Proyecto Final Análisis

Shamyr Quishpe, Freddy Villavicencio, Alejandro Quiroz, Francisco Caero y Cristian Chulde.

ESFOT, Escuela Politécnica Nacional
Quito, Ecuador

Resumen —

En este texto se podrá observar la realización del trabajo final del grupo 1 en la materia de análisis de datos de la Escuela Politécnica Nacional. A través de este se demostrarán y aplicarán los conocimientos adquiridos durante el semestre mediante realización de cinco casos de estudio. Los casos seleccionados y discutidos con el ingeniero de la materia fueron: música a nivel mundial, Juegos en línea por países, películas, Deportes a nivel mundial, Eventos o noticias mundiales.

Para el desarrollo de estos casos de estudio se realizó la recopilación de información de 12 fuentes diferentes, estas fuentes fueron almacenadas dentro de diversas bases de datos, limpiadas y unificadas para luego realizar la creación de un Data lake que permitiera realizar el tratado de datos de la mejor manera posible.

Una vez hecho el tratado de datos y el almacenamiento dentro del datalake se procedió a generar una conexión con Power Bi, herramienta que fue utilizada para generar objetos visuales de los datos que permitieran sustentar las conclusiones generadas.

A continuación, se podrán observar los resultados obtenidos de cada caso de estudio, junto con explicaciones generales de la realización.

INDÍCE

I. OBJETIVOS GENERAL

El objetivo general del presente informe es presentar de manera detallada el desarrollo y los resultados del proyecto, desde la definición del caso de estudio hasta las conclusiones y recomendaciones obtenidas.

II. OBJETIVOS ESPECÍFICOS

- Definir los casos de estudio, especificando las temáticas seleccionadas para el análisis de datos.
- Presentar el análisis de la información obtenida, incluyendo hallazgos significativos y las tendencias identificadas en cada caso de estudio.
- Mostrar las visualizaciones de información generadas a partir de los datos.

III. DISTRIBUCIÓN DE TRABAJO

- Shamyr Sebastián Quishpe Cuadrado

Líder del Proyecto: Coordinación general del equipo, planificación de actividades y seguimiento del cronograma.
Extracción de Datos: Encargado de desarrollar y ejecutar los scripts para la extracción de datos de las diferentes fuentes.
Análisis de Información: Responsable de realizar el análisis exploratorio de datos.

- Freddy Antonio Villavicencio Rosendo

Diseñador de Arquitectura: Encargado de diseñar la arquitectura del Data Lake, seleccionando las bases de datos y herramientas adecuadas.
Conexiones: Encargado de la conexión entre el data lake y el power BI.

Conclusiones y Recomendaciones: Encargado de elaborar las conclusiones y recomendaciones finales basadas en los resultados obtenidos.

- Alejandro Quiroz

Gestor de Bases de Datos: Responsable de administrar y gestionar las bases de datos relacionales y NoSQL utilizadas en el proyecto.
Desarrollador de Scripts: Encargado de escribir y mantener los scripts para la transformación y carga de datos en el Data Lake.
Desafíos y Problemas: Encargado de identificar y resolver los desafíos y problemas técnicos encontrados durante el desarrollo del proyecto.

- Francisco Caero

Analista de Datos: Responsable de realizar análisis estadísticos avanzados y modelado de datos para identificar patrones y tendencias.
Coordinador de Presentación: Encargado de coordinar la preparación y diseño de la presentación del proyecto.
Revisión de Contenido: Responsable de revisar y asegurar la calidad del contenido del informe y la presentación.

- Cristian Ermel Usiña Chulde

Investigador de Fuentes de Datos: Encargado de investigar y recopilar información adicional sobre las fuentes de datos seleccionadas.
Apoyo en Extracción: Asiste en el proceso de extracción de datos, proporcionando soporte técnico y resolviendo problemas.
Asistente de Presentación: Ayuda en la preparación de la presentación, creando diapositivas y asegurando la coherencia visual.

- Trabajo individual de cada persona.

IV. CRONOGRAMA

Fecha	Actividades	Encargados
24/2/2024	Reunión inicial del equipo para discutir el alcance del proyecto y asignación de roles.	Todo el grupo
25/2/2024	Investigación sobre las fuentes de datos seleccionadas.	Alejandro
26/2/2024	Desarrollo de scripts de extracción de datos.	Alejandro, Francico
27/2/2024	Extracción de datos de las diferentes fuentes.	Freddy, Cristian
28/2/2024	Procesamiento y transformación de datos en el Data Lake.	Shamyr
29/2/2024	Diseño y creación de visualizaciones en PowerBI.	Shamyr, Freddy
1/3/2024	Análisis de la información obtenida y generación de insights.	Cristian
2/3/2024	Elaboración del informe final y revisión de contenidos	Todo el grupo

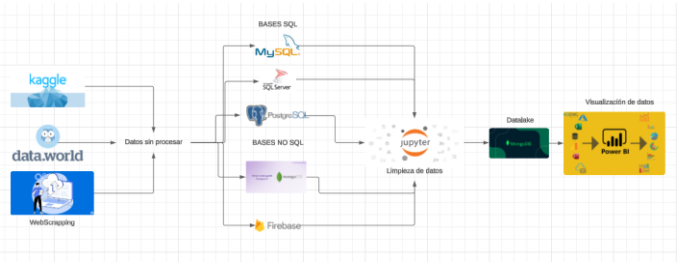
V. RECURSOS Y HERRAMIENTAS UTILIZADAS

Para el desarrollo de este proyecto, utilizamos una variedad de recursos y herramientas que nos permitieron llevar a cabo las diferentes etapas de análisis de datos de manera eficiente. Entre los recursos principales destacan las plataformas de colaboración y almacenamiento como GitHub, que facilitó la gestión del código y la colaboración entre los miembros del equipo. También recurrimos a Kaggle y DataWorld para acceder a conjuntos de datos relevantes y complementar nuestra fuente de datos.

En cuanto a las bases de datos, implementamos una arquitectura que incluye MongoDB Atlas, MongoDB Compass y Firebase para la gestión de bases de datos NoSQL. Además, integramos MySQL, PostgreSQL y Microsoft SQL Server para bases de datos relacionales, lo que nos permitió trabajar con una amplia gama de datos estructurados.

Para el análisis y la visualización de datos, utilizamos Jupyter Notebook como entorno de programación interactivo, lo que nos permitió realizar análisis estadísticos y modelado de datos de manera eficiente. Finalmente, empleamos Power BI para la creación de dashboards interactivos y la visualización de resultados, proporcionando una herramienta poderosa para comunicar nuestros hallazgos de manera clara y efectiva

VI. ARQUITECTURA DE LA SOLUCIÓN



VII. CASOS DE USO

a. Freddy: MUSICA A TRAVÉS DEL MUNDO

Este caso de estudio se enfoca en examinar la relación entre diversas variables relacionadas con la popularidad de canciones y artistas en plataformas digitales como Spotify y YouTube. A partir de una recopilación de datos que incluye información detallada sobre distintas canciones, artistas y métricas de popularidad, se busca comprender cómo factores como el género musical, la duración de la canción y la disponibilidad de un video oficial pueden influir en la recepción y la difusión de la música en línea. Este análisis proporcionará insights valiosos para artistas, sellos discográficos y otros actores de la industria musical.

Objetivo General:

El objetivo general de este caso de estudio es analizar la relación entre diversas variables y la popularidad de las canciones en plataformas digitales, utilizando datos reales de la industria musical.

Objetivos Específicos:

- Identificar patrones y tendencias en la popularidad de canciones en función del género musical.
- Evaluar la influencia de la duración de la canción en su recepción por parte del público.
- Determinar si la disponibilidad de un video oficial afecta la difusión y el alcance de una canción en plataformas como YouTube.
- Proporcionar recomendaciones basadas en los hallazgos para artistas y sellos discográficos que buscan maximizar el impacto de su música en el entorno digital.

Metodología:

Los datos sin procesar de este caso de uso fueron adquiridos de la plataforma Kaggle. Estos conjuntos de datos se cargaron inicialmente en la base de datos de Google Firebase. Para extraer todas las colecciones de datos de Firebase, se utilizó el siguiente código:

Una vez completada la extracción de datos, se procedió a limpiar todos los dataframes generados utilizando la biblioteca Pandas. Esto implicó la eliminación de valores nulos, la corrección de errores de formato y cualquier otro tipo de limpieza necesaria para asegurar la calidad de los datos. Posteriormente, se generó un archivo CSV para poder subir los datos tratados al Data Lake de mongo db Atlas.

El código se encuentra en el siguiente repositorio:

https://github.com/ShamyrQuishpe/scriptFreddy

Análisis de datos:

Como mis datos están bastante diversificados, quise realizar un análisis exhaustivo que abarcara diferentes aspectos de la industria musical. Desde la exploración de la diversidad cultural y la distribución de géneros a lo largo del tiempo hasta la evaluación de factores como la conexión emocional con el público y la influencia de la promoción en las ventas de canciones. Este enfoque integral me permitió obtener una visión completa de la dinámica de la industria y proporcionar insights valiosos para los actores involucrados en la producción y promoción de música en plataformas digitales.

Primera visualización:

En base a los datos recopilados en mi investigación podemos observar que los países con más diversidad de géneros y artistas son estados y reino unidos lo que nos podría llevar a pensar en que su presencia en la industria musical es bastante grande.

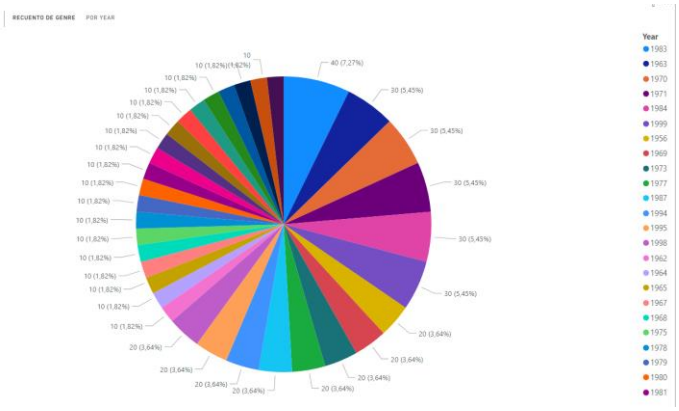
Realicé un análisis para confirmar la afirmación de que Estados Unidos y Reino Unido son los países con mayor cantidad de géneros musicales y artistas. Para ello, exploré diversas fuentes de información disponibles en línea y consulté informes de la industria musical y estadísticas de plataformas de streaming. [11]

Los datos recopilados revelaron que, si bien Estados Unidos y Reino Unido son dos de los países más influyentes en la industria musical a nivel mundial, la afirmación de que son los países con mayor cantidad de géneros musicales y artistas no es completamente precisa. Si bien ambos países tienen una rica historia musical y han contribuido significativamente al desarrollo de diversos géneros, otros países también tienen una presencia destacada en la escena musical global.

Por ejemplo, países como Brasil, Japón, Alemania y China son reconocidos por su diversidad musical y por ser cunas de géneros y artistas influyentes en la música mundial. Además, la globalización y la accesibilidad a través de plataformas de streaming han permitido que artistas de una amplia variedad de países y culturas alcancen audiencias internacionales. Que es un tema que será analizado próximamente [10].

En conclusión, la primera visualización realizada en base a los datos recopilados nos deja una afirmación que tiene una precisión bastante alta

Segunda visualización:



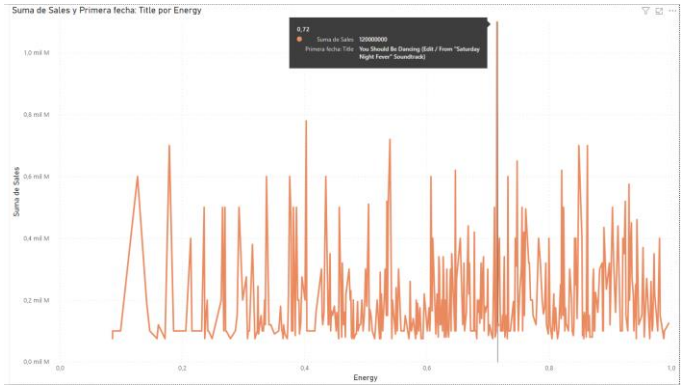
Después de realizar un análisis detallado de la distribución de géneros musicales por año puedo confirmar que el año 1983 destaca como un período con una notable diversidad en géneros musicales. Durante este año, se observa una amplia gama de géneros musicales que abarcan desde el pop y el rock hasta el hip-hop, el reggae y la música electrónica, entre otros. Esta diversidad puede atribuirse a una serie de factores, como el

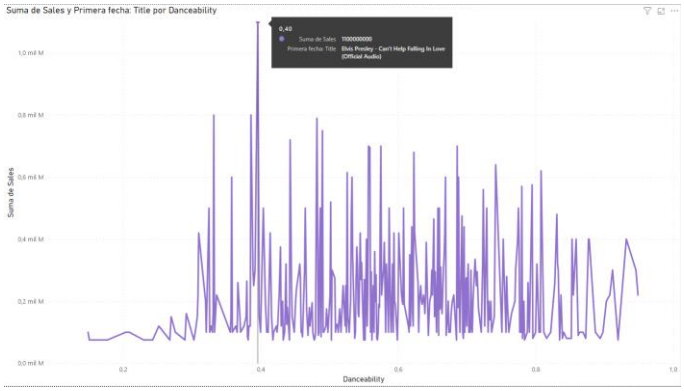
surgimiento de nuevos estilos musicales, la influencia de movimientos culturales y sociales, y el avance tecnológico en la producción musical.

Sin embargo, es importante destacar que la distribución de géneros musicales por año tiende a ser bastante uniforme en general. Aunque el año 1983 destaque por su diversidad, otros años pueden mostrar una distribución similar de géneros musicales, lo que sugiere que la cantidad de géneros generados por año no varía significativamente a lo largo del tiempo. Esta uniformidad puede reflejar la naturaleza cambiante pero constante de la industria musical, donde diferentes géneros coexisten y evolucionan de manera continua. [11]

En conclusión, si bien el año 1983 destaca por su diversidad en géneros musicales, la distribución uniforme de géneros a lo largo de los años sugiere que no hay una variación significativa en la cantidad de géneros generados por año. Esto subraya la constante evolución y adaptación de la música a lo largo del tiempo, así como la riqueza y diversidad de la escena musical en general.

Tercera visualización:





El gráfico que he creado compara las ventas de canciones en función de su bailabilidad y energía. Para representar esta comparación, he utilizado un sistema de coordenadas donde el eje x representa la bailabilidad de las canciones en una escala del 0 al 1, y el eje y representa la energía de las canciones, también en una escala del 0 al 1.

Cada punto en el gráfico representa una canción específica, donde su posición en los ejes x e y corresponde a su nivel de bailabilidad y energía respectivamente. Los puntos se distribuyen en el gráfico de acuerdo con los valores de bailabilidad y energía de cada canción.
La ausencia de una correlación significativa entre la bailabilidad, la energía y las ventas sugiere que otros factores pueden influir más en el éxito comercial de una canción, como la promoción, la popularidad del artista o la calidad de la producción.

La presencia de una zona muerta en la escala de bailabilidad entre 0 y 3.1 indica que las canciones con una baja bailabilidad pueden tener una recepción menos favorable por parte del público, lo que puede afectar sus ventas.

La identificación de picos de energía en ciertas canciones sugiere que estas pueden generar un mayor interés y compromiso emocional por parte del público, lo que podría traducirse en un aumento de las ventas.

La importancia del repertorio y la identidad del artista se refleja en la presencia de canciones específicas de un mismo artista en áreas destacadas del gráfico. Esto sugiere que los fanáticos pueden estar más inclinados a comprar canciones de artistas que les gustan, independientemente de su bailabilidad o energía.

En conclusión, aunque la bailabilidad y la energía son aspectos importantes en la música, su relación con las ventas de canciones puede no ser directa. Otros factores, como la promoción, la identidad del artista y la conexión emocional con el público, pueden desempeñar un papel más significativo en el éxito comercial de una canción.

Resultados obtenidos:

- Diversidad de géneros y artistas: Aunque Estados Unidos y el Reino Unido tienen una presencia significativa en la industria musical, otros países como Brasil, Japón, Alemania y China también son importantes en términos de diversidad musical y contribución a la escena global. Esta diversidad refleja la naturaleza multicultural de la música y la globalización de la industria.

- Distribución de géneros por año: Aunque el año 1983 destacó por su diversidad en géneros musicales, la distribución general de géneros a lo largo de los años muestra una uniformidad relativa. Esto sugiere una constante evolución y adaptación de la música a lo largo del tiempo, con diferentes géneros coexistiendo y evolucionando de manera continua.

- Correlación entre bailabilidad, energía y ventas de canciones: No se encontró una correlación significativa entre la bailabilidad, la energía y las

ventas de canciones. Esto sugiere que otros factores como la promoción, la popularidad del artista y la calidad de la producción pueden tener un impacto más significativo en el éxito comercial de una canción.

- Importancia de la identidad del artista y la conexión emocional: Se observó que los fanáticos pueden estar más inclinados a comprar canciones de artistas que les gustan, independientemente de su bailabilidad o energía. Esto resalta la importancia de la conexión emocional y la lealtad de los seguidores hacia los artistas en el éxito de ventas.

b. Alejandro: VALOR DE JUGADORES Y ÉXITO EN EL FÚTBOL.

En el mundo del fútbol, el éxito deportivo de un equipo está influenciado por una variedad de factores, incluyendo la calidad de sus jugadores y su rendimiento en la cancha. En este informe, analizaremos la relación entre el precio de los jugadores y el éxito deportivo de sus respectivos equipos, así como una comparativa entre regiones y su éxito deportivo. Para lograrlo, utilizaremos datos históricos de ranking de equipos y detalles de jugadores de diferentes regiones para comprender mejor cómo el valor económico de un equipo se relaciona con su desempeño en el campo.

Objetivos Generales:

Evaluar la relación entre el precio máximo de los jugadores y el éxito deportivo de sus equipos.
Comparar el rendimiento deportivo de equipos de diferentes regiones y analizar las posibles disparidades en el éxito deportivo.

Objetivos Específicos:

- Comparación entre precio y éxito deportivo de los equipos:
- Determinar si existe una correlación significativa entre el precio máximo de los jugadores de un equipo y su posición en el ranking.
- Identificar patrones o tendencias en la relación entre el precio de los jugadores y el cambio en el ranking del equipo a lo largo del tiempo.
- Comparativa entre regiones y su éxito deportivo:
- Analizar la distribución del éxito deportivo (puntuación/ranking) entre equipos de diferentes regiones.
- Investigar posibles diferencias en el rendimiento deportivo entre regiones y su relación con factores económicos, demográficos o geográficos.

Metodología:

Recopilación y Preprocesamiento de Datos:

Los datos utilizados en este análisis fueron obtenidos de Kaggle, una plataforma en línea que alberga conjuntos de datos de diversas fuentes. Los conjuntos de datos relevantes consisten en archivos CSV que contienen información sobre el ranking de equipos y detalles de jugadores de fútbol.
Antes de realizar el análisis, los archivos CSV fueron limpiados y preprocesados utilizando Python y la biblioteca Pandas. Durante este proceso, se llevaron a cabo las siguientes tareas:

- Eliminación de datos faltantes o incompletos.
- Conversión de tipos de datos para garantizar la coherencia y la compatibilidad entre las columnas.
- Eliminación de columnas irrelevantes o redundantes que no contribuyan al análisis previsto.

Almacenamiento de Datos:
Una vez que los conjuntos de datos fueron limpiados y preparados en formato CSV, fueron cargados en una base de datos MongoDB local. Esta base de datos local se utilizó como un paso intermedio antes de transferir los datos a una instancia de MongoDB Atlas, una base de datos en la nube que ofrece una infraestructura escalable y segura para almacenar y gestionar grandes volúmenes de datos.
Análisis de Datos:
Los datos almacenados en MongoDB Atlas fueron accesibles a través de una conexión establecida desde el entorno de desarrollo de Python. Utilizando bibliotecas de análisis de datos como Pandas y NumPy, se llevaron a cabo diversas operaciones de análisis para explorar y comprender los conjuntos de datos.
Repositorio del script de limpieza de datos:

https://github.com/AlejoQuiroz08/Analisis-FutbolData

Limitaciones y Consideraciones:

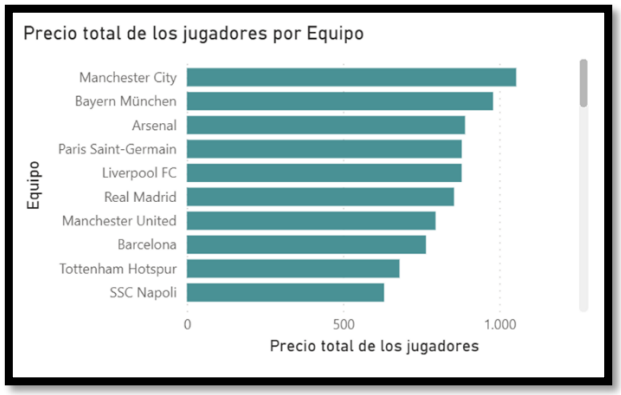
Es importante tener en cuenta algunas limitaciones y consideraciones en relación con la metodología utilizada en este análisis. Estas incluyen:
La calidad y la integridad de los datos dependen en gran medida de la precisión y la actualidad de los conjuntos de datos disponibles en Kaggle.
El proceso de limpieza y preprocesamiento de datos puede influir en los resultados del análisis, y se deben tener en cuenta las decisiones tomadas durante este proceso.
La transferencia de datos entre la base de datos local y MongoDB Atlas puede estar sujeta a restricciones de ancho de banda y velocidad de conexión a Internet, lo que puede afectar el tiempo necesario para completar el análisis.

Análisis de datos:

Antes de presentar los resultados de las visualizaciones, es importante contextualizar el análisis que se ha realizado. En este estudio, nos enfocamos en examinar el precio total de los jugadores por equipo en el ámbito del fútbol profesional. Esta exploración nos permite entender mejor las tendencias de gasto y la competitividad entre los principales clubes europeos durante el periodo analizado.

Utilizando datos de precios de transferencia de jugadores, hemos creado visualizaciones que nos proporcionan una visión clara de cómo se distribuyen estos gastos entre diferentes equipos destacados.

Primera visualización:



Manchester City y Bayern lideran en gastos: Manchester City y Bayern son los equipos con los precios totales de jugadores más altos. Esto sugiere que están dispuestos a invertir significativamente en la adquisición de talento futbolístico, lo que podría reflejar su ambición por el éxito tanto en sus ligas nacionales como en competiciones internacionales.

Equipos de la Premier League dominan el top: Manchester City, Arsenal, Liverpool y Manchester United están entre los primeros lugares en términos de gastos en jugadores. Esto podría indicar la alta competencia financiera y deportiva dentro de la Premier League, con varios equipos grandes compitiendo por el talento y tratando de fortalecer sus equipos para alcanzar el éxito.

Presencia de clubes de élite europeos: Equipos como PSG, Real Madrid y Barcelona también se encuentran entre los principales en términos de precios totales de jugadores. Esto no es sorprendente, ya que estos clubes son conocidos por su estatus de élite en el fútbol europeo y mundial, y a menudo realizan inversiones significativas para mantener su competitividad.

Segunda visualización:

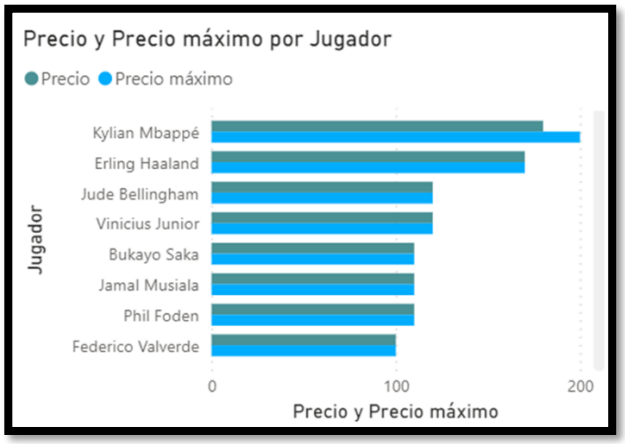
Dominio del mercado europeo: La concentración del precio total de los jugadores en Europa sugiere que este continente es el epicentro del fútbol mundial en términos de inversión en jugadores. Esto puede deberse a la presencia de numerosas ligas de alto nivel en Europa, como la Premier League inglesa, La Liga española, la Serie A italiana, entre otras, que atraen una inversión considerable de clubes de todo el mundo.

Importancia de Argentina y Brasil: La presencia significativa de Argentina y Brasil en términos de precio total de los jugadores indica la riqueza de talento futbolístico en estos países sudamericanos. Argentina y Brasil son conocidos por su historial de producción de jugadores de clase mundial y son exportadores importantes de talento al mercado europeo y otras ligas en todo el mundo.

Factores socioeconómicos y culturales: La concentración del precio total de los jugadores en ciertos países puede estar relacionada con factores socioeconómicos y culturales específicos. Por ejemplo, en países con fuerte tradición futbolística y una cultura arraigada en el deporte, es probable que se invierta más en el desarrollo y adquisición de talento futbolístico.

Impacto en la competencia internacional: La concentración del precio total de los jugadores en Europa, Argentina y Brasil puede influir en la competitividad en el ámbito internacional. Los equipos y selecciones nacionales de estos países pueden beneficiarse de tener acceso a una amplia base de talento futbolístico, lo que potencialmente les otorga una ventaja competitiva en torneos internacionales.

Tercera visualización:



Juventud y potencial: Es notable que la lista esté dominada por jóvenes talentosos como Kylian Mbappé, Erling Haaland, Jude Bellingham, Vinicius Jr. y Bukayo Saka. Esto indica un cambio en la tendencia del mercado, donde los clubes están dispuestos a invertir en jugadores jóvenes con gran potencial de desarrollo y rendimiento a largo plazo.

Valor de mercado en ascenso: La presencia de estos jugadores en el top de los más caros del mundo sugiere un aumento en el valor de mercado de los talentos emergentes en el fútbol profesional. Esto puede atribuirse a factores como el rendimiento destacado en competiciones nacionales e internacionales, así como a un mayor reconocimiento y demanda por parte de los clubes.

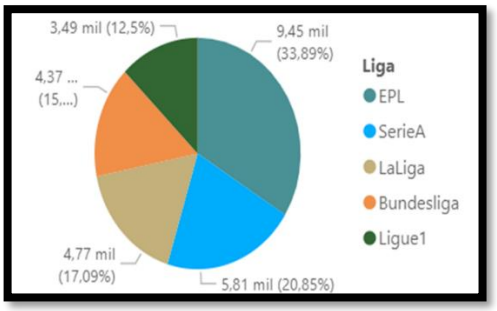
Competencia por talento joven: La inclusión de jugadores como Haaland, Mbappé y otros en la lista resalta la intensa competencia entre los clubes por asegurar el talento joven y prometedor. Esto puede conducir a ofertas y transferencias significativas en el mercado de fichajes, ya que los clubes

buscan garantizar el futuro éxito deportivo y financiero a través de la adquisición de jugadores clave.

Diversidad geográfica: La lista muestra una diversidad geográfica con jugadores de diferentes nacionalidades, lo que refleja la globalización del mercado futbolístico y la búsqueda de talento en todo el mundo. Esta diversidad también puede indicar la creciente importancia de ligas y clubes de diferentes países en el panorama futbolístico global.

Cuarta visualización:

Porcentaje de precio de las 5 grandes ligas:



Premier League domina en gastos: La Premier League lidera en términos de porcentaje de gastos, representando aproximadamente el 33.89% del gasto total en jugadores entre las ligas europeas consideradas. Esto refleja la posición financiera sólida y la alta competitividad de la Premier League en el mercado de fichajes, con los clubes ingleses dispuestos a invertir considerablemente en la adquisición de talento futbolístico.

Serie A y La Liga siguen de cerca: La Serie A y La Liga también tienen una participación significativa en los gastos, representando aproximadamente el 20.85% y el 17.09% respectivamente. Esto demuestra que tanto Italia como España mantienen una presencia destacada en el mercado de fichajes europeo, con los clubes de Serie A y La Liga realizando inversiones considerables para fortalecer sus equipos y competir a nivel nacional e internacional.

Bundesliga y Ligue 1 también tienen una participación importante: Aunque en menor medida en comparación con las ligas mencionadas anteriormente, la Bundesliga y la Ligue 1 también contribuyen significativamente al gasto total en jugadores, representando aproximadamente el 15.67% y el 12.5% respectivamente. Esto muestra que Alemania y Francia también son actores importantes en el mercado de fichajes europeo, con los clubes de la Bundesliga y la Ligue 1 invirtiendo recursos en la adquisición de talento futbolístico.

Resultados obtenidos:

Relación entre precio de los jugadores y éxito deportivo:

Existe una correlación entre los precios de los jugadores y el éxito deportivo, aunque no es necesariamente causal. Los equipos con mayores inversiones en jugadores suelen tener más recursos para construir equipos competitivos y aspirar a logros deportivos.

Sin embargo, el éxito deportivo no está determinado únicamente por el precio de los jugadores. La gestión del equipo, la estrategia táctica, la cohesión del equipo y otros factores también desempeñan un papel crucial en los resultados deportivos.

Distribución geográfica de precios de jugadores y éxito deportivo:

Los precios de los jugadores están significativamente influenciados por la región geográfica. Por ejemplo, Europa, Argentina y Brasil son regiones destacadas en términos de precios de jugadores, lo que refleja la abundancia de talento futbolístico en estas áreas.

Además, estas regiones también suelen tener un historial exitoso en el fútbol internacional, con equipos y selecciones nacionales que han logrado grandes éxitos en competiciones tanto a nivel de clubes como a nivel internacional.

Inversión en talento joven y potencial:

La presencia de jóvenes talentosos como Haaland, Mbappé, Bellingham, Vinicius Jr. y Saka entre los jugadores más caros del mundo sugiere una tendencia hacia la inversión en talento joven y potencial.

Los clubes están dispuestos a invertir en jugadores jóvenes con el objetivo de asegurar el éxito deportivo a largo plazo y potencialmente obtener beneficios financieros significativos a medida que estos jugadores desarrollan su carrera y aumentan su valor en el mercado.

Competencia financiera y deportiva en diferentes regiones:

La distribución geográfica de los precios de los jugadores también refleja la competencia financiera y deportiva en diferentes regiones del mundo. Por ejemplo, la Premier League lidera en términos de gastos en jugadores, lo que coincide con su posición como una de las ligas más competitivas y financieramente poderosas del mundo.

La competencia entre las ligas y los clubes por el talento futbolístico puede influir en la distribución de los precios de los jugadores y, en última instancia, en el éxito deportivo tanto a nivel nacional como internacional.

En resumen, la relación entre el precio de los jugadores, el éxito deportivo y la región es compleja y multifacética, influenciada por una variedad de factores que incluyen la inversión financiera, la distribución geográfica del talento futbolístico y la competencia entre ligas y clubes.

c. Shamyr: RELACIÓN DE FACTORES ASOCIADOS A LOS VIDEOJUEGOS

En la industria de los videojuegos el rating, los precios y las reseñas juegan un papel crucial sobre el éxito del videojuego. Este caso de estudio examina como estos factores interactúan y como a lo largo del tiempo el desempeño en el desarrollo de videojuegos ha mejorado en el mercado.

Objetivo general:

Analizar la relación entre el rating, los precios y las reseñas sobre los videojuegos.

Objetivos específicos:

- Examinar el rating de los videojuegos por años.
- Comparar el máximo de precios en años.
- Comparar los precios entre los mejores y peores videojuegos.
- Identificar los juegos con más reseñas.
- Identificar el sistema operativo más usado en videojuegos.

Metodología:

Los datos sin procesar de este caso de uso fueron adquiridos mediante un dataset de la plataforma Kaggle y mediante web scrapping directamente en la plataforma de videojuegos de Steam. A este conjunto de datos se le realizo una limpieza y equivalencia de datos en Jupyter usando la biblioteca de Pandas, también se usó la biblioteca Selenium se imoporto un webdriver el cual fue necesario al momento de realizar el webscrapping ya que Steam recarga la página automáticamente al bajar en busca de más resultados, este driver nos permite definir una cantidad de veces que queremos se realice estas cargas [1].

Finalmente, se concateno el dataset de Kaggle y el de webscraping como se muestra en la figura Figura 1.c.1 una vez limpios se cargaron a la base de datos en MySQL y se extrajeron para así cargarlos en DataLake en MongoDB ATLAS.

El script se encuentra en el repositorio de GitHub: https://github.com/ShamyrQuishpe/protecto_analisis

	app_id	title	date_release	win	mac	linux	rating	positive_ratio	user_reviews	price_final
0	13500	Prince of Persia: Warrior Within™	2008	True	False	False	Very Positive	84	2199	9.99
1	22364	BRINK: Agents of Change	2011	True	False	False	Positive	85	21	2.99
2	113020	Monaco: What's Yours Is Mine	2013	True	True	True	Very Positive	92	3722	14.99
3	226560	Escape Dead Island	2014	True	False	False	Mixed	61	873	14.99
4	249050	Dungeon of the ENDLESS™	2014	True	True	False	Very Positive	88	8784	11.99
...	...	...	...	...	...	...	...	...	...	...
52067	2754340	Lagoon Lounge 2 : The Secret Roommate	2024	True	False	False	Very Positive	98	73	7.79
52068	762174	Monster Hunter: World - Gesto: pistolas dobles...	2018	True	False	False	Very Positive	100	95	3.99
52069	886900	Chef: A Restaurant Tycoon Game	2020	True	False	False	Mostly Positive	73	1220	19.99
52070	2383760	Monopoly Madness	2023	True	False	False	Mixed	41	73	23.99
52071	2485180	Cats Hidden in Georgia	2024	True	False	False	Very Positive	91	236	0.89

Figura 1.c.1. Dataset Videojuegos

Análisis:

Primer análisis:

En el primer análisis podemos visualizar un gráfico de barras que denota como ha aumentado la cantidad de rating entre negativas, positivas, muy positivas, muy negativas, mixtas, extremadamente positivas y negativas entre el 2000 al 2024 como se muestra en la figura.

Además, tenemos otra comparativa entre los mismos años, pero comparado con un gráfico de la suma de user reviews, como se muestra en la figura.

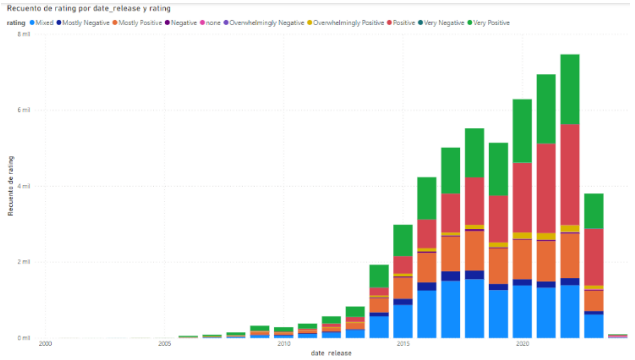


Figura 1.c.2. Rating por año (2000-2024)

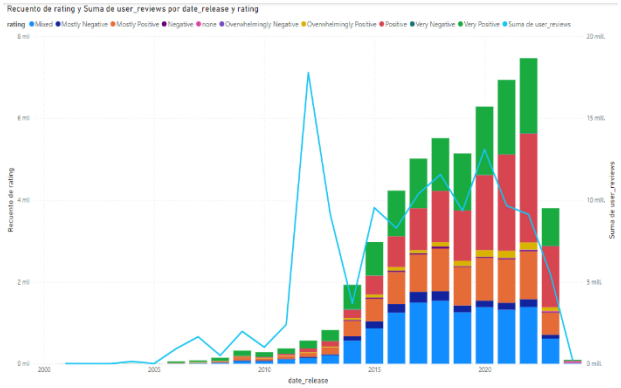


Figura 1.c.3. Rating por año (2000-2024) comparado con cantidad de reseñas

Segundo análisis:

Para el segundo análisis relacionamos los años del 2000 al 2024 con el máximo precio de los videojuegos existentes por cada año como se muestra en la figura.

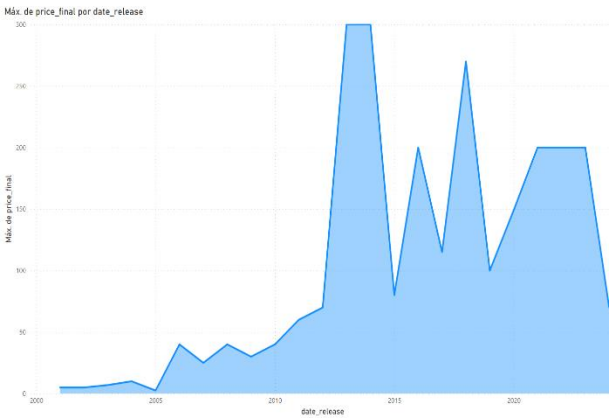


Figura c.4. Máximo precios por año

Tercer análisis:

En cuanto a los juegos con más reseñas relacionamos los juegos con la mayor cantidad de reseñas realizadas por los usuarios y los títulos respectivos y realizamos un top 12 como se muestra en la figura .

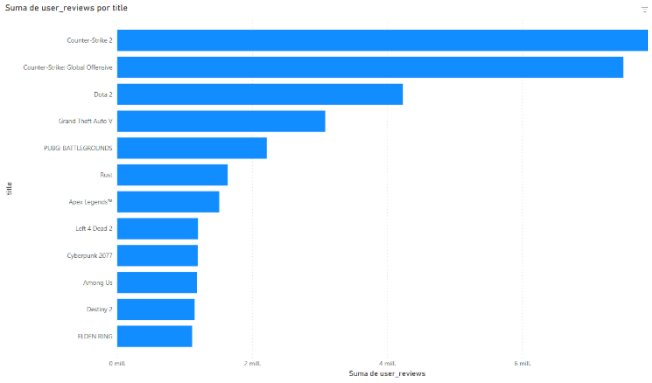


Figura c.5. Reseñas de usuarios por título (Top 12)

Cuarto análisis:

Se genero un top de juegos mediante un rating de clasificación como “Extremadamente positivo” y su comparación con el precio máximo entre los años 2021 al 2023 como se observa en la figura.

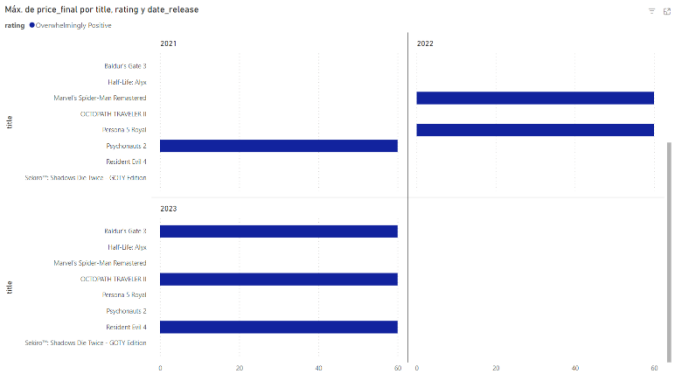


Figura c.6. Top mejores juegos según el rating y el precio máximo

Quinto análisis:

Se genero un top de juegos con un rating de clasificación como “Extremadamente negativo” y su comparación con los precios máximos entre los años 2019 al 2024 como se observa en la figura .

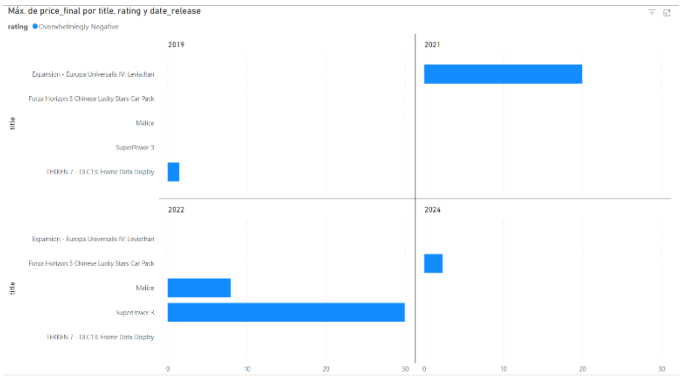


Figura c.7. Top peores juegos según el rating y el precio máximo

Sexto análisis:

Ahora para poder verificar los sistemas operativos en los que más existen juegos y usuarios, entre los sistemas operativos tenemos a “Windows”, “MacOS” y “Linux”, con gráficos de pastel de verdadero y falso donde si existen juegos para cada sistema operativo y la métrica es la suma de reseñas existentes como se muestra en las figuras.

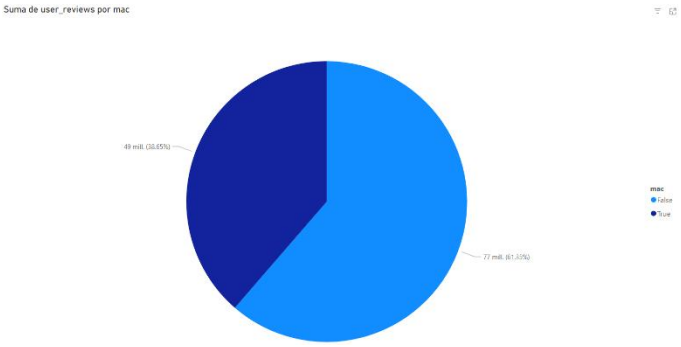


Figura c.8. Reseñas de usuarios en MacOS

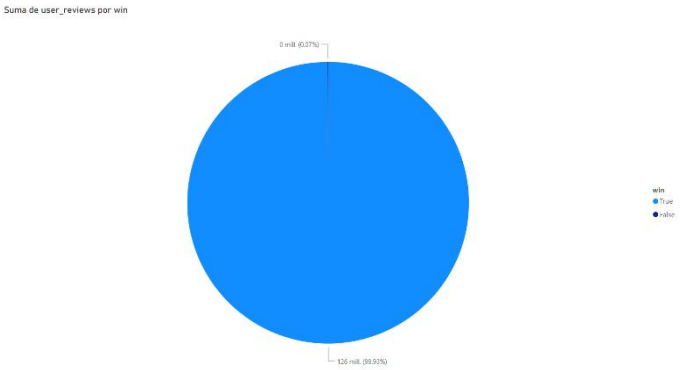


Figura c.9. Reseñas de usuarios en Windows

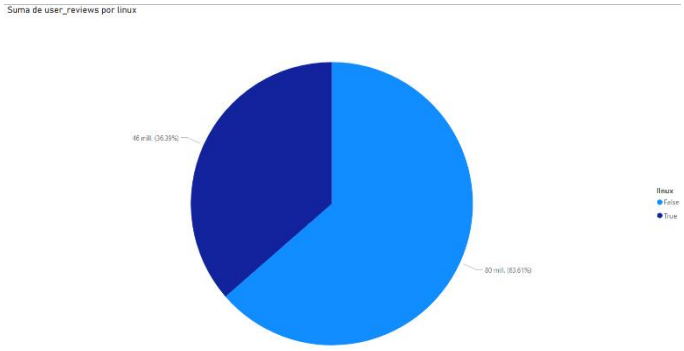


Figura c.10. Reseñas de usuarios en Linux

Resultados obtenidos:

Primer análisis:

En el primer análisis podemos visualizar un gráfico de barras que denota como ha aumentado la cantidad de rating los últimos años, pero llega a un pico en conteo de rating en el año 2022. En ese año la industria de videojuegos empezó a generar más ganancias todo justo en la pandemia de Covid-19 se estima que solo esta industria recaudo 180 millones de dólares esto justifica el aumento en la cantidad de rating de dicho año [3].

En cuanto a la segunda grafica al combinarla con la suma de reseñas por año podemos notar que hay un pico en el 2012 con una cantidad de reseñas de 17784205, esto se debe a que en dicho año las ventas de ordenadores dedicados al gaming aumento nueve veces sus ventas también esto debido al crecimiento que empezó a tener los ESports o deportes electrónicos [4].

Segundo análisis:

Podemos visualizar que del 2000 al 2024 entre 2013 a 2014 existen juegos con un máximo precio de $299.99 y en la actualidad se han mantenido en un precio estándar de entre $199.99 a $69, sin embargo, los precios en los primeros años no eran tan altos si tenemos en cuenta los años 2000 al 2010 los precios se mantenían por debajo de los $40, esto debido a diferentes ediciones de juegos que aumentan su precio con el contenido descargable adicional que ofrecen por más horas de juego [5].

Tercer análisis:

En cuanto a los juegos con más reseñas podemos ver un top 12 de los juegos más aceptados por la crítica según la cantidad de reseñas hechas por los usuarios, podemos visualizar que los 3 primeros son “Counter Stike 2”, “Counter Strike Global Ofensive” y “Dota 2”. Con unas cantidades de reseñas muy altas, 7925413, 7494460 y 4231951 respectivamente a los títulos ofrecidos.

Cuarto análisis:

Al analizar el top de juegos con un rating de clasificación como “Extremadamente positivo” y su comparación con precios entre los años 2021 al 2023, los mejores juegos mantienen el precio de $60 y 2 de ellos fueron nominados a los GOTY 2023 que son “Baldur´s Gate 3” y “Resident Evil 4” [2].

Quinto análisis:

Al analizar el top de juegos con un rating de clasificación como “Extremadamente negativo” y su comparación con precios entre los años 2019 al 2024, donde el máximo precio al que llegan es de $29.99, en su mayoría son DLC’s o contenido descargable los que tienen las peores reseñas, esto debido a que divide el juego en múltiples partes y llega a ser muy criticado por los usuarios [6].

Sexto análisis:

Podemos visualizar que el sistema operativo más usado para gaming por los usuarios es “Windows”, sin embargo, existe gran aceptación de los demás sistemas para poder disfrutar de los videojuegos. Windows a recibido gran aceptación y un aumento de usuarios del mismo desde el 2018 mayor al 50% de usuarios que ocupan este sistema operativo, además los demás sistemas tales como “MacOS” y “Linux” también han ido aumentando en un 0.22% [7].

Conclusiones:

En conclusión, el análisis detallado de diversos aspectos de la industria de los videojuegos revela una serie de tendencias y patrones significativos. El aumento en la cantidad de ratings en 2022, coincidiendo con el auge de la industria durante la pandemia de COVID-19, sugiere una correlación entre eventos externos y el compromiso de los usuarios con los videojuegos. Además, la relación entre las ventas de hardware para gaming y la cantidad de reseñas destaca la influencia de los eSports y el crecimiento general del sector en la actividad de los jugadores.

La evolución de los precios de los juegos a lo largo del tiempo, especialmente en relación con el contenido descargable, refleja una adaptación continua de los desarrolladores para satisfacer las demandas del mercado y maximizar los ingresos. Asimismo, el análisis de los juegos mejor y peor valorados junto con sus precios ofrece información valiosa sobre las preferencias y expectativas de los consumidores, así como sobre las estrategias de precios que pueden influir en la recepción de un juego.

Finalmente, la preferencia por el sistema operativo Windows para gaming, aunque predominante, no es excluyente, ya que

otros sistemas como MacOS y Linux también han experimentado un crecimiento en su base de usuarios. Este fenómeno sugiere una mayor apertura y accesibilidad en la industria de los videojuegos, donde los desarrolladores pueden encontrar oportunidades para expandir su alcance a través de múltiples plataformas. En conjunto, estos análisis proporcionan una visión integral de la dinámica y evolución de la industria de los videojuegos, subrayando su complejidad y su continua adaptación a las demandas y preferencias de los jugadores.

d. Francisco Caero: Eventos y Noticias Mundiales

Con una temática establecida de eventos y Noticias Mundiales necesitaba encontrar un área de la que tuviera suficiente información, y pudiera estudiar a fondo. Tras una larga investigación decidí orientar mi estudio a la población mundial y sus variaciones. Fue un tópico del que encontré información confiable y que pudiera relacionar con otras fuentes, para desplegar datos más a detalle. Entonces, mi trabajo iría destinado a visualizar la demografía mundial: su división, su transformación y más características. Me pareció un tema realmente importante porque ver estos datos a través de un largo periodo nos pueden decir muchas cosas, como que existan factores que pueden haber afectado a ciertas regiones y beneficiado a otras. Podemos observar cómo en algunas partes del mundo se concentran gran parte de la población y compararla con el territorio, para conocer su densidad de población. Conocer la diminución y aumento de habitantes permite establecer relaciones y conclusiones que ayudan a localizar problemas o situaciones positivas para lograr superar estos contratiempos y dedicarse a la mejora de condiciones de vida.

Objetivo general:

Analizar la relación entre el rating, los precios y las reseñas sobre los videojuegos.

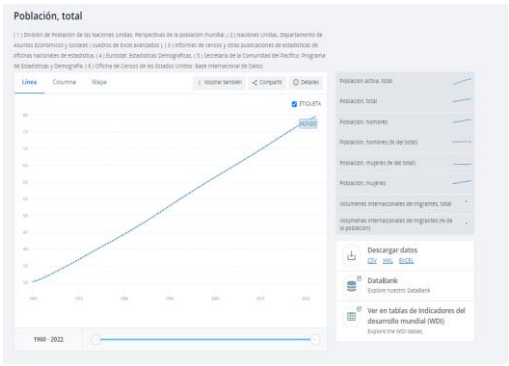
Objetivos específicos:

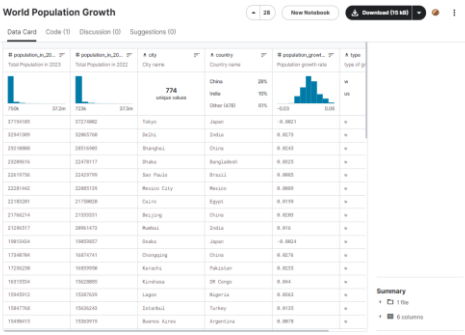
- Examinar el rating de los videojuegos por años.
- Comparar el máximo de precios en años.
- Comparar los precios entre los mejores y peores videojuegos.
- Identificar los juegos con más reseñas.
- Identificar el sistema operativo más usado en videojuegos.

Metodología:

Este tema fue elegido después de una larga indagación: por páginas y fuentes distintas, buscando datos confiables que pudieran ser comparados y analizados juntos. Se encontró varios sitios que me podrían servir, como al siguiente web de noticias; pero al no disponer de información similar, continué buscando.

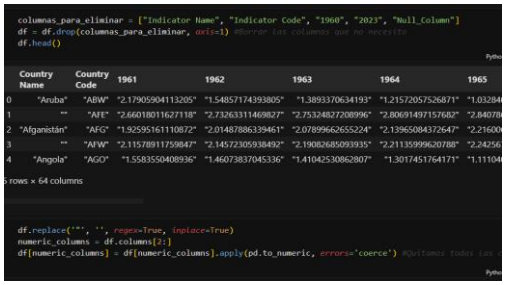
Finalmente se encontró, en el banco de datos mundial un indicador detallado: crecimiento poblacional en el mundo, me pareció interesante y viable, mi otra fuente de información sería Kaggle, un gran repositorio de datos, que tras una buena investigación se halló un csv que trataba la misma temática.

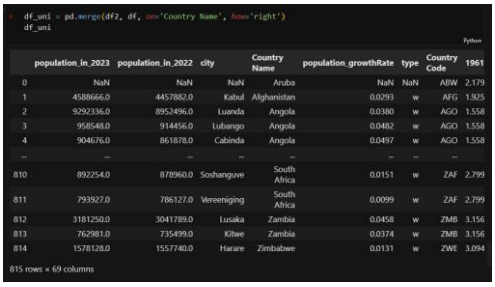




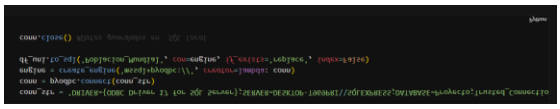
A pesar de su parecido, no sería tan fácil la concatenación, pues sí hizo falta un buen proceso previo, para poder unir la información de las diferentes fuentes. Incluso tuve dificultades a la hora de subir el primero y tuve que buscar la manera de separarlo por líneas, borrando algunas que no necesitara.

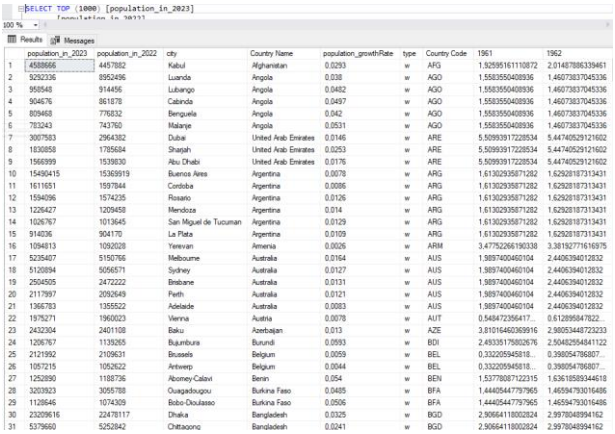
Una vez se terminó de subir ese y el segundo dataset, se procedió al procesamiento de estos, el cual se basó principalmente en cambiar formato, borrar columnas inservibles, y algo muy importante fue, traducir una columna que contenía los mismos datos, pero en diferente idioma, esta parte fue la clave para concatenar la información.



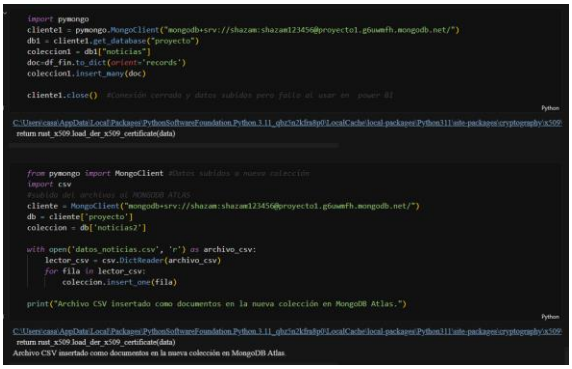


La siguiente etapa, con el debido proceso es mediante un ODBC mandar los datos a mi base de datos designada, en mi caso fue a SQLServer.





Para terminar con la subida de datos, se tomó la información ya subida a esa base de datos y se agregó a la data lake que establecimos para el grupo en MongoDb Atlas. Lo siguiente solo sería tomar es información para hacer un dashboard en Power BI.



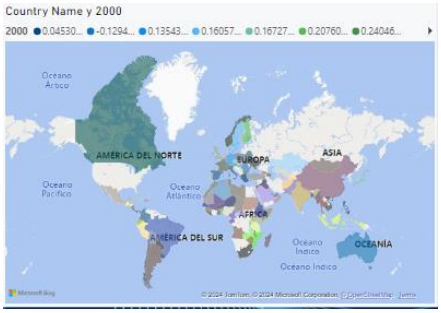
Repositorio con todo el proceso: https://github.com/franciscocaero/Proyecto_Noticias_Poblacion

Limitaciones y consideraciones:

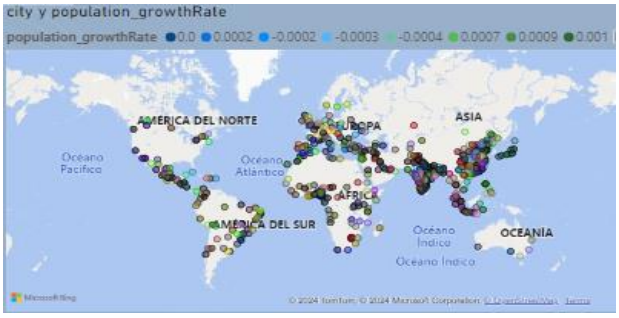
Los mayores problemas que enfrenté como expliqué un poco más arriba fue para subir el primer dataset a pandas, ya que me daba errores, y tuve que buscar bien la manera, por lo tanto, es importante tener en cuenta el formato y la configuración del archivo, para evitar complicaciones como en este caso.

La segunda gran problemática, también tuvo que ver un poco con el formato del archivo, se intentó varias veces antes de poder crear un dashboard usando los datos importados en Mongo. En conclusión, es importante tener en cuenta el método como se va a subir los archivos porque pueden afectar la integridad.

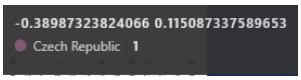
Análisis de datos:

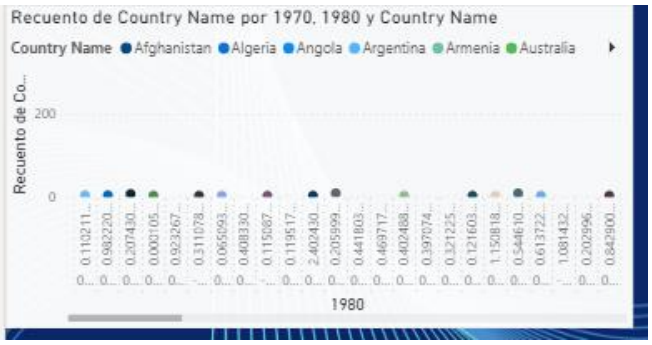


Por la temática del estudio, me centré en usar principalmente mapas y localizaciones, en la primera quise crear elementos que fueran más interactivos, que me mostraran datos según quisieras ver, ya que son regiones muy extensas. Y puedan ser distinguidas mediante colores, en el mapa veremos que aquellos valores más bajos serán los que tengan alguna tonalidad azul clara, siendo la mayoría de estos países europeos. Esto quiere decir que países como Suecia, Rumanía y otros de la zona fueron los que menos crecieron por la época. Investigando un poco más, por aquella época sucedieron cambios importantes que realmente pueden haber afectado a estos ciertos países.

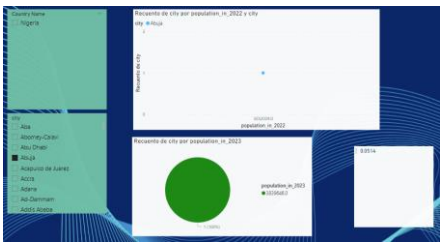


En la segunda pantalla, traté el tema desde la tasa de crecimiento por ciudad entre 2022 y 2023, las que menos crecimiento tuvieron, fueron varias al este de Asia, unas de Centroamérica y ciertas ciudades de Europa: podemos mencionar, Managua Belgrado, Atenas, Tianshui. Aunque sacando un promedio, en general, la tasa de crecimiento por ciudad no es realmente abrumadora.

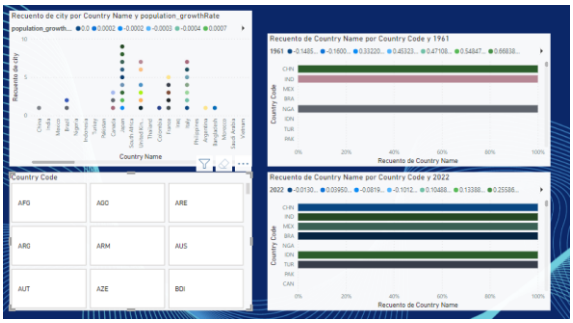




Agrupando estos datos países pude realizar una comparación del crecimiento, con 10 años de diferencia, como en este caso República Checa que logró levantarse de un mal momento tras guerras y crisis de la zona.



Esta es toda una pantalla interactiva entre sí, usando los mismos elementos, se nos permite ver a detalle en forma de distintos gráficos, la población en 2022, en 2023 y cuál fue su tasa de crecimiento.



Como última pantalla me pareció interesante detallar, me parece importante hablar de esto, analizando año por año, ya que es ahí donde poco a poco realmente veremos una diferencia, porque entre estos dos años, ha habido y sigue habiendo un crecimiento exponencial de la población mundial inmenso, para hablar de porcentajes, deberíamos hablar también de población, no es lo mismo crecer 1.2 en 1960 que hacerlo en 2022, con más razón se evidencia que aquellos países que han tenido un buen porcentaje de crecimiento a lo

largo de todos estos años, han tenido un desarrollo en otras áreas.

Resultados Obtenidos:

Gracias a la forma detallada que nos proporcionan etas herramientas, podemos ubicar puntos temporales o geográficos en los que se pierde la curva de crecimiento en algún país, gracias a las noticias y otras fuentes podemos encontrar razones que ayuden a explicar a qué se deben estos valores negativos, como en el caso de países que han vivido regímenes políticos difíciles, crisis financieras o simplemente desastres, probándose así que estos problemas afectan a cada sector de la población y a todos los ámbitos. Aunque hay que tener cuidado, no tomar literalmente el porcentaje de crecimiento como un valor absoluto, ya que la población nunca lo es, menos en una diferencia temporal tan grande como la que se estudia aquí, el crecimiento va de acuerdo a la población por lo tanto no sería correcto decir que un 1% pueda ser menor que un 2% en un mundo cada día más poblado, por eso es importante observar bien las variables, antes de realizar ciertas conjeturas, la demografía es algo que puede ayudar a tratar muchos problemas si es observada cuidadosamente.

e. Cristian Usiña: películas

El séptimo arte es una de las industrias más grandes que hay en la actualidad, y los datos que se trafican en todo este universo son demasiados es decir para el análisis de datos es un gran campo para procesar, en este caso vamos a incursionar en los premios más famosos que se otorgan por este maravilloso tema, los premios de la “Academia” o premios “Oscar”, la base del tema son las nominaciones y los premios ganados por cada película, desde el año 1980 hasta el 2021.

Objetivo general:

Conocer mediante el procesamiento de datos y la visualización gráfica cuantas nominaciones y cuantos premios han ganado las películas desde 1980 hasta 2021.

Objetivos específicos:

- Analizar cuáles son las películas con más nominaciones a estos premios.
- Conocer cuál es la película con más premios ganados.
- Encontrar un modelo visual que me permita mostrar el país en donde fueron filmadas estas películas.
- Identificar las películas con más nominaciones y más premios.
- Analizar en que año se llevó una película más premios y más nominaciones.

Metodología:

Recolectar un conjunto de datos que incluya información detallada sobre cada ceremonia de los Premios Oscar, incluyendo las películas nominadas, los premios, los ganadores y otros datos relevantes. Luego, dedico tiempo a explorar estos datos para comprender su estructura y contenido.
Hay que revisar de que los datos estén limpios y consistentes, corrigiendo cualquier error de entrada, eliminando valores atípicos y manejando los valores faltantes de manera adecuada. Esto es crucial para garantizar la precisión de mis análisis posteriores. [1]
El script se encuentra en el repositorio de GitHub:
https://github.com/VnCriis/proyectoAnalisisPeliculas.git

id	year	pelicula	premios	nominaciones	pais
1	1980	Ordinary People	4	6	Estados Unidos
2	1980	Raging Bull	2	8	Estados Unidos
3	1980	Tess	3	6	Reino Unido
4	1980	Coal Miner's Daughter	1	7	Estados Unidos
5	1980	The Elephant Man	0	8	Reino Unido
...	...	...	...	...	...
531	2021	El agente topo	0	2	Chile
532	2021	Another Round	1	2	Dinamarca
533	2021	Emma	0	2	Reino Unido
534	2021	Collective	0	2	Rumania
535	2021	Una joven prometedora	1	5	Reino Unido

536 rows × 5 columns

Figura 1.e.1. Dataset Películas

Análisis:

Primer análisis:

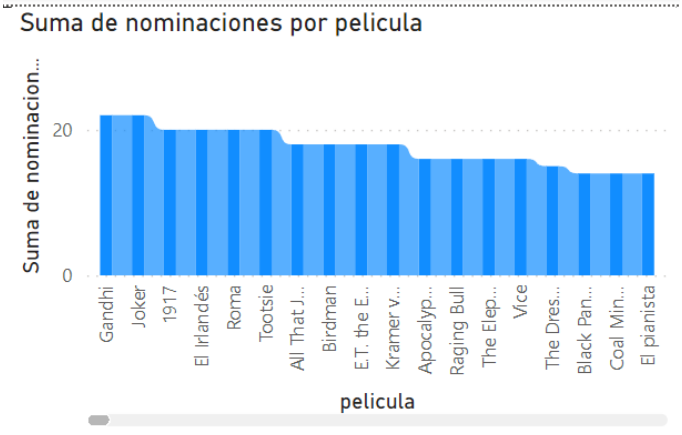


Figura 1.e.2. Nominaciones de películas desde la que tiene más de estas.

Segundo análisis:

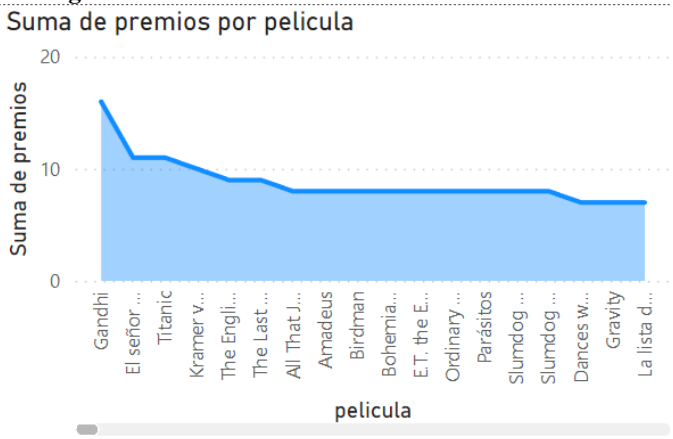


Figura e.3. Premios ganados por películas desde la que tiene más de estas.

Tercer análisis:

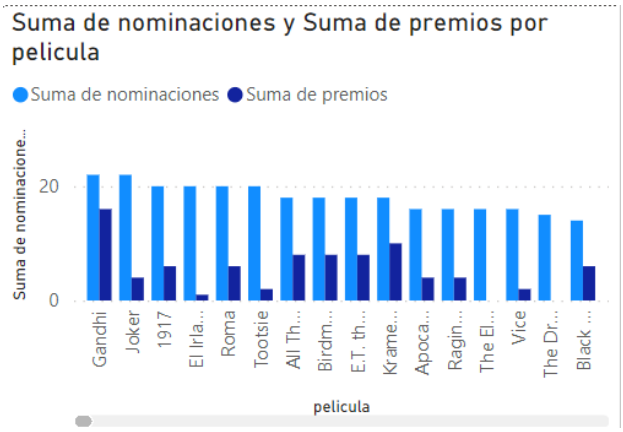


Figura e.4. Suma de nominaciones y premios ganados por película.

Cuarto análisis:

Figura e.5. Países en donde fueron filmadas las películas nominadas.

Quinto análisis:

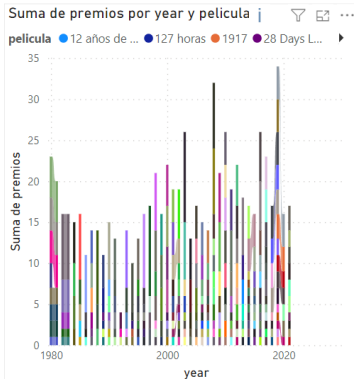


Figura e.6. suma de todos los premios por película en cada año.

Resultados obtenidos

Primer análisis:

Al examinar los datos de las nominaciones a lo largo de los años en los Premios Oscar, podemos observar algunas tendencias interesantes. Por ejemplo, en la década de 1990, películas como 'Titanic' y 'Shakespeare in Love' destacaron con múltiples nominaciones en sus respectivos años. Esto sugiere un período en el que las películas épicas y los dramas históricos capturaron la atención tanto de la audiencia como de la academia.

En contraste, en la década de 2000, vemos un aumento en la diversidad de géneros representados entre las películas con más nominaciones. Por ejemplo, mientras que películas como 'The Lord of the Rings: The Return of the King' continúan con la tradición de las películas épicas, también vemos comedias como 'Chicago' y películas de ciencia ficción como 'Avatar' recibiendo múltiples nominaciones. Esto refleja una mayor apertura por parte de la academia hacia una variedad de géneros cinematográficos.

Sin embargo, la ganadora en este caso es la película ‘Gandhi’ que se acerca a las 20 nominaciones, al igual que ‘Joker’ superando por mucho a sus seguidores.
En general, este análisis nos ofrece una visión fascinante de la evolución de las películas con más nominaciones en los Premios Oscar a lo largo del tiempo, y nos ayuda a comprender mejor las tendencias y los cambios en la industria cinematográfica.

Segundo análisis:
Desde 1980, los Premios Oscar han servido como un barómetro de la evolución y diversidad en la industria cinematográfica. Durante este período, hemos presenciado una amplia gama de películas ganadoras que reflejan diferentes géneros, estilos y temas. En la década de 1980, hubo una variedad notable, desde dramas históricos como 'Gandhi' hasta películas biográficas como 'Amadeus', demostrando una apertura hacia diferentes estilos cinematográficos.
A medida que avanzamos hacia los años 90, la diversidad en las películas ganadoras se hizo aún más evidente. Películas como 'Rain Man' (1988) y 'Dances with Wolves' (1990) abordaron temas sociales y culturales de manera profunda, mientras que 'The Silence of the Lambs' (1991) destacó en el género del thriller, mostrando una amplia variedad de preferencias y gustos dentro de la academia.
En las últimas décadas, hemos visto un enfoque renovado en la inclusión y la diversidad en las películas galardonadas. Títulos como '12 Years a Slave' (2013) y 'Moonlight' (2016) han sido aclamados por su representación auténtica y poderosa de experiencias subrepresentadas en la industria cinematográfica. Estas películas no solo han sido reconocidas por su excelencia artística, sino que también han marcado un cambio significativo en las preferencias y valores de la audiencia y la academia.

Tercer análisis:

En la década de 1980, películas como 'Gandhi' y 'Amadeus' no solo acumularon un alto número de nominaciones, sino que también se llevaron a casa múltiples premios, demostrando su destacada contribución a la cinematografía de la época. Esto sugiere un período en el que las películas no solo eran reconocidas por su excelencia técnica y artística, sino también por su éxito comercial y crítico.
A medida que avanzamos hacia los años 90, películas como 'The Silence of the Lambs' y 'Schindler's List' continuaron esta tendencia, recibiendo numerosas nominaciones y premios por su destacada calidad cinematográfica y su impacto cultural. Estas películas no solo fueron aclamadas por la crítica, sino que también resonaron con el público, lo que las convirtió en referentes en la industria del cine.
En las últimas décadas, hemos visto un aumento en la diversidad de películas que han logrado tanto nominaciones como premios. Títulos como 'The Lord of the Rings: The Return of the King' (2003), 'Slumdog Millionaire' (2008) y 'The Shape of Water' (2017) no solo recibieron numerosas nominaciones, sino que también se llevaron múltiples premios, destacando su impacto y reconocimiento en la industria cinematográfica contemporánea.
Cuarto análisis:

En la década de 1980, Estados Unidos continuó siendo un epicentro importante de la producción cinematográfica, con Hollywood liderando la carga en la creación de películas galardonadas. Sin embargo, también vimos películas de otros países, como el Reino Unido con 'Chariots of Fire' (1981), que mostró la diversidad geográfica de las producciones premiadas.
A medida que avanzamos hacia los años 90 y el nuevo milenio, vimos un aumento en la globalización de la industria cinematográfica, con películas ganadoras provenientes de una variedad de países de todo el mundo. Desde el Reino Unido con 'Shakespeare in Love' (1998) hasta Nueva Zelanda con 'The Lord of the Rings: The Return of the King' (2003), y México con 'Birdman' (2014), la diversidad de países en donde se filmaron películas ganadoras refleja la creciente influencia global en el cine.
En las últimas décadas, hemos visto una mayor representación de películas de países fuera de los tradicionales centros de producción cinematográfica. Por ejemplo, películas como 'Slumdog Millionaire' (2008), filmada en India, y 'Parasite' (2019), filmada en Corea del Sur, han recibido reconocimiento internacional y han demostrado que el talento y la creatividad no están limitados por las fronteras geográficas.
Quinto análisis:

En el nuevo milenio, hemos presenciado una mayor diversidad en las películas ganadoras, con éxitos de taquilla y películas independientes recibiendo reconocimiento por igual. Películas como 'The Lord of the Rings: The Return of the King' y 'Slumdog Millionaire' se llevaron múltiples premios en sus respectivos años, demostrando la amplia gama de historias y estilos que han sido celebrados en la industria cinematográfica en las últimas décadas.
En los últimos años, hemos visto un énfasis renovado en la inclusión y la diversidad en los premios ganados, con películas como '12 Years a Slave' y 'Moonlight' recibiendo elogios por su representación auténtica de experiencias subrepresentadas. Esto refleja un cambio significativo en la industria cinematográfica hacia una mayor sensibilidad y conciencia de la diversidad de voces y perspectivas en el cine.
En resumen, el análisis de los premios ganados en cada año en los Premios Oscar desde 1980 destaca la diversidad y la evolución de la industria cinematográfica, así como los valores y las preferencias que han guiado el reconocimiento de la excelencia en el cine a lo largo de las décadas.

Conclusiones:

Diversidad de películas premiadas: El análisis de datos revela una diversidad significativa en las películas premiadas en los

Premios Oscar desde 1980. Esto se refleja en la variedad de géneros, temas y estilos cinematográficos que han sido reconocidos, lo que indica una industria cinematográfica dinámica y en constante evolución.

Globalización de la industria cinematográfica: La ubicación de filmación de las películas premiadas muestra una tendencia hacia la globalización de la industria cinematográfica. A lo largo de las décadas, hemos visto películas ganadoras provenientes de una amplia gama de países, lo que sugiere una mayor diversidad cultural y geográfica en la producción cinematográfica.

Énfasis en la inclusión y la diversidad: El análisis de datos también destaca un énfasis creciente en la inclusión y la diversidad en los premios ganados en los últimos años. Películas como '12 Years a Slave' y 'Moonlight' han sido aclamadas por su representación auténtica de experiencias subrepresentadas, lo que refleja un cambio significativo en la sensibilidad y la conciencia de la diversidad en la industria cinematográfica. [2]

VIII. CONCLUSIONES Y RECOMENDACIONES

a. Conclusiones:

Durante el desarrollo de este proyecto, hemos logrado diseñar e implementar una arquitectura de Data Lake que integra datos de diversas fuentes, incluyendo redes sociales, portales de datos abiertos y archivos estáticos. Esta arquitectura nos ha permitido almacenar, procesar y analizar grandes volúmenes de datos de manera eficiente y escalable.

Gracias a la variedad de herramientas y recursos utilizados, hemos podido abordar con éxito los diferentes desafíos del proyecto, desde la extracción y transformación de datos hasta la visualización de resultados. La combinación de bases de datos relacionales y NoSQL, junto con herramientas como Jupyter Notebook y Power BI, ha demostrado ser efectiva para el análisis y la generación de insights.

Los resultados obtenidos de los cinco casos de estudio seleccionados proporcionan información valiosa sobre temas como el pulso político, los juegos en línea, eventos mundiales y la ciencia en Ecuador. Las visualizaciones generadas nos han permitido identificar tendencias y patrones significativos en los datos, proporcionando una comprensión más profunda de los fenómenos analizados.

b. Recomendaciones:

Para futuros proyectos similares, recomendamos continuar explorando nuevas fuentes de datos y expandir la variedad de herramientas utilizadas. Esto podría incluir la incorporación de tecnologías emergentes como el aprendizaje automático y la inteligencia artificial para mejorar el análisis de datos y la generación de insights.

Se sugiere también realizar un seguimiento continuo de los datos y actualizar regularmente la arquitectura de Data Lake para adaptarse a cambios en las fuentes de datos y las necesidades del negocio. Esto garantizará la relevancia y precisión de los análisis realizados.

Finalmente, recomendamos compartir los resultados obtenidos con las partes interesadas relevantes y utilizarlos como base para la toma de decisiones estratégicas. La información generada a partir de este proyecto puede ser útil para informar políticas públicas, estrategias de marketing y otras iniciativas empresariales.

IX. DESAFIOS Y PROBLEMAS ENCONTRADOS

- Durante el proceso de concatenación de columnas, observamos una pérdida significativa de registros, lo que potencialmente limitó la amplitud de nuestro caso de estudio. Esta limitación puede haber surgido debido a la presencia de valores nulos o inconsistencias en los datos originales, lo que requiere una mayor atención en futuros procesos de manipulación de datos para minimizar la pérdida de información.

- Otro desafío al que nos enfrentamos fue al intentar cargar un archivo CSV con 20,000 datos en la base de datos de Google. En este caso, nos encontramos con la limitación de la capacidad de la base de datos, que no pudo manejar la cantidad masiva de información. Esto subraya la importancia de evaluar cuidadosamente los límites de capacidad y considerar alternativas de almacenamiento más robustas para conjuntos de datos de gran tamaño.

- Asimismo, al cargar la información en MongoDB, nos encontramos con la necesidad de realizar primero una transformación a un formato CSV antes de subir los datos directamente. Este paso adicional fue necesario debido a problemas de compatibilidad con Power BI al cargar los datos directamente desde MongoDB. Esta experiencia destaca la importancia de la interoperabilidad entre diferentes herramientas y plataformas, así como la necesidad de ajustar los procesos de carga para garantizar la integridad y la eficiencia del análisis de datos.

- Uno de los principales desafíos y además de ser un problema fue al momento de realizar el web scrapping debido a que no había una clase en HTML que me

muestre cuando recarga la página, sin embargo, con la biblioteca de selenium fue posible con el webdriver.

- Es necesario tener en cuenta los tipos de datos antes de ser analizados en el powerbi ya que pueden presentar complicaciones, es decir datos numéricos identificados como números porque si están como String se vuelve un problema al analizar los datos.

- El formato de los archivos y los datos que estos manejen deben ser manejados correctamente para trabajar con ellos sin perder información, tener eso en cuenta me complicó un poco más la transformación y subida de los mismos, además de la búsqueda de información verificada que pueda correlacionarse.
- El equipo y la conexión juegan un papel muy importante en el uso de estos datos, sin el cuidado necesario se podría perder información en medio del proceso de la subida, es fundamental un chequeo constante por seguridad.
- El primer desafío fue la búsqueda de la información correspondiente para el respectivo análisis, ya que si bien existe mucha información no toda se encuentra de forma gratuita es por ello que se necesita varias fuentes para la recolección de datos.
- Conocer las funciones y herramientas que nos ayudan al análisis es fundamental ya que como POWER BI existen muchas mas herramientas que nos ayudan a hacer análisis acertados y dar buenas explicaciones para la toma de decisiones.

X. BIBLIOGRAFÍA

[1] “WebDriver”. Selenium. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://www.selenium.dev/documentation/webdriver/

[2] “Todos los nominados de los GOTY 2023: lista completa de The Game Awards”. Meristation. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://as.com/meristation/noticias/todos-los-nominados-de-los-goty-2023-lista-completa-de-the-game-awards-n-2/

[3] “Tema: Industria mundial del videojuego”. Statista. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://es.statista.com/temas/9150/industria-mundial-del-videojuego/#:~:text=Los%20videojuegos%20han%20demostrado%20año,180.000%20millones%20de%20dólares%20estadounidenses.

[4]S. Fernández. ““Desde 2012, las ventas de ordenadores para ‘gaming’ se han multiplicado por nueve””. elconfidencial.com. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://www.elconfidencial.com/tecnologia/2017-04-04/gaming-videojuegos-esports_1351892/

[5]I. Fdez. “Los DLC arruinaron mi tarjeta de crédito: lo que cuesta de verdad un videojuego completo hoy día”. Xataka - Tecnología y gadgets, móviles, informática, electrónica. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://www.xataka.com/videojuegos/los-dlc-arruinaron-mi-tarjeta-de-credito-lo-que-cuesta-de-verdad-un-videojuego-completo-hoy-dia

[6]“¿Qué son y cómo funcionan los DLC de videojuegos? | U-tad”. U-tad. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://u-tad.com/que-son-y-como-funcionan-los-dlc/

[7]G. González. “Más del 50% de los gamers prefieren Windows 10 para jugar, según los últimos datos de Steam”. Genbeta - Software, descargas, aplicaciones web y móvil, desarrollo. Accedido el 4 de marzo de 2024. [En línea]. Disponible: https://www.genbeta.com/windows/mas-del-50-de-los-gamers-prefieren-windows-10-para-jugar-segun-los-ultimos-datos-de-steam

[8]KAGGLE, «KAGGLE,» DATASETS, [En línea]. Available: https://www.kaggle.com/.

[9]GOOGLE, «DATASEARCH,» [En línea]. Available: https://datasetsearch.research.google.com/.

[10]T. Parvez, “Los 10 principales mercados mundiales de música digital,” *SonoSuite*, Apr. 28, 2023. https://sonosuite.com/es/blog/10-principales-mercados-musica-digital-global/

[11] “Industry data - IFPI,” *IFPI*, Jul. 21, 2023. https://www.ifpi.org/our-industry/industry-data/