Informe Detallado: Análisis de Datos sobre IA

Autor: Equipo de Análisis de Datos

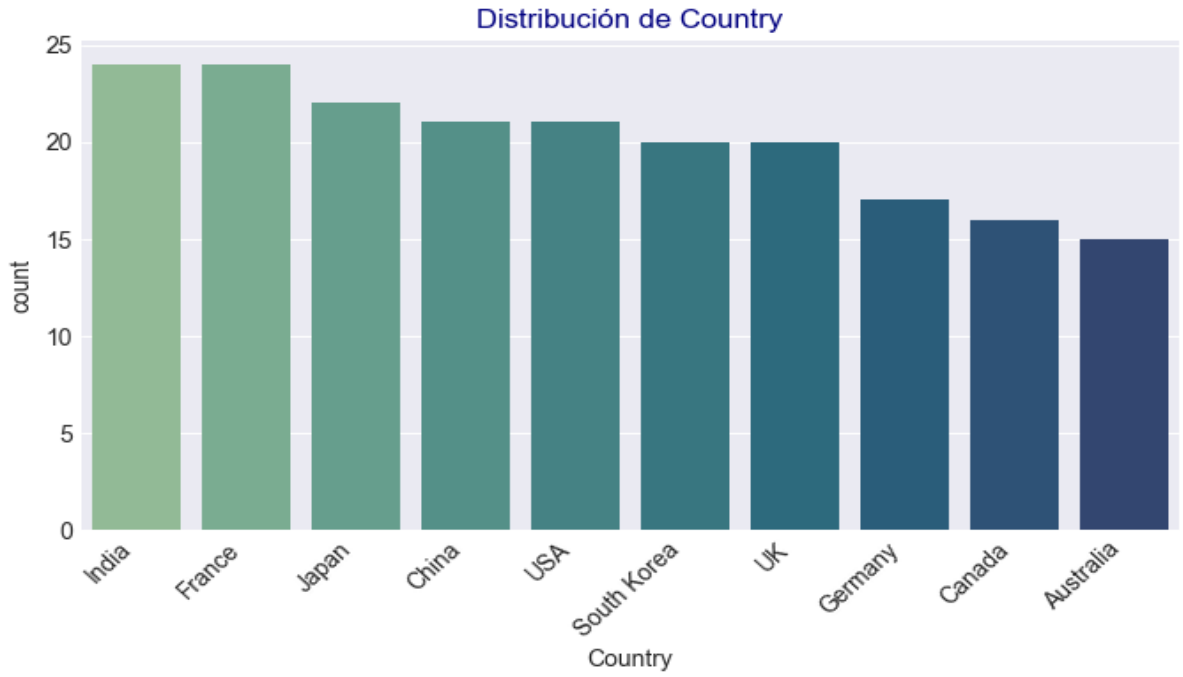
Contenido: ETL, EDA inicial y EDA final con explicaciones detalladas de cada gráfico.

1. Resumen del Proceso ETL

Se trabajó con un dataset original de dimensiones (200, 12). Durante el proceso ETL se realizaron las siguientes acciones principales: Renombrado y estandarización de columnas para mejorar legibilidad. Verificación y eliminación de duplicados: **No se encontraron duplicados**. Comprobación de valores faltantes: **No se detectaron valores nulos** en las columnas analizadas. Validación de consistencia de variables y generación del archivo limpio: **Global_AI_Content_Clean.csv**. Este documento presenta los resultados del EDA inicial y final, con cada gráfico acompañado de una explicación detallada y recomendaciones operativas.

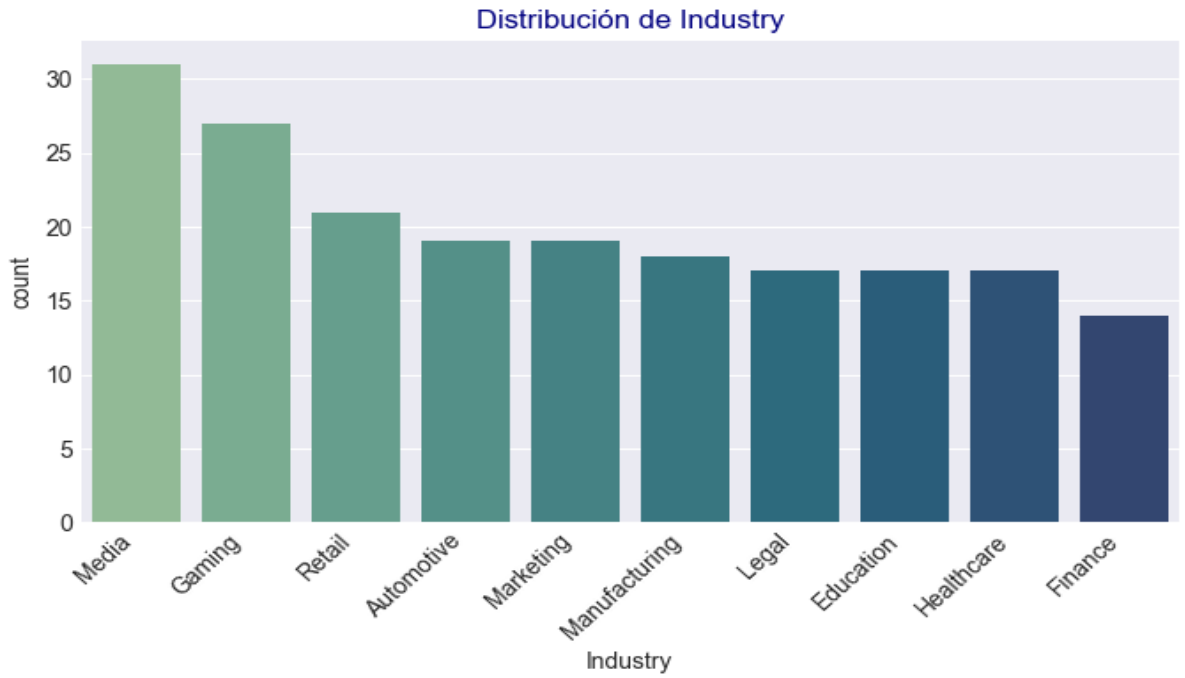
2. EDA Inicial - Gráficos y explicaciones

