Estadística Aplicada. Master Big Data

Miguel Ángel Montero Valero

Febrero 2024

Índice

1. Práctica de evaluación

Instrucciones: Realizar la práctica contestando a las preguntas planteadas en este mismo documento para posteriormente cambiar mi nombre por el tuyo y guardarlo como PracticaEstadística_TuNombre.PDF y adjuntarlo a la entrega.

2. Introducción

En la carpeta de práctica hay un conjunto de datos sobre *Fuga de Clientes* en determinada compañía telefónica, que contiene información sobre las características de los clientes y el estado de fuga. Esta información puede servir a la compañía para construir un modelo predictivo para la fuga de sus clientes antes ante de que ocurra para poder emprender acciones de recaptación.

Previo al modelado predictivo es fundamental un estudio de los datos para la evaluación de su calidad y relaciones en términos estadísticos.

En esta práctica vamos aclarando ciertos comportamientos de los clientes.

Nombre Variable	Descripción
ID	Identificador único de Cliente
Genero	Género binario (Male/Female)
Conyuge	Cliente con conyuge/compañero/compañera
PersCargo	Cliente con personas a su cargo
Antigüedad	Antigüedad en meses del contrato
Telf_serv	Cliente tiene servicios de teléfono contratados
VariasLineas	Cliente tiene varias líneas de teléfono
Int_serv	Cliente tiene servicios de internet contratados
Seguridad	Cliente tiene servicio de seguridad contratado
CopiaSeguridad	Cliente tiene servicio de copia de seguridad contratado
Antivirus_disp	Cliente tiene antivirus en dispositivo
Soporte_tecnico	Cliente tiene soporte técnico contratado
TV_streaming	Cliente tiene TV en streaming contratado
Peliculas	Cliente tiene servicio de películas contratado
Contrato	Tipo de contrato de cliente
Fact_sinPapel	Factura sin papel
MetodoPago	Método de pago
FacturaMes	Importe facturado mensual
FacturaTotal	Importe facturado total
Fuga (objetivo)	Cliente abandona la compañía

Fig1. Descripción de los datos

Se propone el siguiente esquema de estudio.

- Lectura del archivo y comprobación de tipos
- Descripción inicial del archivo
- Estudio univariante
  - Variable continua
  - Variable categórica
- Estudio bivariado
  - Continua-continua
  - Continua-categórica
- Análisis Inferencial
  - Una media
  - Una proporción
  - Diferencia de medias
  - Asociación de variables nominales
- Anova
- Regresión Logística

3. Lectura del archivo y comprobación de tipos

Importa el conjunto **FugaClientes.csv** adecuadamente y realiza la primera inspección del archivo para contestar a las siguientes preguntas:

Código:

```
data <- read.csv("C:/Users/migue/OneDrive/Desktop/FugaClientes.csv",
      sep=",", header=TRUE, stringsAsFactors=FALSE)data <-
read.csv(text=gsub('"', '',
```

```
readLines("C:/Users/migue/OneDrive/Desktop/FugaClientes.csv")), sep=",",
header=TRUE, stringsAsFactors=FALSE)
```

Vemos la estructura de los datos, **código:**

```
str(data)
```

```
data.frame':    3927 obs. of  20 variables:
 $ ID           : chr  "3668-QPYBK" "9763-GRSKD" "8091-TTVAX" "8191-XWSZG" ...
 $ Genero       : chr  "Male" "Male" "Male" "Female" ...
 $ Conyuge      : chr  "No" "Yes" "Yes" "No" ...
 $ PersCargo    : chr  "No" "Yes" "No" "No" ...
 $ Antigüedad   : int  2 13 58 52 10 1 58 30 47 1 ...
 $ Telf_serv    : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ VariasLineas : chr  "No" "No" "Yes" "No" ...
 $ Int_serv     : chr  "DSL" "DSL" "Fiber optic" "No" ...
 $ Seguridad    : chr  "Yes" "Yes" "No" "No" ...
 $ CopiaSeguridad : chr  "Yes" "No" "No" "No" ...
 $ Antivirus_disp : chr  "No" "No" "Yes" "No" ...
 $ Soporte_tecnico: chr  "No" "No" "No" "No" ...
 $ TV_streaming  : chr  "No" "No" "Yes" "No" ...
 $ Peliculas    : chr  "No" "No" "Yes" "No" ...
 $ Contrato     : chr  "Month-to-month" "Month-to-month" "One year" "One year" ...
 $ Fact_sinPapel : chr  "Yes" "Yes" "No" "No" ...
 $ MetodoPago   : chr  "Mailed check" "Mailed check" "Credit card (automatic)"
"Mailed check" ...
 $ FacturaMes   : num  53.9 50 100.3 20.6 55.2 ...
 $ FacturaTotal  : num  108 587 5681 1023 528 ...
 $ Fuga         : int  1 0 0 0 1 1 0 0 1 1 ...
```

Y ahora vamos a ver el resumen de los datos:

```
summary(data)
```

3.1 ¿Cuántas variables tiene el conjunto de datos?

Este conjunto de datos tiene 19 variables.

3.2 ¿Cuántos clientes?

Tiene 3.927 clientes

3.3 Número de variables continuas/categóricas

De variables continuas hay 3 y de variables categóricas hay 16.

3.4 ¿Todas las variables están en formato correcto?

No, hay algunas variables según he comprobado del dataset que no lo están:

- Conyuge
- Género
- PersCargo
- Telf_serv
- VariasLineas
- Int_serv
- Seguridad
- CopiaSeguridad
- Antivirus_disp
- Soporte_tecnico
- TV_streaming
- Peliculas
- Contrato

- Fact_sinPapel
- MetodoPago

3.5 Corrige los tipos que consideres necesario para continuar con el estudio.

Corrigo mediante código los tipos de variables anteriormente mencionados.

4. Descripción inicial del archivo

Llega el momento de estudiar descriptivamente las variables. Inspecciona la distribución con **summary** y responde a las siguientes cuestiones.

```
summary(data)
ID                Genero      Conyuge     PersCargo     Antigüedad     Telf_serv  VariasLinea
 Length:3927        Female:1966   No :2045   No :2744   Min.   : 1.00   No : 382   No :2
 Class :character   Male  :1961   Yes:1882   Yes:1183   1st Qu.: 9.00   Yes:3545   Yes:1
 Mode  :character                                       Median :28.00
                                                        Mean   :32.02
                                                        3rd Qu.:55.00
                                                        Max.   :72.00
       Int_serv    Seguridad  CopiaSeguridad Antivirus_disp Soporte_tecnico TV_streamin
 DSL        :1364   No :2813   No :2596       No :2601       No :2800        No :2432
 Fiber optic:1739   Yes:1114   Yes:1331       Yes:1326       Yes:1127        Yes:1495
 No         : 824


                Contrato    Fact_sinPapel                     MetodoPago     FacturaMes
Fuga
       Month-to-month:2178   No :1595      Bank transfer (automatic): 857   Min.   : 18.
0:2870
       One year      : 830   Yes:2332      Credit card (automatic)  : 873   1st Qu.: 38.
1:1057
       Two year      : 919                 Electronic check         :1331   Median : 70.
```

```
 Mailed check             : 866   Mean   : 64.
                                  3rd Qu.: 89.
                                 Max.   :118.7
```

4.1 ¿Consideras que estamos ante una muestra de clientes balanceada por sexo?

```
table(data$Genero)
```

```
Female   Male
  1966   1961
```

La muestra está bien balanceada, casi hay el mismo número de mujeres que de hombres.

4.2 ¿Cuántos clientes fugados se han recogido?

Código:

```
clientes_fugados <- sum(data$Fuga == 1)

print(paste("Total de clientes fugados:", total_fugados))
```

Total de clientes fugados: 1057

4.3 ¿Cuál es el método más frecuente de pago?

El método más frecuente de pago es Electronic check, que es una forma de pago por internet, donde el dinero se retira eletrónicamente del pagador al beneficiario.

4.3.1 ¿Qué estadístico has utilizado para determinarlo?

He usado el estadístico de frecuencia de cada pago

Código:

```
metodo_pago_mas_frecuente <- names(metodo_mas_frecuente_de_pago)
[which.max(metodo_mas_frecuente_de_pago)]
```

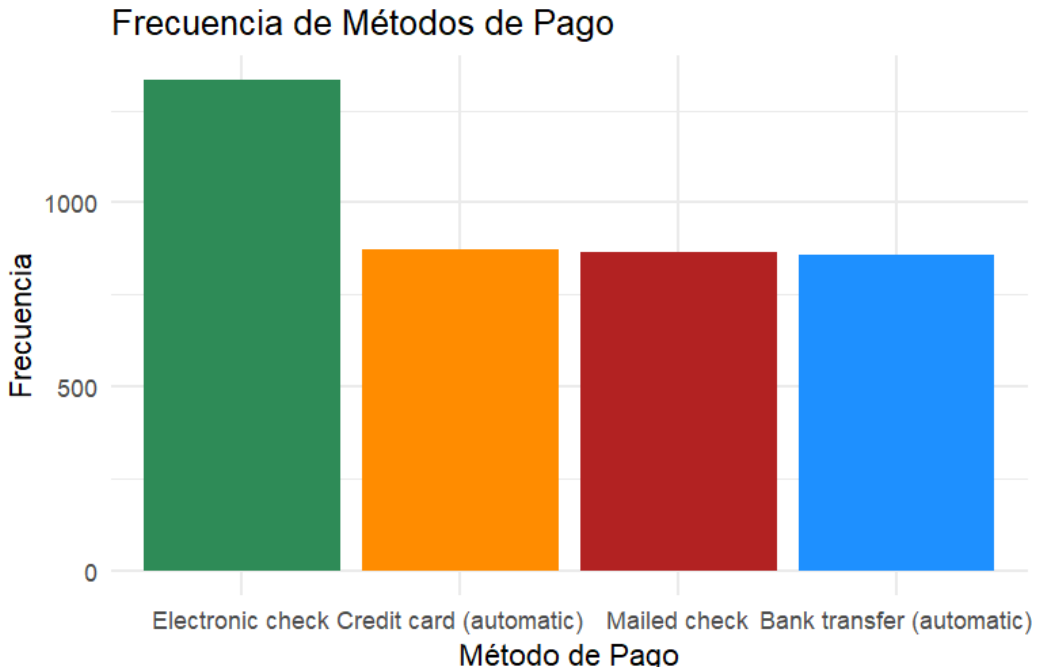
Dataframe para crear un gráfico:

```
metodos_de_pago_df <- as.data.frame(metodo_mas_frecuente_de_pago)
colnames(metodos_de_pago_df) <- c("MetodoPago", "Frecuencia")
```

Código para crear un gráfico de barras basándonos en la frecuencia de los métodos de pago:

```
ggplot(metodos_de_pago_df, aes(x = reorder(MetodoPago, -Frecuencia), y =
Frecuencia, fill = MetodoPago)) + geom_bar(stat = "identity") +
labs(title = "Frecuencia de Métodos de Pago", x = "Método de Pago", y =
"Frecuencia") + theme_minimal() + theme(legend.position = "none") +
scale_fill_manual(values = c("dodgerblue", "darkorange", "seagreen",
"firebrick"))
```

Gráfico:



4.4 ¿Cuál es la media de la variable FacturaTotal?

La media de la variable factura total es 2240.52678889738

4.4.1 ¿Y su mediana?

La mediana de la variable factura total es 1380.1

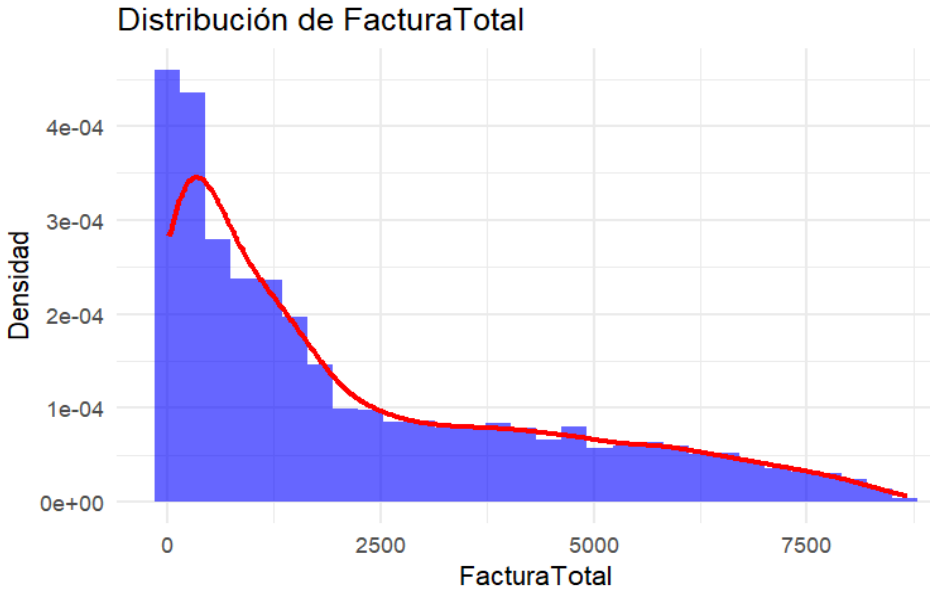
4.4.2 ¿Tiene asimetría?

La asimetría de la variable factura total es 0.98.

Vamos a crear el histograma de FacturaTotal con una curva de densidad para poder visualizar la asimetría:

```
ggplot(data, aes(x = FacturaTotal)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "blue", alpha =
 0.6) +
geom_density(color = "red", size = 1) +
labs(title = "Distribución de FacturaTotal",
x = "FacturaTotal",
y = "Densidad") +
theme_minimal()
```

Gráfico



4.4.3 ¿En qué sentido?

Como vemos en el gráfico, el valor de la asimetría es positivo, lo que indica que la distribución está sesgada a la derecha. Hay

algunos valores que son relativamente altos y estan empujando la cola hacia la derecha.

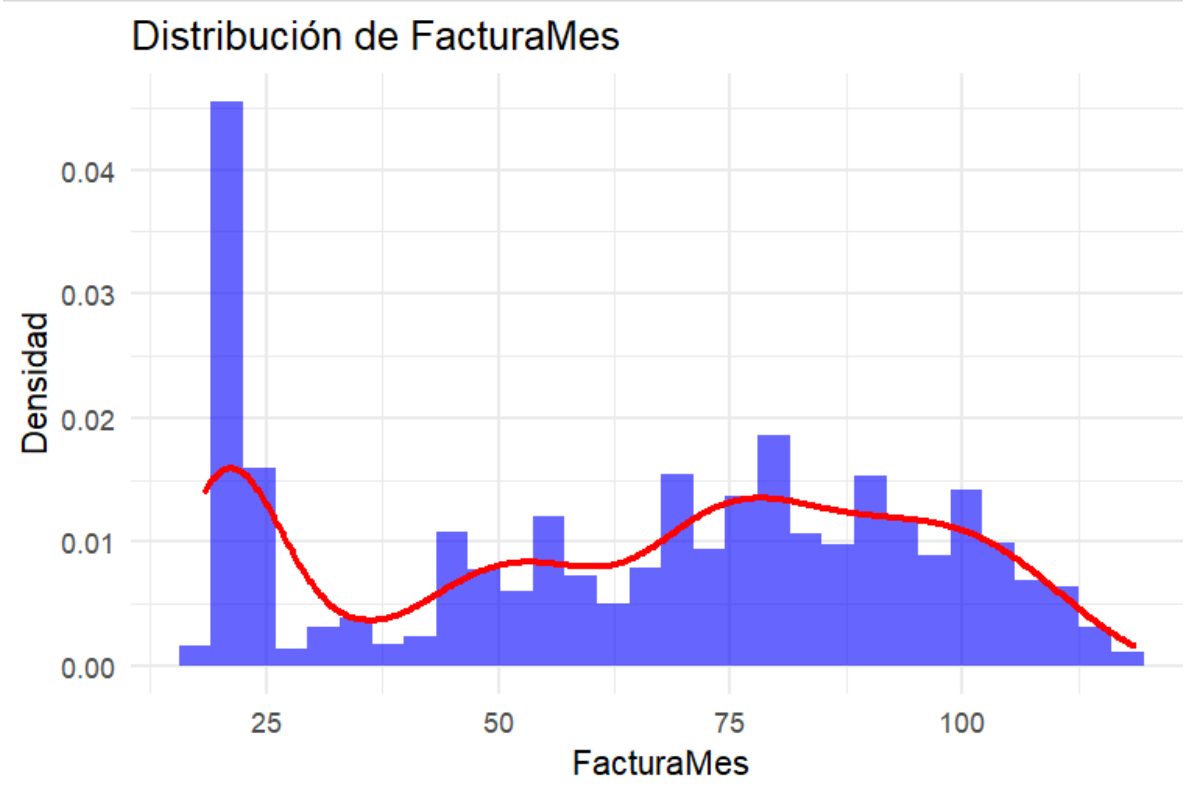
4.5 ¿Y si miramos a FacturaMes?

En el caso de la variable FacturaMes, el valor de la asimetría es negativo, lo que indica que la distribución está sesgada a la izquierda. Esto sugiere que hay algunos valores relativamente bajos que están influyendo en la media.

Vamos a crear el histograma de FacturaMes con una curva de densidad para poder visualizar la asimetría:

```
ggplot(data, aes(x = FacturaMes)) + geom_histogram(aes(y = ..density..), bins = 30, fill = "blue", alpha = 0.6) + geom_density(color = "red", size = 1) + labs(title = "Distribución de FacturaMes", x = "FacturaMes", y = "Densidad") + theme_minimal()
```

Gráfico:



5. Estudio univariante

5.1 Variables numéricas

5.1.1 ¿Cuál es el coeficiente de asimetría de FacturaTotal?

La asimetría de FacturaTotal es: 0.983984519291272

5.1.2 ¿Sería más adecuado interpretar la desviación típica en relación a la Regla Empírica o a la Desigualdad de Chebyshev?

Como el coeficiente de asimetría de la variable FacturaTotal es significativa, casi de un punto, con una distribución bastente sesgada, la interpretación de la desviación típica no sería la más adecuada. Yo me decantaría más por la desigualdad de Chebyshev, que es la más adecuada para cualquier distribución.

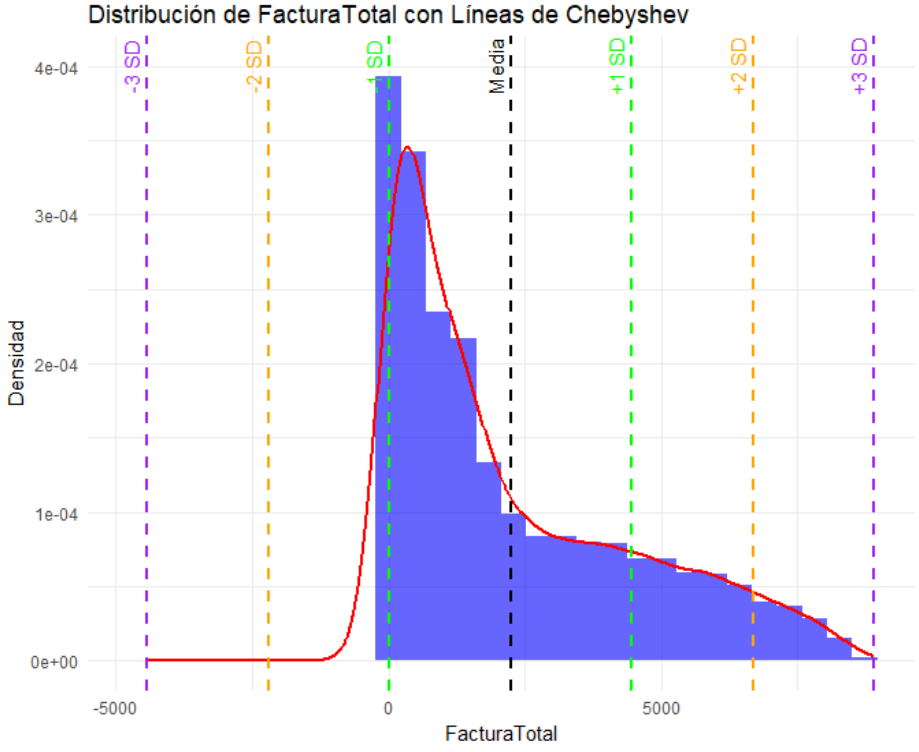
5.1.3 Con el gráfico adecuado para FacturaTotal.

Hacemos histograma de FacturaTotal con una curva de densidad y líneas de Chebyshev

Código:

```
ggplot(data, aes(x = FacturaTotal)) +

geom_histogram(aes(y = ..density..), bins = 30, fill = "blue", alpha =
0.6) +

geom_density(color = "red", size = 1) +

geom_vline(aes(xintercept = mean_factura_total), color = "black",
linetype    = "dashed", size = 1) +

geom_vline(aes(xintercept = mean_factura_total + sd_factura_total), color
= "green", linetype = "dashed", size = 1) +

geom_vline(aes(xintercept = mean_factura_total - sd_factura_total), color
= "green", linetype = "dashed", size = 1) +

geom_vline(aes(xintercept = mean_factura_total + 2 * sd_factura_total),
color = "orange", linetype = "dashed", size = 1) +

geom_vline(aes(xintercept = mean_factura_total - 2 * sd_factura_total),
color = "orange", linetype = "dashed", size = 1) +
```

```
geom_vline(aes(xintercept = mean_factura_total + 3 * sd_factura_total),
color = "purple", linetype = "dashed", size = 1) +

geom_vline(aes(xintercept = mean_factura_total - 3 * sd_factura_total),
color = "purple", linetype = "dashed", size = 1) +

labs(title = "Distribución de FacturaTotal con Líneas de Chebyshev",

x = "FacturaTotal",

y = "Densidad") +

theme_minimal() +

annotate("text", x = mean_factura_total, y = 0.0004, label = "Media",
angle = 90, vjust = -0.5, color = "black") +

annotate("text", x = mean_factura_total + sd_factura_total, y = 0.0004,
label       = "+1 SD", angle = 90, vjust = -0.5, color = "green") +

annotate("text", x = mean_factura_total - sd_factura_total, y = 0.0004,
label       = "-1 SD", angle = 90, vjust = -0.5, color = "green") +

annotate("text", x = mean_factura_total + 2 * sd_factura_total, y =
0.0004, label = "+2 SD", angle = 90, vjust = -0.5, color = "orange") +

annotate("text", x = mean_factura_total - 2 * sd_factura_total, y =
0.0004,     label = "-2 SD", angle = 90, vjust = -0.5, color = "orange")
+

annotate("text", x = mean_factura_total + 3 * sd_factura_total, y =
0.0004, label = "+3 SD", angle = 90, vjust = -0.5, color = "purple") +

annotate("text", x = mean_factura_total - 3 * sd_factura_total, y =
0.0004,     label = "-3 SD", angle = 90, vjust = -0.5, color = "purple")
```



5.1.4 ¿Parece que presenta un comportamiento cercano a la distribución normal?

Como vemos en el gráfico, la distribución no sigue una forma simétrica de campana y hay una larga cola hacia la derecha, que nos está indicando que hay una distribución sesgada positivamente. Además, las líneas de Chebyshev indican que una proporción de datos muy significativa se encuentra lejos de la media, lo que nos viene a indicar que hay una alta variabilidad y dispersión.

5.1.5 ¿Cuál es el p-valor del contraste de Shapiro-Wilk?

Código:

```
shapiro_test <- shapiro.test(data$FacturaTotal)
print(shapiro_test)
```

Resultado:

Shapiro-Wilk normality test

data: data$FacturaTotal

W = 0.86036, p-value < 2.2e-16

5.1.6 ¿Cuál es la conclusión?

El valor que ha arrojado el test de Shapiro-Wilk es muy pequeño, hablamos de 2.2 x $10^{-16,}$= 0.00000000000000022. Esto es un indicativo evidente y fuerte en contra de la hipótesis nula (distribución normal de los datos). Por tanto, nos encontramos ante unos datos que no siguen una distribución normal, así que hablamos de hipótesis alternativa.

5.2 Variables categóricas

Ya que es una de las variables objetivo en el futuro estudio predictivo, merece la pena estudiar la variable **Fuga** mediante una *tabla de frecuencias* que nos indicará la distribución de este factor.

