PROMPT1:  
I am a data analyst expert for an Healthcare Insurance company ,where my domain is Payer Industry .I need to analyse different datasets which will give me information on Population health in US Healthcare from the Payer Perspective. I want to provide insight on Clinical conditions ,Access of Care ,Social Determinants of Health ,Spending level,Information on Struggling states,Value based care Impact in reducing the admissions and per capita cost ,Promote wellness,system-wise improvements ,Funding based practice ,Focus on Preventive Care ,Follow up with Primary Care Provider,Emergency Visits. Create data in such a way so that One patient can have multiple visits and One provider can treat multiple patients. Indicate FACT tables and dimension tables Rows should be 2 k in one table and accordingly adjust other datasets My data should be uncleaned which has multiple anomalies which can further be cleaned using Python. Now after analysing all the above requirements I want you to provide me dummy datasets which which will be logically correlated with all the tables in such a way so that I can able to find out above insights using that datatsets .Provide me datasets files zipped in one folder.

Update Download Healthcare Data ​and provide me new zipped folder after introducing the below updation in that dataset Diagnosis code description in separate table which could be logically joined later with Primary table In visist.csv based on visits date add discharge date which should be in future Add zip code in patients.csv Add one more table in which hospital ,clinic name of provider is mentioned ,which is logically correlated with other datasets using primary and foreign key Also add details of claim type

I want to provide insight on Clinical conditions ,Access of Care ,Social Determinants of Health ,Spending level,Information on Struggling states,Value based care Impact in reducing the admissions and per capita cost ,Promote wellness,system-wise improvements ,Funding based practice ,Focus on Preventive Care ,Follow up with Primary Care Provider,Emergency Visits. give sql queries for above conditions using this dataset.

###############################################################################

PROMPT :

I am a data analyst expert for an Healthcare Insurance company ,where my domain is Payer Industry .I need to analyse different datasets which will give me information on Population health in US Healthcare from the Payer Perspective.

I want to provide insight on Clinical conditions ,Access of Care ,Social Determinants of Health ,Spending level,Information on Struggling states,Value based care Impact in reducing the admissions and per capita cost ,Promote wellness,system-wise improvements ,Funding based practice ,Focus on Preventive Care ,Follow up with Primary Care Provider,Emergency Visits.

Create data in such a way so that One patient can have multiple visits and One provider can treat multiple patients.

Indicate FACT tables and dimension tables

Rows should be 2 k in one table and accordingly adjust other datasets

My data should be uncleaned which has multiple anomalies which can further be cleaned using Python.

Now after analysing all the above requirements I want you to provide me dummy datasets which which will be logically correlated with all the tables in such a way so that I can able to find out above insights using that datatsets .Provide me datasets files zipped in one folder.

#########################################----

Now after analysing all the above requirements I want you to provide me dummy datasets which which will be logically correlated with all the tables in such a way so that I can able to find out above insights using that datatsets .Provide me datasets files zipped in one folder .

Please provide me updated folder accordingly

-In admission data put admission date as well as discharge date

-Provide Diagnosis\_code description in another table .

-Total\_Healthcare\_Cost Total\_Admissions Total\_Readmissions Population\_Covered in Value\_based\_care.csv file should not be direct but ,Make sure we will calculating this cost using the datasets available .So update my values accordingly in files

-Also don't provide direct information on Total\_Medicaid\_Spending Total\_Medicare\_Spending Hospital\_Admissions Readmission\_Rate

But manipulate data in such a way So that later on we will calculate these values using Data Analysis.

-Create data in such a way so that One patient can have multiple visits and One provider can treat multiple patients.

-Indicate FACT tables and dimension tables

-Rows should be 2 k in one table and accordingly adjust other datasets

-My data should be uncleaned which has multiple anomalies which can further be cleaned using Python.

-In Value\_based\_care.csv and state\_health\_matrics.csv files ,i want information in such a way so that using data and after applying calculations i should be able to find out Cost\_Reduction and Hospital re-admission states ,I don't want direct values

I am a data analyst expert for a company .I need to analyze different dataset which will give me information of Population health in US Healthcare.Provide me idea what kind of data i should have to analyze end to end population health ,Starting from patient primary visit to hospitalization ,care and discharge.

To analyze population health in the U.S. healthcare system comprehensively, you should have a dataset that covers the entire patient journey, from primary visits to hospitalization, care, and discharge. Here's a breakdown of the data categories and elements you should focus on:

**1. Patient Demographics**

* **What to Collect:**
  + Age, gender, race/ethnicity
  + Socioeconomic status (income, education level)
  + Geographic location (urban/rural, ZIP code)
  + Insurance status (private, Medicaid, Medicare, uninsured)
* **Why Important:**
  + To identify disparities in access, care quality, and health outcomes among different population groups.

**2. Primary Care Visits**

* **What to Collect:**
  + Date and reason for the visit (diagnosis codes - ICD-10)
  + Preventive care services (vaccinations, screenings)
  + Lab tests and results (blood tests, cholesterol levels)
  + Prescribed medications and treatments
* **Why Important:**
  + To analyze the role of primary care in managing chronic diseases, early intervention, and preventive measures.