Figura 1: Distribución de Country



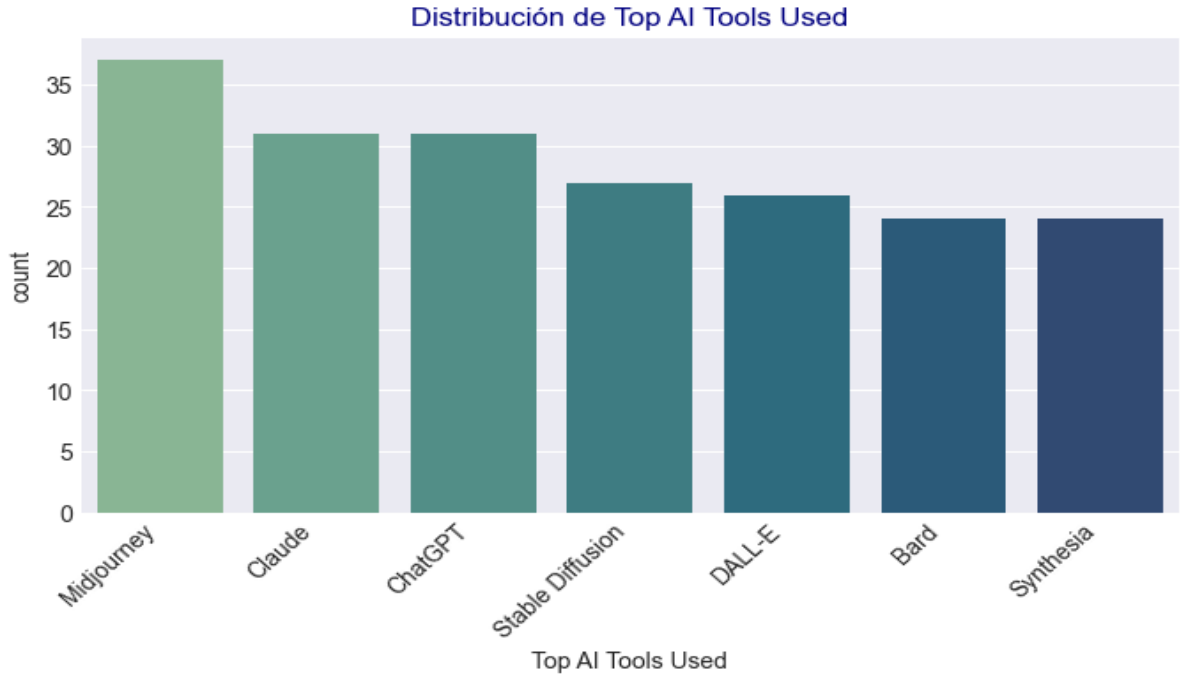
Este gráfico de barras muestra la frecuencia de registros por país. India y Francia son los países con mayor representación (~24 registros cada uno), lo que indica un sesgo de muestra hacia esas regiones. Implicaciones: tener mayor representación de unos países puede influir en métricas agregadas (por ejemplo, adopción media) si esos países comparten características comunes. Recomendación: si se busca una muestra más balanceada, considerar recolectar más datos de países con baja representación (Australia, Canadá) o aplicar ponderación en análisis posteriores.

Figura 2: Distribución de Industry



Bar chart que muestra la cantidad de registros por industria. 'Media' es la industria más frecuente (≈31), seguida por 'Gaming' (≈27) y 'Retail' (≈21). Interpretación: el dataset está sesgado hacia industrias creativas/entretenimiento, lo que explica la fuerte presencia de herramientas generativas observadas más adelante. Impacto analítico: métricas como 'ai_generated_content_volume' pueden estar infladas por la predominancia de sectores que generan mucho contenido. Sugerencia: segmentar análisis por industria para evitar generalizaciones que no apliquen a sectores menos representados (ej. Finance, Healthcare).

