```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

sc = SparkContext.getOrCreate()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
dyf =
glueContext.create_dynamic_frame.from_catalog(database='database_name'
, table_name='table_name')
dyf.printSchema()
df = dyf.toDF()
df.show()
import matplotlib.pyplot as plt

# Set X-axis and Y-axis values
x = [5, 2, 8, 4, 9]
y = [10, 4, 8, 5, 2]

# Create a bar chart
plt.bar(x, y)

# Show the plot
s3output = glueContext.getSink(
  path="s3://bucket_name/folder_name",
  connection_type="s3",
  updateBehavior="UPDATE_IN_DATABASE",
  partitionKeys=[],
  compression="snappy",
  enableUpdateCatalog=True,
  transformation_ctx="s3output",
)
s3output.setCatalogInfo(
  catalogDatabase="demo", catalogTableName="populations"
)
s3output.setFormat("glueparquet")
s3output.writeFrame(DyF)
df =
spark.read.option("header",True).option("inferSchema",True).csv("s3://
group5-final-transformed-6th-aug/fully-transformed-data/part-00000-db6
60cd7-59f1-4599-b24f-1aaac466ef14-c000.csv")
df.columns
df.printSchema()
df = df.withColumnRenamed("occurrence_start_date", "occurred_date") \
       .withColumnRenamed("occurrence_start_time", "occurred_time")
from pyspark.sql.functions import to_date, col

df = df.withColumn("report_date", to_date(col("report_date"))) \
       .withColumn("occurred_date", to_date(col("occurred_date")))
df = df.drop("occurrence_end_date", "occurrence_end_time")
len(df.columns)
df1 = df
df1.printSchema()
df1.columns
df1 = df1.drop("suspect_race")
df2 = df1.drop("case_id")
df2.printSchema()
df2.select("crime_code").count()
from pyspark.sql.functions import monotonically_increasing_id

dim_crime = df2.select("crime_code", "crime_category",
"weapon_category", "source") \
               .dropDuplicates() \
               .withColumn("crime_id", monotonically_increasing_id())

dim_victim = df2.select("victim_age", "Victim_Sex",
"victim_race_group") \
                .dropDuplicates() \
                .withColumn("victim_id",
monotonically_increasing_id())

dim_suspect = df2.select("suspect_age", "Suspect_Sex",
"suspect_race_grouped") \
                 .dropDuplicates() \
                 .withColumn("suspect_id",
monotonically_increasing_id())

dim_location = df2.select("latitude", "longitude", "city",
"location_category") \
                  .dropDuplicates() \
                  .withColumn("location_id",
monotonically_increasing_id())

dim_jurisdiction = df2.select("jurisdiction") \
                      .dropDuplicates() \
                      .withColumn("jurisdiction_id",
monotonically_increasing_id())

dim_jurisdiction.count()
dim_location.count()
df2.count()
from pyspark.sql.functions import col

# Join with dim_crime
df_fact = df2.join(dim_crime, on=["crime_code", "crime_category",
"weapon_category", "source"], how="left") \
             .join(dim_victim, on=["victim_age", "Victim_Sex",
"victim_race_group"], how="left") \
             .join(dim_suspect, on=["suspect_age", "Suspect_Sex",
"suspect_race_grouped"], how="left") \
             .join(dim_location, on=["latitude", "longitude", "city",
"location_category"], how="left") \
             .join(dim_jurisdiction, on=["jurisdiction"], how="left")

fact_crime_cases = df_fact.select(
    "case_num", "report_date", "occurred_date", "occurred_time",
    "arrest_made", "domestic_incident",
    "crime_id", "victim_id", "suspect_id", "location_id",
"jurisdiction_id"
)
fact_crime_cases.show()
dim_crime.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \
    .csv("s3://group5-final-transformed-6th-aug/facts&dimension/")

fact_crime_cases.count()
dim_victim.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \

.csv("s3://group5-final-transformed-6th-aug/facts&dimension/Facts/Vict
im Dimensions/")

dim_suspect.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \
.csv("s3://group5-final-transformed-6th-aug/facts&dimension/Facts/Susp
ect Dimensions/")

dim_location.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \

.csv("s3://group5-final-transformed-6th-aug/facts&dimension/Facts/Loca
tion Dimensions/")
dim_jurisdiction.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \

.csv("s3://group5-final-transformed-6th-aug/facts&dimension/Facts/Juri
diction Dimensions/")
dim_jurisdiction.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \

.csv("s3://group5-final-transformed-6th-aug/facts&dimension/Facts/Juri
diction Dimensions/")
fact_crime_cases.coalesce(1).write \
    .mode("overwrite") \
    .option("header", True) \

.csv("s3://group5-final-transformed-6th-aug/facts&dimension/Facts/fact
Crime-cases/")
job.commit()
```