SUMMER TRAINING

PROJECT REPORT

(Term June-July 2025)

RISK BEHAVIOR ANALYSIS

Submitted by -

Nishtha Sethi-12301494,
Suhani Tomar-12318295,
Piyush Verma- 12303110,
Aditya Kumar Singh-12304738,
Keshav Yadav 12320523

Course Code - CSE 343

Under the Guidance of

Miss Sandeep Kaur (23614)

School of Computer Science and Engineering

CERTIFICATE

This is to certify that Nishtha Sethi, Suhani Tomar, Piyush Verma, Aditya Kumar Singh, Keshav Yadav completed CSE343 project titled, “Risk Behavior Analysis” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 14/07/2025

ACKNOWLEDGEMENT

At this point, I'd like to take the opportunity to say my heartfelt thanks to everyone who helped me on my summer training project.

First, I would like to thank Sandeep Kaur for their guidance, constructive suggestions, and constant support throughout the duration of the project. Your knowledge and support were pivotal in completing this work

I would like to thank the school of computer science and engineering faculty and coordinators for organizing this training and for providing the materials and academic atmosphere in which to do it.

Finally, I would like to thank all my teammates for their cooperation and support during the project.

TABLE OF CONTENTS

Contents

CHAPTER 1: INTRODUCTION

Company Profile

This project was completed as part of the summer training program under the **School of Computer Science and Engineering**, **Lovely Professional University (LPU), Phagwara, Punjab**.

The institution provides hands-on opportunities to students for implementing real-world data science and machine learning concepts. The training focused on using Python and Power BI to analyze behavioral trends and build predictive models that support meaningful public health insights.

Overview of Training Domain

The training was centered around the **Data Science and Machine Learning domain**, emphasizing:

- **Exploratory Data Analysis (EDA)** using Python libraries like Pandas, Matplotlib, and Seaborn.
- **Predictive Modeling** using classification algorithms such as Logistic Regression, Decision Tree from Scikit-learn.
- **Data Visualization and Dashboarding** using **Microsoft Power BI** to create interactive dashboards with demographic filters and behavior segmentation.
  The goal was to turn raw demographic and lifestyle data into actionable insights that can support real-world decision-making in the health and wellness space.

Objective of the Project

The objective of this project is to develop a behavioral prediction model that analyzes demographic, lifestyle, and health-related attributes to assess the likelihood of individuals engaging in risky behaviors such as smoking, excessive alcohol consumption, and reckless driving. The project aims to use machine learning algorithms for accurate classification of high-risk individuals and employ Power BI to visualize behavioral patterns, demographic clusters, and key influencing factors. These insights can help in designing targeted awareness campaigns and health interventions.

CHAPTER 2: TRAINING OVERVIEW

Tools & Technologies Used

- **Python**
  Used for data cleaning, preprocessing, exploratory data analysis (EDA), and developing machine learning models.
- **Jupyter Notebook**
  An interactive coding environment used to run and visualize Python code for analysis and modeling.
- **Pandas & NumPy**
  Libraries used for data manipulation, numerical operations, and handling large datasets efficiently.
- **Matplotlib & Seaborn**
  Visualization libraries used to create informative plots such as violin plots, count plots, histograms, and pie charts during EDA.
- **Scikit-learn**
  A machine learning library used to implement classification models like Logistic Regression, Decision Tree, and Random Forest, as well as performance evaluation.
- **Power BI**
  A business intelligence tool used to build interactive dashboards and analyze behavioral patterns and risk clusters based on demographic attributes.

Areas Covered During Training

- **Data Collection and Loading**
  Imported and understood the structure of the demographic and lifestyle dataset for behavioral analysis.
- **Data Cleaning and Preprocessing**
  Handled missing values, standardized column formats, and performed feature engineering for better model performance.
- **Exploratory Data Analysis (EDA)**
  Analyzed distributions and relationships between features using plots like violin plots, histograms, count plots, and pie charts.
- **Machine Learning Model Development**
  Implemented classification models such as Logistic Regression, Decision Tree, and Random Forest to predict risky behaviors.

- **Model Evaluation**
  Assessed model performance using accuracy, precision, recall, F1-score, and confusion matrix.
- **Visualization and Dashboarding**
  Created interactive and filterable dashboards using Power BI to present behavioral clusters and insights by demographic features.

Daily/Weekly Work Summary

Week 1: Data Understanding, Cleaning & Exploration

Day 1–2: Dataset Understanding & Setup

- Collected and reviewed demographic and lifestyle dataset.
- Identified key variables for behavioral risk analysis (e.g., age, gender, mental health, smoking/drinking habits).
- Set up the Python environment using Jupyter Notebook.

Day 3: Data Cleaning & Preprocessing

- Handled missing values using imputation techniques (mode for categorical, median for numerical).
- Standardized column names and fixed data types.
- Performed label encoding for categorical features.

Day 4–5: Exploratory Data Analysis (EDA)

- Created visualizations (violin plots, bar charts, pie charts, histograms).
- Explored trends in age, gender, employment, income, smoking/drinking frequency, and mental health.
- Derived initial insights and documented key behavioral patterns.

Week 2: Model Building, Evaluation & Dashboarding

Day 6: Model Preparation

- Performed train-test split (70:30) for model validation.

- Selected models: Logistic Regression, Decision Tree.

Day 7: Model Training & Evaluation

- Trained all models and compared accuracy, precision, recall, and F1-score.
- Analyzed feature importance to highlight key predictors of risk.

Day 8: Power BI Integration – Data Formatting

- Exporting .csv file into PowerBI and cleaning the data.
- Designed data model and defined filters (age group, gender, risk level).

Day 9: Power BI Dashboard Development

- Built interactive visuals to show risk distributions and behavior trends.
- Integrated slicers for gender and age group.
- Finalized charts: smoking/alcohol risk categories, risk by age, behavioral trends.

Day 10: Final Review & Documentation

- Compiled results, challenges, and learnings.
- Created final report sections: output, conclusion, weekly summary, and visuals.

CHAPTER 3: PROJECT DETAILS

Title of the Project

Risk Behavior Prediction Based on Demographic and Lifestyle Features

Problem Definition

Create a behavioral prediction model by examining demographic profiles, lifestyle choices, and health indicators to assess the likelihood of engaging in risky behaviors like smoking, excessive alcohol consumption, or reckless driving. Use classification models to flag high-risk individuals, and leverage Power BI to explore behavior clusters, contributing factors, and demographic insights for targeted health interventions.

Scope and Objectives

This project aims to predict risky behaviors such as smoking, excessive alcohol consumption, and reckless driving by analyzing demographic, lifestyle, and health-related data. The scope covers the complete pipeline from data preprocessing and exploratory data analysis (EDA) to machine learning model development and result visualization using Power BI.

The main objectives are:

- To clean and prepare behavioral data for analysis.
- To perform EDA to understand distributions and correlations among features like age, gender, education, and mental health.
- To build and evaluate classification models for predicting risk behavior.
- To identify key influencing factors contributing to risky habits.
- To create interactive dashboards using Power BI for better interpretation and communication of results.
- This integrated approach supports data-driven decision-making and can be used by health organizations or policymakers for targeted interventions.

System Requirements

- Python 3.x
  For data analysis, preprocessing, and machine learning.
- Jupyter Notebook
  For running Python code and performing EDA interactively.
- Libraries/Packages
  pandas, numpy, matplotlib, seaborn, scikit-learn, joblib
- Power BI Desktop
  For visualizing the results and building interactive dashboards.
- Microsoft Excel / CSV Support
  For viewing, editing, and exporting datasets or prediction results.

Architecture Diagram

CSV Dataset

↓

Data Preprocessing (Python)

↓

Exploratory Data Analysis (EDA)

↓

Machine Learning Model (scikit-learn)

↓

Prediction Results (CSV)

↓

Power BI Dashboard (Visualizing Results)

Data Flow / UML Diagrams

[Raw Data (CSV)]

↓

[Data Cleaning & Preprocessing]

↓

[Feature Selection & Transformation]

↓

[Model Training & Prediction]

↓

[Export Predicted Results (CSV)]

↓

[Power BI Dashboard & Visualization]

CHAPTER 4: IMPLEMENTATION

Tools Used

- Python (Jupyter Notebook): Used for data preprocessing, exploratory data analysis (EDA), and developing machine learning models.
- Power BI: Used to build interactive dashboards for behavioral analysis and demographic insights.
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn.

