correlation in the social sciences [**16**], the quantitative nature of the data would ideally

require stronger evidence of correlation. In addition to this, by studying the graphs generated, some local maxima seem very unrealistic and are not representative of reality. The removal of all anomalous datapoints would improve the model, as would access to a wider dataset. "Totshequity" and "Profits" could strenuously be described as correlated through these means.

5.3 Neural Networks

The NN was the most successful methodology. Although it doesn't prove any direct dependency on revenue, it shows that the combined data can predict revenues accurately. The very low error margins supplant this as evidence. The model could be improved by producing more consistent results, and by a reduction in time complexity. Future developments would include streamlining the code to increase overall efficiency and increasing the available data so that the model would work effectively for many companies' finances. Additionally, more research into what data is provided to the NN could be conducted. This might reduce errors or allow for a more comprehensive model.

References

[1] David Arthur and Sergei Vassilvitskii. *k-means++: The Advantages of Careful Seeding*. Technical Report 2006-13. Stanford InfoLab, June 2006. URL: http://ilpubs.stanford.edu:8090/778/.

[2] Claire Boyte-White. *Revenue vs. profit: What's the difference?* URL: https://www.investopedia.com/ask/answers/122214/what-difference-between-revenue-and-profit.asp.

[3] J Carlis and K Bruso. "RSQRT: AN HEURISTIC FOR ESTIMATING THE NUMBER OF CLUSTERS TO REPORT". In: (), pp. 152–158. DOI: 10.1016/j.elerap.2011.12.006.

[4] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 100–108. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/2346830 (visited on 12/09/2022).

[5] Or Herman-Saffar. *An Approach for Choosing Number of Clusters for K-Means*. URL: https://towardsdatascience.com/an-approach-for-choosing-number-of-clusters-for-k-means-c28e614ecb2c.

[6] Hestry Humaira and Rasyidah Rasyidah. "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm". In: Jan. 2020. DOI: 10.4108/eai.24-1-2018.2292388.

[7] Eva Ostertagová. "Modelling using Polynomial Regression". In: *Procedia Engineering* 48 (2012). Modelling of Mechanical and Mechatronics Systems, pp. 500–506. ISSN: 1877-7058. DOI: https://doi.org/10.1016/j.proeng.2012.09.545. URL: https://www.sciencedirect.com/science/article/pii/S1877705812046085.

[8] Dilek Penpece and Emre Elma. "Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey". In: *International Journal of Trade, Economics and Finance* 5 (Oct. 2014), pp. 435–440. DOI: 10.7763/IJTEF.2014.V5.411.

[9] Christophe Pere. *What are Loss Functions?* URL: https://towardsdatascience.com/what-is-loss-function-1e2605aeb904.

[10] Richard R. Picard and Kenneth N. Berk. "Data Splitting". In: *The American Statistician* 44.2 (1990), pp. 140–147. DOI: 10.1080/00031305.1990.10475704.

[11] Thomas Schreiber and Andreas Schmitz. "Improved Surrogate Data for Nonlinearity Tests". In: *Phys. Rev. Lett.* 77 (4 July 1996), pp. 635–638. DOI: 10.1103/PhysRevLett.77.635. URL: https://link.aps.org/doi/10.1103/PhysRevLett.77.635.

[12] Phil Sherrod. *Missing Values*. URL: https://www.dtreg.com/technical/missing-values.

[13] Sun-Chong Wang. "Artificial Neural Network". In: *Interdisciplinary Computing in Java Programming*. Boston, MA: Springer US, 2003, pp. 81–100. ISBN: 978-1-4615-0377-4. DOI: 10.1007/978-1-4615-0377-4_5. URL: https://doi.org/10.1007/978-1-4615-0377-4_5.

[14] *When should I remove an outlier from my dataset*. URL: https://www.scribbr.com/frequently-asked-questions/when-to-remove-an-outlier/.

[15] Chunhui Yuan and Haitao Yang. "Research on K-Value Selection Method of K-Means Clustering Algorithm". In: *J* 2.2 (2019), pp. 226–235. ISSN: 2571-8800. DOI: 10.3390/j2020016. URL: https://www.mdpi.com/2571-8800/2/2/16.

6 Appendices

Sector dictionary:
1: Retailing
2: Financials
3: Technology
4: Energy
5: Wholesalers
6: Health Care
7: Motor Vehicles Parts
8: Telecommunications
9: Industrials
10: Food Drug Stores
11: Aerospace Defense
12: Household Products
13: Food, Beverages Tobacco
14: Transportation
15: Media
16: Chemicals
17: Apparel
18: Hotels, Restaurants Leisure
19: Materials
20: Business Services
21: Engineering Construction

Industry dictionary:
1: General Merchandisers
2: Insurance: Property and Casualty (Stock)
3: Computers, Office Equipment
4: Petroleum Refining
5: Wholesalers: Health Care
6: Health Care: Insurance and Managed Care

7: Health Care: Pharmacy and Other Services
8: Motor Vehicles and Parts
9: Telecommunications
10: Internet Services and Retailing
11: Industrial Machinery
12: Food and Drug Stores
13: Diversified Financials
14: Commercial Banks
15: Specialty Retailers: Other
16: Aerospace and Defense
17: Computer Software
18: Information Technology Services
19: Insurance: Property and Casualty (Mutual)
20: Pharmaceuticals
21: Household and Personal Products
22: Insurance: Life, Health (Stock)
23: Food Consumer Products
24: Food Production
25: Mail, Package, and Freight Delivery
26: Semiconductors and Other Electronic Components
27: Entertainment
28: Wholesalers: Food and Grocery
29: Network and Other Communications Equipment
30: Chemicals
31: Health Care: Medical Facilities
32: Beverages
33: Insurance: Life, Health (Mutual)
34: Airlines
35: Electronics, Electrical Equipment
36: Construction and Farm Machinery
37: Pipelines
38: Specialty Retailers: Apparel
39: Apparel
40: Utilities: Gas and Electric
41: Miscellaneous
42: Energy
43: Tobacco
44: Wholesalers: Electronics and Office Equipment

45: Food Services
46: Mining, Crude-Oil Production
47: Automotive Retailing, Services
48: Packaging, Containers
49: Medical Products and Equipment
50: Railroads
51: Scientific, Photographic and Control Equipment
52: Temporary Help
53: Engineering, Construction
54: Computer Peripherals
55: Hotels, Casinos, Resorts
56: Metals
57: Oil and Gas Equipment, Services
58: Advertising, Marketing
59: Wholesalers: Diversified
60: Financial Data Services
61: Transportation and Logistics
62: Diversified Outsourcing Services
63: Waste Management
64: Home Equipment, Furnishings
65: Real Estate
66: Homebuilders
67: Securities
68: Publishing, Printing
69: Forest and Paper Products
70: Trucking, Truck Leasing
71: Building Materials, Glass
72: Transportation Equipment
73: Toys, Sporting Goods

HQ city dictionary:
1: Bentonville, 2: Omaha, 3: Cupertino, 4: Irving, 5: San Francisco, 6: Minnetonka, 7: Woonsocket, 8: Detroit, 9: Dallas, 10: Dearborn, 11: Chesterbrook, 12: Seattle, 13: Boston, 14: New York, 15: Dublin, 16: Issaquah, 17: Deerfield, 18: Cincinnati, 19: San Ramon, 20: Washington, 21: St. Louis, 22: Atlanta, 23: Chicago, 24: Charlotte, 25: Mountain View, 26: Redmond, 27: Indianapolis, 28: Philadelphia, 29: Armonk, 30: Bloomington, 31: Houston, 32: New Brunswick, 33: San Antonio, 34: Minneapolis, 35: McLean,