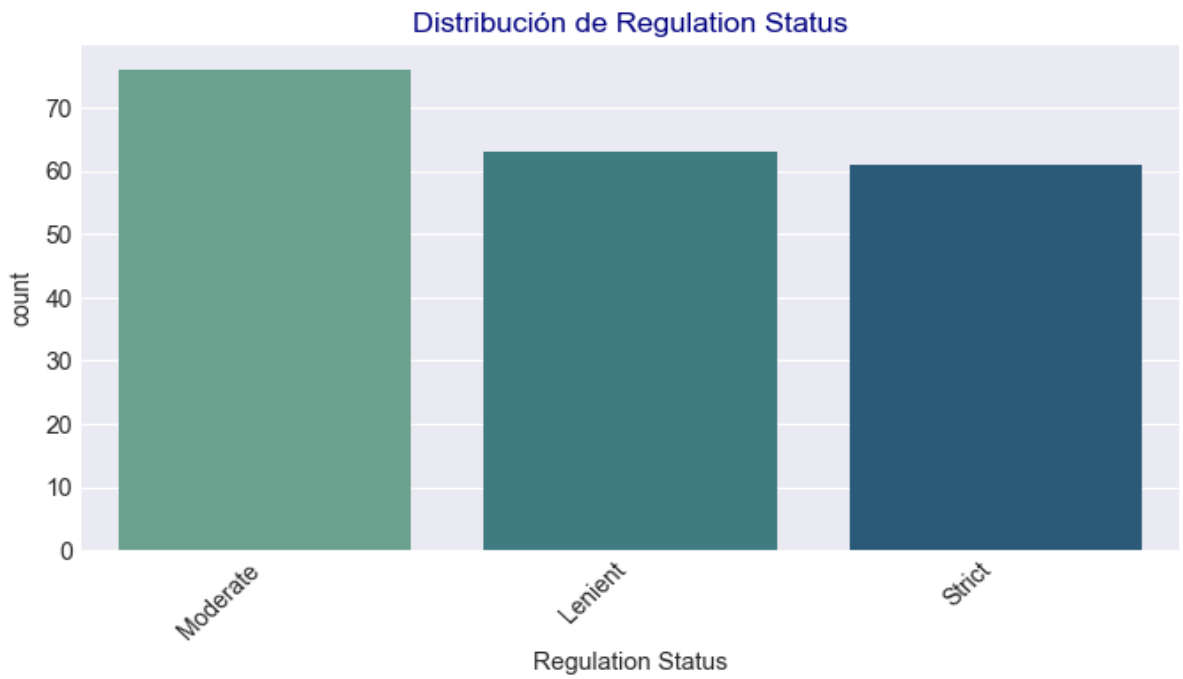
Figura 3: Distribución de Top AI Tools Used



Este gráfico muestra las herramientas de IA más reportadas. Midjourney encabeza la lista (~37), seguido por Claude y ChatGPT (~31 cada uno). También aparecen Stable Diffusion, DALL-E, Bard y Synthesia. Conclusión: fuerte sesgo hacia herramientas de generación de imágenes y modelos conversacionales. Esto sugiere que el dataset recoge mayormente casos de uso creativos y generativos. Recomendación: en análisis de impacto (ej.

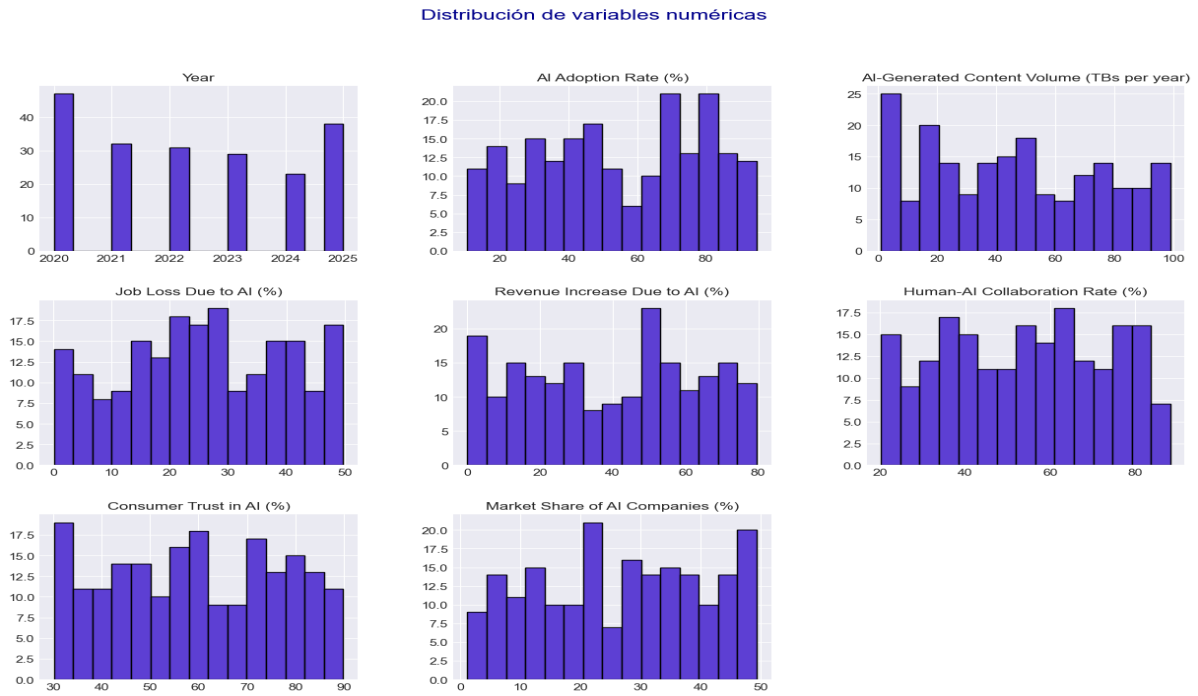
revenue, job loss) controlar por tipo de herramienta (generativa vs analítica) ya que el efecto esperado puede variar sensiblemente.

