Proyecto Final

Variabilidad y volatilidad de precios alimentarios por mercado en Colombia y su impacto en el costo de vida.

Problema y Estrategia

Definición del problema

El comportamiento de los precios de los alimentos en Colombia es altamente variable debido a factores económicos, climáticos, logísticos y estacionales. Estos precios fluctúan entre mercados, ciudades y regiones, generando incertidumbre tanto para los consumidores como para los actores de la cadena productiva y comercial. Esta variabilidad incluye incrementos abruptos, disminuciones inesperadas y diferencias significativas entre plazas de mercado que no siempre se explican de manera transparente.

Actualmente, los datos disponibles sobre precios están dispersos en múltiples fuentes, muchas de ellas en formatos heterogéneos, con periodicidades distintas y en ocasiones con problemas de calidad (faltantes, duplicados, inconsistencia entre mercados). Esto genera una dificultad adicional para quienes necesitan obtener una visión clara, comparable y actualizada del comportamiento del mercado alimentario.

Necesidad del negocio

Las organizaciones públicas, privadas y sociales requieren información clara, confiable y actualizada para tomar decisiones relacionadas con el abastecimiento y la gestión de precios de alimentos. La volatilidad en los precios afecta directamente la estabilidad económica y operativa de múltiples actores, desde hogares y cadenas minoristas hasta productores, distribuidores y entidades encargadas de monitorear la seguridad alimentaria.

A nivel institucional, se requiere una solución que consolide datos provenientes de fuentes oficiales y que permita:

- Generar estadísticas confiables para la toma de decisiones.
- Apoyar el diseño de políticas públicas basadas en evidencia.
- Anticipar periodos críticos de desabastecimiento,
- Evaluar el impacto de cambios climáticos, logísticos o económicos.
- Comunicar información al público de forma transparente.

Para la industria y el sector logístico, existe la necesidad de:

- Optimizar procesos de compra y abastecimiento.
- Ajustar inventarios según tendencias y estacionalidad.
- Negociar de manera más estratégica con proveedores,

- Gestionar riesgos asociados a productos críticos con alta volatilidad,
- Reducir pérdidas económicas por decisiones tardías o mal informadas.

Asimismo, los consumidores se ven afectados por fluctuaciones constantes del costo de vida, y una herramienta analítica puede ayudar a mejorar la planificación financiera y el acceso a información sobre precios en diferentes mercados.

En general, el negocio requiere una solución que:

1. **Integre datos dispersos** en un solo pipeline reproducible.
2. **Genere insights accionables** que permitan anticiparse, no reaccionar tarde.
3. **Permita comparar mercados** en tiempo real o con frecuencia actualizada.
4. **Identifique productos críticos** con base en métricas objetivas.
5. **Ofrezca visualizaciones claras** y de fácil consumo para perfiles técnicos y no técnicos.
6. **Aumente la eficiencia operativa** mediante decisiones basadas en datos.

Con esta solución, se logra una visión integral del comportamiento de los precios, fortaleciendo la capacidad de respuesta de las organizaciones y permitiendo un manejo más inteligente de riesgos asociados a la cadena alimentaria.

Estrategia Corporativa

La estrategia corporativa se basa en aprovechar datos abiertos de precios de alimentos para mejorar la toma de decisiones, optimizar procesos y anticipar variaciones que afecten el costo de vida. La iniciativa fortalece la capacidad de la organización para responder de forma rápida y eficiente a cambios en el mercado.

Se articula en tres ejes principales:

1. Decisiones basadas en datos

Transformar datos dispersos en información estructurada permite identificar tendencias, comparar mercados y detectar productos con alta volatilidad. Esto facilita decisiones oportunas y mejor fundamentadas.

2. Optimización operativa

El análisis continuo y las visualizaciones dinámicas ayudan a mejorar procesos como abastecimiento, compras, planificación logística e inventarios, reduciendo riesgos y costos innecesarios.

3. Transparencia y trazabilidad

El uso de pipelines reproducibles, datos abiertos y documentación en un repositorio colaborativo permite garantizar calidad de información, trazabilidad del proceso y acceso claro para todos los involucrados.

Propuesta inicial del modelado de procesos (BPM)

Descripción general del proceso.

El proceso propuesto consiste en transformar datos abiertos de precios de alimentos en información clara, analítica y útil para apoyar la toma de decisiones. Inicia con la obtención de los datos desde fuentes oficiales y continúa con una etapa de limpieza y normalización que garantiza su calidad y consistencia.

Posteriormente, los datos procesados se analizan de manera exploratoria para identificar tendencias, variaciones entre mercados y comportamientos relevantes. Con esta base, se generan indicadores clave que sintetizan la información y permiten evaluar volatilidad, cambios temporales y diferencias regionales.

Finalmente, los resultados se presentan mediante visualizaciones y dashboards que facilitan la interpretación y la comunicación efectiva de hallazgos. Este flujo asegura un paso a paso organizado, reproducible y alineado con los objetivos estratégicos del proyecto.

Objetivo del proceso

Transformar datos abiertos de precios en indicadores confiables, análisis exploratorio y resultados visuales que permitan comparar mercados y detectar tendencias relevantes.

Etapas del proceso

1. Ingesta y consolidación de datos abiertos

Tareas:

- Conexión a fuente oficial (SIPSA – DANE).
- Descarga manual o automática del archivo CSV.
- Validación de estructura (columnas: Fecha, Producto, Mercado, Precio).
- Verificación de tipos de datos y delimitadores.

Actor: Data Engineer.
Artefacto: dataset crudo (raw_data).

2. Limpieza y normalización

Tareas:

- Conversión de fechas a formato estándar.
- Eliminación de valores nulos, duplicados o inconsistentes.
- Corrección de nombres de mercados/productos.
- Homologación de unidades (kg).
- Generación de dataset limpio (clean_data).

Actor: Data Analyst.
Artefacto: dataset procesado (processed_data).

3. Transformación y agregación

Tareas:

- Agrupación por períodos (semanal/mensual).
- Cálculo de métricas clave (promedio, variación porcentual, volatilidad).
- Estandarización de columnas y estructura final.
- Preparación de tableros analíticos.

Actor: Data Analyst / Data Scientist.
Artefacto: dataset normalizado (normalized_data).

4. Análisis exploratorio (EDA)

Tareas:

- Visualización de tendencias por producto y mercado.
- Comparación entre regiones.
- Identificación de patrones estacionales o anómalos.
- Generación de insights preliminares.

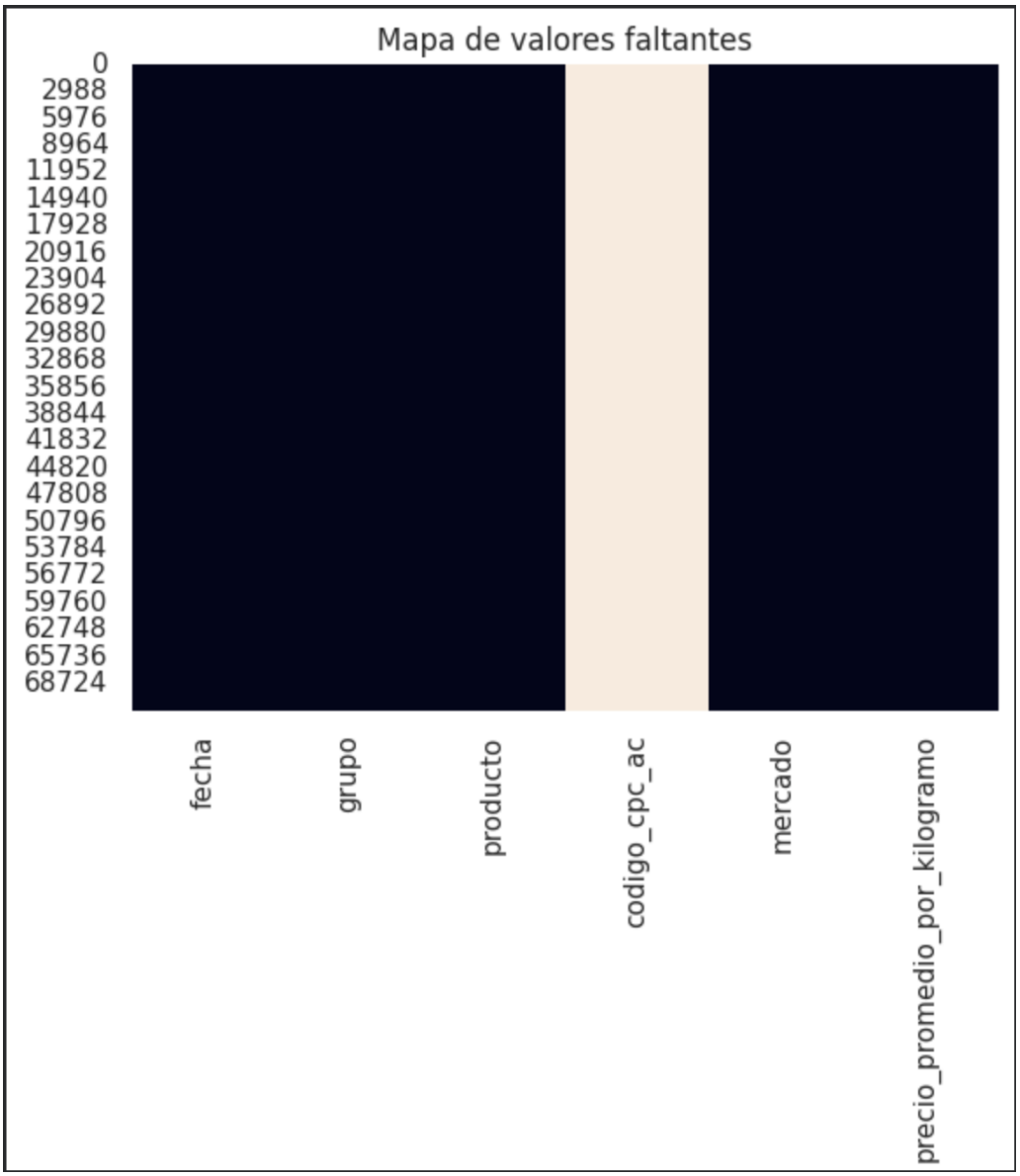
Actor: Data Scientist.
Artefacto: reporte exploratorio (eda_report).

```
df = pd.read_csv("data/processed/sipsa_master.csv", parse_dates=["fecha"])
df.head()
```

	fecha	grupo	producto	codigo_cpc_ac	mercado	precio_promedio_por_kilogramo
0	2022-02-01	PROCESADOS	ACEITE DE PALMA	NaN	ARMENIA, MERCAR	7204.0
1	2022-02-01	PROCESADOS	ACEITE DE PALMA	NaN	MONTERÍA, MERCADO DEL SUR	7234.0
2	2022-02-01	PROCESADOS	ACEITE GIRASOL	NaN	BARRANQUILLA, BARRANQUILLITA	15163.0
3	2022-02-01	PROCESADOS	ACEITE GIRASOL	NaN	BARRANQUILLA, GRANABASTOS	15985.0
4	2022-02-01	PROCESADOS	ACEITE GIRASOL	NaN	CARTAGENA, BAZURTO	12299.0

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 71698 entries, 0 to 71697
Data columns (total 6 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   fecha                         71698 non-null  datetime64[ns]
 1   grupo                         71698 non-null  object
 2   producto                      71698 non-null  object
 3   codigo_cpc_ac                 0 non-null      float64
 4   mercado                       71698 non-null  object
 5   precio_promedio_por_kilogramo  71698 non-null  float64
dtypes: datetime64[ns](1), float64(2), object(3)
memory usage: 3.3+ MB
```

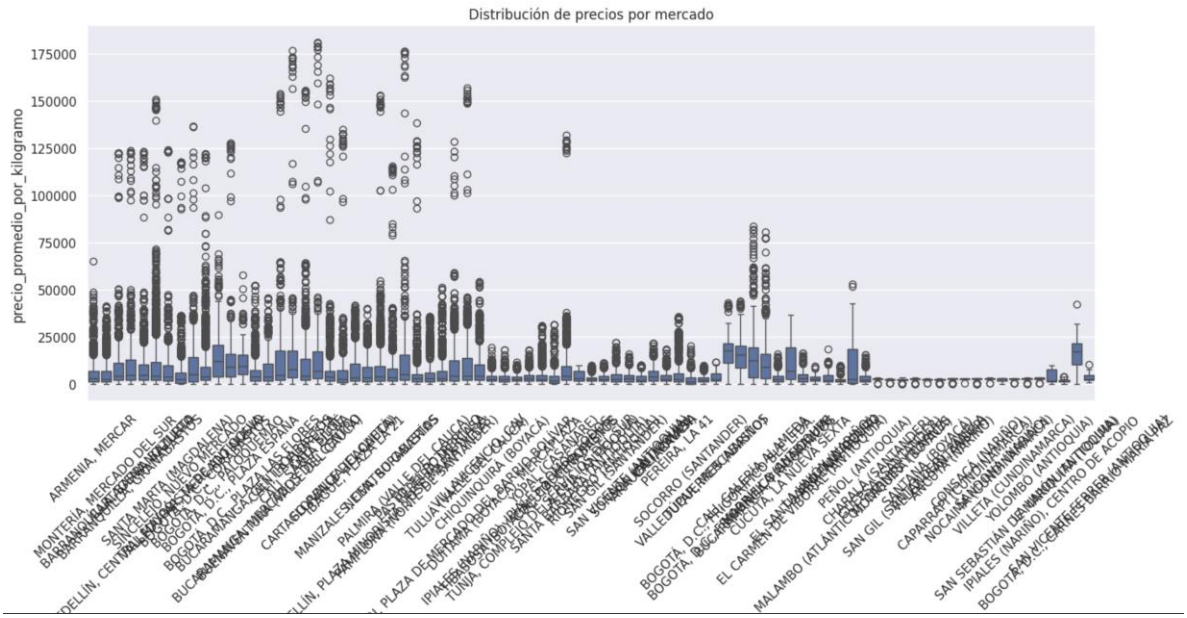


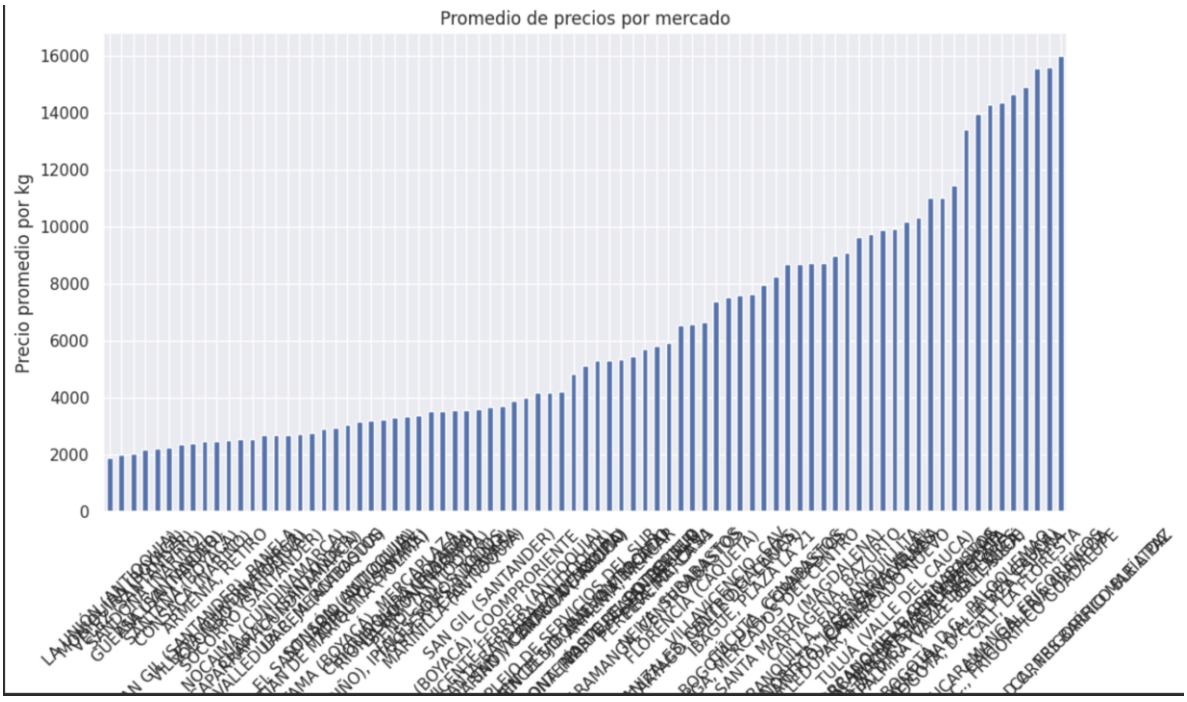
