Conociendo al Cliente 360°: Datos, Opiniones y Tendencias

Nombre de la autora: Vanina A. Cavallin.

Email: vaninacavallin@gmail.com

Cohorte: PT01.

Fecha de entrega: 04/09/2025.

INTRODUCCIÓN

El presente informe documenta el proceso de análisis realizado a partir de múltiples fuentes de datos, con el fin de extraer insights relevantes sobre los hábitos de consumo en restaurantes en Estados Unidos, y en particular en la ciudad de Chicago.

Se trabajó con dos fuentes principales:

- Base de usuarios (30.000 registros) con datos demográficos, socioeconómicos y de consumo.
- Datos de la API de Yelp con información de restaurantes en Chicago (categorías, reseñas, precios, ubicación y calificaciones).

Objetivos del PI:

→ Explorar y comprender estructuras de datos reales provenientes de fuentes diversas.

→ Aplicar técnicas de limpieza y transformación de datos para preparar datasets para análisis.

→ Utilizar librerías de Python como pandas, numpy, matplotlib y seaborn para realizar análisis exploratorio de datos (EDA).

→ Conectar con una API externa (Yelp) para enriquecer un dataset con información contextual e implementar técnicas de web scraping para extraer datos relevantes de páginas web.

→ Interpretar los resultados del análisis para generar recomendaciones accionables.

→ Documentar el proceso de análisis de forma clara, estructurada y reproducible.

Detalle de la consigna:

Se hará una selección técnica para el rol de Científico de Datos Junior en Insight Reach. Dicha empresa tiene por objetivo optimizar su estrategia de segmentación para mejorar la efectividad de sus campañas.

Para ello, se ha diseñado un reto técnico dirigido a candidatos para el rol de Científico de Datos Junior, cuyo objetivo es demostrar su capacidad para integrar múltiples fuentes de datos, analizarlas y generar insights accionables.
El desafío es demostrar tu capacidad para analizar, enriquecer e interpretar bases de datos complejas para generar recomendaciones accionables. El rol simula el trabajo de un profesional en un entorno real, donde se deberá construir análisis sólidos, justificar decisiones y comunicar resultados de manera clara y estructurada.

Dataset inicial proporcionado por Henry (input): base_datos_restaurantes_USA_v2.csv

Datasets generados a partir del procesamiento del dataset inicial (outputs):

- chicago_restaurantes.csv (Datos crudos de API - Avance_API_YELP)
- Datos_YELP_limpios_VC.csv (Datos limpios de API - Avance_API_YELP)
- df_restaurante_limpio.csv (Dataset limpio - Avance_EDA)
- usuarios_restaurante_chicago_limpio.csv (Datos por ciudad - Avance_EDA)

Librerías Principales Utilizadas:

- Pandas: Manipulación y análisis de datos.
- Numpy: Operaciones numéricas.
- Matplotlib/seaborn: Visualización de datos.
- FuzzyWuzzy: Análisis de similitud de strings.

Paso a paso del desarrollo de las distintas instancias del PI

Avance 1: Conexión y limpieza de datos

Input: base_datos_restaurantes_USA_v2.csv

Outputs:

- df_restaurante_limpio.csv (Dataset limpio - Avance_EDA)
- usuarios_restaurante_chicago_limpio.csv (Datos por ciudad - Avance_EDA)

Descripción del proceso

Se comenzó con la carga y exploración inicial del dataset Se comenzó con la carga y exploración inicial del dataset base_datos_restaurantes_USA_v2.csv. Para ello, se importaron las librerías necesarias para la manipulación y visualización de datos (pandas, numpy, matplotlib, seaborn) y se cargó el archivo en un DataFrame. Posteriormente, se revisaron las primeras filas y la estructura general, identificando que el dataset contiene 30.000 registros y 17 columnas, con información demográfica, socioeconómica y de consumo en restaurantes.

A continuación, se clasificaron las variables según su tipo: 12 categóricas (nombre, género, ciudad_residencia, preferencias alimenticias, etc.) y 5 numéricas (edad, frecuencia de visita, gasto promedio en comida, ingresos mensuales, id_persona).

Se realizó un análisis de calidad de los datos, detectando inconsistencias como las que se detallan a continuación:

- Edad: se encontraron 207 valores fuera de rango (<18 o >90).
- Frecuencia de visita: presencia de valores anómalos (ej. negativos como -3).
- Promedio de gasto en comida: casos con valores nulos o iguales a 0.
- Preferencias alimenticias, correo electrónico y teléfono: contienen datos faltantes en proporción considerable.
- Ingresos mensuales: dentro de un rango lógico (800 a 18.000 USD), aunque con gran dispersión.

