

test_final

Diana Yarihuaman

Aldo Salazar

Hugo Murgueza

Marcelo Narvaez

r Sys.Date()

Introducción:

El telemarketing es una estrategia clave para muchas empresas, ya que permite interactuar directamente con los clientes, promoviendo productos o servicios de manera efectiva. Sin embargo, para optimizar estas estrategias, es fundamental analizar los datos generados durante las campañas.

En el presente informe vamos a trabajar con una base de datos “Bank Marketing” de una campaña de marketing realizada por una entidad bancaria portuguesa mediante llamadas telefónicas. La campaña pretende vender suscripciones de un depósito bancario a plazo representado por la variable “y” (suscripción o no suscripción). Para ver la inferencia estadística de la base de datos vamos a realizar un modelo de regresión logística binaria con esto predecimos si un cliente comprará o no una suscripción en función de las variables predictoras es decir, los atributos del cliente, como la información demográfica y los indicadores macroeconómicos.

Exploración y revisión:

Ánalisis exploratorio de las BBDD:

En este apartado se procede a realizar una exploración inicial de la base mediante el programa R-Studio, para poder comprender mejor la base de datos. Como se puede observar tras cargar los datos, la base original con la que partimos cuenta con 4119 observaciones y 21 variables.

```
## [1] 4119   21
```

Descripción de las variables

Las variables explicativas utilizadas en esta base de datos abarcan diversos aspectos relacionados con los clients, que nos permitirán su análisis detallado. A continuación, vamos a enumararlas con sus correspondientes interpretaciones:

- 1) **age**: edad
- 2) **job**: tipo de trabajo
- 3) **marital**: estado civil
- 4) **education**: nivel educativo
- 5) **default**: ¿tiene crédito impagado?
- 6) **housing**: ¿tiene un préstamo de vivienda?
- 7) **loan**: ¿tiene un préstamo personal?
- 8) **contact**: tipo de comunicación de contacto
- 9) **month**: mes del último contacto del año
- 10) **day_of_week**: día de la semana del último contacto

- 11) **duration**: duración del último contacto (en segundos)
- 12) **campaign**: número de contactos realizados durante esta campaña y para este cliente
- 13) **pdays**: número de días que han pasado desde que el cliente fue contactado por última vez en una campaña anterior (999 significa que no ha habido contactos previos)
- 14) **previous**: número de contactos realizados antes de esta campaña para este cliente
- 15) **poutcome**: resultado de la campaña de marketing anterior
- 16) **emp.var.rate**: tasa de variación del empleo (indicador trimestral)
- 17) **cons.price.idx**: índice de precios al consumidor (indicador mensual)
- 18) **cons.conf.idx**: índice de confianza del consumidor (indicador mensual)
- 19) **euribor3m**: tasa euribor a 3 meses (indicador diario)
- 20) **nr.employed**: número de empleados (indicador trimestral)
- 21) **y**: ¿el cliente ha suscrito un depósito a plazo fijo?

Ademas para poder llevar a cabo un mejor análisis, se ha decidido clasificar cada variable según su tipo en una tabla.

Table 1: Tabla de variables y sus tipos

	variable	tipo
age	age	numerical
job	job	nominal
marital	marital	nominal
education	education	ordinal
default	default	nominal
housing	housing	nominal
loan	loan	nominal
contact	contact	nominal
month	month	ordinal
day_of_week	day_of_week	ordinal
duration	duration	numerical
campaign	campaign	numerical
pdays	pdays	numerical
previous	previous	numerical
poutcome	poutcome	nominal
emp.var.rate	emp.var.rate	numerical
cons.price.idx	cons.price.idx	numerical
cons.conf.idx	cons.conf.idx	numerical
euribor3m	euribor3m	numerical
nr.employed	nr.employed	numerical
y	y	nominal

Revisión de incongruencias

Tras una revisión inicial de la BBDD, se pueden observar varias incongruencias en el resumen descriptivo de nuestra fuente .

Por un lado, se observa que en la variable **age** sus valores máximos y mínimos son 350 y 4 años de edad, por lo que decidimos al no estar seguros a que se deben estos datos, convertirlos en NA's. También vemos que en la variable **Pdays** hay muchos valores 999, es decir que no habido contactos previos, por lo que decidimos convertir dichos valores en 0, ya que no ha habido ningún contacto previo con los clientes. Y por último, en la variable **marital** se observa que hay varios errores ortográficos con respecto a la categoría -married- dando lugar a un mayor número de categorías de las que realmente hay, por lo que corregimos dichos errores ortográficos.

Imputación de datos faltantes

Una vez corregidos los errores superficiales, procedemos a imputar los datos faltantes de nuestra BBDD. Para ello primero realizaremos un recuento de cuantas valores -unknown- tenemos para posteriormente convertirlos en NA's. Y luego analizar dichos valores faltantes en base a las observaciones y a las variables, y así poder determinar si es necesario eliminar observaciones o variables con más de un 20% de datos faltantes.

Recuento de unknown por variable:

Exploración del número de faltantes por variable

Tras convertir los valores desconocidos en NA's, procedemos a determinar la cantidad y su porcentaje en cada una de las variables de la BBDD.

Table 2: Na's por variable y su porcentaje

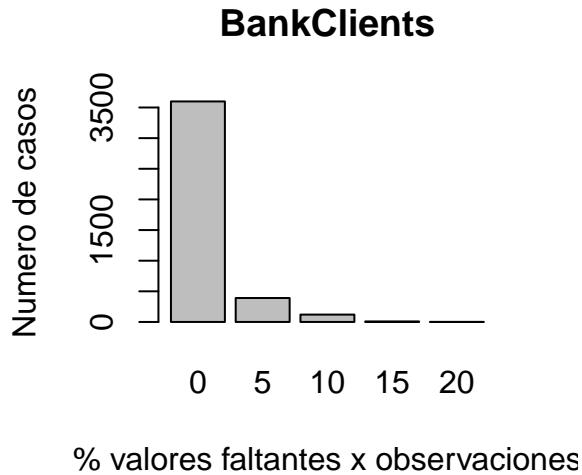
	variable	numNA	percNA
age	age	12	0.29
job	job	99	2.40
marital	marital	53	1.29
education	education	205	4.98
default	default	803	19.50
housing	housing	105	2.55
loan	loan	105	2.55
contact	contact	0	0.00
month	month	0	0.00
day_of_week	day_of_week	0	0.00
duration	duration	0	0.00
campaign	campaign	40	0.97
pdays	pdays	0	0.00
previous	previous	0	0.00
poutcome	poutcome	0	0.00
emp.var.rate	emp.var.rate	0	0.00
cons.price.idx	cons.price.idx	0	0.00
cons.conf.idx	cons.conf.idx	0	0.00
euribor3m	euribor3m	0	0.00
nr.employed	nr.employed	0	0.00
y	y	40	0.97

Se decide eliminar la variable default por no ser relevante para este análisis ya que al realizar nuestra tabla de frecuencias el porcentaje de NA's roza un 20% dejando un porcentaje del 80 % en No del cual no aportaría información relevante para el análisis. Para concluir se actualiza los cambios tanto en la BBDD original como la tabla auxiliar.

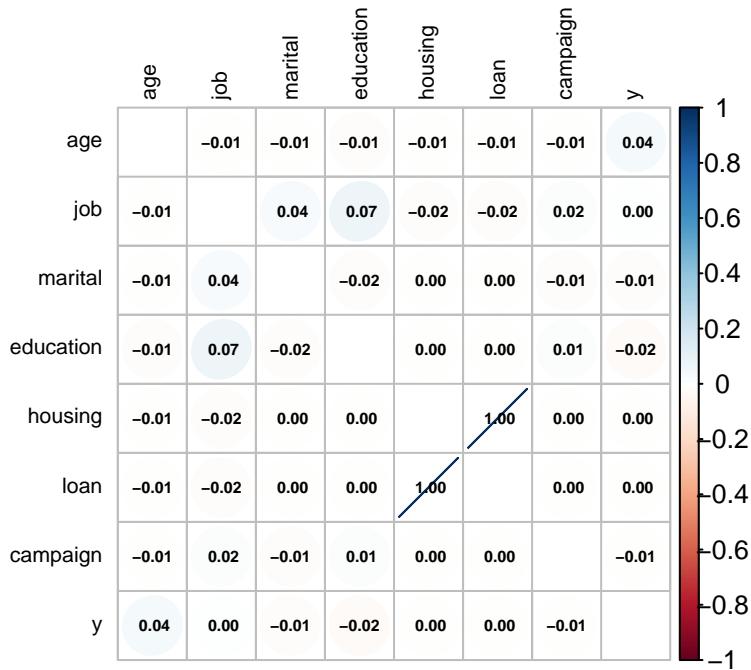
Exploración del número de faltantes por observaciones

En este punto, se procede analizar la cantidad de NA's por observaciones, y realizar las acciones necesarias con respecto a los resultados obtenidos.

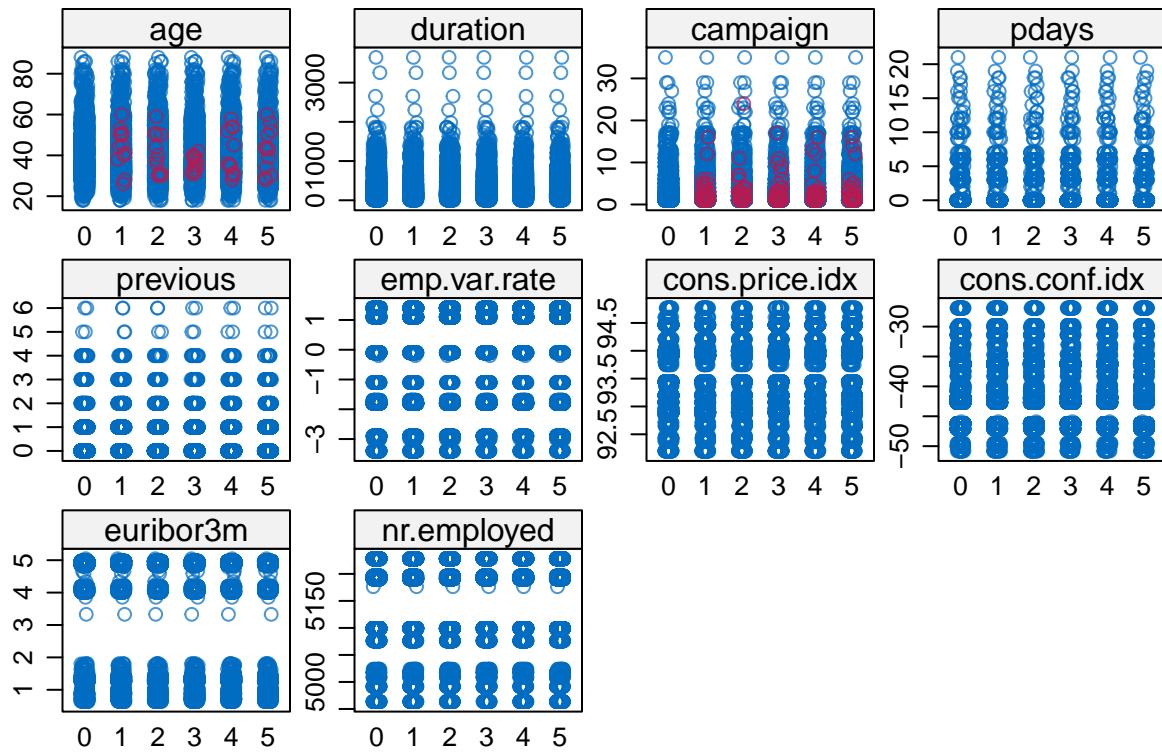
Extrapolamos los datos en un grafico de barras, y se puede confirmar que ninguna observacion cuenta con más de un 20% de datos faltantes, por lo que en principio no sería necesario suprimir ninguna observación.



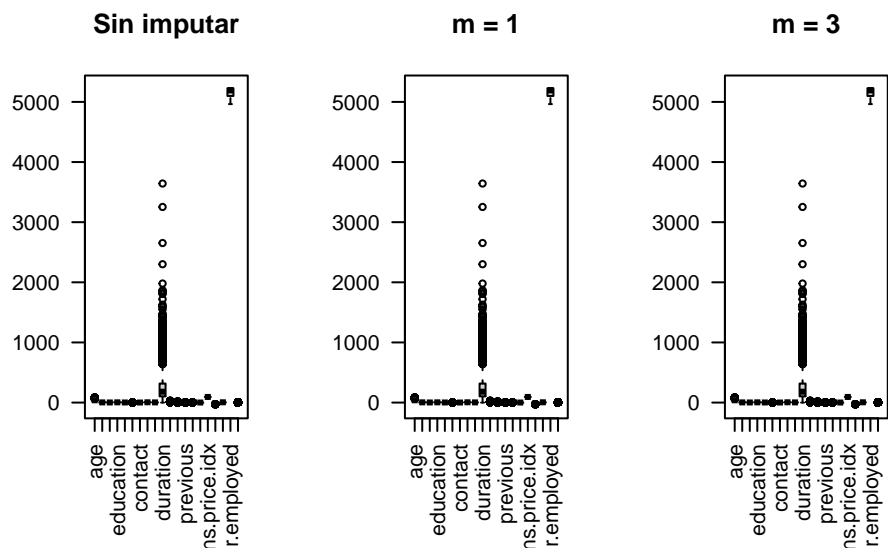
Tambien analizamos si la distribución de los NA's es aleatoria, y si existe una maxima correlación entre las variables. Y podemos observar que hay una fuerte correlacion entre las variables **housing** y **loan**, pero no se decide a aliminarlas ya que esto, no necesariamente implica que haya una redundancia en los valores observados de las variables, sino que su patrón de valores faltantes es idéntico.



Tras eliminar las observaciones y/o variables con un gran porcentaje de NA's, procedemos a imputar dicho valores para poder continuar con nuestro proyecto. Para ello mediante la libreria ‘MICE’ el metodo de imputación predictiva “pmm”, que asegura que los valores imputados se ajusten a los rangos y distribuciones observados en las variables correspondientes. En este caso se han realizado 5 imputaciones diferentes



Ahora procedemos a comparar la base original con los datos faltantes, y las imputaciones creadas, como se observa en el grafico de cajas y bigotes, no hay mucha diferencia entre las distintas imputaciones, por lo que decidimos seleccionar la primera imputacion creada para nuestra base de datos, ya que rellena los datos manteniendo la distribución original.



Identificación de variables

Veamos que variables no pueden o no deben ser incluidas en los análisis estadísticos, procedemos con las variables numéricicas identificando variables constantes, casi constante y valores anómalos.

Variables numéricas

Variables constantes, usaremos parámetros de dispersión: desviación típica y coeficiente de desviación. Se calculan la desviación estándar y el coeficiente de variación para identificar variables con baja dispersión.

Table 3: Resumen de Variables Numéricas

age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
Min. :18.00	Min. : 0.0	Min. : 1.000	Min. : 0.0000	Min. : 0.0000	Min. :-3.40000	Min. :92.20	Min. :-50.8	Min. :0.635	Min. :4964
1st Qu.:32.00	1st Qu.: 103.0	1st Qu.: 1.000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: -1.80000	1st Qu.: 93.08	1st Qu.: -42.7	1st Qu.: 1.334	1st Qu.: 5099
Median :38.00	Median : 181.0	Median : 2.000	Median : 0.0000	Median : 0.0000	Median : 1.10000	Median : 93.75	Median : -41.8	Median : 4.857	Median : 5191
Mean :40.12	Mean : 256.8	Mean : 2.542	Mean : 0.2277	Mean : 0.1903	Mean : 0.08497	Mean : 93.58	Mean : -40.5	Mean : 3.621	Mean : 5166
3rd Qu.:47.00	3rd Qu.: 317.0	3rd Qu.: 3.000	3rd Qu.: 0.0000	3rd Qu.: 0.0000	3rd Qu.: 1.40000	3rd Qu.: 93.99	3rd Qu.: 36.4	3rd Qu.: 4.961	3rd Qu.: 5228
Max. :88.00	Max. :3643.0	Max. :35.000	Max. :21.0000	Max. :6.0000	Max. :1.40000	Max. :94.77	Max. :-26.9	Max. :5.045	Max. :5228

Se observa de la tabla 3 que no podemos observar variables constantes.

Para ver si tenemos variables “casi constantes” utilizaremos la desviación estándar o coeficiente de variación: Creamos un bucle para calcular la desviación estándar y coeficiente de variación.

Variable	SD	CV
age	10.32	0.26
duration	254.70	0.99
campaign	2.58	1.01
pdays	1.37	6.01
previous	0.54	2.85
emp.var.rate	1.56	18.40
cons.price.idx	0.58	0.01
cons.conf.idx	4.59	-0.11
euribor3m	1.73	0.48
nr.employed	73.67	0.01

Observamos que `duration` tienen una dispersión alta pero las mantenemos por ser inherentes de las variables. La variable `Cons.price.idx` es una variable numérica que si bien su dispersión es pequeña, no es discreta.

La variable `Cons.price.idx` es una variable numérica que si bien su dispersión es pequeña, no es discreta

Variables categóricas

Se crean tablas de frecuencia (absolutas y relativas) para variables categóricas ordinales y nominales, destacando variables con muchas categorías o bajas frecuencias que podrían requerir recodificación.

Variable	Categoría	Porcentaje
education	university.degree	32.022
education	high.school	23.185
education	basic.9y	14.591
education	professional.course	13.523
education	basic.4y	10.876
education	basic.6y	5.778
education	illiterate	0.024
month	may	33.455
month	jul	17.261
month	aug	15.441
month	jun	12.867
month	nov	10.828
month	apr	5.220
month	oct	1.675
month	sep	1.554
month	mar	1.165
month	dec	0.534
day_of_week	thu	20.879
day_of_week	mon	20.757
day_of_week	tue	20.418
day_of_week	wed	19.301
day_of_week	fri	18.645

Variable	Categoría	Porcentaje
job	admin.	24.739
job	blue-collar	21.462
job	technician	16.970
job	services	9.638
job	management	8.060
job	retired	4.079
job	self-employed	3.957
job	entrepreneur	3.714
job	unemployed	2.695
job	housemaid	2.598
job	student	2.088
marital	married	60.889
marital	single	28.211
marital	divorced	10.901
housing	yes	54.261
housing	no	45.739
loan	no	83.418
loan	yes	16.582
contact	cellular	64.385
contact	telephone	35.615
poutcome	nonexistent	85.530
poutcome	failure	11.022
poutcome	success	3.447
y	no	88.978
y	yes	11.022

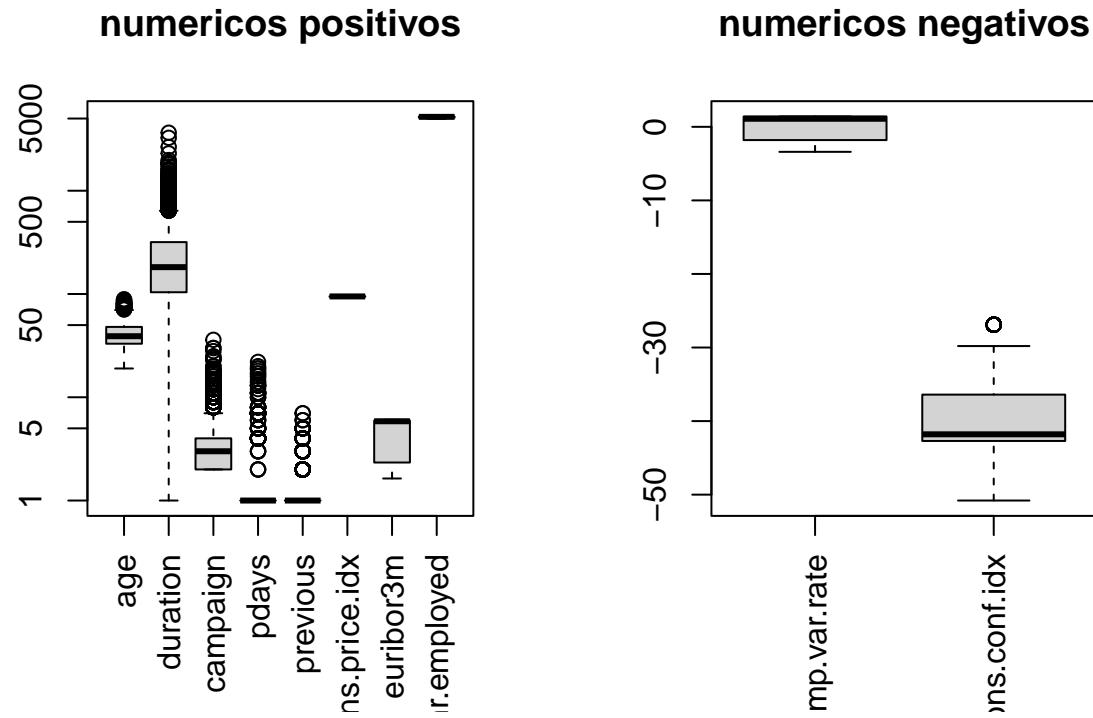
No se observa nada notable en la variabilidad de estas variables categóricas. Únicamente destacar que la variable job y education tiene muchas categorías (lo que puede dificultar su análisis e interpretación) y algunas de ellas con muy baja frecuencia. Se estudiará más adelante como recodificar esta variable para agrupar las categorías de forma que ninguna tenga una frecuencia demasiado baja.

El resumen descriptivo de las variables categóricas y binarias mediante tablas de frecuencias no mostró ningún valor inconsistente.

Datos Anómalos

Separación de variables numéricas en positivas y negativas para facilitar su análisis mediante boxplots y evitar distorsiones en la representación logarítmica. Eliminación de observaciones específicas no relevantes, como valores de duración de llamadas iguales a 0, que afectan la normalidad.

Variables numéricas como se observó en las variables numéricas hay 2 variables donde su mínimo valores son negativos y para poder graficar boxplot usando log no se permite que sean negativos, se realiza un análisis separando los numéricos positivos y a parte los negativos para no distorsionar los valores.



Observando el boxplot de numéricos positivos y tratándose “duration” de una variable que expresa la duración de la llamada, y teniendo solo una observación con valor 0, se considera no relevante y se procede a eliminarla. Además, afecta muy significativamente a la normalidad de esta variable.

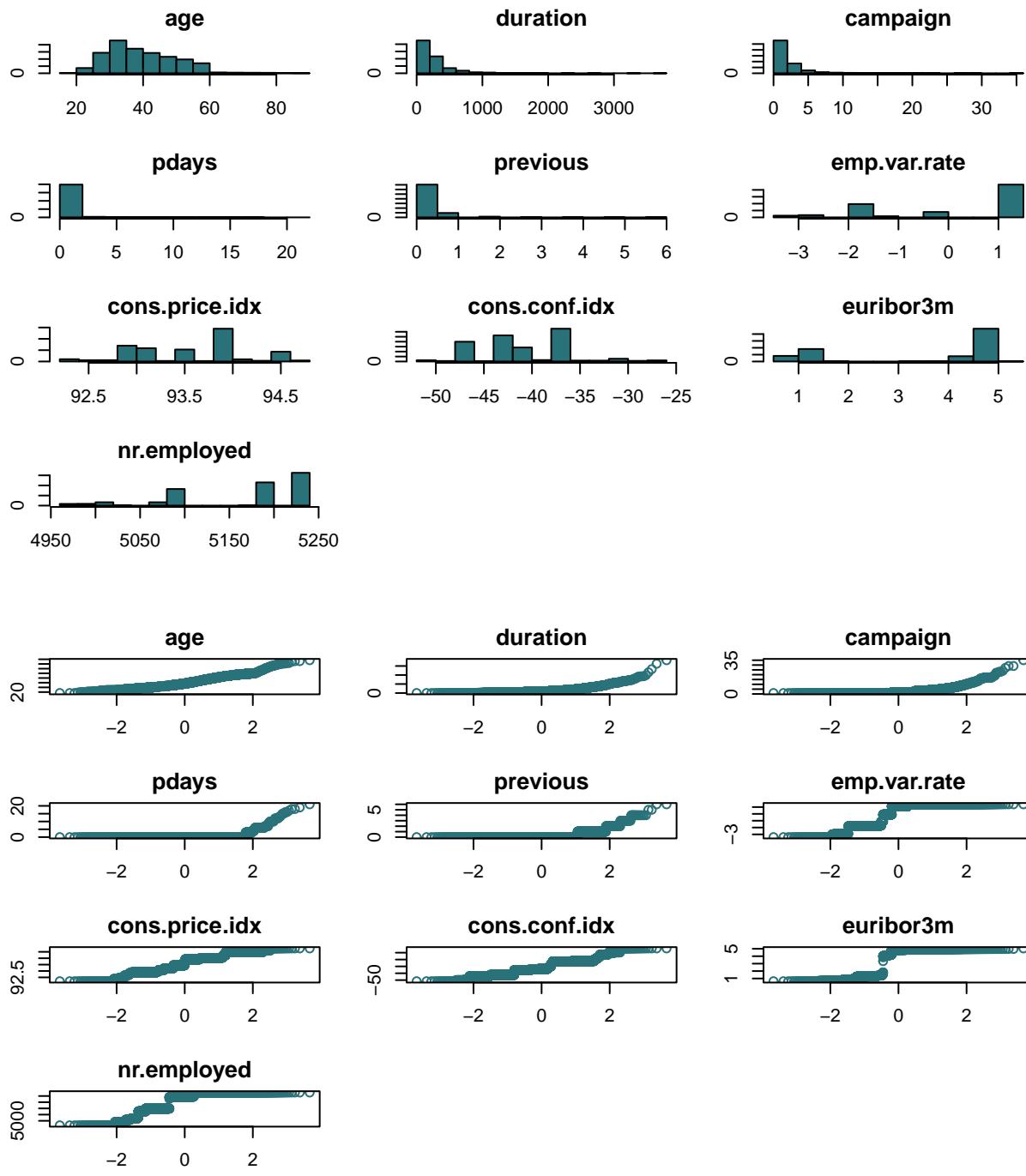
Transformación y Categorización de Variables

En esta sección, transformamos variables numéricas del dataset para mejorar su distribución y facilitar el análisis estadístico. La transformación nos ayudara a reducir el **sesgo** en variables como **duration** y facilitara la aplicación de técnicas estadísticas que asumen **normalidad**.

Variables numéricas

Para cada variable numérica, evaluamos:

- **Distribución inicial** mediante histogramas.
- **Ajuste a normalidad** con gráficos Q-Q y pruebas estadísticas.
- **Transformaciones** que mejoran la normalidad.

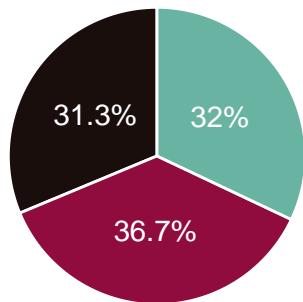


Del gráfico global de las variables numericas podemos decir que la variables duration,campaign son valores que requieren transformaciones mientras que para previous se observa primero una alta concentración sobre grupos entre 0 y 1 , entonces es recomendable categorizarlo y para la variable age podremos intentar transformarlo para obtener una mejor distribución.

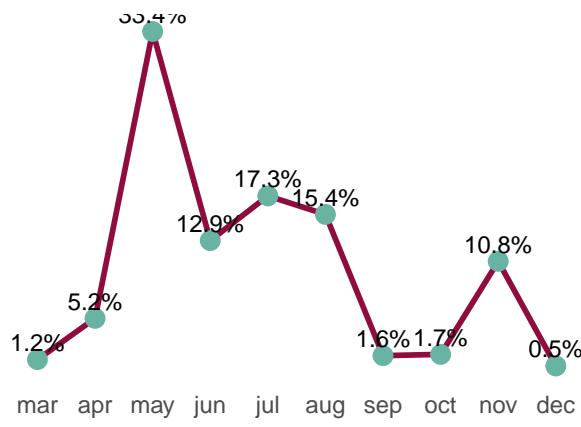
Variables categoricas

Se procede a analizar la distribución de las variables categóricas. El objetivo es revisar si estas variables se encuentran balanceadas. En particular, para la variable education, se busca analizar su distribución, ya que esta puede influir significativamente en los resultados posteriores. Debido a que la variable presenta una gran cantidad de categorías, se agruparán niveles similares para facilitar el análisis y la interpretación.

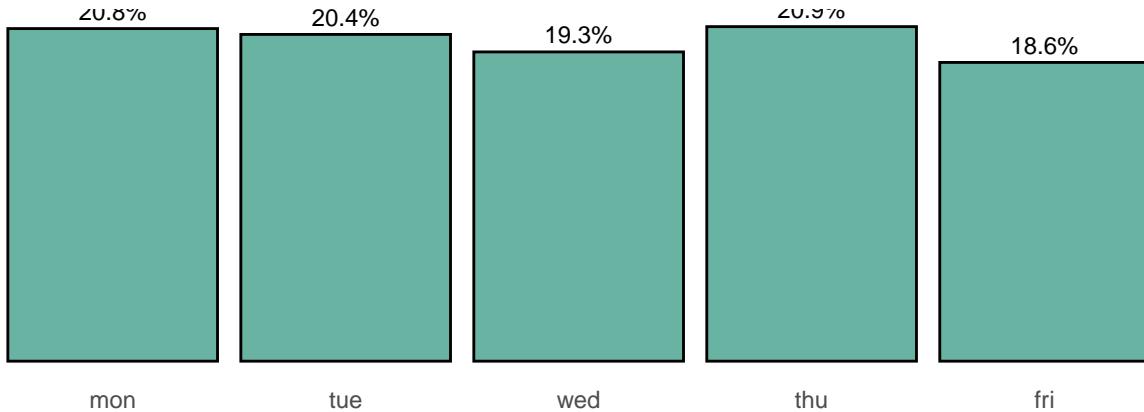
Distribucion por Niveles de Educacion



Distribucion % por Mes



Distribucion % por Dia de la semana



Como observamos en los gráficos la distribución de la variable educación es relativamente balanceada entre los tres niveles, lo que garantiza representatividad en los diferentes segmentos educativos , el gráfico de línea muestra cómo se distribuyen por mes, destacando un mayor porcentaje que podría estar relacionada con campañas específicas o estrategias del banco en ese período y para la variable day_of_week muestra una distribución relativamente uniforme.

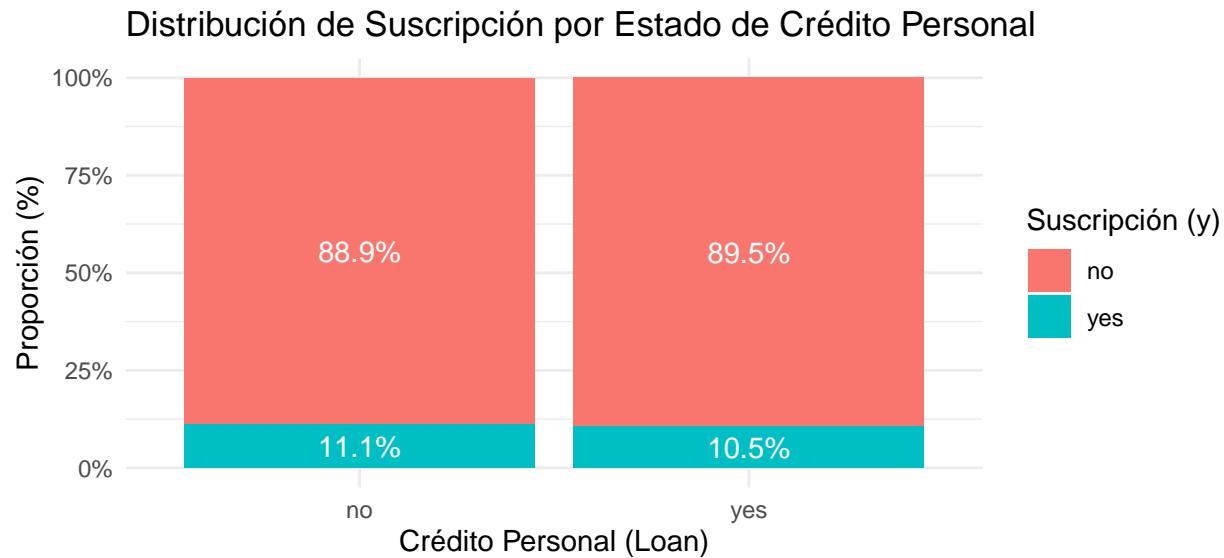
Inferencia estadística

En este apartado de inferencia estadística se evaluaron hipótesis relacionadas con la probabilidad de que los clientes de un banco se suscriban a un depósito a plazo, utilizando un conjunto de variables categóricas (como Loan, Month, Job, Education, Previous) y numéricas (como Age, Campaign y Duration). A partir de estas variables, se formularon ocho hipótesis que exploran cómo estas características afectan la probabilidad de suscripción. Las pruebas estadísticas empleadas incluyen el test de Chi-cuadrado para evaluar la relación entre variables categóricas y la respuesta (y), así como pruebas no paramétricas como el test de Wilcoxon para comparar medianas entre diferentes grupos. Los resultados se visualizan mediante gráficos de barras apiladas y análisis post-hoc, como la prueba de Tukey, para analizar diferencias significativas entre subgrupos.

Relación entre aceptación de crédito y posesión de un crédito personal

Se analiza la relación entre la posesión de un crédito personal (Loan) y la probabilidad de que un cliente se suscriba a un depósito a plazo. Las herramientas estadísticas incluyen el test de **Chi-cuadrado** para evaluar la

independencia entre estas variables y gráficos de barras apiladas para observar visualmente las proporciones de aceptación.



```
## 
## Pearson's Chi-squared test with Yates' continuity correction
## 
## data: contingency_table_loan
## X-squared = 0.1402, df = 1, p-value = 0.7081
```

El p-valor no es inferior a 0,05, por lo que no existe una relación significativa entre tener un crédito personal y la probabilidad de suscribirse a un depósito a plazo.

Relación entre aceptación de crédito y el mes de contacto

Este apartado examina si el mes en el que se contacta al cliente influye en su decisión de aceptar un depósito a plazo. Se utilizan el test de **Chi-cuadrado** y el **test de Wilcoxon** pareado para evaluar diferencias entre meses, y gráficos de barras apiladas para visualizar tendencias estacionales.