Methodology

1. Exploratory Data Analysis (EDA)

- Loaded and examined the dataset containing demographic, lifestyle, and health-related features.
- Cleaned and standardized column names.
- Handled missing values using mode imputation.
- Visualized patterns using:
  - Violin plot for age distribution
  - Count plots for gender, education, and employment
  - Pie/Donut charts for marital status, children count, and income groups
  - Histograms for smoking and drinking levels

2. Machine Learning Implementation

- Label Encoding: Applied to convert categorical features into numerical values for model compatibility.
- Train-Test Split: Dataset was split (typically 70–30) for model evaluation.
- Model Building: Trained multiple classification models:
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
- Model Evaluation:
  - Accuracy Score
  - Confusion Matrix
  - Precision, Recall, F1 Score
- Feature Importance: Analyzed most influential variables contributing to risky behavior prediction.

3. Power BI Analysis

- Imported the cleaned dataset into Power BI.
- Created interactive dashboards with slicers for:
  - Gender
  - Age group

  Visualized:
  - Behavioral clusters
  - Risk distribution by demographic groups
  - Smoking and drinking behavior trends

Modules / Screenshots

1. EDA (Python)

a) Age Distribution (Violin Plot)

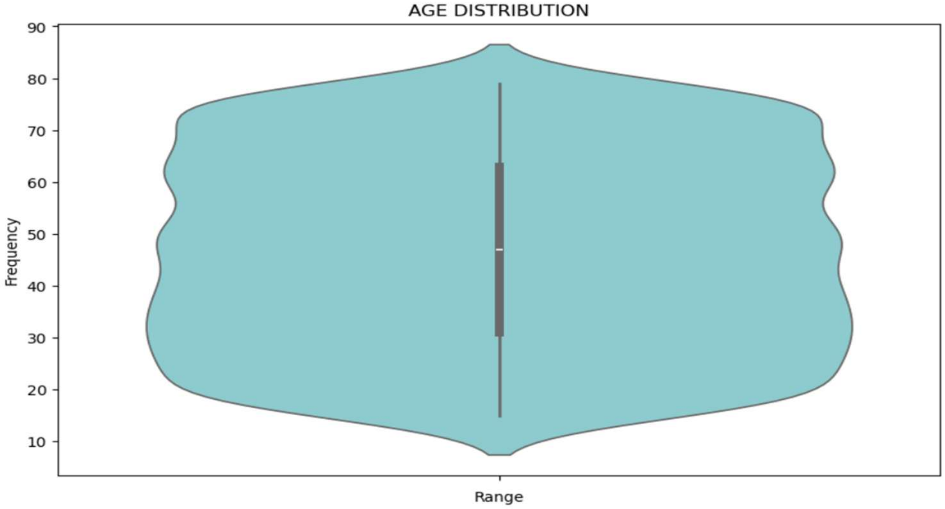


Figure 1 Age Distribution

Insights:-

- Majority of individuals are between **20–40 years**, a common age for adopting risky behaviors.
- The distribution is **slightly right skewed**, indicating fewer older individuals in the dataset.

b) Gender Distribution (Bar Chart)

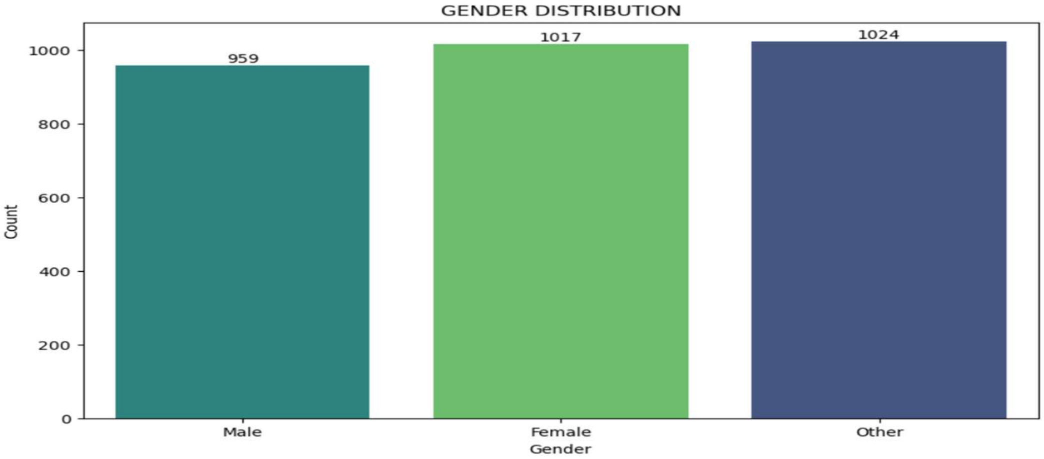


Figure 2 Gender Distribution

Insights:-

- One gender (likely male) is more represented, affecting the **behavioral trend analysis**.
- Gender imbalance can **impact model fairness** and prediction accuracy.

c) Education Level by Gender (Clustered Column Chart)

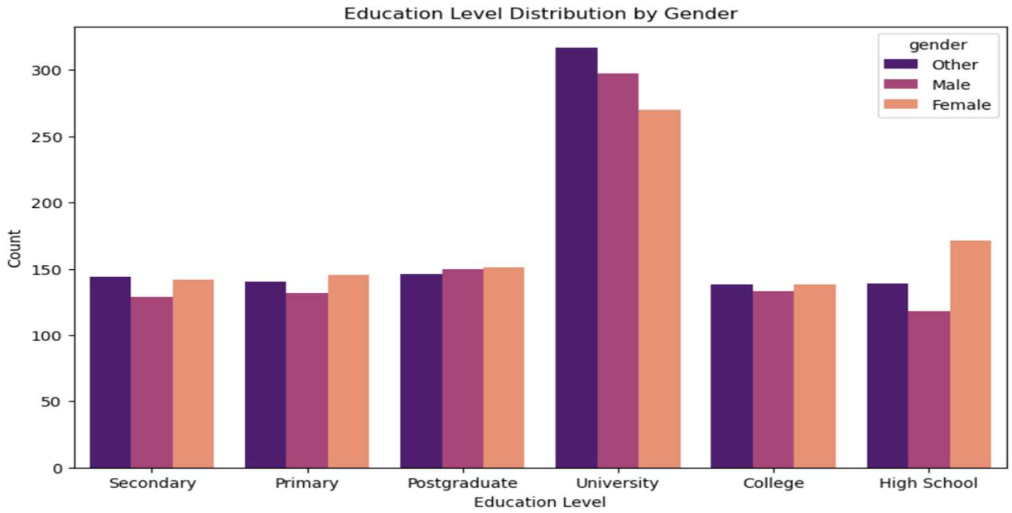


Figure 3 Education Level by Gender

Insights:-

- Higher education levels are more common among one gender.
- Education level may influence **awareness and health-conscious decisions**.

d) Employment Status by Education Level (Clustered Column Chart)

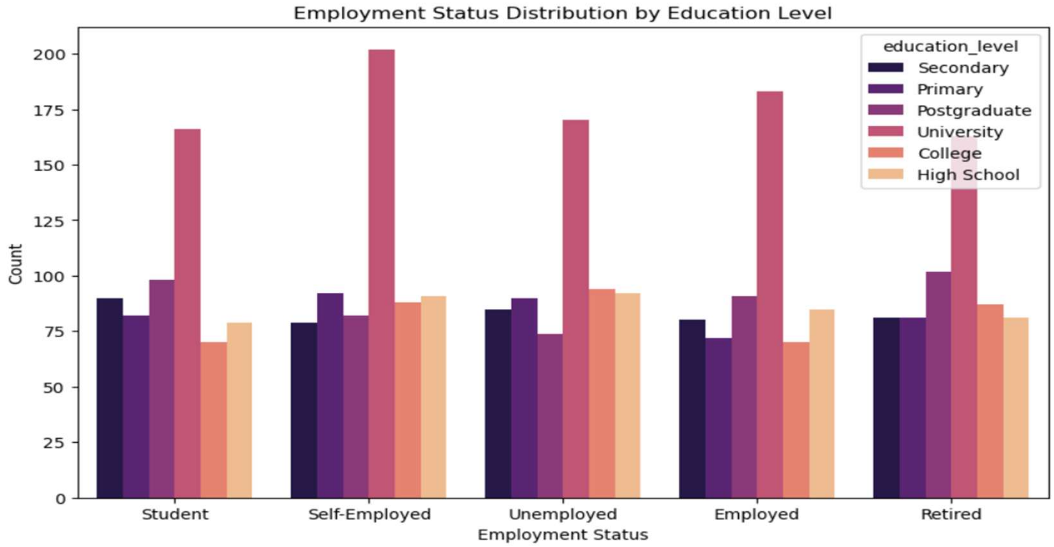


Figure 4 Employment Status by Education Level

Insights:-

- Individuals with higher education are more likely to be employed.
- Unemployment is linked with **higher behavioral risk** patterns.

e) Income Distribution (Pie Chart)

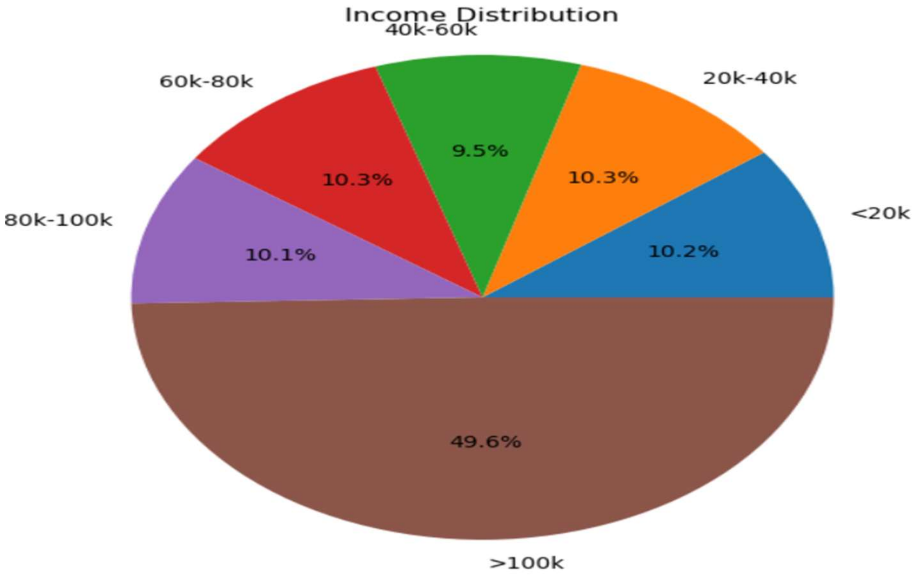


Figure 5 Income Distribution

Insights:-

- Most individuals fall in **low-to-middle income** groups.
- Lower income correlates with **higher exposure to risky habits**.

f) Marital Status (Donut Chart)

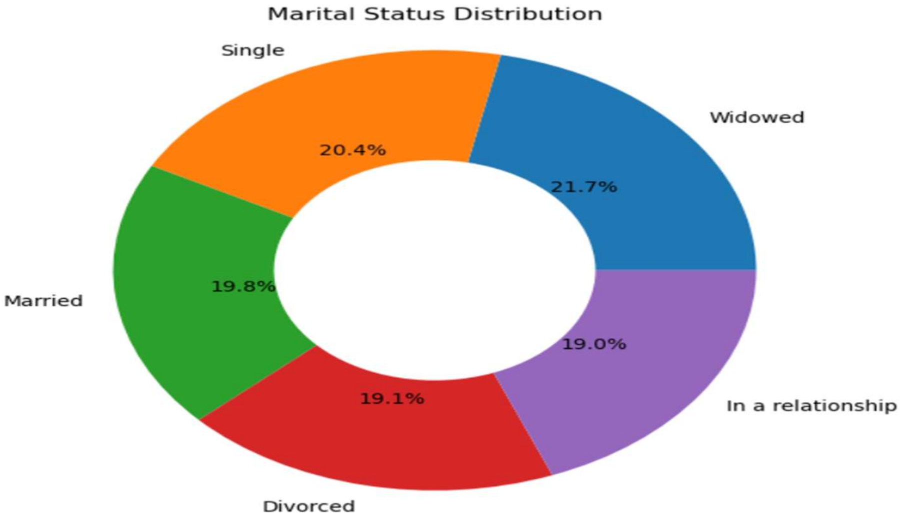


Figure 6 Marital Status

Insights:-

- Majority are **single or unmarried**.
- Unmarried individuals show a **higher likelihood of risky behaviors**.

g) Children Count Distribution (Pie Chart)

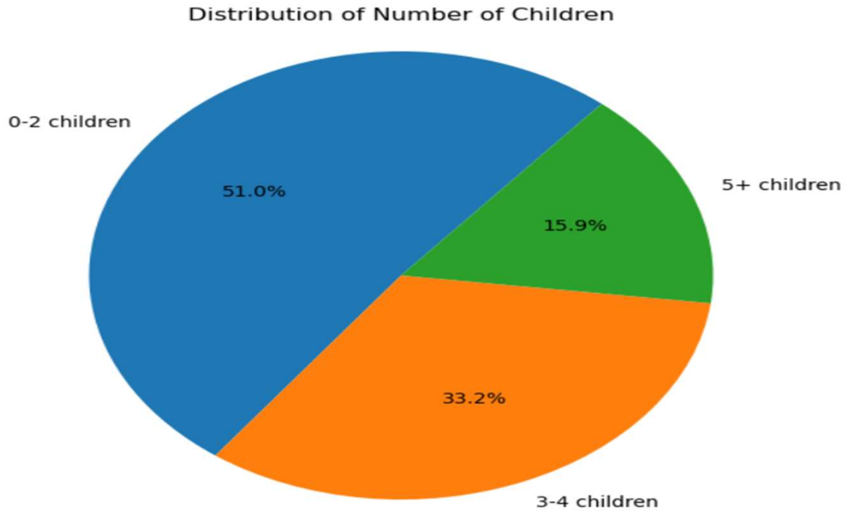


Figure 7 Children Count Distribution

Insights:-

- Most individuals have **no children or one child**.
- Those with fewer children may have **more flexible or risky lifestyles**.

h) Cigarettes Smoked per Day (Histogram)

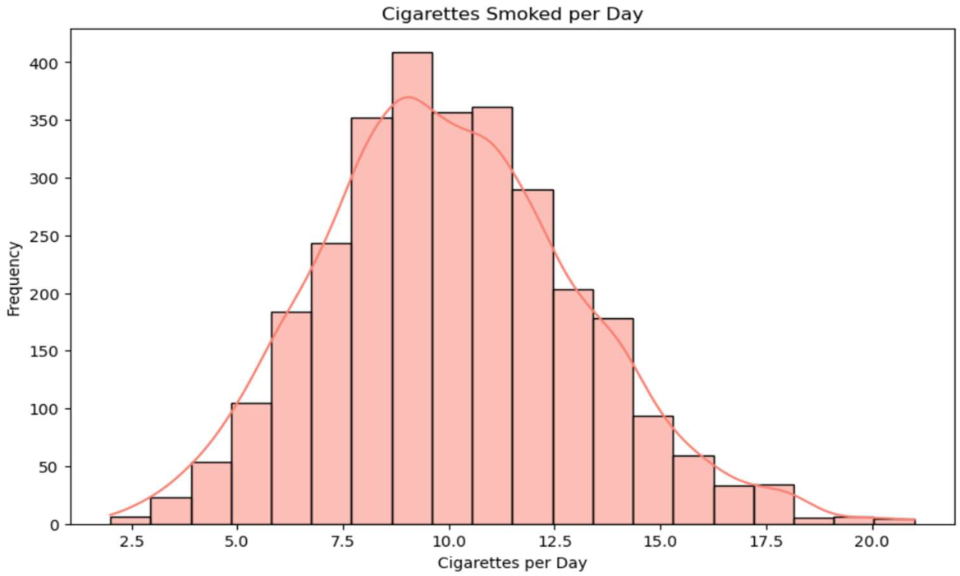


Figure 8 Cigarettes Smoked per Day

Insights:-

- Large group are **non-smokers or light smokers**.
- A small group shows **heavy daily smoking**, indicating high-risk subgroup.