**3. Specialist Referrals and Follow-ups**

* **What to Collect:**
  + Referrals made by primary care providers (specialty and reason)
  + Diagnostic tests conducted (MRI, CT scans, biopsies)
  + Treatment plans and compliance
* **Why Important:**
  + To understand care coordination and whether patients are receiving timely specialist attention.

**4. Emergency Department (ED) Visits**

* **What to Collect:**
  + Date and time of ED visit
  + Chief complaint and final diagnosis
  + Triage level (urgent, non-urgent, critical)
  + Disposition (admitted, discharged, transferred)
* **Why Important:**
  + To identify patterns of emergency care usage and preventable visits.

**5. Hospitalization Data**

* **What to Collect:**
  + Admission date, reason (ICD-10 codes), and discharge date
  + Length of stay and unit (ICU, general ward)
  + Procedures performed (CPT/HCPCS codes)
  + Patient outcomes (recovery, death, readmission within 30 days)
* **Why Important:**
  + To analyze inpatient care efficiency, resource utilization, and health outcomes.

**6. Post-Discharge Care**

* **What to Collect:**
  + Discharge summary (diagnosis, treatments, medications)
  + Readmission rates and reasons
  + Follow-up visits (primary or specialist)
  + Rehabilitation and home care services
* **Why Important:**
  + To evaluate care transitions and post-discharge support for reducing readmissions.

**7. Chronic Disease Management**

* **What to Collect:**
  + Chronic conditions (diabetes, hypertension, asthma, etc.)
  + Long-term medication adherence
  + Regular monitoring and check-ups
* **Why Important:**
  + To assess the impact of care plans on managing chronic diseases.

**8. Social Determinants of Health (SDoH)**

* **What to Collect:**
  + Access to transportation, housing, and food security
  + Employment status and workplace conditions
  + Neighborhood safety and environmental factors
* **Why Important:**
  + To understand external factors affecting patient health and well-being.

**9. Behavioral Health Data**

* **What to Collect:**
  + Mental health diagnoses (depression, anxiety, etc.)
  + Substance use (alcohol, drugs)
  + Behavioral health interventions and outcomes
* **Why Important:**
  + To evaluate the integration of mental health into population health strategies.

**10. Quality and Outcomes Metrics**

* **What to Collect:**
  + Patient satisfaction scores
  + Care quality metrics (e.g., HEDIS measures)
  + Health outcomes (e.g., control of chronic conditions, mortality rates)
  + Cost of care (per visit, hospitalization, overall)
* **Why Important:**
  + To measure healthcare system performance and identify areas for improvement.

**11. Population-Level Data**

* **What to Collect:**
  + Disease prevalence and incidence rates
  + Immunization coverage
  + Trends in hospitalization and ED visits
  + Mortality rates
* **Why Important:**
  + To monitor public health trends and guide policy-making.

**12. Data Integration and Analysis Tools**

* **Key Systems:**
  + **Electronic Health Records (EHRs):** For patient-specific data
  + **Claims Data:** For insurance billing and utilization patterns
  + **Public Health Databases:** CDC, CMS, and state-level datasets
  + **Survey Data:** Behavioral Risk Factor Surveillance System (BRFSS), National Health Interview Survey (NHIS)
* **Analytical Techniques:**
  + Predictive modeling to identify high-risk populations
  + Root cause analysis for preventable hospitalizations
  + Social network analysis for care coordination

By combining data from these categories, you can gain a holistic understanding of the population's health and provide actionable recommendations to your clients on improving access, quality, and outcomes while reducing costs.

Can you provide me the Exact resources to get data for each set of category

Here’s a list of **exact resources** where you can obtain datasets for each category of population health data you are looking to analyze:

**1. Patient Demographics**

* **U.S. Census Bureau**  
  [Census Data](https://www.census.gov/data.html) – Provides demographic data by location, age, race, income, and more.
* **American Community Survey (ACS)**  
  [ACS Data](https://www.census.gov/programs-surveys/acs/) – Includes socioeconomic and demographic data.

**2. Primary Care Visits**

* **National Ambulatory Medical Care Survey (NAMCS)**  
  [NAMCS Data](https://www.cdc.gov/nchs/ahcd/ahcd_questionnaires.htm) – Data on outpatient care, including primary care visits.
* **HealthData.gov**  
  [Primary Care Data](https://healthdata.gov/) – Datasets related to healthcare services, preventive care, and chronic disease management.

**3. Specialist Referrals and Follow-Ups**

* **HCUP (Healthcare Cost and Utilization Project)**  
  [HCUP Data](https://www.hcup-us.ahrq.gov/) – Provides hospital-level and outpatient-level datasets, including referrals.
* **National Center for Health Statistics (NCHS)**  
  [Specialty Care Statistics](https://www.cdc.gov/nchs/index.htm) – Includes data on specialist visits and treatment outcomes.

**4. Emergency Department (ED) Visits**

* **National Emergency Department Sample (NEDS)**  
  [NEDS Data](https://www.hcup-us.ahrq.gov/nedsoverview.jsp) – The largest U.S. database on ED visits, part of HCUP.
* **CDC WONDER System**  
  [CDC WONDER](https://wonder.cdc.gov/) – Tracks emergency department visit trends and outcomes.

**5. Hospitalization Data**

* **Medicare Provider Utilization and Payment Data**  
  [CMS Hospital Data](https://data.cms.gov/provider-data/) – Detailed information on inpatient hospital services, including diagnoses, procedures, and costs.
* **HCUP National Inpatient Sample (NIS)**  
  [NIS Data](https://www.hcup-us.ahrq.gov/nisoverview.jsp) – Comprehensive hospitalization data for the U.S.

**6. Post-Discharge Care**

* **CMS Chronic Conditions Data Warehouse (CCW)**  
  [CCW Data](https://www2.ccwdata.org/) – Contains data on care transitions, follow-ups, and post-acute services for Medicare beneficiaries.
* **Home Health Compare**  
  [Home Health Data](https://data.cms.gov/provider-data/topics/home-health) – Data on post-discharge home healthcare providers and outcomes.

**7. Chronic Disease Management**

* **CDC Chronic Disease Indicators (CDI)**  
  [CDI Data](https://chronicdata.cdc.gov/) – Data on chronic diseases like diabetes, hypertension, and obesity at state and national levels.
* **Behavioral Risk Factor Surveillance System (BRFSS)**  
  [BRFSS Data](https://www.cdc.gov/brfss/) – Data on health-related risk behaviors and chronic diseases.

**8. Social Determinants of Health (SDoH)**

* **County Health Rankings & Roadmaps**  
  [County Health Rankings](https://www.countyhealthrankings.org/) – Includes data on transportation, housing, income inequality, and more.
* **Social Vulnerability Index (SVI)**  
  [CDC SVI](https://www.atsdr.cdc.gov/placeandhealth/svi/index.html) – Provides data on socioeconomic and environmental factors.

**9. Behavioral Health Data**

* **Substance Abuse and Mental Health Services Administration (SAMHSA)**  
  [SAMHSA Data](https://www.samhsa.gov/data) – Data on mental health and substance abuse.
* **National Survey on Drug Use and Health (NSDUH)**  
  [NSDUH Data](https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health) – Insights into substance use and behavioral health trends.

**10. Quality and Outcomes Metrics**

* **Healthcare Effectiveness Data and Information Set (HEDIS)**  
  [HEDIS Measures](https://www.ncqa.org/hedis/) – Quality performance data for healthcare providers.
* **Medicare Hospital Compare**  
  [Hospital Compare Data](https://www.medicare.gov/care-compare/) – Data on hospital performance, including readmission and mortality rates.
* **AHRQ Quality Indicators**  
  [AHRQ Data](https://www.ahrq.gov/data/qualityindicators/index.html) – Measures to evaluate healthcare quality and outcomes.

**11. Population-Level Data**

* **CDC Wide-ranging Online Data for Epidemiologic Research (CDC WONDER)**  
  [CDC WONDER](https://wonder.cdc.gov/) – Tracks disease prevalence, mortality rates, and public health trends.
* **National Health and Nutrition Examination Survey (NHANES)**  
  [NHANES Data](https://www.cdc.gov/nchs/nhanes/) – Population health data, including nutrition, chronic conditions, and biomarkers.

**12. Data Integration and Analytical Tools**

* **Kaggle Datasets**  
  [Kaggle Healthcare Datasets](https://www.kaggle.com/datasets) – Free and community-curated datasets for integration and modeling.
* **AWS Public Dataset Program**  
  [AWS Healthcare Data](https://registry.opendata.aws/) – Publicly available healthcare and genomic datasets for analysis.
* **HealthData.gov API**  
  [API Access](https://healthdata.gov/developers) – Programmatic access to public health data.

**Tips for Success:**

1. Combine multiple datasets to create a holistic view of the patient journey.
2. Use tools like **SQL, Python (Pandas/NumPy), Power BI, and Tableau** for data integration and visualization.
3. Ensure compliance with **HIPAA** and other privacy regulations when working with patient-level data.

These resources will help you access the data needed for end-to-end population health analysis, from primary care to hospitalization, care, and discharge.

COLUMNS :

Here’s a set of **dummy datasets** (tables with column names) for each category in population health analysis. These tables are designed to represent a relational database structure, with logical relationships and an Entity-Relationship (ER) model in mind.

**1. Patient Demographics Table**

**Table Name**: Patients  
**Columns**:

* PatientID (Primary Key, Unique)
* FirstName
* LastName
* DateOfBirth
* Gender
* RaceEthnicity
* Address
* City
* State
* ZIPCode
* InsuranceType (Private, Medicaid, Medicare, Uninsured)

**2. Primary Care Visits Table**

**Table Name**: PrimaryCareVisits  
**Columns**:

* VisitID (Primary Key)
* PatientID (Foreign Key from Patients)
* VisitDate
* ReasonForVisit (ICD-10 Code)
* LabTestsOrdered
* MedicationsPrescribed
* ProviderID (Foreign Key from Providers)

**3. Specialist Referrals Table**

**Table Name**: Referrals  
**Columns**:

* ReferralID (Primary Key)
* PatientID (Foreign Key from Patients)
* ReferringProviderID (Foreign Key from Providers)
* SpecialistProviderID (Foreign Key from Providers)
* ReferralDate
* ReasonForReferral (ICD-10 Code)
* FollowUpCompleted (Yes/No)

**4. Emergency Department (ED) Visits Table**

**Table Name**: EmergencyVisits  
**Columns**:

* VisitID (Primary Key)
* PatientID (Foreign Key from Patients)
* VisitDateTime
* ChiefComplaint
* TriageLevel (Critical, Urgent, Non-Urgent)
* FinalDiagnosis (ICD-10 Code)
* Disposition (Admitted, Discharged, Transferred)

**5. Hospitalizations Table**

**Table Name**: Hospitalizations  
**Columns**:

* AdmissionID (Primary Key)
* PatientID (Foreign Key from Patients)
* AdmissionDate
* DischargeDate
* LengthOfStay (Calculated: Days)
* PrimaryDiagnosis (ICD-10 Code)
* ProceduresPerformed (CPT/HCPCS Codes)
* Outcome (Recovered, Readmitted, Death)

**6. Post-Discharge Care Table**

**Table Name**: PostDischargeCare  
**Columns**:

* FollowUpID (Primary Key)
* PatientID (Foreign Key from Patients)
* AdmissionID (Foreign Key from Hospitalizations)
* FollowUpDate
* ProviderID (Foreign Key from Providers)
* CareType (Home Care, Rehabilitation, Primary Care Visit)
* Readmitted (Yes/No)

**7. Chronic Disease Management Table**

**Table Name**: ChronicConditions  
**Columns**:

* ConditionID (Primary Key)
* PatientID (Foreign Key from Patients)
* ConditionName (Diabetes, Hypertension, etc.)
* DiagnosisDate
* ManagementPlan (Lifestyle Changes, Medications, Regular Monitoring)

**8. Social Determinants of Health (SDoH) Table**

**Table Name**: SocialDeterminants  
**Columns**:

* SDoHID (Primary Key)
* PatientID (Foreign Key from Patients)
* TransportationAccess (Yes/No)
* HousingStatus (Stable, Homeless, Temporary)
* FoodSecurity (Yes/No)
* EmploymentStatus (Employed, Unemployed)
* NeighborhoodSafety (Safe, Unsafe)

**9. Behavioral Health Table**

**Table Name**: BehavioralHealth  
**Columns**:

* BehavioralHealthID (Primary Key)
* PatientID (Foreign Key from Patients)
* MentalHealthDiagnosis (Anxiety, Depression, etc.)
* SubstanceUse (Alcohol, Drugs)
* InterventionType (Counseling, Medication, Group Therapy)
* Outcome (Improved, No Change, Deteriorated)

**10. Quality Metrics Table**

**Table Name**: QualityMetrics  
**Columns**:

* MetricID (Primary Key)
* PatientID (Foreign Key from Patients)
* ProviderID (Foreign Key from Providers)
* CareQualityScore (1-100)
* PatientSatisfactionScore (1-5)
* ReadmissionWithin30Days (Yes/No)

**11. Providers Table**

**Table Name**: Providers  
**Columns**:

* ProviderID (Primary Key)
* ProviderName
* Specialty (Primary Care, Cardiology, Oncology, etc.)
* ClinicName
* City
* State
* ZIPCode

**12. Population-Level Data Table**

**Table Name**: PopulationHealthMetrics  
**Columns**:

* RegionID (Primary Key)
* RegionName (County, State, ZIP Code Level)
* PopulationSize
* ChronicDiseasePrevalenceRate (%)
* AverageLifeExpectancy (Years)
* HealthInsuranceCoverageRate (%)
* HospitalizationRate (%)

**Logical Relationships and ER Model**

1. Patients is the central table and has relationships with almost all other tables via PatientID.
2. Providers links with PrimaryCareVisits, Referrals, PostDischargeCare, and QualityMetrics via ProviderID.
3. Hospitalizations links to PostDischargeCare via AdmissionID.
4. SocialDeterminants, ChronicConditions, and BehavioralHealth are patient-level tables linked via PatientID.
5. PopulationHealthMetrics is aggregated data and does not directly link to patient-level tables but is used for regional analysis.

This schema provides the groundwork for a fully relational dataset suitable for analyzing population health end-to-end, from primary care visits to hospital discharge and follow-ups.

