Theses project - python

October 21, 2021

[1]:

```python
#import library
import pandas as pd
import numpy as np
import missingno as msno
import gender_guesser.detector as gender
gen = gender.Detector()
import plotly.graph_objects as go
import plotly.express as px
```

[2]:

```python
#Load the csv file
df = pd.read_csv("theses_v2.csv", low_memory=False)
df.head(3)
```

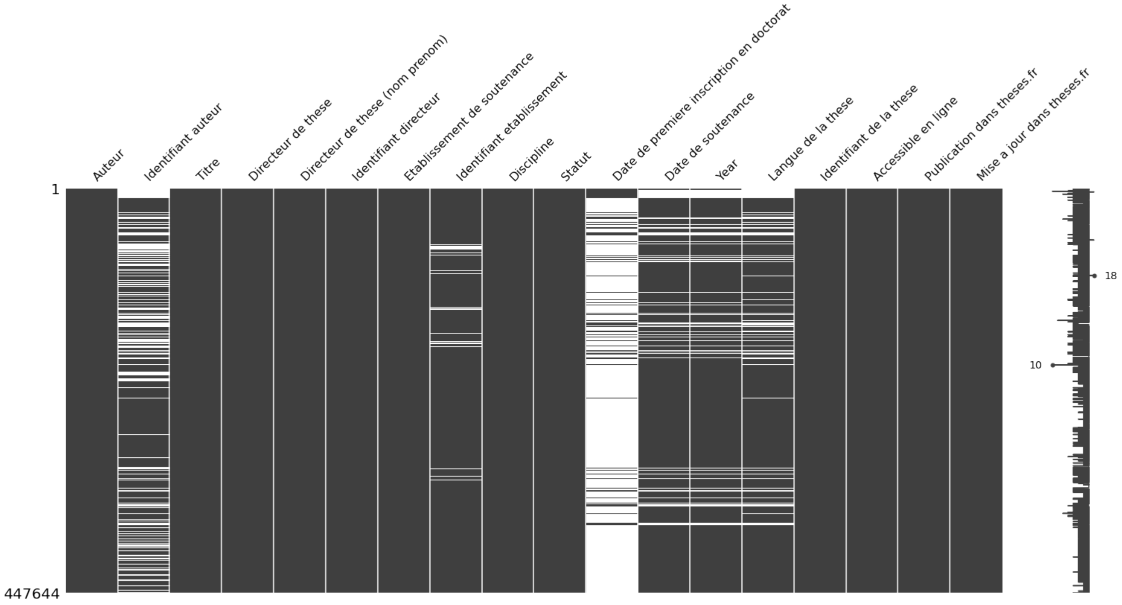
[2]:

```
                Auteur Identifiant auteur  \
0       Saeed Al marri                NaN
1    Andrea Ramazzotti          174423705
2  OLIVIER BODENREIDER                NaN

                                               Titre  \
0  Le credit documentaire et l'onopposabilite des…
1  Application de la PGD a la resolution de probl…
2  Conception d'un outil informatique d'etude des…

                           Directeur de these  \
0                         Philippe Delebecque
1  Jean-Claude Grandidier,Marianne Beringhier
2                             Francois Kohler

              Directeur de these (nom prenom) Identifiant directeur  \
0                         Delebecque Philippe              29561248
1  Grandidier Jean-Claude,Beringhier Marianne           715,441,511
2                             Kohler Francois              57030758

                         Etablissement de soutenance  \
0                                            Paris 1
1  Chasseneuil-du-Poitou, Ecole nationale superie…
2                                            Nancy 1
```

```
  Identifiant etablissement  \
0                  27361802
1                  28024400
2                       NaN

                                          Discipline    Statut  \
0                                        Driot prive   enCours
1  Mecanique des solides, des materiaux, des stru…   enCours
2                                           Medecine  soutenue

  Date de premiere inscription en doctorat Date de soutenance    Year  \
0                                 30-09-11                NaN     NaN
1                                 01-10-12                NaN     NaN
2                                      NaN           01-01-93  1993.0

  Langue de la these Identifiant de la these Accessible en ligne  \
0                NaN                  s69480                 non
1                NaN                  s98826                 non
2                 fr            1993NAN19006                 non

  Publication dans theses.fr Mise a jour dans theses.fr
0                   26-01-12                   26-01-12
1                   22-11-13                   22-11-13
2                   24-05-13                   17-11-12
```

[3]:
```
msno.matrix(df)
```

[3]:
```
<AxesSubplot:>
```



```
[4]: #Chose df from 2010
     df.dropna(subset=['Date de soutenance'], inplace=True)
     df['Date de soutenance'] = pd.DatetimeIndex(df['Date de soutenance'])
     df = df[df['Date de soutenance'].apply(lambda x: np.logical_and(x.year > 2009,␣
      ↪np.logical_or(x.day != 1, x.month != 1)))]
```

```
[5]: df.head(4)
```

```
[5]:                           Auteur Identifiant auteur  \
     8   Jennifer Guiraud (McKELLIPS)                NaN
     9     Nathalie Warcholak (David)                NaN
     10  Scheherazade Pinilla canadas                NaN
     15                Elodie Demaret                NaN

                                                     Titre      Directeur de these  \
     8   L'autobiographie sans frontieres : espace et d…  Anne-Emmanuelle Berger
     9                Interoperabilite et droits du marche.     Jean-Pierre Clavier
     10  Les cites reapparaissantes: L'heroisme du gran…        Patrice Vermeren
     15  La mediation comme facteur de maitrise intelle…       Emile-Henri Riard

        Directeur de these (nom prenom) Identifiant directeur  \
     8           Berger Anne-Emmanuelle              32574088
     9              Clavier Jean-Pierre              35557060
     10                Vermeren Patrice              28251873
     15               Riard Emile-Henri             137391919

        Etablissement de soutenance Identifiant etablissement  \
     8                      Paris 8                  26403552
     9                       Nantes                  26403447
     10                     Paris 8                  26403552
     15                      Amiens                  26403714

                                                Discipline   Statut  \
     8                                     Etudes de genre  enCours
     9                                         Droit prive  enCours
     10  Philosophie (metaphysique, epistemologie, esth…  enCours
     15                                        Psychologie  enCours

        Date de premiere inscription en doctorat Date de soutenance    Year  \
     8                                  01-11-03         2013-10-01  2013.0
     9                                  01-12-02         2011-06-24  2011.0
     10                                 01-03-03         2010-11-26  2010.0
     15                                 01-11-03         2011-06-10  2011.0

        Langue de la these Identifiant de la these Accessible en ligne  \
     8                 NaN                 s11354                 non
     9                 NaN                  s9544                 non
     10                NaN                 s11451                 non
     15                NaN                  s9649                 non

        Publication dans theses.fr Mise a jour dans theses.fr
     8                    26-09-11                   04-04-16
     9                    26-09-11                   05-04-16
     10                   26-09-11                   02-04-12
     15                   26-09-11                   06-02-12
```

[6]:

```
#Create col "Month" & "Year"
df['Month'] = df['Date de soutenance'].apply(lambda x: x.month)
df['Year'] = df['Date de soutenance'].apply(lambda x: x.year)
df.head(4)
```

