# Actividad 1 – Análisis de datos de aguacate

### Grupo 3 - Lote 3 - Alejandro Campos, Leticia Florido, Alejandro Cita, Iván Gómez



# 1 1. Carga de datos y preparación

En esta sección se carga la base de datos de precios de aguacate, se eliminan algunas regiones agregadas de Estados Unidos (Total U.S., West, etc.) y se preparan las variables para el análisis.

Se asume que el fichero `avocado-updated-2020.csv` está en la misma carpeta que este archivo `.Rmd`.

```
# Cargar datos desde el archivo CSV
avocado <- read.csv("avocado-updated-2020.csv")

# Estructura inicial de la base
str(avocado)
```

```
## 'data.frame':    33045 obs. of  13 variables:
##  $ date         : chr  "2015-01-04" "2015-01-04" "2015-01-04" "2015-01-04" ...
##  $ average_price: num  1.22 1.79 1 1.76 1.08 1.29 1.01 1.64 1.02 1.83 ...
##  $ total_volume : num  40873 1374 435021 3847 788025 ...
##  $ X4046        : num  2819.5 57.4 364302.4 1500.2 53987.3 ...
##  $ X4225        : num  28287 154 23821 938 552906 ...
##  $ X4770        : num  49.9 0 82.2 0 39995 ...
##  $ total_bags   : num  9716 1163 46816 1408 141137 ...
##  $ small_bags   : num  9187 1163 16707 1071 137146 ...
##  $ large_bags   : num  530 0 30109 337 3991 ...
##  $ xlarge_bags  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ type         : chr  "conventional" "organic" "conventional" "organic" ...
##  $ year         : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ geography    : chr  "Albany" "Albany" "Atlanta" "Atlanta" ...
```

```
names(avocado)
```

```
##  [1] "date"          "average_price" "total_volume"  "X4046"
##  [5] "X4225"         "X4770"         "total_bags"    "small_bags"
##  [9] "large_bags"    "xlarge_bags"   "type"          "year"
## [13] "geography"
```

```r
# Conteo de observaciones por región
avocado %>%
  count(geography, sort = TRUE)
```

```
##               geography   n
## 1                Albany 612
## 2               Atlanta 612
## 3  Baltimore/Washington 612
## 4                 Boise 612
## 5                Boston 612
## 6     Buffalo/Rochester 612
## 7            California 612
## 8             Charlotte 612
## 9               Chicago 612
## 10    Cincinnati/Dayton 612
## 11             Columbus 612
## 12     Dallas/Ft. Worth 612
## 13               Denver 612
## 14              Detroit 612
## 15         Grand Rapids 612
## 16          Great Lakes 612
## 17  Harrisburg/Scranton 612
## 18 Hartford/Springfield 612
## 19              Houston 612
## 20         Indianapolis 612
## 21         Jacksonville 612
## 22            Las Vegas 612
## 23          Los Angeles 612
## 24           Louisville 612
## 25 Miami/Ft. Lauderdale 612
## 26             Midsouth 612
## 27            Nashville 612
## 28   New Orleans/Mobile 612
## 29             New York 612
## 30            Northeast 612
## 31 Northern New England 612
## 32              Orlando 612
## 33         Philadelphia 612
## 34       Phoenix/Tucson 612
## 35           Pittsburgh 612
## 36               Plains 612
## 37             Portland 612
## 38   Raleigh/Greensboro 612
## 39     Richmond/Norfolk 612
## 40              Roanoke 612
## 41           Sacramento 612
## 42            San Diego 612
## 43        San Francisco 612
## 44              Seattle 612
## 45       South Carolina 612
## 46        South Central 612
## 47            Southeast 612
## 48              Spokane 612
## 49            St. Louis 612
## 50             Syracuse 612
## 51                Tampa 612
## 52           Total U.S. 612
## 53                 West 612
## 54  West Tex/New Mexico 609
```

```r
# Lista de regiones agregadas que se van a excluir
regiones_grandes <- c(
  "Total U.S.", "West", "South Central", "Northeast",
  "Southeast", "Midsouth", "Great Lakes", "Plains", "California"
)

# Filtramos para quedarnos con regiones de mercado más desagregadas
avocado_limpio <- avocado %>%
  filter(!geography %in% regiones_grandes)

# Convertimos algunas variables a factor (categóricas)
avocado_limpio$type      <- as.factor(avocado_limpio$type)
avocado_limpio$geography <- as.factor(avocado_limpio$geography)
avocado_limpio$year      <- as.factor(avocado_limpio$year)

# Renombramos las columnas de PLU a nombres más legibles (si existen)
avocado_limpio <- avocado_limpio %>%
  rename(
    plu_small  = `X4046`,
    plu_large  = `X4225`,
    plu_xlarge = `X4770`
  )

# Estructura después de la limpieza y renombrado
str(avocado_limpio)
```

```
## 'data.frame':    27537 obs. of  13 variables:
##  $ date         : chr  "2015-01-04" "2015-01-04" "2015-01-04" "2015-01-04" ...
##  $ average_price: num  1.22 1.79 1 1.76 1.08 1.29 1.01 1.64 1.02 1.83 ...
##  $ total_volume : num  40873 1374 435021 3847 788025 ...
##  $ plu_small    : num  2819.5 57.4 364302.4 1500.2 53987.3 ...
##  $ plu_large    : num  28287 154 23821 938 552906 ...
##  $ plu_xlarge   : num  49.9 0 82.2 0 39995 ...
##  $ total_bags   : num  9716 1163 46816 1408 141137 ...
##  $ small_bags   : num  9187 1163 16707 1071 137146 ...
##  $ large_bags   : num  530 0 30109 337 3991 ...
##  $ xlarge_bags  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ type         : Factor w/ 2 levels "conventional",..: 1 2 1 2 1 2 1 2 1 2 ...
##  $ year         : Factor w/ 6 levels "2015","2016",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ geography    : Factor w/ 45 levels "Albany","Atlanta",..: 1 1 2 2 3 3 4 4 5 5
...
```

**Comentario:**

- `date` es una variable de fecha (frecuencia semanal).
- `average_price`, `total_volume`, `total_bags`, `small_bags`, `large_bags`, `xlarge_bags`, `plu_small`, `plu_large`, `plu_xlarge` son variables numéricas continuas.
- `type`, `geography` y `year` son variables categóricas (factor).

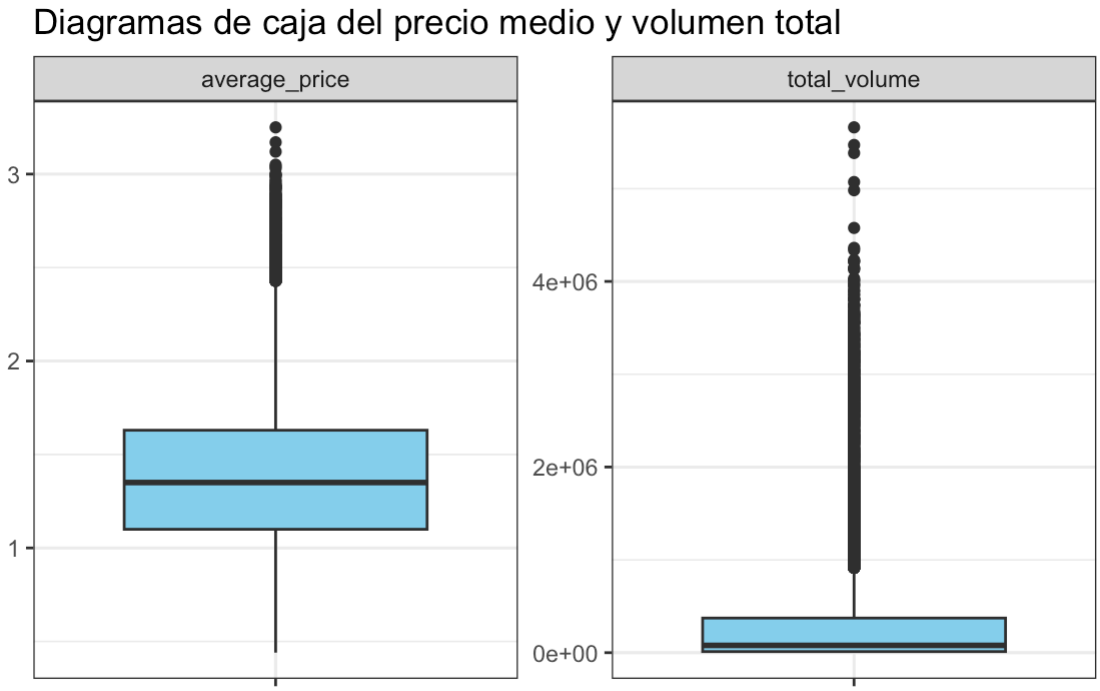
# 2 2. Tipos de variables y análisis exploratorio de las variables numéricas

La actividad pide identificar los tipos de variables y realizar un análisis exploratorio de las variables numéricas (media, varianza, diagramas de caja, etc.). Nos centraremos en `average_price` y `total_volume` por ser las variables clave en el resto del ejercicio.

```
# Seleccionamos las variables numéricas principales
main_vars_num <- c("average_price", "total_volume")

# Subconjunto numérico
datos_num <- avocado_limpio[ , main_vars_num]

# Función para obtener un resumen extendido de cada variable
resumen_extendido <- function(x) {
  qs <- quantile(x, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
  c(
    Q1      = qs[1],
    mediana = qs[2],
    Q3      = qs[3],
    IQR     = IQR(x, na.rm = TRUE),
    media   = mean(x, na.rm = TRUE),
    sd      = sd(x, na.rm = TRUE),
    var     = var(x, na.rm = TRUE)
  )
}

tabla_desc <- t(apply(datos_num, 2, resumen_extendido))
tabla_desc_redonda <- round(tabla_desc, 3)

tabla_desc_redonda
```

```
##                Q1.25% mediana.50%    Q3.75%       IQR      media         sd
## average_price     1.10        1.35      1.63      0.53      1.387      0.388
## total_volume  11838.56    78185.95 372777.66 360939.10 283735.488 475378.648
##                       var
## average_price 1.510000e-01
## total_volume  2.259849e+11
```

```r
# Pasamos a formato largo para hacer boxplots de ambas variables a la vez
avocado_long <- avocado_limpio %>%
  select(all_of(main_vars_num)) %>%
  pivot_longer(cols = everything(),
               names_to = "variable",
               values_to = "valor")

ggplot(avocado_long, aes(x = "", y = valor)) +
  geom_boxplot(fill = "skyblue") +
  theme_bw() +
  facet_wrap(~ variable, scales = "free_y") +
  labs(
    title = "Diagramas de caja del precio medio y volumen total",
    x = "",
    y = ""
  )
```




**Comentario:**

En la muestra, el precio medio de un aguacate es de aproximadamente 1,39 dólares, con una mediana muy similar (1,35). Esto indica que la distribución de `average_price` es bastante simétrica y que no hay una cola muy pronunciada hacia valores extremos. El rango intercuartílico (IQR) es de unos 0,53 dólares, por lo que el 50 % central de las observaciones se sitúa aproximadamente entre 1,10 y 1,63 dólares, lo que refleja una dispersión moderada del precio.

En cambio, el volumen total vendido ( `total_volume` ) presenta una media cercana a 283.735 unidades y una mediana mucho más baja (alrededor de 78.186). La diferencia tan grande entre media y mediana, junto con un IQR muy elevado, indica una distribución muy asimétrica a la derecha: la mayoría de las semanas tienen volúmenes relativamente bajos, pero hay algunas semanas con ventas muy altas que empujan la media hacia arriba. Los diagramas de caja confirman esta situación, mostrando una concentración de datos cerca del cero y varios valores atípicos en la parte superior del boxplot de `total_volume` .

# 3 3. Submuestras: precio de los aguacates orgánicos en Albany y Boston

La actividad pide extraer el precio de venta de los aguacates orgánicos vendidos en Albany y Boston. Para ello creamos una submuestra con esas dos regiones y tipo orgánico y calculamos también el precio medio global por ciudad.

```
# Definimos un vector con las regiones de interés
albanyboston <- c("Albany", "Boston")

# Submuestra: aguacates orgánicos en Albany y Boston
precios_org_albanyboston <- subset(
  avocado_limpio,
  type == "organic" & geography %in% albanyboston,
  select = c(geography, average_price)
)

head(precios_org_albanyboston)
```

```
##     geography average_price
## 2      Albany          1.79
## 10     Boston          1.83
## 92     Albany          1.77
## 100    Boston          1.94
## 182    Albany          1.93
## 190    Boston          2.00
```

```
# Resumen: precio medio global por ciudad
resumen_precios_albanyboston <- aggregate(
  average_price ~ geography,
  data = precios_org_albanyboston,
  FUN  = mean
)

names(resumen_precios_albanyboston)[2] <- "precio_medio_global"

resumen_precios_albanyboston
```

```
##   geography precio_medio_global
## 1    Albany            1.683529
## 2    Boston            1.742778
```

**Comentario:**

La submuestra contiene las observaciones de aguacates orgánicos vendidos en Albany y Boston. A partir del resumen, se observa que el precio medio global del aguacate orgánico en Albany es de aproximadamente 1,68 dólares por unidad, mientras que en Boston es algo mayor, en torno a 1,74 dólares.

Por tanto, en promedio el aguacate orgánico resulta ligeramente más caro en Boston que en Albany. La diferencia no es muy grande, pero sugiere que el mercado de Boston soporta precios algo más altos para este tipo de producto.

---

# 4 4. Covarianza y matriz de correlación

Como paso previo al modelado, la actividad pide calcular la covarianza y la matriz de correlación del precio de los aguacates orgánicos, convencionales y su volumen de ventas. Para ello construimos una tabla en la que, para cada combinación de fecha y región, aparezcan las observaciones de ambos tipos (orgánico y convencional).

```
# 1) Separar orgánico y convencional
org  <- subset(avocado_limpio, type == "organic")
conv <- subset(avocado_limpio, type == "conventional")

# 2) Nos quedamos con las variables necesarias
org_small  <- org[ , c("date", "geography", "average_price", "total_volume")]
conv_small <- conv[ , c("date", "geography", "average_price", "total_volume")]

# 3) Unir por fecha y región: sólo casos donde existan los dos tipos
datos_merge <- merge(
  org_small,
  conv_small,
  by = c("date", "geography"),
  suffixes = c("_org", "_conv")
)

# 4) Crear las 3 variables de interés
vars_cor <- data.frame(
  price_org  = datos_merge$average_price_org,
  price_conv = datos_merge$average_price_conv,
  vol_total  = datos_merge$total_volume_org + datos_merge$total_volume_conv
)

head(vars_cor)
```

```
##   price_org price_conv vol_total
## 1      1.79       1.22  42247.23
## 2      1.76       1.00 438868.18
## 3      1.29       1.08 807162.34
## 4      1.64       1.01  81539.44
## 5      1.83       1.02 493930.13
## 6      1.73       1.40 116633.26
```

```
# Matriz de covarianzas
cov_matrix <- cov(vars_cor, use = "complete.obs")
cov_matrix
```

```
##                price_org    price_conv     vol_total
## price_org   1.253106e-01  4.664413e-02 -1.020610e+04
## price_conv  4.664413e-02  6.460976e-02 -3.347256e+04
## vol_total  -1.020610e+04 -3.347256e+04  3.312079e+11
```

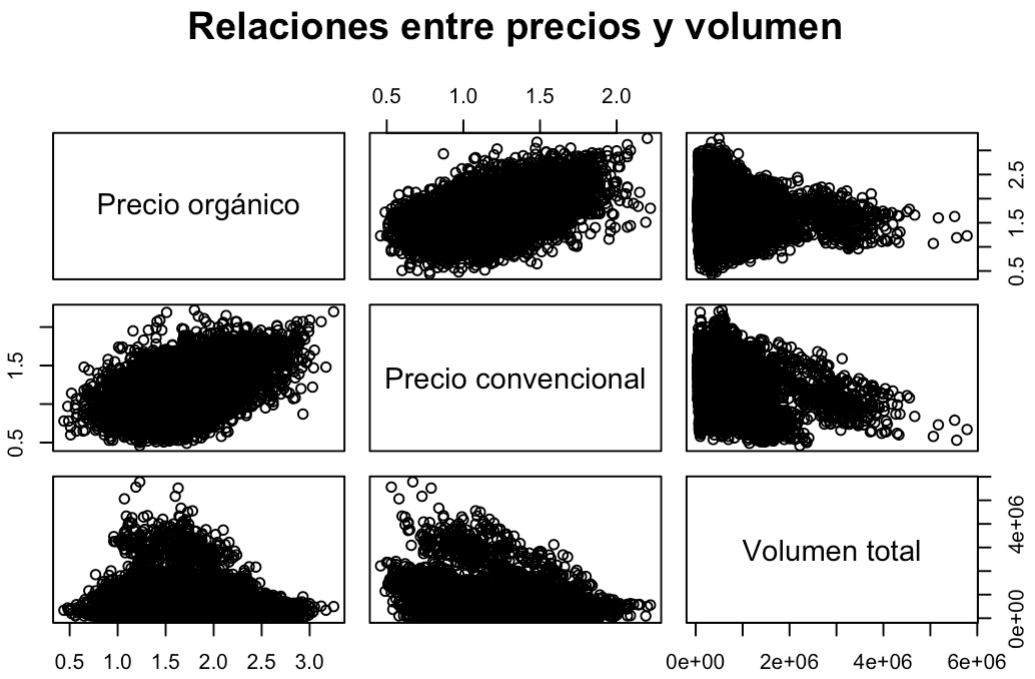
```
# Matriz de correlaciones
cor_matrix <- cor(vars_cor, use = "complete.obs")
cor_matrix
```

```
##              price_org price_conv   vol_total
## price_org   1.00000000  0.5183869 -0.05009745
## price_conv  0.51838688  1.0000000 -0.22881766
## vol_total  -0.05009745 -0.2288177  1.00000000
```

```
# Matriz de correlaciones con p-values (Hmisc::rcorr)
rcorr(as.matrix(vars_cor), type = "pearson")
```

```
##            price_org price_conv vol_total
## price_org       1.00       0.52     -0.05
## price_conv      0.52       1.00     -0.23
## vol_total      -0.05      -0.23      1.00
##
## n= 13767
##
##
## P
##            price_org price_conv vol_total
## price_org             0          0
## price_conv  0                    0
## vol_total   0         0
```

```
pairs(
  vars_cor,
  main   = "Relaciones entre precios y volumen",
  labels = c("Precio orgánico", "Precio convencional", "Volumen total")
)
```




**Comentario e**

**interpretación**

La matriz de correlaciones muestra una correlación positiva moderada (≈ 0,52) entre el precio de los aguacates orgánicos y el de los convencionales. Esto indica que, cuando sube el precio de uno de los tipos, suele subir también el del otro, lo cual es razonable si ambos responden a factores comunes de mercado

(costes, demanda, etc.).

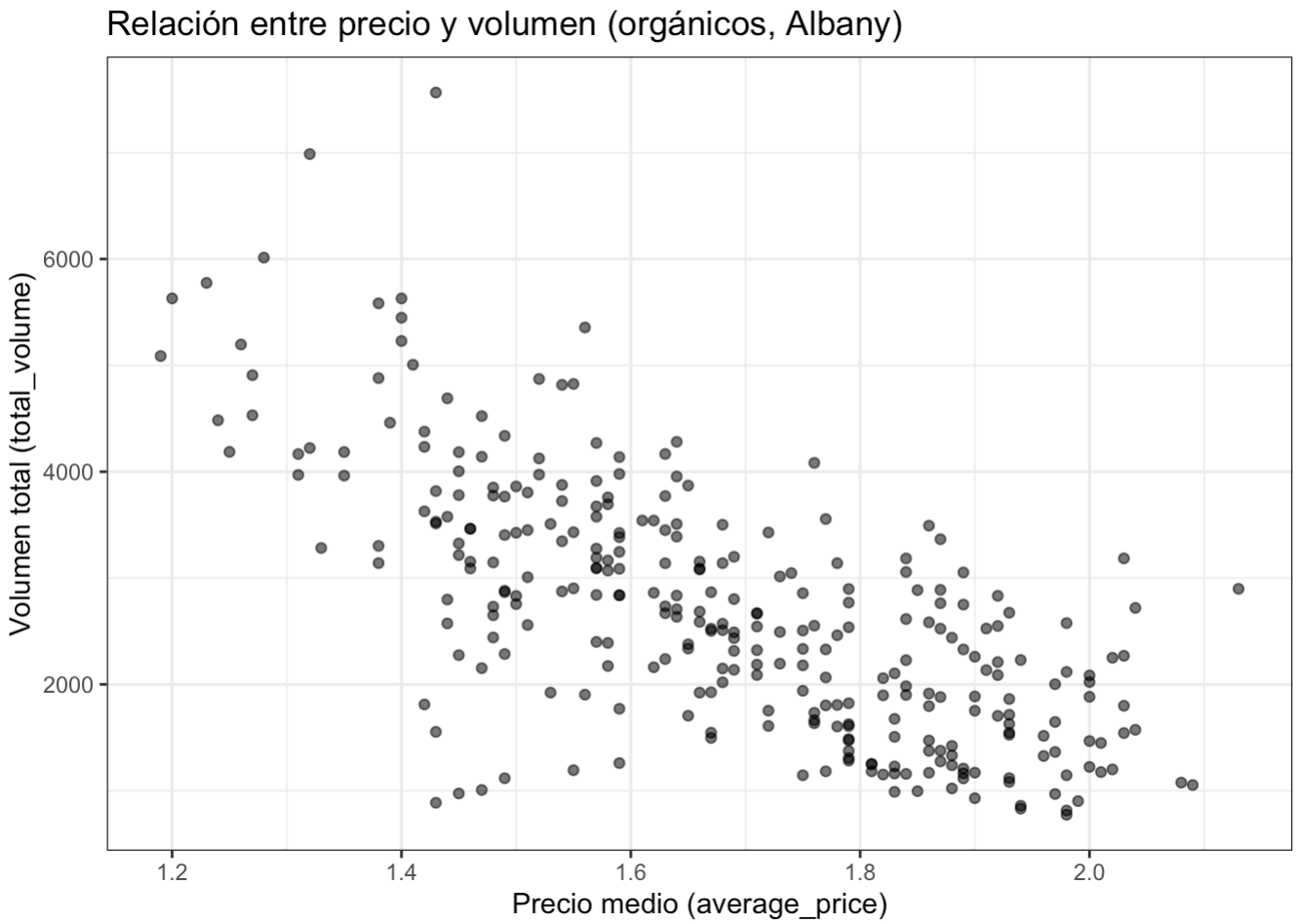
En cuanto a la relación con el volumen total de ventas, las correlaciones son negativas: alrededor de −0,05 entre el precio orgánico y el volumen, y de −0,23 entre el precio convencional y el volumen. La relación es débil para el orgánico y algo más marcada para el convencional. Este signo negativo es coherente con la teoría económica: precios más altos tienden a asociarse con menores cantidades vendidas. Sin embargo, la magnitud no es muy grande, por lo que el precio por sí solo no explica toda la variación observada en las ventas. Los p-values prácticamente nulos se deben al tamaño muestral muy elevado, pero desde el punto de vista económico las correlaciones deben interpretarse como moderadas (precio–precio) y débiles/medias (precio–volumen).

