Analyzing Tobacco Use in the United States

Group 26:

1. **Rahul Gundekari – 11636305**
2. **Nagakalyan Kalavalapudi – 11601574**
3. **Rakesh Mangalarapu – 11667551**

Table of Contents

1. Introduction

Tobacco use continues to cause challenges to individuals worldwide leading to various preventable illnesses and premature deaths. In the United States, the effects of tobacco abuse are still prevalent hence the need to address the issue and maintain a healthy population. For this reason, we have embarked on a process to conduct data analysis from the CDC to establish ways to control and address this public health challenge.

Our primary objective in this study is to examine the dataset and explore aspects of tobacco use behavior. Through data analysis and modeling techniques, we aim to address key research questions such as the distribution of smokers among various demographic categories regional differences in smoking rates factors influencing attempts to quit smoking, and the potential for predicting successful smoking cessation based on demographic characteristics.

By enhancing our understanding of tobacco use patterns and related influences we aim to contribute towards reducing tobacco usage prevalence and enhancing outcomes in smoking cessation efforts. Ultimately our analysis aims to support health initiatives by providing insights that can guide strategies and interventions aimed at promoting behaviors while lessening the impact of tobacco-related illnesses, in the United States.

2. Literature Review

Research, on the use of tobacco and efforts to quit smoking have delved deeply into how common smoking addiction impacts people trying to quit and how well interventions work in helping people smoke less. Many studies have shown the effects of smoking on health and why it's crucial to have measures in place to control tobacco use. Moreover, studies have looked at factors like age, gender, and education level when studying how people smoke and stop smoking. Various methods to help people quit smoking, such as counseling, medications, and policies have been thorough if put into practice (West, 2017). While there has been progress in

reducing smoking rates there are still challenges in addressing differences in how much different groups smoke and making sure everyone has access, to support when trying to quit.

3. Data Description

The data used for this study was collected from the CDC website, via the State Tobacco Activities Tracking and Evaluation System originating from the Behavioral Risk Factor Surveillance System (BRFSS) survey. It includes details on the rates of cigarette and e-cigarette use, characteristics, frequency of tobacco use, and attempts to quit smoking. The dataset contains 43,341 entries with 31 attributes covering information such as year, location, demographics, and various aspects related to tobacco usage (CDC, 2023).

This dataset provides a view of tobacco consumption patterns along with factors over time and across diverse population segments in the United States. By conducting exploratory data analysis and modeling techniques the objective is to study user habits linked to tobacco usage including variations based on demographics and geography. Moreover, it aims to investigate factors that impact efforts to quit smoking. The main goal is to forecast the likelihood of quitting smoking and offer suggestions to individuals and communities, on reducing tobacco prevalence while enhancing cessation initiatives. This will contribute to improving public health outcomes and overall well-being.

To achieve this, we will mainly employ various major variables from our dataset. These include:

i. **Gender:** This is a fundamental demographic variable indicating whether an individual identifies as male or female. In this analysis, gender is essential for understanding disparities in tobacco use behaviors and cessation efforts between men and women. By examining gender differences, we can identify potential gender-specific interventions and policies to address tobacco-related health disparities.

ii. Age: This is a critical demographic variable representing the chronological age of individuals. Age plays a crucial role in tobacco use behaviors, as smoking prevalence tends to vary across different age groups. Understanding age-related patterns in smoking initiation, cessation, and relapse can inform targeted interventions aimed at specific age cohorts, such as youth prevention programs or smoking cessation initiatives for older adults.

iii. Race: This is a socio-demographic variable reflecting individuals' racial or ethnic backgrounds. In tobacco research, race is significant for understanding disparities in smoking prevalence and cessation outcomes among different racial and ethnic groups. By examining racial disparities in tobacco use, we can identify social, economic, and cultural factors influencing smoking behaviors and develop culturally tailored interventions to address disparities and promote equitable access to smoking cessation resources.

iv. Education: Education level is a socio-demographic variable indicating individuals' educational attainment, such as a high school diploma, college degree, or higher education. Education level is associated with smoking behaviors, with higher levels of education often linked to lower smoking prevalence and higher rates of smoking cessation. Understanding the relationship between education and tobacco use can inform public health interventions targeting vulnerable populations with lower levels of education to reduce smoking prevalence and promote smoking cessation.

v. Smoking Status: Smoking status is a behavioral variable indicating individuals' current smoking behavior, such as current smoking, former smoking, or never smoking. In tobacco research, smoking status is crucial for assessing smoking prevalence, cessation efforts, and relapse rates. By categorizing individuals into smoking status groups, we can analyze trends over time, identify predictors of smoking initiation and cessation, and evaluate the effectiveness of tobacco control policies and interventions.

4. Research Questions

i. What is the distribution of smokers in terms of age, gender, race, and education level?

Hypothesis: The distribution of smokers will vary significantly across different demographic factors, including age, gender, race, and education level. Specifically, we hypothesize that there will be differences in smoking prevalence among different age groups, genders, racial and ethnic groups, and educational attainment levels.

ii. What is the frequency of smoking among different demographics?

Hypothesis: The frequency of smoking will differ among various demographic groups. We predict that certain demographics, such as younger individuals, males, certain racial or ethnic groups, and those with lower education levels, will exhibit higher frequencies of smoking compared to others.

iii. Are there parts of the country where smokers smoke more or less?

Hypothesis: There will be regional variations in smoking prevalence across different parts of the country. We hypothesize that some geographical areas may have higher smoking rates due to differences in cultural norms, socioeconomic factors, and tobacco control policies.

iv. What demographic characteristics, such as age, gender, race, and education level, influence the frequency of quit attempts among smokers?

Hypothesis: Demographic characteristics, including age, gender, race, and education level, will influence the frequency of quit attempts among smokers. We predict that younger individuals, females, certain racial or ethnic minorities, and those with higher education levels will exhibit higher rates of quit attempts compared to their counterparts.

v. What is the percentage of quit attempts among smokers

Hypothesis: The percentage of quit attempts among smokers will vary based on demographic factors. We hypothesize that certain demographic groups, such as older individuals, females,

and those with higher education levels, will have higher percentages of quit attempts compared to others.

vi. Can we predict if a smoker will successfully quit based on factors such as age, gender, or education level?

Hypothesis: It is possible to predict whether a smoker will successfully quit based on factors such as age, gender, and education level. We hypothesize that individuals who are older, female, and have higher education levels will have higher success rates in quitting smoking compared to younger, male, and less educated individuals.

5. Data Preprocessing

Data preprocessing plays a role, in getting the dataset ready for analysis and modeling. It includes tasks like cleaning, transforming, and organizing the data to ensure its quality and suitability for analysis (Kumar, 2021). Here are the key steps taken during the data preprocessing phase:

ii. Dealing with Missing Values

We identified any missing values in the dataset and handled them appropriately by using strategies such as imputation, where missing values were replaced with the mean or median of the respective feature, or deletion, where rows or columns with missing values were removed if deemed appropriate based on the extent of missingness and the impact on the analysis.

iii. Converting Categorical Variables

Categorical variables were converted into a suitable format using methods such as one-hot encoding or label encoding to ensure compatibility with logistic regression algorithms, which require numerical input. This conversion process allowed us to represent categorical variables as binary or numerical values, facilitating their inclusion in the modeling process.

iv. Creating New Features

To enhance the predictive power of our models, we generated new features from existing ones by combining, transforming, or extracting additional information. This step aimed to capture more nuanced relationships within the data and create features that better represent the underlying patterns and trends present in the dataset. The variables that underwent this preprocessing included 'Age,' 'Gender,' 'Education,' 'TopicDesc,' and 'MeasureDesc,' which were carefully engineered to extract specific variables of factors influencing smoking behavior and cessation attempts.

v. Standardizing Data

Numeric variables were standardized to a common scale to prevent features with larger magnitudes from overshadowing others during the model training process. Standardization ensures that all features contribute equally to the model fitting process and helps improve the stability and convergence of the algorithms.

vi. Splitting Data

We partitioned the dataset into training and testing sets to assess the performance of our models. This approach allowed us to train the models on a subset of the data and evaluate their performance on unseen data, providing valuable insights into their generalization ability and potential issues such as overfitting.

vii. Feature Selection

Given the dataset's wide range of features, we employed various methods such as ranking feature importance and dimensionality reduction techniques to identify and select the most significant features for modeling. This step helped streamline the modeling process by focusing on the most informative features while reducing computational complexity and potential overfitting.

The rationale behind feature selection for the variables 'Age', 'Gender', 'Education', 'TopicDesc', and 'MeasureDesc' lies in identifying the most influential factors that contribute to the target variable of interest. In this context, these variables serve as potential predictors for understanding smoking behavior, cessation attempts, and other relevant outcomes. By selecting these variables for modeling, we aim to capture the demographic characteristics and behavioral patterns that are most strongly associated with smoking habits and cessation efforts. This selection process allows us to focus on the factors that are most likely to impact the target variable, thereby improving the predictive power and interpretability of the model. Additionally, by including these variables in the modeling process, we can uncover valuable insights into the complex interplay between demographic factors and smoking behavior, ultimately facilitating more targeted public health interventions and policy initiatives.

6. Result and Discussion

Exploration Data Analysis and Visualization

In this process, our main aim is to identify how various variables or features interact with each other and probably depict patterns or relationships that we can discern to establish how various factors affect tobacco users.

a) Distribution of smokers by demographics

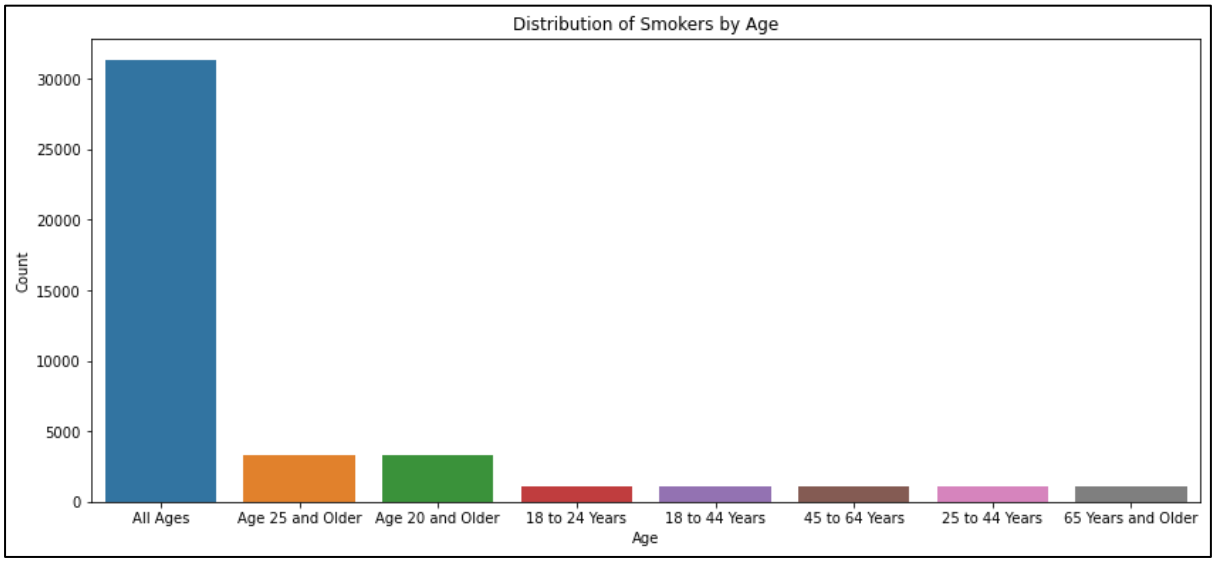

Figure 1: Distribution of Smokers by Age

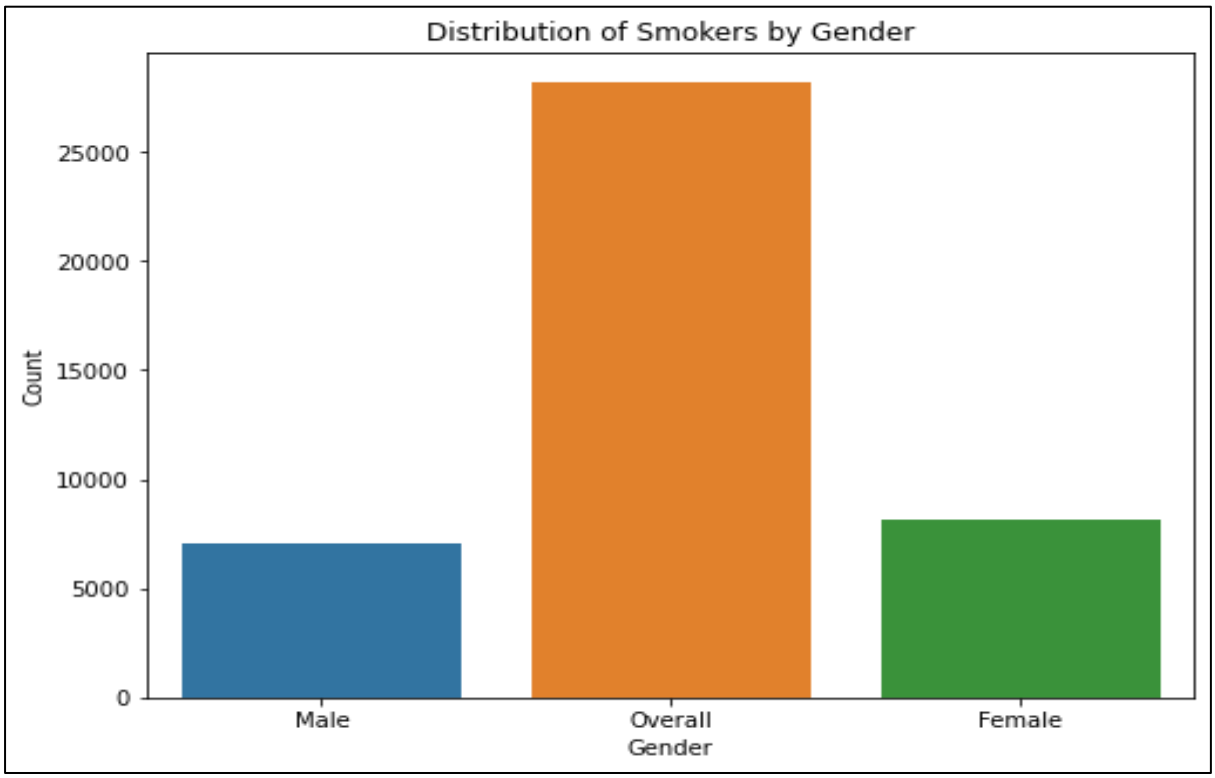

Figure 2: Distribution of Smokers by Gender

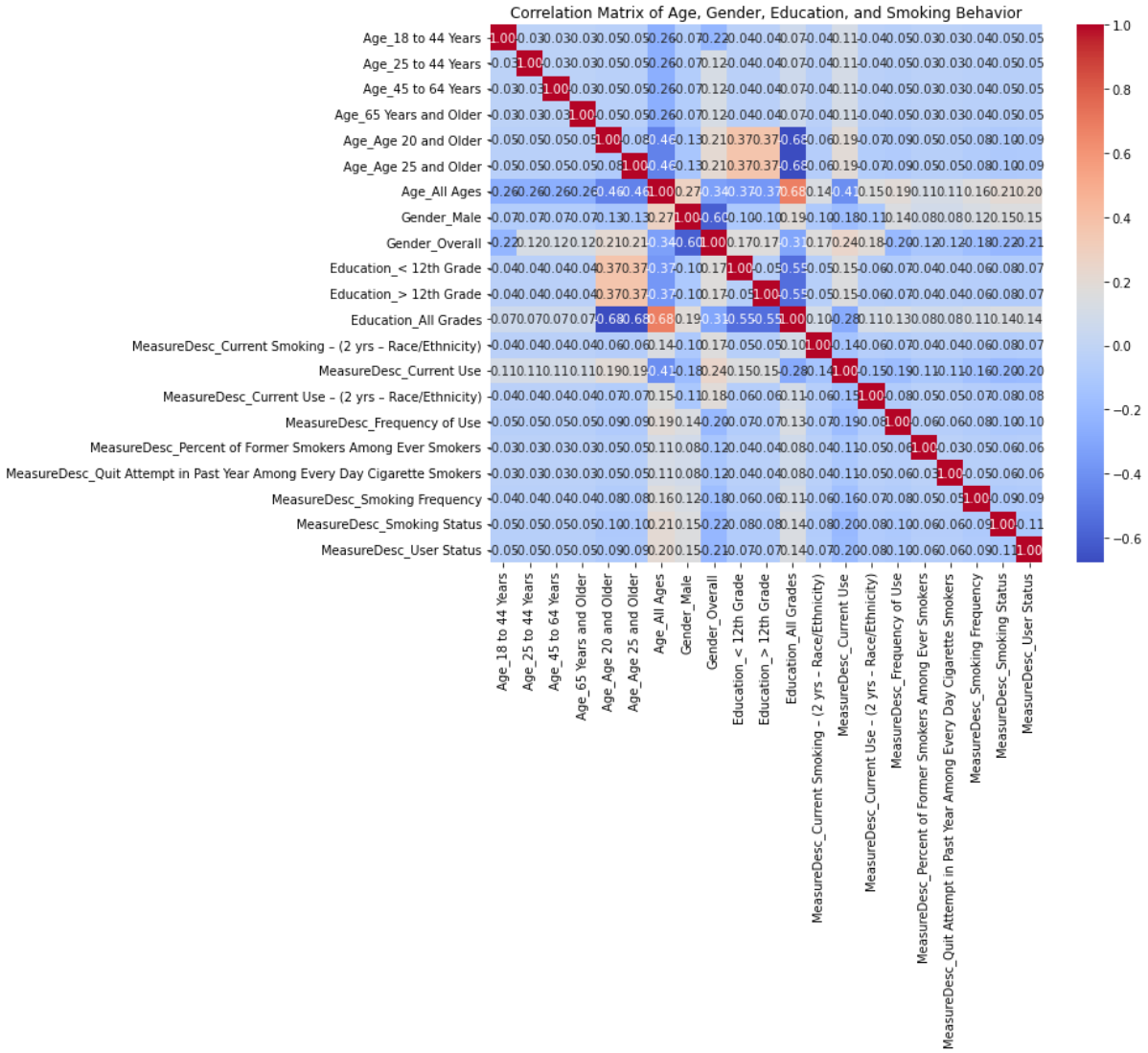

Figure 2.1: Correlation Matrix of Age, Gender, Education, and Smoking Behavior

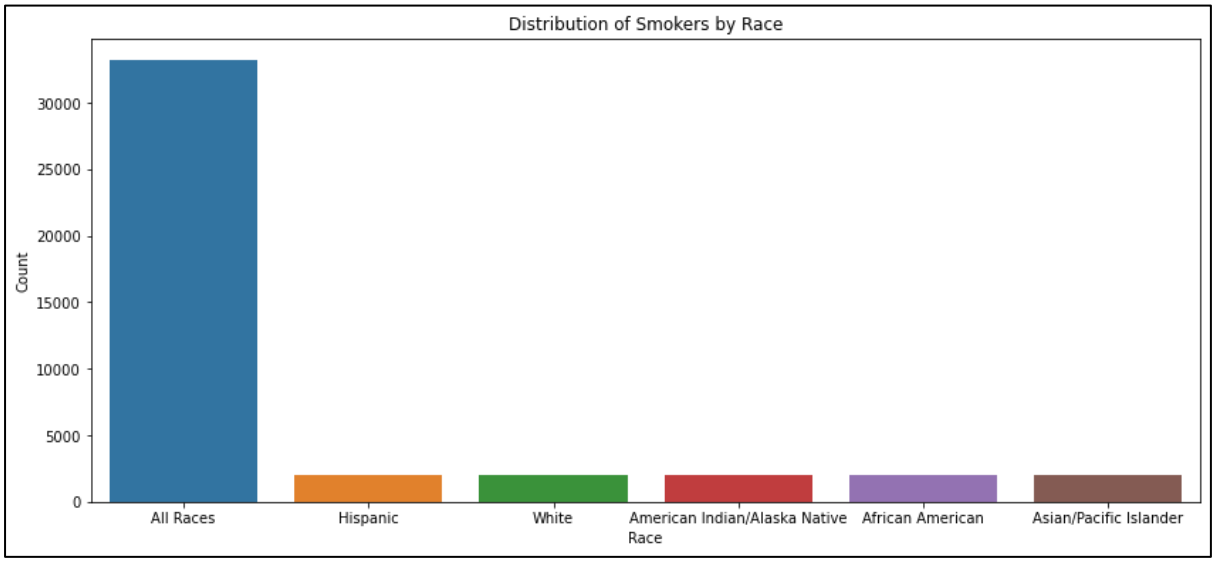

Figure 3: Distribution of Smokers by Race

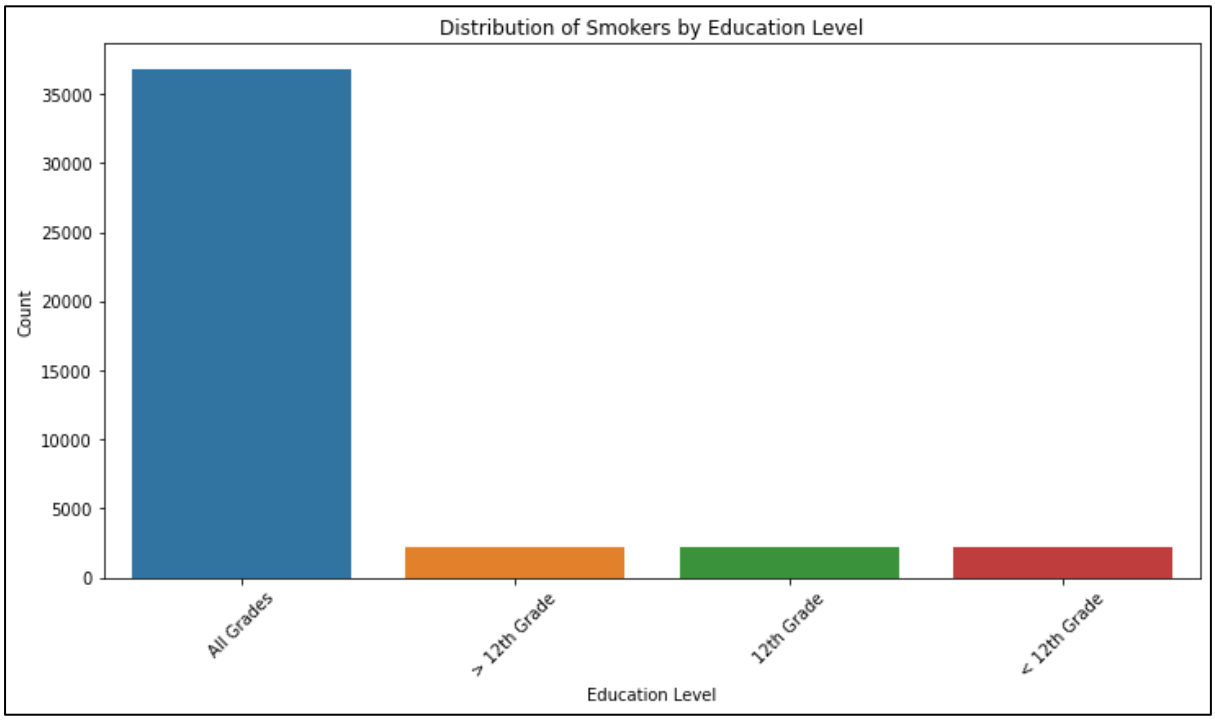

Figure 4: Distribution of Smokers by Education Level

The visualizations provide valuable insights into the distribution of demographic attributes within the dataset. Figure 1 illustrates that the majority of samples are drawn from a broad age range, encompassing individuals of all ages. Additionally, specific age groups, such as those aged 20 and older or 25 and older, are also represented, allowing for targeted analysis within these cohorts. Similarly, the distribution of gender in Figure 1 reveals a balanced representation of data from both male and female participants, with relatively few instances exclusively focusing on one gender. This gender inclusivity is essential for capturing a comprehensive understanding of smoking behavior across diverse populations.

Moving on to the correlation analysis depicted in Figure 2.1, a noteworthy correlation coefficient of 0.68 is observed between education and age, particularly among individuals aged 25 years or older. This moderate positive correlation suggests that as individuals progress through adulthood, their level of education tends to increase. Possible explanations for this

trend may include increased access to educational opportunities or the pursuit of higher education later in life. Understanding this relationship is crucial for tailoring interventions aimed at specific age groups to address smoking behavior effectively.

