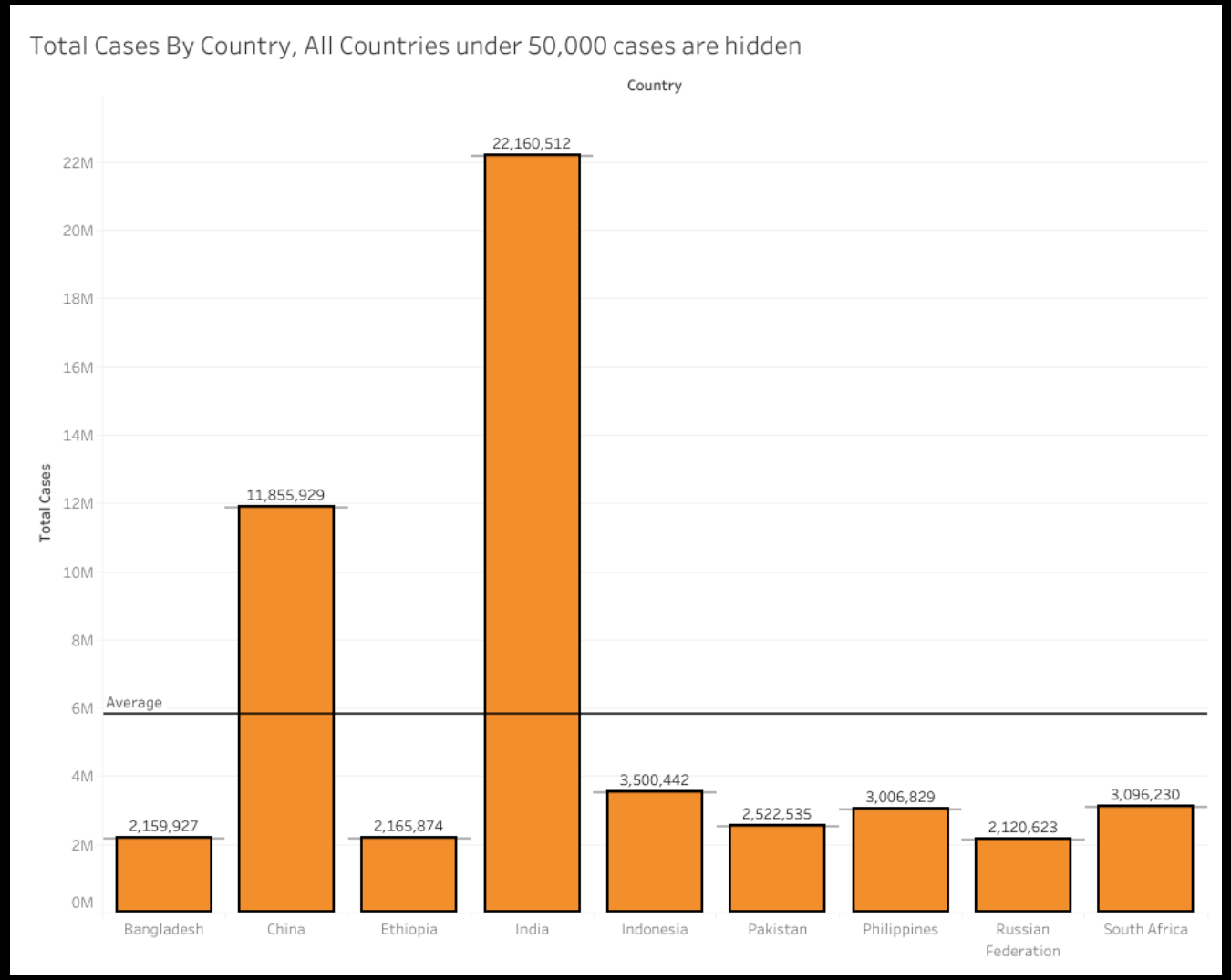
Case Study 7: Tuberculosis Insights

Presented By Wil Jones

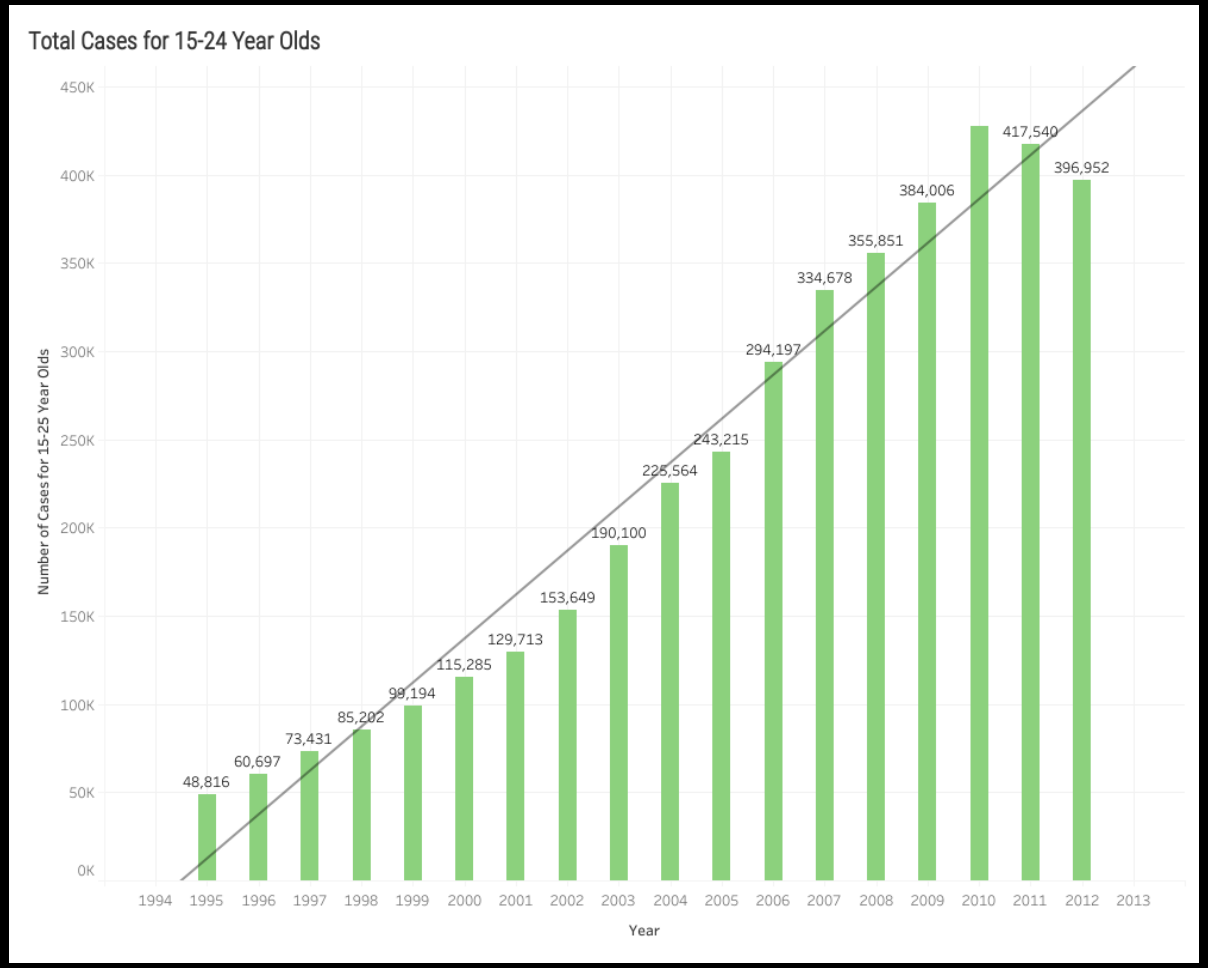
Which countries require our attention?

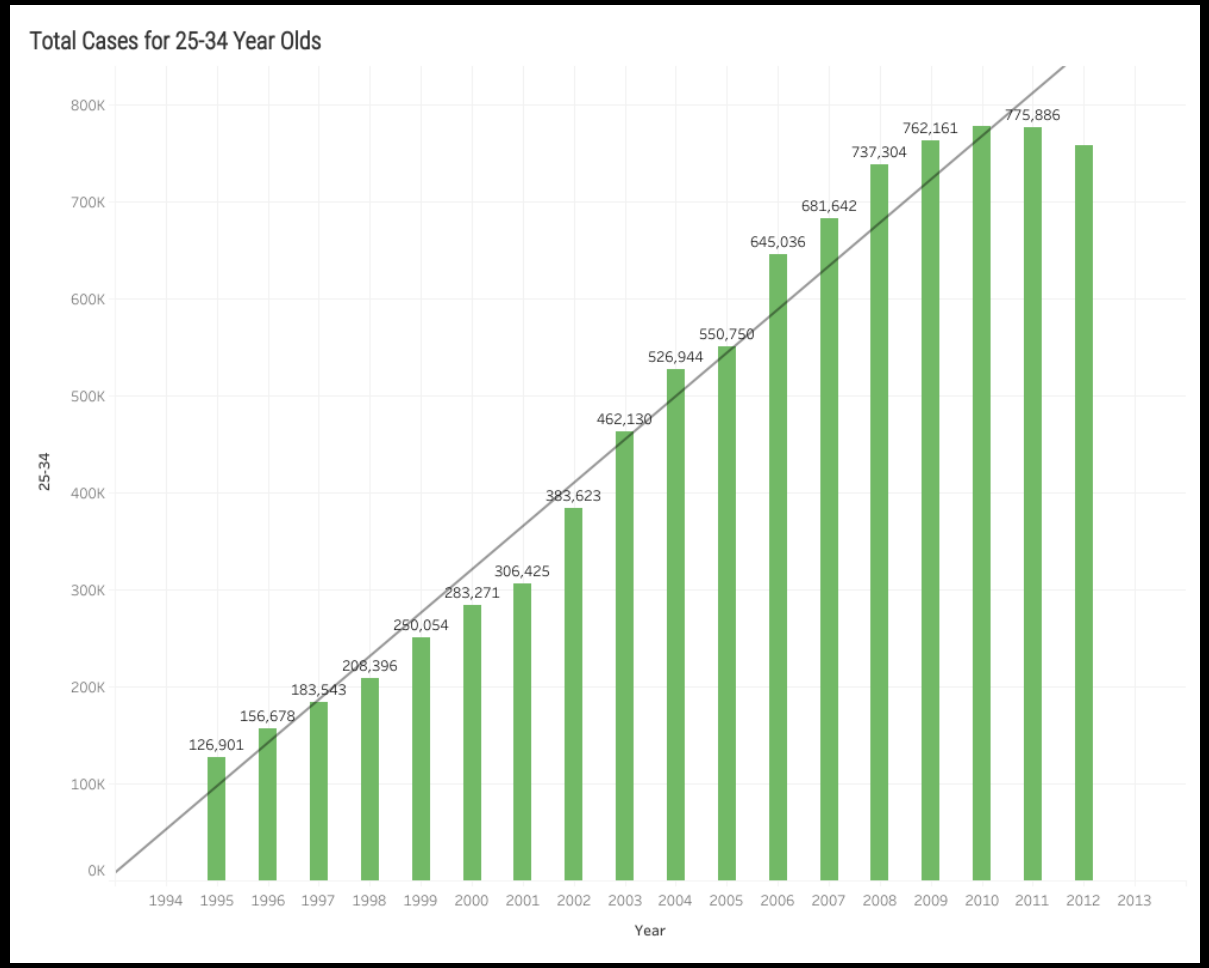
Insights

- **India** and **China** account for the **vast majority** of global TB cases.
- India leads with **22,160,512** total TB cases, followed by China with **11,855,929**.
- Other high impact countries include **Indonesia**, **Pakistan**, and **South Africa**.
- The countries with the highest case counts are **far above** the global average, with India being the **biggest contributor**.



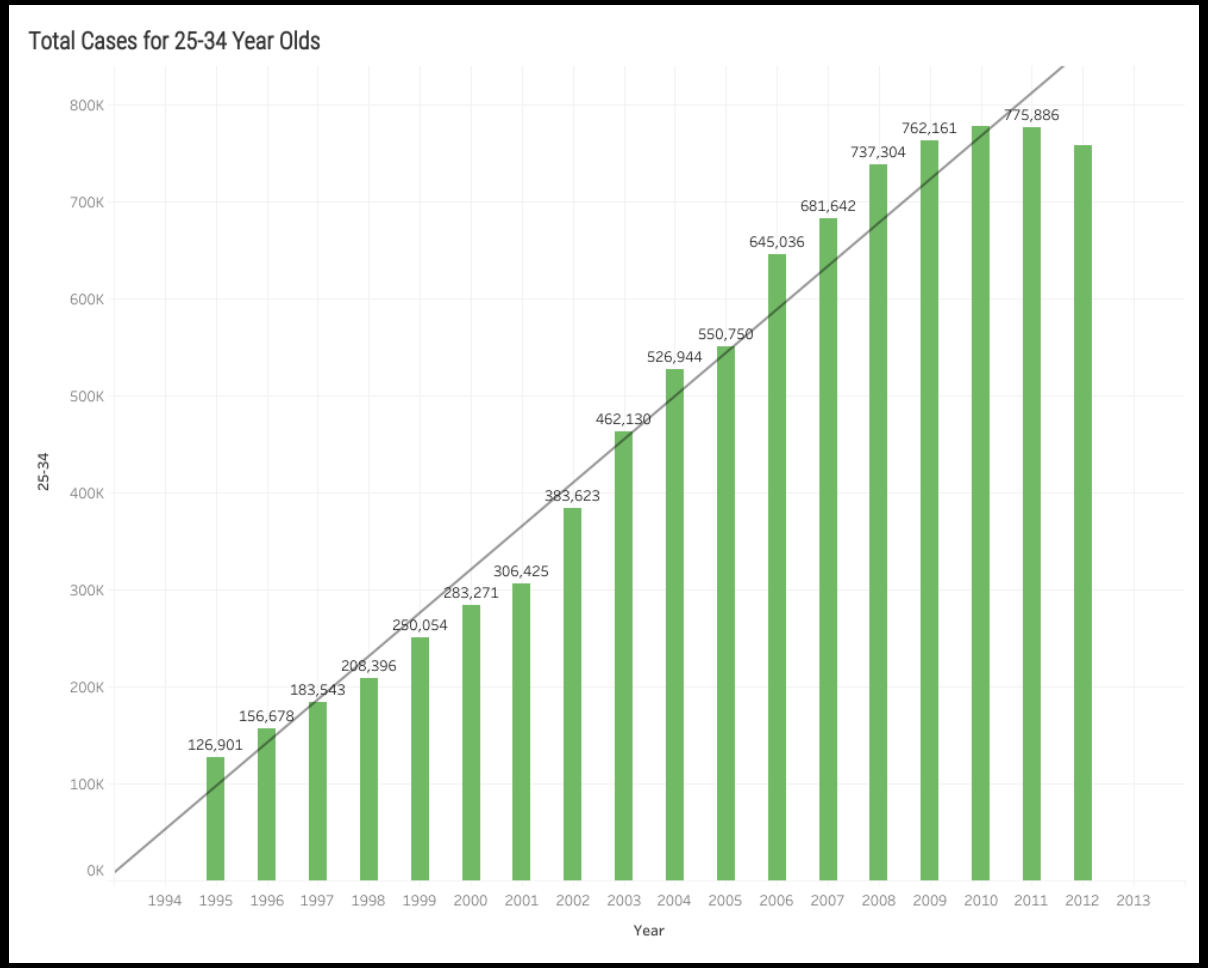
Which age group requires our attention?

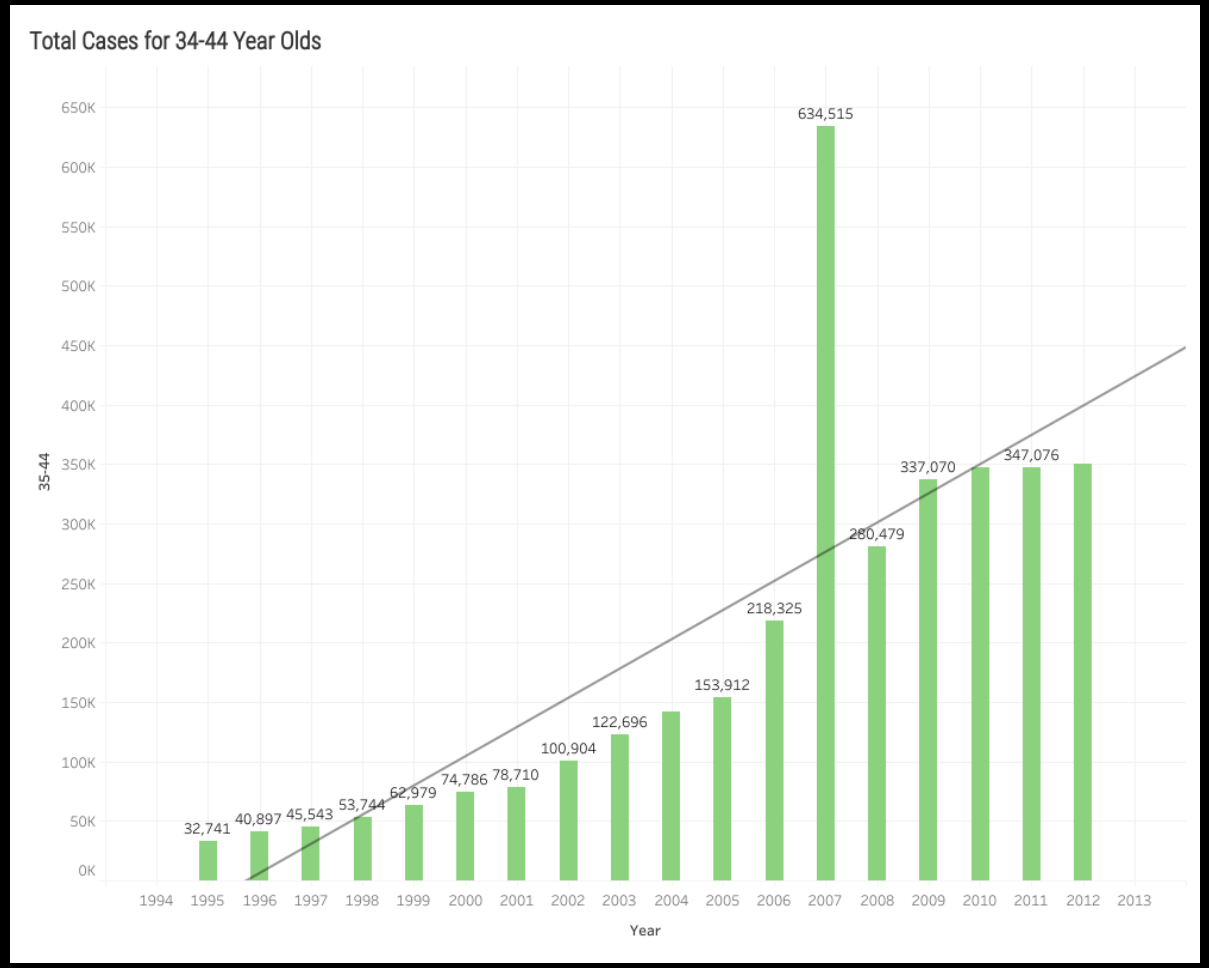




Insights

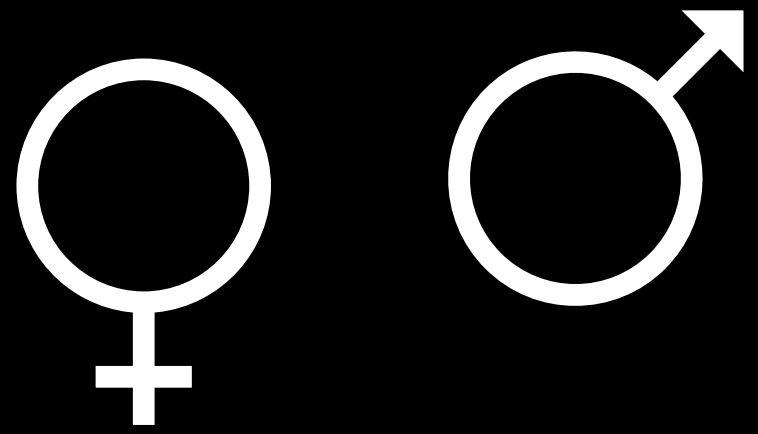
- Between **1994** and **2013**, both age groups show a **steady rise** in TB cases, but the **25-34** cohort consistently reports **higher** numbers than the **15-24** group. By 2013, **25-34 year olds** reach nearly **double** the number of cases compared to **15-24 year olds**, suggesting that TB risk increases further into adulthood. These findings emphasize the need for targeted interventions and preventive strategies tailored to young and middle-aged adults.



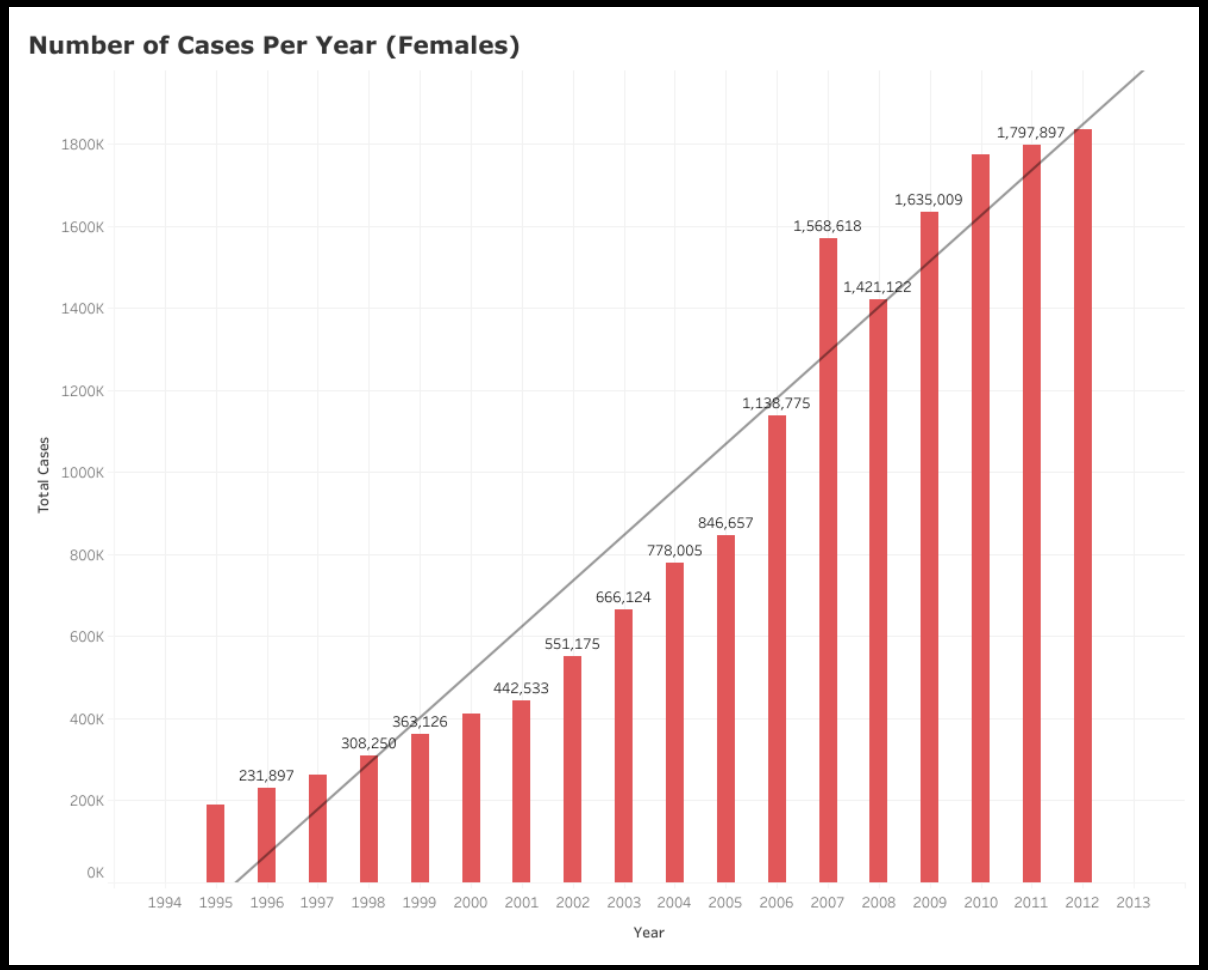


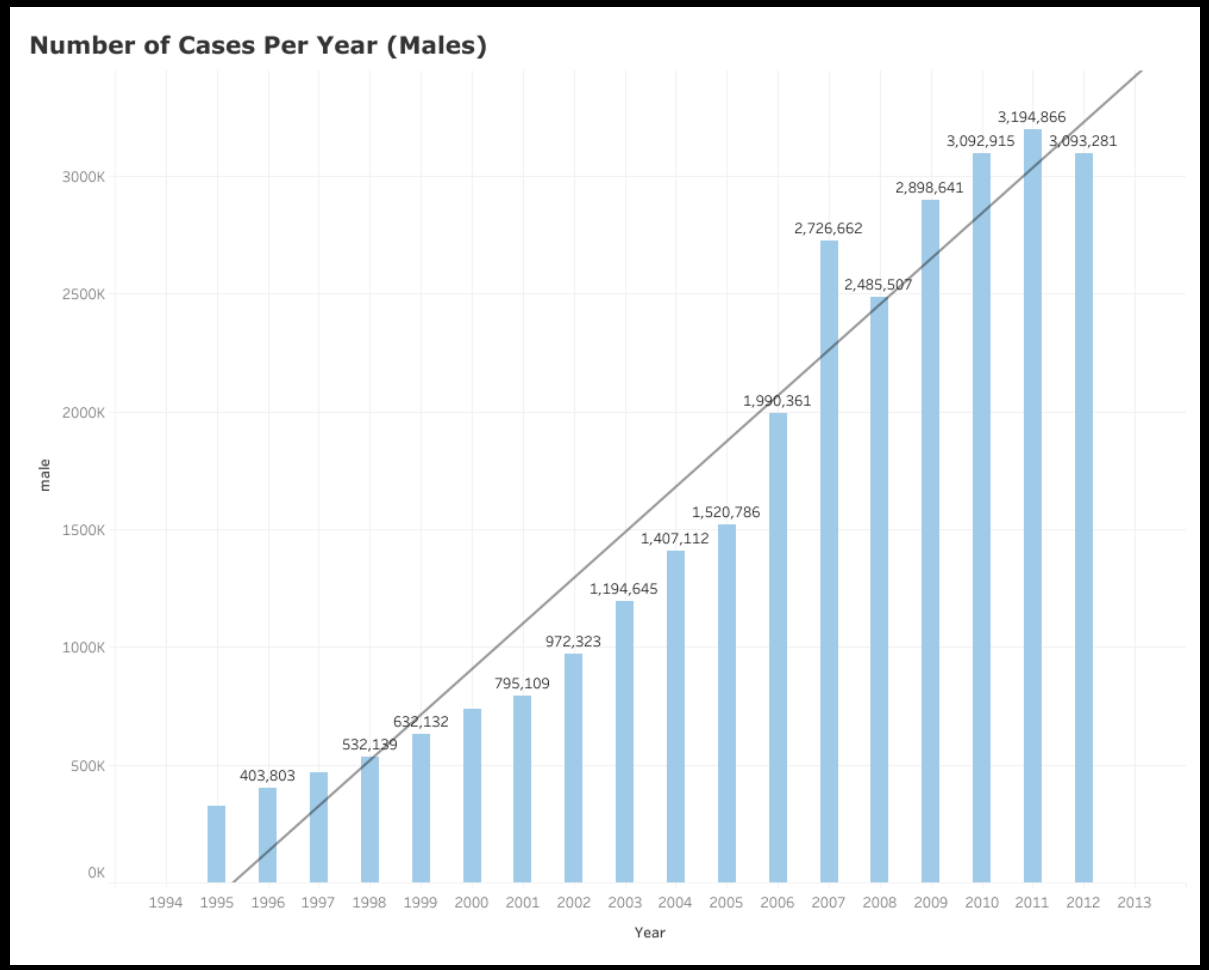
Insights

- Both groups show a clear upward trend in TB cases over the years, but **25-34 year olds** reach a higher overall peak (**~775k cases by 2012-2013**). In contrast, **34-44 year olds** experience a **sharp spike** around **2007 (~634k)** before cases drop below **350k** by 2013. While both age groups remain high-risk, the **25-34** cohort maintains consistently **higher total cases** throughout this time frame, indicating a particularly critical need for targeted interventions in that age range.



Are there differences between genders?





Insights

- Both males and females display a steady rise in TB cases over time. However, the male population consistently has higher total cases, exceeding 3 million by 2013, compared to roughly 1.8 million among females in the same year. This gap underscores the importance of gender-specific strategies and interventions, while still recognizing that TB remains a major concern for both groups.

Where do we go from here?

How should we precede after filtering through this data?

- After reviewing the data from the World Health Organization's TB reports, we discovered that much of it is **fragmented, inconsistently labeled, and littered with missing values**. These problems hinder robust analysis and can lead to misleading conclusions if left unaddressed. Therefore, a thorough cleaning and normalization process is essential to ensure the data is accurate, consistent, and ready for deeper insights.

- **Python** serves as an **excellent tool for managing and cleaning messy data** before transferring it to a **MySQL database**. Using libraries like **Pandas**, analysts can efficiently address missing values, standardize column names, and ensure consistent formats across multiple files. The data can then be **normalized** (reshaped) for easier querying by transforming wide datasets into a tidy, long format. Once cleaned, Python's MySQL connectors or **SQLAlchemy** enable seamless insertion of the resulting tables into a **MySQL database,** creating a single, centralized source for ongoing analysis and reporting.