INTRODUCCIÓN AL ANÁLISIS DE DATOS

M.Sc. Henry López

"LOS DATOS SON EL NUEVO PETRÓLEO. ES VALIOSO, PERO SI NO ESTÁ REFINADO, REALMENTE NO SE PUEDE USAR. TIENE QUE CAMBIARSE A GAS, PLÁSTICO, PRODUCTOS QUÍMICOS, ETC. PARA CREAR UNA ENTIDAD VALIOSA QUE IMPULSE LA ACTIVIDAD RENTABLE; ENTONCES LOS DATOS DEBEN DESGLOSARSE, ANALIZARSE PARA QUE TENGAN VALOR ".

– Clave humby 2006 & Michael Palmer

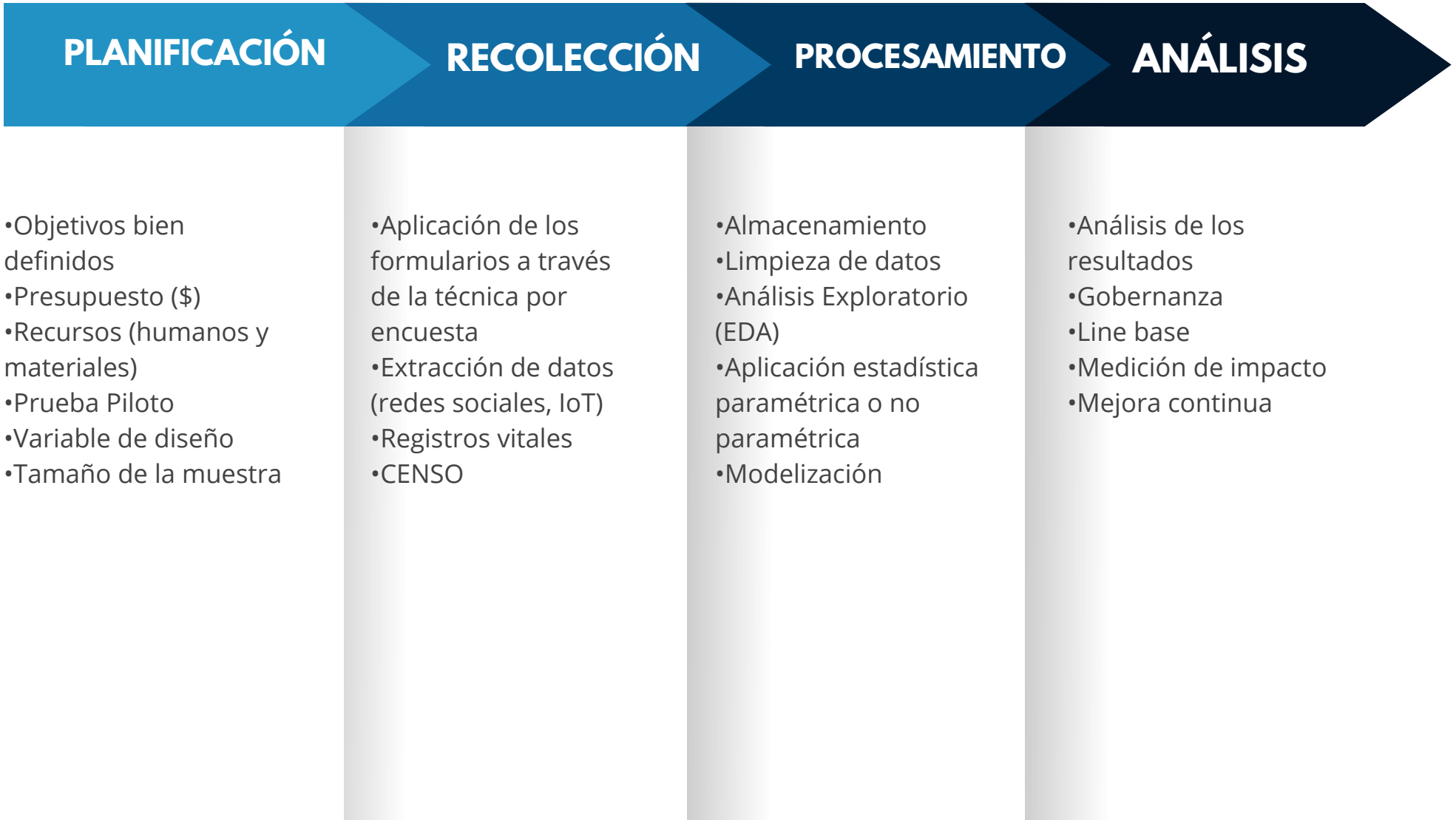
"SEGÚN TUKEY, EL ANÁLISIS EXPLORATORIO DE DATOS (EDA) ES UN PROCESO DE EXAMINAR Y COMPRENDER LOS DATOS UTILIZANDO ESTADÍSTICA, GRÁFICOS Y OTRAS TÉCNICAS PARA EXPLORAR PATRONES, DESCUBRIR RELACIONES Y SACAR CONCLUSIONES ÚTILES PARA LA INCORPORACIÓN POSTERIOR EN LA MODELIZACIÓN Y LA TOMA DE DECISIONES. ".

– Tukey (1977)

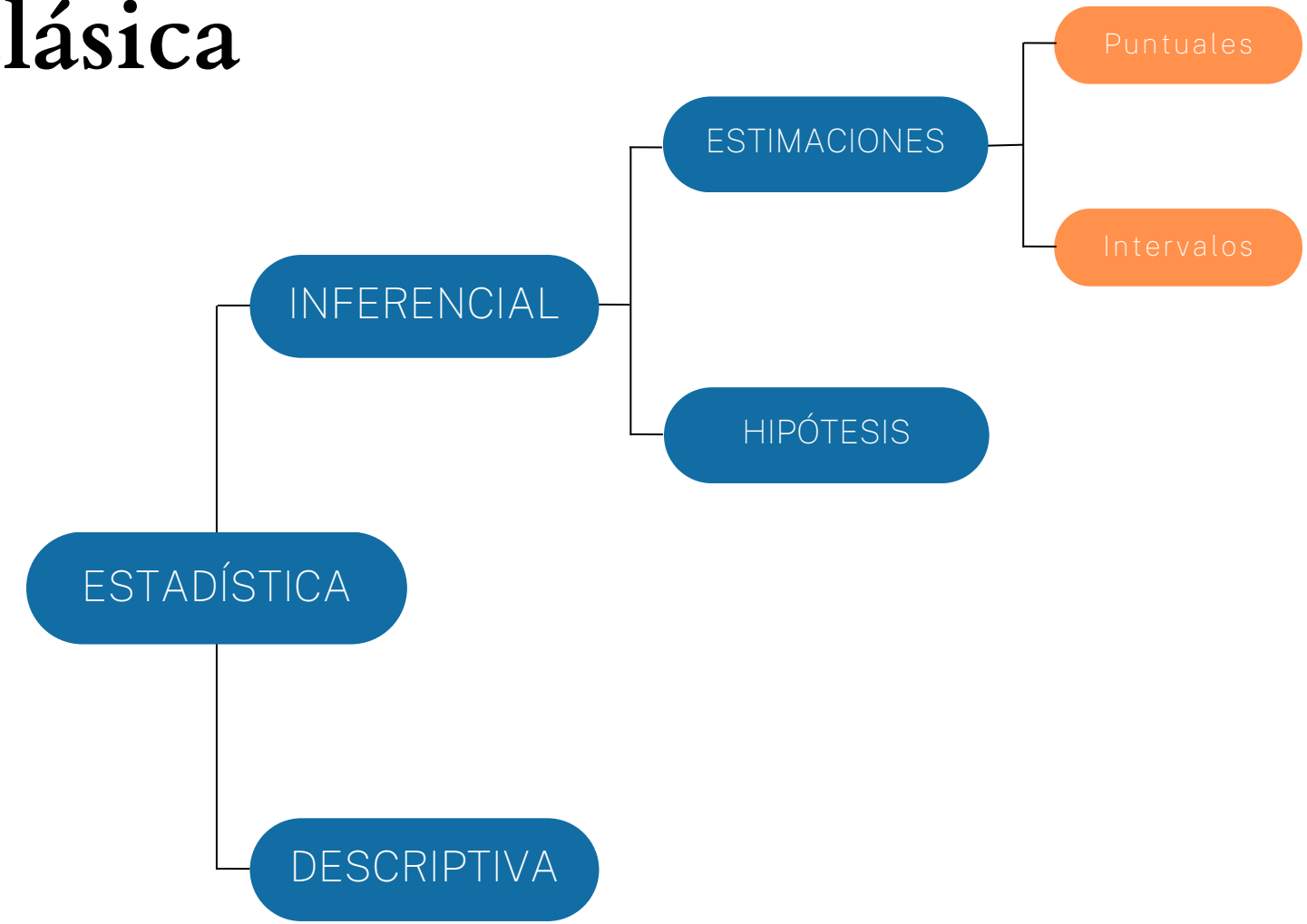
Competencia

Aplica la estadística descriptiva para la toma decisión, utilizando de manera responsable las herramientas tecnológicas para el análisis de datos.

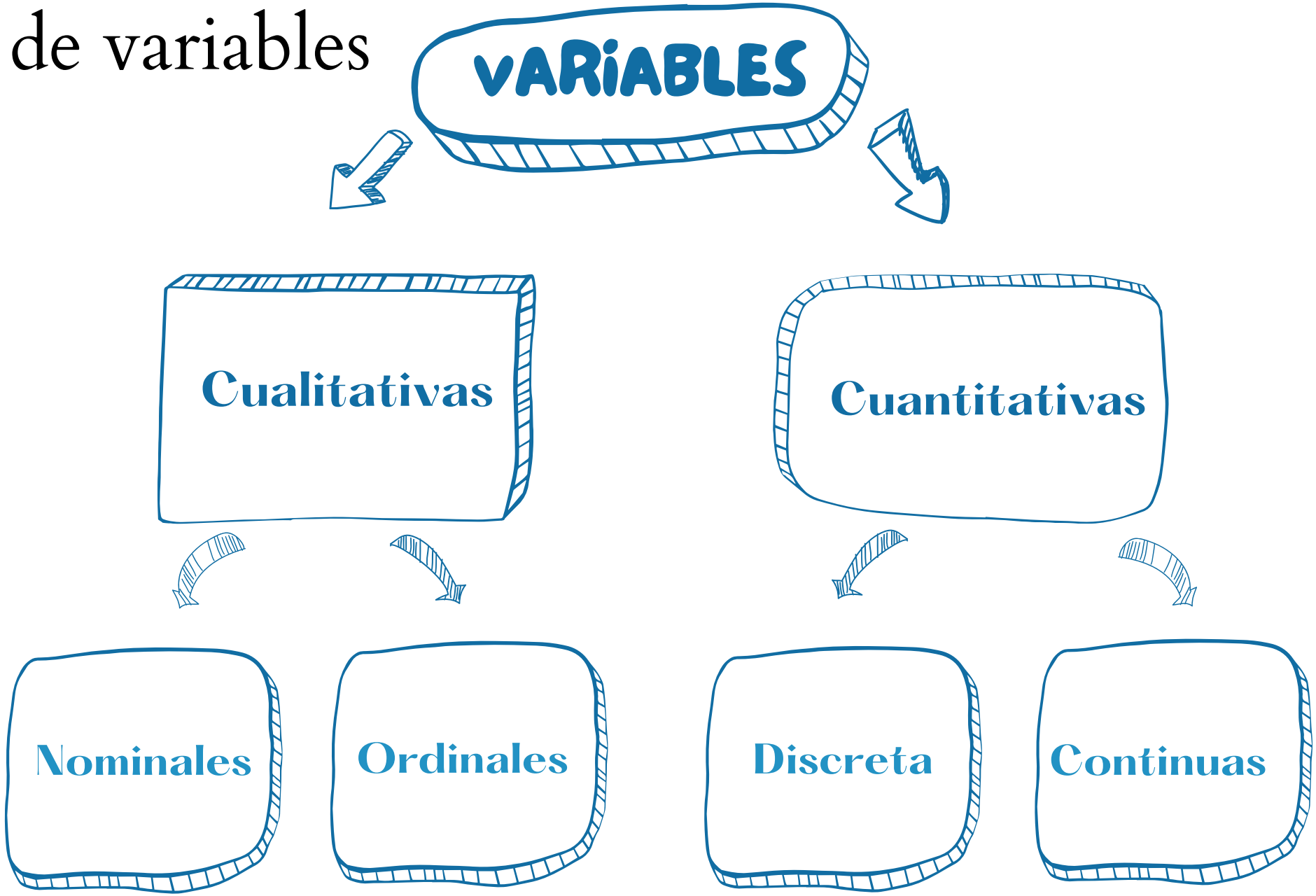
Estadística



División clásica

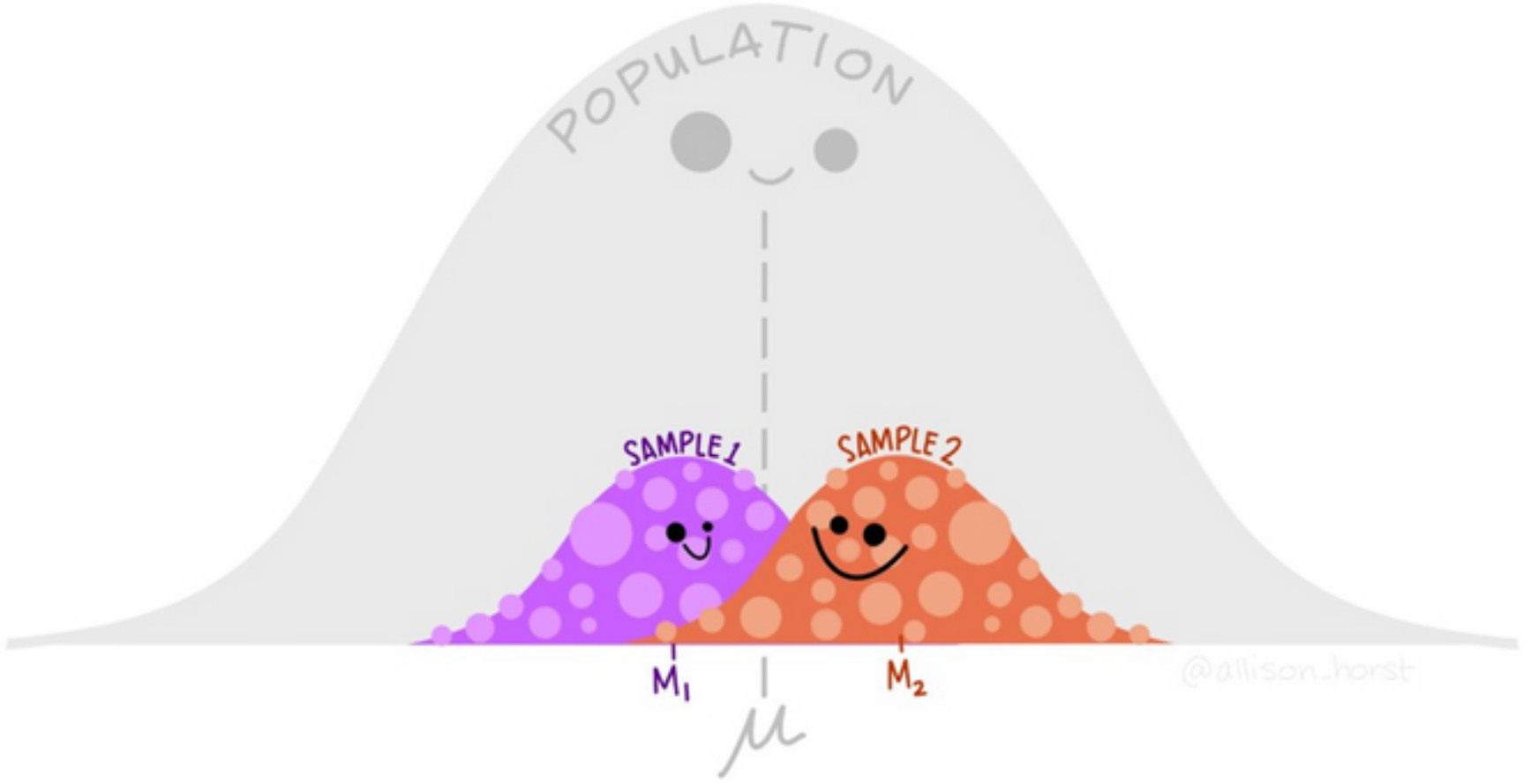


Tipos de variables

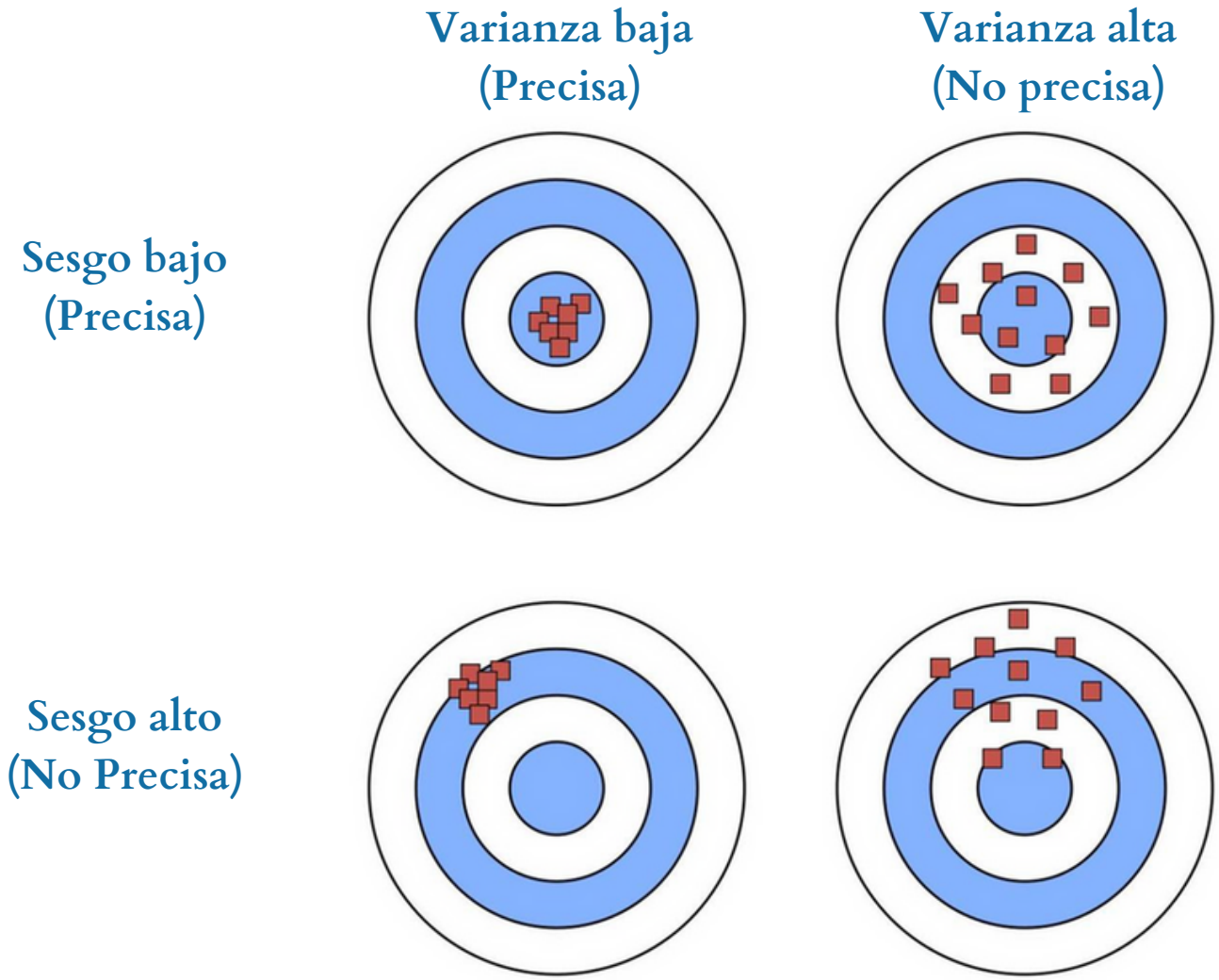


Tipos de variables

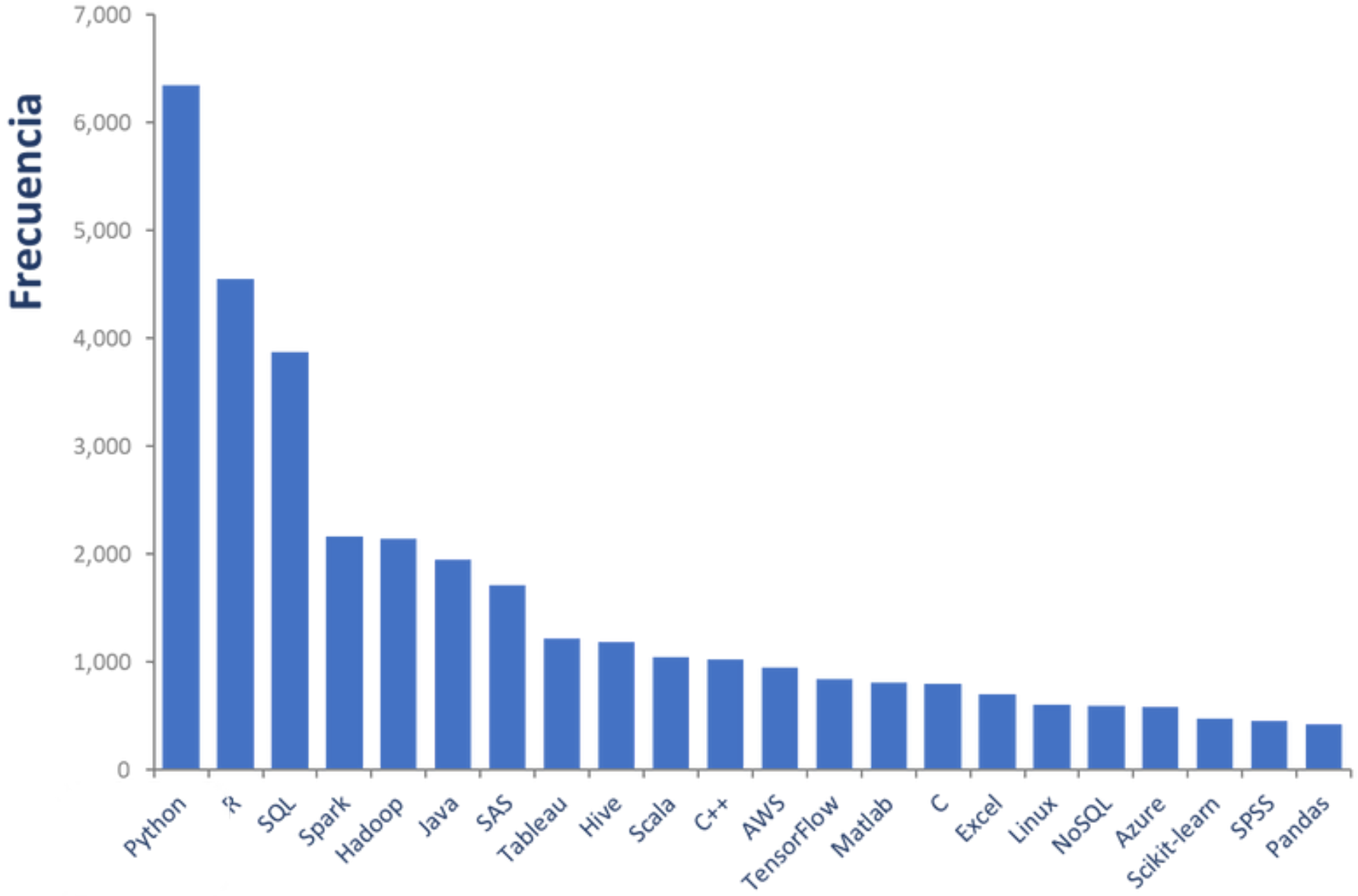
Población y muestra



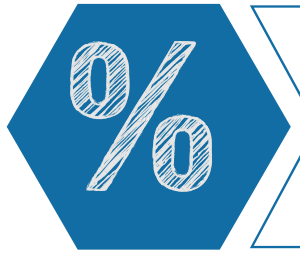
Descripción de datos



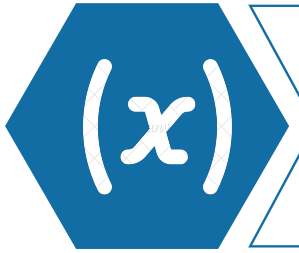
Software más usados



Métodos estadísticos



Tabulación



Descripción de datos

Métodos gráficos

Tabulación de datos categóricos

Tabla 1

Unidades agrícolas por niveles de bienestar

Pobreza	Año	Tamaño de la unid Agricola (manzanas)					
		<2	2 to 5	5 to 20	20 to 50	>50	Total
Pobre	2001	16.7	22.1	36.4	24.7	0	100
	2011	43.3	32.7	21.2	2.9	0	100
No pobre	2001	0	0	0	0	100	100
	2011	0	0	30.1	30.4	39.5	100
Total	2001	12.3	16.3	26.9	18.2	26.2	100
	2011	25.4	19.2	24.9	14.3	16.3	100

Source: Castro-Leal and Laguna (2015) using CENAGRO 2001 and 2011

