ESCUELA POLITÉCNICA NACIONAL

TECNOLOGIA EN DESARROLLO DE SOFTWARE

INFORME FINAL DE PROYECTO DE

ANÁLISIS DE DATOS

Autores:

María Paula Girón Cedeño

Eduardo Fabricio Ganchala Castillo

Kevin Alexis Simbaña Sanguña

Profesor

Ing. Lorena Chulde

Quito -Ecuador

05 agosto de 2025

Informe Proyecto de Análisis

María Girón, Eduardo Ganchala, Kevin Simbaña

(*Escuela Politécnica Nacional, Quito, Ecuador)*

maria.giron@epn.edu.ec

eduardo.ganchala@epn.edu.ec

kevin.simbana06@epn.edu.ec

Resumen – Este informe presenta un análisis exploratorio de datos aplicados a un conjunto de datos relacionados con Eventos de futbol a nivel mundial, Actividades y Hobbies, Noticias Mundiales Recorrido de Restaurantes y Factores de colaboración de estudiantes. Para el desarrollo del proyecto se utilizaron herramientas como gestores de bases de datos relacionales y no relacionales, el lenguaje de programación Python y Power BI. Se aplico la metodología estándar de análisis de datos. El objetivo principal fue identificar patrones y elementos clave permitan extraer conclusiones significativas y facilitar la toma de decisiones basada en datos.

I. INTRODUCCIÓN

Los datos se han convertido en uno de los recursos más valiosos para comprender fenómenos sociales, económicos y culturales a escala global. El análisis de datos permite extraer información relevante de grandes volúmenes de datos y transformarla en conocimiento útil para la toma de decisiones. En este contexto, se desarrolló un proyecto orientado al análisis exploratorio de datos provenientes de diversas categorías temáticas: *Eventos de fútbol a nivel mundial, Actividades y Hobbies, Noticias Mundiales, Recorrido de Restaurante y Factores de colaboración de estudiantes.*

El propósito principal de este proyecto es aplicar los conocimientos adquiridos durante el proceso de formación académica en análisis de datos, abarcando etapas como la limpieza, estructuración, migración, visualización e interpretación de la información.

Este informe detalla cada etapa del proceso, describe las técnicas empleadas y presenta los principales hallazgos obtenidos a partir del análisis realizado. La experiencia permitió fortalecer habilidades analíticas y técnicas, contribuyendo a la formación integral del estudiante en el ámbito del análisis de datos.

II. DESARROLLO

Caso de Estudio

(Llenar por dataframe seleccionado)

Objetivo General

Aplicar de manera práctica los conocimientos de extracción, limpieza, transformación, análisis y visualización de datos en un entorno de Big Data heterogéneo, integrando múltiples fuentes de información y bases de datos, para generar conclusiones accionables en las cinco temáticas escogidas.

Objetivo Especifico

1. Extraer datos de al menos 10 fuentes distintas (Kaggle, CSV, JSON, Internet
2. Depurar y normalizar los datasets, garantizando la calidad de la información.
3. Diseñar una arquitectura que combine SQL y MongoDB Atlas, asegurando interoperabilidad.
4. Implementar análisis de sentimientos y modelos descriptivos para cada temática.
5. Construir 15 casos de estudio y dashboards en Power BI que resuman los hallazgos.
6. Documentar todo el proceso en un repositorio GitHub siguiendo buenas prácticas.

Equipo de Trabajo

El proyecto fue desarrollado por un equipo de trabajo conformado por tres integrantes: **María Girón**, **Eduardo Ganchala** y **Kevin Simbaña**, quienes participaron activamente en todas las etapas del proceso de análisis de datos. Cada miembro aportó de manera equitativa en las tareas clave del proyecto, incluyendo la extracción, limpieza, migración y visualización de la información.

La responsabilidad técnica fue asumida de forma conjunta, permitiendo un trabajo colaborativo en el procesamiento de los datos, la construcción de las visualizaciones y la formulación de las conclusiones finales presentadas en este informe.

Cronograma de Actividades

Para la planificación y gestión del proyecto, se utilizó el diagrama de Gantt como herramienta clave. Esta herramienta permitió organizar las actividades de forma estructurada, asignando responsabilidades específicas a cada integrante del

equipo y distribuyendo las tareas a lo largo de un periodo de cuatro semanas.

Gracias a esta planificación visual, se logró una visión clara del avance del proyecto, los tiempos estimados para cada etapa y la carga de trabajo por integrante. Esto facilitó la coordinación eficiente del equipo y permitió el cumplimiento ordenado de cada fase del sistema, desde la extracción y limpieza de datos hasta la visualización y presentación final.

La planificación detallada se encuentra representada en el Anexo A, donde se presenta el diagrama de Gantt completo aplicado al desarrollo del proyecto.