Limpieza de datos

- En la variable edad, los valores fuera de rango fueron reemplazados con la mediana de la columna, normalizando así la distribución.
- Se verificaron las estadísticas descriptivas, confirmando que los valores extremos fueron corregidos (min = 18, max = 80).

Conclusiones preliminares

El dataset requiere un proceso adicional de normalización para ser usado en análisis avanzados. Las principales variables con inconsistencias son:

1. Edad: valores fuera de rango (ya corregidos).
2. Frecuencia de visita: casos negativos que deben ser revisados.

3. Promedio de gasto en comida: valores nulos o cero.
4. Preferencias alimenticias: datos faltantes.
5. Teléfono y correo electrónico: alta proporción de nulos.

Con estas correcciones iniciales, se avanzó hacia dos datasets más limpios y consistentes, apto para posteriores análisis exploratorios, segmentación y enriquecimiento con fuentes externas (ej. API de Yelp). Uno de ellos general (df_restaurante_limpio.csv) mientras que el otro se enfocó en la ciudad de Chicago (usuarios_restaurante_Chicago_limpio.csv).

Avance 2: Conexión y limpieza de datos

Input: API de Yelp.

Outputs:

- chicago_restaurantes.csv (Datos crudos de API - Avance_API_YELP)
- Datos_YELP_limpios_VC.csv (Datos limpios de API - Avance_API_YELP)

Descripción del proceso

En esta segunda etapa se trabajó con la conexión a la API de Yelp con el objetivo de enriquecer la información del proyecto mediante datos reales de restaurantes.

Se comenzó con la carga de librerías necesarias (pandas, numpy, requests, ast) y la configuración de las credenciales proporcionadas (API Key e ID de cliente). A partir de estas credenciales se definieron los parámetros de consulta, entre ellos: ciudad de análisis (Chicago), término de búsqueda (“restaurants”) y límite de resultados por solicitud (50).

Posteriormente se enviaron las solicitudes HTTP a la API y se recibieron las respuestas en formato JSON. Estos datos fueron transformados en estructuras de Python (diccionarios/listas) y luego en un DataFrame de pandas para su manipulación y análisis.

El dataset resultante incluyó variables como:

- Nombre y ubicación del restaurante.
- Categoría gastronómica.

- Rango de precios.
- Calificación promedio (rating).
- Cantidad de reseñas (reviews).
- Coordenadas geográficas (latitud y longitud).

Se realizó una primera validación de la calidad de los datos, observando que:

- Las categorías y ratings se encontraban completos y consistentes.
- El campo de precios presentó valores faltantes en algunos restaurantes.
- La cantidad de reseñas varía mucho entre negocios (desde muy pocas hasta miles).

Este proceso permitió construir un dataset enriquecido de la oferta gastronómica de Chicago, complementando la información de los usuarios analizada en el **Avance 1**.
Conclusiones preliminares:

- La API de Yelp brinda datos actualizados y relevantes sobre la oferta gastronómica, lo que posibilita vincular hábitos de consumo de usuarios con la oferta real de restaurantes en una ciudad determinada.
- El dataset de Yelp aporta variables clave (categorías, calificación, reseñas, precios) que permiten segmentar la oferta y analizar su alineación con las preferencias de los usuarios.
- Se detecta la necesidad de integrar esta información con el dataset de clientes para generar análisis comparativos y detectar oportunidades de negocio.

Avance 3: Conexión y limpieza de datos

Input: Datasets procesados de Avance EDA y API YELP.
Outputs: Análisis, visualizaciones y Recomendaciones.pdf

Descripción del proceso

En esta tercera etapa se avanzó con la integración de los datasets previamente trabajados:

1. La base de usuarios depurada en el Avance 1.

2. La base de restaurantes de Chicago obtenida mediante la API de Yelp en el Avance 2.

El propósito central fue combinar ambas fuentes para identificar coincidencias, tendencias y oportunidades de negocio a partir de la relación entre la demanda de los clientes y la oferta gastronómica disponible en la ciudad.

Se realizaron diversas visualizaciones y análisis comparativos:

- Perfil demográfico y socioeconómico de los usuarios: distribución de edad, ingresos y estrato socioeconómico.
- Preferencias alimenticias: cruce entre edad, género y tipo de cocina (vegetariano, vegano, carnes, mariscos).
- Oferta gastronómica Yelp: categorías predominantes, rangos de precios, ratings y número de reseñas.
- Relación oferta–demanda: comparación de zonas y categorías gastronómicas más buscadas por los usuarios frente a la disponibilidad de restaurantes en Chicago.

