

# Informe Consultoria

Luis Hernández y Juan Carvajal

2025-02-05

Table 1: Valores faltantes para cada variable

Variables	Valores faltantes
estado_vital_5anos	0
estado_vital_2	0
tiempo_evento_bx_5anos	0
tiempo_evento_bx_2	0
ciudad	0
edad	0
edad_cat	0
edad_cat2	0
estrato	0
estrato_cat	0
educacion	0
educacion_cat	0
afiliacion	0
lateralidad_cat	2
tipo_histologico	4
tipo_histol_cat	4
grado_histologico	31
grado_nuclear	9
gh_gn	33
t	15
n	16
m	9
estadio	15
estadio_cat	14
estadio_cat3	15
estadio_early_late	15
er	10

Table 1: Valores faltantes para cada variable

Variables	Valores faltantes
pr	11
her2	13
subtipo_molecular_definitivo	10
eur	7
nam	7
afr	7
eur_cat	7
nam_cat	7
afr_cat	7
recaidas	131
fecha_corte_seguimiento	0
fecha_dx	0
ano_dx	0
cuartil_fecha_dx	0
tiempo_supervivencia_dias	0
tiempo_supervivencia_anos	0
fecha_dx_paciente	0
fecha_bx	0
anos_supervivencia_dx	0
anos_supervivencia_bx	0
tiempo_supervivencia_5_anos_dx	0
pd_l1	191
area_ocupada_por_los_ti_ls_estromales_percent_total	139
interaccion_reg_stage	15
pd_l1_ti_ls_si_no	0
missing_clinical_data	0

De la Table 1 podemos concluir que:

#### Variables sin valores faltantes

Muchas variables clave no tienen datos faltantes, lo que indica una base de datos bien estructurada en su mayoría. Ejemplos:

- estado\_vital\_5anos
- estado\_vital\_2
- tiempo\_evento\_bx\_5anos
- edad, ciudad, afiliacion, fecha\_dx, tiempo\_supervivencia\_dias, etc.

#### Variables con algunos valores faltantes

Algunas variables presentan valores faltantes moderados (menores a 20 casos), lo que puede impactar el análisis dependiendo de la variable. Ejemplos:

- `grado_histologico` (31 valores faltantes)
- `t` (15), `n` (16), `m` (9)
- `estadio` (15), `er` (10), `pr` (11), `her2` (13)
- `subtipo_molecular_definitivo` (10)

### Variables con muchos valores faltantes

Algunas variables tienen un número considerable de datos faltantes, lo que puede representar un problema para el análisis. Ejemplos:

- `recaidas` (131 valores faltantes)
- `pd_l1` (191)
- `area_ocupada_por_los_tis_estromales_percent_total` (139)

## Análisis Descriptivo

Para obtener correlaciones entre variables categóricas se utilizó el **coeficiente de correlación de Cramer's V**, a continuación se da una breve introducción.

El coeficiente de correlación de Cramer's V es una medida de asociación entre dos variables categóricas. Es especialmente útil cuando las variables tienen más de dos categorías (politómicas) y no son ordinales (es decir, no tienen un orden inherente). Cramer's V se basa en la prueba de chi-cuadrado ( $X^2$ ) y es una extensión del coeficiente phi ( $\phi$ ), que se usa para tablas de contingencia de 2x2.

Cramer's V cuantifica la fuerza de la asociación entre dos variables categóricas. Su valor oscila entre 0 y 1:

- 0: Indica que no hay asociación entre las variables (son independientes).
- 1: Indica una asociación perfecta (las variables están completamente relacionadas).

Cramer's V se calcula a partir de la prueba de chi-cuadrado ( $X^2$ ) y ajusta el tamaño de la tabla de contingencia. La fórmula es:

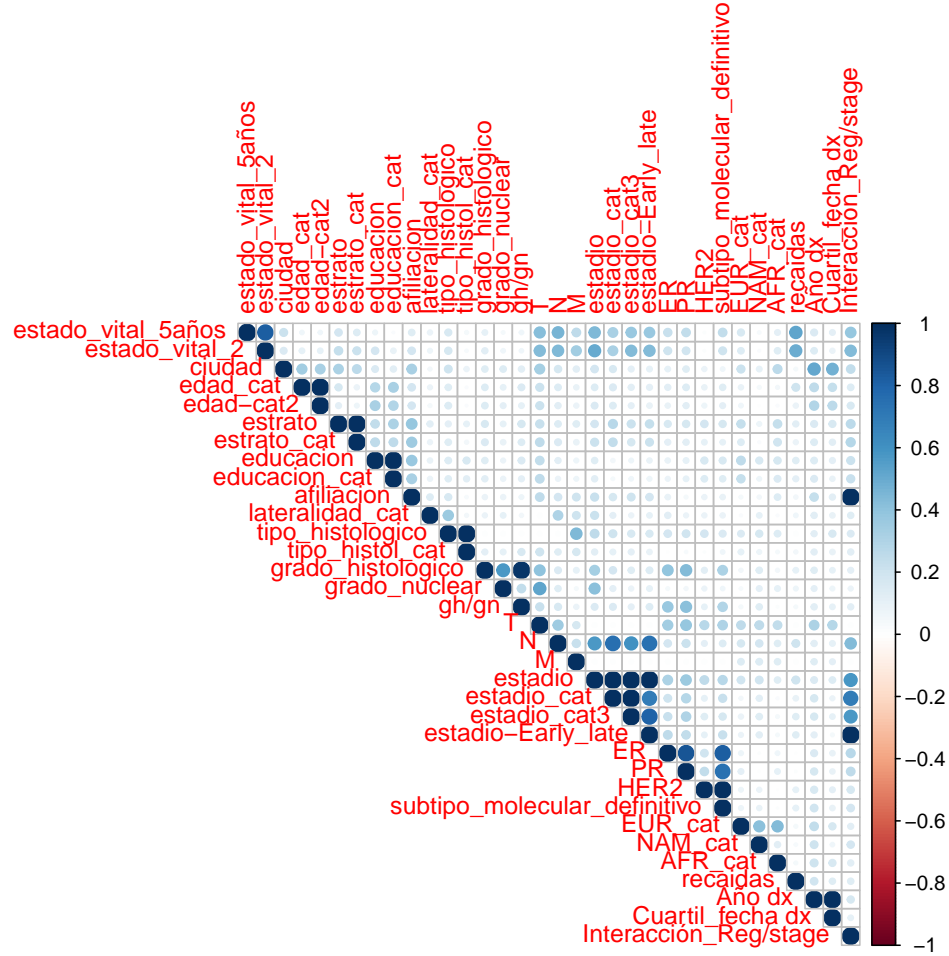
$$V = \sqrt{\frac{x^2}{n \cdot \min(r-1, c-1)}}$$

Donde:

- $X^2$ : Es el valor de la prueba de chi-cuadrado.
- $n$ : Es el número total de observaciones

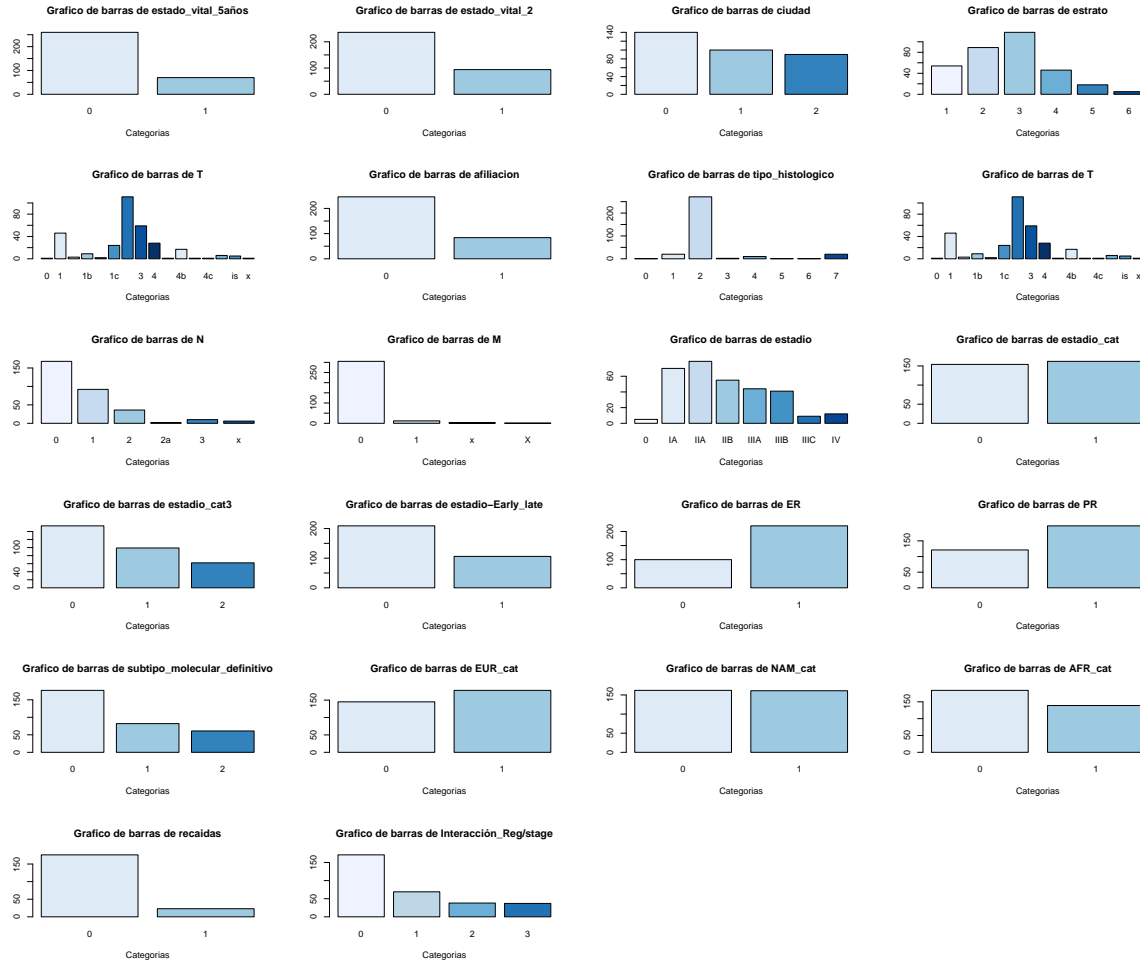
- $r$ : Es el numero de filas de la tabla de contingencia.
- $c$ : Es el numero de columnas de la tabla de contingencia.
- $\min(r-1, c-1)$ : Es el minimo entre el numero de filas menos 1 y el numero de columnas menos 1.

## Correlaciones entre las variables



La matriz de correlación presentada muestra las relaciones entre las variables categóricas de nuestra base de datos. Dado que el objetivo del estudio es determinar el estado vital a cinco años en mujeres con cáncer de mama utilizando variables predictoras categóricas, analizamos cuáles se relacionan mejor con la variable de respuesta. Además, identificamos aquellas variables predictoras con altas correlaciones entre sí. Estas correlaciones nos permiten, como primer paso, seleccionar variables para evaluar su comportamiento tanto de manera individual como en conjunto con la variable de respuesta.

## Análisis descriptivos de variables individuales



En los diagramas de barras presentados, podemos observar las frecuencias de los distintos niveles de cada variable predictora. Un aspecto relevante que notamos es que, en su mayoría, las variables categóricas analizadas presentan entre dos y tres niveles, mientras que solo unas pocas tienen más de tres categorías. Este patrón sugiere que las variables con un menor número de niveles podrían ser más fáciles de interpretar y modelar en relación con la variable de respuesta. Además, hemos identificado que aquellas variables con tres o menos factores tienden a mostrar una mejor correlación con la respuesta, lo que indica su posible relevancia en el estudio. Esto nos permite enfocar el análisis en estas variables, explorando su impacto tanto de manera individual como en combinación con otras, con el objetivo de mejorar la capacidad predictiva del modelo.

## Análisis descriptivo de variables en conjunto con la respuesta



Para finalizar el análisis descriptivo, observamos cómo se relacionan los niveles de las variables predictoras más influyentes con los niveles de la variable de respuesta. Los gráficos de mosaico muestran, de manera visual, la frecuencia con la que los distintos niveles de la variable de respuesta coinciden con cada una de las categorías de las variables predictoras. Esto nos permite identificar patrones en la distribución de los datos y evaluar posibles asociaciones entre las variables, lo que será útil para el análisis posterior.

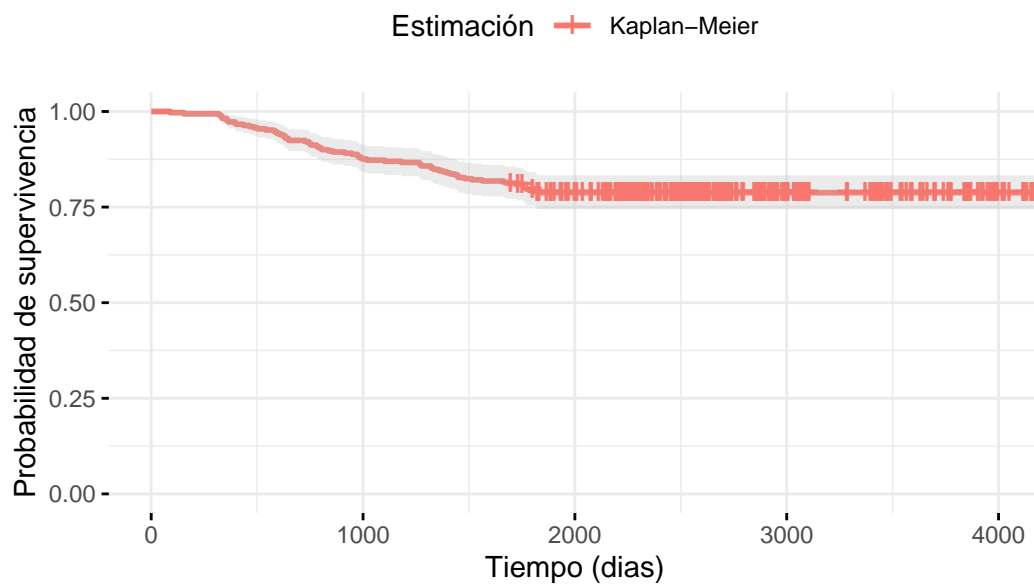
## Análisis de Supervivencia

Para las curvas de supervivencia utilizaremos el estimador de Kaplan-Meier.

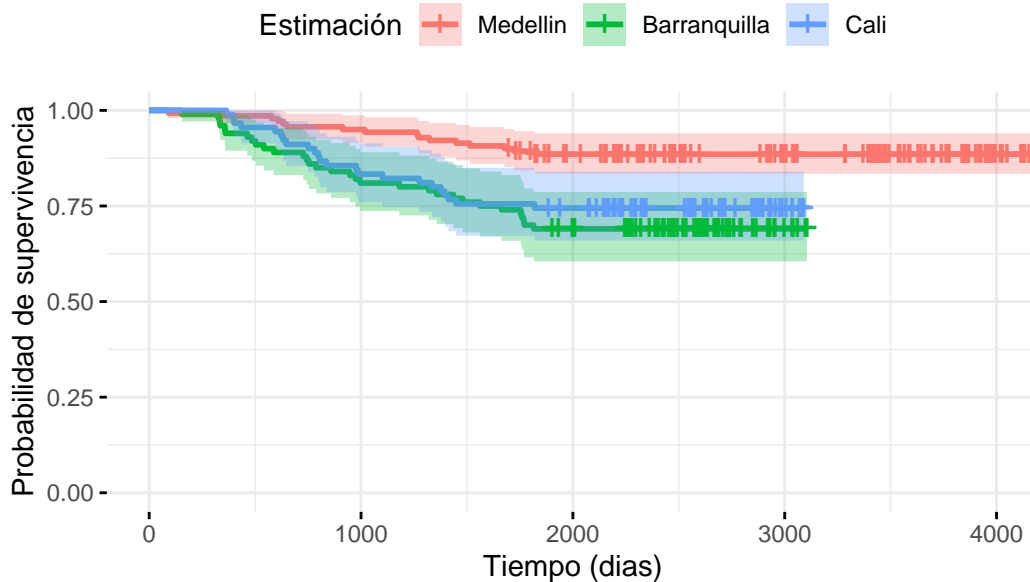
### Estimador de Kaplan-Meier.

