Data Analysis (Spark)

Once we have made the data ready for analysis, we have to perform below analysis queries

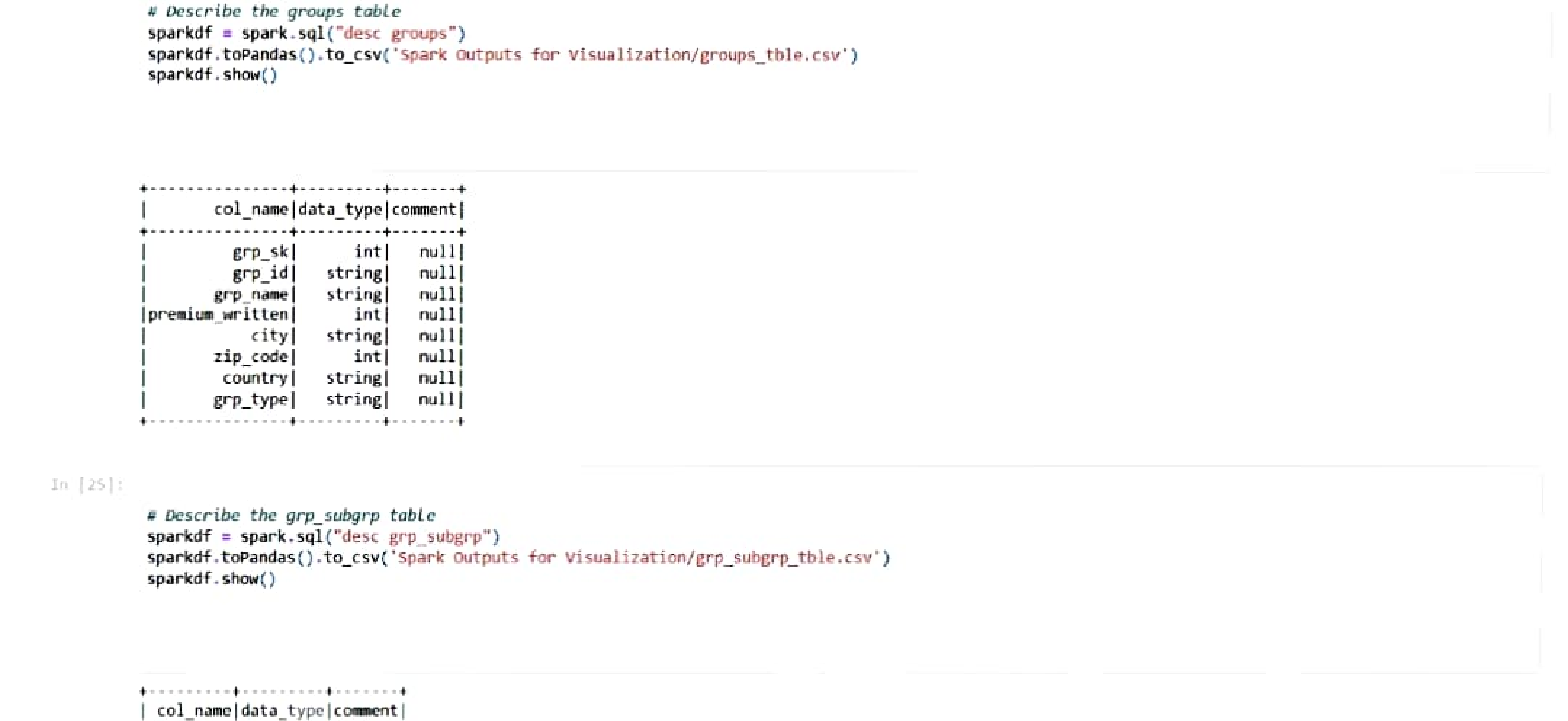
In [19]:

```
# Show all the Databases of Hive
df = spark.sql("show databases").show()
```

```
+-----------------+
|        namespace|
+-----------------+
|         airlines|
|            cog41|
|          default|
|healthcare_system|
|          sumitdb|
+-----------------+
```

In [20]:

```
# Use healthcare_system Database
spark.sql("use healthcare_system").show()
```



