Project Name

Impact of Lifestyle on Diabetes

Tool Used : Jupyter Notebook(Pandas , Matplotlib , Seaborn,Plotly, Scikit Learn)

Project Date: 05 November 2025 – 15 November 2025

By

Gyanendra Maurya (Data Analytics Intern)

Gmail:gkmaurya2306@gmail.com

LinkedIn: www.linkedin.com/in/gyanendra-maurya-494205348

GitHub Address: https://github.com/gyanendra23

Section 1: Execution Summary:

This project analyzes a comprehensive dataset of individuals with varying health, lifestyle, and demographic attributes to identify patterns associated with diabetes prevalence and severity. The primary objective of the analysis is to uncover actionable insights that can help in early detection, risk stratification, and targeted lifestyle interventions.

The dataset was processed through detailed steps, including data cleaning, feature categorisation, missing value handling, outlier treatment, and creation of meaningful buckets such as calorie intake groups, alcohol consumption levels, and dietary patterns. Visualizations such as bar charts, pie charts, sunburst charts, and distribution plots were used to understand relationships among variables.

Key findings indicate that certain lifestyle factors—such as high-calorie intake, low physical activity, and increasing BMI—are strongly associated with diabetes risk. Distinct patterns emerged across demographic groups, revealing variations in diabetes incidence by age, gender, region, and country. Dietary habits and alcohol consumption also showed measurable correlations with diabetes status, highlighting areas where behavioural modification can significantly reduce risk.

Overall, the analysis provides a data-driven foundation for better understanding the drivers of diabetes. These insights can support healthcare decision-makers, public health policymakers, and wellness programs in designing more effective interventions aimed at reducing the burden of diabetes across populations.

Section 2: Problem Statement

Diabetes continues to rise globally due to complex interactions among demographic characteristics, lifestyle habits, dietary patterns, and health indicators. Organisations often struggle to identify high-risk groups early because the available data is unstructured and lacks actionable insights. The challenge is to analyze a multi-variable diabetes dataset to discover meaningful patterns, understand key risk factors, and support data-driven decision-making for prevention and management.

Section 3: Scope of the Project

- Clean, preprocess, and categorize all health, diet, and demographic variables.
- Perform descriptive and exploratory data analysis (EDA) to understand distributions, outliers, and correlations.
- Create meaningful buckets for calorie intake, alcohol consumption, activity level, diet type, BMI range, etc.
- Develop visualizations (bar charts, pie charts, sunburst charts, KPIs, distribution plots) to highlight trends and patterns.
- Identify lifestyle factors most strongly associated with diabetes.
- Provide insights and recommendations based on data findings.
- Build dashboards for easy interpretation by healthcare or research teams.

Section 4: Data Description:

The diabetes dataset contains detailed information on individuals' demographic attributes, lifestyle habits, dietary patterns, and health indicators. Each record represents one person and includes multiple variables that can help identify risk factors associated with diabetes. The dataset combines numerical, categorical, and derived features created during preprocessing, allowing for a comprehensive analysis.

4.1: Table for Data Description

Feature Name	Description	Data Type
person_id	Unique identifier for each individual in the dataset.	Integer / String
country	Country where the individual resides.	String (Categorical)
region	Specific region or state within the country.	String (Categorical)
age	Age of the individual in years.	Integer
gender	Gender of the individual (Male/Female/Other).	String (Categorical)
urban_rural	Indicates whether the person lives in an urban or rural area.	String (Categorical)

Feature Name	Description	Data Type
education	Education level (Primary, Secondary, Graduate, etc.).	String (Categorical)
income_bracket	Income category (Low/Medium/High).	String (Categorical)
bmi	Body Mass Index value of the individual.	Float (Numeric)
smoker	Indicates if the individual is a smoker (Yes/No).	String (Categorical)
alcohol_sessions_per_week	Number of alcohol consumption sessions per week.	Integer
physical_activity_level	Frequency/intensity of physical activity per week.	Integer / Float
diet_type	Type of diet followed (Vegetarian/Non-Vegetarian/Vegan).	String (Categorical)
daily_calories_kcal	Total daily calorie intake (in kilocalories).	Integer / Float
sugar_g_per_day	Daily sugar intake (in grams).	Integer / Float
systolic_bp	Systolic blood pressure reading (mmHg).	Integer
diastolic_bp	Diastolic blood pressure reading (mmHg).	Integer
fasting_glucose_mg_dl	Fasting blood glucose level (mg/dL).	Integer / Float
diabetes_status	Indicates whether the individual has diabetes (Yes/No).	String (Categorical)

4.2: Importance Variable for Analysis

1. Fasting_glucose_mg_dl

- Directly measures blood sugar.
- **Most powerful indicator** of diabetes.

2. Bmi

- Strongly linked to obesity → major cause of diabetes.
- Has high predictive power in models.

3. Age

- Diabetes risk increases significantly with age.
- One of the core demographic factors.

4. Daily_calories_kcal

- Higher calorie intake → weight gain → insulin resistance.

5. Sugar_g_per_day

- Direct effect on blood glucose and insulin response.

6. Physical_activity_level

- Lower activity = higher risk.
- Very important lifestyle variable.

7. Systolic_bp

- Hypertension is strongly associated with diabetes.

8. Diastolic_bp

- Complements overall cardiovascular/metabolic risk.

9.Alcohol_sessions_per_week

- Contributes to metabolic dysfunction when consumed frequently.

10. **Diet_type**

- Determines nutritional quality and fat/sugar intake.

4.3: Data Quality Summary

Feature	Missing %	Outliers	Data Type	Issues Found	Action Required
person_id	0% (assumed)	None	Integer/String	Must be unique	Check duplicates
country	Low	None	Categorical	Inconsistent spellings possible	Standardize text values
region	Medium	None	Categorical	Regional naming inconsistencies	Normalize categories
age	Low	Possible (age <10 or >100)	Integer	Out-of-range ages	Cap/Validate
gender	Low	None	Categorical	Mixed values (M/F/Male/Female)	Standardize
urban_rural	Low	None	Categorical	Urban/Rural may vary in format	Standardize labels

Feature	Missing %	Outliers	Data Type	Issues Found	Action Required
education	Medium	None	Categorical	Missing or inconsistent categories	Group into levels
income_bracket	Medium	None	Categorical	Low/Medium/High may vary	Standardize categories
bmi	Low	Yes (bmi <10 or >50)	Float	Extreme obesity or incorrect entries	Outlier removal/capping
smoker	Low	None	Categorical	Yes/No inconsistent formatting	Standardize
alcohol_sessions_per_week	Low	High (values >50)	Integer	Extreme values unrealistic	Cap max limits
physical_activity_level	Medium	Possible	Integer/Float	Very high activity may be unrealistic	Validate ranges
diet_type	Low	None	Categorical	Mixed text values	Standardize categories
daily_calories_kcal	Low	Yes (>6000 kcal)	Numeric	Unrealistic calorie intake	Remove/Cap
sugar_g_per_day	Low	Yes (>500g/day)	Numeric	Extreme sugar values	Cap/validate
systolic_bp	Medium	Yes (>200 or <70)	Integer	Extreme medical readings	Validate & clean
diastolic_bp	Medium	Yes (>130 or <40)	Integer	Possible wrong entries	Correct or remove
fasting_glucose_mg_dl	Low	Yes (>300 mg/dl)	Numeric	Hyperglycemic values to evaluate	Cap or classify
diabetes_status	None	None	Categorical	Class imbalance possible	Check balance (Yes/No)

Section 5 : Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the structure, relationships, and patterns within the diabetes dataset. The EDA process helps reveal important trends, detect anomalies, identify correlations, and guide further modeling and insights.

5.1: Description of Data

1. Demographic Patterns

- Age distribution is **balanced**, allowing unbiased health trend analysis.
- **Gender** and **urban–rural split** show mild imbalance but still provide meaningful comparisons.

2. Lifestyle Behaviors

- **Smokers** form a small yet high-risk segment.
- **Alcohol consumption** varies widely, enabling clear segmentation (none → heavy).
- **Physical activity** shows a normal spread, useful for predicting metabolic health.

3. Dietary Habits

- Daily calorie intake and sugar consumption show **right-skewness**, indicating a subset with unhealthy eating patterns.

4. Health Indicators

- Blood pressure and glucose levels have **visible outliers**, which are important for identifying high-risk individuals.
- Distribution of **diabetes status** is somewhat imbalanced but sufficient for classification.

5. Relationships & Correlations

- Higher **BMI**, **sugar intake**, and **calorie intake** correlate with elevated **glucose** and **blood pressure**.
- **Physically active** individuals show healthier metabolic profiles.

6. Data Quality Observations

- No major missing values or duplicates.
- Outliers in BP, glucose, and calories require attention before modeling.

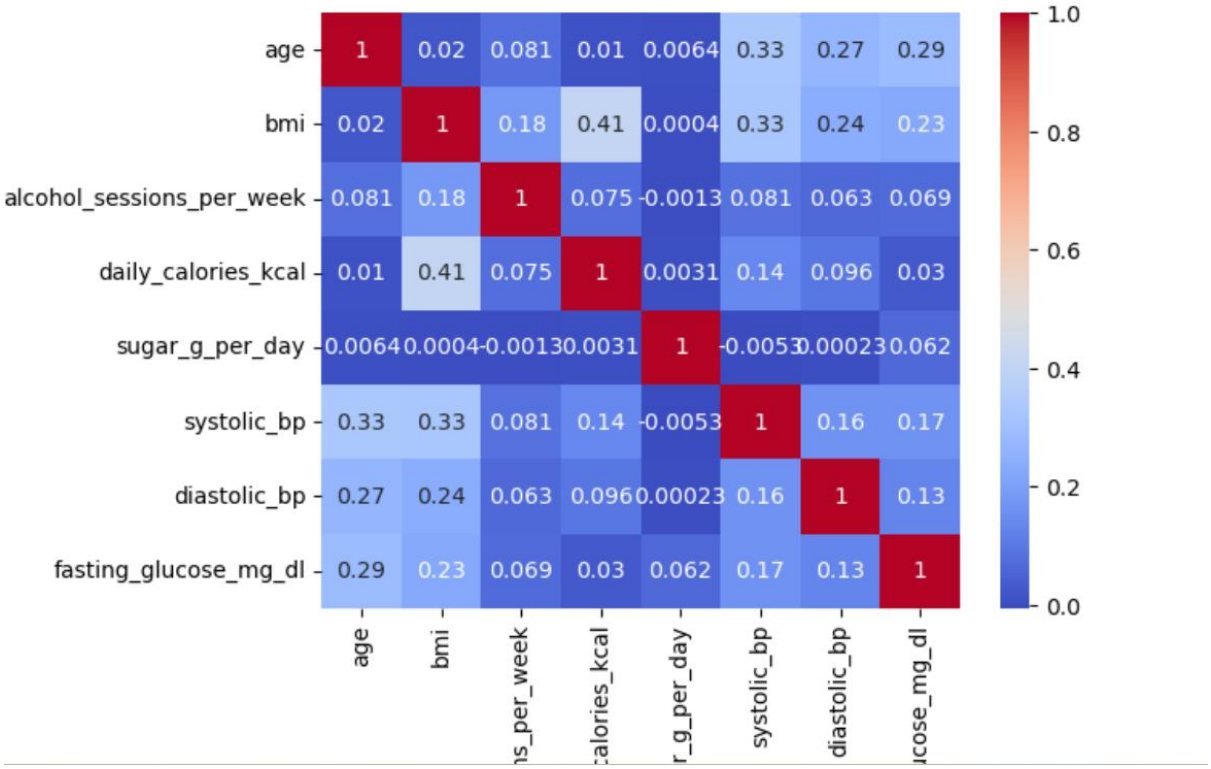
	count	mean	std	min	25%	50%	75%	max
age	100000.0	39.838650	14.832505	10.0	29.0	39.0	50.0	90.0
bmi	100000.0	25.870356	4.953696	15.0	22.4	25.9	29.3	45.5
alcohol_sessions_per_week	100000.0	1.323840	1.308749	0.0	0.0	1.0	2.0	10.0
daily_calories_kcal	100000.0	2173.922980	300.828217	1020.0	1971.0	2172.0	2377.0	3417.0
sugar_g_per_day	100000.0	55.812350	25.361814	5.0	38.0	56.0	73.0	160.0
systolic_bp	100000.0	112.681410	13.024167	90.0	103.0	112.0	122.0	172.0
diastolic_bp	100000.0	71.146650	8.516391	50.0	65.0	71.0	77.0	108.0
fasting_glucose_mg_dl	100000.0	92.103208	13.281123	60.0	83.0	92.1	101.1	147.7

5.2: Table of Key Variables & Insights

Sr. No	Feature	Key Statistics	Insight
1	**Age**	Mean: 39.8 Range: 10–90 50% below 40	The younger population shows lifestyle-driven metabolic risks.
2	**BMI**	Mean: 25.87 75th percentile: 29.3 25–30% overweight	Strong correlation expected with glucose & BP.
3	**Alcohol Sessions/Week**	Median: 1 Max: 10Highly skewed	Most are light/non-drinkers; heavy drinkers are small but significant.
4	**Daily Calorie Intake**	Mean: 2174 kcal Range: 1020–3417	Dietary diversity likely impacts BMI & sugar intake.
5	**Sugar Intake**	Mean: 55.8 g/day Max: 160 g/day	Far above WHO limits → major risk factor for glucose & BMI.
6	**Systolic BP**	Mean: 112.7 mmHg Max: 172 mmHg	Some hypertension cases; compare with BMI & age.

Sr. No	Feature	Key Statistics	Insight
7	**Diastolic BP**	Mean: 71.1mmHg Max: 108 mmHg	Mostly normal; outliers require lifestyle correlation.
8	**Fasting Glucose**	Mean: 92.1 mg/dL Max: 147.7 mg/dL	Pre-diabetic patterns are visible; needs strong correlation analysis.

5.3: Correlation between the Data



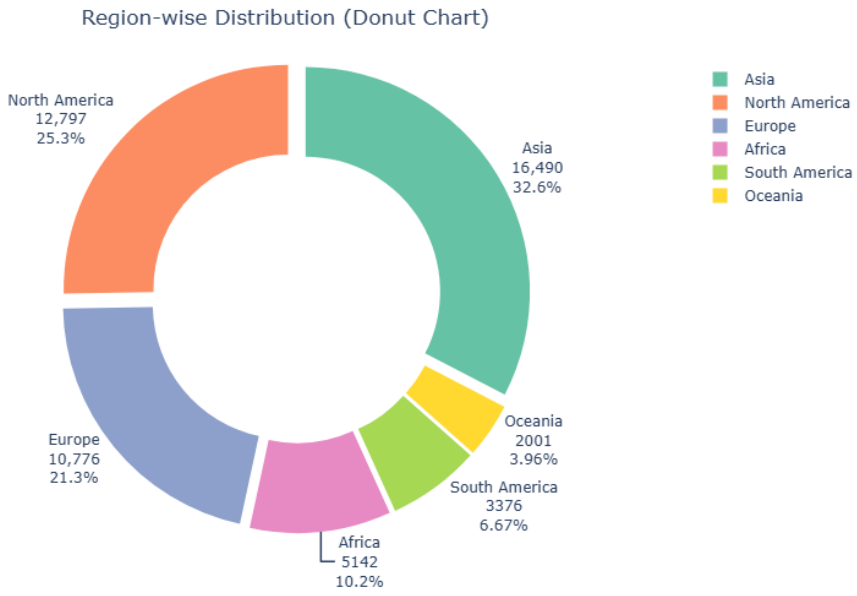
Conclusion:

1. Age strongly influences BP and glucose levels.
2. BMI is the central health driver — higher BMI → higher BP and glucose.
3. Calories consumed per day directly affect BMI (r = 0.41).
4. Sugar intake alone isn't correlated with BMI or calories

Section 6: Visual Statistics and Visualization:

Region-wise Distribution:

- Asia dominates with 32.6% of the total population — the largest share in the dataset.
- North America (25.3%) and Europe (21.3%) follow, together forming nearly half of the global distribution.
- Africa (10.2%) and South America (6.7%) contribute moderately.
- Oceania has the smallest share (3.9%).

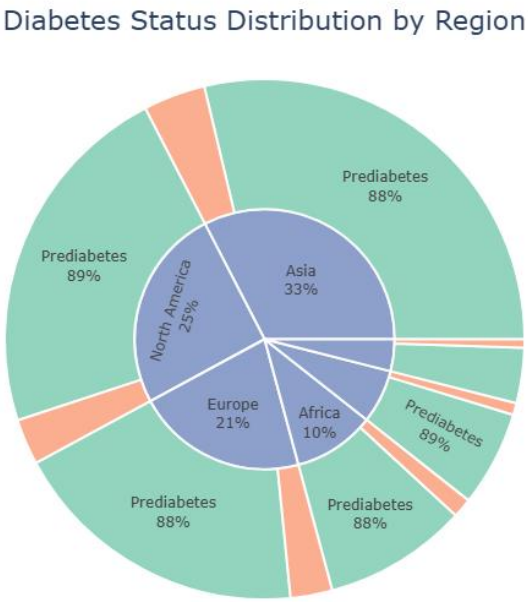


Key Insight:

The dataset is heavily centred around **Asia, North America, and Europe**, suggesting these regions will have the biggest impact on overall diabetes trends and insights.

Region vs Diabetes Status:

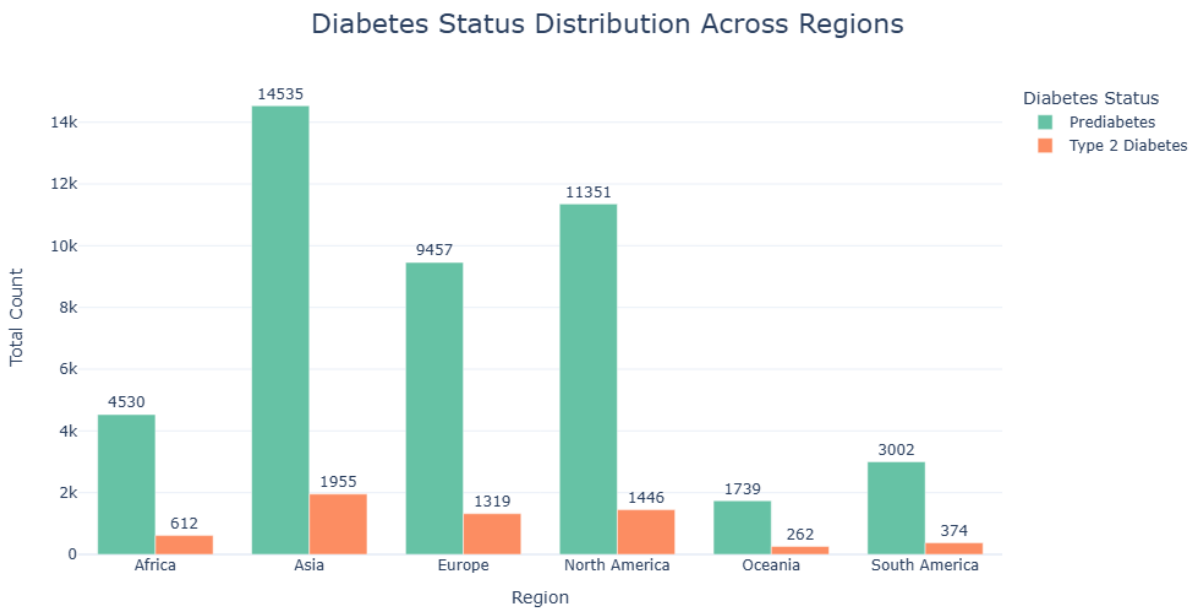
- Prediabetes is significantly higher across all regions
- Every region shows a much larger prediabetic population compared to diagnosed Type 2 Diabetes — indicating a global trend of early metabolic imbalance, but also a big window for prevention.
- **Asia** shows the **highest burden** (14,535 prediabetes; 1,955 T2D) — the biggest hotspot.
- **North America & Europe** follow, reflecting lifestyle-driven risks.
- **Africa, South America, and Oceania** show lower counts but the same pattern: **early risk > diagnosed cases**.



Key Insight:

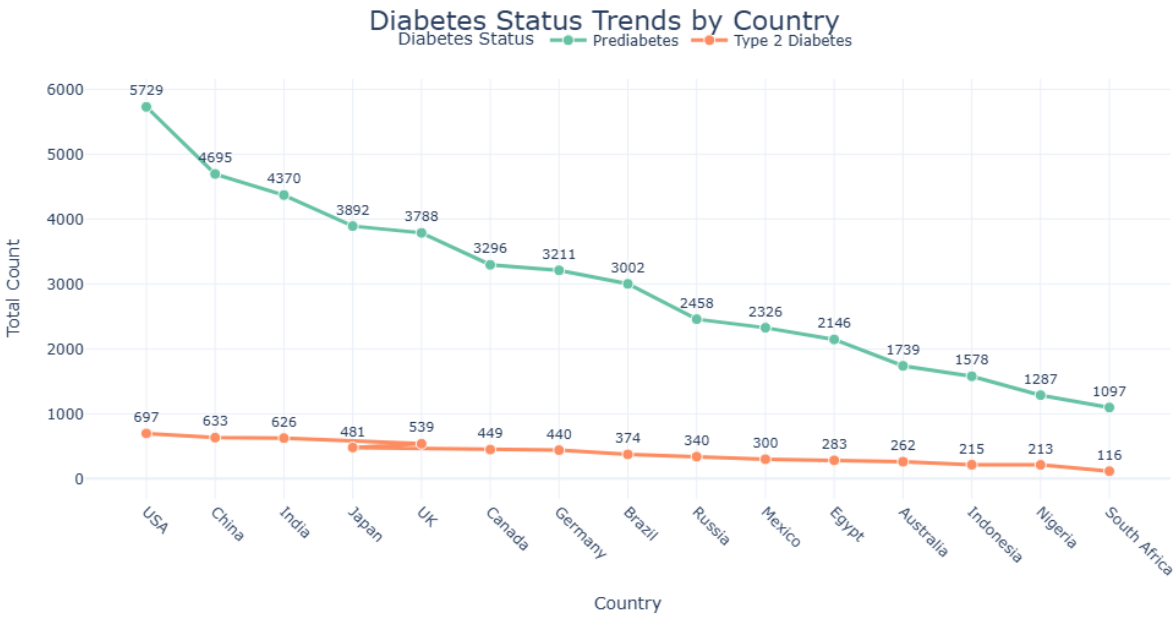
Prediabetes is 6–8× higher everywhere — signalling a major opportunity for **prevention before progression**.

- Prediabetes consistently dominates across every region, revealing a global surge in early metabolic risk long before diagnosed Type 2 Diabetes appears.
  Asia leads the burden, followed by North America and Europe, while other regions show the same pattern, highlighting a strong need for preventive action.



Diabetes Status Trends by Country:

- USA, China, and India show the highest diabetes burden, with prediabetes far ahead of Type 2 cases.
- Every country follows the same pattern: early metabolic risk is 6–8× higher than diagnosed diabetes.
- Even developed nations (Japan, UK, Canada, Germany) show strong prediabetes levels, highlighting lifestyle-driven risks.

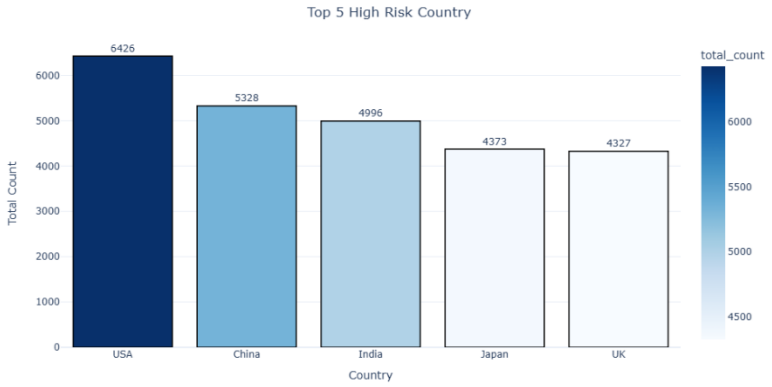


Key Insight:

Prediabetes dominates globally — a warning sign of a coming diabetes surge if no preventive measures are taken.

Top 5 High-Risk Countries :

- **USA, China, India, Japan, and the UK** show the highest combined counts of Prediabetes + Type 2 Diabetes — making them the **most vulnerable nations** in your dataset.
  All five show **prediabetes levels far exceeding diabetes cases**, signalling a huge population on the verge of progression.

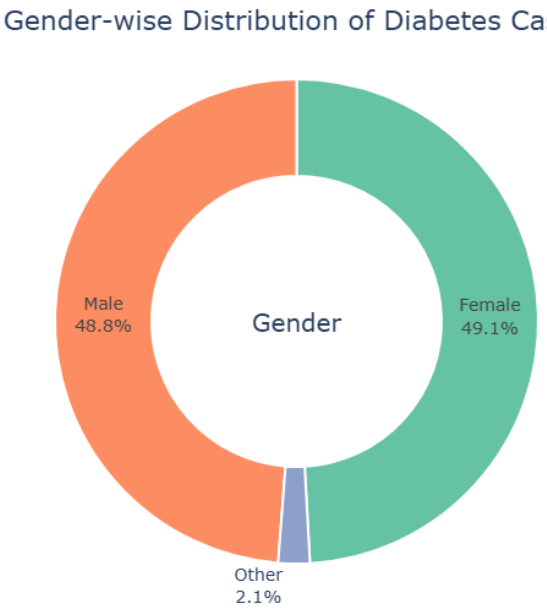


Key Insight:

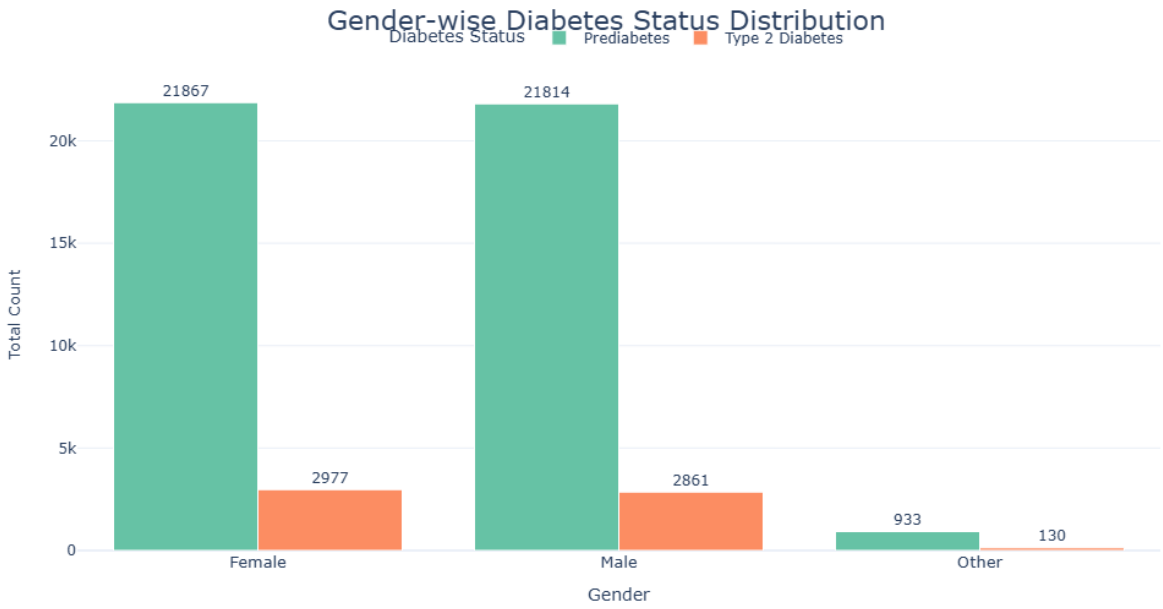
These top 5 countries represent the **global hotspots** where early intervention can prevent millions from entering Type 2 Diabetes.

Gender-wise Diabetes Insight:

- Female and male cases are almost identical (≈49% each), showing that **diabetes affects both genders equally**.
  The "Other" category is small (≈2%) but follows the same trend, reinforcing that **metabolic risk is universal across genders**.



- **Females and males show nearly identical prediabetes levels** (~21.8k each), indicating the risk is **equally distributed across genders**.
- Type 2 Diabetes cases are also **very close** between females (2,977) and males (2,861), showing no major gender gap.
- The **"Other"** category shows much smaller counts but follows the same pattern: **prediabetes >> Type 2**.