Recursos y Herramientas
Durante el desarrollo del proyecto, se emplearon diversas herramientas tecnológicas y analíticas que facilitaron cada una de las etapas del análisis de datos. A continuación, se detallan las principales:

- *Lenguajes:*

Python: Es un lenguaje de programación informático que se utiliza para crear software, automatizar tareas y realizar análisis de datos. Se puede utilizar para crear una variedad de programas diferentes y no está especializado en ningún problema específico.

SQL: Es un lenguaje de programación que sirve para poder trabajar con sistemas de gestión de bases de datos relacionales, permitiendo realizar consultas para mostrar información que necesita el usuario.

- *Bases de datos:*

SQLite: SQLite es un sistema de gestión de bases de datos relacionales ligero y de código abierto, integrado en muchas aplicaciones. A diferencia de otros sistemas, SQLite no requiere un servidor separado, sino que se implementa como una biblioteca que se enlaza directamente a la aplicación

SQL Server: Base de datos relacional propietaria de Microsoft, basada en transact-SQL, orientada **a** entornos empresariales y con múltiples ediciones para la nube e instalación local.

MongoDB: Es una base de datos no relacional de documentos (BSON) o tipo colecciones con esquemas flexibles, diseñada para facilitar el desarrollo ágil**.**

Redis: Redis es un almacén de datos de código abierto en memoria, conocido por su alta velocidad y flexibilidad. Se utiliza principalmente como base de datos, caché y mensajería.

- *Librerías:*

Pandas: Biblioteca de Python para análisis y manipulación de datos tabulares. Ofrece estructuras (Series y DataFrame) y operaciones eficientes para limpiar, transformar y combinar datos.

NumPy: La base del cómputo científico en Python. Proporciona el tipo ndarray (arrays multidimensionales de alto rendimiento) y funciones para álgebra lineal, transformadas, estadísticas, generación de números aleatorios, y más. Usada por miles de bibliotecas.

TextBlob: Librería sencilla de Python para procesamiento de lenguaje natural (NLP). Permite análisis de sentimiento, etiquetado gramatical, extracción de frases nominales, corrección ortográfica, traducción y más, con una interfaz muy amigable. Ideal para prototipos rápidos**.**

- *Visualización*

Power BI: Herramienta de inteligencia empresarial (Business Intelligence – BI) de Microsoft para **conectar, transformar, modelar y visualizar datos**, ya sea en un entorno local o en la nube, de forma interactiva y colaborativa. Está diseñada para usuarios de negocio (analistas, ejecutivos) con poco requerimiento de código, pero también ofrece funcionalidades avanzadas para modelado de datos y escenarios corporativos complejos.

- *Otras Herramientas:*

GitHub: Plataforma colaborativa basada en Git para alojar código, automatizar flujos de trabajo y coordinar proyectos de desarrollo. Ideal para equipos que comparten, revisión de proyectos con control, transparencia y CI/CD integrado.

Kaggle: Comunidad y plataforma de ciencia de datos de Google donde puedes explorar datasets, crear notebooks colaborativos y participar en competencias de machine learning para resolver problemas reales mientras aprendes y compites.

Arquitectura de Solución
El proyecto se desarrolló en equipo, aplicando una arquitectura de solución basada en el enfoque ETL (Extracción, Transformación y Carga), complementada con análisis y visualización de datos. Este enfoque permitió estructurar el trabajo de manera eficiente y distribuir las tareas de forma equitativa entre los integrantes. A continuación, se describen las principales fases del proceso:

1. *Extracción de Datos*

Para la extracción de los datos se utilizaron principalmente datasets obtenidos desde la plataforma Kaggle, lo que permitió acceder a grandes volúmenes de información confiable, diversa y estructurada, adecuada para cada uno de los cinco temas seleccionados:

- Factores de colocación entre estudiantes
- Noticias mundiales
- Actividades y hobbies

- Restaurantes y lugares de esparcimiento
- Eventos de futbol a nivel mundial

Cada uno de estos temas aportó un conjunto relevante de datos, principalmente en formatos CSV y JSON, los cuales fueron extraídos directamente para ser utilizados en las siguientes fases del proyecto.

La extracción inicial permitió reunir un volumen superior al millón de registros en bruto, constituyendo así una base sólida para el posterior procesamiento, limpieza y análisis de datos.

Anexos por tema:

- Anexo B-1: Factores de colaboración entre estudiantes.
- Anexo B-2: Recorrido de estudiantes
- Anexo B-3: Actividades y hobbies
- Anexo B-4: Recorrido de restaurantes
- Anexo B-5: Eventos de futbol a nivel mundial

2. *Procesamiento intermedio Python ETL (Jupyter notebook).*

Una vez realizada la extracción de datos, se llevó a cabo el procesamiento intermedio utilizando Jupyter Notebook como entorno de desarrollo y la librería Pandas como herramienta principal para el tratamiento de datos.

Las tareas aplicadas de forma general para todos los temas incluyeron:

- Eliminación de datos duplicados.
- Detección y tratamiento de valores nulos o vacíos.
- Normalización de columnas y estandarización de formatos.
- Generación de scripts que automatizan el proceso de limpieza.

Este procesamiento permitió dejar los datasets listos para su posterior análisis, asegurando su calidad y consistencia.

A continuación, se detallan los temas procesados y sus respectivos anexos con evidencia del trabajo realizado:

- Anexo C-1: Limpieza – Factores de colaboración entre estudiantes.
- Anexo C-2: Limpieza – Noticias Mundiales
- Anexo C-3: Limpieza – Actividades y hobbies
- Anexo C-4: Limpieza – Recorrido de restaurantes
- Anexo C-5: Limpieza – Eventos de fútbol

3. Almacenamiento

Los datos pasaron por distintos entornos de bases de datos dependiendo del tipo de análisis y destino final:

Eventos mundiales de fútbol: (CSV) a MongoDB como (JSON)

Factores de colaboración entre estudiantes y noticias mundiales: de MongoDB y Redis (CSV → JSON)

Actividades y hobbies y recorrido de restaurantes: de SQLite y MongoDB (CSV → JSON)

Posteriormente, todos los temas fueron centralizados en MongoDB Atlas, desde donde se migraron a SQL Server para finalmente ser visualizados en Power BI.

Anexos de flujos de integración por tema:

- Anexo D-1: Flujo técnico – Factores de colaboración entre estudiantes.
- Anexo D-2: Flujo técnico – Recorrido de estudiantes
- Anexo D-3: Flujo técnico – Actividades y hobbies
- Anexo D-4: Flujo técnico – Recorrido de restaurantes
- Anexo D-5: Flujo técnico – Eventos de futbol a nivel mundial

Análisis de Información

Durante la fase de análisis se aplicaron técnicas exploratorias para comprender los patrones subyacentes en los datos recopilados. Se realizaron estudios independientes por temática, los cuales se detallan a continuación:

a) Factores de Colaboración entre Estudiantes

Se identificó que los estudiantes con mayores puntajes en habilidades de comunicación mostraron una mayor probabilidad de ser colocados en empleos. Esto se visualiza en la Fig. 1, donde la suma de habilidades de comunicación es considerablemente mayor para quienes lograron colocación laboral.

b) *Actividades y Hobbies*

Se encontró una relación inversa entre el precio de las actividades y la cantidad de participantes, como se evidencia en la Fig. 2. Las actividades más costosas (mayores a $100) tienden a tener una participación individual o muy reducida. En cambio, las actividades de bajo costo fueron más comunes entre grupos grandes, como lo demuestra la Fig. 3 al mostrar que el tipo "social" lidera en volumen de participación.

c) *Recorrido de Restaurantes*

Los usuarios tienden a dejar reseñas cuando su experiencia fue muy buena o muy mala, confirmando un sesgo de polarización. La Fig. 4 lo muestra mediante una concentración de puntajes extremos. Asimismo, la calificación promedio ha aumentado a lo largo del tiempo (Fig. 5), y las palabras clave en los

resúmenes permiten anticipar el puntaje otorgado (Fig. 6), con términos como “delicious” o “excellent” asociados a altas calificaciones.

d) Eventos de Fútbol

El análisis mostró un aumento en el volumen de eventos registrados desde 2008, indicando mayor cobertura y participación mediática. Estos hallazgos fueron útiles para comprender la evolución del fútbol como fenómeno global.

e) Noticias Mundiales

A través del análisis de sentimientos se determinó que ciertos países presentan una mayor cantidad de noticias con polaridad negativa. Se aplicaron filtros y segmentaciones geográficas para profundizar en los hallazgos por región.

Visualización de Información

Se hicieron 15 visualizaciones (1 por caso de estudio), que resume los aspectos importantes que influyen en resolver el caso de estudio planteado previamente.

A continuación, se muestran las visualizaciones anexadas por tema:

- Anexo E-1:Caso 01: Usuarios Con Mayor Impacto
- Anexo E-2:Caso 02: Noticias más Vistas
- Anexo E-3:Caso 03 Influencia de las Habilidades de Comunicación en la Colocación
- Anexo E-4:Caso 04: Rol de la Experiencia en Pasantías en los Resultados de Colocación
- Anexo E-5:Caso 05: Existe relación entre el resumen (Summary) y el puntaje dado
- Anexo E-6:Caso 06: Tienden los usuarios a dejar solo reseñas con puntajes extremos
- Anexo E-7:Caso 07: ¿Cómo ha cambiado la calificación promedio con el tiempo?
- Anexo E-8:Caso 08: ¿Las actividades más caras tienden a ser individuales o grupales?
- Anexo E-9:Caso 09: ¿Qué tipo de actividad es más común para grupos grandes?
- Anexo E-10:Caso 10: Acciones Realizadas Por Jugador
- Anexo E-11:Caso 11: Acciones por hora de juego
- Anexo E-12:Caso 12: Equipos con mayores interacciones en partido
- Anexo E-13:Caso 13: Equipos con mayor tiempo de juego por Jugador

III. CONCLUSIONES

En base a los datos, visualizaciones y casos de estudio presentados, hemos llegado a los siguientes resultados y conclusiones.

A. CASOS DE ESTUDIO

Caso 01: Usuarios como Mezvan y Ratón Colorado dominan en karma y número de clics en una plataforma de noticias, lo que indica que tienen una gran capacidad para generar contenido influyente. Estas métricas pueden utilizarse para segmentar influencers informativos dentro de una comunidad y facilitar campañas o estrategias de contenido.

Caso 02: Las publicaciones que abordan experiencias personales profundas o usan títulos provocativos reciben más clics, aunque no necesariamente votos positivos. Esto demuestra la importancia de considerar tanto la reacción emocional como la percepción de calidad al analizar impacto mediático.

Caso 03 Los estudiantes con mejores habilidades de comunicación tienen tasas significativamente mayores de colocación laboral. Esto sugiere que este tipo de habilidades blandas son tan importantes como el rendimiento académico y deben ser reforzadas desde etapas tempranas de la formación profesional.

Caso 04: Se evidenció que tener experiencia en pasantías incrementa considerablemente las oportunidades de colocación. Las instituciones educativas pueden utilizar esta información para fortalecer sus convenios empresariales y promover activamente las prácticas preprofesionales.

Caso 05: Se identificó que palabras como *"delicious"*, *"excellent"*, *"love it"* y *"great"* están claramente asociadas con calificaciones altas. Este patrón confirma que el lenguaje positivo en los resúmenes de reseñas es un buen predictor del sentimiento y de la puntuación final, lo cual es útil para modelos de análisis de sentimientos automatizados.

Caso 06: Los usuarios tienden a dejar reseñas cuando su experiencia es muy buena o muy mala, lo que genera un sesgo de polarización. Esto se observó en la distribución de puntuaciones acumuladas, donde predominan valores extremos. Este hallazgo es clave para interpretar los análisis de reseñas en plataformas de consumidores, y se pueden aplicar filtros de normalización o ponderación en los modelos para ajustar dicho sesgo.

Caso 07: Desde 2008 en adelante, se evidenció un aumento constante en la cantidad de calificaciones en línea, alcanzando un pico en 2012. Esto podría estar vinculado al crecimiento de plataformas digitales y mayor participación de usuarios. Este crecimiento en la participación implica mayor diversidad de opiniones y datos más representativos para análisis de tendencias.

Caso 08: Las actividades de bajo costo (menores a $20) tienden a atraer más participantes, mientras que las costosas (más de $100) son predominantemente individuales. Esto sugiere que al diseñar actividades comunitarias, sociales o escolares, se deben priorizar opciones accesibles económicamente para maximizar la participación.

Caso 09: Las actividades del tipo “social” lideran en número de participantes, seguidas de las “recreativas” y “educativas”. Esto permite enfocar el diseño de eventos hacia estas categorías cuando se busca alta asistencia, como en programas municipales o escolares.

Caso 10: Excluyendo pases, las acciones más comunes fueron

los duelos y el control individual del balón. Esto evidencia un estilo de juego de alta intensidad, con múltiples enfrentamientos por la posesión, útil para definir estrategias de marcaje y recuperación en el medio campo.
Caso 11: Se observó que la mayoría de acciones se registraron en la primera mitad del partido. Esta caída en la segunda mitad podría explicarse por fatiga, ajustes tácticos o desgaste físico. Este tipo de análisis ayuda a planificar mejor las rotaciones y la preparación física.
Caso 12: Arabia Saudita y Uruguay fueron los equipos con mayor cantidad de interacciones por partido, lo que refleja un estilo de juego más dinámico o de presión alta. Esta información es valiosa para preparar estrategias defensivas ante este tipo de selecciones.
Caso 13: Rumania y Suiza destacaron con más de 40.000 minutos jugados, lo que sugiere un uso intensivo de su plantilla titular. Si bien esto mejora la cohesión táctica, también puede aumentar el riesgo de fatiga y lesiones acumuladas.

B. ANALISIS DE SENTIMIENTOS

Se aplicaron técnicas de procesamiento de lenguaje natural (NLP) y análisis de sentimientos a textos relacionados con el rendimiento de equipos de fútbol. Se utilizó Python con bibliotecas como TextBlob, VADER y/o Transformers.
Se visualiza en el Anexo E-14

IV. DIFICULTADES

Durante la realización de este proyecto nos encontramos con varios desafíos, los cuáles se detallan a continuación:

V. ANEXOS

Anexos-A Cronograma

Anexo A-1 Diagrama de Gantt del cronograma de Actividades.

Diagrama de Gantt – Proyecto Análisis de Datos 2025A

Actividad	Responsable	Semana 1	Semana 2	Semana 3	Semana 4	6
Selección de temas y casos de estudio	Todos					
Búsqueda y descarga de datasets	María Giron					
Limpieza y transformación de datós	María Giron					
Conversión de CSV a JSON	María Girón					
Subida de datos a MongoDB Atlas	Todos					
Análisis de datos y generación de reporrtes	Todos					
Elaboración del informe final	Todos					
Preparación del video explicativ	Marìa Girón					
Presentación y entrega del proyecto	Todos					

Anexos-B Extracción de datos

Anexo B-1 – Dataset extraído: Factores de colaboración entre estudiantes

```
import pandas as pd

# 1. Cargar el dataset
df = pd.read_csv("C:\ProyectAnalisis2025A\extract\college_student_placement_dataset.csv")

# 2. Revisamos la estructura del dataset
print("Columnas:", df.columns.tolist())
print("Primeras filas:\n", df.head())
print("\nValores nulos:\n", df.isnull().sum())
print("\nDuplicados:", df.duplicated().sum())

# 3. Renombramos las columnas (evitar espacios o caracteres especiales)
df.columns = [col.strip().replace(" ", "_").replace("'", "").lower() for col in df.columns]

# 4. Eliminamos duplicados de la data
df.drop_duplicates(inplace=True)

# 5. Manejar valores nulos (ejemplo: rellenar numéricos con 0, categóricos con 'Desconocido')
for col in df.select_dtypes(include=["float64", "int64"]):
    df[col].fillna(0, inplace=True)

for col in df.select_dtypes(include=["object"]):
    df[col].fillna("Desconocido", inplace=True)

# 6. Exportar CSV limpio
df.to_csv("collage_students_placement_clean.csv", index=False)
print("Archivo limpio exportado como students_clean.csv")
```

Anexo B-3 – Dataset extraído: Actividades y hobbies

```
import pandas as pd
import json


# RUTA DEL ARCHIVO JSON
ruta = 'C:/ProyectAnalisis2025A/extract/activity.json'
df = pd.read_json(ruta, lines=True)
print("Dimensiones del DataFrame:", df.shape)
df.head()
```

```
Dimensiones del DataFrame: (32000, 4)
```

	activity	type	participants	price
0	Meditate for five minutes	relaxation	1	0.0
1	Uninstall unused apps from your devices	busywork	1	0.0
2	Learn to greet someone in a new language	education	1	0.1
3	Go on a long drive with no music	relaxation	1	0.1
4	Sit in the dark and listen to your favorite mu...	relaxation	1	0.0

Anexo B-4 – Dataset extraído: Recorrido de restaurantes

```
import pandas as pd


# DEFINICION DE LA RUTA DEL ARCHIVO
ruta_origen = 'C:/ProyectAnalisis2025A/extract/Reviews.csv'
df = pd.read_csv(ruta_origen)
# VISUALIZACION DE LOS DATOS
print("Filas originales: ",len(df))
df.head()
```

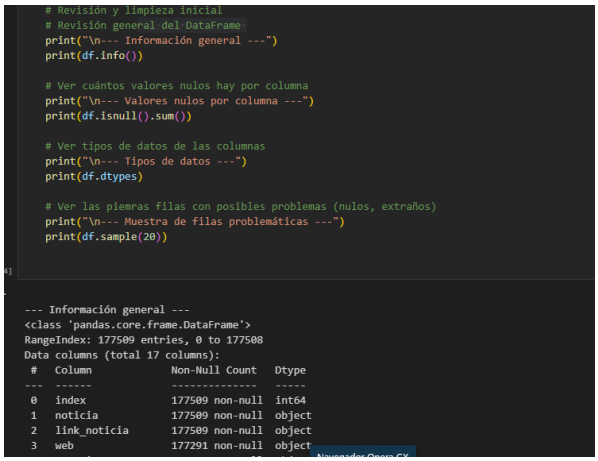
Anexo B-5 – Dataset extraído: Eventos mundiales de fútbol

```
#Importar libreria
import pandas as pd


#llamar archivos principales
event1=pd.read_csv("events_England.csv")
event2=pd.read_csv("events_European_Championship.csv")
event3=pd.read_csv("events_France.csv")
event4=pd.read_csv("events_Germany.csv")
event5=pd.read_csv("events_Italy.csv")
event6=pd.read_csv("events_Spain.csv")
event7=pd.read_csv("events_World_Cup.csv")


#llamar archivos secundarios
player_df=pd.read_csv('player_games.csv')
tags_df=pd.read_csv('tags2name.csv')
teams_df=pd.read_csv('teams.csv')
event_df=pd.read_csv('events_England.csv')
```

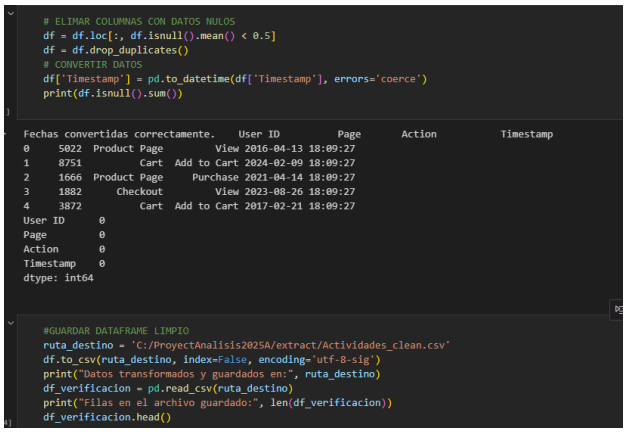
Anexos-C Procesamiento intermediario

Anexo C-1 – Limpieza y procesamiento: Factores de colaboración entre estudiantes

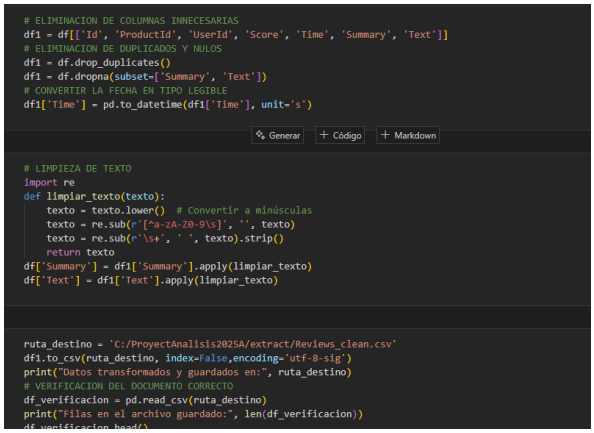
Anexo C-2 – Limpieza y procesamiento: Noticias a nivel mundial



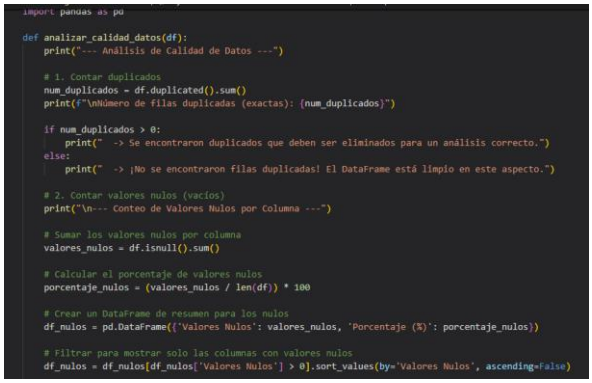
Anexo C-3 – Limpieza y procesamiento: Actividades y hobbies



Anexo C-4 – Limpieza y procesamiento: Recorrido de restaurantes



Anexo C-5 – Limpieza y procesamiento: Eventos de fútbol



Anexos-D Almacenamiento

Anexo D-1 – Integración de datos: Factores de colaboración entre estudiantes

Anexo D-2 – Integración de datos: Noticias a Nivel Mundial.

```python
Haga clic para agregar un punto de interrupción
#Seleccionar base de datos
db= client["Datas"]
coleccion = db["noticias"]


# Leer archivo JSON
with open(r'C:\Users\ALIEN PC EC\Desktop\Universidad\Proyectos\ProyectAnalisis2025A\extract/data_clean.json', "r", en
    data = json.load(f)
# Insertar datos
if isinstance(data, list):
    coleccion.insert_many(data)
else:
    coleccion.insert_one(data)

print("Datos subidos exitosamente a MongoDBAtlas")
```

```
Datos subidos exitosamente a MongoDBAtlas
```

Anexo D-3 – Integración de datos: Actividades y hobbies

```python
#Eduardo Ganchala
from pymongo import MongoClient
import json
# Conexion con mongo db atlas
client=MongoClient("mongodb+srv://eduardoganchala29:jyxu5JIMqQk9JLtS@adcluster.hz5cw.mongodb.net/?retryWrites=tru

#Seleccionar base de datos
db= client["Datas"]
coleccion=db["Actividades_Hobbies"]


# Leer archivo JSON
with open("C:/ProyectAnalisis2025A/extract/Actividades.json", "r", encoding="utf-8") as f:
    data = json.load(f)
# Insertar datos
if isinstance(data, list):
    coleccion.insert_many(data)
else:
    coleccion.insert_one(data)

print("Datos subidos exitosamente a MongoDB Atlas")
```

Anexo D-4 – Integración de datos: Recorrido de restaurantes

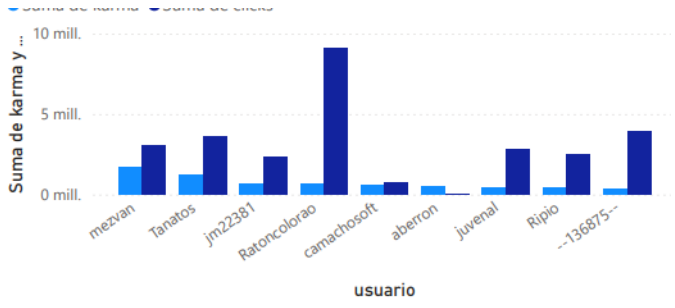
```python
from pymongo import MongoClient
import json
# Conexion con mongo db atlas
client=MongoClient("mongodb+srv://eduardoganchala29:jyxu5JIMqQk9JLtS@adcluster.hz5cw.mongodb.net/?retryWrites

#Seleccionar base de datos
db= client["Datas"]
coleccion = db["Restaurantes"]


# Leer archivo JSON
with open(r'C:/ProyectAnalisis2025A/extract/Reviews1.json', "r", encoding="utf-8") as f:
    data = json.load(f)
# Insertar datos
if isinstance(data, list):
    coleccion.insert_many(data)
else:
    coleccion.insert_one(data)

print("Datos subidos exitosamente a MongoDBAtlas")
```

Anexo D-5 – Integración de datos: Eventos de fútbol

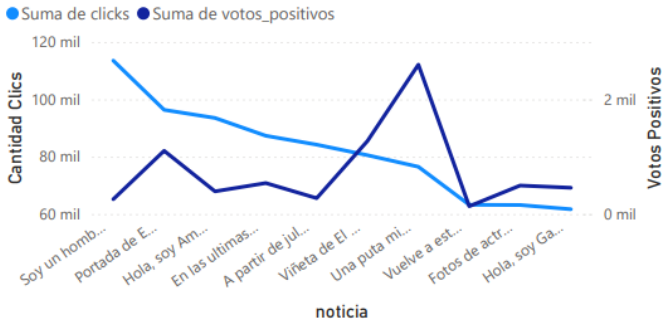
```python
# Crear cadena de conexión con sqlalchemy

# Cargar CSV
Cargando...
df = pd.read_csv("eventos_final.csv")

# Crear conexión
engine = create_engine('postgresql+psycopg2://postgres:12345@localhost:5432/Eventos_Futbol')

# Subir DataFrame como nueva tabla (se crea automáticamente)
df.to_sql('Events', con=engine, if_exists='replace', index=False)

#Leer la tabla para verificar que se creó correctamente (opcional)
df_pg = pd.read_sql('SELECT * FROM "Events"', engine)
print(df_pg.head())
```

Anexos-E Visualizaciones

Anexo E-1:Figura 01: Usuarios Con Mayor Impacto

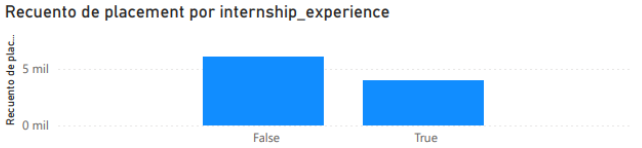


Anexo E-2:Caso 02: Noticias más Vistas

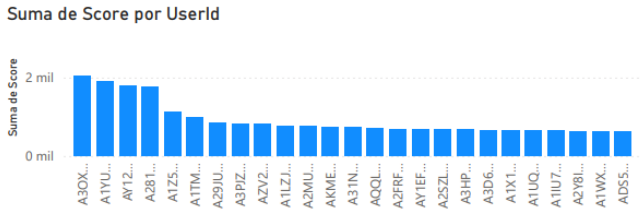


Anexo E-3:Caso 03 Influencia de las Habilidades de Comunicación en la Colocación

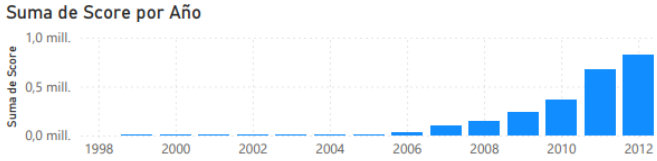
Anexo E-4:Caso 04: Rol de la Experiencia en Pasantías en los Resultados de Colocación



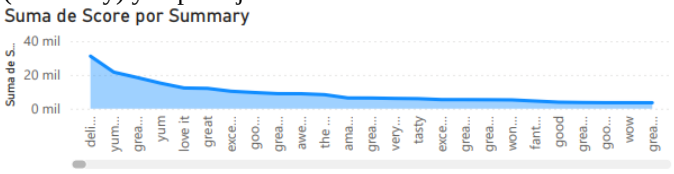
Anexo E-6:Caso 05: Tienden los usuarios a dejar solo reseñas con puntajes extremos



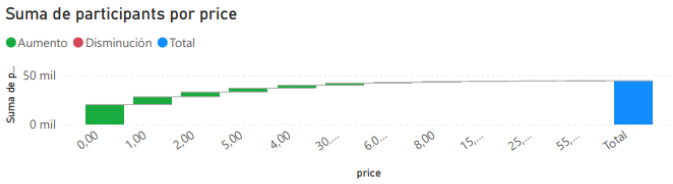
Anexo E-7:Caso 06: ¿Cómo ha cambiado la calificación promedio con el tiempo?



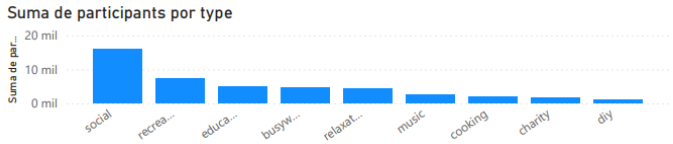
Anexo E-5:Caso 07: Existe relación entre el resumen (Summary) y el puntaje dado



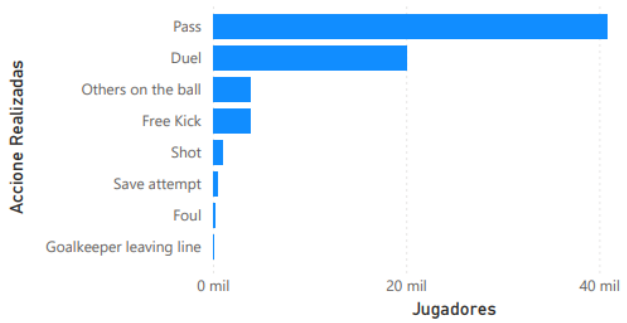
Anexo E-8:Caso 08: ¿Las actividades más caras tienden a ser individuales o grupales?



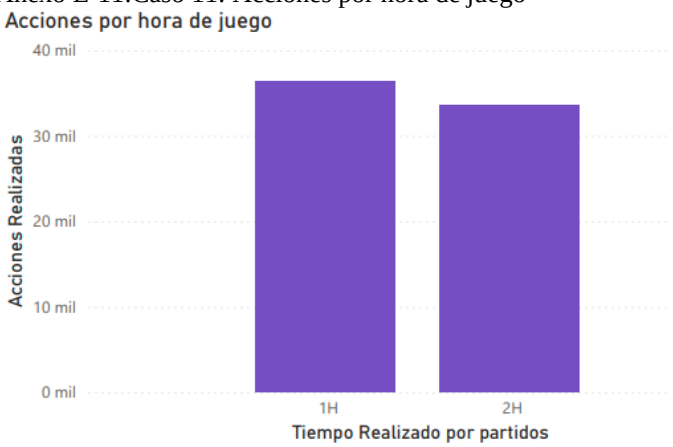
Anexo E-9:Caso 09: ¿Qué tipo de actividad es más común para grupos grandes?



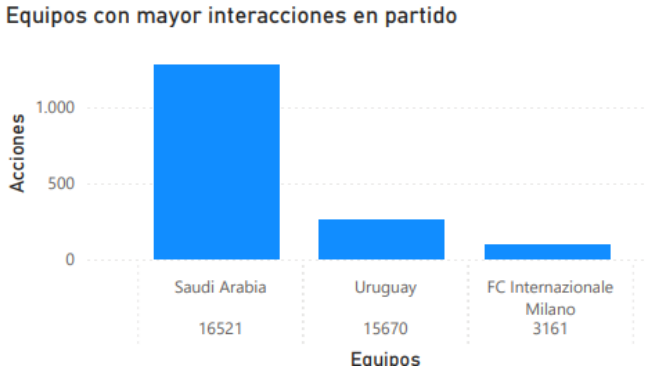
Anexo E-10:Caso 10: Acciones Realizadas Por Jugador



Anexo E-11:Caso 11: Acciones por hora de juego

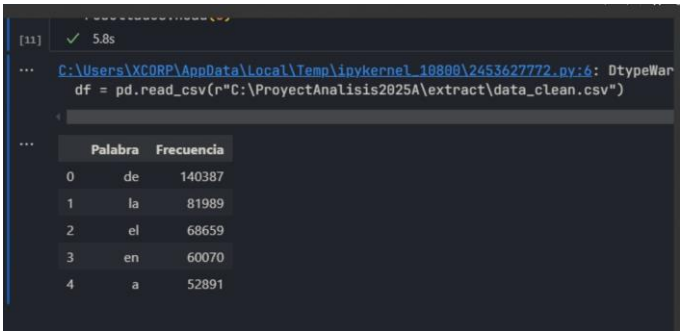


Anexo E-12:Caso 12: Equipos con mayores interacciones en partido



Anexo E-13:Caso 13: Equipos con mayor tiempo de juego por Jugador

Anexo E-14: Análisis de sentimientos



VI. ANEXOS

Enlace a Github: https://github.com/MariaGiron-code/ProyectoIntegrador_BasesDatos.git

Recopilación de datos: https://drive.google.com/file/d/1ho0j7np6O38K6xRsM9t6CIhgRIKoEm_V/view?usp=drive_link

Videos de presentación:

Extracción de datos: https://drive.google.com/file/d/1ho0j7np6O38K6xRsM9t6CIhgRIKoEm_V/view?usp=drive_link

Resultado obtenido de la extracción de datos: https://youtu.be/2VasxBssBE8