Tabulación de datos categóricos

Tabla 2

Distribución de la población según nivel de alfabetismo por macro región

Macro Región	EMNV 2009			Total	EMNV 2014			Total
	Lee y escribe	Solo sabe leer	No sabe ni leer ni escribir		Lee y escribe	Solo sabe leer	No sabe ni leer ni escribir	
- Managua	57.8	0.5	4.1	62.4	36.9	0.5	2.4	39.7
- Pacífico	14.1	0.2	1.9	16.3	20.0	0.3	1.8	22.1
- Central	8.9	0.1	2.3	11.4	20.0	0.3	2.6	22.9
- Atlántico	7.2	0.3	2.5	10.0	12.7	0.4	2.2	15.3
Total	88.0	1.1	10.9	100.0	89.6	1.4	9.1	100.0

Nota: Se refleja el porcentaje de la población por nivel de alfabetismo según macro región.

Datos abiertos. (INIDE- EMNV´s 2011,2016).

Descripción de datos

Muestra fragmento de código

```
   dat <- read_excel("henry/PredictingToyota.xlsx")
   library(flextable)

   descr<- summarizor(dat[,c(1:4)], by = "FuelType",
overall_label = NULL)
   descr

   as_flextable( descr)
```

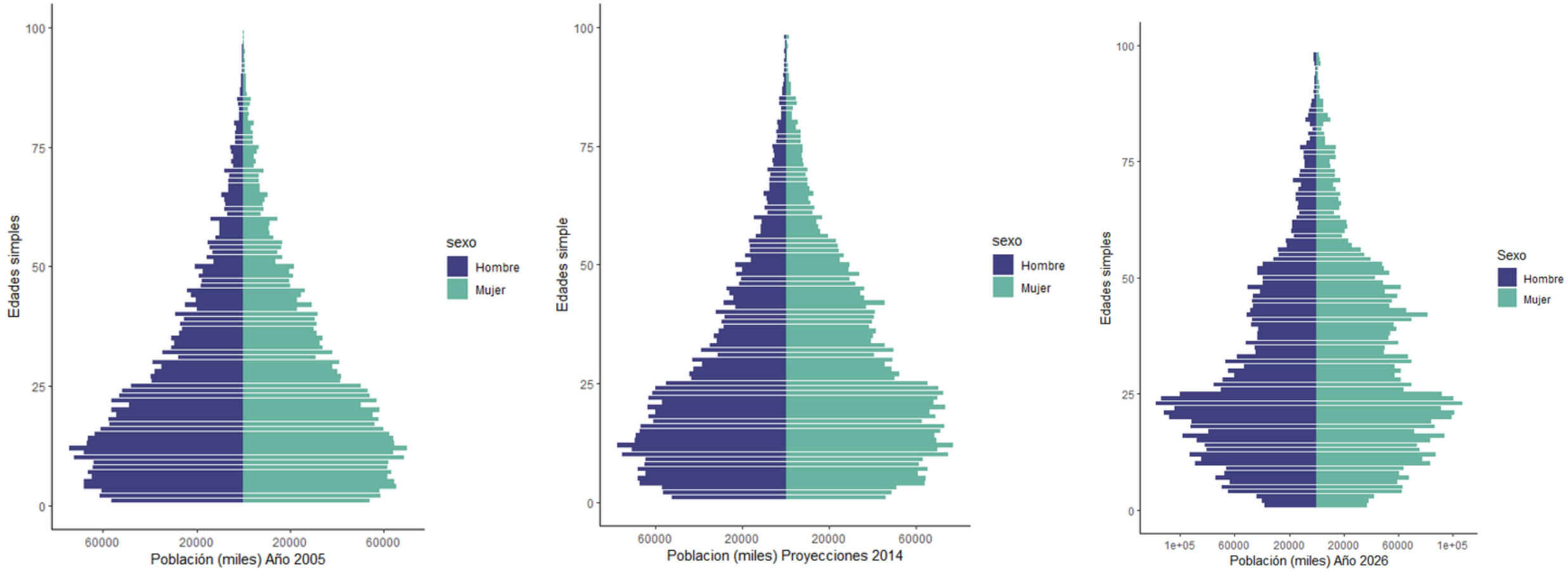
Tabla 3.

