In [1]:

```python
# Initialize Spark Session (Auto-configured for Dataproc + GCS)
from pyspark.sql import SparkSession
spark = SparkSession.builder \
    .appName("HealthDataAnalysis") \
    .getOrCreate()

# 1. Load data from GCS
gcs_path = "gs://bidhan/home/health_data.csv"
df = spark.read.csv(gcs_path, header=True, inferSchema=True)

# 2. Data Preview
print("Schema:")
df.printSchema()
print("\nFirst 5 rows:")
df.show(5)

# 3. Data Processing
from pyspark.sql.functions import col, when, avg, count

processed_df = df.withColumn(
    "age_group",
    when(col("Age") < 30, "18-29")
    .when(col("Age") < 40, "30-39")
    .when(col("Age") < 50, "40-49")
    .when(col("Age") < 60, "50-59")
    .otherwise("60+")
)

# 4. Analysis (Avg Blood Pressure by Age Group)
results = processed_df.groupBy("age_group") \
    .agg(
        avg("BloodPressure").alias("avg_blood_pressure"),
        count("*").alias("patient_count")
    ) \
    .orderBy("age_group")

# 5. Display Results
print("\nAverage Blood Pressure by Age Group:")
results.show()

# 6. Save Results to GCS
output_path = "gs://bidhan/output/health_analysis_results"
results.write \
    .mode("overwrite") \
    .option("header", "true") \
    .csv(output_path)

print(f"\nResults saved to: {output_path}")

# Optional: Plotting (requires pandas)
import matplotlib.pyplot as plt
pd_results = results.toPandas()
pd_results.plot(kind='bar', x='age_group', y='avg_blood_pressure',
               title='Average Blood Pressure by Age Group')
plt.ylabel('Blood Pressure (mmHg)')
plt.show()

# Cleanup
spark.stop()
```

```
25/07/15 22:02:32 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
```

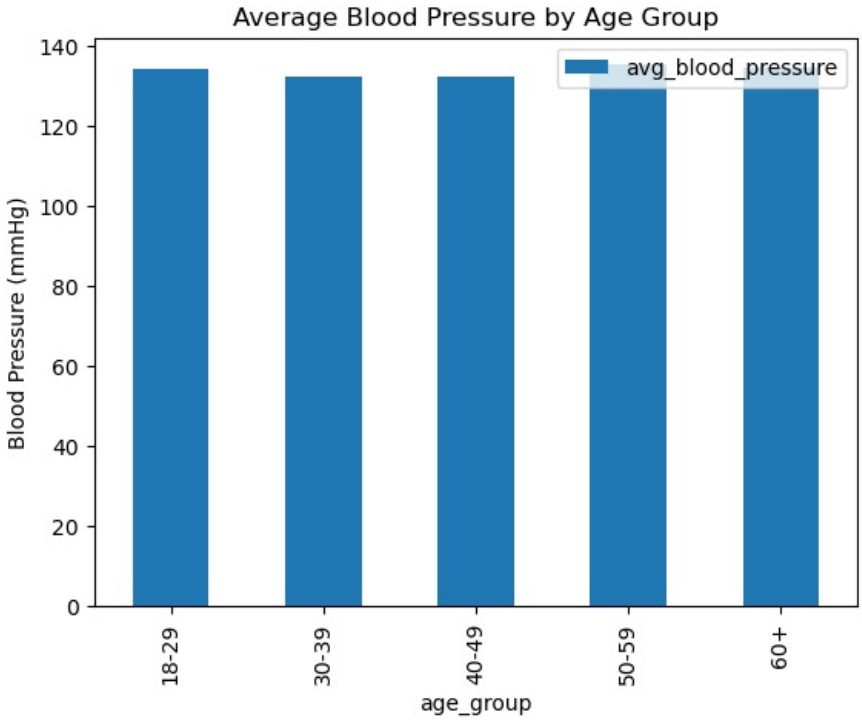
```
Schema:
root
 |-- PatientID: integer (nullable = true)
 |-- Age: integer (nullable = true)
 |-- BloodPressure: integer (nullable = true)
 |-- Cholesterol: integer (nullable = true)
 |-- BMI: double (nullable = true)
 |-- SmokingStatus: integer (nullable = true)
 |-- Disease: integer (nullable = true)


First 5 rows:
+---------+---+-------------+-----------+----------------+-------------+-------+
|PatientID|Age|BloodPressure|Cholesterol|             BMI|SmokingStatus|Disease|
+---------+---+-------------+-----------+----------------+-------------+-------+
|        1| 56|          100|        155|26.4533508153497|            0|      0|
|        2| 69|          174|        296| 23.776217016552|            1|      0|
|        3| 46|          115|        280|32.8263069003821|            1|      0|
|        4| 32|          152|        150|27.2624119044368|            0|      0|
|        5| 60|          178|        207|30.8932253037749|            1|      0|
+---------+---+-------------+-----------+----------------+-------------+-------+
only showing top 5 rows


Average Blood Pressure by Age Group:

+---------+------------------+-------------+
|age_group|avg_blood_pressure|patient_count|
+---------+------------------+-------------+
|    18-29| 134.2078651685393|          178|
|    30-39|132.13970588235293|          136|
|    40-49|132.45142857142858|          175|
|    50-59|135.22155688622755|          167|
|      60+|134.49127906976744|          344|
+---------+------------------+-------------+


Results saved to: gs://bidhan/output/health_analysis_results
```



In [ ]: