**Project Overview**

In recent years, smoking and drug use among young people have turned into a significant public health problem, affecting both society and the economy. This project plans to examine data on young people's behaviors, focusing on factors like age, gender, socioeconomic background, and habits related to smoking and drug use. By analyzing this information, the project aims to identify key patterns and influential factors that could help provide valuable insights for public health officials, educators, and policymakers who work to address these issues.

**Personal Objective**

The goal of this project is to conduct a detailed analysis to uncover patterns and connections in youth smoking and drug use behaviors. Using the dataset provided, the project will look at important questions related to factors such as age, gender, income, and lifestyle choices that may be linked to these behaviors. The aim is to build a strong data-based foundation that can support public health recommendations. This analysis can contribute to a better academic understanding of youth smoking and drug use and has practical value for creating focused intervention programs that target youth groups at higher risk.

Through this project, I hope to answer questions like:

* What are the main demographic factors linked with youth smoking and drug use?
* How do socioeconomic factors relate to these behaviors?
* Are there visible trends across different age groups or regions?

The insights gained should help build a foundation for data-driven decisions that can strengthen public health programs and youth-oriented initiatives

**Intended Outcomes**

The project intends to deliver several clear outcomes that could be useful for various groups:

1. **Identify Key Influencers**: By analyzing the data, we aim to find the main demographic and socioeconomic factors linked to youth smoking and drug use. This may include influences like age, income, and education level, which can help us understand what drives these behaviors and guide prevention efforts.
2. **Highlight High-Risk Demographics**: Visualizations will be created to show which groups or regions are at higher risk of smoking and drug use. This is important because it will turn complex data into understandable information that stakeholders can use.
3. **Provide Actionable Recommendations**: Based on the insights discovered, we will create a list of recommendations tailored to meet the specific needs of the high-risk groups. These recommendations can support public health officials and educators in setting up more targeted and effective preventive programs aligned with the identified risk factors.

By achieving these outcomes, this project will offer meaningful, data-driven insights that can guide public health efforts aimed at reducing youth smoking and drug use.

**Audience Needs**

This project has been designed with three main groups in mind: public health officials, educators, and policymakers. Each of these groups has specific needs that this project aims to address.

1. **Public Health Officials**: These professionals lead the way in creating and applying health strategies. They need data-driven insights to allocate resources effectively and design health programs that tackle the core reasons for youth smoking and drug use. This analysis will give them the statistical and demographic insights they need to plan targeted interventions.
2. **Educators**: Schools play a vital role in shaping young people’s behavior. By understanding the factors that influence smoking and drug use, educators can introduce awareness programs and support systems that address these risks within school settings. The data insights will also help educators communicate these issues more effectively to both students and parents.
3. **Policymakers**: Policymakers control funding and create regulations that impact public health. They need clear, data-based recommendations to form policies that support preventive measures and fund public health programs focused on youth health. This project will aid policymakers in prioritizing funds, designing awareness campaigns, and potentially creating regulatory measures.

These targeted insights and recommendations will help each of these groups make well-informed decisions that encourage healthier lifestyle choices for youth.

**Foreseeable Challenges**

The project may face several challenges that we need to address to ensure reliable analysis and useful insights:

* **Missing Data**: Some information may be missing from the dataset, so we’ll need to decide on ways to fill in or manage these gaps.
* **Data Consistency**: Some data entries might not align properly, so it’s important to ensure that all information is consistent.
* **Outliers**: Some values may be significantly different from others, which could affect our results. These will need careful handling to prevent skewed analysis.

By identifying and addressing these challenges in advance, we can improve the quality of the analysis and make sure that the insights produced are both useful and ethically sound.

**Description of the Dataset to Be Used**

The dataset chosen for this project contains essential information on youth behaviors, particularly focusing on smoking and drug use patterns. It includes several key variables:

* **Age**: The ages of individuals, allowing us to observe trends across different age groups.
* **Gender**: Information on gender, helping us explore any behavioral patterns or differences that may exist between genders.
* **Socioeconomic Status**: Data on economic backgrounds, as factors like income level and economic standing often play a role in health-related behaviors.
* **Smoking and Drug Use Habits**: Specific information about smoking and drug use behaviors, central to our analysis.

This data set provides a thorough view of the factors potentially linked to smoking and drug use among youth. By examining these variables, the project aims to uncover meaningful patterns and relationships that can guide public health efforts and support educators and policymakers in creating informed, effective interventions.

**Milestone 3: Data Cleaning and Analysis**

**1. Exploratory Analysis**

* **Summary Statistics**: Identify distributions, outliers, and key metrics.
* **Data Types**: Inspect data types and ensure consistency.
* **Visualizations**: Create histograms, scatterplots, and boxplots for an initial understanding.

**2. Identify Metrics and KPIs**

* **KPIs (Key Performance Indicators)**:
  + Smoking prevalence by age group and gender.
  + Drug experimentation rates for high-risk demographics.
  + Percentage of missing data.

**3. Clean, Filter, and Edit Data**

* Fix inconsistencies in categorical variables.
* Detect and handle outliers using IQR.
* Remove duplicates if found.

**4. Ensure Data Quality**

* Validate categorical labels.
* Ensure numeric data aligns with expected ranges.

**5. Impute Missing Data**

* **Linear Regression**: Predict missing values based on correlated variables.
* **Mean Imputation**: Replace missing values with the mean.
* **Ratio Imputation**: Estimate missing values using proportional relationships.

**Milestone 4: Structuring the Data Story**

**1. Skewness Calculation**

* Calculate skewness for numeric variables.
* Interpret skewness:
  + **Symmetric**: Close to 0.
  + **Positive Skew**: Tail is on the right.
  + **Negative Skew**: Tail is on the left.

**2. Histogram and Normality Assessment**

* Use histograms and boxplots to visually assess distributions.
* Identify normal and skewed variables.

**3. Data Transformation**

* **Log Transformation**: For highly positively skewed data.
* **Square Root Transformation**: For mild skew.
* **Box-Cox Transformation**: For complex distributions.

**A graph of a number of drugs

Description automatically generated**A graph showing a number of purple balls

Description automatically generatedA blue and red graph

Description automatically generatedA graph of smoking prevalence

Description automatically generatedA graph of smoking

Description automatically generated with medium confidence

A diagram of a drug experiment

Description automatically generated