Data Exploration Final Project

Paula Lozano Gonzalo

April 1, 2025

1 Introduction

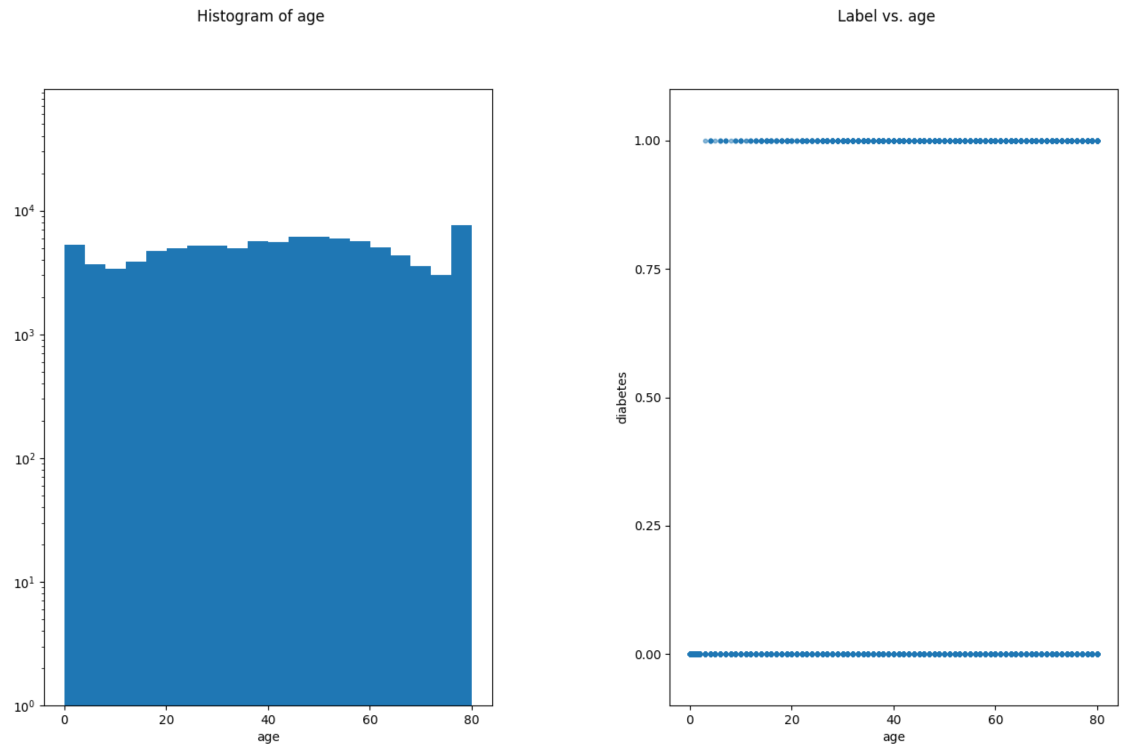
This project aims to predict diabetes based on various health indicators and demographic factors. The dataset contains medical records of patients including their age, BMI, blood glucose levels, and other relevant features.

2 Data Exploration

2.1 Loading the Data

The dataset was downloaded from kaggle: `https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows` I had already briefly looked at the features for the project idea so now let's look even deeper by plotting the data.

2.2 Feature Analysis

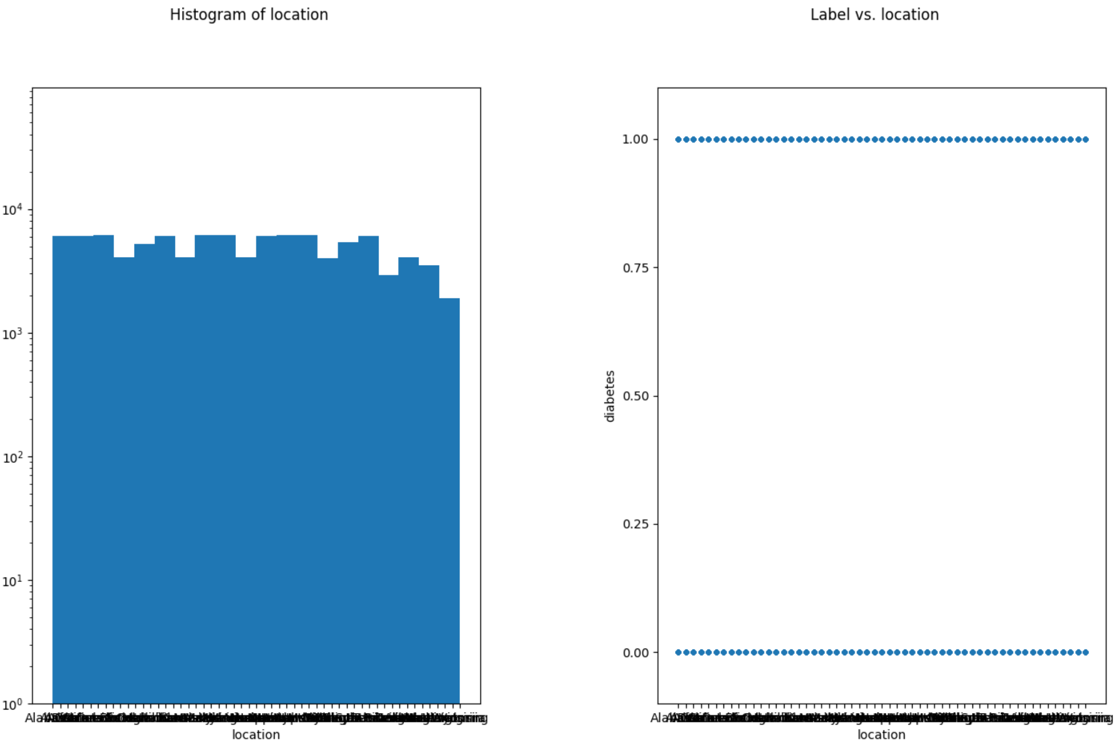


(a) Age distribution

(b) Age vs. diabetes

Figure 1: Age analysis

From the visual analysis, age doesn't appear to be strongly associated with diabetes, evidenced by the evenly distributed data points across all ages.

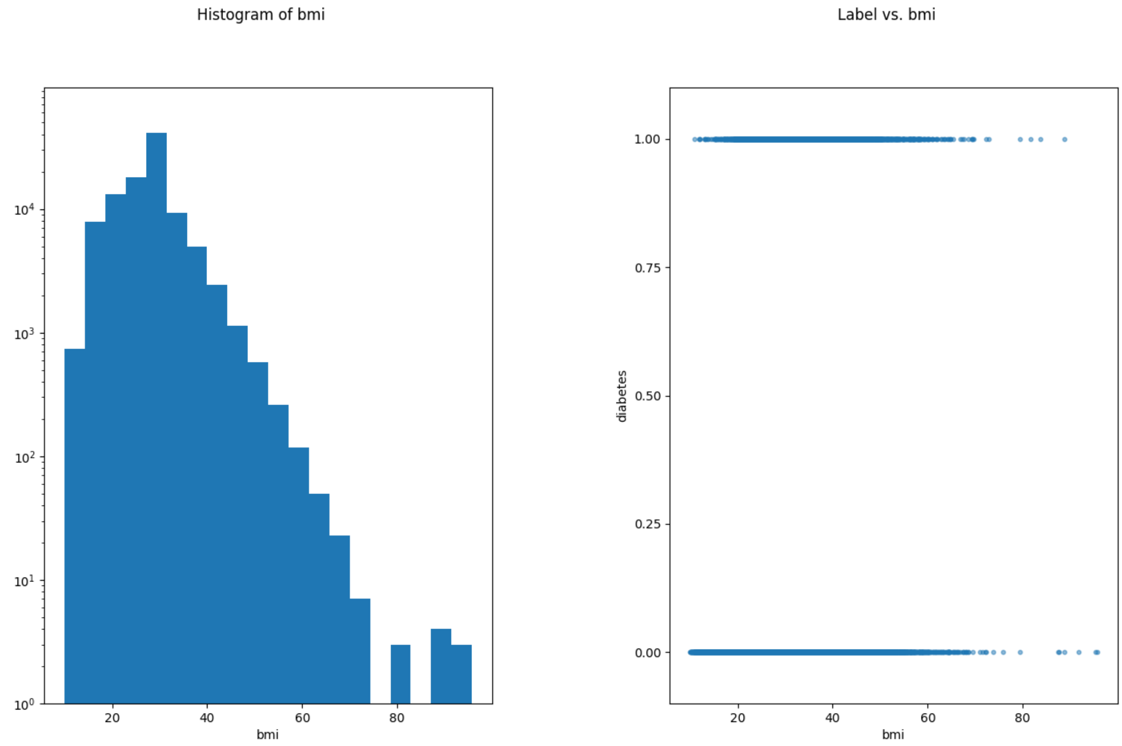


(a) Location distribution

(b) Location vs. diabetes

Figure 2: Location analysis

There are so many states that they all go over each other on the bottom of the plots, that is why it looks so messy. I don't see any trends regarding with the location and I also think there shouldn't really be any strong relationship with the label.

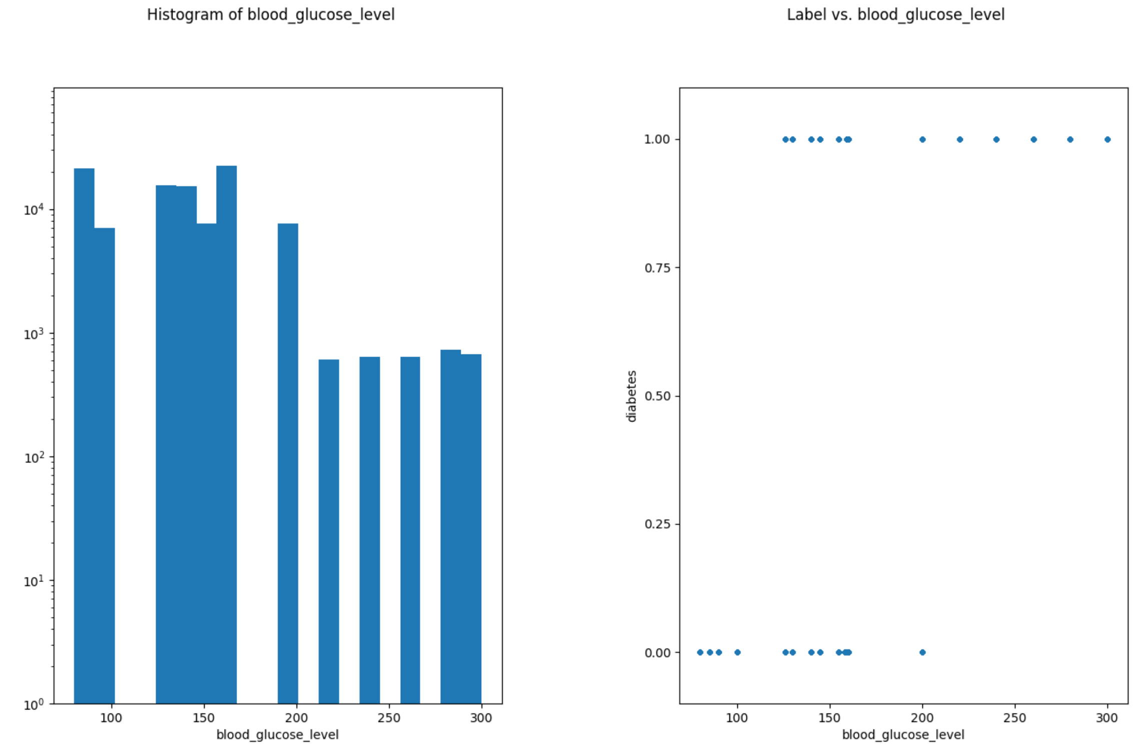


(a) BMI distribution

(b) BMI vs. diabetes

Figure 3: BMI analysis

The data shows a wider BMI distribution compared to other factors, yet diabetes cases appear randomly distributed without concentration at higher BMI levels like I was expecting.



