Data Analytics Project 1

Team Members:

Fernando Figueiredo

Christopher Guthrie

Rachel Teigen

**Update this document in Google Docs and download as needed.**

**Due date:** 09/05/2017

Link to a submission document where we add a link to our presentation and GitHub repository:

<https://docs.google.com/forms/d/e/1FAIpQLSeKY7-0QrxyOmmoahmnRsacOdjM0yIt4sx0WdW6ksSo2tHHSQ/viewform?usp=sf_link>

**Project Scope:**

**See Pavan’s project requirements PDF in our repo docs or Google drive folder.**

1. What are the interesting questions we want to explore and answer?
2. Data sources
   1. CrunchBase data searches/queries and csv downloads.
   2. API data interface.
   3. Screen scraping similarweb.com.
3. In Jupyter Notebook, create Pandas data frames to organize and prepare data for charting and analysis.
4. Use descriptive statistics to make sense of the data (you're probably going to want to use inferential statistics too) mean, median, mode, range, iqr, outliers, std deviation, variance.
5. Create charts using Matplotlib, and Seaborn.
6. Presentation slide show using PowerPoint?
7. Big text and few sentences.
8. Data decision and data gathering slides.
9. Graphics. Charts and analysis slides.
10. It should be fun to watch (subjective requirement).
11. *Appendix with footnotes giving credit for data, pictures, and information.*
12. We also need a "solid markdown" file of our notebook.
13. Project files and folders should be in github.
    1. We want file and code sharing mainly through GitHub, and it would be good to try Google Drive, and Google Docs.

**Project Approval:**

* Approval by default.

**Issues and Assumptions:**

1. This is our first project, so we need a simple project with readily available data.
2. Crunchbase data availability is licensed ($) at the Application and Enterprise levels and is accessed through a user search/query interface and API.

* We can’t access some API structures even though we purchased access.
* Crunch Support won’t respond to requests for help.
* The user interface for searches and queries has export limitation.
* There are data and usability differences between the API and user interfaces.

1. We have software issues to overcome with data scraping.
2. We haven't yet covered descriptive or inferential statistics.

#------------------------------------------------------------------------------------------------------------------------------

**Questions, Exploration, and Answers Section:**

### **StartUp Viability Assessment**

1. What is the average Duration of Startups?
   1. Create a data frame of closed companies with an average of the days between the Founded On Date and Closed On Date.
   2. Could be a scatter plot, with the mean as one axis. What is the other axis?
2. What is the Average Investment $?
   1. Group by funding type and take the amount.mean().
   2. Create a bar chart of the data.
3. Is there a correlation from duration (shelf-life) of startup, investment capital, investment firm?
   1. What data, x, y axes, and chart do we use?
4. What determines the rate Decline: 1 year with no new funding, 6 months or no new employees added?
   1. What data, x, y axes, and chart do we use?
5. Is there a prediction for failure for startups (time markers: for new new series funding & adding employees, location, type of product)?
   1. Is there? Data, x, y axes, chart?
6. Is there a prediction/correlation of Success for startups (duration of business, new investments of capital, acquisition, IPO, location and types of product)
   1. Is there? Data, x, y axes, chart?
7. What determines the rate of Growth: Internet traffic vs years in business or initial funding vs new funding?
   1. The data needed is derived from screen scraping and is still under development.
   2. Similarweb.com scraping data available is whatever was produced last month or was produced in the past if the company is no longer in business.
   3. Past data is available with date ranges, but that is beyond our screen scraping abilities at this time.
   4. How do we compare current and last available data and to what do we compare the data? X, y axes? Chart type?

**Questions that we need to ask:**

* % of Start ups that are in the Seed Stage?
* % of Start ups that advance to Series A?
* % of Start ups that advance to Series B?
* % of Start ups that advance to Series C?
* % of Start ups that advance to Series D?
* Average time it takes a startup to request next series of funding?
* Average amount of funding for the seed stage?
* Maximum funding for the seed stage?
* What do we compare the scrape statistics against?

**What Charts?**

1. Funding type sums in a pie chart.
2. Funding type percentages in a pie chart.
3. Funding type investment mean in a bar chart.
4. Average duration of startups.
5. Series D chart of the 59 companies in that phase.

**Parameters for Crunchbase - Not all endpoints are available.**

Founded Date:

Number of Employees:

Location:

Series of Funding:

Closed Date:

Acquired Date:

IPO Date:

#------------------------------------------------------------------------------------------------------------------------------

**Rest API**

Enterprise and Applications licensees have access to the full Crunchbase API.

<https://data.crunchbase.com/v3/reference> - complete endpoint documentation

**Basic Access licensees are limited to the**

* [/odm-organizations](https://data.crunchbase.com/docs/odm-organizations)
* [/odm-people](https://data.crunchbase.com/docs/odm-people)
* [Open Data Map](https://data.crunchbase.com/v3/docs/open-data-map) - Link has documentation for downloading 103 mb odm.csv.tar.gz file.

Enter the url below in a browser, I used Firefox to download the file.

https://api.crunchbase.com/v/3/odm/odm.csv.tar.gz?user\_key=8a3efb5b5136f95bff0593ebb94994e3

* For Windows open Git-Bash, cd to the directory with the file, enter the tar command, and it creates two csv files. For Mac users the file should expand when opened.
* Chrisg@Chrisg-PC MINGW64 ~/bootcampclasswork/Project-1/data (master)
* $ tar -xzvf odm.csv.tar.gz
* organizations.csv
* people.csv
* crunchbase\_license.txt

For simple testing / inspection, the Crunchbase API can also be accessed through any web browser capable of displaying JSON.

All API calls require a user key which is emailed to you following registration. If you lose your key, contact data@crunchbase.com.

**Register for Crunchbase and receive a Rest API user key:**

All API calls require a user key which is emailed to you following registration.

If you lose your key, contact data@crunchbase.com.

**Authentication**

The Crunchbase API uses token-based authentication, which means that developers

must pass their individual user\_key parameter with every request. An example

request for the /organizations endpoint would read as follows:

**Example URLs**

url = "https://api.crunchbase.com/v/3/odm-organizations?user\_key="

url = "https://api.crunchbase.com/v/3/odm-people?user\_key="

Use a for x in range(1,11): # to get 10 pages of data

# Build query URL

query\_url = url + api\_key + "&page=" + str(x)

for i in range(0,len(org\_json["data"]["items"])):

oname = org\_json["data"]["items"][i]["properties"]["name"]

This user\_key is used to validate each developer's access to the API and ensure

that any rate limits or quotas are respected. If you omit the user\_key, overrun

a rate limit, or exceed a quota, the system will respond with an error message.

**Using the REST API**

We encourage you to leverage the API for your internal business and research needs. Unless otherwise noted in your license (see [License Agreement](https://data.crunchbase.com/v3/docs/license) and [Data Access Terms](https://data.crunchbase.com/v3/docs/terms)), you may not license, sublicense, sell, offer to sell, distribute or otherwise provide any of the Crunchbase data to any third parties.

**Attribution Requirements**

When sharing information about the data, follow these attribution requirements.

* Attribution must be accompanied by a hyperlink to Crunchbase
* If content being attributed is primarily about one entity, link must point to that entity on Crunchbase.
* Link must be plainly visible to end user
* Link must be in close proximity to the attributed data
* Link must be visible to Internet spiders (e.g. Googlebot)
* Link cannot include the “nofollow” tag"

**Access Crunchbase API using Google Sheets**  
<http://www.benlcollins.com/apps-script/crunchbase/>

**Crunchbase Getting Started**

<https://data.crunchbase.com/docs>

**Getting a API Key**

From the link above, on the **Crunchbase Getting Started** page in the Basic Access section, click on the " Sign up for Crunchbase Basic [here](https://about.crunchbase.com/forms/apply-basic-access/).", then click on the Get Started button, fill out the form, check the agree to terms box, and click on Submit. I received an email with the key in a couple of minutes.

**chrisg Crunchbase API key - Basic Access**

8a3efb5b5136f95bff0593ebb94994e3

**API Key Crunchbase**

c7601d6088382d4848d12f30b6baeba4

**Pro Access Crunchbase**

username: [rmwalker512@yahoo.com](mailto:rmwalker512@yahoo.com)

Password: UCBd!t@

**Complete Node List**

Crunchbase makes a CSV export of all node keys available for developers seeking an

alternative to paginating through the collection endpoints. The export is updated

each morning and includes separate files for each key node type and includes an

updated\_at timestamp with each key. Crunchbase Enterprise and/or Crunchbase

Platform is required to download the CSV files.

The CSV export is a compressed TAR file containing the following files:

organizations.csv - All organization names and permalinks

people.csv - All people names and permalinks

funding\_rounds.csv - All funding round UUIDs

investors.csv - All investor UUIDs

investments.csv - All investment UUIDs

acquisitions.csv - All acquisition UUIDs

ipos.csv - All IPO UUIDs

funds.csv - All fund UUIDs

**Accessing the Node Export**

See this link: <https://data.crunchbase.com/v3/docs/node-export>

To access the CSV files, follow these steps:

Locate your User Key. The user key is a 32 character string that you should have

received by email after signing up for Crunchbase data. If you have not received

your user key, please contact support.

Enter URL Into Your Browser. Once you’ve located your user key, type the following

URL address into your browser and replacing "user\_key" with the value that you

received in your email.

https://api.crunchbase.com/v/3/node\_keys/node\_keys.tar.gz?user\_key=user\_key

For example, if your "user\_key" is 1a2b3c4d5e6f7g8h9i0j, you would type:

https://api.crunchbase.com/v/3/node\_keys/node\_keys.tar.gz?user\_key=1a2b3c4d5e6f7g8h9i0j

Press Enter Key to Begin Download. Once the correct URL is in the address bar,

press “Enter” (or "Return") and the download should begin automatically. If the

download does not begin, please contact support.

**Accessing the CSV Export**

See this link: <https://data.crunchbase.com/v3/docs/daily-csv-export>

The CSV export is a compressed TAR file containing the following files:

* **organizations.csv** - All organizations in Crunchbase
* **people.csv** - All people in Crunchbase
* **funding\_rounds.csv** - Detail for each funding round in the dataset
* **investors.csv** - Active investors including organizations and individuals
* **investments.csv** - Mapping between investors and investments
* **investment\_partners.csv** - Partners who are responsible for their firm's investments
* **acquisitions.csv** - Detail for each acquisition in the dataset
* **ipos.csv** - Detail for each IPO in the dataset
* **org\_parents.csv** - Parent-child mapping for each organization
* **organization\_descriptions.csv** - Long descriptions for organizations
* **customers.csv** - List of customers for each organization
* **competitors.csv** - List of competitors for each organization
* **jobs.csv** - List of all job and advisory roles
* **events.csv** - Detail for each event in the dataset
* **event\_relationship.csv** - Detail for each event participant in the dataset
* **category\_groups.csv** - Mapping between categories and category groups

To access the CSV files, follow these steps:

**Locate your User Key.** The user key is a 32 character string that you should have received by email after signing up for Crunchbase data. If you have not received your user key, please contact support.

**Enter URL Into Your Browser.** Once you’ve located your user key, type the following URL address into your browser and replacing "user\_key" with the value that you received in your email.

https://api.crunchbase.com/v/3/csv\_export/csv\_export.tar.gz?user\_key=user\_key

For example, if your "user\_key" is 1a2b3c4d5e6f7g8h9i0j, you would type:

https://api.crunchbase.com/v/3/csv\_export/csv\_export.tar.gz?user\_key=1a2b3c4d5e6f7g8h9i0j

**Press Enter Key to Begin Download.** Once the correct URL is in the address bar, press “Enter” (or "Return") and the download should begin automatically. If the download does not begin, please contact support.

**Cross Reference with Jobs posted for start up - Is this possible?**

<https://www.programmableweb.com/category/jobs/apis?category=20080>

What is the prediction of Growth: Employees #?

What is the a measure of Growth: Number of Job Postings?