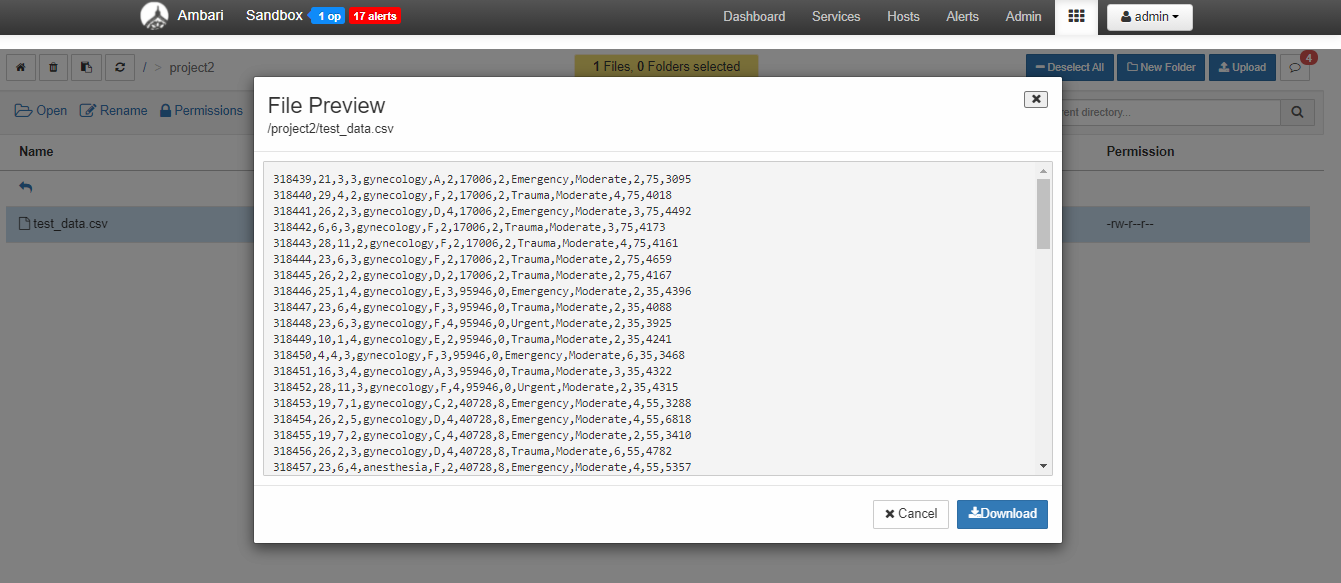
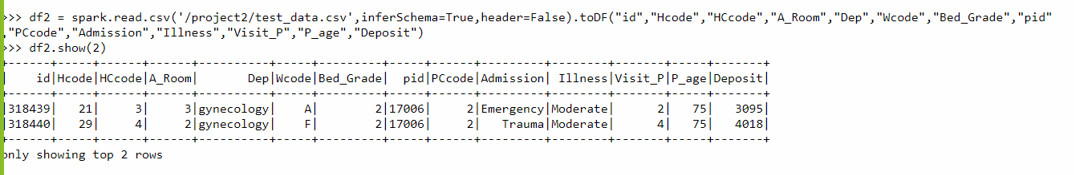
>>> import pyspark

>>> from pyspark.sql import sparksession



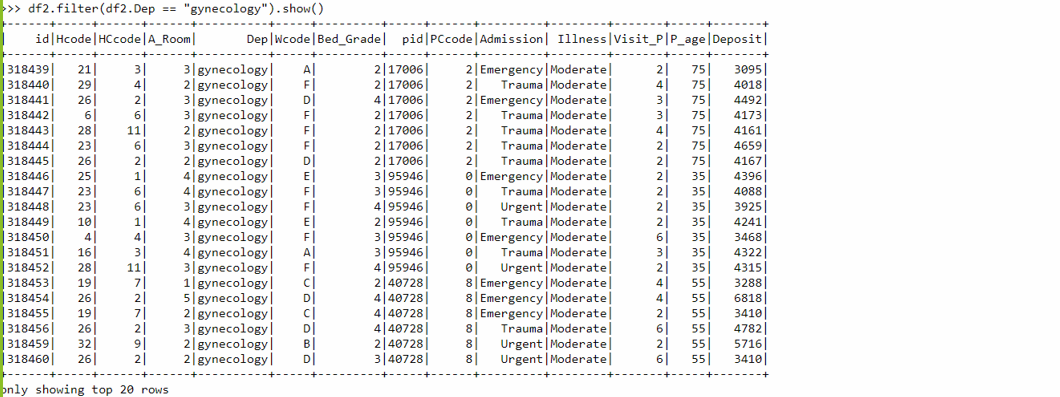
df2 = spark.read.csv('/project2/test\_data.csv',inferSchema=True,header=False).toDF("id","Hcode","HCcode","A\_Room","Dep","Wcode","Bed\_Grade","pid"

,"PCcode","Admission","Illness","Visit\_P","P\_age","Deposit")

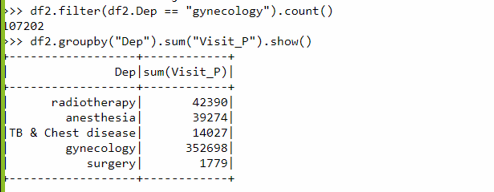


>>> df2.show(2)

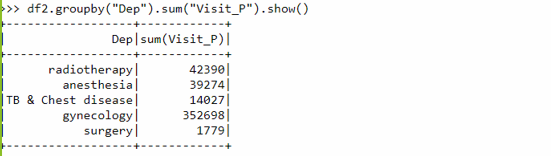
df2.filter(df2.Dep == "gynecology").show()



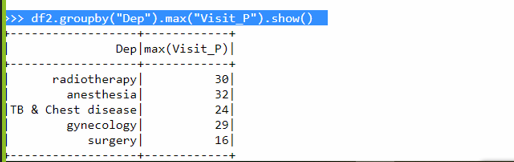
df2.filter(df2.Dep == "gynecology").count()



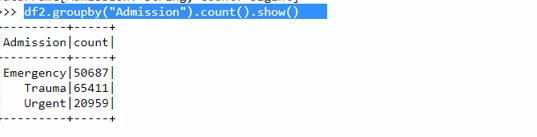
df2.groupby("Dep").sum("Visit\_P").show()



df2.groupby("Dep").max("Visit\_P").show()



df2.groupby("Admission").count().show()



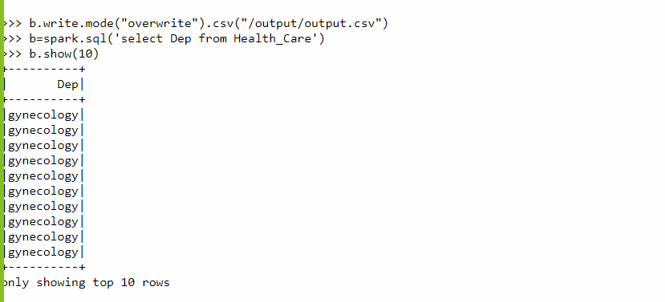
df3=df2.createOrReplaceTempView("Health\_Care")

df4=spark.sql('select \* from Health\_Care where Dep="gynecology"')

df2.write.mode("overwrite").csv("/output/output.csv")

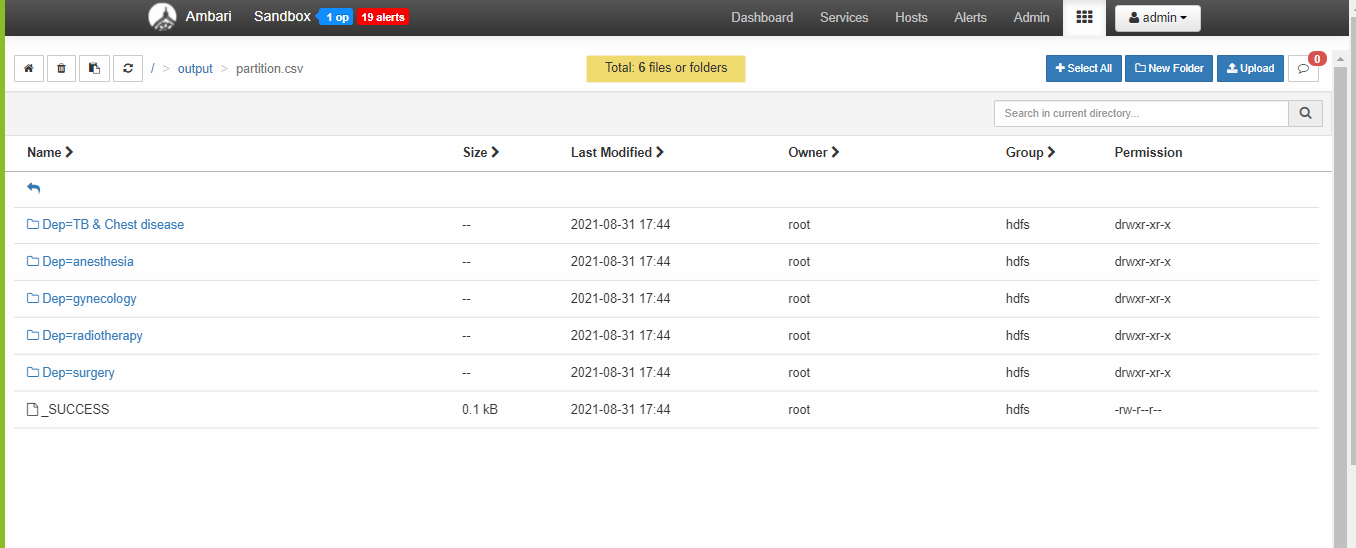
>>> b.write.mode("overwrite").csv("/output/output.csv")

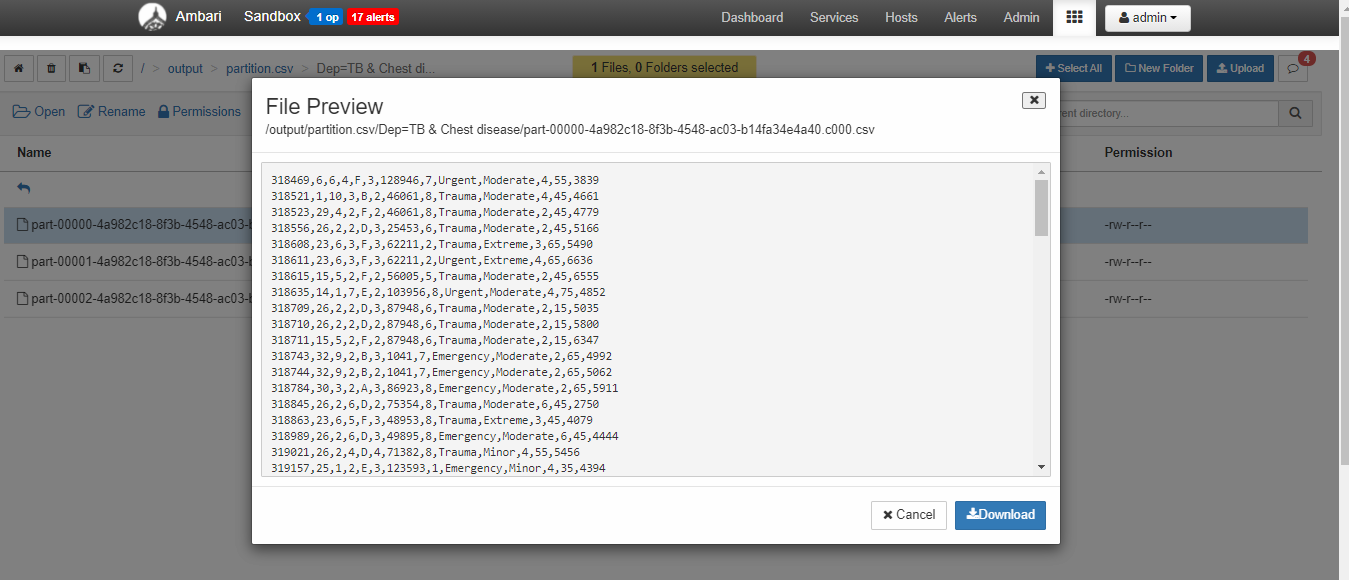
>>> b=spark.sql('select Dep from Health\_Care')



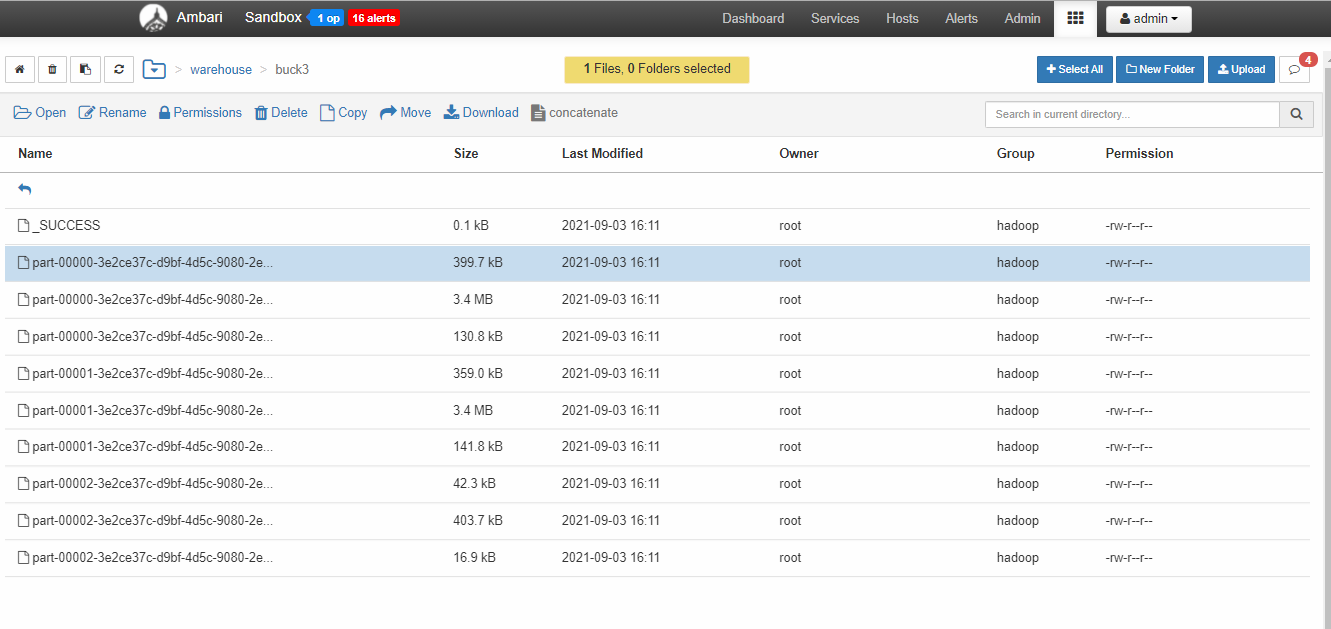
Partition :

