Análisis Detallado de la Base de Datos de Diabetes

Hemos recibido una base de datos completa sobre diabetes para realizar un estudio exhaustivo de su información. Como primer paso, hemos analizado minuciosamente los datos para evaluar su calidad y viabilidad para el análisis estadístico.

Estructura del Dataset

La base de datos contiene **768 filas** y **9 columnas**, proporcionando un volumen de datos suficiente para realizar análisis estadísticos robustos y obtener conclusiones significativas sobre los patrones de diabetes.

Calidad de los Datos

Verificamos la integridad de los datos identificando valores faltantes y datos sospechosos. Detectamos múltiples valores con **0** que resultan médicamente improbables, requiriendo un proceso de limpieza de datos.

Proceso de Limpieza

Implementamos un proceso de limpieza riguroso para eliminar valores anómalos y datos erróneos, permitiendo una comparativa entre los datos originales y los datos procesados para mayor precisión analítica.

```
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Análisis de Valores Faltantes

La evaluación de la completitud de los datos reveló patrones específicos de valores faltantes que requieren atención especial. La identificación de estos valores es crucial para mantener la integridad del análisis estadístico.

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Detección de Datos Sospechosos

Identificamos valores médicamente improbables que podrían comprometer la validez del análisis. Estos datos requieren un tratamiento especial para asegurar la precisión de nuestras conclusiones.

```
Glucose            5
BloodPressure     35
SkinThickness    227
Insulin          374
BMI               11
dtype: int64
```

Limpieza de datos eliminando los datos Sospechosos

Hicimos una limpieza de los datos originales elimiado los datos sospechosos y esto es con los datos que hemos podido trabajar para hacer una analisis mas exacto.

Total registros originales: 768

Total registros después de limpiar: 532

Registros eliminados: 236

A continuación presentamos los hallazgos más significativos de nuestro estudio, incluyendo comparativas entre datos originales y procesados, análisis de distribuciones por grupos de edad, y patrones de prevalencia de diabetes en la población estudiada.

Resultados del Análisis Comparativo

A continuación presentamos los hallazgos más significativos de nuestro estudio, incluyendo comparativas entre datos originales y procesados, análisis de distribuciones por grupos de edad, y patrones de prevalencia de diabetes en la población estudiada.

Comparativa niveles de Glucosa:

Detectamos 5 datos con Glucosa 0 y esto es imposible sabemos que esos datos son o falsos o erróneos al hacer la limpieza de esos 5 valores cambia todo el grafico la Media y la media en el gráfico de la derecha son mas reales.

Personas con diabetes:

Vemos como los datos varían por los datos recibidos , Hay una diferencia de 30,2% en el valor entre los diabéticos y no en el primer grafico mientras que en el segundo grafico es un 23.2% de diferencia.

Distribución de Glucosa por Rangos de Edad

Datos Originales: Muestra la distribución inicial de glucosa segmentada por grupos etarios, revelando patrones preliminares de comportamiento glucémico.

Datos Procesados: Presenta una visión más clara y precisa de cómo los niveles de glucosa varían según la edad tras la limpieza de datos.

Comparativa Resumen Estadístico Detallado

```
RESUMEN DETALLADO
==================================================
Total personas con diabetes: 268


≤30: 90 personas
  - <100: 7 personas (7.8%)
  - 100-140: 42 personas (46.7%)
  - 140-180: 32 personas (35.6%)
  - 180+: 9 personas (10.0%)

31-40: 76 personas
  - <100: 4 personas (5.3%)
  - 100-140: 37 personas (48.7%)
  - 140-180: 25 personas (32.9%)
  - 180+: 9 personas (11.8%)

41-50: 64 personas
  - <100: 5 personas (7.8%)
  - 100-140: 28 personas (43.8%)
  - 140-180: 23 personas (35.9%)
  - 180+: 7 personas (10.9%)

51-60: 31 personas
  - <100: 2 personas (6.5%)
  - 100-140: 8 personas (25.8%)
  - 140-180: 13 personas (41.9%)
  - 180+: 8 personas (25.8%)

60+: 7 personas
  - <100: 0 personas (0.0%)
  - 100-140: 1 personas (14.3%)
  - 140-180: 4 personas (57.1%)
  - 180+: 2 personas (28.6%)
```

```
ESUMEN DETALLADO Datos Limpios
==================================================
Total registros en df_limpio: 532


≤30: 321 personas
  - <100: 117 personas (36.4%)
  - 100-140: 148 personas (46.1%)
  - 140-180: 47 personas (14.6%)
  - 180+: 9 personas (2.8%)

31-40: 102 personas
  - <100: 22 personas (21.6%)
  - 100-140: 47 personas (46.1%)
  - 140-180: 25 personas (24.5%)
  - 180+: 8 personas (7.8%)

41-50: 68 personas
  - <100: 20 personas (29.4%)
  - 100-140: 27 personas (39.7%)
  - 140-180: 16 personas (23.5%)
  - 180+: 5 personas (7.4%)

51-60: 32 personas
  - <100: 3 personas (9.4%)
  - 100-140: 8 personas (25.0%)
  - 140-180: 14 personas (43.8%)
  - 180+: 7 personas (21.9%)

60+: 9 personas
  - <100: 0 personas (0.0%)
  - 100-140: 5 personas (55.6%)
  - 140-180: 3 personas (33.3%)
  - 180+: 1 personas (11.1%)
```

Comparativa Distribución Porcentual de Niveles de Glucosa por Edad

Los resúmenes estadísticos proporcionan una visión comprehensiva de las características principales del dataset, incluyendo medidas de tendencia central, dispersión y distribución de todas las variables relevantes para el estudio de diabetes.

Visualizaciones Alternativas del Análisis

Representación alternativa que ofrece una perspectiva diferente de los patrones identificados en los datos de diabetes.

Visualización complementaria que refuerza los hallazgos principales del análisis estadístico realizado.

Conclusión del Análisis: El proceso de limpieza y análisis de datos ha revelado patrones significativos en la base de datos de diabetes. La comparativa entre datos originales y procesados demuestra la importancia de un tratamiento adecuado de los datos para obtener conclusiones válidas y confiables en estudios médicos.

Me faltaron datos para poder comparar la causa del del diabetes en este análisis quería ver si hay alguna relación con tener la Glucose alta y tener la BloodPressure o el BMI altos causan la diabetics. Lo que si hemos notado es que no hay ninguna persona en el rango de edad mas de 60 amos que tenga los niveles de Glucose menos de 100. Ese es un dato a tener en cuenta