(a) Blood glucose level distribution

(b) Blood glucose vs. diabetes

Figure 4: Blood glucose analysis

I did some research when doing the project idea and found that:
Blood_Glucose_Level: Fasting blood sugar level measurement:

- Normal: Less than 100 mg/dL
- Prediabetes: 100-125 mg/dL
- Diabetes: 126 mg/dL or higher

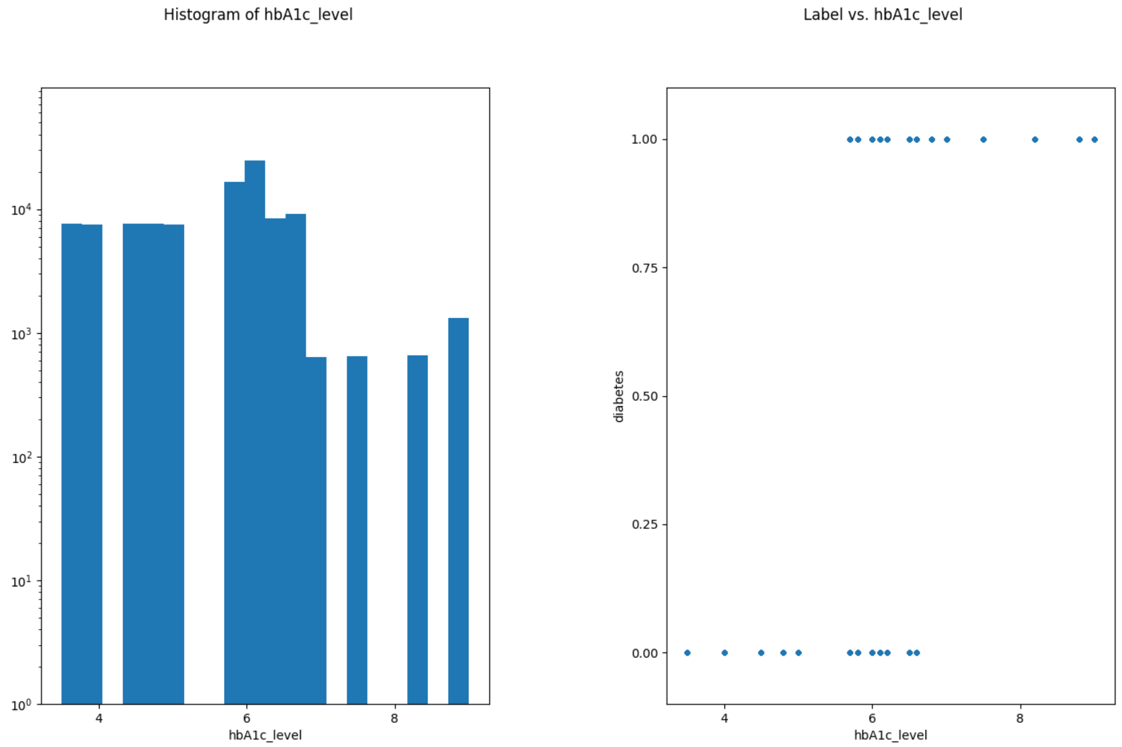
So I am expecting that all the higher glucose levels are diabetic. Looking at the scatter plot you can see that in the normal glucose level (< 100) nobody is diabetic which is fantastic, in the prediabetes range (100-125) there are some diabetic and in the diabetes range (126 $>$) some are diabetic but some others aren't. The scatter plot reveals three notable observations that both confirm and challenge expectations:

Clear Normal Range: No diabetes cases appear below 100 mg/dL, perfectly aligning with clinical standards. (No outliers)

Prediabetes Paradox: In the 100-125 mg/dL range shows some diabetes cases.

Diagnostic Ambiguity: Above 126 mg/dL, we find mixed results—while many cases confirm diabetes diagnosis, a significant portion of high-glucose individuals don't show diabetes.

This suggests that while blood glucose level serves as an excellent negative predictor (ruling out diabetes below 100 mg/dL), its predictive power weakens in higher ranges where other factors may influence outcomes.



