-

DataScientest

QUIT: Quantifying 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

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

-

Data and Project Discovery

Sources

The data to be used is the WHO Global Tobacco Control Data (2000-2022).
From this data the following tables will be used:

- MPOWER Overview - Policy compliance scores
- National tobacco control programmes - Tobacco resource allocation figures
- Retail price + national tax - Prices and taxes on tobacco
- Non-age-standardized estimates of current tobacco use
- Age-standardised estimates of current tobacco use

Here is a summary in spreadsheet form of the different datasets to be used: Data Audit Tobacco

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-standardised tobacco 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

Important Note: The WHO tables contained the variable ‘countries, territories and areas’ initially. To simplify things we changed the name to ‘Region’ so when we refer to ‘Region’ in our study we are referring to a country, territory or area.

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 tobacco use 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 tobacco use prevalence.
   Effect/Bias: The effects of individual policies may be confounded, as simultaneous policy changes make it challenging to attribute changes in tobacco use 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.

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: Aiming to control for contextual factors.

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

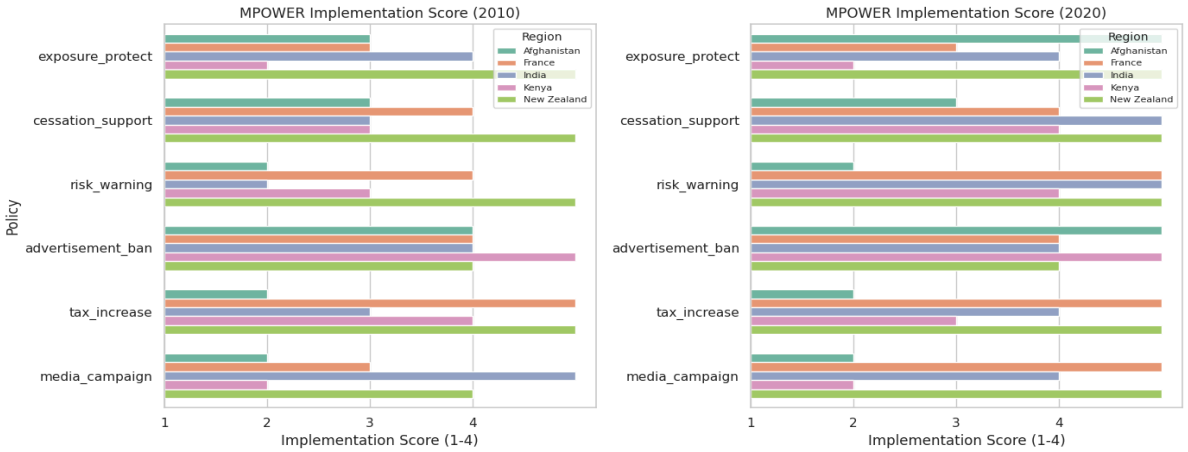
-

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

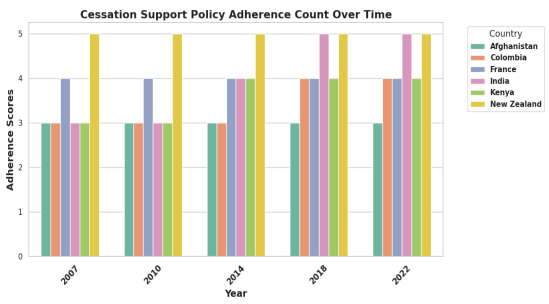
To systematically explore our dataset, we created separate visualizations for each key variable, allowing us to examine trends in policy implementation, tobacco use distribution, and pricing over time. Below, we present these visualizations along with key observations.

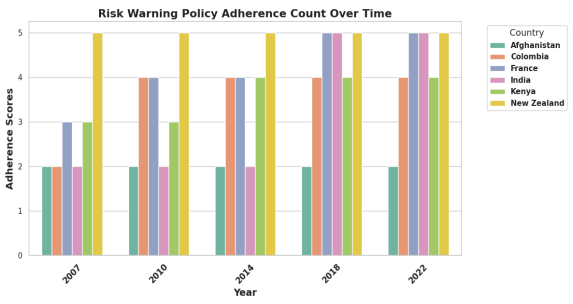
MPOWER Policies Implementation Score 2010 vs. 2022 across 5 countries



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.

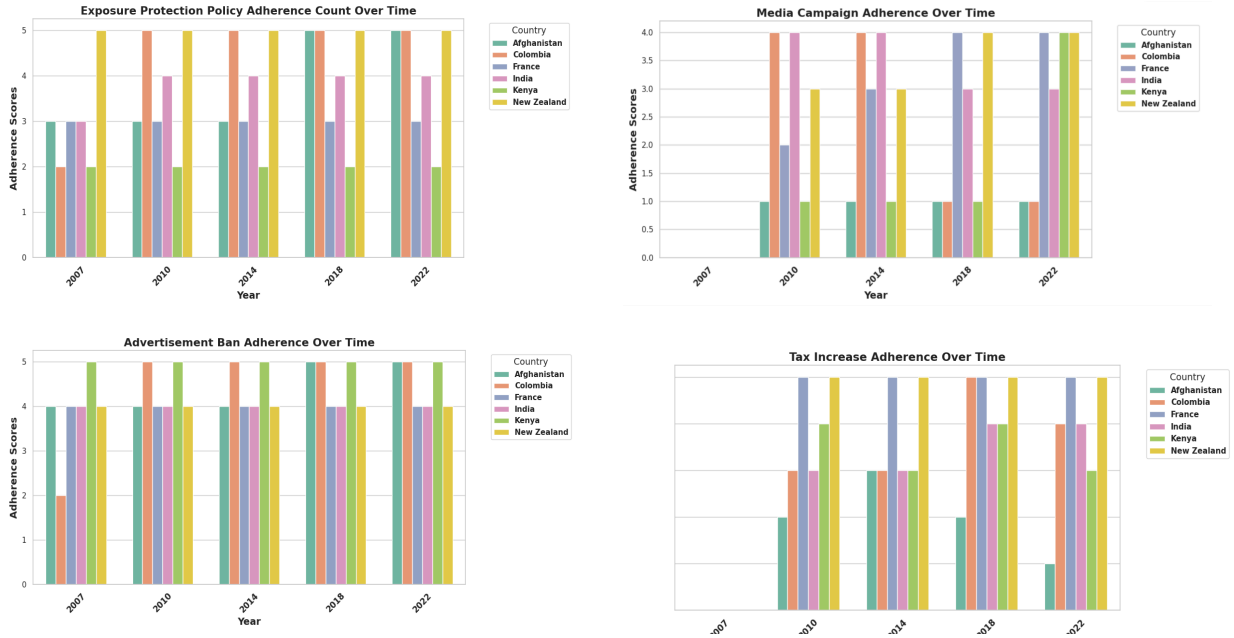
MPOWER Policies implementation 2007 until 2022 across 5 Countries





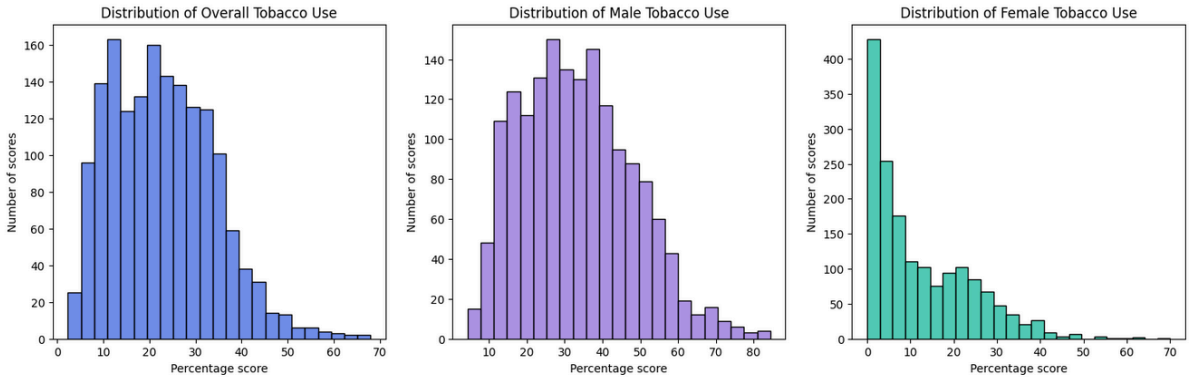
-

DataScientest



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

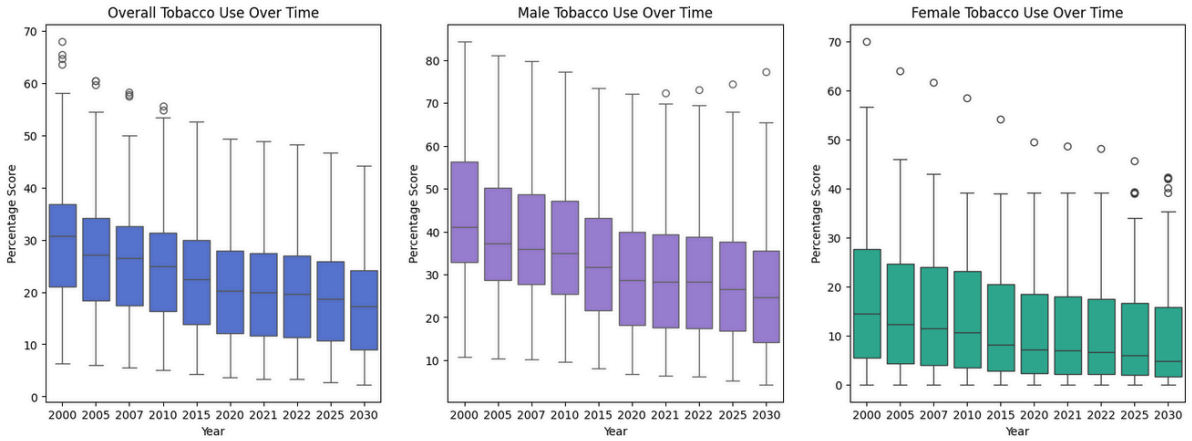


The above graphs show the distribution of the target variables (Standardised scores for overall, male, and female tobacco use). We can see that the majority of percentage scores for overall tobacco use lie between roughly 5 and 35%. When we look at the

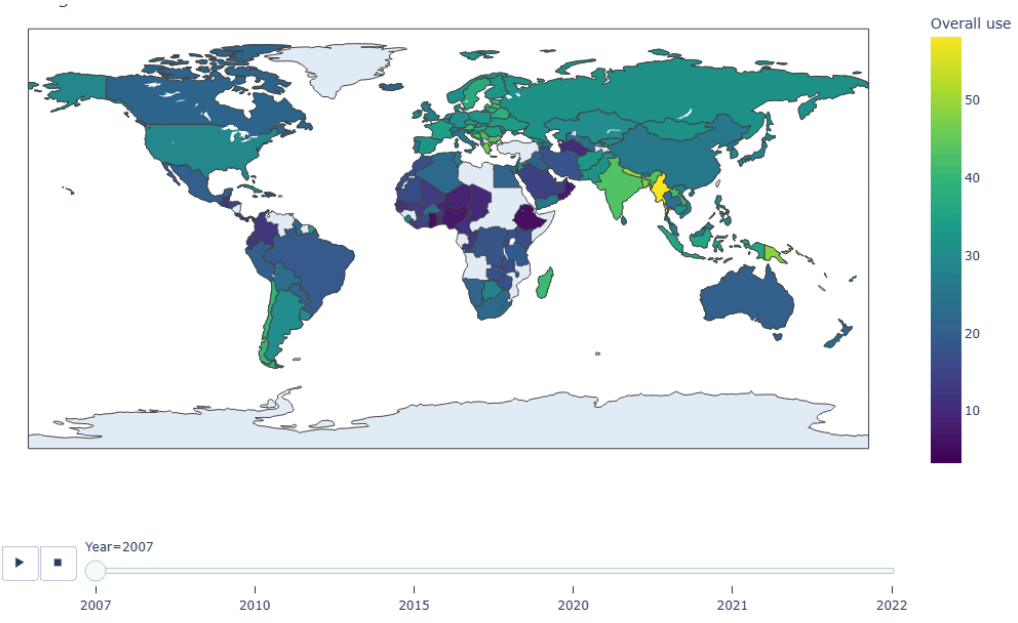
-

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

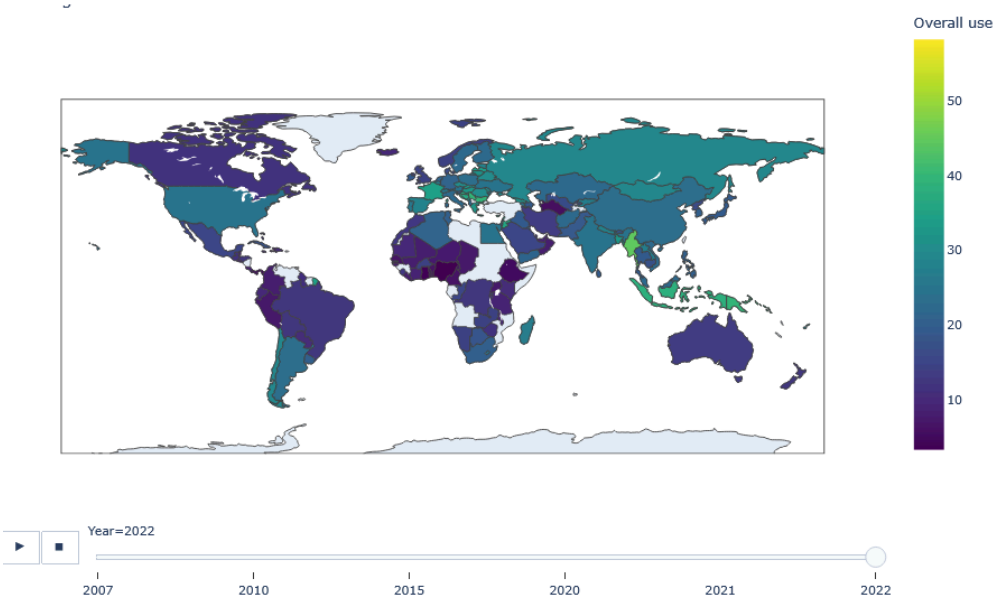


In the box plots above we can observe that for all groups (Overall population, Males, and Females) tobacco use decreased (and is predicted to continue decreasing until 2030) over time. Again, we can clearly see the trend that males tend to use tobacco more than females in general by looking at the median 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.

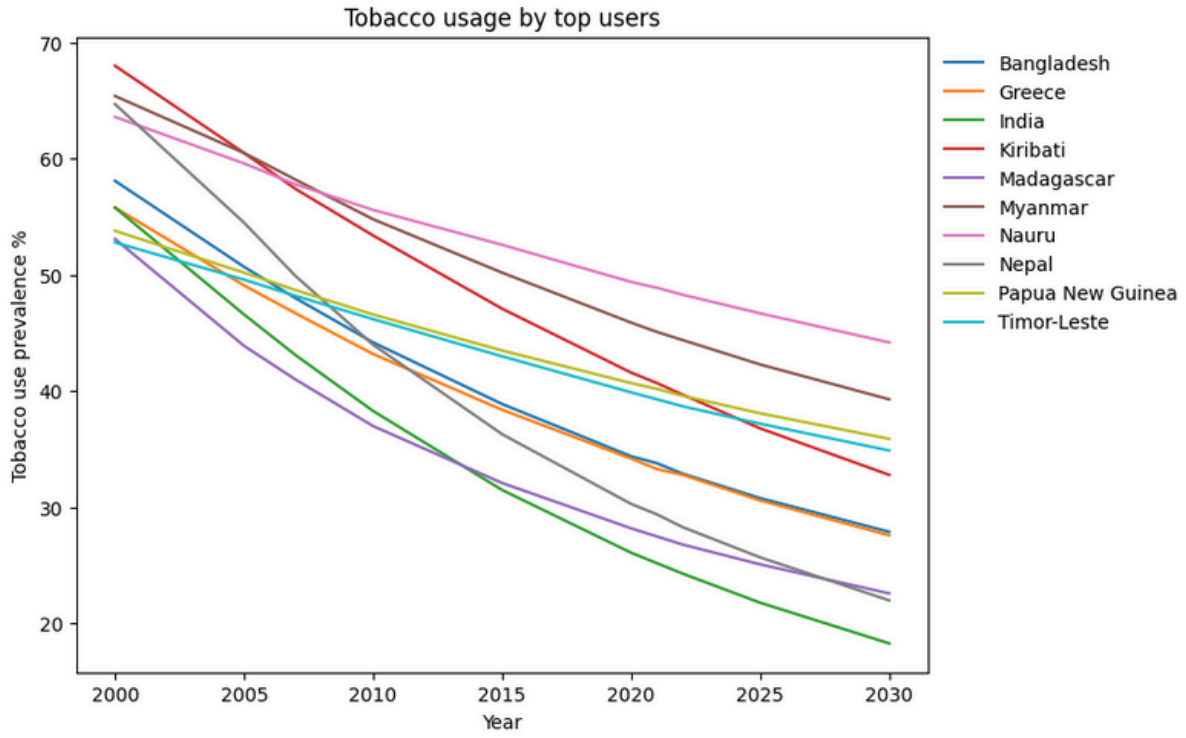


-

DataScientest



Above is a comparison from 2007 to 2022 of tobacco usage across the world, where once more we can see a general trend of decreasing tobacco usage globally.

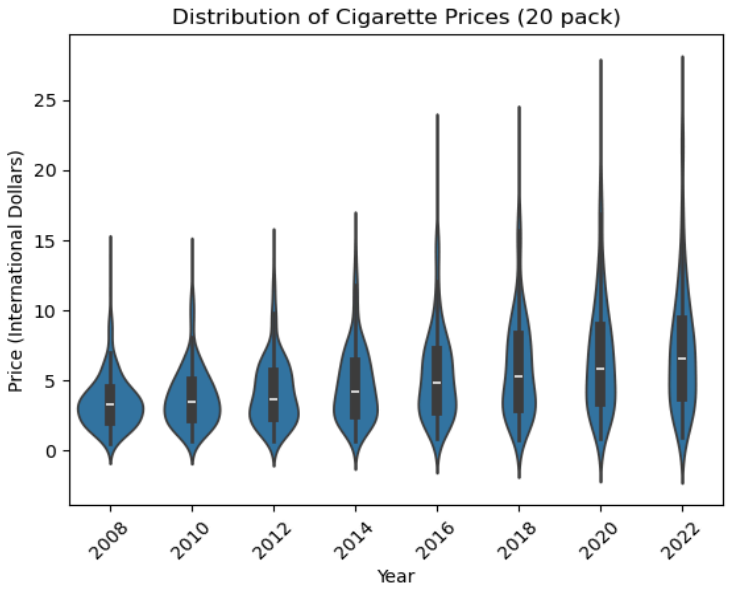


Even the countries with the heaviest tobacco use reduced their tobacco usage over time (and are predicted to continue to do so until 2030). The countries that showed the

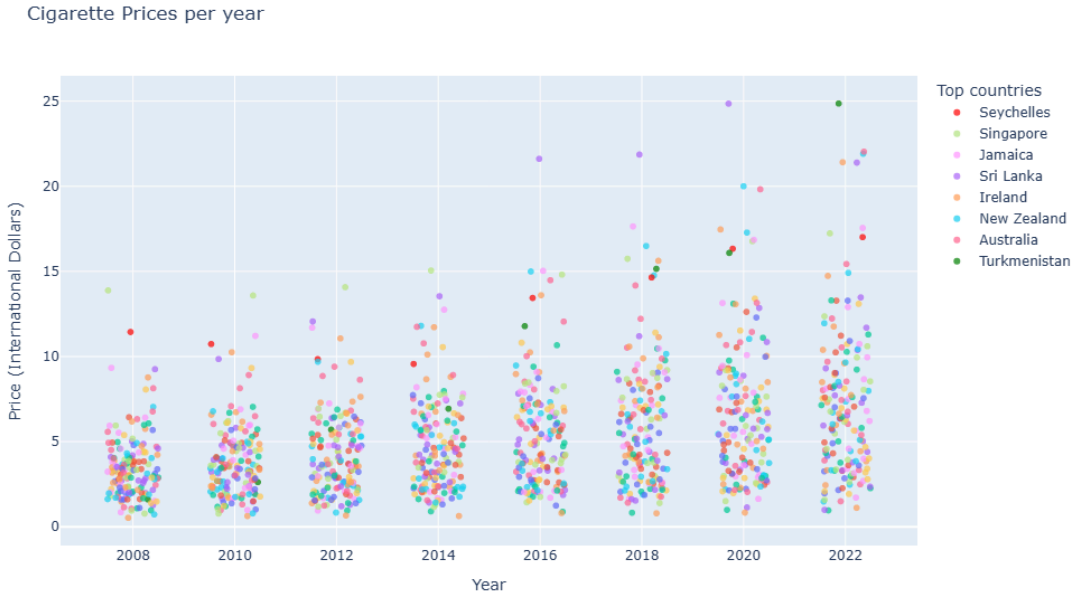
-

greatest decrease among the top users were Nepal (65 - 25%), India (55 - 20%), and Kiribati (67 - 35%).

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.



-

DataScientest

This visualization helps identify which countries have the highest cigarette prices, potentially allowing us to explore whether these regions 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

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.

2. **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 standardising 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.

3. **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.

4. **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 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.

2. **Enhancing Data with New Variables**
   - To explore the role of a country's Income Group and wider geographics (beyond country-variable:"region") on the effectiveness of policy implementation on tobacco use prevalence we introduced two additional variables "IncomeGroup" and "Continental Classification" from a dataset from the world bank (also mentioned at the 'data cleaning' step).

3. **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 2020 data showed an accuracy within 0.55%, 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:

-

DataScientest

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1296 entries, 0 to 1295
Data columns (total 22 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Region                                  1296 non-null   object
 1   Year                                    1296 non-null   int64
 2   Overall use                             1296 non-null   float64
 3   Male                                    1296 non-null   float64
 4   Female                                  1296 non-null   float64
 5   Non_age_standardised_tobacco_use        1296 non-null   float64
 6   Male(Non_age_standardised_tobacco_use)  1296 non-null   float64
 7   Female(Non_age_standardised_tobacco_use) 1296 non-null  float64
 8   Objectives exist                        1296 non-null   object
 9   Ntl agency exists                       1296 non-null   object
 10  No. staff                               1296 non-null   float64
 11  Cigarette_price                         1296 non-null   float64
 12  Monitor                                 1296 non-null   float64
 13  exposure_protect                        1296 non-null   float64
 14  cessation_support                       1296 non-null   float64
 15  risk_warning                            1296 non-null   float64
 16  advertisement_ban                       1296 non-null   float64
 17  tax_increase                            1296 non-null   float64
 18  media_campaign                          1296 non-null   float64
 19  Continental Classification              1296 non-null   object
 20  Income Group                            1296 non-null   object
 21  Total_compliance                        1296 non-null   float64
dtypes: float64(16), int64(1), object(5)
memory usage: 222.9+ KB
```

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

-

DataScientest

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	OneHot Encoding	6 unordered categories, OneHotEncoding was efficient
Cigarette Prices	num	Unique continuous values	Scaling	Standard Scaler	Values varied greatly and Scaling ensured compatibility with Linear Regression, which is sensitive to varying values.
MPOWER	num	0 - 4 Scale	Scaling	Standard Scaler	

3. **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.

4. **Implementation Feature Selection**

   We dropped the variables from the Tobacco Control table (No. staff, Objectives exist, Ntl agency exists) 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.

5. **Result**

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

Model	Train MAE	Test MAE
Linear Regression	1.02	1.09
DecisionTreeRegressor	0.00	1.61
RandomForestRegressor	0.24	1.22

   We chose **Mean Absolute Error** to evaluate model performance because it is calculated in the same scale as the target variable, providing more intuitive interpretability. Aside from gender-specific modelling, our models are trained on **Overall use** as the target variable as this data is most representative.

   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 generalisability, 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'.

6. **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 tobacco use 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 1.25, 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.

7. **Removing 'Region' (=Countries/Region) and Continental 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 both respectively 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 Sub-Saharan Africa (28%), Latin America & Caribbean (22%), and Middle East & North Africa (11%) showed that geography continued to dominate. Additionally, our Linear Regression showed much greater overfitting when continent was included, with an MAE of 4.94 on the training set and 7.23 on the test set.
     Removed Continental Classification: led to a sharp increase in MAE:

Model	Train MAE	Test MAE
Linear Regression	7.21	7.94
DecisionTreeRegressor	0.00	10.500
RandomForestRegressor	1.64	8.79

-

Feature importance without Region can be seen in the table below (under Overall Use Coefficient)

8. **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 tobacco use, 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 tobacco use, 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.

9. **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, in order to group countries with similar policy, economic and price environments. Continent was excluded because the One-Hot Encoding introduced many new columns, and K-means struggles with high dimensionality. The elbow method indicated three 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: 7.21, Test: 7.86), 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 tobacco use behavior, we introduced a time-lag to better capture delayed policy impact and improve model interpretability.
   - **Implementation:**
     A two-year time lag was applied by shifting the values for policy-related variables forward in time, ensuring that policies from year *t* were aligned with tobacco use prevalence in year *t*+2.
   - **Result:**
     Despite this adjustment, the model did not perform noticeably differently (Train MAE: 7.15, Test MAE: 7.83 on our linear regression model), once again supporting our conclusion that policy variables alone do not fully explain tobacco use prevalence. We were hesitant to increase the time lag as this would have resulted in less data (in an already limited data set) as whenever we moved the values forward in time it resulted in NaN's in the row where they previously were and we had no reasonable way of filling these in.

- **Model Tuning**

To improve model performance, we tuned the Random Forest Regressor using RandomSearchCV to optimize 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.

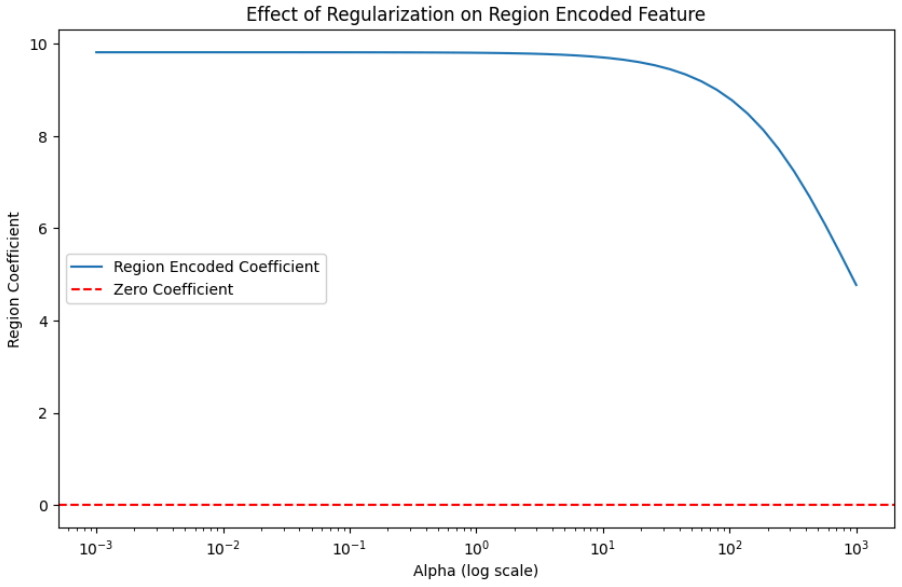
Model	Train MAE	Test MAE

-

RandomForestRegressor (default)	1.64	8.79
RandomForestRegressor (tuned)	3.78	8.65
XGBoost	5.48	8.07

3. 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.



4. 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 (countries labelled as 'HIC' or 'LIC' from our World Bank data) and carried out a feature importances analysis.
- **Result:**
  Model performance improved compared to previous region-removed models, with test 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 tobacco use 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.

5. Stratification by Continental Class

Furthermore, we decided to train individual models for each continental class to 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 stratification by geographic group can improve performance and reveal region-specific policy effects. The results were as follows:

Continental Classification	Test MAE	Greatest Coefficient	Second Great Coefficient	Third Greatest Coefficient
South Asia	6.20	Year (-5.561)	Advertisement Ban (2.157)	Risk Warning (1.992)
Europe & Central Asia	6.57	Cigarette price (-3.875)	Tax Increase (3.057)	Advertisement Ban (-1.739)
Middle East & North Africa	5.47	Risk Warning (-3.570)	Tax Increase (3.226)	Exposure Protect (3.085)
East Asia & Pacific	12.21	Risk Warning (-4.612)	Cessation Support (-3.296)	Advertisement Ban (2.859)

-

DataScientest

Americas	5.69	Tax Increase (3.509)	Exposure Protect (-2.142)	Risk Warning (1.948)
Sub-Saharan Africa	6.14	Cigarette price (3.000)	Risk Warning (-1.858)	Tax Increase (1.137)

Machine Learning Conclusions

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.

Next steps:

- Machine learning regression models have allowed us to determine the importance of our explanatory variables in predicting tobacco use, but do not provide evidence of the statistical significance of these relationships.
- We will utilize statistical models to see whether the observed coefficients are robust across different levels of the data (such as across income groups).

Statistical Modeling

After exploring predictive patterns through machine learning, we applied statistical models to better understand the individual effects of tobacco control policies and improve interpretability.

1. Fixed Effect Model

- **Rationale:**
  We applied a multiple linear regression model to estimate the effect of several policy variables on a continuous outcome—tobacco use prevalence. Given the panel structure of our dataset, which follows countries over time, we used a Fixed Effects (FE) model to control for unobserved, time-invariant country characteristics (e.g., culture, enforcement capacity) and global time trends (global shifts, including economic changes or international tobacco control momentum). This approach helps isolate the within-country effects of policy changes and accounts for the non-independence of repeated observations over time, improving the interpretability of causal relationships.

-

- **Implementation:**
  Multiple linear regression with fixed effects for both country (C(Region)) and year (C(Year)).
- **Result:**
  The model explained 98.2% of the variance in tobacco use ($R^2 = 0.982$), with an adjusted R^2 of 0.979. While this indicates a strong overall fit, the extremely high values suggest potential multicollinearity and risk of overfitting. 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 tobacco use prevalence. Advertisement bans showed a significant but unexpected positive effect, warranting further investigation. Other policies 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**
  To assess whether high correlations between policy variables distorted individual effect estimates, we checked for multicollinearity. This is particularly relevant in our model, as overlapping effects (e.g., between cigarette prices and tax increases) may inflate standard errors and mask true associations.
- **Results**
  Several variables showed high Variance Inflation Factors (VIF), especially *cessation support* (11.32) and *risk warnings* (10.38), indicating strong multicollinearity. Other variables such as *advertisement bans* and *tax increases* showed moderate VIFs (~8–9), while *media campaigns* and *cigarette prices* were within acceptable ranges (<5). These results confirm that multicollinearity likely contributed to the insignificance of otherwise impactful policies.

-

- **Next Step:**
  Since dropping variables would undermine the purpose of assessing individual policy effects, we explored adding interaction terms to better capture potential policy synergies.

3. Interaction Terms Analysis

- **Rationale**
  Policies may have synergistic effects—such as media campaigns reinforcing the impact of risk warnings. Including interaction terms can improve model fit and reduce omitted variable bias.
- **Implementation:**
  Interaction terms were selected using a two-step approach. First, we used a correlation matrix to identify moderately correlated policy variables ($0.3 < r < 0.8$). Then, we filtered these based on policy logic to ensure that only theoretically meaningful combinations were tested (e.g., risk warnings with media campaigns). Selected interaction terms were introduced into the Fixed Effects model two at a time to assess their individual impact without introducing excessive multicollinearity.
- **Result**
  Three interaction terms showed statistically significant or borderline effects when introduced two at a time into the model:

Interaction	Coefficient	p-value	Interpretation
Media Campaign × Risk Warning	-0.0758	0.026	Significant negative effect; suggests that media campaigns enhance the impact of risk warnings.
Price × Advertisement Ban	-0.0669	0.072	Borderline effect; potential policy synergy, though not robust.
Cessation Support × Risk Warning	0.0638	0.086	Borderline positive effect; may reflect increased quit attempts without necessarily reducing overall smoking prevalence.

When all three interaction terms were included in the same model, only **Media Campaign × Risk Warning** remained statistically significant (*coefficient* = –0.0727, $p = 0.031$). A general observation across all interaction models was that individual policy coefficients fluctuated notably, suggesting instability in the estimates and reinforcing the presence of multicollinearity.

4. Clustering Standard Errors

- **Rationale**
  Due to the instability of policy coefficients in the interaction models, we applied clustered standard errors to improve estimate reliability. While fixed effects control for time-invariant differences, clustering additionally accounts for within-country correlation and heteroskedasticity, resulting in more robust and conservative standard errors.producing more conservative standard errors and

-

reducing the risk of overstating policy effects and reducing the risk of overstating policy effects.

- **Implementation:**
  We re-estimated the Fixed Effects model using clustered standard errors at the country level, both with and without the interaction term Media Campaign × Risk Warning.
- **Result:**
  Clustering produced more conservative but statistically reliable estimates. Standard errors increased, leading to non-significant results for most policies. Risk Warning remained the only policy with a statistically significant effect (p = 0.034). The interaction term Media Campaign × Risk Warning lost robustness and was no longer significant (p = 0.105).

5. Stratifying by Income Groups

- **Rationale:**
  After applying clustered standard errors, policy effects became more conservative and often non-significant, though more statistically reliable. To explore whether policy effectiveness varies by economic context—and to identify effects that may be masked in the pooled model—we stratified the data by income group. This approach reduces within-group heterogeneity and allows for more precise estimation of group-specific policy impacts.
- **Implementation:**
  We ran separate Fixed Effects models with clustered standard errors at the country level for each of the four World Bank income groups: LIC, LMIC, UMIC, and HIC. No interaction terms were introduced.
- **Result:**
  Tobacco use declined across all income groups over the observation time, with the largest reductions in LICs and LMICs. Risk Warnings were significantly associated with reduced smoking only in LMICs. In HICs, Media Campaigns showed a borderline negative effect. No significant policy impacts were observed in UMICs or LICs, suggesting possible implementation challenges or limited policy reach in those settings.

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

-

6. Stratifying by Continental Classification

- **Rationale:**
  To explore region-specific policy effects, we stratified the data by continental classification. This reduces within-group variation and helps identify context-dependent impacts that may be obscured in the overall model.
- **Implementation:**
  We ran separate Fixed Effects models with clustered standard errors by country for each continental class (North America was merged with South America to Americas)
- **Result:**
  Tobacco use decreased across all continents over the observation period, with the most significant decline in Europe & Central Asia (–5.49, $p < 0.001$). Among specific policies, advertisement bans were statistically significant in Europe & Central Asia and the Middle East & North Africa, though both showed counterintuitive positive associations with tobacco use. In East Asia & Pacific, risk warnings were significantly associated with a reduction in tobacco use 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 tobacco use 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 tobacco use prevalence significantly in this region. Borderline increase in tobacco use (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 tobacco use (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.

Key Takeaways Stratification

- Risk Warnings show the strongest negative correlation with tobacco use prevalence, in LMIC and East Asia & Pacific .
- Advertisement Bans are positively correlated with tobacco use 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
  To address our research question, we examined whether tobacco control policies affect male and female smoking behavior differently, given known variations in social norms and economic factors. Identifying these differences can inform more targeted and effective policy design.
- **Implementation**
  1. Ran Fixed Effects models, including clustered standard errors at the regional level, separately for **male and female tobacco use prevalence,**both without and with the Media Campaign × Risk Warning interaction term.
  2. Stratified the analysis by **Income Group**.
  3. Stratified the analysis by **Continental Classification**.

- **Results and Interpretation**
  1. Without interaction term: For **men**, **risk warnings** significantly but minor reduce tobacco use prevalence (coef=-0.4738, p = 0.015), while all other policies 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 tobacco use behavior more strongly.
     With interaction term: Introducing **Media Campaign × Risk Warning** reduced the direct effect of risk warnings on **male tobacco use** (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 tobacco use**, no policies were significant before or after adding the interaction (p = 0.368), reinforcing that female tobacco use behavior is less responsive to these policies.
  2. Stratified by Income Group: Risk warnings were significantly effective in **LMICs** for both men and women. Cigarette prices significantly reduced tobacco use among **women in HICs**, indicating higher price sensitivity in this group. No significant policy effects were observed in **UMICs** or **LICs**. Overall, female tobacco use appeared less responsive to policy measures consistent with non-stratified findings.

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 tobacco use for both genders.
LIC	None	None	No significant effects observed for either gender.

3. Stratified by Continential Class: Tobacco use declined most among men in **Europe & Central Asia** and least among women in **Middle East & North Africa**. **Risk warnings** were the most consistently effective policy in **East Asia & Pacific** for both genders, while effects in other regions were weaker or absent. **Advertisement bans** showed counterintuitive positive associations in several regions—especially for women in **Europe & Central Asia** and **Sub-Saharan Africa**, and for men in **Middle East & North Africa**. Female tobacco use was generally less responsive to policies, with some unexpected increases observed in the **Americas** for **cessation support** and **risk warnings**. After stratifying by continent, the emergence of counterintuitive policy effects could be due to smaller sample sizes, reduced variation within groups, and region-specific dynamics—such as policy timing and implementation quality—that become more influential when broader controls are removed.

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 tobacco use from media campaigns. Counterintuitive increase in female tobacco use from ad bans.
Middle East & North Africa	Advertisement Ban: +1.41 (P = 0.004)	None	Counterintuitive increase in male tobacco use from ad bans.
Americas	None	Cessation Support: +0.51 (P = 0.049) Risk Warning: +0.32 (P = 0.011)	Unexpected increase in female tobacco use 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 tobacco use 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 tobacco use from ad bans. Borderline decrease from exposure protection.

-

DataScientest

Overview of Policy Impact Statistical Models

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

Statistical Modelling Conclusions

Key takeaways:

- **Risk Warnings** were the most consistently effective policy across all models—particularly in LMICs and East Asia & Pacific—likely due to high implementation fidelity and lower baseline awareness.
- **Advertisement Bans** frequently showed counterintuitive positive associations with tobacco use, especially in gender- and region-stratified models. These effects are unlikely due to reverse causality (controlled for baseline prevalence) and may reflect industry adaptation, policy loopholes, or reactive implementation timing.
- **Cigarette Prices** were only significantly effective for women in HICs, suggesting gendered differences in price sensitivity. Their limited impact in LICs/LMICs may be due to low price elasticity, illicit market access, and weak tax enforcement.
- **Media Campaigns** showed limited effectiveness overall, with borderline effects observed only in HICs and men in Europe & Central Asia, potentially reflecting differences in media reach or campaign quality.
- **Cessation Support** showed unexpected positive associations in some cases (e.g., women in the Americas), potentially indicating increased quit attempts without long-term cessation success or substitution effects.

-

- **Exposure Protection** showed a borderline effect in the full model, with a slightly stronger impact among females in Sub-Saharan Africa, possibly due to gendered exposure patterns in public settings.
- **Tax Increases** were not statistically significant in any model. Their lack of effect may reflect delayed behavior change, compensatory strategies, or weak enforcement.
- Interaction Terms revealed a potential synergy between **Media Campaigns and Risk Warnings,** which remained significant when tested in isolation (p = 0.026) but lost robustness in the full model (p = 0.105). This highlights both the potential for policy interaction and the instability caused by multicollinearity in small samples.
- **Clustering** standard errors improved statistical reliability but made many policy effects non-significant, confirming that some previously observed effects may have been overstated.
- **Stratified models** (by income, continent) helped uncover context-specific effects, but also introduced instability and counterintuitive findings, likely due to reduced sample size, limited within-group variation, and unobserved regional dynamics (e.g., timing, enforcement).

Conclusion

This project explored global tobacco use trends and the impact of control policies through both machine learning and statistical modeling approaches. Our findings reinforce well-documented patterns—such as higher male smoking prevalence and a general decline in tobacco use over time—while also uncovering nuanced, region-specific policy effects.

Based on our machine learning models, country-specific factors were the strongest predictors of tobacco use, overshadowing the direct influence of policy measures and cigarette prices. Simple linear regression outperformed more complex models, highlighting the challenges of capturing tobacco consumption trends through non-linear relationships and the risk of overfitting. While our efforts to optimise predictive performance across the entire dataset were largely unsuccessful, stratification by income group and continental classification proved valuable, improving model accuracy and revealing variations in feature importance across different economic and country groups.

Key insights suggest that tax increases often coincide with rising tobacco use in some regions, indicating they may be implemented reactively rather than preventively. Cigarette prices had strong negative effects in high-income countries, but showed unexpected positive associations in low-income countries, likely due to affordability dynamics and informal markets. Risk warnings consistently emerged as one of the most effective policies, while advertising bans showed unintended positive correlations with tobacco use in some continental areas, possibly due to enforcement challenges or

-

industry adaptation. These trends were broadly consistent across both machine learning and statistical models.

Statistical modeling helped explain these patterns more clearly by focusing on interpretability and causal structure. Fixed effects were used to control for differences between countries and over time—factors that strongly influence tobacco use—while clustering improved the reliability of standard error estimates. However, high multicollinearity between policy variables and model instability in smaller subgroups limited the strength of individual policy conclusions. Testing interaction terms revealed a potential synergy between media campaigns and risk warnings, though this effect was not robust across models. Stratification proved most valuable, uncovering how policy impacts vary by income level, region, and gender, and reinforcing the need for targeted, context-sensitive strategies.

These findings reaffirmed the machine learning results while adding clearer interpretability and causal insights. Risk warnings showed the strongest and most consistent negative association with tobacco use, particularly in LMICs and East Asia & Pacific, where baseline awareness may be lower and implementation tends to be stronger. Advertising bans, on the other hand, exhibited counterintuitive positive associations in some regions, possibly due to industry adaptation, policy loopholes, or reactive implementation. Cigarette prices and tax increases had limited effects outside of high-income settings, suggesting that price elasticity, illicit markets, and enforcement challenges weaken their impact in lower-income countries. Media campaigns and exposure protection policies had mixed effects, with some gender- and region-specific variations, while cessation support showed unexpected positive associations in certain cases, potentially reflecting increased quit attempts without long-term success. Stratified models also revealed inconsistencies in some subgroups, likely due to limited sample sizes and unobserved confounding.

Policy Implications

Our findings emphasize that tobacco control is not a one-size-fits-all solution, but requires context-sensitive, coordinated strategies to be effective across diverse populations and settings.

1. **Implementation quality and reach matter.**
   Risk warnings consistently stood out as one of the most effective and reliable measures, particularly in LMICs and East Asia & Pacific. Their simplicity, low cost, and direct targeting of smokers highlight the importance of policy fidelity, coverage, and behavioral relevance.

2. **Policies interact and should be designed as integrated packages**
   Our analysis shows that it is difficult to isolate the effect of individual measures, as tobacco control policies often operate in synergy. They should therefore be

-

understood and implemented as part of a comprehensive policy mix, rather than as isolated interventions.

3. **Context matters**
   The limited impact of cigarette pricing in LICs, compared to its stronger effects in HICs, underscores the need for context-specific policy design that accounts for economic conditions, market structures, and enforcement capacity.

4. **Targeting matters**
   Gender-specific effects—such as the impact of pricing and exposure protection—show that tobacco control policies must be responsive to social norms and differential behavioral patterns. Interventions should be designed with **equity and inclusivity** in mind to reach underserved or less responsive groups.

Ultimately, while global tobacco use is declining, our findings highlight the complexity of policy effects and the importance of aligning interventions with contextual realities and population needs. Moving forward, future research should explore policy interactions and long-term behavioral dynamics to strengthen the global response to tobacco use.

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 increasing the size of the dataset. Expanding the dataset to include more countries and historical data from earlier years could provide a broader perspective on tobacco use trends and policy impacts. A larger dataset may also help mitigate overfitting by allowing the model to learn more generalisable patterns rather than being overly influenced by a limited set of observations.

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 tobacco use 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 tobacco use 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 tobacco use 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.