[6]:

```
                          Auteur Identifiant auteur  \
8   Jennifer Guiraud (McKELLIPS)                NaN
9     Nathalie Warcholak (David)                NaN
10  Scheherazade Pinilla canadas                NaN
15                Elodie Demaret                NaN

                                                Titre      Directeur de these  \
8   L'autobiographie sans frontieres : espace et d…  Anne-Emmanuelle Berger
9               Interoperabilite et droits du marche.     Jean-Pierre Clavier
10  Les cites reapparaissantes: L'heroisme du gran…        Patrice Vermeren
15  La mediation comme facteur de maitrise intelle…       Emile-Henri Riard

   Directeur de these (nom prenom) Identifiant directeur  \
8          Berger Anne-Emmanuelle              32574088
9             Clavier Jean-Pierre              35557060
10               Vermeren Patrice              28251873
15              Riard Emile-Henri             137391919

   Etablissement de soutenance Identifiant etablissement  \
8                      Paris 8                  26403552
9                       Nantes                  26403447
10                     Paris 8                  26403552
15                      Amiens                  26403714

                                           Discipline   Statut  \
8                                     Etudes de genre  enCours
9                                         Droit prive  enCours
10  Philosophie (metaphysique, epistemologie, esth…  enCours
15                                        Psychologie  enCours

   Date de premiere inscription en doctorat Date de soutenance  Year  \
8                                   01-11-03         2013-10-01  2013
9                                   01-12-02         2011-06-24  2011
10                                  01-03-03         2010-11-26  2010
15                                  01-11-03         2011-06-10  2011

   Langue de la these Identifiant de la these Accessible en ligne  \
8                 NaN                  s11354                 non
9                 NaN                   s9544                 non
10                NaN                  s11451                 non
15                NaN                   s9649                 non

   Publication dans theses.fr Mise a jour dans theses.fr  Month
8                    26-09-11                   04-04-16     10
9                    26-09-11                   05-04-16      6
10                   26-09-11                   02-04-12     11
15                   26-09-11                   06-02-12      6
```

[7]:
```
years = df.groupby('Year').count().reset_index().reindex(['Year', 'Titre'],␣
 ↪axis=1).set_index('Year')
years.head(4)
```

[7]:
```
      Titre
Year
2010   4326
2011   7505
2012   9587
2013  10631
```

[8]:
```
#Create df_months from df and calculate the percentage
df_months = df.groupby(['Year', 'Month']).count().reset_index().
 ↪reindex(['Year', 'Month', 'Titre'], axis=1)
df_months['nb_Year'] = df_months['Year'].apply(lambda x: years.loc[x])
df_months['Percentage'] = df_months['Titre'] / df_months['nb_Year'] * 100
df_months['Time'] = pd.to_datetime(df_months[['Year', 'Month']].assign(day=1))
df_months.head(4)
```

[8]:
```
   Year  Month  Titre  nb_Year  Percentage       Time
0  2010      1    268     4326    6.195099 2010-01-01
1  2010      2    176     4326    4.068423 2010-02-01
2  2010      3    287     4326    6.634304 2010-03-01
3  2010      4    205     4326    4.738789 2010-04-01
```

[9]:
```
#Calculate the mean value
df_test=pd.DataFrame(df_months.groupby(['Month'])['Percentage'].mean())
df_test
```

```
[9]:        Percentage
     Month
     1        7.832324
     2        5.455898
     3        6.341093
     4        5.071927
     5        6.620925
     6        9.864329
     7        4.563513
     8        3.515359
     9       10.941472
     10       9.794319
     11      14.314484
     12      15.684357
```

```
[10]: #Calculate the std value
      df_test2 = pd.DataFrame(df_months.groupby(['Month'])['Percentage'].std())
      df_test2
```

```
[10]:       Percentage
      Month
      1        6.391824
      2        3.050201
      3        0.671306
      4        0.530827
      5        1.708365
      6        1.683027
      7        0.752937
      8        0.889911
      9        3.083066
      10       1.802613
      11       3.889338
      12       4.522215
```

```
[11]: df_test2.rename(columns={"Percentage":"sd"},inplace=True)
      df_test2
```

```
[11]:             sd
      Month
      1      6.391824
      2      3.050201
      3      0.671306
      4      0.530827
      5      1.708365
      6      1.683027
      7      0.752937
      8      0.889911
```

```
9      3.083066
10     1.802613
11     3.889338
12     4.522215
```

[12]:

```
df_thesis = pd.merge(df_test,df_test2,on="Month")
df_thesis.reset_index(inplace=True)
df_thesis
```

