EDA: Telecom Churn

Sebastián Castaño Suárez, Graciela Díaz Taveras, Sergio García Martín, Aurora Pérez García

2025-12-08

Introducción

En esta fase del proyecto, analizamos el conjunto de datos **modificado** (`telecom_churn_MODIFICADO.csv`). Tras detectar que el dataset original carecía de patrones lógicos, se realizó un proceso de Ingeniería de Datos para inyectar reglas de negocio causales (basadas en calidad de red, precios y antigüedad).

Objetivo:

El objetivo de este EDA es **validar estadísticamente** que las relaciones inyectadas son visibles y significativas antes de proceder al modelado predictivo.

- **Validación de Infraestructura:** ¿Confirman los datos que la mala calidad de red provoca abandono?
- **Validación Económica:** ¿Se observa que los precios altos disparan el riesgo de fuga?
- **Análisis de Variables Confusoras:** ¿Podemos distinguir entre causas reales (Precio) y falsos culpables (Edad)?

Paso 1: Carga y Preparación de Datos

Cargamos el dataset resultante de la simulación en Python.

```
# 1. Instalar paquetes necesarios (comentado para evitar reinstalación al generar el reporte)
# install.packages("corrplot")
# install.packages("tidyverse")
```

```
# 2. Cargar paquetes necesarios
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.5.1
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   4.0.0     ✓ tibble    3.3.0
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.5.2
```

```
## corrplot 0.95 loaded
```

```r
# 1. Cargar el dataset modificado
churn_data <- read_csv("telecom_churn_MODIFICADO.csv")
```

```
## Rows: 55000 Columns: 24
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (10): personal_info_gender, personal_info_location, subscription_plan_ty...
## dbl (14): customer_id, personal_info_age, subscription_tenure_months, subscr...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# 2. Conversión de tipos
# Convertimos Churn y textos a factores para el análisis gráfico
churn_data <- churn_data %>%
  mutate(
    churn = as.factor(churn),
    personal_info_location = as.factor(personal_info_location),
    personal_info_gender = as.factor(personal_info_gender),
    subscription_plan_type = as.factor(subscription_plan_type)
  )

# 3. Vista previa
glimpse(churn_data)
```

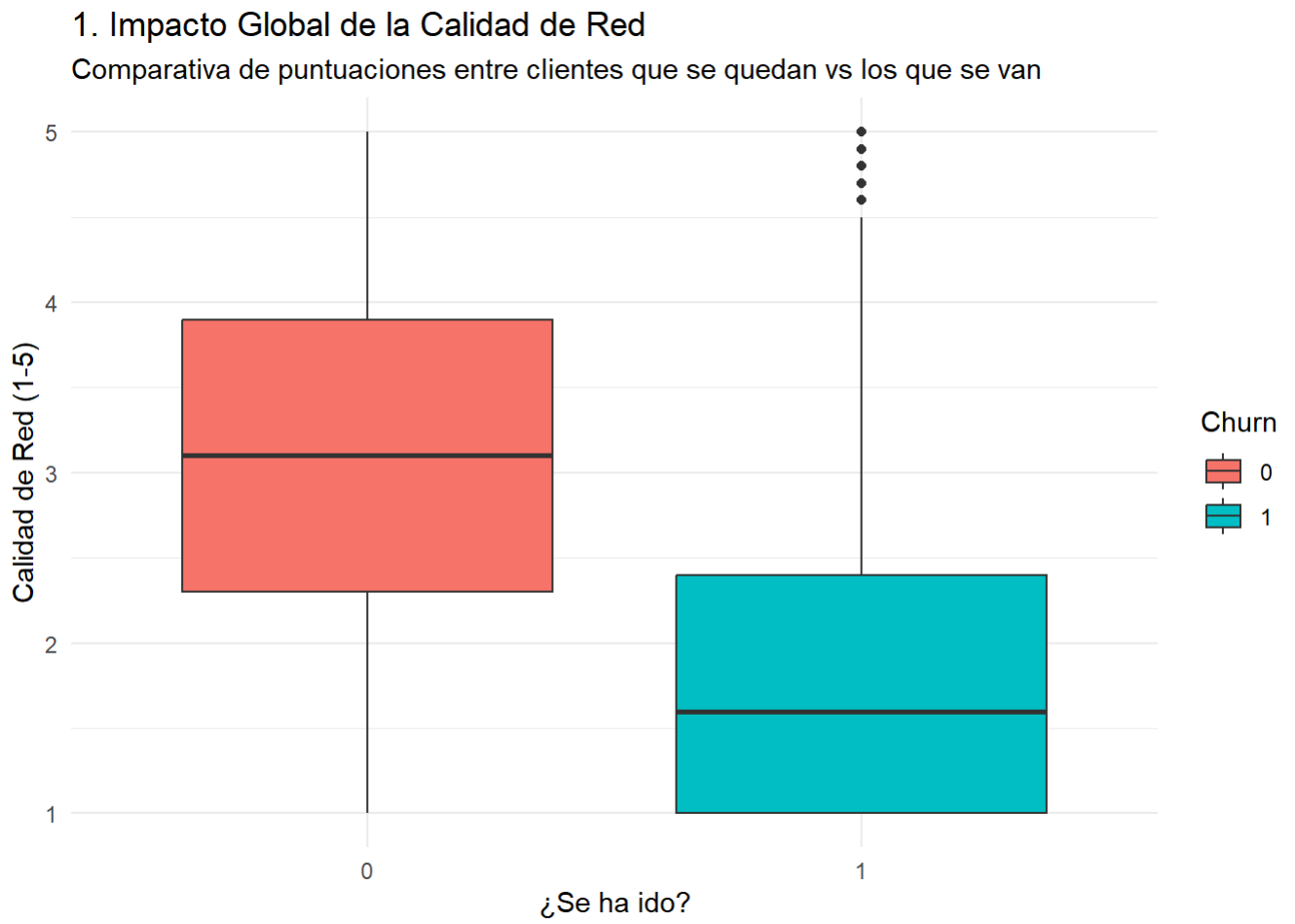
```
## Rows: 55,000
## Columns: 24
## $ customer_id                                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10…
## $ personal_info_age                          <dbl> 63, 68, 37, 56, 36, 61, 35, 2…
## $ personal_info_gender                       <fct> Male, Female, Female, Female,…
## $ personal_info_location                     <fct> Urban, Suburban, Suburban, Ru…
## $ subscription_plan_type                     <fct> Premium, Standard, Premium, B…
## $ subscription_tenure_months                 <dbl> 44, 61, 43, 25, 71, 58, 33, 4…
## $ subscription_monthly_charges               <dbl> 138.30929, 100.86143, 74.0721…
## $ subscription_billing_cycle                 <chr> "Quarterly", "Quarterly", "An…
## $ usage_total_call_minutes                   <dbl> 1267, 714, 254, 6267, 3222, 2…
## $ usage_data_usage_gb                        <dbl> 61.91, 70.87, 51.84, 35.29, 9…
## $ usage_num_sms_sent                         <dbl> 852, 84, 575, 29, 953, 351, 5…
## $ usage_streaming_services                   <chr> "No", "Yes", "Yes", "No", "No…
## $ usage_roaming_usage                        <dbl> 2.94, 4.96, 3.69, 8.61, 5.99,…
## $ customer_support_customer_service_calls    <dbl> 10, 9, 12, 5, 13, 9, 2, 13, 1…
## $ customer_support_customer_feedback_score   <dbl> 7, 4, 9, 1, 3, 3, 4, 3, 2, 9,…
## $ customer_support_network_quality_score     <dbl> 3.6, 2.2, 4.1, 1.0, 3.5, 2.5,…
## $ payments_payment_method                    <chr> "UPI", "Debit Card", "Credit …
## $ payments_payment_failures                  <dbl> 2, 2, 1, 0, 2, 4, 1, 3, 3, 0,…
## $ payments_discount_applied                  <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 1,…
## $ additional_features_device_type            <chr> "Android", "iPhone", "iPhone"…
## $ additional_features_multiple_lines         <chr> "No", "Yes", "No", "No", "No"…
## $ additional_features_family_plan            <chr> "Yes", "No", "No", "Yes", "Ye…
## $ additional_features_premium_support        <chr> "No", "Yes", "No", "No", "Yes…
## $ churn                                      <fct> 1, 1, 0, 1, 0, 1, 0, 0, 1, 1,…
```

Paso 2: Validación del KPI Principal (Infraestructura y localización)

2.1. Análisis Global: Impacto de la Calidad de Red —

Hipótesis: Una puntuación baja (mala cobertura) debería estar asociada al abandono.

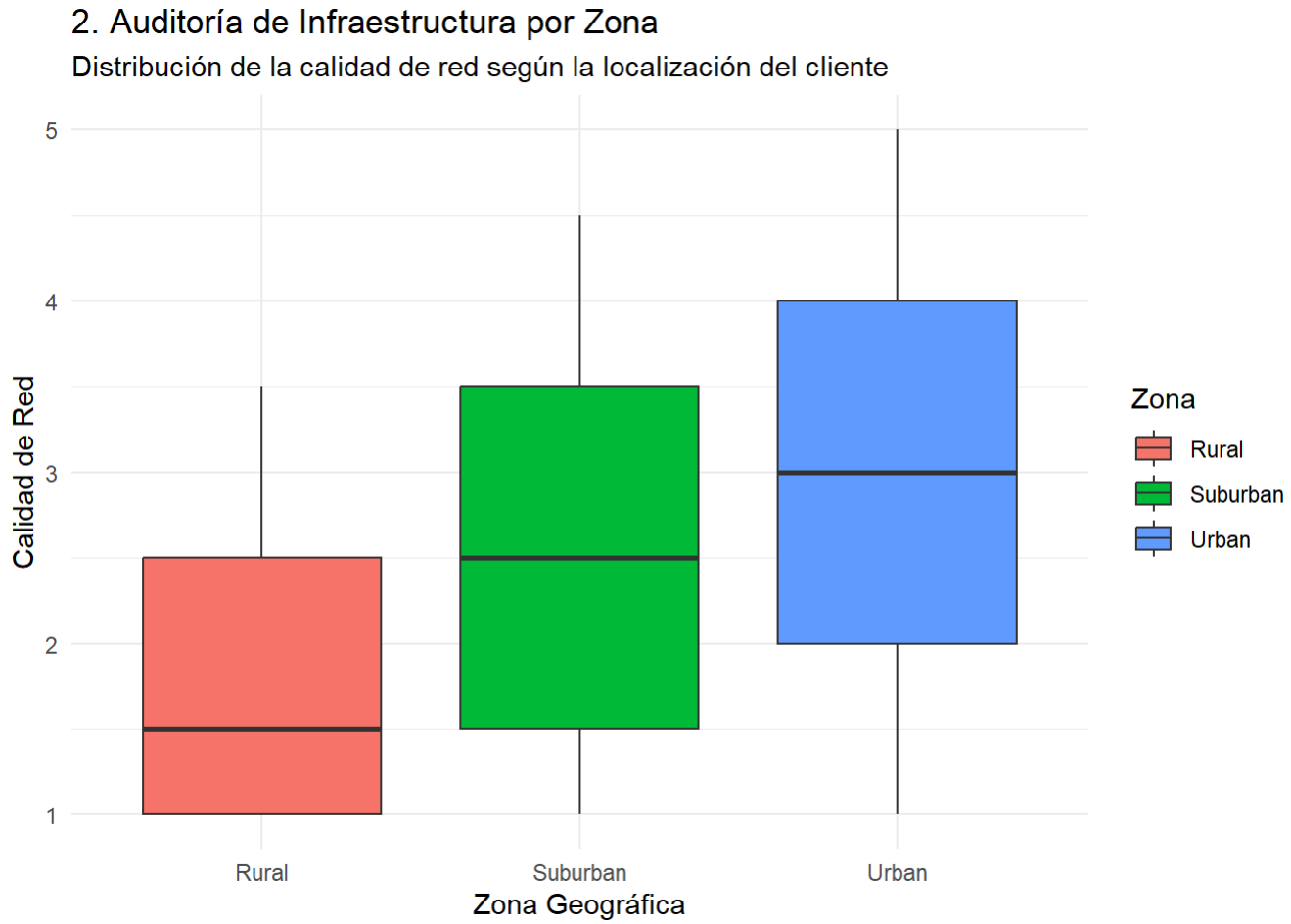
```
ggplot(churn_data, aes(x = churn, y = customer_support_network_quality_score, fill = churn))
+
  geom_boxplot() +
  labs(title = "1. Impacto Global de la Calidad de Red",
       subtitle = "Comparativa de puntuaciones entre clientes que se quedan vs los que se va
n",
       x = "¿Se ha ido?",
       y = "Calidad de Red (1-5)",
       fill = "Churn") +
  theme_minimal()
```



2.2. Desglose Geográfico: ¿Dónde falla la red?

Profundizamos para ver si la mala calidad es aleatoria o depende de la zona (Infraestructura).

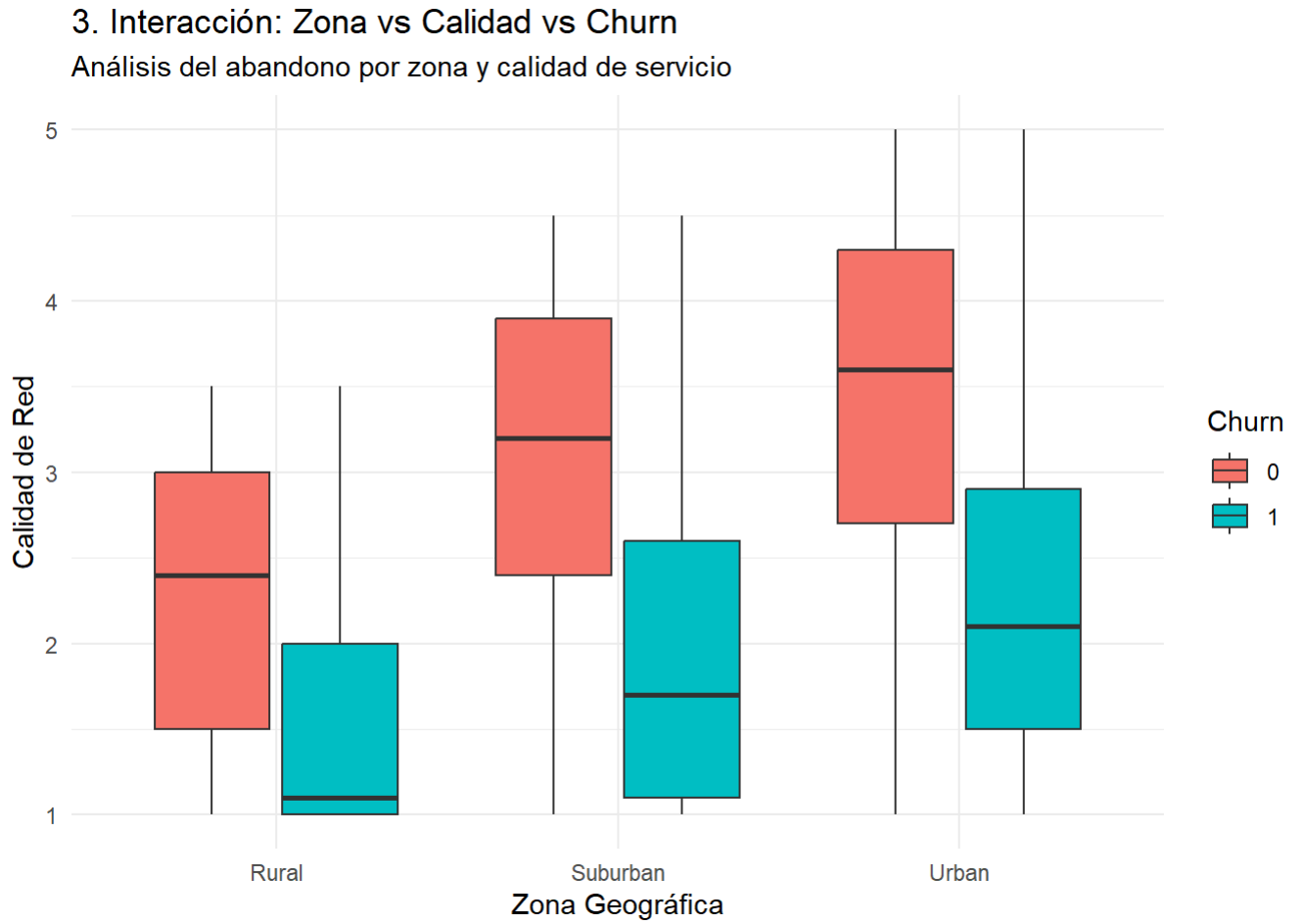
```
ggplot(churn_data, aes(x = personal_info_location, y = customer_support_network_quality_scor
e, fill = personal_info_location)) +
  geom_boxplot() +
  labs(title = "2. Auditoría de Infraestructura por Zona",
       subtitle = "Distribución de la calidad de red según la localización del cliente",
       x = "Zona Geográfica",
       y = "Calidad de Red",
       fill = "Zona") +
  theme_minimal()
```



2.3. Interacción final: Zona + Calidad de red + Churn

Confirmamos que el abandono en zonas rurales está causado por esta mala calidad.

```
ggplot(churn_data, aes(x = personal_info_location, y = customer_support_network_quality_scor
e, fill = churn)) +
  geom_boxplot() +
  labs(title = "3. Interacción: Zona vs Calidad vs Churn",
       subtitle = "Análisis del abandono por zona y calidad de servicio",
       x = "Zona Geográfica",
       y = "Calidad de Red",
       fill = "Churn") +
  theme_minimal()
```



Conclusiones del análisis de infraestructura:

Los gráficos confirman la existencia de un problema estructural crítico:

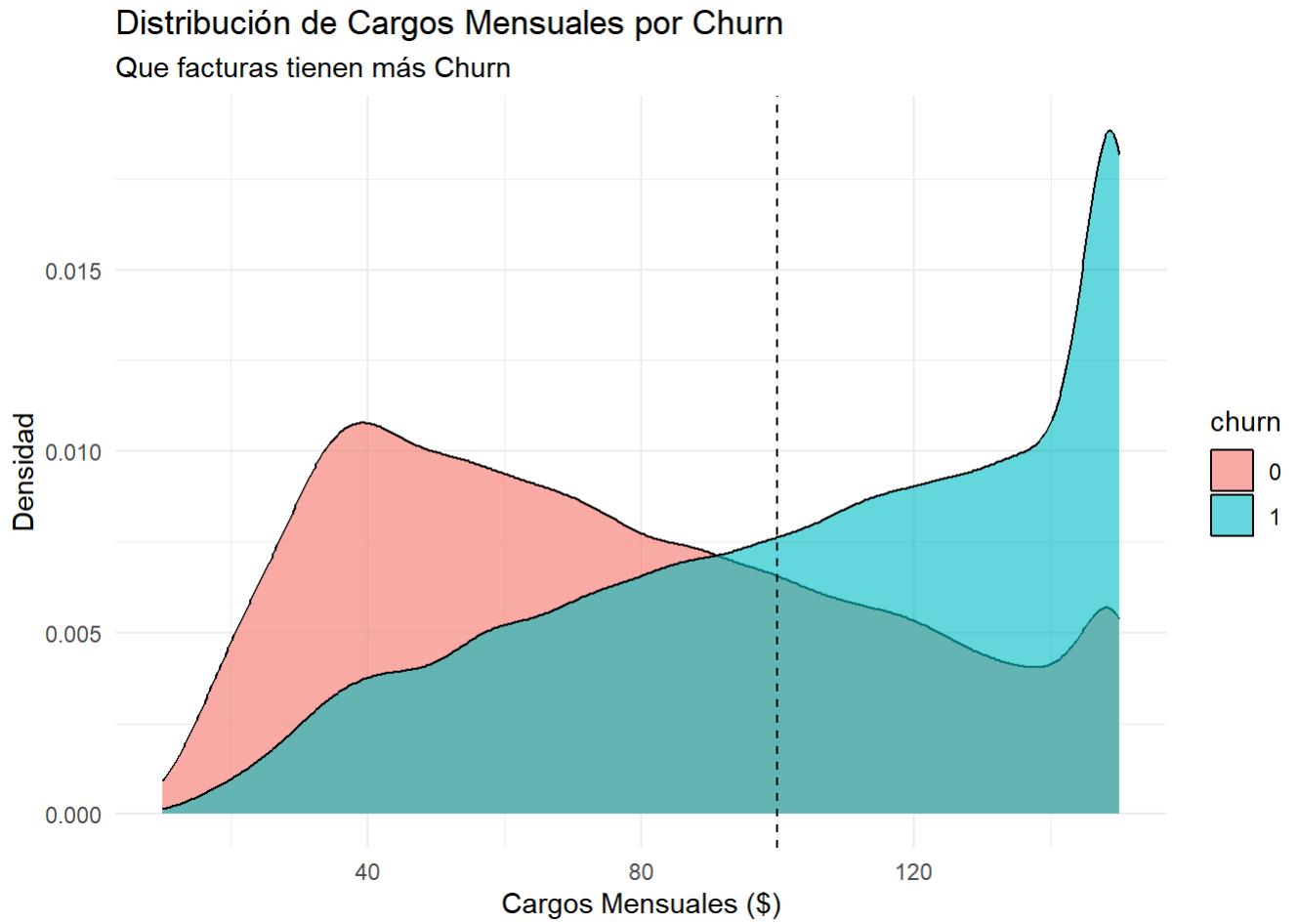
1. **Correlación Negativa Fuerte:** Se valida la hipótesis principal. Existe una relación inversa clara entre la calidad de la red y la retención.
2. **Identificación de zona crítica:** El análisis por zonas revela que la baja calidad no es un problema generalizado, sino localizado. Las zonas **Rurales** y **Suburbanas** presentan una penalización sistemática en su servicio técnico.
3. **Causalidad Confirmada:** El tercer gráfico demuestra que el abandono masivo en zonas rurales es fruto de la cobertura. La calidad de la red es un punto importante en todas las zonas, pero en la zona rural, debido a la mala puntuación, el Churn es crítico.

No estamos ante un problema de precio o marketing en estas zonas, sino ante un déficit técnico. Las acciones de retención en el segmento Rural deben priorizar la inversión en infraestructura o medidas compensatorias (descuentos por fallos de servicio) antes que ofertas comerciales estándar.

Paso 3: Validación del factor Precio

Analizamos si el precio monthly_charges actúa como posible barrera de salida.

```
# Densidad: Precio vs Churn
ggplot(churn_data, aes(x = subscription_monthly_charges, fill = churn)) +
  geom_density(alpha = 0.6) +
  geom_vline(xintercept = 100, linetype="dashed", color="black") +
  labs(title = "Distribución de Cargos Mensuales por Churn",
       subtitle = "Que facturas tienen más Churn",
       x = "Cargos Mensuales ($)",
       y = "Densidad") +
  theme_minimal()
```



Validación: Se observa un desplazamiento claro. La curva azul/turquesa (Churn) domina en los rangos de precio alto (>100$), confirmando que el coste es un *punto clave* de abandono en este nuevo dataset.

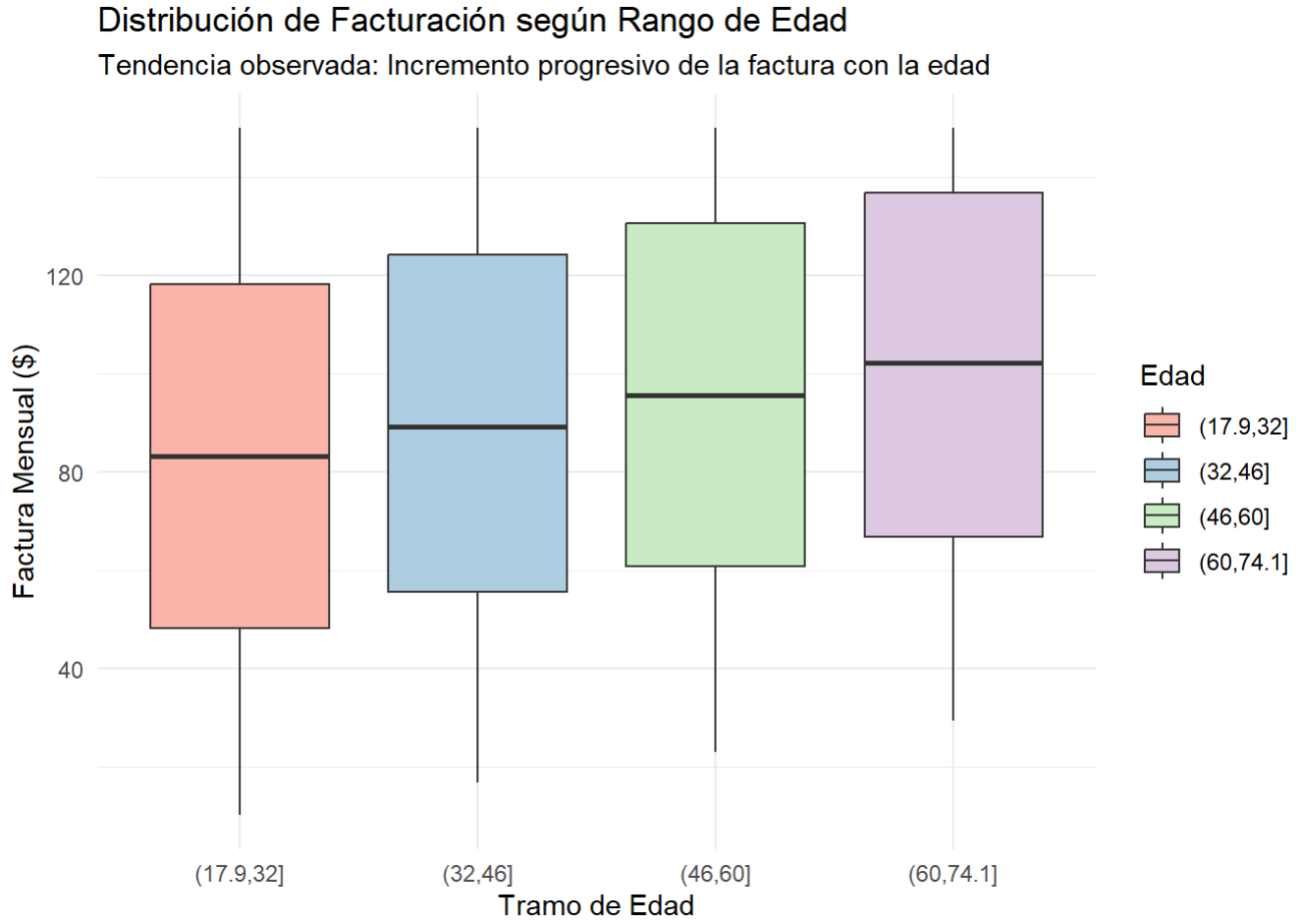
Paso 4: Análisis de la Variable confusora (Edad vs Precio)

Como se planeó, se observa una relación entre el precio y la edad.

Para facilitar la interpretación visual, segmentamos la variable continua ‘Edad’ en 6 intervalos y analizamos la distribución del gasto en cada tramo.

```r
# 1. Segmentación de la edad
churn_data <- churn_data %>%
  mutate(rango_edad = cut(personal_info_age, breaks = 4))

# 2. Visualización de la tendencia
# Añadimos 'fill = rango_edad' para que R ponga sus colores por defecto
ggplot(churn_data, aes(x = rango_edad, y = subscription_monthly_charges, fill = rango_edad))
+
  geom_boxplot() +
  labs(title = "Distribución de Facturación según Rango de Edad",
       subtitle = "Tendencia observada: Incremento progresivo de la factura con la edad",
       x = "Tramo de Edad",
       y = "Factura Mensual ($)",
       fill = "Edad") + # Leyenda automática
  theme_minimal() +
   scale_fill_brewer(palette = "Pastel1")
```



Conclusiones del Análisis de Edad-Precio:

1. **Confirmación de tendencia:** A medida que avanzamos en los rangos de edad, la mediana de la factura mensual se incrementa sistemáticamente.
2. **Implicación para el modelo:** El modelo predictivo deberá ser capaz de desentrañar esta relación para asignar correctamente la probabilidad de abandono al factor económico y no al demográfico.

Conclusiones del EDA (Dataset Modificado)

El nuevo dataset presenta:

1. **Señal Fuerte en Infraestructura:** Correlación negativa clara entre Calidad de Red y Churn intensificado en la zona Rural.
2. **Sensibilidad al Precio:** Los clientes con tarifas altas muestran mayor propensión al abandono.
3. **Complejidad Estadística:** La presencia de correlaciones entre variables independientes (Edad-Precio) asegura un escenario realista y desafiante para el entrenamiento de modelos de Machine Learning.

Decisión: El dataset es válido y está listo para la fase de modelado predictivo.