

Start-up Venture Funding Dataset Analysis

Lauren Paige Collins - lcollins.134209@student.uwtsd.ac.uk

Abstract

The chosen dataset is going to be analysed for the use of the data mining processes. To conduct and complete a thorough analysis, the software programs Tableau and Weka are going to be applied to complete the testing phase and the investigation. The following four questioned were asked of the data.

Question One - Is there a trend within the market category and where in the country the companies are operating?

Question Two - Can the countries with the most funding be identified using data mining techniques?

Question Three - Can data mining techniques be used to predict whether the organisations are going to be operating or not, based on the organisation having a website?

Question Four - Can the amount and quarter that a company

Dataset Description

The Start-up Venture Funding dataset was originally collected by a company called CrunchBase. Data World then extracted the data from CrunchBase on 11th May 2017. This dataset contains information regarding start-up businesses from all over the world and was originally collected for analysis and creating comparisons. The data is described as “Information about start-up companies, investment and acquisitions via CrunchBase.” [1]

The main focus of this dataset is on the funding totals (business owners contributing capital to buy "shares" of the company) and funding rounds (money offered by investors, in return for an equity stake in the business). The original size of this dataset was 48.63MB and contained 18 Columns, 49,439 Records and 791,024 cells. The original fields were:-

- Permalink
 - Name
 - Homepage URL
 - Category List
 - Market
 - Funding Total (USD)
 - Status
 - Country Code

- State Code
 - Region
 - City
 - Funding Rounds
 - Founded At
 - Founded Quarter
 - First Funding At
 - Last Funding At

Pre-processing of the Data

Empty Values

Due to the data set containing empty values throughout the table, a Ctrl-Find was conducted within the file to resolve this issue. For each attribute that had empty values, Excel was then used to insert the word ‘Missing’ into each cell that was blank. The Funding Totals attribute had previously been auto filled with a ‘-’. During the cleaning process, these were then replaced with ‘0’ to ensure successful launch within both Tableau and Weka.

Deleted Records

There were a few records that were removed from the dataset, due to incorrect date information. It was unclear if the incorrect dates were a typing error or incorrect data. As a result of this, 10 records with this issue were removed and documented.

Added Attributes

Six attributes were able to be added to dataset, splitting some of the fields, allowing for a more concise investigation to be conducted on the data. These were created by performing calculations on the fields that already existed within the dataset

- Quarter
 - Founded Year Quarter
 - First Funding Quarter

- First Funding Year
 - Last Funding Quarter
 - Last Funding Year

Inserting data into Weka

The first issue was result of some of the elements containing commas within the punctuation of the words e.g. A find and replace was conducted.

operating	USA	FL	Florida's Space Coast	Rockledge	1
operating	SWE		Stockholm	Stockholm	3

Another issues arose regarding the data that had been classified as String. To resolve this, each data record which was string, was surrounded by inverted commas e.g. ‘waywire’. To achieve this, a formula (=""&A2&"") was applied within the excel sheet.

For a sufficient data visualisation, the input of the data within Notepad++ was required to be modified. For the relevant attributes, the options that were included within those fields needed to be entered alongside the data type. This procedure was done for attributes country code, state code and status.

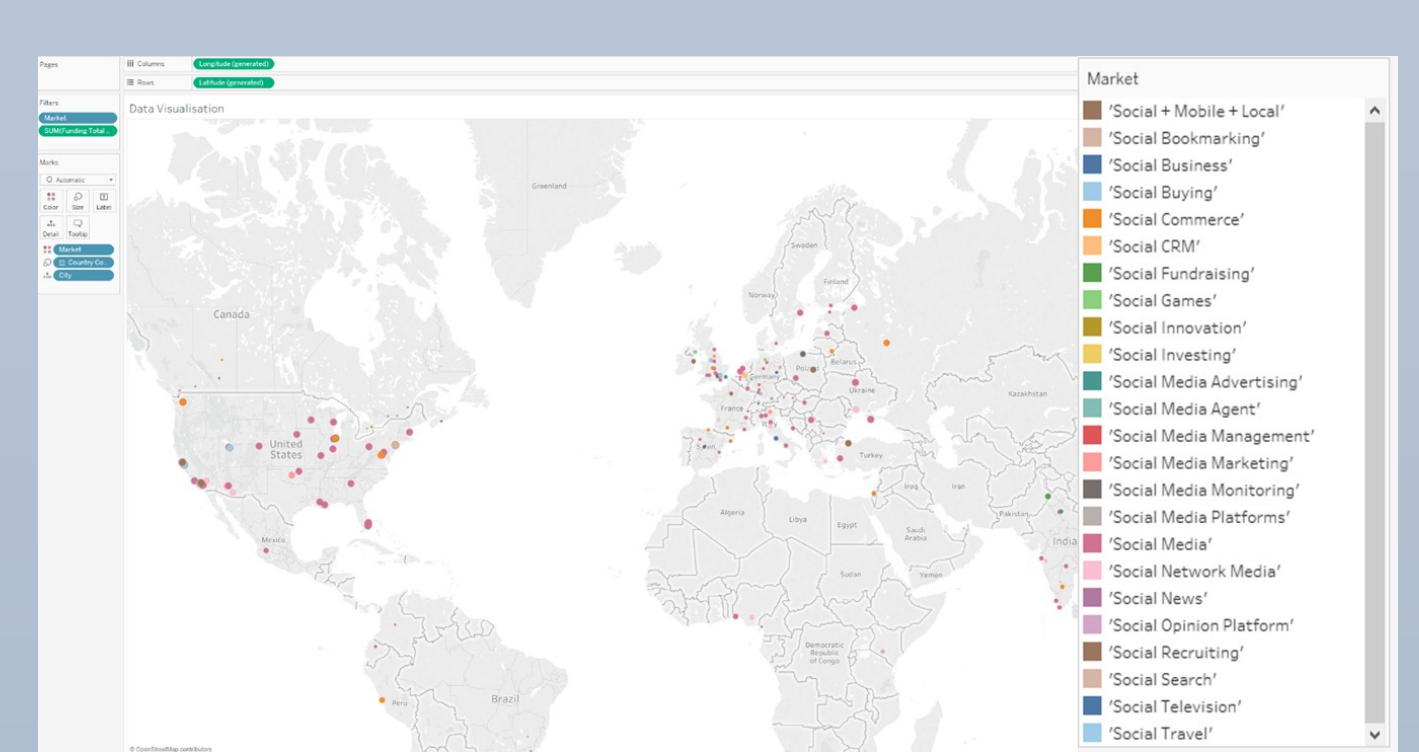
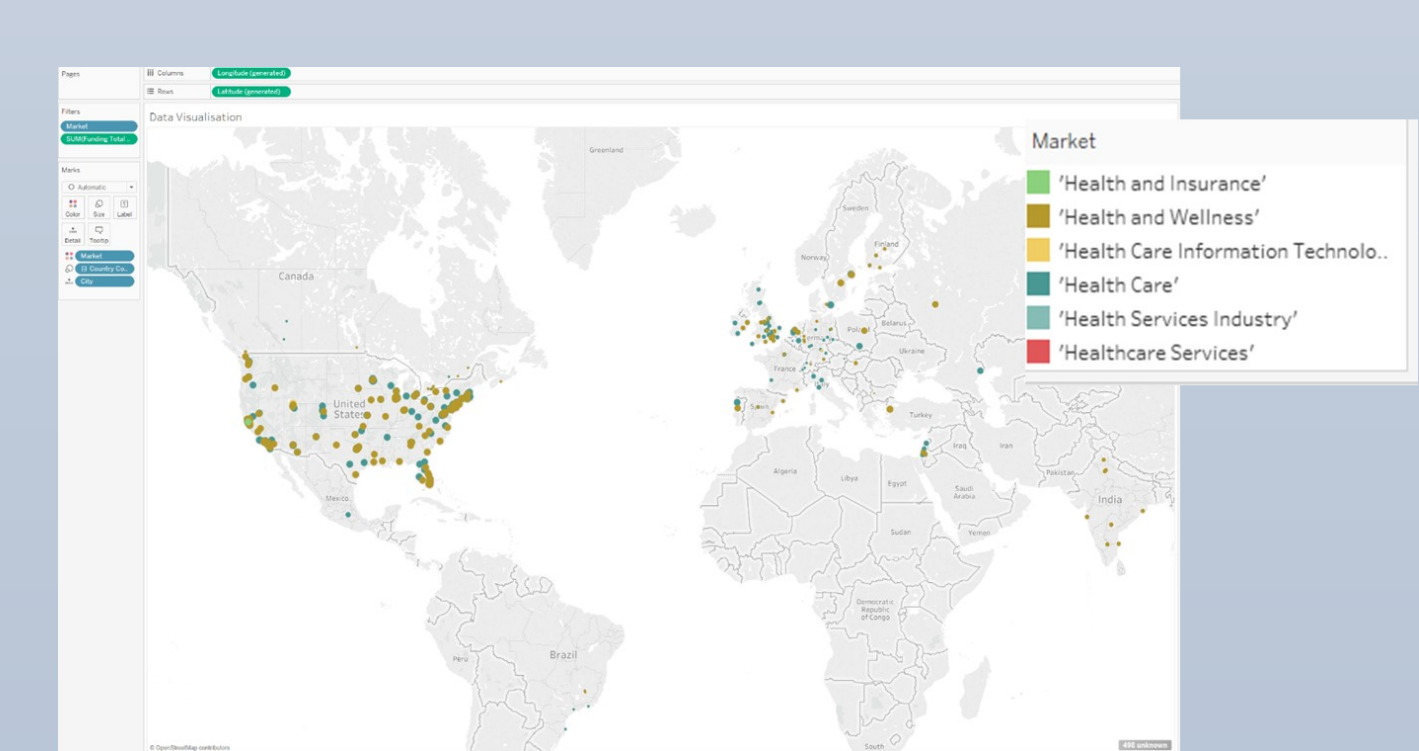
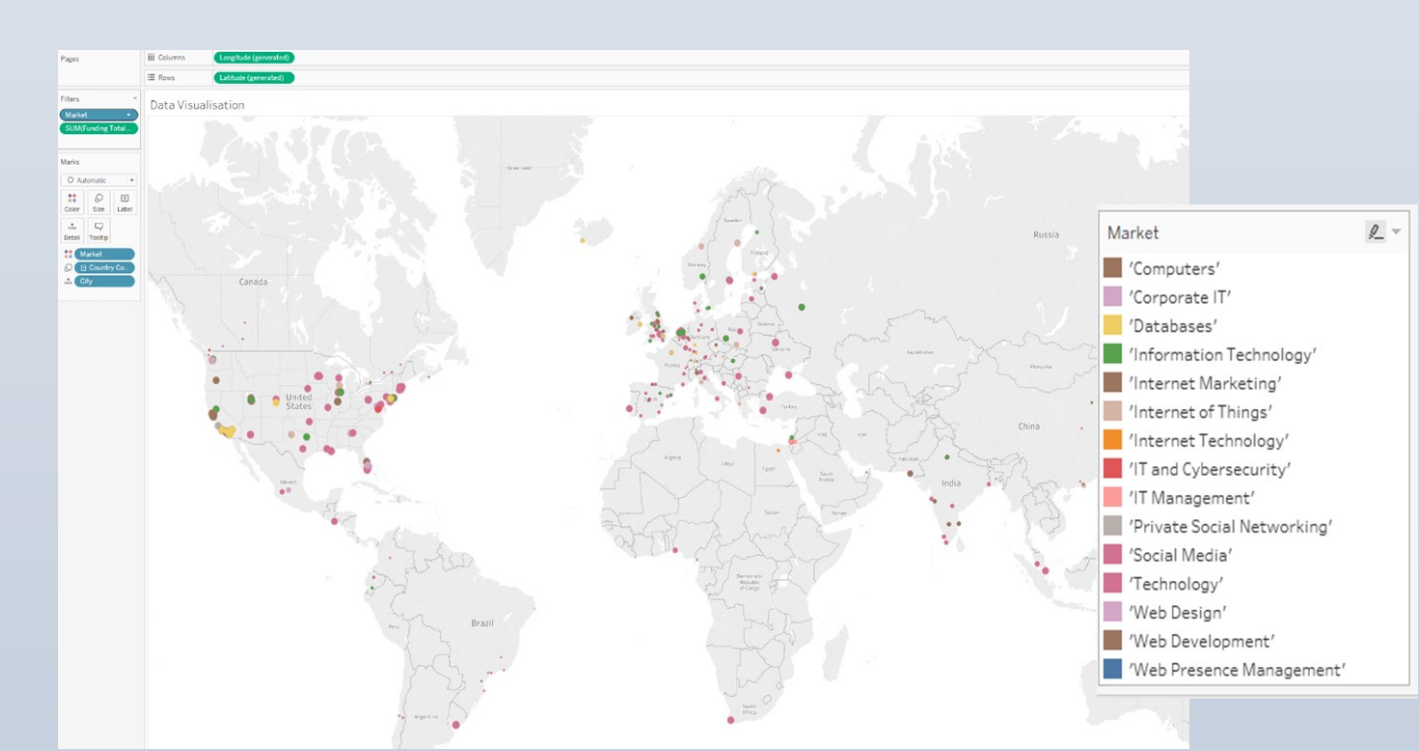
```
1. [RELOCATION #startup
2.
3. ATTRIBUTE permalink STRING
4. ATTRIBUTE name STRING
5. ATTRIBUTE homepage_url STRING
6. ATTRIBUTE category_list STRING
7. ATTRIBUTE market STRING
8. ATTRIBUTE funding_total_usd NUMERIC
9. ATTRIBUTE status (operating, acquired, missing, closed)
