ECS 171

Data analytics

Goal: Parse and select columns from the crunchbase.companies table

We based our decision processes on

Finding the most important columns of data in terms of analysis
Weeding out the data that's inconsistent or repeated

Columns we deleted

Column	**Description**
permalink	Not relevant to analysis
homepage_url	Not relevant to analysis
category_list	This contains extra data that might be useful, but difficult to analyze or to categorize in discrete values because it contains multiple values; we used market instead, which only contains one particular category e.g. |Entertainment|Politics|Social Media|News|
city	Generally metropolitan areas is good enough to indicate the geographical area
founded_month *founded_quarter* *founded_year*	Redundant, and far less granular than the founded_at column which contains the exact day the company was founded
state_codes	Anything with this not blank is basically within the United States, so we don't actually need this column to determine the geographical location of a company Also redundant with the city and even country

Columns that will act as keys

Column	**Description**
name	Uniquely identifies the name of each startup

Rows we are removing

Anything missing data for column funding_total_usd – this is an important metric of the success or at least the funding status of a company

Anything missing data for column founded_at because we should know how much time has elapsed between its funding date and its founding time

Anything missing the key (name)

Anything missing data for column region – we used the country column to determine the general geographical areas each startup belongs to

Questions? Contact Jason or Philson

Code

PostgreSQL on ModeAnalytics

```
SELECT c.name, c.market, c.funding_total_usd AS funding_amount, c.status,
c.country_code AS country, c.region, c.funding_rounds, extract(epoch from
c.founded_at) AS founded_at, extract(epoch from c.first_funding_at) AS
first_funding_at, extract(epoch from c.last_funding_at) AS last_funding_at

FROM crunchbase.companies AS c

WHERE market IS NOT null AND funding_total_usd != ' -    ' AND founded_at IS NOT null
AND status IS NOT null AND region IS NOT null


ORDER BY name
```