Diabetes Health Indicators

Context

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in

10 prediabetics are unaware of their risk. While there are different types of diabetes, type II diabetes is the most common form and its prevalence varies by age, education, income, location, race, and other social determinants of health. Much of the burden of the disease falls on those of lower socioeconomic status as well. Diabetes also places a massive burden on the economy, with diagnosed diabetes costs of roughly $327 billion dollars and total costs with undiagnosed diabetes and prediabetes approaching $400 billion dollars annually.

Content

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used.

Some of the following research questions:

- What percentage of patients have diabetes?
- Which risk groups have the highest number of diabetes cases?
- What are the most common risk factors among patients?

Familiarization with the data:

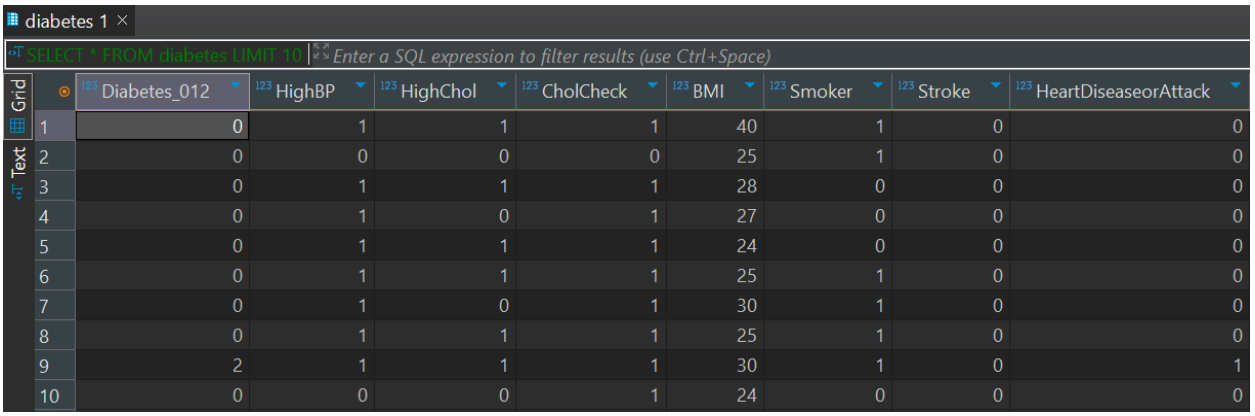
Column	Description
Diabetes_012	0 = No diabetes, 1 = Prediabetes, 2 = Diabetes
HighBP	0 = No, 1 = Yes
HighChol	0 = No, 1 = Yes
CholCheck	0 = Cholesterol not checked in the last 5 years, 1 = Checked
BMI	Numerical value representing Body Mass Index (e.g., 24.5, 30.2).
Smoker	0 = No, 1 = Yes
Stroke	0 = No, 1 = Yes
HeartDiseaseorAttack	0 = No, 1 = Yes
PhysActivity	0 = No physical activity in the last 30 days, 1 = Yes
Fruits	0 = Does not regularly consume fruits, 1 = Regularly consumes fruits
Veggies	0 = Does not regularly consume vegetables, 1 = Regularly consumes vegetables
HvyAlcoholConsump	0 = No, 1 = Yes
AnyHealthcare	0 = No access to healthcare, 1 = Access to healthcare
NoDocbcCost	0 = Did not avoid seeing a doctor due to cost, 1 = Avoided seeing a doctor due to cost
GenHlth	1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor
MentHlth	Numerical value: Number of mentally unhealthy days in the last 30 days (0–30).
PhysHlth	Numerical value: Number of physically unhealthy days in the last 30 days (0–30).
DiffWalk	0 = No difficulty walking, 1 = Difficulty walking
Sex	0 = Female, 1 = Male
Age	1 = 18–24 years, 2 = 25–29 years, 3 = 30–34 years, 4 = 35–39 years, 5 = 40–44 years, 6 = 45–49 years, 7 = 50–54 years, 8 = 55–59 years, 9 = 60–64 years, 10 = 65–69 years, 11 = 70–74 years, 12 = 75–79 years, 13 = 80 years or older
Education	1 = No schooling, 2 = Elementary school, ..., 6 = College graduate
Income	1 = Less than $10,000, 2 = $10,000–$15,000, 3 = $15,000–$20,000, 4 = $20,000–$25,000, 5 = $25,000–$35,000, 6 = $35,000–$50,000, 7 = $50,000–$75,000, 8 = More than $75,000

The values are already aggregated, so many columns have binary values (0 or 1) that correspond to a particular state or category.

We have **253 680 rows** and **22 columns** of REAL-type data (for numbers with medium precision).

```
PRAGMA table_info(diabetes);

SELECT * FROM diabetes LIMIT 10;
```

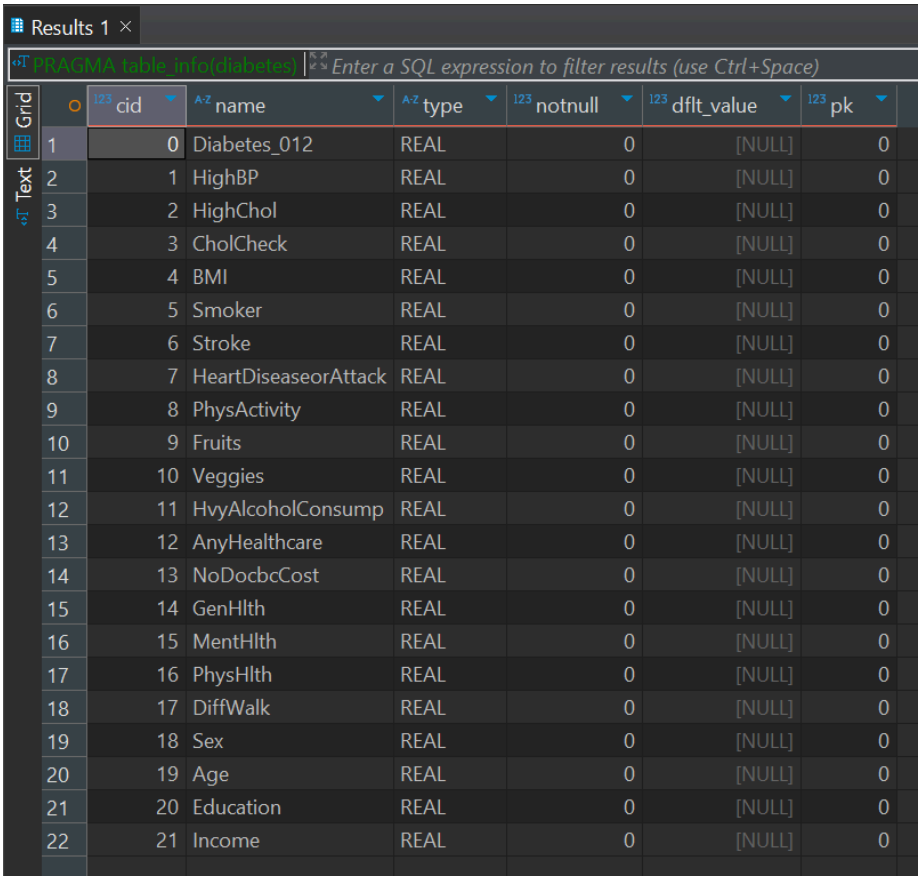
Results:



diabetes 1

SELECT * FROM diabetes LIMIT 10 | Enter a SQL expression to filter results (use Ctrl+Space)

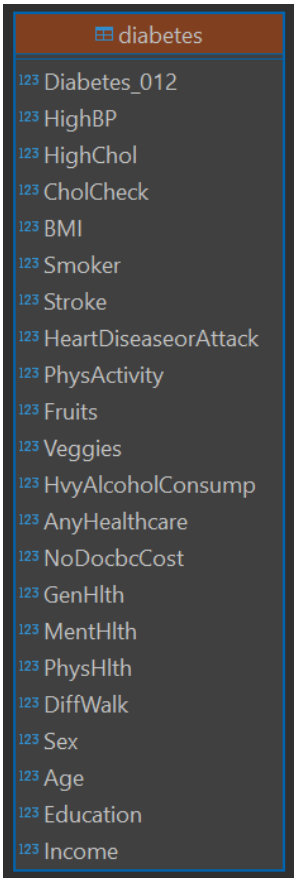
	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack
1	0	1	1	1	40	1	0	0
2	0	0	0	0	25	1	0	0
3	0	1	1	1	28	0	0	0
4	0	1	0	1	27	0	0	0
5	0	1	1	1	24	0	0	0
6	0	1	1	1	25	1	0	0
7	0	1	0	1	30	1	0	0
8	0	1	1	1	25	1	0	0
9	2	1	1	1	30	1	0	1
10	0	0	0	1	24	0	0	0



Results 1

PRAGMA table_info(diabetes) | Enter a SQL expression to filter results (use Ctrl+Space)

	cid	name	type	notnull	dflt_value	pk
1	0	Diabetes_012	REAL	0	[NULL]	0
2	1	HighBP	REAL	0	[NULL]	0
3	2	HighChol	REAL	0	[NULL]	0
4	3	CholCheck	REAL	0	[NULL]	0
5	4	BMI	REAL	0	[NULL]	0
6	5	Smoker	REAL	0	[NULL]	0
7	6	Stroke	REAL	0	[NULL]	0
8	7	HeartDiseaseorAttack	REAL	0	[NULL]	0
9	8	PhysActivity	REAL	0	[NULL]	0
10	9	Fruits	REAL	0	[NULL]	0
11	10	Veggies	REAL	0	[NULL]	0
12	11	HvyAlcoholConsump	REAL	0	[NULL]	0
13	12	AnyHealthcare	REAL	0	[NULL]	0
14	13	NoDocbcCost	REAL	0	[NULL]	0
15	14	GenHlth	REAL	0	[NULL]	0
16	15	MentHlth	REAL	0	[NULL]	0
17	16	PhysHlth	REAL	0	[NULL]	0
18	17	DiffWalk	REAL	0	[NULL]	0
19	18	Sex	REAL	0	[NULL]	0
20	19	Age	REAL	0	[NULL]	0
21	20	Education	REAL	0	[NULL]	0
22	21	Income	REAL	0	[NULL]	0



Creating tables and building relations

We can divide the *diabetes* data into the following logical tables:

patients	health_conditions	lifestyle	healthcare_access	health_status
patient_id (PK) Age Sex Education Income	patient_id (FK) HighBP HighChol BMI HeartDiseaseorAttack Diabetes_012	patient_id (FK) Smoker PhysActivity Fruits Veggies HvyAlcoholConsump	patient_id (FK) AnyHealthcare NoDocbcCost	patient_id (FK) GenHlth MentHlth PhysHlth DiffWalk

Creating tables using SQL

patients

```
CREATE TABLE patients(
        patient_id INTEGER PRIMARY KEY AUTOINCREMENT
        , Age INTEGER
        , Sex INTEGER
        , Education INTEGER
        , Income
);
```

health_conditions

```
CREATE TABLE health_conditions(
        patient_id INTEGER
        , HighBP INTEGER
        , HighChol INTEGER
        , BMI INTEGER
        , HeartDiseaseorAttack INTEGER
        , Diabetes_012
        , FOREIGN KEY(patient_id) REFERENCES patients(patient_id)
);
```

lifestyle

```
CREATE TABLE lifestyle(
        patient_id INTEGER
        , Smoker INTEGER
        , PhysActivity INTEGER
        , Fruits INTEGER
        , Veggies INTEGER
        , HvyAlcoholConsump INTEGER
        , FOREIGN KEY (patient_id) REFERENCES patients (patient_id)
```

```
);
```

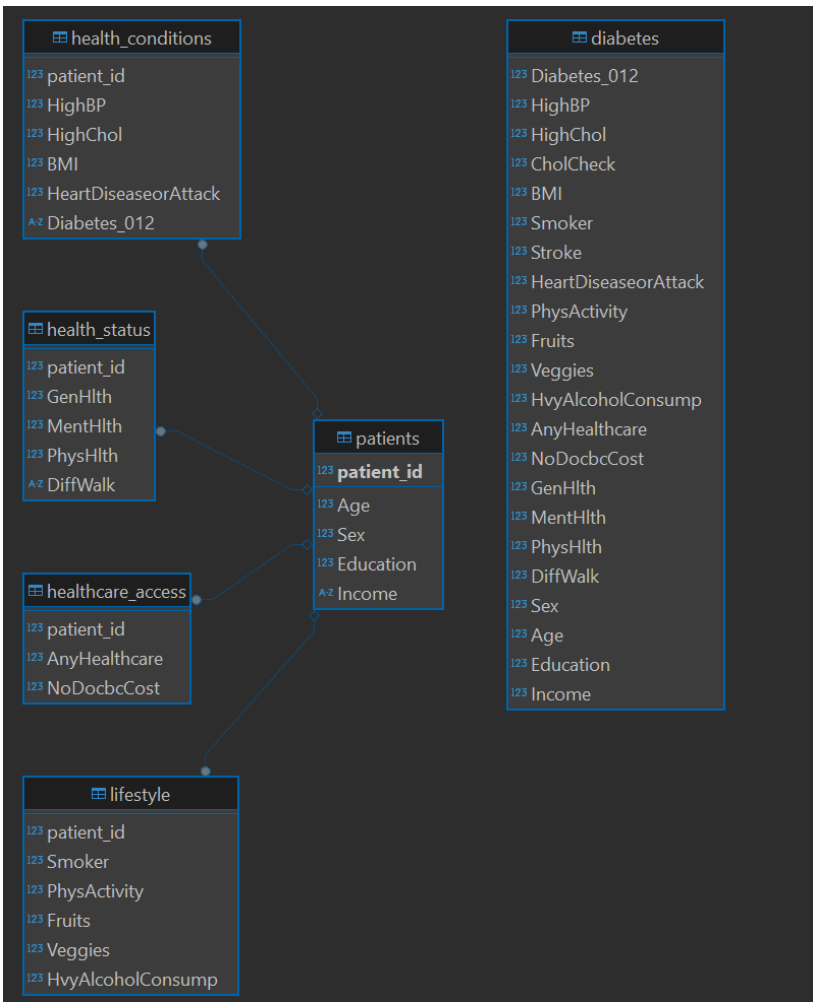
healthcare_access

```
CREATE TABLE healthcare_access(
        patient_id INTEGER
        , AnyHealthcare INTEGER
        , NoDocbcCost INTEGER
        , FOREIGN KEY(patient_id) REFERENCES patients(patient_id)
);
```

health_status

```
CREATE TABLE health_status (
        patient_id INTEGER
        , GenHlth INTEGER
        , MentHlth INTEGER
        , PhysHlth INTEGER
        , DiffWalk
        , FOREIGN KEY (patient_id) REFERENCES patients(patient_id)
);
```

Great! We've created new tables, take a look at our chart **ER-diagram**:



Enter values into these tables.

patients

```
INSERT INTO patients (patient_id, Age, Sex, Education, Income)
SELECT rowid, Age, Sex, Education, Income
FROM diabetes;
```

health_conditions

```
INSERT INTO health_conditions (patient_id, HighBP, HighChol, BMI, HeartDiseaseorAttack,
Diabetes_012)
SELECT rowid, HighBP, HighChol, BMI, HeartDiseaseorAttack, Diabetes_012
FROM diabetes;
```

lifestyle

```
INSERT INTO lifestyle(patient_id, Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump)
SELECT rowid, Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump
FROM diabetes;
```

healthcare_access

```
INSERT INTO healthcare_access (patient_id, AnyHealthcare, NoDocbcCost)
SELECT rowid, AnyHealthcare, NoDocbcCost
FROM diabetes;
```

health_status

```
INSERT INTO health_status(patient_id, GenHlth, MentHlth, PhysHlth, DiffWalk)
SELECT rowid, GenHlth, MentHlth, PhysHlth, DiffWalk
FROM diabetes;
```

And eventually check-up our new tables:

```
SELECT * FROM patients LIMIT 5;
SELECT * FROM health_conditions LIMIT 5;
SELECT * FROM lifestyle LIMIT 5;
SELECT * FROM healthcare_access LIMIT 5;
SELECT * FROM health_status LIMIT 5;
```

Data Analysis

A. Simple analysis

1. Share people with diabetes.

```
SELECT
        Diabetes_012
        , COUNT(*)
        , (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM health_conditions)) AS percentage
FROM health_conditions
GROUP BY Diabetes_012;
```

2. Age distribution

```
SELECT
        Age
        , COUNT(*) AS patient_count
FROM patients
GROUP BY Age
ORDER BY Age;
```

B. Deeper analysis

3. Patients with high blood pressure who also have diabetes.

```
SELECT
        p.Age
        , hc.HighBP
        , hc.Diabetes_012
FROM patients p
JOIN health_conditions hc ON p.patient_id = hc.patient_id
WHERE hc.HighBP = 1 AND hc.Diabetes_012 = 2;
```

4. Lifestyle analysis: how physical activity is related to diabetes.

```
SELECT
        l.PhysActivity
        , hc.Diabetes_012
        , COUNT(*) AS patient_count
FROM lifestyle l
JOIN health_conditions hc ON hc.patient_id = l.patient_id
GROUP BY l.PhysActivity, hc.Diabetes_012
ORDER BY patient_count
```

Visualisation

For visualization I'll use **Python** (**pandas** and **matplotlib / seaborn** library).
Firstly we need to import libraries for work and create connection with data base.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sqlite3
```

```
# DB connection
conn = sqlite3.connect(r"C:\Users\Nickel-Man\Desktop\SQL\diabeth.db")
```

Next step create necessary query:

```
# Виконання SQL-запиту
q_diabetes = """
SELECT
    Diabetes_012
    , COUNT(*)
    , (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM health_conditions)) AS percentage
FROM health_conditions
GROUP BY Diabetes_012;
"""



# Завантаження результатів у DataFrame
df = pd.read_sql_query(q_diabetes, conn)

# Виведення результатів
print(df)
```

We get the result:

```
   Diabetes_012  COUNT(*)  Percentage
0           0.0    213703   84.241170
1           1.0     4631    1.825528
2           2.0     35346   13.933302
```

Create plots in one cells, to receive data "people with diabetes":

```
# Rename columns for clarity (if not already done)
df_diabetes.rename(columns={"Diabetes_012": 'Diabetes', "COUNT(*)": "Count", "percentage":
"Percentage"}, inplace=True)
```

```
# Create a figure with two subplots
fig, axes = plt.subplots(1, 2, figsize=(12, 5))  # 1 row, 2 columns, adjust `figsize` as needed

# Plot 1: Count by Diabetes Category
sns.barplot(x="Diabetes", y="Count", data=df_diabetes, ax=axes[0])
axes[0].set_title("Count by Diabetes Category")
axes[0].set_xlabel("Diabetes Category")
axes[0].set_ylabel("Count")

# Plot 2: Percentage by Count
sns.barplot(x="Percentage", y="Count", data=df_diabetes, ax=axes[1])
axes[1].set_title("Percentage by Count")
axes[1].set_xlabel("Percentage")
axes[1].set_ylabel("Count")

# Adjust layout
plt.tight_layout()
plt.show()
```

“Age distribution” code.

```
sns.barplot(x='Age', y='patient_count', data=df_age, palette='viridis')

plt.xlabel('Age Group')
plt.ylabel('Number of Patients')
plt.title('Patient Distribution by Age Group')
plt.show()
```

Patients with high blood pressure who also have diabetes.

```
sns.histplot(df_highBp['Age'],bins=20,kde=True, color='skyblue', edgecolor='black')
plt.title('Age distribution of Patients with High BP and Diabetes')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```

How physical activity is related to diabetes

```
sns.barplot(x='PhysActivity', y='patient_count', hue='Diabetes_012',data=df_activity)

plt.title('Patient Count by Physical Activity and Diabetes Status')
```

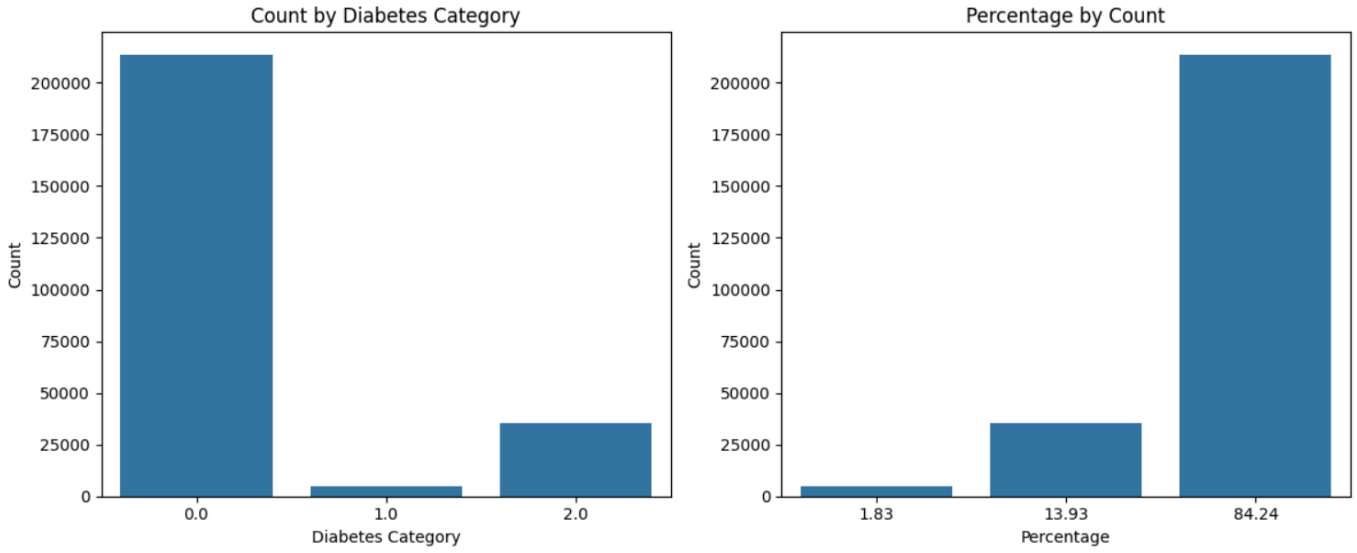
```
plt.xlabel('Physical Activity Level')
plt.ylabel('Patient Count')
plt.legend(title='Diabetes Status')
plt.show()
```

or heatmap:

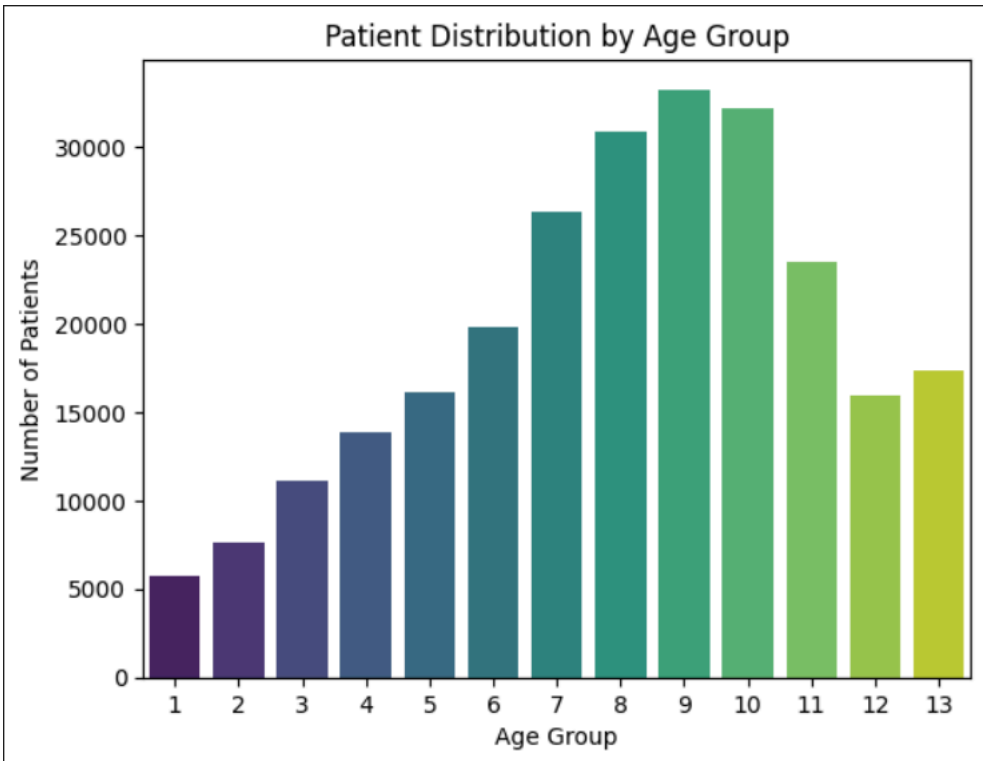
```
# Pivot data for the heatmap
heatmap_data = df_activity.pivot(index="PhysActivity", columns="Diabetes_012",
values="patient_count")

# Create the heatmap
sns.heatmap(heatmap_data, annot=True, fmt="d", cmap="YlGnBu")

# Add titles and labels
plt.title("Patient Count Heatmap: Physical Activity vs Diabetes Status")
plt.xlabel("Diabetes Status")
plt.ylabel("Physical Activity Level")
plt.show()
```

Visual Diagrams and Bars

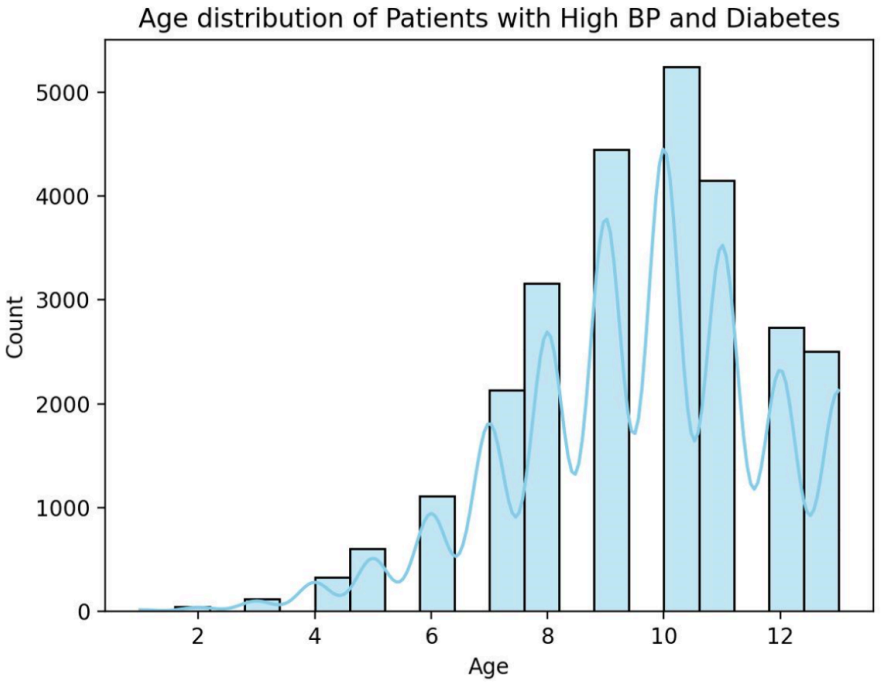
Share people with diabetes.



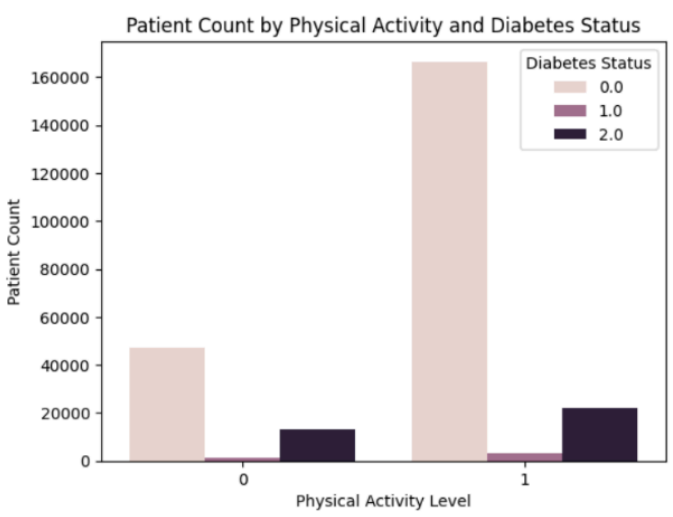
Age distribution.

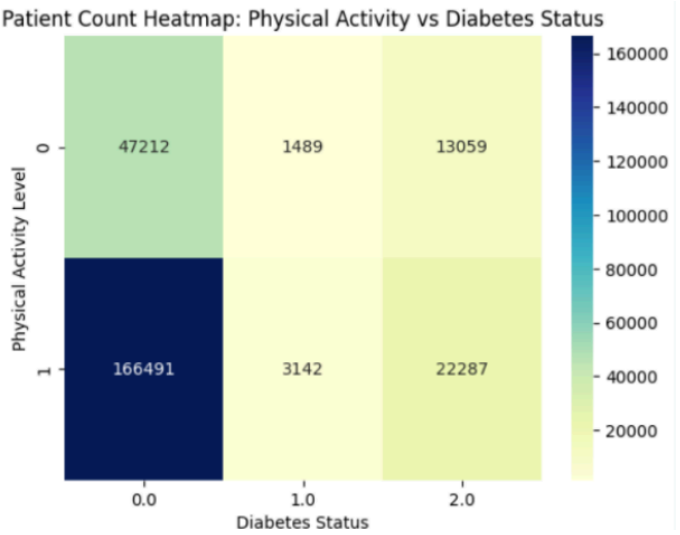


Patients with high blood pressure who also have diabetes.



How physical activity is related to diabetes





Conclusion

Physically Inactive People:

- The highest number of diabetic patients was observed among individuals who do not engage in physical activity (PhysActivity=0).

Vulnerable Age Groups:

- Older age groups (particularly 65+) have a significantly higher prevalence of diabetes, especially among those who do not engage in physical activity.
- Younger age groups who are physically active (PhysActivity=1) show much lower diabetes levels, underscoring the preventive role of physical activity.

Impact of Physical Activity:

- Physical activity (PhysActivity=1) reduces the risk of diabetes regardless of age, as evident from comparisons between active and inactive groups.

Prevalence of Diabetes Categories:

- Category Diabetes_012=2 (Type 2 Diabetes) is the most prevalent among all patient groups, particularly among older, physically inactive individuals.

Recommendations:

- Implement programs aimed at increasing physical activity levels, especially among vulnerable age groups.
- Conduct regular medical screenings for older individuals with sedentary lifestyles.
- Actively promote the benefits of physical activity for diabetes prevention.