Exploración De Datos

Documentación Del Proceso De Exploración De Datos (EDA) En La Base De Datos CIRI De Derechos Humanos (1981-2011)

Metadatos Del Proyecto

- Dataset: Cingranelli-Richards Human Rights Database (CIRI)
- Período: 1981-2011 (31 años)
- Observaciones: 6,262 registros (países-año)
- Variables: 26 columnas
- Estado: Datos crudos (pre-limpieza)
- Fecha de análisis: 30-Septiembre-2025

Objetivos Del EDA

Objetivo Principal

Evaluar la estructura, calidad y características fundamentales de los datos CIRI en su estado original, identificando patrones, anomalías y desafíos técnicos para fundamentar el proceso de limpieza de datos.

Objetivos Específicos

1. **Caracterización estructural:** Dimensiones, tipos de datos y esquema general.
2. **Calidad de datos:** Identificación de valores faltantes y atípicos.
3. **Análisis de esquemas de codificación:** Interpretación de códigos especiales.
4. **Verificación de consistencia:** Verificar cobertura por año y país.
5. **Documentación base:** Establecimiento de línea base para decisiones de limpieza.

Metodología De Exploración

Enfoque Por Fases

- Fase 1: Análisis Estructural.
- Fase 2: Evaluación de Calidad de Datos.

- Fase 3: Identificación de Patrones Específicos

Herramientas Utilizadas

- Python 3.8+ con bibliotecas pandas, numpy, matplotlib, seaborn
- Google Colab para análisis interactivo
- Git para control de versiones del análisis

Criterios De Evaluación

- **Completitud:** Porcentaje de valores no faltantes.
- **Consistencia:** Coherencia en formatos y rangos.
- **Validez:** Cumplimiento con esquemas de codificación documentados.
- **Uniformidad:** Consistencia temporal y geográfica.

Hallazgos Del EDA - Estado Actual

1. Estructura General Del Conjunto De Datos

Procedimiento realizado:
Se cargó el dataset completo y se examinaron las dimensiones fundamentales para comprender el alcance del conjunto de datos.

Hallazgos iniciales:

```
======================================================================
CARGA Y DIMENSIONES BÁSICAS
======================================================================
Dimensiones del dataset: (6262, 26)
Rango de años: 1981 - 2011
Países únicos: 202
```

Interpretación:

- El dataset tiene 6,262 registros y 26 variables, lo que coincide con lo esperado.
- El período cubierto es de 1981 a 2011, que son 31 años.
- Hay 202 países únicos, lo que está dentro del rango esperado.

2. Evaluación De Calidad De Datos

Procedimiento realizado:
Se examinó la estructura de tipos de datos para cada variable, identificando la naturaleza de las columnas y su distribución.

Hallazgos iniciales:

```
======================================================================
ANÁLISIS DE TIPOS DE DATOS
======================================================================
Data columns (total 26 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Year                                  6262 non-null   int64
 1   CIRI Country ID                       6262 non-null   int64
 2   Country                               6262 non-null   object
 3   UN Country ID                         6138 non-null   float64
 4   UN Region                             6262 non-null   object
 5   UN Subregion                          6262 non-null   object
 6   Physical Integrity Rights Index       4915 non-null   float64
 7   Disappearance                         5045 non-null   float64
 8   Extrajudicial Killing                 5043 non-null   float64
 9   Political Imprisonment                5043 non-null   float64
 10  Torture                               5043 non-null   float64
 11  Empowerment Rights Index (Old)        3975 non-null   float64
 12  Empowerment Rights Index (New)        4933 non-null   float64
 13  Freedom of Assembly and Association   5043 non-null   float64
 14  Freedom of Foreign Movement           5738 non-null   float64
 15  Freedom of Domestic Movement          5737 non-null   float64
 16  Freedom of Movement (Old)             4082 non-null   float64
 17  Freedom of Speech                     5043 non-null   float64
 18  Electoral Self-Determination          5043 non-null   float64
 19  Freedom of Religion (Old)             4074 non-null   float64
 20  Freedom of Religion (New)             5640 non-null   float64
 21  Employment Rights                     5043 non-null   float64
 22  Women’s Economic Rights               5042 non-null   float64
 23  Women’s Political Rights              5042 non-null   float64
 24  Women’s Social Rights                 3882 non-null   float64
 25  Independence of Judiciary             5689 non-null   float64

Resumen de tipos de datos:
   - Numéricas (float64): 21
   - Enteras (int64): 2
   - Categóricas (object): 3

Variables categóricas específicas:
   - Country: 202 valores únicos
   - UN Region: 6 valores únicos
     Valores: ['Asia', 'Europe', 'Africa', 'Latin America and the Caribbean',
```

```
'Oceania', 'North America']
   - UN Subregion: 21 valores únicos
```

Interpretación:
La mayoría de las variables son numéricas (float64), lo que coincide con la naturaleza de índices y puntuaciones de derechos humanos. Las variables categóricas se limitan a metadatos geográficos, con una estructura de clasificación regional de la ONU bien definida.

- **Variables numéricas continuas (float64)**: 21 variables - Representan índices y scores de derechos humanos
- **Variables numéricas enteras (int64)**: 2 variables - Identificadores y años
- **Variables categóricas (object)**: 3 variables - Metadatos geográficos y de clasificación

3. Evaluación De Valores Faltantes

Procedimiento realizado:
Se realizó un análisis sistemático de valores faltantes en todas las variables, calculando porcentajes y patrones de distribución.

Hallazgos cuantificados:

```
======================================================================
ANÁLISIS DE VALORES FALTANTES
======================================================================
Variables con valores faltantes:
   - UN Country ID: 124 valores faltantes (2.0%)
   - Physical Integrity Rights Index: 1347 valores faltantes (21.5%)
   - Disappearance: 1217 valores faltantes (19.4%)
   - Extrajudicial Killing: 1219 valores faltantes (19.5%)
   - Political Imprisonment: 1219 valores faltantes (19.5%)
   - Torture: 1219 valores faltantes (19.5%)
   - Empowerment Rights Index (Old): 2287 valores faltantes (36.5%)
   - Empowerment Rights Index (New): 1329 valores faltantes (21.2%)
   - Freedom of Assembly and Association: 1219 valores faltantes (19.5%)
   - Freedom of Foreign Movement: 524 valores faltantes (8.4%)
   - Freedom of Domestic Movement: 525 valores faltantes (8.4%)
   - Freedom of Movement (Old): 2180 valores faltantes (34.8%)
   - Freedom of Speech: 1219 valores faltantes (19.5%)
   - Electoral Self-Determination: 1219 valores faltantes (19.5%)
   - Freedom of Religion (Old): 2188 valores faltantes (34.9%)
   - Freedom of Religion (New): 622 valores faltantes (9.9%)
   - Employment Rights: 1219 valores faltantes (19.5%)
   - Women’s Economic Rights: 1220 valores faltantes (19.5%)
   - Women’s Political Rights: 1220 valores faltantes (19.5%)
```

```
   - Women's Social Rights: 2380 valores faltantes (38.0%)
   - Independence of Judiciary: 573 valores faltantes (9.2%)

Total variables con valores faltantes: 21/26
Porcentaje de variables con valores faltantes: 80.8%
Completitud global del dataset: 83.9%

Variables sin valores faltantes: 5
   - Year
   - CIRI Country ID
   - Country
   - UN Region
   - UN Subregion
```

Interpretación:
Se identifica un patrón heterogéneo de completitud, donde los metadatos básicos mantienen integridad total, mientras que variables específicas de derechos muestran niveles significativos de valores faltantes, particularmente en categorías relacionadas con derechos de la mujer y libertades civiles.

4. Distribución De Datos En Columnas Clave

Procedimiento realizado:
Se analizó la distribución estadística de variables fundamentales del dataset para comprender los rangos, valores típicos y distribución de los indicadores de derechos humanos antes de proceder con la identificación de valores especiales.

Resultados del análisis:

```
DISTRIBUCIÓN DE DATOS EN COLUMNAS CLAVE:
================================================

Physical Integrity Rights Index:
  - Valores no nulos: 4915 (78.5%)
  - Rango: 0.0 a 8.0
  - Media: 4.91
  - Valores únicos: [np.float64(0.0), np.float64(1.0), np.float64(2.0),
np.float64(3.0), np.float64(4.0), np.float64(5.0), np.float64(6.0),
np.float64(7.0), np.float64(8.0)]

Disappearance:
  - Valores no nulos: 5045 (80.6%)
  - Rango: -999.0 a 2.0
  - Media: -2.54
  - Valores únicos: [np.float64(-999.0), np.float64(-77.0), np.float64(-66.0),
```

```
np.float64(0.0), np.float64(1.0), np.float64(2.0)]

Torture:
  - Valores no nulos: 5043 (80.5%)
  - Rango: -999.0 a 2.0
  - Media: -2.61
  - Valores únicos: [np.float64(-999.0), np.float64(-77.0), np.float64(-66.0),
np.float64(0.0), np.float64(1.0), np.float64(2.0)]

Empowerment Rights Index (New):
  - Valores no nulos: 4933 (78.8%)
  - Rango: 0.0 a 14.0
  - Media: 8.28
  - Valores únicos: [np.float64(0.0), np.float64(1.0), np.float64(2.0),
np.float64(3.0), np.float64(4.0), np.float64(5.0), np.float64(6.0),
np.float64(7.0), np.float64(8.0), np.float64(9.0), np.float64(10.0),
np.float64(11.0), np.float64(12.0), np.float64(13.0), np.float64(14.0)]

Freedom of Speech:
  - Valores no nulos: 5043 (80.5%)
  - Rango: -77.0 a 2.0
  - Media: -0.43
  - Valores únicos: [np.float64(-77.0), np.float64(-66.0), np.float64(0.0),
np.float64(1.0), np.float64(2.0)]
```

Interpretación:
El análisis de distribución revela hallazgos críticos sobre la calidad de los datos. Mientras que el `Physical Integrity Rights Index` muestra una distribución esperada (rango 0-8, media 4.91) sin valores anómalos, se identifican valores negativos extremos (-999, -77, -66) en variables clave como `Disappearance` y `Torture`.

Estos valores negativos distorsionan significativamente las medidas estadísticas: la media de `Disappearance` resulta en -2.54 y la de `Torture` en -2.61, valores ilógicos para escalas que deberían oscilar entre 0-2. El `Empowerment Rights Index` presenta un rango expandido (0-14) con media de 8.28, mostrando una distribución más favorable pero requiriendo validación de la escala documentada.

5. Análisis De Valores Especiales

Procedimiento realizado:
Se investigó la presencia de valores especiales codificados (-999, -77, -66) que podrían representar valores faltantes o casos especiales en el esquema de codificación original.

Resultados del análisis:

```
======================================================================
ANÁLISIS DE VALORES ESPECIALES (-999, -77, -66)
======================================================================
CONTEO GLOBAL DE VALORES ESPECIALES:
   - Valor -999: 532 ocurrencias totales
     En 14 variables diferentes
   - Valor -77: 1130 ocurrencias totales
     En 17 variables diferentes
   - Valor -66: 491 ocurrencias totales
     En 17 variables diferentes

RESUMEN:
   • Total valores -999: 532
   • Total valores -77: 1130
   • Total valores -66: 491

ANÁLISIS DE IMPACTO:
   Ejemplo en 'Disappearance':
   - Media actual (con valores especiales): -2.54
   - Valores especiales en esta variable: {-999: np.int64(14), -77: np.int64(67),
-66: np.int64(29)}
   - Media limpia (sin valores especiales): 1.67
   - Diferencia absoluta: 4.21
   - Distorsión porcentual: 252.4%

DISTRIBUCIÓN EN 'Disappearance':
   Valores únicos y sus frecuencias:
     -999.0: 14 ocurrencias (ESPECIAL)
     -77.0: 67 ocurrencias (ESPECIAL)
     -66.0: 29 ocurrencias (ESPECIAL)
     0.0: 460 ocurrencias (normal)
     1.0: 718 ocurrencias (normal)
     2.0: 3757 ocurrencias (normal)
```

Interpretación:
Los valores especiales constituyen una codificación no estándar para valores faltantes que distorsiona significativamente las estadísticas descriptivas. Su presencia requiere un proceso de limpieza antes de cualquier análisis cuantitativo.

