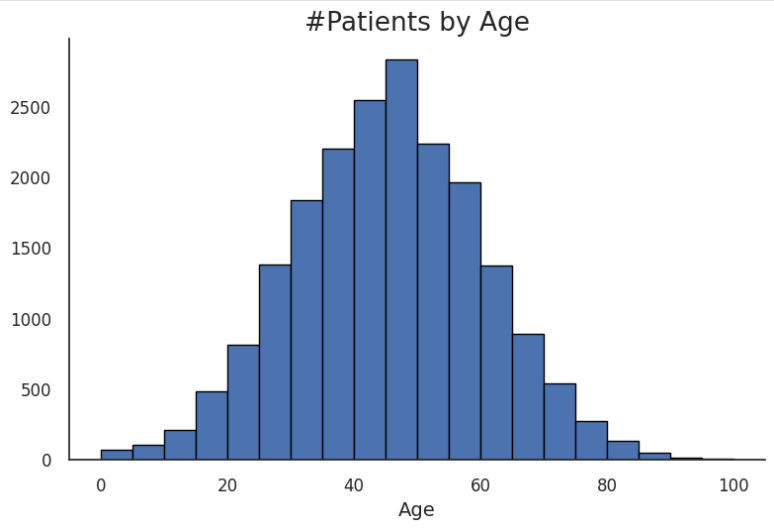
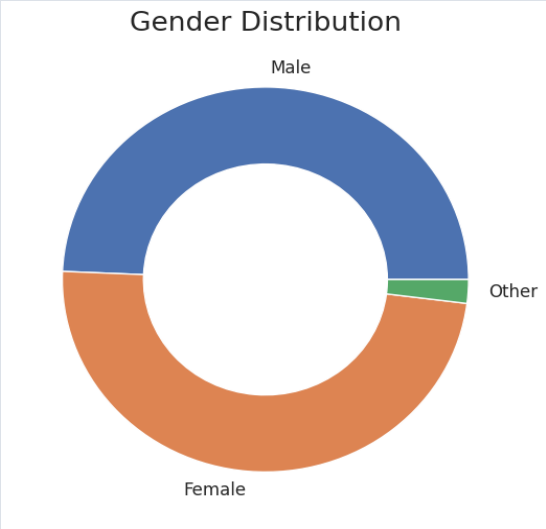
Healthcare Data Analysis Overview

Presented by Hamid Lu

June 2025

Patient Demographics



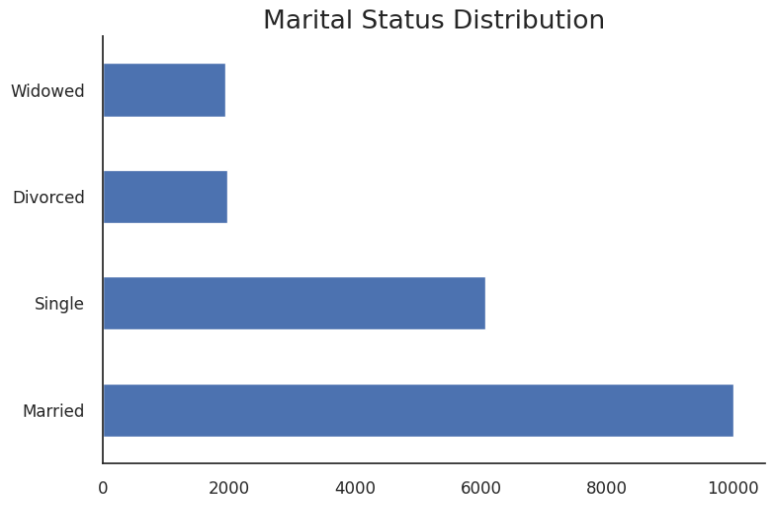


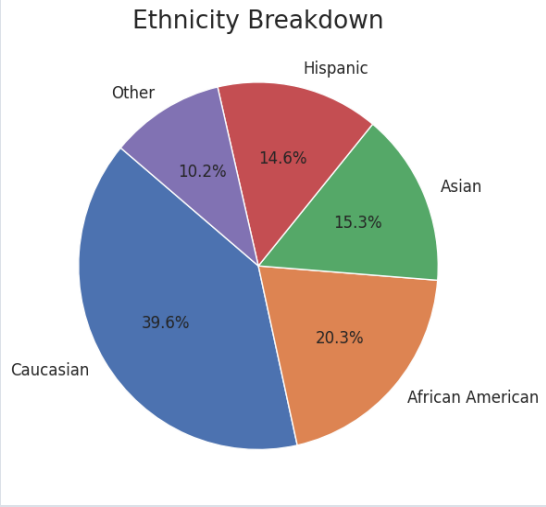
Age Distribution

- Most patients are middle-aged.

Gender Distribution

- The number of male and female patients is about equal.