Hallazgos principales

- Los usuarios con ingresos altos y membresía premium presentan mayor frecuencia de visitas y prefieren restaurantes con precios altos y buenas calificaciones.
- Los usuarios jóvenes (<35 años) tienden a consumir en establecimientos de tipo “fast casual” y de menor precio, aunque gastan más por visita en relación a otros grupos.
- Los clientes vegetarianos y veganos se concentran en ciertos barrios, los mismos donde Yelp registra mayor cantidad de restaurantes de esas categorías.
- Se detectaron áreas donde existe alta demanda de ciertas preferencias alimenticias (ej. mariscos, vegano) pero baja oferta en Yelp, lo que sugiere oportunidades para expansión o alianzas comerciales.

Conclusiones preliminares

- La integración de ambas fuentes permitió identificar patrones valiosos de segmentación y detectar desajustes entre la oferta y la demanda gastronómica en Chicago.
- Los resultados obtenidos pueden orientar campañas de marketing más precisas, diseñadas en función de variables como edad, ingresos y preferencias alimenticias.
- Se abren oportunidades de negocio tanto en la optimización de campañas publicitarias como en la expansión estratégica de restaurantes en zonas específicas.

Resultados por Avance

Avance 1: Análisis Exploratorio de Usuarios

Dataset procesado: 30,000 → 28,537 registros limpios.

- Valores nulos imputados inteligentemente por ciudad y estrato.
- Rangos de edad normalizados (18-90 años).
- Tipos de datos estandarizados.
- Ciudad foco seleccionada: Chicago.

Archivos generados:

- df_restaurante_limpio.csv (Dataset completo limpio).
- usuarios_restaurante_chicago_limpio.csv (Dataset completo limpio con sólo usuarios de Chicago).

Avance 2: API YELP - Datos de Restaurantes

Dataset obtenido: 200 restaurantes de Chicago.

- Conexión exitosa a API de Yelp.
- Análisis de similitud con FuzzyWuzzy (>70%).
- Imputación contextual de precios por categoría.
- Estructuras JSON normalizadas.

Archivos generados:

- chicago_restaurantes.csv (Datos crudos de la API).
- Datos_YELP_limpios_VC.csv (Datos limpios de la API).

Avance Análisis Final: Sistema de Recomendaciones

- Integración de datasets de usuarios y restaurantes.
- Sistema de recomendaciones personalizado.
- Desarrollo de visualizaciones para análisis.

Recomendaciones Estratégicas

Archivo generado: Recomendaciones.pdf (propuestas de segmentación).

- Insights más valiosos identificados.
- Recomendaciones estratégicas para InsightReach.
- Conclusiones basadas en un análisis integral.
- Propuestas para campañas de marketing dirigidas.

Características técnicas destacadas

En cuanto a las metodologías implementadas:

- Exploración y Limpieza de Datos (EDA): detección y corrección de valores fuera de rango en variables críticas como edad y frecuencia_visita.
- Imputación de valores faltantes: reemplazo de edades inválidas con la mediana para mantener la consistencia de la distribución.
- Clasificación de variables: diferenciación entre categóricas y numéricas para definir el tipo de tratamiento a aplicar en limpieza y análisis.
- Integración Multifuente: combinación de datos de usuarios (dataset base) con la API de Yelp para enriquecer información con oferta gastronómica real.
- Validación de calidad: revisión de duplicados, rangos lógicos y presencia de valores nulos en variables clave.

En cuanto a la calidad de los datos alcanzada:

- Completitud: las variables críticas (edad, frecuencia_visita, promedio_gasto_comida) fueron corregidas para evitar distorsiones.
- Consistencia: los tipos de datos fueron revisados y estandarizados para su posterior análisis y visualización.
- Validez: se establecieron rangos realistas en las principales métricas (edad entre 18 y 90 años; ingresos entre 800 y 18.000 USD).
- Representatividad: integración con Yelp permitió contextualizar los datos de usuarios en función de la oferta real de restaurantes.

Próximos desarrollos propuestos:

- Modelos de Machine Learning:
    - Segmentación de clientes (clustering).
    - Modelos predictivos para gasto promedio y frecuencia de visita.
    - Sistemas de recomendación de restaurantes basados en perfil socioeconómico + preferencias.
- Dashboard Interactivo: visualización dinámica de insights mediante librerías como Plotly/Dash o Power BI.
- API interna de recomendaciones: servicio que conecte usuarios y restaurantes en tiempo real.
- Framework de validación (A/B Testing): para evaluar impacto de campañas de marketing segmentadas.

Soporte y Contacto

Autora principal: Vanina, A. Cavallin.
Email: vaninacavallin@gmail.com
LinkedIn: Vanina Cavallin.