(a) HbA1c level distribution

(b) HbA1c vs. diabetes

Figure 5: HbA1c analysis

HbA1c (glycated hemoglobin) reflects average blood glucose levels over approximately 3 months, with established diagnostic thresholds:

HbA1c_level: This measures average blood sugar levels over the past 2-3 months. Higher values indicate poorer blood sugar control:

- Normal: Below 5.7%
- Prediabetes: 5.7% to 6.4%
- Diabetes: 6.5% or higher

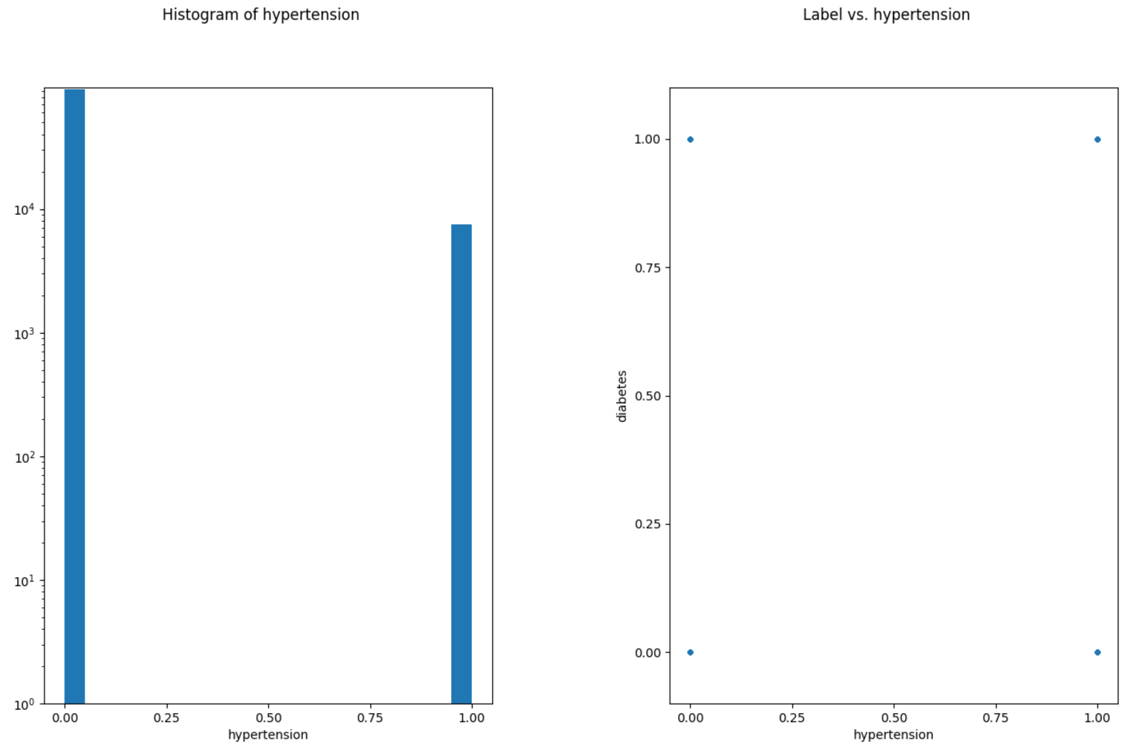
The data exhibits strong alignment with clinical expectations:

Definitive Normal Range: All subjects below 5.7% were non-diabetic, confirming the threshold's reliability for ruling out diabetes.

Transition Zone: The prediabetes range (5.7%–6.4%) shows expected diagnostic overlap, containing both diabetic and non-diabetic cases.

Diagnostic Certainty: All values $\geq$ 6.5% correctly identified diabetic cases without exceptions.

This pattern mirrors but improves upon the blood glucose observations, demonstrating HbA1c's superior diagnostic clarity - particularly in maintaining perfect specificity at the diabetes threshold. The clean separation at 6.5

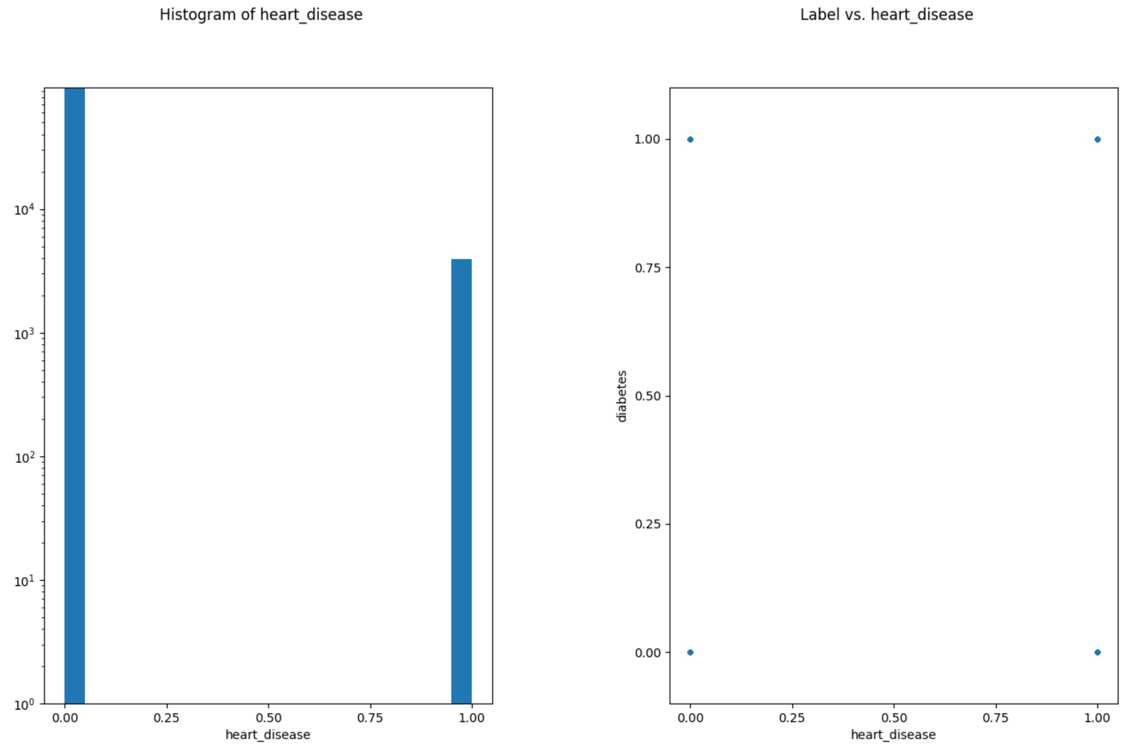


(a) Hypertension distribution

(b) Hypertension vs. diabetes

Figure 6: Hypertension analysis

Hypertension shows no clear association with diabetes in these visualizations, as diabetic and non-diabetic cases are equally distributed among those with and without high blood pressure.

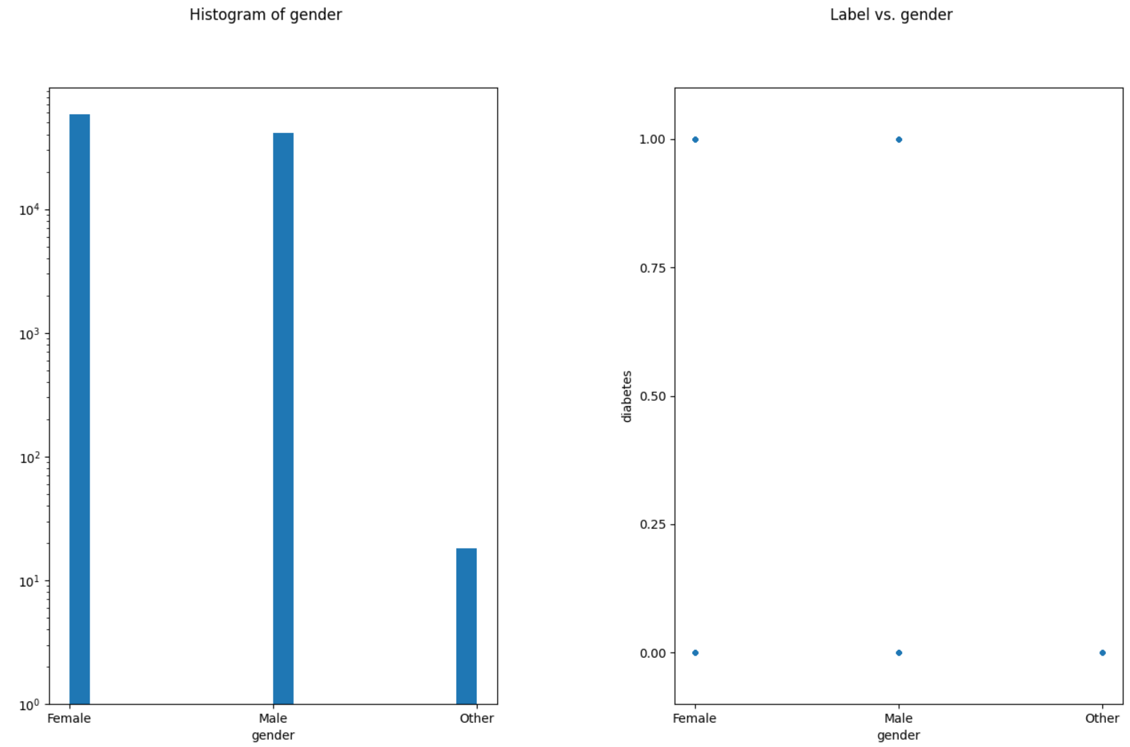


(a) Heart disease distribution

(b) Heart disease vs. diabetes

Figure 7: Heart disease analysis

The data doesn't show a strong connection between heart disease and diabetes in these visualizations. Both diabetic and non-diabetic groups include individuals with and without heart conditions. A different plotting approach might uncover subtler trends, but based on these plots alone, the relationship isn't clear.