5.2.1 ¿Qué porcentaje de clientes fugados hay en los datos?

```
data$Fuga <- as.numeric(as.character(data$Fuga))

total_clientes <- nrow(data)>

total_fugados <- sum(data$Fuga == 1, na.rm = TRUE)>

 porcentaje_fugados <- (total_fugados / total_clientes) * 100

 print(paste("Porcentaje de clientes fugados:", round(porcentaje_fugados,
2), "%"))
```

Porcentaje de clientes fugados: 26.92%

5.2.2 ¿Cuál es el gráfico adecuado para mostrar esta variable?

Para este tipo de variable, podemos usar un gráfico de barras o un gráfico pastel.

Creo un dataframe con los porcentajes de fuga y el conteo:

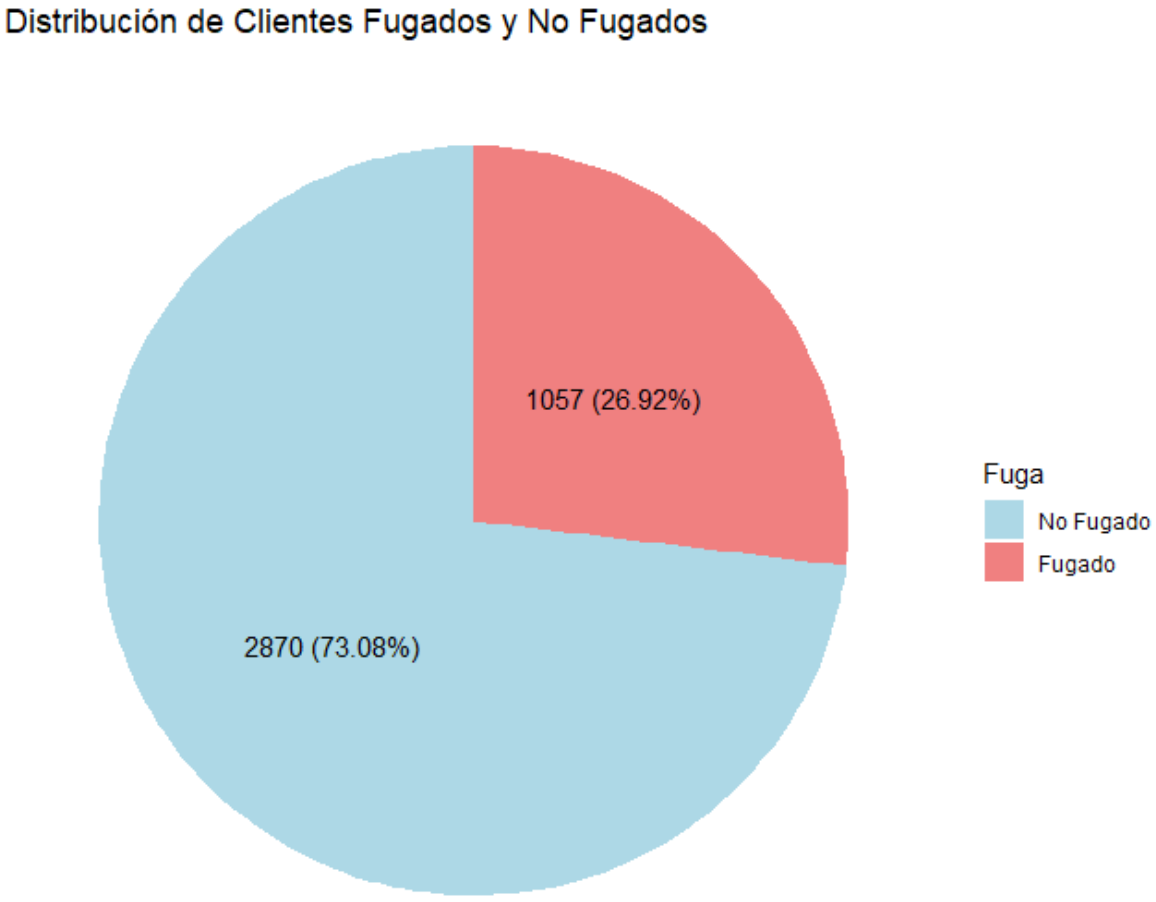
```
frecuencias_fuga_df <- as.data.frame(tabla_frecuencias_fuga)

colnames(frecuencias_fuga_df) <- c("Fuga", "Frecuencia")
```

Calculamos los porcentajes:

```
fuga_por_genero$Percentage <- round(100 * fuga_por_genero$Count /
ave(fuga_por_genero$Count,
fuga_por_genero$Genero, FUN = sum), 1)
```

Creamos el **gráfico** pastel diferenciando entre géneros, optamos por azul y rojo para que resalten más los resultados.

```
ggplot(frecuencias_fuga_df, aes(x = "", y = Frecuencia, fill =
factor(Fuga))) +

geom_bar(width = 1, stat = "identity") +

coord_polar(theta = "y") +

labs(title = "Distribución de Clientes Fugados y No Fugados",

fill = "Fuga",

x = NULL,

y = NULL) +

theme_minimal() +

geom_text(aes(label = paste0(Frecuencia, " (", round((Frecuencia /
sum(Frecuencia)) * 100, 2), "%)")),

position = position_stack(vjust = 0.5), color = "black") +

scale_fill_manual(values = c("lightblue", "lightcoral"), labels = c("No
Fugado", "Fugado")) +

theme(axis.title.x = element_blank(),

axis.title.y = element_blank(),

panel.grid = element_blank(),

axis.ticks = element_blank(),

axis.text.x = element_blank())
```



6. Estudio bivariado

6.1 Continua-Continua

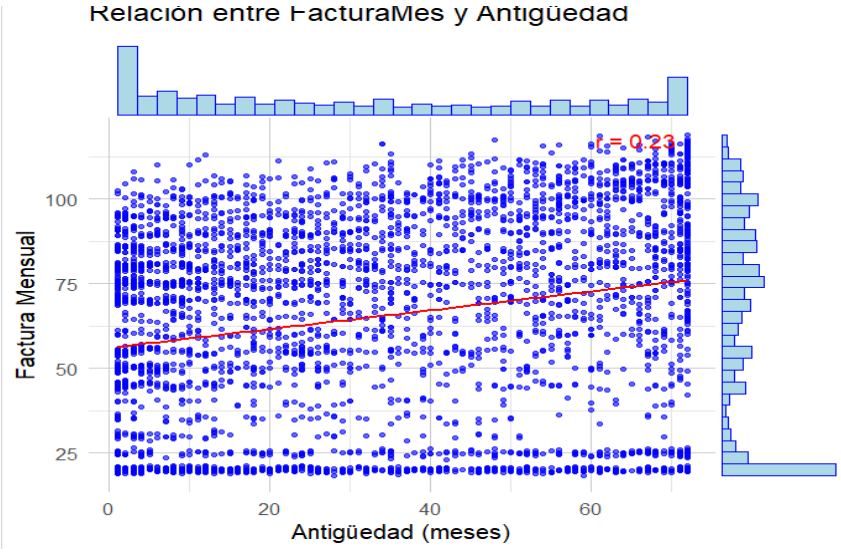
6.1.1 Realiza el gráfico adecuado para el estudio descriptivo de la relación entre FacturaMes y Antiguëdad.

Vamos a visualizar la relación de estas dos variables usando un gráfico de dispersión.

Código:

```
p <- ggplot(data, aes(x = Antigüedad, y = FacturaMes)) +
 geom_point(color = "blue", alpha = 0.6) +
 labs(title = "Relación entre FacturaMes y Antigüedad",
 x = "Antigüedad (meses)",
 y = "Factura Mensual") +
theme_minimal(base_size = 15) +
 geom_smooth(method = "lm", color = "red", se = FALSE) +
 annotate("text", x = Inf, y = Inf, label = paste("r =",
round(coef_pearson, 2)),
 hjust = 1.5, vjust = 2, color = "red", size = 5) +
 theme(panel.grid.major = element_line(color = "grey80"),
 panel.grid.minor = element_line(color = "grey90"))p_marginal ←
 ggMarginal(p, type = "histogram", fill = "lightblue", color =
 "blue"print(p_marginal)
```

15. Gráfico:



6.1.2 ¿Cuál es el coeficiente de correlación lineal de ambas variables?

El coeficiente de correlación lineal (de Pearson) es: 0.226536806808559

El gráfico de dispersión con la línea de regresión y el coeficiente de correlación de Pearson de 0.227, sugieren una relación positiva débil entre FacturaMes y Antigüedad. Esto nos viene a decir que a medida que aumenta la antigüedad del cliente, también aumenta ligeramente la factura mensual.

6.1.3 ¿La relación entre FacturaTotal y FacturaMes es más o menos intensa que la de FacturaTotal con Antigüedad?

Vamos a calcular el coeficiente de correlación lineal entre FacturaTotal y Antigüedad, nos da este resultado: 0.825368599815635

Ahora, vamos a ver gráficamente la combinación de ambas variables en un gráfico de dispersión.

Código:

```
coef_pearson_ft <- cor(data$FacturaMes, data$FacturaTotal, use =
"complete.obs", method = "pearson")

pearson_2 <- ggplot(data, aes(x = FacturaMes, y = FacturaTotal)) +

geom_point(color = "blue", alpha = 0.6) +

labs(title = "Relación entre FacturaMes y FacturaTotal",

 x = "Factura Mensual",

 y = "Factura Total") +

theme_minimal(base_size = 15) +

geom_smooth(method = "lm", color = "red", se = FALSE) +

annotate("text", x = Inf, y = Inf, label = paste("r =",
round(coef_pearson_ft, 2)),

hjust = 1.5, vjust = 2, color = "red", size = 5) +

theme(panel.grid.major = element_line(color = "grey80"),

panel.grid.minor = element_line(color = "grey90"))p_marginal2 <-
ggMarginal(p2, type = "histogram", fill = "lightblue", color =
"blue"print(p_marginal2)
```

Gráfico:

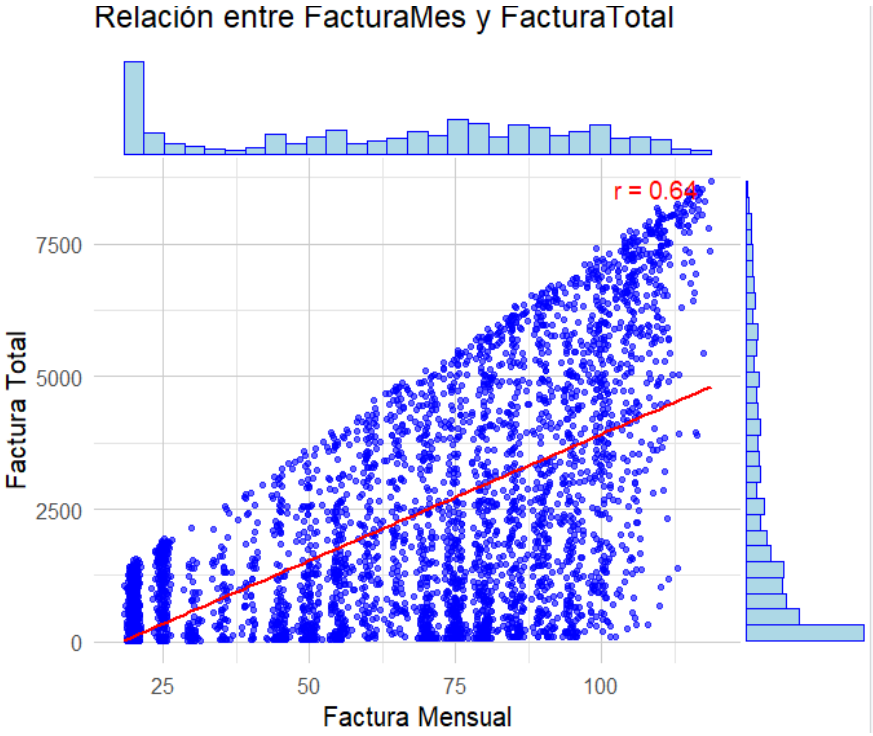


Tabla clasificación coeficientes de correlación de Pearson

Rango de Correlación	Descripción
0.00 - 0.19	Muy débil
0.20 – 0.39	Débil
0.40 – 0.59	Moderada
0.60 – 0.79	Fuerte
0.80 – 1.00	Muy Fuerte

El coeficiente de correlación lineal entre FacturaMes y Antigüedad es 0.227, por lo que sería débil y el coeficiente de correlación lineal entre FacturaTotal y FacturaMes es 0.825, que seria muy fuerte.

6.2 Continua-Categórica

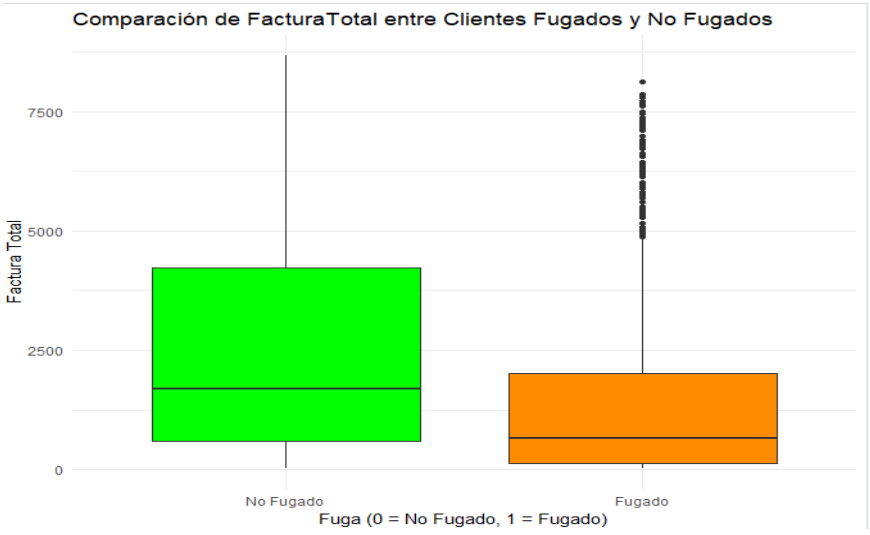
6.2.1 La distribución de la FacturaTotal en los clientes fugados es, en términos generales, superior, inferior o igual que para los clientes no fugados?

Vamos a usar un gráfico box plot para que podamos comparar a simple vista las distribuciones de FacturaTotal entre estos dos grupos de clientes, los fugados y los no fugados.

Código:

```
geom_boxplot(fill = c("green", "green")) +
 abs(title = "Comparación de FacturaTotal entre Clientes Fugados y No
 Fugados",
 x = "Fuga (0 = No Fugado, 1 = Fugado)",
 y = "Factura Total") +
 scale_x_discrete(labels = c("No Fugado", "Fugado")) +
 theme_minimal()
```

Gráfico:



6.2.2 ¿Podrían encontrarse diferencias de facturación entre ambos grupos?

Vamos a realizar el test de Mann-Whitney U para determinar este punto:

Código:

```
test_mann_whitney <- wilcox.test(FacturaTotal ~ Fuga, data = data)
print(test_mann_whitney)
```

Resultado:

Wilcoxon rank sum test with continuity correction

data: FacturaTotal by Fuga

W = 2029509, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

Este test nos da un resultado muy bajo, $2.2 \times 10^{-16} = 0.00000000000000022$, lo que indica que rechazamos la hipótesis nula, lo que significa que hay diferencias bastante significativas en las distribuciones de FacturaTotal entre los clientes fugados y no fugados.

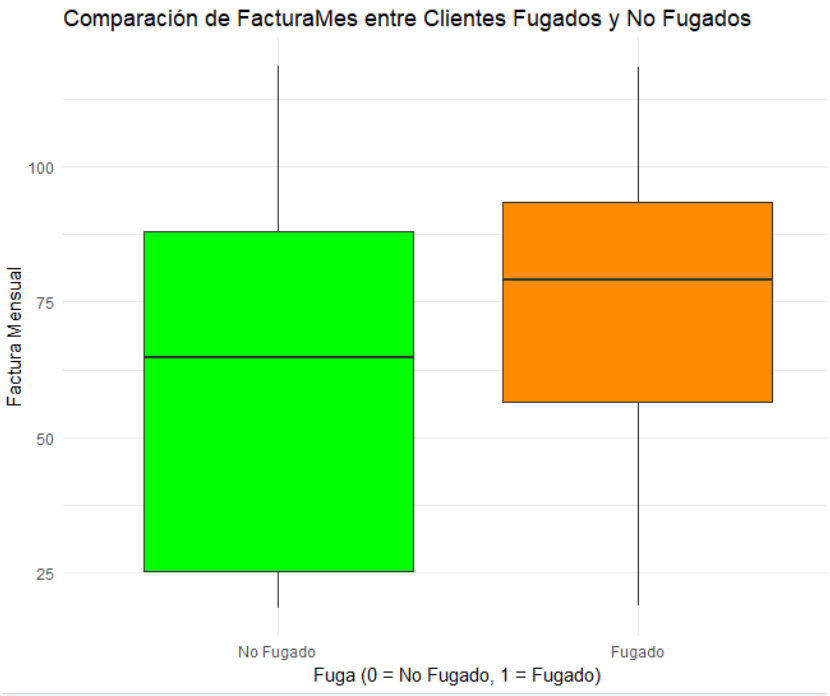
6.2.3 ¿Y para la variable FacturaMes se mantiene el mismo sentido de la relación?

Vamos a usar un gráfico box plot para que podamos comparar a simple vista las distribuciones de FacturaMes entre estos dos grupos de clientes, los fugados y los no fugados.

Código:

```
ggplot(data, aes(x = as.factor(Fuga), y = FacturaMes)) +
geom_boxplot(fill = c("red", "blue")) +
labs(title = "Comparación de FacturaMes entre Clientes Fugados y No
Fugados",
x = "Fuga (0 = No Fugado, 1 = Fugado)",
y = "Factura Mensual") +
scale_x_discrete(labels = c("No Fugado", "Fugado")) +
theme_minimal()
```

Gráfico:



Código:

```
test_mann_whitney_factura_mes <- wilcox.test(FacturaMes ~ Fuga, data =
      data)

print(test_mann_whitney_factura_mes)
```

Resultado:

Wilcoxon rank sum test with continuity correction

data: FacturaMes by Fuga

W = 1159100, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

Este test nos da un resultado muy bajo, $2.2 \times 10^{-16} = 0.00000000000000022$, lo que indica que rechazamos la hipótesis nula, lo que significa que hay diferencias bastante

significativas en las distribuciones de FacturaMes entre los clientes fugados y no fugados.

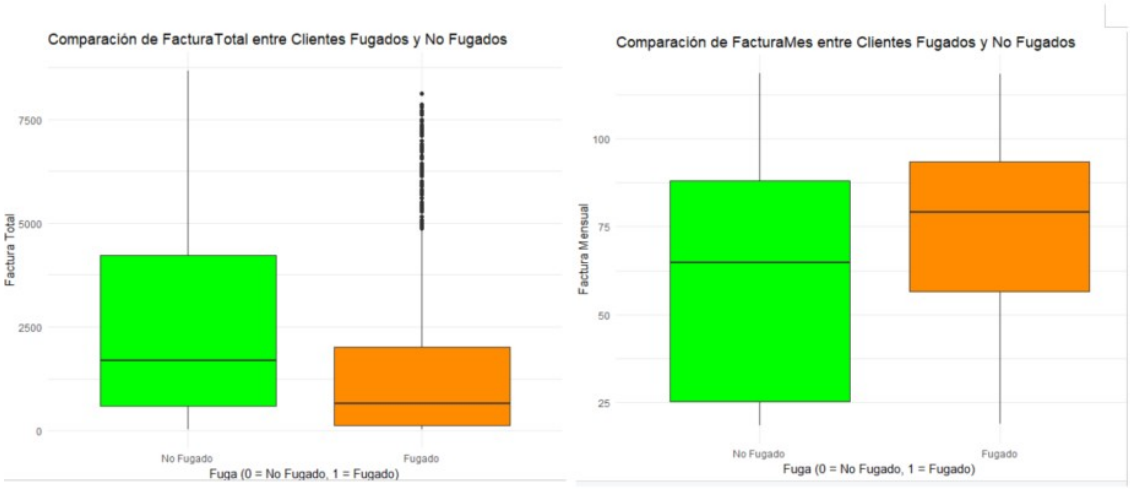


Tabla comparativa según observaciones de los gráficos:

Variable	Grupo	Mediana	Rango intercuartil (IQR)	Dispersión y valores atípicos
FacturaTotal	No fugados	Más alta	Más grande	Menor dispersión, menos valores atípicos
	Fugados	Más baja	Más pequeño	Mayor dispersión, muchos valores atípicos
FacturaMes	No fugados	Más baja	Más grande	Menor dispersión, menos valores atípicos
	Fugados	Más alta	Más pequeño	Menor dispersión, pocos valores atípicos

Según interpreto en los gráficos, en cuanto a FacturaTotal, los clientes fugados tienen más valores que son atípicos y una mediana más baja, sugiriendo que entre esos clientes que se fugan hay, tanto clientes con facturas totales muy altas, como muy bajas. De los clientes no fugados, tienen facturas más altas en promedio, con menor variabilidad y menos valores atípicos.

De FacturaMes, los clientes fugados tienen facturas mensuales más bajas en promedio y menos valores atípicos, mientras que los clientes no fugados tienen facturas mensuales más bajas en promedio, pero con mayor variabilidad dentro del IQR.

Según este análisis, los clientes que se fugan tienen patrones de facturación más variados y extremos en lo que a FacturaTotal se refiere, pero tienen facturas mensuales más altas en promedio en FacturaMes. Estos datos pueden sugerir que los clientes que tienen facturas mensuales más altas tienden a fugarse.

7. Análisis inferencial

7.1 Una media

Supongamos que extraemos una muestra aleatoria de 300 clientes como representativa del conjunto. El objetivo es descubrir relaciones en la muestra y contrastar si encajan con las hipótesis nulas formuladas en cada caso.

La población objetivo será el dataset completo. Veamos los valores de media y varianza.

Código

```
set.seed(123)  # Para reproducibilidad

muestra <- data %>% sample_n(300)

media_muestral <- mean(muestra$FacturaTotal, na.rm = TRUE)

cat("Media muestral de FacturaTotal:", media_muestral, "\n")
```

7.1.1 Contrasta la hipótesis de que la verdadera media poblacional de la FacturaTotal tome el valor de 2600, tomando como conocida la varianza de la población.

Extraemos la muestra de 300 clientes, calculamos la media y la varianza muestral y realizamos el test Z donde:

$\bar{X}$ es la media muestral.

μ_0 es la media poblacional supuesta (2600).

σ es la desviación estándar poblacional.

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

n es el tamaño de la muestra.

Para hacer este test en R, **código:**

```
Z <- (media_muestral - media_poblacional_sup) / (sigma / sqrt(n))
p_value <- 2 * pnorm(-abs(Z))
print(paste("Estadístico Z:", Z))
print(paste("P-valor:", p_value))
```

Resultados:

Estadístico Z: -1.49429468839792"

P-valor: 0.135098612838642"

7.1.2 ¿Cuál es H0?

La hipótesis nula de la media poblacional de FacturaTotal es 2.600

7.1.3 ¿Se encuentran evidencias para rechazarla al 95% de nivel de confianza?

No, ya que el p-valor es mayor que 0.05.

7.2 Repite lo anterior sin suponer conocida la varianza poblacional.

Esta vez, para hacer el mismo análisis pero sin la varianza poblacional, usaremos el test t de Student, donde:

X es la media muestral.

$\mu 0$ es la media poblacional supuesta (2600).

s es la desviación estándar muestral.

n es el tamaño de la muestra.

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Para hacer este test en R, **código:**

```
media_poblacional_sup = 2600 n = 300
s = desviacion_estandar_muestral
t <- (media_muestral - media_poblacional_sup) / (s / sqrt(n))
p_value <- 2 * pt(-abs(t), df = n – 1)
```

```
print(paste("Estadístico t:", t))
print(paste("Intervalo de confianza al 95%: [", ci_lower, ", ", ci_upper, "]",  sep=""))
```

Resultado:

Estadístico t: -1.48696608542477"

P-valor: 0.138077506160964

7.2.1 ¿El resultado es el mismo?

Sí, el resultado es similar, el p-valor en el caso anterior era de 0.135 y en este caso, sin la varianza poblacional, el resultado ha sido de 0.138. Ambos valores son mayores de 0,05, lo que llevan a la misma conclusión de rechado de la hipótesis nula.

7.2.2 ¿Cuál es el intervalo de confianza propuesto?

Código:

```
Alpha <- 0.05 t_crit <- qt(1 - alpha/2, df = n - 1)
ci_lower <- media_muestral - t_crit * (s / sqrt(n))
ci_upper <- media_muestral + t_crit * (s / sqrt(n))
```

Resultado:

Intervalo de confianza al 95%: [2154.32764471803, 2662.04302194864]

7.3 Considerando que la normalidad es importante para este tipo de contrastes paramétricos, ¿Se puede asumir normalidad en la variable FacturaTotal en la muestra?

Para poder responder a esta cuestión sobre la variable FacturaTotal en la muestra, vamos a usar el test de Shapiro-Wilk.

Código:

```
shapiro_test <- shapiro.test(sample_data$FacturaTotal)
print(shapiro_test)
```

Resultado:

Shapiro-Wilk normality test

data: sample_data$FacturaTotal

W = 0.8861, p-value = 3.553e-14

7.3.1 ¿Qué contraste alternativo podría ser más adecuado?

Como alternativa y dada la distribución de los datos, el método adecuado sería el test de Wilcoxon Signed-Rank, para comparar la mediana con el valor supuesto de 2.600.

Test, código:

```
test_wilcoxon <- wilcox.test(sample_data$FacturaTotal, mu = 2600)
print(wilcoxon_test)
```

Resultado:

Wilcoxon signed rank test with continuity correction

data: sample_data$FacturaTotal

V = 18871, p-value = 0.01378

alternative hypothesis: true location is not equal to 2600

7.3.2 ¿Cuáles son las conclusiones?

El P-valor: 0.01378 y es menor que el nivel de Significancia: 0.05, rechazamos la hipótesis nula. Hay evidencias bastante significativas para concluir que la mediana poblacional de FacturaTotal no es 2.600.

Para visualizar mejor estos datos, vamos a sacar un histograma y un gráfico que compare los cuartiles (gráfico Q-Q)

Histograma, **código:**

```
ggplot(sample_data, aes(x = FacturaTotal)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "blue", alpha =
0.6) +
geom_density(color = "red", size = 1) +
labs(title = "Distribución de FacturaTotal",
x = "FacturaTotal",
```

```
y = "Densidad") +
theme_minimal()
```

Resultado:

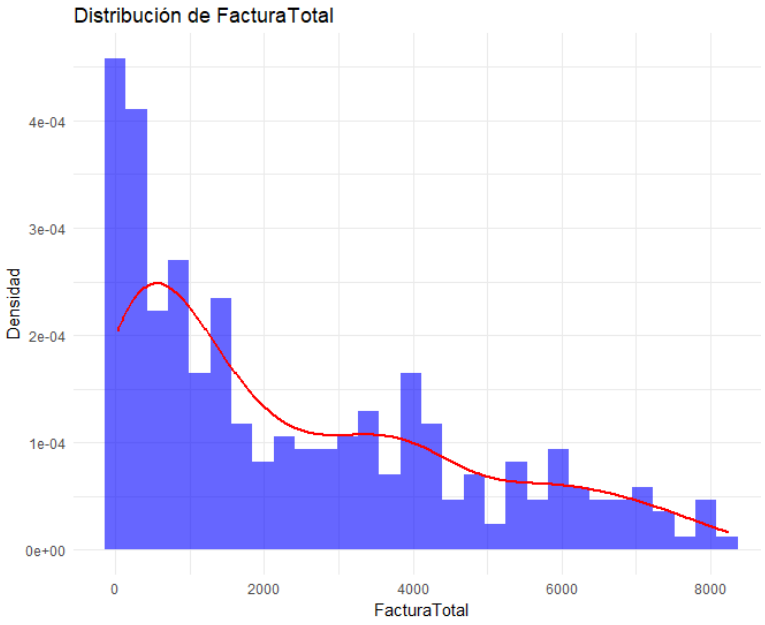
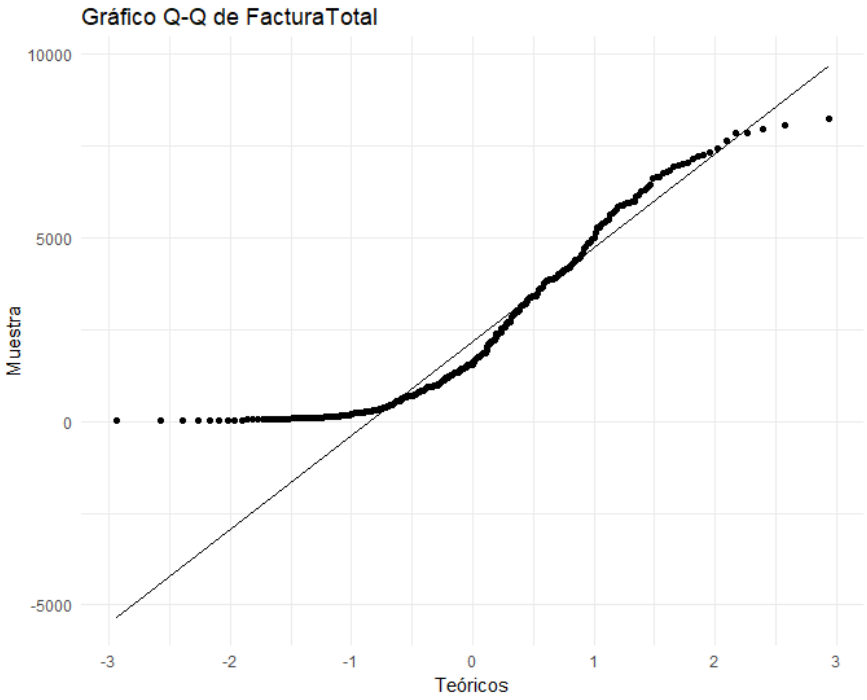


Gráfico Q-Q, **código:**

```
ggplot(sample_data, aes(sample = FacturaTotal)) +
stat_qq() +
stat_qq_line() +
labs(title = "Gráfico Q-Q de FacturaTotal",
 x = "Teóricos",
 y = "Muestra") +
 theme_minimal()
```

Resultado:



En el gráfico Q-Q, comparando los cuantiles observados en FacturaTotal, lo s puntos se desvían de forma significativa de la línea diagonal, sobre todo en los extremos, indicando que los datos no siguen una distribución normal.

7.2 Una proporción

7.2.1 Contrasta la hipótesis de que la proporción de clientes fugados es mayor que el 30%. Utiliza la distribución exacta.

Código:

```
total_clientes <- nrow(data)
total_fugados <- sum(data$Fuga == 1, na.rm = TRUE)
prop_fugados <- total_fugados / total_clientes



test_binomial <- binom.test(total_fugados, total_clientes, p = 0.30,
alternative = "greater")
test_binomial
```

Resultado:

Exact binomial test

data: total_fugados and total_clientes

number of successes = 1057, number of trials = 3927, p-value = 1

alternative hypothesis: true probability of success is greater than 0.3

95 percent confidence interval:

0.257523 1.000000

sample estimates:

probability of success

0.2691622

7.2.2 ¿Cuál es el p-valor?

```
> test_binomial$p.value
```

Resultado:

[1] 0.9999905

7.2.3 ¿Se puede rechazar H0?

Código:

```
if (test_binomial$p.value < 0.05) {
+     resultado <- "Se rechaza H0"
+ } else {
+     resultado <- "No se rechaza H0"
+ }
> resultado
```

[1] "No se rechaza H0"

7.2.4 ¿Cuál es el intervalo de confianza al 99% para la proporción de clientes fugados según el test asintótico?

Código:

```
prop_test <- prop.test(total_fugados, total_clientes, conf.level = 0.99)
prop_test$conf.int
```

Resultado:

[1] 0.2512077 0.2879009, **este es el intervalo de confianza.**

attr(,"conf.level")

[1] 0.99

7.3 Dos muestras. Diferencia de medias.

7.3.1 Realiza el contraste paramétrico adecuado para evaluar asociación entre las variables FacturaTotal y Fuga, asumiendo varianzas iguales.

Código:

```
t_test <- t.test(FacturaTotal ~ Fuga, data = data, var.equal = TRUE)
t_test$p.value
```

Resultado:

[1] 7.917043e-46 P-valor obtenido: 7.917043×10^{-46}

7.3.2 ¿Se puede decir que las medias de facturación de ambos grupos son similares?

No, hay una diferencia significativa entre las medias de la FacturacionTotal y clientes fugados y no fugados.

7.3.3 ¿Se puede asumir la hipótesis de homocedasticidad de FacturaTotal en ambos grupos de Fuga?

Código:

```
levene_test <- leveneTest(FacturaTotal ~ Fuga, data = data)

levene_test

p_value_levene <- levene_test$`Pr(>F)`[1]

p_value_levene
```

Resultado:

[1] 1.024562e-28; El p-valor del test de Levene es muy bajo, 1.024562×10^{-28}, por lo que no se puede asumir homocedasticidad.

7.3.4 En caso contrario, ¿qué dice en test no paramétrico sobre la relación?

Código:

```
wilcox_test <- wilcox.test(FacturaTotal ~ Fuga, data = data)

wilcox_test$p.value
```

Resultado:

[1] 1.598177e-59; El p-valor del test no paramétrico de Wilcoxon es también muy bajo, $1.598177 \times 10^{-59,}$ por lo que está indicando que hay una diferencia significativa de las distribuciones de la facturación total entre los clientes fugados y no fugados.

7.4 Asociación de variables nominales

7.4.1 Realiza un test χ^2 para la asociación de entre las variables Fuga y Método de pago.

Código:

```
tabla_contingencia <- table(data$Fuga, data$MetodoPago)

chi_test <- chisq.test(tabla_contingencia)

chi_test
```

Resultado:

earson's Chi-squared test

data: tabla_contingencia

X-squared = 349.23, df = 3, p-value < 2.2e-16

Código para acceder a p-valor:

```
p_value_chi <- chi_test$p.value

p_value_chi
```

Resultado:

[1] 2.185189e-75; p-valor del test de chi-cuadrado 2,185189 x 10^{-75}.

7.4.2 ¿Hay razones para pensar que existe un patrón de asociación?

Sí, al descartar la hipótesis nula dado que el p-valor es mucho más bajo que 0.05, nos indica que hay ua asociación significativa entre las variables de fuga y método de pago.

7.4.3 ¿Cuál es la casilla con un residuo estandarizado positivo mayor?

Código:

```
residuos_estandarizados <- chi_test$stdres

residuos_estandarizados
```

Resultado:

	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check
0	7.811065	8.910833	-18.451089	4.347225
1	-7.811065	-8.910833	18.451089	-4.347225

Código para indentificar la casilla con residuo positivo:

```
max_residuo <- max(residuos_estandarizados)
max_residuo
```

Resultado:

[1] 18.45109, la casilla con el residuo estandarizado positivo mayor corresponde a la de los clientes fugados y que utilizanel método de pago "Electronic check".

7.4.4 ¿Qué significa?

El resultado sugiere que el método de pago de cheque electrónico tiene una fuerte asociación con los clientes que se fugan.

8. ANOVA

Toma una muestra de tamaño 100 por cada tipo de contrato con la semilla 1234.

8.1 ¿Se puede rechazar la hipótesis de misma facturación media en los clientes de los tres tipos de contrato?

Código:

```
levene_test_anova <- leveneTest(FacturaTotal ~ Contrato, data = muestra_anova)

levene_test_anova

summary(anova_result)
```

Resultado:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)

Contrato 2 3.211e+08 160556804 38.34 1.54e-15 ***

Residuals 297 1.244e+09 4187605

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

El p-valor obtenido es 1.54 x 10^{-15}, es mucho menor que 0.05, **por lo que la hipótesis nula queda descartada**. Hay diferencias significativas en la facturación media entre los diferentes tipos de contrato.

8.1.2 ¿Se pueden suponer varianzas iguales al 95% de confianza?

Código:

```
p_value_levene <- levene_test_anova$`Pr(>F)`[1]

p_value_levene
```

Resultado:

[1] 1.233557e-13, el p-valor obtenido con el test de Levene es 1.233557 s 10^{-13}. **Como el valor es mucho menor que 0.05, no se pueden suponer varianzas iguales.**

8.1.3 ¿Se cumple la hipótesis de normalidad?

Código:

```
shapiro_test_anova <- shapiro.test(residuals(anova_result))
shapiro_test_anova$p.value
```

Resultado:

[1] 2.112888e-08, el p-valor obtenido con el test de Shapiro es 2.112888 x 10^{-8}, por lo que al ser más bajo que 0.05, indica que no sigue una distribución normal y **no cumple la hipótesis de normalidad.**

8.2 Si consideras que las hipótesis no se cumplen. Aplica un test no paramétrico adecuado para la evaluación de la relación del anterior apartado.

Código:

```
kruskal_test <- kruskal.test(FacturaTotal ~ Contrato, data = muestra_anova)
kruskal_test$p.value
```

Resultado:

[1] 7.888066e-17, el p-valor obtenido con el test de Krustal es 7.888066 x 10^{-17}. El valor obtenido es mucho menor que 0.05, lo que indica que hay diferencias significativas en la facturación media entre los diferentes tipos de contrato.

8.2.1 ¿Cuál es la conclusión?

Según las pruebas realizadas, las suposiciones del ANOVA no se cumplen. Al utilizar el test de Krustal-Wallis, éste muestra también que existe una diferencia significativa entre los diferentes tipos de contrato y la facturación media.

9. Regresión Logística

Se toma una muestra balanceada por la variable fuga, de 200 clientes de cada grupo con la semilla 2345.

Código para sacar la muestra:

```
set.seed(2345)
muestra_logistica <- data %>% group_by(Fuga) %>% sample_n(200)
```

9.1 Ajusta un modelo de regresión logística para la clasificación de clientes fugados en relación exclusivamente a su antigüedad.

Código:

```
modelo_logistico <- glm(Fuga ~ Antigüedad, data = muestra_logistica,
family = binomial)
summary(modelo_logistico)
```

Resultado:

Call:

glm(formula = Fuga ~ Antigüedad, family = binomial, data = muestra_logistica)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.255559	0.170427	7.367	1.74e-13	***
Antigüedad	-0.048886	0.005449	-8.971	< 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: 554.52 on 399 degrees of freedom

Residual deviance: 445.15 on 398 degrees of freedom

AIC: 449.15

Number of Fisher Scoring iterations: 4

9.1.1 ¿Es significativo este efecto en el modelo?

El p-valor para antigüedad es < 2e-16, así que **es muy significativo** al ser <0.001.

9.1.2 ¿Cuál es el sentido de influencia sobre la probabilidad de Fuga?

Código:

```
exp(coef(modelo_logistico))
```

Resultado:

(Intercept) Antigüedad

3.5097984 0.9522898

Según los datos de intercept, con un odds ratio de 3.5097984, los clientes con antigüedad de 0 tienen 3.509 veces más probabilidades de fugarse que de no fugarse.

Según el odds ratio de antigüedad, que es de 0.9522898, nos refleja que por cada año que el cliente permanece en la compañía, las probabilidades de fugarse disminuyen en un factor de 0.952. Así que, a mayor antigüedad, menos posibilidades de perder al cliente.

9.2 ¿Cuál es la capacidad predictiva del modelo dada por la tasa de aciertos o Accuracy en una muestra de test de 100 clientes por grupo tomada con la semilla 2345?

Código para la predicción:

```
set.seed(2345)
muestra_test <- data %>% group_by(Fuga) %>% sample_n(100)
predicciones <- predict(modelo_logistico, newdata = muestra_test, type =
"response")
predicciones_binarias <- ifelse(predicciones > 0.5, 1, 0)
accuracy <- mean(predicciones_binarias == muestra_test$Fuga)
accuracy
```

Resultado:

[1] 0.67

Según este resultado, **el modelo clasificó correctamente el 67% de los casos** en la muestra de test.