Distribución Real (excluyendo valores especiales):

- **0.0** (Nunca ocurrió): 460 casos (8.8%)
- **1.0** (Ocasionalmente): 718 casos (13.7%)
- **2.0** (Frecuentemente): 3,757 casos (71.7%)

5.1. Análisis De Patrones Temporales y Geográficos

Procedimiento realizado:
Se realizó un análisis exhaustivo de la distribución temporal y geográfica de los valores faltantes para identificar patrones sistemáticos en la calidad de los datos a lo largo del tiempo y entre diferentes regiones.

Resultados del análisis:

```
======================================================================
PATRONES TEMPORALES Y GEOGRÁFICOS
======================================================================

PATRÓN TEMPORAL: Valores faltantes por año:
   1981: 22.7%
   1982: 22.5%
   1983: 22.4%
   1984: 22.0%
   1985: 22.0%
   1986: 21.9%
   1987: 21.8%
   1988: 21.7%
   1989: 21.7%
   1990: 21.8%
   1991: 21.9%
   1992: 15.6%
   1993: 14.5%
   1994: 14.5%
   1995: 14.5%
   1996: 14.2%
   1997: 14.3%
   1998: 14.4%
   1999: 14.3%
   2000: 14.2%
   2001: 3.9%
   2002: 14.0%
   2003: 4.1%
   2004: 4.1%
   2005: 7.6%
   2006: 7.3%
   2007: 14.5%
   2008: 18.2%
   2009: 17.9%
   2010: 17.9%
   2011: 17.6%
```

```
PATRÓN GEOGRÁFICO: Valores Faltantes por región:
   Africa: 9.5%
   Asia: 13.2%
   Europa: 23.5%
   Latinoamérica y el Caribe: 14.0%
   Norte America: 2.6%
   Oceania: 33.3%
```

Interpretación:
El análisis temporal revela una evolución marcada en la calidad de los datos a lo largo de las tres décadas. Se identifican cuatro períodos claramente diferenciados: una fase inicial (1981-1991) con alta incidencia de valores faltantes (~22%), seguida de una mejora moderada (1992-2000) con tasas alrededor del 14-15%. El período 2001-2006 destaca por su excelente calidad, con mínimos históricos del 3.9-4.1% en 2001-2004, aunque se observa una anomalía en 2002 con un repunto al 14%. Finalmente, se constata un deterioro progresivo en la fase final (2007-2011) con tasas que alcanzan el 18%.

Geográficamente, la distribución muestra disparidades significativas. Norteamérica presenta la mejor cobertura (2.6%), seguida de África (9.5%), mientras Europa (23.5%) y especialmente Oceanía (33.3%) muestran deficiencias críticas en la completitud de datos. Estas variaciones regionales podrían reflejar diferencias en la capacidad de recolección, prioridades de investigación o acceso a información.

5.2. Análisis Final De Patrones

Procedimiento realizado:
Se integraron los hallazgos de los análisis previos para desarrollar una comprensión holística de los patrones de calidad de datos, identificando correlaciones entre las dimensiones temporal, geográfica y temática.

Resultados del análisis:

```
======================================================================
ANÁLISIS FINAL DE PATRONES - RESUMEN EJECUTIVO
======================================================================
RESUMEN EJECUTIVO DE PATRONES IDENTIFICADOS:

PATRÓN TEMPORAL:
   - 1981-1991: ~22% valores faltantes (alta falta de datos)
   - 1992-2000: ~14-15% valores faltantes (mejora moderada)
   - 2001-2006: 4-8% valores faltantes (mejora significativa)
   - 2007-2011: ~18% valores faltantes (retroceso)
```

```
PATRÓN GEOGRÁFICO:
   - Norteamérica: 2.6% valores faltantes (excelente)
   - África: 9.5% valores faltantes (bueno)
   - Asia: 13.2% valores faltantes (regular)
   - Latinoamérica: 14.0% valores faltantes (regular)
   - Europa: 23.5% valores faltantes (problemático)
   - Oceanía: 33.3% valores faltantes (crítico)

PATRÓN DE 19.5%:
   - 10 variables afectadas con datos faltantes casi idénticos
   - 1,215 registros (19.4%) faltan las 10 variables simultáneamente
   - 5,040 registros (80.5%) completos para estas variables

IMPLICACIONES PARA ANÁLISIS:
   - Ventaja: Patrones sistemáticos facilitan estrategias de manejo
   - Desafío: Valores Faltantes no aleatorios pueden introducir sesgos
   - Oportunidad: Período 2001-2006 tiene mejor calidad de datos
   - Acción: Oceanía y Europa requieren atención especial

COMPLETITUD POR CATEGORÍA DE VARIABLES:
   Metadatos: 100.0%
   Índices Principales: 78.6%
   Derechos Fundamentales: 80.5%
```

Interpretación:
La integración de los patrones identificados revela una estructura sistemática en la calidad de los datos que trasciende las dimensiones individuales. La correlación entre los patrones temporales y el agrupamiento del 19.5% sugiere que los problemas de recolección afectaron conjuntos específicos de variables durante períodos definidos, particularmente en ciertas regiones.

