from os import truncate  
  
from pyspark.sql import SparkSession, Window  
from pyspark.sql.functions import col, explode, count, size, split, lit, concat, when, isnan, \  
 countDistinct, sum, desc, dense\_rank  
import pyspark.sql.functions as F  
  
spark = SparkSession.builder.master(**"local"**).appName(**"city\_report"**).getOrCreate()  
  
presDF = spark.read.format(**"csv"**).option(**"header"**,**"true"**).option(**"inferSchema"**,**"true"**) \  
 .load(**"C:**\\**Users**\\**zakeerm**\\**Downloads**\\**PysparkLearning**\\**resources**\\**USA\_Presc\_Medicare\_Data\_12021.csv"**)  
  
presDF.printSchema()  
*#df1.limit(10).show(truncate=False)*presDF.show()  
  
print(**"Clean Prescriber Data"**)  
  
cleansedDF = presDF.select(**"npi"**,**"nppes\_provider\_last\_org\_name"**,**"nppes\_provider\_first\_name"**, **"nppes\_provider\_city"**,  
 **"nppes\_provider\_state"**, **"specialty\_description"**,**"drug\_name"**, **"total\_claim\_count"**, **"total\_day\_supply"**,  
 **"total\_drug\_cost"**)  
cleansedDF.show()  
  
print(**"Rename the above fields to shorter names"**)  
colRenameDF = cleansedDF.withColumnRenamed(**"nppes\_provider\_last\_org\_name"**, **"last\_name"**) \  
 .withColumnRenamed(**"nppes\_provider\_first\_name"**, **"first\_name"**) \  
 .withColumnRenamed(**"nppes\_provider\_city"**, **"city"**) \  
 .withColumnRenamed(**"nppes\_provider\_state"**, **"state"**) \  
 .withColumnRenamed(**"specialty\_description"**, **"specialty"**) \  
 .withColumnRenamed(**"drug\_name"**, **"drug"**) \  
 .withColumnRenamed(**"total\_claim\_count"**, **"claim\_count"**) \  
 .withColumnRenamed(**"total\_day\_supply"**, **"day\_supply"**) \  
 .withColumnRenamed(**"total\_drug\_cost"**, **"drug\_cost"**)  
  
print(**"Add a Country Field 'USA' to the above data"**)  
addCountryDF = colRenameDF.withColumn(**"country"**,lit(**"USA"**))  
addCountryDF.show()  
  
print(**"Clean the “years\_of\_exp” to extract only the numbers."**)  
cleaneYeardDF1 = presDF.withColumn(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **"="**, **""**))\  
 .withColumn(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **"\.0"**, **""**))\  
 .withColumn(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **" "**, **""**))  
cleaneYeardDF1.show()  
cleaneYeardDF1.printSchema()  
  
print(**"Convert the years\_of\_exp field to integer"**)  
convertDF = cleaneYeardDF1.withColumn(**"years\_of\_exp"**, cleaneYeardDF1[**"years\_of\_exp"**].cast(**"int"**))  
convertDF.show()  
convertDF.printSchema()  
  
print(**"Combine First Name and Last Name into a single field and remove the individual columns"**)  
combineDF = convertDF.withColumn(**"full\_name"**, F.concat\_ws(**" "**, convertDF.nppes\_provider\_first\_name, convertDF.nppes\_provider\_last\_org\_name))  
combineDF1 = combineDF.select(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **"="**, **""**)),(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **"\.0"**, **""**))  
(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **" "**, **""**))  
*#combineDF1 = combineDF.drop("nppes\_provider\_first\_name", "nppes\_provider\_last\_org\_name")*combineDF = convertDF.withColumn(**"full\_name"**, F.concat\_ws(**" "**, convertDF.nppes\_provider\_first\_name, convertDF.nppes\_provider\_last\_org\_name))  
combineDF = combineDF.drop(**"nppes\_provider\_first\_name"**, **"nppes\_provider\_last\_org\_name"**)  
combineDF = combineDF.withColumn(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **"="**, **""**))  
combineDF = combineDF.withColumn(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **"\.0"**, **""**))  
combineDF = combineDF.withColumn(**"years\_of\_exp"**, F.regexp\_replace(**"years\_of\_exp"**, **" "**, **""**))  
combineDF1 = combineDF.select(**"years\_of\_exp"**, **"full\_name"**)  
combineDF1.show()  
  
print(**"Count the number of null values for each column"**)  
null\_counts = combineDF1.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in combineDF1.columns])  
null\_counts.show()  
  
print(**"clean all the Null/Nan Values"**)  
presCityDF = cleaneYeardDF1.dropna(subset=[**"npi"**, **"drug\_name"**])  
presCityDF.show()  
presCityDF.printSchema()  
  
print(**"Calculate the number of distinct Prescribers assigned for each City."**)  
joined\_df1 = presDF.groupBy(**"nppes\_provider\_state"**,**"nppes\_provider\_city"**).agg(countDistinct(**"npi"**).alias(**"distinct\_prescribers"**),  
 sum(**"total\_claim\_count"**).alias(**"total\_trx\_cnt"**))  
  
joined\_df1.show()  
  
print(**"Calculate total total\_claim\_count prescribed for each city."**)  
result\_df2 = presDF.groupBy(**"nppes\_provider\_city"**).agg(sum(**"total\_claim\_count"**).alias(**"total\_claim\_count"**))  
result\_df2.show()  
  
print(**"Do not report a city in the final report if no prescriber is assigned to it."**)  
  
joined\_df2 = presDF.groupBy(**"nppes\_provider\_state"**, **"nppes\_provider\_city"**).agg(  
 countDistinct(**"npi"**).alias(**"distinct\_prescribers"**),  
 sum(**"total\_claim\_count"**).alias(**"total\_trx\_cnt"**))  
  
joined\_df2.show()  
  
*# Convert the result to a DataFrame*result\_df = joined\_df2.select(**"nppes\_provider\_city"**, **"distinct\_prescribers"**).filter(**"distinct\_prescribers > 0"**)  
  
*# Show the result*result\_df.show()  
  
print(**"Top 5 Prescribers with highest total\_claim\_count per each state."**)  
window\_spec = Window.partitionBy(**"nppes\_provider\_state"**).orderBy(col(**"total\_claim\_count"**).desc())  
  
*# Add a new column 'rnk' using row\_number over the defined window*ranked\_pres\_df = presCityDF.withColumn(**"dense\_rank"**, F.dense\_rank().over(window\_spec))  
  
*# Filter for rows where the rank is less than or equal to 5*result\_df = ranked\_pres\_df.filter(col(**"dense\_rank"**) <= 5)  
*# Show the result*result\_df.show()  
result\_df.printSchema()  
  
print(**"Consider the prescribers only from 20 to 50 years of experience"**)  
result\_df.createOrReplaceTempView(**"result\_table"**)  
  
finalDF = result\_df.select(**"npi"**, **"nppes\_provider\_state"**, **"years\_of\_exp"**, **"total\_claim\_count"**, **"total\_day\_supply"**, **"total\_drug\_cost"**) \  
 .filter((result\_df[**'years\_of\_exp'**] >= 20) & (result\_df[**'years\_of\_exp'**] <= 50))  
finalDF.show()  
  
finalDF.format(**"orc"**).mode(**"overwrite"**).option(**"compression"**,**"snappy"**).save(**"hdfs\_path"**)

Output =>

C:\Users\zakeerm\IdeaProjects\Spark-Scala\venv\Scripts\python.exe C:/Users/zakeerm/Downloads/PysparkLearning/pysparkprograms/prescriber.py

23/11/16 13:14:28 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP\_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

23/11/16 13:14:28 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

root

|-- npi: integer (nullable = true)

|-- nppes\_provider\_last\_org\_name: string (nullable = true)

|-- nppes\_provider\_first\_name: string (nullable = true)

|-- nppes\_provider\_city: string (nullable = true)

|-- nppes\_provider\_state: string (nullable = true)

|-- specialty\_description: string (nullable = true)

|-- description\_flag: string (nullable = true)

|-- drug\_name: string (nullable = true)

|-- generic\_name: string (nullable = true)

|-- bene\_count: integer (nullable = true)

|-- total\_claim\_count: integer (nullable = true)

|-- total\_30\_day\_fill\_count: double (nullable = true)

|-- total\_day\_supply: integer (nullable = true)

|-- total\_drug\_cost: double (nullable = true)

|-- bene\_count\_ge65: integer (nullable = true)

|-- bene\_count\_ge65\_suppress\_flag: string (nullable = true)

|-- total\_claim\_count\_ge65: integer (nullable = true)

|-- ge65\_suppress\_flag: string (nullable = true)

|-- total\_30\_day\_fill\_count\_ge65: double (nullable = true)

|-- total\_day\_supply\_ge65: integer (nullable = true)

|-- total\_drug\_cost\_ge65: double (nullable = true)

|-- years\_of\_exp: string (nullable = true)

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

| npi|nppes\_provider\_last\_org\_name|nppes\_provider\_first\_name|nppes\_provider\_city|nppes\_provider\_state|specialty\_description|description\_flag| drug\_name| generic\_name|bene\_count|total\_claim\_count|total\_30\_day\_fill\_count|total\_day\_supply|total\_drug\_cost|bene\_count\_ge65|bene\_count\_ge65\_suppress\_flag|total\_claim\_count\_ge65|ge65\_suppress\_flag|total\_30\_day\_fill\_count\_ge65|total\_day\_supply\_ge65|total\_drug\_cost\_ge65|years\_of\_exp|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ATORVASTATIN CALCIUM|ATORVASTATIN CALCIUM| NULL| 13| 15.0| 450| 139.32| NULL| \*| 13| NULL| 15.0| 450| 139.32| = 45.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| CIPROFLOXACIN HCL| CIPROFLOXACIN HCL| NULL| 11| 11.0| 96| 80.99| NULL| \*| NULL| \*| NULL| NULL| NULL| = 43.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| DOXYCYCLINE HYCLATE| DOXYCYCLINE HYCLATE| 20| 20| 20.0| 199| 586.12| NULL| #| NULL| #| NULL| NULL| NULL| = 33.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| ELIQUIS| APIXABAN| NULL| 17| 17.0| 510| 6065.02| NULL| \*| 17| NULL| 17.0| 510| 6065.02| = 44.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| FUROSEMIDE| FUROSEMIDE| 12| 17| 17.0| 405| 45.76| NULL| #| NULL| #| NULL| NULL| NULL| = 55.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| HYDRALAZINE HCL| HYDRALAZINE HCL| NULL| 16| 16.0| 420| 169.48| NULL| #| NULL| #| NULL| NULL| NULL| = 36.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ISOSORBIDE MONONI...|ISOSORBIDE MONONI...| 16| 33| 35.0| 1005| 372.63| NULL| #| NULL| #| NULL| NULL| NULL| = 54.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LEVOFLOXACIN| LEVOFLOXACIN| 26| 26| 26.0| 159| 222.41| NULL| #| NULL| #| NULL| NULL| NULL| = 53.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LISINOPRIL| LISINOPRIL| 14| 31| 32.0| 960| 129.24| NULL| #| NULL| #| NULL| NULL| NULL| = 35.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METOPROLOL TARTRATE| METOPROLOL TARTRATE| 14| 33| 37.0| 1050| 183.29| 14| NULL| 33| NULL| 37.0| 1050| 183.29| = 51.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METRONIDAZOLE| METRONIDAZOLE| 11| 12| 12.0| 127| 152.66| 11| NULL| 12| NULL| 12.0| 127| 152.66| = 44.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PANTOPRAZOLE SODIUM| PANTOPRAZOLE SODIUM| NULL| 15| 15.0| 450| 140.83| NULL| #| NULL| #| NULL| NULL| NULL| = 52.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PREDNISONE| PREDNISONE| 17| 20| 20.0| 121| 59.96| NULL| #| NULL| #| NULL| NULL| NULL| = 52.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| WARFARIN SODIUM| WARFARIN SODIUM| NULL| 12| 12.0| 236| 197.69| NULL| \*| 12| NULL| 12.0| 236| 197.69| = 36.0|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| XARELTO| RIVAROXABAN| 12| 34| 34.0| 943| 12110.2| NULL| #| NULL| #| NULL| NULL| NULL| = 52.0|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S|ACETAMINOPHEN-COD...|ACETAMINOPHEN WIT...| 17| 51| 51.0| 1398| 577.96| NULL| #| 37| NULL| 37.0| 978| 437.25| = 42.0|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| AMITRIPTYLINE HCL| AMITRIPTYLINE HCL| NULL| 29| 29.0| 870| 254.2| 0| NULL| 0| NULL| 0.0| 0| 0.0| = 41.0|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| BACLOFEN| BACLOFEN| 43| 104| 112.0| 3194| 2106.97| 17| NULL| 31| NULL| 31.0| 890| 484.56| = 40.0|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| BUTRANS| BUPRENORPHINE| 19| 63| 63.0| 1764| 24514.23| NULL| #| 41| NULL| 41.0| 1150| 16188.72| = 33.0|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| CYCLOBENZAPRINE HCL| CYCLOBENZAPRINE HCL| NULL| 11| 11.0| 300| 31.88| NULL| \*| NULL| \*| NULL| NULL| NULL| = 31.0|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

only showing top 20 rows

Clean Prescriber Data

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+--------------------+-----------------+----------------+---------------+

| npi|nppes\_provider\_last\_org\_name|nppes\_provider\_first\_name|nppes\_provider\_city|nppes\_provider\_state|specialty\_description| drug\_name|total\_claim\_count|total\_day\_supply|total\_drug\_cost|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+--------------------+-----------------+----------------+---------------+

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine|ATORVASTATIN CALCIUM| 13| 450| 139.32|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| CIPROFLOXACIN HCL| 11| 96| 80.99|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| DOXYCYCLINE HYCLATE| 20| 199| 586.12|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| ELIQUIS| 17| 510| 6065.02|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| FUROSEMIDE| 17| 405| 45.76|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| HYDRALAZINE HCL| 16| 420| 169.48|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine|ISOSORBIDE MONONI...| 33| 1005| 372.63|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| LEVOFLOXACIN| 26| 159| 222.41|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| LISINOPRIL| 31| 960| 129.24|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| METOPROLOL TARTRATE| 33| 1050| 183.29|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| METRONIDAZOLE| 12| 127| 152.66|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| PANTOPRAZOLE SODIUM| 15| 450| 140.83|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| PREDNISONE| 20| 121| 59.96|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| WARFARIN SODIUM| 12| 236| 197.69|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| XARELTO| 34| 943| 12110.2|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology|ACETAMINOPHEN-COD...| 51| 1398| 577.96|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| AMITRIPTYLINE HCL| 29| 870| 254.2|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| BACLOFEN| 104| 3194| 2106.97|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| BUTRANS| 63| 1764| 24514.23|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| CYCLOBENZAPRINE HCL| 11| 300| 31.88|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+--------------------+-----------------+----------------+---------------+

only showing top 20 rows

Rename the above fields to shorter names

Add a Country Field 'USA' to the above data

+----------+---------+----------+----------+-----+-----------------+--------------------+-----------+----------+---------+-------+

| npi|last\_name|first\_name| city|state| specialty| drug|claim\_count|day\_supply|drug\_cost|country|

+----------+---------+----------+----------+-----+-----------------+--------------------+-----------+----------+---------+-------+

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine|ATORVASTATIN CALCIUM| 13| 450| 139.32| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| CIPROFLOXACIN HCL| 11| 96| 80.99| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| DOXYCYCLINE HYCLATE| 20| 199| 586.12| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| ELIQUIS| 17| 510| 6065.02| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| FUROSEMIDE| 17| 405| 45.76| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| HYDRALAZINE HCL| 16| 420| 169.48| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine|ISOSORBIDE MONONI...| 33| 1005| 372.63| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| LEVOFLOXACIN| 26| 159| 222.41| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| LISINOPRIL| 31| 960| 129.24| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| METOPROLOL TARTRATE| 33| 1050| 183.29| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| METRONIDAZOLE| 12| 127| 152.66| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| PANTOPRAZOLE SODIUM| 15| 450| 140.83| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| PREDNISONE| 20| 121| 59.96| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| WARFARIN SODIUM| 12| 236| 197.69| USA|

|2006000252|ENKESHAFI| ARDALAN|CUMBERLAND| MD|Internal Medicine| XARELTO| 34| 943| 12110.2| USA|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology|ACETAMINOPHEN-COD...| 51| 1398| 577.96| USA|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| AMITRIPTYLINE HCL| 29| 870| 254.2| USA|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| BACLOFEN| 104| 3194| 2106.97| USA|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| BUTRANS| 63| 1764| 24514.23| USA|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| CYCLOBENZAPRINE HCL| 11| 300| 31.88| USA|

+----------+---------+----------+----------+-----+-----------------+--------------------+-----------+----------+---------+-------+

only showing top 20 rows

Clean the “years\_of\_exp” to extract only the numbers.

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

| npi|nppes\_provider\_last\_org\_name|nppes\_provider\_first\_name|nppes\_provider\_city|nppes\_provider\_state|specialty\_description|description\_flag| drug\_name| generic\_name|bene\_count|total\_claim\_count|total\_30\_day\_fill\_count|total\_day\_supply|total\_drug\_cost|bene\_count\_ge65|bene\_count\_ge65\_suppress\_flag|total\_claim\_count\_ge65|ge65\_suppress\_flag|total\_30\_day\_fill\_count\_ge65|total\_day\_supply\_ge65|total\_drug\_cost\_ge65|years\_of\_exp|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ATORVASTATIN CALCIUM|ATORVASTATIN CALCIUM| NULL| 13| 15.0| 450| 139.32| NULL| \*| 13| NULL| 15.0| 450| 139.32| 45|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| CIPROFLOXACIN HCL| CIPROFLOXACIN HCL| NULL| 11| 11.0| 96| 80.99| NULL| \*| NULL| \*| NULL| NULL| NULL| 43|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| DOXYCYCLINE HYCLATE| DOXYCYCLINE HYCLATE| 20| 20| 20.0| 199| 586.12| NULL| #| NULL| #| NULL| NULL| NULL| 33|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| ELIQUIS| APIXABAN| NULL| 17| 17.0| 510| 6065.02| NULL| \*| 17| NULL| 17.0| 510| 6065.02| 44|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| FUROSEMIDE| FUROSEMIDE| 12| 17| 17.0| 405| 45.76| NULL| #| NULL| #| NULL| NULL| NULL| 55|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| HYDRALAZINE HCL| HYDRALAZINE HCL| NULL| 16| 16.0| 420| 169.48| NULL| #| NULL| #| NULL| NULL| NULL| 36|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ISOSORBIDE MONONI...|ISOSORBIDE MONONI...| 16| 33| 35.0| 1005| 372.63| NULL| #| NULL| #| NULL| NULL| NULL| 54|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LEVOFLOXACIN| LEVOFLOXACIN| 26| 26| 26.0| 159| 222.41| NULL| #| NULL| #| NULL| NULL| NULL| 53|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LISINOPRIL| LISINOPRIL| 14| 31| 32.0| 960| 129.24| NULL| #| NULL| #| NULL| NULL| NULL| 35|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METOPROLOL TARTRATE| METOPROLOL TARTRATE| 14| 33| 37.0| 1050| 183.29| 14| NULL| 33| NULL| 37.0| 1050| 183.29| 51|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METRONIDAZOLE| METRONIDAZOLE| 11| 12| 12.0| 127| 152.66| 11| NULL| 12| NULL| 12.0| 127| 152.66| 44|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PANTOPRAZOLE SODIUM| PANTOPRAZOLE SODIUM| NULL| 15| 15.0| 450| 140.83| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PREDNISONE| PREDNISONE| 17| 20| 20.0| 121| 59.96| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| WARFARIN SODIUM| WARFARIN SODIUM| NULL| 12| 12.0| 236| 197.69| NULL| \*| 12| NULL| 12.0| 236| 197.69| 36|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| XARELTO| RIVAROXABAN| 12| 34| 34.0| 943| 12110.2| NULL| #| NULL| #| NULL| NULL| NULL| 52|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S|ACETAMINOPHEN-COD...|ACETAMINOPHEN WIT...| 17| 51| 51.0| 1398| 577.96| NULL| #| 37| NULL| 37.0| 978| 437.25| 42|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| AMITRIPTYLINE HCL| AMITRIPTYLINE HCL| NULL| 29| 29.0| 870| 254.2| 0| NULL| 0| NULL| 0.0| 0| 0.0| 41|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| BACLOFEN| BACLOFEN| 43| 104| 112.0| 3194| 2106.97| 17| NULL| 31| NULL| 31.0| 890| 484.56| 40|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| BUTRANS| BUPRENORPHINE| 19| 63| 63.0| 1764| 24514.23| NULL| #| 41| NULL| 41.0| 1150| 16188.72| 33|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| CYCLOBENZAPRINE HCL| CYCLOBENZAPRINE HCL| NULL| 11| 11.0| 300| 31.88| NULL| \*| NULL| \*| NULL| NULL| NULL| 31|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

only showing top 20 rows

root

|-- npi: integer (nullable = true)

|-- nppes\_provider\_last\_org\_name: string (nullable = true)

|-- nppes\_provider\_first\_name: string (nullable = true)

|-- nppes\_provider\_city: string (nullable = true)

|-- nppes\_provider\_state: string (nullable = true)

|-- specialty\_description: string (nullable = true)

|-- description\_flag: string (nullable = true)

|-- drug\_name: string (nullable = true)

|-- generic\_name: string (nullable = true)

|-- bene\_count: integer (nullable = true)

|-- total\_claim\_count: integer (nullable = true)

|-- total\_30\_day\_fill\_count: double (nullable = true)

|-- total\_day\_supply: integer (nullable = true)

|-- total\_drug\_cost: double (nullable = true)

|-- bene\_count\_ge65: integer (nullable = true)

|-- bene\_count\_ge65\_suppress\_flag: string (nullable = true)

|-- total\_claim\_count\_ge65: integer (nullable = true)

|-- ge65\_suppress\_flag: string (nullable = true)

|-- total\_30\_day\_fill\_count\_ge65: double (nullable = true)

|-- total\_day\_supply\_ge65: integer (nullable = true)

|-- total\_drug\_cost\_ge65: double (nullable = true)

|-- years\_of\_exp: string (nullable = true)

Convert the years\_of\_exp field to integer

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

| npi|nppes\_provider\_last\_org\_name|nppes\_provider\_first\_name|nppes\_provider\_city|nppes\_provider\_state|specialty\_description|description\_flag| drug\_name| generic\_name|bene\_count|total\_claim\_count|total\_30\_day\_fill\_count|total\_day\_supply|total\_drug\_cost|bene\_count\_ge65|bene\_count\_ge65\_suppress\_flag|total\_claim\_count\_ge65|ge65\_suppress\_flag|total\_30\_day\_fill\_count\_ge65|total\_day\_supply\_ge65|total\_drug\_cost\_ge65|years\_of\_exp|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ATORVASTATIN CALCIUM|ATORVASTATIN CALCIUM| NULL| 13| 15.0| 450| 139.32| NULL| \*| 13| NULL| 15.0| 450| 139.32| 45|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| CIPROFLOXACIN HCL| CIPROFLOXACIN HCL| NULL| 11| 11.0| 96| 80.99| NULL| \*| NULL| \*| NULL| NULL| NULL| 43|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| DOXYCYCLINE HYCLATE| DOXYCYCLINE HYCLATE| 20| 20| 20.0| 199| 586.12| NULL| #| NULL| #| NULL| NULL| NULL| 33|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| ELIQUIS| APIXABAN| NULL| 17| 17.0| 510| 6065.02| NULL| \*| 17| NULL| 17.0| 510| 6065.02| 44|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| FUROSEMIDE| FUROSEMIDE| 12| 17| 17.0| 405| 45.76| NULL| #| NULL| #| NULL| NULL| NULL| 55|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| HYDRALAZINE HCL| HYDRALAZINE HCL| NULL| 16| 16.0| 420| 169.48| NULL| #| NULL| #| NULL| NULL| NULL| 36|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ISOSORBIDE MONONI...|ISOSORBIDE MONONI...| 16| 33| 35.0| 1005| 372.63| NULL| #| NULL| #| NULL| NULL| NULL| 54|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LEVOFLOXACIN| LEVOFLOXACIN| 26| 26| 26.0| 159| 222.41| NULL| #| NULL| #| NULL| NULL| NULL| 53|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LISINOPRIL| LISINOPRIL| 14| 31| 32.0| 960| 129.24| NULL| #| NULL| #| NULL| NULL| NULL| 35|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METOPROLOL TARTRATE| METOPROLOL TARTRATE| 14| 33| 37.0| 1050| 183.29| 14| NULL| 33| NULL| 37.0| 1050| 183.29| 51|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METRONIDAZOLE| METRONIDAZOLE| 11| 12| 12.0| 127| 152.66| 11| NULL| 12| NULL| 12.0| 127| 152.66| 44|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PANTOPRAZOLE SODIUM| PANTOPRAZOLE SODIUM| NULL| 15| 15.0| 450| 140.83| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PREDNISONE| PREDNISONE| 17| 20| 20.0| 121| 59.96| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| WARFARIN SODIUM| WARFARIN SODIUM| NULL| 12| 12.0| 236| 197.69| NULL| \*| 12| NULL| 12.0| 236| 197.69| 36|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| XARELTO| RIVAROXABAN| 12| 34| 34.0| 943| 12110.2| NULL| #| NULL| #| NULL| NULL| NULL| 52|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S|ACETAMINOPHEN-COD...|ACETAMINOPHEN WIT...| 17| 51| 51.0| 1398| 577.96| NULL| #| 37| NULL| 37.0| 978| 437.25| 42|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| AMITRIPTYLINE HCL| AMITRIPTYLINE HCL| NULL| 29| 29.0| 870| 254.2| 0| NULL| 0| NULL| 0.0| 0| 0.0| 41|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| BACLOFEN| BACLOFEN| 43| 104| 112.0| 3194| 2106.97| 17| NULL| 31| NULL| 31.0| 890| 484.56| 40|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| BUTRANS| BUPRENORPHINE| 19| 63| 63.0| 1764| 24514.23| NULL| #| 41| NULL| 41.0| 1150| 16188.72| 33|

| NULL| KHALIL| RASHID| TOLEDO| OH| Anesthesiology| S| CYCLOBENZAPRINE HCL| CYCLOBENZAPRINE HCL| NULL| 11| 11.0| 300| 31.88| NULL| \*| NULL| \*| NULL| NULL| NULL| 31|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

only showing top 20 rows

root

|-- npi: integer (nullable = true)

|-- nppes\_provider\_last\_org\_name: string (nullable = true)

|-- nppes\_provider\_first\_name: string (nullable = true)

|-- nppes\_provider\_city: string (nullable = true)

|-- nppes\_provider\_state: string (nullable = true)

|-- specialty\_description: string (nullable = true)

|-- description\_flag: string (nullable = true)

|-- drug\_name: string (nullable = true)

|-- generic\_name: string (nullable = true)

|-- bene\_count: integer (nullable = true)

|-- total\_claim\_count: integer (nullable = true)

|-- total\_30\_day\_fill\_count: double (nullable = true)

|-- total\_day\_supply: integer (nullable = true)

|-- total\_drug\_cost: double (nullable = true)

|-- bene\_count\_ge65: integer (nullable = true)

|-- bene\_count\_ge65\_suppress\_flag: string (nullable = true)

|-- total\_claim\_count\_ge65: integer (nullable = true)

|-- ge65\_suppress\_flag: string (nullable = true)

|-- total\_30\_day\_fill\_count\_ge65: double (nullable = true)

|-- total\_day\_supply\_ge65: integer (nullable = true)

|-- total\_drug\_cost\_ge65: double (nullable = true)

|-- years\_of\_exp: integer (nullable = true)

Combine First Name and Last Name into a single field and remove the individual columns

+------------+-----------------+

|years\_of\_exp| full\_name|

+------------+-----------------+

| 45|ARDALAN ENKESHAFI|

| 43|ARDALAN ENKESHAFI|

| 33|ARDALAN ENKESHAFI|

| 44|ARDALAN ENKESHAFI|

| 55|ARDALAN ENKESHAFI|

| 36|ARDALAN ENKESHAFI|

| 54|ARDALAN ENKESHAFI|

| 53|ARDALAN ENKESHAFI|

| 35|ARDALAN ENKESHAFI|

| 51|ARDALAN ENKESHAFI|

| 44|ARDALAN ENKESHAFI|

| 52|ARDALAN ENKESHAFI|

| 52|ARDALAN ENKESHAFI|

| 36|ARDALAN ENKESHAFI|

| 52|ARDALAN ENKESHAFI|

| 42| RASHID KHALIL|

| 41| RASHID KHALIL|

| 40| RASHID KHALIL|

| 33| RASHID KHALIL|

| 31| RASHID KHALIL|

+------------+-----------------+

only showing top 20 rows

Count the number of null values for each column

+------------+---------+

|years\_of\_exp|full\_name|

+------------+---------+

| 1| 0|

+------------+---------+

clean all the Null/Nan Values

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

| npi|nppes\_provider\_last\_org\_name|nppes\_provider\_first\_name|nppes\_provider\_city|nppes\_provider\_state|specialty\_description|description\_flag| drug\_name| generic\_name|bene\_count|total\_claim\_count|total\_30\_day\_fill\_count|total\_day\_supply|total\_drug\_cost|bene\_count\_ge65|bene\_count\_ge65\_suppress\_flag|total\_claim\_count\_ge65|ge65\_suppress\_flag|total\_30\_day\_fill\_count\_ge65|total\_day\_supply\_ge65|total\_drug\_cost\_ge65|years\_of\_exp|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ATORVASTATIN CALCIUM|ATORVASTATIN CALCIUM| NULL| 13| 15.0| 450| 139.32| NULL| \*| 13| NULL| 15.0| 450| 139.32| 45|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| CIPROFLOXACIN HCL| CIPROFLOXACIN HCL| NULL| 11| 11.0| 96| 80.99| NULL| \*| NULL| \*| NULL| NULL| NULL| 43|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| DOXYCYCLINE HYCLATE| DOXYCYCLINE HYCLATE| 20| 20| 20.0| 199| 586.12| NULL| #| NULL| #| NULL| NULL| NULL| 33|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| ELIQUIS| APIXABAN| NULL| 17| 17.0| 510| 6065.02| NULL| \*| 17| NULL| 17.0| 510| 6065.02| 44|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| FUROSEMIDE| FUROSEMIDE| 12| 17| 17.0| 405| 45.76| NULL| #| NULL| #| NULL| NULL| NULL| 55|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| HYDRALAZINE HCL| HYDRALAZINE HCL| NULL| 16| 16.0| 420| 169.48| NULL| #| NULL| #| NULL| NULL| NULL| 36|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S|ISOSORBIDE MONONI...|ISOSORBIDE MONONI...| 16| 33| 35.0| 1005| 372.63| NULL| #| NULL| #| NULL| NULL| NULL| 54|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LEVOFLOXACIN| LEVOFLOXACIN| 26| 26| 26.0| 159| 222.41| NULL| #| NULL| #| NULL| NULL| NULL| 53|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| LISINOPRIL| LISINOPRIL| 14| 31| 32.0| 960| 129.24| NULL| #| NULL| #| NULL| NULL| NULL| 35|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METOPROLOL TARTRATE| METOPROLOL TARTRATE| 14| 33| 37.0| 1050| 183.29| 14| NULL| 33| NULL| 37.0| 1050| 183.29| 51|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| METRONIDAZOLE| METRONIDAZOLE| 11| 12| 12.0| 127| 152.66| 11| NULL| 12| NULL| 12.0| 127| 152.66| 44|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PANTOPRAZOLE SODIUM| PANTOPRAZOLE SODIUM| NULL| 15| 15.0| 450| 140.83| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| PREDNISONE| PREDNISONE| 17| 20| 20.0| 121| 59.96| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| WARFARIN SODIUM| WARFARIN SODIUM| NULL| 12| 12.0| 236| 197.69| NULL| \*| 12| NULL| 12.0| 236| 197.69| 36|

|2006000252| ENKESHAFI| ARDALAN| CUMBERLAND| MD| Internal Medicine| S| XARELTO| RIVAROXABAN| 12| 34| 34.0| 943| 12110.2| NULL| #| NULL| #| NULL| NULL| NULL| 52|

|2006000293| ESCOBAR| JULIO| DAYTON| NV| Dentist| T| AMOXICILLIN| AMOXICILLIN| 19| NULL| 21.0| 121| 71.33| NULL| #| NULL| #| NULL| NULL| NULL| 43|

|2006000293| ESCOBAR| JULIO| DAYTON| NV| Dentist| T|HYDROCODONE-ACETA...|HYDROCODONE/ACETA...| NULL| 11| 11.0| 27| 63.43| NULL| \*| NULL| \*| NULL| NULL| NULL| 40|

|2006000408| BLAKEMORE| ROSIE| NASHVILLE| TN| Nurse Practitioner| S| CALCITRIOL| CALCITRIOL| NULL| 12| 12.0| 217| 129.17| NULL| \*| 12| NULL| 12.0| 217| 129.17| 32|

|2006000408| BLAKEMORE| ROSIE| NASHVILLE| TN| Nurse Practitioner| S|LEVOTHYROXINE SODIUM|LEVOTHYROXINE SODIUM| NULL| 14| 18.0| 399| 198.87| NULL| #| NULL| #| NULL| NULL| NULL| 43|

|2006000533| GIRARDI| DAVID| BROOKVILLE| PA| Family Practice| S| ADVAIR DISKUS|FLUTICASONE/SALME...| NULL| 14| 14.0| 420| 5415.66| NULL| \*| 14| NULL| 14.0| 420| 5415.66| 42|

+----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+

only showing top 20 rows

root

|-- npi: integer (nullable = true)

|-- nppes\_provider\_last\_org\_name: string (nullable = true)

|-- nppes\_provider\_first\_name: string (nullable = true)

|-- nppes\_provider\_city: string (nullable = true)

|-- nppes\_provider\_state: string (nullable = true)

|-- specialty\_description: string (nullable = true)

|-- description\_flag: string (nullable = true)

|-- drug\_name: string (nullable = true)

|-- generic\_name: string (nullable = true)

|-- bene\_count: integer (nullable = true)

|-- total\_claim\_count: integer (nullable = true)

|-- total\_30\_day\_fill\_count: double (nullable = true)

|-- total\_day\_supply: integer (nullable = true)

|-- total\_drug\_cost: double (nullable = true)

|-- bene\_count\_ge65: integer (nullable = true)

|-- bene\_count\_ge65\_suppress\_flag: string (nullable = true)

|-- total\_claim\_count\_ge65: integer (nullable = true)

|-- ge65\_suppress\_flag: string (nullable = true)

|-- total\_30\_day\_fill\_count\_ge65: double (nullable = true)

|-- total\_day\_supply\_ge65: integer (nullable = true)

|-- total\_drug\_cost\_ge65: double (nullable = true)

|-- years\_of\_exp: string (nullable = true)

Calculate the number of distinct Prescribers assigned for each City.

+--------------------+-------------------+--------------------+-------------+

|nppes\_provider\_state|nppes\_provider\_city|distinct\_prescribers|total\_trx\_cnt|

+--------------------+-------------------+--------------------+-------------+

| CA| SALINAS| 194| 26601|

| MT| HELENA| 73| 10565|

| TN| BRENTWOOD| 64| 8366|

| CA| ANAHEIM| 498| 82010|

| MO| LAMAR| 11| 4254|

| WV| GRANTSVILLE| 10| 2324|

| VA| BEDFORD| 18| 2867|

| NJ| PATERSON| 87| 19438|

| MN| SHAKOPEE| 34| 3827|

| FL| OVIEDO| 49| 10904|

| MI| TRAVERSE CITY| 256| 32627|

| FL| MIRAMAR| 96| 10627|

| SD| CUSTER| 6| 1798|

| GA| SNELLVILEL| 1| 44|

| DE| REHOBOTH BEACH| 19| 2919|

| MD| SNOW HILL| 5| 352|

| TX| PEARLAND| 82| 11527|

| IL| COULTERVILLE| 3| 816|

| IL| PRINCETON| 14| 3836|

| PA| NARBERTH| 9| 7221|

+--------------------+-------------------+--------------------+-------------+

only showing top 20 rows

Calculate total total\_claim\_count prescribed for each city.

