DUNAEC CENSUS ANALYZING (1778/1897)

Github link: https://github.com/tbaraniuk/dunaec-cencuc-analyzing/blob/main/main.ipynb

Presented by: Taras Baraniuk

PROJECT GOALS

- Explore the gender ratio of different age groups, how it has changed over a hundred years
- Investigate family structure and size; how it has changed over the past hundred years
- Find are there surnames or families from both databases that can be linked

STEPS

Read the datasets

Preprocess the data

Visualize the graphs

USED TOOLS

- Numpy, Pandas – for processing the data;
- Dotenv – for reading environment data;
- Deepl – for translating the surnames;
- Wordcloud – for displaying the words;
- Nbformat – for formatting the files in Jupyter Notebook;
- Matplotlib, Seaborn, Plotly – for visualizing the graphs;

PREPROCESS DATA

```
people1_df = people1_df.loc[:, ['ID строки в базе', 'ID Домохозяйс
глава семьи', 'Возраст', 'Семейный статус', 'Сословие, состояние и
проживает', 'Умеет ли читать', 'Обучение', 'Профессия главное', 'П
повинности']]
```

✓ 0.0s

```
people2_df = people2_df.loc[:, ['ID наскрізна', 'ID домогосподарство', 'Чолові
'Прізвище', 'Родиний статус', 'Категорія', 'Клас', 'Соціальний статус', 'Вік']
```

✓ 0.0s

```
household1_df.rename(columns={
    'ID Домохозяйствао': 'id',
    'Село/деревня': 'village/hamlet',
    'Хозяин': 'host_name',
    'Сколько жилых строений': 'residential_buildings_count',
    'Всего наличного мужеского населения': 'total_male_count',
    'Всего наличнаго женского населения': 'total_female_count',
    'Постоянно живущаго М': 'permanent_male_count',
    'Постоянно живущаго Ж': 'permanent_female_count',
    'Приписанного здесь М': 'attributed_male_count',
    'Приписанного здесь Ж': 'attributed_female_count'
}, inplace=True)
```

✓ 0.0s

```
people1_df.loc[:, 'name'] = people1_df.loc[:, 'name'].apply(str.strip)
people1_df.loc[:, 'surname'] = people1_df.loc[:, 'name'].str.extract(r'(\w+)\s', expand=False)
people1_df.loc[:, 'first_name'] = people1_df.loc[:, 'name'].str.extract(r'\s(\w+)\s', expand=False)

people1_df.patronymic = ''
people1_df.loc[people1_df.name.apply(lambda x: len(x.split(' '))) == 3, 'patronymic'] = people1_df \
    .loc[people1_df.name.apply(lambda x: len(x.split(' '))) == 3, 'name'] \
    .str.extract(r'\s\w+\s(\w+)', expand=False)

people1_df.drop(labels=['name'], axis=1, inplace=True)
```

✓ 0.0s

```
# Delete last rows with full empty data
people1_df = people1_df[people1_df.name.notna()]
```

✓ 0.0s

```
people1_df.loc[~people1_df.age.str.isnumeric(), 'age'] = '0'
```

✓ 0.0s

```
people1_df.age = people1_df.age.astype(np.int32)
```

✓ 0.0s

PREPROCESS DATA (CONTINUATION)

```
people2_df['sex'] = ''
people2_df.loc[people2_df.men_count.notna(), 'sex'] = 'm'
people2_df.loc[people2_df.women_count.notna(), 'sex'] = 'f'
```

✓ 0.0s

```
people2_df.drop(axis=1, labels=['men_count', 'women_count'], inplace=True)
```

✓ 0.0s

```
people2_df.loc[:, ['household_id', 'category', 'class', 'surname']] = people2_df \
    .loc[:, ['household_id', 'category', 'class', 'surname']] \
    .ffill()

people2_df.household_id = people2_df.household_id.astype(np.int32)
```

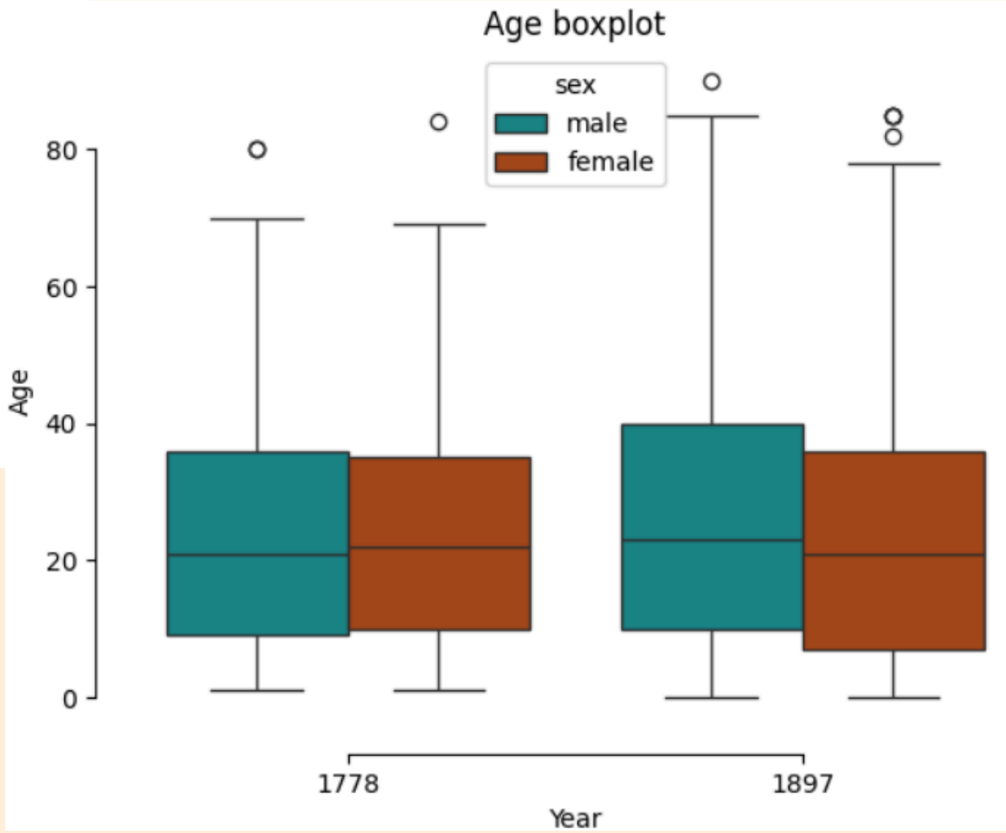
✓ 0.0s

```
family_total_structure_2_df.loc[:, 'category_total_count'] = family_total_structure_2_df.loc[:, 'category_total_count'].
bfill()
```