Figura 4: Distribución de Regulation Status



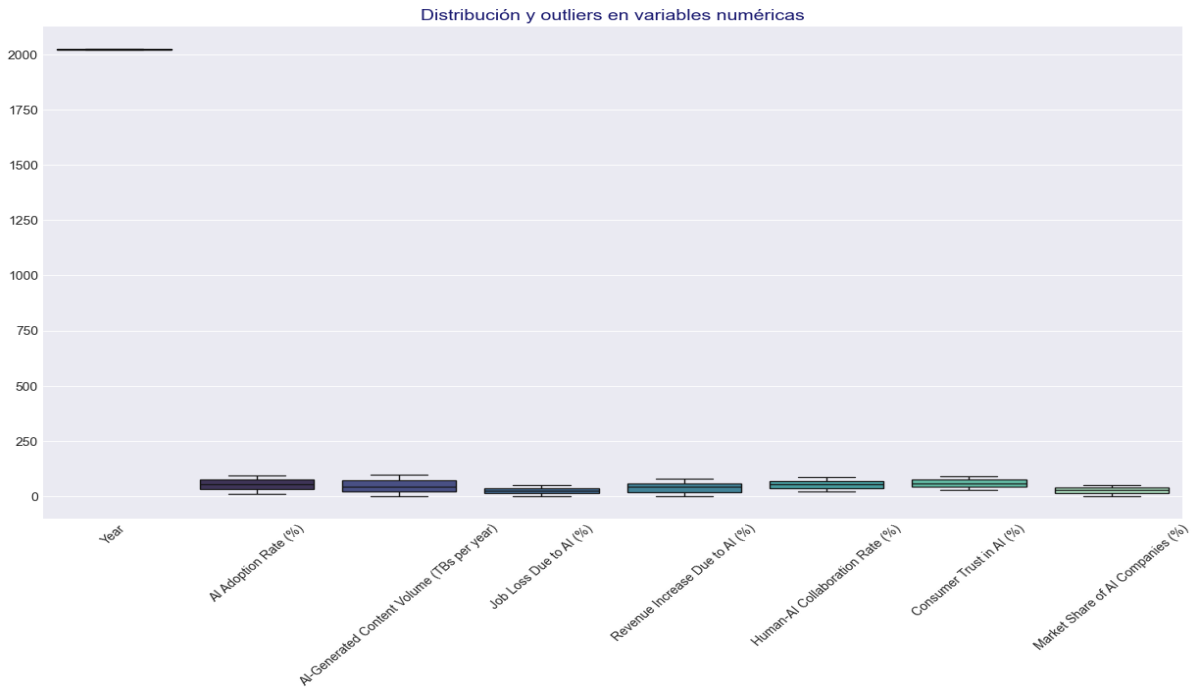
Bar chart con el estado regulatorio reportado. El estado 'Moderate' es el más frecuente (~76), seguido por 'Lenient' (~63) y 'Strict' (~61). Interpretación: predominio de marcos regulatorios intermedios, lo que puede favorecer adopciones prácticas de IA con ciertas restricciones. Implicaciones: los hallazgos sobre adopción y confianza deben leerse considerando el contexto regulatorio; sectores en países 'Strict' podrían mostrar comportamientos distintos. Acción sugerida: cruzar 'regulation_status' con 'ai_adoption_rate' para medir efectos regulatorios sobre la adopción.

Figura 5: Histogramas de variables numéricas



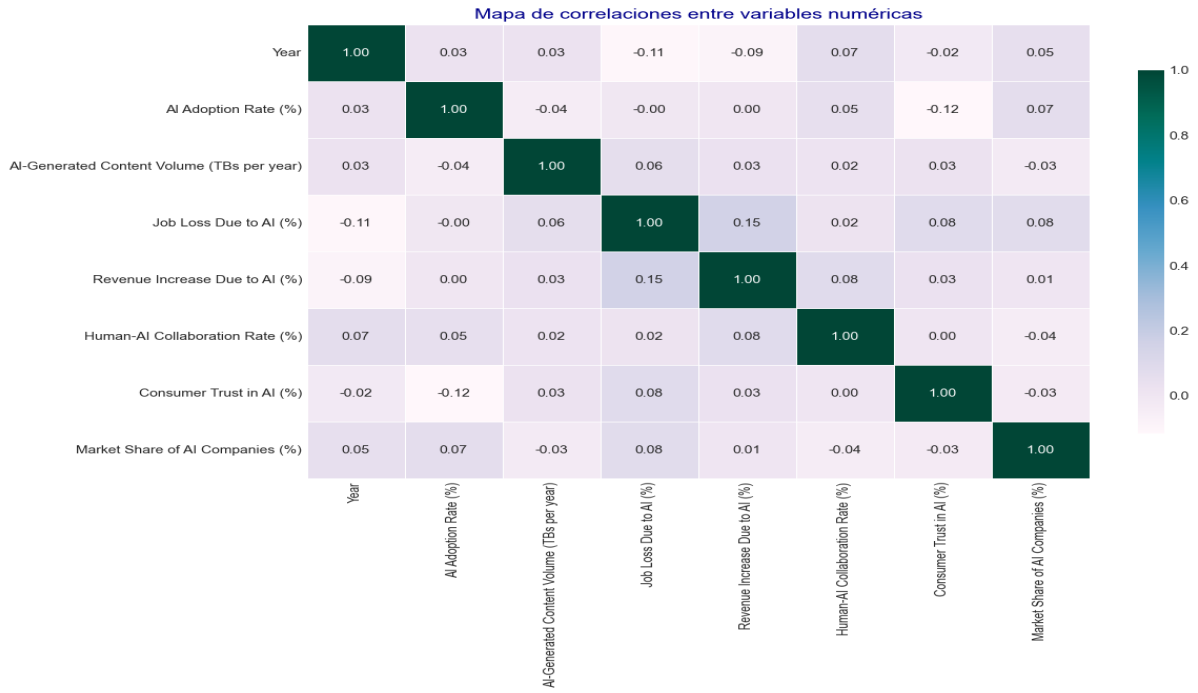
Panel de histogramas que muestra la distribución de las variables numéricas principales: - 'Year': registros distribuidos entre 2020 y 2025, con picos en 2020 y 2025. - 'AI Adoption Rate': amplio rango (aprox. 10% a 90%), con varias densidades intermedias que sugieren heterogeneidad entre organizaciones. - 'AI-Generated Content Volume (TBs/year)': amplia dispersión de 0 a ~100 TBs, con acumulaciones en valores bajos y medios. - 'Job Loss', 'Revenue Increase', 'Human-AI Collaboration', 'Consumer Trust' y 'Market Share': muestran variaciones y diferentes sesgos (algunas ligeramente sesgadas a la derecha/izquierda). Interpretación: varias variables presentan multimodalidad o dispersión alta, lo que sugiere subgrupos dentro de la muestra (p. ej. organizaciones con adopción alta vs baja). Recomendación: aplicar transformaciones (log) para variables con sesgo fuerte y crear segmentaciones por industria/región para análisis más finos.

Figura 6: Boxplots y detección de outliers



Boxplots que muestran la mediana, cuartiles y outliers por variable. Se observan: - Outliers prominentes en 'AI-Generated Content Volume' (organizaciones que generan cantidades muy superiores a la mediana). - El eje 'Year' aparece con escala distinta (valores discretos 2020–2025), por lo que su boxplot no es directamente comparable con las demás métricas numéricas. Impacto: los outliers pueden sesgar medidas resumen (media, desviación). Es necesario decidir si tratarlos (capar, winsorize) o mantenerlos según el objetivo del análisis. Sugerencia: documentar y revisar casos extremos (posible error de entrada o empresas con volumes excepcionalmente altos) antes de excluirlos.

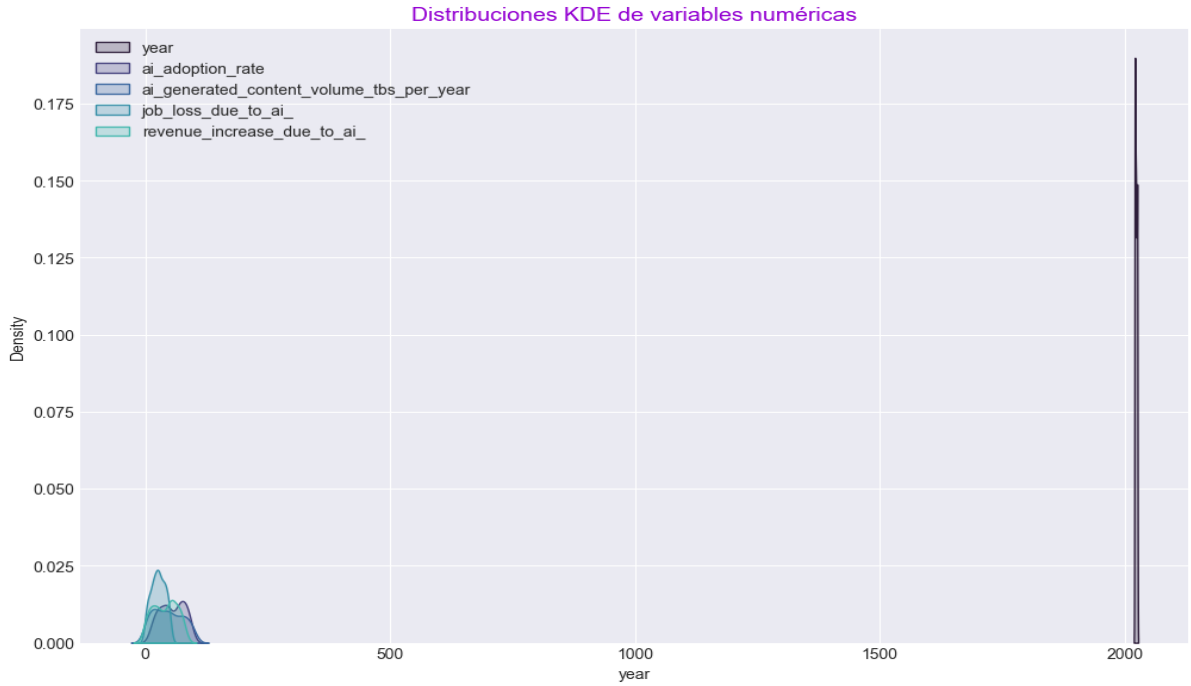
Figura 7: Mapa de correlaciones



Heatmap que muestra coeficientes de correlación entre variables numéricas. Observaciones clave: - Correlaciones débiles en general (la mayoría cercanas a 0). - La mayor correlación positiva observada es entre 'Job Loss Due to AI' y 'Revenue Increase' (≈+0.15). - Existe una correlación negativa leve entre 'AI Adoption Rate' y 'Consumer Trust' (≈-0.12). Interpretación: las variables no están fuertemente correlacionadas en este conjunto de datos, lo que sugiere que los factores que determinan cada métrica pueden ser independientes o no lineales. Recomendación: usar análisis adicionales (regresión multivariante, modelos no lineales o técnicas de causalidad) para explorar relaciones más complejas y controlar variables confusoras.

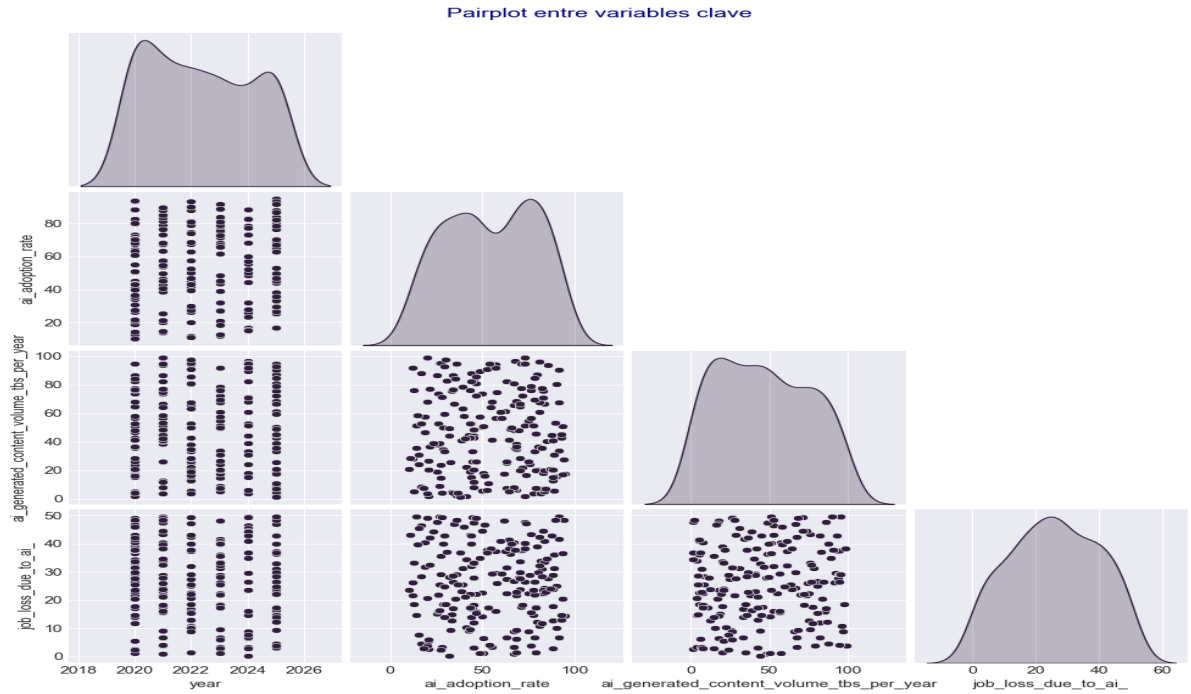
3. EDA Final - Gráficos y explicaciones detalladas

Figura 8: Distribuciones KDE de variables numéricas (EDA Final)



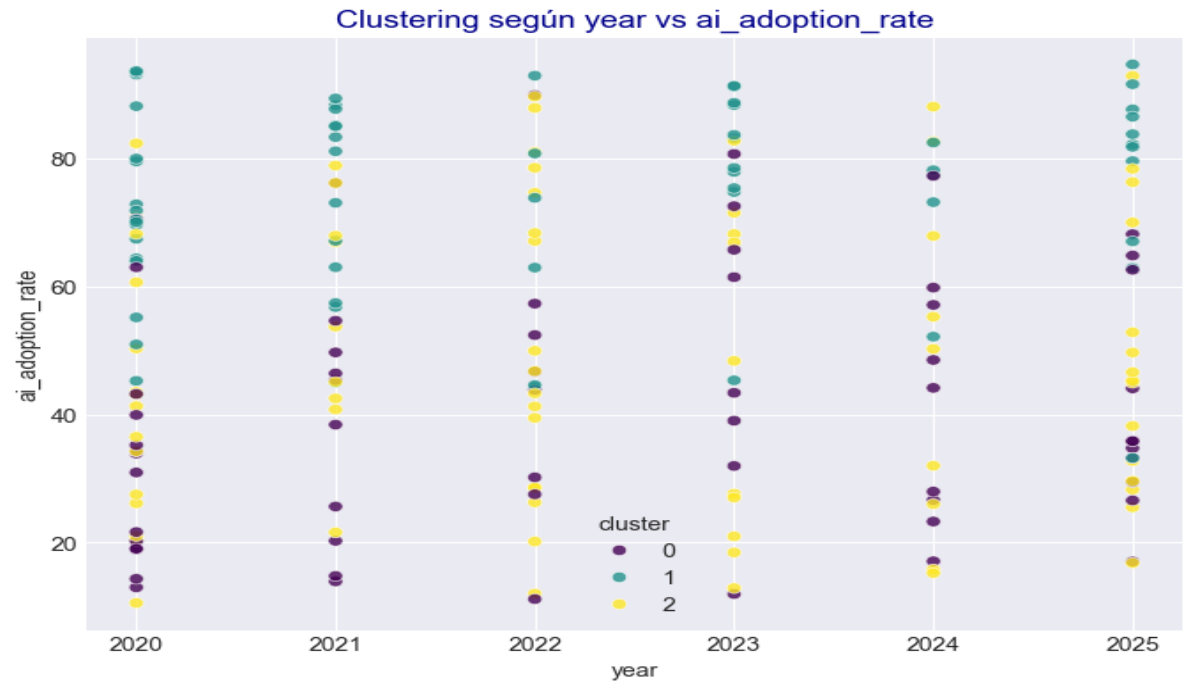
KDE (Kernel Density Estimate) que muestra la 'densidad' estimada de varias variables continuas: - 'AI Adoption Rate' presenta picos bien definidos en rangos medios (~40–70%), lo que sugiere grupos de adopción intermedia y alta. - 'Job Loss' y 'Revenue Increase' muestran picos más discretos, lo que sugiere concentraciones de valores (p.ej. 10–30% para job loss y 10–60% para revenue increase). - La variable 'Year' aparece concentrada en valores discretos (2020–2025), por lo que su densidad no es comparable con las métricas continuas. Interpretación: las KDE permiten ver multimodalidad y la presencia de subpoblaciones; aquí se confirman grupos de adopción distintos. Acción: usar estos insights para definir segmentos (ej. campañas de adopción dirigidas a organizaciones del segmento medio).