10. ATTRIBUTE country_code (BEL, USA, Missing, ESP, ITA, DEU, JPN, GBR, ESP, DNK, BEL, MEX, CAN, YRM, IND, NLD,
11. ATTRIBUTE state_code (Missing, MA, PA, IL, CA, MD, FL, KY, AL, NY, OH, TN, AR, WI, TX, MI, NM, MO, NC, CO,
12. ATTRIBUTE region STRING
13. ATTRIBUTE city STRING
14. ATTRIBUTE founded_rounds NUMERIC
15. ATTRIBUTE founded_at DATE 'yyyy-mm-dd'
16. ATTRIBUTE founded_year NUMERIC
17. ATTRIBUTE founded_quarter STRING
18. ATTRIBUTE quarter NUMERIC
19. ATTRIBUTE founded_year_quarter DATE 'yyyy-mm-dd'
20. ATTRIBUTE first_funding_date 'yyyy-mm-dd'
21. ATTRIBUTE first_funding_quarter NUMERIC
22. ATTRIBUTE first_funding_year NUMERIC
23. ATTRIBUTE last_funding_at DATE 'yyyy-mm-dd'
24. ATTRIBUTE last_funding_quarter NUMERIC
25. ATTRIBUTE last_funding_year NUMERIC
26.
27.
```

Analysis and Results

Question One:

Is there a trend within the market category and where in the country the companies are operating?

To conduct a fair investigation for this question, three different popular market trends were chosen which had a majority of markets. After analysing three market trends, there was clear indication that USA (United States of America) dominates these particular trends and the markets included.

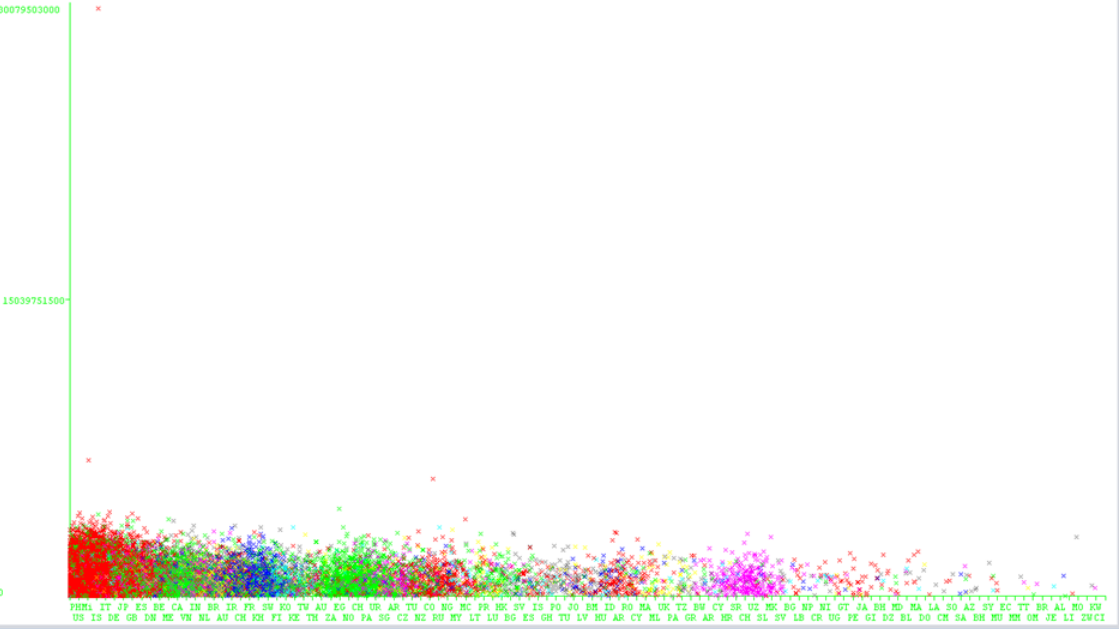


Data Mining - MSc Applied Computing

Question Two:

Can the countries with the most funding be identified using data mining techniques?

The algorithm outputted that USA and China are the countries with the most funding within this dataset. By analysing the below model, the clusters can be easily identified due to the variation of colours. USA is red and China is green. The jitter tool was used below to gain a clearer indication of the cluster sizes.



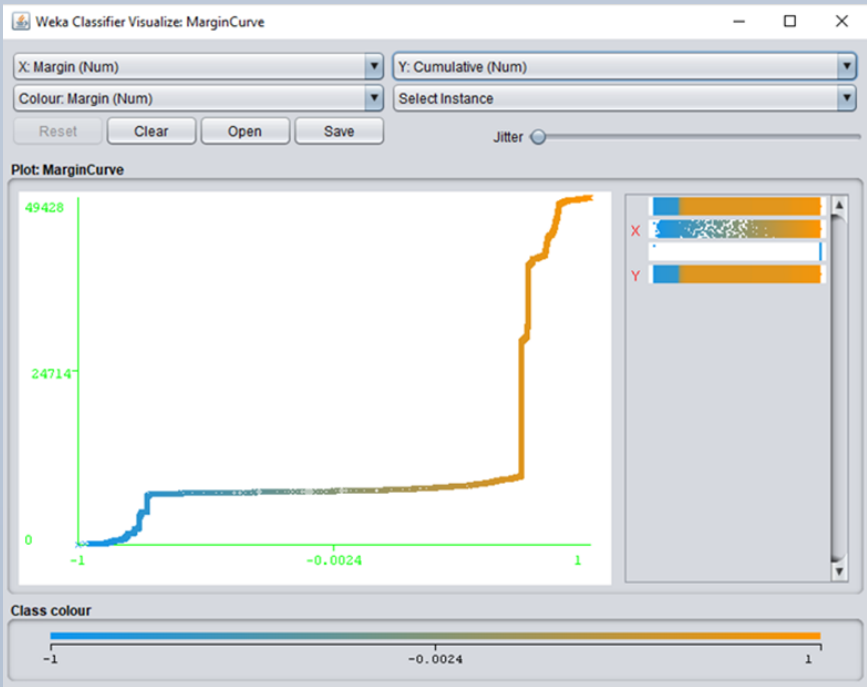
Question Three:

Can data mining techniques be used to predict whether the organisations are going to be operating or not, based on the organisation having a website?