[12]:

```
    Month  Percentage        sd
0       1    7.832324  6.391824
1       2    5.455898  3.050201
2       3    6.341093  0.671306
3       4    5.071927  0.530827
4       5    6.620925  1.708365
5       6    9.864329  1.683027
6       7    4.563513  0.752937
7       8    3.515359  0.889911
8       9   10.941472  3.083066
9      10    9.794319  1.802613
10     11   14.314484  3.889338
11     12   15.684357  4.522215
```

[13]:

```
#Covert month to name
import datetime
def monthnum_toname(x):
    month = datetime.date(1900, x, 1).strftime('%B')
    return month
```

[14]:

```
df_thesis["month_name"]=df_thesis["Month"].apply(lambda x:monthnum_toname(x))
df_thesis["month_name"]
```

[14]:

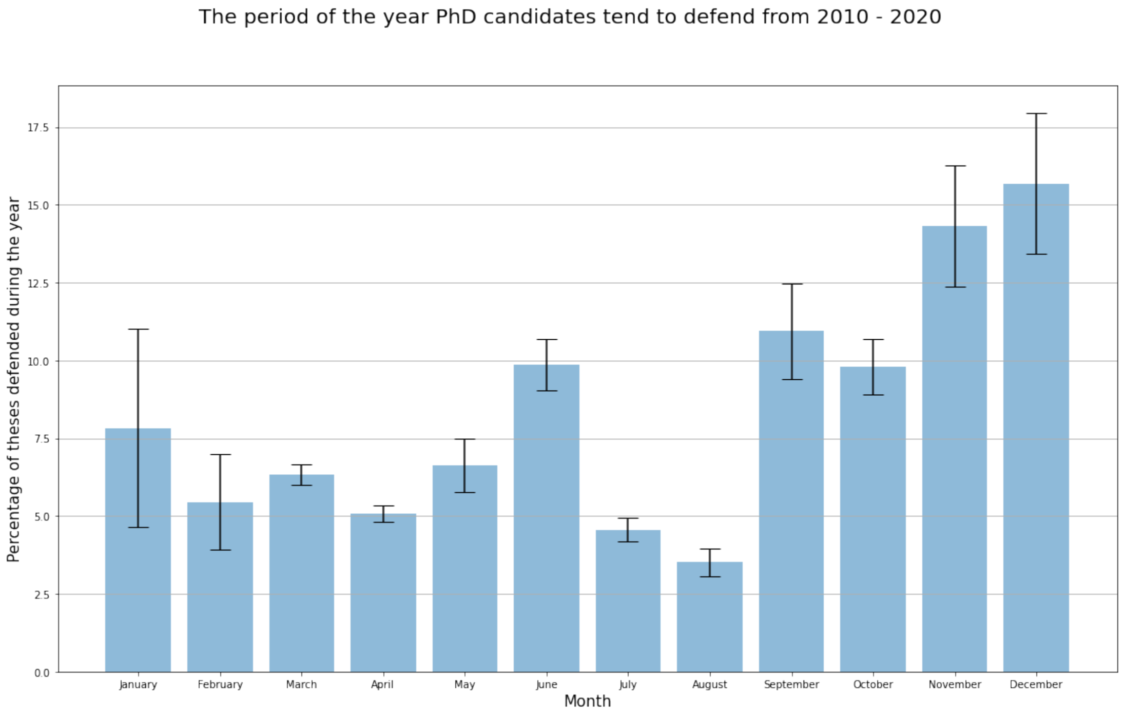
```
0       January
1      February
2         March
3         April
4           May
5          June
6          July
7        August
8     September
9       October
10     November
11     December
Name: month_name, dtype: object
```

[15]:

```
df_thesis.to_csv("df_thesis.csv",)
```

[16]:
```python
import matplotlib.pyplot as plt
```

[17]:
```python
#Plot
fig, ax = plt.subplots()
ax.bar(df_thesis["month_name"], df_thesis["Percentage"], yerr=df_thesis["sd"]/
 ↪2, align='center', alpha=0.5, ecolor='black', capsize=10)
ax.set_ylabel('Percentage of theses defended during the year',fontsize=15)
ax.set_xlabel('Month',fontsize=15)
ax.set_xticks(df_thesis["month_name"])
ax.set_xticklabels(df_thesis["month_name"])
fig.suptitle('The period of the year PhD candidates tend to defend from 2010 -␣
 ↪2020', fontsize=20)
ax.yaxis.grid(True)
fig.set_size_inches(18.5, 10.5)
```



[18]:
```python
gender = df[["Auteur","Date de soutenance"]]
```

[19]:
```python
#Split the column Auteur in order to create "First name" column
gender['First_name']=gender.loc[:, ('Auteur')].str.split(expand=True)[[0]]
gender.head(4)
```

```
<ipython-input-19-39e05f11f8ea>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender['First_name']=gender.loc[:, ('Auteur')].str.split(expand=True)[[0]]
```

[19]:

```
                           Auteur Date de soutenance   First_name
8   Jennifer Guiraud (McKELLIPS)         2013-10-01     Jennifer
9     Nathalie Warcholak (David)         2011-06-24     Nathalie
10  Scheherazade Pinilla canadas         2010-11-26  Scheherazade
15                Elodie Demaret         2011-06-10       Elodie
```

[20]:

```python
#Create function to find the gender by using gender_guesser.detector library
def get_gender(x,gen):
    return gen.get_gender(u"{}".format(x))
```

[21]:

```python
#Apply function get_gender
gender["Gender"] = gender['First_name'].apply(lambda x:get_gender(x,gen))
gender.head(4)
```

```
<ipython-input-21-ce6313b0b0ce>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender["Gender"] = gender['First_name'].apply(lambda x:get_gender(x,gen))
```

[21]:

```
                           Auteur Date de soutenance   First_name   Gender
8   Jennifer Guiraud (McKELLIPS)         2013-10-01     Jennifer   female
9     Nathalie Warcholak (David)         2011-06-24     Nathalie   female
10  Scheherazade Pinilla canadas         2010-11-26  Scheherazade  unknown
15                Elodie Demaret         2011-06-10       Elodie   female
```

[22]:

```python
#Create col "Year" in gender
gender['Year'] = pd.DatetimeIndex(gender["Date de soutenance"]).year
```

```
<ipython-input-22-9adc5d2e4d46>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender['Year'] = pd.DatetimeIndex(gender["Date de soutenance"]).year
```

[23]:

```python
#Cleaning data
gender.dropna(subset=['Year'],how='all',inplace=True)
gender.isnull().sum()
gender.head(4)
```

```
<ipython-input-23-8a78a2a48150>:2: SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  gender.dropna(subset=['Year'],how='all',inplace=True)
```

[23]:

```
                            Auteur Date de soutenance   First_name   Gender  \
8   Jennifer Guiraud (McKELLIPS)           2013-10-01     Jennifer   female
9     Nathalie Warcholak (David)           2011-06-24     Nathalie   female
10  Scheherazade Pinilla canadas           2010-11-26  Scheherazade  unknown
15                Elodie Demaret           2011-06-10       Elodie   female

    Year
8   2013
9   2011
10  2010
15  2011
```

[24]:

```
gender_df = gender.groupby(['Gender','Year']).count().reset_index()
gender_df
```

[24]:

```
     Gender  Year  Auteur  Date de soutenance  First_name
0      andy  2010      87                  87          87
1      andy  2011     155                 155         155
2      andy  2012     217                 217         217
3      andy  2013     242                 242         242
4      andy  2014     279                 279         279
..      …     …     …                 …           …
61  unknown  2016    2327                2327        2327
62  unknown  2017    2582                2582        2582
63  unknown  2018    2400                2400        2400
64  unknown  2019    2080                2080        2079
65  unknown  2020     214                 214         214

[66 rows x 5 columns]
```

[25]:

```
fig = px.area(gender_df,title='The evolution of gender among PhD candidates␣
 ↪over the past decades', x="Year",␣
 ↪y="Auteur",color="Gender",line_group="Gender")
fig.show()
```

[ ]:

[ ]: