Fundamentos del Análisis Programático de Datos

Funciones avanzadas: conceptos y casos de uso

2 de septiembre de 2025

Objetivos de la clase

- Comprender cuándo y por qué usar funciones avanzadas
- Identificar problemas que requieren reestructuración de datos
- Reconocer situaciones que necesitan lógica condicional compleja
- Entender los desafíos del manejo de series temporales
- Conceptualizar la integración de múltiples fuentes de datos
- Analizar relaciones temporales entre observaciones

El salto conceptual: de operaciones básicas a

Hasta ahora trabajamos con:

- **Una base de datos** a la vez
- **Operaciones directas**: seleccionar, filtrar, crear variables
- **Agrupaciones simples**: resumir por categorías
- **Lógica binaria**: verdadero o falso

Ahora enfrentamos problemas más complejos:

- **Múltiples estructuras** de datos
- **Decisiones con múltiples criterios**
- **Información temporal** y secuencial
- **Relaciones entre diferentes bases de datos**
- **Cálculos que dependen de observaciones relacionadas**

¿Cuándo los datos no se adaptan al análisis?

Problemas frecuentes en datos económicos:

Estructura inadecuada

- Años como columnas separadas
- Variables múltiples en nombres de columnas
- Información "ancha" cuando necesitamos "larga"

Información dispersa

- Datos de empleo en una base
- Información de salarios en otra
- Datos demográficos en una tercera

Lógica compleja

- Clasificaciones socioeconómicas con múltiples criterios
- Condiciones que dependen de varias variables
- Definiciones técnicas complejas

Relaciones temporales

- Comparar con períodos anteriores
- Calcular tasas de crecimiento
- Detectar ciclos económicos

Pivots: Cuando la estructura obstaculiza el análisis

El problema de los datos "anchos"

Situación típica: datos macroeconómicos

País	2020	2021	2022	2023
Argentina	2.1	10.4	72.4	133.0
Brasil	3.2	8.3	9.3	4.6
Chile	3.0	4.5	11.6	7.6

¿Qué preguntas NO podemos responder fácilmente?

- ¿Cuál es la inflación promedio por año en la región?
- ¿Qué país tiene mayor volatilidad inflacionaria?
- ¿Hay convergencia o divergencia entre países?
- ¿Cómo graficar la evolución temporal comparativa?

El problema: los años son *valores*, no *variables*

Casos de uso para pivot_longer()

¿Cuándo necesitamos "alargar" los datos?

1. **Análisis de series temporales**

- Datos con períodos como columnas (años, trimestres)
- Calcular estadísticas por período
- Crear gráficos de evolución temporal

2. **Análisis comparativo entre unidades**

- Múltiples países, provincias, sectores como columnas
- Análisis de convergencia o divergencia
- Estudios de correlación entre unidades

3. **Preparación para modelado econométrico**

- Panel data requiere formato largo
- Variables instrumentales en formato tidy
- Análisis de efectos fijos

Casos de uso para pivot_wider()

¿Cuándo necesitamos "ensanchar" los datos?

1. **Matrices de análisis**

- Tablas input-output
- Matrices de correlación entre variables
- Análisis de componentes principales

2. **Reportes comparativos**

- Comparar indicadores año a año
- Benchmarking entre regiones
- Antes vs después de políticas económicas

3. **Preparación para análisis específicos**

- Algunos modelos econométricos requieren formato ancho
- Cálculo de ratios entre variables
- Exportar a sistemas estadísticos especializados

Identificando cuándo pivotar

Preguntas guía:

¿Necesito pivot_longer()?

- ¿Tengo información temporal en múltiples columnas?
- ¿Las columnas representan niveles de una misma variable?
- ¿Quiero agrupar o comparar entre estos "períodos/unidades"?
- ¿Necesito analizar estas "columnas" como una serie temporal?

¿Necesito pivot_wider()?

- ¿Quiero comparar valores lado a lado?
- ¿Necesito crear una matriz para análisis posterior?
- ¿Debo calcular diferencias o ratios entre categorías?
- ¿El análisis posterior requiere variables separadas?

Lógica condicional: Más allá de lo binario

Limitaciones del pensamiento binario

En análisis básico hacemos:

```
SI PIB_per_capita > 15000 ENTONCES "Desarrollado" SINO "En desarrollo"
```

Pero la realidad económica es más compleja:

- **Trabajadores**: No solo formal vs informal
- **Países**: Múltiples niveles de desarrollo
- **Sectores**: Diferentes grados de formalización
- **Políticas**: Varios niveles de intervención

El problema: clasificaciones con múltiples dimensiones

- ¿Cómo clasifico un trabajador cuentapropia con ingresos altos?
- ¿Qué categoría para países de renta media con alta desigualdad?
- ¿Cómo manejo sectores con características mixtas?

if_else vs case_when: ¿Cuándo usar cada uno?

if_else: Para decisiones simples y claras

Casos ideales:

- **Dicotomías reales**: Público/Privado, Urbano/Rural
- **Transformaciones directas**: Convertir códigos a etiquetas
- **Cálculos condicionales**: Aplicar subsidio o no
- **Validaciones**: Dentro/Fuera del rango esperado

Ventajas:

- Más eficiente para casos binarios
- Fuerza a pensar en dicotomías cuando corresponde
- Menos propenso a errores de lógica

case_when: Para la complejidad

Casos que requieren case_when:

1. **Clasificación de trabajadores**

- Múltiples variables: ingresos, sector, estabilidad
- Categorías jerárquicas: Formal pleno > Formal precario > Informal
- Casos especiales: Monotributistas, cooperativistas

2. **Tipología de países/regiones**

- Desarrollo: Desarrollado > Renta media alta > Renta media baja > Bajo desarrollo
- Competitividad: Líder > Seguidor > Rezagado
- Estructura productiva: Industrial > Servicios > Primario > Mixto

3. **Clasificación sectorial**

- Productividad: Alto > Medio > Bajo rendimiento
- Intensidad tecnológica: Intensivo > Medio > Extensivo
- Orientación: Exportador > Mercado interno > Mixto

Pensando en lógica jerárquica

Principio: Orden de evaluación importa

Ejemplo conceptual: Clasificación de trabajadores

1. **Primero**: ¿Tiene registración formal completa? → Formal
2. **Segundo**: ¿Trabaja por cuenta propia con registración? → Cuentapropia formal
3. **Tercero**: ¿Ingresos > salario mínimo? → Informal de ingresos medios
4. **Resto**: Informal de bajos ingresos

Errores comunes:

- **Orden incorrecto**: Evaluar condiciones más específicas al final
- **Solapamiento**: Categorías que se superponen sin jerarquía clara
- **Casos faltantes**: No contemplar todas las situaciones laborales posibles
- **Complejidad excesiva**: Demasiados criterios simultáneos

Fechas: El desafío de la dimensión temporal

¿Por qué las fechas son complicadas en economía?

Múltiples formatos institucionales

- "2024-Q3" (Trimestres)
- "2024M08" (Mensual)
- "15/08/2024" (DD/MM/YYYY)
- "Ago-2024" (Texto)
- Datos anuales: "2024"

Complejidades específicas

- **Calendarios fiscales**: Año fiscal vs año calendario
- **Frecuencias mixtas**: Datos anuales vs trimestrales vs mensuales
- **Ajustes estacionales**: Datos desestacionalizados
- **Revisiones**: Datos preliminares vs definitivos

Conceptos clave en análisis temporal económico

Componentes temporales relevantes

- **Año, trimestre, mes**: Ubicación en ciclos económicos
- **Estacionalidad**: Patrones que se repiten anualmente
- **Tendencia**: Dirección de largo plazo
- **Ciclo**: Fluctuaciones de mediano plazo

Operaciones temporales en economía

- **Parsing**: Convertir formatos institucionales a fechas
- **Sincronización**: Alinear frecuencias diferentes
- **Rezagos**: Comparar con períodos anteriores
- **Agregación temporal**: De mensual a trimestral
- **Interpolación**: Completar datos faltantes

Casos de uso para análisis temporal

1. **Análisis de ciclos económicos**

- ¿El PIB muestra patrones cíclicos?
- ¿Hay correlación entre indicadores adelantados?
- ¿Qué variables predicen recesiones?

2. **Evaluación de políticas**

- Antes vs después de implementación
- Efectos inmediatos vs efectos de largo plazo
- Análisis de interrupciones (structural breaks)

3. **Modelado econométrico**

- Variables rezagadas como regresores
- Análisis de cointegración
- Modelos de corrección de errores

Patrones temporales en economía

Preguntas que guían el análisis:

Estacionalidad

- ¿Hay trimestres con mejor performance?
- ¿El empleo varía sistemáticamente por época del año?
- ¿Los precios siguen patrones estacionales?

Tendencias

- ¿La productividad crece sostenidamente?
- ¿La desigualdad aumenta o disminuye?
- ¿Hay cambios estructurales en la economía?

Ciclos

- ¿Cuánto duran las recesiones típicamente?
- ¿Hay patrones en los ciclos políticos?
- ¿Los shocks externos tienen efectos persistentes?

Estrategias para trabajar con datos temporales

Principios de diseño:

1. **Estandarización temprana**

- Convertir todas las fechas al mismo formato
- Decidir frecuencia de análisis (mensual, trimestral)
- Documentar origen y metodología de las series

2. **Extracción sistemática de componentes**

- Crear variables para cada dimensión temporal relevante
- Año, trimestre, mes, período presidencial
- Variables dummy para crisis, reformas, eventos

3. **Validación constante**

- ¿Las fechas corresponden a períodos de publicación?
- ¿Hay datos faltantes por feriados o crisis?
- ¿Los rangos temporales son consistentes entre variables?

Joins: Integrando información dispersa

El mundo real: información fragmentada en economía

Situación típica en análisis económico:

INDEC

- EPH, PIB, inflación, comercio exterior
- Diferentes periodicidades y metodologías

BCRA

- Variables monetarias, cambiarias, financieras
- Series de alta frecuencia

Ministerios sectoriales

- Empleo, educación, salud, infraestructura
- Datos administrativos y registros

Tipos de relaciones entre bases económicas

1. **Uno a uno** (1:1)

- Un país → Un indicador de desarrollo humano
- Una provincia → Un dato de población
- **Join perfecto**: Cada unidad tiene exactamente un valor por período

2. **Uno a muchos** (1:N)

- Un país → Múltiples provincias
- Una industria → Múltiples empresas
- **El más común**: Unidades agregadas y sus componentes

3. **Muchos a muchos** (M:N)

- Trabajadores → Sectores (un trabajador puede tener múltiples empleos)
- Productos → Clasificaciones (un producto en varias categorías)
- **El más complejo**: Requiere tablas de correspondencia

Estrategia de joins: ¿Qué conservar?

Inner join: Solo coincidencias completas

Cuándo usar:

- Análisis de series balanceadas
- Cuando la coincidencia es requisito metodológico
- Estudios de países/regiones con datos completos

Riesgo: **Perder observaciones por datos faltantes**

Estrategia de joins: ¿Qué conservar?

Left join: Conservar base principal

Cuándo usar:

- Enriquecer base de datos principal
- Mantener integridad del panel de datos
- La tabla de la izquierda contiene las unidades de análisis

Ventaja: **No perdemos observaciones principales**

Estrategia de joins (continuación)

Right join: Conservar tabla secundaria

Cuándo usar:

- Menos común, generalmente se prefiere reordenar y usar left
- Cuando la tabla "secundaria" es realmente la de referencia

Full join: Conservar toda la información

Cuándo usar:

- Análisis de completitud de datos
- Identificar qué información falta en cada fuente
- Auditorías de calidad de bases oficiales

Cuidado: **Puede generar muchos valores faltantes**

Problemas comunes con joins en datos económicos

1. **Joins que explotan**

- Una base tiene datos duplicados por error de metodología
- Resultado: más observaciones de las esperadas
- **Solución**: Validar unicidad antes de join

2. **Claves que no coinciden**

- Diferencias en codificación: "CABA" vs "Ciudad de Buenos Aires"
- Cambios en clasificaciones industriales (CIIU)
- **Solución**: Crear tablas de correspondencia

3. **Múltiples criterios de unión**

- No basta con código de país, también necesitamos año
- Claves compuestas: región + sector + período
- **Solución**: Identificar todas las dimensiones relevantes

Diseñando estrategia de joins

Preguntas antes de hacer join:

¿Qué relación espero?

- ¿1:1, 1:N, o M:N?
- ¿Todas las observaciones deberían tener coincidencia?

¿Qué datos son críticos?

- ¿Puedo perder observaciones sin coincidencia?
- ¿La información faltante afecta la representatividad?

¿Cómo valido el resultado?

- ¿Cuántas observaciones espero después del join?
- ¿Se mantienen los totales consistentes?
- ¿Las distribuciones tienen sentido?

Window Functions: Cálculos relacionales

Más allá de las agregaciones simples

Limitación de group_by + summarise:

- **Colapsa las observaciones** en un solo resultado por grupo
- Perdemos la variabilidad individual
- No podemos comparar cada observación con su contexto

Window functions permiten:

- **Mantener todas las observaciones**
- Calcular estadísticas "mirando alrededor"
- Comparar cada dato con su grupo de referencia
- Hacer cálculos secuenciales manteniendo el detalle

Concepto clave: **"Ventana" de observación**

- Cada observación puede "ver" otras observaciones relacionadas
- La "ventana" puede ser todo el grupo o un rango temporal específico

Tipos de cálculos con window functions

1. **Comparaciones intertemporales**

- PIB vs trimestre anterior
- Tasa de crecimiento período a período
- Detectar aceleraciones y desaceleraciones

2. **Rankings y posiciones relativas**

- Países con mayor crecimiento por región
- Posición relativa en distribución de ingresos
- Identificar líderes y rezagados

3. **Acumulados y promedios móviles**

- PIB acumulado del año
- Inflación promedio de últimos 12 meses
- Suavizar volatilidad y ver tendencias

Casos de uso: lag y lead

lag: Análisis retrospectivo

Casos típicos:

- **Crecimiento económico**: ¿Cómo cambió vs trimestre anterior?
- **Detección de crisis**: ¿Este valor es anómalamente diferente?
- **Análisis de recuperación**: ¿Cuánto tardamos en volver al nivel pre-crisis?
- **Efectos rezagados**: Impacto de políticas con delay

lead: Análisis prospectivo

Casos típicos:

- **Indicadores adelantados**: ¿Qué pasó después de este evento?
- **Análisis de efectos**: ¿La política tuvo impacto en períodos siguientes?
- **Validación de predicciones**: ¿Las expectativas se cumplieron?

Casos de uso: rankings

row_number(): Posición única

- **Top N performers** económicos sin empates
- Seleccionar el "mejor" de cada grupo temporal
- Crear identificadores únicos dentro de períodos

rank(): Rankings con empates

- **Clasificaciones internacionales**: mismo nivel, misma posición
- Índices de competitividad donde empates son válidos
- Análisis de distribución por percentiles

dense_rank(): Rankings compactos

- Sin "huecos" después de empates
- Útil para categorización por niveles de desarrollo
- Análisis de movilidad entre estratos

Casos de uso: funciones acumulativas

cumsum(): Suma acumulada

- **PIB acumulado** del año
- Déficit fiscal acumulado
- Inversión pública ejecutada año a la fecha

cummean(): Promedio acumulado

- **Inflación promedio** hasta la fecha
- Evolución del desempleo promedio anual
- Performance económica acumulada

cummax()/cummin(): Extremos acumulados

- **Picos y valles** históricos
- Niveles máximos de reservas alcanzados
- Mínimos históricos de pobreza

Pensando en ventanas temporales

Preguntas clave para window functions:

¿Qué grupo define mi "ventana"?

- ¿Por país, provincia, sector económico?
- ¿Todos los datos juntos para comparación global?

¿Qué orden tiene sentido?

- Cronológico para análisis de series temporales
- Por valor para rankings de performance
- Por nivel de desarrollo para análisis de convergencia

¿Qué información necesito de otras observaciones?

- Solo el período anterior (lag)
- Todo el grupo histórico (rank)
- Acumulado desde el inicio del período (cumsum)

Implementación práctica de funciones avanzadas

Aplicación a casos reales con datos económicos

¿Preguntas?