Exploratory Data Analysis (EDA) Memo

Checkpoint 5

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

Question Snapshot

The project aims to answer the question 'Which international markets present the highest potential for sustainable NFL expansion, based on a balance of fan interest, economic strength, and infrastructure readiness?'. Success means identifying a small set or cluster of countries that have the right mix of factors and would make good candidates for the NFL's international expansion.

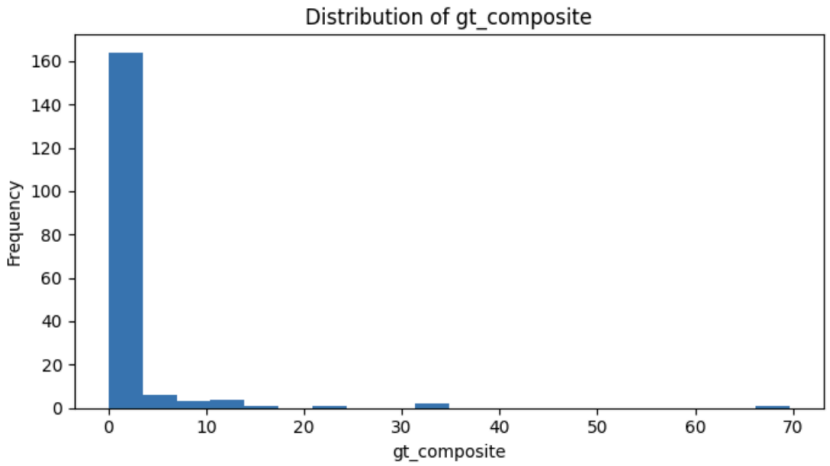
Data Used

The project is based on the cleaned country-level dataset from CP4, where each row represents a single country. Variables include Google Trends signals (to measure interest in the NFL), economic indicators (GDP per capita), demographics (urban population), stadium infrastructure, and aviation connectivity metrics.

Univariate Distributions

The key variables analyzed were:

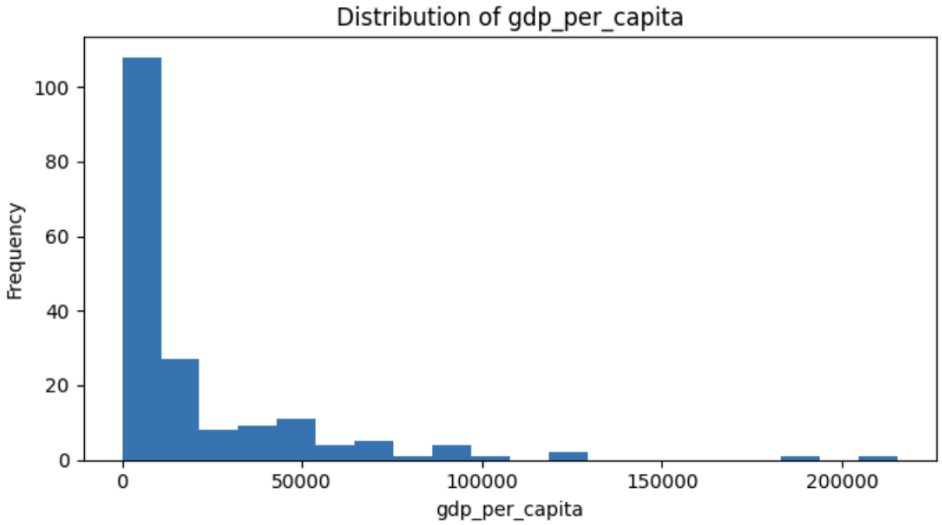
- **gt_composite:** the composite score that measures interest in the NFL per country from Google Trends



Many countries show low to moderate interest and a long tail of countries with high NFL-related search volume. A clear **floor at 0–5** reflects countries with minimal search activity, which is expected for a

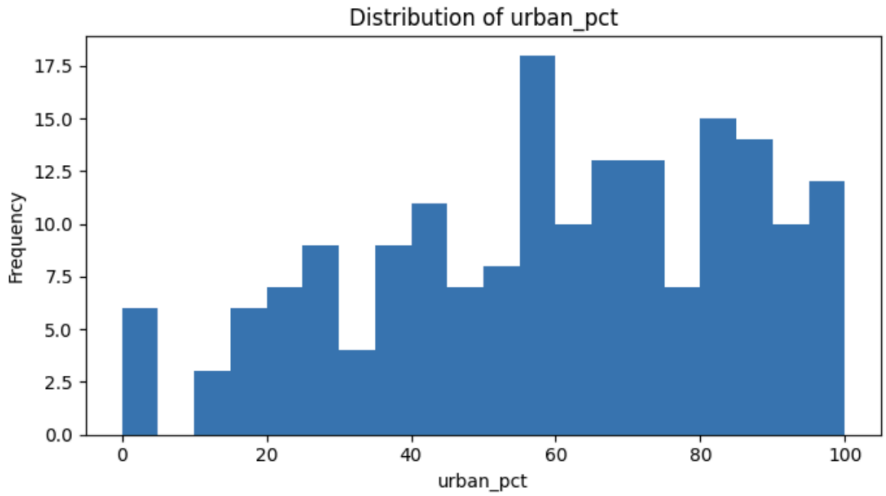
U.S.-centric sport. The main outlier (nearly 70 composite score) corresponds to the U.S., which will be removed from the analysis since the research question pertains to international expansion. Another outlier, at 33 was identified, corresponding to Papua New Guinea and attributed to data error, so it was also removed.

- **gdp_per_capita**: the country's GDP per capita



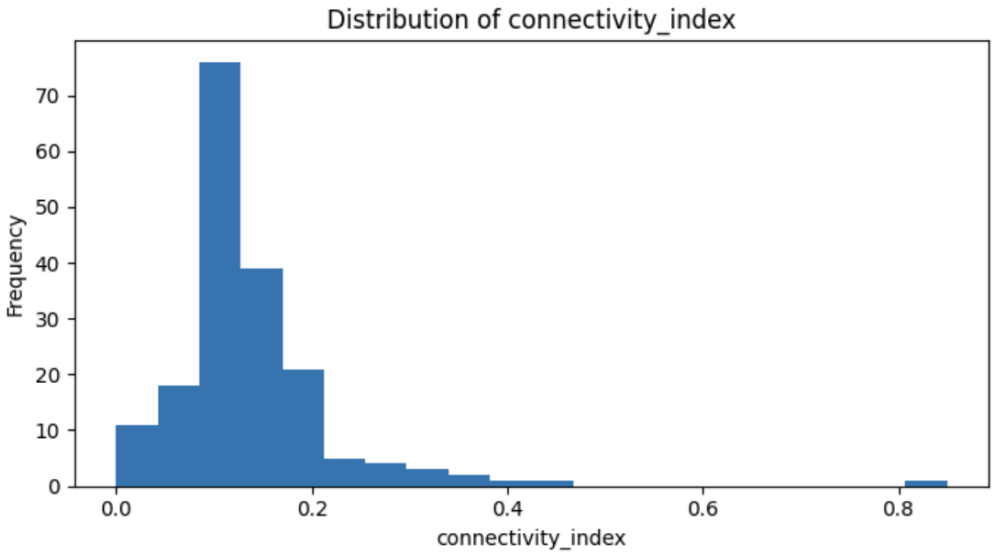
Most countries are clustered in the lower-to-mid income range and a small group of very wealthy outliers (e.g., Luxembourg, Qatar, Norway, Switzerland). This reflects real-world global inequality rather than data issues. No artificial ceilings or floors appear. The skew suggests that modeling should consider log-scaling GDP or using standardized percentiles to avoid extreme values dominating relationships.

- **urban_pct:** percentage of population in urban areas



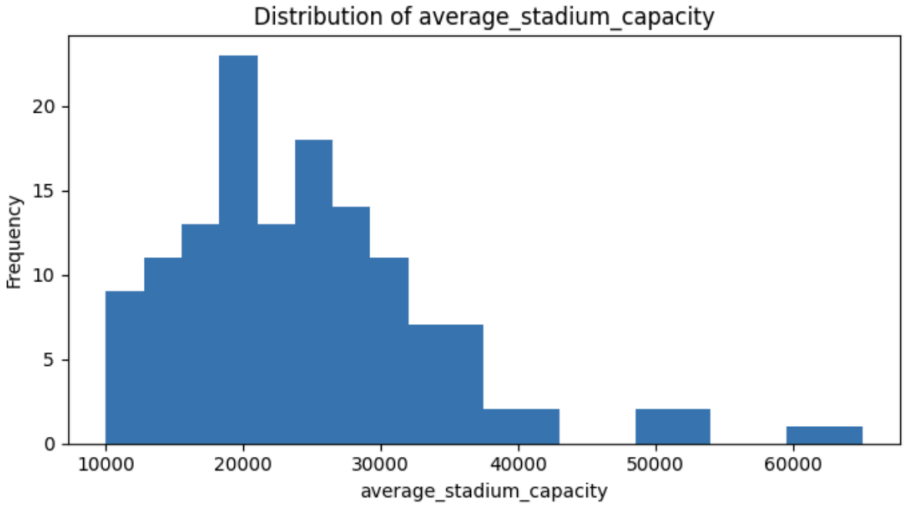
Most countries are highly urbanized. Typical values fall between 55% and 80%, consistent with global patterns. A few countries are extreme outliers on the low end (e.g., Papua New Guinea, Burundi), which are real values but may not be relevant NFL markets. No anomalies or formatting issues visible. This suggests urbanization will not function as a main differentiator among top candidate markets.

- **connectivity_index:** composite index of number of airports and routes



Most countries have low to median air connectivity and only a handful acting as major global hubs (e.g., USA, UK, Germany, UAE, Singapore). This distribution seems to reflect real aviation patterns. There's a strong floor at zero or near-zero for small island nations or countries with minimal international flight networks, and no apparent errors in data. The concentration of connectivity in a few nations suggests that infrastructure readiness is highly concentrated and may serve as a decisive variable for NFL expansion.

- **average_stadium_capacity:** average number of seats per stadium in the country