estado_vital_5anos	tiempo_supervivencia_dias
1	1271
0	4237
0	4323
0	4286
0	4293
1	1677

### Curva de Supervivencia



## Curva de Supervivencia por Ciudad



Call:

```
survdif(formula = Surv(tiempo_supervivencia_dias, estado_vital_5anos) ~
  ciudad, data = bd, rho = 0)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
ciudad=0	140	16	31.5	7.62	13.86
ciudad=1	100	31	20.0	6.07	8.50
ciudad=2	90	23	18.5	1.08	1.47

Chisq= 14.8 on 2 degrees of freedom, p= 6e-04

## Modelo de Riesgos Proporcionales de Cox

En las situaciones experimentales en las que deseamos estudiar la supervivencia de un conjunto de sujetos en función de un conjunto  $X = (X_1, \dots, X_p)$  de variables predictoras, es decir, variables que pueden afectar o caracterizar su supervivencia, es necesario establecer modelos estadísticos capaces de analizar dichas relaciones. La construcción de este tipo de modelos que depende del tiempo y de las predictoras se hace a través del análisis de la función hazard asociada  $h(t; X)$ .

El modelo más habitual en esta situación es el **modelo hazard proporcional** que separa en dos componentes la función hazard, una correspondiente al tiempo de supervivencia y otra a



las variables predictoras. La finalidad de este modelo es para identificar factores que influyen en la supervivencia.

A manera de ejemplo se ajustara un modelo con algunas variables, las variables a considerar al modelo final, y se tranda en cuenta tambien el criterio de Akaike.

Call:

```
coxph(formula = Surv(tiempo_supervivencia_dias, estado_vital_5anos ==
  1) ~ ciudad + edad_cat + estrato_cat + educacion_cat + afiliacion,
  data = bd)
```

n= 330, number of events= 70

	coef	exp(coef)	se(coef)	z	Pr(> z )
ciudad1	1.0247	2.7862	0.3470	2.953	0.003147 **
ciudad2	1.2044	3.3347	0.3567	3.377	0.000733 ***
edad_cat1	-0.2012	0.8177	0.3809	-0.528	0.597242
edad_cat2	-0.3109	0.7328	0.4132	-0.753	0.451714
edad_cat3	-0.1475	0.8629	0.4144	-0.356	0.721891
estrato_cat1	0.8648	2.3746	0.7397	1.169	0.242356
estrato_cat2	1.3519	3.8648	0.7696	1.757	0.078980 .
educacion_cat1	0.1487	1.1603	0.3092	0.481	0.630482
educacion_cat2	0.5548	1.7415	0.4227	1.313	0.189335
afiliacion1	0.4964	1.6428	0.2795	1.776	0.075733 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
ciudad1	2.7862	0.3589	1.4114	5.500
ciudad2	3.3347	0.2999	1.6575	6.709
edad_cat1	0.8177	1.2229	0.3876	1.725
edad_cat2	0.7328	1.3647	0.3261	1.647
edad_cat3	0.8629	1.1589	0.3830	1.944
estrato_cat1	2.3746	0.4211	0.5571	10.121
estrato_cat2	3.8648	0.2587	0.8551	17.466
educacion_cat1	1.1603	0.8618	0.6330	2.127
educacion_cat2	1.7415	0.5742	0.7606	3.987
afiliacion1	1.6428	0.6087	0.9499	2.841

Concordance= 0.678 (se = 0.031 )

Likelihood ratio test= 27.26 on 10 df, p=0.002

Wald test = 24.51 on 10 df, p=0.006

Score (logrank) test = 26.51 on 10 df, p=0.003