To collect a predicated result from the classification, the status and homepage_url were applied. This chosen algorithm predicted that if a company does have a website in place that the organisation will be operating.

Confusion Matrix

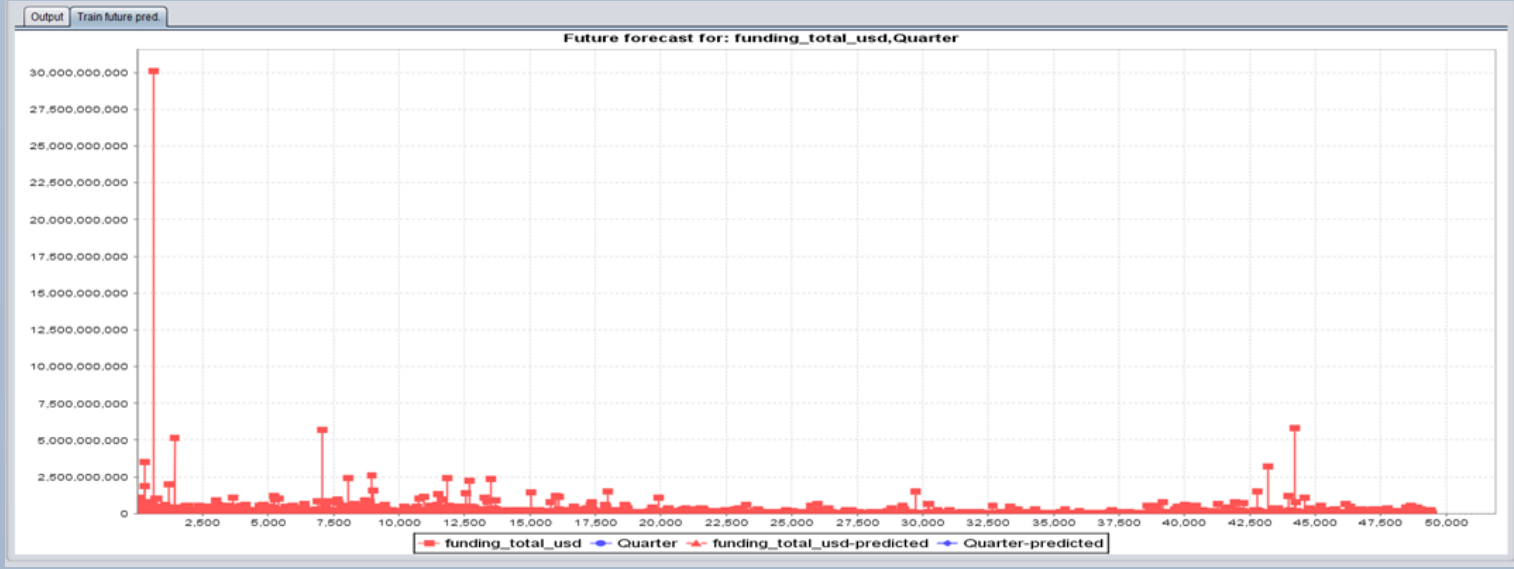
	a	b	c	d	classified as
a	41730	6	63	21	a = operating
b	3691	0	1	9	b = acquired
c	1214	0	100	0	c = Missing
d	2568	2	3	30	d = closed



Question Four:

Can the amount and quarter that a company may receive funding be predicted?

The test result predicated based on the data that the average amount that would be paid to an organisation is 14,642,002 and would be within the first quarter of the year (April – June).



1. “Startup Venture Funding,” 11 May 2017. [Online]. Available: <https://data.world/datanerd/startup-venture-funding>. [Accessed 21 November 2018].