"BASED ON THE ABOVE ATTRIBUTES AND SCHEMA INFORMATION,PROVIDE ME DUMMY DATASETS in .csv format WITH 1 K rows WHICH IS NOT CLEANED AND REQUIRED MULTIPLE CLEANING USING PYTHON SCRIPTS; CONSIDERING FACT TABLES AND OTHER COULD BE DIMENSION TABLES WHICH ESTABLISHED LOGICAL RELATIONSHIP BETWEEN THE DATASETS.

import pandas as pd

import random

import numpy as np

from faker import Faker

fake = Faker()

# Parameters

num\_patients = 1000

num\_providers = 200

num\_visits = 1500

num\_referrals = 800

num\_emergency\_visits = 600

num\_hospitalizations = 500

num\_post\_discharge = 300

num\_chronic\_conditions = 700

num\_sdh = 900

num\_behavioral\_health = 400

num\_quality\_metrics = 1000

# Helper function to introduce unclean data

def introduce\_noise(value, prob=0.1):

return value if random.random() > prob else None

# Generate Patients Table

patients = pd.DataFrame({

"PatientID": range(1, num\_patients + 1),

"FirstName": [introduce\_noise(fake.first\_name()) for \_ in range(num\_patients)],

"LastName": [introduce\_noise(fake.last\_name()) for \_ in range(num\_patients)],

"DateOfBirth": [introduce\_noise(fake.date\_of\_birth(minimum\_age=0, maximum\_age=90)) for \_ in range(num\_patients)],

"Gender": [introduce\_noise(random.choice(["Male", "Female", "Other"])) for \_ in range(num\_patients)],

"RaceEthnicity": [introduce\_noise(fake.random\_element(elements=("White", "Black", "Asian", "Hispanic", "Other"))) for \_ in range(num\_patients)],

"Address": [introduce\_noise(fake.address().replace("\n", ", ")) for \_ in range(num\_patients)],

"City": [introduce\_noise(fake.city()) for \_ in range(num\_patients)],

"State": [introduce\_noise(fake.state()) for \_ in range(num\_patients)],

"ZIPCode": [introduce\_noise(fake.zipcode()) for \_ in range(num\_patients)],

"InsuranceType": [introduce\_noise(random.choice(["Private", "Medicaid", "Medicare", "Uninsured"])) for \_ in range(num\_patients)],

})

# Generate Providers Table

providers = pd.DataFrame({

"ProviderID": range(1, num\_providers + 1),

"ProviderName": [introduce\_noise(fake.name()) for \_ in range(num\_providers)],

"Specialty": [introduce\_noise(random.choice(["Primary Care", "Cardiology", "Oncology", "Pediatrics"])) for \_ in range(num\_providers)],

"ClinicName": [introduce\_noise(fake.company()) for \_ in range(num\_providers)],

"City": [introduce\_noise(fake.city()) for \_ in range(num\_providers)],

"State": [introduce\_noise(fake.state()) for \_ in range(num\_providers)],

"ZIPCode": [introduce\_noise(fake.zipcode()) for \_ in range(num\_providers)],

})

# Generate PrimaryCareVisits Table

primary\_care\_visits = pd.DataFrame({

"VisitID": range(1, num\_visits + 1),

"PatientID": [introduce\_noise(random.randint(1, num\_patients)) for \_ in range(num\_visits)],

"VisitDate": [introduce\_noise(fake.date\_this\_decade()) for \_ in range(num\_visits)],

"ReasonForVisit": [introduce\_noise(fake.lexify("?????")) for \_ in range(num\_visits)], # Simulating ICD-10

"LabTestsOrdered": [introduce\_noise(", ".join(fake.words(nb=random.randint(0, 3)))) for \_ in range(num\_visits)],

"MedicationsPrescribed": [introduce\_noise(", ".join(fake.words(nb=random.randint(0, 3)))) for \_ in range(num\_visits)],

"ProviderID": [introduce\_noise(random.randint(1, num\_providers)) for \_ in range(num\_visits)],

})

# Repeat similar logic for other tables...

# Example: Referrals Table

referrals = pd.DataFrame({

"ReferralID": range(1, num\_referrals + 1),

"PatientID": [introduce\_noise(random.randint(1, num\_patients)) for \_ in range(num\_referrals)],

"ReferringProviderID": [introduce\_noise(random.randint(1, num\_providers)) for \_ in range(num\_referrals)],

"SpecialistProviderID": [introduce\_noise(random.randint(1, num\_providers)) for \_ in range(num\_referrals)],

"ReferralDate": [introduce\_noise(fake.date\_this\_year()) for \_ in range(num\_referrals)],

"ReasonForReferral": [introduce\_noise(fake.lexify("?????")) for \_ in range(num\_referrals)],

"FollowUpCompleted": [introduce\_noise(random.choice(["Yes", "No"])) for \_ in range(num\_referrals)],

})

# Save CSVs

datasets = {

"Patients.csv": patients,

"Providers.csv": providers,

"PrimaryCareVisits.csv": primary\_care\_visits,

"Referrals.csv": referrals,

# Add other tables here...

}

for filename, df in datasets.items():

df.to\_csv(filename, index=False)

LOGICAL RELATIOSNHIP FOR THIS SCRIPT :  
**Key Relationships Summary**

| **Table** | **Related Table** | **Relationship Type** |
| --- | --- | --- |
| Patients | PrimaryCareVisits | One-to-Many (Patient → Visits) |
| Patients | Referrals | One-to-Many (Patient → Referrals) |
| Providers | PrimaryCareVisits | One-to-Many (Provider → Visits) |
| Providers | Referrals | Many-to-Many (Provider ↔ Referrals) |
| Patients and Providers | Geography | Comparable by City, State, ZIPCode |

**Key Relationships Summary**

| **Table** | **Primary Key** | **Foreign Key Constraints** |
| --- | --- | --- |
| **Patients** | PatientID | None |
| **Providers** | ProviderID | None |
| **PrimaryCareVisits** | VisitID | PatientID → Patients.PatientID  ProviderID → Providers.ProviderID |
| **Referrals** | ReferralID | PatientID → Patients.PatientID  ReferringProviderID → Providers.ProviderID  SpecialistProviderID → Providers.ProviderID |

**Diagram of Relationships**

Here’s a conceptual outline of the relationships:

1. **One Patient** can have **many Visits** and **many Referrals**.
2. **One Provider** can handle **many Visits** and can be involved in **many Referrals** (as a referrer or specialist).
3. The Referrals table bridges two providers (referring and specialist) and links them to a patient.

Got it! I'll help create the dummy data for the additional tables: **Chronic Conditions**, **SDH (Social Determinants of Health)**, **Behavioral Health**, and **Quality Metrics**, ensuring they have logical relationships with the previously generated tables. Here's how the relationships will be structured:

**Proposed Relationships**

1. **Chronic Conditions**:
   * Relates to Patients.csv via PatientID.
   * Each patient may have multiple chronic conditions (e.g., diabetes, hypertension).
2. **SDH (Social Determinants of Health)**:
   * Relates to Patients.csv via PatientID.
   * Includes data on social factors affecting health (e.g., housing, employment).
3. **Behavioral Health**:
   * Relates to Patients.csv via PatientID.
   * Tracks mental health diagnoses or behavioral health visits.
4. **Quality Metrics**:
   * Relates to PrimaryCareVisits.csv via VisitID.
   * Captures care quality indicators (e.g., medication adherence, follow-up rates).

**Logical Relationships**

Here’s how the relationships are established:

* **Chronic Conditions**: Links to patients (PatientID) to capture chronic diseases.
* **SDH**: Links to patients (PatientID) to record social health factors.
* **Behavioral Health**: Links to patients (PatientID) for mental health diagnoses.
* **Quality Metrics**: Links to primary care visits (VisitID) to evaluate care quality.

Let me know if you'd like to refine the relationships or add more details!

##########PROMPT1 RESULT#########################################################

Here are SQL queries to derive insights based on your updated dataset:

**1. Clinical Conditions Prevalence**

**Find the top 5 most common diagnoses and their descriptions.**

sql

CopyEdit

SELECT v.Diagnosis\_Code, d.Diagnosis\_Description, COUNT(\*) AS Total\_Visits

FROM Visits v

JOIN Diagnosis\_Codes d ON v.Diagnosis\_Code = d.Diagnosis\_Code

GROUP BY v.Diagnosis\_Code, d.Diagnosis\_Description

ORDER BY Total\_Visits DESC

LIMIT 5;

**2. Access to Care (Provider Density by State)**

**Find the number of providers per state.**

sql

CopyEdit

SELECT State, COUNT(\*) AS Total\_Providers

FROM Providers

GROUP BY State

ORDER BY Total\_Providers DESC;

**3. Social Determinants of Health Impact on Chronic Conditions**

**Analyze how social factors impact chronic conditions.**

sql

CopyEdit

SELECT

p.Has\_Chronic\_Condition,

s.Food\_Insecurity,

s.Housing\_Insecurity,

COUNT(\*) AS Total\_Patients

FROM Patients p

JOIN SDOH s ON p.Patient\_ID = s.Patient\_ID

GROUP BY p.Has\_Chronic\_Condition, s.Food\_Insecurity, s.Housing\_Insecurity

ORDER BY Total\_Patients DESC;

**4. Spending Level by State**

**Calculate the average healthcare spending per visit per state.**

sql

CopyEdit

SELECT p.State, AVG(v.Total\_Cost) AS Avg\_Spending\_Per\_Visit

FROM Visits v

JOIN Patients p ON v.Patient\_ID = p.Patient\_ID

GROUP BY p.State

ORDER BY Avg\_Spending\_Per\_Visit DESC;

**5. Struggling States with High Admission Rates**

**Identify states with the highest admission rates.**

sql

CopyEdit

SELECT p.State, COUNT(\*) AS Admission\_Count

FROM Visits v

JOIN Patients p ON v.Patient\_ID = p.Patient\_ID

WHERE v.Outcome = 'Admitted'

GROUP BY p.State

ORDER BY Admission\_Count DESC

LIMIT 5;

**6. Value-Based Care Impact (Reducing Admissions & Per Capita Cost)**

**Compare admission rates between insured vs. uninsured patients.**

sql

CopyEdit

SELECT p.Insurance\_Type, COUNT(\*) AS Admission\_Count, AVG(v.Total\_Cost) AS Avg\_Cost

FROM Visits v

JOIN Patients p ON v.Patient\_ID = p.Patient\_ID

WHERE v.Outcome = 'Admitted'

GROUP BY p.Insurance\_Type

ORDER BY Admission\_Count DESC;

**7. Wellness Promotion (Preventive vs Emergency Visits)**

**Compare the ratio of routine checkups vs emergency visits.**

sql

CopyEdit

SELECT

SUM(CASE WHEN Visit\_Type = 'Routine Checkup' THEN 1 ELSE 0 END) AS Preventive\_Visits,

SUM(CASE WHEN Visit\_Type = 'Emergency' THEN 1 ELSE 0 END) AS Emergency\_Visits

FROM Visits;

**8. System-Wide Improvements**

**Find the average wait time (proxy: time between visits per patient).**

sql

CopyEdit

SELECT Patient\_ID,

AVG(DATEDIFF(LEAD(Visit\_Date) OVER (PARTITION BY Patient\_ID ORDER BY Visit\_Date), Visit\_Date)) AS Avg\_Days\_Between\_Visits

FROM Visits;

**9. Funding-Based Practices**

**Find the spending breakdown for Medicaid, Medicare, and Private insurance.**

sql

CopyEdit

SELECT p.Insurance\_Type, SUM(v.Total\_Cost) AS Total\_Spending

FROM Visits v

JOIN Patients p ON v.Patient\_ID = p.Patient\_ID

GROUP BY p.Insurance\_Type

ORDER BY Total\_Spending DESC;

**10. Preventive Care Focus**

**Find the proportion of preventive visits per patient.**

sql

CopyEdit

SELECT Patient\_ID,

COUNT(CASE WHEN Visit\_Type = 'Routine Checkup' THEN 1 ELSE NULL END) \* 1.0 / COUNT(\*) AS Preventive\_Visit\_Ratio

FROM Visits

GROUP BY Patient\_ID

ORDER BY Preventive\_Visit\_Ratio DESC;

**11. Follow-up with Primary Care Provider**

**Find the percentage of patients who had a primary care follow-up after an emergency visit.**

sql

CopyEdit

WITH EmergencyPatients AS (

SELECT Patient\_ID, MIN(Visit\_Date) AS Emergency\_Date

FROM Visits

WHERE Visit\_Type = 'Emergency'

GROUP BY Patient\_ID

)

SELECT COUNT(DISTINCT ep.Patient\_ID) AS Patients\_With\_Followup,

COUNT(DISTINCT v.Patient\_ID) AS Total\_Emergency\_Patients,

COUNT(DISTINCT ep.Patient\_ID) \* 100.0 / COUNT(DISTINCT v.Patient\_ID) AS Followup\_Percentage

FROM EmergencyPatients ep

JOIN Visits v ON ep.Patient\_ID = v.Patient\_ID

WHERE v.Visit\_Type = 'Routine Checkup'

AND v.Visit\_Date > ep.Emergency\_Date;

**12. Emergency Visits Trend Analysis**

**Find the monthly trend of emergency visits.**

sql

CopyEdit

SELECT DATE\_FORMAT(Visit\_Date, '%Y-%m') AS Month, COUNT(\*) AS Emergency\_Visits

FROM Visits

WHERE Visit\_Type = 'Emergency'

GROUP BY Month

ORDER BY Month;

**13. Facility Utilization Analysis**

**Find which facilities have the highest number of visits.**

sql

CopyEdit

SELECT h.Facility\_Name, h.Facility\_Type, COUNT(v.Visit\_ID) AS Total\_Visits

FROM Visits v

JOIN Providers p ON v.Provider\_ID = p.Provider\_ID

JOIN Hospital\_Clinic h ON p.Provider\_ID = h.Provider\_ID

GROUP BY h.Facility\_Name, h.Facility\_Type

ORDER BY Total\_Visits DESC;

**14. Insurance Claim Type Distribution**

**Find the number of visits based on claim type.**

sql

CopyEdit

SELECT Claim\_Type, COUNT(\*) AS Claim\_Count

FROM Visits

GROUP BY Claim\_Type

ORDER BY Claim\_Count DESC;

**15. Readmission Rates**

**Find patients who were admitted more than once.**

sql

CopyEdit

SELECT Patient\_ID, COUNT(\*) AS Readmission\_Count

FROM Visits

WHERE Outcome = 'Admitted'

GROUP BY Patient\_ID

HAVING Readmission\_Count > 1

ORDER BY Readmission\_Count DESC;

These SQL queries will help analyze: ✔ Clinical Conditions  
✔ Access to Care  
✔ Social Determinants of Health  
✔ Spending Levels  
✔ Struggling States  
✔ Value-Based Care  
✔ Wellness Promotion  
✔ System-Wide Improvements  
✔ Funding-Based Practices  
✔ Preventive Care  
✔ Follow-Ups with PCPs  
✔ Emergency Visits

Let me know if you need additional queries or modifications! 🚀

**You said:**

also describe the relationship between each table

**ChatGPT said:**

ChatGPT

**Entity-Relationship Description of the Updated Dataset**

Your dataset follows a **star schema** structure where the **Fact Table** (Visits) stores transactional data, and the **Dimension Tables** provide descriptive attributes. Below is the detailed description of relationships among the tables:

**1. Fact Table: Visits**

* **Primary Key:** Visit\_ID
* **Foreign Keys:**
  + Patient\_ID → Patients.Patient\_ID
  + Provider\_ID → Providers.Provider\_ID
  + Diagnosis\_Code → Diagnosis\_Codes.Diagnosis\_Code
  + Claim\_Type → Classification of claims
* **Description:** Stores all patient visits, including visit type, diagnosis, cost, claim type, and discharge date.

**2. Dimension Table: Patients**

* **Primary Key:** Patient\_ID
* **Description:** Contains demographic and insurance-related details about patients.
* **Relationships:**
  + One **patient** can have **multiple visits** (1:M with Visits).
  + One **patient** can have one **SDOH profile** (1:1 with SDOH).

**3. Dimension Table: Providers**

* **Primary Key:** Provider\_ID
* **Description:** Contains provider details such as specialty and experience.
* **Relationships:**
  + One **provider** can treat **multiple patients** (1:M with Visits).
  + One **provider** can be affiliated with a **facility** (1:M with Hospital\_Clinic).

**4. Dimension Table: Diagnosis\_Codes**

* **Primary Key:** Diagnosis\_Code
* **Description:** Contains descriptions for each diagnosis code.
* **Relationships:**
  + One **diagnosis code** can appear in **multiple visits** (1:M with Visits).

**5. Dimension Table: SDOH (Social Determinants of Health)**

* **Primary Key:** Patient\_ID
* **Description:** Stores social factors affecting health (food insecurity, housing issues, etc.).
* **Relationships:**
  + One **patient** has **one set** of SDOH factors (1:1 with Patients).

**6. Dimension Table: Hospital\_Clinic**

* **Primary Key:** Facility\_ID
* **Foreign Key:** Provider\_ID → Providers.Provider\_ID
* **Description:** Stores details about healthcare facilities such as hospitals, clinics, and private practices.
* **Relationships:**
  + One **facility** can have **multiple providers** (1:M with Providers).
  + One **provider** is affiliated with **one facility** (M:1 with Hospital\_Clinic).

**Entity-Relationship Diagram (ERD)**

lua

CopyEdit

+-----------------+

| Patients |

|----------------|

| Patient\_ID (PK) |

| Age |

| Gender |

| Race |

| Income\_Level |

| ZIP\_Code |

| Insurance\_Type |

| Has\_Chronic\_Cond|

+-----------------+

|

| 1

|

| ∞

v

+----------------+

| Visits |

|----------------|

| Visit\_ID (PK) |

| Patient\_ID (FK)|

| Provider\_ID (FK)|

| Diagnosis\_Code (FK) |

| Claim\_Type |

| Visit\_Date |

| Discharge\_Date |

| Visit\_Type |

| Total\_Cost |

| Outcome |

+----------------+

|

+------------+------------+------------+

| | |

∞ ∞ ∞

| | |

+----------------+ +----------------+ +-----------------+

| Providers | | Diagnosis\_Codes | | SDOH |

|--------------| |----------------| |----------------|

| Provider\_ID (PK)| | Diagnosis\_Code (PK)| | Patient\_ID (FK) |

| Provider\_Type | | Diagnosis\_Desc | | Food\_Insecurity |

| State | | | | Housing\_Issues |

| Hospital\_Affil| +-------------------+ | Employment |

| Experience | | Transport\_Access|

+----------------+ +----------------+

|

| 1

|

| ∞

v

+------------------------+

| Hospital\_Clinic |

|------------------------|

| Facility\_ID (PK) |

| Provider\_ID (FK) |

| Facility\_Name |

| Facility\_Type |

| State |

| ZIP\_Code |

+------------------------+

**Key Relationships**

1. A **patient can have multiple visits** (1:M between Patients and Visits).
2. A **provider can treat multiple patients** (1:M between Providers and Visits).
3. A **visit is associated with a diagnosis code** (M:1 between Visits and Diagnosis\_Codes).
4. A **patient has one set of social determinants of health** (1:1 between Patients and SDOH).
5. A **provider is affiliated with a hospital/clinic** (1:M between Hospital\_Clinic and Providers).
6. A **visit has a claim type**, allowing for analysis on inpatient, outpatient, and emergency claims.

This schema allows for deep analysis of: ✔ Clinical Conditions  
✔ Access to Care  
✔ Social Determinants of Health  
✔ Spending Levels  
✔ Struggling States  
✔ Value-Based Care  
✔ Wellness Promotion  
✔ System-Wide Improvements  
✔ Funding-Based Practices  
✔ Preventive Care  
✔ Follow-Ups with PCPs  
✔ Emergency Visits

Would you like me to refine or expand on any part? 🚀