---

# 5 5. Relación precios–volumen y uso de logaritmos

La actividad pide determinar la relación entre los precios y el volumen de ventas, y analizar cómo cambia dicha relación al tomar logaritmos. Para ilustrarlo, se trabaja con la submuestra de aguacates orgánicos vendidos en Albany.

```
# Submuestra: orgánicos en Albany
albany_org <- subset(
  avocado_limpio,
  type == "organic" & geography == "Albany"
)

# Eliminamos posibles NA
albany_org <- na.omit(albany_org)

# Dispersión precio vs volumen
ggplot(albany_org,
       aes(x = average_price, y = total_volume)) +
  geom_point(alpha = 0.6) +
  theme_bw() +
  labs(title = "Relación entre precio y volumen (orgánicos, Albany)",
       x = "Precio medio (average_price)",
       y = "Volumen total (total_volume)")
```




```
# Modelo lineal en niveles
modelo_lin <- lm(total_volume ~ average_price, data = albany_org)
summary(modelo_lin)
```

```
##
## Call:
## lm(formula = total_volume ~ average_price, data = albany_org)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2822.9  -584.7   -48.9   541.6  3855.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9511.3      407.3   23.35   <2e-16 ***
## average_price  -4056.7      240.2  -16.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 855 on 304 degrees of freedom
## Multiple R-squared:  0.4841, Adjusted R-squared:  0.4824
## F-statistic: 285.2 on 1 and 304 DF,  p-value: < 2.2e-16
```

```r
# Modelo lineal con logaritmos (elasticidad)
albany_org2 <- subset(albany_org, total_volume > 0)

modelo_log <- lm(log(total_volume) ~ log(average_price),
                 data = albany_org2)
summary(modelo_log)
```

```
##
## Call:
## lm(formula = log(total_volume) ~ log(average_price), data = albany_org2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39458 -0.21847  0.03444  0.24617  0.78862
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.07982    0.08495  106.88   <2e-16 ***
## log(average_price) -2.50803    0.16089  -15.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3473 on 304 degrees of freedom
## Multiple R-squared:  0.4443, Adjusted R-squared:  0.4424
## F-statistic:   243 on 1 and 304 DF,  p-value: < 2.2e-16
```

**Comentario e interpretación**

En el modelo lineal en niveles, la pendiente asociada al precio ( `average_price` ) es de aproximadamente −4057. Esto significa que, manteniendo el resto constante, un aumento de 1 dólar en el precio medio se asocia con una reducción de unas 4.057 unidades en el volumen semanal vendido de aguacates orgánicos en Albany. El coeficiente es estadísticamente significativo (p-value < 0,001) y el $R^2 \approx 0{,}48$ indica que casi la mitad de la variabilidad del volumen se explica únicamente por el precio.

