HE2: Consultoría Económica e IAR

# Clase 4 - Ciclo de Vida de Datos

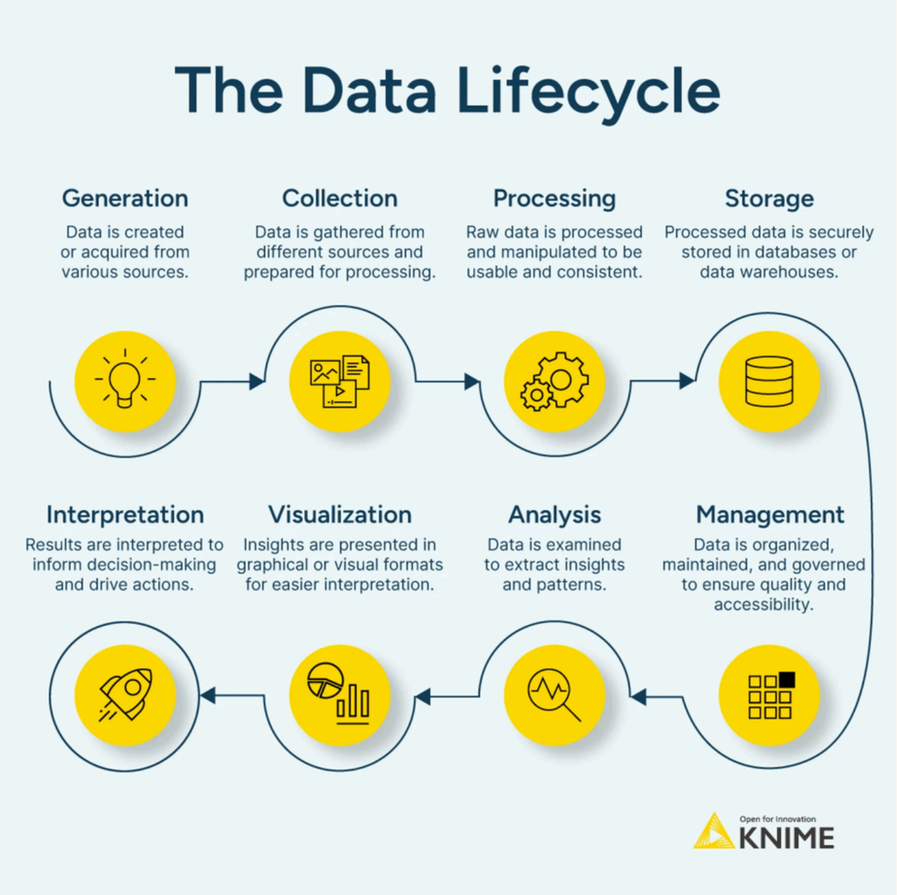
Catalina Bernal
Santiago Neira

Enero 29 2026
Semana 2

Universidad de los Andes | Facultad de Economía

# Contenido

# Ciclo de vida de los datos (Data Cycle)




Ciclo de vida - KNime

# ¿Por qué el promedio puede mentir?

**× Crudo**

**Adult Income:**

- Edad: 38.6 años
- Horas: 40.4/sem
- >50K: 24%

**✓ Analizado**

**>50K por género:**

- Hombres: 31%
- Mujeres: 11%

**Por educación:**

- Sin bach.: 5%
- Maestría: 62%

# ¿Por qué el promedio puede mentir?

**× Crudo**

**Adult Income:**

- Edad: 38.6 años
- Horas: 40.4/sem
- >50K: 24%

**✓ Analizado**

**>50K por género:**

- Hombres: 31%
- Mujeres: 11%

**Por educación:**

- Sin bach.: 5%
- Maestría: 62%

El promedio esconde **desigualdades**

# Construcción de Indicadores

**Ratios**

Relación entre variables. Por ejemplo: ingresos/horas

**Agregaciones**

Resumir grupos. Por ejemplo: promedio por nivel de educación

**Categorización**

Crear grupos. Por ejemplo: tiempo parcial, tiempo completo

# Construcción de Indicadores

**horas/ingresos**

- Valores altos = muchas horas por poco dinero
- ¿qué ocupaciones son más “eficientes”?

**promedio por nivel de educación**

- Sensible a outliers

**Categorización**

¿Por qué esto es útil?

# Construcción de Indicadores

**horas/ingresos**

- Valores altos = muchas horas por poco dinero
- ¿qué ocupaciones son más “eficientes”?

**promedio por nivel de educación**

- Sensible a outliers

**Categorización**

¿Por qué esto es útil?

Diseñen 2 indicadores que revelen **desigualdad**

# Construcción de Indicadores

**horas/ingresos**

- Valores altos = muchas horas por poco dinero
- ¿qué ocupaciones son más “eficientes”?

**promedio por nivel de educación**

- Sensible a outliers

**Categorización**

¿Por qué esto es útil?

- Calculables
- Interpretables
- Accionables

# La Paradoja de Simpson

"Las mujeres, en promedio, ganan más que los hombres"

**× Total**

Promedio:

- Mujeres: $42K
- Hombres: $41.5K

**✓ Por ocupación**

**Tech:** M: $55K (10%) — H: $56K (40%)
**Service:** M: $38K (90%) — H: $37K (60%)

En CADA ocupación, los hombres ganan más. El promedio general es engañoso ¿por qué?

# Cuándo usar media vs mediana vs moda

ESCENARIO: Ingresos de 10 personas
Datos: [30K, 32K, 35K, 35K, 38K, 40K, 42K, 45K, 48K, 500K]

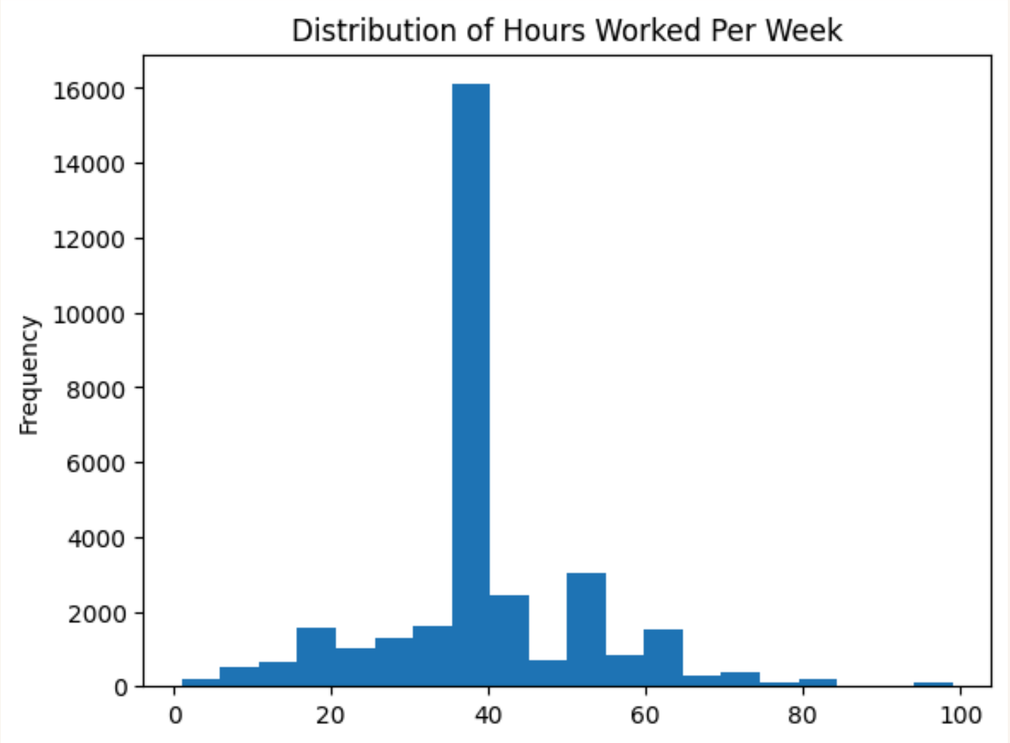
| Media | Mediana | Moda |
|---|---|---|
| $84.5K | $39K | $35K |
| ! Distorsionada | ✓ Típico | Más común |

# Cuándo usar media vs mediana vs moda

| Métrica | Cuándo usarla | Cuándo NO usarla |
|---|---|---|
| **Media** | Distribuciones simétricas, sin outliers | Datos con valores extremos, distribuciones asimétricas |
| **Mediana** | Distribuciones asimétricas, con outliers | Cuando necesitas sensibilidad a todos los valores |
| **Moda** | Variables categóricas, identificar patrones | Distribuciones continuas sin picos claros |

# Cuándo usar media vs mediana vs moda

| Métrica | Cuándo usarla | Cuándo NO usarla |
|---|---|---|
| **Media** | Distribuciones simétricas, sin outliers | Datos con valores extremos, distribuciones asimétricas |
| **Mediana** | Distribuciones asimétricas, con outliers | Cuando necesitas sensibilidad a todos los valores |
| **Moda** | Variables categóricas, identificar patrones | Distribuciones continuas sin picos claros |



# Un caso...

Quieren calcular "el salario promedio de personas con educación universitaria" en el dataset.

**Opción A:** Filtrar education=='Bachelors' y sacar mean(income)
**Opción B:** Hacer groupby por [education, sex, race] y luego promediar los promedios
**Opción C:** Hacer groupby por [education, occupation] primero, luego agregar

1. ¿Cuál opción elegirían y por qué?
2. ¿Qué sesgos podría introducir cada método?
3. ¿Para qué pregunta de investigación sirve cada una?

# Sesgos en los Datos

Un sesgo es una desviación sistemática de la realidad que queremos medir, causada por cómo recolectamos, procesamos o interpretamos los datos. No es ruido aleatorio, es error direccional.

¿Por qué importa?

