# PREDICCIÓN DE PRECIOS DE VUELOS SEGÚN AEROLÍNEA Y RUTA

**TRABAJO FINAL HECHO EN : DICIEMBRE DE 2025**
**GONZALO GASTÓN VÁZQUEZ- 904203**
**ROCÍO PEREZ GREGORINI - 905323**

# HIPOTESIS

El precio no es aleatorio; está determinado por la aerolínea y la ruta (origen, destino,tiempo y escalas).

**OBJETIVO:** Desarrollar un modelo de regresión para predecir el precio del pasaje (Price) basado en características observables.

Fuente: Kaggle

01 Script 01 – Carga de datos

02 Script 02 - Limpieza

03 Script 03 - Exploracion

04 Script 04 - Análisis principal

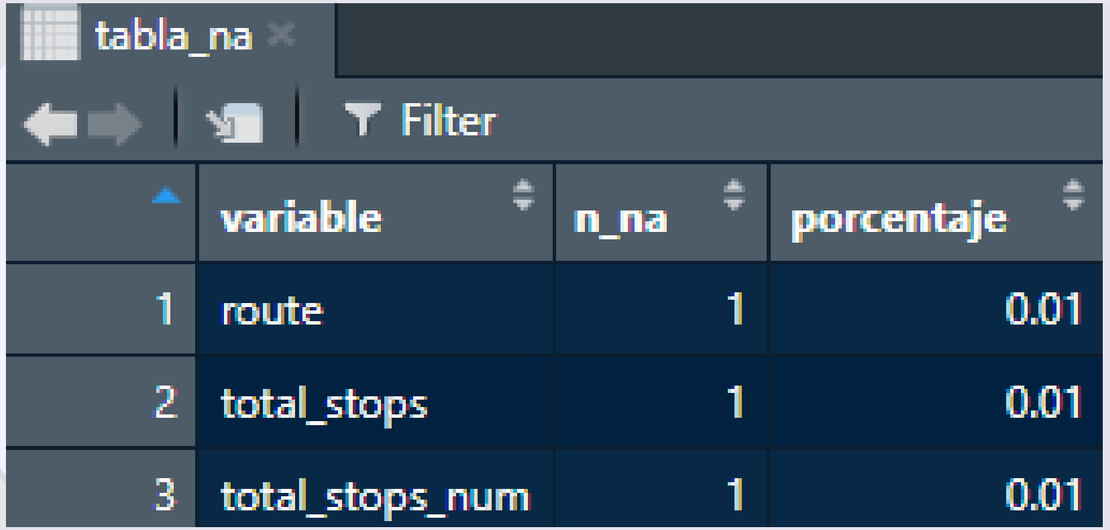
05 Script 05 - Reportes

# ESTRUCTURA Y VALORES FALTANTES

datos_raw

Filter

| | Airline | Date_of_Journey | Source | Destination | Route | Dep_Time | Arrival_Time | Duration | Total_Stops | Additional_Info | Price |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IndiGo | 24/03/2019 | Banglore | New Delhi | BLR → DEL | 22:20 | 01:10 22 Mar | 2h 50m | non-stop | No info | 3897 |
| 2 | Air India | 1/05/2019 | Kolkata | Banglore | CCU → IXR → BBI → BLR | 05:50 | 13:15 | 7h 25m | 2 stops | No info | 7662 |
| 3 | Jet Airways | 9/06/2019 | Delhi | Cochin | DEL → LKO → BOM → COK | 09:25 | 04:25 10 Jun | 19h | 2 stops | No info | 13882 |
| 4 | IndiGo | 12/05/2019 | Kolkata | Banglore | CCU → NAG → BLR | 18:05 | 23:30 | 5h 25m | 1 stop | No info | 6218 |
| 5 | IndiGo | 01/03/2019 | Banglore | New Delhi | BLR → NAG → DEL | 16:50 | 21:35 | 4h 45m | 1 stop | No info | 13302 |
| 6 | SpiceJet | 24/06/2019 | Kolkata | Banglore | CCU → BLR | 09:00 | 11:25 | 2h 25m | non-stop | No info | 3873 |
| 7 | Jet Airways | 12/03/2019 | Banglore | New Delhi | BLR → BOM → DEL | 18:55 | 10:25 13 Mar | 15h 30m | 1 stop | In-flight meal not included | 11087 |
| 8 | Jet Airways | 01/03/2019 | Banglore | New Delhi | BLR → BOM → DEL | 08:00 | 05:05 02 Mar | 21h 5m | 1 stop | No info | 22270 |
| 9 | Jet Airways | 12/03/2019 | Banglore | New Delhi | BLR → BOM → DEL | 08:55 | 10:25 13 Mar | 25h 30m | 1 stop | In-flight meal not included | 11087 |
| 10 | Multiple carriers | 27/05/2019 | Delhi | Cochin | DEL → BOM → COK | 11:25 | 19:15 | 7h 50m | 1 stop | No info | 8625 |
| 11 | Air India | 1/06/2019 | Delhi | Cochin | DEL → BLR → COK | 09:45 | 23:00 | 13h 15m | 1 stop | No info | 8907 |
| 12 | IndiGo | 18/04/2019 | Kolkata | Banglore | CCU → BLR | 20:20 | 22:55 | 2h 35m | non-stop | No info | 4174 |
| 13 | Air India | 24/06/2019 | Chennai | Kolkata | MAA → CCU | 11:40 | 13:55 | 2h 15m | non-stop | No info | 4667 |
| 14 | Jet Airways | 9/05/2019 | Kolkata | Banglore | CCU → BOM → BLR | 21:10 | 09:20 10 May | 12h 10m | 1 stop | In-flight meal not included | 9663 |



tabla_na

Filter

| | variable | n_na | porcentaje |
|---|---|---|---|
| 1 | route | 1 | 0.01 |
| 2 | total_stops | 1 | 0.01 |
| 3 | total_stops_num | 1 | 0.01 |

N = 10.683 vuelos, K= 11 variables

Se crean variables como:
duration_min y total_stop_num

**DECISION**

Debido a la nula incidencia de casos (<0.01%), se descartaron técnicas de imputación compleja.
Eliminación puntual (drop) exclusivamente en la etapa de modelado.

El impacto principal de la limpieza fue la mejora en la calidad de los datos

# GESTION DE OUTLIERS

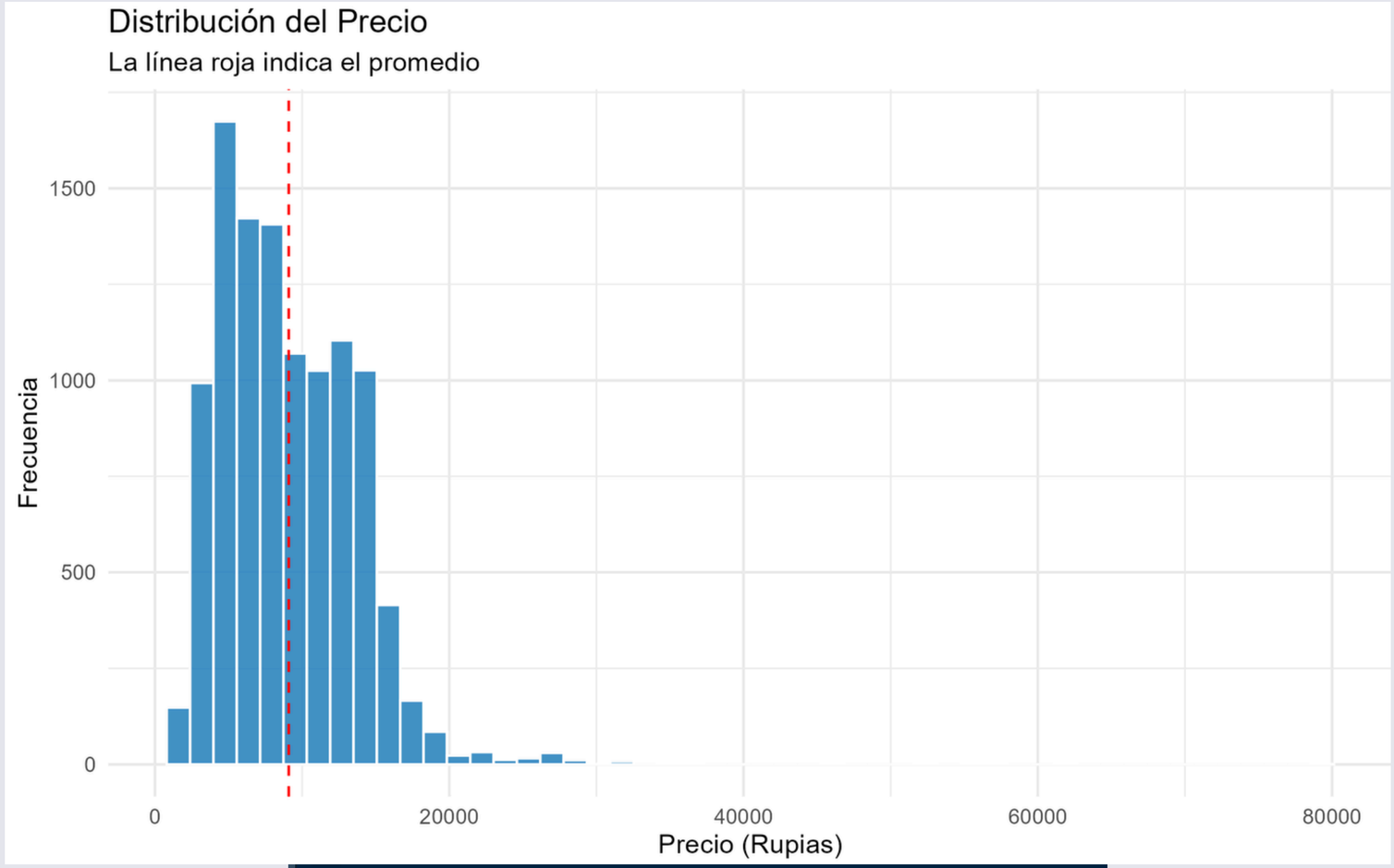
Se encontraron 94 vuelos con tarifas que superan los ₹23,017, lo que representa apenas el 0.88% de toda la muestra.

