**Exploratory Data Analysis (EDA) Report:**

**Healthcare-Diabetes Dataset**

**1. Data Cleaning Overview**

* **Initial Issues Detected:**
  + Missing values in key variables:
    - **Glucose:** 17 missing entries
    - **Blood Pressure:** 24 missing entries
    - **Skin Thickness:** 17 missing entries
  + Placeholder values like **0** in variables such as **Glucose**, **Blood Pressure**, **Skin Thickness**, and **Insulin**, which could indicate unmeasured or missing data.
* **Action Taken:**
  + Missing values in **Glucose**, **Blood Pressure**, and **Skin Thickness** were replaced with their **median values** to minimize skewness caused by extreme values.
  + Placeholder zeros were retained for now but may require further imputation or exclusion based on visual inspection.
* **Post-Cleaning Status:**
  + No columns have missing values anymore.

**2. Statistical Summary (After Data Cleaning)**

**Key Numerical Insights:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Variable** | **Mean** | **Min** | **Max** | **Key Observations** | | --- | --- | --- | --- | --- | | **Glucose** | 121.1 | 0 | 199 | Some zero values may indicate missing data. High glucose correlates with diabetes. | | **Blood Pressure** | 69.2 | 0 | 122 | Zero values may indicate missing measurements. Normal BP is around 80. | | **Skin Thickness** | 20.8 | 0 | 110 | Many zero values suggest unrecorded measurements. | | **Insulin** | 80.1 | 0 | 846 | Extreme max value suggests outliers. | | **BMI** | 32.1 | 0 | 80.6 | High BMI is a known risk factor for diabetes. | | **Age** | 33.1 | 21 | 81 | Patients range from young adults to elderly individuals. | | **Outcome** | 0.34 | 0 | 1 | About 34% of patients are diabetic. | |

**Observations:**

* Variables like **Glucose**, **Blood Pressure**, **Skin Thickness**, and **Insulin** have zeros, suggesting either data entry issues or unmeasured observations.
* The dataset represents a relatively young population (mean age ~33 years).
* Approximately **34% of individuals** in the dataset have diabetes.

**3. Correlation Analysis**

**Top Correlations with Diabetes Outcome:**

* **Glucose:** 0.46 → Strong positive correlation, indicating higher glucose levels are strongly associated with diabetes.
* **BMI:** 0.28 → Moderate positive correlation; individuals with higher BMI are at a higher risk.
* **Age:** 0.24 → Older age correlates moderately with diabetes prevalence.
* **Insulin:** 0.12 → Weak positive correlation, though irregularities in Insulin data may impact this.

**Other Notable Correlations:**

* **Skin Thickness ↔ Insulin:** 0.44 → Individuals with thicker skinfold measurements tend to have higher insulin levels.
* **Blood Pressure ↔ BMI:** 0.28 → Suggests a connection between higher BMI and elevated blood pressure.

**Visual Correlation Summary:**

* The strongest predictor of diabetes is **Glucose levels**, followed by **BMI** and **Age**.
* Insulin and Skin Thickness need further inspection due to potential anomalies in their data distribution.

**4. Key Insights**

1. **High-Risk Indicators for Diabetes:**
   * Elevated **Glucose**, **BMI**, and **Age** are significant risk factors.
2. **Data Quality Issues:**
   * Variables like **Insulin**, **Skin Thickness**, and **Blood Pressure** have multiple zeros, which may require better handling (e.g., imputation based on predictive models).
3. **Demographic Distribution:**
   * The dataset predominantly represents individuals aged **21–81**, with a mean age of **33 years**.
4. **Outcome Imbalance:**
   * Approximately **34%** of patients are diabetic, indicating a slight imbalance in class distribution.

**5. Next Steps and Recommendations**

1. **Data Visualization:**
   * Create histograms and boxplots for **Glucose**, **Insulin**, and **BMI** to detect anomalies and patterns.
   * Generate scatter plots to visualize relationships between correlated variables.
2. **Outlier Detection:**
   * Investigate extreme values in **Insulin** and **Skin Thickness** for better data handling.
3. **Feature Engineering:**
   * Derive additional insights by creating new features or transforming existing ones (e.g., Age bins).
4. **Predictive Modeling:**
   * Build a predictive model (e.g., Logistic Regression or Random Forest) to classify diabetic and non-diabetic patients effectively.