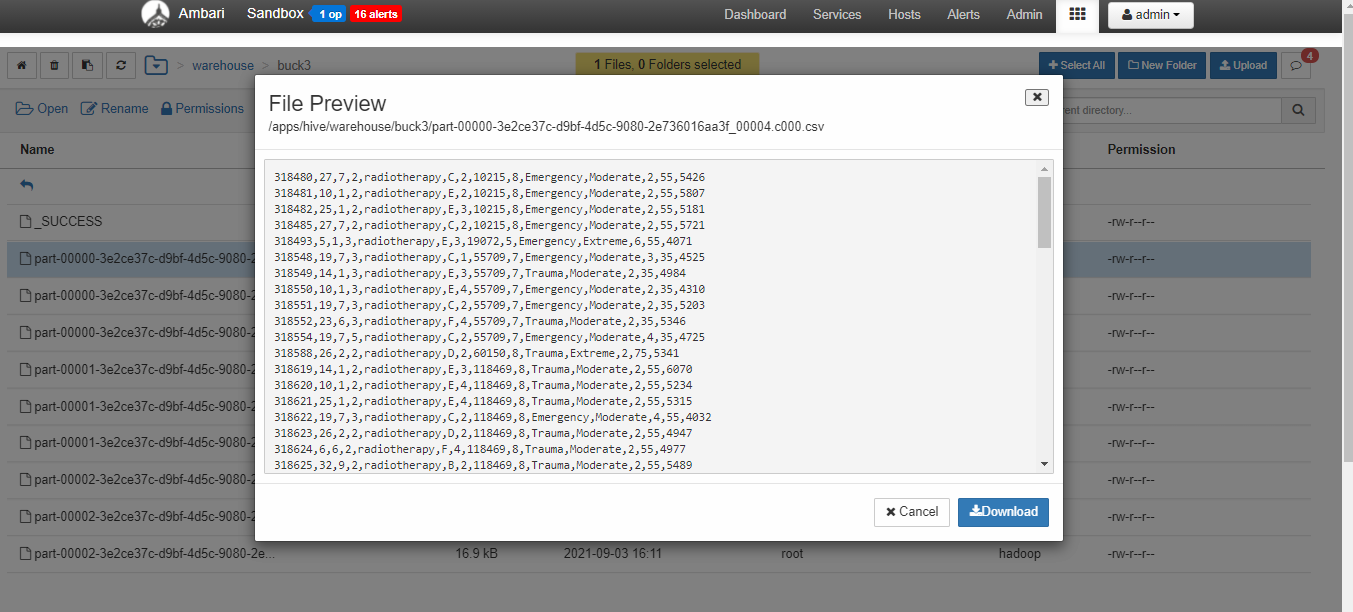
df2.write.partitionBy("Dep").mode("overwrite").csv("/output/partition.csv")

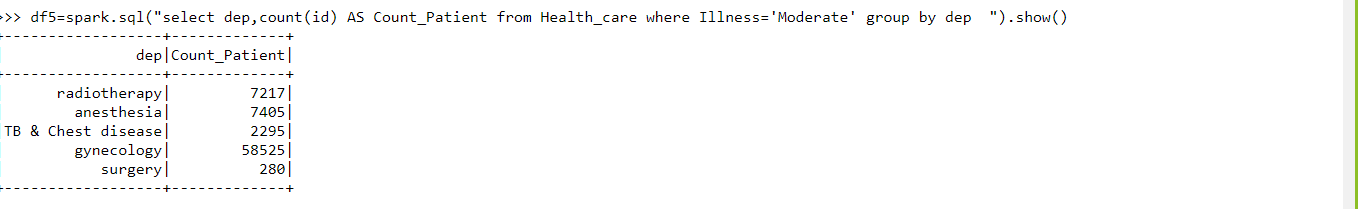




df2.write.format("csv").bucketBy(10,"Dep").saveAsTable("Buck3")





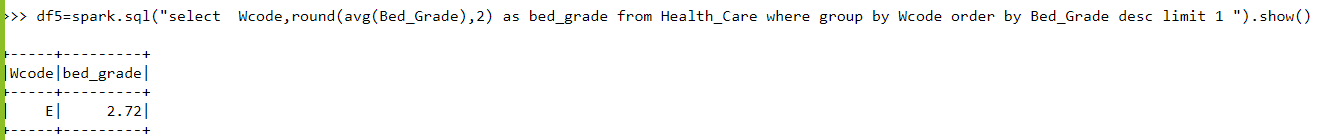
df5=spark.sql("select dep,count(id) AS Count\_Patient from Health\_care where Illness='Moderate' group by dep ").show() 

df5=spark.sql("select pid,P\_age,(select max(Deposit) max\_deposit from Health\_Care) from Health\_Care group by pid,p\_age order by pid desc").show(

)



