

**Air University Islamabad**

**FACULTY COMPUTING & ARTIFICAL**

**INTELLENGENCE**

**Department of creative technology**

**Project Report**

**Probability & Statistics**

**Class: BSDS-3A**

**Submitted to**: Sir Tahir Sher

**Submitted by**

Abdullah Ayaz 232573

Abdul Wahab 232542

Due Date: 20th Dec , 2024

:

**Title:** Analysis of Smoking Prevalence and Drug Experimentation among Youth

**Goal of the Project:**

The goal of this project is to analyze smoking prevalence, drug experimentation, and peer influence among youth. The analysis aims to understand the trends over time, assess gender and socioeconomic factors, identify influential age groups, and uncover relationships between key variables. This study uses statistical measures such as mean, median, mode, standard deviation, and skewness to summarize and analyze the data effectively. By examining these trends and relationships, we can identify patterns that help target awareness campaigns and prevention strategies. This research provides essential information for policymakers, healthcare organizations, and educators to develop programs that address smoking and drug experimentation issues.

**Introduction**:

Smoking and drug experimentation are significant public health concerns worldwide, particularly among youth populations. Adolescence and early adulthood are critical periods during which individuals are most susceptible to external influences such as peer pressure, socioeconomic conditions, and accessibility to harmful substances. These behaviors not only affect immediate health outcomes but also have long-term implications on physical, mental, and emotional well-being.

This study explores smoking prevalence, drug experimentation, and peer influence using data sourced from Kaggle. The dataset comprises demographic information such as age, gender, and socioeconomic status, as well as measures of smoking and drug use prevalence. By employing statistical analysis and data visualization techniques, this study aims to:

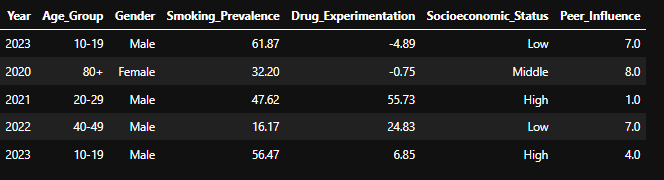
* Identify trends in smoking and drug experimentation over the years (2020-2024).
* Understand the influence of gender, age group, and socioeconomic status on smoking and drug use.
* Analyze peer influence and its correlation with smoking and drug experimentation.
* Provide actionable insights that can help reduce smoking and drug use among youth.

**Dataset Description** The dataset, sourced from Kaggle, provides a comprehensive overview of smoking and drug experimentation trends. It includes the following variables:

* **Year**: The year of observation, ranging from 2020 to 2024.
* **Age\_Group**: Different age groups such as 10-19, 20-29, 30-39, etc.
* **Gender**: Male, Female, and Both.
* **Smoking\_Prevalence**: The percentage of individuals who report smoking.
* **Drug\_Experimentation**: The percentage of individuals who report experimenting with drugs.
* **Socioeconomic\_Status**: Classification of individuals into High, Middle, and Low socioeconomic groups.
* **Peer\_Influence**: A numerical measure of the influence peers have on individuals.

We use this dataset as it captures critical variables necessary to explore youth behaviors comprehensively. By analyzing the relationships between these variables, we can uncover patterns and trends that inform interventions. The dataset is suitable for statistical analysis because it is well-organized, contains relevant information, and is clean for analysis.

We use this dataset as it captures critical variables necessary to explore youth behaviors comprehensively. By analyzing the relationships between these variables, we can uncover patterns and trends that inform interventions. The dataset is suitable for statistical analysis because it is well-organized, contains relevant information, and is clean for analysis.



**Statistical Measures:**

To understand the underlying trends in smoking prevalence, drug experimentation, and peer influence, various statistical measures were computed. These measures are used because they provide insights into the behavior of the data distribution, allowing us to summarize trends and variability in the data effectively.  
I have written some results that I find .If you like to see code than run notebook .I attached the link below.

**Measures of Central Tendency**

Central tendency measures such as mean, median, and mode are used to summarize the data:

**Mean**:

* Smoking Prevalence: 27.93
* Drug Experimentation: 33.86
* Peer Influence: 11.58

***Why Use Mean?*** The mean shows the average value, which helps us understand the general trend in the data.

**Median**:

* Smoking Prevalence: 27.76
* Drug Experimentation: 33.90
* Peer Influence: 12.10

***Why Use Median****?* The median is not affected by extreme values (outliers) and gives a better measure of central value when the data is skewed.

**Mode**:

* Smoking Prevalence: 36.88
* Drug Experimentation: 18.80
* Peer Influence: 15.10

***Why Use Mode****?* The mode shows the most frequently occurring values in the dataset.

### ****Measures of Dispersion****

Measures of dispersion such as range, variance, and standard deviation show how spread out the data is:

**Range**:

* Smoking Prevalence: 99.54
* Drug Experimentation: 142.37
* Peer Influence: 21.10

**Why Use Range**? Range helps show the difference between the highest and lowest values.

**Standard Deviation**:

* Smoking Prevalence: 18.26
* Drug Experimentation: 27.68
* Peer Influence: 5.41

**Why Use Standard Deviation?** It measures how spread out the data is around the mean.

### ****Skewness and Kurtosis****

**Skewness**:

* Smoking Prevalence: 0.038 (approximately symmetrical)
* Drug Experimentation: -0.031 (slightly negatively skewed)
* Peer Influence: -0.102 (slightly negatively skewed)

**Kurtosis**:

* Smoking Prevalence: -0.469 (platykurtic)
* Drug Experimentation: -0.504 (platykurtic)
* Peer Influence: -0.767 (platykurtic)

**Grouped Data Analysis**

The data was grouped into intervals to create **frequency tables**, making it easier to interpret patterns within the ranges of smoking, drug use, and peer influence.

1. **Gender Analysis**:

* Smoking Prevalence: Males (68.11%) have higher smoking rates than females (31.89%).
* Drug Experimentation: Males (67.48%) have higher drug use rates compared to females (32.52%).

1. **Age Group Analysis**:

* The age group 20-29 shows the highest peer influence and smoking prevalence.

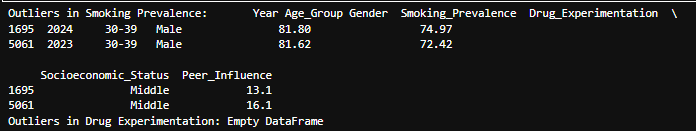
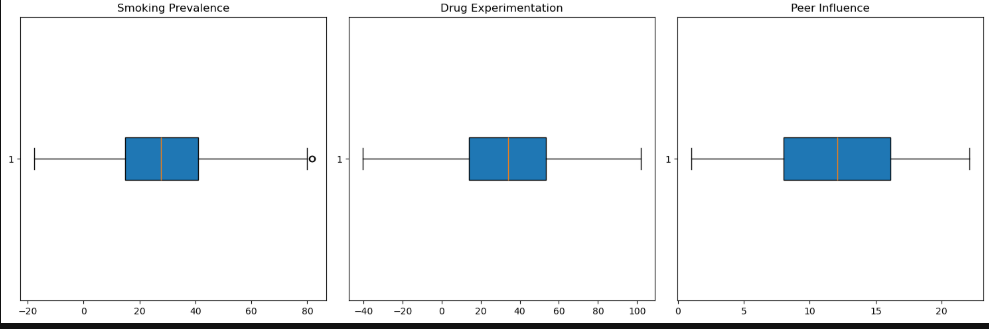
1. **Socioeconomic Status**:

* Smoking is highest in the middle class (34.1%).
* Drug experimentation is highest in the high socioeconomic class (46.0%).

**Outliers**

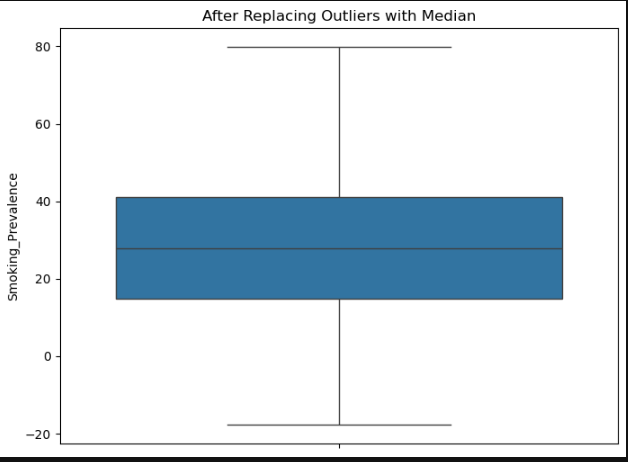
**1- Outlier Detection**

Outliers were detected in the **Smoking Prevalence** column using statistical methods. Outliers are values that deviate significantly from the general pattern of the data. These extreme values can affect the overall analysis, particularly measures such as the mean.