5. Generación de indicadores

Tareas:

- Cálculo de volatilidad por producto.
- Construcción de índice de costo de vida alimentario.
- Ranking de productos críticos.
- Consolidación de métricas para dashboard.

Actor: Data Scientist.
Artefacto: tabla de indicadores (metrics_table).


Distribución de precios por mercado
precio_promedio_por_kilogramo
175000
150000
125000
100000
75000
50000
25000
0

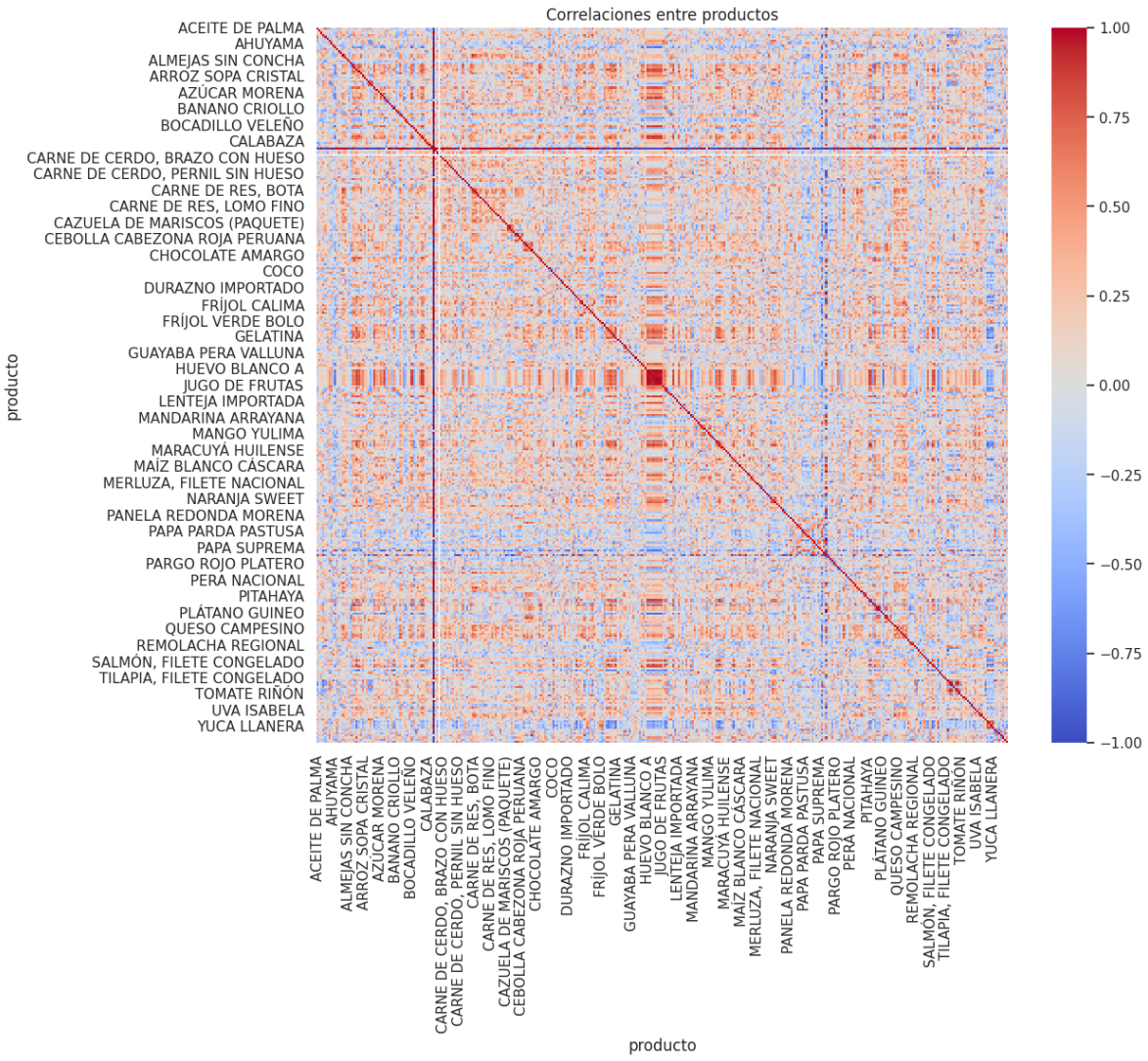

Promedio de precios por mercado
Precio promedio por kg
16000
14000
12000
10000
8000
6000
4000
2000
0

```
top_expensive
```

producto	precio_promedio_por_kilogramo
CAFÉ INSTANTÁNEO	115342.660661
LANGOSTINO U12	64857.935484
CORVINA, FILETE CONGELADO NACIONAL	44768.269231
LANGOSTINO 16-20	40357.080000
SALMÓN, FILETE CONGELADO	35979.701493
MERLUZA, FILETE NACIONAL	35290.026316
SOPA DE POLLO (CAJA)	34602.705882
LECHE EN POLVO	34466.031161
LOMITOS DE ATÚN EN LATA	27011.899204
GELATINA	26579.092166

```
top_cheap
```

producto	precio_promedio_por_kilogramo
CAPAZ MAGDALENA FRESCO	180.000000
PAPA TOCARREÑA	1006.714286
MAÍZ AMARILLO CÁSCARA IMPORTADO	1290.000000
RABADILLAS DE POLLO	1346.000000
PAPA R-12 NEGRA	1462.432432
SAL YODADA	1543.036053
CALAMAR MORADO ENTERO	1570.000000
BADEA	1595.818182
BANANO BOCADILLO	1621.825000
ZANAHORIA BOGOTANA	1673.759124



6. Visualización y presentación de resultados

Tareas:

- Diseño de dashboard interactivo (Streamlit / Power BI / Tableau).
- Integración de gráficos y mapas.
- Exportación de reportes e insights.

Actor: Data Visualization Specialist / Data Analyst.
Artefacto: dashboard final (dashboard_app).

Roles principales

Rol	Responsabilidades
Data Engineer	- Conectar con fuentes oficiales (SIPSA–DANE) - Descargar y validar estructura de datos - Organizar carpetas y estructura del proyecto - Preparar scripts de ingesta
Data Analyst	- Limpieza y normalización del dataset - Estandarización de campos y unidades - Elaboración de tablas procesadas - Preparación de datos para EDA
Data Scientist	- Análisis exploratorio (EDA) - Cálculo de métricas (variación, volatilidad, índices) - Modelos de predicción - Generación de insights
Data Visualization Specialist (o Data Analyst)	- Diseño del dashboard final - Integración de gráficas e indicadores - Presentación y narrativa visual
Project Lead / Coordinador	- Asegurar coherencia entre etapas - Documentación en GitHub - Revisión de calidad - Preparación del pitch final

Búsqueda y selección de datos

Identificación de fuentes potenciales

Para iniciar el proyecto, se realizó una búsqueda exhaustiva de fuentes de datos abiertas que permitieran analizar la variación de precios de alimentos en Colombia. El objetivo fue encontrar un conjunto de datos que ofreciera:

- periodicidad constante
- cobertura geográfica amplia
- buena calidad y trazabilidad
- disponibilidad en formato estructurado
- documentación mínima sobre su contenido.

Dentro de las fuentes revisadas se incluyeron:

a) Portales oficiales del DANE

El DANE proporciona múltiples bases de datos relacionadas con precios, inflación y abastecimiento agrícola. Sin embargo, no todas cuentan con resolución temporal adecuada o con desagregación por mercado y producto.

Los microdatos del SIPSA fueron identificados como la fuente más granular y útil para el propósito del proyecto.

b) Datos Abiertos Colombia (datos.gov.co)

Se encontraron bases vinculadas a agricultura, producción y comercio, pero muchas carecen de consistencia temporal o no incluyen el detalle de precios mayoristas.

c) Otras fuentes complementarias

Kaggle, FAO y Our World in Data ofrecen datasets relacionados con alimentos, pero sus datos no tienen la especificidad necesaria para mercados locales colombianos.

Tras este análisis, se seleccionó una fuente robusta, confiable y con alto nivel de detalle.

Dataset seleccionado (SIPSA-P 2013–2024)

El dataset escogido corresponde al **Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA-P)**, disponible en la plataforma de microdatos del DANE.

Características principales:

- **Cobertura temporal:** 2013 a 2024
- **Periodicidad:** mensual
- **Nivel de detalle:** precios mayoristas por producto y mercado
- **Formato de descarga:** ZIP con archivos CSV
- **Fuente:** DANE (microdatos)
- **Acceso:** libre, requiere solo registro gratuito
- **Estructura:** incluye códigos CPC, mercados, productos y precios promedio

Este dataset permite seguir el comportamiento de precios mayoristas con suficiente profundidad como para construir series temporales comparables y generar análisis exploratorio, indicadores y modelos predictivos.

También ofrece:

- diversificación de productos en múltiples categorías alimentarias,
- información organizada y estandarizada,
- buena frecuencia para análisis longitudinal de variaciones.

Justificación de selección

La elección del dataset se fundamenta en los siguientes elementos clave:

Relevancia

Es la base de datos más completa y detallada sobre precios mayoristas disponible en Colombia. Permite abordar de forma directa el problema definido en el proyecto.

Consistencia

La recolección de información sigue metodologías estandarizadas del DANE, lo que garantiza uniformidad en la estructura y calidad de datos.

Accesibilidad técnica

Los archivos CSV son fáciles de procesar en un pipeline de análisis con Python, lo que facilita:

- ingesta,
- limpieza,
- normalización,
- análisis exploratorio,
- visualización posterior.

Valor analítico

La cobertura temporal amplia permite:

- identificar patrones estacionales,
- detectar tendencias largas,
- hacer comparativas entre años y mercados,
- evaluar volatilidad con métricas robustas.

Trazabilidad y control

Al ser datos oficiales, se pueden justificar y respaldar fácilmente las decisiones y conclusiones derivadas.

Limitaciones del dataset

Es importante documentar las limitaciones para transparencia y rigor metodológico:

- Puede haber mercados con datos incompletos o ausentes en algunos periodos.
- La periodicidad mensual limita algunos análisis de alta frecuencia.
- Existen valores atípicos que requieren depuración o suavizado.
- La clasificación de productos puede variar ligeramente entre años.

Estas limitaciones se pueden mitigar mediante:

- limpieza y preprocesamiento,
- interpolación temporal,
- filtrado por mercados con alta completitud de datos.

Bocetos de arquitectura de datos

La arquitectura de datos propuesta se estructura en capas secuenciales que permiten la ingesta, procesamiento, almacenamiento y análisis de la información proveniente del sistema SIPSA-P del DANE. El objetivo es garantizar un flujo ordenado, reproducible y transparente para obtener indicadores confiables sobre precios de alimentos y sus variaciones temporales.

1. Capa de Fuentes de Datos

La fuente principal corresponde al conjunto de microdatos del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA-P), disponible en el portal de microdatos del DANE. Este dataset incluye registros de precios mayoristas por producto, mercado y período, con cobertura temporal entre 2013 y 2024. Los datos se descargan en formato CSV comprimido en archivos ZIP.

2. Capa de Ingesta

En esta capa se realiza la adquisición inicial de los archivos y su verificación estructural. El proceso incluye la descarga, descompresión y revisión de las columnas definidas para garantizar consistencia con el estándar proporcionado por la fuente oficial. Los archivos brutos se almacenan en la carpeta /data/raw/, preservando la integridad del dataset original.

3. Capa de Procesamiento (ETL)

La fase de procesamiento comprende la limpieza, normalización y transformación de los datos. Las actividades principales incluyen: conversión de fechas a formato estándar, eliminación de duplicados, tratamiento de valores faltantes, estandarización de nombres de productos y mercados, así como la agregación por períodos y cálculo de métricas relevantes (promedios, variaciones y medidas de volatilidad). El resultado se almacena en /data/processed/, con una estructura clara y documentada.

4. Capa de Almacenamiento Analítico

Una vez procesados, los datos se organizan en un repositorio analítico estructurado que permite su consulta, reutilización y análisis posterior. Este almacenamiento facilita la trazabilidad y la separación lógica entre datos brutos, procesados y resultados analíticos. Se crean carpetas dedicadas para métricas, series temporales y resultados de análisis exploratorio.

5. Capa de Consumo y Análisis

En esta etapa se desarrollan los análisis exploratorios (EDA), los modelos de predicción y las visualizaciones. Los notebooks se utilizan como herramienta principal de ejecución

secuencial y documentación técnica. Aquí se generan indicadores clave, gráficos comparativos entre mercados y tendencias temporales, así como modelos predictivos basados en series temporales.

6. Capa de Visualización y Comunicación

Los resultados se integran en un dashboard interactivo, permitiendo a los usuarios explorar las tendencias de precios, comparar mercados y monitorear indicadores relevantes. Esta capa facilita la comunicación de hallazgos de forma clara y accesible, apoyando la toma de decisiones basada en datos.

7. Capa de Gobernanza y Documentación

La arquitectura incluye mecanismos de gobernanza que aseguran transparencia, control de versiones y trazabilidad. Estas actividades se desarrollan en un repositorio colaborativo en GitHub, donde se registra la documentación técnica, el diccionario de datos, las políticas de calidad y los archivos necesarios para reproducir el análisis.

Gobierno de Datos

1. Roles y responsabilidades

Data Owner

- Define el alcance de productos/mercados a analizar y valida la publicación de resultados (gráficos/tablas).

Data Steward

- Mantiene este documento, el **diccionario de datos** y los **catálogos** (lista de mercados y productos).
- Revisa que se cumplan las reglas de calidad (abajo) y aprueba los cambios de nombres/formatos.

Data Engineer

- Ejecuta la ingesta desde /content/data/raw/*.csv con sep=";" y encoding="latin-1".
- Aplica las transformaciones que ya están en el notebook:
    - Normaliza nombres de columnas.
    - Convierte fecha a datetime (formato %b-%y).
    - Convierte precio_promedio_por_kilogramo a numérico (reemplazo de coma decimal y espacios → to_numeric(errors="coerce")).
    - Elimina duplicados y NaN en campos clave.

- Concatena y exporta el **master** a /content/data/processed/sipsa_master.csv.

Analista

- Carga el dataset procesado (data/processed/sipsa_master.csv, parse_dates=["fecha"]) para hacer EDA:
    - Histograma de precio_promedio_por_kilogramo.
    - Boxplot por mercado.
    - Tendencias por producto (ej. “ACELGA”).
    - Matriz de correlación tras pivot.

Nota práctica: hoy el notebook escribe en /content/data/processed/... pero lee desde data/processed/.... **Regla** abajo para unificar ruta.

2. Reglas de calidad

Q-01 — Estructura mínima por archivo *(aplica en crudo antes de concatenar)*

- **Se esperan estas columnas**: ["fecha", "grupo", "producto", "codigo_cpc_ac", "mercado", "precio_promedio_por_kilogramo"].
- **Acción**: si falta alguna, ese archivo no entra al lote.

Q-02 — Tipos y codificación

- Lectura de cada CSV con sep=";" y encoding="latin-1".
- fecha → datetime usando pd.to_datetime(..., format="%b-%y", errors="coerce").
- precio_promedio_por_kilogramo → numérico tras str.replace (coma→punto, quitar espacios) y to_numeric(errors="coerce").
- **Acción**: filas que no se puedan tipificar quedan como NaN y se eliminan en Q-04.

Q-03 — Duplicados

- **Por archivo** y **tras concatenar** se ejecuta drop_duplicates().
- **PK efectiva hoy**: todas las columnas (porque no hay subset explícito).
- **Recomendación no disruptiva**: cuando edites el notebook, usa subset=["fecha","mercado","producto"] para que quede explícita la unicidad de observaciones.

Q-04 — Completitud en claves

- Se elimina cualquier fila con NaN en fecha o precio_promedio_por_kilogramo (dropna(subset=[...])).

- **Métrica a reportar**: % de filas descartadas por cada motivo (tipo/NaN/duplicado).

Q-05 — Orden reproducible

- El dataset se ordena por ["fecha", "producto", "mercado"] antes de exportar.
- **Beneficio**: salidas deterministas entre corridas.

Q-06 — Consumo analítico consistente

- El EDA debe usar el **master exportado** (no “fragmentos” previos). En el notebook ya se hace un read_csv("data/processed/sipsa_master.csv", parse_dates=["fecha"]).
- **Regla**: todos los gráficos/tablas se derivan de ese **único archivo procesado**.

3. Estándares de nomenclatura y formato (exactos a lo que hoy existe)

3.1 Columnas:

- fecha *(datetime)*
- grupo *(str)*
- producto *(str)*
- codigo_cpc_ac *(str/categórico; si llega vacío, mantener vacío, no inventar)*
- mercado *(str)*
- precio_promedio_por_kilogramo *(float; COP por kg)*

3.2 Rutas y archivos:

- **Entrada (raw):** /content/data/raw/*.csv
- **Salida (procesado maestro): usar una sola** de estas dos y estandarizar:
  - O bien /content/data/processed/sipsa_master.csv (Colab),
  - o bien data/processed/sipsa_master.csv (ruta relativa del repo).
- **Regla concreta:** dejar **ambas** invocaciones apuntando al **mismo** path; si se ejecuta en Colab, leer y escribir en /content/…. Si se ejecuta local, en data/…. (Cambia **una** línea, no la lógica).

3.3 Formato de archivo:

- **CSV**, delimitador , al exportar (por defecto pandas).
- **Codificación**: no se especifica en el notebook al escribir; se asume utf-8 por defecto de pandas.
- **Regla**: no añadir BOM; no incluir índice (index=False ya está).

3.4 Convenciones de valores

- Textos (grupo, producto, mercado) **str** limpios (sin espacios extra, ya se normalizan).
- precio_promedio_por_kilogramo siempre en **COP/kg** (la notebook ya lo deja en numérico, no cambies la unidad).
- Fechas a nivel mensual según el origen (%b-%y), convertidas a datetime y reutilizadas para pivot, groupby y series de tiempo.

4. Diccionario de datos

- **fecha**: fecha de referencia (mes) del precio; tipo datetime64[ns].
- **grupo**: agrupación del producto (texto).
- **producto**: nombre del producto (texto).
- **codigo_cpc_ac**: código CPC asociado si existe (texto).
- **mercado**: plaza/mercado mayorista (texto).
- **precio_promedio_por_kilogramo**: precio medio en COP/kg (float).

5. Publicación y trazabilidad mínima

- **Artefacto oficial de análisis**: sipsa_master.csv.
- **EDA**: todos los gráficos salen de ese archivo (ya ocurre en el cuaderno).
- **Registro liviano**: al cerrar cada sesión, anota **fecha de corrida** y **cantidad de filas** del master (sirve para reproducibilidad sin tocar el código).