



Dr. Vishwanath Karad
**MIT WORLD PEACE
UNIVERSITY** | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

School of Computer Engineering & Technology
T.Y. B.Tech. (CSE) Trimester 9 (2019-2020)

BDA Mini-Project Report

--- Summary ---

Faculty: Prof Sheetal Girase

Submission Date: 19th June, 2021

Group - 08

Group members

PE25 Ajinkya Karnik - 1032180678

PD05 Rushikesh Kothawade - 1032180122

PD Shreyash Shinde - 1032170009

Topic

Immigrants data set preprocessing, cleaning and making appropriate data visualizations with D3.js

This report is a short summary of the project and will primarily include a few screenshots from the Notebook and from D3.js viz :

Data Preprocessing and cleaning :

BDA_mini_project_report - x Report.pdf - Google Drive x WhatsApp x Home Page - Select or cre x BDA mini project data proc x bda analysis - Jupyter Note x +

localhost:8889/notebooks/BDA%20mini%20project%20data%20proc.ipynb

jupyter BDA mini project data proc Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run

	0	2017.0	1.0	Clutat Vella	1.0	el Raval	Spain	1109.0	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni
1	2017.0	1.0	Clutat Vella	2.0	el Barri Gòtic	Spain	482.0	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
2	2017.0	1.0	Clutat Vella	3.0	la Barceloneta	Spain	414.0	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
3	2017.0	1.0	Clutat Vella	4.0	Sant Pere, Santa Caterina i la Ribera	Spain	537.0	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
4	2017.0	2.0	Eixample	5.0	el Fort Pienc	Spain	663.0	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
...	
41051	NaN	NaN	NaN	NaN	NaN	NaN	NaN	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
41052	NaN	NaN	NaN	NaN	NaN	NaN	NaN	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
41053	NaN	NaN	NaN	NaN	NaN	NaN	NaN	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
41054	NaN	NaN	NaN	NaN	NaN	NaN	NaN	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	
41055	NaN	NaN	NaN	NaN	NaN	NaN	NaN	/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Pro...	NaN	NaN	NaN	Ni	

41056 rows x 15 columns

Joining from-to datasets [d4 and d5]

```
In [33]: immigrants_1 = pd.merge(d4,d5,how="outer")
```

```
In [34]: immigrants_1
```

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all

BDA_mini_project_report - x Report.pdf - Google Drive x WhatsApp x Home Page - Select or cre x BDA mini project data proc x bda analysis - Jupyter Note x +

localhost:8889/notebooks/BDA%20mini%20project%20data%20proc.ipynb

jupyter BDA mini project data proc Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run

```
In [148]: immigrants_1
```

```
Out[148]:
```

	from	to	weight
400	Abroad	Barcelona	53085
415	Barcelona	Catalonia	35036
430	Catalonia	Barcelona	28779
311	Abroad	Eixample	11473
310	Abroad	Clutat Vella	8586
...
7	Clutat Vella	Ceuta	0
67	Horta-Guinardó	Ceuta	0
274	Ceuta	Les Corts	0
47	Gràcia	Ceuta	0
107	Nou Barris	Ceuta	0

440 rows x 3 columns

```
In [137]: immigrants_1.reset_index(drop=True, inplace=True)
```

Merging [d1+d2] with d3

```
In [125]: immigrants_3 = pd.merge(immigrants_2, d3, on=['District Name', 'Neighborhood Code', 'Year', 'Immigrants', 'Emigrants', 'Dist
```

Handling missing values

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all



BDA mini project data proc Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

440 rows x 3 columns

```
In [137]: immigrants_1.reset_index(drop=True, inplace=True)
```

Merging [d1+d2] with d3

```
In [125]: immigrants_3 = pd.merge(immigrants_2, d3, on=['District Name', 'Neighborhood Code', 'Year', 'Immigrants', 'Emigrants', 'District Name'])
```

Handling missing values

```
In [155]: immigrants_3 = immigrants_3.sort_values(['Number', 'Immigrants', 'Emigrants'], ascending=[0,0,0])
immigrants_3 = immigrants_3.dropna()
immigrants_3
```

Out[155]:

index	Neighborhood Code	Year	District Code	District Name	Nationality	Number	Age	Immigrants	Emigrants	Neighborhood Name	Gender	
8	8	9	2017	2	Example	Spain	1593	0-4	123.0	130.0	la Nova Esquerra de l'Eixample	Male
30	30	31	2017	6	Gràcia	Spain	1415	0-4	88.0	104.0	la Vila de Gràcia	Male
5	5	6	2017	2	Example	Spain	1181	0-4	111.0	95.0	la Sagrada Família	Male
7	7	8	2017	2	Example	Spain	1177	0-4	97.0	63.0	l'Antiga Esquerra de l'Eixample	Male
25	25	26	2017	5	Sarrià-Sant Gervasi	Spain	1146	0-4	141.0	82.0	Sant Gervasi - Galvany	Male
...	
221	221	99	2017	99	No consta	China	0	10-14	0.0	0.0	la Barceloneta	Female
295	295	99	2017	99	No consta	Colombia	0	15-19	0.0	0.0	Sant Pere, Santa Caterina i la Ribera	Male
...	

BDA_mini_project_report.pdf BDA_mini_project_report.pdf immigrants.csv stacked area.html treemap.html Show all

BDA mini project data proc Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Merging [d1+d2] with d3

```
In [125]: immigrants_3 = pd.merge(immigrants_2, d3, on=['District Name', 'Neighborhood Code', 'Year', 'Immigrants', 'Emigrants', 'District Name'])
```

Handling missing values

```
In [155]: immigrants_3 = immigrants_3.sort_values(['Number', 'Immigrants', 'Emigrants'], ascending=[0,0,0])
immigrants_3 = immigrants_3.dropna()
immigrants_3
```

7	7	8	2017	2	Exemple	Spain	1177	0-4	97.0	63.0	l'Antiga Esquerra de l'Eixample	Male
25	25	26	2017	5	Sarrià-Sant Gervasi	Spain	1146	0-4	141.0	82.0	Sant Gervasi - Galvany	Male
...	
221	221	99	2017	99	No consta	China	0	10-14	0.0	0.0	la Barceloneta	Female
295	295	99	2017	99	No consta	Colombia	0	15-19	0.0	0.0	Sant Pere, Santa Caterina i la Ribera	Male
369	369	99	2017	99	No consta	Venezuela	0	20-24	0.0	0.0	el Fort Pienc	Female
591	591	99	2017	99	No consta	France	0	35-39	0.0	0.0	l'Antiga Esquerra de l'Eixample	Male
665	665	99	2017	99	No consta	Peru	0	40-44	0.0	0.0	la Nova Esquerra de l'Eixample	Female

730 rows x 12 columns

```
In [139]: immigrants_1.dropna()
immigrants_1.reset_index()
```

Out[139]:

index	from	to	weight
-------	------	----	--------

BDA_mini_project_report.pdf BDA_mini_project_report.pdf immigrants.csv stacked area.html treemap.html Show all



BDA_mini_project_report - x Report.pdf - Google Drive x WhatsApp x Home Page - Select or cre x BDA mini project data proc x bda analysis - Jupyter Not x +

localhost:8889/notebooks/BDA%20mini%20project%20data%20proc.ipynb

jupyter BDA mini project data proc Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

440 rows x 4 columns

Checking for null values

```
In [156]: immigrants_1.isnull().values.any()
Out[156]: False

In [158]: immigrants_3.isnull().values.any()
Out[158]: False

In [ ]:
```

Final datasets

```
In [140]: immigrants_1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ---
 0   from    440 non-null     object
 1   to      440 non-null     object
 2   weight  440 non-null     int64
dtypes: int64(1), object(2)
memory usage: 10.4+ KB

In [149]: immigrants_1.head()

Out[149]:
```

	from	to	weight
--	------	----	--------

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all

BDA_mini_project_report - x Report.pdf - Google Drive x WhatsApp x Home Page - Select or cre x BDA mini project data proc x bda analysis - Jupyter Not x +

localhost:8889/notebooks/BDA%20mini%20project%20data%20proc.ipynb

jupyter BDA mini project data proc Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Final datasets

```
In [140]: immigrants_1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ---
 0   from    440 non-null     object
 1   to      440 non-null     object
 2   weight  440 non-null     int64
dtypes: int64(1), object(2)
memory usage: 10.4+ KB

In [149]: immigrants_1.head()

Out[149]:
```

	from	to	weight
400	Abroad	Barcelona	53085
415	Barcelona	Catalonia	35036
430	Catalonia	Barcelona	28779
311	Abroad	Example	11473
310	Abroad	Ciutat Vella	8586

```
In [150]: immigrants_1.tail()

Out[150]:
```

	from	to	weight
7	Ciutat Vella	Ceuta	0
67	Horta-Guinardó	Ceuta	0
274	Ceuta	Les Corts	0

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all



BDA mini project data proc Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
In [143]: immigrants_3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 35224 entries, 0 to 35223
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  -
 0   index               35224 non-null  int64
 1   Neighborhood Code    35224 non-null  int64
 2   Year                35224 non-null  int64
 3   District Code        35224 non-null  int64
 4   District Name        35224 non-null  object
 5   Nationality          35224 non-null  object
 6   Number              35224 non-null  int64
 7   Age                 4662 non-null   object
 8   Immigrants           4662 non-null   float64
 9   Emigrants            4662 non-null   float64
10   Neighborhood Name     730 non-null    object
11   Gender               730 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 3.5+ MB
```

```
In [165]: immigrants_3.head(100)
```

```
Out[165]:
```

	index	Neighborhood Code	Year	District Code	District Name	Nationality	Number	Age	Immigrants	Emigrants	Neighborhood Name	Gender
8	8	9	2017	2	Eixample	Spain	1593	0-4	123.0	130.0	la Nova Esquerra de l'Eixample	Male
30	30	31	2017	6	Gràcia	Spain	1415	0-4	88.0	104.0	la Vila de Gràcia	Male
5	5	6	2017	2	Eixample	Spain	1181	0-4	111.0	95.0	la Sagrada Família	Male
7	7	8	2017	2	Eixample	Spain	1177	0-4	97.0	63.0	l'Antiga Esquerra de l'Eixample	Male
25	25	26	2017	5	Sarrià-Sant Gervasi	Spain	1146	0-4	141.0	82.0	Sant Gervasi - Galvany	Male

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all

BDA mini project data proc Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
In [173]: immigrants_3.tail()
```

```
Out[173]:
```

	index	Neighborhood Code	Year	District Code	District Name	Nationality	Number	Age	Immigrants	Emigrants	Neighborhood Name	Gender
221	221	99	2017	99	No consta	China	0	10-14	0.0	0.0	la Barceloneta	Female
295	295	99	2017	99	No consta	Colombia	0	15-19	0.0	0.0	Sant Pere, Santa Caterina i la Ribera	Male
369	369	99	2017	99	No consta	Venezuela	0	20-24	0.0	0.0	el Fort Pienc	Female
591	591	99	2017	99	No consta	France	0	35-39	0.0	0.0	l'Antiga Esquerra de l'Eixample	Male
665	665	99	2017	99	No consta	Peru	0	40-44	0.0	0.0	la Nova Esquerra de l'Eixample	Female

```
In [ ]:
```

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all



BDA mini project data proc Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
In [173]: immigrants_3.tail()
```

Out[173]:

	index	Neighborhood Code	Year	District Code	District Name	Nationality	Number	Age	Immigrants	Emigrants	Neighborhood Name	Gender
221	221	99	2017	99	No consta	China	0	10-14	0.0	0.0	la Barceloneta	Female
295	295	99	2017	99	No consta	Colombia	0	15-19	0.0	0.0	Sant Pere, Santa Caterina i la Ribera	Male
369	369	99	2017	99	No consta	Venezuela	0	20-24	0.0	0.0	el Fort Pienc	Female
591	591	99	2017	99	No consta	France	0	35-39	0.0	0.0	l'Antiga Esquerra de l'Eixample	Male
665	665	99	2017	99	No consta	Peru	0	40-44	0.0	0.0	la Nova Esquerra de l'Eixample	Female

In []:

Saving the processed files

```
In [130]: immigrants_1.to_csv('/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Project/immigrants_processed_1.csv')
```

```
In [134]: immigrants_3.to_csv('/Users/ajinkyakarnik/Desktop/tri9/BDA/Mini Project/immigrants_processed_2.csv')
```

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all

BDA mini project data proc Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
In [66]: d1.reset_index(inplace=True)
```

```
In [67]: d2.reset_index(inplace=True)
```

Merging d1 and d2

```
In [78]: immigrants_2 = pd.merge(d1, d2, on=['District Name', 'Neighborhood Code', 'Year', 'District Code'], how='outer', left_index=True, right_index=True)
```

```
In [79]: immigrants_2
```

Out[79]:

	Neighborhood Code	Year	District Code	District Name	Nationality	Number	Age	Immigrants	Emigrants
0	1	2017	1	Ciutat Vella	Spain	1109	0-4	154.0	108.0
1	2	2017	1	Ciutat Vella	Spain	482	0-4	58.0	33.0
2	3	2017	1	Ciutat Vella	Spain	414	0-4	38.0	37.0
3	4	2017	1	Ciutat Vella	Spain	537	0-4	56.0	55.0
4	5	2017	2	Eixample	Spain	663	0-4	79.0	60.0
...
35219	70	2015	10	Sant Martí	No information	0	NaN	NaN	NaN
35220	71	2015	10	Sant Martí	No information	0	NaN	NaN	NaN
35221	72	2015	10	Sant Martí	No information	0	NaN	NaN	NaN
35222	73	2015	10	Sant Martí	No information	1	NaN	NaN	NaN
35223	99	2015	99	No consta	No information	0	NaN	NaN	NaN

35224 rows x 9 columns

```
In [80]: immigrants_2.head(10)
```

BDA_mini_projec....pdf BDA_mini_projec....pdf immigrants.csv stacked area.html treemap.html Show all

The code and the Jupyter notebook for the same is included in the project file for further reference.

D3 Visualizations :