**2- Outlier Replacement**

To maintain the integrity of the dataset, the outliers in the Smoking Prevalence column were replaced with the median value of Smoking Prevalence. This approach ensures that the overall distribution of the data is not heavily skewed by extreme values.

**General Findings and Insights**

### ****3-Why Replace Outliers with Median?****

The **median** is a robust measure of central tendency that is not affected by extreme values. Replacing outliers with the median:

* Prevents the mean and standard deviation from being distorted.
* Ensures that the data analysis remains reliable and accurate.

**1-Smoking and Drug Trends**:

* Smoking prevalence and drug experimentation percentages are higher among **males** compared to females.
* Smoking and drug experimentation vary across years, with significant peaks observed in **2021** for females and **2023** for males.

**2-Age Group Influence**:

* The **20-29 age group** shows the highest Peer Influence, particularly among males.
* Drug experimentation and smoking percentages are also higher in younger age groups.

**3-Economic Class Analysis**:

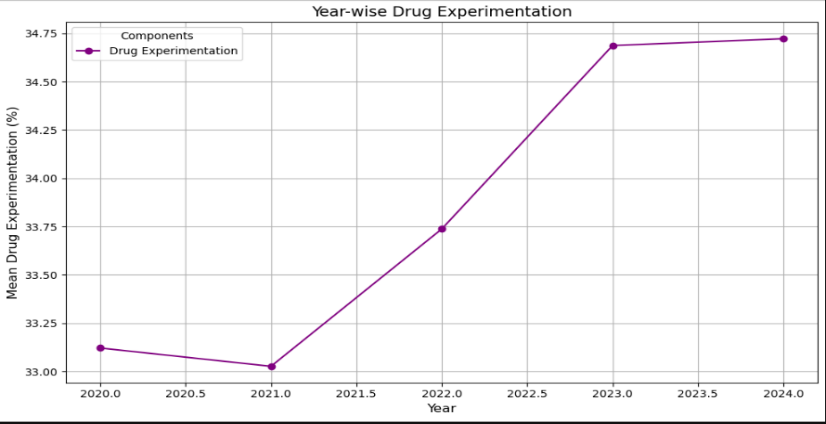
* Smoking prevalence is highest in the **Middle Class** with a ratio of **34.1%**.
* Drug experimentation is most common in the **High Class** with a ratio of **46.0%**

**4-Gender Comparison**:

* Males dominate smoking and drug experimentation trends across all years and age groups.
* Females show lower percentages but still highlight significant variations.

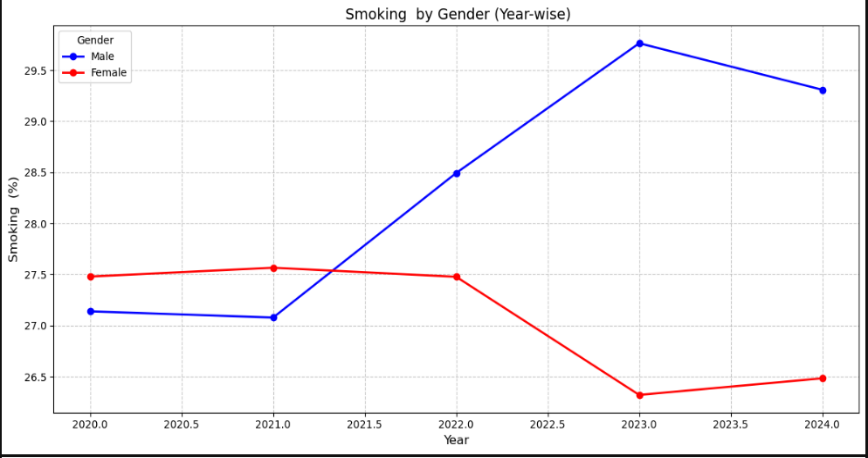
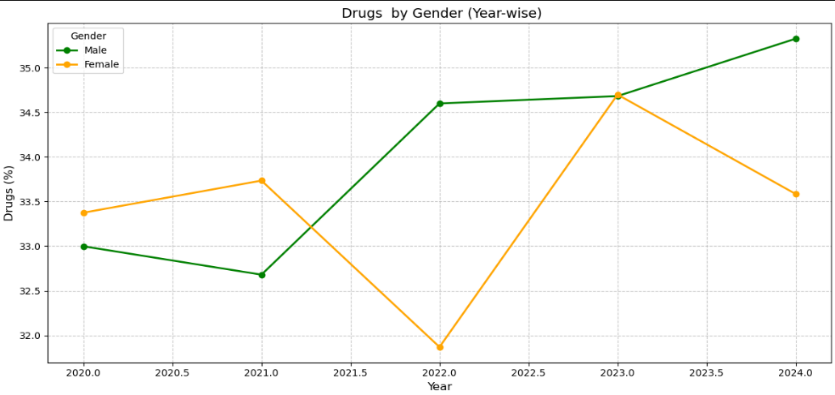
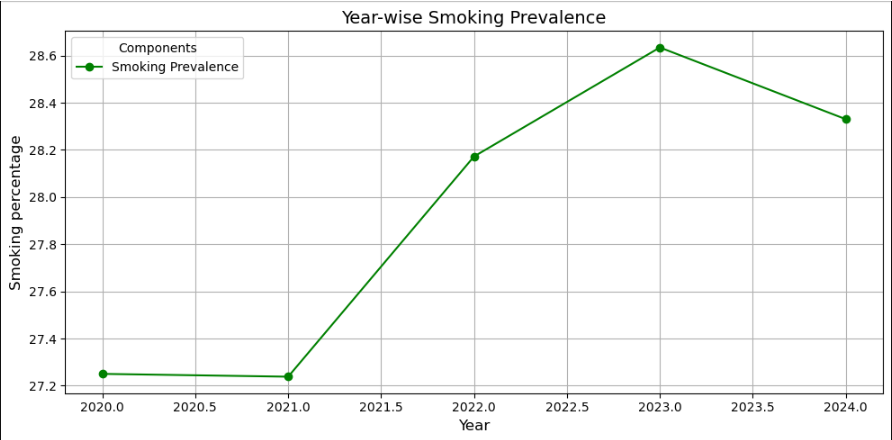
**Visualizations**

**1-Line Graphs**:

* Year-wise Drug Experimentation
* Year-wise Smoking Prevalence
* Smoking Prevalence by Gender (Year-wise)

The graph shows an **overall increase** in drug experimentation among youth from 2020 to 2024.

The level of drug experimentation appears to be **relatively low in 2020 and 2021**, with a noticeable increase starting from 2022.

**2- Bar Charts:**

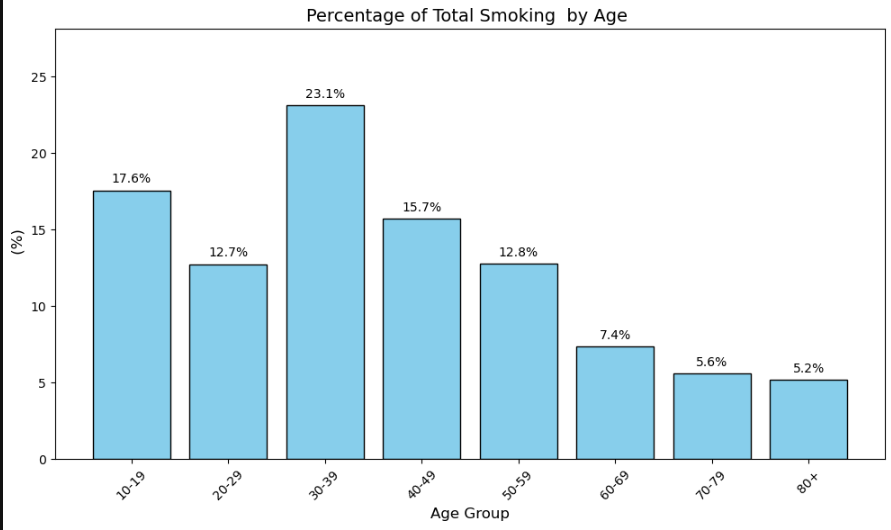
The graph shows drug experimentation rates by gender from **2020 to 2024.** Male drug experimentation rates were consistently higher than female rates throughout the period, with both showing an overall increasing trend.

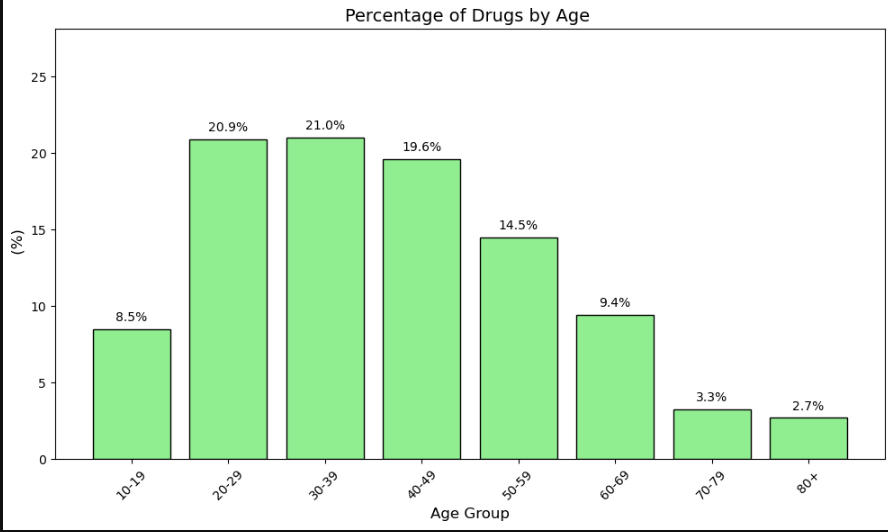
The graph shows an **overall increase** in smoking prevalence among youth from 2020 to 2023, followed by a slight decrease in 2024.

The level of smoking prevalence appears to be **relatively low in 2020 and 2021**, with a noticeable increase starting from 2022.

The graph illustrates smoking prevalence by gender **from 2020 to 2024**. Male smoking rates were consistently higher than female rates throughout the period. Both genders showed an overall increasing trend in smoking prevalence until 2023, followed by a slight **decrease in 2024**.

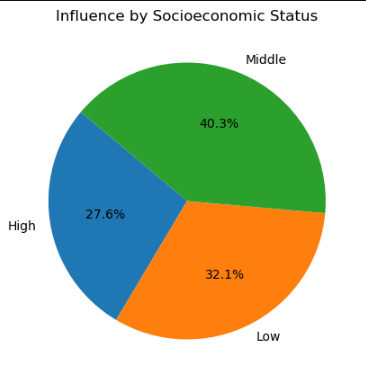
* Smoking and Drug Percentage by Age Group
* Peer Influence by Age Group

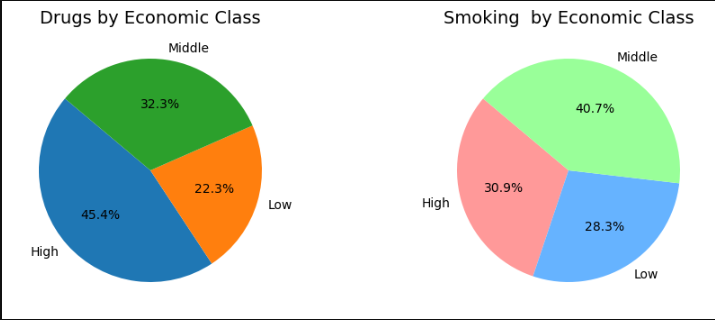




3- **Pie Charts**:

* Smoking and Drug Experimentation by Socioeconomic Status
* Peer Influence by Socioeconomic Status





The pie charts show the prevalence of drug use and smoking across different economic classes.  
 The "High" economic class has the highest percentage of drug use **(45.4%),**   
 The "Middle" class has the highest percentage of smoking **(40.7%).**

chart shows that peer influence is highest among individuals in the middle socioeconomic class (40.3%), followed by the low (32.1%) and high (27.6%) classes.

**Conclusion**

The analysis shows that:

* Males have higher smoking and drug experimentation rates compared to females.
* The age group 20-29 is most influenced by peer pressure and has the highest smoking prevalence.
* Socioeconomic status impacts smoking and drug experimentation rates, with the middle class having higher smoking prevalence and the high class showing higher drug experimentation.

This project demonstrates how basic statistical methods can help analyze real-world data to extract meaningful insights. By applying measures of central tendency, dispersion, and outlier treatment, we ensured the accuracy and reliability of results.

The visualizations simplified the understanding of trends in smoking, drug use, and peer influence, providing a clearer picture of the impact across different demographic groups. Such analysis is critical for identifying areas where intervention is needed to address public health concerns effectively.