**QUIT: Qu**antifying the **I**mpact of **T**obacco control measures – An Assessment of Policy Effectiveness, Pricing, and Resource Allocation

Jef van Heijster, Owen Molloy, Kathrin Müller

****

Table of Contents

[Introduction 2](#_Toc192493354)

[Project Framework 3](#_Toc192493355)

[Goal 3](#_Toc192493356)

[Research Questions 3](#_Toc192493357)

[Objectives 3](#_Toc192493358)

[Data and Project Discovery 3](#_Toc192493359)

[Sources 3](#_Toc192493360)

[Variables 4](#_Toc192493361)

[Difficulties and Possible Challenges 4](#_Toc192493362)

[Data Exploration and Visualizations 6](#_Toc192493363)

[Cleaning and Pre-processing 11](#_Toc192493364)

[Data Cleaning 11](#_Toc192493365)

[Merging Datasets 12](#_Toc192493366)

[Modeling 13](#_Toc192493367)

# Introduction

Tobacco use is one of the leading causes of preventable death, causing over 8 million deaths annually. It is linked to serious health issues, including lung cancer, cardiovascular diseases, respiratory conditions, and stroke, and imposes a significant economic burden through healthcare costs and lost productivity. Prevalence varies widely across countries, from 0.4% to 35.7%. In response, the WHO launched the MPOWER program to monitor key tobacco control measures such as smoke-free laws, advertising bans, and national initiatives. These measures are tracked using standardized compliance scores to assess the effectiveness of policies and support countries in reducing tobacco-related harm globally.

# Project Framework

## Goal

To assess the effectiveness of tobacco control measures, tobacco pricing, and resource allocation on smoking prevalence, with a focus on regional and gender differences, in order to identify best practices and cost-effective strategies for tobacco control.

## Research Questions

1. **How effective are tobacco control measures in reducing smoking prevalence and which are the most impactful?**
2. **What role do gender, region, and the economic status of a country play in the level of smoking prevalence?**
3. **How does resource allocation affect the effectiveness and cost-effectiveness of tobacco control policies in reducing smoking prevalence?**

## Objectives

1. Assess the tobacco use prevalence and the coverage of MPOWER policies worldwide.
2. Evaluate the effectiveness of tobacco control measures in reducing smoking prevalence.
3. Examine regional, gender, and economic status differences in the effectiveness of tobacco control measures.
4. Examine the impact of tobacco pricing and taxation on smoking prevalence.
5. Assess the impact of resource allocation on the implementation and effectiveness of tobacco control policies.
6. Identify cost-effective tobacco control strategies based on resource allocation and smoking prevalence reduction.

# Data and Project Discovery

## Sources

The data to be used is the [WHO Global Tobacco Control Data (2000-2022)](https://apps.who.int/gho/data/node.main.Tobacco?lang=en).

From this data the following tables will be used:

* [MPOWER Overview](https://apps.who.int/gho/data/node.main.TOBMPOWER?lang=en) - Policy compliance scores
* [National tobacco control programmes](https://apps.who.int/gho/data/node.main.TOB1312?lang=en) - Tobacco resource allocation figures
* [Retail price + national tax](https://www.who.int/data/gho/data/indicators/indicator-details/GHO/gho-tobacco-control-raise-taxes-national-taxes-pack-of-20) - Prices and taxes on tobacco
* [Non-age-standardized estimates of current tobacco use](https://www.who.int/data/gho/data/indicators/indicator-details/GHO/gho-tobacco-control-monitor-current-tobaccouse-tobaccosmoking-cigarrettesmoking-nonagestd-tobnonagestdcurr)
* [Age-standardised estimates of current tobacco use](https://apps.who.int/gho/data/node.main.TOBAGESTDCURR?lang=en)

Here is a summary in spreadsheet form of the different tables to be used: [Data Audit Tobacco](https://docs.google.com/spreadsheets/d/1IEH8lHvP2CVBLz-Bp2WgJNw29s9bt-C5VdGszi9w7OI/edit?usp=sharing)

## Variables

**The Explanatory Variables can be grouped as follows:**

* Implemented Policies - MPOWER: Six tobacco control measures (1-5 scale)
* Tobacco pricing and taxes
* Resource Allocation and National Strategy on Tobacco Control: Annual budget, number of employees, existence of a national agency

**The Outcome Variables are as follows:**

* Non-age-standardised tobacco use prevalence (%)
* Age-standardisedtobacco use prevalence (%)
* Gender-stratified estimates of tobacco use prevalence (%)

A preview of the different datasets containing above variables incl. details can be found here: [Combined Table Preview](https://docs.google.com/spreadsheets/d/1ZpKaZv-KTXdETQnn2plA-efCKWIb-4iSupM8ovNNuqo/edit?gid=0#gid=0)

## Difficulties and Possible Challenges

1. **Data Quality and Measurement Inconsistencies**

Challenge: The WHO dataset relies on self-reported data from countries, which may have varying levels of accuracy and completeness.

Effect/Bias: Differences in data collection methodologies and reporting standards across countries and years can introduce measurement error and inconsistencies, affecting the reliability of compliance scores and smoking prevalence estimates.

Solution: Acknowledge and report limitations in the data and focus on identifying broader trends rather than relying on precise values.

**2.** **Time Lag Effect in Policy Impact**

Challenge: The effects of tobacco control policies (e.g., tax increases, advertising bans) may take years to manifest in prevalence rates.

Effect/Bias: Short-term analysis may underestimate policy effectiveness, while long-term trends may be confounded by socio-economic or cultural factors.

Solution: Focus on long-term trend analysis (e.g., 5-10 years) to better capture delayed policy effects and incorporate lag variables to account for delayed impacts.

**3.**  **Simultaneous Policy Implementation**

Challenge: Multiple tobacco control policies were often implemented at the same time, making it difficult to isolate the effects of each individual policy on smoking prevalence.

Effect/Bias: The effects of individual policies may be confounded, as simultaneous policy changes make it challenging to attribute changes in smoking rates to a specific intervention.

Solution: Use techniques like difference-in-differences (DiD) to assess the impact of policies separately.

**4.** **Comparability Across Countries**

Challenge: Variations in healthcare systems, policy environments, and socio-economic contexts across countries make direct cross-country comparisons difficult.

Effect/Bias: Disparities in healthcare infrastructure, economic conditions, and cultural factors can obscure the true effects of tobacco control policies.

Solution: Acknowledge and report limitations in the data.

**5.** **Temporal Heterogeneity**

Challenge: The prevalence dataset contains data for different years across countries, while policy adherence is often measured at different time points.

Effect/Bias: This temporal misalignment can make it difficult to directly compare the impact of policies across time and between countries.

Solution: Focus on common time points (e.g., 2010 and 2020) to conduct before-and-after analyses of policy effects.

**6. E-cigarette use data not included**

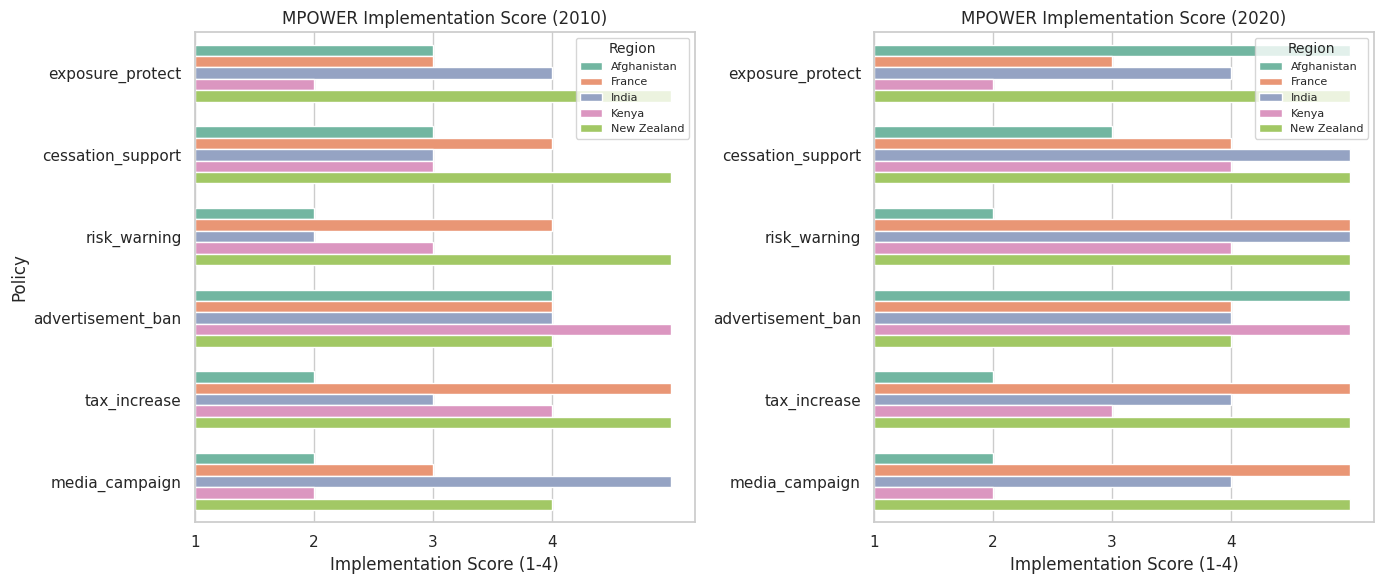
Challenge: The tobacco prevalence scores we are using refer to data regarding the use of tobacco products, which include cigarettes, pipes, cigars, cigarillos, waterpipes (hookah, shisha), bidis, kretek, heated tobacco products, and all forms of smokeless (oral and nasal) tobacco. The data excludes e-cigarettes, “e-cigars”, “e-hookahs”, JUUL and “e-pipes”.

Effect/Bias: This is a potential limitation of the data as e-cigarettes have become incredibly popular, particularly among younger people so our study does not capture this.

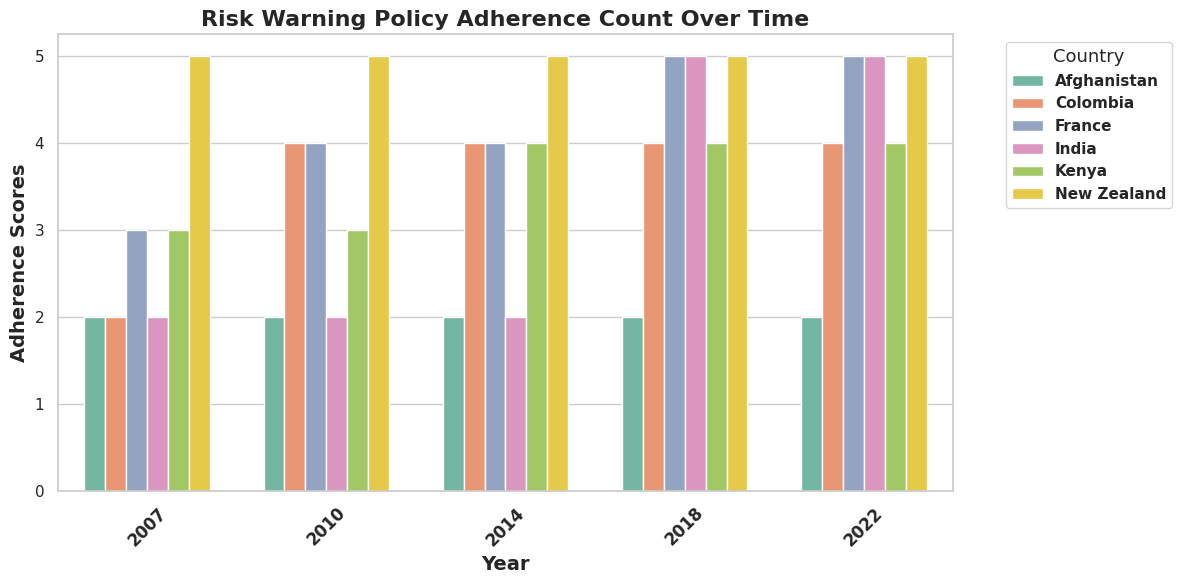
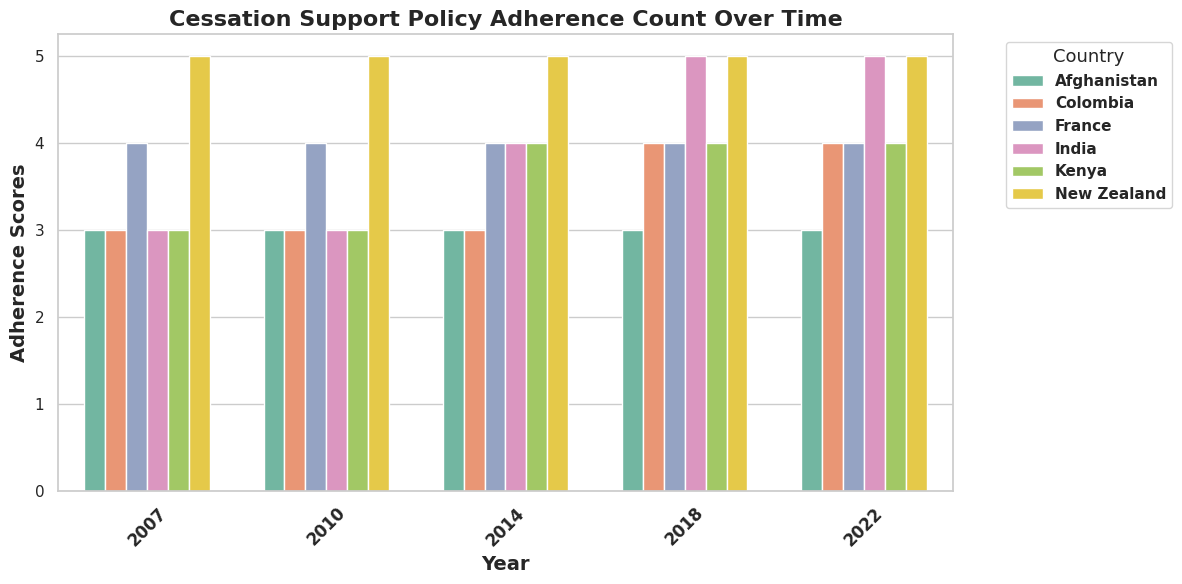
Solution: Acknowledge and report limitations in the data.

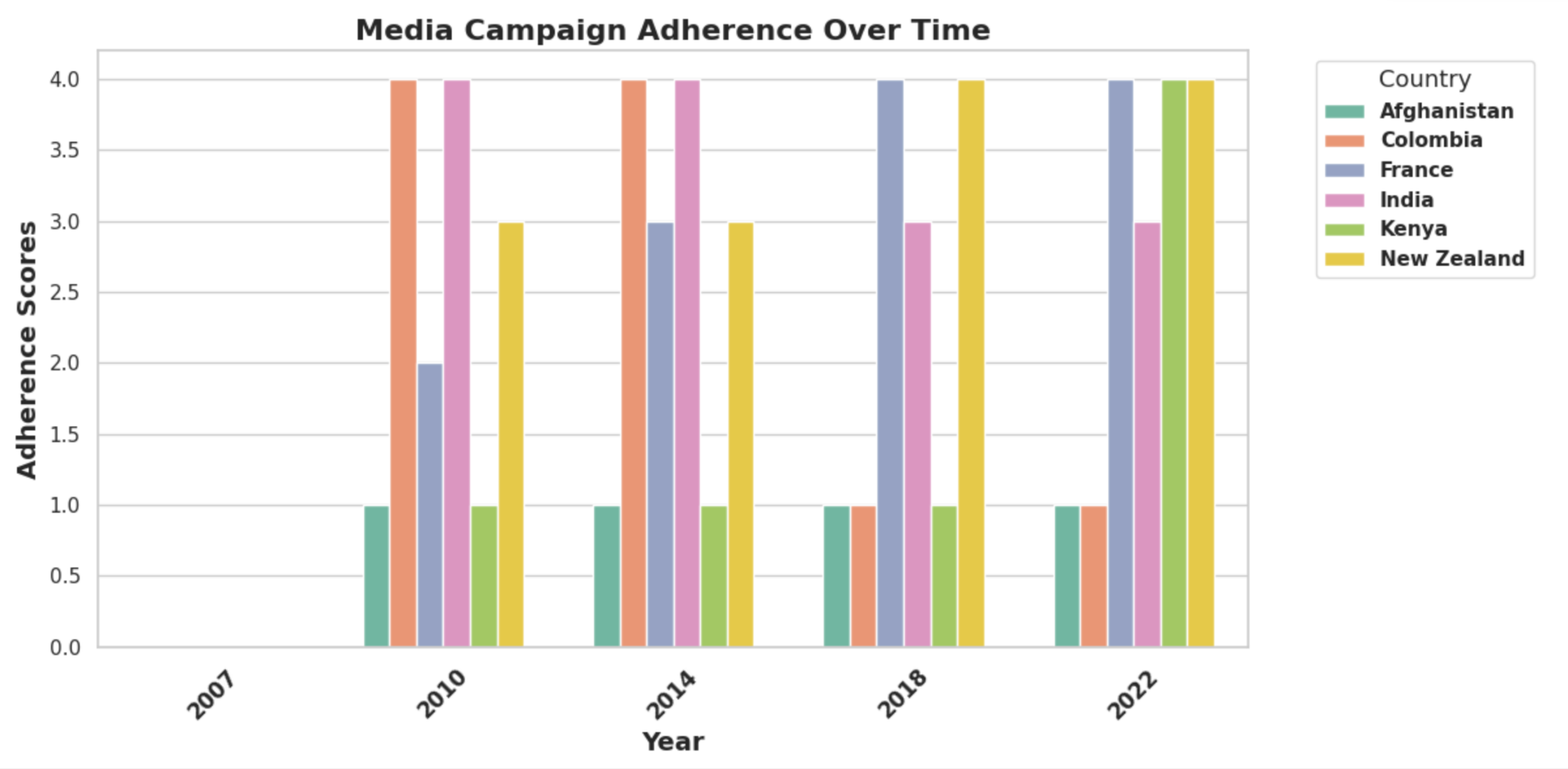
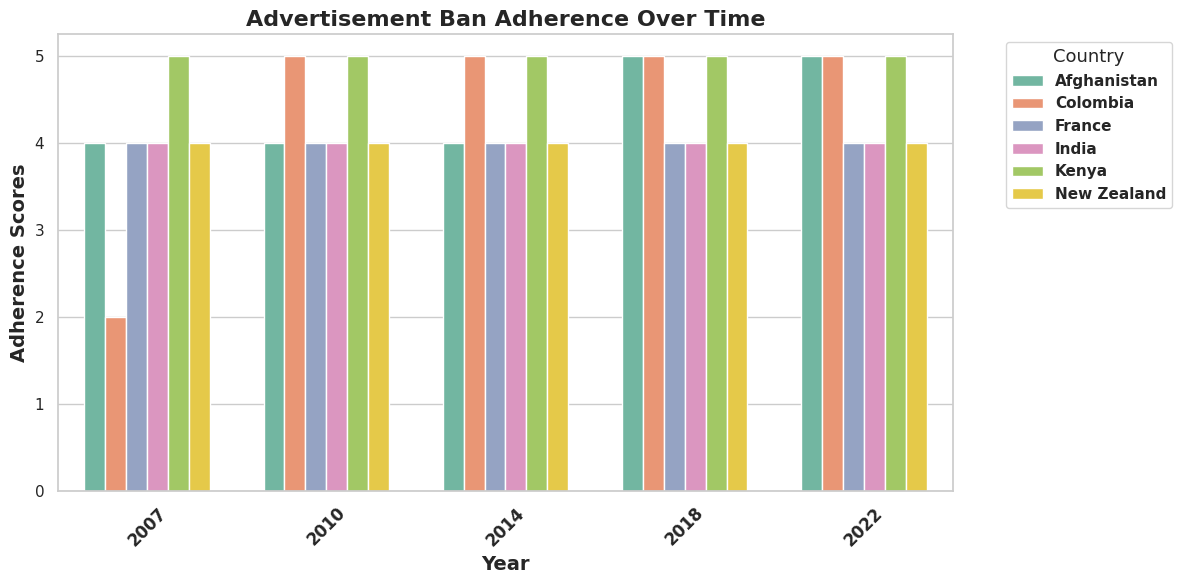
# Data Exploration and Visualizations

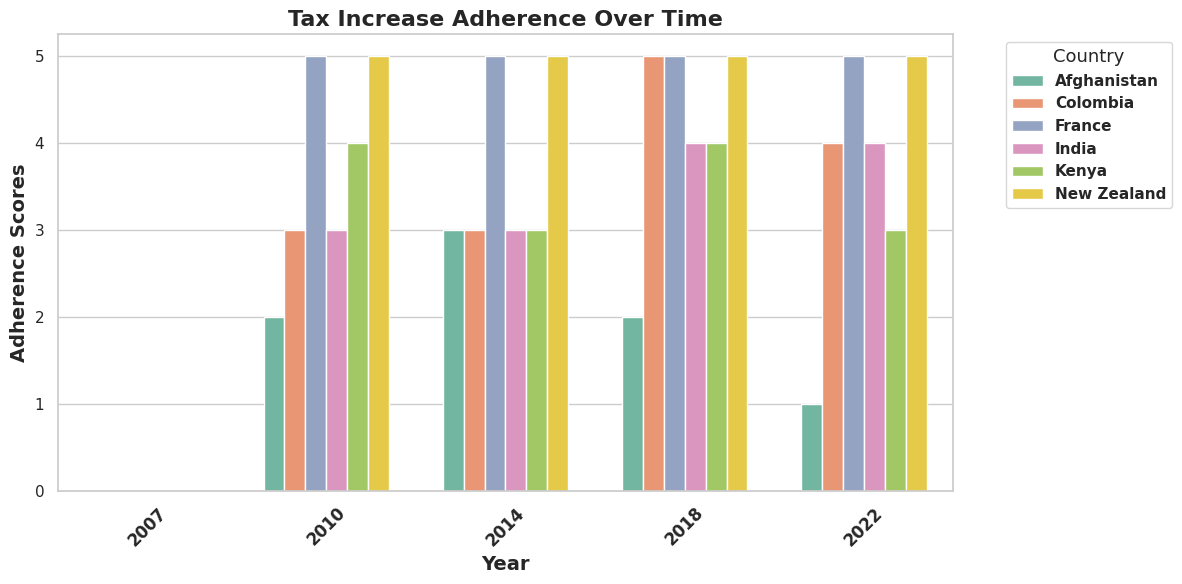
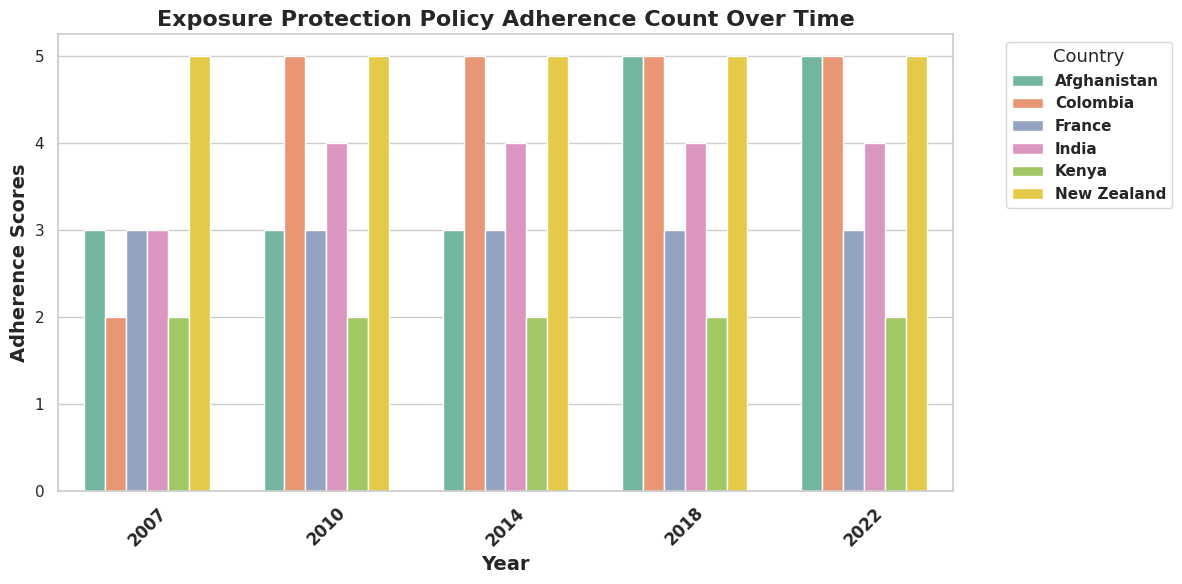
**MPOWER Policies Implementation Score 2010 vs. 2022 across 5 countries**

****

In 2010, Advertisement Bans were the most implemented policies, while Media Campaigns and Risk Warnings had the lowest adherence. Over the next decade, policy improvements varied across countries, ranging from 0 to 5 points. Risk Warnings showed the greatest progress (+5), whereas Tax Increases remained unchanged. New Zealand had already reached high adherence in 2010, achieving full implementation (level 5) by 2020, while India made the most substantial overall improvements.

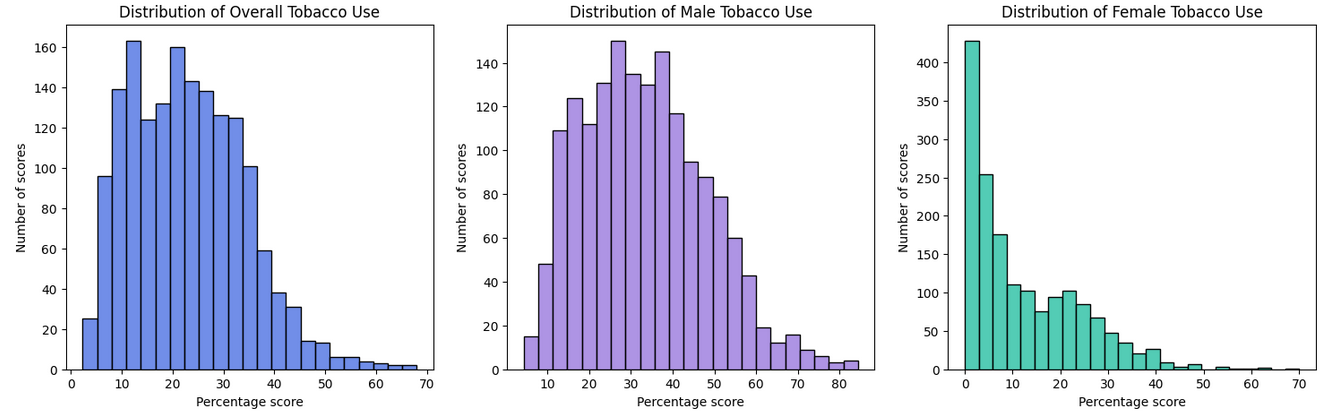
**MPOWER Policies implementation 2007 until 2022 across 5 Countries**



****

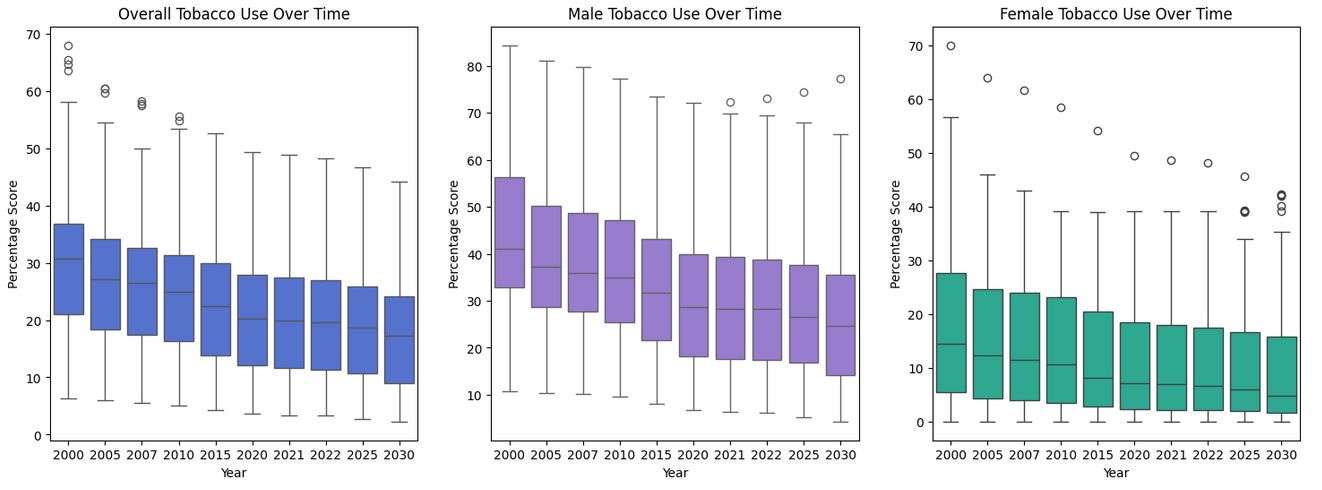
Across all policies, Health Warnings saw the most significant global improvement, with most countries advancing from low (levels 2-3) to high implementation (levels 4-5) by 2022. Conversely, Tax Increases showed minimal progress, with most countries maintaining their initial levels. While some policies saw steady, widespread improvements, advancements were uneven—Colombia and India made the greatest strides, whereas New Zealand had already achieved near-full adherence from the start.

**Distribution of Tobacco use**

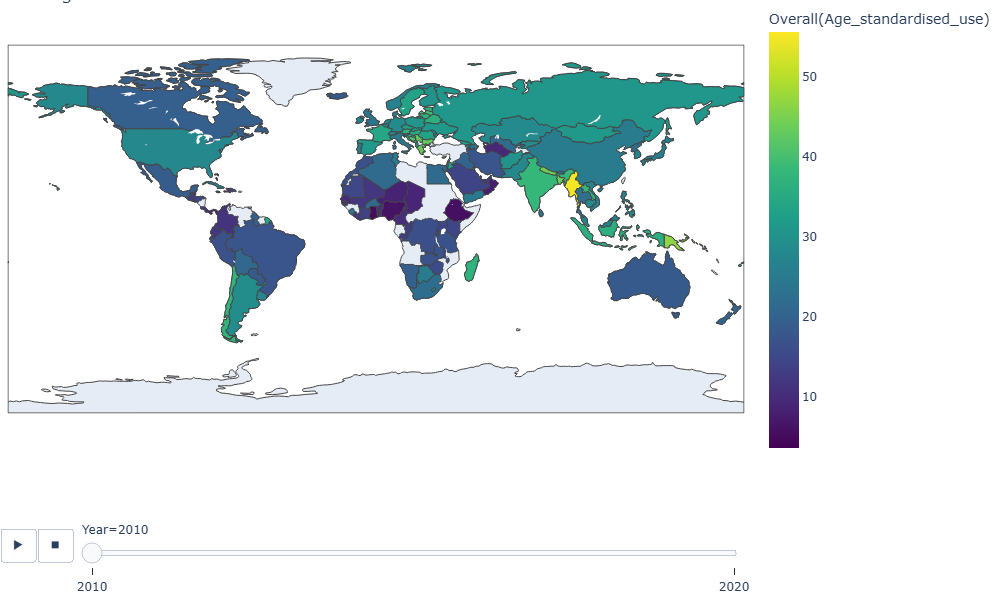
****

Here we see that the majority of percentage scores for Overall tobacco use lie between roughly 5 and 35%. When we look at the distribution of male tobacco vs female tobacco use we can see that males are more likely to be heavier users of tobacco with the majority of scores falling between roughly 13 and 50% whereas for females the majority of scores lie under about 18%. Are there certain policies or measures that influence the lower scores for females or is it related more to cultural, regional or socio-economic factors potentially?

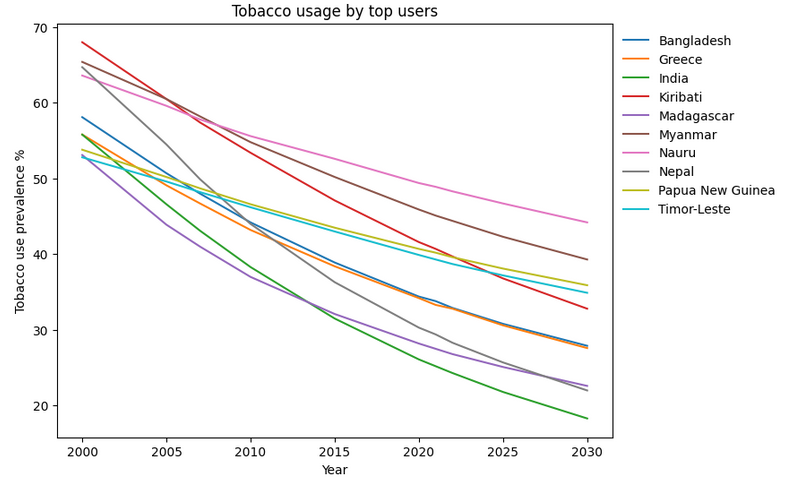
**Distribution of Tobacco use Over Time**

****

We can see that for all groups (Overall population, Males, and Females) tobacco use decreased over time. Again, we can clearly see the trend that males tend to use tobacco more than females in general by looking at the median scores and the interquartile ranges. For males we can see that the upper whiskers are quite long, indicating a larger spread of scores in the upper range. For females there is a consistent presence of high outliers. It would be interesting to see which country/countries these belong to.

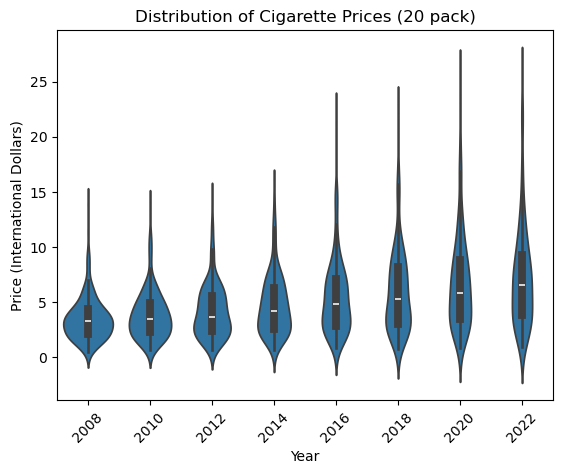


## 

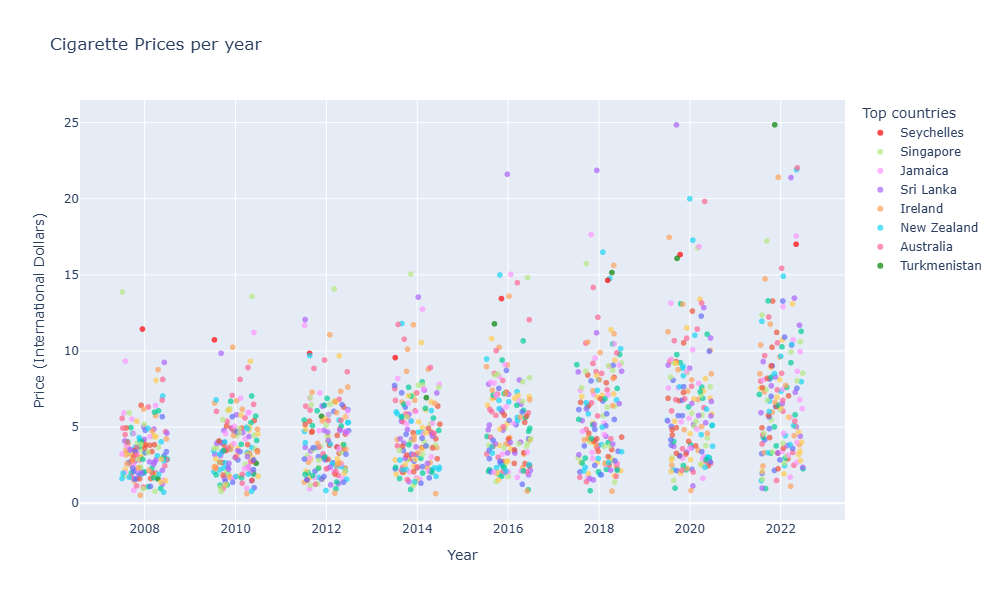
Above is a comparison from 2010 to 2020 of tobacco usage across the world, where once more we can see a general trend of decreasing tobacco usage globally.

Even the countries with the heaviest tobacco use reduced their tobacco usage over time. The countries that showed the greatest decrease among the top users were Nepal (65 - 25%), India (55 - 20%) , and Kiribati (67 - 35%). For further analysis it would be interesting to look at the countries that decreased their usage the most over time and then analyse to see if there were some common factors that influenced this decrease.

**Cigarette Pricing**

****

At the beginning of our time frame, international cigarette prices were largely clustered around the median value, with the highest values in the data being more extreme than the lowest values. Over time, the median value does not change too much, but there is a drastic increase in our maximum values, and also greater spread in our data as countries cluster less around the median.

****

This graph allows us to view the countries for whom these more extreme values belong. The aim will be to see, for these countries with notably high cigarette prices, if those countries also have any notably lower tobacco use prevalence.

# Cleaning and Pre-processing

## Data Cleaning

As we examined our separate tables, we removed irrelevant columns, checked and converted data-types, renamed columns for clarity and consistency across the different tables, and checked for duplicates and missing values, replacing them where necessary. A summary of the data cleaning steps for each table is available in the first table: [Data Cleaning/Merging Tables](https://docs.google.com/spreadsheets/d/1JPhyIRO1mJDPgj_pSe2m2juYC-oHmHCCBnVLNGV0Fyo/edit?usp=sharing)

Some of the main challenges we faced were:

1. **MPOWER Table**

* Policy implementation was originally rated on a 1-5 scale, with 1 indicating no data. To improve clarity, we adjusted the scale to 0-4, where 0 now represents missing data.

1. **Tobacco Control Table**

* The data in the “Annual budget for tobacco control in currency reported" column was kept as a key indicator of national tobacco control funding, but 45% of values were missing, and each value was in a non-standardized, country-specific currency. Converting to a common currency would require adjusting for yearly exchange rates and standardizing for purchasing power. The decision was made to drop this column, along with ‘Budget year’ and 'National tobacco control budget - currency reported', as missing values resulted from unavailable data rather than partial reporting. However, we may reintroduce this variable for a targeted analysis of countries with complete data.

1. **Cigarette Price**

* Most of the columns in this table were removed. The table originally included both taxes and cigarette prices, but we kept only the latter, as taxes expressed as a percentage of prices did not provide additional meaningful insights. We also removed cigarette prices in local currency and US dollars, opting instead for prices in international dollars—a hypothetical unit that reflects the same purchasing power parity (PPP) as the US dollar at a given point in time. This allows for more accurate cross-country comparisons without exchange rate distortions.

1. **IncomeGroup and Continental Classification**

* With WHO datasets for 162 countries (‘Region’), we aimed to categorize them into smaller groups by integrating World Bank data on Income Group (Low, Lower Middle, Upper Middle, High) and continental classification (South Asia, Europe & Central Asia, Middle East & North Africa, East Asia & Pacific, Sub-Saharan Africa, Latin America & Caribbean, North America). Since naming conventions differed, we used FuzzyWuzzy for approximate matching and filled unmatched entries using a dictionary linking countries to their income group and continent.

## Merging Datasets

Merging our datasets presented several challenges, requiring adjustments for consistency and completeness. The steps we took to merge our tables can be found [here](https://docs.google.com/spreadsheets/d/1JPhyIRO1mJDPgj_pSe2m2juYC-oHmHCCBnVLNGV0Fyo/edit?gid=931052335#gid=931052335) on the second tab and the main challenges and mitigations are also summarised in following:

1. **Data Removal**

* Tobacco usage tables included predictions beyond 2022 and data before 2007, which lacked corresponding policy or price data. These were removed (predictions could possibly be compared with machine learning predictions).
* To maintain consistency, we kept only countries present across all datasets, reducing the total to 162 countries.

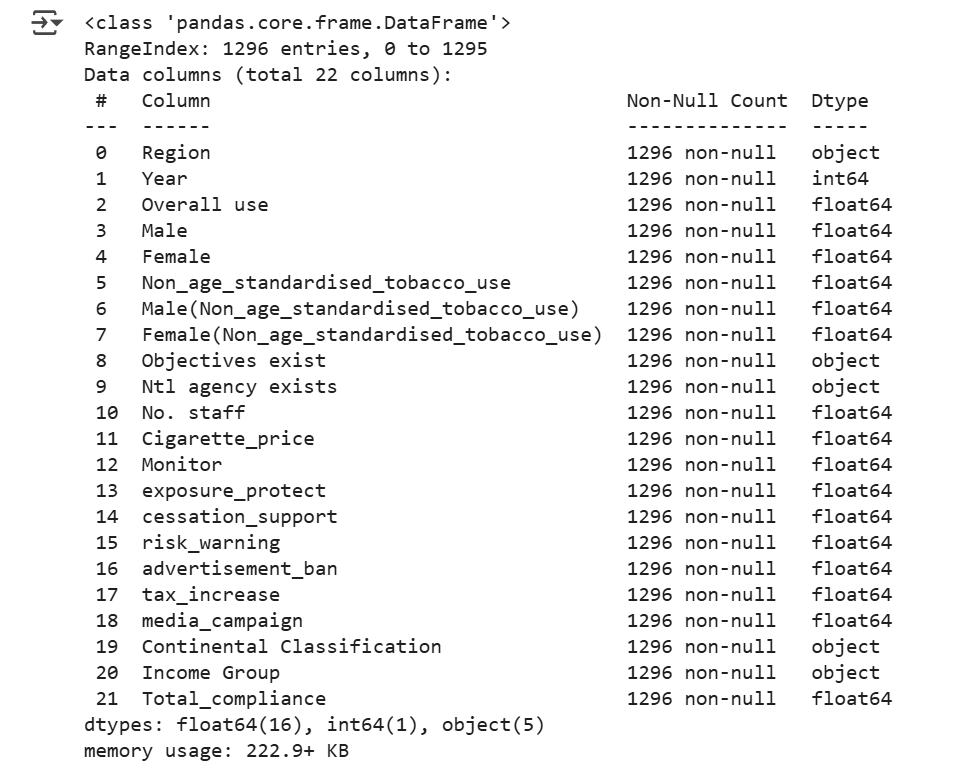
1. **Enhancing Data with New Variables**

* We created a Total Policy Adherence column, summing policy scores for each country and year, allowing for a broader assessment of overall MPOWER policy implementation.

1. **Handling Missing Values**

* Our explanatory variable tables contained data at two-year intervals (2008-2022), but tobacco usage data had mismatched years.
* Instead of dropping large portions of data, we applied linear interpolation to estimate missing values. Testing on 2022 data showed an accuracy within 0.3%, validating this approach.
* To ensure a consistent time series, we dropped odd years (2007, 2015, 2021), resulting in a dataset with biennial data from 2008 to 2022.

Below can be seen a screenshot of the .info() method as applied to our merged table.



# Modeling

**After merging the data, we proceeded with machine learning.**

1. **Methodology**Since our target variable is quantitative (tobacco use percentage per country), we chose regression models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor for comparison, using ‘Overall use’ as the target variable.
2. **Preprocessing**To prepare the data, we applied encoding, scaling, test-train splitting, and feature selection, ensuring optimal model performance, see table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Type** | **Modality** | **Preprocessing** | | **Causality** |
| Region | object | 162 unique values | Encoding | Target Encoding | With 162 high-cardinality values, OneHotEncoding was impractical. |
| Income Group | object | 1,Low,  2,Lower Middle,  3,Upper Middle,  4,High | Encoding | Ordinal Encoding | 4 ordered categories |
| Continental Classification | object | 1, South Asia, Europe & Central Asia,  2, Middle East & North Africa,  3,East Asia & Pacific, 4,Sub-Saharan Africa,  5,Latin America & Caribbean,  6,North America | Encoding | OneHotEncoding | 6 unordered categories, OneHotEncoding was efficient |
| Cigarette Prices | num |  | Scaling | StandardScaler | values varied greatly and Scaling ensured compatibility with Linear Regression, which is sensitive to varying values. |
| MPOWER | num |  | Scaling | StandardScaler |

#### **Test-Train Split**

* To prevent data leakage, we typically interpolate after splitting the data. However, since interpolation required complete country-specific data, we first applied interpolation and then used GroupShuffleSplit to ensure each country’s data (2008-2022) remained within the same set. This prevented cross-contamination between training and test data while maintaining valid interpolations.

#### **Feature Selection**

#### We dropped tobacco control table columns as they had low variance and were reserved for cost-effectiveness analysis, making them unsuitable for machine learning.

1. **Result**

Model performance was evaluated using **Mean Absolute Error (MAE):**

* Linear Regression: 1.08
* DecisionTreeRegressor: 1.71
* RandomForestRegressor: 1.24

→ Feature importance analysis showed that **‘Region’** dominated predictions (95%). Dropping it led to a sharp increase in MAE:

1. Linear Regression: 7.23
2. DecisionTreeRegressor: 8.39
3. RandomForestRegressor: 7.36

#### **Conclusions**

Country-specific factors are the strongest predictors of tobacco use, while policy adherence and cigarette prices alone are insufficient. Surprisingly, Linear Regression outperformed more complex models, likely due to the linear trend in tobacco usage within each country.

#### **Next Steps**

We will **tune tree-based models** for better performance and explore **Gradient Boosting** for improved accuracy. We will also run some statistical models such as DiD and Regressions.