



UNIVERSIDAD
NEBRIJA

Máster de Estadística Aplicada con R

Análisis cluster no jerárquico de los indicadores de salud relacionados con el consumo de alcohol del año 2021 en los estados de EE. UU.

AUTOR: Teodoro José Martínez Arán

DIRECTOR: Juan Luis López Garrancho

FECHA: 14 agosto 2024

ENTIDAD COLABORADORA



Resumen

Justificación del análisis

El uso nocivo del alcohol es una de las causas de mortalidad evitable más importantes a nivel mundial. EE. UU. presenta una de las tasas de mortalidad relacionadas con alcohol más altas del mundo. Comprender los determinantes de salud que condicionan los valores de estos indicadores puede permitir el desarrollo de políticas de salud pública más efectivas, enfocadas a modificar los principales factores de riesgo.

Metodología utilizada

00. Configuración

En el análisis se utilizó el software **R version 4.4.1 (2024-06-14)**. Se generaron scripts de configuración e instalación del entorno de análisis para facilitar la reproductibilidad.

01. Ingesta

Se analizaron los Indicadores de salud relacionados con el consumo de alcohol ([U.S. Chronic Disease Indicators \(CDI\), 2023 Release](#)), y las tasas de mortalidad relacionadas con el consumo de alcohol ([U.S. Underlying Cause of Death, 2018-2022, Single Race](#)), para el año 2021.

02. Limpieza

Se identificó la información sucia, incorrecta, incompleta, imprecisa, irrelevante o incómoda y se reingestó, modificó, reemplazó o borró dicha información de acuerdo con la necesidad.

03. Análisis exploratorio de datos

Se evaluó una visión general de los *data.frame*, se exploraron las variables categóricas (*Fisher* y Chi-cuadrado), las variables numéricas (*Estadísticos descriptivos*), las distribuciones (*test de sesgo* y *kurtosis*), la normalidad (*QQ-plots* y *Shapiro-Wilk*), se compararon grupos (*Boxplots*, *test no-paramétricos*), se exploraron correlaciones, se realizó una modelización exploratoria (*Modelos lineales y no-lineales*) y se exploraron NAs y outliers.

04. Transformaciones de datos

Se crearon dos subconjuntos de datos, con o sin outliers, para valorar el impacto de estos en la clasificación, y se estandarizaron los datos numéricos para reducir el impacto de la diferencia de magnitud de las distintas variables sobre las técnicas de agrupación.

05. Análisis

Se generaron varios análisis cluster no jerárquicos con k -means, para investigar patrones de agrupación en los indicadores relacionados con el consumo de alcohol en EE. UU. durante el año 2021, por sexo, y se evaluó la adecuación, calidad interna y validez externa, así como las características definitorias de cada uno de los clusters, en cada uno de los modelos generados.

Interpretación de resultados

Se identificaron 2 modelos cluster de interés:

- Un modelo con $k=2$ clusters, para datos recortados, que sugiere la correlación positiva entre el sexo (masculino) y valores altos en los indicadores relacionados con el consumo de alcohol.
- Un modelo con $k=3$ clusters, para datos recortados, que no se relaciona claramente con ninguna variable del conjunto de datos, y que sugiere la existencia de tres grupos de estados en los que, además del sexo, existe una cierta influencia del número de grandes bebedores en un estado para el valor el resto de los indicadores relacionados con el consumo de alcohol.

Conclusiones y recomendaciones

Los valores de los indicadores relacionados con el consumo de alcohol presentan características comunes que los hacen candidatos a los estudios de agrupación. Las intervenciones de salud pública destinadas a reducir el daño derivado del consumo de alcohol deben contemplar acciones específicas dirigidas a cada uno de los sexos. Existen factores no identificados en el dataset que podrían estar condicionando los valores de los indicadores relacionados con el consumo del alcohol. El presente análisis cluster podría utilizarse para herramienta para la estratificación del muestreo de futuros estudios destinados a evaluar las características diferenciales entre los grupos que no se han podido identificar.

Palabras clave

Determinantes de la salud, mortalidad relacionada con el alcohol, análisis cluster no jerárquico, k -means, EE. UU.

Agradecimientos

Este trabajo se ha podido realizar gracias al trabajo, apoyo, consejo o recursos de varias personas o instituciones a las que el autor quiere mostrar su gratitud.

- Mis profesores, Rosana Ferrero y Juan Luis López, por su profesionalidad, amabilidad y cercanía a lo largo de todo el curso
- Máxima Formación, por la calidad de los contenidos del Máster, y la seriedad en la gestión de este
- El *Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health*, por la publicación en abierto de los datos utilizados en este trabajo
- Mi familia, por el apoyo y comprensión demostrados durante todo el desarrollo del máster

Índice

RESUMEN.....	2
<i>Justificación del análisis</i>	2
<i>Metodología utilizada</i>	2
00. Configuración	2
01. Ingesta	2
02. Limpieza	2
03. Análisis exploratorio de datos.....	2
04. Transformaciones de datos	2
05. Análisis	3
<i>Interpretación de resultados</i>	3
<i>Conclusiones y recomendaciones</i>	3
<i>Palabras clave</i>	3
AGRADECIMIENTOS.....	4
ÍNDICE	5
INTRODUCCIÓN	9
A - <i>Justificación de la necesidad para la organización</i>	9
A1 - Alineación del proyecto de investigación con los objetivos estratégicos de la organización	10
B - <i>Descripción del estudio</i>	11
B1 - Objetivo del estudio y preguntas de investigación.....	11
B2 - Metodología de análisis de datos	11
B3 - Datos utilizados en el análisis	12
C - <i>Gestión del proyecto de investigación</i>	12
C1 - Gobernanza de datos.....	12
C2 - Promotores del estudio.....	12
C3 - Destinatarios del estudio	12
C4 - Tecnología utilizada para el análisis	12
MATERIAL Y MÉTODOS	13
<i>Metodología utilizada</i>	14
01 - Ingesta	14
02 - Limpieza	14
03 - Análisis exploratorio de datos.....	14
04 - Transformaciones de datos	15
05 - Análisis	15
RESULTADOS	16
Subproceso 00 - <i>Configuración del equipo</i>	16
Descripción del subproceso	16
Acciones del subproceso	16
00a. Definir una configuración de R y RStudio que garantice la reproductibilidad de los resultados	16
00b. Instalar los paquetes de R necesarios para el análisis	16
Salidas del subproceso.....	16
Subproceso 01 - <i>Ingesta</i>	17
Descripción del subproceso	17
Acciones del subproceso	17
Salidas del subproceso.....	18
Subproceso 02 - <i>Limpieza</i>	19
Descripción del subproceso	19
Acciones del subproceso	19
Salidas del subproceso.....	19
Subproceso 03 - <i>Análisis exploratorio de datos (EDA)</i>	20

Descripción del subprocesso	20
Salidas del subprocesso.....	20
Subproceso 04 - Transformación	21
Descripción del subprocesso	21
Acciones del subprocesso	21
4a - Tratamiento de valores faltantes.....	21
4b - Tratamiento de valores atípicos (outliers)	21
4c - Estandarización de valores numéricos.....	21
Salidas del subprocesso.....	21
Subproceso 05f - Análisis	23
Descripción del análisis	23
Acciones incluidas en el análisis	23
Resultados - Análisis cluster.....	24
Descripción del subprocesso	24
Acciones del subprocesso.....	24
05fa - Selección de los datos adecuados para el análisis cluster	24
05fb - Estandarización de valores numéricos.....	24
05fc - Cálculo de la distancia entre observaciones	24
05fd - Análisis de tendencia de agrupación.....	26
05fe - Elección del método y la vinculación de grupos	27
05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación.....	28
05fg - Representación e interpretación de los resultados.....	30
05fh - Evaluación de la importancia de las variables	31
05fi - Visualización de las agrupaciones cluster	34
05fj - Validación de la agrupación.....	37
05fk - Resumen de resultados obtenidos	42
DISCUSIÓN.....	44
<i>Limitaciones del estudio.....</i>	44
<i>Discusión de los resultados obtenidos del estudio</i>	44
CONCLUSIONES	46
BIBLIOGRAFÍA.....	47
ANEXO 00 – PAQUETES DE R UTILIZADOS EN EL ANÁLISIS	51
ANEXO 1 – SUBPROCESO DE INGESTA – DETALLE DE ACCIONES DEL PROCESO.....	53
1 - <i>Indicadores de salud relacionados con el consumo de alcohol.....</i>	53
2 - <i>Tasas de mortalidad relacionadas con el consumo de alcohol.....</i>	53
3 - <i>Códigos FIPS de los estados de EE. UU.....</i>	55
01b - Definir el método y la configuración de la ingesta.....	55
01c - Crear los data.frame de datos crudos	55
01d - Validar la fase de ingesta	56
ANEXO 02 – SUBPROCESO DE LIMPIEZA - EVALUACIÓN DE LOS DATASETS CRUDOS Y ACCIONES DE LIMPIEZA EJECUTADAS	57
<i>02a - Identificación de la información sucia, incorrecta, irrelevante, incompleta, imprecisa o incómoda</i>	57
Validación de rawCdiAlcohol	57
Validación de rawUnderlyingCauseOfDeathAlcohol	59
Validación de rawFipsCodes	60
<i>02b - Reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo a la necesidad.....</i>	61
ANEXO 03 – ACCIONES DEL SUBPROCESO DE ANÁLISIS EXPLORATORIO DE DATOS (EDA)	63
<i>03a - Visión general del data.frame: summarytools::dfSummary().....</i>	63
<i>03b - Explorar variables categóricas: SmartEDA::ExpCatViz()</i>	63

<i>03c - Explorar variables numéricas (Estadística descriptiva): SmartEDA::ExpNumStat()</i>	63
<i>03d - Explorar distribuciones (skewness and kurtosis tests)</i>	64
<i>03e - Explorar normalidad (QQ-plots and Shapiro-Wilk)</i>	66
Objeto data	66
Test gráfico - QQ plot (DataExplorer::plot_qq()).....	66
Test de hipótesis (Shapiro-Wilk)	68
Objeto data_gender	68
Test gráfico - QQ plot (DataExplorer::plot_qq()).....	68
Test de hipótesis (Shapiro-Wilk)	70
Objeto data_overall	70
Test gráfico - QQ plot (DataExplorer::plot_qq()).....	70
Test de hipótesis (Shapiro-Wilk)	71
<i>03f - Comparar grupos (Boxplots, non-parametric tests)</i>	71
Objeto data	71
03fa - Valoración gráfica: DataExplorer::plot_boxplot()	71
03fb - Test de hipótesis: ggstatsplot::ggbetweenstats()	72
Objeto data_gender	77
03fa - Valoración gráfica: DataExplorer::plot_boxplot()	77
03fb - Test de hipótesis: ggstatsplot::ggbetweenstats()	77
Objeto data_overall	82
03fa - Valoración gráfica: DataExplorer::plot_boxplot()	82
03fb - Test de hipótesis: ggstatsplot::ggbetweenstats()	82
<i>03g - Explorar correlaciones</i>	82
Objeto data	82
03ga - Correlation matrix	82
03gb - Correlograma (visualización de la correlación)	84
03gc - Test de hipótesis ggstatsplot::ggcorrmat()	85
Objeto data_gender	86
03ga - Correlation matrix	86
03gb - Correlograma (visualización de la correlación)	87
03gc - Test de hipótesis ggstatsplot::ggcorrmat()	88
Objeto data_overall	89
03ga - Correlation matrix	89
03gb - Correlograma (visualización de la correlación)	90
03gc - Test de hipótesis ggstatsplot::ggcorrmat()	91
<i>03h - Explorar modelos de datos para las correlaciones estadísticamente significativas</i>	92
Objeto data	92
Population y AgeAdjustedDeathRate	92
Deaths y Population	93
Deaths y PercentageOfTotalDeaths	93
Population y PercentageOfTotalDeaths	94
HeavyDrinkingAdults y AgeAdjustedDeathRate	95
BingeDrinkingPrevalenceAdults y HeavyDrinkingAdults	96
BingeDrinkingPrevalenceAdults y BingeDrinkingFrecuencyAdults	97
BingeDrinkingPrevalenceAdults y BingeDrinkingIntensityAdults	98
BingeDrinkingIntensityAdults y BingeDrinkingFrecuencyAdults	99
Objeto data_gender	99
Population y Deaths	100
Deaths y PercentageOfTotalDeaths	100
Deaths y BingeDrinkingIntensityAdults	101
Deaths y BingeDrinkingPrevalenceAdults	102
Population y AgeAdjustedDeathRate	103
Population y PercentageOfTotalDeaths	103

Population y BingeDrinkingFrecuencyAdults	104
AgeAdjustedDeathRate y HeavyDrinkingAdults	104
AgeAdjustedDeathRate y BingeDrinkingFrecuencyAdults.....	105
AgeAdjustedDeathRate y BingeDrinkingIntensityAdults.....	106
AgeAdjustedDeathRate y BingeDrinkingPrevalenceAdults.....	107
PercentageOfTotalDeaths y BingeDrinkingIntensityAdults.....	108
PercentageOfTotalDeaths y BingeDrinkingPrevalenceAdults.....	109
HeavyDrinkingAdults y BingeDrinkingIntensityAdults	110
HeavyDrinkingAdults y BingeDrinkingPrevalenceAdults.....	111
BingeDrinkingFrecuencyAdults y BingeDrinkingPrevalenceAdults.....	112
BingeDrinkingIntensityAdults y BingeDrinkingFrecuencyAdults	112
Objeto data_overall	113
Deaths y HeavyDrinkingAdults	113
Deaths y Population	114
Deaths y PercentageOfTotalDeaths	114
Population y AgeAdjustedDeathRate.....	115
Population y HeavyDrinkingAdults	115
Population y PercentageOfTotalDeaths	116
PercentageOfTotalDeaths y HeavyDrinkingAdults	117
HeavyDrinkingAdults y BingeDrinkingIntensityAdults	117
BingeDrinkingFrecuencyAdults y BingeDrinkingPrevalenceAdults.....	118
AgeAdjustedDeathRate y HeavyDrinkingAdults	118
03i - Análisis de valores faltantes (NA's) y outliers	119
Funciones de R utilizadas en el análisis de datos faltantes (NA's)	119
Objeto data	120
03ia - Análisis de valores faltantes NA's.....	120
03ib - Exploración de Outliers	123
Objeto data_gender	124
03ia - Análisis de valores faltantes NA's.....	124
03ib - Exploración de Outliers	127
Objeto data_overall	128
03ia - Análisis de valores faltantes NA's.....	128
03ib - Exploración de Outliers	130
ANEXO 4 - TRANSFORMACIÓN	132
Descripción del subproceso	132
Acciones del subproceso	132
4a - Tratamiento de valores faltantes.....	132
4b - Tratamiento de valores atípicos (outliers)	132
1- Objeto data_lab	132
2- Objeto data_overall	135
Salidas del subproceso	138
ANEXO 5 - CÓDIGO DE R UTILIZADO EN EL ÁNALISIS	139

Introducción

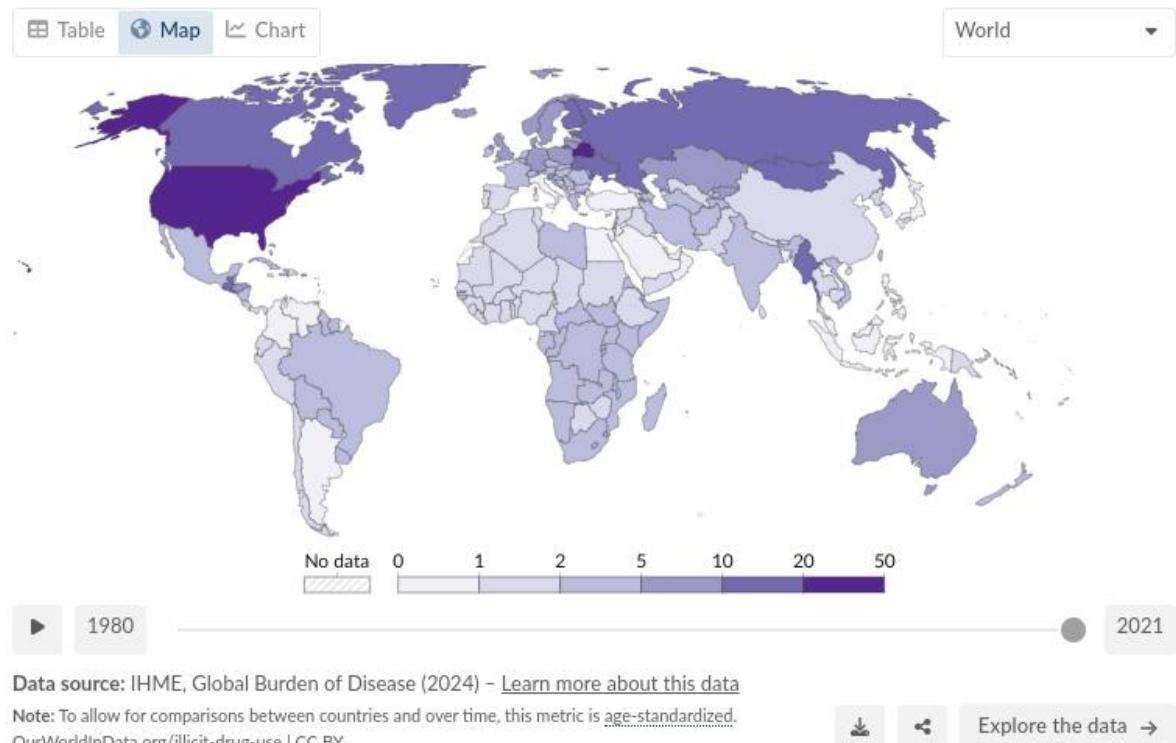
A - Justificación de la necesidad para la organización

- Se estima que el uso nocivo del alcohol causa cada año 2,5 millones de muertes a nivel mundial (Organización Mundial de la Salud 2010). Unas 178,000 personas mueren cada año en los Estados Unidos debido a una ingesta excesiva de alcohol (Esser, Sherk, Liu, and Naimi 2024), con una de las tasas más altas del mundo (Ortiz-Ospina and Roser 2016).
- En Europa, se estima que el alcohol es responsable de unas 195.000 muertes al año. Los países nórdicos y de Europa del este presentan tasas de mortalidad más altas que los del sur. Hasta 1 de cada 4 muertes en hombres jóvenes y 1 de cada 10 muertes en mujeres jóvenes se deben al consumo excesivo de alcohol.
- En España, el alcohol fue la sustancia psicoactiva más consumida en 2017 (Ministerio de Sanidad and Instituto Nacional de Estadística, n.d.). Se estima que durante el periodo 2010-2017 produjo anualmente unas 15.489 muertes (74% en hombres).

Alcohol and drug use disorders death rate, 2021

Our World
in Data

Estimated annual number of deaths from alcohol and drug use disorders per 100,000 people. These include only direct deaths from these disorders, meaning they do not include suicide deaths which can in some cases be connected or attributed to them.



- El concepto de bebida excesiva incluye la ingesta episódica de grandes cantidades de alcohol (botellones), grandes bebedores, y cualquier ingesta durante el embarazo o por personas menores de 21 años.

- La ingesta de alcohol contribuye al desarrollo de más de 200 problemas de salud y lesiones, así como a muerte prematura y es uno de los principales factores de riesgo asociado a enfermedades crónicas o no transmisibles ([International Agency for Research on Cancer and International Agency for Research on Cancer 1988](#)). Las muertes relacionadas con la ingesta excesiva de alcohol se deben tanto a patologías directamente relacionadas con el consumo (intoxicaciones etílicas, cirrosis), como a patologías parcialmente relacionadas con el consumo, como los accidentes de tráfico ([Esser, Sherk, Liu, and Naimi 2024](#)), o el cáncer ([Bagnardi et al. 2014](#)).
- Las políticas que disminuyen la accesibilidad al alcohol, o que lo encarecen, pueden prevenir el consumo excesivo de alcohol y sus consecuencias negativas. Los individuos, las organizaciones, las comunidades y los estados pueden apoyarse en estrategias probadas para reducir los daños relacionados con el alcohol, y mejorar la salud y la seguridad ([Esser, Sherk, Liu, Henley, et al. 2024](#)).

A1 - Alineación del proyecto de investigación con los objetivos estratégicos de la organización

- Este estudio está alineado con distintas estrategias mundiales, europeas y nacionales enfocadas a reducir los riesgos para la salud derivados del consumo de alcohol:

Ámbito	Estrategia	Descripción
Mundial	Estrategia Mundial para la Reducción del consumo de alcohol (Organización Mundial de la Salud 2010)	Estrategia enfocada en la concienciación, mejora de la evidencia científica, apoyo técnico a la formulación de política de salud pública, fortalecimiento de alianzas y mejora de sistemas de vigilancia y monitorización de los efectos negativos para la salud relacionados con el alcohol
Mundial	Estrategia SAFER: Un mundo sin daños relacionados con el alcohol (El Paquete Técnico SAFER. Un Mundo Libre de Los Daños Relacionados Con El Alcohol 2020)	Iniciativa enfocada en disminuir de la disponibilidad del alcohol, reducir las situaciones que fomentan el consumo, implementar acciones rápidas de screening, intervenciones breves o tratamiento, reducir el patrocinio y publicidad, y elevar el precio mediante impuestos y políticas de precios
Unión Europea	Estrategia RARHA: Reducing Alcohol Related Harm	Estrategia enfocada en crear mejor evidencia relacionada con el consumo del alcohol, e implementar herramientas de monitorización

		de los efectos negativos en la salud relacionados con el alcohol
España	<u>Estrategia de Promoción de la Salud y Prevención en el Sistema Nacional de Salud</u>	La Estrategia de Promoción de la Salud y Prevención en el SNS propone el desarrollo progresivo de intervenciones dirigidas a ganar salud y a prevenir las enfermedades, las lesiones y la discapacidad. Entre ellas, existe una línea dedicada al consumo de alcohol

B - Descripción del estudio

B1 - Objetivo del estudio y preguntas de investigación

- EL objetivo del estudio fue mejorar el conocimiento de los determinantes de salud que se relacionan con un consumo excesivo de alcohol en los ciudadanos del país con mayor letalidad por esta causa (Estados Unidos), para posibilitar el diseño de políticas de salud pública más efectivas.
- El estudio se encaminó a responder a la siguiente pregunta
- A - ¿Pueden clasificarse los pacientes en varios grupos similares entre sí, en función de los resultados de los indicadores relacionados con el consumo de alcohol?

B2 - Metodología de análisis de datos

- Para responder a las preguntas de investigación se utilizaron las siguientes técnicas de análisis de datos:
- Análisis exploratorio de los datos
- Análisis cluster para identificar patrones de agrupación

B3 - Datos utilizados en el análisis

Datos	Fuente	Justificación
1. Indicadores de salud relacionados con el consumo de alcohol	U.S. Chronic Disease Indicators (CDI), 2023 Release	Datos necesarios para hacer el análisis de agrupación (variables predictoras)
2. Tasas de mortalidad relacionadas con el consumo de alcohol	U.S. Underlying Cause of Death, 2018-2022, Single Race	Datos de la variable respuesta a predecir en el análisis de regresión
3. Códigos FIPS de los estados de EE. UU.	Dataset with FIPS codes for US states and counties	Tabla maestra con los códigos FIPS identificativos de los estados y condados de EE. UU.

C - Gestión del proyecto de investigación

C1 - Gobernanza de datos

- Se utilizaron fuentes de datos abiertas con datos agregados para el análisis, por lo que no fue necesario gestionar aspectos de privacidad de datos, acceso, seguridad o propiedad intelectual.

C2 - Promotores del estudio

- El estudio fue desarrollado como ejercicio para el Máster de Estadística aplicada con R, (Máxima Formación y Universidad de Nebrija).

C3 - Destinatarios del estudio

- Claustro docente del Máster.

C4 - Tecnología utilizada para el análisis

- Los datos se obtuvieron de fuentes de datos abiertas, o de datasets incluidos en paquetes de R
- Todas las fases del análisis (ingesta, limpieza, análisis exploratorio, transformación, análisis y comunicación) se realizaron con el software R

Material y métodos

El análisis se realizó siguiendo la siguiente secuencia de subprocessos:

ID	Denominación	Finalidad
00	<u>Configuración</u>	<ul style="list-style-type: none"> · 00a - Definición de la configuración de R y RStudio que garantice la reproductibilidad de los resultados · 00b - Instalación los paquetes de R necesarios para el análisis
01	<u>Ingesta</u>	<ul style="list-style-type: none"> · 01a - Identificación de los datos necesarios para el análisis y sus fuentes, si existen · 01b - Definición del método y la configuración de la ingesta (codificación de caracteres, valores faltantes, datos numéricos, categóricos, lógicos, fechas...) · 01c - Creación de los <i>data.frame</i> de datos crudos · 01d - Validación de la fase de ingestra
02	<u>Limpieza</u>	<ul style="list-style-type: none"> · 02a - Identificación de la información sucia, incorrecta, incompleta, imprecisa, irrelevante o incómoda · 02b - Reingesta, modificación, reemplazo o borrado de esta información no deseada, de acuerdo con la necesidad
03	<u>Análisis exploratorio de datos</u>	<ul style="list-style-type: none"> · 03a - Visión general del <i>data.frame</i> · 03b - Exploración variables categóricas (χ^2 y <i>test de Fisher</i>) · 03c - Exploración variables numéricas (<i>Estadística descriptiva</i>) · 03d - Exploración distribuciones (<i>test de sesgo</i> y <i>curtosis</i>) · 03e - Exploración normalidad (<i>QQ-plots</i> y <i>Shapiro-Wilk</i>) · 03f - Comparación de grupos (<i>Boxplots</i>, <i>test no-paramétricos</i>) · 03g - Exploración de correlaciones · 03h - Exploración de datos (<i>Modelos lineales</i> y <i>no lineales</i>) · 03i - Exploración NA's y outliers
04	<u>Transformación de datos</u>	<ul style="list-style-type: none"> · 4a - Tratamiento de valores faltantes · 4b - Tratamiento de valores atípicos (<i>outliers</i>)
05	<u>Análisis</u>	<ul style="list-style-type: none"> · 05fa - Selección de los datos adecuados para el análisis cluster · 05fb - Estandarización de valores numéricos · 05fc - Cálculo de la distancia entre observaciones · 05fd - Análisis de tendencia de agrupación

	<ul style="list-style-type: none"> · 05fe - Elección del método y la vinculación de grupos · 05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación. · 05fg - Representación e interpretación de los resultados. · 05fh - Evaluación de la importancia de las variables · 05fi - Visualización de las agrupaciones cluster · 05fj - Validación de la agrupación · 05fk - Interpretación
--	--

Metodología utilizada

01 - Ingesta

Se utilizaron como origen de datos los siguientes conjuntos de datos:

- Indicadores de salud relacionados con el consumo de alcohol ([U.S. Chronic Disease Indicators \(CDI\), 2023 Release](#))
- Tasas de mortalidad relacionadas con el consumo de alcohol ([U.S. Underlying Cause of Death, 2018-2022, Single Race](#))
- Códigos FIPS de los estados de EE. UU. ([Dataset with FIPS codes for US states and counties](#))

02 - Limpieza

Para la limpieza de los datos crudos se siguió una metodología sistemática para identificar la información sucia, incorrecta, incompleta, imprecisa, irrelevante o incómoda, en una primera fase, para posteriormente reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo a la necesidad.

03 - Análisis exploratorio de datos

El análisis exploratorio siguió la metodología propuesta por ([Data Science 2021](#)).

Incluye las siguientes actividades:

- 03a - Visión general del *data.frame*
- 03b - Explorar variables categóricas (χ^2 y test de Fisher)
- 03c - Explorar variables numéricas (*Estadística descriptiva*)
- 03d - Explorar distribuciones (*Test de sesgo y curtosis*)
- 03e - Explorar normalidad (*QQ-plots* y *test de Shapiro-Wilk*)
- 03f - Comparar grupos (*Boxplots, test no paramétricos*)
- 03g - Explorar correlaciones
- 03h - Explorar datos (*Modelización exploratoria sobre las correlaciones*)
- 03i - Explorar NAs y outliers

04 - Transformaciones de datos

Se realizaron las siguientes tareas de transformación:

- 4a - Tratamiento de valores faltantes
- 4b - Tratamiento de valores atípicos (*outliers*)
- 4c - Estandarización de las variables numéricas

05 - Análisis

Se llevó a cabo un conjunto de análisis cluster (5 en total), siguiendo la siguiente metodología

- 05fa - Selección de los datos adecuados para el análisis cluster
- 05fb - Estandarización de valores numéricos
- 05fc - Cálculo de la distancia entre observaciones
- 05fd - Análisis de tendencia de agrupación
- 05fe - Elección del método y la vinculación de grupos
- 05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación.
- 05fg - Representación e interpretación de los resultados.
- 05fh - Evaluación de la importancia de las variables
- 05fi - Visualización de las agrupaciones cluster
- 05fj - Validación de la agrupación
- 05fk - Interpretación

Durante la fase de análisis exploratorio se evidenció una marcada diferencia en la mortalidad relacionada con alcohol entre ambos sexos, por lo que se analizó el efecto de la variable *Sex*.

Además, se observaron problemas de valores atípicos en los distintos conjuntos de datos considerados para el análisis. Dado el potencial interés para nuestro análisis de estas observaciones atípicas, los análisis se realizaron sobre datasets completos (con *outliers*) y recortados (sin *outliers*).

Resultados

Subproceso 00 - Configuración del equipo

Descripción del subprocesso

Subproceso destinado a establecer una configuración de R y RStudio que garantice la reproductibilidad de los resultados, e instalar los paquetes de R necesarios para el análisis.

Incluye las siguientes acciones:

- 00a - Definir una configuración de R y RStudio que garantice la reproductibilidad de los resultados
- 00b - Instalar los paquetes de R necesarios para el análisis

Acciones del subprocesso

00a. Definir una configuración de R y RStudio que garantice la reproductibilidad de los resultados

Para facilitar la reproductibilidad del análisis y la coherencia de los resultados obtenidos en distintos equipos, se han incorporado las siguientes opciones de configuración:

- Establecer una semilla aleatoria para el análisis:
- Impedir que los números grandes se muestren con notación científica

00b. Instalar los paquetes de R necesarios para el análisis

Los paquetes de R necesarios para este análisis están recogidos en el objeto `paquetesNecesariosAnalisis`, y están disponibles en el Anexo 1.

Salidas del subprocesso

- El equipo queda correctamente configurado para reproducir los resultados del análisis
 - Configuración de R estandarizada
 - Paquetes de R necesarios instalados en el equipo
- Se documentan los paquetes necesarios para el análisis
 - Objeto `paquetesNecesariosAnalisis`

Subproceso 01 - Ingesta

Descripción del subproceso

Subproceso destinado a recopilar datos desde sus fuentes originales, y ubicarlos en un entorno en el que se pueda acceder a ellos, usarlos o analizarlos.

El subproceso incluye las siguientes acciones:

- 01a - Identificar los datos necesarios para el análisis y sus fuentes, si existen
- 01b - Definir el método y la configuración de la ingesta (ingesta de datos existentes o creación de datasets personalizados) (codificación de caracteres, valores faltantes, datos numéricos, categóricos, lógicos, fechas...)
- 01c - Crear los *data.frame* de datos crudos
- 01d - Validar la fase de ingesta

Acciones del subproceso

Para el análisis se utilizaron los siguientes datos:

Datos	Fuente	Justificación
1 - Indicadores de salud relacionados con el consumo de alcohol	U.S. Chronic Disease Indicators (CDI), 2023 Release	Datos necesarios para hacer el análisis de agrupación (variables predictoras)
2 - Tasas de mortalidad relacionadas con el consumo de alcohol	U.S. Underlying Cause of Death, 2018-2022, Single Race	Datos de la variable respuesta a predecir en el análisis de regresión
3 - Códigos FIPS de los estados de EE. UU.	Dataset with FIPS codes for US states and counties	Tabla maestra con los códigos FIPS identificativos de los estados y condados de EE. UU.

En el Anexo 1 se puede consultar el detalle del subproceso de ingesta.

Salidas del subprocesso

Objeto	Descripción del <i>data.frame</i>	Filas	Columnas
rawCdiAlcohol	Indicadores de enfermedades crónicas (CDI) del área de interés ‘Alcohol’, por estado y año (2010-2022)	66091	34
rawUnderlyingCauseOfDeathAlcohol	Tasas de mortalidad por Alcohol, por sexo, estado y año (2018-2022)	817	18
rawFipsCodes	Maestra de códigos de estados y condados de EE. UU.	3256	5

Subproceso 02 - Limpieza

Descripción del subprocesso

Subproceso destinado a identificar la información sucia, incorrecta, incompleta, imprecisa, irrelevante o incómoda, y a reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo con la necesidad

Acciones del subprocesso

El subprocesso incluye las siguientes acciones:

- 02a - Identificar la información sucia, incorrecta, incompleta, imprecisa, irrelevante o incómoda
- 02b - Reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo con la necesidad

En el anexo 3 puede consultarse el resultado de las dos acciones para los datasets evaluados.

Salidas del subprocesso

Objeto	Descripción del <i>data.frame</i>	Filas	Columnas
data	Indicadores de enfermedades crónicas (CDI) del área de interés ‘Alcohol’ y de mortalidad por alcohol, año 2021, por estado de EE. UU. y sexo	163	10
data_overall	Indicadores de enfermedades crónicas (CDI) del área de interés ‘Alcohol’ y de mortalidad por alcohol, año 2021, por estado de EE. UU. y sexo	153	9
data_gender	Indicadores de enfermedades crónicas (CDI) del área de interés ‘Alcohol’ y de mortalidad por alcohol, año 2021, por estado de EE. UU. y sexo	106	10

Subproceso 03 - Análisis exploratorio de datos (EDA)

Descripción del subprocesso

Proceso de investigación del conjunto de datos para descubrir patrones y anomalías, y establecer hipótesis basadas en la comprensión del dataset.

El proceso ejecuta las siguientes acciones ([Data Science 2021](#)):

- 03a - Visión general del *data.frame*
- 03b - Explorar variables categóricas (*Fisher* y *chi-cuadrado*)
- 03c - Explorar variables numéricas (Estadística descriptiva)
- 03d - Explorar distribuciones (*tests de sesgo* y *curtosis*)
- 03e - Explorar normalidad (*QQ-plots* y *Shapiro-Wilk*)
- 03f - Comparar grupos (*Boxplots*, *test no-paramétricos*)
- 03g - Explorar correlaciones
- 03h - Explorar datos (Modelos lineales y no lineales)
- 03i - Explorar NAs y *outliers*

Puede verse una descripción detallada del proceso en el anexo 4

Salidas del subprocesso

- Análisis de variables categóricas
- Análisis de variables numéricas
- Análisis de distribución de variables aleatorias
- Estudio de normalidad de las variables numéricas
- Comparación de los valores de las variables numéricas según niveles de las variables categóricas
- Estudio de correlación lineal entre variables numéricas
- Exploración de modelos de datos para correlaciones estadísticamente significativas
- Análisis de datos faltantes
- Análisis de datos extremos (*outliers*)

Los resultados detallados del análisis exploratorio pueden consultarse en el Anexo 3.

Subproceso 04 - Transformación

Descripción del subproceso

Subproceso destinado a convertir los datos crudos ya limpiados al formato o estructura que requiere el tipo de análisis que se va a realizar en nuestros datos.

Acciones del subproceso

Se realizaron las siguientes tareas de transformación:

- 4a - Tratamiento de valores faltantes
- 4b - Tratamiento de valores atípicos (*outliers*)
- 4c - Estandarización de valores numéricos

4a - Tratamiento de valores faltantes

Se creó un dataset de trabajo sin datos faltantes, uno para cada dataset de interés.

Tras la omisión de NA's, los dos datasets `data_lab` y `data_gender_lab` son idénticos, y sólo difieren en los atributos que se han ido creando durante el proceso de limpieza. Por tanto, podemos trabajar exclusivamente con `data_lab` (para datos por sexos) y `data_overall` (para datos globales):

4b - Tratamiento de valores atípicos (outliers)

Las observaciones con valores extremos para las variables estudiadas podrían ser muy interesantes para nuestro análisis, porque pueden contener información sobre los factores de riesgo más asociados a la mortalidad por alcohol.

4c - Estandarización de valores numéricos

Se estandarizaron los valores numéricos para evitar que las diferencias de magnitud de las distintas variables distorsionasen el resultado del análisis cluster.

Salidas del subproceso

Se crearon los siguientes objetos, diferenciados entre sí por la presencia o ausencia de tres características: datos estratificados por sexo, Inliers y Outliers:

Objeto	Datos por sexo	Inliers	Outliers
<code>data_lab</code>	Sí	Sí	Sí
<code>data_inliers_lab</code>	Sí	Sí	No
<code>data_outliers_lab</code>	Sí	No	Sí
<code>data_overall_lab</code>	No	Sí	Sí
<code>data_overall_inliers_lab</code>	No	Sí	No
<code>data_overall_outliers_lab</code>	No	No	Sí

Subproceso 05f - Análisis

Descripción del análisis

Metodología destinada a agrupar observaciones según su similitud, de modo que las observaciones de cada grupo tengan características similares.

Acciones incluidas en el análisis

Se llevaron a cabo un conjunto de análisis cluster (5 en total), siguiendo la siguiente metodología:

- 05fa - Selección de los datos adecuados para el análisis cluster
- 05fb - Estandarización de valores numéricos
- 05fc - Cálculo de la distancia entre observaciones
- 05fd - Análisis de tendencia de agrupación
- 05fe - Elección del método y la vinculación de grupos
- 05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación.
- 05fg - Representación e interpretación de los resultados.
- 05fh - Evaluación de la importancia de las variables
- 05fi - Visualización de las agrupaciones cluster
- 05fj - Validación de la agrupación
- 05fk - Resumen de los resultados obtenidos

Resultados - Análisis cluster

Descripción del subprocesso

Subproceso destinado a agrupar observaciones según su similitud, de modo que las observaciones de cada grupo tengan características similares.

Acciones del subprocesso

05fa - Selección de los datos adecuados para el análisis cluster

Durante la fase de análisis exploratorio se evidenció una marcada diferencia en la mortalidad relacionada con alcohol entre ambos sexos, por lo que se analizó el dataset `data_lab` para controlar el efecto de la variable `Sex`.

Además, se observaron problemas de valores atípicos en los distintos conjuntos de datos considerados para el análisis. Estas observaciones podrían ser muy interesantes para nuestro análisis, porque pueden contener información sobre los factores de riesgo más asociados a la mortalidad por alcohol.

- `data_lab`: con todos los datos (incluyendo *outliers*), y
- `data_inliers_lab`: con datos recortados (sin *outliers*).

05fb - Estandarización de valores numéricos

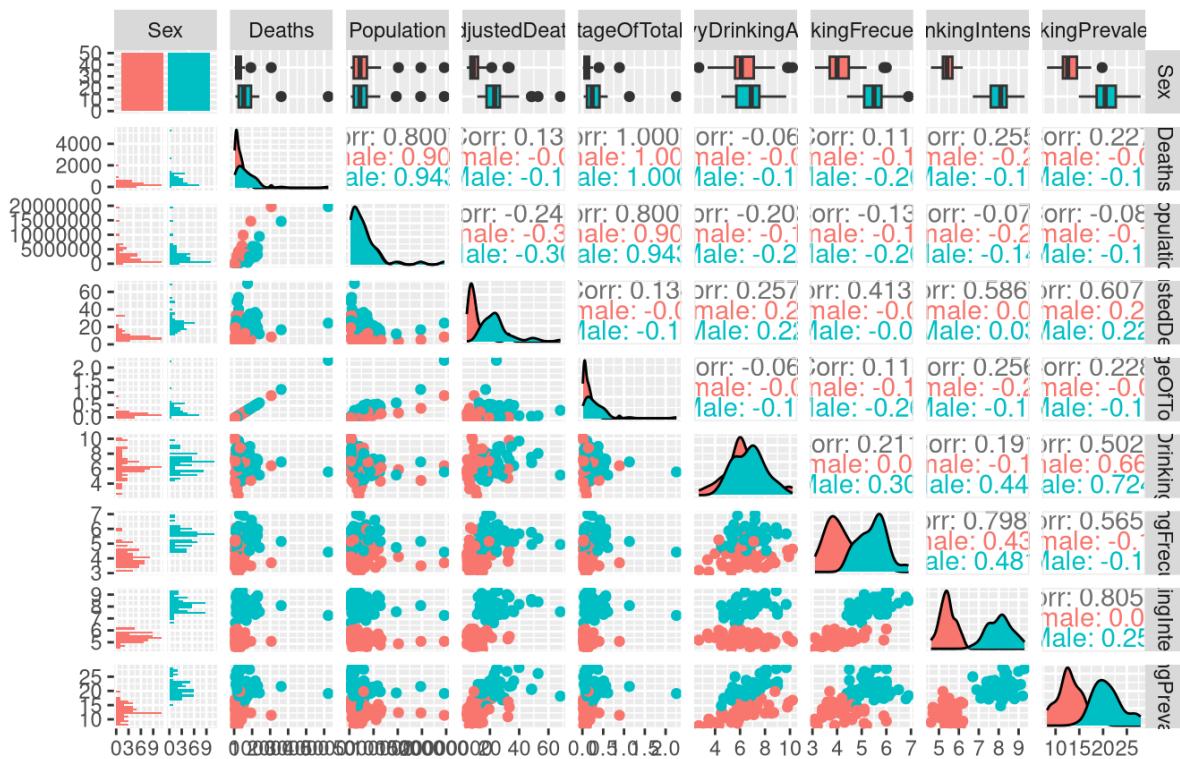
Para impedir que las diferencias de magnitud entre las variables numéricas alterasen la agrupación, se escalaron los valores de ambos datasets.

05fc - Cálculo de la distancia entre observaciones

Se utilizó la función `stat::dist()` con los parámetros por defecto (distancia euclídea):

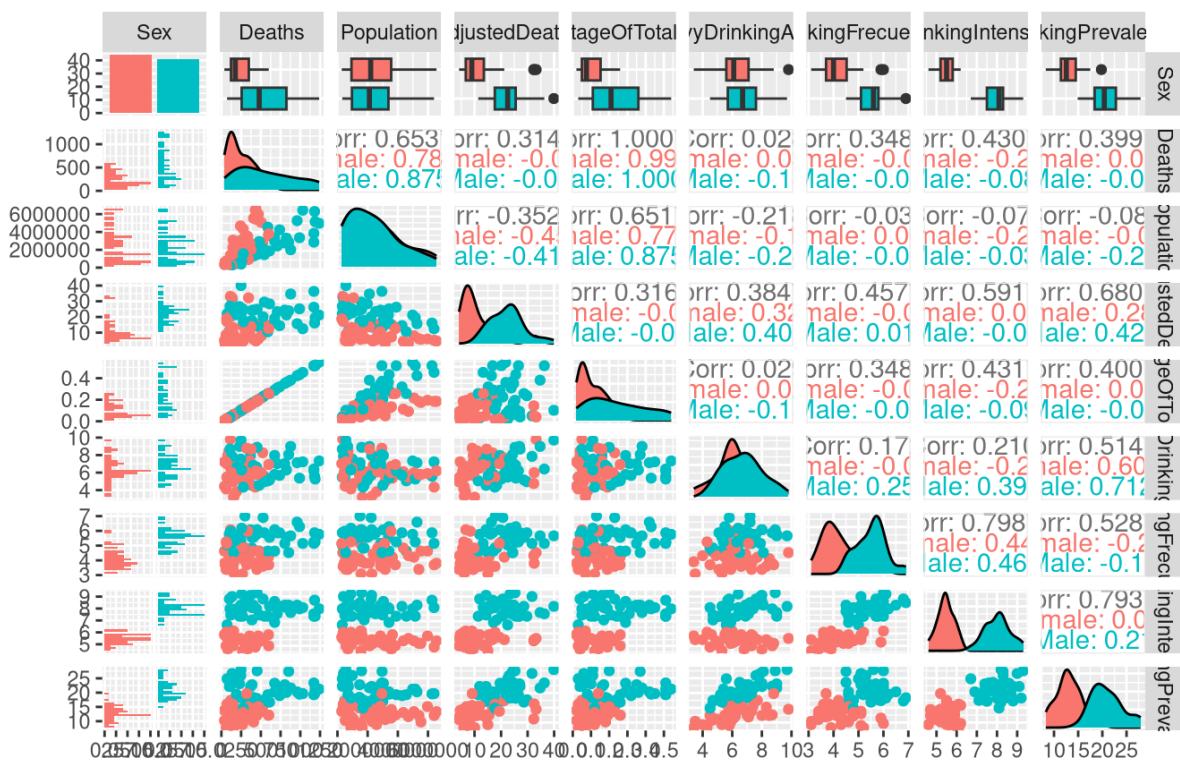
Objeto data_lab

Análisis de correlación



Objeto data_inliers_lab

Análisis de correlación



05fd - Análisis de tendencia de agrupación

Valoramos en primer lugar si es pertinente realizar un análisis de agrupación de los datos. Para ello:

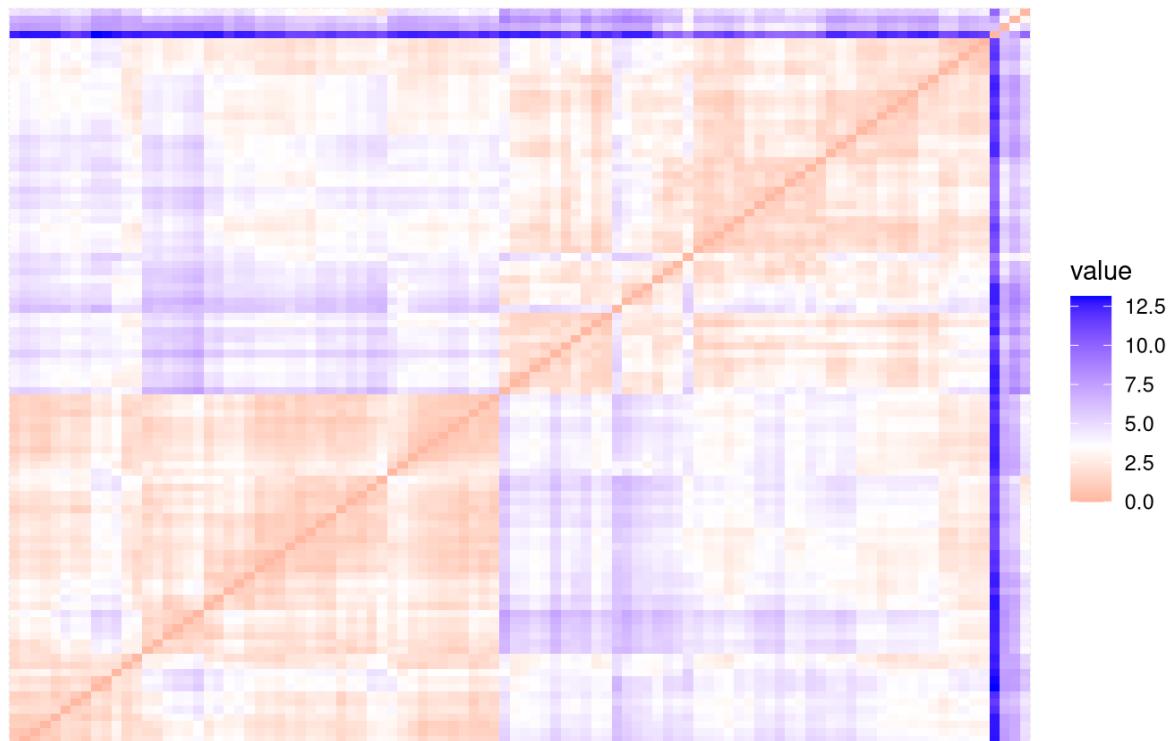
- Analizamos visualmente los clústeres de datos con el análisis visual de tendencia (VAT)
- Evaluamos la tendencia de agrupación con el estadístico de Hopkins.

Evaluación visual de tendencia (VAT)

Este mapa del calor reordena la matriz de tal manera que observaciones similares se localizan cerca. Visualmente, se observan entre tres y cinco grandes clusters, que son más evidentes cuando se eliminan los outliers.

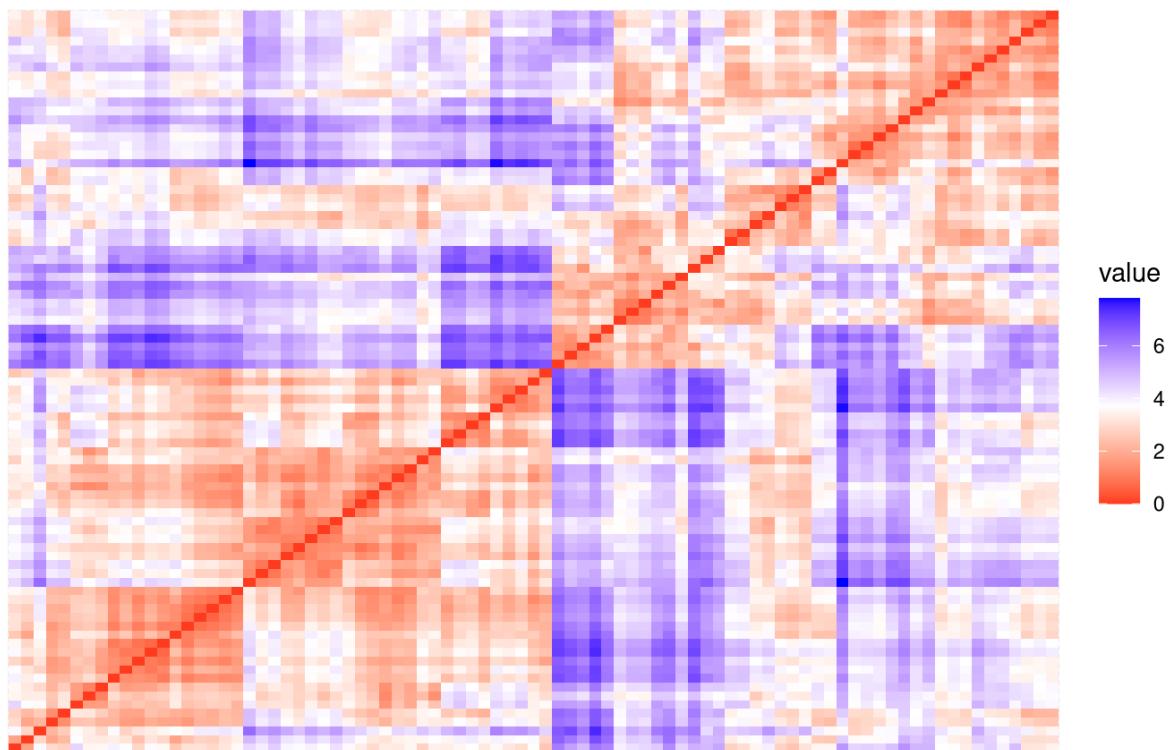
Evaluación visual de tendencia de agrupación (VAT)

Datos completos (`data_lab`)



Evaluación visual de tendencia de agrupación (VAT)

Datos sin outliers (`data_inliers_lab`)



Estadística de Hopkins

En ambos supuestos (datos totales y recortados), el valor es distinto de 0.5, por lo que suponemos que las distancias observadas entre el conjunto de datos aleatorio y el conjunto de datos real no se debe al azar, y, por tanto, existe tendencia de agrupación.

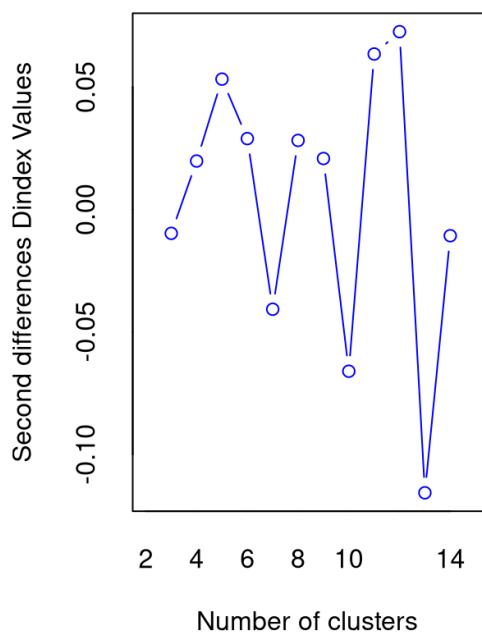
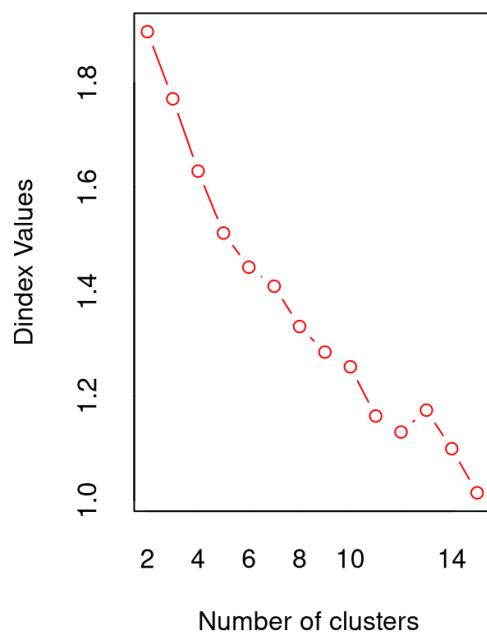
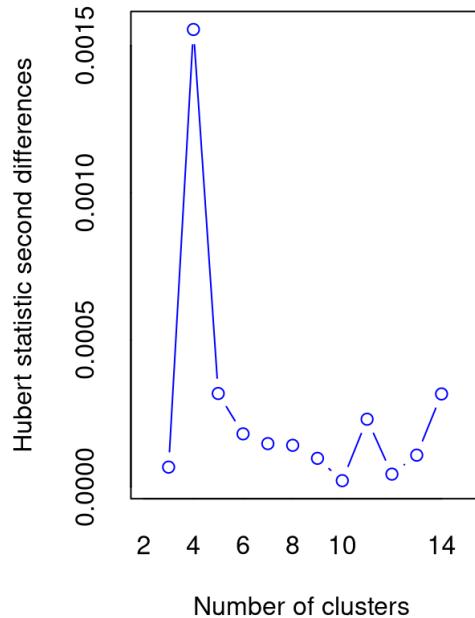
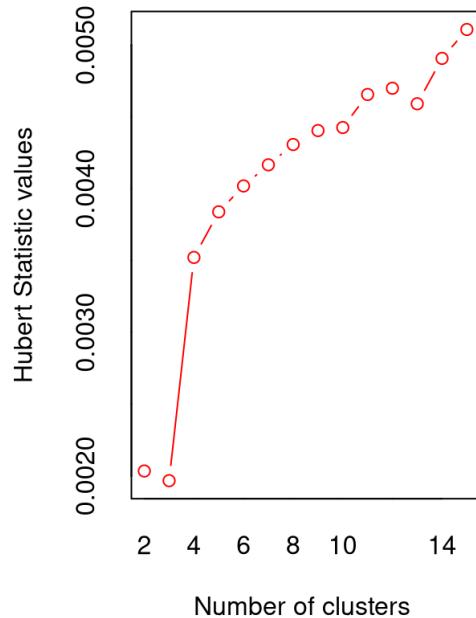
05fe - Elección del método y la vinculación de grupos

Se utilizó el método de agrupación por k-medias.

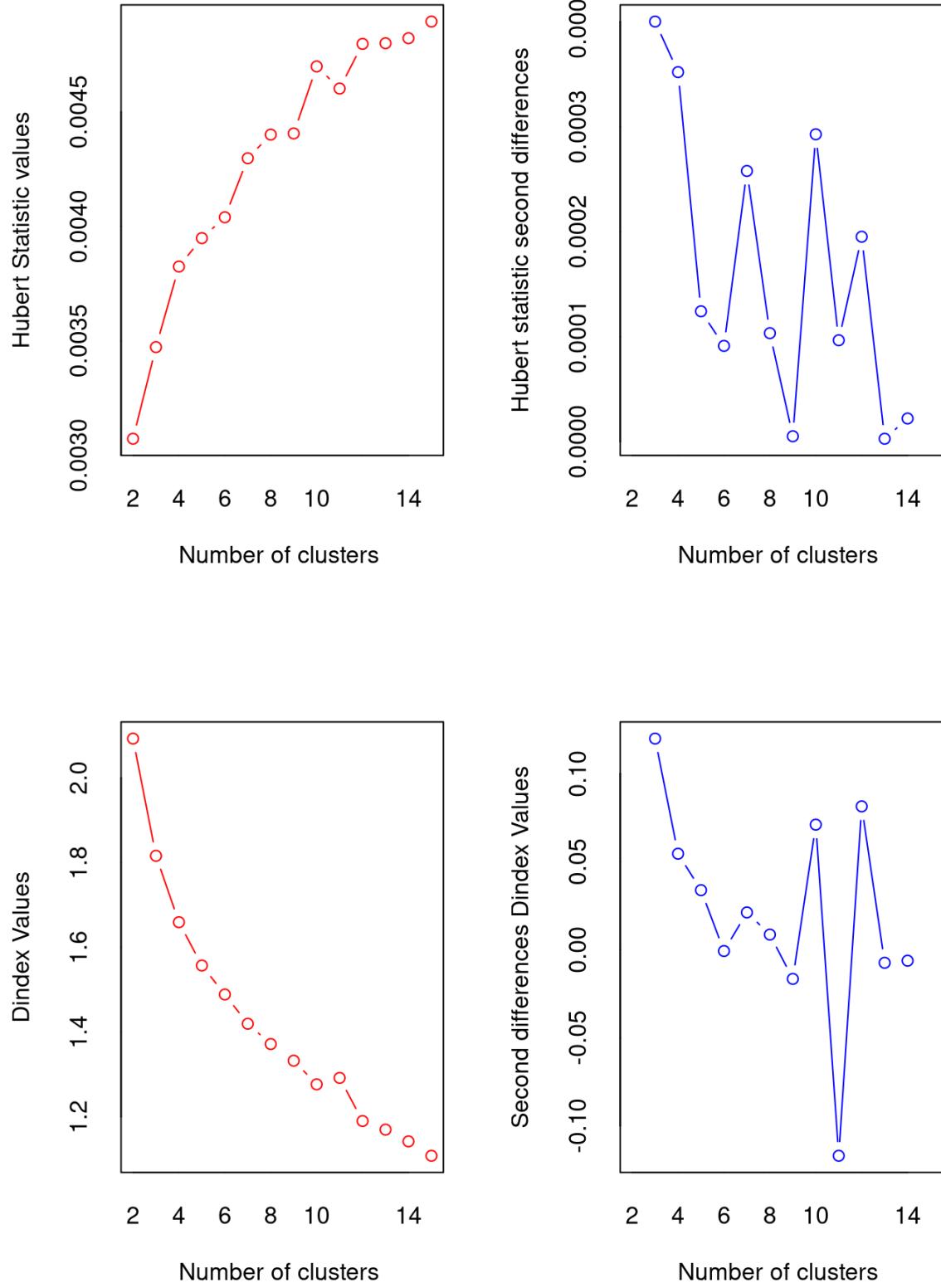
05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación.

Según criterios de calidad interna

En el dataset completo, la mayoría de los métodos sitúa el óptimo de clústeres entre 2 y 4.



En el dataset sin outliers, el número óptimo de clústeres está entre 2 y 3.



Según criterios de estabilidad

Dado que no podemos encontrar un nivel de clústeres óptimo en base a los resultados, se exploraron las opciones más repetidas:

- 2, 3 y 4 clústeres para `data_lab`, y
- 2 y 3 clústeres para `data_inliers_lab`.

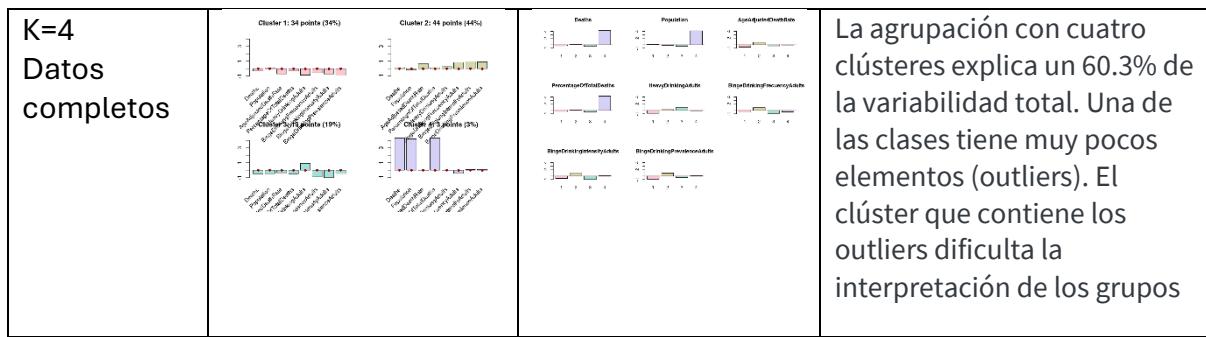
05fg - Representación e interpretación de los resultados.

Datos completos `data_lab`

Resultados de los modelos de agrupación para `data_lab`

- Los modelos explican un porcentaje de la variabilidad total observada entre el 34.2% y el 60.3%. El modelo con mejor explicación de los datos observados es el de $k=4$ grupos
- Todos los modelos explorados son algo difíciles de interpretar, porque los outliers tienden a agruparse en uno de los grupos del modelo, y dificultan la comprensión de la lógica de la agrupación.

Modelo	Comportamiento variables en cluster	Variables por cluster	Valoración
K=2 Datos completos			La agrupación con dos clústeres explica un 34.2% de la variabilidad total. Ambos clústeres tienen aproximadamente el mismo número de elementos. Se diferencian entre sí esencialmente por los valores del área de interés Alcohol1 del CDI: valores elevados frente a valores bajos.
K=3 Datos completos			La agrupación con tres clústeres explica un 53.9% de la variabilidad total. Los clústeres están muy desequilibrados, con uno de ellos con 3 elementos (los valores <i>outliers</i>). La presencia del grupo con los outliers (cluster 1) dificulta la interpretación visual de los otros dos grupos, tanto en el gráfico global como por variables. El resultado genera una agrupación muy desbalanceada.



Datos recortados data_inliers_lab

Resultados de los modelos de agrupación para data_inliers_lab

- Los modelos explican un porcentaje de la variabilidad total observada entre el **38.32%** y el **54%**. El modelo con mejor explicación de los datos observados es el de $k=3$ grupos
- Los modelos con datos recortados explican un menor porcentaje de la variabilidad que los de datos completos; los outliers capturan una parte considerable de la información, y deben estudiarse con detalle.

Modelo	Comportamiento variables en cluster	Variables por cluster	Valoración
K=2 Datos recortados			<p>La agrupación con dos clústeres explica un 38.32% de la variabilidad total. El primer clúster tiene un poco más de elementos que el segundo. Los dos clústeres se diferencian entre sí por los valores del área de interés Alcohol del CDI: valores elevados frente a valores bajos</p>
K=3 Datos recortados			<p>La agrupación con tres clústeres explica un 54% de la variabilidad total. El primer cluster tiene más elementos que los otros dos juntos.</p>

05fh - Evaluación de la importancia de las variables

Datos completos data_lab

Los resultados de la evaluación de la importancia de las variables para los modelos para datos completos fueron los siguientes:

- En los modelos de 2 y 3 clústeres para datos completos, las variables más importantes para establecer la agrupación fueron las relacionadas con las

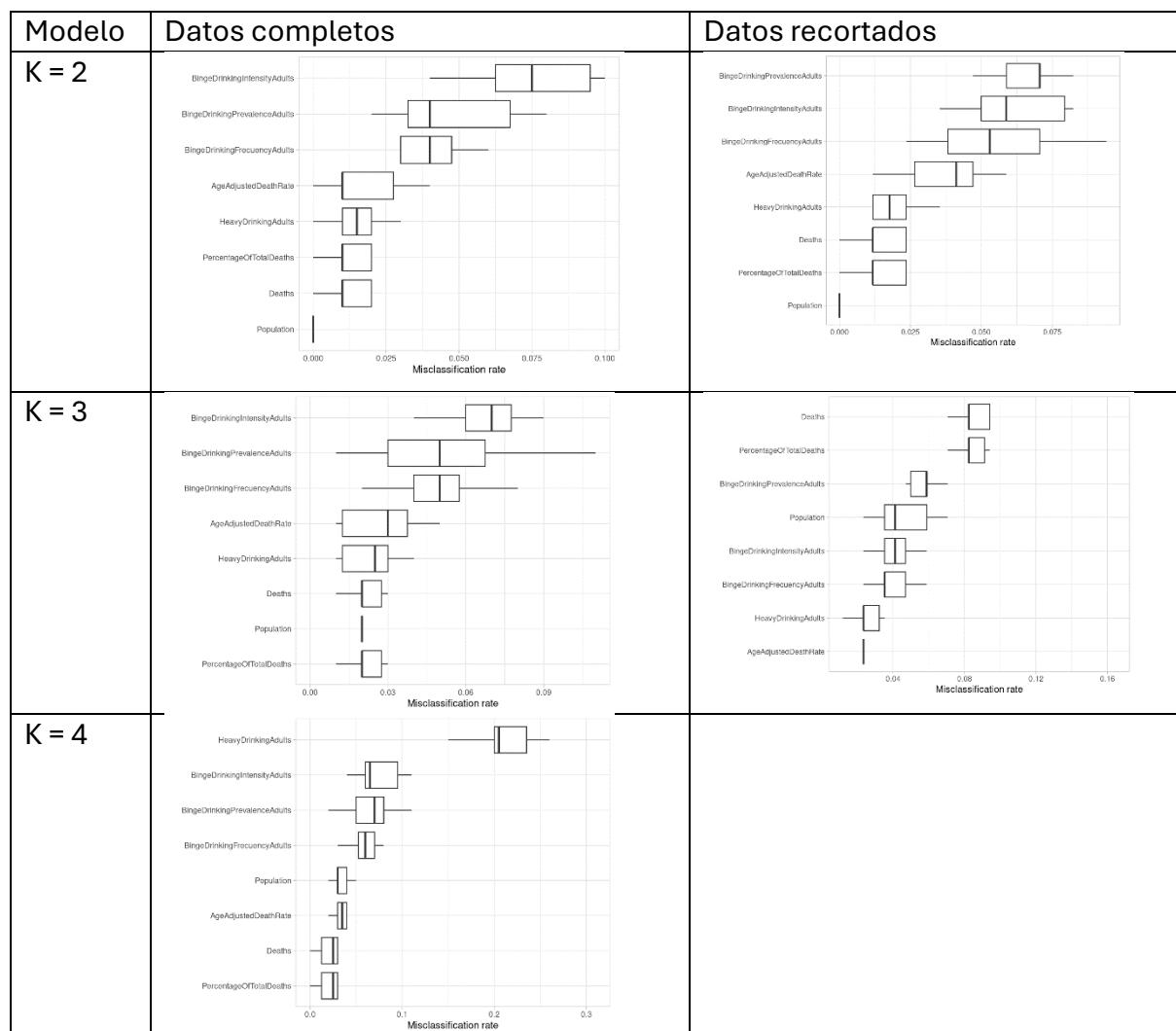
características de las borracheras (BingeDrinkingIntensityAdults, BingeDrinkingPrevalenceAdults, BingeDrinkingFrecuencyAdults)

- Para el modelo con 4 clústeres, la variable más importante con diferencia fue la de grandes bebedores (HeavyDrinkingAdults), seguida de las tres variables relacionadas con borracheras (BingeDrinkingIntensityAdults, BingeDrinkingPrevalenceAdults, BingeDrinkingFrecuencyAdults)

Datos recortados data_inliers_lab

Los resultados de la evaluación de la importancia de las variables para los modelos para datos recortados fueron los siguientes:

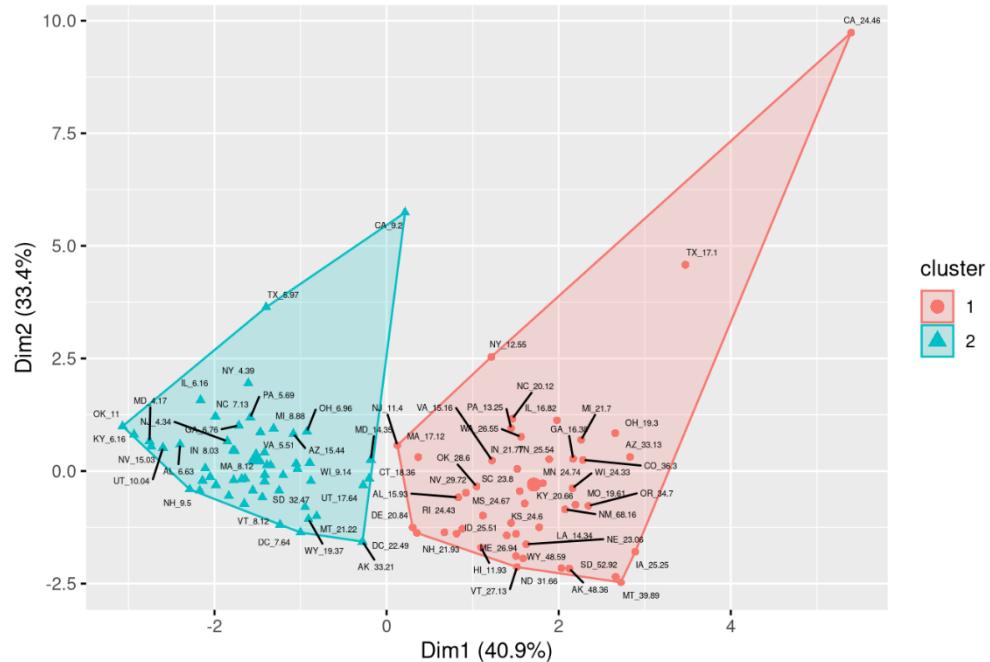
- En el modelo de 2 clústeres para datos recortados, las variables más importantes para establecer la agrupación fueron las relacionadas con las características de las borracheras (BingeDrinkingIntensityAdults, BingeDrinkingPrevalenceAdults, BingeDrinkingFrecuencyAdults)
- Para el modelo con 3 clústeres, las variables más importantes para establecer la agrupación fueron las relacionadas el número de muertes relacionadas con alcohol (Deaths y PercentageOfTotalDeaths), seguido de la prevalencia de borracheras (BingeDrinkingPrevalenceAdults).



05fi - Visualización de las agrupaciones cluster

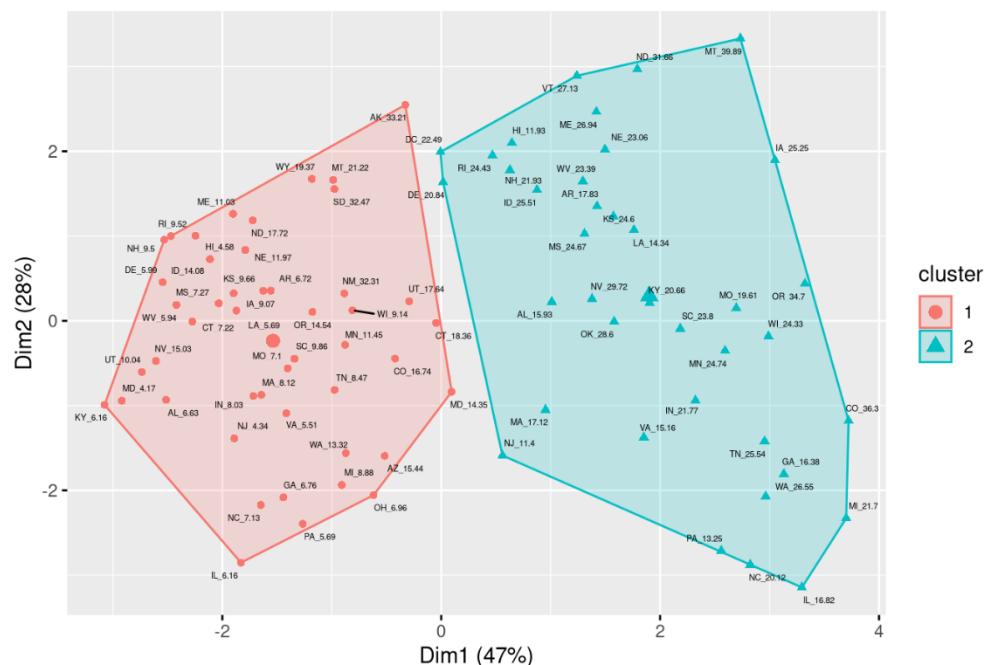
Modelos con K=2

k=2 grupos, datos completos



El modelo de 2 clústeres en los datos completos está fuertemente influenciado por los outliers, y crea unos clústeres con poco sentido.

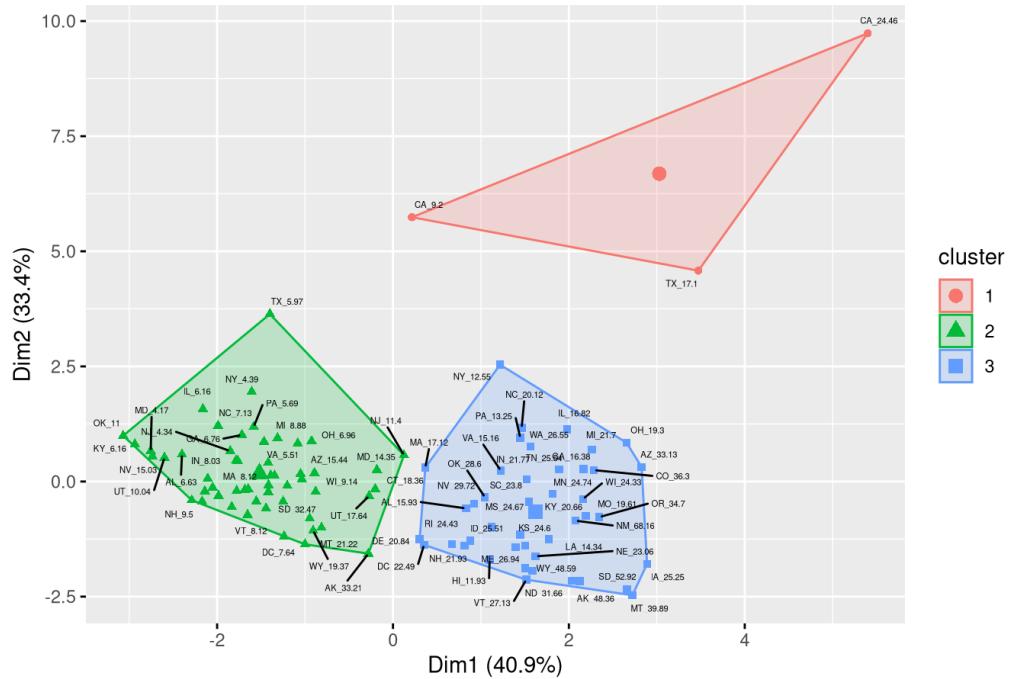
k=2 grupos, datos recortados (sin outliers)



Al eliminar los outliers, el modelo agrupa los datos en dos grandes bloques, sin solapamientos

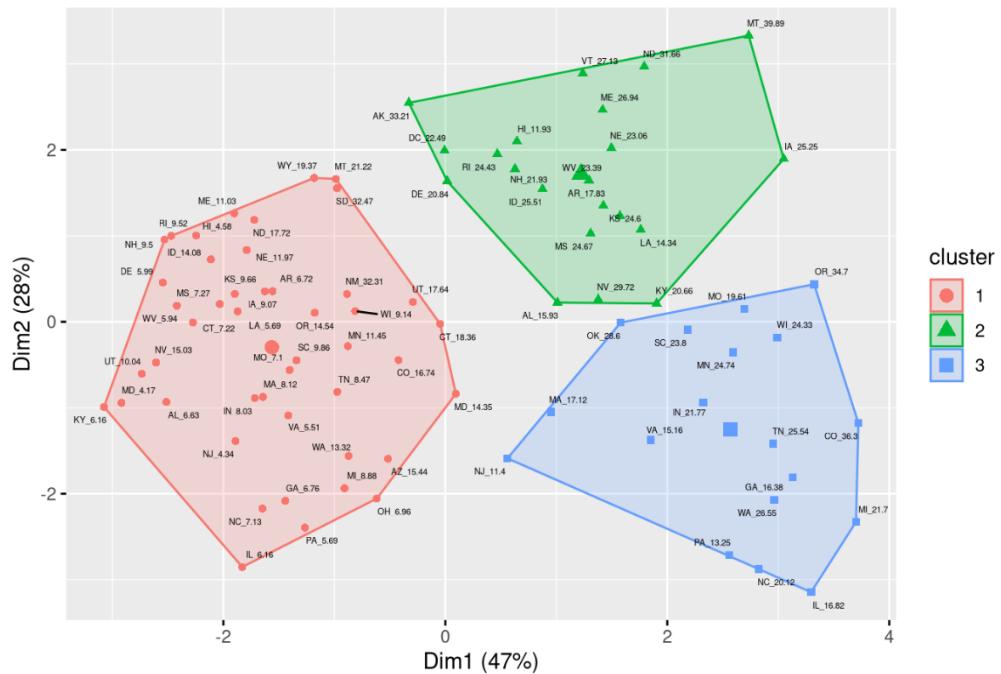
Modelos con K=3

k=3 grupos (datos completos)



El modelo de tres clústeres para los datos completos separa un grupo con los outliers, y otros dos grupos dentro del resto de los datos.

k=3 grupos, datos recortados (sin outliers)

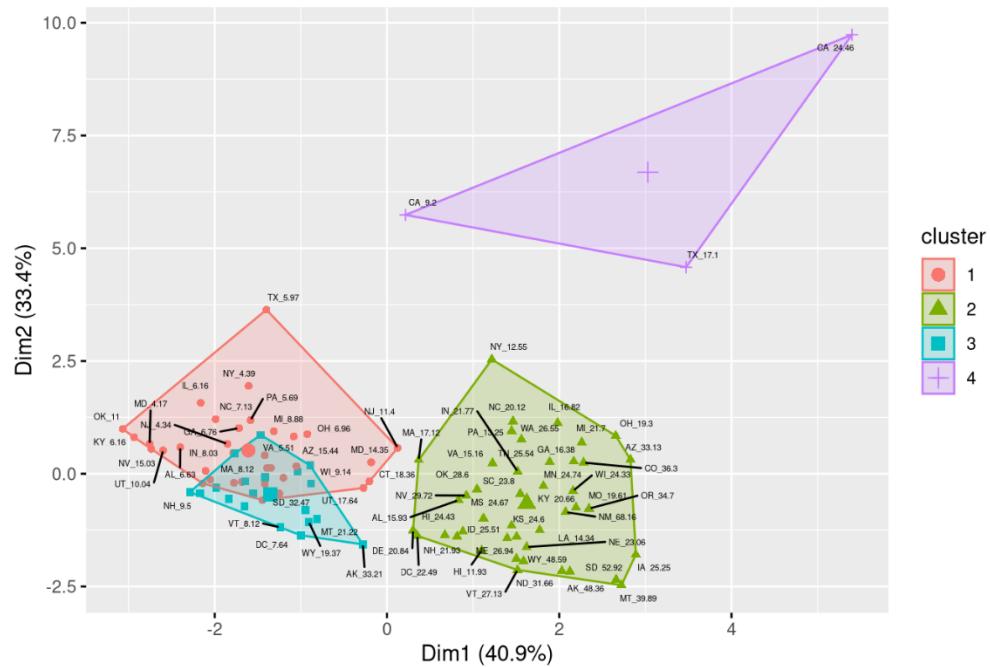


Al eliminar los outliers, el modelo de tres clústeres es capaz de separar tres grupos de datos con una cierta coherencia visual.

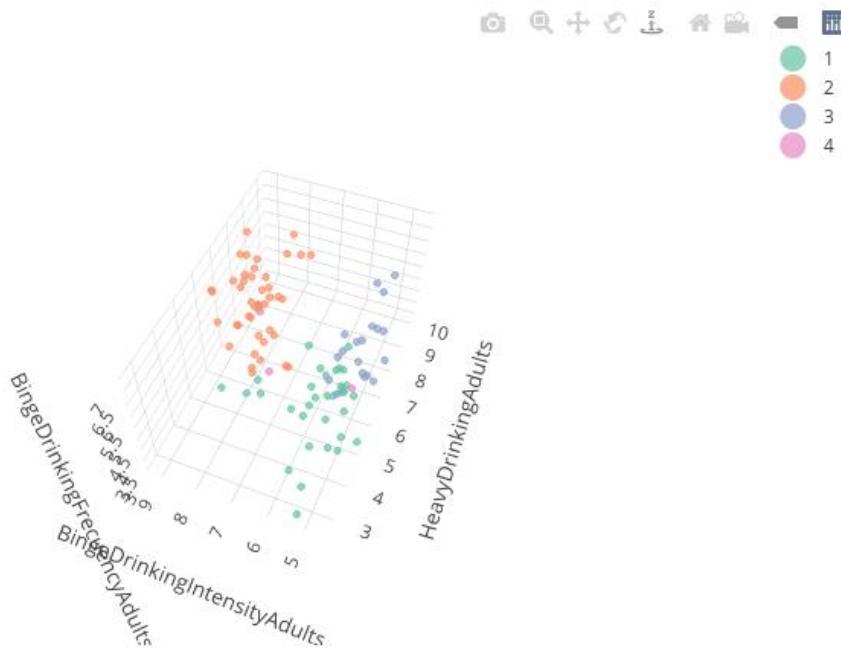
Modelos con K=4

El modelo de cuatro clústeres para los datos completos separa un grupo con los outliers, y dos de los grupos presentan un alto grado de solapamiento en la representación bidimensional.

k=4 grupos (datos completos)



Al representarlo en tres dimensiones, se observa que el solapamiento es menor. Por ejemplo, eligiendo las tres variables con mayor importancia para la agrupación del modelo $k=4$, se puede obtener este gráfico:

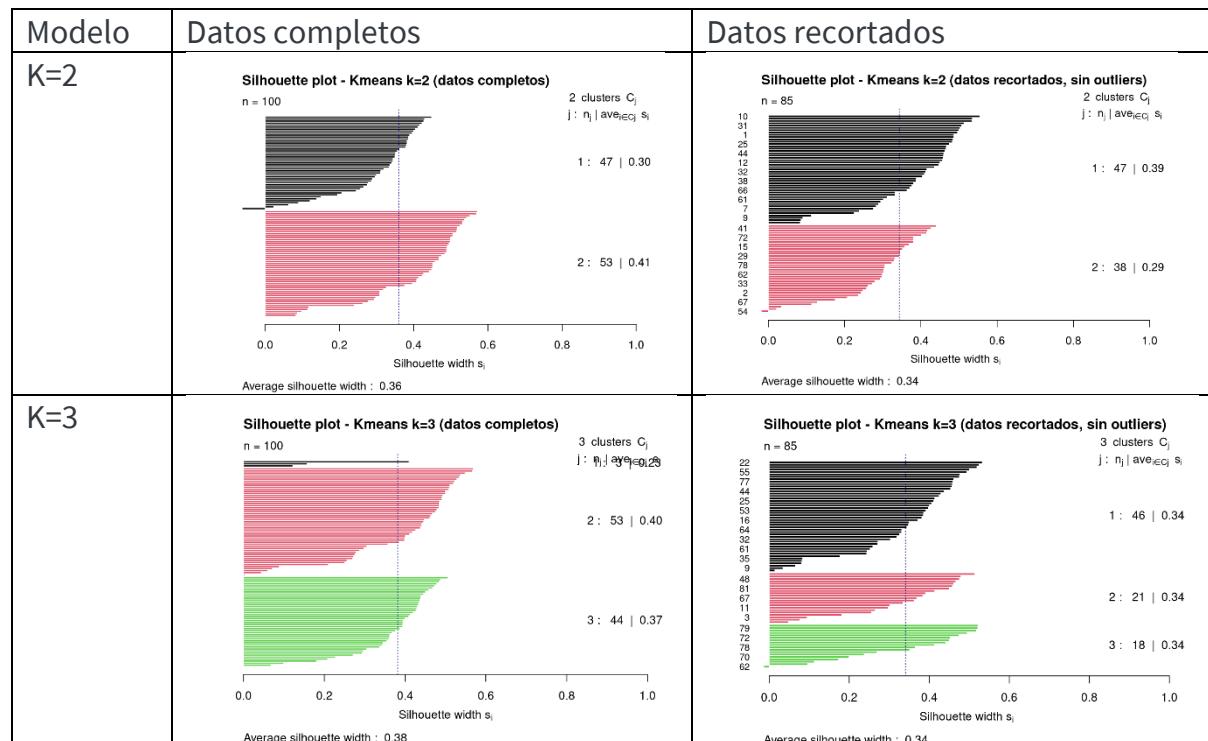


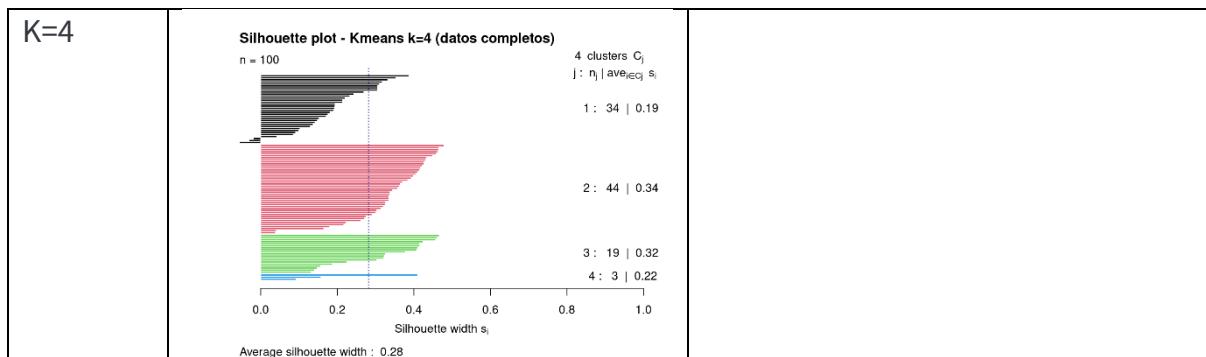
05fj - Validación de la agrupación

Interna

Para la validación interna se utilizó el diagrama de silueta, con los siguientes resultados:

Modelo	Datos	Evaluación del gráfico de silueta
$k=2$	Completos	Los dos clústeres tienen un rendimiento aceptable, aunque en el caso del cluster 1 es inferior a silueta media.
$k=3$	Completos	Dos de los clústeres, tienen un rendimiento bueno, y el cluster más pequeño tiene un rendimiento muy inferior a lo esperado
$k=4$	Completos	Sólo los clústeres 2 y 3 superaron la silueta media; todos los demás quedaron por debajo de lo deseable.
$k=2$	Recortados	Ambos dos clústeres tienen un rendimiento bueno, por encima de la silueta media
$k=3$	Recortados	Bastante equilibrado; todos los clústeres tienen un ancho de silueta medio igual al ancho de silueta medio.





Externa

Se utilizaron las siguientes variables para la validación externa de las agrupaciones:

- Para los modelos $k=2$, se utilizó la variable Sex.
- Para los modelos $k=3$ se creó una variable instrumental que discretizaba en tres niveles (alta, media y baja) la tasa de mortalidad ajustada por edad AgeAdjustedDeathRate.
- Para el modelo $k=4$ se creó una variable instrumental que discretizaba en cuatro niveles (muy alta, alta, baja y muy baja) la tasa de mortalidad ajustada por edad AgeAdjustedDeathRate.

Los dos modelos $k=2$, tanto para datos completos como recortados, se ajustan bastante bien a los niveles de la variable Sex, por lo que capturan una información similar a esta variable.

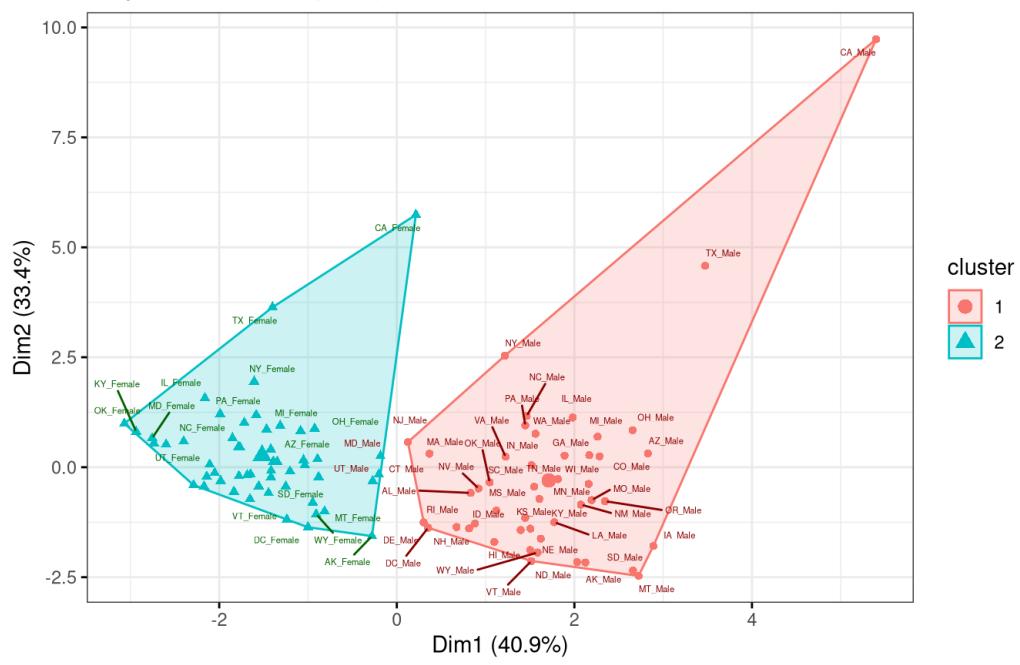
Los modelos $k=3$ y $k=4$ se relacionan mal con la variable instrumental creada discretizando los valores de la variable AgeAdjustedDeathRate, con lo que es razonable suponer que capturan información no contenida en estas variables.

Modelos de 2 grupos

Ambos modelos cluster (tanto el de datos completos como el de datos recortados) separan perfectamente a las mujeres, y se equivocan con un pequeño porcentaje de los hombres (6% en datos completos, 7.3% en datos recortados). Se observan unos valores elevados del estadístico de Rand, por lo que las observaciones incluidas en los clústeres son muy similares entre sí, tanto para los modelos de datos completos como para los de datos recortados.

Modelo con K=2 grupos

Conjunto de datos completo (con outliers)

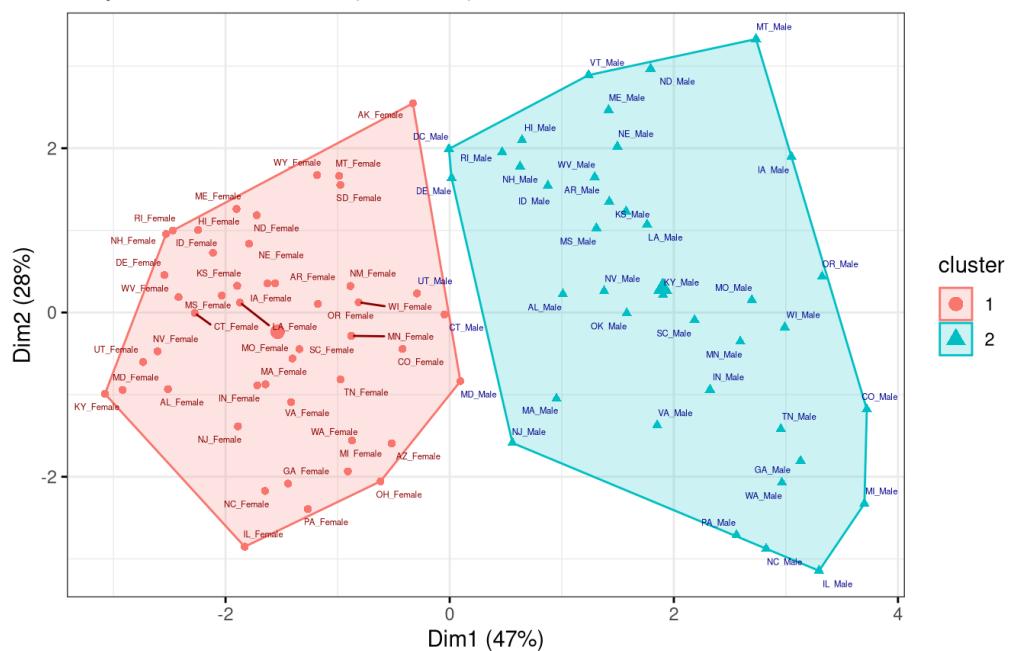


El modelo

con datos completos separa hombres y mujeres, sobre ajustándose por los outliers detectados. Comete errores en la clasificación de tres observaciones de hombres, con valores anormalmente bajos de los indicadores relacionados con el alcohol.

Modelo con K=2 grupos

Conjunto de datos recortado (sin outliers)



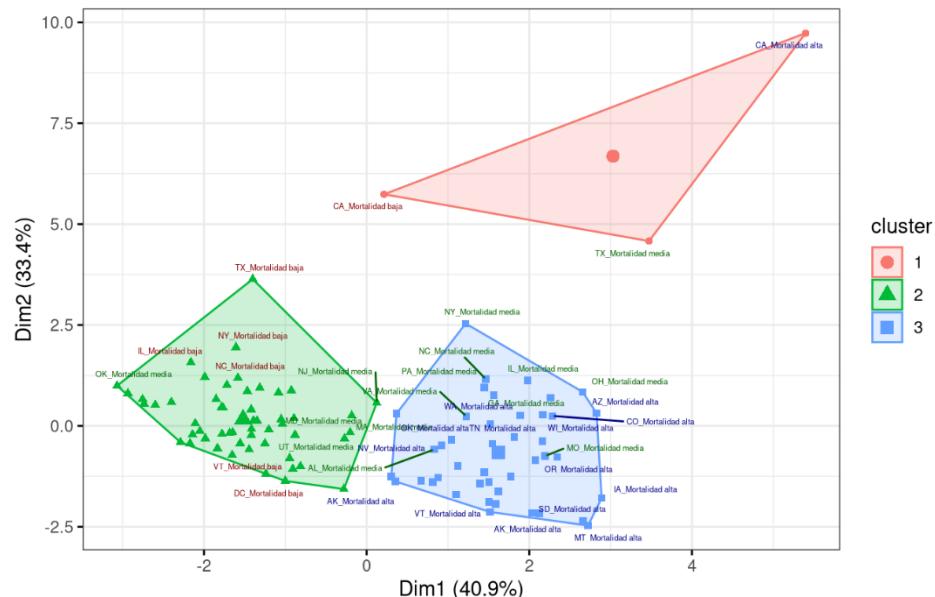
El modelo con datos recortados separa hombres y mujeres, sin el sobreajuste impuesto por los outliers. Comete errores en la clasificación de tres observaciones de hombres, con valores anormalmente bajos de los indicadores relacionados con el alcohol.

Modelos de 3 categorías

Ambos modelos cluster (tanto el de datos completos como el de datos recortados) separan mal los tres niveles de la variable AgeAdjustedDeathRate_fct3, con un importante número de discordancias entre lo esperado y lo observado. Se observan unos valores pobres del estadístico de Rand, por lo que las observaciones incluidas en los clústeres son muy distintas entre sí, tanto para los modelos de datos completos como para los de datos recortados.

Modelo con K=3 grupos

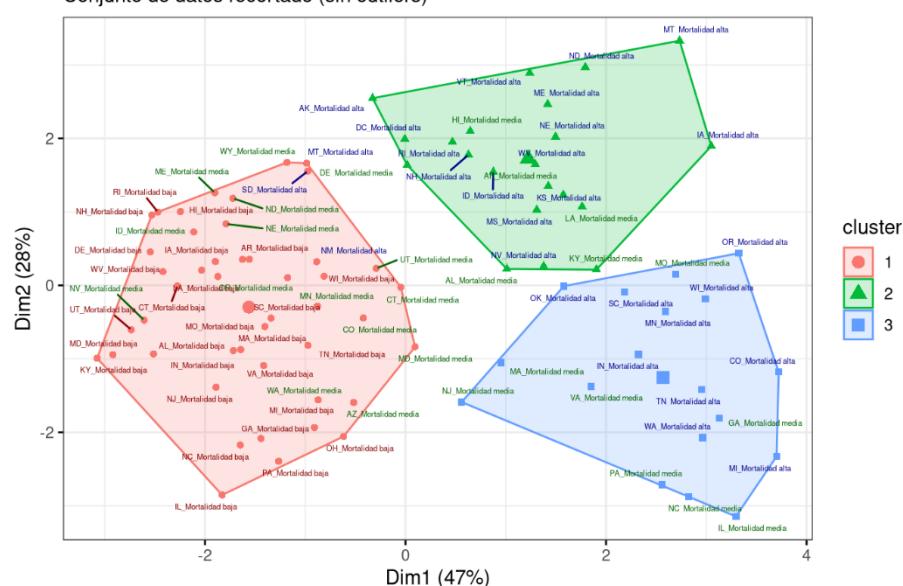
Conjunto de datos completo (con outliers)



El modelo de datos completos identifica razonablemente bien a las observaciones con mortalidad alta, pero a costa de equivocarse mucho en las que tiene mortalidad media y baja.

Modelo con K=3 grupos

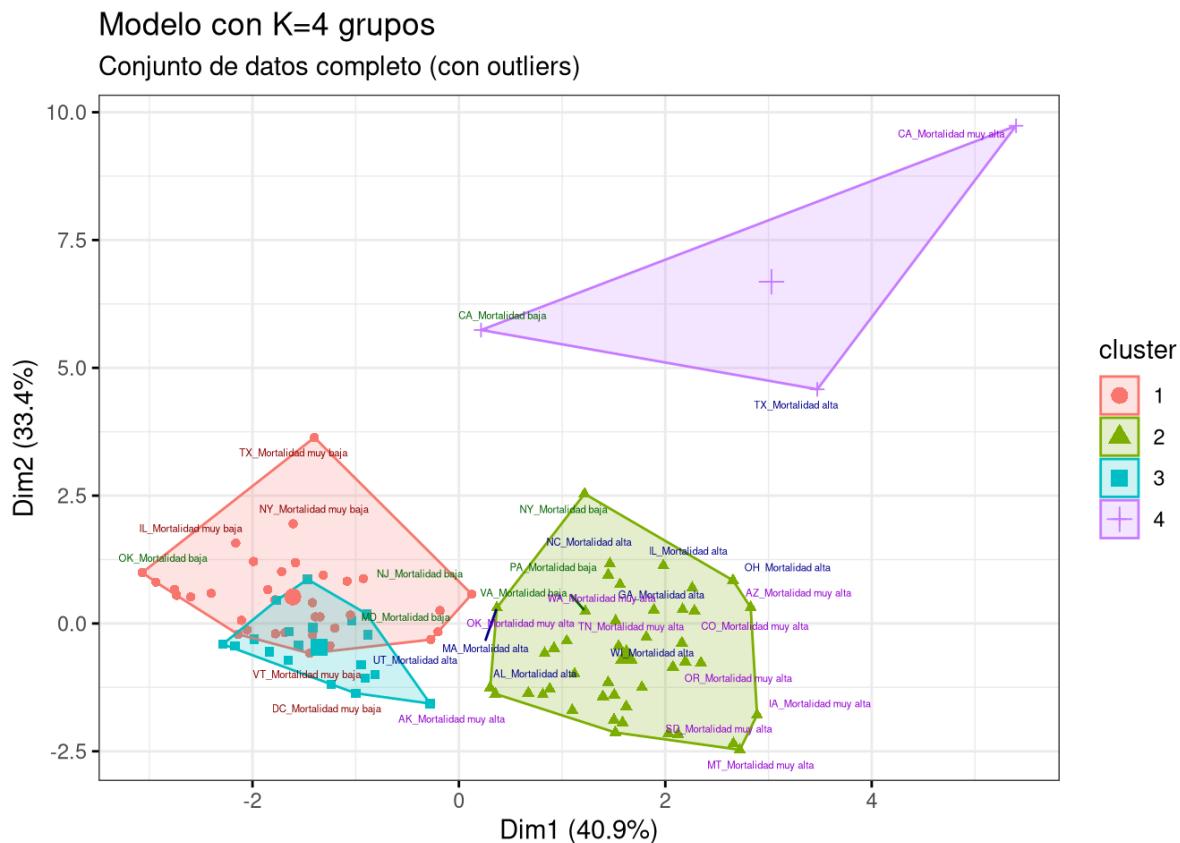
Conjunto de datos recortado (sin outliers)



El modelo con datos recortados no está separando adecuadamente los niveles de la variable **AgeAdjustedDeathRate_fct3**. La información separada en los clústeres tiene poco que ver con los niveles de esta variable categórica.

Modelo de 4 categorías

Ambos modelos cluster (tanto el de datos completos como el de datos recortados) separan mal los tres niveles de la variable **AgeAdjustedDeathRate_fct3**, con un importante número de discordancias entre lo esperado y lo observado. Se observa un valor bajo del estadístico de Rand, por lo que las observaciones incluidas en los clústeres son muy distintas entre sí. El modelo no está separando la información contenida en la variable **AgeAdjustedDeathRate_fct4**.



El modelo de datos completos para $k=4$ no identifica correctamente los niveles de la variable **AgeAdjustedDeathRate_fct4**.

05fk - Resumen de resultados obtenidos

- En ambos supuestos (datos totales y recortados), se observó una tendencia a la agrupación, tanto estadísticamente (Hopkins <0.5), por lo que se justifica realizar un análisis de agrupación.
- Para nuestro caso se utilizó un análisis cluster jerárquico. En la fase de análisis exploratorio se detectó la presencia de *outliers*, por lo que se replicó el análisis con o sin los datos, para valorar la influencia de estos.
- Se determinó que el número óptimo de clústeres se encontraba entre 2 y 4, para el conjunto de datos completo, y entre 2 y 3, para el modelo recortado sin *outliers*.
- Se crearon cinco modelos de agrupación, utilizando el método *k-means*, 3 para los datos completos, y 2 para los datos recortados, para los valores óptimos de cluster identificados.
- Respecto a la importancia de las variables para establecer la agrupación:
 - En los modelos de 2 y 3 clústeres para datos completos, fueron las relacionadas con las características de las borracheras (*BingeDrinkingIntensityAdults*, *BingeDrinkingPrevalenceAdults*, *BingeDrinkingFrecuencyAdults*)
 - Para el modelo con 4 clústeres, fue la de grandes bebedores (*HeavyDrinkingAdults*), seguida de las tres variables relacionadas con borracheras (*BingeDrinkingIntensityAdults*, *BingeDrinkingPrevalenceAdults*, *BingeDrinkingFrecuencyAdults*)
- Visualmente, los clústeres de los modelos pueden agrupar los datos sin solapamientos. El modelo de \emptyset presenta solapamientos en la representación en 2-D, pero evidencia buena capacidad discriminatoria en los modelos 3-D.
- Respecto a la evaluación de la validez de los modelos:
 - En lo que concierne a la validez interna, los modelos con mejor resultado han sido los $K=2$ y $K=3$ para datos recortados.
 - En lo tocante a validación externa:
 - Los dos modelos $K=2$, tanto para datos completos como recortados, se ajustan bastante bien a los niveles de la variable *Sex*, por lo que capturan una información similar a esta variable.
 - Los modelos \emptyset y \emptyset se relacionan mal con la variable instrumental creada discretizando los valores de la variable

`AgeAdjustedDeathRate`, con lo que es razonable suponer que capturan información no contenida en estas variables

Discusión

Limitaciones del estudio

Se han identificado las siguientes limitaciones del estudio, debido a la metodología seguida, que deben tenerse en cuenta a la hora de evaluar los resultados obtenidos:

- Fase de ingesta:
 - Se ingestaron los datos disponibles en las fuentes originales a fecha 13-Ago-2024
- Fase limpieza:
 - No se analizó la evolución temporal de cada indicador a lo largo del tiempo
 - No se analizó el detalle a nivel de condado de los indicadores
 - Sólo se analizó el año 2021
- Fase de transformación:
 - No se ejecutaron técnicas de imputación para los datos faltantes. El análisis omitió las observaciones del dataset que no tenían datos
 - No se disponía en el dataset de ninguna variable categórica que permitiera la validación externa de los modelos de 3 y 4 clústeres, por lo que se utilizó una variable instrumental sintética para dicha validación.

Discusión de los resultados obtenidos del estudio

- Existe una notable influencia del sexo en los indicadores de salud relacionados con el alcohol, para todos los estados analizados. Esto parece condicionar que los hombres tengan una mayor incidencia y prevalencia de factores deletéreos determinantes de la salud relacionados con el alcohol mayor que las mujeres, y un mayor número de muertes relacionadas con el alcohol. Esto es consistente con lo observado en estudios previos ([White 2020](#)).
- Algunos estados presentan unos valores para los indicadores anormalmente elevados, algo ya conocido ([CDC 2024](#)). Deben estudiarse en detalle estos estados, pues pueden aportar información valiosa sobre los factores de riesgo con mayor impacto en el desarrollo de efectos perjudiciales para la salud relacionados con el alcohol.
- Los indicadores relacionados con el consumo de alcohol de los datos analizados presentan una tendencia a la agrupación en clústeres, tanto en el conjunto de datos completo como en el recortado (*sin outliers*).
- Cuando se utiliza la agrupación no jerárquica por k -means, dos de los modelos de agrupación para datos recortados presentan suficiente solidez para ser tomados en consideración: $k=2$, y $k=3$.

- La variable del dataset que más se acerca a la agrupación de modelos con $k=2$ es Sex.
- No se ha identificado ninguna variable en el dataset que se solape con la agrupación de $k=3$, por lo que asumimos que esta clusterización captura información no registrada en el conjunto de datos original.

Conclusiones

Este estudio ofrece dos modelos de agrupación no jerárquica mediante k -means para indicadores de salud relacionados con el consumo de alcohol, que podrían utilizarse como apoyo a técnicas de reducción de la dimensionalidad en el análisis de datasets extensos en este dominio. Deben pro Las limitaciones identificadas en la metodología del estudio lastran la generalización de las conclusiones. Sería interesante profundizar en las características de los estados *outliers*, y en la utilidad de los dos modelos de análisis cluster para utilizarlos en modelos predictivos.

Bibliografía

- Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2024. *rmarkdown: Dynamic Documents for r*.
<https://github.com/rstudio/rmarkdown>.
- Bagnardi, V, M Rota, E Botteri, I Tramacere, F Islami, V Fedirko, L Scotti, et al. 2014. “Alcohol Consumption and Site-Specific Cancer Risk: A Comprehensive Doseresponse Meta-Analysis.” *British Journal of Cancer* 112 (3): 580–93.
<https://doi.org/10.1038/bjc.2014.579>.
- Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, and Toby Hocking. 2024. *data.table: Extension of “data.frame”*. <https://CRAN.R-project.org/package=data.table>.
- CDC. 2024. “Addressing Excessive Alcohol Use: State Fact Sheets.”
<https://www.cdc.gov/alcohol/fact-sheets/states/excessive-alcohol-use-united-states.html>.
- Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* 61 (6): 1–36.
<https://www.jstatsoft.org/v61/i06/>.
- Comtois, Dominic. 2022. *summarytools: Tools to Quickly and Neatly Summarize Data*.
<https://CRAN.R-project.org/package=summarytools>.
- Cui, Boxuan. 2024. *DataExplorer: Automate Data Exploration and Treatment*.
<https://CRAN.R-project.org/package=DataExplorer>.
- Data Science, yuzaR. 2021. “R Demo | Deep Exploratory Data Analysis (EDA) | Explore Your Data and Start to Test Hypotheses,” April.
https://www.youtube.com/watch?v=Swcp0_l65lw.
- Dayanand Ubrangala, Kiran R, Ravi Prasad Kondapalli, and Sayan Putatunda. 2024. *SmartEDA: Summarize and Explore the Data*. <https://CRAN.R-project.org/package=SmartEDA>.
- Dolnicar, Sara, Bettina Grün, and Friedrich Leisch. 2018. *Market Segmentation Analysis—Understanding It, Doing It, and Making It Useful*. Singapore: Springer.
<https://doi.org/10.1007/978-981-10-8818-6>.
- El Paquete Técnico SAFER. Un Mundo Libre de Los Daños Relacionados Con El Alcohol.*
2020. Organización Panamericana de la Salud.
<https://doi.org/10.37774/9789275321959>.
- Esser, Marissa B., Adam Sherk, Yong Liu, S. Jane Henley, and Timothy S. Naimi. 2024. “Reducing Alcohol Use to Prevent Cancer Deaths: Estimated Effects Among U.S. Adults.” *American Journal of Preventive Medicine* 66 (4): 725–29.
<https://doi.org/10.1016/j.amepre.2023.12.003>.

- Esser, Marissa B., Adam Sherk, Yong Liu, and Timothy S. Naimi. 2024. "Deaths from Excessive Alcohol Use United States, 2016–2021." *MMWR. Morbidity and Mortality Weekly Report* 73 (8): 154–61. <https://doi.org/10.15585/mmwr.mm7308a1>.
- Gohel, David, and Panagiotis Skintzos. 2024. *flextable: Functions for Tabular Reporting*. <https://CRAN.R-project.org/package=flextable>.
- Harrell Jr, Frank E. 2024. *Hmisc: Harrell Miscellaneous*. <https://CRAN.R-project.org/package=Hmisc>.
- Hennig, Christian. 2024. *fpc: Flexible Procedures for Clustering*. <https://CRAN.R-project.org/package=fpc>.
- Holt, James B., Sara L. Huston, Khosrow Heidari, Randy Schwartz, Charles W. Gollmar, Annie Tran, Leah Bryan, Yong Liu, and Janet B. Croft. 2015. "Indicators for chronic disease surveillance - United States, 2013." *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports* 64 (RR-01): 1–246.
- International Agency for Research on Cancer, and International Agency for Research on Cancer, eds. 1988. *Alcohol drinking: views and expert opinions: this publication represents the views and expert opinions of an IARC Working Group on the Evaluation of the Carcinogenic Risk of Chemicals to Humans which met in Lyon, 13 - 20 October 1987*. IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans 44. Lyon: International Agency for Research on Cancer.
- Kassambara, Alboukadel, and Fabian Mundt. 2020. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. <https://CRAN.R-project.org/package=factoextra>.
- Komsta, Lukasz, and Frederick Novomestky. 2022. *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. <https://CRAN.R-project.org/package=moments>.
- Landis, Justin. 2024. *ggside: Side Grammar Graphics*. <https://CRAN.R-project.org/package=ggside>.
- Leisch, Friedrich. 2006. "A Toolbox for k-Centroids Cluster Analysis." *Computational Statistics and Data Analysis* 51 (2): 526–44. <https://doi.org/10.1016/j.csda.2005.10.006>.
- . 2010. "Neighborhood Graphs, Stripes and Shadow Plots for Cluster Visualization." *Statistics and Computing* 20: 457–69. <https://doi.org/10.1007/s11222-009-9137-8>.
- Leisch, Friedrich, and Bettina Grün. 2006. "Extending Standard Cluster Algorithms to Allow for Group Constraints." In *Compstat 2006—Proceedings in Computational*

Statistics, edited by Alfredo Rizzi and Maurizio Vichi, 885–92. Physica Verlag, Heidelberg, Germany.

- Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. “performance: An R Package for Assessment, Comparison and Testing of Statistical Models.” *Journal of Open Source Software* 6 (60): 3139. <https://doi.org/10.21105/joss.03139>.
- Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2023. *cluster: Cluster Analysis Basics and Extensions*. <https://CRAN.R-project.org/package=cluster>.
- Ministerio de Sanidad, and Instituto Nacional de Estadística. n.d. “Encuesta Nacional de Salud de España 2017.” <https://www.sanidad.gob.es/estadEstudios/estadisticas/encuestaNacional/encuesta2017.htm>.
- Morgan, Martin, and Marcel Ramos. 2024. *BiocManager: Access the Bioconductor Project Package Repository*. <https://CRAN.R-project.org/package=BiocManager>.
- Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.
- Organización Mundial de la Salud. 2010. “Estrategia mundial para reducir el uso nocivo del alcohol,” 40. <https://iris.who.int/handle/10665/44486>.
- Ortiz-Ospina, Esteban, and Max Roser. 2016. “Alcohol and Drug Use Disorders Death Rate.” <https://ourworldindata.org/grapher/death-rates-substance-disorders>.
- Patil, Indrajeet. 2021. “Visualizations with statistical details: The ‘ggstatsplot’ approach.” *Journal of Open Source Software* 6 (61): 3167. <https://doi.org/10.21105/joss.03167>.
- Pfaffel, Oliver. 2021. *FeatureImpCluster: Feature Importance for Partitional Clustering*. <https://CRAN.R-project.org/package=FeatureImpCluster>.
- Ryu, Choonghyun. 2024. *dlookr: Tools for Data Diagnosis, Exploration, Transformation*. <https://CRAN.R-project.org/package=dlookr>.
- Scharl, Theresa, and Friedrich Leisch. 2006. “The Stochastic QT-Clust Algorithm: Evaluation of Stability and Variance on Time-Course Microarray Data.” In *Compstat 2006—Proceedings in Computational Statistics*, edited by Alfredo Rizzi and Maurizio Vichi, 1015–22. Physica Verlag, Heidelberg, Germany.
- Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2024. *GGally: Extension to ggplot2*. <https://CRAN.R-project.org/package=GGally>.
- Slowikowski, Kamil. 2024. *ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2”*. <https://CRAN.R-project.org/package=ggrepel>.

- Tierney, Nicholas, and Dianne Cook. 2023. “Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations.” *Journal of Statistical Software* 105 (7): 1–31. <https://doi.org/10.18637/jss.v105.i07>.
- van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate Imputation by Chained Equations in r.” *Journal of Statistical Software* 45 (3): 1–67. <https://doi.org/10.18637/jss.v045.i03>.
- Wei, Taiyun, and Viliam Simko. 2021. *R Package “corrplot”: Visualization of a Correlation Matrix*. <https://github.com/taiyun/corrplot>.
- White, A. 2020. “Gender Differences in the Epidemiology of Alcohol Use and Related Harms in the United States.” *Alcohol Research: Current Reviews* 40 (2). <https://doi.org/10.35946/arcr.v40.2.01>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. *devtools: Tools to Make Developing r Packages Easier*. <https://CRAN.R-project.org/package=devtools>.
- Wikipedia. 2020. “Gráfico q-q — Wikipedia, La Enciclopedia Libre.” https://es.wikipedia.org/w/index.php?title=Gr%C3%A1fico_Q-Q&oldid=130441335.
- William Revelle. 2024. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. <https://CRAN.R-project.org/package=psych>.
- Wright, Kevin, Luo YiLan, and Zeng RuTong. 2023. *clustertend: Check the Clustering Tendency*. <https://CRAN.R-project.org/package=clustertend>.
- Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.
- Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>.
- Zhu, Hao. 2024. *kableExtra: Construct Complex Table with “kable” and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.

Anexo 00 – Paquetes de R utilizados en el análisis

	Title	Version
arsenal	An Arsenal of ‘R’ Functions for Large-Scale Statistical Summaries	3.6.3
BiocManager	Access the Bioconductor Project Package Repository	1.30. 23
clustertend	Check the Clustering Tendency	1.7
data.table	Extension of <code>data.frame</code>	1.15. 4
DataExplorer	Automate Data Exploration and Treatment	0.8.3
devtools	Tools to Make Developing R Packages Easier	2.4.5
dlookr	Tools for Data Diagnosis, Exploration, Transformation	0.6.3
dplyr	A Grammar of Data Manipulation	1.1.4
factoextra	Extract and Visualize the Results of Multivariate Data Analyses	1.0.7
FeatureImpCluster	Feature Importance for Partitional Clustering	0.1.5
flexclust	Flexible Cluster Algorithms	1.4- 2
flextable	Functions for Tabular Reporting	0.9.6
fpc	Flexible Procedures for Clustering	2.2- 12
GGally	Extension to ‘ggplot2’	2.2.1
ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	3.5.1
ggpubr	‘ggplot2’ Based Publication Ready Plots	0.6.0
ggrepel	Automatically Position Non-Overlapping Text Labels with ‘ggplot2’	0.9.5
ggsome	Side Grammar Graphics	0.3.1
graph	graph: A package to handle graph data structures	1.82. 0
grateful	Facilitate Citation of R Packages	0.2.4
here	A Simpler Way to Find Your Files	1.0.1
kableExtra	Construct Complex Table with ‘kable’ and Pipe Syntax	1.4.0
mice	Multivariate Imputation by Chained Equations	3.16. 0
moments	Moments, Cumulants, Skewness, Kurtosis and Related Tests	0.14. 1

naniar	Data Structures, Summaries, and Visualisations for Missing Data	1.1.0
NbClust	Determining the Best Number of Clusters in a Data Set	3.0.1
plotly	Create Interactive Web Graphics via ‘plotly.js’	4.10. 4
psych	Procedures for Psychological, Psychometric, and Personality Research	2.4.6 .26
rstantools	Tools for Developing R Packages Interfacing with ‘Stan’	2.4.0
SmartEDA	Summarize and Explore the Data	0.3.1 0
summarytools	Tools to Quickly and Neatly Summarize Data	1.0.1
tidycensus	Load US Census Boundary and Attribute Data as ‘tidyverse’ and ‘sf’-Ready Data Frames	1.6.5
tidyrr	Tidy Messy Data	1.3.1
tinytex	Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents	0.52
usmap	US Maps Including Alaska and Hawaii	0.7.1
usmapdata	Mapping Data for ‘usmap’ Package	0.3.0
utils	The R Utils Package	4.4.1

Anexo 1 – Subproceso de ingesta – Detalle de acciones del proceso

Las bases de datos utilizadas en el análisis fueron las siguientes:

1 - Indicadores de salud relacionados con el consumo de alcohol

El conjunto de datos *Cronic Disease Indicators (CDI)* ha sido elaborado por el *Centres of Disease Control* de USA. Contiene un conjunto de 124 indicadores desarrollados por consenso entre todos los estados miembros de US, utilizado para definir, recoger e informar sobre las enfermedades crónicas de manera uniforme entre distintos estados y territorios. Los indicadores se agrupan en 17 áreas de interés. Puede consultarse una descripción detallada de los indicadores que contiene en el artículo *Indicators for Chronic Disease Surveillance — United States, 2013* (Holt et al. 2015). Para una descripción más detallada, puede consultarse la [web del CDI](#).

Para nuestro análisis, se seleccionó el subconjunto de indicadores del área de interés Alcohol.

2 - Tasas de mortalidad relacionadas con el consumo de alcohol

El conjunto de datos *Underlying Cause of Death* contiene datos de mortalidad y población para todos los condados de EE. UU. Los datos provienen de los certificados de defunción de los residentes de EE. UU. Cada certificado de muerte identifica una única causa de muerte, junto con un conjunto de datos demográficos.

Para nuestro análisis, se seleccionaron los datos de mortalidad por alcohol, ajustada por edad, estado y año. Se descargaron los datos globales para comparar la media de cada nivel de las variables con la media global.

Los datos para el análisis se obtuvieron con la siguiente configuración de la herramienta de búsqueda [CDC WONDER](#):

Sección	Subsección	Parámetro	Configuración
1. Organize table layout	Group Results By	· Group Results By · And By · And By	State Year Gender
	Default Measures	· Deaths · Population · Crude Rate	TRUE TRUE TRUE
	For Crude Rates	· 95% Confidence Interval · Standard Error	TRUE TRUE
	Age Adjusted Rate	· Age Adjusted Rate · 95% Confidence Interval · Standard Error	TRUE TRUE TRUE

	Total Deaths	· Percent of Total Deaths	TRUE
	Additional Rate Options	· Calculate Rates Per · Standard Population	100,000 2000 U.S. Std. Population
2. Select location	Grouping method	· Click a button to choose locations by US-Mexico Border Region, Border State Area, State, Census Region or HHS Region	States
	Selected codes	· Browse or search to find items in the States Finder Tool, then highlight the items to use for this request.	All (The United States)
	Urbanization	· Pick between	2013 Urbanization, All categories
3. Select demographic s	Age Groups	· Pick between	All ages
	Gender	· Gender	All Genders
	Hispanic origin	· Hispanic origin	All Origins
	Single Race	· Pick between	Single Race 6 , All Races
4. Select year and month		· Year/Month	All (All Dates)
5. Select weekday, autopsy and place of death	Weekday	· Weekday	All Weekdays
	Autopsy	· Autopsy	All values
	Place of death	· Place of death	All places
6. Select cause of death	Grouping method	· Click a button to select ICD codes by Chapters or by Groups	Drug/Alcohol Induced Causes
	Selected codes	· Browse or search to find items in the Drug/Alcohol Induced Causes Finder Tool, then highlight the items to use for this request	Alcohol-induced causes

7. Other options	<ul style="list-style-type: none"> · Export results · Show totals · Show Zero values · Show suppressed values · Precision (decimal places) · Data Access Timeout 	TRUE FALSE TRUE TRUE 2 10
------------------	--	--

3 - Códigos FIPS de los estados de EE. UU.

Para garantizar una adecuada estandarización de los datos, se utilizó como tabla maestra el dataset `fips_states` del paquete **tidycensus** [Dataset with FIPS codes for US states and counties](#).

01b - Definir el método y la configuración de la ingesta

Conjunto de datos	Método	Configuración
1 - Indicadores de salud relacionados con el consumo de alcohol	Ingesta directa desde origen con <code>data.table::fread()</code>	Opciones por defecto de la función, salvo <code>encoding = "UTF-8"</code>
2 - Tasas de mortalidad relacionadas con el consumo de alcohol	<ul style="list-style-type: none"> · Descarga del conjunto de datos crudo con la configuración especificada · Ingesta desde descarga local 	<code>encoding = "UTF-8"</code> <code>header = TRUE</code> <code>nrow = 306</code>
3 - Códigos FIPS de los estados de EE. UU.	Carga directa con <code>data()</code>	Opciones por defecto de la función

01c - Crear los data.frame de datos crudos

Se crearon los siguientes *data.frame* de datos crudos:

Objeto	Fuente	Descripción
<code>rawCdiAlcohol</code>	<i>Chronic disease indicators (CDI), CDC, 2023</i>	Evaluación de los indicadores de enfermedades crónicas relacionados con el consumo excesivo de alcohol, ajustados por sexo y edad, durante el periodo 2010-21, por estado y año
<code>rawMortalityAlcohol</code>	<i>Underlying Cause of Death, 2018-2022, Single Race</i>	Tasa de mortalidad relacionada con alcohol, ajustada por sexo y edad, durante el periodo 2018-22, por estado y año
<code>rawFipsCodes</code>	Paquete tidycensus	Tabla maestra de códigos para estados y condados de EE. UU.

01d - Validar la fase de ingestá

En la validación de la fase de ingestá se realizaron las siguientes actividades:

- Verificar la completitud de la ingestá `head()`, `tail()`, `dim()`
- Normalizar los nombres de las variables
- Comprobar la estructura de los `data.frame` (`str()`)
- Características generales del `data.frame`

1 - Validación de `rawCdiAlcohol`

Dimensión	Evaluación
Completitud	Ingestá correcta
Normalización del nombre de variables	No es necesaria
Estructura del <code>data.frame</code>	Algunas variables deben convertirse a tipo <code>factor</code>
Resumen del <code>data.frame</code>	Existen algunas variables innecesarias (sin datos, o con todos los valores iguales) Hay valores faltantes dispersos (NA)

2 - Validación de `rawUnderlyingCauseOfDeathAlcohol`

Dimensión	Evaluación
Completitud	Faltan datos para algunos estados en algunos años
Normalización del nombre de variables	· Los nombres de las variables del dataset original no están normalizados. · Deben renombrarse las variables.
Estructura del <code>data.frame</code>	· Algunas variables deben convertirse a tipo <code>factor</code> · Alguna variable debe convertirse a tipo <code>numeric</code>
Resumen del <code>data.frame</code>	· Debe convertirse a numérico el valor % of Total Deaths · Hay valores faltantes dispersos (NA)

3 - Validación de `rawFipsCodes`

Dimensión	Evaluación
Completitud	Ingestá correcta
Normalización del nombre de variables	Debe convertirse a Pascal Case
Estructura del <code>data.frame</code>	Deben convertirse algunas variables al tipo <code>factor</code>
Resumen del <code>data.frame</code>	Sin problemas adicionales

Anexo 02 – Subproceso de limpieza - Evaluación de los datasets crudos y acciones de limpieza ejecutadas

02a - Identificación de la información sucia, incorrecta, irrelevante, incompleta, imprecisa o incómoda

Validación de rawCdiAlcohol

Problema del dato	Subtipo de problema	Valoración del dataset	Acción correctiva
Sucio	Codificación de caracteres incorrecta	CORRECTO	
	Símbolos innecesarios (\$, €, %, ...)	CORRECTO	
	Nombre de <i>data.frame</i> no acorde a estilo	CORRECTO	
	Nombre de variables (columnas) no acorde a estilo	CORRECTO	
	Nombre de observaciones (filas) no acorde a estilo	NO APLICA	
Incorrecto	Errores del subproceso de ingestión	CORRECTO	
	Datos no ordenados - Más de una variable por columna	CORRECTO	
	Datos no ordenados - Más de una observación por fila	CORRECTO	
	Datos no ordenados - Más de un dato por registro	CORRECTO	
Irrelevante	Variables con todas las observaciones faltantes	INCORRECTO	Eliminación de variables
	Variables con todas las observaciones con el mismo valor	INCORRECTO	Eliminación de variables
	Variables innecesarias para el análisis (columnas)	INCORRECTO	Eliminación de variables
	Observaciones innecesarias para el análisis (filas)	INCORRECTO	Agrupación de observaciones
Incompleto	Cobertura incompleta de observaciones (filas)	INCORRECTO	Selección de año 2021 (mejor cobertura)
	Cobertura incompleta de variables (columnas)	CORRECTO	

	Cobertura incompleta de datos (NA)	INCORRECTO	Análisis de variables completas
	Cobertura incompleta de periodos (series temporales)	CORRECTO	
Impreciso	Tipado de variable incorrecto	INCORRECTO	Retipado de variables
	Precisión decimal insuficiente	CORRECTO	
	Cardinalidad inadecuada (demasiadas o insuficientes categorías)	CORRECTO	
	Datos impuntuales (<i>punctuality</i>) (series temporales)	NO APLICA	
	Datos desactualizados (<i>freshness</i>) (series temporales)	CORRECTO	
Incómodo	Formato inadecuado para el análisis (largo / ancho)	CORRECTO	
	Orden incorrecto de variables (columnas)	CORRECTO	
	Orden incorrecto de observaciones (filas)	CORRECTO	

Validación de rawUnderlyingCauseOfDeathAlcohol

Problema del dato	Subtipo de problema	Problemas del dataset	Acción
Sucio	Codificación de caracteres incorrecta	CORRECTO	
	Símbolos innecesarios (\$, €, %, ...)	INCORRECTO	Transformación de variable
	Nombre de <i>data.frame</i> no acorde a estilo	CORRECTO	
	Nombre de variables (columnas) no acorde a estilo	INCORRECTO	Renombrado de variables
	Nombre de observaciones (filas) no acorde a estilo	NO APLICA	
Incorrecto	Errores del subproceso de ingesta	CORRECTO	
	Datos no ordenados - Más de una variable por columna	CORRECTO	
	Datos no ordenados - Más de una observación por fila	CORRECTO	
Irrelevante	Variables con todas las observaciones faltantes	CORRECTO	
	Variables con todas las observaciones con el mismo valor	CORRECTO	
	Observaciones innecesarias para el análisis (filas)	INCORRECTO	Filtrado de observaciones
	Variables innecesarias para el análisis (columnas)	INCORRECTO	Eliminación de variables
Incompleto	Cobertura incompleta de observaciones (filas)	CORRECTO	
	Cobertura incompleta de variables (columnas)	CORRECTO	
	Cobertura incompleta de datos (NA)	CORRECTO	
	Cobertura incompleta de periodos (series temporales)	CORRECTO	
Impreciso	Tipado de variable incorrecto	INCORRECTO	Retipado de variables

	Precisión decimal insuficiente	CORRECTO	
	Cardinalidad inadecuada (demasiadas o insuficientes categorías)	CORRECTO	
	Datos impuntuales (punctuality) (series temporales)	NO APLICA	
	Datos desactualizados (freshness) (series temporales)	NO APLICA	
Incómodo	Formato inadecuado para el análisis (largo / ancho)	CORRECTO	
	Orden incorrecto de variables (columnas)	CORRECTO	
	Orden incorrecto de observaciones (filas)	CORRECTO	

Validación de rawFipsCodes

Problema del dato	Subtipo de problema	Problemas del dataset	Acción
Sucio	Codificación de caracteres incorrecta	CORRECTO	
	Símbolos innecesarios (\$, €, %, ...)	CORRECTO	
	Nombre de <i>data.frame</i> no acorde a estilo	CORRECTO	
	Nombre de variables (columnas) no acorde a estilo	INCORRECTO	Renombrado de variables
	Nombre de observaciones (filas) no acorde a estilo	NO APLICA	
Incorrecto	Errores del subprocesso de ingestá	CORRECTO	
	Datos no ordenados - Más de una variables por columna	CORRECTO	
	Datos no ordenados - Más de una observación por fila	CORRECTO	
	Datos no ordenados - Más de un dato por registro	CORRECTO	

Irrelevante	Variables con todas las observaciones faltantes	CORRECTO	
	Variables con todas las observaciones con el mismo valor	CORRECTO	
	Variables innecesarias para el análisis (columnas)	INCORRECTO	Eliminación de variables
	Observaciones innecesarias para el análisis (filas)	CORRECTO	
Incompleto	Cobertura incompleta de observaciones (filas)	CORRECTO	
	Cobertura incompleta de variables (columnas)	CORRECTO	
	Cobertura incompleta de datos (NA)	NO APLICA	
	Cobertura incompleta de periodos (series temporales)	CORRECTO	
Impreciso	Tipado de variable incorrecto	CORRECTO	
	Precisión decimal insuficiente	CORRECTO	
	Cardinalidad inadecuada (demasiadas o insuficientes categorías)	CORRECTO	
	Datos impuntuales (<i>punctuality</i>) (series temporales)	NO APLICA	
	Datos desactualizados (<i>freshness</i>) (series temporales)	NO APLICA	
Incómodo	Formato inadecuado para el análisis (largo / ancho)	CORRECTO	
	Orden incorrecto de variables (columnas)	CORRECTO	
	Orden incorrecto de observaciones (filas)	CORRECTO	

02b - Reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo a la necesidad

Una vez limpios los tres data.frame de origen, se crearon tres conjuntos de datos para el análisis:

- `data`: Datos completos
- `data_overall`: Datos globales, sin estratificación por sexo
- `data_gender`: Datos estratificados por sexo

Anexo 03 – Acciones del subprocesso de Análisis exploratorio de datos (EDA)

03a - Visión general del *data.frame*: `summarytools::dfSummary()`

Se realizó un resumen de cada *data.frame* con los siguientes elementos:

- Nombre de variables y tipos,
- Etiquetas (si existían)
- Niveles de los factores,
- Frecuencias y / o estadísticos de resumen numéricos,
- Gráficos de barras / histogramas, y
- Conteos y proporciones de observaciones válidas / faltantes.

03b - Explorar variables categóricas: `SmartEDA::ExpCatViz()`

En el análisis se revisó para cada variable categórica:

- Frecuencia relativa de cada nivel de la variable respecto al total de observaciones
- La frecuencia relativa de datos faltantes, si existían

Se identificaron 2 variables categóricas (Sex, 2 niveles y State, con 53 niveles). La distribución de los niveles de las variables fue homogénea en todo el *dataset*, con algunos valores faltantes identificados.

03c - Explorar variables numéricas (Estadística descriptiva): `SmartEDA::ExpNumStat()`

Para las variables numéricas de cada *data.frame* se exploraron los siguientes aspectos:

- Número de variables (columnas) y de observaciones (filas)
- Número y porcentaje de valores faltantes (NA)
- Valoración de la escala de magnitud de cada variable, y comparación relativa entre las variables
- Evaluación del sesgo y la kurtosis
- Análisis básico de *outliers*

Los resultados de la evaluación fueron los siguientes:

Data.frame	Resultado de la evaluación
data	Respecto a las variables numéricas del <i>data.frame</i> :

	<ul style="list-style-type: none"> • Existen 8 variables numéricas, con 163 observaciones en total • Existen valores faltantes NA en todas las variables, con un rango de valores faltantes entre el 2.454% y el 38.65% • Las variables tienen una escala de magnitud muy diferente entre sí • La mitad de las variables están más o menos centradas, y la otra mitad sesgadas. Sólo una de las variables tiene una kurtosis similar a la de la distribución normal. Cuatro de las variables son más achataadas, y tres más picudas. • Algunas variables tienen bastantes <i>outliers</i>. Las más afectadas son <i>AgeAdjustedDeathRate</i>, <i>Deaths</i>, <i>PercentageOfTotalDeaths</i> y <i>Population</i>.
<i>data_gender</i>	Respecto a las variables numéricas del <i>data.frame</i> : <ul style="list-style-type: none"> • Existen 8 variables numéricas, con 106 observaciones en total • Existen valores faltantes NA en cinco variables, con un rango de valores faltantes entre el 0% y el 5.66% • Las variables tienen una escala de magnitud muy diferente entre sí • La mitad de las variables están más o menos centradas, y la otra mitad sesgadas. Sólo una de las variables tiene una kurtosis similar a la de la distribución normal. Cuatro de las variables son más achataadas, y tres más picudas. • Las variables con mayor número de <i>outliers</i> son <i>AgeAdjustedDeathRate</i>, <i>Deaths</i>, <i>PercentageOfTotalDeaths</i> y <i>Population</i>.
<i>data_overall</i>	Respecto a las variables numéricas del <i>data.frame</i> : <ul style="list-style-type: none"> • Existen 8 variables numéricas, con 153 observaciones en total • Existen valores faltantes NA en cuatro variables, con un rango de valores faltantes entre el 0% y el 1.961% • Las variables tienen una escala de magnitud muy diferente entre sí • La mitad de las variables están más o menos centradas, y la otra mitad sesgadas. Sólo una de las variables tiene una kurtosis similar a la de la distribución normal. Cuatro de las variables son más achataadas, y tres más picudas. • Las variables con mayor número de <i>outliers</i> son <i>AgeAdjustedDeathRate</i>, <i>Deaths</i>, <i>PercentageOfTotalDeaths</i> y <i>Population</i>.

03d - Explorar distribuciones (skewness and kurtosis tests)

Se evaluó la distribución de probabilidad de las variables numéricas. Para ello, se utilizaron las siguientes técnicas:

- 03da - Visualización de las distribuciones (histograma y función de densidad)
- 03db - Test de hipótesis de la centralidad (kurtosis y sesgo)

Cuando ambos test son no significativos, la variable puede seguir una distribución aproximadamente normal.

Los resultados obtenidos al aplicar los test a las variables numéricas del objeto `data` fueron los siguientes:

Variable	Test kurtosis Anscombe-Glynn	Test sesgo D'Agostino
Deaths	<0.05	<0.05
Population	<0.05	<0.05
AgeAdjustedDeathRate	<0.05	<0.05
PercentageOfTotalDeaths	<0.05	<0.05
HeavyDrinkingAdults	No significativo	No significativo
BingeDrinkingFrecuencyAdults	No significativo	No significativo
BingeDrinkingIntensityAdults	<0.05	No significativo
BingeDrinkingPrevalenceAdults	No significativo	No significativo

Los resultados obtenidos al aplicar los test a las variables numéricas del objeto `data_gender` fueron los siguientes:

Variable	Test kurtosis Anscombe-Glynn	Test sesgo D'Agostino
Deaths	<0.05	<0.05
Population	<0.05	<0.05
AgeAdjustedDeathRate	<0.05	<0.05
PercentageOfTotalDeaths	<0.05	<0.05
HeavyDrinkingAdults	<0.05	No significativo
BingeDrinkingFrecuencyAdults	No significativo	No significativo
BingeDrinkingIntensityAdults	<0.05	No significativo
BingeDrinkingPrevalenceAdults	<0.05	No significativo

Los resultados obtenidos al aplicar los test a las variables numéricas del objeto `data_overall` fueron los siguientes:

Variable	Test kurtosis Anscombe-Glynn	Test sesgo D'Agostino
Deaths	<0.05	<0.05
Population	<0.05	<0.05
AgeAdjustedDeathRate	<0.05	<0.05
PercentageOfTotalDeaths	<0.05	<0.05
HeavyDrinkingAdults	No significativo	No significativo
BingeDrinkingFrecuencyAdults	No significativo	<0.05
BingeDrinkingIntensityAdults	<0.05	No significativo
BingeDrinkingPrevalenceAdults	No significativo	No significativo

03e - Explorar normalidad (QQ-plots and Shapiro-Wilk)

Para explorar la normalidad de las variables se utilizaron las siguientes técnicas:

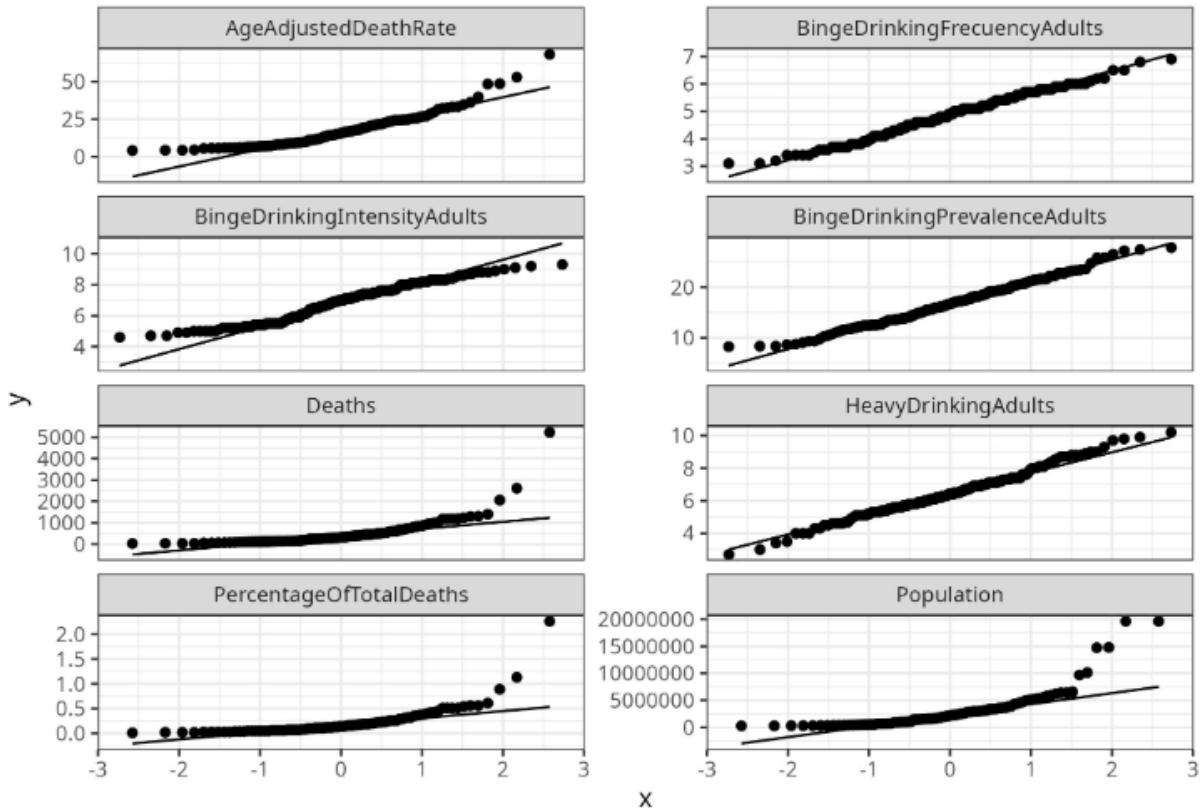
Técnica	Descripción
Evaluación gráfica: QQ-Plot	Método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación (Wikipedia 2020)
Test de Shapiro-Wilks	Se considera el test más potente para testar la normalidad, seguido de cerca por el Anderson–Darling Si la muestra es muy grande, es posible que el test detecte desviaciones mínimas frente a la normal, que no tengan importancia práctica. Por eso el test debe interpretarse siempre conjuntamente con el gráfico QQ-plot.

Objeto `data`

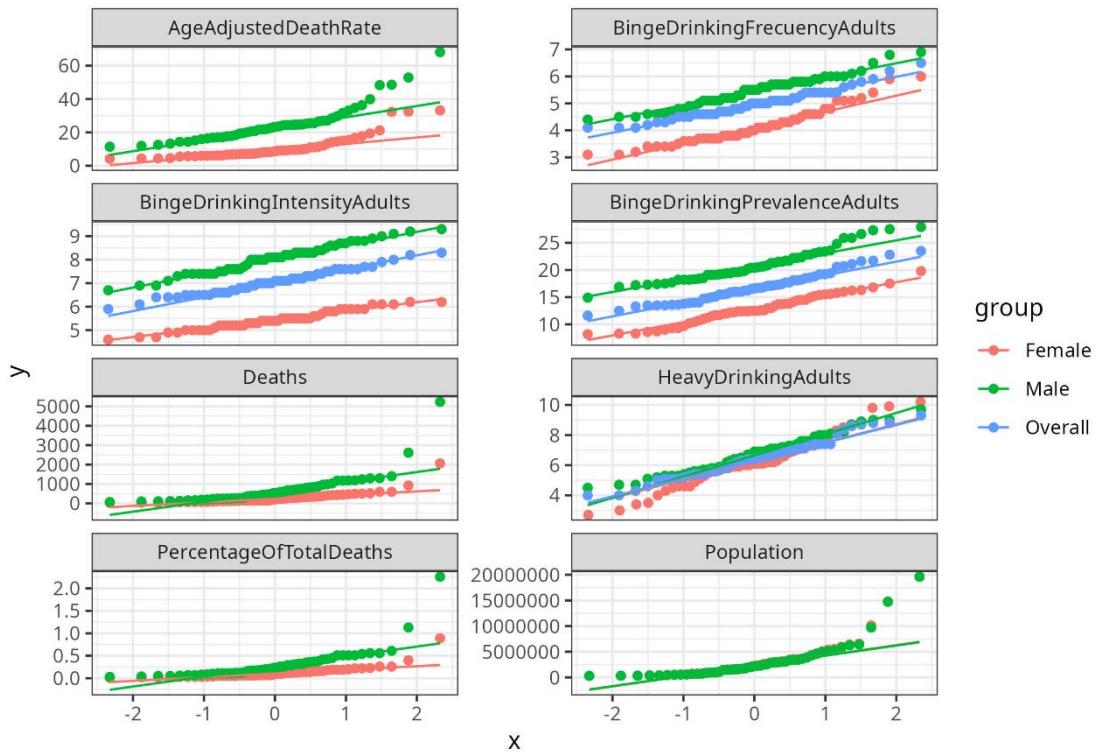
Test gráfico - QQ plot (`DataExplorer::plot_qq()`)

Hay tres variables que, visualmente se ajustan aproximadamente a una distribución normal: `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults` y

`BingeDrinkingPrevalenceAdults`. Son las mismas tres variables con los test de kurtosis y sesgo no significativos.



Cuando estratificamos por la variable sexo, estas variables mantienen su tendencia a la normalidad.



Test de hipótesis (Shapiro-Wilk)

Hay tres variables con un test no significativo, y, por tanto, no es posible rechazar la hipótesis nula de normalidad: HeavyDrinkingAdults, BingeDrinkingFrecuencyAdults y BingeDrinkingPrevalenceAdults.

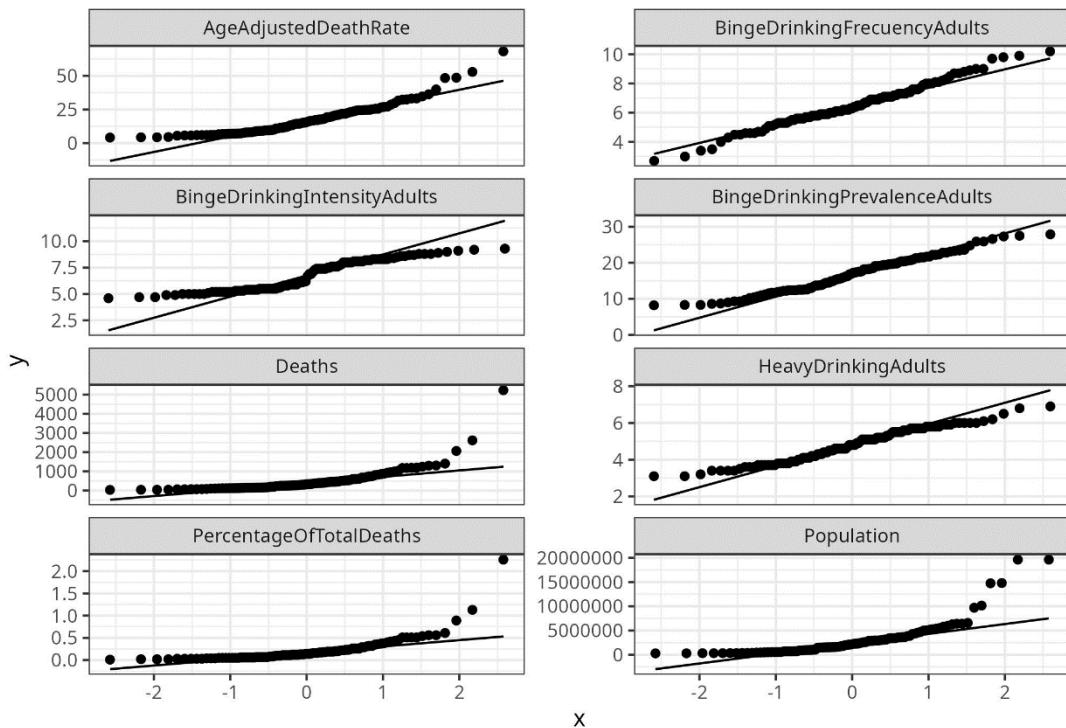
vars	statistic	p_value	sample
Deaths	0.591	0.000	163
Population	0.682	0.000	163
AgeAdjustedDeathRate	0.877	0.000	163
PercentageOfTotalDeaths	0.592	0.000	163
HeavyDrinkingAdults	0.992	0.491	163
BingeDrinkingFrecuencyAdults	0.990	0.288	163
BingeDrinkingIntensityAdults	0.962	0.000	163
BingeDrinkingPrevalenceAdults	0.989	0.232	163

Objeto data_gender

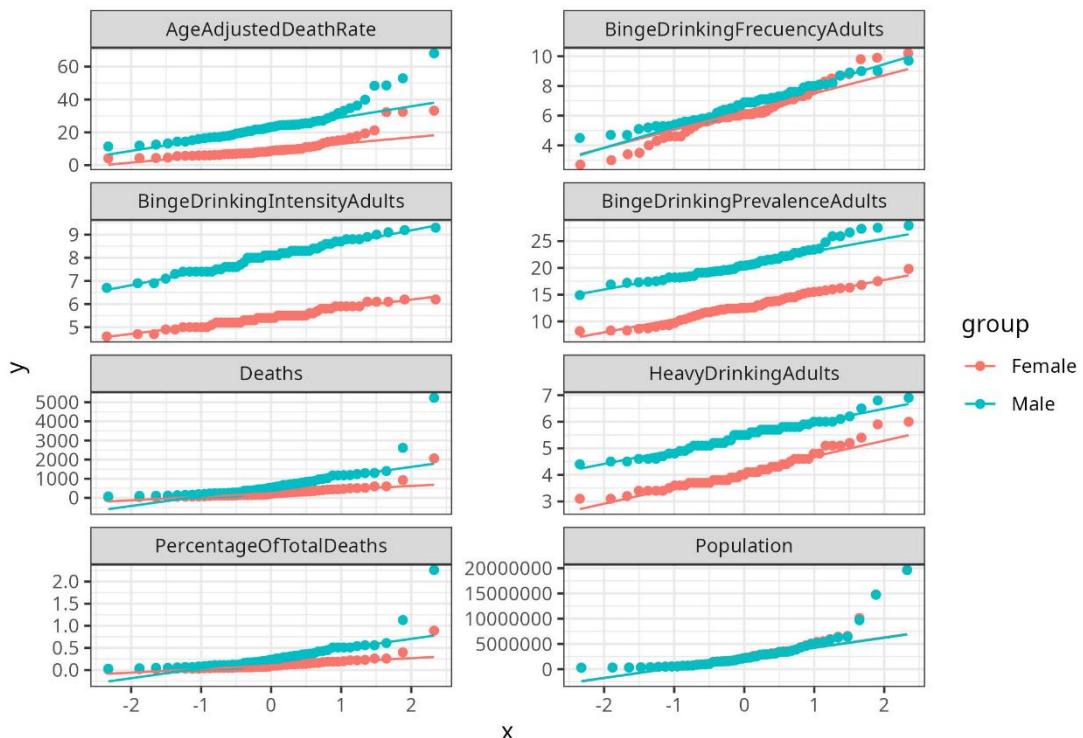
Hay una variable que, visualmente se ajusta aproximadamente a una distribución normal: BingeDrinkingFrecuencyAdults. Hay muchas variables que son aproximadamente normales en el centro de la distribución, pero que no lo son en los extremos (*outliers*).

Test gráfico - QQ plot (DataExplorer::plot_qq())

Hay una variable que, visualmente se ajustan aproximadamente a una distribución normal: BingeDrinkingFrecuencyAdults. Hay muchas variables que son aproximadamente normales en el centro de la distribución, pero que no lo son en los extremos (*outliers*).



Cuando estratificamos por la variable sexo, estas variables mantienen su tendencia a la normalidad, y persiste la influencia de los outliers, especialmente para los varones.



Test de hipótesis (Shapiro-Wilk)

Hay una variable con un test no significativo, y, por tanto, no es posible rechazar la hipótesis nula de normalidad: BingeDrinkingFrecuencyAdults.

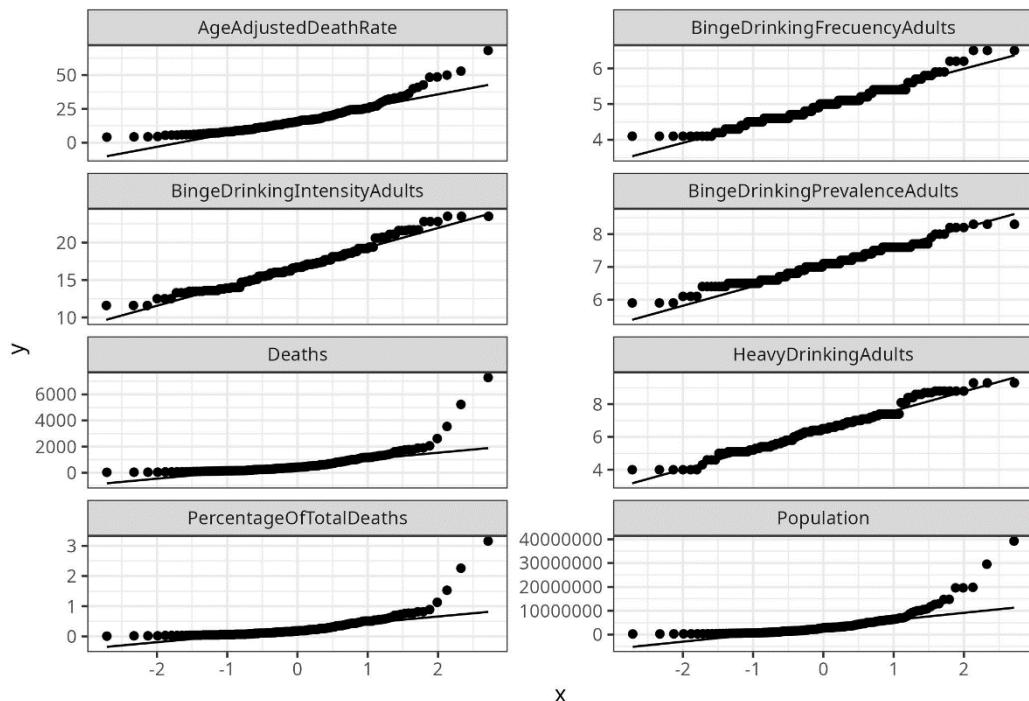
vars	statistic	p_value	sample
Deaths	0.591	0.000	106
Population	0.682	0.000	106
AgeAdjustedDeathRate	0.877	0.000	106
PercentageOfTotalDeaths	0.592	0.000	106
HeavyDrinkingAdults	0.971	0.020	106
BingeDrinkingFrecuencyAdults	0.990	0.622	106
BingeDrinkingIntensityAdults	0.896	0.000	106
BingeDrinkingPrevalenceAdults	0.975	0.040	106

Objeto data_overall

Al considerar los datos globales, ninguna variable se ajusta ni visualmente ni en el test de hipótesis a una distribución normal.

Test gráfico - QQ plot (DataExplorer::plot_qq())

Visualmente, ninguna variable se ajusta a lo esperado en una distribución normal.



Test de hipótesis (Shapiro-Wilk)

Ninguna variable tiene un test no significativo, y, por tanto, no es posible afirmar la hipótesis nula de normalidad.

vars	statistic	p_value	sample
Deaths	0.599	0.000	153
Population	0.644	0.000	153
AgeAdjustedDeathRate	0.875	0.000	153
PercentageOfTotalDeaths	0.599	0.000	153
HeavyDrinkingAdults	0.975	0.007	153
BingeDrinkingFrequencyAdults	0.961	0.000	153
BingeDrinkingIntensityAdults	0.972	0.003	153
BingeDrinkingPrevalenceAdults	0.980	0.023	153

03f - Comparar grupos (*Boxplots, non-parametric tests*)

Para la comparación entre grupos, se utilizaron dos técnicas:

- La valoración gráfica, mediante boxplots
- El test de hipótesis no paramétrico

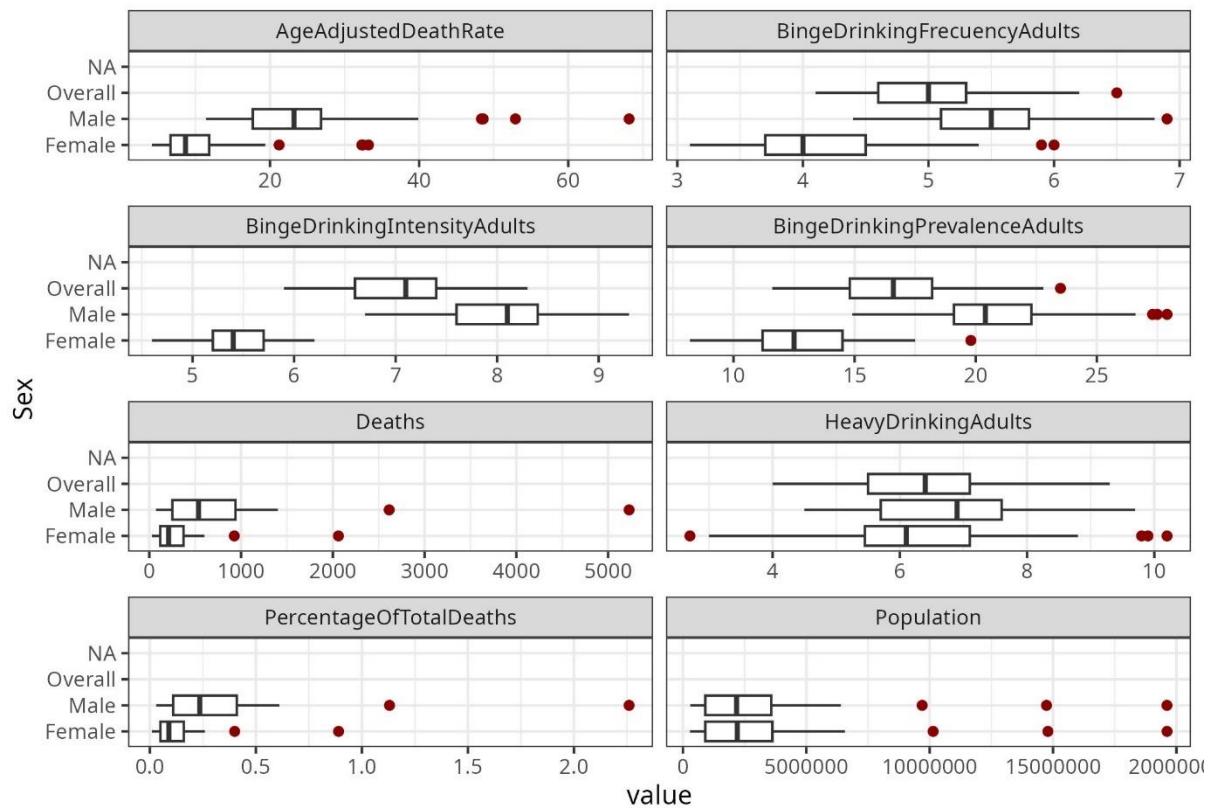
Objeto data

El objeto data tiene la variable Sex, con tres niveles que son redundantes entre sí: Hombres, Mujeres y valoración global.

En la fase exploratoria, esta variable se utilizó para comparar los indicadores entre hombres y mujeres, y entre cada uno de ellos con la media global.

03fa - Valoración gráfica: `DataExplorer::plot_boxplot()`

Se observa una diferencia entre los dos sexos, y entre cada sexo con la media, para las variables, BingeDrinkingFrequencyAdults, BingeDrinkingIntensityAdults y BingeDrinkingPrevalenceAdults. También se ha observado una diferencia entre hombres y mujeres para la variable AgeAdjustedDeathRate.

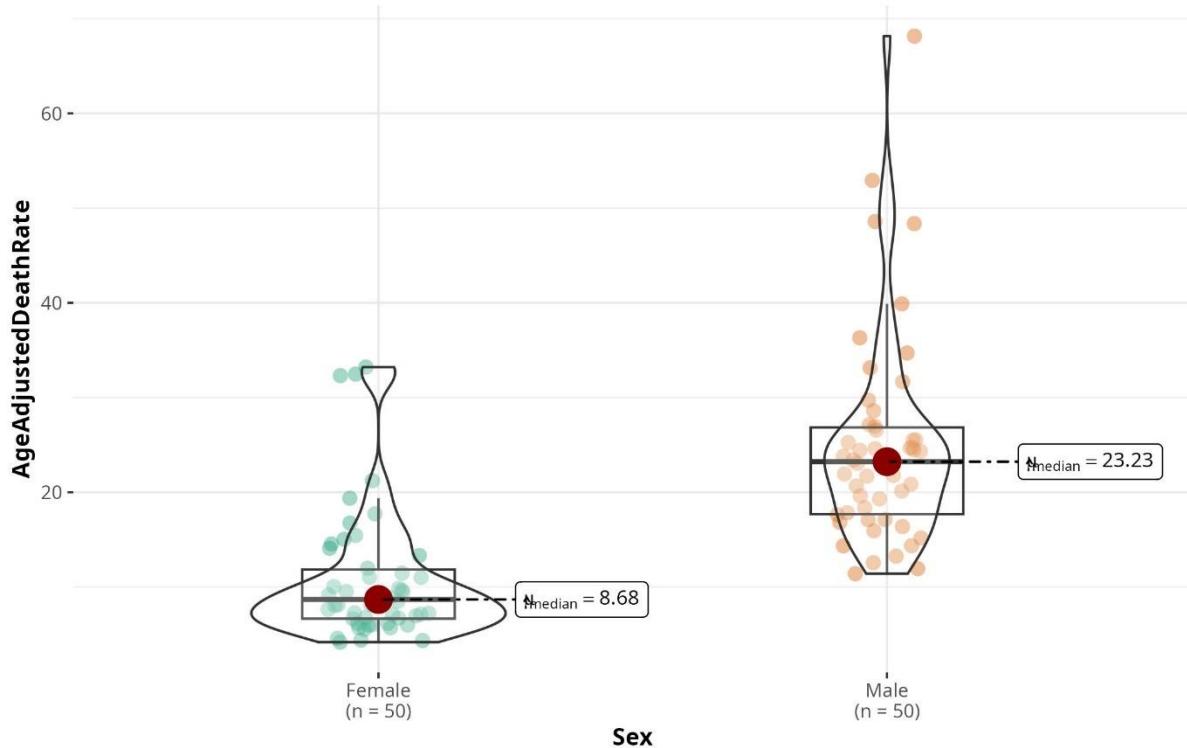


03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`

Se evidenciaron diferencias estadísticamente significativas entre hombres y mujeres, y entre cada uno de ellos con la media general, para las variables `BingeDrinkingFrequencyAdults`, `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults`. También se ha evidenciado una diferencia significativa entre hombres y mujeres para la variable `AgeAdjustedDeathRate`.

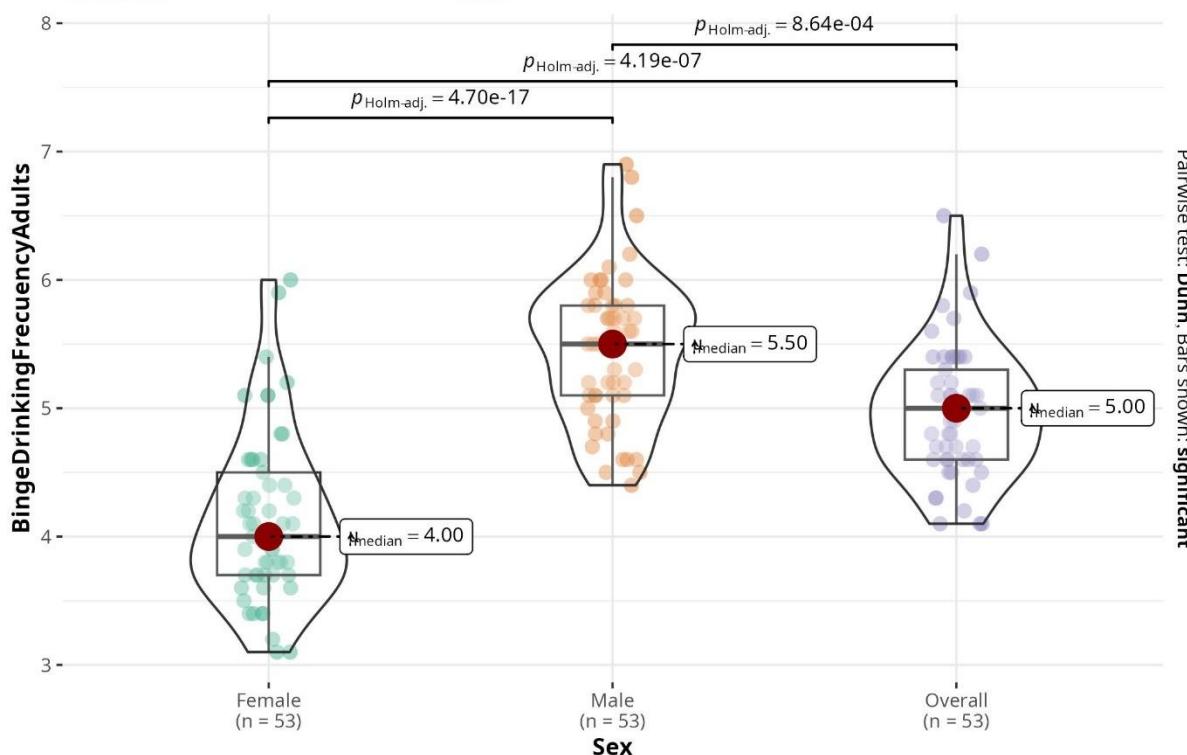
Age Adjusted Death Rate, by sex

$W_{\text{Mann-Whitney}} = 215.00, p = 9.91e-13, \hat{\rho}_{\text{biserial}}^{\text{rank}} = -0.83, \text{CI}_{95\%} [-0.89, -0.74], n_{\text{obs}} = 100$

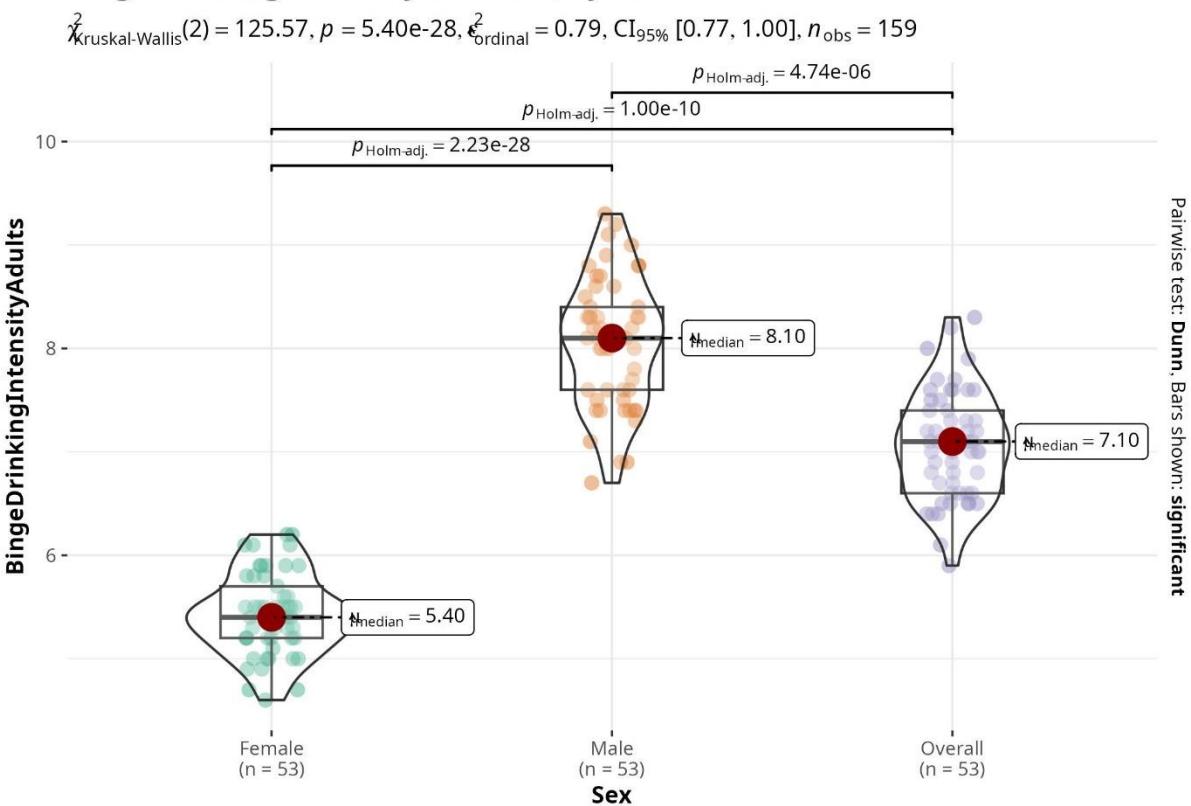


Binge Drinking Frequency in Adults, by sex

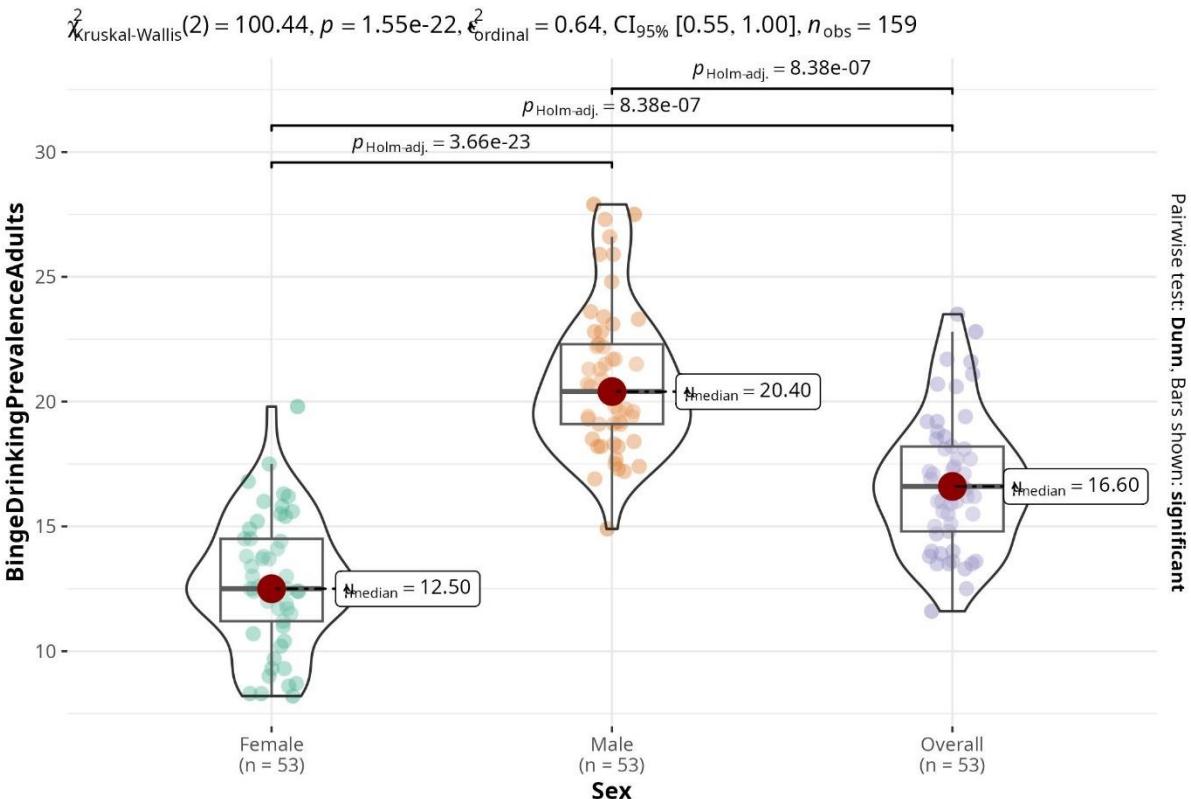
$\chi^2_{\text{Kruskal-Wallis}}(2) = 73.78, p = 9.52e-17, \hat{\rho}_{\text{ordinal}}^{\text{rank}} = 0.47, \text{CI}_{95\%} [0.36, 1.00], n_{\text{obs}} = 159$



Binge Drinking Intensity in Adults, by sex

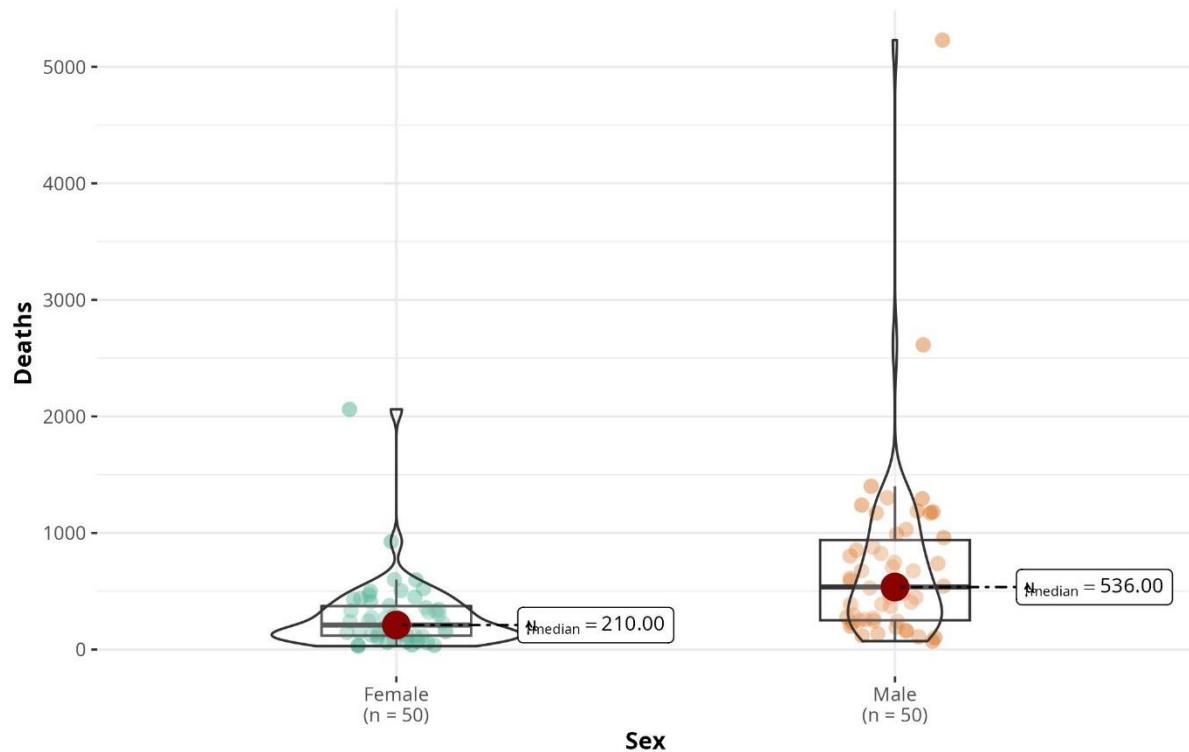


Binge Drinking Prevalence in Adults, by sex



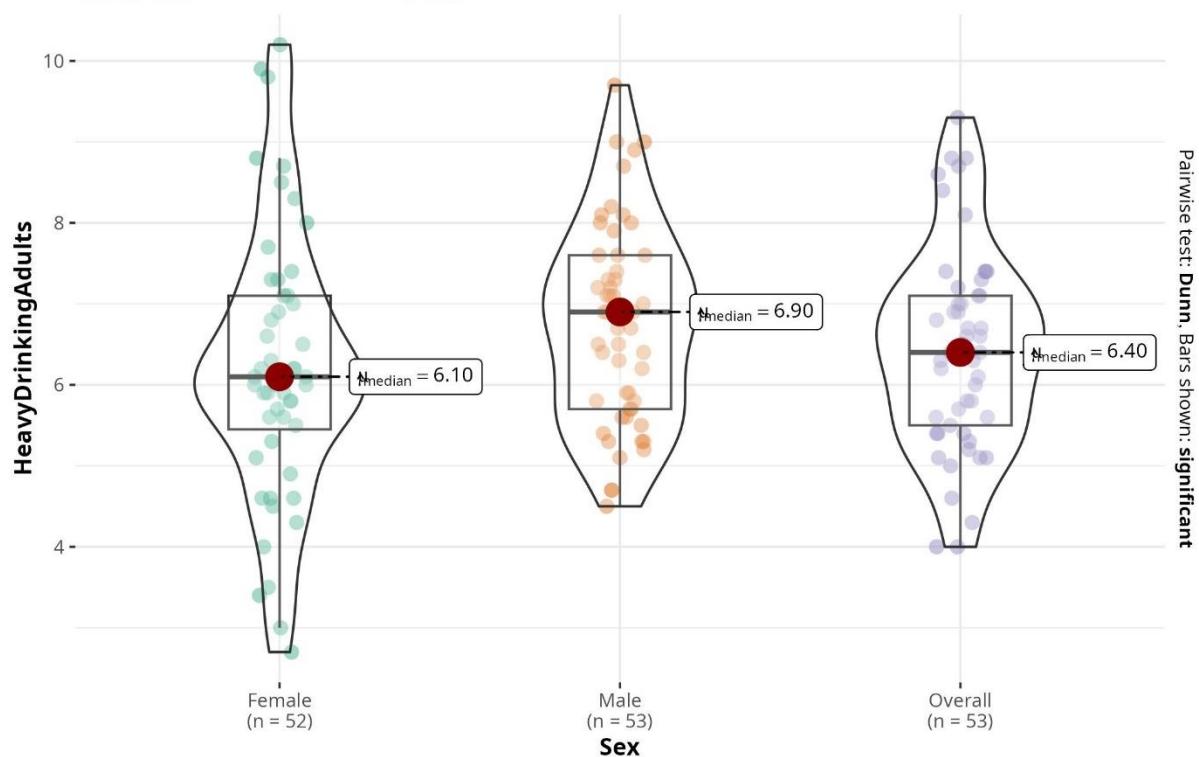
Deaths, by sex

$W_{\text{Mann-Whitney}} = 585.50, p = 4.70e-06, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.53, \text{CI}_{95\%} [-0.67, -0.35], n_{\text{obs}} = 100$



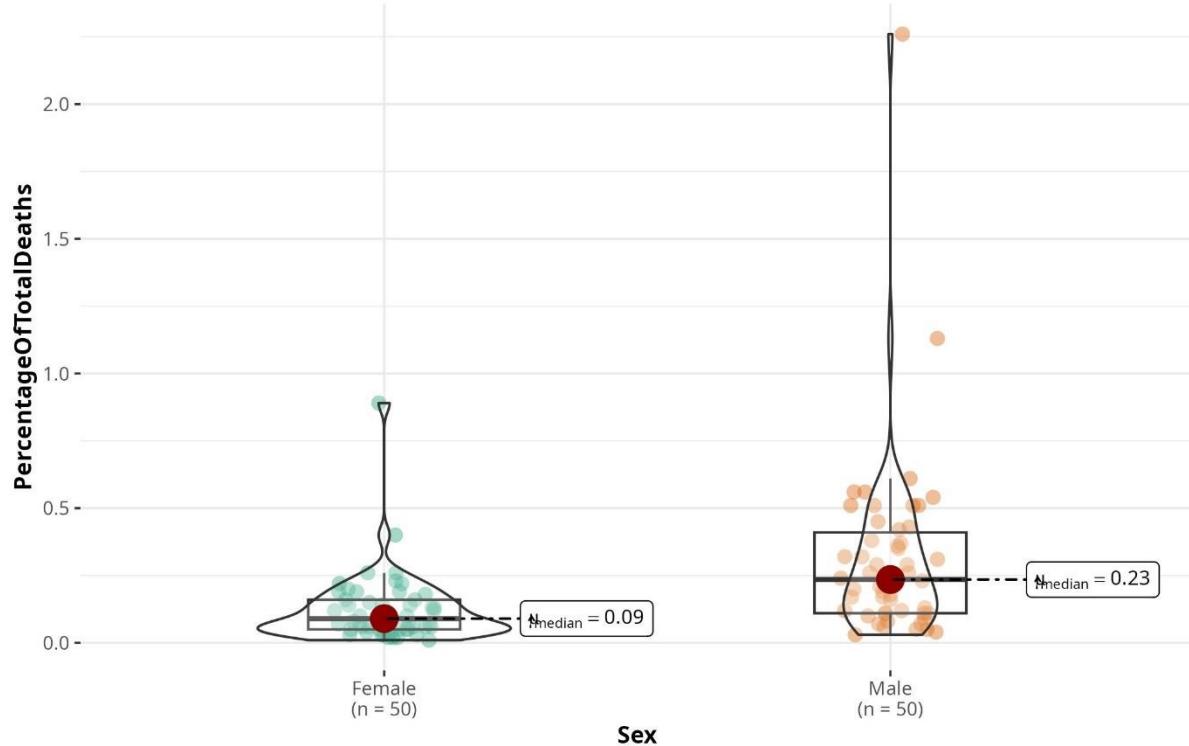
Heavy Drinking in Adults, by sex

$\chi^2_{\text{Kruskal-Wallis}}(2) = 3.93, p = 0.14, \epsilon^2_{\text{ordinal}} = 0.03, \text{CI}_{95\%} [7.73e-03, 1.00], n_{\text{obs}} = 158$



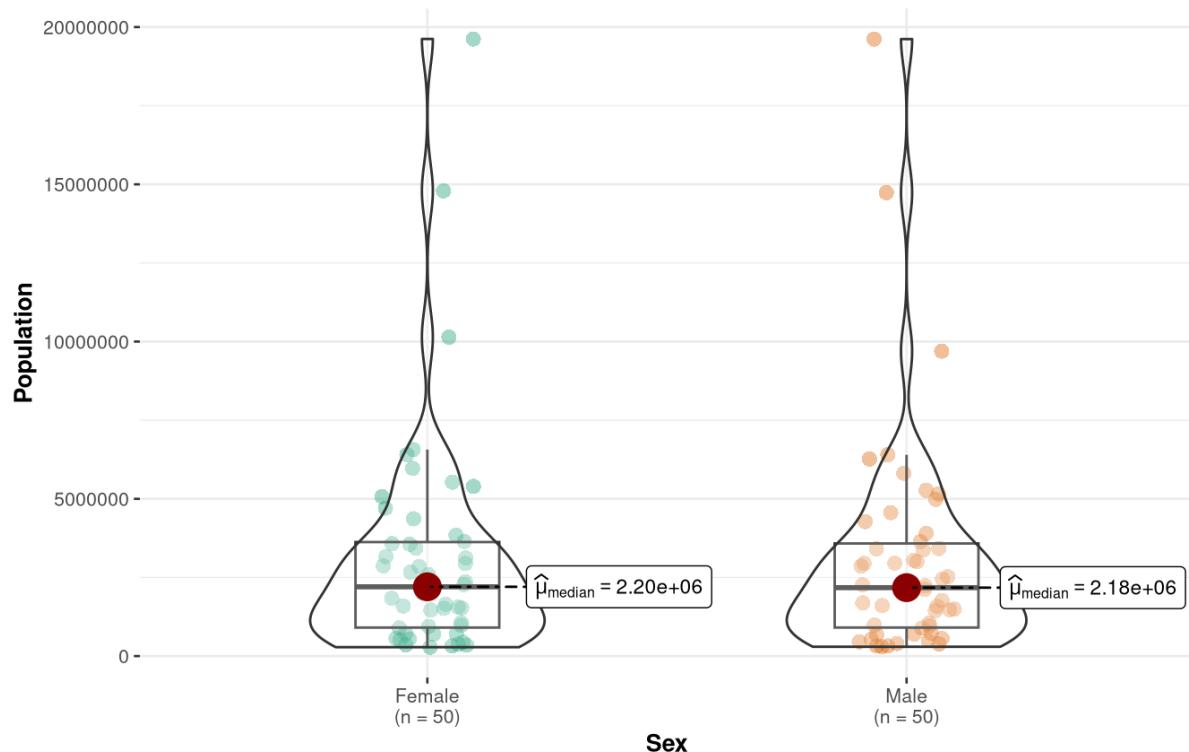
Percentage of Total Deaths, by sex

$W_{\text{Mann-Whitney}} = 590.00, p = 5.37e-06, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.53, \text{CI}_{95\%} [-0.67, -0.35], n_{\text{obs}} = 100$



Population, by sex

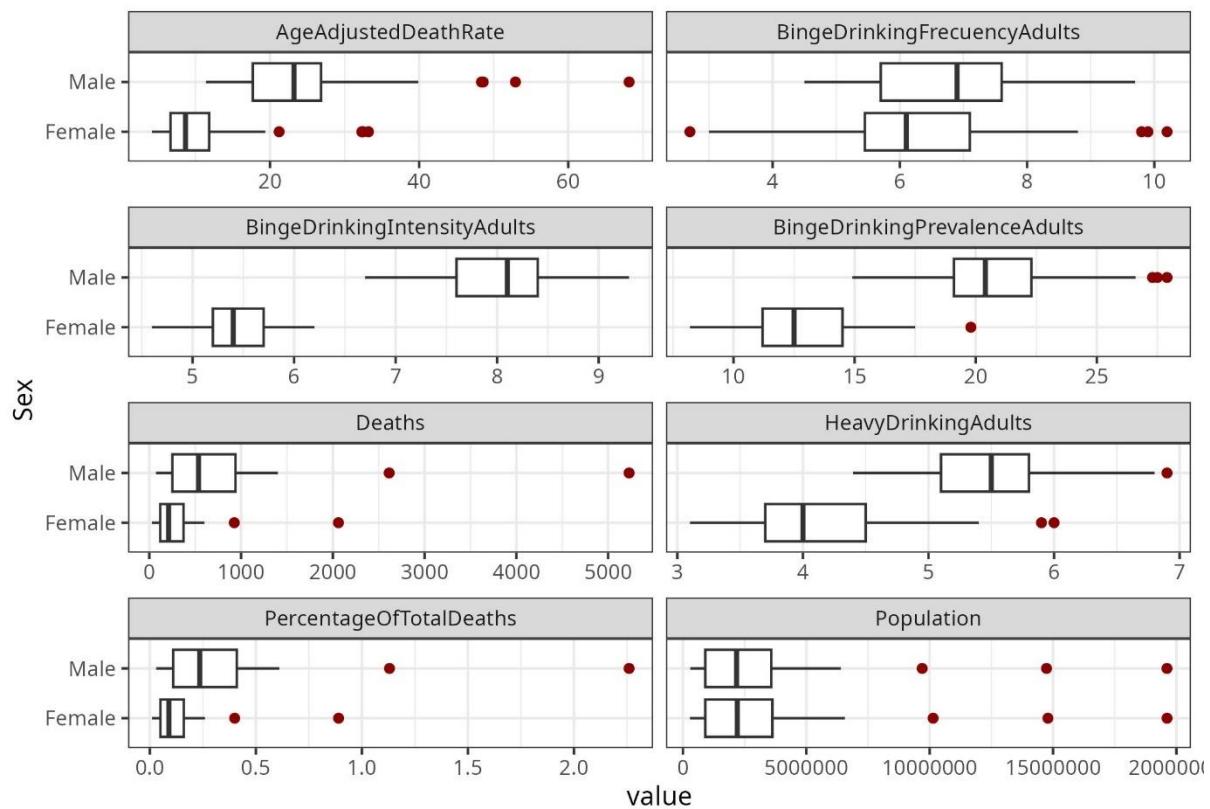
$W_{\text{Mann-Whitney}} = 1269.00, p = 0.90, \hat{r}_{\text{biserial}}^{\text{rank}} = 0.02, \text{CI}_{95\%} [-0.21, 0.24], n_{\text{obs}} = 100$



Objeto `data_gender`

03fa - Valoración gráfica: `DataExplorer::plot_boxplot()`

Se observa una diferencia entre los dos性 para las variables, HeavyDrinkingAdults, AgeAdjustedDeathRate, BingeDrinkingPrevalenceAdults y BingeDrinkingIntensityAdults.

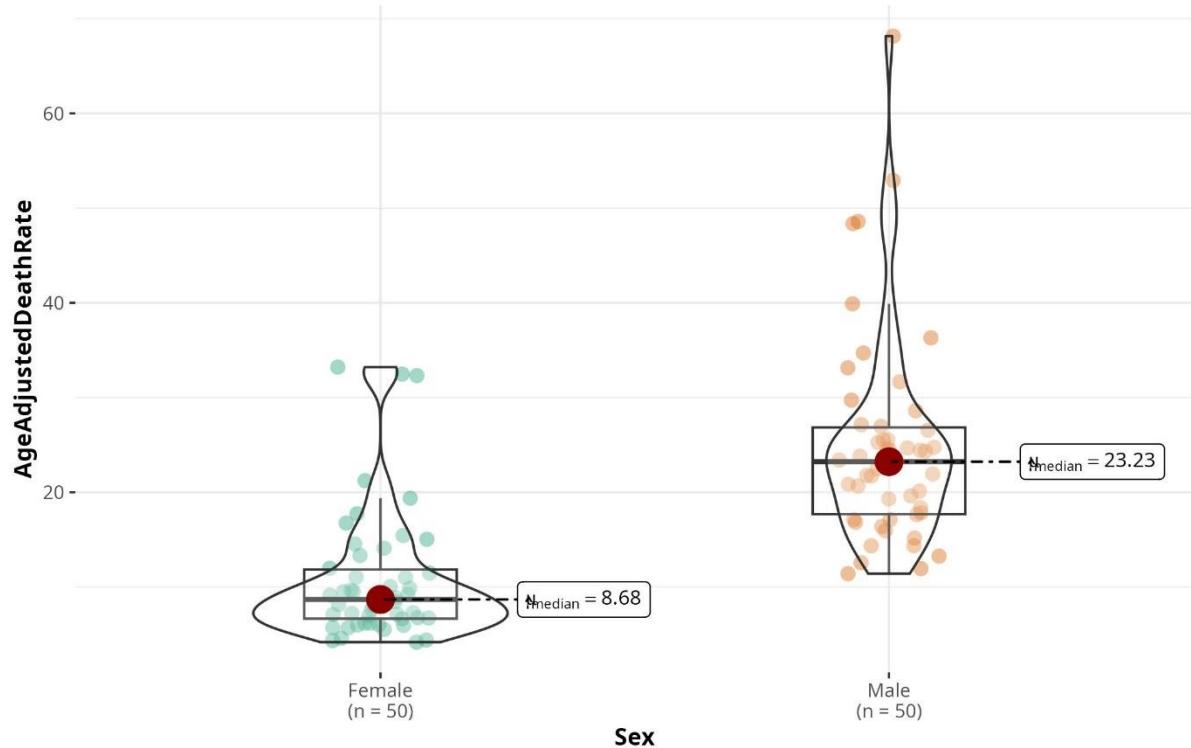


03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`

Se evidenciaron diferencias estadísticamente significativas entre hombres y mujeres, y entre cada uno de ellos con la media general, para las variables BingeDrinkingFrecuencyAdults, BingeDrinkingIntensityAdults y BingeDrinkingPrevalenceAdults. También se ha evidenciado una diferencia significativa entre hombres y mujeres para la variable AgeAdjustedDeathRate.

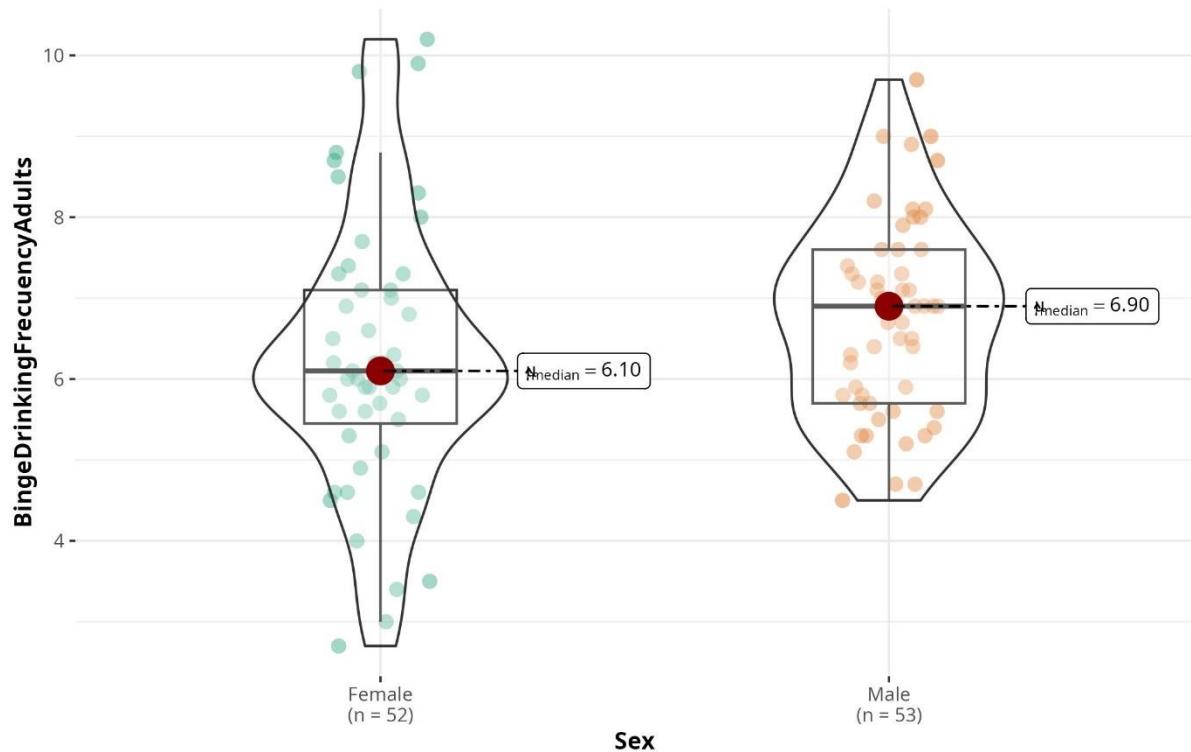
Age Adjusted Death Rate, by sex

$W_{\text{Mann-Whitney}} = 215.00, p = 9.91e-13, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.83, \text{CI}_{95\%} [-0.89, -0.74], n_{\text{obs}} = 100$



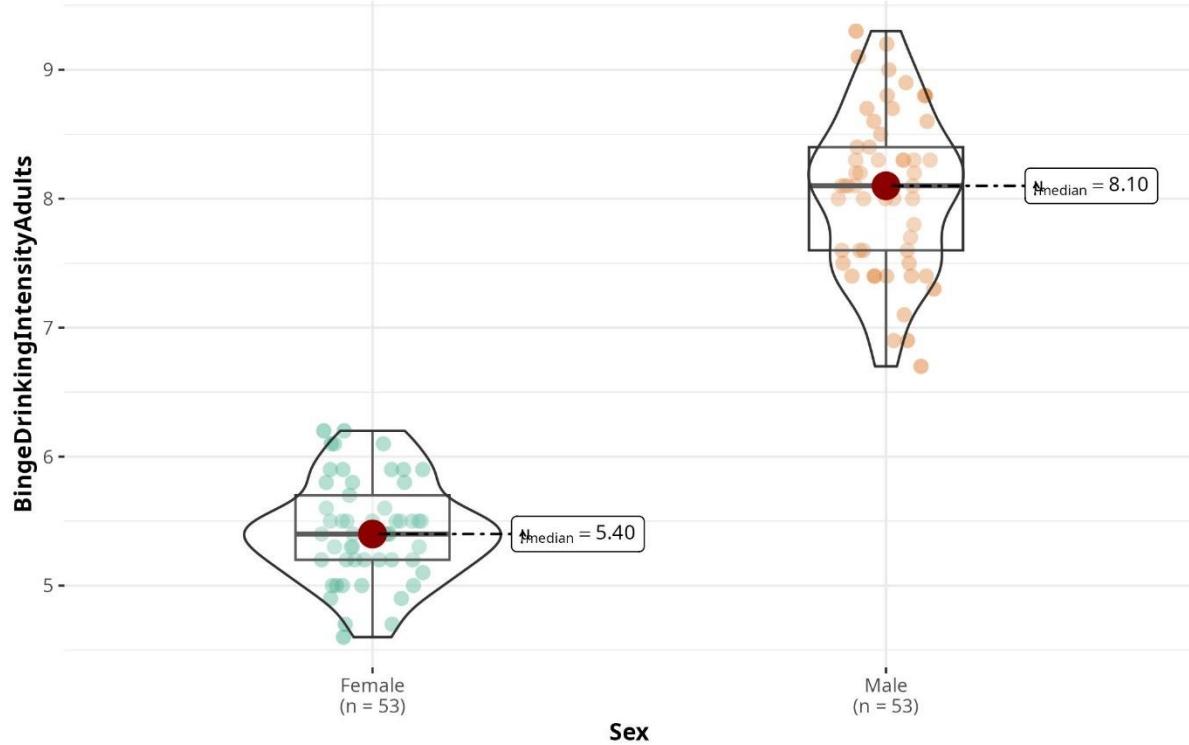
Binge Drinking Frequency in Adults, by sex

$W_{\text{Mann-Whitney}} = 1090.50, p = 0.07, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.21, \text{CI}_{95\%} [-0.41, 0.01], n_{\text{obs}} = 105$



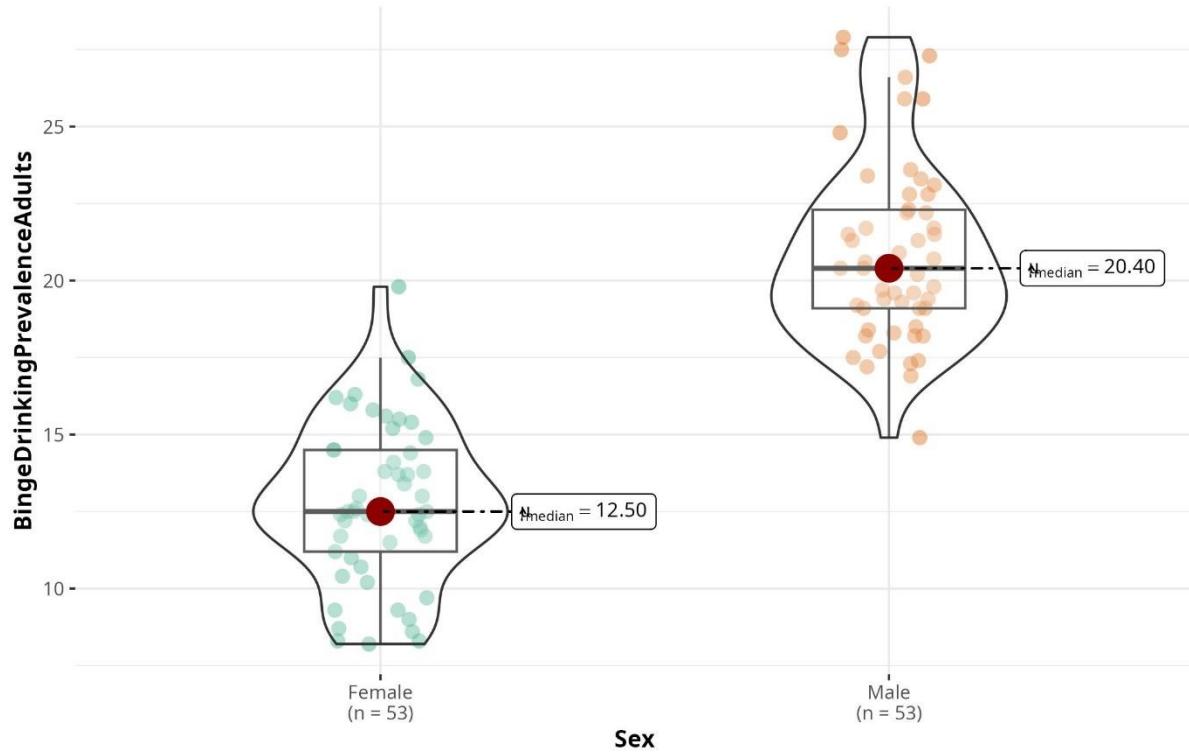
Binge Drinking Intensity in Adults, by sex

$W_{\text{Mann-Whitney}} = 0.00, p = 6.66e-19, \hat{r}_{\text{biserial}}^{\text{rank}} = -1.00, \text{CI}_{95\%} [-1.00, -1.00], n_{\text{obs}} = 106$



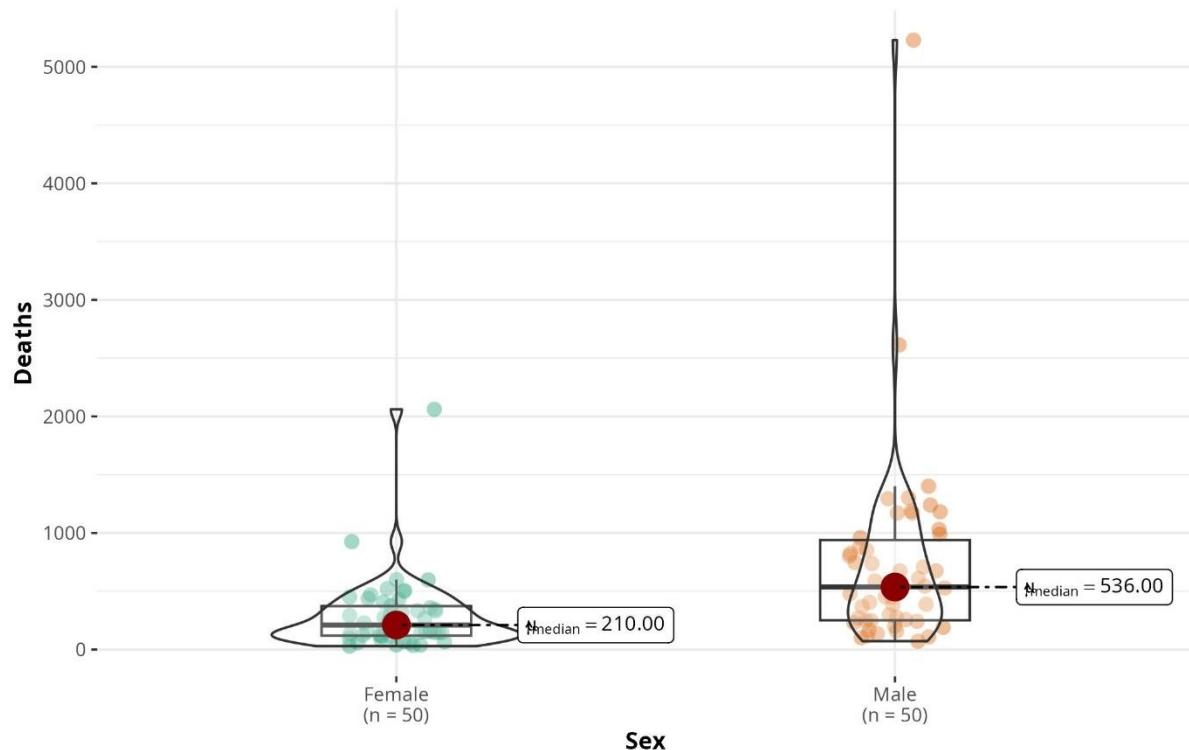
Binge Drinking Prevalence in Adults, by sex

$W_{\text{Mann-Whitney}} = 38.50, p = 6.19e-18, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.97, \text{CI}_{95\%} [-0.98, -0.96], n_{\text{obs}} = 106$



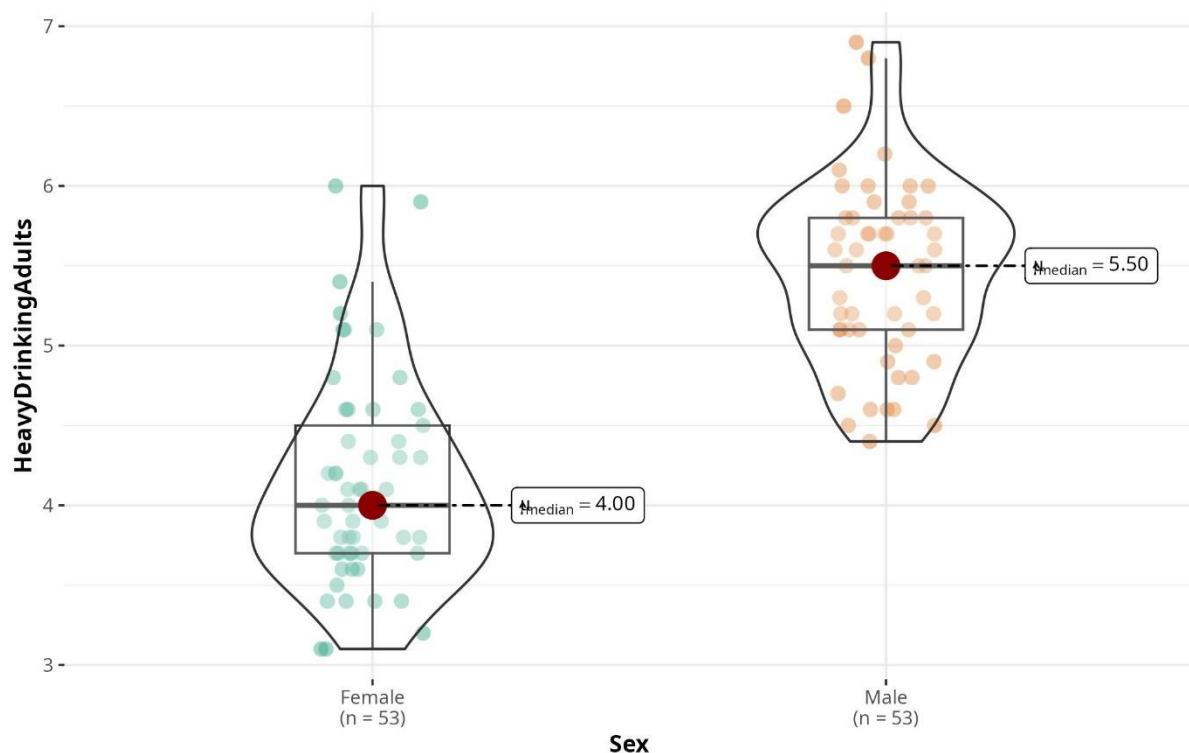
Deaths, by sex

$W_{\text{Mann-Whitney}} = 585.50, p = 4.70e-06, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.53, \text{CI}_{95\%} [-0.67, -0.35], n_{\text{obs}} = 100$



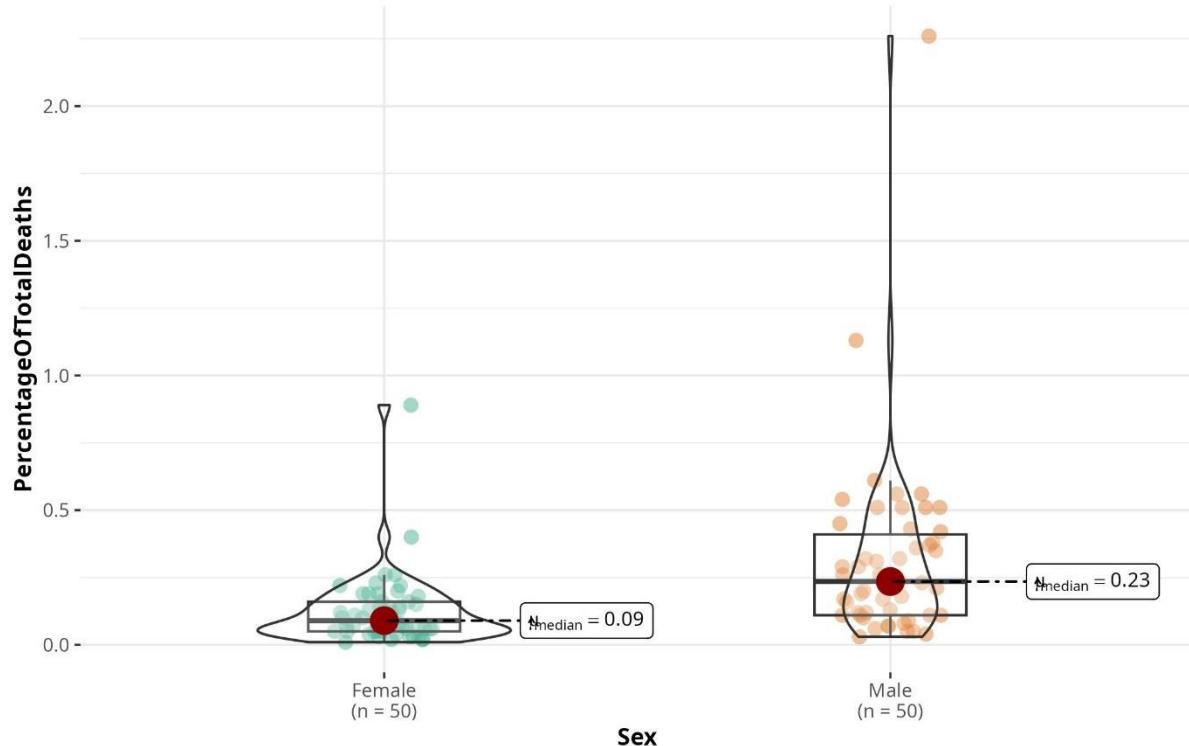
Heavy Drinking in Adults, by sex

$W_{\text{Mann-Whitney}} = 211.50, p = 4.62e-14, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.85, \text{CI}_{95\%} [-0.90, -0.78], n_{\text{obs}} = 106$



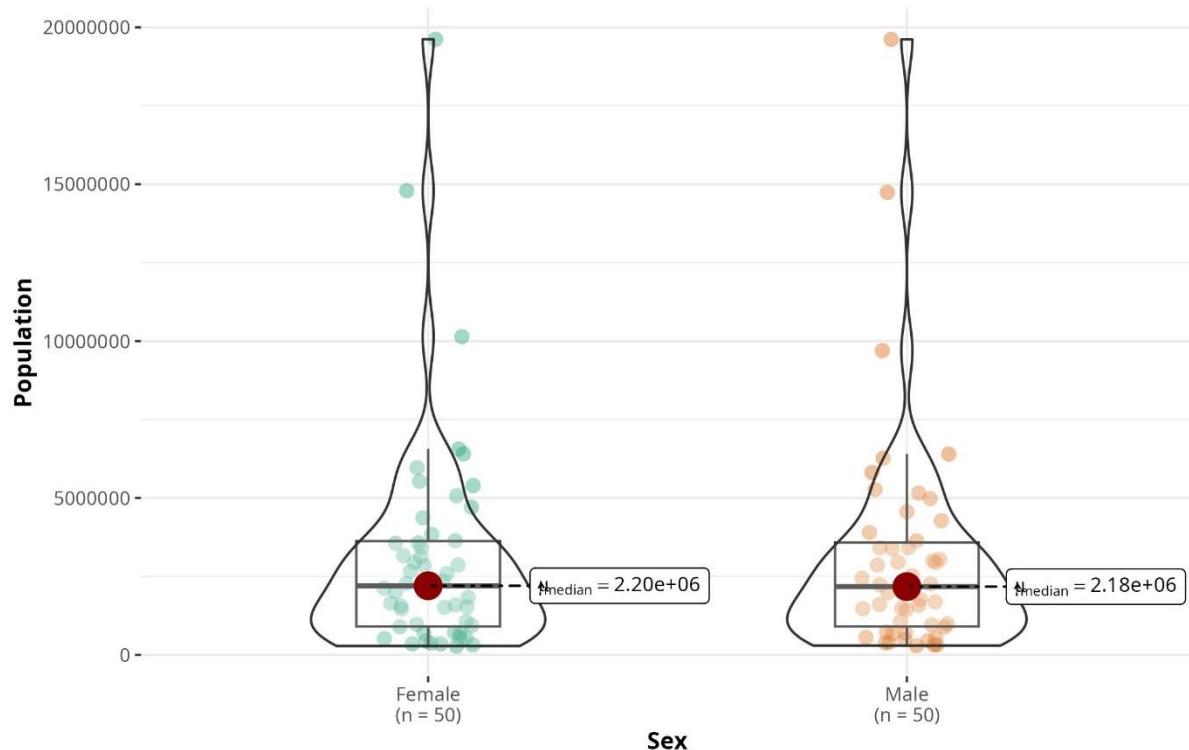
Percentage of Total Deaths, by sex

$W_{\text{Mann-Whitney}} = 590.00, p = 5.37e-06, \hat{r}_{\text{biserial}}^{\text{rank}} = -0.53, \text{CI}_{95\%} [-0.67, -0.35], n_{\text{obs}} = 100$



Population, by sex

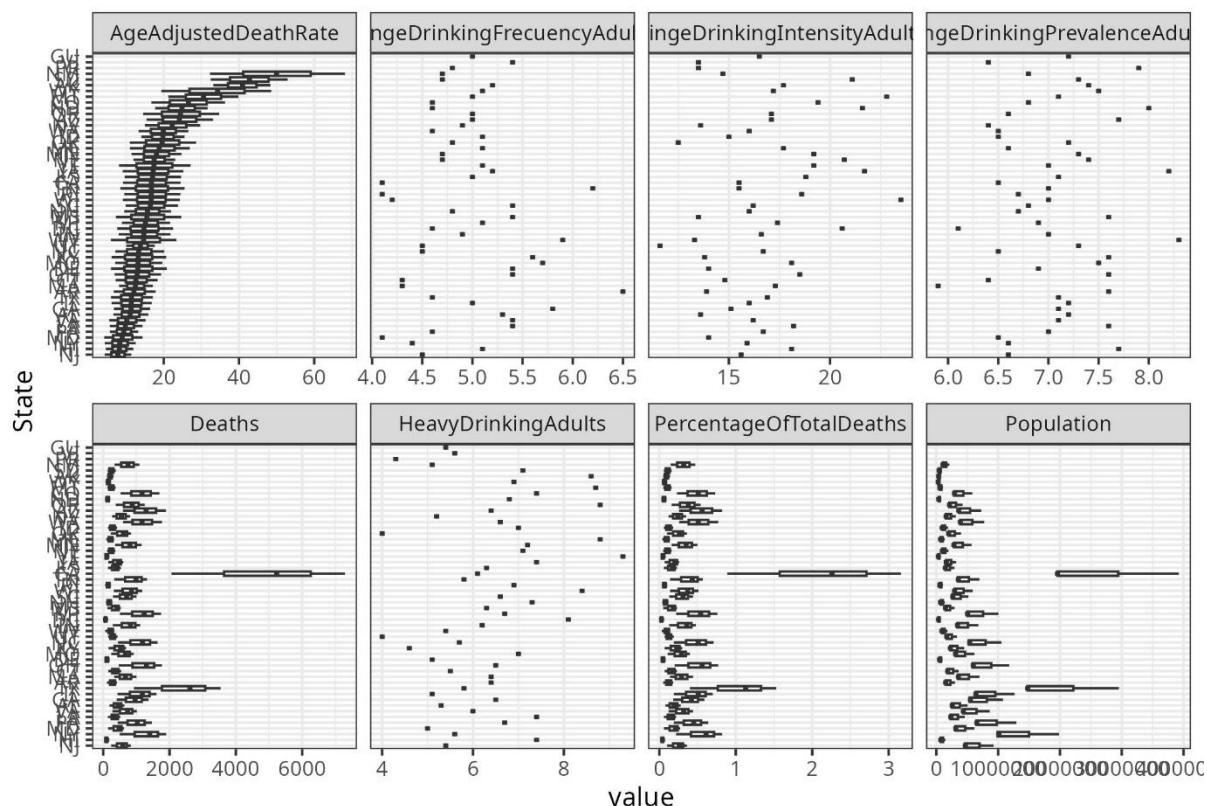
$W_{\text{Mann-Whitney}} = 1269.00, p = 0.90, \hat{r}_{\text{biserial}}^{\text{rank}} = 0.02, \text{CI}_{95\%} [-0.21, 0.24], n_{\text{obs}} = 100$



Objeto data_overall

03fa - Valoración gráfica: `DataExplorer::plot_boxplot()`

Se observan diferencias en el valor de los indicadores para todos los estados, para las distintas variables, identificándose estados con valores en torno a la media, y otros con valor muy superior.



03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`

En este dataset no hay observaciones suficientes para hacer un análisis comparativo de las variables numéricas por los niveles de la variable State.

03g - Explorar correlaciones

Se utilizaron las siguientes técnicas para explorar la correlación:

- 03ga - Matriz de correlación, mediante el test de Spearman
- 03gb - Correlograma, para visualizar la fuerza, la significación estadística y la dirección de la correlación
- 03gc - El test de hipótesis estadístico para la correlación

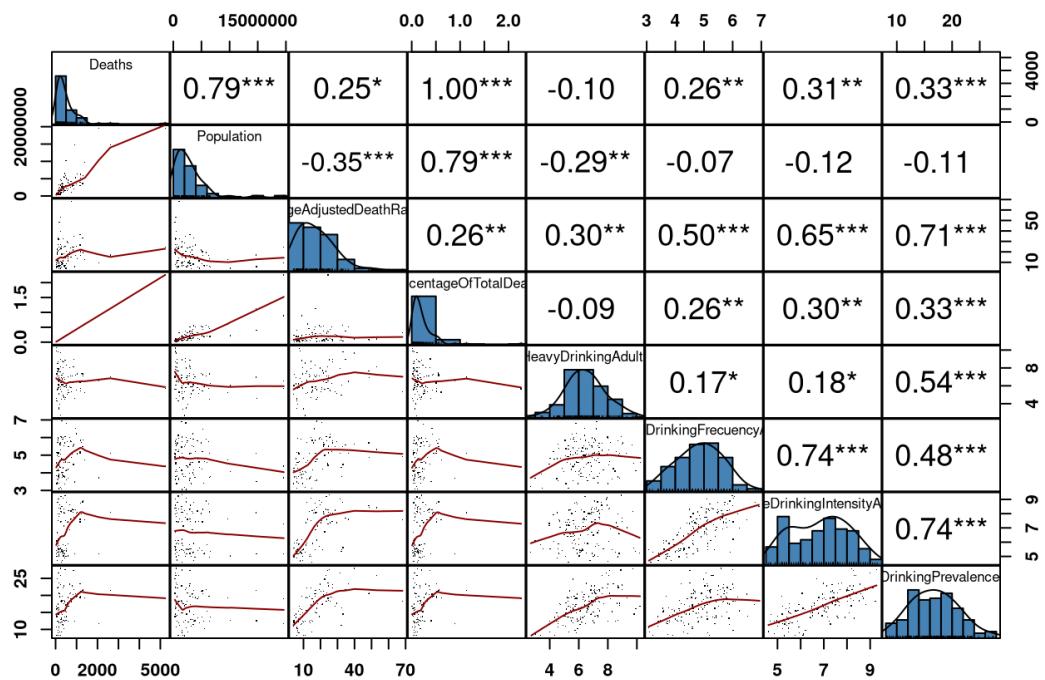
Objeto data

03ga - Correlation matrix

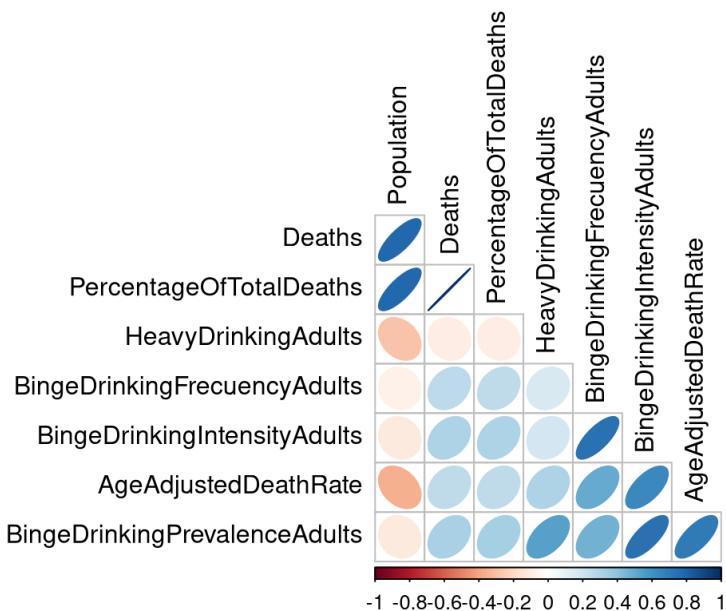
Se exploraron las correlaciones entre variables numéricas con el test de Spearman. Se obtuvieron los siguientes resultados (en verde, las correlaciones estadísticamente significativas):

row	column	cor	p
Deaths	Population	0.79	0.00
Deaths	AgeAdjustedDeathRate	0.25	0.01
Population	AgeAdjustedDeathRate	-0.35	0.00
Deaths	PercentageOfTotalDeaths	1.00	0.00
Population	PercentageOfTotalDeaths	0.79	0.00
AgeAdjustedDeathRate	PercentageOfTotalDeaths	0.26	0.01
Deaths	HeavyDrinkingAdults	-0.10	0.33
Population	HeavyDrinkingAdults	-0.29	0.00
AgeAdjustedDeathRate	HeavyDrinkingAdults	0.30	0.00
PercentageOfTotalDeaths	HeavyDrinkingAdults	-0.09	0.36
Deaths	BingeDrinkingFrequencyAdults	0.26	0.01
Population	BingeDrinkingFrequencyAdults	-0.07	0.48
AgeAdjustedDeathRate	BingeDrinkingFrequencyAdults	0.50	0.00
PercentageOfTotalDeaths	BingeDrinkingFrequencyAdults	0.26	0.01
HeavyDrinkingAdults	BingeDrinkingFrequencyAdults	0.17	0.04
Deaths	BingeDrinkingIntensityAdults	0.31	0.00
Population	BingeDrinkingIntensityAdults	-0.12	0.24
AgeAdjustedDeathRate	BingeDrinkingIntensityAdults	0.65	0.00
PercentageOfTotalDeaths	BingeDrinkingIntensityAdults	0.30	0.00
HeavyDrinkingAdults	BingeDrinkingIntensityAdults	0.18	0.02
BingeDrinkingFrequencyAdults	BingeDrinkingIntensityAdults	0.74	0.00
Deaths	BingeDrinkingPrevalenceAdults	0.33	0.00
Population	BingeDrinkingPrevalenceAdults	-0.11	0.27
AgeAdjustedDeathRate	BingeDrinkingPrevalenceAdults	0.71	0.00
PercentageOfTotalDeaths	BingeDrinkingPrevalenceAdults	0.33	0.00
HeavyDrinkingAdults	BingeDrinkingPrevalenceAdults	0.54	0.00
BingeDrinkingFrequencyAdults	BingeDrinkingPrevalenceAdults	0.48	0.00
BingeDrinkingIntensityAdults	BingeDrinkingPrevalenceAdults	0.74	0.00

Correlation matrix



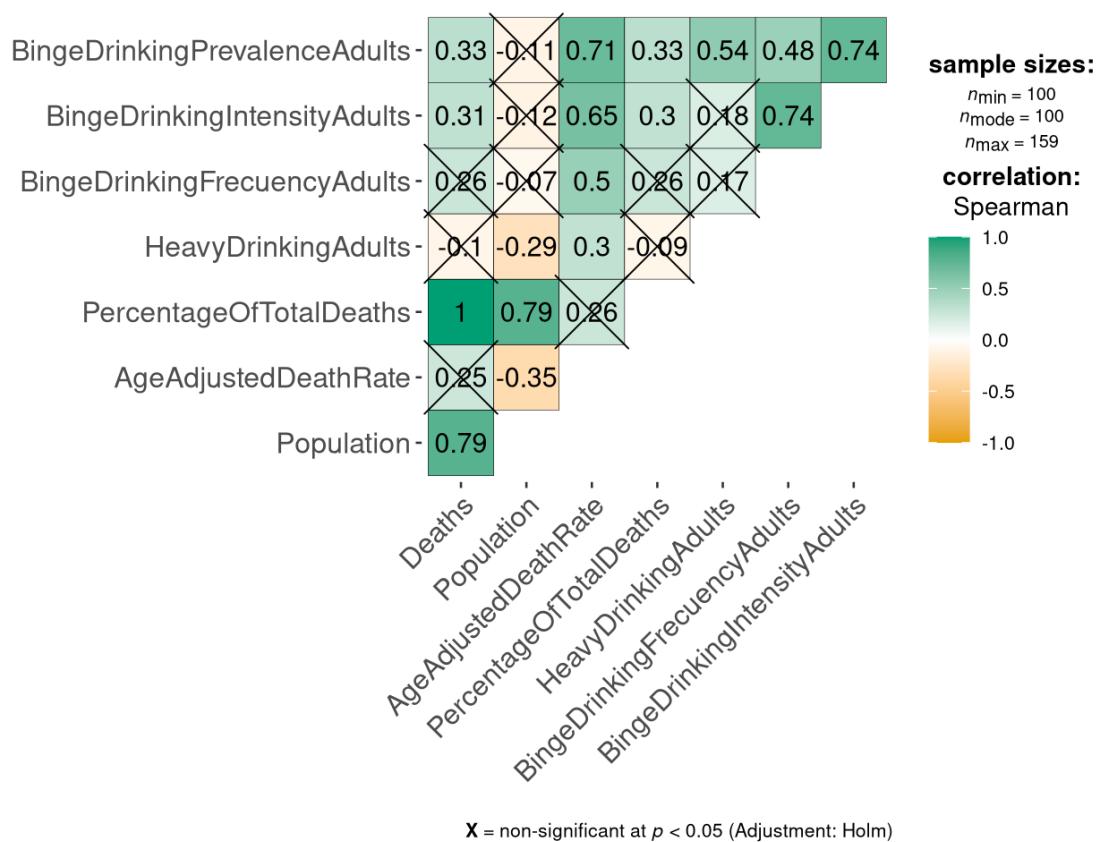
03gb - Correlograma (visualización de la correlación)



03gc - Test de hipótesis ggstatsplot::ggcorrmat()

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre Population y AgeAdjustedDeathRate
- Una correlación positiva entre:
 - Deaths y Population
 - Deaths y PercentageOfTotalDeaths
 - Population y PercentageOfTotalDeaths
 - HeavyDrinkingAdults y AgeAdjustedDeathRate
 - BingeDrinkingPrevalenceAdults y HeavyDrinkingAdults
 - BingeDrinkingPrevalenceAdults y BingeDrinkingFrecuencyAdults
 - BingeDrinkingPrevalenceAdults y BingeDrinkingIntensityAdults
 - BingeDrinkingIntensityAdults y BingeDrinkingFrecuencyAdults



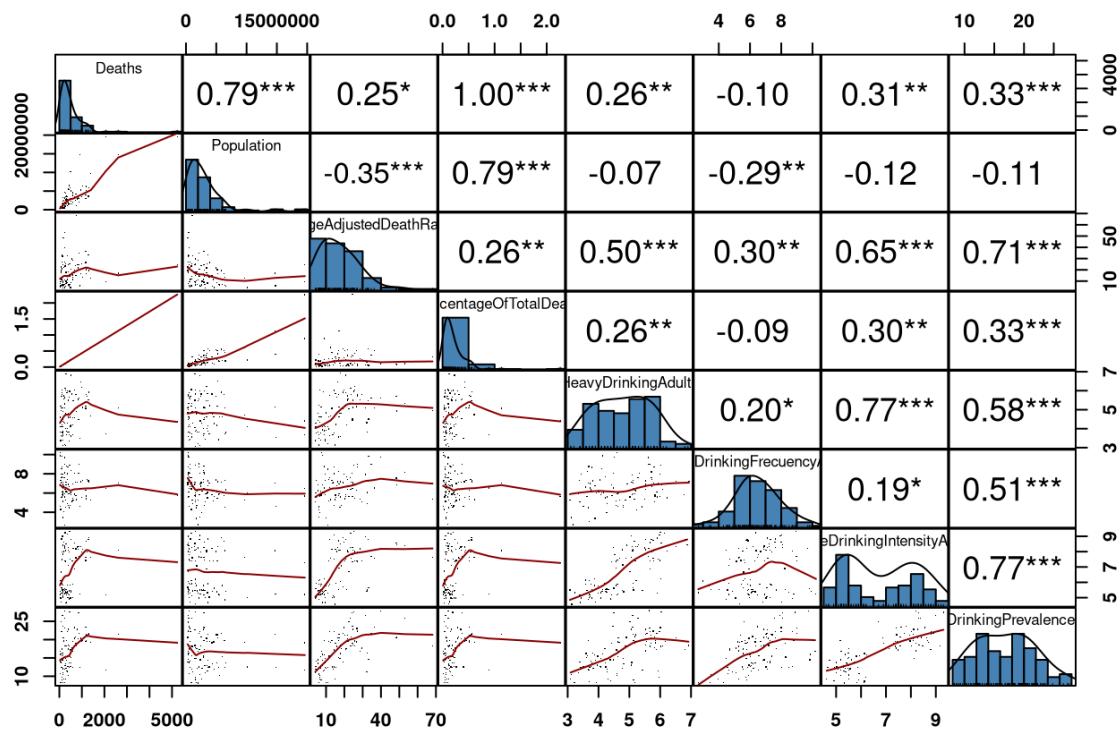
Objeto data_gender

03ga - Correlation matrix

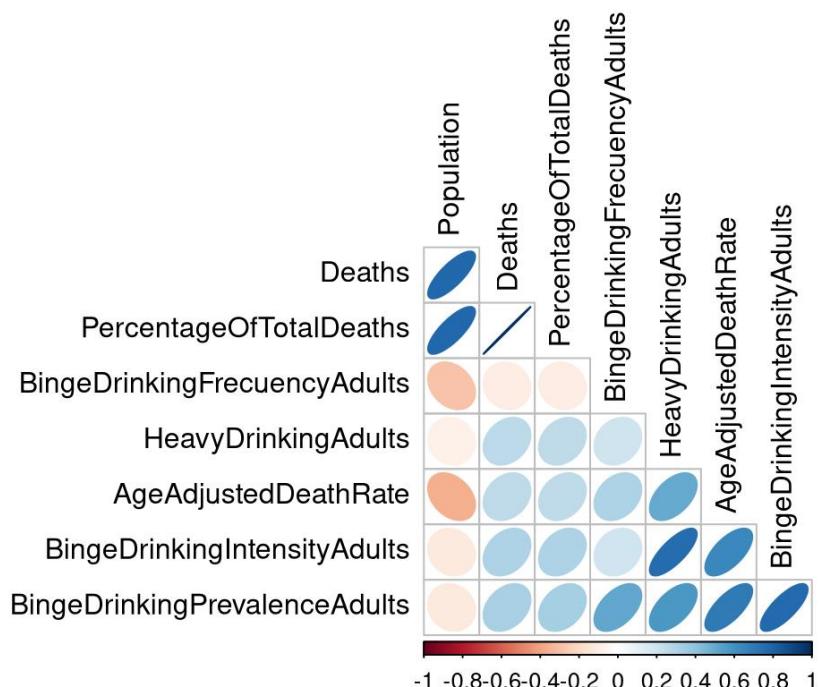
Se exploraron las correlaciones entre variables numéricas con el test de Spearman. Se obtuvieron los siguientes resultados (en verde, las correlaciones estadísticamente significativas):

row	column	cor	p
Deaths	Population	0.79	0.00
Deaths	AgeAdjustedDeathRate	0.25	0.01
Population	AgeAdjustedDeathRate	-0.35	0.00
Deaths	PercentageOfTotalDeaths	1.00	0.00
Population	PercentageOfTotalDeaths	0.79	0.00
AgeAdjustedDeathRate	PercentageOfTotalDeaths	0.26	0.01
Deaths	HeavyDrinkingAdults	0.26	0.01
Population	HeavyDrinkingAdults	-0.07	0.48
AgeAdjustedDeathRate	HeavyDrinkingAdults	0.50	0.00
PercentageOfTotalDeaths	HeavyDrinkingAdults	0.26	0.01
Deaths	BingeDrinkingFrequencyAdults	-0.10	0.33
Population	BingeDrinkingFrequencyAdults	-0.29	0.00
AgeAdjustedDeathRate	BingeDrinkingFrequencyAdults	0.30	0.00
PercentageOfTotalDeaths	BingeDrinkingFrequencyAdults	-0.09	0.36
HeavyDrinkingAdults	BingeDrinkingFrequencyAdults	0.20	0.04
Deaths	BingeDrinkingIntensityAdults	0.31	0.00
Population	BingeDrinkingIntensityAdults	-0.12	0.24
AgeAdjustedDeathRate	BingeDrinkingIntensityAdults	0.65	0.00
PercentageOfTotalDeaths	BingeDrinkingIntensityAdults	0.30	0.00
HeavyDrinkingAdults	BingeDrinkingIntensityAdults	0.77	0.00
BingeDrinkingFrequencyAdults	BingeDrinkingIntensityAdults	0.19	0.05
Deaths	BingeDrinkingPrevalenceAdults	0.33	0.00
Population	BingeDrinkingPrevalenceAdults	-0.11	0.27
AgeAdjustedDeathRate	BingeDrinkingPrevalenceAdults	0.71	0.00
PercentageOfTotalDeaths	BingeDrinkingPrevalenceAdults	0.33	0.00
HeavyDrinkingAdults	BingeDrinkingPrevalenceAdults	0.58	0.00
BingeDrinkingFrequencyAdults	BingeDrinkingPrevalenceAdults	0.51	0.00
BingeDrinkingIntensityAdults	BingeDrinkingPrevalenceAdults	0.77	0.00

Correlation matrix



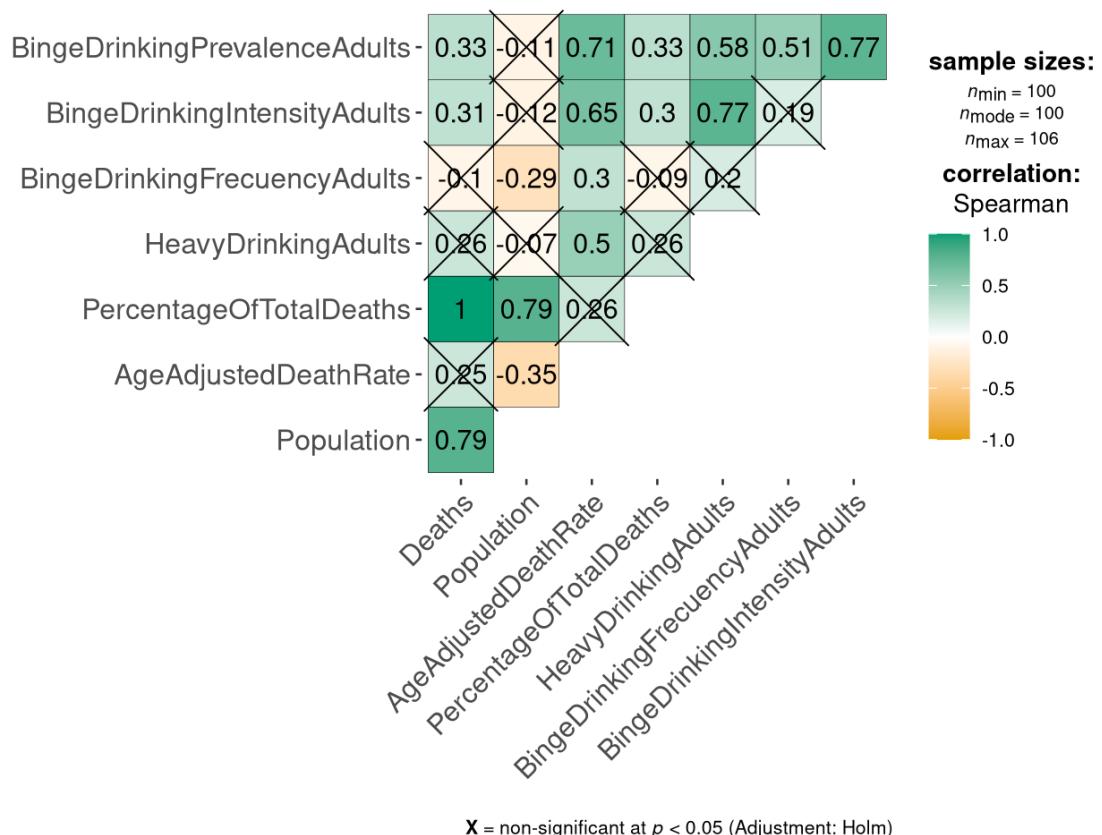
03gb - Correlograma (visualización de la correlación)



03gc - Test de hipótesis ggstatsplot::ggcorrmat()

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre Population y Deaths.
- Una correlación positiva entre:
 - Deaths y PercentageOfTotalDeaths.
 - Deaths y BingeDrinkingIntensityAdults.
 - Deaths y BingeDrinkingPrevalenceAdults.
 - Population y AgeAdjustedDeathRate.
 - Population y PercentageOfTotalDeaths.
 - Population y BingeDrinkingFrecuencyAdults.
 - AgeAdjustedDeathRate y HeavyDrinkingAdults.
 - AgeAdjustedDeathRate y BingeDrinkingFrecuencyAdults.
 - AgeAdjustedDeathRate y BingeDrinkingIntensityAdults.
 - AgeAdjustedDeathRate y BingeDrinkingPrevalenceAdults.
 - PercentageOfTotalDeaths y BingeDrinkingIntensityAdults.
 - PercentageOfTotalDeaths y BingeDrinkingPrevalenceAdults.
 - HeavyDrinkingAdults y BingeDrinkingIntensityAdults.
 - HeavyDrinkingAdults y BingeDrinkingPrevalenceAdults.
 - BingeDrinkingFrecuencyAdults y BingeDrinkingPrevalenceAdults.
 - BingeDrinkingIntensityAdults y BingeDrinkingPrevalenceAdults.



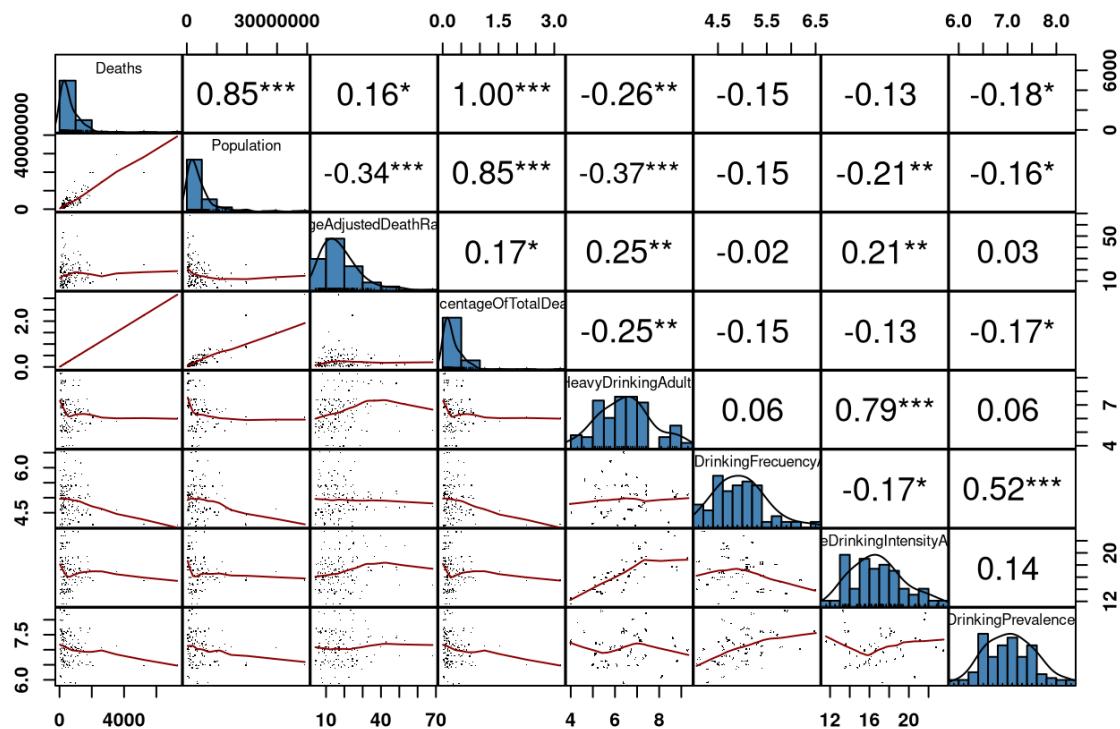
Objeto data_overall

03ga - Correlation matrix

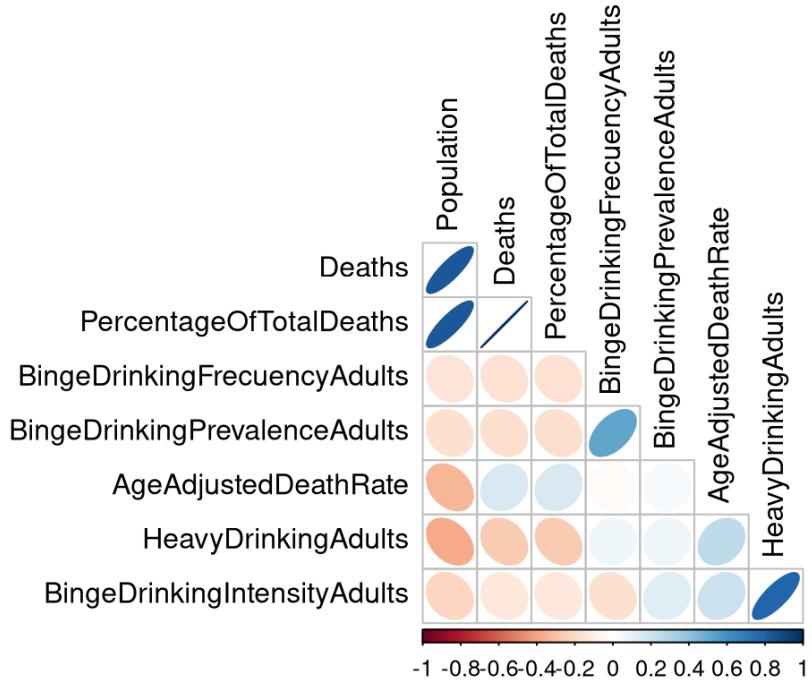
Se exploraron las correlaciones entre variables numéricas con el test de Spearman. Se obtuvieron los siguientes resultados (en verde, las correlaciones estadísticamente significativas):

row	column	cor	p
Deaths	Population	0.85	0.00
Deaths	AgeAdjustedDeathRate	0.16	0.04
Population	AgeAdjustedDeathRate	-0.34	0.00
Deaths	PercentageOfTotalDeaths	1.00	0.00
Population	PercentageOfTotalDeaths	0.85	0.00
AgeAdjustedDeathRate	PercentageOfTotalDeaths	0.17	0.04
Deaths	HeavyDrinkingAdults	-0.26	0.00
Population	HeavyDrinkingAdults	-0.37	0.00
AgeAdjustedDeathRate	HeavyDrinkingAdults	0.25	0.00
PercentageOfTotalDeaths	HeavyDrinkingAdults	-0.25	0.00
Deaths	BingeDrinkingFrequencyAdults	-0.15	0.06
Population	BingeDrinkingFrequencyAdults	-0.15	0.07
AgeAdjustedDeathRate	BingeDrinkingFrequencyAdults	-0.02	0.76
PercentageOfTotalDeaths	BingeDrinkingFrequencyAdults	-0.15	0.06
HeavyDrinkingAdults	BingeDrinkingFrequencyAdults	0.06	0.43
Deaths	BingeDrinkingIntensityAdults	-0.13	0.11
Population	BingeDrinkingIntensityAdults	-0.21	0.01
AgeAdjustedDeathRate	BingeDrinkingIntensityAdults	0.21	0.01
PercentageOfTotalDeaths	BingeDrinkingIntensityAdults	-0.13	0.13
HeavyDrinkingAdults	BingeDrinkingIntensityAdults	0.79	0.00
BingeDrinkingFrequencyAdults	BingeDrinkingIntensityAdults	-0.17	0.04
Deaths	BingeDrinkingPrevalenceAdults	-0.18	0.03
Population	BingeDrinkingPrevalenceAdults	-0.16	0.04
AgeAdjustedDeathRate	BingeDrinkingPrevalenceAdults	0.03	0.67
PercentageOfTotalDeaths	BingeDrinkingPrevalenceAdults	-0.17	0.04
HeavyDrinkingAdults	BingeDrinkingPrevalenceAdults	0.06	0.46
BingeDrinkingFrequencyAdults	BingeDrinkingPrevalenceAdults	0.52	0.00
BingeDrinkingIntensityAdults	BingeDrinkingPrevalenceAdults	0.14	0.10

Correlation matrix



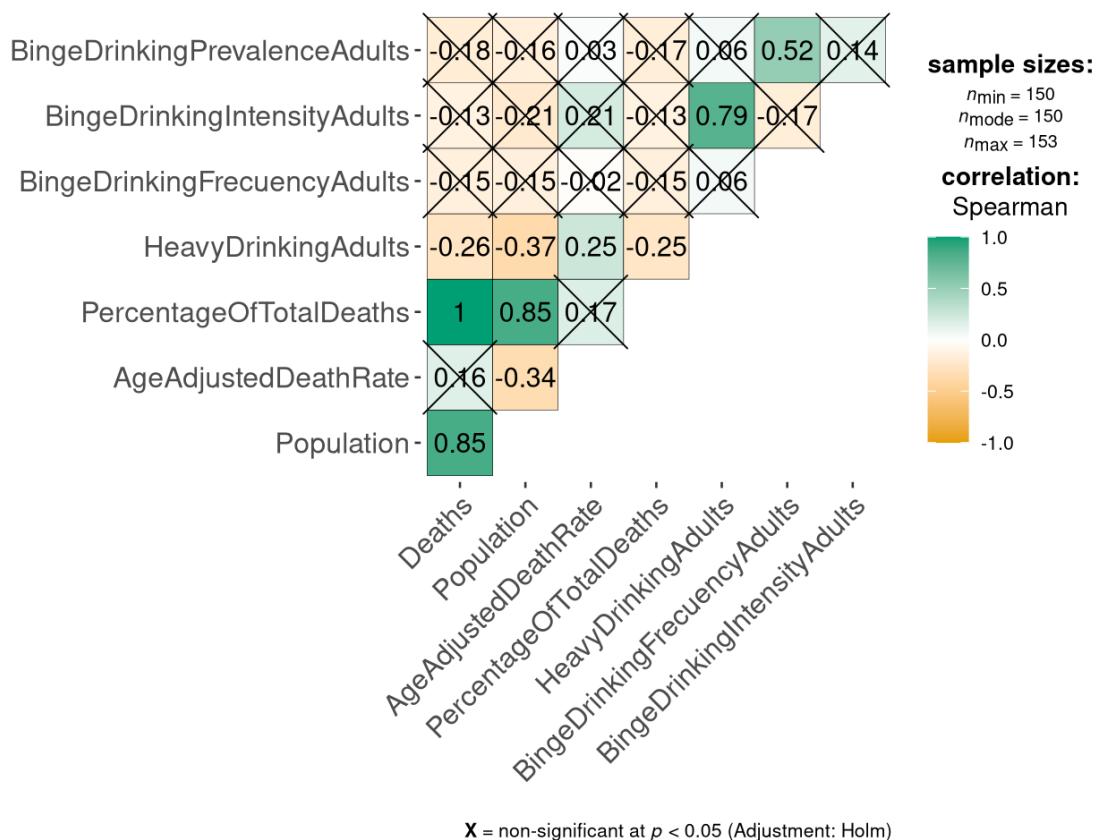
03gb - Correlograma (visualización de la correlación)



03gc - Test de hipótesis ggstatsplot::ggcorrmat()

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre:
 - Deaths y HeavyDrinkingAdults.
 - Population y AgeAdjustedDeathRate.
 - Population y HeavyDrinkingAdults.
- Una correlación positiva entre:
 - Deaths y Population.
 - Deaths y PercentageOfTotalDeaths.
 - PercentageOfTotalDeaths y HeavyDrinkingAdults.
 - Population y PercentageOfTotalDeaths.
 - AgeAdjustedDeathRate y HeavyDrinkingAdults.
 - HeavyDrinkingAdults y BingeDrinkingIntensityAdults.
 - BingeDrinkingFrequencyAdults y BingeDrinkingPrevalenceAdults.



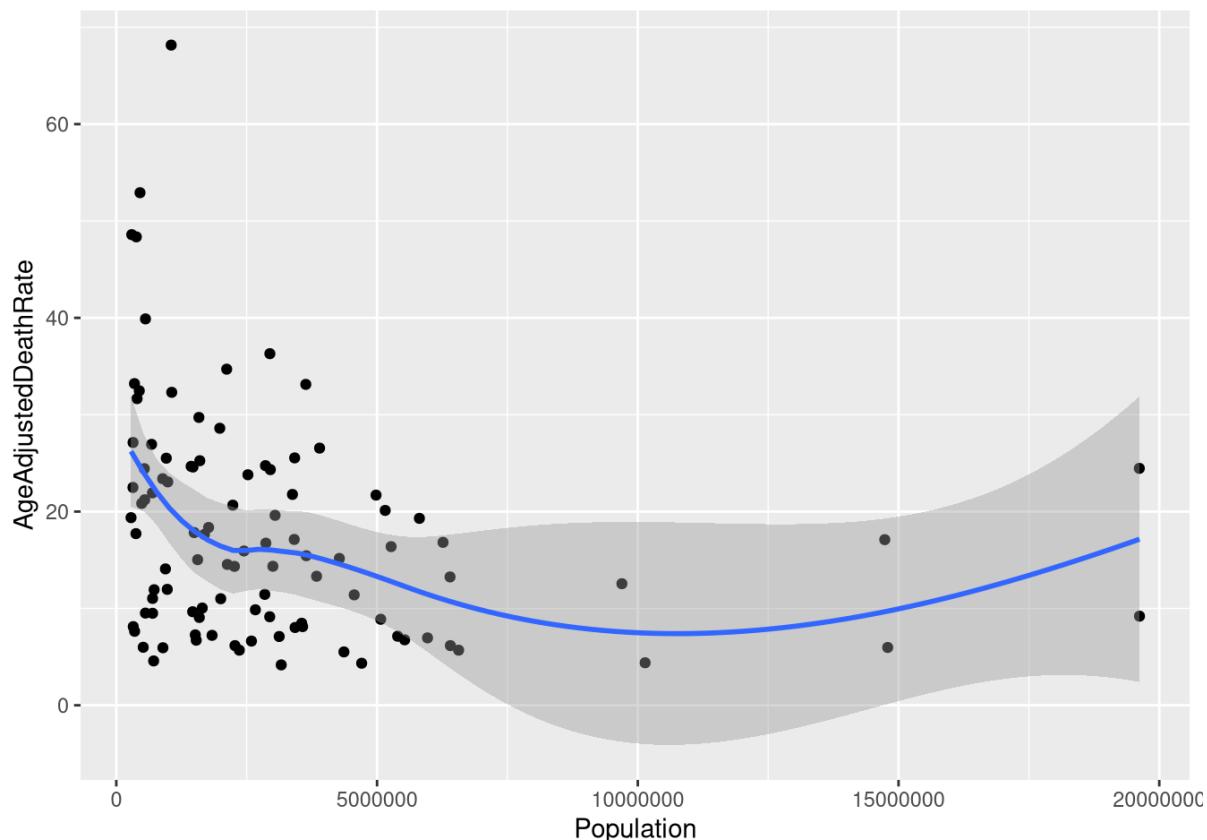
03h - Explorar modelos de datos para las correlaciones estadísticamente significativas
Se crearon modelos exploratorios para todos los pares de variables en las que se ha obtenido una correlación lineal estadística significativa, para cada uno de los *data.frame*.

Objeto data

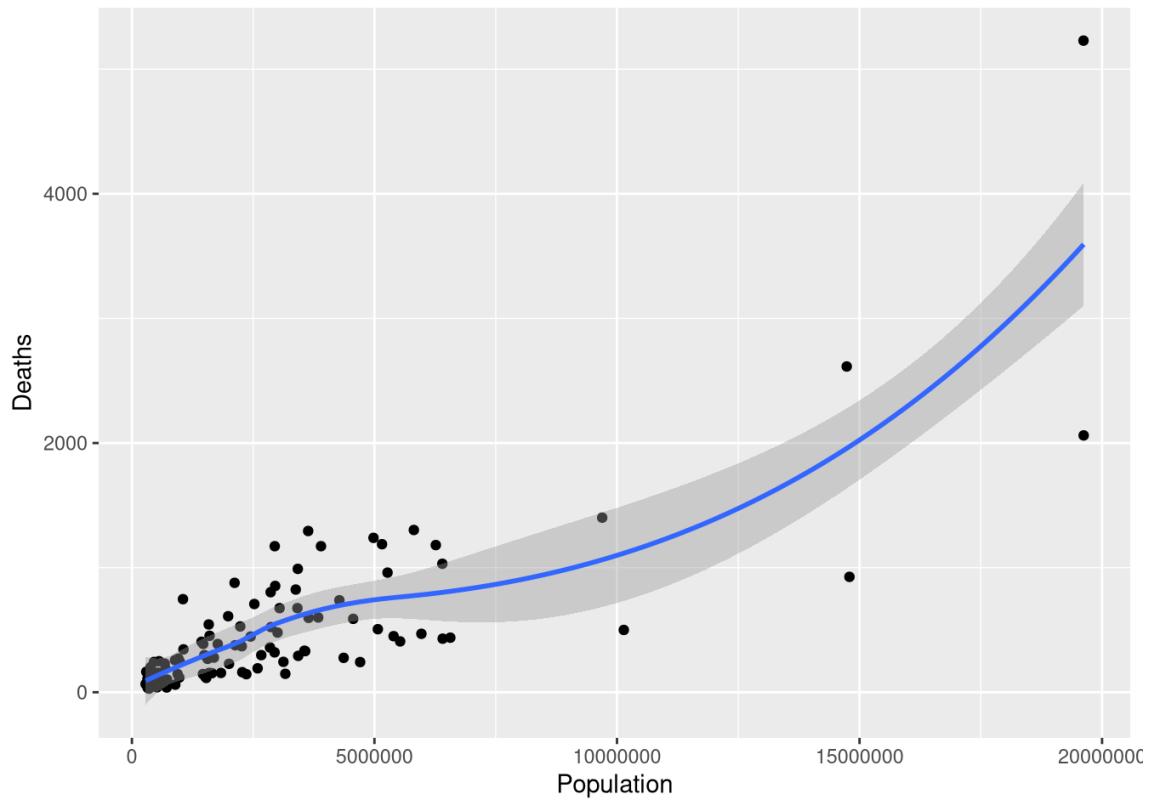
Se identificaron correlaciones estadísticamente significativas en los siguientes pares de variables:

- Correlación negativa entre Population y AgeAdjustedDeathRate
- Correlación positiva entre:
 - Deaths y Population
 - Deaths y PercentageOfTotalDeaths
 - Population y PercentageOfTotalDeaths
 - HeavyDrinkingAdults y AgeAdjustedDeathRate
 - BingeDrinkingPrevalenceAdults y HeavyDrinkingAdults
 - BingeDrinkingPrevalenceAdults y BingeDrinkingFrequencyAdults
 - BingeDrinkingPrevalenceAdults y BingeDrinkingIntensityAdults
 - BingeDrinkingIntensityAdults y BingeDrinkingFrequencyAdults

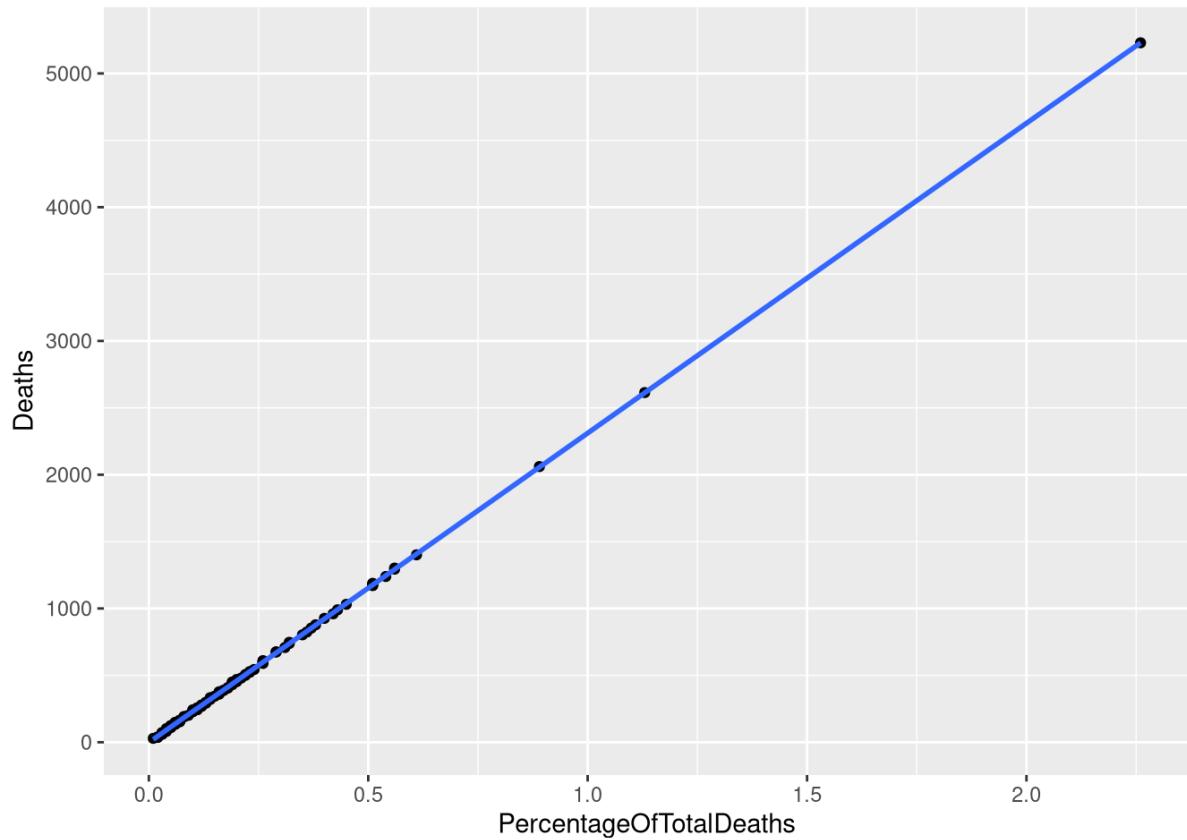
Population y AgeAdjustedDeathRate



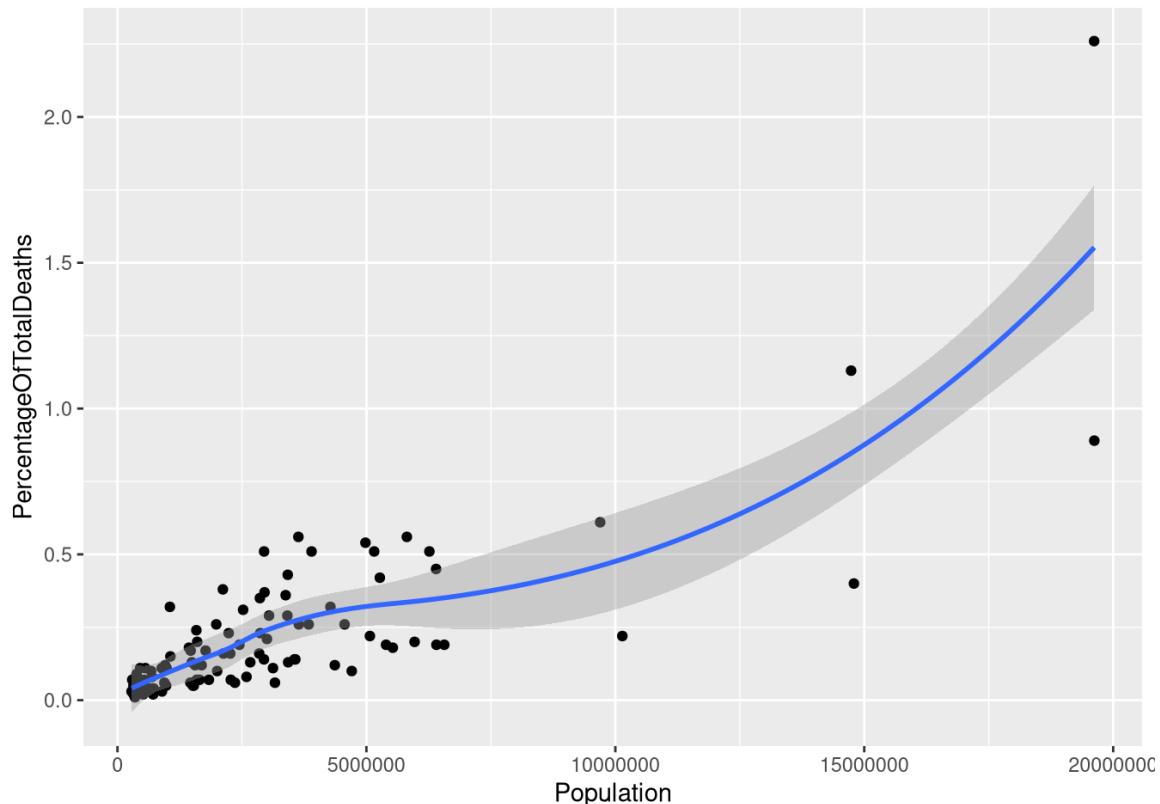
Deaths y Population



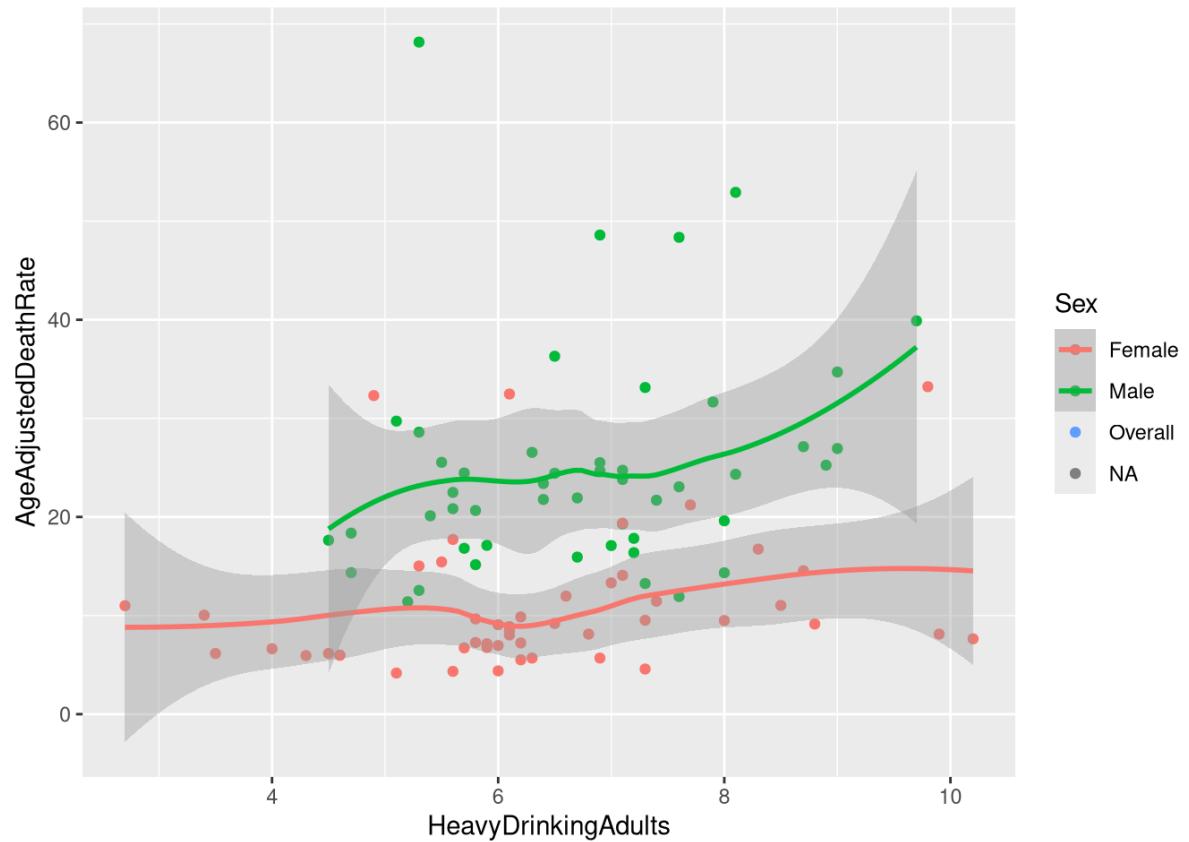
Deaths y PercentageOfTotalDeaths



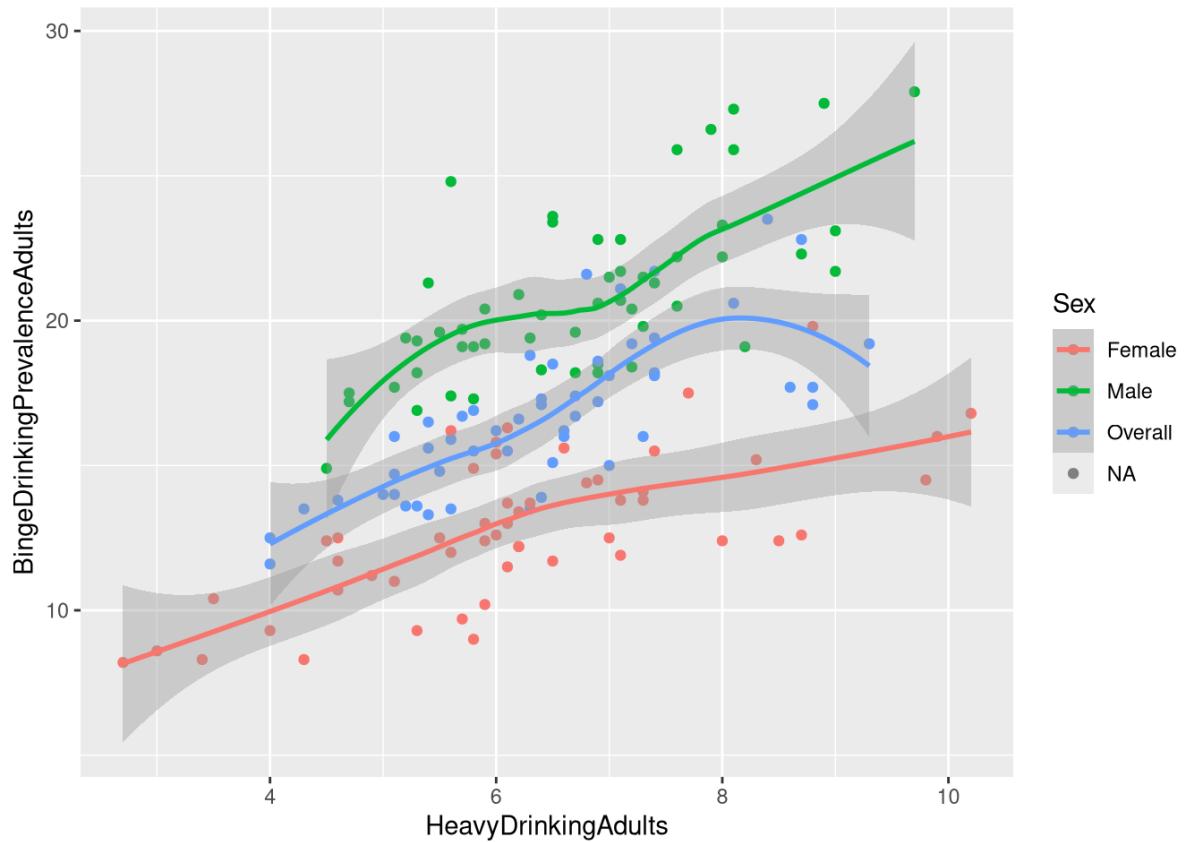
Population y PercentageOfTotalDeaths



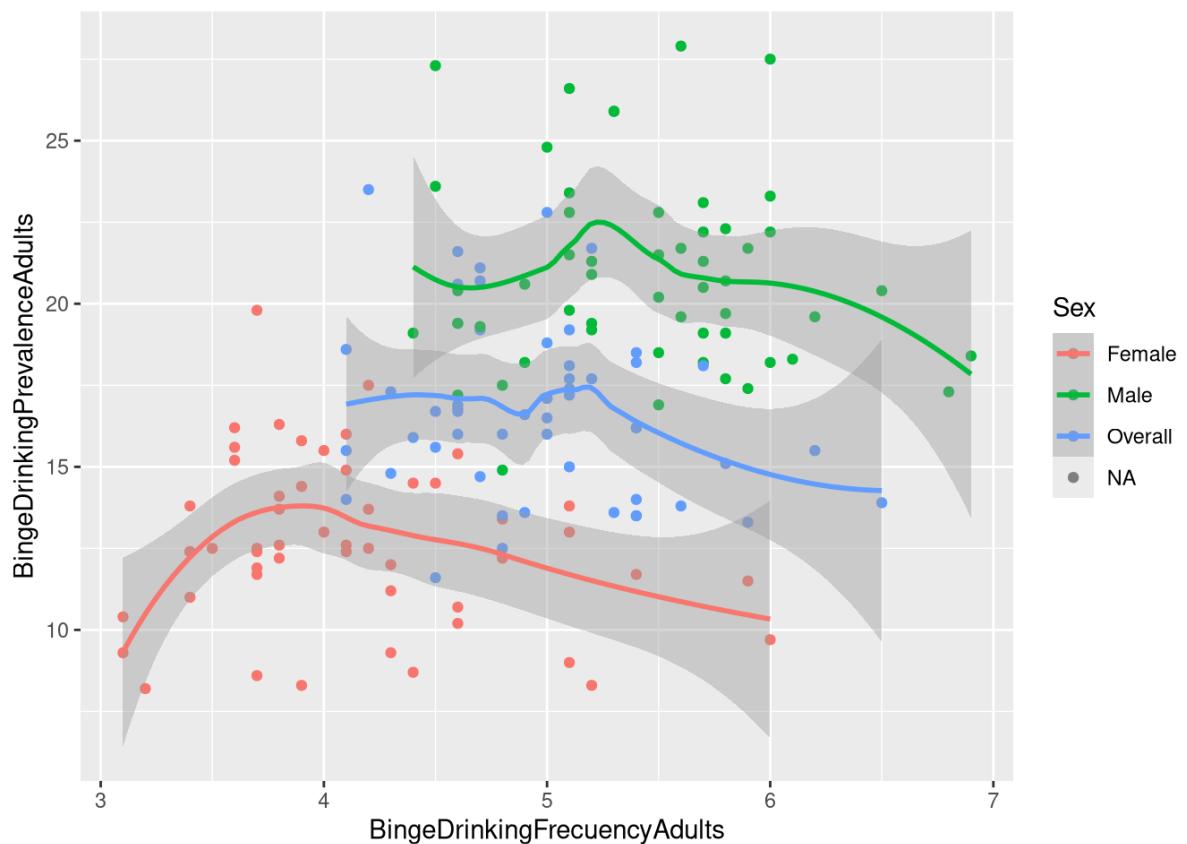
HeavyDrinkingAdults y AgeAdjustedDeathRate



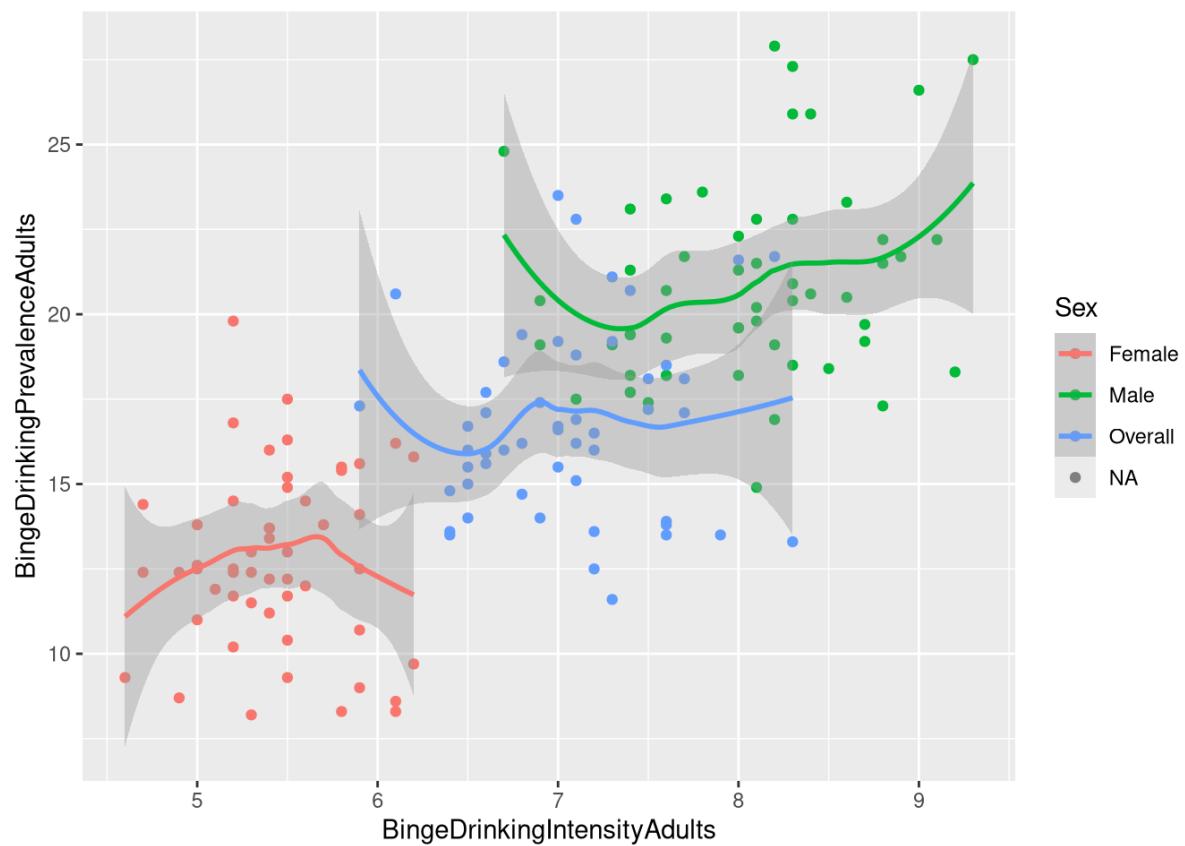
BingeDrinkingPrevalenceAdults y HeavyDrinkingAdults



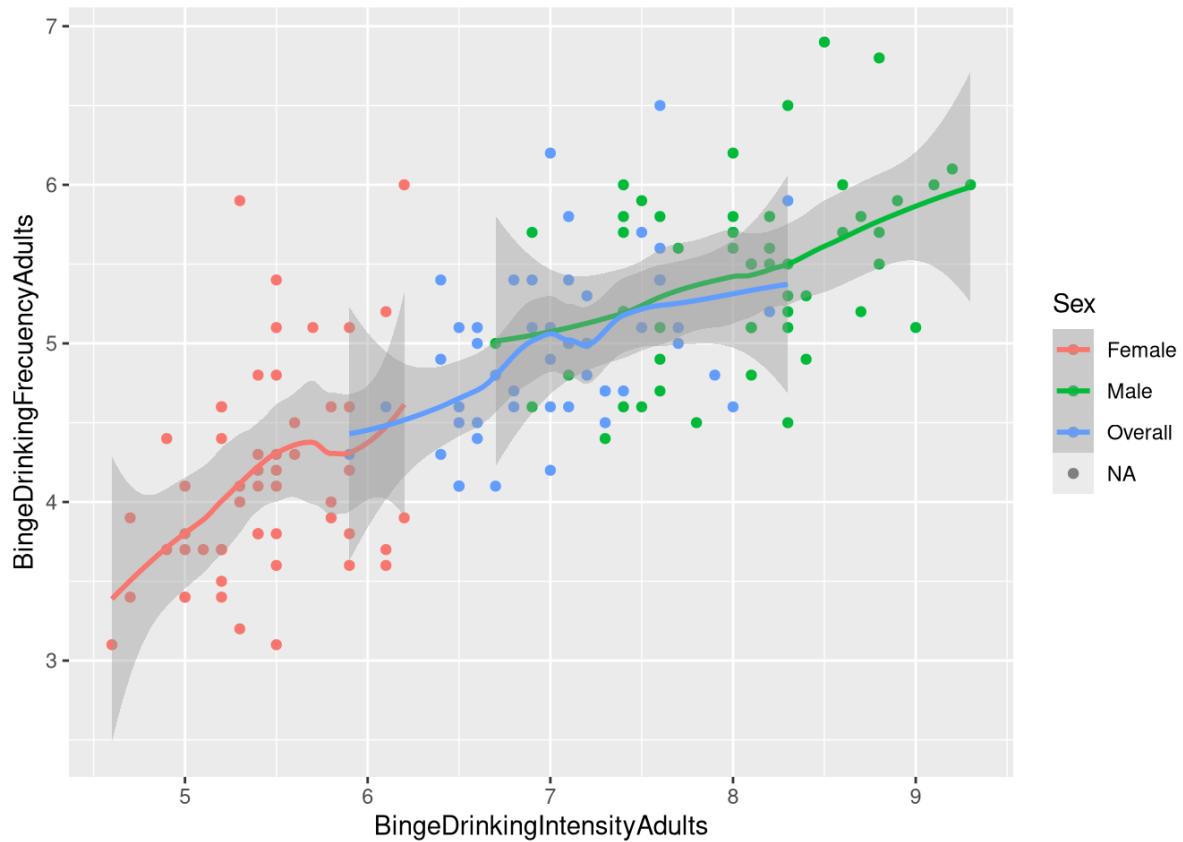
BingeDrinkingPrevalenceAdults y BingeDrinkingFrequencyAdults



BingeDrinkingPrevalenceAdults y BingeDrinkingIntensityAdults



BingeDrinkingIntensityAdults y BingeDrinkingFrecuencyAdults

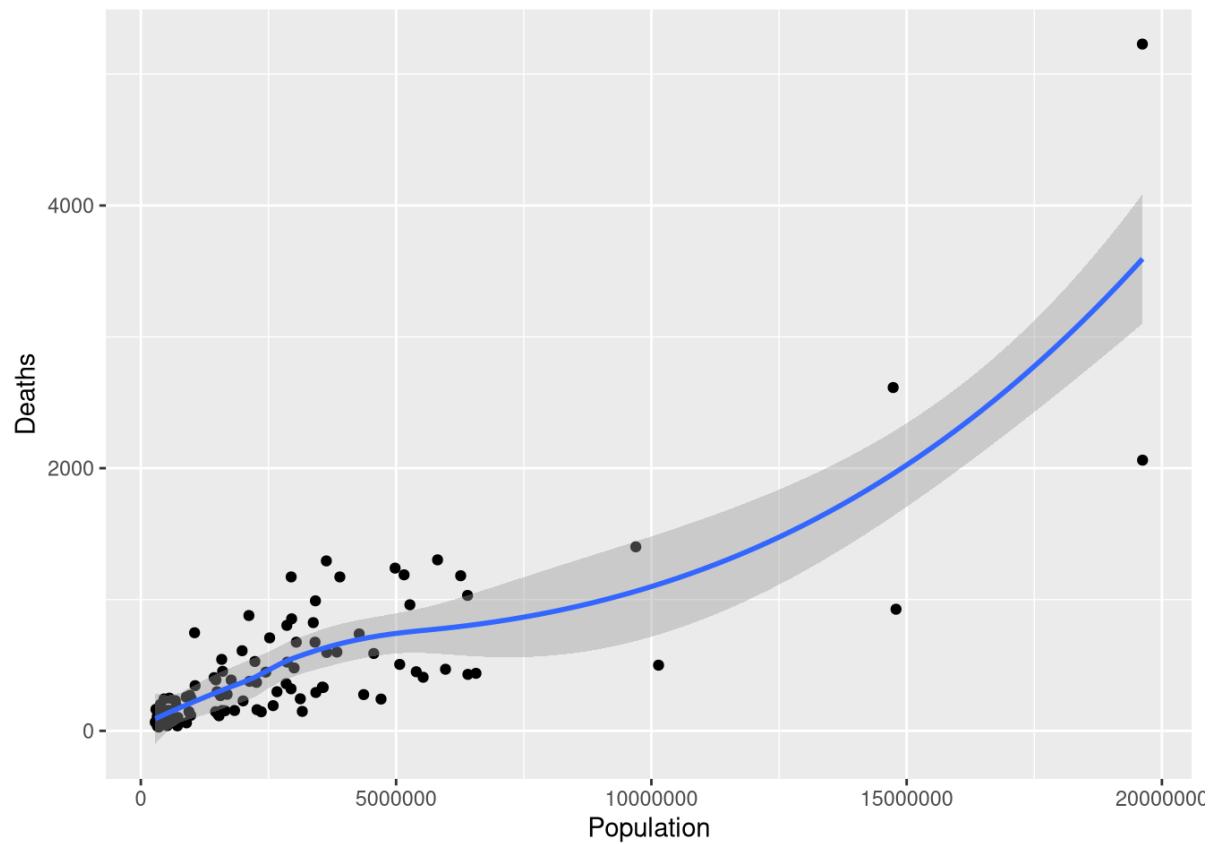


Objeto data_gender

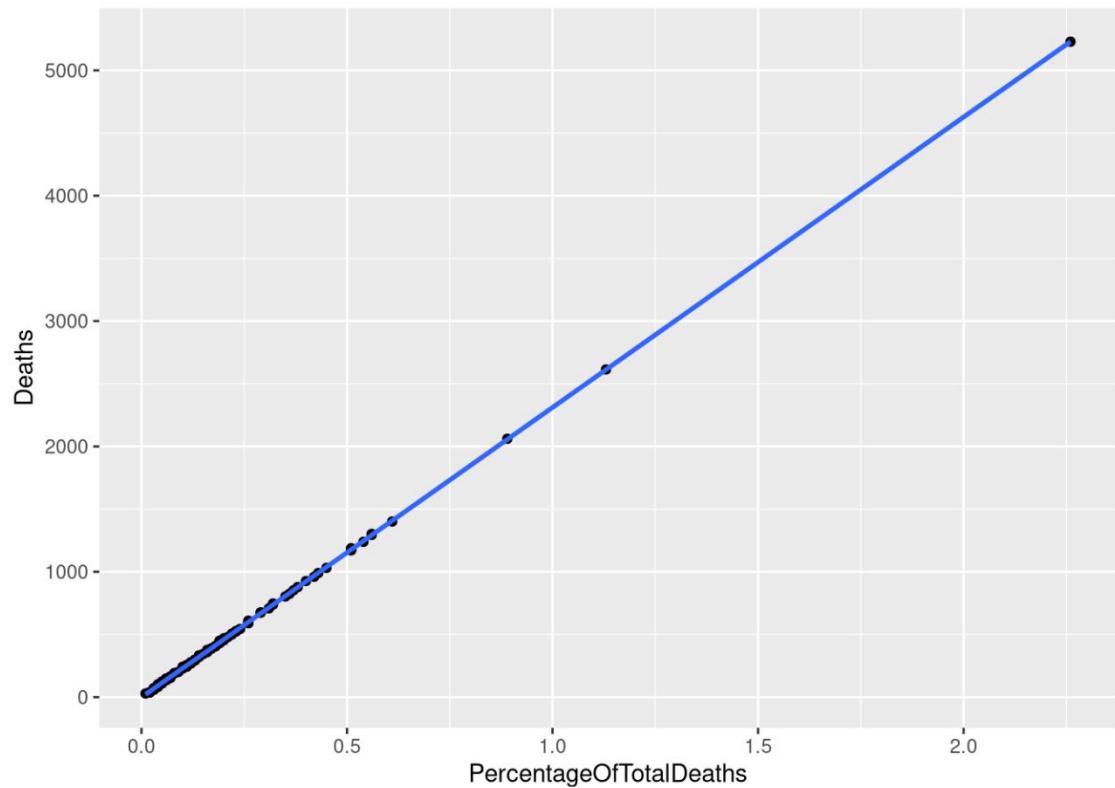
Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre Population y Deaths.
- Una correlación positiva entre:
 - Deaths y PercentageOfTotalDeaths.
 - Deaths y BingeDrinkingIntensityAdults.
 - Deaths y BingeDrinkingPrevalenceAdults.
 - Population y AgeAdjustedDeathRate.
 - Population y PercentageOfTotalDeaths.
 - Population y BingeDrinkingFrecuencyAdults.
 - AgeAdjustedDeathRate y HeavyDrinkingAdults.
 - AgeAdjustedDeathRate y BingeDrinkingFrecuencyAdults.
 - AgeAdjustedDeathRate y BingeDrinkingIntensityAdults.
 - AgeAdjustedDeathRate y BingeDrinkingPrevalenceAdults.
 - PercentageOfTotalDeaths y BingeDrinkingIntensityAdults.
 - PercentageOfTotalDeaths y BingeDrinkingPrevalenceAdults.
 - HeavyDrinkingAdults y BingeDrinkingIntensityAdults.
 - HeavyDrinkingAdults y BingeDrinkingPrevalenceAdults.
 - BingeDrinkingFrecuencyAdults y BingeDrinkingPrevalenceAdults.
 - BingeDrinkingIntensityAdults y BingeDrinkingPrevalenceAdults.

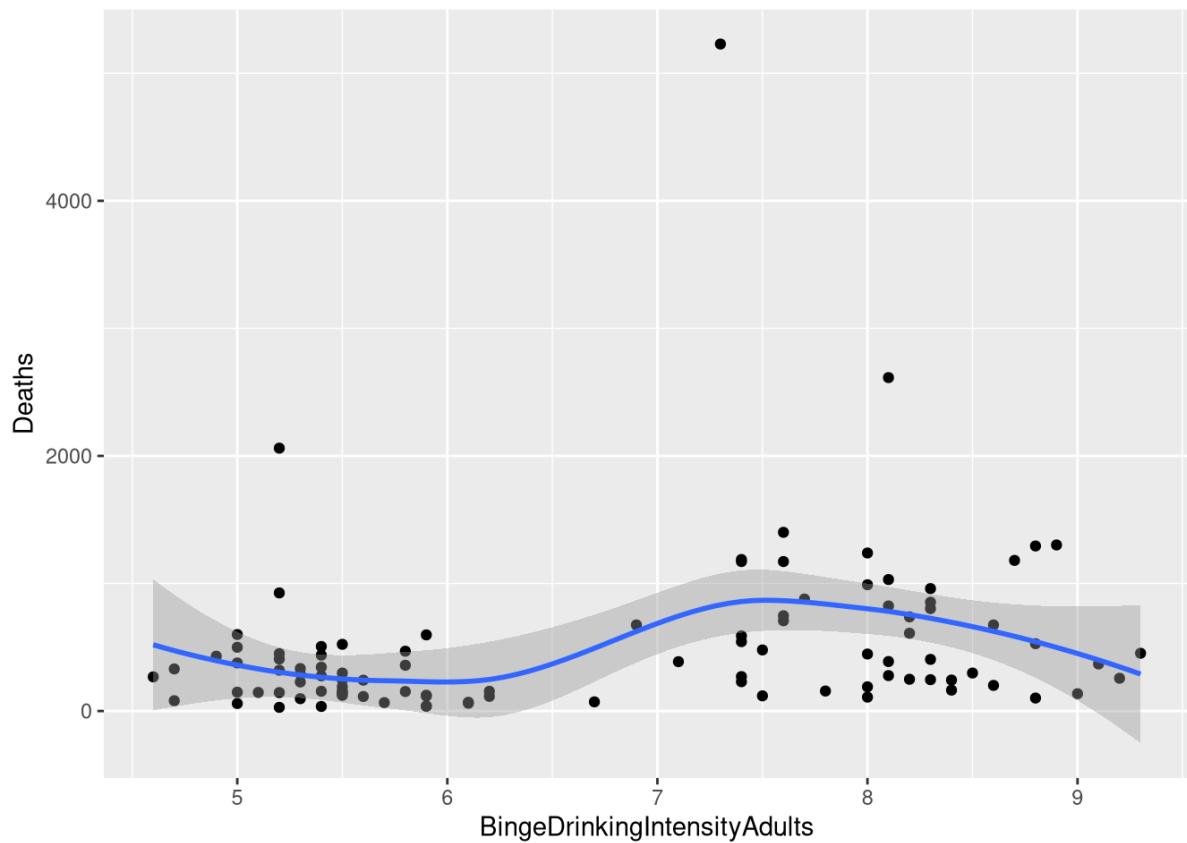
Population y Deaths



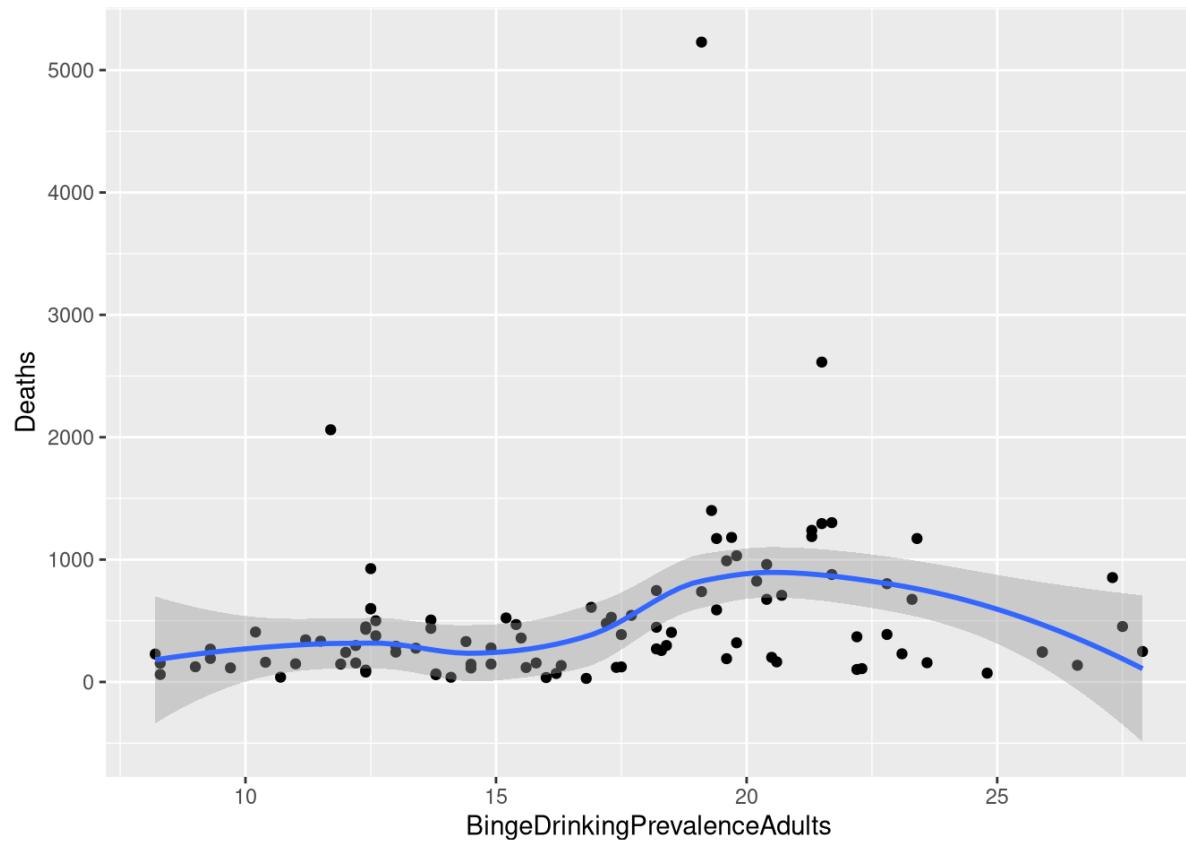
Deaths y PercentageOfTotalDeaths



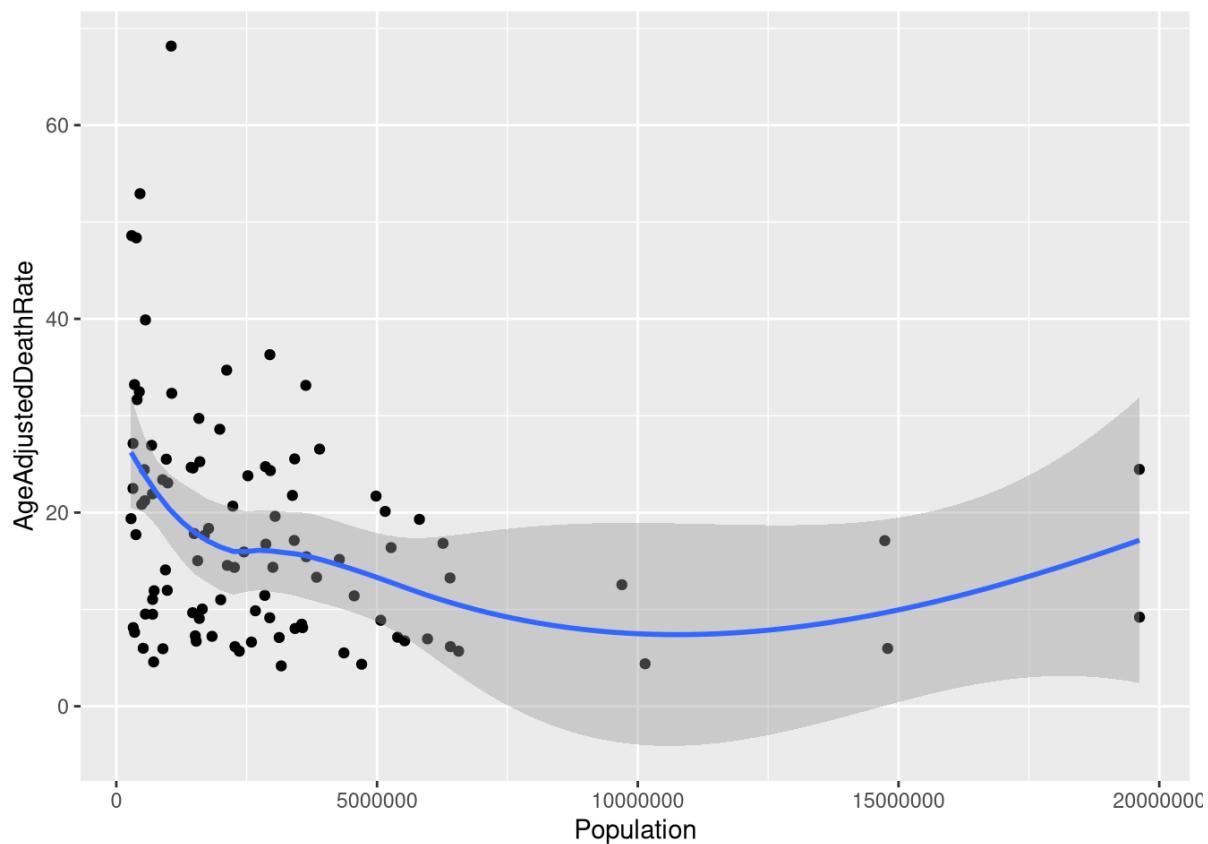
Deaths y BingeDrinkingIntensityAdults



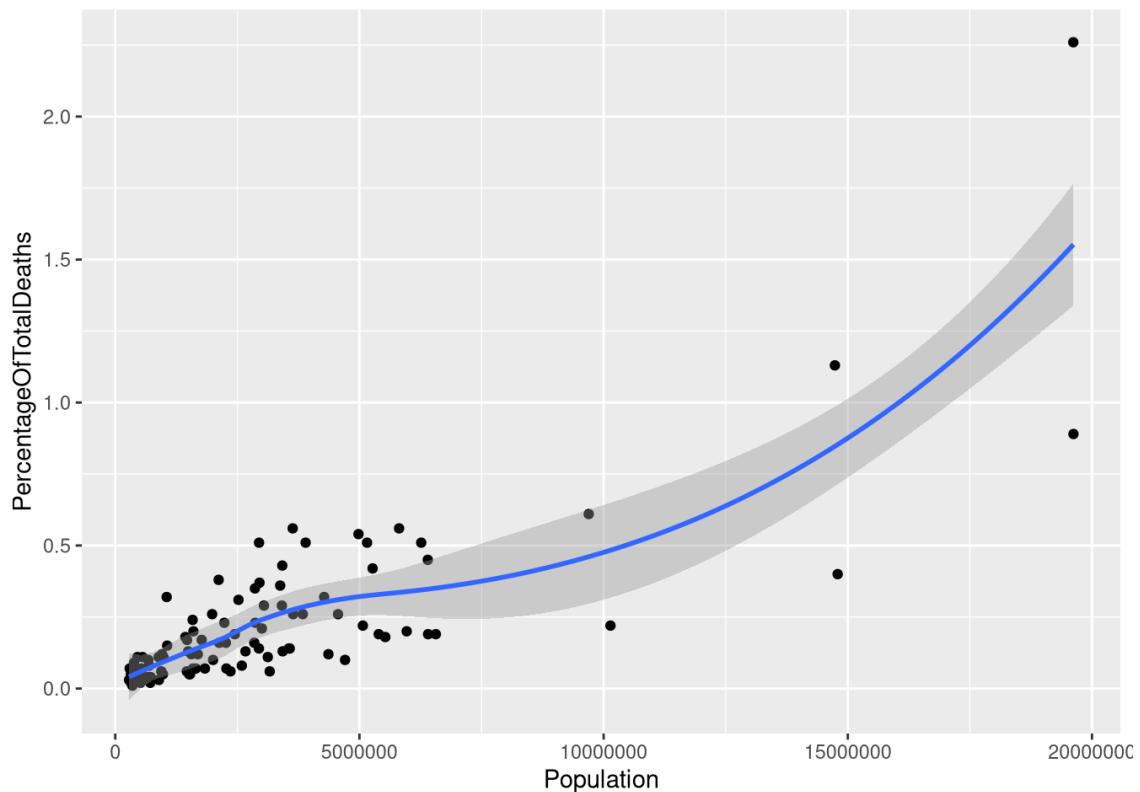
Deaths y BingeDrinkingPrevalenceAdults



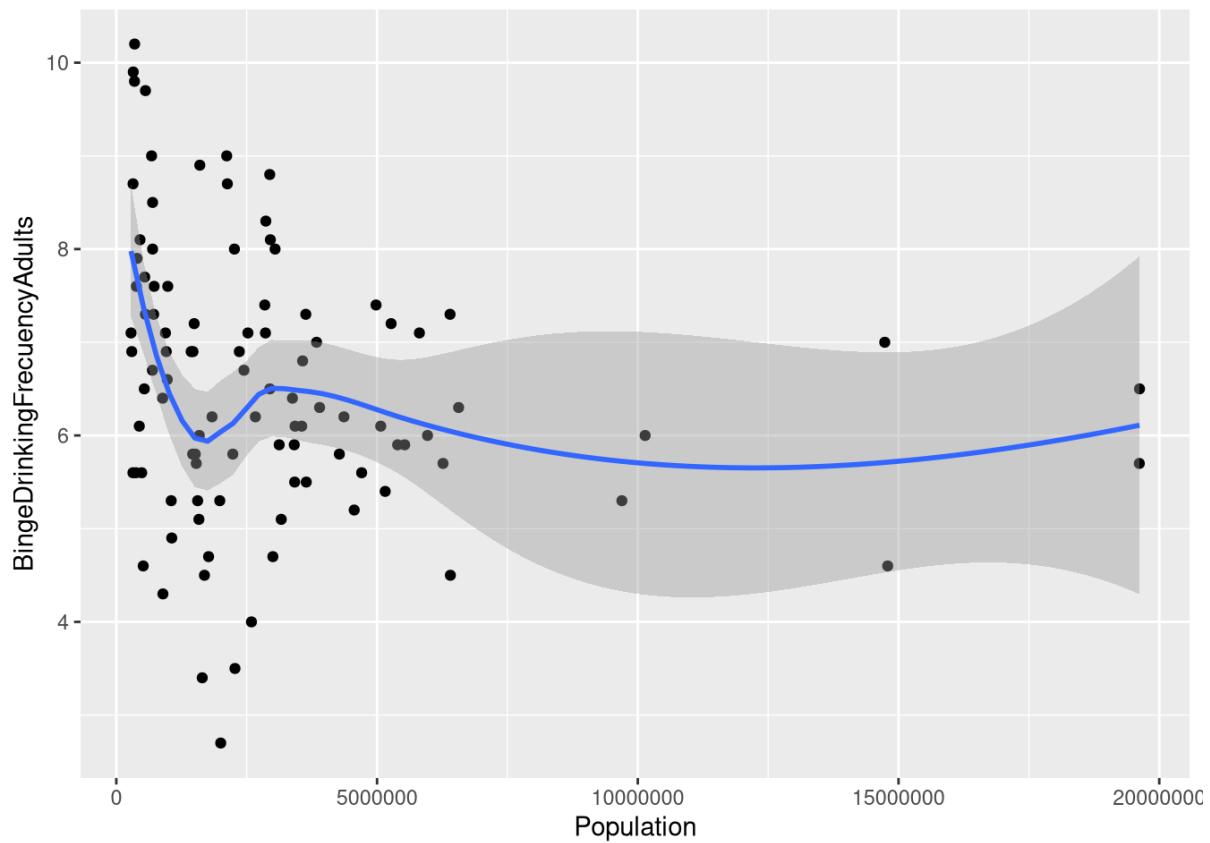
Population y AgeAdjustedDeathRate



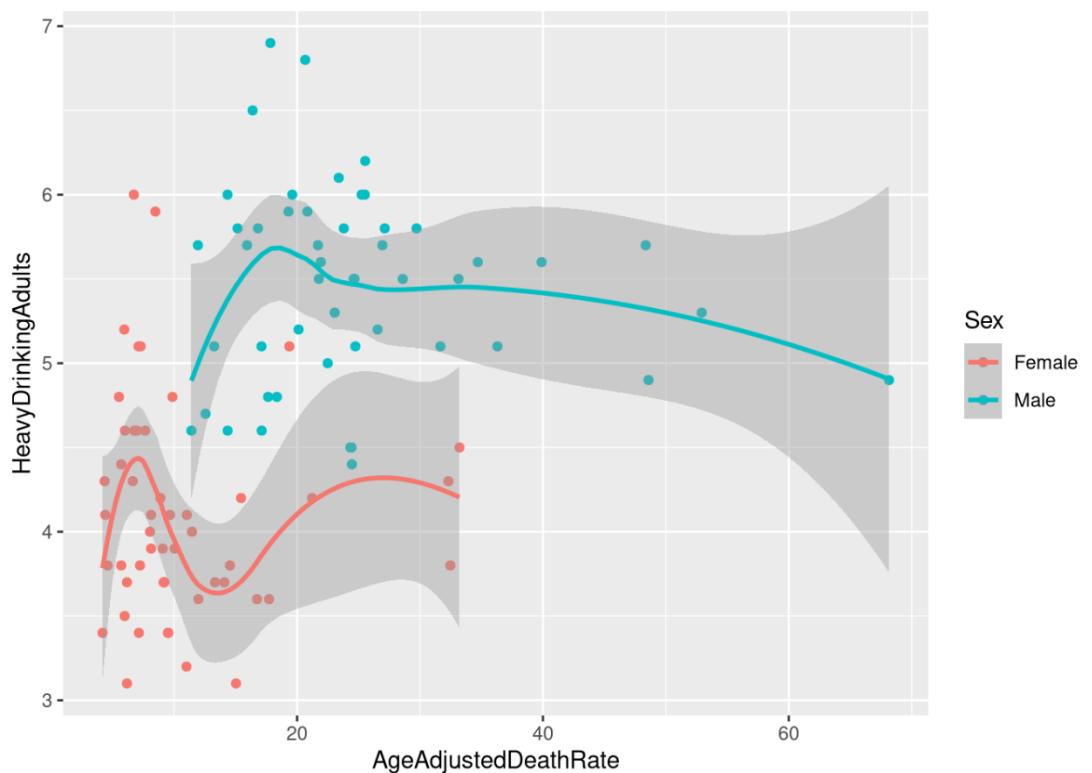
Population y PercentageOfTotalDeaths



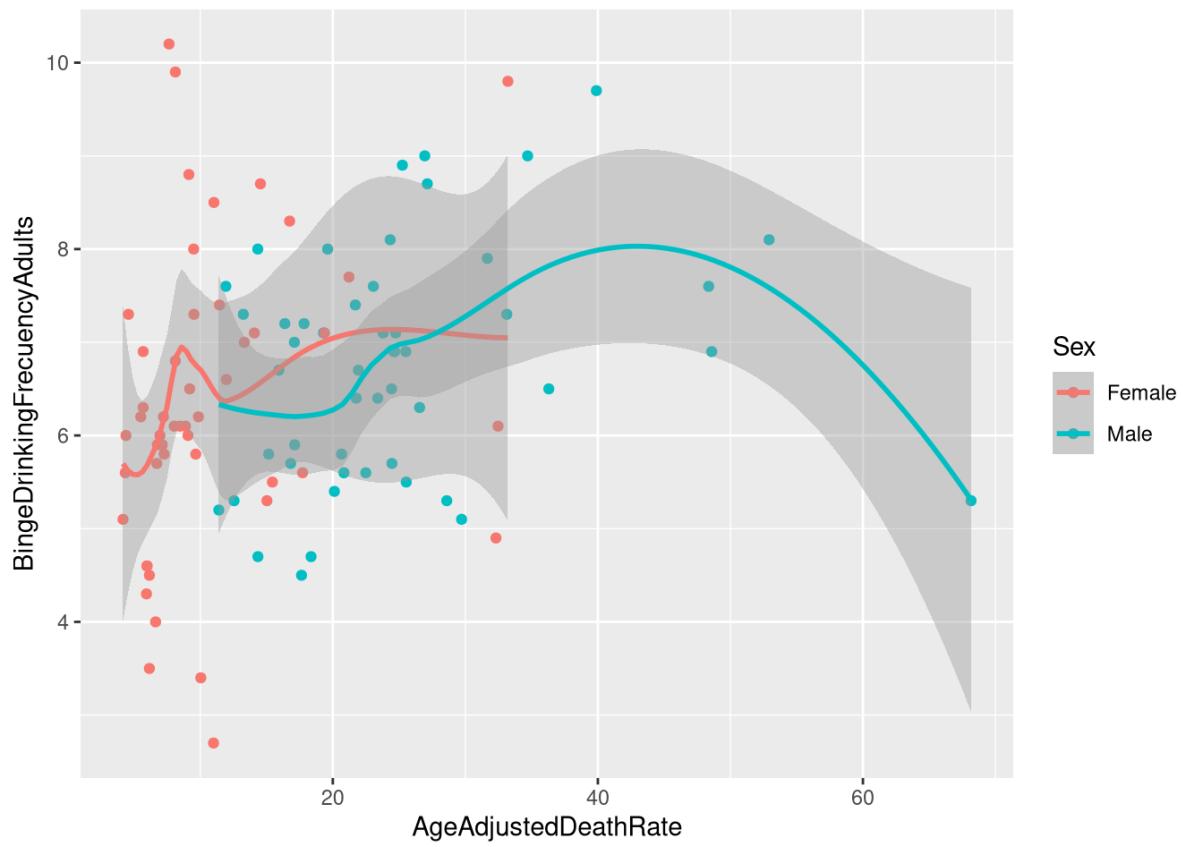
Population y BingeDrinkingFrequencyAdults



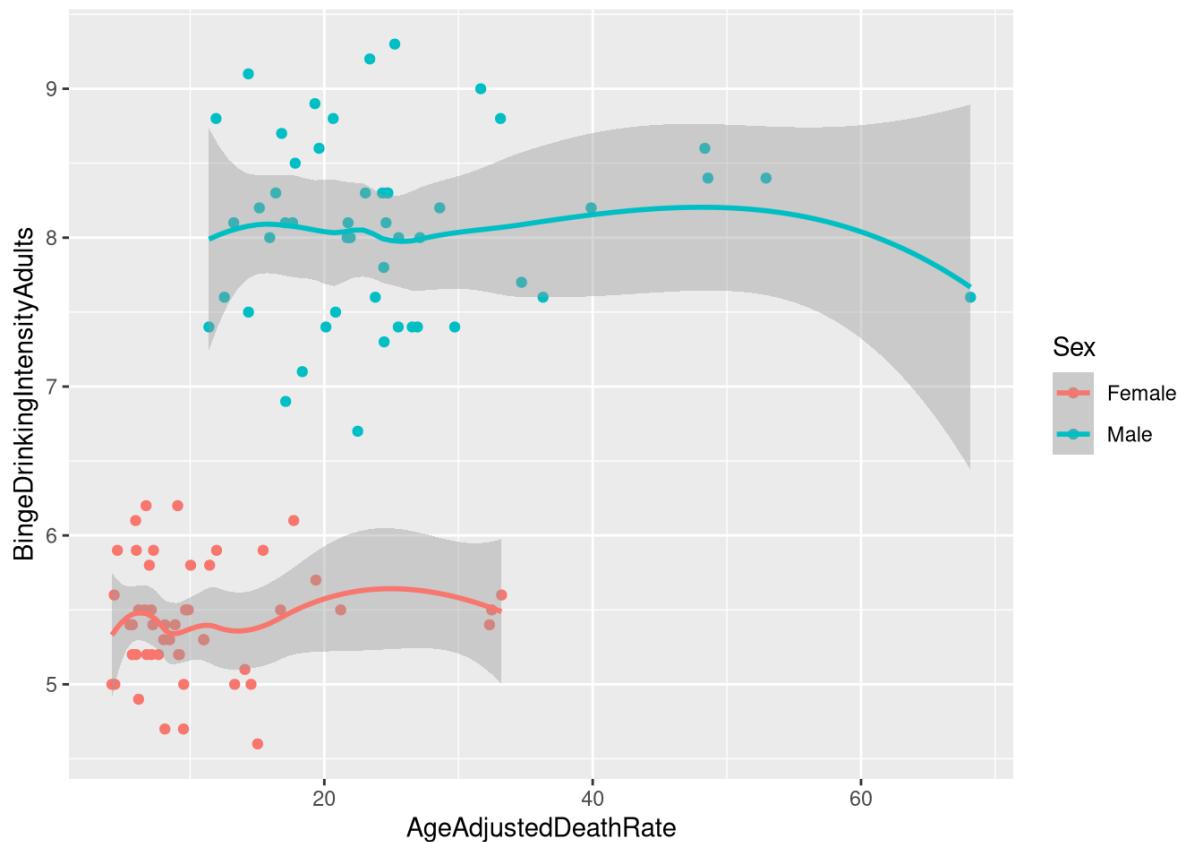
AgeAdjustedDeathRate y HeavyDrinkingAdults



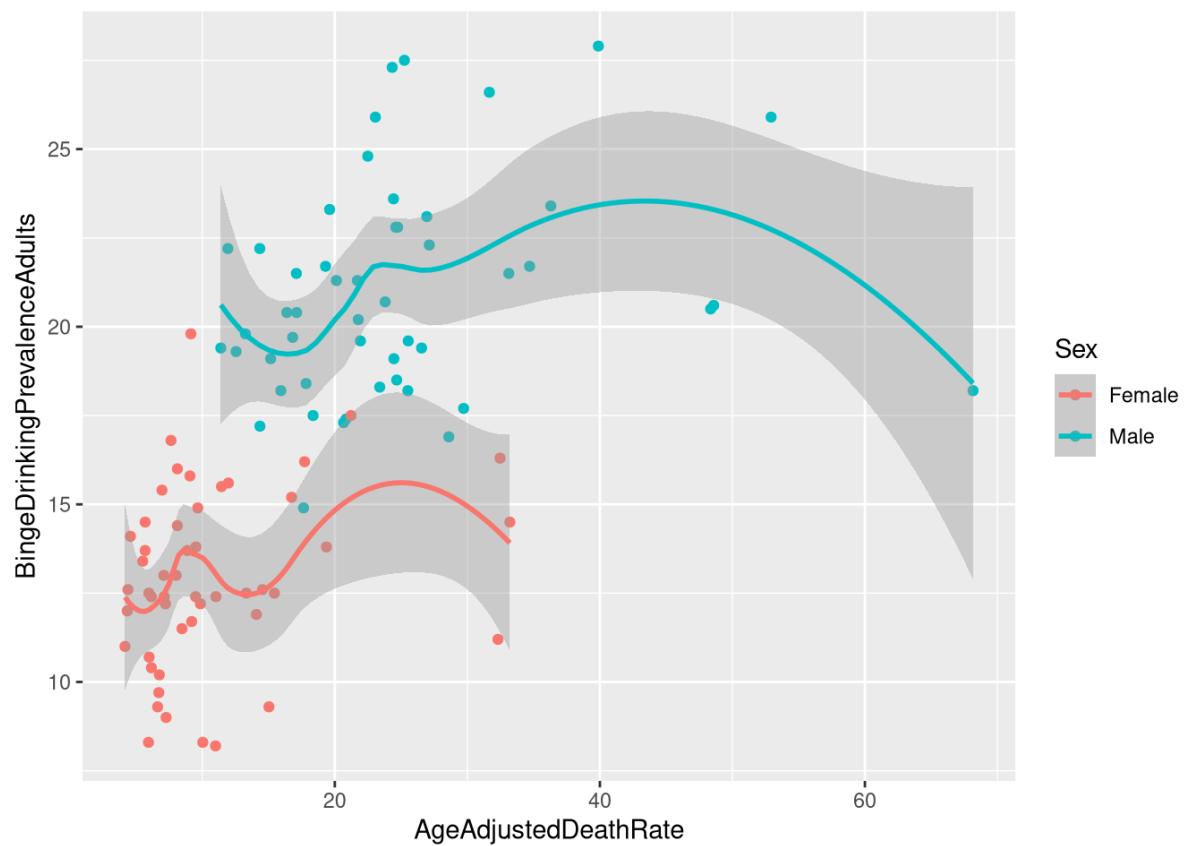
AgeAdjustedDeathRate y BingeDrinkingFrecuencyAdults



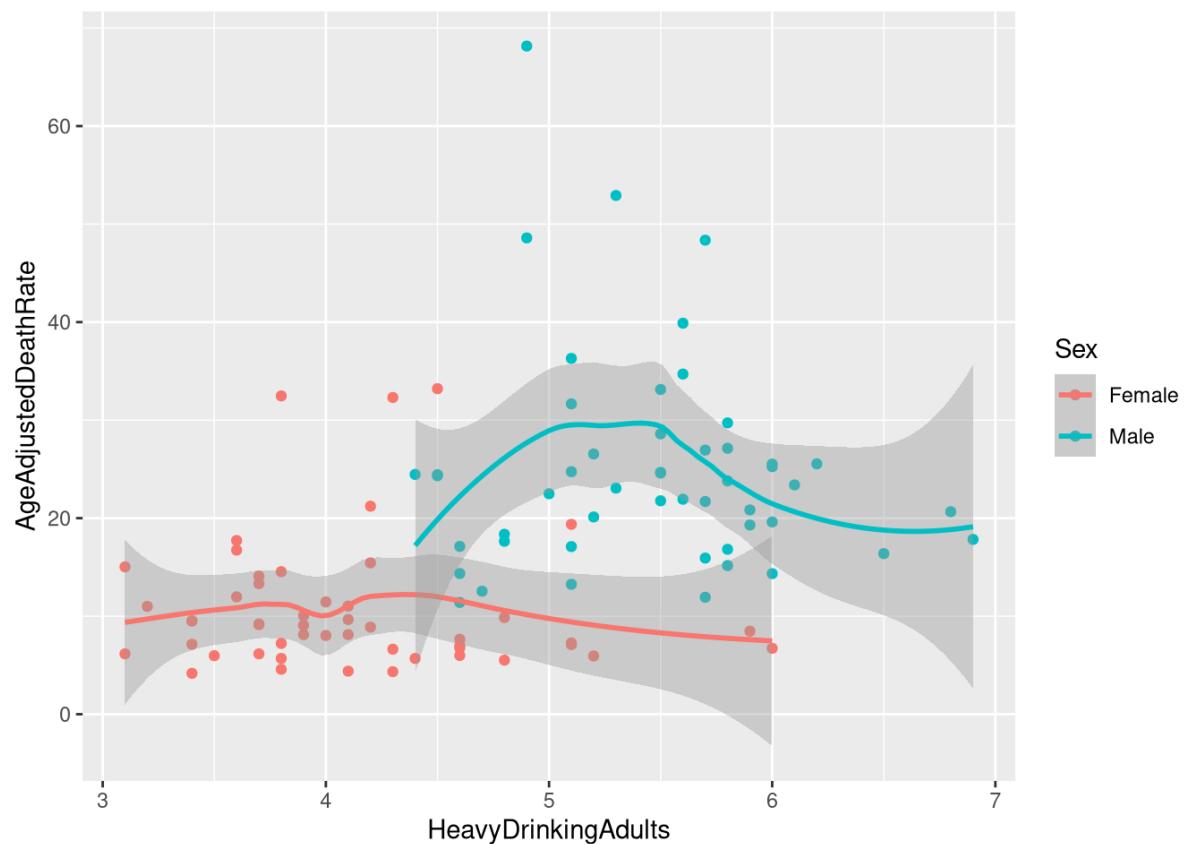
AgeAdjustedDeathRate y BingeDrinkingIntensityAdults



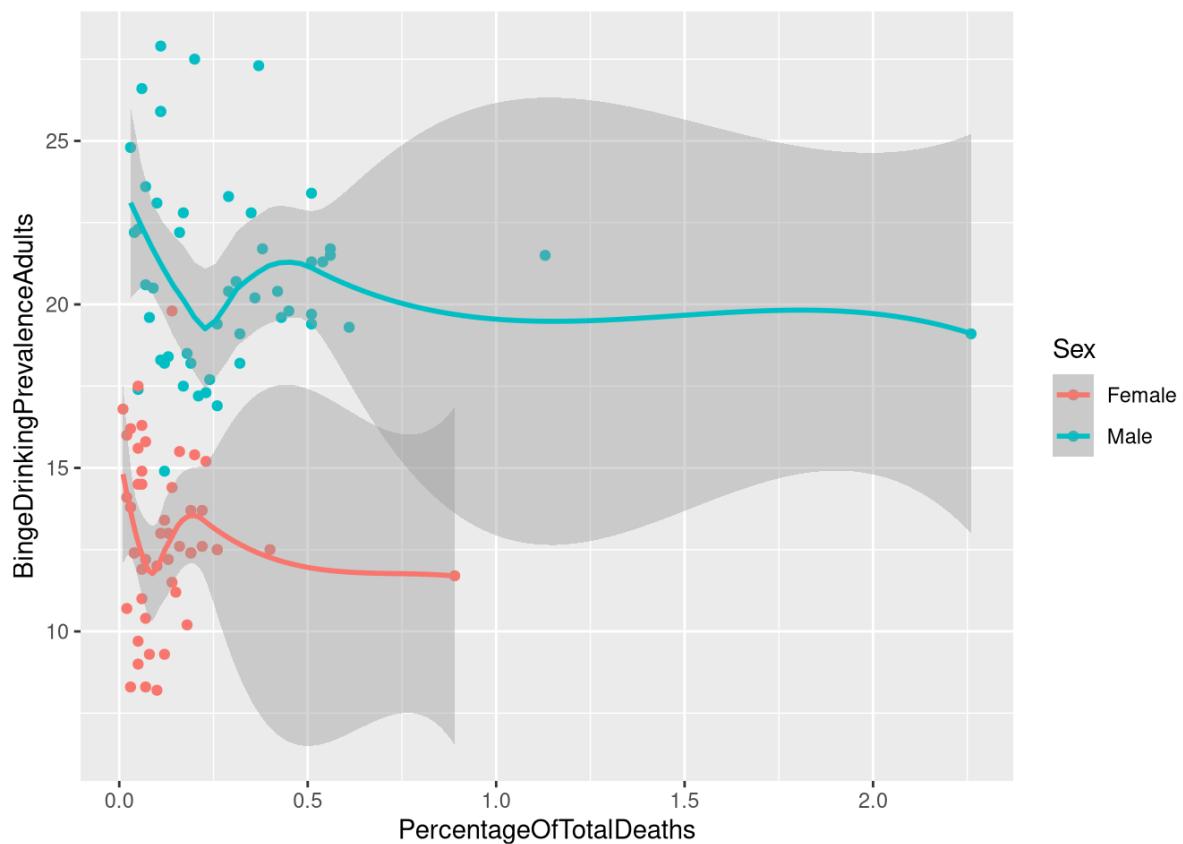
AgeAdjustedDeathRate y BingeDrinkingPrevalenceAdults



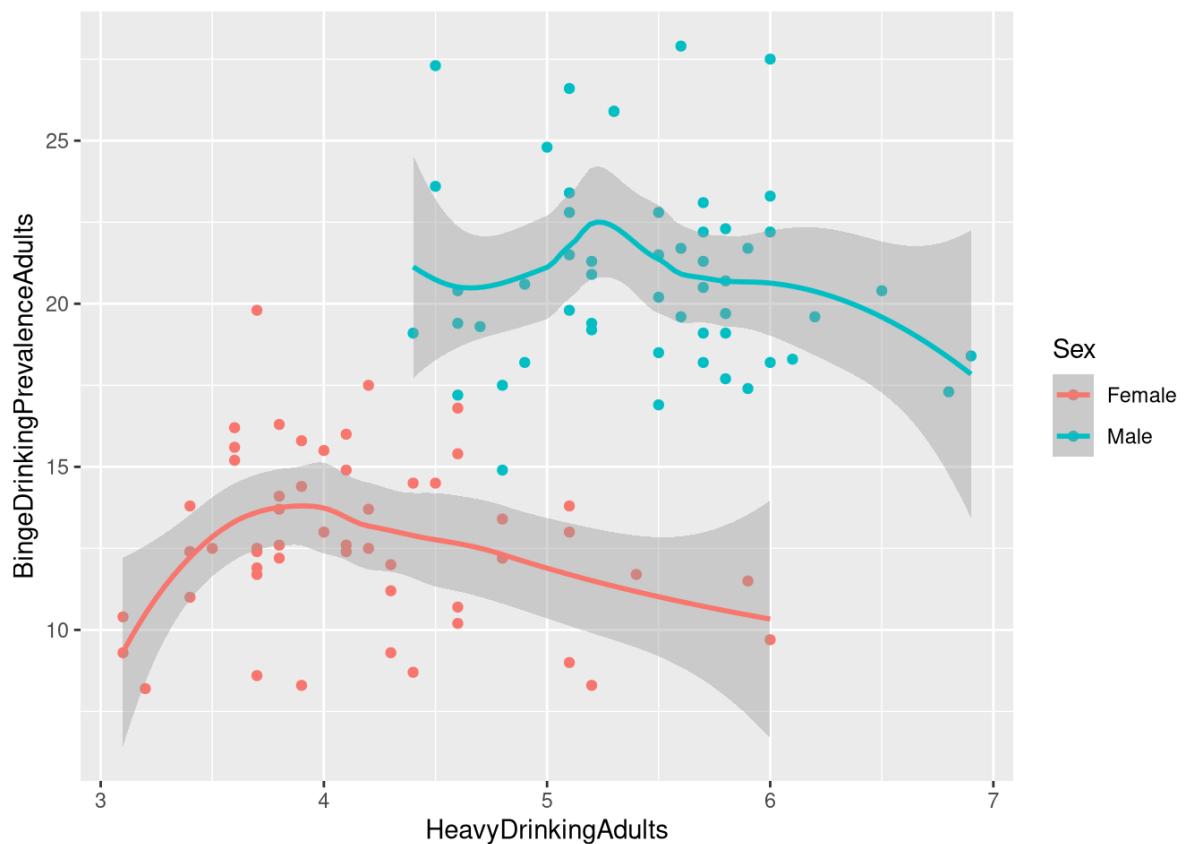
PercentageOfTotalDeaths y BingeDrinkingIntensityAdults



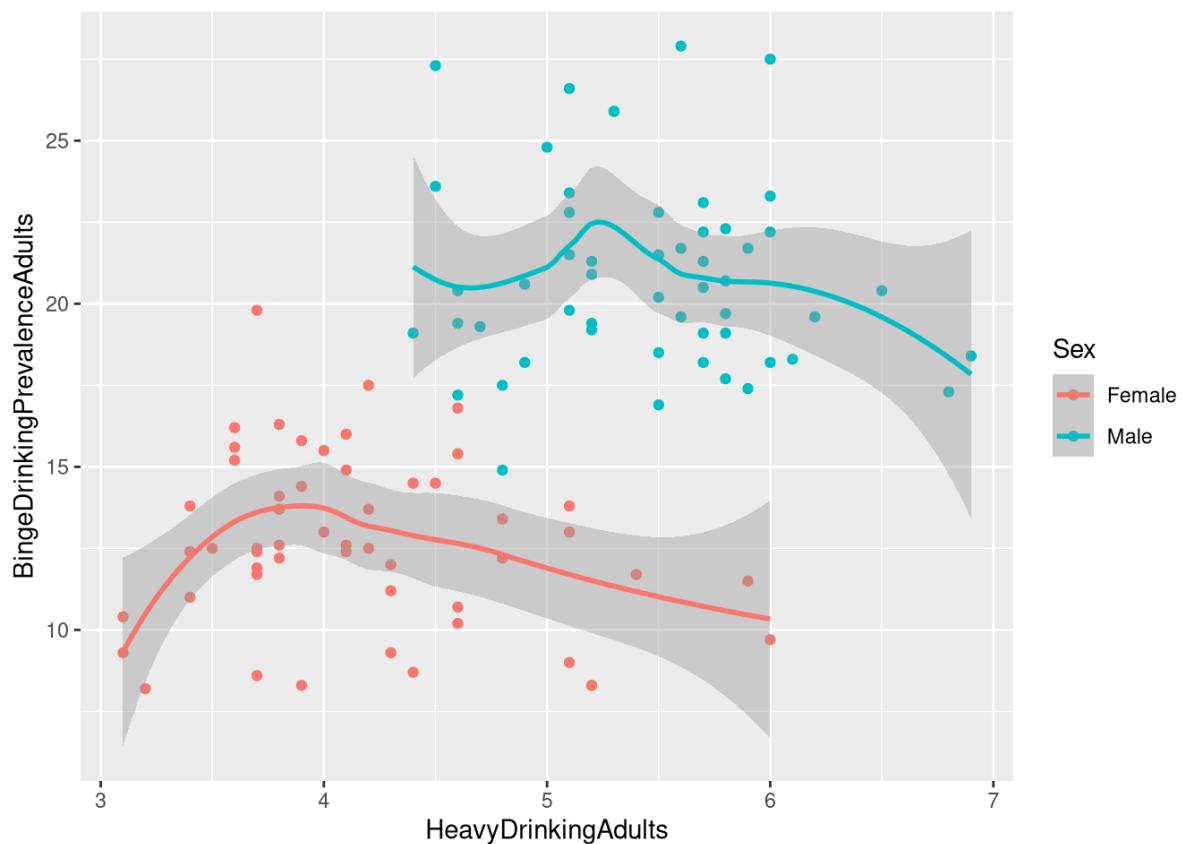
PercentageOfTotalDeaths y BingeDrinkingPrevalenceAdults



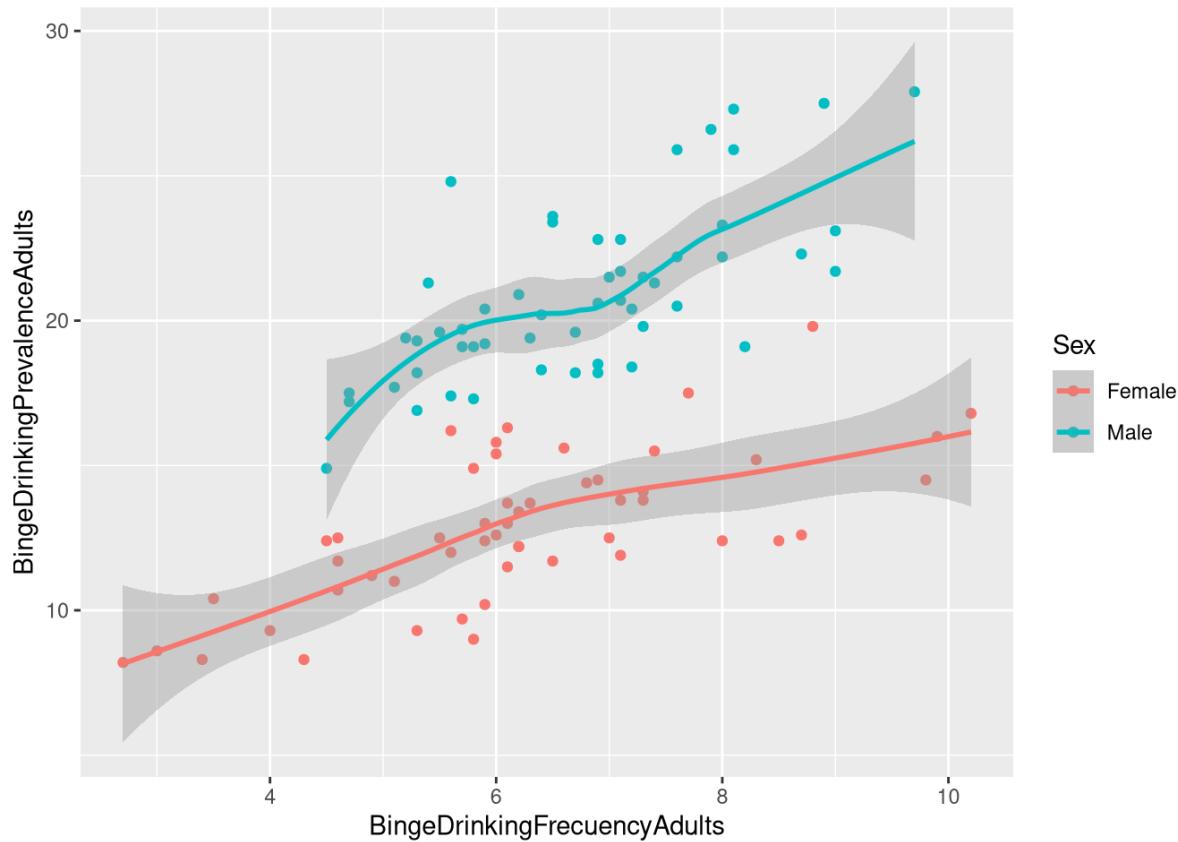
HeavyDrinkingAdults y BingeDrinkingIntensityAdults



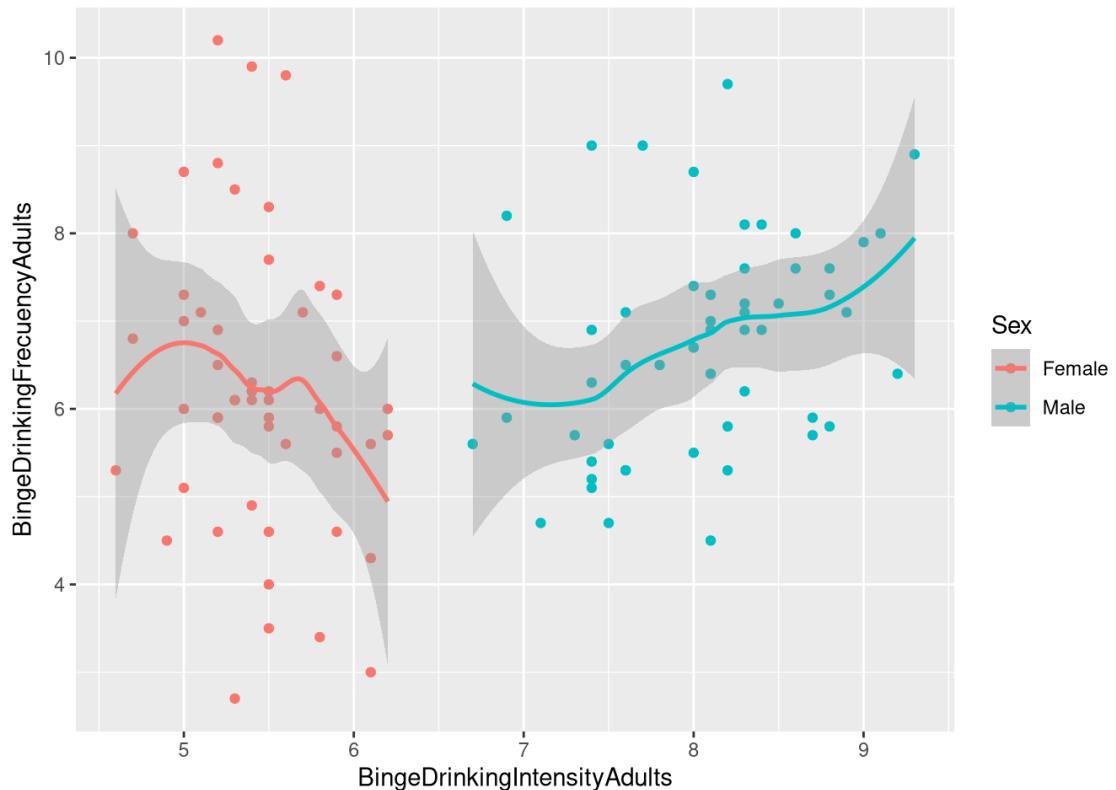
HeavyDrinkingAdults y BingeDrinkingPrevalenceAdults



BingeDrinkingFrequencyAdults y BingeDrinkingPrevalenceAdults



BingeDrinkingIntensityAdults y BingeDrinkingFrequencyAdults

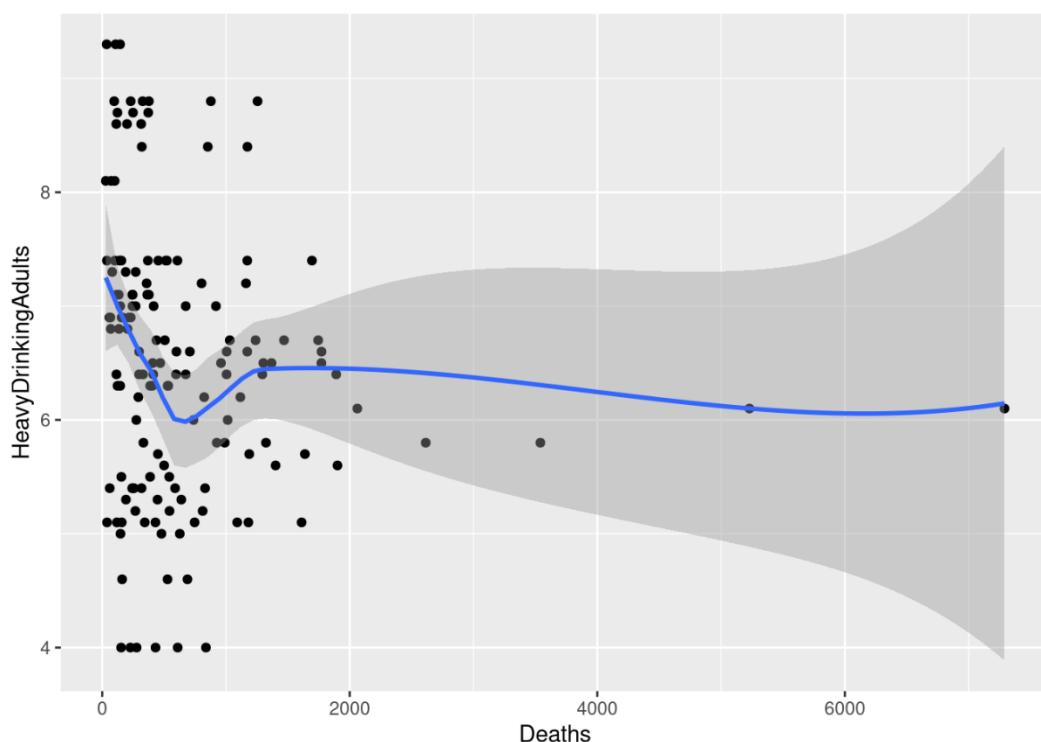


Objeto data_overall

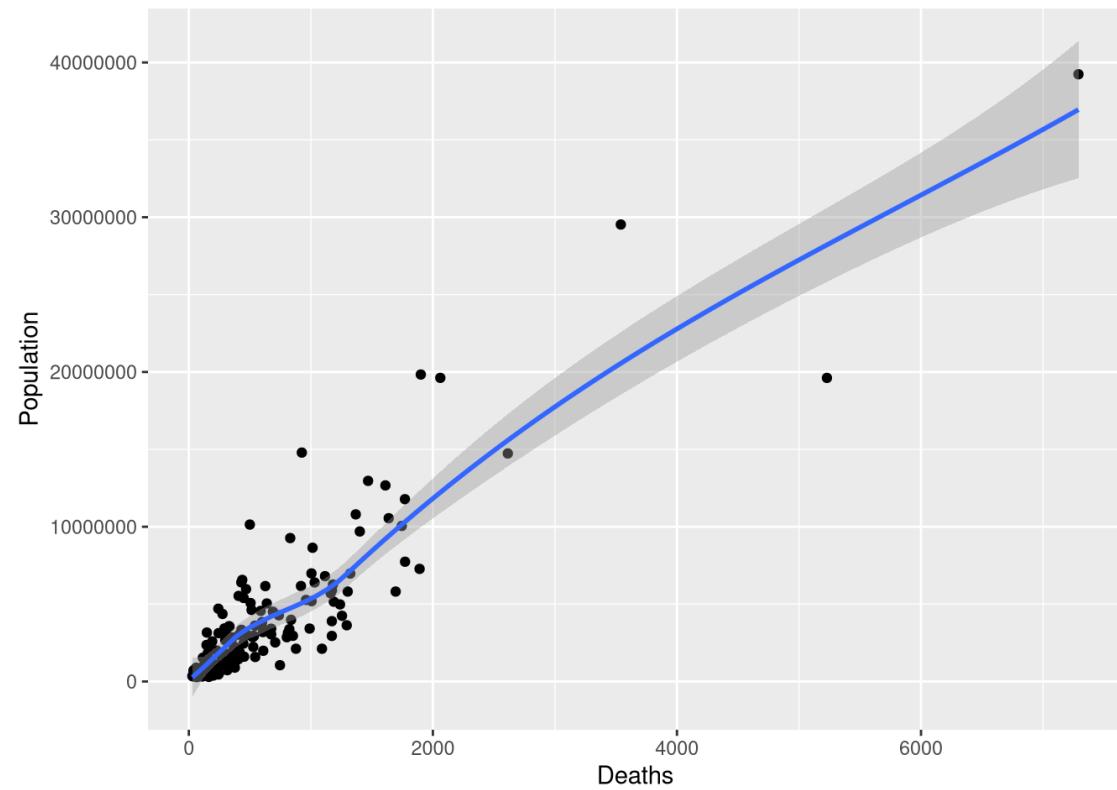
Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre:
 - Deaths y HeavyDrinkingAdults.
 - Population y AgeAdjustedDeathRate.
 - Population y HeavyDrinkingAdults.
- Una correlación positiva entre:
 - Deaths y Population.
 - Deaths y PercentageOfTotalDeaths.
 - Population y PercentageOfTotalDeaths.
 - AgeAdjustedDeathRate y HeavyDrinkingAdults.
 - PercentageOfTotalDeaths y HeavyDrinkingAdults.
 - HeavyDrinkingAdults y BingeDrinkingIntensityAdults.
 - BingeDrinkingFrecuencyAdults y BingeDrinkingPrevalenceAdults.

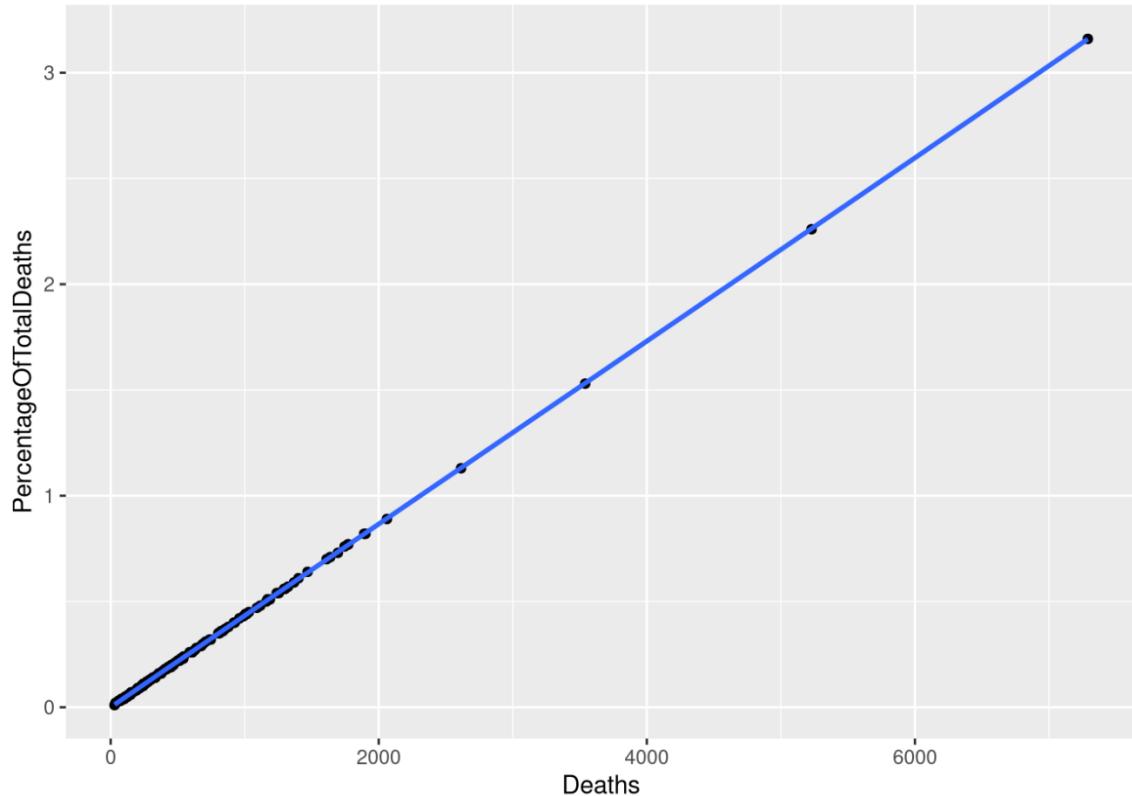
Deaths y HeavyDrinkingAdults



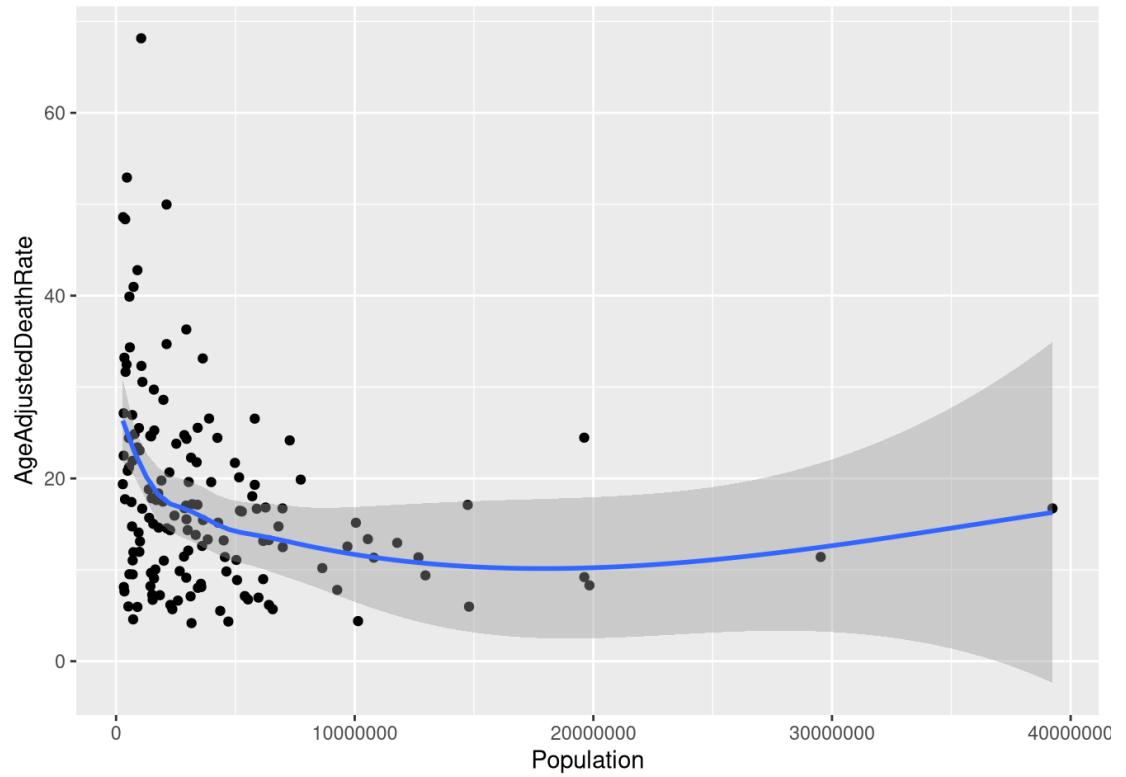
Deaths y Population



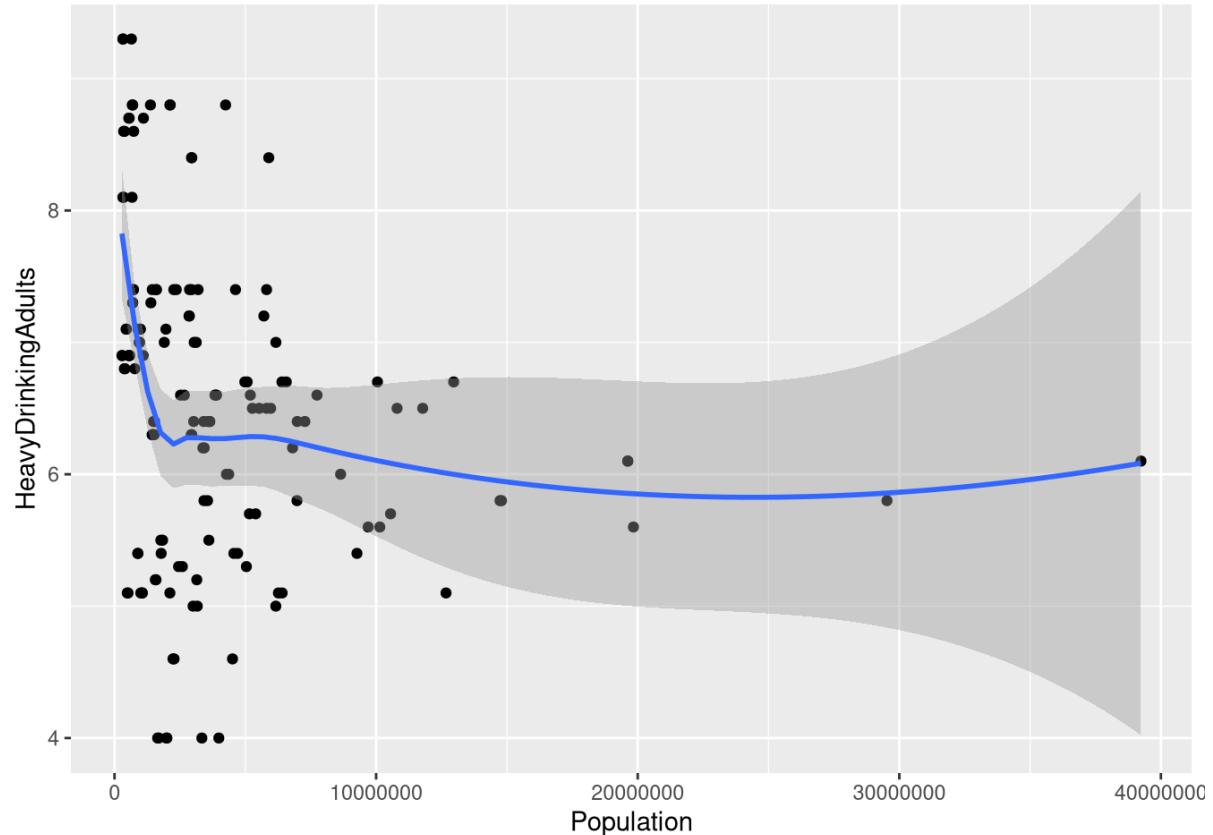
Deaths y PercentageOfTotalDeaths



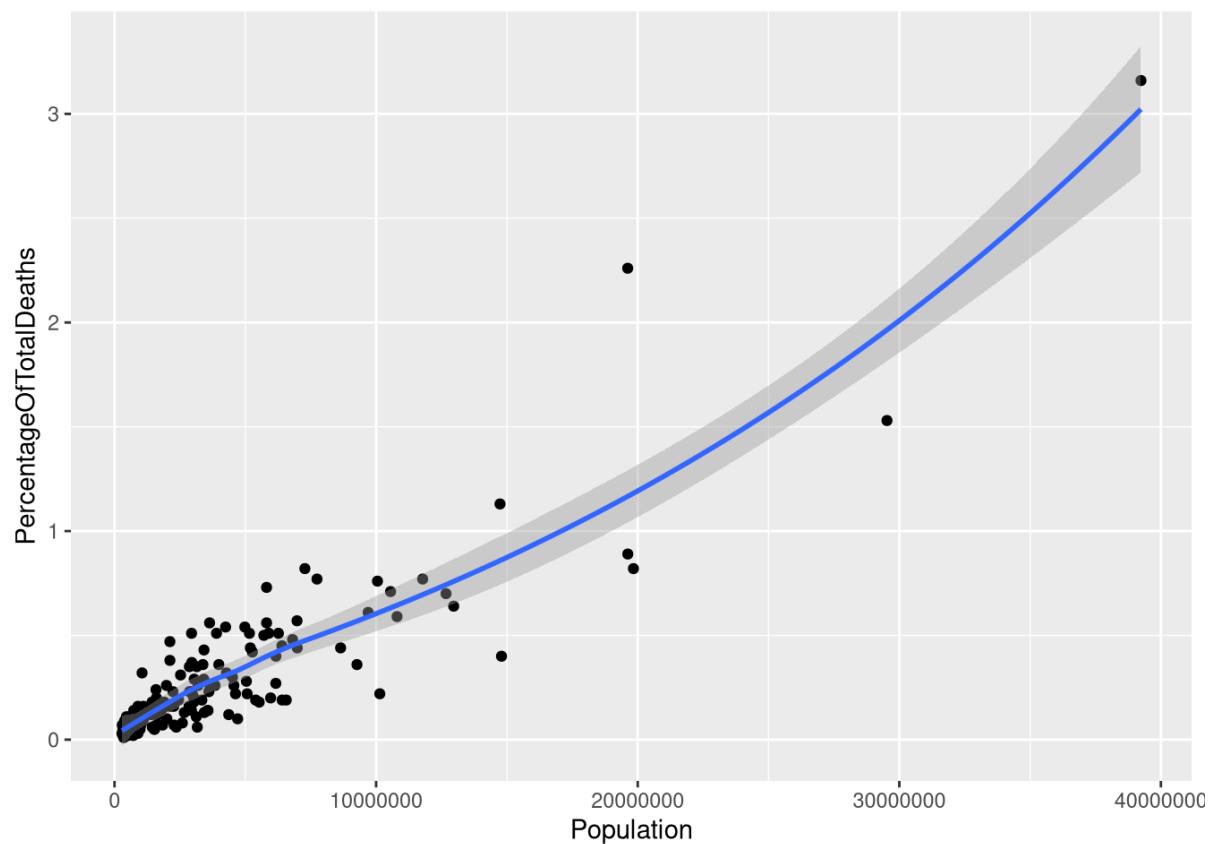
Population y AgeAdjustedDeathRate



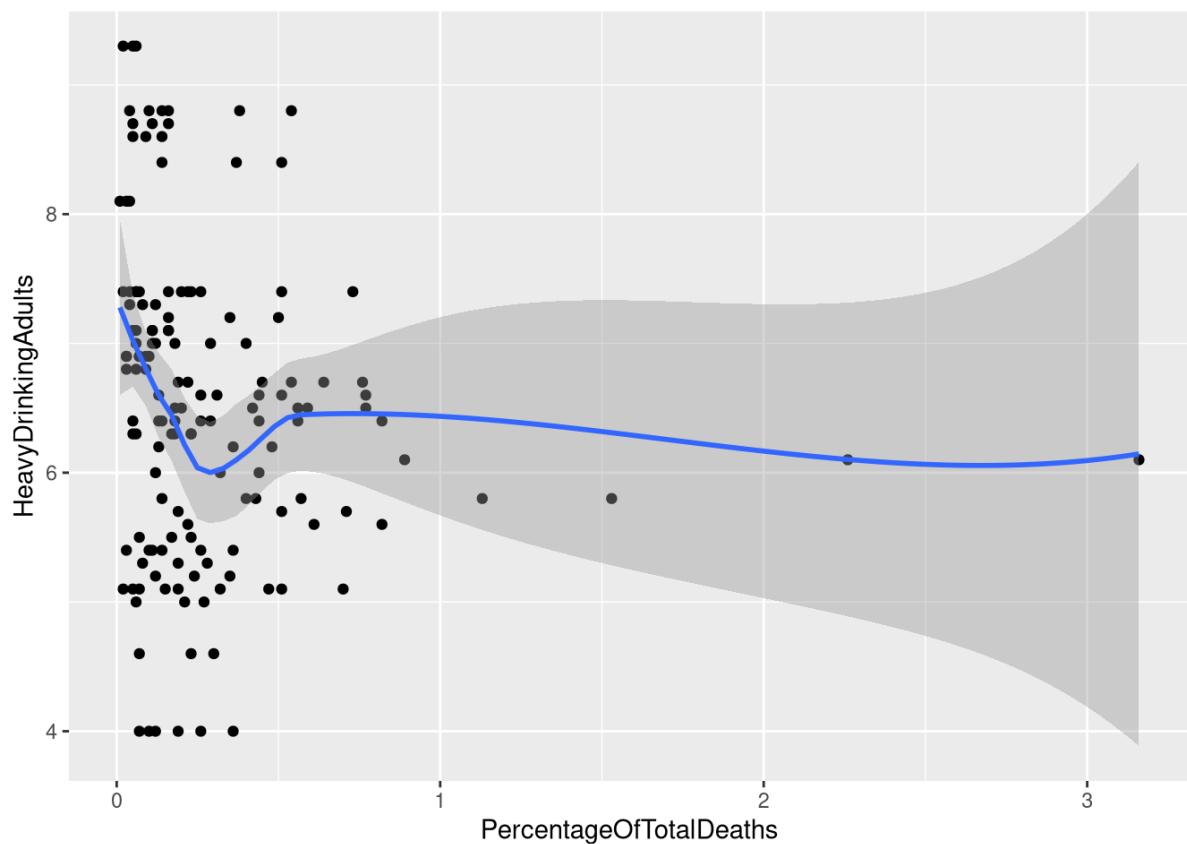
Population y HeavyDrinkingAdults



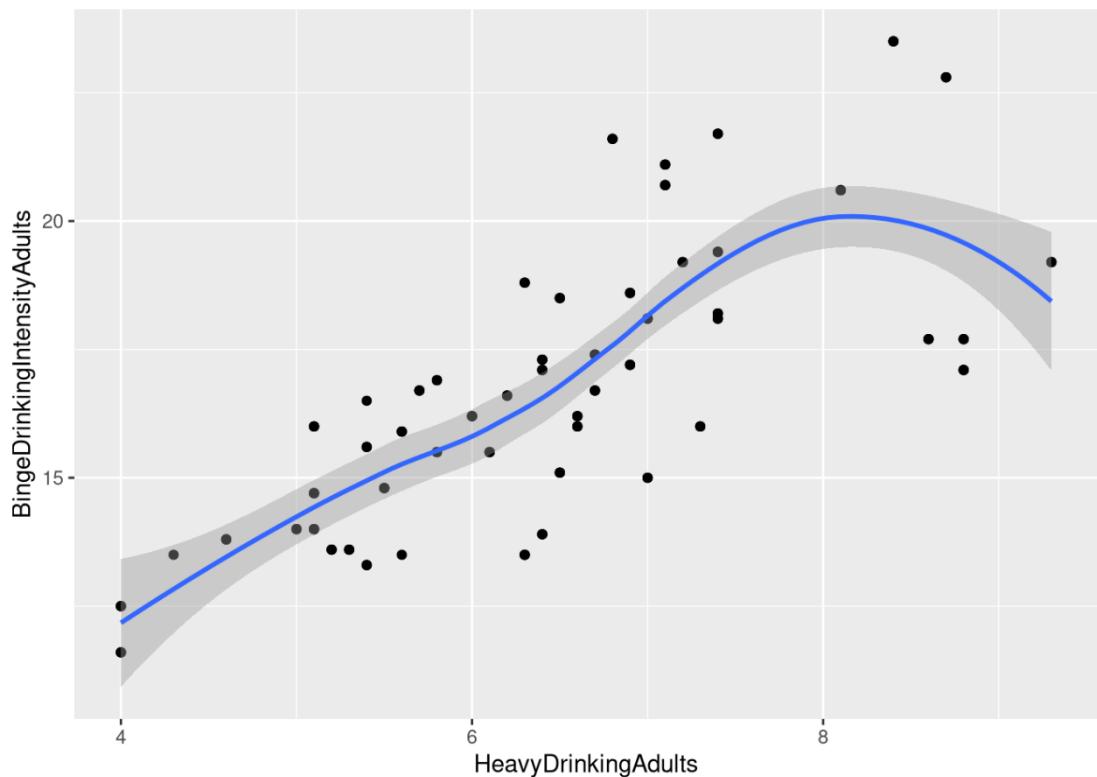
Population y PercentageOfTotalDeaths



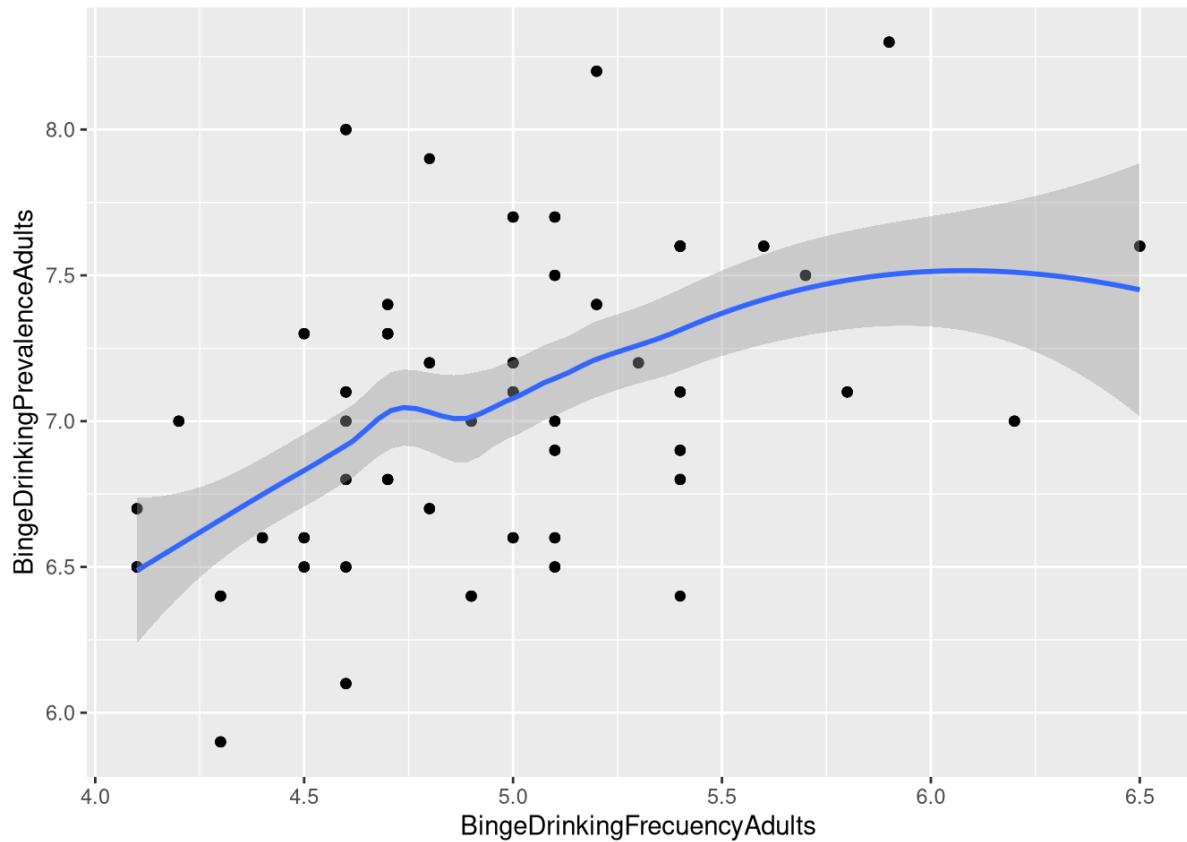
PercentageOfTotalDeaths y HeavyDrinkingAdults



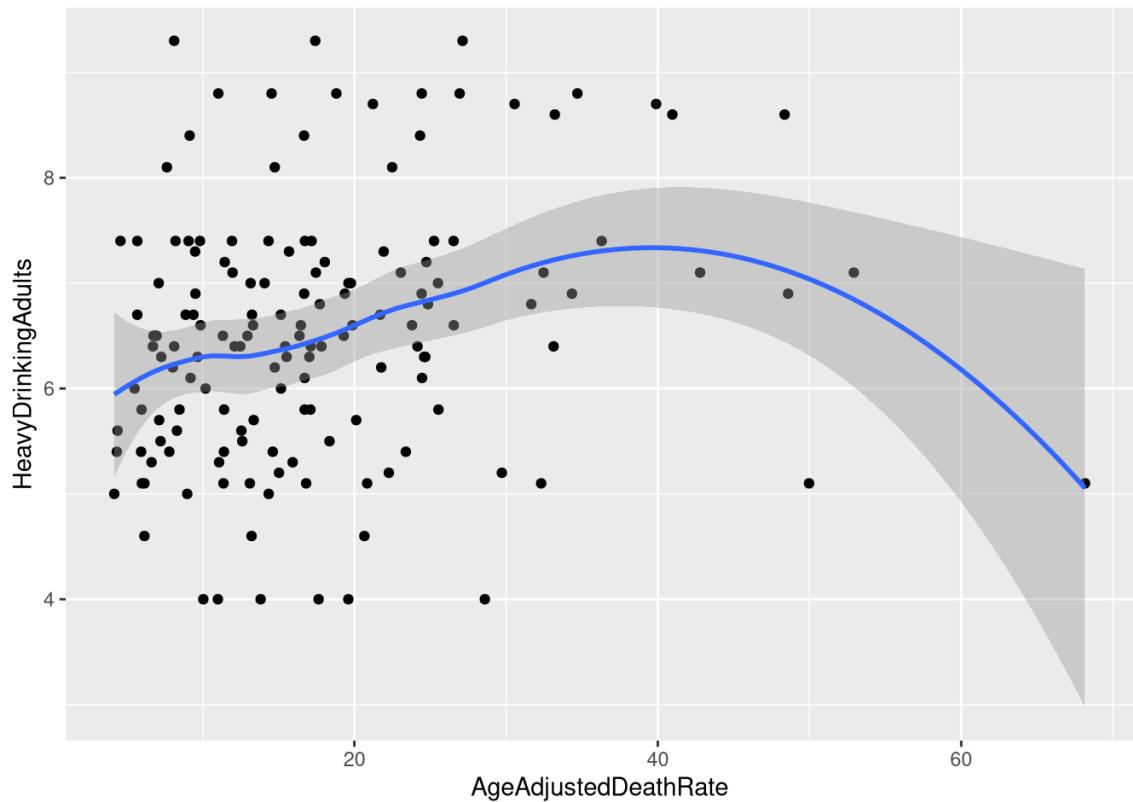
HeavyDrinkingAdults y BingeDrinkingIntensityAdults



BingeDrinkingFrequencyAdults y BingeDrinkingPrevalenceAdults



AgeAdjustedDeathRate y HeavyDrinkingAdults



03i - Análisis de valores faltantes (NA's) y outliers

Se ejecutaron las siguientes acciones

- 03ia - Análisis de valores faltantes NAs
- 03ib - Exploración de Outliers

Funciones de R utilizadas en el análisis de datos faltantes (NA's)

Para la evaluación de datos faltantes (NA's) se evaluaron las siguientes dimensiones:

Dimensión	Función	Resultado evaluación
Existencia de algún valor valor faltante NA (sí/no)	naniar::any_na()	TRUE
Número total de NAs	naniar::n_miss()	
Variables afectadas por la presencia de NAs	is.na() > colSums()	Sex, Deaths, Population, AgeAdjustedDeathRate, PercentageOfTotalDeaths, HeavyDrinkingAdults, BingeDrinkingFrequencyAdults, BingeDrinkingIntensityAdults y BingeDrinkingPrevalenceAdults
Número de NA por variable (n y %)	naniar::miss_var_summary() naniar::miss_var_table()	Número: Rango Porcentaje: Rango
Número de NA por observación (n y %)	naniar::miss_case_summary() naniar::miss_case_table()	Número: Rango Porcentaje: Rango
Ranking de variables más afectadas por NA's	naniar::gg_miss_var()	Deaths, Population, PercentageOfTotalDeaths, AgeAdjustedDeathRate
Tipología de los valores faltantes (MAR, MNAR, MCAR)	naniar::vis_miss()	Los valores faltantes se concentran en las últimas observaciones del dataset (MNAR)
Relación entre valores faltantes de distintas variables	naniar::gg_miss_upset()	Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: Deaths, Population, AgeAdjustedDeathRate y PercentageOfTotalDeaths

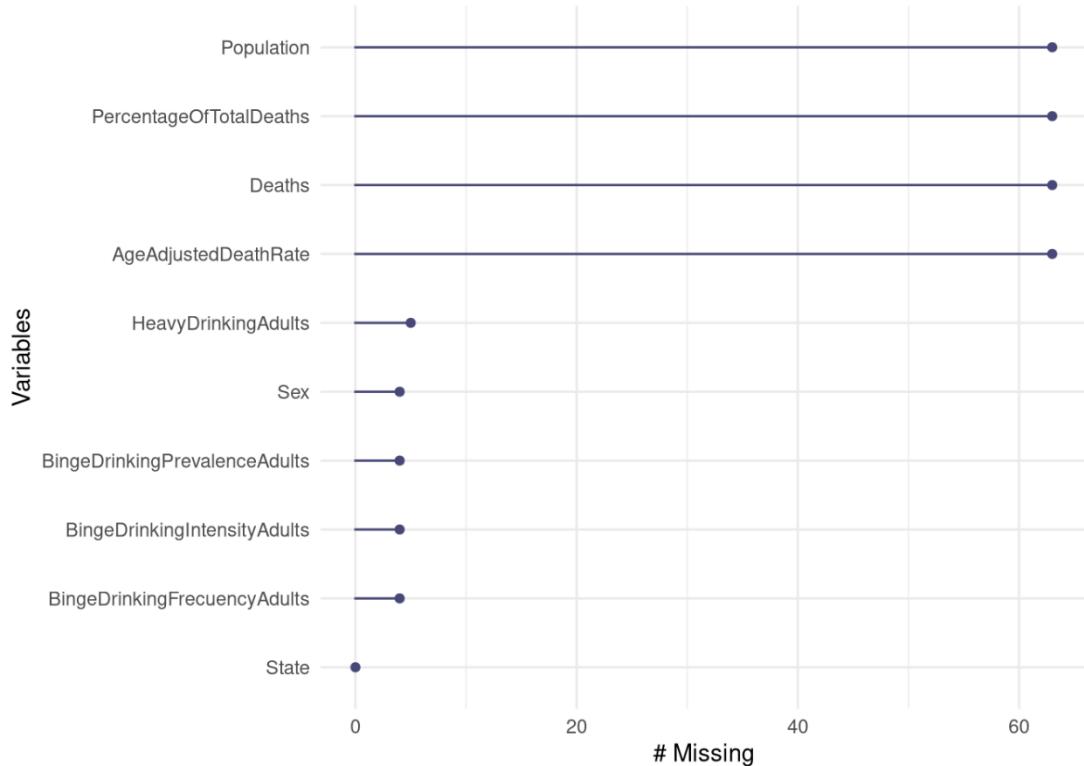
Relación entre valores faltantes y niveles de las variables categóricas	<code>naniar::gg_miss_fct()</code>	Algunos estados concentran todos los valores faltantes. No hay influencia del sexo
---	------------------------------------	--

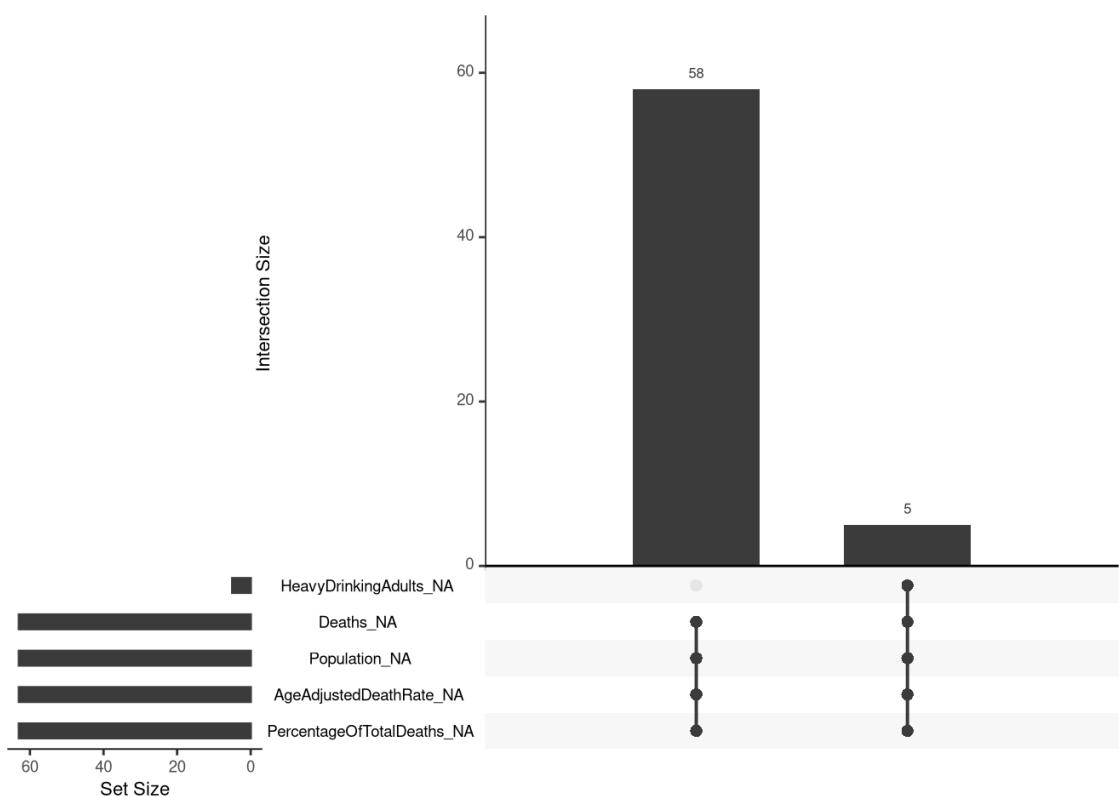
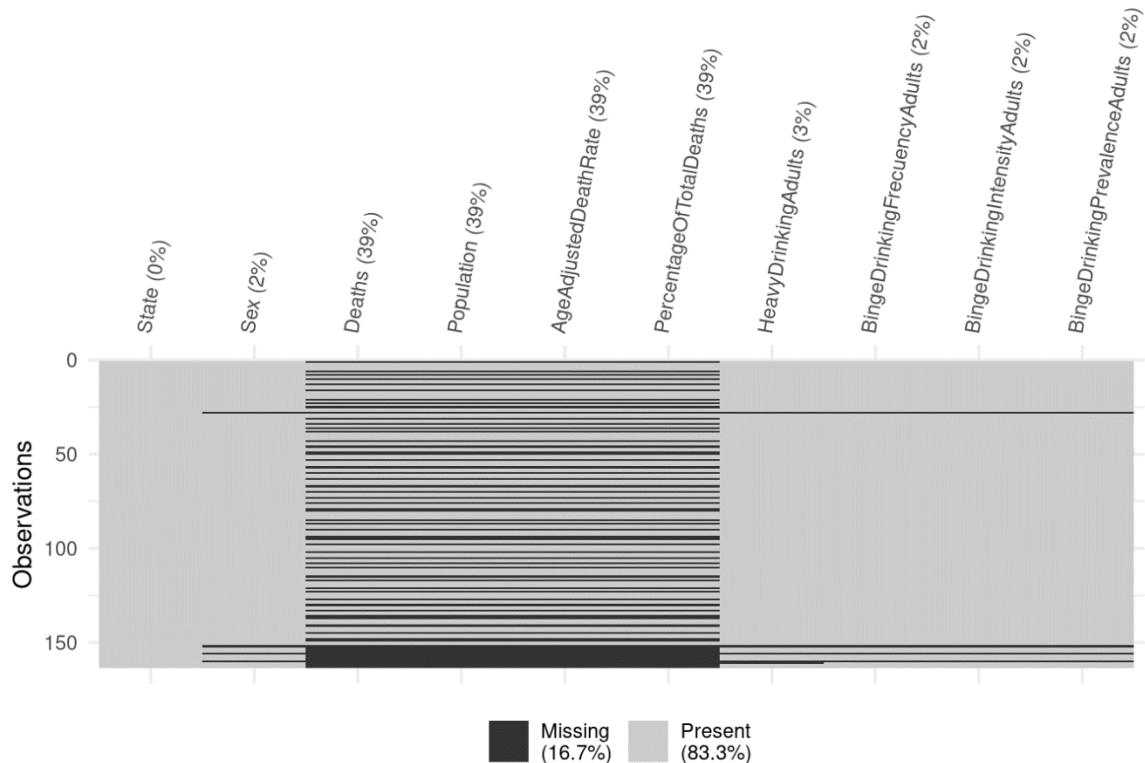
Objeto data

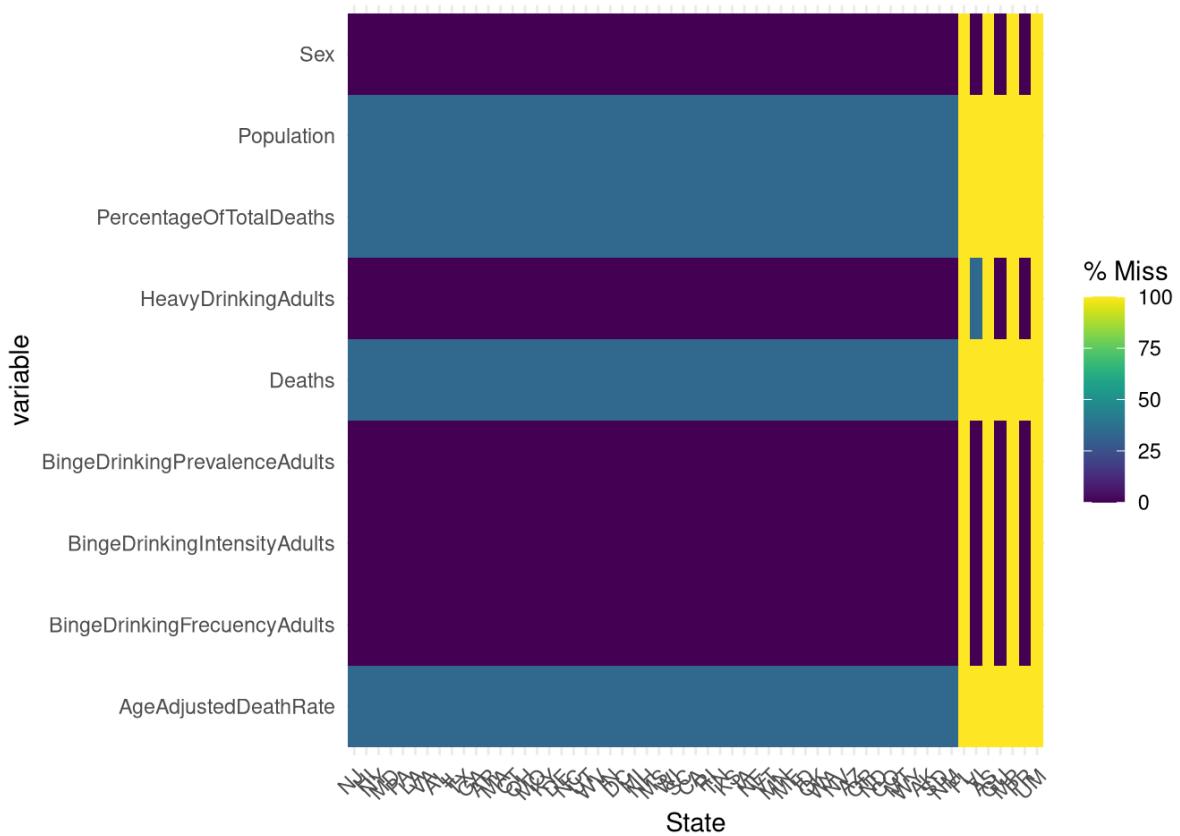
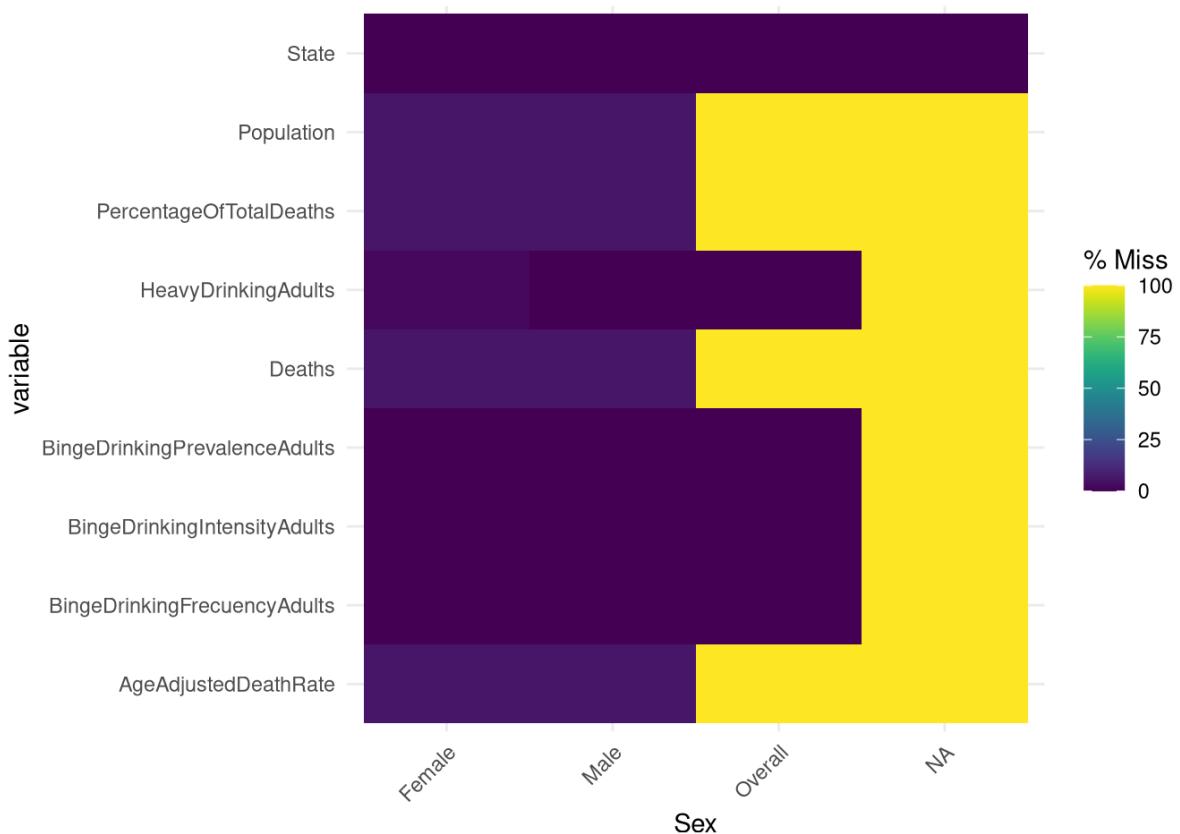
03ia - Análisis de valores faltantes NA's

Hallazgos:

- Los valores faltantes se concentran en las últimas observaciones del dataset , para estados concretos (*Missing not at random*, MNAR).
- Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: Deaths, Population, AgeAdjustedDeathRate y PercentageOfTotalDeaths
- Algunos estados concentran todos los valores faltantes. No hay influencia del sexo.



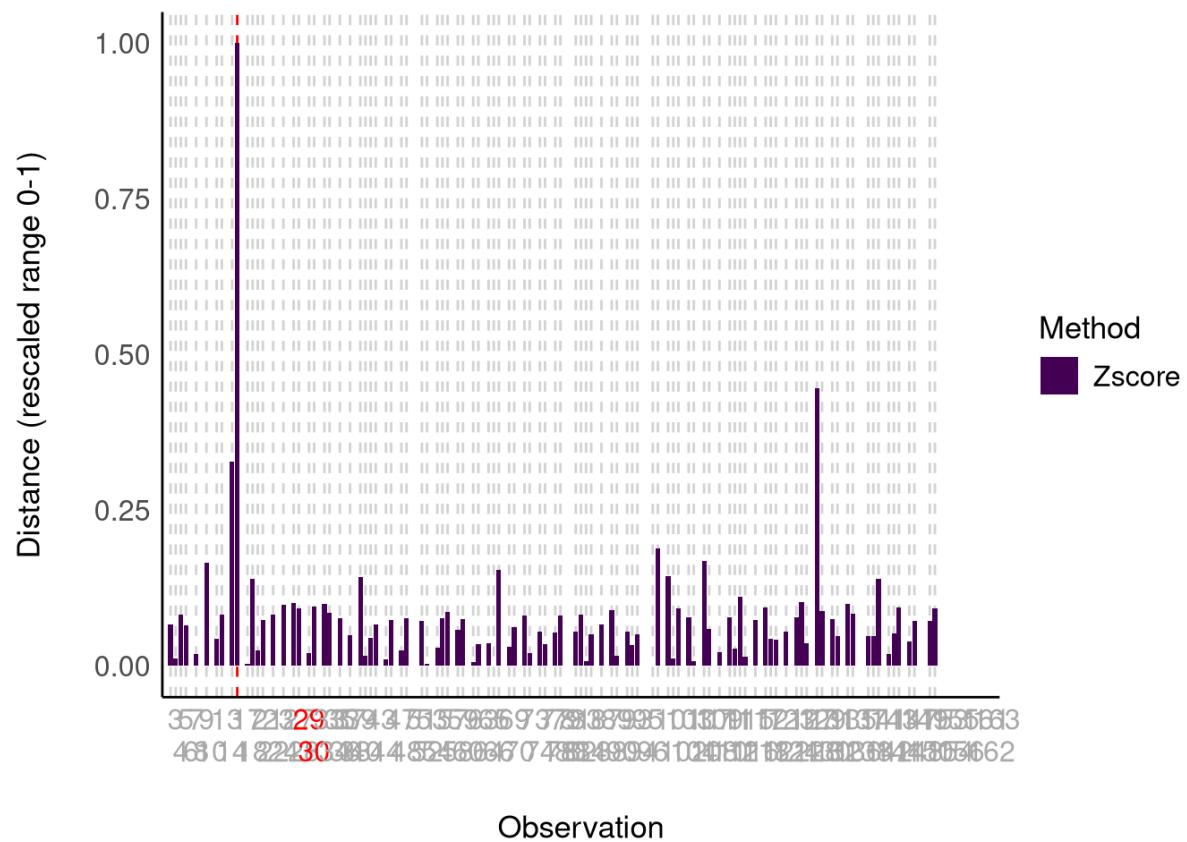




03ib - Exploración de Outliers

Existe un número significativo de outliers en el *data.frame*. Eso deberá tenerse en cuenta para el análisis cluster.

variables	outliers_cnt	outliers_ratio	outliers_mean
Deaths	6	3.680982	2,316.833333
Population	6	3.680982	14,766,948.333333
AgeAdjustedDeathRate	4	2.453988	54.507500
PercentageOfTotalDeaths	6	3.680982	1.001667
HeavyDrinkingAdults	4	2.453988	6.450000
BingeDrinkingFrequencyAdults	0	0.000000	
BingeDrinkingIntensityAdults	0	0.000000	
BingeDrinkingPrevalenceAdults	0	0.000000	



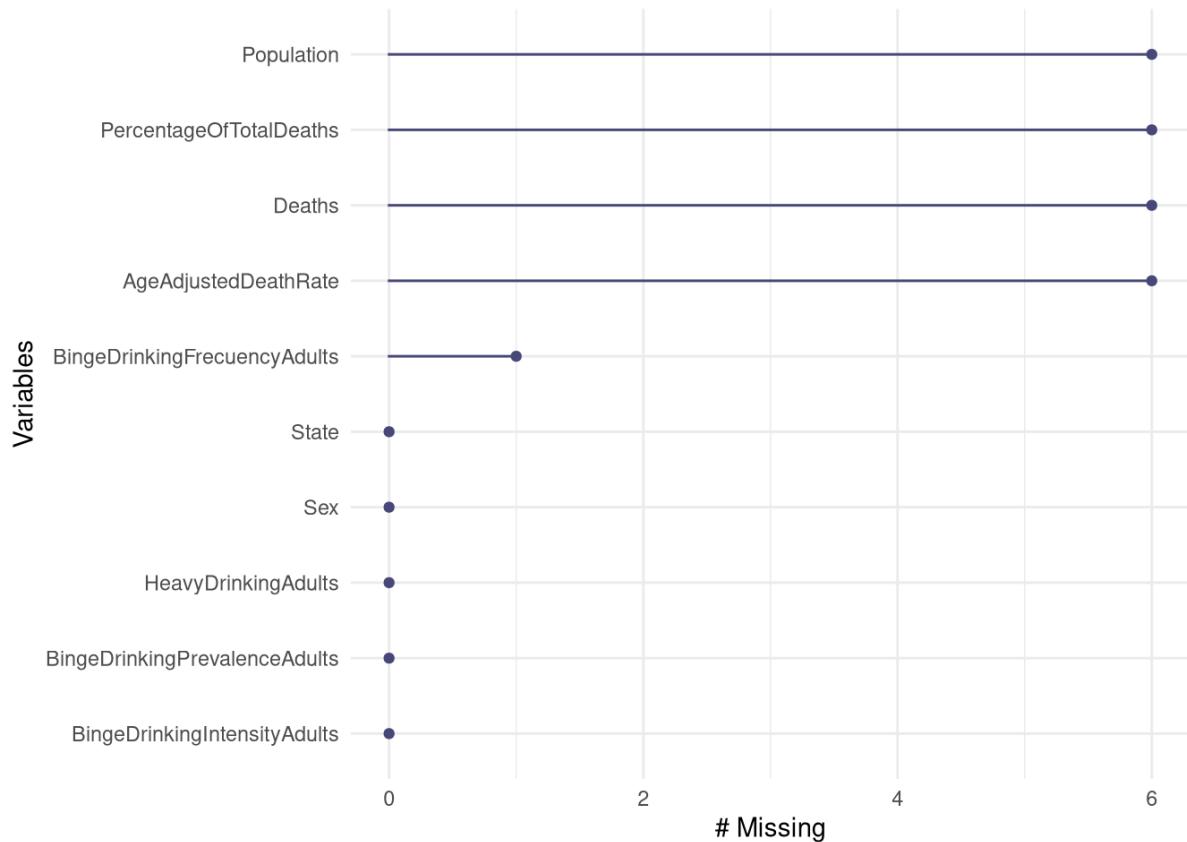
Ninguna variable tiene más de un 5% de outliers en sus valores.

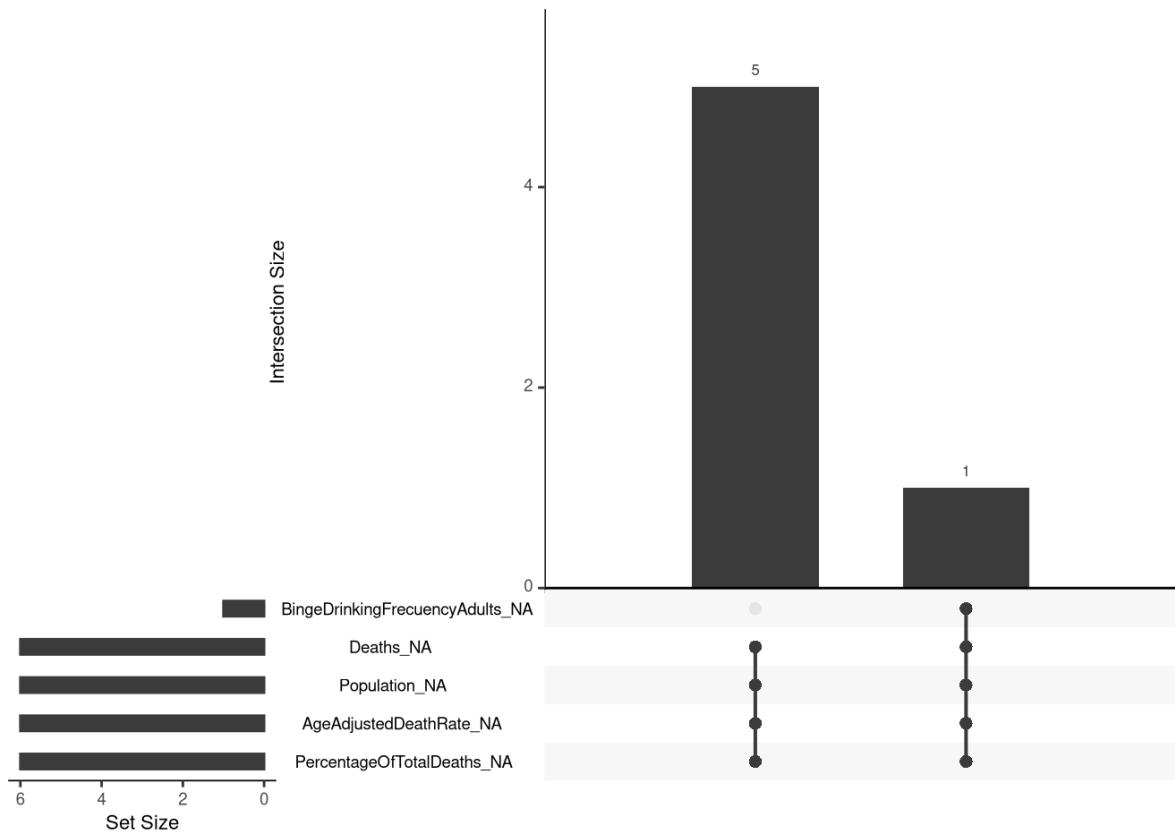
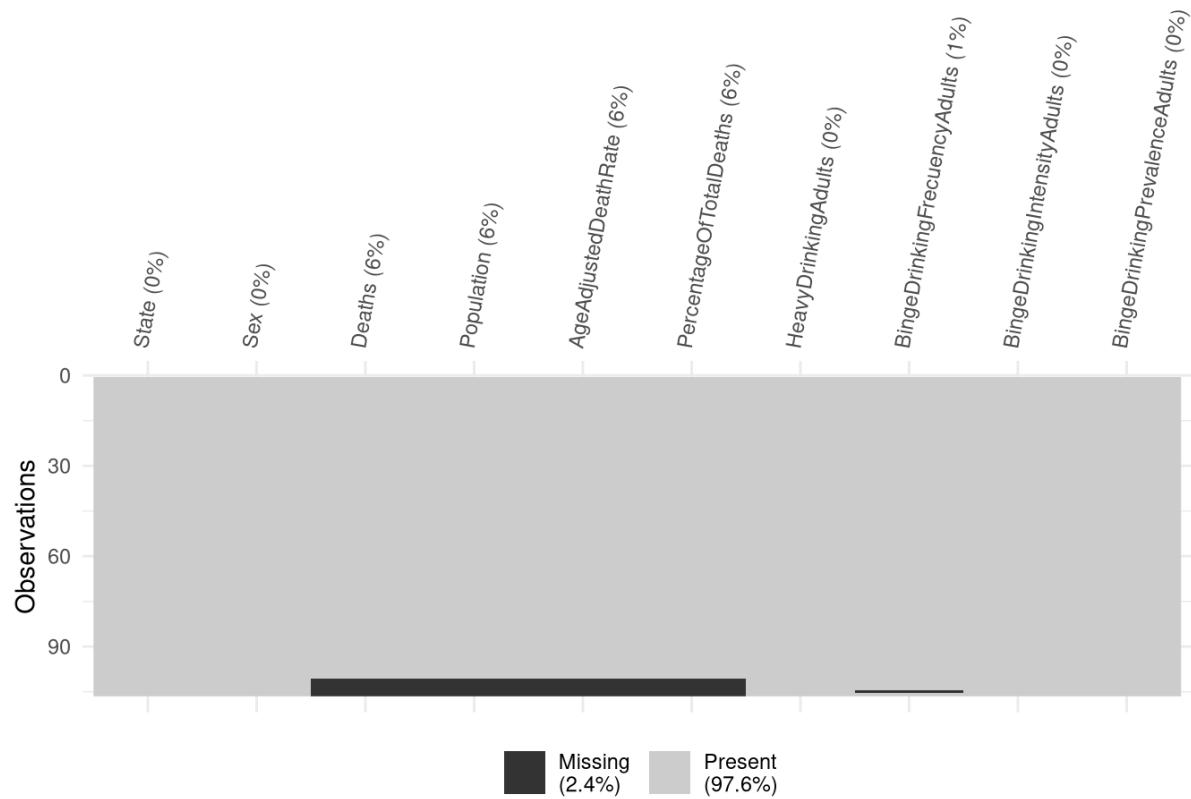
Objeto data_gender

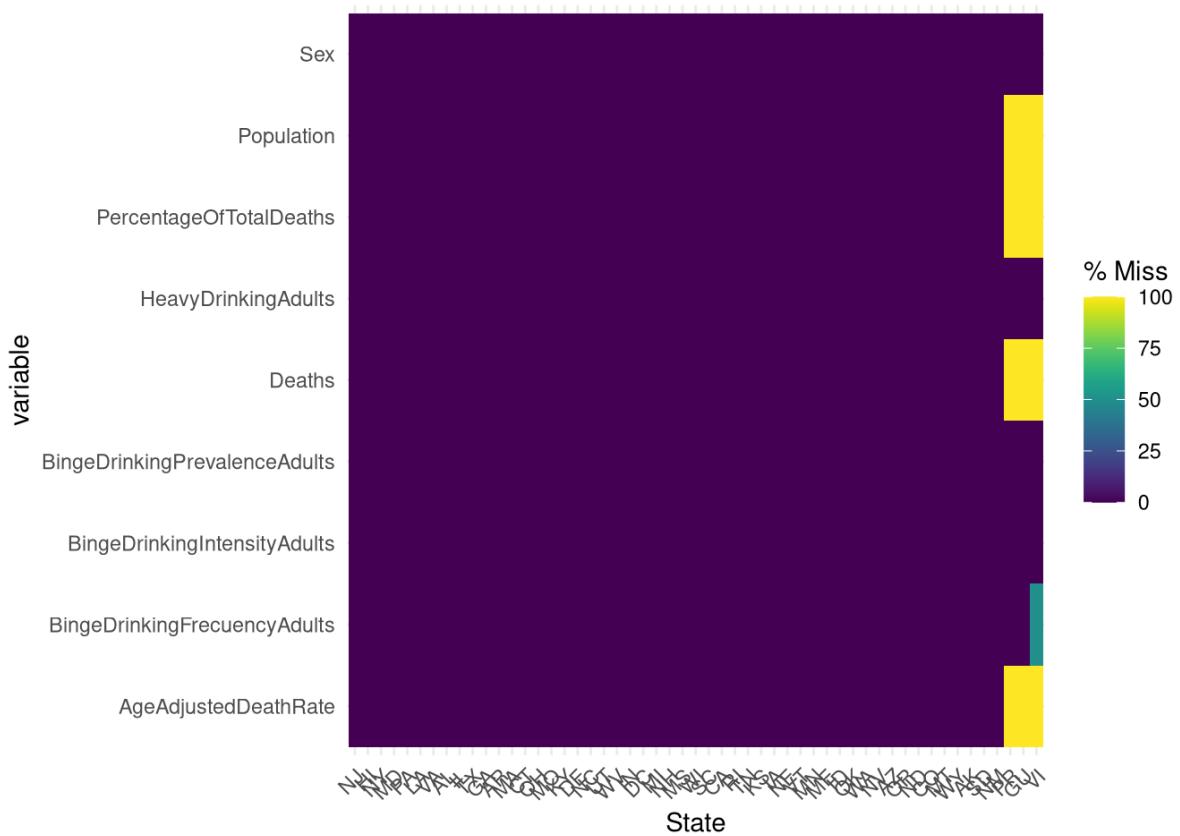
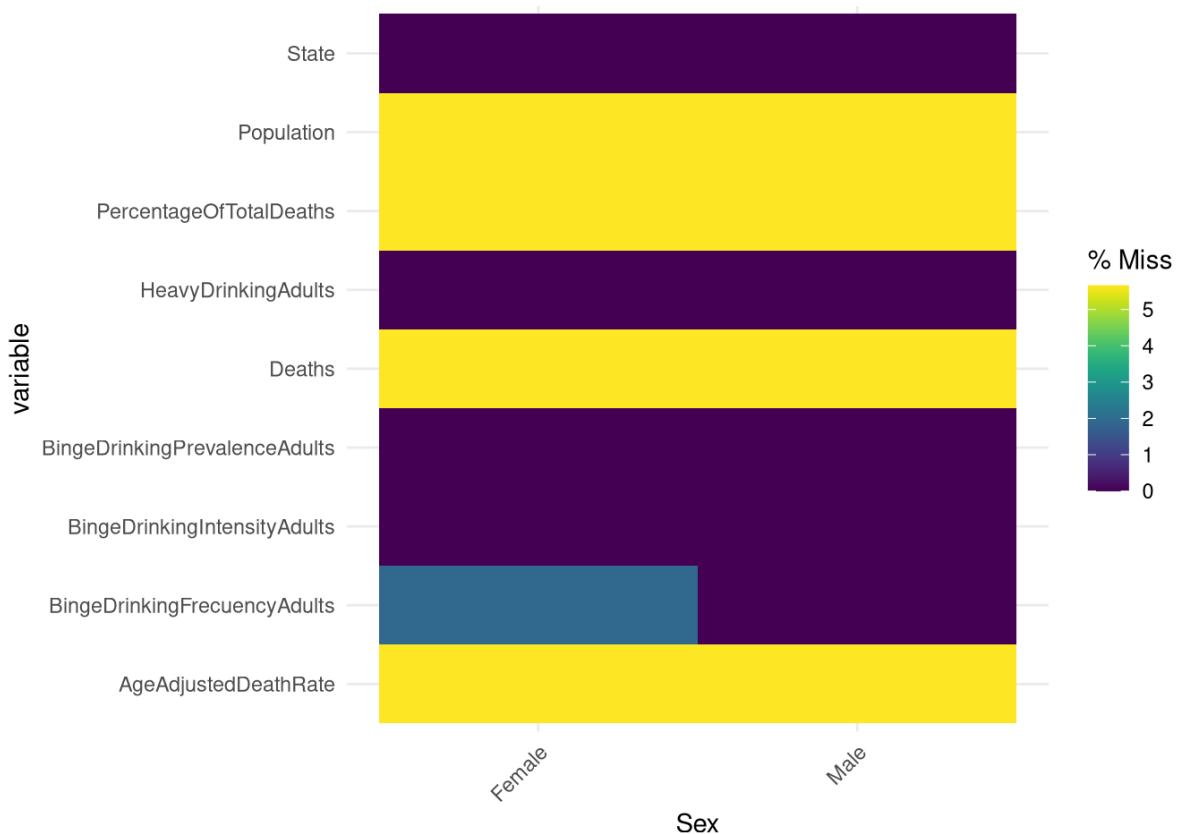
03ia - Análisis de valores faltantes NA's

Hallazgos:

- Los valores faltantes se concentran en las últimas observaciones del dataset , para estados concretos (*Missing not at random*, MNAR).
- Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: Deaths, Population, AgeAdjustedDeathRate y PercentageOfTotalDeaths
- Algunos estados concentran todos los valores faltantes. No hay influencia del sexo.



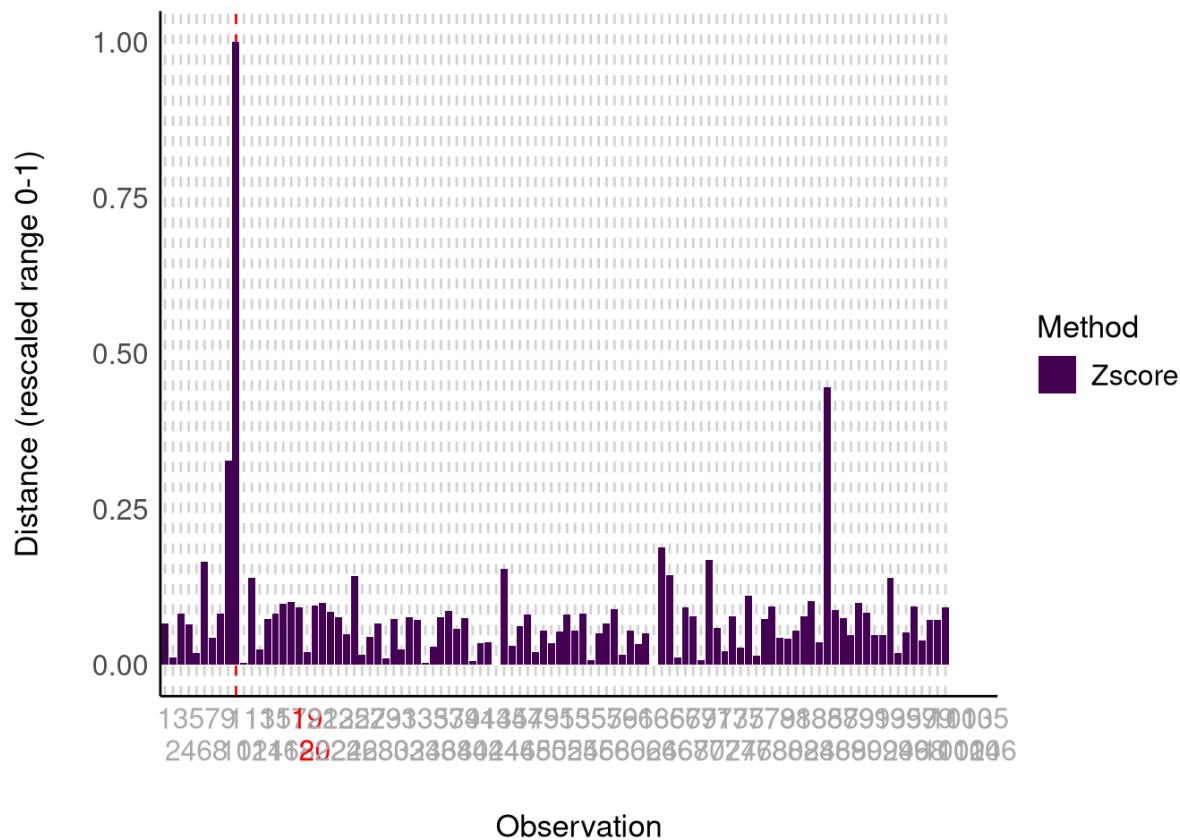




03ib - Exploración de Outliers

Existe un número significativo de outliers en el *data.frame*. Eso deberá tenerse en cuenta para el análisis cluster.

variables	outliers_cnt	outliers_ratio	outliers_mean
Deaths	6	5.660377	2,316.833333
Population	6	5.660377	14,766,948.333333
AgeAdjustedDeathRate	4	3.773585	54.507500
PercentageOfTotalDeaths	6	5.660377	1.001667
HeavyDrinkingAdults	0	0.000000	
BingeDrinkingFrequencyAdults	4	3.773585	6.450000
BingeDrinkingIntensityAdults	0	0.000000	
BingeDrinkingPrevalenceAdults	0	0.000000	



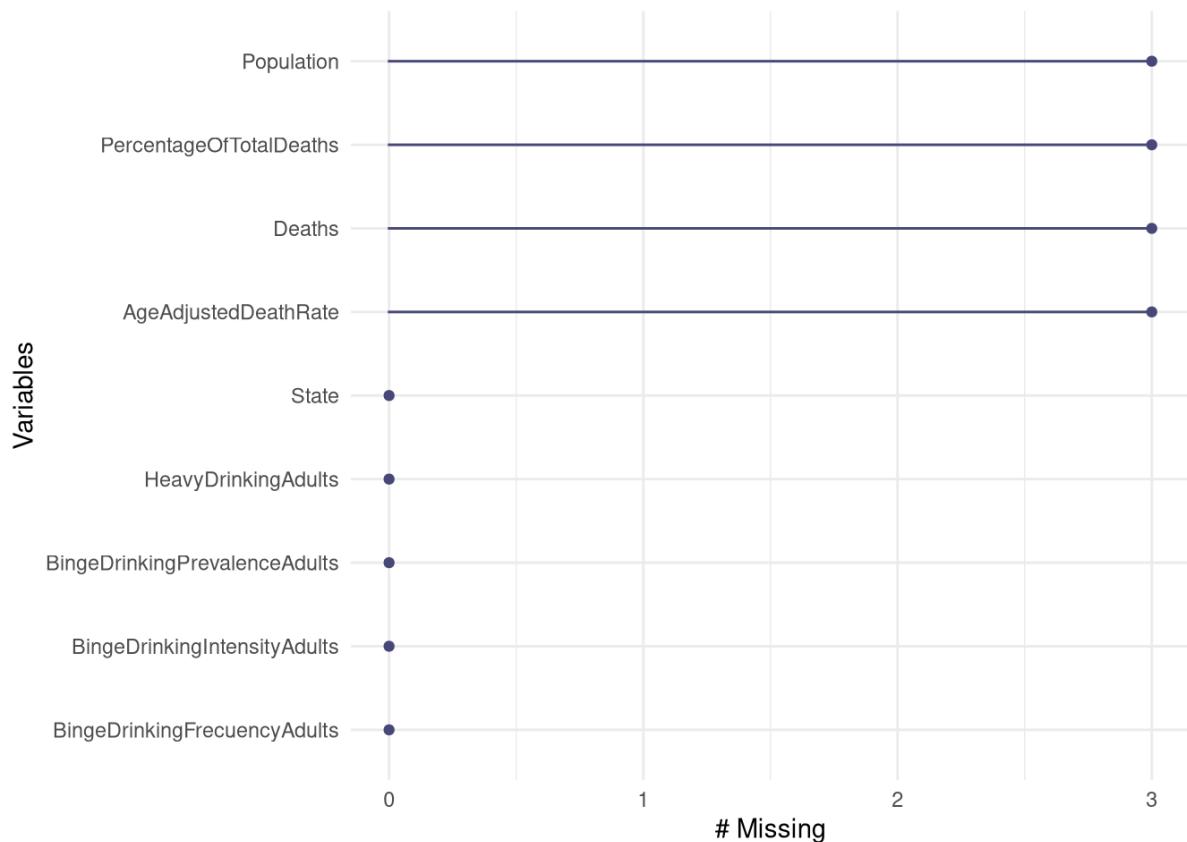
Ninguna variable tiene más de un 5% de outliers en sus valores.

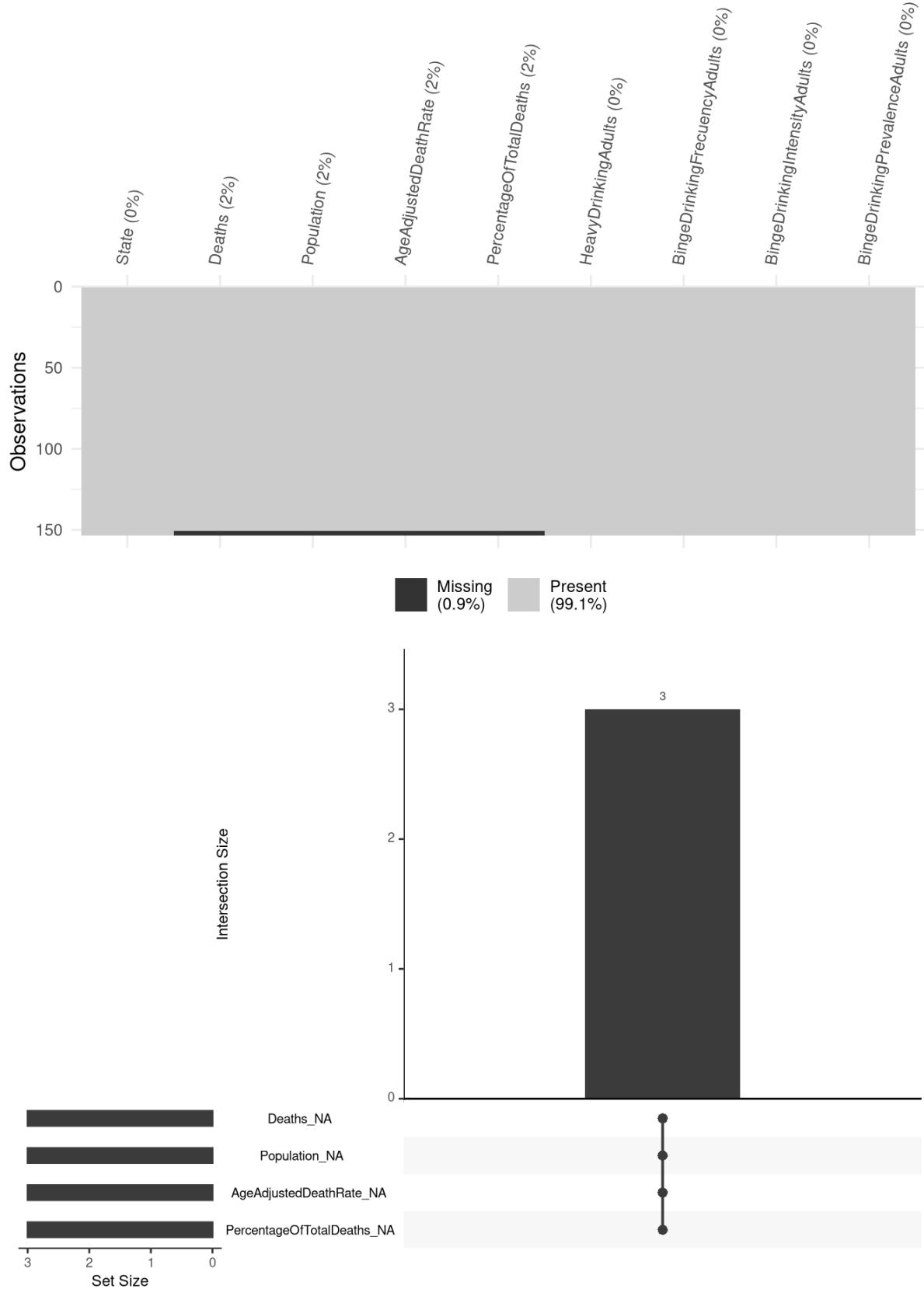
Objeto data_overall

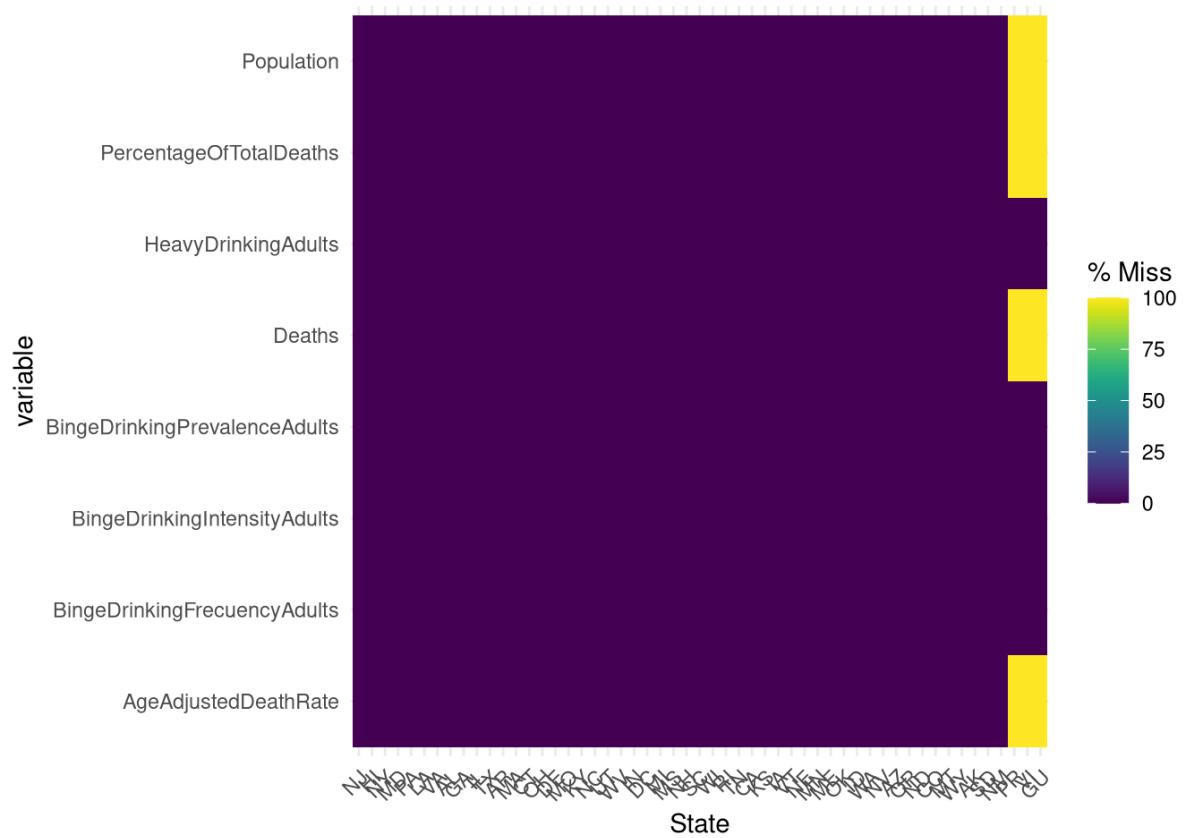
03ia - Análisis de valores faltantes NA's

Hallazgos:

- Los valores faltantes se concentran en las últimas observaciones del dataset , para estados concretos (*Missing not at random*, MNAR).
- Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: Deaths, Population, AgeAdjustedDeathRate y PercentageOfTotalDeaths
- Algunos estados concentran todos los valores faltantes. No hay influencia del sexo.



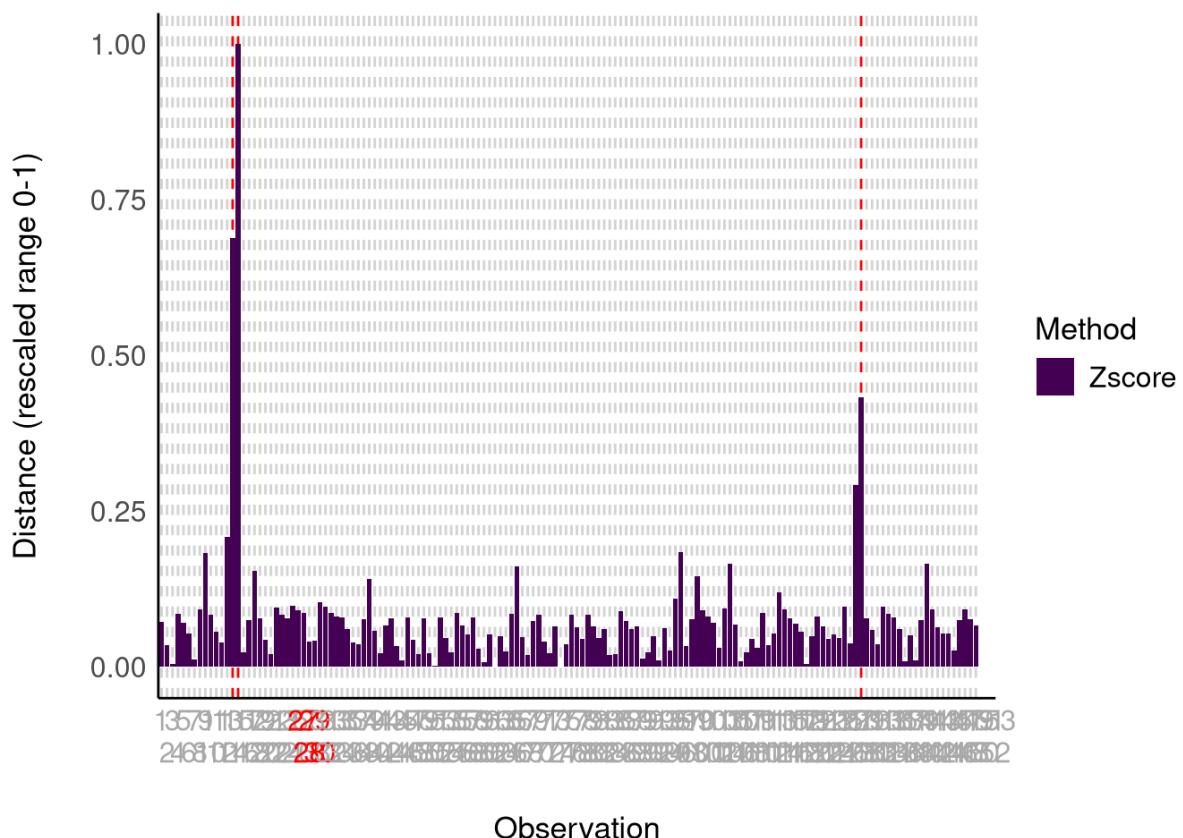




03ib - Exploración de Outliers

Existe un número significativo de outliers en el *data.frame*. Eso deberá tenerse en cuenta para el análisis cluster.

variables	outliers_cnt	outliers_ratio	outliers_mean
Deaths	6	3.921569	3,772.500000
Population	10	6.535948	19,478,300.900000
AgeAdjustedDeathRate	5	3.267974	53.600000
PercentageOfTotalDeaths	7	4.575163	1.515714
HeavyDrinkingAdults	0	0.000000	
BingeDrinkingFrequencyAdults	3	1.960784	6.500000
BingeDrinkingIntensityAdults	0	0.000000	
BingeDrinkingPrevalenceAdults	0	0.000000	



Ninguna variable tiene más de un 5% de outliers en sus valores.

Anexo 4 - Transformación

Descripción del subprocesso

Subproceso destinado a convertir los datos crudos ya limpiados al formato o estructura que requiere el tipo de análisis que se va a realizar en nuestros datos.

Code

Acciones del subprocesso

Se realizaron las siguientes tareas de transformación:

- 4a - Tratamiento de valores faltantes
- 4b - Tratamiento de valores atípicos (*outliers*)

4a - Tratamiento de valores faltantes

Se creó un dataset de trabajo sin datos faltantes, una para cada dataset de interés.

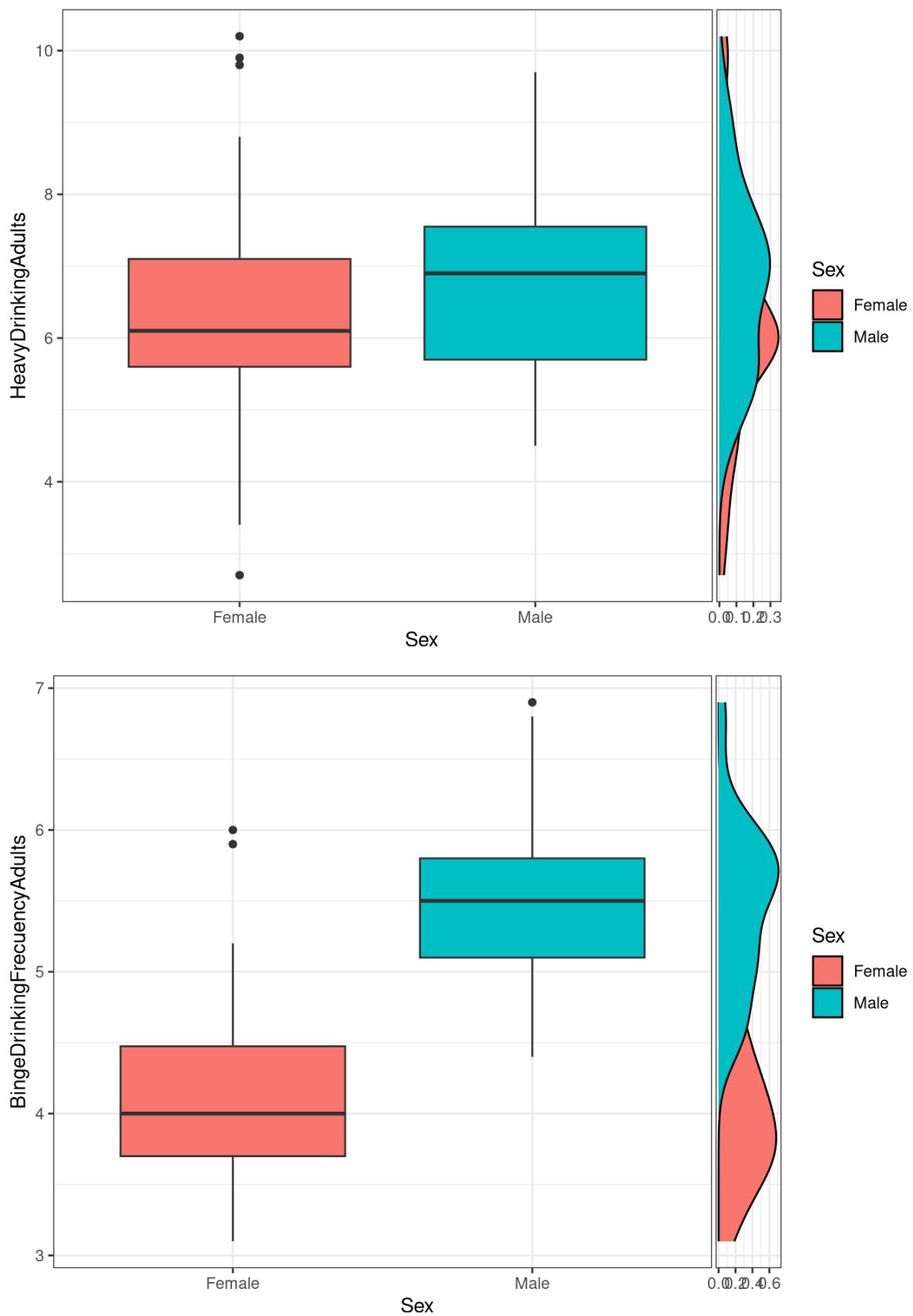
Tras la omisión de NA's, los dos datasets `data_lab` y `data_gender_lab` son idénticos, y sólo difieren en los atributos que se han ido creando durante el proceso de limpieza. Por tanto, podemos trabajar exclusivamente con `data_lab` (para datos por sexos) y `data_overall` (para datos globales):

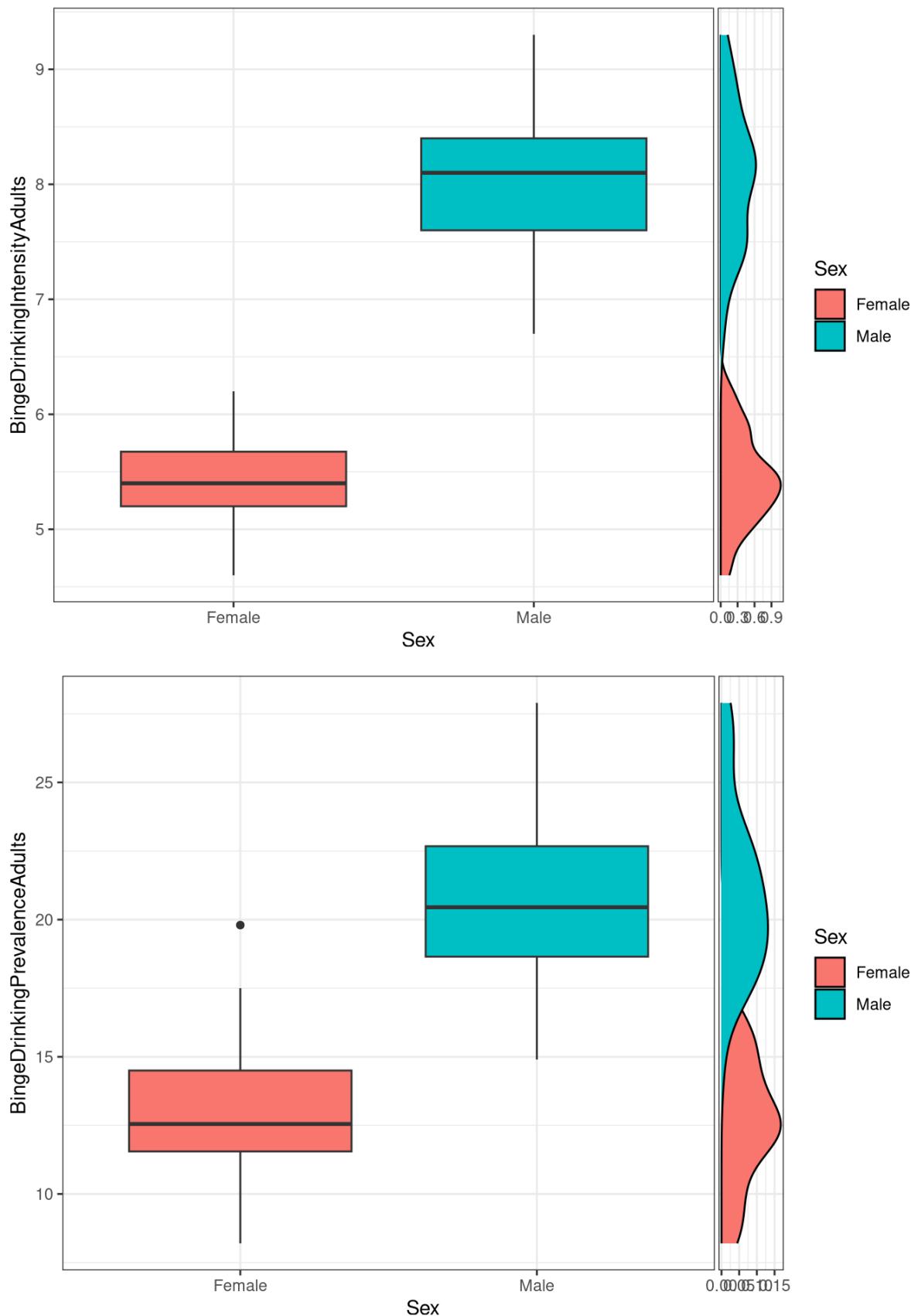
4b - Tratamiento de valores atípicos (outliers)

Las observaciones con valores extremos para las variables estudiadas podrían ser muy interesantes para nuestro análisis, porque pueden contener información sobre los factores de riesgo más asociados a la mortalidad por alcohol.

1- Objeto `data_lab`

En el subprocesso de EDA se identificaron problemas de valores atípicos en cinco variables del objeto `data`: `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths`.

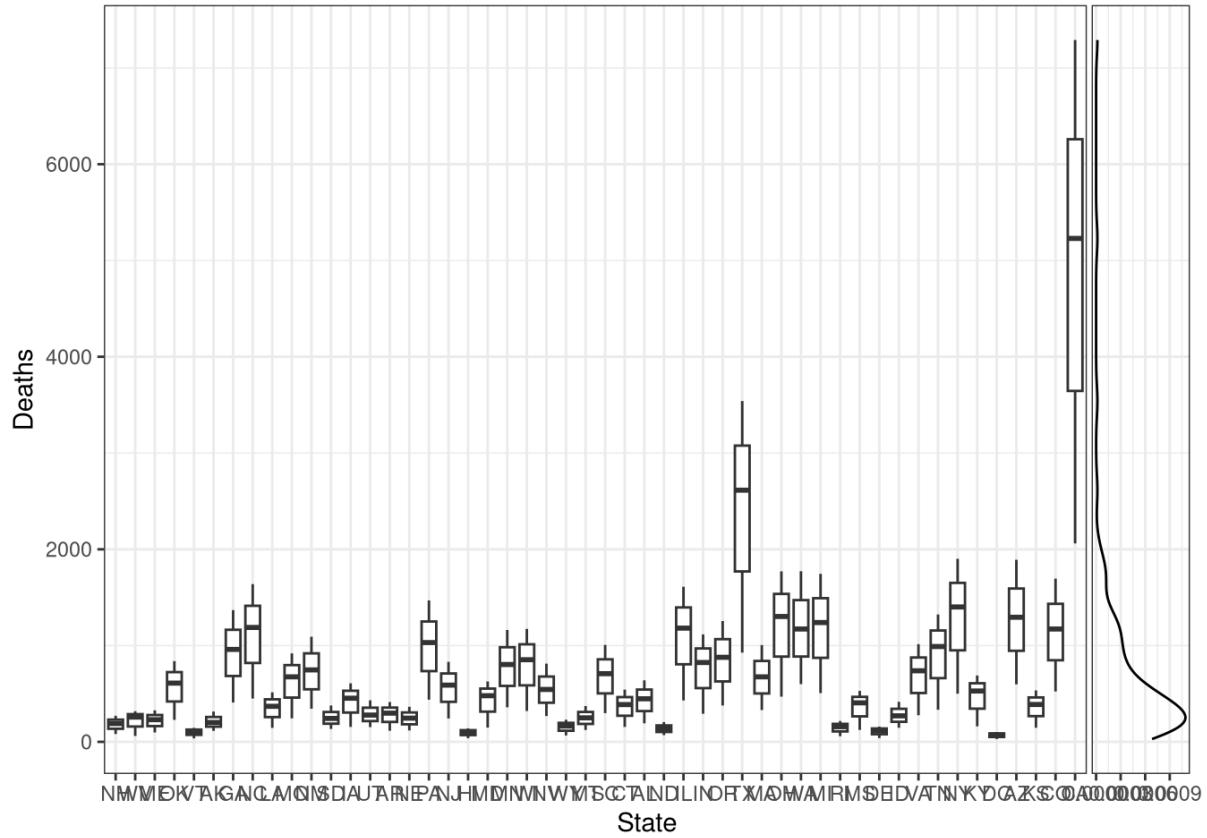


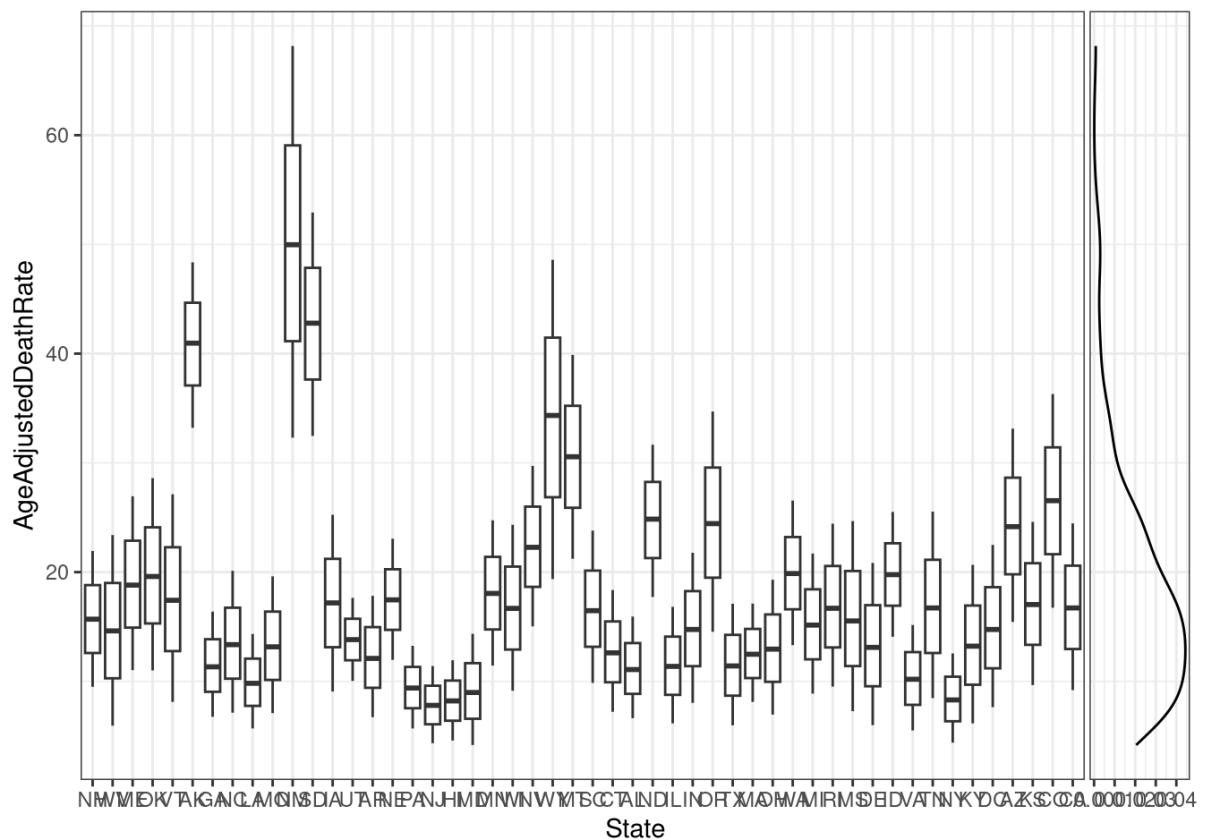
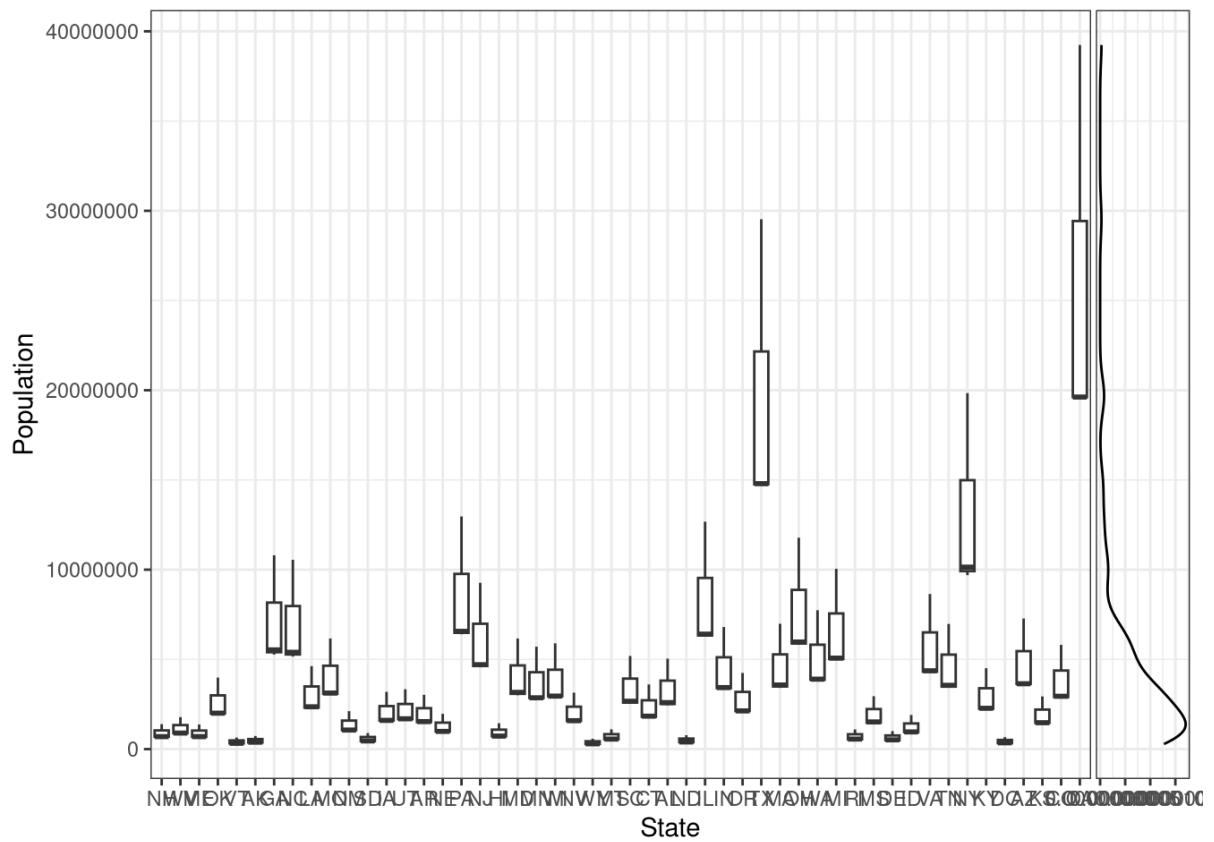


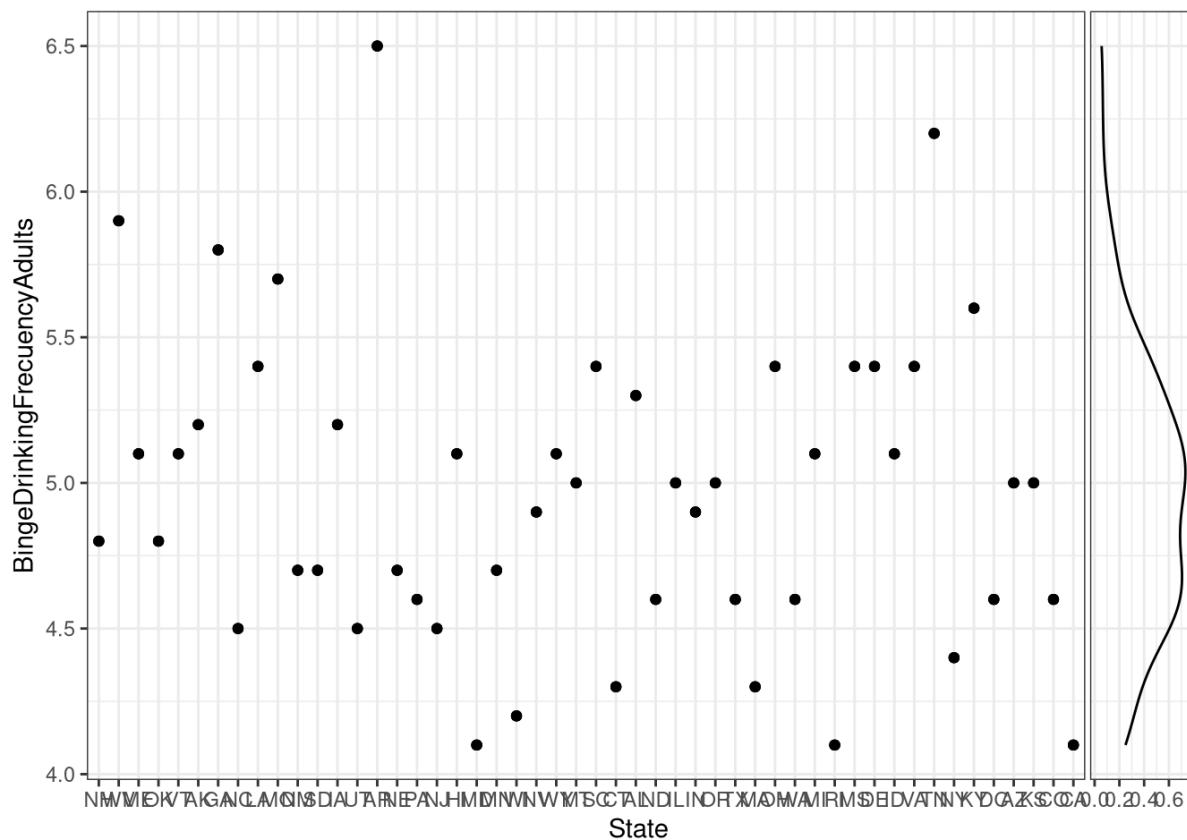
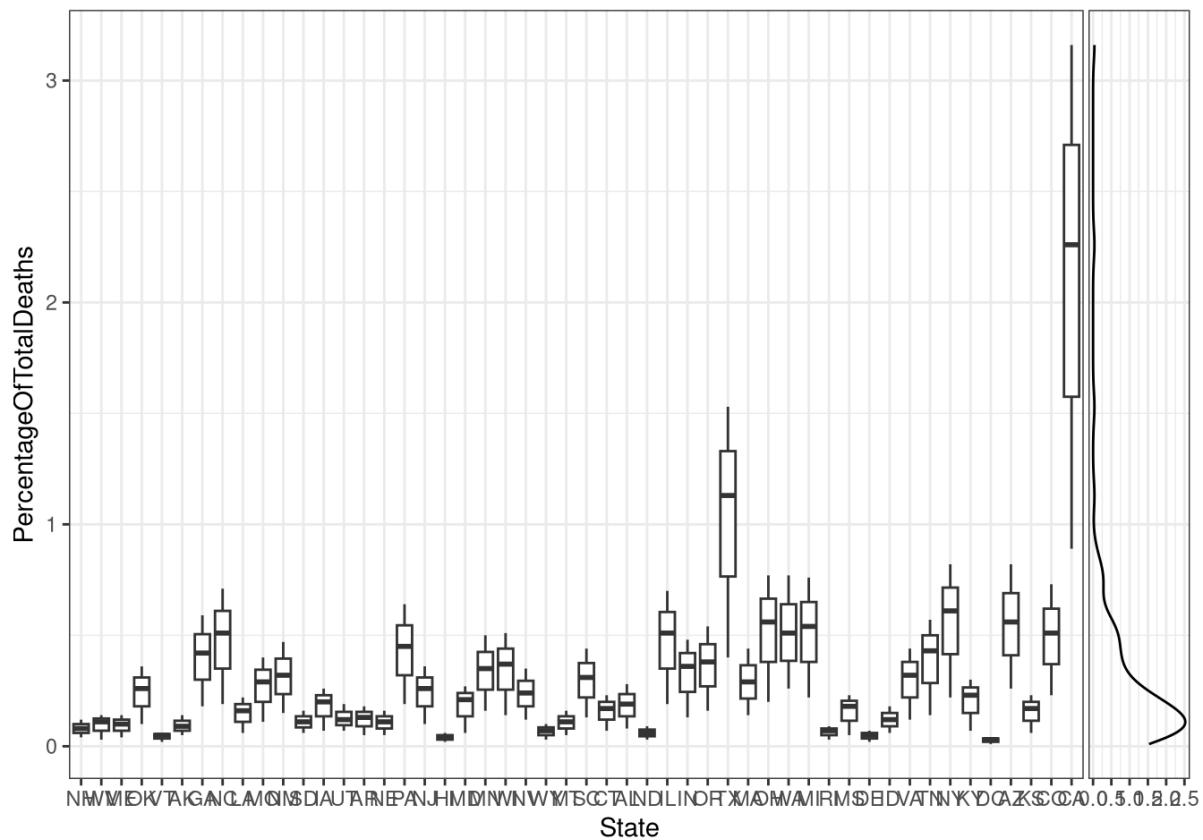
Se crearon dos conjuntos de datos para poder realizar el análisis de agrupación con y sin datos atípicos.

2- Objeto data_overall

En el subprocesso de EDA se identificaron problemas de valores atípicos en cinco variables del objeto `data_overall`: `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths` y `BingeDrinkingFrecuencyAdults`.







Se crearon dos conjuntos de datos para poder realizar el análisis de agrupación con y sin datos atípicos.

Salidas del subprocesso

Se crearon los siguientes objetos, diferenciados entre sí por la presencia o ausencia de tres características: datos estratificados por sexo, Inliers y Outliers:

Objeto	Datos por sexo	Inliers	Outliers
data_lab	Sí	Sí	Sí
data_inliers_lab	Sí	Sí	No
data_outliers_lab	Sí	No	Sí
data_overall_lab	No	Sí	Sí
data_overall_inliers_lab	No	Sí	No
data_overall_outliers_lab	No	No	Sí

Anexo 5 - Código de R utilizado en el análisis

subtitle: TFM - Máster de estadística aplicada para la ciencia de datos con R software.

author: Teodoro José Martínez Arán

editor: source

date: 2024-08-13

date-format: iso

toc: true

toc-depth: 4

toc-location: right

toc-title: Tabla de contenidos

number-sections: true

number-offset: [1,1,1,1]

Anexo A - Código de R utilizado en el análisis {.unnumbered}

Configuración

00a. Definir una configuración de R y RStudio que garantice la reproductibilidad de los resultados

Para facilitar la reproductibilidad del análisis y la coherencia de los resultados obtenidos en distintos equipos, se han incorporado las siguientes opciones de configuración:

- Establecer una semilla aleatoria para el análisis: `set.seed = 2024`
- Impedir que los números grandes se muestren con notación científica: `scipen = 999`

```

```{r Configuracion}
#| code-fold: true

00a. Configuración de R y RStudio

Establecer una semilla aleatoria para el análisis
set.seed(2024)

Impedir que los números grandes se muestren con notación científica
options(scipen = 999)
```

```

00b. Instalar los paquetes de R necesarios para el análisis

```

:::{.callout-note title="Referencias de los paquetes utilizados en el análisis"
collapse="true"}

Los paquetes utilizados en el análisis han sido: BiocManager v. 1.30.23
[@BiocManager], cluster v. 2.1.6 [@cluster], clustertend v. 1.7 [@clustertend], corrplot
v. 0.92 [@corrplot2021], data.table v. 1.15.4 [@datatable], DataExplorer v. 0.8.3
[@DataExplorer], devtools v. 2.4.5 [@devtools], dlookr v. 0.6.3 [@dlookr], factoextra v.
1.0.7 [@factoextra], FeatureImpCluster v. 0.1.5 [@FeatureImpCluster], flexclust v. 1.4.2
[@flexclust2006a; @flexclust2006c; @flexclust2006d; @flexclust2010b;
@flexclust2018e], flextable v. 0.9.6 [@flextable], fpc v. 2.2.12 [@fpc], GGally v. 2.2.1
[@GGally], ggrepel v. 0.9.5 [@ggrepel], ggsignif v. 0.3.1 [@ggsignif], ggstatsplot v. 0.12.4
[@ggstatsplot], here v. 1.0.1 [@here], Hmisc v. 5.1.3 [@Hmisc], kableExtra v. 1.4.0
[@kableExtra], mice v. 3.16.0 [@mice], moments v. 0.14.1 [@moments], naniar v. 1.1.0
[@naniar], NbClust v. 3.0.1 [@NbClust], performance v. 0.12.2 [@performance], psych
v. 2.4.6.26 [@psych], rmarkdown v. 2.27 [@rmarkdown2018; @rmarkdown2020;
@rmarkdown2024], SmartEDA v. 0.3.10 [@SmartEDA], summarytools v. 1.0.1
[@summarytools], tidyverse v. 2.0.0 [@tidyverse].
```

:::

Los paquetes de R necesarios para este análisis están recogidos en el objeto `paquetesNecesariosAnalisis`, y han sido los siguientes:

```
```{r}
#| code-fold: true

Instalación de paquetes necesarios para el análisis

Listado de paquetes a instalar

Paquetes de CRAN

libsCran <- c(
 'arsenal', # An Arsenal of 'R' Functions for Large-Scale Statistical Summaries
 'BiocManager', # Access the Bioconductor Project Package Repository
 'clustertend', # Check the Clustering Tendency
 'DataExplorer', # Automate Data Exploration and Treatment
 'data.table', # Extension of 'data.frame'
 'devtools', # Tools to Make Developing R Packages Easier
 'dlookr', # Tools for Data Diagnosis, Exploration, Transformation
 'dplyr', # A Grammar of Data Manipulation
 'factoextra', # Extract and Visualize the Results of Multivariate Data Analyses
 'flexclust', # Flexible Cluster Algorithms
 'flextable', # Functions for Tabular Reporting
 'FeatureImpCluster', # Feature Importance for Partitional Clustering
 'fpc', # Flexible Procedures for Clustering
 'GGally', # Extension to 'ggplot2'
 'ggplot2', # Create Elegant Data Visualisations Using the Grammar of Graphics
 'ggpubr', # 'ggplot2' Based Publication Ready Plots
 'ggside', # Side Grammar Graphics
```

```

'ggrepel', # Automatically Position Non-Overlapping Text Labels with 'ggplot2'
'grateful', # Facilitate Citation of R Packages
'here', # A Simpler Way to Find Your Files
'kableExtra', # Construct Complex Table with 'kable' and Pipe Syntax
'mice', # Multivariate Imputation by Chained Equations
'moments', # Moments, Cumulants, Skewness, Kurtosis and Related Tests
'naniar', # Data Structures, Summaries, and Visualisations for Missing Data
'NbClust', # Determining the Best Number of Clusters in a Data Set
'plotly', # Create Interactive Web Graphics via 'plotly.js'
'psych', # Procedures for Psychological, Psychometric, and Personality Research
'rstantools', # Tools for Developing R Packages Interfacing with 'Stan'
'SmartEDA', # Summarize and Explore the Data
'summarytools',# Tools to Quickly and Neatly Summarize Data
'tidycensus', # Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames
'tidyr', # Tidy Messy Data
'tinytex', # Helper Functions to Install and Maintain TeX Live
'usmap', # US Maps Including Alaska and Hawaii
'usmapdata', # Mapping Data for 'usmap' Package
'utils' # Various Programming Utilities
)

```

## Paquetes de BioConductor

```

libsBioconductor <- c(
 'graph'
)

```

## Paquetes en repositorios de Github

```

libsGitHub <- c(
 'tinytex'
)

reposGitHub <- c(
 'rstudio/tinytex'
)

paquetesGitHub <- data.frame(
 libsGitHub,
 reposGitHub
)

Identificación de los paquetes que ya están instalados

isInstalledLibCran <- libsCran %in% rownames(utils::installed.packages())
isInstalledLibGitHub <- libsGitHub %in% rownames(utils::installed.packages())
isInstalledLibBioConductor <-
 libsBioconductor %in% rownames(utils::installed.packages())

Instalación de los paquetes faltantes

Paquetes CRAN

if (any(isInstalledLibCran == F)) {
 utils::install.packages(libsCran[!isInstalledLibCran])
}

Paquetes Bioconductor

if (any(isInstalledLibBioConductor == F)) {
 BiocManager::install(libsBioconductor[!isInstalledLibBioConductor], ask = F)
}

```

```
Paquetes en repositorios GitHub

if (any(isInstalledLibGitHub == F)) {

 sapply(
 paquetesGitHub$reposGitHub[!isInstalledLibGitHub],
 devtools::install_github,
 upgrade = 'ask',
 build_manual = TRUE,
 build_vignettes = TRUE
)
}
```

```
Elaboración de un data.frame con la lista de paquetes instalados

Listado de paquetes gestionados

libs <- c(libsCran, libsBioconductor, libsGitHub) |> unique()

libs <- libs[order(libs)]
```

```
Dataframe de paquetes
```

```
paquetes <- sapply(
 libs,
 utils::packageDescription,
 fields = c(
 'Package',
 'Title',
 'Version',
 'Author',
 'Description',
 'License',
```

```

'URL',
'BugReports',
'Depends',
'Imports',
'Suggests',
'Date/Publication'
)
) |>
as.data.frame()

```

```
paquetesNecesariosAnalisis <- do.call(rbind.data.frame, paquetes)
```

```

(OPCIONAL) - Muestra resultado en salida de quarto
Eliminamos los saltos de carro que incluyen algunos títulos de paquetes '\n'
paquetesNecesariosAnalisis$Title <- gsub(
 pattern = "\n",
 replacement = " ",
 x = paquetesNecesariosAnalisis$Title
)

```

```

paquetesNecesariosAnalisis |>
 dplyr::select(
 Title,
 Version
) |>
 kableExtra::kable()

```

```
```
```

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
(OPCIONAL) - Crear una bibliografía de los paquetes de R utilizados
```

```
grateful::cite_packages(
```

```
 out.dir = here::here('notebooks'),
```

```
 out.format = 'Rmd')
```

```
```
```

```
#### Limpieza de objetos temporales del subprocesso
```

```
```{r}
```

```
#| code-fold: true
```

```
Limpieza
```

```
rm(list = c(
```

```
 'isInstalledLibCran',
```

```
 'isInstalledLibGitHub',
```

```
 'isInstalledLibBioConductor',
```

```
 'libs',
```

```
 'libsCran',
```

```
 'libsBioconductor',
```

```
 'libsGitHub',
```

```
 'reposGitHub',
```

```

'paquetesGitHub',
'paquetes'
)
)
```
## Ingesta

```

01a - Identificar los datos necesarios para el análisis y sus fuentes

Para el análisis se utilizaron los siguientes datos:

| Datos | Fuente |
|--|--------|
| Justificación | |
| ----- ----- | |
| ----- ----- | |
| 1 - Indicadores de salud relacionados con el consumo de alcohol [U.S. Chronic Disease Indicators (CDI), 2023 Release](https://data.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-2023-Release/g4ie-h725/about_data) Datos necesarios para hacer el análisis de agrupación (variables predictoras) | |
| 2 - Tasas de mortalidad relacionadas con el consumo de alcohol [U.S. Underlying Cause of Death, 2018-2022, Single Race](https://wonder.cdc.gov/ucd-icd10-expanded.html) Datos de la variable respuesta a predecir en el análisis de regresión | |
| 3 - Códigos FIPS de los estados de EEUU states and counties](https://search.r-project.org/CRAN/refmans/tidycensus/html/fips_codes.html) [Dataset with FIPS codes for US states and counties] Tabla maestra con los códigos FIPS identificativos de los estados y condados de EEUU | |

1 - Indicadores de salud relacionados con el consumo de alcohol

[El conjunto de datos *Cronic Disease Indicators (CDI)*](<https://data.cdc.gov/resource/g4ie-h725.csv>) ha sido elaborado por el *Centres of Disease Control* de USA. Contiene un conjunto de 124 indicadores desarrollados por consenso entre todos los estados miembros de US, utilizado para definir, recoger e informar sobre las enfermedades crónicas de manera uniforme entre distintos estados y territorios. Los indicadores se agrupan en 17 áreas de interés. Puede consultarse una descripción detallada de los indicadores que contiene en el artículo [Indicators for Chronic Disease Surveillance — United States, 2013]([@holt2015\]. Para una descripción más detallada, puede consultarse la \[web del CDI\]\(\[www.cdc.gov/cdi\]\(http://www.cdc.gov/cdi\)\).](https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6401a1.htm)

Para nuestro análisis, se seleccionó el subconjunto de indicadores del área de interés `Alcohol`.

:::{.callout-note title="Áreas de interés del conjunto de datos CDI del CDC" collapse="true"}

Además de las áreas de interés analizadas en el análisis, existen otras áreas de interés en el dataset:

| | |
|---------------|-------|
| Topic area | |
| Justification | |
| ----- | ----- |
| ----- | ----- |
| ----- | ----- |
| ----- | ----- |
| ----- | ----- |
| ----- | ----- |

| [Alcohol](<https://www.cdc.gov/cdi/indicator-definitions/alcohol.html>)
| About 178,000 people die from excessive alcohol use each year in the United States.
Excessive drinking includes binge drinking, heavy drinking, and any drinking during
pregnancy or by people younger than 21. Policies that make alcohol less available and
less affordable can prevent excessive drinking and related harms. Individuals,
organizations, communities, and states can support proven solutions to reduce

alcohol-related harms and improve health and safety.

|

| [Arthritis](<https://www.cdc.gov/cdi/indicator-definitions/arthritis.html>)

| About 53 million U.S. adults have arthritis. That number is expected to increase as people live longer. Arthritis is a general term for conditions that affect the joints, tissues around the joint, and other connective tissues. Managing arthritis symptoms is important to reduce pain, prevent or delay disability, and improve overall quality of life. Public health professionals and others can help build awareness of proven self-management strategies to reduce arthritis pain so patients can pursue the activities that are important to them.

|

| [Asthma](<https://www.cdc.gov/cdi/indicator-definitions/asthma.html>)

| About 25 million U.S. children and adults have asthma, which kills over 3,500 people each year. Asthma affects the lungs, causing wheezing, breathlessness, chest tightness, and coughing. CDC's National Asthma Control Program (NACP) helps people control and manage their asthma for better health. Public health and other professionals can play a role in asthma-related surveillance, public education, and provider training.

|

| [Cancer](<https://www.cdc.gov/cdi/indicator-definitions/cancer.html>)

| Cancer is the second leading cause of death in the United States, causing 1 in every 6 deaths. We can reduce cancer cases and deaths by limiting behavioral and environmental risks, making screening and treatment available to all, supporting medically underserved populations, and improving quality of life for those who have survived cancer. The Community Preventive Services Task Force recommends several patient- and provider-focused interventions to increase screening for breast, cervical, and colorectal cancers.

|

| [Cardiovascular Disease](<https://www.cdc.gov/cdi/indicator-definitions/cardiovascular-disease.html>)

| Heart disease is the number one cause of death and disability in the United States. Cardiovascular disease not only includes heart disease, but also stroke, heart failure, and atrial fibrillation. Common behaviors and health conditions put people at risk, like smoking, unhealthy diet, inactivity, excessive alcohol, high blood pressure, high cholesterol, and obesity. We can prevent cardiovascular disease and reduce related death and disability with lifestyle changes, control of risk factors, and timely, effective treatment.

|

| [Chronic Kidney Disease](<https://www.cdc.gov/cdi/indicator-definitions/chronic-kidney-disease.html>)

| More than 1 in 7 U.S. adults has chronic kidney disease (CKD), though most do not know it. CKD is caused by damaged kidneys that

cannot properly filter blood, causing fluids and waste to build up in the body. CKD usually gets worse over time, but treatment and lifestyle changes can slow it down. Public health strategies—from public and provider education, to monitoring CKD and its risk factors—can promote kidney health.

|

| [Chronic Obstructive Pulmonary Disease](<https://www.cdc.gov/cdi/indicator-definitions/chronic-obstructive-pulmonary-disease.html>) | Chronic obstructive pulmonary disease (COPD) is 1 of the top 10 causes of death in the United States. Nearly 16 million U.S. adults have COPD, and many more do not know they have it. COPD prevents airflow to the lungs, causing breathing problems. There is no cure for COPD, but it can be managed and treated. Public health professionals and others can help build awareness of COPD and support prevention, early diagnosis, treatment, and management strategies.

|

| [Cognitive Health Caregiving](<https://www.cdc.gov/cdi/indicator-definitions/cognitive-health-caregiving.html>) | About 1 in 10 adults (45 and older) reports worsening memory loss (or cognitive decline) and another 1 in 4 reports caring for someone who has a cognitive impairment. Dementias, like Alzheimer's disease, start with mild memory loss and can lead to an inability to carry on conversations or respond to one's environment. Caregiving can be rewarding, but it can also put a burden on caregivers, at the risk of their own health. As our aging population grows, so does the need for public health strategies to reduce the impact of dementias and raise awareness of caregiving as a public health issue.

|

| [Diabetes](<https://www.cdc.gov/cdi/indicator-definitions/diabetes.html>) | About 38 million U.S. adults have diabetes, and 1 in 5 of them don't know it. Diabetes is a long-lasting condition that can lead to serious disease and death over time. There is no cure for diabetes, but we can prevent or manage the most common type (type 2) by eating healthy, being active, and keeping a healthy weight. Public health efforts can offer proven lifestyle change programs to reduce diabetes risk for people.

|

| [Disability](<https://www.cdc.gov/cdi/indicator-definitions/disability.html>) | More than 1 in 4 U.S. adults (27%) has a disability. A disability is any condition of the body or mind that makes it more difficult to do certain tasks. People with disabilities are more likely to report poor health than those without disabilities. We can best serve people with disabilities by including them in all health care, health promotion, and disease prevention programs to help them stay active, healthy, and thrive.

|

| [Health Status](<https://www.cdc.gov/cdi/indicator-definitions/health-status.html>)

| About 1 in 6 U.S. adults report fair or poor health, which is notable since self-rated health can predict hospitalization and death. Health—including physical, mental, emotional, and functional health—is fundamental for everyday living. We can help improve overall health by addressing chronic disease risk factors and social determinants of health, and by identifying and linking those in poor health to community resources.

|

| [Immunization](<https://www.cdc.gov/cdi/indicator-definitions/immunization.html>)

| Each year in the United States, many people get diseases that vaccines could have prevented. Vaccines help the body create proteins that fight off infections (called antibodies), protecting us from potentially life-threatening diseases. Yet some children and adults do not get their recommended vaccines. The types of vaccines needed are based on age, and in some cases, certain chronic diseases or risk factors.

|

| [Maternal Health](<https://www.cdc.gov/cdi/indicator-definitions/maternal-health.html>) | Maternal health refers to women's health and well-being during pregnancy, childbirth, and postpartum (after childbirth). Health

problems—like diabetes, high blood pressure, and depression—can arise before, during or after pregnancy, putting the mother's or infant's health at risk. Alcohol and tobacco use during pregnancy can also harm the mother's and baby's health. Access to good health care can help women adopt healthier habits, prevent and address complications, and improve pregnancy and postpartum health.

|

| [Mental Health](<https://www.cdc.gov/cdi/indicator-definitions/mental-health.html>)

| Nearly 1 in 4 U.S. adults had a mental illness in the past year, and almost 3 in 10 high school students reported poor mental health. Mental health—which includes emotional, psychological, and social well-being—is important for overall health and well-being. Mental health conditions (like depression) increase the risk of chronic conditions, which in turn can increase the risk of mental health conditions. Public health strategies can address these conditions and promote mental health and emotional well-being.

|

| [Nutrition, Physical Activity, and Weight Status](<https://www.cdc.gov/cdi/indicator-definitions/npao.html>) | Poor nutrition and physical inactivity increase the

risk of chronic conditions like obesity, depression, type 2 diabetes, heart disease, and some cancers—which can lead to disability and premature death. Fewer than 1 in 10 children and adults eat their recommended vegetables. A quarter of adults (25%) and even fewer adolescents (16%) meet U.S. physical activity guidelines. As many as 40% of adults and 20% of adolescents have obesity. Public health approaches, including

surveillance, education, policy and environmental strategies, and resources are needed to make healthy eating and active living accessible for everyone. |

| [Older Adults](<https://www.cdc.gov/cdi/indicator-definitions/older-adults.html>)
| The U.S. population is aging, with almost a quarter of the population expected to be 65 or older by 2060. Aging increases the risk of chronic diseases like dementia, heart disease, type 2 diabetes, arthritis, and cancer. Older adults are also more vulnerable to severe illness from infections, including flu and pneumonia. Effective strategies for healthy aging are needed to improve the length and quality of life of older adults, and their ability to live independently.

|

| [Oral Health](<https://www.cdc.gov/cdi/indicator-definitions/oral-health.html>)
| Nearly all Americans have had tooth decay in their lifetime. Left untreated, oral diseases can cause severe pain, infections, tooth loss, and (rarely) even death. Poor oral health limits a person's ability to eat, learn, and work. It unevenly affects members of racial and ethnic groups and those with low income and education. Most oral diseases can be prevented by limiting risks like tobacco, alcohol, and sugary foods and drinks. Public health strategies—like community water fluoridation and school sealant programs—are safe, cost-effective ways proven to prevent cavities and improve oral health equity. |

| [Sleep](<https://www.cdc.gov/cdi/indicator-definitions/sleep.html>)
| About one-third of U.S. adults and children (under 14) and three-quarters of high schoolers do not get enough sleep. Insufficient sleep is linked to increased risk of anxiety, depression, obesity, heart disease, injury, and other serious conditions. We can promote sleep health through research and surveillance, public and provider education, clinical guidance, and traffic safety education.

|

| [Social Determinants of Health](<https://www.cdc.gov/cdi/indicator-definitions/sdoh.html>) | Social determinants of health (SDOH) are nonmedical factors and conditions in our environment that influence health, well-being, and quality of life. They include things like economic stability; our social, community, and built environments; and access to quality health care and education. Addressing differences in SDOH can advance health equity so everyone has the opportunity for optimal health. Action and collaboration across sectors (e.g., public health, transportation, education, housing, health care) and from public, private, and community agencies are needed to address SDOH. |

| [Student Health](<https://www.cdc.gov/cdi/indicator-definitions/student-health.html>)
| Youth risk behaviors—like physical inactivity, unhealthy diet, and tobacco, alcohol, and other drug use—are linked to lower academic achievement and poorer mental and

physical health. On the other hand, healthy students are better learners, and academic achievement bears a lifetime of health benefits. With 8 hours a day of direct contact with students, U.S. schools are uniquely positioned to promote student health behaviors and reduce unhealthy behaviors for better health and well-being. They can do this using strategies from the Whole School, Whole Community, Whole Child model.

|

| [Tobacco](<https://www.cdc.gov/cdi/indicator-definitions/tobacco.html>)

| Cigarette smoking is the leading preventable cause of disease, death, and disability in the United States. Smoking and secondhand smoke exposure cause over 480,000 U.S. deaths each year. Nearly 12% of U.S. adults currently smoke cigarettes and 10% of middle and high school students currently use a tobacco product. Increasing the use of proven tobacco control strategies can reduce tobacco use and related diseases.

|

:::

2 - Tasas de mortalidad relacionadas con el consumo de alcohol

El conjunto de datos [Underlying Cause of Death](<https://wonder.cdc.gov/ucd-icd10-expanded.html>) contiene datos de mortalidad y población para todos los condados de EEUU. Los datos provienen de los certificados de defunción de los residentes de EEUU. Cada certificado de muerte identifica una única causa de muerte, junto con un conjunto de datos demográficos.

Para nuestro análisis, se seleccionaron los datos de mortalidad por alcohol, ajustada por edad, estado y año. Se descargaron además, los datos globales, para poder comparar la media de cada uno de los niveles de las variables con la media global.

:::{.callout-note title="Configuración de la herramienta de descarga de datos WONDER del CDC para obtener el conjunto de datos utilizado en el análisis" collapse="true"}

Los datos para el análisis se obtuvieron con la siguiente configuración de la herramienta de búsqueda [CDC WONDER](<https://wonder.cdc.gov/controller/datarequest/D158>):

| Sección | Subsección | Parámetro |
|---------------|------------|-----------|
| Configuración | | |

| | | |
|---|---|--|
| | | |
| | | |
| 1\. Organize table layout
` And By`
· ` And By`
State
Year
Gender | Group Results By
 · ` Group Results By`

 | |
|
` Crude Rate`
TRUE
TRUE
TRUE | Default Measures
 · ` Deaths`
· ` Population`

 | |
|
` Standard Error` | For Crude Rates
 · ` 95% Confidence Interval`

 TRUE
TRUE | |
|
TRUE
TRUE
TRUE | Age Adjusted Rate
 · ` Age Adjusted Rate`
· ` 95%
Confidence Interval`
· ` Standard Error`
 | |
|
 Total Deaths | · ` Percent of Total Deaths`
 | |
|
 Additional Rate Options
` Standard Population`
2000 U.S. Std. Population | · ` Calculate Rates Per`

 100,000
 | |
| 2\. Select location
 Grouping method
locations by US-Mexico Border Region, Border State Area, State, Census Region or HHS
Region` | · ` Click a button to choose
 States | |
|
 All (The United States) | Selected codes
 · ` Browse or search to find items in the
States Finder Tool, then highlight the items to use for this request.`
 | |
|
 2013 Urbanization, All categories | Urbanization
 · ` Pick between`
 | |
| 3\. Select demographics
 All ages | Age Groups
 · ` Pick between`
 | |
|
 All Genders | Gender
 · ` Gender`
 | |
|
 All Origins | Hispanic origin
 · ` Hispanic origin`
 | |

| | | |
|--|------------------------|---|
| | Single Race | · ` Pick between` |
| Single Race 6 , All Races | | |
| 4\. Select year and month | | · `Year/Month` |
| *All* (All Dates) | | |
| 5\. Select weekday, autopsy and place of death | Weekday | · `Weekday` |
| All Weekdays | | |
| | Autopsy | · `Autopsy` |
| All values | | |
| | Place of death | · `Place of death` |
| All places | | |
| 6\. Select cause of death | Grouping method | · `Click a button to select ICD codes by Chapters or by Groups` |
| Drug/Alcohol Induced Causes | | |
| | Selected codes | · `Browse or search to find items in the Drug/Alcohol Induced Causes Finder Tool, then highlight the items to use for this request` |
| | Alcohol-induced causes | |
| 7\. Other options | | · `Export results`
· `Show totals`
· `Show Zero values`
· `Show supressed values`
· `Precision (decimal places)`
· `Data Access Timeout` |
| | | TRUE
FALSE
TRUE
TRUE
2
10 |
| ::: | | |

3 - Códigos FIPS de los estados de EEUU

Para garantizar una adecuada estandarización de los datos, se utilizó como tabla maestra el dataset `fips_states` del paquete **tidycensus** [Dataset with FIPS codes for US states and counties](https://search.r-project.org/CRAN/refmans/tidycensus/html/fips_codes.html).

01b - Definir el método y la configuración de la ingesta

| | |
|-------------------|--------|
| Conjunto de datos | Método |
| Configuración | |

| | |
|--|--|
| | |
| | |
| | |
| 1 - Indicadores de salud relacionados con el consumo de alcohol Ingesta directa desde origen con `data.table::fread()` | Opciones por defecto de la función, salvo `encoding = "UTF-8"` |
| 2 - Tasas de mortalidad relacionadas con el consumo de alcohol · Descarga del conjunto de datos crudo con la configuración especificada
· Ingesta desde descarga local `encoding = "UTF-8"``
`header = TRUE`
`nrow = 306` | |
| 3 - Códigos FIPS de los estados de EEUU Opciones por defecto de la función | Carga directa con `data()` |

01c - Crear los *data.frame* de datos crudos

Se crearon los siguientes *data.frame* de datos crudos:

| Objeto | Fuente | Descripción |
|--|---|-------------|
| | | |
| ----- ----- ----- | | |
| ----- ----- ----- | | |
| ----- ----- ----- | | |
| `rawCdiAlcohol` [*Chronic disease indicators (CDI), CDC*, 2023](https://data.cdc.gov/resource/g4ie-h725.csv) | Evaluación de los indicadores de enfermedades crónicas relacionados con el consumo excesivo de alcohol, ajustados por sexo y edad, durante el periodo 2010-21, por estado y año | |
| `rawMortalityAlcohol` [Underlying Cause of Death, 2018-2022, Single Race](https://wonder.cdc.gov/wonder/help/ucd-expanded.html#) | Tasa de mortalidad relacionada con alcohol, ajustada por sexo y edad, durante el periodo 2018-22, por estado y año | |
| `rawFipsCodes` Paquete **tidycensus** | | |
| Tabla maestra de códigos para estados y condados de EEUU | | |
| | | |

```

::: {.callout-note title="1 - Indicadores de salud relacionados con el consumo de alcohol - `rawCdiAlcohol` " collapse="true"}

```{r}

Indicadores de salud relacionados con el consumo de alcohol - `rawCdiAlcohol`

rawCdiAlcohol <- data.table::fread(
 file = here::here(
 "data", "raw", "US_Chronic_Disease_Indicators_CDI_2023_Release_20240716.csv"
),
 encoding = "UTF-8"
) |>

Filtrado de indicadores de la dimensión 'Alcohol'

dplyr::filter(Topic == "Alcohol")

```

:::

::: {.callout-note title="2 - Tasas de mortalidad relacionadas con el consumo de alcohol - `rawUnderlyingCauseOfDeathAlcohol` " collapse="true"}

[Documentación del dataset](https://wonder.cdc.gov/wonder/help/ucd-expanded.html#)

```{r}

#| code-overflow: wrap

Tasas de mortalidad relacionadas con el consumo de alcohol por sexo y estado-
`rawUnderlyingCauseOfDeathAlcohol`

rawUnderlyingCauseOfDeathAlcohol <- data.table::fread(
 here::here(
 'data', 'raw', 'Underlying Cause of Death, 2018-2022, Single Race.txt'
),
```

```

```
encoding = "UTF-8",  
header = TRUE,  
nrow = 817 # Se ingestan únicamente las observaciones, y se desprecian los  
comentarios  
)
```

```

:::

```
:::{.callout-note title="3 - Códigos FIPS de los estados de EEUU `rawFipsCodes` "
collapse="true"}
```{r}
```

```
# Códigos FIPS de los estados de EEUU - `fips_codes`  
data("fips_codes", package = "tidycensus")  
rawFipsCodes <- fips_codes # Normalización del nombre del data.frame  
rm(fips_codes)
```

```

:::

```
01d - Validar la fase de ingesta
```

En la validación de la fase de ingesta se realizaron las siguientes actividades:

- Verificar la completitud de la ingesta `head()`, `tail()`, `dim()`
- Normalizar los nombres de las variables
- Comprobar la estructura de los \*data.frame\* (`str()`)
- Características generales del \*data.frame\*

#### #### 1 - Validación de `rawCdiAlcohol`

| Dimensión                             | Evaluación                                                                                                                    |
|---------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Completitud                           | Ingesta correcta                                                                                                              |
| Normalización del nombre de variables | No es necesaria                                                                                                               |
| Estructura del *data.frame*           | Algunas variables deben convertirse a tipo `factor`                                                                           |
| Resumen del *data.frame*              | Existen algunas variables innecesarias (sin datos, o con todos los valores iguales)<br>Hay valores faltantes dispersos (`NA`) |

```
::: {.callout-note title="Completitud de la ingesta (`head()`, `tail()`, `dim()`)" collapse="true"}
```

La ingestá se ha realizado correctamente, con el número de filas y variables esperado.

```
```{r}
```

```
# Verificación de que se han ingestado correctamente las primeras filas
```

```
head(rawCdiAlcohol)
```

```
# Verificación de que se han ingestado correctamente las últimas filas
```

```
tail(rawCdiAlcohol)
```

```
# Valoración del número de filas y columnas ingestadas
```

```
dim(rawCdiAlcohol)
```

```
```
```

```
:::
```

::: {.callout-note title="Normalización de los nombres de las variables" collapse="true"}

En nuestro \*data.frame\* las variables del dataset original tenían un nombre normalizado siguiendo el estándar

[PascalCase]([https://en.wikipedia.org/wiki/Camel\\_case](https://en.wikipedia.org/wiki/Camel_case)), por lo que no fue necesario renombrarlas.

```
```{r}
```

```
#| code-fold: true
```

```
#### Normalización de los nombres de las variables
```

```
(rawVarNamesCdi <- names(rawCdiAlcohol))
```

```
```
```

```
:::
```

::: {.callout-note title="Estructura del \*data.frame\* (`str()`)" collapse="true"}

Deben convertirse algunas variables al tipo `factor`.

```
```{r}
```

```
#| code-fold: true
```

```
#### Comprobar la estructura de los *data.frame* (`str()`)
```

```
str(rawCdiAlcohol)
```

```
```
```

```
:::
```

::: {.callout-note title="Resumen del \*data.frame\* " collapse="true"}

Deben convertirse algunas variables al tipo `factor` y `numeric`.

- Existen algunas variables innecesarias (sin datos, o con todos los valores iguales)
- Hay valores faltantes dispersos (`NA`)

```

```{r}
#| code-fold: true

#### Explorar las características generales de los datos ingestados

summary(rawCdiAlcohol)

```
:::
```

#### 2 - Validación de `rawUnderlyingCauseOfDeathAlcohol`

| Dimensión                             | Evaluación                                                                                                       |
|---------------------------------------|------------------------------------------------------------------------------------------------------------------|
| Completitud                           | Faltan datos para algunos estados en algunos años                                                                |
| Normalización del nombre de variables | · Los nombres de las variables del dataset original no están normalizados.<br>· Deben renombrarse las variables. |
| Estructura del *data.frame*           | · Algunas variables deben convertirse a tipo `factor`<br>· Alguna variable debe convertirse a tipo `numeric`     |
| Resumen del *data.frame*              | · Debe convertirse a numérico el valor `% of Total Deaths`<br>· Hay valores faltantes dispersos (`NA`)           |

::: {.callout-note title="Completitud de la ingesta (`head()`, `tail()`, `dim()`)" collapse="true"}

Faltan datos para algunos de los estados en algunos años.

```

```{r}
# Verificación de que se han ingestado correctamente las primeras filas
head(rawUnderlyingCauseOfDeathAlcohol)

# Verificación de que se han ingestado correctamente las últimas filas
tail(rawUnderlyingCauseOfDeathAlcohol)
```

```
# Valoración del número de filas y columnas ingestadas  
dim(rawUnderlyingCauseOfDeathAlcohol)
```

```

:::

:::{.callout-note title="Normalización de los nombres de las variables" collapse="true"}

Los nombres de las variables del dataset original no están normalizados. Deben renombrarse las variables.

```{r}

```
#| code-fold: true
```

```
## Guardar los nombres de variable originales
```

```
(rawVarNamesCauseOfDeath <- names(rawUnderlyingCauseOfDeathAlcohol))
```

```

:::

:::{.callout-note title="Estructura del \*data.frame\* (`str()`)" collapse="true"}

- Deben convertirse algunas variables al tipo `factor`, y alguna variable debe convertirse a tipo `numeric` .

```{r}

```
#| code-fold: true
```

```
#### Comprobar la estructura de los *data.frame* (`str()`)
```

```
str(rawUnderlyingCauseOfDeathAlcohol)
```

```

:::

```
::: {.callout-note title="Resumen del *data.frame* " collapse="true"}
- Debe convertirse a numérico el valor `% of Total Deaths`
- Hay valores faltantes dispersos (`NA`)
```

```
```{r}
```

```
#| code-fold: true
```

```
#### Explorar las características generales de los datos ingestados
```

```
summary(rawUnderlyingCauseOfDeathAlcohol)
```

```
```
```

```
:::
```

```
3 - Validación de `rawFipsCodes`
```

| Dimensión                             | Evaluación                                           |  |
|---------------------------------------|------------------------------------------------------|--|
| Completitud                           | Ingesta correcta                                     |  |
| Normalización del nombre de variables | Debe convertirse a Pascal Case                       |  |
| Estructura del *data.frame*           | Deben convertirse algunas variables al tipo `factor` |  |
| Resumen del *data.frame*              | Sin problemas adicionales                            |  |

```
::: {.callout-note title="Compleitud de la ingesta (`head()`, `tail()`, `dim()`)" collapse="true"}
```

No hay problemas con la completitud de la ingesta

```
```{r}
```

```
# Verificación de que se han ingestado correctamente las primeras filas
```

```
head(rawFipsCodes)

# Verificación de que se han ingestado correctamente las últimas filas

tail(rawUnderlyingCauseOfDeathAlcohol)

# Valoración del número de filas y columnas ingestadas

dim(rawUnderlyingCauseOfDeathAlcohol)

```

:::
```

::: {.callout-note title="Normalización de los nombres de las variables" collapse="true"}

Los nombres de las variables del dataset original están en formato [snake\_case]([https://es.wikipedia.org/wiki/Snake\\_case](https://es.wikipedia.org/wiki/Snake_case)). Se convierten a formato Pascal Case, para mantener la coherencia con el resto de dataframes

```
```{r}

#| code-fold: true

## Guardar los nombres de variable originales
```

```
(rawVarNamesrawFipsCodes <- names(rawFipsCodes))
```

```
```
```

```
:::
```

::: {.callout-note title="Estructura del \*data.frame\* (`str()`)" collapse="true"}

- Deben convertirse algunas variables al tipo `factor`.

```
```{r}

#| code-fold: true

##### Comprobar la estructura de los *data.frame* (`str()`)

str(rawFipsCodes)

```
```

:::

:::{.callout-note title="Resumen del \*data.frame\* " collapse="true"}

Sin hallazgos

```{r}

#| code-fold: true

Explorar las características generales de los datos ingestados

summary(rawFipsCodes)

```

:::

Objeto	Descripción del *data.frame*
Filas	Columnas
-----	-----
-----	-----
-----	-----
`rawCdiAlcohol`   Indicadores de enfermedades crónicas (CDI) del área de interés 'Alcohol', por estado y año (2010-2022)   `r nrow(rawCdiAlcohol)`   `r ncol(rawCdiAlcohol)`	
`rawUnderlyingCauseOfDeathAlcohol`   Tasas de mortalidad por Alcohol, por sexo, estado y año (2018-2022)   `r nrow(rawUnderlyingCauseOfDeathAlcohol)`   `r ncol(rawUnderlyingCauseOfDeathAlcohol)`	
`rawFipsCodes`   Maestra de códigos de estados y condados de EEUU   `r nrow(rawFipsCodes)`   `r ncol(rawFipsCodes)`	

```{r}

#| output: false

#| code-fold: true

```

# Copia de seguridad de los ficheros crudos
saveRDS(
  object = rawCdiAlcohol,
  file = here::here('data', 'raw', 'rawCdiAlcohol.rds')
)
saveRDS(
  object = rawUnderlyingCauseOfDeathAlcohol,
  file = here::here('data', 'raw', 'rawUnderlyingCauseOfDeathAlcohol.rds')
)
saveRDS(
  object = rawFipsCodes,
  file = here::here('data', 'raw', 'rawFipsCodes.rds')
)
```

```

## ## Limpieza

El subprocesso incluye las siguientes acciones:

- 02a - Identificar la información sucia, incorrecta, incompleta, imprecisa, irrelevante o incómoda
- 02b - Reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo a la necesidad

### 02a - Identificación de la información sucia, incorrecta, irrelevante, incompleta, imprecisa o incómoda

#### Validación de `rawCdiAlcohol`

::: {.callout-note title="1- Información sucia" collapse="true"}

#### ##### 1- Información sucia

- Correcto - No existe ningún problema de suciedad de datos

::: {.callout-caution title="Codificación de caracteres incorrecta" collapse="true"}

#### ##### Codificación de caracteres incorrecta

- Correcto:

- Los datasets se han ingestado en la codificación estándar \*\*UTF-8\*\*
- No se han registrado advertencias durante la ingesta
- La apariencia de los datos ingestados es correcta

:::

::: {.callout-caution title="Símbolos innecesarios (\$, €, %, ... )" collapse="true"}

#### ##### Símbolos innecesarios (\\$, €, %, ... )

- Correcto - Sin problemas en ninguna de las variables

:::

::: {.callout-caution title="Nombre de \*data.frame\* no acorde a estilo" collapse="true"}

#### ##### Nombre de \*data.frame\* no acorde a estilo

- Correcto - Nombre del \*\*data.frame\*\* en formato CamelCase

:::

::: {.callout-caution title="Nombre de variables (columnas) no acorde a estilo" collapse="true"}

## ##### Nombre de variables (columnas) no acorde a estilo

- Correcto - Variables con nombre en formato CamelCase

:::

::: {.callout-caution title="Nombre de observaciones (filas) no acorde a estilo" collapse="true"}

## ##### Nombre de observaciones (filas) no acorde a estilo

- No aplica - Las filas no tienen nombre

:::

:::

::: {.callout-note title="2 - Datos incorrectos" collapse="true"}

## ##### 2 - Datos incorrectos

- Correcto - No se detectan problemas con la corrección de los datos

::: {.callout-caution title="Errores del subprocesso de ingestá" collapse="true"}

## ##### Errores del subprocesso de ingestá

- Correcto - No se detectan problemas de ingestá

:::

::: {.callout-caution title="Datos desestructurados - Más de una variable por columna" collapse="true"}

## ##### Datos desestructurados - Más de una variables por columna

- Correcto - Cada columna tiene una única variable

:::

::: {.callout-caution title="Datos desestructurados - Más de una observación por fila" collapse="true"}

#### ##### Datos desestructurados - Más de una observación por fila

- Correcto - Cada fila tiene una única observación

:::

:::

::: {.callout-note title="3 - Datos irrelevantes" collapse="true"}

#### ##### 3 - Datos irrelevantes

Se han detectado los siguientes problemas con datos irrelevantes:

- Existen variables con todos los datos faltantes
- Existen variables con todas las observaciones con el mismo valor
- Existen variables innecesarias para el análisis (columnas)
- Existen observaciones innecesarias para el análisis (filas)

::: {.callout-caution title="Variables con todas las observaciones faltantes" collapse="true"}

#### ##### Variables con todas las observaciones faltantes

- Incorrecto: Existen variables con todos los datos faltantes

Se analizó el patrón de datos faltantes de los datos crudos con la función

`mice::md.pattern()`

```
```{r}
#| code-fold: true

mice:::md.pattern(rawCdiAlcohol, plot = T, rotate.names = T)
```
```

Se observó que:

- Diez variables no tenían ningún valor en ninguna observación
- En tres de ellas había un elevado número de `NA` .
- El resto de variables no presentaba este problema

:::

:::{.callout-caution title="Variables con todas las observaciones con el mismo valor" collapse="true"}

##### Variables con todas las observaciones con el mismo valor

- Incorrecto: Existe al menos una variable con todas las observaciones con el mismo valor (`Topic` )

:::

:::{.callout-caution title="Variables innecesarias para el análisis (columnas)" collapse="true"}

##### Variables innecesarias para el análisis (columnas)

- Incorrecto: Varias variables tienen información innecesaria o redundante para el análisis

Las siguientes variables del dataset original no son necesarias:

| Variable                  | Justificación                                          |
|---------------------------|--------------------------------------------------------|
| `YearEnd`                 | Información redundante con `YearStart`                 |
| `DataSource`              | Irrelevante para el análisis                           |
| `DataValue`               | Información redundante con `DataValueAlt`              |
| `DataValueFootnoteSymbol` | Irrelevante para el análisis                           |
| `DatavalueFootnote`       | Irrelevante para el análisis                           |
| `GeoLocation`             | Irrelevante para el análisis                           |
| `LocationID`              | Irrelevante para el análisis                           |
| `DataValueTypeID`         | Información redundante con `DataValueType`             |
| `DataValueUnit`           | Irrelevante para el análisis                           |
| `LocationDesc`            | Información redundante con `LocationAbbr`              |
| `Question`                | Información redundante con `QuestionID`                |
| `StratificationCategory1` | Información redundante con `StratificationCategoryID1` |
| `Stratification1`         | Información redundante con `StratificationID1`         |
| ...                       |                                                        |

:::{.callout-caution title="Observaciones innecesarias para el análisis (filas)" collapse="true"}

##### Observaciones innecesarias para el análisis (filas)

- Incorrecto: Sobran observaciones en el dataset para el análisis

No son necesarias todas las observaciones del dataset para el análisis. Necesitamos únicamente los datos correspondientes al año con más cobertura de información para todos los estados

...

:::

::: {.callout-note title="4 - Datos incompletos" collapse="true"}

#### ##### 4 - Datos incompletos

Se han detectado los siguientes problemas de incompletitud:

- Faltan datos para algunos de los estados de EEUU
- Faltan datos para algunos de los años; el año con una completitud similar para todos los estados es 2021

::: {.callout-caution title="Cobertura incompleta de observaciones (filas)" collapse="true"}

#### ##### Cobertura incompleta de observaciones (filas)

- Incorrecto - Existen problemas de completitud de observaciones para algunos estados, y para algunos años

```{r}

#| output: false

#| code-fold: true

Lista de estados de cada tabla

```
listaEstadosMaestra <- levels(as.factor(rawFipsCodes$state))
```

```
listaNombresMaestra <- levels(as.factor(rawFipsCodes$state_name))
```

```
listaEstadosCdiAlcohol <- levels(as.factor(rawCdiAlcohol$LocationAbbr))
```

```
listaEstadosCauseOfDeath <-
```

```
levels(as.factor(rawUnderlyingCauseOfDeathAlcohol$State))
```

```

# Estados sin datos en rawCdiAlcohol

esEstadoMaestraConDatosCdi <- listaEstadosMaestra %in% listaEstadosCdiAlcohol
esDatoCdiConEstadoMaestra <- listaEstadosCdiAlcohol %in% listaEstadosMaestra

listaEstadosMaestraSinDatosCdi <-
listaEstadosMaestra[!esEstadoMaestraConDatosCdi]

listaEstadosDatosCdiSinEstadoMaestra <-
listaEstadosCdiAlcohol[!esDatoCdiConEstadoMaestra]

# Estados sin datos en rawUnderlyingCauseOfDeathAlcohol

esEstadoMaestraConDatosCauseOfDeath <- listaNombresMaestra %in%
listaEstadosCauseOfDeath

esDatoCauseOfDeathConEstadoMaestra <- listaEstadosCauseOfDeath %in%
listaNombresMaestra

listaEstadosMaestraSinDatosCauseOfDeath <-
listaNombresMaestra[!esEstadoMaestraConDatosCauseOfDeath]

listaEstadosDatosCauseOfDeathSinEstadosMaestra <-
listaEstadosCauseOfDeath[!esDatoCauseOfDeathConEstadoMaestra]

tmp <- rawCdiAlcohol |>
dplyr::group_by(
  YearStart,
  LocationAbbr
) |>
dplyr::summarise(
  n = dplyr::n()
)
tmp$YearStart <- paste0('y', tmp$YearStart)

```

```

tmp <- tmp |>

tidyr::pivot_wider(names_from = YearStart, values_from = n)

tmp <- dplyr::left_join(rawFipsCodes, tmp, by = dplyr::join_by(state == LocationAbbr)) |>

dplyr::select(-state_code, -county, -county_code) |>

unique()

```

```

data.table	Variable	Descripción	Datos faltantes
----- ----- ----- -----			
`rawCdiAlcohol`	LocationAbbr	Datos CDI que no corresponden a un estado	`r listaEstadosDatosCdiSinEstadoMaestra`
`rawFipsCodes`	state	Estados sin datos en CDI	`r listaEstadosMaestraSinDatosCdi`
`rawFipsCodes`	state	Estados sin datos en Cause of Death	`r listaEstadosMaestraSinDatosCauseOfDeath`
`rawUnderlyingCauseOfDeathAlcohol`	State	Datos Cause of Death que no corresponden a un estado	`r listaEstadosDatosCauseOfDeathSinEstadosMaestra`

```{r}

```
#| code-fold: true
```

```
kableExtra::kable(tmp)
```

```

:::

```
::: {.callout-caution title="Cobertura incompleta de variables (columnas)" collapse="true"}
```

```
Cobertura incompleta de variables (columnas)
```

- Correcto - No faltan ninguna variable de interés

:::

::: {.callout-caution title="Cobertura incompleta de datos (`NA`)" collapse="true"}

##### Cobertura incompleta de datos (`NA`)

- Incorrecto - Existen datos faltantes

Se analizó el patrón de datos faltantes de los datos crudos con la función

`mice:::md.pattern()` :

```
```{r}
```

```
#| code-fold: true
```

```
mice:::md.pattern(rawCdiAlcohol, plot = T, rotate.names = T)
```

```

:::

::: {.callout-caution title="Cobertura incompleta de periodos (series temporales)" collapse="true"}

##### Cobertura incompleta de periodos (series temporales)

- No aplica en nuestro estudio

```
```{r}
```

```

:::

:::

::: {.callout-note title="5 - Datos imprecisos" collapse="true"}

#### ##### 5 - Datos imprecisos

Se han detectado los siguientes problemas de imprecisión de los datos:

- Tipado de variables incorrecto; deben retipificarse las variables tipo `character`

::: {.callout-caution title="Tipado de variable incorrecto" collapse="true"}

#### ##### Tipado de variable incorrecto

- Incorrecto - Deben retipificarse las variables tipo `character`

:::

::: {.callout-caution title="Precisión decimal insuficiente" collapse="true"}

#### ##### Precisión decimal insuficiente

- Correcto - No se han detectado problemas de precisión en números

:::

::: {.callout-caution title="Cardinalidad inadecuada (demasiadas o insuficientes categorías)" collapse="true"}

#### ##### Cardinalidad inadecuada (demasiadas o insuficientes categorías)

- Correcto - No se han detectado problemas de cardinalidad

:::

:::{.callout-caution title="Datos impuntuales \*(punctuality)\* (series temporales)" collapse="true"}

#### ##### Datos impuntuales \*(punctuality)\* (series temporales)

- No aplica (análisis de series temporales)

:::

:::{.callout-caution title="Datos desactualizados \*(freshness)\* (series temporales)" collapse="true"}

#### ##### Datos desactualizados \*(freshness)\* (series temporales)

- Correcto - Datos suficientemente actualizados para el propósito del estudio

:::

:::

:::{.callout-note title="6 - Datos incómodos" collapse="true"}

#### ##### 6 - Datos incómodos

- Correcto - No se han detectado problemas

:::{.callout-caution title="Formato inadecuado para el análisis (largo / ancho)" collapse="true"}

#### ##### Formato inadecuado para el análisis (largo / ancho)

- Correcto - No se han detectado problemas

:::

:::{.callout-caution title="Orden incorrecto de variables (columnas)" collapse="true"}

## ##### Orden incorrecto de variables (columnas)

- Correcto - No se han detectado problemas

:::

:::{.callout-caution title="Orden incorrecto de observaciones (filas)" collapse="true"}

## ##### Orden incorrecto de observaciones (filas)

- Correcto - No se han detectado problemas

:::

:::

## ##### Resumen del resultado de la validación

| Problema del dato                    | Subtipo de problema                                | Acción correctiva | Valoración del |
|--------------------------------------|----------------------------------------------------|-------------------|----------------|
| *dataset*                            |                                                    | Acción correctiva |                |
| Sucio                                | Codificación de caracteres incorrecta              |                   |                |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                   |                |
|                                      | Símbolos innecesarios (\\$, €, %, ... )            |                   |                |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                   |                |
|                                      | Nombre de *data.frame* no acorde a estilo          |                   |                |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                   |                |
|                                      | Nombre de variables (columnas) no acorde a estilo  |                   |                |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                   |                |
|                                      | Nombre de observaciones (filas) no acorde a estilo |                   | [*NO           |
| APLICA*]{style="color:darkgray;"}    |                                                    |                   |                |
| Incorrecto                           | Errores del subproceso de ingestá                  |                   |                |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                   |                |

|                                      |                                                                 |                                        |
|--------------------------------------|-----------------------------------------------------------------|----------------------------------------|
|                                      | Datos no ordenados - Más de una variable por columna            |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
|                                      | Datos no ordenados - Más de una observación por fila            |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
|                                      | Datos no ordenados - Más de un dato por registro                |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
| Irrelevante                          | Variables con todas las observaciones faltantes                 |                                        |
| [INCORRECTO]{style="color:darkred;"} | Eliminación de variables                                        |                                        |
|                                      | Variables con todas las observaciones con el mismo valor        |                                        |
| [INCORRECTO]{style="color:darkred;"} | Eliminación de variables                                        |                                        |
|                                      | Variables innecesarias para el análisis (columnas)              |                                        |
| [INCORRECTO]{style="color:darkred;"} | Eliminación de variables                                        |                                        |
|                                      | Observaciones innecesarias para el análisis (filas)             |                                        |
| [INCORRECTO]{style="color:darkred;"} | Agrupación de observaciones                                     |                                        |
| Incompleto                           | Cobertura incompleta de observaciones (filas)                   |                                        |
| [INCORRECTO]{style="color:darkred;"} | Selección de año 2021 (mejor cobertura)                         |                                        |
|                                      | Cobertura incompleta de variables (columnas)                    |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
|                                      | Cobertura incompleta de datos (`NA`)                            |                                        |
| [INCORRECTO]{style="color:darkred;"} | Análisis de variables completas                                 |                                        |
|                                      | Cobertura incompleta de periodos (series temporales)            |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
| Impreciso                            | Tipado de variable incorrecto                                   |                                        |
| [INCORRECTO]{style="color:darkred;"} | Retipado de variables                                           |                                        |
|                                      | Precisión decimal insuficiente                                  |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
|                                      | Cardinalidad inadecuada (demasiadas o insuficientes categorías) |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
|                                      | Datos impuntuales *(punctuality)* (series temporales)           | [*NO APlica*]{style="color:darkgray;"} |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |
|                                      | Datos desactualizados *(freshness)* (series temporales)         |                                        |
| [CORRECTO]{style="color:darkgreen;"} |                                                                 |                                        |

|                                                                |
|----------------------------------------------------------------|
| Incómodo   Formato inadecuado para el análisis (largo / ancho) |
| [CORRECTO]{style="color:darkgreen;"}                           |
| Orden incorrecto de variables (columnas)                       |
| [CORRECTO]{style="color:darkgreen;"}                           |
| Orden incorrecto de observaciones (filas)                      |
| [CORRECTO]{style="color:darkgreen;"}                           |

#### Validación de `rawUnderlyingCauseOfDeathAlcohol`

::: {.callout-note title="1- Información sucia" collapse="true"}

#### ##### 1- Información sucia

Se han detectado los siguientes problemas de suciedad de los datos:

- Variable ` % of Total Death` con valores seguidos del símbolo ` %`
- Variables con nombre no normalizado

::: {.callout-caution title="Codificación de caracteres incorrecta" collapse="true"}

#### ##### Codificación de caracteres incorrecta

- Correcto:
  - Los datasets se han ingestado en la codificación estándar \*\*UTF-8\*\*
  - No se han registrado advertencias durante la ingesta
  - La apariencia de los datos ingestados es correcta

:::

::: {.callout-caution title="Símbolos innecesarios (\$, €, %, ... )" collapse="true"}

#### ##### Símbolos innecesarios (\\$, €, %, ... )

- Incorrecto - Variable ` % of Total Death` con valores seguidos del símbolo ` %`

:::

::: {.callout-caution title="Nombre de \*data.frame\* no acorde a estilo" collapse="true"}

##### Nombre de \*data.frame\* no acorde a estilo

- Correcto - Nombre del \*\*data.frame\*\* en formato CamelCase

:::

::: {.callout-caution title="Nombre de variables (columnas) no acorde a estilo" collapse="true"}

##### Nombre de variables (columnas) no acorde a estilo

- Incorrecto - Variables con nombre no normalizado

:::

::: {.callout-caution title="Nombre de observaciones (filas) no acorde a estilo" collapse="true"}

##### Nombre de observaciones (filas) no acorde a estilo

- No aplica - Las filas no tienen nombre normalizado

:::

:::

::: {.callout-note title="2 - Datos incorrectos" collapse="true"}

##### 2 - Datos incorrectos

- Correcto - No se han identificado problemas de incorrección de datos

::: {.callout-caution title="Errores del subprocesso de ingestión" collapse="true"}

#### ##### Errores del subprocesso de ingestión

- Correcto - No se han detectado problemas de ingestión

:::

::: {.callout-caution title="Datos desestructurados - Más de una variable por columna" collapse="true"}

#### ##### Datos desestructurados - Más de una variables por columna

- Correcto - Cada variable está en una columna

:::

::: {.callout-caution title="Datos desestructurados - Más de una observación por fila" collapse="true"}

#### ##### Datos desestructurados - Más de una observación por fila

- Correcto - Cada observación está en una fila

:::

:::

::: {.callout-note title="3 - Datos irrelevantes" collapse="true"}

#### ##### 3 - Datos irrelevantes

Se han identificado los siguientes valores irrelevantes:

- Deben eliminarse varias variables innecesarias para el análisis

:::{.callout-caution title="Variables con todas las observaciones faltantes" collapse="true"}

#### ##### Variables con todas las observaciones faltantes

- Correcto - No se han identificado variables con todas las observaciones faltantes

:::

:::{.callout-caution title="Variables con todas las observaciones con el mismo valor" collapse="true"}

#### ##### Variables con todas las observaciones con el mismo valor

- Correcto - No se han identificado variables con todas las observaciones con el mismo valor

:::

:::{.callout-caution title="Observaciones innecesarias para el análisis (filas)" collapse="true"}

#### ##### Observaciones innecesarias para el análisis (filas)

- Incorrecto: deben eliminarse las observaciones con el valor 'Total' en la variable `Notes`, los valores redundantes de la variable `Stratification1` y conservarse únicamente las observaciones del año '2021'

:::

:::{.callout-caution title="Variables innecesarias para el análisis (columnas)" collapse="true"}

#### ##### Variables innecesarias para el análisis (columnas)

- Incorrecto: Varias variables tienen información innecesaria o redundante para el análisis

Las siguientes variables del dataset original no son necesarias:

| Variable                                          | Justificación                                                 |
|---------------------------------------------------|---------------------------------------------------------------|
| `Notes`                                           | Irrelevante para el análisis                                  |
| `State Code`                                      | Información redundante con `State`                            |
| `Year`                                            | Información redundante con `Year Code`                        |
| `Crude Rate`                                      | Irrelevante para el análisis                                  |
| `Crude Rate Lower 95% Confidence Interval`        | Irrelevante para el análisis                                  |
| `Crude Rate Upper 95% Confidence Interval`        | Irrelevante para el análisis                                  |
| `Crude Rate Standard Error`                       | Irrelevante para el análisis                                  |
| `Age Adjusted Rate Lower 95% Confidence Interval` | Información redundante con `Age Adjusted Rate Standard Error` |
| `Age Adjusted Rate Upper 95% Confidence Interval` | Información redundante con `Age Adjusted Rate Standard Error` |
| :::                                               |                                                               |
| :::                                               |                                                               |

::: {.callout-note title="4 - Datos incompletos" collapse="true"}

#### ##### 4 - Datos incompletos

- Correcto - No se han detectado problemas

```
::: {.callout-caution title="Cobertura incompleta de observaciones (filas)" collapse="true"}
```

```
Cobertura incompleta de observaciones (filas)
```

- Correcto - Todos los estados tienen datos en el dataset para el año seleccionado

```
:::
```

```
::: {.callout-caution title="Cobertura incompleta de variables (columnas)" collapse="true"}
```

```
Cobertura incompleta de variables (columnas)
```

- Correcto - Están disponibles todas las variables necesarias

```
:::
```

```
::: {.callout-caution title="Cobertura incompleta de datos (`NA`)" collapse="true"}
```

```
Cobertura incompleta de datos (`NA`)
```

- Correcto - No existen problemas de datos faltantes. Los únicos `NA` están asociados a las filas de totales

```
```{r}
```

```
#| code-fold: true
```

```
mice:::md.pattern(rawUnderlyingCauseOfDeathAlcohol, plot = T, rotate.names = T)
```

```
```
```

```
:::
```

```
::: {.callout-caution title="Cobertura incompleta de periodos (series temporales)" collapse="true"}
```

## ##### Cobertura incompleta de periodos (series temporales)

- No aplica

:::

:::

:::{.callout-note title="5 - Datos imprecisos" collapse="true"}

## ##### 5 - Datos imprecisos

Se han detectado los siguientes problemas de imprecisión:

- Se debe cambiar el tipado de las variables tipo `character`

:::{.callout-caution title="Tipado de variable incorrecto" collapse="true"}

## ##### Tipado de variable incorrecto

- Incorrecto - Se debe cambiar el tipado de las variables tipo `character`

:::

:::{.callout-caution title="Precisión decimal insuficiente" collapse="true"}

## ##### Precisión decimal insuficiente

- Correcto - No se detectan problemas

:::

:::{.callout-caution title="Cardinalidad inadecuada (demasiadas o insuficientes categorías)" collapse="true"}

## ##### Cardinalidad inadecuada (demasiadas o insuficientes categorías)

- Correcto - No se detectan problemas

:::

::: {.callout-caution title="Datos impuntuales \*(punctuality)\* (series temporales)" collapse="true"}

##### Datos impuntuales \*(punctuality)\* (series temporales)

- No aplica

:::

::: {.callout-caution title="Datos desactualizados \*(freshness)\* (series temporales)" collapse="true"}

##### Datos desactualizados \*(freshness)\* (series temporales)

- No aplica

:::

:::

::: {.callout-note title="6 - Datos incómodos" collapse="true"}

##### 6 - Datos incómodos

::: {.callout-caution title="Formato inadecuado para el análisis (largo / ancho)" collapse="true"}

##### Formato inadecuado para el análisis (largo / ancho)

- Correcto - No se detectan problemas

:::

::: {.callout-caution title="Orden incorrecto de variables (columnas)" collapse="true"}

#### ##### Orden incorrecto de variables (columnas)

- Correcto - No se detectan problemas

:::

::: {.callout-caution title="Orden incorrecto de observaciones (filas)" collapse="true"}

#### ##### Orden incorrecto de observaciones (filas)

- Correcto - No se detectan problemas

:::

:::

#### ##### Resumen del resultado de la validación

| Problema del dato                    | Subtipo de problema                                | Acción                     | Problemas del |
|--------------------------------------|----------------------------------------------------|----------------------------|---------------|
| *dataset*                            |                                                    |                            |               |
| -----                                | -----                                              | -----                      | -----         |
| Sucio                                | Codificación de caracteres incorrecta              |                            |               |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                            |               |
|                                      | Símbolos innecesarios (\\$, €, %, ... )            |                            |               |
| [INCORRECTO]{style="color:darkred;"} |                                                    | Transformación de variable |               |
|                                      | Nombre de *data.frame* no acorde a estilo          |                            |               |
| [CORRECTO]{style="color:darkgreen;"} |                                                    |                            |               |
|                                      | Nombre de variables (columnas) no acorde a estilo  |                            |               |
| [INCORRECTO]{style="color:darkred;"} |                                                    | Renombrado de variables    |               |
|                                      | Nombre de observaciones (filas) no acorde a estilo |                            | [*NO          |
| APLICA*]{style="color:darkgray;"}    |                                                    |                            |               |

|                                                                  |                                                 |                                        |
|------------------------------------------------------------------|-------------------------------------------------|----------------------------------------|
| Incorrecto                                                       | Errores del subproceso de ingestá               |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Datos no ordenados - Más de una variables por columna            |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Datos no ordenados - Más de una observación por fila             |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Irrelevante                                                      | Variables con todas las observaciones faltantes |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Variables con todas las observaciones con el mismo valor         |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Observaciones innecesarias para el análisis (filas)              |                                                 |                                        |
| [INCORRECTO]{style="color:darkred;"}   Filtrado de observaciones |                                                 |                                        |
| Variables innecesarias para el análisis (columnas)               |                                                 |                                        |
| [INCORRECTO]{style="color:darkred;"}   Eliminación de variables  |                                                 |                                        |
| Incompleto                                                       | Cobertura incompleta de observaciones (filas)   |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Cobertura incompleta de variables (columnas)                     |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Cobertura incompleta de datos (`NA`)                             |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Cobertura incompleta de periodos (series temporales)             |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Impreciso                                                        | Tipado de variable incorrecto                   |                                        |
| [INCORRECTO]{style="color:darkred;"}   Retipado de variables     |                                                 |                                        |
| Precisión decimal insuficiente                                   |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Cardinalidad inadecuada (demasiadas o insuficientes categorías)  |                                                 |                                        |
| [CORRECTO]{style="color:darkgreen;"}                             |                                                 |                                        |
| Datos impuntuales *(punctuality)* (series temporales)            | [*NO APLICA*]{style="color:darkgray;"}          | [*NO APLICA*]{style="color:darkgray;"} |
| Datos desactualizados *(freshness)* (series temporales)          |                                                 | [*NO APLICA*]{style="color:darkgray;"} |

|                                      |                                                     |  |
|--------------------------------------|-----------------------------------------------------|--|
| Incómodo                             | Formato inadecuado para el análisis (largo / ancho) |  |
| [CORRECTO]{style="color:darkgreen;"} |                                                     |  |
|                                      | Orden incorrecto de variables (columnas)            |  |
| [CORRECTO]{style="color:darkgreen;"} |                                                     |  |
|                                      | Orden incorrecto de observaciones (filas)           |  |
| [CORRECTO]{style="color:darkgreen;"} |                                                     |  |

#### Validación de `rawFipsCodes`

::: {.callout-note title="1- Información sucia" collapse="true"}

#### ##### 1- Información sucia

Se han encontrado los siguientes problemas con la suciedad de los datos:

- Variables con nombre en formato snake\_case

::: {.callout-caution title="Codificación de caracteres incorrecta" collapse="true"}

#### ##### Codificación de caracteres incorrecta

- Correcto:

- Los datasets se han ingestado en la codificación estándar \*\*UTF-8\*\*

- No se han registrado advertencias durante la ingesta

- La apariencia de los datos ingestados es correcta

:::

::: {.callout-caution title="Símbolos innecesarios (\$, €, %, ...)" collapse="true"}

## ##### Símbolos innecesarios (\\$, €, %, ... )

- Correcto - Sin problemas en ninguna de las variables

:::

::: {.callout-caution title="Nombre de \*data.frame\* no acorde a estilo" collapse="true"}

## ##### Nombre de \*data.frame\* no acorde a estilo

- Correcto - Nombre del \*\*data.frame\*\* en formato CamelCase

:::

::: {.callout-caution title="Nombre de variables (columnas) no acorde a estilo" collapse="true"}

## ##### Nombre de variables (columnas) no acorde a estilo

- Incorrecto - Variables con nombre en formato snake\_case

:::

::: {.callout-caution title="Nombre de observaciones (filas) no acorde a estilo" collapse="true"}

## ##### Nombre de observaciones (filas) no acorde a estilo

- No aplica - Las filas no tienen nombre

:::

:::

::: {.callout-note title="2 - Datos incorrectos" collapse="true"}

## ##### 2 - Datos incorrectos

- Correcto - No se han detectado problemas de incorrección de datos

::: {.callout-caution title="Errores del subprocesso de ingestión" collapse="true"}

#### ##### Errores del subprocesso de ingestión

- Correcto - No se han detectado problemas de ingestión

:::

::: {.callout-caution title="Datos desestructurados - Más de una variable por columna" collapse="true"}

#### ##### Datos desestructurados - Más de una variables por columna

- Correcto - Una variable por columna

:::

::: {.callout-caution title="Datos desestructurados - Más de una observación por fila" collapse="true"}

#### ##### Datos desestructurados - Más de una observación por fila

- Correcto - Una observación por fila

:::

:::

::: {.callout-note title="3 - Datos irrelevantes" collapse="true"}

#### ##### 3 - Datos irrelevantes

- Correcto - No se han identificado problemas de irrelevancia de datos

:::{.callout-caution title="Variables con todas las observaciones faltantes" collapse="true"}

##### Variables con todas las observaciones faltantes

- Correcto - No hay problemas de datos faltantes

:::

:::{.callout-caution title="Variables con todas las observaciones con el mismo valor" collapse="true"}

##### Variables con todas las observaciones con el mismo valor

- Correcto - No hay variables con todas las observaciones iguales

:::

:::{.callout-caution title="Variables innecesarias para el análisis (columnas)" collapse="true"}

##### Variables innecesarias para el análisis (columnas)

- Incorrecto: Varias variables tienen información innecesaria o redundante para el análisis

Las siguientes variables del dataset original no son necesarias:

| Variable | Justificación |
|----------|---------------|
|----------|---------------|

|       |       |
|-------|-------|
| ----- | ----- |
|-------|-------|

|               |                              |
|---------------|------------------------------|
| `county_code` | Irrelevante para el análisis |
|---------------|------------------------------|

|          |                              |
|----------|------------------------------|
| `county` | Irrelevante para el análisis |
|----------|------------------------------|

:::

::: {.callout-caution title="Observaciones innecesarias para el análisis (filas)" collapse="true"}

##### Observaciones innecesarias para el análisis (filas)

:::

:::

::: {.callout-note title="4 - Datos incompletos" collapse="true"}

##### 4 - Datos incompletos

::: {.callout-caution title="Cobertura incompleta de observaciones (filas)" collapse="true"}

##### Cobertura incompleta de observaciones (filas)

:::

::: {.callout-caution title="Cobertura incompleta de variables (columnas)" collapse="true"}

##### Cobertura incompleta de variables (columnas)

:::

::: {.callout-caution title="Cobertura incompleta de datos (`NA`)" collapse="true"}

##### Cobertura incompleta de datos (`NA`)

:::

::: {.callout-caution title="Cobertura incompleta de periodos (series temporales)" collapse="true"}

##### Cobertura incompleta de periodos (series temporales)

:::

:::

::: {.callout-note title="5 - Datos imprecisos" collapse="true"}

#### ##### 5 - Datos imprecisos

::: {.callout-caution title="Tipado de variable incorrecto" collapse="true"}

#### ##### Tipado de variable incorrecto

:::

::: {.callout-caution title="Precisión decimal insuficiente" collapse="true"}

#### ##### Precisión decimal insuficiente

:::

::: {.callout-caution title="Cardinalidad inadecuada (demasiadas o insuficientes categorías)" collapse="true"}

#### ##### Cardinalidad inadecuada (demasiadas o insuficientes categorías)

:::

::: {.callout-caution title="Datos impuntuales \*(punctuality)\* (series temporales)" collapse="true"}

#### ##### Datos impuntuales \*(punctuality)\* (series temporales)

:::

::: {.callout-caution title="Datos desactualizados \*(freshness)\* (series temporales)" collapse="true"}

#### ##### Datos desactualizados \*(freshness)\* (series temporales)

:::

:::

::: {.callout-note title="6 - Datos incómodos" collapse="true"}

## ##### 6 - Datos incómodos

:::{.callout-caution title="Formato inadecuado para el análisis (largo / ancho)" collapse="true"}

##### Formato inadecuado para el análisis (largo / ancho)

:::

:::{.callout-caution title="Orden incorrecto de variables (columnas)" collapse="true"}

##### Orden incorrecto de variables (columnas)

:::

:::{.callout-caution title="Orden incorrecto de observaciones (filas)" collapse="true"}

##### Orden incorrecto de observaciones (filas)

:::

:::

## ##### Resumen del resultado de la validación

| Problema del dato   Subtipo de problema           | Problemas del                         |
|---------------------------------------------------|---------------------------------------|
| *dataset*                                         | Acción                                |
| ----- -----                                       | ----- -----                           |
| Sucio                                             | Codificación de caracteres incorrecta |
| [CORRECTO]{style="color:darkgreen;"}              |                                       |
| Símbolos innecesarios (\\$, €, %, ... )           |                                       |
| [CORRECTO]{style="color:darkgreen;"}              |                                       |
| Nombre de *data.frame* no acorde a estilo         |                                       |
| [CORRECTO]{style="color:darkgreen;"}              |                                       |
| Nombre de variables (columnas) no acorde a estilo |                                       |
| [INCORRECTO]{style="color:darkred;"}              | Renombrado de variables               |

|                                   |                                                                                                         |                          |
|-----------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------|
|                                   | Nombre de observaciones (filas) no acorde a estilo<br>APLICA*]{style="color:darkgray;"}                 | [*NO                     |
| Incorrecto                        | Errores del subprocesso de ingestá<br>[CORRECTO]{style="color:darkgreen;"}                              |                          |
|                                   | Datos no ordenados - Más de una variables por columna<br>[CORRECTO]{style="color:darkgreen;"}           |                          |
|                                   | Datos no ordenados - Más de una observación por fila<br>[CORRECTO]{style="color:darkgreen;"}            |                          |
|                                   | Datos no ordenados - Más de un dato por registro<br>[CORRECTO]{style="color:darkgreen;"}                |                          |
| Irrelevante                       | Variables con todas las observaciones faltantes<br>[CORRECTO]{style="color:darkgreen;"}                 |                          |
|                                   | Variables con todas las observaciones con el mismo valor<br>[CORRECTO]{style="color:darkgreen;"}        |                          |
|                                   | Variables innecesarias para el análisis (columnas)<br>[INCORRECTO]{style="color:darkred;"}              | Eliminación de variables |
|                                   | Observaciones innecesarias para el análisis (filas)<br>[CORRECTO]{style="color:darkgreen;"}             |                          |
| Incompleto                        | Cobertura incompleta de observaciones (filas)<br>[CORRECTO]{style="color:darkgreen;"}                   |                          |
|                                   | Cobertura incompleta de variables (columnas)<br>[CORRECTO]{style="color:darkgreen;"}                    |                          |
|                                   | Cobertura incompleta de datos (`NA`)                                                                    | [*NO                     |
| APLICA*]{style="color:darkgray;"} |                                                                                                         |                          |
|                                   | Cobertura incompleta de periodos (series temporales)<br>[CORRECTO]{style="color:darkgreen;"}            |                          |
| Impreciso                         | Tipado de variable incorrecto<br>[CORRECTO]{style="color:darkgreen;"}                                   |                          |
|                                   | Precisión decimal insuficiente<br>[CORRECTO]{style="color:darkgreen;"}                                  |                          |
|                                   | Cardinalidad inadecuada (demasiadas o insuficientes categorías)<br>[CORRECTO]{style="color:darkgreen;"} |                          |

|                                      |                                                         |                                        |  |
|--------------------------------------|---------------------------------------------------------|----------------------------------------|--|
|                                      | Datos impuntuales *(punctuality)* (series temporales)   | [*NO APLICA*]{style="color:darkgray;"} |  |
|                                      | Datos desactualizados *(freshness)* (series temporales) | [*NO APLICA*]{style="color:darkgray;"} |  |
| Incómodo                             | Formato inadecuado para el análisis (largo / ancho)     |                                        |  |
| [CORRECTO]{style="color:darkgreen;"} |                                                         |                                        |  |
|                                      | Orden incorrecto de variables (columnas)                |                                        |  |
| [CORRECTO]{style="color:darkgreen;"} |                                                         |                                        |  |
|                                      | Orden incorrecto de observaciones (filas)               |                                        |  |
| [CORRECTO]{style="color:darkgreen;"} |                                                         |                                        |  |

### 02b - Reingestar, modificar, reemplazar o borrar esta información no deseada de acuerdo a la necesidad

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
# Creación de los data.frame temporal para la limpieza
```

```
tmpCdiAlcohol <- rawCdiAlcohol
```

```
tmpUnderlyingCauseOfDeathAlcohol <- rawUnderlyingCauseOfDeathAlcohol
```

```
tmpFipsCodes <- rawFipsCodes
```

```
```
```

#### 1- Limpieza de `rawCdiAlcohol`

::: {.callout-note title="1a - Eliminación de datos irrelevantes" collapse="true"}

##### 1a - Eliminación de datos irrelevantes

Se realizaron las siguientes acciones para corregir los datos irrelevantes:

- 1aa - Eliminación de variables sin datos
- 1ab - Eliminación de variables con todas las observaciones con el mismo valor
- 1ac - Eliminación de variables innecesarias
- 1ad - Eliminación de observaciones innecesarias
- 1ae - Limpieza final

:::{.callout-caution title="1a1 - Eliminación de variables sin datos" collapse="true"}

#### ##### 1aa - Eliminación de variables sin datos

Se analizó el patrón de datos faltantes de los datos crudos con la función `mice:::md.pattern()` :

```
```{r}
## code-fold: true
## output: false

mice:::md.pattern(tmpCdiAlcohol, plot = T, rotate.names = T)
```
```

Se observó que:

- Diez variables no tenían ningún valor en ninguna observación
- En tres de ellas había un elevado número de `NA` .
- El resto de variables no presentaba problemas de valores faltantes

Para identificar las variables sin datos, se utilizó el siguiente procedimiento:

- Paso 1 - Valoración del número total de datos faltantes de cada variable (`NA` )

- Paso 2 - Identificación de las variables cuyo número total de `NA` sea igual al número de observaciones (filas) del dataset
- Paso 3 - Eliminación de las variables sin datos

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
## Identificación de variables sin ningún dato
```

```
### Paso 1 - Valoración del número total de datos faltantes de cada variable (`NA`)
```

```
nMissingDataByCol <- colSums(is.na(tmpCdiAlcohol))
```

```
### Paso 2 - Identificación de las variables cuyo número total de `NA` sea igual al número de observaciones (filas) del dataset
```

```
namesVarNoData <-
```

```
names(nMissingDataByCol)[nMissingDataByCol == nrow(tmpCdiAlcohol)]
```

```
idxVarNoData <- names(tmpCdiAlcohol) %in% namesVarNoData
```

```
namesVarData <- names(tmpCdiAlcohol)[!idxVarNoData]
```

```
### Paso 3 - Eliminación de las variables sin datos
```

```
tmpCdiAlcohol <- tmpCdiAlcohol |>
```

```
dplyr::select(
```

```
dplyr::all_of(namesVarData)
```

```
)
```

```
mice::md.pattern(tmpCdiAlcohol, plot = T, rotate.names = T)
```

```
```
```

```
:::
```

```
:::{.callout-caution title="1ab - Eliminación de variables con todas las observaciones con el mismo valor" collapse="true"}
1ab - Eliminación de variables con todas las observaciones con el mismo valor
```

Para eliminar las variables con todas las observaciones iguales, se utilizó el siguiente procedimiento:

- Para las variables categóricas:
  - Paso 1 - Identificar las variables de tipo `character`
  - Paso 2 - Evaluar la cardinalidad (número de categorías) de cada variable
  - Paso 3 - Eliminar las variables con cardinalidad de 1

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
# Eliminación de variables con todas las observaciones con el mismo valor  
## Paso 1 - Identificar las variables de tipo `character`  
### Tipo de variables del dataframe  
sapply(tmpCdiAlcohol, class)  
### Número de tipos de variable en el dataframe  
table(sapply(tmpCdiAlcohol, class))  
### Variables de tipo character  
varCharacterNames <- names(which(sapply(tmpCdiAlcohol, is.character)))  
isCharacter <- names(tmpCdiAlcohol) %in% varCharacterNames  
  
## Paso 2 - Evaluar la cardinalidad (número de categorías) de cada variable
```

```

### Función para calcular la cardinalidad de una variable
cardinality <- function(x){length(levels(as.factor(x)))}

### Cardinalidad de las variables categóricas
cardinalityCdiAlcohol <- tmpCdiAlcohol |>
  dplyr::select(
    dplyr::all_of(varCharacterNames)
  ) |>
  sapply(cardinality)

## Paso 3 - Eliminar las variables con cardinalidad de 1
### Variables categóricas con cardinalidad de 0 (sin valores) o 1 (con todos iguales)
esVarCardinalidadBaja <- cardinalityCdiAlcohol %in% c(0,1)
namesVarCardinalidadNormal <- names(tmpCdiAlcohol)[!esVarCardinalidadBaja]
tmpCdiAlcohol <- tmpCdiAlcohol |>
  dplyr::select(
    dplyr::all_of(namesVarCardinalidadNormal)
  )
```
:::
```
```

::: {.callout-caution title="1ac - Eliminación de variables innecesarias para el análisis" collapse="true"}

1ac - Eliminación de variables innecesarias para el análisis

Se eliminaron las siguientes variables del dataset original:

Variable	Justificación	
----------	---------------	--

----- -----		
`YearEnd`	Información redundante con `YearStart`	
`DataSource`	Irrelevante para el análisis	
`DataValue`	Información redundante con `DataValueAlt`	
`DataValueFootnoteSymbol` `DatavalueFootnote`	Irrelevante para el análisis	
`GeoLocation`	Irrelevante para el análisis	
`LocationID`	Irrelevante para el análisis	
`DataValueTypeID`	Información redundante con `DataValueType`	
`DataValueUnit`	Irrelevante para el análisis	

A pesar de que tenían información idéntica, se mantuvieron en el dataset las siguientes variables, para facilitar el etiquetado de gráficos con etiquetas cortas:

| Variable A (descripción larga) | Variable B (etiqueta corta) |

----- -----		
-------------	--	--

| `LocationDesc` | `LocationAbbr` |

| `Question` | `QuestionID` |

| `StratificationCategory1` | `StratificationCategoryID1` |

| `Stratification1` | `StratificationID1` |

```{r}

#| code-fold: true

#| output: false

# Variables a mantener en el dataset

varCdiAlcohol <- c(

"YearStart",

```

 ##"YearEnd",
 "LocationAbbr",
 ##"DataSource",
 "Question",
 "DataValueType",
 "DataValueAlt",
 ##"DataValueUnit",
 ##"DataValue",
 ##"DataValueFootnoteSymbol",
 ##"DatavalueFootnote",
 ##"LowConfidenceLimit",
 ##"HighConfidenceLimit",
 "StratificationCategory1",
 "Stratification1"# ,
 ##"StratificationID1"

 ##"GeoLocation",
 ##"LocationID",
 # "DataValueTypeID",
)

Selección de variables relevantes
tmpCdiAlcohol <- tmpCdiAlcohol |>
 dplyr::select(
 dplyr::all_of(varCdiAlcohol)
)

```

```

:::

::: {.callout-caution title="1ad - Eliminación de observaciones innecesarias para el análisis (filas)" collapse="true"}

1ad - Eliminación de observaciones innecesarias para el análisis (filas)

Se filtraron las siguientes observaciones del dataset original:

Variable	Filtro aplicado	Justificación
StratificationCategory1 ` %in% c("Gender", "Overall")`		La estratificación 'Race/Ethnicity' no tiene datos para un número significativo de estados
DataValueTypeID ` %in% c("AGEADJMEAN", "AGEADJPREV", "AGEADJRATE")`	Las tasas ajustadas por edad son más adecuadas para comparar indicadores entre poblaciones con pirámides de población diferentes	
YearStart ` == 2021`		Año con datos más recientes, y con una mayor completitud de datos para todos los estados
LocationAbbr ` != "US"`		Datos correspondientes al país completo

Al revisar el patrón de datos faltantes de los resultados, se comprobó que:

- El estado de Florida tenía una tasa de datos faltantes muy superior al de los demás estados, para todos los indicadores
- La estratificación 'Race/Ethnicity' no tiene datos para un número significativo de estados
- Los niveles de 'Race' son redundantes (Male, Female, Overall)

```{r}

```

#| code-fold: true
#| output: false

Datos faltantes por estado

tmpCdiAlcohol[is.na(tmpCdiAlcohol$LocationAbbr),] |>
 dplyr::group_by(LocationAbbr) |>
 dplyr::summarise(n = dplyr::n())

Datos faltantes por tipo de estratificación

tmpCdiAlcohol[is.na(tmpCdiAlcohol$DataValueAlt),] |>
 dplyr::group_by(StratificationCategory1) |>
 dplyr::summarise(n = dplyr::n())

```

```

Por este motivo,

- Se eliminaron los datos de la estratificación por raza y etnia

```

```
{r}

#| code-fold: true
#| output: false

tmpCdiAlcohol_overall <- tmpCdiAlcohol |>
 dplyr::filter(
 Stratification1 == 'Overall'
) |>
 dplyr::select(
 -StratificationCategory1,

```

```

-Stratification1
)

tmpCdiAlcohol_byGender <- tmpCdiAlcohol |>

dplyr::filter(
 Stratification1 != 'Overall'
) |>
dplyr::select(
 -StratificationCategory1
)

tmpCdiAlcohol <- tmpCdiAlcohol |>

dplyr::filter(
 StratificationCategory1 %in% c("Gender", "Overall"),
 DataValueType %in% c("Age-adjusted Mean", "Age-adjusted Prevalence", "Age-
adjusted Rate"),
 YearStart == 2021,
 LocationAbbr != "US",
 # Stratification1 != "Overall"
)

```
```
```
```
{r}

#| code-fold: true
#| output: false

tmpCdiAlcohol <- tmpCdiAlcohol |>
```

```

dplyr::filter(
 StratificationCategory1 %in% c("Gender", "Overall"),
 DataValueType %in% c("Age-adjusted Mean", "Age-adjusted Prevalence", "Age-
adjusted Rate"),
 LocationAbbr != "FL",
 YearStart == 2021
)
mice::md.pattern(tmpCdiAlcohol, plot = T, rotate.names = T)
```

```

Tras estas transformaciones, sólo un estado del dataset final queda sin datos (Florida), lo que se tendrá en cuenta al interpretar el análisis.

```

```{r}
#| code-fold: true
tmpCdiAlcohol[is.na(tmpCdiAlcohol$DataValueAlt),]
```
:::
### {.callout-caution title="1ae - Limpieza final" collapse="true"}
##### 1ae - Limpieza final

```

Una vez eliminados los datos, innecesarios, algunas variables se vuelven superfluas:

| Variables eliminadas Transformación |
|--|
| ----- ----- |
| `DataValueType` Información Innecesaria para el análisis |

```

```{r}
#| code-fold: true
#| output: false

tmpCdiAlcohol <- tmpCdiAlcohol |>

dplyr::select(
 #-StratificationCategory1,
 #-StratificationCategoryID1,
 -DataValueType
)
```
:::

:::
:::

#### 1b - Retipado de variables

##### 1b - Retipado de variables
```

Se pospuso la tarea hasta la fusión de datasets de trabajo.

:::

2 - Limpieza de `rawUnderlyingCauseOfDeathAlcohol`

```

#### 2a - Limpieza de datos sucios

##### 2a - Limpieza de datos sucios
```

Se realizaron las siguientes acciones para corregir los datos sucios:

- 2aa - Transformación de la variable ` % of Total Death`

- 2ab - Renombrado de variables

```
:::{.callout-caution title="2aa - Transformación de la variable `% of Total Death`" collapse="true"}
```

```
##### 2aa - Transformación de la variable `% of Total Death`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
2aa - Transformación de la variable `% of Total Death`
```

```
tmpUnderlyingCauseOfDeathAlcohol$`% of Total Deaths` <- gsub(
```

```
pattern = "%",
```

```
replacement = "",
```

```
tmpUnderlyingCauseOfDeathAlcohol$`% of Total Deaths`
```

```
) |>
```

```
as.numeric()
```

```
```
```

```
:::
```

```
:::{.callout-caution title="2ab - Renombrado de variables" collapse="true"}
```

```
##### 2ab - Renombrado de variables
```

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
Nombres originales
```

```

rawNamesUnderlyingCauseOfDeath <- names(tmpUnderlyingCauseOfDeathAlcohol)

Nombres normalizados

names(tmpUnderlyingCauseOfDeathAlcohol) <- c(
 "Notes",
 "State",
 "StateCode",
 "Year",
 "YearCode",
 "Gender",
 "GenderCode",
 "Deaths",
 "Population",
 "CrudeRate",
 "CrudeRateLower95ConfidenceInterval",
 "CrudeRateUpper95ConfidenceInterval",
 "CrudeRateStandardError",
 "AgeAdjustedRate",
 "AgeAdjustedRateLower95ConfidenceInterval",
 "AgeAdjustedRateUpper95ConfidenceInterval",
 "AgeAdjustedRateStandardError",
 "PercentageOfTotalDeaths"
)

```
:::
```
:::

```

::: {.callout-note title="2b - Eliminación de información irrelevante" collapse="true"}

## ##### 2b - Eliminación de información irrelevante

Se realizaron las siguientes acciones para corregir los datos irrelevantes:

- 2ba - Eliminación de observaciones innecesarias
- 2bb - Eliminación de variables innecesarias

:::{.callout-caution title="2ba - Eliminación de observaciones innecesarias" collapse="true"}

### ##### 2ba - Eliminación de observaciones innecesarias

Se seleccionaron las observaciones del año 2021

```
```{r}
#| code-fold: true
#| output: false
```

```
tmpUnderlyingCauseOfDeathAlcohol <- tmpUnderlyingCauseOfDeathAlcohol |>
```

```
dplyr::filter(
  YearCode == 2021
)
```

```
```
```

```
:::
```

:::{.callout-caution title="2bb - Eliminación de variables innecesarias" collapse="true"}

### ##### 2bb - Eliminación de variables innecesarias

Se eliminaron las siguientes variables del dataset original:

| Variable                                                                                                                                                                   | Justificación |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
|                                                                                                                                                                            |               |
| ----- -----                                                                                                                                                                |               |
| ` Notes`<br>para el análisis                                                                                                                                               | Irrelevante   |
| ` State`<br>redundante con ` State Code`                                                                                                                                   | Información   |
| ` Year`<br>` YearCode`<br>Irrelevante para el análisis<br>Sólo se va analizar el año 2021                                                                                  |               |
| ` Crude Rate`<br>` Crude Rate Lower 95% Confidence Interval`<br>` Crude Rate Upper 95% Confidence Interval`<br>` Crude Rate Standard Error`   Irrelevante para el análisis |               |
| ` Age Adjusted Rate Lower 95% Confidence Interval`<br>` Age Adjusted Rate Upper 95% Confidence Interval`   Información redundante con ` Age Adjusted Rate Standard Error`  |               |

```
```{r}
#| code-fold: true
#| output: false
```

```
tmpUnderlyingCauseOfDeathAlcohol <- tmpUnderlyingCauseOfDeathAlcohol |>
  dplyr::select(
    # Notes,
    # State,
    StateCode,
    # Year,
    # YearCode,
    Gender,
    # GenderCode,
```

```
Deaths,  
Population,  
# CrudeRate,  
# CrudeRateLower95ConfidenceInterval,  
# CrudeRateUpper95ConfidenceInterval,  
# CrudeRateStandardError,  
AgeAdjustedRate,  
# AgeAdjustedRateLower95ConfidenceInterval,  
# AgeAdjustedRateUpper95ConfidenceInterval,  
AgeAdjustedRateStandardError,  
PercentageOfTotalDeaths  
)
```

```

:::

:::

```
2c - Retipado de variables
```

```{r}

```
#| code-fold: true
```

```
#| output: false
```

```
tmpUnderlyingCauseOfDeathAlcohol$AgeAdjustedRate <-
```

```
tmpUnderlyingCauseOfDeathAlcohol$AgeAdjustedRate |> as.numeric()
```

```

```
3 - Limpieza de `rawFipsCodes`
```

```
::: {.callout-note title="3a - Limpieza de datos sucios" collapse="true"}
```

#### ##### 3a - Limpieza de datos sucios

Se realizaron las siguientes acciones para corregir los datos sucios:

- 3aa - Renombrado de variables

```
::: {.callout-caution title="3aa - Renombrado de variables" collapse="true"}
```

#### ##### 3aa - Renombrado de variables

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
rawNamesFipsCodes <- names(tmpFipsCodes)
```

```
names(tmpFipsCodes) <- c("State", "StateCode", "StateName", "CountyCode",  
"County" )
```

```
...
```

```
:::
```

```
:::
```

```
::: {.callout-note title="3b - Eliminación de datos irrelevantes" collapse="true"}
```

3b - Eliminación de datos irrelevantes

```
::: {.callout-caution title="3ba - Eliminación de variables innecesarias" collapse="true"}
```

3ba - Eliminación de variables innecesarias

Se eliminaron las siguientes variables del dataset original:

Variable	Justificación
`county_code`	Irrelevante para el análisis
`county`	Irrelevante para el análisis

```{r}

```
#| code-fold: true
#| output: false
```

```
tmpFipsCodes <- tmpFipsCodes |>
```

```
dplyr::select(
 State,
 StateCode,
 StateName #,
 # CountyCode,
 # County
```

```
) |>
```

```
unique()
```

```

:::

:::

4 - Combinación de los tres *data.frame* originales

Una vez limpios los tres data.frame de origen, se crearon tres conjuntos de datos para el análisis:

- `data`: Datos completos
- `data_overall`: Datos globales, sin estratificación por sexo
- `data_gender`: Datos estratificados por sexo

```
```{r}
```

```
#| code-fold: true
```

```
#| output: false
```

```
tmp <- dplyr::left_join(
 tmpFipsCodes,
 tmpCdiAlcohol,
 by = dplyr::join_by(State == LocationAbbr)
)
tmp$StateCode <- as.numeric(tmp$StateCode)
```

```
tmp_overall <- tmp |>
dplyr::filter(
 Stratification1 == "Overall"
)
tmp_gender <- tmp |>
dplyr::filter(
 Stratification1 != "Overall"
)
```

```
tmp <- dplyr::left_join(
 tmp,
 tmpUnderlyingCauseOfDeathAlcohol,
 by = dplyr::join_by(
```

```

StateCode,
Stratification1 == Gender
)
) |>
dplyr::select(
 State,
 # StateCode,
 StateName,
 # YearStart,
 # LocationDesc,
 # QuestionID,
 Question,
 QuestionValue = DataValueAlt,
 #QuestionValueLowConfidenceLimit = LowConfidenceLimit,
 #QuestionValueHighConfidenceLimit = HighConfidenceLimit,
 # Gender,
 Sex = Stratification1,
 Deaths,
 Population,
 AgeAdjustedDeathRate = AgeAdjustedRate,
 AgeAdjustedDeathRateStandardError = AgeAdjustedRateStandardError,
 PercentageOfTotalDeaths
)

```

```

tmp_gender <- dplyr::left_join(
 tmp_gender,

```

```

tmpUnderlyingCauseOfDeathAlcohol,
by = dplyr::join_by(
 StateCode,
 Stratification1 == Gender
)
) |>
dplyr::select(
 State,
 # StateCode,
 StateName,
 # YearStart,
 # LocationDesc,
 # QuestionID,
 Question,
 QuestionValue = DataValueAlt,
 #QuestionValueLowConfidenceLimit = LowConfidenceLimit,
 #QuestionValueHighConfidenceLimit = HighConfidenceLimit,
 # Gender,
 Sex = Stratification1,
 Deaths,
 Population,
 AgeAdjustedDeathRate = AgeAdjustedRate,
 AgeAdjustedDeathRateStandardError = AgeAdjustedRateStandardError,
 PercentageOfTotalDeaths
)

```

```

tmp_overall <- dplyr::left_join(
 tmp_overall,

```

```

tmpUnderlyingCauseOfDeathAlcohol,
by = dplyr::join_by(
 StateCode
)
) |>
dplyr::select(
 State,
 # StateCode,
 StateName,
 # YearStart,
 # LocationDesc,
 # QuestionID,
 Question,
 QuestionValue = DataValueAlt,
 #QuestionValueLowConfidenceLimit = LowConfidenceLimit,
 #QuestionValueHighConfidenceLimit = HighConfidenceLimit,
 Gender,
 #Sex = Stratification1,
 Deaths,
 Population,
 AgeAdjustedDeathRate = AgeAdjustedRate,
 AgeAdjustedDeathRateStandardError = AgeAdjustedRateStandardError,
 PercentageOfTotalDeaths
)
data <- tmp |>
dplyr::select(
 State,
 Question,

```

```
 QuestionValue,
 Sex,
 Deaths,
 Population,
 AgeAdjustedDeathRate,
 PercentageOfTotalDeaths
) |>
tidyR::pivot_wider(names_from = Question, values_from = QuestionValue) |>
dplyr::select(-`NA`)
```

```
data_overall <- tmp_overall |>
dplyr::select(
 State,
 Question,
 QuestionValue,
 # Sex,
 Deaths,
 Population,
 AgeAdjustedDeathRate,
 PercentageOfTotalDeaths
) |>
tidyR::pivot_wider(names_from = Question, values_from = QuestionValue)
```

```
data_gender <- tmp_gender |>
dplyr::select(
 State,
 Question,
```

```

 QuestionValue,
 Sex,
 Deaths,
 Population,
 AgeAdjustedDeathRate,
 PercentageOfTotalDeaths
) |>
tidyrr::pivot_wider(names_from = Question, values_from = QuestionValue)

data$State <- factor(
 data$State,
 levels = levels(as.factor(data$State))[order(data$AgeAdjustedDeathRate)])

data_overall$State <- factor(
 data_overall$State,
 levels =
 levels(as.factor(data_overall$State))[order(data_overall$AgeAdjustedDeathRate)])

data_gender$State <- factor(
 data_gender$State,
 levels =
 levels(as.factor(data_gender$State))[order(data_gender$AgeAdjustedDeathRate)])

Limpieza del dataset combinado

data$Deaths <- as.numeric(data$Deaths)
data$Population <- as.numeric(data$Population)
data_overall$Deaths <- as.numeric(data_overall$Deaths)
data_overall$Population <- as.numeric(data_overall$Population)

```

```
data_gender$Deaths <- as.numeric(data_gender$Deaths)

data_gender$Population <- as.numeric(data_gender$Population)

Renombrado

names(data_gender) <- c(
 'State',
 'Sex',
 'Deaths',
 'Population',
 'AgeAdjustedDeathRate',
 'PercentageOfTotalDeaths',
 'HeavyDrinkingAdults',
 'BingeDrinkingFrequencyAdults',
 'BingeDrinkingIntensityAdults',
 'BingeDrinkingPrevalenceAdults'
)
```

```
names(data) <- c(
 'State',
 'Sex',
 'Deaths',
 'Population',
 'AgeAdjustedDeathRate',
 'PercentageOfTotalDeaths',
 'HeavyDrinkingAdults',
 'BingeDrinkingFrequencyAdults',
 'BingeDrinkingIntensityAdults',
```

```
'BingeDrinkingPrevalenceAdults'
```

```
)
```

```
names(data_overall) <- c(
 'State',
 #'Sex',
 'Deaths',
 'Population',
 'AgeAdjustedDeathRate',
 'PercentageOfTotalDeaths',
 'HeavyDrinkingAdults',
 'BingeDrinkingFrequencyAdults',
 'BingeDrinkingIntensityAdults',
 'BingeDrinkingPrevalenceAdults')
```

```
)
```

```
```
```

```
### Salidas del subprocesso
```

| Objeto | Descripción del *data.frame* |
|--|------------------------------|
| Filas | Columnas |
| ----- ----- | |
| `data` Indicadores de enfermedades crónicas (CDI) del área de interés 'Alcohol' y de mortalidad por alcohol, año 2021, por estado de EEUU y sexo `r nrow(data)` `r ncol(data)` | |
| `data_overall` Indicadores de enfermedades crónicas (CDI) del área de interés 'Alcohol' y de mortalidad por alcohol, año 2021, por estado de EEUU y sexo `r nrow(data_overall)` `r ncol(data_overall)` | |

```

| `data_gender` | Indicadores de enfermedades crónicas (CDI) del área de interés
'Alcohol' y de mortalidad por alcohol, año 2021, por estado de EEUU y sexo | `r
nrow(data_gender)` | `r ncol(data_gender)` |

````{r}
#| code-fold: true
#| output: false

saveRDS(
 data,
 file = here::here('data', 'lab', 'data.rds')
)

saveRDS(
 data_gender,
 file = here::here('data', 'lab', 'data_gender.rds')
)

saveRDS(
 data_overall,
 file = here::here('data', 'lab', 'data_overall.rds')
)

````

````{r}
summary(data)
summary(data_gender)
summary(data_overall)
````
```

```
### Limpieza de las variables intermedias del subprocesso
```

```
```{r Limpieza - limpieza}

#| code-fold: true
#| output: false

rm(list = c(
 'esDatosCauseOfDeathConEstadoMaestra',
 'esDatosCdiConEstadoMaestra',
 'esEstadoMaestraConDatosCauseOfDeath',
 'esEstadoMaestraConDatosCdi',
 'esVarCardinalidadBaja',
 'esVarCardinalidadNormal',
 'idxVarNoData',
 'listaEstadosCauseOfDeath',
 'listaEstadosCdiAlcohol',
 'listaEstadosDatosCdiSinEstadoMaestra',
 'listaEstadosMaestra',
 'listaEstadosMaestraSinDatosCauseOfDeath',
 'listaEstadosMaestraSinDatosCdi',
 'listaNombresMaestra',
 'namesVarNoData',
 'namesVarData',
 'namesVarCardinalidadNormal',
 'nMissingDataByCol',
 'rawCdiAlcohol',
 'rawFipsCodes',
 'rawUnderlyingCauseOfDeathAlcohol',
```

```

'tmp',
'tmp_gender',
'tmp_overall',
'tmpCdiAlcohol_byGender',
'tmpCdiAlcohol_overall',
'cardinality',
'cardinalityCdiAlcohol',
'esIrrelevanteCdiAlcohol',
'isCharacter',
'listaEstadosDatosCauseOfDeathSinEstadosMaestra',
'rawNamesFipsCodes',
'rawNamesUnderlyingCauseOfDeath',
'varCdiAlcohol',
'varCharacterNames',
'tmpCdiAlcohol',
'tmpFipsCodes',
'tmpUnderlyingCauseOfDeathAlcohol'

)))
gc()
```

```

Análisis exploratorio de datos

03a - Visión general del *data.frame*: `summarytools::dfSummary()`

Se realizó un resumen de cada *data.frame* con los siguientes elementos:

- Nombre de variables y tipos,

- Etiquetas (si existían)
- Niveles de los factores,
- Frecuencias y / o estadísticos de resumen numéricos,
- Gráficos de barras / histogramas, y
- Conteos y proporciones de observaciones válidas / faltantes.

```
:::{.callout-note title="1- Objeto `data`" collapse="true"}
```

```
#### Objeto `data`
```

```
```{r}
```

```
#| code-fold: true
```

```
data |>
```

```
summarytools::dfSummary()
```

```
```
```

```
:::
```

```
:::{.callout-note title="2- Objeto `data_gender`" collapse="true"}
```

```
#### Objeto `data_gender`
```

```
```{r}
```

```
#| code-fold: true
```

```
data_gender |>
```

```
summarytools::dfSummary()
```

```
```
```

```
:::
```

```
::: {.callout-note title="3- Objeto `data_overall` " collapse="true"}
#### Objeto `data_overall`

``{r}
#| code-fold: true

data_overall |>
  summarytools::dfSummary()
```
:::
```

### 03b - Explorar variables categóricas: `SmartEDA::ExpCatViz()`

En el análisis se revisó para cada variable categórica:

- Frecuencia relativa de cada nivel de la variable respecto al total de observaciones
- La frecuencia relativa de datos faltantes, si existían

```
::: {.callout-note title="1- Objeto `data` " collapse="true"}
```

```
Objeto `data`

``{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```

tmp <- data[,c("State", "Sex")] |>

```

SmartEDA::ExpCatViz(
  clim = 60
)
ggplot2::ggsave(
  plot = tmp[[1]],
  here::here('notebooks', 'images', 'eda_data_03b_01_sex.jpg')
)
ggplot2::ggsave(
  plot = tmp[[2]],
  here::here('notebooks', 'images', 'eda_data_03b_02_state.jpg')
)
rm(tmp)
```

```

Existen 2 variables categóricas: `Sex` y `State` , con la siguiente distribución de niveles entre las observaciones:

```

```

```

```

```
:::
```

```
::: {.callout-note title="2- Objeto `data_gender` " collapse="true"}
```

```
Objeto `data_gender`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

tmp <- data_gender[,c("State", "Sex")] |>
  SmartEDA::ExpCatViz(
    clim = 60
  )

ggplot2::ggsave(
  plot = tmp[[1]],
  here::here('notebooks', 'images', 'eda_data_gender_03b_01_sex.jpg')
)
ggplot2::ggsave(
  plot = tmp[[2]],
  here::here('notebooks', 'images', 'eda_data_gender_03b_02_state.jpg')
)
rm(tmp)
```
```



```
```
::: {.callout-note title="3- Objeto `data_overall`" collapse="true"}
#### Objeto `data_overall`


```
```

```

```

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

tmp <- data_overall[,c("State")] |>
  SmartEDA::ExpCatViz(
    clim = 60
  )

ggplot2::ggsave(
  plot = tmp[[1]],
  here::here('notebooks', 'images', 'eda_data_overall_03b_01_state.jpg')
)
rm(tmp)
```
```


:::

#### 03c - Explorar variables numéricas (*Estadística descriptiva*):

`SmartEDA::ExpNumStat()`

```

Para las variables numéricas de cada *data.frame* se exploraron los siguientes aspectos:

- Número de variables (columnas) y de observaciones (filas)
- Número y porcentaje de valores faltantes (`NA`)

- Valoración de la escala de magnitud de cada variable, y comparación relativa entre las variables
- Evaluación del sesgo y la kurtosis
- Análisis básico de *outliers*

```
:::{.callout-note title="1- Objeto `data`" collapse="true"}
```

```
#### Objeto `data`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
smartEda_data <- SmartEDA::ExpNumStat(
```

```
 data,
```

```
 # Qnt = c(.25, .5, 0.75),
```

```
 Outlier = TRUE
```

```
)
```

```
smartEda_data
```

```
tmp <- smartEda_data |>
```

```
 flextable::flextable()
```

```
tmp <- flextable::autofit(tmp)
```

```
 flextable::save_as_image(
```

```
 x = tmp,
```

```
 path = here::here("notebooks", "images", "eda_data_03c.png")
```

```
)
```

```
rm(tmp)
```

```
```
```

Respecto a las variables numéricas del *data.frame*:

- Existen `r nrow(smartEda_data)` variables numéricas, con `r max(smartEda_data\$TN)` observaciones en total
- Existen valores faltantes `NA` en todas las variables, con un rango de valores faltantes entre el `r min(smartEda_data\$Per_of_Missing)` % y el `r max(smartEda_data\$Per_of_Missing)` %
- Las variables tienen una escala de magnitud muy diferente entre sí
- La mitad de las variables están más o menos centradas, y la otra mitad sesgadas. Sólo una de las variables tiene una kurtosis similar a la de la distribución normal. Cuatro de las variables son más achacadas, y tres más picudas.
- Algunas variables tienen bastantes *outliers*. Las más afectadas son `AgeAdjustedDeathRate`, `Deaths`, `PercentageOfTotalDeaths` y `Population` .

```
:::
```

```
::: {.callout-note title="2- Objeto `data_gender` " collapse="true"}
```

```
#### Objeto `data_gender`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
smartEda_data_gender <- SmartEDA::ExpNumStat(
```

```

data_gender,
Qnt = c(.25, .5, 0.75),
Outlier = TRUE
)
smartEda_data_gender

tmp <- smartEda_data_gender |>
flextable::flextable()

tmp <- flextable::autofit(tmp)
flextable::save_as_image(
x = tmp,
path = here::here("notebooks", "images", "eda_data_gender_03c.png")
)

rm(tmp)
```

```

Respecto a las variables numéricas del *data.frame*:

- Existen `r nrow(smartEda_data_gender)` variables numéricas, con `r max(smartEda_data_gender\$TN)` observaciones en total
- Existen valores faltantes `NA` en cinco variables, con un rango de valores faltantes entre el `r min(smartEda_data_gender\$Per_of_Missing)` % y el `r max(smartEda_data_gender\$Per_of_Missing)` %
- Las variables tienen una escala de magnitud muy diferente entre sí
- La mitad de las variables están más o menos centradas, y la otra mitad sesgadas. Sólo una de las variables tiene una kurtosis similar a la de la distribución normal. Cuatro de las variables son más achataadas, y tres más picudas.

- Las variables con mayor número de *outliers* son `AgeAdjustedDeathRate` , `Deaths` , `PercentageOfTotalDeaths` y `Population` .

:::

```
:::{.callout-note title="3- Objeto `data_overall` " collapse="true"}
```

```
#### Objeto `data_overall`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
smartEda_data_overall <- SmartEDA::ExpNumStat(
```

```
 data_overall,
```

```
 # Qnt = c(.25, .5, 0.75),
```

```
 Outlier = TRUE
```

```
)
```

```
smartEda_data_overall
```

```
tmp <- smartEda_data_overall |>
```

```
 flextable::flextable()
```

```
tmp <- flextable::autofit(tmp)
```

```
 flextable::save_as_image(
```

```
 x = tmp,
```

```
 path = here::here("notebooks", "images", "eda_data_overall_03c.png")
```

```
)
```

```
rm(tmp)
```

```
...
```

Respecto a las variables numéricas del \*data.frame\*:

- Existen `r nrow(smartEda\_data\_overall)` variables numéricas, con `r max(smartEda\_data\_overall\$TN)` observaciones en total
- Existen valores faltantes `NA` en cuatro variables, con un rango de valores faltantes entre el `r min(smartEda\_data\_overall\$Per\_of\_Missing)` % y el `r max(smartEda\_data\_overall\$Per\_of\_Missing)` %
- Las variables tienen una escala de magnitud muy diferente entre sí
- La mitad de las variables están más o menos centradas, y la otra mitad sesgadas. Sólo una de las variables tiene una kurtosis similar a la de la distribución normal. Cuatro de las variables son más achataadas, y tres más picudas.
- Las variables con mayor número de \*outliers\* son `AgeAdjustedDeathRate` , `Deaths` , `PercentageOfTotalDeaths` y `Population` .

```
:::
```

```
03d - Explorar distribuciones (*skewness and kurtosis tests*)
```

Se evaluó la distribución de probabilidad de las variables numéricas. Para ello, se utilizaron las siguientes técnicas:

- 03da - Visualización de las distribuciones (histograma y función de densidad)
- 03db - Test de hipótesis de la centralidad (kurtosis y sesgo)

Cuando ambos test son no significativos, la variable puede seguir una distribución aproximadamente normal.

```
::: {.callout-note title="1- Objeto `data`" collapse="true"}
```

```
Objeto `data`
```

```
::: {.callout-caution title="03da - Visualización de las distribuciones" collapse="true"}
```

Se utilizaron dos tipos de gráfico para evaluar la distribución de las variables aleatorias numéricas:

- Histogramas
- Funciones de densidad

```
03da - Visualización de las distribuciones (histograma y función de densidad)
```

```
Gráfico - histogramas
```

La función `DataExplorer::plot\_histogram()` crea histogramas para las variables de clase `numeric` o `integer` de un dataset.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

data |>
  DataExplorer::plot_histogram(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
```

```

ncol = 2L
)
ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03da_histograma.jpg')
)
```


```

## ##### Gráfico - Función de densidad de probabilidad

La función `DataExplorer::plot\_density()` dibuja la función de densidad de probabilidad para las variables de clase `numeric` o `integer` de un dataset.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```

data |>
  DataExplorer::plot_density(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
)
ggplot2::ggsave(

```

```
here::here('notebooks', 'images', 'eda_data_03db_densidad.jpg')
```

```
)
```

```
...
```

```

```

```
:::
```

```
::: {.callout-caution title="03db - Test de hipótesis de la centralidad - Kurtosis y Sesgo (*skewness*)" collapse="true"}
```

```
##### 03db - Test de hipótesis de la centralidad (kurtosis y sesgo)
```

Los test de hipótesis de centralidad (Kurtosis y sesgo) orientan para saber qué variables podrían seguir una distribución normal. Se utilizan dos test:

- Test de hipótesis de la kurtosis (Anscombe-Glynn)
- Test de hipótesis del sesgo (D'Agostino)

La hipótesis nula de los dos test asumen que la variable numérica evaluada sigue una distribución con un sesgo (o una kurtosis) similar a la de una distribución normal.

```
::: {.callout-tip title="Test de hipótesis de la kurtosis (Anscombe-Glynn)" collapse="true" icon="false"}
```

```
##### Test de hipótesis de la kurtosis - Anscombe-Glynn
```

Contrasta dos hipótesis:

- Hipótesis nula - La distribución tiene una kurtosis igual a 3 (como una normal)
- Hipótesis alternativa - La distribución no tiene una kurtosis igual a 3

Resultados:

- Para las variables `Deaths`, `Population`, `AgeAdjustedDeathRate`, `PercentageOfTotalDeaths` y `BingeDrinkingIntensityAdults` el p -valor del test de Anscombe-Glynn es < 0.05 , por lo que rechazamos la hipótesis nula y aceptamos la alternativa, asumiendo que la distribución tiene una curtosis diferente de 3 (distribución normal), por lo que no están distribuidas normalmente.
- Para las variables `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults` el p -valor del test de Anscombe-Glynn es > 0.05 , por lo que no podemos rechazar la hipótesis nula de que la variable tiene curtosis igual a 3 (distribución normal), por lo que pueden estar distribuidas normalmente.

```
```{r}
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
Variable Deaths
```

```
data$Deaths |>
```

```
moments::anscombe.test()
```

```
Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
Variable Population
```

```
data$Population |>
```

```
moments::anscombe.test()
```

```
Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
Variable AgeAdjustedDeathRate
```

```
data$AgeAdjustedDeathRate |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable PercentageOfTotalDeaths
data$PercentageOfTotalDeaths |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable HeavyDrinkingAdults
data$HeavyDrinkingAdults |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable BingeDrinkingFrecuencyAdults
data$BingeDrinkingFrecuencyAdults |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable BingeDrinkingIntensityAdults
data$BingeDrinkingIntensityAdults |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable BingeDrinkingPrevalenceAdults
data$BingeDrinkingPrevalenceAdults |>
moments::anscombe.test()

````
```

:::

```
:::{.callout-tip title="Test de hipótesis de sesgo (D'Agostino)" collapse="true" icon="false"}
```

Test de hipótesis de sesgo - D'Agostino

Contrasta dos hipótesis:

- Hipótesis nula - La distribución no está sesgada (como una distribución normal)
- Hipótesis alternativa - La distribución está sesgada

Resultados:

- Para las variables `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths` el p -valor del test D'Agostino es < 0.05 , por lo que rechazamos la hipótesis nula y aceptamos la alternativa, asumiendo que la distribución está sesgada significativamente, y no está distribuida normalmente; es posible que existan outliers.
- Para `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults`, `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults` el p -valor del test D'Agostino es > 0.05 , por lo que no podemos rechazar la hipótesis nula y afirmar que exista sesgo.

```
```{r}
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
Test de hipótesis de sesgo - D'Agostino
```

```
Variable Deaths
```

```
data$Deaths |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable Population

data$Population |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable AgeAdjustedDeathRate

data$AgeAdjustedDeathRate |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable PercentageOfTotalDeaths

data$PercentageOfTotalDeaths |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable HeavyDrinkingAdults

data$HeavyDrinkingAdults |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable BingeDrinkingFrecuencyAdults

data$BingeDrinkingFrecuencyAdults |>
moments::agostino.test()
```

```

Test de hipótesis de sesgo - D'Agostino
Variable BingeDrinkingIntensityAdults
data$BingeDrinkingIntensityAdults |>
moments::agostino.test()

```

```

Test de hipótesis de sesgo - D'Agostino
Variable BingeDrinkingPrevalenceAdults
data$BingeDrinkingPrevalenceAdults |>
moments::agostino.test()
```
```

```

Los resultados obtenidos al aplicar los test a las variables numéricas fueron los siguientes:

Variable	Test kurtosis Anscombe-Glynn	Test sesgo D'Agostino
`Deaths`	\$< 0.05\$	\$< 0.05\$
`Population`	\$< 0.05\$	\$< 0.05\$
`AgeAdjustedDeathRate`	\$< 0.05\$	\$< 0.05\$
`PercentageOfTotalDeaths`	\$< 0.05\$	\$< 0.05\$
`HeavyDrinkingAdults`	No significativo	No significativo
`BingeDrinkingFrecuencyAdults`	No significativo	No significativo
`BingeDrinkingIntensityAdults`	\$< 0.05\$	No significativo
`BingeDrinkingPrevalenceAdults`	No significativo	No significativo
...		
...		

```
::: {.callout-note title="2- Objeto `data_gender` " collapse="true"}
```

```
Objeto `data_gender`
```

```
::: {.callout-caution title="03da - Visualización de las distribuciones" collapse="true"}
```

Se utilizaron dos tipos de gráfico para evaluar la distribución de las variables aleatorias numéricas de nuestra muestra:

- Histogramas
- Funciones de densidad

```
03da - Visualización de las distribuciones (histograma y función de densidad)
```

```
Gráfico - histogramas
```

La función `DataExplorer::plot\_histogram()` crea histogramas para las variables de clase `numeric` o `integer` de un data\_overallset.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false
```

```
data_gender |>
```

```
  DataExplorer::plot_histogram(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
```

```

ncol = 2L
)
ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_gender_03da_histograma.jpg')
)
```


```

## ##### Gráfico - Función de densidad de probabilidad

La función `data\_genderExplorer::plot\_density()` dibuja la función de densidad de probabilidad para las variables de clase `numeric` o `integer` de un data\_genderset.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```

data_gender |>
  DataExplorer::plot_density(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
)
ggplot2::ggsave(

```

```
here::here('notebooks', 'images', 'eda_data_gender_03db_densidad.jpg')
```

```
)
```

```
...
```

```

```

```
:::
```

```
::: {.callout-caution title="03db - Test de hipótesis de la centralidad - Kurtosis y Sesgo (*skewness*)" collapse="true"}
```

```
##### 03db - Test de hipótesis de la centralidad (kurtosis y sesgo)
```

Los test de hipótesis de centralidad (Kurtosis y sesgo) orientan para saber qué variables podrían seguir una distribución normal. Se utilizan dos test:

- Test de hipótesis de la kurtosis (Anscombe-Glynn)
- Test de hipótesis del sesgo (D'Agostino)

La hipótesis nula de los dos test asumen que la variable numérica evaluada sigue una distribución con un sesgo (o una kurtosis) similar a la de una distribución normal.

```
::: {.callout-tip title="Test de hipótesis de la kurtosis (Anscombe-Glynn)" collapse="true" icon="false"}
```

```
##### Test de hipótesis de la kurtosis - Anscombe-Glynn
```

Contrasta dos hipótesis:

- Hipótesis nula - La distribución tiene una kurtosis igual a 3 (como una normal)
- Hipótesis alternativa - La distribución no tiene una kurtosis igual a 3

Resultados:

- Para las variables `Deaths`, `Population`, `AgeAdjustedDeathRate`, `PercentageOfTotalDeaths` y `BingeDrinkingIntensityAdults` el p -valor del test de Anscombe-Glynn es < 0.05 , por lo que rechazamos la hipótesis nula y aceptamos la alternativa, asumiendo que la distribución tiene una curtosis diferente de 3 (distribución normal), por lo que no están distribuidas normalmente.
- Para las variables `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults` el p -valor del test de Anscombe-Glynn es > 0.05 , por lo que no podemos rechazar la hipótesis nula de que la variable tiene curtosis igual a 3 (distribución normal), por lo que pueden estar distribuidas normalmente.

```
```{r}
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
Variable Deaths
```

```
data_gender$Deaths |>
```

```
moments::anscombe.test()
```

```
Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
Variable Population
```

```
data_gender$Population |>
```

```
moments::anscombe.test()
```

```
Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
Variable AgeAdjustedDeathRate
```

```
data_gender$AgeAdjustedDeathRate |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable PercentageOfTotalDeaths
data_gender$PercentageOfTotalDeaths |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable HeavyDrinkingAdults
data_gender$HeavyDrinkingAdults |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable BingeDrinkingFrecuencyAdults
data_gender$BingeDrinkingFrecuencyAdults |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable BingeDrinkingIntensityAdults
data_gender$BingeDrinkingIntensityAdults |>
moments::anscombe.test()

Test de hipótesis de la kurtosis (Anscombe-Glynn)

Variable BingeDrinkingPrevalenceAdults
data_gender$BingeDrinkingPrevalenceAdults |>
moments::anscombe.test()

````
```

:::

```
:::{.callout-tip title="Test de hipótesis de sesgo (D'Agostino)" collapse="true" icon="false"}
```

Test de hipótesis de sesgo - D'Agostino

Contrasta dos hipótesis:

- Hipótesis nula - La distribución no está sesgada (como una distribución normal)
- Hipótesis alternativa - La distribución está sesgada

Resultados:

- Para las variables `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths` el p -valor del test D'Agostino es < 0.05 , por lo que rechazamos la hipótesis nula y aceptamos la alternativa, asumiendo que la distribución está sesgada significativamente, y no está distribuida normalmente; es posible que existan outliers.
- Para `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults`, `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults` el p -valor del test D'Agostino es > 0.05 , por lo que no podemos rechazar la hipótesis nula y afirmar que exista sesgo.

```
```{r}
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
Test de hipótesis de sesgo - D'Agostino
```

```
Variable Deaths
```

```
data_gender$Deaths |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable Population

data_gender$Population |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable AgeAdjustedDeathRate

data_gender$AgeAdjustedDeathRate |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable PercentageOfTotalDeaths

data_gender$PercentageOfTotalDeaths |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable HeavyDrinkingAdults

data_gender$HeavyDrinkingAdults |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino

Variable BingeDrinkingFrecuencyAdults

data_gender$BingeDrinkingFrecuencyAdults |>
moments::agostino.test()
```

```

Test de hipótesis de sesgo - D'Agostino
Variable BingeDrinkingIntensityAdults
data_gender$BingeDrinkingIntensityAdults |>
moments::agostino.test()

Test de hipótesis de sesgo - D'Agostino
Variable BingeDrinkingPrevalenceAdults
data_gender$BingeDrinkingPrevalenceAdults |>
moments::agostino.test()

```
```
```
```

Los resultados obtenidos al aplicar los test a las variables numéricas fueron los siguientes:

| Variable | Test kurtosis
Anscombe-Glynn | Test sesgo
D'Agostino |
|---------------------------------|---------------------------------|--------------------------|
| `Deaths` | \$< 0.05\$ | \$< 0.05\$ |
| `Population` | \$< 0.05\$ | \$< 0.05\$ |
| `AgeAdjustedDeathRate` | \$< 0.05\$ | \$< 0.05\$ |
| `PercentageOfTotalDeaths` | \$< 0.05\$ | \$< 0.05\$ |
| `HeavyDrinkingAdults` | \$< 0.05\$ | No significativo |
| `BingeDrinkingFrecuencyAdults` | No significativo | No significativo |
| `BingeDrinkingIntensityAdults` | \$< 0.05\$ | No significativo |
| `BingeDrinkingPrevalenceAdults` | \$< 0.05\$ | No significativo |
| ... | | |
| ... | | |

```
::: {.callout-note title="3- Objeto `data_overall` " collapse="true"}
```

```
#### Objeto `data_overall`
```

```
::: {.callout-caution title="03da - Visualización de las distribuciones" collapse="true"}
```

Se utilizaron dos tipos de gráfico para evaluar la distribución de las variables aleatorias numéricas de nuestra muestra:

- Histogramas
- Funciones de densidad

```
##### 03da - Visualización de las distribuciones (histograma y función de densidad)
```

```
##### Gráfico - histogramas
```

La función `DataExplorer::plot_histogram()` crea histogramas para las variables de clase `numeric` o `integer` de un dataset.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false
```

```
data_overall |>
```

```
 DataExplorer::plot_histogram(
 ggtheme = ggplot2::theme_bw(),
 nrow = 4L,
```

```

ncol = 2L
)
ggplot2::ggsave(
 here::here('notebooks', 'images', 'eda_data_overall_03da_histograma.jpg')
)
```


```

Gráfico - Función de densidad de probabilidad

La función `data_genderExplorer::plot_density()` dibuja la función de densidad de probabilidad para las variables de clase `numeric` o `integer` de un dataset.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```

data_overall |>
 DataExplorer::plot_density(
 ggtheme = ggplot2::theme_bw(),
 nrow = 4L,
 ncol = 2L
)
ggplot2::ggsave(

```

```
here::here('notebooks', 'images', 'eda_data_overall_03db_densidad.jpg')
```

```
)
```

```
...
```

```

```

```
:::
```

```
::: {.callout-caution title="03db - Test de hipótesis de la centralidad - Kurtosis y Sesgo (*skewness*)" collapse="true"}
```

```
03db - Test de hipótesis de la centralidad (kurtosis y sesgo)
```

Los test de hipótesis de centralidad (Kurtosis y sesgo) orientan para saber qué variables podrían seguir una distribución normal. Se utilizan dos test:

- Test de hipótesis de la kurtosis (Anscombe-Glynn)
- Test de hipótesis del sesgo (D'Agostino)

La hipótesis nula de los dos test asumen que la variable numérica evaluada sigue una distribución con un sesgo (o una kurtosis) similar a la de una distribución normal.

```
::: {.callout-tip title="Test de hipótesis de la kurtosis (Anscombe-Glynn)" collapse="true" icon="false"}
```

```
Test de hipótesis de la kurtosis - Anscombe-Glynn
```

Contrasta dos hipótesis:

- Hipótesis nula - La distribución tiene una kurtosis igual a 3 (como una normal)
- Hipótesis alternativa - La distribución no tiene una kurtosis igual a 3

## Resultados:

- Para las variables `Deaths`, `Population`, `AgeAdjustedDeathRate`, `PercentageOfTotalDeaths` y `BingeDrinkingIntensityAdults` el  $p$ -valor del test de Anscombe-Glynn es  $< 0.05$ , por lo que rechazamos la hipótesis nula y aceptamos la alternativa, asumiendo que la distribución tiene una curtosis diferente de 3 (distribución normal), por lo que no están distribuidas normalmente.
- Para las variables `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults` el  $p$ -valor del test de Anscombe-Glynn es  $> 0.05$ , por lo que no podemos rechazar la hipótesis nula de que la variable tiene curtosis igual a 3 (distribución normal), por lo que pueden estar distribuidas normalmente.

```
```{r}
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
### Variable Deaths
```

```
data_overall$Deaths |>
```

```
moments::anscombe.test()
```

```
## Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
### Variable Population
```

```
data_overall$Population |>
```

```
moments::anscombe.test()
```

```
## Test de hipótesis de la kurtosis (Anscombe-Glynn)
```

```
### Variable AgeAdjustedDeathRate
```

```

data_overall$AgeAdjustedDeathRate |>
moments::anscombe.test()

## Test de hipótesis de la kurtosis (Anscombe-Glynn)

### Variable PercentageOfTotalDeaths
data_overall$PercentageOfTotalDeaths |>
moments::anscombe.test()

## Test de hipótesis de la kurtosis (Anscombe-Glynn)

### Variable HeavyDrinkingAdults
data_overall$HeavyDrinkingAdults |>
moments::anscombe.test()

## Test de hipótesis de la kurtosis (Anscombe-Glynn)

### Variable BingeDrinkingFrecuencyAdults
data_overall$BingeDrinkingFrecuencyAdults |>
moments::anscombe.test()

## Test de hipótesis de la kurtosis (Anscombe-Glynn)

### Variable BingeDrinkingIntensityAdults
data_overall$BingeDrinkingIntensityAdults |>
moments::anscombe.test()

## Test de hipótesis de la kurtosis (Anscombe-Glynn)

### Variable BingeDrinkingPrevalenceAdults
data_overall$BingeDrinkingPrevalenceAdults |>
moments::anscombe.test()
```

```

:::

```
:::{.callout-tip title="Test de hipótesis de sesgo (D'Agostino)" collapse="true" icon="false"}
```

## ##### Test de hipótesis de sesgo - D'Agostino

Contrasta dos hipótesis:

- Hipótesis nula - La distribución no está sesgada (como una distribución normal)
- Hipótesis alternativa - La distribución está sesgada

Resultados:

- Para las variables `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths` el  $p$ -valor del test D'Agostino es  $< 0.05$ , por lo que rechazamos la hipótesis nula y aceptamos la alternativa, asumiendo que la distribución está sesgada significativamente, y no está distribuida normalmente; es posible que existan outliers.
- Para `HeavyDrinkingAdults`, `BingeDrinkingFrecuencyAdults`, `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults` el  $p$ -valor del test D'Agostino es  $> 0.05$ , por lo que no podemos rechazar la hipótesis nula y afirmar que exista sesgo.

```
```{r}
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Test de hipótesis de sesgo - D'Agostino
```

```
### Variable Deaths
```

```
data_overall$Deaths |>
moments::agostino.test()

## Test de hipótesis de sesgo - D'Agostino

### Variable Population

data_overall$Population |>
moments::agostino.test()

## Test de hipótesis de sesgo - D'Agostino

### Variable AgeAdjustedDeathRate

data_overall$AgeAdjustedDeathRate |>
moments::agostino.test()

## Test de hipótesis de sesgo - D'Agostino

### Variable PercentageOfTotalDeaths

data_overall$PercentageOfTotalDeaths |>
moments::agostino.test()

## Test de hipótesis de sesgo - D'Agostino

### Variable HeavyDrinkingAdults

data_overall$HeavyDrinkingAdults |>
moments::agostino.test()

## Test de hipótesis de sesgo - D'Agostino

### Variable BingeDrinkingFrecuencyAdults

data_overall$BingeDrinkingFrecuencyAdults |>
moments::agostino.test()
```

```

## Test de hipótesis de sesgo - D'Agostino
### Variable BingeDrinkingIntensityAdults
data_gender$BingeDrinkingIntensityAdults |>
moments::agostino.test()

```

```

## Test de hipótesis de sesgo - D'Agostino
### Variable BingeDrinkingPrevalenceAdults
data_overall$BingeDrinkingPrevalenceAdults |>
moments::agostino.test()
```
:::
```

Los resultados obtenidos al aplicar los test a las variables numéricas fueron los siguientes:

| Variable                        | Test kurtosis<br>Anscombe-Glynn | Test sesgo<br>D'Agostino |
|---------------------------------|---------------------------------|--------------------------|
| `Deaths`                        | \$< 0.05\$                      | \$< 0.05\$               |
| `Population`                    | \$< 0.05\$                      | \$< 0.05\$               |
| `AgeAdjustedDeathRate`          | \$< 0.05\$                      | \$< 0.05\$               |
| `PercentageOfTotalDeaths`       | \$< 0.05\$                      | \$< 0.05\$               |
| `HeavyDrinkingAdults`           | No significativo                | No significativo         |
| `BingeDrinkingFrecuencyAdults`  | No significativo                | \$< 0.05\$               |
| `BingeDrinkingIntensityAdults`  | \$< 0.05\$                      | No significativo         |
| `BingeDrinkingPrevalenceAdults` | No significativo                | No significativo         |
| :::                             |                                 |                          |
| :::                             |                                 |                          |

### ### 03e - Explorar normalidad (\*QQ-plots and Shapiro-Wilk\*)

Para explorar la normalidad de las variables se utilizaron las siguientes técnicas:

| Técnica                                                                                                                                                                                                                                                | Descripción                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Evaluación gráfica: QQ-Plot   Método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación [@eswiki:130441335] |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Test de Shapiro-Wilks                                                                                                                                                                                                                                  | Se considera el test más potente para testar la normalidad, seguido de cerca por el Anderson--Darling<br>- \$H_0\$: Los datos provienen de una variable normalmente distribuida<br>- \$H_1\$: Los datos no están distribuidos según una distribución normal.<br>Si la muestra es muy grande, es posible que el test detecte desviaciones mínimas frente a la normal, que no tengan importancia práctica. Por eso el test debe interpretarse siempre conjuntamente con el gráfico QQ-plot. |

::: {.callout-note title="1- Objeto `data` " collapse="true"}

#### Objeto `data`

Hay tres variables que se ajustan aproximadamente a una distribución normal, tanto visualmente como en el test de hipótesis de Shapiro-Wilks: `HeavyDrinkingAdults`, `BingeDrinkingFrequencyAdults` y `BingeDrinkingPrevalenceAdults`. Son las mismas tres variables con los test de kurtosis y sesgo no significativos.

##### Test gráfico - QQ plot (`DataExplorer::plot\_qq()`)

Hay tres variables que, visualmente se ajustan aproximadamente a una distribución normal: `HeavyDrinkingAdults`, `BingeDrinkingFrequencyAdults` y `BingeDrinkingPrevalenceAdults`. Son las mismas tres variables con los test de kurtosis y sesgo no significativos.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

data |>
  DataExplorer::plot_qq(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
  )

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03e_qqplot_general.jpg')
)
```

```

Cuando estratificamos por la variable sexo, estas variables mantienen su tendencia a la normalidad

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

data |>
  DataExplorer::plot_qq(
    by = 'Sex',
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
  )

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03e_qqplot_bySex.jpg')
)
```


Test de hipótesis (Shapiro-Wilk)
```

Hay tres variables con un test no significativo, y, por tanto, no es posible rechazar la hipótesis nula de normalidad: `HeavyDrinkingAdults` , `BingeDrinkingFrequencyAdults` y `BingeDrinkingPrevalenceAdults` .

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

data |>
  dlookr::normality() |>
  dplyr::mutate_if(is.numeric, ~round(., digits = 3)) |>
  flextable::flextable()

```
:::

:::

:::

::: {.callout-note title="2- Objeto `data_gender` " collapse="true"}

Objeto `data_gender`
```

Hay una variable que, visualmente se ajusta aproximadamente a una distribución normal: `BingeDrinkingFrequencyAdults` . Hay muchas variables que son aproximadamente normales en el centro de la distribución, pero que no lo son en los extremos (outliers).

```
Test gráfico - QQ plot (`DataExplorer::plot_qq()`)
```

Hay una variable que, visualmente se ajustan aproximadamente a una distribución normal: `BingeDrinkingFrequencyAdults` . Hay muchas variables que son

aproximadamente normales en el centro de la distribución, pero que no lo son en los extremos (outliers).

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

data_gender |>

  DataExplorer::plot_qq(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
  )

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_gender_03e_qqplot_general.jpg')
)
```

```



Cuando estratificamos por la variable sexo, estas variables mantienen su tendencia a la normalidad, y persiste la influencia de los outliers, especialmente para los varones.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

data_gender |>

  DataExplorer::plot_qq(
    by = 'Sex',
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
  )

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_gender_03e_qqplot_bySex.jpg')
)
```


Test de hipótesis (Shapiro-Wilk)
```

Hay una variable con un test no significativo, y, por tanto, no es posible rechazar la hipótesis nula de normalidad: `BingeDrinkingFrequencyAdults`.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

data_gender |>
  dlookr::normality() |>
  dplyr::mutate_if(is.numeric, ~round(., digits = 3)) |>
  flextable::flextable()

```
:::
::: {.callout-note title="3- Objeto `data_overall` collapse="true"}
Objeto `data_overall`
```

Al considerar los datos globales, ninguna variable se ajusta ni visualmente ni en el test de hipótesis a una distribución normal.

```
Test gráfico - QQ plot (`DataExplorer::plot_qq()`)
```

Visualmente, ninguna variable se ajusta a lo esperado en una distribución normal..

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```

#| output: false

data_overall |>

  DataExplorer::plot_qq(
    ggtheme = ggplot2::theme_bw(),
    nrow = 4L,
    ncol = 2L
  )

  ggplot2::ggsave(
    here::here('notebooks', 'images', 'eda_data_overall_03e_qqplot_general.jpg')
  )
```

```



#### ##### Test de hipótesis (Shapiro-Wilk)

Ninguna variable tiene un test no significativo, y, por tanto, no es posible afirmar la hipótesis nula de normalidad.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

```

```

data_overall |>
dlookr::normality() |>
dplyr::mutate_if(is.numeric, ~round(., digits = 3)) |>
flextable::flextable()
```
:::
03f - Comparar grupos (*Boxplots, non-parametric tests*)

```

Para la comparación entre grupos, se utilizaron dos técnicas:

- La valoración gráfica, mediante boxplots
- El test de hipótesis no paramétrico

:::{.callout-note title="1- Objeto `data` " collapse="true"}

#### Objeto `data`

El objeto `data` tiene la variable `Sex`, con tres niveles que son redundantes entre sí: Hombres, Mujeres y valoración global.

En la fase exploratoria, esta variable se utilizó para comparar los indicadores entre hombres y mujeres, y entre cada uno de ellos con la media global.

:::{.callout-caution title="03fa - Valoración gráfica: `DataExplorer::plot\_boxplot()` " collapse="true"}

##### 03fa - Valoración gráfica: `DataExplorer::plot\_boxplot()`

Se observa una diferencia entre los dos sexos, y entre cada sexo con la media, para las variables, `BingeDrinkingFrequencyAdults`, `BingeDrinkingIntensityAdults` y

`` `BingeDrinkingPrevalenceAdults` . También se ha observa una diferencia entre hombres y mujeres para la variable `AgeAdjustedDeathRate` .

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

## Reordenamos los niveles de los factores para una visualización más equilibrada
data$State <- with(
  data,
  reorder(State , AgeAdjustedDeathRate, median , na.rm = T)
)

data |>
  DataExplorer::plot_boxplot(
    geom_boxplot_args = list('outlier.colour' = 'darkred'),
    ggtheme = ggplot2::theme_bw(),
    by = 'Sex',
    nrow = 4L,
    ncol = 2L
  )

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03fa_boxplot.jpg')
)
```

```



:::

::: {.callout-caution title="03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`" collapse="true"}

##### 03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`

Se evidenciaron diferencias estadísticamente significativas entre hombres y mujeres, y entre cada uno de ellos con la media general, para las variables `BingeDrinkingFrequencyAdults`, `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults`. También se ha evidenciado una diferencia significativa entre hombres y mujeres para la variable `AgeAdjustedDeathRate`.

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

#| output: false

ggstatsplot::ggbetweenstats(

data = data,

x = Sex,

y = AgeAdjustedDeathRate,

type = 'np' # Nonparametric

) +

ggplot2::ggtitle(

```
"Age Adjusted Death Rate, by sex"  
)  
ggplot2::ggsave(  
  here::here('notebooks', 'images', 'eda_data_03fb_testAgeAdjustedDeathRate.jpg')  
)  
```
```

```

```

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
#| output: false
```

```
ggstatsplot::ggbetweenstats(  
  data = data,  
  x = Sex,  
  y = BingeDrinkingFrecuencyAdults,  
  type = 'np'    # Nonparametric  
) +  
  ggplot2::ggtitle(  
  "Binge Drinking Frecuency in Adults, by sex")  
)  
ggplot2::ggsave(  
  here::here('notebooks', 'images',  
  'eda_data_03fb_testBingeDrinkingFrecuencyAdults.jpg')
```

```
)
```

```
```
```

```

```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
#| output: false
```

```
ggstatsplot::ggbetweenstats(
```

```
  data = data,
```

```
  x = Sex,
```

```
  y = BingeDrinkingIntensityAdults,
```

```
  type = 'np'    # Nonparametric
```

```
) +
```

```
  ggplot2::ggtitle(
```

```
    "Binge Drinking Intensity in Adults, by sex"
```

```
)
```

```
  ggplot2::ggsave(
```

```
    here::here('notebooks', 'images',
    'eda_data_03fb_testBingeDrinkingIntensityAdults.jpg')
```

```
)
```

```
```
```

```

```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data,
  x = Sex,
  y = BingeDrinkingPrevalenceAdults,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Binge Drinking Prevalence in Adults, by sex"
  )
  ggplot2::ggsave(
    here::here('notebooks', 'images',
    'eda_data_03fb_testBingeDrinkingPrevalenceAdults.jpg')
  )
```


```

```

```{r}
#| code-fold: true
#| info: false

```

```

#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data,
  x = Sex,
  y = Deaths,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Deaths, by sex"
  )
  ggplot2::ggsave(
    here::here('notebooks', 'images', 'eda_data_03fb_testDeaths.jpg')
  )
  ```



```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```
ggstatsplot::ggbetweenstats(
```

```

data = data,
x = Sex,
y = HeavyDrinkingAdults,
type = 'np'    # Nonparametric
) +
ggplot2::ggtitle(
  "Heavy Drinking in Adults, by sex"
)
ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03fb_testHeavyDrinkingAdults.jpg')
)
```


```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data,
  x = Sex,
  y = PercentageOfTotalDeaths,
  type = 'np'    # Nonparametric
) +

```

```

ggplot2::ggtitle(
  "Percentage of Total Deaths, by sex"
)
ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03fb_testPercentageOfTotalDeaths.jpg')
)
```


```
`{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

ggstatsplot::ggbetweenstats(
  data = data,
  x = Sex,
  y = Population,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Population, by sex"
)

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_03fb_testPopulation.jpg')
)

```

```
)
```

```
...
```

```

```

```
:::
```

```
:::
```

```
::: {.callout-note title="2- Objeto `data_gender` " collapse="true"}
```

```
#### Objeto `data_gender`
```

```
::: {.callout-caution title="03fa - Valoración gráfica: `DataExplorer::plot_boxplot()` " collapse="true"}
```

```
##### 03fa - Valoración gráfica: `DataExplorer::plot_boxplot()`
```

Se observa una diferencia entre los dos sexos para las variables, `HeavyDrinkingAdults`, `AgeAdjustedDeathRate`, `BingeDrinkingPrevalenceAdults` y `BingeDrinkingIntensityAdults`.

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
#| output: false
```

```
Reordenamos los niveles de los factores para una visualización más equilibrada
```

```
data_gender$State <- with(
```

```
 data_gender,
```

```

reorder(State , AgeAdjustedDeathRate, median , na.rm = T)
)

data_gender |>

 DataExplorer::plot_boxplot(
 geom_boxplot_args = list('outlier.colour' = 'darkred'),
 ggtheme = ggplot2::theme_bw(),
 by = 'Sex',
 nrow = 4L,
 ncol = 2L
)

ggplot2::ggsave(
 here::here('notebooks', 'images', 'eda_data_gender_03fa_boxplot.jpg')
)
```
```

```
```
::: {.callout-caution title="03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()` " collapse="true"}
03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`

```

Se evidenciaron diferencias estadísticamente significativas entre hombres y mujeres, y entre cada uno de ellos con la media general, para las variables `BingeDrinkingFrequencyAdults`, `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults`. También se ha evidenciado una diferencia significativa entre hombres y mujeres para la variable `AgeAdjustedDeathRate`.

```

```{r}

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data_gender,
  x = Sex,
  y = AgeAdjustedDeathRate,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Age Adjusted Death Rate, by sex"
  )
  ggplot2::ggsave(
    here::here('notebooks', 'images',
    'eda_data_gender_03fb_testAgeAdjustedDeathRate.jpg')
  )
```



```{r}

#| code-fold: true
#| info: false

```

```

#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data_gender,
  x = Sex,
  y = BingeDrinkingFrecuencyAdults,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Binge Drinking Frecuency in Adults, by sex"
  )
  ggplot2::ggsave(
    here::here('notebooks', 'images',
    'eda_data_gender_03fb_testBingeDrinkingFrecuencyAdults.jpg')
  )
  ...

```



```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```

ggstatsplot::ggbetweenstats(
 data = data_gender,
 x = Sex,
 y = BingeDrinkingIntensityAdults,
 type = 'np' # Nonparametric
) +
 ggplot2::ggtitle(
 "Binge Drinking Intensity in Adults, by sex"
)
 ggplot2::ggsave(
 here::here('notebooks', 'images',
 'eda_data_gender_03fb_testBingeDrinkingIntensityAdults.jpg')
)
```
```

```



```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```

ggstatsplot::ggbetweenstats(
  data = data_gender,
  x = Sex,
  y = BingeDrinkingPrevalenceAdults,

```

```

type = 'np'    # Nonparametric
) +
ggplot2::ggtitle(
  "Binge Drinking Prevalence in Adults, by sex"
)
ggplot2::ggsave(
  here::here('notebooks', 'images',
  'eda_data_gender_03fb_testBingeDrinkingPrevalenceAdults.jpg')
)
```

```

``

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

```

```

ggstatsplot::ggbetweenstats(
  data = data_gender,
  x = Sex,
  y = Deaths,
  type = 'np'    # Nonparametric
) +
ggplot2::ggtitle(
  "Deaths, by sex"
)

```

```

)
ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_gender_03fb_testDeaths.jpg')
)
```


```
{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data_gender,
  x = Sex,
  y = HeavyDrinkingAdults,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Heavy Drinking in Adults, by sex"
  )
ggplot2::ggsave(
  here::here('notebooks', 'images',
  'eda_data_gender_03fb_testHeavyDrinkingAdults.jpg')
)

```

```

```

```

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
#| output: false
```

```
ggstatsplot::ggbetweenstats(
```

```
  data = data_gender,
```

```
  x = Sex,
```

```
  y = PercentageOfTotalDeaths,
```

```
  type = 'np'    # Nonparametric
```

```
) +
```

```
  ggplot2::ggtitle(
```

```
    "Percentage of Total Deaths, by sex"
```

```
)
```

```
  ggplot2::ggsave(
```

```
    here::here('notebooks', 'images',
```

```
    'eda_data_gender_03fb_testPercentageOfTotalDeaths.jpg')
```

```
)
```

```

```

```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| output: false

ggstatsplot::ggbetweenstats(
  data = data_gender,
  x = Sex,
  y = Population,
  type = 'np'    # Nonparametric
) +
  ggplot2::ggtitle(
    "Population, by sex"
  )

ggplot2::ggsave(
  here::here('notebooks', 'images', 'eda_data_gender_03fb_testPopulation.jpg')
)
```

:::
:::

::: {.callout-note title="3- Objeto `data_overall` " collapse="true"}

```

```
Objeto `data_overall`

:::{.callout-caution title="03fa - Valoración gráfica: `DataExplorer::plot_boxplot()`" collapse="true"}
03fa - Valoración gráfica: `DataExplorer::plot_boxplot()`
```

Se observan diferencias en el valor de los indicadores para todos los estados, para las distintas variables, identificándose estados con valores en torno a la media, y otros con valor muy superior.

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
#| output: false  
  
## Reordenamos los niveles de los factores para una visualización más equilibrada  
data_overall$State <- with(  
  data_overall,  
  reorder(State , AgeAdjustedDeathRate, median , na.rm = T)  
)  
  
data_overall |>  
  DataExplorer::plot_boxplot(  
    by = 'State',  
    ggtheme = ggplot2::theme_bw(),  
    ncol = 4L,  
    nrow = 2L
```

```
)
```

```
  ggplot2::ggsave(  
    here::here('notebooks', 'images', 'eda_data_overall_03fa_boxplot.jpg')  
)
```

```
```
```

```

```

```
:::
```

```
{.callout-caution title="03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`"
collapse="true"}
```

```
03fb - Test de hipótesis: `ggstatsplot::ggbetweenstats()`
```

En este dataset no hay observaciones suficientes para hacer un análisis comparativo de las variables numéricas por los niveles de la variable `State` .

```
:::
```

```
:::
```

```
03g - Explorar correlaciones (*Several methods*)
```

Se utilizaron las siguientes técnicas para explorar la correlación:

- 03ga - Matriz de correlación, mediante el test de Spearman
- 03gb - Correlograma, para visualizar la fuerza, la significación estadística y la dirección de la correlación
- 03gc - El test de hipótesis estadístico para la correlación

```
```{r}
```

```

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

## Función para convertir la tabla de correlación en un data.frame
## Tomada de http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-
guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor = (cormat)[ut],
    p = pmat[ut]
  )
}

```
```
:::{.callout-note title="1- Objeto `data`" collapse="true"}
#### Objeto `data`


:::{.callout-caution title="03ga - Correlation matrix" collapse="true"}
##### 03ga - Correlation matrix

```

Se exploraron las correlaciones entre variables numéricas con el test de Spearman. Se obtuvieron los siguientes resultados (en verde, las correlaciones estadísticamente significativas):

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Análisis de correlación con `Hmisc::rcorr()`

corrAnalysis_data <- Hmisc::rcorr(
 as.matrix(data[,3:10]),
 type = c("spearman")
)

Tabla para visualización de resultados

Se colorean en verde las correlaciones significativas (test np Spearman)

Método: https://stackoverflow.com/questions/62730125/flextable-basic-conditional-formatting

tmpTbl_data <- flattenCorrMatrix(
 round(corrAnalysis_data$r, 2),
 round(corrAnalysis_data$P, 2)
)

colormatrix_data <- ifelse(tmpTbl_data$p < .05, "lightgreen", "white")

tmpTbl_data |>
 flextable::flextable() |>
 flextable::bg(bg = colormatrix_data)

```

```

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

psych::pairs.panels(
 data[,3:10],
 method = "spearman",
 hist.col = "steelblue",
 show.points = TRUE,
 stars = TRUE,
 gap = 0.05,
 pch = ".",
 ellipses = FALSE,
 scale = FALSE,
 jiggle = TRUE,
 factor = 2,
 main = "Correlation matrix",
 col = "darkred",
 pty = "m",
 font = 2,
)
```

```

:::

```

::: {.callout-caution title="03gb - Correlograma (visualización de la correlación)"
collapse="true"}

##### 03gb - Correlograma (visualización de la correlación)

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

corrplot::corrplot(
 corrAnalysis_data$r,
 method = "ellipse",
 type = "lower",
 order = "hclust",
 tl.col = "black",
 diag = FALSE
)
```
:::

::: {.callout-caution title="03gc - Test de hipótesis `ggstatsplot::ggcorrmat()`"
collapse="true"}

##### 03gc - Test de hipótesis `ggstatsplot::ggcorrmat()`


```

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre `Population` y `AgeAdjustedDeathRate`
- Una correlación positiva entre:
 - `Deaths` y `Population`
 - `Deaths` y `PercentageOfTotalDeaths`
 - `Population` y `PercentageOfTotalDeaths`
 - `HeavyDrinkingAdults` y `AgeAdjustedDeathRate`
 - `BingeDrinkingPrevalenceAdults` y `HeavyDrinkingAdults`
 - `BingeDrinkingPrevalenceAdults` y `BingeDrinkingFrecuencyAdults`
 - `BingeDrinkingPrevalenceAdults` y `BingeDrinkingIntensityAdults`
 - `BingeDrinkingIntensityAdults` y `BingeDrinkingFrecuencyAdults`

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

ggstatsplot::ggcorrmat(
 data = data,
 type = 'np', ## Non-parametric Spearman correlation
 output = 'dataframe'
)
```
:::
:::
```

```
::: {.callout-note title="2- Objeto `data_gender`" collapse="true"}
```

```
#### Objeto `data_gender`
```

```
::: {.callout-caution title="03ga - Correlation matrix" collapse="true"}
```

```
##### 03ga - Correlation matrix
```

Se exploraron las correlaciones entre variables numéricas con el test de Spearman. Se obtuvieron los siguientes resultados (en verde, las correlaciones estadísticamente significativas):

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
Análisis de correlación con `Hmisc::rcorr()`
```

```
corrAnalysis_data_gender <- Hmisc::rcorr(
```

```
as.matrix(data_gender[,3:10]),
```

```
type = c("spearman")
```

```
)
```

```
Tabla para visualización de resultados
```

```
Se colorean en verde las correlaciones significativas (test np Spearman)
```

```
Método: https://stackoverflow.com/questions/62730125/flextable-basic-conditional-formatting
```

```
tmpTbl_data_gender <- flattenCorrMatrix(
```

```
round(corrAnalysis_data_gender$r, 2),
```

```
round(corrAnalysis_data_gender$P, 2)
```

```
)

colormatrix_data_gender <- ifelse(tmpTbl_data_gender$p < .05, "lightgreen", "white")

tmpTbl_data_gender |>
 flextable::flextable() |>
 flextable::bg(bg = colormatrix_data_gender)

...

```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
  
psych::pairs.panels(  
  data_gender[,3:10],  
  method = "spearman",  
  hist.col = "steelblue",  
  show.points = TRUE,  
  stars = TRUE,  
  gap = 0.05,  
  pch = ".",  
  ellipses = FALSE,  
  scale = FALSE,  
  jiggle = TRUE,  
  factor = 2,  
  main = "Correlation matrix",  
  col = "darkred",
```

```

pty = "m",
font = 2,
)

````

:::

:::{.callout-caution title="03gb - Correlograma (visualización de la correlación)" collapse="true"}

03gb - Correlograma (visualización de la correlación)

``{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

corrplot::corrplot(
 corrAnalysis_data_gender$r,
 method = "ellipse",
 type = "lower",
 order = "hclust",
 tl.col = "black",
 diag = FALSE
)
````

:::

```

```
:::{.callout-caution title="03gc - Test de hipótesis `ggstatsplot::ggcorrmat()`"  
collapse="true"}
```

```
##### 03gc - Test de hipótesis `ggstatsplot::ggcorrmat()`
```

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre `Population` y `Deaths` .
- Una correlación positiva entre:
 - `Deaths` y `PercentageOfTotalDeaths` .
 - `Deaths` y `BingeDrinkingIntensityAdults` .
 - `Deaths` y `BingeDrinkingPrevalenceAdults` .
 - `Population` y `AgeAdjustedDeathRate` .
 - `Population` y `PercentageOfTotalDeaths` .
 - `Population` y `BingeDrinkingFrecuencyAdults` .
 - `AgeAdjustedDeathRate` y `HeavyDrinkingAdults` .
 - `AgeAdjustedDeathRate` y `BingeDrinkingFrecuencyAdults` .
 - `AgeAdjustedDeathRate` y `BingeDrinkingIntensityAdults` .
 - `AgeAdjustedDeathRate` y `BingeDrinkingPrevalenceAdults` .
 - `PercentageOfTotalDeaths` y `BingeDrinkingIntensityAdults` .
 - `PercentageOfTotalDeaths` y `BingeDrinkingPrevalenceAdults` .
 - `HeavyDrinkingAdults` y `BingeDrinkingIntensityAdults` .
 - `HeavyDrinkingAdults` y `BingeDrinkingPrevalenceAdults` .
 - `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults` .
 - `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults` .

