In [111]:

```
import pandas as pd
import numpy as np
import scipy
import matplotlib.pyplot as plt
import sklearn
```

In [112]:

```
# LoadDatabase

file_path = './names/names/yob{year}.txt'
columns = ['Name', 'Sex', 'Number','Year']
dfs = pd.read_csv(file_path.format(year=1880), names=columns)
dfs['Year'] = 1880

# load Dataset yang lainya
# for year in range(1881, 2015):
#     df = pd.read_csv(file_path.format(year=year), names=columns)
#     df['Year'] = year
#     dfs.append(df)
```

In [113]:

```
dfs.head(10)
```

Out[113]:

| | Name | Sex | Number | Year |
|---|---|---|---|---|
| **0** | Mary | F | 7065 | 1880 |
| **1** | Anna | F | 2604 | 1880 |
| **2** | Emma | F | 2003 | 1880 |
| **3** | Elizabeth | F | 1939 | 1880 |
| **4** | Minnie | F | 1746 | 1880 |
| **5** | Margaret | F | 1578 | 1880 |
| **6** | Ida | F | 1472 | 1880 |
| **7** | Alice | F | 1414 | 1880 |
| **8** | Bertha | F | 1320 | 1880 |
| **9** | Sarah | F | 1288 | 1880 |

In [114]:

```
dfs.describe()
```

Out[114]:

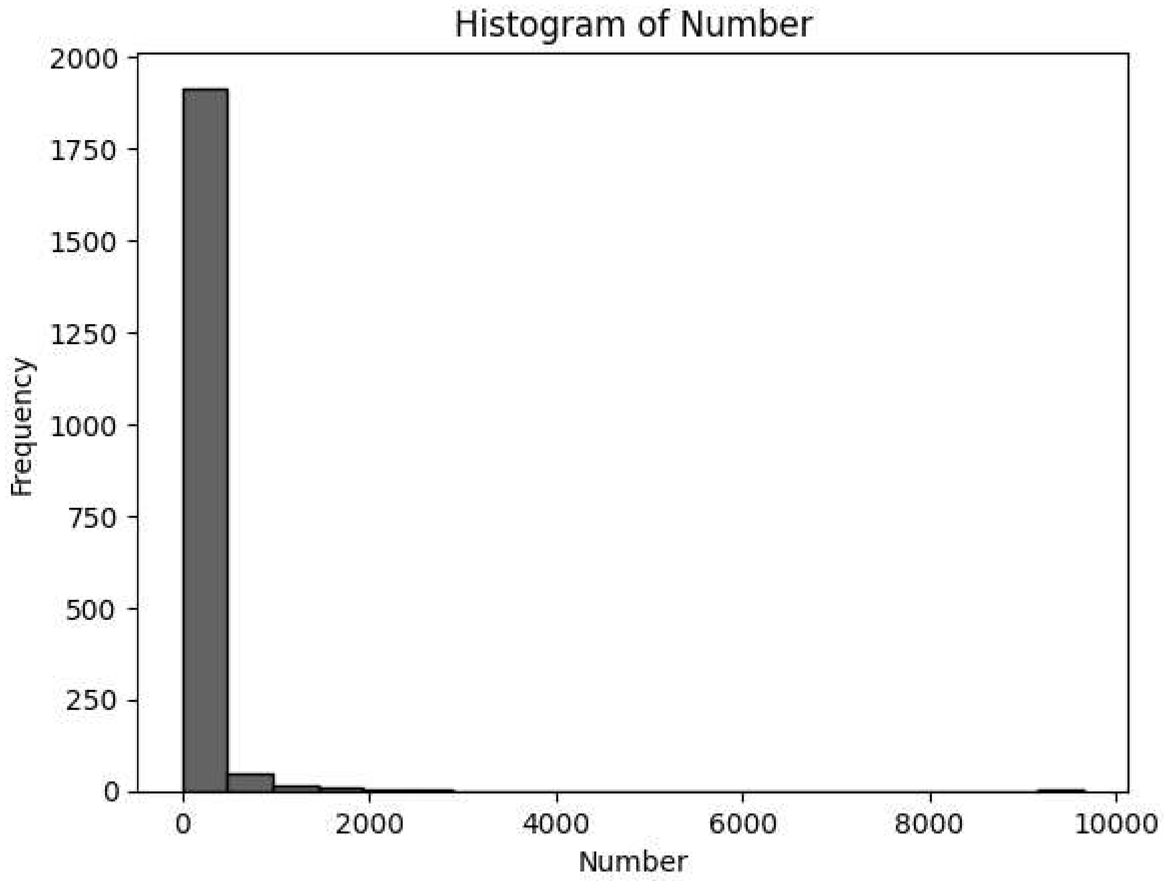
|  | Number | Year |
|---|---|---|
| **count** | 2000.000000 | 2000.0 |
| **mean** | 100.742000 | 1880.0 |
| **std** | 466.108732 | 0.0 |
| **min** | 5.000000 | 1880.0 |
| **25%** | 7.000000 | 1880.0 |
| **50%** | 13.000000 | 1880.0 |
| **75%** | 41.250000 | 1880.0 |
| **max** | 9655.000000 | 1880.0 |

# PReprocessing

# ANALIZE NUMBER

In [115]:

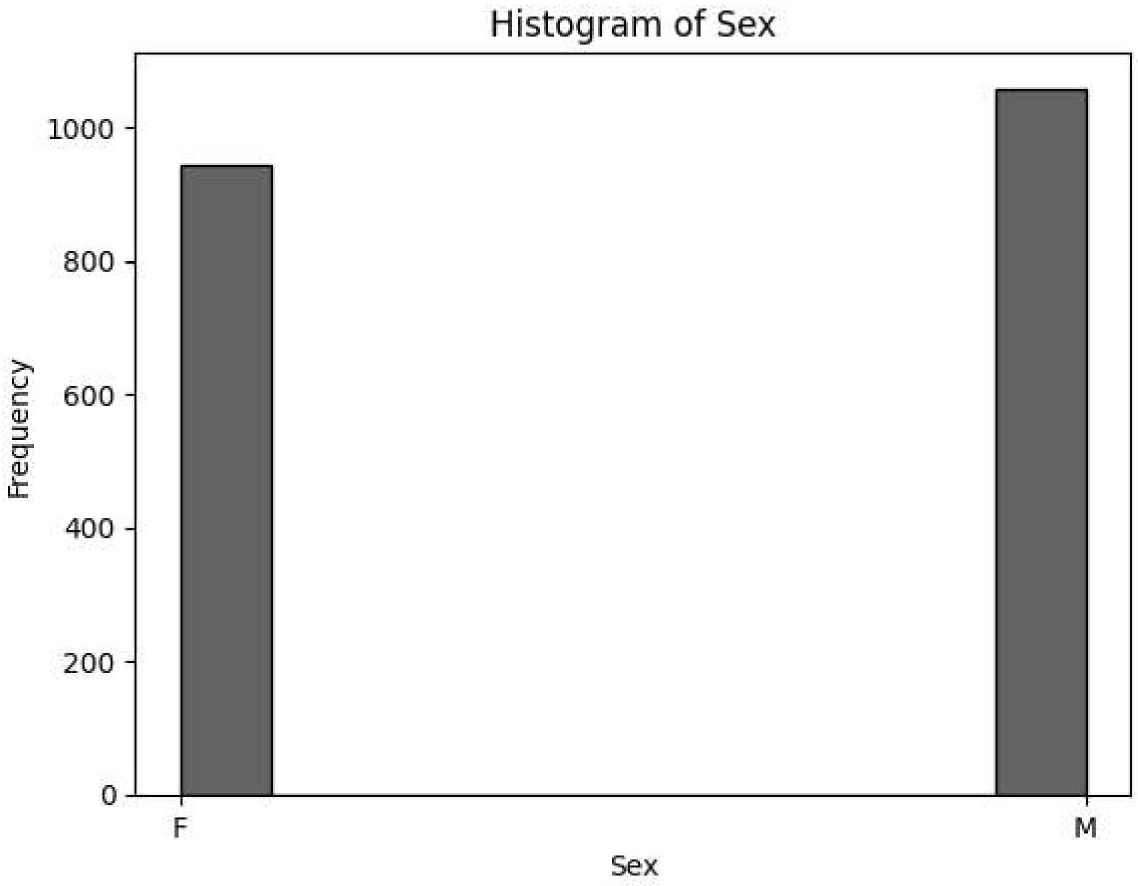
```
#  histogram column 'Number'
bins_ = 20
plt.hist(dfs['Number'], bins=20, edgecolor='black')

#  labels
plt.xlabel('Number')
plt.ylabel('Frequency')
plt.title('Histogram of Number')

plt.show()
```



# Analyze Sex

In [116]:

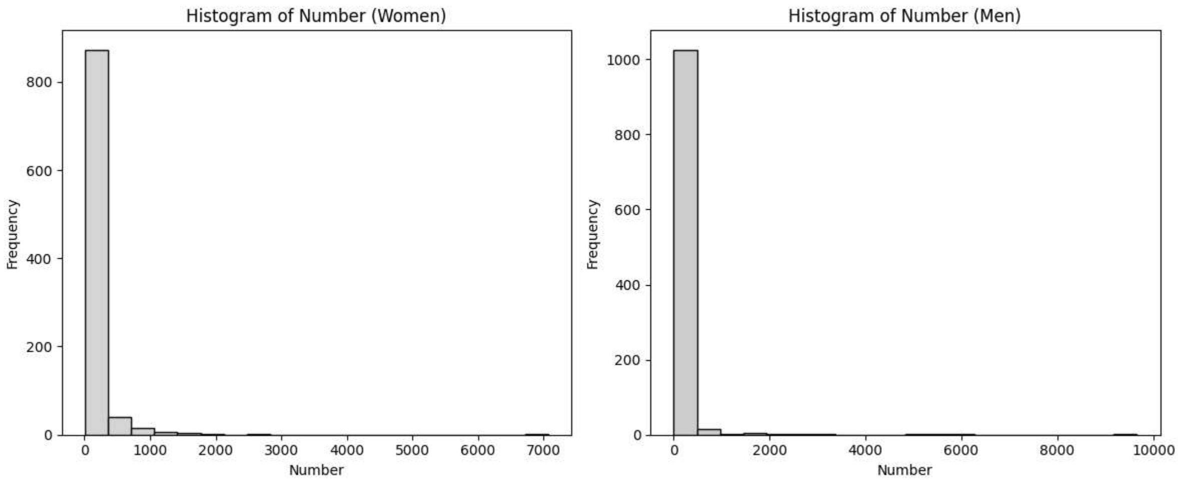
```
#  histogram column 'Numbe '
plt.hist(dfs['Sex'], edgecolor='black')

#  labels
plt.xlabel('Sex')
plt.ylabel('Frequency')
plt.title('Histogram of Sex')

plt.show()
```



In [120]:

```
# Separate data by sex
women_df = dfs[dfs['Sex'] == 'F']
men_df = dfs[dfs['Sex'] == 'M']


print("Women : ", women_df.count())
print("Men : ", men_df.count())
# subplots for women and men
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

bins_ = 20
# Plot histogram for women
ax1.hist(women_df['Number'], bins=bins_, edgecolor='black', color='pink')
ax1.set_xlabel('Number')
ax1.set_ylabel('Frequency')
ax1.set_title('Histogram of Number (Women)')

# Plot histogram for men
ax2.hist(men_df['Number'], bins=bins_, edgecolor='black', color='lightblue')
ax2.set_xlabel('Number')
ax2.set_ylabel('Frequency')
ax2.set_title('Histogram of Number (Men)')


# Show the plots
plt.tight_layout()
plt.show()
```

```
Women :  Name      942
Sex       942
Number    942
Year      942
dtype: int64
Men :  Name      1058
Sex       1058
Number    1058
Year      1058
dtype: int64
```



In [121]:

```
print("Women : ", women_df.head(10))
print("Men : ", men_df.head(10))
```

```
Women :          Name Sex  Number  Year
0       Mary   F    7065  1880
1       Anna   F    2604  1880
2       Emma   F    2003  1880
3  Elizabeth   F    1939  1880
4     Minnie   F    1746  1880
5   Margaret   F    1578  1880
6        Ida   F    1472  1880
7      Alice   F    1414  1880
8     Bertha   F    1320  1880
9      Sarah   F    1288  1880
Men :          Name Sex  Number  Year
942     John   M    9655  1880
943  William   M    9532  1880
944    James   M    5927  1880
945  Charles   M    5348  1880
946   George   M    5126  1880
947    Frank   M    3242  1880
948   Joseph   M    2632  1880
949   Thomas   M    2534  1880
950    Henry   M    2444  1880
951   Robert   M    2415  1880
```

# insight

Distribusi data M dan F mirip dengan Men sejumlah 1058 dan Female 9

nama paling populer untuk laki adalah john william james

nama paling populer untuk perempuan adalah mary anna dan emma

Type *Markdown* and LaTeX: $\alpha^2$