Asociación entre Datos Médicos y Diagnóstico de Alzheimer

Análisis Estadístico de Datos Clínicos con Enfoque Tidyverse

Matías Elier Labraña Abarca

Invalid Date

Table of contents

1 Introducción

Este estudio tiene como objetivo identificar los factores clínicos y demográficos asociados al diagnóstico de la enfermedad de Alzheimer a partir del análisis del conjunto de datos *Alzheimer's Disease Data* (Kaggle). Se emplea un enfoque sistemático basado en el ecosistema **Tidyverse**, el cual facilita tanto la limpieza de los datos como la exploración de relaciones significativas entre variables como edad, nivel educativo, antecedentes familiares y puntuaciones cognitivas (MMSE).

1.0.1 Objetivos del análisis

- **Preprocesamiento de datos** utilizando las herramientas del Tidyverse para garantizar su calidad y consistencia.

  Análisis exploratorio de datos (AED) enfocado en la identificación de patrones y relaciones entre variables relevantes.

  Desarrollo de una función reproducible para automatizar el análisis y facilitar su aplicación a futuros datasets similares.

2 Definición del Problema

2.0.0.1 Problema de Investigación

La enfermedad de Alzheimer es una condición neurodegenerativa progresiva que afecta significativamente la calidad de vida de quienes la padecen. La identificación de factores de riesgo tempranos constituye un desafío clave en el ámbito de la salud pública y la investigación biomédica, ya que permitiría diseñar estrategias preventivas más eficaces y personalizadas.

2.0.0.2 Objetivo General

El objetivo de este estudio es explorar las relaciones existentes entre características demográficas, factores clínicos y el diagnóstico de la enfermedad de Alzheimer, evaluando además la factibilidad de desarrollar un modelo predictivo basado en dichas variables.

2.0.0.3 Variables Clave

- **Variables Cuantitativas**:
  - **Age (Edad)**: Edad del paciente en años. Se plantea la hipótesis de que una mayor edad está asociada con un riesgo más elevado de desarrollar Alzheimer
  - **MMSE (Mini-Mental State Examination)**: Puntaje obtenido en la evaluación cognitiva. Se espera que puntuaciones más bajas se correlacionen con un diagnóstico de Alzheimer.

  Variables Cualitativas:

- **FamilyHistoryAlzheimers**: Antecedentes familiares de Alzheimer (sí/no). Se plantea que contar con antecedentes familiares incrementa el riesgo de desarrollar la enfermedad.
- **EducationLevel**: Nivel educativo alcanzado. Se investigará la posible relación entre los niveles educativos (bajo, medio, alto) y el diagnóstico de Alzheimer.

3 Datos y Metodología

3.0.1 Descripción del Dataset

En primer lugar, se listan los archivos con extensión **.csv** en el directorio de trabajo actual, con el propósito de verificar la presencia del archivo requerido: `alzheimers_disease_data.csv`. Esta verificación inicial es fundamental para evitar errores durante el proceso de carga.

A continuación, se procede a cargar el dataset utilizando la función **`read_csv`** del paquete **readr**, que forma parte del ecosistema **tidyverse**. Se implementa un manejo de errores robusto para asegurar que la carga se realice de manera exitosa y detener el proceso en caso de fallos, garantizando la confiabilidad de los datos desde la etapa inicial.

Una vez completada la carga, se inspeccionan las dimensiones del dataset, el número de filas y columnas para obtener un panorama general de su tamaño. Posteriormente, se efectúa una revisión preliminar de valores ausentes (**NA**) en todas las columnas, generando un resumen con la cantidad de datos faltantes por variable. Esta información es crucial para detectar problemas de calidad de datos que deberán abordarse en etapas posteriores.

Finalmente, se presenta un listado de las principales variables utilizadas en el análisis, con una breve descripción de cada una, incluyendo su tipo (cuantitativa o cualitativa) y su relevancia en el estudio.

```
print(list.files(pattern = "*.csv"))
```

```
[1] "alzheimers_disease_data.csv"
```

```
# Asigna el nombre del .CSV a una variable
archivo <- "alzheimers_disease_data.csv"

# Inicializar una bandera para rastrear el éxito de la carga del archivo.
carga_exitosa <- TRUE

# Intentar cargar el archivo CSV utilizando readr::read_csv.
# Se incluye manejo de errores para capturar problemas durante la carga.
alzheimer_raw <- tryCatch(
  {
    # show_col_types = FALSE evita mensajes sobre los tipos de columna.
    readr::read_csv(archivo, show_col_types = FALSE)
  },
  error = function(captura_error) {
    # En caso de error, mostrar un mensaje descriptivo.
    message("Error al cargar el archivo: ", captura_error$message)
    # Actualizar la bandera para indicar que la carga falló.
    carga_exitosa <<- FALSE # <<- para modificar la variable en el entorno global del chunk.
    # Devolver NULL como resultado de la operación fallida.
    NULL
  }
)
```

```r
# Verificar si la carga fue exitosa y el objeto de datos no es NULL.
if (carga_exitosa && !is.null(alzheimer_raw)) {
  message("Carga exitosa del archivo.")
} else {
  message("No se pudo cargar el archivo. Verifica la ruta y el nombre.")
  # Detener la ejecución del documento si la carga del archivo falla.
  knitr::knit_exit()
}


# Verificar que el dataset 'alzheimer_raw' exista y no sea NULL antes de proceder.
if (exists("alzheimer_raw") && !is.null(alzheimer_raw)) {
  # Imprimir las dimensiones del dataset.
  cat(paste0("El dataset original contiene ", nrow(alzheimer_raw), " filas y ",
      ncol(alzheimer_raw), " columnas.\n"))
}
```

```
El dataset original contiene 2149 filas y 35 columnas.
```

3.1 Procesamiento

Esta sección se centra en la preparación y limpieza del dataset original (`alzheimer_raw`). A continuación, se ejecutan pasos clave para garantizar la calidad de los datos y su correcta tipificación antes de los análisis posteriores. Primero, se realiza una revisión exhaustiva de los valores ausentes y se documentan las variables que presentan datos faltantes. Luego, se transforman variables categóricas y numéricas a sus formatos adecuados, facilitando la consistencia en las etapas de análisis y modelado. Finalmente, se ajusta el dataset para excluir columnas irrelevantes o identificadores únicos, obteniendo un conjunto final de datos (`alzheimer_analisis`) listo para las fases de exploración y modelado.

3.1.1 Limpieza Inicial

```r
# Revisión inicial de valores ausentes (NA).
# Verificar que el dataset 'alzheimer_raw' exista y no sea NULL.
if (exists("alzheimer_raw") && !is.null(alzheimer_raw)) {
  # Calcular la cantidad de valores NA por columna.
  missing_values_summary <- alzheimer_raw %>%
    summarise(
      # Aplicar la función sum(is.na(.)) a todas las columnas.
      across(
        everything(),
        ~ sum(is.na(.))
      )
    ) %>%
    # Convertir el resumen de formato ancho a largo para facilitar el filtrado y visualización.
    pivot_longer(
      everything(),
      names_to = "columna",
      values_to = "cantidad_na"
```

```
    ) %>%
    # Filtrar para mostrar solo las columnas que tienen al menos un valor NA.
    filter(cantidad_na > 0)

  # Visualización de los resultados del conteo de NAs.
  # Si se encontraron columnas con valores ausentes, mostrar un resumen.
  if (nrow(missing_values_summary) > 0) {
    cat("Se identificaron valores ausentes en las siguientes columnas:\n")
    # Imprimir la tabla de valores perdidos usando kable para un formato legible.
    print(knitr::kable(
      missing_values_summary,
      caption = "Resumen de Valores Perdidos Iniciales"
    ))
    cat("Estos valores serán considerados en etapas posteriores del análisis.\n")
  } else {
    # Si no se encontraron NAs, informar al usuario.
    cat("No se encontraron valores perdidos en una revisión inicial del dataset.\n")
  }
}
```

```
No se encontraron valores perdidos en una revisión inicial del dataset.
```

```
# Descripción breve de las variables clave (reiteración de la sección de Objetivos).
cat("\nVariables principales del estudio (reiteración):\n")
```

```
Variables principales del estudio (reiteración):
```

```
cat("- Age (Numérica): Edad del paciente en años.\n")
```

```
- Age (Numérica): Edad del paciente en años.
```

```
cat("- MMSE (Numérica): Puntaje Mini-Mental State Examination.\n")
```

```
- MMSE (Numérica): Puntaje Mini-Mental State Examination.
```

```
cat("- FamilyHistoryAlzheimers (Categórica Binaria): Antecedentes familiares.\n")
```

```
- FamilyHistoryAlzheimers (Categórica Binaria): Antecedentes familiares.
```

```
cat("- EducationLevel (Categórica Ordinal): Nivel educativo.\n")
```

```
- EducationLevel (Categórica Ordinal): Nivel educativo.
```

```r
cat("- Diagnosis (Categórica Binaria): Diagnóstico de Alzheimer (1 = Sí, 0 = No).\n")
```

```
- Diagnosis (Categórica Binaria): Diagnóstico de Alzheimer (1 = Sí, 0 = No).
```

