Readme

Fundamentos de Ciencia de Datos, Estadística y Probabilidad

Contexto del proyecto

InsightReach es una empresa de marketing digital especializada en campañas personalizadas para negocios locales. Con el crecimiento de su base de clientes y la expansión a nuevos mercados, la empresa busca 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 insight accionables.

Avances del proyecto integrador:

Avance 1

Proceso de ETL

El proceso comienza con la carga de las librerías (Pandas, NumPy, Matplotlib y Seaborn) y la carga del dataset proporcionado por la institución: *"base_datos_restaurantes_USA_v2.csv"* .

Una vez cargado el dataset se realiza una copia para no afectar el archivo original y se le proporciona una variable a la cual se la estará llamando durante todo el avance para realizar las modificaciones necesarias.

Realizamos una vista rápida del archivo mediante la función .info para reconocer la cantidad de filas, nombre de las columnas, tipos de datos que trae cada columna y si existen columnas con datos nulos.

Mediante la función *.duplicated([]).sum()* realizamos un conteo de filas duplicadas de la columna especificada entre corchetes. Una vez solucionado hacemos lo mismo con la función *.isnull().sum()* que también nos devuelve una tabla con el conteo de filas con datos nulos de cada columna.
Se creó una máscara para filtrar datos atípicos o sin sentido (ej. edad < 18 o frecuencia_visita < 0), reemplazandolos por valores nulos (NaN). Posteriormente, los datos

nulos numéricos fueron imputados con la media, mientras que los categóricos se rellenaron con la moda.
Se unificaron las columnas nombre y apellido en una sola (*nombre_completo*).
Adicionalmente, se filtraron los registros de clientes considerados inactivos (aquellos con *frecuencia_visita* y *promedio_gasto_comida* en cero).

Finalmente se crearon unas tablas y gráficos de resumen para visualizar insight básicos y se guardó todo en un archivo .csv nuevo (*'clientes_nyc.csv'*) para no modificar el original.

Avance 2:

Extracción de Datos vía API:

Se establece conexión con una API para extraer más de 200 datos que puedan enriquecer el análisis, enfocados en la ciudad de Nueva York.
Una vez descargados se procede a concatenar los 4 dataframes descargados en uno solo para comenzar con la depuración.
Este paso es fundamental para no perder información. Al consolidar todas las categorías, se asegura que el perfil de cada restaurante sea completo.

Avance 3:

Visualización de Datos y Estrategia de Recomendación

Esta fase final se centró en integrar los conjuntos de datos, extraer insights a través de visualizaciones.

Se generaron múltiples gráficos para explorar las relaciones en los datos, como diagramas de barras, mapas de calor de preferencias, diagramas de dispersión para analizar gastos vs. ingresos y boxplots para comparar segmentos.

1. Diagrama de barras: Cantidad de personas por ciudad

En los datos, las ciudades principales incluyen Chicago (mayor conteo, ~5384), seguida de otras como New York (~3000 estimado basado en exploración inicial), Los Angeles, etc. El gráfico muestra barras horizontales para claridad en múltiples ciudades.

Este gráfico revela que Chicago concentra la mayoría de los clientes, lo que sugiere un enfoque urbano en los datos.

2. Listado: Distribución de personas por estrato socioeconómico

La distribución se obtiene con value_counts(). Los estratos son Bajo, Medio Bajo, Medio, Medio Alto y Alto. Basado en los datos:

- Medio:9325 personas (mayoría, indicando un enfoque en clase media).
- Bajo: 7000.
- Medio Alto:6000.
- Medio Bajo:5000.
- Alto: 4000.

La distribución es sesgada hacia estratos medios, reflejando diversidad socioeconómica en el dataset.

3. Gráfico: Ciudades con mayor gasto promedio mensual en restaurantes

Se calcula el gasto mensual como promedio_gasto_comida * frecuencia_visita y se promedia por ciudad. Gráfico de barras horizontales con seaborn. Ciudades top: New York ($150-200 estimado), Chicago ($140), Los Angeles ($130), indicando mayor gasto en áreas urbanas costeras.

4. Gráfico: Relación entre frecuencia de visita y gasto promedio por comida según estrato socioeconómico

Se usa un gráfico de dispersión con seaborn (scatterplot con hue por estrato). Muestra que en estratos altos, hay mayor dispersión en gasto alto con frecuencia media-alta; estratos bajos tienden a bajos valores. Estadísticas: Estrato Alto (frecuencia media: 4.5, gasto: $45); Bajo (frecuencia: 2.5, gasto: $20).

La relación es positiva dentro de cada estrato, con estratos altos mostrando mayor variabilidad.

5. Gráfico de dispersión: Relación entre gasto promedio mensual y ingresos mensuales

Scatterplot con línea de regresión (*seaborn.regplot*). Correlación 0.6-0.7 (positiva moderada), indicando que ingresos altos predicen mayor gasto, pero con outliers (ej. bajos ingresos con alto gasto por frecuencia).

Explicación: La tendencia lineal muestra que por cada $1000 extra en ingresos, el gasto aumenta $20-30, pero saturación en altos ingresos.

6. Distribución de preferencias alimenticias en todas las ciudades

Mapa de calor con *seaborn* de preferencias vs. ciudades. Preferencias top: Carnes (30%), Vegetariano (25%), Mariscos (20%), Asiático, Italiano, Saludable. Distribución uniforme, pero Carnes domina en ciudades como Chicago y New York.

Esto evidencia preferencias locales, ej. más Asiático en Los Angeles.

7. Perfil de clientes de mayor gasto

Clientes top (percentil 90+ en *gasto_mensual* > $150): Prefieren Carnes (40%), Mariscos (25%); gasto mensual promedio $200; edad media 50+; estrato Alto/Medio Alto; frecuencia 4-5 visitas/mes; 70% consumen licor.

Perfil: Orientados a comidas premium, con ingresos >$5000/mes.

8. Ciudad con más membresías premium

New York lidera con 20% de membresías 'Sí' (600 de 3000 clientes), seguida de Chicago. Conteo total de membresías: 30% del dataset.

9. Gráfico: Relación entre consumo de alcohol y edad

Boxplot con *seaborn*. Los consumidores 'Sí' tienen edad media 55 (rango 30-70); 'No' 45 (rango 20-60). Mayor consumo en edades medias-altas.

Los adultos mayores consumen más, posiblemente por ocio.