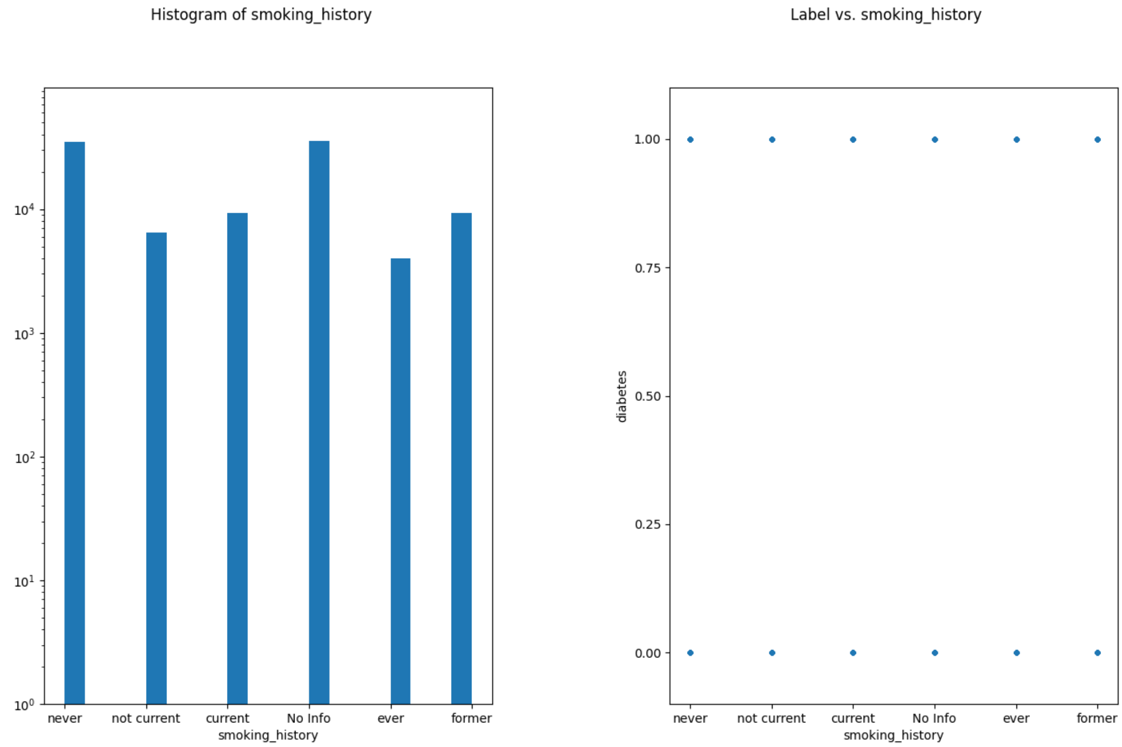


(a) Gender distribution

(b) Gender vs. diabetes

Figure 8: Gender analysis

The dataset shows an unexpected pattern where only Male and Female genders have diabetes cases, while the ”Other” gender category contains zero diabetes occurrences. This could potentially introduce bias if the model learns to associate diabetes exclusively with binary genders. However, without knowing the sample size of the ”Other” category, it’s unclear whether this reflects a true biological pattern or simply data limitations.



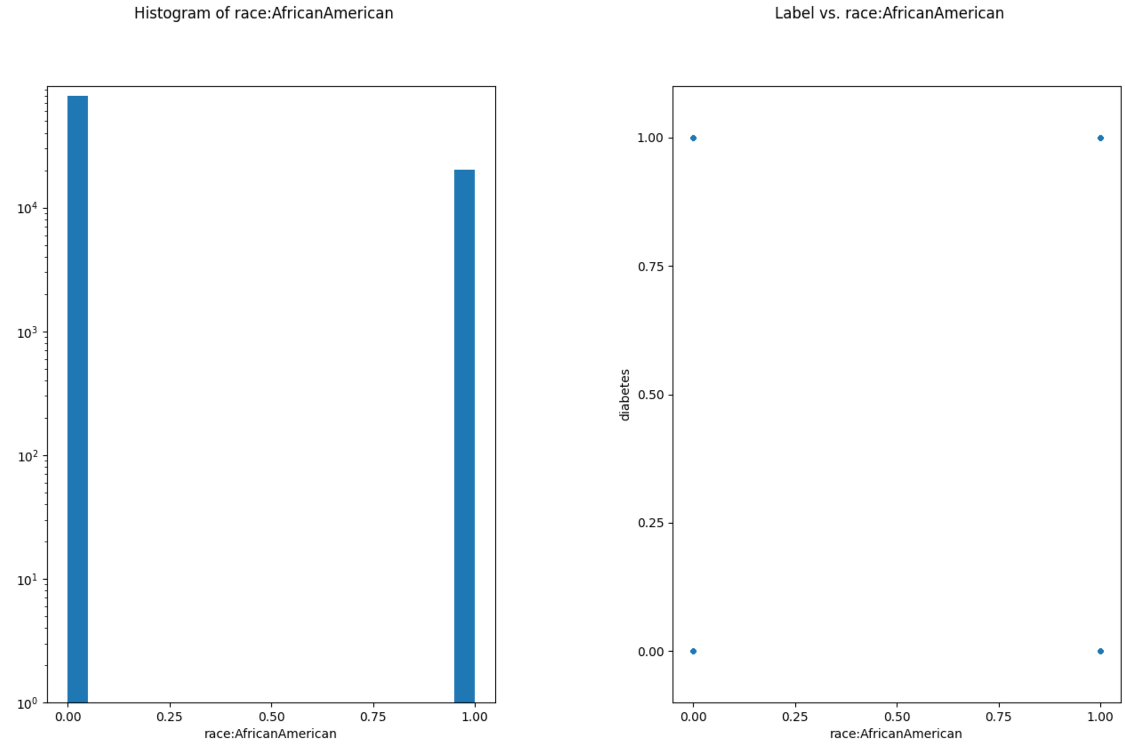
(a) Smoking history distribution

(b) Smoking history vs. diabetes

Figure 9: Smoking history analysis

Like heart disease, smoking history shows no immediate visual correlation with diabetes in these plots. Quantitative analysis of rates per smoking category might provide clearer insights than the current visualizations.

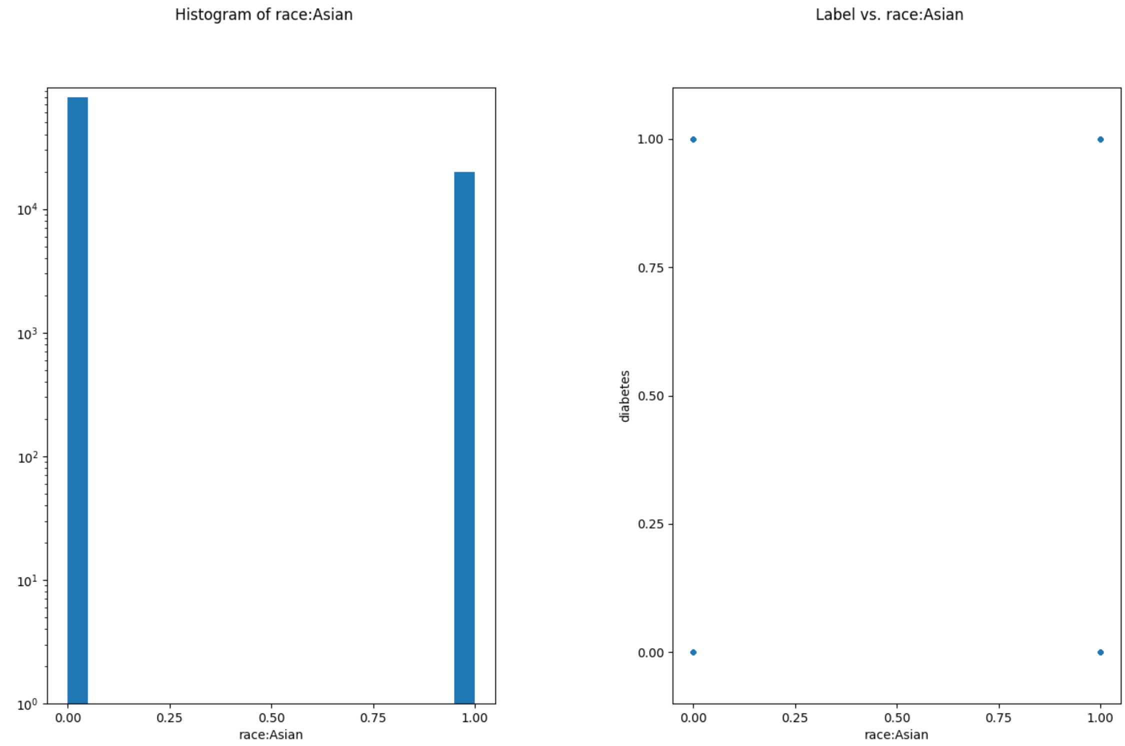
The race variable appears to be one-hot encoded already so we have five distinct categories: African American, Asian, Caucasian, Hispanic, and Other. However, the current plots don't reveal any obvious racial patterns in diabetes distribution.



