Report

1 Exploratory Data Analysis Steps

1.1 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.

1.2 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).

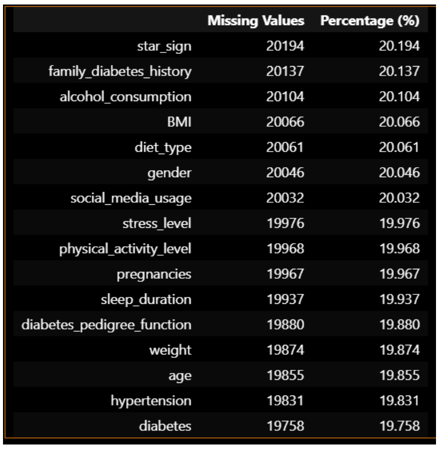
1.3 Handle Missing Data

- **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 strategy. The missing values summary has been presented. From the table, approximately 20% of the data is missing for most columns.



	Missing Values	Percentage (%)
star_sign	20194	20.194
family_diabetes_history	20137	20.137
alcohol_consumption	20104	20.104
BMI	20066	20.066
diet_type	20061	20.061
gender	20046	20.046
social_media_usage	20032	20.032
stress_level	19976	19.976
physical_activity_level	19968	19.968
pregnancies	19967	19.967
sleep_duration	19937	19.937
diabetes_pedigree_function	19880	19.880
weight	19874	19.874
age	19855	19.855
hypertension	19831	19.831
diabetes	19758	19.758

- **Steps**
  - **High Missing Percentage**

    Columns like **star sign** may not be crucial for analysis and might be dropped if they don’t add significant value.
  - **Numerical Columns:**

    Missing data in **age, BMI, and diabetes pedigree function** can be handled using imputation (mean/median).
  - **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.

---

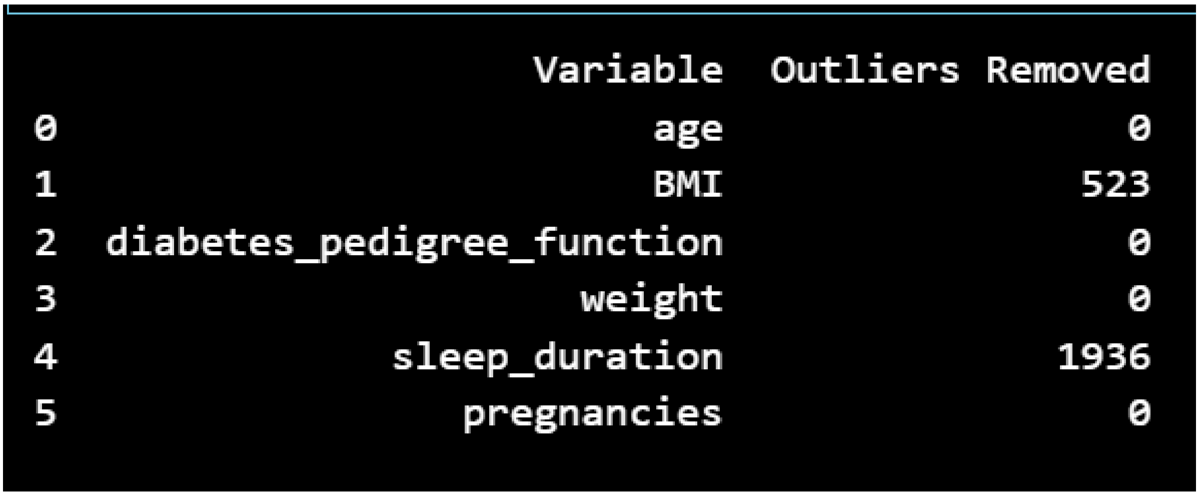
2 Handling outliers

2.1 Interquartile Range (IQR) Method:

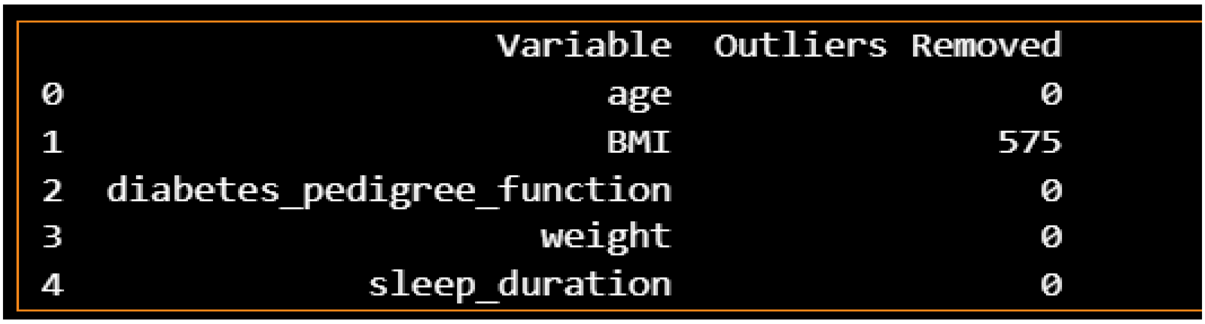
- Compute IQR: IQR=Q3−Q1\text {IQR} = Q3 - Q1IQR=Q3−Q1.
- Identify outliers: Values outside [Q1−1.5×IQR, Q3+1.5×IQR] [Q1 - 1.5 \times \ text {IQR}, Q3 + 1.5 \times \ text {IQR}] [Q1−1.5×IQR, Q3+1.5×IQR].

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.

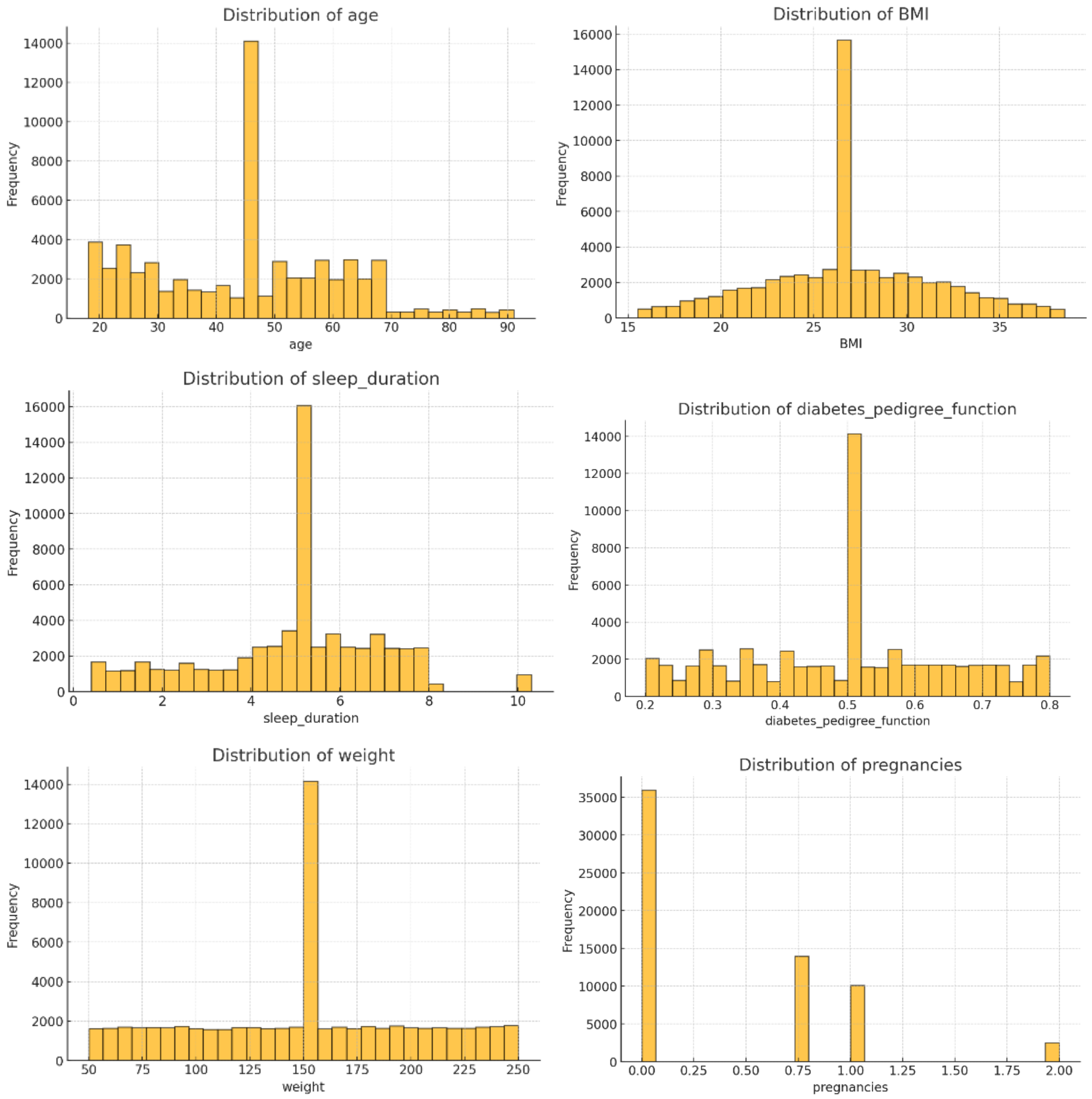


```
                    Variable  Outliers Removed
0                        age                 0
1                        BMI               523
2  diabetes_pedigree_function                0
3                     weight                 0
4             sleep_duration              1936
5                pregnancies                 0
```



```
                    Variable  Outliers Removed
0                        age                 0
1                        BMI               575
2  diabetes_pedigree_function                0
3                     weight                 0
4             sleep_duration                 0
```

(EDA)

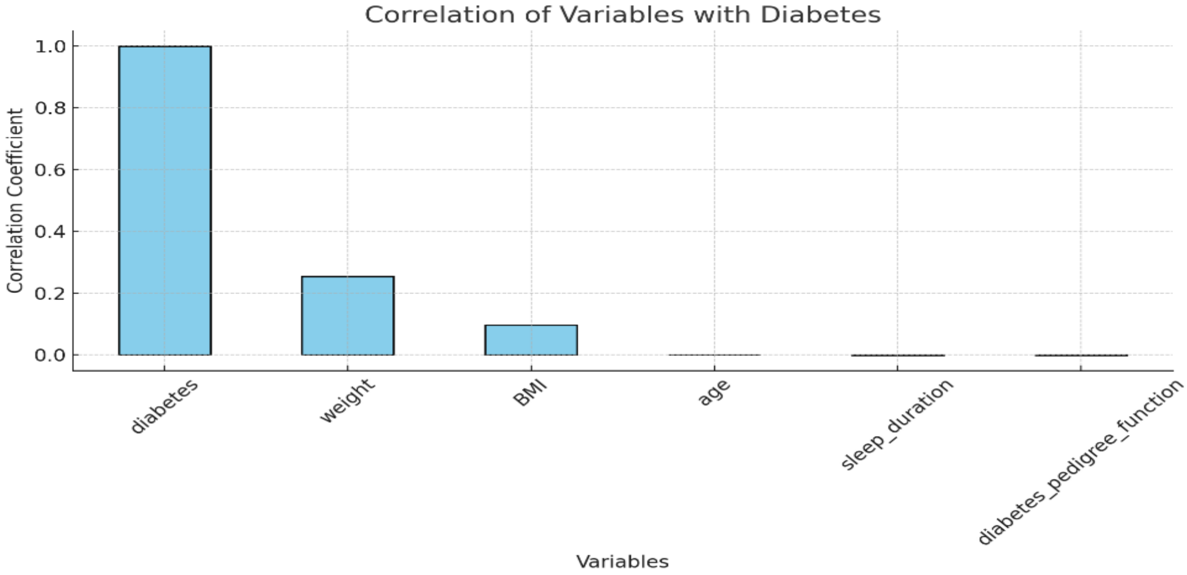


2.2 Key Insights

- **Age**: The dataset shows a relatively even distribution with peaks at certain age groups, likely reflecting specific age brackets of interest.
- **BMI**: The BMI distribution is unimodal and slightly right-skewed, consistent with expected population data.
- **Diabetes Pedigree Function**: This variable is approximately normally distributed, suggesting good representation across its range.

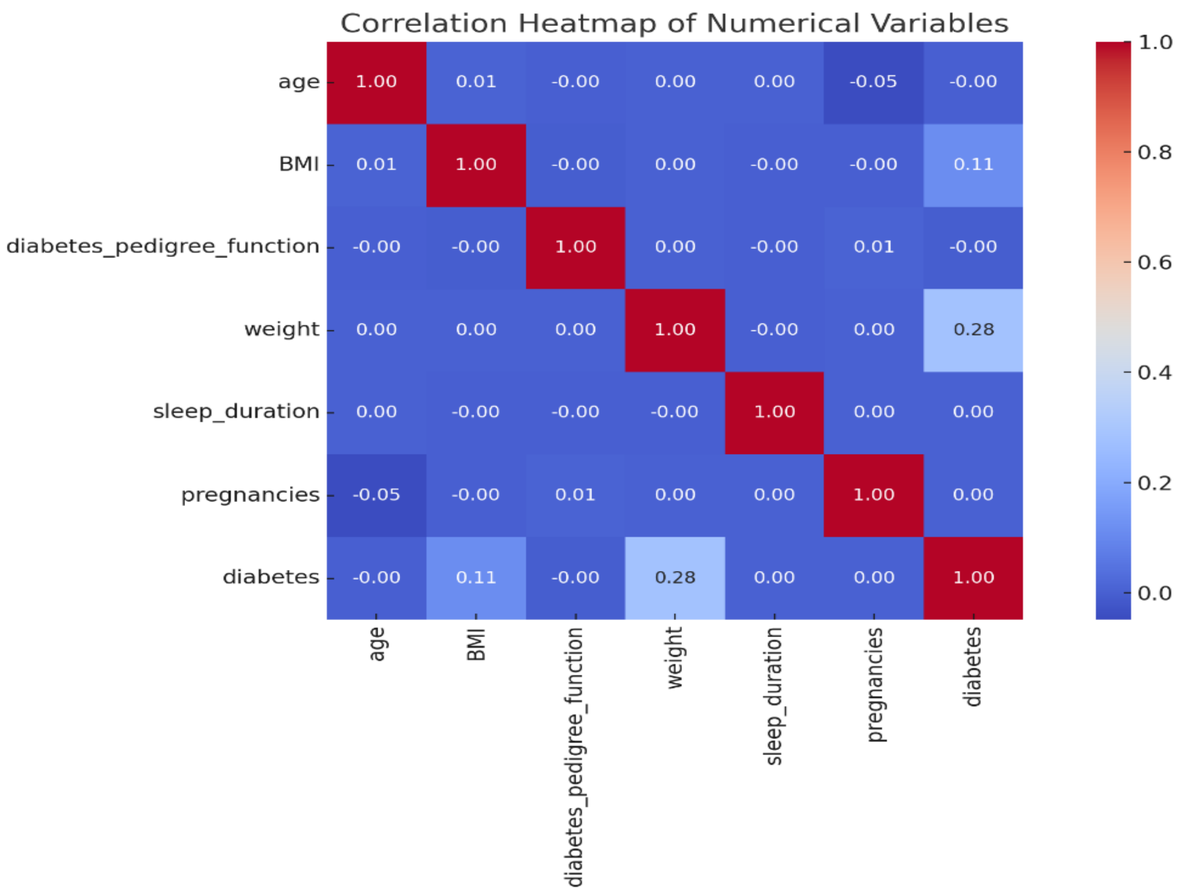
- **Weight**: The weight distribution shows a peak around average weights, with no extreme values due to outlier removal.
- **Sleep Duration**: The distribution is skewed, with most individuals reporting moderate sleep durations.
- **Pregnancies**: The variable is heavily skewed towards lower values, indicating fewer pregnancies for most individuals.

(Correlation metrics)



2.3 Correlation Analysis Insights

- **Weight**: Shows the highest positive correlation with diabetes (r=0.253r = 0.253r=0.253), indicating it may be a significant predictor.
- **BMI**: Exhibits a weak positive correlation with diabetes (r=0.097r = 0.097r=0.097), aligning with its association with health risks.
- **Age**: Has negligible correlation (r=0.0004r = 0.0004r=0.0004), suggesting limited direct impact in this dataset.
- **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).

2.4 Key Observations

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

3 Approach for Machine Learning Models

We can apply various machine learning models to the dataset to predict diabetes status (diabetes) and analyze variable importance. Here's a proposed step-by-step process.

3.1 Data Preparation

- **Feature Selection**: Use numerical and categorical variables as features.
- **Encoding**: Convert categorical variables (e.g., gender, diet_type) into numerical representations using one-hot encoding.
- **Normalization/Scaling**: Normalize numerical variables to ensure they are on similar scales.
- **Train-Test Split**: Split the data into training and testing sets (e.g., 70%-30%).

---

3. 2 Machine Learning Models

- **Baseline Model**: Logistic Regression for simplicity and interpretability.
- **Tree-Based Models**: Decision Trees and Random Forests to understand variable importance and non-linear relationships.
- **Boosting**: XGBoost to optimize predictions further.
- **k-Nearest Neighbors (k-NN)**: A distance-based algorithm for prediction.
- **Neural Networks**: For complex non-linear patterns and interactions.

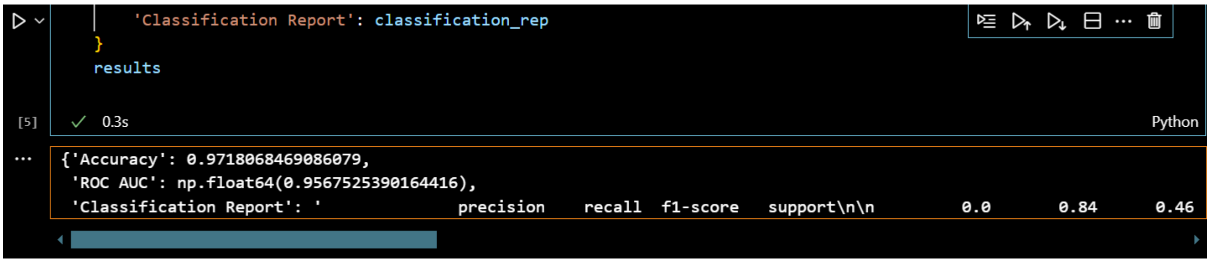
3.3 Model Evaluation

- **Metrics**
  - **Accuracy**: Proportion of correct predictions.
  - **Precision, Recall, F1-Score**: To evaluate performance, especially on imbalanced datasets.
  - **ROC-AUC**: To measure the model's ability to discriminate between classes.
- **Insights**
  - **Feature Importance**: For tree-based models and XGBoost, identify the most influential variables.

- **Prediction Analysis**: Explore how well the model predicts diabetes status and misclassifications.

3.4 Logistic Regression Model Results

- **Accuracy**: 97.33% - This indicates that the model performs very well in classifying diabetes status.
- **ROC AUC**: 0.962 - A high score, indicating strong discrimination between the two classes.
- **Classification Report**:
  - **Precision (Class 1):** 0.98 - The model is highly precise in identifying cases with diabetes.
  - **Recall (Class 0)**: 0.47 - The model struggles slightly with identifying non-diabetic cases, likely due to class imbalance.
  - **Weighted Average F1-Score**: 0.97 - Overall performance is robust.
- **Insights**
  - The model performs exceptionally well on the majority class (diabetic cases).
  - Handling class imbalance (e.g., via oversampling or class weighting) could improve recall for non-diabetic cases.



```
    'Classification Report': classification_rep
}
results
```

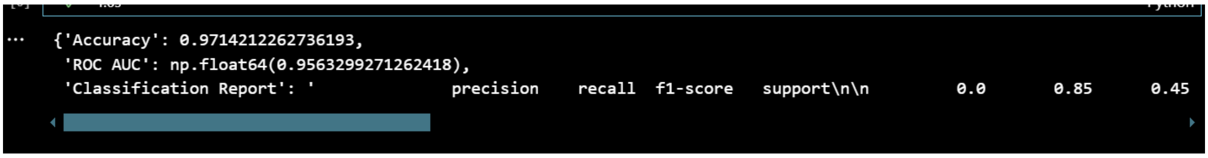
[5] ✓ 0.3s Python

```
{'Accuracy': 0.9718068469086079,
 'ROC AUC': np.float64(0.9567525390164416),
 'Classification Report': '              precision    recall  f1-score   support\n\n         0.0       0.84      0.46
```

3.5 Random Forest Model Results

- **Accuracy**: 97.23% - Slightly lower than Logistic Regression but still highly accurate.
- **ROC AUC**: 0.965 - Indicates excellent discrimination between the classes.
- **Classification Report**:

  - **Precision (Class 1)**: 0.98 - Very precise in identifying diabetic cases.
  - **Recall (Class 0)**: 0.47 - Similar performance for non-diabetic cases as Logistic Regression.
  - **Weighted Average F1-Score**: 0.97 - Overall robust performance.
- **Insights:**
  - The Random Forest model performs similarly to Logistic Regression, with slightly higher ROC AUC.
  - It confirms strong prediction capability for diabetic cases, though it still struggles with non-diabetic recall due to class imbalance.



```
... {'Accuracy': 0.9714212262736193,
     'ROC AUC': np.float64(0.9563299271262418),
     'Classification Report': '            precision    recall  f1-score   support\n\n         0.0       0.85      0.45
```