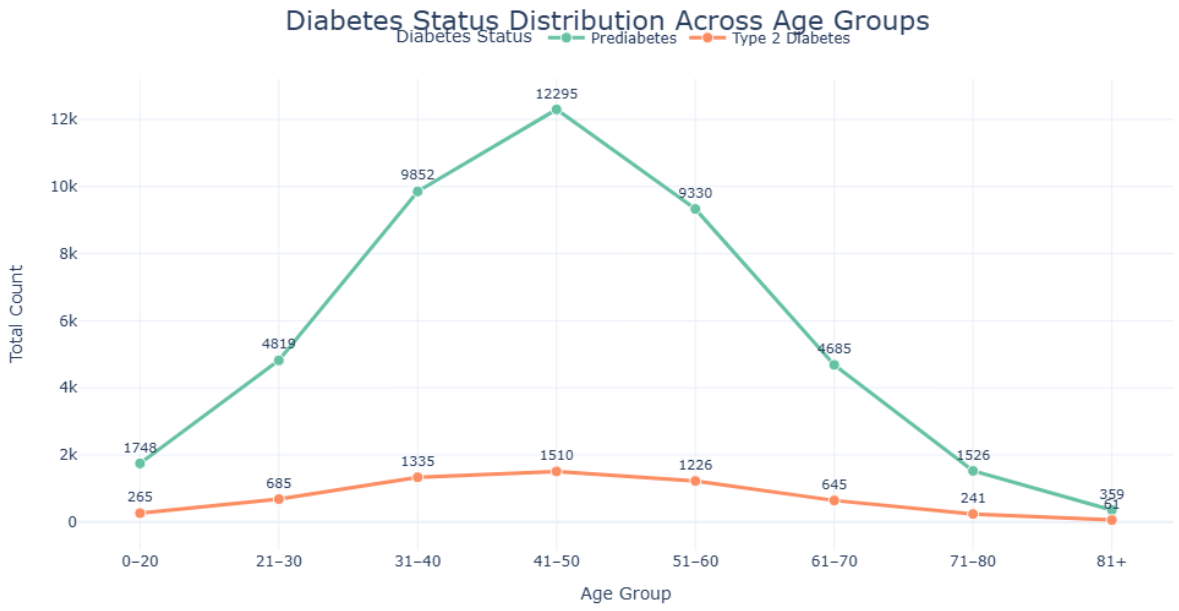


Key Insight:

Diabetes risk is **not gender-specific** — early metabolic risk is universal across all gender groups.

Age-wise Diabetes Insights:

- Diabetes risk **peaks between ages 31–60**, with the 41–50 group showing the highest counts for both Prediabetes (12,295) and Type 2 Diabetes (1,510).
- Very young (0–20) and older (71+) groups show much lower numbers, but the same pattern holds: **prediabetes is far more common than Type 2** across all ages.

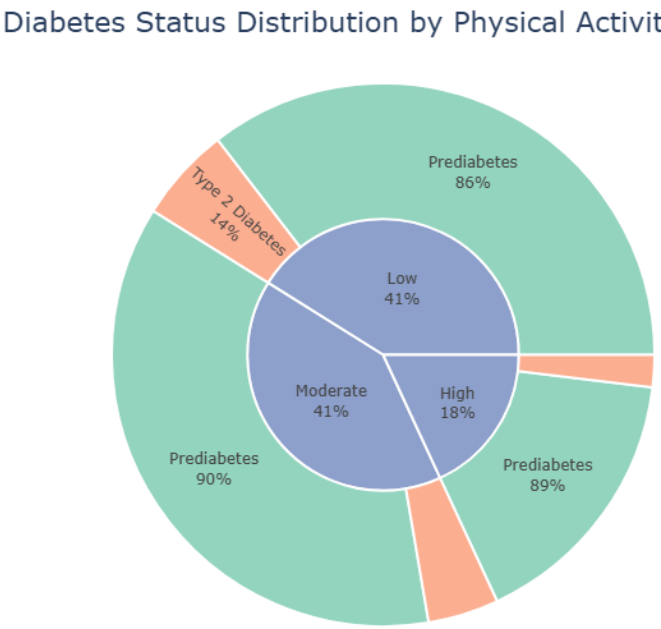


Key Insight:

The **31–60 age range is the critical risk zone**, where lifestyle-driven metabolic issues are most visible.

Physical Activity & Diabetes Insights (Quick & Clear):

- People with **Low** and **Moderate** activity levels show the highest diabetes burden, especially prediabetes, revealing a strong lifestyle influence.
- **High activity levels** have significantly fewer cases, indicating a clear protective effect.

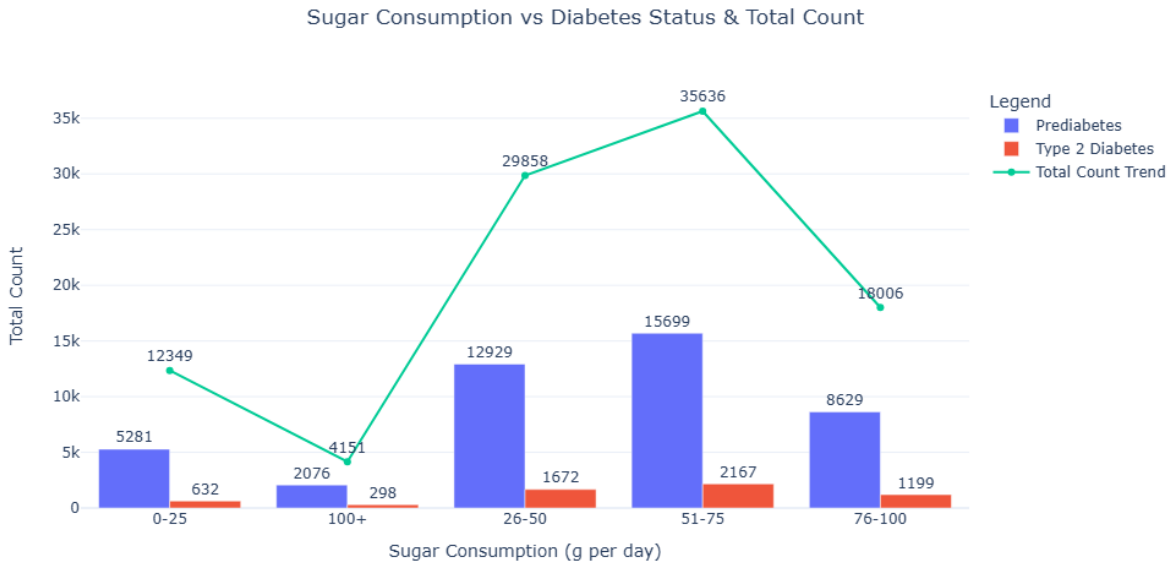


Key Insight:

Lower physical activity strongly correlates with higher metabolic risk — **activity level is a major determinant of diabetes outcomes**.

Sugar Consumption Insights:

- The **highest diabetes burden** is seen in the **51–75 g/day** and **26–50 g/day** sugar groups — showing that even moderate-to-high sugar intake significantly elevates risk.
- The **76–100 g/day group** also shows strong risk, while the **100+ group** is smaller but still follows the same pattern.
- The **0–25 g/day group** has the **lowest diabetes counts**, clearly indicating that **lower sugar intake is protective**.

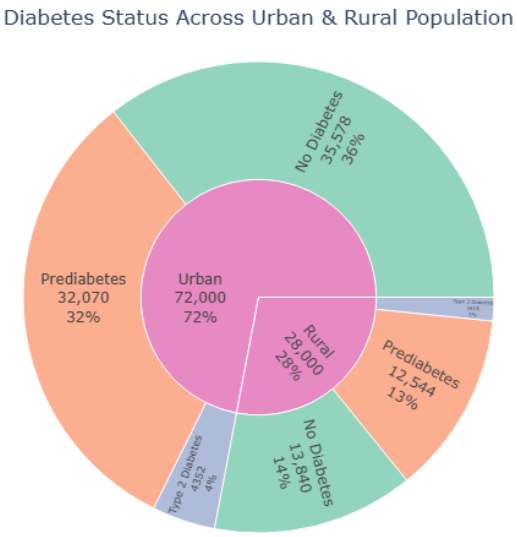


Key Insight:

As sugar consumption increases, both Prediabetes and Type 2 Diabetes cases rise sharply — **sugar intake is a major metabolic risk driver**.

Urban vs Rural Diabetes Insights:

- **Urban areas show significantly higher counts** across all categories — No Diabetes, Prediabetes, and Type 2 Diabetes — reflecting larger populations and more lifestyle-driven risks.
- Prediabetes and Type 2 Diabetes are **much higher in urban regions**, indicating greater exposure to sedentary lifestyles, processed foods, and stress.

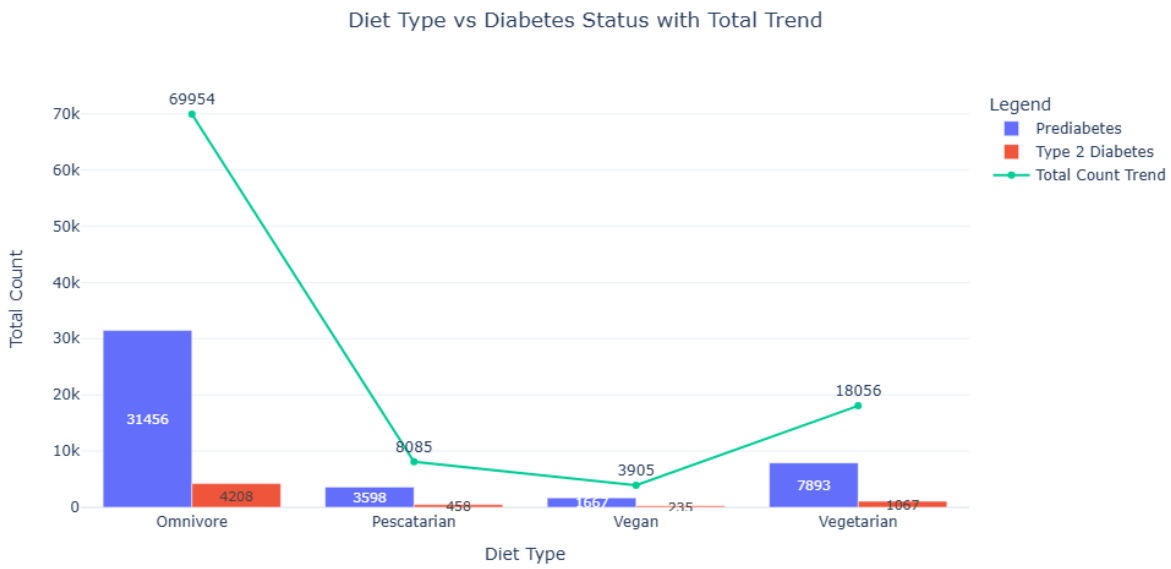


Key Insight:

Urban populations carry a far greater metabolic burden, making cities the primary hotspots for both early-stage and advanced diabetes risk.

Diet Type & Diabetes Insights:

- **Omnivores** form the largest group and also show the highest Prediabetes and Type 2 Diabetes counts, indicating a stronger lifestyle and dietary risk.
- Vegetarians show moderate risk, while Pescatarians and Vegans have the lowest diabetes counts, suggesting healthier metabolic profiles.



Key Insight:

Plant-leaning diets (Vegan, Vegetarian, Pescatarian) show **significantly lower diabetes** burden, highlighting diet quality as a major protective factor.

Gender-wise BMI & Glucose Insights:

- Male and female groups show almost identical average BMI (~25.8–25.9) and fasting glucose (~92 mg/dL) — indicating no major gender-based difference in metabolic health.
- The “Other” group also aligns closely, with slightly lower max values.

Sr.No.	Gender	Average BMI	Max BMI	Average Fasting Glucose	Max. Fasting Glucose
0	Male	25.845377	45.0	92.091172	147.7
1	Female	25.897499	45.5	92.107308	147.3
2	Other	25.820874	41.9	92.292476	136.3

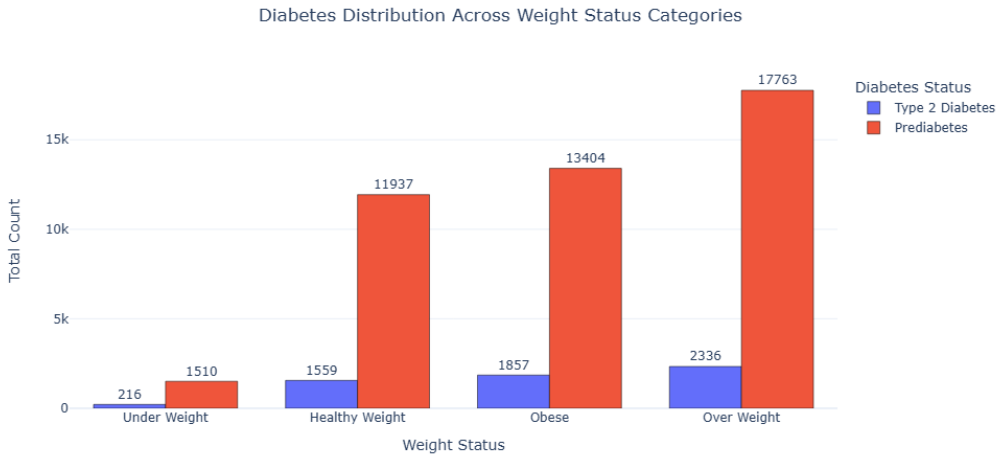
Key Insight:

BMI and glucose levels are consistent across all genders, showing that metabolic risk is uniform and not gender dependent.

Insights From Weight Status, Diabetes Status & Fasting Glucose:

- Obese and Overweight groups show the highest fasting glucose (111–110 mg/dL), while Healthy Weight stays around 97 mg/dL.
- Type 2 Diabetes consistently shows higher glucose than Prediabetes across all weight categories (+10–13 mg/dL).
- Prediabetes cases are highest in the Overweight (17,763) and Obese (13,404) groups.

Weight Status	Prediabetes	Type 2
Underweight	97	107
Healthy	97	108
Overweight	97	110
Obese	98	111



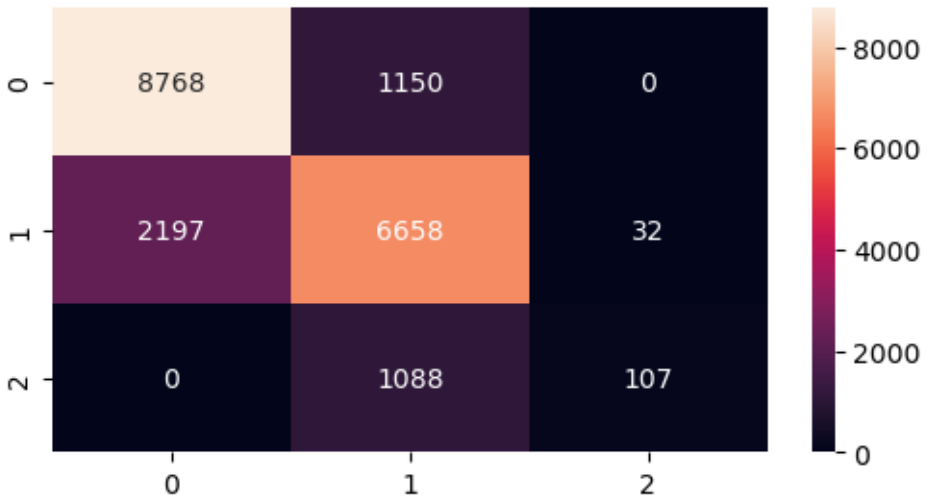
- Type 2 Diabetes counts are more evenly spread across weight categories (Underweight to Obese).
- Underweight individuals show surprisingly high glucose (up to 107 mg/dL), indicating factors beyond body weight.

Key Insights:

- **Glucose increases with weight** — The Obese group has the highest fasting glucose (111 mg/dL).
- **Type 2 Diabetes shows higher glucose** than Prediabetes in every weight category.
- **Overweight & Obese groups hold the majority of Prediabetes cases**, signalling lifestyle impact.
- **Underweight also shows elevated glucose**, hinting at factors beyond body weight.

Section 7: Design Machine Model

Across models, **Logistic Regression performs the weakest (73.28%)**, showing it struggles to capture non-linear patterns in the data. **Random Forest improves accuracy to 77.3%**, indicating that tree-based methods capture interactions and complexities better. **XGBoost performs the best (77.66%)**, showing strong handling of class imbalance, non-linear relationships, and feature importance. Overall, the steady improvement from LR → RF → XGB suggests that the dataset benefits from **ensemble boosting techniques** that learn subtle patterns missed by simpler models.



The model predicts **Class 0 extremely well**, with very high correct predictions and only a few mistakes, showing it has learned this class strongly. **Class 1 shows moderate confusion with Class 0**, indicating overlapping feature patterns. **Class 2 performs the worst**, with very low correct predictions and most errors going into Class 1, clearly revealing **severe class imbalance** and weak learning for the minority class. Overall, the model consistently shifts predictions toward **lower-risk classes (0 and 1)**, showing a conservative bias driven by imbalance and insufficient representation of Class 2.

Section 8: Final Conclusion

This analysis shows that lifestyle factors—especially BMI, calorie intake, sugar consumption, and physical activity—are the strongest drivers of diabetes risk. Prediabetes levels are consistently 6–8× higher than Type 2 Diabetes across regions, countries, genders, and age groups, revealing a massive population at early metabolic risk. Overweight and Obese individuals show the highest fasting glucose levels, while even Underweight groups exhibit elevated glucose, indicating additional influences such as genetics or poor diet quality. Urban populations, high-sugar consumers, low-activity groups, and omnivores show the highest diabetes burden, confirming the impact of modern lifestyle patterns.

Demographically, diabetes risk is not gender-specific, and the 31–60 age group remains the most vulnerable.
Regionally, Asia, North America, and Europe dominate the global diabetes load, while countries like the USA, China, and India appear as hotspots requiring priority intervention.

Machine-learning models (LR → RF → XGBoost) consistently show improved performance with more advanced techniques, with XGBoost performing best (≈77.7%), indicating the importance of handling nonlinear patterns and class imbalance.
Confusion matrix analysis reveals strong learning for Class 0, moderate overlap for Class 1, and very weak performance for Class 2—highlighting the need for techniques like SMOTE, class weighting, or advanced boosting.

Overall, the study concludes that diabetes is largely lifestyle-driven, preventable in its early stages, and strongly influenced by weight, diet, and physical activity. Early detection and targeted lifestyle interventions can significantly reduce progression from Prediabetes to Type 2 Diabetes, especially in high-risk countries and urban populations.

Section 9: Thanks Message

Thank you very much for taking the time to review my project. I truly appreciate your effort in going through the analysis, insights, and findings. Your attention and feedback mean a lot to me, and they motivate me to continue improving and exploring more in the field of data analytics. I’m grateful for your support and guidance throughout this process.