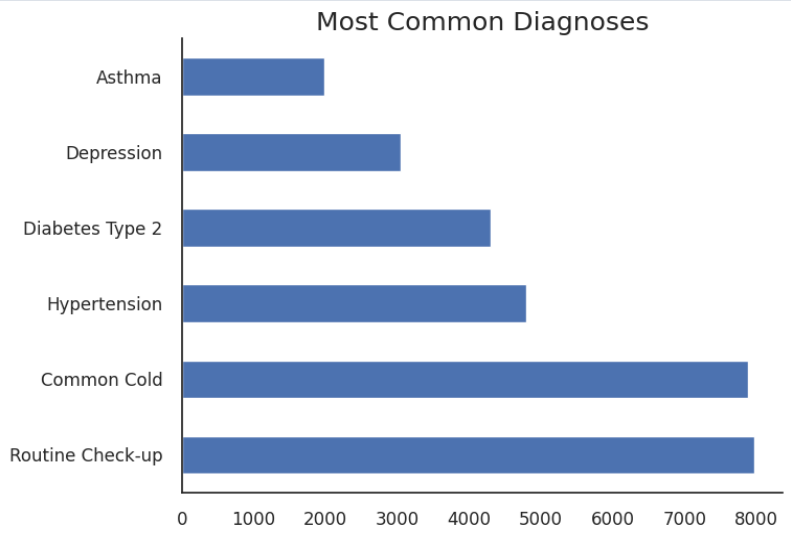
Marital Status Distribution

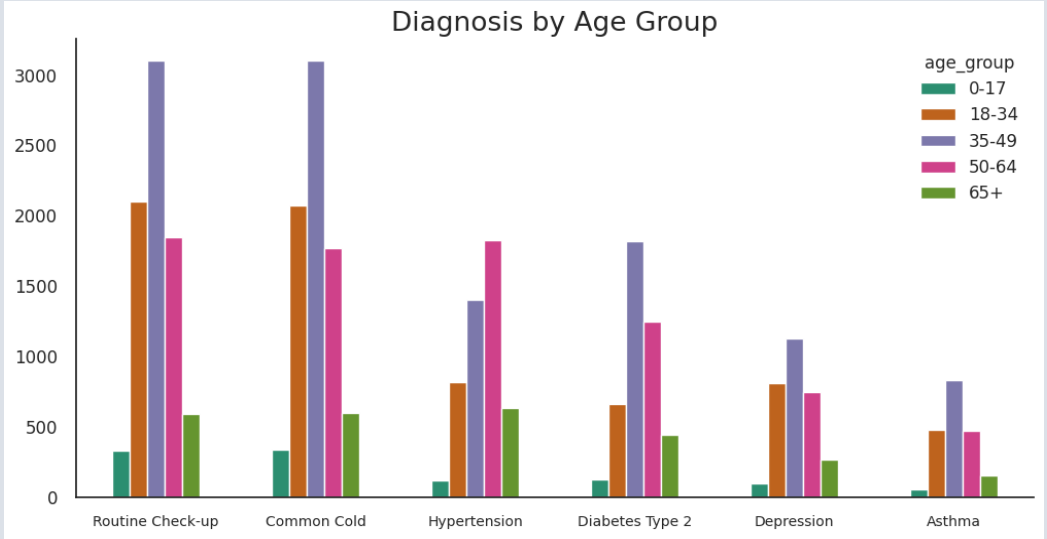
- Most patients are married.

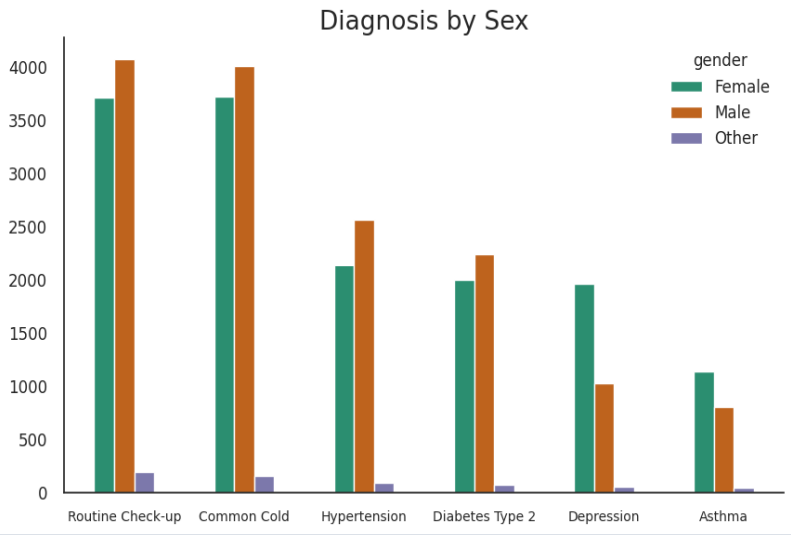
Ethnicity Breakdown

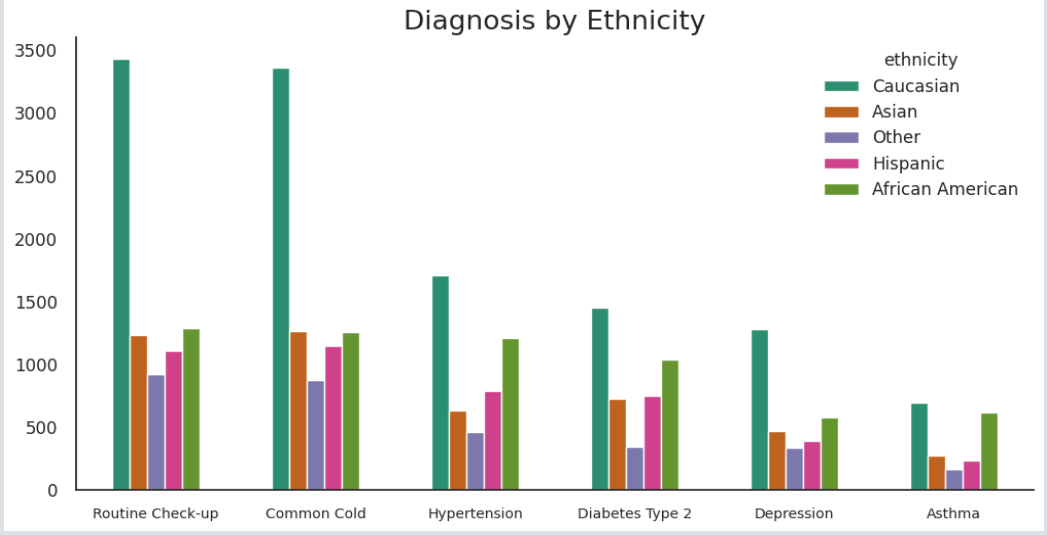
- Caucasian patients make up the largest group.
- There is a good mix of African American, Asian, Hispanic, and Other ethnicities.

