Reporte Técnico: Análisis Exploratorio de Datos (EDA) para Máquina Expendedora de Café

Fecha: 19 de agosto de 2025 | **Estado:** EDA Completado - Listo para Fase de Modelado

---

Resumen Ejecutivo

El análisis exploratorio de datos (EDA) realizado constituye una fase fundamental en el desarrollo de un sistema predictivo para ventas de una máquina expendedora de café. El código implementa un pipeline robusto de procesamiento y análisis de datos que abarca desde la carga y validación inicial hasta la preparación de características para modelos de machine learning.

1. Calidad y Estructura de los Datos

1.1. Dataset Original

Total de registros	**Período cubierto**
3,636 transacciones	**1 marzo 2024 - 23 marzo 2025**
Productos únicos	**Calidad general**
8 tipos de café	**97.5% datos válidos**

Problemas de calidad detectados:

- 9 registros con fechas inválidas (0.25%)
- 89 valores faltantes en 'card' (2.4%)
- 3,569 valores faltantes en 'holiday_name' (98.4%)

1.2. Limpieza y Procesamiento

El proceso de limpieza eliminó 9 registros (0.25%), resultando en un dataset limpio de 3,627 transacciones con un revenue total de 115,144.90 UAH .

2. Análisis de Outliers y Distribución

2.1. Distribución de Montos

Métrica	Valor (UAH)
Media	31.75
Mediana	32.82
Desviación estándar	4.92
Sesgo	-0.52 (ligera asimetría izquierda)
Curtosis	-0.66 (distribución platicúrtica)

2.2. Detección de Outliers

- **Método IQR:** 0 outliers detectados
- **Método Z-Score:** 0 outliers detectados
- **Método percentil:** 273 outliers (7.5% de los datos)

3. Patrones Temporales y Estacionalidad

3.1. Revenue por Período

Promedio diario
296.77 UAH

Promedio semanal
2,056.16 UAH

Promedio mensual
8,857.30 UAH

Coeficiente de variación
0.55 (moderada-alta)

3.2. Patrones Horarios

- **Horas pico:** 13:00, 12:00, 18:00
- **Horas de menor actividad:** 8:00, 9:00, 1:00

Distribución por período:

- **Mañana:** Cortado, Americano con Leche, Americano
- **Tarde:** Americano, Cortado, Espresso
- **Noche:** Chocolate Caliente, Cocoa, Cappuccino

3.3. Patrones por Día de la Semana

- **Día más activo:** Martes
- **Día menos activo:** Domingo
- **Distribución consistente** con patrones de consumo de oficina

4. Análisis de Productos (Pareto)

4.1. Concentración de Revenue

- **Principio 80/20:** 4 productos (50%) generan el 74.7% del revenue
- **Producto más rentable:** Latte
- **Producto menos rentable:** Espresso
- **Coeficiente de Gini:** 0.32 (concentración moderada)

4.2. Correlación entre Productos

Análisis de correlación muestra patrones de consumo complementarios:

- Americano y Espresso: alta correlación (0.90)
- Cappuccino y Cocoa: alta correlación (0.87)
- Grupos de productos con patrones de consumo similares

5. Cobertura Temporal y Gaps

5.1. Continuidad de Datos

- **Total de gaps significativos:** 1,017
- **Gap máximo:** 65.03 horas
- **Horario operativo:** 0:00 - 23:00 (cobertura completa)

5.2. Estabilidad Operativa

Métrica	Valor
Coeficiente de variación diaria	0.53
Mejor día	796.00 UAH
Peor día	23.02 UAH
Días outliers	16 (4.2% del período)

6. Preparación para Machine Learning

6.1. Feature Engineering

Total de características

38

Características numéricas

29

Características categóricas

9

Tipos de features:

- Temporales (fecha, hora, día de semana)
- Cíclicas (seno/coseno para periodicidad)
- Lag features (valores históricos)
- Rolling statistics (medias móviles)
- Market share por producto

6.2. Calidad para Modelado

- **Balance de la variable objetivo:**
  - Transacciones promedio/día: 1.17
  - Días sin ventas: 44.7%
  - Desviación estándar: 1.52
- **Problemas de correlación:** 6 pares de features con alta correlación (>0.95)
- **Varianza:** 0 features con varianza cero

7. Recomendaciones Operativas

7.1. Gestión de Inventario

- Priorizar stock para: Americano con Leche, Latte, Americano
- Revisar viabilidad de: Espresso, Cocoa
- Sistema de alerta con stock mínimo de 453 unidades

7.2. Optimización de Horarios

- Mantenimiento en horas de baja actividad: 8:00, 9:00
- Máxima atención en horas pico: 13:00, 12:00, 18:00

7.3. Estrategia Comercial

- Alta concentración de revenue (74.7% en 4 productos)
- Baja variación de precios (oportunidad para estrategias diferenciales)
- Patrones de consumo predecibles por hora y día

8. Próximos Pasos para Modelado Predictivo

8.1. Preparación de Modelos

- **Validación temporal** (no k-fold aleatorio)
- **Modelos especializados:** Zero-Inflated o Hurdle para manejar ceros
- **Ensemble methods:** XGBoost, Random Forest
- **Feature selection** para reducir dimensionalidad

8.2. Integración de Datos Externos

- Datos climáticos (ya disponibles en el dataset)
- Información de festivos (necesita mejorarse)
- Eventos especiales o temporadas

8.3. Sistema de Alertas

- Detección de anomalías en tiempo real
- Alertas de mantenimiento predictivo
- Optimización de reposición de inventario

9. Conclusiones Técnicas

El EDA realizado demuestra un pipeline robusto de procesamiento de datos con:

1. **Manejo robusto de errores** en la carga y conversión de datos
2. **Análisis comprehensivo** de outliers y distribución
3. **Detección efectiva** de patrones temporales y de consumo
4. **Feature engineering avanzado** para modelado predictivo
5. **Sistema de validación** de calidad de datos integral

Los datos presentan calidad suficiente para avanzar a la fase de modelado predictivo, con algunas consideraciones especiales para el manejo de la alta proporción de días sin ventas y la estacionalidad marcada.

10. Archivos Generados

El proceso generó múltiples recursos para la siguiente fase:

- Dataset limpio (`coffee_clean_dataset.csv`)
- Features para ML (`coffee_ml_features.csv`)
- Metadatos del procesamiento (`datasets_metadata.json`)
- Gráficos de análisis en formato PNG
- Reporte completo de recomendaciones operativas
