**Description of the folder**

The folder contains the results of the cleaning, geolocating, and collapsing the Deprivation and Okhrana data on the province-year level. It contains the three subfolders “python\_code”, “raw\_data” and “results”.

* Folder **“python\_code”**: File ***okhrana\_merge.ipynb*** contains Python **code** that merges several files containing Deprivation/Okhrana data. Some cleaning is performed as well. After that, it geolocates the observations based on the locations of the courts (for Deprivation) or gendarme custodies (for Okhrana). The results are further collapsed at the province(governorate)-year level. ***Below you can find more detailed description of the code. Note that the Python file includes fairly detailed comments for further reference.***
  + ***Deprivation part of the code***. Based on *Deprivation\_Rights\_MinJust\_1894-1903.xlsx* and *Deprivation\_Rights\_MinJust\_1904-1914.xlsx*, I merge the data to create the unique dataset. Then, I fix some typos in the court names and add cites where the courts are located using *deprivation\_fix\_courtnames.xlsx.* Then, I drop duplicates and geolocate the data based on cities. For the geolocation, I rely on *Administrative Structure and Coordinates\_3.dta.* The data is then collapsed at the province-year level. The dependent variable *count\_deprivation* shows number of observations for that province and year in Deprivation data.
  + ***Okhrana part of the code*.** Basedon data sources described below, we append the data from several sources, clean it and drop the duplicates. I fix some typos in the court names and add cites where the gendarme custodies are located using *fix\_gendarme\_custody\_investigation.xlsx* and *okhrana\_fix\_gendarme.xlsx.* After that, the observations are geolocated based on the gendarme custody and collapsed at the province-year level similar to the Deprivation part of the code. The dependent variable *count\_okhrana* shows number of observations for that province and year in Okhrana data.
* Folder **“*Data\_result”*** contains 4 resulting data files: Deprivation and Okhrana datasets at individual (*results\_deprivation\_geolocated.xlsx* and *results\_okhrana\_geolocated.xlsx)* and province-year levels (*results\_deprivation\_collapsed.xlsx* and *results\_okhrana\_collapsed.xlsx).*
* Folder **“*auxiliary”*** contains ***auxiliary files*** that are used by the code. In particular, Administrative *Structure and Coordinates\_3.dta* is used for geolocation, *deprivation\_fix\_courtnames.xlsx* is used to fix the court names in the Deprivation data, *okhrana\_fix\_gendarme.xlsx* and *fix\_gendarme\_custody\_investigation.xlsx* do the same for the gendarme custodies in 2 parts of the Okhrana data. In the last 3 files variable that ends with \_old denotes the initial court name / gendarme custody, \_new is the updated (unified) name, and city is the place where gendarme custody / court is located. City is then used for geolocation.
* Folder **“*Data\_raw”*** contains the raw data that is used for the data sets construction.

**Details on the data**

* **Deprivation data** (*deprivation\_geolocated.xlsx* and *deprivation\_collapsed.xlsx*) is based on "Deprivation\_Rights\_MinJust\_1894-1903" and "Deprivation\_Rights\_MinJust\_1904-1914".
  + ***Deprivation\_geolocated.xlsx*** contains observations at the individual level. *Column court\_new* essentially fixes the typos in the column "Каким судебным учреждением" and *city* denotes the city where the court is located. *Year\_deprivation* is the year of deprivation. Remaining variables started from *gub* are taken from *Administrative Structure and Coordinates\_3.dta.* The most important of them are governorate’s id (*gnr*) and name (*gub\_new*)
  + ***Deprivation\_collapsed.xlsx*** contains observations at the province-year level. Governorate’s id (*gnr*) and name (*gub\_new*) are taken from *Administrative Structure and Coordinates\_3.dta. year\_deprivation* is the time variable. *Count\_deprivation* denotes the number of observations for this province and year.
  + Some years do not exist in raw data: there is no year 1894 in the file "Deprivation\_Rights\_MinJust\_1894-1903". In file "Deprivation\_Rights\_MinJust\_1904-1914" years 1910, 1911, 1913, and 1914 are missing.
  + Deprivation data has been geolocated using the file "Administrative divisions/Administrative Structure and Coordinates\_3.dta". I kept all the data stored in that file. Only 11 out of 1513 observations have not been successfully geolocated. 6 of them have several court locations, e.g., Moscow and Riga, 1 has no court location. 1 (Выборг) is also not in the data as, if I understand correctly, it was a part of Finland at that point. The remaining 4 are Порт Артур & Пограничный (x2) which are not presented in the data due to the history of expansion in Asia.
* **Okhrana data** (*okhrana\_geolocated.xlsx* and *okhrana\_collapsed.xlsx*) is based on several sources. These sources can be split into 3 parts:

1. ***Accused*** individuals from files *"* *Raw Data on the List of the Accused by the Okhrana, 1895-1899.xlsx", "* *Raw Data on the List of the Accused by the Okhrana, 1900-1901.xlsx"*
2. ***Revolutionaries’*** data *from "* *Raw Data on the List of Revolutionaries (State Crimes) by the Okhrana (Vedomost), 1887-1897.xlsx”*
3. People under ***investigation*** (4 files with names starting from *Raw Data on the Alphabet of People under Investigation*…).

* In a previous version, the data also included observations from “*names mentioned in Obzor 1900-1901”.* We dropped this data as it contains nothing apart from names and surnames
* To reflect the differences in the data sources, I created column ***data\_source*** in Okhrana dataset with 3 values (accused / revolutionaries / investigation) respectively.
* Additionally, column ***year\_okhrana*** is obtained from file or sheet names, e.g. for the file " Alphabet\_People Under Investigation\_1899" years="1899"
  + ***Okhrana\_geolocated.xlsx*** contains observations at the individual level. *Column gendarme\_new* essentially fixes the typos in the column "Gendarme custody" and *city* denotes the city where the court is located. *Year\_okhrana* is the year of observation. Variable *data\_*source is described above. Remaining variables started from *gub* are taken from *Administrative Structure and Coordinates\_3.dta.* The most important of them are governorate’s id (*gnr*) and name (*gub\_new*)
  + ***Okhrana \_collapsed.xlsx*** contains observations at the province-year level. Governorate’s id (*gnr*) and name (*gub\_new*) are taken from *Administrative Structure and Coordinates\_3.dta. year\_okhrana* is the time variable. *Count\_okhrana* denotes the number of observations for this province and year.
  + *Investigation data* (data\_source = investigation) is taken from 4 files named *Raw Data on the Alphabet of People under Investigation…*. Column *investigation\_type* is defined only for this part of the data. It is taken from the first row of the corresponding sheets. i.e.
    - "politically unreliable" = "Список розыскиваемых политиески неблагонадежных лиц",
    - "fled" = "Список лиц скрывшихся, которые по розыску их подлежат подчинению негласному надзору (207-223)',
    - "article 246" = "Список лиц, обвиняемых в совершении преступления, предусмотренного 246 ст. Улож. о Наказ. и скрывшихся (225)",
    - "non political" = "Список лиц, розыскиваемых Департаментом Полиции по делам неполитическим (227-229)" or "Список лиц, розыскиваемых Департаментом Полиции по делам не политическим и помещенным в циркуляры на имя Губернаторов, Градоначальников и Обер-Полицмейстеров"
    - "all" = "Алфавитный указатель к розыскному списку и циркулярам Департамента Полиции"
    - "arrested" = "Список лиц, задержанных во время печатания настоящего списка"
    - "police department" = "Список лиц, розыскиваемых по делам Департамента Полиции"
    - "measures 1889" = Список лиц, розыскиваемых, в отношении которых, по обнаружении места жительства, надлежить принять меры, указанные в циркуляре Департамента Полиции от 06.09.1889 за №2938
  + The *duplicate dropping* is organized in the following way. First, within the pair of gendarme custody and year, we drop all the observations with the same full name (surname, name, and patronymic). This approach is pretty conservative and may leave out some observations with the persons of the same full name but different age though it should not be a common situation. We also do additionally search for duplicates accounting for capital VS lowercase letters and absence of patronymic. Note that this refinement is done only for observations with non-empty surname. This approach leaves us with 13 956 out of 14 881 initial observations (93.7%)
  + There are 13 956 observations in the Okhrana dataset. 5 201 do not have data on Gendarme custody. Out of the remaining 8755, I geolocate all but 18. The remaining observations have either several cities or very vague depictions ("nearest governorate")