Healthcare Information System

Fall 2017

CMS Medicare and Medicaid EHR Incentive Program, Electronic Health Record Products Used for Attestation using Python

Submitted to: Dr. Shilpa Balan

Team Members: Alekhya Raidu Bojja Venkata CIN:305829386

A. Data sets

URL of dataset:

https://www.healthdata.gov/dataset/cms-medicare-and-medicaid-ehr-incentive-program-electronic-health-record-products-used

Data Description

The Medicare Electronic Health Record (EHR) Incentive Program provides incentives to eligible clinicians and hospitals to adopt electronic health records. This dataset combines meaningful use attestations from the Medicare EHR Incentive Program and certified health IT product data from the ONC Certified Health IT Product List (CHPL) to identify the unique vendors, products, and product types of each certified health IT product used to attest to meaningful use. (data, 2017)

Data set merges information about the Centers for Medicare and Medicaid Services, Medicare and Medicaid EHR Incentive Programs attestations with the Office of the National Coordinator for Health IT Certified Health IT Products List. This new dataset enables systematic analysis of the distribution of certified EHR vendors and products among those providers that have attested to meaningful use within the CMS EHR Incentive Programs. The data set can be analyzed by state, provider type, provider specialty, and practice setting. (Technology, 2017)

The dataset also includes important provider-specific data, related to the provider's participation and status in the program, unique provider identifiers, and other characteristics unique to each provider, like geography and provider type. Because providers may declare more than one EHR product when attesting, this list also provides a unique ID (i.e. NPI) for each provider. The Medicare EHR Incentive Program provides incentive payments to eligible providers as they adopt, implement, upgrade, or demonstrate meaningful use of certified EHR technology. The CHPL

provides the authoritative, comprehensive listing of certified health IT products that have been tested under the ONC Certification Program. (data, 2017)

The complete dataset exceeds 1 million rows of data. This data is intended to provide names of EHR products and their vendors, the certification classification of each product (Complete or Modular), the healthcare setting for which the product was certified (Ambulatory or Inpatient), the type of provider attesting to "meaningful use" of an HER, the Incentive Program the provider attested in (Medicare or Medicare/Medicaid), Unique ID for each attestation, Version of EHR product and the Stage of Meaningful Use that the provider attested to (Stage 1/Stage 2). The size of the dataset is 370 MB with 23 columns giving all the necessary information about it. The information in this dataset is from April 2011 till present which is very useful for finding interesting trends from this dataset.

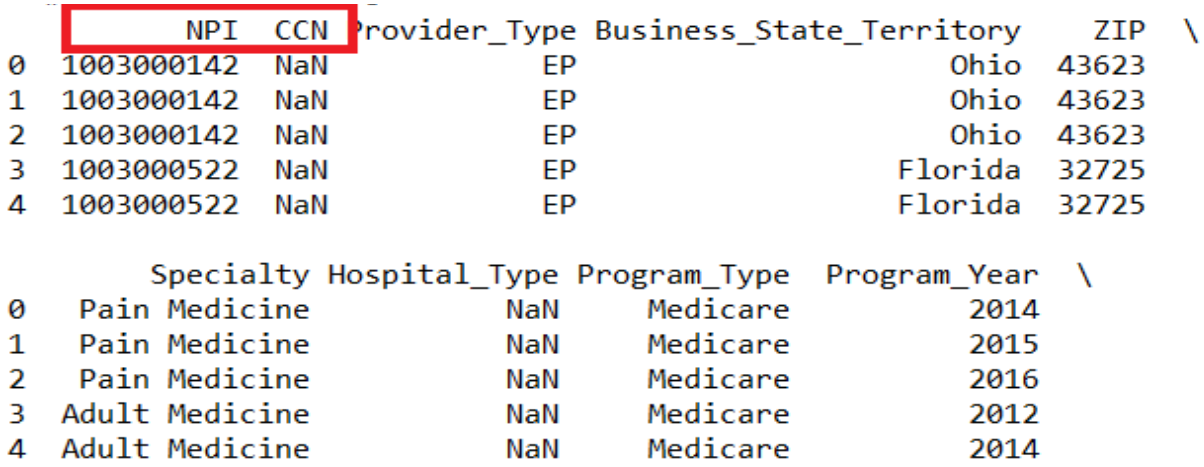
B. Data Cleaning

1. Rename columns: Renamed NPI and CCN columns to 'National Provider Identifier' and 'CMS Certified Number'. **List** has been used to store the new list of column names and then assigned that list as column names.

   (Highlights from Python script – **Pandas Data Frame**, **List**)

```
@author: bhagy
"""
import pandas as pd

new_lables=['National_Provider_Identifier','CMS_Certified_Number', 'Provider_Type',
'Business_State_Territory', 'ZIP', 'Specialty', 'Hospital_Type', 'Program_Type',
'Program_Year', 'Provider_Stage_Number', 'Payment_Year', 'Attestation_Month',
'Attestation_Year', 'MU_Definition_2014', 'Stage_2_Scheduled_2014',
'EHR_Certification_Number', 'EHR_Product_CHP_Id', 'Vendor_Name', 'EHR_Product_Name',
'EHR_Product_Version', 'Product_Classification', 'Product_Setting', 'Product_Certification_Edition_Yr']
df=pd.read_csv('MU_REPORT.csv', header=0,names= new_lables)
print(df.head())
```

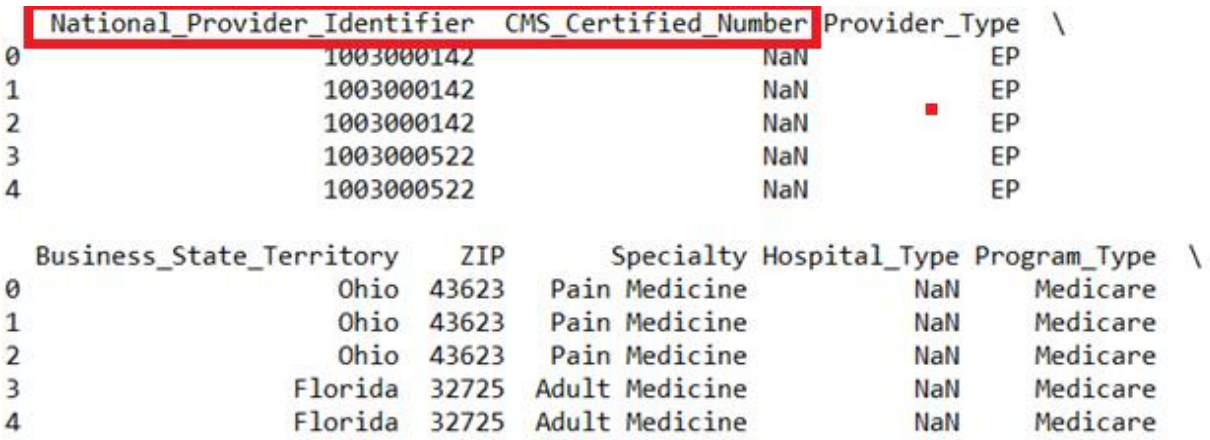
Before Cleaning:



	NPI	CCN	Provider_Type	Business_State_Territory	ZIP	\
0	1003000142	NaN	EP	Ohio	43623	
1	1003000142	NaN	EP	Ohio	43623	
2	1003000142	NaN	EP	Ohio	43623	
3	1003000522	NaN	EP	Florida	32725	
4	1003000522	NaN	EP	Florida	32725	

	Specialty	Hospital_Type	Program_Type	Program_Year	\
0	Pain Medicine	NaN	Medicare	2014	
1	Pain Medicine	NaN	Medicare	2015	
2	Pain Medicine	NaN	Medicare	2016	
3	Adult Medicine	NaN	Medicare	2012	
4	Adult Medicine	NaN	Medicare	2014	

After Cleaning:



	National_Provider_Identifier	CMS_Certified_Number	Provider_Type	\
0	1003000142	NaN	EP	
1	1003000142	NaN	EP	
2	1003000142	NaN	EP	
3	1003000522	NaN	EP	
4	1003000522	NaN	EP	

	Business_State_Territory	ZIP	Specialty	Hospital_Type	Program_Type	\
0	Ohio	43623	Pain Medicine	NaN	Medicare	
1	Ohio	43623	Pain Medicine	NaN	Medicare	
2	Ohio	43623	Pain Medicine	NaN	Medicare	
3	Florida	32725	Adult Medicine	NaN	Medicare	
4	Florida	32725	Adult Medicine	NaN	Medicare	

2. Split Column EHR Product CHP ID:

Column EHR Product CHP ID had product ID. Here, **String inbuilt function 'split'** has been used to split the column into 'EHR Product Chip' and 'EHR Product ID'. By splitting this column values of product can be used for further analysis.

