Data Quality & Source Validation

Checkpoint 3

Mapping Global Growth — A Data-Driven Strategy for NFL Market Entry

Project Title & Data Overview

Title: Mapping Global Growth: A Data-Driven Framework for NFL International Expansion

This project develops a quantitative model to identify and rank international markets most suitable for NFL expansion. It integrates fan engagement, economic capacity, and infrastructure readiness into a unified Market Attractiveness Index to support strategic decisions on future global growth.

Datasets Used:

1. **World Bank Open Data:** GDP per capita, population, urbanization rate (2019–2024), to understand economic and demographic viability

2. **Google Trends:** Country-level search interest in ‘NFL’, ‘American Football’, and ‘Super Bowl’, to measure fan engagement

3. **World of Stadiums (Kaggle):** Stadium capacities and locations, to assess sports infrastructure capacity

4. **OpenFlights database:** Two datasets, one with flight routes and one with airport information (i.e., city and country), to measure international connectivity for logistic feasibility

These datasets were selected for their accessibility, reliability, international coverage, and alignment with key market attractiveness factors.

Data Source Summary

The following table contains details on the source and collection method for the datasets:

Source	Collection	Notes
World Bank Open Data	Official international economic database (https://api.worldbank.org/v2/country/all)	Collected and maintained by World Bank; data reported annually by country
Google Trends	Publicly accessible search interest data, obtained through Google Trends API (obtained using pytrends)	Real-time search volume aggregated at the country level; measures fan engagement online
World of Stadiums	Crowd-sourced dataset from Kaggle (downloaded from https://www.kaggle.com/datasets/rahuldabholkar/world-of-stadiums?select=all_stadiums.csv)	Compiles stadium name, capacity, and location; verified by contributors
OpenFlights	Public airline and airport database (https://openflights.org)	Routes dataset provides number of airports and route connectivity; Country information gathered from Airports dataset; maintained by aviation enthusiasts

Data Structure & Content

- **World Bank:** 250 × 6 years (2019–2024), from which we compute the average of: GDP per capita (USD), total population, urban percentage.
- **Google Trends:** 183 countries, from which we compute the normalized search interest (0–100 scale) for terms such as ‘NFL’, ‘American Football’ and ‘Super Bowl’.
- **World of Stadiums:** 2,100 stadiums with variables: stadium name, city, country, and capacity. This will then be aggregated to country level for stadiums ≥45,000 seats, computing average stadium capacity for each country.
- **OpenFlights:** over 10,000 airports and routes globally; variables include airport location, routes, and country code. This is then aggregated to country level, where we compute a connectivity index.

The merged dataset is at the country level, merged according to ISO 3 country code (unique identifier from World Bank dataset, then replicated to the remaining datasets), with each row representing one country and columns corresponding to economic, engagement, and infrastructure features.

Data Completeness & Consistency

Coverage:

- **World Bank Data:** Nearly all recognized countries are included, providing annual economic and demographic indicators (GDP per capita, population, urbanization) from 2019–2024. This ensures strong baseline coverage for global market analysis.
- **Google Trends:** Provides normalized search interest at the country level, but coverage is limited to countries with sufficient search volume for the terms “NFL” and “Super Bowl.” Smaller countries or regions with low online engagement may be missing or show near-zero values. These gaps could underrepresent latent fan interest in some markets. This does not seem critical, as there seems to be a positive correlation between most countries with low search volume and low infrastructure or demographic variables, which would not make them top candidates for the Market Attractiveness Index.
- **World of Stadiums:** The dataset includes over several stadiums globally but is crowd-sourced and uneven in coverage. Some countries, particularly in Africa or smaller nations, may lack complete records of large venues. Only stadiums with capacity ≥45,000 seats were included for analysis to align with NFL hosting feasibility, which further reduces coverage in regions with fewer large venues. This should not present a serious issue for the final index, similarly to the Google Trends data, since it is excluding candidates that were rather inaccessible or low-performing in economic/demographic metrics.
- **OpenFlights:** Contains large amounts of data on airports and routes worldwide, but it is crowd-sourced and maintained by aviation enthusiasts. This could overstate/understate connectivity in some countries.

Given the different data sources, in the final dataset, countries that are not present across all datasets will be dropped.

Consistency:

- All datasets can be standardized to numeric values where applicable (e.g., GDP in USD, stadium capacity as integers, Google Trends scaled 0–100).
- Country identifiers can be harmonized using ISO 3 codes to ensure consistent merging across datasets.
- Stadium capacities will be cleaned to remove non-numeric entries, converted to integers, and filtered to ≥45,000 seats. Aggregations such as number of stadiums, and average capacity will be computed per country consistently.

Duplicates:

- The stadium dataset contains some repeated entries due to multiple sources or alternate naming conventions. These can be identified and removed during aggregation by counting unique stadium names per country.

- World Bank and Google Trends datasets inherently have unique country-year entries, so duplicates are not present there.

Quality Issues & Potential Biases

- **Data quality issues:** all datasets were reviewed and corrected for data quality, through identifying outliers and missing values. Standardized numeric formats and consistent country codes enable reliable merging and index computation.
- **Potential bias**: Even though the missing or partial coverage in Google Trends and stadium datasets may introduce some bias, it does not seem to have a critical impact in the analysis, since the Market Attractiveness Index will focus on countries with high online engagement and existing large infrastructure.

Initial Cleaning / Preparation Plan

Missing Data:

Assess whether to input missing data using regional averages or dropping missing value entries, depending on the size/economics of the country, for the following datasets:

- Google trends
- World of Stadiums

Aggregation:

- Stadiums ≥45,000 seats aggregated by country
- Airports aggregated to country-level connectivity scores
- World Bank and Google Trends data aggregated to country-level with averages for the past 5 years

Standardization:

- Convert all numeric variables to consistent units
- Ensure all countries have consistent ISO 3 country codes assigned
- Normalize variables to enable aggregation into a composite index (for the connectivity metric and for the overall dataset to enable the computation of the Market Attractiveness Index)

Merging:

- Merge World Bank, Google Trends, stadium, and flight datasets on ISO3 country codes
- Verify unique country rows and remove duplicates

Tools / Methods: Python (pandas, sklearn) for cleaning, normalization, and index calculation (for the connectivity index).

Next steps

- Complete data merging into a single country-level table.
- Conduct exploratory data analysis (EDA): distributions, correlations, missing value patterns.
- Compute a preliminary Market Attractiveness Index using a weighted combination of normalized features.