La identificación del período 2001-2006 como ventana de máxima calidad ofrece oportunidades analíticas significativas, mientras que las deficiencias en Oceanía y Europa requieren consideración especial en el diseño metodológico. El patrón de "todo o nada" observado en el 19.5% de los registros indica que las decisiones sobre manejo de valores faltantes pueden implementarse de manera consistente para bloques completos de variables.

Estos hallazgos proporcionan una base sólida para desarrollar estrategias diferenciadas de limpieza y análisis, permitiendo optimizar el uso de los datos disponibles mientras se mitigan los sesgos introducidos por los patrones no aleatorios de valores faltantes.

6. Resumen Ejecutivo Final

Procedimiento realizado:
Se consolidaron todos los hallazgos del análisis exploratorio para generar una visión integral del estado del dataset, integrando las evidencias recopiladas en las fases anteriores y formulando recomendaciones estratégicas para las siguientes etapas del proyecto.

Resultados del análisis:

```
======================================================================
6. RESUMEN EJECUTIVO FINAL - EXPLORACIÓN DE DATOS CIRI
======================================================================
METRICS GLOBALES CONFIRMADAS:
   • Registros: 6,262 observaciones país-año
   • Variables: 26 indicadores
   • Período: 1981-2011 (31 años)
   • Países: 202 países únicos

CALIDAD DE DATOS - RESUMEN:
   • Variables sin valores faltantes: 5/26 (19.2%)
   • Variables con >30% de valores faltantes: 4/26 (15.4%)
   • Completitud global: 83.9%
   • Valores especiales identificados: 2,153 (-999, -77, -66)

PATRONES CLAVE IDENTIFICADOS:
   1. TEMPORAL: Mejor calidad 2001-2006 (4-8% valores faltantes)
   2. GEOGRÁFICO: Oceanía (33.3%) y Europa (23.5%) críticos
   3. SISTEMÁTICO: 10 variables con 19.5% valores faltantes agrupados
   4. DISTORSIÓN: Valores especiales afectan estadísticas >250%

RECOMENDACIONES INMEDIATAS:
   • Limpieza prioritaria de valores especiales (-999, -77, -66)
   • Estrategias diferenciadas por patrón de valores faltantes
   • Enfocar análisis en períodos/regiones de alta calidad
   • Documentar todas las decisiones de limpieza

ESTADO: EDA COMPLETADO - LISTO PARA FASE DE LIMPIEZA
   Se han identificado y cuantificado todos los patrones relevantes
   para informar decisiones técnicas en la siguiente fase del proyecto.
```

Interpretación:
El análisis exploratorio ha proporcionado una caracterización exhaustiva del dataset CIRI, confirmando su viabilidad para análisis avanzados mientras se identifican limitaciones específicas que requieren atención. La completitud global del 83.9% representa un nivel aceptable para datos históricos de derechos humanos, aunque la distribución desigual temporal y geográficamente exige consideraciones metodológicas cuidadosas.

Los patrones sistemáticos identificados - particularmente el agrupamiento del 19.5% de valores faltantes y la distorsión estadística por valores especiales - constituyen hallazgos críticos que deben guiar el proceso de limpieza. La ventana temporal 2001-2006 emerge como un período de referencia por su excelente calidad, mientras que las regiones de Oceanía y Europa requieren aproximaciones específicas dado su bajo nivel de completitud.

El proyecto se encuentra en condiciones óptimas para avanzar a la fase de limpieza de datos, con una base documental sólida que permitirá tomar decisiones informadas sobre el manejo de valores faltantes, la corrección de distorsiones estadísticas y la optimización del conjunto de datos para análisis subsiguientes.