Y Combinator Startup Database

Team A7

Yanan Sun, Xinru Zhao, Wenqi Wang, Gaohan Lin

Agenda

1 **Background**

Introduce Y Combinator

2 **Dataset**

Introduce Metadata of Entity and Attributes

3 **ER Diagram**

Show ERD and relationships

4 **Relational Data Model**

Show ERD with PKs and FKs

5 **As-is Dependency Diagrams**

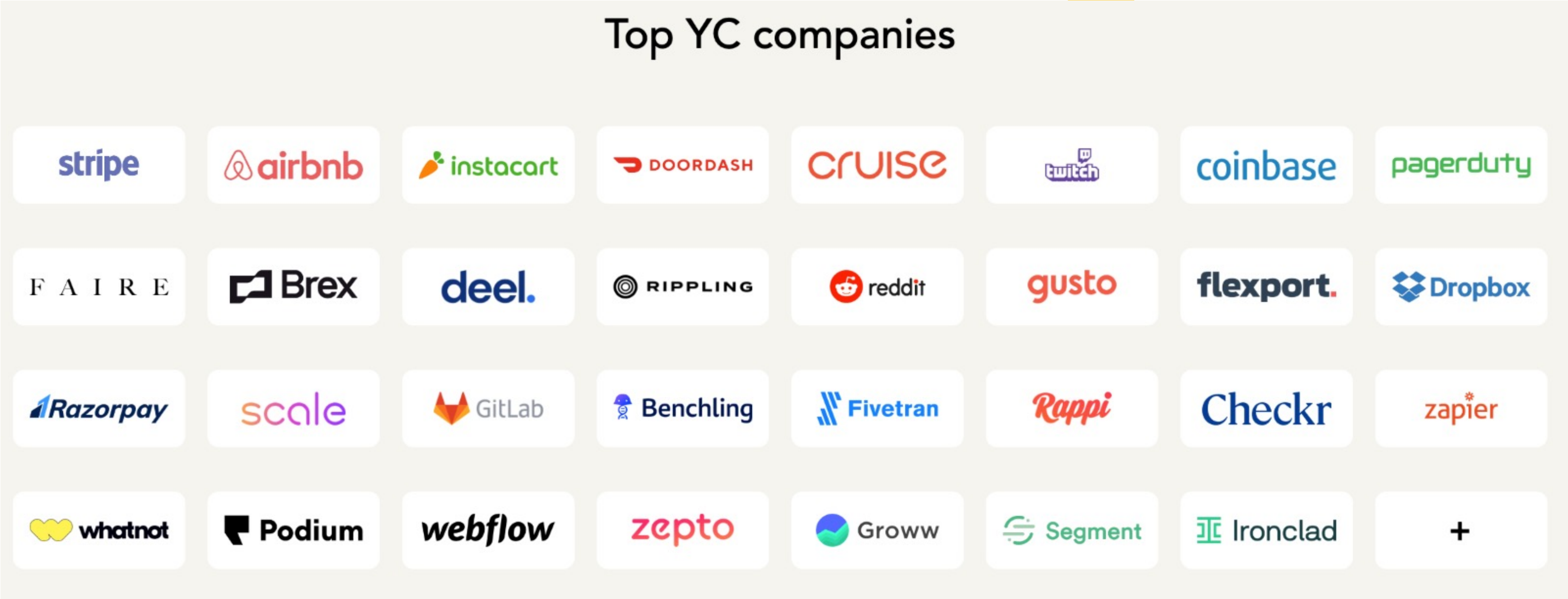
Show in 3rd normal way

6 **SQL Query & Analysis**

The importance of finding

Y Combinator

A driving force behind the startup ecosystem, shaping the future of innovation.



Top YC companies

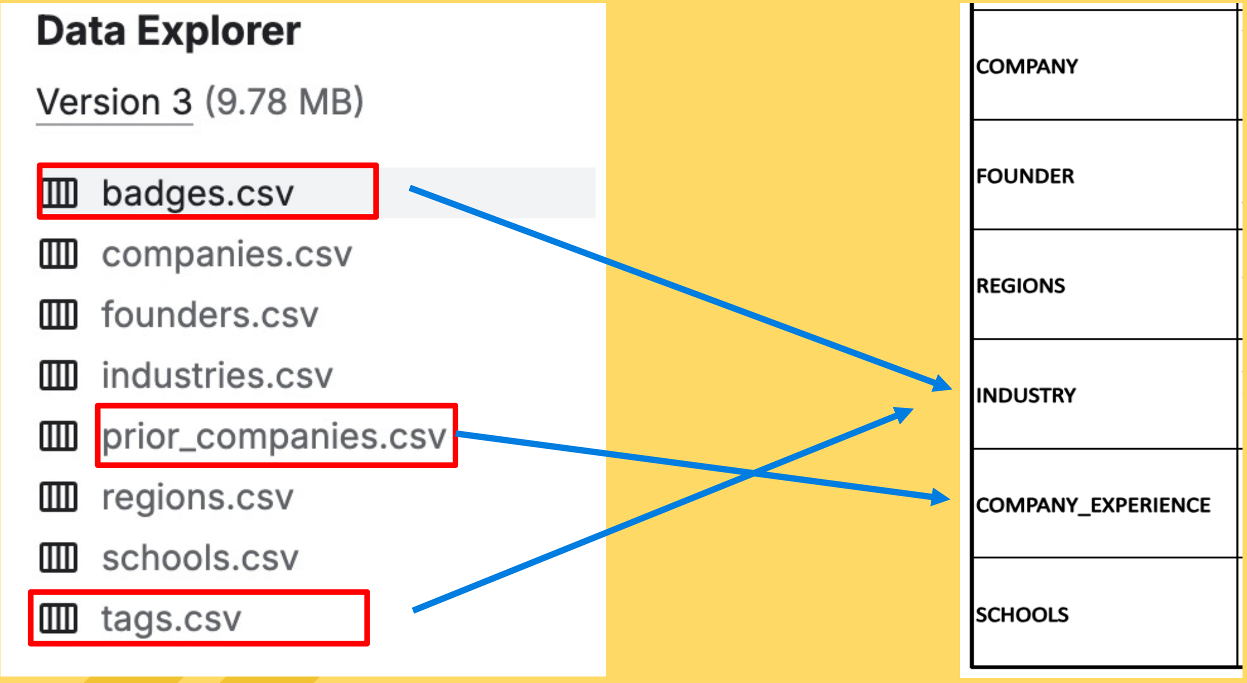
Which Company/Organization will benefit from this dataset?

a. **Venture Capital Firms & Angel Investors**: They could use this data to identify successful patterns among YC startups, analyze historical performance by batch, track founder backgrounds, and discover promising investment opportunities.

b. **Market Research Companies**: Organizations that analyze startup ecosystems could use this data to create industry reports, track trends, and provide insights on the evolution of YC companies.

c. **Enterprise Companies Looking for Acquisitions**: Large tech firms seeking to acquire startups could use this database to identify potential acquisition targets.

Dataset &Preprocessing



Data Explorer

Version 3 (9.78 MB)

- badges.csv
- companies.csv
- founders.csv
- industries.csv
- prior_companies.csv
- regions.csv
- schools.csv
- tags.csv

COMPANY
FOUNDER
REGIONS
INDUSTRY
COMPANY_EXPERIENCE
SCHOOLS

Metadata-Entities

Entities and Definitions

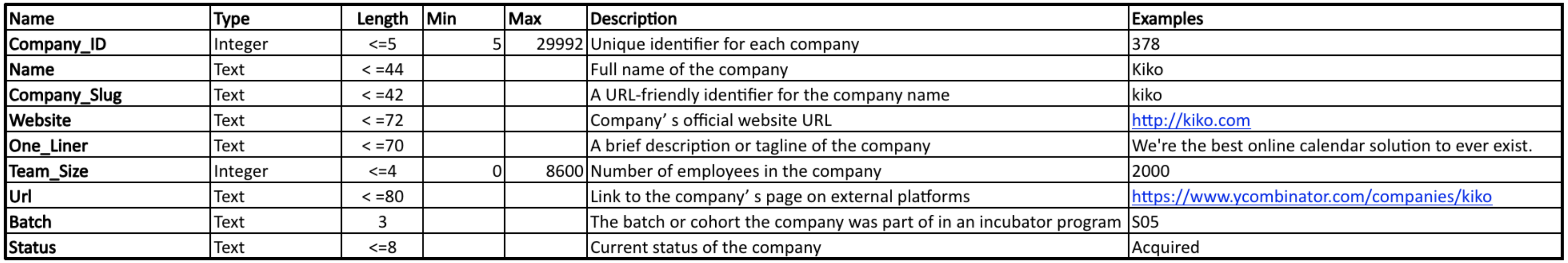
Entity	Definition	Example
COMPANY	Y Combinator startups, including their identifiers, descriptions, team sizes, batch participation, and current status.	E.g. Reddit (Company ID: 379), known as “the front page of the internet,” participated in Y Combinator’s S05 batch, had a team size of 2000, and was later acquired.
FOUNDER	Startup founders, including their names, unique identifiers, current working companies, current job titles, company foundered.	E.g. Alexis Ohanian, co-founder of Reddit (a top Y Combinator company), is currently a General Partner at Initialized Capital.
REGIONS	A specific geographical area where a startup operates or is headquartered.	E.g. Company ID 380 is located in London, UK, in Europe.
INDUSTRY	A category that describes the primary business sector in which a startup operates, such as healthcare, fintech, or e-commerce.	E.g. Company ID 5, operating in the Industrials industry with a focus on Manufacturing and Robotics, is hiring.
COMPANY_EXPERIENCE	Companies experience that Y Combinator startup previously had before their current role.	E.g HN_ID 110 had company experience working in Khoj, Microsoft and Hillhacks before starting own enterprise.
SCHOOLS	Educational institutions where Y Combinator startups receive formal education, including primary schools, secondary schools, universities, and specialized training institutions.	E.g HN_ID 110 studied Electronics and Instrumentation at Birla Institute of Technology and Science from 2009 to 2014, completing a 6-year program.

Key Conceptual Elements

Key Entities	Key Attributes	Relationship
COMPANY	Company_Name Status	A COMPANY is belong to exactly one INDUSTRY. A INDUSTRY has one or more COMPANY.
INDUSTRY	Business_Model Industry Sub_Industry	

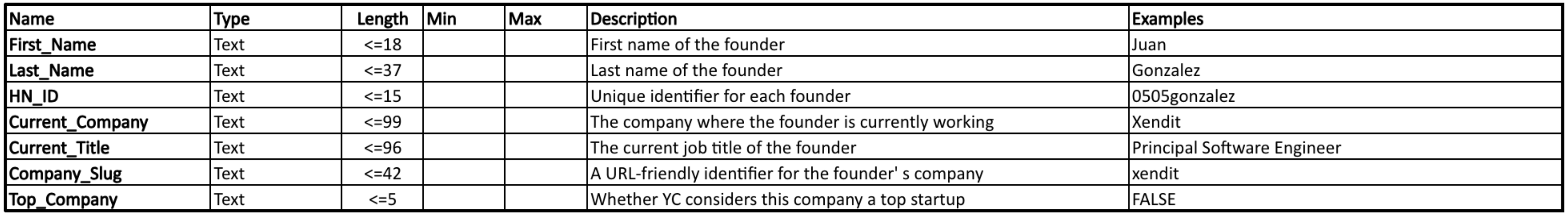
Metadata-Attributes

COMPANY



Name	Type	Length	Min	Max	Description	Examples
Company_ID	Integer	<=5	5	29992	Unique identifier for each company	378
Name	Text	< =44			Full name of the company	Kiko
Company_Slug	Text	< =42			A URL-friendly identifier for the company name	kiko
Website	Text	< =72			Company’ s official website URL	http://kiko.com
One_Liner	Text	< =70			A brief description or tagline of the company	We're the best online calendar solution to ever exist.
Team_Size	Integer	<=4	0	8600	Number of employees in the company	2000
Url	Text	< =80			Link to the company’ s page on external platforms	https://www.ycombinator.com/companies/kiko
Batch	Text	3			The batch or cohort the company was part of in an incubator program	S05
Status	Text	<=8			Current status of the company	Acquired

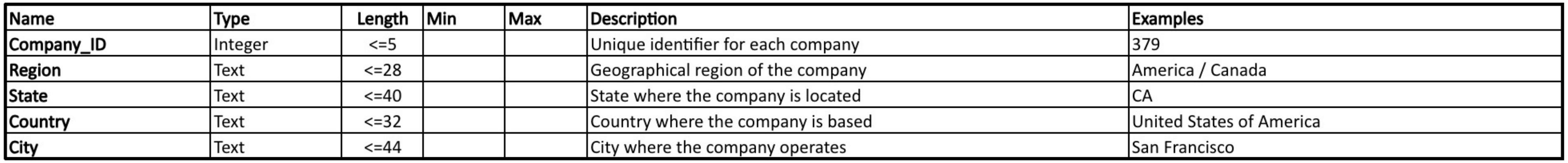
FOUNDER



Name	Type	Length	Min	Max	Description	Examples
First_Name	Text	<=18			First name of the founder	Juan
Last_Name	Text	<=37			Last name of the founder	Gonzalez
HN_ID	Text	<=15			Unique identifier for each founder	0505gonzalez
Current_Company	Text	<=99			The company where the founder is currently working	Xendit
Current_Title	Text	<=96			The current job title of the founder	Principal Software Engineer
Company_Slug	Text	<=42			A URL-friendly identifier for the founder' s company	xendit
Top_Company	Text	<=5			Whether YC considers this company a top startup	FALSE

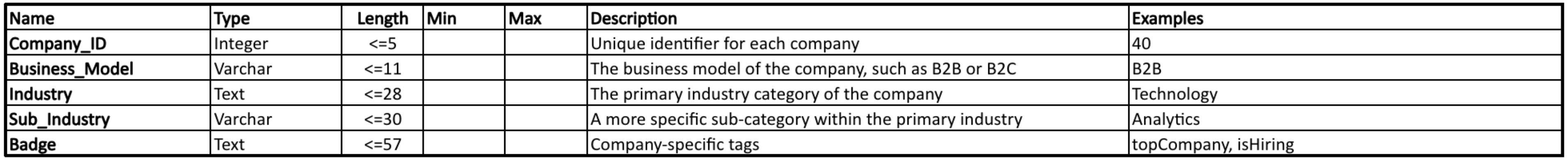
Metadata-Attributes

REGIONS



Name	Type	Length	Min	Max	Description	Examples
Company_ID	Integer	<=5			Unique identifier for each company	379
Region	Text	<=28			Geographical region of the company	America / Canada
State	Text	<=40			State where the company is located	CA
Country	Text	<=32			Country where the company is based	United States of America
City	Text	<=44			City where the company operates	San Francisco

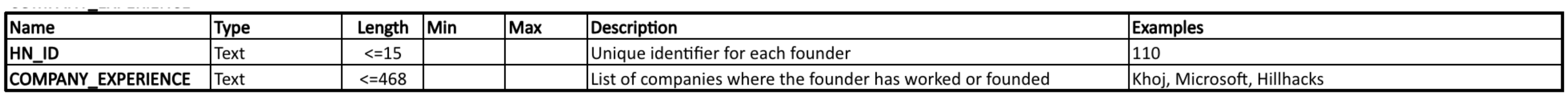
INDUSTRIES



Name	Type	Length	Min	Max	Description	Examples
Company_ID	Integer	<=5			Unique identifier for each company	40
Business_Model	Varchar	<=11			The business model of the company, such as B2B or B2C	B2B
Industry	Text	<=28			The primary industry category of the company	Technology
Sub_Industry	Varchar	<=30			A more specific sub-category within the primary industry	Analytics
Badge	Text	<=57			Company-specific tags	topCompany, isHiring

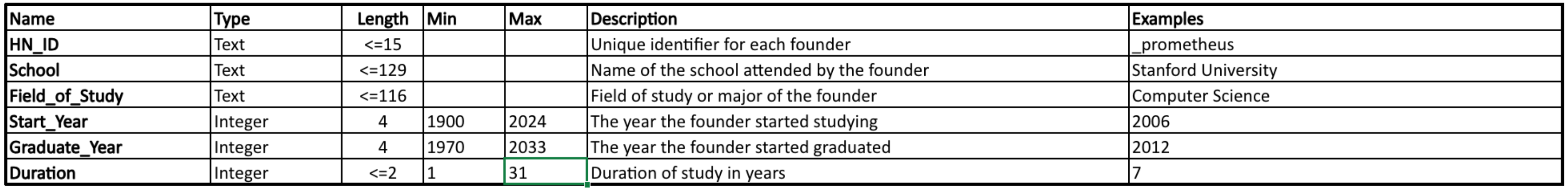
Metadata-Attributes

COMPANIES_EXPERIENCE



Name	Type	Length	Min	Max	Description	Examples
HN_ID	Text	<=15			Unique identifier for each founder	110
COMPANY_EXPERIENCE	Text	<=468			List of companies where the founder has worked or founded	Khoj, Microsoft, Hillhacks

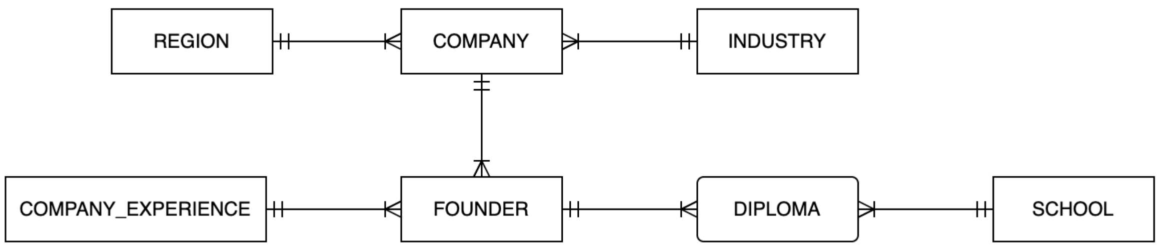
SCHOOLS



Name	Type	Length	Min	Max	Description	Examples
HN_ID	Text	<=15			Unique identifier for each founder	_prometheus
School	Text	<=129			Name of the school attended by the founder	Stanford University
Field_of_Study	Text	<=116			Field of study or major of the founder	Computer Science
Start_Year	Integer	4	1900	2024	The year the founder started studying	2006
Graduate_Year	Integer	4	1970	2033	The year the founder started graduated	2012
Duration	Integer	<=2	1	31	Duration of study in years	7

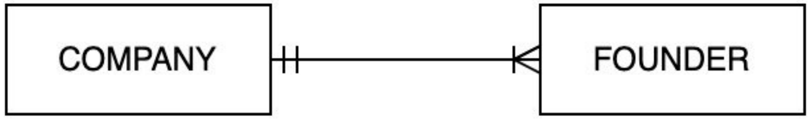
ER Diagram & Relationship

Entity	Relationship
COMPANY - FOUNDER	A COMPANY is founded by one or many FOUNDERS. Each FOUNDER founds exactly one COMPANY.
COMPANY - REGION	A COMPANY is located at exactly one REGION. EACH REGION has one or many COMPANY
COMPANY - INDUSTRY	A COMPANY is belong to exactly one INDUSTRY. A INDUSTRY has one or more COMPANY.
FOUNDERS - COMPANY_EXPERIENCE	A FOUNDER has one or many COMPANY_EXPERIENCE. A COMPANY_EXPERIENCE is associated with one or many FOUNDERS.
FOUNDERS - SCHOOLS	A FOUNDER has attended one or many SCHOOLS. A SCHOOL have one or many FOUNDERS.



ER Diagram, Relationship, Business Rule

Entity	**Relationship**
COMPANY - FOUNDER	A COMPANY is founded by one or many FOUNDERS. Each FOUNDER founds exactly one COMPANY.



COMPANY

FOUNDER

Business Rule:

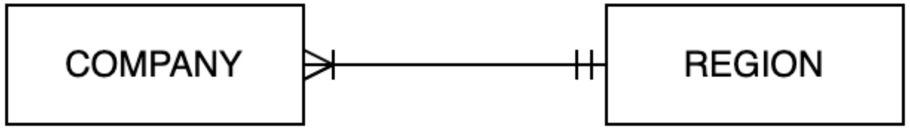
For Company entity:

Each **Company** must have a unique **Company_ID** that serves as its identifier.

This rule causes any attempt to insert a **Company** without a **Company_ID** or with a duplicate **Company_ID** to be rejected.

ER Diagram, Relationship, Business Rule

Entity	Relationship
COMPANY - REGION	A COMPANY is located at exactly one REGION. EACH REGION has one or many COMPANY

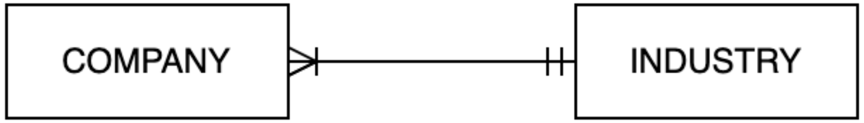


Business Rule:

For Region entity:
Each **Company** must belong to exactly one **Region,and Each Region must have an associated State, Country, and City**
This rule enforces that no company can exist without being assigned a **Region,and ensures complete geographical data for every company on Ycombiantor.**

ER Diagram, Relationship, Business Rule

Entity	Relationship
COMPANY - INDUSTRY	A COMPANY is belong to exactly one INDUSTRY. A INDUSTRY has one or more COMPANY.



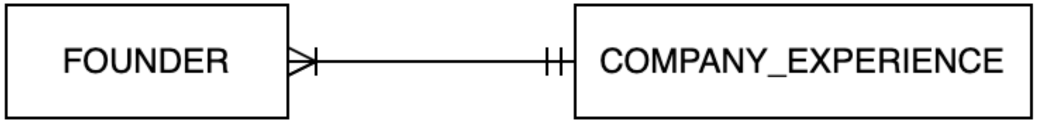
Business Rule:

For Industry entity:

Each **Industry** may have multiple **Sub_Industries**, but a **Sub_Industry** must belong to exactly one **Industry**. This rule prevents a **Sub_Industry** from existing without being associated with an **Industry**.

ER Diagram, Relationship, Business Rule

Entity	Relationship
FOUNDERS - COMPANY_EXPERIENCE	A FOUNDER has one or many COMPANY_EXPERIENCE. A COMPANY_EXPERIENCE is associated with one or many FOUNDERS.



Business Rule:

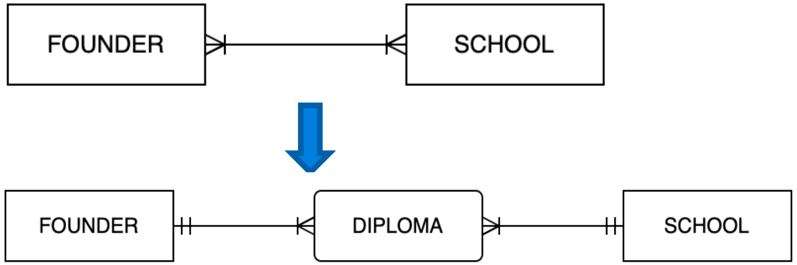
For Founder entity:

Each **Founder** must have a unique **HN_ID** assigned to them.

This rule prevents any registration from inserting a **Founder** without an **HN_ID** or with a duplicate **HN_ID**.

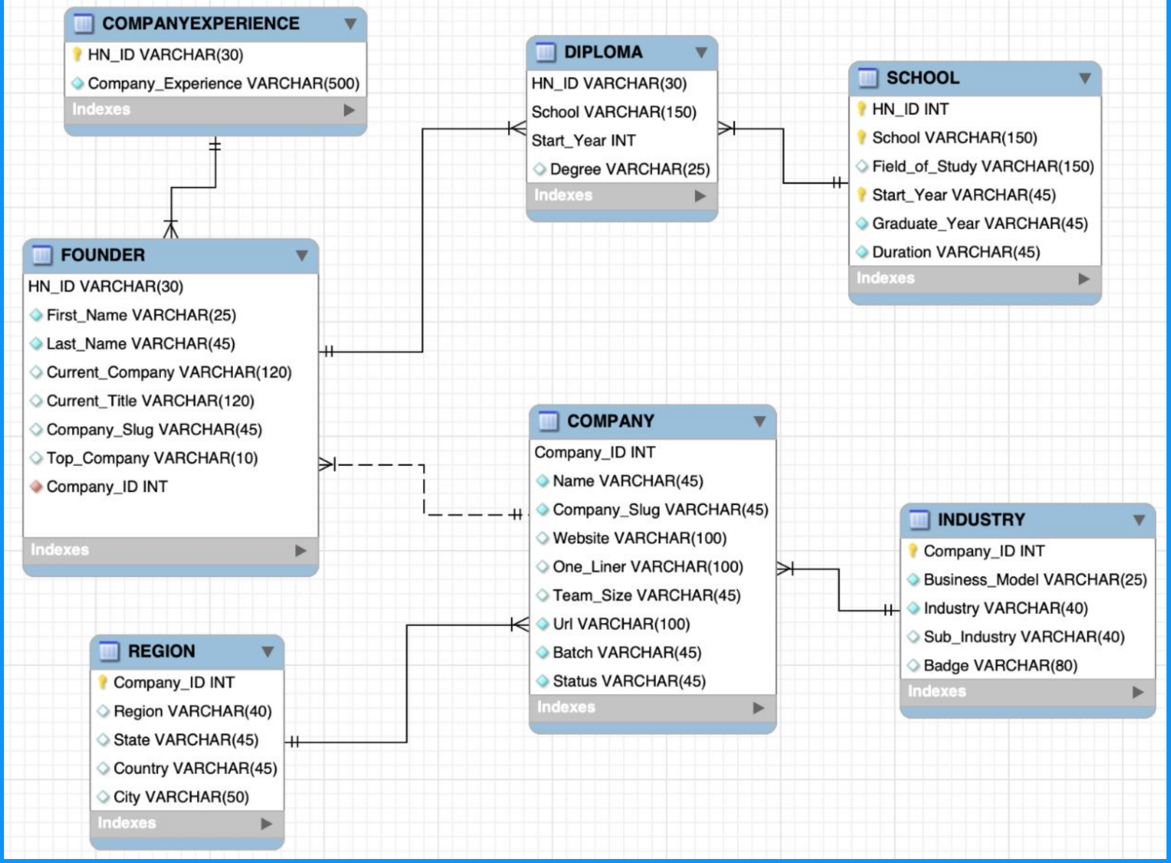
ER Diagram, Relationship, Business Rule

Entity	Relationship
FOUNDERS - SCHOOLS	A FOUNDER has attended one or many SCHOOLS. A SCHOOL have one or many FOUNDERS.



Entity	Relationship
FOUNDERS - SCHOOLS	A FOUNDER can obtain one or many DIPLOMA. A SCHOOL can issue one or many DIPLOMAS to different FOUNDER. A DIPLOMA is associated with one and only one SCHOOL and FOUNDER.

Relational Data Model



COMPANYEXPERIENCE
HN_ID VARCHAR(30)
Company_Experience VARCHAR(500)
Indexes

DIPLOMA
HN_ID VARCHAR(30)
School VARCHAR(150)
Start_Year INT
Degree VARCHAR(25)
Indexes

SCHOOL
HN_ID INT
School VARCHAR(150)
Field_of_Study VARCHAR(150)
Start_Year VARCHAR(45)
Graduate_Year VARCHAR(45)
Duration VARCHAR(45)
Indexes

FOUNDER
HN_ID VARCHAR(30)
First_Name VARCHAR(25)
Last_Name VARCHAR(45)
Current_Company VARCHAR(120)
Current_Title VARCHAR(120)
Company_Slug VARCHAR(45)
Top_Company VARCHAR(10)
Company_ID INT
Indexes

COMPANY
Company_ID INT
Name VARCHAR(45)
Company_Slug VARCHAR(45)
Website VARCHAR(100)
One_Liner VARCHAR(100)
Team_Size VARCHAR(45)
Url VARCHAR(100)
Batch VARCHAR(45)
Status VARCHAR(45)
Indexes

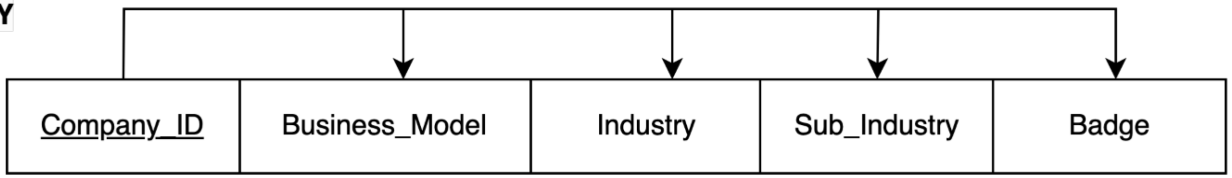
INDUSTRY
Company_ID INT
Business_Model VARCHAR(25)
Industry VARCHAR(40)
Sub_Industry VARCHAR(40)
Badge VARCHAR(80)
Indexes

REGION
Company_ID INT
Region VARCHAR(40)
State VARCHAR(45)
Country VARCHAR(45)
City VARCHAR(50)
Indexes

As-is Dependency Diagrams

Full dependency

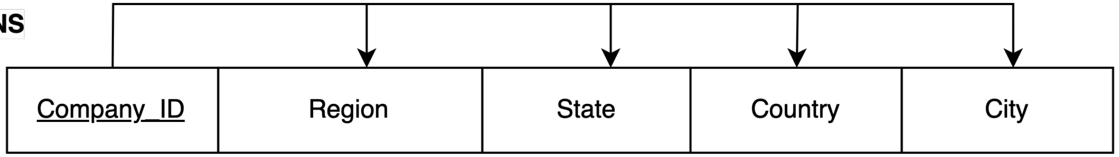
INDUSTRY



Company_ID	Business_Model	Industry	Sub_Industry	Badge

Full dependency

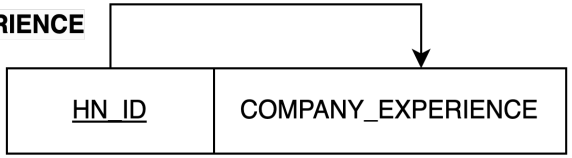
REGIONS



Company_ID	Region	State	Country	City

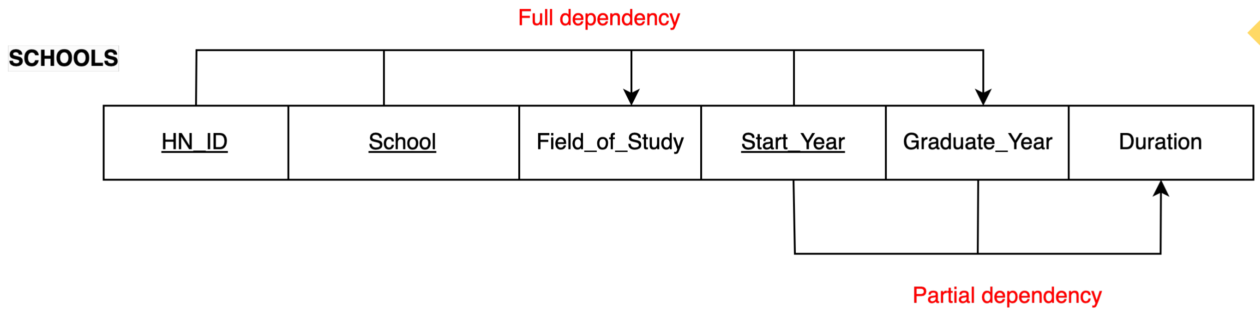
Full dependency

COMPANY EXPERIENCE



HN_ID	COMPANY_EXPERIENCE

As-is Dependency Diagrams

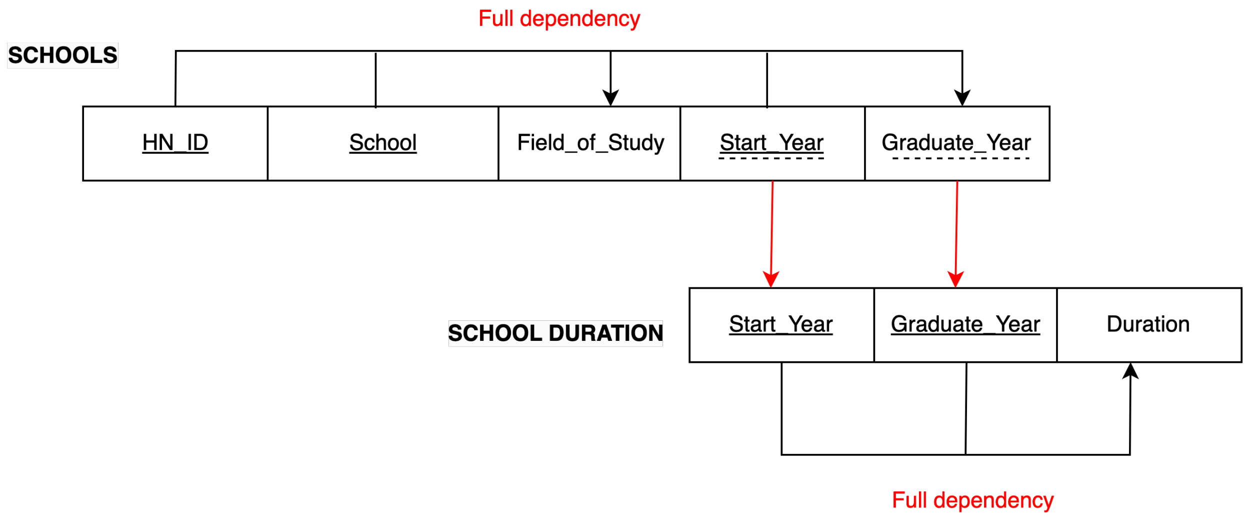


Full dependency

SCHOOLS

HN_ID	School	Field_of_Study	Start_Year	Graduate_Year	Duration

Partial dependency



Full dependency

SCHOOLS

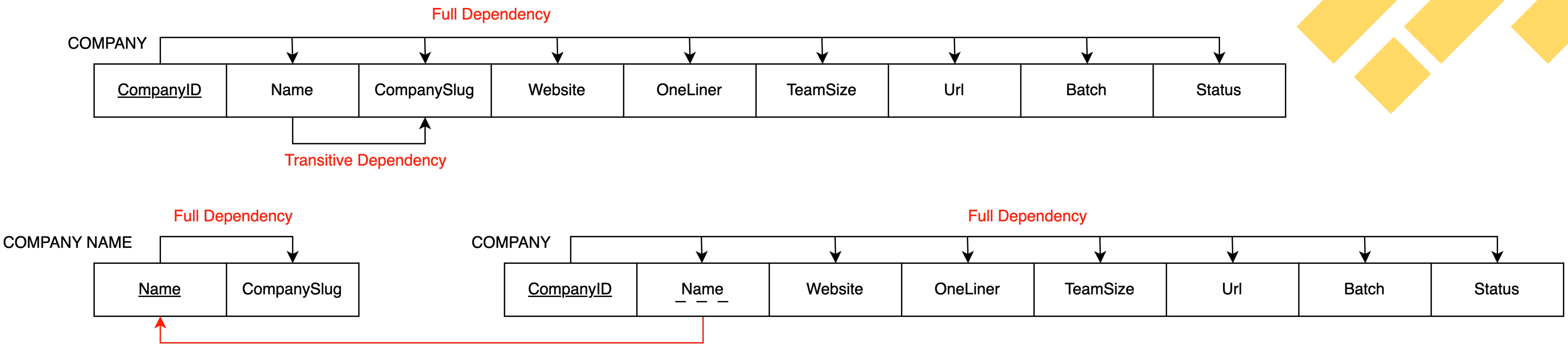
HN_ID	School	Field_of_Study	Start_Year	Graduate_Year

SCHOOL DURATION

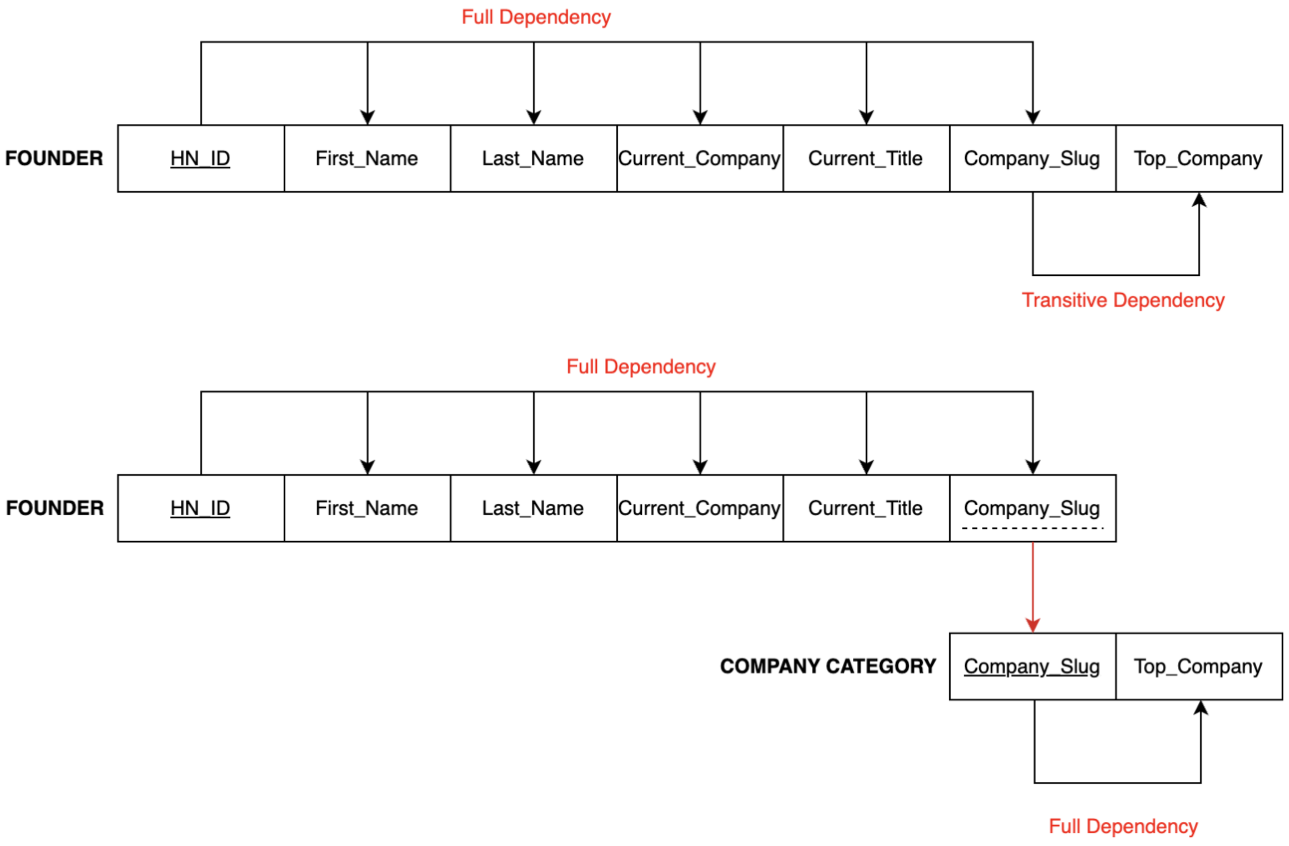
Start_Year	Graduate_Year	Duration

Full dependency

As-is Dependency Diagrams



As-is Dependency Diagrams



Full Dependency

FOUNDER	<u>HN_ID</u>	First_Name	Last_Name	Current_Company	Current_Title	Company_Slug	Top_Company

Transitive Dependency

Full Dependency

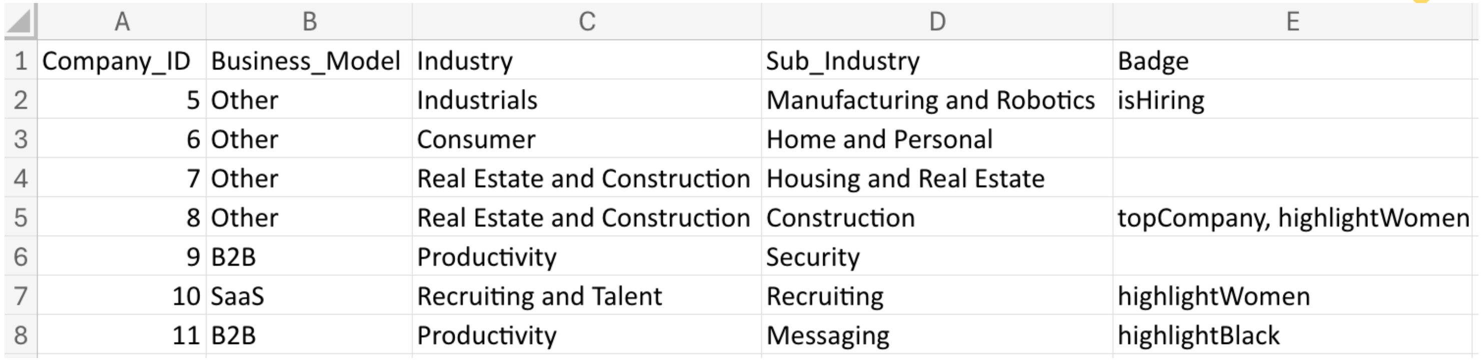
FOUNDER	<u>HN_ID</u>	First_Name	Last_Name	Current_Company	Current_Title	Company_Slug

COMPANY CATEGORY	<u>Company_Slug</u>	Top_Company

Full Dependency

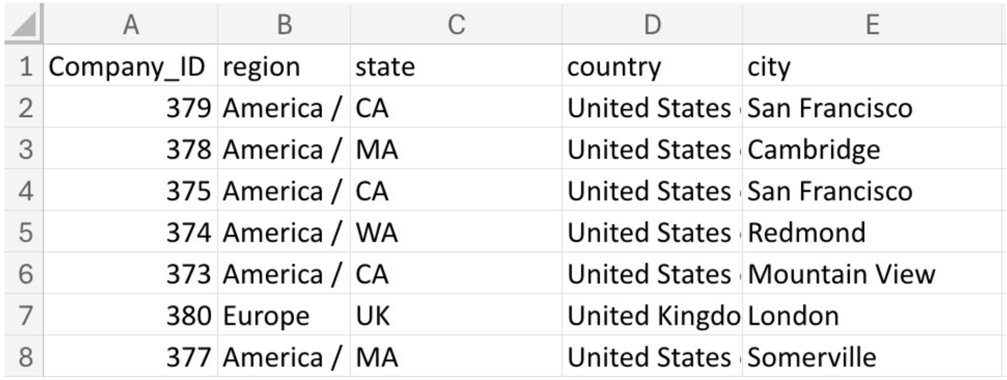
As-is Dependency Diagrams

INDUSTRY:



	A	B	C	D	E
1	Company_ID	Business_Model	Industry	Sub_Industry	Badge
2	5	Other	Industrials	Manufacturing and Robotics	isHiring
3	6	Other	Consumer	Home and Personal	
4	7	Other	Real Estate and Construction	Housing and Real Estate	
5	8	Other	Real Estate and Construction	Construction	topCompany, highlightWomen
6	9	B2B	Productivity	Security	
7	10	SaaS	Recruiting and Talent	Recruiting	highlightWomen
8	11	B2B	Productivity	Messaging	highlightBlack

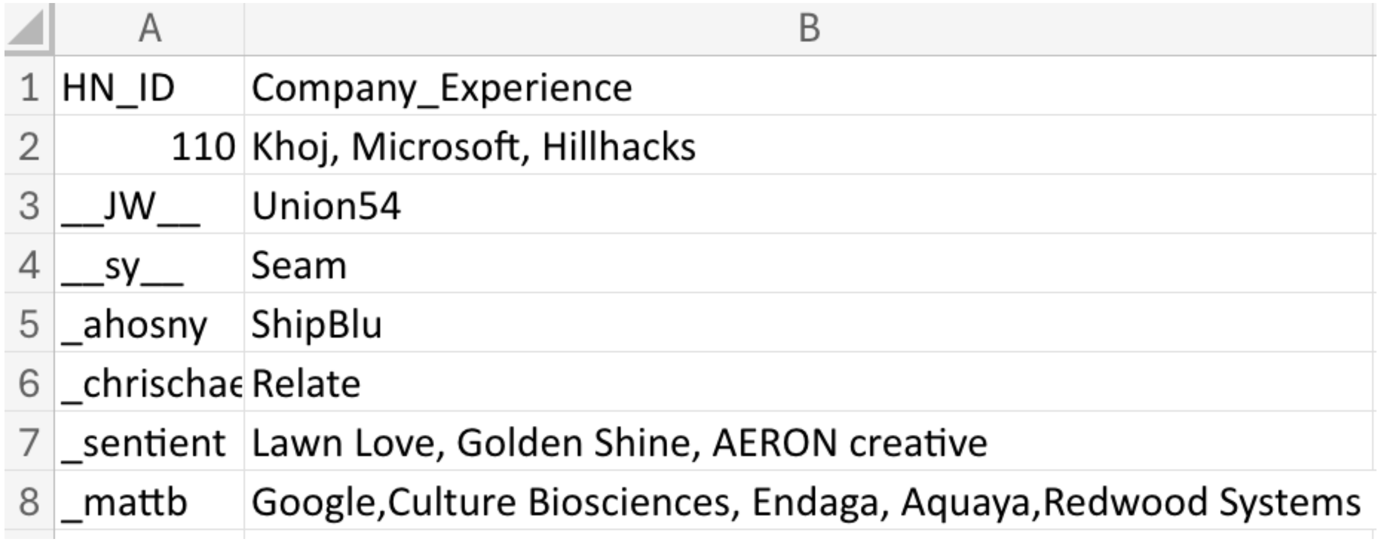
REGIONS:



	A	B	C	D	E
1	Company_ID	region	state	country	city
2	379	America /	CA	United States	San Francisco
3	378	America /	MA	United States	Cambridge
4	375	America /	CA	United States	San Francisco
5	374	America /	WA	United States	Redmond
6	373	America /	CA	United States	Mountain View
7	380	Europe	UK	United Kingdo	London
8	377	America /	MA	United States	Somerville

As-is Dependency Diagrams

COMPANY EXPERIENCE



	A	B
1	HN_ID	Company_Experience
2	110	Khoj, Microsoft, Hillhacks
3	__JW__	Union54
4	__sy__	Seam
5	_ahosny	ShipBlu
6	_chrischae	Relate
7	_sentient	Lawn Love, Golden Shine, AERON creative
8	_mattb	Google,Culture Biosciences, Endaga, Aquaya,Redwood Systems

Analysis with SQL

1

COMPANY STATUS

Focus on Non-Inactive Startups

2

COMPANY INDUSTRY

Company Industry Distribution

3

EDUCATION

Successful Founders' Education Experience

4

REGION

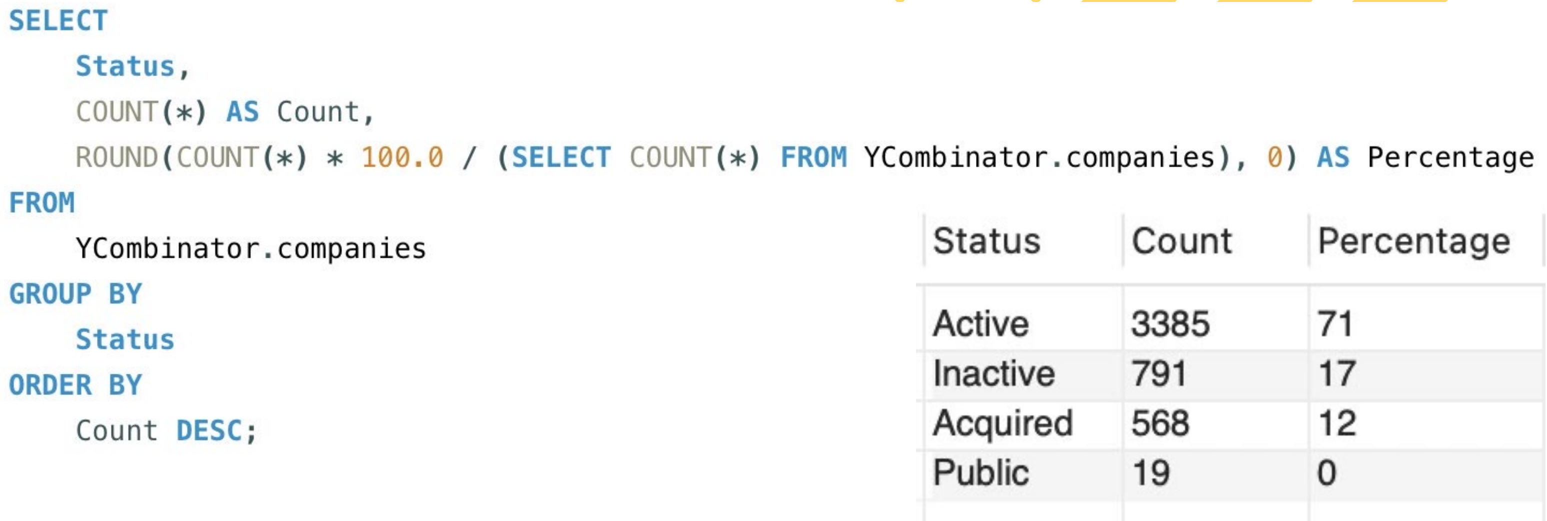
Successful Founders' Region Distribution

5

WORKING EXPERIENCE

Successful Founders' Prior Working Companies

COMPANY STATUS



```sql
SELECT
    Status,
    COUNT(*) AS Count,
    ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM YCombinator.companies), 0) AS Percentage
FROM
    YCombinator.companies
GROUP BY
    Status
ORDER BY
    Count DESC;
```

