

Proyecto Final de Análisis de Datos

- Integrantes:
- Elvis Agila
- Jonathan Alquina
- Yomara Diaz
- Guillermo Rivera

Definición de caso de estudio

- **Pulso político por provincia y ciudades.**
 - Los datos tomados tuvieron el fin de analizar la postura, opinión y preferencia de la población en base a temas relacionados con las Elecciones Ecuador 2021.
- **Eventos o noticias mundiales**
 - Se recopiló datos relacionados con el tema del Covid-19 alrededor del mundo, para comprender como ha avanzado esta problemática a nivel mundial.

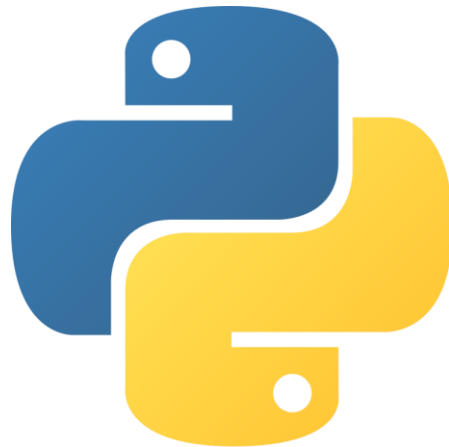
Descripción del equipo de trabajo y sus tareas a realizar.

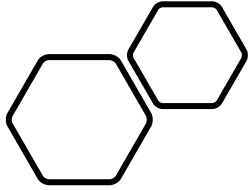
Nombre	Tareas
Elvis Agila	<ul style="list-style-type: none">• Recopilar datos sobre los videojuegos por países utilizando las herramientas con bases de datos públicas como las páginas de Kaggle.com, Tableau Public Resources y UNdata.• Se exportara los datos a la base de datos SQL en este caso MySQL.• Las visualizaciones de los datos obtenidos se realizaran utilizando la herramienta de PowerBI ya que la misma cuenta con la función de extraer datos de MySQL.
Yomara Diaz	<ul style="list-style-type: none">• Recopilación de datos desde kaggle para posterior almacenado en una base de datos tipo SQL• Conectar Power Bi con MySQL• Realización de visualizaciones en Power Bi.• Análisis de las visualizaciones en el dashboard de Power Bi
Jonathan Alquina	<ul style="list-style-type: none">• Recopilar datos sobre el pulso político por provincias en Ecuador, listas y candidatos y presidenciales, a través de Twitter utilizando un script de Python.• Guardar la información obtenida en un Base de Datos NoSQL, en este caso CouchDB.• Exportar esta información a Elasticsearch para crear un nuevo índice y poder realizar las visualizaciones en Kibana.• A partir del tema de preferencia elegido que es “Test de Personalidad”, se realizara las respectivas tareas mencionadas anteriormente.
Guillermo Rivera	<ul style="list-style-type: none">• Recopilar datos sobre el pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados, a través de Twitter utilizando un script de Python.• Guardar la información obtenida en un Base de Datos NoSQL, en este caso MongoDB.• Exportar esta información a Elasticsearch para crear un nuevo índice y poder realizar las visualizaciones en tiempo real utilizando Kibana.

Recursos y Herramientas utilizadas



kaggle

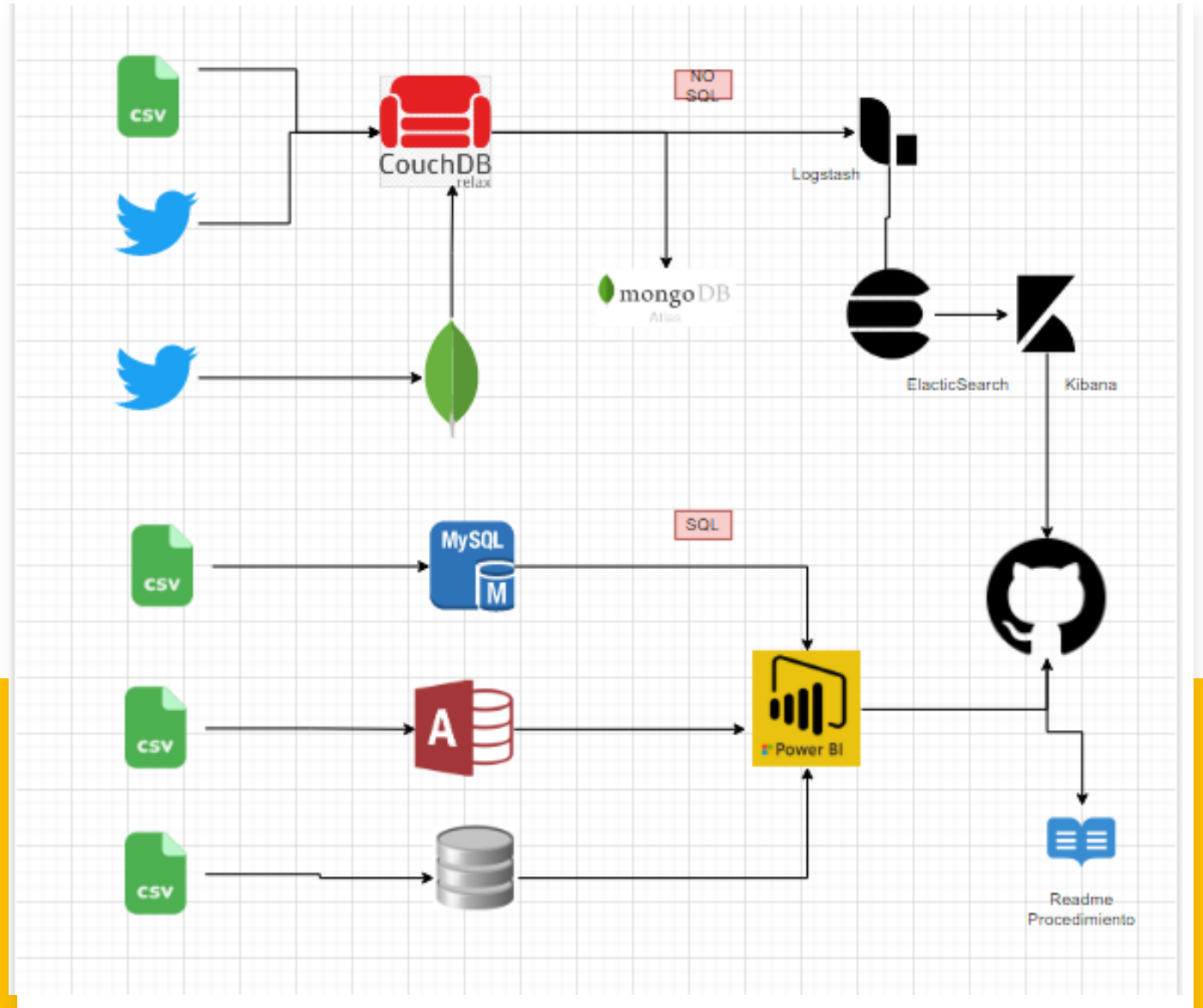




Cronograma

Objetivos	Actividades	Semana 1	Semana 2	Semana 3	Total
OBJ 1: Definir las formas en las cuales se va a recolectar los datos, incluyendo los parámetros en aquellos en los que se tiene que extraer información de fuentes específicas. Definir las bases de datos en las cuales se almacenará la información. OBJ 2: Establecer la arquitectura con la cual se desarrollará el proyecto, tomando en cuenta la vía más factible para cumplir con la problemática expuesta.	Definir el numero de Bases de Datos SQL y NoSQL	4			4
	Definir las herramientas	4			4
	Definir los recursos que se va a utilizar	4			4
	Definir el timepo del proyecto	4			4
	Crear script para la recoleccion de datos	6			6
	Buscar script CSV	6			6
	Ejecutar los script	20	10		30
	Analizar los datos de los script y CSV		10		10
	Cargar los datos en una Base de Datos		6	4	10
	Crear indices en ElasticSearch		6	4	10
OBJ 3: Concentrar los datos en la herramienta Kibana para realizar su respectivo análisis.	Crear y ejecutar los script para transportar los datos a ElasticSearch		4	4	8
OBJ 4: Recopilar data con respecto a los temas expuestos en la problemática.	Cargar los datos en la herramienta kibana		4	4	8
	Realizar la limpieza de datos			4	4
OBJ 5:• Recopilada la información, establecer un punto en el cual se incorporará el conjunto de dicha información. Realizar las visualizaciones acordes a los resultados que presente la información receptada.	Analizar las visualizaciones y dashboards			10	10
	Interpretar las visualizaciones y dashboards			14	14
OBJ 6: Interpretar las visualizaciones y los resultados de tal manera que permita a todo tipo de usuario entender de qué trata el tema expuesto y las características que presente	Documentar los resultados obtenidos			8	8
Total de horas por semana		48	40	62	150

Arquitectura de Solución

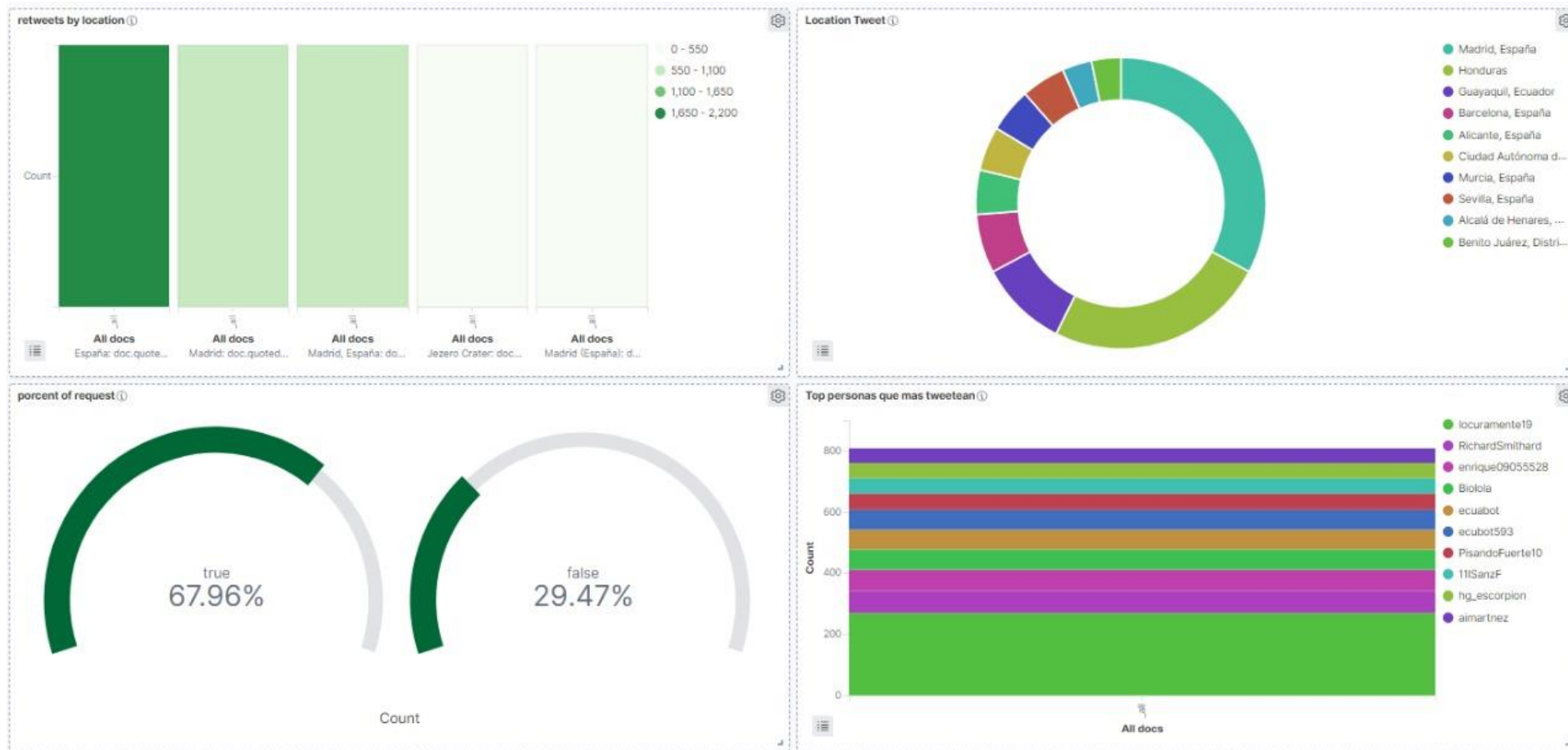


Pulso político de las 20
Ciudades Principales del
país.

Colecciones en Couch

Databases				inglés	español	name	Actions
				Google Translate			
Name	Size	# of Docs					
_replicator	10.4 KB	14			No		
ambato	43.4 MB	15903			No		
chone	10.6 MB	3842			No		
cuenca	43.4 MB	15903			No		
esmeraldas	11.6 MB	4178			No		
guaranda	26.9 MB	9809			No		
guayaquil	43.4 MB	15903			No		
ibarra	10.9 MB	3977			No		
latacunga	22.1 MB	8024			No		
loja	12.9 MB	4737			No		
losrios	8.7 MB	3132			No		
machala	10.1 MB	3712			No		
manta	9.5 MB	3482			No		
otavalo	0.6 MB	239			No		
peliculas	22.4 KB	30			No		
pichincha	0 bytes	0			No		
portoviejo	10.8 MB	3930			No		
quito	43.3 MB	15903			No		
salinas	9.6 MB	3548			No		
stodomingo	1.0 MB	356			No		
tena	9.9 MB	3506			No		

ANÁLISIS DE LAS VISUALIZACIONES DE KIBANA





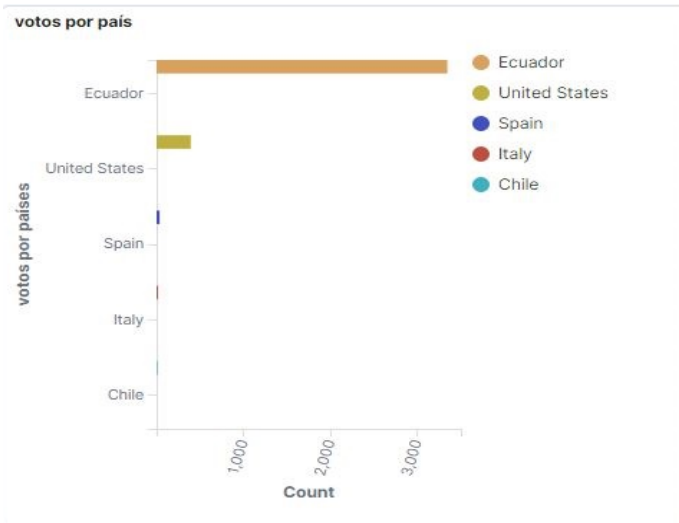
Pulso político para las provincias del Ecuador

Pulso Político para las Provincias del Ecuador

- ❖ Los datos se recopilaron por el script de “Tweets Provincias.py”, los que se almacenarían en la base de datos NoSQL CouchDB
- ❖ Se aplica un filtro por palabras con las palabras claves de: “ Elecciones Ecuador 2021, Nombre_Candidato, Partido Político” y su ubicación geográfica, eje: -79.3709,-0.7211,-77.8397,0.3273 (Pichincha)
- ❖ Losa datos recolectados se los enviara a ElasticSearch mediante un archivo de configuración “Nom_provincia.conf”
- ❖ Los datos recopilados serán analizados y visualizados mediante Kibana.

Database nar		
1.8 MB	712	No
1.8 MB	735	No
0 bytes	0	No
0 bytes	0	No
2.8 MB	1197	No
2.7 MB	1147	No
1.4 MB	594	No
335.3 KB	145	No
218.9 KB	86	No
216.6 KB	85	No
240.3 KB	96	No
234.5 KB	94	No
32.8 KB	13	No

Visualizaciones en Kibana



- Votos por países

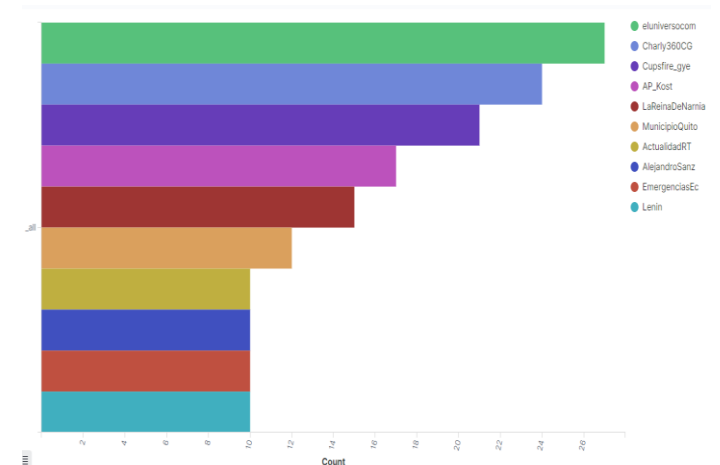
- Los hashtags más utilizados



- 10 usuarios que realizan más twist



- Votos por candidato.



Resultados obtenidos

- ❖ Votos por candidato
 - ❖ Mediante el CSV del diario, las visualizaciones que se obtuvieron son muy representativas porque como candidato principal de este datasets se tiene a Rafael Correa como un postulante y que el va a ganar las elecciones, pero cabe recalcar que solo es un datasets de encuestas realizados a las personas del diario, en esa fecha todavía no se tenían candidatos a presidentes.
- ❖ Voto por cada país
 - ❖ Igual estos resultados se obtuvieron de del data set del diario, lo interesante es que esta encuesta llego a 12 países de Latinoamérica, pero para esta visualización solo se analizan 5 que se mostro en la visualización y a simple vista se nota la participación de los ecuatorianos
- ❖ Los hashtags más utilizados
 - ❖ En esta visualización lo que nos muestra es los hashtag que los usuarios ponían para poder Repostear una publicación en Twitter, Aquí como dato interesante pusimos un top de lo 100 hashtag , pero se puede modificar a 10 o depende del Análisis que le vamos a proporcionar
- ❖ 10 usuarios que realizan más twist
 - ❖ Realizamos un top 10 de los usuarios que mas postean en Twitter los que están apoyando los diferentes partidos políticos que existe en el Ecuador

Segunda parte



Es una herramienta que se utiliza principalmente para crear cuadros de mando que faciliten la toma de decisiones, nos ayudan a las organizaciones a recopilar, administrar y analizar datos de una variedad

Herramientas

MySQL sirve para almacenar toda la información que se desee en bases de datos relacionales, como también para administrar todos estos datos sin apenas complicaciones.



Es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios.

Es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos.

Herramientas

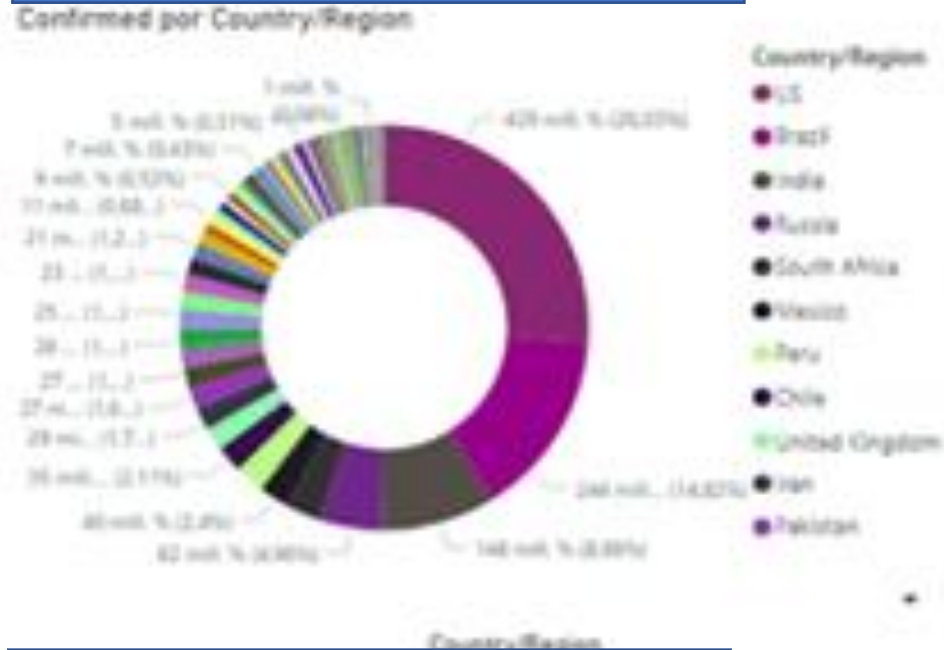


Plataforma gratuita para obtener datasets.

Plataforma gratuita para obtener datasets.

Covid mundialmente

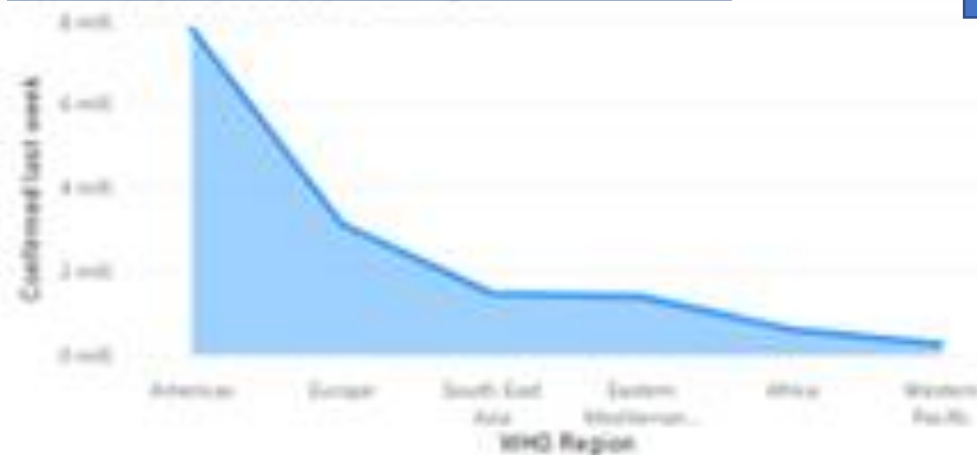
Casos confirmados por país



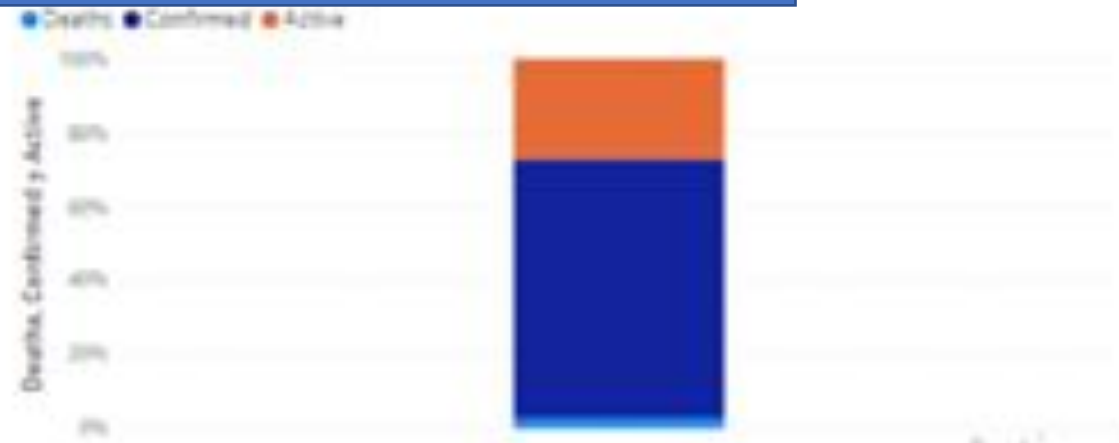
Muertes por cada continente



Casos confirmados por continente

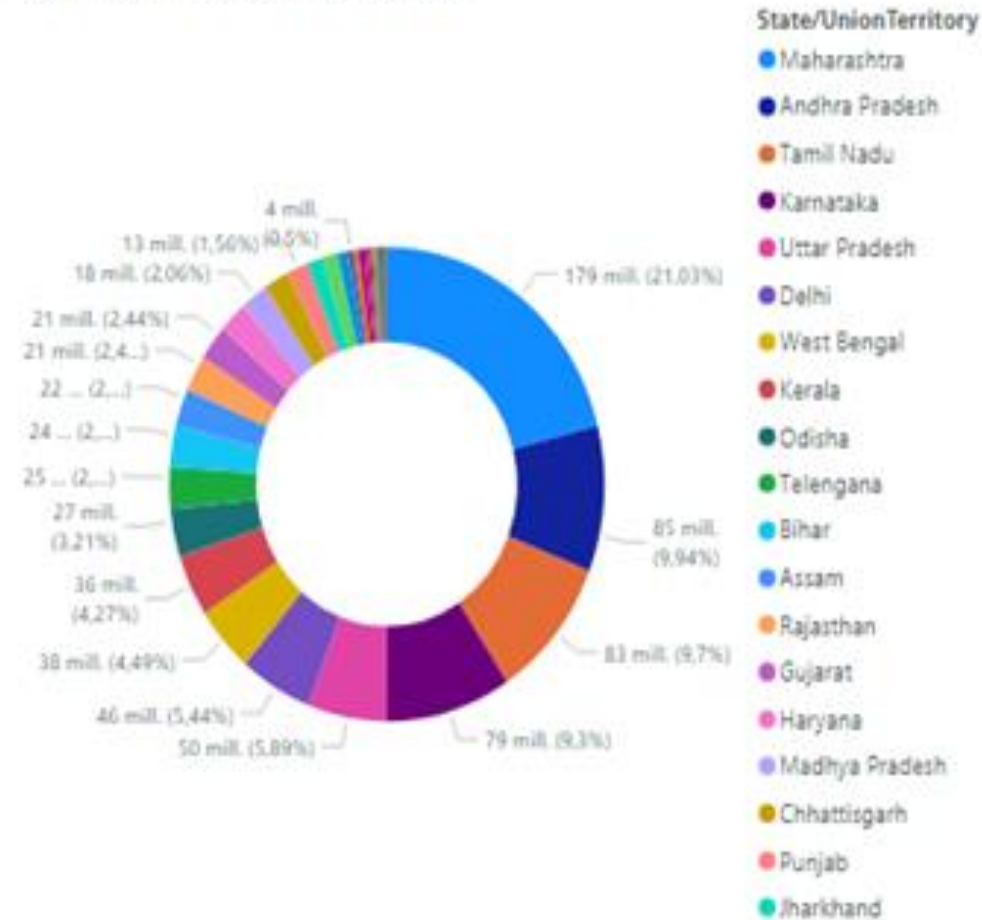


Casos de muerte, confirmados y activos.

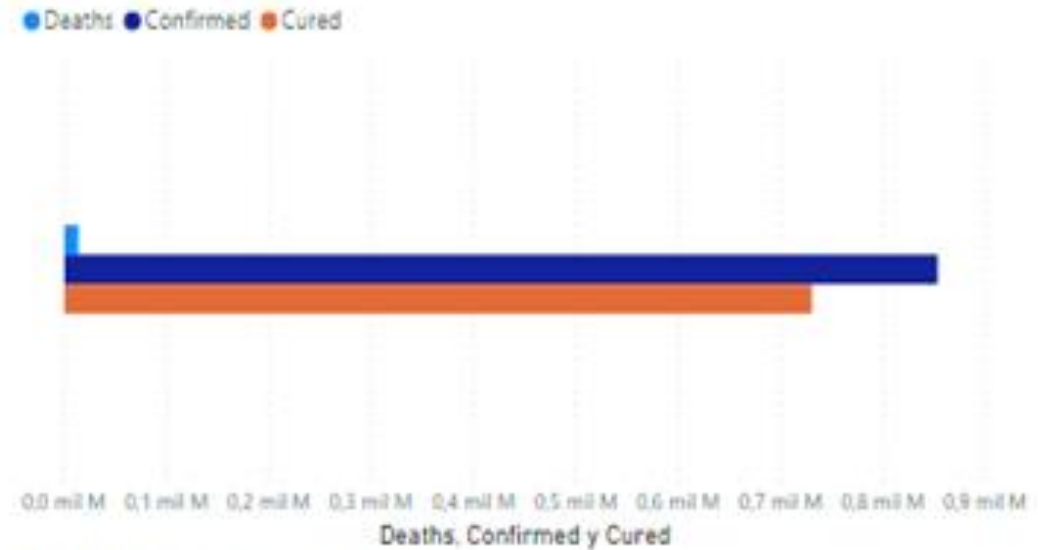


Covid India

Confirmed por State/Union Territory



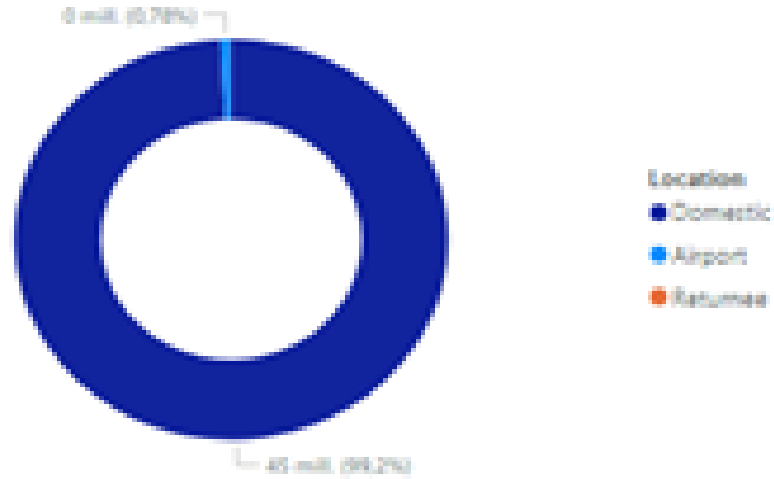
Deaths, Confirmed y Cured



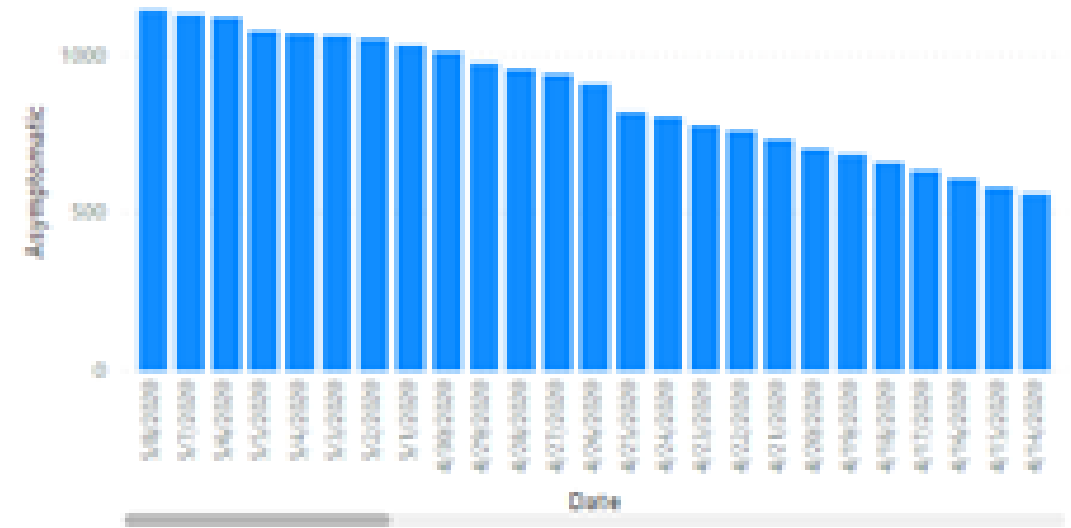
Confirmed por State/Union Territory



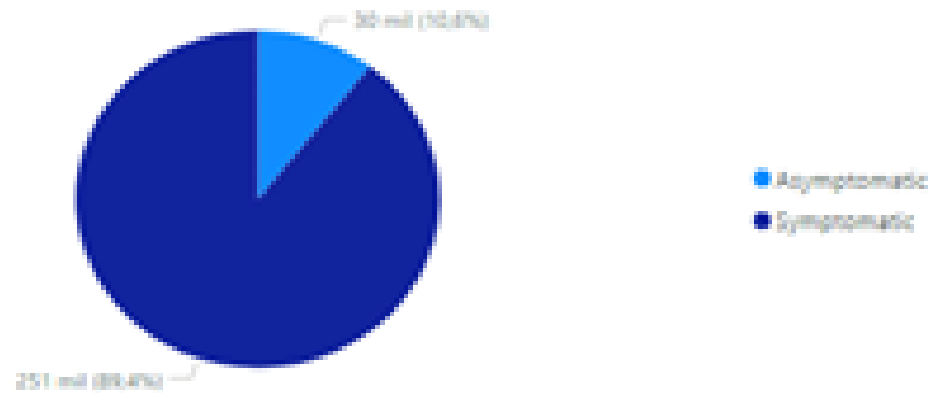
Covid Japón



Asymptomatic per Date



Asymptomatic y Symptomatic

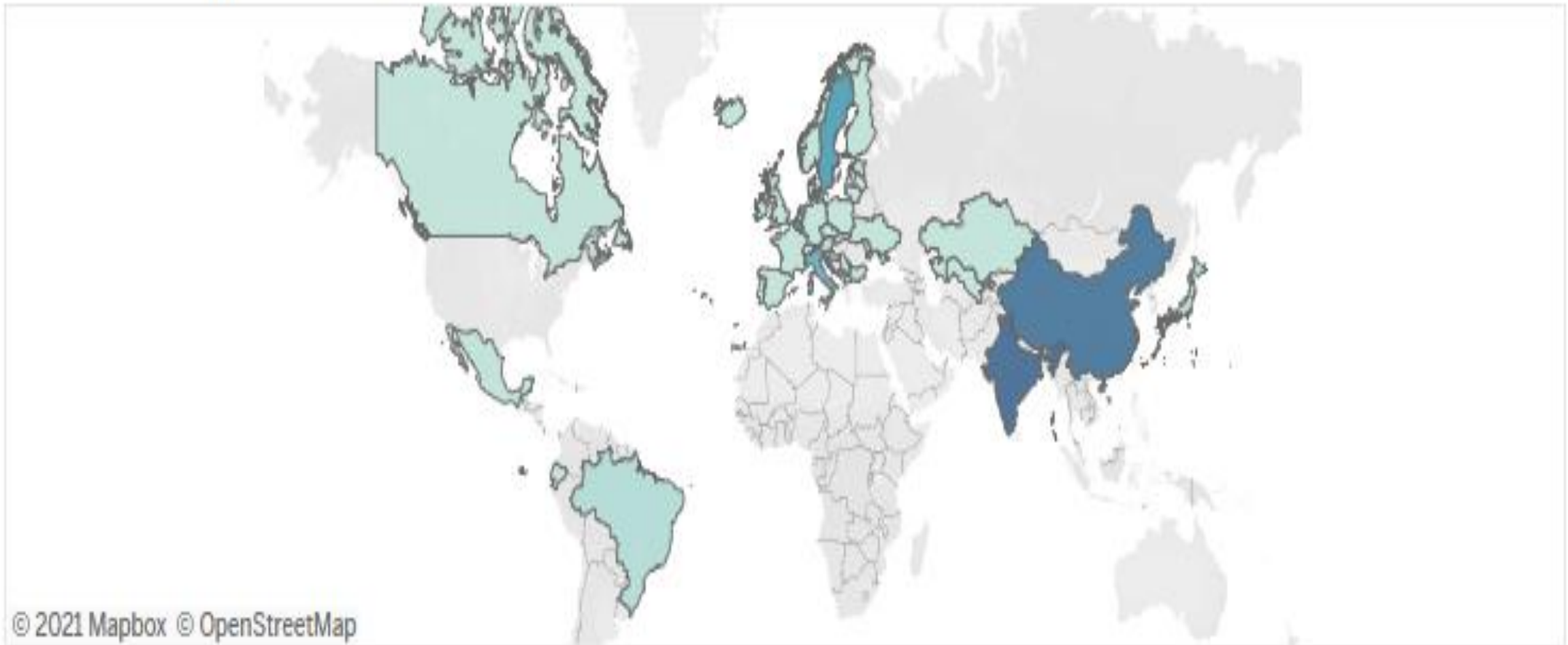


Resultados Obtenidos

- **Covid mundial.-** podemos identificar que Estados Unidos tiene el primer lugar en porcentaje de pacientes covid el cual es el 26.03%, como segundo país esta Brazil con el 14.82% y en tercer lugar India con el 8.98% de contagios.
- **Covid india.-** nos muestra el porcentaje de casos contagiados de covid por cada región tomando el primer lugar Maharashtra (21.03%), segundo Andhra Pradesh (9.94%), tercer lugar Tamil Nadu (9,7%).
- **Covid Japón.-** se analizó los casos positivos de covid en tres localidades: Domestic (99.2%), airport (0.78%), return (0.02%), es decir la detección de casos mas grande se dio dentro del país. También se analizó los casos de covid con personas asintomáticas y con síntomas la cual lleva al análisis que existen más personas que presentan síntomas de covid con el 88.6% y el resto de las personas 9.25%

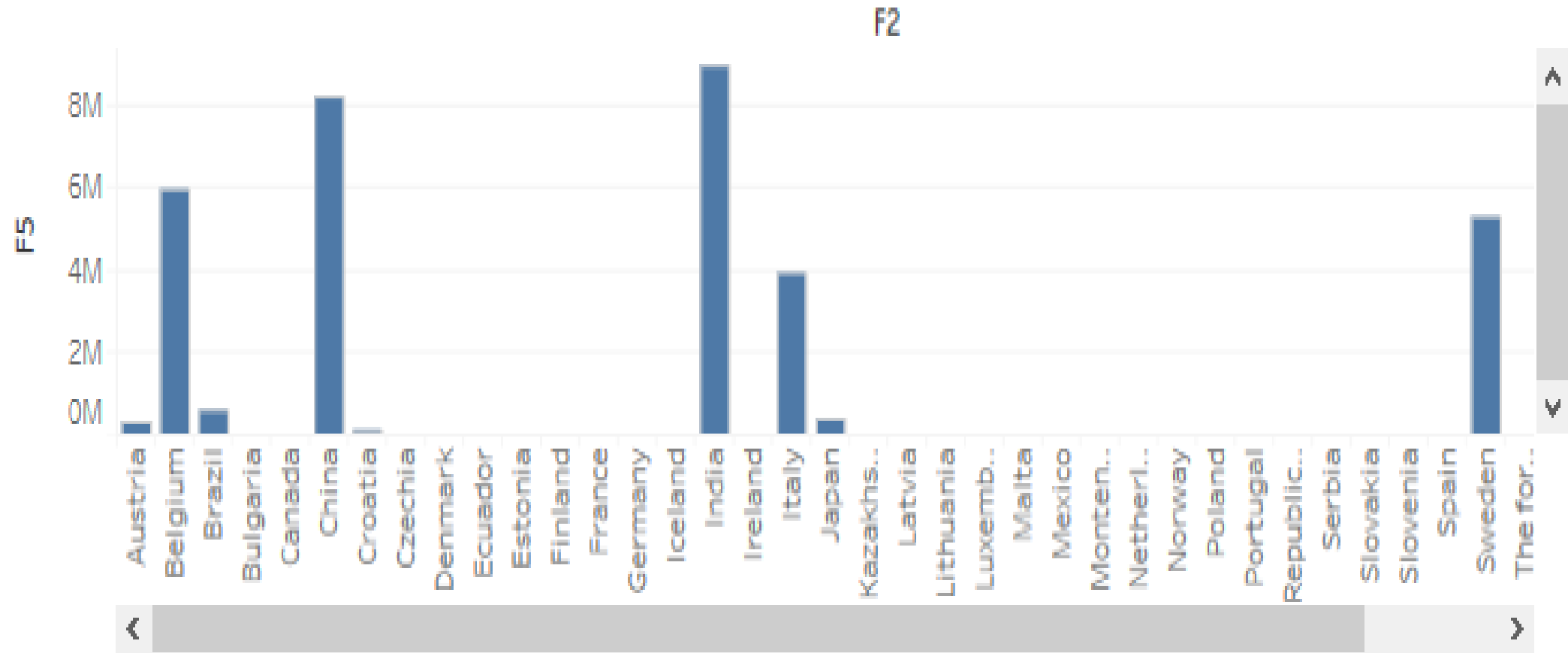
Visualizaciones de los datos sobre video juegos por países

Unidades exportadas vista geografica



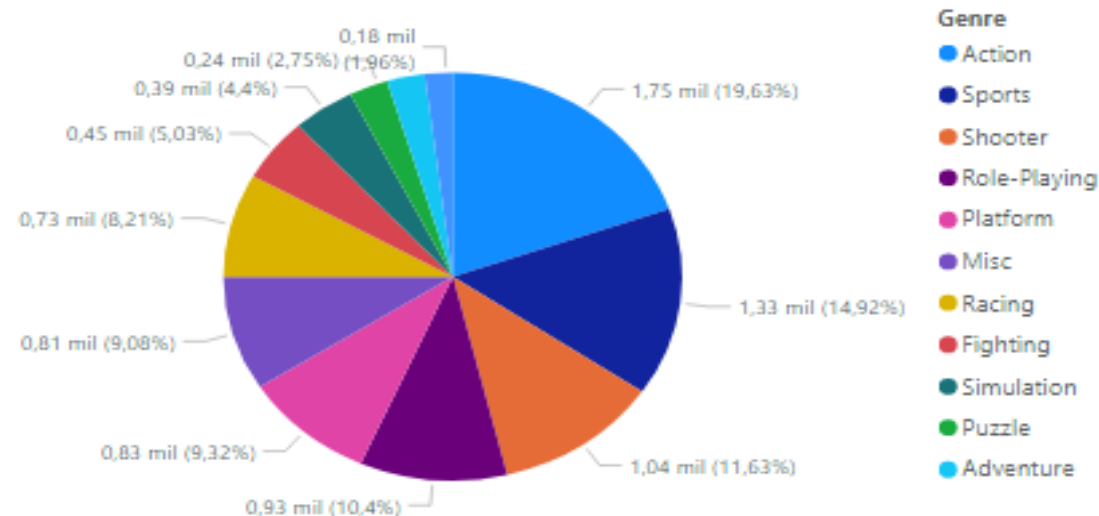
Visualizaciones de los datos sobre video juegos por países

Países/Exportaciones(USD)



Visualizaciones de los datos sobre video juegos por países

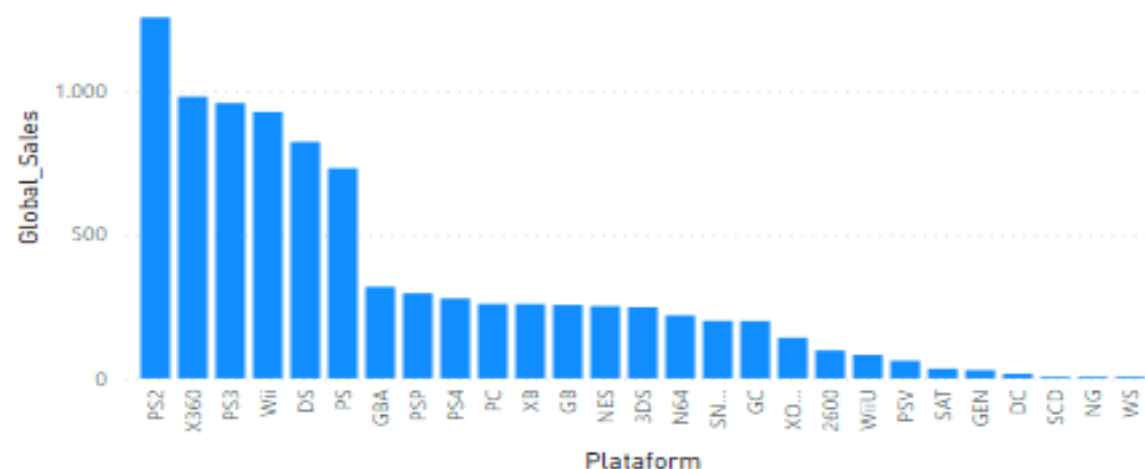
Global_Sales por Genre



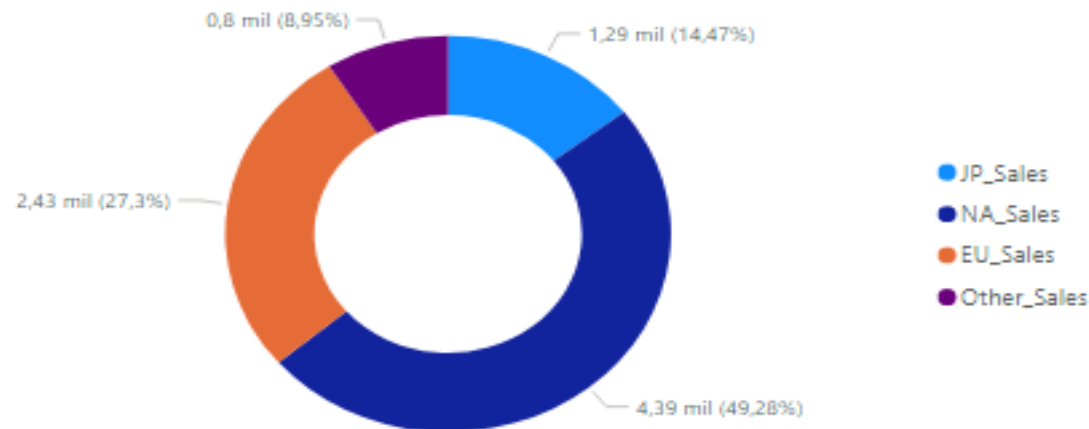
Publisher Global_Sales

Publisher	Global_Sales
Nintendo	1.786,56
Electronic Arts	1.110,32
Activision	727,46
Sony Computer Entert	632,57
Ubisoft	474,72
Take-Two Interactive	399,54
THQ	340,77
Konami Digital Enter	283,64
Sega	272,99
Namco Bandai Games	254,09
Microsoft Game Studi	245,79
Capcom	200,89
Atari	157,22
Warner Bros. Interac	153,89
Square Enix	145,18
Total	8.918,95

Global_Sales por Plataforma



JP_Sales, NA_Sales, EU_Sales y Other_Sales



Resultados Obtenidos

- **Visualización geográfica.-** Con los datos recopilados en la vista geográfica se puede apreciar como determinados países exportan video juegos en una gran escala y corroborando que las potencias mundiales de Europa, Asia y Norte América lideran la exportación de videojuegos consolas y mas tecnología para el entretenimiento.
- **Países/Exportaciones.-** A diferencia de la primera grafica primera grafica en esta se aprecia el ingreso que tuvieron determinados países en cuanto a la exportación de video juegos, con el avance de la tecnología los video juegos se han vuelto una fuente de entretenimiento que genera grandes ganancias para los desarrolladores de juegos y de consolas para los mismos.
- **Graficas de PowerBI.-** Las graficas realizados en PowerBI se realizaron con una segunda base de datos donde se clasifica juegos, consolas y el numero de ventas en las grandes potencias se obtuvieron diferentes resultados en cuanto a ventas por géneros de juegos, ventas y por el tipo de juego.



Conclusiones y recomendaciones

- La importación y exportación de datos es algo que se debe hacer para cumplir la arquitectura del proyecto, así que, hay que saber bien la estructura de los datos para no tener problemas al momento de exportar o importar estos.
 - La limpieza de datos es un tema importante que se debe hacer para que al momento de realizar las visualizaciones estas puedan ser lo más exacta posibles y puedan cumplir con su propósito.
 - La información que se puede obtener en base a los datos obtenidos es muy importante y variada, hemos visto como la información construye y nos da una idea clara de preferencias e ideas, de que se está hablando actualmente.
-

Desafíos y Problemas

- Uno de los problemas encontrados fue que para poder cargar todos los datos del dataset a la base de datos MySQL, se tuvo que cambiar el tiempo máximo de ejecución, ya que, de manera predeterminada está establecido en 30 segundos lo que no permitía que los datos fueran cargados totalmente.