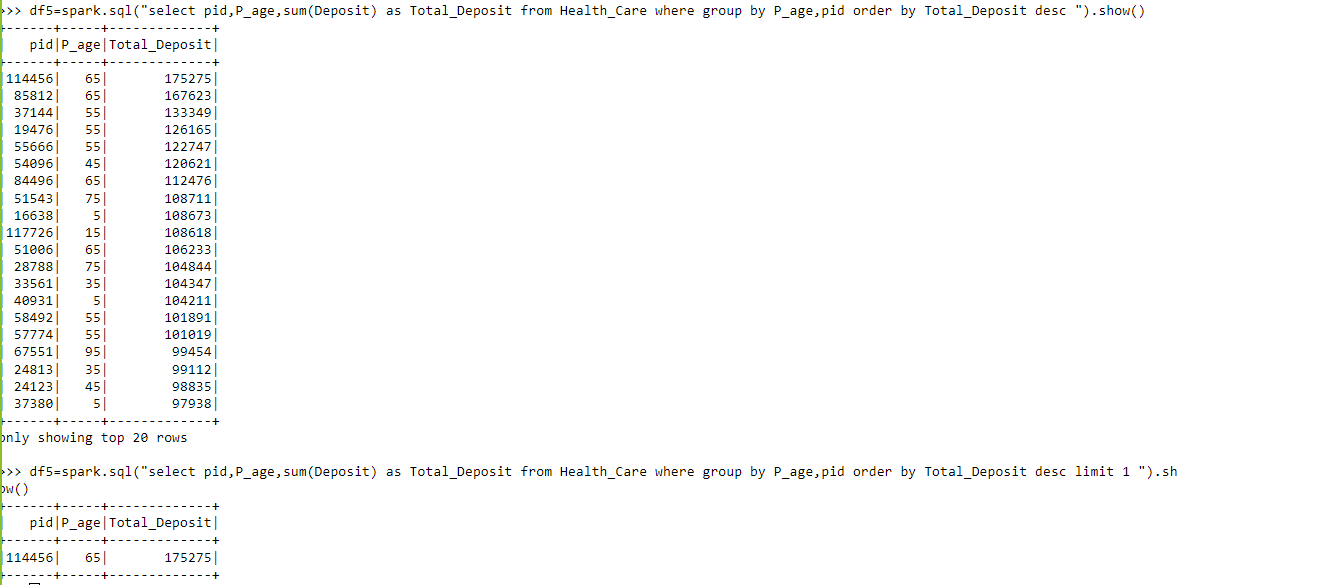
df5=spark.sql("select Wcode,round(avg(Bed\_Grade),2) as bed\_grade from Health\_Care where group by Wcode order by Bed\_Grade desc limit 1 ").show()



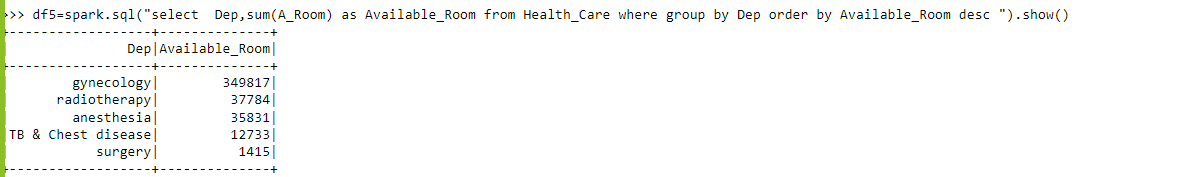
df5=spark.sql("select pid,P\_age,sum(Deposit) as Total\_Deposit from Health\_Care where group by P\_age,pid order by Total\_Deposit desc ").show()

df5=spark.sql("select pid,P\_age,sum(Deposit) as Total\_Deposit from Health\_Care where group by P\_age,pid order by Total\_Deposit desc limit 1 ").sh

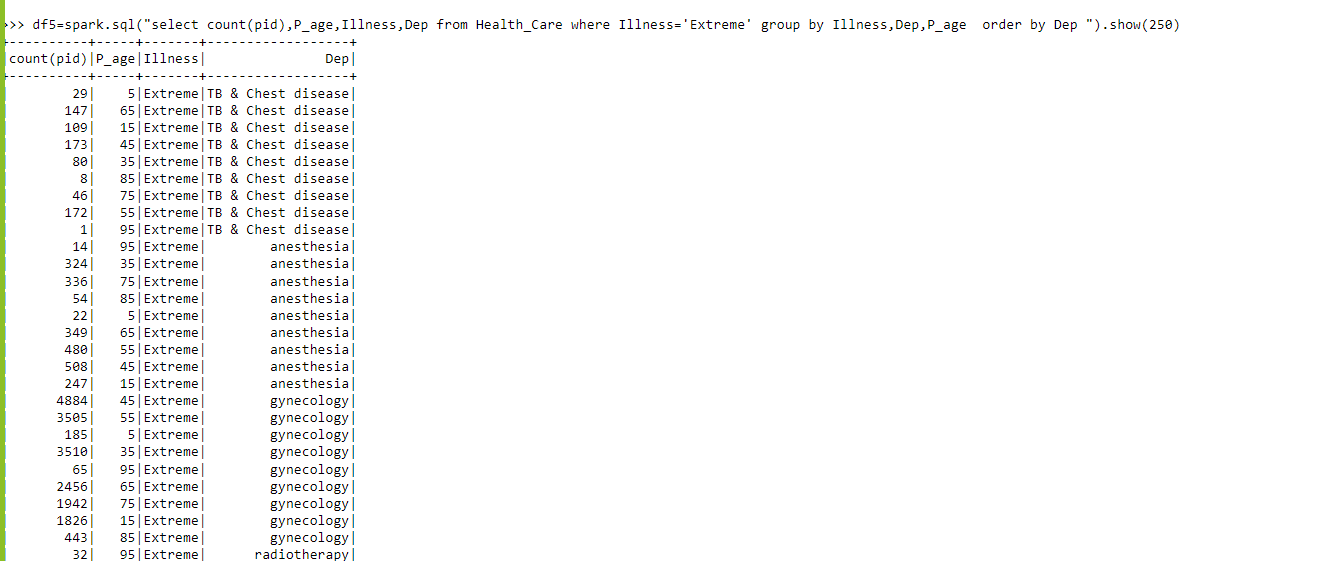
ow()



df5=spark.sql("select Dep,sum(A\_Room) as Available\_Room from Health\_Care where group by Dep order by Available\_Room desc ").show()

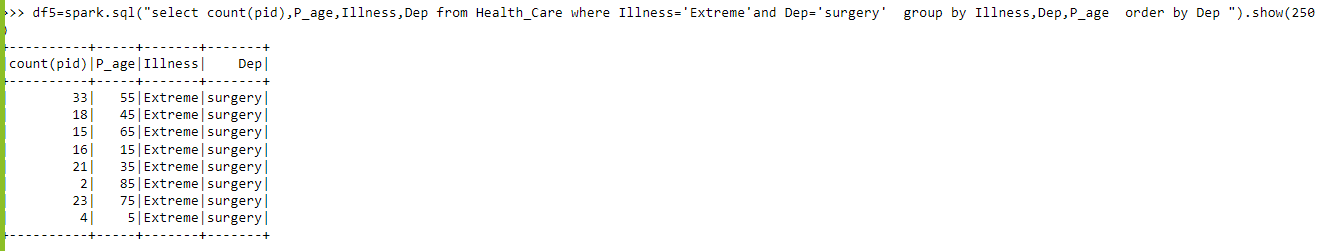


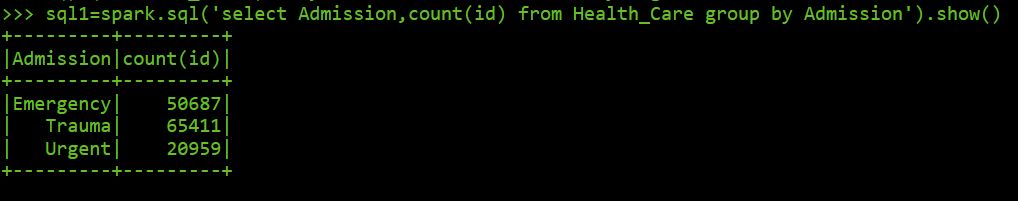
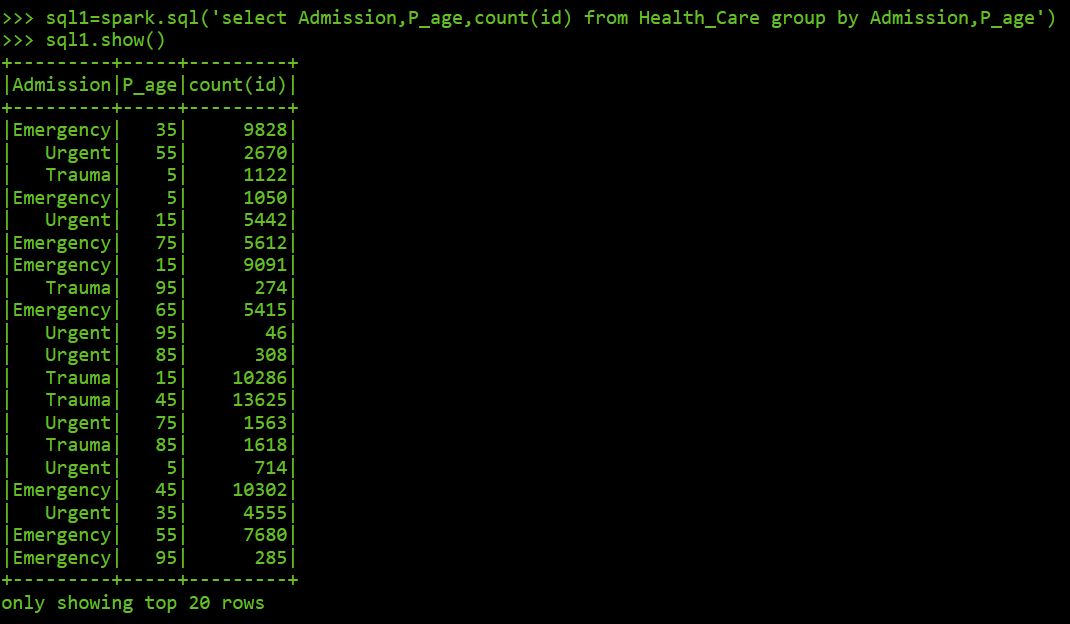
df5=spark.sql("select count(pid),P\_age,Illness,Dep from Health\_Care where Illness='Extreme' group by Illness,Dep,P\_age order by Dep ").show(250)

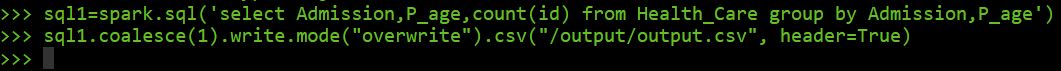


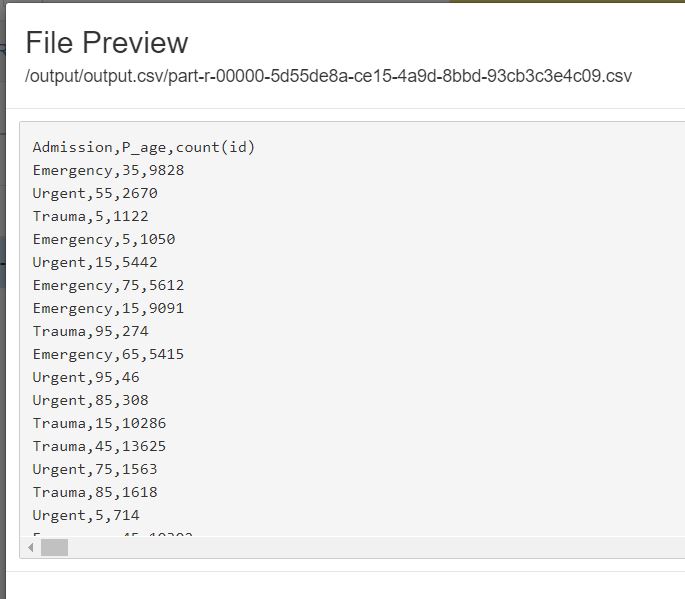
df5=spark.sql("select count(pid),P\_age,Illness,Dep from Health\_Care where Illness='Extreme'and Dep='surgery' group by Illness,Dep,P\_age order by Dep ").show(250

)



* 
* 





* 