Task

Charging and Gas Station Data Cleaning

1. Make a geospatial ID list, which contains ID, longitude, latitude, first year observed, and last year observed.
2. Aggregated data structure (columns)

Year	Province	City	GS	CS	GCS	entry of GS	entry of CS	exit of GS	exit of CS	switch GS -> CS	switch CS -> GS	CS (added to existing GS)	GS (added to existing CS)

Process

- How to build ID: baed on pname(province), cityname(city), adname(ad division), wgs84_x(longitude), wgs84_y(latitute)
- But if there is any problem with longitude and latitude ID?
- Longitude and latitude precision: 1e-6 (40,000km / 360 * 1e-6 = 0.11meter)
- If it is exactly same, must be the same CS or GS.
- But if it is not exactly same, there is small differece, what's the case(discussed later)

Statistics

- charging station count for each year

```
2015 2039
2016 2092
2017 4533
2018 33972
2019 57113
2020 73320
2021 94209
2022 98813
2023 120660
2024 185617
2025 225899
```

- check:if there is any duplicate x and y; check if x and y are all in reasonable range(China's Longitude range:73.55-135.08, Latitude range:3.85-53.55)
- In most years, the duplicated data is acceptable(I checked the data, and they are same CS or GS)

```
Year 2015:
  Total rows: 2039  Duplicate rows: 98
Year 2016:
  Total rows: 2092  Duplicate rows: 202
Year 2017:
  Total rows: 4533  Duplicate rows: 98
Year 2018:
  Total rows: 33972  Duplicate rows: 820
Year 2019:
  Total rows: 57113  Duplicate rows: 1391
Year 2020:
  Total rows: 73320  Duplicate rows: 1393
Year 2021:
  Total rows: 94209  Duplicate rows: 94209
Year 2022:
  Total rows: 98813  Duplicate rows: 3952
Year 2023:
  Total rows: 120660  Duplicate rows: 4568
Year 2024:
  Total rows: 185617  Duplicate rows: 2855
Year 2025:
  Total rows: 225899  Duplicate rows: 6796
```

```
Year 2013:
  Total rows: 101816  Duplicate rows: 480
Year 2014:
  Total rows: 104542  Duplicate rows: 472
Year 2015:
  Total rows: 118645  Duplicate rows: 6270
Year 2016:
  Total rows: 120030  Duplicate rows: 8998
Year 2017:
  Total rows: 120669  Duplicate rows: 684
Year 2018:
  Total rows: 107356  Duplicate rows: 415
Year 2019:
  Total rows: 113770  Duplicate rows: 167
Year 2020:
  Total rows: 120313  Duplicate rows: 50
Year 2021:
  Total rows: 122005  Duplicate rows: 122005
Year 2022:
  Total rows: 111608  Duplicate rows: 472
Year 2023:
  Total rows: 119012  Duplicate rows: 50
Year 2024:
  Total rows: 119029  Duplicate rows: 48
Year 2025:
  Total rows: 107755  Duplicate rows: 612
```

- There is something wrong with 2021 data(there isn't wgs84_x and wgs84_y),try to use 2020 and 2022 data to fill
- use [pname, cityname, adname, address, name] to match but can only fill less than half

```
Filling year 2021 data, total rows: 94209, rows with missing coordinates: 94209
Filled 38579 rows from previous year (2020)
Filled 6920 rows from next year (2022)
```

Problem

1. 2021 data: all wgs84_x and wgs84_y are empty

2. even the same charging station, wgs84_x and wgs84_y in different year have slight difference(which means hard to set id, so exit, entry, switch, add are all hard to calculate)

```
filter_2019 = (CS_data_set[2019]['name'] == '星星充电汽车充电站(亳州之星汽车有限公司)')
CS_data_set[2019][filter_2019]
```

✓ 0.0s Python

	id	name	address	# wgs84_x	# wgs84_y	tel	pname	cityname	adname	大类	中...	小类
1	B0FFGXT8LP	星星充电汽车充电站(亳州之星汽车有限公司)	药都大道西38号	115.796044848	33.8399597577	[]	安徽省	亳州市	谯城区	汽车服务	充电站	充电站

1 rows x 12 cols 10 per page Page 1 of 1

```
filter_2020 = (CS_data_set[2020]['name'] == '星星充电汽车充电站(亳州之星汽车有限公司)')
CS_data_set[2020][filter_2020]
```

✓ 0.0s Python

	id	name	address	# wgs84_x	# wgs84_y	tel	pname	cityname	adname	大类	中...	小类
9458	B0FFGXT8LP	星星充电汽车充电站(亳州之星汽车有限公司)	药都大道西38号	115.7960576397839	33.8399686633858	[]	安徽省	亳州市	谯城区	汽车服务	充电站	充电站

3. if use [pname, cityname, adname, name, address] to merge, the merged ratio is much lower than expected(41%)

```
data_2021_filled_CS = fill_specific_year_x_y_data_efficient(CS_data_set, 2021, 'wgs84_x', 'wgs84_y', matching_cols=['pname', 'cityname','adname','address','name'])
data_2021_filled_CS
```

✓ 0.3s 在 Data Wrangler 中打开"data_2021_filled_CS"

```
Filling year 2021 data, total rows: 94209, rows with missing coordinates: 94209
Filled 38579 rows from previous year (2020)
Filled 6920 rows from next year (2022)
```

Result

- seperate cs & gs count by city and year(cannot be merged because x and y problem)
- small problem about pname empty(Beijing, Shanghai, Chongqin, etc)(solved)
- original_aggregated data wsa storaged as "GS_aggregated_original", "CS_aggregated_original"
- But there is still some problem: about 4k and 3k rows: manual solved

330	2013	香港特	香港特	207
1709	2017	香港特别行	香港特别行	203
674	2014	香港特别行政区	香港特别行政区	197
1019	2015	香港特别行政区	香港特别行政区	222
1364	2016	香港特别行政区	香港特别行政区	225
2076	2018	香港特别行政区	香港特别行政区	253
2444	2019	香港特别行政区	香港特别行政区	190
2813	2020	香港特别行政区	香港特别行政区	178
3183	2021	香港特别行政区	香港特别行政区	177
3552	2022	香港特别行政区	香港特别行政区	178
3922	2023	香港特别行政区	香港特别行政区	172
4292	2024	香港特别行政区	香港特别行政区	173
4628	2025	香港特别行政区	香港特别行政区	178

- The political region is not stable for some province(liake , 内蒙古, 自治区自治州), but this part is quite hard to solve
- after manual check, sort again and calculate change between years
- change the Chinese name to English version