

Web Scrapping & Data Analysis

Forbes Top Companies in India

Executive Summary

This project focuses on web scraping and analyzing data on the world's 2,000 largest publicly traded companies, based on Forbes 2000 rankings. The dataset includes key attributes such as company name, industry, headquarters location, revenue net profit, total assets, and market value of each company. The analysis aims to provide insights into the distribution of top companies across industries and locations, allowing for a deeper understanding of the Indian corporate landscape.

Introduction

The Indian economy boasts numerous large corporations that play vital roles in the global market. This analysis aims to uncover patterns in revenue distribution, industry dominance, and geographic concentration of company headquarters. By analyzing this data, we can gain valuable insights into the industries driving economic growth in India.

Objective

- To identify the leading industries by revenue.
- To analyze the distribution of company headquarters across Indian cities.
- To uncover insights into the industry-wise average revenue and company representation.
- To provide data-driven recommendations for a better understanding of India's corporate environment.

stakeholder

This analysis will benefit:

- **Business Analysts:** To identify leading industries and key players.
- **Economists and Market Researchers:** To understand trends in corporate growth across different sectors and regions.
- **Business Development Teams:** To identify potential opportunities for collaboration with top companies.

Data Description

Dataset Source

This dataset was scraped from a [List of the largest companies in India](#) Wikipedia Page.

Data Fields

The key columns in the dataset include:

- **Rank:** Ranked an Indian Company based on Forbes 2000
- **Forbes 2000 Rank:** Annual ranking of the top 2000 public companies in the world
- **Company Name:** Name of the company.
- **HQ:** The city where the company is headquartered
- **Revenue, Profit, Assets, Value:** In billions of US dollars of the company
- **Industry:** The industry in which the company operates.

Data Preprocessing

Data Scraping

Scrap the data using Pandas, BeautifulSoup, and Request from the Wikipedia page. On that particular page, I just picked out the 2024 Forbes list table for this analysis.

Data Cleaning

- To know the information about Datafram
- Rename the Columns
- Convert numerical columns to appropriate data types
- Set the Rank Column as an index

Analyze the Data with Ploty

Key Visualization and Insights

1. **Revenue Generating by Companies:** Filtered Top 5 Revenue Generating Companies. As a result, **Reliance Industries Limited** is generating huge revenue among other Indian companies.
2. **Profit Distribution of Companies:** **Reliance Industries Limited** creates more profit from their business followed by **State Bank of India** and **HDFC Bank**
3. **Top Companies Ranked by their Asset Value:** **State Bank of India(SBI)** holds huge assets among its fellow public traded companies.
4. **The Top-performing companies within each industry:** List the Companies and Average Revenue with those related Industries.
5. **Top 10 Industry-wise Average Revenue:** The industries generating the highest average revenue were identified. The **conglomerate** and **Oil and Gas** sectors showed the highest average revenue among top companies.
6. **Distribution of Company Headquarters:** The geographic analysis of company headquarters shows **Mumbai** as the dominant location, followed by **New Delhi** and **Bangalore**.

Correlation Analysis

A correlation analysis was conducted between revenue, Profit, assets, and Market Value. In this discovery, all the variables display a positive correlation with one another, indicating an absence of negative correlations within the datasets.

Methodology

The descriptive analytical approach was applied using basic statistics and visualizations. The project relied on Python's web scraping libraries (e.g., **Pandas**, **BeautifulSoup**, **requests**) for data extraction and visualization tools like **Plotly** for creating interactive charts.