



PROYECTO FINAL - ANÁLISIS DE DATOS


ESCUELA POLITÉCNICA NACIONAL

ARIEL SÁNCHEZ
RICHARD PADILLA
LENIN TACO
FRANCIS GUAMAN



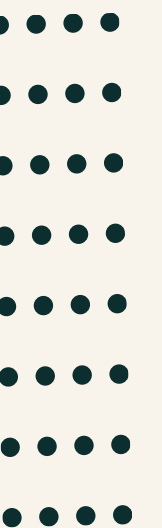


INTRODUCCIÓN



El análisis de datos es un proceso fundamental en la toma de decisiones en diversas áreas. Este proyecto explora el ciclo completo del análisis de datos, desde la recolección y limpieza hasta la transformación y visualización. Se han utilizado herramientas como Power BI y Python para facilitar este proceso.

Además, se han aplicado técnicas de análisis para descubrir patrones en diferentes conjuntos de datos, abarcando temas como deportes, seguridad vial, crowdfunding y ciberseguridad.





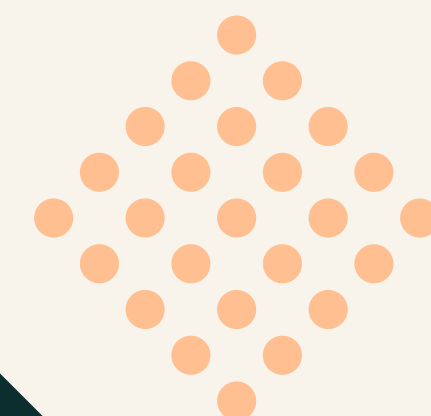
DEFINICION DEL CASO DE ESTUDIO

Este estudio se basa en técnicas avanzadas de procesamiento de datos para extraer información relevante de múltiples fuentes. Se busca identificar patrones y tendencias en diversas áreas de interés, con el fin de generar conocimiento valioso y facilitar la toma de decisiones informadas.

Los datos analizados provienen de fuentes heterogéneas y se han procesado utilizando bases de datos relacionales (SQL Server, SQLite) y NoSQL (MongoDB, CouchDB).

Fuentes:

- Informatica - Trafico
- Arte - Kickstarter
- Noticias Mundiales - Cybersecurity
- Deportes - Futbol
- Noticias - Incendios





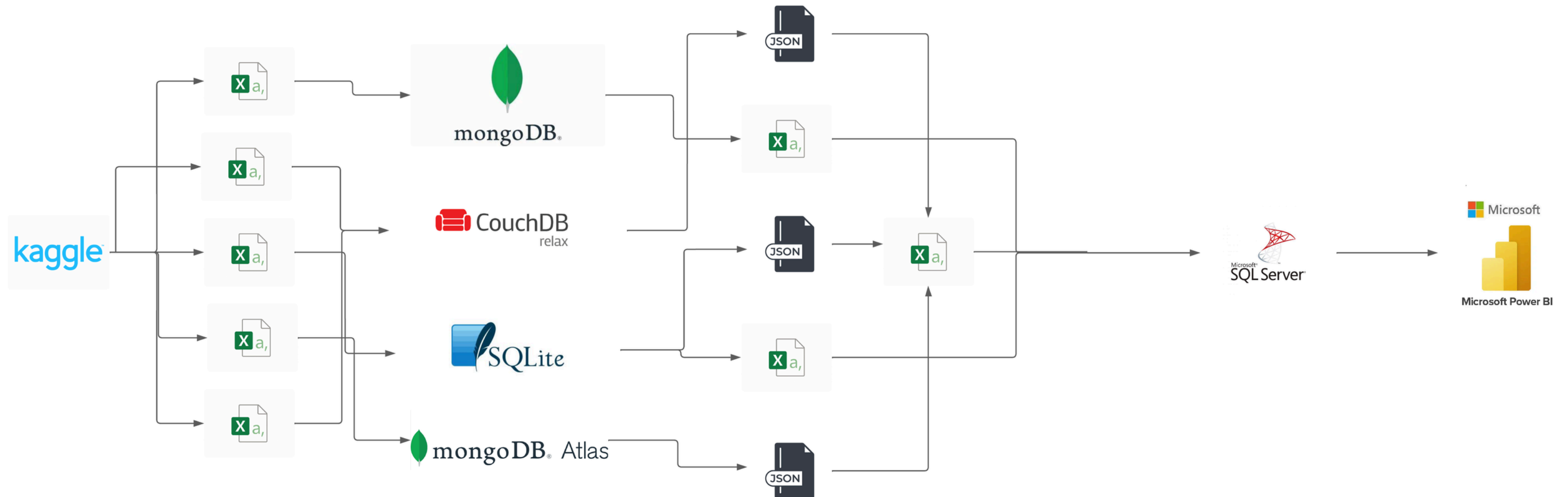
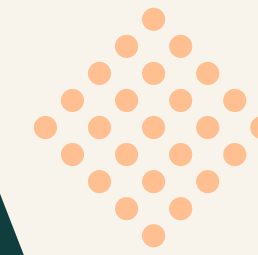
RECURSOS Y HERRAMIENTAS



- Software:
 - Python → Utilizado para la limpieza, transformación y conexión con bases de datos.
 - Power BI → Creación de dashboards interactivos para visualizar patrones y tendencias en los datos.
- Plataformas:
 - Jupyter Notebook → Plataforma utilizada para escribir código en Python, procesar los datos y hacer análisis exploratorio.
 - Power BI Desktop → Aplicación usada para conectar las bases de datos y diseñar visualizaciones interactivas.
- Bases de datos:
 - SQLite → Base de datos relacional utilizada para el almacenamiento temporal de datos antes de la migración a otros sistemas.
 - MySQL → Base de datos relacional usada para estructurar la información y realizar consultas SQL avanzadas.
 - MongoDB Atlas → Base de datos NoSQL para manejar datos semi-estructurados y escalables en la nube.
 - CouchDB → Base de datos NoSQL utilizada para la integración de datos y su transformación a JSON.
- Extras:
 - Excel → Gestionó tareas del equipo y permitió una primera revisión de los datasets antes de cargarlos en bases de datos.



ARQUITECTURA DE LA SOLUCION



EXTRACCION DE DATOS

Se seleccionaron 5 conjuntos de datos provenientes de Kaggle, cubriendo diferentes áreas temáticas

Fuentes:

- Informatica - Trafico
- Arte - Kickstarter
- Noticias Mundiales - Cybersecurity
- Deportes - Futbol
- Noticias - Incendios





LIMPIEZA DE DATOS

El proceso inicia con la conexión a SQLite desde Jupyter, donde se almacenan temporalmente los datos crudos. Aquí se realiza la limpieza utilizando herramientas como pandas para eliminar duplicados, corregir formatos y manejar valores

```
# Arte - Realizando Requisitos
import pandas as pd
import re
# CONECTAR A SQL LITE Y PASAR REGISTROS
import sqlite3

# 1. Cargar el CSV en un DataFrame de Pandas
csv_file = "male_players_23.csv"
df = pd.read_csv(csv_file)

# 2. Conectar o crear la base de datos SQLite
db_name = "BDDLite.db" # Nombre de la base de datos SQLite
conn = sqlite3.connect(db_name)

# 3. Insertar los datos del CSV en SQLite
table_name = "tabla" # Nombre de la tabla
df.to_sql(table_name, conn, if_exists="replace", index=False)

# 4. Cerrar la conexión
conn.close()

print(f"Datos del archivo '{csv_file}' importados con éxito a la base de datos SQLite")

Datos del archivo 'male_players_23.csv' importados con éxito a la base de datos SQLite
```





MIGRACION DE DATOS

Los datos limpios se migran a CouchDB, una base de datos NoSQL flexible que permite el manejo de datos no estructurados. Posteriormente, los datos se exportan desde CouchDB a un archivo JSON utilizando comando de terminal, evitando los límites de exportación de la interfaz web.

	_id	age	attacking_cro	attacking_fini	attacking_he
<input type="checkbox"/>	049f0d527300d5344e...	25	53	49	59
<input type="checkbox"/>	049f0d527300d5344e...	22	41	52	74
<input type="checkbox"/>	049f0d527300d5344e...	27	44	38	67
<input type="checkbox"/>	049f0d527300d5344e...	27	39	69	63
<input type="checkbox"/>	049f0d527300d5344e...	24	50	43	64
<input type="checkbox"/>	049f0d527300d5344e...	25	73	62	47
<input type="checkbox"/>	049f0d527300d5344e...	25	41	65	68
<input type="checkbox"/>	049f0d527300d5344e...	20	10	9	14
<input type="checkbox"/>	049f0d527300d5344e...	28	58	41	55
<input type="checkbox"/>	049f0d527300d5344e...	21	65	60	54

Showing 5 of 112 columns. ☐ Show all columns.

Showing document 801 - 900. Documents per page: 100





CARGA EN MONGODB ATLAS

El archivo JSON generado en CouchDB se importa a MongoDB Atlas, aprovechando la escalabilidad y el manejo de datos semiestructurados que ofrece MongoDB.

Select delimiter

Comma

☒ Ignore empty strings

☒ Stop on errors

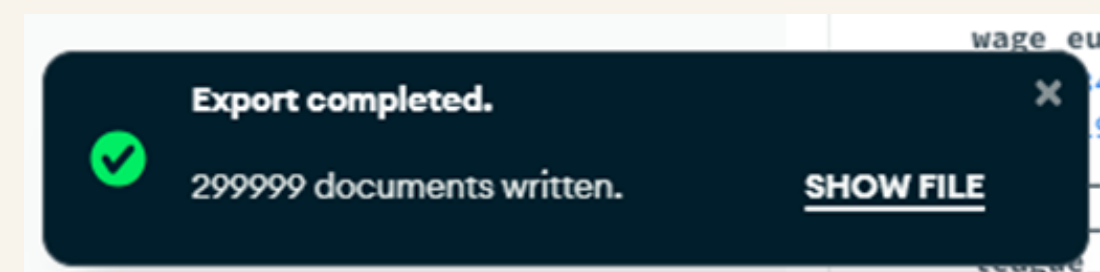
Specify Fields and Types

[Learn more about data types](#)

	<input checked="" type="checkbox"/> player_id	<input checked="" type="checkbox"/> player_url	<input checked="" type="checkbox"/> fifa_version	<input checked="" type="checkbox"/> fifa_update	<input checked="" type="checkbox"/> fifa_up
	Int32	String	Int32	Int32	Date
1	158023	/player/158023/lionel-messi/230009	23	9	2023-01-13
2	165153	/player/165153/karim-benzema/230009	23	9	2023-01-13
3	188545	/player/188545/robert-lewandowski/230...	23	9	2023-01-13
4	192985	/player/192985/kevin-de-bruyne/230009	23	9	2023-01-13

Cancel

Import





TRASLADO A SQL SERVER

Los datos se trasladan a SQL Server para su estructuración en un entorno relacional. Se realizan consultas SQL desde Jupyter para validar la carga, como el conteo de registros, antes de proceder a la integración con Power BI.

```
1 create database BDD_SERVER_FINAL;  
2 use BDD_SERVER_FINAL;
```





TRASLADO A SQL SERVER

Los datos se trasladan a SQL Server para su estructuración en un entorno relacional. Se realizan consultas SQL desde Jupyter para validar la carga, como el conteo de registros, antes de proceder a la integración con Power BI.

```
1 create database BDD_SERVER_FINAL;  
2 use BDD_SERVER_FINAL;  
3 SELECT * FROM serverfinal;
```

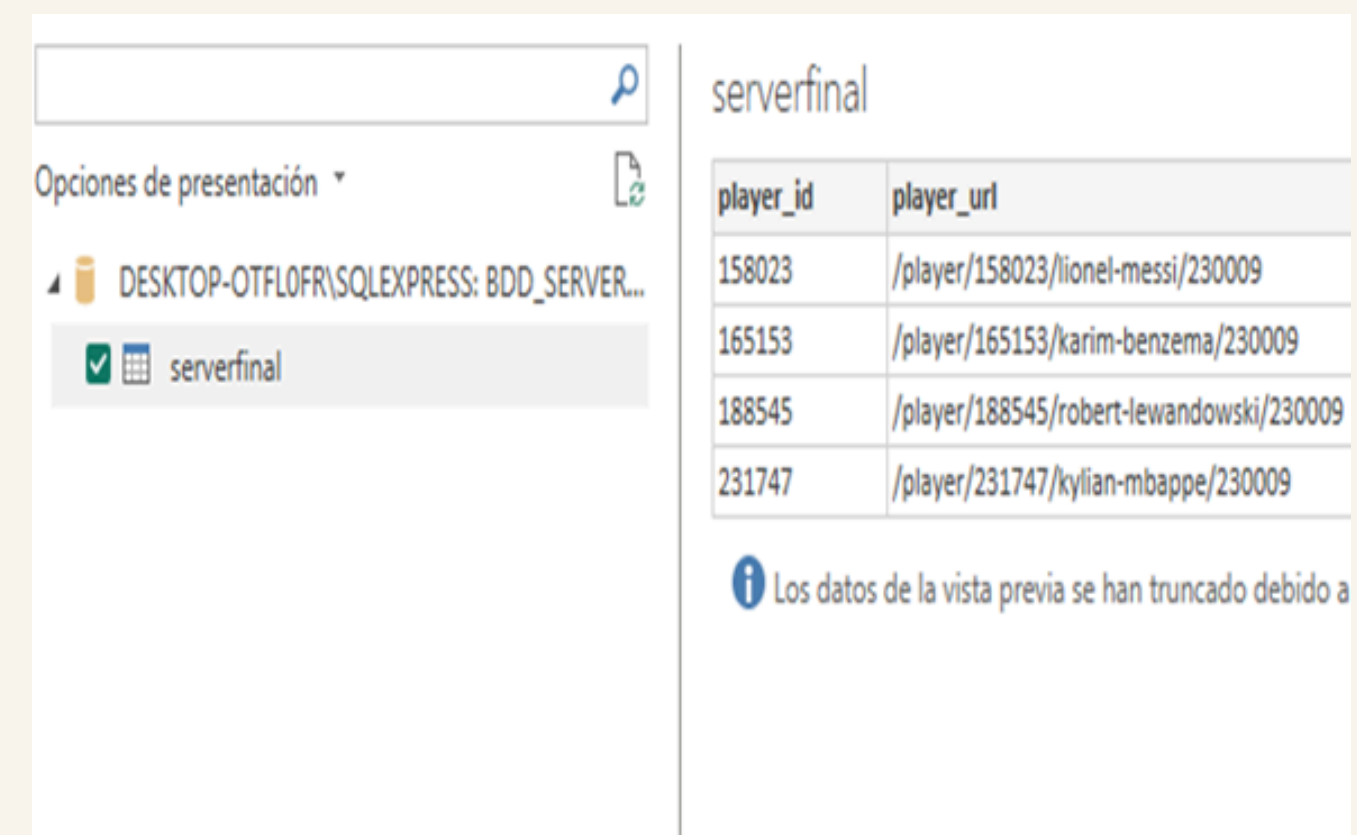
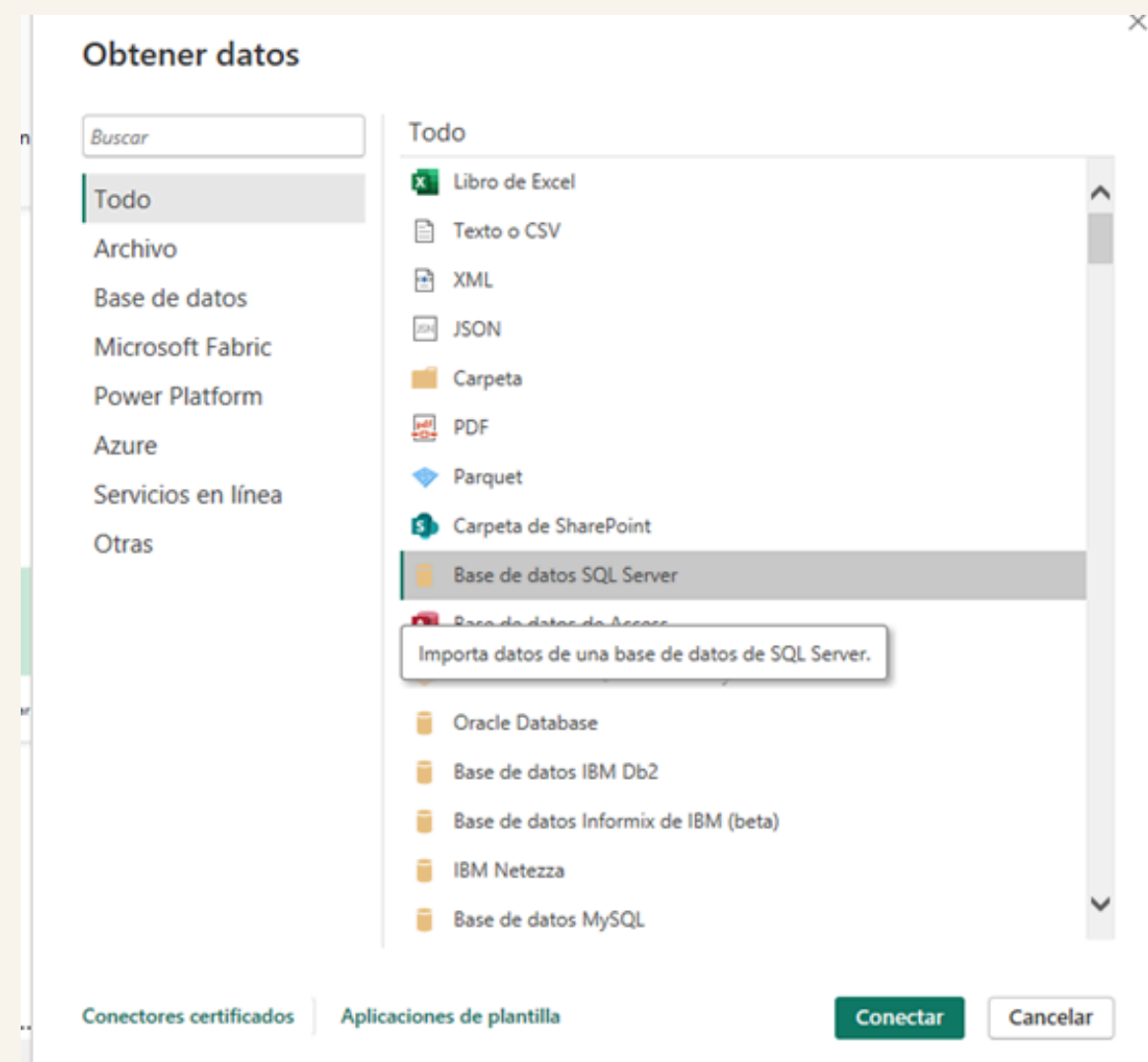
	player_id	player_url	ffa_version	ffa_update	ffa_update_date	short_name	long_name	player_positions	overall	potentia
1	158023	/player/158023/lionel-messi/230009	23	9	2023-01-13T00:00:00.000Z	L. Messi	Lionel Andrés Messi Cuccittini	RW	91	91
2	165153	/player/165153/karim-benzema/230009	23	9	2023-01-13T00:00:00.000Z	K. Benzema	Karim Benzema	CF, ST	91	91
3	188545	/player/188545/robert-lewandowski/230009	23	9	2023-01-13T00:00:00.000Z	R. Lewandowski	Robert Lewandowski	ST	91	91
4	231747	/player/231747/kyllian-mbappe/230009	23	9	2023-01-13T00:00:00.000Z	K. Mbappé	Kyllian Mbappé Lottin	ST, LW	91	95
5	192119	/player/192119/thibaut-courtois/230009	23	9	2023-01-13T00:00:00.000Z	T. Courtois	Thibaut Nicolas Marc Courtois	GK	90	91
6	192985	/player/192985/kevin-de-bruyne/230009	23	9	2023-01-13T00:00:00.000Z	K. De Bruyne	Kevin De Bruyne	CM, CAM	91	91
7	209331	/player/209331/mohamed-salah/230009	23	9	2023-01-13T00:00:00.000Z	M. Salah	Mohamed Salah Ghaly	RW	90	90
8	167495	/player/167495/manuel-neuer/230009	23	9	2023-01-13T00:00:00.000Z	M. Neuer	Manuel Peter Neuer	GK	89	89
9	190871	/player/190871/neyymar-da-silva-santos-jr/230009	23	9	2023-01-13T00:00:00.000Z	Neymar Jr	Neymar da Silva Santos Júnior	LW	89	89
10	210257	/player/210257/ederson-santana-de-moraes/230009	23	9	2023-01-13T00:00:00.000Z	Ederson	Ederson Santana de Moraes	GK	89	91
11	200145	/player/200145/carlos-henrique-venancio-casimiro...	23	9	2023-01-13T00:00:00.000Z	Casemiro	Carlos Henrique Venancio Casimiro	CDM	89	89
12	200389	/player/200389/jan-oblak/230009	23	9	2023-01-13T00:00:00.000Z	J. Oblak	Jan Oblak	GK	89	90
13	202126	/player/202126/harry-kane/230009	23	9	2023-01-13T00:00:00.000Z	H. Kane	Harry Kane	ST	89	89
14	203376	/player/203376/virgil-van-dijk/230009	23	9	2023-01-13T00:00:00.000Z	V. van Dijk	Virgil van Dijk	CB	89	89
15	208722	/player/208722/sadio-mane/230009	23	9	2023-01-13T00:00:00.000Z	S. Mané	Sadio Mané	CF, LM	89	89
16	212622	/player/212622/joshua-kimmich/230009	23	9	2023-01-13T00:00:00.000Z	J. Kimmich	Joshua Walter Kimmich	CDM, RB, CM	89	90
17	212831	/player/212831/alisson-ramses-becker/230009	23	9	2023-01-13T00:00:00.000Z	Alisson	Alisson Ramsés Becker	GK	89	90
18	239085	/player/239085/erling-haaland/230009	23	9	2023-01-13T00:00:00.000Z	E. Haaland	Erling Braut Haaland	ST	89	94
19	20801	/player/20801/cristiano-dos-santos-aveiro/230009	23	9	2023-01-13T00:00:00.000Z	Cristiano Ronal...	Cristiano Ronaldo dos Santos Av...	ST	88	88
20	177003	/player/177003/luka-modric/230009	23	9	2023-01-13T00:00:00.000Z	L. Modrić	Luka Modrić	CM	88	88
21	182521	/player/182521/toni-kroos/230009	23	9	2023-01-13T00:00:00.000Z	T. Kroos	Toni Kroos	CM	88	88





INTEGRACIÓN CON POWER BI

Finalmente, los datos se integran en Power BI para su visualización y análisis avanzado. Power BI se conecta directamente a SQL Server utilizando el nombre del servidor y la base de datos creada





INFORMÁTICA TRAFICO





OBJETIVOS

Reducir la mortalidad en carreteras mediante estrategias basadas en datos, como ajustes en límites de velocidad, adaptación de infraestructuras al clima y refuerzo de controles contra el alcohol al volante.

- Analizar la relación entre clima, velocidad y alcohol con la severidad de los accidentes.
- Identificar patrones de riesgo para proponer medidas preventivas.
- Sugerir mejoras en normativas de tránsito y controles de seguridad vial.



CASO DE ESTUDIO 1: IMPACTO DEL CLIMA EN ACCIDENTES

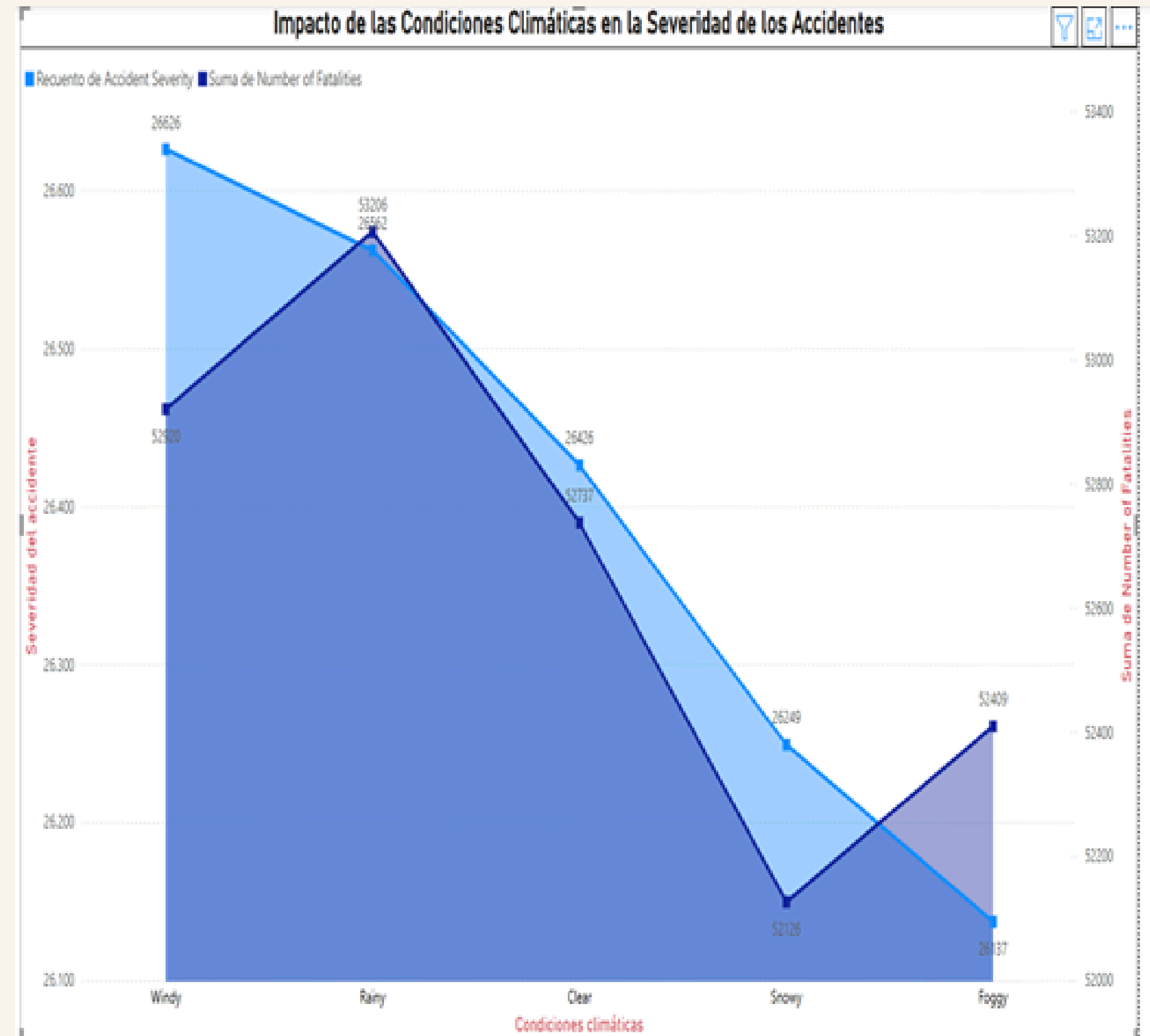
Objetivo: Analizar cómo la lluvia, nieve y niebla afectan la gravedad de los accidentes.

Descripción:

- Se estudiaron datos de informes viales y bases meteorológicas.
- Se identificaron patrones entre el clima y la severidad de los siniestros.

Conclusión:

- Lluvia y nieve aumentan la gravedad de los accidentes por menor agarre en la vía.
- La niebla limita la visibilidad, aumentando el riesgo de colisiones múltiples.



CASO DE ESTUDIO 2: VELOCIDAD Y ACCIDENTES FATALES

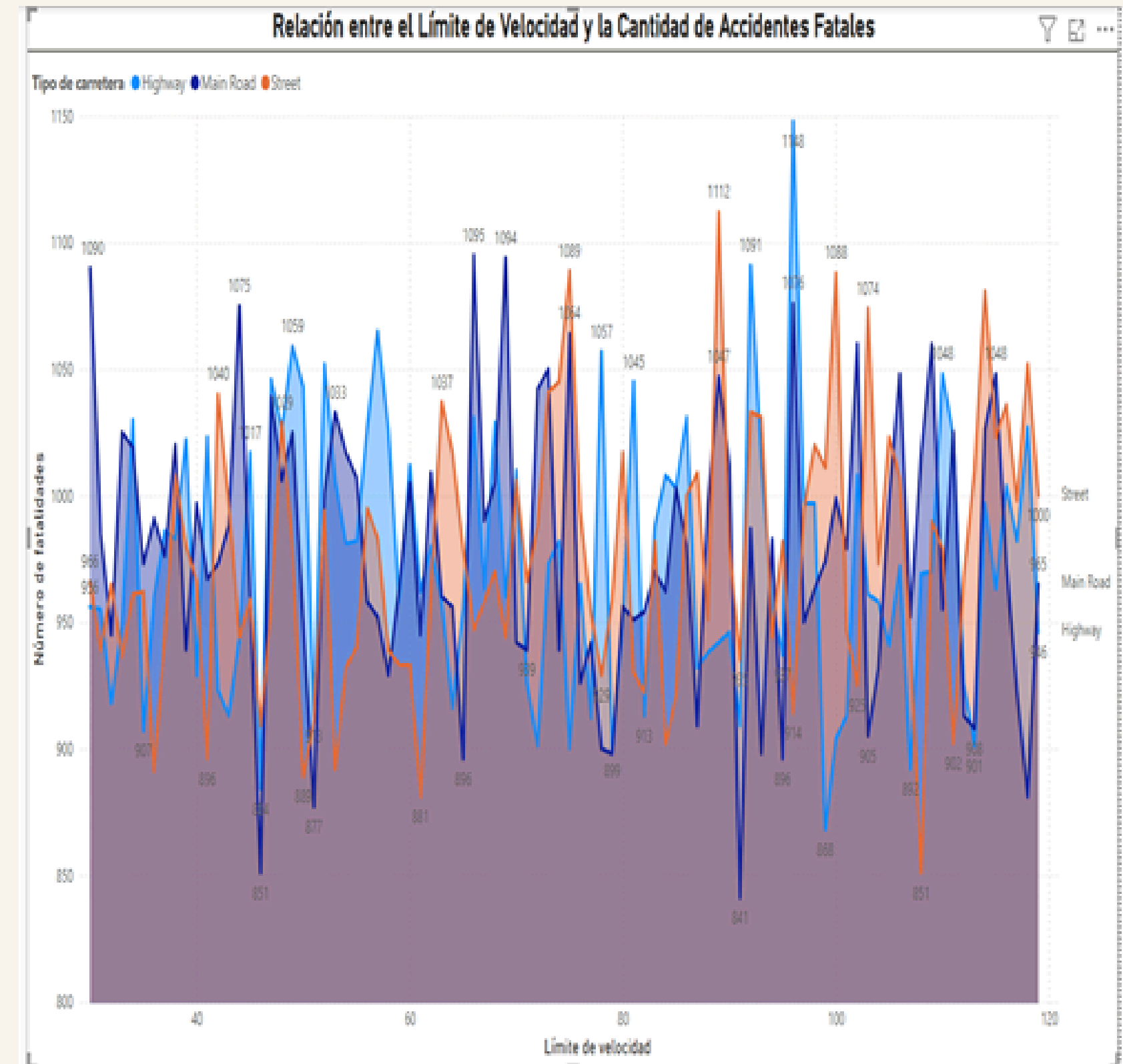
Objetivo: Evaluar la relación entre los límites de velocidad y la frecuencia de accidentes mortales.

Descripción:

- Se analizaron reportes de tráfico y siniestros.
- Se determinó si vías con mayor velocidad tienen más accidentes fatales.

Conclusión:

- Velocidades elevadas reducen el tiempo de reacción.
- Impactos a alta velocidad disminuyen la tasa de supervivencia.



CASO DE ESTUDIO 3: ALCOHOL Y SEVERIDAD DE ACCIDENTES

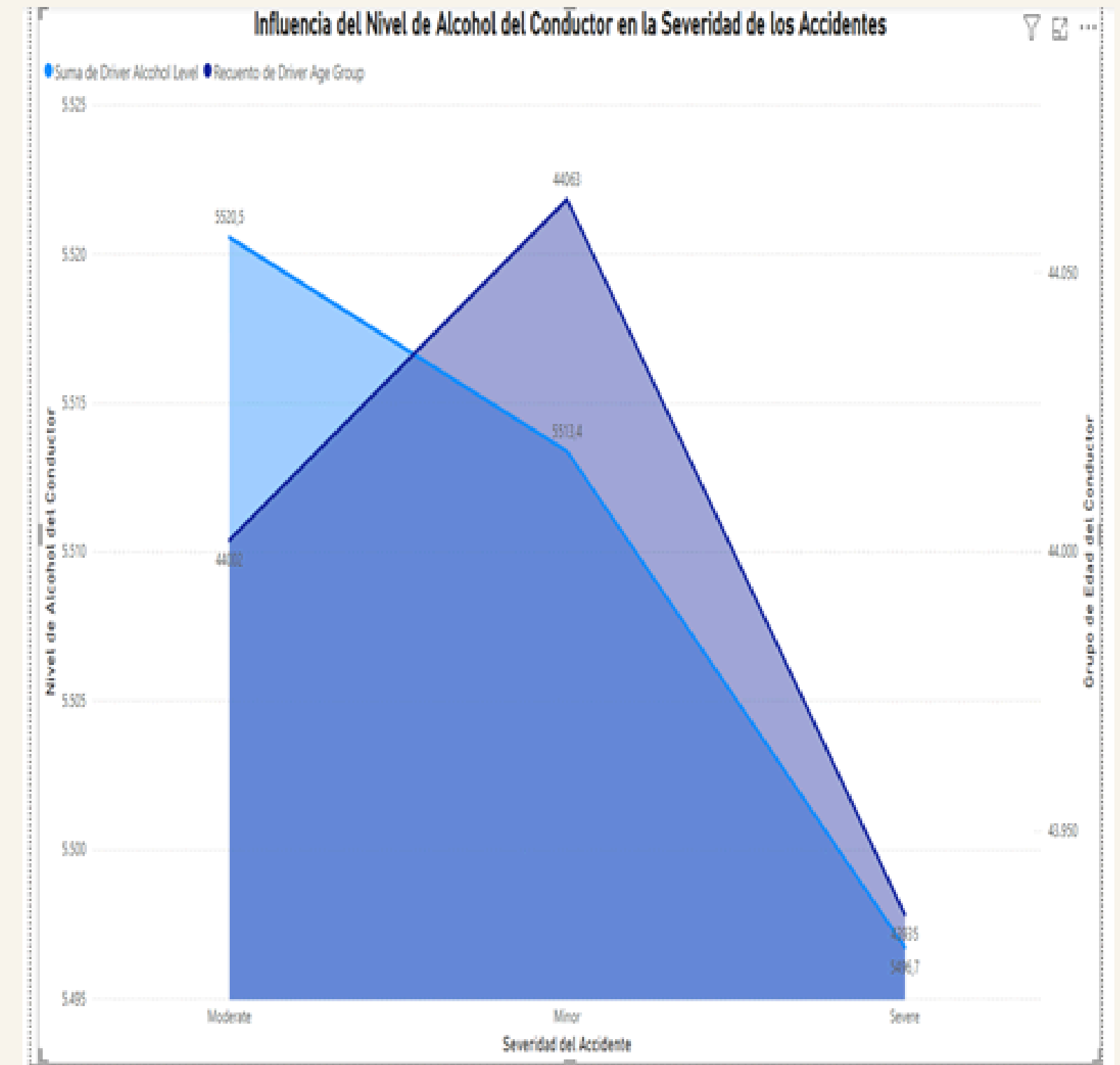
Objetivo: Analizar cómo el nivel de alcohol en sangre influye en la gravedad de los accidentes.

Descripción:

- Se estudiaron reportes de alcoholemia y siniestros viales.
- Se evaluó la relación entre BAC (contenido de alcohol en sangre) y accidentes fatales.

Conclusión:

- Un contenido de alcohol en sangre elevado reduce reflejos y percepción del riesgo.
- La probabilidad de un accidente grave o fatal aumenta con el consumo de alcohol.





ARTE PROYECTOS DE KICKSTARTER





OBJETIVOS

Analizar las tendencias de financiamiento en Kickstarter para identificar categorías exitosas, patrones geográficos y factores clave que influyen en el cumplimiento de metas, y ofrecer recomendaciones para optimizar las campañas de crowdfunding.

- Evaluar las categorías de proyectos con mayor y menor financiamiento, y los factores que afectan su éxito.
- Examinar cómo el país de origen influye en el financiamiento, identificando diferencias regionales y tendencias clave para campañas exitosas.



CASO DE ESTUDIO 1: FINANCIAMIENTO SEGUN CATEGORIA Y PAIS

Objetivo:

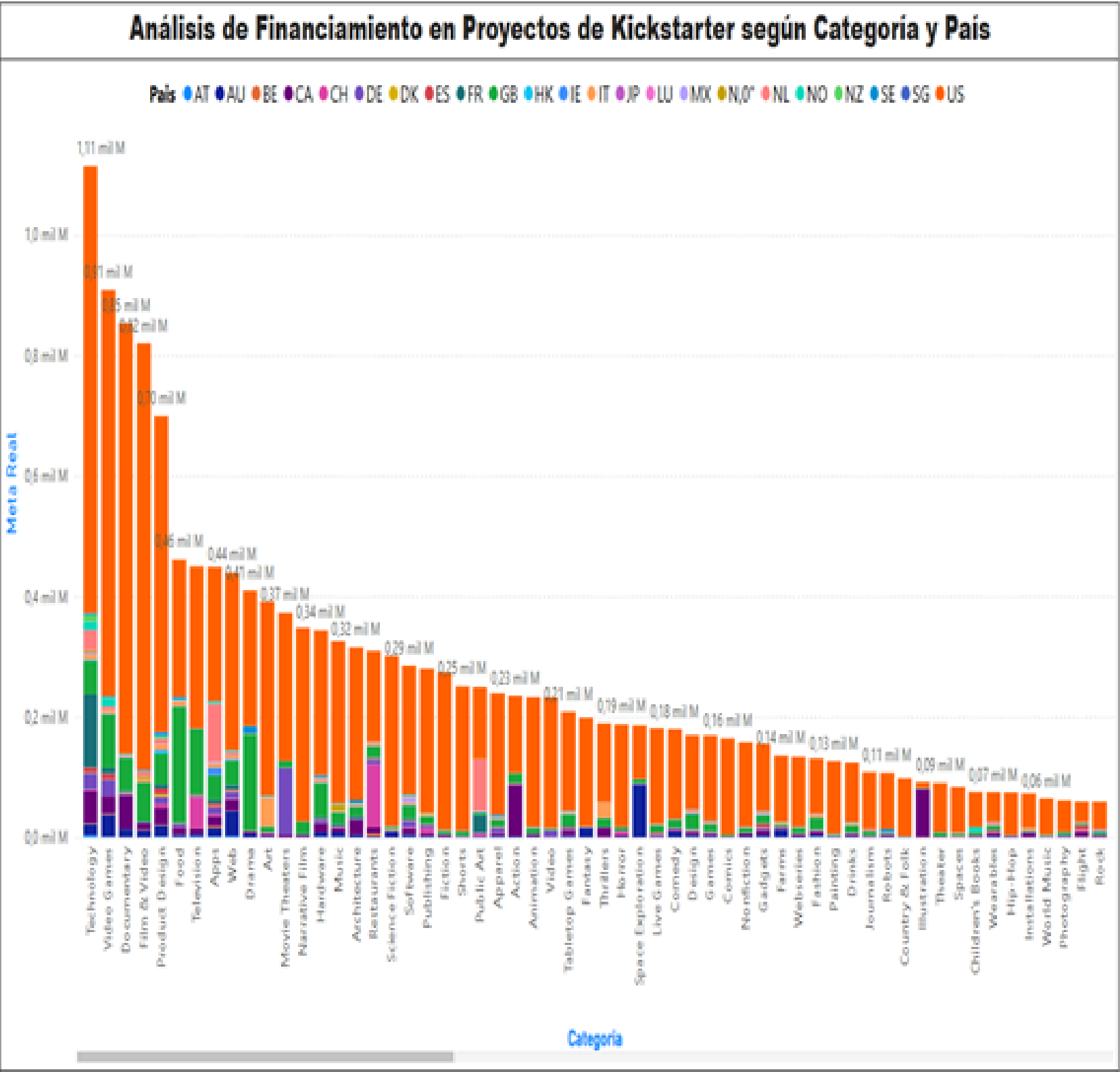
- Identificar categorías con mayor y menor financiamiento.
- Analizar la distribución de fondos por país.
- Proponer estrategias para optimizar la captación de fondos.

Descripción:

- Se analizaron datos de Kickstarter con SQLite y Power BI.
- Se identificaron tendencias de inversión y patrones de éxito en diferentes sectores y regiones.

Conclusión:

- Tecnología, videojuegos y diseño reciben más apoyo.
- Artes escénicas y periodismo tienen menor financiamiento.
- EE.UU., Reino Unido y Canadá lideran en recaudación.



CASO DE ESTUDIO 2 – IMPACTO DE LA CATEGORÍA EN EL CUMPLIMIENTO DE METAS

Objetivo:

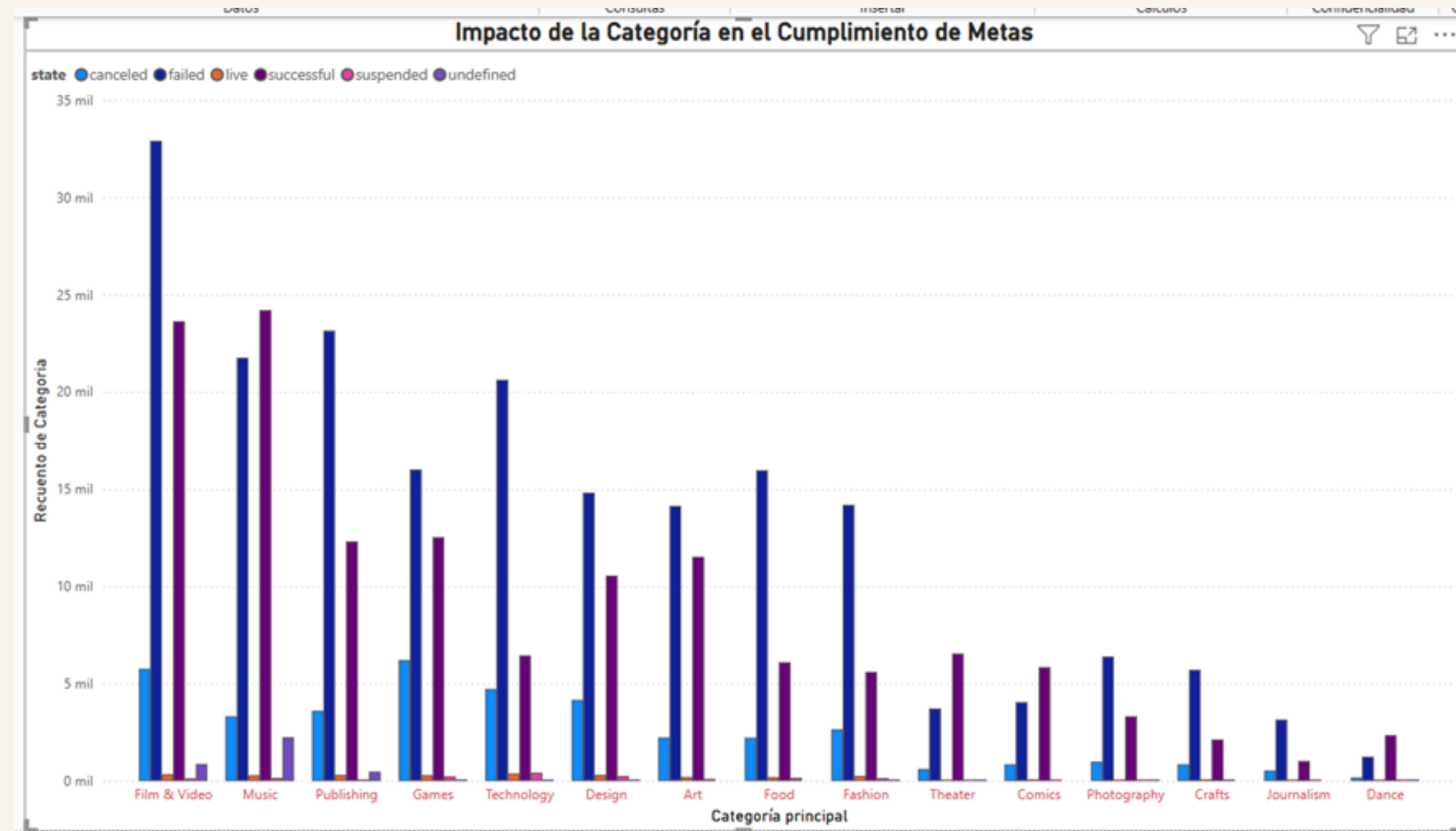
- Evaluar cómo la categoría de un proyecto afecta la probabilidad de alcanzar su meta.

Descripción:

- Se estudiaron tasas de éxito y fracaso en Kickstarter usando Power BI.
- Se analizaron categorías principales y subcategorías específicas.

Conclusión:

- Tecnología y videojuegos tienen mayor éxito por su base de seguidores.
- Arte y moda tienen menor demanda y menor probabilidad de éxito.
- Proyectos con metas muy altas y baja visibilidad suelen fracasar.



CASO DE ESTUDIO 3 – INFLUENCIA DEL PAÍS EN EL FINANCIAMIENTO

Objetivo:

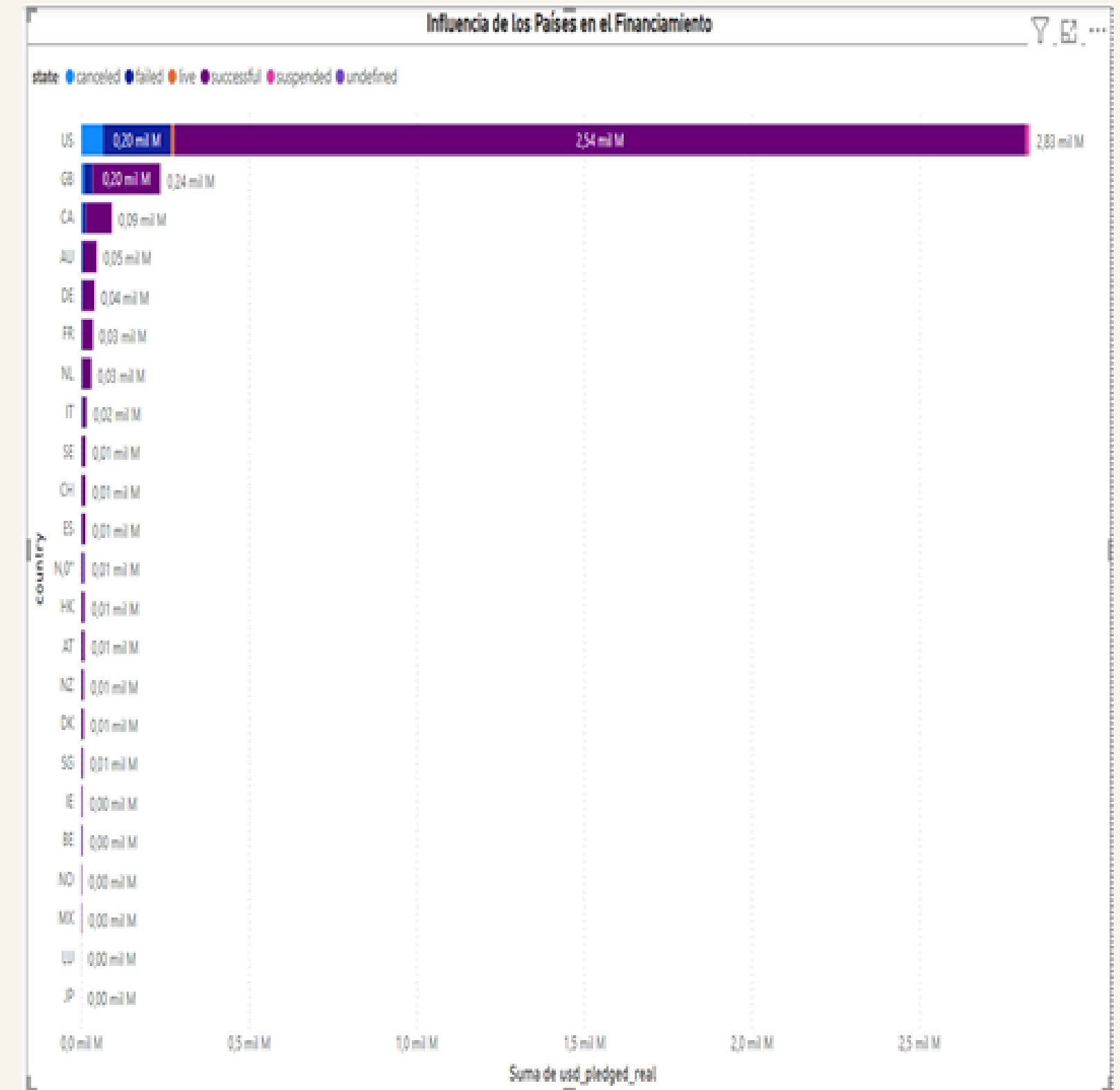
- Analizar el impacto del país de origen en el éxito financiero de un proyecto.

Descripción:

- Se evaluaron tasas de éxito, número de patrocinadores y financiamiento total según el país.
- Se identificaron tendencias regionales en el crowdfunding.

Conclusión:

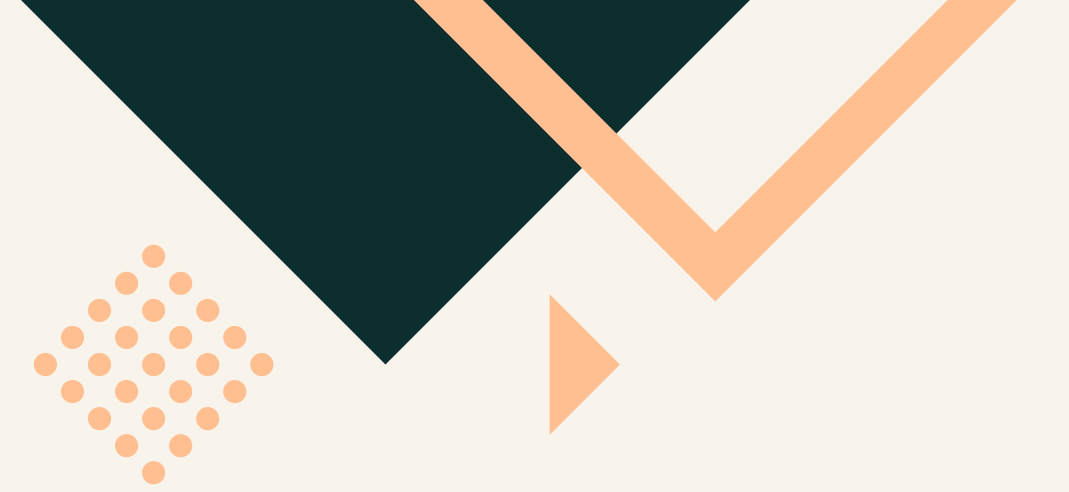
- EE.UU., Reino Unido y Canadá reciben más respaldo financiero.
- Países en desarrollo enfrentan más dificultades en crowdfunding.





NOTICIAS CYBERSEGURIDAD





OBJETIVOS

El análisis de incidentes de ciberseguridad permite identificar patrones en la actividad maliciosa, evaluar amenazas y presencia de malware, proporcionando información clave para anticipar ataques, priorizar la respuesta y mejorar la seguridad organizacional.

- Analizar los días de la semana con mayor frecuencia de ataques y los tipos predominantes para reforzar defensas en momentos críticos.
- Estudiar las alertas generadas según su severidad para priorizar la respuesta ante amenazas de alto riesgo.



CASO DE ESTUDIO 1 – TIPOS DE ATAQUES POR DÍA DE LA SEMANA

Objetivo:

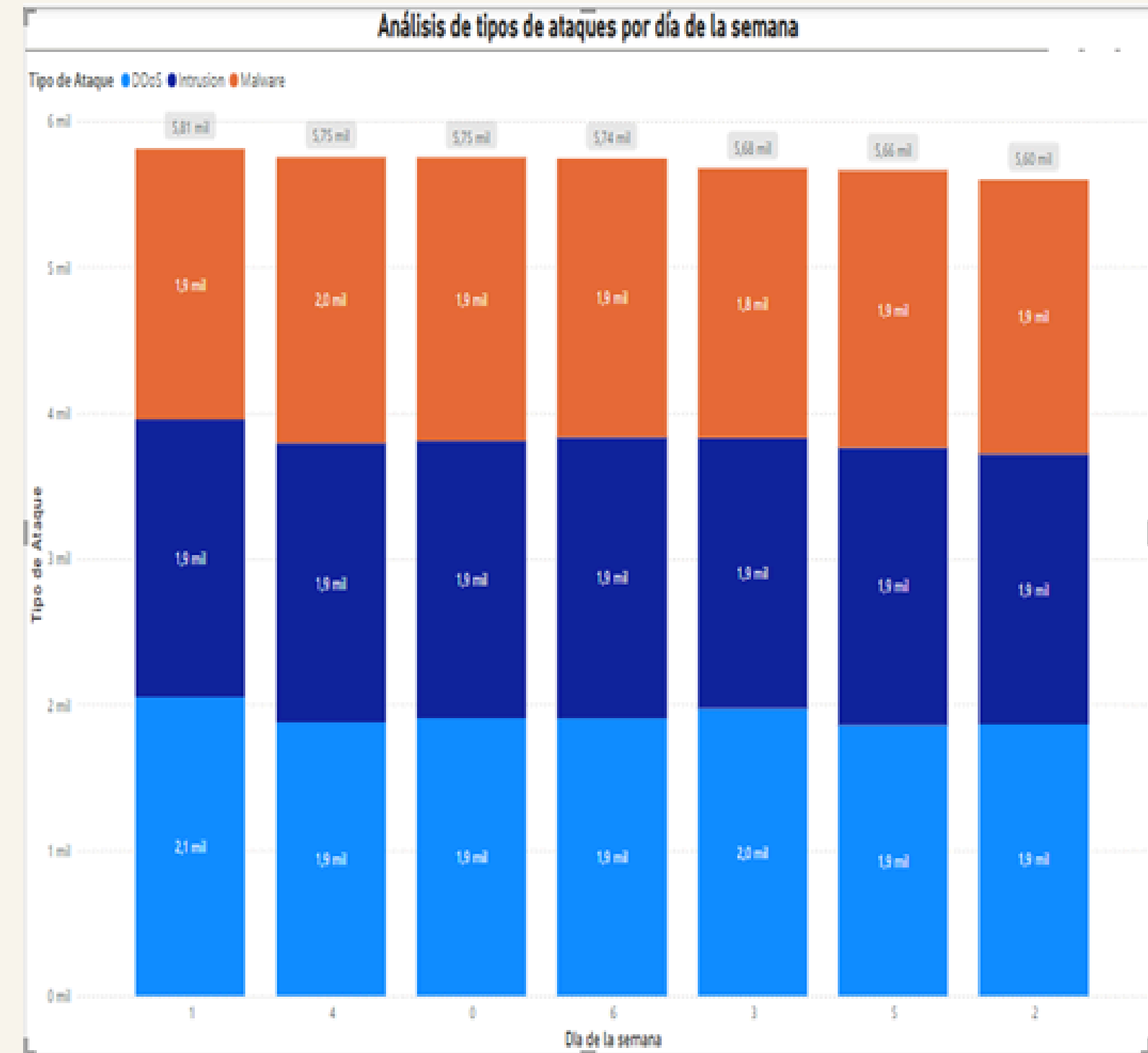
- Identificar los días con mayor actividad de ataques cibernéticos.
- Determinar los tipos de ataques más frecuentes (DDoS, intrusiones, malware).
- Optimizar la respuesta de los equipos de seguridad.

Descripción:

- Se analizaron datos de incidentes cibernéticos para detectar patrones de actividad.
- Se evaluó si los ataques ocurren más en días laborables o fines de semana.

Conclusión:

- Los ataques varían a lo largo de la semana.
- Algunos tipos de ataques ocurren más en días específicos, lo que permite mejorar estrategias de defensa.



CASO DE ESTUDIO 2 – DISTRIBUCIÓN DE ALERTAS POR NIVEL DE SEVERIDAD

Objetivo:

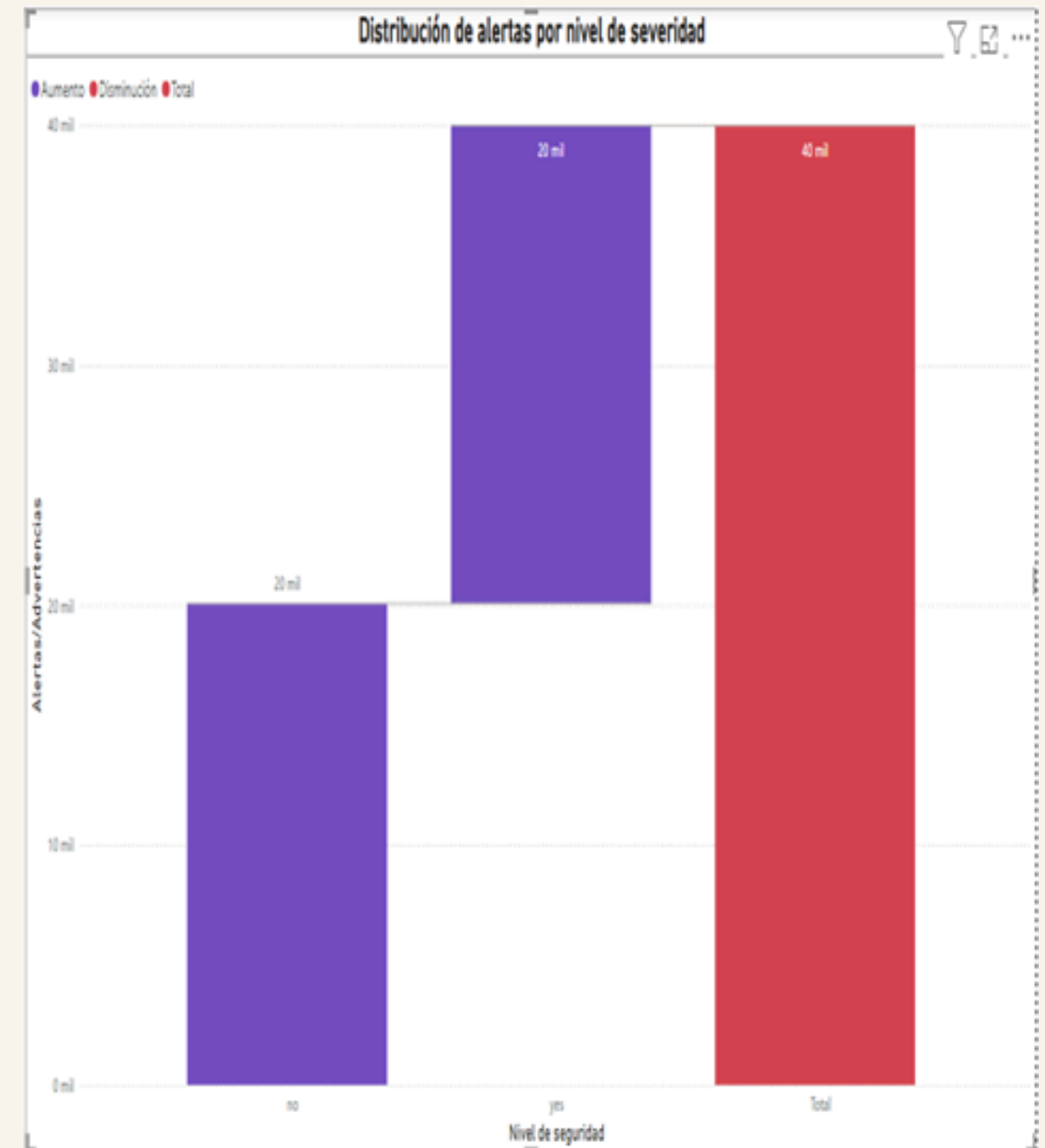
- Evaluar la cantidad y tipo de alertas de seguridad según su nivel de riesgo.
- Priorizar incidentes críticos para mejorar la respuesta ante amenazas.

Descripción:

- Se clasificaron alertas de seguridad en niveles: baja, media y alta severidad.
- Se identificaron tendencias en la cantidad de alertas generadas.

Conclusión:

- La mayoría de las alertas son de baja severidad.
- Solo un pequeño porcentaje representa amenazas graves.
- Se recomienda ajustar estrategias de seguridad para centrarse en las alertas críticas.



CASO DE ESTUDIO 3 – ANÁLISIS DE ACTIVIDAD DE MALWARE EN LA RED

Objetivo:

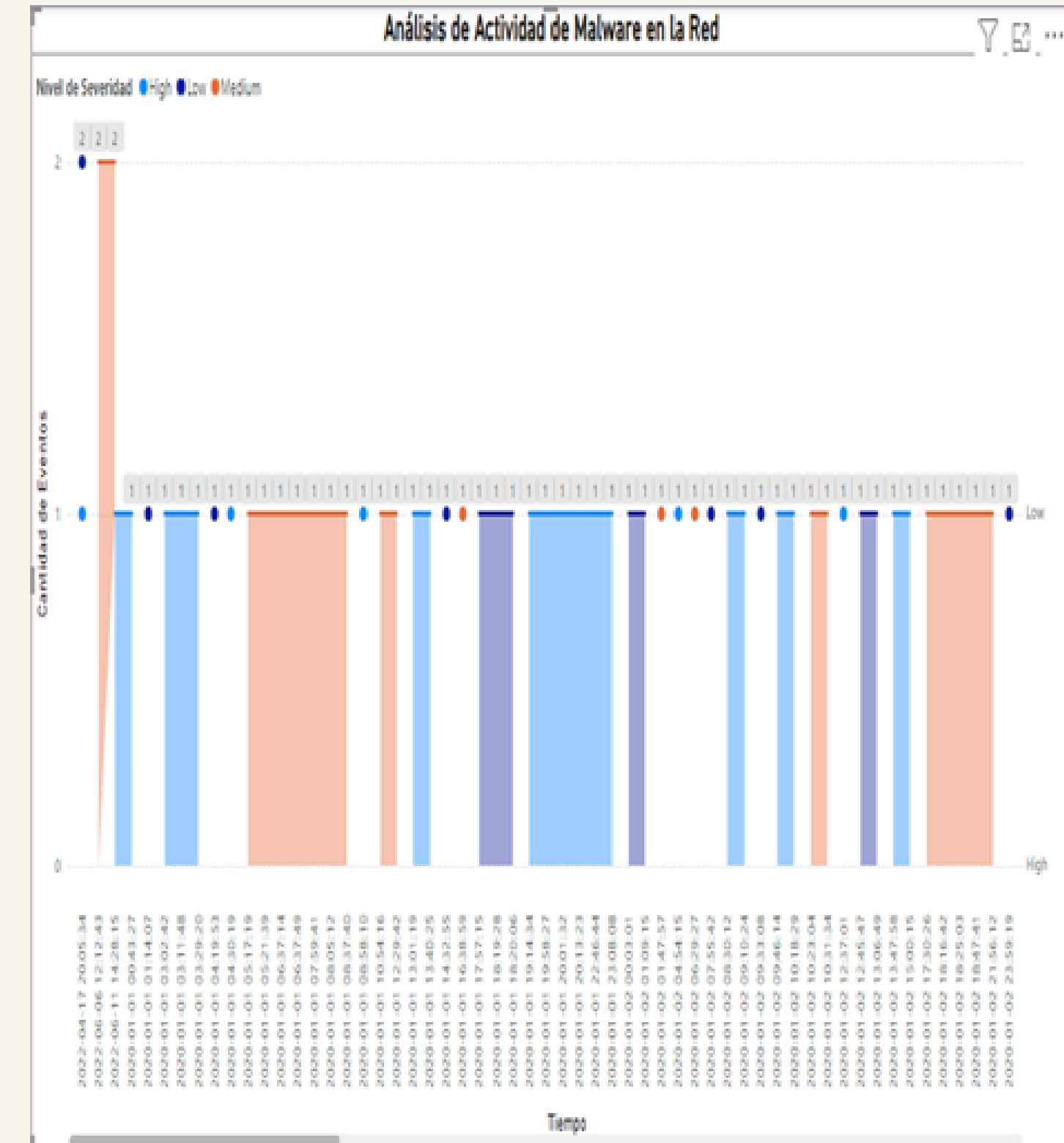
- Identificar los tipos de malware más comunes y su comportamiento a lo largo del tiempo.
- Evaluar la severidad de las detecciones para mejorar la seguridad de la red.

Descripción:

- Se analizaron eventos de malware, fuentes de alerta y niveles de riesgo.
- Se identificaron momentos de mayor actividad de amenazas.

Conclusión:

- La actividad del malware varía con picos en diferentes momentos.
- Se recomienda reforzar medidas preventivas en períodos de alta actividad.





DEPORTES FUTBOL





OBJETIVOS

Analizar las tendencias y relaciones entre la edad, el valor de mercado, el potencial y la posición de los jugadores, identificando patrones relevantes para su desarrollo y valoración en distintas etapas de su carrera.

- Evaluar la relación entre la edad y el potencial promedio de los jugadores, identificando las edades de su máximo desarrollo.
- Analizar cómo la edad y nacionalidad influyen en la valoración promedio y cómo ciertas posiciones son mejor valoradas según el país de origen.



CASO DE ESTUDIO 1 – POTENCIAL DE JUGADORES SEGUN LA EDAD

Objetivo:

- Evaluar la relación entre la edad de los jugadores y su potencial en el campo de juego.
- Identificar en qué etapa los futbolistas alcanzan su máximo desarrollo.

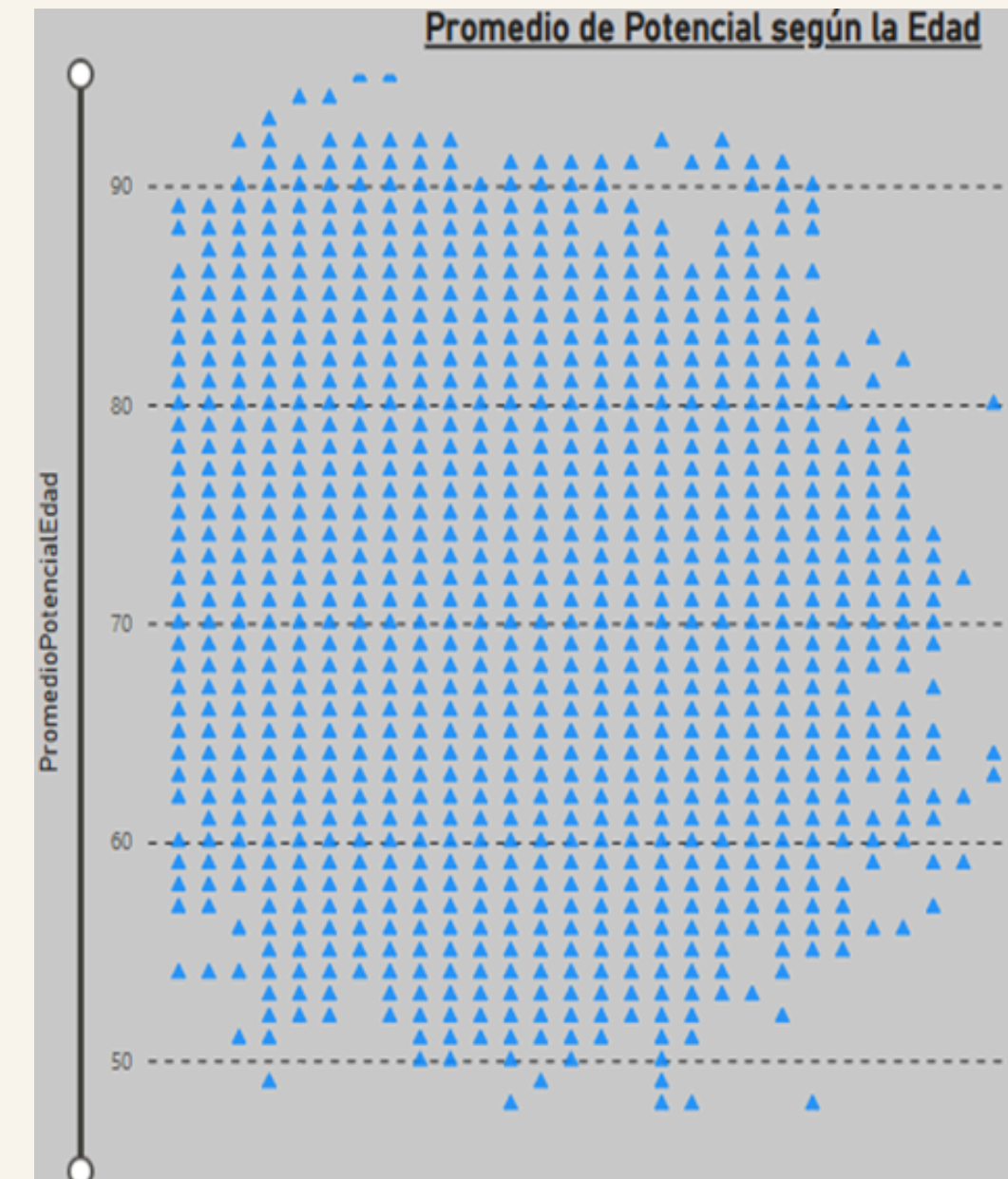
Descripción:

- Se analizó una base de datos con información sobre la edad y el potencial máximo de los jugadores.
- El potencial representa la calificación máxima que un jugador puede alcanzar a lo largo de su carrera.

Conclusión:

- El potencial máximo se alcanza entre los 20 y 25 años.
- A partir de los 25 años, se observa una disminución progresiva del potencial.

```
1 PromedioPotencialEdad = AVERAGEX(VVALUES(serverfinal[age]), CALCULATE(AVERAGE(serverfinal[potential]))))
```



CASO DE ESTUDIO 2 - VALOR DE MERCADO SEGUN LA EDAD

Objetivo:

- Analizar cómo la edad influye en la valoración económica de los jugadores.
- Identificar en qué etapas los futbolistas alcanzan su mayor cotización.

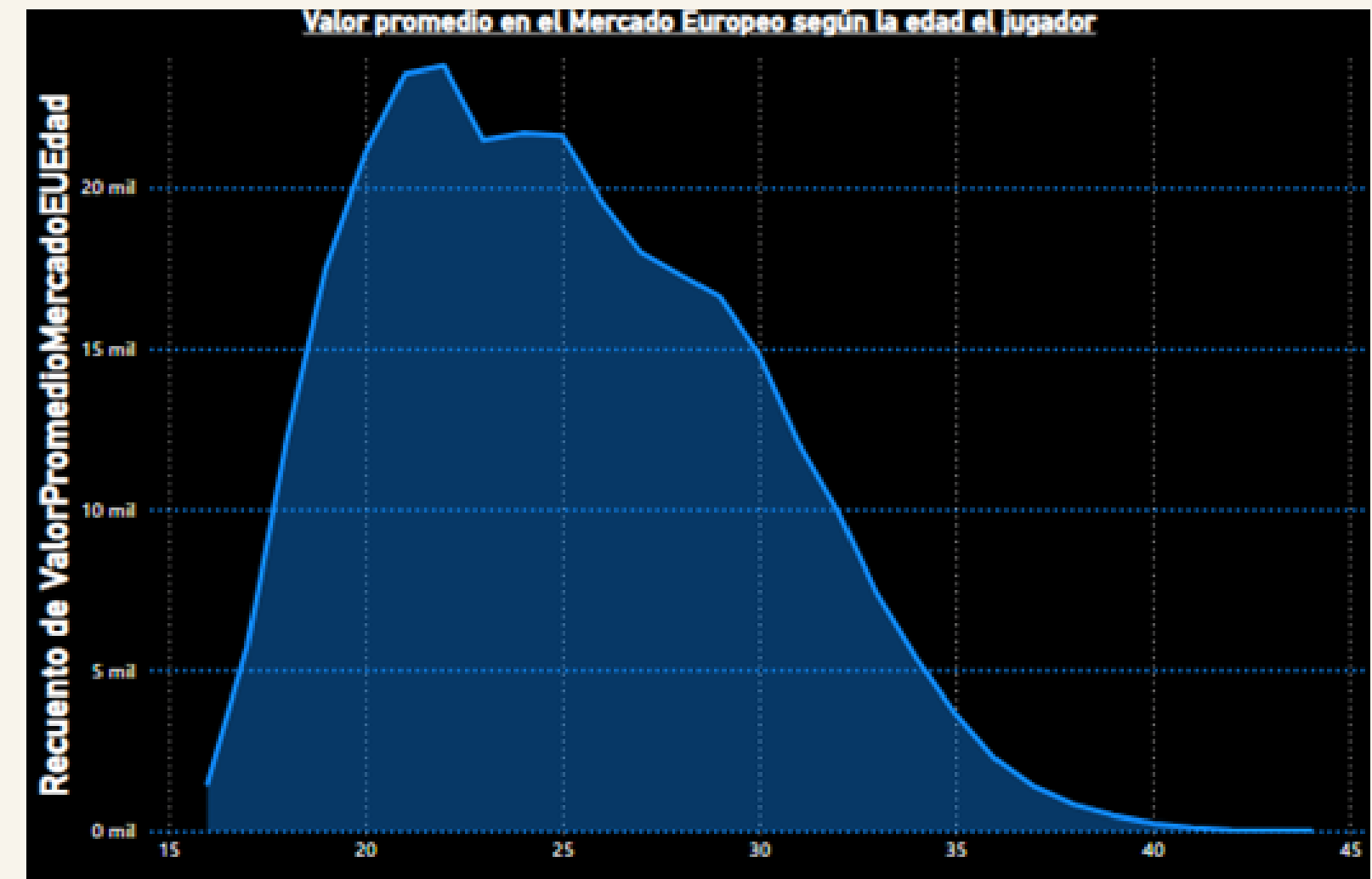
Descripción:

- Se estudió la relación entre la edad de los jugadores y su valor de mercado en el fútbol europeo.
- El valor de mercado se define en función del rendimiento, la popularidad y la demanda en transferencias.

Conclusión:

- El valor de mercado aumenta rápidamente entre los 15 y 17 años.
- Alcanza su punto máximo a los 25 años y disminuye gradualmente con la edad.

```
ValorPromedioMercadoEUEdad =  
CALCULATE(  
    AVERAGE((serverfinal[value_eur])),  
    ALLEXCEPT(serverfinal, serverfinal[age])  
)
```



CASO DE ESTUDIO 3 - VALORACIÓN SEGUN POSICIÓN Y NACIONALIDAD

Objetivo:

- Determinar cómo la posición y la nacionalidad influyen en la valoración de los jugadores.
- Identificar qué países producen jugadores con mayor cotización.

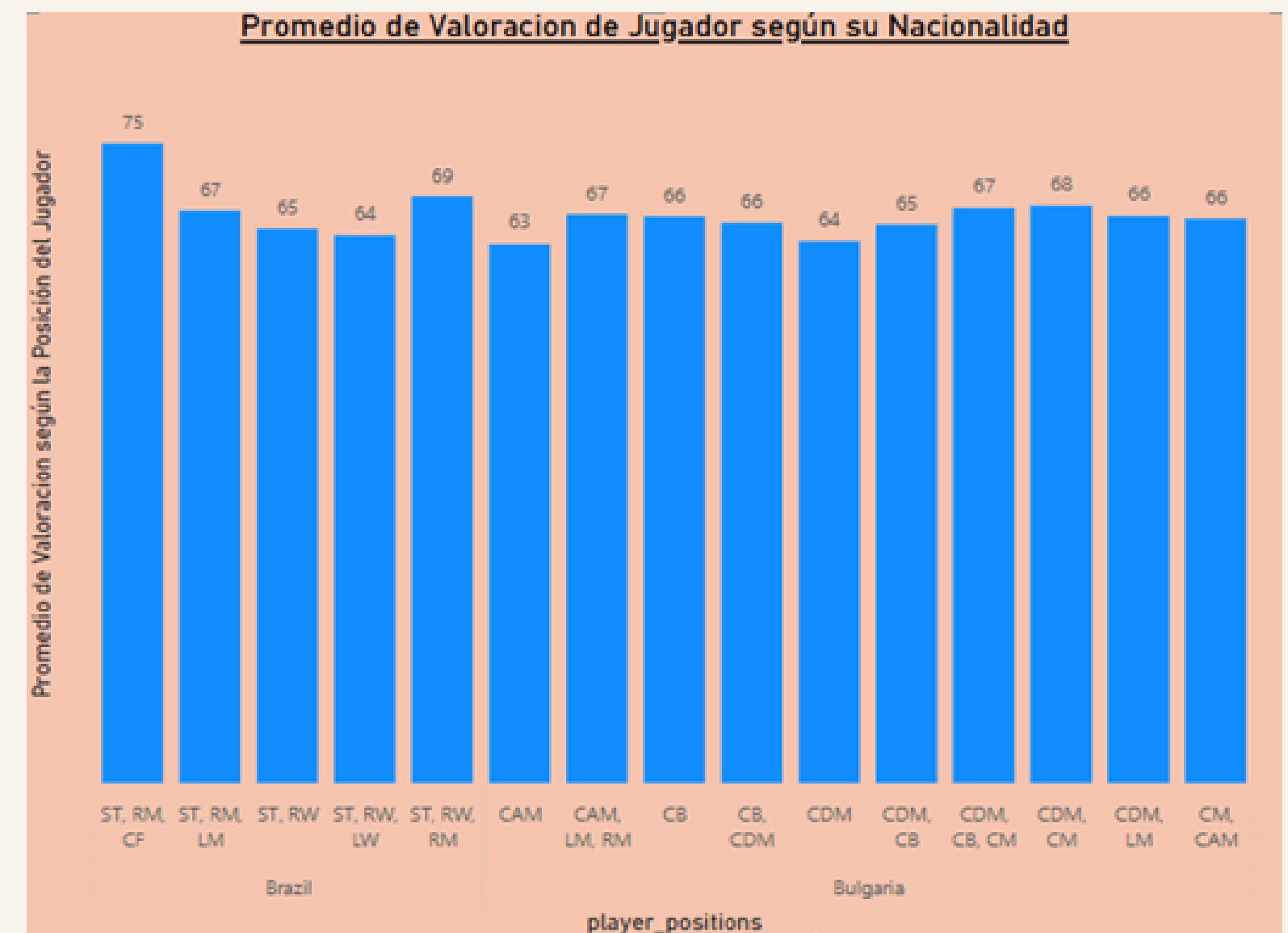
Descripción:

- Se estudió la relación entre el país de origen y la valoración de los jugadores en el mercado.
- Se analizaron factores como rendimiento individual, posición en el campo y éxito en selecciones nacionales.

Conclusión:

- Brasil lidera en posiciones ofensivas (ST, RW, LW).
- La calidad de las ligas y el éxito internacional influyen en la valoración de los jugadores.

```
Promedio_Valoracion_Posicion =  
CALCULATE(  
    AVERAGE((serverfinal[overall])),  
    ALLEXCEPT(serverfinal, serverfinal[player_positions])  
)
```





NOTICAS INCENDIOS





OBJETIVOS

El análisis de los incendios forestales y su impacto en las estructuras permite identificar patrones de daño, evaluar la distribución geográfica de los incendios y la carga de trabajo de las unidades de respuesta, facilitando decisiones informadas para mitigar futuros incendios.

- Analizar el daño causado en diferentes tipos de edificaciones, como residencias unifamiliares, casas móviles y estructuras de servicios públicos.
- Estudiar cómo varía la intensidad y frecuencia de los incendios según la ubicación, identificando las áreas más afectadas y evaluando la carga de trabajo de CAL FIRE.



CASO DE ESTUDIO 1 - ANÁLISIS DEL DANO ESTRUCTURAL POR INCIDENTES DE INCENDIO

Objetivo

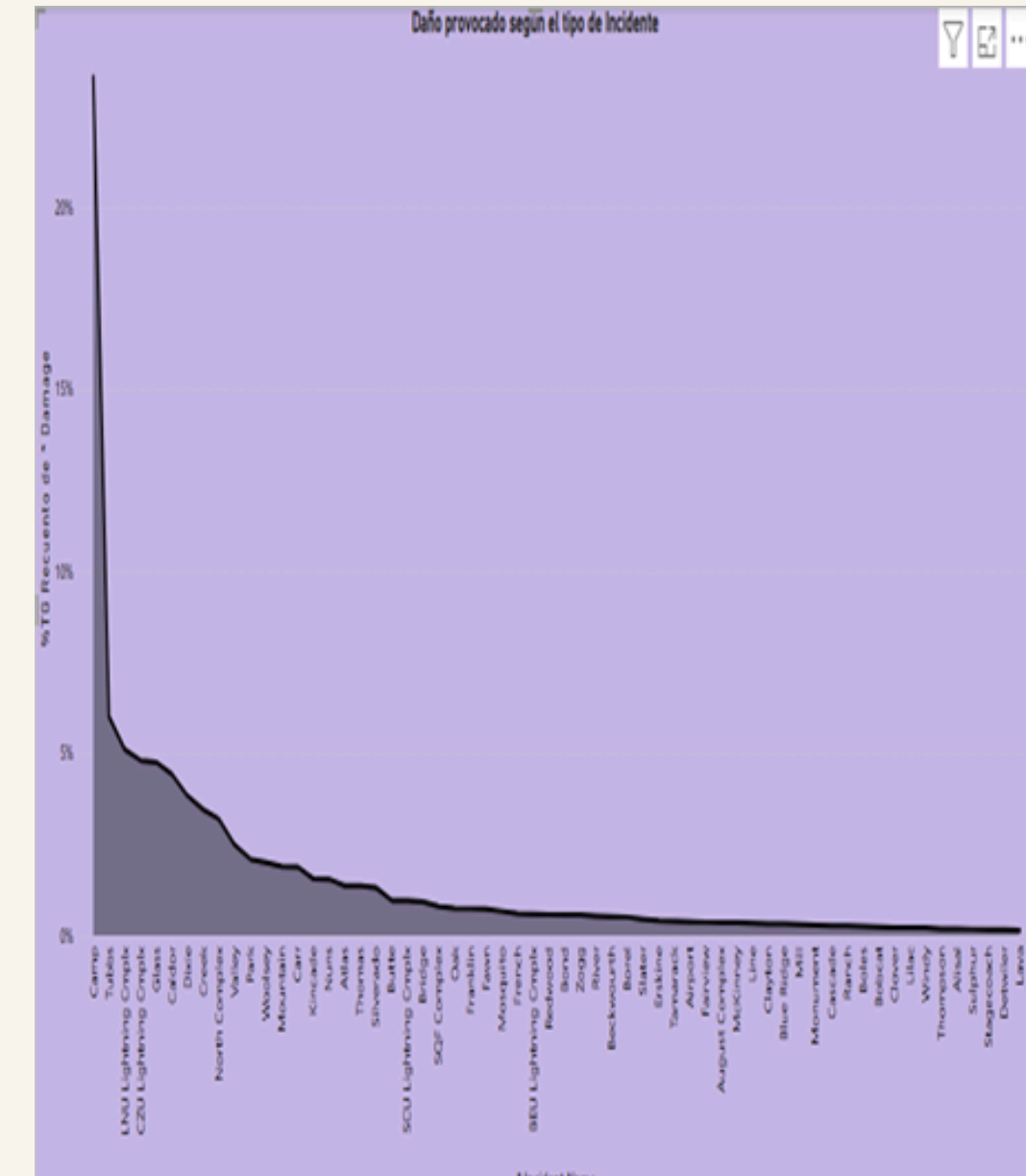
- Analizar el impacto de los incendios en estructuras afectadas según el tipo de incidente.

Descripción

- Agrupación de incendios por nombre del incidente.
- Comparación del impacto en diferentes tipos de estructuras (residencias, casas móviles, estructuras utilitarias).

◆ Conclusión

- Se observa una disparidad en el daño causado por distintos incendios.
- Los incendios forestales han provocado un daño estructural significativo.



CASO DE ESTUDIO 2 - DISTRIBUCIÓN GEOGRÁFICA DE LOS INCENDIOS Y SU IMPACTO

Objetivo

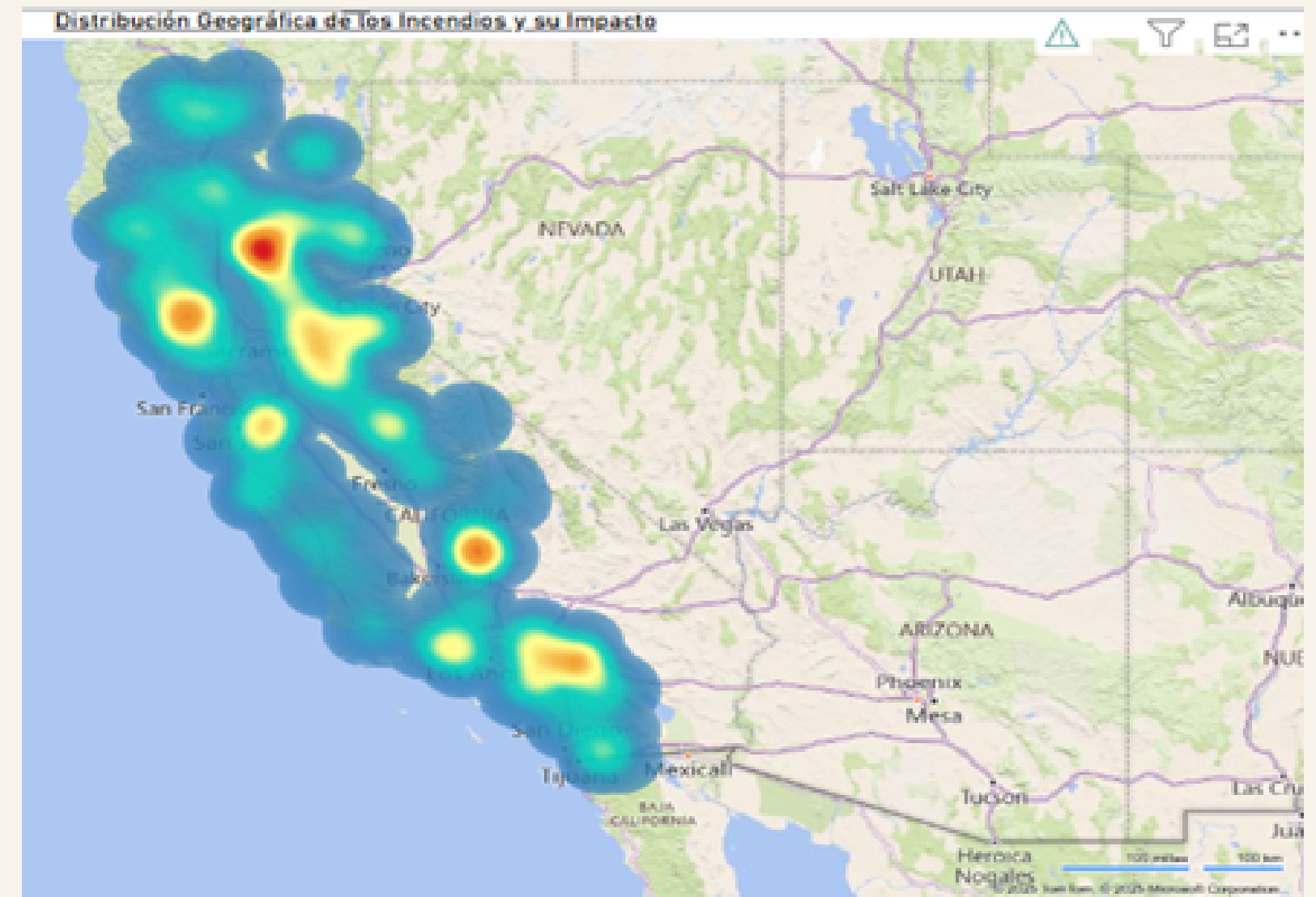
- Analizar la distribución geográfica de los incendios en ciudades y condados.
- Identificar áreas más afectadas y la magnitud del daño estructural.

Descripción

- Estudio basado en datos de localización (ciudad, condado, estado).
- Evaluación del impacto según la magnitud del daño estructural.

Conclusión

- El mapa de calor indica mayor concentración de incendios en California.
- Regiones como Sacramento y Los Ángeles son especialmente propensas a incendios.



CASO DE ESTUDIO 2 - ANÁLISIS DE LA FRECUENCIA DE INCENDIOS POR UNIDAD DE RESPUESTA (CAL FIRE UNIT)

Objetivo

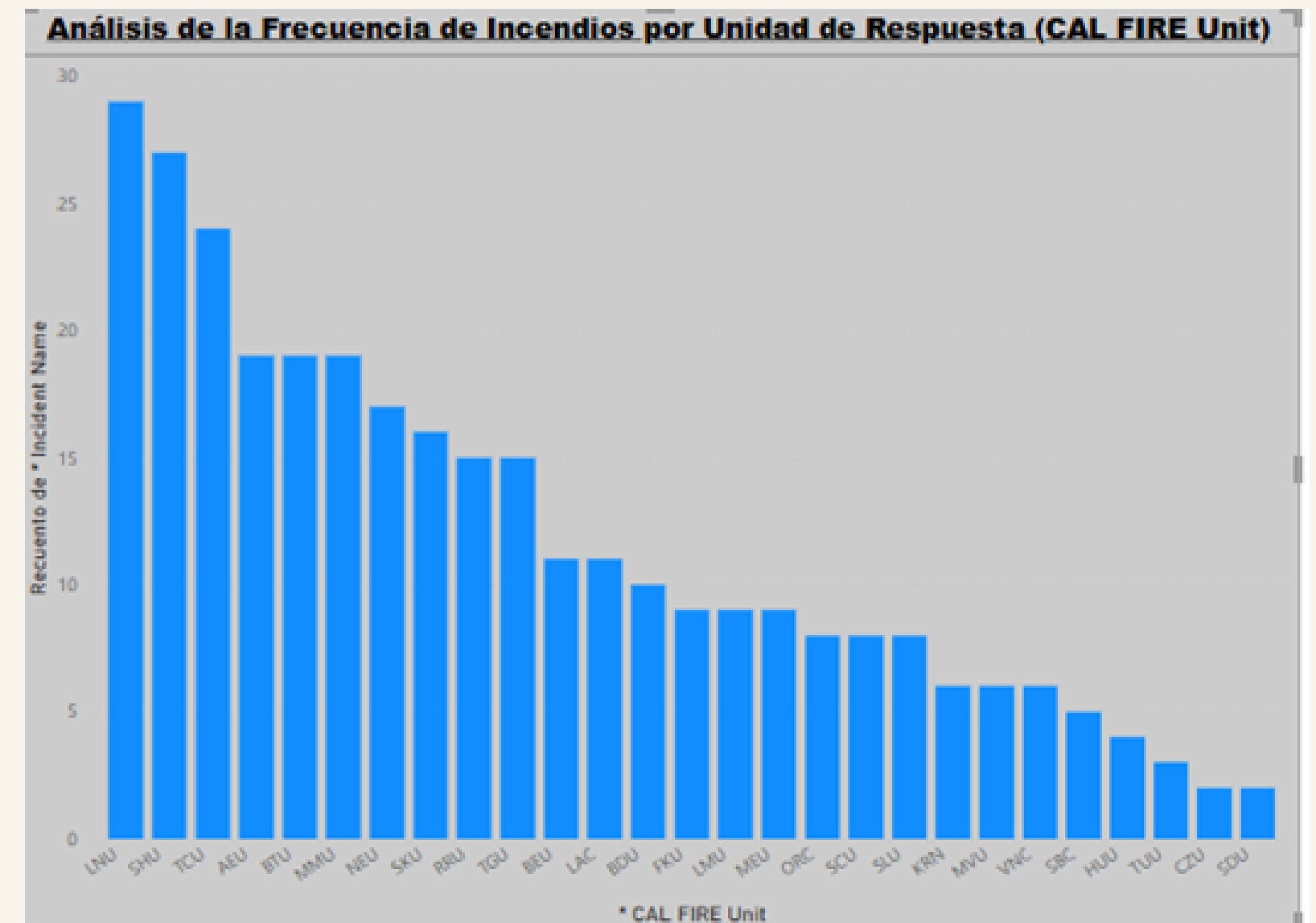
- Evaluar la frecuencia de incendios según unidades de respuesta de CAL FIRE.

Descripción

- Análisis de incendios atendidos por diferentes unidades de CAL FIRE en diversos condados.
- Comparación de la carga de trabajo entre las unidades.

Conclusión

- Unidades LNU, SHU y TCU presentan la mayor cantidad de incidentes.
- Estas unidades atienden una proporción significativa de los incendios en comparación con otras.



CONCLUSION

Identificación de Patrones y Tendencias

Se analizaron distintos factores que permiten anticipar eventos y mejorar la toma de decisiones basadas en datos.

Optimización de Recursos y Estrategias

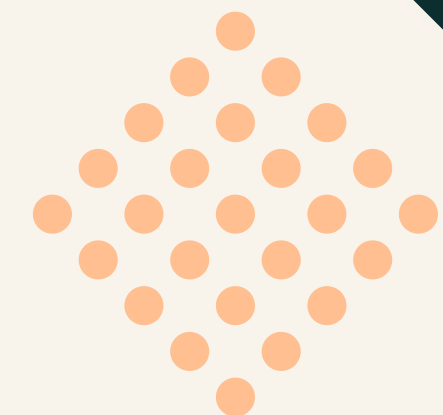
Los resultados obtenidos permiten una mejor asignación de recursos, ajustando estrategias de respuesta y fortaleciendo medidas preventivas en áreas de mayor riesgo.

Mejor Comprensión del Comportamiento de Incidentes

El análisis detallado ayudó a identificar factores recurrentes que influyen en los resultados, facilitando la toma de decisiones informadas.

Aplicación en Diversos Contextos

La metodología utilizada también permitió optimizar estrategias en otros ámbitos, como el desarrollo y entrenamiento en el rendimiento deportivo.





THANK
YOU