Se utilizó el método del Rango Intercuartílico (IQR) para identificar valores atípicos.

| | Variable | Límite Inf | Límite Sup | Cant. Outliers | % Muestra |
|---|---|---|---|---|---|
| 1 | price | -5367 | 23017 | 94 | 0.88% |
| 2 | duration_min | -970 | 2070 | 73 | 0.68% |

**Se decidió NO eliminar estos registros. El análisis cualitativo confirmó que corresponden legítimamente al segmento 'Jet Airways Business', que es fundamental para capturar la variabilidad real del mercado premium.**

# PRECIOS




```
> resumen_num
# A tibble: 1 x 13
      n price_mean price_median price_mode price_sd
  <int>      <dbl>        <dbl>      <dbl>    <dbl>
1 10683      9087.         8372      10262    4611.
```

Se aplico una tranformacion logaritmica para estabilizar la varianza.
La distribucion  muestra fuerte asimetria positiva.

# SUPUESTOS

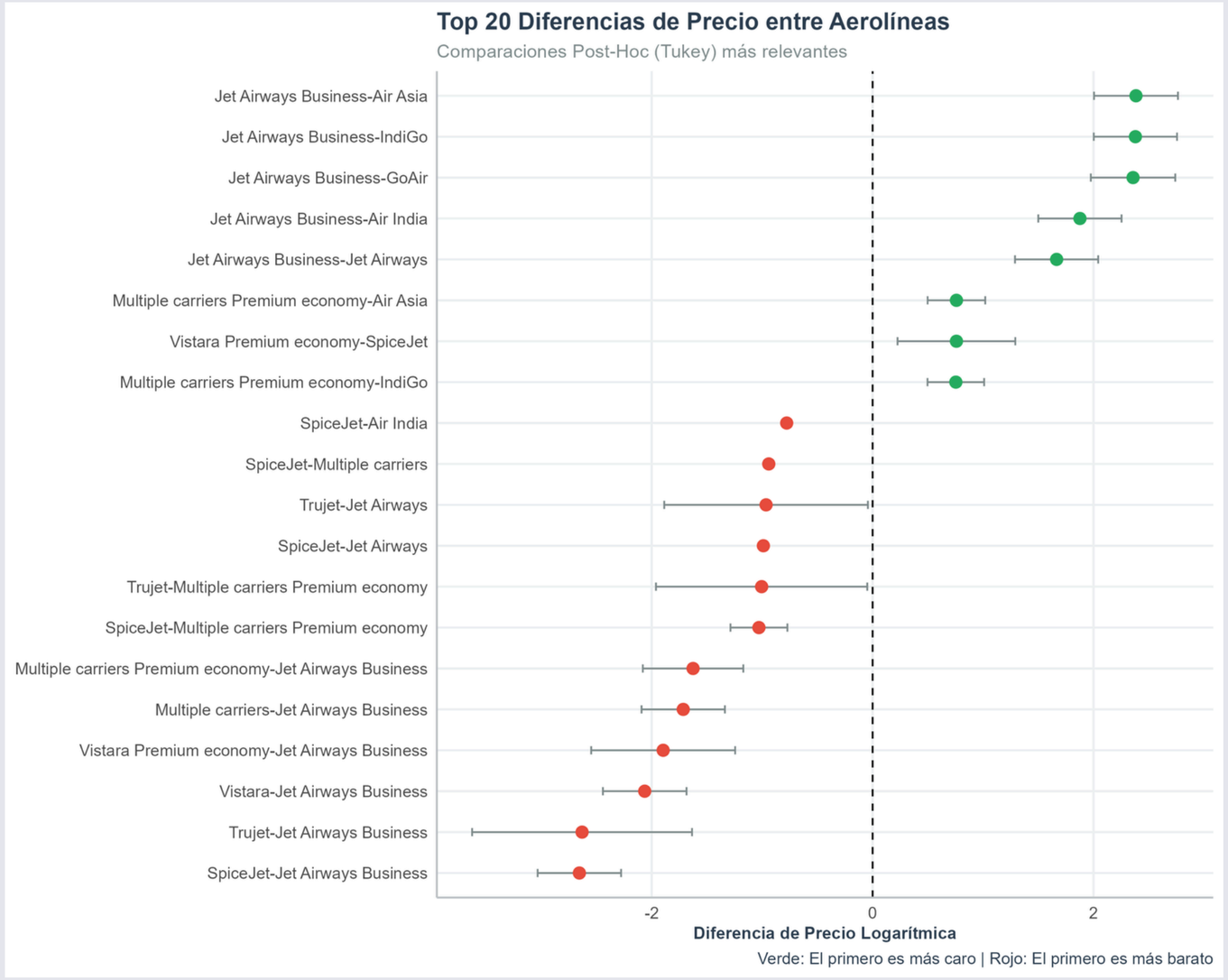
Durbin-Watson: DW = 1.98, p-value > 0.05 → sin autocorrelación.

```
> print(dw_test)

        Durbin-Watson test

data:  modelo_regresion
DW = 1.9799, p-value = 0.1499
alternative hypothesis: true autocorrelation is greater than 0
```

VIF < 5 en todos los predictores → sin colinealidad severa.