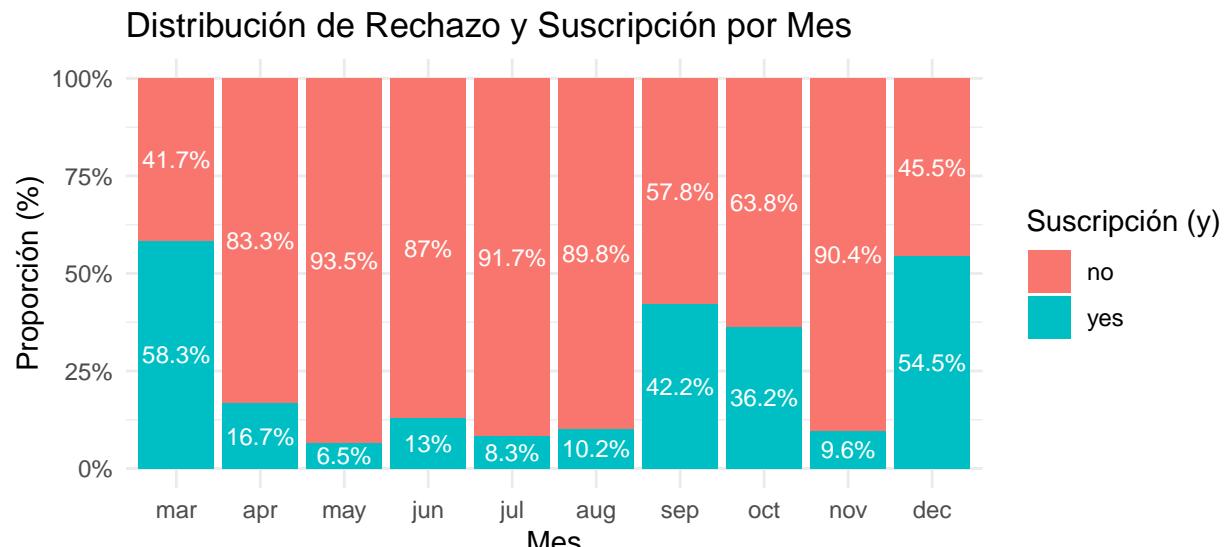


Table 5: Resultados de la prueba de Wilcoxon (formato wide)

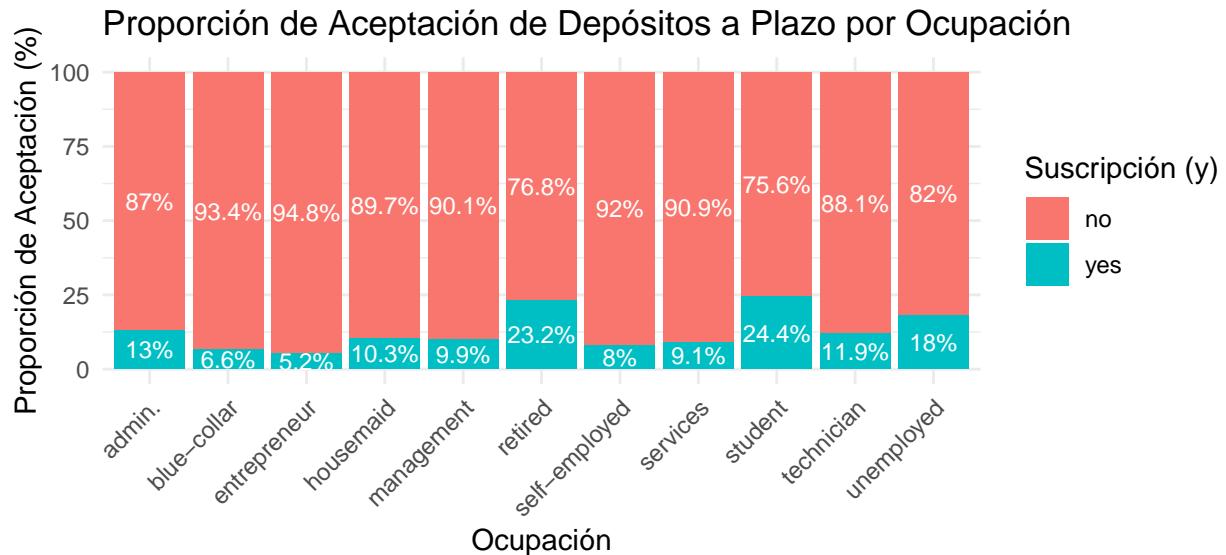
Comparison1	mar	apr	may	jun	jul	aug	sep	oct	nov
apr	0.0000	NA	NA						
may	0.0000	0.0000	NA	NA	NA	NA	NA	NA	NA
jun	0.0000	0.2145	0.0000	NA	NA	NA	NA	NA	NA
jul	0.0000	0.0006	0.1639	0.0106	NA	NA	NA	NA	NA
aug	0.0000	0.0154	0.0064	0.1639	0.2514	NA	NA	NA	NA
sep	0.1229	0.0000	0.0000	0.0000	0.0000	0.0000	NA	NA	NA
oct	0.0265	0.0010	0.0000	0.0000	0.0000	0.0000	0.5078	NA	NA
nov	0.0000	0.0126	0.0388	0.1277	0.4642	0.7719	0.0000	0.0000	NA
dec	0.7735	0.0001	0.0000	0.0000	0.0000	0.0000	0.3521	0.1639	0

```
## 
## Pearson's Chi-squared test
## 
## data: table(BankClients$respuesta_binaria, BankClients$month)
## X-squared = 304.33, df = 9, p-value < 2.2e-16
```

Los valores coloreados en rojo indican un Pvalor inferior a 0,05. Esto indica significación estadística y por lo tanto diferencias entre los meses de contacto respecto a la probabilidad e aceptar el crédito.

Relación entre aceptación de crédito y ocupación del cliente

Se evalúa si la ocupación (Job) afecta la probabilidad de suscripción a depósitos a plazo. Las pruebas incluyen el test de **Chi-cuadrado** para explorar relaciones significativas y el **test de Wilcoxon** pareado para comparar ocupaciones específicas, apoyado en gráficos de barras apiladas.



```
## 
## Pearson's Chi-squared test
## 
## data: contingency_table
## X-squared = 77.886, df = 10, p-value = 1.302e-12
```

Table 6: Resultados de la prueba de Wilcoxon (formato wide)

Comparison1	admin.	blue-collar	entrepreneur	housemaid	management	retired	self-employed	services	student	technician
blue-collar	0.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA
entrepreneur	0.0173	0.5988	NA	NA	NA	NA	NA	NA	NA	NA
housemaid	0.5348	0.2180	0.2078	NA	NA	NA	NA	NA	NA	NA
management	0.2180	0.0941	0.1531	0.9195	NA	NA	NA	NA	NA	NA
retired	0.0017	0.0000	0.0001	0.0188	0.0004	NA	NA	NA	NA	NA
self-employed	0.1358	0.5923	0.4305	0.5923	0.5869	0.0007	NA	NA	NA	NA
services	0.0883	0.1920	0.2180	0.7292	0.7291	0.0001	0.7291	NA	NA	NA
student	0.0097	0.0000	0.0001	0.0222	0.0014	0.8474	0.0014	0.0004	NA	NA
technician	0.5923	0.0011	0.0367	0.6962	0.4589	0.0007	0.2180	0.2180	0.0039	NA
unemployed	0.2180	0.0002	0.0031	0.1830	0.0515	0.4018	0.0296	0.0206	0.3773	0.1358

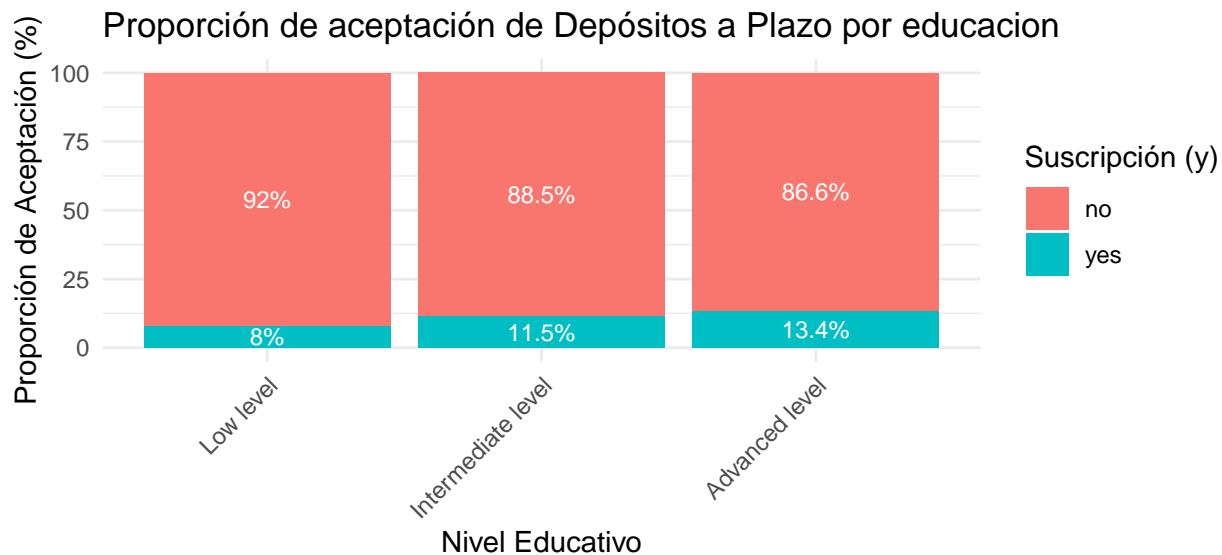
Table 7: Resultados de la prueba de Wilcoxon (formato wide)

Comparison1	Low level	Intermediate level
Intermediate level	0.0028	NA
Advanced level	0.0000	0.1254

Los valores coloreados en rojo indican un Pvalor inferior a 0,05. Esto indica significación estadística y por lo tanto diferencias entre los meses de contacto respecto a la probabilidad e aceptar el crédito.

Relación entre aceptación de crédito y nivel educativo

Se explora el impacto del nivel educativo (Education) en la probabilidad de suscripción. Se aplican el test de **Chi-cuadrado** para evaluar relaciones globales y el **test de Wilcoxon** pareado para identificar diferencias específicas entre niveles educativos. Los resultados se presentan mediante gráficos de barras apiladas.



```
##
## Pearson's Chi-squared test
##
## data: contingency_table
## X-squared = 20.119, df = 2, p-value = 4.279e-05
```

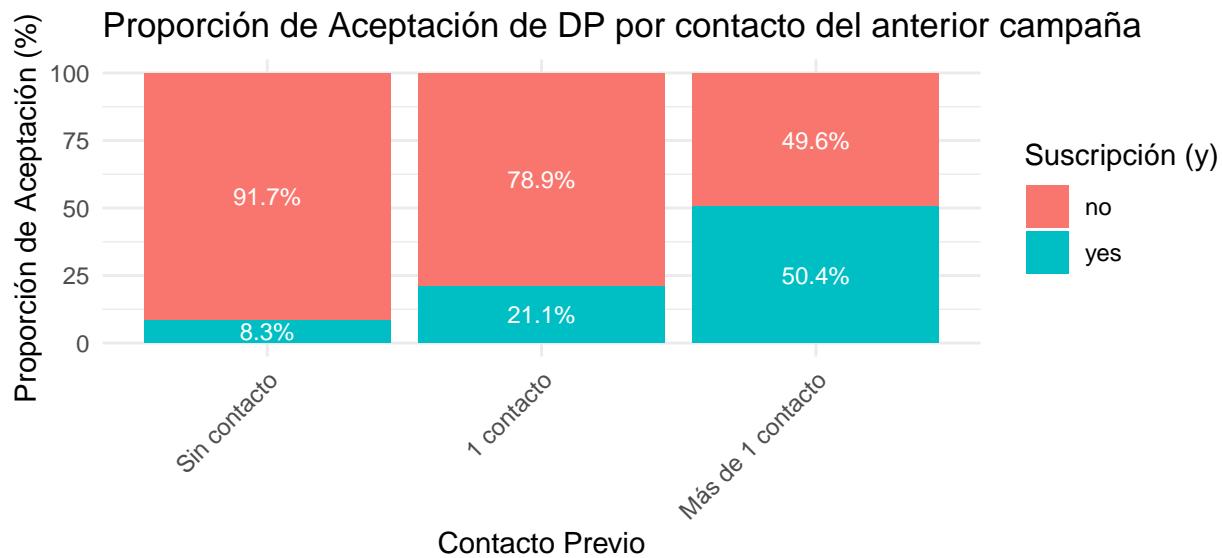
Los valores coloreados en rojo indican un P-valor inferior a 0,05. Esto demuestra significación estadística y, por lo tanto, diferencias entre los niveles educativos en relación con la probabilidad de aceptar el crédito.

Table 8: Resultados de la prueba de Wilcoxon (formato wide)

Comparison1	Sin contacto	1 contacto
1 contacto	0	NA
Más de 1 contacto	0	0

Relación entre aceptación de crédito y contactos previos

Este apartado analiza si haber sido contactado en campañas anteriores **Previous** incrementa la probabilidad de suscripción. Se aplican el test de **Chi-cuadrado** para analizar la asociación entre contactos previos y aceptación, junto con el **test de Wilcoxon** pareado para identificar diferencias significativas.



```
##
## Pearson's Chi-squared test
##
## data: contingency_table
## X-squared = 266.35, df = 2, p-value < 2.2e-16
```

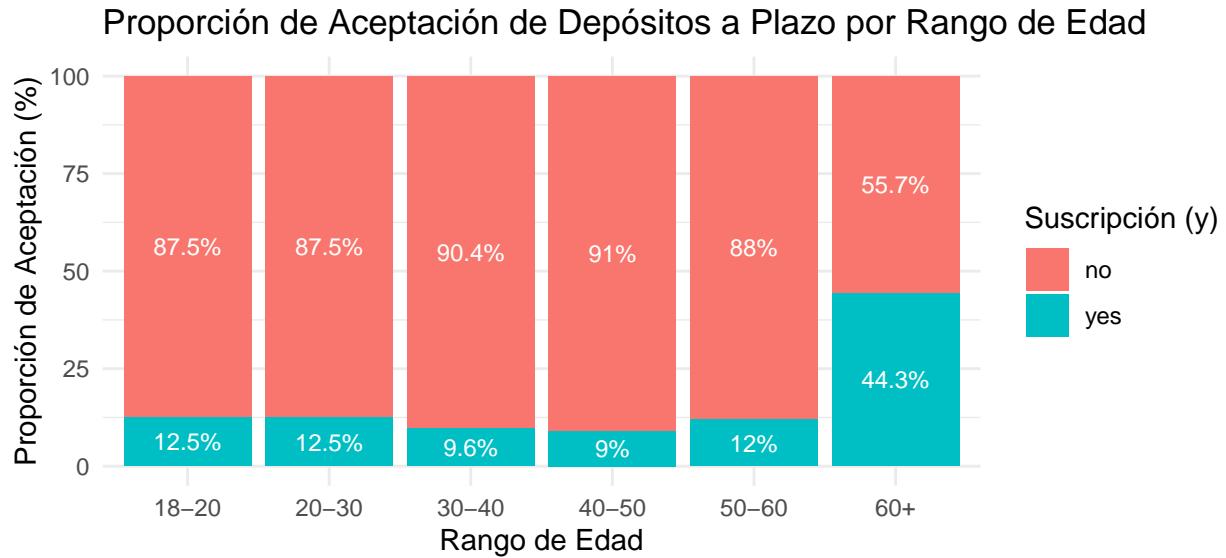
Los valores coloreados en rojo indican un P-valor inferior a 0,05. Esto evidencia significación estadística y, en consecuencia, diferencias entre los grupos de clientes según el número de contactos previos respecto a la probabilidad de aceptar el crédito.

Relación entre aceptación de crédito y rango de edad

Se estudia si ciertos rangos etarios presentan mayores probabilidades de suscripción. La variable Age se recategorizó en intervalos definidos (18-20, 20-30, etc.) para facilitar el análisis. Se usan el test de **Chi-cuadrado** y el **test de Wilcoxon** pareado para identificar diferencias entre grupos de edad.