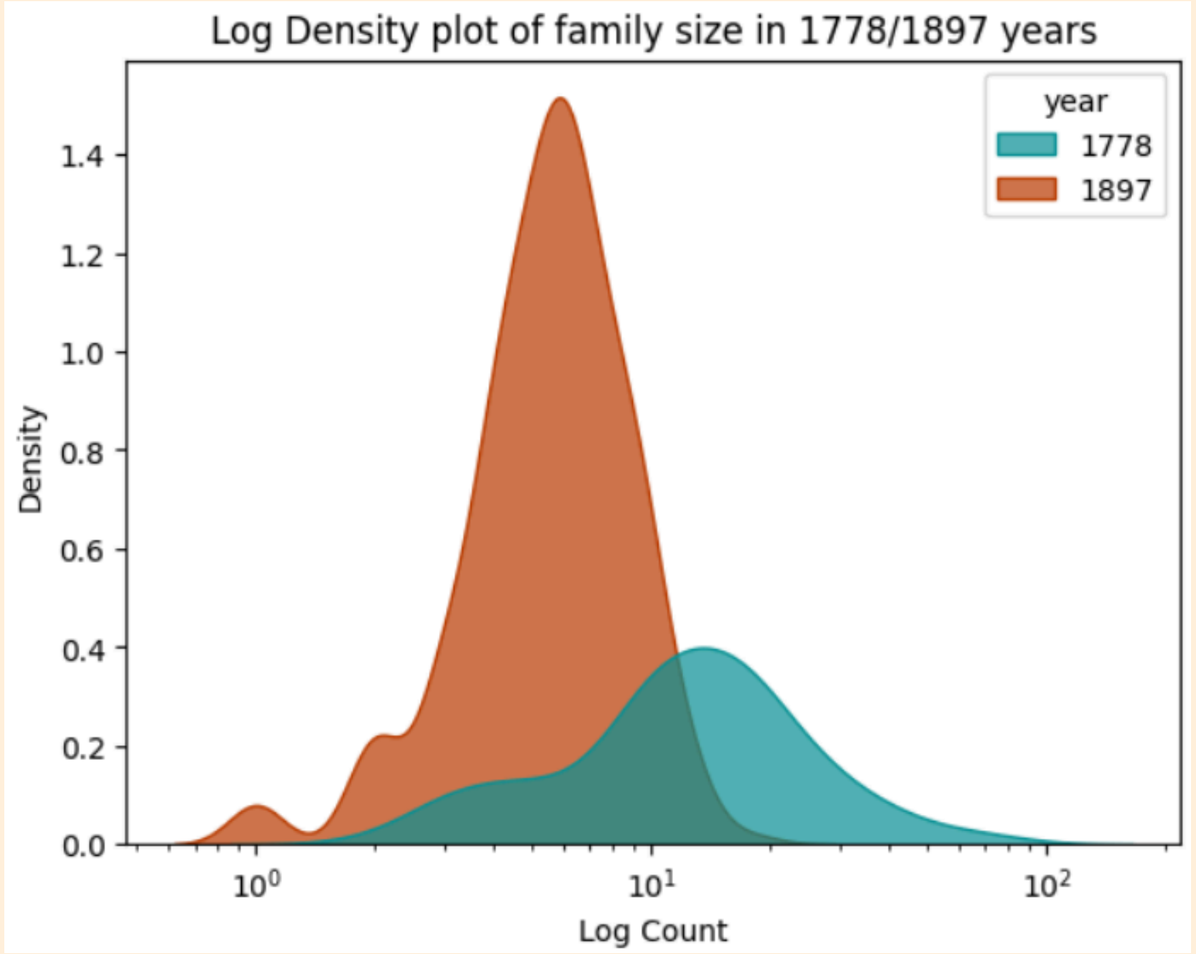
✓ 0.0s

```
# Delete Total Column
family_total_structure_1_df = family_total_structure_1_df[family_total_structure_1_df.category != 'Усього']
family_total_structure_2_df = family_total_structure_2_df[family_total_structure_2_df.category != 'Усього']
```

SEX-AGE DISTRIBUTION

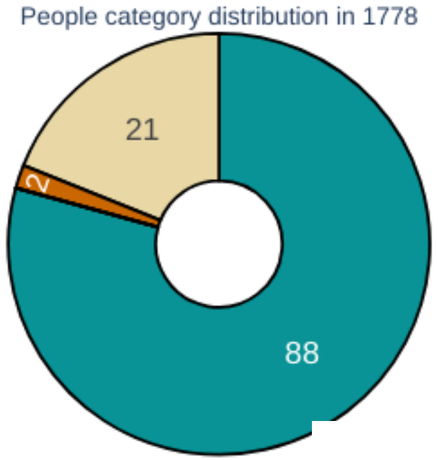


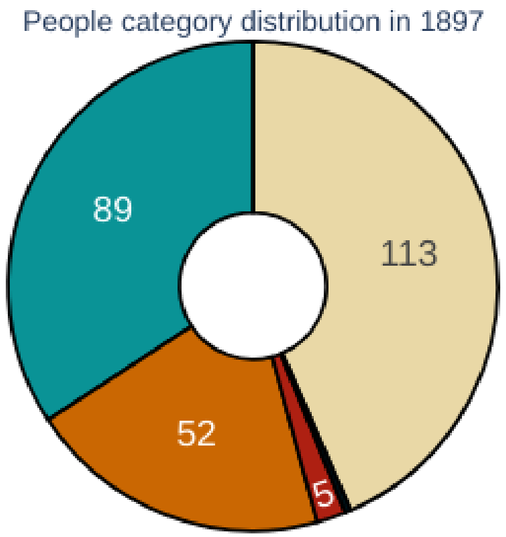
FAMILY SIZE DISTRIBUTION

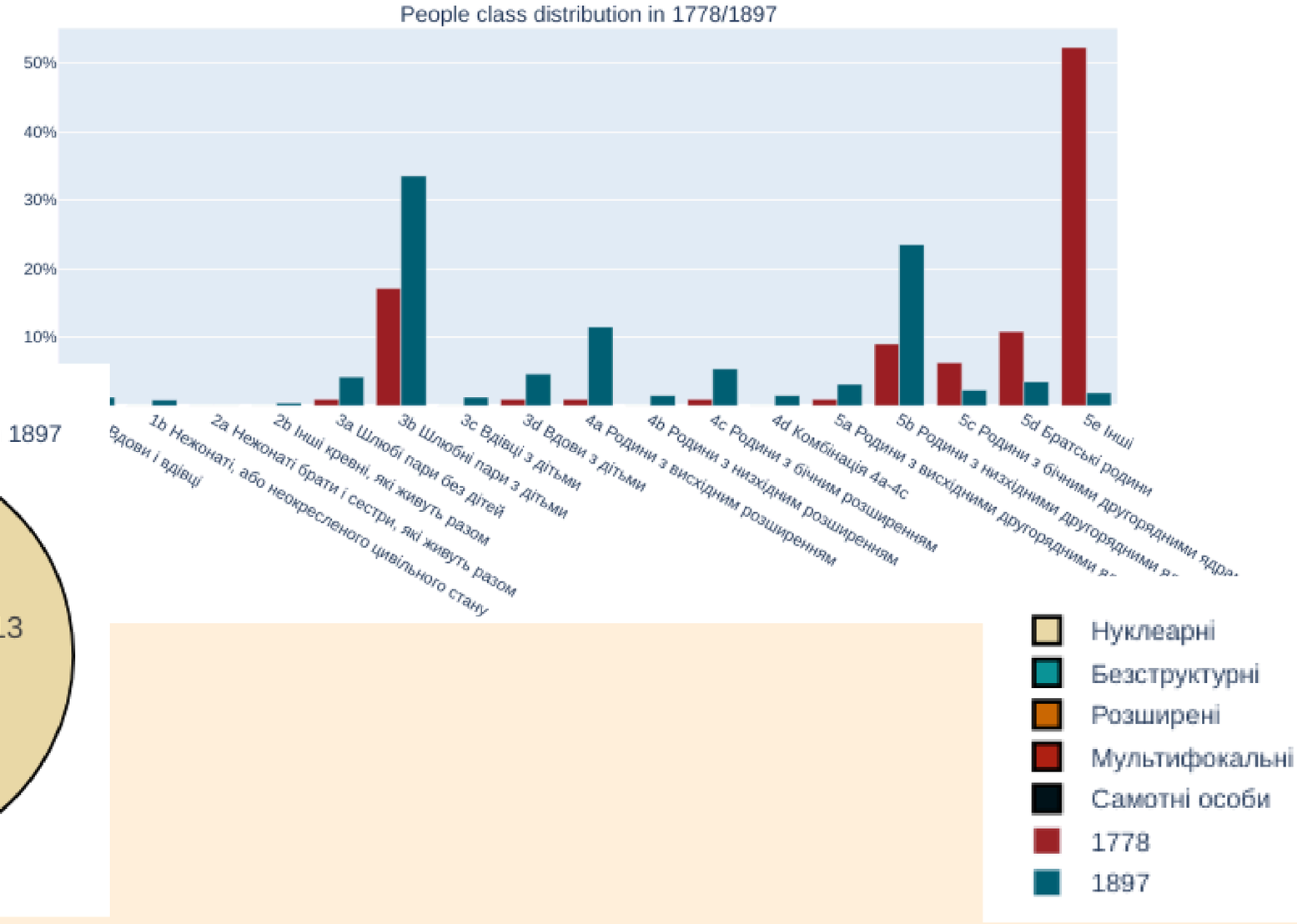



```
The average family size in 1778 is 15.21
And the average family size in 1897 is 5.94
```

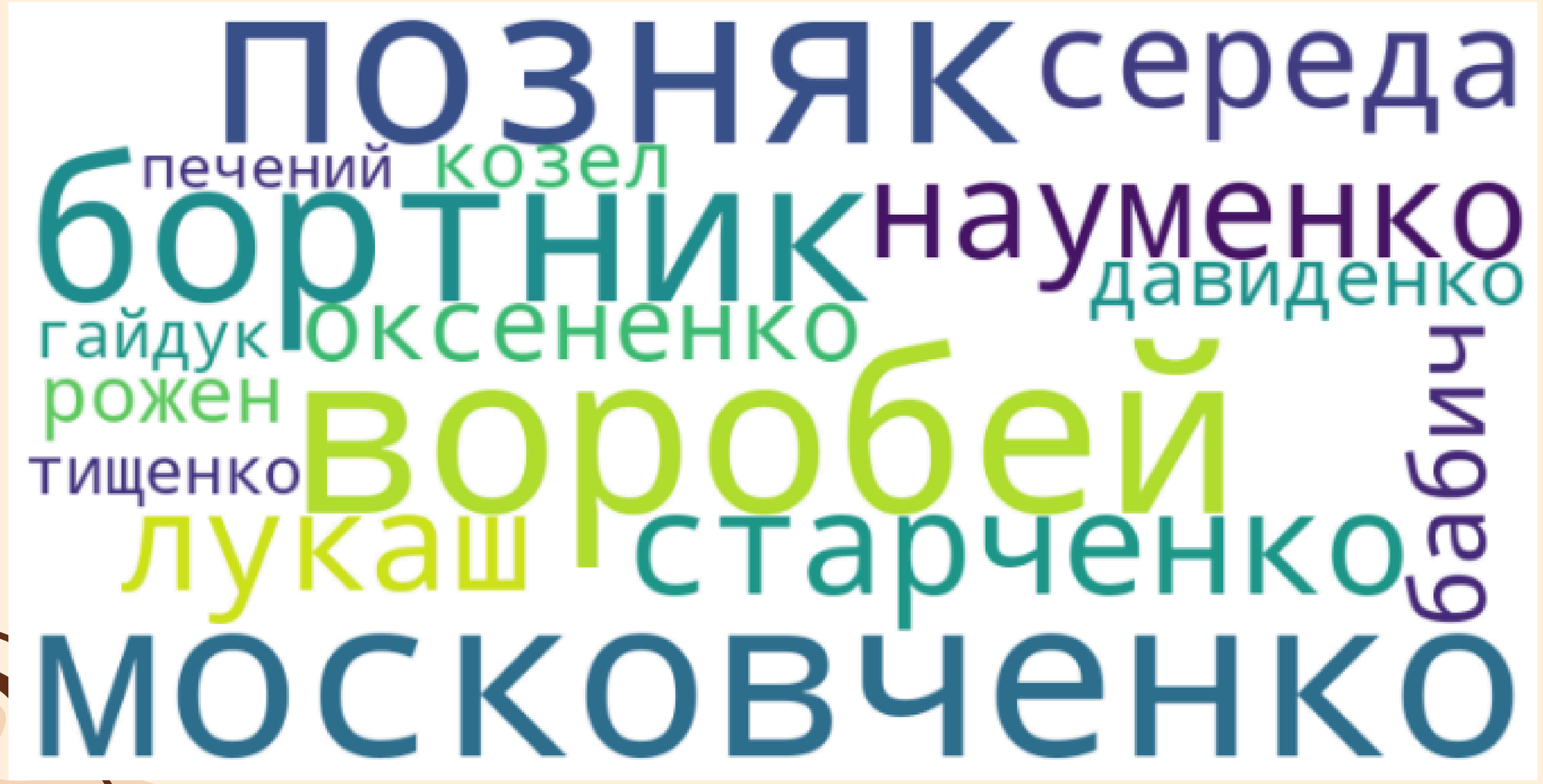
FAMILY TOTAL STRUCTURE COMPARISON







SURNAME COMPARISON



CONCLUSIONS

- **The average age and sex ratio are remained stable ;**
- **The sex ratio proportion is almost the same (≈ 1);**
- **The family size during the period was highly decreased (from 15 => 6);**
- **The percentage of nuclear and expanded families was highly increased (from 19% => 43.5%, and from 1.8% => 20% respectively)**
- **The percentage of "other" family class was extremely decreased (from 58% => 2%), and so the percentage of the others classes was increase (particularly for "3b** Шлюбні пари з дітьми" **and for "5b** Родини з низхідними другорядними ядрама")
- **The most common surnames in both years are:** Позняк, Московченко, Воробей, Бортник...

THANK YOU