Data Analysis

Loaded the transformed datasets to AWS Athena and used Hive query language for data retrieval

Table creation:

```
CREATE EXTERNAL TABLE IF NOT EXISTS `health_care`.`health_gdp` (
  `country` string,
  `code` string,
  `year` int,
  `exp` double
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT 'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
    OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION 's3://bunny970/Datasets_updated/'
TBLPROPERTIES ('classification' = 'parquet');
```

Query to get average healthcare expenditure as a percentage of GDP for each country

Query 1

```
1  SELECT country, AVG(exp) AS avg_health_exp
2  FROM health_gdp
3  GROUP BY country;
```

Query to find the year with highest overall healthcare expenditure across all countries

Query 1

```
1  SELECT year, SUM(exp) AS total_health_exp
2  FROM health_gdp
3  GROUP BY year
4  ORDER BY total_health_exp DESC
5  LIMIT 1;
```

Query to get countries with insurance coverage less than 20%

Query 1

```
1  SELECT country, year, cov
2  FROM insurance
3  WHERE cov < 20 AND cov IS NOT NULL;
```

Query to get countries with decrease in healthcare insurance coverage from 1910 to 1975

Query 1

```sql
SELECT country, cov_1910, cov_1975, cov_1975 - cov_1910 AS coverage_change
FROM (
    SELECT country,
           MAX(CASE WHEN year = 1910 THEN cov END) AS cov_1910,
           MAX(CASE WHEN year = 1975 THEN cov END) AS cov_1975
    FROM insurance
    GROUP BY country
) subquery
WHERE cov_1975 < cov_1910;
```

Query to get top 3 regions with lowest healthcare expenditure

Query 1

```sql
SELECT Entity AS region, year, exp
FROM health_exp
WHERE code IS NULL
ORDER BY exp ASC
LIMIT 3;
```

Query to calculate yearly percentage change in health expenditure for each region

Query 1

```sql
SELECT Entity AS region, year, exp,
       (exp - LAG(exp, 1) OVER (PARTITION BY Entity ORDER BY year)) / LAG(exp, 1) OVER (PARTITION BY Entity ORDER BY year) * 100 AS percentage_change
FROM health_exp
WHERE code IS NULL
ORDER BY Entity, year;
```

Query to find the years with lowest and highest percentage of people without health insurance

Query 1

```sql
SELECT year, without_ins
FROM no_ins
ORDER BY without_ins DESC
LIMIT 1
UNION
SELECT year, without_ins
FROM no_ins
ORDER BY without_ins ASC
LIMIT 1;
```