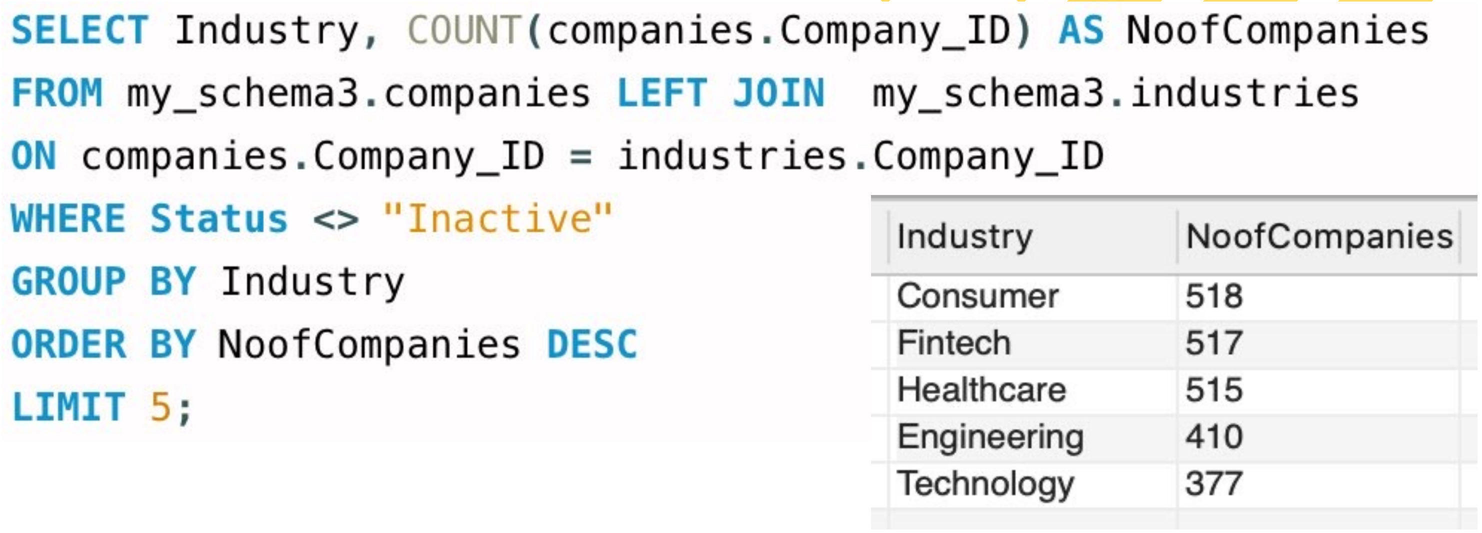
Status	Count	Percentage
Active	3385	71
Inactive	791	17
Acquired	568	12
Public	19	0

Insights:

1. Most companies (71%) remain active---a good overall survival rate.
2. Companies(12%) have been acquired---a successful exit strategy.
3. Only 19 companies(<1%) have achieved public status---high-risk nature of startups
4. Companies(17%) are inactive

Analytic Object: Non-Inactive Companies and their Founders

COMPANY INDUSTRY



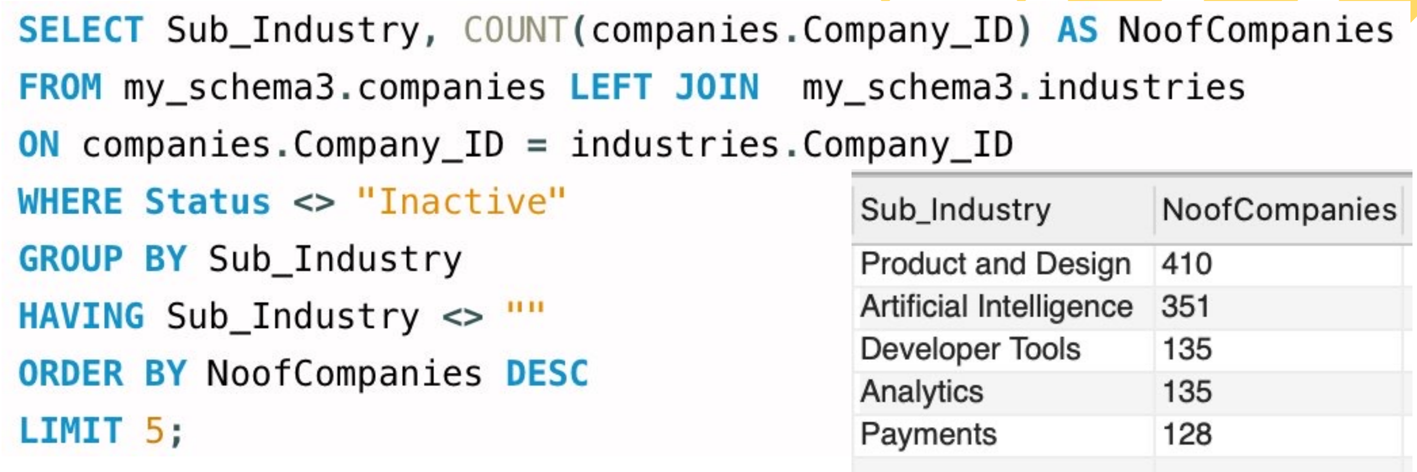
```sql
SELECT Industry, COUNT(companies.Company_ID) AS NoofCompanies
FROM my_schema3.companies LEFT JOIN  my_schema3.industries
ON companies.Company_ID = industries.Company_ID
WHERE Status <> "Inactive"
GROUP BY Industry
ORDER BY NoofCompanies DESC
LIMIT 5;
```

Industry	NoofCompanies
Consumer	518
Fintech	517
Healthcare	515
Engineering	410
Technology	377

Insights:

1. Consumer, Fintech and Healthcare industry(>510) show Significant Presence---attractive for YC's investors
2. Tech-driven industries (Fintech, Technology) dominate the list---ongoing importance of technology in startup ecosystems

COMPANY INDUSTRY



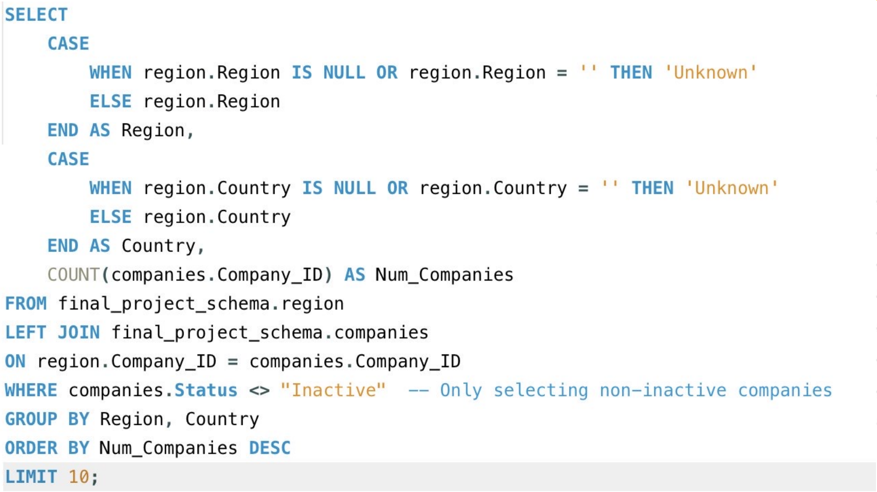
```sql
SELECT Sub_Industry, COUNT(companies.Company_ID) AS NoofCompanies
FROM my_schema3.companies LEFT JOIN  my_schema3.industries
ON companies.Company_ID = industries.Company_ID
WHERE Status <> "Inactive"
GROUP BY Sub_Industry
HAVING Sub_Industry <> ""
ORDER BY NoofCompanies DESC
LIMIT 5;
```

Sub_Industry	NoofCompanies
Product and Design	410
Artificial Intelligence	351
Developer Tools	135
Analytics	135
Payments	128

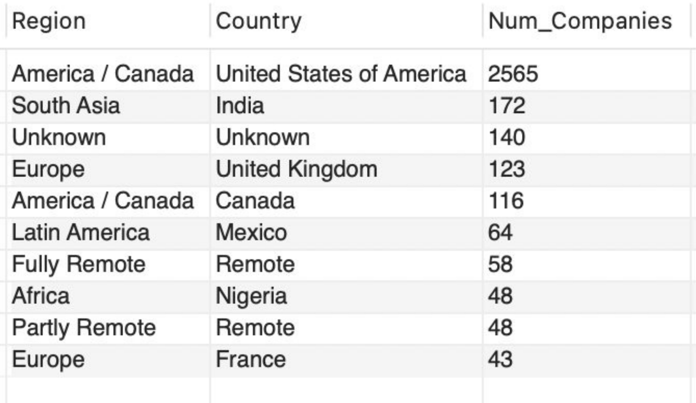
Insights:

1. Product and Design(410) Leads the Sub-Industries---high demand for product innovation and user experience design in startups
2. High potential growth for AI Industry(351)---increasing importance of AI technologies in various sectors

REGION



```
SELECT
    CASE
        WHEN region.Region IS NULL OR region.Region = '' THEN 'Unknown'
        ELSE region.Region
    END AS Region,
    CASE
        WHEN region.Country IS NULL OR region.Country = '' THEN 'Unknown'
        ELSE region.Country
    END AS Country,
    COUNT(companies.Company_ID) AS Num_Companies
FROM final_project_schema.region
LEFT JOIN final_project_schema.companies
ON region.Company_ID = companies.Company_ID
WHERE companies.Status <> "Inactive"  -- Only selecting non-inactive companies
GROUP BY Region, Country
ORDER BY Num_Companies DESC
LIMIT 10;
```

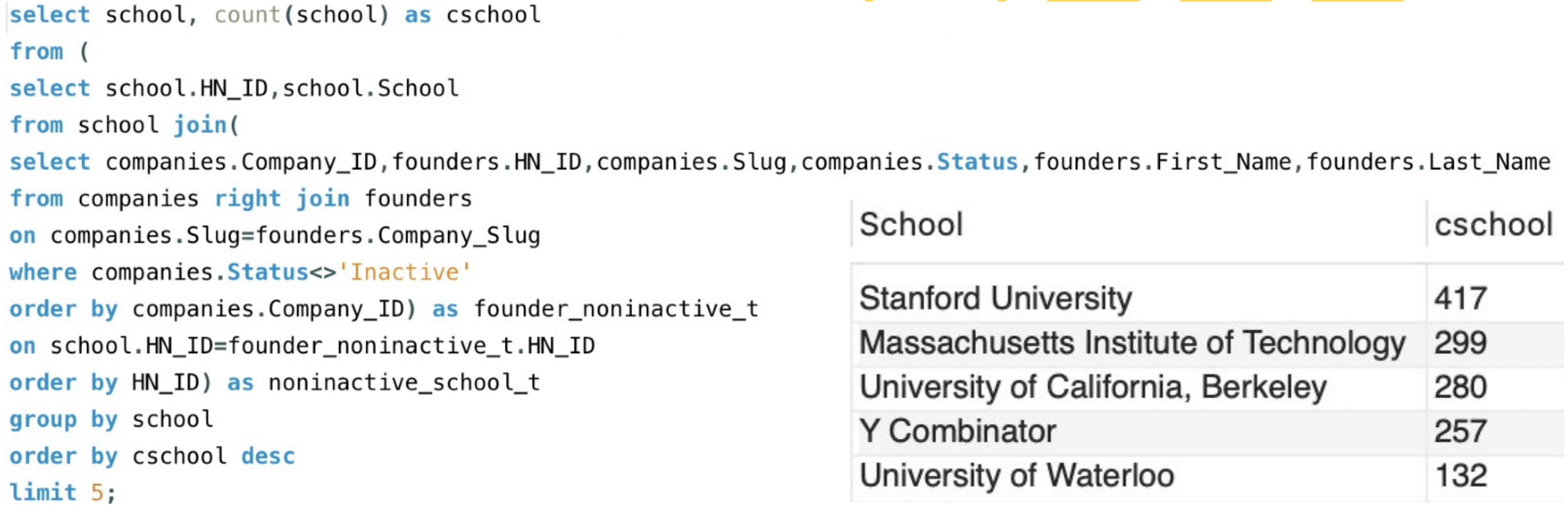


Region	Country	Num_Companies
America / Canada	United States of America	2565
South Asia	India	172
Unknown	Unknown	140
Europe	United Kingdom	123
America / Canada	Canada	116
Latin America	Mexico	64
Fully Remote	Remote	58
Africa	Nigeria	48
Partly Remote	Remote	48
Europe	France	43

Insights:

1. Dominance(2565) of the U.S. in YC-backed Startups, Canada(116) also has increasing startups--- well-developed startup ecosystem of North America
2. United Kingdom(123) & India(172) have Emerging Startup Hubs

EDUCATION



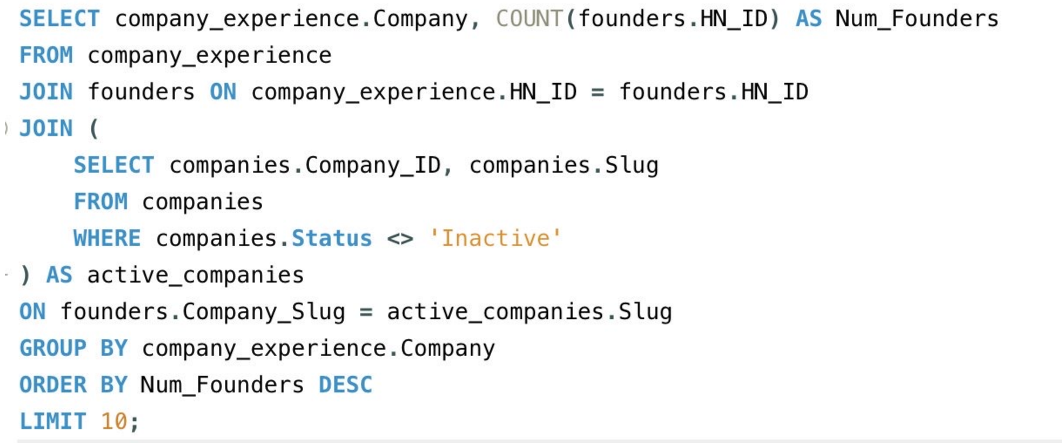
```sql
select school, count(school) as cschool
from (
select school.HN_ID,school.School
from school join(
select companies.Company_ID,founders.HN_ID,companies.Slug,companies.Status,founders.First_Name,founders.Last_Name
from companies right join founders
on companies.Slug=founders.Company_Slug
where companies.Status<>'Inactive'
order by companies.Company_ID) as founder_noninactive_t
on school.HN_ID=founder_noninactive_t.HN_ID
order by HN_ID) as noninactive_school_t
group by school
order by cschool desc
limit 5;
```

School	cschool
Stanford University	417
Massachusetts Institute of Technology	299
University of California, Berkeley	280
Y Combinator	257
University of Waterloo	132

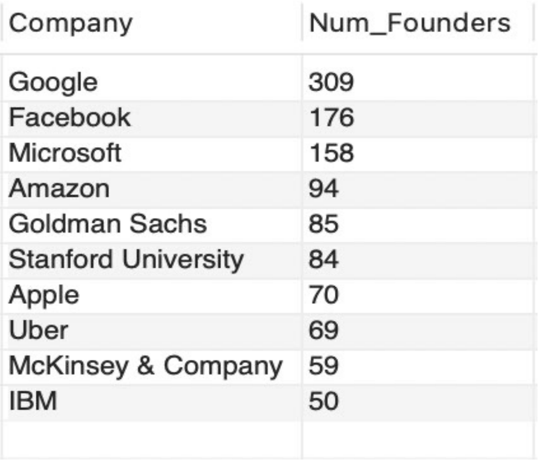
Insights:

1. Stanford University accounts for most(417 founders)---top school for entrepreneurs
2. More than 250 founders have graduated from MIT,UC Bekeley and Y Combinator(course/program)
3. About 132 entrepreneurs have graduated from University of Waterloo

WORKING EXPERIENCE



```
SELECT company_experience.Company, COUNT(founders.HN_ID) AS Num_Founders
FROM company_experience
JOIN founders ON company_experience.HN_ID = founders.HN_ID
JOIN (
    SELECT companies.Company_ID, companies.Slug
    FROM companies
    WHERE companies.Status <> 'Inactive'
) AS active_companies
ON founders.Company_Slug = active_companies.Slug
GROUP BY company_experience.Company
ORDER BY Num_Founders DESC
LIMIT 10;
```



Company	Num_Founders
Google	309
Facebook	176
Microsoft	158
Amazon	94
Goldman Sachs	85
Stanford University	84
Apple	70
Uber	69
McKinsey & Company	59
IBM	50

Insights:

1. Tech giants are the main feeders, including Google (309), Facebook (176), Microsoft (158), Amazon (94) and Apple (70).
2. Goldman Sachs (85) and McKinsey (59) ---YC founders don't just come from the tech sector.
3. Stanford University (84) aligns perfectly with the insights from EDUCATION part---connection with education and working.

Insights Summary

1. To speed up the development of Y Combinator and rise the survival rate of startups, **Non-Inactive Companies and their founders** are our analytic objects.

2. Y Combinator should allocate more funds into companies from **Consumer, Fintech and Healthcare Industry** or **Product and Design and AI industry** with various sectors.

3. Y Combinator should pay more attention to **North America(USA and Canada)**, **South Asia(India)** and **Europe(United Kingdom)**.

4. Founders who were graduated from **Stanford University, MIT, University of California Berkey and University of Waterloo** and have attended **course or program from Y Combinator** should be invested to keep high rate of outstanding startups.

5. Y Combinators should allocate more funds to founders who have worked or now is working at **Google(Alphabet), Facebook(Meta) and Microsoft**.

Scope for Extension

If we have more time and resources, below are what we can extend:

1. Figuring out what should Y Combinator improved for reducing inactive company

2. Finding out what feature will affect and improve the operation of the YC most.

3. Searching for more resources about the YC's business model and getting the most effective way to rise the survival rate of YC's startups.

Thank You!!