Figura 9: Pairplot entre variables clave



El pairplot combina histogramas/densidades en la diagonal y scatterplots en las celdas inferiores, mostrando relaciones pairwise: - 'AI Adoption Rate' vs 'AI-Generated Content Volume': se aprecia una leve tendencia positiva (mayor adopción asociada a mayor volumen de contenido). - 'AI Adoption Rate' vs 'Job Loss': dispersión amplia sin patrón claro, lo que sugiere poca correlación lineal. - Diagonales (densidades) muestran la forma de distribución de cada variable (multimodalidades y sesgos). Sugerencia: el pairplot es útil para identificar relaciones iniciales y sospechas de no-linealidad que requerirán modelos específicos.

Figura 10: Clustering: year vs ai_adoption_rate (k=3)



Scatterplot con clusters (k=3) basado en 'ai_adoption_rate' y 'year'. Resultados observados: - Cluster 0 (adopción baja): puntos concentrados por debajo de ~40%. - Cluster 1 (adopción media): valores entre ~40% y ~70%. - Cluster 2 (adopción alta): valores por encima de ~70%. Interpretación: los clusters son consistentes a lo largo de los años, lo que indica que el nivel de adopción es una característica persistente por organización más que un

efecto temporal puro. Aplicación: estas segmentaciones se pueden usar para diseñar políticas o estrategias por nivel de madurez (formación, inversión, regulación diferenciada).

Figura 11: Relación entre Year y AI Adoption Rate (regresión local)

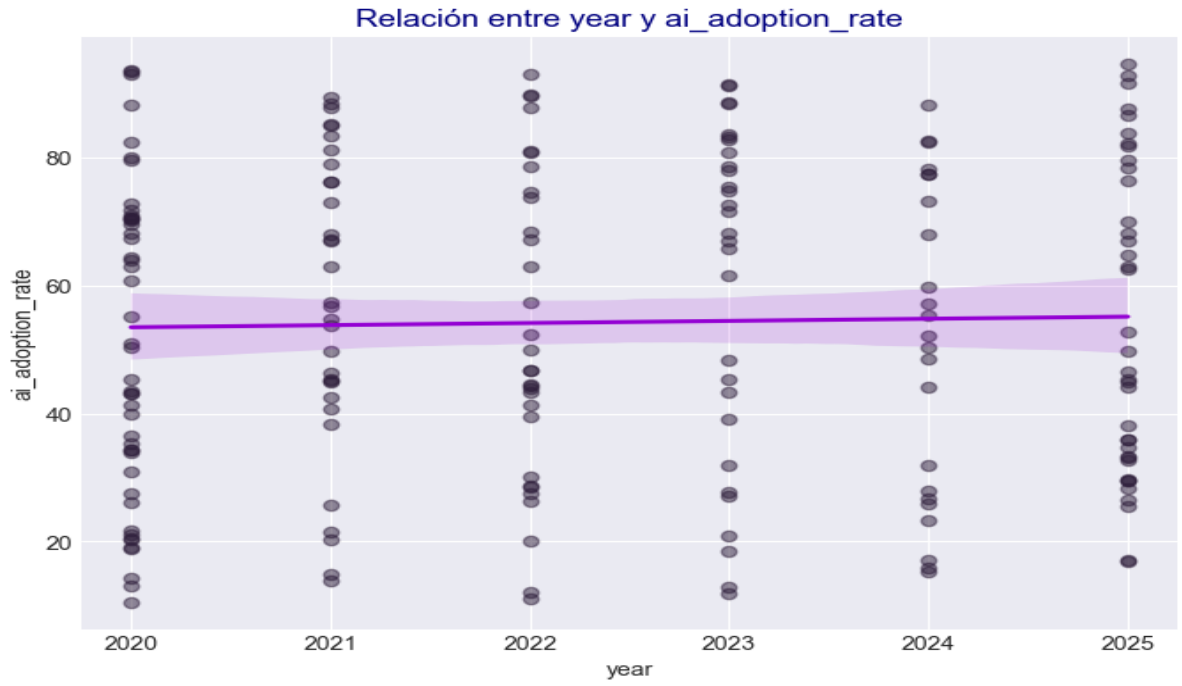


Gráfico scatter con línea de tendencia y banda de confianza que muestra la evolución temporal de la adopción: - Observamos una ligera pendiente positiva entre 2020 y 2025, lo que implica un aumento moderado en la tasa de adopción promedio. - La banda de confianza es amplia: alta varianza entre organizaciones en cada año. Interpretación: aunque hay un incremento promedio en adopción, la heterogeneidad por organización es grande; por tanto, políticas universales pueden no ser igualmente efectivas. Recomendación: modelar la tendencia con regresiones que incluyan variables de control (industria, país, regulación) para aislar el efecto temporal.

Figura 12: Distribución de registros por Year (balance temporal)

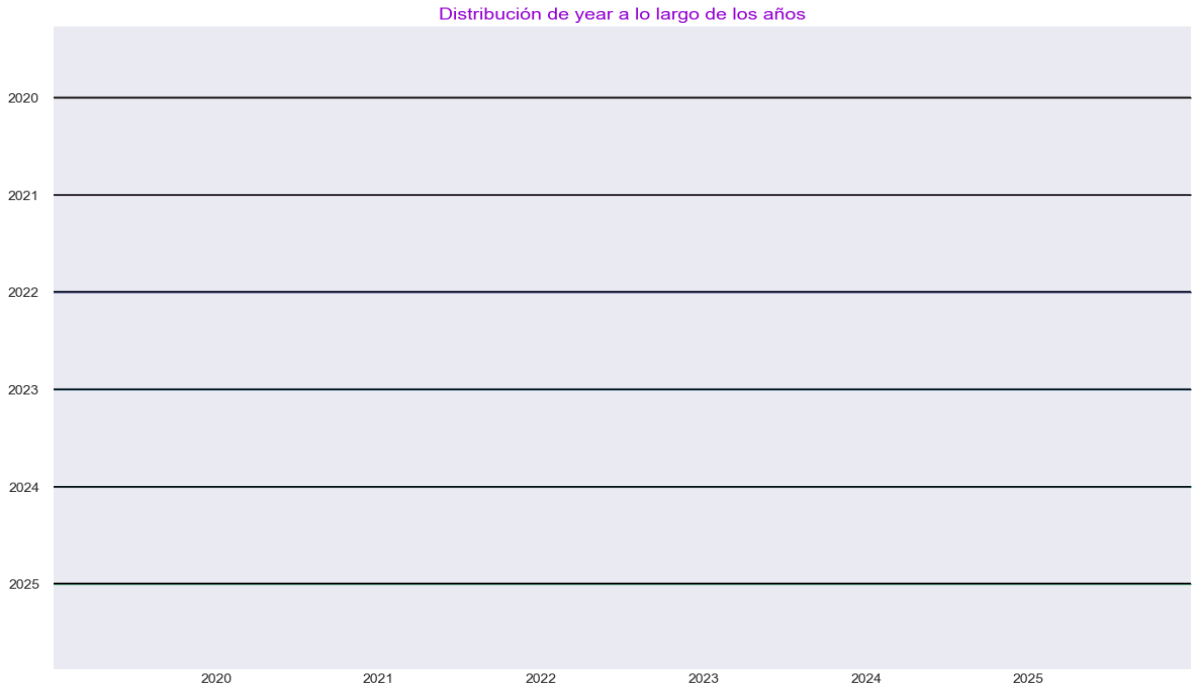


Gráfico que confirma distribución equilibrada de registros entre 2020 y 2025. No hay años con sobre-representación clara ni vacíos de datos. Consecuencia: validan las comparaciones interanuales y la estimación de tendencias temporales sin necesidad de reponderación. Siguiente paso: a pesar del balance, siempre verificar si cambios regulatorios o eventos exógenos por año afectan las métricas observadas (p. ej. cambios legislativos, crisis económicas).

4. Conclusiones, limitaciones y siguientes pasos

Conclusiones principales: La adopción de IA muestra una tendencia creciente entre 2020 y 2025, aunque con gran variabilidad entre organizaciones. Existen al menos tres niveles de madurez (baja, media, alta) estables a lo largo del tiempo. Las correlaciones lineales entre las principales métricas son débiles; pueden existir relaciones no lineales o efectos mediadores. Limitaciones del análisis: El dataset está sesgado hacia ciertas industrias (Media, Gaming) y países (India, Francia) lo que limita la generalización. Algunas variables presentan outliers significativos (p.ej. volumen de contenido) que requieren validación caso a caso. El análisis descriptivo no prueba causalidad; se recomienda análisis adicional para inferencias causales. Siguientes pasos recomendados: Realizar análisis segmentados por industria y país. Aplicar modelos multivariantes (regresión, random forest) para identificar drivers de adopción y revenue. Validar y documentar outliers antes de decidir su exclusión o tratamiento. Considerar series temporales o panel data si se dispone de observaciones longitudinales por organización.