A la caza del phi: explicando qué obras “dan la talla” áurea

Manuela Lopez Cambron, 1673688

2026-01-30

CONTENIDOS

1. Introducción

1.2 Motivación del estudio

La razón áurea (también llamada proporción áurea, sección áurea o phi) es una proporción numérica aproximada de 1,618. Se define cuando dividimos un segmento en dos partes de forma que la relación entre la parte mayor y la menor sea la misma que la relación entre el total y la parte mayor. Esa proporción aparece en geometría y se relaciona con la sucesión de Fibonacci, porque el cociente entre términos consecutivos de Fibonacci se aproxima a 1,618 a medida que los números crecen.

En arte y diseño, la razón áurea se ha popularizado como una regla de composición asociada a armonía visual. En el discurso histórico se vincula con tradiciones artísticas (especialmente desde el Renacimiento) y con la idea de que ciertas proporciones resultan equilibradas o agradables a la vista. Sin embargo, que algo se use como herramienta o que aparezca en ejemplos puntuales no implica que automáticamente sea una regla universal ni que explique por sí sola la "belleza" o la preferencia estética. Precisamente por eso es interesante tratarlo como un problema empírico: medir, comparar y contrastar hipótesis con datos.

La investigación psicológica y experimental sobre si preferimos la razón áurea ha dado resultados mixtos. Hay estudios experimentales donde se comparan versiones de una misma imagen/pintura ajustadas a distintas proporciones y se observa preferencia por la sección áurea en ciertos contextos y muestras, pero también hay trabajos que cuestionan que exista una preferencia automática o universal por esta proporción, especialmente cuando se usan pruebas implícitas y estímulos más variados.

Además una línea importante de literatura advierte que muchas afirmaciones populares sobre la razón áurea en arte se apoyan en selecciones de ejemplos o en mediciones discutibles y que cuando se analiza de manera sistemática (por ejemplo, proporciones de formato en pinturas), la proporción áurea no siempre aparece como la proporción dominante.

1.3 Objetivos e hipótesis

Objetivos

El objetivo general de este trabajo es construir un modelo explicativo que nos ayude a entender qué factores se asocian con que una obra cumpla o no la proporción áurea. No buscamos hacer predicción "para acertar", sino explicar con qué características de las obras (como la época, el tamaño, la orientación, la técnica o el soporte) es más probable encontrar la proporción áurea y cómo estas características se relacionan entre sí.

Como objetivo específico, queremos describir si el cumplimiento de la proporción áurea es un fenómeno frecuente dentro del conjunto de obras analizadas en el museo del prado y si dicho cumplimiento se concentra en determinados contextos (por ejemplo, en ciertos periodos históricos o en ciertos formatos).

Aunque nuestro conjunto de datos no mide directamente la "belleza" ni la "calidad artística", sí nos permite plantear la pregunta de manera crítica: si la proporción áurea fuera una regla compositiva muy general en el arte, deberíamos observar patrones claros y consistentes en las obras analizadas; y si aparece de forma irregular o depende fuertemente del contexto, ello refuerza la idea de que su presencia no es universal y que conviene interpretarla como una herramienta posible, pero no como un criterio determinante.

Hipótesis

Consideramos las siguientes hipótesis acerca de variables que podrían relacionarse con la probabilidad de cumplimiento de la razón aurea en las pinturas:

1) Posiblemente la fecha de creación de la obra de relacione con el evento de interés debido a corrientes artísticas ir al innegable cambio es los conceptos de belleza.

2) Presuponemos que posiblemente el tamaño de la obra mantenga relación con la decisión de sus medidas y su relación (proporción aurea) debido a la perspectiva con que esta se mira. Es decir, cuando una obra

es más pequeña podemos verla en su totalidad más rápidamente, mientras que si esta es de dimensiones más grandes nos vemos forzados a recorrerla con la mirada. Por esta razón queremos estudiar si la proporción aurea pudiera estar relacionada con estos aspectos y quizás marcar la semilla de una futura investigación.

3) Muy relacionado con el punto anterior, se nos planteó la pregunta de si quizás las dimensiones de la obra por si solas no fueran el punto clave. Consideramos que el material de soporte utilizado es un factor influyente en sus dimensiones y quizás más limitante en unos tamaños que en otros, por lo tanto relacionado con la proporción aurea en función del tamaño de la obra

4) Relacionado con el apartado anterior, también se considera estudiar si la iconografia de la pintura podría estar relacionada debido a aspectos de composición en su interior que acaben repercutiendo en las dimensiones totales.

1.4 Presentación de los datos

Para el propósito del estudio se ha decidido estudiar la coleción de obras de arte del Prado, concretamente aquellas clasificadas como pinturas. Mediante web scraping, se ha extraido la ficha técnica de las 7.117 pinturas, la cual contiene la siguiente información:

Siempre (en todas las obras):

- Número de catálogo (asignado por el Prado, único por obra)
- Título
- Fecha

Casi siempre:

- Técnica (7.114/7.117)
- Dimensión (7.107/7.117)
- Procedencia (7.098/7.117)
- Soporte (7.096/7.117)
- A veces:
  - Serie (1.415/7.117)

Muy raros:

- Materia, Lugar de producción, Edición / Estado

Están son las variables originales de las que se disponía, y después de la correspondiente gestión de datos se ha obtenido la base de datos final con un total de 7.002 registros y 12 variables preparadas para el análisis. En el archivo original, los valores faltantes se condificaban como “0”, después de su apropiada gestión, se proporciona un dataset sin valores faltantes.

Diccionario de variables

- exito: Indicador de cumplimiento de “razón aurea”, con error del 5% (categórica binaria, 2 niveles: 0 no, 1 sí)

- area: Área de la pintura a partir de las dimensiones, tamaño en formato numérico (numérica)
- tam_cat: Tamaño categorizado a partir del área usant cuantiles (categórica ordinal, utilizada como nominal, 3 niveles: pequena, mediano, grane)
- orientación: Forma según comparación de dimensiones (categórica nominal, 3 niveles:vertical, horizontal, cuadrado)
- soporte_grp: Agrupación de tipos de soporte en familias de material (categórica nominal, niveles: Lienzo, Tabla/Panel, Metal, Mural, Otros)
- sop_montaje: Indicador de "montaje/transferencia" detectado en el campo soporte (cotegórica binaria, 2 niveles: 0 no, 1 sí)
- tecnica: Agrupación de técnicas en grupos genéricos (categórica nominal, 3 niveles: mixta, oleo, otras)
- tipo_autor: A partir de autor/autora/autores se creó un tipo de autoría (categórica nominal, 4 niveles: hombre, mujer, varios, anonimo)
- serie: Indicador de pertenencia a serie (categórica binaria, 2 niveles: 0 no, 1 sí)
- fecha_est: Año estimado a partir de la datación convertida en intervalo, año central de este (numérica)
- fecha_ancho: Incertidumbre/ancho del intervalo temporal de datación (numérica)
- tema: Tema asignado en función de términos clave encontrados en el texto original de título (categórica, 10 niveles: religioso, mitologia, retrato_corte, historia_aleagoria, paisajes_lugares, vida_cotidiana, bodegon_floral, caza_animales, proceso_obra, otros)

2. Metodología

Como enfoque general se decidieron adoptar procedimientos que permitieran contar una historia en relación a las variables que finalmente se involucren. No solo obtener resultados automatizados sinó estudiar en profundidad cómo unas variables, en presencia o no de otras, afectan a la probabilidad de observar el evento de interés. Por lo tanto, no se presentará un único modelo, pues consideramos que dado el gran alcance de nuestra base de datos, no existe un único modelo óptimo para estudiarla, sino diversas combinaciones de variables que respondan a preguntas de distinto enfoque.

Como aprendimos en la última práctica, gracias a la lectura del artículo "A hypotesis is a liability", la persecución de nuestras hipótesis no debe ser el único objetivo de un buen estudio. Debemos recordar la importancia de trabajar los datos tanto bajo la ciencia diurna como la nocturna, de manera que mantengamos la mente y los ojos abiertos a nuevas posibilidades aunque siempre en presencia y ayuda de procedimientos y técnicas adecuadas que nos permitan analizarlos rigurosamente.

Por esta razón, se descartan los métodos de selección automática tipo stepwise con el fin de controlar cada paso interno que este tipo de procesos esconden. En su lugar, se hará uso de una filosofía de selección basada en bloques conceptuales que definen las variables. Estos bloques se introducirán uno por uno por uno de manera que se irá construyendo un modelo principal sobre el que trabajaremos pero también se irá guardando la información obtenida en cada paso no fructifero de manera que se utilizará para generar otros modelos alternativos que nos permitan estudiar la variable respuesta desde otro enfoque. Esta filosofía pretende ser adecuada para explorar al máximo nuestra base de datos y darnos información potencialmente interesante sobre todas y cada una de las variables.

Por otro lado debemos poder actuar de forma estadísticamente correcta, por lo que en esta sección se detallan los procedimientos y técnicas utilizadas, además de los criterios preestablecimos en relación a las sucesivas decisiones que se tomarán para derivar tanto el modelo principal como los alternativos.

Regla estructural 'cuadrado'

Primera deberemos modificar nuestra abse de datos de manera que se eliminarán las observaciones de pinturas cuadradas. Como se informó anteriormente, la categoría 'cuadrado' de la variable "orientacion" genera ceros estructurales, ya que por definición los cuadros cuadrados nunca seguirán la proporción aurea. La presencia de una categoría con ausencia completa del evento induce separación perfecta en modelos binarios, lo que puede producir estimaciones inestables o no finitas y distorsionar la estimación de efectos del resto de covariables. Por estos motivos, se restringió el análisis inferencial a la subpoblación con orientacion $\neq$ "cuadrado". Gracias a esta modificación podemos introducir la variable "orientación" sin problemas, pero la inferencia de nuestro estudio solo será aplicable a esta subpoblación.

Modelos y correcciones

La base para el análisis serán Modelos Lineales Generalizados (GLZ/GLM). No se contempla la utilización de modelos mixtos dado que no hay presencia aparente de estructura por bloques. Las observaciones se consideran independientes entre sí, puesto que disponemos de una única medición de cada variable para cada obra, ni ninguna variable que naturalmente las pueda estructurar. Nuestra variable respuesta es de tipo binaria, por lo que asumiremos una distribución Binomial con enlace logit (enlace canónico de la distribución). Es decir trabajaremos en todo momento con regresiones logísticas.

Dado el desbalance y la posible separación a la que podremos enfrentarnos, se considerará añadir una reducción de sesgo con un enfoque Firth.

Selección de efectos principales

Es en este punto donde se remarca la filosofía por bloques utilizada: para la selección de las covariables que formarán parte del modelo principal, y las que se mantendrán para los modelos alternativos.

Las variables han sido clasificadas en los siguientes bloques según su interpretación conceptual:

1) Datación + incertidumbre (fecha_est, fecha_ancho)
2) Morfología (log(area), orientacion) + (tam_cat)
3) Material y tecnica (soporte_grp, tecnica) + (sop_montaje, como "control")
4) Iconografia (tema)
5) Autoría y serie (tipo_autor, serie)

Comenzaremos con un modelo nulo sobre el cual se irán añadiendo estos bloques de variables sucesivamente y uno por uno. En cada paso se valorará si el bloque demuestra o no mejora respecto el modelo anterior. Se analizará si éste aporta información adicional mediante comparación de modelos anidados y el coste en complejidad que refleja. Si el veredicto es positivo, entonces se mantendrá el bloque en el modelo principal y se añadirá el siguiente sobre este. Si un bloque o una variable no demuestra aportar información al modelo ajustado por los anteriores, se descartará automáticamente. Si por el contrario existe aporte de información pero coste de complejidad excesivo, se mantendra el bloque para ser utilizado en un futuro modelo alternativo.

Selección de interacciones

Una vez escogidas las covariables principales, se procederá a determinar la inclusión de algunas interacciones al modelo principal. No se considerarán interacciones de dos variables. Aunque algunas de ellas son de interés para nosotros con el fin de dar respuesta a nuestras hipótesis, la principal herramiento para su selección serán los gráficos de analisis descriptivo, donde se hará una primera criva. Las candidatas serán sometidas a una segunda criva en esta sección del análisis de interacciones. Primeramente se graficarán los gráficos de interacción de los modelos con cada una de ellas. Seguidamente, aquellas que resulten significativas se verificarán mediante la inspección de sus combinaciones de niveles para descartar relaciones engañosas devido a celdas vacías (ya que se han observado incidios de desbalances muchas de ellas). Finalmente las seleccionadas se someterán a pruebas formales mediante modelos anidados.

Trnasformación de variables

Según se visualice en el análisis descriptivo, se decidirá si algunas de las variables continuas deben ser transformadas o no para entrar en el modelo.

Además, para estas variables se considerarán siempre tanto una especificación lineal como una flexible mediante spline con 3 nudos, ya que no se espera una relación que exista una relación del todo lineal entre "fecha_est" y "area" con la respuesta, pero ambas son variables que queremos que entren en el modelo. Abas especificaciones se compararan con modelos anidados.

Comparación de modelos

Para las comparaciones de modelos anidados se empleará el test de razón de verosimilitudes (LRT) basado en diferencia de devianza con distribución χ^2. Con una confianza del 95% en el análisis de efectos principales y con una confianza del 99% en el análisis de interacciones.

También de emplearán los criteriors de información Akaike y Bayesiano (AIC/BIC) para evaluar el coste de complejidad, consideran un aumento de 2 puntos como significativo.

Para cada modelo anidado se proporcionará también el resumen estadístico de este (summary) para visualizar rápidamente si vemos indicios problemáticos, concretamente para verificar que los errore estándares de los coeficientes no se disparán y que el modelo se ha ajustado correctamente.

Validación de modelos

Para la validación de los modelos se ha analizado la desviación i pseudoR2. La capacidad discriminativa mediante el área bajor la curva y las gráficas ROC. Analizado los residuos de los modelos y revisado los problemas de multicolinealidad mediante el VIF y el GVIF ajustado.

Para el modelo principal también se ha analizado la distància de cook al detectar mediante el summary que havia interacciones influyentes. Posteriormente, se ha creado la tabla para detectar las causas.

3. Gestión de datos e ingeniería de características

En esta sección se trabajará la información obtenida del web scrapping con el objetivo de crear la base de datos final que analizaremos, la cual ha sido descrita en la sección anterior.

3.1 Manejo de valores faltantes

Tal y como se ha generado la base de datos, los valores faltantes estan indicados como "0". Veamos el porcentaje de estos para cada variables:

```
##            variable    prop_cero
## 6            estado 0.9998582566
## 8  lugar_produccion 0.9961729270
## 9           materia 0.9958894401
## 4            autora 0.9916371368
## 5           autores 0.9780297661
## 12            serie 0.8005669738
## 3             autor 0.0303330971
## 11      procedencia 0.0026931254
## 13          soporte 0.0026931254
## 14          tecnica 0.0004252303
## 1              alto 0.0000000000
## 2             ancho 0.0000000000
## 7             fecha 0.0000000000
## 10  numero_catalogo 0.0000000000
```

```
## 15           titulo 0.0000000000
## 16              url 0.0000000000
```

Eliminamos “estado”, “lugar_produccion” y “materia”, debido a su gran presencia de valores faltantes no podemos extraer información. Eliminamos también los casos donde haya ausencia de información para las variable “procedencia”, “soporte” o “tecnica”. Justificamos el análisis de casos completos considerando esta ausencia totalmente aleatoria provocada por fenomenos sociales o históricos a cerca de la conservaciónd de estas obras que no mantienen relación con nuestro objetivo, la proporción aurea.

Los valores faltantes de las otras variables se manejaran automaticamente mediante las transformaciones pertinentes de estas en la siguiente sección.

Eliminar variables no deseadas

```
##  [1] "alto"            "ancho"           "autor"           "autora"
##  [5] "autores"         "fecha"           "numero_catalogo" "procedencia"
##  [9] "serie"           "soporte"         "tecnica"         "titulo"
## [13] "url"
```

Eliminamos casos incompletos

```
## Se han eliminado 20 casos
```

3.2 Transformación de variables

Aunque nuestra base de datos ya aporta la información necesaria, sus variables no presentan la estructura que necesitamos para trabajarla. Por esta razón, en esta sección se procederá a transformar y crear nuevas variables de manera que queden listas para ser utilizadas en el modelaje.

Recodificación de variables de autoría

Agruparemos las variables “autor”, autora”, “autores” creando una variable “tipo_autor” que indicará si se trata de ‘hombre/mujer/varios/anonimo’ y además se conservarán los nombres en una nueva variable llamada “nombre_autor”.

```
##
## anonimo  hombre   mujer  varios
##     661    6161      59     154
##    tipo_autor nombre_autor
## 1     anonimo      Anónimo
## 2     anonimo      Anónimo
## 3     anonimo      Anónimo
## 4     anonimo      Anónimo
## 5     anonimo      Anónimo
## 6     anonimo      Anónimo
## 7     anonimo      Anónimo
## 8     anonimo      Anónimo
## 9     anonimo      Anónimo
## 10    anonimo      Anónimo
##      tipo_autor                              nombre_autor
## 7046     hombre Álvarez de Sotomayor y Zaragoza, Fernando
## 7047     hombre          Carretero Cepeda, Francisco José
## 7048     hombre          Carretero Cepeda, Francisco José
## 7049     hombre          Carretero Cepeda, Francisco José
```

```
## 7050     hombre          Carretero Cepeda, Francisco José
## 7051     hombre                              Gaya, Ramón
## 7052     hombre                Werboff, Michel Alexander
## 7053     hombre Álvarez de Sotomayor y Zaragoza, Fernando
## 7054     hombre Álvarez de Sotomayor y Zaragoza, Fernando
## 7055     hombre          Carretero Cepeda, Francisco José
```

Recodificación de variable "serie"

Consideramos apropiado mantener "serie" en calidad de conocer si la pintura pertenece (1) o no (0) a una serie, sin importar a cuál, por lo que la convertiremos en una variable binaria indicadora.

```
##
##    0    1
## 5628 1407
```

Recodificación de variable "tecnica"

Estamos frente a una variable que podría ser interpretada como un factor pero presenta demasiados niveles.

```
## Niveles de 'tecnica':
##
##                                  Dorado con pan de oro; Óleo
##                                                           1
##                                              Dorado; Óleo
##                                                           1
##                                            Dorado; Témpera
##                                                           3
##                                             Dorado; Temple
##                                                           2
##                                                 Enconchado
##                                                           8
##                                            Enconchado; Óleo
##                                                          23
##                                                    Grisalla
##                                                           1
##                                              Grisalla; Óleo
##                                                          14
##                                            Grisalla; Temple
##                                                           1
##                                                        Óleo
##                                                        6802
##                     Óleo; Pastillaje; Dorado con pan de oro
##                                                           1
##                                             Óleo; Témpera
##                                                           5
##                                              Óleo; Temple
##                                                           3
##                                              Pastel; Óleo
##                                                           1
##                                          Pintura al fresco
##                                                          35
## Pintura al fresco; Pintura al seco (falso fresco o a la cal)
##                                                           1
```

```
##                                          Técnica mixta
##                                                     66
##                                  Técnica mixta; Temple
##                                                      1
##                               Técnicas de fotografía; Óleo
##                                                      1
##                                                Témpera
##                                                      3
##                                           Témpera; Aguada
##                                                      1
##                                                 Temple
##                                                     58
##                                            Temple de cola
##                                                      1
##                                              Temple graso
##                                                      2
```

La solución óptima, dado que muchos casos presentan niveles multi-etiqueta, será recodificar en varias variables dummies como indicadores interpretables de la presencia de cada categoría. De esta manera podremos asociar la presencia de cada técnica concreta con el aumento o no en la probabilidad de observar nuestro evento de interés. Obviamos las técnicas poco frecuentes 'Pastillaje', 'Pastel', 'Aguada', 'Técnicas de fotografía'

```
## Frecuencia de cada técnica:
##       tec_oleo     tec_temple      tec_mixta     tec_fresco tec_enconchado
##           6852             68             67             36             31
##   tec_grisalla    tec_tempera     tec_dorado tec_pastillaje     tec_pastel
##             16             12              8              1              1
##     tec_aguada       tec_foto
##              1              1
```

Vemos que hay una serie de técnicas con poca frecuencia, por el bien de nuestros futuros modelos, agruparemos estas variables dummies en una nueva llamada "tec_otras". La única con baja frecuencia que conservaremos será 'fresco' debido a su posible efecto sobre la respuesta, ja que envuelve un tipo de obra con características propias. También unificaremos 'temple', 'tempera' y 'aguada' en "tec_acuosas" por sus características similares.

```
## Frecuencia de cada técnica:
##    tec_oleo tec_acuosas   tec_mixta   tec_otras  tec_fresco
##        6852          80          67          57          36
```

A continuación proporcionamos un listado de los niveles originales y que dummies activa cada uno:

```
##                                              tecnica        activas
## 1                  Dorado con pan de oro; Óleo    oleo, otras
## 2                              Dorado; Óleo    oleo, otras
## 3                            Dorado; Témpera acuosas, otras
## 4                             Dorado; Temple acuosas, otras
## 5                                 Enconchado          otras
## 6                               Enconchado; Óleo    oleo, otras
## 7                                      Grisalla          otras
## 8                                Grisalla; Óleo    oleo, otras
```

```
## 9                                                Grisalla; Temple acuosas, otras
## 10                                                          Óleo           oleo
## 11                         Óleo; Pastillaje; Dorado con pan de oro    oleo, otras
## 12                                                 Óleo; Témpera  oleo, acuosas
## 13                                                  Óleo; Temple  oleo, acuosas
## 14                                                  Pastel; Óleo    oleo, otras
## 15                                              Pintura al fresco         fresco
## 16 Pintura al fresco; Pintura al seco (falso fresco o a la cal)         fresco
## 17                                                 Técnica mixta          mixta
## 18                                       Técnica mixta; Temple acuosas, mixta
## 19                                    Técnicas de fotografía; Óleo    oleo, otras
## 20                                                        Témpera        acuosas
## 21                                                Témpera; Aguada        acuosas
## 22                                                         Temple        acuosas
## 23                                                  Temple de cola        acuosas
## 24                                                    Temple graso        acuosas
```

Para confirmar si el tipo de codificación muultietiqueta con dummies es útil o no vamos a detectar cuantas observacines hacen uso de tal carácter contando cuantas ativan más de 1 dummy:

```
## [1] 57
```

Vemos que el porcentaje de observaciones que activan más una técnica (después de nuetsra reagrupación de estas) es obviamente despreciable. Por esta razón y con el objetivo de simplificar nuestra base de datos, procedemos a reestablecer un único factor indicativo de la técnica utilizada, con los actuales valores de las dummies, y clasificaremos estas observaciones multietiqueta dentro de la categoria 'tec_mixta'.

```
##
##    oleo   mixta acuosas  fresco   otras
##    6802     123      65      36       9
```

Recodificación de variable "soporte"

Esta variable presenta el mismo problema de exceso de niveles.

```
##
## Niveles de 'soporte':
##
##                                  Cartón                   Cartón sobre lienzo
##                                     27                                     2
##                      Cartón sobre tabla            Cartón; Lienzo sobre cartón
##                                      3                                     1
##                  Cartulina sobre lienzo                         Contrachapado
##                                      1                                     2
##                   Contrachapado; Lienzo                               Corcho
##                                      1                                     1
##                                Hojalata                        Lámina de cobre
##                                      8                                   146
##                                Lienzo               Lienzo al aguazo (sarga)
##                                  5396                                     4
##                  Lienzo pasado a tabla                 Lienzo pegado a lienzo
##                                      1                                    49
##                  Lienzo pegado a tabla                      Lienzo sin forrar
```

```
##                                     2                                    27
##                     Lienzo sobre cartón                     Lienzo sobre tabla
##                                    33                                    23
##               Lienzo sobre tabla; Tabla                                 Papel
##                                    23                                     2
##                  Papel pegado en cartón                Papel pegado en lienzo
##                                    21                                    43
##                   Papel pegado en tabla                       Papel verjurado
##                                     2                                     1
##       Papel; Tabla de roble del Báltico                Papel; Tabla; Hojalata
##                                     1                                     1
##                                 Piedra                               Pizarra
##                                     1                                     6
##                                  Raso                   Revestimiento mural
##                                     2                                     2
## Revestimiento mural trasladado a lienzo                                 Sarga
##                                    48                                     5
##                                 Tabla               Tabla de madera de álamo
##                                  1032                                     1
##           Tabla de madera de cedro rojo               Tabla de madera de chopo
##                                     1                                     5
##                Tabla de madera de nogal                Tabla de madera de pino
##                                     7                                    13
##                Tabla de madera de roble            Tabla de roble del Báltico
##                                    27                                    33
##                   Tabla pasada a lienzo                              Táblex
##                                    21                                     9
##                                 Vitela
##                                     1
```

En el caso de "soporte" vemos que son escasas las observaciones en que aparecen etiquetas multi-nivel, además es lógico interpretar que existe un soporte principal (aunque haya variaciones como montajes o transferencias), por lo que el sentido natural es comparar tipos de soporte y no su presencia. En consecuencia, transformaremos esta variable en un solo factor llamado "soporte_grp" agrupando sus niveles en un conjunto más reducido y de categorías más generales y proporcionando un nivel llamado 'ambiguo' para manejar aquellos niveles de soporte multi-etiqueta donde no se puede determinar el soporte principal. También se agregará una variable binaria llamada 'sop_montaje' para indicar si se trata de un soporte puro (0) o de un montaje (1), de esta manera podremos conservar información sobre este aspecto material relevante sin multiplicar categorías.

```
##                                 soporte soporte_grp sop_montaje
## 1 Revestimiento mural trasladado a lienzo       Mural           1
## 2 Revestimiento mural trasladado a lienzo       Mural           1
## 3 Revestimiento mural trasladado a lienzo       Mural           1
## 4 Revestimiento mural trasladado a lienzo       Mural           1
## 5 Revestimiento mural trasladado a lienzo       Mural           1
## 6 Revestimiento mural trasladado a lienzo       Mural           1
## Frecuencia de cada soporte:
##
##           Lienzo      Tabla/Panel            Metal     Papel/Vitela
##             5542             1151              154               70
##            Mural Cartón/Cartulina          Ambiguo   Piedra/Pizarra
##               50               33               27                7
##           Corcho
```

```
##                1
## Frecuencia de cada soporte puro/montaje:
##                    sop_montaje
## soporte_grp            0    1
##   Lienzo            5434  108
##   Tabla/Panel       1130   21
##   Metal              154    0
##   Papel/Vitela         4   66
##   Cartón/Cartulina    27    6
##   Mural                2   48
##   Piedra/Pizarra       7    0
##   Corcho               1    0
##   Ambiguo              0   27
```

Debido a la baja frecuencia observada en algunas clases, deberiamos volver a reagrupar estas minoritarias en un mismo grupo. Decidimos conservar 'mural' y 'papel/vileta' como últimos de frecuencia aceptable y devido a su potencial efecto sobre la respesta. Reagrupamos el resto en un nivel llamadp 'otros'.

```
##
##       Lienzo  Tabla/Panel        Metal Papel/Vitela        Otros        Mural
##         5542         1151          154           70           68           50
```

A continuación proporcionamos un listado con los niveles originales y a que nuevo nivel han sido asignados:

```
##                                    soporte  soporte_grp sop_montaje
## 1                                  Lienzo       Lienzo           0
## 2                 Lienzo al aguazo (sarga)       Lienzo           0
## 3                    Lienzo pasado a tabla       Lienzo           1
## 4                   Lienzo pegado a lienzo       Lienzo           1
## 5                    Lienzo pegado a tabla       Lienzo           1
## 6                        Lienzo sin forrar       Lienzo           0
## 7                      Lienzo sobre cartón       Lienzo           1
## 8                      Lienzo sobre tabla       Lienzo           1
## 9                                    Raso       Lienzo           0
## 10                                 Sarga       Lienzo           0
## 11                          Contrachapado  Tabla/Panel           0
## 12                                  Tabla  Tabla/Panel           0
## 13               Tabla de madera de álamo  Tabla/Panel           0
## 14          Tabla de madera de cedro rojo  Tabla/Panel           0
## 15               Tabla de madera de chopo  Tabla/Panel           0
## 16               Tabla de madera de nogal  Tabla/Panel           0
## 17                Tabla de madera de pino  Tabla/Panel           0
## 18               Tabla de madera de roble  Tabla/Panel           0
## 19             Tabla de roble del Báltico  Tabla/Panel           0
## 20                  Tabla pasada a lienzo  Tabla/Panel           1
## 21                                 Táblex  Tabla/Panel           0
## 22                               Hojalata        Metal           0
## 23                          Lámina de cobre        Metal           0
## 24                    Revestimiento mural        Mural           0
## 25 Revestimiento mural trasladado a lienzo        Mural           1
## 26                                  Papel Papel/Vitela           0
## 27                  Papel pegado en cartón Papel/Vitela           1
## 28                  Papel pegado en lienzo Papel/Vitela           1
```

```
## 29                   Papel pegado en tabla Papel/Vitela           1
## 30                        Papel verjurado Papel/Vitela           0
## 31                                Vitela Papel/Vitela           0
## 32                                Cartón        Otros           0
## 33                  Cartón sobre lienzo        Otros           1
## 34                   Cartón sobre tabla        Otros           1
## 35            Cartón; Lienzo sobre cartón        Otros           1
## 36                 Cartulina sobre lienzo        Otros           1
## 37                  Contrachapado; Lienzo        Otros           1
## 38                                Corcho        Otros           0
## 39              Lienzo sobre tabla; Tabla        Otros           1
## 40      Papel; Tabla de roble del Báltico        Otros           1
## 41                  Papel; Tabla; Hojalata        Otros           1
## 42                                Piedra        Otros           0
## 43                               Pizarra        Otros           0
```

Recodificación de variable "fecha"

La datación de las pinturas es muy inexacta y provoca que tengamos fechas de muschos formatos distintos. La complejidad de esta avriable escapa a nuestras posibilidades pero pensamos que es realmente valiosa, por lo que hemos recurrido a insteligencias artificiales para generar el codigo a continuación que nos permita reexplicar la información de la siguiente manera:

```
##        fecha fecha_tipo fecha_inicio fecha_fin fecha_est fecha_ancho
## 1  Siglo XII    century         1101      1200      1150          99
## 2  Siglo XII    century         1101      1200      1150          99
## 3  Siglo XII    century         1101      1200      1150          99
## 4  Siglo XII    century         1101      1200      1150          99
## 5  Siglo XII    century         1101      1200      1150          99
## 6  Siglo XII    century         1101      1200      1150          99
## 7  Siglo XII    century         1101      1200      1150          99
## 8  Siglo XII    century         1101      1200      1150          99
## 9  Siglo XII    century         1101      1200      1150          99
## 10 Siglo XII    century         1101      1200      1150          99
## 11 Siglo XII    century         1101      1200      1150          99
## 12 Siglo XII    century         1101      1200      1150          99
##                                                        fecha
## 1                                                   Siglo XII
## 101                                                1445 - 1460
## 201                                                 Hacia 1495
## 301                                                  Siglo XVI
## 401                                                 Hacia 1510
## 502                                                1525 - 1530
## 602                                                1543 - 1550
## 702 Segunda mitad del siglo XVI - Primer tercio del siglo XVII
## 803                                                        1566
## 903                                                 Hacia 1586
##             fecha_tipo fecha_inicio fecha_fin fecha_est fecha_ancho
## 1              century         1101      1200      1150          99
## 101         year_range         1445      1460      1452          15
## 201              circa         1495      1495      1495           0
## 301             century         1501      1600      1550          99
## 401              circa         1510      1510      1510           0
## 502         year_range         1525      1530      1528           5
```

```
## 602         year_range         1543      1550      1546           7
## 702 century_part_range         1551      1633      1592          82
## 803         year_exact         1566      1566      1566           0
## 903              circa         1586      1586      1586           0
```

Ahora con estas nuevas variables podemos tomar una decisión teniendo en cuenta nuestras necesidades y el posterior modelaje. Nos decantamos por mantener las variables "fecha" original, como información de posible interés futuro, "anio_est" como una aproximación y "fecha_ancho" que nos permitirá almacenar ifnormación sobre la anterior aproximación, ya que en muchos casos contiene mucha variabilidad. Utilizando estas dos últimas variables en un modelo podremos hacer que este no interprete como iguales observaciones que comparten el año estimado si su rango es distinto.

Echamos un primer vistazo a la distribución de esta nueva variable "fecha_ancho" y nos damos cuenta de algo interesante:

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 4000  259  209  101   60  231   59   53   19   15  127    1    3    2    4   12
##   16   19   20   21   22   24   25   28   30   32   33   34   35   37   38   39
##    5  123   23    1   11   88   26    2   14   52   26    5   10    1    1   24
##   40   44   48   49   50   52   53   57   58   69   71   74   79   82   83   84
##    4    3    9  222   12    9    3    4    1   24    1    5    1    1    1    1
##   85   99  111  113  119  129  130  155  171  174  184  191  199  214  216  231
##    2 1039    1    1    3    1    1    1    1    1    1    1   33    1    1    1
##  234  235  269  274  276  279  299  310  311  327  341  342  344  349  361  363
##    1    1    1    1    1    4    2    1    1    1    1    1    2    2    1    1
##  364  365  367  369  374  377  385  387  388  399  404  411  424  425  434  445
##    1    1    1    3    1    1    1    1    2    1    1    1    1    1    1    5
##  446  448  460  470  478  480  494  511  517  518  519  528  534  547  548  560
##    1    3    2    1    1    1    1    2    1    2    1    1    1    1    1    1
##  568  587  610  620  643  644  785  806
##    1    1    1    4    2    1    1    1
```

Vemos que la distribución de esta avriable está fuertemente influenciada por algunos picos, esto deberá tenerse ne cuenta en la modleización y se extenderá en el análisis descriptivo

Recodificación de variable "procedencia"

visualizarmeos algunos casos:

```
## [1] "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948"
## [2] "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948."
## [3] "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948."
## [4] "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948."
## [5] "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948."
## [6] "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948."
## [1] "Posible donación del autor; Museo Español de Arte Contemporáneo; Museo Nacional Centro de Arte R
## [2] "Donación de Isabel Verdejo, viuda de Ramón Gaya, al Museo del Prado, 2024"
## [3] "Donación Michel Alexander Werboff, 1960"
## [4] "Donación de los herederos del pintor para el Museo de Arte Moderno, 1961; Museo Español de Arte
## [5] "Donación de los herederos del pintor con destino al Museo de Arte Moderno, 1961; Museo de Arte R
## [6] "Museo Español de Arte Contemporáneo, hasta 1995; Museo Nacional Centro de Arte Reina Sofía, 201
```

Podemos observar que no existe ningun tipo de estructura que nos sea de utilidad en esta variable, además simplemente aporta información de la posesión de la obra, no de su lugar de producción, por lo que no la

consideramos una característica propia de esta. Decidimos prescindir de esta variable, no se transformará para ver utilizada pero se conservará en la base de datos como mera información consultable.

Recodificación de variable "titulo"

Haremos uso de esta avriable para hacer una aproximación de la temática de cada obra utilizando sus palabras más frecuentes

```
## palabras
##       san   paisaje    virgen     maria   retrato      juan     santa      nino
##       907       378       288       205       195       194       192       163
## francisco    cristo   bodegon     reina     pedro    felipe    carlos       rey
##       156       153       150       112       110       107       103        94
##   familia    isabel   florero  fernando   antonio    borbon     jesus    pintor
##        90        84        83        81        79        78        77        76
## adoracion  bautista   austria      jose    espana caballero
##        72        71        69        69        66        65
```

La idea es crear un diccionario que almacene palabras clave de cada temática creada y asigne el tema correspondiente en la nueva variable "tema". Este proceso se ha repetido reiteradamente hasta lograr una clasificación satisfactoria, ya que en los primeros intentos un gran porcentaje de los casos caía en la categoría 'otros' por no ser clasificable. En cada iteración se obtubieron de nuevo las palabras más repetidas de este nivel 'otros' y se incluian en el diccionario para actualizarlo. Finalmente se ha aplicado el definitivo obteniendo las seguientes temáticas, con sus respectivas frecuencias:

```
##
##         religioso             otros     retrato_corte   paisaje_lugares
##              2512              1216               995               884
##    vida_cotidiana    bodegon_floral historia_alegoria         mitologia
##               389               325               258               227
##      proceso_obra     caza_animales
##               123               106
```

Variables no analizables

Finalmente listamos las variables que se han conservado en la abse de datos a modo de posibles consultas futuras, pero que no participaran de ningun modo en el analisis: "numero_catalalogo", "titulo", "url", "nombre_autor", "procedencia", "fecha".

3.3 Creación de nuevas variables

En la sección anterior hemos transformado y también creado algunas variable con el fin único de reestructurar la información ya presente. Sin embargo, en esta sección podrecederemos a de alguna manera "generar" nueva información, que consideramos de posible utilidad, a partir de la que ya disponemos.

Variable respuesta

Para comenzar generaremos nuestra variabe objetivo dele studio que es aquella relacionada con la proporción aurea. Haremos uso de las variables de dimension "alto" y "ancho" para crear la que será nuestra avriable respuesta llamada "exito". Esta será un indicador binaria que determinará si la pintura sigue o no la proporción aurea. El error aceptado es del 5% y en todo momento se utiliza el largo largo como numerador.

```
##  Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
```

Variable "orientacion"

A partir de las variables de dimensión "alto" y "ancho" creamos una nueva variable que nos indique si la forma de la pintura es vertical, horizontal o cuadrada.

```
##
##   cuadrado horizontal   vertical
##         33       3001       4001
```

Variables "area" y "tam_cat"

También trabajamos con las dimensiones para crear 2 variables relacionadas con el tamaño: "area", para conocer su superficie, y "tam_cat" para clasificarlas en categorías de tamaño "pequeño/mediano/grande".

Aunque los niveles de tamaño son ordenados por definición, decidimos no establecer orden en el factor, por simplicidad de interpretación.

3.4 Manejo de desbalances

Presentaremos las frecuencias de cada variable factor:

```
##
## ====================
## Variable: tam_cat
##
## pequeno mediano  grande
##    2322    2324    2389
##
## ====================
## Variable: orientacion
##
##   cuadrado horizontal   vertical
##         33       3001       4001
##
## ====================
## Variable: soporte_grp
##
##       Lienzo  Tabla/Panel        Metal        Mural Papel/Vitela        Otros
##         5542         1151          154           50           70           68
##
## ====================
## Variable: sop_montaje
##
##    0    1
## 6759  276
##
## ====================
## Variable: tecnica
##
## acuosas  fresco   mixta    oleo   otras
##      65      36     123    6802       9
##
## ====================
## Variable: tipo_autor
```

```
##
## anonimo  hombre   mujer  varios
##     661    6161      59     154
##
## ====================
## Variable: serie
##
##    0    1
## 5628 1407
##
## ====================
## Variable: tema
##
##    bodegon_floral     caza_animales historia_alegoria         mitologia
##               325               106               258               227
##             otros   paisaje_lugares      proceso_obra         religioso
##              1216               884               123              2512
##     retrato_corte    vida_cotidiana
##               995               389
```

Para detectar roblemas reales debemos fijarnos no en el conteo marginal, sino cruzado con la variable objetivo. Fijaremos como indicio de problema cuando el recuento de "exito"='1' sea inferior a 5 en una categoría.

```
##
## *** POSIBLE PROBLEMA (min celda < 5 ) en: orientacion ***
##
## y   cuadrado horizontal vertical
##   0       33       2559     3594
##   1        0        442      407
##
## *** POSIBLE PROBLEMA (min celda < 5 ) en: soporte_grp ***
##
## y   Lienzo Tabla/Panel Metal Mural Papel/Vitela Otros
##   0   4843        1024   146    44           64    65
##   1    699         127     8     6            6     3
##
## *** POSIBLE PROBLEMA (min celda < 5 ) en: tecnica ***
##
## y   acuosas fresco mixta oleo otras
##   0      60     32   109 5976     9
##   1       5      4    14  826     0
```

Vemos que las únicas variables realmente problematicas son "orientacion" con el nivel 'cuadrado', "soporte_grp" con el nivel 'otros' y "tecnica" con los niveles 'acuosas', 'fresco' y 'otras'.

Comenzando con "orientación" podemos ver que es obvio el hecho de que en el nivel 'cuadrado' no va a aparecer ningún éxito por definición: si es cuadrado es imposible que siga la proporción aurea. No podemso simplemente eliminar estos casos, aunque sean pocos, porque estariamos generando un sesgo sistemático, pero lo que sí haremos será aplicar esta regla en el ajuste. Es decir, los cuadrados siempre serán catalogados como exito=0 y las interpretaciones se haran teniendo en cuenta la modificación en la inferencia, solo podremos inferir en aquellas obras no cuadradas.

Las dos otras variables se solucionarán simplemente reagrupando sus niveles. Para "soporte_grp" reagruparemos 'Papel/Vitela' en 'Otros', y para "tecnica" reagruparemos únicamente en 'oleo', 'mixta' y 'otras'.