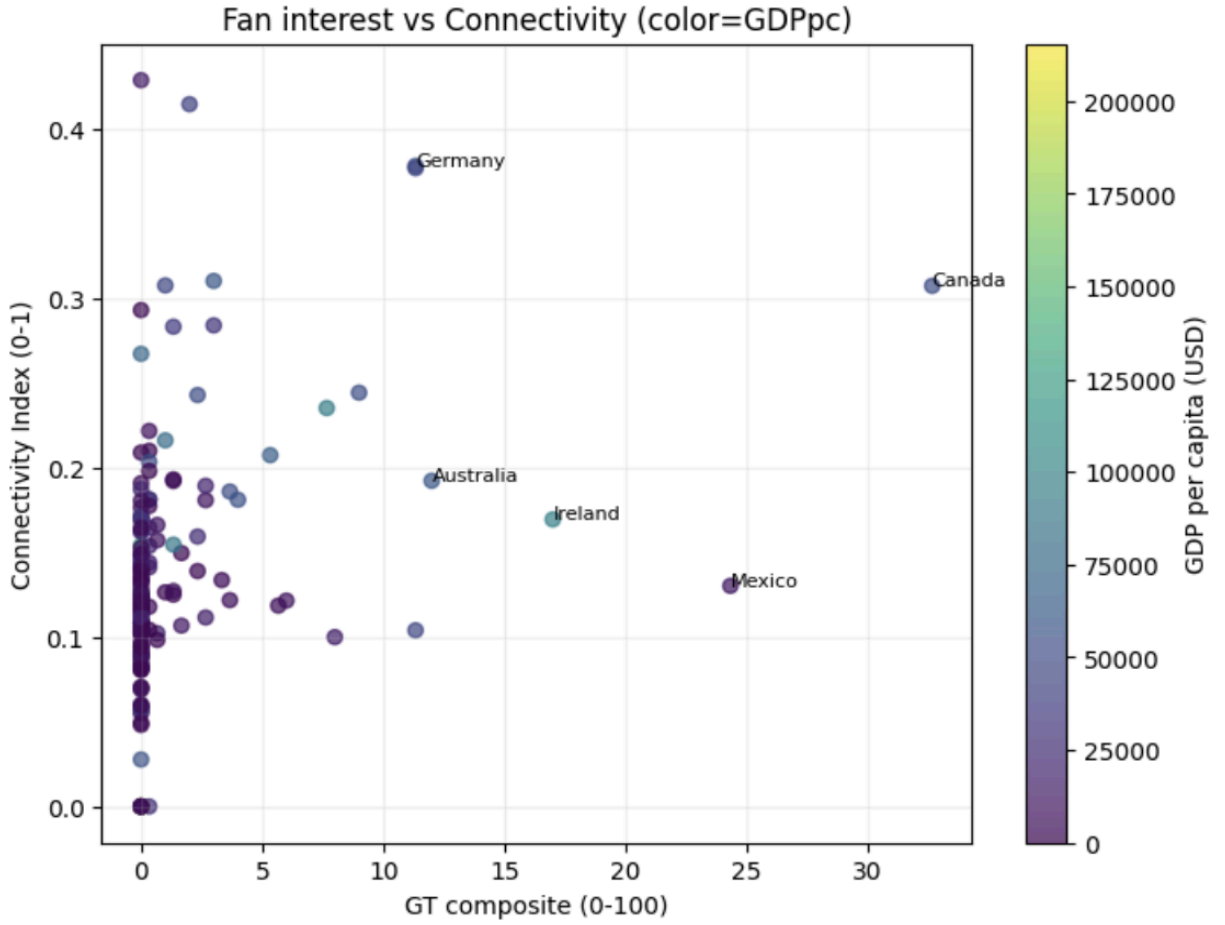
The majority of countries have relatively small or mid-size stadiums, with only a few hosting very large venues. This is expected, as mega-stadiums are rare, typically found in countries with strong football, cricket, or Olympic traditions. A handful of extremely large stadiums create a long upper tail but no

apparent data errors. This variable aligns with expectations and reflects capacity constraints that could impact where the NFL can realistically host large-scale events.

Relationships

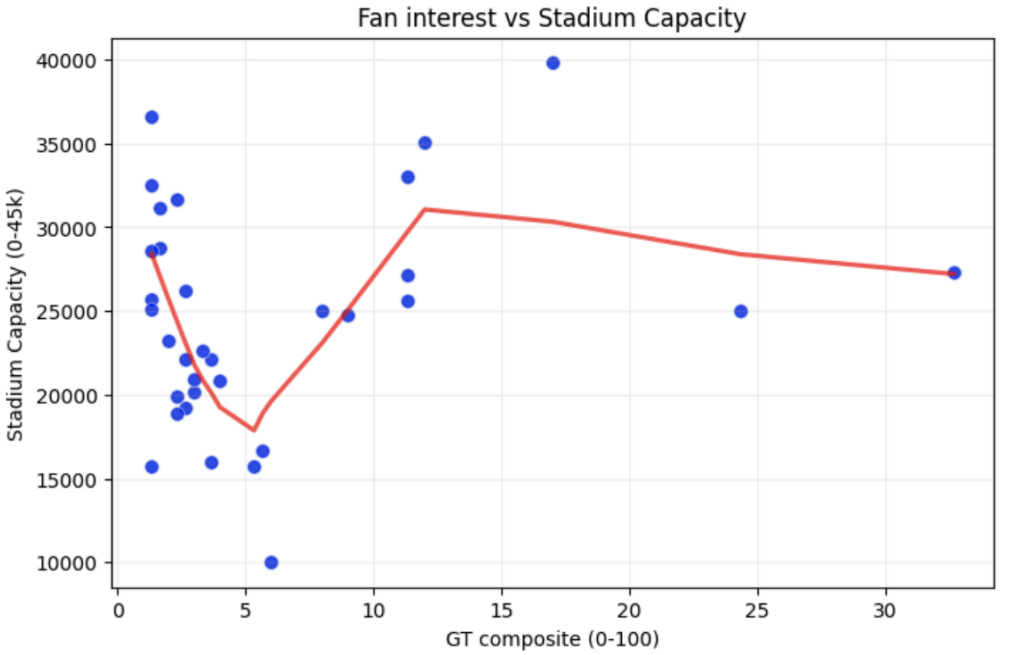
The following relationships were analyzed:

- **Fan interest** (gt_composite) relationship to **connectivity** (connectivity_index) and **GDP per capita** (gdp_per_capita)



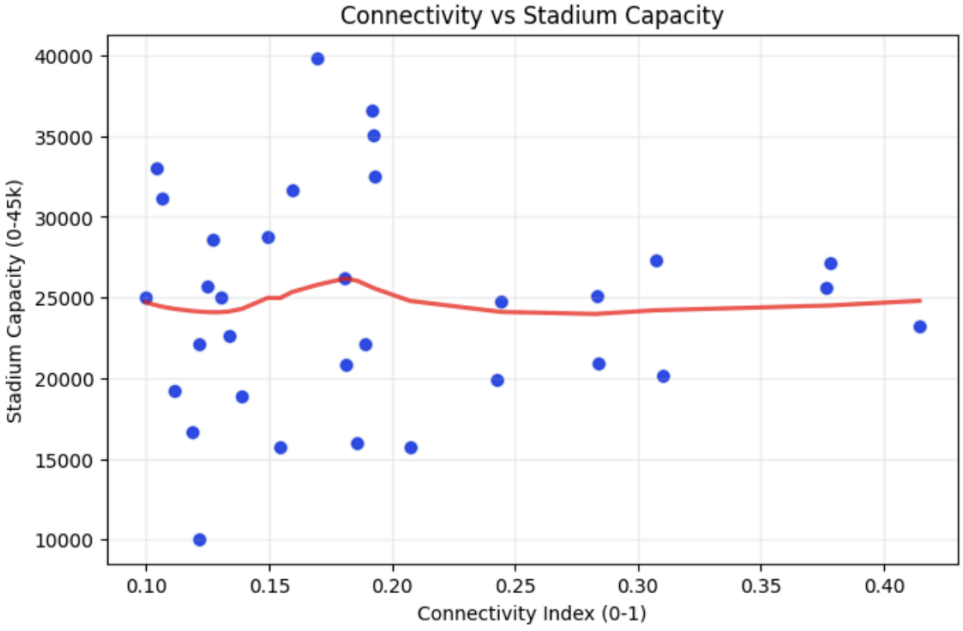
The plot is useful to assess potential countries for expansion, since connectivity plays a major role in bringing spectators in from nearby countries. After removing the outliers (US and Papua New Guinea), certain countries (Mexico, Canada, Ireland) emerge as possible candidates. Germany will also be removed from the final analysis, given that the NFL already conducts games in Frankfurt.
There does not seem to be a strong correlation between the variables, which was expected given that Football is a US-centric sport, and that most likely neighboring or anglo-saxonic countries will have higher interest in the NFL (Mexico, Canada, Ireland, Australia), yet it is still useful to assess the relationship between the two dimensions.

- **Fan interest** (gt_composite) vs **Stadium Capacity** (average_stadium_capacity)



Again, there does not seem to be a strong relationship between the two variables given the specificity of American Football. However, it is good to note that there is stadium capacity above 25,000 in countries with relatively high interest in the NFL. In line with the previous relationship analysis, this seems to indicate that there will be different candidates that can qualify for the NFL's international expansion.

- **Stadium capacity** (average_stadium_capacity) vs **connectivity** (connectivity_index)



The relationship between connectivity and stadium capacity (very minimally positive, almost no correlation) seems to reflect the fact that many countries have modern connectivity networks even if their stadiums are small (e.g., small European countries with strong soccer culture). Conversely, some countries have very large stadiums (e.g., South America) but only moderate connectivity.

Subgroup Comparisons

Splitting was done by 'economic strength' to help isolate whether interest and infrastructure are high in countries that can afford to support the NFL.

- **Split Column:** gdp_per_capita.
- **Subgroups:**
  - **High-GDP Countries:** Countries with a gdp_per_capita **above** the median ($6,752.03).
  - **Low-GDP Countries:** Countries with a gdp_per_capita **below** the median.
- **Comparison:** Compare the average **fan interest** (gt_composite) and **infrastructure readiness** (connectivity_index) between these two groups.

gdp_subgroup	Avg_NFL_Interest_Index	Avg_Connectivity_Index	Avg_GDP_Per_Capita
High-GDP Countries	2.92674	0.159350	$37,727.98
Low-GDP Countries	0.63370	0.115237	$2,725.26

- **Higher Group:** The **High-GDP Countries** group is significantly higher across all three key metrics:
  - **Economic Strength:** Their average GDP per capita is over 13x that of the Low-GDP group.
  - **Fan Interest:** Their average NFL interest index is ~4.6x higher.
  - **Infrastructure:** Their average connectivity index is ~1.4x higher.
- **Lower Group:** The Low-GDP Countries group is consistently lower in all three metrics.

The difference is meaningful and suggests a positive correlation between economic strength, fan interest, and infrastructure readiness.

- **The Key Insight:** Markets that are already economically strong (High-GDP) are currently the most attractive, as they also have significantly higher levels of Fan Interest and Infrastructure Readiness.

- This indicates that NFL expansion efforts should primarily target High-GDP Countries for the highest potential for sustainable expansion, as these markets meet the three required criteria (fan interest, economic strength, and infrastructure readiness) in a balanced way. The low-GDP group, while potentially offering future growth, currently lacks the economic foundation and infrastructure to be considered a priority for sustainable expansion.

Outliers and Anomalies

The only outlier identified as a data error was Papua New Guinea's Google Trends (fan interest) data, which was removed from the analysis. The US were also an outlier in terms of fan interest, not due to errors but to being the main country where there is engagement with the NFL. It was also removed from the analysis since it pertains to international expansion.

Missingness & Coverage

Some small countries lack data on connectivity and stadiums (e.g., Andorra, Eswatini, Kosovo, Liechtenstein, Monaco, North Macedonia, San Marino, St. Lucia, Timor-Leste), the vast majority due to lack of actual airports/stadiums. The countries affected do not seem critical for the analysis.

A very small subset of countries do not have GDP information but will also be disregarded given that they have a none or very small levels of fan engagement (measured through Google Trends)

Early Takeaways & Next Steps

Early Takeaways:

- **Strong Alignment of Potential:** The three pillars of potential for sustainable expansion (Economic Strength, Fan Interest, and Infrastructure Readiness) are all significantly higher in the High-GDP Countries group. This strong positive correlation may indicate an initial targeting strategy.
- **Fan Interest Follows Wealth:** Fan Interest (Avg. NFL Index) is higher in the High-GDP group (2.93) than in the Low-GDP group (0.63). This suggests that the NFL brand is already reaching and resonating most effectively in affluent, developed markets, and bringing games to those locations will expand engagement in those countries.
- **Infrastructure is Ready:** Connectivity and infrastructure are also higher in the High-GDP group, minimizing the logistical hurdle for staging future international games or establishing permanent presences.

Next steps:

- To move beyond simple paired comparisons, a composite **Market Attractiveness Index** will be constructed. This index provides a single, balanced score for each country, reflecting its overall potential for sustainable NFL expansion based on the core project requirements.
- We plan to perform K-Means Clustering on a normalized feature set, to objectively group countries into homogeneous segments.