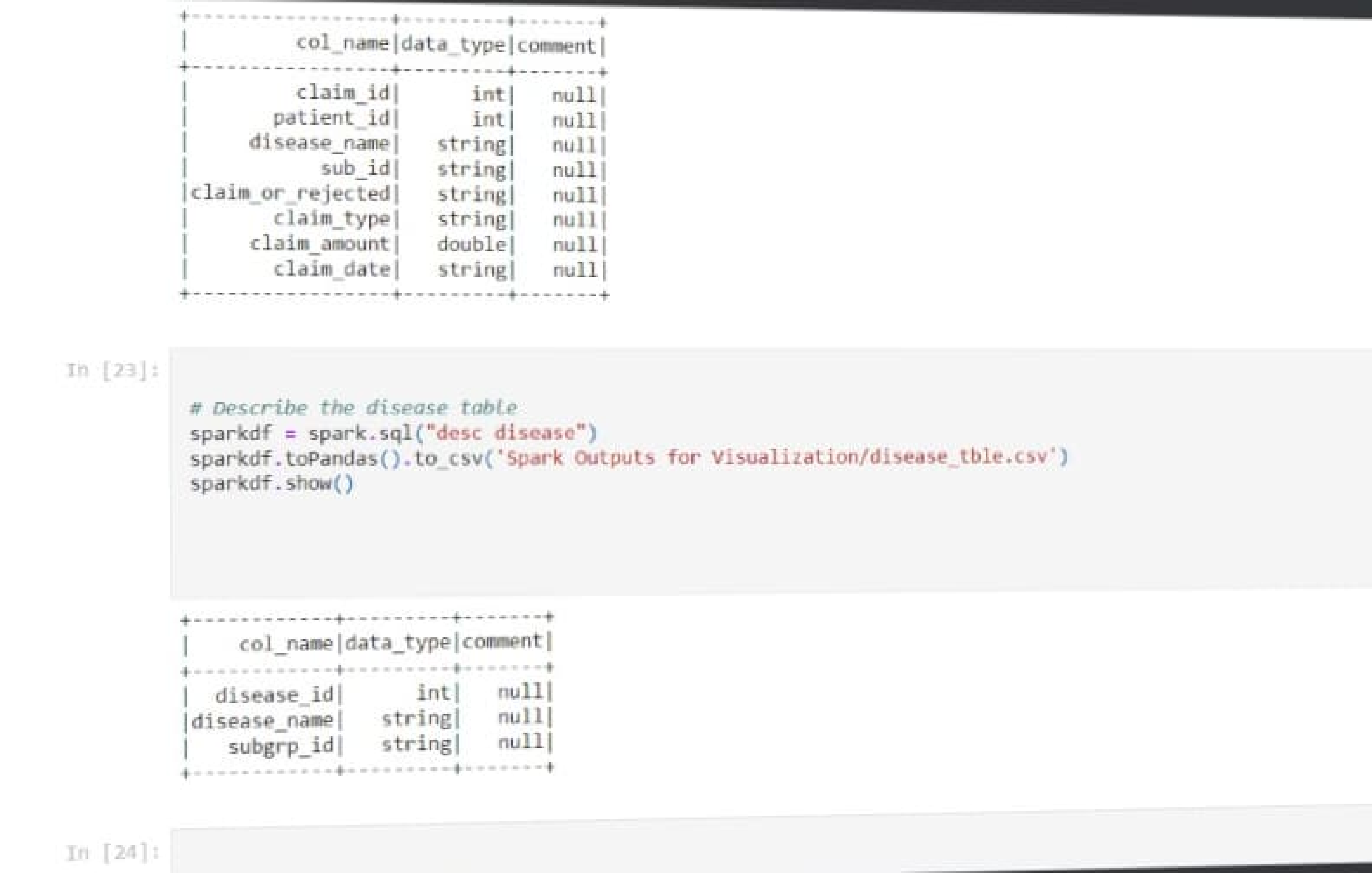
```
# Describe the groups table
sparkdf = spark.sql("desc groups")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/groups_tble.csv')
sparkdf.show()
```

```
+---------------+---------+-------+
|       col_name|data_type|comment|
+---------------+---------+-------+
|         grp_sk|      int|   null|
|         grp_id|   string|   null|
|       grp_name|   string|   null|
|premium_written|      int|   null|
|           city|   string|   null|
|       zip_code|      int|   null|
|        country|   string|   null|
|       grp_type|   string|   null|
+---------------+---------+-------+
```

In [25]:

```
# Describe the grp_subgrp table
sparkdf = spark.sql("desc grp_subgrp")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/grp_subgrp_tble.csv')
sparkdf.show()
```

```
+----------+---------+-------+
|  col_name|data_type|comment|
```