```
##
## ====================
## Frecuencias margincales:
##      Lienzo       Metal       Mural       Otros Tabla/Panel
##        5542         154          50         138        1151
##
## mixta  oleo otras
##   123  6802   110
##
## ====================
## Frecuencias cruzadas:
##     Lienzo Metal Mural Otros Tabla/Panel
##   0   4843   146    44   129        1024
##   1    699     8     6     9         127
##
##     mixta oleo otras
##   0   109 5976   101
##   1    14  826     9
```

BASE DE DATOS FINAL

Finalmente obtenemos la base de datos ya tratada ('df_completa') y una extracción solo con las variables útiles para el análisis ('df')

```
## 'data.frame':    7035 obs. of  21 variables:
##  $ numero_catalogo: chr  "P007269" "P007270" "P007271" "P007272" ...
##  $ titulo         : chr  "Parte superior del Pantocrátor sostenido por cuatro ángeles. Pintura mural
##  $ url            : chr  "https://www.museodelprado.es/coleccion/obra-de-arte/parte-superior-del-pan
##  $ nombre_autor   : chr  "Anónimo" "Anónimo" "Anónimo" "Anónimo" ...
##  $ procedencia    : chr  "Ermita de la Vera Cruz, Maderuelo, Segovia, 1948" "Ermita de la Vera Cruz,
##  $ fecha          : chr  "Siglo XII" "Siglo XII" "Siglo XII" "Siglo XII" ...
##  $ alto           : num  249 249 200 200 185 185 200 200 140 181 ...
##  $ ancho          : num  327 327 263 185 243 220 200 245 75 283 ...
##  $ exito          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ area           : num  81423 81423 52600 37000 44955 ...
##  $ tam_cat        : Factor w/ 3 levels "pequeno","mediano",..: 3 3 3 3 3 3 3 3 2 3 ...
##  $ orientacion    : Factor w/ 3 levels "cuadrado","horizontal",..: 2 2 2 3 2 2 1 2 3 2 ...
##  $ soporte        : chr  "Revestimiento mural trasladado a lienzo" "Revestimiento mural trasladado a
##  $ soporte_grp    : Factor w/ 5 levels "Lienzo","Metal",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ sop_montaje    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ tecnica        : Factor w/ 3 levels "mixta","oleo",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ tipo_autor     : Factor w/ 4 levels "anonimo","hombre",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ serie          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ fecha_est      : int  1150 1150 1150 1150 1150 1150 1150 1150 1150 1150 ...
##  $ fecha_ancho    : int  99 99 99 99 99 99 99 99 99 99 ...
##  $ tema           : Factor w/ 10 levels "bodegon_floral",..: 8 8 8 8 8 8 8 8 8 8 ...
## 'data.frame':    7035 obs. of  12 variables:
##  $ exito      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ area       : num  81423 81423 52600 37000 44955 ...
##  $ tam_cat    : Factor w/ 3 levels "pequeno","mediano",..: 3 3 3 3 3 3 3 3 2 3 ...
##  $ orientacion: Factor w/ 3 levels "cuadrado","horizontal",..: 2 2 2 3 2 2 1 2 3 2 ...
##  $ soporte_grp: Factor w/ 5 levels "Lienzo","Metal",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ sop_montaje: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ tecnica    : Factor w/ 3 levels "mixta","oleo",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
##  $ tipo_autor : Factor w/ 4 levels "anonimo","hombre",..: 1 1 1 1 1 1 1 1 1 ...
##  $ serie      : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ fecha_est  : int  1150 1150 1150 1150 1150 1150 1150 1150 1150 1150 ...
##  $ fecha_ancho: int  99 99 99 99 99 99 99 99 99 99 ...
##  $ tema       : Factor w/ 10 levels "bodegon_floral",..: 8 8 8 8 8 8 8 8 8 ...
```

Por último, comprobamos no haber introducido algún dato faltante:

```
##       exito        area     tam_cat orientacion soporte_grp sop_montaje
##           0           0           0           0           0           0
##     tecnica  tipo_autor       serie   fecha_est fecha_ancho        tema
##           0           0           0           0           0           0
```

4. Análisis descriptivo

Empezamos el análisis descriptivo construyendo para cada variable, una tabla de frecuencias y el gráfico de barras correspondiente, la forma más directa de entender cómo se reparte la información dentro del conjunto de datos. Antes de comparar con la variable respuesta (exito) o plantear cualquier modelo, necesitamos saber qué categorías aparecen, con qué peso y si existe algún patrón evidente en la composición de la muestra.

Como control de calidad permite detectar problemas típicos en datos obtenidos por web scraping, como etiquetas duplicadas, niveles inesperados o valores codificados de manera especial.

Las frecuencias son necesarias para interpretar correctamente lo que venga después. Si tenemos una categoría muy mayoritaria cualquier resultado puede estar condicionado por esa dominancia. Tener la proporción de cada nivel ayuda a contextualizar comparaciones; no es lo mismo observar un patrón en una categoría con miles de obras que en otra con muy pocas.

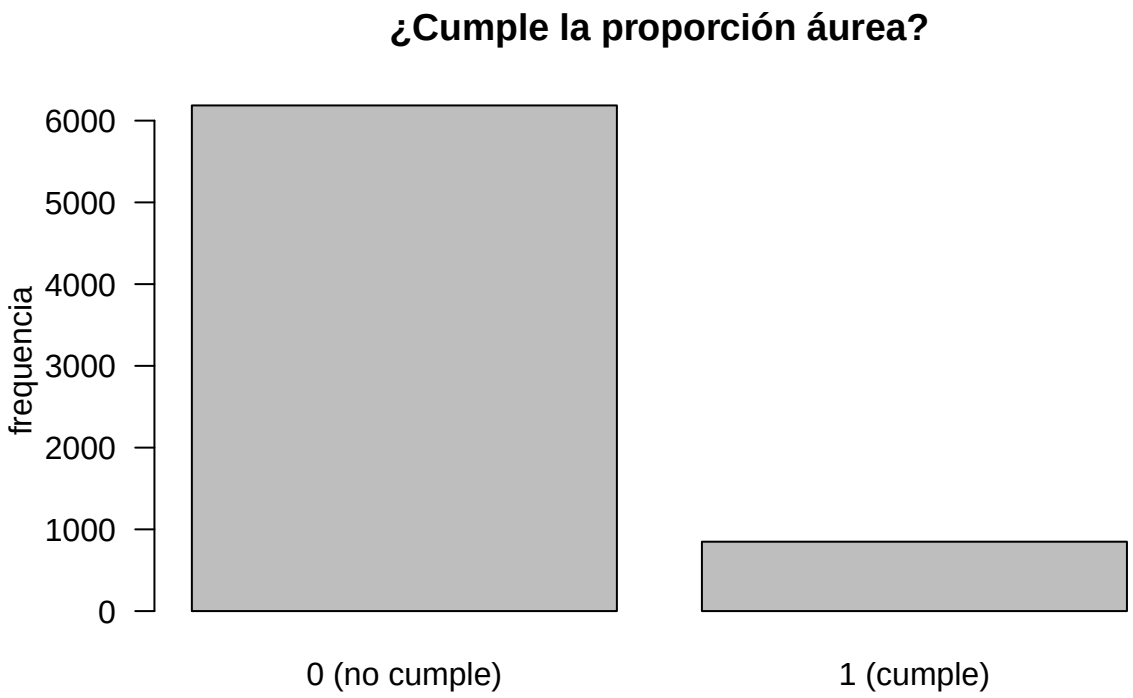
Las clases minoritarias aun que ess habitual que algunas categorías tengan pocas observaciones.Si tenemos demasiados niveles con baja frecuencia hacen que los gráficos sean ilegibles y no podamos extraer conclusiones.

En resumen, trabajar variable por variable es una etapa imprescindible porque permite verificar la coherencia de las variables, comprender la estructura real de los datos y anticipar problemas derivados de categorías raras. Esto asegura que las comparaciones posteriores con exito sean interpretables, estables y defendibles.

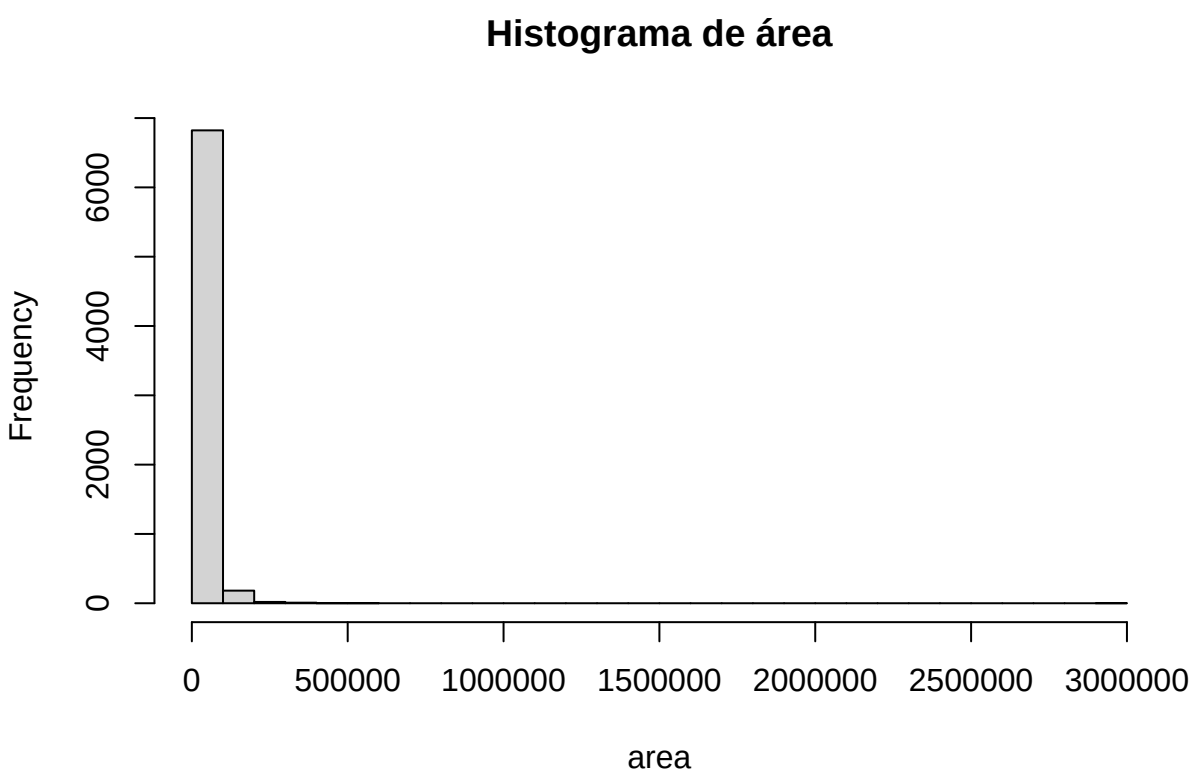
4.1 Variables individuales

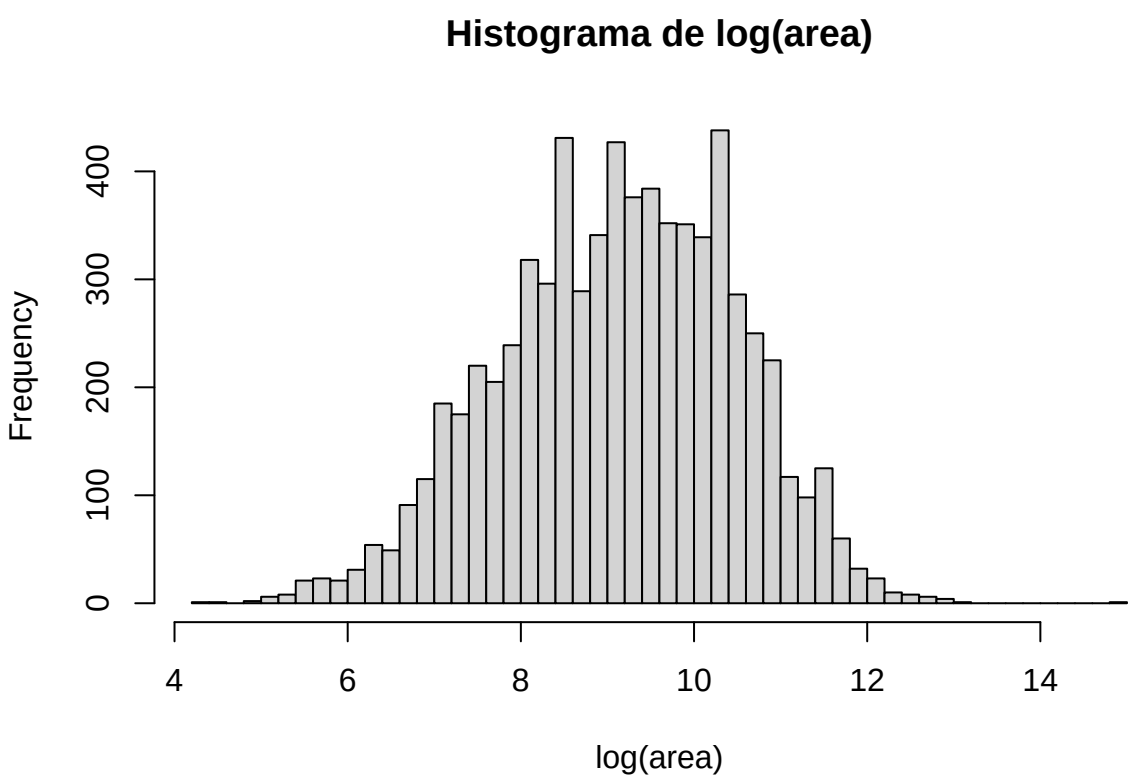
EXITO

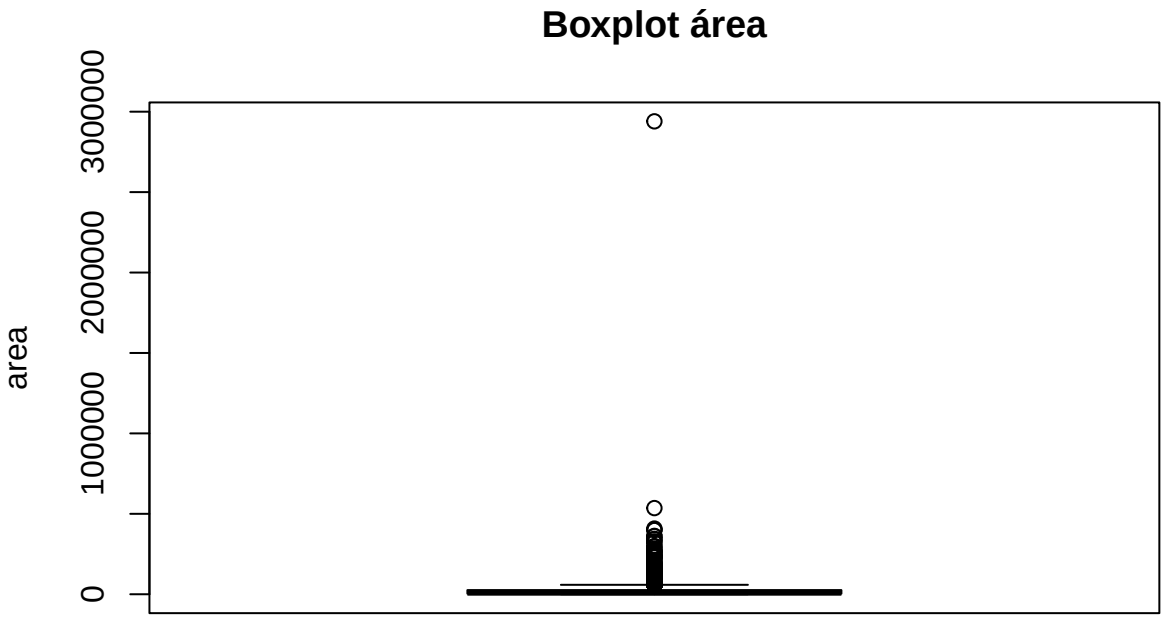
```
##
##    0    1
## 6186  849
```

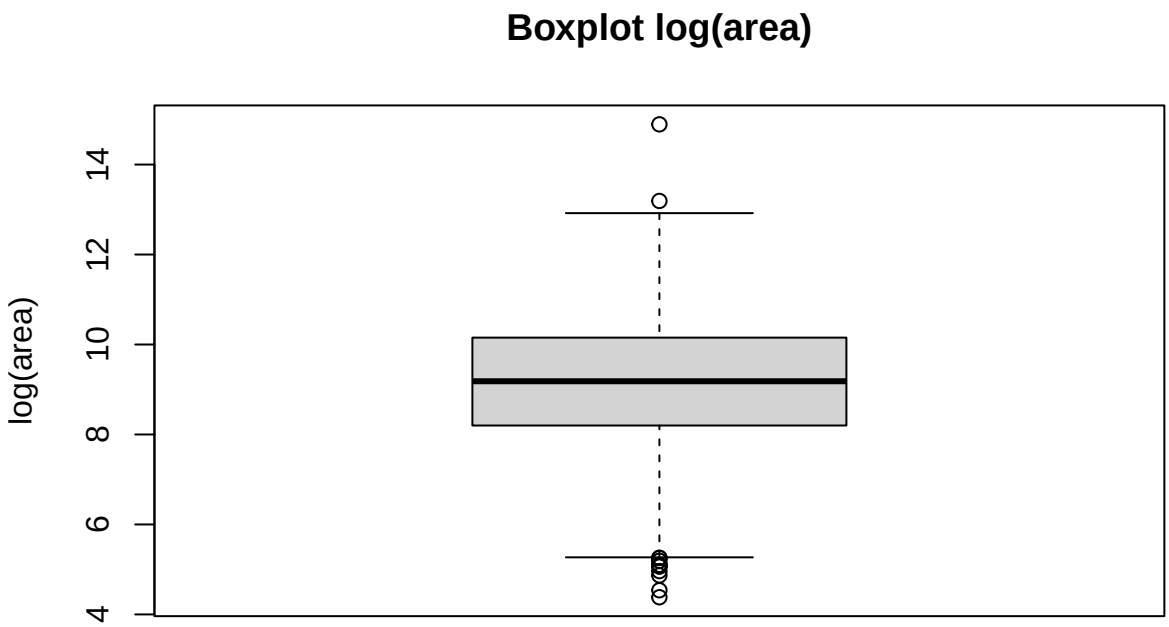

¿Cumple la proporción áurea?
6000
5000
4000
3000
2000
1000
0
frequencia
0 (no cumple)
1 (cumple)

AREA


Histograma de área
Frequency
0
2000
4000
6000
0
500000
1000000
1500000
2000000
2500000
3000000
area


Histograma de log(area)
Frequency
0
100
200
300
400
4
6
8
10
12
14
log(area)

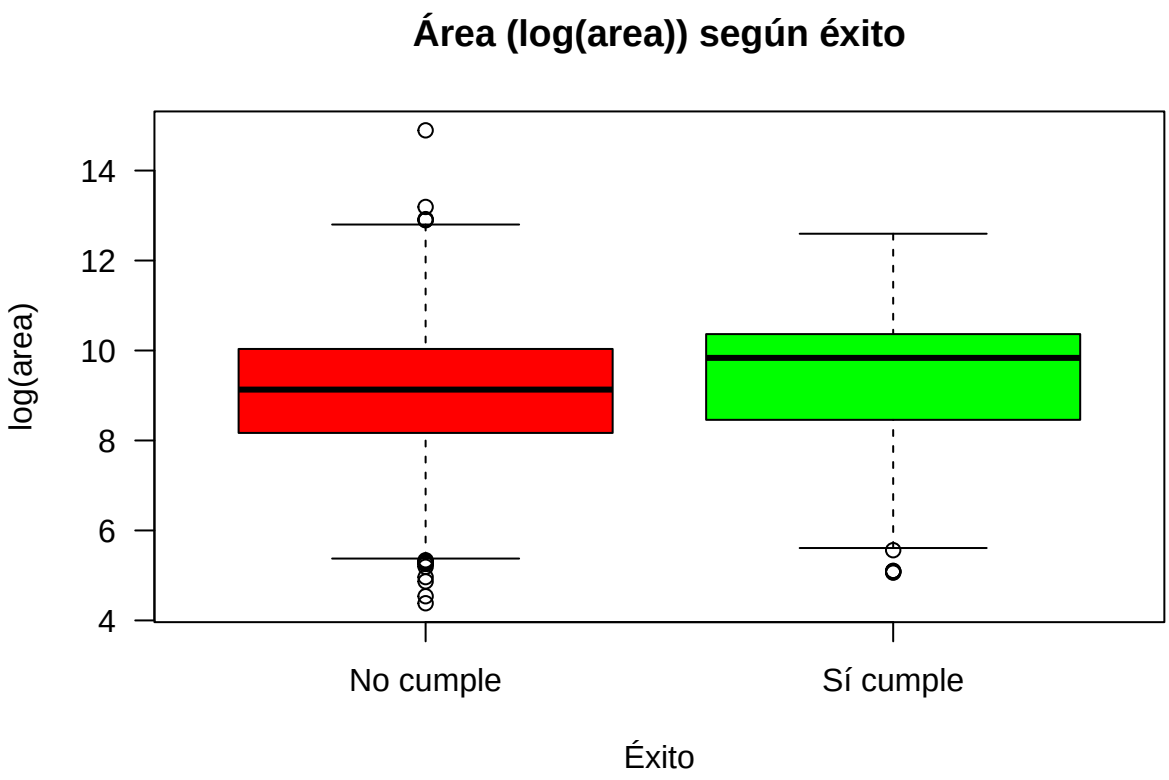

Boxplot área
area
0
1000000
2000000
3000000

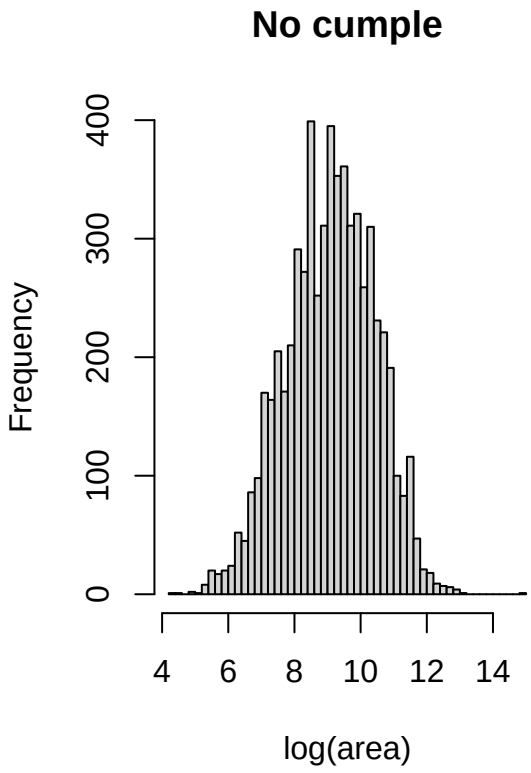


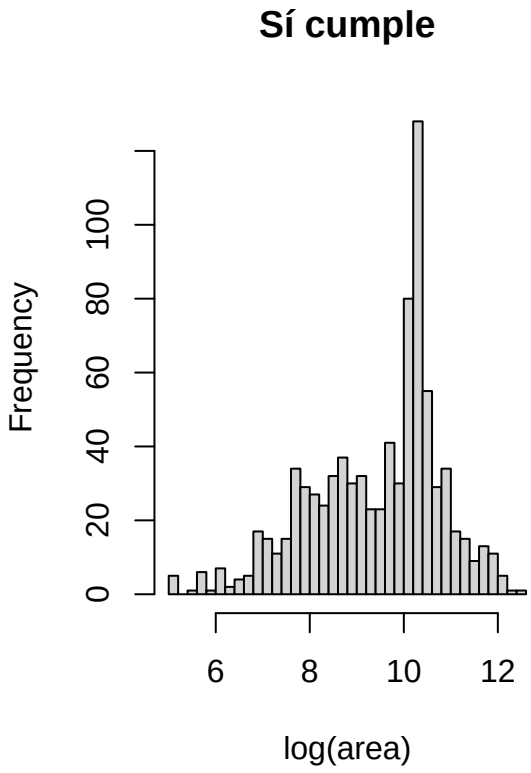
La variable area esta muy sesgada a la derecha: la mayoría de obras tienen áreas pequeñas y hay pocas con áreas enormes (por eso el histograma "normal" queda aplastado y el boxplot de area sale con mil outliers).

Al hacer log(area) la distribución queda mucho más "normalita" (campana) y el boxplot se vuelve legible.

AREA CON ÉXITO

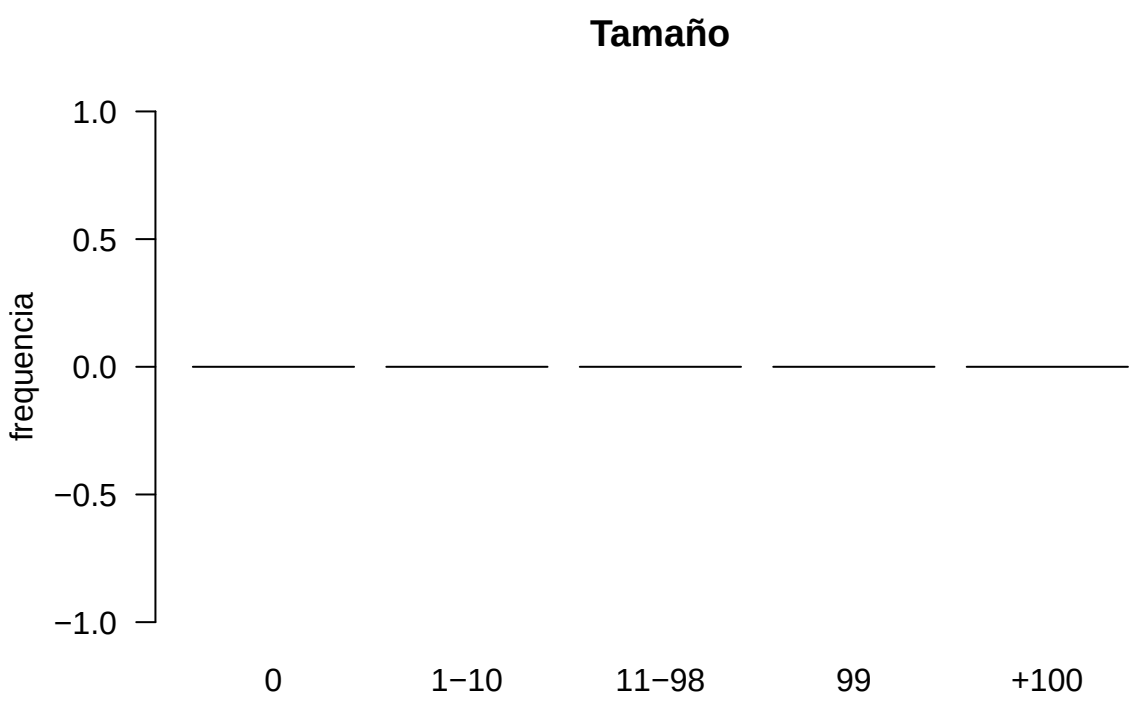

Área (log(area)) según éxito
log(area)
14
12
10
8
6
4
No cumple
Sí cumple
Éxito



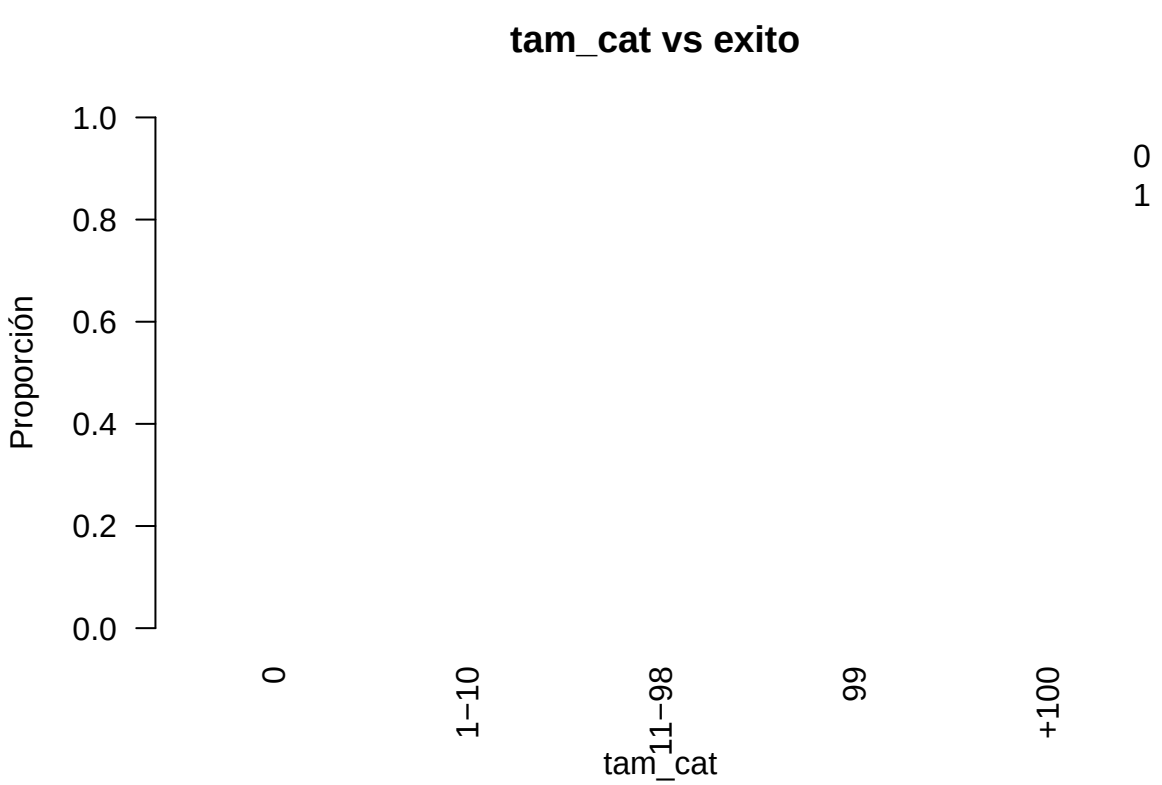


TAM_CAT

```
##
##     0  1-10 11-98    99  +100
##     0     0     0     0     0
```

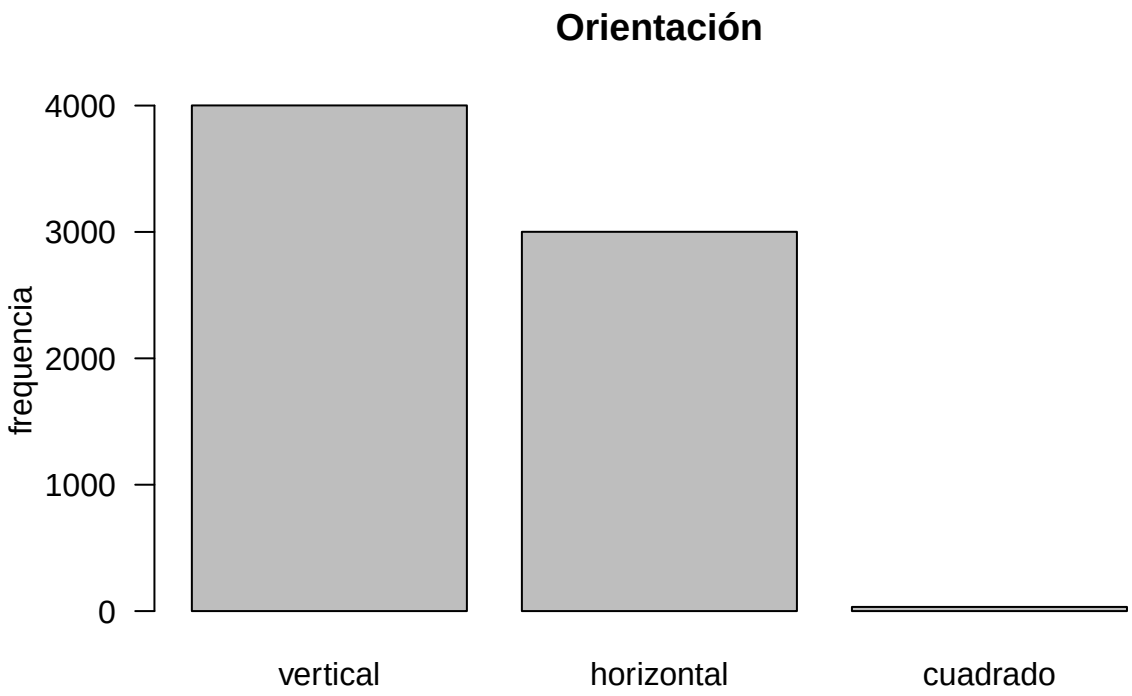


TAM_CAT CON ÉXITO

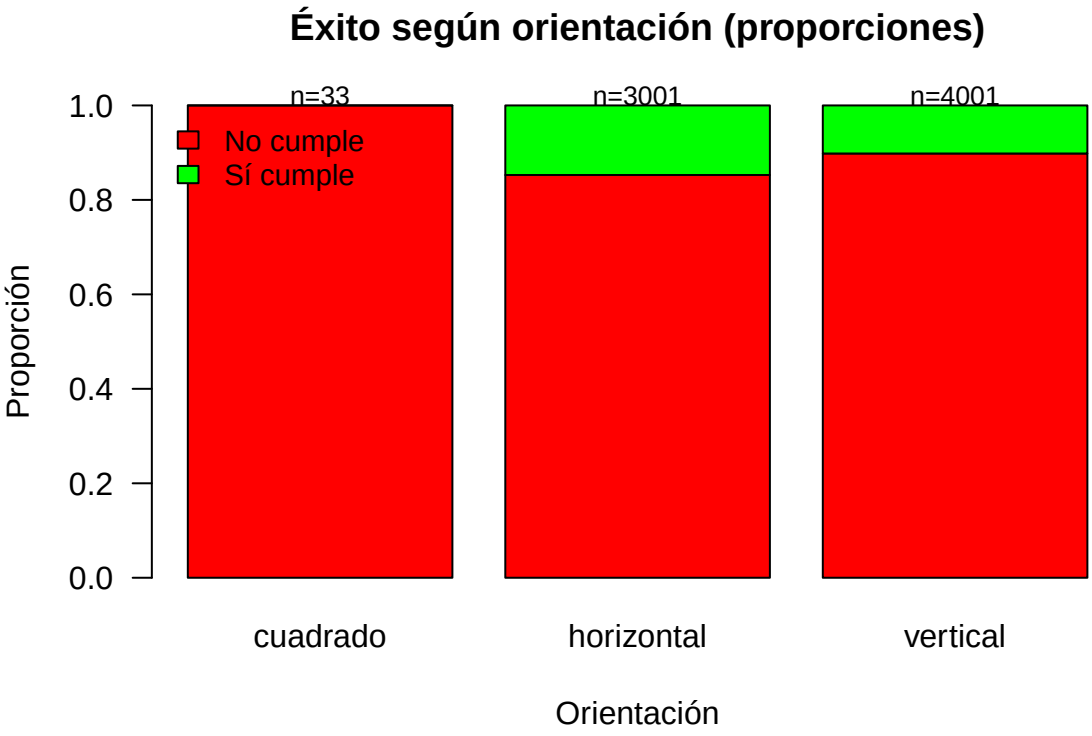


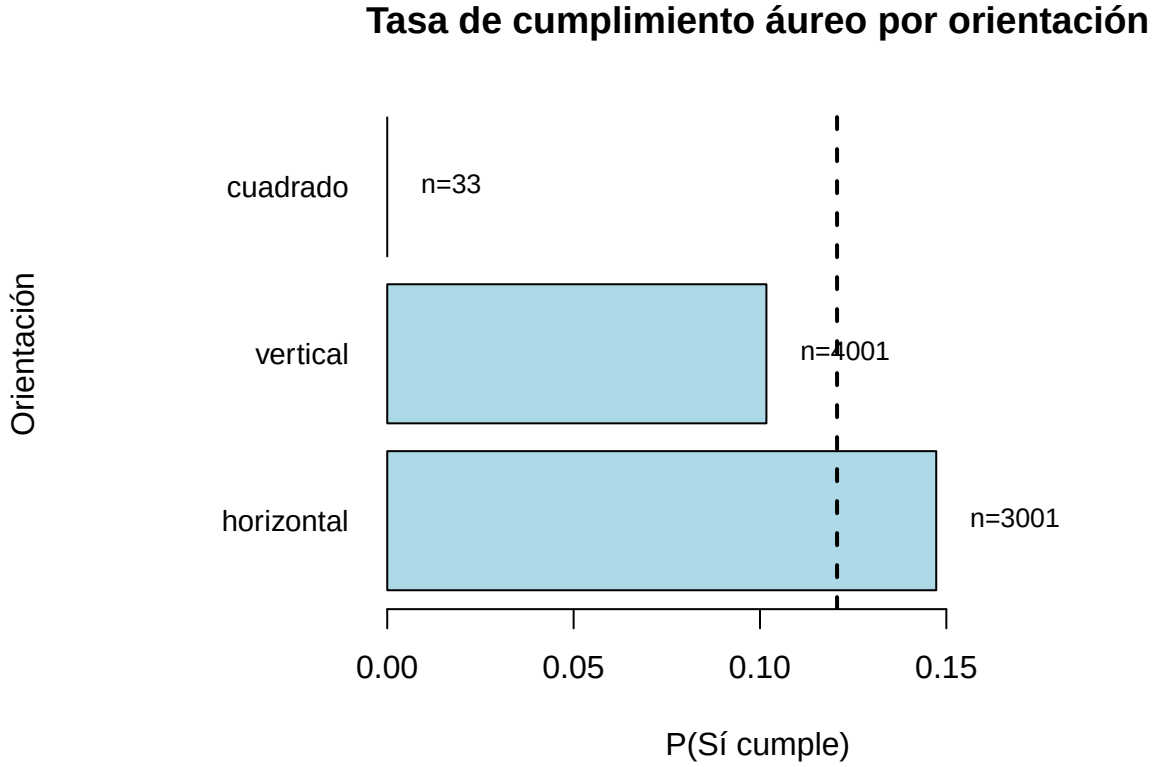
ORIENTACIÓN

```
##
##   cuadrado horizontal   vertical
##         33       3001       4001
```



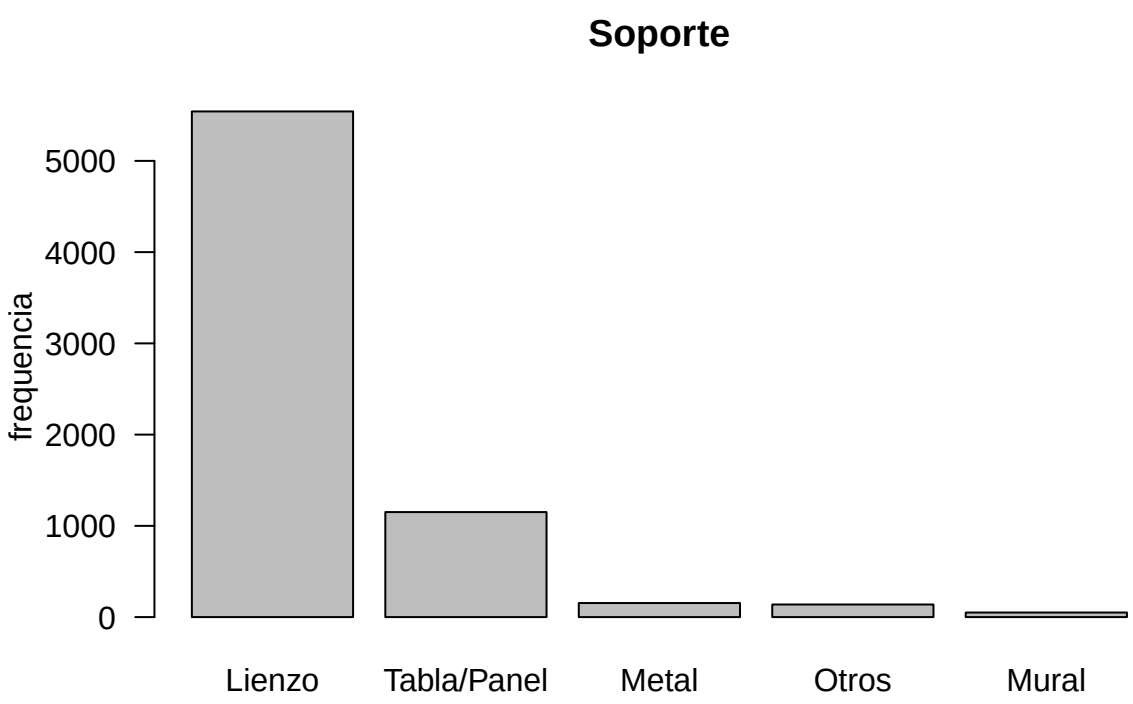
EXITO SEGÚN ORIENTACIÓN


Éxito según orientación (proporciones)
n=33
n=3001
n=4001
No cumple
Sí cumple
Proporción
1.0
0.8
0.6
0.4
0.2
0.0
cuadrado
horizontal
vertical
Orientación



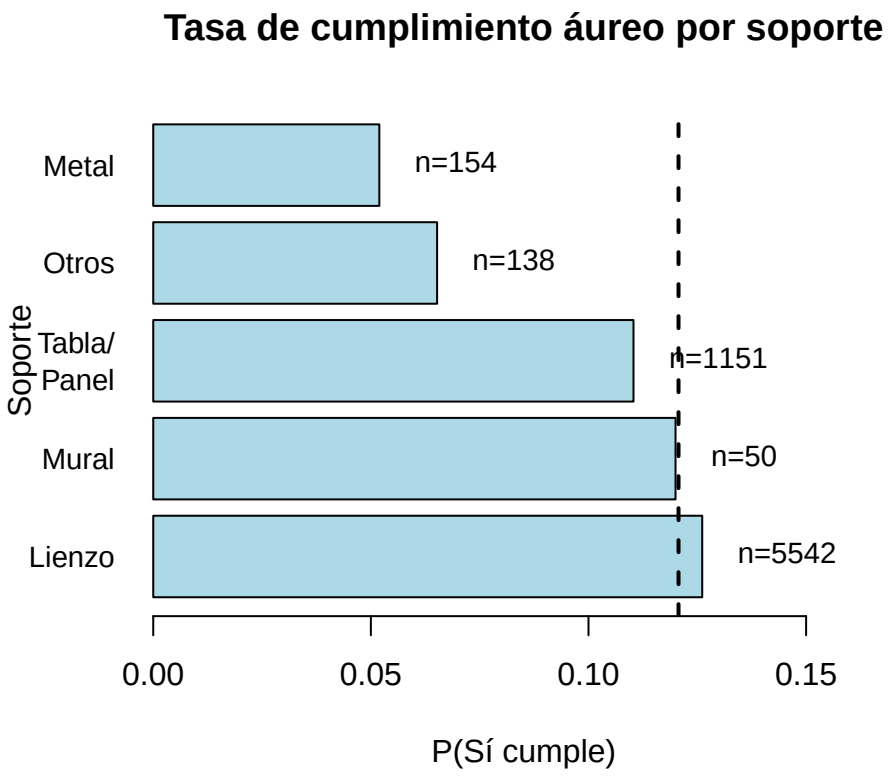
SOPORTE_GRP

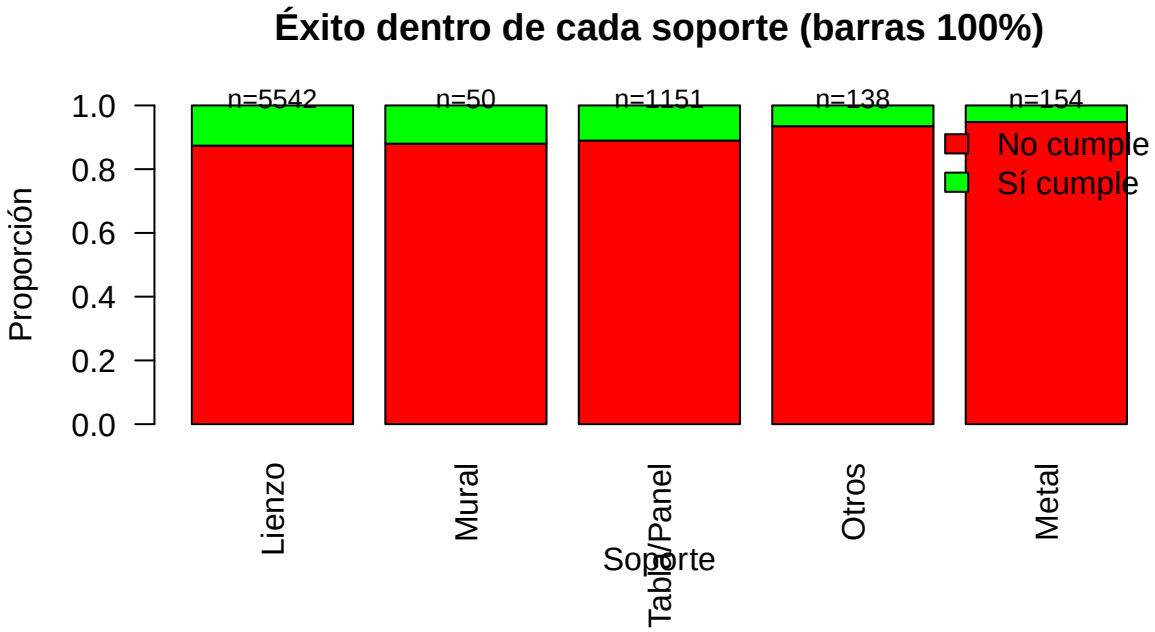
```
##
##      Lienzo       Metal       Mural       Otros Tabla/Panel
##        5542         154          50         138        1151
```



SOPORTE_GRP CON EXITO

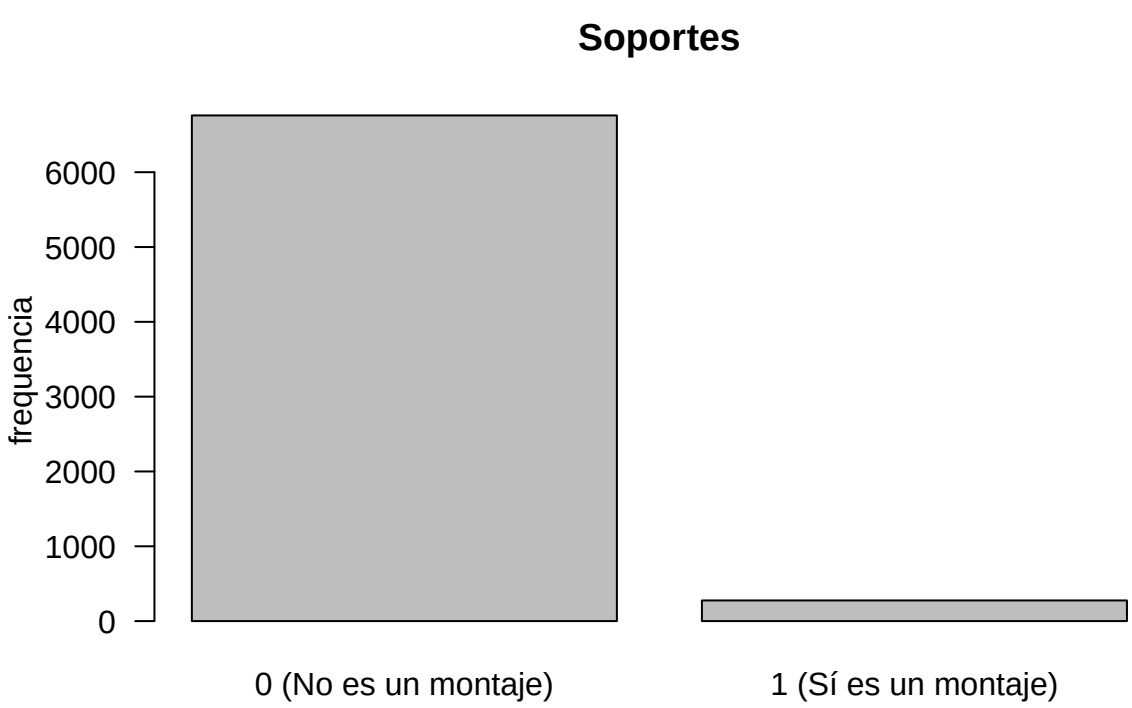
```
## integer(0)
```


Tasa de cumplimiento áureo por soporte
Soporte
Metal
Otros
Tabla/
Panel
Mural
Lienzo
n=154
n=138
n=1151
n=50
n=5542
0.00
0.05
0.10
0.15
P(Sí cumple)



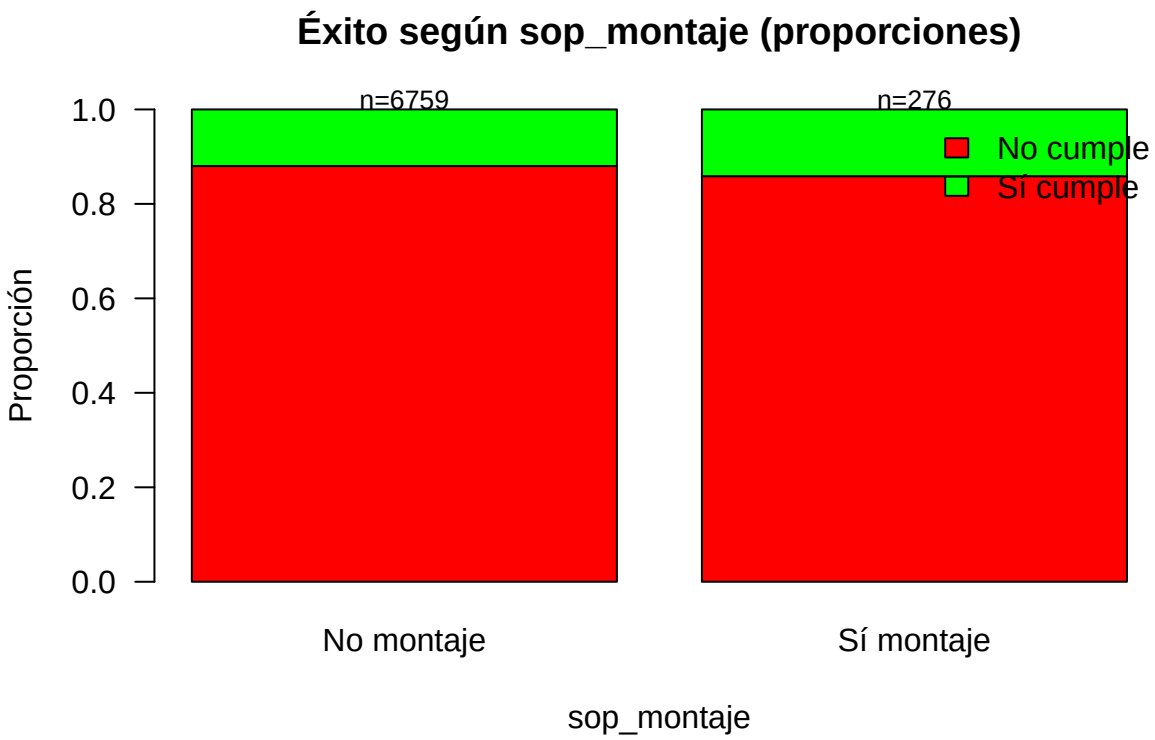
SOP_MONTAJE

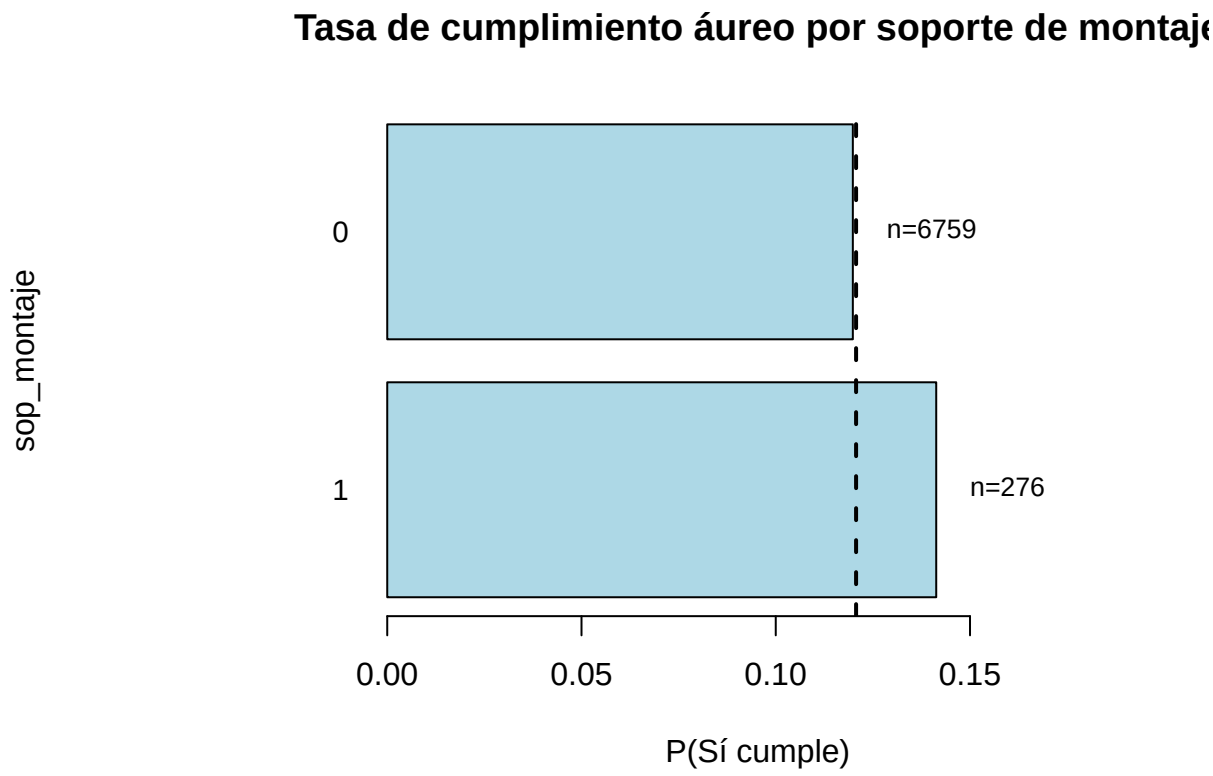
```
##
##    0    1
## 6759  276
```



SOP_MONTAJE Y EXITO

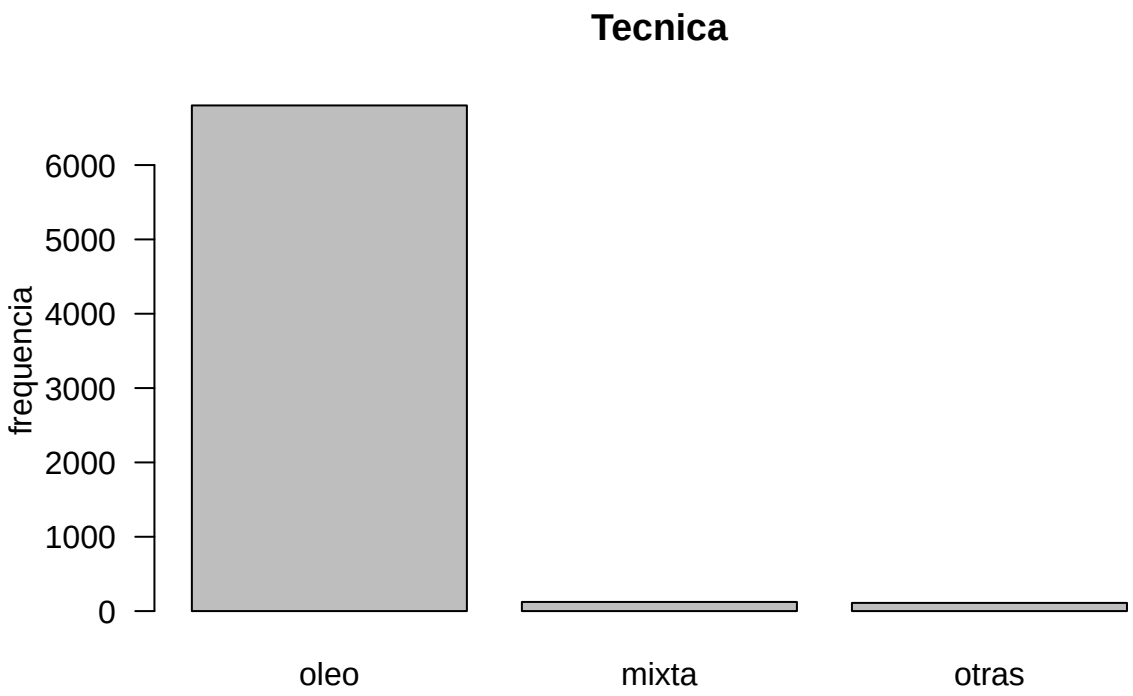
```
##
##              No cumple Sí cumple
##   No montaje      5949       810
##   Sí montaje       237        39
##
##              No cumple Sí cumple
##   No montaje     0.880     0.120
##   Sí montaje     0.859     0.141
```


Éxito según sop_montaje (proporciones)
n=6759
n=276
No cumple
Sí cumple
1.0
0.8
0.6
0.4
0.2
0.0
Proporción
No montaje
Sí montaje
sop_montaje



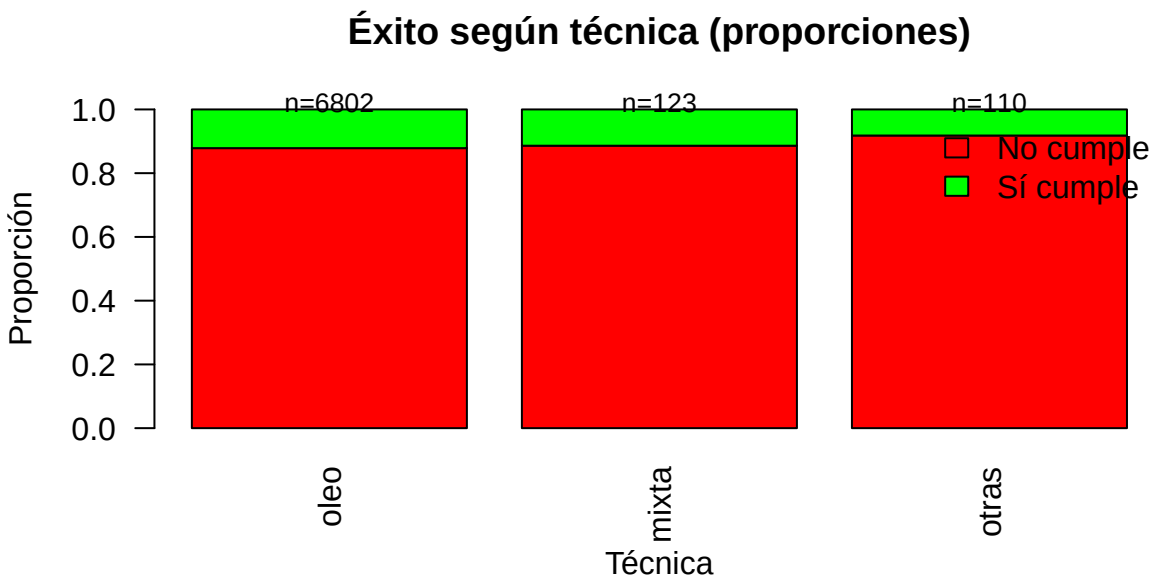
TECNICA

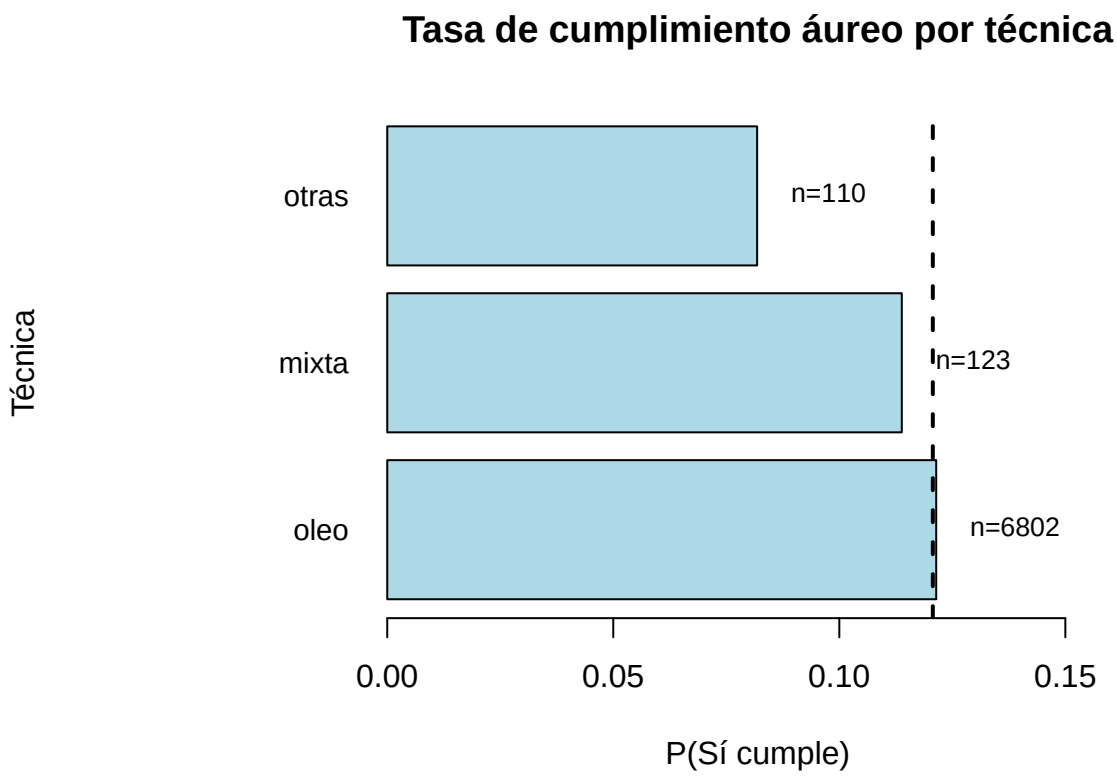
```
##
## mixta  oleo otras
##   123  6802   110
```



TECNICA CON EXITO

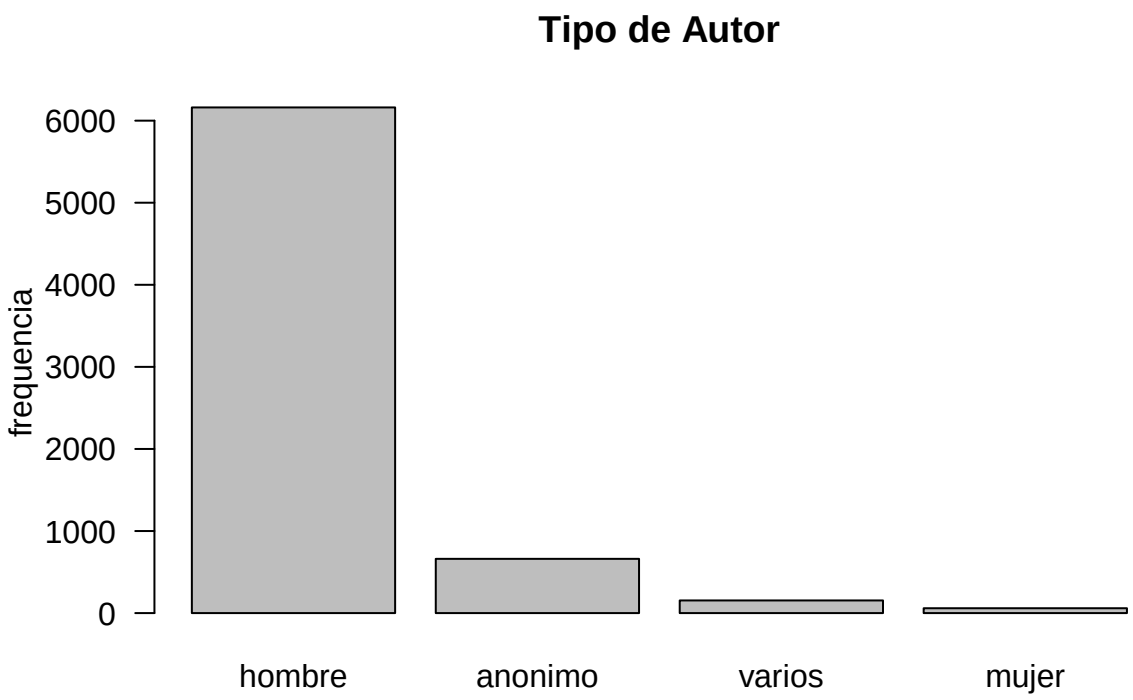
```
##
##         No cumple Sí cumple
##   mixta       109        14
##   oleo       5976       826
##   otras       101         9
##
##         No cumple Sí cumple
##   oleo      0.879     0.121
##   mixta     0.886     0.114
##   otras     0.918     0.082
```


Éxito según técnica (proporciones)
n=6802
n=123
n=110
No cumple
Sí cumple
Proporción
1.0
0.8
0.6
0.4
0.2
0.0
oleo
mixta
otras
Técnica



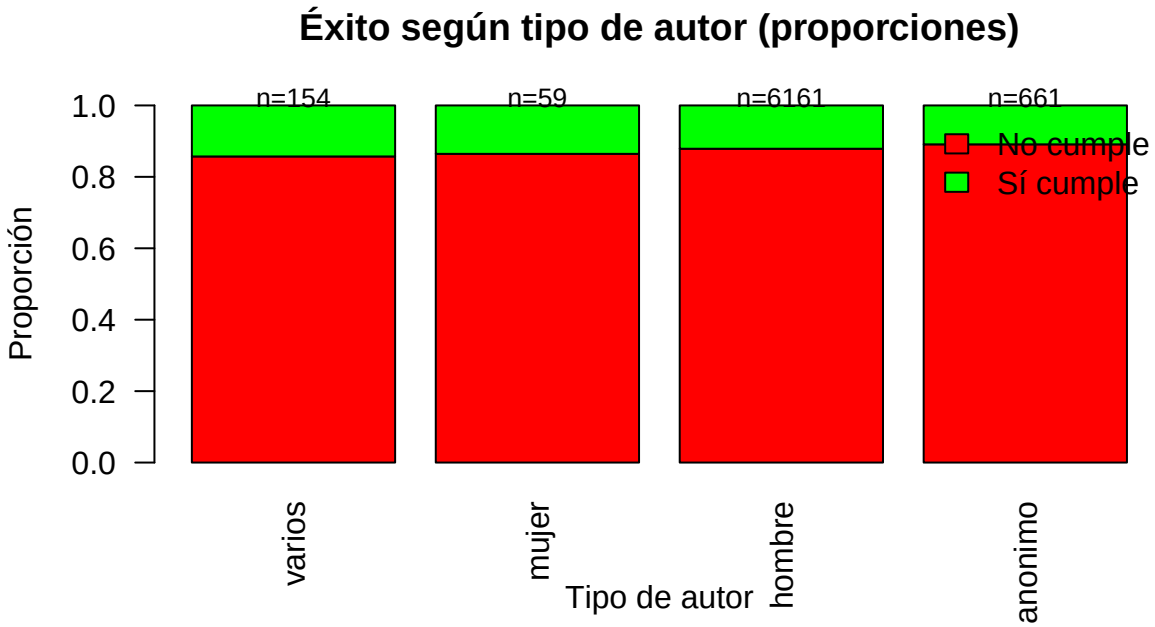
TIPO_AUTOR

```
##
## anonimo  hombre   mujer  varios
##     661    6161      59     154
```

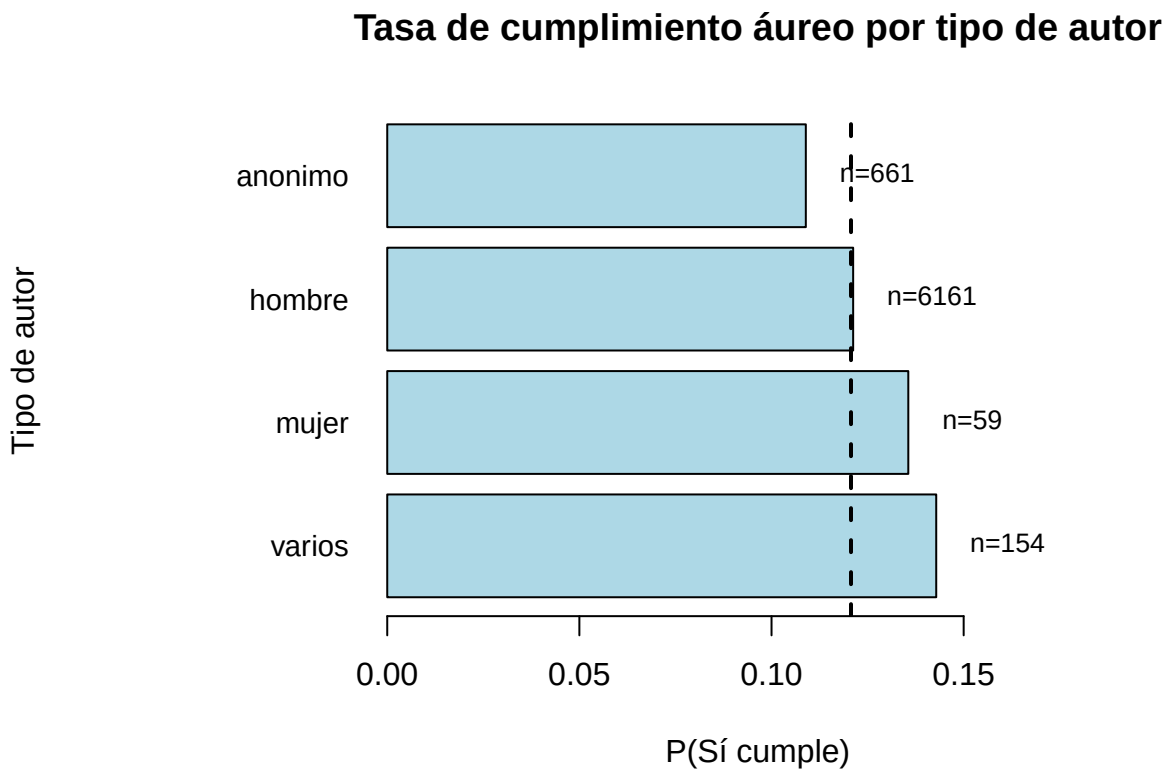


TIPO AUTOR CON EXITO

```
##
##           No cumple Sí cumple
##   anonimo       589        72
##   hombre       5414       747
##   mujer          51         8
##   varios        132        22
##
##           No cumple Sí cumple
##   varios       0.857     0.143
##   mujer        0.864     0.136
##   hombre       0.879     0.121
##   anonimo      0.891     0.109
```

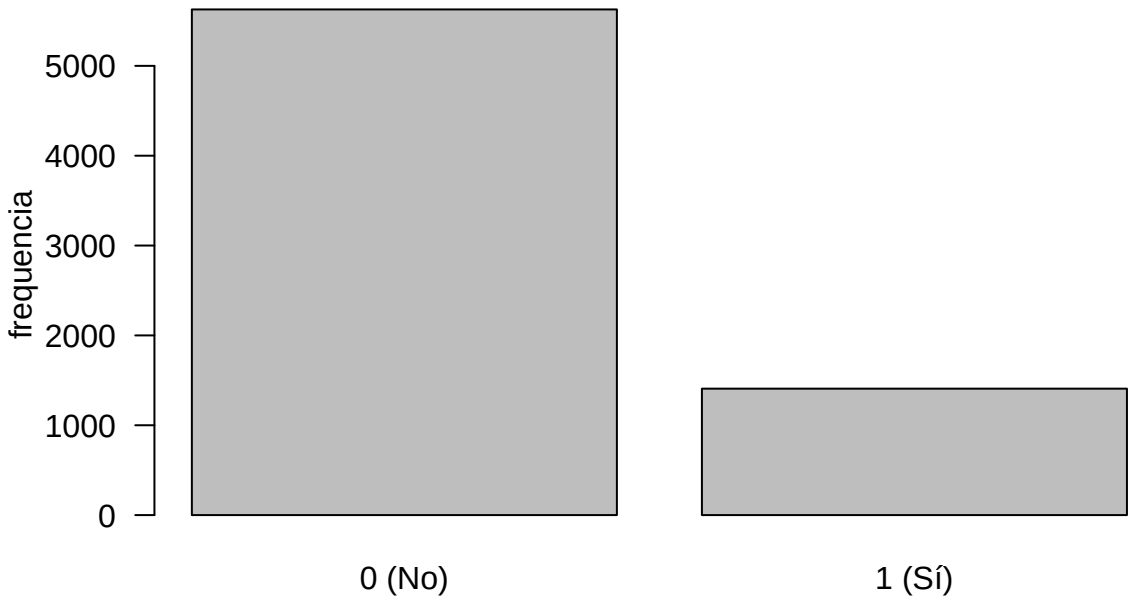


En todos los grupos (varios, mujer, hombre y anónimo) la mayoría de las obras no cumple la razón áurea, y la parte verde es siempre pequeña. Las proporciones de "sí cumple" son muy parecidas entre tipos de autor, sin un grupo que destaque claramente por cumplir mucho más que los demás. Esto sugiere que el éxito en cumplir la razón áurea no depende de quién sea el autor, sino que el incumplimiento es lo normal independientemente del tipo de autor.



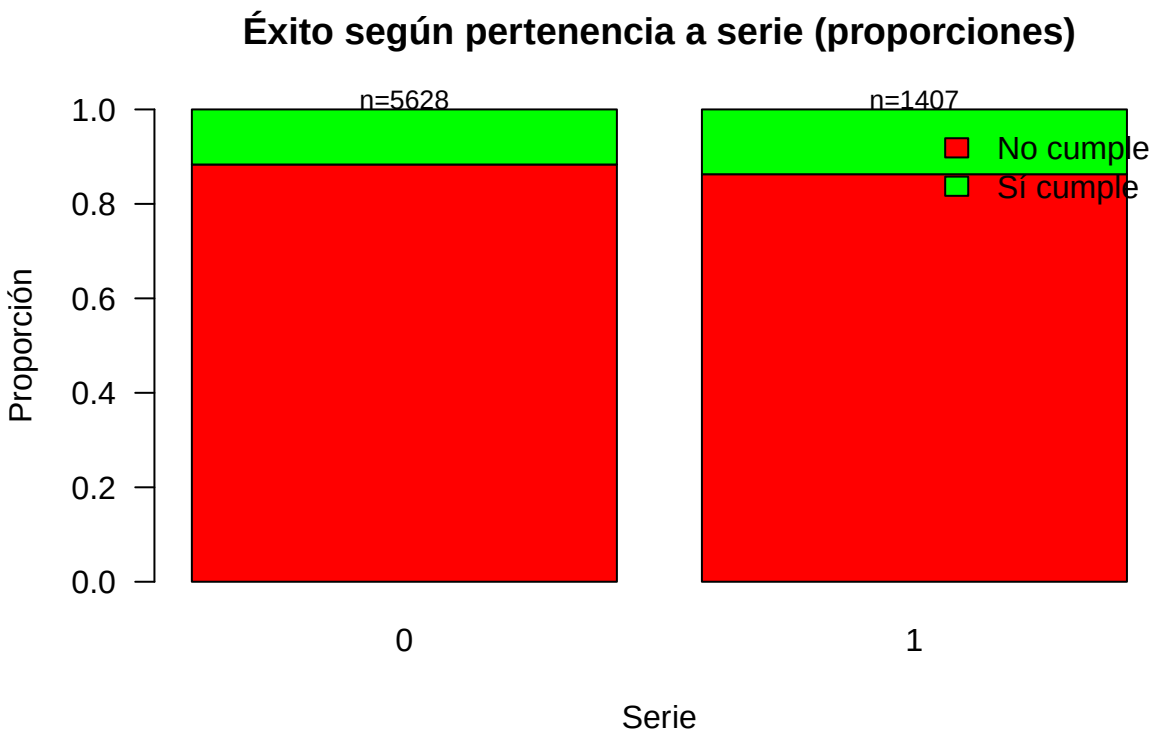
SERIE

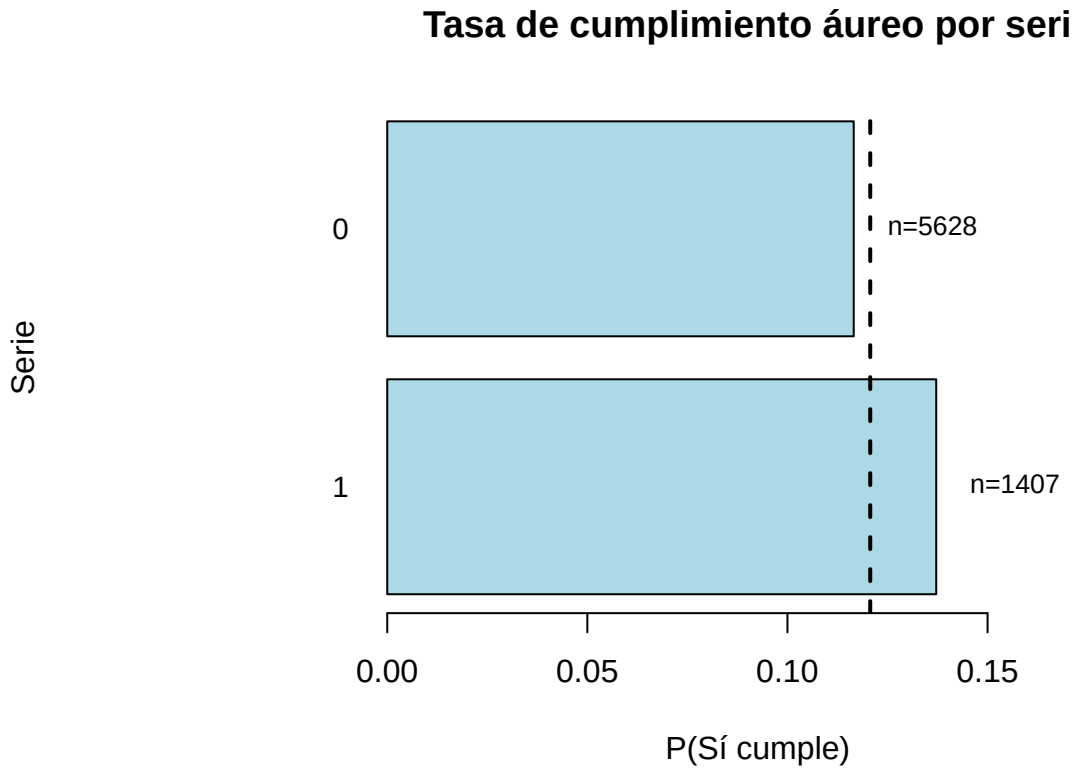
```
##
##    0    1
## 5628 1407
```



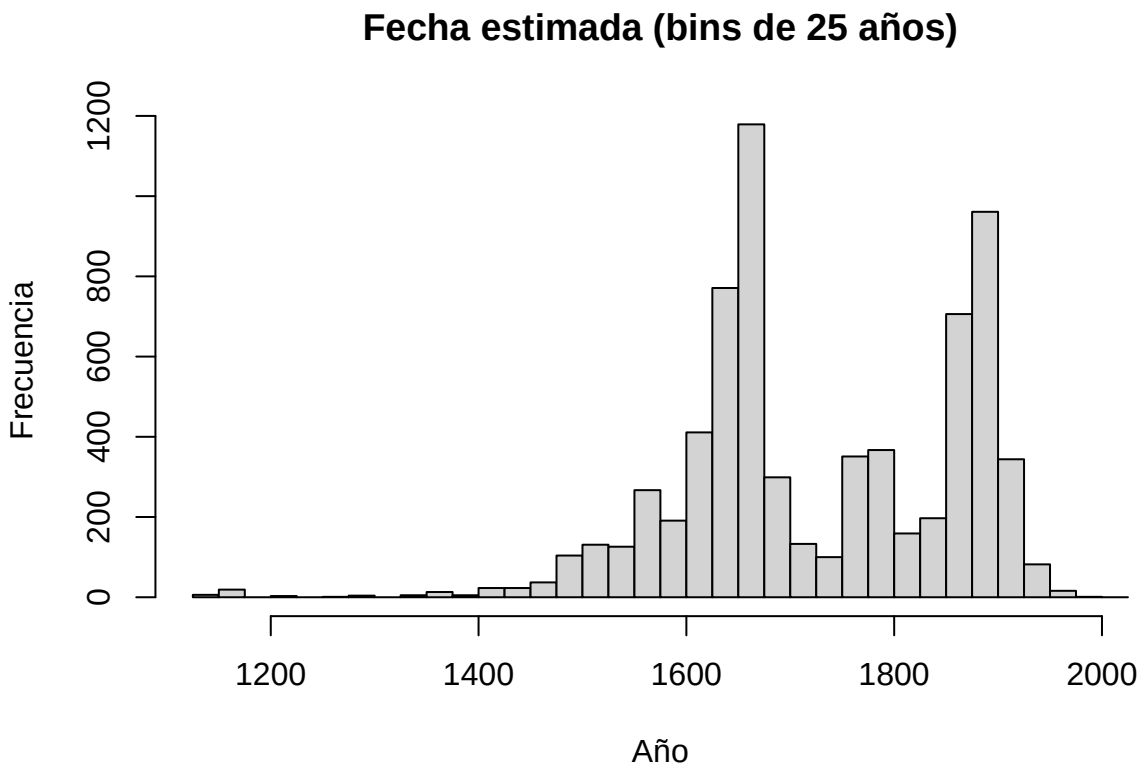
SERIE CON EXITO

```
##
##     No cumple Sí cumple
##   0      4972       656
##   1      1214       193
##
##     No cumple Sí cumple
##   0     0.883     0.117
##   1     0.863     0.137
```

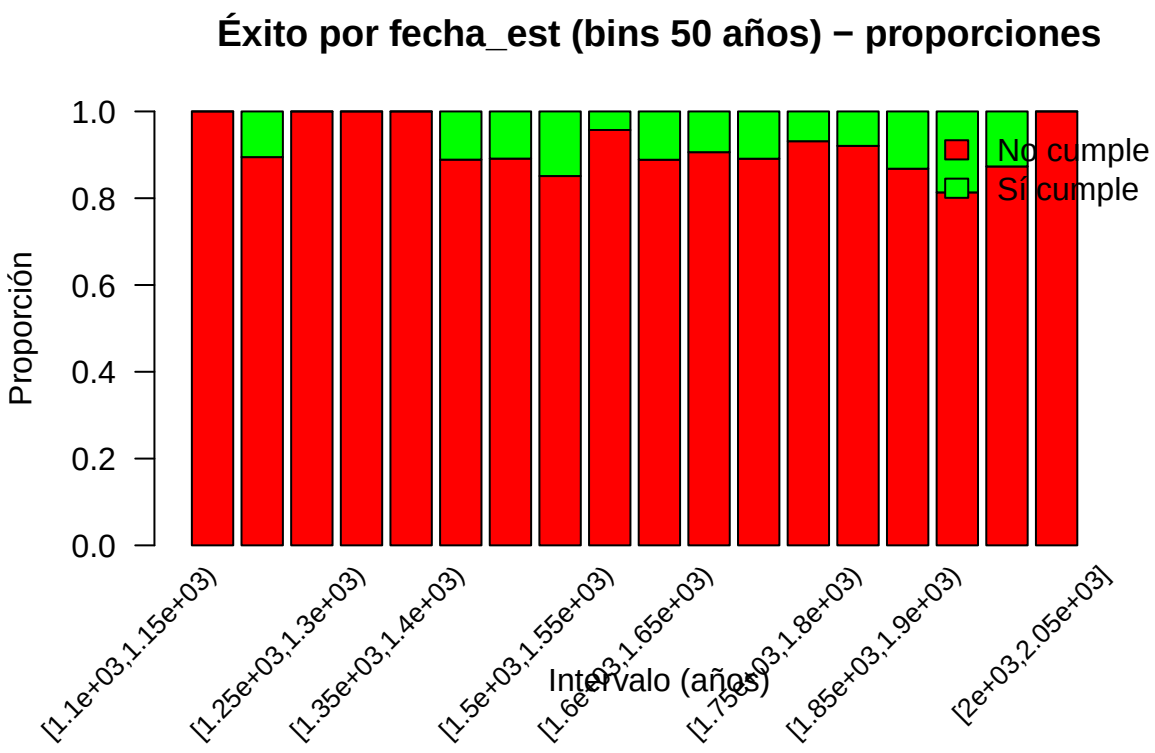

Éxito según pertenencia a serie (proporciones)
n=5628
n=1407
No cumple
Sí cumple
1.0
0.8
0.6
0.4
0.2
0.0
Proporción
0
1
Serie



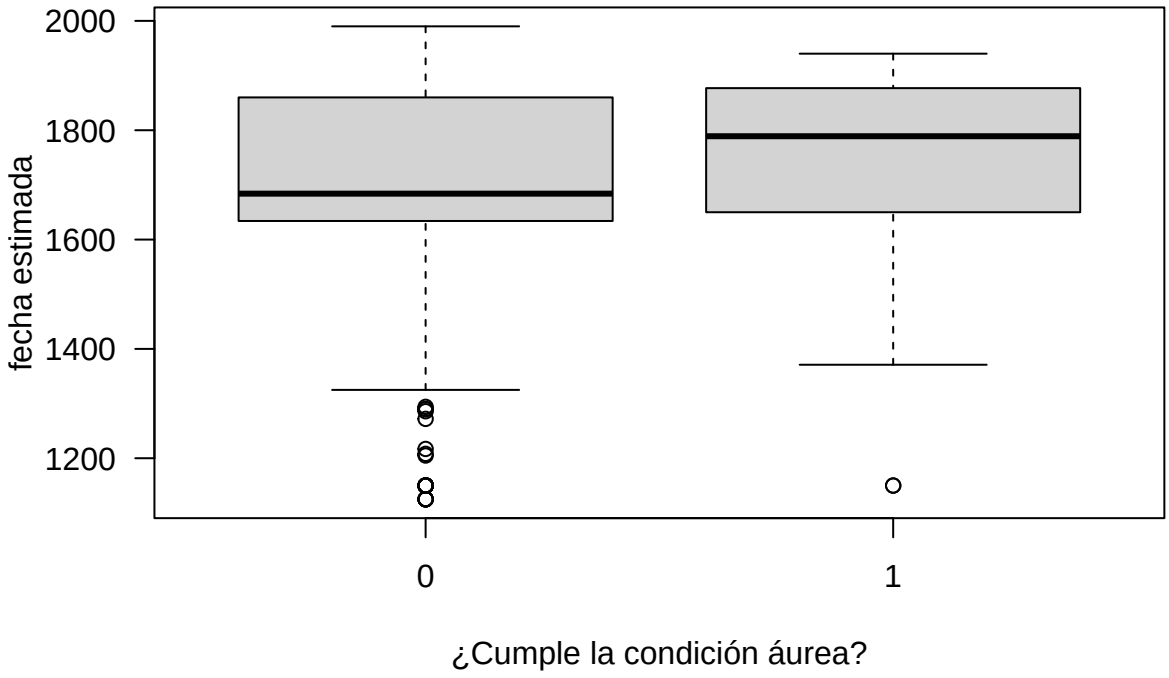
FECHA_ESTIMADA



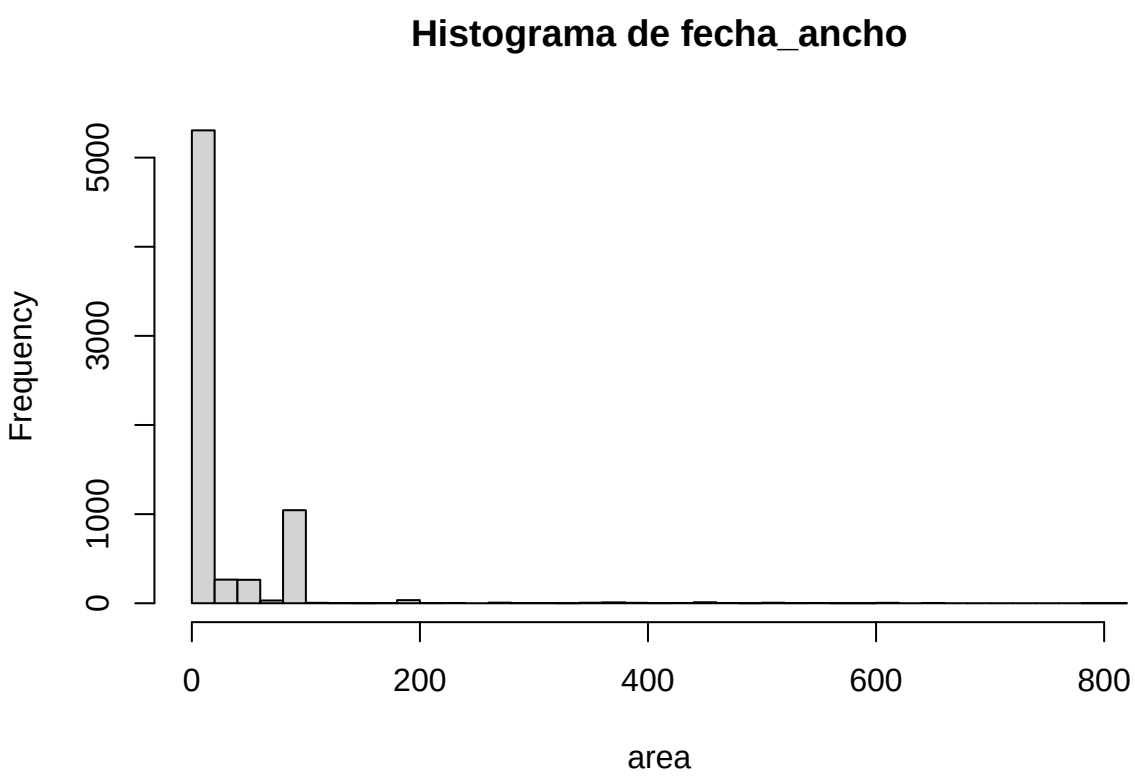
FECHA Y ÉXITO


Éxito por fecha_est (bins 50 años) − proporciones
Proporción
1.0
0.8
0.6
0.4
0.2
0.0
No cumple
Sí cumple
[1.1e+03,1.15e+03)
[1.25e+03,1.3e+03)
[1.35e+03,1.4e+03)
[1.5e+03,1.55e+03)
[1.6e+03,1.65e+03)
[1.75e+03,1.8e+03)
[1.85e+03,1.9e+03)
[2e+03,2.05e+03]
Intervalo (años)

Distribución de la fecha estimada según cumplimiento de la condición a

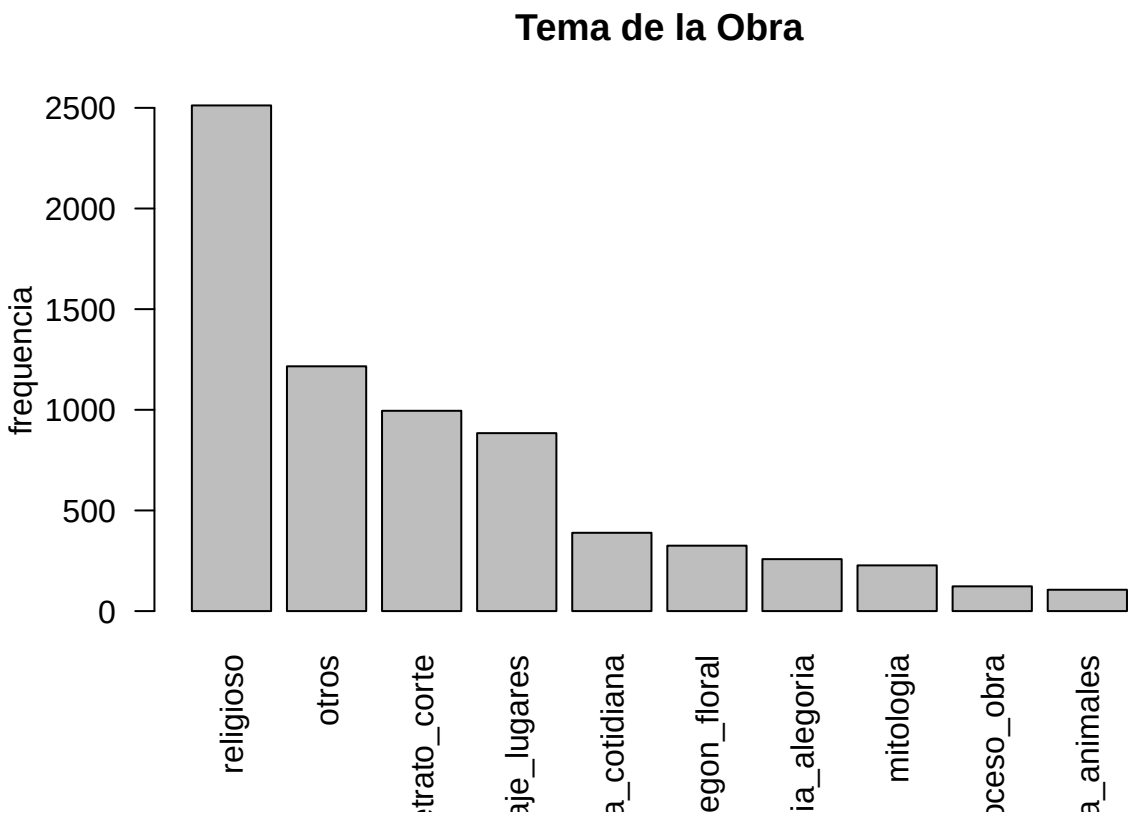


FECHA_ANCHO

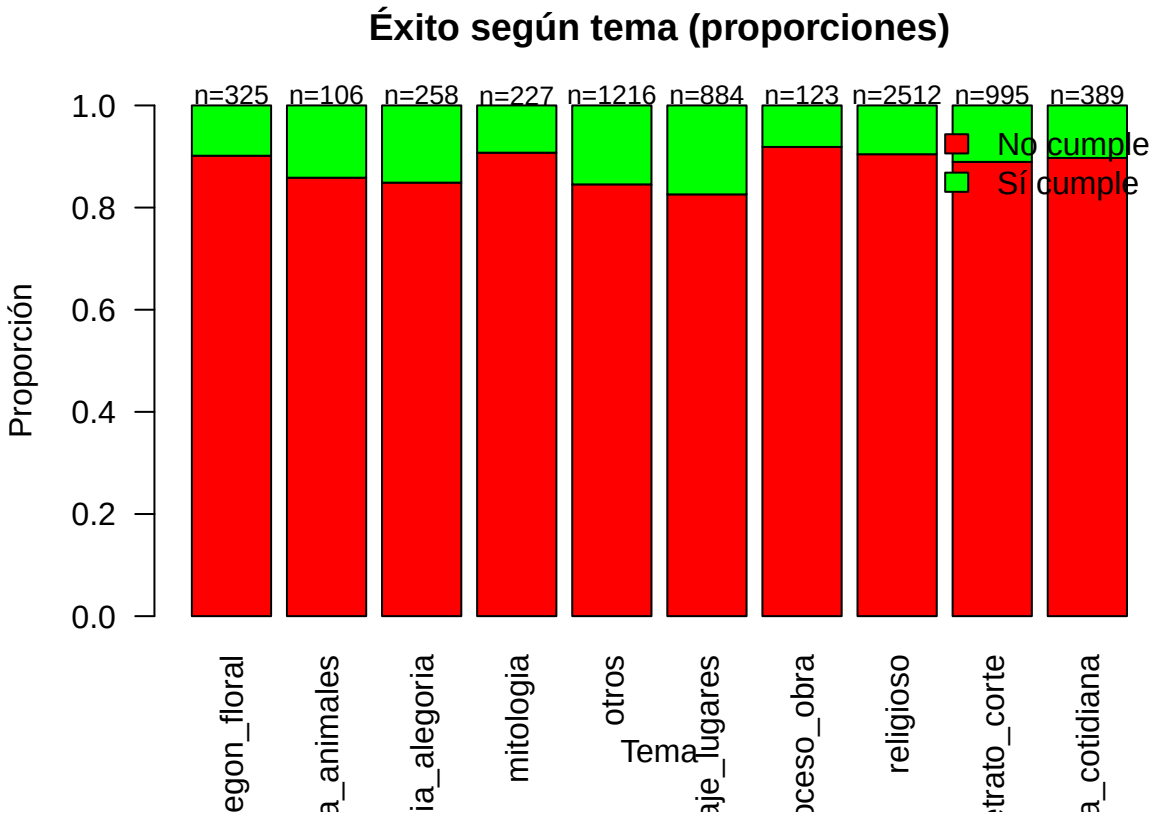


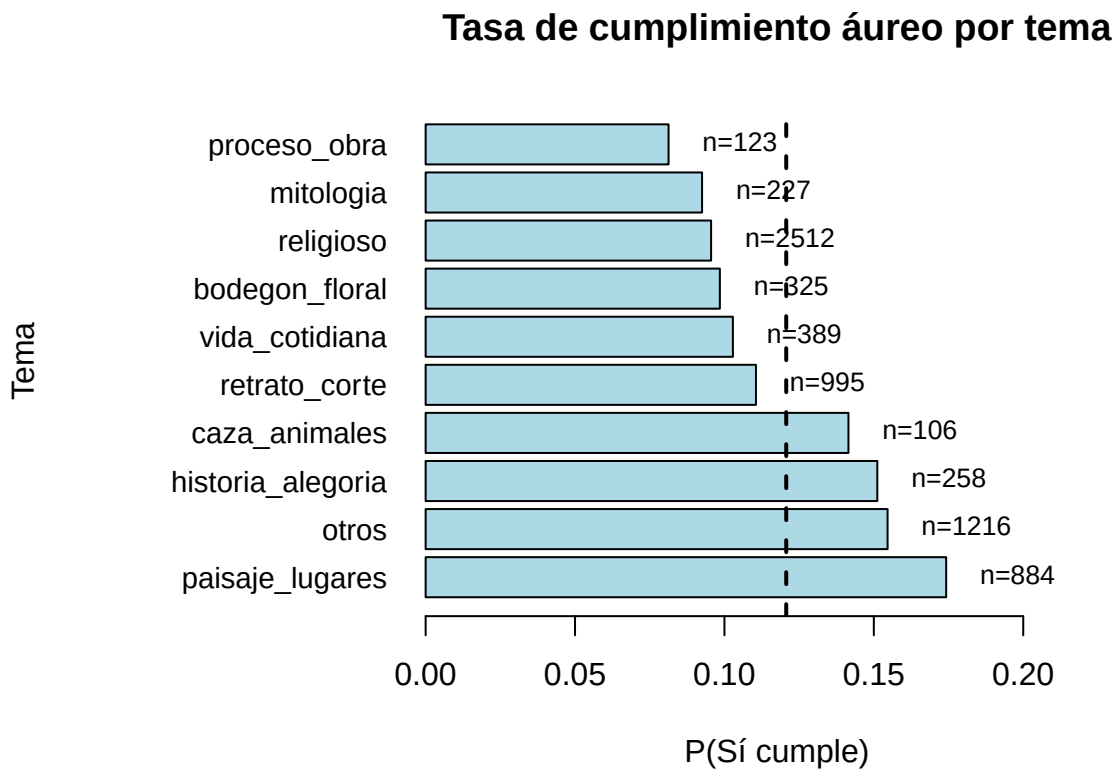
TEMA

```
##
##    bodegon_floral     caza_animales historia_alegoria         mitologia
##               325               106               258               227
##            otros   paisaje_lugares      proceso_obra         religioso
##              1216               884               123              2512
##     retrato_corte    vida_cotidiana
##               995               389
```



TEMA CON EXITO

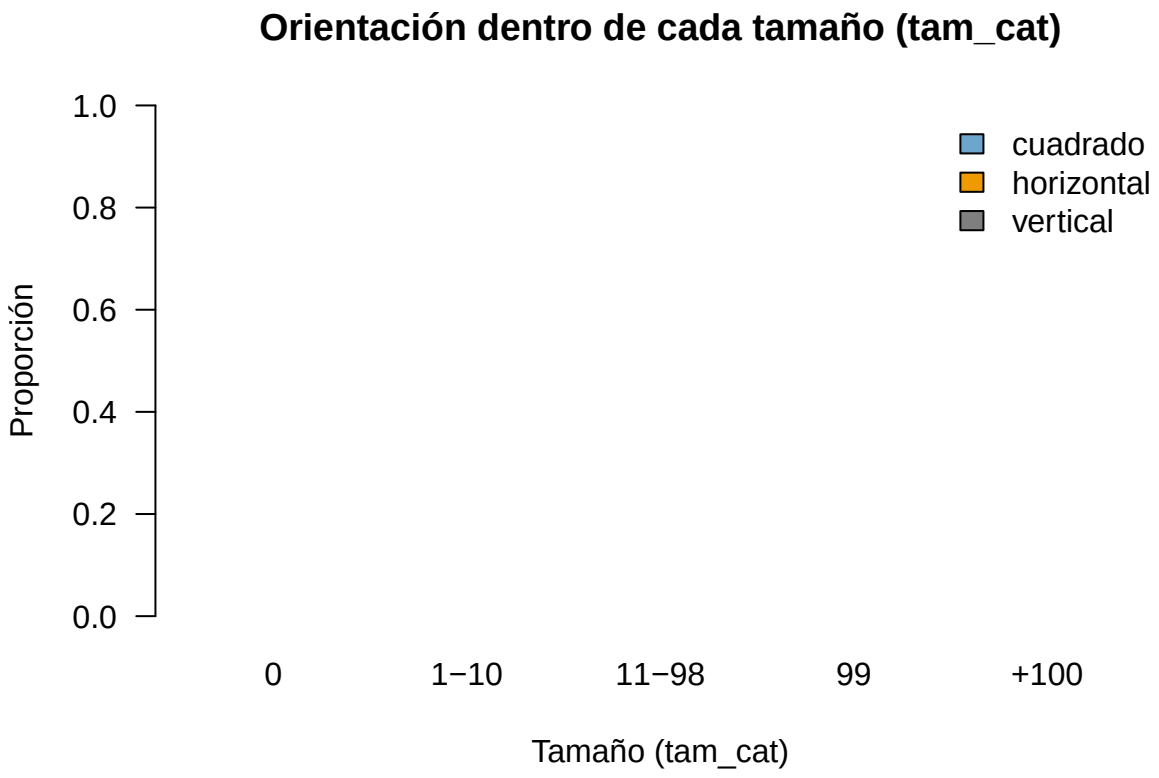

Éxito según tema (proporciones)
n=325
n=106
n=258
n=227
n=1216
n=884
n=123
n=2512
n=995
n=389
1.0
0.8
0.6
0.4
0.2
0.0
Proporción
No cumple
Sí cumple
egon_floral
a_animales
ia_alegoria
mitologia
otros
aje_lugares
oceso_obra
religioso
trato_corte
a_cotidiana
Tema



4.2 Combinaciones dos a dos

Orientación VS Tam_cat

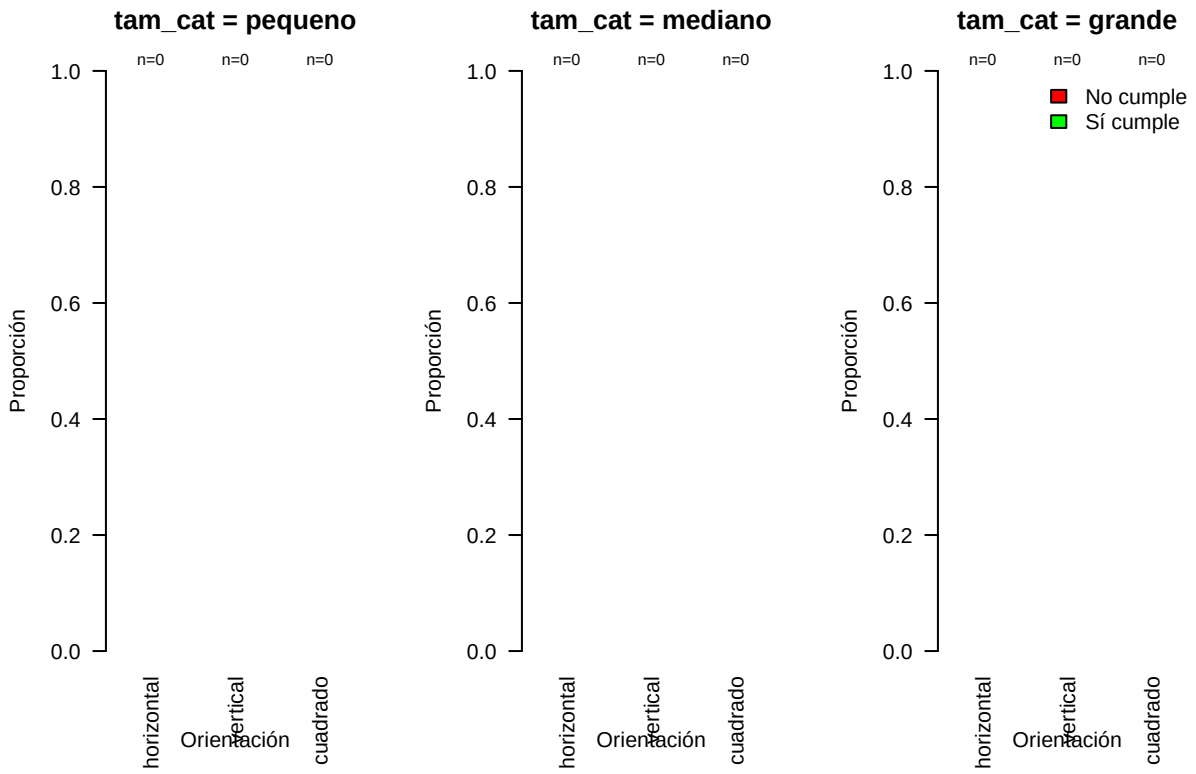
```
##
##              0 1-10 11-98 99 +100
##   cuadrado
##   horizontal
##   vertical
```



En los tres tamaños (pequeño, mediano y grande) prácticamente no aparecen obras "cuadrado" (proporción casi 0), así que la orientación dominante es horizontal/vertical.

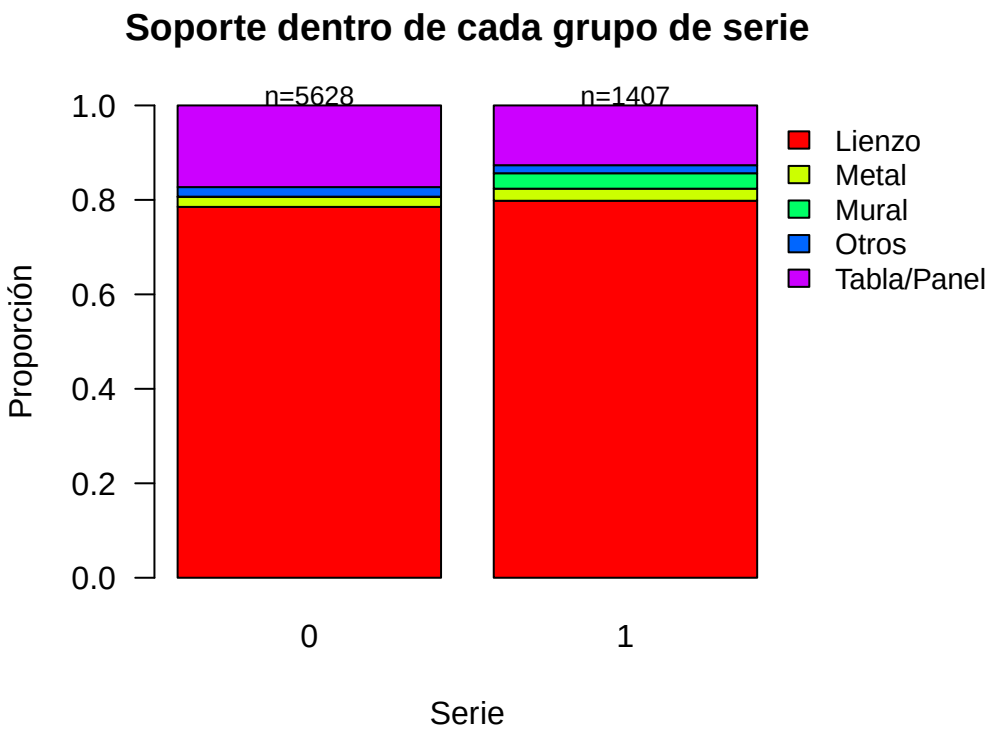
Sí hay diferencias entre categorías: en "mediano" aumenta la proporción de verticales (y bajan las horizontales) respecto a "grande" y "pequeño", pero las diferencias no parecen enormes (el patrón general se mantiene).

Orientación VS Tam_cat con Éxito

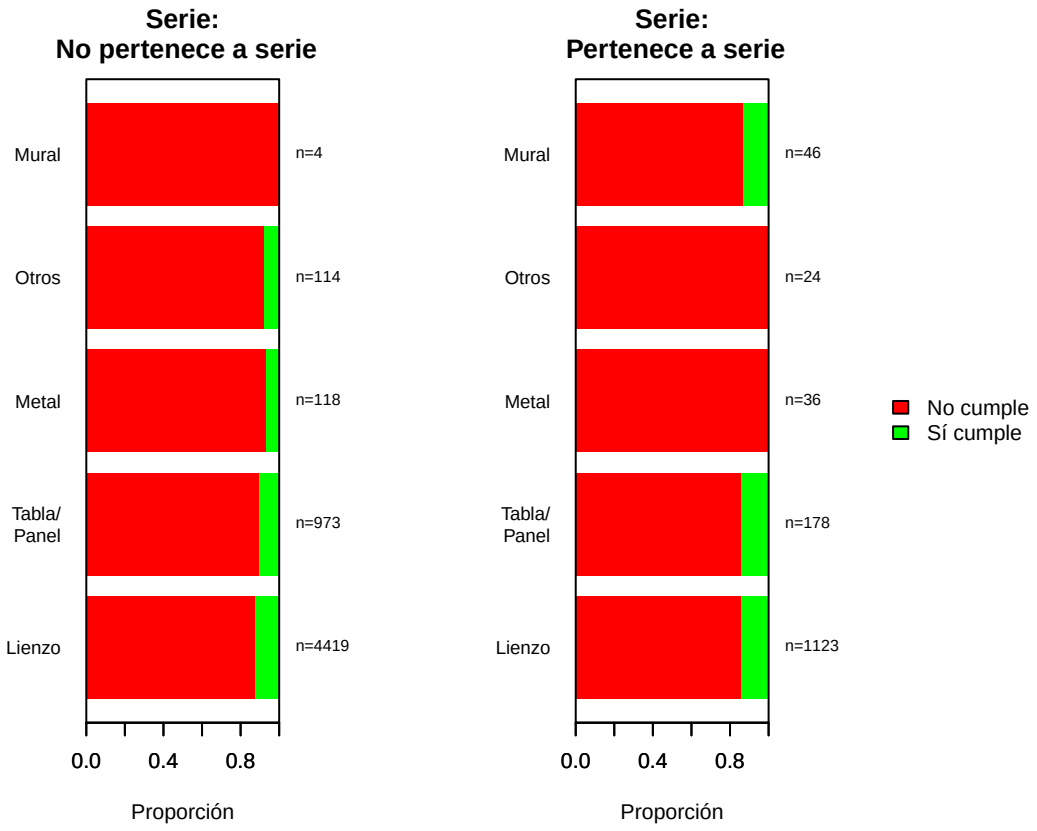


Série VS Soporte_grp

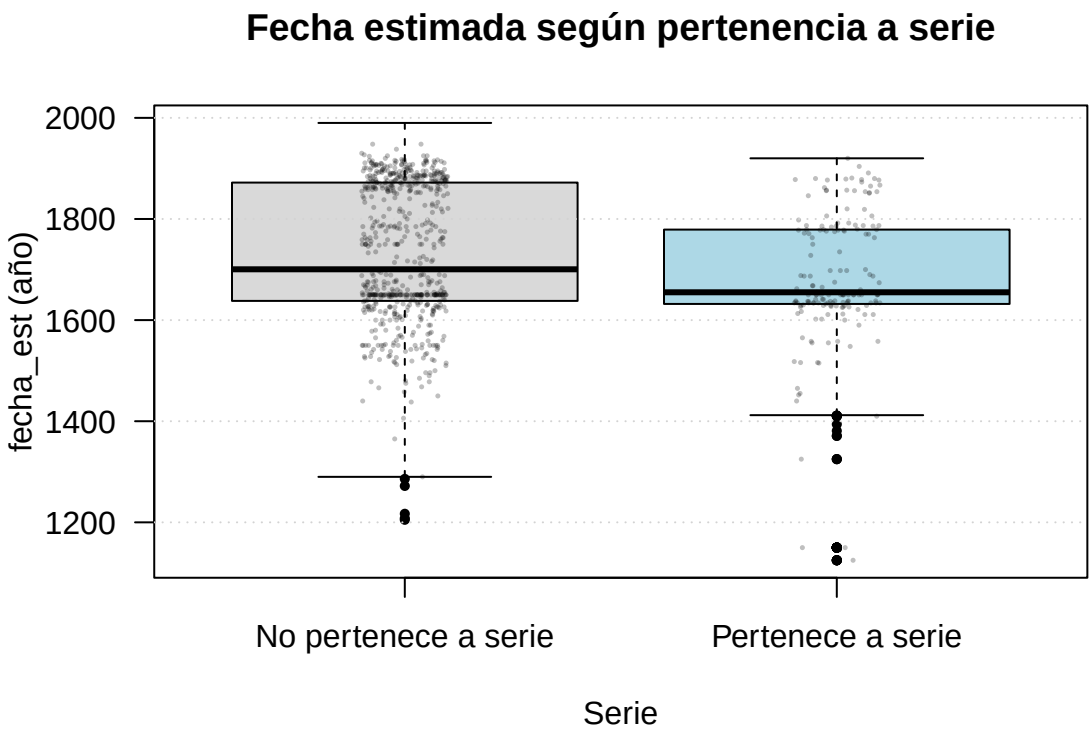
```
##
##                No pertenece a serie Pertenece a serie
##   Lienzo                       4419                1123
##   Metal                         118                  36
##   Mural                           4                  46
##   Otros                         114                  24
##   Tabla/Panel                   973                 178
```



Soporte_grp VS Serie y Éxito

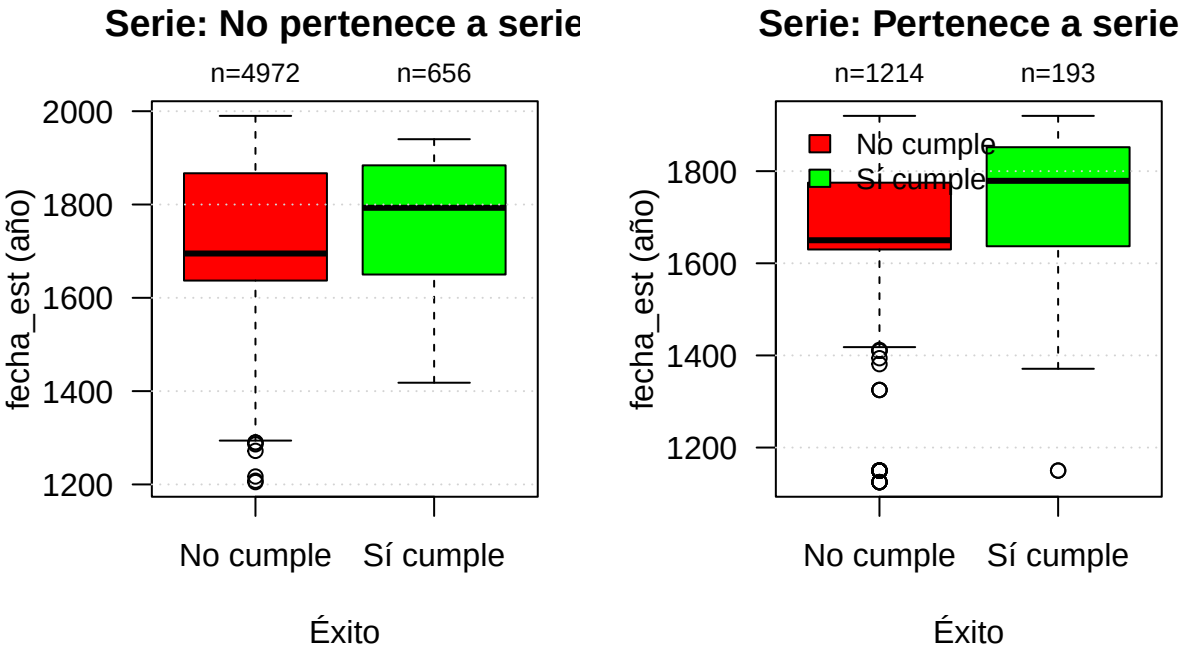


Serie VS Fecha_est



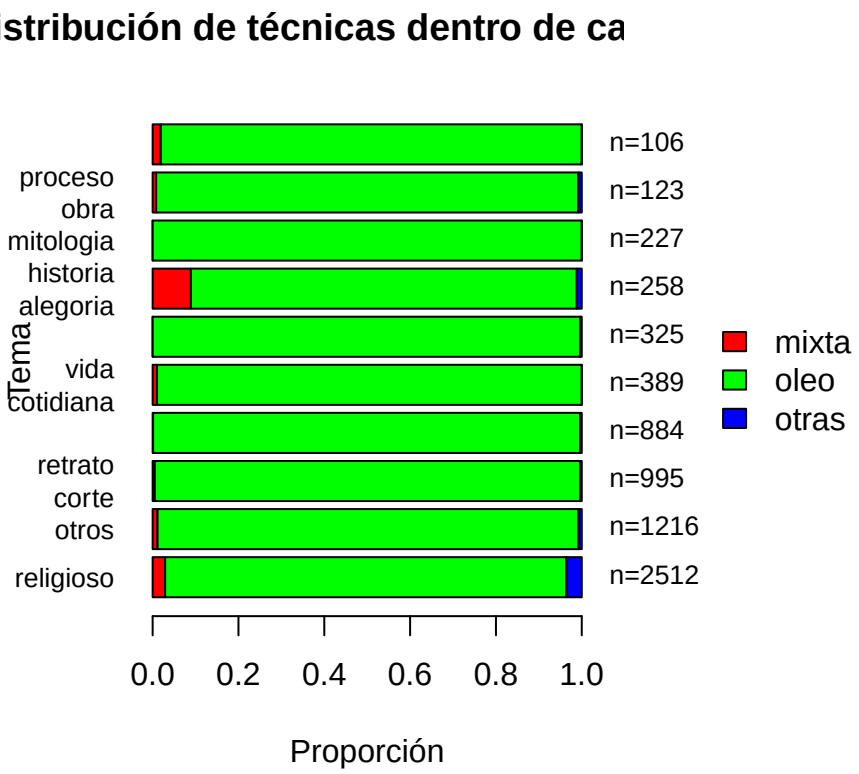
Fecha_est VS Série con Éxito

Fecha estimada por éxito, separando si pertenece a serie



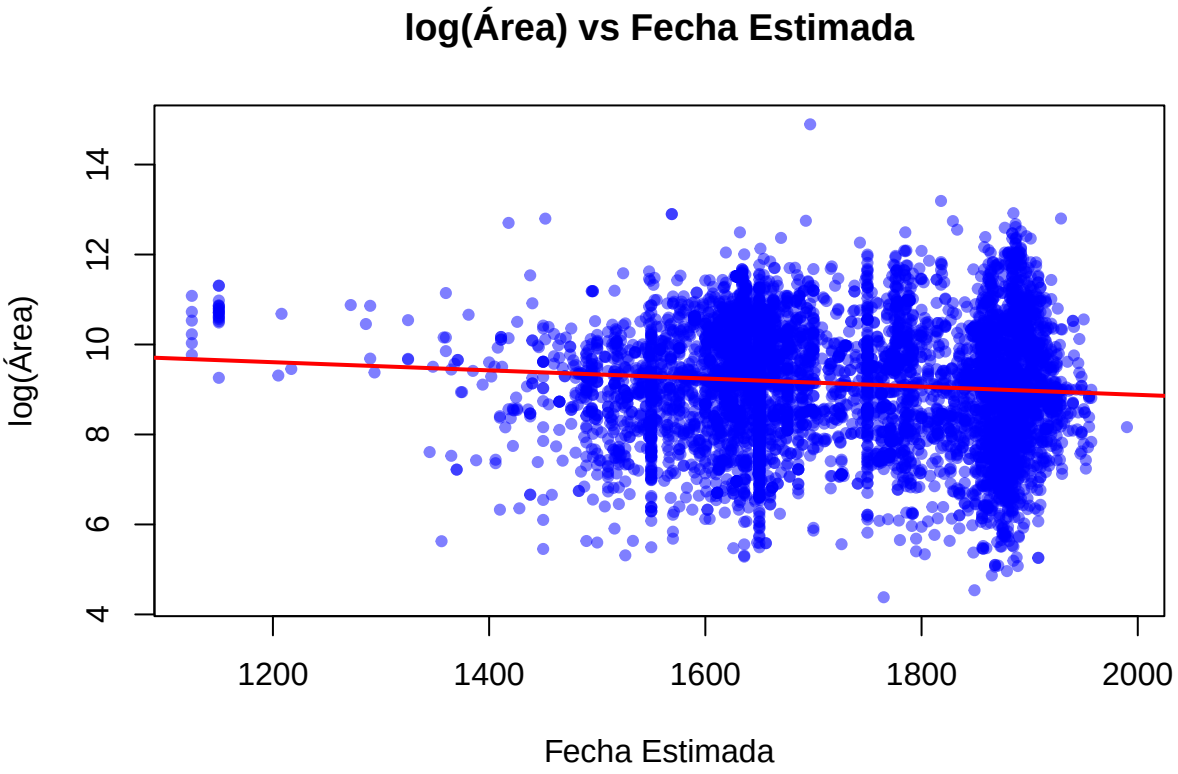
Tema VS Técnica

```
##                     tec
## tema                 mixta oleo otras
##   bodegon_floral         0  324     1
##   caza_animales          2  104     0
##   historia_alegoria     23  232     3
##   mitologia              0  227     0
##   otros                 14 1193     9
##   paisaje_lugares        1  880     3
##   proceso_obra           1  121     1
##   religioso             73 2349    90
##   retrato_corte          5  987     3
##   vida_cotidiana         4  385     0
##                     tec
## tema                 mixta  oleo otras
##   religioso          0.029 0.935 0.036
##   otros              0.012 0.981 0.007
##   retrato_corte      0.005 0.992 0.003
##   paisaje_lugares    0.001 0.995 0.003
##   vida_cotidiana     0.010 0.990 0.000
##   bodegon_floral     0.000 0.997 0.003
##   historia_alegoria  0.089 0.899 0.012
##   mitologia          0.000 1.000 0.000
##   proceso_obra       0.008 0.984 0.008
##   caza_animales      0.019 0.981 0.000
```



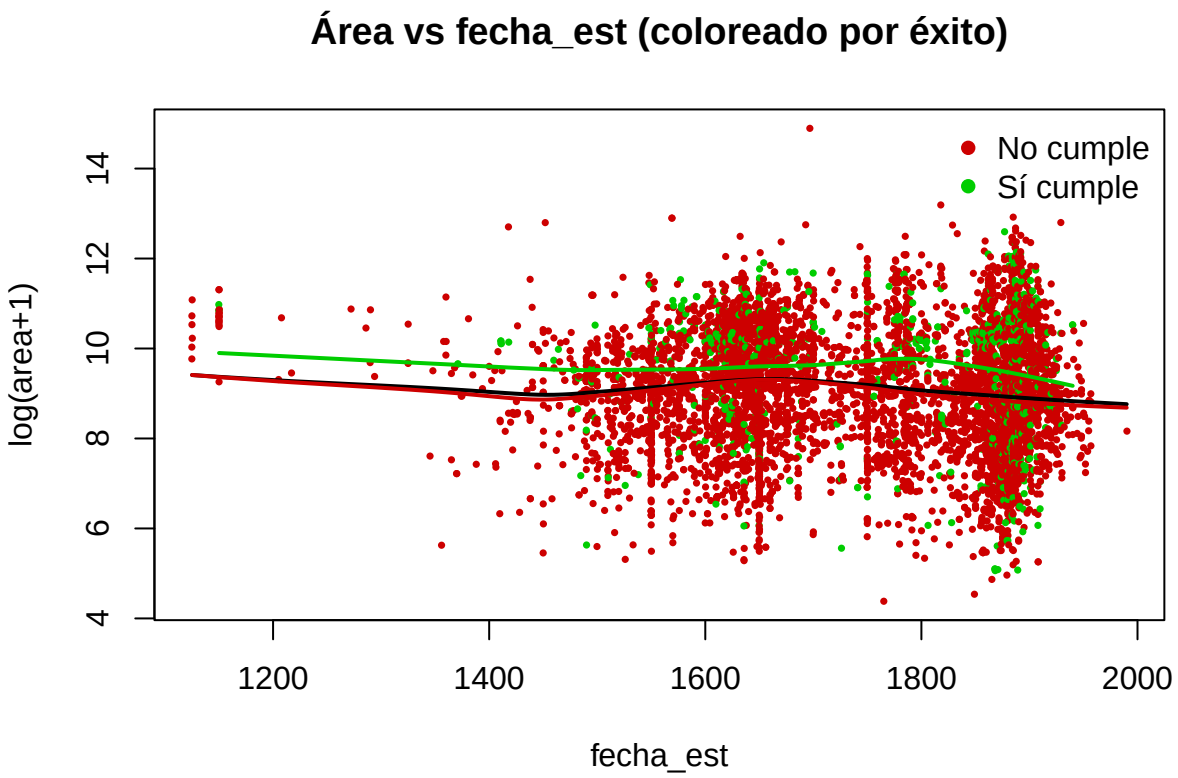
Tema VS Técnica y Éxito

Área VS Fecha

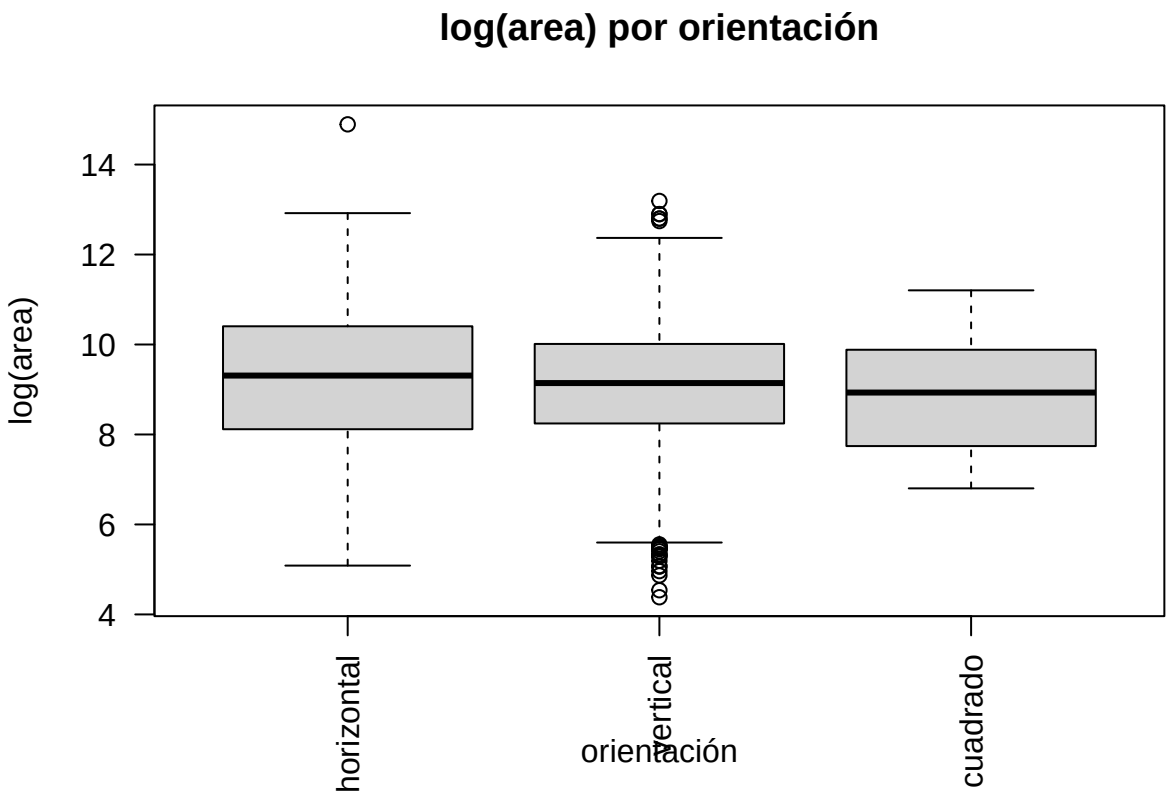


La línea LOWESS es la tendencia: como está casi plana (ligeramente hacia abajo), no parece que el área cambie mucho con la fecha, las verticales salen porque muchos años se repiten.

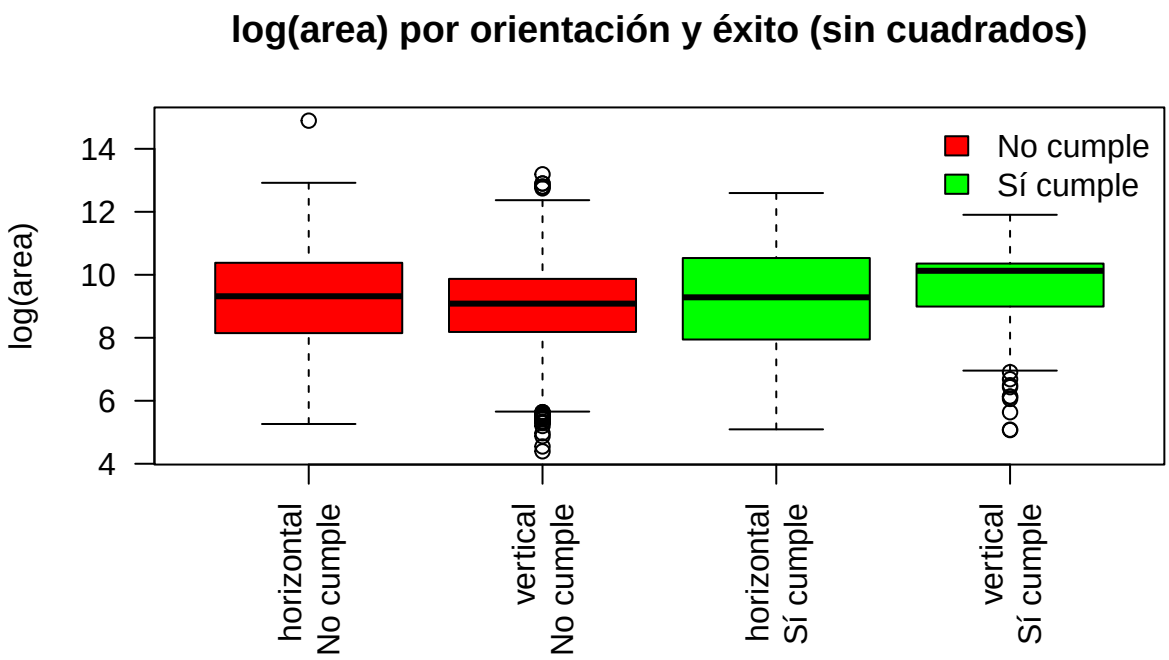
Área VS Fecha_est con Éxito



Área VS Orientación

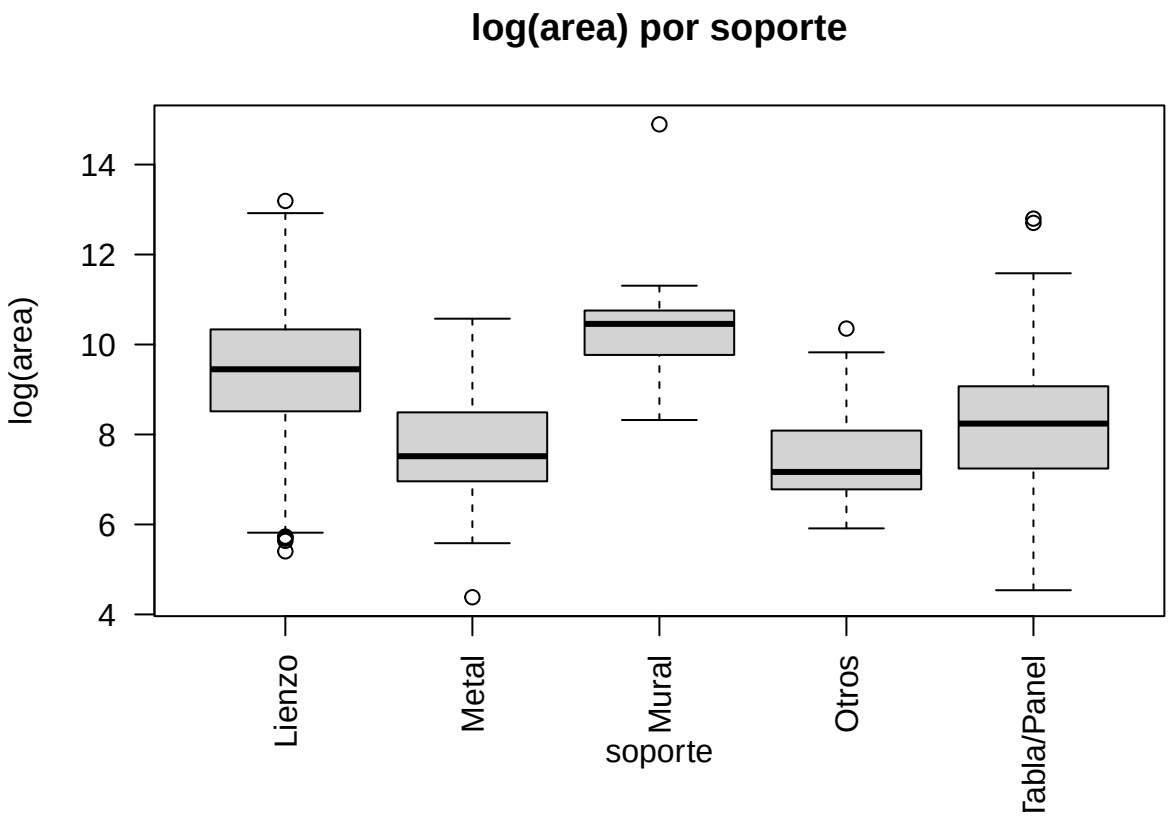


Área VS Orientación con Éxito

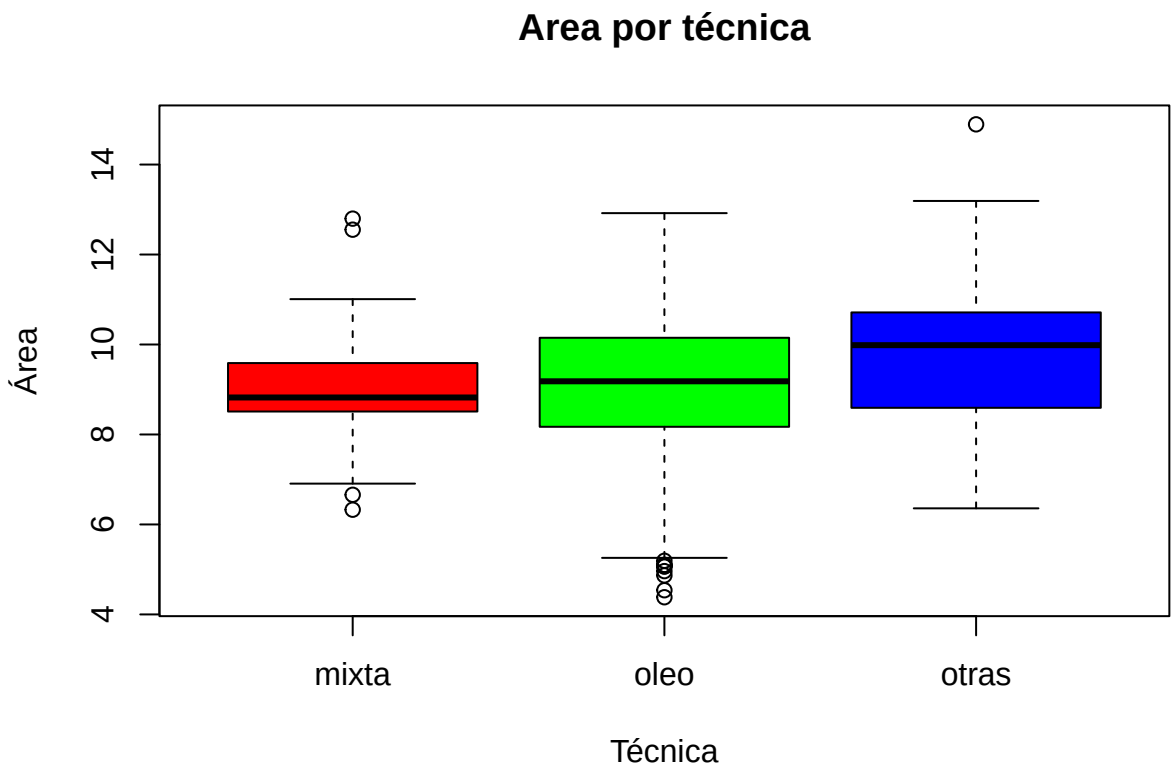


Área VS Soporte

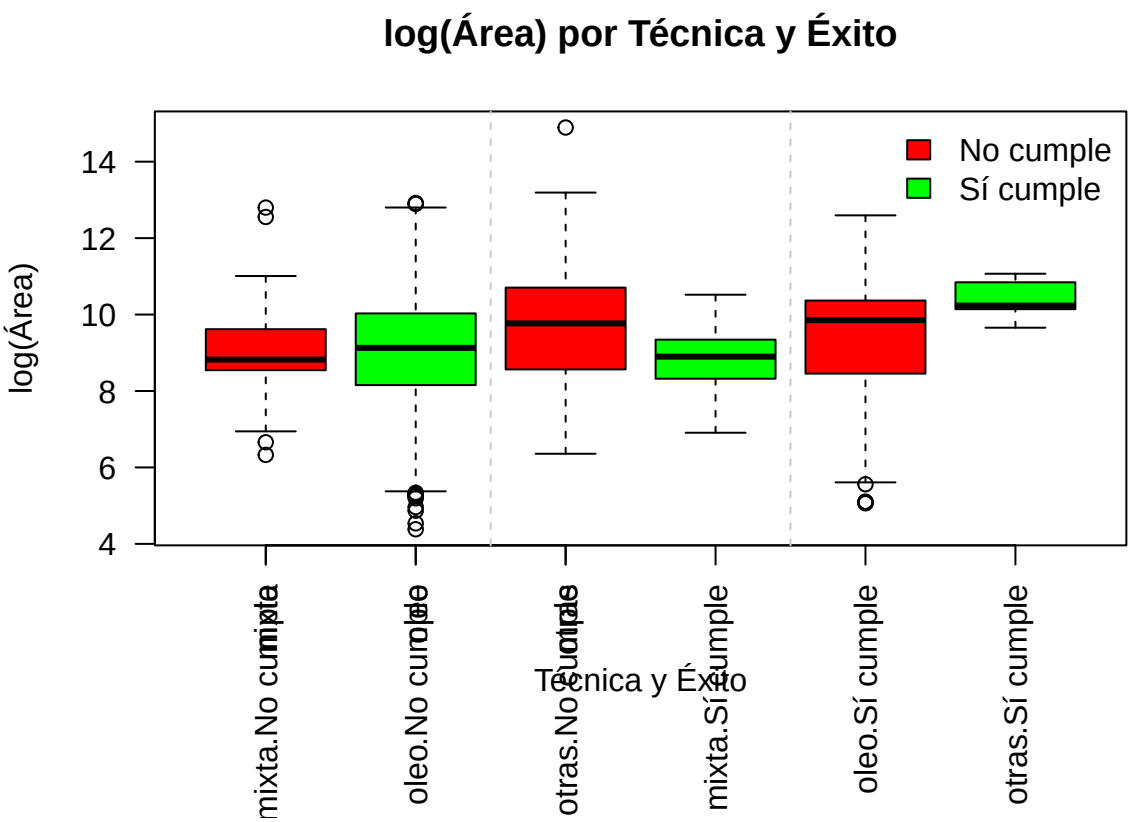
```
##             exito_f
## orientacion  No cumple Sí cumple
##   horizontal      2559       442
##   vertical        3594       407
##   cuadrado          33         0
```



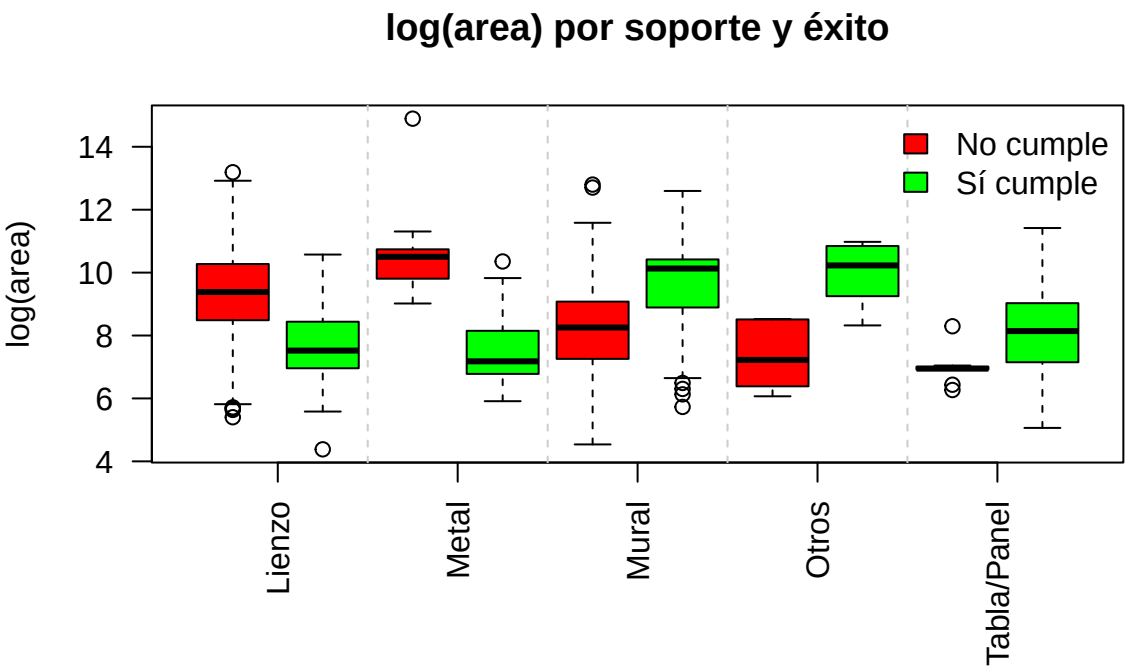
Area VS Técnica



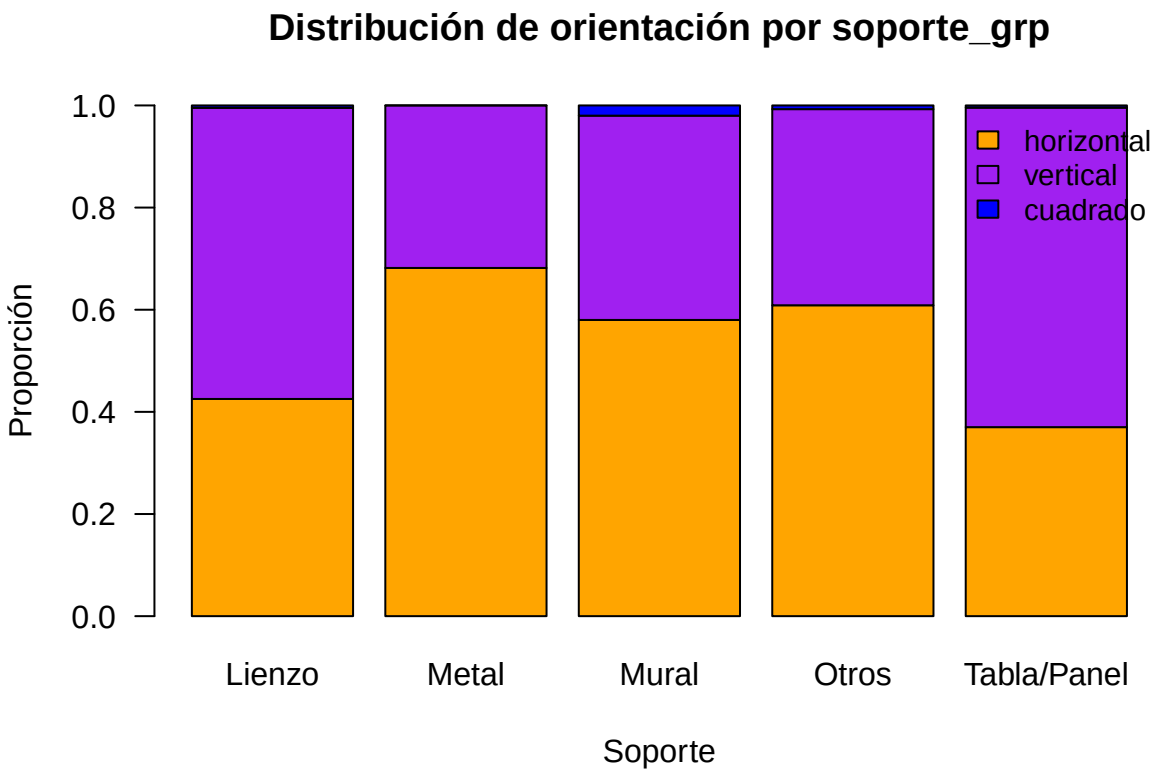
Area VS Técnica y Éxito



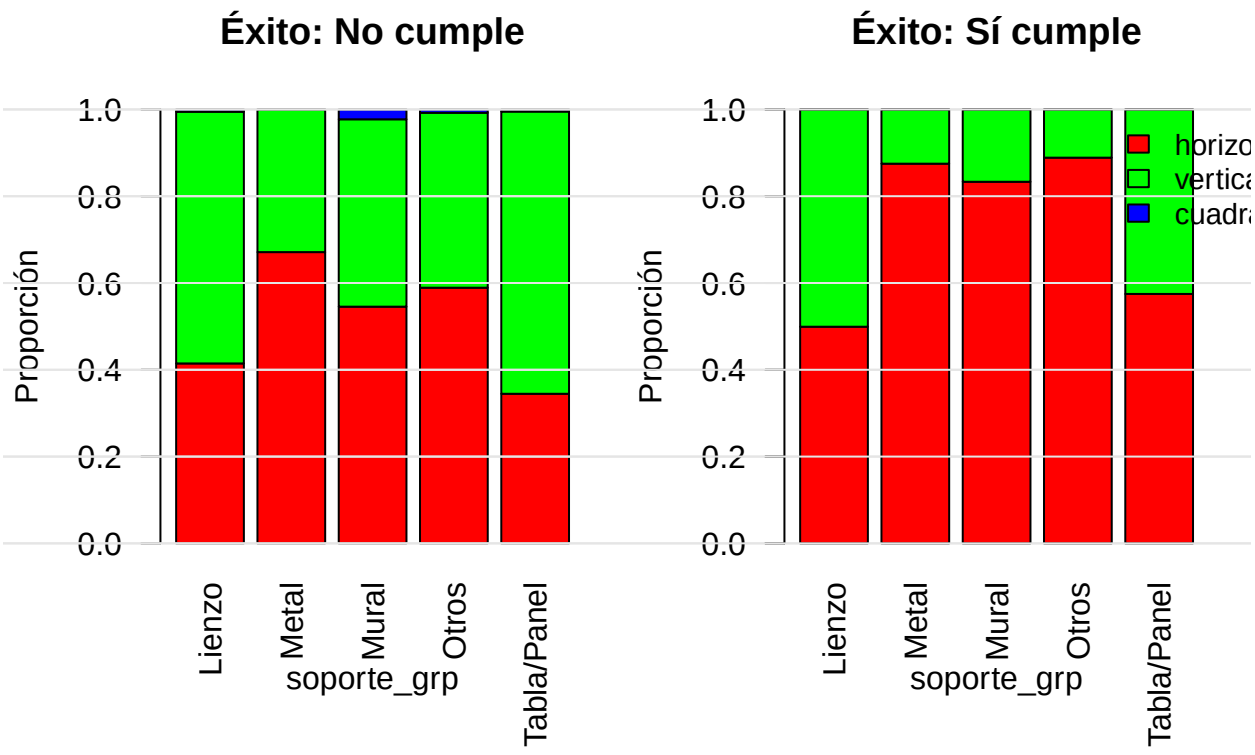
Soporte VS Área y Éxito



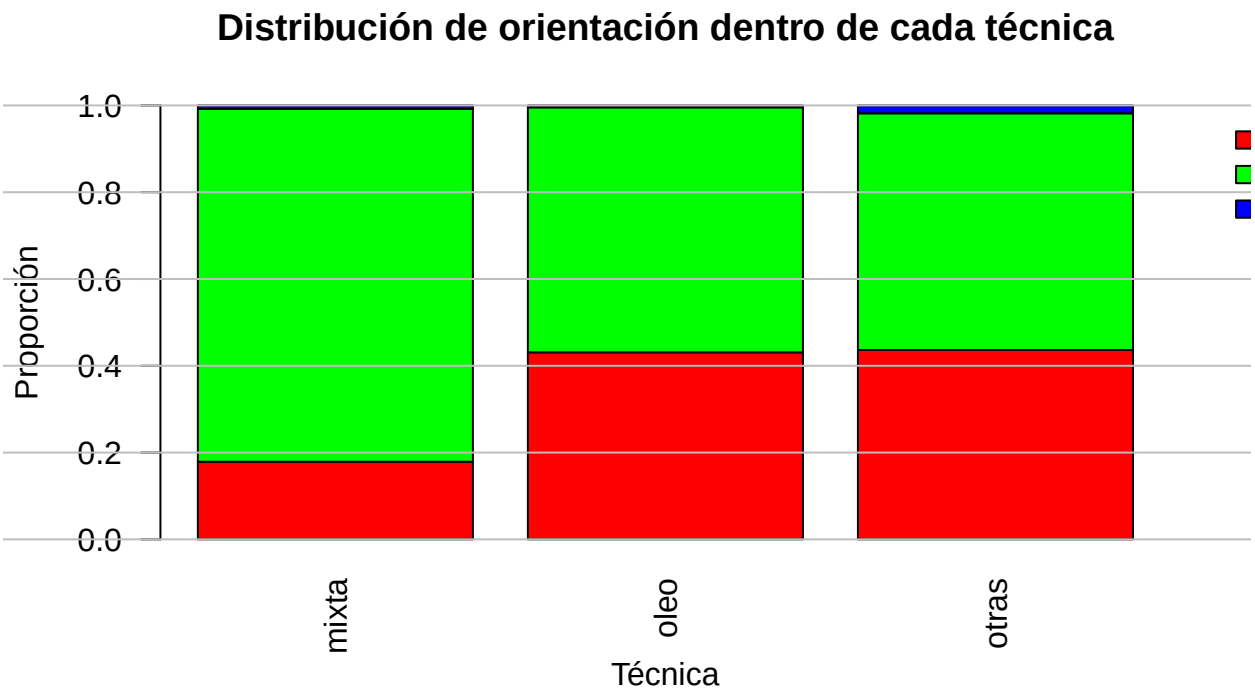
Orientación VS soporte_grp



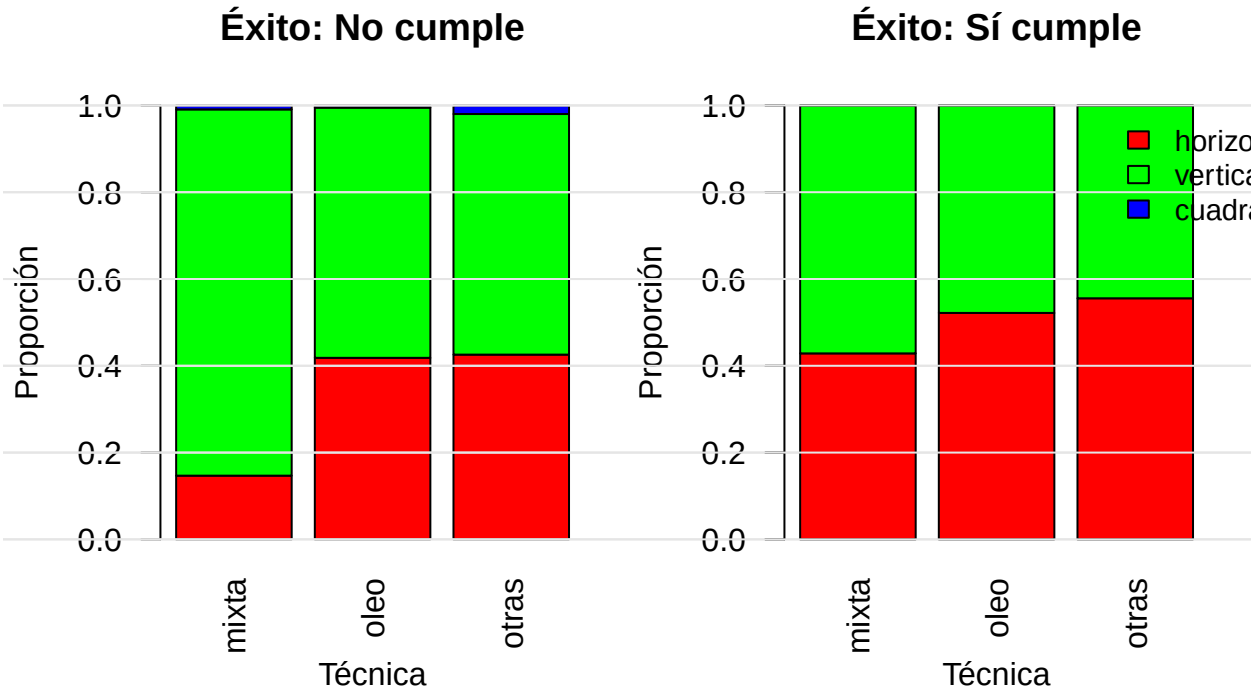
Orientación VS Soporte_grp con Éxito



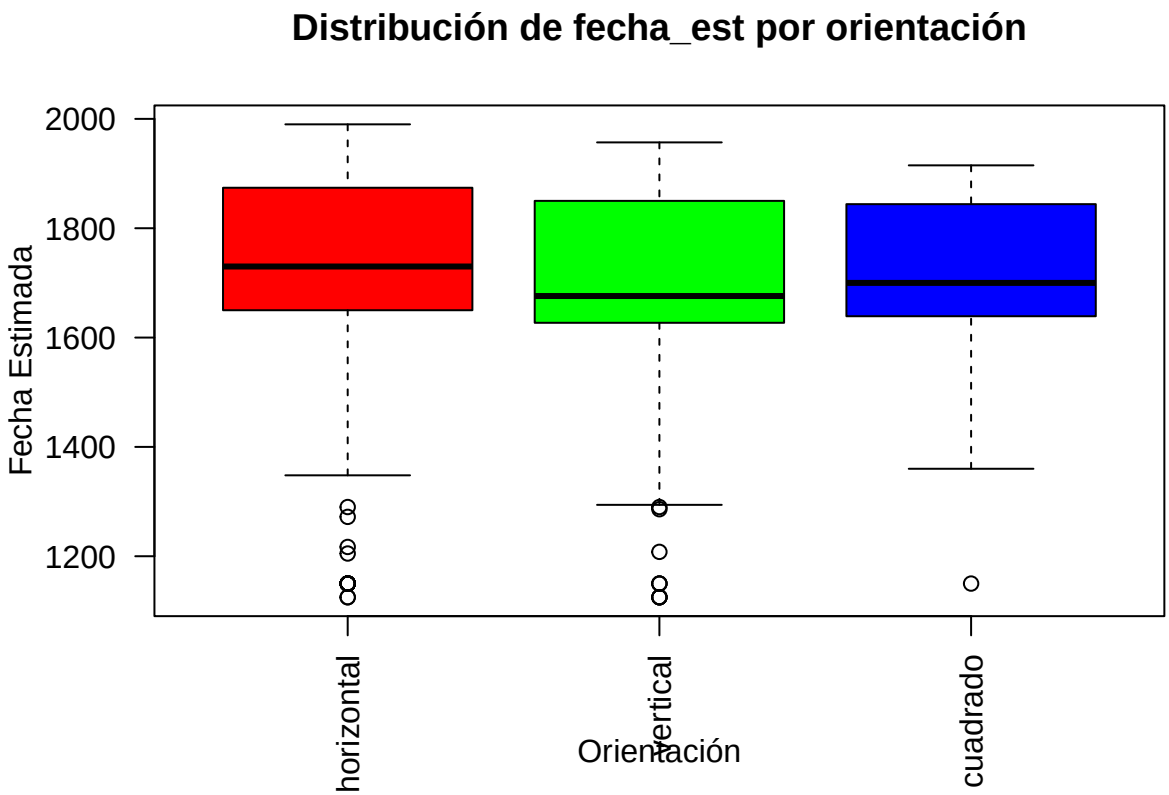
Orientación VS técnica



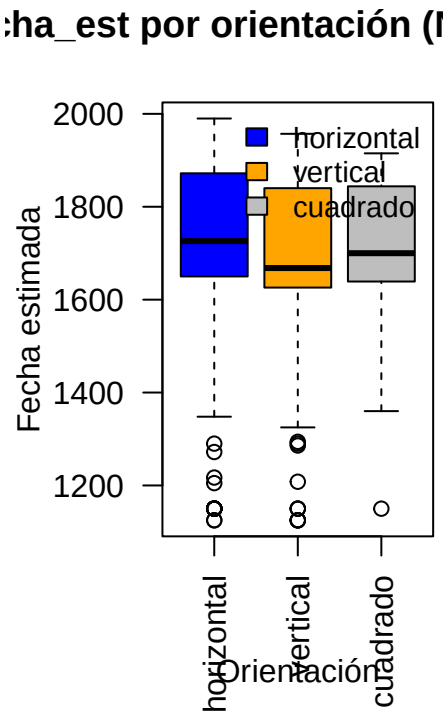
Orientación VS Técnica con Éxito

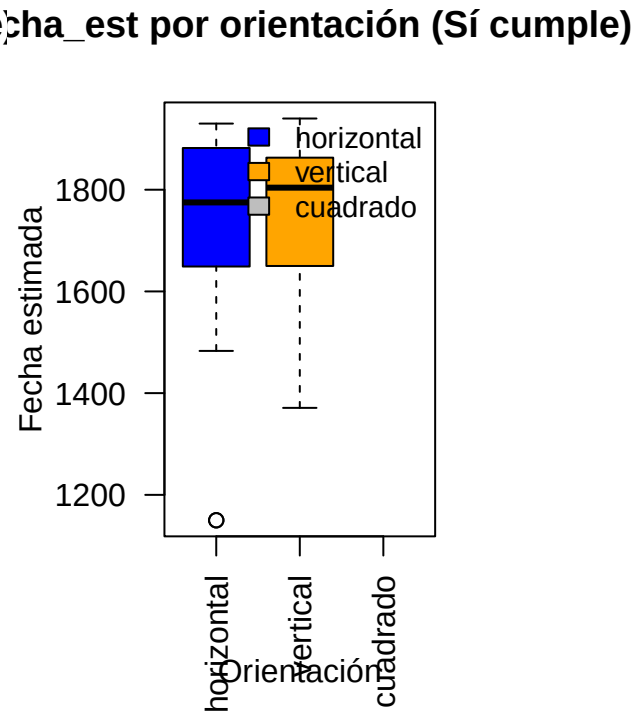


Orientación VS Fecha_est

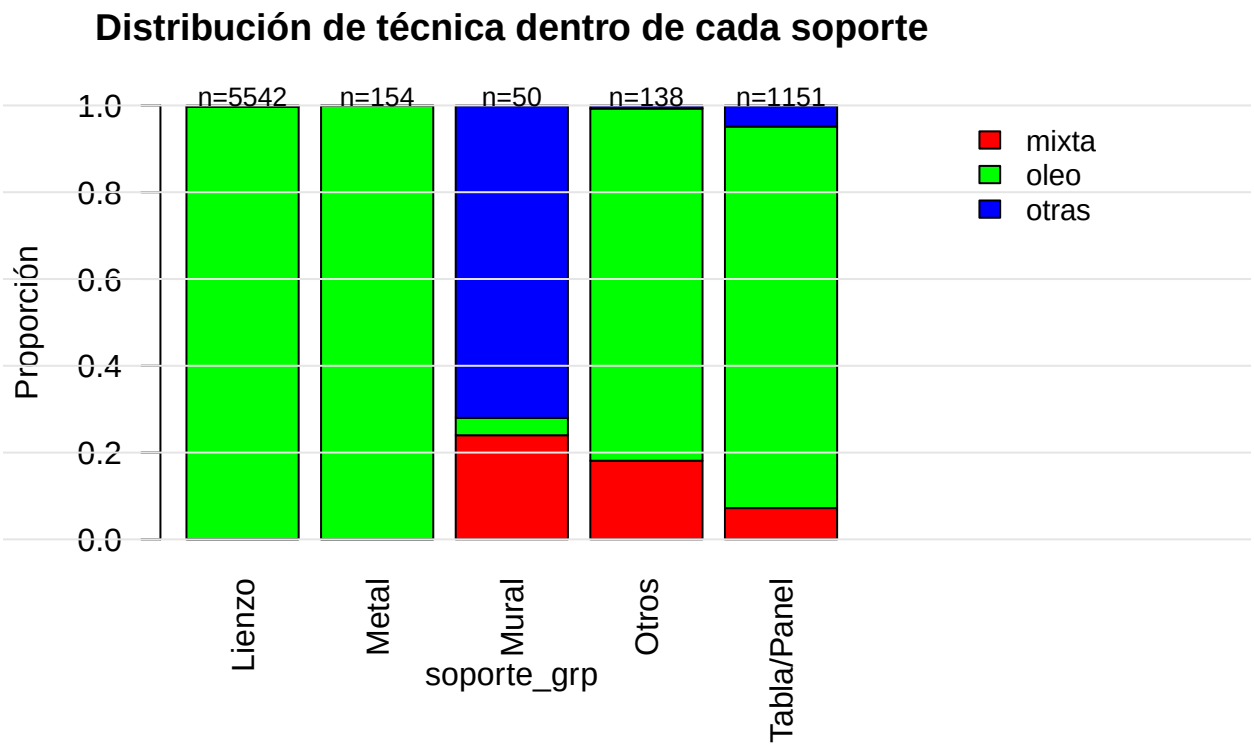


Orientación VS Fecha_est con Éxito

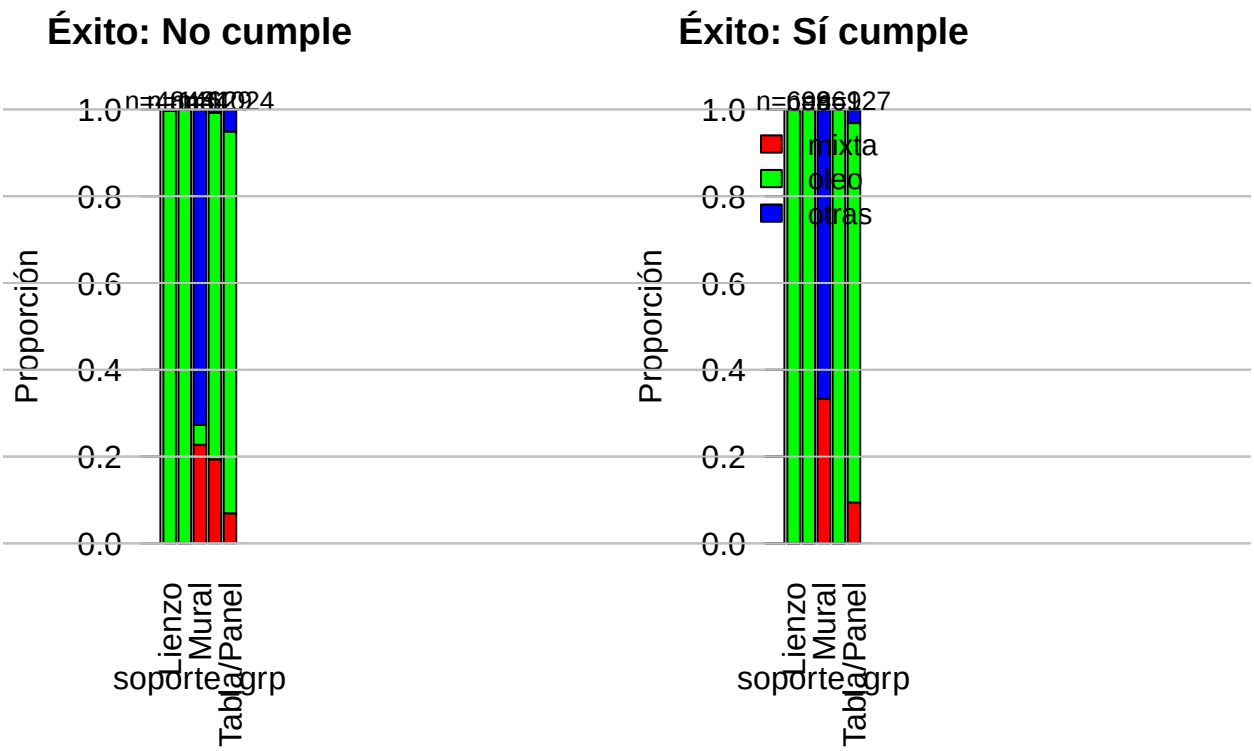




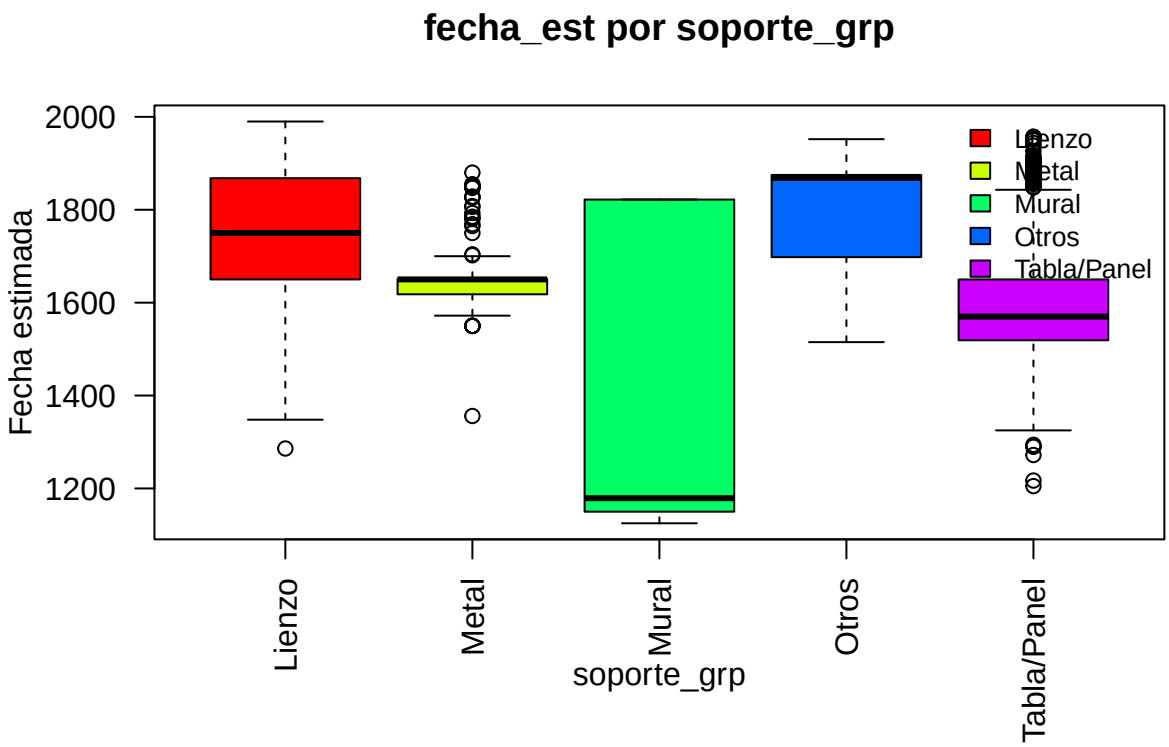
soporte_grp VS Técnica



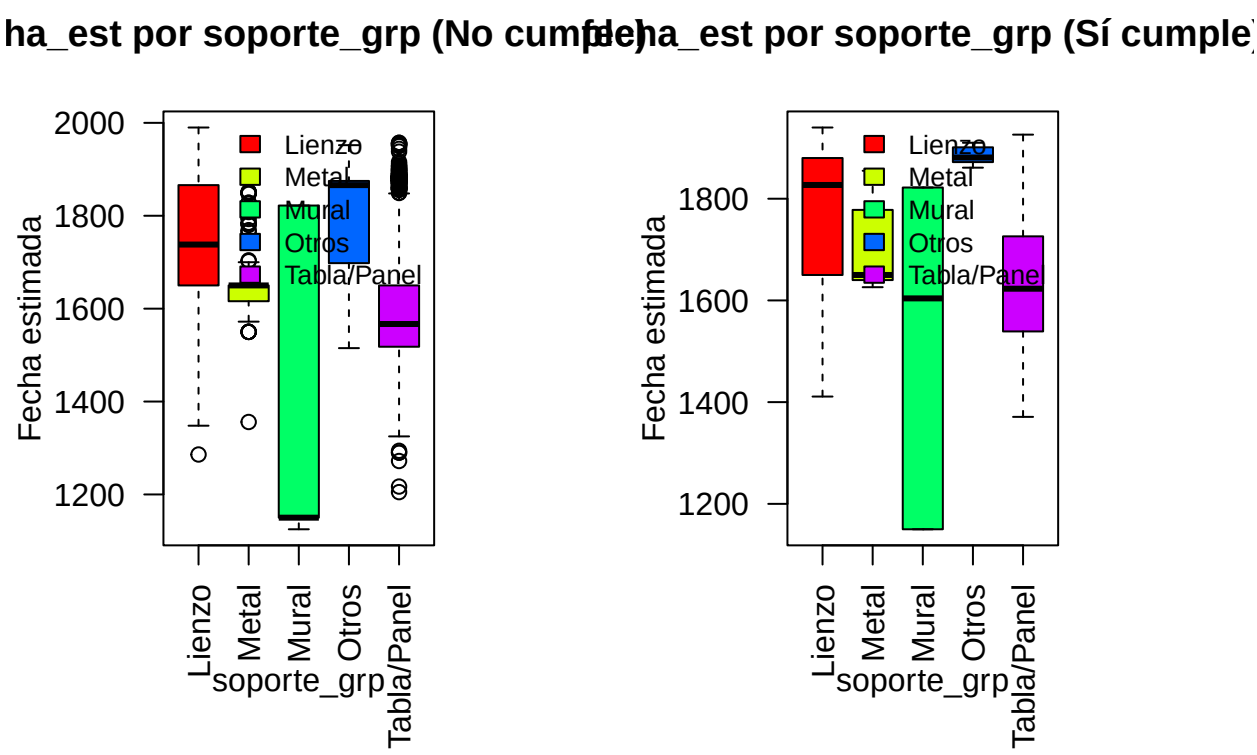
soporte_grp VS Técnica con Éxito



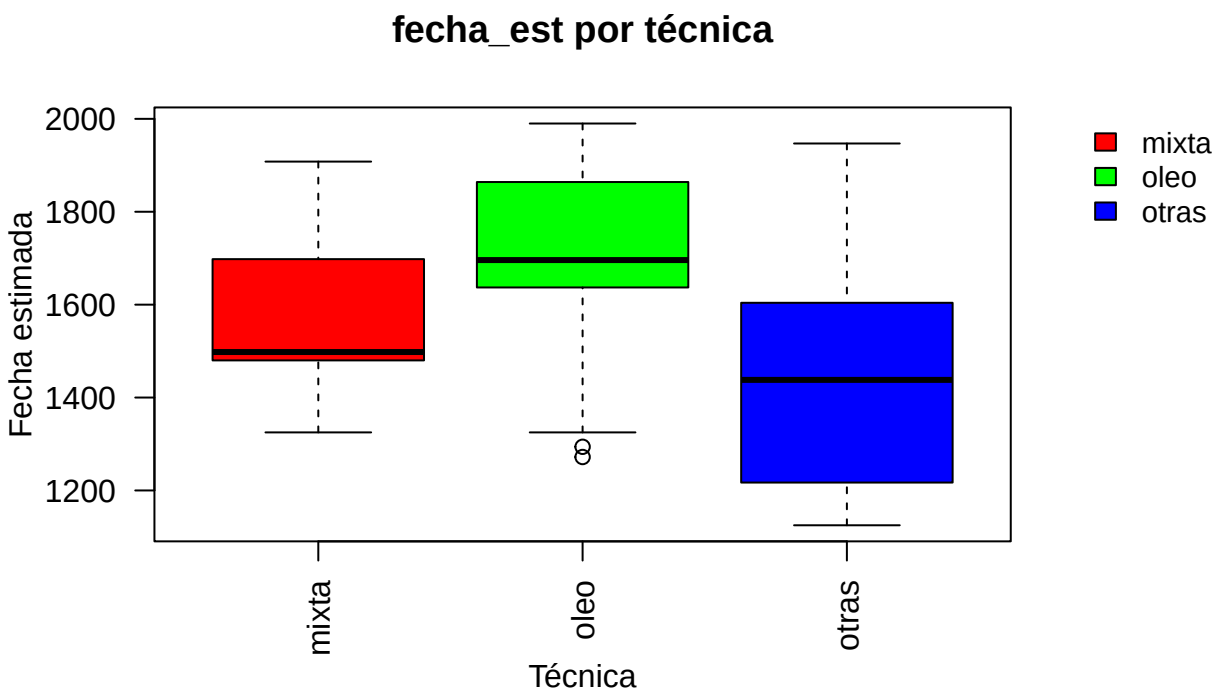
Soporte_grp VS Fecha_estimada



Soporte_grp VS Fecha_estimada con Éxito

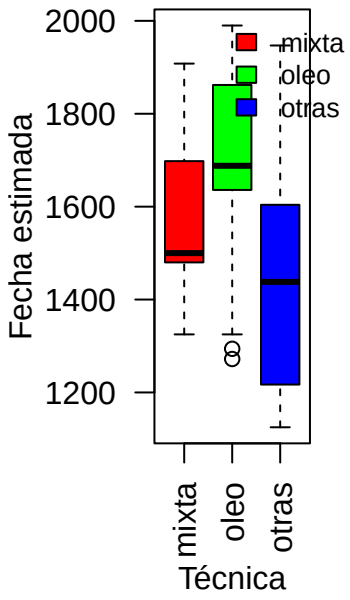


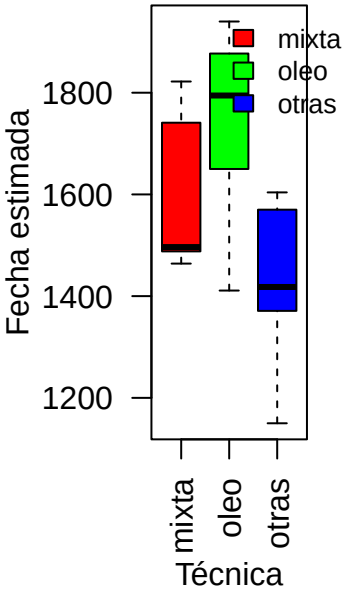
Técnica VS Fecha_estimada



Técnica VS Fecha_estimada con Éxito

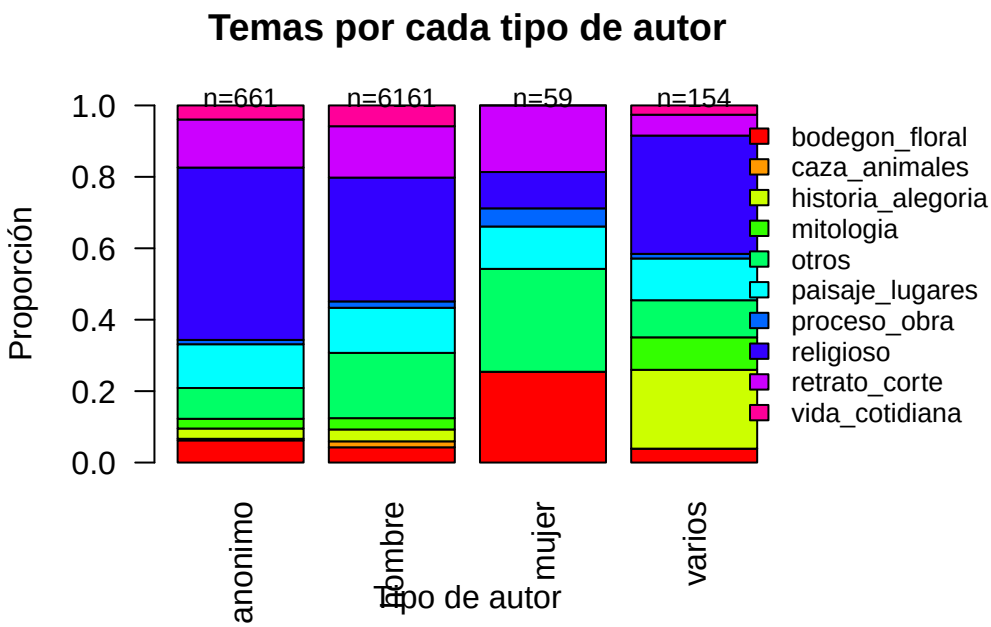
ıa_est por técnica (No cumple) **fecha_est por técnica (Sí cumple)**





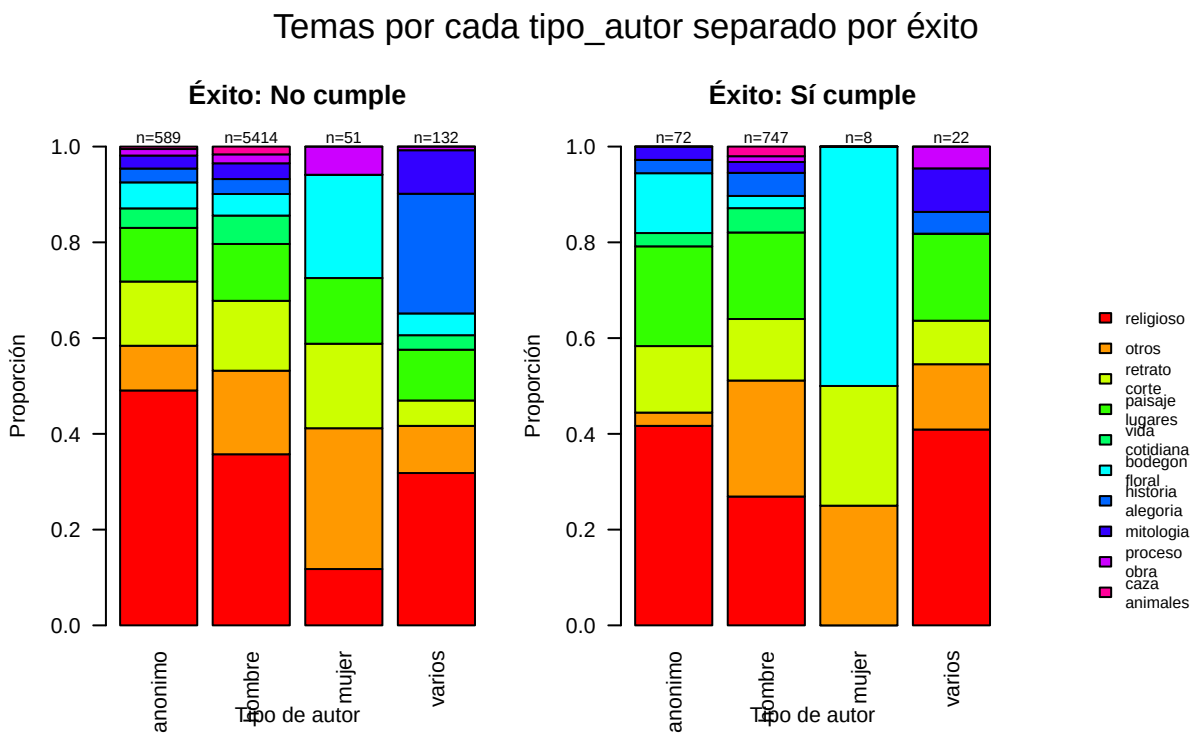
Tipo_autor VS Tema