+-------------------+-----------------+

|nppes\_provider\_city|total\_claim\_count|

+-------------------+-----------------+

| CUMBERLAND| 36549|

| MCMINNVILLE| 15746|

| ANNA| 5785|

| MARLTON| 13640|

| PRINCETON| 46943|

| NESCONSET| 3632|

| NOVATO| 8265|

| EDMONDS| 13496|

| SOUTHLAKE| 4720|

| HANCOCK| 4188|

| WATERTOWN| 34119|

| MINFORD| 206|

| OSAGE| 983|

| RARITAN| 1047|

| WALLED LAKE| 1325|

| OLD TOWN| 2383|

| TOWANDA| 4379|

| OAKBROOK TERRACE| 3203|

| EMINENCE| 3575|

| MAYVILLE| 3631|

+-------------------+-----------------+

only showing top 20 rows

Do not report a city in the final report if no prescriber is assigned to it.

+--------------------+-------------------+--------------------+-------------+

|nppes\_provider\_state|nppes\_provider\_city|distinct\_prescribers|total\_trx\_cnt|

+--------------------+-------------------+--------------------+-------------+

| CA| SALINAS| 194| 26601|

| MT| HELENA| 73| 10565|

| TN| BRENTWOOD| 64| 8366|

| CA| ANAHEIM| 498| 82010|

| MO| LAMAR| 11| 4254|

| WV| GRANTSVILLE| 10| 2324|

| VA| BEDFORD| 18| 2867|

| NJ| PATERSON| 87| 19438|

| MN| SHAKOPEE| 34| 3827|

| FL| OVIEDO| 49| 10904|

| MI| TRAVERSE CITY| 256| 32627|

| FL| MIRAMAR| 96| 10627|

| SD| CUSTER| 6| 1798|

| GA| SNELLVILEL| 1| 44|

| DE| REHOBOTH BEACH| 19| 2919|

| MD| SNOW HILL| 5| 352|

| TX| PEARLAND| 82| 11527|

| IL| COULTERVILLE| 3| 816|

| IL| PRINCETON| 14| 3836|

| PA| NARBERTH| 9| 7221|

+--------------------+-------------------+--------------------+-------------+

only showing top 20 rows

+-------------------+--------------------+

|nppes\_provider\_city|distinct\_prescribers|

+-------------------+--------------------+

| SALINAS| 194|

| HELENA| 73|

| BRENTWOOD| 64|

| ANAHEIM| 498|

| LAMAR| 11|

| GRANTSVILLE| 10|

| BEDFORD| 18|

| PATERSON| 87|

| SHAKOPEE| 34|

| OVIEDO| 49|

| TRAVERSE CITY| 256|

| MIRAMAR| 96|

| CUSTER| 6|

| SNELLVILEL| 1|

| REHOBOTH BEACH| 19|

| SNOW HILL| 5|

| PEARLAND| 82|

| COULTERVILLE| 3|

| PRINCETON| 14|

| NARBERTH| 9|

+-------------------+--------------------+

only showing top 20 rows

Top 5 Prescribers with highest total\_claim\_count per each state.

+-----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+----------+

| npi|nppes\_provider\_last\_org\_name|nppes\_provider\_first\_name|nppes\_provider\_city|nppes\_provider\_state|specialty\_description|description\_flag| drug\_name| generic\_name|bene\_count|total\_claim\_count|total\_30\_day\_fill\_count|total\_day\_supply|total\_drug\_cost|bene\_count\_ge65|bene\_count\_ge65\_suppress\_flag|total\_claim\_count\_ge65|ge65\_suppress\_flag|total\_30\_day\_fill\_count\_ge65|total\_day\_supply\_ge65|total\_drug\_cost\_ge65|years\_of\_exp|dense\_rank|

+-----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+----------+

|-1662759025| COOMES| MARK| APO| AA| General Surgery| S| AZITHROMYCIN| AZITHROMYCIN| 28| 30| 30.0| 150| 222.68| NULL| #| NULL| #| NULL| NULL| NULL| 42| 1|

|-2046463267| CASSIDY| GEORGE| FPO| AA| Physician Assistant| S|LIDOCAINE HCL VIS...| LIDOCAINE HCL| 13| 13| 13.0| 62| 148.02| NULL| \*| NULL| \*| NULL| NULL| NULL| 52| 2|

|-2046463267| CASSIDY| GEORGE| FPO| AA| Physician Assistant| S| PREDNISONE| PREDNISONE| 12| 12| 12.0| 62| 36.5| NULL| \*| NULL| \*| NULL| NULL| NULL| 35| 3|

| 2086924743| VOLNEK| JILLIAN| APO| AE| Physician Assistant| S|LEVOTHYROXINE SODIUM|LEVOTHYROXINE SODIUM| 32| 246| 307.1| 9031| 3412.29| 18| NULL| 124| NULL| 167.0| 4989| 1669.89| 36| 1|

|-1996260772| VARONE| RICKY| FPO| AE| Physician Assistant| S|HYDROCODONE-ACETA...|HYDROCODONE/ACETA...| 59| 125| 125.0| 2430| 2818.97| 38| NULL| 79| NULL| 79.0| 1495| 1832.53| 30| 2|

| 2086924743| VOLNEK| JILLIAN| APO| AE| Physician Assistant| S| LISINOPRIL| LISINOPRIL| 11| 61| 77.0| 2310| 358.8| NULL| \*| 29| NULL| 33.0| 990| 118.92| 35| 3|

| 2086924743| VOLNEK| JILLIAN| APO| AE| Physician Assistant| S| HUMALOG| INSULIN LISPRO| 14| 58| 75.0| 2159| 46544.78| NULL| \*| 27| NULL| 41.9| 1183| 17831.35| 41| 4|

| 2086924743| VOLNEK| JILLIAN| APO| AE| Physician Assistant| S|HUMALOG KWIKPEN U...| INSULIN LISPRO| 14| 58| 68.0| 1997| 39025.32| NULL| \*| 37| NULL| 42.9| 1262| 24055.81| 51| 4|

| 2096867263| SAUNDERS| SHEREE| FPO| AE| Internal Medicine| S| GABAPENTIN| GABAPENTIN| 14| 48| 63.8| 1800| 607.79| NULL| \*| 21| NULL| 34.5| 920| 264.96| 41| 5|

|-1440346106| KIM| KWANGHEE| ANCHORAGE| AK| Internal Medicine| S| SIMVASTATIN| SIMVASTATIN| 109| 482| 1045.1| 29009| 3830.52| NULL| #| NULL| #| NULL| NULL| NULL| 53| 1|

|-2046512683| MUNGARA| MAHIJA| KETCHIKAN| AK| Psychiatry| S| QUETIAPINE FUMARATE| QUETIAPINE FUMARATE| 58| 412| 442.8| 13181| 21544.69| NULL| \*| 77| NULL| 92.0| 2760| 7890.78| 50| 2|

|-1379786104| CARRICABURU| RACHAEL| ANCHORAGE| AK| Physician Assistant| S| METFORMIN HCL| METFORMIN HCL| 55| 394| 559.3| 13041| 2261.29| 37| NULL| 205| NULL| 318.3| 7899| 1115.67| 34| 3|

|-2046512683| MUNGARA| MAHIJA| KETCHIKAN| AK| Psychiatry| S| INVEGA SUSTENNA|PALIPERIDONE PALM...| 55| 378| 387.2| 10900| 666869.03| NULL| \*| 42| NULL| 42.0| 1212| 63517.85| 46| 4|

|-1945712863| MAILER| KAREN| ANCHORAGE| AK| Family Practice| S| OMEPRAZOLE| OMEPRAZOLE| 32| 329| 329.1| 6732| 3008.15| NULL| #| 201| NULL| 201.0| 4127| 1632.5| 47| 5|

|-1571906923| LACKEY| THOMAS| GADSDEN| AL| Anesthesiology| S|HYDROCODONE-ACETA...|HYDROCODONE/ACETA...| 579| 4708| 4716.4| 137802| 126481.51| 234| NULL| 1860| NULL| 1864.7| 54544| 49766.26| 40| 1|

|-1773030142| PARKER| CECIL| MOBILE| AL| Internal Medicine| S|HYDROCODONE-ACETA...|HYDROCODONE/ACETA...| 359| 3157| 3168.0| 93581| 53403.62| 211| NULL| 1696| NULL| 1706.0| 50104| 29864.04| 37| 2|

| 2127074787| MANCHIKES| LLOYD| MOBILE| AL| General Practice| S|HYDROCODONE-ACETA...|HYDROCODONE/ACETA...| 284| 2187| 2193.0| 64899| 57628.95| 110| NULL| 909| NULL| 913.5| 26934| 25414.89| 34| 3|

|-1763926571| LACHMAN| MARILYN| ALBERTVILLE| AL| Psychiatry| S| CLONAZEPAM| CLONAZEPAM| 259| 1888| 2011.0| 59226| 13799.27| 37| NULL| 213| NULL| 273.2| 8090| 1970.99| 46| 4|

|-1844686870| RUSSELL| BRUCE| SUMITON| AL| General Practice| S|HYDROCODONE-ACETA...|HYDROCODONE/ACETA...| 197| 1776| 1778.4| 46121| 63588.39| 147| NULL| 1298| NULL| 1300.0| 34004| 45639.0| 34| 5|

|-1834465056| SANTIAGO| ERIC| FPO| AP| Internal Medicine| S|LISINOPRIL-HYDROC...|LISINOPRIL/HYDROC...| 20| 58| 134.0| 4020| 390.08| NULL| #| NULL| #| NULL| NULL| NULL| 40| 1|

+-----------+----------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+-----------------------+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+----------+

only showing top 20 rows

root

|-- npi: integer (nullable = true)

|-- nppes\_provider\_last\_org\_name: string (nullable = true)

|-- nppes\_provider\_first\_name: string (nullable = true)

|-- nppes\_provider\_city: string (nullable = true)

|-- nppes\_provider\_state: string (nullable = true)

|-- specialty\_description: string (nullable = true)

|-- description\_flag: string (nullable = true)

|-- drug\_name: string (nullable = true)

|-- generic\_name: string (nullable = true)

|-- bene\_count: integer (nullable = true)

|-- total\_claim\_count: integer (nullable = true)

|-- total\_30\_day\_fill\_count: double (nullable = true)

|-- total\_day\_supply: integer (nullable = true)

|-- total\_drug\_cost: double (nullable = true)

|-- bene\_count\_ge65: integer (nullable = true)

|-- bene\_count\_ge65\_suppress\_flag: string (nullable = true)

|-- total\_claim\_count\_ge65: integer (nullable = true)

|-- ge65\_suppress\_flag: string (nullable = true)

|-- total\_30\_day\_fill\_count\_ge65: double (nullable = true)

|-- total\_day\_supply\_ge65: integer (nullable = true)

|-- total\_drug\_cost\_ge65: double (nullable = true)

|-- years\_of\_exp: string (nullable = true)

|-- dense\_rank: integer (nullable = false)

Consider the prescribers only from 20 to 50 years of experience

+-----------+--------------------+------------+-----------------+----------------+---------------+

| npi|nppes\_provider\_state|years\_of\_exp|total\_claim\_count|total\_day\_supply|total\_drug\_cost|

+-----------+--------------------+------------+-----------------+----------------+---------------+

|-1662759025| AA| 42| 30| 150| 222.68|

|-2046463267| AA| 35| 12| 62| 36.5|

| 2086924743| AE| 36| 246| 9031| 3412.29|

|-1996260772| AE| 30| 125| 2430| 2818.97|

| 2086924743| AE| 35| 61| 2310| 358.8|

| 2086924743| AE| 41| 58| 2159| 46544.78|

| 2096867263| AE| 41| 48| 1800| 607.79|

|-2046512683| AK| 50| 412| 13181| 21544.69|

|-1379786104| AK| 34| 394| 13041| 2261.29|

|-2046512683| AK| 46| 378| 10900| 666869.03|

|-1945712863| AK| 47| 329| 6732| 3008.15|

|-1571906923| AL| 40| 4708| 137802| 126481.51|

|-1773030142| AL| 37| 3157| 93581| 53403.62|

| 2127074787| AL| 34| 2187| 64899| 57628.95|

|-1763926571| AL| 46| 1888| 59226| 13799.27|

|-1844686870| AL| 34| 1776| 46121| 63588.39|

|-1834465056| AP| 40| 58| 4020| 390.08|

|-1834465056| AP| 32| 52| 1737| 16760.74|

| 2086605152| AP| 50| 50| 258| 176.15|

|-1834465056| AP| 50| 35| 327| 270.4|

+-----------+--------------------+------------+-----------------+----------------+---------------+

only showing top 20 rows

Process finished with exit code 0