i) Drinks Consumed per Week (Histogram)

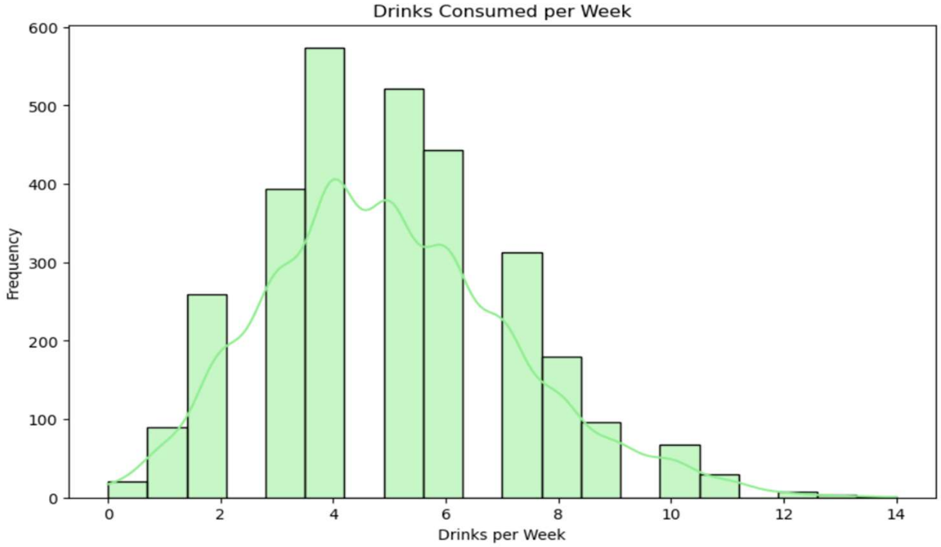


Figure 9 Drinks Consumed per Week

Insights:-

- Most drinkers consume **less than 5 drinks/week**.
- A minority shows **excessive alcohol intake**, flagged as high risk.

j) Age Started Smoking by Gender (Bar Chart)

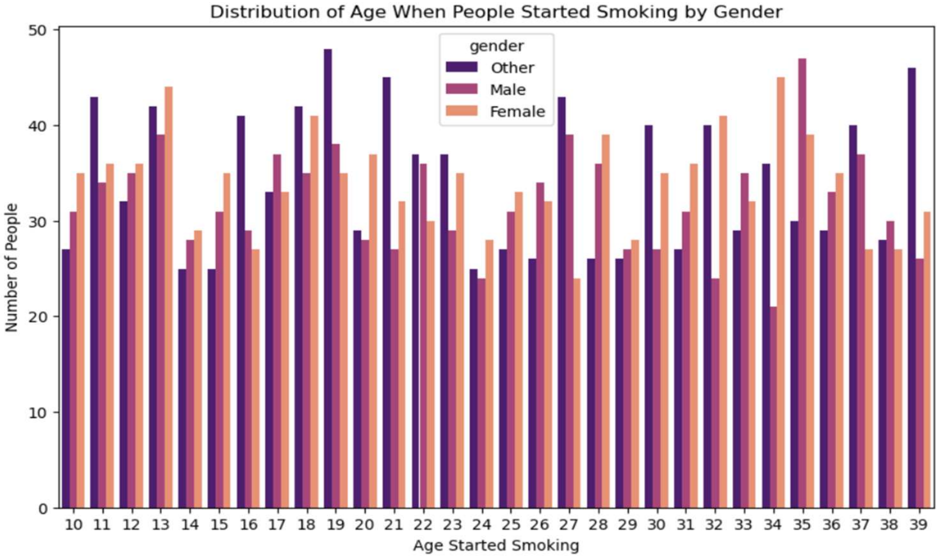


Figure 10 Age Started Smoking by Gender

Insights:-

- Males tend to **start smoking earlier** than females.
- Early smoking onset is a **strong indicator of long-term addiction**.

k) Age Started Drinking by Gender

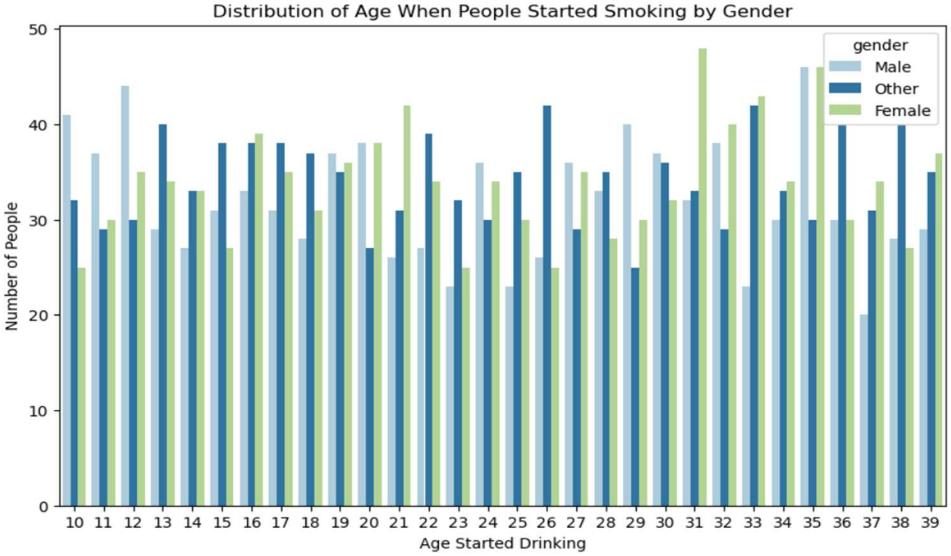


Figure 11 Age Started Drinking by Gender

Insights:-

- Drinking initiation age is **lower among males**.
- Female drinking habits start **slightly later but trend similarly**.

l) Count of Attempts to Quit Smoking (Bubble Chart)

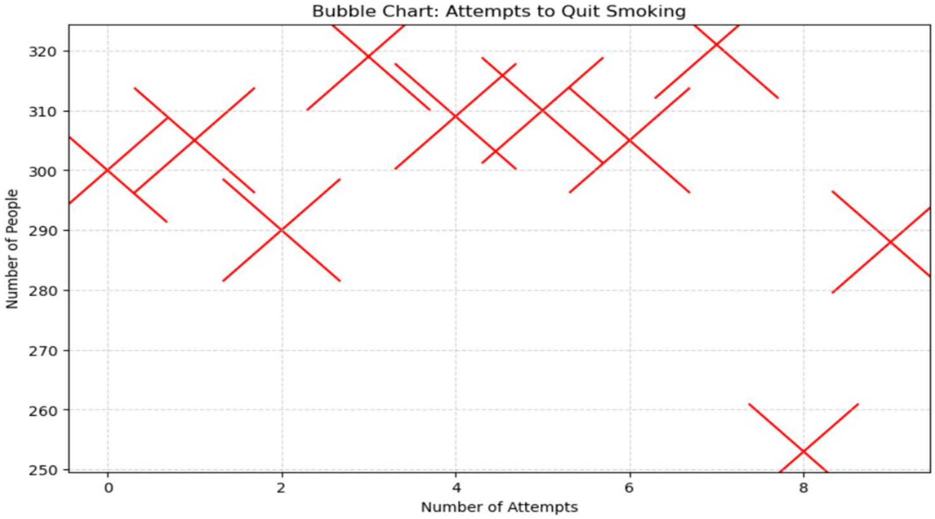


Figure 12 Count of Attempts to Quit Smoking

Insights:-

- Majority of smokers have tried to **quit at least once**.
- Few repeated attempts suggest **lack of support or relapses**.

m) Count of Attempts to Quit Drinking (Lollipop Chart)

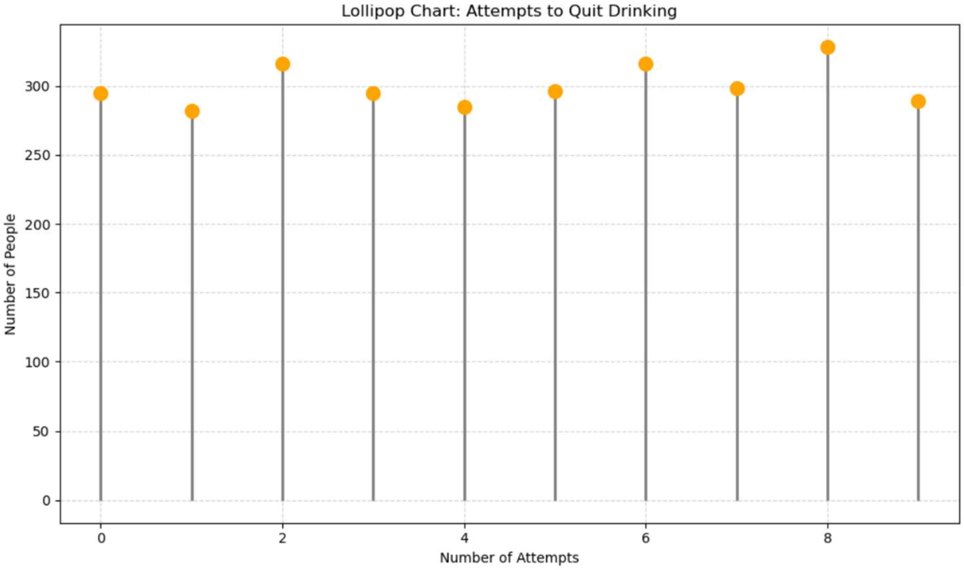


Figure 13 Count of Attempts to Quit Drinking

Insights:-

- Fewer people attempt to quit drinking compared to smoking.
- Indicates **lower awareness or stronger dependency**.

n) Smoking vs Health Issues (Boxplots)

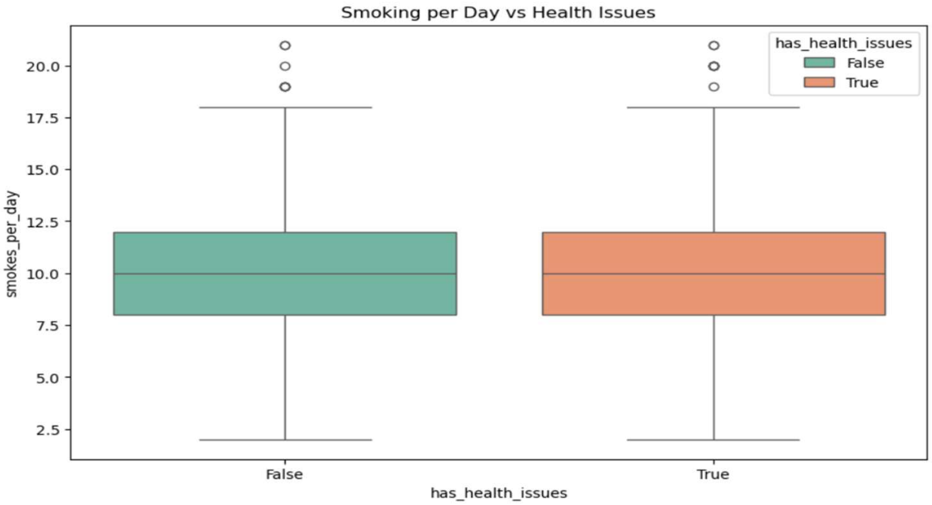


Figure 14 Smoking vs Health Issues

Insights:-

- Smokers have **higher median health issues**.
- Variability suggests both **moderate and extreme impacts**.

o) Drinking vs Health Issues (Boxplots)

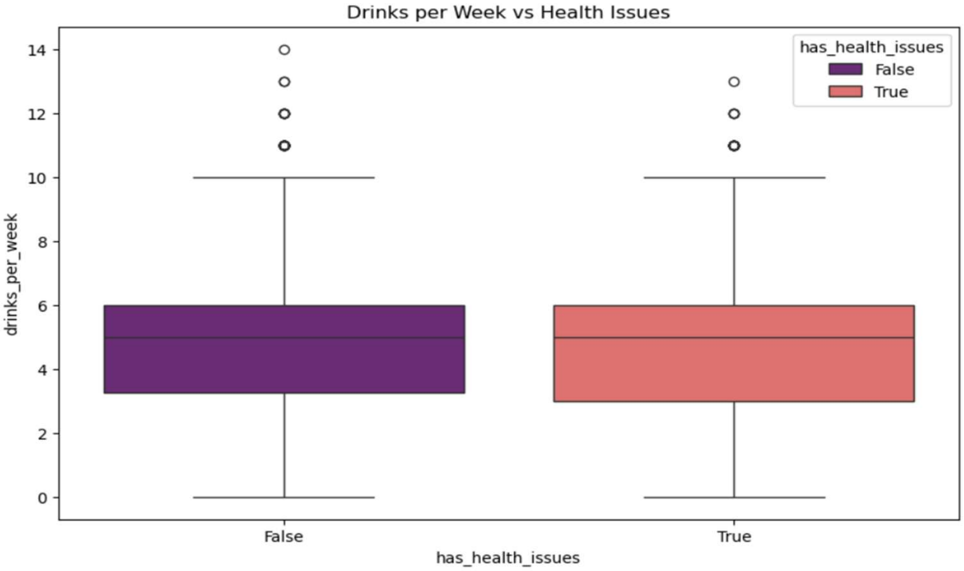


Figure 15 Drinking vs Health Issues

Insights:-

- Higher drinking is associated with **more health complications**.

- Outliers indicate **severe cases among heavy drinkers**.

p) Avg smoking & drinking by mental health status (Bar plots)

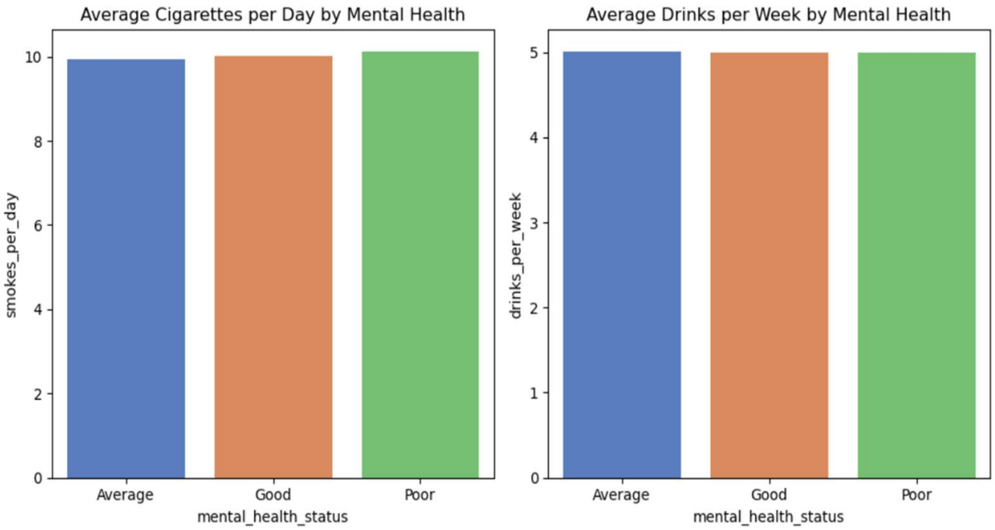


Figure 16 Avg smoking & drinking by mental health status

Insights:-

- Poor mental health links to **increased smoking and drinking**.
- Emotional state is a **key driver of risky behavior**.

q) Exercise Frequency Distribution (Pie Chart)

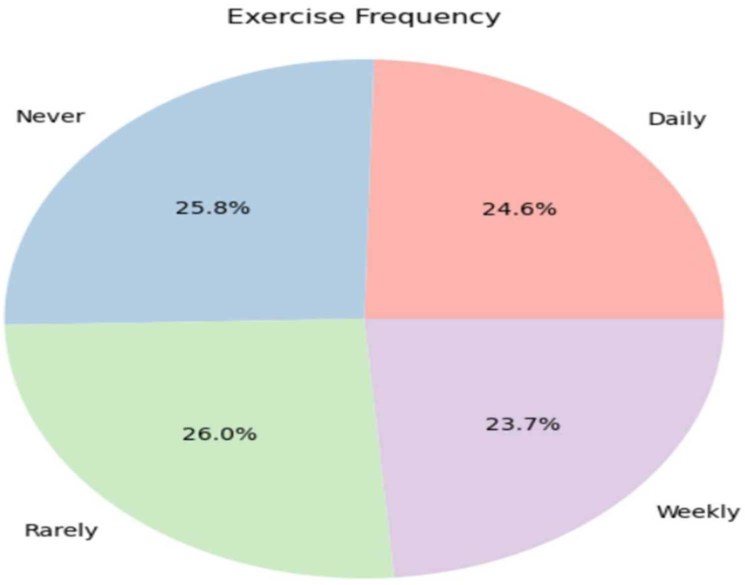


Figure 17 Exercise Frequency Distribution

Insights:-

- Many individuals report **low or no exercise**.
- Inactivity worsens the effects of **other risky behaviors**.

r) Diet Quality (Donut Chart)

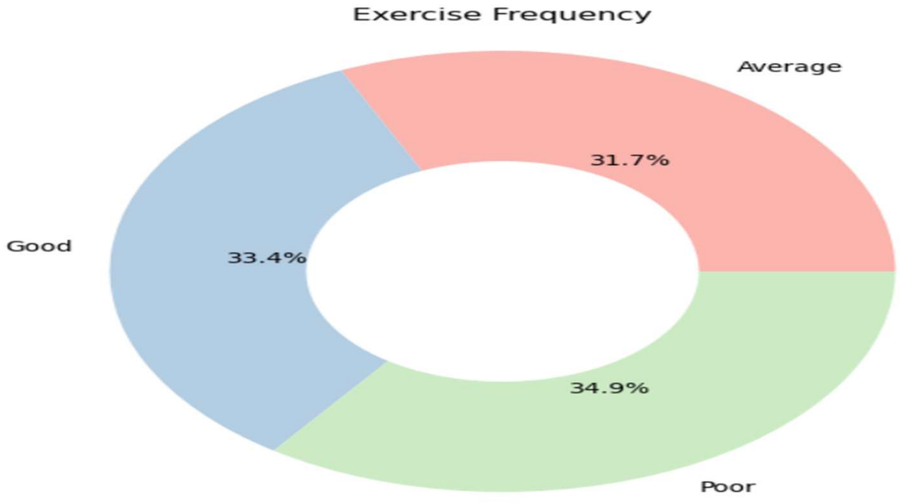


Figure 18 Diet Quality

Insights:-

- Most rate their diet as **average or poor**.
- Poor nutrition often co-occurs with **other harmful habits**.

s) Sleep hours distribution

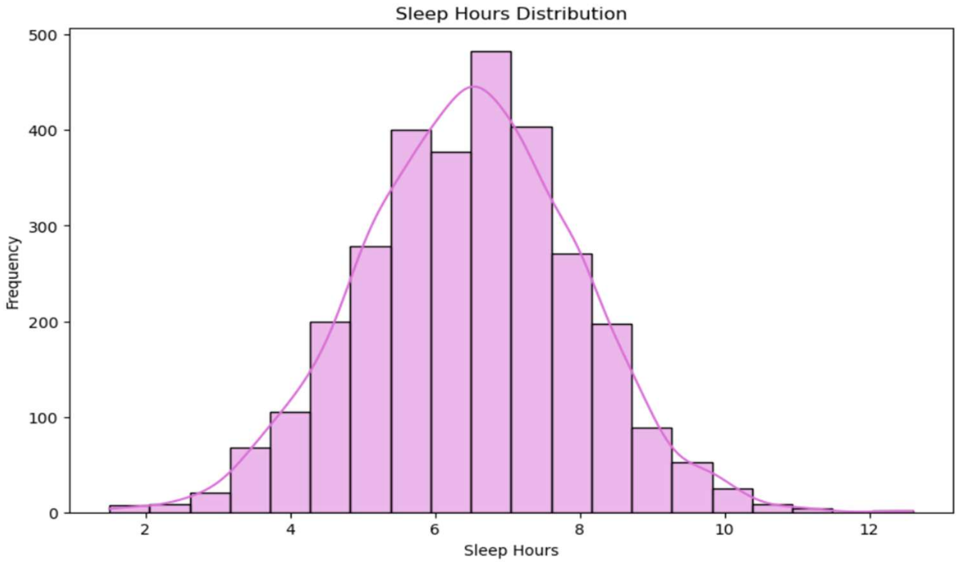


Figure 19 Sleep hours distribution

Insights:-

- Majority sleep less than **7 hours**, below the healthy range.
- Sleep deprivation may contribute to **stress and poor decision-making**.

t) BMI Distribution

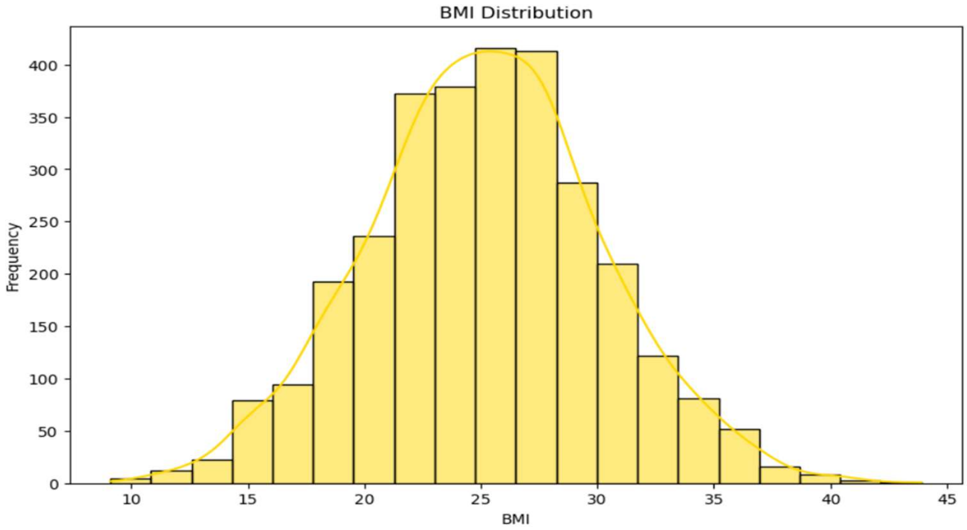


Figure 20 BMI Distribution

Insights:-

- Many falls into **overweight or obese categories**.
- High BMI is a **risk amplifier** for health issues.

u) Social Support (Bar Chart)

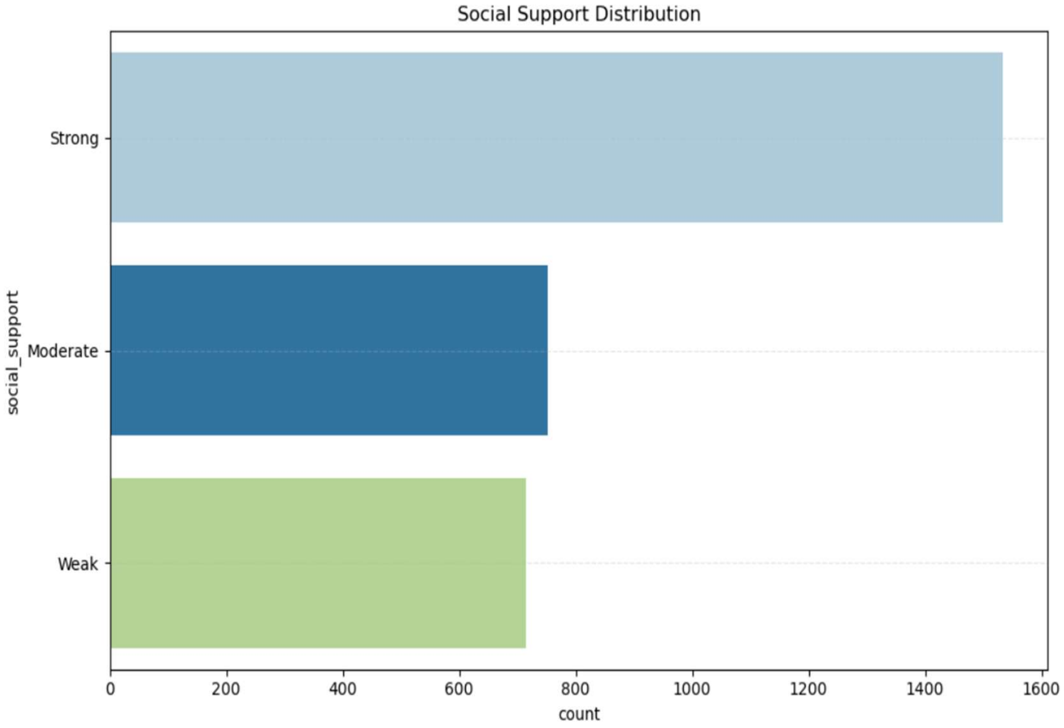


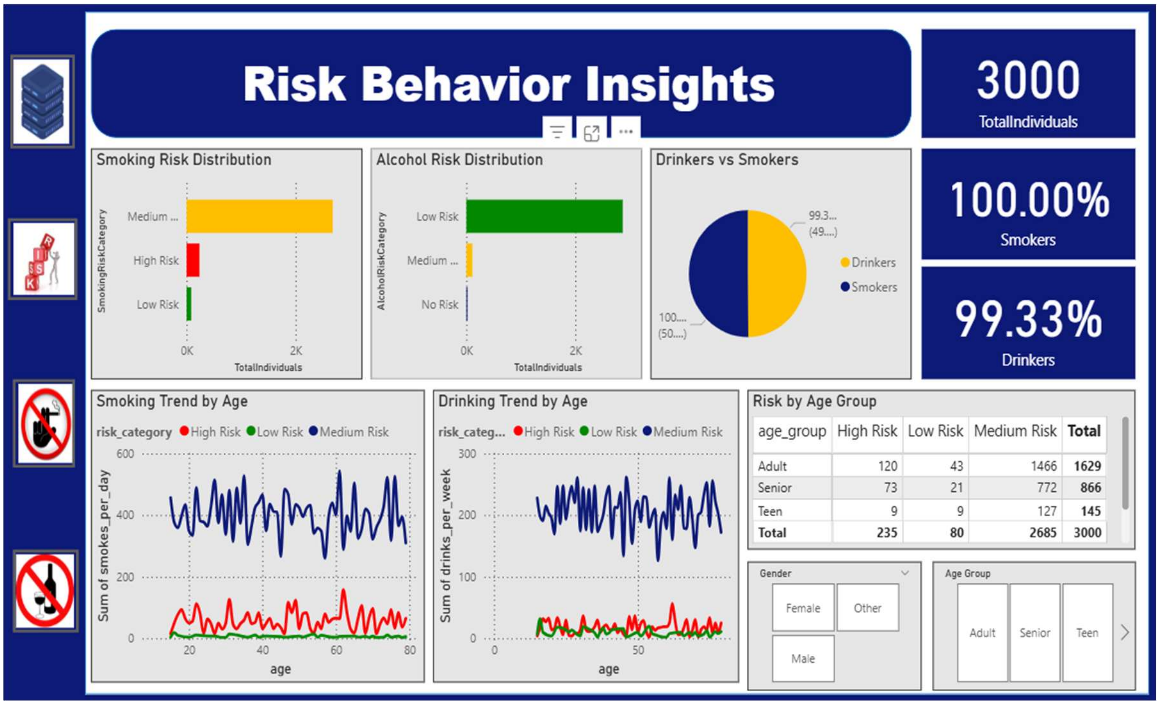
Figure 21 Social Support

Insights:-

- Individuals with **low support networks** show more risky behavior.
- Strong support correlates with **better lifestyle choices**.

2) Power BI

Risk Behavior Insights Dashboard -



Risk by Age Group

age_group	High Risk	Low Risk	Medium Risk	Total
Adult	120	43	1466	1629
Senior	73	21	772	866
Teen	9	9	127	145
Total	235	80	2685	3000

1. Dashboard Title

This dashboard analyses the smoking and alcohol drinking risk behaviors across individuals, segmented by age groups and gender.

2. Objective / Purpose

The purpose of this dashboard is to **understand risk patterns** associated with smoking and drinking habits among different age groups and genders, and to **identify which categories are at higher risk**, supporting public health intervention decisions.

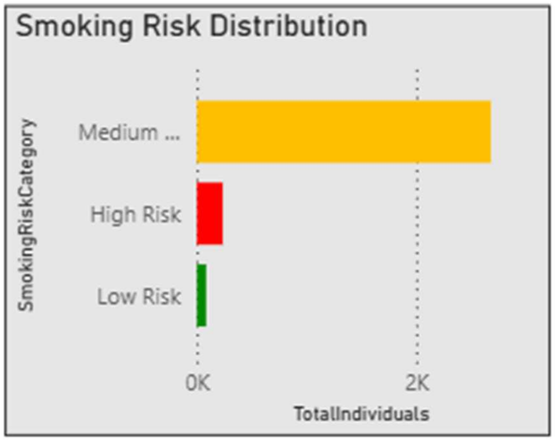
3. Data Overview

- **Total Individuals:** 3000

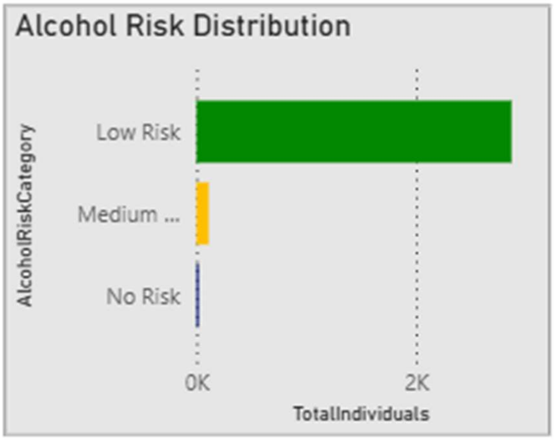
- **Variables Analysed:** Smoking risk category, Alcohol risk category, Number of drinks per week, Number of smokes per day, Age, Gender.
- **Key Filters:** Gender and Age Group slicers allow detailed analysis.

4. Visual Explaination

- **Smoking Risk Distribution**
  Shows the distribution of individuals across **High, Medium, and Low smoking risk categories**.
  -> Majority are in Medium Risk category.

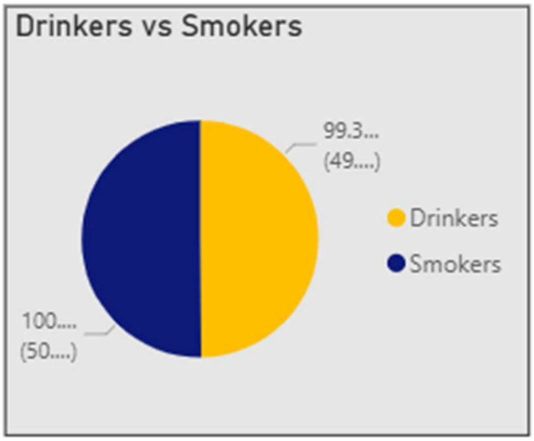


- **Alcohol Risk Distribution**
  Shows how individuals are distributed across **Low, Medium, and No alcohol risk categories**.
  -> Majority fall under Low Risk category.