(a) African American distribution

(b) African American vs. diabetes

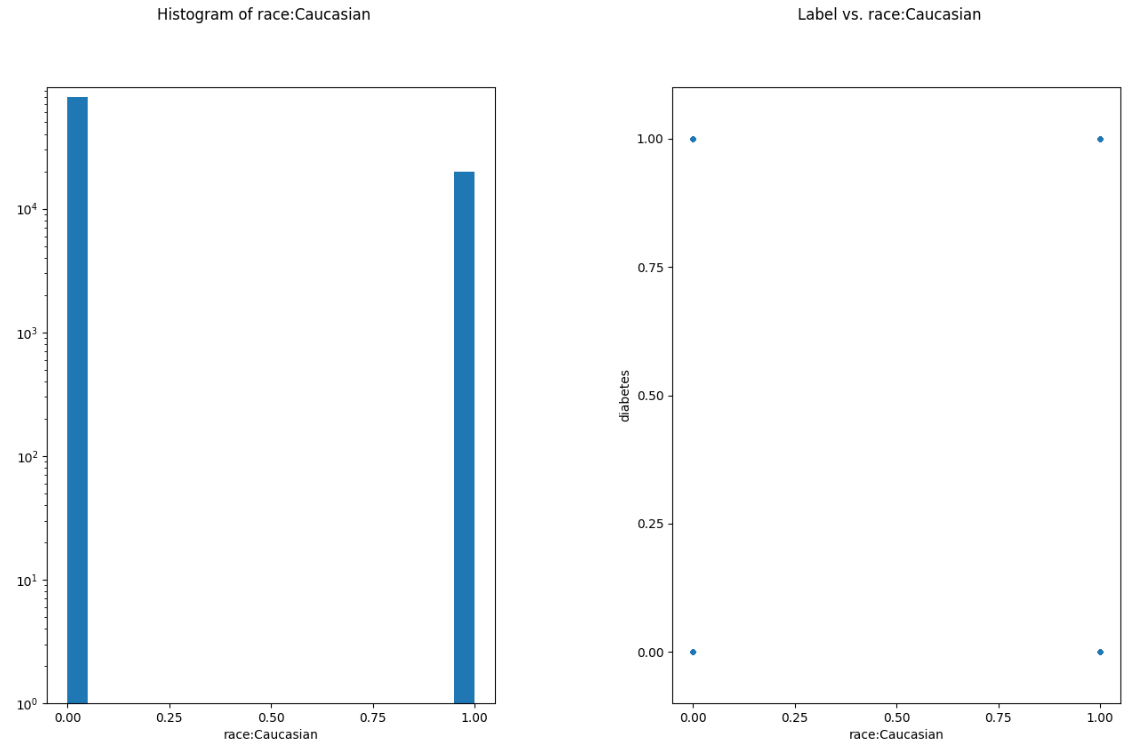
Figure 10: African American analysis



(a) Asian distribution

(b) Asian vs. diabetes

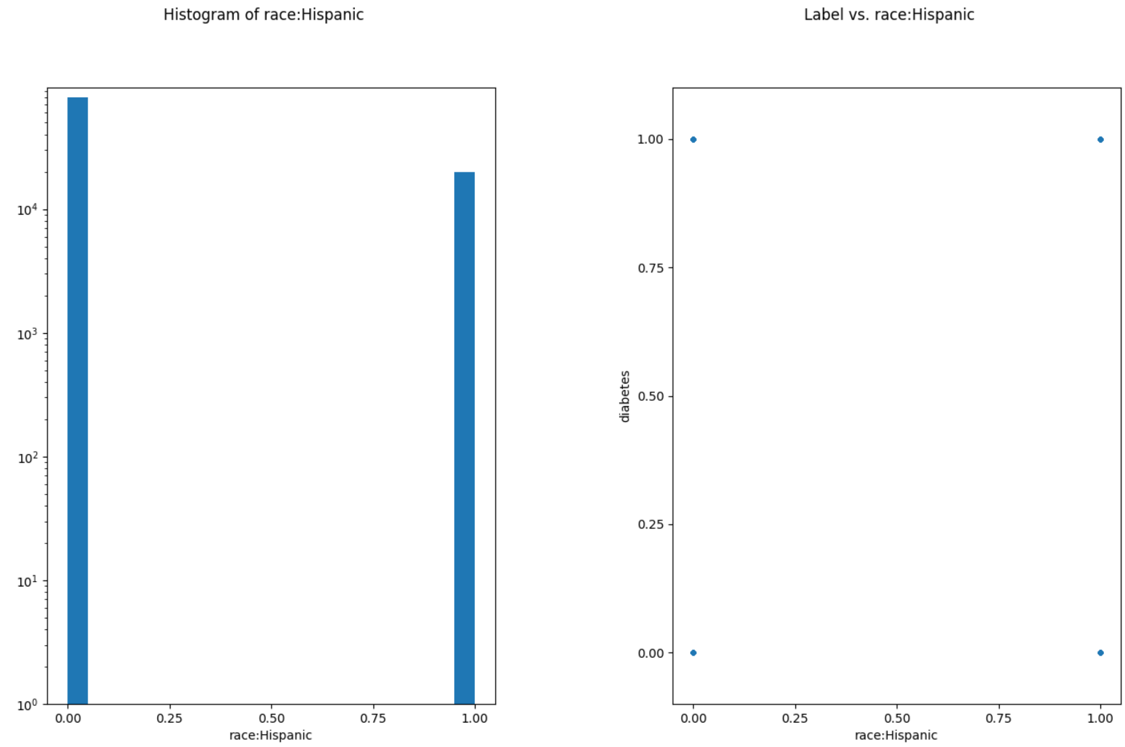
Figure 11: Asian analysis



(a) Caucasian distribution

(b) Caucasian vs. diabetes

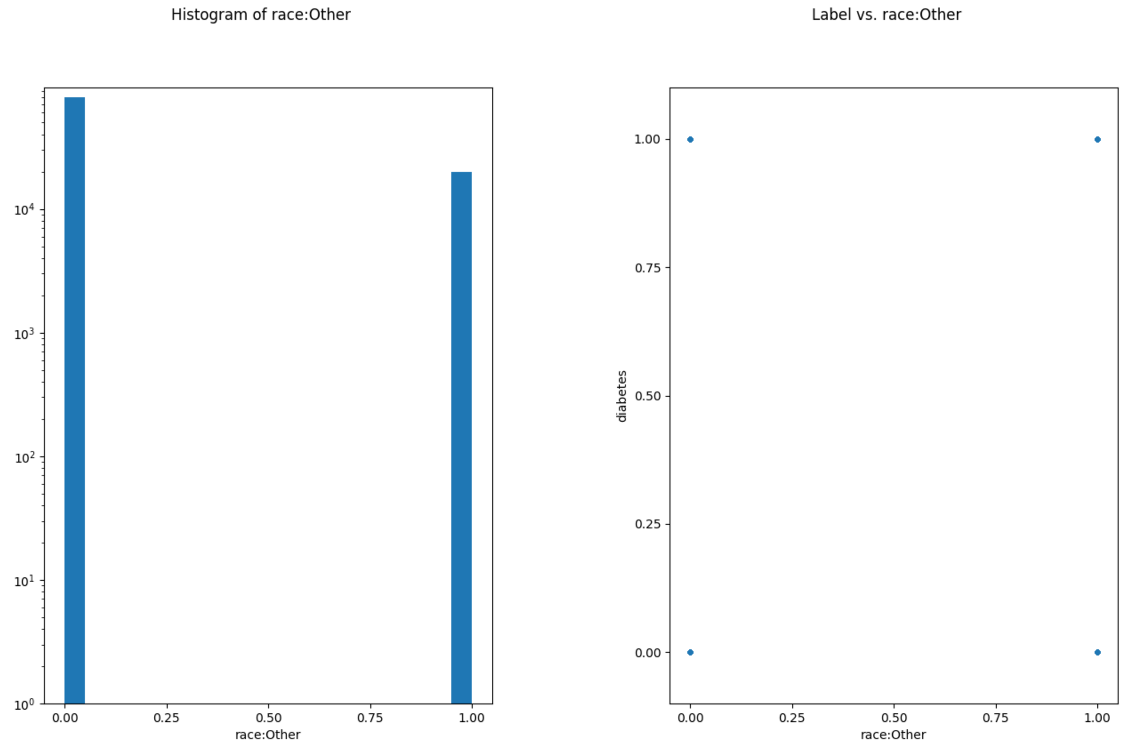
Figure 12: Caucasian analysis



(a) Hispanic distribution

(b) Hispanic vs. diabetes

Figure 13: Hispanic analysis



(a) Other races distribution

(b) Other races vs. diabetes

Figure 14: Other races analysis

3 Data Preparation

- There was no missing values (I used `grep -n ",," data/diabetes_dataset.csv` and there were no missing values). However, I still left the missing data pipeline elements just in case. In numerical the strategy is median and in categorical its constant.
- Normalized numerical features using the standar scaler.
- Encoded categorical variables using one-hot encoding
- Split data into training (70%), validation (10%), and test (20%) sets.

4 Initial Dataset Shapes

- Original data shape: (100000, 17)
- Test data shape: (20000, 17)
- Training data shape final: (72000, 17)
- Validation data shape: (8000, 17)

5 Final Dataset Shapes

1. Test preprocessed data shape: (20000, 77)
2. Train preprocessed data shape: (72000, 77)
3. Validation preprocessed data shape: (8000, 77)

Some comments about the preprocessed data: I dropped the label clinical notes because each one of them is different and there is 100,000 data so we dont want that many features as well as I don't know how the model could learn anything from it. I was thinking on dropping the location as well since that would bump up the features after one hot encoding but I decided to leave it after all. Maybe I will come back to drop it and see how much better or worse the models do.

6 Conclusion

The dataset shows clear relationships between several features and diabetes outcomes. The preprocessing steps have prepared the data effectively for modeling. Based on the exploratory analysis, this dataset appears suitable for building predictive models for diabetes.