Descripción precio, edad y km recorridos vehículos Toyota

		CNG (N=17)	Diesel (N=155)	Petrol (N=1,264)	overall (N=1,436)
Price	Mean (SD)	9421.2 (2492.1)	11294.6 (5535.9)	10679.3 (3326.6)	10730.8 (3627.0)
	Median (IQR)	8950.0 (3550.0)	8950.0 (5750.0)	9940.0 (3400.0)	9900.0 (3500.0)
	Range	5250.0 - 14950.0	4350.0 - 32500.0	5250.0 - 24500.0	4350.0 - 32500.0
	Missing	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Age	Mean (SD)	56.4 (13.3)	50.7 (20.6)	56.6 (18.3)	55.9 (18.6)
	Median (IQR)	58.0 (20.0)	55.0 (33.0)	61.0 (26.2)	61.0 (26.0)
	Range	37.0 - 80.0	4.0 - 80.0	1.0 - 80.0	1.0 - 80.0
	Missing	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
KM	Mean (SD)	117865.6 (45070.2)	111977.6 (54473.0)	62542.3 (30173.8)	68533.3 (37506.4)
	Median (IQR)	115191.0 (61257.0)	117000.0 (77522.5)	60716.0 (40290.2)	63389.5 (44020.8)
	Range	41499.0 - 207114.0	1.0 - 243000.0	1.0 - 194545.0	1.0 - 243000.0
	Missing	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)

Visualización

Figura 1. Distribución de la población por sexo y edades simples

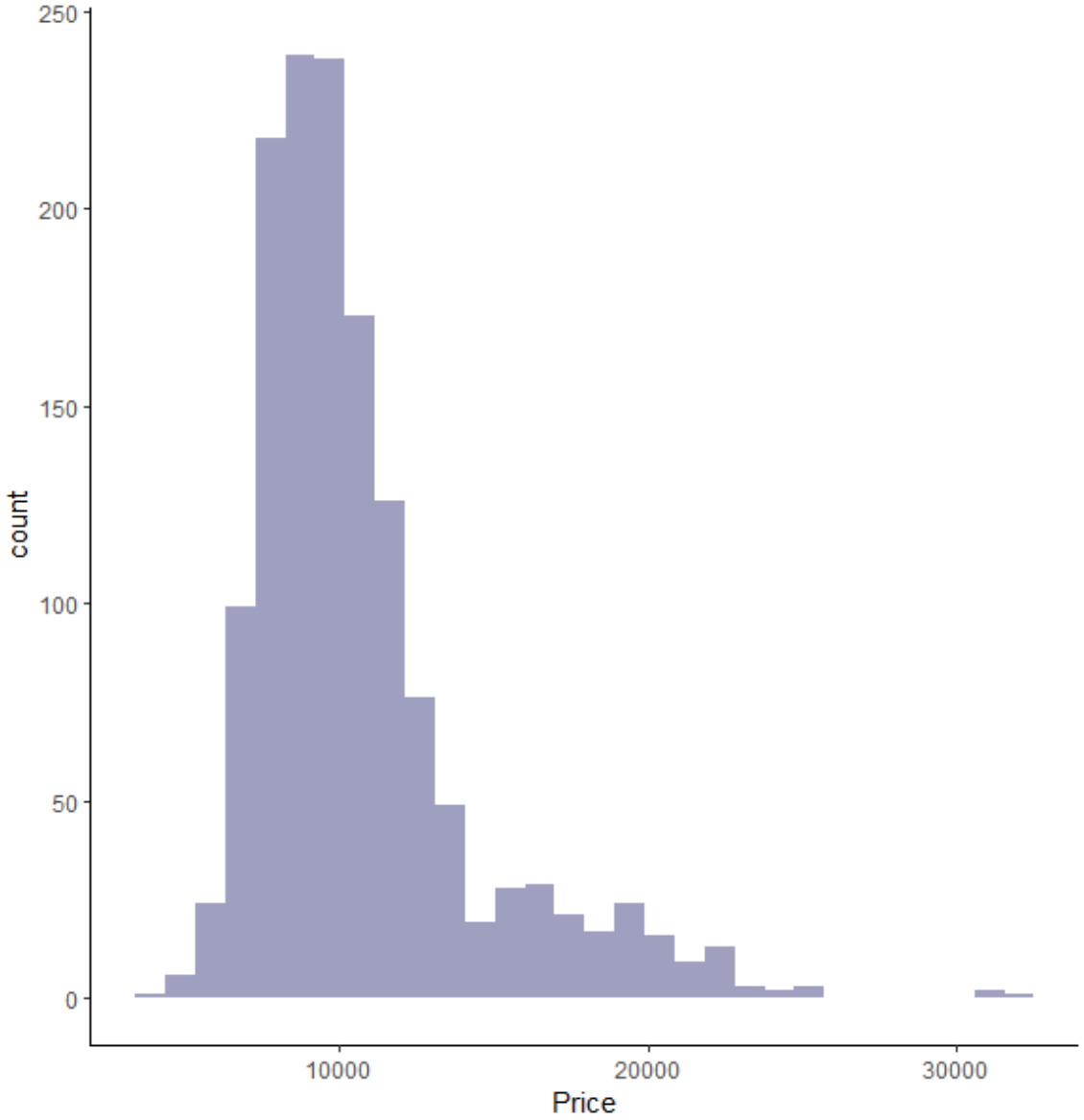


Visualización

Muestra fragmento de código

```
dat <- read_excel("henry/PredictingToyota.xlsx")

ggplot(dat, aes(x = Price))+
geom_histogram(alpha = 0.5, fill="#69b3a2")+
theme_classic
```