3.1.2 Trasformación de Variable

```r
# Verificar que el dataset 'alzheimer_raw' exista y no sea NULL.
if (exists("alzheimer_raw") && !is.null(alzheimer_raw)) {

  # Lista de columnas a convertir en factor.
  cols_to_factor <- c(
  "BehavioralProblems",       "CardiovascularDisease",
  "Confusion",                "Depression",
  "Diabetes",                 "DifficultyCompletingTasks",
  "Disorientation",           "EducationLevel",
  "Ethnicity",                "FamilyHistoryAlzheimers",
  "Forgetfulness",            "Gender",
  "HeadInjury",               "Hypertension",
  "MemoryComplaints",         "PersonalityChanges",
  "Smoking"
)

# Crear el dataset 'alzheimer' aplicando las transformaciones.
alzheimer <- alzheimer_raw %>%
  # Convertir las columnas en 'cols_to_factor' a tipo factor.
  mutate(across(all_of(cols_to_factor), as.factor)) %>%
  # Mutaciones específicas con niveles y etiquetas definidos.
  mutate(
    Diagnosis = factor(Diagnosis,
                       levels = c(0, 1),
                       labels = c("No Alzheimer",
                                  "Alzheimer"
                       )),
    Gender = factor(Gender,
                    levels = c(0, 1),
                    labels = c("Masculino",
                               "Femenino"
                       )),
    Ethnicity = factor(Ethnicity,
                       levels = c(0,1,2,3),
                       labels = c("Caucásico",
                                  "Afroamericano",
                                  "Asiático",
                                  "Otro"
                       )),
    EducationLevel = factor(EducationLevel,
                            levels = c(0,1,2,3),
                            labels = c("Ninguno",
                                       "Secundaria",
                                       "Universitario",
                                       "Superior"),
```

```
                            ordered = TRUE
                        ))

  # Mostrar estructura del dataset tras preprocesamiento inicial.
  cat("\nEstructura del dataset 'alzheimer' tras preprocesamiento inicial:\n")
  dplyr::glimpse(alzheimer)
  }
```

```
Estructura del dataset 'alzheimer' tras preprocesamiento inicial:
Rows: 2,149
Columns: 35
$ PatientID                 <dbl> 4751, 4752, 4753, 4754, 4755, 4756, 4757, 47~
$ Age                       <dbl> 73, 89, 73, 74, 89, 86, 68, 75, 72, 87, 89, ~
$ Gender                    <fct> Masculino, Masculino, Masculino, Femenino, M~
$ Ethnicity                 <fct> Caucásico, Caucásico, Otro, Caucásico, Caucá~
$ EducationLevel            <ord> Universitario, Ninguno, Secundaria, Secundar~
$ BMI                       <dbl> 22.92775, 26.82768, 17.79588, 33.80082, 20.7~
$ Smoking                   <fct> 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0,~
$ AlcoholConsumption        <dbl> 13.2972177, 4.5425238, 19.5550845, 12.209265~
$ PhysicalActivity          <dbl> 6.3271125, 7.6198845, 7.8449878, 8.4280014, ~
$ DietQuality               <dbl> 1.34721431, 0.51876714, 1.82633466, 7.435604~
$ SleepQuality              <dbl> 9.025679, 7.151293, 9.673574, 8.392554, 5.59~
$ FamilyHistoryAlzheimers   <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
$ CardiovascularDisease     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,~
$ Diabetes                  <fct> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1,~
$ Depression                <fct> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,~
$ HeadInjury                <fct> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
$ Hypertension              <fct> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,~
$ SystolicBP                <dbl> 142, 115, 99, 118, 94, 168, 143, 117, 117, 1~
$ DiastolicBP               <dbl> 72, 64, 116, 115, 117, 62, 88, 63, 119, 78, ~
$ CholesterolTotal          <dbl> 242.3668, 231.1626, 284.1819, 159.5822, 237.~
$ CholesterolLDL            <dbl> 56.15090, 193.40800, 153.32276, 65.36664, 92~
$ CholesterolHDL            <dbl> 33.68256, 79.02848, 69.77229, 68.45749, 56.8~
$ CholesterolTriglycerides  <dbl> 162.18914, 294.63091, 83.63832, 277.57736, 2~
$ MMSE                      <dbl> 21.4635324, 20.6132673, 7.3562486, 13.991127~
$ FunctionalAssessment      <dbl> 6.5188770, 7.1186955, 5.8950773, 8.9651063, ~
$ MemoryComplaints          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
$ BehavioralProblems        <fct> 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,~
$ ADL                       <dbl> 1.72588346, 2.59242413, 7.11954774, 6.481225~
$ Confusion                 <fct> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1,~
$ Disorientation            <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1,~
$ PersonalityChanges        <fct> 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,~
$ DifficultyCompletingTasks <fct> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
$ Forgetfulness             <fct> 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0,~
$ Diagnosis                 <fct> No Alzheimer, No Alzheimer, No Alzheimer, No~
$ DoctorInCharge            <chr> "XXXConfid", "XXXConfid", "XXXConfid", "XXXC~
```

3.1.3 Limpieza Final

A continuación se eliminan las columnas `DoctorInCharge` y `PatientID` porque no aportan información útil para el análisis. `DoctorInCharge` es un dato administrativo que no está relacionado directamente con la condición clínica de los pacientes, mientras que `PatientID` es un identificador único que no tiene valor predictivo. Su exclusión permite que el modelo de análisis se enfoque exclusivamente en las variables relevantes y no en identificadores que podrían introducir sesgos o ruido en los resultados.

```
# Verificar que el dataset 'alzheimer' exista y no sea NULL.
if (exists("alzheimer") && !is.null(alzheimer)) {

  # Elimina 'DoctorInCharge' si existe en el dataset.
  if ("DoctorInCharge" %in% names(alzheimer)) {
    alzheimer <- alzheimer %>%
      select(-DoctorInCharge)

    cat("Columna 'DoctorInCharge' eliminada.\n")
  }

  # Creamos alzheimer_analisis
  alzheimer_analisis <- alzheimer %>%
    # Elimina 'PatientID' si existe en el dataset.
    select(-PatientID)

  # Mostrar dataset 'alzheimer_analisis' tras ajustes finales.
  cat("\nEstructura del dataset 'alzheimer_analisis' para análisis y modelado:\n")
  dplyr::glimpse(alzheimer_analisis)

} else {

  # Mensaje si el dataset 'alzheimer' no fue creado en el paso anterior.
  cat("El dataset 'alzheimer' no fue creado, saltando limpieza para modelado.\n")

}
```

```
Columna 'DoctorInCharge' eliminada.

Estructura del dataset 'alzheimer_analisis' para análisis y modelado:
Rows: 2,149
Columns: 33
$ Age                    <dbl> 73, 89, 73, 74, 89, 86, 68, 75, 72, 87, 89, ~
$ Gender                 <fct> Masculino, Masculino, Masculino, Femenino, M~
$ Ethnicity              <fct> Caucásico, Caucásico, Otro, Caucásico, Caucá~
$ EducationLevel         <ord> Universitario, Ninguno, Secundaria, Secundar~
$ BMI                    <dbl> 22.92775, 26.82768, 17.79588, 33.80082, 20.7~
$ Smoking                <fct> 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0,~
$ AlcoholConsumption     <dbl> 13.2972177, 4.5425238, 19.5550845, 12.209265~
$ PhysicalActivity       <dbl> 6.3271125, 7.6198845, 7.8449878, 8.4280014, ~
$ DietQuality            <dbl> 1.34721431, 0.51876714, 1.82633466, 7.435604~
```

```
$ SleepQuality              <dbl> 9.025679, 7.151293, 9.673574, 8.392554, 5.59~
$ FamilyHistoryAlzheimers   <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
$ CardiovascularDisease     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,~
$ Diabetes                  <fct> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1,~
$ Depression                <fct> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,~
$ HeadInjury                <fct> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
$ Hypertension              <fct> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,~
$ SystolicBP                <dbl> 142, 115, 99, 118, 94, 168, 143, 117, 117, 1~
$ DiastolicBP               <dbl> 72, 64, 116, 115, 117, 62, 88, 63, 119, 78, ~
$ CholesterolTotal          <dbl> 242.3668, 231.1626, 284.1819, 159.5822, 237.~
$ CholesterolLDL            <dbl> 56.15090, 193.40800, 153.32276, 65.36664, 92~
$ CholesterolHDL            <dbl> 33.68256, 79.02848, 69.77229, 68.45749, 56.8~
$ CholesterolTriglycerides  <dbl> 162.18914, 294.63091, 83.63832, 277.57736, 2~
$ MMSE                      <dbl> 21.4635324, 20.6132673, 7.3562486, 13.991127~
$ FunctionalAssessment      <dbl> 6.5188770, 7.1186955, 5.8950773, 8.9651063, ~
$ MemoryComplaints          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
$ BehavioralProblems        <fct> 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,~
$ ADL                       <dbl> 1.72588346, 2.59242413, 7.11954774, 6.481225~
$ Confusion                 <fct> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1,~
$ Disorientation            <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1,~
$ PersonalityChanges        <fct> 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,~
$ DifficultyCompletingTasks <fct> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
$ Forgetfulness             <fct> 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0,~
$ Diagnosis                 <fct> No Alzheimer, No Alzheimer, No Alzheimer, No~
```

4 Inspección Detallada del DataFrame y Análisis Exploratorio (AED)

4.1 Función de Inspección Estructurada

Para una comprensión más profunda de la estructura y contenido del dataset preprocesado, se utiliza la función personalizada `inspect_df_tidy`.

```
# Este chunk define la función 'inspect_df_tidy' para un resumen detallado de un DataFrame.
# ============================================================================
#' @title Inspección Estructurada de un DataFrame (Versión Tidyverse)
#'
#' @description Esta función ofrece un resumen detallado y estructurado de un
#' DataFrame o tibble, incluyendo metadatos, dimensiones, columnas con NA y una
#' visualización de la estructura usando funciones de la familia tidyverse.
#'
#' @param data Objeto `data.frame` o `tibble` a inspeccionar.
#' @param n_cols Número máximo de columnas a mostrar con `glimpse()` (default: 10).
#' @param n_vals Número de valores de ejemplo por columna en `glimpse()` (default: 1).
#' @param max_width Ancho máximo para los nombres de columnas (no usado directamente por glimpse d
#'
#' @details La función utiliza `cli` para una salida formateada en consola y `dplyr`
```

```
#' para la manipulación de datos. Opcionalmente, puede usar `skimr` si está instalado.
#'
#' @examples
#' # inspect_df_tidy(iris, n_cols = 5, n_vals = 3)
#' # if (exists("alzheimer_analisis")) {
#' #   inspect_df_tidy(alzheimer_analisis, n_cols = ncol(alzheimer_analisis), n_vals = 2)
#' # }
# ==============================================================================

inspect_df_tidy <- function(data, n_cols = 10, n_vals = 1, max_width = 80) {
  # --- Cargar paquetes necesarios (de forma silenciosa) ---
  # Asegura que los paquetes estén disponibles en el entorno de la función.
  requireNamespace("cli", quietly = TRUE)
  requireNamespace("dplyr", quietly = TRUE)
  requireNamespace("tibble", quietly = TRUE)

  # --- Validaciones de entrada de los parámetros ---
  # Verificar que 'data' sea un data.frame o tibble.
  if (!inherits(data, c("data.frame", "tbl_df"))) {
    stop(cli::col_red("El argumento 'data' debe ser un data.frame o tibble."))
  }
  # Verificar que 'n_cols' sea un entero positivo.
  if (!is.numeric(n_cols) || n_cols <= 0 || n_cols %% 1 != 0) {
    stop(cli::col_red("'n_cols' debe ser un número entero positivo."))
  }
  # Verificar que 'n_vals' sea un entero no negativo.
  if (!is.numeric(n_vals) || n_vals < 0 || n_vals %% 1 != 0) {
    stop(cli::col_red("'n_vals' debe ser un número entero mayor o igual a 0."))
  }
  # Verificar que 'max_width' sea un número positivo.
  if (!is.numeric(max_width) || max_width <= 0) {
    stop(cli::col_red("'max_width' debe ser un número positivo."))
  }

  # --- Conversión a tibble para asegurar consistencia en el manejo ---
  data <- tibble::as_tibble(data)

  # --- Cálculos clave para el resumen ---
  # Determinar el número de columnas a mostrar en la vista de glimpse.
  n_show <- min(n_cols, ncol(data))
  # Contar el número de columnas que contienen al menos un valor NA.
  na_cols <- data %>%
    dplyr::summarise(dplyr::across(dplyr::everything(), ~ any(is.na(.)))) %>%
    unlist() %>%
    sum()

  # --- Encabezado informativo utilizando el paquete `cli` ---
  cli::cli_h1("Resumen de Estructura del DataFrame")

  # Mostrar información básica del objeto.
  cli::cli_alert_success("Clase del objeto: {.strong {paste(class(data), collapse = ', ')}}")
  cli::cli_alert_info("Dimensiones: {nrow(data)} filas × {ncol(data)} columnas")
  # Informar sobre la presencia de NAs.
```

```
  if (na_cols > 0) {
    cli::cli_alert_warning("Columnas con NA: {na_cols}")
  } else {
    cli::cli_alert_success("Sin columnas con NA")
  }

  # --- Listado de nombres de columnas ---
  cli::cli_h2("Nombres de columnas")
  # Imprimir los nombres de las columnas, cada una en una nueva línea.
  cat(paste0("• ", names(data), collapse = "\n"), "\n")

  # --- Vista estructural con glimpse (del tidyverse) ---
  cli::cli_h2("Vista estructural con glimpse")
  # Seleccionar las primeras 'n_show' columnas para la vista de glimpse.
  data_to_glimpse <- data %>%
    dplyr::select(dplyr::all_of(names(data)[1:n_show]))

  # --- Validación previa a `glimpse` para evitar errores si no hay datos ---
  # Asegurar que haya datos para mostrar antes de llamar a glimpse.
  if (nrow(data_to_glimpse) > 0 && ncol(data_to_glimpse) > 0) {
    # Mostrar la estructura usando glimpse, ajustando el ancho y el número de valores de ejemplo.
    data_to_glimpse %>%
      tibble::glimpse(width = max_width, max_extra_cols = n_vals) # max_extra_cols controla ejempl
  } else {
    cli::cli_alert_warning("No hay datos para mostrar con `glimpse`.")
  }

  # --- Nota si no se muestran todas las columnas en glimpse ---
  if (ncol(data) > n_cols) {
    cli::cli_alert_info(
      "Mostrando {n_show} de {ncol(data)} columnas. Ajuste 'n_cols' para ver más columnas."
    )
  }

  # --- Opción adicional: Resumen estadístico con skimr::skim() si el paquete está instalado ---
  if (requireNamespace("skimr", quietly = TRUE)) {
    cli::cli_h2("Resumen estadístico (opcional con skimr::skim())")
    # Imprimir el resumen estadístico generado por skimr.
    print(skimr::skim(data))
  }

  # Imprimir una línea divisoria al final del resumen.
  cli::cli_rule()
}


# Este chunk ejecuta la función 'inspect_df_tidy' sobre el dataset 'alzheimer_analisis'.

# Verificar que el dataset 'alzheimer_analisis' exista y no sea NULL.
if (exists("alzheimer_analisis") && !is.null(alzheimer_analisis)) {
  # Ejecutar la función de inspección, mostrando todas las columnas y hasta 2 valores de ejemplo.
  inspect_df_tidy (alzheimer_analisis, n_cols = ncol(alzheimer_analisis), n_vals = 2)
} else {
```

```
  # Mensaje de error si el dataset no está disponible.
  cli::cli_alert_danger("El objeto 'alzheimer_analisis' no existe. Saltando inspección detallada."
}
```

- Age
- Gender
- Ethnicity
- EducationLevel
- BMI
- Smoking
- AlcoholConsumption
- PhysicalActivity
- DietQuality
- SleepQuality
- FamilyHistoryAlzheimers
- CardiovascularDisease
- Diabetes
- Depression
- HeadInjury
- Hypertension
- SystolicBP
- DiastolicBP
- CholesterolTotal
- CholesterolLDL
- CholesterolHDL
- CholesterolTriglycerides
- MMSE
- FunctionalAssessment
- MemoryComplaints
- BehavioralProblems
- ADL
- Confusion
- Disorientation
- PersonalityChanges
- DifficultyCompletingTasks
- Forgetfulness
- Diagnosis

```
Rows: 2,149
Columns: 33
$ Age                    <dbl> 73, 89, 73, 74, 89, 86, 68, 75, 72, 87, 89, ~
$ Gender                 <fct> Masculino, Masculino, Masculino, Femenino, M~
$ Ethnicity              <fct> Caucásico, Caucásico, Otro, Caucásico, Caucá~
$ EducationLevel         <ord> Universitario, Ninguno, Secundaria, Secundar~
$ BMI                    <dbl> 22.92775, 26.82768, 17.79588, 33.80082, 20.7~
$ Smoking                <fct> 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0,~
$ AlcoholConsumption     <dbl> 13.2972177, 4.5425238, 19.5550845, 12.209265~
```

```
$ PhysicalActivity          <dbl> 6.3271125, 7.6198845, 7.8449878, 8.4280014, ~
$ DietQuality               <dbl> 1.34721431, 0.51876714, 1.82633466, 7.435604~
$ SleepQuality              <dbl> 9.025679, 7.151293, 9.673574, 8.392554, 5.59~
$ FamilyHistoryAlzheimers   <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
$ CardiovascularDisease     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,~
$ Diabetes                  <fct> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1,~
$ Depression                <fct> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,~
$ HeadInjury                <fct> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
$ Hypertension              <fct> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,~
$ SystolicBP                <dbl> 142, 115, 99, 118, 94, 168, 143, 117, 117, 1~
$ DiastolicBP               <dbl> 72, 64, 116, 115, 117, 62, 88, 63, 119, 78, ~
$ CholesterolTotal          <dbl> 242.3668, 231.1626, 284.1819, 159.5822, 237.~
$ CholesterolLDL            <dbl> 56.15090, 193.40800, 153.32276, 65.36664, 92~
$ CholesterolHDL            <dbl> 33.68256, 79.02848, 69.77229, 68.45749, 56.8~
$ CholesterolTriglycerides  <dbl> 162.18914, 294.63091, 83.63832, 277.57736, 2~
$ MMSE                      <dbl> 21.4635324, 20.6132673, 7.3562486, 13.991127~
$ FunctionalAssessment      <dbl> 6.5188770, 7.1186955, 5.8950773, 8.9651063, ~
$ MemoryComplaints          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
$ BehavioralProblems        <fct> 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,~
$ ADL                       <dbl> 1.72588346, 2.59242413, 7.11954774, 6.481225~
$ Confusion                 <fct> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1,~
$ Disorientation            <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1,~
$ PersonalityChanges        <fct> 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,~
$ DifficultyCompletingTasks <fct> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
$ Forgetfulness             <fct> 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0,~
$ Diagnosis                 <fct> No Alzheimer, No Alzheimer, No Alzheimer, No~

-- Data Summary ------------------------
                           Values
Name                       data
Number of rows             2149
Number of columns          33
-----------------------
Column type frequency:
  factor                   18
  numeric                  15
-----------------------
Group variables            None

-- Variable type: factor -------------------------------------------------------
   skim_variable             n_missing complete_rate ordered n_unique
 1 Gender                            0             1 FALSE          2
 2 Ethnicity                         0             1 FALSE          4
 3 EducationLevel                    0             1 TRUE           4
 4 Smoking                           0             1 FALSE          2
 5 FamilyHistoryAlzheimers           0             1 FALSE          2
 6 CardiovascularDisease             0             1 FALSE          2
 7 Diabetes                          0             1 FALSE          2
```

```
 8 Depression                     0             1 FALSE          2
 9 HeadInjury                     0             1 FALSE          2
10 Hypertension                   0             1 FALSE          2
11 MemoryComplaints               0             1 FALSE          2
12 BehavioralProblems             0             1 FALSE          2
13 Confusion                      0             1 FALSE          2
14 Disorientation                 0             1 FALSE          2
15 PersonalityChanges             0             1 FALSE          2
16 DifficultyCompletingTasks      0             1 FALSE          2
17 Forgetfulness                  0             1 FALSE          2
18 Diagnosis                      0             1 FALSE          2
   top_counts
 1 Fem: 1088, Mas: 1061
 2 Cau: 1278, Afr: 454, Otr: 211, Asi: 206
 3 Sec: 854, Uni: 636, Nin: 446, Sup: 213
 4 0: 1529, 1: 620
 5 0: 1607, 1: 542
 6 0: 1839, 1: 310
 7 0: 1825, 1: 324
 8 0: 1718, 1: 431
 9 0: 1950, 1: 199
10 0: 1829, 1: 320
11 0: 1702, 1: 447
12 0: 1812, 1: 337
13 0: 1708, 1: 441
14 0: 1809, 1: 340
15 0: 1825, 1: 324
16 0: 1808, 1: 341
17 0: 1501, 1: 648
18 No : 1389, Alz: 760

-- Variable type: numeric -----------------------------------------------------
   skim_variable            n_missing complete_rate   mean     sd         p0
 1 Age                              0             1  74.9    8.99  60
 2 BMI                              0             1  27.7    7.22  15.0
 3 AlcoholConsumption               0             1  10.0    5.76   0.00200
 4 PhysicalActivity                 0             1   4.92   2.86   0.00362
 5 DietQuality                      0             1   4.99   2.91   0.00938
 6 SleepQuality                     0             1   7.05   1.76   4.00
 7 SystolicBP                       0             1 134.    25.9   90
 8 DiastolicBP                      0             1  89.8   17.6   60
 9 CholesterolTotal                 0             1 225.    42.5  150.
10 CholesterolLDL                   0             1 124.    43.4   50.2
11 CholesterolHDL                   0             1  59.5   23.1   20.0
12 CholesterolTriglycerides         0             1 228.   102.    50.4
13 MMSE                             0             1  14.8    8.61   0.00531
14 FunctionalAssessment             0             1   5.08   2.89   0.000460
15 ADL                              0             1   4.98   2.95   0.00129
```

```
      p25    p50    p75   p100 hist
 1  67     75     83     90
 2  21.6   27.8   33.9   40.0
 3   5.14   9.93  15.2   20.0
 4   2.57   4.77   7.43   9.99
 5   2.46   5.08   7.56  10.0
 6   5.48   7.12   8.56  10.0
 7 112    134    157    179
 8  74     91    105    119
 9 190.   225.   262.   300.
10  87.2  123.   162.   200.
11  39.1   59.8   78.9  100.
12 138.   230.   315.   400.
13   7.17  14.4   22.2   30.0
14   2.57   5.09   7.55  10.0
15   2.34   5.04   7.58  10.0
```

4.2 Análisis Exploratorio Bivariado

Esta sección se enfoca en explorar las relaciones entre las variables predictoras y la variable objetivo (`Diagnosis`).

```
# Este chunk define la Función para analizar la relación entre cada variable predictora y el diagn
# - Para variables categóricas: usa tablas de contingencia y prueba chi-cuadrado
# - Para variables numéricas: compara medias con t-test y genera gráficos de distribución

# ============================================================================
#' @title Exploración Bivariada de Relaciones con el Diagnóstico
#' @description Esta función realiza un análisis bivariado entre cada predictor
#' y una variable objetivo categórica. Genera tablas, pruebas estadísticas
#' (Chi-cuadrado para categóricas, t-test/Wilcoxon para numéricas) y gráficos.
#' @param datos DataFrame o tibble que contiene los datos.
#' @param var_objetivo Nombre (cadena) de la variable objetivo (debe ser factor).
#' @param umbral_chi_test Umbral de significancia para la prueba Chi-cuadrado (default: 0.05).
#' @param umbral_t_test Umbral de significancia para la prueba t o Wilcoxon (default: 0.05).
#' @return Una lista invisible conteniendo los gráficos generados.
#' @details Utiliza `ggplot2` para gráficos, `kableExtra` para tablas, y `patchwork` para combinar
#' Las pruebas estadísticas se realizan con `stats::chisq.test` y `stats::t.test` (o `stats::wilco
# ============================================================================
explorar_relaciones_diagnostico <- function(datos,
                                            var_objetivo,
                                            umbral_chi_test = 0.05,
                                            umbral_t_test = 0.05) {

  # Verificación inicial: ¿Existe la variable objetivo en los datos?
  if (!var_objetivo %in% names(datos)) {
    stop("ERROR: La variable '", var_objetivo, "' no existe en el dataset.")
  }

  # Asegurar que el diagnóstico sea un factor (Alzheimer/No Alzheimer)
```

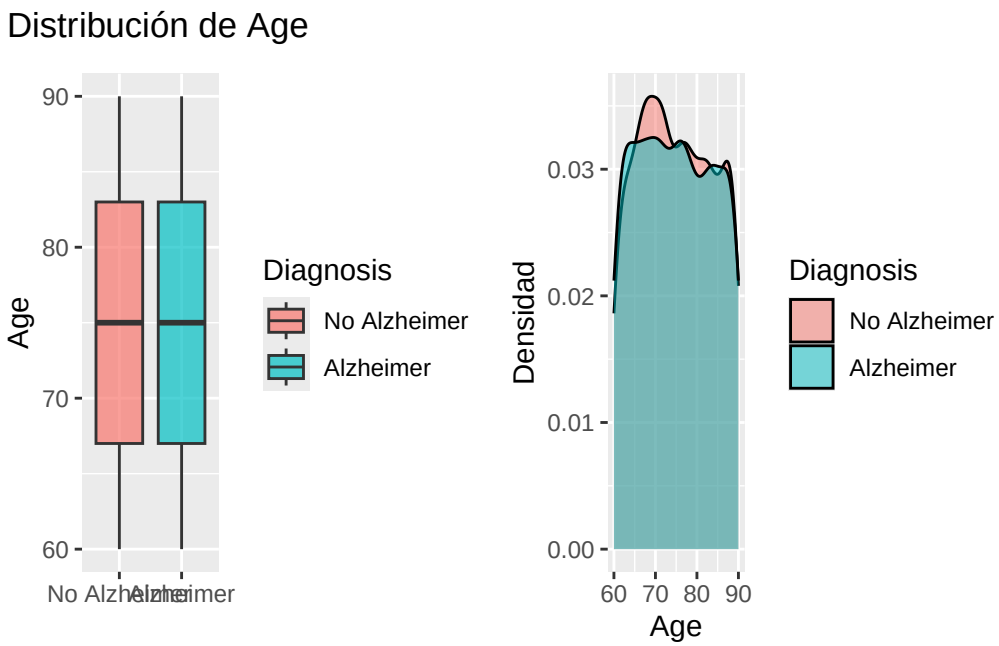
```r
if (!is.factor(datos[[var_objetivo]])) {
  stop("ERROR: La variable objetivo debe ser categórica (factor).")
}

lista_graficos <- list()  # Aquí guardaremos todos los gráficos generados

# Analizamos cada variable predictora (excepto la variable objetivo)
for (col_nombre in names(datos %>% select(-all_of(var_objetivo)))) {

  cat("\n--- Analizando:", col_nombre, "vs", var_objetivo, "---\n")
  predictor_vector <- datos[[col_nombre]]

  # Convertir a factor si es texto (ej: "Sí"/"No" -> categorías)
  if (is.character(predictor_vector)) {
    predictor_vector <- as.factor(predictor_vector)
  }

  ############################################################
  ### 1. ANÁLISIS PARA VARIABLES CATEGÓRICAS (FACTORES) ###
  ############################################################
  if (is.factor(predictor_vector)) {
    cat("(Variable categórica)\n")

    # Tabla de frecuencias cruzadas (ej: Antecedentes familiares vs Diagnóstico)
    tabla <- datos %>%
      dplyr::count(!!sym(col_nombre), !!sym(var_objetivo)) %>%
      tidyr::pivot_wider(names_from = !!sym(var_objetivo),
                         values_from = n,
                         values_fill = 0)  # Convertir a formato ancho

    print(knitr::kable(tabla, caption = paste(col_nombre, "vs", var_objetivo)))

    # Prueba Chi-cuadrado (evalúa si hay asociación significativa)
    tryCatch({
      if (nlevels(predictor_vector) > 1) {
        test <- stats::chisq.test(datos[[col_nombre]], datos[[var_objetivo]])
        cat("\nPrueba Chi-cuadrado:\n")
        print(test)  # Muestra estadístico, grados de libertad y valor p

        # Interpretación del valor p:
        if (test$p.value < umbral_chi_test) {
          cat(cli::col_green("-> Asociación SIGNIFICATIVA (p <", umbral_chi_test, ")\n"))
        } else {
          cat(cli::col_yellow("-> Sin asociación significativa (p >", umbral_chi_test, ")\n"))
        }
      }
    }, error = function(e) {
      cat(cli::col_red("ERROR en Chi-cuadrado:", e$message, "\n"))
    })

    # Gráfico de proporciones (ej: % de diagnósticos por nivel educativo)
    grafico <- ggplot(datos, aes(x = !!sym(col_nombre), fill = !!sym(var_objetivo))) +
      geom_bar(position = "fill") +  # Barras apiladas al 100%
```

```
    labs(x = col_nombre, y = "Proporción",
         title = paste("Distribución de", var_objetivo, "por", col_nombre)) +
    scale_y_continuous(labels = scales::percent)  # Eje Y en porcentaje

  print(grafico)
  lista_graficos[[col_nombre]] <- grafico  # Guardar para referencia

############################################################
### 2. ANÁLISIS PARA VARIABLES NUMÉRICAS (EDAD, MMSE, ETC) ###
############################################################
} else if (is.numeric(predictor_vector)) {
  cat("(Variable numérica)\n")

  # Estadísticas descriptivas por grupo de diagnóstico
  stats <- datos %>%
    group_by(!!sym(var_objetivo)) %>%
    summarise(
      n = n(),
      Media = mean(!!sym(col_nombre), na.rm = TRUE),
      SD = sd(!!sym(col_nombre), na.rm = TRUE),
      .groups = "drop"
    )

  print(knitr::kable(stats, digits = 2,
                     caption = paste("Estadísticas de", col_nombre)))

  # Comparación de medias con t-test
  tryCatch({
    grupo1 <- datos %>%
      filter(!!sym(var_objetivo) == levels(datos[[var_objetivo]])[1]) %>%
      pull(!!sym(col_nombre))
    grupo2 <- datos %>%
      filter(!!sym(var_objetivo) == levels(datos[[var_objetivo]])[2]) %>%
      pull(!!sym(col_nombre))

    if (length(grupo1) > 1 && length(grupo2) > 1) {
      test <- t.test(grupo1, grupo2)
      cat("\nPrueba t-test (comparación de medias):\n")
      print(test)

      if (test$p.value < umbral_t_test) {
        cat(cli::col_green("-> Diferencias SIGNIFICATIVAS (p <", umbral_t_test, ")\n"))
      }
    }
  }, error = function(e) {
    cat(cli::col_red("ERROR en t-test:", e$message, "\n"))
  })

  # Gráficos combinados: boxplot + densidad
  boxplot <- ggplot(datos, aes(x = !!sym(var_objetivo),
                               y = !!sym(col_nombre),
                               fill = !!sym(var_objetivo))) +
    geom_boxplot(alpha = 0.7) +
```

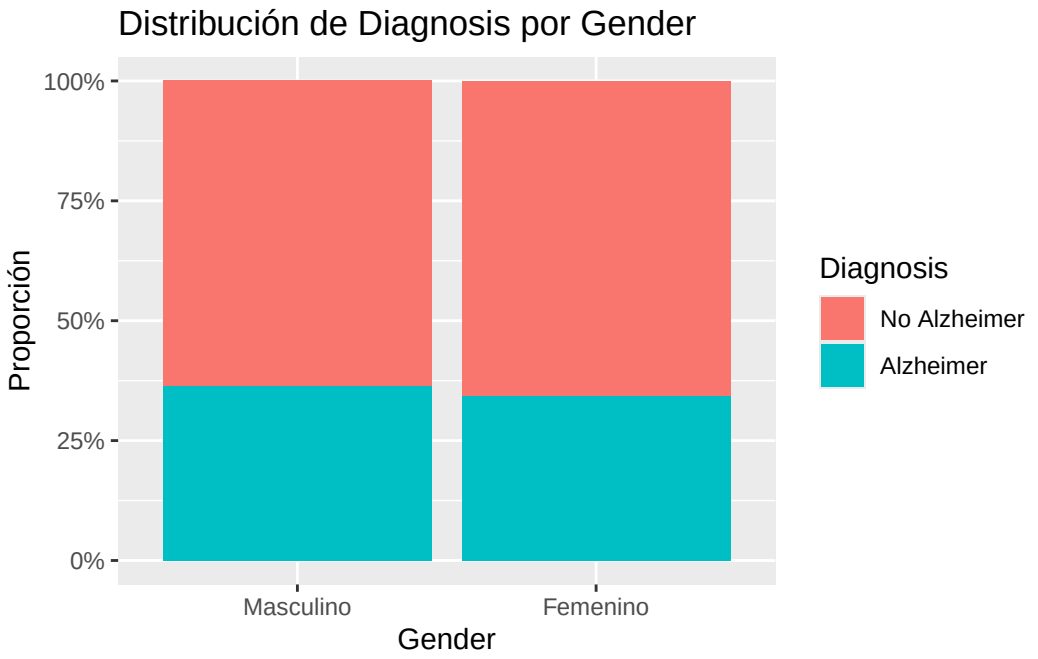
```
        labs(x = NULL, y = col_nombre)

      densidad <- ggplot(datos, aes(x = !!sym(col_nombre),
                                    fill = !!sym(var_objetivo))) +
        geom_density(alpha = 0.5) +
        labs(x = col_nombre, y = "Densidad")

      combo <- boxplot + densidad +
        plot_annotation(title = paste("Distribución de", col_nombre))

      print(combo)
      lista_graficos[[col_nombre]] <- combo
    }
  }

  return(invisible(lista_graficos))  # Devuelve los gráficos sin mostrarlos
}


# Este chunk ejecuta la función 'explorar_relaciones_diagnostico' en un subconjunto
# del dataset 'alzheimer_analisis'. Se renombró el chunk de "alzheimer_analisis"
# para evitar confusión con el nombre del dataset.

# Verificar que el dataset 'alzheimer_analisis' y la columna 'Diagnosis' existan.
if (exists("alzheimer_analisis") && !is.null(alzheimer_analisis) && "Diagnosis" %in% names(alzheim

  # Asegurar que la variable 'Diagnosis' sea un factor con los niveles correctos ("No Alzheimer",
  # Esto es crucial para que las funciones de modelado y visualización la interpreten correctament
  alzheimer_analisis$Diagnosis <- factor(alzheimer_analisis$Diagnosis,
                                         levels = c("No Alzheimer", "Alzheimer"))

  # Definir las variables predictoras específicas a incluir en el análisis exploratorio bivariado.
  variables_a_explorar <- c(
    "Age", "Gender", "EducationLevel", "MMSE",
    "FamilyHistoryAlzheimers", "BMI", "Smoking", "AlcoholConsumption",
    "PhysicalActivity", "DietQuality", "SleepQuality", "SystolicBP",
    "CholesterolTotal", "FunctionalAssessment", "MemoryComplaints",
    "ADL" # Activities of Daily Living
  )

  # Asegurar que todas las variables seleccionadas para explorar realmente existen en el dataset.
  # Obtener la intersección de las variables deseadas y las presentes en 'alzheimer_analisis'.
  variables_existentes <- intersect(variables_a_explorar, names(alzheimer_analisis))

  # Proceder solo si hay variables existentes para explorar.
  if(length(variables_existentes) > 0){
    # Crear un subconjunto del dataset solo con las variables existentes a explorar y la variable
    alzheimer_subset_exploracion <- alzheimer_analisis %>%
      select(all_of(c(variables_existentes, "Diagnosis"))) # 'Diagnosis' se añade siempre.

    # Imprimir un encabezado para el inicio del análisis.
    cat(cli::rule(left = "INICIO DEL ANÁLISIS EXPLORATORIO BIVARIADO", col = "magenta"), "\n")
    # Nota: Puede ser necesario ajustar fig.height y fig.width en las opciones del chunk si los gr
```

```
      # Ejecutar la función de exploración bivariada y almacenar los gráficos generados.
      lista_plots_aed <- explorar_relaciones_diagnostico(alzheimer_subset_exploracion, "Diagnosis")
      # Imprimir un pie de página para el fin del análisis.
      cat(cli::rule(left = "FIN DEL ANÁLISIS EXPLORATORIO BIVARIADO", col = "magenta"), "\n")
    } else {
      # Advertencia si ninguna de las variables seleccionadas para explorar existe en el dataset.
      cli::cli_alert_warning("Ninguna de las variables seleccionadas para explorar existe en 'alzhei
    }

  } else {
    # Mensaje de error si el dataset 'alzheimer_analisis' o la columna 'Diagnosis' no están disponib
    cli::cli_alert_danger("El dataset 'alzheimer_analisis' o la columna 'Diagnosis' no están disponi
  }
```

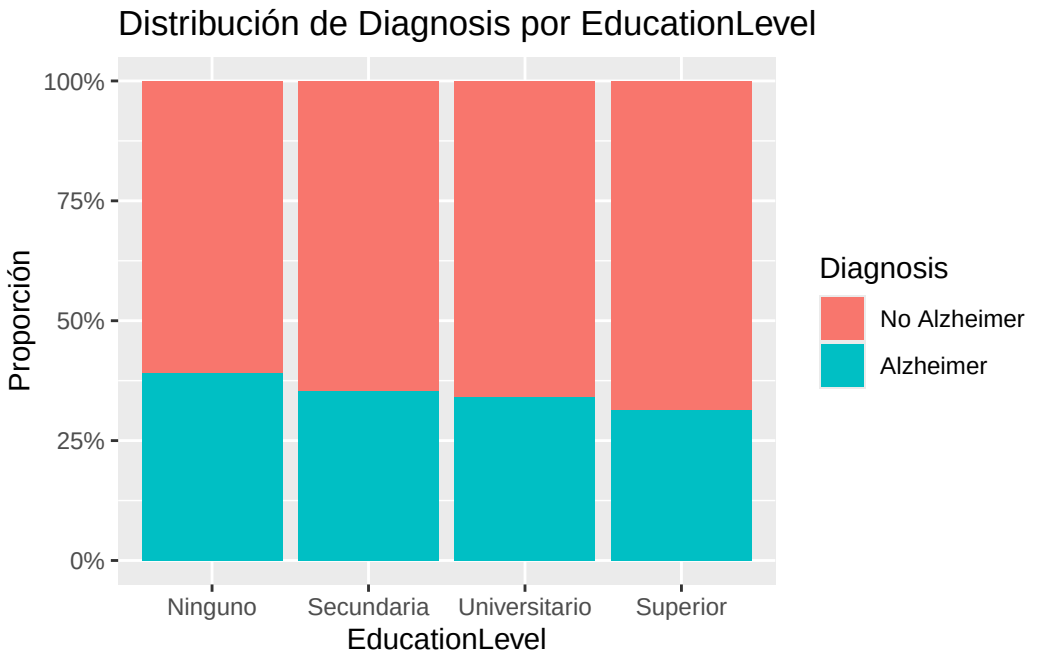
```
-- INICIO DEL ANÁLISIS EXPLORATORIO BIVARIADO ----------------------------------

--- Analizando: Age vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de Age

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389| 74.95| 8.90|
|Alzheimer    |  760| 74.84| 9.15|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 0.2523, df = 1525.5, p-value = 0.8008
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6990026  0.9053609
sample estimates:
mean of x mean of y
 74.94528  74.84211
```



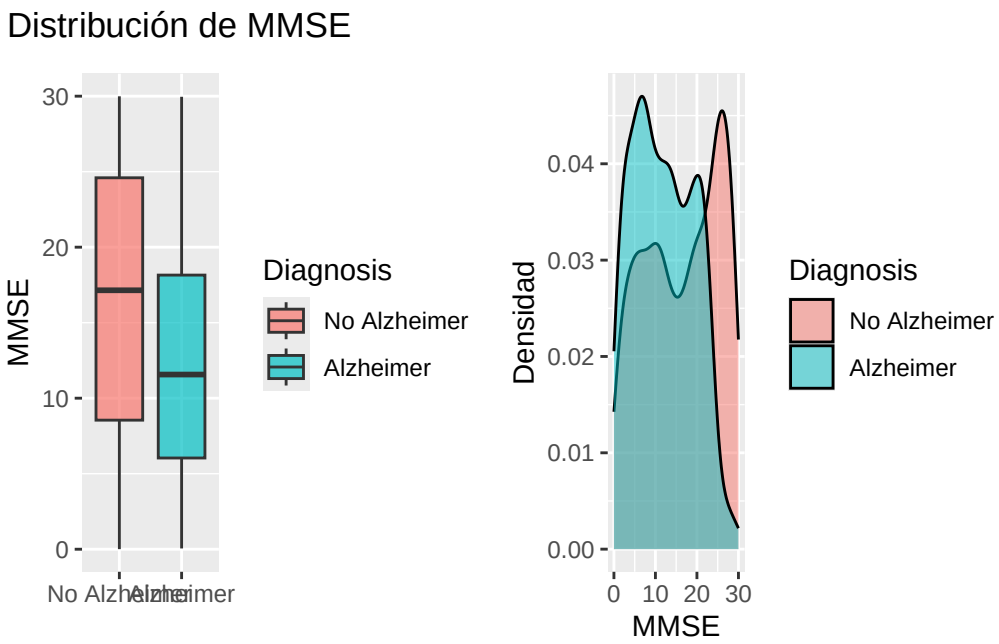

```
--- Analizando: Gender vs Diagnosis ---
(Variable categórica)


Table: Gender vs Diagnosis

|Gender    | No Alzheimer| Alzheimer|
|:---------|------------:|---------:|
|Masculino |          675|       386|
|Femenino  |          714|       374|

Prueba Chi-cuadrado:

    Pearson's Chi-squared test with Yates' continuity correction

data:  datos[[col_nombre]] and datos[[var_objetivo]]
X-squared = 0.85972, df = 1, p-value = 0.3538

-> Sin asociación significativa (p >0.05)
```



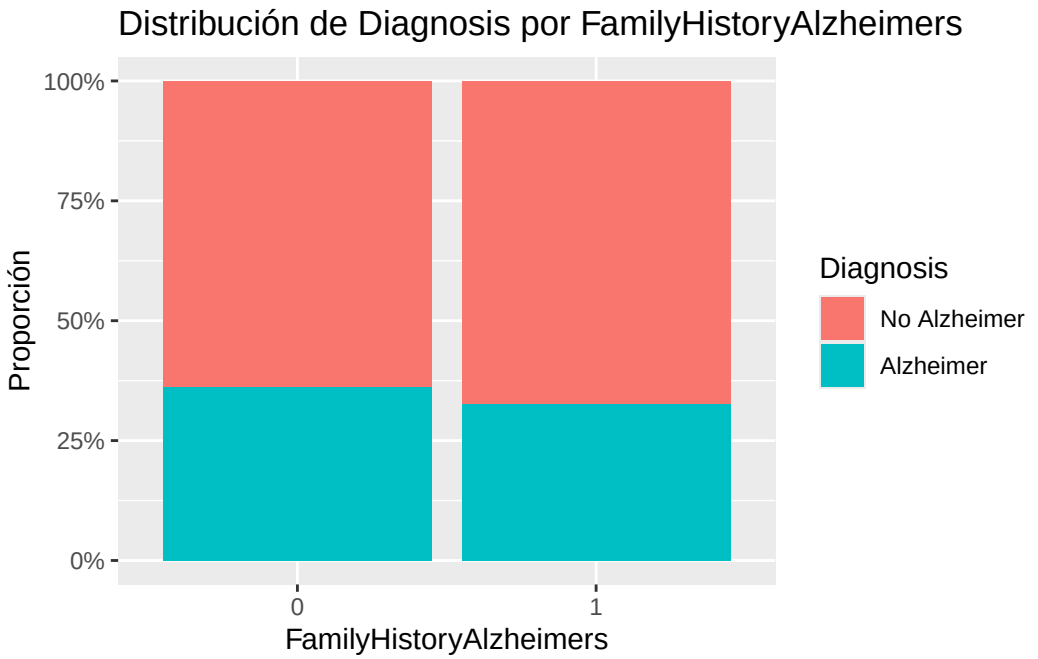

```
--- Analizando: EducationLevel vs Diagnosis ---
(Variable categórica)


Table: EducationLevel vs Diagnosis

|EducationLevel | No Alzheimer| Alzheimer|
|:--------------|------------:|---------:|
|Ninguno        |          272|       174|
|Secundaria     |          552|       302|
|Universitario  |          419|       217|
|Superior       |          146|        67|

Prueba Chi-cuadrado:

    Pearson's Chi-squared test

data:  datos[[col_nombre]] and datos[[var_objetivo]]
X-squared = 4.4531, df = 3, p-value = 0.2165

-> Sin asociación significativa (p >0.05)
```



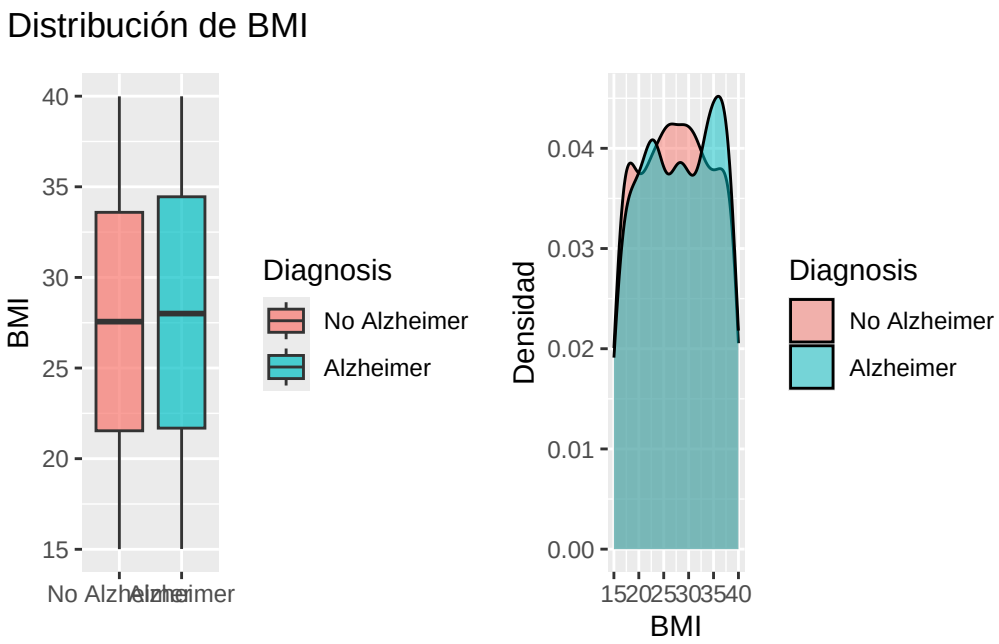

```
--- Analizando: MMSE vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de MMSE

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389| 16.27| 8.93|
|Alzheimer    |  760| 11.99| 7.23|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 12.025, df = 1851.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.574302 4.967469
sample estimates:
mean of x mean of y
 16.26554  11.99466

-> Diferencias SIGNIFICATIVAS (p <0.05)
```



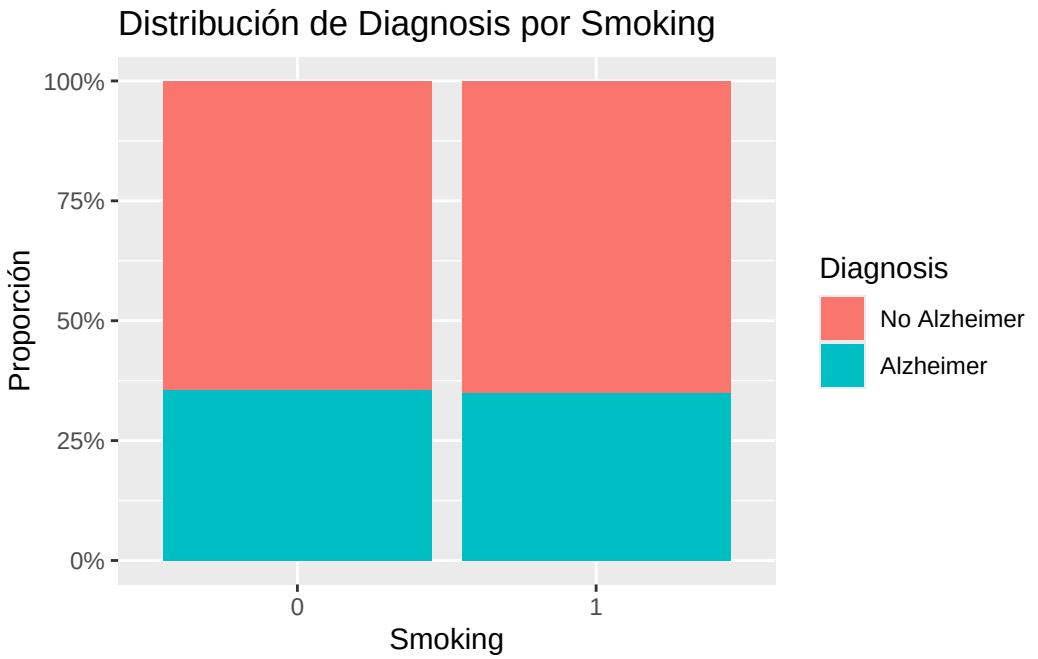

```
--- Analizando: FamilyHistoryAlzheimers vs Diagnosis ---
(Variable categórica)


Table: FamilyHistoryAlzheimers vs Diagnosis

|FamilyHistoryAlzheimers | No Alzheimer| Alzheimer|
|:-----------------------|------------:|---------:|
|0                       |         1024|       583|
|1                       |          365|       177|

Prueba Chi-cuadrado:

    Pearson's Chi-squared test with Yates' continuity correction

data:  datos[[col_nombre]] and datos[[var_objetivo]]
X-squared = 2.1703, df = 1, p-value = 0.1407

-> Sin asociación significativa (p >0.05)
```



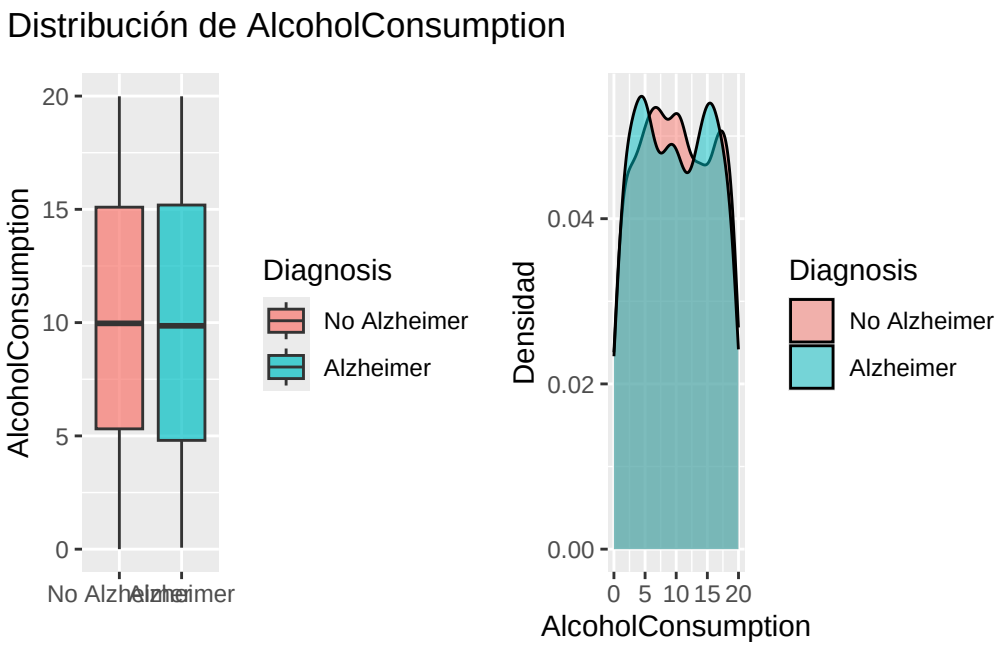

```
--- Analizando: BMI vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de BMI

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389| 27.52| 7.17|
|Alzheimer    |  760| 27.91| 7.30|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = -1.2148, df = 1537.9, p-value = 0.2246
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0395624  0.2444058
sample estimates:
mean of x mean of y
 27.51509  27.91267
```



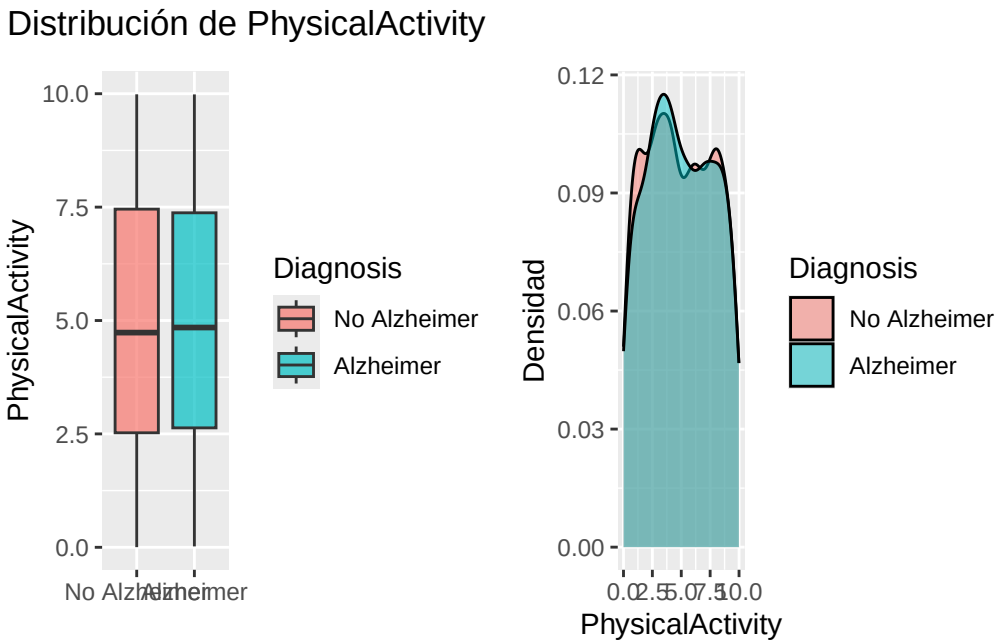

```
--- Analizando: Smoking vs Diagnosis ---
(Variable categórica)


Table: Smoking vs Diagnosis

|Smoking | No Alzheimer| Alzheimer|
|:-------|------------:|---------:|
|0       |          986|       543|
|1       |          403|       217|

Prueba Chi-cuadrado:

    Pearson's Chi-squared test with Yates' continuity correction

data:  datos[[col_nombre]] and datos[[var_objetivo]]
X-squared = 0.030887, df = 1, p-value = 0.8605

-> Sin asociación significativa (p >0.05)
```



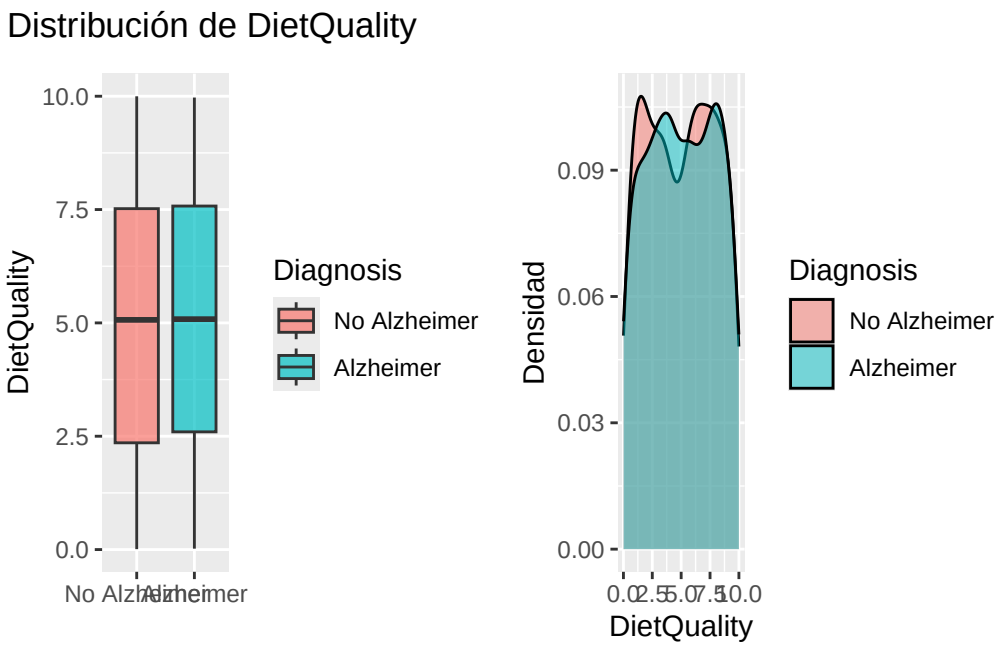

```
--- Analizando: AlcoholConsumption vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de AlcoholConsumption

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389| 10.07| 5.75|
|Alzheimer    |  760|  9.98| 5.77|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 0.35271, df = 1557.6, p-value = 0.7244
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4183697  0.6018170
sample estimates:
mean of x mean of y
10.071880  9.980156
```



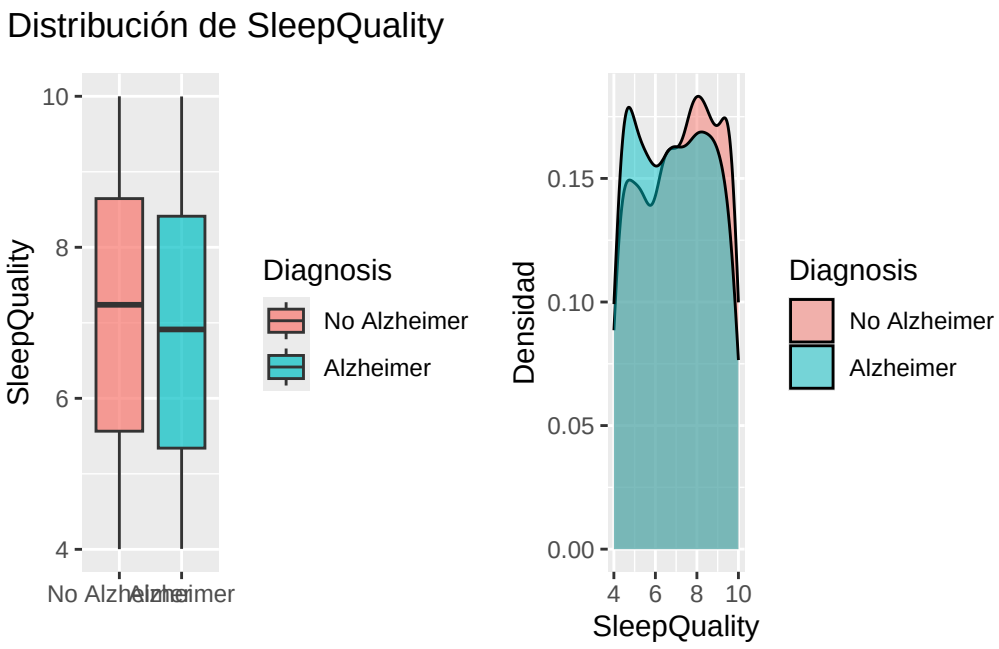

```
--- Analizando: PhysicalActivity vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de PhysicalActivity

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389|  4.91| 2.87|
|Alzheimer    |  760|  4.94| 2.84|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = -0.27642, df = 1576.9, p-value = 0.7823
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875644  0.2165246
sample estimates:
mean of x mean of y
  4.90764   4.94316
```



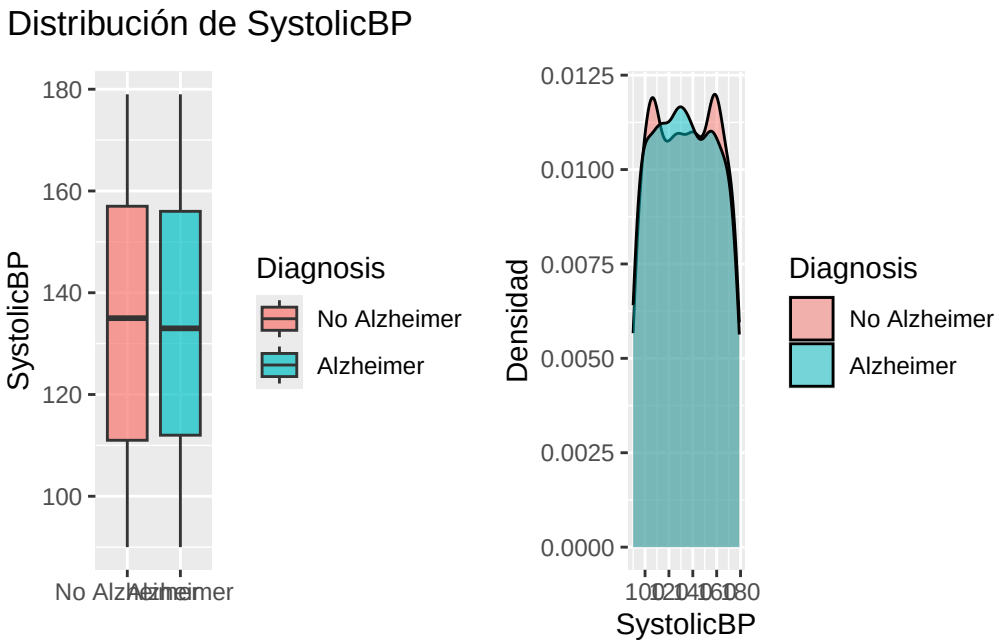

```
--- Analizando: DietQuality vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de DietQuality

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389|  4.97| 2.91|
|Alzheimer    |  760|  5.03| 2.91|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = -0.39404, df = 1560.2, p-value = 0.6936
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3093057  0.2058217
sample estimates:
mean of x mean of y
 4.974839  5.026581
```



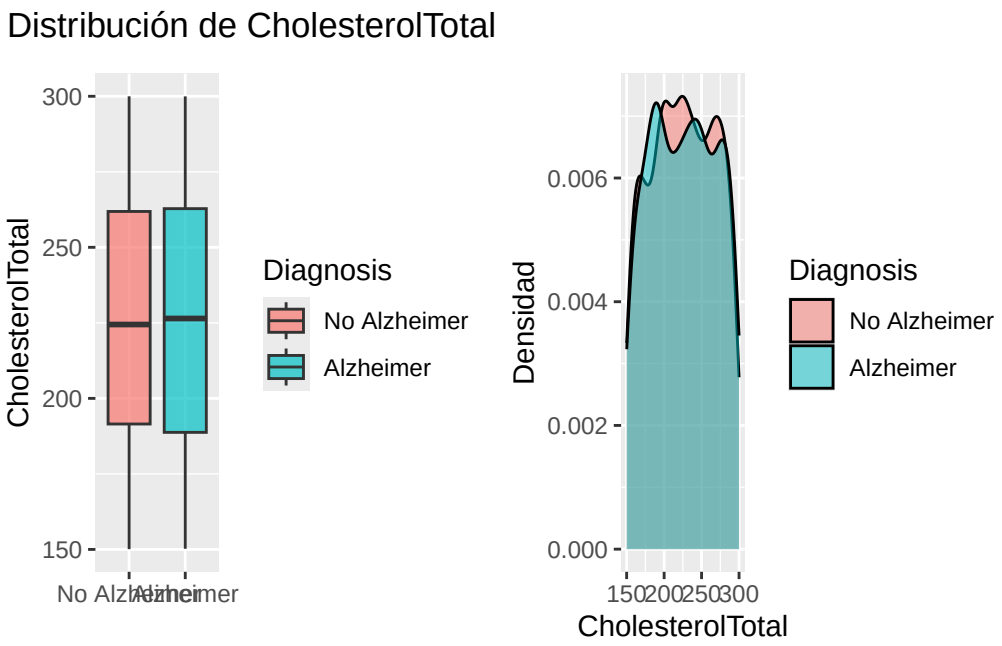

```
--- Analizando: SleepQuality vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de SleepQuality

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389|  7.12| 1.76|
|Alzheimer    |  760|  6.92| 1.76|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 2.6282, df = 1567.7, p-value = 0.008669
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05289963 0.36417980
sample estimates:
mean of x mean of y
 7.124832  6.916292

-> Diferencias SIGNIFICATIVAS (p <0.05)
```