(Highlights from Python script – **Pandas Data Frame**, **String function 'Split'**)

```
@author: bhagy
"""

import pandas as pd

df=pd.read_csv("MU_REPORT.csv")
#Using 'String' inbuilt split function to split 'EHR_Product_CHP_Id'
df[['EHR_Product_Chip', 'EHR_Product_ID']] = df['EHR_Product_CHP_Id'].str.split('-', expand=True)

print(df.head())
```

Before Cleaning:

EHR_Product_CHP_	Vendor
CHP-022045	Epic Syster
CHP-022044	Epic Syster
CHP-027887	Epic Syster
CHP-007425	NextGen H
CHP-018548	NextGen H
CHP-021999	NextGen H
CHP-023369	NextGen H

After Cleaning:

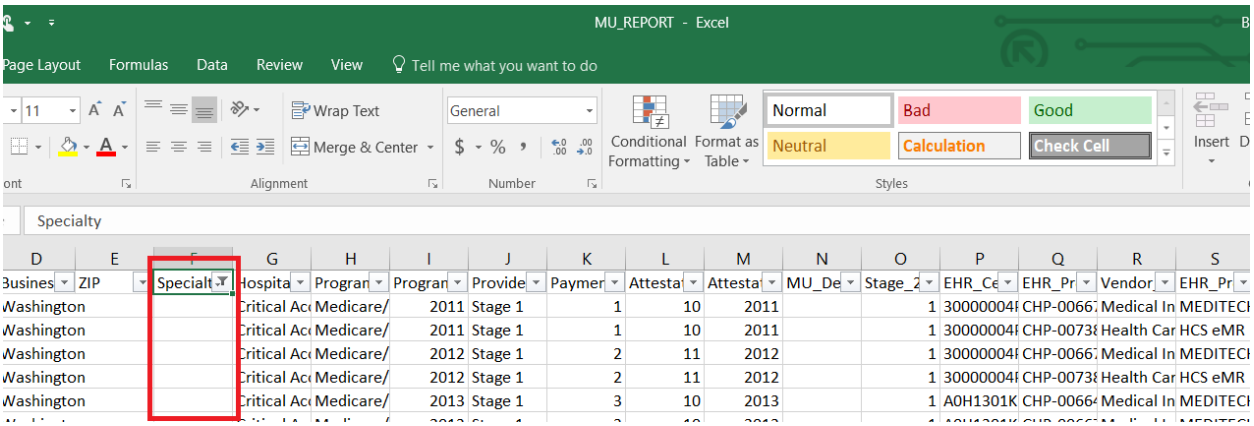
```
  Product_Classification Product_Setting Product_Certification_Edition_Yr  \
0           Complete EHR      Ambulatory                              2014
1           Complete EHR      Ambulatory                              2014
2           Complete EHR      Ambulatory                              2014
3           Complete EHR      Ambulatory                              2011
4           Complete EHR      Ambulatory                              2011

  EHR_Product_Chip EHR_Product_ID
0              CHP         022045
1              CHP         022044
2              CHP         027887
3              CHP         007425
4              CHP         018548
```

3. Remove missing values in the column ‘Specialty’: Removed missing values from the Specialty column and saved cleaned file in a new csv file.

   (Highlights from Python script – **Pandas Data Frame**, **File, Dropna**)

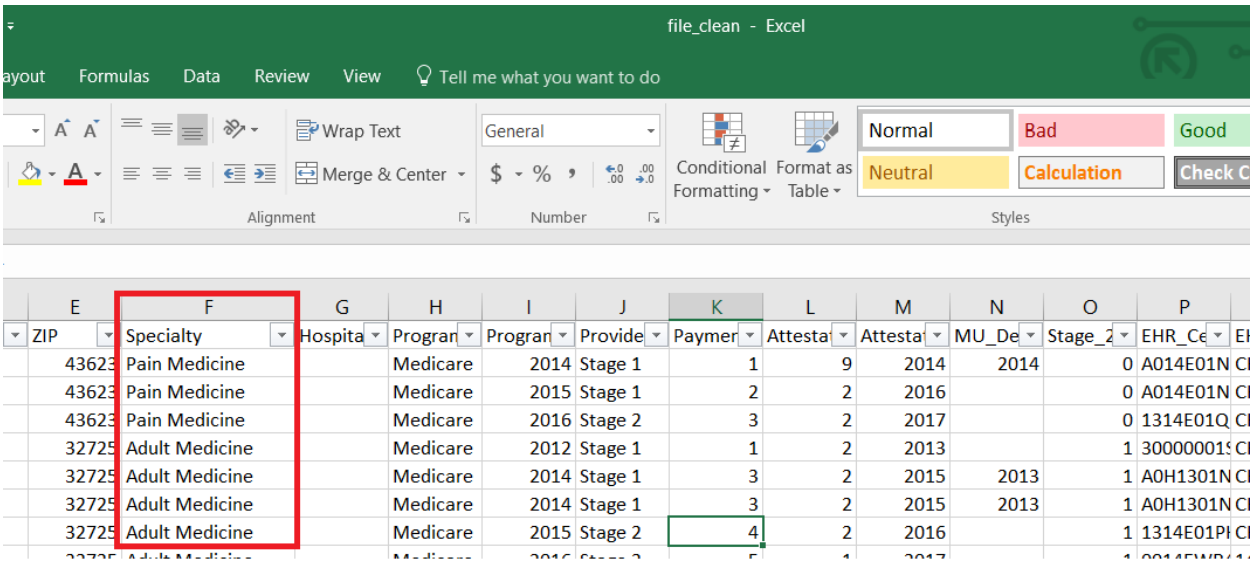
```
@author: bhagy
"""
import pandas as pd

df=pd.read_csv("MU_REPORT.csv")
new_df=df.dropna(subset=['Specialty'], how='any')
#Using 'file' attribute to save the dataset in new CSV file
new_df.to_csv("file_clean.csv",index=False)
```

Before Cleaning:



D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Busines	ZIP	Specialt	Hospita	Progran	Progran	Provide	Paymer	Attesta	Attesta	MU_De	Stage_2	EHR_Ce	EHR_Pr	Vendor	EHR_Pr
Washington			Critical Ac	Medicare/	2011	Stage 1	1	10	2011		1	30000004	CHP-0066	Medical In	MEDITECH
Washington			Critical Ac	Medicare/	2011	Stage 1	1	10	2011		1	30000004	CHP-0073	Health Car	HCS eMR
Washington			Critical Ac	Medicare/	2012	Stage 1	2	11	2012		1	30000004	CHP-0066	Medical In	MEDITECH
Washington			Critical Ac	Medicare/	2012	Stage 1	2	11	2012		1	30000004	CHP-0073	Health Car	HCS eMR
Washington			Critical Ac	Medicare/	2013	Stage 1	3	10	2013		1	A0H1301K	CHP-0066	Medical In	MEDITECH

After cleaning:



E	F	G	H	I	J	K	L	M	N	O	P
ZIP	Specialty	Hospita	Progran	Progran	Provide	Paymer	Attesta	Attesta	MU_De	Stage_2	EHR_Ce
43623	Pain Medicine		Medicare	2014	Stage 1	1	9	2014	2014	0	A014E01N
43623	Pain Medicine		Medicare	2015	Stage 1	2	2	2016		0	A014E01N
43623	Pain Medicine		Medicare	2016	Stage 2	3	2	2017		0	1314E01Q
32725	Adult Medicine		Medicare	2012	Stage 1	1	2	2013		1	30000001
32725	Adult Medicine		Medicare	2014	Stage 1	3	2	2015	2013	1	A0H1301N
32725	Adult Medicine		Medicare	2014	Stage 1	3	2	2015	2013	1	A0H1301N
32725	Adult Medicine		Medicare	2015	Stage 2	4	2	2016		1	1314E01P

C. Show/Apply Summary Statistics

```
import pandas as pd

df=pd.read_csv('MU_REPORT.csv')

#summary statistics

stat=df.EHR_Product_Name.value_counts() // return object containing count of unique values

print('Min:')

print(stat.min())

print('Max:')

print(stat.max())

print('Mean:')

print(stat.mean())

print('Standard Deviation:')

print(stat.std())

print(stat.describe())// Generates descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.
```

```
  # -*- coding: utf-8 -*-
Min:
1
Max:
128614
Mean:
719.680851064
Standard Deviation:
4818.89966014
count      1457.000000
mean        719.680851
std        4818.899660
min           1.000000
25%           7.000000
50%          37.000000
75%         187.000000
max      128614.000000
Name: EHR_Product_Name, dtype: float64
```

D. Analysis & Visualizations

1. What are the top 10 EHR vendors who developed the product certified for meaningful use attestation?

Code:

import pandas as pd

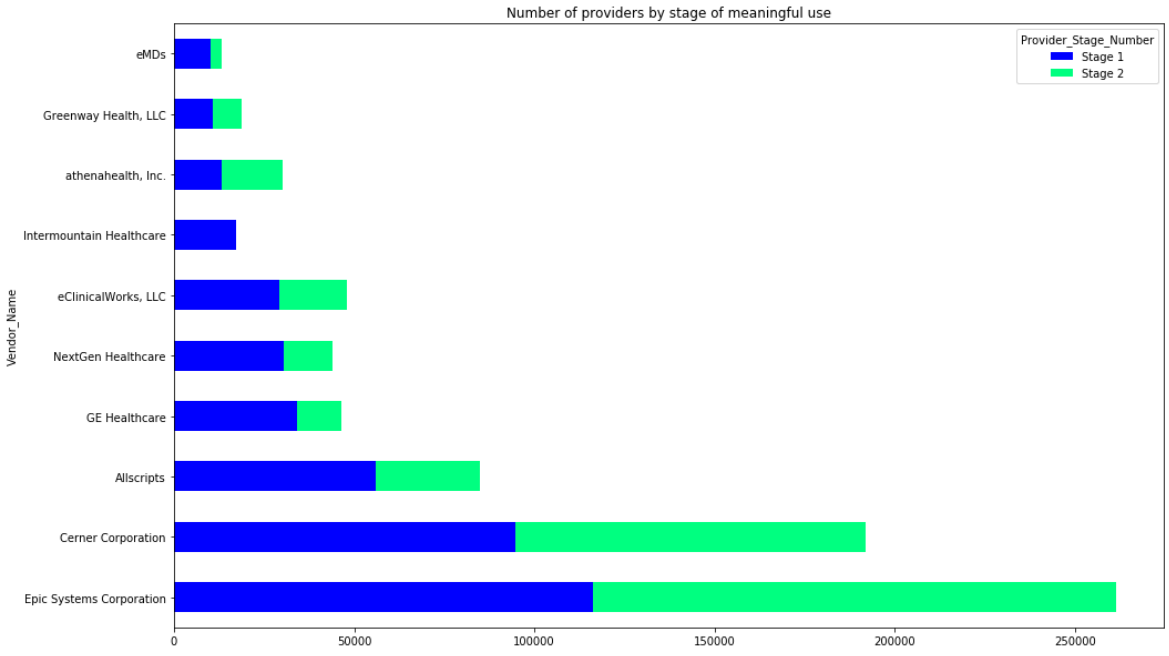
import matplotlib. pyplot as plt // `matplotlib.pyplot` is a collection of command style functions that make matplotlib work like MATLAB. Each `pyplot` function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc

df=pd.read_csv("file_clean.csv")

vendors=df.groupby('Vendor_Name').Provider_Stage_Number.value_counts().unstack()

```
new_vendor=vendors.sort_values(by='Stage 1', ascending=False)[:10]

new_vendor.plot(kind='barh', stacked=True, figsize=[16,10], colormap='winter',

title='Number of providers by stage of meaningful use')

plt.show()
```

Screenshot:



(Highlights from Python script – **Pandas Data Frame**, Stacked bar chart, different colors)

The above Stacked Bar chart shows top 10 Electronic Health Record (EHR) vendor names by their stage. There are three stages over the life of Meaningful Use. Stage 2 began in 2014. Only stage 1 and stage 2 appear in this data set. We can see that Epic System Corporation is a vendor who has highest number of products. This vendor has more than 250,000 products from which 116,226 products are at Stage 1 and 145,282 are at stage 2. This shows that more number of products were at the stage of meaningful use attestation after 2014. 'Cerner Corporation', 'Allscripts', 'GE

