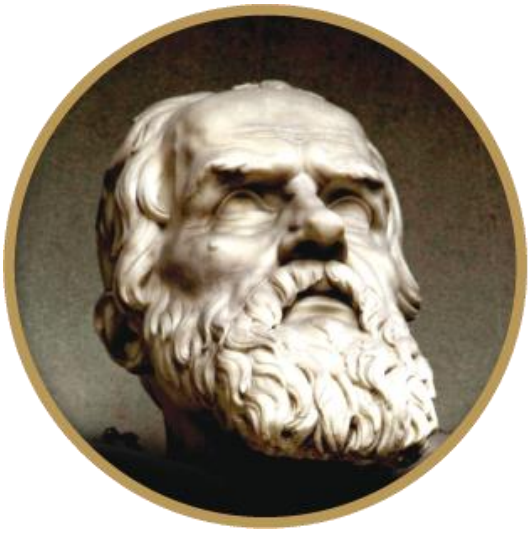
Universidad Galileo
Ciencia de Datos en Python
Sección B
Ing. Preng Bibas Solares



Proyecto final: ingeniería de Datos en Python

22003738 Franz Schubert Castillo Colocho
22005266 Doris Andrea Paz García

Guatemala abril año 2023

Alcance del proyecto y fuentes de información

El proyecto se basa en el análisis y creación de una base de datos que contenga tablas dimensionales y tabla de hechos, utilizando para su desarrollo y carga de datos las herramientas Python, AWS y PostgreSQL. Los datos por utilizar en el proyecto reflejan la situación en la Región de las Américas sobre la tendencia de casos de la viruela del mono, de los meses de marzo a noviembre año 2022, que su análisis y utilización pueda apoyar en tomar decisiones acertadas basadas en datos reales.

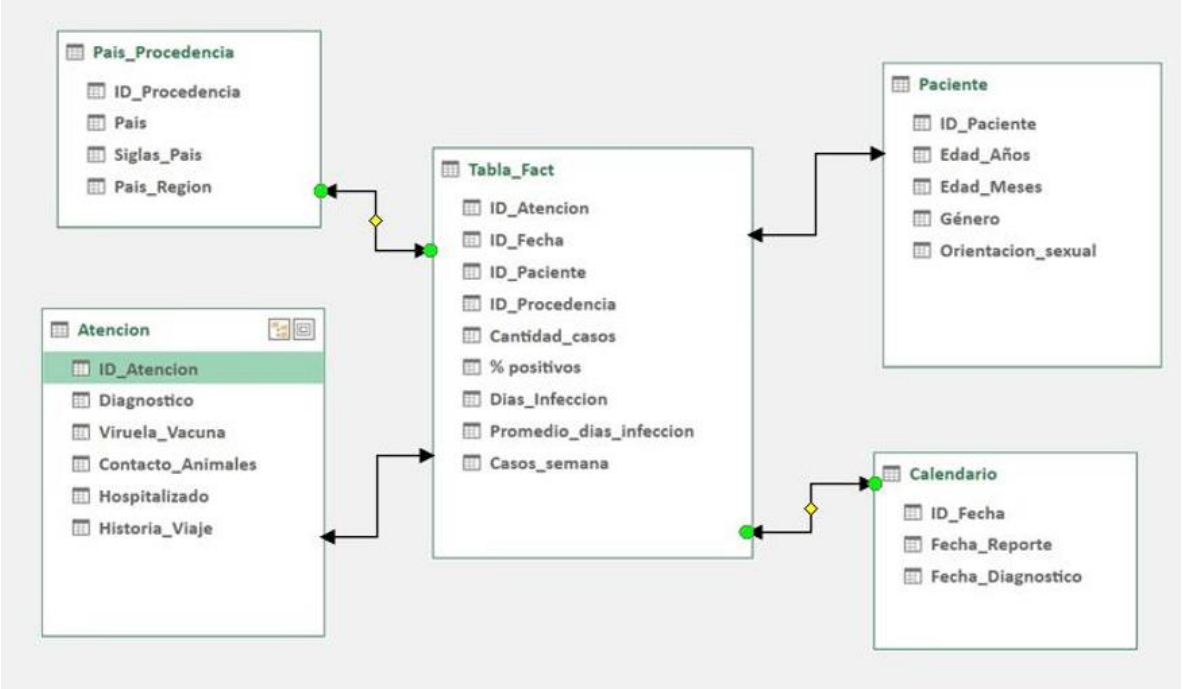
Se utilizarán cinco dataset en formato .CSV obtenidos de la página oficial de la OPS, que contiene los datos oficiales de casos de viruela del mono de los países de la Región de las Américas; Los dataset contiene la siguiente información: paciente, atención, país de procedencia y calendario.

Los datasets contiene las variables siguientes:

Variable	Tipo de dato	Descripción
Correlativo	Numérico	Secuencia correlativa de cada registro
País	Texto	País de procedencia del paciente
Siglas_Pais	Texto	Iniciales que identifican al país de procedencia
Fecha_Reporte	Fecha	Fecha en que se reportó al paciente en el informe
Pais_Region	Texto	Ciudad de procedencia del paciente
Diagnóstico	Texto	Si el paciente fue o no fue confirmado con el virus
Viruela_Vacuna	Texto	Si el paciente ya ha sido vacunado previamente contra la viruela
Fecha_Diagnostico	Fecha	Fecha en la que se le hizo el diagnóstico al paciente
Edad_Años	Numérico	Edad del paciente
Edad_Meses	Numérico	Si el paciente tiene menos de un año, la edad en meses
Contacto_Animales	Texto	Si el paciente tuvo algún contacto con animales
Género	Texto	Género del paciente
Orientación_Sexual	Texto	La orientación sexual del paciente
Hospitalizado	Texto	Si el paciente ha sido hospitalizado
Fecha_Inicio_Sintomas	Fecha	Fecha en la que el paciente empezó con los síntomas
Historia_Viaje	Texto	Si el paciente ha viajado recientemente

Modelo dimensional

De acuerdo con el análisis realizado sobre las variables disponibles en los datasets, se estableció un modelo dimensional de esquema estrella, que consta de cuatro tablas dimensionales y una tabla de hechos, tal como se muestra a continuación:



Este modelo de datos permite generar diversidad de información y responder preguntas comunes en los usuarios, como por ejemplo:

1. ¿Cuántos casos corresponden a mujeres y cuantos a hombres?
2. ¿Cuántos casos corresponden a Guatemala?
3. ¿En qué mes existe la mayor cantidad de casos?
4. ¿En qué edad predomina la mayor cantidad de casos?
5. ¿Cuántos casos reportan haber tenido historial de viaje previo a enfermarse?
6. ¿ Cuántos casos reportados hay según orientación sexual?