How do common diagnoses vary across age, gender, and ethnicity?









Most Common Diagnoses

- Routine check-ups and common colds are the top reasons patients visit.
- High blood pressure and diabetes are also common.

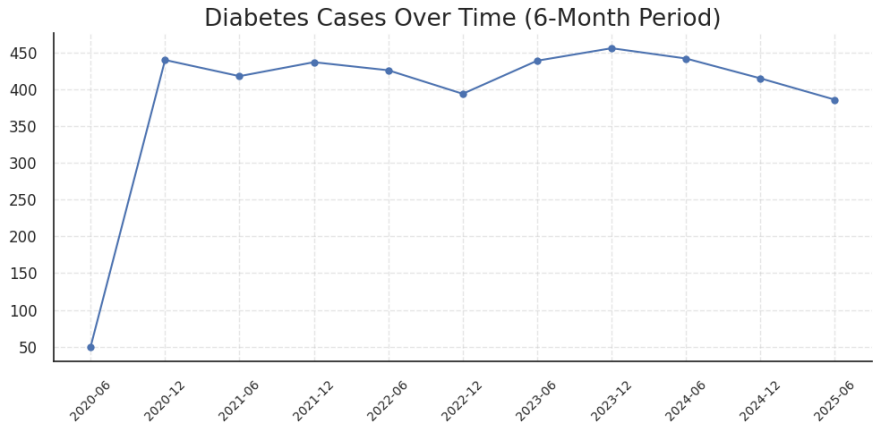
Diagnosis by Gender

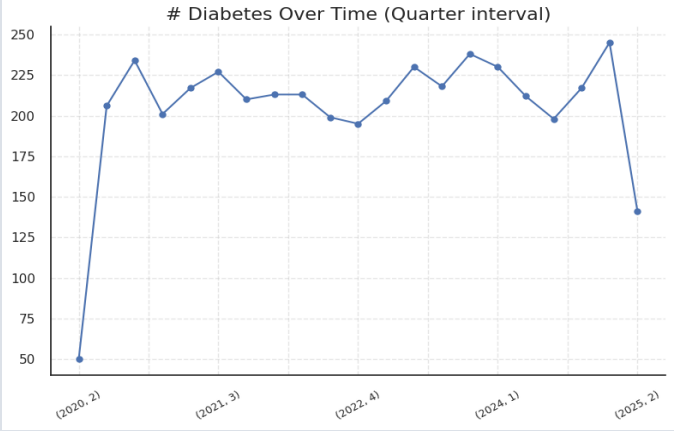
- Men have a bit more cases than women for most diseases.
- Asthma is more common in women.
- Depression shows a notably higher number of cases in females compared to males.

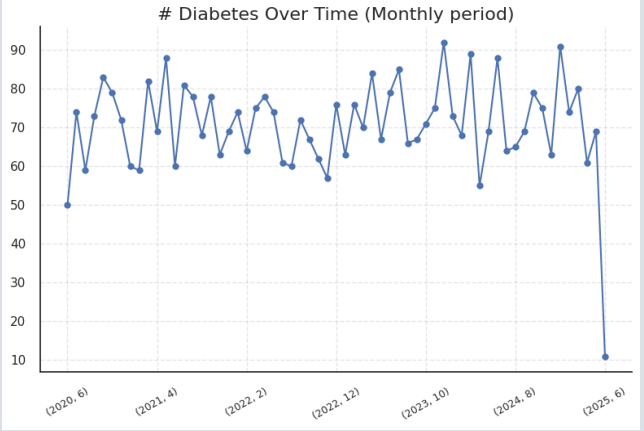
Diagnosis by Age Group

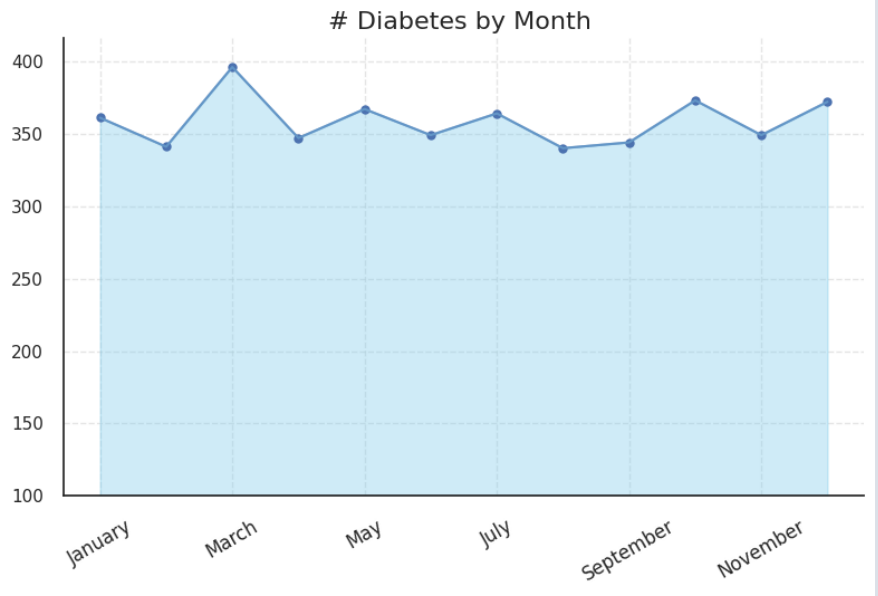
- Most cases are in people aged 35 to 64.
- High blood pressure is most common in people aged 50 to 64.
- Young people get colds more often than other diseases.

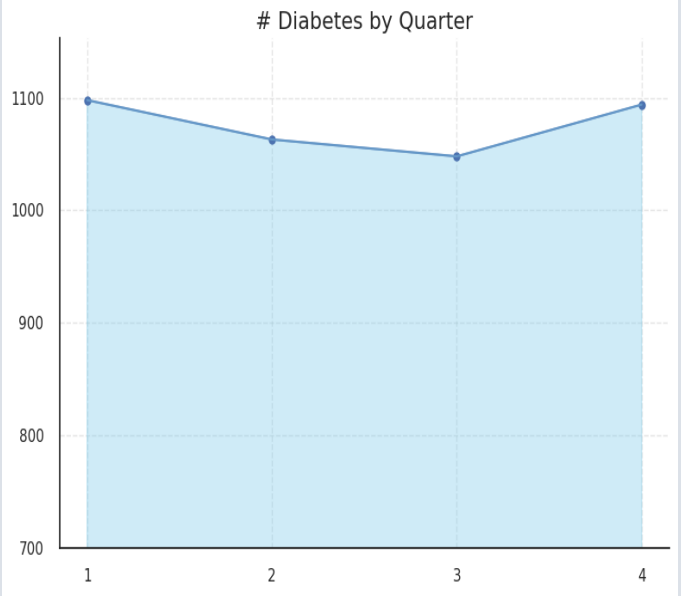
Trend of Diabetes Cases Over Time











Time Trend (Yearly & 6-Month)

- Diabetes cases have slightly dropped in recent years.
- The drop is not sharp but shows a gradual downward trend.

Seasonality – Quarterly

- Q3 (July–Sep) shows slightly more cases than other quarters.
- No major seasonal spikes across other quarters.

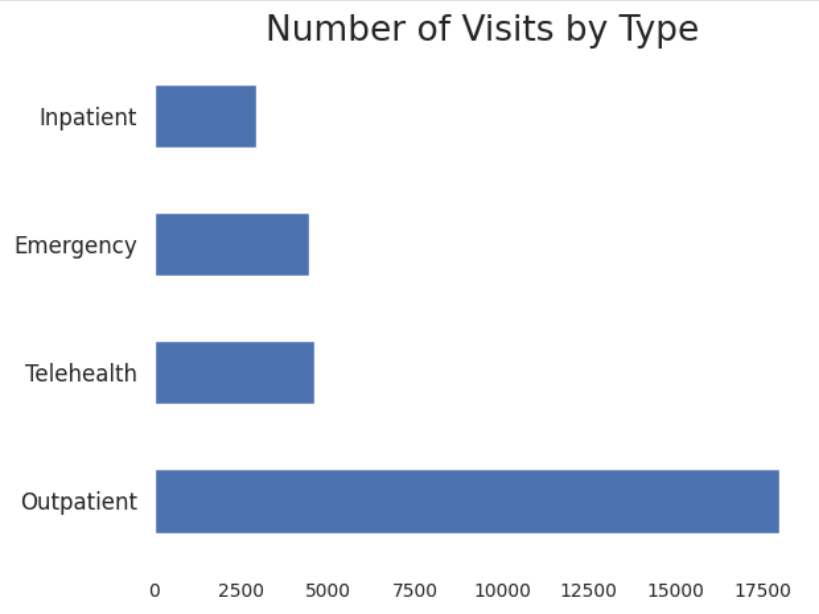
Seasonality – Monthly

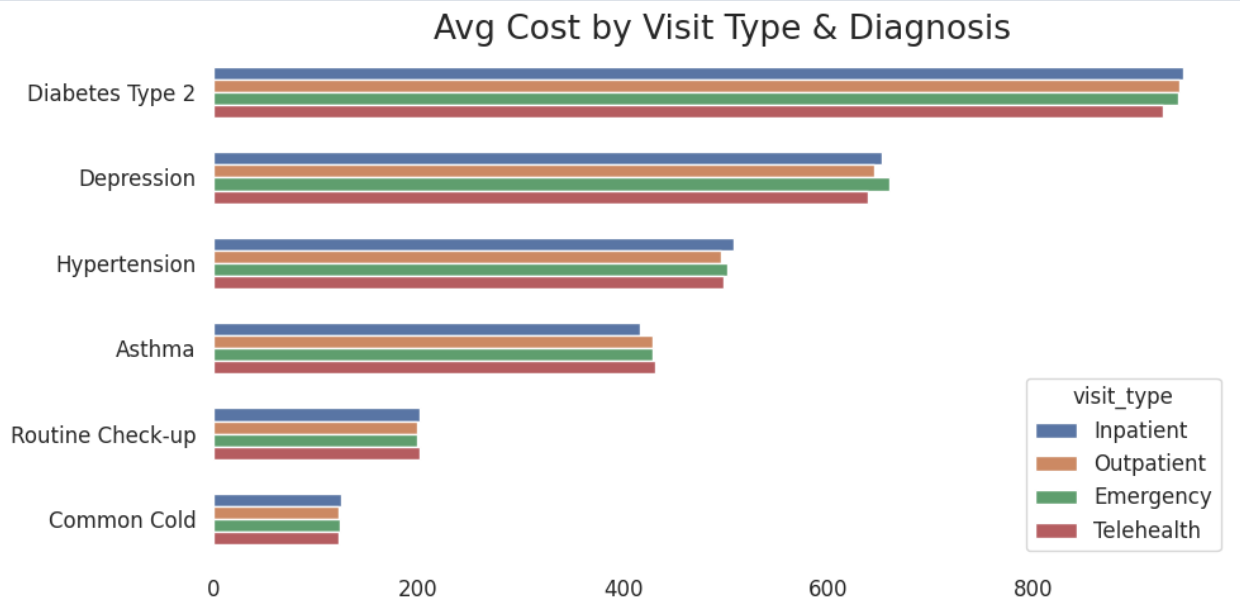
- Some ups and downs across months, but no strong pattern.
- No month stands out with a sharp rise or drop.

Summary

The prevalence of diabetes has slightly decreased over the past few years, with no strong seasonal patterns, except for a small rise in Q3.

What visit types are most common?
How much do they typically cost by condition?





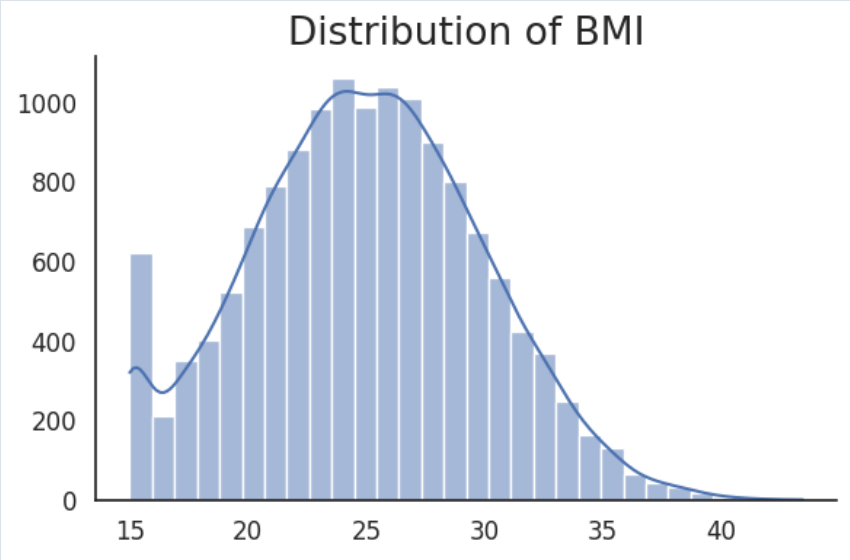
Visit Type Distribution

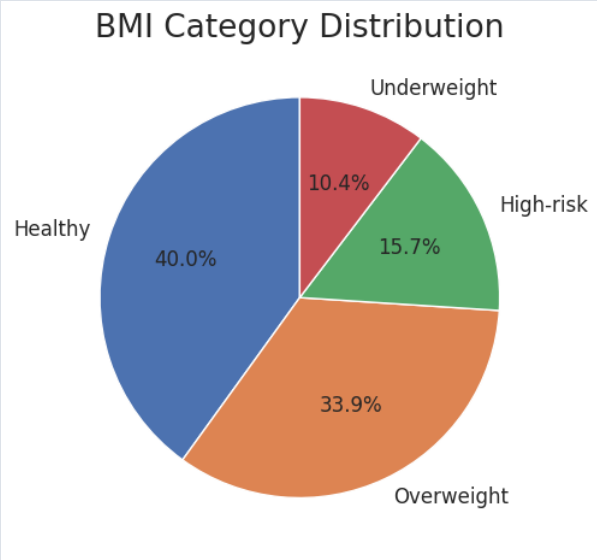
- Outpatient visits are most common.
- Telehealth, Emergency, and Inpatient visits are less frequent.

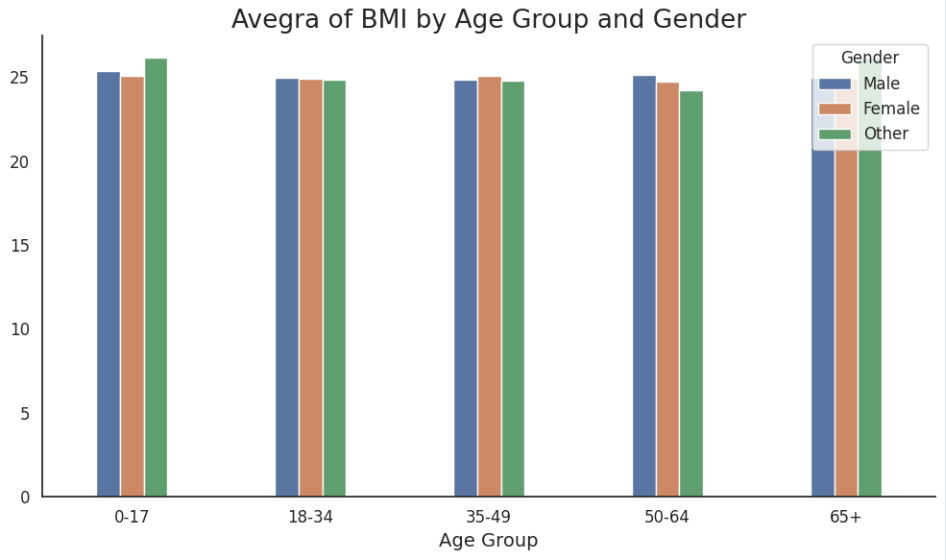
Cost Analysis

- Diabetes and Depression visits cost the most on average.
- Cost is fairly similar across visit types — no big price gap.

BMI: What percent of patients fall into healthy, overweight, or high-risk categories? How does BMI vary by age or gender?







BMI

- Most patients have BMI between 22 and 30, which is in the normal to overweight range.
- There's a small spike around 16–17, suggesting a group of underweight patients.
- The distribution is slightly right-skewed — a few patients have very high BMI.

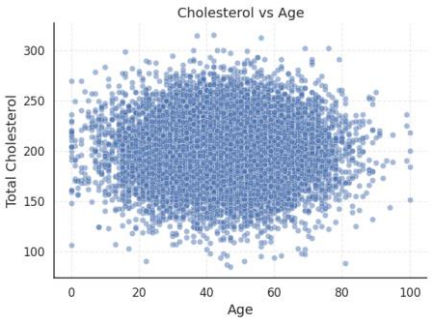
BMI Category Distribution

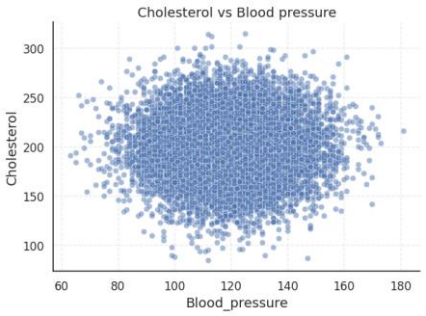
- Most patients fall in the Healthy or Overweight range.
- About significant 16% patients are in the High-risk (Obese) category.
- About 10% are Underweight.

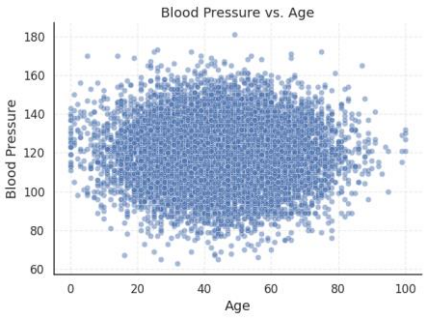
BMI Category Distribution

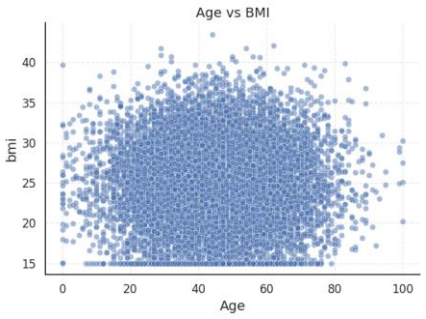
- BMI values are very similar across all age groups and genders.
- There’s no significant variation between males, females, or others.
- Age does not appear to influence average BMI in this population.

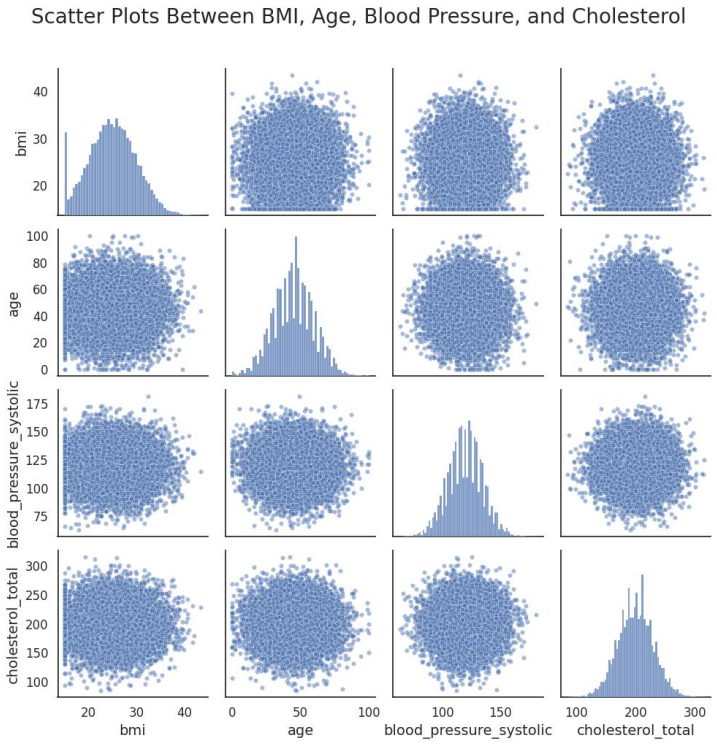
Are there any strong correlations between patients' BMI, age, blood pressure, and cholesterol?

