Andrea Reyes 20190265
Katherine García 20190418
Business Intelligence
20/05/2022

Data Slicerss
Proof Of Concept

Link a repositorio en Git: https://github.com/AndreaNathalia/BIproject

Descripción del proyecto:
Data Slicerss ha sido contratado por la ONG llamada Health and Well Being (HWB), que tiene el objetivo de evaluar el impacto de la pandemia de Covid 19 durante 2020 y 2021 y proporcionar al gobierno de los Estados Unidos un informe detallando cómo se vieron afectados los diferentes estados y condados.

Criterios de éxito para el proyecto
Nuestras métricas y criterios de éxito son definidas en base al valor que se da al cliente (HWB) y en base al cumplimiento de los tasks definidos por el equipo. El objetivo principal de Data Slicerss es responder a todas las preguntas formuladas por HWB para obtener los detalles necesarios y llevar a cabo el informe para el gobierno de Estados Unidos.

- Uso de la metodología Scrum.
- Cumplimiento de los sprints (entregas a tiempo por parte del equipo).
- Cumplir con las expectativas del cliente.
- Llegar a una conclusión y a una recomendación concreta sobre los estados afectados por Covid-19.
- Contar con visualización entendible de la data a presentar al finalizar el proyecto.
- Cumplir con los objetivos del workflow completo.

Estimado de Duración y Esfuerzo
Este proyecto se llevará a cabo en las siguientes 4 semanas, culminando el 17 de mayo del 2022. En este periodo de tiempo se espera resolver las diferentes preguntas planteadas e inquietudes sobre el Covid-19.

El proyecto se manejará con la metodología de Scrum. Los sprints serán semanales y se llevará el control desde la plataforma Trello. El esfuerzo semanal por cada integrante del equipo es de 3 a 5 horas semanales. Esto ayudará a que el proyecto se desarrolle de manera continua y sin atrasos.

El objetivo de la semana 1 es completar las etapas de comprensión del problema y de los datos escogidos para la resolución del proyecto. Seguido de esto, en la semana 2 se buscará terminar con la preparación de los datos y todo lo que ésta conlleva. Esta etapa se puede extender hasta la semana 3. En la cual se buscará refinar y mejorar la data para su posterior análisis. Finalmente, a finales de la tercera semana y la última, se realizará el análisis y las herramientas visuales que sean necesarias para una comprensión profunda y una presentación detallada de las observaciones y hallazgos que surjan de la investigación.

Alcance de este documento (POC)

El propósito de este documento es demostrar el plan definido para este proyecto. Data Slicerss fue contratado por una ONG para que se evalúe el impacto que la pandemia de Covid-19 ha tenido durante los años 2020 y 2021. Por lo que, se plantearon diferentes objetivos para demostrar la viabilidad de este proyecto y lograr resultados precisos.

El alcance que se espera lograr es el de contestar las preguntas planteadas por la ONG con base en la investigación que será realizada en los siguientes sprints. Como objetivos principales se encuentran:

- Conocimiento de los datos
- Análisis profundo de los datos
- Presentación clara y comprensible

Para conocer los datos se revisarán las preguntas y los datasets que se poseen para escoger los adecuados y los que nos proporcionen mayor información. Luego, para el manejo de los datos, se evaluó el mejor modelo a seguir.

Entre un ETL y un ELT, se decidió por seguir un ETL, en donde se transformarán primero los datos previo a su carga en la base de datos. Se decidió esto ya que, se considera que la base de datos se encontrará más organizada y limpia si desde el principio se poseen datos que ya se encuentren listos para su uso. Asimismo proveen más fiabilidad y seguridad en ciertos aspectos de los datos, que al realizar un ELT.

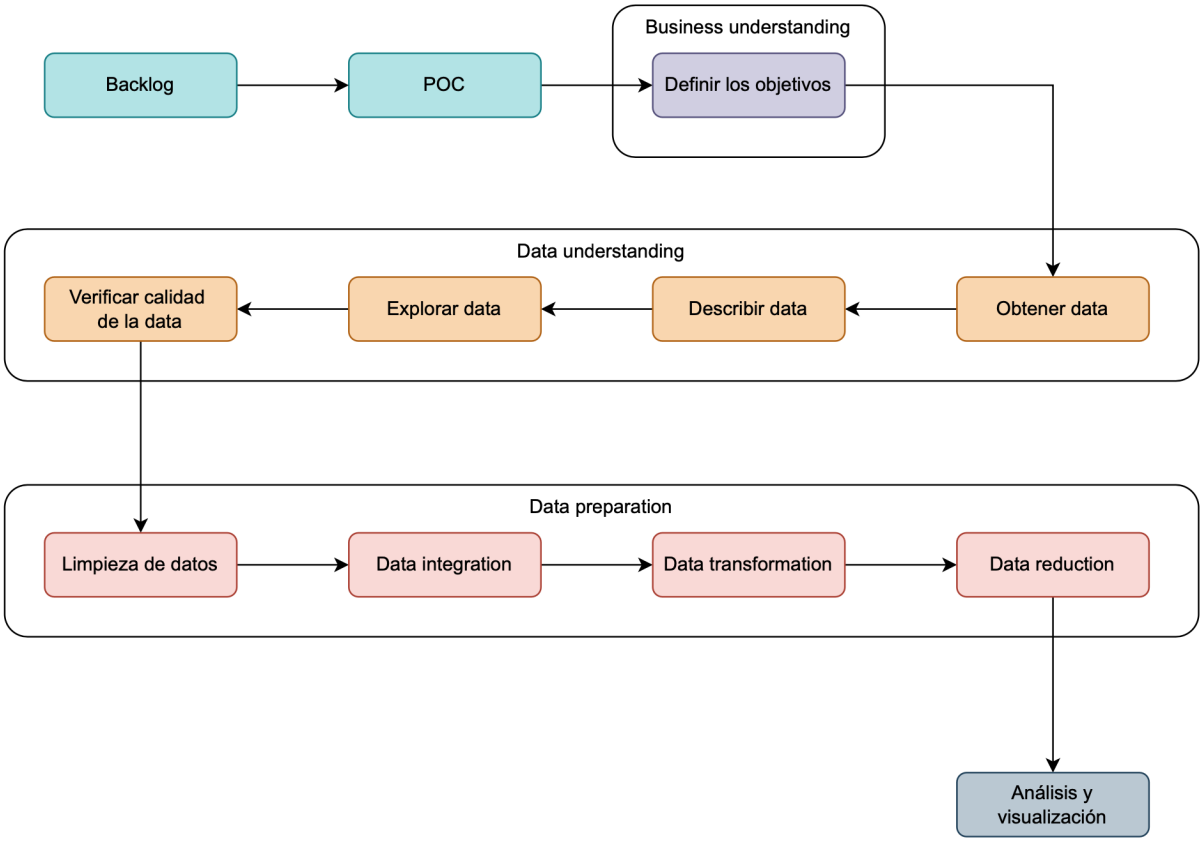
El análisis se hará con el objetivo principal de resolver las preguntas sobre el impacto de la pandemia de Covid-19. Seguido de esto, se espera poder encontrar otros hallazgos que aporten significativamente a nuestras observaciones para la ONG. Finalmente, se realizará una presentación amigable de los datos para presentarle a Health Well Being.

Recursos a utilizar

Para este proyecto se hará uso de un listado de tecnologías y herramientas para lograr cada objetivo y task definido.

- Scrum como metodología.
- Trello para el control y proceso del proyecto.
- GitHub para versionamiento y merge de distribución del trabajo.
- Modelo ERD para el modelo DB, con los features necesarios (no el dataset completo).
- PostgreSQL para la base de datos.
- Python - Pandas para la transformación de datos.
- Datasets para el ETL, las preguntas planteadas y los dashboards a presentar.
- Tableau para la integración, análisis y visualización de la información.

Workflow del proyecto



Datasets a utilizar para cada pregunta:

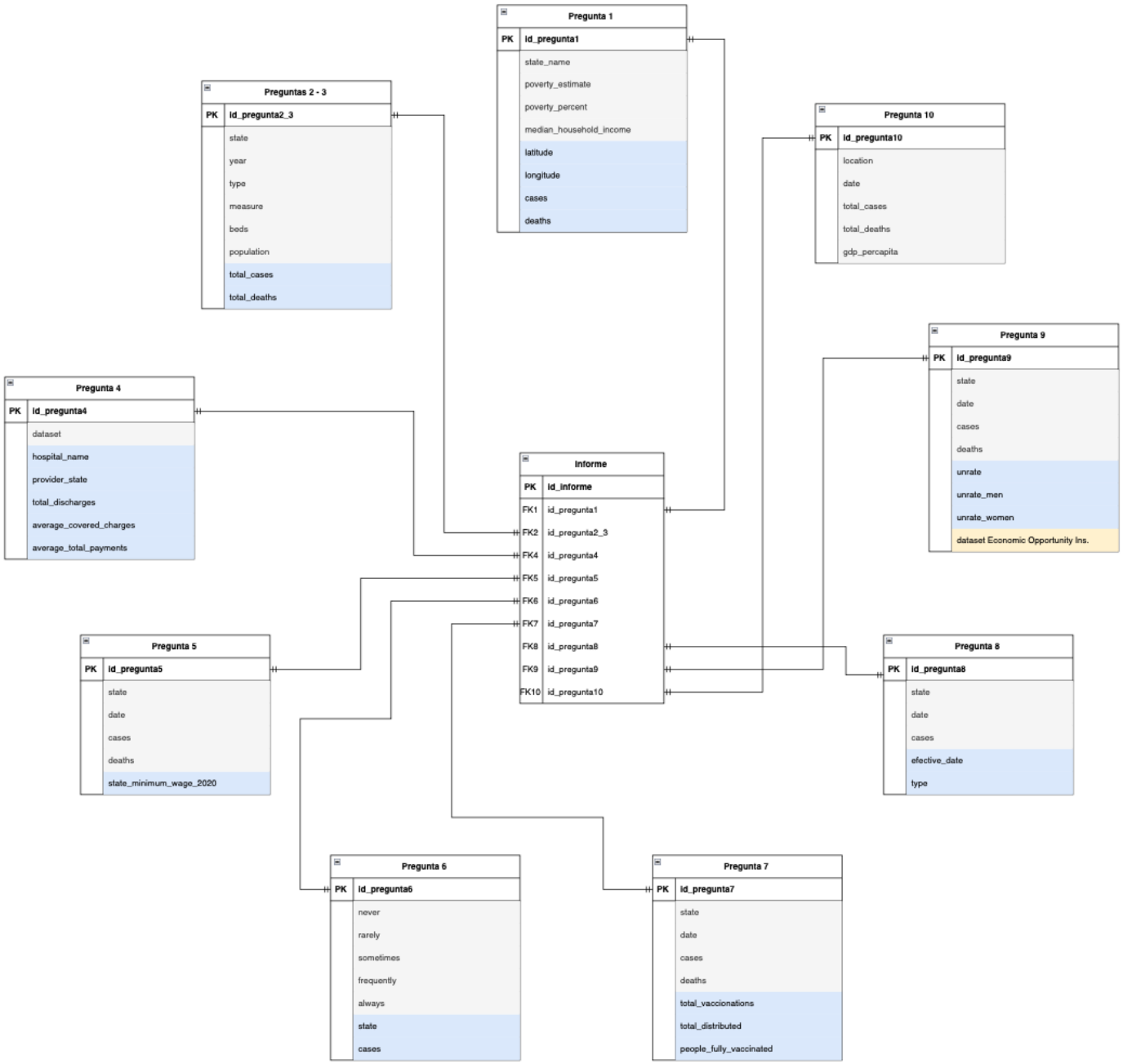
No.	Pregunta	Datasets
1	¿Existe una relación directa entre la pobreza, la infección y la muerte provocada por el Covid 19?	• Poverty Data 2020 • COVID-19 US County JHU Data & Demographics
2	¿Cuáles son los estados de EE.UU. con menos camas de hospital y que se vieron desbordados por el Covid 19?	• Hospital Bed Capacity by State and County • Impact of Covid 19
3	¿Cuáles son los estados de EE.UU. con más camas de hospital y que se vieron menos afectados por el Covid 19?	• Hospital Bed Capacity by State and County • Impact of Covid 19
4	¿Cuáles son los costes medios por hospitalización por Covid 19, cuál es el estado más barato para tratar el Covid 19 y cuál es el estado más caro para tratar el Covid 19?	• Covid 19 Deaths, hospitalization and other stats • Hospital Charges in US
5	¿Está el salario mínimo directamente relacionado con las infecciones y las muertes?	• COVID-19 US County JHU Data & Demographics

		• US Minimum Wage by State
6	El uso de mascarillas en los diferentes estados, ayudó a reducir las infecciones, ¿hay estados con alto uso de mascarillas y altas infecciones?	• Mask use by county • COVID-19 US County JHU Data & Demographics
7	¿Cuál ha sido el impacto de los avances de la vacunación en la reducción de las infecciones y las muertes?	• COVID-19 US County JHU Data & Demographics • US Vaccination Progress By State
8	¿Ayudaron realmente los cierres a reducir las infecciones?	• COVID-19 US County JHU Data & Demographics • US Lockdowns by State and County
9	Cómo afecta el Covid 19 la tasa de desempleo. ¿Qué periodo es el más crítico para el desempleo?	• COVID-19 US County JHU Data & Demographics • Us Unemployment Data • US Economic Opportunity Insides
10	Cómo afectó la pandemia a la economía mundial, utilice indicadores como el PIB, para analizar el impacto.	• Impact of Covid 19

Modelo Entidad-Relación Base de Datos

Se decidió por un modelo de entidad relación por su fácil entendimiento y visibilidad de sus componentes y relaciones entre campos. Y por la facilidad en cuanto a la manipulación y extracción de los datos en la BD, pues cada miembro del equipo tendrá conocimiento de qué campo de cada tabla se tienen que utilizar en las cláusulas y con qué otros elementos se relacionan.

El siguiente diagrama representa el modelo relacional del proyecto, dónde cada tabla representa una o dos de las preguntas a responder en el informe. Cada tabla se formó en base a los datasets seleccionados en la tabla anterior, de los cuáles se seleccionaron únicamente los features necesarios para responder la pregunta correspondiente. Los colores en las tablas representan a qué dataset corresponde cada campo.



POC

Para la prueba y testeo de los recursos y concepto del proyecto propuestos, se implementó el workflow completo en una porción pequeña de la información. Se hizo el ETL, integración y visualización de la información correspondiente a la pregunta 1.

Flujo implementado

ETL:

- Extracción y transformación:

```
poverty = poverty.drop(labels=["Age 0-17", "Age 5-17 in Families", "Age 0-4"], axis=1, level=0)
poverty = poverty.droplevel(level=0, axis=1)
poverty = poverty.drop(labels=["State FIPS Code", "Postal Code", "90% CI Lower Bound", "90% CI Upper Bound",
                               "90% CI Lower Bound.1", "90% CI Upper Bound.1", "90% CI Lower Bound", "90% CI Upper Bound"]), axis=1)
[5]  ✓ 0.3s
```

```
covid = covid[covid.date.str.contains("2020")]
covid = covid.drop(labels=["fips", "county", "state_code", "date"], axis=1)
[6]  ✓ 0.9s
```

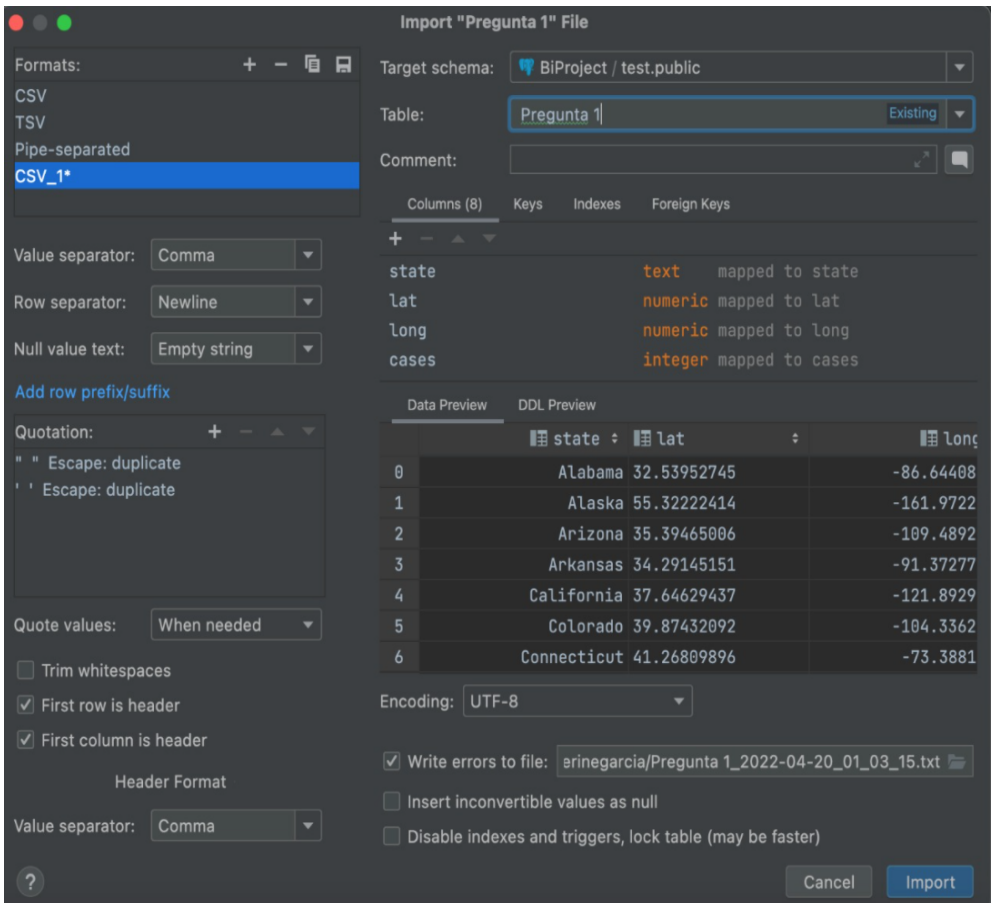
```
cases = covid.groupby(["state"])["cases"].sum().tolist()
deaths = covid.groupby(["state"])["deaths"].sum().tolist()
[7]  ✓ 0.1s
```

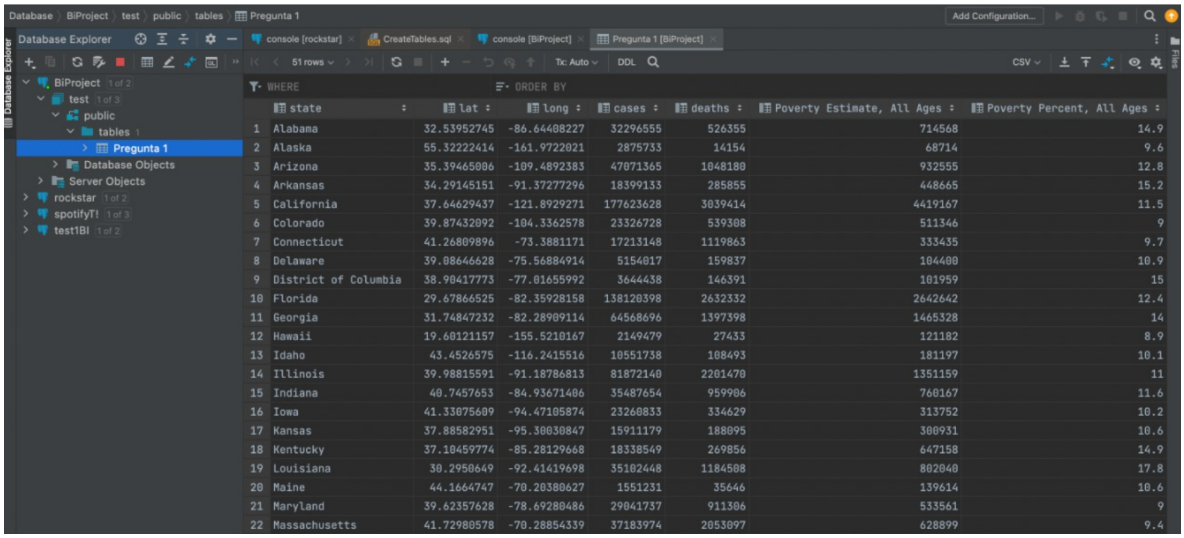
+ Code + Markdown

```
newCovid = covid.groupby(["state"]).first().reset_index()

newCovid["cases"] = cases
newCovid["deaths"] = deaths
[10]  ✓ 0.9s
```

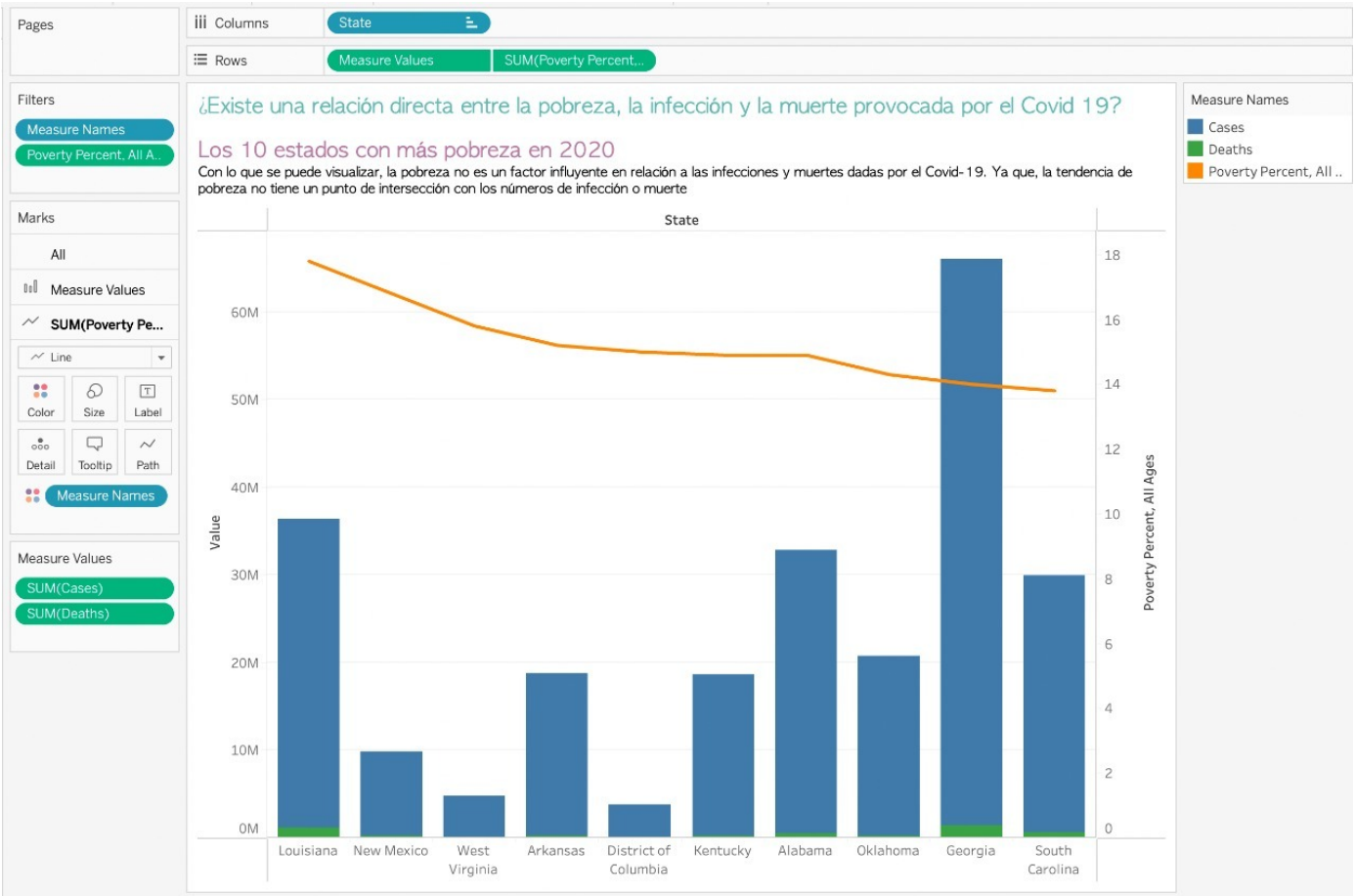
- Carga de datos:





	state	lat	long	cases	deaths	Poverty Estimate, All Ages	Poverty Percent, All Ages
1	Alabama	32.53952745	-86.64408227	32296555	526355	714568	14.9
2	Alaska	55.32222414	-161.9722021	2875733	14154	68714	9.6
3	Arizona	35.39465086	-109.4892383	47071365	1048180	932555	12.8
4	Arkansas	34.29145151	-91.37277296	18399133	285855	448665	15.2
5	California	37.64629437	-121.8929271	177623628	3039414	4419167	11.5
6	Colorado	39.87432092	-104.3362578	23326728	539308	511346	9
7	Connecticut	41.26809896	-73.3881171	17213148	1119863	333435	9.7
8	Delaware	39.08646628	-75.56884914	5154017	159837	104400	10.9
9	District of Columbia	38.90417773	-77.01655992	3644438	146391	101959	15
10	Florida	29.67866525	-82.35928158	138120398	2632332	2642642	12.4
11	Georgia	31.74847232	-82.28909114	64568696	1397398	1465328	14
12	Hawaii	19.60121157	-155.5210167	2149479	27433	121182	8.9
13	Idaho	43.4526575	-116.2415516	10551738	108493	181197	10.1
14	Illinois	39.98815591	-91.18786813	81872140	2201470	1351159	11
15	Indiana	40.7457653	-84.93671486	35487654	959906	760167	11.6
16	Iowa	41.33075609	-94.47105874	23260833	334629	313752	10.2
17	Kansas	37.88582951	-95.30030847	15911179	188095	300931	10.6
18	Kentucky	37.10459774	-85.28129668	18338549	269856	647158	14.9
19	Louisiana	30.2950649	-92.41419698	35102448	1184508	802040	17.8
20	Maine	44.1664747	-70.20380627	1551231	35646	139614	10.6
21	Maryland	39.62357628	-78.69280486	29041737	911306	533561	9
22	Massachusetts	41.72980578	-70.28854339	37183974	2053097	628899	9.4

Tableau Integration:



Partiendo de cómo fluyó la prueba y los resultados que se obtuvieron, se continuará trabajando de esta manera. Realizando ETLs, Postgres y la misma estructura para la base de datos.