- Los algoritmos amplifican sesgos en datos
- Perpetúan desigualdades históricas
- Causan daño real a personas y comunidades
- Generan riesgo legal y reputacional para empresas

# Sesgos de Selección

La muestra no representa adecuadamente a la población objetivo

Manifestaciones comunes:

- Survivorship bias (solo vemos los que “sobrevivieron”)
- Sampling bias (método de muestreo defectuoso)
- Non-response bias (quienes no responden son diferentes)

**Ejemplo**

**Caso: Encuesta de satisfacción online**

- Solo responden usuarios con acceso a internet
- Usuarios muy insatisfechos (cancelaron) no están en la base
- Usuarios muy satisfechos tienen más motivación para responder

**Resultado:** Sobreestimamos satisfacción.

# Sesgos de Medición

Errores sistemáticos en cómo se capturan los datos

- Encuestas con preguntas sesgadas
- Proxies incorrectos
- Instrumentos de medición descalibrados

**Ejemplo**

**Caso: Predicción de riesgo criminal (COMPAS)**

- Variable objetivo: "arrestos futuros"
- Problema: Arrestos reflejan *actividad policial*, no necesariamente criminalidad
- Comunidades sobre-vigiladas tienen más arrestos
- El modelo "aprende" que ser de cierta raza = más riesgo

**Resultado:** Profecía autocumplida.

# Sesgos Históricos

Los datos reflejan desigualdades del pasado

**Ejemplo**

**Caso Amazon Recruiting (recordar intro)**

- Datos históricos: 10 años de contrataciones
- Realidad histórica: Industria tech dominada por hombres
- Modelo aprende: "Hombre" = característica positiva

**Ejemplo**

**Caso: Préstamos bancarios**

- Históricamente, mujeres tenían menos acceso a crédito
- Datos muestran que mujeres tienen menos historial crediticio
- Modelo penaliza a mujeres por falta de historial
- Perpetúa exclusión financiera

# Cómo las decisiones de limpieza introducen sesgos

Limpieza de datos es necesaria, pero cada decisión puede introducir sesgo.

Qué se hace usualmente:

1. Eliminar filas con datos faltantes
2. Imputar con media/mediana
3. Imputar con un modelo de ajuste / predictivo

**¿Dónde está el sesgo?**

# Cómo las decisiones de limpieza introducen sesgos - NANs

**Ejemplo**

**Escenario:** Dataset de solicitudes de empleo. 20 % no reportan salario anterior.
**Si eliminamos esas filas:**

- Perdemos desempleados de largo plazo
- Perdemos personas con trabajos informales
- Nuestro modelo asume “todos tienen historial formal”

**Si imputamos con mediana:**

- Asumimos que faltantes = promedio
- Invisibilizamos desigualdad salarial
- Reducimos artificialmente varianza

# Cómo las decisiones de limpieza introducen sesgos - outliers

**Ejemplo**

**Escenario:** Modelo de scoring crediticio. Algunas personas tienen 15+ tarjetas de crédito (outliers).
**Si eliminamos outliers:**

- Removemos casos "anómalos"
- Pero: ¿Y si esos outliers son empresarios legítimos?
- El modelo aprende: "Comportamiento atípico = malo"
- Penalizamos innovación financiera

**Punto Clave**

**Regla de oro:**
Antes de eliminar datos, pregunta: ¿Estos datos son *erróneos* o solo *incómodos para mi modelo*? La diferencia es crucial.

# Cómo las decisiones de limpieza introducen sesgos - features

**Ejemplo**

**Escenario:** Predecir deserción escolar. Incluimos "código postal" como feature.
**Problema:**

- Código postal correlaciona con raza y nivel socioeconómico
- Estamos usando un *proxy* de características protegidas
- El modelo puede discriminar "legalmente" por geografía
- Pero realmente discrimina por raza/clase

# Otros riesgos

Privacidad y Protección de Datos

Seguridad, Reidentificación

*Data Drift*

# Checklist del Consultor para Detectar Sesgos

Antes de iniciar cualquier proyecto:

1. Sobre la recolección:

- ¿Quién está ausente en estos datos?
- ¿El método de captura favorece a ciertos grupos?
- ¿Tenemos representación balanceada de poblaciones clave?

2. Sobre las variables:

- ¿Qué estamos realmente midiendo?
- ¿Nuestro feature/target es un proxy sesgado?
- ¿Refleja desigualdades históricas?

# Checklist del Consultor para Detectar Sesgos

3. Más sobre las variables:

- ¿Alguna variable es proxy de características protegidas?
- ¿Estamos usando datos que perpetúan estereotipos?
- ¿Las correlaciones tienen sentido causal o son espurias?

4. Sobre limpieza:

- ¿A quién afectan desproporcionadamente nuestras decisiones de limpieza?
- ¿Estamos eliminando datos “inconvenientes” vs. erróneos?
- ¿Documentamos el impacto de cada transformación?

5. Sobre el impacto:

- ¿Quién se beneficia de este sistema?
- ¿Quién podría ser perjudicado?
- ¿Cómo auditamos el sistema una vez en producción?

HE2: Consultoría Económica e IAR

# Clase 4 - Ciclo de Vida de Datos

Santiago Neira
Catalina Bernal

Enero 29 2026
Semana 2

Universidad de los Andes | Facultad de Economía