

Case-Study Day (Hospital Data)

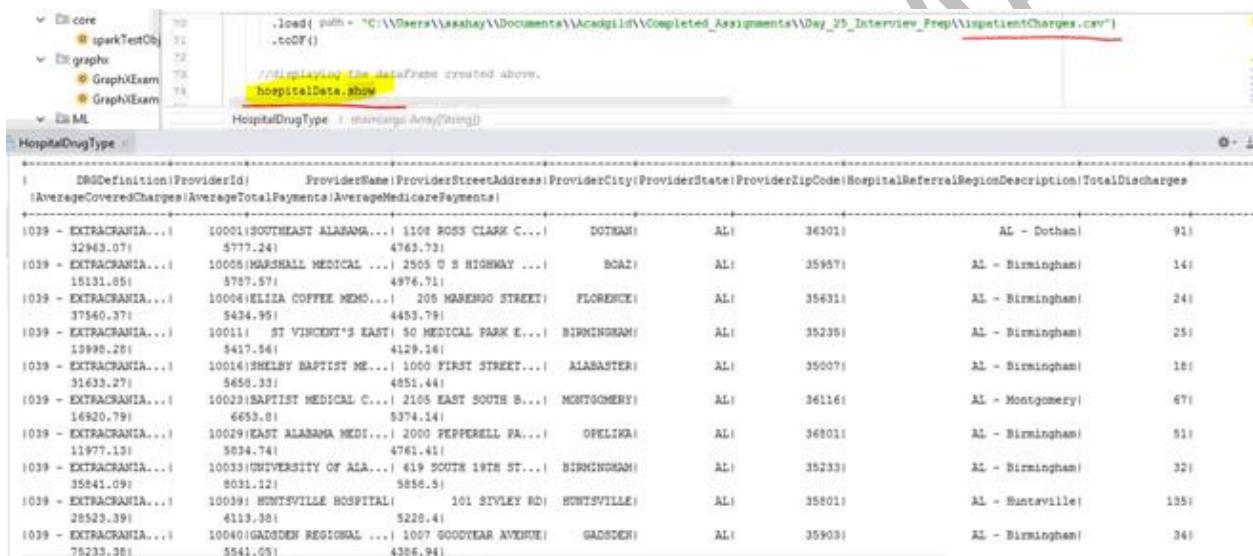
Objective 1:

Load file into spark

Ans:-

Note: Program files are properly documented for a detailed description of each instruction used within the program.

ScreenShot:-



The screenshot shows a Jupyter Notebook interface. The code cell contains the following code:

```
load(path = "C:\\Users\\ashay\\Documents\\Acadgild\\Completed_Assignments\\Day_25_Interview_Frep\\InpatientCharges.csv")
.toDF()

//displaying the dataframe created above.
hospitalData.show
```

The output cell displays a table with the following columns: DRGDefinition, ProviderId, ProviderName, ProviderStreetAddress, ProviderCity, ProviderState, ProviderZipCode, HospitalReferralRegionDescription, and TotalDischarges. The table contains 10 rows of data, each representing a different hospital and its associated charges.

DRGDefinition	ProviderId	ProviderName	ProviderStreetAddress	ProviderCity	ProviderState	ProviderZipCode	HospitalReferralRegionDescription	TotalDischarges
1039 - EXTRACRANIA...	10001	SOUTHEAST ALABAMA...	1108 ROSS CLARK C...	DOTMAN	AL	36301	AL - Dothan	91
32943.07	5777.24	4763.73						
1039 - EXTRACRANIA...	10005	MARSHALL MEDICAL ...	2505 U S HIGHWAY ...	BOAZ	AL	35957	AL - Birmingham	14
15131.85	5787.57	4976.71						
1039 - EXTRACRANIA...	10004	ELIZA COFFE MEMO...	205 MOSENGO STREET	FLORENCE	AL	35431	AL - Birmingham	24
37540.37	5434.95	4453.79						
1039 - EXTRACRANIA...	10011	ST VINCENT'S EAST	50 MEDICAL PARK E...	BIRMINGHAM	AL	35235	AL - Birmingham	25
13990.28	5417.54	4129.16						
1039 - EXTRACRANIA...	10016	SHELBY BAPTIST ME...	1000 FIRST STREET...	ALABASTER	AL	35007	AL - Birmingham	18
31633.27	5658.33	4851.44						
1039 - EXTRACRANIA...	10023	BAPTIST MEDICAL C...	2105 EAST SOUTH S...	MONTGOMERY	AL	36116	AL - Montgomery	67
14920.79	6653.81	5374.14						
1039 - EXTRACRANIA...	10029	EAST ALABAMA MED...	2000 PEPPERELL PA...	OPELIKA	AL	36801	AL - Birmingham	51
11977.13	5834.74	4761.41						
1039 - EXTRACRANIA...	10033	UNIVERSITY OF ALA...	619 SOUTH 19TH ST...	BIRMINGHAM	AL	35233	AL - Birmingham	32
35841.09	8031.12	5858.5						
1039 - EXTRACRANIA...	10039	HUNTSVILLE HOSPITAL	101 STIVLEY RD	HUNTSVILLE	AL	35801	AL - Huntsville	135
28523.39	4113.38	5228.4						
1039 - EXTRACRANIA...	10040	GADSDEN REGIONAL ...	1007 GOODYEAR AVENUE	GADSDEN	AL	35903	AL - Birmingham	34
75233.38	3541.05	4386.94						

Objective 2:

What is the average amount of **AverageCoveredCharges** per state?

Ans:

Note: Program files are properly documented for a detailed description of each instruction used within the program.

ScreenShot:-

core 81 //grouping by each state calculates average per state. Round function with 2 as a second parameter keeps only two values after decimal.
 sparkTestObj 82 val a2 = spark.sql(sqlText = "select ProviderState,round(avg(AverageCoveredCharges),2) as 'Avg_Amount/State' from hospital group by ProviderSta
 graphx 83
 GraphXExam 84 //Displaying the result
 GraphXExam 85 a2.show

HospitalDrugType > main(args: Array[String])

ProviderState	Avg_Amount/State
AZ	41200.06
SC	35842.49
LA	33085.37
MI	27894.36
WJ	66125.49
DC	40116.46
OR	27390.11
VA	29222.0
RI	29942.7
KY	24523.81
WY	28700.6
HI	27059.02
NI	24124.25
WV	41047.12
WI	26149.33
ID	25545.55
CA	67508.42
CT	31318.41
NE	31736.43
MT	22670.02

only showing top 20 rows

Find out the **AverageTotalPayments** charges per state

core 91 //grouping by each state calculates average per state. Round function with 2 as a second parameter keeps only two values after decimal.
 sparkTestObj 92 val a3 = spark.sql(sqlText = "select ProviderState,round(sum(cast(AverageTotalPayments as decimal))/cast(pow(10,2) as decimal),2) as 'Total_Paym
 graphx 93
 GraphXExam 94 //Displaying the result
 GraphXExam 95 a3.show

HospitalDrugType > main(args: Array[String])

ProviderState	Total_Payment/State
AZ	289506.28
SC	260000.46
LA	261492.69
MI	224034.81
WJ	515368.48
DC	60051.02
OR	135566.46
VA	385017.99
RI	61796.34
KY	267315.91
WY	28154.28
HI	76454.02
NI	528592.49
WV	123706.69
WI	262732.20
ID	54147.78
CA	1649942.14
CT	228559.40
NE	99102.66
MT	46819.20

only showing top 20 rows

Find out the **AverageMedicarePayments** charges per state

```

100 //grouping by each state calculates average per state. Round function with 2 as a second parameter keeps only two values after decimal.
101 val a4 = spark.sql( sqlText = "select ProviderState,round(sum(cast(AverageMedicarePayments as decimal))/cast(pow(10,2) as decimal)),2) as 'Total_Medicare_Payment/State'"
102 //Displaying the result
103 a4.show
104
HospitalDrugType > main(args: Array[String])
HospitalDrugType >
ProviderState|Total_Medicare_Payment/State|
-----+-----+
AZ|251621.67|
SC|224239.61|
LA|223626.26|
MI|194104.94|
NJ|462666.61|
DC|54571.33|
OR|117368.35|
VA|326583.58|
RI|54789.50|
KY|232011.54|
WY|23562.37|
NE|66864.97|
HI|469403.22|
WV|105146.54|
WI|226794.06|
ID|46625.52|
CA|1501628.09|
CT|203203.55|
SE|84881.92|
MT|40384.33|
-----+-----+
only showing top 20 rows

```

Objective 3:

Find out the total number of **Discharges** per state and for each disease

Ans:

Note: Program files are properly documented for a detailed description of each instruction used within the program.

ScreenShot:-

```

109 //grouping the result based on DRGDefinition,ProviderState
110 val a5 = spark.sql( sqlText = "select DRGDefinition,ProviderState, sum(TotalDischarges) as '#Discharges/State/Disease' from hospital group by DRGDefinition,ProviderState")
111 //Displaying the result
112 a5.show
113
HospitalDrugType > main(args: Array[String])
HospitalDrugType >
DRGDefinition|ProviderState|#Discharges/State/Disease|
-----+-----+
1064 - INTRACRANIA...|AR|659|
1069 - TRANSIENT I...|OR|3382|
1189 - PULMONARY E...|OR|1102|
1193 - SIMPLE PRES...|KS|1213|
1200 - RESPIRATORY...|NE|230|
1200 - RESPIRATORY...|SD|184|
1254 - OTHER VASCU...|TX|1999|
1280 - ACUTE MYOCA...|MA|1956|
1300 - PERIPHERAL ...|WV|511|
1309 - CARDIAC ARR...|OR|1205|
1378 - G.I. MEMOR...|VT|268|
1392 - ESOPHAGITIS...|KY|9822|
1394 - OTHER DISES...|RI|14|
1418 - LARYNGOSCOPI...|MA|309|
1481 - HIP & FEMUR...|MT|294|
1641 - MISC DISORD...|WV|1271|
1884 - RENAL FAILD...|IN|541|
1889 - KIDNEY & UR...|NE|92|
1812 - RED BLOOD C...|MI|161|
1897 - ALCOHOL/DRU...|RI|314|
-----+-----+

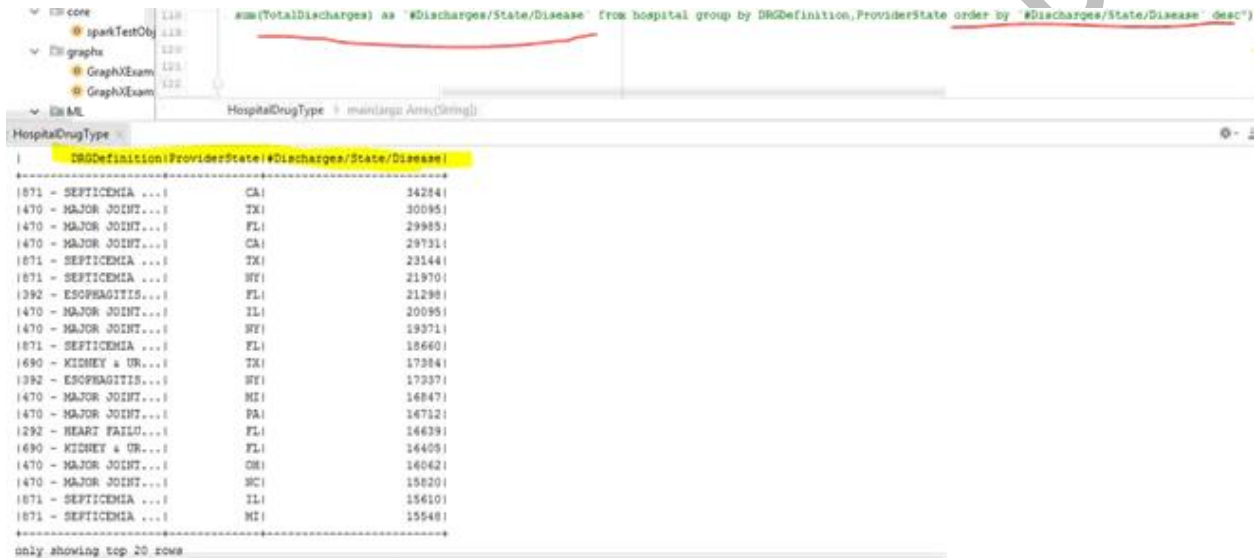
```

Sort the output in descending order of **totalDischarges**

Ans:

Note: Program files are properly documented for a detailed description of each instruction used within the program.

Screenshot:-



```
sum(TotalDischarges) as '#Discharges/State/Disease' from hospital group by DRGDefinition,ProviderState order by '#Discharges/State/Disease' desc
```

DRGDefinition	ProviderState	#Discharges/State/Disease
071 - SEPTICEMIA ...	CA	34284
470 - MAJOR JOINT...	TX	30095
470 - MAJOR JOINT...	FL	29985
470 - MAJOR JOINT...	CA	29731
071 - SEPTICEMIA ...	TX	23144
071 - SEPTICEMIA ...	NY	21970
392 - ESOPHAGITIS...	FL	21298
470 - MAJOR JOINT...	IL	20095
470 - MAJOR JOINT...	NY	19371
071 - SEPTICEMIA ...	FL	18660
690 - KIDNEY & UR...	TX	17384
392 - ESOPHAGITIS...	NY	17337
470 - MAJOR JOINT...	ME	16847
470 - MAJOR JOINT...	PA	16712
292 - HEART FAILU...	FL	16639
690 - KIDNEY & UR...	FL	16405
470 - MAJOR JOINT...	OR	16062
470 - MAJOR JOINT...	NC	15820
071 - SEPTICEMIA ...	IL	15610
071 - SEPTICEMIA ...	NE	15548

only showing top 20 rows