In [1]:

```python
# Import required libraries
import pandas as pd
import plotly.express as px
import gender_guesser.detector as gender
```

In [2]:

```python
# import datset
df = pd.read_csv('../data/theses_v2.csv')
df.head()
```

```
C:\Users\Administrator\miniconda3\lib\site-packages\IPython\core\interactiveshell.p
y:3441: DtypeWarning: Columns (10) have mixed types.Specify dtype option on import o
r set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[2]:

	Auteur	Identifiant auteur	Titre	Directeur de these	Directeur de these (nom prenom)	Identifiant directeur	Etal s
0	Saeed Al marri	NaN	Le credit documentaire et l'onopposabilite des...	Philippe Delebecque	Delebecque Philippe	29561248	
1	Andrea Ramazzotti	174423705	Application de la PGD a la resolution de probl...	Jean-Claude Grandidier,Marianne Beringhier	Grandidier Jean-Claude,Beringhier Marianne	715,441,511	Cl
2	OLIVIER BODENREIDER	NaN	Conception d'un outil informatique d'etude des...	Francois Kohler	Kohler Francois	57030758	
3	Emmanuel Porte	NaN	Socio-histoire des politiques publiques en mat...	Gilles Pollet	Pollet Gilles	na	
4	Arthur Devriendt	NaN	LES TECHNOLOGIES DE L'INFORMATION ET DE LA COM...	Gabriel Dupuy	Dupuy Gabriel	na	

In [3]:

```python
# function to detect gender
d = gender.Detector()

def get_gender_by_name(x,d):
    return d.get_gender(u"{}".format(x))
```

In [4]:

```python
# function to set text in title case
def title_case(x):
    if x is None:
        pass
    else:
        return x.title()
```

Authors

In [5]:

```
# select authors
df_gender = df[["Auteur","Date de soutenance"]]
df_gender.head()
```

Out[5]:

	Auteur	Date de soutenance
0	Saeed Al marri	NaN
1	Andrea Ramazzotti	NaN
2	OLIVIER BODENREIDER	01-01-93
3	Emmanuel Porte	NaN
4	Arthur Devriendt	NaN

In [6]:

```
# get the first name of the author and set it to title case
df_gender['first_name']=df_gender.Auteur.str.split(expand=True)[[0]]
df_gender["first_name"]=df_gender["first_name"].apply(lambda x: title_case(x))
df_gender.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/3831665226.py:2: SettingWithCop
yWarning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df_gender['first_name']=df_gender.Auteur.str.split(expand=True)[[0]]
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/3831665226.py:3: SettingWithCop
yWarning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df_gender["first_name"]=df_gender["first_name"].apply(lambda x: title_case(x))
```

Out[6]:

	Auteur	Date de soutenance	first_name
0	Saeed Al marri	NaN	Saeed
1	Andrea Ramazzotti	NaN	Andrea
2	OLIVIER BODENREIDER	01-01-93	Olivier
3	Emmanuel Porte	NaN	Emmanuel
4	Arthur Devriendt	NaN	Arthur

In [7]:

```
# get the gender of each author
df_gender["gender"] = df_gender['first_name'].apply(lambda x:get_gender_by_name(x,d)
df_gender.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/3239357764.py:2: SettingWithCop
yWarning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df_gender["gender"] = df_gender['first_name'].apply(lambda x:get_gender_by_name(x,
d))
```

Out[7]:

	Auteur	Date de soutenance	first_name	gender
0	Saeed Al marri	NaN	Saeed	male
1	Andrea Ramazzotti	NaN	Andrea	female
2	OLIVIER BODENREIDER	01-01-93	Olivier	male
3	Emmanuel Porte	NaN	Emmanuel	male
4	Arthur Devriendt	NaN	Arthur	male

In [8]:

```
# get the year of defence and drop na values in year
df_gender['year'] = pd.DatetimeIndex(df_gender["Date de soutenance"]).year
df_gender.dropna(subset=['year'],how='all',inplace=True)
df_gender['year'] = df_gender['year'].astype(int) # set year as integer
df_gender.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/1668690746.py:2: SettingWithCop
yWarning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df_gender['year'] = pd.DatetimeIndex(df_gender["Date de soutenance"]).year
C:\Users\Administrator\miniconda3\lib\site-packages\pandas\util\_decorators.py:311:
 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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  return func(*args, **kwargs)
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/1668690746.py:4: SettingWithCop
yWarning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df_gender['year'] = df_gender['year'].astype(int) # set year as integer
```

Out[8]:

	Auteur	Date de soutenance	first_name	gender	year
2	OLIVIER BODENREIDER	01-01-93	Olivier	male	1993
5	Elmantsr Briak	24-11-08	Elmantsr	unknown	2008
6	Jae-hyun Park	01-07-05	Jae-Hyun	male	2005
7	Laurent david Benoiton	08-12-09	Laurent	male	2009
8	Jennifer Guiraud (McKELLIPS)	10-01-13	Jennifer	female	2013

In [9]:

```
# group data by gender and year to get frequency
df_gender_count = df_gender.groupby(['gender','year']).count().reset_index()
df_gender_count.head()
```

Out[9]:

	gender	year	Auteur	Date de soutenance	first_name
0	andy	1985	47	47	47
1	andy	1986	102	102	102
2	andy	1987	204	204	204
3	andy	1988	251	251	251
4	andy	1989	294	294	294

In [10]:

```
# select data for years between 1998 and 2019
df_gender_count = df_gender_count.query('year > 1987 & year < 2020')
df_gender_count.head()
```

Out[10]:

	gender	year	Auteur	Date de soutenance	first_name
3	andy	1988	251	251	251
4	andy	1989	294	294	294
5	andy	1990	278	278	278
6	andy	1991	256	256	256
7	andy	1992	267	267	267

In [11]:

```
# select and rename required columns
df_gender_count = df_gender_count[['gender', 'year', 'Auteur']]
df_gender_count.rename(columns={'Auteur':'Number of authors (Genderwise)'},inplace=T
df_gender_count.head()
```

Out[11]:

	gender	year	Number of authors (Genderwise)
3	andy	1988	251
4	andy	1989	294
5	andy	1990	278
6	andy	1991	256
7	andy	1992	267

In [12]:

```
# get tot
df_date_count = df_gender.groupby(['year']).count().reset_index()
df_date_count = df_date_count.query('year >= 1988 & year < 2020')
df_date_count.head()
```

Out[12]:

	year	Auteur	Date de soutenance	first_name	gender
11	1988	11045	11045	11045	11045
12	1989	11102	11102	11102	11102
13	1990	11011	11011	11011	11011
14	1991	10831	10831	10831	10831
15	1992	12065	12065	12065	12065

In [13]:

```
df_date_count = df_date_count[['year', 'Auteur']]
df_date_count.rename(columns={'Auteur':'Number of authors (Yearwise)'},inplace=True)
df_date_count.head()
```

Out[13]:

	year	Number of authors (Yearwise)
11	1988	11045
12	1989	11102
13	1990	11011
14	1991	10831
15	1992	12065

In [14]:

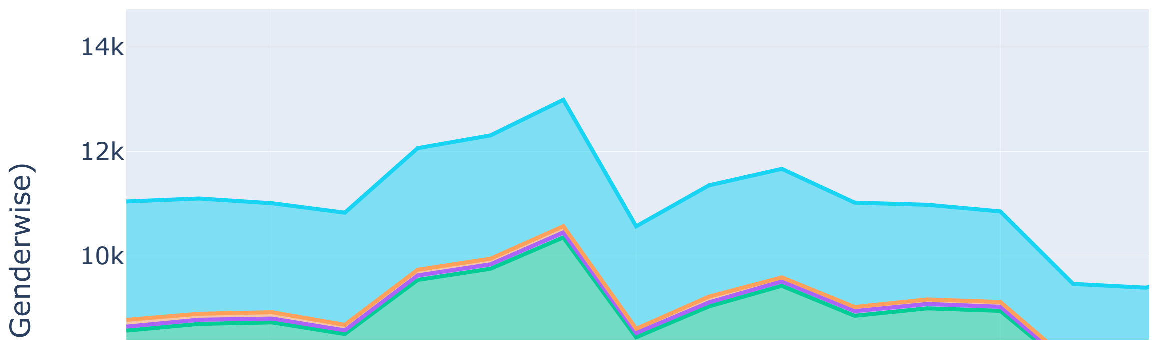
```
df_gender_prec = pd.merge(df_gender_count, df_date_count, on='year', how = 'outer')
df_gender_prec['Percentage of Authors'] = df_gender_prec['Number of authors (Genderw
df_gender_prec.head()
```

Out[14]:

	gender	year	Number of authors (Genderwise)	Number of authors (Yearwise)	Percentage of Authors
0	andy	1988	251	11045	2.272522
1	female	1988	3080	11045	27.885921
2	male	1988	5244	11045	47.478497
3	mostly_female	1988	76	11045	0.688094
4	mostly_male	1988	129	11045	1.167949

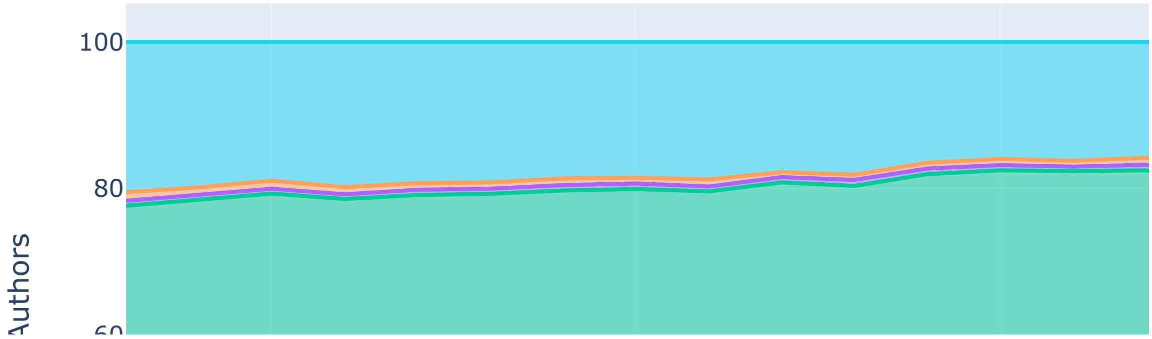
In [15]:

```
fig = px.area(df_gender_count, x="year", y="Number of authors (Genderwise)", color="

fig.show()
```



In [16]:

```
fig = px.area(df_gender_prec, x="year", y="Percentage of Authors",color="gender")

fig.show()
```



Supervisors

In [17]:

```
# select supervisors
df_gender_2 = df[["Directeur de these","Date de soutenance"]]
df_gender_2['first_name'] = df_gender_2["Directeur de these"].str.split(expand=True)
df_gender_2.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/1201103762.py:3: SettingWithCop
yWarning:


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/u
```

```
ser_guide/indexing.html#returning-a-view-versus-a-copy
```

Out[17]:

	Directeur de these	Date de soutenance	first_name
0	Philippe Delebecque	NaN	Philippe
1	Jean-Claude Grandidier,Marianne Beringhier	NaN	Jean-Claude
2	Francois Kohler	01-01-93	Francois
3	Gilles Pollet	NaN	Gilles
4	Gabriel Dupuy	NaN	Gabriel

In [18]:

```
# get the first name of the supervisor and set it to title case
df_gender_2["first_name"]=df_gender_2["first_name"].apply(lambda x: title_case(str(x
df_gender_2.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/1787802300.py:2: SettingWithCop
yWarning:


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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
```

Out[18]:

	Directeur de these	Date de soutenance	first_name
0	Philippe Delebecque	NaN	Philippe
1	Jean-Claude Grandidier,Marianne Beringhier	NaN	Jean-Claude
2	Francois Kohler	01-01-93	Francois
3	Gilles Pollet	NaN	Gilles
4	Gabriel Dupuy	NaN	Gabriel

In [19]:

```
# get the gender of each supervisor
df_gender_2["gender"] = df_gender_2['first_name'].apply(lambda x:get_gender_by_name(
df_gender_2.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/3494067728.py:2: SettingWithCop
yWarning:


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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
```

Out[19]:

	Directeur de these	Date de soutenance	first_name	gender
0	Philippe Delebecque	NaN	Philippe	male
1	Jean-Claude Grandidier,Marianne Beringhier	NaN	Jean-Claude	male
2	Francois Kohler	01-01-93	Francois	unknown

	Directeur de these	Date de soutenance	first_name	gender
3	Gilles Pollet	NaN	Gilles	male
4	Gabriel Dupuy	NaN	Gabriel	male

In [20]:

```
# get the year of defence and drop na values in year
df_gender_2['year'] = pd.DatetimeIndex(df_gender_2["Date de soutenance"]).year
df_gender_2.dropna(subset=['year'],how='all',inplace=True)
df_gender_2['year'] = df_gender_2['year'].astype(int)
df_gender_2.head()
```

```
C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/324897322.py:2: SettingWithCopy
Warning:


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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy

C:\Users\Administrator\miniconda3\lib\site-packages\pandas\util\_decorators.py:311:
 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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy

C:\Users\ADMINI~1\AppData\Local\Temp/ipykernel_13500/324897322.py:4: SettingWithCopy
Warning:


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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
```

Out[20]:

	Directeur de these	Date de soutenance	first_name	gender	year
2	Francois Kohler	01-01-93	Francois	unknown	1993
5	Edmond Jouve	24-11-08	Edmond	male	2008
6	Pierre Comte	01-07-05	Pierre	male	2005
7	Laurent Sermet	08-12-09	Laurent	male	2009
8	Anne-Emmanuelle Berger	10-01-13	Anne-Emmanuelle	unknown	2013

In [21]:

```
# group data by gender and year to get frequency
df_gender_count_2 = df_gender_2.groupby(['gender','year']).count().reset_index()
df_gender_count_2 = df_gender_count_2.query('year >= 1988 & year < 2020')
df_gender_count_2 = df_gender_count_2[['gender', 'year', 'Directeur de these']]
df_gender_count_2.rename(columns={'Directeur de these':'Number of supervisors (Gende
df_gender_count_2.head()
```

Out[21]:

	gender	year	Number of supervisors (Genderwise)

	gender	year	Number of supervisors (Genderwise)
3	andy	1988	96
4	andy	1989	115
5	andy	1990	126
6	andy	1991	135
7	andy	1992	151

In [22]:

```
# select data for years between 1998 and 2019
df_date_count_2 = df_gender_2.groupby(['year']).count().reset_index()
df_date_count_2 = df_date_count_2.query('year >= 1988 & year < 2020')
df_date_count_2.head()
```

Out[22]:

	year	Directeur de these	Date de soutenance	first_name	gender
11	1988	11045	11045	11045	11045
12	1989	11101	11102	11102	11102
13	1990	11011	11011	11011	11011
14	1991	10831	10831	10831	10831
15	1992	12063	12065	12065	12065

In [23]:

```
# select and rename required columns
df_date_count_2 = df_date_count_2[['year', 'Directeur de these']]
df_date_count_2.rename(columns={'Directeur de these':'Number of supervisors (Yearwis
df_date_count_2.head()
```

Out[23]:

	year	Number of supervisors (Yearwise)
11	1988	11045
12	1989	11101
13	1990	11011
14	1991	10831
15	1992	12063

In [24]:

```
# merge to get percentage
df_gender_prec_2 = pd.merge(df_gender_count_2, df_date_count_2, on='year', how = 'ou
df_gender_prec_2['Percentage of Supervisors'] = df_gender_prec_2['Number of supervis
df_gender_prec_2.head()
```

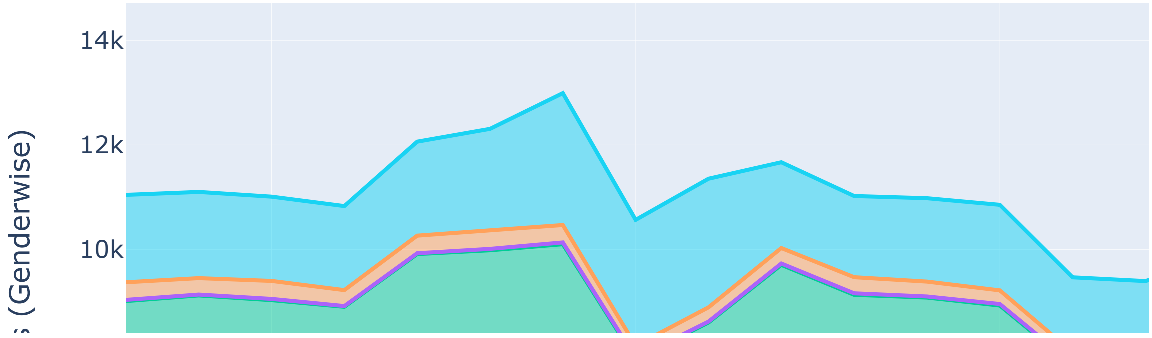
Out[24]:

	gender	year	Number of supervisors (Genderwise)	Number of supervisors (Yearwise)	Percentage of Supervisors
0	andy	1988	96	11045	0.869172
1	female	1988	664	11045	6.011770
2	male	1988	8253	11045	74.721593
3	mostly_female	1988	23	11045	0.208239

	gender	year	Number of supervisors (Genderwise)	Number of supervisors (Yearwise)	Percentage of Supervisors
4	mostly_male	1988	337	11045	3.051154

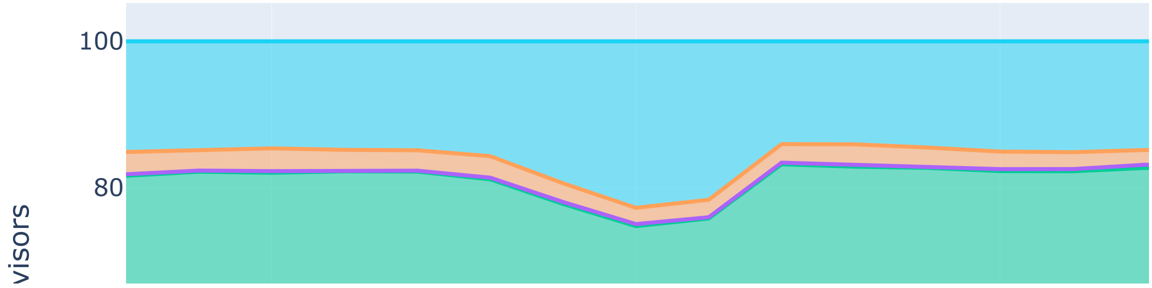
In [25]:

```
# plot for sum
fig = px.area(df_gender_count_2, x="year", y="Number of supervisors (Genderwise)", c

fig.show()
```



In [26]:

```
# plot for percentage

fig = px.area(df_gender_prec_2, x="year", y="Percentage of Supervisors",color="gende

fig.show()
```



```
In [ ]:
```