Group 2 Data Project

INTRODUCTION

During the initiation stage of our project, the UK Parliament was in the process of approving a smoking ban, which from 2027 would ban cigarette sales to anyone born after January 2009. The aim of our project was originally to examine the evidence that the proposed bill being put before the UK Parliament will be effective in reducing smoking rates in the UK by analysing historic data on the effects of government mitigations.

The Bill was set to prohibit the sale of cigarettes to people born after 2008. Each year the age at which people can legally buy cigarettes would increase, resulting in people born after 31st December 2008 never being able to buy them. The government believed that this would reduce the harm caused by smoking, which is still the number one cause of preventable deaths in the UK and costs the NHS £17bn a year.

However, the Prime Minister called a general election on 22nd May to be held on the 4th July. The proposed bill did not come into law before this happened. The focus of our project has therefore changed slightly to examine the evidence that any future prevention strategies implemented by the next government will be effective in further reducing smoking rates.

Aims & Objectives

How effective does historic data suggest that government mitigations have been on affecting smoking prevalence? Based on this analysis, how effective will the new ban be on reducing smoking prevalence?

Our goal is to present data analysis and visualisations to central government officials involved in designing future prevention strategies. By examining historical data, we aim to predict the potential effectiveness of prevention measures in reducing smoking rates.

Roadmap of the project

- **Stage 1:** Scope project ideas and formulate purposeful questions.
- **Stage 2:** Collect the raw datasets.
- **Stage 3:** Clean the datasets.
- **Stage 4:** Evaluate and visualise our datasets. At this stage we had identified it was best to analyse our data by different factors affecting/potentially affecting the smoking rate:
    - 1. Analyse effectiveness of government measures on reducing global smoking rates, using the WHO API MPOWER framework.
    - 2. Financial factors, looking at tobacco affordability, household expenditure and increased tax rates against the smoking rate.
    - 3. Increasing e-cigarette use, particularly in the younger age group.
    - 4. Social/educational factors, analysing smoking rates against education level.
- **Stage 5:** In-depth analysis. Looking further in depth and expanding our visualisations and analysis for each factor. Implementing machine learning to predict future smoking rates based on our data.
- **Stage 6:** Present findings and conclusions from holistic analysis and visualisation of our data.

BACKGROUND

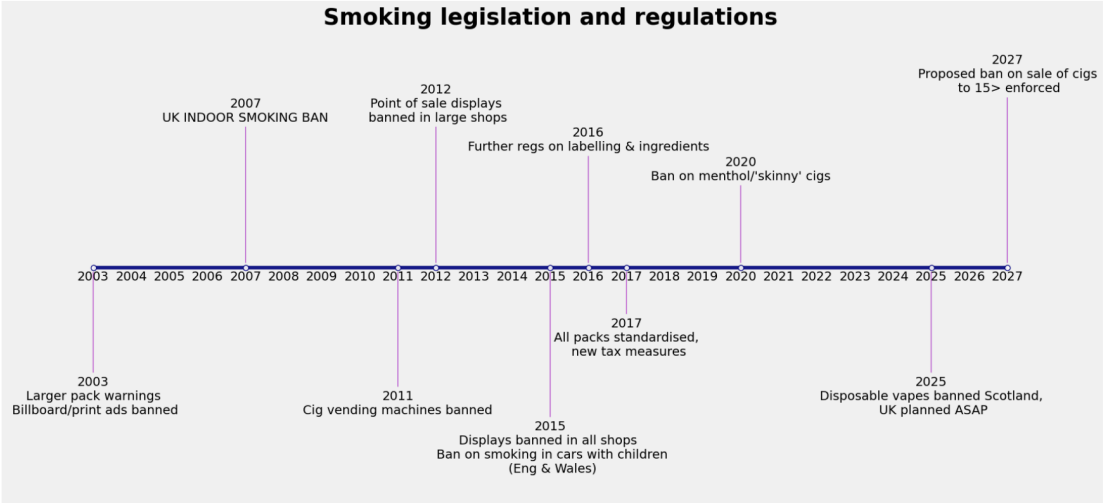


Fig 1 - Timeline of key government smoking regulations and legislation.

In the UK, smoking has claimed the lives of millions due to the presence of nearly 70 carcinogenic chemicals in each cigarette (Djordjevic, Stellman, and Zang, 2000). Tobacco-related diseases continue to impose a significant burden on the NHS and other public entities due to their debilitating effects. In 2023, England spent £17 billion on tobacco-related expenses alone. This includes £6.6 billion on smoking-related unemployment, £6.1 billion on smoking-related loss of earnings, and £1.3 billion on smoking-related early deaths. According to NHS Digital, in 2017 alone, there were 489,300 hospital admissions attributed to smoking and 77,800 smoking-related deaths (NHS Digital, 2019).

The UK has made significant strides to reduce the tobacco mortality rate by introducing legislation to limit advertisement, outlawing indoor smoking in public areas, banning smoking in cars with children present, and many other measures. Our project aims to assess the success of these legislative efforts in reducing the prevalence of smoking in the population.

To support existing smokers in quitting, the government is more than doubling the budget for stop smoking services, investing an additional £70 million per year aimed at supporting around 360,000 people to quit each year (Department of Health and Social Care, 2023). Further funding includes £5 million in 2023 and £15 million in 2024 for new national campaigns to explain the legal changes, the benefits of quitting, and the support available (Department of Health and Social Care, 2023). Additionally, £30 million a year will be allocated to enforcement agencies to ensure new regulations around tobacco use and sale are enforced.

While the dangers of smoking have become more well known in recent decades, there has been a long history of government interventions taken to curb smoking. The first significant measure came with the 1964 Television Act, which banned cigarette advertising on television. This was followed by the introduction of health warning labels on cigarette packs in 1971 (Department of Health and Social Care, 2023). Further restrictions on the visibility and advertisement of cigarettes came in 2003 with the Tobacco Advertising and Promotion Act, which banned most forms of tobacco advertising, sponsorship, and promotion (Department of Health and Social Care, 2023). The introduction of plain packaging regulations in 2012 required standardised packaging for all tobacco products.

Various smoking bans have been implemented over the past 50 years, including the ban on smoking in the London Underground in 1986, all NHS premises in 1991, and the controversial ban on smoking in all public spaces and workplaces in 2007 (Department of Health and Social Care, 2023). This was further extended to private vehicles carrying children under 18 in 2015, following extensive research on the deadly consequences of second-hand cigarette smoke on the health and development of young children.

SPECIFICATIONS AND DESIGN

Technical requirements

The primary technical requirements for the project were an ability to work with and effectively utilise Jupyter notebook using a Python kernel. As we were importing and analysing extant datasets, we needed the technical ability to interact with and effectively utilise these datasets - using Python to export data from APIs and CSV files and render them into a format in which we could perform analyses. We also needed to install and import relevant Python libraries to assist in data analysis and visualisation. We needed the means to share information and ideas, so an understanding of Github, Google Colab and Miro was required.

General requirements

Whilst we all had foundational knowledge of the technical requirements due to our work throughout the CFG degree, there were also a number of soft skills that allowed us to effectively complete the project. We needed to ensure that we were able to effectively communicate and collaborate with each other, understanding our roles in the project and being open to everyone’s ideas to push the project into new directions.

We needed to have good organisational and time management skills, able to run through the various stages of the project systematically so that each stage led onto the next.In addition to this, we needed to understand how to use the tools that were at our disposal in order to best analyse the data. We needed a degree of pattern recognition and the ability to conduct meaningful statistical analysis in a way that best utilises the data that we had at our disposal.

Design and architecture

Our main project file was a Jupyter notebook, hence we used Python to code and markdown to share our findings and analyses. We used:

- Jupyter notebook using Python kernel for general coding
- Python libraries for analysis and visualisation
- Github and Miro for collaboration

IMPLEMENTATION AND EXECUTION

We approached the project with the view that each team member would have a role throughout each stage of the process. We would schedule weekly meetings where we discussed the stages we were at and what the next tasks would be. We all collaborated on cleaning data and creating data visualisations, and made adjustments based on group feedback where required.

In addition to this, some team members took on additional logistical tasks in order to ensure the project was able to run smoothly. Bonnie focussed on tracking the project through Miro task boards, the group activity log and project documentation. Rhona managed all of the GitHub files, assembling all of the raw datasets in a GitHub repository and consolidating our findings into a single Jupyter Notebook, and Funmi created the presentation outline deck and PDF submission template so that we could all view and contribute information.

We also took advantage of our unique backgrounds and experiences when assigning tasks. Norah's Science background meant she was ideal for the more in-depth background research into smoking rates, while Rhona's knowledge of taxation and statistics allowed her to provide discipline-specific knowledge around the aforementioned areas. Funmi's background as an account executive meant she was able to help with setting up the project PDF and provide valuable feedback on the presentation deck. Katy's background as archaeological supervisor meant she had a lot of experience managing large scale projects, creative problem solving and pattern recognition. Katy was able to provide great insights into the data we were analysing - especially the socioeconomic background. Bonnie's experience as an intelligence analyst helped with drawing out succinct insights to communicate results.

We began by independently researching project ideas, and quickly narrowing it down first to the theme of healthcare due to three of our team members having a background in that sector. We settled on the more specific theme of smoking primarily based on the existence of multiple good datasets. We brainstormed question ideas to further narrow our focus within the theme of smoking. We gravitated towards the more social aspects rather than the effects of smoking on health as we felt that the dangers of smoking had already been comprehensively explored.

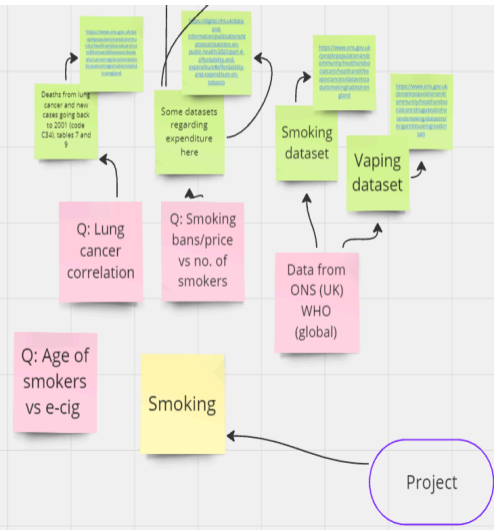


Fig 2 - Mind map Miro board to scope ideas for the project.

We used Miro to brainstorm ideas and shared datasets we thought were viable for the project. We used a Miro board to weigh the pros and cons of the datasets we discovered to find the most applicable ones to our project. Once we narrowed down our datasets, we each took 1-2 datasets to clean and uploaded our code and the cleaned datasets on ro Github. At this stage, we independently performed some rudimentary analysis and visualisations on our datasets. We used a Kanban board to track our task progress and what each group member was assigned.

We then met to discuss our findings, the qualities of our datasets and how we could implement some more in-depth analysis. It was at this point that we were able to further refine our datasets based on limitations we found while cleaning. We kept an Agile development mindset throughout, completing our project in iterative stages and making flexible adjustments to our direction throughout to account for our findings and constraints to produce the best outputs.

Dataset no	Dataset	Subset no	Subsets	Dates	Constraints/notes
1	ONS - Adult smoking habits in Great Britain	1.1	• Proportion of smokers, those quite smoking, and those never smoked by sex and age	1974 - 2022	• General note about covid impact on extent of data collected, but not specific • ONS data generally much better documented with data quality notes than NH
		1.2	• Time to first smoke of the day after waking	2015 - 2022	
		1.3	• Proportion of smokers intending to quit	2015 - 2022	
		1.4	• Intention to quit vs time of first smoke of day	2015 - 2022	
2	NHS Statistics on Smoking 2020	2.1	• Total NHS hospital admissions with primary diagnosis caused by/estimated attributable to smoking, for people >35	2009 - 2020 (looks like financial year 2020, noted as '2019/20')	• Divided into general categories of cancers, respiratory diseases, circulatory d by smoking
		2.2	• Total NHS hospital admissions with primary diagnosis caused by/estimated attributable to smoking, for people >35 *by gender*	2018 - 2020 (looks like financial year 2020, noted as '2019/20')	• Limited data available for admissions/deaths by gender for year, but data is r not just categories
		2.3	• Total registered deaths from diseases caused by/estimated attributable to smoking for people >35	2009 - 2019	• Discrepancies between hospital admissions and deaths in date recording - wi
		2.4	• Total registered deaths from diseases caused by/estimated attributable to smoking for people >35 *by gender*	2018-2019	• Limited data for deaths by gender
		2.5	• Prescription items and net ingredient cost of prescriptions in primary care to help people quit smoking, by type received	2009-2020 (financial years)	

Fig 3 - Dataset Miro board to compile and assess data sources..

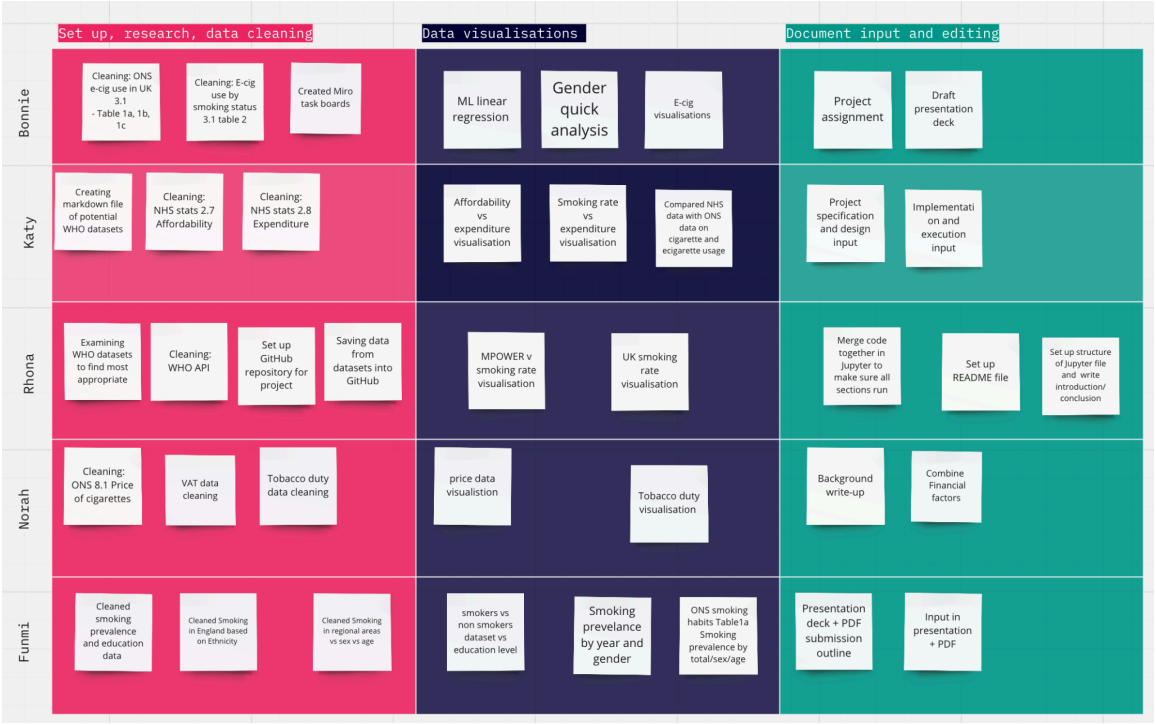


Fig 4 - Miro Kanban board for task tracking and assignment.

Implementation challenges

- Communication and task allocation: varying levels of free time and obligations and work schedules between team members required robust communication to coordinate and make sure tasks were equitable to available time for team members.
- Interacting with the WHO API: This API has a huge amount of information, Rhona had the challenge of thoroughly assessing the dataset to draw out data relevant to our question.
- Limited datasets: Funmi discovered that one of her ONS datasets regarding the prevalence of smokers in different areas of the UK had updated its survey methods which resulted in an unusual drop in the percentage of ex-smokers.
- Some of our data is for the UK and some is just for England so some differences when comparing
- Date range of the data: The years data was collected varied between our datasets - the team came to a decision to restrict all data to after 2000 for sake of comparison. Lack of recency with e-cigarette data limited the scope of our conclusions in this area.
- As the data is collected from surveys, there are some issues with continuity between the years due to a change in survey questions, but most of such cases were acceptable for continuity.
- We largely focussed on change in data over time, so we are limited in the type of visualisations that we can use to effectively display the data, with line and bar plots being the most appropriate. When analysing different factors at a single point in time such as demographics, we had more options for visualisations.
- Our data is aggregated data, not ‘raw’ data entries from the surveys themselves, hence limiting the type and extent of some analysis we can do. Due to the aggregation/time series data, we are mostly looking at correlation between variables and distribution. Statistical analysis is not as relevant as we expect a range/skewed distribution over time.
- Conclusions are often difficult between correlation/causation as there are many affecting factors to consider.

Tools & Libraries

- Jupyter notebook - used to compile all visualisations.
- Miro - track tasks and deadlines.
- Google Collab - used to easily work on the main jupyter notebook.
- Visual studio code - used to write relevant code.
- GitHub - used to create a repository where all relevant code was stored for easy access for the entire team.
- Python libraries: Pandas, Numpy, Matplotlib, Seaborn, Scikit Learn

DATA COLLECTION

We required data on smoking prevalence in the UK and other countries, including e-cigarette usage in the UK and

smoking prevalence according to the highest level of education. Additionally, we sought information on affordability, household expenditure, and the correlation between cigarette prices and smoking prevalence. We found an abundance of publicly available data sources that were well suited to our research question and provided good analytical opportunities. We collected the data via either an open API for the WHO, or downloads of XLS files that we converted to CSV for the other sources. We also took a small amount of data from written documents (all that was available) of the ASH Smokefree GB youth survey 2023 to mitigate against recency constraints in youth e-cigarette data.

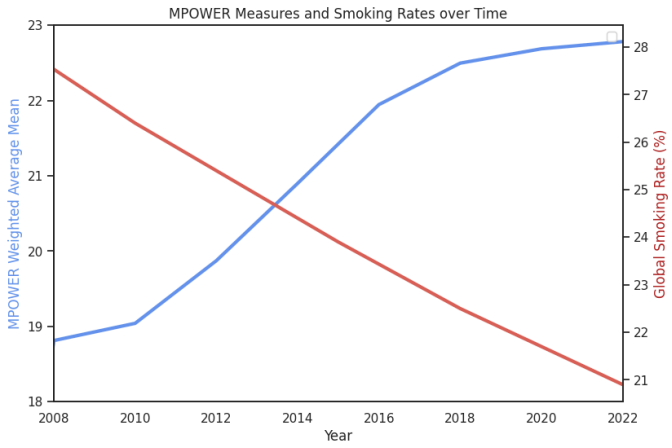
Data sources used:
WHO GHO OData API. The WHO adopted the WHO Framework Convention on Tobacco Control on 21 May 2003 as a response to the global tobacco epidemic. It uses the MPOWER package to help countries reduce their smoking rates by reducing the demand for tobacco products from their populations. It has six different strands:

- **M**: Monitor tobacco use and prevention policies
- **P**: Protect people from tobacco smoke
- **O**: Offer help to quit tobacco use
- **W**: Warn about the dangers of tobacco
- **E**: Enforce bans on tobacco advertising, promotion and sponsorship
- **R**: Raise taxes on tobacco

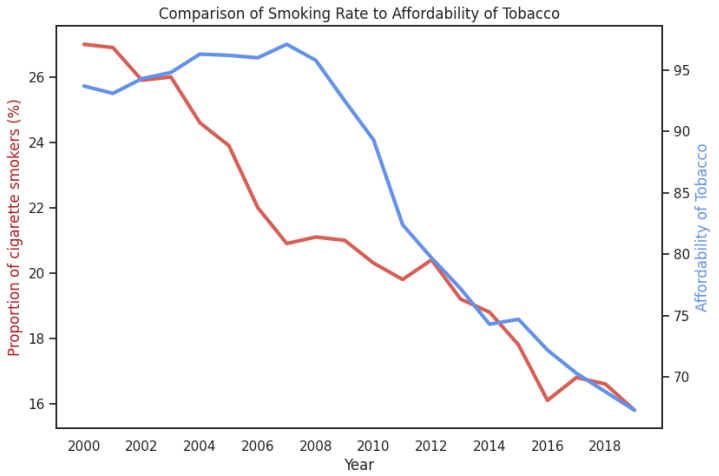
ONS: Adult smoking habits in Great Britain
NHS: Tobacco affordability and expenditure
ONS: E-cigarette use in Great Britain
ONS: Smoking habits in the UK and its constituent countries
Historical Tobacco Duty rates - GOV.UK
Institute for Fiscal Studies - historic VAT rates
ASH Smokefree GB adults and youth survey

CONCLUSION

We set out to analyse the effect that government interventions have historically had on reducing smoking prevalence, both worldwide and more specifically in the UK. Using a variety of data sources we were able to comprehensively analyse a variety of factors that contributed to a reduction in the smoking rate, as well as external factors that must be taken into consideration when considering future government interventions.

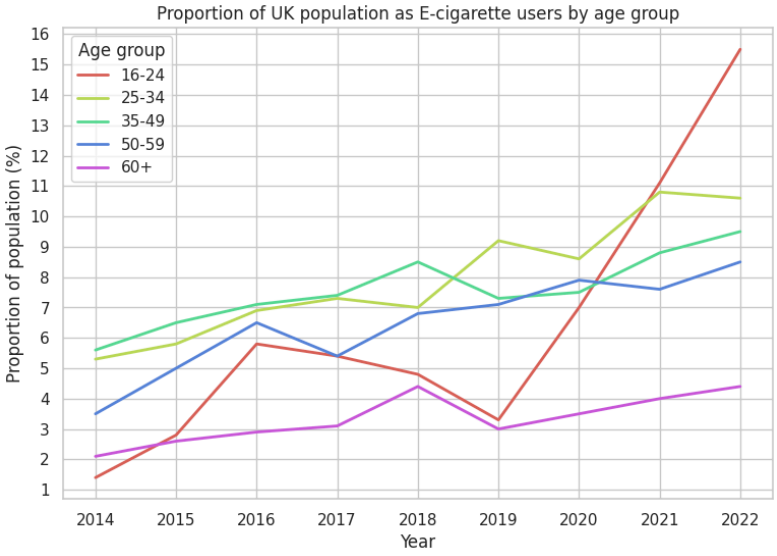


Firstly, we determined that global smoking rates reduced from 28.1% in 2007 to 20.9% in 2022 and that this strongly correlated to government intervention measures increasing over the same period. When we focused our analysis on the UK, we saw a similar reduction in smoking rates from 27.0% in 2000 to 15.6% in 2019. This can be seen to link with government intervention measures around packaging and advertising becoming more stringent.



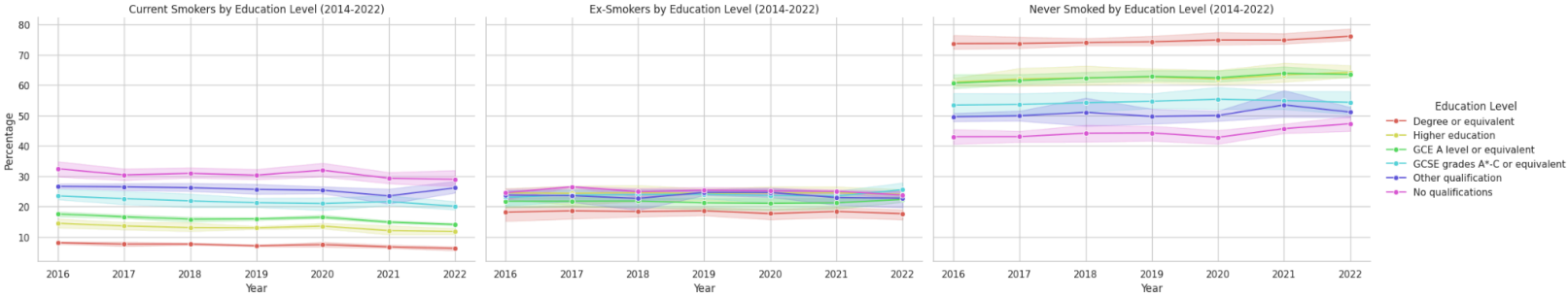
We then investigated the effect of financial mitigations on smoking rates in the UK. We found that increased taxation on tobacco has significantly increased the price of cigarettes, leading to a reduction in the affordability of tobacco. This strongly correlates with the decrease in smoking rates and from this we can infer that decreasing the affordability of tobacco by increasing taxes may be an effective method of reducing the overall smoking rate. However, we also found that the correlation is not as strong between reduced affordability and a reduced expenditure on cigarettes. This suggests that increased taxation is potentially less effective in reducing the tobacco consumption of current smokers.