36: Mooresville, 37: Round Rock, 38: Hartford, 39: Purchase, 40: Santa Clara, 41: Newark, 42: Boise, 43: Farmington, 44: Findlay, 45: Burbank, 46: Louisville, 47: Bethesda, 48: Memphis, 49: Palo Alto, 50: San Jose, 51: Midland, 52: Nashville, 53: Fort Worth, 54: Columbus, 55: Kenilworth, 56: Bloomfield, 57: Richfield, 58: Morris Plains, 59: Peoria, 60: Springfield, 61: Redwood City, 62: Springdale, 63: Northbrook, 64: Lakeland, 65: Framingham, 66: Beaverton, 67: Falls Church, 68: Camp Hill, 69: Foster City, 70: Inver Grove Heights, 71: St. Paul, 72: Stamford, 73: Milwaukee, 74: Menlo Park, 75: Miami, 76: Moline, 77: Pittsburgh, 78: Clearwater, 79: Phoenix, 80: North Chicago, 81: Oak Brook, 82: Wilmington, 83: Waltham, 84: Centennial, 85: San Diego, 86: Mayfield Village, 87: Thousand Oaks, 88: Rosemont, 89: Hoffman Estates, 90: Goodlettsville, 91: Fort Lauderdale, 92: Abbott Park, 93: Chesapeake, 94: Benton Harbor, 95: Bloomfield Hills, 96: Richmond, 97: Menomonee Falls, 98: Southfield, 99: St. Petersburg, 100: Long Beach, 101: Eden Prairie, 102: Monroe, 103: Los Angeles, 104: Norwalk, 105: Bellevue, 106: Juno Beach, 107: Parsippany, 108: Austin, 109: Denver, 110: Akron, 111: Englewood, 112: Greenwich, 113: Arlington, 114: Tampa, 115: Fremont, 116: Lincolnshire, 117: Providence, 118: Glenview, 119: Teaneck, 120: Radnor, 121: Hoboken, 122: Arden Hills, 123: Battle Creek, 124: Irvine, 125: Plano, 126: Woodland Hills, 127: Winston-Salem, 128: Franklin Lakes, 129: Des Moines, 130: Princeton, 131: Greensboro, 132: Oklahoma City, 133: Union, 134: Rosemead, 135: Cleveland, 136: Tulsa, 137: Roseland, 138: Melville, 139: Wayne, 140: Chesterfield, 141: South San Francisco, 142: Cambridge, 143: Las Vegas, 144: New Britain, 145: Kalamazoo, 146: Summit, 147: Jacksonville, 148: Chattanooga, 149: New Orleans, 150: Danbury, 151: King of Prussia, 152: Riverwoods, 153: Lake Forest, 154: Auburn Hills, 155: Norfolk, 156: Mechanicsville, 157: The Woodlands, 158: Burlington, 159: El Dorado, 160: Roanoke, 161: Allentown, 162: Estero, 163: Corning, 164: Newport Beach, 165: Broomfield, 166: Kingsport, 167: Calhoun, 168: Los Gatos, 169: Madison, 170: Medford, 171: Grapevine, 172: Marlborough, 173: Greenwood Village, 174: Beverly Hills, 175: Lisle, 176: Downers Grove, 177: Camden, 178: Coraopolis, 179: Orrville, 180: Fort Wayne, 181: El Paso, 182: Byron Center, 183: Warsaw, 184: Melbourne, 185: Hershey, 186: Duluth, 187: Taylor, 188: Brentwood, 189: Plymouth, 190: Newport News, 191: Reston, 192: Erie, 193: Mahwah, 194: Orlando, 195: Durham, 196: Wichita, 197: Santa Monica, 198: Lansing, 199: Perrysburg, 200: Des Peres, 201: Long Island City, 202: San Mateo, 203: Lowell, 204: Victor, 205: Silver Spring, 206: Evansville, 207:

Little Rock, 208: Jackson, 209: Ankeny, 210: Wallingford, 211: Oshkosh, 212: Glendale, 213: Birmingham, 214: West Chester, 215: Rye, 216: Westport, 217: Maumee, 218: Oakland, 219: Kennett Square, 220: Buffalo, 221: Westchester, 222: Toledo, 223: Glen Allen, 224: Itasca, 225: Santa Ana, 226: Clayton, 227: Sunnyvale, 228: Westlake, 229: Brookfield, 230: Tempe, 231: El Segundo, 232: Fairfield, 233: Merriam, 234: Troy, 235: Cedar Rapids, 236: Horsham,

HQ State dicionary:
1: AR, 2: NE, 3: CA, 4: TX, 5: MN, 6: RI, 7: MI, 8: PA, 9: WA, 10: MA, 11: NY, 12: OH, 13: IL, 14: DC, 15: MO, 16: GA, 17: NC, 18: IN, 19: NJ, 20: VA, 21: CT, 22: ID, 23: KY, 24: MD, 25: TN, 26: FL, 27: OR, 28: WI, 29: AZ, 30: DE, 31: CO, 32: LA, 33: IA, 34: OK, 35: NV, 36: KS, 37: AL,