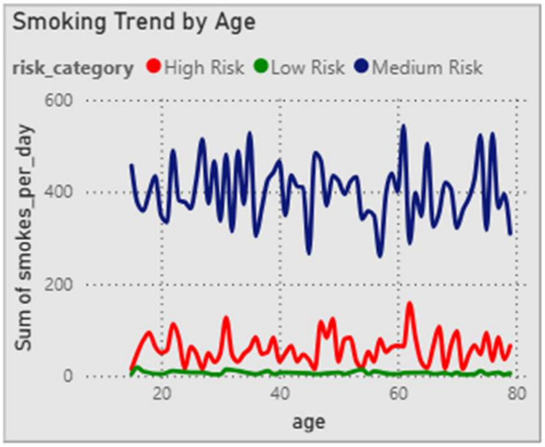


- **Drinkers vs Smokers Pie Chart**
  Compares the percentage of drinkers vs smokers in the dataset.

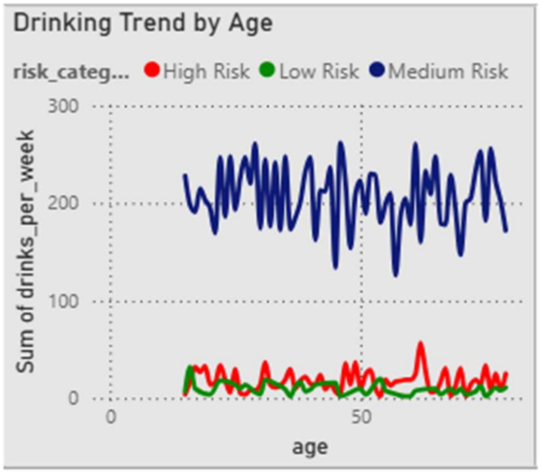
->Almost all individuals are smokers (100%) and 99.33% are drinkers.



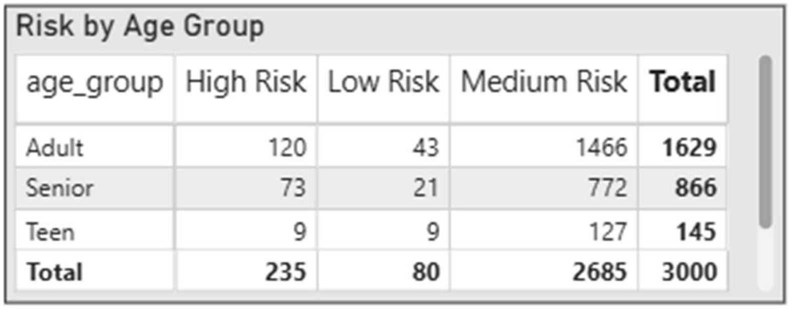
- **Smoking Trend by Age**
  A line chart showing **sum of smokes per day by age** for each risk category.
  -> Medium risk individuals consistently smoke more across ages.



- **Drinking Trend by Age**
  A line chart showing **sum of drinks per week by age** for each risk category.
  ->Medium risk category again has the highest drink count across ages.



- **Risk by Age Group Table**
  Shows **High, Medium, and Low risk counts across age groups (Adult, Senior, Teen)**.
  ->Adults have the highest count in Medium risk.

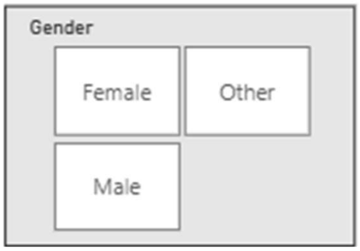


Risk by Age Group

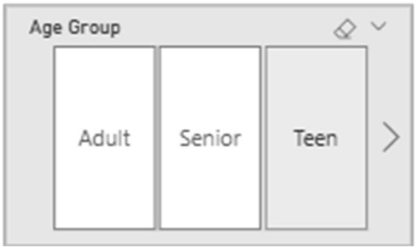
age_group	High Risk	Low Risk	Medium Risk	**Total**
Adult	120	43	1466	**1629**
Senior	73	21	772	**866**
Teen	9	9	127	**145**
Total	**235**	**80**	**2685**	**3000**

Filters

- **Gender Filter:** To analyse data by Male, Female, or Other.



- **Age Group Filter:** To focus analysis on Adult, Senior, or Teen categories.



5. Key Insights

1. **Smoking:** Medium risk is predominant; very few are in Low risk.
2. **Drinking:** Most are at Low risk; very few have no risk.
3. **Age Impact:** Adults are the highest risk group for both smoking and drinking.

6. Conclusion

This dashboard provides a **clear overview of risk behavior patterns** among individuals for smoking and drinking. It helps identify **target age groups** and **risk categories** that require attention for health interventions and awareness programs.

CHAPTER 5: RESULTS AND DISCUSSION

Output / Report

A behavioral prediction model was built using demographic, lifestyle, and health data to classify individuals into risk levels. Most individuals were found to be in the **medium smoking risk** and **low alcohol risk** categories. Adults showed the **highest risk**, especially those with **poor mental health**, **low exercise**, and **unhealthy diets**. The dashboard effectively highlighted these patterns with interactive visualizations segmented by **age and gender**.

Challenges Faced

During the development of the Risk Behavior Analysis project, we encountered several challenges across both the data exploration (EDA) and visualization (Power BI) phases:

Exploratory Data Analysis (EDA)

- **Missing and Inconsistent Data**: Many columns had null values or inconsistent formats, requiring careful imputation using techniques like mode replacement.
- **Categorical Encoding**: Converting categorical variables (e.g., gender, education level) into numerical form for analysis without losing meaning was complex.
- **Complex Visuals**: Creating clear and interpretable violin plots and boxplots across multiple variables sometimes resulted in overlapping or cluttered visuals.

Power BI Visualization

- **Data Mapping**: Mapping the processed data and model outputs to Power BI visuals required careful formatting and consistency checks.
- **Visual Clarity**: Designing visuals that clearly distinguished between high, medium, and low-risk categories required additional formatting.

Learnings

Throughout the course of this project, we acquired valuable technical and analytical skills that contributed to our growth as data science practitioners. Key learnings include:

Technical & Analytical Skills

- **Proficiency in Python for EDA**: Learned to use libraries such as pandas, seaborn, and matplotlib to clean data, analyze patterns, and visualize complex relationships effectively.

- **Feature Engineering & Encoding**: Understood the importance of transforming raw data into meaningful features and applying encoding techniques to handle categorical variables.

- **Model Evaluation Techniques**: Acquired knowledge of evaluating machine learning models using accuracy, precision, recall, F1 score, and confusion matrix.

Visualization & Interpretation

- **Power BI Dashboarding**: Developed interactive dashboards with filters and slicers to analyze risk behavior by demographics and visualize trends in a user-friendly format.

- **Data Storytelling**: Learned to extract key insights from visual data and communicate them clearly to support decision-making.

- **Design Thinking in Visuals**: Gained experience in organizing visuals for clarity, impact, and interactivity in Power BI.

Team Collaboration

- **Problem-Solving**: Tackled real-world challenges like missing data, class imbalance, and performance issues using critical thinking.

- **Collaboration**: Coordinated with team members on tasks like cleaning, modeling, and visualization to ensure consistent progress.

- **Documentation**: Learned to document the entire project pipeline clearly for reporting and presentation purposes.

CHAPTER 6: CONCLUSION

Summary

The Risk Behavior Analysis project successfully integrated **Exploratory Data Analysis (EDA)**, **Machine Learning**, and **Power BI visualization** to identify and predict risky behavioral patterns such as smoking and alcohol consumption based on demographic and lifestyle data.

Through detailed EDA using Python, we explored relationships between variables like age, gender, mental health, and physical habits. This helped uncover critical trends, such as higher risk behavior among adults and individuals with poor mental health, limited exercise, or unhealthy diets.

Using classification models like **Logistic Regression, Decision Tree**, and **Random Forest** Classifier we built predictive models to categorize individuals into **low, medium, or high-risk groups**.

Finally, the insights were translated into a visually interactive format using **Power BI dashboards**. These dashboards allowed real-time filtering by age and gender, helping to identify target groups for health awareness campaigns.

Overall, the project provided a strong foundation in data analysis and visualization, and it demonstrated how data-driven methods can support effective **public health interventions**. The integrated approach we followed can be extended for broader behavioral studies and decision-making applications in health and wellness sectors.