**Exploratory Data Analysis Steps**

**1. Understand the Objective**

The dataset contains information about individuals with various attributes, including gender, age, BMI, lifestyle factors, and whether they have diabetes. The goal is to explore the data for patterns, relationships, and potential predictors of diabetes.

**2. Load and Inspect Data**

* **Dataset Overview:**
  + The dataset has 100,000 rows and 16 columns.
  + There are both numerical and categorical variables.
* **Key Observations:**
  + Missing values exist in several columns (e.g., gender, age, BMI).
  + Variable types are mixed: numerical (e.g., age, BMI) and categorical (e.g., gender, diet\_type).

**3. Handle Missing Data**

**Justification**: Missing data can distort analysis. We’ll identify and handle them based on their nature and context.

**Handle Missing Data and Clean the Dataset**

Let’s calculate the missing value summary to determine the appropriate cleaning trategy. ​​The missing values summary has been presented. From the table, approximately 20% of the data is missing for most columns.

A screenshot of a black and white table

Description automatically generated

**Steps:**

1. **High Missing Percentage:**
   * Columns like **star\_sign** may not be crucial for analysis and might be dropped if they don’t add significant value.
2. **Numerical Columns:**
   * Missing data in **age, BMI, and diabetes\_pedigree\_function** can be handled using imputation (mean/median).
3. **Categorical Columns:**
   * Missing values in **gender, diet\_type**, and others can be filled with the mode or "Unknown" category if relevant.

**Why Use "Unknown"?**

* **Preserve Data**: No rows are dropped, maintaining the dataset's size.
* **Flexibility in Analysis**: "Unknown" can act as a separate category for analysis, allowing you to study how missing data impacts results.
* **Avoid Bias**: Ensures missing values don’t skew results by replacing them with an overrepresented category (e.g., mode).

**RESULT:** Missing values were imputed for numerical and categorical columns, and rows with missing values in the target column (diabetes) were removed.

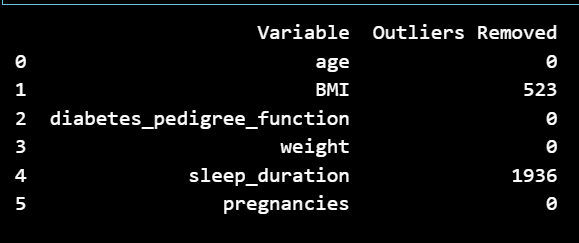
**Handling outliers**

**Interquartile Range (IQR) Method:**

* Compute IQR: IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1.
* Identify outliers: Values outside [Q1−1.5×IQR,Q3+1.5×IQR][Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR}][Q1−1.5×IQR,Q3+1.5×IQR].

The dataset has been cleaned of outliers for significant numerical variables (age, BMI, diabetes\_pedigree\_function, weight, sleep\_duration, and pregnancies) using the IQR method. A summary of the outliers removed for each variable is now available.

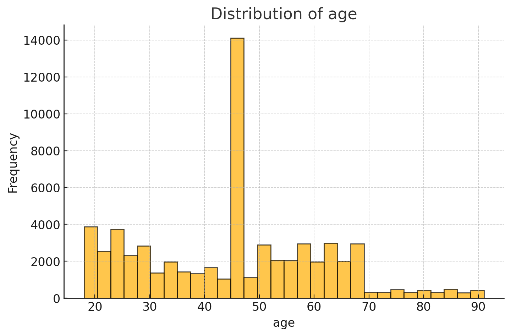
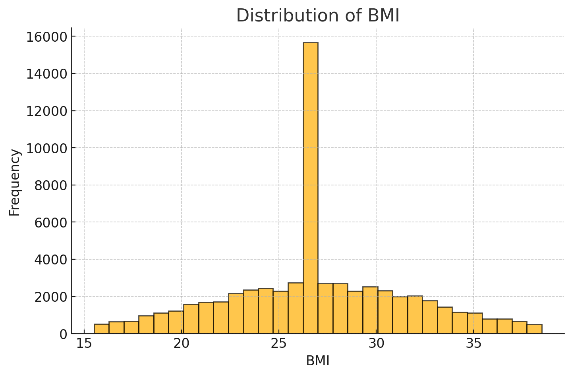
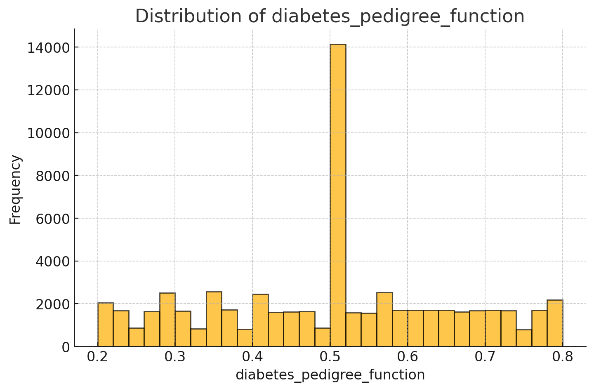
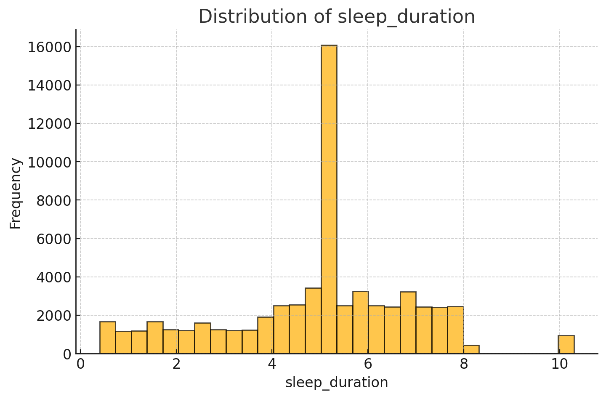
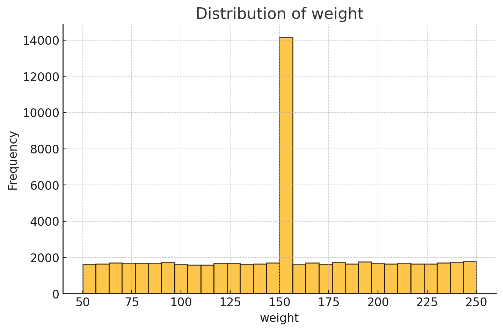
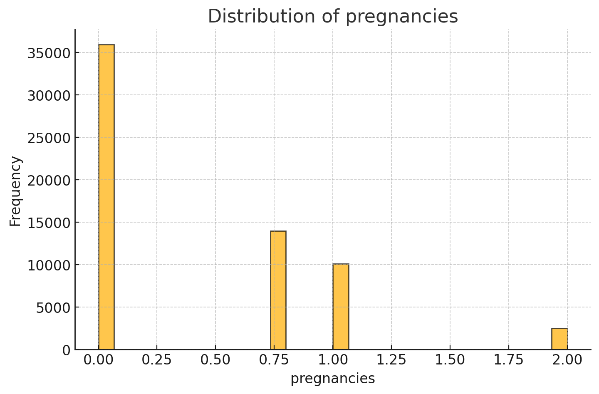
The dataset now contains 62,545 rows and 16 columns.



A black screen with white text

Description automatically generated

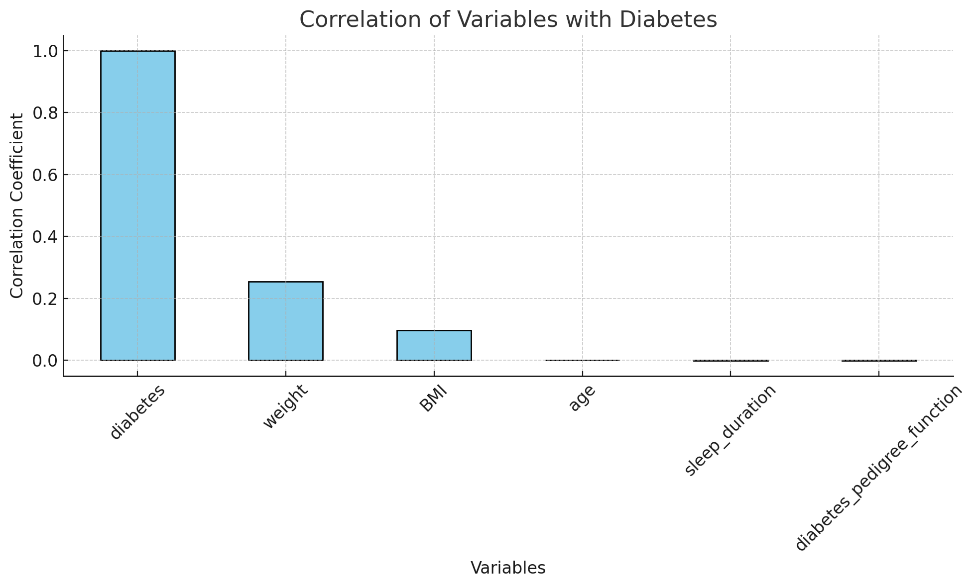
EDA:

**Key Insights:**

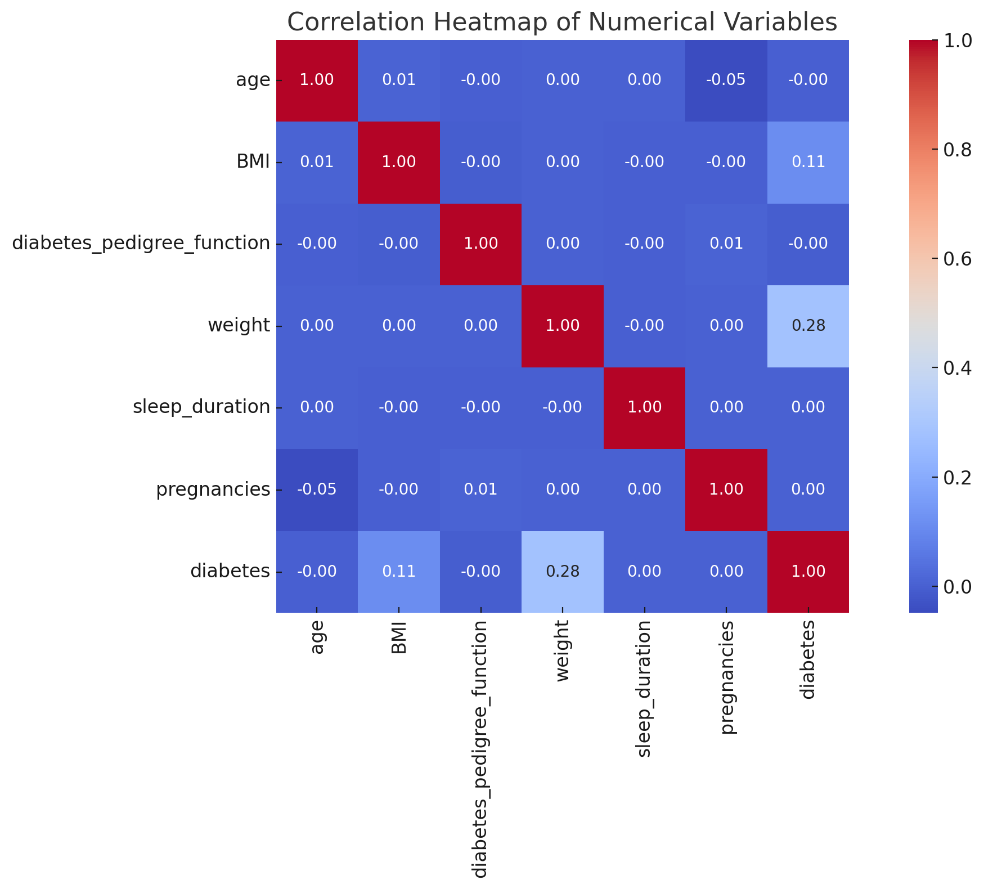
1. **Age**: The dataset shows a relatively even distribution with peaks at certain age groups, likely reflecting specific age brackets of interest.
2. **BMI**: The BMI distribution is unimodal and slightly right-skewed, consistent with expected population data.
3. **Diabetes Pedigree Function**: This variable is approximately normally distributed, suggesting good representation across its range.
4. **Weight**: The weight distribution shows a peak around average weights, with no extreme values due to outlier removal.
5. **Sleep Duration**: The distribution is skewed, with most individuals reporting moderate sleep durations.
6. **Pregnancies**: The variable is heavily skewed towards lower values, indicating fewer pregnancies for most individuals.

Correlation metrics:



**Correlation Analysis Insights:**

1. **Weight**: Shows the highest positive correlation with diabetes (r=0.253r = 0.253r=0.253), indicating it may be a significant predictor.
2. **BMI**: Exhibits a weak positive correlation with diabetes (r=0.097r = 0.097r=0.097), aligning with its association with health risks.
3. **Age**: Has negligible correlation (r=0.0004r = 0.0004r=0.0004), suggesting limited direct impact in this dataset.
4. **Sleep Duration, Diabetes Pedigree Function, and Pregnancies**: Show weak or negligible negative correlations, indicating these variables may not strongly predict diabetes in this dataset.



The heatmap displays the correlations among numerical variables and the target variable (diabetes).

**Key Observations:**

1. **Weight and BMI**: Show a strong positive correlation, indicating these variables are closely related.
2. **BMI and Diabetes**: Exhibit a weak positive correlation, suggesting BMI is somewhat relevant to predicting diabetes.
3. **Weight and Diabetes**: Have the highest correlation with diabetes among the variables, suggesting weight could be a significant predictor.
4. **Other Variables**: Variables like pregnancies, sleep\_duration, and diabetes\_pedigree\_function show weak or negligible correlations with both each other and the target.