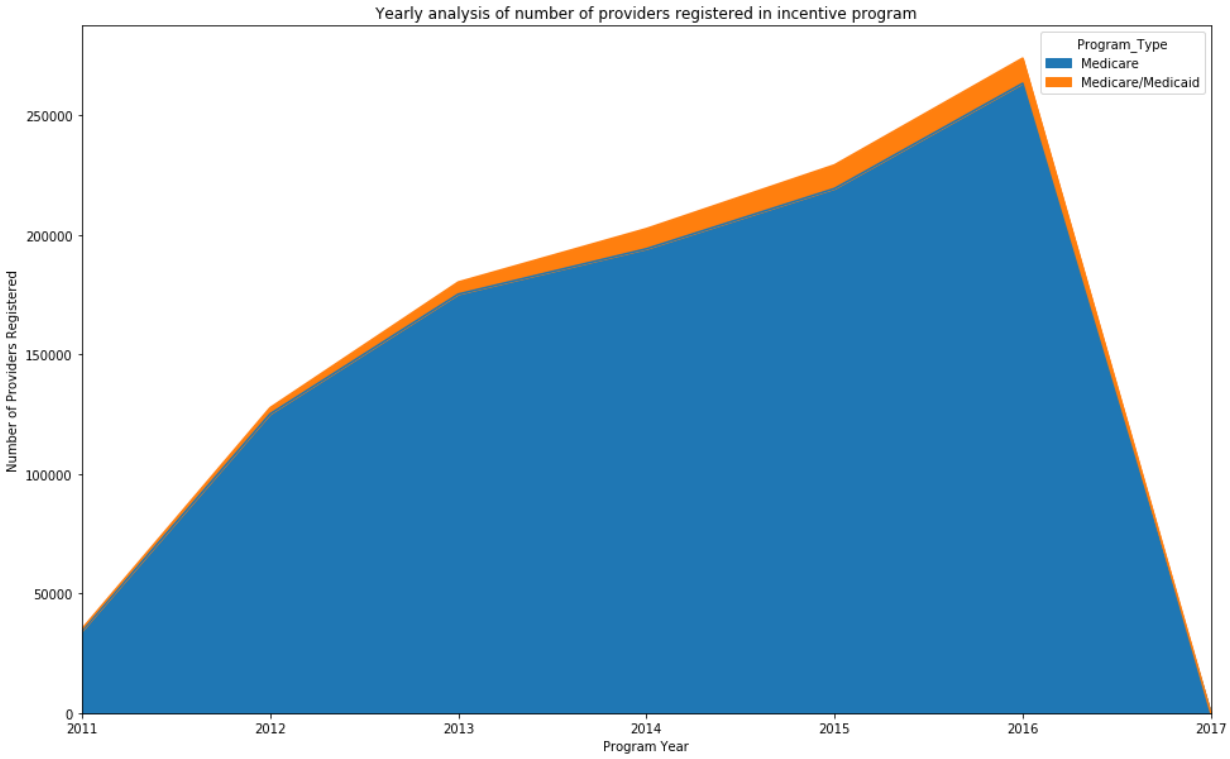
Healthcare', 'Nextgen Healthcare', 'eClinicalWorks', 'Intermountain Healthcare', Athenahealth', 'Greenway Health', 'eMDs' are some of the top 10 vendors.

2. How number of providers registered in incentive program are changing every year from 2011?

   Code:

```
import pandas as pd
import matplotlib.pyplot as plt
#Pandas Data Frame has been used here
df=pd.read_csv("file_clean.csv")
#User defined function has been used
def prog_year(df):
    df1=df.groupby('Program_Year').Program_Type.value_counts().unstack()
    df1.plot(kind='area', figsize=[16,10], stacked=True, title='Yearly analysis of number of providers registered in incentive program')
    plt.ylabel('Number of Providers Registered')
    plt.xlabel('Program Year')
prog_year (df)
```

   Screenshot:



(Highlights from Python script – **Pandas Data Frame**, **User Defined function**, Area Line chart, different colors, Matplotlib)

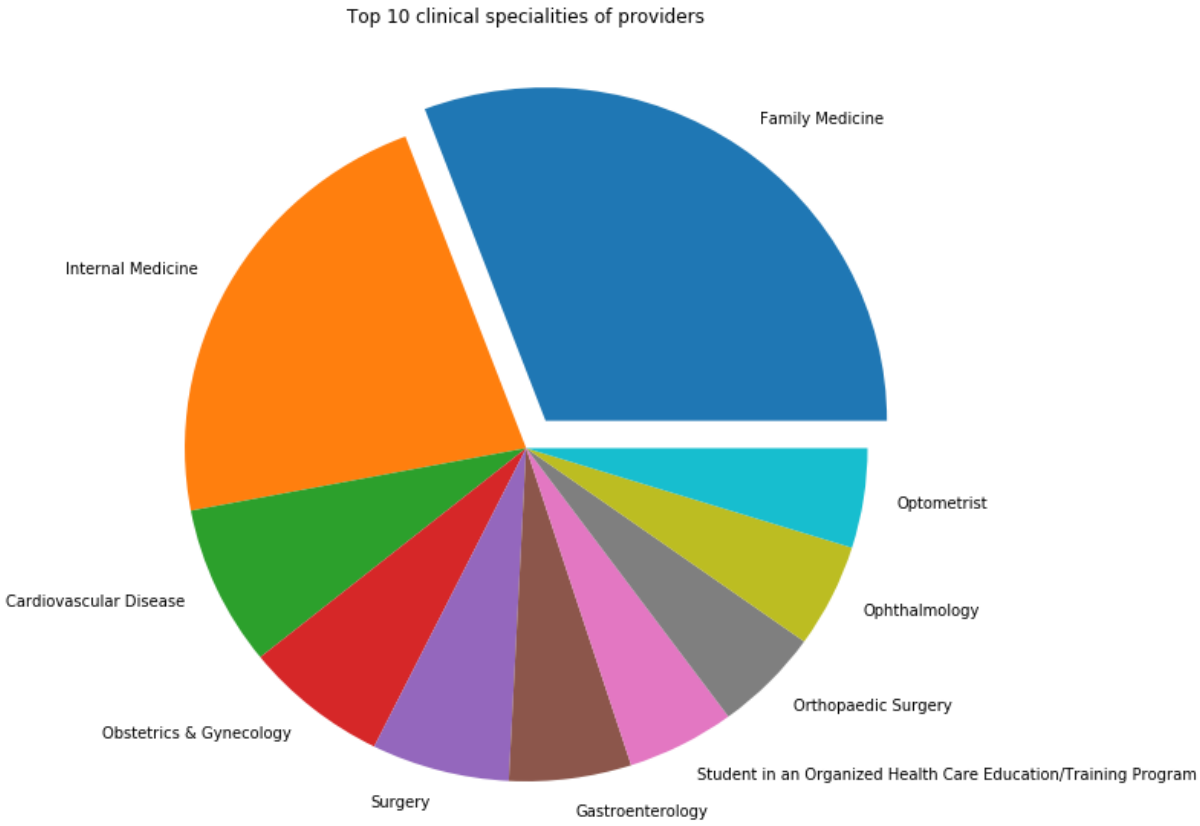
As seen from the above time series area chart, number of providers registered in incentive program have been increasing every year since 2011. This shows that more and more providers are getting registered and adopting, implementing, upgrading or demonstrating meaningful use of certified EHR technology. Area covered in chart is divided into CMS Incentive Program in which Provider is Registered which are 'Medicare' and 'Medicare/Medicaid'. Orange colored area shows Medicare/Medicaid program. Because hospitals may be eligible and participating in both the Medicaid and Medicare programs, some hospitals have Medicare/Medicaid program. The graph goes down in 2017 because the complete data for the year 2017 is not available yet.

3. Which clinical specialties have large number of providers?

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

df=pd.read_csv("file_clean.csv")

plot1=df.Specialty.value_counts()

new_plot1=plot1.sort_values(ascending=False)[:10]

#Using tuple to store values in explode

explode = (0.1, 0, 0, 0,0,0,0,0,0,0)

new_plot1.plot.pie(figsize=(10, 10), explode=explode, title='Top 10 clinical specialities of providers')

plt.ylabel(' ')
```

Screenshot:



(Highlights from Python script – **Pandas Data Frame**, **Tuple**, Pie chart, different colors, Matplotlib)

Each health care provider has a clinical Specialty assigned to them. It is important to find out which Specialties have large number of providers. Above pie chart states the same thing, it provides top 10 clinical specialties of health care providers. Family Medicine is a topmost clinical specialty. 'Internal Medicine', 'Cardiovascular Disease', 'Obstetrics & Gynecology' and 'Surgery' are some other specialties which have considerable number of providers. This shows that providers for critical diseases like Cardiovascular Disease, Gynecology and Surgery are more.

Python Code: Project 2

#Data Cleaning 1

```
import pandas as pd

new_lables=['National_Provider_Identifier','CMS_Certified_Number', 'Provider_Type',

'Business_State_Territory', 'ZIP', 'Specialty', 'Hospital_Type', 'Program_Type',

'Program_Year', 'Provider_Stage_Number', 'Payment_Year', 'Attestation_Month',

'Attestation_Year', 'MU_Definition_2014', 'Stage_2_Scheduled_2014',

'EHR_Certification_Number', 'EHR_Product_CHP_Id', 'Vendor_Name', 'EHR_Product_Name',

'EHR_Product_Version', 'Product_Classification', 'Product_Setting',

'Product_Certification_Edition_Yr']

df=pd.read_csv('MU_REPORT.csv', header=0,names= new_lables)

print(df.head())
```

#**Data cleaning 2**

```
import pandas as pd

df=pd.read_csv("MU_REPORT.csv")

#Using 'String' inbuilt split function to split 'EHR_Product_CHP_Id'

df[['EHR_Product_Chip', 'EHR_Product_ID']] = df['EHR_Product_CHP_Id'].str.split('-',

expand=True)

print(df.head())
```

#**Data Cleaning 3**

```
import pandas as pd
```

```
df=pd.read_csv("MU_REPORT.csv")

new_df=df.dropna(subset=['Specialty'], how='any')

#Using 'file' attribute to save the dataset in new CSV file

new_df.to_csv("file_clean.csv",index=False)
```

#**Statistical Averages**

```
import pandas as pd

df=pd.read_csv('MU_REPORT.csv')

#summary statistics

stat=df.EHR_Product_Name.value_counts()

print('Min:')

print(stat.min())

print('Max:')

print(stat.max())

print('Mean:')

print(stat.mean())

print('Standard Deviation:')

print(stat.std())

print(stat.describe())
```

#**Visualization 1**

```
import pandas as pd

import matplotlib.pyplot as plt

df=pd.read_csv("file_clean.csv")

vendors=df.groupby('Vendor_Name').Provider_Stage_Number.value_counts().unstack()

new_vendor=vendors.sort_values(by='Stage 1', ascending=False)[:10]

print(new_vendor)

new_vendor.plot(kind='barh', stacked=True, figsize=[16,10], colormap='winter',

title='Number of providers by stage of meaningful use')

plt.show()
```

#**Visualization 2**

```
import pandas as pd

import matplotlib.pyplot as plt

#Pandas Data Frame has been used here

df=pd.read_csv("file_clean.csv")

#User defined function has been used

def prog_year(df):

    df1=df.groupby('Program_Year').Program_Type.value_counts().unstack()
```

```
    df1.plot(kind='area', figsize=[16,10], stacked=True, title='Yearly analysis of number of providers registered in incentive program')
    plt.ylabel('Number of Providers Registered')
    plt.xlabel('Program Year')
prog_year (df)
```

#**Visualization 3**

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("file_clean.csv")
plot1=df.Specialty.value_counts()
new_plot1=plot1.sort_values(ascending=False)[:10]
#Using tuple to store values in explode
explode = (0.1, 0, 0, 0,0,0,0,0,0,0)
new_plot1.plot.pie(figsize=(10, 10), explode=explode, title='Top 10 clinical specialities of providers')
plt.ylabel(' ')
```

References

CMS Medicare and Medicaid EHR Incentive Program: Electronic Health Record. 21 10 2015. Web technical memo. 11 11 2017.

data, Medicare EHR Incentive Program. *EHR Products Used for Meaningful Use Attestation*. 01 11 2017. Web Article. 11 11 2017. <https://dashboard.healthit.gov/datadashboard/documentation/ehr-products-mu-attestation-data-documentation.php>.

Technology, Office of the National Coordinator for Health Information. *CMS Medicare and Medicaid EHR Incentive Program, Electronic Health Record Products Used for Attestation*. 04 06 2017. Web dataset. 10 11 2017. <https://www.healthdata.gov/dataset/cms-medicare-and-medicaid-ehr-incentive-program-electronic-health-record-products-used>.