```
```{r}
```

```
#| code-fold: true
```

```

#| info: false
#| warning: false
#| code-overflow: wrap

ggstatsplot::ggcorrmat(
 data = data_gender,
 type = 'np', ## Non-parametric Spearman correlation
 output = 'dataframe'
)

```
:::
```
:::
```
::: {.callout-note title="3- Objeto `data_overall` collapse="true"}
#### Objeto `data_overall`

::: {.callout-caution title="03ga - Correlation matrix" collapse="true"}
##### 03ga - Correlation matrix

```

Se exploraron las correlaciones entre variables numéricas con el test de Spearman. Se obtuvieron los siguientes resultados (en verde, las correlaciones estadísticamente significativas):

```

```
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

```

```
Análisis de correlación con `Hmisc::rcorr()`

corrAnalysis_data_overall <- Hmisc::rcorr(
 as.matrix(data_overall[,2:9]),
 type = c("spearman")
)

Tabla para visualización de resultados

Se colorean en verde las correlaciones significativas (test np Spearman)

Método: https://stackoverflow.com/questions/62730125/flextable-basic-
conditional-formatting

tmpTbl_data_overall <- flattenCorrMatrix(
 round(corrAnalysis_data_overall$r, 2),
 round(corrAnalysis_data_overall$P, 2)
)

colormatrix_data_overall <- ifelse(tmpTbl_data_overall$p < .05, "lightgreen", "white")

tmpTbl_data_overall |>
 flextable::flextable() |>
 flextable::bg(bg = colormatrix_data_overall)

```
```

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap
```

```
psych::pairs.panels(  
  data_overall[,2:9],  
  method = "spearman",  
  hist.col = "steelblue",  
  show.points = TRUE,  
  stars = TRUE,  
  gap = 0.05,  
  pch = ".",  
  ellipses = FALSE,  
  scale = FALSE,  
  jiggle = TRUE,  
  factor = 2,  
  main = "Correlation matrix",  
  col = "darkred",  
  pty = "m",  
  font = 2,  
)
```

```

:::

```
::: {.callout-caution title="03gb - Correlograma (visualización de la correlación)"
collapse="true"}
```

```
03gb - Correlograma (visualización de la correlación)
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

corrplot::corrplot(
  corrAnalysis_data_overall$r,
  method = "ellipse",
  type = "lower",
  order = "hclust",
  tl.col = "black",
  diag = FALSE
)
```
:::
```
::: {.callout-caution title="03gc - Test de hipótesis `ggstatsplot::ggcorrmat()`" collapse="true"}
##### 03gc - Test de hipótesis `ggstatsplot::ggcorrmat()`

```

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre:
 - `Deaths` y `HeavyDrinkingAdults` .
 - `Population` y `AgeAdjustedDeathRate` .
 - `Population` y `HeavyDrinkingAdults` .
- Una correlación positiva entre:
 - `Deaths` y `Population` .
 - `Deaths` y `PercentageOfTotalDeaths` .

```
- `PercentageOfTotalDeaths` y `HeavyDrinkingAdults`.  
- `Population` y `PercentageOfTotalDeaths`.  
- `AgeAdjustedDeathRate` y `HeavyDrinkingAdults`.  
- `HeavyDrinkingAdults` y `BingeDrinkingIntensityAdults`.  
- `BingeDrinkingFrequencyAdults` y `BingeDrinkingPrevalenceAdults`.
```

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

```
ggstatsplot::ggcorrmat(
 data = data_overall,
 type = 'np', ## Non-parametric Spearman correlation
 output = 'dataframe'
)
...
:::
:::
```

### 03h - Explorar modelos de datos para las correlaciones estadísticamente significativas

Se crearon modelos exploratorios para todos los pares de variables en las que se ha obtenido una correlación lineal estadística significativa, para cada uno de los \*data.frame\*.

:::{.callout-note title="1- Objeto `data` " collapse="true"}

```
Objeto `data`
```

Se identificaron correlaciones estadísticamente significativas en los siguientes pares de variables:

- Correlación negativa entre `Population` y `AgeAdjustedDeathRate`
- Correlación positiva entre:
  - `Deaths` y `Population`
  - `Deaths` y `PercentageOfTotalDeaths`
  - `Population` y `PercentageOfTotalDeaths`
  - `HeavyDrinkingAdults` y `AgeAdjustedDeathRate`
  - `BingeDrinkingPrevalenceAdults` y `HeavyDrinkingAdults`
  - `BingeDrinkingPrevalenceAdults` y `BingeDrinkingFrecuencyAdults`
  - `BingeDrinkingPrevalenceAdults` y `BingeDrinkingIntensityAdults`
  - `BingeDrinkingIntensityAdults` y `BingeDrinkingFrecuencyAdults`

```
`Population` y `AgeAdjustedDeathRate`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
##### `Population` y `AgeAdjustedDeathRate`
```

```
data |>
```

```
ggplot2::ggplot(
```

```
mapping = ggplot2::aes(
```

```
x = Population,  
y = AgeAdjustedDeathRate  
)  
) +  
ggplot2::geom_point() +  
ggplot2::geom_smooth()  
` ` `
```

```
##### `Deaths` y `Population`
```

```
` ` ` {r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap
```

```
##### `Population` y `Deaths`
```

```
data |>  
ggplot2::ggplot(  
mapping = ggplot2::aes(  
x = Population,  
y = Deaths  
)  
) +  
ggplot2::geom_point() +  
ggplot2::geom_smooth()  
` ` `
```

```
##### `Deaths` y `PercentageOfTotalDeaths`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
`PercentageOfTotalDeaths` y `Deaths`
```

```
data |>
```

```
ggplot2::ggplot(
```

```
mapping = ggplot2::aes(
```

```
x = PercentageOfTotalDeaths,
```

```
y = Deaths
```

```
)
```

```
) +
```

```
ggplot2::geom_point() +
```

```
ggplot2::geom_smooth()
```

```
```
```

```
##### `Population` y `PercentageOfTotalDeaths`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```

`Population` y `PercentageOfTotalDeaths`

data |>

 ggplot2::ggplot(

 mapping = ggplot2::aes(

 x = Population,

 y = PercentageOfTotalDeaths

)

) +

 ggplot2::geom_point() +

 ggplot2::geom_smooth()

...

`HeavyDrinkingAdults` y `AgeAdjustedDeathRate`

...{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

`HeavyDrinkingAdults` y `AgeAdjustedDeathRate`

data |>

 ggplot2::ggplot(

 mapping = ggplot2::aes(

 x = HeavyDrinkingAdults,

 y = AgeAdjustedDeathRate,

 color = Sex

)

```

```

) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
```

##### `BingeDrinkingPrevalenceAdults` y `HeavyDrinkingAdults`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

`BingeDrinkingPrevalenceAdults` y `HeavyDrinkingAdults`

data |>
 ggplot2::ggplot(
 mapping = ggplot2::aes(
 x = HeavyDrinkingAdults,
 y = BingeDrinkingPrevalenceAdults,
 color = Sex
)
) +
 ggplot2::geom_point() +
 ggplot2::geom_smooth()
```

##### `BingeDrinkingPrevalenceAdults` y `BingeDrinkingFrequencyAdults`
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

`BingeDrinkingPrevalenceAdults` y `BingeDrinkingFrecuencyAdults`

data |>
 ggplot2::ggplot(
 mapping = ggplot2::aes(
 x = BingeDrinkingFrecuencyAdults,
 y = BingeDrinkingPrevalenceAdults,
 color = Sex
)
) +
 ggplot2::geom_point() +
 ggplot2::geom_smooth()
```

##### `BingeDrinkingPrevalenceAdults` y `BingeDrinkingIntensityAdults`  

``{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `BingeDrinkingPrevalenceAdults` y `BingeDrinkingIntensityAdults`
```

```

data |>

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = BingeDrinkingIntensityAdults,
    y = BingeDrinkingPrevalenceAdults,
    color = Sex
  )
) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
````

`BingeDrinkingIntensityAdults` y `BingeDrinkingFrequencyAdults` `

````{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `BingeDrinkingFrequencyAdults` y `BingeDrinkingIntensityAdults` `

data |>

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = BingeDrinkingIntensityAdults,
    y = BingeDrinkingFrequencyAdults,
    color = Sex
  )
)

```

```

) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
```
:::

Objeto `data_gender`
```

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre `Population` y `Deaths` .
- Una correlación positiva entre:
  - `Deaths` y `PercentageOfTotalDeaths` .
  - `Deaths` y `BingeDrinkingIntensityAdults` .
  - `Deaths` y `BingeDrinkingPrevalenceAdults` .
  - `Population` y `AgeAdjustedDeathRate` .
  - `Population` y `PercentageOfTotalDeaths` .
  - `Population` y `BingeDrinkingFrecuencyAdults` .
  - `AgeAdjustedDeathRate` y `HeavyDrinkingAdults` .
  - `AgeAdjustedDeathRate` y `BingeDrinkingFrecuencyAdults` .
  - `AgeAdjustedDeathRate` y `BingeDrinkingIntensityAdults` .
  - `AgeAdjustedDeathRate` y `BingeDrinkingPrevalenceAdults` .
  - `PercentageOfTotalDeaths` y `BingeDrinkingIntensityAdults` .
  - `PercentageOfTotalDeaths` y `BingeDrinkingPrevalenceAdults` .
  - `HeavyDrinkingAdults` y `BingeDrinkingIntensityAdults` .
  - `HeavyDrinkingAdults` y `BingeDrinkingPrevalenceAdults` .
  - `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults` .

```

- `BingeDrinkingIntensityAdults` y `BingeDrinkingPrevalenceAdults`.

#####
`Population` y `Deaths`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

#####
`Population` y `Deaths`


data_gender |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = Population,
      y = Deaths
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()

```
#####`Deaths` y `PercentageOfTotalDeaths`


```{r}
#| code-fold: true
#| info: false
#| warning: false

```

```

##| code-overflow: wrap

##### `PercentageOfTotalDeaths` `y` `Deaths` `

data_gender |>

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = PercentageOfTotalDeaths,
    y = Deaths
  )
) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
```
```
##### `Deaths` `y` `BingeDrinkingIntensityAdults` `

```
```
{r}
##| code-fold: true
##| info: false
##| warning: false
##| code-overflow: wrap

##### `BingeDrinkingIntensityAdults` `y` `Deaths` `

data_gender |>

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = BingeDrinkingIntensityAdults,
    y = Deaths
  )
)

```

```

)
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
````

#####
`Deaths` y `BingeDrinkingPrevalenceAdults`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

#####
`BingeDrinkingPrevalenceAdults` y `Deaths`  

data_gender |>
ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = BingeDrinkingPrevalenceAdults,
    y = Deaths
  )
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
````

#####
`Population` y `AgeAdjustedDeathRate`
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `Population` y `AgeAdjustedDeathRate`  

data_gender |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = Population,
      y = AgeAdjustedDeathRate
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
```

```

```
`Population` y `PercentageOfTotalDeaths`
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `Population` y `PercentageOfTotalDeaths`  

data_gender |>
```

```

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = Population,
    y = PercentageOfTotalDeaths
  )
) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
````

`Population` y `BingeDrinkingFrecuencyAdults`


```
##| code-fold: true
##| info: false
##| warning: false
##| code-overflow: wrap

##### `Population` y `BingeDrinkingFrecuencyAdults`


data_gender |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = Population,
      y = BingeDrinkingFrecuencyAdults
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()

```

```
```
```

```
`AgeAdjustedDeathRate` y `HeavyDrinkingAdults`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
##### `AgeAdjustedDeathRate` y `HeavyDrinkingAdults`
```

```
data_gender |>
```

```
ggplot2::ggplot(
```

```
mapping = ggplot2::aes(
```

```
x = AgeAdjustedDeathRate,
```

```
y = HeavyDrinkingAdults,
```

```
color = Sex
```

```
)
```

```
) +
```

```
ggplot2::geom_point() +
```

```
ggplot2::geom_smooth()
```

```
```
```

```
`AgeAdjustedDeathRate` y `BingeDrinkingFrecuencyAdults`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

##### `AgeAdjustedDeathRate` `y` `BingeDrinkingFrecuencyAdults`  

data_gender |>  

ggplot2::ggplot(  

mapping = ggplot2::aes(  

x = AgeAdjustedDeathRate,  

y = BingeDrinkingFrecuencyAdults,  

color = Sex  

)  

)+  

ggplot2::geom_point() +  

ggplot2::geom_smooth()  

```

```

```
`AgeAdjustedDeathRate` `y` `BingeDrinkingIntensityAdults`
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
##### `AgeAdjustedDeathRate` `y` `BingeDrinkingIntensityAdults`
```

```

data_gender |>  

ggplot2::ggplot(  

mapping = ggplot2::aes(  


```

```

x = AgeAdjustedDeathRate,
y = BingeDrinkingIntensityAdults,
color = Sex

)
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
````

`AgeAdjustedDeathRate` y `BingeDrinkingPrevalenceAdults`

`{r}`

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

`AgeAdjustedDeathRate` y `BingeDrinkingPrevalenceAdults`

data_gender |>
ggplot2::ggplot(
 mapping = ggplot2::aes(
 x = AgeAdjustedDeathRate,
 y = BingeDrinkingPrevalenceAdults,
 color = Sex
)
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()

```

```

```
##### `PercentageOfTotalDeaths` y `BingeDrinkingIntensityAdults`
```

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
`HeavyDrinkingAdults` y `AgeAdjustedDeathRate`
```

```
data_gender |>
```

```
ggplot2::ggplot(
```

```
mapping = ggplot2::aes(
```

```
x = HeavyDrinkingAdults,
```

```
y = AgeAdjustedDeathRate,
```

```
color = Sex
```

```
)
```

```
) +
```

```
ggplot2::geom_point() +
```

```
ggplot2::geom_smooth()
```

```

```
##### `PercentageOfTotalDeaths` y `BingeDrinkingPrevalenceAdults`
```

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

`PercentageOfTotalDeaths` y `BingeDrinkingPrevalenceAdults`

data_gender |>

ggplot2::ggplot(

mapping = ggplot2::aes(

x = PercentageOfTotalDeaths,

y = BingeDrinkingPrevalenceAdults,

color = Sex

)

)+

ggplot2::geom_point() +

ggplot2::geom_smooth()

````  

##### `HeavyDrinkingAdults` y `BingeDrinkingIntensityAdults`  

```{r}  

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

`BingeDrinkingPrevalenceAdults` y `HeavyDrinkingAdults`

data_gender |>

ggplot2::ggplot(

mapping = ggplot2::aes(


```

```

x = HeavyDrinkingAdults,
y = BingeDrinkingPrevalenceAdults,
color = Sex
)
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
```

```

```
##### `HeavyDrinkingAdults` y `BingeDrinkingPrevalenceAdults`
```

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
`BingeDrinkingPrevalenceAdults` y `BingeDrinkingFrequencyAdults`
```

```
data_gender |>
ggplot2::ggplot(
 mapping = ggplot2::aes(
 x = HeavyDrinkingAdults,
 y = BingeDrinkingPrevalenceAdults,
 color = Sex
)
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
```

```

```
##### `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults`
```

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
`BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults`
```

```
data_gender |>
```

```
ggplot2::ggplot(
```

```
mapping = ggplot2::aes(
```

```
 x = BingeDrinkingFrecuencyAdults,
```

```
 y = BingeDrinkingPrevalenceAdults,
```

```
 color = Sex
```

```
)
```

```
) +
```

```
ggplot2::geom_point() +
```

```
ggplot2::geom_smooth()
```

```

```
##### `BingeDrinkingIntensityAdults` y `BingeDrinkingFrecuencyAdults`
```

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

`BingeDrinkingFrecuencyAdults` y `BingeDrinkingIntensityAdults`

data_gender |>

ggplot2::ggplot(

 mapping = ggplot2::aes(

 x = BingeDrinkingIntensityAdults,

 y = BingeDrinkingFrecuencyAdults,

 color = Sex

)

) +

 ggplot2::geom_point() +

 ggplot2::geom_smooth()

...

:::

::: {.callout-note title="3- Objeto `data_overall` collapse="true"}

Objeto `data_overall`
```

Se han encontrado las siguientes correlaciones estadísticamente significativas:

- Una correlación negativa entre:
  - `Deaths` y `HeavyDrinkingAdults` .
  - `Population` y `AgeAdjustedDeathRate` .
  - `Population` y `HeavyDrinkingAdults` .
- Una correlación positiva entre:
  - `Deaths` y `Population` .

```
- `Deaths` y `PercentageOfTotalDeaths`.
- `Population` y `PercentageOfTotalDeaths`.
- `AgeAdjustedDeathRate` y `HeavyDrinkingAdults`.
- `PercentageOfTotalDeaths` y `HeavyDrinkingAdults`.
- `HeavyDrinkingAdults` y `BingeDrinkingIntensityAdults`.
- `BingeDrinkingFrequencyAdults` y `BingeDrinkingPrevalenceAdults`.
```

```
`Deaths` y `HeavyDrinkingAdults`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
##### `Deaths` y `HeavyDrinkingAdults`
```

```
data_overall |>
```

```
  ggplot2::ggplot(
```

```
    mapping = ggplot2::aes(
```

```
      x = Deaths,
```

```
      y = HeavyDrinkingAdults
```

```
    )
```

```
  ) +
```

```
  ggplot2::geom_point() +
```

```
  ggplot2::geom_smooth()
```

```
```
```

```
`Deaths` y `Population`
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `Deaths` y `Population`

data_overall |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = Deaths,
      y = Population
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()

```
`Deaths` y `PercentageOfTotalDeaths`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `Deaths` y `PercentageOfTotalDeaths`
```

```

data_overall |>

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = Deaths,
    y = PercentageOfTotalDeaths
  )
) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
````

`Population` y `AgeAdjustedDeathRate`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `Population` y `AgeAdjustedDeathRate`


data_overall |>

ggplot2::ggplot(
  mapping = ggplot2::aes(
    x = Population,
    y = AgeAdjustedDeathRate
  )
) +
  ggplot2::geom_point() +

```

```
ggplot2::geom_smooth()  
` ` `  
  
##### `Population` y `HeavyDrinkingAdults`  
  
` ` ` {r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
  
##### `Population` y `HeavyDrinkingAdults`  
data_overall |>  
ggplot2::ggplot(  
  mapping = ggplot2::aes(  
    x = Population,  
    y = HeavyDrinkingAdults  
  )  
  ) +  
  ggplot2::geom_point() +  
  ggplot2::geom_smooth()  
` ` `  
  
##### `Population` y `PercentageOfTotalDeaths`  
  
` ` ` {r}  
#| code-fold: true  
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

##### `Population` y `PercentageOfTotalDeaths`
data_overall |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = Population,
      y = PercentageOfTotalDeaths
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
````

`PercentageOfTotalDeaths` y `HeavyDrinkingAdults`

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `PercentageOfTotalDeaths` y `HeavyDrinkingAdults`
data_overall |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = PercentageOfTotalDeaths,

```

```

y = HeavyDrinkingAdults
)
) +
ggplot2::geom_point() +
ggplot2::geom_smooth()
````

`HeavyDrinkingAdults` y `BingeDrinkingIntensityAdults`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `Deaths` y `HeavyDrinkingAdults`


data_overall |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = HeavyDrinkingAdults,
      y = BingeDrinkingIntensityAdults
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
````

`BingeDrinkingFrequencyAdults` y `BingeDrinkingPrevalenceAdults`
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `BingeDrinkingFrecuencyAdults` y `BingeDrinkingPrevalenceAdults`
data_overall |>
  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = BingeDrinkingFrecuencyAdults,
      y = BingeDrinkingPrevalenceAdults
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
```

```

##### `AgeAdjustedDeathRate` y `HeavyDrinkingAdults`

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

##### `AgeAdjustedDeathRate` y `HeavyDrinkingAdults`
```

```

data_overall |>

  ggplot2::ggplot(
    mapping = ggplot2::aes(
      x = AgeAdjustedDeathRate,
      y = HeavyDrinkingAdults
    )
  ) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth()
``````

:::

```

### 03i - Análisis de valores faltantes (`NA`'s) y outliers

Se ejecutaron las siguientes acciones

- 03ia - Análisis de valores faltantes `NA`'s
- 03ib - Exploración de Outliers

::: {.callout-caution title="Funciones de R utilizadas en el análisis de datos faltantes (`NA`'s)" collapse="true"}

Para la evaluación de datos faltantes (`NA`'s) se evaluaron las siguientes dimensiones:

| Dimensión            | Función |  |
|----------------------|---------|--|
| Resultado evaluación |         |  |
| ----- ----- -----    |         |  |

|                                                                                                                         |                                       |
|-------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| -----                                                                                                                   |                                       |
| Existencia de algún valor valor faltante `NA` (sí/no)                                                                   | `naniar::any_na()`                    |
| `TRUE`                                                                                                                  |                                       |
|                                                                                                                         |                                       |
| Número total de `NA`'s                                                                                                  | `naniar::n_miss()`                    |
| \$69\$                                                                                                                  |                                       |
|                                                                                                                         |                                       |
| Variables afectadas por la presencia de `NA`'s                                                                          | `is.na()  > colSums()`                |
| `Sex`, `Deaths`, `Population`, `AgeAdjustedDeathRate`,                                                                  |                                       |
| `PercentageOfTotalDeaths`, `HeavyDrinkingAdults`,                                                                       |                                       |
| `BingeDrinkingFrequencyAdults`, `BingeDrinkingIntensityAdults` y                                                        |                                       |
| `BingeDrinkingPrevalenceAdults`                                                                                         |                                       |
| Número de `NA` por variable (n y %)                                                                                     |                                       |
| `naniar::miss_var_summary()` <br> `naniar::miss_var_table()`   Número: Rango \$0-12\$<br>Porcentaje: Rango \$0-7.36\%\$ |                                       |
|                                                                                                                         |                                       |
| Número de `NA` por observación (n y %)                                                                                  |                                       |
| `naniar::miss_case_summary()` <br> `naniar::miss_case_table()`   Número: Rango \$0-9\$<br>Porcentaje: Rango \$0-90\%\$  |                                       |
|                                                                                                                         |                                       |
| Ranking de variables más afectadas por `NA`'s                                                                           | `naniar::gg_miss_var()`               |
| `Deaths`, `Population`, `PercentageOfTotalDeaths`, `AgeAdjustedDeathRate`                                               |                                       |
|                                                                                                                         |                                       |
| Tipología de los valores faltantes (MAR, MNAR, MCAR)                                                                    | `naniar::vis_miss()`                  |
| Los valores faltantes se concentran en las últimas observaciones del dataset (MNAR)                                     |                                       |
|                                                                                                                         |                                       |
| Relación entre valores faltantes de distintas variables                                                                 |                                       |
| `naniar::gg_miss_upset()`                                                                                               | Existe una cierta tendencia a agrupar |
| valores faltantes para ciertas variables: `Deaths`, `Population`,                                                       |                                       |
| `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths`                                                                      |                                       |
|                                                                                                                         |                                       |
| Relación entre valores faltantes y niveles de las variables categóricas                                                 |                                       |
| `naniar::gg_miss_fct()`                                                                                                 | Algunos estados concentran todos los  |
| valores faltantes. No hay influencia del sexo                                                                           |                                       |
|                                                                                                                         |                                       |
| ...:                                                                                                                    |                                       |

```

::: {.callout-note title="1- Objeto `data`" collapse="true"}
Objeto `data`

::: {.callout-caution title="03ia - Análisis de valores faltantes `NA`s" collapse="true"}
03ia - Análisis de valores faltantes `NA`s

```{r eda - Explorar NAs y outliers}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

## Existencia de algún valor valor faltante `NA` (sí/no)
naniar::any_na(data)

## Número total de `NA`s
naniar::n_miss(data)

## Variables afectadas por la presencia de `NA`s
data |>
  is.na() |>
  colSums()

## Número de `NA` por variable (n y %)
naniar::miss_var_summary(data)
naniar::miss_var_table(data)

```

```

## Número de `NA` por observación (n y %)

naniar::miss_case_summary(data)

naniar::miss_case_table(data)

## Ranking de variables afectadas por `NA`'s

naniar::gg_miss_var(data)

## Tipología de los valores faltantes (MAR, MNAR, MCAR)

naniar::vis_miss(data) +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 80))

## Relación entre valores faltantes de distintas variables

naniar::gg_miss_upset(data)

## Relación entre valores faltantes y niveles de las variables categóricas

naniar::gg_miss_fct(data, fct = Sex)

naniar::gg_miss_fct(data, fct = State)

```

```

Hallazgos:

- Los valores faltantes se concentran en las últimas observaciones del dataset , para estados concretos (\*Missing not at random\*, MNAR).
  - Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: `Deaths` , `Population` , `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths`
  - Algunos estados concentran todos los valores faltantes. No hay influencia del sexo.
- ```

```
::: {.callout-caution title="03ib - Exploración de Outliers" collapse="true"}
```

#### ##### 03ib - Exploración de Outliers

Existe un número significativo de outliers en el \*data.frame\*. Eso deberá tenerse en cuenta para el análisis cluster.

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
data |>
```

```
dlookr::diagnose_outlier() |>
```

```
flextable::flextable()
```

```
```
```

```
`performance:: check_outliers()`
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
performance::check_outliers(
```

```
  data$Deaths,
```

```
  method = 'zscore'
```

```
) |>
plot()

```
Visualize variables with a ratio of outliers greater than 5%
`dlookr::plot_outlier()`
```

Ninguna variable tiene más de un 5% de outliers en sus valores.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

## Visualize variables with a ratio of outliers greater than 5%
data |>
dlookr::diagnose_outlier() |>
dplyr::filter(outliers_ratio > 5) |>
dplyr::select(variables) |>
dplyr::pull()

```
:::
:::
```

::: {.callout-note title="2- Objeto `data\_gender`" collapse="true"}

#### Objeto `data\_gender`

```

:::{.callout-caution title="03ia - Análisis de valores faltantes `NA`s" collapse="true"}

03ia - Análisis de valores faltantes `NA`s

```{r}

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap


## Existencia de algún valor valor faltante `NA` (sí/no)
naniar::any_na(data_gender)

## Número total de `NA`s
naniar::n_miss(data_gender)

## Variables afectadas por la presencia de `NA`s
data_gender |>
  is.na() |>
  colSums()

## Número de `NA` por variable (n y %)
naniar::miss_var_summary(data_gender)
naniar::miss_var_table(data_gender)

## Número de `NA` por observación (n y %)
naniar::miss_case_summary(data_gender)
naniar::miss_case_table(data_gender)

```

```

## Ranking de variables afectadas por `NA`'s
naniar::gg_miss_var(data_gender)

## Tipología de los valores faltantes (MAR, MNAR, MCAR)
naniar::vis_miss(data_gender) +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 80))

## Relación entre valores faltantes de distintas variables
naniar::gg_miss_upset(data_gender)

## Relación entre valores faltantes y niveles de las variables categóricas
naniar::gg_miss_fct(data_gender, fct = Sex)
naniar::gg_miss_fct(data_gender, fct = State)
```

```

Hallazgos:

- Los valores faltantes se concentran en las últimas observaciones del dataset , para estados concretos (\*Missing not at random\*, MNAR).
- Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: `Deaths` , `Population` , `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths`
- Algunos estados concentran todos los valores faltantes. No hay influencia del sexo.

```

::: {.callout-caution title="03ib - Exploración de Outliers" collapse="true"}

03ib - Exploración de Outliers

Existe un número significativo de outliers en el *data.frame*. Eso deberá tenerse en cuenta para el análisis cluster.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

data_gender |>
 dlookr::diagnose_outlier() |>
 flextable::flextable()
```

#####
`performance:: check_outliers()`


```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

performance::check_outliers(
 data_gender$Deaths,
 method = 'zscore'
) |>
 plot()

```

```

```
##### Visualize variables with a ratio of outliers greater than 5%
`dlookr::plot_outlier()`
```

Ninguna variable tiene más de un 5% de outliers en sus valores.

```{r}

```
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
Visualize variables with a ratio of outliers greater than 5%
```

```
data_gender |>
 dlookr::diagnose_outlier() |>
 dplyr::filter(outliers_ratio > 5) |>
 dplyr::select(variables) |>
 dplyr::pull()
```

```

:::

:::

```
::: {.callout-note title="3- Objeto `data_overall` " collapse="true"}
```

```
#### Objeto `data_overall`
```

```
::: {.callout-caution title="03ia - Análisis de valores faltantes `NA`s" collapse="true"}
```

```
##### 03ia - Análisis de valores faltantes `NA`s
```

```

```{r}

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Existencia de algún valor valor faltante `NA` (sí/no)
naniar::any_na(data_overall)

Número total de `NA`'s
naniar::n_miss(data_overall)

Variables afectadas por la presencia de `NA`'s
data_overall |>
 is.na() |>
 colSums()

Número de `NA` por variable (n y %)
naniar::miss_var_summary(data_overall)
naniar::miss_var_table(data_overall)

Número de `NA` por observación (n y %)
naniar::miss_case_summary(data_overall)
naniar::miss_case_table(data_overall)

Ranking de variables afectadas por `NA`'s
naniar::gg_miss_var(data_overall)

```

```

Tipología de los valores faltantes (MAR, MNAR, MCAR)

naniar::vis_miss(data_overall) +
 ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 80))

Relación entre valores faltantes de distintas variables

naniar::gg_miss_upset(data_overall)

Relación entre valores faltantes y niveles de las variables categóricas

naniar::gg_miss_fct(data_overall, fct = State)
```

```

Hallazgos:

- Los valores faltantes se concentran en las últimas observaciones del dataset , para estados concretos (*Missing not at random*, MNAR).
- Existe una cierta tendencia a agrupar valores faltantes para ciertas variables: `Deaths` , `Population` , `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths`
- Algunos estados concentran todos los valores faltantes. No hay influencia del sexo.

```

::: {.callout-caution title="03ib - Exploración de Outliers" collapse="true"}

##### 03ib - Exploración de Outliers

Existe un número significativo de outliers en el \*data.frame\*. Eso deberá tenerse en cuenta para el análisis cluster.

```{r}

#| code-fold: true

```

#| info: false
#| warning: false
#| code-overflow: wrap

data_overall |>
  dlookr::diagnose_outlier() |>
  flextable::flextable()
  ```

`performance::check_outliers()`

``{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

performance::check_outliers(
 data_overall$Deaths,
 method = 'zscore'
) |>
 plot()

```

##### Visualize variables with a ratio of outliers greater than 5%
`dlookr::plot_outlier()`

```

Ninguna variable tiene más de un 5% de outliers en sus valores.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Visualize variables with a ratio of outliers greater than 5%
data_overall |>
 dlookr::diagnose_outlier() |>
 dplyr::filter(outliers_ratio > 5) |>
 dplyr::select(variables) |>
 dplyr::pull()
```
:::
:::
```

Salidas del subprocesso

- Análisis de variables categóricas
- Análisis de variables numéricas
- Análisis de distribución de variables aleatorias
- Estudio de normalidad de las variables numéricas
- Comparación de los valores de las variables numéricas según niveles de las variables categóricas
- Estudio de correlación lineal entre variables numéricas
- Exploración de modelos de datos para correlaciones estadísticamente significativas

- Análisis de datos faltantes
- Análisis de datos extremos (*outliers*)

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Limpieza de los objetos temporales del subprocesso
rm(list = c(
 'data',
 'data_gender',
 'data_overall',
 'tmpTbl_data',
 'tmpTbl_data_gender',
 'tmpTbl_data_overall',
 'colormatrix_data',
 'colormatrix_data_gender',
 'colormatrix_data_overall',
 'flattenCorrMatrix'
))

```
## Transformación
```

Se realizaron las siguientes tareas de transformación:

- 4a - Tratamiento de valores faltantes
- 4b - Tratamiento de valores atípicos (*outliers*)

4a - Tratamiento de valores faltantes

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Configuración
Establecer una semilla aleatoria para el análisis
set.seed(2024)
Impedir que los números grandes se muestren con notación científica
options(scipen = 999)

Ingesta
data <- readRDS(
 here::here('data', 'lab', 'data.rds')
)
data_overall <- readRDS(
 here::here('data', 'lab', 'data_overall.rds')
)
data_gender <- readRDS(
 here::here('data', 'lab', 'data_gender.rds')
)
```

```

Se creo un dataset de trabajo sin datos faltantes, una para cada dataset de interés.

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

## Datos totales

data\_lab <- na.omit(data)

data\_gender\_lab <- na.omit(data\_gender)

data\_overall\_lab <- na.omit(data\_overall)

## Limpieza de datos intermedios

rm(list = c(

'data',

'data\_gender',

'data\_overall'

)

)

```

Tras la omisión de `NA`'s, los dos datasets `data_lab` y `data_gender_lab` son idénticos, y sólo difieren en los atributos que se han ido creando durante el proceso de limpieza. Por tanto, podemos trabajar exclusivamente con `data_lab` (para datos por sexos) y `data_overall` (para datos globales):

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
arsenal::comparedf(data_lab, data_gender_lab)
```

```
```
```

4b - Tratamiento de valores atípicos (*outliers*)

Las observaciones con valores extremos para las variables estudiadas podrían ser muy interesantes para nuestro análisis, porque pueden contener información sobre los factores de riesgo más asociados a la mortalidad por alcohol.

```
:::{.callout-note title="1- Objeto `data_lab` " collapse="true"}
```

En el subprocesso de EDA se identificaron problemas de valores atípicos en cinco variables del objeto `data`: `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths`).

```
```{r}
```

```
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
Diagnóstico de outliers
```

```
dlookr::diagnose_outlier(data_lab)
```

```
```
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

data_lab |> ggplot2::ggplot(
 ggplot2::aes(
 x = Sex,
 y = HeavyDrinkingAdults,
 fill = Sex
)
) +
 ggplot2::geom_boxplot() +
 ggside::geom_ysidedensity() +
 ggplot2::theme_bw()

```

```

data_lab |> ggplot2::ggplot(
 ggplot2::aes(
 x = Sex,
 y = BingeDrinkingFrecuencyAdults,
 fill = Sex
)
) +
 ggplot2::geom_boxplot() +
 ggside::geom_ysidedensity() +

```

```

ggplot2::theme_bw()

data_lab |> ggplot2::ggplot(
 ggplot2::aes(
 x = Sex,
 y = BingeDrinkingIntensityAdults,
 fill = Sex
)
) +
 ggplot2::geom_boxplot() +
 ggside::geom_ysidedensity() +
 ggplot2::theme_bw()

data_lab |> ggplot2::ggplot(
 ggplot2::aes(
 x = Sex,
 y = BingeDrinkingPrevalenceAdults,
 fill = Sex
)
) +
 ggplot2::geom_boxplot() +
 ggside::geom_ysidedensity() +
 ggplot2::theme_bw()
```

```

Se crearon dos conjuntos de datos para poder realizar el análisis de agrupación con y sin datos atípicos.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Valores extremos por variable

outliersDeaths <- boxplot.stats(data_lab$Deaths)$out
outliersPopulation <- boxplot.stats(data_lab$Population)$out
outliersAgeAdjustedDeathRate <- boxplot.stats(data_lab$AgeAdjustedDeathRate)$out
outliersPercentageOfTotalDeaths <-
 boxplot.stats(data_lab$PercentageOfTotalDeaths)$out
outliersHeavyDrinkingAdults <- boxplot.stats(data_lab$HeavyDrinkingAdults)$out

Índice de los outliers

idxOutliersDeaths <-
 which(data_lab$Deaths %in% outliersDeaths)
idxOutliersPopulation <-
 which(data_lab$Population %in% outliersPopulation)
idxOutliersAgeAdjustedDeathRate <-
 which(data_lab$AgeAdjustedDeathRate %in% outliersAgeAdjustedDeathRate)
idxOutlierPercentageOfTotalDeaths <-
 which(data_lab$PercentageOfTotalDeaths %in% outliersPercentageOfTotalDeaths)
idxOutlierHeavyDrinkingAdults <-
 which(data_lab$HeavyDrinkingAdults %in% outliersHeavyDrinkingAdults)

Vector de ayuda para identificar outliers

idxOutlierAll <- c(

```

```

idxOutliersDeaths,
idxOutlierHeavyDrinkingAdults,
idxOutlierPercentageOfTotalDeaths,
idxOutliersAgeAdjustedDeathRate,
idxOutliersPopulation

) |>
unique() |>
sort()

esOutlier <- 1:nrow(data_lab) %in% idxOutlierAll

Dataset sin outliers

data_outliers_lab <- data_lab[esOutlier,]
data_inliers_lab <- data_lab[!esOutlier,]

data_inliers_lab$Sex <- as.factor(data_inliers_lab$Sex)
data_inliers_lab$Sex <- as.factor(data_inliers_lab$Sex)
```
:::
```{r}
En el subprocesso de EDA se identificaron problemas de valores atípicos en cinco variables del objeto `data_overall`: `Deaths`, `Population`, `AgeAdjustedDeathRate` y `PercentageOfTotalDeaths` y `BingeDrinkingFrequencyAdults`).

```

```

```{r}
#| code-fold: true
#| info: false

```

```

#| warning: false
#| code-overflow: wrap

# Diagnóstico de outliers
dlookr::diagnose_outlier(data_overall_lab)
```
```
```
```

data_overall_lab |> ggplot2::ggplot(
  ggplot2::aes(
    x = State,
    y = Deaths
  )
) +
  ggplot2::geom_boxplot() +
  ggside::geom_ysidedensity() +
  ggplot2::theme_bw()

data_overall_lab |> ggplot2::ggplot(
  ggplot2::aes(
    x = State,
    y = Population

```

```
)  
)+  
ggplot2::geom_boxplot() +  
ggside::geom_ysidedensity() +  
ggplot2::theme_bw()
```

```
data_overall_lab |> ggplot2::ggplot(  
  ggplot2::aes(  
    x = State,  
    y = AgeAdjustedDeathRate  
)  
)+  
  ggplot2::geom_boxplot() +  
  ggside::geom_ysidedensity() +  
  ggplot2::theme_bw()
```

```
data_overall_lab |> ggplot2::ggplot(  
  ggplot2::aes(  
    x = State,  
    y = PercentageOfTotalDeaths  
)  
)+  
  ggplot2::geom_boxplot() +  
  ggside::geom_ysidedensity() +  
  ggplot2::theme_bw()
```

```
data_overall_lab |> ggplot2::ggplot(  
  ggplot2::aes(
```

```

x = State,
y = BingeDrinkingFrecuencyAdults
)
) +
ggplot2::geom_point() +
ggsidetoplot::geom_ysidedensity() +
ggplot2::theme_bw()
```

```

Se crearon dos conjuntos de datos para poder realizar el análisis de agrupación con y sin datos atípicos.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Valores extremos por variable
outliersDeaths <- boxplot.stats(data_overall_lab$Deaths)$out
outliersPopulation <- boxplot.stats(data_overall_lab$Population)$out
outliersAgeAdjustedDeathRate <-
boxplot.stats(data_overall_lab$AgeAdjustedDeathRate)$out
outliersPercentageOfTotalDeaths <-
boxplot.stats(data_overall_lab$PercentageOfTotalDeaths)$out
outliersBingeDrinkingFrecuencyAdults <-
boxplot.stats(data_overall_lab$HBingeDrinkingFrecuencyAdults)$out

```

```

# Índice de los outliers

idxOutliersDeaths <-
  which(data_overall_lab$Deaths %in% outliersDeaths)

idxOutliersPopulation <-
  which(data_overall_lab$Population %in% outliersPopulation)

idxOutliersAgeAdjustedDeathRate <-
  which(data_overall_lab$AgeAdjustedDeathRate %in% outliersAgeAdjustedDeathRate)

idxOutlierPercentageOfTotalDeaths <-
  which(data_overall_lab$PercentageOfTotalDeaths %in%
outliersPercentageOfTotalDeaths)

idxOutlierBingeDrinkingFrecuencyAdults <-
  which(data_overall_lab$BingeDrinkingFrecuencyAdults %in%
outliersBingeDrinkingFrecuencyAdults)

# Vector de ayuda para identificar outliers

idxOutlierAll <- c(
  idxOutliersDeaths,
  idxOutliersPopulation,
  idxOutliersAgeAdjustedDeathRate,
  idxOutlierPercentageOfTotalDeaths,
  idxOutlierBingeDrinkingFrecuencyAdults
) |>
unique() |>
sort()

esOutlier <- 1:nrow(data_overall_lab) %in% idxOutlierAll

# data_overallset sin outliers

data_overall_outliers_lab <- data_overall_lab[esOutlier,]

```

```

data_overall_inliers_lab <- data_overall_lab[!esOutlier,]

```
```

```{r}
Limpieza de objetos intermedios de la tarea

rm(list = c(
 'esOutlier',
 'idxOutlierAll',
 'idxOutlierBingeDrinkingFrecuencyAdults',
 'idxOutlierHeavyDrinkingAdults',
 'idxOutlierPercentageOfTotalDeaths',
 'idxOutliersAgeAdjustedDeathRate',
 'idxOutliersDeaths',
 'idxOutliersPopulation',
 'outliersAgeAdjustedDeathRate',
 'outliersBingeDrinkingFrecuencyAdults',
 'outliersDeaths',
 'outliersHeavyDrinkingAdults',
 'outliersPercentageOfTotalDeaths',
 'outliersPopulation'
))

```
```

:::

Salidas del subprocesso

```

Se crearon los siguientes objetos, diferenciados entre sí por la presencia o ausencia de tres características: datos estratificados por sexo, Inliers y Outliers:

Objeto	Datos por sexo	Inliers	Outliers
`data_lab`	Sí	Sí	Sí
`data_inliers_lab`	Sí	Sí	No
`data_outliers_lab`	Sí	No	Sí
`data_overall_lab`	No	Sí	Sí
`data_overall_inliers_lab`	No	Sí	No
`data_overall_outliers_lab`	No	No	Sí

```{r}

```
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
# Objeto `data_lab`
saveRDS(
  data_lab,
  here::here('data', 'lab', 'data_lab.rds')
)
saveRDS(
  data_inliers_lab,
  here::here('data', 'lab', 'data_inliers_lab.rds')
)
saveRDS(
  data_outliers_lab,
  here::here('data', 'lab', 'data_outliers_lab.rds')
```

```
)
```

```
# Objeto `data_overall_lab`  
saveRDS(  
  data_overall_lab,  
  here::here('data', 'lab', 'data_overall_lab.rds')  
)  
saveRDS(  
  data_overall_inliers_lab,  
  here::here('data', 'lab', 'data_overall_inliers_lab.rds')  
)  
saveRDS(  
  data_overall_outliers_lab,  
  here::here('data', 'lab', 'data_overall_outliers_lab.rds')  
)  
```
```

```
Análisis clúster no jerárquico (k-means)
```

Se llevó a cabo un conjunto de análisis cluster (5 en total), siguiendo la siguiente metodología

- 05fa - Selección de los datos adecuados para el análisis cluster
- 05fb - Estandarización de valores numéricos
- 05fc - Cálculo de la distancia entre observaciones
- 05fd - Análisis de tendencia de agrupación
- 05fe - Elección del método y la vinculación de grupos

- 05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación.
- 05fg - Representación e interpretación de los resultados.
- 05fh - Evaluación de la importancia de las variables
- 05fi - Visualización de las agrupaciones cluster
- 05fj - Validación de la agrupación
- 05fk - Resumen de los resultados obtenidos

#### ### 05fa - Selección de los datos adecuados para el análisis cluster

Durante la fase de análisis exploratorio se evidenció una marcada diferencia en la mortalidad relacionada con alcohol entre ambos sexos, por lo que se analizará el dataset `data\_lab` para controlar el efecto de la variable `Sex` .

Además, se observaron problemas de valores atípicos en los distintos conjuntos de datos considerados para el análisis. Estas observaciones podrían ser muy interesantes para nuestro análisis, porque pueden contener información sobre los factores de riesgo más asociados a la mortalidad por alcohol.

- `data\_lab` : con todos los datos (incluyendo \*outliers\*), y
- `data\_inliers\_lab` : con datos recortados (sin \*outliers\*).

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Ingesta
data_lab <- readRDS(here::here('data', 'lab', 'data_lab.rds'))
```

```
data_inliers_lab <- readRDS(here::here('data', 'lab', 'data_inliers_lab.rds'))
```

```
```
```

### ### 05fb - Estandarización de valores numéricos

Para impedir que las diferencias de magnitud entre las variables numéricas alterase la agrupación, se escalaron los valores de ambos datasets.

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Datos totales
```

```
data_std <- scale(data_lab[,-(1:2)])
```

```
## Datos recortados
```

```
data_inliers_std <- scale(data_inliers_lab[,-(1:2)])
```

```
```
```

### ### 05fc - Cálculo de la distancia entre observaciones

Se utilizó la función `stat::dist()` con los parámetros por defecto (distancia euclídea):

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```

#| warning: false
#| code-overflow: wrap

## Datos totales
data_dist <- dist(data_std)

## Datos recortados
data_inliers_dist <- dist(data_inliers_std)
```

Visualización de la relación entre las variables estandarizadas

:::{.callout-note title="1- Objeto `data_lab`" collapse="true"}
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| eval: false

## Datos totales
GGally::ggpairs(data_lab[2:10], ggplot2::aes(colour = Sex)) +
  ggplot2::ggtitle('Objeto data_lab', subtitle = 'Análisis de correlación')
```
:::

:::{.callout-note title="2- Objeto `data_inliers_lab`" collapse="true"}
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| eval: false
```

```

```

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
#| eval: false

Datos recortados

GGally::ggpairs(data_inliers_lab[2:10], ggplot2::aes(colour = Sex)) +
 ggplot2::ggtitle('Objeto data_inliers_lab', subtitle = 'Análisis de correlación')
```
```
:::

```

### 05fd - Análisis de tendencia de agrupación

Valoramos en primer lugar si es pertinente realizar un análisis de agrupación de los datos. Para ello:

- Analizamos visualmente los clústeres de datos con el análisis visual de tendencia (VAT)
- Evaluamos la tendencia de agrupación con el estadístico de Hopkins.

#### #### Evaluación visual de tendencia (VAT)

Este mapa del calor reordena la matriz de tal manera que observaciones similares se localizan cerca. Visualmente, se observan entre tres y cinco grandes clusters, que son más evidentes cuando se eliminan los outliers.

```

::: {.callout-note title="1- Objeto `data_lab` " collapse="true"}
```
{r}

```

```

## Datos totales

factoextra::fviz_dist(
  data_dist,
  lab_size = .1,
  show_labels = FALSE
) +
  ggplot2::ggtitle(
    label = "Evaluación visual de tendencia de agrupación (VAT)",
    subtitle = 'Datos completos (`data_lab`)'
  )
  ...
  :::

:::{.callout-note title="2- Objeto `data_inliers_lab`" collapse="true"}
``{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

## Datos recortados

factoextra::fviz_dist(
  data_inliers_dist,
  lab_size = .1,
  show_labels = FALSE
) +
  ggplot2::ggtitle(
    label = "Evaluación visual de tendencia de agrupación (VAT)",

```

```
    subtitle = 'Datos sin outliers (` data_inliers_lab `)'  
  )
```

```

:::

#### #### Estadística de Hopkins

En ambos supuestos (datos totales y recortados), el valor es distinto de 0.5, por lo que suponemos que las distancias observadas entre el conjunto de datos aleatorio y el conjunto de datos real no se debe al azar, y, por tanto, existe tendencia de agrupación.

```
::: {.callout-note title="1- Objeto ` data_lab ` " collapse="true"}
```

```{r}

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Datos totales
```

```
clustertend::hopkins(
```

```
  data_std,
```

```
  n = nrow(data_std) - 1
```

```
)
```

```

:::

```
::: {.callout-note title="2- Objeto ` data_inliers_lab ` " collapse="true"}
```

```

```{r}
## Datos recortados
clustertend::hopkins(
  data_inliers_std,
  n = nrow(data_inliers_std) - 1
)
```
:::
```

### 05fe - Elección del método y la vinculación de grupos

Se utilizó el método de agrupación por  $k$ -medias.

### 05ff - Elección del número de grupos finales de forma arbitraria basados en ciertos estadísticos de agrupación.

##### Según criterios de calidad interna

::: {.callout-note title="1- Objeto `data\_lab` " collapse="true"}

En el dataset completo, la mayoría de métodos sitúa el óptimo de clústeres entre 2 y 4.

```

```{r}
#| info: false
#| warning: false
#| code-overflow: wrap
```

Datos totales

`nb <- NbClust::NbClust(`

```
data = data_std,  
distance = 'euclidean',  
min.nc = 2,  
max.nc = 15,  
method = "kmeans",  
index = "all"  
)  
...  
:::
```

:::{.callout-note title="2- Objeto `data_inliers_lab` " collapse="true"}

En el dataset sin outliers, el número óptimo de clústeres está entre 2 y 3.

```
```{r}  
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
Datos recortados
nb_inliers <- NbClust::NbClust(
 data = data_inliers_std,
 distance = 'euclidean',
 min.nc = 2,
 max.nc = 15,
 method = "kmeans",
 index = "all")
```

```

:::

Según criterios de estabilidad

Dado que no podemos encontrar un nivel de clústeres óptimo en base a los resultados, exploraremos las opciones más repetidas:

- 2, 3 y 4 clústeres para `data_lab`, y
- 2 y 3 clústeres para `data_inliers_lab`.

::: {.callout-note title="1- Objeto `data_lab` " collapse="true"}

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

## Datos totales

data\_km2 <- kmeans(data\_std, centers = 2)

data\_km3 <- kmeans(data\_std, centers = 3)

data\_km4 <- kmeans(data\_std, centers = 4)

```

:::

::: {.callout-note title="2- Objeto `data_inliers_lab` " collapse="true"}

```{r}

#| code-fold: true

```

#| info: false
#| warning: false
#| code-overflow: wrap

Datos recortados

data_inliers_km2 <- kmeans(data_inliers_std, centers = 2)
data_inliers_km3 <- kmeans(data_inliers_std, centers = 3)
```
```
:::

```

### 05fg - Representación e interpretación de los resultados.

```

::: {.callout-note title="1- Objeto `data_lab` " collapse="true"}
Datos completos `data_lab`
```

Resultados de los modelos de agrupación para `data\_lab`

- Los modelos explican un porcentaje de la variabilidad total observada entre el \$34.2\%\$ y el \$60.3\%\$. El modelo con mejor explicación de los datos observados es el de \$k=4\$ grupos
- Todos los modelos explorados son algo difíciles de interpretar, porque los outliers tienden a agruparse en uno de los grupos del modelo, y dificultan la comprensión de la lógica de la agrupación.

```

::: {.callout-caution title="1a- Agrupación $k = 2$" collapse="true"}
Agrupación $k = 2$
```

El resultado de la agrupación con 2 clústeres es el siguiente

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
## Datos totales
data_km2
table(data_km2$cluster)
```

```
```
```

La agrupación con dos clústeres explica un 34.2% de la variabilidad total. Ambos clústeres tienen aproximadamente el mismo número de elementos.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
## Datos totales
cl_kcca2 <- flexclust::as.kcca(data_km2, data_std)
flexclust::barplot(cl_kcca2)
```

```
```
```

- Los dos clústeres se diferencian entre sí esencialmente por los valores del área de interés `Alcohol` del CDI: valores elevados frente a valores bajos

Podemos explorar cómo se comporta cada cluster variable a variable con el siguiente gráfico

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

flexclust::barplot(cl_kcca2, bycluster = FALSE)
````
```

El resultado explica sólo un tercio de la variabilidad total, por lo que no es el ideal.

...:

```
:::{.callout-caution title="1b- Agrupación $k = 3$" collapse="true"}
Agrupación $k = 3$
```

La agrupación con tres clústeres explica un 53.9% de la variabilidad total. Los clústeres están muy desequilibrados, con uno de ellos con 3 elementos (los valores \*outliers\*)

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
## Datos totales  
data_km3  
table(data_km3$cluster)
```

```

La presencia del grupo con los outliers (cluster 1) dificulta la interpretación visual de los otros dos grupos, tanto en el gráfico global como por variables

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap
```

```
## Datos totales  
cl_kcca3 <- flexclust::as.kcca(data_km3, data_std)  
flexclust::barplot(cl_kcca3)
```

```

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap
```

```
flexclust::barplot(cl_kcca3, bycluster = FALSE)
```

```

El resultado genera una agrupación muy desbalanceada, por lo que tampoco parece el modelo óptimo.

:::

::: {.callout-caution title="1c- Agrupación \$k = 4\$" collapse="true"}

##### Agrupación \$k = 4\$

La agrupación con cuatro clústeres explica un 60.3% de la variabilidad total. De nuevo, una de las clases tiene muy pocos elementos (outliers)

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

Datos totales

data_km4

table(data_km4\$cluster)

```

El clúster que contiene los outliers también dificulta la interpretación de los grupos, aunque se intuye que se diferencian los siguientes cuatro grupos de estados en relación con los indicadores:

- Indicadores con valores bajos
- Indicadores con valores altos
- Indicadores con valores bajos en general, pero con valores muy altos en el número de adultos grandes bebedores (`HeavyDrinkingAdults`)
- \*Outliers\*

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Datos totales
```

```
cl_kcca4 <- flexclust::as.kcca(data_km4, data_std)
```

```
flexclust::barplot(cl_kcca4)
```

```
```
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
flexclust::barplot(cl_kcca4, bycluster = FALSE)
```

```
```
```

```
:::
```

```
:::
```

```
:::{.callout-note title="2- Objeto `data_inliers_lab` " collapse="true"}
Datos recortados `data_inliers_lab`
```

Resultados de los modelos de agrupación para `data\_inliers\_lab`

- Los modelos explican un porcentaje de la variabilidad total observada entre el  $38.32\%$  y el  $54\%$ . El modelo con mejor explicación de los datos observados es el de  $k=3$  grupos
- Los modelos con datos recortados explican un menor porcentaje de la variabilidad que los de datos completos; los outliers una capturan parte considerable de la información, y deben estudiarse con detalle.

```
:::{.callout-caution title="1a- Agrupación $k = 2$" collapse="true"}
Agrupación $k = 2$
```

El resultado de la agrupación con 2 clústeres es el siguiente

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
  
## Datos recortados  
data_inliers_km2  
table(data_inliers_km2$cluster)  
  
```
```

La agrupación con dos clústeres explica un 38.32% de la variabilidad total. El primer clúster tiene un poco más de elementos que el segundo.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

## Datos recortados

cl_inliers_kcca2 <- flexclust::as.kcca(data_inliers_km2, data_inliers_std)

flexclust::barplot(cl_inliers_kcca2)
```

```

- Los dos clústeres se diferencian entre sí por los valores del área de interés `Alcohol` del CDI: valores elevados frente a valores bajos

Podemos explorar cómo se comporta cada cluster variable a variable con el siguiente gráfico

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

flexclust::barplot(cl_inliers_kcca2, bycluster = FALSE)

```

```

El resultado explica un porcentaje muy pequeño de la variabilidad total, por lo que descartamos este modelo.

:::

```
::: {.callout-caution title="1 b- Agrupación $k = 3$" collapse="true"}
```

```
Agrupación $k = 3$
```

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Datos recortados
```

```
data_inliers_km3
```

```
table(data_inliers_km3$cluster)
```

```

La agrupación con dos clústeres explica un 54% de la variabilidad total. El primer cluster tiene más elementos que los otros dos juntos.

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap

## Datos totales

cl_inliers_kcca3 <- flexclust::as.kcca(data_inliers_km3, data_inliers_std)

flexclust::barplot(cl_inliers_kcca3)

````
```

- El primer cluster agrupa estados con indicadores con bajos bajos para todas las variables.
- El segundo incluye aquellos estados con indicadores por encima de la media, excepto el número de adultos grandes bebedores, pero en los que la tasa de muerte ajustada por edad está por debajo de la media
- El tercero agrupa a los estados con indicadores por encima de la media, incluido el porcentaje de adultos grandes bebedores, y en el que la tasa de muerte también es alta.

```
```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap
```

```
flexclust::barplot(cl_inliers_kcca3, bycluster = FALSE)
```

```
````

:::

:::
```

### ### 05fh - Evaluación de la importancia de las variables

```
:::{.callout-note title="1- Objeto `data_lab` " collapse="true"}
```

Los resultados de la evaluación de la importancia de las variables para los modelos para datos completos fueron los siguientes:

- En los modelos de 2 y 3 clústeres para datos completos, las variables más importantes para establecer la agrupación fueron las relacionadas con las características de las borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrecuencyAdults`)
- Para el modelo con 4 clústeres, la variable más importante con diferencia fue la de grandes bebedores (`HeavyDrinkingAdults`), seguida de las tres variables relacionadas con borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrecuencyAdults`)

```
:::{.callout-caution title="1a- Agrupación k = 2" collapse="true"}
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Datos totales
```

```
importance <- FeatureImpCluster::FeatureImpCluster(
```

```
cl_kcca2,
```

```
data.table::as.data.table(data_std)
```

```
)
```

```
flexclust::plot(importance)
```

```
```
```

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

## Datos recortados

importance_inliers <- FeatureImpCluster::FeatureImpCluster(
  cl_inliers_kcca2,
  data.table::as.data.table(data_inliers_std))
flexclust::plot(importance_inliers)
```

```

En el modelo de 2 clústeres para datos completos, las variables más importantes para establecer la agrupación fueron las relacionadas con las características de las borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrequencyAdults`).

:::

::: {.callout-caution title="1b- Agrupación k = 3" collapse="true"}

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

```

Datos totales

```

importance <- FeatureImpCluster::FeatureImpCluster(
  cl_kcca3,

```

```

data.table::as.data.table(data_std)
)

flexclust::plot(importance)

## Datos recortados

importance_inliers <- FeatureImpCluster::FeatureImpCluster(
  cl_inliers_kcca3,
  data.table::as.data.table(data_inliers_std))

flexclust::plot(importance_inliers)

```

```

En el modelo de 3 clústeres para datos completos, las variables más importantes para establecer la agrupación fueron las relacionadas con las características de las borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrequencyAdults`).

:::

::: {.callout-caution title="1c- Agrupación k = 4" collapse="true"}

```{r}

```

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

```

Datos totales

```

importance <- FeatureImpCluster::FeatureImpCluster(
  cl_kcca4,
  data.table::as.data.table(data_std)
)

```

```
flexclust::plot(importance)
```

...

En el modelo con 4 clústeres, la variable más importante con diferencia fue la de grandes bebedores (` HeavyDrinkingAdults`), seguida de las tres variables relacionadas con borracheras (` BingeDrinkingIntensityAdults` , ` BingeDrinkingPrevalenceAdults` , ` BingeDrinkingFrecuencyAdults`)

:::

:::

::: {.callout-note title="2- Objeto ` data_inliers_lab` " collapse="true"}

Los resultados de la evaluación de la importancia de las variables para los modelos para datos recortados fueron los siguientes:

- En el modelo de 2 clústeres para datos recortados, las variables más importantes para establecer la agrupación fueron las relacionadas con las características de las borracheras (` BingeDrinkingIntensityAdults` , ` BingeDrinkingPrevalenceAdults` , ` BingeDrinkingFrecuencyAdults`)
- Para el modelo con 3 clústeres, las variables más importantes para establecer la agrupación fueron las relacionadas el número de muertes relacionadas con alcohol (` Deaths` y ` PercentageOfTotalDeaths`), seguido de la prevalencia de borracheras (` BingeDrinkingPrevalenceAdults`).

::: {.callout-caution title="1a- Agrupación k = 2" collapse="true"}

```{r}

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```

Datos recortados

importance_inliers <- FeatureImpCluster::FeatureImpCluster(
 cl_inliers_kcca2,
 data.table::as.data.table(data_inliers_std))

flexclust::plot(importance_inliers)

```

```

En el modelo de 2 clústeres para datos recortados, las variables más importantes para establecer la agrupación fueron las relacionadas con las características de las borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrequencyAdults`).

```

::: {.callout-caution title="1b- Agrupación k = 3" collapse="true"}

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

```

```

## Datos recortados

importance_inliers <- FeatureImpCluster::FeatureImpCluster(
  cl_inliers_kcca3,
  data.table::as.data.table(data_inliers_std))

flexclust::plot(importance_inliers)

```

```

En el modelo de 3 clústeres para datos recortados, las variables más importantes para establecer la agrupación fueron las relacionadas el número de muertes relacionadas

con alcohol (`Deaths` y `PercentageOfTotalDeaths`), seguido de la prevalencia de borracheras (`BingeDrinkingPrevalenceAdults`).

:::

:::

### ### 05fi - Visualización de las agrupaciones cluster

::: {.callout-caution title="1a- Modelos \$k = 2\$" collapse="true"}

#### Modelos con \$k = 2\$

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

Datos totales

factoextra::fviz_cluster(

 data_km2,

 data_std,

 labelsize = 5,

 main = "k=2 grupos, datos completos",

 geom = "point"

) +

 ggrepel::geom_text_repel(

 label = paste(

 data_lab\$State,

 data_lab\$AgeAdjustedDeathRate,

```
sep = "_"),
size = 1.5)
```

```

El modelo de 2 clústeres en los datos completos está fuertemente influenciado por los outliers, y crea unos clústeres con poco sentido.

```
```{r}
```

```
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
## Datos totales
```

```
factoextra::fviz_cluster(
  data_inliers_km2,
  data_inliers_std,
  labelsize = 5,
  main = "k=2 grupos, datos recortados (sin outliers)",
  geom = "point"
) +
  ggrepel::geom_text_repel(
    label = paste(
      data_inliers_lab$State,
      data_inliers_lab$AgeAdjustedDeathRate,
      sep = "_"),

```

```
size = 1.5)
```

```
```
```

Al eliminar los outliers, el modelo agrupa los datos en dos grandes bloques, sin solapamientos

```
:::
```

```
::: {.callout-caution title="Modelos con $k=3$" collapse="true"}
```

```
Modelos con $k=3$
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
## Datos totales
```

```
factoextra::fviz_cluster(
```

```
  data_km3,
```

```
  data_std,
```

```
  labelsize = 5,
```

```
  main = "k=3 grupos (datos completos)",
```

```
  geom = "point"
```

```
) +
```

```
  ggrepel::geom_text_repel(
```

```
    label = paste(
```

```
  data_lab$State,  
  data_lab$AgeAdjustedDeathRate,  
  sep = "_"),  
  size = 1.5)  
  
```
```

El modelo de tres clústeres para los datos completos separa un grupo con los outliers, y otros dos grupos dentro del resto de los datos.

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
  
## Datos recortados  
factoextra::fviz_cluster(  
  data_inliers_km3,  
  data_inliers_std,  
  labelsize = 5,  
  main = "k=3 grupos, datos recortados (sin outliers)",  
  geom = "point"  
) +  
  ggrepel::geom_text_repel(  
  label = paste(  
    data_inliers_lab$State,
```

```
  data_inliers_lab$AgeAdjustedDeathRate,  
  sep = "_"),  
  size = 1.5)  
```
```

Al eliminar los outliers, el modelo de tres clústeres es capaz de separar tres grupos de datos con una cierta coherencia visual.

:::

```
::: {.callout-caution title="Modelos con $k=4$" collapse="true"}
Modelo con $k=4$
```

El modelo de cuatro clústeres para los datos completos separa un grupo con los outliers, y dos de los grupos presentan un alto grado de solapamiento en la representación bidimensional.

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap
```

```
## Datos totales  
factoextra::fviz_cluster(  
  data_km4,  
  data_std,  
  labelsize = 5,  
  main = "k=4 grupos (datos completos)",  
  geom = "point"
```

```

) +
ggrepel::geom_text_repel(
  label = paste(
    data_lab$State,
    data_lab$AgeAdjustedDeathRate,
    sep = "_"),
  size = 1.5)
```

```

Al representarlo en tres dimensiones, se observa que el solapamiento es menor. Por ejemplo, eligiendo las tres variables con mayor importancia para la agrupación del modelo  $k = 4\$$ , se puede obtener este gráfico interactivo:

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

df <- data_lab[1:10]
df$cluster <- factor(data_km4$cluster)

p <- plotly::plot_ly(
  df,
  x = ~HeavyDrinkingAdults,
  y = ~BingeDrinkingIntensityAdults,
  z = ~BingeDrinkingFrequencyAdults,

```

```

mode = 'markers',
color = ~cluster,
hoverinfo = 'text',
text = ~paste(
  '</br> State:', State,
  '</br> Sex:', Sex,
  '</br> Heavy Drinking Adults:', HeavyDrinkingAdults,
  '</br> Binge Drinking Intensity Adults:', BingeDrinkingIntensityAdults,
  '</br> Binge Drinking Frecuency Adults:', BingeDrinkingFrecuencyAdults
)
) |>
plotly::add_markers(size = 1.5)

```

p

...

:::

05fj - Validación de la agrupación

Interna

Para la validación interna se utilizó el diagrama de silueta, con los siguientes resultados:

| Modelo | Datos | Evaluación del gráfico de silueta

|

|-----|-----|-----|

| \$k=2\$ | Completos | Los dos clústeres tienen un rendimiento aceptable, aunque en el caso del cluster 1 es inferior a silueta media. |

| \$k=3\$ | Completos | Dos de los clústeres, tienen un rendimiento bueno, y el cluster más pequeño tiene un rendimiento muy inferior a lo esperado |

| \$k=4\$ | Completos | Sólo los clústeres 2 y 3 superaron la silueta media; todos los demás quedaron por debajo de lo deseable. |

| \$k=2\$ | Recortados | Ambos dos clústeres tienen un rendimiento bueno, por encima de la silueta media |

| \$k=3\$ | Recortados | Bastante equilibrado; todos los clústeres tienen un ancho de silueta medio igual al ancho de silueta medio. |

```
::: {.callout-caution title="1- Modelos $k=2$" collapse="true"}
```

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
Datos totales
```

```
sk2 <- cluster::silhouette(
```

```
 data_km2$cluster,
```

```
 data_dist)
```

```
sk2_mean <- mean(sk2[,3])
```

```
gráfico de silueta
```

```
flexclust::plot(
```

```
 sk2,
```

```
 main = "Silhouette plot - Kmeans k=2 (datos completos)",
```

```
cex.names = 0.8,
col = 1:2,
nmax = 100,
do.clust.stat = TRUE)

abline(v = sk2_mean, col = "darkblue", lty = 3)
```

```

Para el modelo de $k=2$ con datos completos, los dos clústeres tienen un rendimiento aceptable, aunque en el caso del cluster 1 es inferior a silueta media.

```
```{r}  
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

## Datos totales

```
sk2_inliers <- cluster::silhouette(
 data_inliers_km2$cluster,
 data_inliers_dist)

sk2_inliers_mean <- mean(sk2_inliers[,3])
```

```
gráfico de silueta
flexclust::plot(
 sk2_inliers,
 main = "Silhouette plot - Kmeans k=2 (datos recortados, sin outliers)",
```

```
cex.names = 0.8,
col = 1:2,
nmax = 100,
do.clust.stat = TRUE)

abline(v = sk2_inliers_mean, col = "darkblue", lty = 3)
```

```

Para el modelo de $k=2$ con datos recortados, ambos dos clústeres tienen un rendimiento bueno, por encima de la silueta media.

:::

```
:::{.callout-caution title="2- Modelos  $k=3$ " collapse="true"}
```

```
```{r}
```

```
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
Datos totales
```

```
sk3 <- cluster::silhouette(
 data_km3$cluster,
 data_dist)

sk3_mean <- mean(sk3[,3])
```

```
gráfico de silueta
```

```
flexclust::plot(
```

```

sk3,
main = "Silhouette plot - Kmeans k=3 (datos completos)",
cex.names = 0.8,
col = 1:3,
nmax = 100,
do.clust.stat = TRUE)

abline(v = sk3_mean, col = "darkblue", lty = 3)

```

```

Para el modelo de $k=3$ con datos completos, dos de los clústeres, tienen un rendimiento bueno, y el cluster más pequeño tiene un rendimiento muy inferior a lo esperado.

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

```

**## Datos totales**

```

sk3_inliers <- cluster::silhouette(
 data_inliers_km3$cluster,
 data_inliers_dist)

sk3_inliers_mean <- mean(sk3_inliers[,3])

```

# gráfico de silueta

```

flexclust::plot(
 sk3_inliers,
 main = "Silhouette plot - Kmeans k=3 (datos recortados, sin outliers)",
 cex.names = 0.8,
 col = 1:3,
 nmax = 100,
 do.clust.stat = TRUE)

abline(v = sk3_inliers_mean, col = "darkblue", lty = 3)

```

```

El modelo de $k=3$ con datos recortados es bastante equilibrado. Todos los clústeres tienen un ancho de silueta medio igual al ancho de silueta medio.

```

::: {.callout-caution title="3- Modelos  $k=4$ " collapse="true"}

```{r}

#| code-fold: true

#| info: false

#| warning: false

#| code-overflow: wrap

Datos totales

```

sk4 <- cluster::silhouette(
  data_km4$cluster,
  data_dist)

sk4_mean <- mean(sk4[,3])

```

```

# gráfico de silueta
flexclust::plot(
  sk4,
  main = "Silhouette plot - Kmeans k=4 (datos completos)",
  cex.names = 0.8,
  col = 1:4,
  nmax = 100,
  do.clust.stat = TRUE)
abline(v = sk4_mean, col = "darkblue", lty = 3)

```

```

Para el modelo de  $k=4$  con datos completos, sólo los clústeres 2 y 3 superaron la silueta media. Todos los demás quedaron por debajo de lo deseable.

:::

#### Externa

Se utilizaron las siguientes variables para la validación externa de las agrupaciones:

- Para los modelos  $k=2$ , se utilizó la variable `Sex` .
- Para los modelos  $k=3$  se creó una variable instrumental que discretizaba en tres niveles (alta, media y baja) la tasa de mortalidad ajustada por edad `AgeAdjustedDeathRate` .
- Para el modelo  $k=4$  se creó una variable instrumental que discretizaba en cuatro niveles (muy alta, alta, baja y muy baja) la tasa de mortalidad ajustada por edad `AgeAdjustedDeathRate` .

Los dos modelos  $k=2$ , tanto para datos completos como recortados, se ajustan bastante bien a los niveles de la variable `Sex`, por lo que capturan una información similar a esta variable.

Los modelos  $k=3$  y  $k=4$  se relacionan mal con la variable instrumental creada discretizando los valores de la variable `AgeAdjustedDeathRate`, con lo que es razonable suponer que capturan información no contenida en estas variables.

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Creación de las variables instrumentales para data_lab

# Variable instrumental de 2 niveles
data_lab$Sex <- as.factor(data_lab$Sex)

# Variable instrumental de 3 niveles
AgeAdjustedDeathRate_3levels_cuts <- recipes::discretize(
  data_lab$AgeAdjustedDeathRate,
  cuts = 3,
  labels = c('Mortalidad baja', 'Mortalidad media', 'Mortalidad alta'),
  prefix = ""
)
data_lab$AgeAdjustedDeathRate_fct3 <-
  predict(AgeAdjustedDeathRate_3levels_cuts, data_lab$AgeAdjustedDeathRate) |>
  as.factor()
```

```

# Variable instrumental de 4 niveles

AgeAdjustedDeathRate_4levels_cuts <- recipes::discretize(
  data_lab$AgeAdjustedDeathRate,
  cuts = 4,
  labels = c('Mortalidad muy baja', 'Mortalidad baja', 'Mortalidad alta', 'Mortalidad muy
alta'),
  prefix = "
)

data_lab$AgeAdjustedDeathRate_fct4 <-
predict(AgeAdjustedDeathRate_4levels_cuts, data_lab$AgeAdjustedDeathRate) |>
as.factor()

# Limpieza de variables temporales intermedias

rm(list = c(
  'AgeAdjustedDeathRate_3levels_cuts',
  'AgeAdjustedDeathRate_4levels_cuts'
))

```
```{r}

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Creación de las variables instrumentales para data_inliers_lab

```

```

# Variable instrumental de 2 niveles

data_inliers_lab$Sex <- as.factor(data_inliers_lab$Sex)

# Variable instrumental de 3 niveles

AgeAdjustedDeathRate_inliers_3levels_cuts <- recipes::discretize(
  data_inliers_lab$AgeAdjustedDeathRate,
  cuts = 3,
  labels = c('Mortalidad baja', 'Mortalidad media', 'Mortalidad alta'),
  prefix = "
)

data_inliers_lab$AgeAdjustedDeathRate_fct3 <-
predict(
  AgeAdjustedDeathRate_inliers_3levels_cuts,
  data_inliers_lab$AgeAdjustedDeathRate
) |>
as.factor()

# Limpieza de variables temporales intermedias

rm(list = c(
  'AgeAdjustedDeathRate_inliers_3levels_cuts'
))

```
:::{.callout-caution title="1 - Modelos con $k=2$" collapse="true"}
Modelos de 2 categorías

```

Ambos modelos cluster (tanto el de datos completos como el de datos recortados) separan perfectamente a las mujeres, y se equivocan con un pequeño porcentaje de los hombres ( $6\%$  en datos completos,  $7.3\%$  en datos recortados):

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Modelo de datos completos
table(
  data_lab$Sex,
  data_km2$cluster
)

# Modelo de datos recortados
table(
  data_inliers_lab$Sex,
  data_inliers_km2$cluster
)

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
```

```
#| code-overflow: wrap

factoextra::fviz_cluster(
  data_km2,
  data_std,
  labelsize = 5,
  # main = "k=2 grupos",
  geom = "point",
  ggtheme = ggplot2::theme_bw()
) +
ggplot2::ggtitle(
  label = 'Modelo con K=2 grupos',
  subtitle = 'Conjunto de datos completo (con outliers)'
) +
ggrepel::geom_text_repel(
  label = paste(
    data_lab$State,
    data_lab$Sex,
    sep = "_"),
  size = 1.5,
  colour = c("darkgreen", 'darkred')[data_lab$Sex]
)
```

```

El modelo con datos completos separa hombres y mujeres, sobreajustándose por los outliers detectados. Comete errores en la clasificación de tres observaciones de hombres, con valores anormalmente bajos de los indicadores relacionados con el alcohol.

```
```{r}
```

```

#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Modelo de datos completos

factoextra::fviz_cluster(
  data_inliers_km2,
  data_inliers_std,
  labelsize = 5,
  # main = "k=2 grupos",
  geom = "point",
  ggtheme = ggplot2::theme_bw()
) +
  ggplot2::ggtitle(
    label = 'Modelo con K=2 grupos',
    subtitle = 'Conjunto de datos recortado (sin outliers)'
) +
  ggrepel::geom_text_repel(
    label = paste(
      data_inliers_lab$State,
      data_inliers_lab$Sex,
      sep = " _"),
    size = 1.5,
    colour = c("darkred", 'darkblue')[data_inliers_lab$Sex]
)
```

```

El modelo con datos recortados separa hombres y mujeres, sin el sobreajuste impuesto por los outliers. Comete errores en la clasificación de tres observaciones de hombres, con valores anormalmente bajos de los indicadores relacionados con el alcohol.

Para ver el nivel de acuerdo de la agrupación con la clasificación, utilizamos el índice de Rand

```
```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Datos completos
rand2 <- fpc::cluster.stats(
  d = data_dist,
  alt.clustering = as.numeric(data_lab$Sex),
  clustering = as.numeric(data_km2$cluster))

rand2$corrected.rand

# Datos recortados
rand2_inliers <- fpc::cluster.stats(
  d = data_inliers_dist,
  alt.clustering = as.numeric(data_inliers_lab$Sex),
  clustering = as.numeric(data_inliers_km2$cluster))

rand2_inliers$corrected.rand
```

```

Se observan unos valores elevados del estadístico de Rand, por lo que las observaciones incluidas en los clústeres son muy similares entre sí, tanto para los modelos de datos completos como para los de datos recortados.

:::

::: {.callout-caution title="2 - Modelos con \$k=3\$" collapse="true"}

#### ##### Modelos de 3 categorías

Ambos modelos cluster (tanto el de datos completos como el de datos recortados) separan mal los tres niveles de la variable `AgeAdjustedDeathRate\_fct3`, con un importante número de discordancias entre lo esperado y lo observado:

```
```{r}
```

```
#| code-fold: true
```

```
#| info: false
```

```
#| warning: false
```

```
#| code-overflow: wrap
```

```
# Modelo de datos completos
```

```
table(
```

```
  data_lab$AgeAdjustedDeathRate_fct3,
```

```
  data_km3$cluster
```

```
)
```

```
# Modelo de datos recortados
```

```
table(
```

```
  data_inliers_lab$AgeAdjustedDeathRate_fct3,
```

```
data_inliers_km3$cluster  
)  
  
```
```

```
```{r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap
```

```
factoextra::fviz_cluster(  
  data_km3,  
  data_std,  
  labelsize = 5,  
  # main = "k=2 grupos",  
  geom = "point",  
  ggtheme = ggplot2::theme_bw()  
) +  
  ggplot2::ggtitle(  
    label = 'Modelo con K=3 grupos',  
    subtitle = 'Conjunto de datos completo (con outliers)'  
) +  
  ggrepel::geom_text_repel(  
    label = paste(  
      data_lab$State,  
      data_lab$AgeAdjustedDeathRate_fct3,  
      sep = " _"),
```

```
size = 1.5,  
colour = c("darkred", 'darkgreen', 'darkblue')[data_lab$AgeAdjustedDeathRate_fct3])  
` ` `
```

El modelo de datos completos identifica razonablemente bien a las observaciones con mortalidad alta, pero a costa de equivocarse mucho en las que tiene mortalidad media y baja..

```
` ` ` {r}  
#| code-fold: true  
#| info: false  
#| warning: false  
#| code-overflow: wrap  
  
# Modelo de datos completos  
factoextra::fviz_cluster(  
  data_inliers_km3,  
  data_inliers_std,  
  labelsize = 5,  
  geom = "point",  
  ggtheme = ggplot2::theme_bw()  
) +  
  ggplot2::ggtitle(  
    label = 'Modelo con K=3 grupos',  
    subtitle = 'Conjunto de datos recortado (sin outliers)'  
) +  
  ggrepel::geom_text_repel(  
    label = paste(
```

```

  data_inliers_lab$State,
  data_inliers_lab$AgeAdjustedDeathRate_fct3,
  sep = "_"),
  size = 1.5,
  colour = c("darkred", 'darkgreen',
'darkblue')[data_inliers_lab$AgeAdjustedDeathRate_fct3])
```

```

El modelo con datos recortados no está separando adecuadamente los niveles de la variable `AgeAdjustedDeathRate\_fct3` . La información separada en los clústeres tiene poco que ver con los niveles de esta variable categórica.

Para ver el nivel de acuerdo de la agrupación con la clasificación, utilizamos el índice de Rand

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

# Datos completos
rand3 <- fpc::cluster.stats(
  d = data_dist,
  alt.clustering = as.numeric(data_lab$AgeAdjustedDeathRate_fct3),
  clustering = as.numeric(data_km3$cluster))

rand3$corrected.rand

```

```

# Datos recortados

rand3_inliers <- fpc::cluster.stats(
  d = data_inliers_dist,
  alt.clustering = as.numeric(data_inliers_lab$AgeAdjustedDeathRate_fct3),
  clustering = as.numeric(data_inliers_km3$cluster))

rand3_inliers$corrected.rand
```

```

Se observan unos valores pobres del estadístico de Rand, por lo que las observaciones incluidas en los clústeres son muy distintas entre sí, tanto para los modelos de datos completos como para los de datos recortados. Los modelos no están separando la información contenida en la variable `AgeAdjustedDeathRate\_fct3` .

```

::: {.callout-caution title="3 - Modelo con \$k=4\$" collapse="true"}

Modelo de 4 categorías

Ambos modelos cluster (tanto el de datos completos como el de datos recortados) separan mal los tres niveles de la variable `AgeAdjustedDeathRate_fct3` , con un importante número de discordancias entre lo esperado y lo observado:

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Modelo de datos completos

```

```
table(
 data_lab$AgeAdjustedDeathRate_fct4,
 data_km4$cluster
)
```

```

```
```{r}  
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap
```

```
factoextra::fviz_cluster(
 data_km4,
 data_std,
 labelsize = 5,
 geom = "point",
 ggtheme = ggplot2::theme_bw()
) +
 ggplot2::ggtitle(
 label = 'Modelo con K=4 grupos',
 subtitle = 'Conjunto de datos completo (con outliers)'
) +
 ggrepel::geom_text_repel(
 label = paste(
 data_lab$State,
 data_lab$AgeAdjustedDeathRate_fct4,
```

```

sep = "_"),
size = 1.5,
colour = c(
 "darkred",
 'darkgreen',
 'darkblue',
 'darkviolet'
)[data_lab$AgeAdjustedDeathRate_fct4])

```

```

El modelo de datos completos para $k=4$ no identifica correctamente los niveles de la variable `AgeAdjustedDeathRate_fct4` .

Para ver el nivel de acuerdo de la agrupación con la clasificación, utilizamos el índice de Rand

```

```{r}
#| code-fold: true
#| info: false
#| warning: false
#| code-overflow: wrap

Datos completos
rand4 <- fpc::cluster.stats(
 d = data_dist,
 alt.clustering = as.numeric(data_lab$AgeAdjustedDeathRate_fct4),
 clustering = as.numeric(data_km4$cluster))

```

```
rand4$corrected.rand
```

```
...
```

Se observa un valor bajo del estadístico de Rand, por lo que las observaciones incluidas en los clústeres son muy distintas entre sí. El modelo no está separando la información contenida en la variable `AgeAdjustedDeathRate\_fct4`.

```
:::
```

```
05fk - Resumen de resultados obtenidos
```

- En ambos supuestos (datos totales y recortados), se observó una tendencia a la agrupación, tanto estadísticamente (Hopkins \$<0.5\$), por lo que se justifica realizar un análisis de agrupación.
- Para nuestro caso se utilizó un análisis cluster no jerárquico. En la fase de análisis exploratorio se detectó la presencia de \*outliers\*, por lo que se replicó el análisis con o sin los datos, para valorar la influencia de los mismos.
- Se determinó que el número óptimo de clústeres se encontraba entre 2 y 4, para el conjunto de datos completo, y entre 2 y 3, para el modelo recortado sin \*outliers\*.
- Se crearon cinco modelos de agrupación, utilizando el método \$k\$-means, 3 para los datos completos, y 2 para los datos recortados, para los valores óptimos de cluster identificados.
- Respecto a la importancia de las variables para establecer la agrupación:
  - En los modelos de 2 y 3 clústeres para datos completos, fueron las relacionadas con las características de las borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrecuencyAdults`)
  - Para el modelo con 4 clústeres, fue la de grandes bebedores (`HeavyDrinkingAdults`), seguida de las tres variables relacionadas con borracheras (`BingeDrinkingIntensityAdults`, `BingeDrinkingPrevalenceAdults`, `BingeDrinkingFrecuencyAdults`)
  - Visualmente, los clústeres de los modelos son capaces de agrupar los datos sin solapamientos. El modelo de \$k=4\$ presenta solapamientos en la representación en 2-D, pero evidencia buena capacidad discriminatoria en los modelos 3-D.

- Respecto a la evaluación de la validez de los modelos:
  - En lo que concierne a la validez interna, los modelos con mejor resultado han sido los  $k=2$  y  $k=3$  para datos recortados.
  - En lo tocante a validación externa:
    - Los dos modelos  $k=2$ , tanto para datos completos como recortados, se ajustan bastante bien a los niveles de la variable `Sex`, por lo que capturan una información similar a esta variable.
    - Los modelos  $k=3$  y  $k=4$  se relacionan mal con la variable instrumental creada discretizando los valores de la variable `AgeAdjustedDeathRate`, con lo que es razonable suponer que capturan información no contenida en esta variable.

### ### Salidas del subprocesso

Se incorporaron los resultados de los modelos a los respectivos datasets, y se crearon dos nuevos objetos con la información del cluster:

- `data\_cluster`, incorporando los modelos  $k=2$ ,  $k=3$  y  $k=4$  a los datos completos, y
- `data\_inliers\_cluster`, incorporando los modelos  $k=2$  y  $k=3$  a los datos recortados

```
```{r}
#| code-fold: true
#| output: false
#| eval: false

## Añadimos la clasificación cluster a los datos de trabajo
data_lab$cluster2 <- data_km2$cluster
data_lab$cluster3 <- data_km3$cluster
data_lab$cluster4 <- data_km4$cluster
```

```
## Añadimos la clasificación cluster a los datos de trabajo (recortados)
data_inliers_lab$cluster2 <- data_inliers_km2$cluster
data_inliers_lab$cluster3 <- data_inliers_km3$cluster

## Grabamos los nuevos dtos
saveRDS(
  data_lab,
  file = here::here('data', 'lab', 'data_cluster.rds')
)

saveRDS(
  data_inliers_lab,
  file = here::here('data', 'lab', 'data_inliers_cluster.rds')
)

````
```