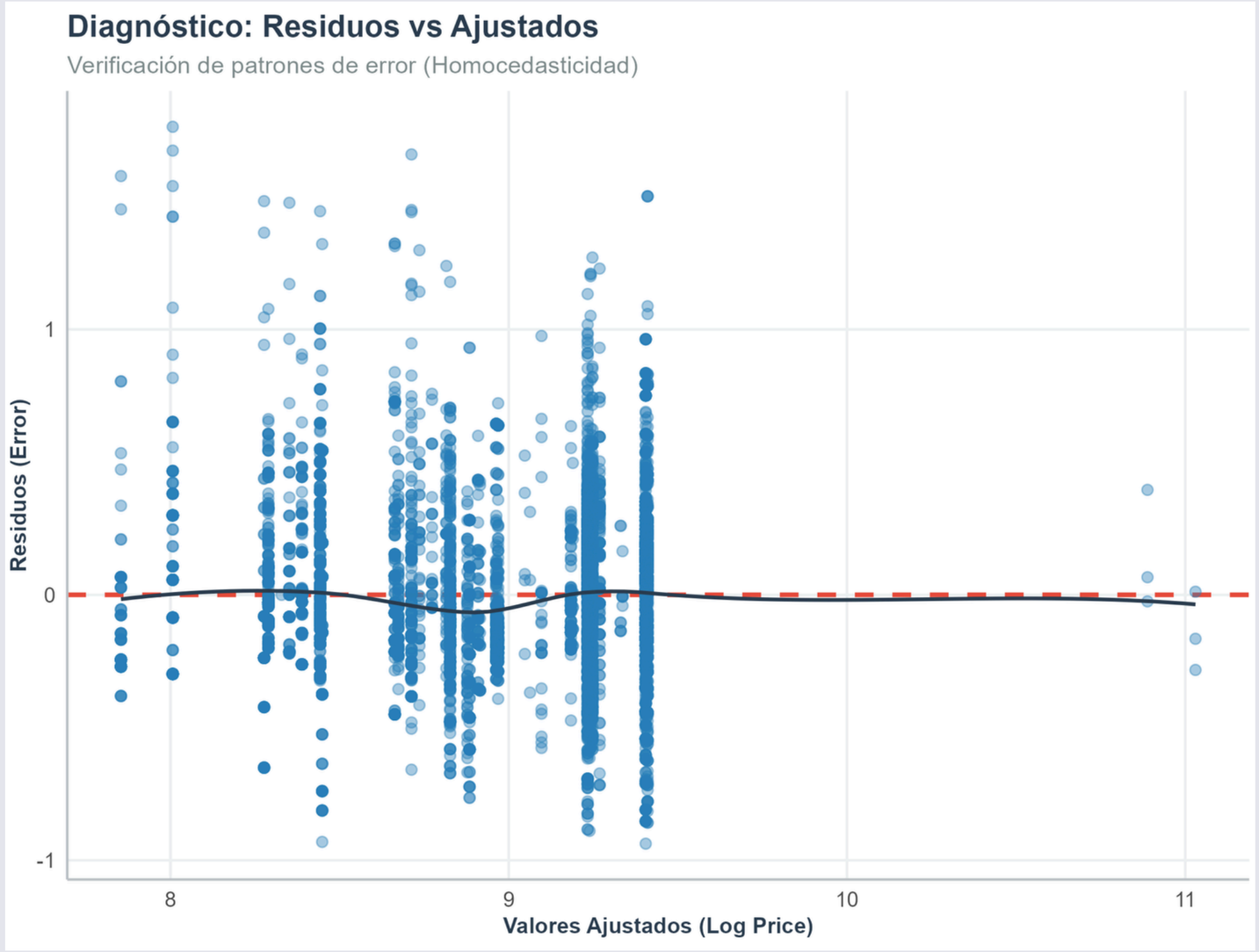
```
> # Calculamos el VIF
> vif_valores <- car::vif(modelo_regresion)
>
> print(vif_valores)
                    GVIF Df GVIF^(1/(2*Df))
airline         1.710426 11        1.024697
duration_min    2.505591  1        1.582906
total_stops_num 2.448488  1        1.564764
momento_dia     1.105237  3        1.016817
```

# ¿TODAS LAS AEROLINEAS COBRAN LO MISMO?

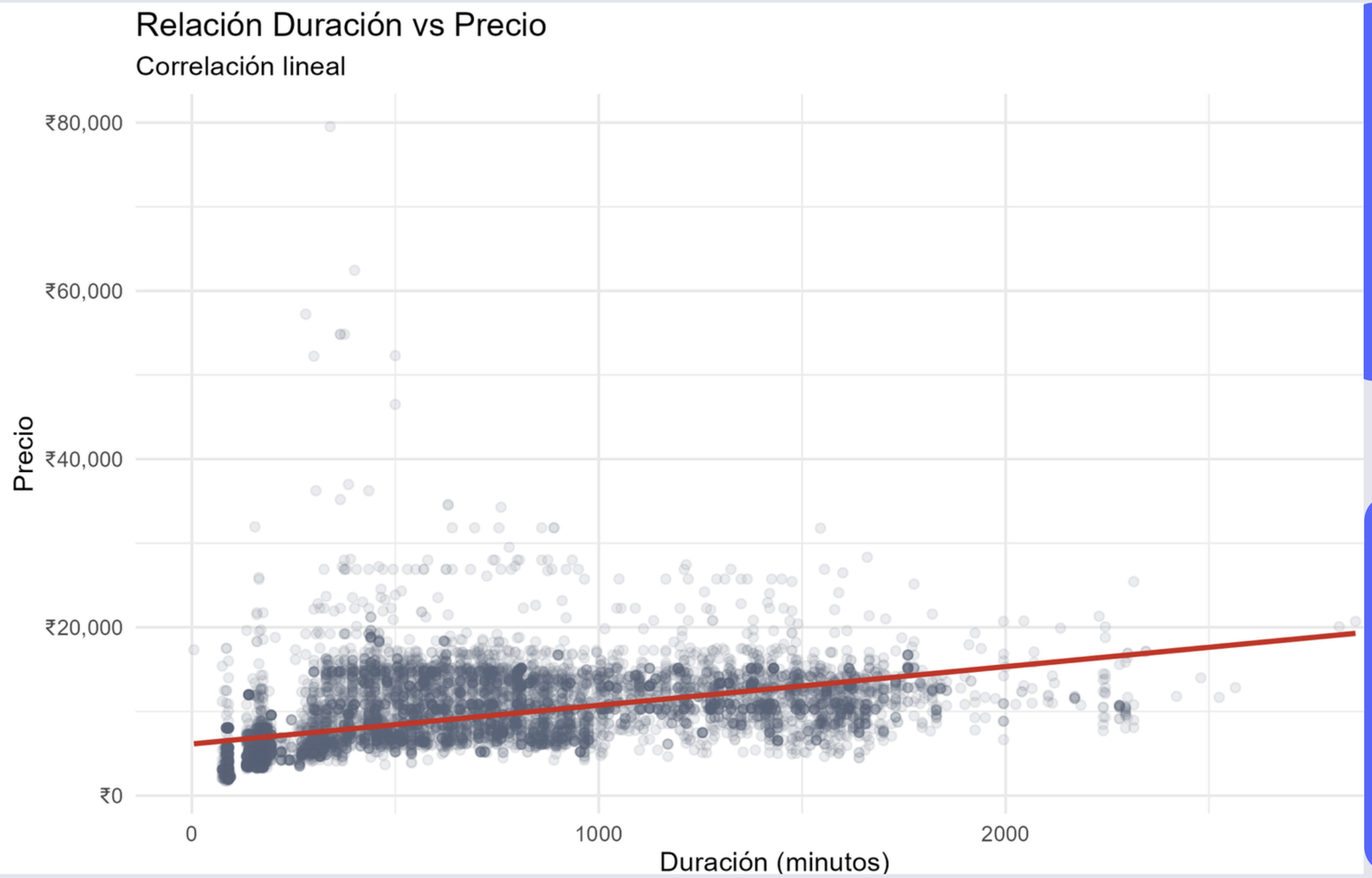



El ANOVA nos confirmó que la variable Aerolínea influía en el precio, pero es una prueba global. Aplicamos Tukey para identificar específicamente entre qué aerolíneas existían esas diferencias. Esto nos permitió validar estadísticamente que existe una brecha de precios real entre el segmento Premium (Jet Airways/Air India) y el Low-Cost.

# Por otro lado, el gráfico de residuos no evidencia patrones graves.



# DURACION VS PRECIO




Existe una relacion positiva entre duracion del vuelo y precio, sin embargo la nube de puntos muestra mucha dispersión, la duracion sola no alcanza para explicar el precio.

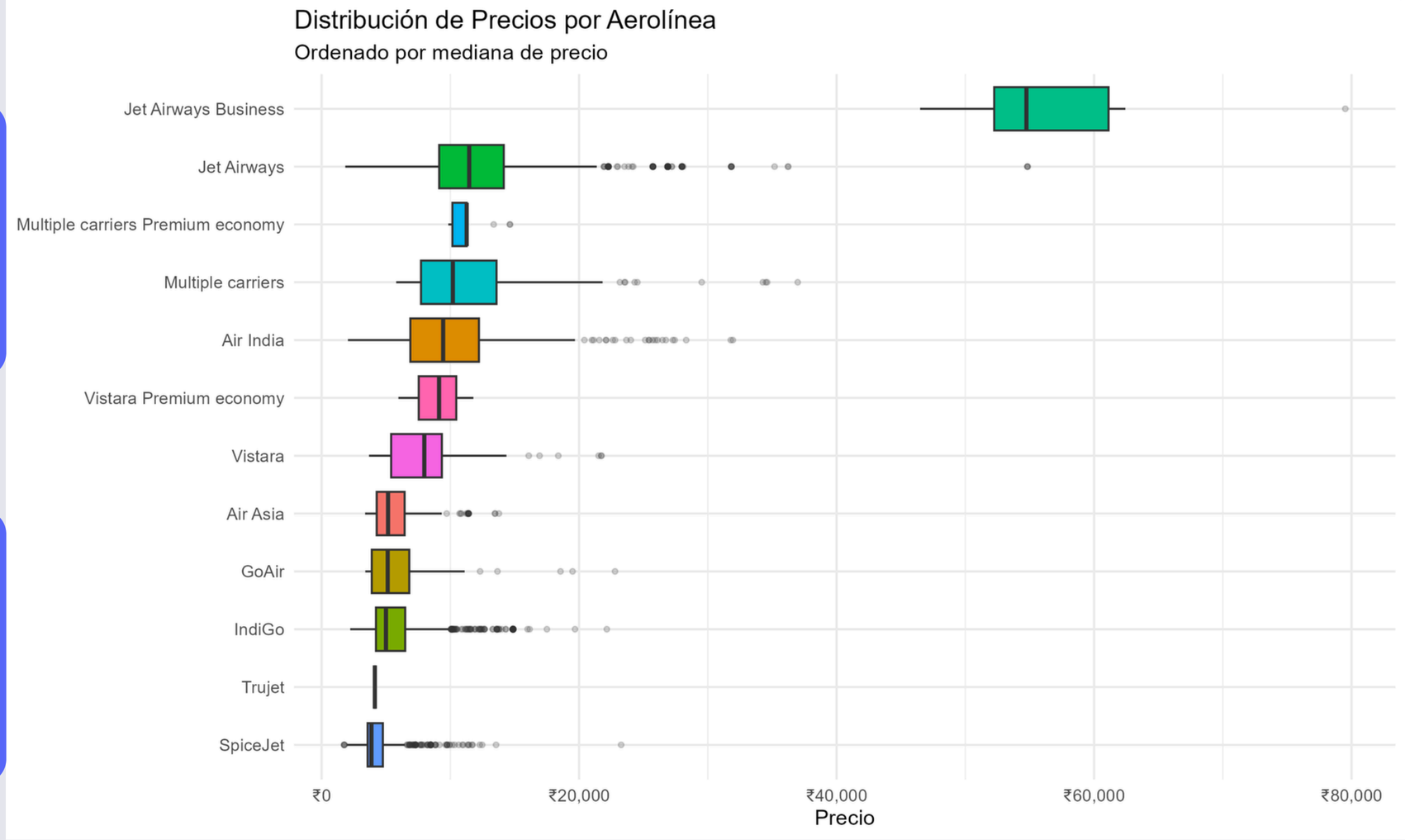
**DECISION:**

Incluir aerolínea, escalas y momento del día junto con duración en una regresión múltiple.

# DIFERENCIA DE PRECIOS POR AEROLINEA

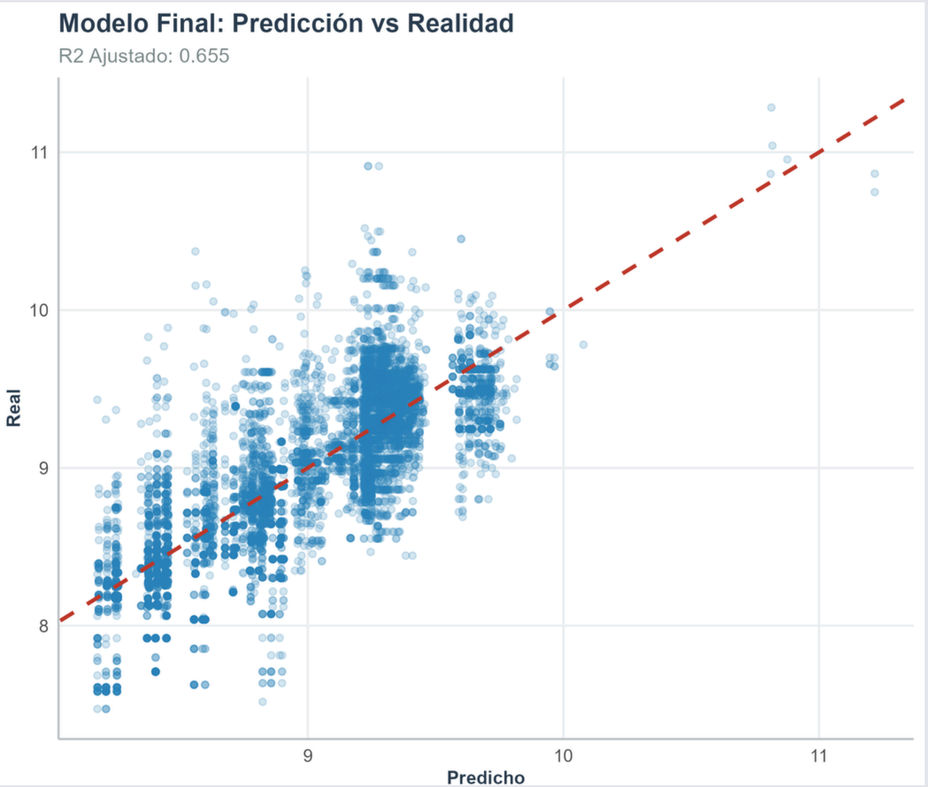
Las aerolineas como Jet Airways presentan medianas y precios muchos mas altos.

Las aerolineas low cost presentan precios mas bajos y menor dispersion.



# REGRESION MULTIPLE

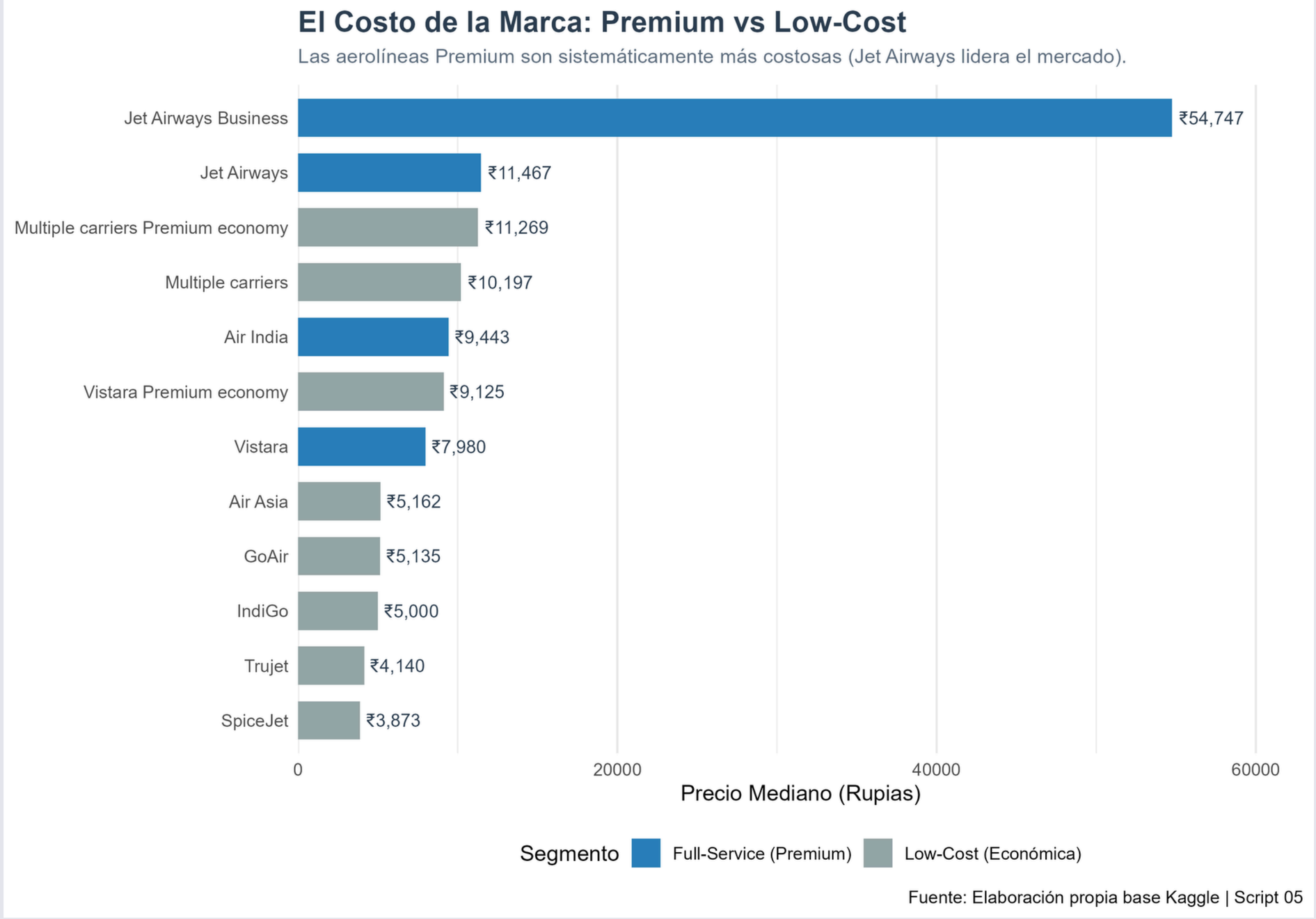
| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 8,34 | 0,02 | 431,34 | 0,00 |
| airlineAir India | 0,18 | 0,02 | 9,51 | 0,00 |
| airlineGoAir | -0,01 | 0,03 | -0,20 | 0,84 |
| airlineIndiGo | 0,03 | 0,02 | 1,57 | 0,12 |
| airlineJet Airways | 0,48 | 0,02 | 26,80 | 0,00 |
| airlineJet Airways Business | 2,09 | 0,12 | 16,81 | 0,00 |
| airlineMultiple carriers | 0,42 | 0,02 | 21,86 | 0,00 |
| airlineMultiple carriers Premium economy | 0,53 | 0,09 | 6,21 | 0,00 |
| airlineSpiceJet | -0,17 | 0,02 | -8,24 | 0,00 |
| airlineTrujet | -0,46 | 0,30 | -1,51 | 0,13 |
| airlineVistara | 0,30 | 0,02 | 13,42 | 0,00 |
| airlineVistara Premium economy | 0,65 | 0,18 | 3,69 | 0,00 |
| duration_min | 0,00 | 0,00 | 7,45 | 0,00 |
| total_stops_num | 0,36 | 0,01 | 53,27 | 0,00 |
| momento_diaMañana | 0,07 | 0,01 | 6,42 | 0,00 |
| momento_diaTarde | 0,08 | 0,01 | 6,92 | 0,00 |
| momento_diaNoche | 0,03 | 0,01 | 3,03 | 0,00 |



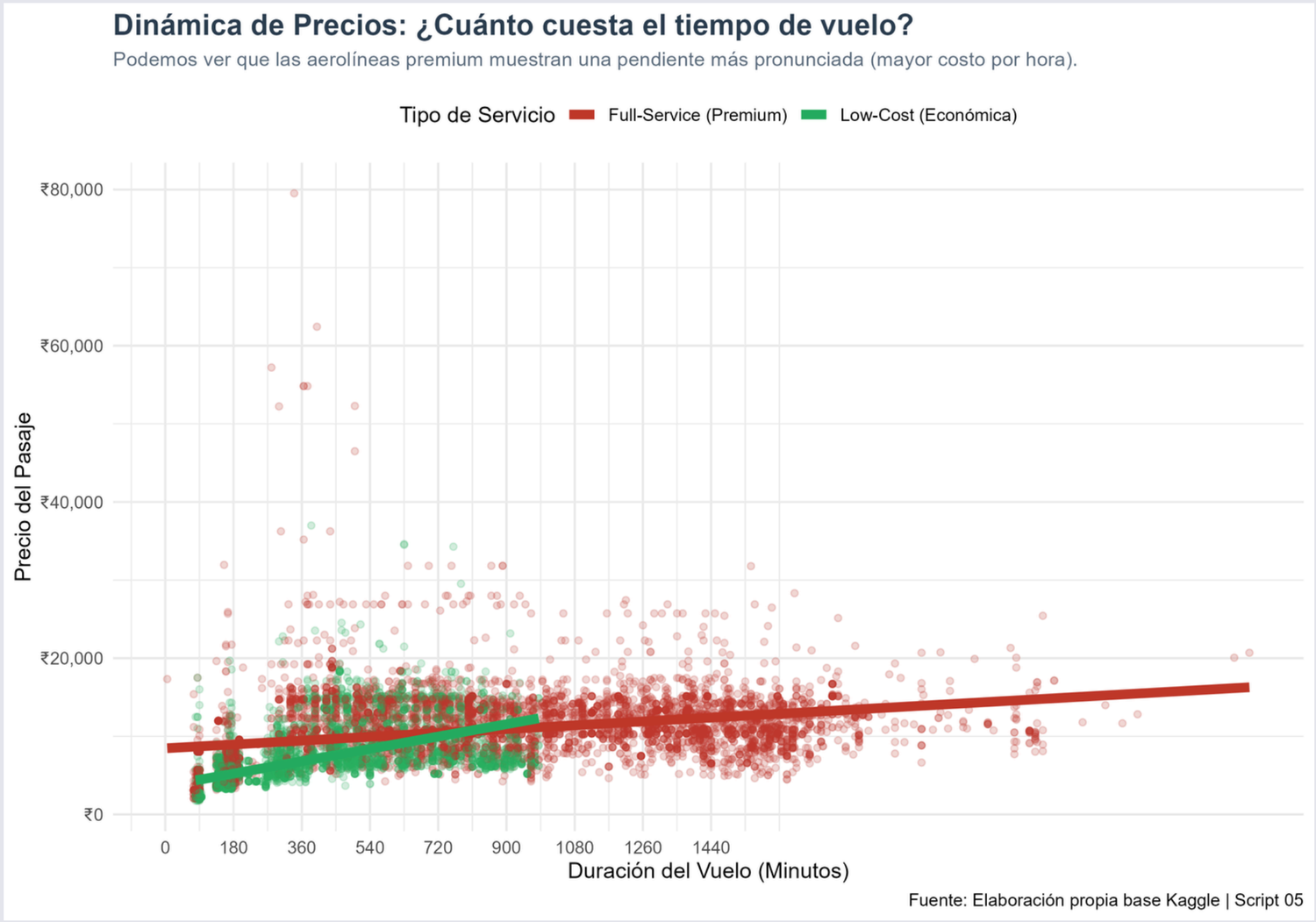

$R^2$ ajustado ≈ 0.655 → el modelo explica ~65.5 % de la varianza de log_price.

Jet Airways Business es la categoría con mayor sobreprecio (premium).

Salidas por la mañana/tarde tienen precios más altos que vuelos nocturnos, controlando por aerolínea, duración y escalas.




**Las diferencias aquí mostradas son consistentes con las diferencias significativas detectadas por Tukey HSD**




**Cada hora adicional de vuelo se cobra mas caro en servicios Full-Service que en Low-Cost.**

# ✅ Hipótesis confirmada.

Se confirma la hipótesis de investigación: El precio de los pasajes aéreos en el mercado indio está determinado por la aerolínea, la duración y el número de escalas.
Evidencia: El modelo de Regresión Lineal Múltiple alcanzó un $R^2$ Ajustado de 0.655, lo que indica que las variables seleccionadas explican el 65.5% de la varianza total de los precios.

# ✈️ Hallazgos:

Las aerolíneas Full-Service (como Jet Airways o Air India) son, en promedio, un 58.5% más costosas que las Low-Cost.
Los precios extremos (superiores a ₹50,000) no eran errores de carga, sino que correspondían al segmento "Jet Airways Business". Conservar estos datos permitió modelar correctamente el comportamiento del segmento de lujo.
Existe una correlación positiva entre duración y precio, pero con sensibilidades distintas.
Las aerolíneas Premium muestran una pendiente más pronunciada: cada hora adicional de vuelo encarece el ticket más rápido que en una aerolínea económica.

# ⚠️LIMITACIONES DEL ESTUDIO

Variable Omitida Clave: El modelo no incluye los "días de antelación a la compra”. Esto explica el 34.5% de varianza no capturada.
Ventana Temporal: Los datos cubren un periodo de 4 meses (marzo-junio), lo que impide capturar la estacionalidad anual completa (ej. temporada alta de fin de año).

# 🔮Futuras Líneas de Investigación

- Modelado No Lineal: Implementar algoritmos para capturar interacciones complejas entre Ruta y Aerolínea que el modelo lineal simplifica. Usando Machine Learning
- Datos Exógenos: Incorporar el precio del combustible  y el calendario de festivos locales para mejorar la precisión predictiva ante choques de demanda.