Table 9: Resultados de la prueba de Wilcoxon (formato wide)

Comparison1	18-20	20-30	30-40	40-50	50-60
20-30	0.9988	NA	NA	NA	NA
30-40	0.9061	0.0994	NA	NA	NA
40-50	0.9061	0.0598	0.8846	NA	NA
50-60	0.9988	0.9061	0.1628	0.1063	NA
60+	0.1596	0.0000	0.0000	0.0000	0



```
##
## Pearson's Chi-squared test
##
## data: contingency_table_age
## X-squared = 98.743, df = 5, p-value < 2.2e-16
```

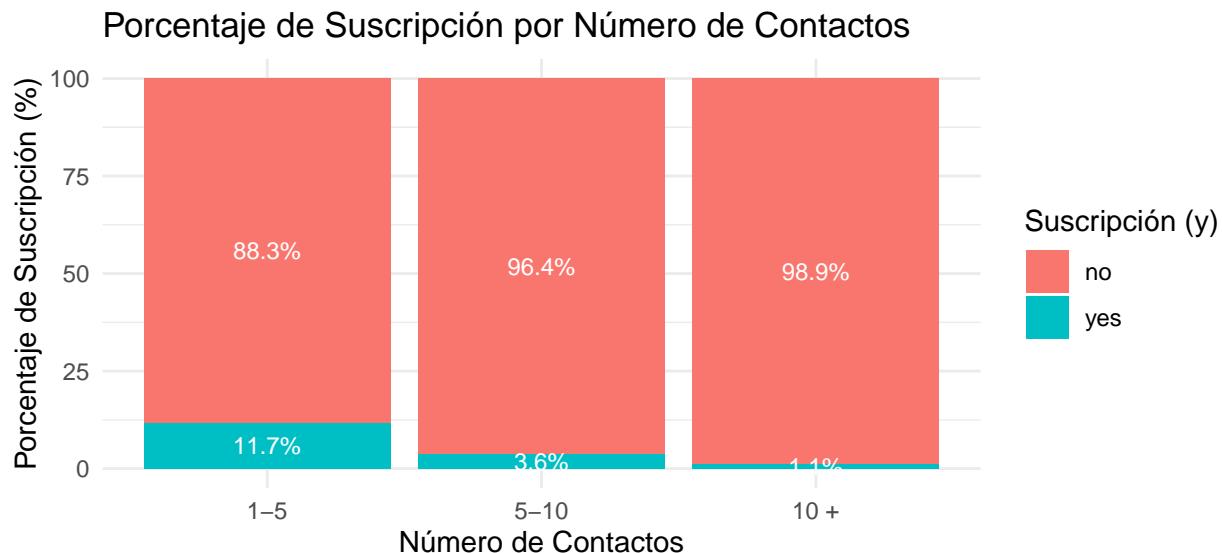
Los valores coloreados en rojo indican un P-valor inferior a 0,05. Esto indica significación estadística y, por lo tanto, diferencias entre los rangos de edad respecto a la probabilidad de aceptar el crédito.

Relación entre aceptación de crédito y número de contactos

Se evalúa el impacto del número de contactos realizados durante la campaña (Campaign) en la probabilidad de suscripción. La variable fue recategorizada en tres grupos (1-5, 5-10, más de 10) para analizar si la frecuencia de contacto afecta la decisión. Las herramientas incluyen el test de **Chi-cuadrado** y el **test de Wilcoxon** pareado.

Table 10: Resultados de la prueba de Wilcoxon (formato wide)

Comparison1	1-5	5-10
5-10	0.0003	NA
10 +	0.0031	0.2369



```
##
## Pearson's Chi-squared test
##
## data: contingency_table
## X-squared = 24.445, df = 2, p-value = 4.919e-06
```

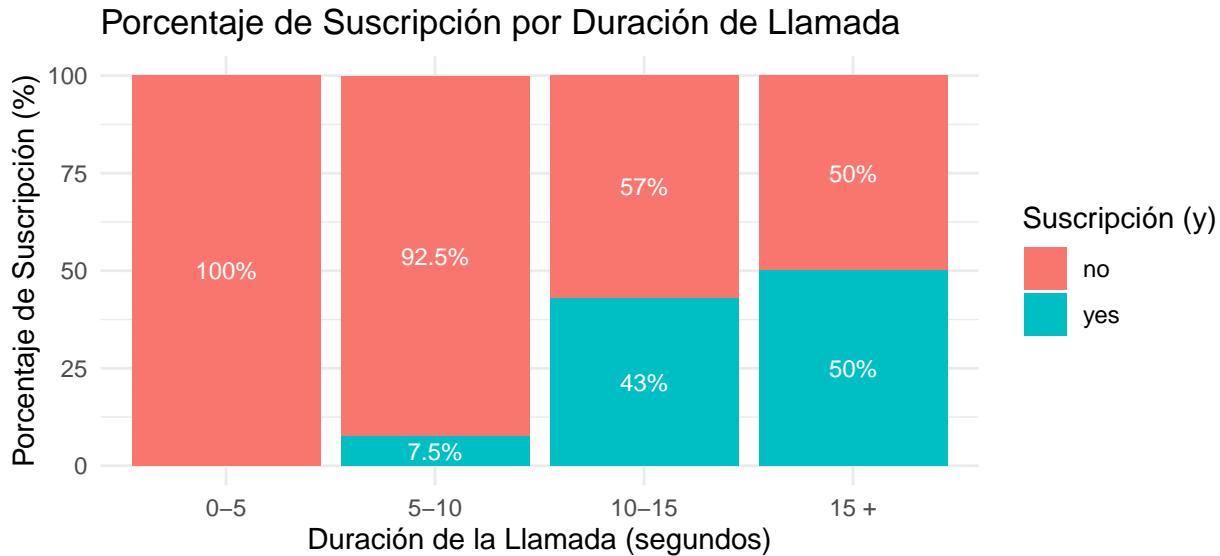
Los valores coloreados en rojo indican un P-valor inferior a 0,05. Esto sugiere significación estadística y, en consecuencia, diferencias entre los grupos según el número de contactos realizados en la probabilidad de aceptar el crédito.

Relación entre aceptación de crédito y duración de la llamada

Este análisis explora cómo la duración de las llamadas (Duration) afecta la probabilidad de aceptación. La variable fue recategorizada en intervalos (0-5, 5-10, etc.) para analizar cómo las diferentes duraciones influyen en las decisiones de los clientes. Se aplican el test de **Chi-cuadrado**, el test de **Wilcoxon** pareado, y gráficos de barras.

Table 11: Resultados de la prueba de Wilcoxon (formato wide)

Comparison1	0-5	5-10	10-15
5-10	0	NA	NA
10-15	0	0.0000	NA
15 +	0	0.0271	0.8447



```
##
## Pearson's Chi-squared test
##
## data: contingency_table_duration
## X-squared = 564.72, df = 3, p-value < 2.2e-16
```

Los valores coloreados en rojo indican un P-valor inferior a 0,05. Esto indica significación estadística y, por ende, diferencias entre los rangos de duración de llamada respecto a la probabilidad de aceptar el crédito.

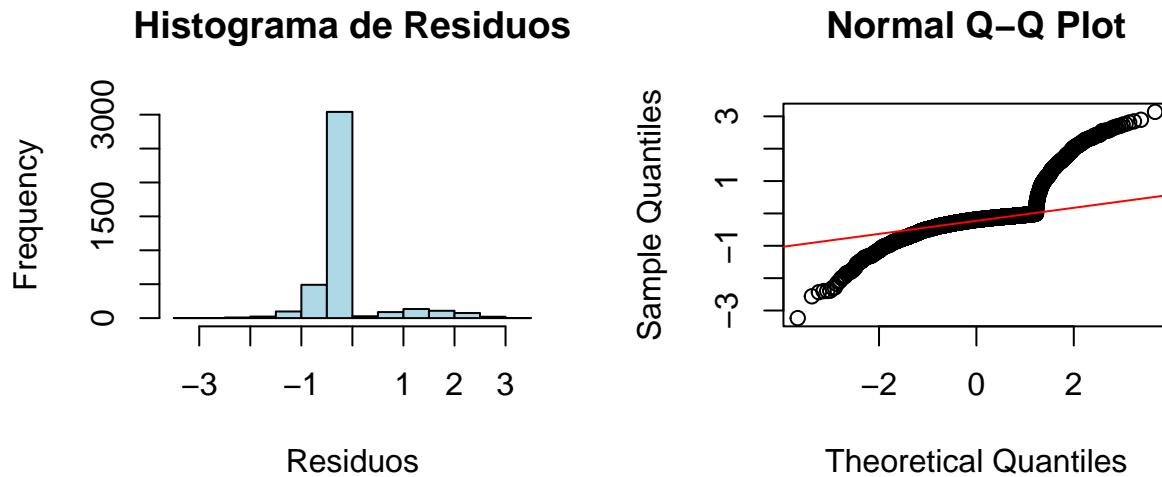
Regresión Binaria

En este análisis se construye un modelo de regresión logística para identificar los factores que influyen en la probabilidad de que un cliente acepte un depósito a plazo. La variable dependiente es binaria, indicando si el cliente acepta (1) o no (0). Las variables independientes incluyen características demográficas (edad, ocupación, nivel educativo), aspectos relacionados con la campaña (duración de la llamada, número de contactos, mes de contacto) y factores financieros (crédito personal, contactos previos). Este modelo permite evaluar la magnitud y dirección del efecto de cada variable sobre la decisión del cliente, proporcionando una herramienta estadística para comprender y predecir el comportamiento de los clientes.

En este análisis se evalúa la validez de un modelo de regresión logística mediante pruebas de diagnóstico que verifican los supuestos necesarios para su correcta interpretación. Se examina la distribución de los residuos deviance para detectar posibles desajustes, la independencia de los residuos utilizando el test de Durbin-Watson, y la bondad de ajuste global mediante la prueba de Hosmer-Lemeshow. Estos procedimientos permiten determinar si el modelo es adecuado para explicar y predecir la variable de interés.

Table 12: Términos Significativos de la Regresión Logística

	Variable	Coeficiente	P-valor
(Intercept)	(Intercept)	-7.7630	0.0000
duration	duration	0.9196	0.0000
campaign	campaign	-0.1468	0.0006
jobblue-collar	jobblue-collar	-0.5052	0.0421
jobentrepreneur	jobentrepreneur	-1.0249	0.0212
monthapr	monthapr	-2.6222	0.0000
monthmay	monthmay	-3.4426	0.0000
monthjun	monthjun	-2.2213	0.0000
monthjul	monthjul	-3.1237	0.0000
monthaug	monthaug	-2.6041	0.0000
monthsep	monthsep	-1.5752	0.0013
monthoct	monthoct	-1.2663	0.0093
monthnov	monthnov	-3.2102	0.0000
educationAdvanced level	educationAdvanced level	0.4663	0.0275
previous1 contacto	previous1 contacto	1.2025	0.0000
previousMás de 1 contacto	previousMás de 1 contacto	2.3012	0.0000



```
##
## Durbin-Watson test
##
## data: modelo
## DW = 1.9853, p-value = 0.3189
## alternative hypothesis: true autocorrelation is greater than 0

##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: BankClients$respuesta_binaria, fitted(modelo)
## X-squared = 11.68, df = 8, p-value = 0.1661
```

El análisis de los resultados obtenidos en la regresión logística revela varios hallazgos importantes sobre los factores que afectan la probabilidad de aceptar el crédito. A continuación, se destacan los puntos clave:

- Intercepto:** El coeficiente del intercepto (-7.7630, $p < 0.0001$) indica que, en ausencia de todas las variables explicativas, la probabilidad de aceptar el crédito es extremadamente baja.
- Duración de la llamada:** La variable `duration` tiene un coeficiente positivo significativo (0.9196, $p < 0.0001$). Esto sugiere que a medida que aumenta la duración de la llamada, también aumenta significativamente la probabilidad de que el cliente acepte el crédito.

- Número de contactos previos (campaign):** Presenta un coeficiente negativo (-0.1468, $p = 0.0006$), lo que indica que contactar al cliente repetidamente reduce ligeramente la probabilidad de aceptación del crédito.

4. Ocupación:

- `jobblue-collar` (-0.5052, $p = 0.0421$) y `jobentrepreneur` (-1.0249, $p = 0.0212$): Estos grupos tienen una menor probabilidad de aceptar el crédito en comparación con la categoría de referencia (presumiblemente otros empleos).

5. Meses del contacto:

- Variables como `monthapr`, `monthmay`, `monthjun`, `monthjul`, `monthaug`, y `monthsep` tienen coeficientes negativos significativos. Esto indica que la probabilidad de aceptar el crédito es considerablemente menor durante estos meses en comparación con el mes de referencia. Por ejemplo, el mes de mayo (`monthmay`) muestra una fuerte disminución en la probabilidad con un coeficiente de -3.4426 ($p < 0.0001$).

Conclusiones del modelo

- **Duración** es el factor más relevante para incrementar la probabilidad de aceptación del crédito, reflejando la importancia de conversaciones largas y detalladas.
- **Mes del contacto y ocupación** también influyen significativamente en la probabilidad de aceptación, lo que puede guiar estrategias de segmentación temporal y profesional.
- **Número de contactos** tiene un impacto negativo, sugiriendo que insistir demasiado puede ser contraproducente.

Estos resultados pueden ser usados para optimizar la estrategia de marketing, priorizando llamadas de mayor duración y evitando un número excesivo de contactos en meses desfavorables.

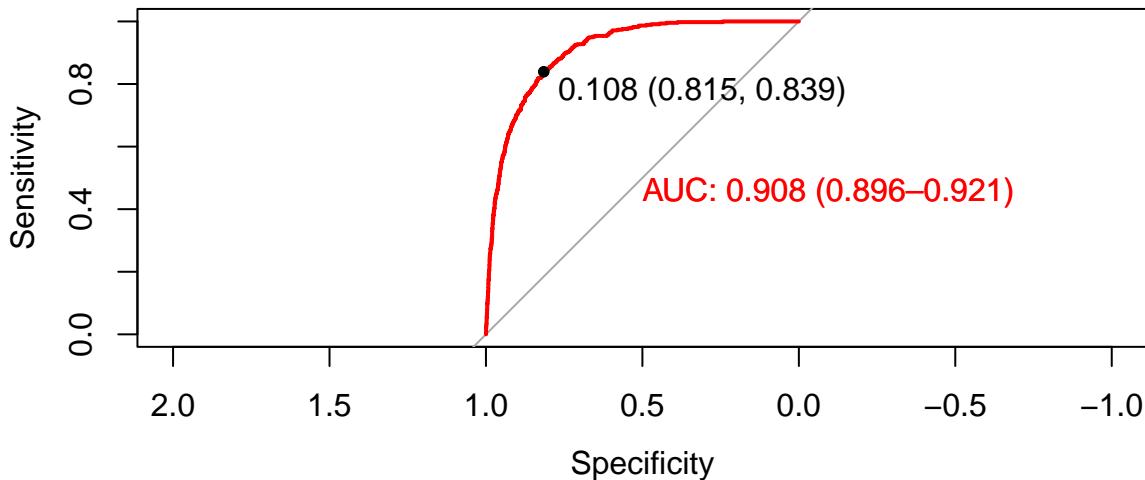
validación del modelo

- **Distribución de Residuos:** o El histograma de los residuos muestra una distribución aproximadamente simétrica, aunque con algunas desviaciones de la normalidad en los extremos, como se evidencia también en el QQ-plot. Esto no es un requisito estricto en regresiones logísticas, pero ayuda a identificar posibles desajustes.
- **Independencia de Residuos:** o El test de Durbin-Watson arrojó un valor cercano a 2, lo que indica que no hay autocorrelación significativa entre los residuos. Esto sugiere que el modelo cumple con el supuesto de independencia.
- **Bondad de Ajuste Global:** o La prueba de Hosmer-Lemeshow resultó en un p-valor de 0.16, lo que implica que no hay evidencia estadísticamente significativa para rechazar la hipótesis nula de un buen ajuste. En otras palabras, las predicciones del modelo se ajustan razonablemente bien a los datos observados.

Curva Roc

En esta sección, se construirá y analizará la **curva ROC** (**Receiver Operating Characteristic**) y se calculará el **AUC** (**Área Bajo la Curva**) para evaluar el desempeño del modelo de regresión logística. La curva ROC permite visualizar la capacidad del modelo para discriminar entre los clientes que aceptan o no aceptan el crédito, mostrando la relación entre la sensibilidad (tasa de verdaderos positivos) y la especificidad (tasa de verdaderos negativos) en diferentes umbrales de probabilidad.

```
##  
## Call:  
## roc.formula(formula = BankClients$y ~ probabilidad_modelo, auc = T,      ci = T)  
##  
## Data: probabilidad_modelo in 3664 controls (BankClients$y no) < 454 cases (BankClients$y yes).  
## Area under the curve: 0.9084  
## 95% CI: 0.8962–0.9206 (DeLong)
```



El AUC, por su parte, proporciona un resumen cuantitativo del rendimiento global del modelo: El AUC es 0.908 (intervalo de confianza: 0.896–0.921), lo que indica un excelente desempeño del modelo. Un AUC cercano a 1 sugiere que el modelo tiene una alta capacidad para diferenciar entre las dos clases. El modelo es robusto y tiene un alto poder discriminativo para predecir la aceptación de depósitos a plazo, siendo útil para la toma de decisiones en el contexto del análisis de clientes bancarios.

Conclusiones finales

El análisis de marketing bancario identificó factores clave que influyen en la aceptación de las campañas. La duración de las llamadas se destacó como determinante del éxito, mientras que un exceso de contactos puede resultar contraproducente, resaltando la importancia de equilibrar calidad y persistencia. Patrones estacionales y demográficos también fueron relevantes: ciertos meses y grupos, como aquellos definidos por educación u ocupación, ofrecen mayores oportunidades de éxito. El modelo de regresión logística mostró un desempeño predictivo sobresaliente ($AUC = 0.908$), validando la utilidad de las variables seleccionadas. Estos hallazgos sugieren priorizar llamadas largas, evitar la saturación de contactos y ajustar las campañas a momentos y segmentos específicos, optimizando así la efectividad del marketing bancario.