Practice Test: Analyzing a Healthcare Dataset Using SAS

Overview

The objective of this practice test is to analyze a healthcare dataset related to diabetes using SAS. The dataset, `Healthcare-Diabetes.csv`, contains information about patients, including various health metrics and an outcome variable indicating whether a patient has diabetes. The columns in the dataset are as follows:

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration after a 2-hour oral glucose tolerance test
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skinfold thickness (mm)
- **Insulin**: 2-hour serum insulin (mu U/ml)
- **BMI**: Body Mass Index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction**: Diabetes pedigree function
- **Age**: Age in years
- **Outcome**: Diabetes outcome (0 = No diabetes, 1 = Diabetes)

Instructions

Importing the Dataset

- Develop a SAS script to import the `Healthcare-Diabetes.csv` file into a new dataset named `diabetes_data`. Ensure the first row of the CSV file contains variable names.
- Display the first 10 records from the dataset to confirm successful import.

Summary Statistics

- Use `PROC MEANS` to calculate summary statistics (mean, standard deviation, minimum, and maximum) for `Glucose`, `BloodPressure`, `BMI`, and `Age`.

Frequency Analysis

- Perform a frequency analysis of the `Outcome` variable to observe the distribution of diabetic and non-diabetic patients.

Correlation Analysis

- Conduct a correlation analysis for the continuous variables: `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction`, and `Age`.

Scatter Plot

- Generate a scatter plot to examine the relationship between `Insulin` and `BMI`, using distinct colors to indicate diabetic and non-diabetic patients.

Categorization

- Categorize the following variables:
  - **Age**: ≤ 20, 21-30, 31-40, and > 40 years
  - **Glucose**: ≤ 100 (normal), 101-125 (pre-diabetes), and ≥ 126 (diabetes)
  - **BMI**: underweight (< 18.5), normal (18.5-24.9), overweight (25-29.9), and obese (≥ 30)

Summary Table

- Create a table with frequency, mean, median, and standard deviation for `Glucose`, `Insulin`, and `BMI`, grouped by the age categories.

Visualizations

- Generate the following plots:
  - Histograms for `Glucose`, `BMI`, and `Insulin` across age groups
  - Boxplot showing `BMI` distribution by age category
  - Scatter plot of `Glucose` versus `Age`, with a trend line

Interpretation

- Provide a brief interpretation for each analysis step and summarize key insights in a paragraph of 150-200 words.

Stratify the Dataset by BMI Categories

- Produce a summary table displaying `Pregnancies`, `Age`, `Glucose`, and `BloodPressure` for each BMI category.

Additional Visualizations

- Create the following visualizations:
  - Histograms comparing `BloodPressure` across BMI categories
  - Scatter plot of `Insulin` versus `Glucose`, highlighting clusters
  - Boxplot of `Glucose` levels for different outcome groups

Correlation Analysis

- Analyze correlations among `Age`, `Glucose`, `BMI`, and `Insulin` and discuss findings.

Report

- Write a report summarizing key findings, focusing on notable patterns, trends, or anomalies identified in the data.

Categorize Pregnancies

- Create categories: 0, 1-2, 3-4, and ≥ 5.

Glucose Distribution Analysis

- Use a histogram to analyze glucose distribution across pregnancy categories.

Summary Table

- Summarize the mean and standard deviation for `Age`, `BMI`, and `Glucose` by pregnancy category.

Scatter Plot

- Create a scatter plot showing the relationship between `Age` and `BMI`, grouped by pregnancy category.

Correlation Matrix

- Generate a correlation matrix for `Pregnancies`, `Glucose`, `Insulin`, and `BMI`.

Summary

- Provide a summary interpreting the exploratory analysis, highlighting significant trends and group differences.