# Reporte EDA – Punto 1 (Estimación de Ventas Retail)

## 1. Exploración inicial

- El dataset contiene **100 tiendas con target (`ventas_m24`)** y 10 sin target.
- Variables de entorno: población en distintos radios, tráfico peatonal y vehicular, competencia, número de comercios y malls, entre otras.
- Variables categóricas relevantes: `store_cat` (tipo de tienda: super, mini, express, hiper) y `malls` (si está en un centro comercial).

### Hallazgos:

- **Distribución de ventas:** sesgada positivamente, media ~2650, mediana ~2350, con outliers estructurales (>8000).
- **Outliers:** 3 tiendas (dos "super" y una "mini") con ventas muy superiores al resto. No son errores, representan hiper-tiendas o ubicaciones premium.
- **Segmentación por store_cat:**
- Express → más estables (~2000–3000).
- Mini → más variables, incluyen un outlier.
- Super → mayor dispersión, incluyen los mayores outliers.
- Hiper → pocos casos (n=4), resultados inconsistentes.
- **Malls:** tiendas en malls (solo 8 casos) muestran ventas significativamente más altas.

## 2. Análisis de correlaciones

- **Población y viviendas** → correlaciones muy altas con ventas (Pearson ~0.9, Spearman ~0.84, Kendall ~0.68). Relación lineal y monótona.
- **Tráfico (car, foot):** correlaciones moderadas (0.6 Pearson), pero más bajas en Spearman/Kendall → la relación no siempre es monótona.
- **Competencia:** correlación casi nula (Pearson ~0.07). Sola no explica las ventas.
- **Interacción población × competencia:** en zonas densas la competencia no reduce ventas; en zonas poco pobladas las ventas son bajas incluso sin competencia.

## 3. Features derivadas propuestas

Para capturar mejor la relación entre densidad, competencia y ventas:

1. **Ratios población/competencia:**
2. `pop_comp_100m = pop_100m / (competencia+1)`
3. `pop_comp_300m = pop_300m / (competencia+1)`
4. `pop_comp_500m = pop_500m / (competencia+1)`
5. `pop_total = pop_100m + pop_300m + pop_500m`
6. `pop_comp_total = pop_total / (competencia+1)`

7. *Interpretación:* mide cuánta población disponible hay por cada competidor cercano.

8. **Ratios oficinas/viviendas por competencia:**

9. `oficinas_comp = oficinas_100m / (competencia+1)`
10. `viviendas_comp = viviendas_100m / (competencia+1)`

11. *Interpretación:* densidad de viviendas/oficinas ajustada por saturación de mercado.

12. **Flag de outliers (solo en train):**

13. `is_outlier = 1` si ventas_m24 > Q3 + 1.5*IQR.
14. *Interpretación:* indica tiendas con ventas excepcionalmente altas (posibles hipermercados).

# 4. Conclusiones del EDA

- Las **ventas se explican principalmente por densidad poblacional y viviendas cercanas.**
- El **tipo de tienda y la presencia en malls** generan diferencias claras en niveles de ventas.
- La **competencia por sí sola no es predictora**, pero su interacción con población sí lo es (mercado denso absorbe más competencia).
- Existen **outliers estructurales** que no deben eliminarse, pero sí tratarse (log-transform o segmentación por categoría).
- Se generaron **features derivadas** para capturar saturación de mercado y mejorar la predicción.

---

**Próximos pasos:** 1. Implementar feature engineering en `data/clean` para train y test. 2. Validar impacto de nuevas features en modelos de predicción (Random Forest, Gradient Boosting). 3. Evaluar transformaciones (ej. log(ventas_m24)) y segmentación por `store_cat`.