En el modelo log-log, el coeficiente de `log(average_price)` es aproximadamente −2,51. Este valor se interpreta como una elasticidad precio de la demanda: un aumento del 1 % en el precio se asocia, en promedio, con una caída de alrededor del 2,5 % en el volumen de ventas. De nuevo, el coeficiente es altamente significativo y el $R^2 \approx 0{,}44$ muestra que el modelo explica una parte importante, aunque no total, de la variabilidad del volumen. En conjunto, ambos modelos confirman una relación negativa entre precio y cantidad demandada, consistente con la teoría de la demanda.

---

# 6 6. Predicción del precio en Albany a 3 meses

Por último, se realiza una predicción del precio de venta de los aguacates orgánicos vendidos en Albany a 3 meses vista. Para ello se ajusta un modelo de tendencia temporal lineal sobre la serie de precios y se extrapola 12 semanas a partir de la última observación disponible.

```r
# Nos aseguramos de que la fecha es de tipo Date
albany_org$date <- as.Date(albany_org$date)

# Ordenamos por fecha
albany_org <- albany_org[order(albany_org$date), ]

# Creamos un índice temporal
albany_org$time_index <- 1:nrow(albany_org)

head(albany_org[, c("date", "average_price", "time_index")])
```

```
##           date average_price time_index
## 2   2015-01-04          1.79          1
## 92  2015-01-11          1.77          2
## 182 2015-01-18          1.93          3
## 272 2015-01-25          1.89          4
## 362 2015-02-01          1.83          5
## 452 2015-02-08          1.59          6
```

```r
tail(albany_org[, c("date", "average_price", "time_index")])
```

```
##             date average_price time_index
## 26999 2020-10-25          1.49        301
## 27089 2020-11-01          1.33        302
## 27179 2020-11-08          1.44        303
## 27269 2020-11-15          1.38        304
## 27359 2020-11-22          1.44        305
## 27449 2020-11-29          1.45        306
```

```r
# Modelo de tendencia lineal
modelo_precio <- lm(average_price ~ time_index, data = albany_org)
summary(modelo_precio)
```

```
##
## Call:
## lm(formula = average_price ~ time_index, data = albany_org)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40222 -0.10540  0.01367  0.11935  0.41590
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.883217   0.019310   97.52   <2e-16 ***
## time_index  -0.001301   0.000109  -11.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1685 on 304 degrees of freedom
## Multiple R-squared:  0.3189, Adjusted R-squared:  0.3167
## F-statistic: 142.3 on 1 and 304 DF,  p-value: < 2.2e-16
```

```
# Última fecha observada
ultima_fecha <- max(albany_org$date)

# 3 meses ≈ 12 semanas
horizonte_semanas <- 12

nuevo_tiempo <- data.frame(
  time_index = nrow(albany_org) + horizonte_semanas
)

pred_3m <- predict(
  modelo_precio,
  newdata  = nuevo_tiempo,
  interval = "prediction"
)

pred_3m
```

```
##        fit      lwr      upr
## 1 1.469532 1.135581 1.803483
```

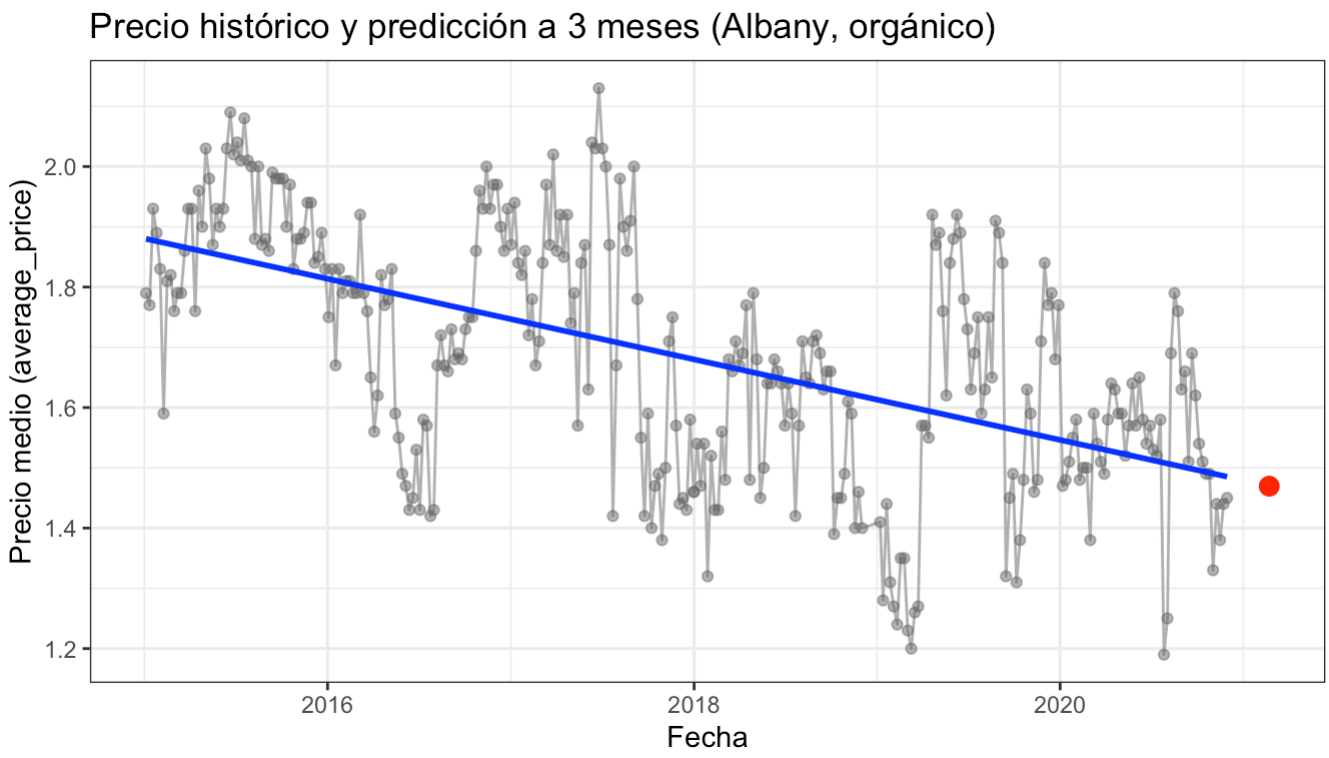
```
ultima_fecha
```

```
## [1] "2020-11-29"
```

```
ultima_fecha + 7 * horizonte_semanas   # fecha aproximada para la predicción
```

```
## [1] "2021-02-21"
```

```r
# ------- Gráfico del modelo y la predicción -------

# Fecha de la predicción
fecha_pred <- ultima_fecha + 7 * horizonte_semanas

# Data frame con el punto de predicción
punto_pred <- data.frame(
  date          = fecha_pred,
  average_price = pred_3m[1, "fit"]
)

# Gráfico: serie histórica + recta de tendencia + punto de predicción
ggplot(albany_org, aes(x = date, y = average_price)) +
  geom_line(color = "grey70") +
  geom_point(alpha = 0.5, color = "grey40") +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  geom_point(data = punto_pred,
             aes(x = date, y = average_price),
             color = "red", size = 3) +
  theme_bw() +
  labs(
    title = "Precio histórico y predicción a 3 meses (Albany, orgánico)",
    x = "Fecha",
    y = "Precio medio (average_price)"
  )
```




**Comentario e interpretación**

La tendencia estimada para el precio de los aguacates orgánicos en Albany es ligeramente descendente: el coeficiente del índice temporal es de aproximadamente −0,0013 dólares por semana, y resulta estadísticamente significativo (p-value < 0,001).

Al extrapolar 12 semanas (≈ 3 meses) a partir de la última observación disponible (29-11-2020), el modelo predice un precio esperado de unos 1,47 dólares por unidad para alrededor del 21-02-2021. El intervalo de predicción al 95 % se sitúa aproximadamente entre 1,14 y 1,80 dólares, lo que indica una incertidumbre considerable en la predicción puntual. En cualquier caso, el modelo sugiere que, si se mantiene la tendencia observada, el precio de los aguacates orgánicos en Albany se mantendría en torno a valores similares o ligeramente inferiores a los actuales.

En el gráfico anterior se representa la evolución del precio medio de los aguacates orgánicos en Albany (puntos y línea gris), junto con la recta de tendencia estimada por el modelo lineal (línea azul). El punto rojo marca la predicción a 3 meses vista obtenida a partir de dicho modelo. Se aprecia que la tendencia general del precio es ligeramente descendente y que la predicción se sitúa en la prolongación natural de esta recta, en torno a 1,47 dólares por unidad para finales de febrero de 2021.