010-Combine_data

June 22, 2020

1 Final Project: Admission Prediction from NHAMCS

1.1 Data exploration notebook

1.1.1 DS5559: Big Data Analysis

1.1.2 Thomas Hartka, Alicia Doan, Michael Langmayr

Created: 6/21/20

In this notebook we read in files from NHAMCS into a pyspark DataFrame for the years 2007-2017, then concatenate these DataFrames. We then determine which years contain data for certain variables. With this information, we select the variables to investigate in our prediction models. Finally, we visualize our data, specifically focusing on the relationship between the predictors and response variables.

1.2 Configuration

[8]:
```
# set data directory
data_dir = "../raw_data"
```

[2]:
```
# import python libraries
import os
import pandas as pd
import numpy as np
from functools import reduce
```

[5]:
```
import os
os.getcwd()

# import sys
# sys.path.append('lib/')
from lib import combineDf
```

[6]:
```
# set up pyspark
from pyspark.sql import *
from pyspark.sql import SparkSession
```

```
from pyspark.sql.functions import *

spark = SparkSession.builder.getOrCreate()
```

1.3 Read in data files and combine

[9]:
```
# create array for DataFrames
df = []

# loop through all files in the directory
for i,filename in enumerate(os.listdir(data_dir)):
    print(i,":", filename)

    df.append(spark.read.csv(data_dir+"//"+filename, inferSchema=True, header =␣
 ↪True))

    # extract year from file name
    year = filename.split(".")[0][-4:]

    # add id
    df[i] = df[i].withColumn("ID", monotonically_increasing_id())
    df[i] = df[i].withColumn("YEAR", lit(year))
```

```
0 : NHAMCS2007.csv
1 : NHAMCS2012.csv
2 : NHAMCS2008.csv
3 : NHAMCS2010.csv
4 : NHAMCS2017.csv
5 : NHAMCS2015.csv
6 : NHAMCS2014.csv
7 : NHAMCS2016.csv
8 : NHAMCS2009.csv
9 : NHAMCS2013.csv
10 : NHAMCS2011.csv
```

[26]:
```
df[0].select(df[0].columns[:5]).show(5)
```

```
+------+-----+--------+---+-------+
|VMONTH|VYEAR|   VDAYR|AGE|ARRTIME|
+------+-----+--------+---+-------+
| April| 2007|Thursday| 49|   1325|
| April| 2007|  Friday| 24|    915|
| April| 2007| Tuesday| 30|    825|
| April| 2007|  Monday| 24|   1815|
| April| 2007|  Friday| 43|   2228|
+------+-----+--------+---+-------+
```

```
only showing top 5 rows
```

[27]:
```
df[0].select("ID","YEAR").show(5)
```

```
+---+----+
| ID|YEAR|
+---+----+
|  0|2007|
|  1|2007|
|  2|2007|
|  3|2007|
|  4|2007|
+---+----+
only showing top 5 rows
```

[10]:
```
# combine data first years
NHAMCS_comb = combineDf.union_d_fs(df[0],df[1])

# add the rest of the years
for i in range(2,len(df)):
    print("Concatentating: ", i)

    NHAMCS_comb = combineDf.union_d_fs(NHAMCS_comb,df[i])
```

```
Concatentating:  2
Concatentating:  3
Concatentating:  4
Concatentating:  5
Concatentating:  6
Concatentating:  7
Concatentating:  8
Concatentating:  9
Concatentating:  10
```

[11]:
```
NHAMCS_comb.count()
```

[11]:
```
305897
```

[12]:
```
NHAMCS_comb.to_csv("combined_data.csv", index = False)
```

```
␣
↪---------------------------------------------------------------------------

AttributeError                            Traceback (most recent call␣
↪last)
```

```
      <ipython-input-12-1384cf2cc7a3> in <module>
----> 1 NHAMCS_comb.to_csv("combined_data.csv", index = False)


      /usr/local/spark/python/pyspark/sql/dataframe.py in __getattr__(self,␣
↪name)
     1302         if name not in self.columns:
     1303             raise AttributeError(
-> 1304                 "'%s' object has no attribute '%s'" % (self.
↪__class__.__name__, name))
     1305         jc = self._jdf.apply(name)
     1306         return Column(jc)


      AttributeError: 'DataFrame' object has no attribute 'to_csv'
```

