**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***](#_heading=h.gjdgxs)

[***Project Framework 3***](#_heading=h.30j0zll)

[**Goal 3**](#_heading=h.1fob9te)

[**Research Questions 3**](#_heading=h.3znysh7)

[**Objectives 3**](#_heading=h.2et92p0)

[***Data and Project Discovery 3***](#_heading=h.tyjcwt)

[**Sources 3**](#_heading=h.3dy6vkm)

[**Variables 4**](#_heading=h.1t3h5sf)

[**Difficulties and Possible Challenges 4**](#_heading=h.4d34og8)

[***Data Exploration and Visualizations 6***](#_heading=h.2s8eyo1)

[***Cleaning and Pre-processing 11***](#_heading=h.17dp8vu)

[**Data Cleaning 11**](#_heading=h.3rdcrjn)

[**Merging Datasets 12**](#_heading=h.26in1rg)

[***Modeling 13***](#_heading=h.lnxbz9)

# 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 on smoking prevalence, with a focus on regional and gender differences, in order to identify best practices for tobacco control.

## Research Questions

1. **How effective are tobacco control measures in reducing smoking prevalence and which are most impactful?**
2. **What role does region, and its Income Group of a country play in the level of smoking prevalence?**
3. **How does gender affect the effectiveness of tobacco control measures 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, including tobacco pricing and taxation.
3. Examine regional, and economomical differences in the effectiveness of tobacco control measures.
4. Examine the impact of gender on the effectiveness of tobacco control measure on smoking prevalence.

# 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 datasets 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.

1. **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.

1. **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 to check interactions between policies or difference-in-differences (DiD) to assess the impact of policies separately.

1. **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: Aiming to control for contextual factors.

1. **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.

1. **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”.

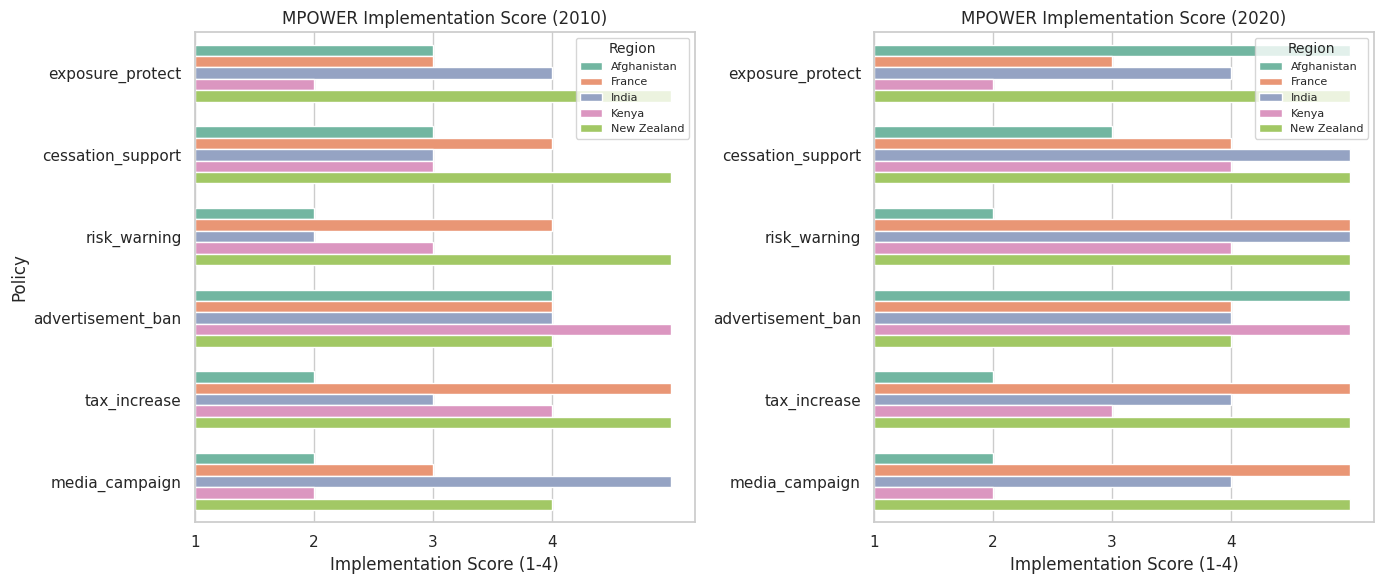
Effect/Bias: This is a potential limitation of the data as e-cigarettes have become incredibly popular, particularly among younger people so our study does not capture this.

Solution: Acknowledge and report limitations in the data.

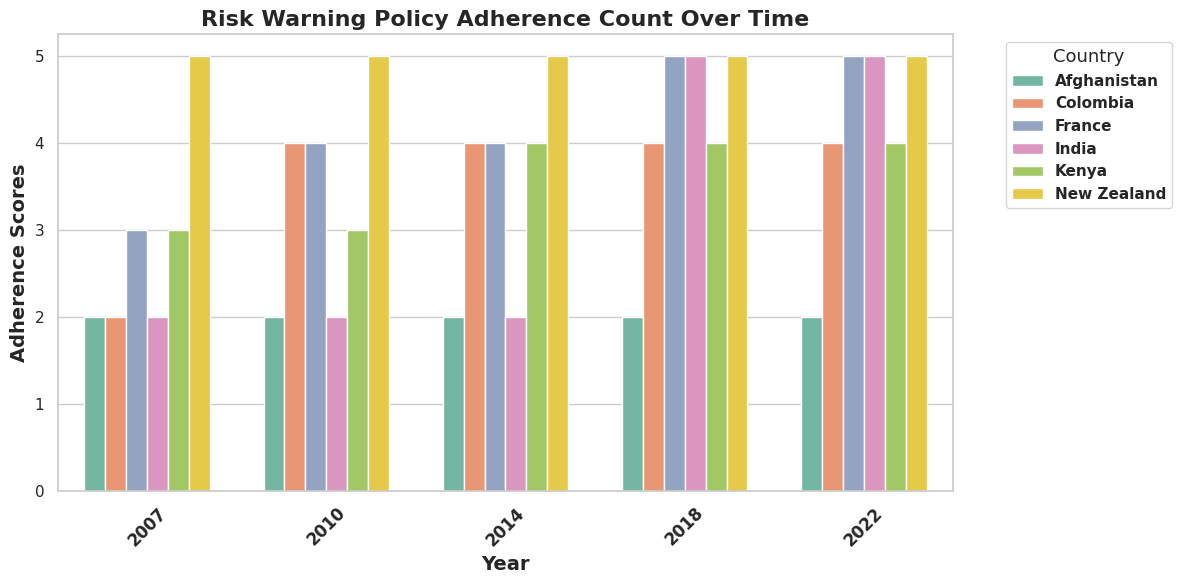
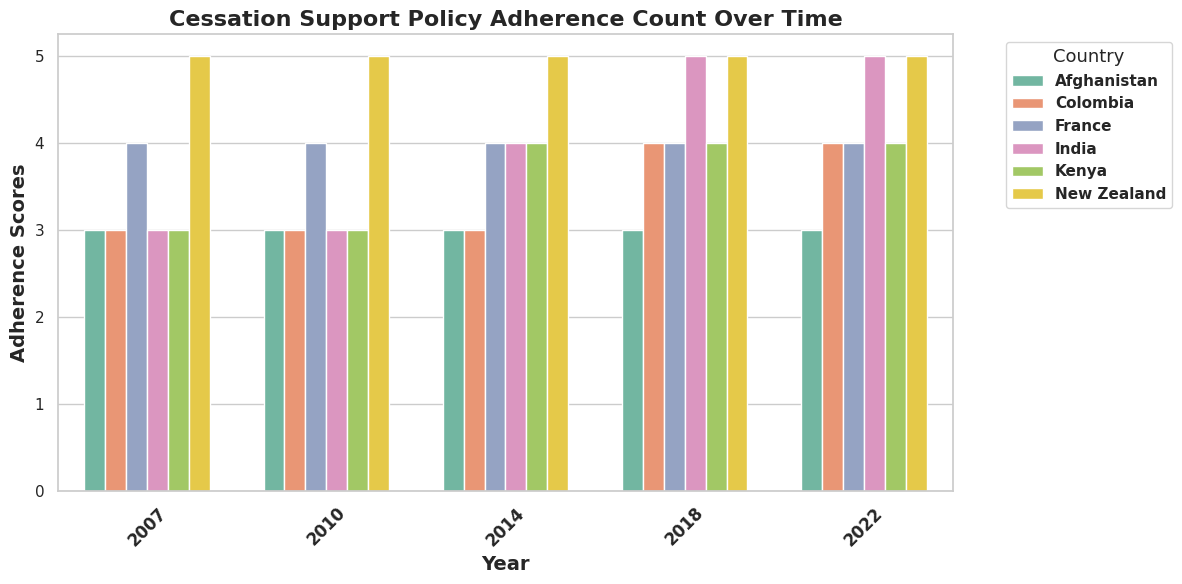
# Data Exploration and Visualizations

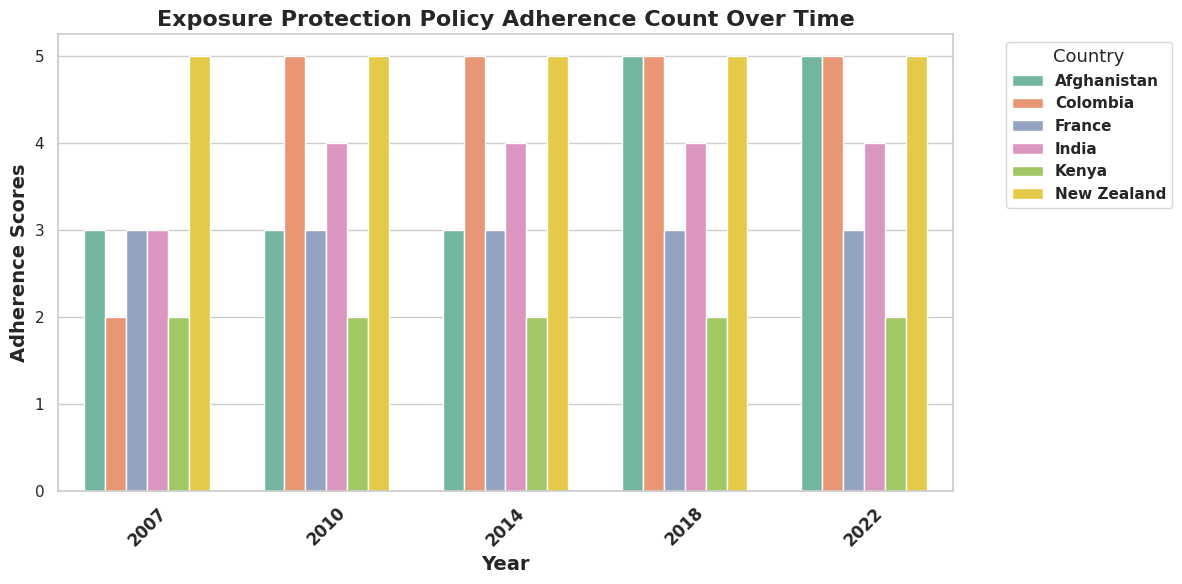
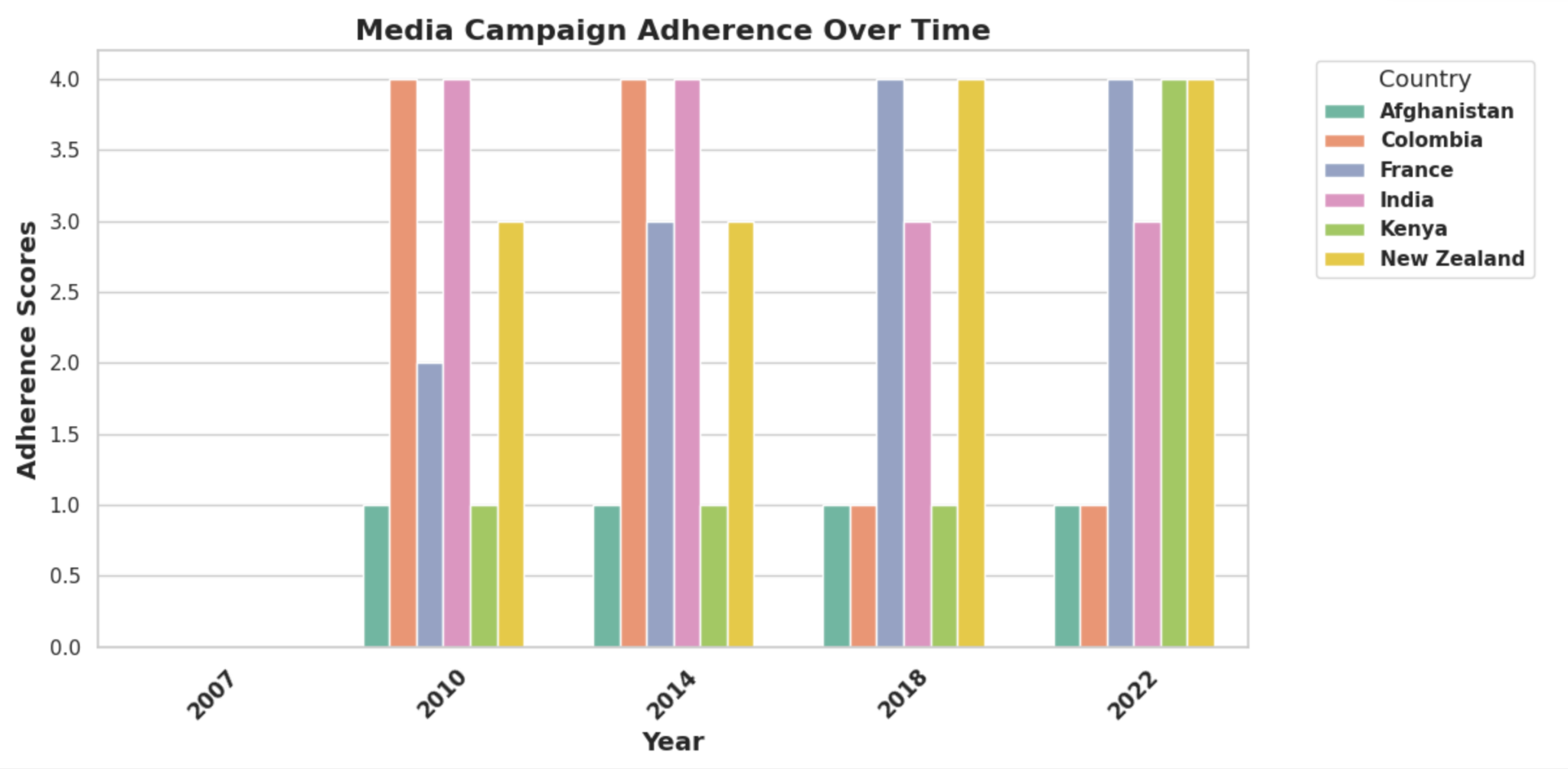
To systematically explore our dataset, we created separate visualizations for each key variable, allowing us to examine trends in policy implementation, tobacco use distribution, and pricing over time. Below, we present these visualizations along with key observations.

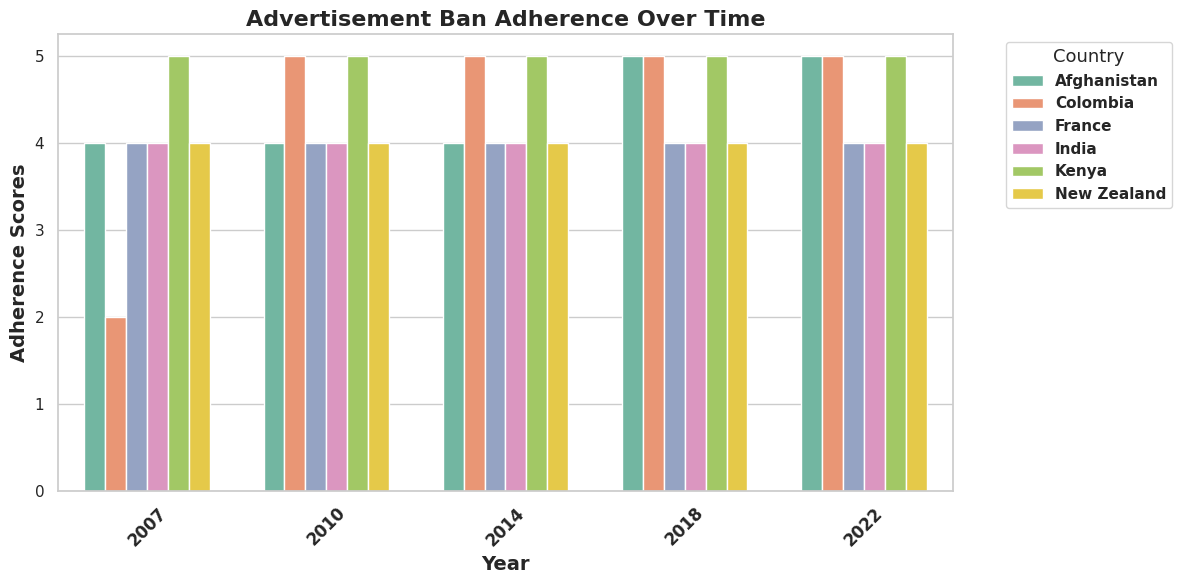
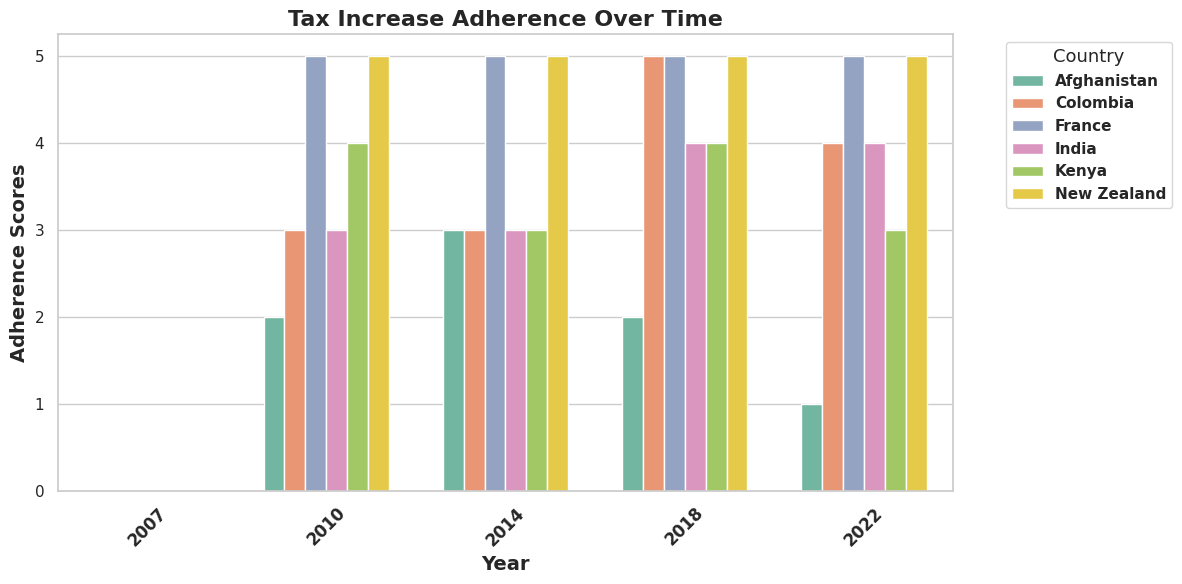
**MPOWER Policies Implementation Score 2010 vs. 2022 across 5 countries**

****

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**

**A graph of a distribution of tobacco use

AI-generated content may be incorrect.**

The above graphs show the distribution of the target variables (Standardised scores for overall, male, and female tobacco use). We can 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%.

**Distribution of Tobacco use Over Time**

**A graph of different sizes and colors

AI-generated content may be incorrect.**

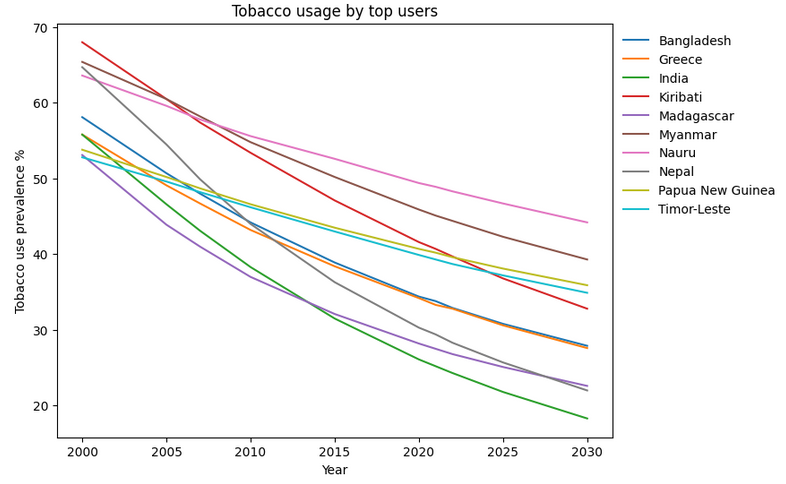
In the box plots above we can observe that for all groups (Overall population, Males, and Females) tobacco use decreased (and is predicted to continue decreasing) over time. Again, we can clearly see the trend that males tend to use tobacco more than females in general by looking at the median scores and the interquartile ranges (median scores for males ranging from 29 - 41% and for females from 7 to 15%). 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 indicating that there are one or more country/countries that have vastly different scores compared to the rest.

A map of the world

AI-generated content may be incorrect.

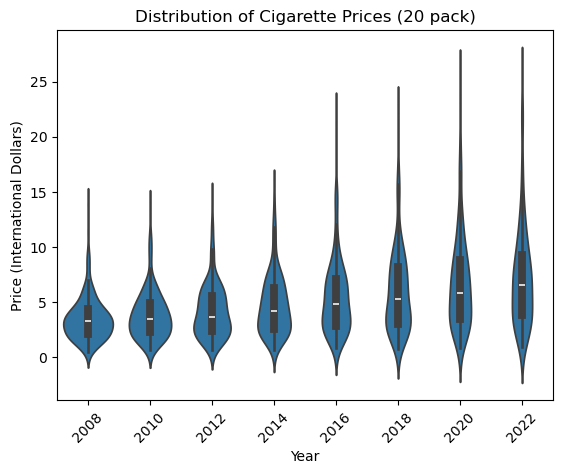
## A map of the world AI-generated content may be incorrect.

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.

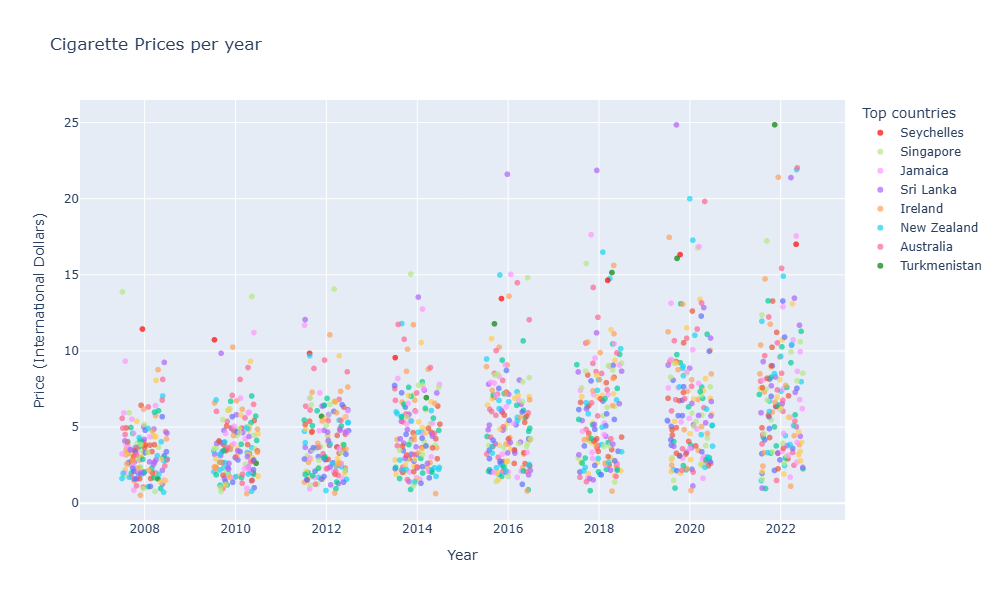


Even the countries with the heaviest tobacco use reduced their tobacco usage over time (and are predicted to continue to do so till 2030). The countries that showed the greatest decrease among the top users were Nepal (65 - 25%), India (55 - 20%), and Kiribati (67 - 35%). For further analysis it would 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**

****

In 2008 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 across countries as countries cluster less around the median.

****

This visualization helps identify which countries have the highest cigarette prices, potentially allowing us to explore whether these outliers also experience lower rates of tobacco use.

# 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 single data cleaning steps for each dataset 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 key steps taken to clean and harmonize the datasets 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 datasets can be found [here](https://docs.google.com/spreadsheets/d/1JPhyIRO1mJDPgj_pSe2m2juYC-oHmHCCBnVLNGV0Fyo/edit?gid=931052335#gid=931052335) on the second table.

Some of the key steps taken to merging the datasets were:

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.

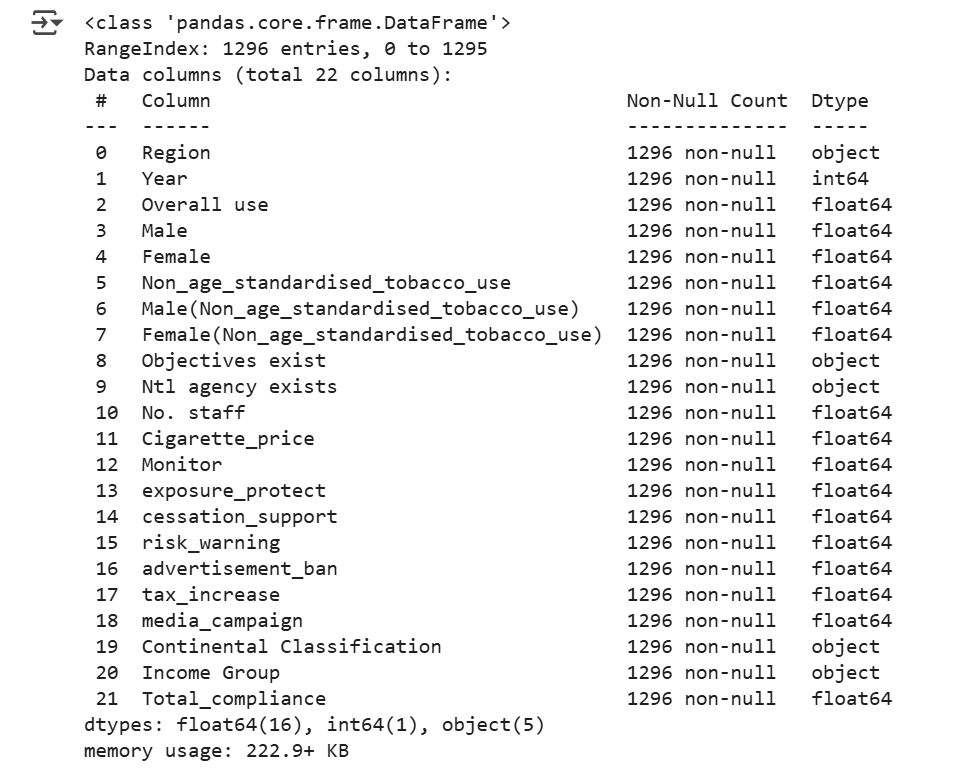
1. **Enhancing Data with New Variables**

* We created a Total Policy Adherence column, summing policy scores for each country and year, potentially allowing for a broader assessment of overall MPOWER policy implementation if needed.
* To explore the role of countrys Income Group and wider geographics (beyond country-variable:"region") on the effectivenes of ploicy implemntation on smoking prevalence we introduced two additional variables "IncomeGroup" and "Continental Classification" from a dataset from the world bank, aslo see data cleaning.

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 is a screenshot of the .info() method applied to our merged dataset:



# Modeling

## Machine Learning

After merging the datasets, 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 | Unique continuos values | Scaling | StandardScaler | Values varied greatly and Scaling ensured compatibility with Linear Regression, which is sensitive to varying values. |
| MPOWER | num | 0 - 4 Scale | 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.

#### **Implemntation Feature Selection**

#### We dropped the variables from the Tobacco Control table—containing data on resources invested in anti-tobacco policies—as they were reserved for a potential cost-effectiveness analysis and showed low variance, making them unsuitable for machine learning.

1. **Result**

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

|  |  |  |
| --- | --- | --- |
| Model | Train MAE | Test MAE |
| Linear Regression | 1.02 | 1.08 |
| DecisionTreeRegressor | 0.00 | 1.56 |
| RandomForestRegressor | 0.24 | 1.23 |

**Linear Regression** showed stable performance with minimal overfitting, making it a reliable baseline. While Decision Tree and Random Forest achieved lower training errors, their higher test MAEs indicate overfitting and limited generalizability, offering no clear advantage over the linear model.

**-->** The **Feature importance analysis** showed that 'Region' dominated predictions, accounting for 95% of model output. This is problematic, especially given our aim to identify the most impactful policies, as the model relied heavily on geography—masking the individual effects of specific tobacco control measures. To reduce this bias and allow policy variables to play a greater role in prediction, our next step was to test alternative methods to encode 'Region'.

#### **Introduction of Leave-One-Encoding**

* + **Rationale:**  
    Given that ‘Region’ overwhelmingly dominated feature importance, we aimed to reduce its influence while still capturing its relationship with smoking prevalence. This was essential to allow policy and pricing variables to contribute more meaningfully to the model and to better assess their impact.
  + **Implementation:**  
    We applied Leave-One-Out (LOO) encoding—a form of target encoding that replaces each region with the mean of the target variable, excluding the current observation to avoid overfitting.
  + **Result:**  
    The LOO-encoded model performed well, with a test MAE of 0.99, indicating good generalisability. However, ‘Region’ still contributed over 95% to feature importance, prompting us to remove it entirely in order to shift the model’s focus toward policy-relevant variables.

1. **Removing ‘Region’ (=Countries/Region) and Continential Class**
   * **Rationale:**  
     To better capture the impact of tobacco control policies we first removed 'Region', and then 'Continental Classification' as the model still relied heavily on respective to make predictions.
   * **Implementation:** we first dropped ‘Region’ and subsequently removed ‘Continental Classification’ as well.
   * **Result:**  
     Removed "region"(countries/region): High feature importance scores for Latin America (20%), Sub-Saharan Africa (16%), and Europe & Central Asia (10%) showed that geography continued to dominate. Additionally, our models showed signs of overfitting, with an MAE of 5.02 on the training set and 6.65 on the test set for our linear regression model.

Removed Continental Classification: led to a sharp increase in MAE:

|  |  |  |
| --- | --- | --- |
| Model | Train MAE | Test MAE |
| Linear Regression | 4.90 | 7.39 |
| DecisionTreeRegressor | 0.00 | 8.50 |
| RandomForestRegressor | 0.98 | 7.47 |

1. **Gender-Specific Modelling without Geographic Variables**
   * **Rationale:**  
     To assess whether policies affect men and women differently, we trained separate models by gender. Geographic variables remained excluded to focus on policy-driven effects.
   * **Implementation:**  
     Linear Regression models were run with ‘Male use’ and ‘Female use’ as target variables, using the same set of explanatory features.
   * **Result:**  
     Cigarette prices had a stronger negative effect on male smoking, indicating higher price sensitivity. Cessation support showed greater impact on women, while exposure protection had modest effects across all groups. Income group was positively associated with female smoking, suggesting cultural or socioeconomic influences.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Overall Use Coefficient** | **Male Coefficient** | **Female Coefficient** | **Interpretation** |
| **Cigarette price** | -0.72 | -2.04 | 0.60 | Higher prices correlate with a significant decrease in male tobacco use, but do not reduce female tobacco use much. This suggests men are more price sensitive. |
| **Income Group** | 1.92 | 0.35 | 3.49 | Higher-income countries have higher female tobacco use rates, while the difference is minimal in men. It may be the case that female tobacco use is more stigmatised in developing countries. |
| **Cessation support** | -1.06 | - 0.33 | -1.80 | Negative correlation with all groups, but more effective with women. |
| **Exposure protection** | -0.66 | -0.78 | -0.53 | Minimal but consistent negative correlation across all groups. |

#### **Key Takeaways**

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. This is likely because tree-based models, particularly DecisionTreeRegressor, showed significant overfitting, and relied too heavily on Region to make accurate splits.

## Machine Learning-Model optimisation

1. **Region Clustering** 
   * **Rationale:**Since the target-encoded 'Region' variable dominated our models, we explored K-means clustering as an alternative—grouping countries based on similarities in policy adherence and country economic level.
   * **Implementation:**We applied K-means clustering using MPOWER policy adherence scores, income group, and cigarette prices. The elbow method indicated two optimal clusters, which we used in place of the original Region variable.
   * **Result:**  
     Substituting Region with clustered groups did not meaningfully improve model performance. For Linear Regression, the MAE remained high (Train: 4.94, Test: 7.22), reinforcing that policy and Income Group alone cannot fully account for variation in tobacco use prevalence.
2. **Time-Lagged Model for Policy Effects:**
   * **Rationale:**  
     Since tobacco control policies may not have an immediate effect on smoking behavior, we introduced a time-lag to better capture delayed policy impact and improve model interpretability.
   * **Implementation:**  
     We applied a two-year lag to all policy-related variables, aligning interventions with their expected outcomes over time.
   * **Result:**  
     Despite this adjustment, model performance worsened (MAE increased to 7.43 on training and 7.75 on test data), supporting the conclusion that policy variables alone do not fully explain tobacco use prevalence.
3. **Model Tuning**

To improve model performance, we tuned the Random Forest Regressor using RandomSearchCV to optimise hyperparameters and also tested a more complex model, XGBoost. However, neither approach outperformed our baseline Linear Regression model. While tuning slightly reduced test error for Random Forest, overall performance remained weaker, with XGBoost showing no advantage despite its complexity.

|  |  |  |
| --- | --- | --- |
| Model | Train MAE | Test MAE |
| RandomForestRegressor  (default) | 0.98 | 7.47 |
| RandomForestRegressor  (tuned) | 1.31 | 7.26 |
| XGBoost | 2.33 | 7.45 |

1. **Ridge Regression Modelling**
   * **Rationale:**  
     To reduce the model’s reliance on ‘Region’, we tested Ridge Regression. We opted for Ridge over Lasso, as Lasso tends to eliminate low-coefficient features entirely, while we aimed to retain all variables in the model.
   * **Implementation:**  
     Ridge Regression applies regularisation by penalising large coefficients, with the strength of the penalty controlled by the alpha parameter. We tested increasing alpha values to assess whether this would reduce the dominance of ‘Region’.
   * **Result:**  
     The coefficient for ‘Region’ remained high until extreme alpha values were applied. Even at an alpha of 1000, ‘Region’ had a coefficient of 4.77—still far above the highest-scoring policy variable (Raise Taxes: 0.90). Model performance declined in parallel, with MAE reaching 4.46, indicating that Ridge Regression was not effective in reducing Region dominance without sacrificing predictive accuracy. We therefore conclude that Ridge regression was not a viable method of balancing Region dominance while retaining model performance.



1. **Stratification by Income Level**
   * **Rationale:**  
     Since previous adjustments failed to fully account for the variation previously captured by ‘Region’, we stratified the data by Income Group to reflect potential differences in policy effectiveness across economic contexts.
   * **Implementation:**

We trained separate Linear Regression models for high-income and low-income countries. and carried out a feature importances analysis.

* + **Result:**  
    Model performance improved compared to previous region-removed models, with MAE of 5.82 for high-income and 7.01 for low-income countries. In addition, training our models separately on high and low income countries allowed us to analyse the differing feature importances, provided by the linear coefficients values gathered from our Linear Regression models:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **LIC Coefficient** | **HIC Coefficient** | **Interpretation** |
| **Raise Taxes** | 4.38 | 4.37 | Strong positive correlation → Perhaps tax increases are implemented in response to high smoking rates, and are a lagging indicator |
| **Cigarette price** | 1.06 | -2.46 | Counter intuitive → Possible explanations include black market alternatives in lower income countries, or lower price elasticity. The small sample size of our low income country group may also skew results. |
| **Cessation Support** | -1.33 | -1.48 | Offering help to quit smoking is one of the few interventions that works consistently across income levels. |
| **Exposure Protection** | -1.31 | 0.12 | More effective in lower income countries, potentially they have more recent bans and a higher degree of public smoking. |

### **Stratification by Continental Class**

Furthermore, we decided to train individual models for each continential class examine how feature importance varies across different continents. Despite the smaller sample sizes, these continent-specific models achieved lower MAE scores compared to the overall model without 'Region', reinforcing the idea that stratifying by geographic group can improve performance and reveal region-specific policy effects. The results were as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Continental Classification** | **MAE** | **Greatest Coefficient** | **Second Great Coefficient** | **Third Greatest Coefficient** |
| **South Asia** | 6.52 | Year (-5.561) | Advertisement Ban (2.157) | Risk Warning (1.992) |
| **Europe & Central Asia** | 6.03 | Cigarette price (-3.875) | Tax Increase (3.057) | Advertisement Ban (-1.739) |
| **Middle East & North Africa** | 5.81 | Risk Warning (-3.570) | Tax Increase (3.226) | Exposure Protect (3.085) |
| **East Asia & Pacific** | 13.29 | Risk Warning (-4.612) | Cessation Support (-3.296) | Advertisement Ban (2.859) |
| **Americas** | 4.54 | Tax Increase (3.509) | Exposure Protect (-2.142) | Risk Warning (1.948) |
| **Sub-Saharan Africa** | 6.56 | Cigarette price (3.000) | Risk Warning (-1.858) | Tax Increase (1.137) |

**Key Takeaways:**

* + Tax increase appears as an important factor for 4 out of the 6 continents, and is positive in all of them. This suggests tax increases to consistently be a lagging factor globally.
  + Cigarette prices are an important factor in Europe & Central Asia, but not universally. In areas like Sub-Saharan Africa, the coefficient for cigarette prices is positive, reflecting what we found in our income stratified models.
  + Risk warnings are important globally, appearing for 5 of the 6 continents, and while the impact varies, on the whole it is negatively associated with tobacco use.
  + The importance of year in South Asia suggests a natural decline in tobacco use independent of policy interventions, unlike for other continents.

## --- to be continued ---

## Statistical Modeling

## **1. Initial Fixed Effects Model**

* **Rationale:**  
  We applied a Fixed Effects (FE) model to control for unobserved heterogeneity across countries and over time. This approach addresses the panel structure of our dataset, where observations are not independent—tobacco use in one year is likely correlated with previous years. By controlling for all time-invariant country characteristics and global time trends, the FE model helps isolate the effect of policy changes on smoking prevalence.
* **Implementation:**  
  We included fixed effects for both country (C(Region)) and year (C(Year)). Country fixed effects control for structural, time-invariant differences between countries, such as cultural norms or health system capacity. Year fixed effects account for global shifts, including economic changes or international tobacco control momentum.
* **Result:**  
  The model explained 98.2% of the variance in tobacco use (R² = 0.982), with an adjusted R² of 0.979. While this indicates a strong overall fit, the extremely high values suggest potential multicollinearity and risk of overfitting. The F-statistic confirmed overall model significance. Country and year fixed effects were mostly significant, with year dummies reflecting a consistent global decline in tobacco use since 2008. Regarding individual policy variables, only **risk warnings** had a robust, significant negative association with smoking prevalence. **Advertisement bans** showed a significant but unexpected positive effect, warranting further investigation. Other policies—including cigarette prices, tax increases, media campaigns, cessation support, and exposure protection—were not statistically significant or showed only borderline effects.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Coefficient** | **P-value** | **Interpretation** |
| **Cigarette\_price** | 0.0247 | 0.408 | Not significant (p > 0.05). Higher cigarette prices did not show a measurable impact on tobacco use. |
| **exposure\_protect** | -0.1501 | 0.051 | Borderline significant (p ≈ 0.05). Exposure protection policies may slightly reduce smoking prevalence. |
| **cessation\_support** | 0.0910 | 0.369 | Not significant (p > 0.05). Cessation programs showed no strong impact on reducing smoking. |
| **risk\_warning** | -0.3382 | 0.000 | Significant (p < 0.01). Risk warning labels were effective in reducing tobacco use. |
| **advertisement\_ban** | 0.2972 | 0.001 | Significant but unexpected. The positive coefficient suggests advertisement bans may have been ineffective or even correlated with increased smoking. Needs further investigation. |
| **tax\_increase** | 0.0471 | 0.644 | Not significant (p > 0.05). Tax increases did not show a strong effect. |
| **media\_campaign** | -0.0146 | 0.743 | Not significant (p > 0.05). Media campaigns did not have a measurable effect on reducing tobacco use |

## **2. Multicollinearity Check (VIF Analysis)**

### **Rationale**

* **Potential Multicollinearity inflates standard errors**, making it difficult to isolate the true effects of individual policies.
* High correlation between policy variables (e.g., tax increases and cigarette prices) could explain why many policies appear insignificant.

### **Results**

* High VIFs (>10) for several policies, particularly cessation\_support (11.32) and risk\_warning (10.38), indicated significant multicollinearity, complicating the isolation of individual policy effects.

|  |  |  |
| --- | --- | --- |
| **Variable** | **VIF** | **Interpretation** |
| **cessation\_support** | 11.32 | High multicollinearity (VIF > 10) → Strong correlation with other variables, likely distorting its estimated effect. |
| **risk\_warning** | 10.38 | High multicollinearity → May cause unstable coefficient estimates, making it difficult to isolate its impact. |
| **advertisement\_ban** | 8.76 | Moderate multicollinearity (VIF ~ 5-10) → Worth investigating, as it may still influence model accuracy. |
| **tax\_increase** | 8.66 | Moderate multicollinearity → Might be correlated with cigarette prices or other policy measures. |
| **exposure\_protect** | 5.63 | Moderate multicollinearity → Manageable but still notable. |
| **media\_campaign** | 3.17 | Low multicollinearity (VIF < 5) → No concern. |
| **Cigarette\_price** | 3.83 | Low multicollinearity → No concern. |

### **Step forward:**

* **Dropping policies is not an option** as we aim to assess their individual effectiveness.
* **Potential Solution**: Introduce **interaction terms** to capture possible **policy synergies**.

## **3. Interaction Terms Analysis**

### **Rationale**

* Policies may work **synergistically** rather than independently (e.g., risk warnings may be more effective when paired with media campaigns).
* **Model Fit:** Including relevant interactions can improve explanatory power, reducing omitted variable bias.
* **Selection Method:**
  + Used a **correlation matrix** to identify moderately correlated policies (0.3 < r < 0.8).
  + Ensured **policy logic** supported interactions.
* **Implementation:** interaction terms were introduced in model two at a time
* **Result and Interpretation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Interaction Term** | **Coefficient** | **P-value** | **Interpretation** |
| **Tax × Media Campaign (tax\_media)** | -0.0396 | 0.325 | Not significant → No strong evidence that tax effectiveness improves with media campaigns. |
| **Cessation × Media Campaign (cessation\_media)** | 0.0302 | 0.538 | Not significant → No clear interaction between media campaigns and cessation programs. |
| **Media Campaign × Risk Warning (media\_warning)** | -0.0758 | 0.026 | Significant → Media campaigns reinforce risk warnings, strengthening smoking deterrence. |
| **Tax Increase × Cigarette Price (tax\_price)** | -0.0122 | 0.580 | Not significant → No evidence that tax-driven price increases have a combined effect. |
| **Exposure Protection × Risk Warning (protect\_warning)** | 0.0085 | 0.858 | Not significant → No strong evidence of an interaction effect. |
| **Tax Increase × Risk Warning (tax\_warning)** | 0.0009 | 0.987 | Completely insignificant → Suggests tax increases and risk warnings act independently. |
| **Media Campaign × Exposure Protection (protect\_media)** | -0.0055 | 0.859 | Not significant → Suggests that media campaigns do not reinforce smoking bans. |

### **→ Introduction of all 3 interaction terms that showed significance in the model:**

**Result Model with all 3 semi robust interaction terms**

* Media Warning (media\_warning): The strongest interaction effect.
* Price × Advertisement Ban: Shows potential but lacks robustness.
* Cessation × Risk Warning: Positive coefficient suggests increased quitting attempts but not necessarily lower smoking prevalence.

→ **No other interactions showed consistent significance**, and policy coefficients **varied greatly** across models, suggesting **instability and multicollinearity effects**.

## **4. Clustering Standard Errors**

### **Rationale**

* **Within-Group Correlation**: Observations within the same country over time are likely correlated due to shared cultural, economic, or policy factors. Clustering adjusts for this correlation, ensuring more reliable estimates.
* **Heteroskedasticity**: Fixed effects do not address differences in error variance across countries. Clustering corrects for heteroskedasticity, producing robust standard errors.
* **Conservative Estimates**: Clustering increases standard errors, leading to more cautious significance testing and reducing the risk of overstating policy impacts.

**Implementation**

* Run FE Model with region clustered with and without interaction term *Media Campaign × Risk Warning*

### **Result and Interpretation**

* Clustering provides **more conservative but reliable estimates**.
* Clustering **increased standard errors**, making previously significant policies **weaker or non-significant**.
* **Risk Warning remains the strongest policy** but decreased in significance (p = 0.034).
* Interaction effects did not remain robust.*Media Campaign × Risk Warning*: Lost robustness (p = 0.105).

## **5. Stratifying by Income Groups**

### **Rationale**

* Policy effectiveness may vary by **economic level**.
* Reduces **within-group variation**, allowing clearer identification of **Income Group-specific** policy effects.

**Result**

Smoking prevalence has declined over the years across all income levels, with the greatest reductions observed in LICs, followed by LMICs, HICs, and UMICs. Risk warnings were the only policy showing a significant correlation in LMICs, while media campaigns showed a borderline negative effect in HICs, suggesting a potential role in reducing smoking rates. In contrast, policy interventions in UMICs and LICs had limited impact, emphasizing the need for tailored, region-specific tobacco control strategies.

|  |  |  |
| --- | --- | --- |
| **Income Group** | **Policy Impact (Policy: Coefficient, P-Value)** | **Interpretation** |
| **HIC** | Media Campaign: −0.16,P= 0.092 | Potential small reduction in smoking; borderline significant. |
| **UMIC** | None significant | Current policies show no measurable impact. |
| **LMIC** | Risk Warning: −0.98,P= 0.003 | Moderate reduction in smoking prevalence; risk warnings are effective. |
| **LIC** | None significant | Policies show no significant effect; possible implementation challenges. |

## **6. Stratifying by Continental Classification**

### **Rationale**

* Policy effectiveness may vary by **continental class**.
* Reduces **within-group variation**, allowing clearer identification of **region-specific** policy effects.

**Result**

Smoking prevalence decreased most significantly over time in Europe & Central Asia, with a reduction of −5.49 (*P* < 0.001) by 2022. Among specific policies, advertisement bans were effective in Europe & Central Asia and the Middle East & North Africa, though these results suggest counterintuitive increases in smoking. Risk warnings in East Asia & Pacific showed significantly associated with reducing smoking prevalence.

|  |  |  |
| --- | --- | --- |
| **Continental Classification** | **Significant Policies** | **Interpretation** |
| **South Asia** | None | Limited responsiveness to tobacco control measures; affordability or informal markets may dominate tobacco use patterns. |
| **Europe & Central Asia** | Advertisement Ban  (r = 0.5763, p = 0.037) | Effective in reducing smoking prevalence when enforcement is strong and culturally accepted. |
| **Middle East & North Africa** | Advertisement Ban  (r = 0.7286, p = 0.015)  Cessation Support:  +0.61 (P = 0.079) | Effective in restricting visibility and promotion of tobacco products, reducing smoking prevalence significantly in this region.  Borderline increase in smoking (unexpected result) |
| **East Asia & Pacific** | Risk Warning  (r = -0.9942, p = 0.004)  Cessation Support:  +0.74 (P = 0.074) | Highly effective in raising awareness of health risks and deterring smoking behavior in this region through targeted campaigns.  Borderline increase in smoking (unexpected result).5 |
| **Americas** | None | Weak responsiveness to policies; socioeconomic factors or enforcement gaps may limit effectiveness of formal measures. |
| **Sub-Saharan Africa** | None | Affordability issues or informal markets likely undermine formal policy impacts in this region. |

## **Takeaways**

* Risk Warnings show the strongest negative correlation with smoking prevalence, in LMIC and East Asia & Pacific .
* Advertisement Bans are positively correlated with smoking prevalence in Europe & Central Asia and Middle East & North Africa suggesting potential unintended effects.
* No significant policy effects are observed in South Asia, the Americas, and Sub-Saharan Africa, likely due to affordability, informal markets, or enforcement gaps.
* Clustering & stratification highlight the variability of policy effectiveness, emphasizing context-dependent policy impact.

### **6. Gender-Specific Policy Effects**

#### **Rationale**

#### Tobacco control policies may impact male and female smoking behavior differently due to social norms and economic factors. Identifying these differences helps refine policy strategies.

#### **Implementation**

1. Ran **fixed-effects models** separately for **male and female smoking prevalence**, controlling for country and year inlc. **clustered standard errors** at the regional level without and with interaction term
2. stratified by Income Group
3. stratified by Continental Class

#### **Results and Interpretation**

1. without interaction term: For **men**, **risk warnings** significantly but minor reduce smoking prevalence (coef=-0.4738, p = 0.015), while all other policies, including cigarette price, advertisement bans, and media campaigns, show no significant effects. For **women**, none of the policies have a statistically significant impact, indicating either lower policy responsiveness or that other social and economic factors influence female smoking behavior more strongly.

without interaction term: Introducing **Media Campaign × Risk Warning** reduced the direct effect of risk warnings on **male smoking** (previously p = 0.015, now p = 0.222), while the interaction was **borderline significant** (p = 0.058), suggesting media campaigns may enhance risk warnings. For **female smoking**, no policies were significant before or after adding the interaction (p = 0.368), reinforcing that female smoking behavior is less responsive to these policies.

1. Risk warnings are the most effective policy in LMICs for both men and women while they show weaker or no effects in other income groups. Cigarette prices significantly reduce smoking prevalence among women in HICs suggesting price sensitivity is higher in this group.No policies significantly impacted smoking prevalence in UMICs. Female smoking prevalence appears less responsive to tobacco policies overall, except for price measures in HICs and risk warnings in LMICs.

|  |  |  |  |
| --- | --- | --- | --- |
| **Income Group** | **Male Effect (β, P-Value)** | **Female Effect (β, P-Value)** | **Interpretation** |
| **HIC** | None | Cigarette Price: -0.27 (P = 0.027) | Higher cigarette prices reduce female smoking. |
| **UMIC** | None | None | No significant effects observed for either gender. |
| **LMIC** | Risk Warning: -1.13 (P = 0.008) | Risk Warning: -0.83 (P = 0.026) | Risk warnings reduce smoking for both genders. |
| **LIC** | None | None | No significant effects observed for either gender. |

1. Smoking prevalence decreased most significantly over time in Europe & Central Asia for males and least in the Middle East & North Africa for females. Risk warnings are the most effective policy in East Asia & Pacific for both men ( and women while showing weaker or no effects in other regions. Advertisement bans show counterintuitive positive associations with smoking prevalence in several regions, particularly for women in Europe & Central Asia and Sub-Saharan Africa and for men in the Middle East & North Africa. Female smoking prevalence appears less responsive to tobacco policies overall, with some unexpected positive associations in the Americas for cessation support and risk warnings.

|  |  |  |  |
| --- | --- | --- | --- |
| **Continental Group** | **Male Effect (β, P-Value)** | **Female Effect (β, P-Value)** | **Interpretation** |
| **South Asia** | None | None | No significant effects observed for either gender. |
| **Europe & Central Asia** | Media Campaign:  -0.21 (P = 0.091) | Advertisement Ban:  +0.70 (P = 0.035) | Borderline decrease in male smoking from media campaigns. Counterintuitive increase in female smoking from ad bans. |
| **Middle East & North Africa** | Advertisement Ban:  +1.41 (P = 0.004) | None | Counterintuitive increase in male smoking from ad bans. |
| **Americas** | None | Cessation Support: +0.51 (P = 0.049) Risk Warning: +0.32 (P = 0.011) | Unexpected increase in female smoking from cessation support and risk warnings. |
| **East Asia & Pacific** | Risk Warning: -1.11 (P = 0.007) | Risk Warning: -0.88 (P = 0.049) | Risk warnings reduce smoking for both genders. |
| **Sub-Saharan Africa** | None | Advertisement Ban: +0.72 (P = 0.041) Exposure Protect: -0.62 (P = 0.086) | Counterintuitive increase in female smoking from ad bans. Borderline decrease from exposure protection. |

Overview of Policy Impact Statistical Models (keep one table !)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Policy** | **FE** | **FE Clustered** | **Income Groups** | **Continental Classification** | **Gender-specific (Male)** | **Gender-specific (Female)** | **Interpretation** |
| **Risk Warning** | -0.3382 (P<0.01) | -0.3382 (P=0.034) | LMIC: 0.98 (P=0.003) | East Asia & Pacific: -0.99 (P=0.004) | LMIC: -1.13 (P=0.008) East Asia & Pacific: -1.11 (P=0.007) | LMIC: -0.83 (P=0.026) East Asia & Pacific: -0.88 (P=0.049) | Consistently effective across models, particularly in LMICs and East Asia & Pacific. (significant overlap between LMICs and the East Asia & Pacific region.) May be due to lower baseline awareness, health literacy in these regions and risk warning policy measure with high fidelity of implementation. |
| **Advertisement Ban** | 0.2972 (P=0.001) | Not significant | None | Europe & Central Asia: +0.58 (P=0.037) Middle East & N. Africa: +0.73 (P=0.015) | Middle East & N. Africa: +1.41 (P=0.004) | Europe & Central Asia: +0.70 (P=0.035) Sub-Saharan Africa: +0.72 (P=0.041) | Delayed effects do not explain the positive association; industry adaptation, policy loopholes, or alternative marketing likely drive counterintuitive increases in smoking, complicating tobacco control efforts. Reverse causality unlikely as controlled for baseline prevalence and fixed effect model |
| **Cigarette Price** | Not significant | Not significant | HIC: -0.27 (P=0.027) | None | None | HIC: -0.27 (P=0.027) | Only effective for females in HICs, possibly due to higher price sensitivity and economic empowerment in these countries. Cigarette price increases might lack might in LICs/LMICs due to low price elasticity, illicit market access, and weak tax enforcement, limiting deterrence despite policy changes. |
| **Media Campaign** | Not significant | Not significant | HIC: -0.16 (P=0.092) | None | Europe & Central Asia: -0.21 (P=0.091) | None | Limited effectiveness, slightly more impactful in HICs and for males in Europe & Central Asia. May reflect media saturation and campaign quality differences. |
| **Cessation Support** | Not significant | Not significant | None | Middle East & N. Africa: +0.61 (P = 0.079) East Asia & Pacific +0.74 (P = 0.074) | None | Americas: +0.51 (P=0.049) | Unexpected positive effect for females in Americas might indicate increased quit attempts rather than successful cessation. The positive association of cessation support with smoking prevalence likely reflects increased engagement with quitting services without sustained cessation success, and potential additional tobacco substitution? rather than reverse causality or policy selection bias? |
| **Exposure Protection** | -0.1501 (P=0.051) | Not significant | None | None | None | Sub-Saharan Africa: -0.62 (P=0.086) | Borderline effect overall, more pronounced for females in Sub-Saharan Africa. Could reflect gender differences in exposure to public smoking. |
| **Tax Increase** | Not significant | Not significant | None | None | None | None | Lack of significance across models may be due to complex price elasticities or implementation challenges |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Policy** | **FE** | **FE Clustered** | **Income Groups** | **Continental Classification** | **Gender-specific (Male)** | **Gender-specific (Female)** |
| **Risk Warning** | -0.3382 (P<0.01) | -0.3382 (P=0.034) | LMIC: -0.98 (P=0.003) | East Asia & Pacific: -0.99 (P=0.004) | LMIC: -1.13 (P=0.008) East Asia & Pacific: -1.11 (P=0.007) | LMIC: -0.83 (P=0.026) East Asia & Pacific: -0.88 (P=0.049) |
| **Advertisement Ban** | 0.2972 (P=0.001) | Not significant | None | Europe & Central Asia: +0.58 (P=0.037) Middle East & N. Africa: +0.73 (P=0.015) | Middle East & N. Africa: +1.41 (P=0.004) | Europe & Central Asia: +0.70 (P=0.035) Sub-Saharan Africa: +0.72 (P=0.041) |
| **Cigarette Price** | Not significant | Not significant | HIC: -0.27 (P=0.027) | None | None | HIC: -0.27 (P=0.027) |
| **Media Campaign** | Not significant | Not significant | HIC: -0.16 (P=0.092) | None | Europe & Central Asia: -0.21 (P=0.091) | None |
| **Cessation Support** | Not significant | Not significant | None | None | None | Americas: +0.51 (P=0.049) |
| **Exposure Protection** | -0.1501 (P=0.051) | Not significant | None | None | None | Sub-Saharan Africa: -0.62 (P=0.086) |
| **Tax Increase** | Not significant | Not significant | None | None | None | None |

**Key takeaways:**

 **Risk Warnings** were consistently effective, especially in LMICs and East Asia & Pacific—likely due to lower baseline awareness and high implementation fidelity.

 **Advertisement Bans** showed counterintuitive positive associations with smoking prevalence, possibly due to industry adaptation, policy loopholes, or alternative marketing strategies. These effects are unlikely due to reverse causality.

 **Cigarette Prices** were only effective for women in high-income countries, suggesting gendered differences in price sensitivity. In LICs/LMICs, their impact was limited—likely due to low price elasticity, access to illicit markets, and weak enforcement.

 **Media Campaigns** had limited overall effectiveness but showed slightly stronger effects in HICs and among males in Europe & Central Asia, potentially reflecting variation in campaign quality or media reach.

 **Cessation Support** showed unexpected positive associations in some groups (e.g., women in the Americas), possibly indicating increased quit attempts without long-term success or substitution effects.

 **Exposure Protection** had a borderline overall effect, with slightly greater impact on females in Sub-Saharan Africa—possibly reflecting gender-specific differences in public exposure.

 **Tax Increases** lacked significant effects across models, potentially due to delayed behavioral responses, weak enforcement, or complex economic dynamics.

* **Overall Conclusion: answernig research question! recomendation how to implement policies !**
* 1, distribution of prevalence and implementation of policies
* 2, Policies interlink difficult to analyse separately and synergies between …
* 3, Time and country has a great impact on the prevalence itself (in all INcome Groups and continental classe prevalence decrease time most often had a decreasing effect on the prevalence, country varied?)
* policies have different impact in different Income Groups
* risk warning advertisement ban: positive effect
* Continential Class ..
* Gender

# **Summary of Core Findings and Policy Recommendation**

Our initial data exploration revealed clear gender differences in tobacco use, with males consistently exhibiting higher prevalence than females. At the same time, overall tobacco use declined globally between 2000 and 2022, with all countries in our dataset showing reductions from 2020 to 2022. Policy implementation varied substantially across countries and policy types. While some measures—such as health warnings—showed significant progress, others, like tax increases, remained largely stagnant.

From a modelling perspective, country-specific factors emerged as the strongest predictors of tobacco use. Policy adherence and cigarette prices alone were not sufficient to explain prevalence. Linear Regression outperformed more complex models, as non-linear models tended to overfit and relied heavily on geographic features.

Optimization failed .. worked ..?

**Key insights across models and stratifications:**

* **Tax increases** were positively associated with smoking in most regions, suggesting they may be implemented reactively, rather than preventively.
* **Cigarette prices** showed regional variation in effect, with strong negative associations in Europe & Central Asia but unexpected positive effects in Sub-Saharan Africa and LICs.
* **Risk warnings** consistently demonstrated negative associations with tobacco use across most regions, making them one of the most reliable policy measures.
* **Year** was a strong predictor in South Asia, indicating a general downward trend in smoking that may occur independently of specific policies.

From our Statistical Modeling ...

# Reflections and Future Research Possibilities

While our analysis provided valuable insights, there are several areas for improvement in future research. Overfitting was a persistent challenge, as models performed well on training data but struggled to generalise to unseen data. One way to address this would be by using Principal Component Analysis (PCA) to reduce feature complexity. Many of our explanatory variables, such as MPOWER policy scores, economic indicators, and geographic classifications, may contain overlapping information. PCA could help by transforming these correlated variables into a smaller set of independent components, allowing the model to capture the underlying patterns in the data while minimising the dominance of any single factor.

Additionally, collecting more detailed and diverse data could improve model performance and allow for a more comprehensive analysis. For example, public opinion surveys on smoking attitudes, enforcement data on tobacco control policies, or health outcome indicators such as lung cancer rates or tobacco related deaths could provide further context on how policies impact smoking behavior beyond prevalence rates alone. More granular economic data, such as disposable income levels, employment rates, and illicit tobacco market estimates, could also help refine our understanding of how financial factors influence smoking trends.

One of the main limitations of our study was the inability to conduct a cost-effectiveness analysis. While we had data on budget allocation for tobacco control and staffing levels in national tobacco control agencies, the data was incomplete, varied in currency, and lacked standardisation. Ideally we would have liked to examine whether countries that invested more resources into tobacco policy enforcement and public health campaigns saw greater reductions in smoking prevalence, allowing us to identify the most cost-effective policy strategies. Future research with more complete financial and resource allocation data could provide crucial insights into how to optimise spending for maximum impact.