ADS-507 – Initial Final Team Project Proposal

Fill out this form and submit it by the end of Module 3 in Canvas.

Team Number: __2__

Team Leader/Representative: Marvin Moran

Full names of team members:

1. Marvin Moran
2. Muris Saab
3. Ravita Kartawinata

Title of your Production-Ready Data Pipeline project: COVID19 data pipeline

Short description of your project and objectives:

We will utilize 3 or more datasets to construct a data pipeline for COVID19 cases and death-toll with ETL approach that would store a final dataset which will be used to perform EDA.

Name of your selected datasets:

1. COVID-19 death by sex and age
2. United States COVID-19 Community Levels by County
3. COVID-19 Case Surveillance Public Use Data with Geography

Description of your selected datasets (data source, format, size of dataset, etc.):

COVID-19 death by sex and age : data.cdc.gov

Column Name	Description	Type
Data As Of	Date of analysis	Date & Time
Start Date	First date of data period	Date & Time
End Date	Last date of data period	Date & Time
Group	Indicator of whether data measured by Month, by Year, or Total	Plain Text
Year	Year in which death occurred	Number
Month	Month in which death occurred	Number
State	Jurisdiction of occurrence	Plain Text
Sex	Sex	Plain Text
Age Group	Age group	Plain Text
COVID-19 Deaths	Deaths involving COVID-19 (ICD-code U07.1)	Number
Total Deaths	Deaths from all causes of death	Number
Pneumonia Deaths	Pneumonia Deaths (ICD-10 codes J12.0-J18.9)	Number
Pneumonia and COVID-19 Deaths	Deaths with Pneumonia and COVID-19 (ICD-10 codes J12.0-J18.9 and U07.1)	Number
Influenza Deaths	Influenza Deaths (ICD-10 codes J09-J11)	Number
Pneumonia, Influenza, or COVID-19	Deaths with Pneumonia, Influenza, or COVID-19 (ICD-10 codes U07.1 or J09-J18.9)	Number
Footnote	Suppressed counts (1-9)	Plain Text

United States COVID-19 Community Levels by County: data.cdc.gov

Column Name	Description	Type
county	County name	Plain Text
county_fips	Federal Information Processing Standards (FIPS) five character county code	Plain Text
state	State name	Plain Text
county_population	County population (2019 Census estimate)	Number
health_service_area_number	Health Service Area (HSA) identifier	Number
health_service_area	Health Service Area (HSA) name	Plain Text
health_service_area_population	Health Service Area population (2019 Census estimate)	Number
covid_inpatient_bed_utilization	Percent of staffed inpatient beds occupied by COVID-19 patients (7-day average)	Number
covid_hospital_admissions_per_100k	New COVID-19 admissions per 100,000 population (7-day total)	Number
covid_cases_per_100k	New COVID-19 cases per 100,000 population (7-day total)	Number
covid-19_community_level	COVID-19 community level [Low, Medium, High]	Plain Text
date_updated	Date of data release	Date & Time

COVID-19 Case Surveillance Public Use Data with Geography : data.cdc.gov

Column Name	Description	Type
case_month	The earlier of month the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC	Plain Text
res_state	State of residence	Plain Text
state_fips_code	State FIPS code	Plain Text
res_county	County of residence	Plain Text
county_fips_code	County FIPS code	Plain Text
age_group	Age group [0 - 17 years; 18 - 49 years; 50 - 64 years; 65 + years; Unknown; Missing; NA, if value suppressed for privacy protection.]	Plain Text
sex	Sex [Female; Male; Other; Unknown; Missing; NA, if value suppressed for privacy protection.]	Plain Text
race	Race [American Indian/Alaska Native; Asian; Black; Multiple/Other; Native Hawaiian/Other Pacific Islander; White; Unknown; Missing; NA, if value suppressed for	Plain Text
ethnicity	Ethnicity [Hispanic; Non-Hispanic; Unknown; Missing; NA, if value suppressed for privacy protection.]	Plain Text
case_positive_specimen_interval	Weeks between earliest date and date of first positive specimen collection	Number
case_onset_interval	Weeks between earliest date and date of symptom onset.	Number
process	Under what process was the case first identified? [Clinical evaluation; Routine surveillance; Contact tracing of case patient; Multiple; Other; Unknown; Missing]	Plain Text
exposure_yn	In the 14 days prior to illness onset, did the patient have any of the following known exposures: domestic travel, international travel, cruise ship or vessel travel as a passenger or crew member, workplace, airport/airplane, adult congregate living facility (nursing, assisted living, or long-term care facility), school/university/childcare center, correctional facility, community event/mass gathering, animal with confirmed	Plain Text
current_status	What is the current status of this person? [Laboratory-confirmed case, Probable case]	Plain Text
symptom_status	What is the symptom status of this person? [Asymptomatic, Symptomatic, Unknown, Missing]	Plain Text
hosp_yn	Was the patient hospitalized? [Yes, No, Unknown, Missing]	Plain Text
icu_yn	Was the patient admitted to an intensive care unit (ICU)? [Yes, No, Unknown, Missing]	Plain Text
death_yn	Did the patient die as a result of this illness? [Yes; No; Unknown; Missing; NA, if value suppressed for privacy protection.]	Plain Text
underlying_conditions_yn	Did the patient have one or more of the underlying medical conditions and risk behaviors: diabetes mellitus, hypertension, severe obesity (BMI>40), cardiovascular disease, chronic renal disease, chronic liver disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autoimmune condition, current	Plain Text

Please provide the link for your GitHub repository here: https://github.com/Pii-USD/ADS507

How many times have your members met in the last two weeks? 2 times

List the specific contributions that each team member is providing for the Final Team Project in the table below.

- **NOTE:** ALL students on the team should contribute equally to the Final Team Project.

Team Member 1 (Marvin)	Team Member 2 (Muris)	Team Member 3 (if applicable) (Ravita)
• Team Leader • Identifying datasets to use • Extract, Transform and Load Data • Generate design document and presentation	• Identifying datasets to use • Extract, Transform and Load Data • Generate design document and presentation. • Create a GitHub	• Identifying datasets to use • Extract, Transform and Load Data • Generate design document and presentation. • Prepare the report

Comments/Roadblocks: _____

_____

References:

CDC, COVID-19 Response. (2023, May 11). *United States covid-19 community levels by county*. Centers for Disease Control and Prevention. https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-Community-Levels-by-County/3nnm-4jni/about_data

CDC, (2024, January 8). *Covid-19 case surveillance public use data with geography*. Centers for Disease Control and Prevention. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/about_data

NCHS. (2023, September 27). *Provisional covid-19 deaths by sex and age*. Centers for Disease Control and Prevention. https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku/about_data