[10]:
```
#NHAMCS_comb.cache()
```

[10]:
```
DataFrame[VMONTH: string, VYEAR: int, VDAYR: string, AGE: string, ARRTIME: int,
WAITTIME: string, LOV: string, RESIDNCE: string, SEX: string, ETHUN: string,
RACEUN: string, ARRIVE: string, PAYPRIV: string, PAYMCARE: string, PAYMCAID:
string, PAYWKCMP: string, PAYSELF: string, PAYNOCHG: string, PAYOTH: string,
PAYDK: string, PAYTYPE: string, TEMPF: string, PULSE: string, RESPR: string,
BPSYS: string, BPDIAS: string, POPCT: string, ORIENTED: string, IMMED: string,
PAIN: string, SEEN72: string, DISCH7DA: string, PASTVIS: string, RFV1: string,
RFV2: string, RFV3: string, RFV13D: string, RFV23D: string, RFV33D: string,
EPISODE: string, INJURY: string, INTENT: string, CAUSE1: string, CAUSE2: string,
CAUSE3: string, CAUSE13D: int, CAUSE23D: int, CAUSE33D: int, VCAUSE: string,
DIAG1: string, DIAG2: string, DIAG3: string, DIAG13D: string, DIAG23D: string,
DIAG33D: string, PRDIAG1: string, PRDIAG2: string, PRDIAG3: string, DIAGSCRN:
string, CBC: string, BUNCREAT: string, CARDENZ: string, ELECTROL: string,
GLUCOSE: string, LFT: string, ABG: string, PTTINR: string, BLOODCX: string, BAC:
string, TOXSCREN: string, OTHERBLD: string, CARDMON: string, EKG: string,
PREGTEST: string, FLUTEST: string, URINE: string, WOUNDCX: string, OTHRTEST:
string, ANYIMAGE: string, XRAY: string, CATSCAN: string, CTHEAD: string,
CTNHEAD: string, CTUNK: string, MRI: string, MRIHEAD: string, MRINHEAD: string,
MRIUNK: string, ULTRASND: string, OTHIMAGE: string, TOTDIAG: string, PROC:
string, IVFLUIDS: string, CAST: string, SPLINT: string, LACREP: string,
INCDRAIN: string, DEBRIDE: string, FBREM: string, NEBUTHER: string, BLADCATH:
string, NGTUBE: string, CPR: string, ENDOINT: string, OTHPROC: string, TOTPROC:
string, MED: string, MED1: string, MED2: string, MED3: string, MED4: string,
MED5: string, MED6: string, MED7: string, MED8: string, GPMED1: string, GPMED2:
string, GPMED3: string, GPMED4: string, GPMED5: string, GPMED6: string, GPMED7:
string, GPMED8: string, NUMGIV: string, NUMDIS: string, NUMMED: int, NOPROVID:
string, ATTPHYS: string, RESINT: string, ONCALL: string, RNLPN: string, NURSEPR:
```

string, PHYSASST: string, EMT: string, OTHPROV: string, NODISP: string, NOFU: string, RETPRN: string, RETREFFU: string, REFSOCS: string, LEFTBMSE: string, LEFTAMSE: string, LEFTAMA: string, DOA: string, DIEDED: string, TRANSFER: string, RFTRANS: string, ADMITOBS: string, ADMITHOS: string, OTHDISP: string, ADMIT: string, LOS: string, HDDIAG: string, HDSTAT: string, ADISP: string, PATWT: string, REGION: string, MSA: string, OWNER: string, BLANK1: string, BLANK2: string, BLANK3: string, HOSPCODE: string, PATCODE: int, BDATEFL: string, SEXFL: string, ETHNICFL: string, RACEFL: string, IMMEDFL: string, DRUGID1: string, PRESCR1: string, CONTSUB1: string, COMSTAT1: string, RX1CAT1: string, RX1CAT2: int, RX1CAT3: int, RX1CAT4: int, RX1V1C1: int, RX1V1C2: int, RX1V1C3: int, RX1V1C4: int, RX1V2C1: int, RX1V2C2: int, RX1V2C3: int, RX1V2C4: int, RX1V3C1: int, RX1V3C2: int, RX1V3C3: int, RX1V3C4: int, DRUGID2: string, PRESCR2: string, CONTSUB2: string, COMSTAT2: string, RX2CAT1: string, RX2CAT2: int, RX2CAT3: int, RX2CAT4: int, RX2V1C1: int, RX2V1C2: int, RX2V1C3: int, RX2V1C4: int, RX2V2C1: int, RX2V2C2: int, RX2V2C3: int, RX2V2C4: int, RX2V3C1: int, RX2V3C2: int, RX2V3C3: int, RX2V3C4: int, DRUGID3: string, PRESCR3: string, CONTSUB3: string, COMSTAT3: string, RX3CAT1: string, RX3CAT2: int, RX3CAT3: int, RX3CAT4: int, RX3V1C1: int, RX3V1C2: int, RX3V1C3: int, RX3V1C4: int, RX3V2C1: int, RX3V2C2: int, RX3V2C3: int, RX3V2C4: int, RX3V3C1: int, RX3V3C2: int, RX3V3C3: int, RX3V3C4: int, DRUGID4: string, PRESCR4: string, CONTSUB4: string, COMSTAT4: string, RX4CAT1: string, RX4CAT2: int, RX4CAT3: int, RX4CAT4: int, RX4V1C1: int, RX4V1C2: int, RX4V1C3: int, RX4V1C4: int, RX4V2C1: int, RX4V2C2: int, RX4V2C3: int, RX4V2C4: int, RX4V3C1: int, RX4V3C2: int, RX4V3C3: int, RX4V3C4: int, DRUGID5: string, PRESCR5: string, CONTSUB5: string, COMSTAT5: string, RX5CAT1: string, RX5CAT2: int, RX5CAT3: int, RX5CAT4: int, RX5V1C1: int, RX5V1C2: int, RX5V1C3: int, RX5V1C4: int, RX5V2C1: int, RX5V2C2: int, RX5V2C3: int, RX5V2C4: int, RX5V3C1: int, RX5V3C2: int, RX5V3C3: int, RX5V3C4: int, DRUGID6: string, PRESCR6: string, CONTSUB6: string, COMSTAT6: string, RX6CAT1: string, RX6CAT2: int, RX6CAT3: int, RX6CAT4: int, RX6V1C1: int, RX6V1C2: int, RX6V1C3: int, RX6V1C4: int, RX6V2C1: int, RX6V2C2: int, RX6V2C3: int, RX6V2C4: int, RX6V3C1: int, RX6V3C2: int, RX6V3C3: int, RX6V3C4: int, DRUGID7: string, PRESCR7: string, CONTSUB7: string, COMSTAT7: string, RX7CAT1: string, RX7CAT2: int, RX7CAT3: int, RX7CAT4: int, RX7V1C1: int, RX7V1C2: int, RX7V1C3: int, RX7V1C4: int, RX7V2C1: int, RX7V2C2: int, RX7V2C3: int, RX7V2C4: int, RX7V3C1: int, RX7V3C2: int, RX7V3C3: int, RX7V3C4: int, DRUGID8: string, PRESCR8: string, CONTSUB8: string, COMSTAT8: string, RX8CAT1: string, RX8CAT2: int, RX8CAT3: int, RX8CAT4: int, RX8V1C1: int, RX8V1C2: int, RX8V1C3: int, RX8V1C4: int, RX8V2C1: int, RX8V2C2: int, RX8V2C3: int, RX8V2C4: int, RX8V3C1: int, RX8V3C2: int, RX8V3C3: int, RX8V3C4: int, EMRED: string, EDEMOGE: string, EPROLSTE: string, ECPOEE: string, EWARNE: string, ESCRIPE: string, ECTOEE: string, EORDERE: string, ERESULTE: string, ERANGEE: string, EIMGRESE: string, EIMAGEE: string, EPNOTESE: string, EHXFUE: string, EREMINDE: string, EPUBHTHE: string, ENOTDISE: string, EMRNEWE: string, INCSHX: string, INCPHYS: string, EXPSPACE: string, SURGDAY: string, BEDCZAR: string, BEDDATA: string, OBSUNIT: string, OBSED: string, BOARD: string, BOARDHOS: string, DIV: string, TOTHRDIVR: string, REGDIV: string, ADMDIV: string, BEDREG: string, CATRIAGE: string, FASTTRAK: string, EDPTOR: string, DASHBORD: string, RFID: string, ZONENURS: string, POOLNURS:

string, FULLCAP: string, NOOPTEFF: string, ETHIM: string, RACEIM: string, RACER: string, RACEETH: string, AGEDAYS: string, AGER: string, CAUSE1R: string, CAUSE2R: string, CAUSE3R: string, INTENTR: string, DIAG1R: string, DIAG2R: string, DIAG3R: string, HDDIAGR: int, WHOCOMP: string, SETTYPE: string, YEAR: string, CSTRATM: int, CPSUM: int, EDWT: string, PCTPOVR: string, HINCOMER: string, PBAMORER: string, URBANRUR: string, ID: bigint, MUINC: string, MEDLISTE1: string, OBSHOS: string, PRESCR10: string, MED9: string, EMSGE: string, PRESCR9: string, INJDETR2: string, RX11V3C2: int, GPMED11: string, RX9V1C3: int, RX9CAT4: string, COPD: string, HDDIAG3: string, MEDLISTE4: string, HDDIAG3R: int, LABRESE4: string, DRUGID12: string, ALGLISTE4: string, RX11V1C2: int, RX10V1C4: int, KIOSELCHK: string, RX9CAT1: int, RX10V1C2: int, COMSTAT9: string, IMMEDR: string, ESCRIPER: string, ERESULTER: string, EWARNER: string, SUTURE: string, RX9CAT2: int, RX11V2C4: int, PRESCR12: string, INJPOISADR2: string, EQOCE: string, IMAGREPE4: string, RX10V3C3: int, RX11V2C1: int, MEDLISTUNKE: string, TOTCHRON: string, EWHOUNKPE: string, OBSPHYSUN: string, BNP: string, RX12V3C1: int, RX10V1C1: int, OBSDIS: string, OBSSTAY: string, EWHOUNKLE: string, DRUGID9: string, RX11V3C3: int, MEDLISTE3: string, DRUGID11: string, RX9V2C3: int, ESHAREE: string, EINSFASTE: string, CANCER: string, IMAGREPUNKE: string, EHLTHINFOER: string, RX10CAT4: int, EMUREPER: string, INJPOISADR1: string, PHYSPRACTRIA: string, ESETSER: string, HDDIAG2: string, PTPROBE3: string, RX10CAT3: int, PAINSCALE: string, RX11V2C3: int, SKINADH: string, PTPROBE1: string, RX10V2C4: int, CONSULT: string, EGRAPHE: string, ESUMER: string, ESMOKEE: string, ESHAREOTHE: string, ESHAREUNKE: string, RX9V3C1: int, CTAB: string, ESETSE: string, EINSE: string, HDDIAG13D: string, RX12CAT2: int, DEMENTIA: string, DVT: string, RX9V1C4: string, HIVTEST: string, INJR2: string, CASTSPLINT: string, EHRINSE: string, EVITALER: string, GPMED9: string, WIRELESS: string, EQOCER: string, INTENDYR: string, PTPROBUNKE: string, HDDIAG33D: string, CONTSUB10: string, ECQMER: string, ALGLISTE2: string, RX12V1C1: int, LABRESREFE: string, ALGLISTE1: string, RX9V3C3: int, HDDIAG23D: string, COMSTAT11: string, LABRESUNKE: string, EGENLISTER: string, RX12CAT1: int, LABRESE3: string, EIMMREGE: string, TRANOTH: string, RX9V1C2: int, RX12V2C2: int, ECTOEER: string, RX12V1C3: string, EMUREPE: string, EWHOREFPE: string, EGENLISTE: string, ECPOEER: string, RX9V2C1: int, RX11CAT3: int, IMAGREPE2: string, NOCHRON: string, OBSUNITS: string, RX11V3C4: int, MIHX: string, RX12V3C3: string, EMSGER: string, EWHOPRACPE: string, OBSHOSP: string, RX9V2C2: int, EDDIAL: string, RX10CAT2: int, RX11V2C2: int, CTCHEST: string, CONTSUB11: string, INJPOISAD: string, RX12V2C4: string, URINECX: string, MHPROV: string, RX9V3C2: int, EVITALE: string, RX9CAT3: int, CONTSUB12: string, AMBDIV: string, RETRNED: string, ALGLISTE3: string, EMEDALGER: string, ADMTPHYS: string, IMBED: string, TRANPSYC: string, FIPSSTHOSP: string, ESHAREREFE: string, ECQME: string, RX12CAT3: string, EPROLSTER: string, CTOTHER: string, CHF: string, RX9V1C1: int, RX12V2C3: string, IMAGREPREFE: string, IMAGREPE3: string, RX10V3C1: int, RACERETH: string, RX11V1C1: int, COMSTAT12: string, RX10V3C4: string, EIMGRESER: string, INJDETR1: string, RX11CAT4: int, HDDIAG1R: int, NOPAY: string, DIABETES: string, RX12V1C2: int, RX9V2C4: string, EHLTHINFOE: string, TRANNH: string, AGEFL: string, LABRESE1: string, PTPROBE4: string, MED12: string, ESUME: string, HDDIAG1: string, RX11CAT1: int, EIMMREGER: string,

BPAP: string, RX11CAT2: int, RX11V3C1: int, INJR1: string, EINSHOWE: string, LACTATE: string, EORDERER: string, EPNOTESER: string, LABRESE2: string, IMAGREPE1: string, EBILLANYE: string, RX10V3C2: int, HLISTED: string, EDHIV: string, MEDLISTREFE: string, ALGLISTREFE: string, RX12V3C4: string, MED10: string, LEFTATRI: string, EWHOOTHLE: string, EWHOOTHEPE: string, CONTSUB9: string, GPMED10: string, CENTLINE: string, RX10V2C1: int, MEDLISTE2: string, PTPROBREFE: string, EGRAPHER: string, LEFTBTRI: string, ONO2: string, DDIMER: string, ESHAREWEBE: string, PELVIC: string, STAY24: string, RX12V2C1: int, ESHAREEHRE: string, DRUGID10: string, RX12V1C4: string, EREMINDER: string, RX11V1C3: int, RACERFL: string, EWHOPRACLE: string, GPMED12: string, ALGLISTUNKE: string, RX10V2C3: int, PAYTYPER: string, COMSTAT10: string, OBSPHYSOT: string, RX10V2C2: int, ESMOKEER: string, RX10CAT1: int, CEBVD: string, RX11V1C4: int, PRESCR11: string, RX12CAT4: string, MED11: string, PTPROBE2: string, OBSPHYSED: string, HLIST: string, EDEMOGER: string, RX9V3C4: string, RX12V3C2: int, EMEDALGE: string, ADVTRIAG: string, LUMBAR: string, RX10V1C3: int, EWHOREFLE: string, IVCONTRAST: string, INJDETR: string, ARREMS: string, HDDIAG2R: int, EHRWHO6E: string, EBILLRECE: string, IMMEDRFL: string, OBSDECMD: string, EHRWHO1E: string, GCS: string, EHRWHO4ER: int, EHRWHO2ER: int, PAYHITH: string, BOARDED: string, EHRWHO5E: string, EHRWHO2E: string, EMEDSE: string, ADVCOMP1: string, EHRWHO5ER: int, EALLERGE: string, BLANK5: string, BLANK4: string, EHRWHO3E: string, EHRWHO7E: string, EHRWHO6ER: int, ERESEHRE: string, EHRWHO7ER: int, PAYYRH: string, ADVCOMP2: string, EHRWHO3ER: int, EHRWHO4E: string, EHRWHO1ER: int, RX28V1C1: int, GPMED20: string, RX27CAT3: string, RX25V2C1: int, RX17V3C3: string, RX28V3C2: string, RX22V1C4: string, RX22V2C2: int, RX30V3C2: string, RX28V2C4: string, RX28V2C3: string, RX20V3C1: int, RX28CAT2: string, RX27CAT2: string, RX20CAT4: string, PRESCR26: string, DIABTYP1: string, RX15CAT4: int, RX28V1C2: string, DIABTYP0: string, RX18V2C2: int, RX21V3C3: int, RX17CAT1: int, RX20V2C2: int, CONTSUB17: string, DRUGID16: string, RX24CAT2: int, COMSTAT13: string, RX13V2C1: int, RX13CAT3: int, RX21V1C4: string, RX15CAT2: int, MED14: string, EDPRIM: string, RX18CAT3: string, RX18V3C3: string, MED18: string, RX23V2C2: int, RX14V1C2: int, RX24V3C3: string, RX27V2C3: string, RX14V3C2: int, COMSTAT23: string, RX16V1C2: int, RX14V1C1: int, RX22V2C1: int, RX25V1C3: string, RX23V1C2: int, MED21: string, RX27V2C2: string, RX30V2C3: string, RX25V3C4: string, RX21V1C1: int, RX19V1C1: int, RX14V1C3: int, RX30V1C1: int, RX13V2C2: int, RX27V1C2: string, OBESITY: string, RX16V3C1: int, INTENT15: string, COMSTAT20: string, RX18CAT1: int, RX28V3C1: int, RX22V1C3: string, RX27V1C3: string, RX26V1C1: int, RX26V1C2: string, ESUMCSRE2: string, VITALSD: string, RX16CAT3: int, EDISCHSRE3: string, MED17: string, RX19V2C4: string, RX16V3C3: int, DRUGID21: string, RX21CAT2: int, MED25: string, GPMED24: string, RX22V3C4: string, RX29CAT2: string, RX17CAT4: string, RX30V1C4: string, GPMED19: string, EDISCHSRE2: string, RX24V2C4: string, RX29V1C1: int, RX26V2C2: string, RX15V1C4: int, RX18V1C4: string, PRESCR27: string, CONTSUB28: string, COMSTAT29: string, RX20CAT3: int, TEMPDF: string, RX24V3C1: int, ECONTSCRIPR: string, ALZHD: string, RX30V3C4: string, RX30V2C2: string, RX14CAT1: int, PRESCR30: string, RX20V2C3: int, DIABTYP2: string, GPMED27: string, RX18V1C2: int, RX19V3C4: string, DIAG5: string, GPMED29: string, PTONLINEE2: string, PTONLINEE5: string, RX20V3C3: string, RX30V2C4:

string, RX14V2C2: int, RX21V2C3: int, RX22CAT3: string, RX15V3C1: int, RX20V2C1: int, RX18V1C1: int, RX15V2C3: int, RX18V2C1: int, RX22V1C2: int, PULSED: string, RX19CAT4: string, COMSTAT14: string, HYPLIPID: string, PRESCR18: string, RX27V1C1: int, SUBSTAB: string, RX18CAT2: int, EEDSRE2: string, RX19CAT3: string, DRUGID23: string, RX19V2C2: int, RX22CAT2: int, RX19V1C3: string, CONTSUB26: string, CONTSUB15: string, GPMED30: string, RX14V2C1: int, COMSTAT18: string, RX23V2C3: string, RX26CAT1: int, RX18V2C4: string, RX22CAT1: int, RX25V2C3: string, RX13V2C4: int, RX29V1C2: string, GPMED15: string, COMSTAT16: string, RX30V1C2: string, RX23V3C2: int, RX27V1C4: string, DRUGID27: string, RX13V1C2: int, RX24V1C4: string, RX14V2C3: int, RX13V1C1: int, RX30CAT1: int, RX19CAT1: int, RX19CAT2: int, RX14CAT4: string, MED30: string, RX29V2C4: string, OTHCX: string, RX28V1C3: string, RX13V3C1: int, COMSTAT24: string, RX20V1C2: int, RX29V2C1: int, RX26CAT3: string, GPMED17: string, EDATAREPER: string, DRUGID19: string, MED20: string, HDDIAG5: string, EEDSRE1: string, HPE: string, RX13V1C3: int, RX17V2C1: int, RX30V1C3: string, HDDIAG4: string, EMEDIDER: string, RX29CAT3: string, RX17CAT2: int, RX18V3C1: int, RX16V3C4: string, RX24V3C2: int, CONTSUB25: string, RX28CAT1: int, RX20V1C4: string, RX23V3C4: string, BPSYSD: string, RX13V3C2: int, RX30V3C1: int, RX22V3C2: int, RX20V2C4: string, RX15V1C3: int, DRUGID22: string, RX26V2C4: string, COMSTAT19: string, PTONLINEE6: string, AMBTRANSFER: string, ESHARERE: string, RX25V1C4: string, EIDPTER: string, DRUGID30: string, RX22V3C1: int, RX24V2C3: string, RFV43D: string, RX24V1C1: int, DRUGID13: string, RX19V1C4: string, MED19: string, RX27CAT4: string, RX24V1C2: int, RX25V2C2: string, MED13: string, DRUGID29: string, RX21V3C4: string, RX22CAT4: string, RX16V1C3: int, PRESCR24: string, RX16CAT4: string, RX28V2C1: int, DRUGID17: string, RX27V3C4: string, RX24V2C1: int, RX28V3C3: string, RX25CAT1: int, RX23CAT4: string, RX16V1C1: int, RX25CAT2: string, RX17V3C2: int, ETOHAB: string, COMSTAT30: string, RX22V2C4: string, RX26V2C1: int, RX20V3C4: string, PRDIAG4: string, RX14V3C3: int, RX14CAT2: int, RX15V3C3: int, DRUGID26: string, RX19V1C2: int, CONTSUB23: string, RX24V1C3: string, RFV5: string, RX29V2C3: string, INJURY_ENC: string, RX17V3C4: string, RX14V3C4: string, RX24CAT1: int, CONTSUB20: string, RX27V3C2: string, RFV4: string, RX13CAT1: int, CONTSUB29: string, OSA: string, RX13CAT2: int, INJURY72: string, RX23CAT1: int, COMSTAT28: string, MED24: string, RX26V1C3: string, PRESCR16: string, DRUGID18: string, RX13V2C3: int, CONTSUB21: string, HTN: string, RX21CAT4: string, RX22V2C3: string, GPMED25: string, RX26V3C4: string, CTCONTRAST: string, MED29: string, RX17V2C4: string, RX21V2C1: int, RX19V3C2: int, RX21V3C2: int, DRUGID28: string, PRESCR15: string, RX16CAT2: int, RX17V1C1: int, RX13CAT4: int, GPMED23: string, PRESCR28: string, MED26: string, LBTC: string, COMSTAT26: string, BPDIASD: string, RX26V2C3: string, RX26V3C3: string, GPMED22: string, CONTSUB27: string, RX15CAT1: int, RX30V2C1: int, RX13V3C4: int, GPMED21: string, MED22: string, COMSTAT15: string, RX23V2C4: string, RX14V1C4: string, RX17V3C1: int, RX16V3C2: int, RX16V2C4: string, COMSTAT22: string, RX18V3C2: int, RX15V1C2: int, RX29V3C2: string, RX21V3C1: int, RX20V3C2: string, RX25CAT3: string, PRESCR17: string, RX29V3C4: string, RX17CAT3: string, RESPRD: string, RX14CAT3: int, MED16: string, RX15V1C1: int, CONTSUB16: string, GPMED28: string, RX16CAT1: int, PRESCR21: string, RX27V2C4: string, RX29V2C2: string, DRUGID20: string, RX23V1C3: string, RX24V3C4: string, PRESCR14: string,

RX15V3C4: int, RX15V2C2: int, RX27V3C3: string, EDISCHSRE1: string, PRESCR22: string, RX19V3C1: int, MED28: string, RX29V3C3: string, RX19V2C3: string, RX20CAT2: int, OBSSEP: string, MRICONTRAST: string, CONTSUB19: string, BMP: string, HHSMUE: string, GPMED16: string, RX29V3C1: string, RX18V3C4: string, CKD: string, ESUMCSRE3: string, RX17V1C3: string, PTONLINEE1: string, RX26CAT2: string, MED15: string, CONTSUB24: string, OSTPRSIS: string, RX16V2C2: int, CONTSUB13: string, RX15CAT3: int, RX30CAT3: string, DEPRN: string, RX21CAT3: int, RX30CAT2: string, RX26V3C2: string, ERADIER: string, MED27: string, LWBS: string, RX23CAT3: string, GPMED13: string, RX23CAT2: int, RX23V3C1: int, RX24V2C2: int, CONTSUB14: string, RX17V1C4: string, RX22V1C1: int, RX28CAT4: string, PTONLINEE3: string, EMEDRES: string, RX28V1C4: string, RX30CAT4: string, RX17V2C2: int, RX20V1C3: int, RX24CAT3: string, RX26CAT4: string, PRESCR13: string, CONTSUB22: string, RX21V2C4: string, RX27V3C1: int, RX21V1C2: int, RX16V2C1: int, PRDIAG5: string, RX18CAT4: string, RX25V2C4: string, RX25V3C3: string, GPMED26: string, ECONTR: string, GPMED14: string, MED23: string, RX16V1C4: string, RX26V1C4: string, COMSTAT21: string, RX14V2C4: string, RX15V2C4: int, ASTHMA: string, RX29V1C3: string, DRUGID14: string, ESRD: string, GPMED18: string, CMP: string, RX25V3C2: string, ESHARESE: string, RX21V1C3: int, RX13V3C3: int, DIAG4: string, RX18V2C3: string, RX29CAT1: int, RFV53D: string, DRUGID24: string, RX17V1C2: int, EDINFO: string, RX20CAT1: int, ESUMCSRE1: string, RX23V1C4: string, RX29CAT4: string, PRESCR20: string, RX23V2C1: int, RX24CAT4: string, RX25V3C1: int, RX23V1C1: int, RX16V2C3: int, RX25V1C2: string, RX17V2C3: string, RX21CAT1: int, EEDSRE3: string, RX18V1C3: string, PRESCR23: string, CONTSUB18: string, CAD: string, RX15V2C1: int, COMSTAT27: string, RX13V1C4: int, RX27V2C1: int, RX23V3C3: string, DRUGID15: string, RX20V1C1: int, RX25V1C1: int, RX29V1C4: string, PTONLINEE4: string, RX15V3C2: int, PRESCR29: string, RX28V3C4: string, RX30V3C3: string, RX28CAT3: string, RX19V2C1: int, RX19V3C3: string, RX26V3C1: int, COMSTAT25: string, COMSTAT17: string, RX14V3C1: int, OBSCLIN: string, RX27CAT1: int, CONTSUB30: string, DRUGID25: string, RX22V3C3: string, RX21V2C2: int, RX25CAT4: string, RX28V2C2: string, PRESCR19: string, TRTCX: string, PRESCR25: string, ESHAREPROVE7: string, EOUTINFOE: string, DIAG53D: string, ERADIE: string, DIAG43D: string, EOUTINCORPE: string, SECURCHCKE: string, ESHAREPROVE6: string, EOUTYPUNK: string, EOUTHOWE3: string, EPTEDUE: string, DIFFEHRE: string, EFORMULAE: string, EOUTHOWE1: string, ESHAREPROVE5: string, EOUTYPREF: string, MUSTAGE1: string, EOUTYPE1: string, EPTRECE: string, ESHAREPROVE2: string, DIAG4R: int, EPTRECER: string, MUSTAGE2: string, EOUTYPE3: string, ESHAREPROVE4: string, EOUTHOWE2: string, EOUTYPE5: string, ESHAREPROVEREF: string, HDDIAG53D: string, ESHAREPROVEUNK: string, ESHAREPROVE1: string, EOUTHOWE4: string, EPTEDUER: string, ESHAREPROVE3: string, HDDIAG5R: int, EOUTHOWREF: string, EMEDIDE: string, EFORMULAER: string, EIDPTE: string, HDDIAG4R: int, HDDIAG43D: string, EOUTYPE4: string, EOUTYPE2: string, EOUTHOWUNK: string, EHRTOEHRE: string, DIAG5R: int, OBSDEC: string, MUYEAR: string, EWHOOTHER: string, EINSELIGE: string, EWHONOTE: string, EWHOOTHE: string, EWHONOTER: string, EXCHSUM5E: string, BLANK7: string, EXCHSUM4E: string, EXCHSUM1E: string, BLANK8: string, EWHOPRACER: string, EWHOPRACE: string, EXCHSUM6E: string, BLANK6: string, EXCHSUM2E: string, EWHOUNKER: string, EXCHSUM3E: string, EXCHSUME: string, EPUBHLTHE: string, EWHOUNKE: string]

1.4 Write data to parquet file

[13]:
```
%%time
# write out data
NHAMCS_comb.write.parquet("../data/NHAMCS.2007-2017")
```

```
CPU times: user 41 ms, sys: 15 ms, total: 56 ms
Wall time: 7min 1s
```

[ ]:
```

```