Figura 2. Precio de los vehículos Toyota

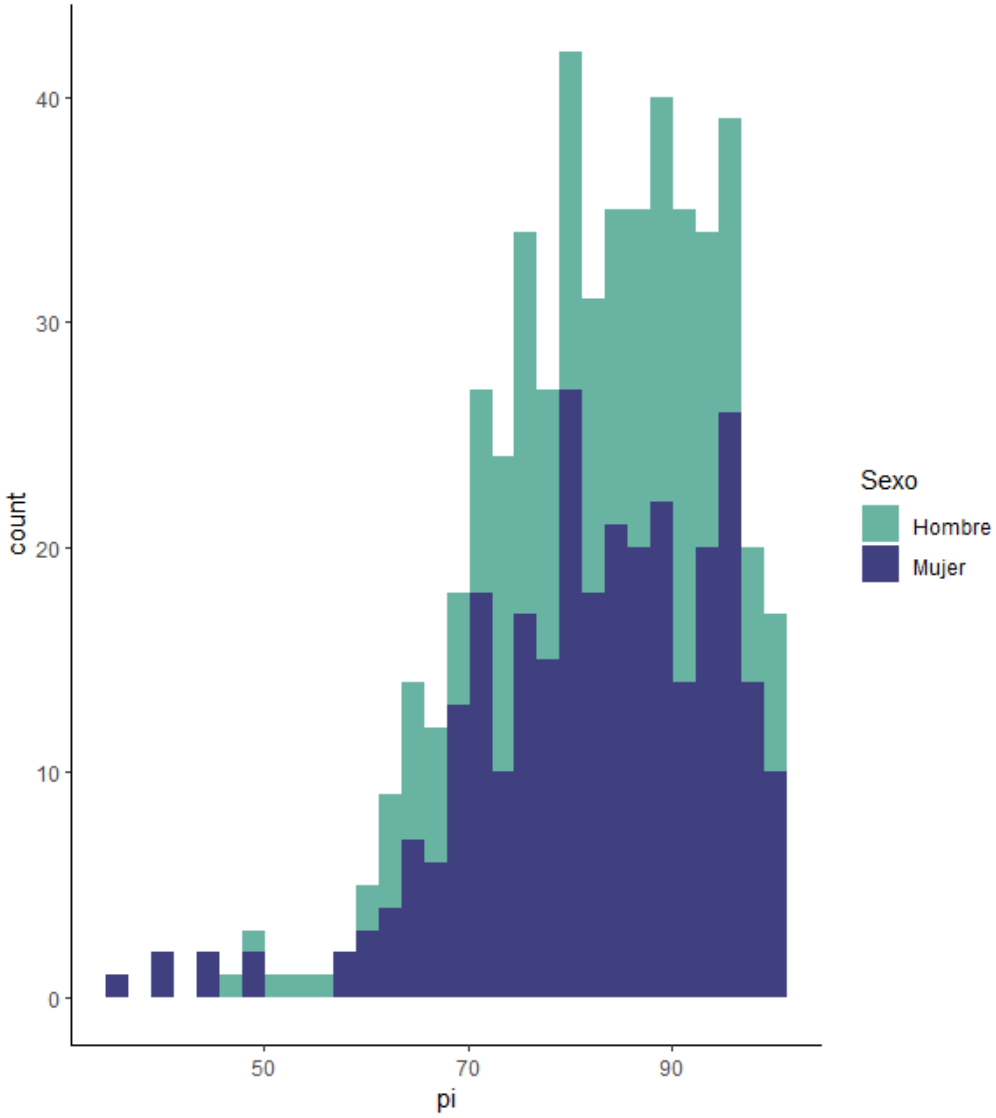


Visualización

Muestra fragmento de código

```
pi<- ((rowSums(selec22[,c(117:140)]))/96)*100

selec22$pi<- pi

hist(selec22$pi)

library(ggplot2)

sub <- as.factor(selec22$Subsistema)
selec22$sub <- sub

ggplot(selec22, aes(x = pi, fill = Sexo)) +
  geom_histogram()+
  scale_fill_manual(values=c("#69b3a2",
"#404080"))
```

Figura 3. Puntuaciones sobre percepción



Visualización

Muestra fragmento de código

```
dat <- read_excel("henry/PredictingToyota.xlsx")

 ggplot(dat, aes(x = Price))+
    geom_boxplot(alpha = 0.5,
fill="#404080",orientation="y")+
    theme_classic()
```

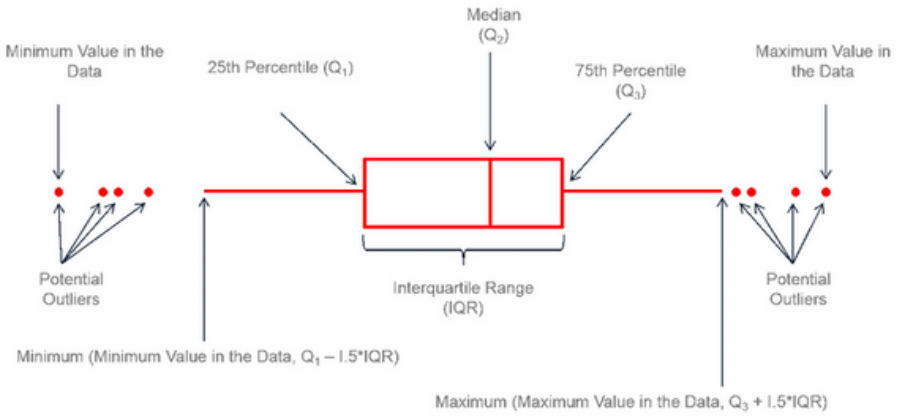
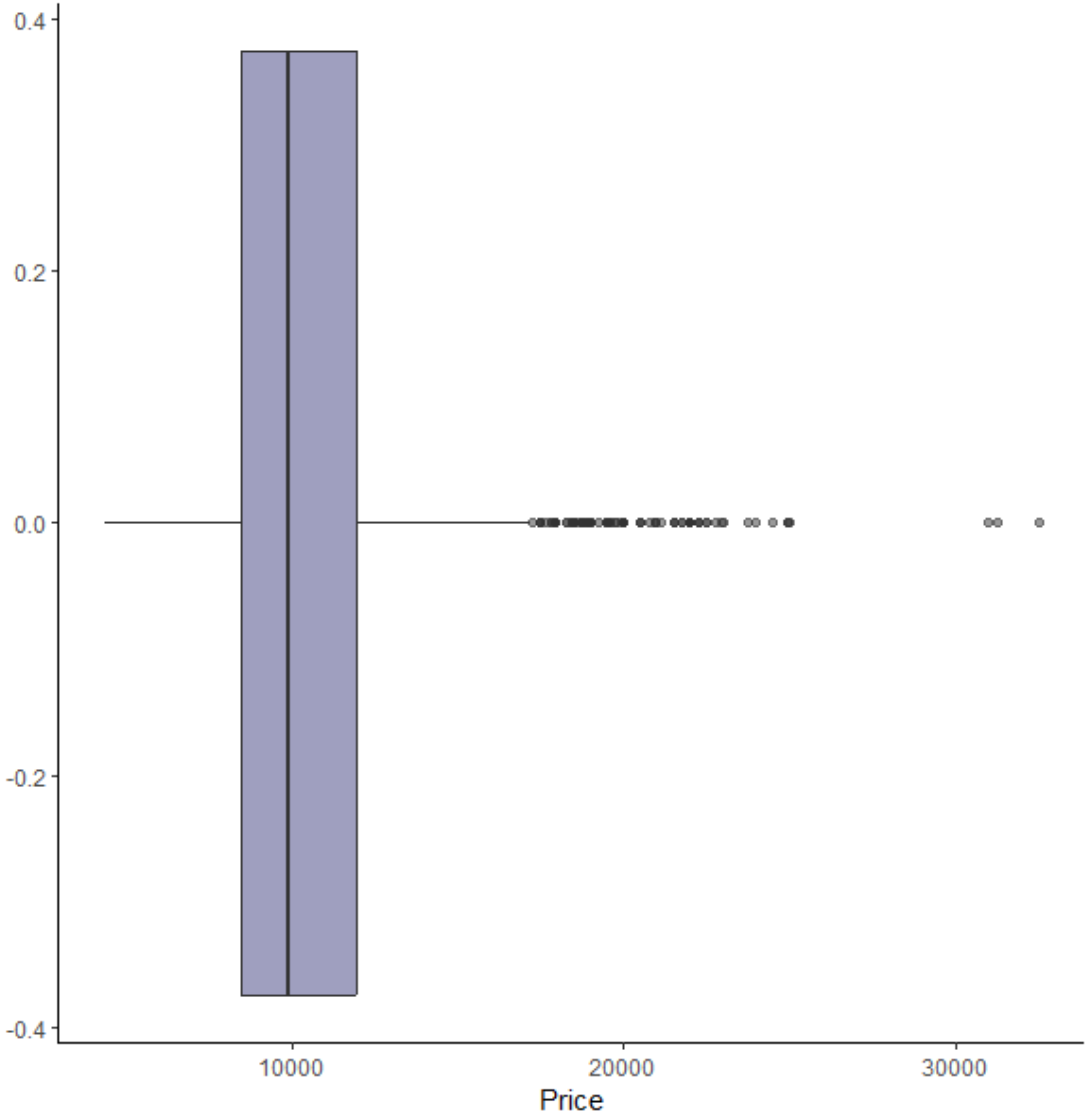