We then looked at the impact of e-cigarettes on changes in the overall tobacco smoking rate as anecdotal evidence and our own personal experiences suggest that there has been a significant increase in e-cigarette use in the past few years. Our analysis shows that as e-cigarette use has increased, the smoking rate has decreased; however, the spike in e-cigarette use is disproportionately due to the younger generation. As such, the impact of e-cigarettes on overall smoking rates is harder to determine.



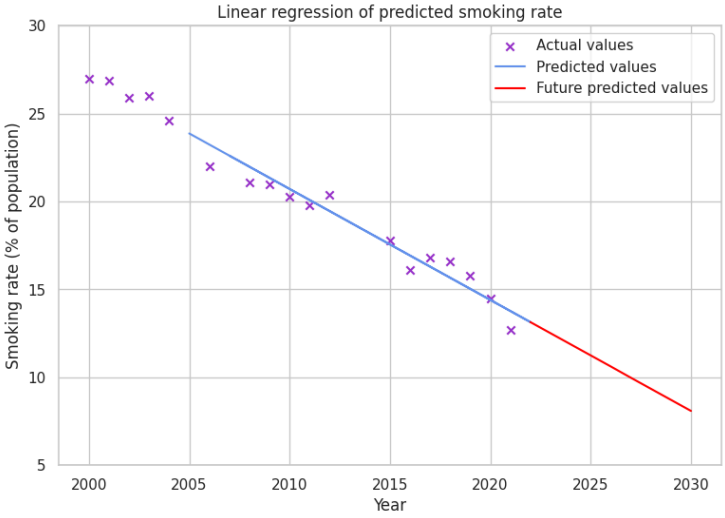
Finally we examined the available data to see if other socio-economic factors could be influencing the declining smoking rate, outside of direct government control. We could see that there was a link between education level and smoking rates; the higher the education level someone has the less likely they are to smoke. This highlights the importance of education as a determinant of health behaviours and suggests that public health interventions aiming to reduce smoking prevalence may benefit from focusing on educational initiatives and targeted support for lower education groups.

We conducted a similar analysis for differences between smoking rates in men and women, but we could not see any similar trends. As such we concluded that any future government interventions should equally consider both men and women when their implementation is planned.



The evidence therefore suggests that smoking rates will continue to fall if the future government, elected in the July 2024 UK general election, continues to tighten legislation and increase tax in a similar way to previous ones. Socio-economic trends influencing e-cigarette use and education levels may also impact this reduction.

Our simple machine learning model shows that without further mitigations, the smoking rate is predicted to continue to fall but is at risk of not reaching the government smokefree target by 2030. If the ban is implemented, the target could realistically be reached.



Reference list

Department of Health and Social Care (2023). *Stopping the start: Our New Plan to Create a Smokefree Generation*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/stopping-the-start-our-new-plan-to-create-a-smokefree-generation/stopping-the-start-our-new-plan-to-create-a-smokefree-generation.

Djordjevic, M.V., Stellman, S.D. and Zang, E. (2000). Doses of Nicotine and Lung Carcinogens Delivered to Cigarette Smokers. *JNCI: Journal of the National Cancer Institute*, 92(2), pp.106–111. doi:https://doi.org/10.1093/jnci/92.2.106.

Lederle, F.A., Nelson, D.B. and Joseph, A.M. (2003). Smokers' relative risk for aortic aneurysm compared with other smoking-related diseases: a systematic review. *Journal of Vascular Surgery*, 38(2), pp.329–334. doi:https://doi.org/10.1016/s0741-5214(03)00136-8.

NHS Digital (2019). *Statistics on Smoking, England - 2019 [NS] [PAS] - NHS Digital*. [online] NHS Digital. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-smoking/statistics-on-smoking-england-2019.