```
##
##          bodegon_floral caza_animales historia_alegoria mitologia otros
##   anonimo             41             3                19        18    57
##   hombre             263           103               205       195  1126
##   mujer               15             0                 0         0    17
##   varios               6             0                34        14    16
##
##          paisaje_lugares proceso_obra religioso retrato_corte vida_cotidiana
##   anonimo              81            8       319            89             26
##   hombre              778          110      2136           886            359
##   mujer                 7            3         6            11              0
##   varios               18            2        51             9              4
##
##          bodegon_floral caza_animales historia_alegoria mitologia otros
##   anonimo          0.062         0.005             0.029     0.027 0.086
##   hombre           0.043         0.017             0.033     0.032 0.183
##   mujer            0.254         0.000             0.000     0.000 0.288
##   varios           0.039         0.000             0.221     0.091 0.104
##
##          paisaje_lugares proceso_obra religioso retrato_corte vida_cotidiana
##   anonimo           0.123        0.012     0.483         0.135          0.039
##   hombre            0.126        0.018     0.347         0.144          0.058
##   mujer             0.119        0.051     0.102         0.186          0.000
##   varios            0.117        0.013     0.331         0.058          0.026
```



Los temas no se reparten igual según el tipo de autor: cambian bastante entre anónimo, hombre, mujer y varios. Pero "mujer" y "varios" tienen muy pocos casos, así que ahí las proporciones pueden engañar. Esto sugiere que tipo_autor y tema están relacionados, así que si metemos uno en el modelo el otro puede aportar poco (o habría que agrupar categorías).

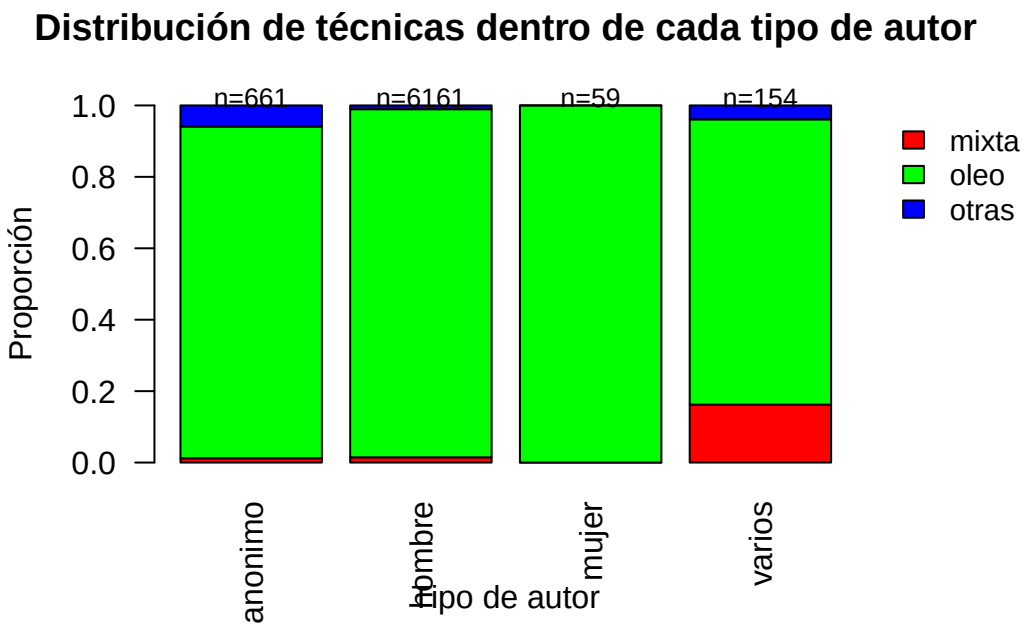
Tipo_autor VS Tema con Éxito



La distribución de temas por tipo de autor es bastante parecida entre "No cumple" y "Sí cumple", no se ve un cambio claro que explique el éxito. Los grupos "mujer" y "varios" tienen muy pocos casos (n muy pequeño), así que cualquier diferencia ahí puede ser casualidad. En el descriptivo, tipo_autor no parece una variable clave para el modelo; si metemos tema, tipo_autor probablemente aportará poco.

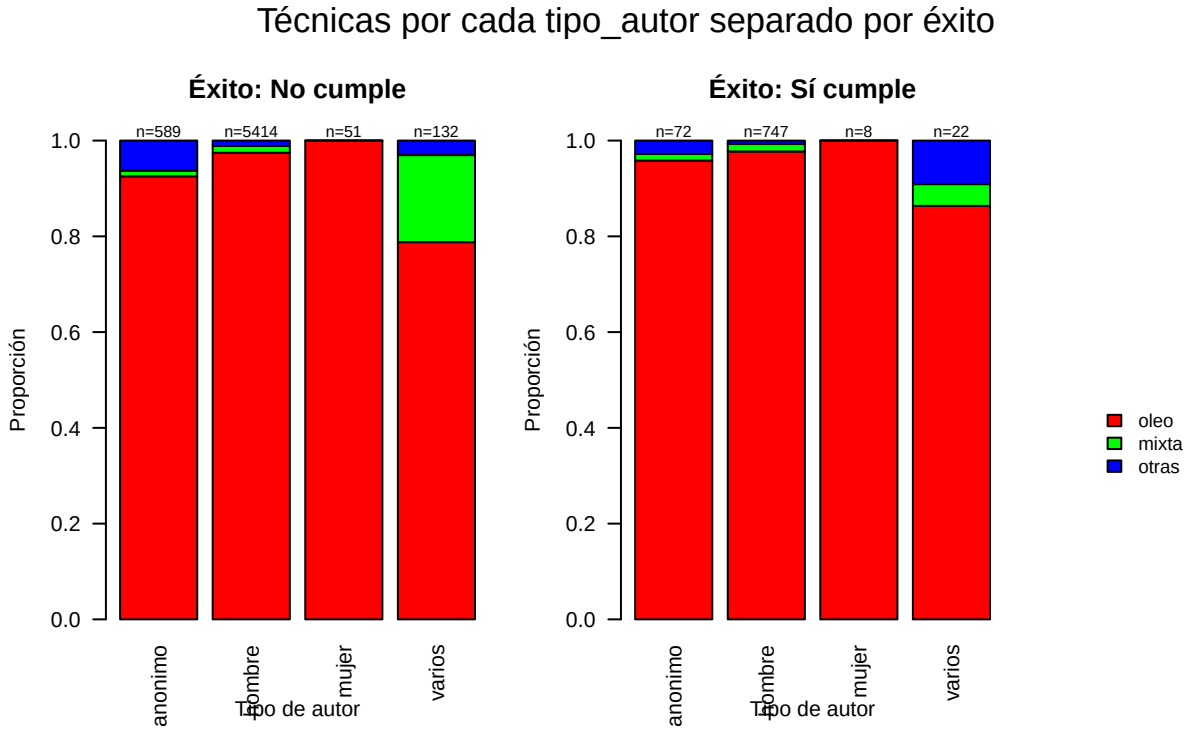
Tipo_autor VS Técnica

```
##
##           mixta oleo otras
##   anonimo     8  614    39
##   hombre     90 6006    65
##   mujer       0   59     0
##   varios     25  123     6
##
##           mixta  oleo otras
##   anonimo 0.012 0.929 0.059
##   hombre  0.015 0.975 0.011
##   mujer   0.000 1.000 0.000
##   varios  0.162 0.799 0.039
```



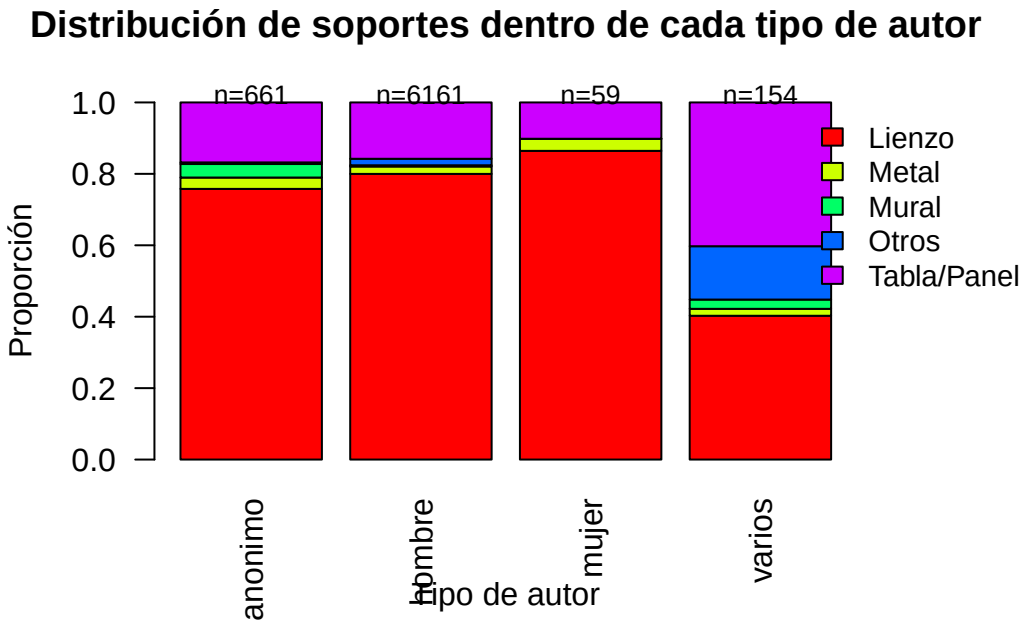
Casi todo es óleo en anónimo/hombre/mujer, así que la técnica apenas cambia entre autores. Con tan poca diferencia, esta variable no parece que vaya a ayudar mucho a explicar el éxito (da poca "señal"). No debería ser prioritaria en el modelo; solo tendría sentido si en el gráfico técnica vs éxito se viera una diferencia clara.

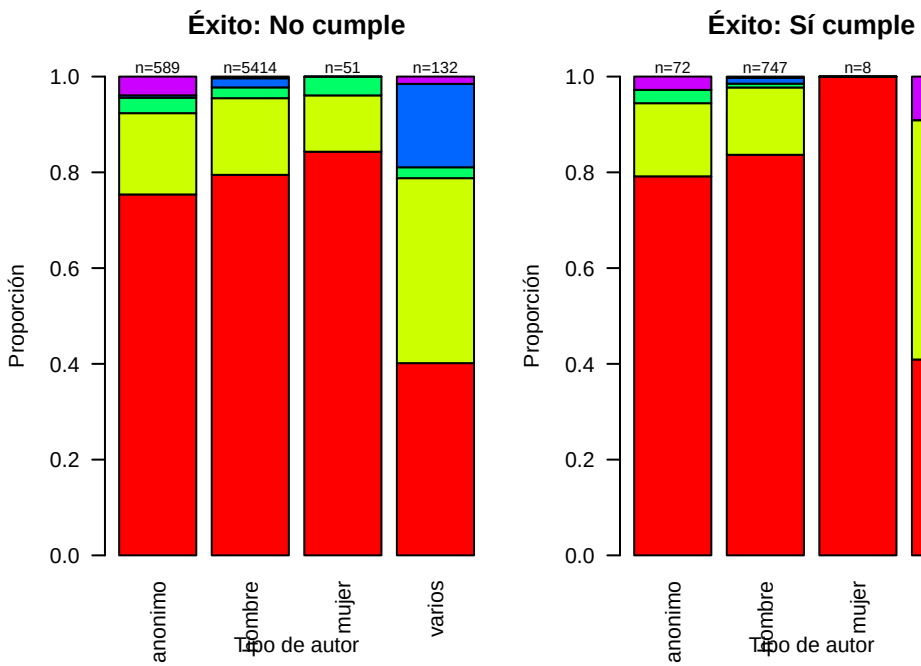
Tipo_autor VS Técnica con Éxito



Tipo_autor VS Soporte_grp

```
##
##           Lienzo Metal Mural Otros Tabla/Panel
##   anonimo    501    21    25     3         111
##   hombre    4928   128    21   112         972
##   mujer       51     2     0     0           6
##   varios      62     3     4    23          62
##
##           Lienzo Metal Mural Otros Tabla/Panel
##   anonimo  0.758 0.032 0.038 0.005       0.168
##   hombre   0.800 0.021 0.003 0.018       0.158
##   mujer    0.864 0.034 0.000 0.000       0.102
##   varios   0.403 0.019 0.026 0.149       0.403
```

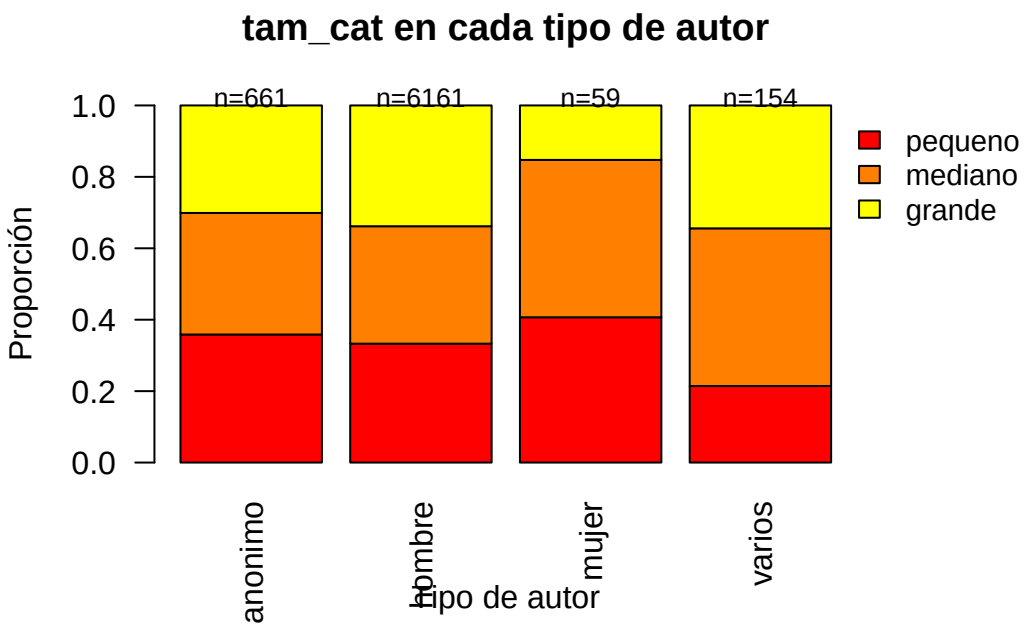

Distribución de soportes dentro de cada tipo de autor
n=661
n=6161
n=59
n=154
1.0
0.8
0.6
0.4
0.2
0.0
Proporción
Lienzo
Metal
Mural
Otros
Tabla/Panel
anonimo
hombre
mujer
varios
Tipo de autor



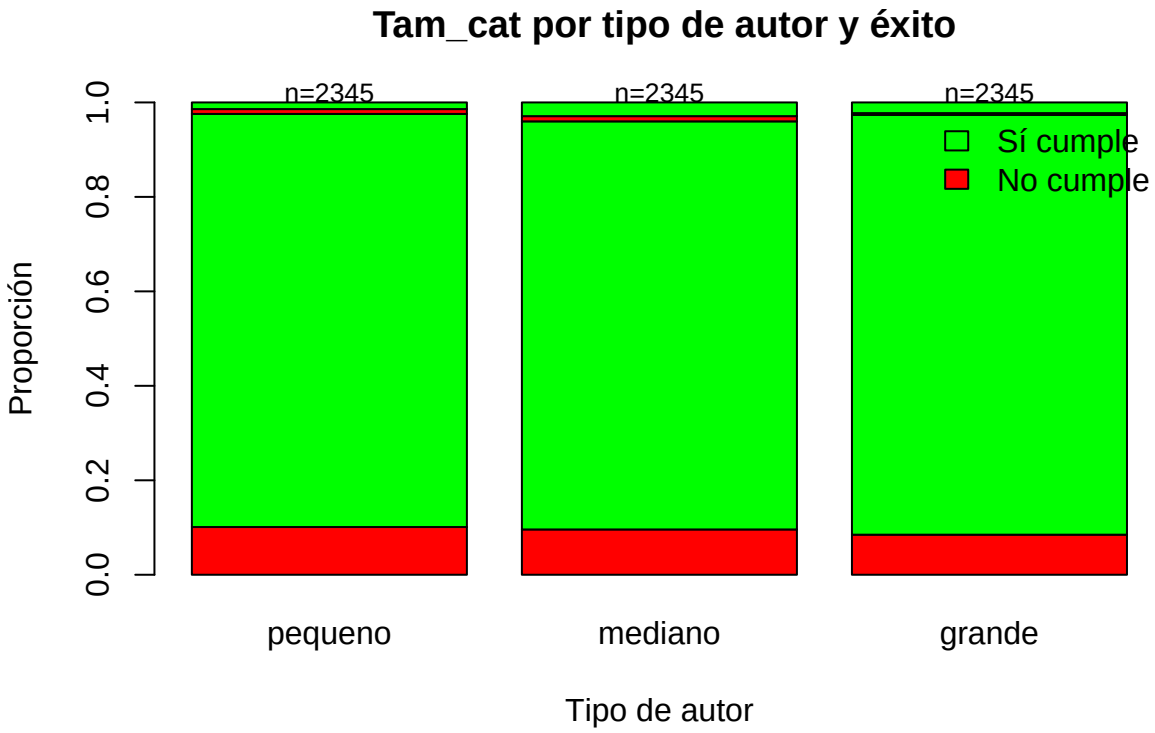
Tipo_autor VS Soporte_grp con Éxito

Tam_cat VS Tipo_autor

```
##
## pequeno mediano  grande    <NA>
##       0       0       0    7035
##
##           pequeno mediano grande
##   anonimo     237     225    199
##   hombre     2051    2026   2084
##   mujer        24      26      9
##   varios       33      68     53
##
##           pequeno mediano grande
##   anonimo   0.359   0.340  0.301
##   hombre    0.333   0.329  0.338
##   mujer     0.407   0.441  0.153
##   varios    0.214   0.442  0.344
```

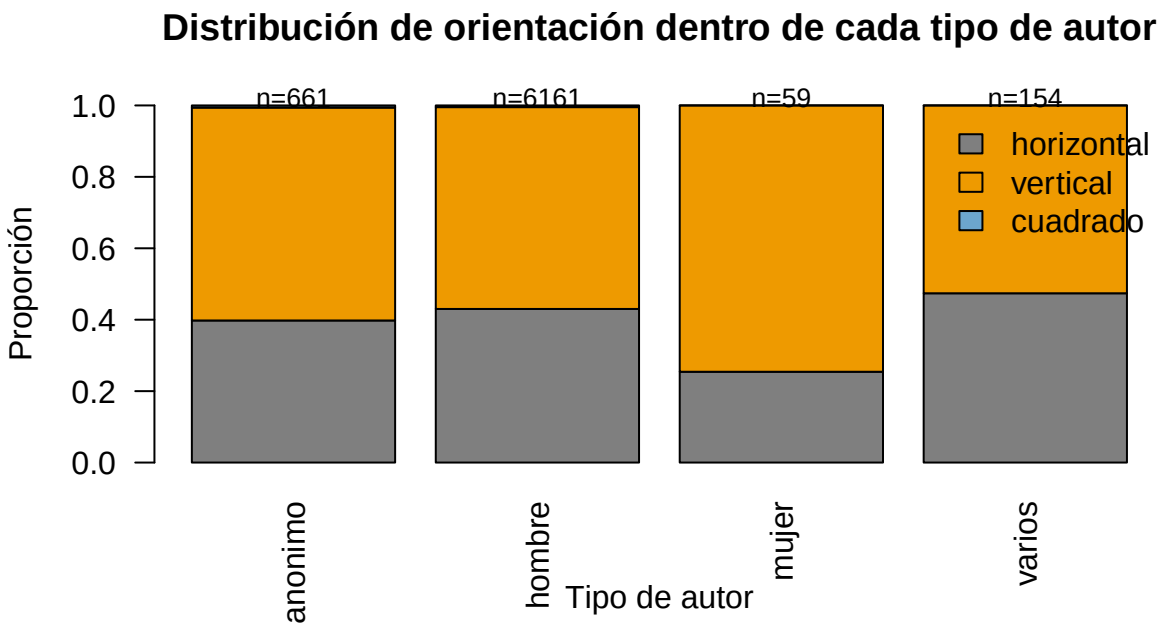


tam_cat VS tipo_autor con Éxito



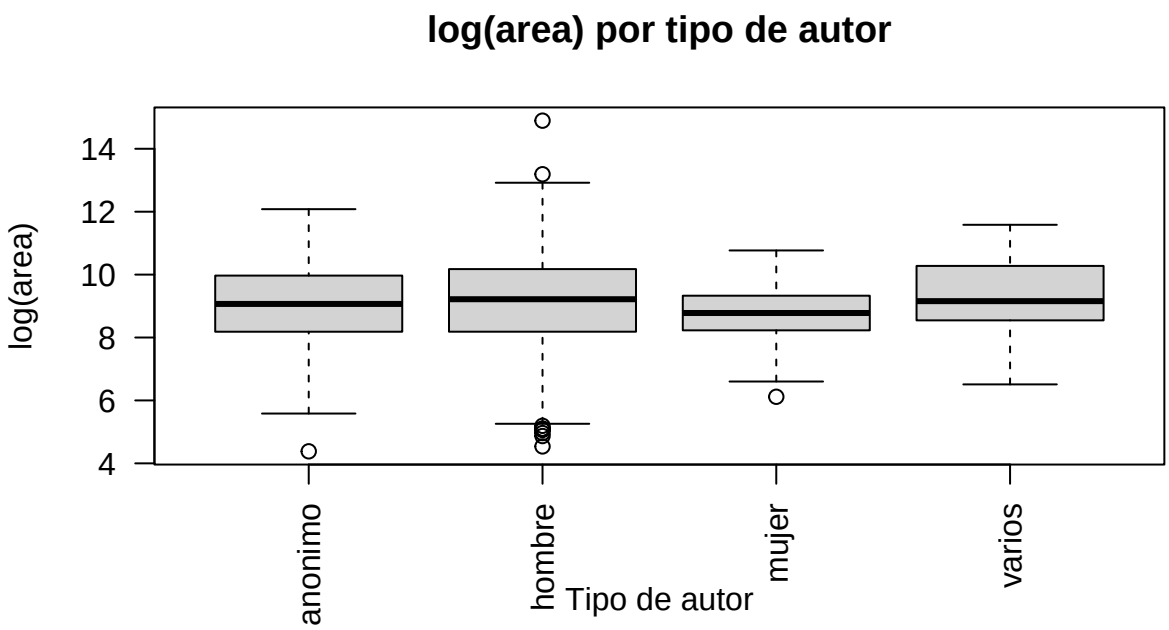
Tipo_autor VS Orientación

```
##
##           horizontal vertical cuadrado
##   anonimo        263      394        4
##   hombre        2650     3482       29
##   mujer           15       44        0
##   varios          73       81        0
##
##           horizontal vertical cuadrado
##   anonimo      0.398    0.596    0.006
##   hombre       0.430    0.565    0.005
##   mujer        0.254    0.746    0.000
##   varios       0.474    0.526    0.000
```

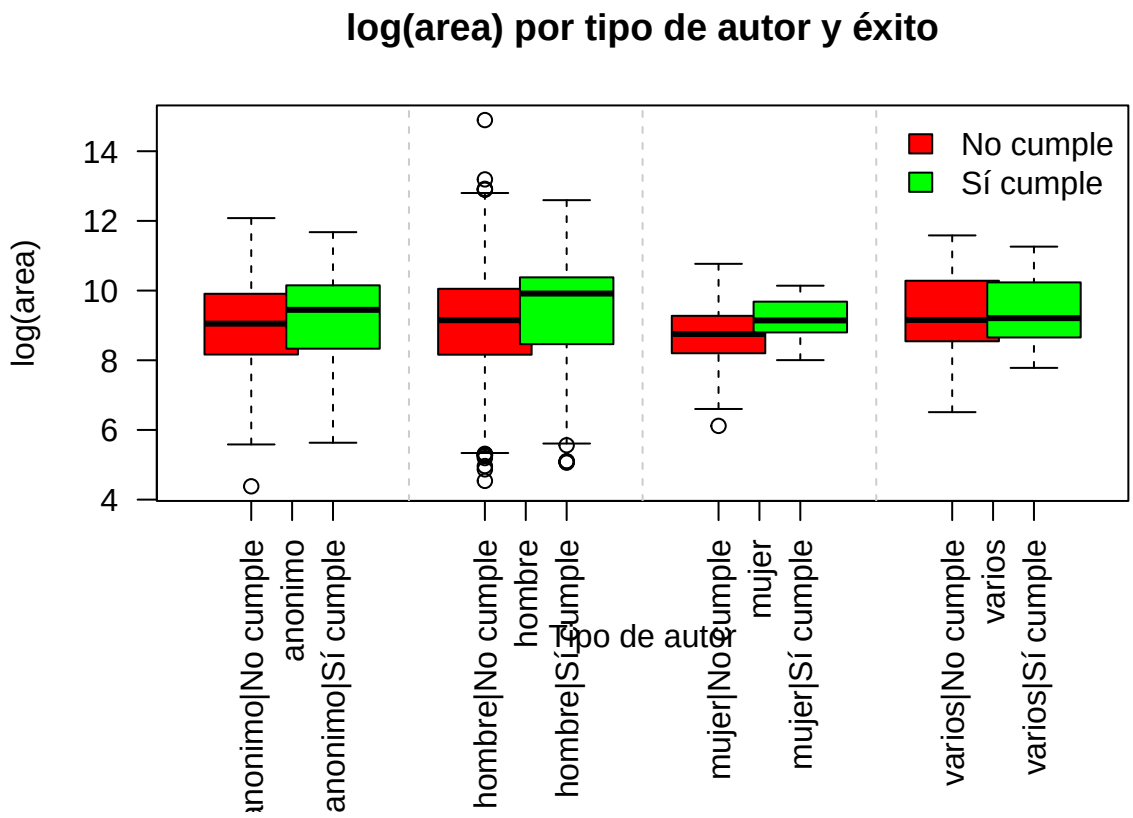


Este gráfico muestra la distribución de la orientación de las obras (horizontal, vertical, cuadrado) según el tipo de autor. Las obras de autores anónimos y hombres son predominantemente horizontales, mientras que las obras de mujeres y varios presentan una mayor proporción de obras verticales. Esta variable (orientación) podría ser relevante para el modelo si se confirma que influye en el éxito, ya que hay diferencias visibles en la distribución por tipo de autor.

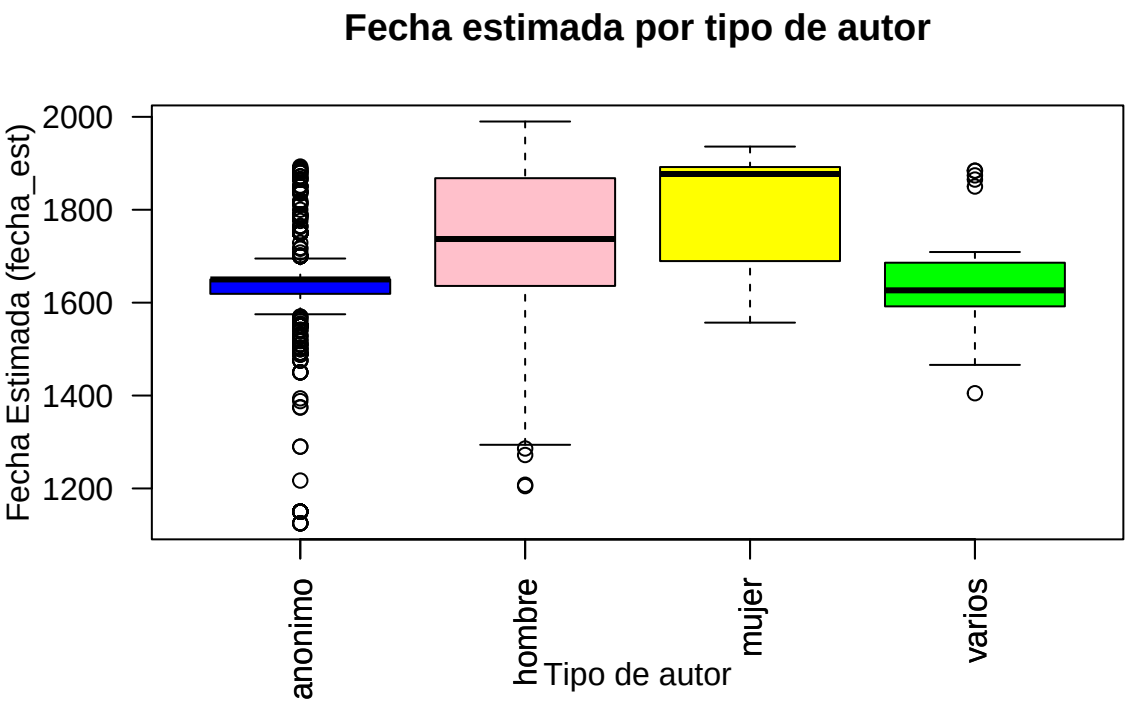
Tipo_autor VS Área

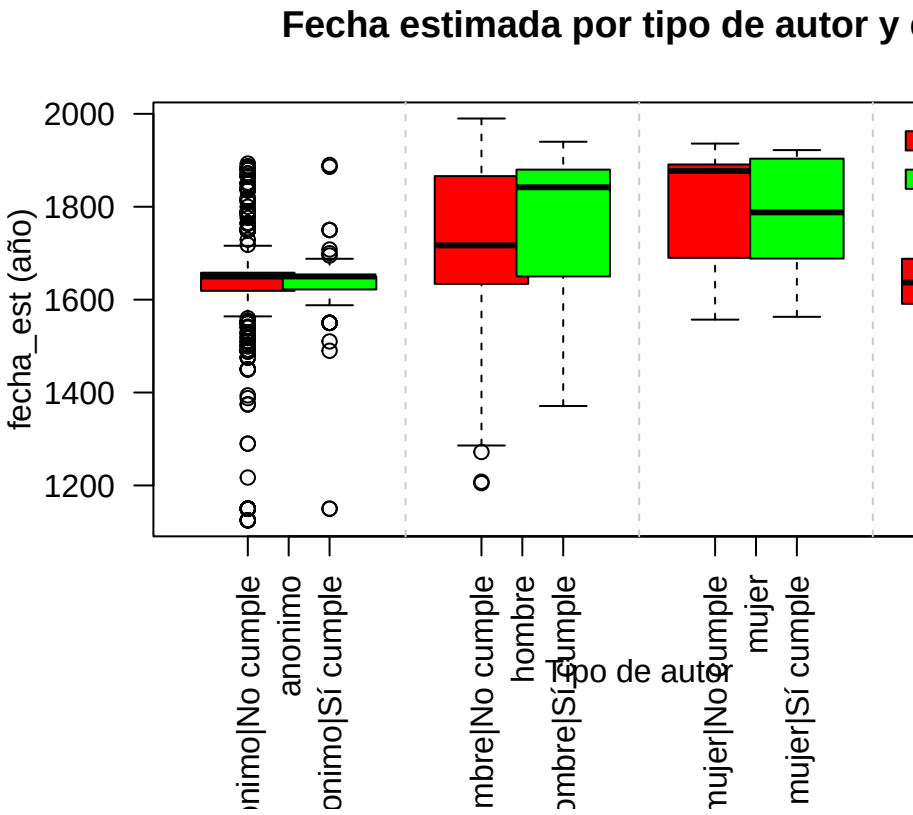


Área por Tipo_autor y Éxito



Tipo_autor VS Fecha_est

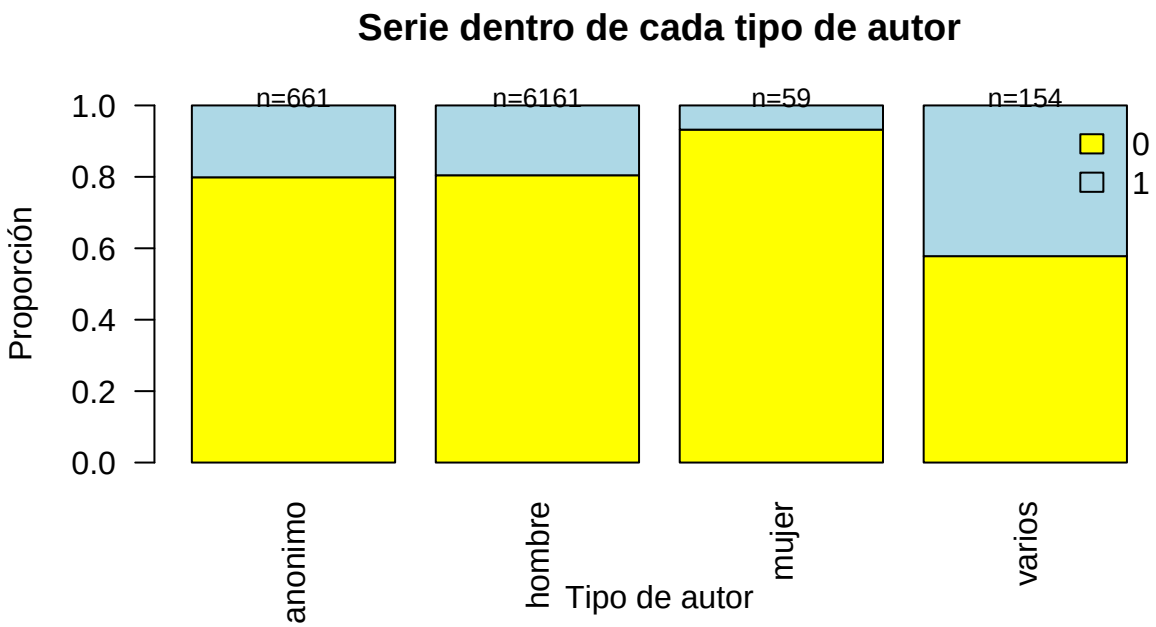

Fecha estimada por tipo de autor
Fecha Estimada (fecha_est)
2000
1800
1600
1400
1200
anonimo
hombre
mujer
varios
Tipo de autor



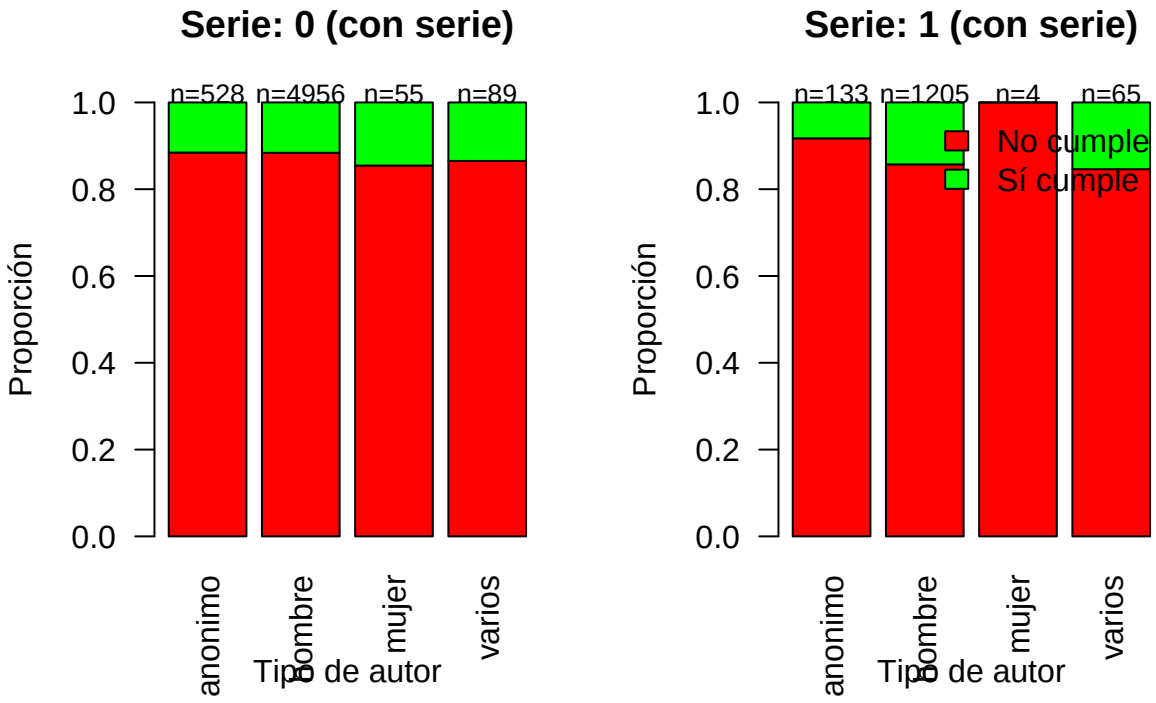
Tipo_autor VS Fecha_est con Éxito

Tipo_autor VS serie

```
##
##              0    1
##   anonimo  528  133
##   hombre  4956 1205
##   mujer     55    4
##   varios    89   65
##
##               0     1
##   anonimo 0.799 0.201
##   hombre  0.804 0.196
##   mujer   0.932 0.068
##   varios  0.578 0.422
```



Tipo_autor VS Serie con Éxito



Resumen:

Hemos revisado todos los gráficos para entender qué variables podrían estar relacionadas con exito, que es nuestra variable respuesta (sí cumple o no cumple la proporción áurea).

En primer lugar, los gráficos y tablas descriptivas muestran que el conjunto de datos está muy desequilibrado por tipo de autor. La gran mayoría de obras aparecen atribuidas a hombre, mientras que los grupos anónimo, varios y especialmente mujer tienen muchos menos casos. Cuando un grupo tiene muy pocas observaciones, cualquier diferencia que parezca grande en un gráfico puede deberse simplemente al tamaño muestral pequeño y no a un efecto real.

Respecto a la variable respuesta, predominan claramente las obras que no cumplen la proporción áurea: hay muchas más observaciones en "no cumple" que en "sí cumple".

En cambio donde se ve un patrón más marcado es en la dimensión temporal. La distribución de fecha_estimada varía mucho según el tipo de autor: en "hombre" hay mucha más dispersión y aparecen más valores extremos (obras muy antiguas), mientras que en "mujer" y "varios" las fechas están más concentradas en un intervalo más estrecho.

Los gráficos indican patrones muy dominantes: casi todas las obras son óleo, el soporte más frecuente es lienzo, y la orientación vertical es la más común. Finalmente pertenecer a una serie es algo minoritario en general, pero en el grupo "varios" aparece con más frecuencia que en el resto.

La variable que más destaca en los gráficos es fecha_estimada. Al separar las obras por exito, observamos que las que sí cumplen tienden a concentrarse más en un rango de fechas mientras que las que no cumplen aparecen más dispersas y con más valores extremos. Además, esta dispersión no es igual para todos: por tipo_autor vemos mucha más variación en hombre y en parte en anónimo, mientras que mujer y varios están más concentrados pero aquí tenemos que insistimos en que mujer y varios tienen pocos datos y eso puede generar ruido. Con area ocurre algo parecido: el área es muy asimétrica y por eso trabajamos con log(area).

En cuanto a las variables categóricas (soporte_grp, tecnica, orientacion, tema, tam_cat y serie) los gráficos describen cómo se distribuyen las obras y que muchas de ellas están relacionadas entre sí. Por ejemplo tema cambia bastante según tipo_autor, y soporte y tecnica están muy conectadas con la fecha (aparecen más en unas épocas que en otras). Esto es importante en un estudio explicativo porque significa que algunas asociaciones aparentes con exito pueden estar "mezcladas" con otra variable (especialmente con la fecha). Por eso más que buscar reglas simples del tipo "este soporte siempre cumple más", interpretamos que el éxito puede estar influido por un conjunto de factores relacionados y que el modelo nos servirá para separar efectos y controlar confusiones entre variables.

Nosotros consideramos interacciones solo cuando los gráficos sugieren que una relación depende de un grupo.

Podemos intuir que hay indicios de que Fecha_estimada afecta a exito, ya que al separar por sí/no se observa un patrón consistente de mayor concentración temporal en los casos que sí cumplen y mayor dispersión en los que no cumplen.

Por otro lado también intuimos que el tamaño afecta a exito, por lo que incluimos log(area) (en lugar de area) para trabajar en una escala más estable y menos dominada por valores extremos.

Respecto a tema creemos que no afecta de manera clara al éxito según el descriptivo, pero lo incluimos como análisis alternativo porque forma parte del planteamiento y comprobamos si añade explicación o si su aparente relación se debe a que tema está asociado a otras variables. También mantenemos técnica y soporte en el modelo porque son variables relevantes del contexto de la obra y porque en los gráficos se relacionan con fecha y tamaño, de modo que pueden actuar como factores explicativos o de ajuste incluso aunque su relación directa con exito no sea uniforme.

Para un enfoque explicativo proponemos como interacción principal soporte_grp $\times$ log(area); la medida no significa lo mismo en todos los soportes (por ejemplo, un mural y un lienzo tienen rangos y usos distintos), por lo que es razonable que la relación entre tamaño y exito dependa del soporte. Somos conscientes de que esta interacción puede generar problemas si hay combinaciones con pocos casos, por lo que la controlamos agrupando categorías poco frecuentes y revisando que no existan celdas casi vacías.

5. Análisis principal

Primeramente generamos la submuestra de la población aplicando la regla estructural de 'cuadrado'.

```
## [1] 7002   12
```

Continuamos con un total de 7002 pinturas

Fijamos también los niveles de referencia para las variables factor. El nivel más frecuente como referencia para todos los factores a excepción de "tam_cat" donde por interpretabilidad se define la referencia en 'pequeno' y las variables binarias "sop_montaje" y "serie" donde se fija el valor "no"

```
## 'data.frame':    7002 obs. of  12 variables:
##  $ exito      : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ area       : num  81423 81423 52600 37000 44955 ...
##  $ tam_cat    : Factor w/ 3 levels "pequeno","mediano",..: 3 3 3 3 3 3 3 2 3 3 ...
##  $ orientacion: Factor w/ 2 levels "vertical","horizontal": 2 2 2 1 2 2 2 1 2 2 ...
##  $ soporte_grp: Factor w/ 5 levels "Lienzo","Metal",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ sop_montaje: Factor w/ 2 levels "no","si": 2 2 2 2 2 2 2 2 2 2 ...
##  $ tecnica    : Factor w/ 3 levels "oleo","mixta",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ tipo_autor : Factor w/ 4 levels "hombre","anonimo",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ serie      : Factor w/ 2 levels "no","si": 2 2 2 2 2 2 2 2 2 2 ...
##  $ fecha_est  : int  1150 1150 1150 1150 1150 1150 1150 1150 1150 1150 ...
##  $ fecha_ancho: int  99 99 99 99 99 99 99 99 99 99 ...
##  $ tema       : Factor w/ 10 levels "religioso","bodegon_floral",..: 1 1 1 1 1 1 1 1 1 1 ...
```

5.1 Efectos principales

Para determinar los efectos principales del modelo se seguirá la estructura por bloques definida en al sección de metodología

Modelo nulo

Como punto de partida se ajustó un modelo nulo (solos intercepto), sin covariables. Este modelo proporcionará la referencia sobre el cual iremos cuantificando el aporte de los bloques que se añadirán sucesivamente. En una regresión logística como la nuestra, el intercepto del modelo nulo estima la probabilidad media de éxito en la muestra (convertido en la escala correcta). Recordemos que estaremos trabajando en todo momento con la sub-muestra no-cuadrado.

```
##
## Resumen del modelo:
##
## Call:
## glm(formula = exito ~ 1, family = binomial(link = "logit"), data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.98064    0.03661   -54.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 5173.2  on 7001  degrees of freedom
## AIC: 5175.2
##
## Number of Fisher Scoring iterations: 4
##
##
## Probabilidad estimada de éxito:
## (Intercept)
##   0.1212511
```

Como ya sabiamos, la probabilidad de observar el evento de interés es baja, concretamente de un 12%.

Bloque 1: Datación y control de incertidumbre

Se incorpora la variable "fecha_est" y su covariable de incertidumbre "fecha_ancho", comparando especificación lineal vs flexible (spline) en ambas variables. Comenzaremos utilizando la especificación flexible para "fecha_est" y provando ambas especificaciones para "fecha_ancho". Además "fecha_ancho" se incorpora mediante "log1p".

Comenzamos decidiendo la especificación del control "fecha_ancho".

```
##
## Resumen del modelo m1_lin (control lineal):
##
## Call:
## glm(formula = exito ~ ns(fecha_est, 3) + log_ancho, family = binomial(link = "logit"),
##     data = df)
##
```

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.196862   0.566571  -3.877 0.000106 ***
## ns(fecha_est, 3)1  0.281381   0.368057   0.765 0.444567
## ns(fecha_est, 3)2  0.558190   1.192596   0.468 0.639751
## ns(fecha_est, 3)3  1.153682   0.297870   3.873 0.000107 ***
## log_ancho          0.005883   0.022782   0.258 0.796244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 5120.5  on 6997  degrees of freedom
## AIC: 5130.5
##
## Number of Fisher Scoring iterations: 4
##
##
## Resumen del modelo m1_spl (control flexible spline):
##
## Call:
## glm(formula = exito ~ ns(fecha_est, 3) + ns(log_ancho, 3), family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.235466   0.571676  -3.910 9.22e-05 ***
## ns(fecha_est, 3)1  0.276625   0.369477   0.749 0.454041
## ns(fecha_est, 3)2  0.730706   1.206027   0.606 0.544596
## ns(fecha_est, 3)3  1.135879   0.300691   3.778 0.000158 ***
## ns(log_ancho, 3)1 -0.283268   0.340571  -0.832 0.405553
## ns(log_ancho, 3)2  0.009094   0.236265   0.038 0.969298
## ns(log_ancho, 3)3  0.431797   0.283258   1.524 0.127410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 5117.3  on 6995  degrees of freedom
## AIC: 5131.3
##
## Number of Fisher Scoring iterations: 4
```

Comprobamos la aportación del bloque comparando ambas especificaciones con el modelo nulo:

```
##
## Aporte de información del Bloque 1 (control lineal):
## Analysis of Deviance Table
##
## Model 1: exito ~ 1
## Model 2: exito ~ ns(fecha_est, 3) + log_ancho
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
```

```
## 1      7001     5173.2
## 2      6997     5120.5  4   52.682 9.933e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de información del Bloque 1 (control flexible spline):
## Analysis of Deviance Table
##
## Model 1: exito ~ 1
## Model 2: exito ~ ns(fecha_est, 3) + ns(log_ancho, 3)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7001     5173.2
## 2      6995     5117.3  6   55.875 3.085e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de complejidad:
##        df      AIC
## m0      1 5175.221
## m1_lin  5 5130.540
## m1_spl  7 5131.346
##        df      BIC
## m0      1 5182.075
## m1_lin  5 5164.809
## m1_spl  7 5179.324
```

Comparamos también las dos formas funcionales del control entre si, mediante modelos anidados:

```
##
## Aporte de información entre opciones (m1_lin vs m1_spl)
## Analysis of Deviance Table
##
## Model 1: exito ~ ns(fecha_est, 3) + log_ancho
## Model 2: exito ~ ns(fecha_est, 3) + ns(log_ancho, 3)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6997     5120.5
## 2      6995     5117.3  2   3.1933   0.2026
```

El aporte del bloque 1 es claramente significativo en las dos opciones, ya que ambas especificaciones mejoraron significativamente el modelo nulo (LRT $p = 9.933e-11$, $p = 3.085e-10$). Sin embargo, la especificación flexible para el control "fecha_ancho" no proporcionó mejora adicional frente a la lineal (LRT $p = 0.2026$) y presentó peor ajuste penalizado por complejidad (AIC y BIC mayores). Por esta razón, se adoptó para los modelos posteriores la especificación lineal $log(1 + fecha\_ancho)$ como ajuste definitivo del control de incertidumbre en la datación.

Para finaizar este bloque, comprobaremos las formas funcionales de "fecha_est" para ver si la especificación flexible de esta variable es necesaria para nuestro modelo o preferimos su versión simplificada (lineal).

```
##
## Aporte de la opción fecha_est flexible:
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho
## Model 2: exito ~ ns(fecha_est, 3) + log_ancho
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6999     5126.6
## 2      6997     5120.5  2   6.0146  0.04943 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Complejidad de la opción fecha_est flexible:
##              df      AIC
## m1_lin        5 5130.540
## m1_lin_fecha  3 5132.554
##              df      BIC
## m1_lin        5 5164.809
## m1_lin_fecha  3 5153.116
```

Vemos que la especificación felxible es preferible frente a la lineal en terminos de información (LRT $p = 0.049$) como AIC ($\triangle AIC \approx -2$). Sin embargo la mejora de AIC es débil frente a una fuerte penalización en BIC ($\triangle AIC \approx -2$ vs $\triangle BIC \approx 12$). Teniendo en cuenta nuestro objetivo descriptivo, y no predictivo, decidimos seleccionar la versión lineal por facilidad interpretativa. Por otro lado, también seleccionamos esta opción por ser más conservadora, ya que añadimos menos parámetros al modelo (2 parámetros menos), y és preferible dado que la mayoría de las futuras covariables són facotres y nuestra variables respuesta está fuertemente desbalanceada. Sin embargo, dejamos constancia del hecho que la especificación flexible con spline par fecha_est podría ser considerada para análisis más exaustivos.

Modelo resultante después de añadir Bloque 1:

Bloque 2: Morfología

Se incorporan las variables morfológicas de tamaño i formato, "area" y "orientación" respectivamente. Se evaluará si aportan infomración adicional, una vez controlado el efecto de datación (bloque 1). La variable "area" se introducirá mediante una transformación logarítmica $log(area)$, pudiendo llegar a ser tratada de manera flexible si fuera necesario. También se proporcinará un modelo secundario sustituyendo "area" por "tam_cat" (categorización de área), y se compararán. Se decidirá el mejor modelo siguiendo las idicaciones de la sección de metodología, que resumidamente dictan lo siguiente:

- si "area" presenta problemas o no mejora sustancialmente más que "tam_cat", permitimos ajuste flexible (spline) y si no funciona elejimos "tam_cat"
- si "area" mejora sustanciamente más que "tam_cat", elejimos "area"
- si "area" y "tam_cat" proporcionan resultados cualitativamente iguales, elejimos "area" pero manteniendo el modelo secundario "tam_cat" para interpretaciones claras.

```
##
## Resumen del modelo m2_area (+ log_area + orientacion):
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + log_area + orientacion,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -7.4602560  0.6658251 -11.205  < 2e-16 ***
## fecha_est              0.0020017  0.0003269   6.124 9.15e-10 ***
## log_ancho              0.0059186  0.0227367   0.260    0.795
## log_area               0.1989546  0.0278371   7.147 8.86e-13 ***
## orientacionhorizontal  0.3267555  0.0748131   4.368 1.26e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 5048.6  on 6997  degrees of freedom
## AIC: 5058.6
##
## Number of Fisher Scoring iterations: 5
##
##
## Resumen del modelo m2_tamcat (+ tam_cat + orientacion):
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -6.0159849  0.6000728 -10.025  < 2e-16 ***
## fecha_est              0.0020633  0.0003304   6.244 4.26e-10 ***
## log_ancho              0.0101612  0.0228402   0.445    0.656
## tam_catmediano        -0.0594074  0.1034091  -0.574    0.566
## tam_catgrande          0.7738667  0.0898896   8.609  < 2e-16 ***
## orientacionhorizontal  0.2962533  0.0751562   3.942 8.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4988.1  on 6996  degrees of freedom
## AIC: 5000.1
##
## Number of Fisher Scoring iterations: 5
##
##
## Aporte de información del Bloque 2 (tamaño continuo):
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho
## Model 2: exito ~ fecha_est + log_ancho + log_area + orientacion
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      6999     5126.6
## 2      6997     5048.6  2   77.907 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de información del Bloque 2 (tamaño categórico):
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion
```

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      6999     5126.6
## 2      6996     4988.1  3   138.45 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de complejidad:
##           df      AIC
## m1         3 5132.554
## m2_area    5 5058.647
## m2_tamcat  6 5000.103
##           df      BIC
## m1         3 5153.116
## m2_area    5 5092.917
## m2_tamcat  6 5041.227
```

Ambas opciones del tamaño (log(area) continua vs. tam_cat categórica) aportan información adicional tras controlar la datación (LRT $p < 2.2e-16$ en ambos casos) y mejorar los criterios de información. Sin embargo, la especificación categórica presenta mejor ajuste–complejidad con AIC y BIC sustancialmente menores (AIC: 5058 vs 5000; BIC: 5092 vs 5042).

Antes de seleccionar "tam_cat" debemos tener en cuenta que podría haber una relación no lineal que actualmente $log(area)$ no esta pudiendo capturar. Por esta razón y dado que "area" ha demostrado mejorar el ajuste, frente al modelo anterior (m1), flexibilizaremos su especificacion (spline) y entonces volveremos a comparar con "tam_cat", para asegurar una decisión justa y cerrada.

```
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + splines::ns(log_area,
##     3) + orientacion, family = binomial(link = "logit"), data = df)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -4.6151418  0.7103089  -6.497 8.17e-11 ***
## fecha_est                   0.0021365  0.0003311   6.453 1.10e-10 ***
## log_ancho                   0.0092053  0.0228401   0.403   0.6869
## splines::ns(log_area, 3)1   1.0901352  0.2248946   4.847 1.25e-06 ***
## splines::ns(log_area, 3)2 -4.0254449  0.9980720  -4.033 5.50e-05 ***
## splines::ns(log_area, 3)3 -2.0573173  0.8145897  -2.526   0.0116 *
## orientacionhorizontal       0.3319577  0.0759994   4.368 1.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 5007.8  on 6995  degrees of freedom
## AIC: 5021.8
##
## Number of Fisher Scoring iterations: 5
##
##
## Aporte de la opción log(area) flexibleAnalysis of Deviance Table
##
```

```
## Model 1: exito ~ fecha_est + log_ancho
## Model 2: exito ~ fecha_est + log_ancho + splines::ns(log_area, 3) + orientacion
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      6999     5126.6
## 2      6995     5007.8  4   118.74 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Complejidad de la opción log(area) flexible            df      AIC
## m2_area      5 5058.647
## m2_area_spl  7 5021.812
## m2_tamcat    6 5000.103
##             df      BIC
## m2_area      5 5092.917
## m2_area_spl  7 5069.789
## m2_tamcat    6 5041.227
```

El modelo con $log(area)$ lineal mejoró el ajuste, y al permitir no linealidad, el ajuste mejoró frente al lineal (AIC: 5021 vs 5058). Sin embargo, la especificación categórica "tam_cat" presentó el mejor compromiso ajuste–complejidad, con AIC y BIC claramente inferiores (AIC: 5000 vs 5021); BIC: 5041 vs 5069), superando también a la versión flexible del tamaño continuo. Por ello, se seleccionó "tam_cat" como representación principal del tamaño para los modelos posteriores.

También se considera manejar en un modelo alternativo la opción de especificación continua del tamaño en versión lineal ($log(area)$ sin spline). El único obejtivo de mantener esta opción como alternativa, en vez de la flexible, es el de aportar un modelo más parsimonioso. Aunque su versión flexible mostró un mejor ajuste, vemos que añade incluso 1 parámetro que "tam_cat" y preferimos no añadir más complejidad para los analisis alernativos, teniendo la opción lineal que ya cumple con el requisito de aporte de información.

Modelo resultante después de añadir Bloque 2:

Opción alternativa: especificación continua del tamaño en versión lineal ($log(area)$)

Bloque 3: Material y técnica

Se incorporan las variables "soporte" y "tecnica" para evaluar si aportan infomración adicional en conjunto, una vez controlados los efectos de datación (Bloque 1) y morfología (Bloque 2). Se evaluará adicionalmente la inclusión de "sop_montaje" como extensión del bloque 3, comparando el modelo con y sin dicha covariable.

```
##
## Resumen del modelo m3_base (soporte_grp + tecnica):
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion +
##     soporte_grp + tecnica, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -7.0808549  0.6673046 -10.611  < 2e-16 ***
## fecha_est              0.0025950  0.0003592   7.224 5.05e-13 ***
## log_ancho              0.0194664  0.0230144   0.846   0.3976
## tam_catmediano        -0.0004402  0.1083753  -0.004   0.9968
## tam_catgrande          0.8764280  0.1014370   8.640  < 2e-16 ***
## orientacionhorizontal  0.3073077  0.0757811   4.055 5.01e-05 ***
## soporte_grpMetal      -0.4347888  0.3736719  -1.164   0.2446
```

```
## soporte_grpMural        0.2739620  0.5241350   0.523   0.6012
## soporte_grpOtros       -0.6011812  0.3581133  -1.679   0.0932 .
## soporte_grpTabla/Panel  0.4854756  0.1241065   3.912 9.16e-05 ***
## tecnicamixta            0.3096232  0.3225271   0.960   0.3371
## tecnicaotras           -0.1599554  0.3953591  -0.405   0.6858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4963.0  on 6990  degrees of freedom
## AIC: 4987
##
## Number of Fisher Scoring iterations: 5
##
##
## Aporte de información del Bloque 3 (base):
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      6996     4988.1
## 2      6990     4963.0  6   25.117 0.0003249 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de complejidad:
##         df      AIC
## m2       6 5000.103
## m3_base 12 4986.987
##         df      BIC
## m2       6 5041.227
## m3_base 12 5069.234
```

La inclusión conjunta de "soporte_grp" y "tecnica" produjo una mejora significativa respecto al modelo con datación y morfología (LRT $p = 0.0003$) también mejoró el ajuste penalizado por AIC ($\triangle AIC \approx -13$), aunque el BIC aumentó ($\triangle BIC \approx +28$) (mayor penalización por el gran número de parámetros añadidos, 6 añadidos). Debido al objetivo descriptivo de nuestro estudio, se necide mantener el bloque por su relevancia teòrica y por la evidencia global de aporte de información.

Extenderemos este bloque añadiendo ahora la variable "sop_montaje", de menos interés conceptual pero con posibles implicaciones en el modelo a nivel de control.

```
##
## Resumen del modelo m3_montaje (+ sop_montaje):
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion +
##     soporte_grp + tecnica + sop_montaje, family = binomial(link = "logit"),
##     data = df)
##
```

```
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -7.1200183  0.6700046 -10.627  < 2e-16 ***
## fecha_est               0.0025881  0.0003602   7.184 6.76e-13 ***
## log_ancho               0.0208326  0.0230627   0.903 0.366364
## tam_catmediano          0.0505267  0.1104297   0.458 0.647279
## tam_catgrande           0.9323677  0.1039707   8.968  < 2e-16 ***
## orientacionhorizontal   0.2891818  0.0762205   3.794 0.000148 ***
## soporte_grpMetal       -0.3862326  0.3741296  -1.032 0.301908
## soporte_grpMural       -0.2720792  0.5551693  -0.490 0.624075
## soporte_grpOtros       -1.0478701  0.3926692  -2.669 0.007617 **
## soporte_grpTabla/Panel  0.5158106  0.1246809   4.137 3.52e-05 ***
## tecnicamixta            0.1796075  0.3279779   0.548 0.583952
## tecnicaotras           -0.2057277  0.3960271  -0.519 0.603427
## sop_montajesi           0.6621291  0.2273603   2.912 0.003588 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4955.2  on 6989  degrees of freedom
## AIC: 4981.2
##
## Number of Fisher Scoring iterations: 5
##
##
## Aporte de informacion de sop_montaje (m3_montaje vs m3_base):
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6990     4963.0
## 2      6989     4955.2  1   7.7395 0.005403 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Complejidad de sop_montaje:
##            df      AIC
## m2          6 5000.103
## m3_base    12 4986.987
## m3_montaje 13 4981.247
##            df      BIC
## m2          6 5041.227
## m3_base    12 5069.234
## m3_montaje 13 5070.348
```

Se observó una mejora significativa del ajuste respecto al modelo sin esta covariable (LRT $p = 0.005$). El AIC disminuyó ($\triangle AIC \approx -5$), indicando también una mejora del ajuste teniendo en cuenta el aporte de complejidad, aunque el BIC aumentó ligeramente ($\triangle BIC \approx +1$), el incremento fue pequeño. Por tanto, se retuvo "sop_montaje" en el modelo para los bloques posteriores.

Sin embargo, debemos recordar que aunque el incremento en BIC para la incluión de "sop_montaje" fue pequeño, la inclusión de Bloque 3 ya produjo aumento fuerte en BIC por lo que la complejidad añadida de todo el bloque más el extra sí representa un valor sustancial ($\triangle_{total} BIC \approx +29$). Por ello se contempla la opción de un modelo alternativo sin este bloque, con el objetivo de proporcionar modelos más convervadores en algunos aspecos que permitan analizar otros, aún siendo conscientes de podrían absorberse los efectos de este bloque.

Modelo resultante después de añadir Bloque 3:

Opción alternativa: modelo sin Bloque 3 (y sin "sop_montaje")

Bloque 4: Iconografía

Se incorpora la variable "tema" para evaluar si la iconografía de la pintura aporta información adicional sobre la probabilidad de exito, una vez controlados los efectos de datacion (Bloque 1), morfología (Bloque 2) y material/tecnica (Bloque 3).

```
##
## Resumen del modelo m4 (+ tema):
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion +
##     soporte_grp + tecnica + sop_montaje + tema, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.4473096  0.6956496  -9.268  < 2e-16 ***
## fecha_est               0.0021170  0.0003842   5.509 3.60e-08 ***
## log_ancho               0.0161148  0.0234213   0.688  0.49143
## tam_catmediano          0.0611689  0.1114788   0.549  0.58321
## tam_catgrande           0.9698516  0.1070041   9.064  < 2e-16 ***
## orientacionhorizontal   0.1920851  0.0823649   2.332  0.01969 *
## soporte_grpMetal       -0.3647577  0.3754039  -0.972  0.33123
## soporte_grpMural       -0.1447200  0.5596334  -0.259  0.79595
## soporte_grpOtros       -1.0758535  0.3947047  -2.726  0.00642 **
## soporte_grpTabla/Panel  0.5114591  0.1257582   4.067 4.76e-05 ***
## tecnicamixta            0.1919487  0.3278557   0.585  0.55823
## tecnicaotras           -0.2149724  0.4000111  -0.537  0.59098
## sop_montajesi           0.5423510  0.2317816   2.340  0.01929 *
## temabodegon_floral      0.2161149  0.2055642   1.051  0.29311
## temacaza_animales       0.2186608  0.2932870   0.746  0.45594
## temahistoria_alegoria   0.4163296  0.1945265   2.140  0.03234 *
## temamitologia          -0.1780479  0.2441280  -0.729  0.46580
## temaotros               0.3379986  0.1154324   2.928  0.00341 **
## temapaisaje_lugares     0.5350909  0.1279326   4.183 2.88e-05 ***
## temaproceso_obra       -0.1095154  0.3453552  -0.317  0.75116
## temaretrato_corte       0.1259847  0.1287362   0.979  0.32776
## temavida_cotidiana     -0.0399076  0.1869047  -0.214  0.83092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4928.5  on 6980  degrees of freedom
```

```
## AIC: 4972.5
##
## Number of Fisher Scoring iterations: 5
##
##
## Aporte de información del Bloque 4:
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + tema
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6989     4955.2
## 2      6980     4928.5  9   26.763  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de complejidad:
##    df      AIC
## m3 13 4981.247
## m4 22 4972.484
##    df      BIC
## m3 13 5070.348
## m4 22 5123.271
```

El bloque iconográfico (tema) mejoró significativamente el ajuste (LRT $p = 0.001$) y redujo el AIC ($\triangle AIC \approx -9$), pero incrementó fuertemente el BIC ($\triangle BIC \approx +53$), reflejando un aumento importante de complejidad por el número de niveles de tema (se añaden 9 parámetros). Dado que el objetivo principal del estudio es caracterizar el éxito con un modelo parsimonioso y fácilmente interpretable, se decidió mantener como modelo principal el que excluye tema.

Sin embargo, dado el interés interpretativo de la iconografía, se mantiene el modelo con tema como análisis complementario específico para interpretar relaciones temáticas.

Por lo tanto, el modelo resultante después del Bloque 4 es el 'm3' y la opción alternativa contempla la inclusión de "tema"

Bloque 5: autoría y serie

Se incorporan las variables "tipo_autor" y "serie" para evaluar si la información de autoría y pertenencia a serie aporta información adicional sobre la probabilidad de éxito, una vez controlados los efectos de datación (Bloque 1), morfología (Bloque 2) y material/técnica (Bloque 3)

```
##
## Resumen del modelo m5 (+ tipo_autor + serie):
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion +
##     soporte_grp + tecnica + sop_montaje + tipo_autor + serie,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -7.7087803  0.6984622 -11.037  < 2e-16 ***
## fecha_est              0.0028765  0.0003735   7.701 1.35e-14 ***
```

```
## log_ancho               0.0197692  0.0247866   0.798  0.42512
## tam_catmediano          0.0439033  0.1105690   0.397  0.69132
## tam_catgrande           0.9209665  0.1045045   8.813  < 2e-16 ***
## orientacionhorizontal   0.3027081  0.0766217   3.951 7.79e-05 ***
## soporte_grpMetal       -0.3951589  0.3745363  -1.055  0.29140
## soporte_grpMural       -0.4827246  0.5611106  -0.860  0.38962
## soporte_grpOtros       -1.1192487  0.3988507  -2.806  0.00501 **
## soporte_grpTabla/Panel  0.5486309  0.1257558   4.363 1.28e-05 ***
## tecnicamixta            0.1537854  0.3294323   0.467  0.64063
## tecnicaotras           -0.2591605  0.3985112  -0.650  0.51548
## sop_montajesi           0.6867481  0.2288773   3.001  0.00270 **
## tipo_autoranonimo       0.1971008  0.1470093   1.341  0.18001
## tipo_autormujer         0.2748502  0.3898887   0.705  0.48084
## tipo_autorvarios        0.3445604  0.2444371   1.410  0.15866
## seriesi                 0.2714728  0.0939767   2.889  0.00387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4943.1  on 6985  degrees of freedom
## AIC: 4977.1
##
## Number of Fisher Scoring iterations: 5
##
##
## Aporte de información del Bloque 5:
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + tipo_autor + serie
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6989     4955.2
## 2      6985     4943.1  4   12.161   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de complejidad:
##    df      AIC
## m3 13 4981.247
## m5 17 4977.086
##    df      BIC
## m3 13 5070.348
## m5 17 5093.603
```

El Bloque 5 (autoría y serie) mejoró significativamente el modelo previo (LRT $p = 0.01$) y redujo el AIC ($\triangle AIC \approx -4$), aunque incrementó fuertemente el BIC ($\triangle BIC \approx +23$), reflejando un aumento de complejidad. En este caso podemos observar algo interesante, los coeficientes pertenecientes a "tipo_autor" no serultaron significativos, mientras que el de "serie=1" sí. Esto refleja que el efecto significativo dentro del bloque parece concentrarse en serie. Para confirmar este hecho, se evaluará la contribución independiente de cada variable, mediante modelos anidados parciales, antes de tomar una decisión formal sobre el modelo

principal.

```
##
## Aporte de información de 'serie':
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6989     4955.2
## 2      6988     4946.9  1   8.3182 0.003925 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Aporte de información de 'tipo_autor':
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + tipo_autor
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6989     4955.2
## 2      6986     4951.2  3   4.0594   0.2551
##
## Aporte de complejidad:
##          df      AIC
## m3       13 4981.247
## m5_serie 14 4974.929
## m5_autor 16 4983.188
## m5       17 4977.086
##          df      BIC
## m3       13 5070.348
## m5_serie 14 5070.884
## m5_autor 16 5092.851
## m5       17 5093.603
```

Efectivamente la variable "serie" sí demostró aportar información adicional (LRT $p = 0.003$) mejorando también el AIC ($\triangle AIC \approx -6$) y manteniendo aproximadamente estable BIC, respecto al anterior modelo aceptado 'm3' y también reduciendo ambos criterios frente al modelo completo 'm5' ($\triangle AIC \approx -2$, $\triangle BIC \approx -23$). Por otro lado, "tipo_autor" no demostró aportar información adicional (LRT $p = 0.26$) además de ser el que aporta los valores más de AIC, superando incluso el valor del modelo completo 'm5'. Se decide prescindir de la variable "tipo_autor" y conservar el modelo únicamente con la inclusión de "serie".

Modelo resultante después de añadir Bloque 5:

Resumen

Después de evaluar los efectos principales, el modelo principal es el siguiente:

```
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion +
```

```
##     soporte_grp + tecnica + sop_montaje + serie, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -7.5545967  0.6911174 -10.931  < 2e-16 ***
## fecha_est               0.0027956  0.0003694   7.567 3.81e-14 ***
## log_ancho               0.0290842  0.0232929   1.249  0.21180
## tam_catmediano          0.0473151  0.1104536   0.428  0.66838
## tam_catgrande           0.9178941  0.1042498   8.805  < 2e-16 ***
## orientacionhorizontal   0.3041996  0.0764413   3.980 6.91e-05 ***
## soporte_grpMetal       -0.4025129  0.3743727  -1.075  0.28230
## soporte_grpMural       -0.4518929  0.5600464  -0.807  0.41973
## soporte_grpOtros       -1.0662903  0.3942547  -2.705  0.00684 **
## soporte_grpTabla/Panel  0.5493312  0.1251844   4.388 1.14e-05 ***
## tecnicamixta            0.1657706  0.3292548   0.503  0.61463
## tecnicaotras           -0.2221573  0.3943927  -0.563  0.57324
## sop_montajesi           0.6860477  0.2283316   3.005  0.00266 **
## seriesi                 0.2744989  0.0937574   2.928  0.00341 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4946.9  on 6988  degrees of freedom
## AIC: 4974.9
##
## Number of Fisher Scoring iterations: 5
```

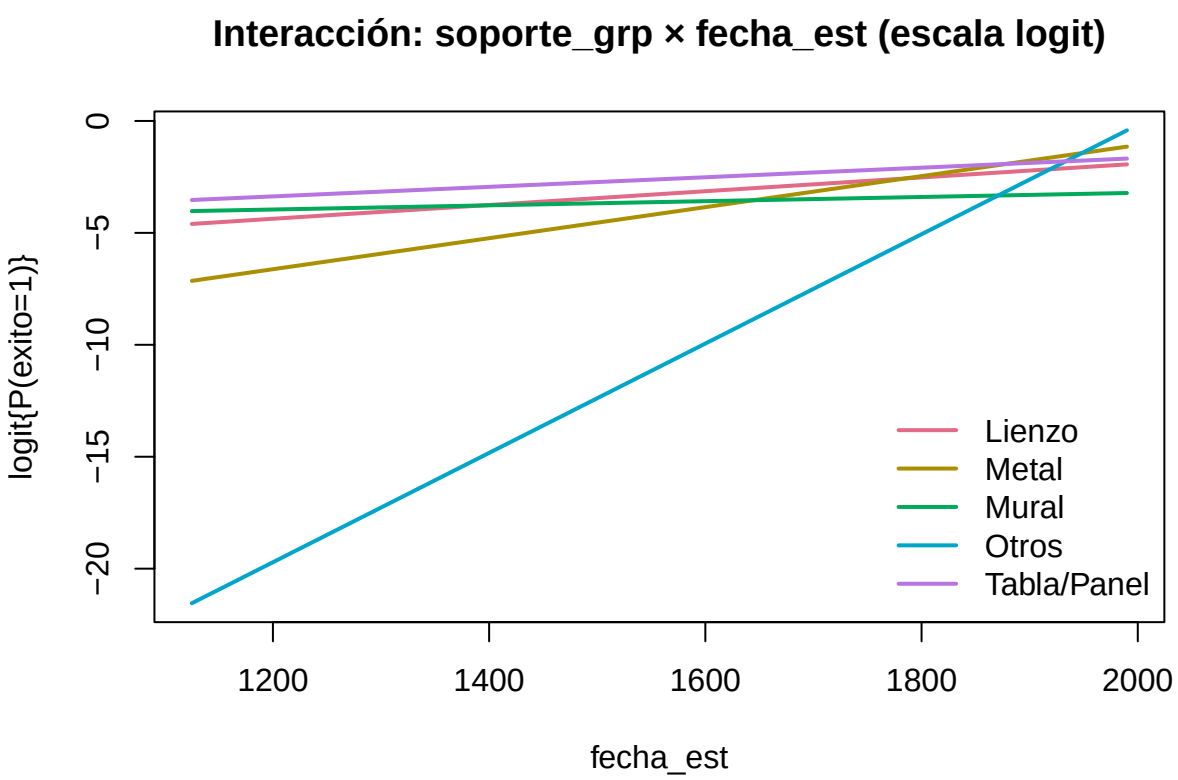
5.2 Interacciones

Para decidir qué variables son candidatas de entrar al modelo primero traeremos las conclusiones sobre estas, obtenidas en el análisis descriptivo. En dicha sección de determinó que las más plausibles eran "soporte_grp x fecha_est", "tecnica x fecha_est", "soporte_grp x tam_cat", "tecnica x tam_cat" y "soporte_grp x orientacion". Graficaremos los correspondientes gráficos de interacción para estas opciones y seleccionaremos las mejores, que seguidamente serán comprobadas con pruebas formales.

Gráficos de interacción

Graficamos $P(exito = 1)$ predicha por el modelo, fijando el resto de covariables en valores de referencia/mediana.

1) soporte_grp x fecha_est



Esta interacción plantea algunas preguntas, ya que aunque podemos observar algunos indicios vemos que existe una categoría, Otros, que presenta una fuerte diferencia. Esto nos hace pensar en que puede tratarse de una escasez de datos en el extremo.

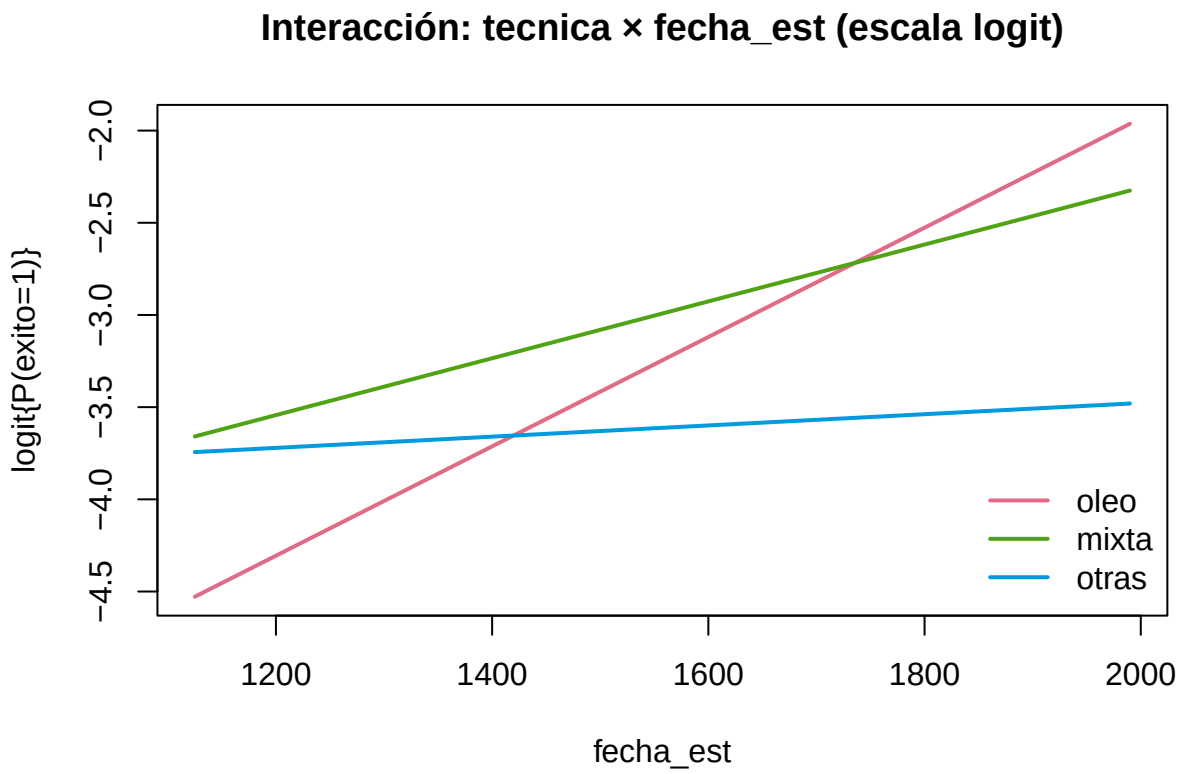
Comprobamos frecuencias y recuento de éxitos por combinación:

```
## $Lienzo
## [1] 1286 1990
##
## $Metal
## [1] 1356 1880
##
## $Mural
## [1] 1125 1822
##
## $Otros
## [1] 1515 1952
##
## $`Tabla/Panel`
## [1] 1205 1957
##
##
## Frecuencias por combinación:
##              grp_fecha
## soporte_grp   [1.1e+03,1.4e+03] (1.4e+03,1.7e+03] (1.7e+03,2e+03]
##   Lienzo                      9              2533            2974
##   Metal                       1               129              24
```

```
##   Mural                      25                10              14
##   Otros                       0                35             102
##   Tabla/Panel                20               884             242
##
## Recuento de éxitos por combinación:
##              grp_fecha
## soporte_grp   [1.1e+03,1.4e+03] (1.4e+03,1.7e+03] (1.7e+03,2e+03]
##   Lienzo                     0               266             433
##   Metal                      0                 5               3
##   Mural                      2                 2               2
##   Otros                      0                 0               9
##   Tabla/Panel                2                91              34
```

Estos resultados explican la fomra rara del gráfico "soporte_grp x fecha_est". Vemos como en el tramo [1100, 1400] todo son frecuencias bajas y celdas vacías. Concretamente vemos que la linea diferenciada de Otros estaba provocada por la inexistencia de esta categoría durante el primer periodo de años. Decidimos descartar esta interacción por inconsistencia de resultados a consecuencia de regiones sin datos.

2) tenica x fecha_est



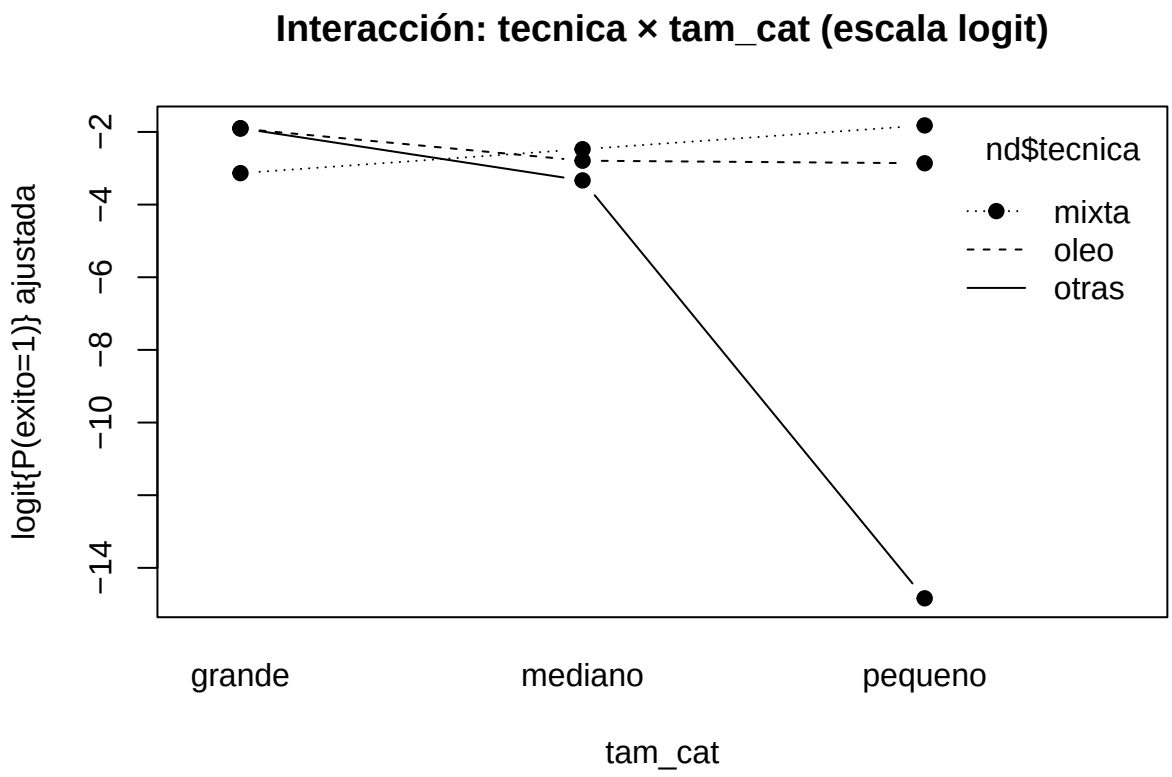
Vemos indicios claros de interacción y un claro cruce: ólea pasa de estar por debajo a por encima. Sin embargo, estas observaciones podrían estar de nuevo sesgadas por falta de datos en algunos periodos. comporbamos recuentos:

```
##
```

```
## Frecuencias por combinación:
##        grp_fecha
## tecnica [1.1e+03,1.4e+03] (1.4e+03,1.7e+03] (1.7e+03,2e+03]
##   oleo                 13              3440            3319
##   mixta                 5                98              19
##   otras                37                53              18
##
## Recuento de éxitos por combinación:
##        grp_fecha
## tecnica [1.1e+03,1.4e+03] (1.4e+03,1.7e+03] (1.7e+03,2e+03]
##   oleo                  0               349             477
##   mixta                 0                10               4
##   otras                 4                 5               0
```

Efectivamente podemos ver como la variable "tecnica" está fuertemente sesgada debido principalmente al nivel 'oleo', que experimentó un increible aumento después del primer tramo. Esto explica su trayectoria ascendente, que no sería a causa de una interacción real sinó a por el desbalance de frecuencias. Los otros dos niveles no experimentaron ningún cruce en el gráfico, por lo que no generan interés teniendo en cuenta estos resultados. Aunque esta variable forma parte de nuestras hipótesis principales, su inclusión ya se aceptó de manera separada asumiendo un aumento de complejidad elevado (incremento significativo en BIC en Bloque 2 de efectos principales), por lo que su interés conceptual e interpretativo ya fue considerado. Por esta razón preferimos descartar su interacción con el objetivo de no viciar nuestro modelo.

3) tecnica x tam_cat

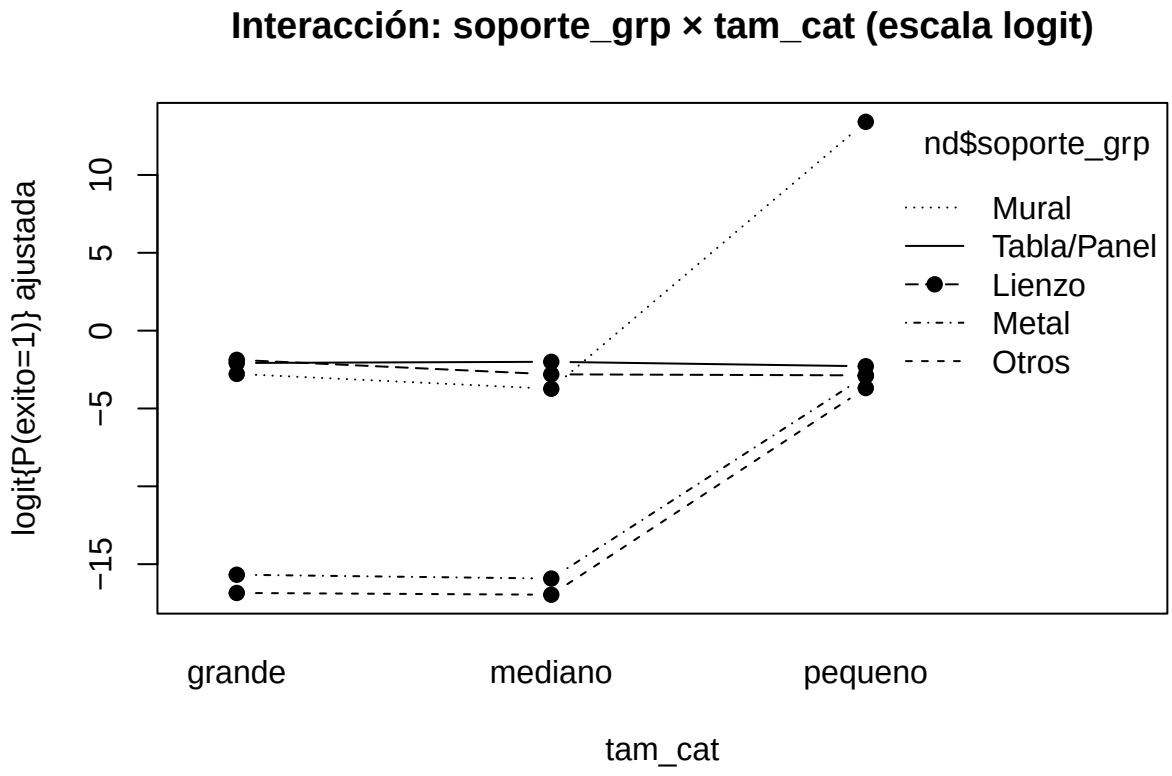


De nuevo vemos puntos extremos de manera que comprobaremos recuentos:

```
##
## Frecuencias por combinación:
##        tam_cat
## tecnica pequeno mediano grande
##   oleo     2259    2216   2297
##   mixta      32      64     26
##   otras      15      37     56
##
## Recuento de éxitos por combinación:
##        tam_cat
## tecnica pequeno mediano grande
##   oleo      217     189    420
##   mixta       6       6      2
##   otras       0       2      7
```

Obtenemos los mismo resultados en esta interacción con "tecnica": nivel 'oloe' fuertemente predominante. Se descarta esta interacción.

4) soporte_grp x tam_cat



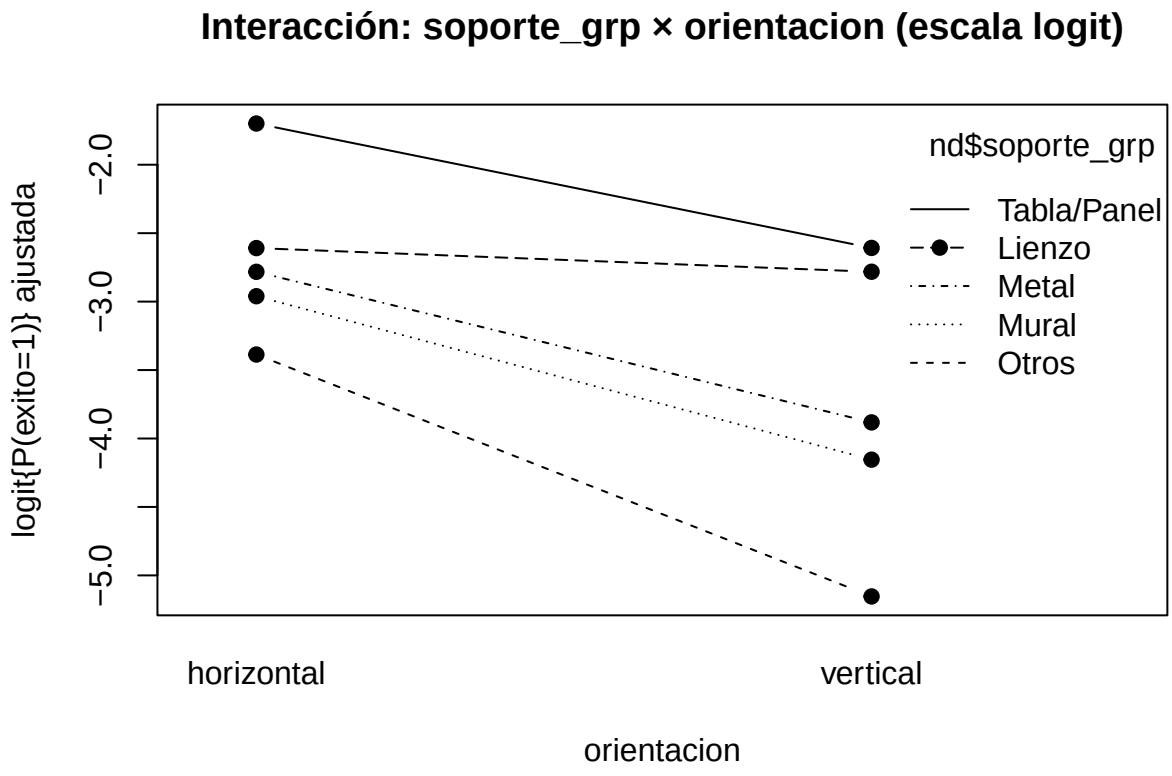
A primera vista podemos ver algunos incidios pero no determinantes de modificación del efecto. Podrían explicarse por la baja frecuencia de algunas combinaciones:

```
##
## Frecuencias por combinación:
```

```
##              tam_cat
## soporte_grp   pequeno mediano grande
##   Lienzo         1402    1884   2230
##   Metal           120      31      3
##   Mural             1      12     36
##   Otros           107      28      2
##   Tabla/Panel     676     362    108
##
## Recuento de éxitos por combinación:
##              tam_cat
## soporte_grp   pequeno mediano grande
##   Lienzo          129     156    414
##   Metal             8       0      0
##   Mural             1       1      4
##   Otros             9       0      0
##   Tabla/Panel      76      40     11
```

Efectivamente vemos que categorías como Metal-grande o Mural-pequeño presentan frecuencias realmente bajas, además los exitos se concentran alrededor de las categoría 'Lienzo' y 'Table/Panel' lo cual es lógico ya que són las categorías mayoritarias. Sin embargo se decide aceptar esta interacción como candidata para las posteriores pruebas formales, ya que es una hipótesis central del estudio.

5) soporte_grp x orientacion



Esta interacción ha resultado la menos relevante, pero aun con incidios de posible interacción. Comprobaremos también los recuentos:

```
##
## Frecuencias por combinación:
##              orientacion
## soporte_grp   vertical horizontal
##   Lienzo          3159       2357
##   Metal             49        105
##   Mural             20         29
##   Otros             53         84
##   Tabla/Panel      720        426
##
## Recuento de éxitos por combinación:
##              orientacion
## soporte_grp   vertical horizontal
##   Lienzo           350        349
##   Metal              1          7
##   Mural              1          5
##   Otros              1          8
##   Tabla/Panel       54         73
```

Vemos el mismo patrón para la variable respuesta: los éxitos se concentran al rededor de 'Lienzo', sin embargo no vemos fuertes desbalances para los grupos de orientación. La mantenemos como posble candidata a pruebas

Pruebas formales

Decidimos testear las siguientes interacciones: "soporte_grp x tam_cat" y "soporte_grp x orientacion". Ambas opciones parecen plausibles tanto por su representación gráfica como por interpretación conceptual, además concretamente "soporte x tam_cat" se incluía en nuestras hipótesis, de manera que consideramos muy apropiada esta selección.

Comprobaremos en primer caso la inclusión de cada interacción de manera separada para estudiar si cada una por separado aporta información al modelo y su compromiso ajuste-complejidad.

```
##
## ==============================
##  soporte_grp x tam_cat
## ==============================
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie + tam_cat:soporte_grp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6988     4946.9
## 2      6980     4925.6  8   21.367 0.006233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                  df      AIC
## m_final          14 4974.929
## m_soporte_tamcat 22 4969.561
##                  df      BIC
## m_final          14 5070.884
## m_soporte_tamcat 22 5120.348
##
```

```
## ==============================
##  soporte_grp x orientacion
## ==============================
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie + orientacion:soporte_grp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6988     4946.9
## 2      6984     4931.6  4   15.335 0.004054 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                        df      AIC
## m_final                14 4974.929
## m_soporte_orientacion 18 4967.594
##                        df      BIC
## m_final                14 5070.884
## m_soporte_orientacion 18 5090.965
```

Vemos que ambas interacciones demuestran mejorar significativamente el modelo aportando informacion (LRT(99) $p < 0.01$) i reduciendo el AIC ($\triangle AIC \approx -7; -10$). Aunque el BIC aumento en los dos casos ($\triangle BIC \approx +47; +17$). Por interés interpretativo decidimos mantener el modelo con la interacción "soporte_grp x tam_cat" como base y procedemos a examinar el modelo completo anidado con la otra interacción.

```
## Analysis of Deviance Table
##
## Model 1: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie
## Model 2: exito ~ fecha_est + log_ancho + tam_cat + orientacion + soporte_grp +
##     tecnica + sop_montaje + serie + tam_cat:soporte_grp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6988     4946.9
## 2      6980     4925.6  8   21.367 0.006233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                    df      AIC
## m_final            14 4974.929
## m_soporte_tamcat 22 4969.561
##                    df      BIC
## m_final            14 5070.884
## m_soporte_tamcat 22 5120.348
```

Vemos que la inclusión de la interacción "soporte_grp x orientación" sigue siendo significativa una vez controlado el efecto de "soporte_grp x tam_cat", sin embargo provoca un brave problema de complejidad ($\triangle BIC \approx +74$) que no consideramos aceptable ni necesario en este punto análisis. En consecuencia descartamos la interaccion con orientación una vez controlado por tamaño. Además ya detectamos anteriormente existencia de celdas problemáticas en soporte, que pueden provocar separación, por lo que algunos coeficientes pueden volverse fuertemente inestables. Por esta razón, debemos mencionar que la interacción con tamaño se mantendrá pero con interpretación principalmente en 'Lienzo' y 'Tabla/Panel', además de tratará de minimizar esta problematica en la siguiente sección.

Finalmente se decide definir el modelo principal con únicamente la interacción “soporte_grp x tam_cat” con el obejtivo de dar respuesta a nuestra hipótesis. Sin embargo, se considera conservar el modelo sin interacciones como alternativo con el fin de explorar más rigurosamente los efectos principales, si se considera oportuno.

Resumen

Después de evaluar las interacciones, el modelo principal es el siguiente:

Y el modelo reducido conservado como alternativo, sin interacciones, es el siguiente:

5.3 Diagnóstico de ajuste y correcciones

En esta sección se presentan diagnósticos preliminares del modelo, centrados en la calidad del ajuste y en la estabilidad de los parámetros, especialmente considerando la baja prevalencia del evento y el uso de múltiples factores e interacciones. No se llevaran a cabo aún procedimientos de validación formales, ya que se abordarán en secciones posteriores.

```
##
## Call:
## glm(formula = exito ~ fecha_est + log_ancho + tam_cat + orientacion +
##     soporte_grp + tecnica + sop_montaje + serie + soporte_grp:tam_cat,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -7.549e+00  6.939e-01 -10.880  < 2e-16
## fecha_est                            2.767e-03  3.733e-04   7.411 1.25e-13
## log_ancho                            2.947e-02  2.332e-02   1.264 0.206323
## tam_catmediano                       6.620e-02  1.303e-01   0.508 0.611445
## tam_catgrande                        9.977e-01  1.158e-01   8.615  < 2e-16
## orientacionhorizontal                2.869e-01  7.685e-02   3.734 0.000189
## soporte_grpMetal                    -3.127e-02  3.834e-01  -0.082 0.934991
## soporte_grpMural                     1.629e+01  1.455e+03   0.011 0.991071
## soporte_grpOtros                    -8.162e-01  3.957e-01  -2.063 0.039139
## soporte_grpTabla/Panel               5.858e-01  1.620e-01   3.615 0.000300
## tecnicamixta                         3.942e-01  3.423e-01   1.152 0.249498
## tecnicaotras                        -1.102e-01  3.933e-01  -0.280 0.779314
## sop_montajesi                        7.855e-01  2.335e-01   3.364 0.000768
## seriesi                              2.799e-01  9.416e-02   2.973 0.002954
## tam_catmediano:soporte_grpMetal     -1.309e+01  2.604e+02  -0.050 0.959920
## tam_catgrande:soporte_grpMetal      -1.377e+01  8.390e+02  -0.016 0.986902
## tam_catmediano:soporte_grpMural     -1.721e+01  1.455e+03  -0.012 0.990563
## tam_catgrande:soporte_grpMural      -1.719e+01  1.455e+03  -0.012 0.990575
## tam_catmediano:soporte_grpOtros     -1.334e+01  2.726e+02  -0.049 0.960978
## tam_catgrande:soporte_grpOtros      -1.416e+01  1.023e+03  -0.014 0.988954
## tam_catmediano:soporte_grpTabla/Panel  2.191e-01  2.513e-01   0.872 0.383262
## tam_catgrande:soporte_grpTabla/Panel  -7.876e-01  3.773e-01  -2.088 0.036839
##
## (Intercept)                         ***
## fecha_est                           ***
## log_ancho
## tam_catmediano
## tam_catgrande                       ***
## orientacionhorizontal               ***
```

```
## soporte_grpMetal
## soporte_grpMural
## soporte_grpOtros                      *
## soporte_grpTabla/Panel                ***
## tecnicamixta
## tecnicaotras
## sop_montajesi                         ***
## seriesi                               **
## tam_catmediano:soporte_grpMetal
## tam_catgrande:soporte_grpMetal
## tam_catmediano:soporte_grpMural
## tam_catgrande:soporte_grpMural
## tam_catmediano:soporte_grpOtros
## tam_catgrande:soporte_grpOtros
## tam_catmediano:soporte_grpTabla/Panel
## tam_catgrande:soporte_grpTabla/Panel  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4925.6  on 6980  degrees of freedom
## AIC: 4969.6
##
## Number of Fisher Scoring iterations: 14
```

Para comenzar vemos valores grandes para los errores de todos los coeficientes y incluso algunos enormes del orden de $10^3 $, lo cual se traduce en una inestabilidad preocupante que resta credibilidad a nuestras conclusiones futuras.

Examinaremos a continuación la capacidad del modelo para estimar los 22 coeficientes presentes en él.

```
## N = 7002   Eventos (1) = 849   Prevalencia = 0.1213
## Num coeficientes (incluye dummies) = 22
## EPV aprox (eventos por coef) = 38.591
```

Con 7002 observaciones y 849 eventos, el modelo dispone de información suficiente para estimar los 22 parámetros que contiene. El coeficiente de ($EPV \approx 38.6$) sugiere que el desnivel en la respuesta no plantea una limitación para el modelo seleccionado.

```
##
## Frecuencias de soporte
##
##      Lienzo       Metal       Mural       Otros Tabla/Panel
##        5516         154          49         137        1146
##
## Frecuencias de tamaño
##
## pequeno mediano  grande
##    2306    2317    2379
##
## Éxitos Y:
##              tam_cat
```

```
## soporte_grp   pequeno mediano grande
##   Lienzo          129     156    414
##   Metal             8       0      0
##   Mural             1       1      4
##   Otros             9       0      0
##   Tabla/Panel      76      40     11
##
## Totales N:
##              tam_cat
## soporte_grp   pequeno mediano grande
##   Lienzo         1402    1884   2230
##   Metal           120      31      3
##   Mural             1      12     36
##   Otros           107      28      2
##   Tabla/Panel     676     362    108
##
## Proporción P=Y/N:
##              tam_cat
## soporte_grp   pequeno mediano grande
##   Lienzo        0.092   0.083  0.186
##   Metal         0.067   0.000  0.000
##   Mural         1.000   0.083  0.111
##   Otros         0.084   0.000  0.000
##   Tabla/Panel   0.112   0.110  0.102
##
## Celdas con 0 éxitos (Y==0):
##              tam_cat
## soporte_grp   pequeno mediano grande
##   Lienzo        FALSE   FALSE  FALSE
##   Metal         FALSE    TRUE   TRUE
##   Mural         FALSE   FALSE  FALSE
##   Otros         FALSE    TRUE   TRUE
##   Tabla/Panel   FALSE   FALSE  FALSE
##
## Celdas con todos éxitos (Y==N):
##              tam_cat
## soporte_grp   pequeno mediano grande
##   Lienzo        FALSE   FALSE  FALSE
##   Metal         FALSE   FALSE  FALSE
##   Mural          TRUE   FALSE  FALSE
##   Otros         FALSE   FALSE  FALSE
##   Tabla/Panel   FALSE   FALSE  FALSE
```

Las tablas de contingencia de exitos y totales por combinación de la interacción muestran celdas con respuestas deterministas lo que puede provocar separación, que explica la inestabilidad de estimación observada. En concreto observamos todo éxitos en Mural-pequeño, lo cual tiene sentido conceptualmete ya que los murales estan asociados a grandes obras de arte; y observamso ausencia total de éxitos en Metal-mediano,grande y Otros-mediano,grande. Estos resultados eran esperables debido a la baja frecuencia de las categorías involucradas Metal/Mural/Otros. Recordamos entonces la importancia de centrar la interpretacion al rededor de las categoría estables Tabl/Panel y Lienzo, aun sabiendo que no estan ajenas a la problemática.

```
##                           GVIF Df GVIF^(1/(2*Df))
## fecha_est          1.662901e+00  1        1.289535
## log_ancho          1.283442e+00  1        1.132891
```

```
## tam_cat             2.123473e+00  2        1.207151
## orientacion         1.061100e+00  1        1.030097
## soporte_grp         2.718990e+07  4        8.497691
## tecnica             1.465668e+00  2        1.100294
## sop_montaje         1.614959e+00  1        1.270810
## serie               1.103178e+00  1        1.050323
## tam_cat:soporte_grp 2.256196e+07  8        2.881284
```

Los indicadores de colinealidad muestran valores muy altos para soporte_grp ($GVIF \approx 8.50$) y elevados para tam_cat:soporte_grp ($GVIF \approx 2.88$), en contraste con el resto de covariables, cuyos valores permanecen cercanos a 1. Esta evidencia afirma una dependencia fuerte entre los bloques de parámetros de los efectos principales y los de la interacción, especialmente en presencia de un diseño como el nuestro, desbalanceado entre niveles y con celdas con baja frecuencia.

```
## Devianza ajustada: 0.7056678
```

La evaluación de la dispersión mediante el parámetro de escala no aporta indicios de sobredispersión ($\varphi \approx 0,706$), ya que es a 1 (la dispersión teórica de la distribución Binomial). Por lo que estos resultados sugieren que no existe un exceso de variabilidad no aceptada por la distribución que justifique adoptar estrategias orientadas a corregir sobredispersión.

Después de este primer análisis vemos que existen dos problematicas que debemos manejar, la separación por celdas vacías y la colinealidad excesiva en soporte.

Primeramente abordaremos la problemática provocada por las celdas deterministas de la interacción y más adelante, si la colinealidad siguiera presente, se adoptarán nuevas medidas.

Separación

Contrastaremos formalmente la existencia de separación en nuestro modelo:

```
## Implementation: ROI | Solver: lpsolve
## Separation: TRUE
## Existence of maximum likelihood estimates
##                             (Intercept)                               fecha_est
##                                       0                                       0
##                               log_ancho                            tam_catmediano
##                                       0                                       0
##                           tam_catgrande                      orientacionhorizontal
##                                       0                                       0
##                         soporte_grpMetal                         soporte_grpMural
##                                       0                                     Inf
##                         soporte_grpOtros                    soporte_grpTabla/Panel
##                                       0                                       0
##                            tecnicamixta                              tecnicaotras
##                                       0                                       0
##                           sop_montajesi                                  seriesi
##                                       0                                       0
##       tam_catmediano:soporte_grpMetal       tam_catgrande:soporte_grpMetal
##                                    -Inf                                    -Inf
##       tam_catmediano:soporte_grpMural       tam_catgrande:soporte_grpMural
##                                    -Inf                                    -Inf
##       tam_catmediano:soporte_grpOtros       tam_catgrande:soporte_grpOtros
##                                    -Inf                                    -Inf
## tam_catmediano:soporte_grpTabla/Panel  tam_catgrande:soporte_grpTabla/Panel
```

```
##                                     0                                     0
## 0: finite value, Inf: infinity, -Inf: -infinity
```

Efectivamente nos enfrentamos a un problema de separación real, lo cual puede traducerse en inexistencia de coeficientes finitos. En consecuencia, decidimos reestimar el mismo modelo mediante regresión logística con reducción de sesgo con un enfoque Firth.

```
##
## Call:
## glm(formula = formula(m_completo), family = binomial("logit"),
##     data = df, method = "brglmFit")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1877  -0.5361  -0.4259  -0.3022   8.4904
##
## Coefficients:
##                                        Estimate Std. Error    z value Pr(>|z|)
## (Intercept)                           -8.346e+00  7.506e-01 -1.112e+01  < 2e-16
## fecha_est                              3.188e-03  4.030e-04  7.909e+00 2.59e-15
## log_ancho                              4.195e-02  2.470e-02  1.698e+00 0.089478
## tam_catmediano                         8.759e-02  1.311e-01  6.680e-01 0.503954
## tam_catgrande                          1.025e+00  1.170e-01  8.762e+00  < 2e-16
## orientacionhorizontal                  3.051e-01  7.978e-02  3.825e+00 0.000131
## soporte_grpMetal                      -1.464e+15  6.126e+06 -2.389e+08  < 2e-16
## soporte_grpMural                       3.251e+15  6.711e+07  4.844e+07  < 2e-16
## soporte_grpOtros                      -8.363e-01  3.934e-01 -2.126e+00 0.033532
## soporte_grpTabla/Panel                 6.109e-01  1.643e-01  3.719e+00 0.000200
## tecnicamixta                           1.167e+00  4.718e-01  2.474e+00 0.013375
## tecnicaotras                          -1.255e+00  9.861e-01 -1.273e+00 0.203130
## sop_montajesi                          8.506e-01  2.454e-01  3.467e+00 0.000527
## seriesi                                3.333e-01  9.881e-02  3.373e+00 0.000744
## tam_catmediano:soporte_grpMetal       -1.172e+15  1.352e+07 -8.672e+07  < 2e-16
## tam_catgrande:soporte_grpMetal        -3.365e+15  3.923e+07 -8.579e+07  < 2e-16
## tam_catmediano:soporte_grpMural       -6.346e+15  6.985e+07 -9.086e+07  < 2e-16
## tam_catgrande:soporte_grpMural        -3.414e+15  6.803e+07 -5.018e+07  < 2e-16
## tam_catmediano:soporte_grpOtros       -1.902e+16  1.268e+07 -1.500e+09  < 2e-16
## tam_catgrande:soporte_grpOtros        -1.689e+15  4.745e+07 -3.559e+07  < 2e-16
## tam_catmediano:soporte_grpTabla/Panel -1.994e+15  3.527e+06 -5.654e+08  < 2e-16
## tam_catgrande:soporte_grpTabla/Panel  -1.393e+15  6.458e+06 -2.157e+08  < 2e-16
##
## (Intercept)                           ***
## fecha_est                             ***
## log_ancho                             .
## tam_catmediano
## tam_catgrande                         ***
## orientacionhorizontal                 ***
## soporte_grpMetal                      ***
## soporte_grpMural                      ***
## soporte_grpOtros                      *
## soporte_grpTabla/Panel                ***
## tecnicamixta                          *
## tecnicaotras
## sop_montajesi                         ***
```

```
## seriesi                                ***
## tam_catmediano:soporte_grpMetal        ***
## tam_catgrande:soporte_grpMetal         ***
## tam_catmediano:soporte_grpMural        ***
## tam_catgrande:soporte_grpMural         ***
## tam_catmediano:soporte_grpOtros        ***
## tam_catgrande:soporte_grpOtros         ***
## tam_catmediano:soporte_grpTabla/Panel ***
## tam_catgrande:soporte_grpTabla/Panel  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 9122.7  on 6980  degrees of freedom
## AIC:  9166.7
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 100
```

En un primer intento nuestro modelo no alcanzó convergencia, lo que se traduce en estimaciones inestables, de manera que se procede a ajustar parámetros de control de las iteraciones:

```
## Convergencia: TRUE
## Número de iteraciones: 24
## Call:
## glm(formula = formula(m_completo), family = binomial("logit"),
##     data = df, control = ctrl, method = "brglmFit", type = "AS_mean")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9587  -0.5496  -0.4409  -0.3547   2.4983
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -7.5188015  0.6902160 -10.893  < 2e-16
## fecha_est                         0.0027539  0.0003714   7.415 1.21e-13
## log_ancho                         0.0294263  0.0232197   1.267 0.205048
## tam_catmediano                    0.0643331  0.1299291   0.495 0.620501
## tam_catgrande                     0.9929847  0.1154692   8.600  < 2e-16
## orientacionhorizontal             0.2860374  0.0765986   3.734 0.000188
## soporte_grpMetal                  0.0189324  0.3747131   0.051 0.959704
## soporte_grpMural                  1.7860819  2.3433219   0.762 0.445940
## soporte_grpOtros                 -0.7732117  0.3885090  -1.990 0.046569
## soporte_grpTabla/Panel            0.5866802  0.1614587   3.634 0.000279
## tecnicamixta                      0.4170149  0.3341610   1.248 0.212051
## tecnicaotras                     -0.0504376  0.3820347  -0.132 0.894966
## sop_montajesi                     0.7901158  0.2314649   3.414 0.000641
## seriesi                           0.2797309  0.0937933   2.982 0.002860
## tam_catmediano:soporte_grpMetal  -1.7264476  1.4984077  -1.152 0.249244
## tam_catgrande:soporte_grpMetal   -0.2107848  1.7900398  -0.118 0.906262
## tam_catmediano:soporte_grpMural  -2.3924557  2.4965511  -0.958 0.337909
## tam_catgrande:soporte_grpMural   -2.6305508  2.3924610  -1.100 0.271543
```

```
## tam_catmediano:soporte_grpOtros       -1.9023583  1.5254641  -1.247 0.212372
## tam_catgrande:soporte_grpOtros        -0.2635839  1.9476521  -0.135 0.892348
## tam_catmediano:soporte_grpTabla/Panel  0.2242579  0.2499618   0.897 0.369629
## tam_catgrande:soporte_grpTabla/Panel  -0.7513967  0.3711931  -2.024 0.042942
##
## (Intercept)                           ***
## fecha_est                             ***
## log_ancho
## tam_catmediano
## tam_catgrande                         ***
## orientacionhorizontal                 ***
## soporte_grpMetal
## soporte_grpMural
## soporte_grpOtros                      *
## soporte_grpTabla/Panel                ***
## tecnicamixta
## tecnicaotras
## sop_montajesi                         ***
## seriesi                               **
## tam_catmediano:soporte_grpMetal
## tam_catgrande:soporte_grpMetal
## tam_catmediano:soporte_grpMural
## tam_catgrande:soporte_grpMural
## tam_catmediano:soporte_grpOtros
## tam_catgrande:soporte_grpOtros
## tam_catmediano:soporte_grpTabla/Panel
## tam_catgrande:soporte_grpTabla/Panel  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4929.9  on 6980  degrees of freedom
## AIC:  4973.9
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 24
```

Gracias a las especificaciones de control, el modelo ahora sí ha alcanzado convergencia. Todavía observamos algunos SE grandes, lo que indica estimaciones débiles para esos niveles, pero una mejora considerable frente al modelo anterior sin este ajuste Firth.

Veamos una compración de SE, y otro test de separación, para el modelo sin ajuste 'm_completo' y el modelo ajustado 'm_firth2':

```
##
## Top 10 SE (MLE):
## tam_catmediano:soporte_grpMural  tam_catgrande:soporte_grpMural
##                    1455.3979311                     1455.3976728
##                soporte_grpMural  tam_catgrande:soporte_grpOtros
##                    1455.3975872                     1022.7042152
##  tam_catgrande:soporte_grpMetal tam_catmediano:soporte_grpOtros
##                     838.9572783                      272.6487034
```

```
## tam_catmediano:soporte_grpMetal                     (Intercept)
##                      260.4130626                       0.6938752
##                soporte_grpOtros                     tecnicaotras
##                       0.3956836                       0.3932892
##
## Top 10 SE (Firth/AS):
## tam_catmediano:soporte_grpMural  tam_catgrande:soporte_grpMural
##                       2.4965511                       2.3924610
##                soporte_grpMural  tam_catgrande:soporte_grpOtros
##                       2.3433219                       1.9476521
##  tam_catgrande:soporte_grpMetal tam_catmediano:soporte_grpOtros
##                       1.7900398                       1.5254641
## tam_catmediano:soporte_grpMetal                     (Intercept)
##                       1.4984077                       0.6902160
##                soporte_grpOtros                     tecnicaotras
##                       0.3885090                       0.3820347
## Separación:
## Implementation: ROI | Solver: lpsolve
## Separation: TRUE
## Existence of maximum likelihood estimates
##                           (Intercept)                            fecha_est
##                                     0                                    0
##                             log_ancho                         tam_catmediano
##                                     0                                    0
##                          tam_catgrande                  orientacionhorizontal
##                                     0                                    0
##                       soporte_grpMetal                       soporte_grpMural
##                                     0                                  Inf
##                       soporte_grpOtros                 soporte_grpTabla/Panel
##                                     0                                    0
##                          tecnicamixta                            tecnicaotras
##                                     0                                    0
##                         sop_montajesi                                seriesi
##                                     0                                    0
##       tam_catmediano:soporte_grpMetal        tam_catgrande:soporte_grpMetal
##                                  -Inf                                 -Inf
##       tam_catmediano:soporte_grpMural        tam_catgrande:soporte_grpMural
##                                  -Inf                                 -Inf
##       tam_catmediano:soporte_grpOtros        tam_catgrande:soporte_grpOtros
##                                  -Inf                                 -Inf
## tam_catmediano:soporte_grpTabla/Panel  tam_catgrande:soporte_grpTabla/Panel
##                                     0                                    0
## 0: finite value, Inf: infinity, -Inf: -infinity
```

Podemos ver como la presencia de separación no se ha eliminado, cual era esperable ya que la reducción de sesgo Firth no cambia la estructura de separación de los dato, pero sí asegura que nuestro modelo es robusto frente a ella. Podemos comprobarlo viendo la significativa reducción de los SE. Sin embargo debemos tener presente que la mayoría de las combinaciones e la interacción no obtienen estimaciones MLE finitas, por lo que de nuevo solo nos centraremos en aquellas finitas y con mucha cautela ya que somos conscientes de que el modelo está afectado.

Colinealidad

Una vez controlada la separación volvamos a realizar las pruebas de colinealidad:

```
##                          GVIF Df GVIF^(1/(2*Df))
## fecha_est            1.671291  1        1.292784
## log_ancho            1.284864  1        1.133518
## tam_cat              2.146515  2        1.210413
## orientacion          1.063647  1        1.031332
## soporte_grp         99.891232  4        1.778038
## tecnica              1.575092  2        1.120280
## sop_montaje          1.675569  1        1.294438
## serie                1.108229  1        1.052725
## tam_cat:soporte_grp 86.923917  8        1.321893
```

Una vez manejada la separación, se elimina la existencia de colinealidad grave y los términos afectados presentan ahora colinealidad moderada que no refleja una preocupación real. Dado uestro objetivo descriptivo, se mantiene el modelo con todos sus términos.

Resumen

Finalmente se concluye con el modelo final de tipo regresión logística incluyendo reducción de sesgo Firth.

```
##
## Call:
## glm(formula = formula(exito ~ fecha_est + log_ancho + tam_cat +
##     orientacion + soporte_grp + tecnica + sop_montaje + serie +
##     soporte_grp:tam_cat), family = binomial("logit"), data = df,
##     control = ctrl, method = "brglmFit", type = "AS_mean")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9587  -0.5496  -0.4409  -0.3547   2.4983
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -7.5188015  0.6902160 -10.893  < 2e-16
## fecha_est                            0.0027539  0.0003714   7.415 1.21e-13
## log_ancho                            0.0294263  0.0232197   1.267 0.205048
## tam_catmediano                       0.0643331  0.1299291   0.495 0.620501
## tam_catgrande                        0.9929847  0.1154692   8.600  < 2e-16
## orientacionhorizontal                0.2860374  0.0765986   3.734 0.000188
## soporte_grpMetal                     0.0189324  0.3747131   0.051 0.959704
## soporte_grpMural                     1.7860819  2.3433219   0.762 0.445940
## soporte_grpOtros                    -0.7732117  0.3885090  -1.990 0.046569
## soporte_grpTabla/Panel               0.5866802  0.1614587   3.634 0.000279
## tecnicamixta                         0.4170149  0.3341610   1.248 0.212051
## tecnicaotras                        -0.0504376  0.3820347  -0.132 0.894966
## sop_montajesi                        0.7901158  0.2314649   3.414 0.000641
## seriesi                              0.2797309  0.0937933   2.982 0.002860
## tam_catmediano:soporte_grpMetal     -1.7264476  1.4984077  -1.152 0.249244
## tam_catgrande:soporte_grpMetal      -0.2107848  1.7900398  -0.118 0.906262
## tam_catmediano:soporte_grpMural     -2.3924557  2.4965511  -0.958 0.337909
## tam_catgrande:soporte_grpMural      -2.6305508  2.3924610  -1.100 0.271543
## tam_catmediano:soporte_grpOtros     -1.9023583  1.5254641  -1.247 0.212372
## tam_catgrande:soporte_grpOtros      -0.2635839  1.9476521  -0.135 0.892348
## tam_catmediano:soporte_grpTabla/Panel  0.2242579  0.2499618   0.897 0.369629
## tam_catgrande:soporte_grpTabla/Panel  -0.7513967  0.3711931  -2.024 0.042942
##
```

```
## (Intercept)                             ***
## fecha_est                               ***
## log_ancho
## tam_catmediano
## tam_catgrande                           ***
## orientacionhorizontal                   ***
## soporte_grpMetal
## soporte_grpMural
## soporte_grpOtros                        *
## soporte_grpTabla/Panel                  ***
## tecnicamixta
## tecnicaotras
## sop_montajesi                           ***
## seriesi                                 **
## tam_catmediano:soporte_grpMetal
## tam_catgrande:soporte_grpMetal
## tam_catmediano:soporte_grpMural
## tam_catgrande:soporte_grpMural
## tam_catmediano:soporte_grpOtros
## tam_catgrande:soporte_grpOtros
## tam_catmediano:soporte_grpTabla/Panel
## tam_catgrande:soporte_grpTabla/Panel  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5173.2  on 7001  degrees of freedom
## Residual deviance: 4929.9  on 6980  degrees of freedom
## AIC:  4973.9
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 24
```