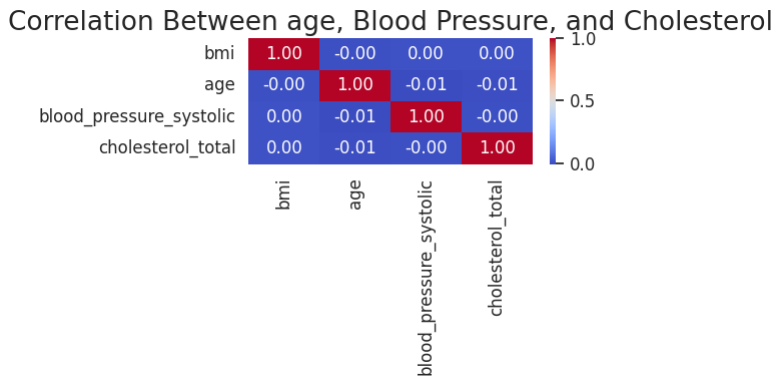






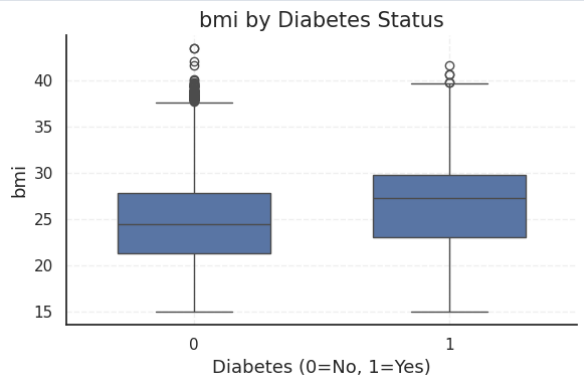


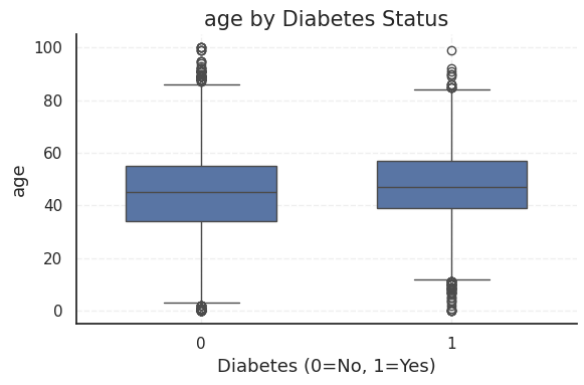




- No strong link between Age, blood pressure, or cholesterol.
- Scatterplots show random spread, no clear trends.
- Correlation heatmap confirms very weak or no correlation.

What are the Risk Factors for certain diseases (Diabetes)?





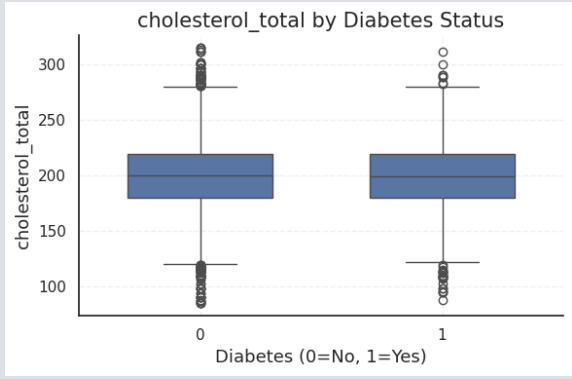
Each boxplot compares the distribution of a specific feature between patients with (1) and without (0) diabetes:

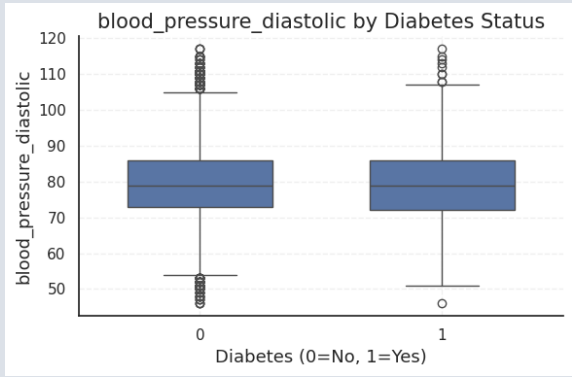
BMI

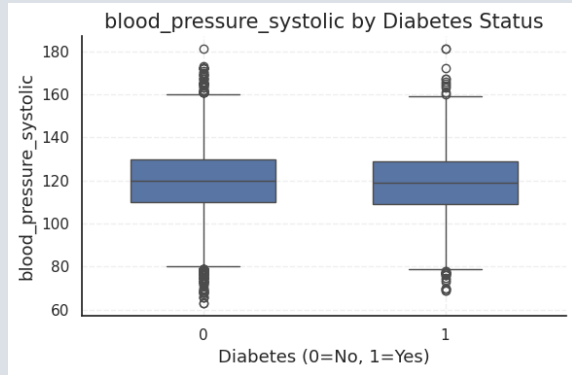
- Visually, diabetic patients tend to have slightly higher BMI values.
- The boxes (interquartile range) for diabetics are shifted up compared to non-diabetics.
- There's also a slightly higher maximum and median BMI for diabetics.

Age

- Diabetics appear to be slightly older on average.
- There is a visible upward shift in the median and interquartile range for age in diabetics.







Cholesterol & Blood Pressure (Systolic and Diastolic)

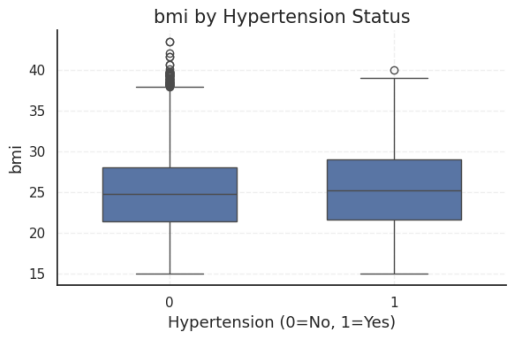
- The boxplots show similar distributions, medians, and spreads, suggesting these features may not have a clear relationship with diabetes based on visual inspection alone.

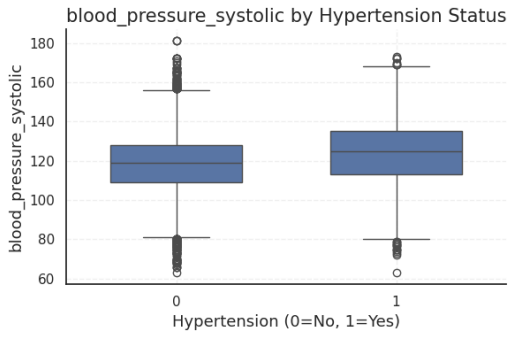
BMI and other Risk Factors for Diabetes

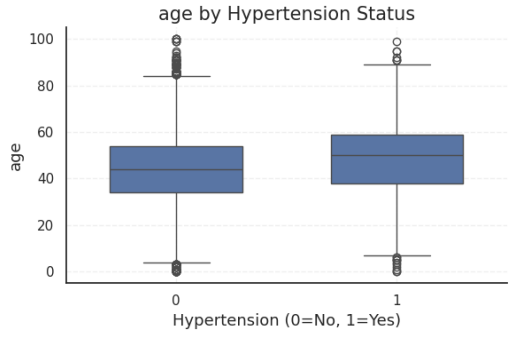
- **BMI and age show statistically significant associations** with diabetes at the 95% confidence level ($p < 0.05$), confirming what we observed in the boxplots.
- **Systolic blood pressure** also plays a minor role. It has weak but statistically significant association
- **Cholesterol and diastolic blood pressure do not show significant effects** ($p > 0.3$), indicating weak or no relationship with diabetes based on both visuals and statistical tests.

- People with **higher BMI and older age** are more likely to have diabetes — this is confirmed both by charts and statistical testing.
- People with **Systolic blood pressure** may have some level of diabetes.
- **Cholesterol and lower blood pressure** don’t seem to make much difference in whether someone has diabetes or not.

What are the Risk Factors for Patients with Hypertension?







Systolic Blood Pressure

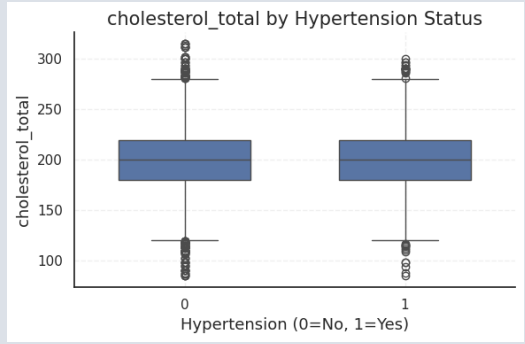
- As expected, visually, Hypertension patients tend to have higher Systolic Blood Pressure values.

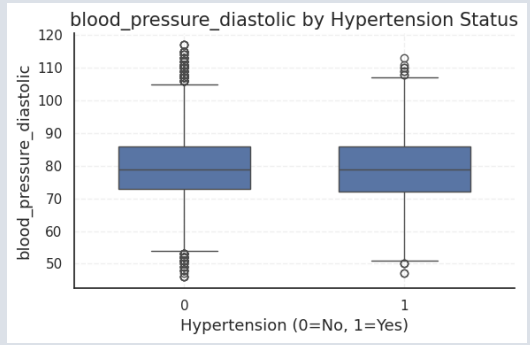
BMI

- Visually, Hypertension patients tend to have slightly higher BMI values.
- The boxes (interquartile range) for diabetics are shifted up compared to non-Hypertensions.

Age

- Hypertension appear to be slightly older on average.
- There is a visible upward shift in the median and interquartile range for age in Hypertensions.





Cholesterol & Blood Pressure (Diastolic)

- The boxplots show similar distributions, medians, and spreads, suggesting these features may not have a clear relationship with Hypertensions based on visual inspection alone.

How we predict if a patient has diabetes based on health data?

Feature Selection

We started with 7 features;

- *BMI, Age, Systolic BP, Diastolic BP, Gender, Marital Status, Ethnicity*

Final selected features:

- BMI
- Age
- Systolic Blood Pressure

Why these?

- Because they showed a strong relationship with diabetes in our analysis.

What was the strategy?

Step 1: Visual and statistical checks

- We removed 3 features: **Gender, Marital Status, Ethnicity**
  ✕ No clear visual or statistical relationship with diabetes.

Step 2: Boxplots and T-tests

- We compared diabetic vs. non-diabetic groups.
- Checked for **normal distribution** and **p-values** to keep valid feature

How we predict if a patient has diabetes based on health data?

Target and Features

- **Target**
  has_diabetes (0 or 1)
- **Selected features**
  bmi, age, blood_pressure_systolic

Model Selection

We tested two models:

- **Logistic Regression:** Simple, interpretable, Binary target
- **XGBoost:** More powerful for numeric data and imbalanced classes

Imbalanced Data

- Our data was imbalanced (fewer diabetic cases), so we used **resampling methods.**

How we predict if a patient has diabetes based on health data?

Results & Takeaways

Metric	Logistic	XGBoost
Accuracy	69%	73%
Recall	62%	78%

Conclusion

- **XGBoost performed better**, especially in detecting diabetic cases.
- **BMI, Age, and Systolic BP** are the strongest predictors.
- Logistic regression is still a valid backup model.
- The model works well and can be improved with more features or tuning.
- We could add lifestyle or family history data for study improvement

Recommendations

I. **Support healthy BMI through weight and nutrition care**
BMI strongly affects conditions like diabetes and hypertension. While it's steady across age and gender, many patients are outside the healthy range. Offer targeted weight management and nutrition programs, especially for those at higher risk..

I. **Expand mental health and asthma programs, especially for women**
Our analysis showed these issues are more frequent among female patients. Targeted programs can improve outcomes and reduce repeat visits.

II. **Make check-ups easier to access**
Outpatient visits are by far the most common. Keeping routine care accessible helps with early diagnosis and long-term cost reduction.

III. **Promote regular blood pressure monitoring and control**
Higher blood pressure is linked to both hypertension and a greater risk of diabetes. Encourage routine checkups and early detection programs, especially for older adults. Provide support for managing systolic pressure through lifestyle changes and medication when needed.

IV. **Watch Q3 (Summer) for seasonal diabetes spikes**
There's a slight increase in diabetes cases during Q3. It may relate to summer lifestyle changes (vacations, inactivity, diet). Monitoring this can guide preventive campaigns or seasonal health messaging.