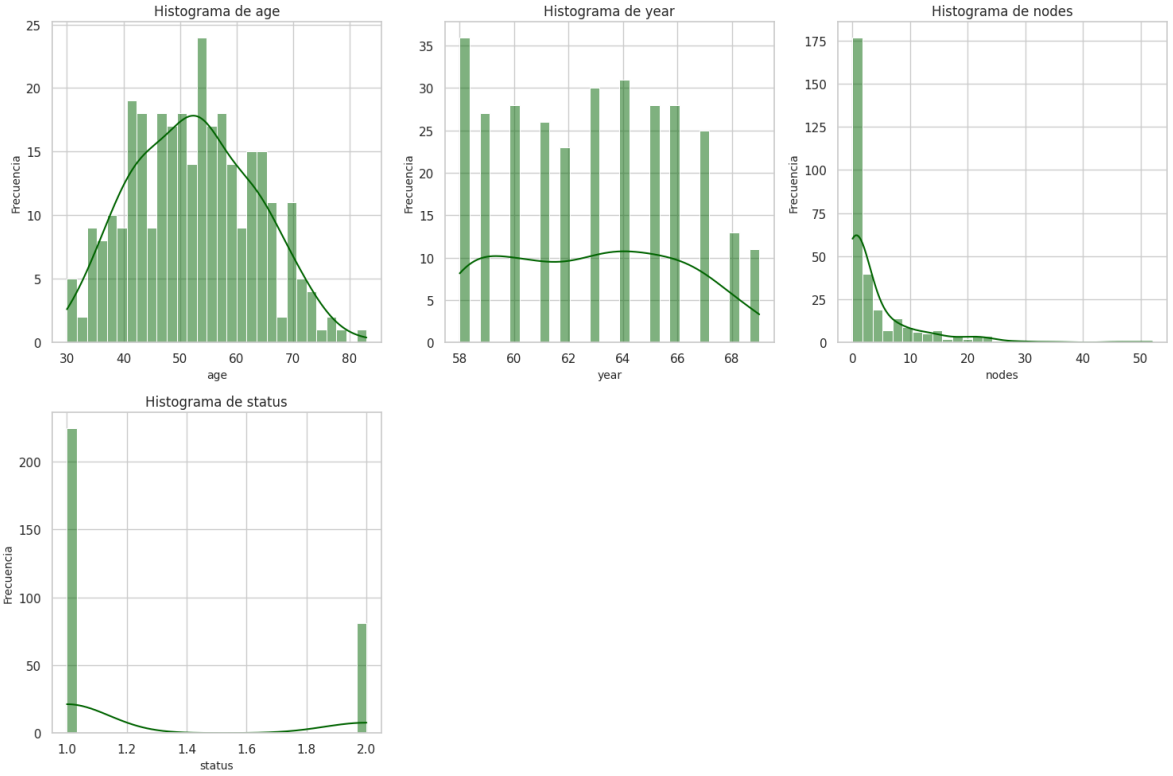
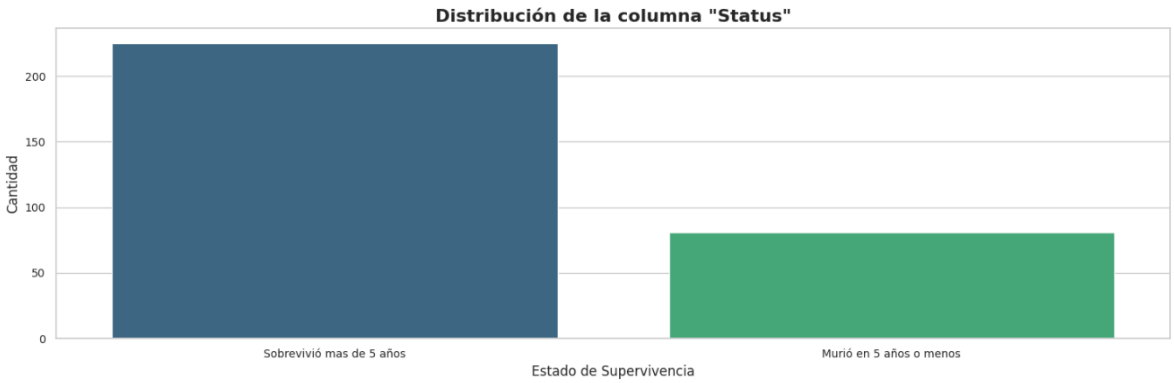
Análisis Exploratorio de Datos (EDA): Base de datos “Haberman’s Survival”

- *¿Qué información relevante obtuviste de los histogramas y gráficas de densidad (PDF)? ¿Alguna variable parece seguir una distribución normal?*



De los histogramas de “age”, “year”, “nodes” y “status”, se puede notar que ninguna sigue una distribución normal o de campana, la más cercana es “age”. La columna “nodes” tiene la mayoría de los valores a la izquierda (en 0) y presenta una larga cola mucho mas baja hasta 52). La columna “Year” tiene valores muy parecidos en todo su rango, y “status” se concentra en los extremos, que es lo que se espera para valores del tipo categórico.

- *¿Qué insights obtuviste de las gráficas de barras para la variable categórica status? ¿Cómo se distribuyen los pacientes según su estado de supervivencia?*



En la gráfica de barras se puede apreciar que aproximadamente un 60% de paciente sobrevivió mas de 5 años, en comparación de los que murieron los primeros 5 años.

Pruebas de Normalidad:

- *¿Qué variables no siguen una distribución normal según las pruebas de Shapiro-Wilk, Anderson-Darling y Kolmogorov-Smirnov? ¿Cómo afecta esto al modelo de regresión lineal?*

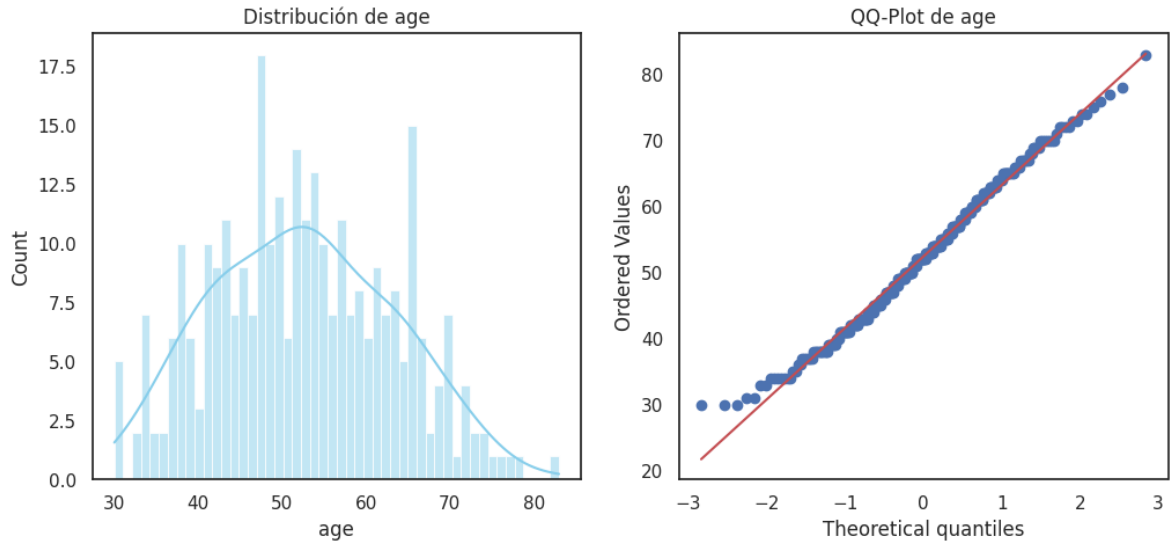
Ninguna variable sigue una distribución normal, lo más cercano es “age” que al nivel de significancia del 5% los datos parecen normales, de ahí en delante pierde la normalidad.

Aplicar un modelo de regresión lineal arrojaría errores, por la falta de normalidad de los datos. La columna “status” es categórica, por lo que se debe someter a otro tipo de modelo, como regresión logística.

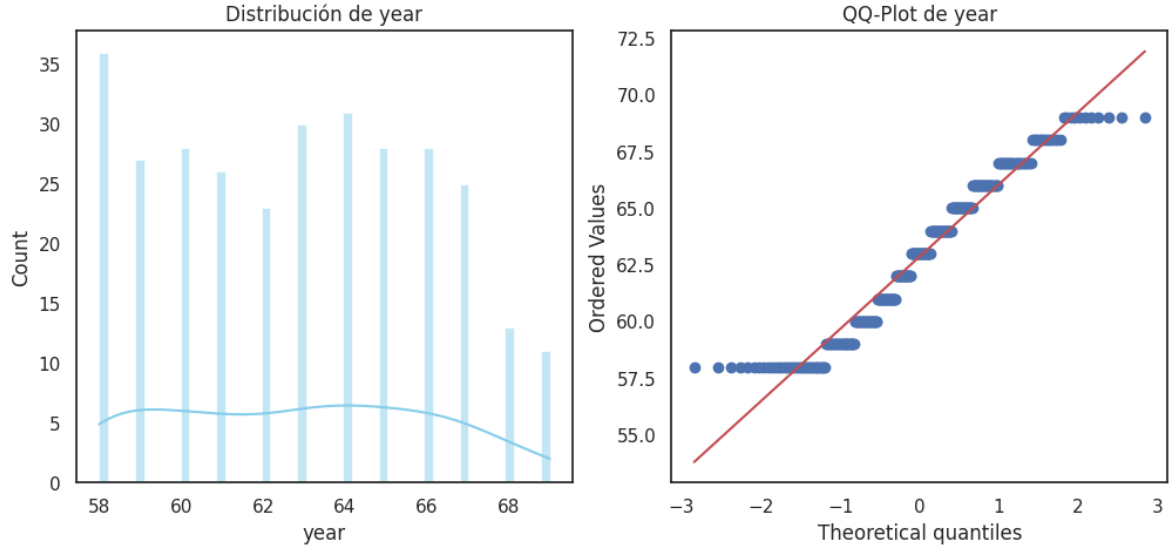
- ¿Qué conclusiones puedes extraer de los QQplots? ¿Qué variables tienen una distribución cercana a la normal?

Las columnas que tienen datos cercanos a la normalidad son la de “age”, que solo se desvía en las colas, “nodes” y “year” están muy desviadas de la línea diagonal. La columna “status” es totalmente indiferente, lo esperado por ser categórica.

```
Pruebas de normalidad
```

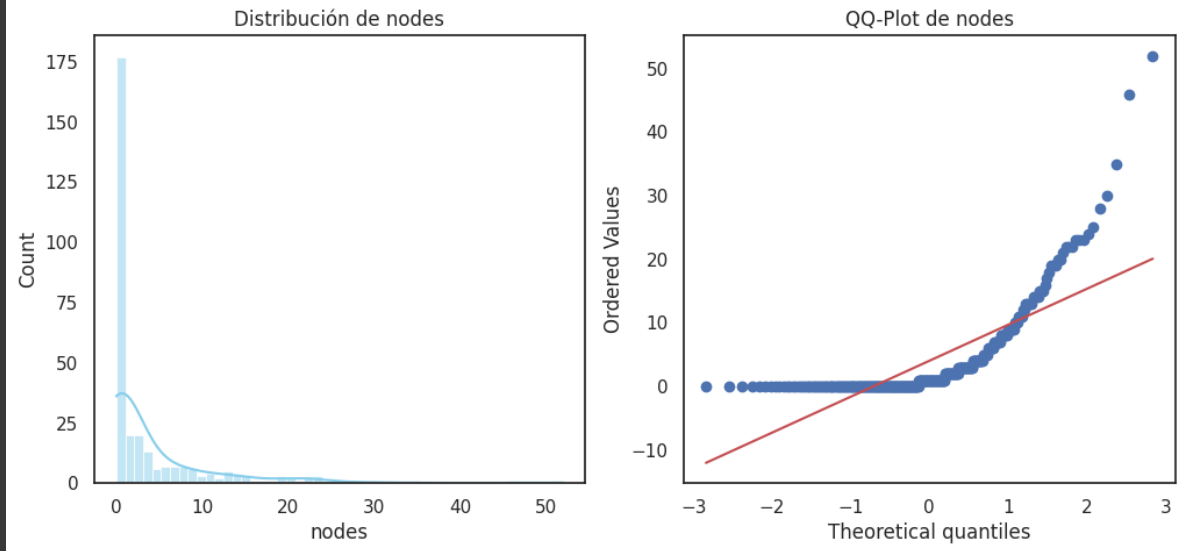



```
Pruebas de normalidad para la columna: age
Shapiro-Wilk Test: p-valor = 0.02605
Kolmogorov-Smirnov Test: p-valor = 0.46980
Anderson-Darling Test: Estadístico = 0.73156
  Al nivel de significancia 15.0%, los datos NO parecen normales.
  Al nivel de significancia 10.0%, los datos NO parecen normales.
  Al nivel de significancia 5.0%, los datos parecen normales.
  Al nivel de significancia 2.5%, los datos parecen normales.
  Al nivel de significancia 1.0%, los datos parecen normales.
D'Agostino-Pearson Test: p-valor = 0.00780
```



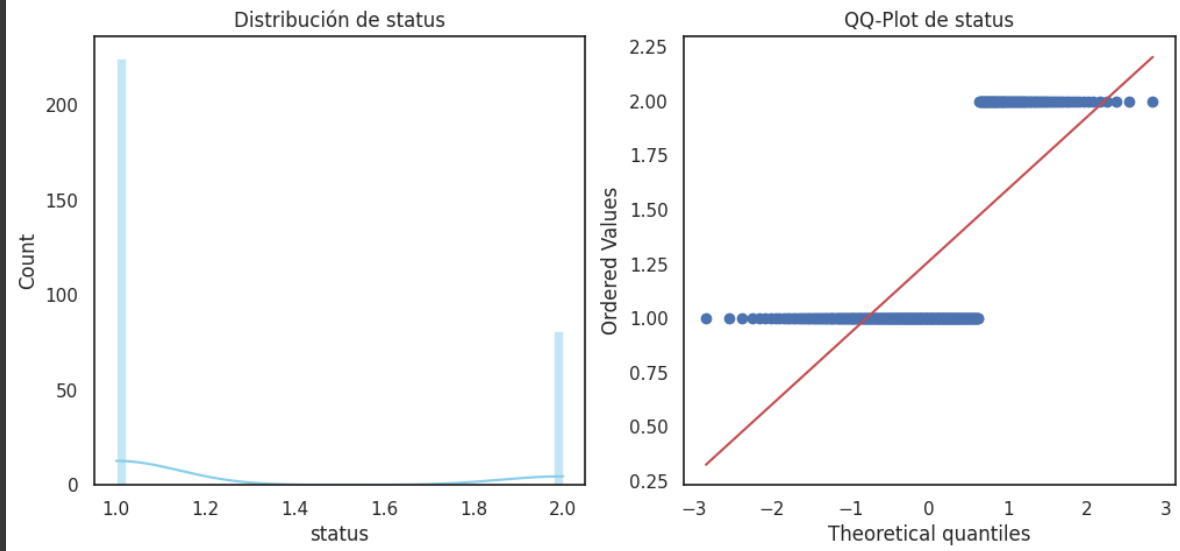

```
Pruebas de normalidad para la columna: year
Shapiro-Wilk Test: p-valor = 0.00000
Kolmogorov-Smirnov Test: p-valor = 0.00158
Anderson-Darling Test: Estadístico = 4.31374
  Al nivel de significancia 15.0%, los datos NO parecen normales.
```

```
  Al nivel de significancia 10.0%, los datos NO parecen normales.
  Al nivel de significancia 5.0%, los datos NO parecen normales.
  Al nivel de significancia 2.5%, los datos NO parecen normales.
  Al nivel de significancia 1.0%, los datos NO parecen normales.
D'Agostino-Pearson Test: p-valor = 0.00000
```




```

Pruebas de normalidad para la columna: nodes
Shapiro-Wilk Test: p-valor = 0.00000
Kolmogorov-Smirnov Test: p-valor = 0.00000
Anderson-Darling Test: Estadístico = 39.68662
  Al nivel de significancia 15.0%, los datos NO parecen normales.
  Al nivel de significancia 10.0%, los datos NO parecen normales.
  Al nivel de significancia 5.0%, los datos NO parecen normales.
  Al nivel de significancia 2.5%, los datos NO parecen normales.
  Al nivel de significancia 1.0%, los datos NO parecen normales.
D'Agostino-Pearson Test: p-valor = 0.00000
```



```
Pruebas de normalidad para la columna: status
Shapiro-Wilk Test: p-valor = 0.00000
Kolmogorov-Smirnov Test: p-valor = 0.00000
Anderson-Darling Test: Estadístico = 71.18230
  Al nivel de significancia 15.0%, los datos NO parecen normales.
  Al nivel de significancia 10.0%, los datos NO parecen normales.
  Al nivel de significancia 5.0%, los datos NO parecen normales.
  Al nivel de significancia 2.5%, los datos NO parecen normales.
  Al nivel de significancia 1.0%, los datos NO parecen normales.
D'Agostino-Pearson Test: p-valor = 0.00000
```

Tratamiento de Datos Faltantes:

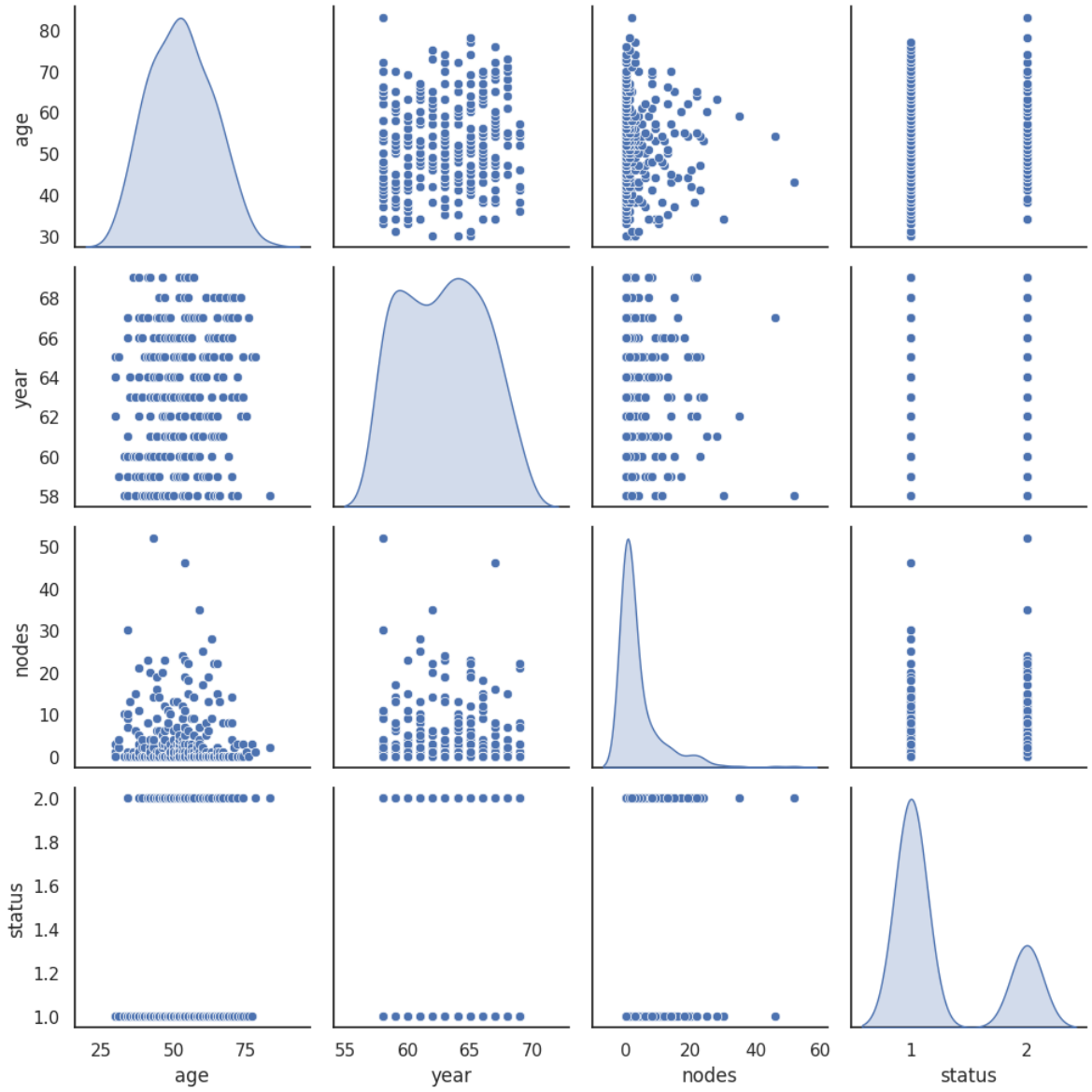
- *¿Qué columnas tenían todos los valores faltantes? ¿Cómo manejaste estas columnas?*

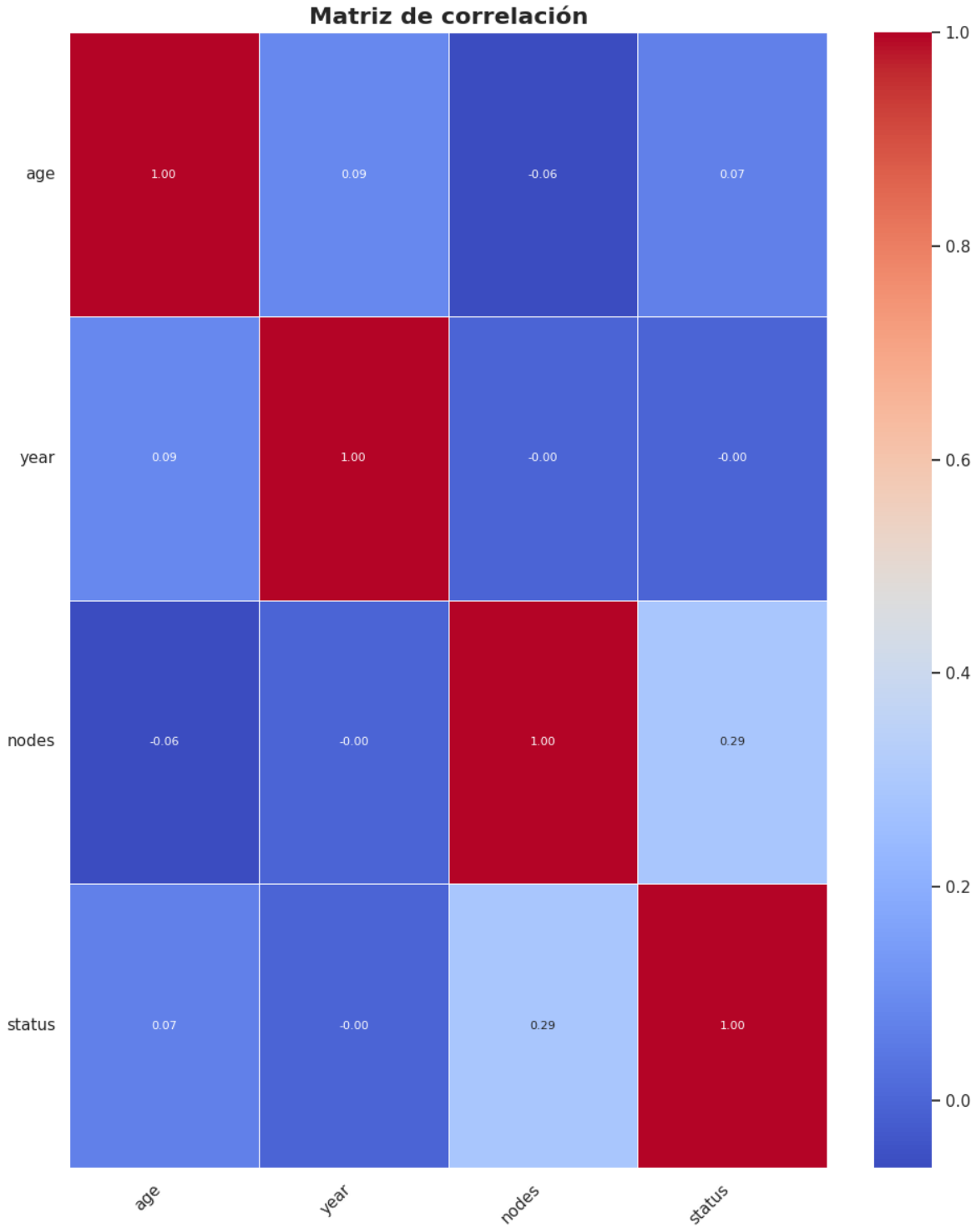
La base de datos no tiene datos faltantes.

- *¿Cómo cambió el EDA después de la imputación de datos? ¿Observaste diferencias significativas en las distribuciones de las variables?*

No hubo cambios en los datos.

Matriz de Correlación y Pairplot:





- *¿Qué relaciones lineales identificaste en la matriz de correlación y el pairplot? ¿Alguna variable tiene una correlación fuerte con la variable objetivo?*

No se encontraron relaciones de las variables con la variable objetivo. Pero cabe mencionar que es importante realizar otro tipo de pruebas, como regresión logística y arboles de decisión. Esto debido a que, clínicamente hablando, la edad del paciente y el numero de ganglios detectados están directamente relacionados con la supervivencia, en el padecimiento de cáncer.

- *¿Cómo podrías utilizar esta información para seleccionar características (features) en un modelo de regresión lineal?*

Esta información es útil para justificar la elección de un modelo distinto, para que el resultado sea válido. Como regresión logística o arboles de decisión.