Furthermore, the equitable distribution of sample data across different racial and educational categories, as highlighted in the visualizations, underscores the importance of avoiding bias or skewness in the dataset. By ensuring representation from diverse racial and educational backgrounds, the dataset becomes more reflective of the broader population, enhancing the validity and generalizability of the analysis outcomes. Additionally, the consistent size of data collected from various states further minimizes potential biases, allowing for robust and reliable analysis of smoking behavior patterns on a regional scale. Deductively, these detailed observations emphasize the significance of comprehensive data collection strategies and rigorous analysis techniques in informing public health interventions and policy decisions aimed at addressing tobacco use.

b) Frequency of Smoking Among Different Demographic Groups

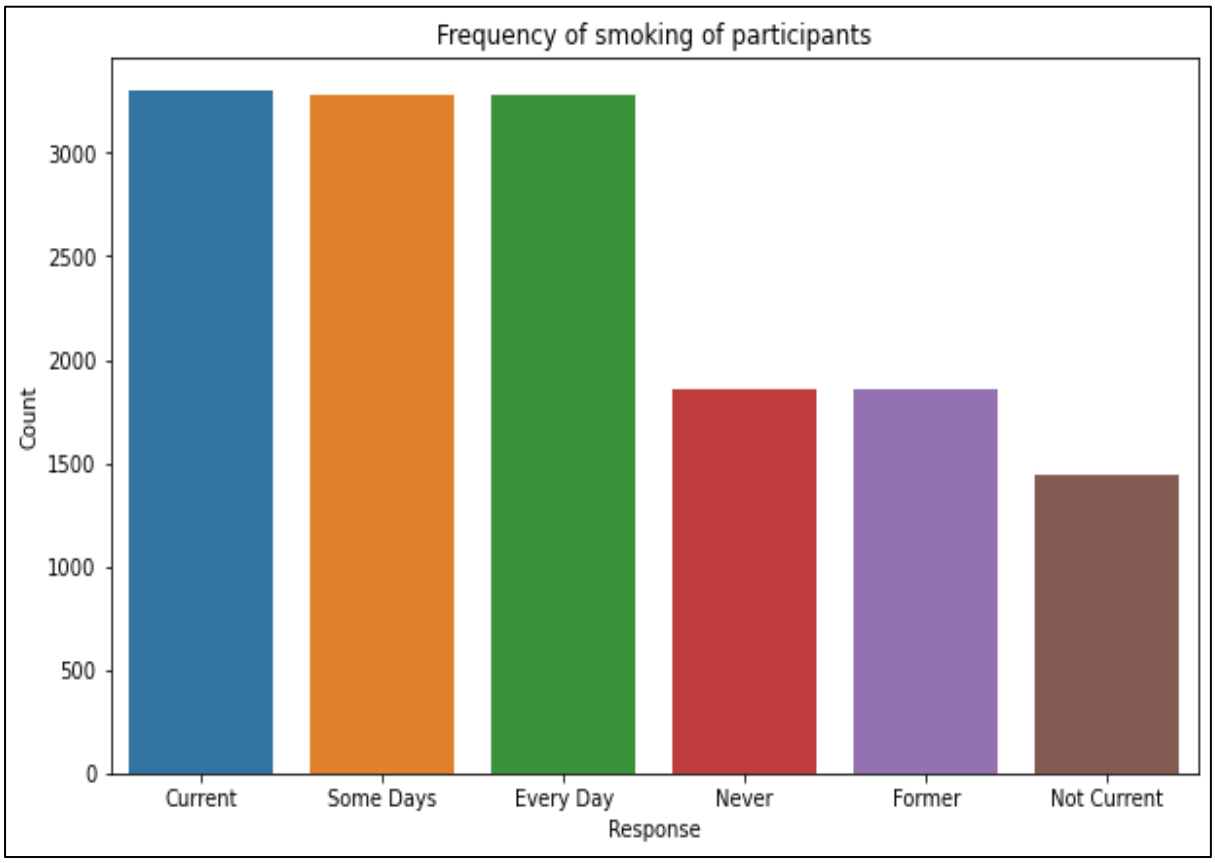

Figure 5: Frequency of Smoking of Participants

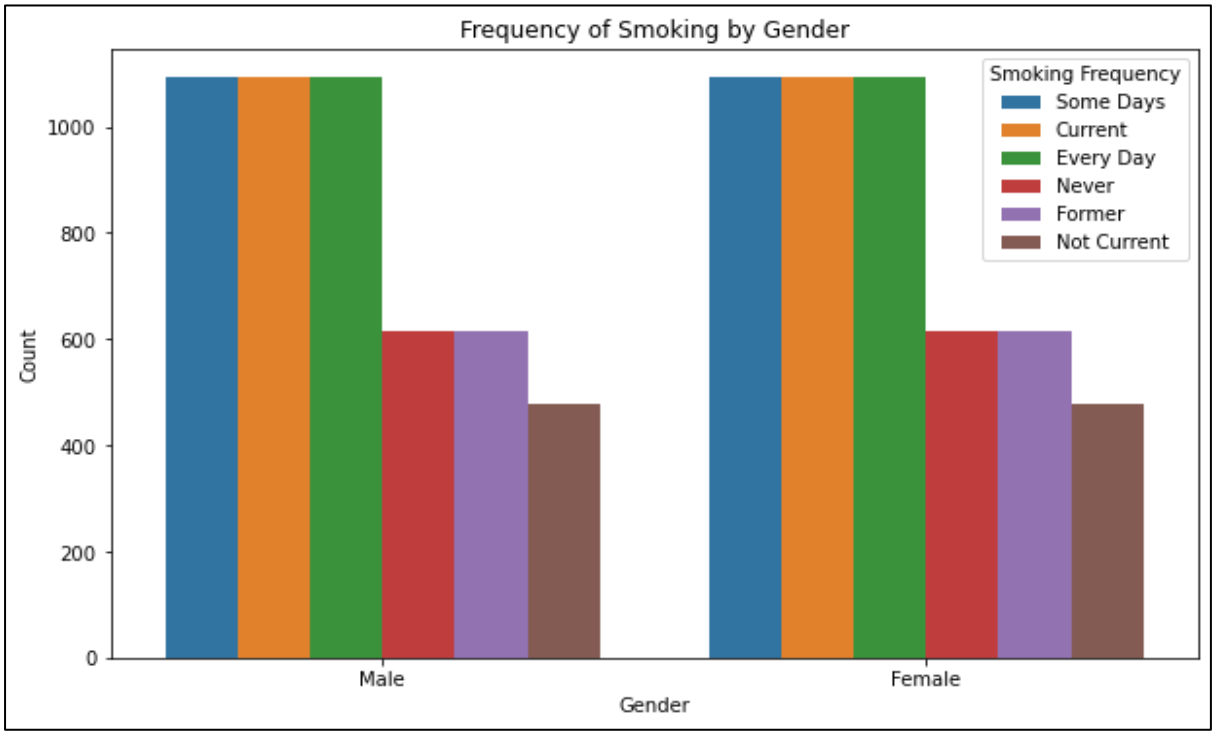

Figure 6: Frequency of Smoking by Gender

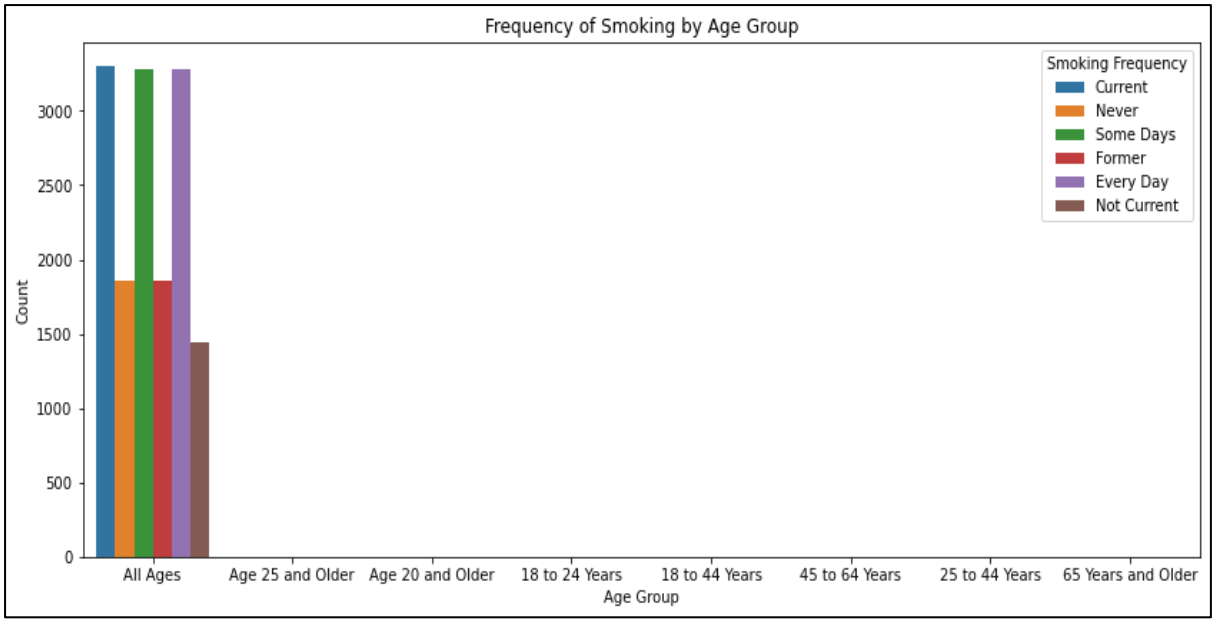

Figure 7: Frequency of Smoking by Age Group.

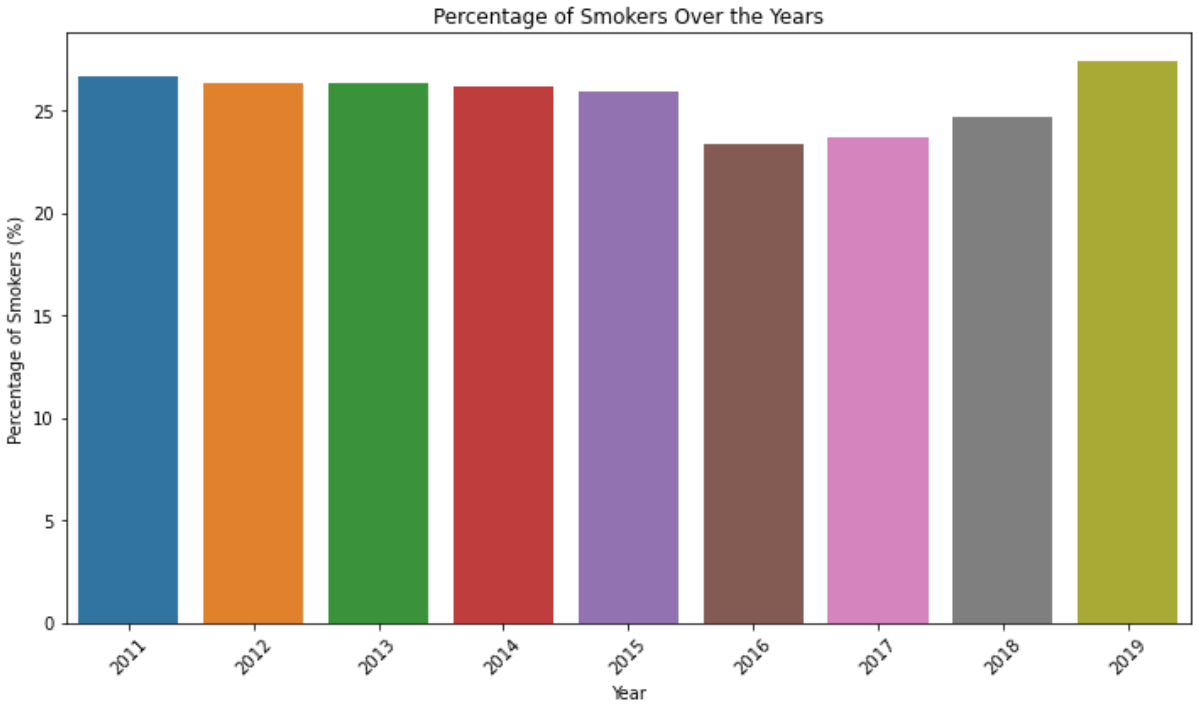

Figure 7.1: Percentage of Smokers Over the Years

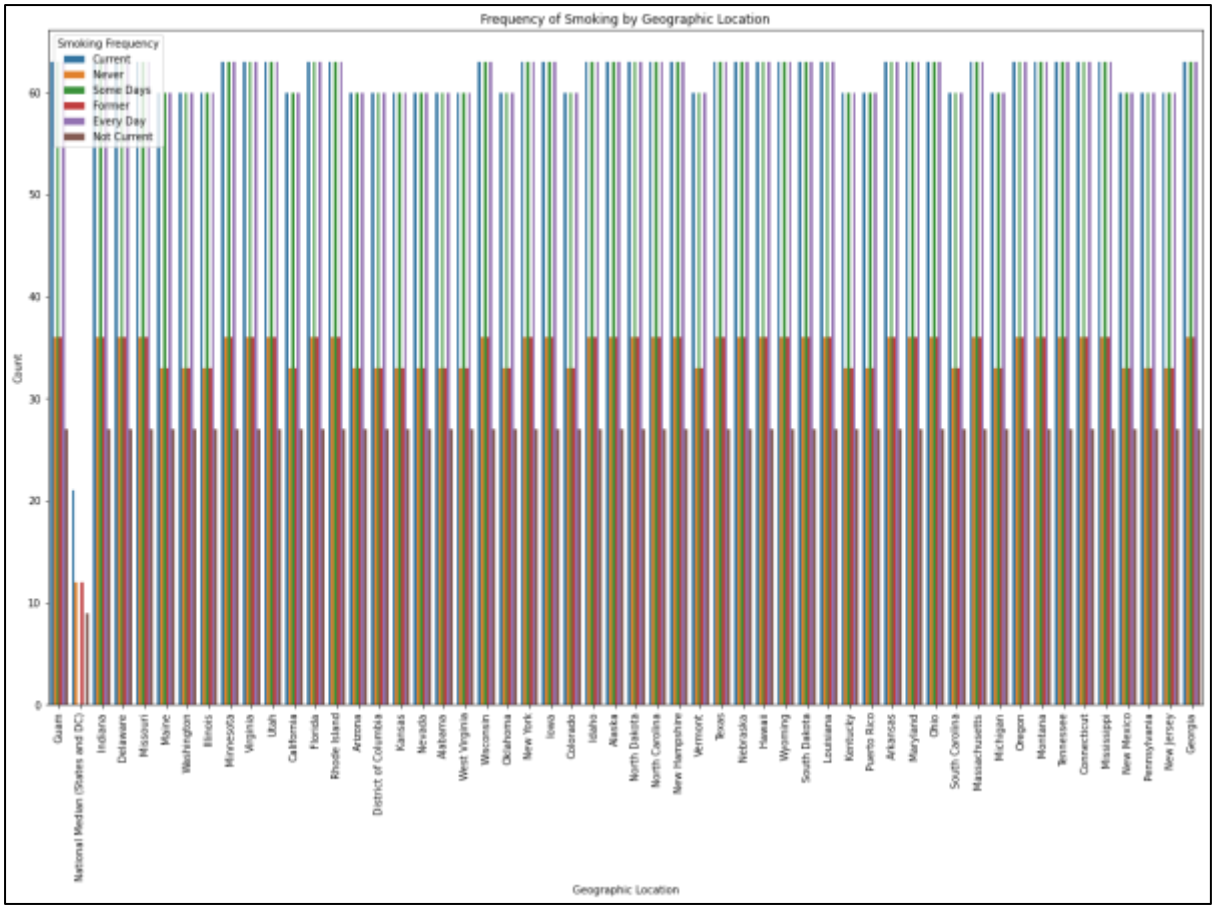

Figure 8: Frequency of Smoking Geographic Location

The comprehensive examination of smoking habits depicted in the figures highlights the widespread and enduring nature of tobacco smoking among the population. Figure 5 presents a thorough analysis of the frequency of smoking among the participants surveyed, indicating a substantial portion of them engage in smoking either on a daily basis or infrequently. Figure 6 provides insight into the distribution of smoking frequency among genders, demonstrating a virtually equal proportion of men and women. Although there is no specific data on how often individuals smoke, the overall perspective across different age groups emphasizes the prevalence of smoking habits among various demographic segments.

Furthermore, the examination of average smoking rates over the years reveals intriguing trends in smoking prevalence. The gradual decline observed from 2011 to 2015 suggests potential successes in public health initiatives aimed at reducing smoking rates. However, the subsequent uptick in smoking rates from 2016 onwards signals a concerning

reversal of this trend, warranting further investigation into the underlying factors driving these fluctuations. These findings underscore the importance of continuous monitoring and assessment of smoking behaviors to inform targeted interventions and policy measures aimed at curbing tobacco use.

Moreover, the analysis highlights the consistent pattern of smoking prevalence across various racial groups and geographical regions. States such as Guam, Indiana, and Delaware exhibit higher smoking frequencies, indicating localized disparities in smoking behaviors that may necessitate tailored interventions. Conversely, states like Maine, Washington, and Illinois demonstrate relatively lower smoking frequencies, suggesting potential success in implementing effective tobacco control measures. Deductively, these insights emphasize the need for multifaceted approaches to address smoking prevalence, encompassing targeted interventions, public awareness campaigns, and policy interventions tailored to specific demographic groups and geographic regions.

c) Frequency of Quit Attempts

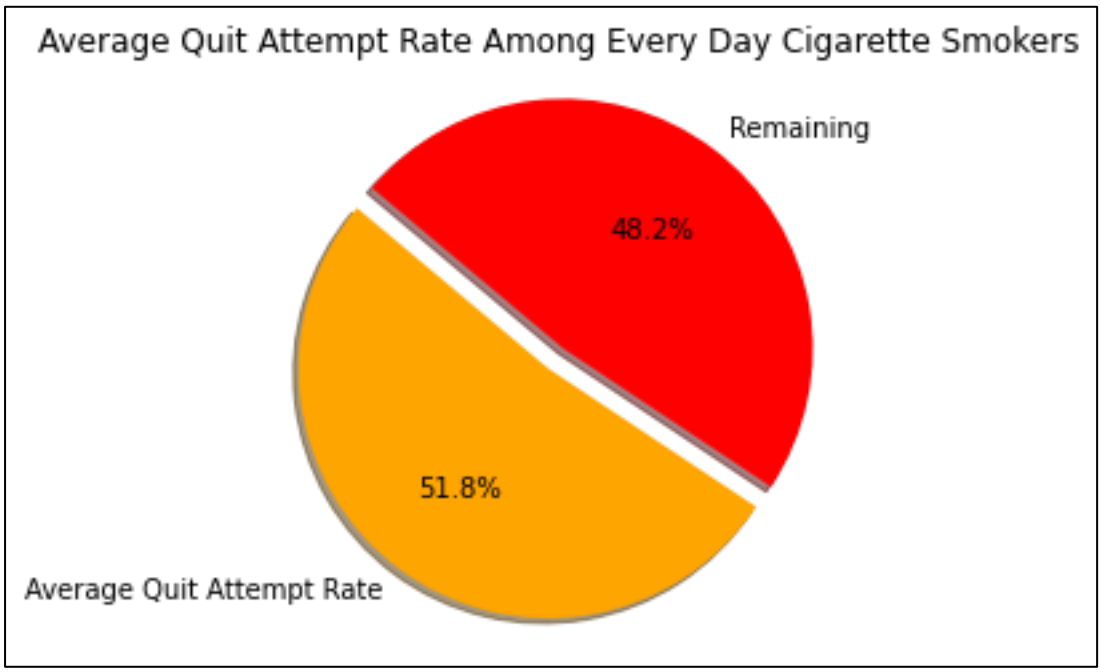

Figure 9: Average Quit Attempt Rate Among Every Day Cigarette Smokers

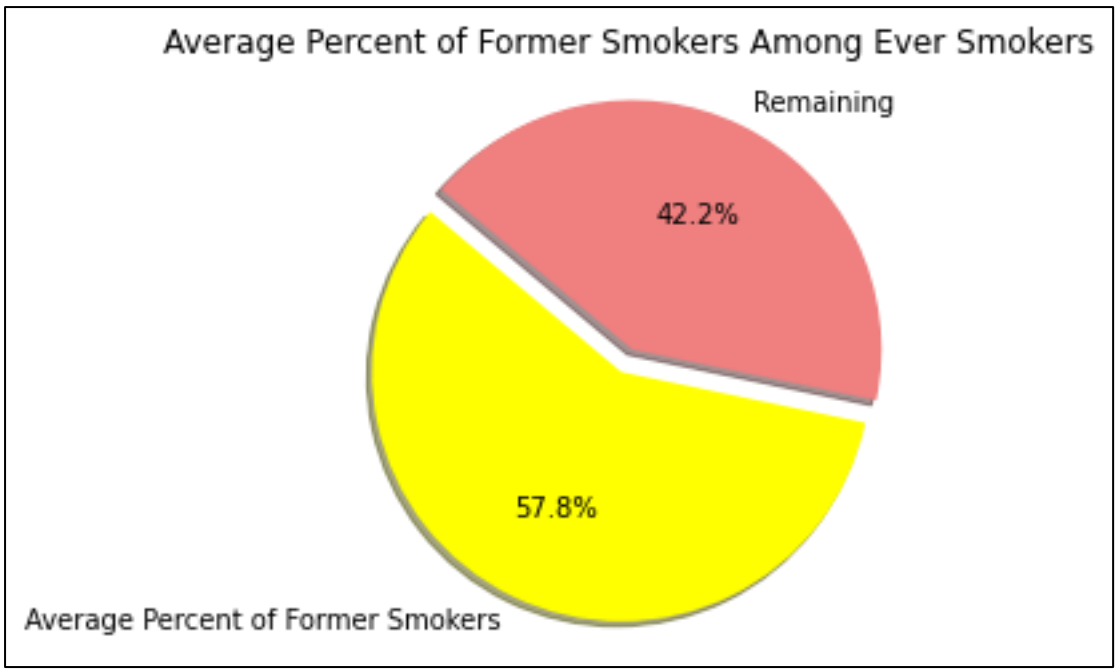

Figure 10: Average Percent of Former Smokers Among Ever Smokers

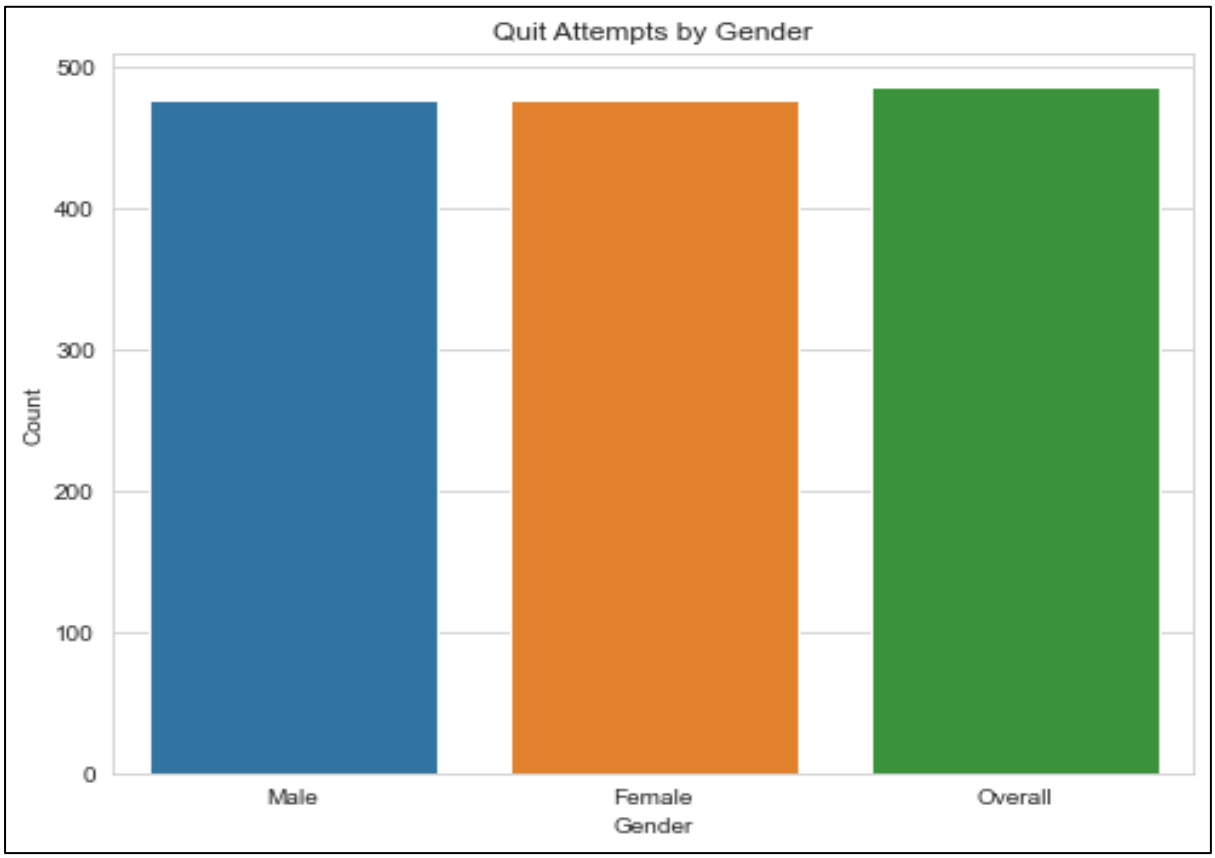

Figure 11: Quit Attempts by Gender

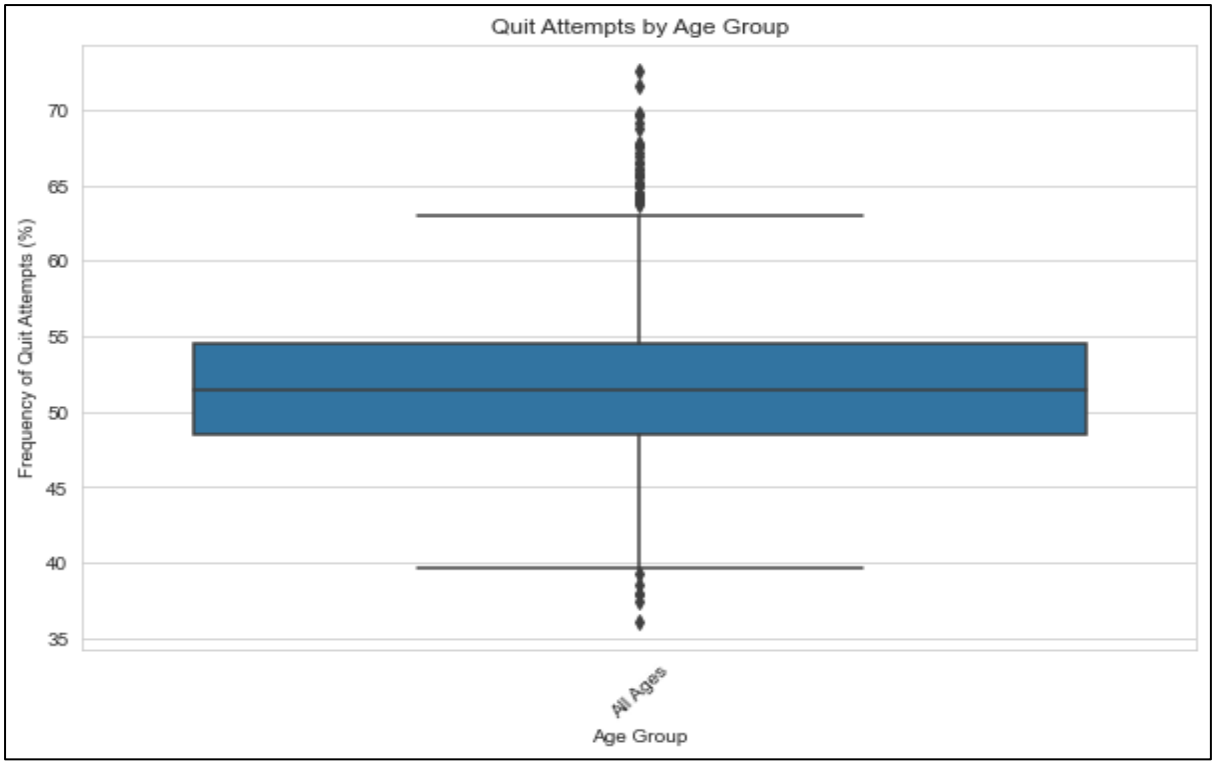

Figure 12: Quit Attempts by Age Group

In the above observations, as depicted in Figure 9, we note a promising 51.8% average quit rate attempt. While not reaching the ideal public health target, it's encouraging to see over half of everyday smokers attempting to quit tobacco use. Furthermore, Figure 10 illustrates a notably higher average percent of former smokers among ever-smokers at 57.8%. This suggests that more than half of ever smokers have successfully quit using tobacco, which is a significant achievement. Gender-wise, quit attempts appear to be evenly distributed between males and females. Additionally, Figure 12 highlights that individuals of all ages are making efforts to quit tobacco use, with an average attempt rate of approximately 52%. These findings reflect encouraging progress in combating tobacco abuse.

d) Modeling and Evaluation

In the modeling phase, we applied two machine learning algorithms to the pre-processed dataset to construct predictive models and address the research question. These algorithms are Logistic Regression and Random Forest Classifier.

```
In [68]: # Evaluating the model
         print ('Logistic Regression: \n')
         print(classification_report(y_test, y_pred))

         Logistic Regression:

                       precision    recall  f1-score   support

                    0       0.82      1.00      0.90      2009
                    1       1.00      0.47      0.64       805

             accuracy                           0.85      2814
            macro avg       0.91      0.73      0.77      2814
         weighted avg       0.87      0.85      0.83      2814
```

Figure 13: Logistic Regression Classification Report