Figura 4. Precio de los vehículos Toyota

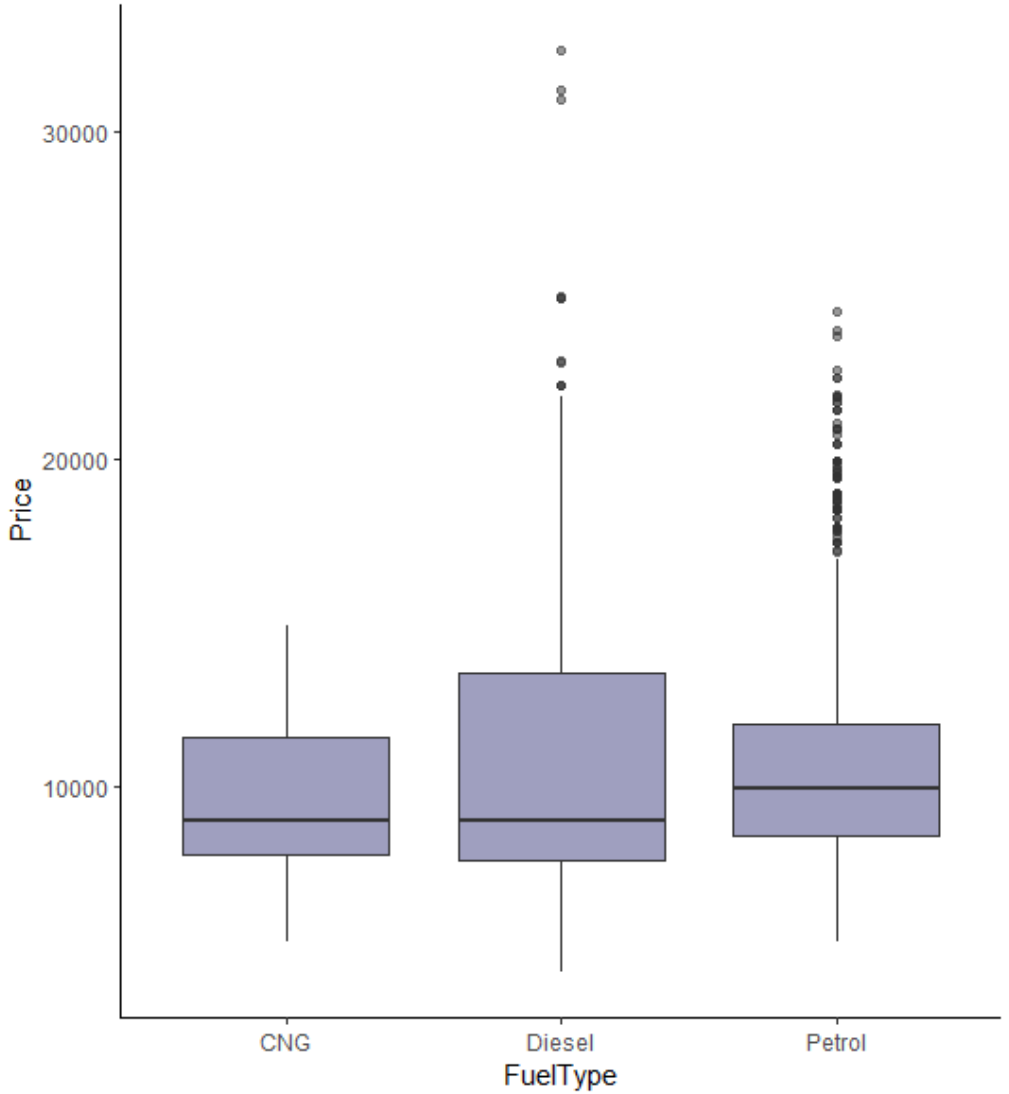


Visualización

Muestra fragmento de código

```
ggplot(dat, aes(x = FuelType, y=Price,
fill=FuelType))+
     geom_boxplot(alpha =
0.5,fill="#404080",orientation="x")+
     theme_classic()
```

Figura 5. Precio de los vehículos Toyota

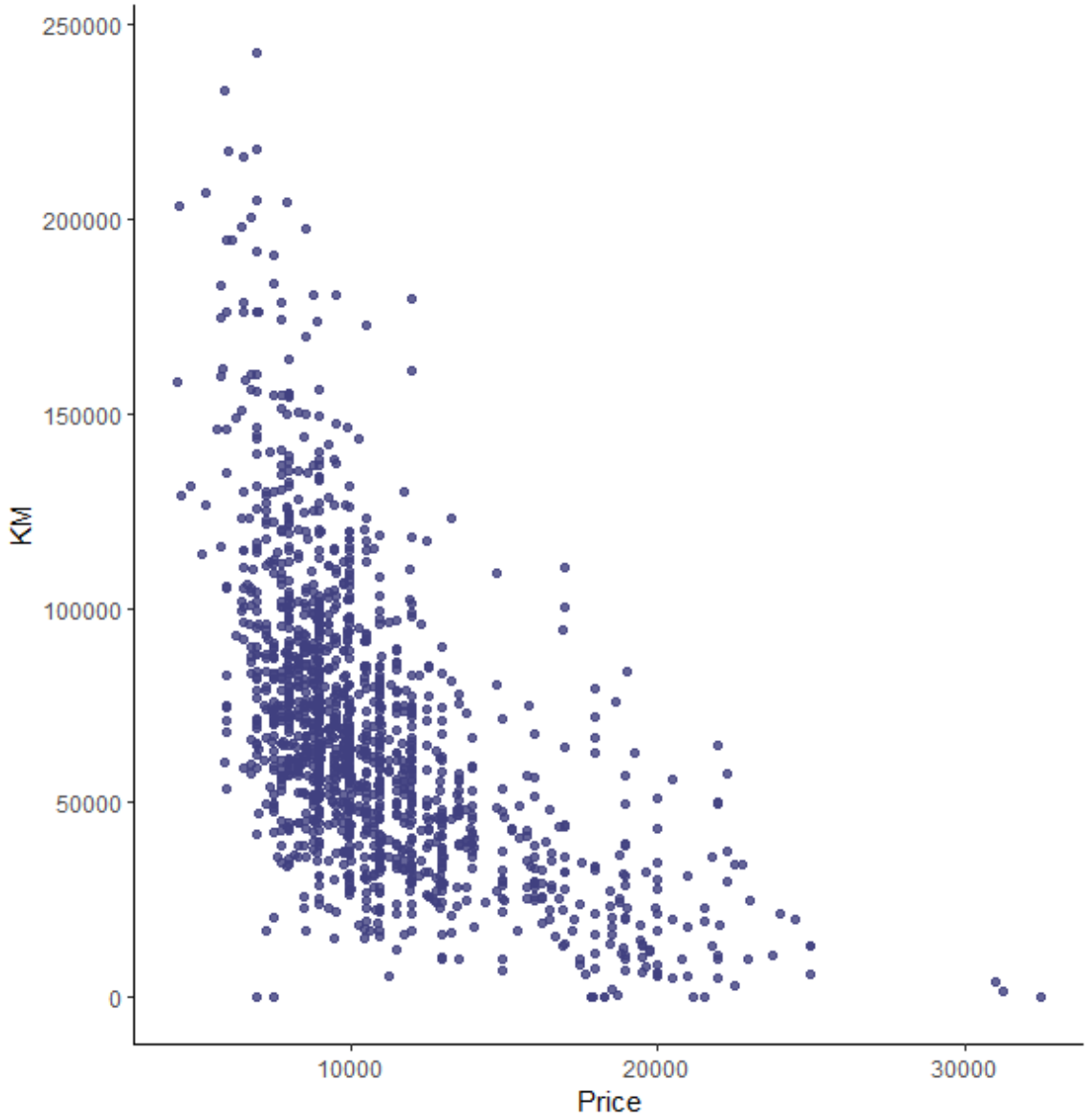


Visualización

Muestra fragmento de código

```
dat <- read_excel("henry/PredictingToyota.xlsx")

    ggplot(dat, aes(x=Price, y=KM))+
     geom_point(color="#404080", alpha=0.8)+
     theme_classic()
```

Figura 6. Comportamiento del precio según km recorrido

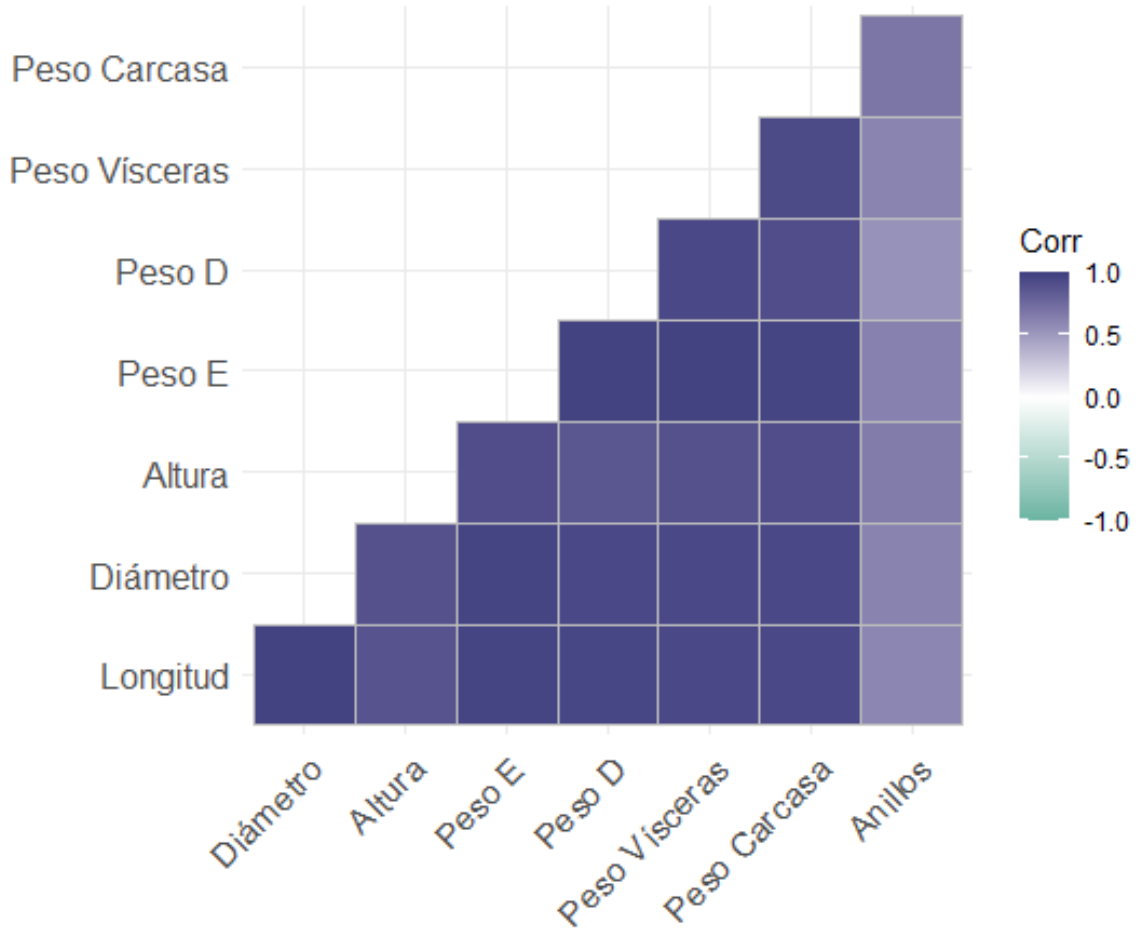


Visualización

Muestra fragmento de código

```
cor1 <- cor(abalone[,c(2:9)], method = "spearman")
   cor1

   ggcorrplot(cor1,
              method = "square",
              type = "lower",
              ggtheme = ggplot2::theme_minimal,
              title = "Matriz de correlacion",
              show.legend = TRUE,
              legend.title = "Corr",
              show.diag = NULL,
              colors = c("#69b3a2", "white",
"#404080"),
              outline.color = "gray",
              hc.order = FALSE,
              hc.method = "complete",
              lab = FALSE,
              lab_col = "black",
              lab_size = 4,
              p.mat = NULL,
              sig.level = 0.05,
              insig = c("pch", "blank"),
              pch = 4,
              pch.col = "black",
              pch.cex = 5,
              tl.cex = 12,
              tl.col = "black",
              tl.srt = 45,
              digits = 2,
              as.is = FALSE)
```

Figura 8. Matriz de correlación



Referencias

1. Binek, R. (2015). Kosaciec szczecinkowaty Iris setosa [Image]. Retrieved from https://commons.wikimedia.org/wiki/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg#/media/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg
2. Chihara, L. M., & Hesterberg, T. C. (2018). *Mathematical Statistics with Resampling and R* (2nd ed.). Wiley.
3. Kloke, J., & McKean, J. W. (2014). *Nonparametric Statistical Methods Using R (Chapman & Hall/CRC The R Series Book 25) (English Edition)* (1.ª ed.). Chapman and Hall/CRC.
4. González, G. C., Liste, V. A., & Felpeto, B. A. (2011). *Tratamiento de datos con R, Statistica y SPSS* (1.ª ed.). Ediciones Diaz de Santos.
5. Rasch, D., Pilz, J., Verdooren, L. R., & Gebhardt, A. (2011). *Optimal Experimental Design with R (English Edition)* (1.ª ed.). Chapman and Hall/CRC.
6. Husson, F., Le, S., & Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R* (2nd ed.). CRC Press.