Формирование итогового датасета

СОДЕРЖАНИЕ

0 ИМПОРТИРОВАНИЕ БИБЛИОТЕК

In [1]:
```
import pandas as pd
import copy
```

Библиотека **pandas** потребуется для работы с данными в табличном представлении.

Библиотека **copy** потребуется для создания глубоких копий объектов.

1 ЗАГРУЗКА РЕЕСТРА МСП

Загружаем ранее сформированный CSV файл, содержащий сведения о субъектах малого и среднего предпринимательства Приморского края.

In [2]:
```
# Создаем переменную names с названием колонок датафрейма
names = ['Наименование МСП', 'ИНН', 'Код региона', 'Вид', 'Категория','Регион',
         'Район', 'Город', 'Населенный пункт', 'Код ОКВЭД', 'Вид деятельности']
# Создадим датафрейм
df = pd.read_csv('reestr_msp_2019.csv', # Указываем название файла
                 sep='[q]',             # Указываем разделитель
                 encoding='utf-8',      # Указываем кодировку
                 names=names,           # Передаем названия колонок
                 engine='python')       # Выбираем движок (python, так как в нем реализованно
                                        # указание регулярного выражения в качестве разделителя)
# Выводим первые 5 строк датафрейма
df.head()
```

Out[2]:

	Наименование МСП	ИНН	Код региона	Вид	Категория	Регион	Район	Город	Населенный пункт	Код ОКВЭД	де
0	СТЕПАНОВА ИРИНА	644934682794	25	Вид 2	Категория 1	КРАЙ ПРИМОРСКИЙ		ГОРОД ФОКИНО		95.21	з быто
1	КУРКИН ЕВГЕНИЙ	870100455199	25	Вид 2	Категория 1	КРАЙ ПРИМОРСКИЙ		ГОРОД УССУРИЙСК		49.32	Д легко ар
2	ГОДУНОВА СВЕТЛАНА	870600600221	25	Вид 2	Категория 1	КРАЙ ПРИМОРСКИЙ		ГОРОД ВЛАДИВОСТОК		47.11	преим пиш
3	НАВРОСЬ ДМИТРИЙ	870600537308	25	Вид 2	Категория 1	Край Приморский		Город Владивосток		95.29	Ре потре
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АГРО...	7017227901	25	Вид 1	Категория 1	КРАЙ ПРИМОРСКИЙ		ГОРОД ВЛАДИВОСТОК		01.25.1	Вы прочи и ягод

В данном представлении данных не очень удобно работать, непонятно к какому муниципальному образованию принадлежит субъект малого и среднего предпринимательства (есть районы, городские округа, населенные пункты), а также непонятно к какому виду и категории относится субъект (вид 1 или вид 2, категория 1, 2 или 3).

Создадим единую колонку с наименованием муниципального образования.

In [3]:
```
# Конкатенируем данные из двух колонок
df['Муниципальное образование'] = df['Район'] + df['Город']
```

Выведем в принт содержание созданной колонки.

In [4]:
```
# Используем set, чтобы оставить только уникальные значения
set(df['Муниципальное образование'])
```

Out[4]:
```
{'      ',
 '   ГОРОД АРСЕНЬЕВ',
 '   ГОРОД АРТЕМ',
 '   ГОРОД БОЛЬШОЙ КАМЕНЬ',
 '   ГОРОД ВЛАДИВОСТОК',
 '   ГОРОД ДАЛЬНЕГОРСК',
 '   ГОРОД ДАЛЬНЕРЕЧЕНСК',
 '   ГОРОД ЛЕСОЗАВОДСК',
 '   ГОРОД НАХОДКА',
 '   ГОРОД ПАРТИЗАНСК',
 '   ГОРОД СПАССК-ДАЛЬНИЙ',
 '   ГОРОД УССУРИЙСК',
 '   ГОРОД ФОКИНО',
 '   Город Арсеньев',
 '   Город Артем',
 '   Город Большой Камень',
 '   Город Владивосток',
 '   Город Дальнегорск',
 '   Город Дальнереченск',
 '   Город Лесозаводск',
 '   Город Находка',
 '   Город Партизанск',
 '   Город Спасск-Дальний',
 '   Город Уссурийск',
 '   Город Фокино',
 'РАЙОН АНУЧИНСКИЙ   ',
 'РАЙОН ДАЛЬНЕРЕЧЕНСКИЙ   ',
 'РАЙОН КАВАЛЕРОВСКИЙ   ',
 'РАЙОН КИРОВСКИЙ   ',
 'РАЙОН КРАСНОАРМЕЙСКИЙ   ',
 'РАЙОН ЛАЗОВСКИЙ   ',
 'РАЙОН МИХАЙЛОВСКИЙ   ',
 'РАЙОН НАДЕЖДИНСКИЙ   ',
 'РАЙОН ОКТЯБРЬСКИЙ   ',
 'РАЙОН ОЛЬГИНСКИЙ   ',
 'РАЙОН ПАРТИЗАНСКИЙ   ',
 'РАЙОН ПОГРАНИЧНЫЙ   ',
 'РАЙОН ПОЖАРСКИЙ   ',
 'РАЙОН СПАССКИЙ   ',
 'РАЙОН ТЕРНЕЙСКИЙ   ',
 'РАЙОН ХАНКАЙСКИЙ   ',
 'РАЙОН ХАСАНСКИЙ   ',
 'РАЙОН ХОРОЛЬСКИЙ   ',
 'РАЙОН ЧЕРНИГОВСКИЙ   ',
 'РАЙОН ЧУГУЕВСКИЙ   ',
 'РАЙОН ШКОТОВСКИЙ   ',
 'РАЙОН ЯКОВЛЕВСКИЙ   ',
 'Район Анучинский   ',
 'Район Дальнереченский   ',
 'Район Кавалеровский   ',
 'Район Кировский   ',
 'Район Красноармейский   ',
 'Район Лазовский   ',
 'Район Михайловский   ',
 'Район Надеждинский   ',
 'Район Октябрьский   ',
 'Район Ольгинский   ',
 'Район Партизанский   ',
 'Район Пограничный   ',
 'Район Пожарский   ',
 'Район Спасский   ',
 'Район Тернейский   ',
 'Район Ханкайский   ',
 'Район Хасанский   ',
 'Район Хорольский   ',
 'Район Черниговский   ',
 'Район Чугуевский   ',
 'Район Шкотовский   ',
 'Район Яковлевский   '}
```

Как можно заметить данные имеют разное написание (что приводит к дублированию) и появились пробелы (чуть позднее все приведем к одному виду). Кроме района и города в датафрейме также есть данные о населенных пунктах, которые необходимо конвертировать в муниципальное образование.

In [5]:
```
set(df['Населенный пункт'])
```

Out[5]:
```
{'   ',
 'ПОСЕЛОК ГОРОДСКОГО ТИПА ПОГРАНИЧНЫЙ',
 'ПОСЕЛОК ГОРОДСКОГО ТИПА СЛАВЯНКА',
 'СЕЛО ВЛАДИМИРО-АЛЕКСАНДРОВСКОЕ',
 'СЕЛО ВОЛЬНО-НАДЕЖДИНСКОЕ',
 'СЕЛО ГЛАЗОВКА',
 'СЕЛО ИЛЬМОВКА',
 'СЕЛО ИННОКЕНТЬЕВКА',
 'СЕЛО ЛЕСНОЕ',
 'СЕЛО НЕВСКОЕ',
 'СЕЛО ПАНТЕЛЕЙМОНОВКА',
 'СЕЛО ПОЛЕВОЕ',
 'СЕЛО РУЖИНО',
 'СЕЛО ТАМГА',
 'СЕЛО ФИЛАРЕТОВКА',
 'СЕЛО ЧУГУЕВКА'}
```

Добавим с помощью фильтрации каждому населенному пункту принадлежность к муниципальному образованию и запишем все в созданную ранее колонку с наименованием муниципального образования к которому принадлежит субъект малого и среднего предпринимательства Приморсеого края.

In [6]:
```
# Если в колонке населенный пункт встречается ПОСЕЛОК ГОРОДСКОГО ТИПА ПОГРАНИЧНЫЙ, то в колонку муницип
альное
# образование записываем РАЙОН ПОГРАНИЧНЫЙ и так дальше по аналогии
df.loc[df['Населенный пункт'] == 'ПОСЕЛОК ГОРОДСКОГО ТИПА ПОГРАНИЧНЫЙ', 'Муниципальное образование'] =
'РАЙОН ПОГРАНИЧНЫЙ'
df.loc[df['Населенный пункт'] == 'ПОСЕЛОК ГОРОДСКОГО ТИПА СЛАВЯНКА', 'Муниципальное образование'] = 'РА
ЙОН ХАСАНСКИЙ'
df.loc[df['Населенный пункт'] == 'СЕЛО ВЛАДИМИРО-АЛЕКСАНДРОВСКОЕ', 'Муниципальное образование'] = 'РАЙО
Н ПАРТИЗАНСКИЙ'
df.loc[df['Населенный пункт'] == 'СЕЛО ВОЛЬНО-НАДЕЖДИНСКОЕ', 'Муниципальное образование'] = 'РАЙОН НАДЕ
ЖДИНСКИЙ'
df.loc[df['Населенный пункт'] == 'СЕЛО ЧУГУЕВКА', 'Муниципальное образование'] = 'РАЙОН ЧУГУЕВСКИЙ'
df.loc[df['Населенный пункт'].isin(['СЕЛО ГЛАЗОВКА',
                                    'СЕЛО ИЛЬМОВКА',
                                    'СЕЛО ИННОКЕНТЬЕВКА',
                                    'СЕЛО ЛЕСНОЕ',
                                    'СЕЛО НЕВСКОЕ',
                                    'СЕЛО ПАНТЕЛЕЙМОНОВКА',
                                    'СЕЛО ПОЛЕВОЕ',
                                    'СЕЛО РУЖИНО',
                                    'СЕЛО ТАМГА',
                                    'СЕЛО ФИЛАРЕТОВКА']), 'Муниципальное образование'] = 'ГОРОД ЛЕСОЗАВ
ОДСК'
```

Уберем пробелы и приведем данные к верхнему регистру.

In [7]:
```
# str.strip() - убирает пробелы
# str.upper() - приводит к верхнему регистру
df['Муниципальное образование'] = df['Муниципальное образование'].str.strip().str.upper()
```

In [8]:
```
set(df['Муниципальное образование'])
```

Out[8]:
```
{'ГОРОД АРСЕНЬЕВ',
 'ГОРОД АРТЕМ',
 'ГОРОД БОЛЬШОЙ КАМЕНЬ',
 'ГОРОД ВЛАДИВОСТОК',
 'ГОРОД ДАЛЬНЕГОРСК',
 'ГОРОД ДАЛЬНЕРЕЧЕНСК',
 'ГОРОД ЛЕСОЗАВОДСК',
 'ГОРОД НАХОДКА',
 'ГОРОД ПАРТИЗАНСК',
 'ГОРОД СПАССК-ДАЛЬНИЙ',
 'ГОРОД УССУРИЙСК',
 'ГОРОД ФОКИНО',
 'РАЙОН АНУЧИНСКИЙ',
 'РАЙОН ДАЛЬНЕРЕЧЕНСКИЙ',
 'РАЙОН КАВАЛЕРОВСКИЙ',
 'РАЙОН КИРОВСКИЙ',
 'РАЙОН КРАСНОАРМЕЙСКИЙ',
 'РАЙОН ЛАЗОВСКИЙ',
 'РАЙОН МИХАЙЛОВСКИЙ',
 'РАЙОН НАДЕЖДИНСКИЙ',
 'РАЙОН ОКТЯБРЬСКИЙ',
 'РАЙОН ОЛЬГИНСКИЙ',
 'РАЙОН ПАРТИЗАНСКИЙ',
 'РАЙОН ПОГРАНИЧНЫЙ',
 'РАЙОН ПОЖАРСКИЙ',
 'РАЙОН СПАССКИЙ',
 'РАЙОН ТЕРНЕЙСКИЙ',
 'РАЙОН ХАНКАЙСКИЙ',
 'РАЙОН ХАСАНСКИЙ',
 'РАЙОН ХОРОЛЬСКИЙ',
 'РАЙОН ЧЕРНИГОВСКИЙ',
 'РАЙОН ЧУГУЕВСКИЙ',
 'РАЙОН ШКОТОВСКИЙ',
 'РАЙОН ЯКОВЛЕВСКИЙ'}
```

Сделаем проверку, по всем ли муниципальным образованиям содержится информация. Всего в Приморском крае 34 муниципальных образований.

In [9]:
```
if len(set(df["Муниципальное образование"])) == 34:
    print('Данные содержат информацию по всем муниципальным образованиям')
else:
    print('Данные содержат информацию не по всем муниципальным образованиям')
```

```
Данные содержат информацию по всем муниципальным образованиям
```

Также отформатируем колонки Вид и Категория, присвоив нужные значения (можно сверить с данными единого реестра субъектов малого и среднего предпринимательства на сайте www.nalog.ru).

In [10]:
```
# Виду 1 присвоим вид МСП юридическое лицо
df.loc[df['Вид'] == 'Вид 1', 'Вид'] = 'Юридическое лицо'
# Виду 2 присвоим вид МСП индивидуальный предприниматель
df.loc[df['Вид'] == 'Вид 2', 'Вид'] = 'Индивидуальный предприниматель'
# Категории 1 присвоим категорию микропредприятие
df.loc[df['Категория'] == 'Категория 1', 'Категория'] = 'Микропредприятие'
# Категории 2 присвоим категорию малое предприятие
df.loc[df['Категория'] == 'Категория 2', 'Категория'] = 'Малое предприятие'
# Категории 3 присвоим категорию среднее предприятие
df.loc[df['Категория'] == 'Категория 3', 'Категория'] = 'Среднее предприятие'
```

Выведем датафрейм.

In [11]:
```
df.head()
```

Out[11]:

	Наименование МСП	ИНН	Код региона	Вид	Категория	Регион	Район	Город	Населенн пу
0	СТЕПАНОВА ИРИНА	644934682794	25	Индивидуальный предприниматель	Микропредприятие	КРАЙ ПРИМОРСКИЙ		ГОРОД ФОКИНО	
1	КУРКИН ЕВГЕНИЙ	870100455199	25	Индивидуальный предприниматель	Микропредприятие	КРАЙ ПРИМОРСКИЙ		ГОРОД УССУРИЙСК	
2	ГОДУНОВА СВЕТЛАНА	870600600221	25	Индивидуальный предприниматель	Микропредприятие	КРАЙ ПРИМОРСКИЙ		ГОРОД ВЛАДИВОСТОК	
3	НАВРОСЬ ДМИТРИЙ	870600537308	25	Индивидуальный предприниматель	Микропредприятие	Край Приморский		Город Владивосток	
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АГРО...	7017227901	25	Юридическое лицо	Микропредприятие	КРАЙ ПРИМОРСКИЙ		ГОРОД ВЛАДИВОСТОК	

Уберем лишние данные (данные, которые в дальнейшем не несут никакой ценности). Для удобства выведем название колонок.

In [12]:
```
df.columns
```

Out[12]:
```
Index(['Наименование МСП', 'ИНН', 'Код региона', 'Вид', 'Категория', 'Регион',
       'Район', 'Город', 'Населенный пункт', 'Код ОКВЭД', 'Вид деятельности',
       'Муниципальное образование'],
      dtype='object')
```

In [13]:
```
# Создадим новый датафрейм, оставив нужные колонки
reestr = df[['Наименование МСП',
             'ИНН',
             'Муниципальное образование',
             'Вид',
             'Категория',
             'Код ОКВЭД',
             'Вид деятельности']]
reestr.head()
```

Out[13]:

	Наименование МСП	ИНН	Муниципальное образование	Вид	Категория	Код ОКВЭД	Вид деятельности
0	СТЕПАНОВА ИРИНА	644934682794	ГОРОД ФОКИНО	Индивидуальный предприниматель	Микропредприятие	95.21	Ремонт электронной бытовой техники
1	КУРКИН ЕВГЕНИЙ	870100455199	ГОРОД УССУРИЙСК	Индивидуальный предприниматель	Микропредприятие	49.32	Деятельность легкового такси и арендованных ле...
2	ГОДУНОВА СВЕТЛАНА	870600600221	ГОРОД ВЛАДИВОСТОК	Индивидуальный предприниматель	Микропредприятие	47.11	Торговля розничная преимущественно пищевыми пр...
3	НАВРОСЬ ДМИТРИЙ	870600537308	ГОРОД ВЛАДИВОСТОК	Индивидуальный предприниматель	Микропредприятие	95.29	Ремонт прочих предметов личного потребления и ...
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АГРО...	7017227901	ГОРОД ВЛАДИВОСТОК	Юридическое лицо	Микропредприятие	01.25.1	Выращивание прочих плодовых и ягодных культур

Сделаем проверку, все ли данные на месте, нет ли ошибок в наименованиях муниципалитетов.

In [14]:
```
# Создадим сводную таблицу
# Сгруппируем данные по наименованию муниципалитета
# В значение считаем количество субъектов МСП
# Сортируем данные по убыванию
pivot = reestr.groupby('Муниципальное образование').count()['Наименование МСП']\
                                                   .reset_index()\
                                                   .sort_values('Наименование МСП', ascending=False)
pivot
```

Out[14]:

	Муниципальное образование	Наименование МСП
3	ГОРОД ВЛАДИВОСТОК	43383
10	ГОРОД УССУРИЙСК	8785
7	ГОРОД НАХОДКА	7020
1	ГОРОД АРТЕМ	4360
0	ГОРОД АРСЕНЬЕВ	1692
19	РАЙОН НАДЕЖДИНСКИЙ	1432
4	ГОРОД ДАЛЬНЕГОРСК	1267
2	ГОРОД БОЛЬШОЙ КАМЕНЬ	1236
6	ГОРОД ЛЕСОЗАВОДСК	1123
9	ГОРОД СПАССК-ДАЛЬНИЙ	1090
8	ГОРОД ПАРТИЗАНСК	1035
28	РАЙОН ХАСАНСКИЙ	942
14	РАЙОН КАВАЛЕРОВСКИЙ	871
5	ГОРОД ДАЛЬНЕРЕЧЕНСК	842
18	РАЙОН МИХАЙЛОВСКИЙ	780
20	РАЙОН ОКТЯБРЬСКИЙ	742
22	РАЙОН ПАРТИЗАНСКИЙ	683
24	РАЙОН ПОЖАРСКИЙ	683
11	ГОРОД ФОКИНО	651
30	РАЙОН ЧЕРНИГОВСКИЙ	592
29	РАЙОН ХОРОЛЬСКИЙ	580
31	РАЙОН ЧУГУЕВСКИЙ	523
25	РАЙОН СПАССКИЙ	523
15	РАЙОН КИРОВСКИЙ	469
27	РАЙОН ХАНКАЙСКИЙ	423
32	РАЙОН ШКОТОВСКИЙ	404
16	РАЙОН КРАСНОАРМЕЙСКИЙ	381
23	РАЙОН ПОГРАНИЧНЫЙ	369
26	РАЙОН ТЕРНЕЙСКИЙ	284
17	РАЙОН ЛАЗОВСКИЙ	279
12	РАЙОН АНУЧИНСКИЙ	273
33	РАЙОН ЯКОВЛЕВСКИЙ	238
21	РАЙОН ОЛЬГИНСКИЙ	221
13	РАЙОН ДАЛЬНЕРЕЧЕНСКИЙ	159

In [15]:
```
# Считаем количество субъектов МСП (должно быть 84 335)
print(f'Данные содержат информацию о {sum(pivot["Наименование МСП"])} субъектов МСП.')
```

```
Данные содержат информацию о 84335 субъектов МСП.
```

2 ЗАГРУЗКА ДАННЫХ ПО ДОХОДАМ И РАСХОДАМ

Загружаем ранее сформированный CSV файл, содержащий сведения о доходах и расходах субъектов малого и среднего предпринимательства Приморского края.

In [16]:
```
# Создаем переменную names_2 с названиями колонок
names_2 = ['Наименование МСП', 'ИНН', 'Доход', 'Расход']

# Создадим датафрейм
company_revenue = pd.read_csv('company_revenue_2019.csv', # Указываем название файла
                 sep='[q]',                               # Указываем разделитель
                 encoding='utf-8',                        # Указываем кодировку
                 names=names_2,                           # Передаем названия колонок
                 engine='python')                         # Выбираем движок (python, так как в нем реал
изованно
                                                          # указание регулярного выражения в качестве р
азделителя)
# Выводим первые пять строк
company_revenue.head()
```

Out[16]:

	Наименование МСП	ИНН	Доход	Расход
0	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ФОРТЕ"	2536299160	14057000.0	5511000.0
1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "НОКС"	2511100506	1934000.0	1934000.0
2	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АВТО...	2540224873	149389000.0	148529000.0
3	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПРОГ...	2543105120	0.0	0.0
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АРГЕСТ"	2540224866	49000.0	1010000.0

Сформируем датафрейм для последующего объединения с реестром.

In [17]:
```
company_revenue = company_revenue[['ИНН', 'Доход', 'Расход']]
company_revenue.head()
```

Out[17]:

	ИНН	Доход	Расход
0	2536299160	14057000.0	5511000.0
1	2511100506	1934000.0	1934000.0
2	2540224873	149389000.0	148529000.0
3	2543105120	0.0	0.0
4	2540224866	49000.0	1010000.0

Выведем количество записей в датафрейме.

In [18]:
```
print(f'Данные содержат {len(company_revenue)} запись(ей).')
```

```
Данные содержат 33319 запись(ей).
```

3 ЗАГРУЗКА ДАННЫХ ПО СИСТЕМАМ НАЛОГООБЛОЖЕНИЯ

Загружаем ранее сформированный CSV файл, содержащий сведения о системах налогообложения, применяющихся субъектами малого и среднего предпринимательства Приморского края.

In [19]:
```
# Создаем переменную names_3 с названиями колонок
names_3 = ['Наименование МСП', 'ИНН', 'ЕСХН', 'УСН', 'ЕНВД', 'СРП']

# Создадим датафрейм
company_tax_regime = pd.read_csv('company_tax_regime_2019.csv', # Указываем название файла
                 sep='[q]',                                     # Указываем разделитель
                 encoding='utf-8',                              # Указываем кодировку
                 names=names_3,                                 # Передаем названия колонок
                 engine='python')                               # Выбираем движок (python, так как в не
м реализованно
                                                                # указание регулярного выражения в каче
стве разделителя)
# Выводим первые пять строк
company_tax_regime.head()
```

Out[19]:

	Наименование МСП	ИНН	ЕСХН	УСН	ЕНВД	СРП
0	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "СПЕК...	2508058043	0	1	0	0
1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "СНАЙ...	2507005666	0	1	1	0
2	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ АУДИТ...	2508052330	0	1	0	0
3	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "КОНЦ...	2508059390	0	1	0	0
4	ЗАКРЫТОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СТРОИТЕЛЬ-4"	2508016999	0	1	0	0

Сформируем датафрейм для последующего объединения с реестром.

In [20]:
```
company_tax_regime = company_tax_regime[['ИНН', 'ЕСХН', 'УСН', 'ЕНВД', 'СРП']]
company_tax_regime.head()
```

Out[20]:

	ИНН	ЕСХН	УСН	ЕНВД	СРП
0	2508058043	0	1	0	0
1	2507005666	0	1	1	0
2	2508052330	0	1	0	0
3	2508059390	0	1	0	0
4	2508016999	0	1	0	0

Выведем количество записей в датафрейме.

In [21]:
```
print(f'Данные содержат {len(company_tax_regime)} запись(ей).')
```

```
Данные содержат 21974 запись(ей).
```

4 ЗАГРУЗКА ДАННЫХ ПО НЕУПЛАЧЕННЫМ НАЛОГАМ

Загружаем ранее сформированный CSV файл, содержащий сведения о неуплаченных суммах налога субъектами малого и среднего предпринимательства Приморского края.

In [22]:
```
# Создаем переменную names_4 с названиями колонок
names_4 = ['Наименование МСП', 'ИНН', 'Налог', 'СумНедНалог', 'СумПени', 'СумШтраф', 'ОбщСумНедоим']

# Создадим датафрейм
df_2 = pd.read_csv('msp_arrears_2019.csv', # Указываем название файла
                 sep='[q]',                # Указываем разделитель
                 encoding='utf-8',         # Указываем кодировку
                 names=names_4,            # Передаем названия колонок
                 engine='python')          # Выбираем движок (python, так как в нем реализованно
                                           # указание регулярного выражения в качестве разделителя)
# Выводим первые пять строк
df_2.head()
```

Out[22]:

	Наименование МСП	ИНН	Налог	СумНедНалог	СумПени	СумШтраф	ОбщСумНедоим
0	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "НАДЕ...	2540221270	НЕНАЛОГОВЫЕ ДОХОДЫ, администрируемые налоговы...	0.00	0.00	1250.0	1250.00
1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "НАДЕ...	2540221270	Налог на имущество организаций	4923.00	215.74	0.0	5138.74
2	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "НАДЕ...	2540221270	Налог на добавленную стоимость	1847609.17	14581.66	0.0	1862190.83
3	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "НАДЕ...	2540221270	Налог на прибыль	290654.00	0.00	0.0	290654.00
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "НАДЕ...	2540221270	Страховые и другие взносы на обязательное пен...	16067203.09	1593577.48	0.0	17660780.57

Сформируем датафрейм для последующего объединения с реестром.

In [23]:
```
# Создадим сводную таблицу
# Групируем данные по ИНН
# В значение считаем сумму недоимки
msp_arrears = df_2.pivot_table(
                         index='ИНН',
                         columns='Налог',
                         values='ОбщСумНедоим',
                         aggfunc=sum).fillna(0).reset_index()
msp_arrears.head()
```

Out[23]:

Налог	ИНН	Акцизы, всего	Водный налог	Единый налог на вмененный доход для отдельных видов деятельности	Единый сельскохозяйственный налог	Задолженность и перерасчеты по ОТМЕНЕННЫМ НАЛОГАМ и сборам и иным обязательным платежам (кроме ЕСН, страх. Взносов)	Земельный налог	НЕНАЛОГОВЫЕ ДОХОДЫ, администрируемые налоговыми органами	Н добав сто
0	1831187536	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	2308178618	0.0	0.0	0.0	0.0	0.0	0.0	200.0	
2	2311228114	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19
3	2320178371	0.0	0.0	8045.7	0.0	0.0	0.0	0.0	
4	2461031942	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Переименуем название колонок. Выведем для удобства текущие названия колонок.

In [24]:
```
msp_arrears.columns
```

Out[24]:
```
Index(['ИНН', ' Акцизы, всего', ' Водный налог',
       ' Единый налог на вмененный доход для отдельных видов  деятельности',
       ' Единый сельскохозяйственный налог',
       ' Задолженность и перерасчеты по ОТМЕНЕННЫМ НАЛОГАМ  и сборам и иным обязательным платежам  (к
роме ЕСН, страх. Взносов)',
       ' Земельный налог',
       ' НЕНАЛОГОВЫЕ ДОХОДЫ, администрируемые налоговыми органами',
       ' Налог на добавленную стоимость',
       ' Налог на добычу полезных ископаемых',
       ' Налог на доходы физических лиц', ' Налог на имущество организаций',
       ' Налог на прибыль',
       ' Налог, взимаемый в связи с  применением упрощенной  системы налогообложения',
       ' Сборы за пользование объектами животного мира  и за пользование объектами ВБР',
       ' Страховые взносы на обязательное медицинское страхование работающего населения, зачисляемые
в бюджет Федерального фонда обязательного медицинского страхования',
       ' Страховые взносы на обязательное социальное страхование на случай временной нетрудоспособнос
ти и в связи с материнством',
       ' Страховые и другие взносы на обязательное пенсионное страхование, зачисляемые в Пенсионный ф
онд Российской Федерации',
       ' Торговый сбор', ' Транспортный налог'],
      dtype='object', name='Налог')
```

In [25]:
```
msp_arrears.columns = ['ИНН',
                       'Акцизы н',
                       'Водный налог н',
                       'ЕНВД н',
                       'ЕСХН н',
                       'Задолженность и перерасчеты по отмененным налогам н',
                       'Земельный налог н',
                       'Неналоговые доходы, администрируемые налоговыми органами н',
                       'НДС н',
                       'НДПИ н',
                       'НДФЛ н',
                       'Налог на имущество организаций н',
                       'Налог на прибыль н',
                       'УСН н',
                       'Сборы за пользование объектами животного мира и за пользование объектами ВБР н'
,
                       'Страховые взносы мед н',
                       'Страховые взносы соц н',
                       'Страховые и другие взносы пенс н',
                       'Торговый сбор н',
                       'Транспортный налог н']
msp_arrears.head()
```

Out[25]:

	ИНН	Акцизы н	Водный налог н	ЕНВД н	ЕСХН н	Задолженность и перерасчеты по отмененным налогам н	Земельный налог н	Неналоговые доходы, администрируемые налоговыми органами н	НДС н	НДПИ н	НДФЛ н	Налог имущест организа
0	1831187536	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	
1	2308178618	0.0	0.0	0.0	0.0	0.0	0.0	200.0	0.00	0.0	0.0	
2	2311228114	0.0	0.0	0.0	0.0	0.0	0.0	0.0	197773.77	0.0	0.0	
3	2320178371	0.0	0.0	8045.7	0.0	0.0	0.0	0.0	1000.05	0.0	3727.0	
4	2461031942	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	

Выведем количество записей в датафрейме.

In [26]:
```
print(f'Данные содержат {len(msp_arrears)} запись(ей).')
```

```
Данные содержат 7922 запись(ей).
```

5 ЗАГРУЗКА ДАННЫХ ПО УПЛАЧЕННЫМ НАЛОГАМ

Загружаем ранее сформированный CSV файл, содержащий сведения об уплаченных суммах налогов субъектми малого и среднего предпринимательства Приморского края.

In [27]:
```
# Создаем переменную names_5 с названиями колонок
names_5 = ['Наименование МСП', 'ИНН', 'Налог', 'СумУплНал']

# Создадим датафрейм
df_3 = pd.read_csv('msp_tax_2019.csv', # Указываем название файла
                 sep='[q]',            # Указываем разделитель
                 encoding='utf-8',     # Указываем кодировку
                 names=names_5,        # Передаем названия колонок
                 engine='python')      # Выбираем движок (python, так как в нем реализованно
                                       # указание регулярного выражения в качестве разделителя)
# Выводим первые пять строк
df_3.head()
```

Out[27]:

	Наименование МСП	ИНН	Налог	СумУплНал
0	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ФЛАГ...	2543077218	Налог на добавленную стоимость	0.0
1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ФЛАГ...	2543077218	Налог на прибыль	0.0
2	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ФЛАГ...	2543077218	Страховые и другие взносы на обязательное пен...	0.0
3	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ФЛАГ...	2543077218	НЕНАЛОГОВЫЕ ДОХОДЫ, администрируемые налоговы...	0.0
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "СЕМЕ...	2501017111	Страховые взносы на обязательное социальное с...	346437.0

Сформируем датафрейм для последующего объединения с реестром.

In [28]:
```
# Создадим сводную таблицу
# Групируем данные по ИНН
# В значение считаем сумму уплаченного налога
msp_tax = df_3.pivot_table(
                         index='ИНН',
                         columns='Налог',
                         values='СумУплНал',
                         aggfunc=sum).fillna(0).reset_index()
msp_tax.head()
```

Out[28]:

Налог	ИНН	Акцизы, всего	Водный налог	Единый налог на вмененный доход для отдельных видов деятельности	Единый сельскохозяйственный налог	Задолженность и перерасчеты по ОТМЕНЕННЫМ НАЛОГАМ и сборам и иным обязательным платежам (кроме ЕСН, страх. Взносов)	Земельный налог	НЕНАЛОГОВЫЕ ДОХОДЫ, администрируемые налоговыми органами	Н добав сто
0	814069428	0.0	0.0	0.0	0.0	0.0	4553.00	0.0	
1	1435260214	0.0	0.0	0.0	0.0	0.0	0.00	0.0	
2	1701050844	0.0	0.0	0.0	0.0	0.0	0.00	0.0	62
3	1831187536	0.0	0.0	0.0	0.0	0.0	221487.77	0.0	
4	2308178618	0.0	0.0	0.0	0.0	0.0	91889.00	0.0	

5 rows × 22 columns

Переименуем название колонок. Выведем для удобства текущие названия колонок.

In [29]:
```
msp_tax.columns
```

Out[29]:
```
Index(['ИНН', ' Акцизы, всего', ' Водный налог',
       ' Единый налог на вмененный доход для отдельных видов  деятельности',
       ' Единый сельскохозяйственный налог',
       ' Задолженность и перерасчеты по ОТМЕНЕННЫМ НАЛОГАМ  и сборам и иным обязательным платежам  (к
роме ЕСН, страх. Взносов)',
       ' Земельный налог',
       ' НЕНАЛОГОВЫЕ ДОХОДЫ, администрируемые налоговыми органами',
       ' Налог на добавленную стоимость',
       ' Налог на добычу полезных ископаемых',
       ' Налог на доходы физических лиц', ' Налог на игорный',
       ' Налог на имущество организаций', ' Налог на прибыль',
       ' Налог, взимаемый в связи с  применением упрощенной  системы налогообложения',
       ' Сборы за пользование объектами животного мира  и за пользование объектами ВБР',
       ' Страховые взносы на обязательное медицинское страхование работающего населения, зачисляемые
в бюджет Федерального фонда обязательного медицинского страхования',
       ' Страховые взносы на обязательное социальное страхование на случай временной нетрудоспособнос
ти и в связи с материнством',
       ' Страховые и другие взносы на обязательное пенсионное страхование, зачисляемые в Пенсионный ф
онд Российской Федерации',
       ' Торговый сбор', ' Транспортный налог', ' Утилизационный сбор'],
      dtype='object', name='Налог')
```

In [30]:
```python
msp_tax.columns = ['ИНН',
                   'Акцизы у',
                   'Водный налог у',
                   'ЕНВД у',
                   'ЕСХН у',
                   'Задолженность и перерасчеты по отмененным налогам у',
                   'Земельный налог у',
                   'Неналоговые доходы, администрируемые налоговыми органами у',
                   'НДС у',
                   'НДПИ у',
                   'НДФЛ у',
                   'Налог на игорный бизнес у',
                   'Налог на имущество организаций у',
                   'Налог на прибыль у',
                   'УСН у',
                   'Сборы за пользование объектами животного мира и за пользование объектами ВБР у',
                   'Страховые взносы мед у',
                   'Страховые взносы соц у',
                   'Страховые и другие взносы пенс у',
                   'Торговый сбор у',
                   'Транспортный налог у',
                   'Утилизационный сбор у']
msp_tax.head()
```
Out[30]:

	ИНН	Акцизы у	Водный налог у	ЕНВД у	ЕСХН у	Задолженность и перерасчеты по отмененным налогам у	Земельный налог у	Неналоговые доходы, администрируемые налоговыми органами у	НДС у	НДПИ у	...	Налог на имущество организаций у
0	814069428	0.0	0.0	0.0	0.0	0.0	4553.00	0.0	0.0	0.0	...	0.0
1	1435260214	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	...	0.0
2	1701050844	0.0	0.0	0.0	0.0	0.0	0.00	0.0	6270866.5	0.0	...	4956.0
3	1831187536	0.0	0.0	0.0	0.0	0.0	221487.77	0.0	0.0	0.0	...	0.0
4	2308178618	0.0	0.0	0.0	0.0	0.0	91889.00	0.0	0.0	0.0	...	0.0

5 rows × 22 columns

Выведем количество записей в датафрейме.

In [31]:
```python
print(f'Данные содержат {len(msp_tax)} запись(ей).')
```
```
Данные содержат 35103 запись(ей).
```

6 ЗАГРУЗКА ДАННЫХ ПО МСП БЕЗ УКАЗАНИЯ ОКВЭД

Загружаем ранее сформированный CSV файл, содержащий сведения о субъектах малого и среднего предпринимательства Приморского края, у которых не указан основной вид деятельности.

In [32]:
```python
# Создаем переменную names_5 с названиями колонок
names_6 = ['Наименование МСП', 'ИНН', 'Признак']

# Создадим датафрейм
activity_list = pd.read_csv('activity_list.csv', # Указываем название файла
                  sep='[q]',                     # Указываем разделитель
                  encoding='utf-8',              # Указываем кодировку
                  names=names_6,                 # Передаем названия колонок
                  engine='python')               # Выбираем движок (python, так как в нем реализованно
                                                 # указание регулярного выражения в качестве разделител
я)
# Выводим первые пять строк
activity_list.head()
```
Out[32]:

	Наименование МСП	ИНН	Признак
0	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "КЛУБ...	2502029222	Оквэд не указан
1	"КРОВЛЯ" ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ	2536013862	Оквэд не указан
2	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЦЕНТ...	2536005773	Оквэд не указан
3	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АВИОР"	2536023155	Оквэд не указан
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ НАУЧН...	2536006174	Оквэд не указан

Сформируем датафрейм для последующего объединения с реестром.

In [33]:
```python
activity_list = activity_list[['ИНН', 'Признак']]
activity_list.head()
```
Out[33]:

	ИНН	Признак
0	2502029222	Оквэд не указан
1	2536013862	Оквэд не указан
2	2536005773	Оквэд не указан
3	2536023155	Оквэд не указан
4	2536006174	Оквэд не указан

Выведем количество записей в датафрейме.

In [34]:
```python
print(f'Данные содержат {len(activity_list)} запись(ей).')
```
```
Данные содержат 135 запись(ей).
```

7 ФОРМИРОВАНИЕ ИТОГОВОГО ДАТАСЕТА

Соберем все данные в единый датасет.

In [35]:
```python
# Создадм единый датафрейм
data = reestr.merge(company_revenue, how='left')\
                 .merge(company_tax_regime, how='left')\
                 .merge(msp_arrears, how='left')\
                 .merge(msp_tax, how='left')\
                 .merge(activity_list, how='left')
# Выводим первые пять строк
data.head()
```
Out[35]:

	Наименование МСП	ИНН	Муниципальное образование	Вид	Категория	Код ОКВЭД	Вид деятельности	Доход	Расхо
0	СТЕПАНОВА ИРИНА	644934682794	ГОРОД ФОКИНО	Индивидуальный предприниматель	Микропредприятие	95.21	Ремонт электронной бытовой техники	NaN	Na
1	КУРКИН ЕВГЕНИЙ	870100455199	ГОРОД УССУРИЙСК	Индивидуальный предприниматель	Микропредприятие	49.32	Деятельность легкового такси и арендованных ле...	NaN	Na
2	ГОДУНОВА СВЕТЛАНА	870600600221	ГОРОД ВЛАДИВОСТОК	Индивидуальный предприниматель	Микропредприятие	47.11	Торговля розничная преимущественно пищевыми пр...	NaN	Na
3	НАВРОСЬ ДМИТРИЙ	870600537308	ГОРОД ВЛАДИВОСТОК	Индивидуальный предприниматель	Микропредприятие	95.29	Ремонт прочих предметов личного потребления и ...	NaN	Na
4	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АГРО...	7017227901	ГОРОД ВЛАДИВОСТОК	Юридическое лицо	Микропредприятие	01.25.1	Выращивание прочих плодовых и ягодных культур	0.0	3000.

5 rows × 54 columns

Проверим, все ли даные корректно попали в итоговый датафрейм.

In [36]:
```python
data.info()
```
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 84335 entries, 0 to 84334
Data columns (total 54 columns):
 #   Column                                                                         Non-Null Count
Dtype
---  ------                                                                         --------------
-----
 0   Наименование МСП                                                               84335 non-null
object
 1   ИНН                                                                            84335 non-null
int64
 2   Муниципальное образование                                                      84335 non-null
object
 3   Вид                                                                            84335 non-null
object
 4   Категория                                                                      84335 non-null
object
 5   Код ОКВЭД                                                                      84335 non-null
object
 6   Вид деятельности                                                               84335 non-null
object
 7   Доход                                                                          33319 non-null
float64
 8   Расход                                                                         33319 non-null
float64
 9   ЕСХН                                                                           21974 non-null
float64
 10  УСН                                                                            21974 non-null
float64
 11  ЕНВД                                                                           21974 non-null
float64
 12  СРП                                                                            21974 non-null
float64
 13  Акцизы н                                                                       7922 non-null
float64
 14  Водный налог н                                                                 7922 non-null
float64
 15  ЕНВД н                                                                         7922 non-null
float64
 16  ЕСХН н                                                                         7922 non-null
float64
 17  Задолженность и перерасчеты по отмененным налогам н                            7922 non-null
float64
 18  Земельный налог н                                                              7922 non-null
float64
 19  Неналоговые доходы, администрируемые налоговыми органами н                     7922 non-null
float64
 20  НДС н                                                                          7922 non-null
float64
 21  НДПИ н                                                                         7922 non-null
float64
 22  НДФЛ н                                                                         7922 non-null
float64
 23  Налог на имущество организаций н                                               7922 non-null
float64
 24  Налог на прибыль н                                                             7922 non-null
float64
 25  УСН н                                                                          7922 non-null
float64
 26  Сборы за пользование объектами животного мира и за пользование объектами ВБР н 7922 non-null
float64
 27  Страховые взносы мед н                                                         7922 non-null
float64
 28  Страховые взносы соц н                                                         7922 non-null
float64
 29  Страховые и другие взносы пенс н                                               7922 non-null
float64
 30  Торговый сбор н                                                                7922 non-null
float64
 31  Транспортный налог н                                                           7922 non-null
float64
 32  Акцизы у                                                                       35103 non-null
float64
 33  Водный налог у                                                                 35103 non-null
float64
 34  ЕНВД у                                                                         35103 non-null
float64
 35  ЕСХН у                                                                         35103 non-null
float64
 36  Задолженность и перерасчеты по отмененным налогам у                            35103 non-null
float64
 37  Земельный налог у                                                              35103 non-null
float64
 38  Неналоговые доходы, администрируемые налоговыми органами у                     35103 non-null
float64
 39  НДС у                                                                          35103 non-null
float64
 40  НДПИ у                                                                         35103 non-null
float64
 41  НДФЛ у                                                                         35103 non-null
float64
 42  Налог на игорный бизнес у                                                      35103 non-null
float64
 43  Налог на имущество организаций у                                               35103 non-null
float64
 44  Налог на прибыль у                                                             35103 non-null
float64
 45  УСН у                                                                          35103 non-null
float64
 46  Сборы за пользование объектами животного мира и за пользование объектами ВБР у 35103 non-null
float64
 47  Страховые взносы мед у                                                         35103 non-null
float64
 48  Страховые взносы соц у                                                         35103 non-null
float64
 49  Страховые и другие взносы пенс у                                               35103 non-null
float64
 50  Торговый сбор у                                                                35103 non-null
float64
 51  Транспортный налог у                                                           35103 non-null
float64
 52  Утилизационный сбор у                                                          35103 non-null
float64
 53  Признак                                                                        135 non-null
object
dtypes: float64(46), int64(1), object(7)
memory usage: 35.4+ MB
```

Разделим датафрейм на два (создадим глубокие копии датафрейма). Один будет содержать информацию по юридическим лицам второй по индивидуальным предпринимателям.

In [37]:
```python
legal_entity = copy.deepcopy(data[data['Вид'] == 'Юридическое лицо'])
entrepreneur = copy.deepcopy(data[data['Вид'] == 'Индивидуальный предприниматель'])
```

Посмотрим еще раз на данные.

In [38]:
```python
legal_entity.info()
```
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38366 entries, 4 to 83991
Data columns (total 54 columns):
 #   Column                                                                         Non-Null Count
Dtype
---  ------                                                                         --------------
-----
 0   Наименование МСП                                                               38366 non-null
object
 1   ИНН                                                                            38366 non-null
int64
 2   Муниципальное образование                                                      38366 non-null
object
 3   Вид                                                                            38366 non-null
object
 4   Категория                                                                      38366 non-null
object
 5   Код ОКВЭД                                                                      38366 non-null
object
 6   Вид деятельности                                                               38366 non-null
object
 7   Доход                                                                          33319 non-null
float64
 8   Расход                                                                         33319 non-null
float64
 9   ЕСХН                                                                           21974 non-null
float64
 10  УСН                                                                            21974 non-null
float64
 11  ЕНВД                                                                           21974 non-null
float64
 12  СРП                                                                            21974 non-null
float64
 13  Акцизы н                                                                       7922 non-null
float64
 14  Водный налог н                                                                 7922 non-null
float64
 15  ЕНВД н                                                                         7922 non-null
float64
 16  ЕСХН н                                                                         7922 non-null
float64
 17  Задолженность и перерасчеты по отмененным налогам н                            7922 non-null
float64
 18  Земельный налог н                                                              7922 non-null
float64
 19  Неналоговые доходы, администрируемые налоговыми органами н                     7922 non-null
float64
 20  НДС н                                                                          7922 non-null
float64
 21  НДПИ н                                                                         7922 non-null
float64
 22  НДФЛ н                                                                         7922 non-null
float64
 23  Налог на имущество организаций н                                               7922 non-null
float64
 24  Налог на прибыль н                                                             7922 non-null
float64
 25  УСН н                                                                          7922 non-null
float64
 26  Сборы за пользование объектами животного мира и за пользование объектами ВБР н 7922 non-null
float64
 27  Страховые взносы мед н                                                         7922 non-null
float64
 28  Страховые взносы соц н                                                         7922 non-null
float64
 29  Страховые и другие взносы пенс н                                               7922 non-null
float64
 30  Торговый сбор н                                                                7922 non-null
float64
 31  Транспортный налог н                                                           7922 non-null
float64
 32  Акцизы у                                                                       35103 non-null
float64
 33  Водный налог у                                                                 35103 non-null
float64
 34  ЕНВД у                                                                         35103 non-null
float64
 35  ЕСХН у                                                                         35103 non-null
float64
 36  Задолженность и перерасчеты по отмененным налогам у                            35103 non-null
float64
 37  Земельный налог у                                                              35103 non-null
float64
 38  Неналоговые доходы, администрируемые налоговыми органами у                     35103 non-null
float64
 39  НДС у                                                                          35103 non-null
float64
 40  НДПИ у                                                                         35103 non-null
float64
 41  НДФЛ у                                                                         35103 non-null
float64
 42  Налог на игорный бизнес у                                                      35103 non-null
float64
 43  Налог на имущество организаций у                                               35103 non-null
float64
 44  Налог на прибыль у                                                             35103 non-null
float64
 45  УСН у                                                                          35103 non-null
float64
 46  Сборы за пользование объектами животного мира и за пользование объектами ВБР у 35103 non-null
float64
 47  Страховые взносы мед у                                                         35103 non-null
float64
 48  Страховые взносы соц у                                                         35103 non-null
float64
 49  Страховые и другие взносы пенс у                                               35103 non-null
float64
 50  Торговый сбор у                                                                35103 non-null
float64
 51  Транспортный налог у                                                           35103 non-null
float64
 52  Утилизационный сбор у                                                          35103 non-null
float64
 53  Признак                                                                        135 non-null
object
dtypes: float64(46), int64(1), object(7)
memory usage: 16.1+ MB
```

In [39]:
```python
entrepreneur.info()
```
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 45969 entries, 0 to 84334
Data columns (total 54 columns):
 #   Column                                                                         Non-Null Count
Dtype
---  ------                                                                         --------------
-----
 0   Наименование МСП                                                               45969 non-null
object
 1   ИНН                                                                            45969 non-null
int64
 2   Муниципальное образование                                                      45969 non-null
object
 3   Вид                                                                            45969 non-null
object
 4   Категория                                                                      45969 non-null
object
 5   Код ОКВЭД                                                                      45969 non-null
object
 6   Вид деятельности                                                               45969 non-null
object
 7   Доход                                                                          0 non-null
float64
 8   Расход                                                                         0 non-null
float64
 9   ЕСХН                                                                           0 non-null
float64
 10  УСН                                                                            0 non-null
float64
 11  ЕНВД                                                                           0 non-null
float64
 12  СРП                                                                            0 non-null
float64
 13  Акцизы н                                                                       0 non-null
float64
 14  Водный налог н                                                                 0 non-null
float64
 15  ЕНВД н                                                                         0 non-null
float64
 16  ЕСХН н                                                                         0 non-null
float64
 17  Задолженность и перерасчеты по отмененным налогам н                            0 non-null
float64
 18  Земельный налог н                                                              0 non-null
float64
 19  Неналоговые доходы, администрируемые налоговыми органами н                     0 non-null
float64
 20  НДС н                                                                          0 non-null
float64
 21  НДПИ н                                                                         0 non-null
float64
 22  НДФЛ н                                                                         0 non-null
float64
 23  Налог на имущество организаций н                                               0 non-null
float64
 24  Налог на прибыль н                                                             0 non-null
float64
 25  УСН н                                                                          0 non-null
float64
 26  Сборы за пользование объектами животного мира и за пользование объектами ВБР н 0 non-null
float64
 27  Страховые взносы мед н                                                         0 non-null
float64
 28  Страховые взносы соц н                                                         0 non-null
float64
 29  Страховые и другие взносы пенс н                                               0 non-null
float64
 30  Торговый сбор н                                                                0 non-null
float64
 31  Транспортный налог н                                                           0 non-null
float64
 32  Акцизы у                                                                       0 non-null
float64
 33  Водный налог у                                                                 0 non-null
float64
 34  ЕНВД у                                                                         0 non-null
float64
 35  ЕСХН у                                                                         0 non-null
float64
 36  Задолженность и перерасчеты по отмененным налогам у                            0 non-null
float64
 37  Земельный налог у                                                              0 non-null
float64
 38  Неналоговые доходы, администрируемые налоговыми органами у                     0 non-null
float64
 39  НДС у                                                                          0 non-null
float64
 40  НДПИ у                                                                         0 non-null
float64
 41  НДФЛ у                                                                         0 non-null
float64
 42  Налог на игорный бизнес у                                                      0 non-null
float64
 43  Налог на имущество организаций у                                               0 non-null
float64
 44  Налог на прибыль у                                                             0 non-null
float64
 45  УСН у                                                                          0 non-null
float64
 46  Сборы за пользование объектами животного мира и за пользование объектами ВБР у 0 non-null
float64
 47  Страховые взносы мед у                                                         0 non-null
float64
 48  Страховые взносы соц у                                                         0 non-null
float64
 49  Страховые и другие взносы пенс у                                               0 non-null
float64
 50  Торговый сбор у                                                                0 non-null
float64
 51  Транспортный налог у                                                           0 non-null
float64
 52  Утилизационный сбор у                                                          0 non-null
float64
 53  Признак                                                                        0 non-null
object
dtypes: float64(46), int64(1), object(7)
memory usage: 19.3+ MB
```

Так как в итоговом датафрейме информация по индивидуальни предпринимателям в большинстве отсутствует, будем проводить последующий анализ по данным о юридических лицах. Кроме того приведем информацию по видам деятельности в соответствие. У субъектов малого и среднего предпринимательства, где основной вид экономической деятельности не указан проставим соответствующий признак.

In [40]:
```python
legal_entity.loc[legal_entity['Признак'] == 'Оквэд не указан', 'Код ОКВЭД'] = 'Не указан'
legal_entity.loc[legal_entity['Признак'] == 'Оквэд не указан', 'Вид деятельности'] = 'Не указан'
```

Больше колонка признак не понадобится, удалим ее.

In [41]:
```python
del legal_entity['Признак']
```

Выгрузим итоговый датасет.

In [42]:
```python
legal_entity.to_csv('data_msp.csv', index=False, encoding='utf-8')
```