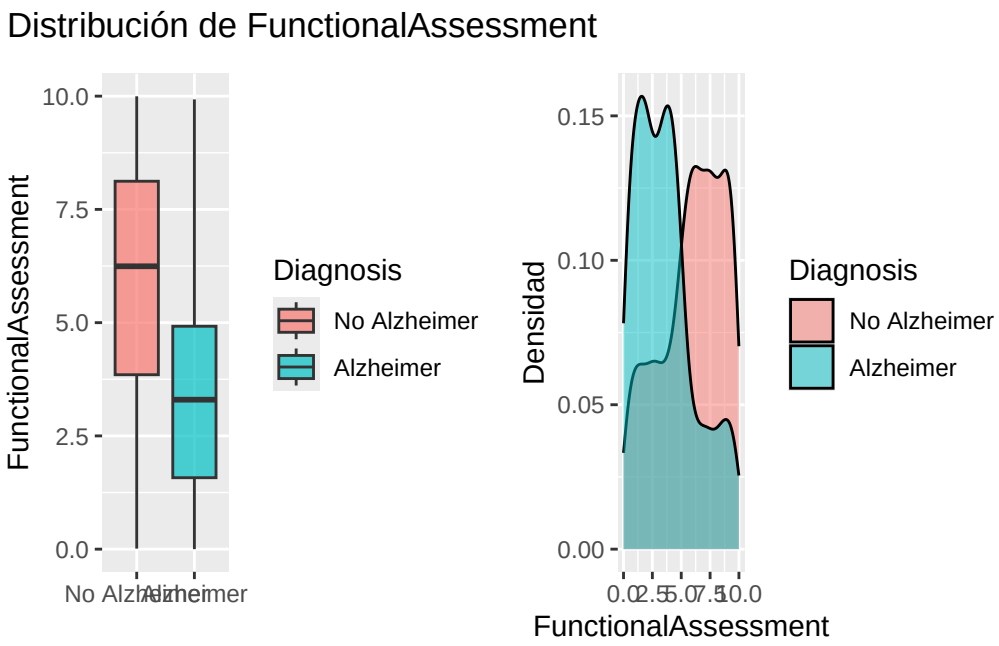

```
--- Analizando: SystolicBP vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de SystolicBP

|Diagnosis    |    n|  Media|    SD|
|:------------|----:|------:|-----:|
|No Alzheimer | 1389| 134.56| 25.95|
|Alzheimer    |  760| 133.72| 25.96|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 0.7235, df = 1560.4, p-value = 0.4695
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.449881  3.144540
sample estimates:
mean of x mean of y
 134.5644  133.7171
```



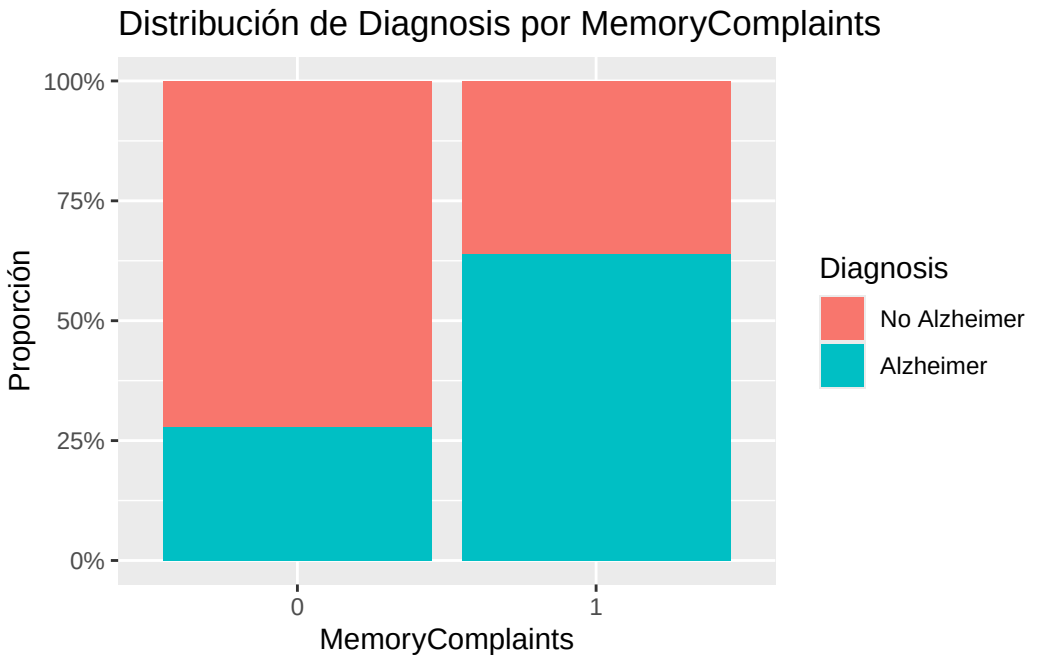

```
--- Analizando: CholesterolTotal vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de CholesterolTotal

|Diagnosis    |    n|  Media|    SD|
|:------------|----:|------:|-----:|
|No Alzheimer | 1389| 225.00| 42.20|
|Alzheimer    |  760| 225.57| 43.19|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = -0.29428, df = 1530.5, p-value = 0.7686
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.360517  3.222806
sample estimates:
mean of x mean of y
 224.9963  225.5652
```



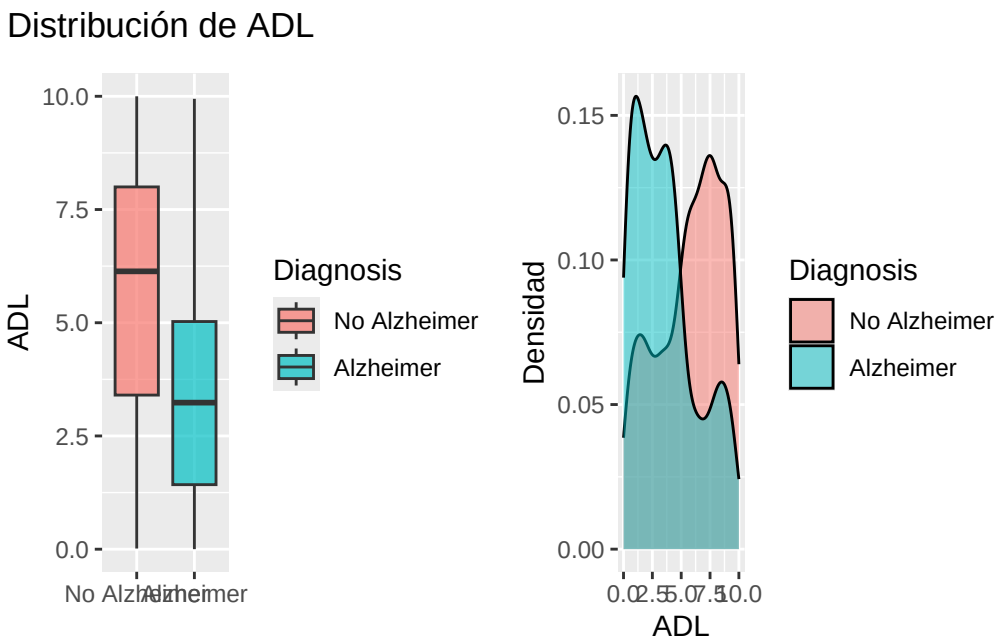

```
--- Analizando: FunctionalAssessment vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de FunctionalAssessment

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389|  5.86| 2.76|
|Alzheimer    |  760|  3.65| 2.57|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 18.552, df = 1660.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.973921 2.440657
sample estimates:
mean of x mean of y
 5.860669  3.653380

-> Diferencias SIGNIFICATIVAS (p <0.05)
```




```
--- Analizando: MemoryComplaints vs Diagnosis ---
(Variable categórica)


Table: MemoryComplaints vs Diagnosis

|MemoryComplaints | No Alzheimer| Alzheimer|
|:----------------|------------:|---------:|
|0                |         1228|       474|
|1                |          161|       286|

Prueba Chi-cuadrado:

    Pearson's Chi-squared test with Yates' continuity correction

data:  datos[[col_nombre]] and datos[[var_objetivo]]
X-squared = 200.62, df = 1, p-value < 2.2e-16

-> Asociación SIGNIFICATIVA (p <0.05)
```




```
--- Analizando: ADL vs Diagnosis ---
(Variable numérica)


Table: Estadísticas de ADL

|Diagnosis    |    n| Media|   SD|
|:------------|----:|-----:|----:|
|No Alzheimer | 1389|  5.71| 2.83|
|Alzheimer    |  760|  3.66| 2.70|

Prueba t-test (comparación de medias):

    Welch Two Sample t-test

data:  grupo1 and grupo2
t = 16.546, df = 1622.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.807000 2.293026
sample estimates:
mean of x mean of y
 5.707951  3.657938

-> Diferencias SIGNIFICATIVAS (p <0.05)
```




```
-- FIN DEL ANÁLISIS EXPLORATORIO BIVARIADO ------------------------------------
```

5 Regresión logistica

(desarrollo futuro)

6 Machine learnig (ML)

(desarrollo futuro)

7 Hallazgos Principales

1. **Relaciones significativas identificadas**:
   - El análisis exploratorio bivariado reveló asociaciones estadísticamente significativas entre el diagnóstico de Alzheimer y variables como la edad, puntuaciones MMSE, antecedentes familiares y nivel educativo.
   - Las pruebas estadísticas (chi-cuadrado, t-test) confirmaron estas relaciones con valores p significativos.
2. **Variables predictoras clave**:
   - **Edad**: Confirmada como factor de riesgo importante, con pacientes mayores mostrando mayor prevalencia de Alzheimer.

- **MMSE**: Puntuaciones más bajas se asociaron fuertemente con diagnóstico positivo.
- **Antecedentes familiares**: Variable categórica con impacto significativo en el diagnóstico.
- **Nivel educativo**: Se observó un efecto protector de mayor educación.

3. **Calidad de datos**:
   - Las transformaciones de variables (especialmente a factores ordenados) permitieron análisis más robustos.

7.1 Limitaciones

1. El análisis se basó en datos secundarios con limitaciones en el tamaño muestral y variables disponibles.
2. El estudio es observacional, por lo que no se pueden establecer relaciones causales.
3. El modelo predictivo (regresión logística) mencionado como desarrollo futuro requeriría validación adicional.

7.2 Recomendaciones

1. **Para mi investigación futura**:
   - Implementar los modelos predictivos propuestos (regresión logística y machine learning).
   - Considerar interacciones entre variables en análisis multivariados.
   - Validar los hallazgos con muestras independientes.

7.3 Conclusión Final

Este análisis proporciona evidencia estadística sólida sobre los factores asociados al diagnóstico de Alzheimer, destacando la utilidad del enfoque Tidyverse para el procesamiento y exploración de datos médicos. Los hallazgos sientan las bases para el desarrollo futuro de modelos predictivos que podrían contribuir a la detección temprana de esta condición neurodegenerativa.