

Project folder location

Google Cloud

Search (/) for resources, docs, products, and more

New Project

Project name * FA25-i535-agr-chronicdisease

Project ID: fa25-i535-agr-chronicdisease. It cannot be changed later.

Organization * iu.edu

Select an organization to attach it to a project. This selection can't be changed later.

Location * FA25-BL-INFO-I535

Parent organization or folder

Bucket and folder creation

Google Cloud FA25-i535-agr-chronicdisease

Cloud Storage Bucket details Go to path Refresh

Overview Buckets Monitoring Settings

Storage Intelligence Insights datasets Configuration

Marketplace Release Notes

Buckets

fa25-i535-agr-chronicdisease-lake

| Location | Storage class | Public access | Protection |
|--------------------|---------------|---------------|-------------|
| us-central1 (Iowa) | Standard | Not public | Soft Delete |

Objects Configuration Permissions Protection Lifecycle Observability New Inventory Reports Operations

Folder browser

Buckets > fa25-i535-agr-chronicdisease-lake > meta

Create folder Upload Transfer data Other services

Filter by name prefix only Filter Filter objects and folders Show Live objects only

| Name | Size | Type | Created | Storage class | Last modified |
|-------------|------|--------|---------|---------------|---------------|
| schema/ | — | Folder | — | — | — |
| validation/ | — | Folder | — | — | — |
| quarantine/ | — | Folder | — | — | — |
| raw/ | — | Folder | — | — | — |

Service account creation

The screenshot shows the Google Cloud IAM & Admin / Service accounts interface. On the left sidebar, under 'Service Accounts', several accounts are listed. In the main pane, a table displays three service accounts:

| Email | Status | Name | Description | Key ID | Actions |
|---|---------|------------------------------------|-------------|--------|---------|
| fa25-i535-agr-chronicdisease@appspot.gserviceaccount.com | Enabled | App Engine default service account | No keys | | ⋮ |
| chronicdisease-pipeline-sa@fa25-i535-agr-chronicdisease.iam.gserviceaccount.com | Enabled | chronicdisease-pipeline-sa | No keys | | ⋮ |
| 693764657658-compute@developer.gserviceaccount.com | Enabled | Default compute service account | No keys | | ⋮ |

Downloading the dataset

The screenshot shows the Google Cloud IAM & Admin / IAM page with a Cloud Shell terminal open. The terminal window displays a command-line session for downloading a dataset:

```

-bash: NLH6-WH6D-604Q-05JE: command not found
agrcloudshell:~ (fa25-i535-agr-chronicdisease)$ gcloud config get-value project
Your active configuration is: [cloudshell-6056]
fa25-i535-agr-chronicdisease
agrcloudshell:~ (fa25-i535-agr-chronicdisease)$ cd ~
agrcloudshell:~ (fa25-i535-agr-chronicdisease)$ ls
README-cloudshell.txt
agrcloudshell:~ (fa25-i535-agr-chronicdisease)$ mkdir -p chronicdisease_project && chronicdisease_project
-bash: chronicdisease_project: command not found
agrcloudshell:~ (fa25-i535-agr-chronicdisease)$ cd ~
mkdir -p cdi_project && cd cdi_project
agrcloudshell:~/cdi_project (fa25-i535-agr-chronicdisease)$ curl -L "https://data.cdc.gov/api/views/hksd-2xuw/rows.csv?accessType=DOWNLOAD" \
-o cdi_raw_full.csv
% Total    % Received =Xfer Average Speed   Time     Time Current
          Dload Upload Total Spent   Left Speed
100  84.0M  0  84.0M  0  4237k      0  --:--:-- 0:00:20  --:--:-- 4269k
agrcloudshell:~/cdi_project (fa25-i535-agr-chronicdisease)$ ls -lh cdi_raw_full.csv
-rw-rw-r-- 1 agr agc 854 Nov 25 01:07 cdi_raw_full.csv
agrcloudshell:~/cdi_project (fa25-i535-agr-chronicdisease)$ ls -lh cdi_raw_full.csv
-rw-rw-r-- 1 agr agc 854 Nov 25 01:07 cdi_raw_full.csv
agrcloudshell:~/cdi_project (fa25-i535-agr-chronicdisease)$ head -5 cdi_raw_full.csv
#CDC_2015_2019_Cancer_Mortality_by_Site_and_Ethnicity_and_Geographic_Level
Topic,Question,Response,DataValueUnit,DataValueType,DataValueAlt,DataValueFootnote,Symbol,DataValueFootnote,LowConfidenceLimit,HighConfidenceLimit,StratificationCategory1,Stratification1,StratificationCategory2,Stratification2,StratificationCategory3,Stratification3,Geolocation,LocationID,TopicID,QuestionID,ResponseID,DataValueTypeID,StratificationCategoryID1,StratificationID1,StratificationCategoryID2,StratificationID2,StratificationCategoryID3,StratificationID3
2020,2020,US,United States,BRFSS,Health Status,Recent activity limitation among adults,Number,Age-adjusted Mean,2.9,2.9,...,2.8,2.9,Sex,Female,...,59,HEA,HEA04,,AGEADJMEAN,SEX,SEXF,,,2015,2019,AR,Arkansas,US Cancer DVT,Cancer,"Invasive cancer (all sites combined), incidence",,Number,Number,9537,9537,,Sex,Male,,,...,POINT (-92.2744907429996 34.7486501240045),05,CA,N,CAN07,,N,SEX,SEXF,,,
2015,2019,CA,California,US Cancer DVT,Cancer,"Cervical cancer mortality among all females, underlying cause",,Number,Number,486,486,,,...,Overall,Overall,,,...,POINT (-120.9999953799971 37.6386401230047),06,CAN,CAN03,,NMRR,OVERALL,OVR,,,
2015,2019,CO,Colorado,US Cancer DVT,Cancer,"Invasive cancer (all sites combined), incidence",,Number,Number,2880,2880,,,...,Race/Ethnicity,Hispanic,,,...,POINT (-106.13361092099967 38.84384075,000468),08,CAN,CAN04,,NMRR,RACE,SEXF,,,
agrcloudshell:~/cdi_project (fa25-i535-agr-chronicdisease)$

```

File Structure after pre processing script

The screenshot shows the Google Cloud Storage interface. On the left, the navigation pane includes Cloud Storage, Overview, Buckets (selected), Monitoring, Settings, Storage Intelligence, Insights datasets, Configuration, Marketplace, Release Notes, and Help. The main area displays a bucket named 'fa25-i535-agr-chronicdisease'. The 'clear/' folder contains subfolders for years from 2015 to 2022. The 'meta/' folder contains 'schema/' and 'validation/'. The 'quarantine/' folder contains 'cdi/' and 'raw/'. A table lists objects: validation_20251125T011904Z.json (291 B, application/json, Nov 24, 2025). Below the table is a 'Cloud Shell' terminal window showing command-line activity related to file uploads and validation.

```

Gemini CLI is available in Cloud Shell terminal! Type gemini to try it. Learn more
Don't show again Dismiss

Uploaded out clean/cdi_clean_year_2021.csv -> gs://fa25-i535-agr-chronicdisease-lake/clean/cdi/year=2021/cdi_clean_year_2021.csv
Uploaded out clean/cdi_clean_year_2022.csv -> gs://fa25-i535-agr-chronicdisease-lake/clean/cdi/year=2022/cdi_clean_year_2022.csv
Uploaded out quarantine/cdi_quarantine_20251125T011904Z.csv -> gs://fa25-i535-agr-chronicdisease-lake/quarantine/cdi/quarantine_20251125T011904Z.csv
Uploaded validation_summary_20251125T011904Z.json -> gs://fa25-i535-agr-chronicdisease-lake/meta/validation/validation_20251125T011904Z.json
Validation summary: {'run_ts': '20251125T011904Z', 'raw_blob': 'raw/cdi_20251125.csv', 'total_rows': 309215, 'valid_rows': 209196, 'invalid_rows': 100019, 'required_fields': ['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'Topic', 'Question', 'DataValue']}
agr@cloudshell:~/cds-project (fa25-i535-agr-chronicdisease)$

```

Sample Bigquery result

The screenshot shows the Google BigQuery interface. The left sidebar shows a project named 'fa25-i535-agr-chronicdisease...' with datasets 'cdi_lake', 'cdi_clean', and 'cdi_raw_external'. The main area displays an 'Untitled...ery' query with the following SQL code:

```

SELECT
  (SELECT COUNT(*) FROM `fa25-i535-agr-chronicdisease.cdi_lake.cdi_raw_external`) AS raw_rows,
  (SELECT COUNT(*) FROM `fa25-i535-agr-chronicdisease.cdi_lake.cdi_clean`) AS clean_rows;

```

The 'Query results' section shows the output of the query:

| Row | raw_rows | clean_rows |
|-----|----------|------------|
| 1 | 309215 | 30312 |

At the bottom, the 'Job history' section is visible.

Chart 1 - Topic coverage

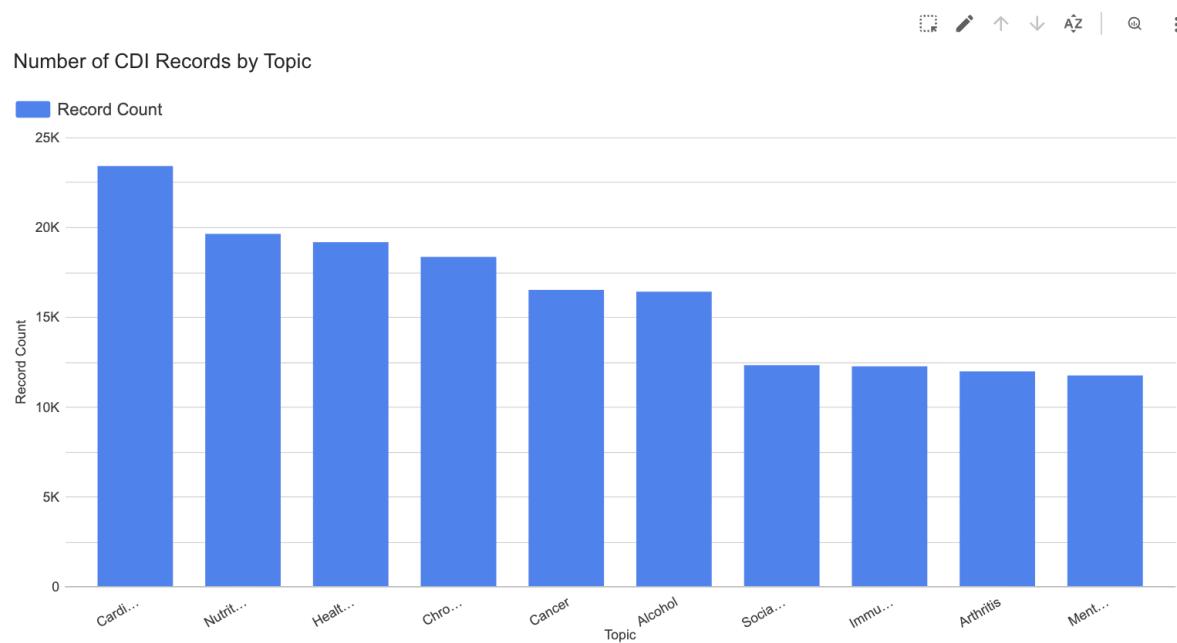


Chart 2 - Records per Year

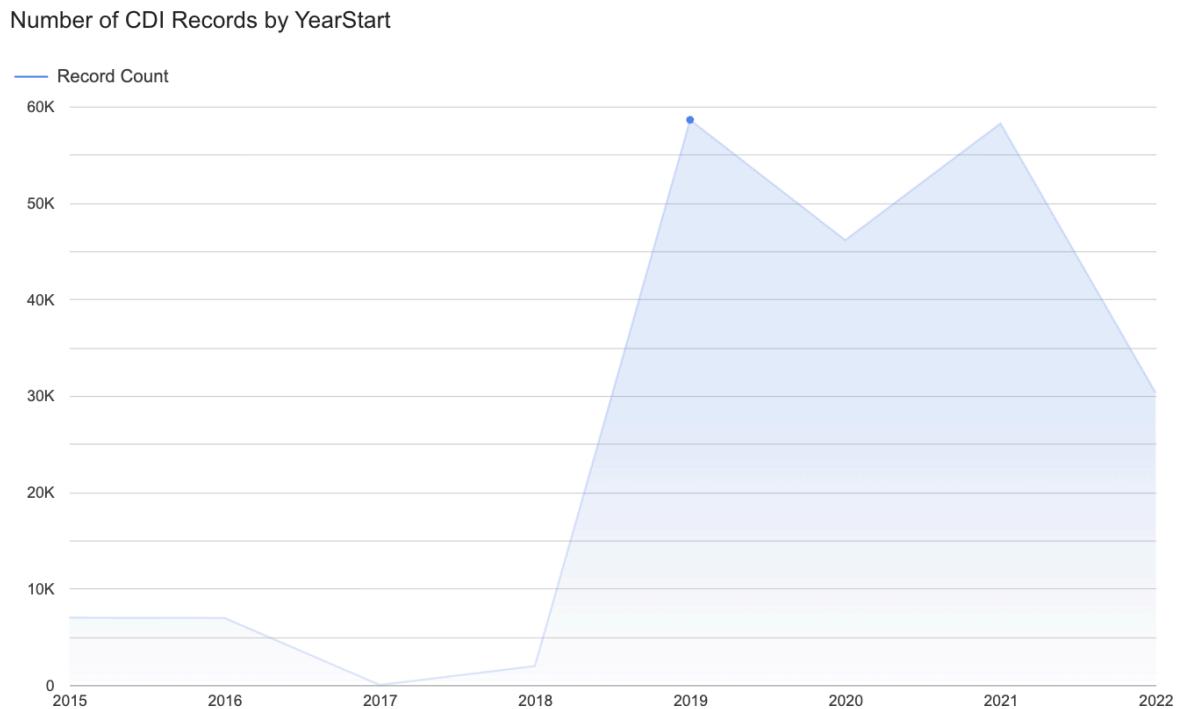


Chart 3 - Choropleth Map: Average

Average Diabetes Indicator (%) by State, 2022

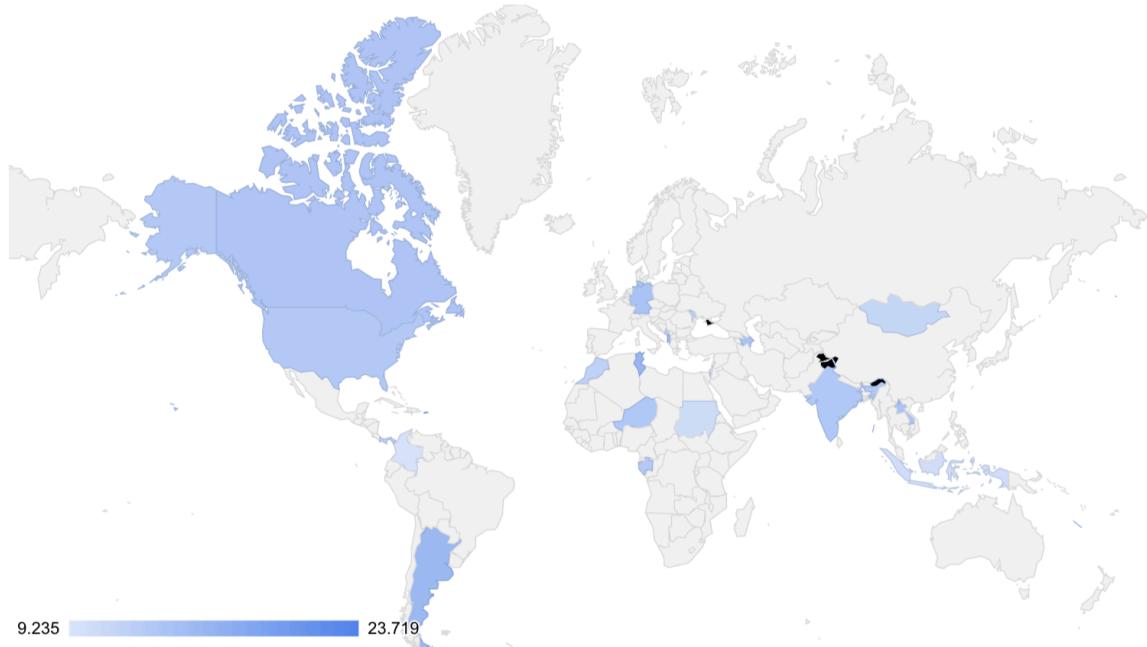


Chart 4 - Multi-Line Trend: Diabetes by Sex in One State

Average Diabetes Indicator (%) by Sex in California, Over Time

