In [ ]:
```
from src.main import generate_csv, read_csv, generate_summary, create_histogram, create_scatter_plot
import pandas as pd
import matplotlib.pyplot as plt
import os
```

In [ ]:
```
default_file_folder = 'data'
csv_file_name = 'random_user.csv'
full_path = os.path.join(os.getcwd(),default_file_folder, csv_file_name)
params = {"results" : 500}
```

In [ ]:
```
file_path = generate_csv(file_name=full_path, params=params)
print(f'CSV generated at - {file_path}')
df = read_csv(file_path, engine_type='pandas')
```

```
CSV data successfully saved to /Users/javidan/Developer/Data Engineering/PandasDescriptiveStatistics/data/random_user.csv
CSV generated at - /Users/javidan/Developer/Data Engineering/PandasDescriptiveStatistics/data/random_user.csv
```

Introduction to data

In [ ]:
```
df.head()
```

Out[ ]:

	gender	name.title	name.first	name.last	location.street.number	location.street.name	location.city	location.state	location.country	location.postcode	...	registered.d
0	female	Miss	Miriam	Lane	7994	Railroad St	Tamworth	Tasmania	Australia	4266	...	2010-24T16:24:10.8
1	male	Mr	Magnus	Rasmussen	9650	Vægterparken	Haslev	Sjælland	Denmark	68442	...	2018-07T10:31:39.2
2	female	Mrs	Ava	Parker	9511	Springfield Road	Thurles	Wicklow	Ireland	64055	...	2011-18T09:52:32.8
3	female	Ms	Gabija	Bøen	6825	Axel Brinchs vei	Judaberg	Oppland	Norway	0985	...	2011-22T06:42:42.0
4	female	Miss	Sofia	Marshall	7717	The Crescent	Tuam	Dublin City	Ireland	14431	...	2019-16T16:42:18.5

5 rows × 34 columns

In [ ]:
```
df.shape
```

Out[ ]:
```
(500, 34)
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 34 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   gender                           500 non-null    object 
 1   name.title                       500 non-null    object 
 2   name.first                       500 non-null    object 
 3   name.last                        500 non-null    object 
 4   location.street.number           500 non-null    int64  
 5   location.street.name             500 non-null    object 
 6   location.city                    500 non-null    object 
 7   location.state                   500 non-null    object 
 8   location.country                 500 non-null    object 
 9   location.postcode                500 non-null    object 
 10  location.coordinates.latitude    500 non-null    float64
 11  location.coordinates.longitude   500 non-null    float64
 12  location.timezone.offset         500 non-null    object 
 13  location.timezone.description    500 non-null    object 
 14  email                            500 non-null    object 
 15  login.uuid                       500 non-null    object 
 16  login.username                   500 non-null    object 
 17  login.password                   500 non-null    object 
 18  login.salt                       500 non-null    object 
 19  login.md5                        500 non-null    object 
 20  login.sha1                       500 non-null    object 
 21  login.sha256                     500 non-null    object 
 22  dob.date                         500 non-null    object 
 23  dob.age                          500 non-null    int64  
 24  registered.date                  500 non-null    object 
 25  registered.age                   500 non-null    int64  
 26  phone                            500 non-null    object 
 27  cell                             500 non-null    object 
 28  id.name                          410 non-null    object 
 29  id.value                         410 non-null    object 
 30  picture.large                    500 non-null    object 
 31  picture.medium                   500 non-null    object 
 32  picture.thumbnail                500 non-null    object 
 33  nat                              500 non-null    object 
dtypes: float64(2), int64(3), object(29)
memory usage: 132.9+ KB
```

In [ ]:
```
res = generate_summary(df)
print(res)
```

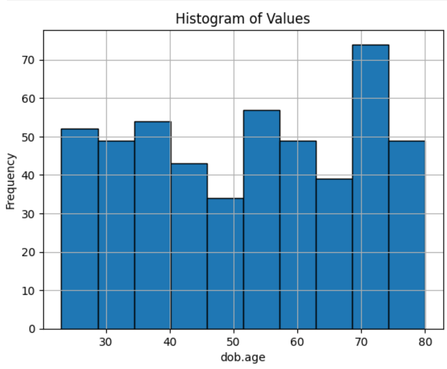
```
       location.street.number  location.coordinates.latitude  \
count              500.000000                     500.000000   
mean              5009.856000                       2.401864   
std               2975.949321                      51.264156   
min                 11.000000                     -89.617700   
25%               2333.250000                     -38.863575   
50%               4855.500000                       2.400400   
75%               7729.000000                      47.162275   
max               9974.000000                      89.715600   

       location.coordinates.longitude     dob.age  registered.age  
count                      500.000000  500.000000      500.000000  
mean                        -2.334455   52.024000       11.630000  
std                         99.085900   16.809657        5.761901  
min                       -179.263700   23.000000        2.000000  
25%                        -85.259400   36.750000        6.750000  
50%                         -1.398000   54.000000       12.000000  
75%                         80.207625   68.000000       17.000000  
max                        179.061400   80.000000       22.000000  
```

In [ ]:

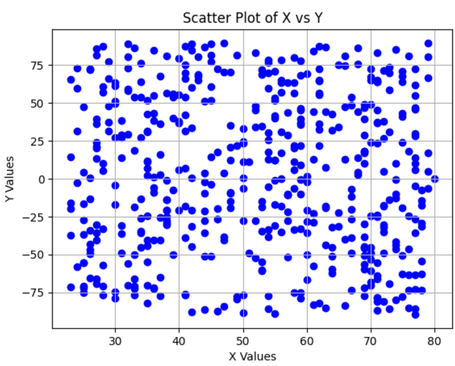
Visualization

In [ ]:
```
create_histogram(df, column = 'dob.age' )
```



Out[ ]:
```
<module 'matplotlib.pyplot' from '/Users/javidan/Developer/Data Engineering/PandasDescriptiveStatistics/venv/lib/python3.12/site-packages/matplotlib/pyplot.py'>
```

In [ ]:
```
create_scatter_plot(df, x_col = 'dob.age', y_col='location.coordinates.latitude')
```




```
Scatter plot saved as 'scatter_plot.png'.
```

Out[ ]:
```
<module 'matplotlib.pyplot' from '/Users/javidan/Developer/Data Engineering/PandasDescriptiveStatistics/venv/lib/python3.12/site-packages/matplotlib/pyplot.py'>
```

In [ ]: