PROYECTO FINAL ANÁLISIS DE DATOS (Septiembre de 2021)

Jorge W. Alba, Juan C. Bolaños, Juan B. Dávila, Byron S. Huaraca

Resumen –

En el presente documento se detalla el proceso de recopilación, concentración, análisis y visualización de datos de varias temáticas propuestas con el objetivo de demostrar todo el conocimiento relacionado al análisis de datos adquiridos durante el semestre y de esta forma tener bases solidas que puedan ser aplicadas en el mundo laboral.

I. Definición de caso de estudio

Se realizará el estudio a datos actualizados sobre las temáticas de pulso político de Ecuador tanto por ciudades como por provincias, juegos online por países, ranking de felicidad por países y noticias o eventos mundiales. Los datos serán recopilados de diferentes fuentes y almacenadas en diferentes bases de datos tanto SQL como NoSQL y posteriormente la información será analizada y visualizada para proporcionar conclusiones relevantes de la información previamente recopilada.

II. Objetivos

- General

  Recopilar datos de diferentes fuentes con el uso de herramientas y métodos para su respectivo análisis y visualización.

- Específicos

  - Analizar los casos de estudio.
  - Diseñar una arquitectura de solución.
  - Establecer el cronograma de actividades.
  - Escoger las herramientas necesarias.
  - Recopilar datos de fuentes como Facebook, Twitter, TikTok, LinkedIn y con métodos como Webscrapping.
  - Concentrar los datos para realizar su respectivo análisis.
  - Visualizar los datos con herramientas de visualización de datos.
  - Documentar las conclusiones obtenidas del previo análisis de datos.

III. Cronograma de actividades

El cronograma de actividades se lo visualiza en la Figura 1.

Cronograma de Actividades

Encargados	Tareas	Semana 1	Semana 2	Semana 3
• Jorge Alba	**Organización del proyecto**			
• Juan Bolaños	Delegación de Actividades	2		
• Bernabé Dávila	Elección de herramientas	2		
• Byron Huaraca	Estimación de tiempos	1		
• Jorge Alba	**Recolección de datos**			
• Juan Bolaños	Búsqueda de datos		10	
• Bernabé Dávila	Generación de Scripts		15	
• Byron Huaraca	Limpieza de datos		5	
• Jorge Alba	**Concentración de datos**			
• Juan Bolaños	Importación de datos en las bases de datos		8	
• Bernabé Dávila	Concentración en elasticsearch		4	
• Byron Huaraca				
• Jorge Alba	**Análisis y visualización de datos**			
• Juan Bolaños	Importación de datos en herramientas de visualización			10
• Bernabé Dávila	Generación de gráficos			10
• Byron Huaraca				
• Jorge Alba	**Resultados**			
• Juan Bolaños	Interpretación de gráficos			10
• Bernabé Dávila	Documentación			15
• Byron Huaraca				
Total, de horas por semana		5	42	45
Total, de horas	92			

Figura 1 Cronograma

IV. Recursos y herramientas

Las herramientas usadas para el proceso son:

- **Web:** Es la fuente mas grande de datos en la cual se pueden recopilar dato actualizados que son de gran ayuda para cualquier tipo de actividad y sobre todo para el análisis de dato.

- **Python:** Lenguaje de programación interpretado que ofrece varios métodos para el análisis de datos y relacionado a bases de datos.

- **SQL Server:** Es un sistema de gestión enfocado en bases de datos relacionales.

- **MySQL**: Sistema de gestión de bases relacionales de código abierto.

- **SQLite:** herramienta para la gestión de bases de datos relacionales contenida en una pequeña biblioteca que se encuentra escrita en C.

- **MongoDB:** sistema de gestión de bases de datos enfocada en bases de datos no relacionales que almacena en estructuras de datos BSON.

- **CouchDB:** Gestor para bases de datos no relacionales de código abierto enfocada en datos que asume la web de forma completa.

- **Elasticsearch:** Herramienta que ofrece un motor de búsqueda con interfaz RESTful y manejo de documentos JSON.

- **Logstash:** Herramienta que permite recolección, análisis y almacenamiento de logs para búsquedas.

- **Kibana:** Herramienta de visualización de datos que trabaja en conjunto con elasticsearch.

- **Tableau:** Software que perite la visualización de datos interactivos con enfoque en inteligencia empresarial.

- **Power BI:** Herramienta para visualizaciones interactivas y que ofrece capacidades de inteligencia empresarial con interfaz sencilla de manejar.

- **Twitter Developer Count:** Cuenta de Twitter que permite el acceso a partes del servicio ofrecido por medio de APIs que permite realizar la búsqueda de datos.

- **Kaggle:** Plataforma que ofrece recursos para Machine Learning y situaciones relacionadas a la ciencia de datos.

V. Arquitectura de solución

La Arquitectura de la solución se la muestra en la Figura 2.

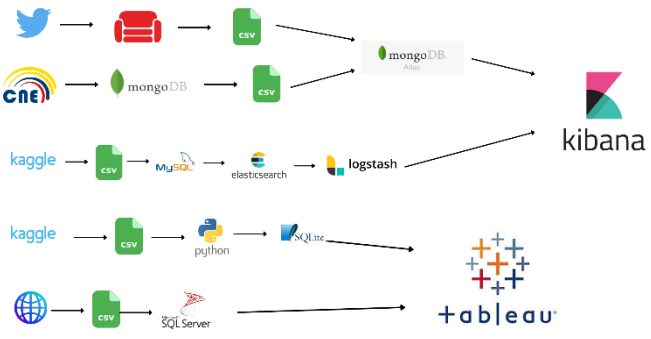

Figura 2 Arquitectura

VI. Extracción de datos

Extracción de datos desde Twitter: La red social Twitter proporciona herramientas a los desarrolladores para facilitar la extracción de datos a través de su aplicación. Entre los principales usos que se la da a la API de Twitter se encuentran:
- Utilizar las potentes API de Twitter para ayudar a una empresa a escuchar, actuar y descubrir.
- Creación de herramientas para que las personas en Twitter integren o mejoren su experiencia en la plataforma.
- Utilizar la API de Twitter para obtener puntos de datos históricos y en tiempo real para su próximo proyecto de investigación [1].

Para poder acceder a esta cuenta de desarrollador, previamente se debe solicitar una cuenta de desarrollador a través de un formulario en el que detallas las intenciones con las que vas a usar su API. De esta forma se aseguran y evitan un mal uso de los datos que proporciona Twitter.
Cuando escribimos un tweet la información que se suele ver es el usuario que lo escribió. Pero la información que no se suele ver es mucho más potente como podemos ver en la Figura 3.

Figura 3 Anatomía de un Tweet

Extracción de datos de la Web: La web proporciona una información inmensa de cualquier temática por lo cual en algún sitio web que muestre información en tablas con la ayuda de Excel permite extraer los datos de dichas tablas y tener los datos disponibles para ser almacenados en una base de datos y posterior a ello ser analizados y presentados en gráficos. Por lo cual los datos sobre gamers por país y gamers de COD son fáciles de reconocer y entender las tendencias presentadas [2, 3].

Los datos fueron extraídos de Kaggle, en este se mostrará el informe mundial sobre la felicidad [4], la cual es una encuesta realizada por las Naciones Unidas que mide cómo ha evolucionado la felicidad de los ciudadanos en los últimos años en 156 países, al descargar de esta página se consigue un archivo .csv. Luego se procedió a utilizar código de Python para poder exportar los datos a SQLite y posterior a ello se descargó un complemento para utilizar la herramienta en la que se realizaran las visualizaciones que es Tableau [5].

Extracción de datos a través de Kaggle: Esta página web proporciona los datos que el usuario necesite, en este caso, se

lo hace en formato CSV para realizar la importación de dichos archivos a la base de datos, en este caso "MySQL". De esta forma se facilita la creación de una tabla asignando los títulos de cada columna del CSV como los índices de la tabla [6].

VII. Análisis de información

a. Pulso Político 20 Ciudades más importantes del Ecuador

Se recaudó un total de 1000 tweets aproximadamente, en los cuales podemos clasificar o filtrar con las herramientas la cantidad de menciones que se obtuvo de cada ciudad que analizamos y así darnos cuenta del volumen de interacciones que tuvo cada ciudad en el proceso de elecciones como se muestra en la Figura 4.También podemos filtrar la información de en cuantos de estos Tweets se mencionaba a Guillermo Lasso para darnos Cuenta de que la gran mayoría de los Tweets que recaudamos lo mencionan ya sea positiva o negativamente como se visualiza en la Figura 5.

b. Pulso en Provincias del Ecuador

A través de la página del Consejo Nacional Electoral se recaudó los datos de la primera Vuelta de las Elecciones presidenciales en donde se puede notar el porcentaje de ciudadanos que votaron por cada candidato. También se especifica que estos datos pueden tener un +-3% de error así que los datos no son del todo precisos. Aun así, estos datos nos sirven para analizar porque los 3 candidatos más solicitados fueron Andrés Arauz, Guillermo Lasso, y Yaku Pérez como podemos ver en la Figura 6 [7].
Para visualizar mejor esto lo dividiremos en regiones del Ecuador. Así por ejemplo tenemos los datos de la Costa Figura 7 y podemos ver que ampliamente hubo una ventaja de Andrés Arauz. Por otra parte, en las regiones de Sierra y Amazonía Figura 8 y Figura 9 podemos ver que Yaku Pérez se llevó una cantidad significante de votos. Lo sorprendente es que Guillermo Lasso que finalmente fue el que ganó las elecciones no se mostró fuerte en ninguna región en particular. Aunque a estos datos debe incorporarse el voto de los extranjeros, que no se encuentra en este estudio.

c. Juegos Online por país

Los datos relacionados a los gamers por país muestran que Estados Unidos es el país con mayor índice de jugadores tanto en general. Por otro lado, existen varios países que cuentan con muy pocos gamers como por ejemplo Madagascar, Aruba o Bermuda, como se visualiza en la Figura 10.
En el análisis de los gamers del juego en específico COD, se puede observar que en Estados Unidos también se tiene el mayor índice de gamers de dicho juego, mientras que para este análisis también se tiene que este juego no es muy famoso en países como Bolivia, El Salvador o Omán, como se muestra en la Figura 11.

d. Ranking de felicidad

Al utilizar el código en Python para ingresar los datos que se tiene en csv y utilizando la herramienta DB Browser (SQLite) se observan todos los datos como se visualiza en la Figura 12. En la Figura 13 se muestra que tanto es el nivel de generosidad con la que se ha calificado en cada uno de los países que se encuentran en el Ranking siendo Indonesia con la gente más generosa y Grecia postula como la que es menos generosa. En la Figura 14 se nos muestra en cual de todas las regiones se encuentra un mayor score que se obtuvo en los índices de felicidad teniendo un mayor porcentaje en África y el menor en Norte América.

e. Eventos o noticias mundiales

El análisis de los datos correspondientes al número de casos de COVID-19 registrados dio como resultado que el país con mayores casos es Estados Unidos con un número cercano a los 30 millones de casos como se puede observar en la Figura 15. Por otro lado, al realizar en análisis de la cantidad de muertos registrada por COVID-19 se puede obtener que Estados Unidos sigue siendo el país con mayor registro de muertos, observando esta información en la Figura 16. Por tanto, el país más afectado durante la pandemia del COVID-19 es Estados Unidos.

El análisis realizado hacia los candidatos estadounidenses en las anteriores votaciones dio como resultado que el estado de Delaware fue el que mayores votos recaudó para ese año como se puede observar en la Figura 17.

El análisis realizado hacia el número de contagiados durante el Ébola en Europa en los años de 2014 – 2016 dio como resultados que Liberia fue el país con un número de casos posibles y casos detectados en nivel medio, apreciándose así en la Figura 18. A su vez se puede observar que este mismo número va en constante incremente hacia países bajos.

El análisis realizado a las olimpiadas de Tokio 2021 posicionó a Estados Unidos como el país con mayor número de medallas de oro obtenidas en las olimpiadas, con un total de 39 medallas, visualizándose así en la Figura 19.

VIII. Visualización de información

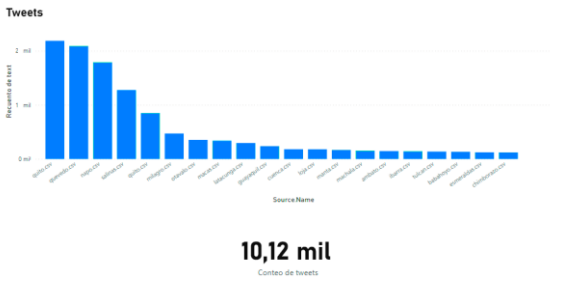

Figura 4 Tweets por ciudades

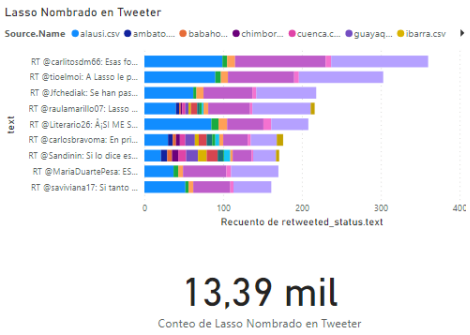

Figura 5 Tweets en los que se menciona a Guillermo Lasso

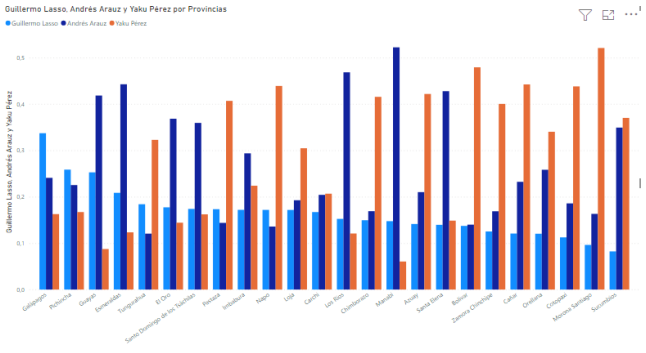

Figura 6 Candidatos por provincias

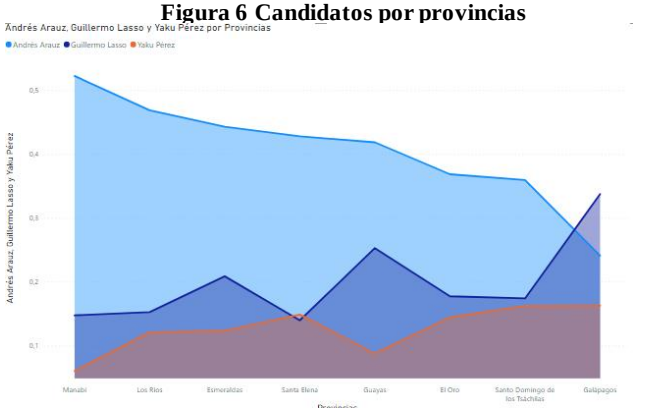

Figura 7 Pulso político en Costa

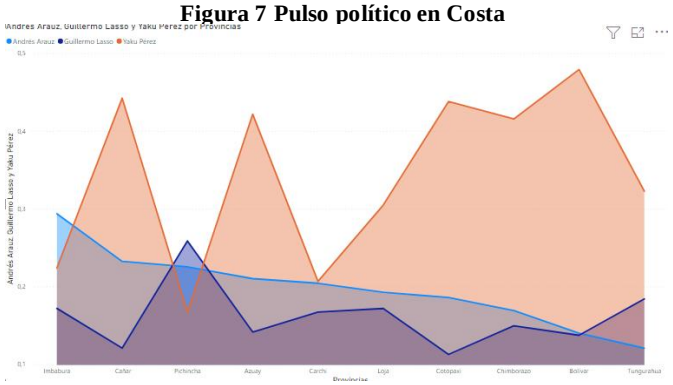

Figura 8 Pulso político en Sierra

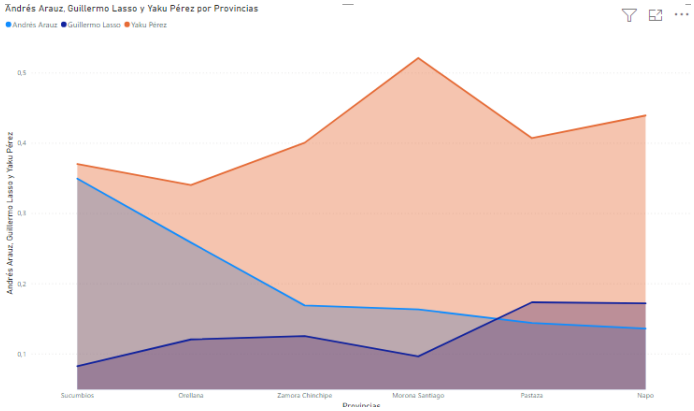

Figura 9 Pulso político en Amazonía

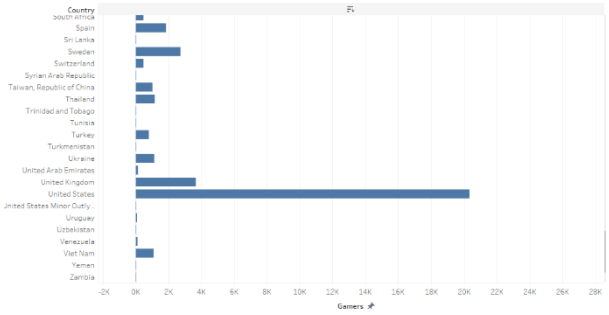

Figura 10 Figura analisis gamers-paises

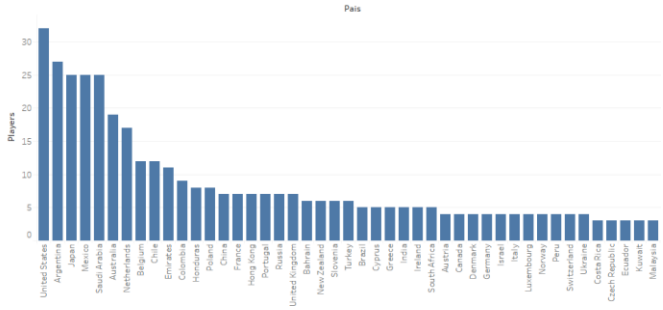

Figura 11 Figura gamers-cod

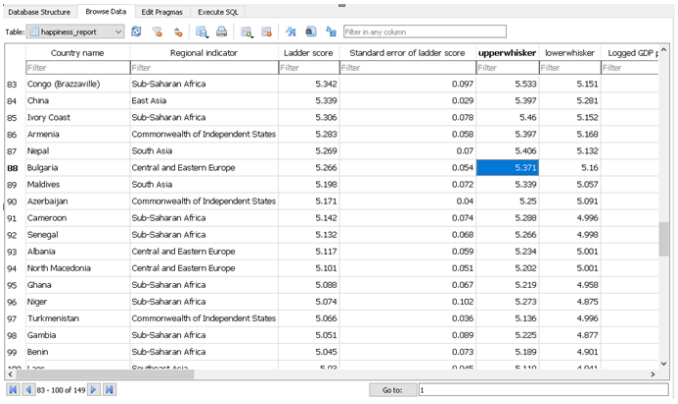

Figura 12 SQLite

Figura 13 Análisis geográfica generosidad

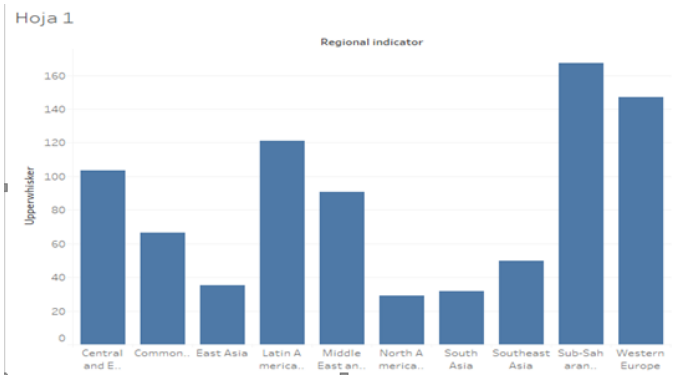

Figura 14 Indicador regional

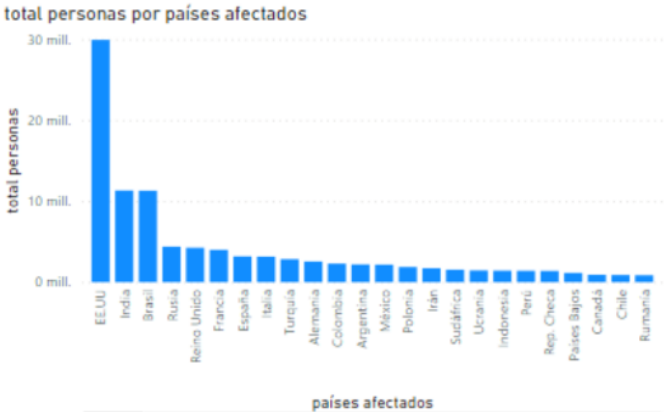

Figura 15 Casos Registrados

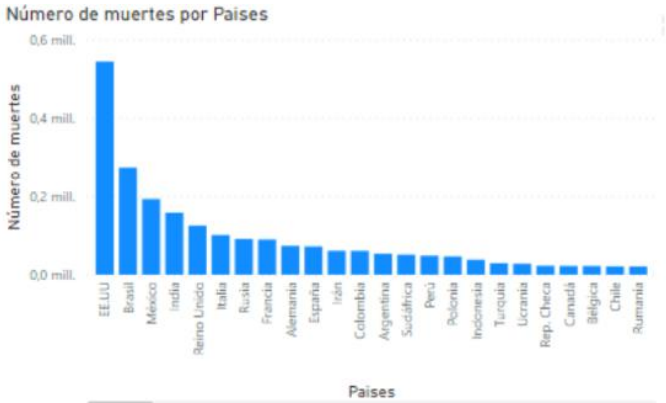

Figura 16 Muertes Registradas

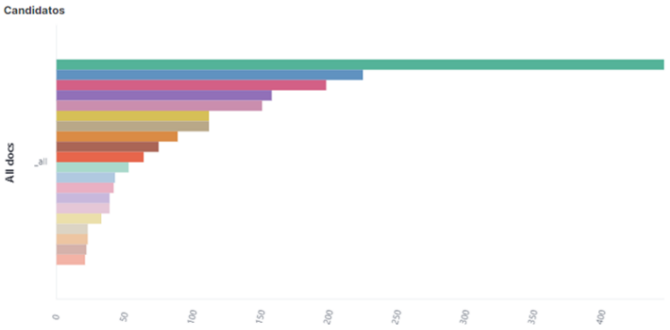

Figura 17 Estado Delaware

Figura 18 Incremento a países bajos

.

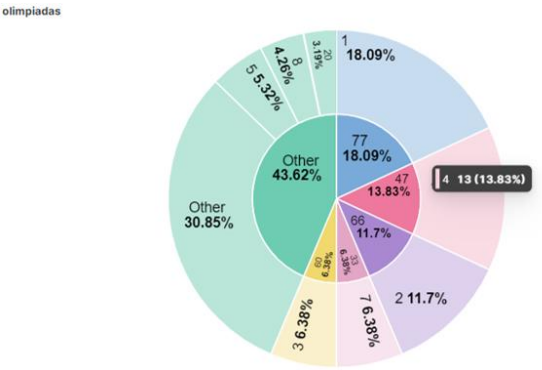

Figura 19 Medallas de oro

IX. RESULTADOS

Los datos que se recolectaron nos muestran el fuerte impacto que el candidato Guillermo Lasso tuvo en la opinión de la gente en las redes sociales. Lo que justifica su elección como presidente del Ecuador. También como podía esperarse la mayor cantidad de interacción vino de la ciudad de Quito.

A través de la información que se obtuvo del CNE por cada provincia del ecuador pudimos realizar el análisis y determinar qué candidato tuvo ventaja en cada región del Ecuador y el resultado es curioso ya que en la primera vuelta electoral Guillermo Lasso no fue el favorito en ninguna de las regiones del Ecuador.

Los datos se encuentran en muchos lugares y para conocer el número de jugadores por países se pueden encontrar varias fuentes las cuales se van actualizando constantemente, como se mostró en la sección de visualizaciones los resultados de los jugadores arrojo que n Estados Unidos se encuentra la mayoría

de estos esto se debe a que en ese país existe un avance tecnológico notable, y el estilo de vida es muy bueno puesto que su economía es estable. Al observar el otro lado de la situación existen países donde no hay gran incidencia de jugadores esto se puede deber a que el nivel de vida de sus habitantes no permite que muchas familias tengan en su hogar consolas de videojuegos o incluso internet.

Al obtener los datos del informe mundial de felicidad se investigó que estos fueron obtenidos de la encuesta mundial de Gallup el cual se pidió a todos los encuestados que puntúen aspectos de su vida dando un valor de 0 a 10, siendo cero la peor vida posible, además, de datos como la generosidad, la corrupción existente en el país, el apoyo social, la dystopia, la libertad entre otros. Así se pudo construir un score de acuerdo para puntuar a todas las ciudades del mundo, además, que en herramientas de visualización como Tableau se puede obtener un mejor sondeo de estos resultados.

Los datos analizados sobre los casos y muertes registradas posicionan a Estados Unidos como el país más afectado, más aún esto supone que aún siendo un país de primer mundo, no se encuentra libre de ser uno de los que más estragos sufrió. Además, al observar los análisis de las olimpiadas de Tokio 2021 Estados Unidos destaca, por lo que se puede concluir que, aun siendo el país más afectado durante la pandemia, no deja de ser unos de los que más destaca en actividades deportivas.

X. Conclusiones y recomendaciones

La Api de Twitter tiene múltiples ventajas para extraer información y utilizando las herramientas que se nos enseñó como Kibana o Power Bi podemos filtrar la información que más nos sirva para realizar un análisis en profundidad.

La web ofrece un sinnúmero de información la cual puede ser recopilada fácilmente con la ayuda de herramientas como Excel y posteriormente ser tratada y analizada para diferentes propósitos.

Cuando se necesita recolectar información de algún ámbito ya sea político, económico o de cualquier índole paginas como Kaggle nos proporcionan datos específicos, auténticos y utilizados para realizar cualquier tipo de mediciones.

El realizar el análisis de información sobre un determinado campo o tema nos permite recolectar datos precisos, esto facilita la toma de decisiones al momento de implementar o desarrollar algún proyecto.

XI. Desafíos y problemas

Para recopilación de datos desde la web con la ayuda de Excel es necesario realizar búsquedas en sitios web en donde se presente la información en tabas puesto que si la información requerida no se encuentra de esta forma su recopilación es muy complicada si se desea usar la herramienta de Excel.

La herramienta de Tableau aunque intuitiva con la mayoría de conexiones a una determinada base de datos, en lo que se tomo en algo complicado fue conectarlo a una base de datos SQLite puesto que era necesario descargar un complemento necesario en la pagina oficial para poder conectarlo a dicha base.

Para recaudar la información desde Twitter hubo la condición de no extraer información demasiado rápido y seguido ya que la API de twitter consideraba que no estaba permitido a realizar esta acción de forma repetida y constante.

XII. Referencias

[1] «Anatomía de un Tweet.,» DIGITALTROUPE, 17 marzo 2019. [En línea]. Available: http://www.digitaltroupe.com/anatomia-de-un-tweet/. [Último acceso: 11 septiembre 2021].

[2] «Call of Duty (CoD) - statistics & facts,» Statista, 13 Agosto 2021. [En línea]. Available: https://www.statista.com/topics/8300/call-of-duty-cod/. [Último acceso: 11 Septiembre 2021].

[3] «Highest Earnings By Country,» Sportsearings, 2021. [En línea]. Available: https://www.esportsearnings.com/countries. [Último acceso: 11 Septiembre 2021].

[4] «World Happiness Report 2021,» Kaggle, [En línea]. Available: https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021. [Último acceso: 11 septiembre 2021].

[5] «SQLite ODBC Driver,» ch-werner, 20 junio 2020. [En línea]. Available: http://www.ch-werner.de/sqliteodbc/. [Último acceso: 11 septiembre 2021].

[6] «Datasets,» Kaggle, [En línea]. Available: https://www.kaggle.com/datasets. [Último acceso: 11 septiembre 2021].

[7] «Consejo Nacional Electoral,» Gobierno del Ecuador, 14 febrero 2021. [En línea]. Available: https://www.gob.ec/. [Último acceso: 11 septiembre 2021].