Reporte Completo – Punto 1 (Estimación de Ventas Retail)

1. Contexto del Problema

Una cadena de retail desea **estimar las ventas en el mes 24 (`ventas_m24`)** de sus tiendas a partir de información **geográfica, sociodemográfica y de competencia**. El objetivo es identificar qué factores impulsan las ventas y predecir el comportamiento de nuevas tiendas.

- **Train (100 tiendas):** contiene covariables + `ventas_m24`.
- **Test (10 tiendas):** mismas covariables, sin `ventas_m24`.

Variables clave: población en distintos radios, tráfico peatonal y vehicular, competencia, viviendas, oficinas, presencia de malls y categoría de tienda.

---

2. Exploración Inicial (EDA)

Distribución de ventas

- Media ~2650, mediana ~2350 → distribución **sesgada positivamente**.
- Existen **outliers estructurales** (>8000) → tiendas con ventas extraordinarias (hipermercados o ubicaciones premium).

Outliers

- 3 casos destacados:
- Tienda_41 (super, ~6779)
- Tienda_43 (super, 10500)
- Tienda_99 (mini, ~8117)
- No son errores → deben tratarse con cuidado (log-transform, segmentación, flag).

Segmentación por `store_cat`

- **Express:** estables (2000–3000), sin outliers fuertes.
- **Mini:** variables, con un outlier grande.
- **Super:** mayor dispersión, incluyen los máximos outliers.
- **Hiper:** pocos casos (n=4), resultados inconsistentes.

Segmentación por `malls`

- Solo 8 tiendas en malls.
- **Ventas significativamente más altas** y más dispersión que las fuera de malls.
- Variable importante aunque desbalanceada.

3. Análisis de Correlaciones

Relación con `ventas_m24`

- **Población y viviendas:** correlaciones muy altas (Pearson ~0.9, Spearman ~0.84, Kendall ~0.68). Relación lineal y monótona.
- **Tráfico (car, foot):** correlaciones moderadas (Pearson ~0.6), pero más bajas en Spearman/Kendall → relación no completamente monótona.
- **Viviendas en pobreza y oficinas:** correlaciones moderadas (~0.75 Pearson).
- **Competencia:** correlación casi nula (Pearson ~0.07). No explica ventas directamente.

Interacción población × competencia

- En zonas densas, aunque haya mucha competencia, las ventas son altas.
- En zonas poco pobladas, incluso sin competencia, las ventas son bajas.
- Conclusión: **la población absorbe la competencia**.

Conclusiones de correlaciones

- Las ventas dependen principalmente de **densidad poblacional y viviendas cercanas**.
- **Tráfico** es relevante pero no suficiente por sí solo.
- **Competencia** solo aporta al considerarse en interacción con población.

---

4. Análisis Univariado y Outliers

- Confirmamos que los outliers son **estructurales y no errores**.
- Decisión: **no eliminarlos**, pero sí tratarlos con:
- Transformaciones logarítmicas de `ventas_m24`.
- Segmentación por `store_cat`.
- Creación de un **flag de outlier** para modelos.

---

5. Relación entre Variables Predictoras

- **Alta multicolinealidad** entre poblaciones y viviendas.
- **Competencia, tiendas pequeñas y comercios** se correlacionan entre sí → indicadores de saturación de mercado.
- **Tráfico peatonal y vehicular** correlacionan moderadamente (0.5) → aportan información complementaria.

Conclusión: en modelos lineales habría que reducir multicolinealidad (PCA o selección). En modelos de ensamble no es crítico.

---

6. Features Derivadas (Feature Engineering)

Para capturar mejor la interacción entre densidad y competencia, se proponen:

Ratios de población por competidor

- `pop_comp_100m = pop_100m / (competencia+1)`
- `pop_comp_300m = pop_300m / (competencia+1)`
- `pop_comp_500m = pop_500m / (competencia+1)`
- `pop_total = pop_100m + pop_300m + pop_500m`
- `pop_comp_total = pop_total / (competencia+1)`
- **Interpretación:** mide la disponibilidad de clientes potenciales por cada competidor.

Ratios de oficinas y viviendas por competencia

- `oficinas_comp = oficinas_100m / (competencia+1)`
- `viviendas_comp = viviendas_100m / (competencia+1)`
- **Interpretación:** densidad ajustada por saturación de mercado.

Flag de outliers (solo train)

- `is_outlier = 1` si ventas_m24 > Q3 + 1.5·IQR.
- **Interpretación:** identifica tiendas con ventas excepcionalmente altas.

---

7. Conclusiones Finales del EDA + Feature Engineering

1. **Determinantes de ventas:** población y viviendas son los factores clave.
2. **Categorías y malls:** influyen significativamente en la dispersión y niveles de ventas.
3. **Competencia:** no es predictor directo, pero sí al ajustarse por densidad poblacional.
4. **Outliers estructurales:** deben mantenerse y tratarse con flags o transformaciones.
5. **Features derivadas** permiten capturar saturación de mercado y mejorar el poder predictivo.
6. **Modelos recomendados:** Random Forest o Gradient Boosting → robustos a multicolinealidad y outliers.

---

8. Próximos pasos

1. Guardar datasets enriquecidos en `data/clean`.
2. Entrenar modelos comparando:
3. Regresión lineal vs Random Forest vs Gradient Boosting.
4. Con y sin features derivadas.
5. Con y sin log-transform de `ventas_m24`.
6. Evaluar importancia de variables y validar qué aportan más (originales vs derivadas).
7. Estimar ventas de las 10 tiendas nuevas.