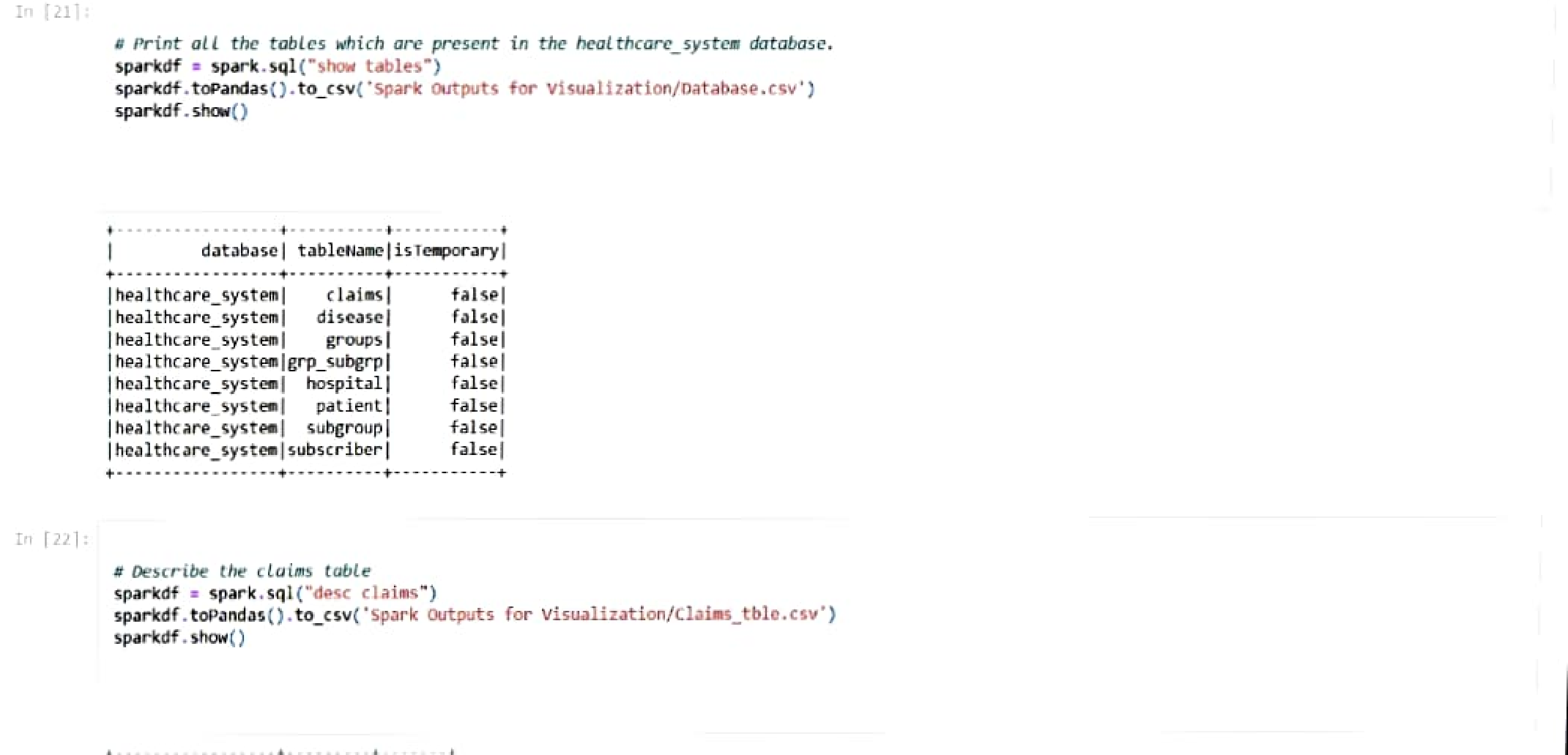
```
+----------------+---------+-------+
|        col_name|data_type|comment|
+----------------+---------+-------+
|        claim_id|      int|   null|
|      patient_id|      int|   null|
|    disease_name|   string|   null|
|          sub_id|   string|   null|
|claim_or_rejected|   string|   null|
|      claim_type|   string|   null|
|    claim_amount|   double|   null|
|      claim_date|   string|   null|
+----------------+---------+-------+
```

In [23]:

```
# Describe the disease table
sparkdf = spark.sql("desc disease")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/disease_tble.csv')
sparkdf.show()
```

```
+------------+---------+-------+
|    col_name|data_type|comment|
+------------+---------+-------+
|  disease_id|      int|   null|
|disease_name|   string|   null|
|   subgrp_id|   string|   null|
+------------+---------+-------+
```

In [24]:

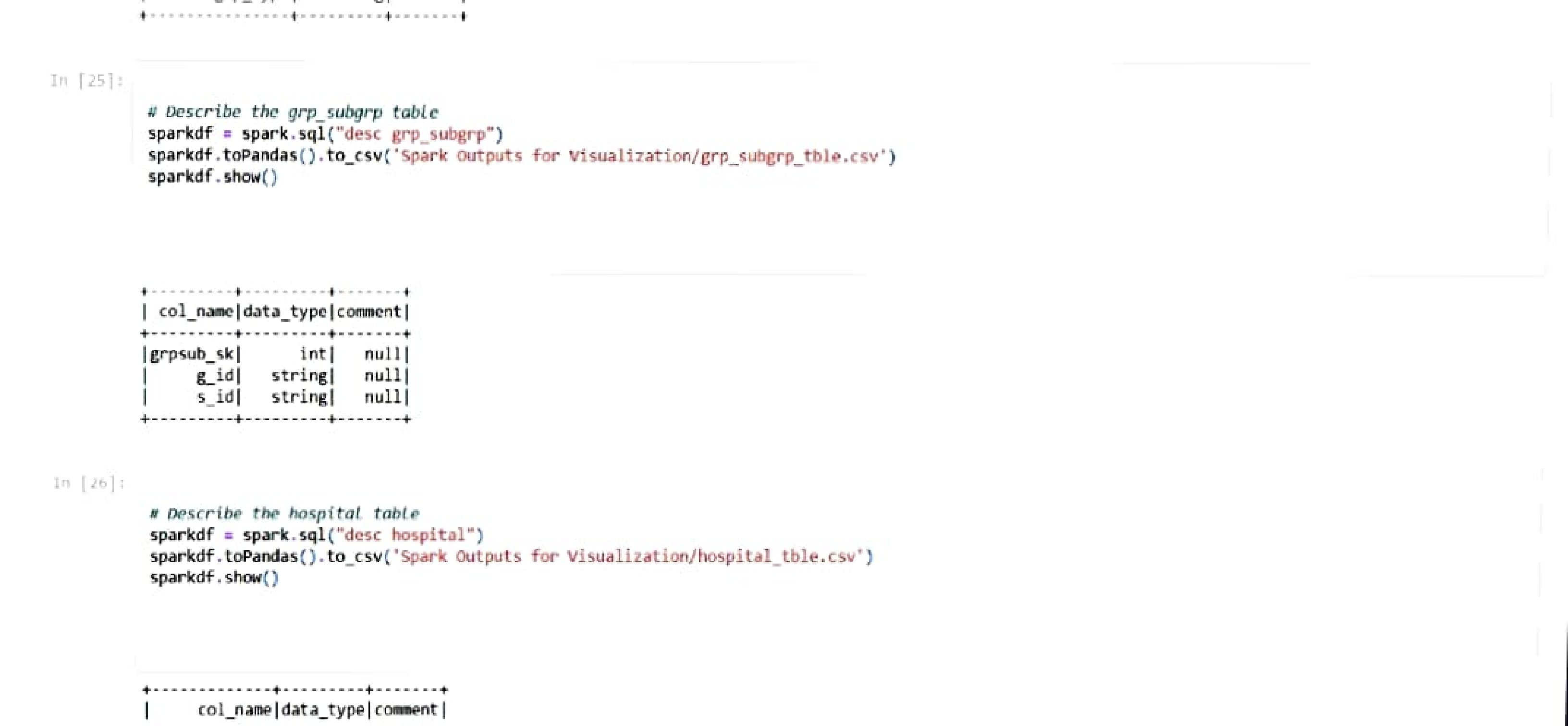


```
In [21]:
# Print all the tables which are present in the healthcare_system database.
sparkdf = spark.sql("show tables")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/Database.csv')
sparkdf.show()
```

```
+----------------+----------+-----------+
|        database| tableName|isTemporary|
+----------------+----------+-----------+
|healthcare_system|    claims|      false|
|healthcare_system|   disease|      false|
|healthcare_system|    groups|      false|
|healthcare_system|grp_subgrp|      false|
|healthcare_system|  hospital|      false|
|healthcare_system|   patient|      false|
|healthcare_system|  subgroup|      false|
|healthcare_system|subscriber|      false|
+----------------+----------+-----------+
```

```
In [22]:
# Describe the claims table
sparkdf = spark.sql("desc claims")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/Claims_tble.csv')
sparkdf.show()
```

```
+----------------+---------+-------+
```



```
+--------------+---------+-------+
```

In [25]:

```
# Describe the grp_subgrp table
sparkdf = spark.sql("desc grp_subgrp")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/grp_subgrp_tble.csv')
sparkdf.show()
```

```
+---------+---------+-------+
| col_name|data_type|comment|
+---------+---------+-------+
|grpsub_sk|      int|   null|
|     g_id|   string|   null|
|     s_id|   string|   null|
+---------+---------+-------+
```

In [26]:

```
# Describe the hospital table
sparkdf = spark.sql("desc hospital")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/hospital_tble.csv')
sparkdf.show()
```

```
+-------------+---------+-------+
|     col_name|data_type|comment|
```