GLOBAL COMPANY LANDSCAPE ANALYSIS

Presented by: Galih Bima Wasena

OVERVIEW

Dataset of 1,320 companies enriched with industry, location, revenue, and employee data

Each record includes location (country, state, city), industry, revenue bracket, employee-size bracket, company type (public/private), and founded year.

Cleaned and transformed the raw API-exported data in Python, modeled it in a normalized SQL schema, and visualized key patterns in Tableau.

The goal is to understand how revenue and workforce size vary across industries and countries, and whether we can predict a company's revenue bracket from basic profile information.

OBJECTIVES

OBJECTIVE 01

Profile how companies are distributed by industry, country, revenue bracket, and employee size.

OBJECTIVE 02

Investigate whether certain industries and countries are more likely to host high-revenue or large-employee companies

OBJECTIVE 03

Use a supervised model (KNN classifier) to predict a company's revenue bracket from its industry, country, and employee size.

OBJECTIVE 04

Normalize the data in SQL and create interactive Tableau dashboards for stakeholders to explore revenue and workforce patterns.

DATA COLLECTION & CLEANING

Data Collection

- Retrieved company data using the CompanyEnrich API
- Exported multiple batches of companies and combined them into a single dataset

Final dataset includes fields such as:

- Name, Domain, Country, State, City
- Industry / Category
- Revenue Bracket & Employee Bracket
- Founded Year, Company Type

Data Cleaning

- Removed irrelevant fields (e.g., logo URLs, social links)
- Standardized text columns: industry, categories, locations
- Handled missing values in:
- founded_year → converted “Unknown”/“None” to NULL
- revenue & employees → preserved as brackets

Created numerical features:

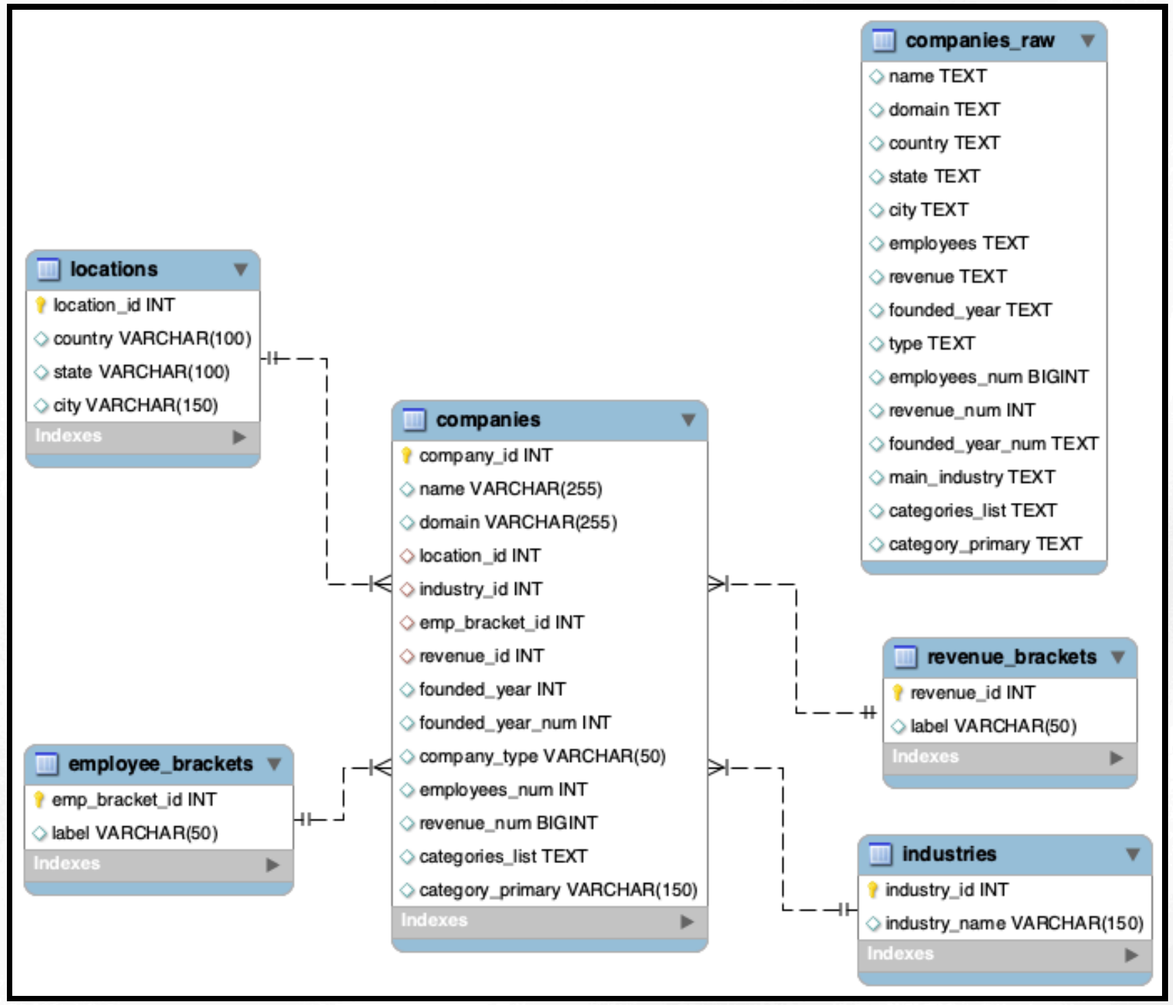
- revenue_num (midpoint of bracket)
- employees_num (midpoint of bracket)

Created cleaned fields:

- main_industry
- category_primary
- categories_list

Identified & removed duplicate companies

SQL NORMALIZATION & DATA MODELING



1. Raw Table Imported (companies_raw)

Contains all original cleaned API fields before transformation.
Used only as a staging table.

2. Separated Repeated Categories into Lookup Tables

Created dimension tables to avoid storing repeated text values:

- locations → unique combination of country, state, city
- industries → unique industry names
- employee_brackets → employee size categories
- revenue_brackets → revenue size categories

Each table has an ID as a primary key, referenced by the fact table.

3. Built a Central Fact Table (companies)

Contains one row per company with:

- Foreign keys → location_id, industry_id, emp_bracket_id, revenue_id
- Cleaned numerical fields → employees_num, revenue_num, founded_year_num
- Category fields → category_primary, categories_list
- Core business info → name, domain, company_type

BIVARIATE ANALYSIS & HYPOTHESIS TESTING

Revenue - Industry

Certain industries show higher concentration of large-revenue companies, especially

- Manufacturing,
- Finance,
- Retail,
- Media & Internet.

Industries like Government and Non-profits cluster more in lower revenue brackets.

Hypothesis Test — Revenue vs. Industry

H_0: Revenue distribution is independent of industry.
H_1: Revenue distribution depends on industry.

Result:

- Chi-square statistic: 405.08
- p-value: 5.0091462766096e-31
- Conclusion: Reject H_0 → Industry and revenue are strongly associated.

BIVARIATE ANALYSIS & HYPOTHESIS TESTING

Employee Size – Industry
Some industries show higher concentrations of large employee brackets, especially:

- Manufacturing
- Finance
- Media & Internet

Industries like Insurance and Government show more small–mid employee clusters.

Hypothesis Test — Employees vs. Industry

H_0: Employee size distribution is independent of industry.
H_1: Employee size distribution depends on industry.

Result:

- Chi-square statistic: 380.49
- p-value: 1.62×10^{-18}
- Conclusion: Reject H_0 → Employee size and industry are significantly associated.

BIVARIATE ANALYSIS & HYPOTHESIS TESTING

Revenue – Country

Countries vary in their revenue distributions. For example, the United States has a noticeably higher share of over-1B revenue companies compared to others.

Hypothesis Test — Revenue vs. Country

H_0: Revenue distribution is independent of country.
H_1: Revenue distribution depends on country.

Result:

- Chi-square statistic: 576.15
- p-value: 1.63×10^{-13}
- Conclusion: Reject H_0 → Country and revenue level are significantly associated.

TABLEAU VISUALIZATION

KEY TAKEAWAY

1. Revenue varies significantly across industries

- Manufacturing, Finance, Retail, and Media & Internet show higher concentrations of large-revenue companies.
- Government and Non-profit sectors tend to have lower revenue brackets.

2. Employee size is industry-dependent

- Manufacturing and Media & Internet show larger employee brackets.
- Government and Hospitality have smaller to mid-sized organizations.

3. Country is linked to revenue distribution

- Countries like the U.S. and U.K. show higher proportions of large-revenue firms.
- Japan and France display more balanced revenue distributions.

PREDICTIVE MODEL

01 Objective

Predict a company's revenue bracket using key categorical business attributes:

- Country
- Industry
- Company Type
- Primary Category
- Employee Size

02 Model Used: K-Nearest Neighbors (KNN)

Selected as an interpretable baseline model for multi-class classification.

Features Included: Country, Main Industry, Type, Category Primary, Employees

Target Variable: Revenue Bracket (under-1m, 1m-10m, 10m-50m, 50m-100m, 200m-1b, over-1b, Unknown)

03 Data Preparation

Converted all categorical features to numerical labels
Standardized KNN-compatible dataset

Split into:

- 80% training
- 20% testing

MODEL PERFORMANCE

04 Classification Report Summary

The model performs best on extreme revenue brackets, especially “over-1b,” which have clearer patterns.

Mid-range brackets (10m–100m) show lower precision and recall due to overlap between company profiles.

Overall scores:

- Accuracy: ~47%
- Macro F1-score: ~35%

Interpretation: The model captures general trends but struggles with fine-grained revenue prediction.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.11      0.14      0.12         7
           1       0.29      0.45      0.36        42
           2       0.11      0.12      0.11        32
           3       0.27      0.16      0.20        19
           4       0.47      0.36      0.41        44
           5       0.77      0.82      0.80        83
           6       0.63      0.32      0.43        37

    accuracy                           0.47       264
   macro avg       0.38      0.34      0.35       264
weighted avg       0.49      0.47      0.47       264

Confusion Matrix:
[[ 1  3  0  1  0  1  1]
 [ 4 19 10  0  5  2  2]
 [ 2 11  4  2  5  5  3]
 [ 1  5  5  3  2  3  0]
 [ 0  9 10  2 16  6  1]
 [ 0  4  4  2  5 68  0]
 [ 1 14  5  1  1  3 12]]
```

CONCLUSION

✓ **Key Relationships Identified**

- Revenue and employee size show significant associations with industry.
- Countries differ meaningfully in their revenue distributions.
- These patterns demonstrate that business characteristics are not randomly distributed, but shaped by structural and regional factors.

✓ **Predictive Modeling Insights**

- A baseline KNN model achieved ~47% accuracy, showing that categorical features provide some predictive power, especially for the largest companies.
- Overlapping mid-range revenue groups limit performance, suggesting the need for richer features.

Larana, Inc.

THANK YOU

Presented by: Galih Bima Wasena