In the logistic regression model evaluation, the precision for predicting non-smokers (class 0) is 0.82, indicating that 82% of instances predicted as non-smokers were correctly classified. On the other hand, the precision for identifying smokers (class 1) is 1.00, suggesting that all instances predicted as smokers were indeed smokers. The recall, or sensitivity, for class 0 is 1.00, indicating that all actual non-smokers were correctly identified, while for class 1, the recall is 0.47, indicating that only 47% of actual smokers were correctly classified.

The overall accuracy of the model is 85%, meaning it correctly predicts the smoking status of individuals 85% of the time. The macro average F1-score, which balances precision and recall across both classes, is 0.77. This indicates a good level of performance in achieving a balance between precision and recall across the two classes. This means that the model's predictive capability is relatively consistent across both smoker and non-smoker categories.

With a weighted average F1-score of 0.83, it suggests that the model performs better in terms of precision and recall for the majority class (non-smokers) compared to the minority class (smokers). While this weighted average F1-score reflects the overall effectiveness of the model in capturing the underlying patterns in the data, it also underscores the importance of

improving the model's ability to correctly identify smokers, especially considering the potential public health implications of tobacco use.

The performance of both logistic regression and Random Forest Classifier models in predicting smoking status demonstrates their effectiveness in analyzing the dataset. The logistic regression model achieved an accuracy of 85%, with a precision of 0.82 for non-smokers and 1.00 for smokers. Similarly, the Random Forest Classifier yielded comparable results, indicating the robustness of both approaches. These findings suggest that both models can accurately predict the smoking status of individuals, with logistic regression offering a simpler interpretation of the results and Random Forest providing robust predictions.

For the random forest model, the cross-validation scores are as follows: [0.7311, 0.7352, 0.7464, 0.7482, 0.7542]. The mean ROC-AUC score across all folds is approximately 0.7430. These scores indicate the performance of the model in distinguishing between positive and negative classes across different validation sets. Similar to logistic regression, the consistent scores across folds suggest that the random forest model generalizes well to unseen data, and the mean score provides an overall estimate of its predictive capability.

The SVM model achieved a ROC-AUC score of approximately 0.734, while the gradient-boosting model achieved a slightly higher ROC-AUC score of approximately 0.745. This indicates that both models perform relatively well in distinguishing between smokers and non-smokers based on the provided features. Among the four algorithms evaluated, logistic regression demonstrated the highest ROC-AUC score, indicating superior performance in predicting smoking status based on the provided features. This suggests that logistic regression may be the most effective algorithm for this particular task.

7. Recommendations

Based on our analysis we would recommend the following for public health initiatives:

a. Improve programs that help people quit smoking by tailoring them to groups based on factors, like age, gender, and education level. These programs should offer support and resources to those trying to kick the habit drawing from quitting experiences across various segments of society.

b. Focus on areas where smoking rates are high, such as Guam, Indiana, and Delawar,e and prioritize targeted actions to reduce tobacco use in these regions. Tailored strategies should consider local factors influencing smoking habits and efforts to quit working closely with community groups and healthcare providers for impact.

c. Launch campaigns to educate people about the prevalence of smoking and the benefits of quitting. Use a mix of media platforms to reach audiences and highlight stories of individuals who have successfully given up smoking as motivation for others. For instance,

The Campaign for Tobacco-Free Kids (CTFK) helped Bangladesh reduce adult tobacco use by 18.5% between 2009 and 2017. Exposure to second-hand smoke declined significantly in homes (-15.9%) and public spaces, with restaurants and healthcare facilities seeing declines of 30% and 11.1%, respectively.

d. Advocate for evidence-based policies that aim to decrease tobacco consumption through measures like taxes on tobacco products, smoke regulations, and limits on tobacco advertising. By implementing these policies, we can create an environment for quitting smoking and deter populations from starting the habit. For example, in Brazil 2009, the state of São Paulo adopted a “smoke-free” law that banned smoking in enclosed public spaces. Ten years after the law was adopted, the compliance rate in São Paulo was 99.7%, and smoking had decreased from 18.8 % in 2006 to 13.5% in 2019. Furthermore, the prevalence of smoking in

Brazil declined from 16.2% in 2006 to 9.8% in 2019. In 2019, the World Health Organization (WHO) recognized Brazil as the second country to adopt all MPOWER measures fully.

e. Encourage collaboration, across sectors. Encourage teamwork, between government agencies, healthcare providers, community groups, and other involved parties to create a strategy, for managing tobacco use. By combining resources and knowledge these stakeholders can put into action thorough tobacco control plans that tackle the elements affecting tobacco usage and quitting.

I. Cost-Effectiveness Analysis and Feasibility Assessment of the Recommendations

Cost-Effectiveness Analysis and Feasibility Assessment		**Feasibility Cost**
a. **Improve Quitting Programs**	Tailoring programs to specific groups can be highly effective, but it may also increase costs due to the need for more specialized resources and personnel. However, the long-term benefits, such as reduced healthcare costs and increased productivity, often outweigh these initial costs. The feasibility of this recommendation is high, as many countries already have some form of smoking cessation	**High**

	programs in place that can be improved upon.	
b. **Focus on High Smoking Rate Areas**	Focusing on areas with high smoking rates, like Guam, Indiana, and Delaware, can be cost-effective as these areas may see the most significant health improvements from reduced smoking. The feasibility is moderate to high, depending on the local resources and infrastructure available for implementing targeted actions.	**Moderate**
c. **Educational Campaigns**	While launching educational campaigns can be costly, especially when using a mix of media platforms, they can also be highly effective in changing public perceptions and behaviors regarding smoking. The success of the CTFK in Bangladesh is a testament to this. The feasibility is high, given the widespread availability and reach of various media platforms today.	**High**

d. **Advocate for Policies**	Advocating for policies like taxes on tobacco products, smoke regulations, and limits on tobacco advertising can be a highly cost-effective way to decrease tobacco consumption. The cost of implementing such policies is often minimal compared to the potential health and economic benefits. The feasibility is moderate, as it often requires navigating complex political landscapes and dealing with potential opposition from the tobacco industry.	**Moderate**
e. **Encourage Collaboration**	Encouraging collaboration across sectors can lead to more comprehensive and effective tobacco control plans. While it may increase coordination costs, it can also lead to better resource allocation and more innovative solutions. The feasibility is high, as it primarily requires effective	**High**

	communication and coordination among existing entities.	

II. Potential Challenges or Barriers to Implementing the Recommendations

a. One of the main challenges is the need for specialized resources and personnel to tailor programs to specific groups. This could include language barriers, cultural differences, and varying levels of education and health literacy. Additionally, tracking and evaluating the effectiveness of these programs can be complex and time-consuming.

b. Also, implementing targeted actions in areas with high smoking rates may face resistance from local communities, particularly if tobacco use is deeply ingrained in the local culture. There may also be logistical challenges in reaching remote or underserved areas.

c. Educational campaign effectiveness can also be hard to measure, and there's a risk that the message may not reach or resonate with the intended audience. There may also be pushback from the tobacco industry, which often uses sophisticated marketing strategies to counter public health messages.

d. Advocacy efforts can face significant political and legal hurdles. The tobacco industry often has considerable influence and may lobby against policies that could hurt their profits. Additionally, implementing and enforcing policies like smoke-free laws and advertising restrictions can be challenging, particularly in regions with weak regulatory systems.

e. Additionally, while collaboration can lead to more effective strategies, it can also be challenging to coordinate efforts across different sectors and organizations. There may be disagreements over priorities, strategies, and resource allocation, and it can be difficult to ensure that all stakeholders are working towards the same goals.

8. Conclusion

In conclusion, examining the patterns of tobacco use and efforts to quit among groups shows the challenges and opportunities in public health initiatives to reduce tobacco-related harm. Despite the smoking seen in population segments, positive trends in quitting smoking attempts and successful cessation rates indicate effective strategies for intervention. By utilizing data on demographics like age, gender, and education level tailored smoking cessation programs and public health campaigns can be developed to cater to diverse community needs.

Moving ahead it is crucial to prioritize evidence-based policies to boost public awareness campaigns and encourage collaboration across sectors to advance tobacco control endeavors. Through investing in cessation programs targeting high-risk areas and implementing policies public health agencies can make significant progress in reducing tobacco use and enhancing overall population well-being. Continuous monitoring of tobacco using trends alongside research and innovation in control methods will be vital for long-term success, against the tobacco crisis.

The analysis presented in this study is subject to several limitations that should be acknowledged. Firstly, the dataset used may contain biases inherent to survey data, including self-reporting biases and underrepresentation of certain demographic groups. Additionally, the modeling approach, while effective, may oversimplify the complex relationships between smoking behavior and demographic factors. Furthermore, the absence of detailed information on individual smoking histories or socioeconomic status may limit the depth of analysis and predictive accuracy of the models. Despite these limitations, the findings provide valuable insights into smoking behavior and highlight the need for further research to address these shortcomings.

Future research endeavors could explore several avenues to enhance our understanding of smoking behavior and cessation efforts. Longitudinal studies tracking individuals over time could provide valuable insights into the long-term effectiveness of cessation programs and factors influencing successful quitting. Additionally, investigating the impact of emerging tobacco products, such as e-cigarettes, on smoking behavior and cessation outcomes could inform public health policies and interventions. Furthermore, examining the intersectionality of demographic factors, including race, gender, and socioeconomic status, in shaping smoking behavior could lead to more targeted and effective smoking cessation strategies tailored to specific populations. Overall, future research endeavors should aim to address the limitations of existing studies and advance our understanding of smoking behavior to inform evidence-based public health interventions.

9. References

Campaign for Tobacco-Free Kids. (2022, July 19). Tobacco control success story: India. https://www.tobaccofreekids.org/problem/toll-global/asia/india/case-study-india

CDC. (2023). Behavioral risk factor data: Tobacco use (2011 to present). Data | Centers for Disease Control and Prevention. https://data.cdc.gov/Survey-Data/Behavioral-Risk-Factor-Data-Tobacco-Use-2011-to-pr/wsas-xwh5/about_data

Coughlin, L. N., Tegge, A. N., Sheffer, C. E., & Bickel, W. K. (2018). A machine-learning approach to predicting smoking cessation treatment outcomes. Nicotine & Tobacco Research, 22(3), 415-422. https://doi.org/10.1093/ntr/nty259

Doogan, N., Roberts, M., Wewers, M., Stanton, C., Keith, D., Gaalema, D., Kurti, A., Redner, R., Cepeda-Benito, A., Bunn, J., Lopez, A., & Higgins, S. (2017). A growing geographic disparity: Rural and urban cigarette smoking trends in the United States. Preventive Medicine, 104, 79-85. https://doi.org/10.1016/j.ypmed.2017.03.011

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 27(2), 265 - 276. https://doi.org/10.1016/j.hrmr.2016.08.003

Kumar, D. (2021, June 20). Introduction to data Preprocessing in machine learning. Medium. https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

Nelson, S. E., Van Ryzin, M. J., & Dishion, T. J. (2014). Alcohol, marijuana, and tobacco use trajectories from age 12 to 24 years: Demographic correlates and young adult substance use problems. Development and Psychopathology, 27(1), 253 277. https://doi.org/10.1017/s0954579414000650

Souza, L. E., Rasella, D., Barros, R., Lisboa, E., Malta, D., & Mckee, M. (2021). Smoking prevalence and economic crisis in Brazil. Revista de Saúde Pública, 55, 3. https://doi.org/10.11606/s1518-8787.2021055002768

West, R. (2017). Tobacco smoking: Health impact, prevalence, correlates, and interventions. Psychology & Health, 32(8), 1018 - 1036. https://doi.org/10.1080/0887 0446.2017.1325890