PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Relacionamiento entre variables

Clase 2

La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE

En esta clase iniciaremos con el estudio de relaciones entre variables, específicamente trataremos sobre distintas funciones y algoritmos para medir la correlación y covarianza entre dos o más variables, así como técnicas para visualizar la correlación.

Un tema importante en la estadística es tratar de entender cómo se relacionan las variables entre sí, para lo cual introduciremos los temas de correlaciones, covarianza y causación. Lo anterior estará apoyado por el conocimiento del marco general lenguaje R y la interface Rstudio mismos que se usarán en los laboratorios propuestos que reforzarán el aprendizaje incorporando algunos enlaces externos. Un tema particular en este caso es la determinación de la existencia de relaciones espurias.

RDA 1: Analizar información de un conjunto de datos, mediante tablas, gráficos y medidas descriptivas.

CLASE 2: Relacionamiento entre variables

En la clase anterior revisamos algunas medidas de tendencia central, sin embargo, esto no debe entenderse como si dos o más conjuntos de datos tienen las misma métrica entonces los datasets son iguales.

Veamos un ejemplo con 3 conjuntos de datos :

Tabla 1
Distintos conjuntos hipotéticos de datos
Creación del autor Alfonso Prado

Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3
0,5,10	4,5,6	5,5,5

Los tres tienen una media de cinco, pero ahí termina toda su similitud, sin ver las observaciones, se puede concluir erróneamente que los conjuntos son iguales. Una imprecisión más notoria de los conjuntos de datos resultaría si se compara el grado en el cual se dispersaron las observaciones individuales en cada conjunto de datos o se expandieron alrededor de la media cinco. Las observaciones en el primer conjunto de datos están muy dispersas por encima y por debajo de la media, mientras que aquellas del segundo grupo de datos están comparativamente cerca de ésta. El primer conjunto de datos tiene una medida de dispersión mayor que la segunda. El tercer conjunto de datos no tiene dispersión, todas las observaciones son iguales a la media. Sabiendo esto, sería poco probable asumir cualquier similitud. En este sentido, las medidas de dispersión son mucho más útiles e informativas.

Relacionamiento entre variables

Varianza , Covarianza y Desviación estándar y Correlación:

Varianza: Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Una serie de datos podría tomar un número infinito de valores, pero en la práctica cuando se analiza un parámetro de interés tienden a estar dentro de un rango más o menos dispersos. La varianza indica la medida de esta dispersión.

La varianza puede tomar cualquier valor real.

Sintaxis de las funciones
Describe el uso de los agumento
Enlace

Figura 1
Fórmula de la Varianza y función programática Prado A. (2025)
Adaptación de la fórmula para la resolución de varianza según Webster (2000) capitulo 3

$Var(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}$	`var(x, y = NULL, na.rm = FALSE, use)`

Desviación estándar: Representa la raíz cuadrada de la varianza. El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que "s" se utiliza para representar la desviación estándar de una muestra. Otras diferencias más notorias las veremos más adelante cuando estudiemos las distribuciones.

Figura 2
Fórmula de la desviación estándar y su funcion programática Prado A. (2025)
Adaptación de la fórmula para la resolución de la desviación estandar según Webster (2000) capitulo 3

$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	sd(x, na.rm=FALSE)

Covarianza: Es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias y ayuda a entender si estas dos variables están relacionadas de alguna forma.

Puede ser positiva (si se mueven juntas), negativa (se mueven en direcciones opuestas), o nula (no se mueven juntas de manera apreciable). Existen varios métodos (algoritmos) para su cálculo. La covarianza se utiliza para comprender cómo se relacionan las variables en un conjunto de datos.

Figura 3
Fórmula de la covarianza y su función programática Prado A. (2025)
Adaptación de la fórmula para la resolución de la covarianza según Webster (2000) capitulo 3

$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$	`cov(x, y = NULL, use = "everything",` `  method = c("pearson", "kendall", "spearman"))`

Correlación: De todas las variables contenidas en un dataset vamos a encontrar que algunas de ellas podrían tener una relación tipo causa-efecto, es decir, cualquier cambio en una variable denominada "predictora" causará un aumento o disminución de la segunda variable denominada variable de respuesta. Este fenómeno, que conocido como correlación, es central en el diseño de modelos estadísticos.

Figura 4
Fórmula de la correlación y su función programática Prado(2025)
Adaptación de la fórmula de https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

r = [nΣ(xy) - ΣxΣy] / sqrt([nΣx^2 - (Σx)^2][nΣy^2 - (Σy)^2])
`cor(x, y = NULL, use = "everything",` `  method = c("pearson", "kendall", "spearman"))`

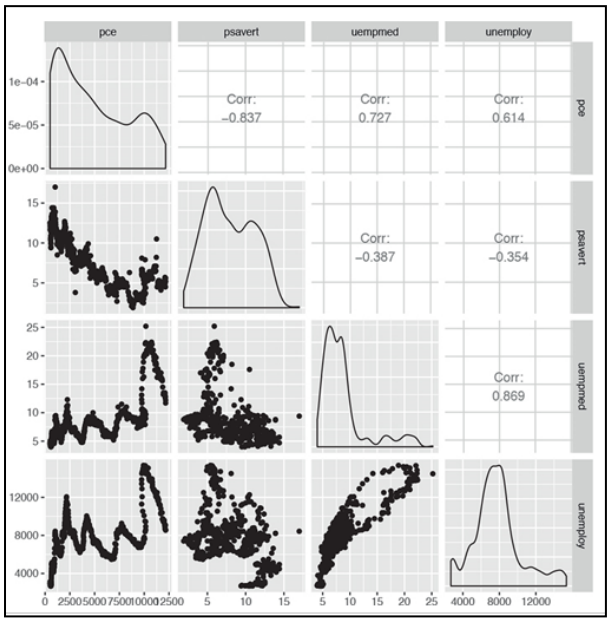
Con frecuencia cuando tratamos de analizar estas relaciones tenemos una cantidad de variables que podrían actuar como variables predictoras y calcular las correlación de cada una puede ser una tarea ardua. La función programática nos ayuda a automatizar dando como resultado una matriz conocida como la matriz de correlación del dataset cuyas característica son :

- ✓ Es una matriz cuadrada nXn donde n representa la cantidad de variables el dataset
- ✓ La diagonal representa la correlación consigo mismo
- ✓ El gráfico es simétrico, la correlación de A y B es igual a la correlación entre B y A

Pueden existir relaciones fuertes positivas y negativas, así como algunos casos no se puee encontrar un patrón, entonces diremos que la correlación es nula.

El problema es que cuando tenemos un dataset extenso por ejemplo, de 100 variables, el análisis de una matriz de 10K valores puede ser extenuante. En estos casos se estila verlo en forma gráfica como se muestra en la siguiente figura.

Figura 5
Representación gráfica de la matriz de correlación.
Fuente: https://cran.r-project.org/



Causación

De acuerdo a Clever Tap, las medidas de correlación y covarianza nos indican como cambia una variable cuando la otra variable cambia. Es decir, solo nos indican como se mueven las variables en el dataset, pero no necesariamente nos indica que una variable causa a la otra variable.

Ejemplo: Usted dispone de un dataset que contiene población de una ciudad, el consumo de carne y el consumo de pescado. Existe una correlación directa entre las 2 últimas variables y la población. Esto es correcto. Pero si se efectúa una correlación entre consumo de pescado y consumo de carne también mostrará correlación directa. ¿Significa esto que el consumo de carne hace que se consuma más pescado? Por supuesto que no. En este caso existe correlación, pero no existe causación.

Por el momento diremos que la causación implica correlación, pero la correlación no implica causación. La pregunta que nos hacemos es como demostramos causación? La respuesta es el diseño de experimentos.

El propósito del diseño de experimentos (DoE) es poder obtener en forma segura la relación causa-efecto entre dos variables. Por ejemplo: saber si un cierto medicamento ayudó (o no) en la recuperación del paciente. Para aseverar que lo anterior es verdadero ciertas consideraciones deben cumplirse.

Asociación

El primer criterio para establecer un efecto causal es una asociación observada entre las variable predictora y de respuesta. Generalmente esto se logra mediante un análisis de correlación. Sijn embaargo, debemos garantizar que existe la relación causa-efecto mediante la validación del ordenamiento en el tiempo. Se debe asegurar que la variación de la variable predictora se produjo antes del tiempo de la variación de la variable de respuesta.

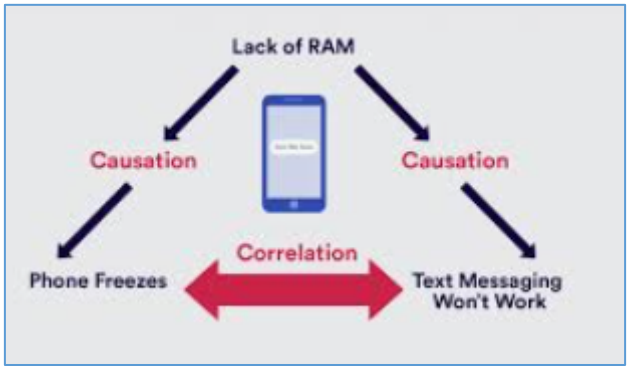
Detección de asociaciones espurias

El segundo criterio es que esta relación no sea espuria, definimos como espuria a una relación que es ficticia o fraudulenta, muchas veces nos topamos con la existencia de una tercera variable que en realidad afecta a las dos primeras. Este tipo de variables se las conoce como “cofounding”.
Variables cofounding o confusoras
Describe el efecto de dichas variables

Enlace

La siguiente figura ilustra la relación entre la falta de memoria en un teléfono celular y sus efectos, en este caso la variable “Lack of RAM” es del tipo cofounding.

Figura 6
Ejemplo de Correlación vs Causación en un teléfono celular
Fuente: (https://clevertap.com/blog)



Para garantizar lo mencionado debemos cumplir las siguientes condiciones:

- ✓ Dos grupos de comparación (en el caso más simple, un grupo experimental y un grupo de control), para establecer asociación
- ✓ Variación en la variable independiente antes del cambio en la variable dependiente, para establecer el orden temporal
- ✓ Asignación aleatoria a los dos (o más) grupos de comparación, para establecer que la relación no es espuria.

Representaciones gráficas de variables de variables cuantitativas, simples y agrupadas

Ggplot provee dos tipos de gráficos de barras: geom_bar () y geom_col (). geom_bar () hace que la altura de la barra sea proporcional al número de casos de cada valor en cada grupo. Si desea que las alturas de las barras representen valores en los datos, use geom_col() en su lugar. Recordando la relacion entre el geom y el stat , podemos decir que geom_bar() usa stat_count() por defecto: cuenta el número de observaciones en cada posición x. Por otro lado, geom_col() usa stat_identity(): deja los datos como están. Veamos un ejemplo: trabajaremos con el dataset midwest que nos provee información demográfica sobre 437 condados en los USA.

Tabla 2
Descripcion el dataset midwest
Creación del autor Alfonso Prado

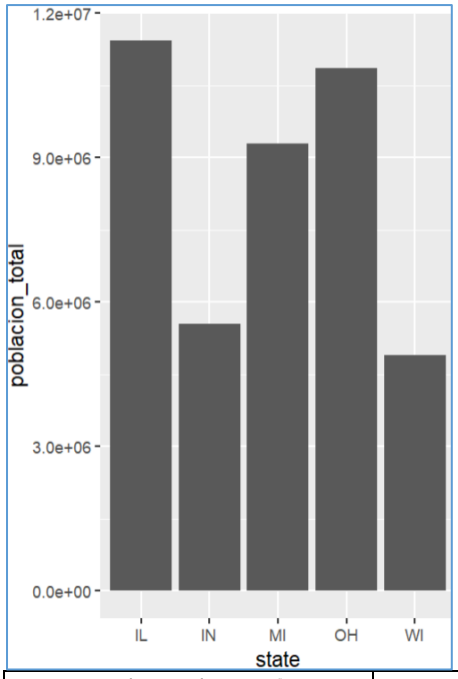
PID	county	state	area	poptotal	popdensity	popwhite	popblack	popamerindian	popasian	popother
<int>	<chr>	<chr>	<dbl>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
561	ADAMS	IL	0.052	66090	1271.	63917	1702	98	249	124
562	ALEXANDER	IL	0.014	10626	759	7054	3496	19	48	9
563	BOND	IL	0.022	14991	681.	14477	429	35	16	34
564	BOONE	IL	0.017	30806	1812.	29344	127	46	150	1139
565	BROWN	IL	0.018	5836	324.	5264	547	14	5	6
566	BUREAU	IL	0.05	35688	714.	35157	50	65	195	221
567	CALHOUN	IL	0.017	5322	313.	5298	1	8	15	0

Nuestro primer gráfico consiste en visualizar la población de cada estado, mediante el siguiente código: (Note que, dado que la población está distribuida por condados hemos tenido que hacer un sumarización previa. Para entender tanto las funciones omo los operadores referirse al recurso Introducción a R parte 2.r

```
mid1 <- midwest %>%
  group_by(state) %>%
  dplyr::summarise(poblacion_total = sum(poptotal, na.rm=TRUE))
gg<- ggplot(data=mid1 , aes( x=state, y=poblacion_total) )+geom_col()
gg
```

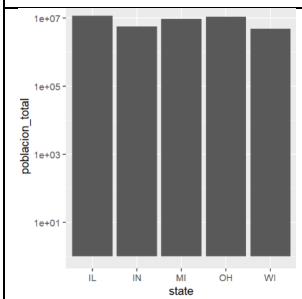
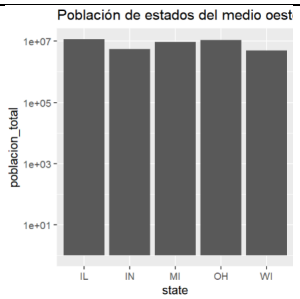
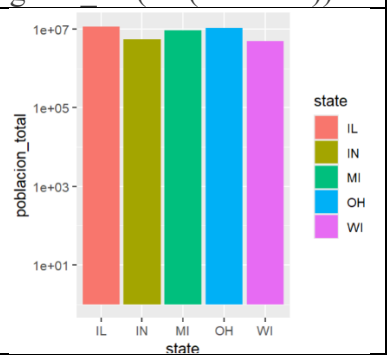
Y su resultado es el siguiente

Figura 7
Resultado figura básica
Creación del autor Alfonso Prado



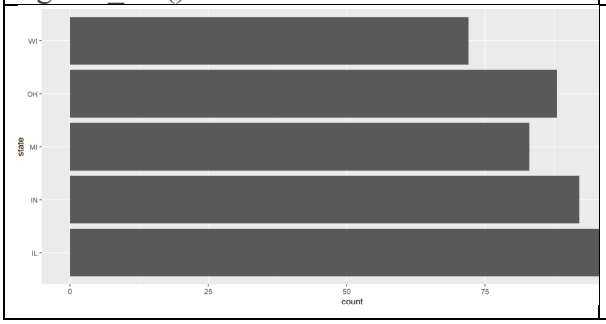
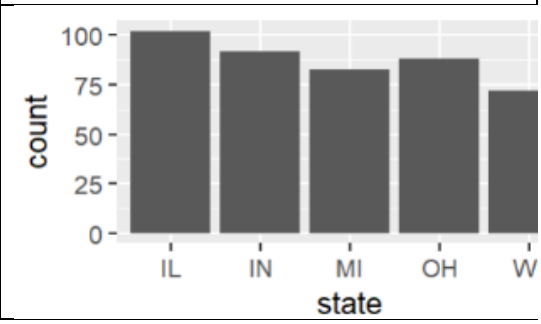
El gráfico anterior contiene algunas deficiencias: Podemos notar que existen diferencias en población pero el gráfico parece mostrar que la diferencia es mayor. En este caso lo que podemos aplicar una transformación logarítmica al eje de las Y . En segundo lugar el gráfico carece de títulos, todo trabajo profesional debe estar acompañado de títulos (inclusive trabajo de clase), En tercer lugar el grafico carece de color, existe varias formas de crear y asignar colores , por el momento podemos utilizar la función aes() para mapear cualquier variable del data para sea asignada a un color distinto

Figura 8
Personalización de la figura 7
Creación del autor Alfonso Prado

gg+ scale_y_log10()	gg + ggtitle("Población de estados del medio oeste")	Gg + geom_col(aes(fill=state))
		

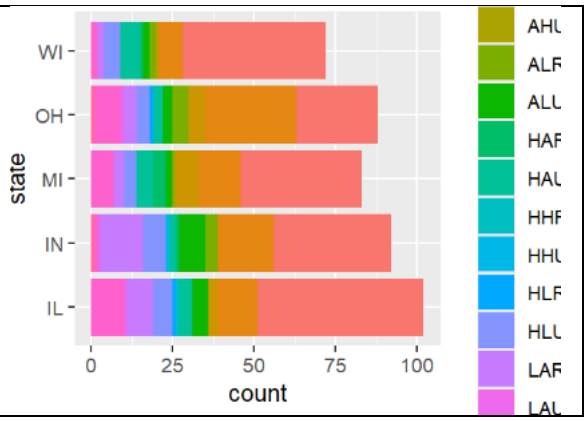
Por otro lado, ggplot nos provee la función que presenta la cantidad de ocurrencias del dataset, por ejemplo, del dataset midwest es posible que nos interese saber cuál es la cantidad de condados por estado. En este caso usar las función geom_col no es adecuada, ya que esta presentaría un valor único de la variable (poptotal). Geom_bar por otro lado esta asociado al stat=count, por lo que sería la adecuada. Note que en este primer gráfico el sistema a seleccionado barras horizontales, pero eso fácilmentente cambiable una transformacion de coordenadas

Figura 9
Personalización del grafico 7
Creación del autor Alfonso Prado

mid3 <- midwest %>% group_by(state) gg <- ggplot(data=mid3, aes(y=state)) +geom_bar()	gg + coord_flip()
	

Barras Apiladas: Eventualmente tenemos otras variables que nos ayudan a desglosar valores y queremos que el gráfico nos ayude a visualizar. Esto lo podemos lograr otra vez mediante un mapeo de terceras variables, por ejemplo en nuestro dataset midwest podemos ver como la variable categoria esta repartida entre los distintos condados de los distintos estados.

Figura 10
Uso del argumento fill para desglosar terceras variables
Creación del autor Alfonso Prado

ggplot(data=midwest, aes(y=state, fill=category)) + geom_bar()	

Diagramas de RADAR

Cuando queremos visualizar múltiples variables numéricas de la misma observación al mismo tiempo, los diagramas de barras apiladas o agrupadas se empiezan a complicar y se vuelven difíciles de leer. En estos casos podemos recurrir a un diagrama de araña o radar. Por ejemplo, usando el mismo dataset anterior, notamos que la población está desglosado por etnicidad, sin embargo visualizar todos y cada uno de los componente étnicos produciría una cantidad de barras apiladas que sería dificil de leer.

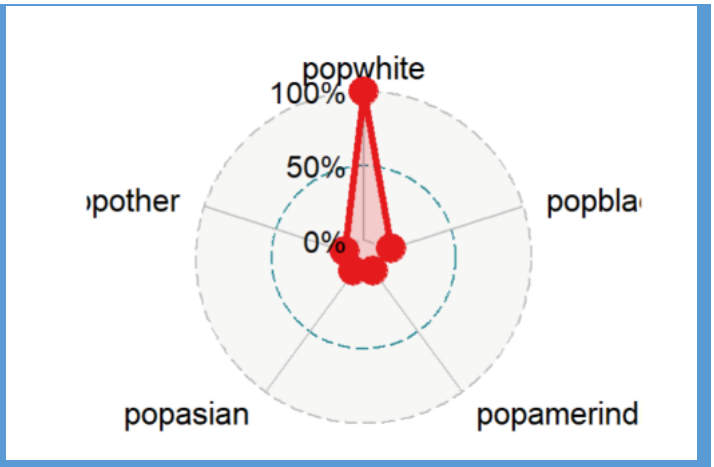
Para esto necesitamos cargar el paquete ggradar que no se encuentra en CRAN por lo que tenemos que bajarlo de github (ricardo-bion/ggradar). Dependiendo de la versión de R es posible que requiera bajar y compilar programas fuente en cuyo caso requiere el paquete adicionalmente del paquete devtools.

Los diagramas de radar requiren una cierta preparacion inicial, aunque, en este ejemplo, el resultado parece mostrar que los 5 estados tienen características etnicas similares como se muestra en la siguiente figura

Figura 11
Diagrama de radar del dataset midwest
Creación del autor Alfonso Prado

```
data(midwest)
mids <- midwest %>%
select(popwhite, popblack
,popamerindian,
popasian,popother) %>%
  as_tibble(rownames =
"group") %>%
  mutate_at(vars(-group),
rescale)
mid_radar <- mids[2:5 , ]

ggradar(mid_radar,
fill=TRUE, fill.alpha=0.2)
```



Tome en cuenta que los diagramas de araña funcionan bien con una pequeña cantidad de variables. Cuando éstas empiezan a incrementarse , el gráfico resultará muy congestionado y será dificil de leer.

Mejores Prácticas de Diagramas de Barras:

Un problema con las barras verticales es que si los nombres de las columnas son largos las etiquetas tienden a sobre escribirse dando un feo aspecto. Es posible solventar parcialmente el problema a través de incluir temas que cambian el ángulo de escritura de las etiquetas, pero igual se ve muy mal. En ese caso la mejor práctica es usar barras horizontales

Figura 12
Ángulo de escritura de las etiquetas para mejorar legibilidad
Creación del autor Alfonso Prado

```
geom_col(aes(fill=drv), position="dodge") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

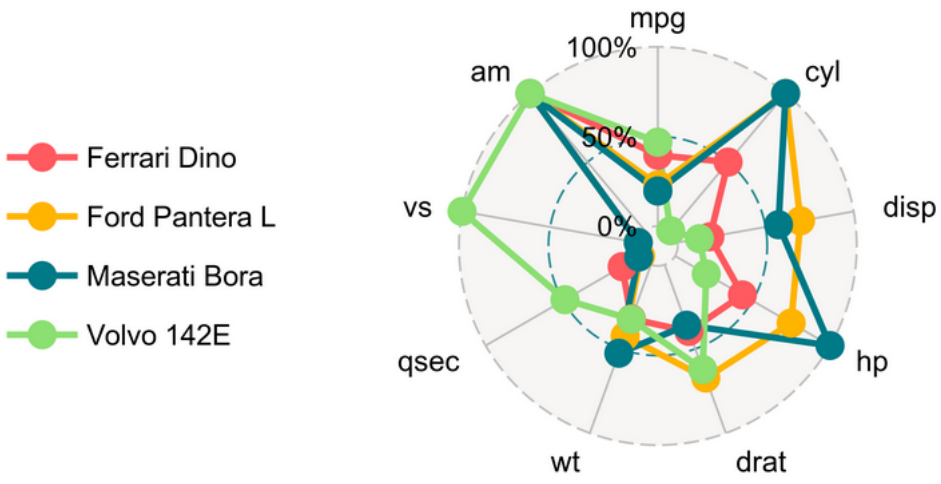
a. Independientemente de si colocamos las barras en vertical u horizontal, debemos prestar atención al orden en el que están dispuestas. Las barras deben estar dispuestas de mayor a menor (principio de pregnancia Gestalt).

b. Sin embargo, si la variable categórica representa edades o fechas esto no puede hacerse porque debe mantenerse el orden natural de la variable categórica. El orden las barras está relacionado con el ordenamiento del factor. Si el factor fue creado en forma desordenada ggplot no puede corregir el orden de salida para lo cual debemos recurrir a la librería "forcats". Las siguientes funciones podrían ser útiles en diferentes casos in_freq() , lump(), reorder(), relevel().

c. Tener en cuenta que existe un límite en la cantidad de observaciones que se puede visualizar, el gráfico no puede estar congestionado. Esto es especialmente crítico con diagramas de radar.

d. En los diagramas de calor, la selección de la escala y el ordenamiento son críticos debe asociarse a escalas continuas.

Figura 13
Diagrama de radar sobre el dataset mtcars, 4 observaciones 10 variables
Creación del autor Alfonso Prado



REFERENCIAS BIBLIOGRÁFICAS

CleverTap,(2018),Correlation vs Causation: Definition, Differences, (https://clevertap.com/blog/correlation-vs-causation/)

The Comprehensive R Archive Network, (https://cran.r-project.org/)

PUCE Virtual

La excelencia no se improvisa

síguenos